MONASH University

Computational modelling and characterisation of cell signalling cross-talks in acquired drug resistance

A. K. M. Azad

M.Sc. in Computational Biology, Gwangju Institute of Science & Technology, South Korea

A thesis submitted for the degree of Doctor of Philosophy

at the School of Mathematical Sciences

Monash Univeristy, Australia

February 2017

Contents

Copyright Notice

Abstract

Initial drug efficacy in controlling cancerous growth and proliferation often recedes as cells acquire resistance mechanisms to escape from the inhibitory effects of RTK (receptor tyrosine kinase)-targeted therapies such as EGFR (epidermal growth factor receptor)-TKI (tyrosine kinase inhibitor). In addition to secondary mutations in targeted oncogenes, signalling cross-talk (interactions among signalling pathways) has been reported to play a vital role as a molecular mechanism of this significant clinical barrier. Therefore, systematic modelling, identification, and characterisation of putative signalling cross-talk in demystifying underlying mechanisms of acquired resistance to EGFR-TKIs *in silico* has become increasingly urgent. In this thesis, I developed a framework combining computational modelling with a fully Bayesian statistical approach to identify perturbations of underlying signalling networks distinguishing between drug-sensitive and drug-resistant conditions. I inferred data-driven signalling networks by analysing gene expression datasets of two breast cancer cell-lines: SKBR3 and BT474 in lapatinib (an EGFR/HER2 (human epidermal growth factor receptor 2) dual inhibitor) treated sensitive (parental) and resistant conditions, and inferred aberrant signalling pairs in resistant-vs-parental conditions using a particular class of Exponential Random Graph Models (ERGMs), called p_1-models. I hypothesised that such aberrant signalling pairs might possess differential probabilities of appearing between the data-driven signalling networks from resistant-vs-parental conditions. I proposed a novel cross-talk categorisation for data-driven signalling networks (Type-I and Type-II cross-talk) and observed that many *compensatory* signalling pathways in SKBR3 and BT474 cell-lines aberrantly cross-talk with EGFR/HER2 signalling

pathway, which is the primary target of lapatinib. In both SKBR3 and BT474 cell-lines, pathway enrichment tests of aberrant pairs with known signalling links revealed that those compensatory pathways from KEGG, Reactome, and WikiPathway databases were significantly *dysregulated* in acquired resistance. Moreover, I proposed and analysed a novel structure of aberrant signalling links, called *V-structures*, and found that many genes were dysregulated in resistant-vs-parental conditions when they were involved in the *dependency switch* from *targeted* to *bypass* signalling events. These results provide further insights into the bypass mechanisms of acquired resistance and have potential to be used in designing novel therapeutics to overcome acquired resistance in cancer.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: ……  …….

Print Name: …..**(A. K. M. Azad)**…….

Date: ……**(01 / 06 / 2017)**…….

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 3 original papers published in peer reviewed journals, 1 conference paper, 1 book chapter and 1 submitted articles. The core theme of the thesis is modelling and characterising signalling cross-talks and elucidating their roles in acquired drug resistance using Bayesian statistical modelling. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Mathematical Science at Monash University under the supervision of A/Prof. Jonathan M Keith, Dr. Alfons Lawen and A/Prof. Tianhai Tian.

In the case of Chapter 3, 4, 5, and 6 my contribution to the work involved the following:

Thesis Chapter	**Publication Title**	**Status** *(published, in press, accepted or returned for revision)*	**Nature and % of student contribution**	**Co-author name(s) Nature and % of Co-author's contribution**	**Co-author(s), Monash student Y/N**
3	*Cross-talk categorisations in data-driven models of signalling networks: a system-level view*	*Submitted (Under review)*	*80%. Concept, collecting data, proposing method and writing manuscript*	*1) Jonathan Keith: 15% (supervised and input into algorithm design)* *2) Alfons Lawen: 5% (Supervised the work)*	*No*
4	*Prediction of signalling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modelling*	*Published*	*70%. Concept, collecting data, proposing and implementing method, analysing data and results, writing manuscript*	*1) Jonathan Keith: 20% (supervised and validated the model, and proofread manuscript)* *2) Alfons Lawen: 10% (supervised and approved validation of results, and proofread manuscript)*	*No*
5	*Bayesian Model of Signal Rewiring Reveals Mechanisms of Gene Dysregulation in acquired Drug Resistance in Breast Cancer*	*Published*	*70%. Concept, collecting data, proposing and implementing method, analysing data and results, writing manuscript*	*1) Jonathan Keith: 20% (supervised and validated the model, and proofread manuscript)* *2) Alfons Lawen: 10% (supervised and approved validation of results, and proofread manuscript and editorial work)*	*No*

6	*Integrating heterogeneous datasets for cancer module identification*	*Published*	*95%. Concept, collecting data and writing manuscript*	*1) Jonathan Keith: 5% (Supervision and editorial work)*	*No*
6	*Uniform Sampling of Directed and Undirected Graphs Conditional on Vertex Connectivity*	*Published*	*30%. Implementing the model, analysing the results and proofreading the manuscript*	*1) Salem A. Alyami: 60% (Concept, collecting data, designed experiment and wrote manuscript) 2) Jonathan Keith: 10% (Supervised the work and editorial work)*	*Yes*
6	*The Neighborhood MCMC sampler for learning Bayesian networks*	*Published*	*30%. Implementing the model and analysing the results*	*1) Salem A. Alyami: 60% (Concept, collecting data, designed experiment and wrote manuscript) 2) Jonathan Keith: 10% (Supervised the work and editorial work)*	*Yes*

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: … [redacted] **……** **Date: 01 / 06 / 2017**

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: … [redacted] **……** **Date: 01 / 06 / 2017**

Acknowledgements

First and foremost I would like to thank my Lord, the Almighty Allah (SWT) for giving me a chance to breathe till now and providing me with enough sustenance and strength to continue my research work.

I would like to express my sincere thanks to my supervisor A/Prof. Jonathan M Keith for supervising me with his immense knowledge, careful guidance, insightful suggestions, motivating presence and friendly attitude. His encouragement, absolute faith in me, and the allowance of abundant freedom throughout my PhD candidature not only assisted me in my tough time but also taught me how to do research independently. I would thank my associate supervisor, Dr Alfons Lawen for being my biochemistry mentor and a resourceful guide to my writing. I am also grateful to my other associate supervisor, A/Prof. Tianhai Tian for his useful suggestions time-to-time whenever I required. I also found myself motivated by the insightful comments and feedbacks from my PhD panel members: Prof. Kate Smith-Miles and Prof. Kais Hamza during my milestone seminars.

I would like to thank Monash University, The School of Mathematical Sciences, and my supervisor A/Prof. Jonathan M Keith for granting me financial and health care supports through various scholarships throughout my PhD candidature.

I would thank my friend and colleague, Salem A. Alyami for teaching me basics of statistics with patience and care which yielded several joint publications. I would also acknowledge the support from our admin, especially Linda Mayer for being very supportive to me throughout my PhD candidature.

Last, but not least, no words would be enough for expressing my gratitude towards the love, care, prayer, sacrifice and support I have had from my loving wife, Monamy as integral parts of this journey.

The List of Publications

1) *Chapter 3* : **Azad A.K.M.**, Lawen A., Keith JM. (2016). Cross-talk categorisations in data-driven models of signalling networks: a system-level view. Submitted. *BMC Research Notes*

2) *Chapter 4* : **Azad A.K.M.**, Lawen A., Keith JM. (2015). Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling. *BMC Systems Biology* 9(1): 1-17. DOI: 10.1186/s12918-014-0135-x. (**citations: 3**)

3) *Chapter 5* : **Azad A.K.M.**, Lawen A., Keith JM. (2017). Bayesian Model of Signal Rewiring Reveals Mechanisms of Gene Dysregulation in Acquired Drug Resistance in Breast Cancer. *PLoS ONE* 12(3): e0173331. https://doi.org/10.1371/journal.pone.0173331

4) *Chapter 6* : **Azad, A.K.M.** (2017). Integrating heterogeneous datasets for cancer module identification. *Bioinformatics Volume II: Structure, Function, and Applications*, pages 119-137. Springer New York, New York, NY

5) *Chapter 6* : Alyami, S., **Azad, A.K.M.**, Keith, JM. (2016). Uniform Sampling of Directed and Undirected Graphs Conditional on Vertex Connectivity. *Electronic Notes in Discrete Mathematics*, 53:43-55. DOI: 10.1016/j.endm.2016.05.005

6) *Chapter 6* : Alyami, S., **Azad, A.K.M.**, Keith, JM. (2016). The Neighborhood MCMC sampler for learning Bayesian networks. *Proc. SPIE 10011, First International Workshop on Pattern Recognition*, 100111K. DOI: 10.1117/12.2242708

Chapter 1

Introduction

'Cancer' refers to a group of diseases characterised by uncontrolled cell growth, migration, survival and differentiation, which is primarily mediated by aberrant activities of cell signalling pathways [1]. In many cancers, these aberrant signalling pathways are initiated and maintained through the up-regulation or mutation of various receptor tyrosine kinases (RTKs), such as the epidermal growth factor receptor (EGFR) and the human EGFR 2 (HER2, also known as ErbB2), thus inducing various cancer-related activities [2, 3]. Therapies targeting these aberrant RTKs by small molecule tyrosine kinase inhibitors (TKIs) have shown great potential in inhibiting cancer cell growth, and are therefore widely used in clinical trials [4]. However, initial success of these inhibitors is often followed by tumour relapse due to the acquired resistance of cancer cells after prolonged treatment. Acquired resistance is thus a significant barrier to achieving drug efficacy in response to advanced cancer [4, 5]. Understanding the mechanisms of acquired resistance in RTK targeted therapies is an important challenge in systems biology in order to facilitate mechanistic insights for developing sustainable cancer therapeutics. Recently, cross-talk among cell signalling pathways have been reported as a potential mechanism of acquired resistance in various cancers [4]. Cross-talk is defined as interactions among signalling pathways where one or more components of one pathway affect the overall activities of another pathway.

In *oncogenic addiction*, cancer cells become dependent upon specific signalling pathways that are controlled by mutation or over-expression of a single protein for their survival and/or proliferation [4, 6]. Kinase inhibitors typically target this *oncogenic behaviour* of cancer cells [4, 7]. However, the inhibitory effects of such therapies are often transient, because cancer cells acquire mechanisms to escape targeted TKIs, and thereby return to their metastatic phenotype. This phenomenon is known as acquired resistance to TKIs.

There are several known mechanisms of acquired resistance that are reported to date, including both genetic and non-genetic factors [7, 8]. Some of the genetic mechanisms of resistance to various TKIs include: 1) secondary mutations of EGFR T790M [9], HER2 kinase domain [10], BCR-ABL [11, 12], and KIT [13, 14], 2) amplifications of MET [15, 16], cyclin E [17], ALK [18], BRAF [19, 20], KRAS [20], BCR-ABL [11, 12], KIT [14], and the androgen receptor gene [21], 3) the loss of PTEN [22, 23], increased expression of IGF-IR [24] and AXL [25]. The non-genetic mechanisms, which are less well studied or poorly characterised, involve epigenetics, alternative RNA (ribonucleic acid) splicing, metabolic changes or specific protein modifications [7]. Evidence supporting such non-genetic mechanisms also includes addiction switching [26], adaptive reprogramming of signalling networks, feedback loops and cross-talk among signalling pathways [7].

In receptor targeted therapies, cross-talk among signalling pathways play a significant role in acquired resistance in various types of cancer. For example, in EGFR family receptor targeted therapies, cancer cells develop acquired resistance since multiple compensatory signalling pathways cross-talk with the EGFR signalling pathway at the receptor, mediator and effector levels [4] (see Chapter 2). The EGFR family of receptors contain four receptor proteins: EGFR (ErbB1 or HER1), ErbB2 (HER2/c-neu), ErbB3 (HER3) and ErbB4 (HER4) [27], some of them (EGFR and HER2) share common downstream signalling components with other alternative RTKs such as MET, AXL, FGFR, IGF-1R, EphA2 [4]. At the receptor level, the amplifications or

altered activations of these alternative RTKs can maintain key signals for cell survival and/or proliferation to the common downstream signalling components that were previously blocked by TKIs targeting EGFR/HER2 signalling [15, 28–31]. Again, at the mediator level, one or more components of two major downstream (of EGFR/HER2 signalling) pathways: RAS/RAF/MEK and PI3K/AKT/mTOR become re-activated by mutations or deletions of genes that act downstream of the receptors, thereby activating downstream effectors [4]. Signalling cross-talk at the effector level are more complex and diverse since there are numerous effectors in RTK signalling pathways. However, cross-talk at the effector level contribute to acquired resistance to EGFR-TKIs when signalling pathways triggered by other RTKs cause an altered response of some critical key effectors (e.g. TSC2, FOXO3) that are involved in cell survival and proliferation [4]. It has been reported that EGFR/HER2 signalling can cross-talk with Notch, Wnt/β-catenin, and TNF-α/IKK/NF-κB in order to nullify the inhibitory effects of EGFR/HER2 targeted therapies [4] (see Chapter 2). Kinome reprogramming in the signalling network is an alternative mechanism for EGFR-TKI resistance [32] (see Chapter 2). Recently, Stuhlmiller *at al.* [32] reported that continued consumption of lapatinib (an EGFR/HER2 dual inhibitor) induces *aberrant signalling activities* through transcriptional up-regulation and altered activation of multiple heterogenous RTKs (e.g. DDR1, FGFRs, IGF1R, MET) to compensate EGFR/HER2 inhibition in breast cancer cell-lines [32]. The idea of a dependency switch (addiction switch) of downstream signalling nodes in acquired resistance was recently studied by Sharifnia *et al.* [33] wherein they found that EGFR-dependent status of the downstream signalling nodes is altered by the transcriptional up-regulation of other redundant kinase-related genes that share those downstream signalling nodes with EGFR-dependent signalling [33].

Given the increasingly recognised importance of signalling cross-talk in contributing to acquired drug resistance to EGFR-TKIs, developing a systematic approach to comprehensively characterise these cross-talk is urgently required yet poorly studied.

Moreover, a network model of signalling activities reflecting the system-level perturbations (i.e. signal rewiring) in resistant-vs-parental conditions has tremendous potential to elucidate the underlying mechanisms of acquired resistance. For example, in cancer drug resistance, some relationships between gene-pairs may evolve in resistant cell-lines to compensate the inhibitory effects of drugs used [5, 34] whereas some relationships that were highly correlated in parental cell-lines may become loosely correlated (or even independent) in resistant cell-lines. *In silico* predictions made using computational modelling of such complex systems can generate biologically plausible hypotheses that can not only save time and cost relative to *in vivo* experiments but also be readily available for validation purposes and for developing novel therapeutics with sustained efficacy. Therefore, considering all the points discussed above, the main objectives of this thesis are to:

develop methods to model, identify, and characterise putative signalling cross-talk in cancer contributing to acquired drug resistance

To achieve these objectives, I employed a statistical modelling approach to characterise the system-level details of the data-driven networks of signalling activities derived from high-throughput datasets from both resistant and parental (sensitive) conditions. In constructing signalling network structure, a data-driven approach learns the relationships among nodes from data by adapting some computational methods. I hypothesised that, given some single-cell high-throughput datasets (e.g. protein or gene expression) in resistant and parental conditions, the contrasting behaviour in their respective *data-driven* network structures inferred using data observed at signalling nodes (e.g. proteins, enzymes) may elucidate the potential *aberrant signalling activities* that ultimately lead cancer cells to develop acquired resistance to targeted therapies. A statistical approach is useful for inferring and analysing *aberrant signalling activities* within the data-driven signalling network structures [35, 36]. The rationale for applying a statistical approach here is that data-driven inference of signalling networks (and other networks) can fail to detect important signalling links or incorrectly identify

links that are not present [35]. Moreover, high-throughput datasets may be noisy. Therefore, a statistical approach for analysing the *uncertain* nature of the data-driven network models can be used to predict posterior edge probabilities (of appearing in the network) by formalising them into a Bayesian model [35]. Once aberrant signalling activities are inferred, computational approaches can identify and characterise putative cross-talk involved in acquired resistance.

Here, I used a fully Bayesian approach involving the p_1-model to quantify uncertainty about which signalling activities are present in a given cell-line. This approach was applied to both parental and resistant cell-lines in breast cancer. Using this technique, I investigated the role of pathway cross-talk among signalling pathways in acquired resistance to lapatinib. The p_1-model is an Exponential Random Graph Model (ERGM) that was originally introduced by Holland and Leinhardt [37]. In general, ERGMs are statistical models for which the global structure of the network emerges as a function of local features called *explanatory variables* [36]. In the p_1-model, the set of explanatory variables includes two edge-level attributes: degree of reciprocity and global density, and two node-level attributes: attractiveness and expansiveness (see Chapter 2). This p_1-model was previously used in modelling the human protein-protein interaction network [35], and other biological systems including metabolic pathways [36] where the edge probabilities were evaluated by summarising the above topological properties of the networks (*explanatory variables*) in a parametric form and associating them with sufficient statistics [35, 37] (see Chapter 2).

In Chapter 2, I discuss: 1) basic terminologies related to acquired resistance to EGFR-TKIs in breast cancer cell-lines (SKBR3 and BT474) and 2) background of modelling signalling rewiring in acquired resistance. In Chapter 3, I propose a possible categorisation of signalling cross-talk for data-driven models of signalling networks and compare this to other state-of-the-art categorisations. In Chapter 4, I propose a computational framework in which I apply the p_1-model to infer posterior probabilities of gene-gene interactions in the networks derived from matched gene expression data

from breast cancer cell-lines (SKBR3 and BT474) under lapatinib-sensitive (parental) and lapatinib-resistant conditions. Next, I identify sets of gene-pairs from the KEGG, Reactome and WikiPathway databases as putative *drug-resistant cross-talk*, where each cross-talk is comprised of a gene in the EGFR/ErbB signalling pathway and a gene from another signalling pathway that appear to be interacting in resistant cells but not in parental cells. This work has been published in the journal BMC Systems Biology [34].

Next, I hypothesised that modelling aberrant networks with differential gene-dependencies occurring in resistant-vs-parental conditions can elucidate signalling rewiring in acquired resistance. In Chapter 5, I use the p_1-model again to model such rewired networks. This facilitates: 1) identifying dysregulated signalling pathways in acquired resistance, and 2) exploring all possible types of cross-talk among all signalling pathways [Chapter 2] involved in drug-resistance; some of which were not covered in my previous framework [Chapter 4]. I propose a novel V-shaped structure of aberrant gene-pairs in rewired networks, called a *V-Structure*, and hypothesised that it can model a possible mechanism of acquired resistance: the dysregulation of genes in acquired resistance can be mediated by the *dependency switch* from targeted signalling to bypass signalling in resistant-vs-parental conditions. Using the same gene expression data from two breast cancer cell-lines: SKBR3 and BT474 as above [Chapter 4], the results indicate that many signalling pathway structures were compromised in acquired resistance and the *V-structures* of aberrant signalling were able to provide detailed insights into the bypass mechanism of targeted inhibition. This work has been accepted for publication in the journal PLoS ONE.

In addition to my primary research focus, I also collaborated with another PhD student (Salem A. Alyami) within our research group on several projects related to *the structure and parameter inference of Bayesian network models of biological networks in systems biology using MCMC methods*. These projects yielded several publications, some of which are already published, and others are in preparation. Moreover, I published

one book chapter that reviews recent methods that integrates multiple heterogenous datasets (e.g. gene expression, copy number aberration, methylation, PPI information) in order to identify cancer modules. All of these additional projects share something in common with this thesis: namely '*Inferring Network Structure*', by which more sophisticated biological hypotheses can be tested to reveal novel signalling activities in various disease conditions. The publications regarding these additional projects are listed in Chapter 6.

Since this thesis is aimed to be written in fulfilment of the requirement for 'Thesis by Publications', Chapter 4 and Chapter 5 are comprised of journal articles with their respective format. In both of these chapters, methodologies were built around the p_1-model in order to infer posterior probabilities of gene-pairs in the data-driven networks derived from the gene expression data sets of breast cancer cell-lines. Therefore, the methods sections of these chapters partially overlap. For each chapter, the bibliography sections are separately included at the end.

In summary, the main objectives of this thesis are to computationally model, identify and characterise the signalling cross-talk in cancer contributing to acquired drug resistance. To achieve these objectives, I have explored the following research programme:

1. Review literature to gain background knowledge about the roles of signalling cross-talk and signalling rewiring in acquired resistance to RTK-targeted therapies. Model the research hypothesis and develop methodological frameworks for the problem [Chapter 2].

2. Model different types of cross-talk in the signalling networks and compare with other state-of-the-art categorisations [Chapter 3].

3. Model data-driven gene-gene relationship (GGR) networks for each of the lapatinib-sensitive (parental) and resistant conditions in breast cancer cell-lines [Chapter 4 and Chapter 5].

4. Apply a fully Bayesian approach using the p_1-model to infer posterior probabilities of gene-gene interactions existing in each of the GGR networks [Chapter 4 and Chapter 5].

5. Predict and characterise the drug resistant signalling cross-talk between EGFR/ErbB signalling and other signalling pathways that have very high probabilities of interacting in resistant cells, but low probabilities in parental cells [Chapter 4)].

6. Model signal rewiring with differential posterior probabilities of gene-pairs emulating differential gene-dependencies in resistant-vs-parental conditions in breast cancer [Chapter 5].

7. Examine the potential of signal rewiring to explain acquired resistance by a) identifying dysregulated signalling pathways in acquired resistance, and b) characterising additional types of signalling cross-talk contributing to the dysregulation of crucial genes involved in breast cancer metastasis and/or developing acquired resistance [Chapter 5].

8. Discuss key findings, issues regarding methodological challenges, and future works [Chapter 7].

Bibliography

[1] Alex Eccleston and Ritu Dhand. Signalling in cancer. *Nature*, 441(7092):423–423, May 2006.

[2] N. S. Dhomen, J. Mariadason, N. Tebbutt, and A. M. Scott. Therapeutic targeting of the epidermal growth factor receptor in human cancer. *Crit Rev Oncog*, 17(1): 31–50, 2012.

[3] J. Baselga and C. L. Arteaga. Critical update and emerging trends in epidermal growth factor receptor targeting in cancer. *J. Clin. Oncol.*, 23(11):2445–2459, Apr 2005.

[4] H. Yamaguchi, S. S. Chang, J. L. Hsu, and M. C. Hung. Signaling cross-talk in the resistance to HER family receptor targeted therapy. *Oncogene*, 33(9):1073–1081, Feb 2014.

[5] Kakajan Komurov, Jen-Te Tseng, Melissa Muller, Elena G Seviour, Tyler J Moss, Lifeng Yang, Deepak Nagrath, and Prahlad T Ram. The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant erbb2-positive breast cancer cells. *Molecular Systems Biology*, 8(1), 2012.

[6] Sreenath V. Sharma, Michael A. Fischbach, Daniel A. Haber, and Jeffrey Settleman. *Clinical Cancer Research*, 12(14):4392s–4395s, 2006.

[7] MR. Lackner, TR Wilson, and J. Settleman. Mechanisms of Acquired Resistance to Targeted Cancer Therapies. *Future Oncol.*, 8(8):999–1014, 2012.

[8] Lihua Huang and Liwu Fu. Mechanisms of resistance to egfr tyrosine kinase inhibitors. *Acta Pharmaceutica Sinica B*, 5(5):390 – 401, 2015.

[9] C. H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K. K. Wong, M. Meyerson, and M. J. Eck. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. U.S.A.*, 105 (6):2070–2075, Feb 2008.

[10] S. E. Wang, A. Narasanna, M. Perez-Torres, B. Xiang, F. Y. Wu, S. Yang, G. Carpenter, A. F. Gazdar, S. K. Muthuswamy, and C. L. Arteaga. HER2 kinase domain mutation results in constitutive phosphorylation and activation of HER2 and EGFR and resistance to EGFR tyrosine kinase inhibitors. *Cancer Cell*, 10 (1):25–38, Jul 2006.

[11] M. E. Gorre, M. Mohammed, K. Ellwood, N. Hsu, R. Paquette, P. N. Rao, and C. L. Sawyers. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, 293(5531):876–880, Aug 2001.

[12] F. X. Mahon, M. W. Deininger, B. Schultheis, J. Chabrol, J. Reiffers, J. M. Goldman, and J. V. Melo. Selection and characterization of BCR-ABL positive cell lines with differential sensitivity to the tyrosine kinase inhibitor STI571: diverse mechanisms of resistance. *Blood*, 96(3):1070–1079, Aug 2000.

[13] C. R. Antonescu, P. Besmer, T. Guo, K. Arkun, G. Hom, B. Koryotowski, M. A. Leversha, P. D. Jeffrey, D. Desantis, S. Singer, M. F. Brennan, R. G. Maki, and R. P. DeMatteo. Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation. *Clin. Cancer Res.*, 11(11):4182–4190, Jun 2005.

[14] M. Debiec-Rychter, J. Cools, H. Dumez, R. Sciot, M. Stul, N. Mentens, H. Vranckx, B. Wasag, H. Prenen, J. Roesel, A. Hagemeijer, A. Van Oosterom, and P. Marynen. Mechanisms of resistance to imatinib mesylate in gastrointestinal stromal tumors and activity of the PKC412 inhibitor against imatinib-resistant mutants. *Gastroenterology*, 128(2):270–279, Feb 2005.

[15] J. A. Engelman, K. Zejnullahu, T. Mitsudomi, Y. Song, C. Hyland, J. O. Park, N. Lindeman, C. M. Gale, X. Zhao, J. Christensen, T. Kosaka, A. J. Holmes, A. M. Rogers, F. Cappuzzo, T. Mok, C. Lee, B. E. Johnson, L. C. Cantley, and P. A. Janne. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science*, 316(5827):1039–1043, May 2007.

[16] A. B. Turke, K. Zejnullahu, Y. L. Wu, Y. Song, D. Dias-Santagata, E. Lifshits, L. Toschi, A. Rogers, T. Mok, L. Sequist, N. I. Lindeman, C. Murphy, S. Akhavanfard, B. Y. Yeap, Y. Xiao, M. Capelletti, A. J. Iafrate, C. Lee, J. G. Christensen, J. A. Engelman, and P. A. Janne. Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. *Cancer Cell*, 17(1):77–88, Jan 2010.

[17] M. Scaltriti, P. J. Eichhorn, J. Cortes, L. Prudkin, C. Aura, J. Jimenez, S. Chandarlapaty, V. Serra, A. Prat, Y. H. Ibrahim, M. Guzman, M. Gili, O. Rodriguez, S. Rodriguez, J. Perez, S. R. Green, S. Mai, N. Rosen, C. Hudis, and J. Baselga. Cyclin E amplification/overexpression is a mechanism of trastuzumab resistance in HER2+ breast cancer patients. *Proc. Natl. Acad. Sci. U.S.A.*, 108(9):3761–3766, Mar 2011.

[18] R. Katayama, T. M. Khan, C. Benes, E. Lifshits, H. Ebi, V. M. Rivera, W. C. Shakespeare, A. J. Iafrate, J. A. Engelman, and A. T. Shaw. Therapeutic strategies to overcome crizotinib resistance in non-small cell lung cancers harboring the fusion oncogene EML4-ALK. *Proc. Natl. Acad. Sci. U.S.A.*, 108(18):7535–7540, May 2011.

[19] R. B. Corcoran, D. Dias-Santagata, K. Bergethon, A. J. Iafrate, J. Settleman, and J. A. Engelman. BRAF gene amplification can promote acquired resistance to MEK inhibitors in cancer cells harboring the BRAF V600E mutation. *Sci Signal*, 3(149):ra84–ra84, 2010.

[20] A. S. Little, K. Balmanno, M. J. Sale, S. Newman, J. R. Dry, M. Hampson, P. A. Edwards, P. D. Smith, and S. J. Cook. Amplification of the driving oncogene, KRAS or BRAF, underpins acquired resistance to MEK1/2 inhibitors in colorectal cancer cells. *Sci Signal*, 4(166):ra17–ra17, 2011.

[21] T. Visakorpi, E. Hyytinen, P. Koivisto, M. Tanner, R. Keinanen, C. Palmberg, A. Palotie, T. Tammela, J. Isola, and O. P. Kallioniemi. In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat. Genet.*, 9(4):401–406, Apr 1995.

[22] Y. Nagata, K. H. Lan, X. Zhou, M. Tan, F. J. Esteva, A. A. Sahin, K. S. Klos, P. Li, B. P. Monia, N. T. Nguyen, G. N. Hortobagyi, M. C. Hung, and D. Yu. PTEN activation contributes to tumor inhibition by trastuzumab, and loss of

PTEN predicts trastuzumab resistance in patients. *Cancer Cell*, 6(2):117–127, Aug 2004.

[23] K. Berns, H. M. Horlings, B. T. Hennessy, M. Madiredjo, E. M. Hijmans, K. Beelen, S. C. Linn, A. M. Gonzalez-Angulo, K. Stemke-Hale, M. Hauptmann, R. L. Beijersbergen, G. B. Mills, M. J. van de Vijver, and R. Bernards. A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer Cell*, 12(4):395–402, Oct 2007.

[24] Y. Lu, X. Zi, Y. Zhao, D. Mascarenhas, and M. Pollak. Insulin-like growth factor-I receptor signaling and resistance to trastuzumab (Herceptin). *J. Natl. Cancer Inst.*, 93(24):1852–1857, Dec 2001.

[25] L. Liu, J. Greger, H. Shi, Y. Liu, J. Greshock, R. Annan, W. Halsey, G. M. Sathe, A. M. Martin, and T. M. Gilmer. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.*, 69(17): 6871–6878, Sep 2009.

[26] H. Aguilar, X. Sole, N. Bonifaci, J. Serra-Musach, A. Islam, N. Lopez-Bigas, M. Mendez-Pertuz, R. L. Beijersbergen, C. Lazaro, A. Urruticoechea, and M. A. Pujana. Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. *Oncogene*, 29(45):6071–6083, Nov 2010.

[27] M. J. Wieduwilt and M. M. Moasser. The epidermal growth factor receptor family: biology driving targeted therapeutics. *Cell. Mol. Life Sci.*, 65(10):1566–1584, May 2008.

[28] Z. Zhang, J. C. Lee, L. Lin, V. Olivas, V. Au, T. LaFramboise, M. Abdel-Rahman, X. Wang, A. D. Levine, J. K. Rho, Y. J. Choi, C. M. Choi, S. W. Kim, S. J. Jang, Y. S. Park, W. S. Kim, D. H. Lee, J. S. Lee, V. A. Miller, M. Arcila, M. Ladanyi, P. Moonsamy, C. Sawyers, T. J. Boggon, P. C. Ma, C. Costa, M. Taron, R. Rosell, B. Halmos, and T. G. Bivona. Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nat. Genet.*, 44(8):852–860, Aug 2012.

[29] L. A. Byers, L. Diao, J. Wang, P. Saintigny, L. Girard, M. Peyton, L. Shen, Y. Fan, U. Giri, P. K. Tumula, M. B. Nilsson, J. Gudikote, H. Tran, R. J. Cardnell, D. J. Bearss, S. L. Warner, J. M. Foulks, S. B. Kanner, V. Gandhi, N. Krett, S. T. Rosen, E. S. Kim, R. S. Herbst, G. R. Blumenschein, J. J. Lee, S. M. Lippman, K. K. Ang, G. B. Mills, W. K. Hong, J. N. Weinstein, I. I. Wistuba, K. R. Coombes, J. D. Minna, and J. V. Heymach. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.*, 19(1):279–290, Jan 2013.

[30] K. Takezawa, V. Pirazzoli, M. E. Arcila, C. A. Nebhan, X. Song, E. de Stanchina, K. Ohashi, Y. Y. Janjigian, P. J. Spitzler, M. A. Melnick, G. J. Riely, M. G. Kris, V. A. Miller, M. Ladanyi, K. Politi, and W. Pao. HER2 amplification: a potential mechanism of acquired resistance to EGFR inhibition in EGFR-mutant lung cancers that lack the second-site EGFRT790M mutation. *Cancer Discov*, 2 (10):922–933, Oct 2012.

[31] G. Zhuang, D. M. Brantley-Sieders, D. Vaught, J. Yu, L. Xie, S. Wells, D. Jackson, R. Muraoka-Cook, C. Arteaga, and J. Chen. Elevation of receptor tyrosine kinase EphA2 mediates resistance to trastuzumab therapy. *Cancer Res.*, 70(1):299–308, Jan 2010.

[32] T. J. Stuhlmiller, S. M. Miller, J. S. Zawistowski, K. Nakamura, A. S. Beltran, J. S. Duncan, S. P. Angus, K. A. Collins, D. A. Granger, R. A. Reuther, L. M. Graves, S. M. Gomez, P. F. Kuan, J. S. Parker, X. Chen, N. Sciaky, L. A. Carey, H. S. Earp, J. Jin, and G. L. Johnson. Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. *Cell Rep*, 11(3):390–404, Apr 2015.

[33] T. Sharifnia, V. Rusu, F. Piccioni, M. Bagul, M. Imielinski, A. D. Cherniack, C. S. Pedamallu, B. Wong, F. H. Wilson, L. A. Garraway, D. Altshuler, T. R.

Golub, D. E. Root, A. Subramanian, and M. Meyerson. Genetic modifiers of EGFR dependence in non-small cell lung cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 111(52):18661–18666, Dec 2014.

[34] A. K. M. Azad, A. Lawen, and J. M. Keith. Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling. *BMC Syst Biol*, 9(1):2, Jan 2015.

[35] S. Bulashevska, A. Bulashevska, and R. Eils. Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. *BMC Bioinformatics*, 11:46, 2010.

[36] Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, Oct 2007.

[37] Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

Chapter 2

Background & Literature Review

2.1 Introduction

This chapter consists of four sections. First, I will introduce: the biological background of signalling pathways in cancer; therapies targeting the receptor tyrosine kinases (RTKs) in signalling pathways, especially the EGFR/HER2/neu signalling pathway; and some of the mechanisms of resistance to those targeted therapies, with a specific focus on the roles of signalling cross-talk and pathway compensation as potential mechanisms of acquired resistance. In the second section, I define cross-talking node (gene) pairs and their different modes of operation among signalling pathways, as discussed in some state-of-the-art cross-talk modelling approaches. In the third section, I formulate the research hypothesis discussed in this thesis. Lastly in the fourth section, I discuss some background regarding the methodologies used in this thesis. These include some previous approaches used for predicting pathway cross-talk and dysregulated pathways, discussions about the statistical model (p_1-model) used in this thesis, basics of the Bayesian statistical modelling approach, and MCMC (Markov chain Monte Carlo) methods for statistical sampling.

2.2 Some background to cancer biology

2.2.1 Signalling pathways in cancer

Signal transduction is a process that transmits molecular signals from the extracellular environment to intracellular components as a series of biochemical events along a pathway in order to perform various activities, such as alteration of cellular metabolism, transcriptional regulation and cell growth. [1]. In practice, signalling pathways (i.e. biological processes) are often collected in databases e.g. KEGG [2], Reactome [3], and WikiPathway [4], where each pathway is annotated as a collection of signalling proteins. However, none of these databases contain perfect pathway annotations [5, 6], and therefore, any pathway-based analyses require the involvement of multiple such databases rather than relying one a single one.

Receptor tyrosine kinases (RTKs) are transmembrane proteins that receive signals with their extracellular ligand-binding domains and in consequence trigger cascades of biochemical events through activation of their intracellular tyrosine kinase domains [7]. RTKs such as epidermal growth factor receptor (EGFR) and EGFR2 (also known as HER2, neu and ErbB2) [Box 2.2.1] perform essential roles in normal cellular process. However, development and progression of many cancers are largely driven by up-regulation and/or alterations of these RTKs [8]. For example, EGFR is often found mutated or over-expressed in lung, colon, head and neck, brain, pancreas and breast cancer [9–12], and HER2 is often found over-expressed in breast, gastric, pancreatic, ovarian and esophageal cancers. [13, 14]. Therefore, alterations in these RTKs trigger aberrant cell signalling that ultimately induces various cancer related activities such as cell proliferation, differentiation, and survival [Box 2.2.1].

Box 2.1.1:

EGFR family of receptors: EGFR family of receptors includes four receptor proteins: ErbB-1 (also known as EGFR or HER1), ErbB-2 (also known as HER2/c-neu), ErbB-3 (also known as HER3), and ErbB-4 (also known as HER4) [15].

Cell Proliferation: Cell proliferation is an increase in cell number as a result of cell growth and cell division [16].

Cell Differentiation: A cellular process by which a less specialised cell changes into a more specialised cell type [17].

Cell Survival: The period of cell viability with sustained capacity to perform certain cellular functions such as metabolism, growth, reproduction, and adaptability [18].

2.2.2 Oncogene addiction & RTK-targeted therapies

Despite numerous genetic alterations and/or epigenetic abnormalities, cancer cell proliferation and survival often rely on a single oncogenic pathway or oncogene and its protein products, controlled impairment of which can significantly inhibit the growth of cancer cells and thereby enhance patient survival [19]. This phenomenon is referred to as *oncogene addiction*. This 'Achilles heel' of cancer cells offers scope to develop novel therapeutics, such as EGFR-family receptor targeted therapies [19]. However, to maintain sustainable efficacy of therapies targeting such oncogenic addiction of cancer cells requires appropriate identification of biomarkers indicative of such addictions and selection of patients possessing such biomarkers [7].

Since many cancer cells exhibit oncogenic addiction to RTKs (particularly growth factor related RTKs) for their proliferation and survival, RTKs possess high potential as novel therapeutic targets [20]. Monoclonal antibodies (mAbs) and TKIs are two kinds

of anticancer therapeutics that target growth factor receptors to block the signalling triggered by RTKs and inhibit growth-related downstream signals [20]. Takeuchi *et al.* [20] listed anticancer therapies targeting growth factor receptors in various cancers as shown in Table 2.1 [20].

Table 2.1: *A list of developed therapeutics targeting growth factor receptors in various cancers. [Source [20]]*

Drug type	**Drug Name**	**Disease**	**Targeted RTKs**
Antibody	Trastuzumab (Herceptin)	Breast cancer	HER2
	Bevacizumab (Avastin)	Metastatic colorectal carcinoma	VEGFR
	Cetuximab (Erbitux)	EGFR-expressing metastatic colorectal cancer	EGFR
	Panitumumab (Vectibix)	Wild-type KRAS-expressing Metastatic colorectal cancer	EGFR
Small molecule inhibitors	Gefitinib (Iressa)	Metastatic non-small-cell lung cancer	EGFR
	Erlotinib (Tarceva)	Metastatic non-small-cell lung cancer	EGFR
	Sorafenib (Nexavar)	Renal cell cancer	VEGFR, PDGFR
	Lapatinib (Tykerb)	HER2-positive Breast cancer	EGFR, HER2

2.2.3 EGFR-family receptor targeted therapies

Because of the well-studied role of EGFR and HER2 in driving cell proliferation and survival signals to their downstream signalling nodes, they have been targeted for designing many therapeutic agents (see Table 2.1). Both mAbs and TKIs have been examined in either clinical trials or advanced pre-clinical studies [21]. MAbs target the extracellular domain of receptors to block their activation while TKIs target their

intracellular ATP-binding sites to inhibit the phosphorylation of the target proteins. Figure 2.1 demonstrates the basic mechanism of these two types of anti-EGFR drugs.

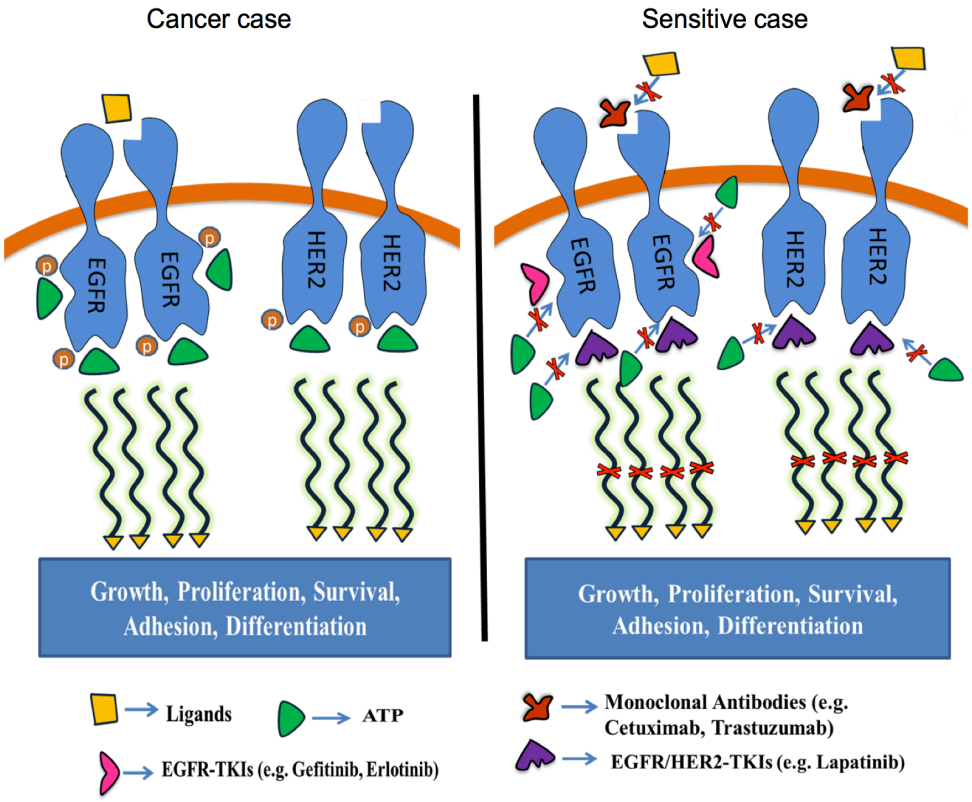


Figure 2.1: *Mechanisms of anti-EGFR and anti-HER2 targeted therapies. Left panel shows the untreated state of cancer cells where key signals from EGFR/HER2 RTKs are maintained in order to induce various cancer related activities. Right panel shows how those key signals are treated with anti-EGFR and anti-HER2 targeted therapies in order to inhibit the cancer related activities by blocking the pathway.*

Trastuzumab (Herceptin)

Trastuzumab (Herceptin) is a humanised recombinant mAb which binds with the extracellular domain of HER2 receptor [22]. It was approved for clinical use by the FDA (Food and Drug Administration) in 1998, and has been proven to reduce the risk of breast cancer recurrence after treatment with adjuvant chemotherapy compared to patients treated with chemotherapy alone [23–25]. Mechanisms of action (MoA) of trastuzumab are two-fold: 1) down-regulation of intracellular signalling pathways via PI3K and MAPK pathway, and 2) activation of immune response via antibody dependent cell-mediated cytotoxicity (ADCC) [22]. Unfortunately, tumours become

resistant to trastuzumab within one year of treatment [26]. The mechanisms of resistance include altered activation of other HER-family receptors (EGFR or HER3), insulin-like growth factor receptor, re-activation of PI3K/AKT/mTOR pathway, over-expression of c-MET or loss of PTEN, and up-regulation of src-kinase activities [22].

Lapatinib

Lapatinib is a TKI which targets both EGFR and HER2 receptors. Unlike trastuzumab, it binds with the intracellular domain (ATP-binding sites) of those receptors and inhibits the phosphorylation of downstream MAPK and Akt [27]. Lapatinib was approved by FDA in 2007 for the treatment of metastatic HER2 over-expressed breast cancer patients in a combination therapy along with other chemotherapeutic agents [28]. Lapatinib also provided improved efficacy over trastuzumab in terms of inhibiting cell proliferation in trastuzumab-resistant breast SKBR3 cell-line [29]. However, like trastuzumab the efficacy of lapatinib is limited since cancer cells acquire resistance to this drug. Characterisation of the underlying mechanisms of resistance is yet to be completed [30].

2.2.4 Mechanisms of acquired resistance to receptor tyrosine kinase-targeted therapies

Resistance: *de novo* and acquired resistance

There are two types of resistance to inhibitor therapies: *de novo* and acquired resistance. *De novo* resistance occurs when a drug with proven efficacy to inhibit tumour cell growth fails to induce any significant response due to some intrinsic characteristics of the cancer [31–33]. In acquired resistance, initial success of inhibitors fails to continue over time as tumour cells acquire escape mechanisms [31]. It is reported that these two types of resistance mechanisms are inter-related, since failure to tackle *de novo* mechanisms may contribute to acquired resistance [31, 33].

Genetic and non-genetic mechanisms

Initial efficacy of RTK-targeted inhibitors may decline because cancer cells acquire *genetic* alterations of key signalling components which in turn lead to changes in corresponding pathway activities [34]. Recently, advancements in array-based technologies provide the potential to profile genomic changes of cancer cells with pre- or post-treatment effects due to targeted therapies. Secondary mutations, genetic amplifications and deletions are some of the *genetic* mechanisms of resistance to various RTK inhibitors. Some of the examples have already been listed on p. 2 [Chapter 2] of this thesis.

On the other hand, *non-genetic* mechanisms of acquired resistance are not driven by mutations [34]. Some *non-genetic* mechanisms are the possible role of epigenetics, alternative RNA splicing, and metabolic changes or post-translational modification of proteins which are not primarily caused by mutations [34]. Alternative mechanisms include epithelial-mesenchymal transition (EMT), proliferation of drug-tolerant cancer stem cells [34], oncogene addiction switching [34, 35], feedback loop and pathway cross-talk [7, 31, 34], reprogramming of cell signalling circuitry [31, 36, 37], and alternative up-regulation of compensatory pathways [36, 38, 39]. In this thesis, I focus on building computational frameworks to elucidate the possible roles of pathway cross-talk, altered activations of compensatory pathways, oncogene addiction switching, and the reprogramming of signalling networks (signal rewiring) as possible mechanisms of acquired resistance to lapatinib.

2.2.5 Signalling cross-talk and pathway compensation in acquired resistance

Crosstalk

Cross-talk is an important aspect of a network of signalling pathways. It is defined as the interactions among pathways whereby one or more components of one pathway

affect(s) the activities of other pathways. For example, the TGF-β/BMP signalling pathway cross-talk with the MAPK, PI3K/AKT, Wnt, Hedgehog and Notch signalling pathways [40, 41]. Again, the EGFR signalling pathway cross-talk with other signalling pathways including insulin, Notch, Wnt and TNFR/IKK/NF-κB pathways, which contribute to acquired resistance to EGFR-TKIs [7]. The term 'cross-talk' is borrowed from the field of electronic circuit design, where it refers to a design flaw resulting in unwanted effects or influence in one circuit caused by another [42]. However, biological cross-talk do not necessarily involve signal interference. Rather, the term refers to complex signal integration between two or more signalling pathways [40]. Therefore, it can be summarised as: signalling are events, signalling pathways are molecular road-maps how that event transmits through their components, and cross-talk among signalling pathways indicate the interactions among those road-maps.

The role of signalling cross-talk and pathway compensation

Signals initiated at a single RTK transduce through a series of biochemical molecules including mediator proteins and effector proteins. Both of these may be enzymes that may result in signal amplification at multiple points along the cascade. In addition, amplification may occur due to cross-talk between different pathways [7]. More specifically, phosphorylated RTKs at the *receptor-level* initiate and amplify the transduced signals by recruiting and phosphorylating multiple target proteins. Next, kinase proteins at the *mediator-level* (downstream of RTKs) also phosphorylate multiple target proteins and amplify the signal by activating or suppressing their activities. Then, further downstream signalling molecules at the *effector-level*, such as transcription factors (TFs) affect the transcription of target gene expression. Thus, signalling from the single RTK can cross-talk with other signalling pathways at multiple stages of its propagation.

Signalling cross-talk occur at all three levels of signal transduction: receptor, mediator and effector-level [7]. Yamaguchi *et al.* [7] reported that cross-talk at these three levels may contribute to acquired resistance mechanisms (Figure 2.2). Inhibition of signals

regarding cell growth, proliferation, and survival by RTK-targeted therapies may fail since cross-talk may affect the targeted signalling cascades and restore the proliferation signal independently of the RTK so that the cell relapses into the tumourigenic phenotype. Cross-talk at the *receptor-level* contributes to acquired resistance when other RTKs with the same common downstream components become aberrantly activated or amplified and compensate for the desired inhibition of those downstream signalling components. For example, resistance to EGFR/HER2 inhibitors in various cancers occurs when the up-regulation or activation of alternate RTKs such as MET (lung [43, 44] and colon [45] cancer), IGF1R (lung [46], breast [47] and colon [48] cancer), AXL (lung [49, 50] and breast [30] cancer), FGFR (lung [51] cancer) or EphA2 (breast [52] cancer) maintain the key signals for cell growth, proliferation and survival to downstream RAS/RAF/MEK and PI3K/AKT pathways [7]. At the mediator-level, mutations or copy number changes of some key kinase-related genes constitutively activate/inactivate the downstream signalling independently of the target RTKs via signalling cross-talk [7]. For example, in two of the major downstream mediator pathways of EGFR/HER2 signalling, the RAS/RAF/MEK and PI3K/AKT pathways, the mutational activation of K-RAS, B-RAF and PI3K, and the inactivation of PTEN by mutations or deletion may cause up-regulation of downstream growth signals and thereby contribute to EGFR TKI resistance in many cancers including colon [53, 54], lung [55] and breast [56, 57]. Finally, cross-talk at the effector-level plays a role in acquired resistance when multiple upstream mediators from the same RTK signalling or other signalling pathways change the activities of the common downstream effectors that are critical for cancer cell proliferation and survival [7]. For example, the IKK/NF-κB signalling pathway cross-talk with the EGFR/HER2 signalling pathway at the effector-level [58–60] by phosphorylating their common downstream targets including FOXO3 and the TSC complex, thereby playing a critical role in acquired resistance to EGKR-TKIs [61–64].

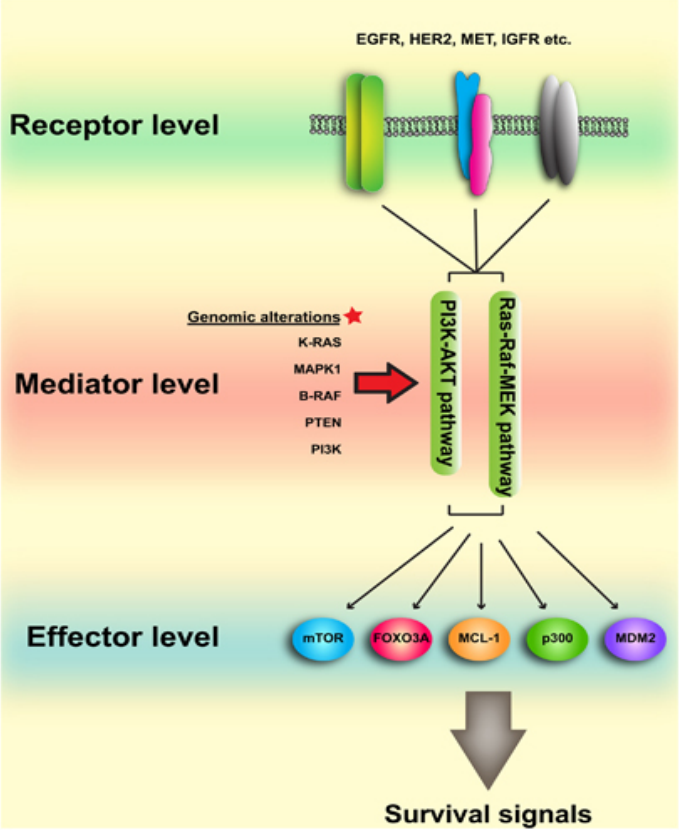


Figure 2.2: *Cross-talk events at multiple levels of EGFR/HER2 signalling pathways. This figure is re-printed from Yamaguchi* et al. *[7] by permission from Macmillan Publishers Ltd: ONCOGENE, copyright 2014.*

2.2.6 Adaptive signalling rewiring and dependency switch in acquired resistance

Tumour cells respond to kinase-targeted inhibitors by rewiring their signalling network to escape the inhibitory effects [31, 37]. Such rearrangements in signalling circuitry may be due to adaptive kinome responses involving altered regulation of various alternate kinase proteins other than the targeted signalling nodes, such as PI3K, AKT, mTOR, BRAF and MEK. These kinases are known to control tumour growth and survival. Recently, Stuhlmiller *at al.* [37] reported that continued consumption of lapatinib induces *aberrant signalling activities* through transcriptional up-regulation and altered activation of multiple heterogenous RTKs (e.g. DDR1, FGFRs, IGF1R, MET) to compensate EGFR/HER2 inhibition in breast cancer cell-lines [37]. In other

words, cancer cells may shift their *oncogenic dependencies* from the targeted (by the inhibitors) signalling nodes to alternate up-regulated kinases in order to continue their proliferation. This bypass mechanism thus provides cancer cells with the necessary signals to recover their tumourigenic phenotype (e.g. abnormal growth, survival, differentiation, migration). These cells thereby acquire resistance to the inhibitors.

2.3 Some background to methodologies

2.3.1 Previous studies on inferring pathway cross-talk

There are several methods that identify cross-talk among signalling pathways. Some of these methods consider cross-talk as part of a broader methodology, but only the cross-talk identification parts are discussed here. XTalk [65] uses a path-based approach that enumerates *shortest-paths* from a predefined list of receptor proteins to a list of transcription factor proteins. This method defines cross-talk as the shortest-paths that connect the receptors of one signalling pathway to the transcription factors of another pathway. After defining a scoring metric for such cross-talk, a novel technique was developed to evaluate their statistical significance.

Applying a signature-based gene-set co-expression analysis (sGSCA), Wang *et al.* [66] inferred a pathway cross-talk network by integrating prior knowledge (e.g. pathway annotations, molecular interactions) with gene expression datasets. A sparse canonical correlation analysis (SCCA) was applied in order to measure gene-set co-expression at the pathway-level, and several important pathway cross-talk that are involved in cancer were identified.

Recently, Andra *et al.* [67] analysed the role of cross-talk between Estrogen signalling and other pathways that influence tamoxifen (a drug used for estrogen positive breast cancer patients) efficacy in breast cancer. Using gene expression datasets of tamoxifen-sensitive and tamoxifen-resistant samples, this method identified cross-talk using the

Jaccard coefficient to quantify pathway overlapping. However, this study considers only shared components between two pathways as cross-talking points, and lacks direct interactions between pathways (see Type-I and Type-II cross-talk in our proposed categorisation). A similar approach to cross-talk definition was also reported by Donato *et al.* [68], where they proposed an additional technique for correcting various pathway analysis methods (e.g. enrichment analysis, functional class scoring, topology-based methods) which are are affected by pathway cross-talk.

2.3.2 Previous studies on inferring dysregulated pathways

There are many methods available to identify dysregulated pathways in context-specific phenotypic changes (i.e. case-vs-control, cancer-vs-normal) [69–71]. These methods include node-centric and edge-centric approaches in order to conduct enrichment analysis of perturbed components within the pathways of interest. Signalling pathway impact analysis (SPIA) considers both classical enrichment of differentially expressed genes and significant perturbation activities in a given signalling pathway topology by analysing cancer-vs-normal gene expression datasets.

DAVID (Database for Annotation, Visualisation and Integrated Discovery) [72] and GATHER (Gene Annotation Tool to Help Explain Relationships) [73] use classical enrichment analysis of differentially expressed genes and thus are applicable to identify dysregulated signalling pathways.

ESEA (Edge Set Enrichment Analysis) [70] and PAGI (Pathway Analysis based on Global Influence) [71] are edge-centric methods that use known pathway structures from popular databases (including KEGG, Reactome, Biocarta). ESEA integrates pathway structure and differential co-expression among genes in order to identify dysregulated pathways in cancer-vs-normal conditions [70]. PAGI detects aberrant pathways by analysing global influences of both intra-pathway and inter-pathway (cross-talk) effects on differentially expressed genes in cancer-vs-normal conditions [71].

2.3.3 p_1-model: a special class of Exponential Random Graph Models (ERGMs)

Exponential Random Graph Models (ERGMs)

The Exponential Random Graph Models (ERGMs) or p^*-models are probability distributions for statistical modelling of various types of network data, where the global structure of the network is expressed as a function of local structural patterns [74, 75]. These local structural features can be some of the network statistics such as *edgecount*, *nodecount*, *trianglecount* and *k-star* for $k = 2, 3, ...$ [74]. However, these statistics can be considered as a set of explanatory variables in order to explain the probability functions of networks [74]. Here, the explanatory variables can be defined as any function from the observed network to the real numbers [74]. Although ERGMs have been extensively studied in social network analyses, they provide enough flexibility and robustness, especially in terms of the number of available local feature choices and their scalability, so that they become applicable in the statistical modelling of biological networks as well [74].

Let $\mathbf{X}$ be a random matrix (matrix-valued random variable) defined on a state space $\mathcal{G}$ containing networks (e.g. biological networks) where each network is represented as a g-by-g adjacency matrix. Here each adjacency matrix is a collection of entries with 0's and 1's where 1 indicates an edge between two nodes (undirected), and 0 indicates otherwise. Let $\mathbf{u}$ be a generic point of $\mathcal{G}$ representing an observed network so that the realisation of $\mathbf{X}$ can be denoted as $\mathbf{X} = \mathbf{u}$. Then the probability function, $P(\mathbf{X} = \mathbf{u})$ can be approximated using a *log-linear* model by summarising all the explanatory variables (i.e. network statistics) and associating corresponding model parameters with those variables. This probability function can be stated as follows:

$$P(\mathbf{X} = \mathbf{u}) = \frac{e^{\sum_p \theta_p z_p(\mathbf{u})}}{\boldsymbol{\kappa}(\boldsymbol{\theta})} \tag{2.3.1}$$

where $z_p(\mathbf{u})$ is the explanatory variable (i.e. network statistic) of type p, which is expressed as a function of the observed network $\mathbf{u}$, θ_p is the model parameter associated with $z_p(\mathbf{u})$, $\boldsymbol{\theta}$ is the vector of all model parameters, and $\boldsymbol{\kappa_\theta}$ is the normalising constant ensuring the probabilities sum to one.

p_1-model

The p_1-model is a particular type of ERGM that was originally proposed for directed graphs by Holland and Leinhardt in 1981 [76]. In a directed graph, the relationship between any two nodes i and j is called a 'dyad' (pair) which can be either *mutual* (both 'i connects to j' and 'j connects to i'), or *assymetric* ('i connects to j', or 'j connects to i', but not both), or *null* (i and j are not connected at all). This study was primarily based on two empirical observations [76]: the parameters 1) the total number of *mutual* relationships in the network and 2) the *in-degree* (the number of relationships connected to node 'i') were repeatedly found in social networks to be significantly *higher* or *lower* than their *expected* values [76]. Inspired by these observations and substantive theoretical predictions, Holland and Leinhardt constructed the p_1-model as follows:

$$P_1(\mathbf{u}) = P(\mathbf{X} = \mathbf{u}) = \frac{e^{\rho m + \theta u_{++} + \sum_i \alpha_i u_{i+} + \sum_j \beta_j u_{+j}}}{\boldsymbol{\kappa}(\rho, \theta, \{\alpha_i\}, \{\beta_j\})} \tag{2.3.2}$$

where, m, u_{++}, u_{i+}, and u_{+j} are the values of the number of mutual relationships, total number of relationships, *in-degree* of node i and the *out-degree* of node j (the number of nodes connected from node j), respectively; all are computed from the observed network $\mathbf{u}$ [76]. The model parameters ρ and θ are two global parameters which are called the global degree of *reciprocity*, and the global *density* parameter, respectively. The terms α_i, and β_j are two local parameters (referring to individual nodes i and j) which are called the *expansiveness* of node i and the *attractiveness* of node j, respectively. The function $\boldsymbol{\kappa}$ maps the network parameters to a normalising constant.

A major limitation of the p_1-model is the difficulty of calculating the above normalising constant, since it is a sum over the entire graph space. Estimating the maximum likelihood of this model becomes intractable as there are $2^{g(g-1)}$ possible directed graphs (or $2^{\frac{g(g-1)}{2}}$ undirected graphs), each having g nodes. A technique called *maximum pseudolikelihood estimation* (MPLE) has been developed to address this problem [77]. This technique employs MCMC methods such as Gibbs or Metropolis-Hastings sampling algorithms [78]. A detailed derivation of the p_1-model for a directed network is described in Appendix B of Chapter 4.

2.3.4 Bayesian Inference

Conventional statistical methods assume that unknown parameters are fixed and not described in terms of their probabilities. However, Bayesian methods treat parameters as random variables and use probabilities to quantify the 'degree of belief'.

Bayesian inference is a statistical learning procedure where initial *prior* probability statements about the parameters can be updated to produce *posterior* knowledge by incorporating the *prior* knowledge with the *data* using Bayes' theorem [79]. Let $\mathcal{D}$ be the observed dataset produced by some generative model $\mathcal{M}$, and let the posterior probability of the parameter θ be $P(\theta|\mathcal{M})$ [75]. Bayes' theorem states:

$$P\left(\theta|\mathcal{M},\mathcal{D}\right) = \frac{P\left(\mathcal{D}|\theta,\mathcal{M}\right) \times P\left(\theta|\mathcal{M}\right)}{\mathcal{Z}} \tag{2.3.3}$$

where $P\left(\mathcal{D}|\theta,\mathcal{M}\right)$ is the likelihood function. The marginal likelihood $\mathcal{Z}$ can be expressed as

$$\mathcal{Z} = P\left(\mathcal{D}|\mathcal{M}\right) = \int P\left(\mathcal{D}|\mathcal{M},\theta\right) \times P\left(\theta|\mathcal{M}\right) d\theta, \tag{2.3.4}$$

Calculation of this normalising constant $\mathcal{Z}$ is often an intractable problem as it is prone to the curse-of-dimensionality [75, 80]. However, simulation techniques (see the following section) can be applied without explicit calculation of $\mathcal{Z}$ [80].

2.3.5 Markov chain Monte Carlo Methods

In Bayesian statistics, the need to integrate complex and high-dimensional functions often arises, such as in calculating 1) the normalising constant of proportionality in Bayes' theorem, 2) the marginal distribution, and 3) inferences in the form of posterior expectations [80]. Explicit computations of such complex integrals are often intractable or at least computationally intensive even with powerful computational resources. Fortunately, Markov chain Monte Carlo (MCMC) methods offer an alternate to such complex computation by sampling from the posterior distribution, and estimating quantities of interest using those samples [80].

Let $\pi(x)$ be a *target* probability distribution of a quantity of interest, where $x \in S$ and S is called *target* state space. If $\pi(x)$ cannot be sampled directly, then the MCMC approach is used to construct a *Markov chain* in the state space S such that its *stationary* distribution is equal to the *target* posterior distribution [80].

A *Markov chain* is defined as a random process which is a sequence of random variables $H_1, H_2, ..., H_n$ with values in a state space S. The key property of a *Markov chain* is that H_{t+1} is conditionally independent of $H_1, H_2, ..., H_{t-1}$, given H_t. A *stationary* distribution for a *Markov chain* on the target space S is invariant for the *transition function* which is the distribution of H_{t+1} conditional on H_t.

After running a *Markov chain* for a sufficient time the chain effectively converges to its stationary distribution, and the samples drawn from that chain can be considered as if they were drawn from the target posterior distribution [80]. Then *Monte Carlo* integration can be applied to approximate the posterior quantities of interest [81].

Monte Carlo integration is a useful technique for computing complex integrals. Let $\int_a^b h(\theta)d\theta$ be the integral to be computed. The *Monte Carlo* technique decomposes $h(\theta)$ into a product of two functions, $f(\theta)$ and $p(\theta)$, where $f(\theta)$ is a function of θ and $p(\theta)$ is a probability density function defined over the interval (a, b). Then the original integral $\int_a^b h(\theta)d\theta$ can be expressed as the expectation of $f(\theta)$ over the density $p(\theta)$ as bellow [81]:

$$\int_a^b h(\theta)d\theta = \int_a^b f(\theta)p(\theta)d\theta = E_{p(\theta)}\left[f(\theta)\right] \tag{2.3.5}$$

Thus if a large number of random variables, $\theta_1, ..., \theta_n$ are drawn from the density $p(\theta)$, then the Monte Carlo integration can be represented as bellow [81]:

$$\int_a^b h(\theta)d\theta = E_{p(\theta)}[f(\theta)] \simeq \frac{1}{n}\sum_{i=1}^{n} f(\theta_i) \tag{2.3.6}$$

There are many MCMC methods available, including the Metropolis-Hastings sampler [82], Gibbs sampler [83], Hit-and-Run sampler [84], and Neighbourhood sampler [85]. In this thesis, I have used the Gibbs sampling technique for parameter inference in the p_1-model.

Gibbs Sampling

Gibbs sampling [83] is an MCMC method for sampling a multivariate probability distribution [81]. The key assumption in Gibbs sampling is that for a given multivariate distribution it is easier to sample from conditional distributions for each parameter in the model than it is to marginalise a joint distribution by integration [81]. The joint distribution over all parameters is decomposed into full conditional distributions for each individual parameter, and these conditionals are sampled sequentially and iteratively.

Suppose M samples of $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ (a parameter vector) need to be drawn from the joint probability distribution $p(\theta_1, \theta_2, ..., \theta_k)$. Let the i-th sample be defined as $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_k^{(i)})$. The sampling steps are as follows:

1. Begin with $i = 0$ and set arbitrary initial values of the parameters in the vector, $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_k^{(i)})$.

2. For the $(i + 1)$-th sample, the parameter vector will be defined as $\boldsymbol{\theta}^{(i+1)} = (\theta_1^{(i+1)}, \theta_2^{(i+1)}, ..., \theta_k^{(i+1)})$. To generate the $(i + 1)$-th sample, each component parameter $\theta_j^{(i+1)}$ is sampled in turn from the distribution specified by $p(\theta_j^{(i+1)} | \theta_1^{(i+1)}, \theta_2^{(i+1)}, ..., \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, ..., \theta_k^{(i)})$

3. Repeat step 2 until $i = M$.

After a period of time known as *burn-in*, the M samples drawn using the above algorithm can be considered as if they were sampled from the posterior joint distribution [81]. Using this sample, *Monte Carlo* integration can be applied to infer the quantities of interest [81].

2.3.6 WinBUGS

WinBUGS is a Microsoft Windows based software that is used for Bayesian inference using Gibbs sampling [86]. This is a high-level software package providing an easy interface for implementing complex Bayesian models. In WinBUGS, users are freed from background lower-level programming details, and only have to precisely express the model, corresponding data, and initial values of model parameters.

Bibliography

[1] H. Lodish, A. Berk, S. L. Zipursky, and et al. *Molecular Cell Biology*. 4th edition. New York: W. H. Freeman, 2000. URL http://www.ncbi.nlm.nih.gov/books/NBK21517/. Section 20.1, Overview of Extracellular Signaling.

[2] M. Kanehisa. The KEGG database. *Novartis Found. Symp.*, 247:91–101, 2002.

[3] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Res.*, 42(Database issue):D472–477, Jan 2014.

[4] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, 40(Database issue):D1301–1307, Jan 2012.

[5] L. Wadi, M. Meyer, J. Weiser, L. D. Stein, and J. Reimand. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, 13(9):705–706, Aug 2016.

[6] M. Masseroli, A. Canakoglu, and M. Quigliatti. Detection of gene annotations and protein-protein interaction associated disorders through transitive relationships between integrated annotations. *BMC Genomics*, 16:S5, 2015.

[7] H. Yamaguchi, S. S. Chang, J. L. Hsu, and M. C. Hung. Signaling cross-talk in the resistance to HER family receptor targeted therapy. *Oncogene*, 33(9):1073–1081, Feb 2014.

[8] E. Zwick, J. Bange, and A. Ullrich. Receptor tyrosine kinase signalling as a target for cancer intervention strategies. *Endocr. Relat. Cancer*, 8(3):161–173, Sep 2001.

[9] M. F. Rimawi, P. B. Shetty, H. L. Weiss, R. Schiff, C. K. Osborne, G. C. Chamness, and R. M. Elledge. Epidermal growth factor receptor expression in breast cancer association with biologic phenotype and clinical outcomes. *Cancer*, 116(5):1234–1242, Mar 2010.

[10] H. Shigematsu and A. F. Gazdar. Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers. *Int. J. Cancer*, 118(2):257–262, Jan 2006.

[11] H. K. Gan, A. H. Kaye, and R. B. Luwor. The EGFRvIII variant in glioblastoma multiforme. *J Clin Neurosci*, 16(6):748–754, Jun 2009.

[12] G. Bronte, M. Terrasi, S. Rizzo, N. Sivestris, C. Ficorella, M. Cajozzo, F. Di Gaudio, G. Gulotta, S. Siragusa, N. Gebbia, and A. Russo. EGFR genomic alterations in cancer: prognostic and predictive values. *Front Biosci (Elite Ed)*, 3:879–887, 2011.

[13] M. R. Sharma and R. L. Schilsky. GI cancers in 2010: New standards and a predictive biomarker for adjuvant therapy. *Nat Rev Clin Oncol*, 8(2):70–72, Feb 2011.

[14] V. Abramson and C. L. Arteaga. New strategies in HER2-overexpressing breast cancer: many combinations of targeted drugs available. *Clin. Cancer Res.*, 17(5): 952–958, Mar 2011.

[15] M. J. Wieduwilt and M. M. Moasser. The epidermal growth factor receptor family: biology driving targeted therapeutics. *Cell. Mol. Life Sci.*, 65(10):1566–1584, May 2008.

[16] Cell proliferation - latest research and news — nature, . URL http://www.nature.com/subjects/cell-proliferation.

[17] Definition of cell differentiation - nci dictionary of cancer terms - national cancer institute, . URL https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=46477.

[18] Cell survival (definition), . URL http://www.reference.md/files/D002/mD002470.html.

[19] I. Bernard Weinstein and Andrew Joe. Oncogene addiction. *Cancer Research*, 68 (9):3077–3080, 2008.

[20] K. Takeuchi and F. Ito. Receptor tyrosine kinases and targeted cancer therapeutics. *Biol. Pharm. Bull.*, 34(12):1774–1780, 2011.

[21] S. E. Wang, A. Narasanna, M. Perez-Torres, B. Xiang, F. Y. Wu, S. Yang, G. Carpenter, A. F. Gazdar, S. K. Muthuswamy, and C. L. Arteaga. HER2 kinase domain mutation results in constitutive phosphorylation and activation of HER2 and EGFR and resistance to EGFR tyrosine kinase inhibitors. *Cancer Cell*, 10 (1):25–38, Jul 2006.

[22] Pernelle Lavaud and Fabrice Andre. Strategies to overcome trastuzumab resistance in her2-overexpressing breast cancers: focus on new data from clinical trials. *BMC Medicine*, 12(1):1–10, 2014.

[23] S. Chia, M. Clemons, L. A. Martin, A. Rodgers, K. Gelmon, G. R. Pond, and L. Panasci. Pegylated liposomal doxorubicin and trastuzumab in HER-2 overexpressing metastatic breast cancer: a multicenter phase II trial. *J. Clin. Oncol.*, 24 (18):2773–2778, Jun 2006.

[24] M. Marty, F. Cognetti, D. Maraninchi, R. Snyder, L. Mauriac, M. Tubiana-Hulin, S. Chan, D. Grimes, A. Anton, A. Lluch, J. Kennedy, K. O'Byrne, P. Conte, M. Green, C. Ward, K. Mayne, and J. M. Extra. Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 study group. *J. Clin. Oncol.*, 23 (19):4265–4274, Jul 2005.

[25] N. Robert, B. Leyland-Jones, L. Asmar, R. Belt, D. Ilegbodu, D. Loesch, R. Raju, E. Valentine, R. Sayre, M. Cobleigh, K. Albain, C. McCullough, L. Fuchs, and

D. Slamon. Randomized phase III study of trastuzumab, paclitaxel, and carboplatin compared with trastuzumab and paclitaxel in women with HER-2-overexpressing metastatic breast cancer. *J. Clin. Oncol.*, 24(18):2786–2792, Jun 2006.

[26] E. H. Romond, E. A. Perez, J. Bryant, V. J. Suman, C. E. Geyer, N. E. Davidson, E. Tan-Chiu, S. Martino, S. Paik, P. A. Kaufman, S. M. Swain, T. M. Pisansky, L. Fehrenbacher, L. A. Kutteh, V. G. Vogel, D. W. Visscher, G. Yothers, R. B. Jenkins, A. M. Brown, S. R. Dakhil, E. P. Mamounas, W. L. Lingle, P. M. Klein, J. N. Ingle, and N. Wolmark. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N. Engl. J. Med.*, 353(16):1673–1684, Oct 2005.

[27] W. Xia, C. M. Gerard, L. Liu, N. M. Baudson, T. L. Ory, and N. L. Spector. Combining lapatinib (GW572016), a small molecule inhibitor of ErbB1 and ErbB2 tyrosine kinases, with therapeutic anti-ErbB2 antibodies enhances apoptosis of ErbB2-overexpressing breast cancer cells. *Oncogene*, 24(41):6213–6221, Sep 2005.

[28] Q. Ryan, A. Ibrahim, M. H. Cohen, J. Johnson, C. W. Ko, R. Sridhara, R. Justice, and R. Pazdur. FDA drug approval summary: lapatinib in combination with capecitabine for previously treated metastatic breast cancer that overexpresses HER-2. *Oncologist*, 13(10):1114–1119, Oct 2008.

[29] R. Nahta, L. X. Yuan, Y. Du, and F. J. Esteva. Lapatinib induces apoptosis in trastuzumab-resistant breast cancer cells: effects on insulin-like growth factor I signaling. *Mol. Cancer Ther.*, 6(2):667–674, Feb 2007.

[30] L. Liu, J. Greger, H. Shi, Y. Liu, J. Greshock, R. Annan, W. Halsey, G. M. Sathe, A. M. Martin, and T. M. Gilmer. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.*, 69(17): 6871–6878, Sep 2009.

[31] J. S. Logue and D. K. Morrison. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. *Genes Dev.*, 26(7):641–650, Apr 2012.

[32] P. A. Bauman, W. S. Dalton, J. M. Anderson, and A. E. Cress. Expression of cytokeratin confers multiple drug resistance. *Proc. Natl. Acad. Sci. U.S.A.*, 91 (12):5311–5314, Jun 1994.

[33] LoriA. Hazlehurst and WilliamS. Dalton. De novo and acquired resistance to antitumor alkylating agents. In BeverlyA. Teicher, editor, *Cancer Drug Resistance*, Cancer Drug Discovery and Development, pages 377–389. Humana Press, 2006. ISBN 978-1-58829-530-9.

[34] MR. Lackner, TR Wilson, and J. Settleman. Mechanisms of Acquired Resistance to Targeted Cancer Therapies. *Future Oncol.*, 8(8):999–1014, 2012.

[35] T. Sharifnia, V. Rusu, F. Piccioni, M. Bagul, M. Imielinski, A. D. Cherniack, C. S. Pedamallu, B. Wong, F. H. Wilson, L. A. Garraway, D. Altshuler, T. R. Golub, D. E. Root, A. Subramanian, and M. Meyerson. Genetic modifiers of EGFR dependence in non-small cell lung cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 111(52):18661–18666, Dec 2014.

[36] Ksenija Asić. Dominant mechanisms of primary resistance differ from dominant mechanisms of secondary resistance to targeted therapies. *Critical Reviews in Oncology/Hematology*, 97:178 – 196, 2016.

[37] T. J. Stuhlmiller, S. M. Miller, J. S. Zawistowski, K. Nakamura, A. S. Beltran, J. S. Duncan, S. P. Angus, K. A. Collins, D. A. Granger, R. A. Reuther, L. M. Graves, S. M. Gomez, P. F. Kuan, J. S. Parker, X. Chen, N. Sciaky, L. A. Carey, H. S. Earp, J. Jin, and G. L. Johnson. Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. *Cell Rep*, 11(3):390–404, Apr 2015.

[38] W. Kolch, M. Halasz, M. Granovskaya, and B. N. Kholodenko. The dynamic control of signal transduction networks in cancer cells. *Nat. Rev. Cancer*, 15(9): 515–527, Sep 2015.

[39] A. K. M. Azad, A. Lawen, and J. M. Keith. Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling. *BMC Syst Biol*, 9(1):2, Jan 2015.

[40] R. Donaldson and M. Calder. Modular modelling of signalling pathways and their cross-talk. *Theor. Comput. Sci.*, 456, October 2012.

[41] X. Guo and X. F. Wang. Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell Res.*, 19(1):71–88, Jan 2009.

[42] I. Catt. Crosstalk (noise) in digital systems. *IEEE Transactions on Electronic Computers*, EC-16(6):743–763, Dec 1967.

[43] J. A. Engelman, K. Zejnullahu, T. Mitsudomi, Y. Song, C. Hyland, J. O. Park, N. Lindeman, C. M. Gale, X. Zhao, J. Christensen, T. Kosaka, A. J. Holmes, A. M. Rogers, F. Cappuzzo, T. Mok, C. Lee, B. E. Johnson, L. C. Cantley, and P. A. Janne. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science*, 316(5827):1039–1043, May 2007.

[44] J. Bean, C. Brennan, J. Y. Shih, G. Riely, A. Viale, L. Wang, D. Chitale, N. Motoi, J. Szoke, S. Broderick, M. Balak, W. C. Chang, C. J. Yu, A. Gazdar, H. Pass, V. Rusch, W. Gerald, S. F. Huang, P. C. Yang, V. Miller, M. Ladanyi, C. H. Yang, and W. Pao. MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. *Proc. Natl. Acad. Sci. U.S.A.*, 104(52):20932–20937, Dec 2007.

[45] D. Liska, C. T. Chen, T. Bachleitner-Hofmann, J. G. Christensen, and M. R. Weiser. HGF rescues colorectal cancer cells from EGFR inhibition via MET activation. *Clin. Cancer Res.*, 17(3):472–482, Feb 2011.

[46] F. Morgillo, W. Y. Kim, E. S. Kim, F. Ciardiello, W. K. Hong, and H. Y. Lee. Implication of the insulin-like growth factor-IR pathway in the resistance of non-small cell lung cancer cells to treatment with gefitinib. *Clin. Cancer Res.*, 13(9): 2795–2803, May 2007.

[47] Y. Lu, X. Zi, Y. Zhao, D. Mascarenhas, and M. Pollak. Insulin-like growth factor-I receptor signaling and resistance to trastuzumab (Herceptin). *J. Natl. Cancer Inst.*, 93(24):1852–1857, Dec 2001.

[48] F. Cappuzzo, M. Varella-Garcia, G. Finocchiaro, M. Skokan, S. Gajapathy, C. Carnaghi, L. Rimassa, E. Rossi, C. Ligorio, L. Di Tommaso, A. J. Holmes, L. Toschi, G. Tallini, A. Destro, M. Roncalli, A. Santoro, and P. A. Janne. Primary resistance to cetuximab therapy in EGFR FISH-positive colorectal cancer patients. *Br. J. Cancer*, 99(1):83–89, Jul 2008.

[49] Z. Zhang, J. C. Lee, L. Lin, V. Olivas, V. Au, T. LaFramboise, M. Abdel-Rahman, X. Wang, A. D. Levine, J. K. Rho, Y. J. Choi, C. M. Choi, S. W. Kim, S. J. Jang, Y. S. Park, W. S. Kim, D. H. Lee, J. S. Lee, V. A. Miller, M. Arcila, M. Ladanyi, P. Moonsamy, C. Sawyers, T. J. Boggon, P. C. Ma, C. Costa, M. Taron, R. Rosell, B. Halmos, and T. G. Bivona. Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nat. Genet.*, 44(8):852–860, Aug 2012.

[50] L. A. Byers, L. Diao, J. Wang, P. Saintigny, L. Girard, M. Peyton, L. Shen, Y. Fan, U. Giri, P. K. Tumula, M. B. Nilsson, J. Gudikote, H. Tran, R. J. Cardnell, D. J. Bearss, S. L. Warner, J. M. Foulks, S. B. Kanner, V. Gandhi, N. Krett, S. T. Rosen, E. S. Kim, R. S. Herbst, G. R. Blumenschein, J. J. Lee, S. M. Lippman, K. K. Ang, G. B. Mills, W. K. Hong, J. N. Weinstein, I. I. Wistuba, K. R. Coombes, J. D. Minna, and J. V. Heymach. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.*, 19(1):279–290, Jan 2013.

[51] S. A. Kono, M. E. Marshall, K. E. Ware, and L. E. Heasley. The fibroblast growth factor receptor signaling pathway as a mediator of intrinsic resistance to EGFR-specific tyrosine kinase inhibitors in non-small cell lung cancer. *Drug Resist. Updat.*, 12(4-5):95–102, 2009.

[52] G. Zhuang, D. M. Brantley-Sieders, D. Vaught, J. Yu, L. Xie, S. Wells, D. Jackson, R. Muraoka-Cook, C. Arteaga, and J. Chen. Elevation of receptor tyrosine kinase EphA2 mediates resistance to trastuzumab therapy. *Cancer Res.*, 70(1):299–308, Jan 2010.

[53] S. Misale, R. Yaeger, S. Hobor, E. Scala, M. Janakiraman, D. Liska, E. Valtorta, R. Schiavo, M. Buscarino, G. Siravegna, K. Bencardino, A. Cercek, C. T. Chen, S. Veronese, C. Zanon, A. Sartore-Bianchi, M. Gambacorta, M. Gallicchio, E. Vakiani, V. Boscaro, E. Medico, M. Weiser, S. Siena, F. Di Nicolantonio, D. Solit, and A. Bardelli. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, 486(7404):532–536, Jun 2012.

[54] L. A. Diaz, R. T. Williams, J. Wu, I. Kinde, J. R. Hecht, J. Berlin, B. Allen, I. Bozic, J. G. Reiter, M. A. Nowak, K. W. Kinzler, K. S. Oliner, and B. Vogelstein. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):537–540, Jun 2012.

[55] K. Ohashi, L. V. Sequist, M. E. Arcila, T. Moran, J. Chmielecki, Y. L. Lin, Y. Pan, L. Wang, E. de Stanchina, K. Shien, K. Aoe, S. Toyooka, K. Kiura, L. Fernandez-Cuesta, P. Fidias, J. C. Yang, V. A. Miller, G. J. Riely, M. G. Kris, J. A. Engelman, C. L. Vnencak-Jones, D. Dias-Santagata, M. Ladanyi, and W. Pao. Lung cancers with acquired resistance to EGFR inhibitors occasionally harbor BRAF gene mutations but lack mutations in KRAS, NRAS, or MEK1. *Proc. Natl. Acad. Sci. U.S.A.*, 109(31):E2127–2133, Jul 2012.

[56] Y. Nagata, K. H. Lan, X. Zhou, M. Tan, F. J. Esteva, A. A. Sahin, K. S. Klos, P. Li, B. P. Monia, N. T. Nguyen, G. N. Hortobagyi, M. C. Hung, and D. Yu.

PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell*, 6(2):117–127, Aug 2004.

[57] K. Berns, H. M. Horlings, B. T. Hennessy, M. Madiredjo, E. M. Hijmans, K. Beelen, S. C. Linn, A. M. Gonzalez-Angulo, K. Stemke-Hale, M. Hauptmann, R. L. Beijersbergen, G. B. Mills, M. J. van de Vijver, and R. Bernards. A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer Cell*, 12(4):395–402, Oct 2007.

[58] H. Yamaguchi, J. L. Hsu, and M. C. Hung. Regulation of ubiquitination-mediated protein degradation by survival kinases in cancer. *Front Oncol*, 2:15, 2012.

[59] J. Y. Yang and M. C. Hung. A new fork for clinical application: targeting forkhead transcription factors in cancer. *Clin. Cancer Res.*, 15(3):752–757, Feb 2009.

[60] J. Y. Yang, C. J. Chang, W. Xia, Y. Wang, K. K. Wong, J. A. Engelman, Y. Du, M. Andreeff, G. N. Hortobagyi, and M. C. Hung. Activation of FOXO3a is sufficient to reverse mitogen-activated protein/extracellular signal-regulated kinase kinase inhibitor chemoresistance in human cancer. *Cancer Res.*, 70(11): 4709–4718, Jun 2010.

[61] D. F. Lee, H. P. Kuo, C. T. Chen, J. M. Hsu, C. K. Chou, Y. Wei, H. L. Sun, L. Y. Li, B. Ping, W. C. Huang, X. He, J. Y. Hung, C. C. Lai, Q. Ding, J. L. Su, J. Y. Yang, A. A. Sahin, G. N. Hortobagyi, F. J. Tsai, C. H. Tsai, and M. C. Hung. IKK beta suppression of TSC1 links inflammation and tumor angiogenesis via the mTOR pathway. *Cell*, 130(3):440–455, Aug 2007.

[62] C. J. Yen, J. G. Izzo, D. F. Lee, S. Guha, Y. Wei, T. T. Wu, C. T. Chen, H. P. Kuo, J. M. Hsu, H. L. Sun, C. K. Chou, N. S. Buttar, K. K. Wang, P. Huang, J. Ajani, and M. C. Hung. Bile acid exposure up-regulates tuberous sclerosis complex 1/mammalian target of rapamycin pathway in Barrett's-associated esophageal adenocarcinoma. *Cancer Res.*, 68(8):2632–2640, Apr 2008.

[63] M. C. Hu, D. F. Lee, W. Xia, L. S. Golfman, F. Ou-Yang, J. Y. Yang, Y. Zou, S. Bao, N. Hanada, H. Saso, R. Kobayashi, and M. C. Hung. IkappaB kinase promotes tumorigenesis through inhibition of forkhead FOXO3a. *Cell*, 117(2): 225–237, Apr 2004.

[64] J. L. Su, X. Cheng, H. Yamaguchi, Y. W. Chang, C. F. Hou, D. F. Lee, H. W. Ko, K. T. Hua, Y. N. Wang, M. Hsiao, P. B. Chen, J. M. Hsu, R. C. Bast, G. N. Hortobagyi, and M. C. Hung. FOXO3a-Dependent Mechanism of E1A-Induced Chemosensitization. *Cancer Res.*, 71(21):6878–6887, Nov 2011.

[65] A. N. Tegge, N. Sharp, and T. M. Murali. Xtalk: a path-based approach for identifying crosstalk between signaling pathways. *Bioinformatics*, 32(2):242–251, Jan 2016.

[66] T. Wang, J. Gu, J. Yuan, R. Tao, Y. Li, and S. Li. Inferring pathway crosstalk networks using gene set co-expression signatures. *Mol Biosyst*, 9(7):1822–1828, Jul 2013.

[67] Guillermo de Anda-Jáuregui, Raúl A. Mejía-Pedroza, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Crosstalk events in the estrogen signaling pathway may affect tamoxifen efficacy in breast cancer molecular subtypes. *Computational Biology and Chemistry*, 59, Part B:42 – 54, 2015. ISSN 1476-9271. Advances in Systems Biology.

[68] M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. Mackenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Draghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.*, 23(11):1885–1893, Nov 2013.

[69] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, Jan 2009.

[70] J. Han, X. Shi, Y. Zhang, Y. Xu, Y. Jiang, C. Zhang, L. Feng, H. Yang, D. Shang, Z. Sun, F. Su, C. Li, and X. Li. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. *Sci Rep*, 5:13044, Aug 2015.

[71] J. Han, C. Li, H. Yang, Y. Xu, C. Zhang, J. Ma, X. Shi, W. Liu, D. Shang, Q. Yao, Y. Zhang, F. Su, L. Feng, and X. Li. A novel dysregulated pathway-identification analysis based on global influence of within-pathway effects and crosstalk between pathways. *J R Soc Interface*, 12(102):20140937, Jan 2015.

[72] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4 (1):44–57, 2009.

[73] J. T. Chang and J. R. Nevins. GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics*, 22(23):2926–2933, Dec 2006.

[74] Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, Oct 2007.

[75] S. Bulashevska, A. Bulashevska, and R. Eils. Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. *BMC Bioinformatics*, 11:46, 2010.

[76] Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[77] D Strauss and M Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, March 1990.

[78] Tom A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3, 2002.

[79] Peter Congdon. *Introduction: The Bayesian Method, its Benefits and Implementation*, pages 1–23. John Wiley & Sons, Ltd, 2006.

[80] Stephen P. Brooks. Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):69–100, 1998.

[81] B Walsh. Markov Chain Monte Carlo and Gibbs Sampling: Lecture notes for EEB 581, 2004.

[82] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[83] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[84] Robert L. Smith. The hit-and-run sampler: A globally reaching markov chain sampler for generating arbitrary multivariate distributions. In *Proceedings of the 28th Conference on Winter Simulation*, WSC '96, pages 260–264. IEEE Computer Society, 1996.

[85] Jonathan Keith, George Sofronov, and Dirk Kroese. *The Generalized Gibbs Sampler and the Neighborhood Sampler*, pages 537–547. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[86] David J. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, October 2000. ISSN 0960-3174.

Chapter 3

Cross-talk categorisations in data-driven models of signalling networks: a system-level view

Chapter Objectives

In this chapter, I review some state-of-the-art approaches for categorising signalling cross-talk and argue that they are not suitable for application to data-driven signalling networks. I propose a novel cross-talk categorisation specific to data-driven network models which can be mapped to all types of signalling cross-talk defined in other state-of-the-art approaches. I also provide a simple but intuitive algorithm called XDaMoSiN (Cross-talks in Data-driven Models of Signalling Networks) to detect all cross-talk between any two given signalling pathways in a data-driven network.

Authorship

A. K. M. Azad[1], Alfons Lawen[2], Jonathan M Keith[1]

1 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

2 Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia

Reference

Azad A.K.M., Lawen A., Keith JM. (2016). Cross-talk categorisations in data-driven models of signalling networks: a system-level view. [Submitted] **BMC Research Notes**

RESEARCH

Cross-talk categorisations in data-driven models of signalling networks: a system-level view

A. K. M. Azad[1*], Alfons Lawen[2] and Jonathan M. Keith[1]

*Correspondence:
a.azad@monash.edu
[1]School of Mathematical Science, Monash University, Wellington Road, Clayton, VIC, Australia
Full list of author information is available at the end of the article

Abstract

Background: Data-driven models of signalling networks are becoming increasingly important in systems biology in order to reflect dynamic patterns of signalling activities in a context-specific manner. State-of-the-art approaches for categorising and detecting signalling cross-talks may not be suitable for such models since they rely on static topologies of cell signalling networks and prior biological knowledge.

Results: In this article, we review state-of-the-art approaches that categorise all possible cross-talks in signalling networks, and propose a novel categorisation specific to data-driven network models. Considering such models as undirected networks, we propose two categories of signalling cross-talks between any two given signalling pathways. In a Type-I cross-talk, a signalling link $\{g_i, g_j\}$ connects two signalling pathways, where g_i and g_j are signalling nodes that belong to two distinct pathways. In a Type-II cross-talk, two signalling links $\{g_i, g_j\}$ and $\{g_j, g_k\}$ meet at the intersection of two signalling pathways at a shared signalling node g_j. We compared our categorisation approach with others, and find that all the types of cross-talks defined by those approaches can be mapped to Type-I and Type-II cross-talks when underlying signalling activities are considered as non-causal relationships. Next, we provided a simple but intuitive algorithm called XDaMoSiN (Cross-talks in Data-driven Models of Signalling Networks) to detect both Type-I and Type-II cross-talks between any two given signalling pathways in a data-driven network model.

Conclusion: By detecting cross-talks in such network models, our approach can be used to analyse and decipher latent mechanisms of various cell phenotypes, such as cancer or acquired drug resistance, that may evolve due to the highly adaptable and dynamic nature of signal transduction networks.

Keywords: signalling cross-talks; data-driven models; signalling network; cancer signalling; signal re-wiring; acquired drug resistance

Background

A signal transduction network is a collection of all cell signalling pathways where each pathway is a series of biochemical events, transmitting input signals from receptor proteins to intracellular target proteins (e.g. transcription factors). The outcomes mediated by signalling pathways include various cellular activities, such as cell growth, proliferation, differentiation, migration, adhesion, and apoptosis [1, 2]. Interactions among distinct signalling pathways are called signalling cross-talks and may also play vital roles in mediating or modulating cellular activities [3] under different disease-related cell conditions, such as cancer and acquired drug resistance.

Models of signal transduction networks often take a qualitative approach that relies on prior biological knowledge obtained from experimental findings in various cell-lines [4, 5]. However, the pattern of cell signalling activities is not static, and can vary in different cell-lines [4, 5]. Moreover, different cell-lines for which the underlying network architectures of signalling activities are conserved may yield different responses even in similar experimental settings [5]. In the same cell, different ligands can produce different signalling connections [5, 6]. Moreover, different drugs and different treatment conditions may also induce different signalling dependencies, and thus create a dynamic re-wiring in the signalling network topology [6–8]. Therefore, understanding a signalling network topology demands a data-driven modelling approach in order to reflect its context-specific nature in a particular cell-type, and a particular experimental configuration. Here, data-driven models of signalling networks are models in which network edges are inferred solely based on signalling data [4] using machine learning approaches, such as least square regression [9], Bayesian networks [10–12], and time-lag correlation [13]. In contrast, static models of signalling networks are based on canonical signalling mechanisms obtained from the literature [4]. Recent advancements in high-throughput data generation techniques facilitate the quantification of signalling responses, and thereby produce large volumes of data measuring protein abundances and activities [4].

Detecting signalling cross-talks using data-driven models of signalling networks is an important task in systems biology since such cross-talks may reveal novel mechanistic details underlying perturbed cellular conditions. RTK (Receptor Tyrosine Kinase) heterodimerisation is one of the forms of signalling cross-talks (also known as receptor function cross-talks [14]), which has been reported to be involved in the processes of tumourigenesis and developing acquired drug resistance in many cancers [6]. Usually, EGFR (Epidermal Growth Factor Receptor) strongly activates ERK (Extracellular signal-Regulated Kinase) signalling, but it is also a weak activator of the PI3K (Phosphatidylinositol 3-Kinase) signalling pathway. Interestingly, when EGFR cross-talks with HER2 (Human Epidermal Growth Factor Receptor 2) through heterodimerisation it activates both signalling pathways significantly [15], and thereby contributes to tumourigenesis by stimulating proliferation and preventing cell death [6]. In another example, the RTK expression of AXL was found to be a mechanism of acquired resistance to EGFR inhibitors [16], and AXL is found to be transactivated by EGFR through heterodimerisation (cross-talk) [6].

In this article, we review existing approaches that have been used in the literature to categorise cross-talks in signalling networks. However, all these methods are limited in application to *static* models of signalling networks, and cannot be used to categorise cross-talks when the types of signalling activities (e.g. reaction, catalysis, or inhibition) are not known. We therefore introduce a novel cross-talk categorisation for a *single cell* model to resolve such issues. We also compare our categorisation with the existing approaches. Lastly, we present an algorithm to computationally detect all signalling cross-talks that are included in our proposed categorisation. Nataranjan *et al.* [17] report that a global analysis of both known and novel cross-talks can reveal system-level insights into context-dependent signalling: many ligand stimuli converge on a relatively small number of signalling molecules to produce unique responses. Thus, we hypothesise that our approach will be useful

to elucidate similar novel system-level aspects of signalling networks derived from context-specific signalling data through the identification of cross-talks.

Existing methods for categorising cross-talks

Only a few studies have attempted to categorise types or modes of cross-talks between two signalling pathways [6, 14, 18]. In reviewing signalling cross-talks between TGF-β/BMP (Transforming Growth Factor-β, Bone Morphogenic Protein) and other signalling pathways, Guo *et al.* [18] distinguish three different modes of signalling cross-talks. According to that study, two pathways: $pathway_1$ and $pathway_2$ cross-talk when 1) some component of $pathway_1$ physically interacts with some component in $pathway_2$ (Mode-A), 2) some component of $pathway_2$ plays a role as an enzymatic or transcriptional target of some component in $pathway_2$ (Mode-B), or 3) signals from $pathway_1$ modulate or compete for a key modulator or mediator protein that is shared between $pathway_1$ and $pathway_2$ (Mode-C).

Donaldson *et al.* [14] proposed five types of signalling cross-talk between any two signalling pathways: $pathway_1$ and $pathway_2$. They are as follows:

- *Signal-flow cross-talk*: An alternative reaction that enhances the signalling in $pathway_1$ by producing, or catalysing, or inhibiting the production of a protein mediated by the signalling of $pathway_2$. For example, there exists signal-flow cross-talk between MAPK (Mitogen-Activated Protein Kinase) and integrin signalling pathways [19] where the increased rate of activation of some key protein in the integrin pathway is mediated by signalling through the MAPK pathway.
- *Receptor function cross-talk*: An alternative reaction to activate/inhibit the receptor of $pathway_1$ by some enzyme of $pathway_2$ without the need of a ligand (a protein that activates a receptor protein). For example, oestrogen receptor may become activated without the need of oestrogen ligand by other signalling pathways [20].
- *Gene expression cross-talk*: A component (typically, a protein) of $pathway_1$ inhibits or modifies the transcription or protein production of genes in $pathway_2$. For example, transcription factor GR (Glucocorticoid Receptor) of hormone signalling pathways translocates to the nucleus and inhibits the transcriptional activities of the transcription factor NF-κB (Nuclear Factor-κB) that is activated in response to inflammatory stimuli and environmental stressors [21].
- *Substrate availability cross-talk*: $pathway_1$ and $pathway_2$ share a protein (or a set of proteins) and both of the pathways compete for the activation of that shared protein(s). For example, two MAPK pathways in the yeast *S. cerevisiae* that share MAPKKK (Mitogen-Activated Protein Kinase Kinase Kinase) protein STE11 (Sterility gene 11) and possess homologous MAPKK (Mitogen-Activated Protein Kinase Kinase) and MAPK proteins compete for the activation of the MAPK cascade [22].
- *Intracellular communication cross-talk*: The gene products of $pathway_1$ act as ligands for the receptor of $pathway_2$. For example, TGF-β and Wnt (Wingless-related integration site) signalling regulate the production of ligands of one another [18].

Donaldson *et al.* [14] also reviewed some computational models that deal with cross-talks between specific pathways including MAPK pathway, AKT pathways, and PKC (Protein Kinase C) pathways. These models [22–24] use Ordinary Differential Equations (ODEs) where the notion of the cross-talk was a part of the system of equations without any explicit way of detecting or categorising them [14].

Kolch *et al.* [6] describe three types of cross-talks: heterodimerisation between signalling proteins, node sharing, and competition for nodes. Signalling protein heterodimerisation is a biochemical process where a protein complex is formed by two different macromolecules; and RTK heterodimerisation is a common example of this type of cross-talk [6]. For example, EGFR heterodimerisation with ErbB2 (Erythroblastic Leukemia Viral Oncogene B2, also known as HER2) or ErbB3 (Erythroblastic Leukemia Viral Oncogene B3) (also known as HER3, Human Epidermal Growth Factor Receptor 3) activates both ERK and PI3K signalling pathways [15], and thereby mediates proliferation and cell survival signals in tumourigenesis [6]. In another example, the transactivation of AXL (an RTK) is caused by EGFR heterodimerisation and the expression of AXL was found to be a mechanism of resistance to EGFR inhibitors [16].

An example of node (i.e. protein) sharing cross-talk is the scaffolding protein (a protein that binds with multiple members of a signalling pathway) GAB (GRB2-associated binding partner) which is shared by two signalling pathways: EGFR and insulin receptor (IR) pathways [25]. Lastly, an example of cross-talk in the form of competition for nodes (i.e. proteins) was recently identified, consisting of a switch-like coordination between proliferation and apoptotic signalling through RAF(Rapidly Accelerated Fibrosarcoma)-ERK signalling and MST2 (Mammalian STE20-like Protein Kinase) signalling [26]. In mammalian cells, RAF1 (Rapidly Accelerated Fibrosarcoma1) inhibits MST2-induced apoptosis (promotes proliferation) [27], whereas RASSF1A (Ras association domain-containing protein 1A) activates MST2 (promotes apoptosis) [28]. Romano *et al.* [26] showed that this signalling coordination is switch-like, since MST2 binds mutually exclusively with its inhibitor RAF1 and activator RASSF1A by changing its binding affinities from low to high.

Identifying the above cross-talk categories requires previous biological knowledge of the nature of signalling links. An essentially *static* model of signal transduction networks is thus assumed. However, in data-driven models of signalling networks, connectivity among signalling nodes may differ from cell to cell [6]. In order to reveal novel signalling dynamics in cell-specific, ligand-specific or treatment-specific contextual data, we define a novel cross-talk categorisation in the following section.

Methods

Proposed cross-talk categorisation in data-driven networks

Approaches for inferring data-driven signalling networks

Although our main focus in this article is to propose a cross-talk categorisation, here we briefly mention some approaches that fit data-driven models of signalling networks to quantitative signalling datasets. Some high-throughput proteomics techniques that quantitatively measure phosphorylation activities of phosphoproteins (signalling proteins) include mass spectrometry, flow-cytometry, RNAi (Ribonucleic Acid Interference) screening, and reverse-phase protein array (RPPA) [13, 29].

Apart from proteomics data, some approaches use gene expression measurements of phosphoproteins as a proxy for protein expression (i.e. protein activity) [30–32] in order to fit data-driven models of signalling networks. However, inference methods include modelling both causal [9–12, 29, 33] and non-causal (simple correlations) relationships [13, 34] among phosphoproteins. To identify causal relationships in a signalling network topology, various approaches have been applied, such as least square regression [9], various models on Bayesian networks [10–12] and dynamic Bayesian networks [29], and maximum entropy [33]. Correlation-based approaches include measuring the simple Pearson correlation [34] and time-lag correlation [13]. The rationale behind applying such simple correlation-based approaches to infer signalling network structure is that individual signals may co-vary with respect to one another [4].

Proposed cross-talk categorisation

In order to generalise our cross-talk categorisation for both causal and non-causal network models, we consider a signalling network as an undirected network. Let $G(V, E)$ be an undirected graph that represents an entire signalling network containing a set of signalling pathways, where V is a set of n signalling components (typically proteins or protein complexes, denoted g_i, for $i = 1, 2, ..., n$) and E is a set of *unordered* pairs of signalling components of the form $\{g_i, g_j\}$ representing signalling links inferred from data. We propose two types of signalling cross-talks between any two signalling pathways, denoted $pathway_1$ and $pathway_2$ [Figure 1]. Here a pathway is defined merely as a list of signalling components, usually obtained from databases such as KEGG [35], Reactome [36], and WikiPathways [37].

Type-I cross-talk: $\{g_i, g_j\} \in E$ is a Type-I cross-talk between $pathway_1$ and $pathway_2$ if $(g_i \in pathway_1 \wedge g_j \in pathway_2) \bigwedge (g_i \notin pathway_2 \wedge g_j \notin pathway_1)$.

Type-II cross-talk: $\{g_i, g_j\} \in E \wedge \{g_j, g_k\} \in E$ is a Type-II cross-talk between $pathway_1$ and $pathway_2$ if $(g_i \in pathway_1 \wedge g_j \in pathway_1) \bigwedge (g_j \in pathway_2 \wedge g_k \in pathway_2)$.

An algorithm for detecting proposed cross-talks

In Table 1, we present a simple but intuitive algorithm for identifying Type-I and Type-II cross-talks in data-driven signalling network models. We refer to our algorithm as XDaMoSiN (Cross-talk in Data-driven Models of Signalling Network). Note that our approach considers data-driven models of signalling networks as undirected networks in order to generalise our categorisation for both causal and non-causal network models. The only assumption we make here is that pathway annotations of signalling pathways are known from pathway databases, such as KEGG [35], Reactome [36] and WikiPathways [37]. In these annotations, a pathway is defined as a list of signalling nodes. Note that the signalling links among these nodes are modelled using data-driven relationships. Therefore, a data-driven model of a signalling network is defined as $G = (V, E)$, where V is a list of n signalling nodes, and E is a list of signalling links $\{g_i, g_j\}$ inferred from data. This algorithm takes two inputs: G (the network) and $PathwayDB$ (a pathway database), and produces two

outputs: $Type\_I\_crosstalk$ and $Type\_II\_crosstalk$, which are two lists containing all Type-I and Type-II cross-talks [Table 1]. Here, we consider $PathwayDB$ as a list, where each element in that list is also a list, containing signalling nodes in a particular pathway, and is indexed by the corresponding pathway ID (typically, the pathway name).

Table 1 Pseudocode for XDaMoSiN Algorithm

```
      XDaMoSiN (G,PathwayDB)
      /* Part#1 : Find  Type − I  crosstalks */
 1      Type_I_crosstalk ← ∅
 2      for each link {g_i, g_j} ∈ E
 3        List_i ← ∅
 4        List_j ← ∅
 5        for each pathway_id ∈ PathwayDB
 6          List_p ← FindList(PathwayDB, pathway_id)
 7          if g_i ∈ List_p
 8            List_i ← List_i ∪ {pathway_id}
 9          end if
10          if g_j ∈ List_p
11            List_j ← List_j ∪ {pathway_id}
12          end if
13        end for
14        if List_i \ List_j is not ∅ and List_j \ List_i is not ∅
15          Type_I_crosstalk ← Type_I_crosstalk ∪ {{g_i, g_j}}
16        end if
17      end for

      /* Part#2 : Find  Type − II  crosstalks */
18      Type_II_crosstalk ← ∅
19      for each g_j ∈ V
20        L_j ← ∅
21        for each pathway_id ∈ PathwayDB
22          L_p ← FindList(PathwayDB, pathway_id)
23          for each g_i ∈ L_p
24            if {g_i, g_j} ∈ E and {g_i, g_j} ⊂ L_p
25              L_j ← L_j ∪ (pathway_id, g_i)
26            end if
27          end for
28        end for
29        for each pair (pathway_id_1, g_i) ∈ L_j
30          for each pair (pathway_id_2, g_k) ∈ L_j
31            if pathway_id_1 is not pathway_id_2
32              Type_II_crosstalk ← Type_II_crosstalk ∪ {{g_i, g_j} ∧ {g_j, g_k}}
33            end if
34          end for
35        end for
36      end for
```

In the first part of the algorithm, we find all the Type-I cross-talks among all the pathways in $PathwayDB$. At first, we initialise the list $Type\_I\_crosstalk$, which collects all such Type-I cross-talks. Then we check each signalling link $\{g_i, g_j\} \in E$ to determine whether it plays a role as Type-I cross-talk. Here, we loop through all pathways, and save pathway IDs that contain g_i or g_j, individually. For this purpose, we maintain two intermediate lists, called $List_i$ and $List_j$, respectively. If $List_i$ contains some pathway IDs that are not in $List_j$, and *vice versa*, then we identify $\{g_i, g_j\}$ as a Type-I cross-talk. Note, we assume here that an intermediate function called $FindList(PathwayDB, pathway\_id)$ exists, which constructs a list of signalling nodes in a particular pathway with ID: $pathwa\_id$ in the $PathwayDB$.

In the second part of the algorithm, we find all Type-II cross-talks. First, we examine each signalling node g_j individually, to determine whether it is shared by more than one pathway and has incident signalling link(s) (from E) in those

pathways. For this purpose, for each signalling node g_j, we construct an intermediate list, called L_j. This list collects ordered pairs of information: each incident signalling link $\{g_i, g_j\} \in E$ is contained in a pathway labelled *pathway_id* and the *pathway_id* itself. Next, for any combination of pairs in the list L_j, such as $(pathway\_id_1, g_i)$ and $(pathway\_id_2, g_k)$, if $pathway\_id_1$ and $pathway\_id_2$ are different, then we define $\{\{g_i, g_j\} \wedge \{g_j, g_k\}\}$ as a Type-II cross-talk between $pathway\_id_1$ and $pathway\_id_2$.

Results

Type-I & Type-II cross-talks include cross-talks from other state-of-the-art categorisations

We compare the cross-talk categorisation approaches, including our proposed methods, in Table 2. This comparison reveals an interesting aspect of these categorisations: *cross-talks between any two pathways can be identified when their corresponding causal relationships are ignored*, i.e. considering the signalling network as an undirected network only. At the same time, we note that our approach can include all types of cross-talks defined by other categorisation.

Type-I cross-talks can represent *signal-flow cross-talks*, *receptor function cross-talks* and *gene-expression cross-talks* from Donaldson *et al.* [14], *Mode-A* and *Mode-B* cross-talks from Guo *et al.* [18], and cross-talk of signalling protein heterodimerisation from Kolch *et al.* [6]. In a cross-talking pair $\{g_i, g_j\}$ in each of these categories, one signalling component g_i belongs to one pathway and g_j belongs to another pathway, or vice versa, but mutually exclusively [Table 2]. Again, Type-II cross-talks represent the cross-talk types of *substrate availability* and *intracellular communications* from Donaldson *et el.* [14], *Mode-C* cross-talks from Guo *et al.* [18], and *Signalling node sharing* and *Competition for nodes* from Kolch *et al.* [6], since in all of these categories there exists a shared component between $pathway_1$ and $pathway_2$ for which the other components of those individual pathways compete for modification or activation of that shared component [Table 2].

Moreover, Donaldson *et al.* [14] reported that their categorisation comprehensively covered all possible types of signalling cross-talks in a single cell model. Since Type-I and Type-II cross-talks include all cross-talks from Donaldson *et al.* [Table 2], we claim that our categorisation is also comprehensive. Moreover, Donaldson *et al.* made a claim that their approach can be useful for detecting cross-talks in data-driven models of signalling networks. However, we note that their proposed algorithm (see the appendix of [14]) was based on qualitative logic only, and is not explicit how that could be used for dealing with network models derived from *high-throughput* quantitative signalling data, such as mass spectometry and RPPA data. Moreover, since they used modular architecture of signal propagation (receptor function, 3-stage cascade, and gene expression [14]) in detecting all signalling cross-talks, their approach is not suitable for models derived from gene expression data only. There are some studies [30–32] that attempted to infer signalling network topology using gene expression as a proxy for signalling protein activities, since gene expression data is usually cheaper to generate and is possible to produce in large-scale [32].

Table 2 Comparative categorisations of signalling cross-talks

Proposed Categorisation	Related Study		
	Donaldson et al. [14]	**Guo et al. [18]**	**Kolch et al. [6]**
1) *Type-I cross-talk*: $\{g_i, g_j\} \in E$ s.t. $(g_i \in pathway_1 \wedge g_j \in pathway_2) \bigwedge (g_i \notin pathway_2 \wedge g_j \notin pathway_1)$	1) *Signal-flow cross-talk*: An alternate reaction to activate a protein 'Y' through an enzyme 'X' 2) *Receptor function cross-talk*: An alternate reaction to activate a receptor 'R' by an enzyme 'X' 3) *Gene expression cross-talk*: Activate/Inhibit the expression of a gene 'g' by a protein 'Y'	1) *Mode-A*: Components of one pathway physically interact with components of another pathway 2) *Mode-B*: Components of one pathway are enzymatic or transcriptional targets of components of another pathway	1) *Heterodimerisation between signalling proteins*: Two different signalling proteins from two different signalling pathways bind with each other
2) *Type-II cross-talk*: $\{g_i, g_j\} \in E \wedge \{g_j, g_k\} \in E$ s.t. $(g_i \in pathway_1 \wedge g_j \in pathway_1) \bigwedge (g_j \in pathway_2 \wedge g_k \in pathway_2)$	4) *Substrate availability cross-talk*: Pathways compete for activation of a shared protein 'Y' 5) *Intra-cellular communication cross-talk*: Output of the expression of a gene 'g' of a pathway acts as ligand of another pathway	3) *Mode-C*: One pathway modulates or competes for a key modulator or mediator of another	1) *Signalling node sharing*: A signalling node that is shared by two different signalling pathways. 2) *Competition for nodes*: Competing protein interactions coordinately regulate a signalling node mutually exclusively

Here, $g_i, g_j, g_k \in V$, V and E are the set of signalling components and signalling links, respectively

Discussion & Conclusion

The data-driven modelling of signalling networks and the detection of cross-talks in those models provide effective ways to elucidate novel mechanisms of perturbed signalling activities in various disease conditions, such as cancer and drug resistance. In this article, we reviewed some state-of-the-art approaches that categorise signalling cross-talks and identified a limitation of their applicabilities to data-driven models, since they rely on a static topology of signalling networks. Here, we propose a novel cross-talk categorisation (Type-I and Type-II) that can not only be applicable to data-driven models, but also generalises all types of cross-talks defined by other approaches. We also present a simple but intuitive algorithm for detecting Type-I and Type-II cross-talks between any two signalling pathways. In combination with other computational and statistical methodologies, our approach is useful in systems biology to generate novel but biologically plausible hypotheses in a data-dependent manner.

Abbreviations

XDaMoSiN: Cross-talks in Data-driven Models of Signalling Networks; **RTK**: Receptor Tyrosine Kinase; **EGFR**: Epidermal Growth Factor Receptor; **ERK**: Extracellular signal-Regulated Kinase; **PI3K**: Phosphatidylinositol 3-Kinase; **HER2**: Human Epidermal Growth Factor Receptor 2; **TGF**-β: Transforming Growth Factor-β; **BMP**: Bone Morphogenic Protein; **MAPK**: Mitogen-Activated Protein Kinase; **GR**: Glucocorticoid Receptor; **NF**-κ**B**: Nuclear Factor-κB; **MAPKKK**: Mitogen-Activated Protein Kinase Kinase Kinase; **MAPKK**: Mitogen-Activated Protein Kinase Kinase; **STE11**: Sterility gene 11; **Wnt**: Wingless-related integration site; **PKC**: Protein Kinase C; **ODEs**: Ordinary Differential Equations; **ErbB2**: Erythroblastic Leukemia Viral Oncogene B2; **ErbB3**: Erythroblastic Leukemia Viral Oncogene B3; **HER3**: Human Epidermal Growth Factor Receptor 3; **GAB**: GRB2-associated binding partner; **IR**: Insulin Receptor; **RAF**: Rapidly Accelerated Fibrosarcoma; **MST2**: Mammalian STE20-like Protein Kinase; **RAF1**: Rapidly Accelerated Fibrosarcoma1; **RASSF1A**: Ras association domain-containing protein 1A; **RNAi**: Ribonucleic Acid Interference; **RPPA**: Reverse-Phase Protein Array;

Declarations

Authors' contributions

AKMA conceived the idea, surveyed literature, proposed the cross-talk categorisation, designed the algorithm, compared with others, and wrote the manuscript; JMK and AL supervised this work and provided guidance in writing the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics (and consent to publish)

Not applicable

Consent to publish

Not applicable

Availability of data and materials

The datasets supporting the results of this article are included in the article

Acknowledgements

This research was supported by Monash International Postgraduate Research Scholarship and Monash Graduate Scholarship at the Monash University, Australia.

Author details

[1]School of Mathematical Science, Monash University, Wellington Road, Clayton, VIC, Australia. [2]Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University, Wellington Road, Clayton, VIC, Australia.

References

1. Brivanlou, A.H., Darnell, J.E.: Signal transduction and the control of gene expression. Science **295**(5556), 813–818 (2002)
2. Sarkar, S., Mandal, M.: Growth factor receptors and apoptosis regulators: signaling pathways, prognosis, chemosensitivity and treatment outcomes of breast cancer. Breast Cancer (Auckl) **3**, 47–60 (2009)

3. Ding, S., Chamberlain, M., McLaren, A., Goh, L., Duncan, I., Wolf, C.R.: Cross-talk between signalling pathways and the multidrug resistant protein MDR-1. Br. J. Cancer **85**(8), 1175–1184 (2001)
4. Janes, K.A., Yaffe, M.B.: Data-driven modelling of signal-transduction networks. Nat. Rev. Mol. Cell Biol. **7**(11), 820–828 (2006)
5. Schoeberl, B., Pace, E., Howard, S., Garantcharova, V., Kudla, A., Sorger, P.K., Nielsen, U.B.: A data-driven computational model of the ErbB receptor signaling network. Conf Proc IEEE Eng Med Biol Soc **1**, 53–54 (2006)
6. Kolch, W., Halasz, M., Granovskaya, M., Kholodenko, B.N.: The dynamic control of signal transduction networks in cancer cells. Nat. Rev. Cancer **15**(9), 515–527 (2015)
7. Santos, S.D., Verveer, P.J., Bastiaens, P.I.: Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nat. Cell Biol. **9**(3), 324–330 (2007)
8. von Kriegsheim, A., Baiocchi, D., Birtwistle, M., Sumpton, D., Bienvenut, W., Morrice, N., Yamada, K., Lamond, A., Kalna, G., Orton, R., Gilbert, D., Kolch, W.: Cell fate decisions are specified by the dynamic ERK interactome. Nat. Cell Biol. **11**(12), 1458–1464 (2009)
9. Janes, K.A., Kelly, J.R., Gaudet, S., Albeck, J.G., Sorger, P.K., Lauffenburger, D.A.: Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. J. Comput. Biol. **11**(4), 544–561 (2004)
10. Woolf, P.J., Prudhomme, W., Daheron, L., Daley, G.Q., Lauffenburger, D.A.: Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics **21**(6), 741–753 (2005)
11. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science **308**(5721), 523–529 (2005)
12. Sachs, K., Itani, S., Carlisle, J., Nolan, G.P., Pe'er, D., Lauffenburger, D.A.: Learning signaling network structures with sparsely distributed data. J. Comput. Biol. **16**(2), 201–212 (2009)
13. Santra, T., Kholodenko, B., Kolch, W.: An integrated Bayesian framework for identifying phosphorylation networks in stimulated cells. Adv. Exp. Med. Biol. **736**, 59–80 (2012)
14. Donaldson, R., Calder, M.: Modular modelling of signalling pathways and their cross-talk. Theor. Comput. Sci. **456** (2012)
15. Birtwistle, M.R., Hatakeyama, M., Yumoto, N., Ogunnaike, B.A., Hoek, J.B., Kholodenko, B.N.: Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Mol. Syst. Biol. **3**, 144 (2007)
16. Meyer, A.S., Miller, M.A., Gertler, F.B., Lauffenburger, D.A.: The receptor AXL diversifies EGFR signaling and limits the response to EGFR-targeted inhibitors in triple-negative breast cancer cells. Sci Signal **6**(287), 66 (2013)
17. Natarajan, M., Lin, K.-M., Hsueh, R.C., Sternweis, P.C., Ranganathan, R.: A global analysis of cross-talk in a mammalian cellular signalling network. Nat. Cell Biol. **8**(6), 571–580 (2006)
18. Guo, X., Wang, X.F.: Signaling cross-talk between TGF-beta/BMP and other pathways. Cell Res. **19**(1), 71–88 (2009)
19. Schwartz, M.A., Ginsberg, M.H.: Networks and crosstalk: integrin signalling spreads. Nat. Cell Biol. **4**(4), 65–68 (2002)
20. Katzenellenbogen, B.S.: Estrogen receptors: bioactivities and interactions with cell signaling pathways. Biol. Reprod. **54**(2), 287–293 (1996)
21. De Bosscher, K., Vanden Berghe, W., Haegeman, G.: Cross-talk between nuclear receptors and nuclear factor kappaB. Oncogene **25**(51), 6868–6886 (2006)
22. McClean, M.N., Mody, A., Broach, J.R., Ramanathan, S.: Cross-talk and decision making in MAP kinase pathways. Nat. Genet. **39**(3), 409–414 (2007)
23. Hatakeyama, M., Kimura, S., Naka, T., Kawasaki, T., Yumoto, N., Ichikawa, M., Kim, J.H., Saito, K., Saeki, M., Shirouzu, M., Yokoyama, S., Konagaya, A.: A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. Biochem. J. **373**(Pt 2), 451–463 (2003)
24. Sreenath, S.N., Soebiyanto, R., Mesarovic, M.D., Wolkenhauer, O.: Coordination of crosstalk between MAPK-PKC pathways: an exploratory study. IET Syst Biol **1**(1), 33–40 (2007)
25. Borisov, N., Aksamitiene, E., Kiyatkin, A., Legewie, S., Berkhout, J., Maiwald, T., Kaimachnikov, N.P., Timmer, J., Hoek, J.B., Kholodenko, B.N.: Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. Mol. Syst. Biol. **5**, 256 (2009)
26. Romano, D., Nguyen, L.K., Matallanas, D., Halasz, M., Doherty, C., Kholodenko, B.N., Kolch, W.: Protein interaction switches coordinate Raf-1 and MST2/Hippo signalling. Nat. Cell Biol. **16**(7), 673–684 (2014)
27. O'Neill, E., Rushworth, L., Baccarini, M., Kolch, W.: Role of the kinase MST2 in suppression of apoptosis by the proto-oncogene product Raf-1. Science **306**(5705), 2267–2270 (2004)
28. Matallanas, D., Romano, D., Yee, K., Meissl, K., Kucerova, L., Piazzolla, D., Baccarini, M., Vass, J.K., Kolch, W., O'neill, E.: RASSF1A elicits apoptosis through an MST2 pathway directing proapoptotic transcription by the p73 tumor suppressor protein. Mol. Cell **27**(6), 962–975 (2007)
29. Hill, S.M., Lu, Y., Molina, J., Heiser, L.M., Spellman, P.T., Speed, T.P., Gray, J.W., Mills, G.B., Mukherjee, S.: Bayesian inference of signaling network topology in a cancer cell line. Bioinformatics **28**(21), 2804–2810 (2012)
30. Amadoz, A., Sebastian-Leon, P., Vidal, E., Salavert, F., Dopazo, J.: Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Sci Rep **5**, 18494 (2015)
31. Sebastian-Leon, P., Carbonell, J., Salavert, F., Sanchez, R., Medina, I., Dopazo, J.: Inferring the functional effect of gene expression changes in signaling pathways. Nucleic Acids Res. **41**(Web Server issue), 213–217 (2013)
32. Neapolitan, R., Xue, D., Jiang, X.: Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks. Cancer Inform **13**, 77–84 (2014)
33. Locasale, J.W., Wolf-Yadlin, A.: Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PLoS ONE **4**(8), 6522 (2009)

34. Imami, K., Sugiyama, N., Imamura, H., Wakabayashi, M., Tomita, M., Taniguchi, M., Ueno, T., Toi, M., Ishihama, Y.: Temporal profiling of lapatinib-suppressed phosphorylation signals in EGFR/HER2 pathways. Mol. Cell Proteomics **11**(12), 1741–1757 (2012)
35. Kanehisa, M.: The KEGG database. Novartis Found. Symp. **247**, 91–101 (2002)
36. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P.: The Reactome pathway knowledgebase. Nucleic Acids Res. **42**(Database issue), 472–477 (2014)
37. Kelder, T., van Iersel, M.P., Hanspers, K., Kutmon, M., Conklin, B.R., Evelo, C.T., Pico, A.R.: WikiPathways: building research communities on biological pathways. Nucleic Acids Res. **40**(Database issue), 1301–1307 (2012)

Figures

[height=4.5cm,width=11.5cm,angle=0]Fig1

Figure 1 Proposed categorisations of signalling cross-talks, Type-I (A) and Type-II (B). Here, each of the pathways is a collection of signalling nodes (typically proteins or protein complexes). A Type-I cross-talk is a signalling link $\{g_i, g_j\}$ that connects two signalling pathways where neither of the two pathways contains both signalling nodes, g_i and g_j. A Type-II signalling cross-talk is a pair of signalling links $\{g_i, g_j\}$ and $\{g_j, g_k\}$ residing at the intersection of two signalling pathways with a shared node g_j.

Chapter 4

Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modelling

Chapter Objectives

The overall objective of this chapter is to conduct computational experiments to conduct computational detection and analysis of Type-I cross-talk [Chapter 3] in data-driven signalling networks derived from gene expression datasets of lapatinib (an EGFR/HER2 dual inhibitor)-treated sensitive (parental) and resistant cell-lines, and their roles in acquired drug resistance. To do that, I would like to apply a fully Bayesian statistical modelling approach with p_1-model to elucidate the role of signalling cross-talk between EGFR and other signalling pathways in acquired lapatinib resistance. All the supplementary files are added in Appendix B.

Authorship

A. K. M. Azad[1], Alfons Lawen[2], Jonathan M Keith[1]

1 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

2 Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia

Reference

Azad A.K.M., Lawen A. Keith JM. (2015). Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling. *BMC Systems Biology* 9(1): 1-17. DOI: 10.1186/s12918-014-0135-x. (**citations: 3**)

Azad *et al. BMC Systems Biology* (2015) 9:2
DOI 10.1186/s12918-014-0135-x

BMC Systems Biology

RESEARCH ARTICLE Open Access

Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling

AKM Azad[1*], Alfons Lawen[2] and Jonathan M Keith[1]

Abstract

Background: Initial success of inhibitors targeting oncogenes is often followed by tumor relapse due to acquired resistance. In addition to mutations in targeted oncogenes, signaling cross-talks among pathways play a vital role in such drug inefficacy. These include activation of compensatory pathways and altered activities of key effectors in other cell survival and growth-associated pathways.

Results: We propose a computational framework using Bayesian modeling to systematically characterize potential cross-talks among breast cancer signaling pathways. We employed a fully Bayesian approach known as the p_1-model to infer posterior probabilities of gene-pairs in networks derived from the gene expression datasets of ErbB2-positive breast cancer cell-lines (parental, lapatinib-sensitive cell-line SKBR3 and the lapatinib-resistant cell-line SKBR3-R, derived from SKBR3). Using this computational framework, we searched for cross-talks between EGFR/ErbB and other signaling pathways from Reactome, KEGG and WikiPathway databases that contribute to lapatinib resistance. We identified 104, 188 and 299 gene-pairs as putative drug-resistant cross-talks, respectively, each comprised of a gene in the EGFR/ErbB signaling pathway and a gene from another signaling pathway, that appear to be interacting in resistant cells but not in parental cells. In 168 of these (distinct) gene-pairs, both of the interacting partners are up-regulated in resistant conditions relative to parental conditions. These gene-pairs are prime candidates for novel cross-talks contributing to lapatinib resistance. They associate EGFR/ErbB signaling with six other signaling pathways: Notch, Wnt, GPCR, hedgehog, insulin receptor/IGF1R and TGF-β receptor signaling. We conducted a literature survey to validate these cross-talks, and found evidence supporting a role for many of them in contributing to drug resistance. We also analyzed an independent study of lapatinib resistance in the BT474 breast cancer cell-line and found the same signaling pathways making cross-talks with the EGFR/ErbB signaling pathway as in the primary dataset.

Conclusions: Our results indicate that the activation of compensatory pathways can potentially cause up-regulation of EGFR/ErbB pathway genes (counteracting the inhibiting effect of lapatinib) via signaling cross-talk. Thus, the up-regulated members of these compensatory pathways along with the members of the EGFR/ErbB signaling pathway are interesting as potential targets for designing novel anti-cancer therapeutics.

Keywords: Drug resisance, Signaling cross-talk, Bayesian statistical modeling, p_1-model, EGFR signaling, Breast cancer, Lapatinib

*Correspondence: aaza7@student.monash.edu
[1] School of Mathematical Science, Monash University, Wellington Road, Clayton, VIC, Australia
Full list of author information is available at the end of the article

Background

Cancer development involves a series of events, ranging from tumorigenesis to metastasis, each of which may be caused by perturbations in crucial signal transduction pathways. Recently, drugs (inhibitors) specifically targeting critical components of signaling pathways known to be up-regulated in specific cancers have been used in the clinic. However, success of these inhibitors is limited by the intrinsic potential of cancer cells to acquire drug resistance. Recent advances in both clinical and laboratory research have reported that cancer cells may adopt several mechanisms against particular treatments including adjusting the signaling circuitry, activation of alternative pathways and cross-talks among various pathways to overcome the effects of inhibitors [1,2]. Resistance to a particular drug such as EGFR (Epidermal Growth Factor Receptor) tyrosine kinase inhibitors, may occur not only due to cross-talks among EGFR-mediated pathways, but also due to cross-talks with pathways triggered by other receptors. Therefore, targeting signaling cross-talks may have the potential to sensitize cancer cells to particular inhibitors.

Drug resistance is a major obstacle in drug efficacy that causes cancer cells to be insensitive to targeted inhibitor therapies and/or conventional chemotherapeutic agents [1,2]. However, there are two categories of resistance to inhibitor therapies: *de novo* and acquired [3]. By definition, *de novo* resistance is a phenotypic characteristic present before drug exposure where drugs with proven efficacy fail to cause tumor cells to respond with any significance [2,4,5]. Acquired resistance refers to a situation where the initial sensitivity of tumor cells to drugs discontinues despite or due to continued consumption [2]. It has been reported that the underlying mechanisms of both types of resistance are related, often due to mutation, loss, or up-regulation of some other important signaling proteins or pathways [2,5]. *De novo* drug resistance can be determined by assessing the genetic profiles of tumors for 1) oncogenic addictions to proteins or pathways and 2) other possible genetic alterations conferring resistance [2]. Therefore, targeting *de novo* resistance can enhance drug efficacy and reduce the chance of acquired resistance [5]. Recently, characterizing drug-resistant tumors, and analyzing cell lines that result from the continuous culture of drug-sensitive cells in the presence of an inhibitor have been shown to be successful approaches for identifying changes responsible for acquired resistance [2].

Cross-talk among signaling pathways may play a vital role in cancer drug resistance, especially in receptor targeted therapies. For example, in EGFR/HER2 signaling pathways, cross-talk with other signaling pathways may occur at various levels of signal transduction: receptor level, mediator level and effector level [1]. At the receptor level, other RTKs (receptor tyrosine kinases) having common downstream targets of EGFR/HER2 may become involved in cross-talk with EGFR/HER2 signaling pathways. In many cancers, these alternative RTKs including MET, IGF1R, FGFR and EphA2 become activated or amplified in order to maintain the signals for cell survival and/or proliferation in common downstream pathways, thus nullifying the inhibition of EGFR kinase [6-10]. Cross-talk at mediator level includes the activation/inactivation of major components of mediator pathways by mutation/deletion of oncogenic driver genes, which eventually activates downstream effectors [1]. These constitutive activations/inactivations of mediator pathways are independent of receptors. The effect of signaling cross-talk in drug resistance at effector level is more complex and diverse since there may be numerous effectors of RTKs signaling pathways. Resistance at the effector level may occur when some critical effectors (i.e. TSC, FOXO3) involved in cell survival and proliferation show an altered phenotype caused by other signaling pathways via RTK signaling cross-talk [1]. Additionally, inhibitor sensitivity can be affected by cross-talk between signaling pathways triggered by the targeted RTK and other signaling pathways (triggered by other RTKs). For example, the EGFR/HER2 signaling pathway can cross-talk with Wnt/β-catenin, Notch, and TNFα/IKK/NF-κB signaling pathways to affect the EGFR/HER2 inhibitors' sensitivities [1]. Cross-talk between effector pathways and feedback inhibition is also responsible for the adaptive and dynamic response of cancer cells against inhibitor therapies, for example, compensating the inhibited components to maintain key downstream functions, such as cell survival, proliferation etc. [11].

Lapatinib is a dual tyrosine kinase inhibitor of EGFR and ErbB2/HER2 receptors [12] that is used in combination therapy of ErbB2/HER2-positive breast cancer patients with advanced or metastatic tumors [13]. Several studies have examined the mechanism underlying lapatinib resistance at the molecular [14-16] and system level [17], active in HER2-positive breast cancer cell-lines through signaling pathways. Garrett *et al.* [14] reported over-expression of *HER2* or *HER3* in lapatinib-resistant SKBR3 and BT474 breast cancer cell lines. Over-expression of AXL tyrosine kinase was found in the BT474 cell-line [16], but interestingly a switched addiction from HER2 to FGFR2 pathway caused the UACC812/LR cell-line to become resistant to lapatinib [15]. Moreover, a detailed analysis of the global cellular network by Komurov *et al.* [17] revealed that up-regulation of the glucose deprivation response pathway compensates for the lapatinib inhibition in SKBR3 cell-line by providing an EGFR/ErbB2-independent mechanism of glucose uptake and survival [17]. Thus, the activation or up-regulation of compensatory pathways confers poor sensitivity of inhibitors (i.e. lapatinib resistance) in EGFR or ErbB2 targeted therapy [1,2,17]. The

identification and analyses of potential cross-talks among the signaling pathways may provide deeper insights into the mechanism of drug resistance, and can facilitate finding a range of compensatory pathways for overcoming resistance in targeted therapy.

In this study, we collected the gene expression values of the ErbB2-positive parental SKBR3 cell-line and the lapatinib-resistant SKBR3-R cell-line, derived from it, in the presence and absence of lapatinib [17]. Then we used a fully Bayesian statistical modeling approach to identify and analyze characteristic drug-resistant cross-talks between EGFR/ErbB and other signaling pathways. ln that process, we considered two gene-gene networks originating from the gene expression matrices of both parental and resistant conditions, individually. To say a gene-pair involved in cross-talk between two particular signaling pathways has high potential of being involved in acquired drug-resistance, our research hypothesis was it should have high probability of appearing in the resistant network and low probability in the parental network. The rationale behind our hypothesis was that in breast cancer cell lines resistant to tamoxifen, a cross-talk mechanism has previously been identified between EGFR and the IGF1R signaling pathway [18]. The schematic diagram of our proposed framework is shown in Figure 1. Like other biological processes, cancer signaling pathway activities and their corresponding network data possess stochasticity such that some gene-gene relationships (i.e. network edges) may not always be present or detected, whereas some other typical relationships may be absent. The stochastic nature of biological systems can be used to predict edge probabilities by formalizing them into a probabilistic model with other network properties [19]. Hill *et al.* reported a data-driven approach that exploits a Dynamic Bayesian Network (DBN) model to infer probabilistic relationships between node-pairs in a context-specific signaling network [20]. This study incorporates existing signaling biology using an informative prior distribution on the network, and its weight of contribution is measured with an empirical Bayes analysis, maximum marginal likelihood. This study predicts a number of known and unexpected signaling links through time that are validated using independent targeted inhibition experiments [20]. Here we have used a fully Bayesian approach for inferring a probabilistic model: a special class of Exponential Random Graph Model, namely the p_1-model. We used Gibbs sampling for estimating model parameters with non-informative priors, in order to estimate the posterior probabilities of edges in gene-gene relationship networks. These identified cross-talks do not appear in the parental network but only in the resistant one, because the signaling network can be 'rewired' in a specific context [21,22]. This idea resembles the approach taken by Hill *et al.* in that they inferred the probabilities of signaling links (gene-pairs) varying through time. Thus, these drug-resistance cross-talks can be informative to elucidate the complex mechanisms underlying drug-insensitivity and can help to develop novel therapeutics targeting signaling pathways.

Materials and method

Dataset

A global gene expression (GE) dataset (GSE38376) from 1) cells sensitive to lapatinib (said to be under "parental conditions") and 2) cells with acquired resistance to lapatinib was obtained from Komurov *et al.* [17]. Expression values were measured using Illumina HumanHT-12 V3.0 expression beadchip (GPL6947). Samples include SKBR3 parental and resistant (SKBR3-R) each under basal conditions and in response to 0.1 μM and 1 μM lapatinib after 24 hours, where the resistant cell line variant (SKBR3-R) showed 100-fold more resistance to lapatinib treatment than the parental SKBR3 cell line, as reported by Komurov *et al.* [17]. These gene expression datasets used probe-level annotation, which we converted into gene-level annotation. To obtain gene-level GE values, probes were mapped to gene symbols using the corresponding annotation file (GPL6947). While mapping, the average GE values were calculated across all probes if the same gene symbol was annotated to multiple probes. Two GE data matrices were constructed for parental SKBR3 cell lines and resistant SKBR3-R cell lines, respectively, where rows were labelled with gene symbols and columns were labelled with different treatment conditions (0, 0.1 μM and 1 μM of lapatinib).

Construction of a gene-gene relationship network

We define the gene-gene relationship network as $GGR := (S, R)$ for each GE data matrix. Here, S is a set of 370 cancer related genes collected from the Cancer Gene Census [23]. R is defined as the set of pair-wise relationships among seed genes. A gene pair ($gene_i$, $gene_j$) is included in R if the corresponding *absolute* Pearson Correlation Coefficient (PCC) is above some threshold, and defined as a pair-wise relationship. These threshold values were empirically chosen for parental and resistant conditions individually, based on the corresponding distributions of all pairwise absolute PCC values. Note PCC values resulting from probes mapped to the same gene were trivially ignored.

Bayesian statistical modeling of *GGR* network

Network model

For statistical modeling of networks, exponential families of distributions offer robust and flexible parametric models [24]. These probabilistic models can be used to

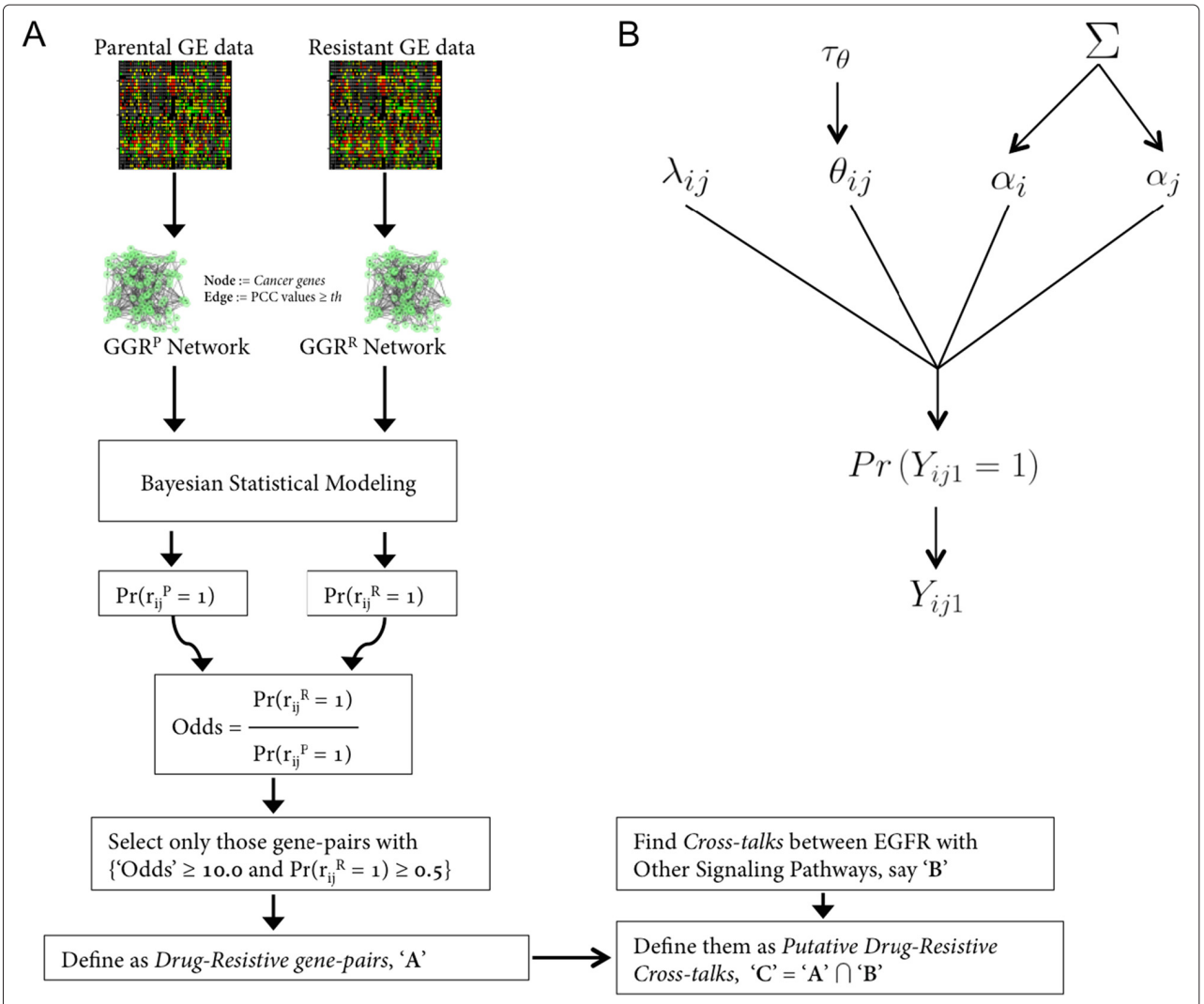


Figure 1 Schematic diagram of our proposed framework. (A) The framework for finding putative drug-resistant cross-talks. At first two gene expression data matrices were generated individually from the samples of both parental and resistant conditions. Next, based on pair-wise correlations of genes' expression values, two gene-gene relationship networks were derived. Then, a Bayesian statistical model called the p_1-model was applied on those two networks to find posterior probabilities of network edges. These posterior probabilities were used to find gene-pairs potentially contributing to drug resistance. Next, these gene-pairs were analyzed for overlap with cross-talks between EGFR/ErbB and other signaling pathways, and thus putative drug-resistant cross-talks were identified. **(B)** Hierarchical Bayesian model for inferring posterior probabilities of network parameters. Here, α represents the propensity (expansiveness/attractiveness) of a gene to be connected in an undirected network, and is dependent on the hyperparameter Σ; θ is the global density parameter; $\lambda_{ij} = log\,(n_{ij})$ is the scaling parameter, which is fixed due to the constraint $\sum_k Y_{ijk} = 1$; the hyperparameter τ_θ represents precision of the normal prior for the parameter θ.

evaluate the probability that an edge is present in the network. They can also be used to quantify topological properties of networks by summarizing them in a parametric form and associating sufficient statistics with those parameters [19,24]. In this study, we use a special class of exponential family distributions known as ERGM (Exponential Random Graph Models), also known as the p_1-model, which was introduced by Holland and Leinhardt [24].

A gene-gene relationship network with g genes can be regarded as a random variable $\mathbf{X}$ taking values from a set $\mathbf{G}$ containing all $2^{g(g-1)}$ possible relationship networks [24,25]. Let $\mathbf{u}$ be a generic point of $\mathbf{G}$ which can alternatively be denoted as the realization of $\mathbf{X}$ by $\mathbf{X} = \mathbf{u}$. Let the binary outcome $u_{ij} = 1$ if $gene_i$ interacts with $gene_j$, or $u_{ij} = 0$ otherwise. Then $\mathbf{u}$ is a binary data matrix [19]. Let $Pr(\mathbf{u})$ be the probability function on G given by

$$Pr(u) = Pr(\mathbf{X} = \mathbf{u}) = \frac{1}{\kappa\,(\boldsymbol{\theta})} \exp \sum_p \boldsymbol{\theta}_p z_p\,(\mathbf{u}) \qquad (1)$$

where $z_p(\mathbf{u})$ is the network statistic of type p, $\boldsymbol{\theta}_p$ is the parameter associated with $z_p(\mathbf{u})$ and $\kappa(\boldsymbol{\theta})$ is the normalizing constant that ensures $Pr(\mathbf{u})$ is a proper probability distribution (sums to 1 over all $\mathbf{u}$ in G) [26]. The parameter $\boldsymbol{\theta}$ is a vector of model parameters associated with network statistics and needs to be estimated. See [24] for further details.

A major limitation of the p_1-model is the difficulty of calculating the normalizing constant, $\kappa(\boldsymbol{\theta})$, since it is a sum over the entire graph space. Estimating the maximum likelihood of this model becomes intractable as there are $2^{g(g-1)}$ possible directed graphs (or $2^{\frac{g(g-1)}{2}}$ undirected graphs), each having g nodes (genes). A technique called *maximum pseudolikelihood estimation* has been developed to address this problem [27]. This technique employs MCMC methods such as Gibbs or Metropolis-Hastings sampling algorithms [28].

The construction of the p_1-model for a directed network is described in an Appendix Additional file 1: Appendix I. For the gene-gene relationship network with undirected edges, the description of the p_1-model can be simplified by using only two Bernoulli variables Y_{ij0} and Y_{ij1} instead of four as follows:

$$Y_{ijk} = \begin{cases} 1 & if \quad u_{ij} = k, \\ 0 & otherwise \end{cases}$$

The simplified p_1-model can then be defined using the following two equations to predict the probability of an edge being present between $gene_i$ and $gene_j$:

$$\log\left\{Pr\left(Y_{ij1}=1\right)\right\} = \lambda_{ij} + \theta + \alpha_i + \alpha_j \tag{2}$$

$$\log\left\{Pr\left(Y_{ij0}=1\right)\right\} = \lambda_{ij} \tag{3}$$

for $i < j$. Note that λ_{ij} is chosen to ensure $Pr\left(Y_{ij0}=1\right) + Pr\left(Y_{ij1}=1\right) = 1$. In this formulation, the expansiveness and attractiveness parameters were reduced to a single parameter, α, which represents the propensity of a gene to be connected in an undirected network. Hence, the p_1-model seeks to find the probabilities of edge formation in a network considering its structural features explicitly.

Bayesian modeling

We used a fully Bayesian approach for modeling our gene-gene relationship network. Parameter estimation is a crucial step in statistical modeling, for which a classical approach is maximum likelihood estimation (MLE). However, unlike MLE, Bayesian techniques involve calculation of posterior probabilities of model parameters by training the model with given data. We assume that the data $\mathcal{D}$ follows the generative model $\mathcal{M}$, and assign a prior probability $P(\theta|\mathcal{M})$ to the parameter vector θ under the model $\mathcal{M}$. Then Bayes' rule for calculating posterior probability is as follows:

$$Pr(\theta|\mathcal{M},\mathcal{D}) = \frac{Pr(\mathcal{D}|\theta,\mathcal{M}) \times Pr(\theta|\mathcal{M})}{\mathcal{Z}} \tag{4}$$

where $Pr(\mathcal{D}|\theta,\mathcal{M})$ is the likelihood function. Now, the marginal likelihood $\mathcal{Z}$ can be expressed as

$$\mathcal{Z} = Pr(\mathcal{D}|\mathcal{M}) = \int Pr(\mathcal{D}|\mathcal{M},\theta) \times P(\theta|\mathcal{M})\, d\theta, \tag{5}$$

Computing the exact solution for the marginal likelihood $\mathcal{Z}$ is often intractable since it is prone to the curse of dimensionality. Fortunately, Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling and Metropolis-Hastings methods do not require $\mathcal{Z}$ to be explicitly computed. In general, MCMC methods are stochastic simulation techniques which generate samples from the joint distribution $P(\mathcal{M},\theta|\mathcal{D})$ for calculating the posterior probabilities of parameters. Here we used Gibbs sampling methods, which sample iteratively, one parameter at a time, from the full conditional distribution given the current and previous values of all other parameters. To implement Gibbs sampling, we employed WinBUGS [29], which is a high-level software package providing an easy interface for implementing complex Bayesian models. In WinBUGS, users are free from background lower-level programming details, and only have to express the model precisely.

We hypothesized that gene-pairs involved in drug resistance are likely to be found with high probabilities in the resistant network but low probabilities in the parental network. Therefore, we built two networks, one from resistant datasets and the other from parental datasets. In this Bayesian approach, the model likelihood is defined in Equations (2) and (3), where Y_k is the data matrix calculated from the observed data $\mathbf{u}$. Here we have two Y_k data matrices, namely a gene-gene relationship network ${Y_k}^R$ derived from resistant samples and ${Y_k}^P$ derived from parental samples.

Our approach is a hierarchical Bayesian model in that model parameters are in turn dependent on *hyperparameters*. We assign the density parameter θ in Equation (2) a normal prior distribution with mean 0 and standard deviation σ_θ.

$$\theta \sim \mathcal{N}\left(0, {\sigma_\theta}^2\right) \tag{6}$$

Note, in WinBUGS the parameter τ, called the *precision*, replaces the standard deviation parameter σ of the normal distribution, where, $\tau = \sigma^{-2}$. For the hyperparameter τ_θ we specify a gamma prior distribution as follows, since it is a conjugate prior for the normal distribution:

$$\tau_\theta \sim Gamma(a_0, b_0) \tag{7}$$

We set $a_0 = 0.001$ and $b_0 = 0.001$ to make the prior for θ *noninformative*, making its standard deviation wide to express large uncertainty [19]. For attractiveness/

expansiveness parameters α_i and α_j, we followed the approach used by Adams *et al.* [30].

$$\begin{pmatrix} \alpha_i^R \\ \alpha_i^P \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \tag{8}$$

$$\Sigma^{-1} \sim Wishart\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2\right) \tag{9}$$

Here, α_i^R and α_i^P represent the expansiveness/attractiveness parameters for the network model of resistant and parental conditions, respectively.

Drug resistant cross-talk prediction

Since, Lapatinib is an EGFR and ErbB inhibitor, we considered the cross-talks between the EGFR/ErbB signaling pathway and other signaling pathways. Here cross-talks can be defined as any gene-pair $(gene_i, gene_j)$ in which $gene_i \in$ {genes in EGFR/ErbB signaling pathway} and $gene_j \in$ {genes in other signaling pathways}, or vice versa [31]. Thus if both genes in any gene-pair were found in the same signaling pathway, that particular gene-pair was trivially ignored. For that purpose, we collected 24 signaling pathways from Reactome [32] (downloaded at 19/05/2014), 35 signaling pathways from KEGG [33,34] (downloaded at 21/10/2014), and 63 signaling pathways from WikiPathway [35] (downloaded at 16/10/2014) databases. Each signaling pathway downloaded from these databases was encoded as tab-delimitated lists of gene symbols.

To determine whether a given gene-pair is involved in drug resistance, we calculated a simple odds ratio of the corresponding two posterior probabilities:

$$odds = \frac{Pr\left(Y_{ij1}^R = 1\right)}{Pr\left(Y_{ij1}^P = 1\right)} \tag{10}$$

where, Y_{ij1}^R and Y_{ij1}^P are gene-gene relationships defined over resistant and parental networks, respectively, and the probabilities are estimated using MCMC sampling. We then selected only those gene-pairs for which the odds score and $Pr\left(u_{ij}^R = 1\right)$ are greater than conservative thresholds, and identified these as the gene-pairs which are potentially involved in drug-resistance.

Results

Developing the network

For building gene-gene relationship networks, we considered the genes (nodes) from the Cancer Gene Census [23] only, since our aim was to find those gene-gene relationships which could be potential cross-talks among cancer signaling pathways. In order to identify such gene-pairs, we applied thresholds on their absolute Pearson Correlation Coefficient (PCC) values. These thresholds were 0.545 and 0.54 for parental and resistant conditions, respectively, which we selected from the corresponding distributions of all-pair absolute PCC values with the purpose of considering approximately the top 20% gene-pairs as pairwise relationships only. Applying these thresholds to the relationship values, 27,865 and 26,865 pair-wise relationships were identified in parental and resistant data matrices, respectively.

Bayesian analysis

For the two gene-gene relationship networks ${Y_k}^R$ and ${Y_k}^P$, Bayesian inference of the parameters of the p_1-model for an undirected network was applied. We used WinBUGS for scripting this inference and our scripts were inspired by Adams *et al.* [30]. We used 6000 MCMC iterations for parameter estimation with the first 5000 as 'burn-in'. All the parameters in the p_1-model appeared to converge rapidly during the burn-in iterations (data not shown). With the above settings, we estimated the posterior probabilities of each edge (gene-gene relationship) $Pr\left(Y_{ij1} = 1\right)$ in the two networks ${Y_k}^R$ and ${Y_k}^P$. For each edge, the proportion of the 1000 sampled networks containing the edge was considered as the posterior probability of that edge being present in the network.

Next, for each edge we calculated the odds ratio of their posterior probabilities as defined above. The rationale behind this calculation was that the edges (gene-pairs) found with high probabilities in resistant conditions but lower probabilities in parental conditions are more likely to be due to acquired resistance in cell lines. Therefore, we chose only gene-pairs with high odds ratio ($\geq$ 10.0) and high posterior probabilities ($\geq$ 0.5) of occurring in resistant conditions. We found 11,515 such gene-pairs (Additional file 2: Table S1) among all 68,265 [$= (370 \times 369) \div 2$] possibilities.

We then observed whether the above gene-pairs overlap with the list of potential cross-talks between EGFR/ErbB signaling and other signaling pathways. Here, we collected 24 signaling pathways from Reactome [32], 35 signaling pathways from KEGG [33,34], and 63 signaling pathways from WikiPathway [35] databases, and respectively identified 1,083 (841 distinct), 2,179 (1,050 distinct) and 3,084 (876 distinct) gene-pairs (Additional file 3: Table S2, Additional file 4: Table S3 and Additional file 5: Table S4) between EGFR/ErbB and other signaling pathways (see Materials and method). Of the 11,515 gene-pairs identified above, we found 104 (97 distinct), 188 (99 distinct) and 299 (96 distinct) gene-pairs overlap with the potential EGFR cross-talks identified using Reactome, KEGG and WikiPathway, respectively. Note the number of potential cross-talks and the number of distinct gene-pairs are different because the same gene-pair can form cross-talks between multiple pathway-pairs (pathways are overlapping). We consider these overlapping gene-pairs as putative drug-resistant cross-talks between EGFR/ErbB and

other signaling pathways. In these 104, 188 and 299 cross-talks, we found candidate EGFR/ErbB cross-talks with 13, 26 and 51 other signaling pathways, respectively. Moreover, among all 104, 188 and 299 cross-talks from Reactome, KEGG and WikiPathway, respectively, we found 32 distinct gene-pairs in at least two of these sets. Primary findings and detailed descriptions of all these putative cross-talks from the analyses of all three pathway sources are listed in Table 1, and Additional file 6: Table S5, Additional file 7: Table S6 and Additional file 8: Table S7, respectively. The network views of all these cross-talk sets from the analyses of individual pathway sources are shown in Figure 2.

Netwalker analyses

We conducted further analyses using Netwalker, a network analysis suite for functional genomics [36]. In this analysis, we observed the changes in GE values for each gene in the identified list of potential cross-talks. This was to verify our expectation that, since lapatinib is an EGFR/ErbB inhibitor, both genes involved in drug-resistant cross-talks should be up-regulated in resistant conditions compared to parental conditions, which may imply that the activation of other compensatory signaling pathways in resistant conditions can play a role in acquired resistance to inhibitors by activating the targeted pathway(s) [1,17]. Therefore, for all 67 genes involved in the above sets of 104, 188 and 299 drug-resistant cross-talks from Reactome, KEGG and WikiPathway, respectively, we made a heatmap image of GE values from both conditions (parental and resistant) (Figure 3A). For both resistant and parental conditions, we first averaged the gene expression values from the three samples corresponding to the three treatment conditions. Then these averaged gene expression values were transformed into z-scores (zero mean, unit standard deviation) and each z-score was normalized with the maximum of the absolute values of the z-scores across that particular gene. We observed that in 28 of these 67 genes (involved in 168 cross-talks), gene expression in one or more resistant conditions (0, 0.1 μM and 1 μM of lapatinib) was up-regulated relative to all the parental conditions (0, 0.1 μM and 1 μM of lapatinib) (Figure 3B) which may signify the insensitivity of these genes to inhibitors under resistant conditions. Note for Figure 3B only those genes are depicted for which both genes in some identified cross-talk had average GE values at resistant conditions greater than the average GE values at parental conditions.

For these 28 selected genes (168 cross-talks), we observed the relative changes in GE values (parental vs resistant conditions) in their candidate signaling pathways. First we analyzed EGFR signaling pathway from Reactome and found that many of the constituent genes were up-regulated in one (or more) resistant conditions whereas in all of their corresponding parental conditions they were down-regulated (Additional file 1: Figure S1). These 168 selected cross-talks associated EGFR (or ErbB) signaling pathways with 6 other signaling pathways that were found in at least two different pathway analyses (i.e. Reactome and KEGG, or KEGG and WikiPathway, or Reactome and WikiPathway). In those 6 other signaling pathways, we also observed a similar phenomenon as above (Additional file 1: Figure S1). These 6 signaling pathways are Notch signaling (in Reactome, KEGG and WikiPathway), Wnt signaling (in Reactome, KEGG and WikiPathway), insulin receptor/IGF1R signaling (in Reactome and WikiPathway), GPCR signaling (in Reactome and WikiPathway), hedgehog (in KEGG and WikiPathway), and TGF-β receptor signaling (in Reactome and WikiPathway). Again, for many of the constituent genes of these 6 signaling pathways, expression was up-regulated in at least one of the resistant conditions whereas in all the corresponding parental conditions they were down-regulated. Primary findings regarding these 168 selected drug-resistant cross-talks are listed in Additional file 9: Table S8, and the top 50 of those 168 cross-talks (based on sorted Odds ratio) are shown in Table 2.

Signaling cross-talk between EGFR/ErbB and other signaling pathways

Cross-talk between EGFR/ErbB and Notch signaling

We investigated literature evidence regarding the putative cross-talks between EGFR/ErbB signaling and other signaling pathways. We found *AKT2*:*MAML2* (in Reactome

Table 1 Primary findings from the analyses using signaling pathways from Reactome, KEGG and WikiPathway in breast cancer cell-line: SKBR3 (GSE38376)

Pathway source	# of signaling pathways	Pathway of interest	All Cross-talks of interest	Distinct gene-pairs[§]	All putative drug-resistant cross-talks	Distinct gene-pairs[¶]	# of other signaling pathways
REACTOME	23	EGFR	1,083	841	104	97	13
KEGG	35	ErbB	2,179	1,050	188	99	26
WikiPathway	63	ErbB	3,084	876	299	96	51

[¶]Number of distinct gene-pairs involved in all EGFR/ErbB cross-talks with all other signaling pathways; [§]Number of distinct gene-pairs commonly involved in all EGFR/ErbB cross-talks and drug resistance.

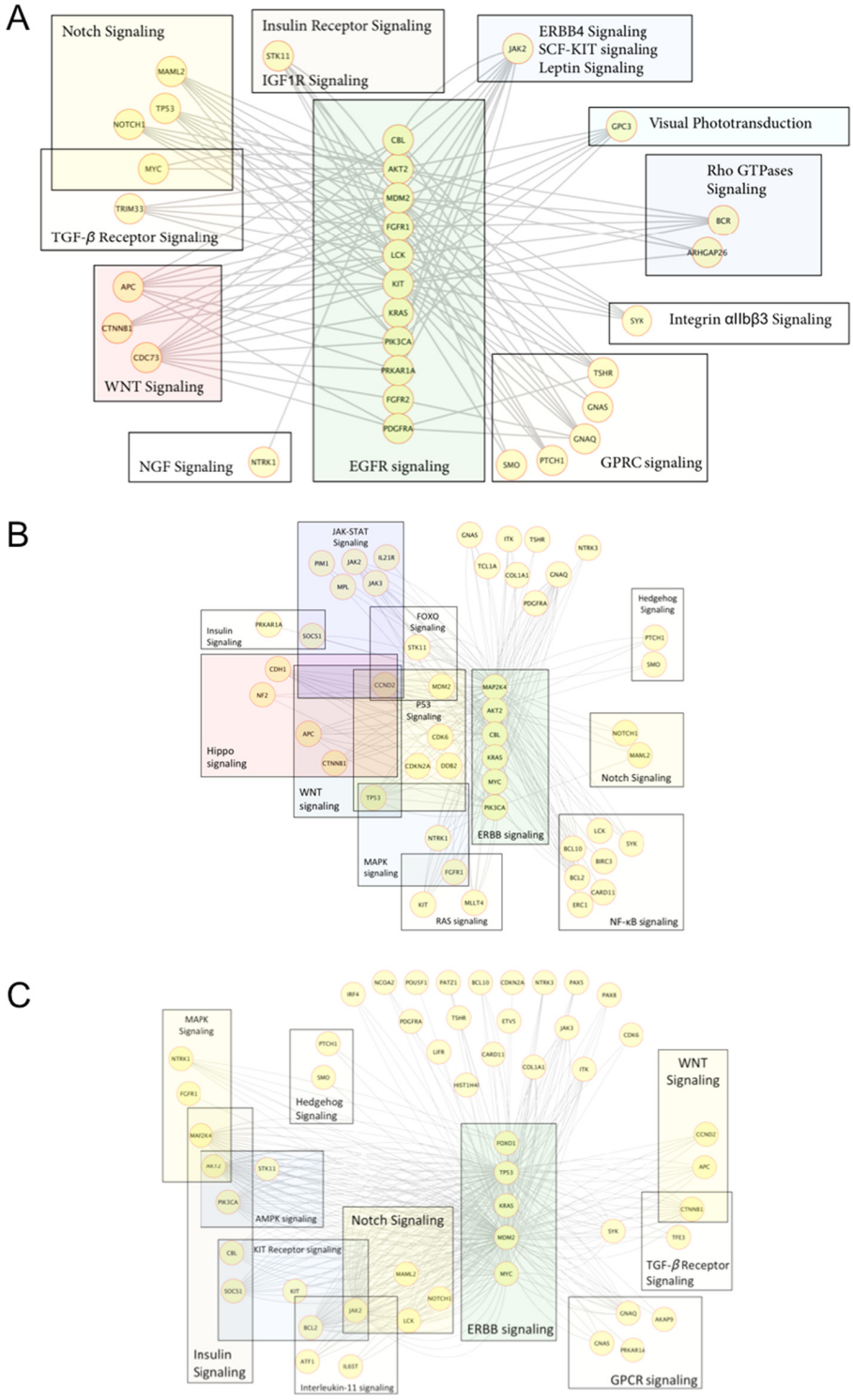


Figure 2 Network view of (A) 104, (B) 188, and (C) 299 putative drug-resistant cross-talks between pathways using Reactome, KEGG, and WikiPathway pathway databases in Breast Cancer Cell-line: SKBR3 (GSE38376). Nodes are genes, and the edges are the cross-talks. Note, all the cross-talks here possess posterior probabilities of appearing in resistant network ≥ 0.5 and Odds Ratio ≥ 10.0, which means the posterior probabilities of that cross-talk for appearing in parental network is ≤ 0.05.

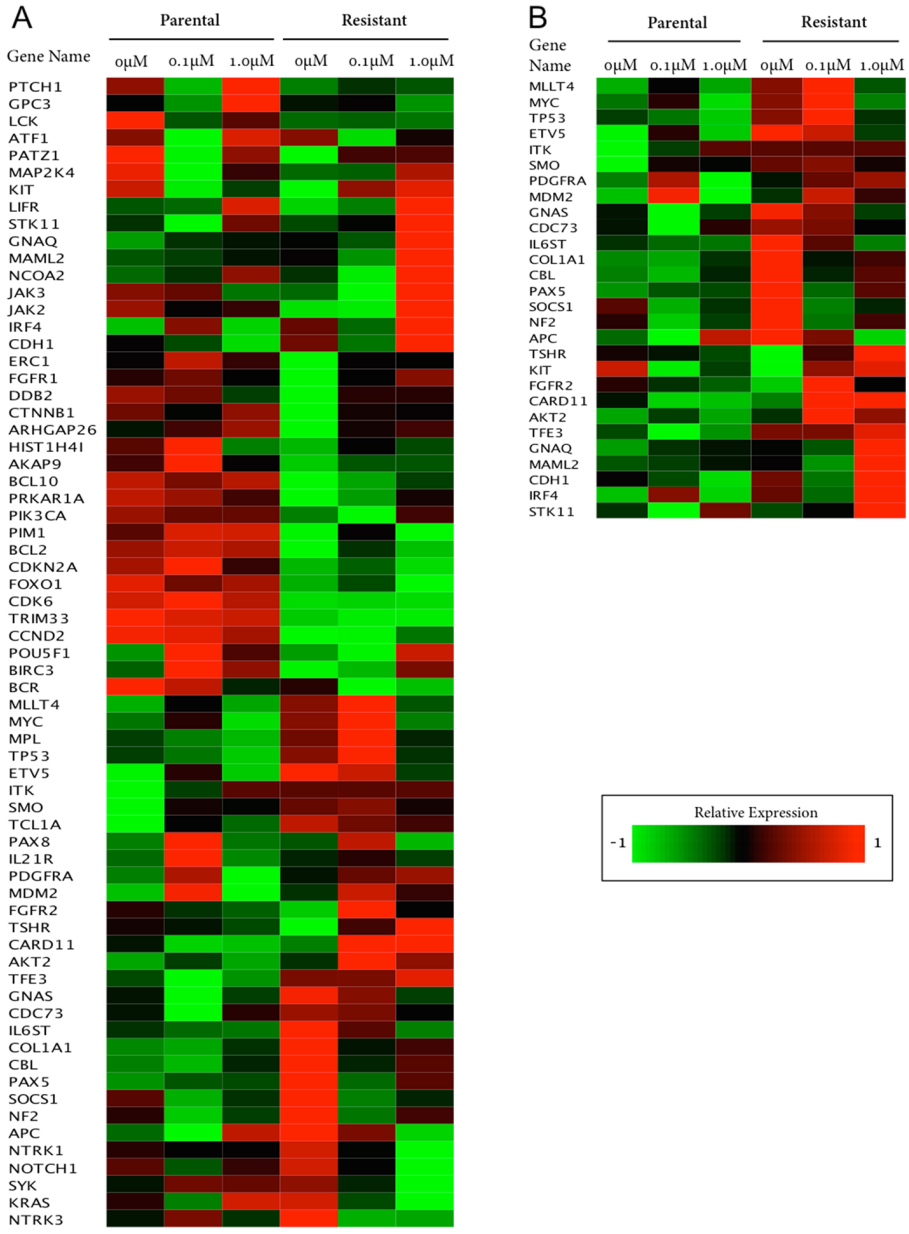


Figure 3 Heatmap of genes in putative drug-resistant cross-talks in breast cancer cell-line: SKBR3 (GSE38376). Heatmap image of comparative gene expression changes of parental and resistant conditions in **(A)** all 67 genes in all 104, 188 and 299 putative drug-resistant cross-talks using signaling pathways from Reactome, KEGG and WikiPathway database, respectively, and **(B)** 28 selected genes based on their differential regulation. Here, for each gene, the expression value at each of the 6 conditions (3 parental conditions, and 3 resistant conditions) is the average value of 3 sample patients [17]. For each gene, these 6 expression values (each of them is the average of 3 samples) were transformed into z-scores (zero mean, unit standard deviation) and each z-score was normalized with the maximum absolute value of the z-scores across that particular gene. Note, **(B)** includes only those genes which belonged to gene-pairs for which the average of GE values at resistant conditions was greater than the average of GE values at parental conditions. For both **(A)** and **(B)**, red and green bars indicate up-regulation and down-regulation, respectively.

and KEGG), *AKT2*:*TP53* (in Reactome), *AKT2*:*MYC* (in Reactome), *KIT*:*MAML2* (in Reactome), *KIT*:*TP53* (in Reactome), *MDM2*:*MAML2* (in Reactome and WikiPathway), *MDM2*:*TP53* (in Reactome), and *TP53*:*MAML2* (in WikiPathway) gene-pairs as putative cross-talks between EGFR/ErbB signaling and Notch signaling pathways. Up-regulation of the Notch signaling pathway inhibits apoptosis and thus contributes to breast carcinogenesis [37]. The Notch signaling pathway cross-talks with EGFR/ErbB signaling at the mediator level [1], e.g. when activated,

Notch1 contributes to cell growth and survival via Akt-activation in melanoma [38]. The Notch1 co-activator complex binds to the HES1 promoter [39] which encodes a transcription repressor that represses the expression of PTEN, a PI3K/Akt pathway inhibitor [40] contributing to tyrosine kinase inhibitor (TKI) resistance. Furthermore, Notch1 stimulates *MYC* transcription [41] and this stimulation can lead to the down-regulation of *MYC* via the Akt-pathway [42,43]. This putative gene-pair, *AKT2*:*MYC* was also found in our results as a potential drug-resistant cross-talk between the EGFR/ErbB and TGF-β receptor signaling pathways. Again, in HER2/*neu*-mediated resistance to DNA-damaging agents, the Akt pathway becomes activated which eventually suppresses p53 functions via enhancing MDM2-mediated ubiquitination [44]. Protein-protein interaction between MDM2 and p53 is evident as contributing to various cancer related activities [45,46].

Cross-talk between EGFR/ErbB and Wnt signaling

We found *MDM2*:*APC* (in Reactome and WikiPathway), *KIT*:*CDC73* (in Reactome), *MDM2*:*CDC73* (in Reactome), *CBL*:*APC* (in Reactome and KEGG), *PDGFRA*:*APC* (in Reactome), and *CBL*:*CDC73* (in Reactome), *AKT2*:*APC* (in KEGG), *AKT2*:*TP53* (in KEGG), and *TP53*:*APC* (in WikiPathway) as putative drug-resistant cross-talks between EGFR/ErbB and Wnt signaling pathways. Deregulation of the Wnt/β-catenin signaling pathway plays a critical role in various cancers including breast, colorectal, pancreatic and colon cancer [47,48], and its association with drug-resistance has been studied by several research groups [47-50]. Recently, it has been reported that resistant cell lines exhibited increased Wnt signaling in both breast and colon cancer [49,50]. Loh *et al.* showed that genes in the Wnt signaling pathway, in both the β-catenin dependent (*AXIN2*, *MYC*, *CSNK1A1*) and the independent arms (*ROR2*, *JUN*), were up-regulated in cell lines resistant to tamoxifen compared to the parental MCF7 cell line [49]. Furthermore, *ROR1*, a constituent gene of Wnt signaling pathway, plays a sustainer role in EGFR-mediated prosurvival signaling in lung adenocarcinoma via signaling cross-talk and was therefore reported to be a potential therapeutic target [51]. *APC* and *MDM2* in the *MDM2*:*APC* cross-talk are both tumor suppressors; they co-regulate DNA polymerase-β [52,53] which is reported to be hyper-activated in a cis-diamminedichloroplatinum(II) resistant P388 murine leukemia cell line [54]. Again, β-catenin whose stability is negatively regulated by *APC* [55], confers resistance to PI3K/Akt inhibitors in colon cancer [56].

Cross-talk between EGFR/ErbB and GPCR signaling

Between EGFR/ErbB and GPCR signaling pathways, we found *KIT*:*GNAQ* (in Reactome), *MDM2*:*GNAQ* (in Reactome and WikiPathway), *CBL*:*GNAQ* (in Reactome), *FGFR2*:*GNAQ* (in Reactome), *PDGFRA*:*GNAQ* (in Reactome), *KIT*:*TSHR* (in Reactome), *MDM2*:*TSHR* (in Reactome), *CBL*:*TSHR* (in Reactome), *PDGFRA*:*TSHR* (in Reactome), *KIT*:*GNAS* (in Reactome), *MDM2*:*GNAS* (in Reactome and WikiPathway), *KIT*:*SMO* (in Reactome), *MDM2*:*SMO* (in Reactome), *TP53*:*GNAQ* (in WikiPathway), and *MYC*:*GNAQ* (in WikiPathway). GPCR-like signaling contributes to acquired drug resistance after being mediated by Smoothened (*SMO*) via activating Gli, a canonical hedgehog (Hh) transcription factor [57]. GPCR and EGFR/ErbB over-expression often contributes to cancer growth. Cross-talk between the two at the receptor level contributes to HNSCC (head and neck squamous cell carcinoma) via triggering EGFR/ErbB signaling by a GPCR ligand [58]. For the *MDM2*:*SMO* cross-talk, found between the EGFR/ErbB and GPCR signaling pathways, a *SMO*-mutant from Hh signal transducer activates PI3K/Akt/Gli pathway that eventually increases MDM2 phosphorylation [59]. This in turn increases MDM2-mediated p53 degradation and thus reduces p53-induced apoptosis [59]. Furthermore, recently it has been reported that *SMO* (Hh signal transducer) functions like a G-protein coupled receptor due to its structural resemblance to GPCRs [60,61] which may be further evidence for a drug-resistant cross-talk between hedgehog signaling and EGFR/ErbB signaling [1].

Cross-talk between EGFR/ErbB and IR (insulin receptor)/IGF1R signaling

Several studies have reported extensive cross-talk between IR (insulin receptor)/IGF1R (insulin-like growth factor-1 receptor) and EGFR/ErbB signaling pathways contributing to acquired drug resistance in various cancers [62-64]. Loduvini *et al.* reported significant correlation between worse disease-free survival and high co-expression of both EGFR/ErbB and IGF1R in NSCLC (non-small-cell lung cancer) patients [65]. EGFR/ErbB can physically interact with other non-ErbB family receptors at the cell surface and can form heterodimers with receptors like IGF1R, PDGFR etc. [62]. Moreover, the EGFR/ErbB and IGF1R pathways can also cross-talk indirectly via physical interactions between their downstream shared-components [62]. It has been reported recently that gefitinib (an EGFR TKI) inhibits the phosphorylation of IRS1 by IR, but also triggers the association between IRS1 and IGF1R which in turn induces drug-resistance [66]. Knowlden *et al.* showed the cross-talk between IGF1R and EGFR signaling pathways occurred in tamoxifen-resistant MCF7 and T47D breast cancer cell-lines but not in non-resistant cells [18]. Our findings suggest *KIT*:*STK11* (in Reactome), *MDM2*:*STK11* (in Reactome), *MDM2*:*AKT2* (in WikiPathway), *MYC*:*AKT2* (in WikiPathway), *TP53*:*AKT2* (in WikiPathway),

MDM2:*CBL* (in WikiPathway), *MDM2*:*SOCS1* (in WikiPathway), and *TP53*:*SOCS1* (in WikiPathway) as putative drug-resistant cross-talks between the IGF1R/IR and EGFR/ErbB signaling pathways. For the *MDM2* and *STK11* (also known as *LKB1*) genes, which we identified as a putative cross-talk between EGFR and IGF1R signaling, we did not find any direct supporting evidence in the literature. However, this association is plausible in the resistant conditions given that Yamaguchi *et al.* suggested EGFR signaling may cross-talk with the AMPK/LKB signaling pathway [1]. Moreover, Levine *et al.* reported interconnections between p53 and IGF1R/AKT/mTOR pathways where both *LKB1* and *MDM2* participate in a series of pathway cross-talks [67].

Validation of the framework using BT474 cell-line (GSE16179)

To further illustrate our method, we analysed a second dataset (GSE16179) containing gene expression profiles of breast cancer cell-line BT474 under two conditions (parental and lapatinib resistant) [16]. The reason for choosing this dataset was that it was obtained using a similar experimental design to the primary dataset GSE38376, but with an additional treatment condition using foretinib (GSK1363089) only and with combined drug use (lapatinib + foretinib). There were three samples per treatment condition. However, to adapt simply and be coherent with the previous experiment, we only considered expression values of parental conditions (3 samples with basal condition: GSM799168, GSM799169 and GSM799170; 3 samples with 1 μM of lapatinib treatment: GSM79917, GSM799172 and GSM799173), and the same conditions with lapatinib resistant cells (3 samples with basal condition: GSM799174, GSM799175 and GSM799176; 3 samples with 1 μM of lapatinib treatment: GSM799180, GSM799181 and GSM799182). Among the 375 cancer genes from Cancer Gene Census [23], there were 357 genes which had gene expression values. We identified 27,358 and 26,292 pair-wise gene-gene relationships (undirected edges) in resistant and parental networks by applying the thresholds 0.71 and 0.81, respectively. Bayesian inference of the p_1-model parameters for an undirected network was applied to these two gene-gene relationship networks as before. Thereafter, among all 63,546 [= $(357\times356)\div2$] possibilities, we found 10,811 gene-pairs (Additional file 10: Table S9) with the same thresholds of odds ratio ($\geq$10.0) as previously, but smaller posterior probability ($\geq$0.15) of occurring in the resistant network. With this set of putative drug-resistant gene-pairs, we also observed the overlap of potential cross-talks of EGFR/ErbB with other signaling pathways using Reactome, KEGG and WikiPathway databases. We found 83 (72 distinct), 133 (87 distinct) and 277 (81 distinct) cross-talks between EGFR/ErbB and other signaling pathways from Reactome, KEGG and WikiPathway (Additional file 11: Table S10, Additional file 12: Table S11 and Additional file 13: Table S12), respectively. The numbers of signaling pathways that were involved in those EGFR/ErbB cross-talks were 10, 18 and 54, respectively. Among the 83, 133 and 277 cross-talks, we found 50 distinct gene-pairs in at least two of these sets. Table 3 shows the comparative findings between our primary dataset (SKBR3 cell-line, GSE38379) and our secondary dataset (BT474 cell-line, GSE16179). In Table 3, we show that some important signaling pathways that were involved in the EGFR/ErbB cross-talks (i.e. Notch, WNT, GPCR, IR/IGF1R, TGF-β signaling pathways) in our primary dataset, have some overlap with our secondary dataset.

There were 78 genes involved in these sets of 83, 133 and 277 putative cross-talks. We performed a similar Netwalker analyses with these 78 genes as we did for the dataset GSE38376, and found 37 genes (involved in 86 cross-talks (Additional file 14: Table S13)) consistent with our hypothesis that both genes in a particular cross-talk should be up-regulated in resistant conditions but down-regulated in parental conditions. In Figure 4, the selected genes from the secondary dataset exhibit an even clearer pattern of up-regulation in resistant conditions than the selected genes from our primary dataset.

Discussion

In this study, we developed a computational framework to systematically predict signaling cross-talks between EGFR/ErbB and other signaling pathways that contribute to lapatinib (an EGFR and ErbB2/HER2 inhibitor) resistance. We hypothesized that gene-pairs (cross-talks) that can potentially cause drug-resistance have a high probability of occurring in the resistant condition(s) but a low probability in parental conditions. We employed a fully Bayesian statistical model: a special class of Exponential Random Graph Model known as the p_1-model, to infer the posterior probabilities of such gene-pairs from corresponding networks inferred using gene expression values [17] of resistant and parental conditions. In selecting gene-pairs as putative cross-talks, threshold values for two parameters: odds and posterior probabilities of edges in resistant networks were empirically selected. However, more robust procedures for the selection of these two parameters can be made in future studies. All other parameters in the p_1-model discussed above were estimated using Gibbs sampling (see Materials and method).

Our results primarily focus on compensatory signaling pathways i.e. Notch signaling, Wnt signaling, GPCR signaling, and IR/IGF1R signaling, which cross-talk with EGFR/ErbB signaling to reduce the inhibiting effect of lapatinib. We present additional literature evidence that the identified cross-talks of the above compensatory signaling pathways with EGFR/ErbB signaling may contribute

Table 2 Description of top 50 (based on sorted Odds ratio) cross-talks among all 168 potential drug-resistant cross-talks between EGFR/ErbB signaling and other pathways from all the analyses using Reactome, KEGG and WikiPathway databases in GSE38376

$gene_i$::$gene_j$	EGFR/ErbB :: Signaling pathway$_j$	$Pr\left(Y_{ij}^R = 1\right)$	$Pr\left(Y_{ij}^P = 1\right)$	Odds ratio	$Avg(GE_i^P)$: $Avg\left(GE_j^P\right)$	$Avg(GE_i^R)$: $Avg\left(GE_j^R\right)$
AKT2::MAML2§,¶	Notch signaling	0.5	0.03	16.67	87.71::76.59	96.84::78.6
MDM2::APC§,$	Wnt signaling	0.5	0.03	16.67	76.33::82.43	77.9::86.76
KIT::CDC73§	Wnt signaling	0.5	0.03	16.67	82.14::104.01	82.68::110.88
MDM2::CDC73§	Wnt signaling	0.5	0.03	16.67	76.33::104.01	77.9::110.88
KIT::GNAQ§	GPCR signaling	0.5	0.03	16.67	82.14::130	82.68::139.33
MDM2::GNAQ§,$	GPCR signaling	0.5	0.03	16.67	76.33::130	77.9::139.33
KIT::TSHR§	GPCR signaling	0.5	0.03	16.67	82.14::71.32	82.68::71.66
MDM2::TSHR§	GPCR signaling	0.5	0.03	16.67	76.33::71.32	77.9::71.66
AKT2::APC¶	Wnt signaling	0.5	0.03	16.67	87.71::82.43	96.84::86.76
AKT2::APC¶	Hippo signaling	0.5	0.03	16.67	87.71::82.43	96.84::86.76
AKT2::CDH1¶	Hippo signaling	0.5	0.03	16.67	87.71::74.2	96.84::79.8
AKT2::GNAQ¶	Gnrh signaling	0.5	0.03	16.67	87.71::130	96.84::139.33
AKT2::GNAQ¶	Calcium signaling	0.5	0.03	16.67	87.71::130	96.84::139.33
AKT2::MDM2¶	p53 signaling	0.5	0.03	16.67	87.71::76.33	96.84::77.9
MDM2::AKT2$	Regulation of toll-like receptor signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	insulin signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	RANKL/RANK signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	AMPK signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	MAPK signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	Tweak signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::AKT2$	Toll-like receptor signaling	0.5	0.03	16.67	76.33::87.71	77.9::96.84
MDM2::APC$	BDNF signaling	0.5	0.03	16.67	76.33::82.43	77.9::86.76
MDM2::APC$	Wnt signaling Netpath	0.5	0.03	16.67	76.33::82.43	77.9::86.76
MDM2::APC$	Wnt signaling and Pluripotency	0.5	0.03	16.67	76.33::82.43	77.9::86.76
MDM2::COL1A1$	Nanoparticle-mediated activation of receptor signaling	0.5	0.03	16.67	76.33::91.44	77.9::102.54
MDM2::COL1A1$	Osteoblast signaling	0.5	0.03	16.67	76.33::91.44	77.9::102.54
MDM2::GNAQ$	TSH signaling	0.5	0.03	16.67	76.33::130	77.9::139.33
MDM2::GNAQ$	Serotonin Receptor 2 and STAT3 signaling	0.5	0.03	16.67	76.33::130	77.9::139.33
MDM2::GNAQ$	Serotonin Receptor 2 and ELK-SRF/GATA4 signaling	0.5	0.03	16.67	76.33::130	77.9::139.33
MDM2::ITK$	T-Cell Receptor and Co-stimulatory signaling	0.5	0.03	16.67	76.33::89.86	77.9::93.27
MDM2::ITK$	Tcr signaling	0.5	0.03	16.67	76.33::89.86	77.9::93.27
MDM2::KIT$	Kit receptor signaling	0.5	0.03	16.67	76.33::82.14	77.9::82.68

Table 2 Description of top 50 (based on sorted Odds ratio) cross-talks among all 168 potential drug-resistant cross-talks between EGFR/ErbB signaling and other pathways from all the analyses using Reactome, KEGG and WikiPathway databases in GSE38376 *(Continued)*

MDM2::PAX5[$]	ID signaling	0.5	0.03	16.67	76.33::68.91	77.9::71.02
MDM2::TSHR[$]	TSH signaling	0.5	0.03	16.67	76.33::71.32	77.9::71.66
AKT2::TP53[§]	Notch signaling	0.5	0.04	12.5	87.71::128.73	96.84::155.09
KIT::APC[§]	Wnt signaling	0.5	0.04	12.5	82.14::82.43	82.68::86.76
KIT::MAML2[§]	Notch signaling	0.5	0.04	12.5	82.14::76.59	82.68::78.6
KIT::STK11[§]	IGF1R signaling	0.5	0.04	12.5	82.14::71.97	82.68::74.95
KIT::STK11[§]	insulin receptor signaling	0.5	0.04	12.5	82.14::71.97	82.68::74.95
KIT::TP53[§]	Notch signaling	0.5	0.04	12.5	82.14::128.73	82.68::155.09
MDM2::MAML2[§,$]	Notch signaling	0.5	0.04	12.5	76.33::76.59	77.9::78.6
MDM2::STK11[§]	IGF1R signaling	0.5	0.04	12.5	76.33::71.97	77.9::74.95
MDM2::STK11[§]	insulin receptor signaling	0.5	0.04	12.5	76.33::71.97	77.9::74.95
MDM2::TP53[§]	Notch signaling	0.5	0.04	12.5	76.33::128.73	77.9::155.09
AKT2::GNAS[¶]	Gnrh signaling	0.5	0.04	12.5	87.71::5465.46	96.84::6212.43
AKT2::GNAS[¶]	Calcium signaling	0.5	0.04	12.5	87.71::5465.46	96.84::6212.43
AKT2::NF2[¶]	Hippo signaling	0.5	0.04	12.5	87.71::85.75	96.84::87.36
AKT2::TP53[¶]	P53 signaling	0.5	0.04	12.5	87.71::128.73	96.84::155.09
AKT2::TP53[¶]	Wnt signaling	0.5	0.04	12.5	87.71::128.73	96.84::155.09
CBL::CDH1[¶]	RAP1 signaling	0.5	0.04	12.5	194.46::74.2	208.45::79.8

Cross-talks found using signaling pathways from [§]Reactome, [¶]KEGG, and [$]WikiPathway Databases; $Pathway_j$ is the pathway containing $gene_j$; $Pr\left(Y_{ij}^R = 1\right)$ and $Pr\left(Y_{ij}^P = 1\right)$ are the posterior probabilities of $gene_i$:$gene_j$ in Resistant and Parental networks, respectively; $Avg\left(GE_i^P\right)$ is the average GE value of all Parental conditions (each of which is an average of 3 samples) for $gene_i$, $Avg\left(GE_i^R\right)$ is similar but with Resistant conditions, and others are likewise similar.

to drug-resistance by maintaining key cell survival and/or proliferation signals in common down-stream pathways, including PI3K/Akt signaling [1].

Komurov *et al.* [17] hypothesized that cross-talks between EGFR/ErbB signaling and metabolic pathways contribute to resistance to lapatinib. More specifically, they identified that glucose deprivation reduces the inhibiting effects of lapatinib by up-regulating constituent genes and thus providing an EGFR/ErbB2-independent mechanism of glucose uptake and cell survival [17]. Here, by using the same gene expression datasets, we found *MDM2*:*STK11* cross-talk between EGFR/ErbB and IGF1R signaling, where STK11 (also known as *LKB1*) phosphorylates and activates *AMPK* in absence of glucose [67]. Again, in the integrated signaling circuitry of pathways: p53-IGF-1-AKT-TSC2-mTOR, a positive feedback loop (p53-PTEN AKT-MDM2-p53) is formed which enhances p53-mediated apoptosis and senses nutrient deprivation [67]. Thus our results complement the findings of Komurov *et al.* by finding *signaling* cross-talks between EGFR/ErbB and IGF1R pathways.

In Netwalker analysis of our primary dataset (SKBR3 cell-line, GSE38376), we compared the expression changes of all the samples in parental conditions (basal, 0.1 μM and 1.0 μM) with those of all the samples in resistant conditions (basal, 0.1 μM and 1.0 μM). However, we conducted another experiment on both of our primary (SKBR3 cell-line, GSE38376) and secondary datasets (BT474 cell-line, GSE16179) in which we first identified genes dysregulated in treatment vs basal conditions in parental samples and then checked if those genes were reversely changed in treatment conditions in resistant samples. To that end, for each sample, first we calculated the fold-change(s) of parental treatment condition(s) compared to parental basal condition, and then we calculated the fold-changes of resistant basal and resistant treatment conditions compared to parental basal condition (Additional file 1: Figure S2A and S3A). Then, we chose only those genes for which, in any of the 3 samples, expressions were dysregulated (up-/down-regulated) in (all the) parental treatment condition(s) (log_2 of fold-changes were positive/negative), and for that particular sample, expressions were reversely changed (the fold-change sign was opposite to that of parental condition) in all the resistant treatment conditions (Additional file 1: Figure S2B and S3B). This may be a strong indicator of sensitivity to an inhibitor in parental conditions and

Table 3 Comparative results between primary dataset (SKBR3 cell-line, GSE38376) and validation dataset (BT474 cell-line, GSE16179)

Pathway name	Found in Pathway source (GSE38376)	Found in Pathway source (GSE16179)	Common cross-talks in both Studies[¶]
Notch Signaling	Reactome, KEGG, WikiPathway	Reactome, KEGG, WikiPathway	*MAP2K4::NOTCH1*
GPCR signaling	Reactome, WikiPathway	Reactome, WikiPathway	***CBL::TSHR***
			FGFR1::TSHR
			PDGFRA::GNAQ
			KIT::TSHR
			LCK::TSHR
			MDM2::TSHR
			PDGFRA::TSHR
WNT Signaling	Reactome, KEGG, WikiPathway	Reactome, KEGG, WikiPathway	*AKT2::CCND2*
			MAP2K4::CCND2
			MAP2K4::TP53
			MDM2::MAP2K4
Insulin (IGF1R) Signaling	Reactome, WikiPathway	Reactome, WikiPathway	*MDM2::MAP2K4*
			TP53::MAP2K4
TGF-β Signaling	Reactome, WikiPathway	Reactome, KEGG, WikiPathway	*MDM2::TFE3*
			TP53::TFE3
MAPK signaling	KEGG, WikiPathway	KEGG, WikiPathway	MDM2::MAP2K4

[¶]These common cross-talks were found using the primary dataset (104, 188 and 299 cross-talks from Reactome, KEGG and WikiPathway databases, respectively) and validation datasets (83, 133 and 277 cross-talks from Reactome, KEGG and WikiPathway databases, respectively). Cross-talks mentioned with **Bold face** are those consistent with our hypothesis that both genes in the particular cross-talk are up-regulated in resistant conditions but down-regulated in parental conditions.

the development of acquired resistance. Next, we compared these selected genes to cross-talks found in results from GSE38379 (104, 188 and 299 EGFR/ErbB cross-talks from Reactome, KEGG and WikiPathway, respectively) and GSE16179 (83, 133 and 277 EGFR/ErbB cross-talks from Reactome, KEGG and WikiPathway, respectively). Although we didn't find any such cross-talks overlapping with the results from the primary dataset (GSE38379), we found 401 from our secondary dataset (GSE16179) (Additional file 15: Table S14).

Currently, our network modeling only considers undirected edges among genes. In future we would like to generalise the approach to identify directed and indirect interactions among genes. In network modeling, a combination of both direct and indirect relationships among gene-pairs was found to provide better insights into biological systems in our previous studies [68]. The rationale for combining these two types of gene-gene relationships in signaling networks is that EGFR/ErbB and IGF1R can both cross-talk (EGFR/IGF1R heterodimerization) directly at the receptor level, and indirectly mediated by GPCR signaling, as reported by Van der Veeken *et al.* [62]. Other high-throughput datasets such as miRNA expression data, copy number aberration data, and methylation data could also be incorporated into our framework to obtain a better understanding of gene dependencies. Note that our methodology exploits a fully data-driven approach for finding putative drug-resistant cross-talks, without incorporating other prior information regarding gene-gene relationships, such as Protein-Protein Interactions. Hence, although our data-driven approach may inherently yield some false-positive predictions, it may also provide the possibilities of finding novel cross-talks contributing to drug- resistance.

Conclusions

Our proposed computational framework is able to predict putative cross-talks among signaling pathways that

Figure 4 Heatmap of genes in putative drug-resistant cross-talks in breast cancer cell-line: BT474 (GSE16179). Heatmap image of comparative gene expression changes of parental and resistant conditions in **(A)** all 78 genes in all 83, 133 and 277 putative drug-resistant cross-talks using signaling pathways from Reactome, KEGG and WikiPathway database, respectively, and **(B)** 37 selected genes based on their differential regulation. Here, for each gene, the expression value at each of the 4 conditions (2 parental conditions, and 2 resistant conditions) is the average value of 3 sample patients [16]. For each gene, these 4 expression values (each of them is the average of 3 samples) were transformed into z-scores (zero mean, unit standard deviation) and each z-score was normalized with the maximum absolute value of the z-scores across that particular gene. For both **(A)** and **(B)**, red and green bars indicate up-regulation and down-regulation, respectively.

play a role in drug resistance in two breast cancer cell-lines, SKBR3 and BT474. Our framework could also be useful for other types of cancer to enhance understanding of the role of signaling cross-talks in drug resistance. Most importantly, we believe our method can be used to find a range of compensatory pathways that nullify/reduce the inhibiting effects of drugs via cross-talk with targeted pathways. These novel compensatory pathways can be further considered as novel targets for single or combination therapies.

Additional files

Additional file 1: Appendix I. Derivation of p_1-model for directed network.

Additional file 2: Table S1. All 11,515 drug-resistant gene-pairs found in GSE38376.

Additional file 3: Table S2. All 1,083 (841 distinct) cross-talks found between EGFR and other 23 signaling pathways from Reactome database.

Additional file 4: Table S3. All 2,179 (1,050 distinct) cross-talks found between ErbB and other 34 signaling pathways from KEGG database.

Additional file 5: Table S4. All 3,084 (876 distinct) cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database.

Additional file 6: Table S5. 104 drug-resistant cross-talks found between EGFR and other 23 signaling pathways from Reactome database [GSE38376].

Additional file 7: Table S6. 188 drug-resistant cross-talks found between ErbB and other 34 signaling pathways from KEGG database [GSE38376].

Additional file 8: Table S7. 299 drug-resistant cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database [GSE38376].

Additional file 9: Table S8. 168 selected cross-talks which associated EGFR (or ErbB) signaling pathways with 6 other signaling pathways that were found in at least two different pathway analyses (i.e. Reactome and KEGG, or KEGG and WikiPathway, or Reactome and WikiPathway) [GSE38376].

Additional file 10: Table S9. All 10,811 drug-resistant gene-pairs found in GSE16179.

Additional file 11: Table S10. 83 drug-resistant cross-talks found between EGFR and other 23 signaling pathways from Reactome database [GSE16179].

Additional file 12: Table S11. 133 drug-resistant cross-talks found between ErbB and other 34 signaling pathways from KEGG database [GSE16179].

Additional file 13: Table S12. 278 drug-resistant cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database [GSE16179].

Additional file 14: Table S13. 86 drug-resistant cross-talks found in all Reactome, KEGG and WikiPathway analyses where both genes in a particular cross-talk was up-regulated in resistant conditions but down-regulated in parental conditions [GSE16179].

Additional file 15: Table S14. 401 cross-talks from Reactome, KEGG and WikiPathway analyses where the genes are dysregulated in parental treatment vs parental basal condition, and reversely changed in resistant basal + resistant treatment vs parental basal condition [GSE16179].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AKMA conceived the idea, collected datasets, designed and conducted experiments, analysed data and results, wrote the manuscript; JMK checked and approved the statistical model; AKMA validated the results and AL approved that validation; JMK and AL supervised this work and provided guidance in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgement
This research was supported by Monash International Postgraduate Research Scholarship and Monash Graduate Scholarship at the Monash University, Australia.

Author details
[1]School of Mathematical Science, Monash University, Wellington Road, Clayton, VIC, Australia. [2]Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University, Wellington Road, Clayton, VIC, Australia.

Received: 15 July 2014 Accepted: 11 December 2014
Published online: 20 January 2015

References
1. Yamaguchi H, Chang SS, Hsu JL, Hung MC. Signaling cross-talk in the resistance to HER family receptor targeted therapy. Oncogene 2014;33(9):1073–81.
2. Logue JS, Morrison DK. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. Genes Dev 2012;26(7):641–50.
3. Jänne PA, Gray N, Settleman J. Factors underlying sensitivity of cancers to small-molecule kinase inhibitors. Nat Rev Drug Discov 2009;8(9):709–23.
4. Bauman PA, Dalton WS, Anderson JM, Cress AE. Expression of cytokeratin confers multiple drug resistance. Proc Nat Acad Sci USA 1994;91(12):5311–4.
5. Hazlehurst L, Dalton W. De Novo and acquired resistance to antitumor alkylating agents In: Teicher B, editor. Cancer Drug Resistance, Cancer Drug Discovery and Development. Humana Press; 2006. p. 377–89.
6. Zhang Z, Lee JC, Lin L, Olivas V, Au V, LaFramboise T, et al. Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. Nat Genet 2012;44(8):852–60.
7. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin Cancer Res 2013;19:279–90.
8. Takezawa K, Pirazzoli V, Arcila ME, Nebhan CA, Song X, de Stanchina E, et al. HER2 amplification: a potential mechanism of acquired resistance to EGFR inhibition in EGFR-mutant lung cancers that lack the second-site EGFRT790M mutation. Cancer Discov 2012;2(10):922–33.
9. Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science 2007;316(5827):1039–43.
10. Zhuang G, Brantley-Sieders DM, Vaught D, Yu J, Xie L, Wells S, et al. Elevation of receptor tyrosine kinase EphA2 mediates resistance to trastuzumab therapy. Cancer Res 2010;70:299–308.
11. Mendoza MC, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends Biochem Sci 2011;36(6):320–8.
12. Higa GM, Abraham J. Lapatinib in the treatment of breast cancer. Expert Rev Anticancer Ther 2007;7(9):1183–92.
13. Medina PJ, Goodin S. Lapatinib: a dual inhibitor of human epidermal growth factor receptor tyrosine kinases. Clin Ther 2008;30(8):1426–47.
14. Garrett JT, Olivares MG, Rinehart C, Granja-Ingram ND, Sanchez V, Chakrabarty A, et al. Transcriptional and posttranslational up-regulation of HER3 (ErbB3) compensates for inhibition of the HER2 tyrosine kinase. Proc Nat Acad Sci USA 2011;108(12):5021–6.
15. Azuma K, Tsurutani J, Sakai K, Kaneda H, Fujisaka Y, Takeda M, et al. Switching addictions between HER2 and FGFR2 in HER2-positive breast tumor cells: FGFR2 as a potential target for salvage after lapatinib failure. Biochem Biophys Res Commun 2011;407:219–24.
16. Liu L, Greger J, Shi H, Liu Y, Greshock J, Annan R, et al. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. Cancer Res 2009;69(17):6871–8.
17. Komurov K, Tseng JT, Muller M, Seviour EG, Moss TJ, Yang L, et al. The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant ErbB2-positive breast cancer cells. Mol Syst Biol 2012;8(1):.
18. Knowlden JM, Hutcheson IR, Barrow D, Gee JM, Nicholson RI. Insulin-like growth factor-I receptor signaling in tamoxifen-resistant breast cancer: a supporting role to the epidermal growth factor receptor. Endocrinology 2005;146(11):4609–18.
19. Bulashevska S, Bulashevska A, Eils R. Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. BMC Bioinformatics 2010;11:46.
20. Hill SM, Lu Y, Molina J, Heiser LM, Spellman PT, Speed TP, et al. Bayesian inference of signaling network topology in a cancer cell line. Bioinformatics 2012;28(21):2804–10.
21. Lee MJ, Ye AS, Gardino AK, Heijink AM, Sorger PK, MacBeath G, et al. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. Cell 2012;149(4):780–94.
22. Pawson T, Warner N. Oncogenic re-wiring of cellular signaling pathways. Oncogene 2007;26(9):1268–75.

23. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4(3):177–83.
24. Holland PW, Leinhardt S. An exponential family of probability distributions for directed graphs. J Am Stat Assoc 1981;76(373):33–50.
25. Katz L, Powell J. A proposed index of the conformity of one sociometric measurement to another. Psychometrika 1953;18(3):249–56.
26. Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika 1996;61(3):401–25.
27. Strauss D, Ikeda M. Pseudolikelihood estimation for social networks. J Am Stat Assoc 1990;85(409):204–12.
28. Snijders TAB. Markov chain monte carlo estimation of exponential random graph models. J Soc Struct 2002;3(2):.
29. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000;10(4):325–37. [http://dx.doi.org/10.1023/A:1008929526011].
30. Adams S, Carter N, Hadlock C, Haughton D, Sirbu G. A time effect in a social network from a Bayesian perspective. Connections (INSNA) 2007. [http://neeo.univ-tlse1.fr/2242/].
31. Li Y, Agarwal P, Rajagopalan D. A global pathway crosstalk network. Bioinformatics 2008;24(12):1442–7.
32. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2014;42(D1):D472—7.
33. Kanehisa M. The KEGG database. Novartis Found Symp 2002;247:91–101.
34. Molecular signature database V4.0 [http://www.broadinstitute.org/gsea/msigdb/index.jsp].
35. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res 2012;40(Database issue):D1301–7.
36. Komurov K, Dursun S, Erdin S, Ram P. NetWalker: a contextual network analysis tool for functional genomics. BMC Genomics 2012;13:282.
37. Capaccione KM, Pine SR. The Notch signaling pathway as a mediator of tumor survival. Carcinogenesis 2013;34(7):1420–30.
38. Liu ZJ, Xiao M, Balint K, Smalley KS, Brafford P, Qiu R, et al. Notch1 signaling promotes primary melanoma progression by activating mitogen-activated protein kinase/phosphatidylinositol 3-kinase-Akt pathways and up-regulating N-cadherin expression. Cancer Res 2006;66(8):4182–90.
39. Jarriault S, Brou C, Logeat F, Schroeter EH, Kopan R, Israel A. Signalling downstream of activated mammalian Notch. Nature 1995;377(6547):355–8.
40. Palomero T, Sulis ML, Cortina M, Real PJ, Barnes K, Ciofani M, et al. Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. Nat Med 2007;13(10):1203–10.
41. Palomero T, Lim WK, Odom DT, Sulis ML, Real PJ, Margolin A, et al. NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. Proc Nat Acad Sci USA 2006;103(48):18261–6.
42. Weng AP, Millholland JM, Yashiro-Ohtani Y, Arcangeli ML, Lau A, Wai C, et al. c-Myc is an important direct target of Notch1 in T-cell acute lymphoblastic leukemia/lymphoma. Genes Dev 2006;20(15):2096–109.
43. Strobl LJ, Hofelmayr H, Marschall G, Brielmeier M, Bornkamm GW, Zimber-Strobl U. Activated Notch1 modulates gene expression in B cells similarly to Epstein-Barr viral nuclear antigen 2. J Virol 2000;74(4):1727–35.
44. Zhou BP, Liao Y, Xia W, Zou Y, Spohn B, Hung MC. HER-2/neu induces p53 ubiquitination via Akt-mediated MDM2 phosphorylation. Nat Cell Biol 2001;3(11):973–82.
45. Vannucchi S, Chiantore MV, Fiorucci G, Percario ZA, Leone S, Affabris E, et al. TRAIL is a key target in S-phase slowing-dependent apoptosis induced by interferon-beta in cervical carcinoma cells. Oncogene 2005;24(15):2536–46.
46. Higashitsuji H, Higashitsuji H, Itoh K, Sakurai T, Nagao T, Sumitomo Y, et al. The oncoprotein gankyrin binds to MDM2/HDM2, enhancing ubiquitylation and degradation of p53. Cancer Cell 2005;8:75–87.
47. Cui J, Jiang W, Wang S, Wang L, Xie K. Role of Wnt/beta-catenin signaling in drug resistance of pancreatic cancer. Curr Pharm Des 2012;18(17):2464–71.
48. Luu HH, Zhang R, Haydon RC, Rayburn E, Kang Q, Si W, et al. Wnt/β-catenin signaling pathway as a novel cancer drug target. Curr Cancer Drug Targets 2004;4(8):653–71.
49. Loh YN, Hedditch EL, Baker LA, Jary E, Ward RL, Ford CE. The Wnt signalling pathway is upregulated in an in vitro model of acquired tamoxifen resistant breast cancer. BMC Cancer 2013;13:174.
50. Chikazawa N, Tanaka H, Tasaka T, Nakamura M, Tanaka M, Onishi H, et al. Inhibition of Wnt signaling pathway decreases chemotherapy-resistant side-population colon cancer cells. Anticancer Res 2010;30(6):2041–8.
51. Yamaguchi T, Yanagisawa K, Sugiyama R, Hosono Y, Shimada Y, Arima C, et al. NKX2-1/TITF1/TTF-1-Induced ROR1 is required to sustain EGFR survival signaling in lung adenocarcinoma. Cancer Cell 2012;21(3):348–61.
52. Neufeld KL. Nuclear APC. Adv Exp Med Biol 2009;656:13–29.
53. Asahara H, Li Y, Fuss J, Haines DS, Vlatkovic N, Boyd MT, et al. Stimulation of human DNA polymerase epsilon by MDM2. Nucleic Acids Res 2003;31(9):2451–9.
54. Kraker AJ, Moore CW. Elevated DNA polymerase beta activity in a cis-diamminedichloroplatinum(II) resistant P388 murine leukemia cell line. Cancer Lett 1988;38(3):307–14.
55. Anastas JN, Moon RT. WNT signalling pathways as therapeutic targets in cancer. Nat Rev Cancer 2012;13:11–26. [http://dx.doi.org/10.1038/nrc3419].
56. Tenbaum SP, Ordonez-Moran P, Puig I, Chicote I, Arques O, Landolfi S, et al. β-catenin confers resistance to PI3K and AKT inhibitors and subverts FOXO3a to promote metastasis in colon cancer. Nat Med 2012;18(6):892–901.
57. Zhan X, Wang J, Liu Y, Peng Y, Tan W. GPCR-like signaling mediated by smoothened contributes to acquired chemoresistance through activating Gli. Mol Cancer 2014;13:4.
58. Thomas A, OHara B, Ligges U, Sturtz S. Making BUGS Open. R News 2006;6:12–7.
59. Abe Y, Oda-Sato E, Tobiume K, Kawauchi K, Taya Y, Okamoto K, et al. Hedgehog signaling overrides p53-mediated tumor suppression by activating Mdm2. Proc Nat Acad Sci USA 2008;105(12):4838–43.
60. Ayers KL, Therond PP. Evaluating Smoothened as a G-protein-coupled receptor for Hedgehog signalling. Trends Cell Biol 2010;20(5):287–98.
61. Philipp M, Caron MG. Hedgehog signaling: is Smo a G protein-coupled receptor? Curr Biol 2009;19(3):R125–7.
62. van der Veeken J, Oliveira S, Schiffelers RM, Storm G, van Bergen En Henegouwen PM, Roovers RC. Crosstalk between epidermal growth factor receptor- and insulin-like growth factor-1 receptor signaling: implications for cancer therapy. Curr Cancer Drug Targets 2009;9(6):748–60.
63. Fidler MJ, Shersher DD, Borgia JA, Bonomi P. Targeting the insulin-like growth factor receptor pathway in lung cancer: problems and pitfalls. Ther Adv Med Oncol 2012;4(2):51–60.
64. Wang Y, Yuan JL, Zhang YT, Ma JJ, Xu P, Shi CH, et al. Inhibition of both EGFR and IGF1R sensitized prostate cancer cells to radiation by synergistic suppression of DNA homologous recombination repair. PLoS ONE 2013;8(8):e68784.
65. Ludovini V, Bellezza G, Pistola L, Bianconi F, Di Carlo L, Sidoni A, et al. High coexpression of both insulin-like growth factor receptor-1 (IGFR-1) and epidermal growth factor receptor (EGFR) is associated with shorter disease-free survival in resected non-small-cell lung cancer patients. Ann Oncol 2009;20(5):842–9.
66. Knowlden JM, Jones HE, Barrow D, Gee JM, Nicholson RI, Hutcheson IR. Insulin receptor substrate-1 involvement in epidermal growth factor receptor and insulin-like growth factor receptor signalling: implication for Gefitinib ('Iressa') response and resistance. Breast Cancer Res Treat 2008;111:79–91.
67. Levine AJ, Feng Z, Mak TW, You H, Jin S. Coordination and communication between the p53 and IGF-1-AKT-TOR signal transduction pathways. Genes Dev 2006;20(3):267–75.
68. Azad AK, Lee H. Voting-based cancer module identification by combining topological and data-driven properties. PLoS ONE 2013;8(8):e70498.

Chapter 5

Bayesian Model of Signal Rewiring Reveals Mechanisms of Gene Dysregulation in Acquired Drug Resistance in Breast Cancer

Chapter Objectives

In this chapter, I continued to identify signalling cross-talk among signalling pathways in data-driven networks and explore their roles in acquired drug resistance. In particular, I build a computational framework to model signalling rewiring in acquired resistance using the p_1-model. After inferring aberrant signalling activities in the rewired signalling network, I investigate two further research objectives: 1) identifying dysregulated signalling pathways in acquired resistance, and 2) detecting, analysing and characterising both Type-I and Type-II cross-talk among all signalling pathways [Chapter 3] involved in acquired lapatinib resistance. Supplementary files are included in Appendix C.

Authorship

A. K. M. Azad[1], Alfons Lawen[2], Jonathan M Keith[1]

1 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

2 Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia

Reference

Azad A.K.M., Lawen A., Keith JM. (2016). Bayesian Model of Signal Rewiring Reveals Mechanisms of Gene Dysregulation in Acquired Drug Resistance in Breast Cancer. [Accepted] **PLoS ONE**

PLOS ONE

RESEARCH ARTICLE

Bayesian model of signal rewiring reveals mechanisms of gene dysregulation in acquired drug resistance in breast cancer

A. K. M. Azad[1]*, Alfons Lawen[2], Jonathan M. Keith[1]

1 School of Mathematical Sciences, Monash University, Clayton, VIC, Australia, **2** Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University, Clayton, VIC, Australia

* aaza7@student.monash.edu

OPEN ACCESS

Citation: Azad AKM, Lawen A, Keith JM (2017) Bayesian model of signal rewiring reveals mechanisms of gene dysregulation in acquired drug resistance in breast cancer. PLoS ONE 12(3): e0173331. https://doi.org/10.1371/journal.pone.0173331

Editor: Aamir Ahmad, University of South Alabama Mitchell Cancer Institute, UNITED STATES

Received: October 13, 2016

Accepted: February 20, 2017

Published: March 13, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by Monash International Postgraduate Research Scholarship and Monash Graduate Scholarship at the Monash University, Australia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Small molecule inhibitors, such as lapatinib, are effective against breast cancer in clinical trials, but tumor cells ultimately acquire resistance to the drug. Maintaining sensitization to drug action is essential for durable growth inhibition. Recently, adaptive reprogramming of signaling circuitry has been identified as a major cause of acquired resistance. We developed a computational framework using a Bayesian statistical approach to model signal rewiring in acquired resistance. We used the p_1-model to infer potential aberrant gene-pairs with differential posterior probabilities of appearing in resistant-vs-parental networks. Results were obtained using matched gene expression profiles under resistant and parental conditions. Using two lapatinib-treated ErbB2-positive breast cancer cell-lines: SKBR3 and BT474, our method identified similar dysregulated signaling pathways including EGFR-related pathways as well as other receptor-related pathways, many of which were reported previously as compensatory pathways of EGFR-inhibition via signaling cross-talk. A manual literature survey provided strong evidence that aberrant signaling activities in dysregulated pathways are closely related to acquired resistance in EGFR tyrosine kinase inhibitors. Our approach predicted literature-supported dysregulated pathways complementary to both node-centric (SPIA, DAVID, and GATHER) and edge-centric (ESEA and PAGI) methods. Moreover, by proposing a novel pattern of aberrant signaling called V-structures, we observed that genes were dysregulated in resistant-vs-sensitive conditions when they were involved in the switch of dependencies from *targeted* to *bypass* signaling events. A literature survey of some important V-structures suggested they play a role in breast cancer metastasis and/or acquired resistance to EGFR-TKIs, where the mRNA changes of *TGFBR2*, *LEF1* and *TP53* in resistant-vs-sensitive conditions were related to the dependency switch from *targeted* to *bypass* signaling links. Our results suggest many signaling pathway structures are compromised in acquired resistance, and V-structures of aberrant signaling within/among those pathways may provide further insights into the bypass mechanism of targeted inhibition.

Introduction

Cell signaling pathways transduce input signals from extracellular to intracellular environments and determine various cell activities, including cell growth, proliferation, differentiation, migration, and apoptosis [1, 2]. Perturbation of a signaling network may occur when there are genetic alterations, such as DNA mutations and/or amplifications/deletions of a genomic region, or changes in gene expression (GE) [3, 4]. For example, the amplification or over-expression of the ErbB2 (HER2/neu) oncogene, that enhances various growth-related signaling activities [5] from receptor-level to effector-level [4], is commonly found in about 25% of breast cancer patients. In the majority of cancers, aberrant activities in signaling pathways are involved in various stages of tumor progression and metastasis [6–9].

Drugs targeting a signaling network, such as EGFR signaling pathway, often become ineffective as acquired resistance develops in cancer cells [10]. Primary reasons for acquired resistance to EGFR family receptor targeted therapies include: secondary mutations of targeted genes (e.g., the EGFR T790M mutation [11]), transcriptional and post-translational up-regulation of RTKs (Receptor Tyrosine Kinases) both within the receptor-family (e.g. ERBB3/HER3 [12, 13]) and other kinases (i.e. IGF1R, MET, FGFR2, FAK, SRC family kinases [14–16]), the over-expression of ABC transporters [3], and the re-activation of targeted pathways [5]. Moreover, tumor cells induce adaptive responses to targeted therapies [5] by *rewiring* in such a way that the adaptive signaling bypasses the inhibiting effects of initial treatments [4, 10, 17–19]. Therefore, rewiring of signaling networks plays a vital role as a non-genetic mechanism of acquired resistance [3, 14, 17, 18, 20]; targeting of which has the potential to improve the response durability of single kinase inhibitors [4, 5, 21]. However, reprogramming of signaling activities in acquired resistance inherently imposes increased uncertainties in the network structure when compared with their sensitive counterparts.

The functionality of biological networks is determined by their underlying architecture. Thus understanding, characterising, and analysing network structures are very important tasks in the field of systems biology [22]. Statistical modeling approaches offer a great deal of flexibility in terms of scalability and the number of local features that can be incorporated [22]. Moreover, as in other biological networks, signaling activities predicted using signaling data may be unreliable, whereas some crucial signaling links may not be predicted [23]. Measurements of the signaling activities often yield noisy data. Therefore, for such data-driven signaling networks a statistical modeling approach such as *exponential random graph models* (ERGMs) or p^* can be a suitable choice [22, 23]. The p_1-model, a special class of ERGMs which was originally proposed by Holland and Leinhardt [24], models the probability of an edge formation in the observed network based on network statistics (e.g. node degree) and associating model parameters with those statistics [22, 23].

Measuring the probabilistic nature of pair-wise relationships is an important aspect of modeling a gene-gene relationship network. Particularly in cancer drug resistance, some relationships between gene-pairs may evolve in the resistant conditions to compensate for the inhibiting effects of the drugs used [10, 19]. Some gene-pairs may have higher probabilities of evolving correlations in resistant conditions than in sensitive conditions. Simultaneously, some gene-pairs having high correlations in sensitive conditions may become loosely correlated (or even independent) in resistant conditions. For example, Komurov *et al.* reported that genes of the *glucose-deprivation response network* are up-regulated in lapatinib- (an EGFR/HER2 dual inhibitor) resistant conditions, thus providing an EGFR-independent mechanism of glucose uptake in cancer cells [19]. ErbB2-positive cancer cells largely depend on EGFR/ErbB2 signaling for their glucose uptake [19] which was recently reported as a major factor in oncogenic KRAS pathway mutations [25, 26]. Lapatinib mediates down-regulation of cell

cycle machinery and up-regulation of cell cycle inhibitory complexes that are downstream of EGFR/ErbB2 signaling [19]. Moreover, the inhibitory effect of lapatinib on EGFR/ErbB2 signaling in the sensitive condition was found to be associated with glucose starvation of cancer cells, and thus induced cancer cell death [19]. However, in resistant conditions, up-regulation of activities involved in the *glucose deprivation response network* (and other hypoglycemic response networks) played an important role as a compensatory mechanism of glucose uptake in cancer cells for which tumors ultimately relapsed. Therefore, it can be hypothesized that genes involved in the process of cell proliferation and survival may evolve, in resistant conditions, to be highly correlated with the genes in the *glucose deprivation response network* in order to establish an alternate mechanism of glucose uptake in cancer cells, even though the inhibiting effects of lapatinib abrogated their dependencies on EGFR/ErbB2 signaling in sensitive conditions (See Fig 1 of [20].) Therefore, studying systematic characterizations of such differential dependencies among gene-pairs in resistant-vs-sensitive conditions, and their combined roles on particular genes' dysregulations (in resistant-vs-sensitive) may reveal novel insights into mechanisms of acquired resistance.

Moreover, Komurov *et al.* [19] suggested that the drug resistance mechanism more likely occurs downstream of growth factor-mediated signaling pathways, such as Ras signaling, PI3K/AKT signaling, mTOR signaling, and others. However, an enormous number of diverse effector pathways may be involved in this process, making the prediction of biologically plausible hypotheses a challenging task. New computational approaches are needed to resolve such challenges in identifying the mechanistic underpinnings of acquired resistance.

Gene dysregulation is associated with aberrant signaling activities that are crucial for both cell growth and apoptosis in breast cancer [27]. For example, dysregulation [28] and/or mutation [28, 29] of apoptosis-related genes may overcome the initial response to apoptotic stimuli, thereby conferring resistance to apoptosis. Sharifnia et al. recently reported that several kinases and kinase-related genes from the Src family (e.g. *FGFR*1, *FGFR*2 and *MOS*) can compensate the loss of EGFR activity across multiple EGFR-dependent models [30]. Using unbiased gene-expression profiles of cells, their study revealed that over-expression of these EGFR-bypass genes plays a critical role in EGFR-independent activation of the MEK-ERK and PI3K-AKT signaling pathways in EGFR-mutant NSCLC cells. Recently, differential dependencies/associations were used to model rewiring in biological networks [31, 32]. Therefore, we hypothesize that differential associations between genes identified by modeling network reprogramming in resistant-vs-sensitive conditions could potentially explain gene dysregulation in acquired resistance.

In this study, we propose a computational framework to identify dysregulated signaling pathways in resistant-vs-sensitive conditions, and a possible mechanism of gene dysregulation in acquired resistance. The schematic diagram of our proposed framework is shown in Fig 1. We used two breast cancer cell-lines, SKBR3 and BT474, each having gene expression values measured under matched lapatinib-sensitive (parental) and lapatinib-resistant conditions. A gene-gene relationship network was constructed for each gene expression dataset by combining data-driven and protein-protein interaction (PPI) information indicative of both direct and indirect relationships between gene-pairs. Then we applied a fully Bayesian approach involving the p_1-model to infer gene-pairs with differential posterior probabilities between these two conditions. Next, statistically significant dysregulated signaling pathways from KEGG, Reactome, and WikiPathway were identified by enriching putative aberrant pairs using literature curated signaling links. Finally, by proposing a novel pattern of aberrant pairs, called a V-structure, we identified possible mechanisms of dysregulation in resistant-vs-sensitive conditions that may be crucial for breast cancer metastasis and/or EGFR-TKI resistance.

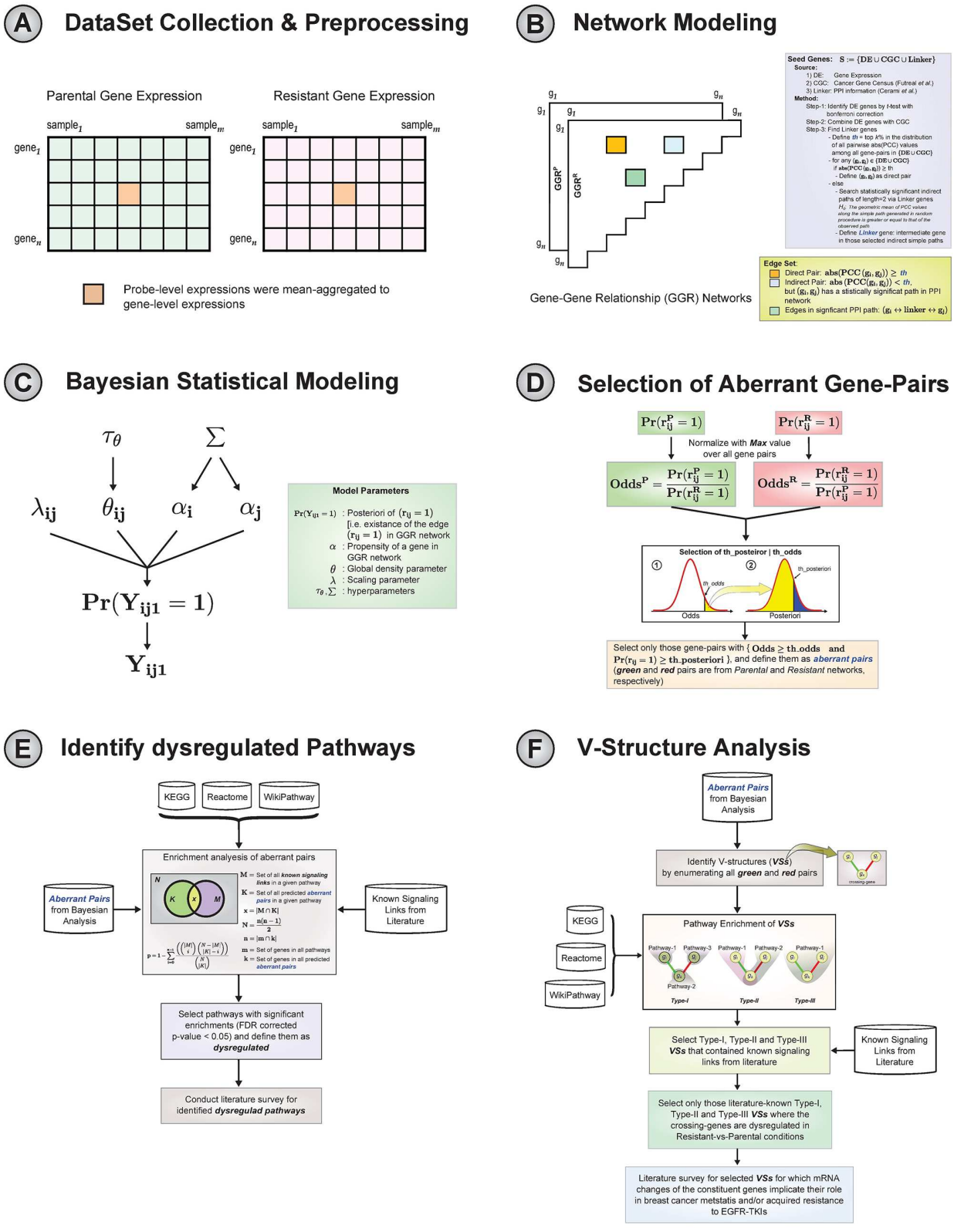

A DataSet Collection & Preprocessing
Parental Gene Expression
Resistant Gene Expression
sample1
samplem
gene1
genen
Probe-level expressions were mean-aggregated to gene-level expressions
B Network Modeling
Gene-Gene Relationship (GGR) Networks
Seed Genes: S := {DE ∪ CGC ∪ Linker}
Source:
1) DE: Gene Expression
2) CGC: Cancer Gene Census (Futreal et al.)
3) Linker: PPI information (Ceramiet al.)
Method:
Step-1: Identify DE genes by t-test with bonferroni correction
Step-2: Combine DE genes with CGC
Step-3: Find Linker genes
- Define th = top k% in the distribution of all pairwise abs(PCC) values among all gene-pairs in {DE ∪ CGC}
- for any (gi, gj) ∈ {DE ∪ CGC}
if abs(PCC(gi, gj)) ≥ th
- Define (gi, gj) as direct pair
- else
- Search statistically significant indirect paths of length=2 via Linker genes
H0: The geometric mean of PCC values along the simple path generated in random procedure is greater or equal to that of the observed path
- Define Linker gene: intermediate gene in those selected indirect simple paths
Edge Set:
Direct Pair: abs(PCC(gi, gj)) ≥ th
Indirect Pair: abs(PCC(gi, gj)) < th, but (gi, gj) has a statistically significat path in PPI network
Edges in signficant PPI path: (gi ↔ linker ↔ gj)
C Bayesian Statistical Modeling
Pr(Yij1 = 1)
Yij1
Model Parameters
Pr(Yij1 = 1) : Posteriori of (rij = 1) [i.e. existance of the edge (rij = 1) in GGR network
α : Propensity of a gene in GGR network
θ : Global density parameter
λ : Scaling parameter
τθ, Σ : hyperparameters
D Selection of Aberrant Gene-Pairs
Normalize with Max value over all gene pairs
Selection of th_posteiror | th_odds
th_odds
th_posteriori
Odds
Posteriori
Select only those gene-pairs with { Odds ≥ th_odds and Pr(rij = 1) ≥ th_posteriori }, and define them as aberrant pairs (green and red pairs are from Parental and Resistant networks, respectively)
E Identify dysregulated Pathways
KEGG
Reactome
WikiPathway
Enrichment analysesis of aberrant pairs
Aberrant Pairs from Bayesian Analysis
Known Signaling Links from Literature
M = Set of all known signaling links in a given pathway
K = Set of all predicted aberrant pairs in a given pathway
x = |M ∩ K|
N = n(n − 1)/2
n = |m ∩ k|
m = Set of genes in all pathways
k = Set of genes in all predicted aberrant pairs
Select pathways with significant enrichments (FDR corrected p-value < 0.05) and define them as dysregulated
Conduct literature survey for identified dysregulad pathways
F V-Structure Analysis
Aberrant Pairs from Bayesian Analysis
Identify V-structures (VSs) by enumerating all green and red pairs
crossing-gene
KEGG
Reactome
WikiPathway
Pathway Enrichment of VSs
Pathway-1
Pathway-3
Pathway-2
Type-I
Type-II
Type-III
Select Type-I, Type-II and Type-III VSs that contained known signaling links from literature
Known Signaling Links from Literature
Select only those literature-known Type-I, Type-II and Type-III VSs where the crossing-genes are dysregulated in Resistant-vs-Parental conditions
Literature survey for selected VSs for which mRNA changes of the constituent genes implicate their role in breast cancer metstatis and/or acquired resistance to EGFR-TKIs

Fig 1. Schematic diagram of our proposed framework to identify and analyse aberrant signaling pathways in acquired resistance. (A) Gene expression datasets of breast cancer cell-lines for both parental and resistant conditions. (B) Two gene-gene relationship networks (GGR) were built from gene expression datasets of breast cancer cell-lines in Parental and resistant conditions. (C) & (D) A fully Bayesian approach was applied for detecting putative aberrant gene-pairs involved in acquired resistance. (E) Using the putative aberrant gene-pairs and a literature-curated signaling network, a statistical test was conducted to identify dysregulated pathways in acquired resistance. (F) Applying the known aberrant signaling links (from literature), we identify and explain the role of a proposed novel structure of aberrant pairs: V-structure (*VS*) in breast cancer metastasis and/or in developing acquired resistance to EGFR-TKIs.

https://doi.org/10.1371/journal.pone.0173331.g001

We hope such patterns revealed using our framework will lead to further insights into aberrant signaling activities in acquired resistance.

Results

A framework for identifying putative aberrant gene-pairs in acquired resistance

We developed a computational framework exploiting Bayesian statistical modeling to identify putative aberrant signaling links involved in acquired resistance. In this study, we hypothesized that aberrant signaling can be detected as differential probabilities of occurrence of gene-pairs in resistant-vs-parental conditions. Thus, after building gene-gene relationship networks individually from both parental and resistant conditions, a comparative study of edge probabilities in those two networks may reveal aberrant relationships due to acquired resistance.

Our framework constructs a gene-gene relationship network, $GGR := (S, R)$ by combining GE and PPI datasets, where S is a set of seed genes and R is a set of pair-wise gene relationships (Fig 1). Table 1 shows primary statistics for the *GGR* networks of both SKBR3 (GSE38376) and BT474 (GSE16179) cell-lines. For SKBR3 cell-lines (Parental and Resistant), we selected 897 seed genes comprised of 345 differentially expressed (DE) genes (Bonferroni corrected p-value ≤ 0.01), 370 genes from the Cancer Gene Census (CGC), and 502 and 479 linker genes from Resistant and Parental cell-lines, respectively. For BT474 cell-lines, we found 875 distinct seed genes comprised of 354 DE genes (Bonferroni corrected p-value ≤ 0.05), 357 CGC genes, and 477 and 489 linker genes from Resistant and Parental cell-lines, respectively. Note that to find DE genes in SKBR3 and BT474 cell-lines, two different p-value thresholds: 0.01 and 0.05 were used, respectively. This was done for two reasons: firstly, because the computational cost of using a conventional threshold of 0.05 with SKBR3 was prohibitive, and secondly, to ensure the numbers of DE genes in the two different cell-lines were comparable, and similarly for the sizes of the seed gene sets [for details see S1 Text].

Table 1. Primary statistics of Gene-Gene Relationship (*GGR*) network construction for both SKBR3 and BT474 cell-lines.

Cell Line	Cell Condition	# of DE Genes	# of CGC Genes	# of DE ∪ CGC Genes	# of All Pairs	# of Linker Genes	# of Total Seed Genes	# of combined Seed Genes	# of Direct Pairs	# of Indirect Pairs	# of PPI Pairs	# of Total Links
SKBR3	Resistant	345	370	704	247456	502	1262	897	49492	1440	1757	52560
	Parental					479	1245			1393	1758	52510
BT474	Resistant	354	357	698	243253	477	1100	875	48651	1572	1895	51998
	Parental					489	1101			1517	1951	51972

https://doi.org/10.1371/journal.pone.0173331.t001

Our approach constructs a *GGR* in a series of stages: an initial set of genes is obtained by combining DE and CGC genes. Edges are added corresponding to direct relationships between pairs of these genes. We then search for indirect relationships among gene-pairs for which direct relationships couldn't be found, and where indirect relationships are found the linker genes and the edges connecting them are added to the network. For the SKBR3 cell-line, the initial gene set contained 704 genes obtained by combining 345 DE and 370 CGC genes, whereas for the BT474 cell-line, the initial gene set contained 698 genes obtained by combining 354 DE and 357 CGC genes. To define direct relationships among the genes in the initial sets, we chose the top 20% from the ranked list of all pair-wise absolute Pearson Correlation Coefficients (PCC). Thus, we identified 49,492 (in both parental and resistant condition) and 48,651 (in both parental and resistant condition) direct relationships in SKBR3 and BT474 cell-lines, respectively. We justified this choice of threshold by applying an approach proposed by Elo *et al.* which analyses the topological properties of a co-expression network in order to find an optimal cutoff value [33] [for details see S1 Text]. In searching for indirect relationships, we found that 502 and 479 linker genes connect 1,440 and 1,393 distinct gene-pairs (for which direct relationships were not found) with the help of 1,757 and 1,758 distinct PPI links, for SKBR3 resistant and parental cell-lines, respectively. Similarly, for BT474 Resistant and Parental cell-lines, 477 and 489 linker genes connect 1,572 and 1,517 distinct indirect gene-pairs along with 1,895 and 1,951 distinct PPI links, respectively. In both datasets (SKBR3 and BT474), to build two *GGR* matrices for resistant and parental conditions with similar sets of genes, we constructed the final set of seed genes as an intersection of the two individual seed gene sets for Resistant and Parental conditions. Hence, 502 and 479 linker genes from SKBR3 resistant and parental conditions were combined with 704 ($DE \cup CGC$) genes to form 1,262 and 1,245 seed genes, respectively, and then finding an intersection of these two sets yielded a set of 897 genes. Similarly, combining 698 ($DE \cup CGC$) with 477 and 489 linker genes from BT474 resistant and parental genes produced 1100 and 1101 seed genes, respectively, and intersecting these resulted in a final set of 875 genes. At the end of this process, the SKBR3 resistant and parental *GGR* networks contained 897 distinct seed genes ($DE \cup CGC \cup Linker$) with 52,560 and 52,510 gene-gene relationships ($direct \cup indirect \cup PPI$), respectively, and the BT474 Resistant and Parental *GGR* network contained 875 distinct seed genes with 51,998 and 51,972 gene-gene relationships, respectively. Note that for both SKBR3 and BT474 cell-lines, although the total number of final seed genes is the same for both resistant and parental conditions, their respective *GGR* networks may contain different numbers of gene-gene relationships.

After building the *GGR* networks for both resistant and parental conditions Y_k^R and Y_k^P separately, we conducted Bayesian inference of parameters using the p_1-model to estimate posterior probabilities of gene-gene relationships in each network. We used a WinBUGS script used in our previous work [10] for this inference. We ran the MCMC (Markov Chain Monte Carlo) method for 15,000 iterations, where the first 10,000 iterations were considered as 'burn-in', and the next 5,000 iterations were used for sampling. Time-series plots indicated that all parameters converged within the first few thousand iterations (data not shown). In both networks, the posterior probability of each edge was estimated to be the proportion of the 5,000 sampled networks in which that edge was present.

We identified a gene pair ($gene_i$, $gene_j$) as putatively aberrant if its posterior probabilities $Pr(Y_{ij1}^R = 1)$ and $Pr(Y_{ij1}^P = 1)$ of appearing in each network (resistant and parental networks, respectively) are significantly different. To determine which gene-pairs had this characteristic, we calculated two odds ratios—$Odds^R$ and $Odds^P$—as shown in Eqs (3) and (4) for each gene-pair ($gene_i$, $gene_j$). Note that since the two posterior probabilities used in these odds ratios may lie in different ranges, we normalized their values by dividing by their respective maximum

values over all the gene-pairs in the respective sets. We then used two thresholds to define significance: first, we constructed the empirical distribution of odds ratios and chose only those gene-pairs which had odds ratios among the top 20%. For SKBR3 cell-lines, these threshold values were 2.53 and 1.66 for resistant and parental conditions, respectively, and for BT474 these values were 12.028 and 2.115, respectively. Next, we constructed empirical distributions of the posterior probabilities of the previously selected gene-pairs, and chose only those gene-pairs whose posterior probabilities were in the top 50% in their respective distributions. For SKBR3 resistant and parental cell-lines, these thresholds of posterior probabilities were 0.212 and 0.252, respectively, and for BT474 cell lines, 0.177 and 0.304, for resistant and parental conditions, respectively. More detailed explanations regarding these two types of thresholds are provided in the Supplementary Methods section in S1 Text. Thus, our framework finally selected 80,372 and 76,476 aberrant gene-pairs for SKBR3 and BT474 cell-lines, respectively, and we hypothesized that these aberrant gene-pairs have the potential to explain the mechanism of acquired resistance in breast cancer. Lists of all identified putative aberrant gene-pairs for both SKBR3 and BT474 cell-lines are shown in S1 Table.

Comparing posterior probabilities to correlation coefficients. To investigate the robustness of our approach, we compared the posterior probabilities with the initial PCC (Pearson Correlation Coefficient) values for each of the putative aberrant gene pairs as shown in Fig 2. We treated the posterior probabilities of the *red* gene-pairs [see Methods] as positive values and the posterior probabilities of the *green* gene-pairs [see Methods] as negative, and plotted their sorted values in descending order (Fig 2). Next, we constructed a scatter plot with *corresponding* absolute PCC values for each of these gene-pairs, sorted based on posterior probabilities. We added a trendline using a moving average with *window size* 25, to investigate whether this trendline was in any way similar to the trend observed in the posterior probabilities. Interestingly, for both SKBR3 and BT474 cell-lines, the trendlines of PCC values revealed a visually similar pattern to that of the corresponding posterior probability values. This confirms our expectation that our Bayesian analysis is sensitive to a signal in the PCC values that would be otherwise difficult to detect.

Many crucial signaling pathways are significantly enriched with aberrant gene-pairs in acquired resistance

To measure the significance of signaling pathways in terms of aberrant signaling activities in acquired resistance, we conducted a hypergeometric test. In this test, we measured how significant was the overlap between the set of literature-supported signaling links [34] found in a particular signaling pathway with the set of putative aberrant gene-pairs in the same pathway. We identified all the signaling pathways from KEGG, Reactome, and WikiPathway databases for which the corresponding q-value (FDR corrected p-value) from the above hypergeometric test was < 0.05 in both SKBR3 and BT474 cell-lines as is shown in Fig 3. For both SKBR3 and BT474 cell-lines, 71.11% (32 out of 45), 62.5% (15 out of 24), and 57.38% (35 out of 61) signaling pathways from KEGG, Reactome, and WikiPathways, respectively, were found to be significantly enriched with aberrant signaling gene-pairs in acquired resistance (Fig 3). Again, for all corresponding KEGG, Reactome, and WikiPathway databases, such high percentages of enriched signaling pathways found in both SKBR3 and BT474 cell-lines indicates that our framework is consistent in terms of finding aberrant gene-pairs in both cell-lines. Complete enrichment results of this hypergeometric test are reported in S2 Table.

We conducted a literature survey for the putative dysregulated signaling pathways, and found that the aberrant activities in most of these pathways are strongly associated with acquired resistance to EGFR tyrosine kinase inhibitors (EGFR-TKIs) [18]. EGFR (also known

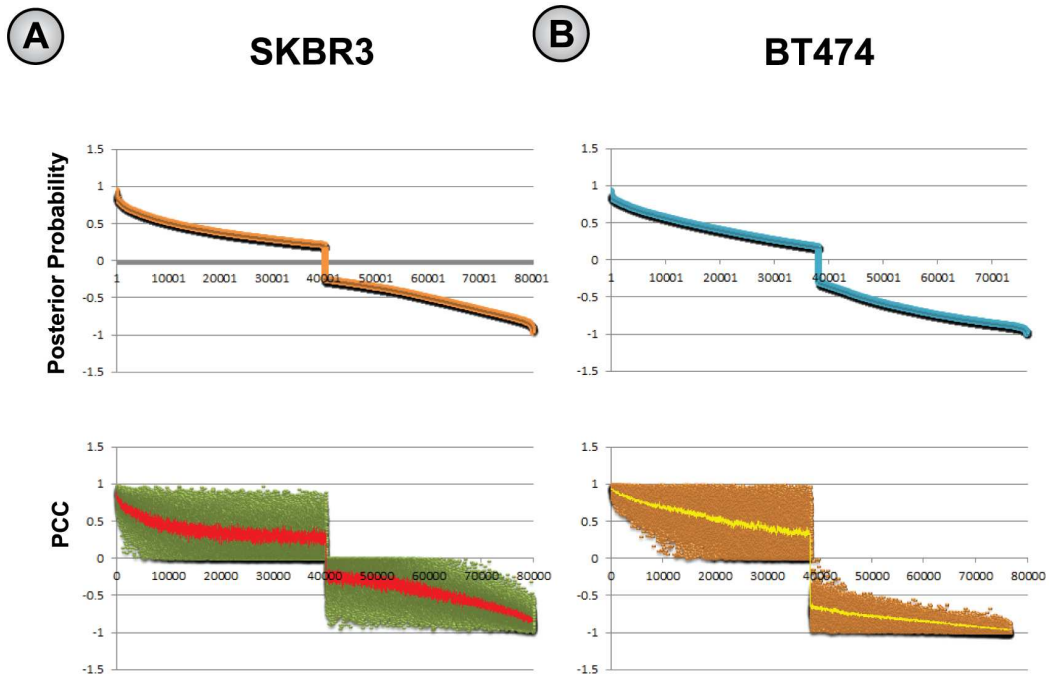


Fig 2. Comparing the posterior probabilities of putative aberrant gene-pairs with corresponding PCC (Pearson Correlation Coefficient) values that were defined among genes prior to the Bayesian analyses, (A) for SKBR3 and (B) BT474 cell-lines. The first figures in (A) and (B) show the sorted posterior probability values of the putative aberrant gene-pairs in descending order, and the second figures of (A) and (B) show the scatter plot of their corresponding PCC values. Note that for both the graphs in (A) and (B) the rank of ordered aberrant pairs is shown in X-axis, and the posterior probabilities and PCC values of *red* gene-pairs are shown in Positive Y-axis and those of *green* pairs are shown in Negative Y-axis, correspondingly. A trendline (red or yellow trendlines for the SKBR3 and BT474 cell-lines, respectively) is drawn for each of the scatter plots (in (A) and (B)) by using a moving average with a *window size* set to 25. For both SKBR3 and BT474, these trendlines clearly show the similarity of the signal contained in the PCC values (defined prior to Bayesian analyses) and the pattern of changes in *a posteriori* values (resulting from Bayesian analyses), and demonstrates the robustness of Bayesian statistical modeling for selecting putative aberrant gene-pairs involved in acquired resistance.

https://doi.org/10.1371/journal.pone.0173331.g002

as HER1, or ErbB1) and EGFR 2 (also known as HER2/neu, or ErbB2) are cell surface transmembrane proteins, and members of the HER family of receptors. EGFR (in KEGG, Reactome, and WikiPathway) and ErbB2 (in Reactome) are reported to be frequently mutated and/or over-expressed in various types of cancer resulting in aberrant activities contributing to abnormal cell growth, survival, migration, and differentiation [35, 36]. However, over-expression and secondary mutations of both EGFR [11, 37–39] and ErbB2 [40] are associated with acquired resistance to EGFR-TKI. Moreover, being key components of cell signaling systems, these RTKs control major downstream signaling pathways, i.e. Ras/Raf/MAPK (in KEGG and WikiPathway), PI3K-Akt (in KEGG and Reactome), FoxO (in KEGG), and Jak-STAT (in KEGG) that are crucial for cancer cell growth and survival [3]. Moreover, as these downstream signaling pathways further regulate multiple downstream *effector* pathways (related to cell growth and survival), aberrant re-activation of those pathways provide a common mechanism to compensate for inhibition of targeted pathways, thereby conferring acquired resistance to EGFR-TKIs [4, 41, 42]. Interestingly, these signaling pathways (i.e. Ras, PI3K-Akt, FoxO, Jak-Stat signaling) were found as the top-most in the list of aberrant signaling pathways in both datasets (SKBR3 and BT474) based on the above hypergeometric test using KEGG database as

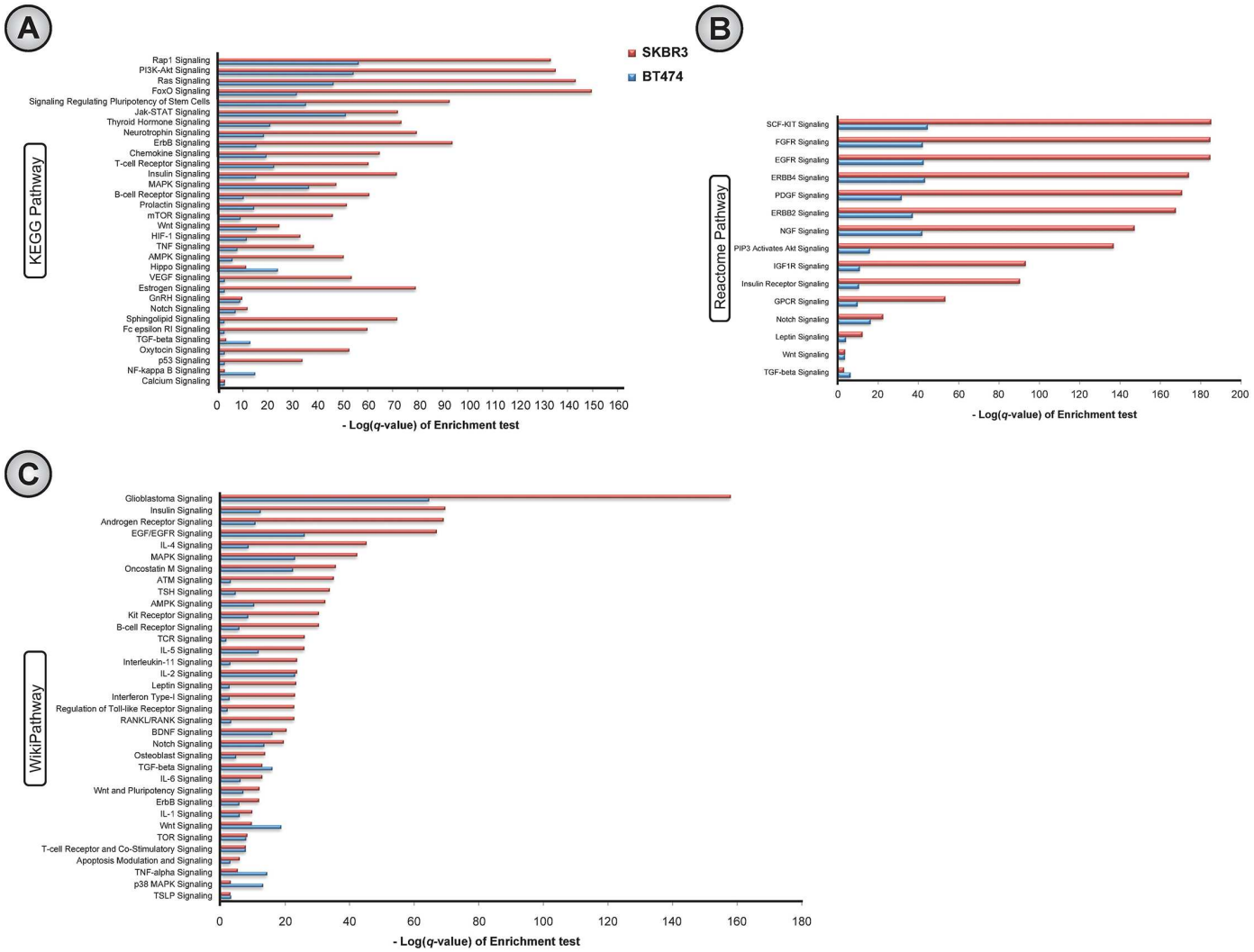


Fig 3. Analysis of dysregulated pathways by conducting pathway enrichment test of aberrant gene-pairs with known signaling links [34] involved in acquired resistance in SKBR3 and BT474 breast cancer cell-lines. Enrichments of all signaling pathways in (A) KEGG, (B) Reactome, and (C) WikiPathway.

https://doi.org/10.1371/journal.pone.0173331.g003

shown in Fig 3. For ErbB4 signaling (in Reactome), recently it has been reported that in ErbB2-positive breast cancer cell-lines, ErbB4 was up-regulated at the protein level *in vitro* and re-activated PI3K-Akt signaling in resistant conditions compared to the sensitive condition, and the knock-down of ErbB4 induced apoptosis in both the lapatinib-resistant and trastuzumab-resistant cell-lines [43]. Rap1 (in KEGG) and ras (in KEGG) signaling are activated by lung cancer oncogene CRKL whose focal amplification (secondary mutation) was reported to be associated with acquired resistance to EGFR inhibitor [44]. Again, signals for cell proliferation and survival from activated AKT may transduce through several phosphorylated transcription factors, such as FoxO (in KEGG) [45], which indicates that the dysregulation of FoxO signaling pathway (in KEGG) may potentially be associated with resistance to EGFR-TKIs.

Our previous study found cross-talks between EGFR signaling and pathways triggered by other types of receptors, e.g. Notch, Wnt, IGF1R, GPCR, etc. which contributed to acquired resistance to lapatinib (an EGFR/Her2 dual inhibitor) [10]. Here, we also found these pathways showing significant aberrant activities in acquired resistance to lapatinib [Fig 3]. The activation of IGF1R signaling (in KEGG, Reactome, and WikiPathway) is commonly reported to induce acquired resistance to EGFR-TKIs by many studies [46, 47], and its inhibition could down-regulate PI3K-Akt signaling, eventually inhibiting cell growth, providing co-inhibition of EGFR and IGF1R signaling a clinical success [3]. Similarly, the Notch signaling pathway (in KEGG, Reactome, and WikiPathway) cross-talks with EGFR signaling in breast cancer, thus maintaining the cancer cell growth signal through MAPK and PI3K-Akt signaling [48]. It is suggested that an improved drug-sensitivity could be achieved by down-regulating the Notch signaling pathway with specific inhibitors [49, 50]. Again, genes involved in Wnt signaling (in KEGG, Reactome, and WikiPathway) were up-regulated in the resistant condition in both breast and colon cancer when compared to the sensitive condition [51, 52], thus contributing to acquired resistance to EGFR-TKIs [51].

Targeting angiogenesis is another important aspect of anticancer therapies [53], as aberrant vascularity and hypoxia are directly associated with tumor growth and survival [3]. In our analysis, we found aberrant angiogenic pathways including signaling by Vascular Endothelial Growth Factors (VEGFs) (in KEGG), Fibroblast Growth Factors (FGFs), and Platelet-Derived Growth Factors (PDGFs). It has been reported that the VEGF/VEGFR-2 feed-forward loop increases VEGF secretion in lung cancer via mTOR-dependent regulation that is required for the activation of downstream signaling [54], and the over-expression of VEGFR-1 reduces EGFR-TKIs sensitivity in different human cancer cells [3, 55]. Alternate activation of the FGFR signaling pathway (in Reactome) through the over-expressions of FGFR1 and FGF2 acts as a compensating mechanism for EGFR-TKIs [56] by maintaining signals for cell survival and proliferation in the downstream signaling pathways [4]. Again, it has been recently reported that, in PDGFR signaling (in Reactome), transcriptional de-repression of PDGFR-β contributed to compensating for the effects of EGFR-TKIs in EGFR-mutant glioblastoma via an mTORC1- and extracellular signal regulated kinase-dependent mechanism [21].

The hippo signaling pathway (in KEGG) is associated with cell proliferation, apoptosis, organ size control, and stem cell self renewal [57]. YAP is a transcription co-activator and oncoprotein [58], and plays a central role in cancer-related activities of the hippo signaling pathway [57]. Huang *et al.* have recently reported that down-regulating YAP expression in various cell-lines can improve the sensitivity of erlotinib (an EGFR-TKI) and cetuximab (anti-EGFR drug) [59]. We found the gene-pair AKT2:MYC as a signaling cross-talk between EGFR/ErbB and the TGF-β signaling pathway (in KEGG, Reactome, and WikiPathway) in our previous study [10]. Recently, it has been reported that combined inhibition of EGFR-TKIs (erlotinib) and TGF-β type I receptor inhibitor may improve sensitivity of EGFR-TKIs in lung cancer without EGFR T790M mutation [60].

For both SKBR3 and BT474 cell-lines, the primary findings in this study with supporting references are summarized in Tables 2 and 3. In this table, for each aberrant pathway, we also show what percentages of predicted gene-pairs from Bayesian analysis were previously defined as direct relationships, indirect relationships, and PPI during the network modeling. It is apparent that substantial proportions of predicted pairs came from direct and indirect relationships in both SKBR3 and BT474 cell-lines. This also indicates the robustness of our Bayesian modeling in inferring gene-pair relationships. Note that in the above calculation, if a predicted pair was defined both as direct and PPI, or both as indirect and PPI, then we counted that as direct or indirect, respectively, since that prediction for that particular pair was made by our framework. Again, some of the predicted pairs (by Bayesian modeling) may not be

Table 2. Summary of predicted dysregulated EGFR and its downstream signaling pathways from KEGG, Reactome and WikiPathway databases in acquired resistance in both SKBR3 and BT474 cell-lines.

Aberrant Pathways in EGFR-TKIs Resistance[k,r,w]	% of Direct Pair[(s,b)k,r,w]	% of Indirect Pair[(s,b)k,r,w]	% of PPI Pair[(s,b)k,r,w]	# of Enriched Pair[(s,b)k,r,w]	Enrichment q-value[(s,b)k,r,w]	Literature References
EGFR and downstream pathways						
EGFR signaling	(53.12%, 71.05%)[k]	(18.75%, 13.16%)[k]	—	(6, 32)[k]	(5.1e-16, 1.5e-94)[k]	[11, 37–39]
	(30.43%, 71.57%)[r]	(8.7%, 6.86%)[r]	(__, 0.49%)[r]	(18, 73)[r]	(3.2e-43, 1.7e-185)[r]	
	(56%, 71.79%)[w]	(12%, 7.69%)[w]	(__, 0.85%)[w]	(2, 4)[w]	(1.0e-26, 1.3e-67)[w]	
ErbB2 signaling	(33.96%, 74.12%)[r]	(7.55%, 4.12%)[r]	(__, 0.59%)[r]	(15, 64)[r]	(7.8e-38, 2.5e-168)[r]	[40]
ErbB4 signaling	(29.09%, 72.19%)[r]	(9.09%, 4.73%)[r]	(__, 0.59%)[r]	(17, 65)[r]	(6.6e-44, 7.5e-175)[r]	[43]
Ras signaling	(34.62%, 66.05%)[k]	(11.54%, 6.17%)[k]	(__, 0.62%)[k]	(22, 60)[k]	(6.5e-47, 6.9e-144)[k]	[4, 41–44]
MAPK signaling	(35.82%, 60.32%)[k]	(8.96%, 7.94%)[k]	—	(19, 23)[k]	(4.4e-37, 4.2e-48)[k]	[3, 4]
	(31.82%, 48.05%)[w]	(9.09%, 12.99%)[w]	—	(12, 19)[w]	(8.7e-24, 5.1e-43)[w]	
PI3K-Akt signaling	(35.27%, 61.85%)[k]	(10.62%, 5.69%)[k]	(__, 0.95%)[k]	(34, 75)[k]	(5.4e-55, 7.2e-136)[k]	[3, 4]
	(26.67%, 73.45%)[r]	(6.67%, 4.42%)[r]	(__, 0.88%)[r]	(6, 46)[r]	(1.5e-16, 1.4e-137)[r]	
Jak-Stat signaling	(25.49%, 71.19%)[k]	(19.61%, 13.56%)[k]	(__, 1.69%)[k]	(20, 7)[k]	(6.8e-52, 8.6e-73)[k]	[3, 4]
Rap1 signaling	(25%, 61.03%)[k]	(11%, 8.09%)[k]	(1%, 0.74%)[k]	(25, 53)[k]	(4.4e-57, 5.5e-134)[k]	[44]
FoxO signaling	(48.15%, 78.45%)[k]	(7.41%, 2.76%)[k]	(1.85%, 1.1%)[k]	(12, 54)[k]	(3.1e-32, 2.7e-150)[k]	[45]

[k] KEGG
[r] Reactome
[w] WikiPathway
[s] SKBR3
[b] BT474;

https://doi.org/10.1371/journal.pone.0173331.t002

defined as direct or indirect previously, because the definitions of the terms *predicted* (based on *posterior probability* from Bayesian modeling), *direct*, and *indirect* were based on thresholds calculated from the distributions of corresponding values [see Methods]. Thus, the enrichment test with literature supported gene-dependencies [34] along with the evidences from the above literature survey confirm that our framework is able to identify significantly dysregulated signaling pathways that have key associations with acquired resistance in cancer.

Comparing with our previous study. To compare the performances of our current framework with our previous one [10], we investigated which of the two frameworks identify a greater number of dysregulated signaling pathways from KEGG, Reactome, and WikiPathway databases, since we used similar gene expression datasets (SKBR3 and BT474) in both approaches. We conducted a hypergeometric test to measure the statistical significance of the overlap between the aberrant pairs and known signaling links [34]. For that purpose, we defined the aberrant pairs in our previous approach [10] with $odds^P$ and $odds^R > 10.0$, and posterior probabilities, $Pr(u_{ij}^P = 1)$ and $Pr(u_{ij}^R = 1) > 0.5$. We found that greater percentages of

Table 3. Summary of predicted dysregulated signaling pathways from KEGG, Reactome and WikiPathway databases that plays a role as compensatory pathway of EGFR/HER2 inhibition in acquired resistance in both SKBR3 and BT474 cell-lines.

Aberrant Pathways in EGFR-TKIs Resistance^{k,r,w}	% of Direct Pair$^{(s,b)k,r,w}$	% of Indirect Pair$^{(s,b)k,r,w}$	% of PPI Pair$^{(s,b)k,r,w}$	# of Enriched Pair$^{(s,b)k,r,w}$	Enrichment q-value$^{(s,b)k,r,w}$	Literature References
Compensating Pathways of EGFR/HER2 inhibition						
Notch signaling	(40%, 75%)^{k}	(__, 25%)^{k}	—	(2, 3)^{k}	(8.6e-08, 1.7e-12)^{k}	[48–50]
	(46.15%, 71.43%)^{r}	(7.69%, 4.76%)^{r}	—	(5, 7)^{r}	(5.2e-17, 3.6e-23)^{r}	
	(35%, 70.37%)^{w}	(__, 7.41%)^{w}	—	(5, 7)^{w}	(3.2e-14, 3.2e-20)^{w}	
Wnt signaling	(25%, 50%)^{k}	(12.5%, 28.57%)^{k}	—	(6, 8)^{k}	(3.4e-16, 3.2e-25)^{k}	[51, 52]
	(21.88%, 66.67%)^{r}	(3.12%, 3.33%)^{r}	—	(2, 2)^{r}	(2.9e-04, 2.7e-04)^{r}	
	(25%, 55.56%)^{w}	(12.5%, 11.11%)^{w}	—	(3, 6)^{w}	(1.7e-19, 2.1e-10)^{w}	
Insulin Receptor/IGF1R signaling	(40%, 70.49%)^{k}	(13.33%, 9.84%)^{k}	—	(6, 25)^{k}	(7.9e-16, 3.1e-72)^{k}	[3, 10, 46, 47]
	(29.41%, 87.93%)^{r}	(5.88%, 3.45%)^{r}	—	(4, 30)^{r}	(1.3e-11, 6.6e-94)^{r}	
	(35.9%, 80%)^{w}	(12.82%, 9.33%)^{w}	(__, 1.33%)^{w}	(6, 28)^{w}	(4.1e-13, 2.9e-70)^{w}	
VEGFR signaling	(40%, 81.82%)^{k}	(__, 4.55%)^{k}	—	(1, 15)^{k}	(1.9e-03, 3.6e-54)^{k}	[3, 55]
FGFR signaling	(32.05%, 71.14%)^{r}	(8.97%, 6.97%)^{r}	(__, 0.5%)^{r}	(18, 72)^{r}	(7.6e-43, 1.7e-185)^{r}	[4, 56]
PDGFR signaling	(40.78%, 71.35%)^{r}	(3.88%, 3.78%)^{r}	(__, 0.54%)^{r}	(15, 67)^{r}	(2.2e-32, 1.9e-171)^{r}	[21]
Others						
Hippo signaling	(41.46%, 72.22%)^{k}	(12.2%, 11.11%)^{k}	—	(9, 4)^{k}	(1.0e-24, 6.7e-12)^{k}	[59]
TGF-β signaling	(22.22%, 100%)^{k}	(11.11%, __)^{k}	—	(4, 1)^{k}	(1.1e-13, 5.0e-04)^{k}	[10, 60]
	(50%, 50%)^{r}	(25%, __)^{r}	—	(2, 1)^{r}	(4.4e-07, 7.7e-04)^{r}	
	(54.55%, 30%)^{w}	(18.18%, 40%)^{w}	—	(5, 4)^{w}	(1.1e-16, 1.2e-13)^{w}	

^{k} KEGG
^{r} Reactome
^{w} WikiPathway
^{s} SKBR3
^{b} BT474;

https://doi.org/10.1371/journal.pone.0173331.t003

pathways from KEGG, Reactome, and WikiPathway databases were found as perturbed (dysregulated) in acquired resistance when the current approach was used compared to the old one [Fig 4].

One of the main differences between these two approaches was in the definitions of the set of edges in *GGR* network models: the current approach used *direct* pairs and *non-direct* pairs (*indirect* pairs and *PPI* pairs), whereas the old approach only used *direct* pairs [10]. Therefore, we conducted two experiments to investigate the importance of non-direct pairs in the new model. First, in aberrant signaling pathways that were detected by our current but not the previous model, we observed what percentages of *enriched links* (i.e. aberrant pairs found as known signaling links) were previously defined as non-direct (indirect and PPI) pairs in our current model. In both SKBR3 and BT474 cell-lines, we found that all such dysregulated

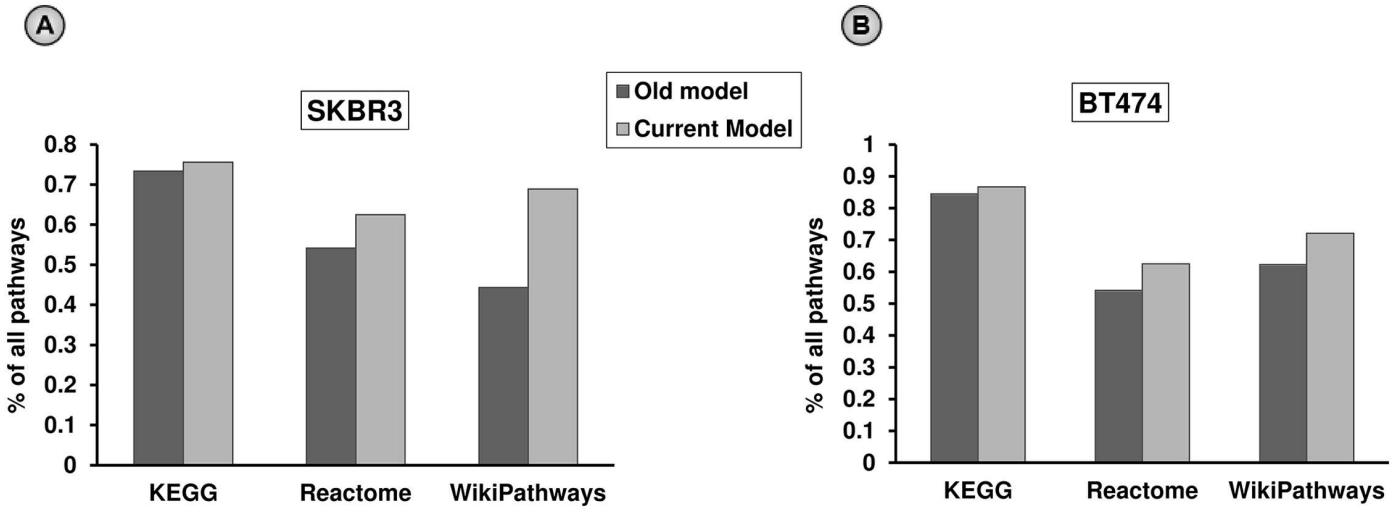


Fig 4. Performance comparison between the current model and our previous model [10] in terms of detecting perturbed signaling in acquired resistance. Percentages of signaling pathways detected as perturbed in acquired resistance by our *current* and *old models* in all KEGG, Reactome and WikiPathway databases: (A) in SKBR3, and (B) in BT474 cell-lines. For both the cell-lines, the performances using KEGG and Reactome pathways are comparable in both approaches, whereas our current model outperforms the old model for pathways from WikiPathway database.

https://doi.org/10.1371/journal.pone.0173331.g004

pathways from KEGG, Reactome, and WikiPathway databases contained high percentages of non-direct (indirect and PPI) *enriched links* [S3 Table]. Second, in aberrant signaling pathways that were detected by both of our current and previous models and were ranked (based on enrichment q-values) *high* in the current model but *low* in the previous model, we observed what percentages of *enriched links* were previously defined as non-direct in our current model. Considering the rank difference ≥ 10 (an empirical cutoff threshold), we found that aberrant pathways from KEGG, Reactome, and WikiPathway databases that showed such behavior in both SKBR3 and BT474 cell-lines, also contained high percentages of non-direct (indirect and PPI) *enriched links* [S4 Table]. Therefore, we claim that our current model demonstrate enhanced performances in detecting dysregulated signaling pathways in acquired resistance compared with our previous model.

Comparing with other methods. Next, we compared our framework with other published methods in terms of identifying the aberrant signaling pathways, specifically SPIA [61], DAVID [62], GATHER [63], ESEA [64] and PAGI [65]. The first three methods (i.e. SPIA, DAVID, and GATHER) are node-centric methods, where the role of differentially expressed (DE) genes was the key to identifying dysregulated pathways. However, ESEA and PAGI are edge-centric methods, where topological information regarding pathway structures was significantly exploited. All of these methods use GE datasets, except DAVID and GATHER which take a list of DE genes as input and identify aberrant pathways, or pathways enriched with given DE genes, respectively. For this comparative analysis, we used KEGG signaling pathways only, and for all the methods default configurations were applied unless specified otherwise.

The SPIA method combines classical enrichment analysis and actual aberrant activities by analysing Cancer-Vs-Normal GE samples [61], and ranks corresponding signaling pathways by calculating a global pathway significance *p*-value, called *pG*. The global *p*-value (*pG*) is obtained by combining the perturbation probability (*p*-value: *pPERT*) and the probability of over-representation of DE genes (using log fold-change) in a particular pathway (*p*-value:

pNDE) by using either Fisher's method or the normal inversion method [61]. Here, we conducted the same analysis but with Resistant-vs-Parental GE samples aiming to capture the aberrant activities responsible for acquired resistance. In the case of multi-probe sets for the same gene, we used the most significant probe to get a single *log2* fold-change value per gene. For SKBR3 cell-lines, we found 5 signaling pathways as significant (raw pG-value ≤ 0.05) including Ras signaling, PI3K-Akt signaling, Rap1 signaling, hippo signaling, thyroid hormone signaling, and TGF-β signaling pathways. Interestingly, we found that the significance ($-log(q\text{-}values)$) of aberrant pathways found by our approach is strongly correlated with the global *p-values* (pG) found by SPIA analysis, both for all pathways (-0.4) and for above 5 signaling pathways only (-0.928). This indicates, in SKBR3 cell-lines, the signaling pathways from our framework with high enrichment of aberrant gene-pairs in acquired resistance are also consistent with the results from SPIA in terms of identifying aberrant activities. Again, for BT474, we found 12 signaling pathways with significant aberration (raw pG-value ≤ 0.05), i.e. hippo, p53, Ras, Rap1, PI3K-Akt, FoxO, Wnt, neurotrophin, insulin, estrogen, ErbB, and MAPK signaling pathways. Moreover, among these 12 signaling pathways, the first 6 had FDR-corrected $pG \leq 0.05$, among which hippo signaling pathway had Bonferroni-corrected $pG \leq 0.05$, as shown in Fig 5A. Among these 12 significantly dysregulated pathways in BT474 cell-line, we chose FoxO signaling to investigate further, since it was found highly perturbed by both SPIA ($pPERT = 0.053$) and our methods (enrichment $q\text{-}value = 2.7 \times 10^{-150}$). We observed perturbation plots for this signaling pathway (KEGG pathway ID = 04068), in which perturbations of all genes were plotted as a function of their initial *log*2 fold-change Fig 5B. Here, non-DE genes were assigned 0 for their *log*2 fold-change. However, many genes were identified as DE, since their absolute *log*2 fold-change values were mostly ~ 2. Again, compared to the null distribution of net accumulated perturbation values, the observed value was also found significant as shown with the red vertical line in Fig 5B. Next, we also drew the network view of the FoxO signaling pathway, where the nodes were the constituent genes (from KEGG), and the edges were the known signaling links from the literature [34]. Here, we found 54 known signaling links that were also identified as aberrant gene-pairs by our method. Next, we plotted the heatmap of the expression values of the genes in these 54 known aberrant signaling links, where each expression value was the mean of all three replicates [66], z-transformed, and normalized with absolute max value (of the z-scores across the particular gene). Here, this heatmap not only shows the differential expression of the genes in aberrant gene-pairs but also indicates the similarities of their expression changes within this signaling pathway, which is a marker of aberrant activities in a modular way. Such differential gene expression in resistant-vs-parental conditions may indicate that pathway dysregulation within the signaling circuitry can be mediated by the corresponding aberrant gene-pairs.

As DAVID and GATHER both take as input a list of presumably differentially expressed genes for their pathway enrichments, we used the list of 703 and 683 distinct genes in the *list of aberrant gene-pairs* which were found by our framework from SKBR3 and BT474 cell lines, respectively. To detect statistically significant pathways using DAVID and GATHER we select those for which the raw *p-values* of their enrichment were < 0.05. For SKBR3 cell-line, DAVID and GATHER identified 15 and 5 signaling pathways as statistically significant, respectively. Again, for BT474 cell-lines they found 13 and 4 pathways as significant, respectively.

For both ESEA and PAGI analyses, we used our Resistant and Parental GE datasets for both SKBR3 and BT474 cell-lines. For both analyses, we used the default running parameters, except for the parameter *nperm* (the number of permutations) which was set to 1000. Both of these methods used a built-in set of topological structures of pathways from known pathway databases including KEGG. After running these methods with our GE datasets, if the identified

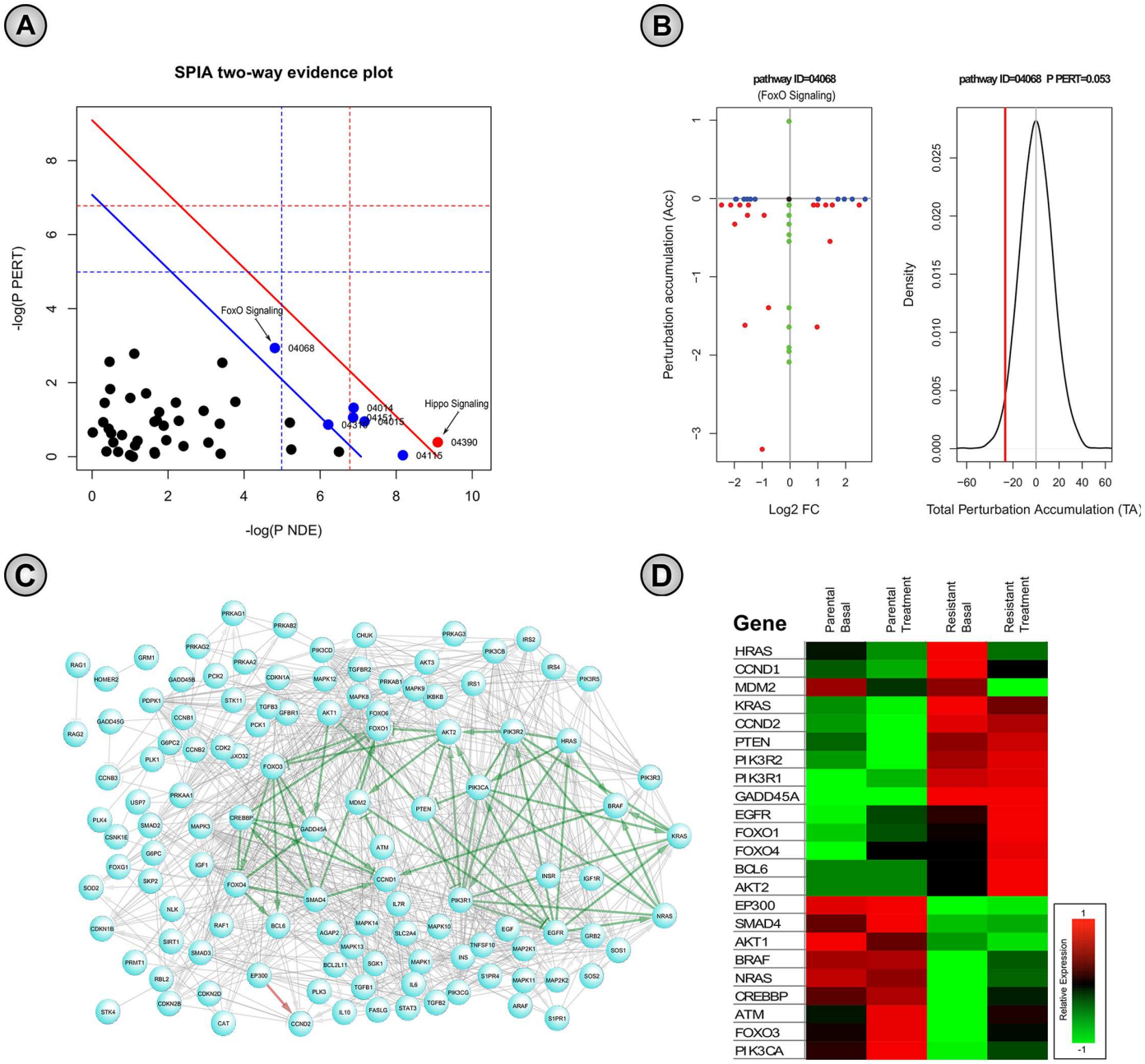


Fig 5. Detection of perturbed pathways with SPIA method. (A) Two-way evidence plot for all 45 KEGG pathways for BT474 cell-line is drawn. Here, pathways are represented with dots and the pathways with red dots and blue dots correspond to perturbed pathways with FDR-corrected and Bonferroni-corrected global *p*-value, $pG < 0.05$, respectively. (B) Next, the perturbation plot for FoxO signaling pathway (KEGG pathway ID = 04068) was also observed, since it contains the lowest perturbation *p*-value among all, $pPERT = 0.053$. In this plot, perturbation of all genes in the FoxO signaling pathway are shown as a function of their initial log2 fold-change (lower-left panel), where each dot indicates a gene in the pathway, and non-differentially expressed genes are assigned 0 as their log2 fold-change value. The null distribution and the observed net accumulated perturbation (red line) are shown in the lower-right panel. (C) Network view of FoxO signaling pathway for BT474 cell-line, where nodes are the constituent genes and the edges are known links collected from literature [34]. Here, *green* and *red* edges are the aberrant gene-pairs found by our method. (D) The heatmap of the genes' expression in aberrant gene-pairs found by our method in FoxO signaling network for BT474 cell-line.

https://doi.org/10.1371/journal.pone.0173331.g005

signaling pathways had nominal *p-value* < 0.05, then we considered them as significantly dysregulated in resistant-vs-parental conditions. Thus, in the SKBR3 cell-line, we found 4 and 15 significantly dysregulated signaling pathways by ESEA and PAGI methods, respectively. For the BT474 cell-line, we found 2 and 12 signaling pathways significantly dysregulated by ESEA and PAGI, respectively.

All the dysregulated KEGG signaling pathways identified by any of these six methods are listed in Table 4. Some pathways were found consistently as dysregulated in both SKBR3 and BT474 cell-lines, but none were common to all six methods. However, our method identifies 33 KEGG pathways in both SKBR3 and BT474 cell-lines among which 17 were also identified by at least one of the other five methods (including both node-centric and edge-centric methods), for example MAPK, insulin (in DAVID, GATHER, and PAGI), ErbB, Wnt, B-cell receptor, Neurotrophin (in DAVID and PAGI), p53 (in DAVID and ESEA), and Jak-Stat signaling (in DAVID and GATHER). Moreover, our method identifies some novel dysregulated pathways in both SKBR3 and BT474 cell-lines which were not detected by any other methods. These pathways include Hif-1, AMPK, TNF and calcium signaling, which were reported to be involved in lapatinib-resistance in ErbB2-positive breast cancer cell-lines [4, 19, 67, 68]. Thus, the comparative identification of dysregulated pathways in resistant-vs-parental conditions in both SKBR3 and BT474 cell-lines indicates that our method is not only comparable to others but also able to detect novel findings which were validated by literature evidence.

V-structures can explain the role of aberrant signaling in acquired resistance

The importance of V-structures. To investigate the potential of the putative aberrant gene-pairs to characterise acquired resistance, we hypothesized that *genes become dysregulated in acquired resistance because of the compensating effect of aberrant signaling that evolves in resistant-vs-parental conditions*. In the simplest cases, this will involve both *red* and *green* aberrant edges incident upon a particular dysregulated gene. To investigate this hypothesis, we identified all genes with at least two aberrant links to observe which of two possible architecture types are associated with a larger number of dysregulated genes: 1) both *red* and *green* aberrant edges incident upon a gene (forming V-structures—see Methods for the definition), or 2) only red or only green aberrant edges incident upon a gene. Next, we identified the dysregulated genes among these for which the following was true: a gene is over-/under-expressed (in any patient sample) in PT-vs-PB conditions, but respectively under-/over-expressed in both RB-vs-PB and RT-vs-PB conditions, where PB, PT, RB and RT stand for 'Parental Basal', 'Parental Treatment', 'Resistant Basal' and 'Resistant Treatment', respectively. The rationale for using only such combinations is as follows. Both expression datasets of SKBR3 (GSE38376) [19] and BT474 (GSE16179) [66] cell-lines contain *steady-state* measurements of signaling activities, for both parental and resistant conditions. Therefore, we hypothesized that the expression changes of dysregulated genes in PT-vs-PB conditions may indicate the sensitivity of Lapatinib drug (EGFR/HER2 dual inhibitor) in the parental (sensitive) conditions whereas the opposite changes in expressions in both RB-vs-PB and RT-vs-PB conditions may indicate two things: 1) the cell-line had already became resistant to the drug for which the tumorigenic phenotype of cancer cells relapsed in the resistant condition (RB-vs-PB), and 2) the resistance characteristics of the cell-line persisted even with further treatment with lapatinib (RT-vs-PB). For each comparison, we examined the *log*2 of fold-change values, and the treatment and basal doses were 1.0 μM and 0 μM, respectively. We found that, for both SKBR3 and BT474 cell-lines, higher percentages of dysregulated genes were identified with both *green* and *red* aberrant signaling links compared to those with only a single type of incident edge (either *red* or

Table 4. Comparative identification of pathway dysregulation in all 45 KEGG signaling pathways in resistant-vs-parental conditions in both SKBR3 and BT474 cell-lines. 'S' for SKBR3 cell-line, and 'B' for BT474 cell-line.

Pathway	SPIA	DAVID	GATHER	ESEA	PAGI	Our Method
MAPK signaling	B	SB	SB		SB	SB
Insulin signaling	B	SB	SB		SB	SB
ErbB signaling	B	SB			SB	SB
p53 signaling	B	SB		SB	B	SB
Wnt signaling	B	SB			SB	SB
Jak-Stat signaling		SB	SB			SB
B-cell receptor signaling		SB			SB	SB
Neorotrophin signaling	B	SB			SB	SB
Ras signaling	SB					SB
Rap1 signaling	SB					SB
Chemokine signaling		SB			S	SB
mTOR signaling		SB			B	SB
PI3K-Akt signaling	SB					SB
TGF-beta signaling	S	S			SB	SB
VEGF signaling		SB		S	S	SB
Hippo signaling	SB					SB
Fc epsilon RI signaling		SB				SB
Calcium signaling						SB
NF-kappa B signaling				S		SB
HIF-1 signaling						SB
FoxO signaling	B			S		SB
Phosphatidylinositol					SB	B
signaling system						
Sphingolipid signaling						SB
AMPK signaling						SB
Notch signaling		S			B	SB
Toll-like receptor signaling			SB		S	B
T-cell receptor signaling		B	S			SB
TNF signaling						SB
GnRH signaling					B	SB
Estrogen signaling	B					SB
Prolactin signaling						SB
Thyroid hormone signaling	S					SB
Oxytocin signaling						SB
Epithelial cell signaling in						B
Helicobacter pylori infection						
PPAR signaling					S	
cGMP-PKG signaling						B
cAMP signaling						B
Adrenergic signaling						B
in cardiomyocytes						
Hedgehog signaling			S			
signaling pathways regulating						SB
pluripotency of stem cells						
NOD-like receptor signaling					S	S
RIG-I-like receptor signaling					S	
Adipocytokine signaling				B		S
Glucagon signaling						B

https://doi.org/10.1371/journal.pone.0173331.t004

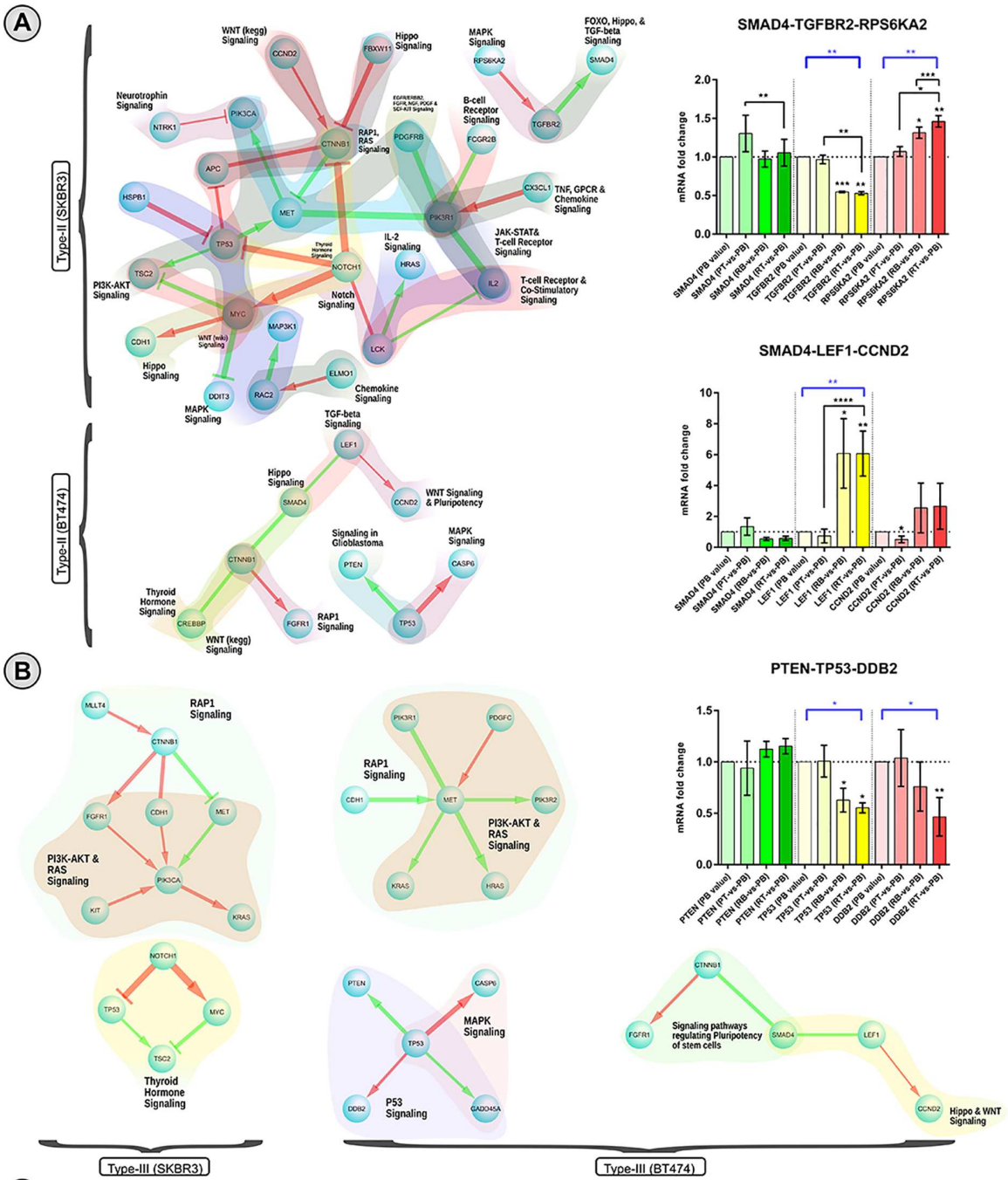
green). For SKBR3 and BT474 cell-lines we identified 111 and 108 genes with degree ≥ 2, respectively. For the SKBR3 cell-line, 90 of the 111 genes had only one type of aberrant signaling link incident upon them, out of which 48 showed dysregulation (53.3%), whereas the remaining 21 of the 111 genes had both *red* and *green* aberrant signaling links, out of which 13 genes were dysregulated (62%). Similarly, for BT474 cell-lines, among the 108 genes with degree ≥ 2, 78 out of 102 (76%) of genes with only one type of aberrant link and 6 out of 6 (100%) of genes with both types of aberrant signaling links, exhibited dysregulation. These results suggest that for a dysregulated gene in resistant-vs-parental conditions, the expression changes that occur upon treatment in parental conditions are likely to be compensated by aberrant signaling link(s) that evolved in resistant conditions. Therefore, the initial effect of inhibitors on oncogene(s)/tumor suppressor gene(s) becomes abrogated by restoring their tumorigenic phenotype once the cell acquires resistance to that inhibitor. This experiment demonstrates that V-structures can explain an interesting mechanism of acquired resistance in cell-lines by associating the dysregulated gene(s) with both *red* and *green* aberrant signaling links.

Type-II and Type-III V-structures provide a possible mechanism of gene dysregulation in acquired resistance. From the list of all putative aberrant gene-pairs (after Bayesian analysis), we enumerated all possible V-structures. We first listed all of the genes in *red* aberrant pairs, and separately listed all of the genes in *green* aberrant pairs. We then identified the genes common to both lists, which we termed *crossing-genes*. Next, we aggregated aberrant gene-pairs incident upon *crossing-genes* and enumerated all possible pairs of a *red* and *green* edge incident upon that gene. Thus, we found 23,156 distinct Type-I V-structures [see Methods for Type-I, Type-II and Type-III V-structure definitions] in SKBR3 cell-lines using signaling pathways from KEGG, Reactome, and WikiPathway, out of which 53 V-structures were found in the literature-curated signaling network [34]. Similarly for BT474, there were 5,271 distinct Type-I V-structures in all KEGG, Reactome, and WikiPathway signaling pathways, and 11 of them overlapped with the literature-curated network [34]. For Type-II V-structures in SKBR3 and BT474 cell-lines, 1,525 and 263 distinct V-structures were found in all KEGG, Reactome and WikiPathway databases, respectively, out of which 29 and 4 V-structures were found in the literature-curated network [34], respectively. For Type-III V-structures in SKBR3 and BT474 cell-lines, 940 and 376 distinct V-structures were found in all KEGG, Reactome, and WikiPathway databases, respectively, where 18 and 10 V-structures overlapped with the literature-curated signaling network [34]. A summary of these results for SKBR3 and BT474 cell-lines is provided in S5 and S6 Tables, respectively. Note that Type-I and Type-II V-structures have the potential to explain the role of signaling cross-talks in acquired resistance, but here we focus on Type-II V-structures only, since we have already investigated the role of signaling cross-talks in acquired resistance in our previous study [10] which are the similar kind of Type-I V-structures.

We investigated whether Type-II and Type-III V-structures can provide insights of a possible mechanism of acquired resistance in cancer cell-lines, focusing on the dysregulations of the *crossing-genes* in resistant-vs-parental conditions and its association with the GE changes of the other two genes in a particular V-structure. Our rationale was that the dysregulation of a *crossing-gene* may provide an indication that significant changes evolved in resistant-vs-parental conditions are associated with acquired resistance of cell-lines to a particular inhibitor. Moreover, significant GE changes in either of the two other genes (in the V-structure) would indicate that their differential associations with crossing-gene(s) may disrupt their functional coherence in signaling activities [30]. Therefore, we considered the above-mentioned 13 and 6 dysregulated genes in SKBR3 and BT474, respectively, for further analyses in which gene-pairs in corresponding V-structures overlapped with known signaling links [34]. Among the 13

dysregulated genes in SKBR3 cell-lines, 8 genes (*CTNNB*1, *TP*53, *MYC*, *RAC*2, *LCK*, *PIK*3*R*1, *PIK*3*CA*, and *TGFBR*2) were found in 22 (out of 29) literature-supported Type-II V-structures and 4 genes (*CTNNB*1, *TP*53, *MYC*, and *PIK*3*CA*) were found in 9 (out of 18) literature-supported Type-III V-structures (S5 Table). Similarly, among 6 dysregulated genes in BT474 cell-lines, 3 genes (*CTNNB*1, *LEF*1, and *TP*53) were found in 4 (out of 4) literature-supported Type-II V-structures and 4 genes (*MET*, *TP*53, *CTNNB*1, and *LEF*1) were found in 10 (out of 10) literature-supported Type-III V-structures (S6 Table). In Fig 6A, we show the network-view of the literature-supported Type-II V-structures incident upon the 8 and 4 dysregulated genes in SKBR3 and BT474 cell-lines, respectively, along with their annotated signaling pathways. Similarly, Fig 6B shows the Type-III literature-supported V-structures in both SKBR3 and BT474 cell-lines. Next, for each of the genes in the selected V-structures in Fig 6 we observed gene expression differences among all four conditions: PB (Parental Basal: 0 μM), PT (Parental Treatment: 1.0 μM), RB (Resistant Basal: 0 μM), and RT (Resistant Treatment: 1 μM) using both two-tailed paired t-tests and one-way ANOVA tests. For these statistical tests we used the mean expression value of all three replicates. In the t-tests, we compared the mean expression of all PT, RB and RT conditions with the mean of PB. Additionally, we also compared the mean of the RT condition with the means of the PT and RB conditions to observe 1) how a gene is behaving differently upon treatment in resistant-vs-parental conditions (RT-vs-PT), and 2) its expression changes upon treatment from its Resistant basal condition (RT-vs-RB). Moreover, one-way ANOVA tests (with the mean of PB as the control condition for the multiple comparison test) may indicate the significance of overall changes in all four groups. All of these statistical tests were done using GraphPad Prism 6.0 software. Concurrently, we also surveyed the literature to determine whether the observed significance of expression changes in resistant-vs-parental conditions were also supported by the literature. We found literature evidence (Fig 6C) supporting a role in breast cancer metastasis and/or in developing acquired resistance to EGFR-TKIs for the *SMAD*4 – *TGFBR*2 – *RPS6KA*2 (Type-II) V-structure in SKBR3, and *SMAD*4 – *LEF*1 – *CCND*2 (Type-II) and *PTEN* – *TP*53 – *DDB*2 (Type-III) V-structures in BT474 cell lines, respectively. Below we discuss these three V-structures in more detail.

- *SMAD*4 – *TGFBR*2 – *RPS6KA*2 (in SKBR3): *TGFBR*2 encodes a transmembrane protein which has been reported as a potent inhibitor of tumor growth and proliferation in breast epithelial cells, and loss of its function has also been associated with tumor malignancies [69]. Moreover, mRNA expression of *TGFBR*2 was reported to be significantly down-regulated in many tumorigenic cell-lines including SKBR3 and BT474 compared to the non-tumorigenic MCF-10F cell-lines [69]. This indicates the tumor-suppressing role of the *TGFBR*2 gene, and the reduction of its mRNA level may confer a resistance to targeted inhibitors by relapsing tumor growth and proliferation. In the GE dataset for the SKBR3 cell-line, the *TGFBR*2 gene was down-regulated in PT-vs-PB conditions without significance, but in resistant conditions it showed significant down-regulation compared to parental conditions (RB-vs-PB: *p-value* = 0.0003; RT-vs-PB: *p-value* = 0.002; RT-vs-PT: *p-value* = 0.001). A one-way ANOVA test also found the overall GE changes to be significant: Sidak corrected *p-value* = 0.0021. Thus, both literature evidence and GE data suggest an association of mRNA down-regulation of *TGFBR*2 gene with lapatinib resistance in SKBR3 cell-lines.
  *RPS6KA*2 (*RSK*3) encodes one of the members of the ribosomal S6 kinase which mediates resistance to PI3K pathway inhibitors in breast cancer [70]. RTK (Receptor Tyrosine Kinase) signaling induces the Ras and PI3K pathways, but upon lapatinib treatment such RTK signaling pathways are disrupted, downstream effectors (e.g. mTOR) are abrogated, and eventually Ras and PI3K signaling become inhibited [20]. Over-expression of *RSK*3 attenuates



1VSs	Green Pair	Odds^{P}	Posteriori^{P}	Ref^{P}	Red Pair	Odds^{R}	Posteriori^{R}	Ref^{R}
SMAD4 - TGFBR2 - RPS6KA2	SMAD4 - TGFBR2	2.63	0.504	Hu *et al.* Wang *et al.*	TGFBR2 - RPS6KA2	2.73	0.356	Hu *et al.* Wang *et al.*
SMAD4 - LEF1 - CCND2	SMAD4 - LEF1	2.48	0.338	Qin *et al.*	LEF1 - CCND2	13.26	0.206	Kanehisa *et al.*
PTEN - TP53 - DDB2	PTEN - TP53	3.12	0.466	Nieto *et al.* Wee *et al.* Chen *et al.*	TP53 - DDB2	14.19	0.332	Sun *et al.* Carson *et al.*

Fig 6. The role of literature-supported Type-II and Type-III V-structures (*VSs*) in explaining gene dysregulation in acquired resistance. (A) Network views of Type-II *VSs* along with their pathway annotations in SKBR3 and BT474 cell-lines. (B) Network views of Type-III *VSs* in SKBR3 and BT474 cell-lines. Note that *VSs* shown here are only those for which the crossing-genes were found as up- or down-regulated in PT-vs-PB conditions, but oppositely regulated in both RB-vs-PB and RT-vs-PB conditions. Nodes are genes, and the edges are known signaling links [34] that were also found as aberrant gene-pairs identified by our framework. Note that the width of edges is proportional to the posterior probability of corresponding pairs. Furthermore, for three *VSs* shown in (A) and (B) (right panels), mRNA changes for their constituent genes were found in the literature, implicating their role in breast cancer metastasis and/or in developing acquired resistance in EGFR-TKIs. (C) Above three *VSs* with their corresponding posterior probabilities, odds, and literature references of gene-pair associations for each of the *red* and *green* pairs. Statistical significance tests were done using t-tests and one-way ANOVA with multiple corrections (Sidak method). All the mRNA values were normalized by corresponding PB expression values in all three replicates. Significance was indicated by * (*p-value* < 0.05), ** (*p-value* < 0.005), and so on.

https://doi.org/10.1371/journal.pone.0173331.g006

the apoptotic response and up-regulates protein translation, and thus promotes cell survival and proliferation under conditions of PI3K/mTOR blockade [70]. Moreover, lapatinib down-regulates the Akt pathway in both SKBR3 and BT474 cell-lines [71]. We observed significant and consistent over-expression of *RSK*3 mRNA in resistant condition compared to parental conditions in our SKBR3 cell-line dataset (RB-vs-PB: *p-value* = 0.011; RT-vs-PB: *p-value* = 0.0046; RT-vs-PT: *p-value* = 0.011; RT-vs-RB: *p-value* = 0.003). Overall expression changes were also found significant: Sidak corrected *p-value* = 0.0011. Therefore, both literature evidence and our experimental data strongly suggest that *RSK*3 over-expression is associated with lapatinib resistance via a PI3K/mTOR signaling blockade.

*SMAD*4 is a downstream mediator of *TGF-β* [72] which plays an important role both in tumor suppression and progression in breast cancer [72, 73]. Liu *et al.* reported that *SMAD*4 expression was decreased in breast cancer cells compared to adjacent normal breast epithelial tissue [72]. Moreover, *SMAD*4 is sensitive to lapatinib according to the COSMIC database [74] with no mutational signature in breast cancer cell-lines. In our GE dataset of SKBR3 cell-lines, *SMAD*4 expression was up-regulated in PT-vs-PB, but was down-regulated in the RB-vs-PB condition, and again up-regulated in the RT-vs-PB condition. Note that however, that none of these comparisons were statistically significant in t-tests at the 0.05 level, and the one-way ANOVA also did not detect significant differences (Sidak corrected *p-value* = 0.101). Interestingly, both *SMAD*4 and *TGFBR*2 mRNA expression changes in PT-vs-PB conditions were non-significant; however, in resistant conditions (RB and RT) both *TGFBR*2 and *RPS6KA*2 showed significant changes in mRNA level compared to parental conditions (PB and PT). This may indicate the dependency switch of *TGFBR*2 from *SMAD*4 to *RPS6KA*2 in resistant-vs-parental conditions.

*TGFBR*2 phosphorylates *SMAD*4 in the TGF-*β* signaling [34, 75], and both of their mRNA changes in parental conditions (PT-vs-PB) were non-significant. However, *TGFBR*2 is an upstream kinase that phosphorylates *RPS6KA*2 [34, 75], and both of their mRNA changes in resistant conditions were very significant compared to parental conditions. Thus, we hypothesize that the gene dysregulation of *TGFBR*2 in acquired resistance can be explained by its significant association with *RPS6KA*2 which evolved in resistant conditions compared to parental conditions.

- *SMAD*4 – *LEF*1 – *CCND*2 (in BT474): *LEF*1 plays an oncogenic role in breast cancer, since both mRNA and protein expression of this gene were found to be higher in breast cancer cell-lines compared to normal cells [76]. A high level of *LEF*1 was also found in *HER*2 expressing BT474 cell-lines [77], where *HER*2-activated *β-catenin* plays a crucial role in producing an increase in the downstream target *LEF*1 [76]. Increased expression of *LEF*1 drives cells towards resistance to TGF-*β*-induced growth inhibitory activities [78]. In our GE

datasets of BT474 cell-lines, *LEF*1 mRNA expressions were significantly increased in resistant conditions compared to the parental basal condition (RB-vs-PB: *p-value* = 0.0178; RT-vs-PB: *p-value* = 0.003). Interestingly, over-expression of *LEF*1 was even more significant in resistant-vs-parental conditions in the presence of lapatinib (RT-vs-PT: *p-value* < 0.0001). Moreover, overall expression changes were also proved to be significant by one-way ANOVA test (Sidak corrected *p-value* = 0.004). Thus, the experimental data and the literature evidences support a role of *LEF*1 gene in lapatinib resistance in the BT474 cell-lines.
*CCND*2 is involved in the cell cycle process, and is a regulatory subunit of a complex formed with *CDK*4 or *CDK*6 that is required for cell cycle G1/S transition [79]. Although *CCND*2 over-expression is found in ovarian, testicular [79] and gastric cancer [80], little is known about its role in breast cancer especially in the presence of lapatinib. In the GE data for the BT474 cell-line, *CCND*2 mRNA expression was significantly down-regulated in the PT-vs-PB condition (*p-value* = 0.024), and this possibly indicates the association of its mRNA down-regulation with lapatinib sensitivity in lapatinib-sensitive BT474 cell-lines. We investigated whether this behaviour is coherent with the literature. Schmidt *et al.* reported that both mRNA and protein expression of *CCND*1 and *CCND*2 were down-regulated when *FOXO3A* induced the process of cell cycle arrest [81]. Such inhibition of *CCND*1 and *CCND*2 perturbs *CDK*4 functionality to inactivate the S-phase repressor Rb [81]. Moreover, Hegde *et al.* reports that mRNA expression of *FOXO*3 and *CCND*1 were significantly up- and down-regulated, respectively, in both SKBR3 and BT474 cell-lines (lapatinib-sensitive) in response to lapatinib treatment [71]. To explain the above-mentioned down-regulation of *CCND*2, we observed *FOXO*3, *CCND*1 and *RB*1 mRNA changes in PT-vs-PB conditions (in BT474 datasets), to determine whether these are coherent with the above literature findings. In SKBR3 cell-lines, *FOXO*3 was significantly up-regulated (*p-value* = 0.0028) and *CCND*1 was significantly down-regulated (*p-value* = 0.0029). In BT474 cell-lines, 2 out of 3 replicates showed a similar pattern of mRNA changes for these two genes (*FOXO*3 and *CCND*1) (*p-values* = 0.042 and 0.017, respectively) as in SKBR3 cell-lines. In BT474 cell-lines *RB*1 mRNA expression was found slightly up-regulated in PT-vs-PB conditions. Moreover, *CCND*2 mRNA expressions are up-regulated in both resistant conditions (RB-vs-PB and RT-vs-PB) compared to the parental basal condition. The above experimental data may indicate the possible reason for *CCND*2 down-regulation in lapatinib-sensitive BT474 cell-lines with lapatinib treatment, and its mRNA up-regulation in both resistant conditions (RB-vs-PB and RT-vs-PB) could possibly be due to acquired resistance of BT474 cell-lines to lapatinib.
*SMAD*4 expression was reported to be decreased in breast cancer cells [72], and the COSMIC database [74] reports *SMAD*4 as sensitive to lapatinib in the BT474 cell-line along with other EGFR-TKI, BIBW2992 and erlotinib [74] with IC_{50} *effect* = 0.225 (*p*-value = 0.000014) and with significant mutational signature in skin cancer, but none in breast cancer cell-lines. However, in the GE data for the BT474 cell-line, mRNA expression of *SMAD*4 was up-regulated in PT-vs-PB conditions, but was down-regulated in resistant-vs-parental conditions, with or without lapatinib treatment (RB-vs-PB and RT-vs-PT), indicating its sensitivity to lapatinib in parental conditions. Note that we observed no significant changes using a one-way ANOVA test (Sidak corrected *p-value* = 0.1212).
SMAD4 binds to LEF1 [82], and the changes in expression of both of their mRNAs indicate sensitivity to lapatinib treatment in parental conditions (PT-vs-PB). Again, LEF1 regulates the transcription of *CCND*2 gene in the Wnt signaling pathway [83], and both genes exhibited up-regulation in resistant conditions compared to parental conditions. Thus, we can hypothesize that the dysregulation of the *LEF*1 gene can be explained by its differential associations with SMAD4 and CCND2 mRNA changes in resistant-vs-parental conditions.

- *PTEN* – *TP*53 – *DDB*2 (in BT474): *PTEN* is one of the most commonly mutated tumor suppressor genes, and the loss of its mRNA and protein expression are found in many metastatic malignancies including breast cancer [84]. *PTEN* modulates lapatinib sensitivity [85], and its loss acts as a marker of poor lapatinib response [58, 86, 87]. In the GE dataset for the BT474 cell-line, no mutation has been detected for *PTEN* and *TP*53 in their corresponding DNA sequences between parental and resistant conditions as reported in the original article associated with this dataset [66], and *PTEN* expression was up-regulated even in resistant-vs-parental conditions with or without lapatinib (RB-vs-PB, and RT-vs-PB), but the overall mRNA changes were not significant as tested with the one-way ANOVA test (*p-value* = 0.264). *TP*53 is another well known tumor suppressor gene, and its inhibition greatly inhibits apoptosis as p53 up-regulates several pro-apoptotic gene products including Puma, Noxa, Apaf-1, and Bax [88]. The loss of p53 is consistently associated with the acquired resistance of EGFR inhibitors cetuximab and erlotinib [89]. However, more experimental evidence is required to claim that p53 loss can be a predictive feature of acquired resistance to EGFR inhibitors [90]. In the GE dataset for the BT7474 cell-line, *TP*53 expression was significantly decreased in both RB-vs-PB (*p-value* = 0.013) and RT-vs-PB (*p-value* = 0.025) conditions, and the overall changes were statistically significant (Sidak corrected *p-value* = 0.01). For the *DDB*2 gene, its under-expression is correlated with poor outcome in ovarian cancer [91]. In breast cancer, although *DDB*2 showed putative oncogenic behaviour by promoting cell-cycle progression [92], it was not over-expressed in ER-negative breast cancer cells [92, 93], e.g. SKBR3 [93]. Moreover, *DDB*2 is down-regulated in lapatinib-resistant cell-lines [94]. This suppression was induced by the over-expression of the hepatitis B viral-encoded X protein (HBX) in the p53/lincRNA-p21 axis and IKK-dependent manner [94]. In our GE dataset for the BT474 cell-line, *DDB*2 was significantly down-regulated in resistant-vs-parental conditions (RT-vs-PB: *p-value* = 0.002) and the over-all changes were significant as well (Sidak corrected *p-value* = 0.046).
  p53 up-regulates or enhances *PTEN* transcription [95–97], and we found both genes' mRNA changes in parental conditions (PT-vs-PB) to be non-significant. Moreover, p53 transcriptionally regulates *DDB*2 expression in a cell cycle-dependant manner [98, 99], and both of their mRNA changes were found to be significant, showing similar phenotypes in resistant-vs-parental conditions. Thus we can claim that the switch in dependency of *TP*53 from *PTEN* to *DDB*2 (in *PTEN* – *TP*53 – *DDB*2) can be a possible mechanism of *TP*53 dysregulation in acquired resistance.

Gene dysregulation plays an important role in developing acquired resistance to EGFR-TKIs in breast cancer [28–30, 100]. Here, along with literature-supported gene-gene associations in V-structures (Fig 6C), we demonstrated that the switch in dependency from the *targeted* signaling link involving *green* gene-pair (with the inhibitor) to the *bypass* signaling link involving *red* gene-pair (evolved in resistant conditions) is a possible mechanism mediating the dysregulation of *crossing-genes* in acquired resistance.

Discussion

In this study, we proposed a computational framework that models signal rewiring by systematically characterizing potential aberrant signaling in acquired resistance. We hypothesized that an aberrant signaling link involved in acquired resistance may have differential probabilities of appearing (either higher, or lower) in resistant-vs-parental networks, where in each network, nodes were genes and the edges were the relationships among genes. In this gene-gene relationship network, called *GGR*, we considered both direct and indirect correlations (via *linker* genes) among genes for defining the edges that combine both data-driven (from gene

expression) and topological (from PPI) information about gene-pairs. Note that the PPI edges in the statistically significant paths [see Methods], defining indirect relationships among gene-pairs for which direct relationships were not found, were also added to the final edge set [Table 1]. The rationale for including those PPI edges was: 1) to retain precise information regarding how indirect relationships were constructed, and 2) to better model the data-driven signaling networks (resistant and parental GGR networks) for the Bayesian statistical analysis (using p_1-model) of their respective global structure formation. We used a fully Bayesian statistical model: a special class of Exponential Random Graph Model, called p_1-model for inferring aberrant gene-pairs with differential posterior probabilities in resistant-vs-parental *GGR* networks, where these networks were constructed from matched gene expression values of resistant and parental conditions, respectively. When selecting aberrant gene-pairs, we chose the thresholds for *Odds* and posterior probabilities from their frequency distributions, sequentially. Firstly, we chose the gene-pairs with top 20% of odd-ratio values from two distribution individually (odd^R and odd^P) by ensuring their mutual exclusivity after selection, and termed them as *red* and *green*, respectively. Then, we further filter *red* and *green* pairs with top 50% of their respective posterior probability values. Note that before calculating the *Odds* values, we normalized both posterior probabilities (from resistant and parental conditions) with their corresponding *max* values over all gene-pairs, individually, in order to achieve same scaling. All other model parameters in p_1-model were estimated with Gibbs sampling [see Methods].

After detecting putative aberrant pairs in resistant-vs-parental conditions, we analyzed them in two-ways, 1) Identifying potentially dysregulated pathways in acquired resistance, 2) Identifying their roles in explaining a possible mechanism of acquired resistance via dysregulation of crucial genes. Using two lapatinib-treated breast cancer cell-lines: SKBR3 and BT474, our method was able to predict similar pathways as dysregulated. The rationale for using these datasets for our experiments was that—to the best of our knowledge—these are only datasets available for responsive and resistant lapatinib-treated ERBB2-positive breast cancer cell-lines. Our results suggested that signal rewiring is a major event in acquired resistance since we found a range of dysregulated pathways in both SKBR3 and BT474 cell-lines including EGFR-related pathways (e.g. EGFR, ErbB2, PI3K-Akt, Mapk, Jak-Stat, FoxO signaling, etc.) as well as other receptor-related pathways (e.g. Notch, Wnt, insulin, PDGFR, FGFR, VEGFR signaling, etc). Although there may be some false-positives in those results, we found literature evidence from Huang *et al.* [3] that aberrant signaling in most of our predicted dysregulated pathways were actually related with acquired resistance in EGFR-TKIs. Furthermore, our predictions of network re-adjustment in multiple signaling pathways were also consistent with the results recently published by Stuhlmiller *et al.* [5]. Their study suggested that multiple heterogenous kinases (e.g. DDR1, FGFRs, IGFI1, MET, etc.) compensate for the ErbB2 inhibition by kinome re-programming induced by lapatinib [5], which provides an indication that aberrant signaling activities in those kinase-related pathways are crucial for such bypass mechanism. Note that since the pathway annotations are still incomplete, we used three pathway databases here: KEGG, Reactome, and WikiPathways to define constituent genes of signaling pathways individually. However, to maintain the same true-relationship among those constituent genes we used literature-supported signaling links (collected from online resources of Wang Lab [34]) since it is the largest manually curated human signaling network as reported.

Gene dysregulation plays crucial roles in acquired resistance by mediating both uncontrolled cell-growth and disrupted apoptosis [27–29]. Here, to evaluate the potentialities of identified aberrant signaling, we conducted an analysis which demonstrated that the greater number of dysregulated genes were found in resistant-vs-parental conditions when they were incident with both *red* and *green*-types of aberrant pairs (V-structures) compared to those with single type only (either *red*, or *green*). Manual literature survey also validated some of the V-

structures, such as *SMAD*4 – *TGFBR*2 – *RPS6KA*2, *SMAD*4 – *LEF*1 – *CCND*2, and *PTEN* – *TP*53 – *DDB*2, as consistent with our hypothesis. Thus, we claim that a mechanism of dependency shift from *targeted signaling* (by inhibitor) towards *bypass signaling* can potentially cause dysregulation of shared genes (crossing-genes). Similar idea of dependency switch was recently reported by Sharifnia et al. [30] that EGFR-dependent status of downstream signaling nodes can be modified by other over-expressed kinase-related genes that shared them (downstream signaling nodes) with EGFR-dependant signaling. However, to the best of our knowledge, our study is the first to emphasise the compensating effects of aberrant signaling upon mRNA expression changes of crucial genes by examining the dependency switch from *targeted* signaling to *bypass* signaling.

We included all the available genes from the Cancer Gene Census (CGC) into the list of seed genes in our framework for which gene expression data was available (see Methods): 370 and 357 genes in SKBR3 and BT474 cell-lines, respectively. Cancer genes are crucial for mediating various cancer related activities and many are hub genes in mammalian signaling networks [101]. Therefore, they are very important in terms of signaling network formation, an aspect which we examine in this study by statistical models (i.e. p_1-model). Note that we combined cancer genes with a set of differentially expressed (DE) genes even though some may not be differentially expressed. However, cancer genes can still be important in network-based analyses of studies comparing two conditions (i.e. resistant-vs-parental). For example, in a network-based classification of breast cancer patients, Chuang *et al.* [102] reported that the subnetworks which can classify metastatic and non-metastatic patients contain genes playing a central role connecting DE genes even though those cancer genes were non-DE themselves [103]. Moreover, we intend to include all CGC genes, not just those ones that are breast cancer related, since no classifications are perfect, and the census is continuously being updated [104]. CGC genes are selected based on the mutational profiles of cancer patients [105], hence their transcriptional profiles may also reveal additional insights into the mechanisms of aberrant signaling activities in acquired resistance. To investigate the influence of CGC genes in our framework, we observed all the genes involved in all the V-Structures (VSs) of aberrant pairs (Type-I, Type-II and Type-III VSs) found in pathways from KEGG, Reactome and WikiPathway databases [See S5 Table]. We found that many of the genes involved in VSs overlapped with genes from CGC, where most of those cancer genes were not identified as DE genes during the formation of the seed gene list [see Methods] [S7 Table]. Thus, we claim that CGC genes were very important in the network-based analyses of our framework.

In this paper, we considered only gene expression values for modeling gene-gene relationship networks (*GGR*). However, we look forward to adapting other appropriate high-throughput datasets, such as miRNA expression, methylation, copy number aberration, and phosphorylation datasets into our framework in order to better model gene-gene dependencies in resistant-vs-parental conditions to reflect greater mechanistic insights. Moreover, the V-structures we have examined in our current study can be called *first-order V-Structures* since they involve only a single aberrant edge of each type (*red* and *green*). In future we intend to examine the role of higher order V-structures in acquired resistance.

Materials and methods

Literature and database search

Our research hypothesis was primarily focused on studying the acquired resistance mechanisms of HER2-positive breast cancer cells to lapatinib (an EGFR/HER2 dual inhibitor). Therefore, we conducted a literature survey in Pubmed database using keywords: 'lapatinib', 'acquired resistance', and 'breast cancer', which lead us to find two articles: Komurov *et al.*

[19] and Liu *et al.* [66]. Both of these articles studied the resistance mechanisms of HER2-positive breast cancer cell-lines by analysing gene expression datasets of lapatinib-treated sensitive (parental) and resistant conditions. To find these gene expression datasets, we also searched GEO (Gene Expression Omnibus) database with the same keywords as above and found two data collections with accession IDs: GSE38376 and GSE16179, respectively. Detailed technical descriptions of cell-line preparation and dataset generation were reported in their respective original articles. The first dataset (GSE38376) included SKBR3 parental and resistant (SKBR3-R) cell-lines, and the second dataset (GSE16179) included BT474 parental and resistant (BT474-J4) cell-lines. In both of these datasets, expression values of both parental and resistant samples were measured first in basal condition (0 μM), and then in treatment conditions (0.1 μM and 1.0 μM for GSE38376; 1.0 μM only for GSE16179). For both GSE38376 (SKBR3) and GSE16179 (BT474), we converted probe-level expression values into gene-level values using the corresponding annotation files: GPL6947 (Illumina HumanHT-12 V3.0 expression beadchip) and GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array), respectively, which were also collected from GEO database. For some genes, multiple probes were mapped to a single gene, and we averaged the GE values of such probes to obtain the final GE values of the corresponding gene. Next, for each collection (GSE38376 and GSE16179) we built two data matrices, one from the parental and another from the resistant GE dataset, where rows were labeled with gene symbols and columns were labeled with samples under different treatment conditions. A protein-protein interaction dataset was obtained from Cerami *et al.* [106]. For the enrichment analysis, we collected gene sets of all 1) the 24 signaling pathways from Reactome [107] (downloaded at 19/05/2014), 2) 45 signaling pathways from KEGG [83, 108] (downloaded at 12/05/2015), and 3) 61 signaling pathways from WikiPathway [109] (downloaded at 16/10/2014) databases. Each signaling pathway downloaded from these databases was encoded as tab-delimited lists of gene symbols. For KEGG signaling pathways, we built a parser program that extracted gene names from the web-responses after making HTTP web-requests to KEGG server using a list of IDs corresponding to signaling pathways.

Constructing gene-gene relationship network

We denote the gene-gene relationship network as $GGR := (S, R)$ for each GE data matrix. Here, S is the set of seed genes, which is the union of a set of differentially expressed (DE) genes, a set of cancer genes collected from the Cancer Gene Census (CGC) [105], and a set of *linker genes* (see below) selected from the PPI network. R is the set of edges defined among the genes in the set S. The sets S and R were constructed as follows.

Defining *S*: The seed genes. We built the set S cumulatively; first a set of DE genes was identified by differential expression analysis of parental and resistant GE data using a two-tailed pooled Student's t-test. For this test, significant p-values were identified using the Bonferroni correction method, and genes with such corrected p-values $\leq$ *threshold* (see Results) were selected as differentially expressed. Next, we added CGC genes for which corresponding GE data was available. The rationale for such inclusion is that CGC genes are well known to be hub genes in mammalian cellular signaling networks [101] where they play key regulatory roles in various cancer related activities. In the process of finding indirect relationships among ($DE \cup CGC$) genes, a set of intermediate genes from the PPI network was identified, which we defined as *linker genes* (see next section). The final set of seed genes consisted of ($DE \cup CGC \cup Linker$) genes.

Defining *R*: The edges. To identify interacting gene pairs, all pair-wise absolute Pearson Correlation Coefficients (PCCs) were calculated for expression levels of the genes in the ($DE \cup CGC$) gene set. The value demarcating the top 20% of absolute PCCs was selected as the

threshold for defining *direct* relationships among the genes in the above set. That is, for each gene pair ($gene_i$, $gene_j$), if the corresponding PCC value was above the threshold then the pair was considered to have a *direct* relationship, and hence added into the set of edges, *R*.

Otherwise, a gene pair was said to have an *indirect* relationship if there was at least one statistically significant simple path in the PPI network between $gene_i$ and $gene_j$ via an intermediate gene (called a *linker* gene). Here, we imposed a path-length threshold of 2 and restricted to paths in the PPI network, otherwise considering all the remaining genes as possible intermediates would convert this searching procedure into an NP-hard problem. Simple paths of length 2 [for details see S1 Text] connecting a given pair ($gene_i$, $gene_j$) in the PPI network were considered statistically significant if one can reject the following null hypothesis: *the geometric mean of pairwise PCC values of constituent edges in the path is distributed as for paths of length 2 between these genes generated by a randomized procedure*. Random paths of the form $gene_i \rightarrow linker \rightarrow gene_j$ were generated by replacing *linker* with any other gene in the network except $gene_i$, $gene_j$ and any gene on a path of length 2 connecting these genes in the PPI network. To evaluate the PCC for a random path, we used the same expression values for the genes as in the observed case. Paths were considered significant if the probability of generating a path using above randomized procedure with a geometric mean of pairwise PCC values greater than or equal to that observed for the PPI network was ≤ 0.05 (an empirical p-value). PPI edges comprising statistically significant simple paths were added to the set *R*. The set of edges *R* was finally composed of direct relationships, indirect relationships, and PPI edges of statistically significant simple paths, which are used for identifying those indirect relationships [see Discussion].

Bayesian statistical modeling of *GGR* network

Exponential Random Graph Models (ERGMs) are parametric probability distributions over spaces of networks [24] that have been successfully used to evaluate probabilities of the presence of each edge in a network [23, 24]. Here, in order to infer edge probabilities in a gene-gene relationship network, we used the p_1-model, a special class of ERGM introduced by Holland and Leinhardt [24]. The p_1-model has previously been used by Bulashevska *et al.* [23] to model human protein-protein interaction networks. In this approach, edge probabilities are evaluated by summarizing topological properties of networks in a parametric form and associating them with sufficient statistics [23, 24]. The definition of the p_1-model for a directed graph is contained in the original article [24]. An equivalent log-linear formulation was proposed by Fienberg and Wasserman [110], in which each directed edge was assigned four Bernoulli variables Y_{ij00}, Y_{ij01}, Y_{ij10} and Y_{ij11}. Since our *GGR* network is an undirected graph, the model can be simplified by using only two Bernoulli variables Y_{ij0} and Y_{ij1} defined as follows:

$$Y_{ijk} = \begin{cases} 1 & if \quad u_{ij} = k, \\ 0 & otherwise \end{cases}$$

where, the binary outcome $u_{ij} = 1$ if $gene_i$ interacts with $gene_j$ in *GGR*, and $u_{ij} = 0$ otherwise. Under this simplified model, the posterior probability of an edge connecting $gene_i$ and $gene_j$ is given by:

$$log\{Pr(Y_{ij1} = 1)\} = \lambda_{ij} + \theta + \alpha_i + \alpha_j \quad (1)$$

$$log\{Pr(Y_{ij0} = 1)\} = \lambda_{ij} \quad (2)$$

for $i < j$. Here, θ is the global density parameter, α_i is the expansiveness/attractiveness of $gene_i$,

and λ_{ij} is the scaling parameter ensuring $\sum_k Y_{ijk} = 1$. We hypothesized that some aberrant gene-pairs involved in acquired resistance may have unusually high probability of appearing in Resistant-vs-Parental conditions, whereas other pairs may have unusually low probabilities. Hence, we used two Y_k data matrices, ${Y_k}^R$ and ${Y_k}^P$, from *GGR* matrices of Resistant and Parental samples, respectively. Note that it is possible to replace the expansiveness and attractiveness parameters by a single parameter α that represents the propensity of a gene to be connected in an undirected network.

We used a fully Bayesian approach, both for modeling the network parameters and their estimation. To estimate the model parameters, we used Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method implemented in WinBUGS [111] which allows users to construct complex Bayesian models in a simple manner. We constructed a hierarchical Bayesian model in which the model parameters were further defined as dependent upon hyperparameters as follows:

$$\begin{aligned}
\theta &\sim \mathcal{N}(0, {\sigma_\theta}^2) \\
\tau_\theta &\sim Gamma(a_0, b_0) \\
\begin{pmatrix} \alpha_i^R \\ \alpha_i^P \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) \\
\Sigma^{-1} &\sim Wishart\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2 \right) \\
a_0 &= 0.001 \\
b_0 &= 0.001
\end{aligned}$$

We assigned the density parameter θ a normal prior distribution with mean zero and standard deviation σ_θ. (In fact, this was implemented in WinBUGS using the precision parameter $\tau_\theta = {\sigma_\theta}^{-2}$). Next, the parameter τ_θ was assigned a gamma prior distribution with hyperparameters $a_0 = 0.001$ and $b_0 = 0.001$. We set $a_0 = 0.001$ and $b_0 = 0.001$ to express large uncertainty regarding the value σ_θ^2, following [23]. For the propensity parameters α_i^R and α_i^P, we selected the above prior following Adam *et al.* [112].

Robust selection of aberrant gene-pairs

One of our primary hypotheses in this study is that aberrant gene-pairs involved in network re-wiring in drug-resistance are likely to have high probabilities of occurring in one network (resistant or parental) but low probabilities in the other network. To determine which gene-pairs exhibit this pattern, we calculated two odds ratios defined in the following equations:

$$odds^R = \frac{Pr(Y_{ij1}^R = 1)}{Pr(Y_{ij1}^P = 1)} \tag{3}$$

$$odds^P = \frac{Pr(Y_{ij1}^P = 1)}{Pr(Y_{ij1}^R = 1)} \tag{4}$$

where, Y_{ij1}^R and Y_{ij1}^P are defined for resistant and parental networks, respectively, and their corresponding posterior probabilities are estimated using MCMC sampling. Before calculating these ratios, we normalized the posterior probabilities by their respective maximum values over all gene-pairs, since two values (Y_{ij1}^R and Y_{ij1}^P) may not be in the same scale. For the sake of

brevity we refer to these ratios as odds ratios, but they are more appropriately called *normalized* odds ratios.

Our intention is to identify gene-pairs for which only one of the two odds ratios (Eqs (3) and (4)) is very high. Additionally, we require that both posterior probabilities exceed a minimum threshold, since very small denominators can yield high odds ratio scores even if the edge has low probability in both networks. We therefore defined two thresholds, one for odds ratio values and another for posterior probabilities. We examined the distributions of all $odds^R$ and $odds^P$ values and set a threshold demarcating the top 20%. Next, we examined the distribution of posterior probabilities for gene-pairs exceeding the odds ratio threshold and set a second threshold to demarcate the top 50%. Finally, we chose only those gene-pairs that had posterior probabilities above that threshold, and identified as putative aberrant gene-pairs that were potentially involved in network rewiring in acquired resistance.

Edges were then grouped into two types: gene-pairs for which the $odds^R$ scores and the $Pr(u_{ij}^R = 1)$ were greater than their respective thresholds in Eq (3) were defined as *red* pairs, whereas gene-pairs for which the $odds^P$ scores and the $Pr(u_{ij}^P = 1)$ were greater than their respective thresholds in Eq (4) were defined as *green* pairs.

Enrichment of aberrant gene-pairs using known signaling links

Putative aberrant gene-pairs from the above Bayesian analyses were then further filtered by comparing them to another set of known (true) signaling links from the literature. For that purpose, we obtained a signaling network from the online resources of Wang Lab [34], which is claimed as the largest manually curated signaling network available to date. This network has more than 6,000 proteins and $\sim$63,000 binary relations defined, including activations, inhibitions and physical interactions. Note that signaling pathways from KEGG, Reactome, and WikiPathway databases were merely genesets, and to define true signaling links among the genes within those genesets we considered the signaling links from Wang Lab [34]. Next, to find dysregulated signaling pathways from KEGG, Reactome, and WikiPathway databases, we searched for significant overlaps between the set of true signaling links and the set of putative aberrant gene-pairs, for the genesets in a specific pathway. To this end, we designed a hypergeometric test as follows:

$$p = 1 - \sum_{i=0}^{x-1} \frac{\left(\binom{|M|}{i}\binom{N-|M|}{|K|-i}\right)}{\binom{N}{|K|}} \tag{5}$$

where N is the number of distinct gene-pairs contained in all of the signaling pathways (from a particular pathway database) and all the predicted aberrant gene-pairs, M is the set of all known signaling links in a given pathway, K is the set of aberrant gene-pairs predicted by our framework, and $x = |M \cap K|$. After measuring p-values using Eq (5), a False Discovery Rate (FDR) multiple comparison correction technique was conducted to obtain q-values. Signaling pathways with q-value <0.05 were considered to be significantly dysregulated in acquired drug resistance. A similar gene-pair enrichment test, called Edge Set Enrichment Analysis (ESEA) using the weighted Kolmogorov-Smirnov statistic was recently proposed by Han *et el.* for detecting dysregulated pathways in the context of gene expression datasets [113].

Identifying V-structures

To investigate the role of signaling rewiring in acquired drug resistance, we searched for a configuration of edges that we call a *V-structure* (Fig 1F). A V-structure consists of three genes connected by one *red* edge and one *green* edge. One gene, called a *crossing-gene*, is connected to both of the other genes, to one by a *red* edge and to the other by a *green* edge. Thus V-structures involve both types of aberrant pairs: one gene-pair present only in Resistant conditions, and another gene-pair present only in Parental conditions, with the *crossing-gene* common to both pairs. Our rationale is that the compensatory kinases may switch the oncogenic-addiction of cancer-related (growth/survival) genes to overcome their dependencies upon their primary driver kinases (e.g. EGFR/HER2) that were initially targeted in parental conditions with inhibitors [19, 30], thereby relapsing into their tumorigenic roles in acquired resistance. We hypothesise that *crossing-genes* that are dysregulated restore their metastatic phenotype (i.e. up- or down-regulation of oncogenes or tumor suppressor genes, respectively) in resistant conditions by forming a V-structure in the rewired signaling network.

Therefore, we define a V-structure to be a pair of aberrant gene-pairs (g_i, g_k) and (g_j, g_k) such that (g_i, g_k) are connected by a *green* edge and (g_j, g_k) are connected by a *red* edge. To identify V-structures, first we identified the set of common genes in the two mutually exclusive sets of aberrant gene-pairs (*red* and *green* gene-pairs). This set of common genes are the *crossing-genes* (see Fig 1). Next, we observed and enumerated all the gene-pairs (*red* and *green*) incident on each of the *crossing-genes*, and enumerated all of the possible pairings of one *red* and one *green* edge to form a V-structure.

Pathway configurations of V-structures: Type-I, Type-II, and Type-III configurations. Next, for each V-structure, we identified signaling pathways from KEGG, Reactome, and WikiPathway databases that contained at least one gene in that V-structure. We classified V-structures into three sub-types based on their configurations relative to these pathways. Firstly, Type-I V-structures are those in which all three member genes belong to different signaling pathways. Type-II V-structures are those in which the two aberrant gene-pairs in a particular V-structure are from two different signaling pathways, with the *crossing-gene* common to both pathways. Type-III V-structures are those in which all three genes are from the same signaling pathway. Note that Type-I and Type-II V-structures may represent signaling pathway cross-talks that play crucial roles in acquired drug-resistance. In our previous study, we investigated and explained the concept of Type-I V-structures, their involvement in the cross-talk between EGFR/ErbB and other signaling pathways, and their contribution to lapatinib resistance [10]. Type-III V-structures can explain the aberrant co-regulation of genes *within a single pathway*. We observed and analysed all the V-structures that overlap with the literature curated signaling network [34].

Supporting information

S1 Text. Supplementary Text 1. Supplementary Methods.
(PDF)

S1 Table. Supplementary Table 1. List of identified putative aberrant gene-pairs (for both SKBR3 and BT474) cell-lines in acquired resistance.
(XLSX)

S2 Table. Supplementary Table 2. Full results of pathway enrichment tests of identified aberrant gene-pairs in acquired resistance from KEGG, Reactome, and WikiPathway databases for both SKBR3 and BT474 cell-lines.
(XLSX)

S3 Table. Supplementary Table 3. Comparing our current model with the previous model by observing the percentages of non-direct (indirect and PPI pair) enriched links (aberrant pairs as known signaling links) in the aberrant signaling pathways from KEGG, Reactome, and WikiPathway databases *that were detected by our current but not the previous model*, for both SKBR3 and BT474 cell-lines.
(XLSX)

S4 Table. Supplementary Table 4. Comparing our current model with the previous model by observing the percentages of non-direct (indirect and PPI pair) enriched links (aberrant pairs as known signaling links) in the aberrant signaling pathways from KEGG, Reactome, and WikiPathway databases *that were detected by both of our current and previous models, and were ranked (based on enrichment q-value) high in the current model but low in the previous model*, for both SKBR3 and BT474 cell-lines.
(XLSX)

S5 Table. Supplementary Table 5. Summary of Type-I, Type-II and Type-III enrichment of V-structures in KEGG, Reactome, and WikiPathway databases in *SKBR3* cell-line.
(XLSX)

S6 Table. Supplementary Table 6. Summary of Type-I, Type-II and Type-III enrichment of V-structures in KEGG, Reactome, and WikiPathway databases in *BT474* cell-line.
(XLSX)

S7 Table. Supplementary Table 7. CGC genes in all the Type-I, Type-II and Type-III V-structures in *SKBR3* and *BT474* cell-lines.
(XLSX)

Acknowledgments

We thank Dr. Tianhai Tian for his initial support in this project.

Author Contributions

Conceptualization: AKMA.

Data curation: AKMA.

Formal analysis: AKMA.

Investigation: AKMA AL.

Methodology: AKMA.

Project administration: AKMA JMK.

Resources: AKMA JMK.

Software: AKMA.

Supervision: JMK AL.

Validation: AKMA AL JMK.

Visualization: AKMA.

Writing – original draft: AKMA.

Writing – review & editing: AL JMK.

References

1. Brivanlou AH, Darnell JE. Signal transduction and the control of gene expression. Science. 2002; 295 (5556):813–818. https://doi.org/10.1126/science.1066355 PMID: 11823631
2. Sarkar S, Mandal M. Growth factor receptors and apoptosis regulators: signaling pathways, prognosis, chemosensitivity and treatment outcomes of breast cancer. Breast Cancer (Auckl). 2009; 3:47–60. PMID: 21556249
3. Huang L, Fu L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. Acta Pharmaceutica Sinica B. 2015; 5(5):390–401. https://doi.org/10.1016/j.apsb.2015.07.001 PMID: 26579470
4. Yamaguchi H, Chang SS, Hsu JL, Hung MC. Signaling cross-talk in the resistance to HER family receptor targeted therapy. Oncogene. 2014; 33(9):1073–1081. https://doi.org/10.1038/onc.2013.74 PMID: 23542173
5. Stuhlmiller TJ, Miller SM, Zawistowski JS, Nakamura K, Beltran AS, Duncan JS, et al. Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. Cell Rep. 2015; 11(3):390–404. https://doi.org/10.1016/j.celrep.2015.03.037 PMID: 25865888
6. Clark GJ, Der CJ. Aberrant function of the Ras signal transduction pathway in human breast cancer. Breast Cancer Res Treat. 1995; 35(1):133–144. https://doi.org/10.1007/BF00694753 PMID: 7612899
7. Shi I, Sadraei N Hashemi, Duan ZH, Shi T. Aberrant signaling pathways in squamous cell lung carcinoma. Cancer Inform. 2011; 10:273–285. https://doi.org/10.4137/CIN.S8283 PMID: 22174565
8. Mao H, Lebrun DG, Yang J, Zhu VF, Li M. Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. Cancer Invest. 2012; 30(1):48–56. https://doi.org/10.3109/07357907.2011.630050 PMID: 22236189
9. McCleary-Wheeler AL, McWilliams R, Fernandez-Zapico ME. Aberrant signaling pathways in pancreatic cancer: a two compartment view. Mol Carcinog. 2012; 51(1):25–39. https://doi.org/10.1002/mc.20827 PMID: 22162229
10. Azad A, Lawen A, Keith J. Prediction of signaling cross-talks contributing to acquired drug resistance in breast cancer cells by Bayesian statistical modeling. BMC Syst Biol. 2015; 9(1):2. https://doi.org/10.1186/s12918-014-0135-x PMID: 25599599
11. Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, et al. EGFR Mutation and Resistance of Non-Small-Cell Lung Cancer to Gefitinib. New England Journal of Medicine. 2005; 352 (8):786–792. https://doi.org/10.1056/NEJMoa044238 PMID: 15728811
12. Amin DN, Sergina N, Ahuja D, McMahon M, Blair JA, Wang D, et al. Resiliency and vulnerability in the HER2-HER3 tumorigenic driver. Sci Transl Med. 2010; 2(16):16ra7. https://doi.org/10.1126/scitranslmed.3000389 PMID: 20371474
13. Garrett JT, Olivares MG, Rinehart C, Granja-Ingram ND, Sanchez V, Chakrabarty A, et al. Transcriptional and posttranslational up-regulation of HER3 (ErbB3) compensates for inhibition of the HER2 tyrosine kinase. Proc Natl Acad Sci USA. 2011; 108(12):5021–5026. https://doi.org/10.1073/pnas.1016140108 PMID: 21385943
14. Azuma K, Tsurutani J, Sakai K, Kaneda H, Fujisaka Y, Takeda M, et al. Switching addictions between HER2 and FGFR2 in HER2-positive breast tumor cells: FGFR2 as a potential target for salvage after lapatinib failure. Biochem Biophys Res Commun. 2011; 407(1):219–224. https://doi.org/10.1016/j.bbrc.2011.03.002 PMID: 21377448
15. Huang C, Park CC, Hilsenbeck SG, Ward R, Rimawi MF, Wang YC, et al. ß1 integrin mediates an alternative survival pathway in breast cancer cells resistant to lapatinib. Breast Cancer Res. 2011; 13 (4):R84. https://doi.org/10.1186/bcr2936 PMID: 21884573
16. Rexer BN, Arteaga CL. Intrinsic and acquired resistance to HER2-targeted therapies in HER2 gene-amplified breast cancer: mechanisms and clinical implications. Crit Rev Oncog. 2012; 17(1):1–16. https://doi.org/10.1615/CritRevOncog.v17.i1.20 PMID: 22471661
17. Kolch W, Halasz M, Granovskaya M, Kholodenko BN. The dynamic control of signal transduction networks in cancer cells. Nat Rev Cancer. 2015; 15(9):515–527. https://doi.org/10.1038/nrc3983 PMID: 26289315
18. Logue JS, Morrison DK. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. Genes Dev. 2012; 26(7):641–650. https://doi.org/10.1101/gad.186965.112 PMID: 22474259
19. Komurov K, Tseng JT, Muller M, Seviour EG, Moss TJ, Yang L, et al. The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant ErbB2-positive breast cancer cells. Molecular Systems Biology. 2012; 8(1). https://doi.org/10.1038/msb.2012.25 PMID: 22864381
20. Locasale JW. Metabolic rewiring drives resistance to targeted cancer therapy. Mol Syst Biol. 2012; 8:597. https://doi.org/10.1038/msb.2012.30 PMID: 22806144

21. Akhavan D, Pourzia AL, Nourian AA, Williams KJ, Nathanson D, Babic I, et al. De-Repression of PDGFR-beta Transcription Promotes Acquired Resistance to EGFR Tyrosine Kinase Inhibitors in Glioblastoma Patients. Cancer Discovery. 2013; 3(5):534–547. https://doi.org/10.1158/2159-8290.CD-12-0502 PMID: 23533263

22. Saul ZM, Filkov V. Exploring biological network structure using exponential random graph models. Bioinformatics. 2007; 23(19):2604–2611. https://doi.org/10.1093/bioinformatics/btm370 PMID: 17644557

23. Bulashevska S, Bulashevska A, Eils R. Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. BMC Bioinformatics. 2010; 11:46. https://doi.org/10.1186/1471-2105-11-46 PMID: 20100321

24. Holland PW, Leinhardt S. An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association. 1981; 76(373):33–50. https://doi.org/10.2307/2287042

25. Yun J, Rago C, Cheong I, Pagliarini R, Angenendt P, Rajagopalan H, et al. Glucose deprivation contributes to the development of KRAS pathway mutations in tumor cells. Science. 2009; 325 (5947):1555–1559. https://doi.org/10.1126/science.1174229 PMID: 19661383

26. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009; 324(5930):1029–1033. https://doi.org/10.1126/science.1160809 PMID: 19460998

27. Roberts CG, Millar EK, O'Toole SA, McNeil CM, Lehrbach GM, Pinese M, et al. Identification of PUMA as an estrogen target gene that mediates the apoptotic response to tamoxifen in human breast cancer cells and predicts patient outcome and tamoxifen responsiveness in breast cancer. Oncogene. 2011; 30(28):3186–3197. https://doi.org/10.1038/onc.2011.36 PMID: 21383694

28. Ng CP, Bonavida B. A new challenge for successful immunotherapy by tumors that are resistant to apoptosis: two complementary signals to overcome cross-resistance. Adv Cancer Res. 2002; 85:145–174. https://doi.org/10.1016/S0065-230X(02)85005-9 PMID: 12374285

29. Debatin KM. Apoptosis pathways in cancer and cancer therapy. Cancer Immunol Immunother. 2004; 53(3):153–159. https://doi.org/10.1007/s00262-003-0474-8 PMID: 14749900

30. Sharifnia T, Rusu V, Piccioni F, Bagul M, Imielinski M, Cherniack AD, et al. Genetic modifiers of EGFR dependence in non-small cell lung cancer. Proc Natl Acad Sci USA. 2014; 111(52):18661–18666. https://doi.org/10.1073/pnas.1412228112 PMID: 25512530

31. Tian Y, Zhang B, Hoffman EP, Clarke R, Zhang Z, Shih IeM, et al. Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks. BMC Syst Biol. 2014; 8:87. https://doi.org/10.1186/s12918-014-0087-1 PMID: 25055984

32. Jung S, Kim S. EDDY: a novel statistical gene set test method to detect differential genetic dependencies. Nucleic Acids Res. 2014; 42(7):e60. https://doi.org/10.1093/nar/gku099 PMID: 24500204

33. Elo LL, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics. 2007; 23 (16):2096–2103. https://doi.org/10.1093/bioinformatics/btm309 PMID: 17553854

34. Wang E. Human signaling network; 2014. Database: Cancer Systems Biology and Bioinformatics. Available from: http://www.cancer-systemsbiology.org/dataandsoftware.htm

35. Dhomen NS, Mariadason J, Tebbutt N, Scott AM. Therapeutic targeting of the epidermal growth factor receptor in human cancer. Crit Rev Oncog. 2012; 17(1):31–50. https://doi.org/10.1615/CritRevOncog.v17.i1.40 PMID: 22471663

36. Baselga J, Arteaga CL. Critical update and emerging trends in epidermal growth factor receptor targeting in cancer. J Clin Oncol. 2005; 23(11):2445–2459. https://doi.org/10.1200/JCO.2005.11.890 PMID: 15753456

37. Costa DB, Halmos B, Kumar A, Schumer ST, Huberman MS, Boggon TJ, et al. BIM mediates EGFR tyrosine kinase inhibitor-induced apoptosis in lung cancers with oncogenic EGFR mutations. PLoS Med. 2007; 4(10):e315. https://doi.org/10.1371/journal.pmed.0040315 PMID: 17973572

38. Toyooka S, Date H, Uchida A, Kiura K, Takata M. The Epidermal Growth Factor Receptor D761Y Mutation and Effect of Tyrosine Kinase Inhibitor. American Association for Cancer Research. 2007; 13 (11):3431–3431. https://doi.org/10.1158/1078-0432.CCR-07-0070 PMID: 17545553

39. Bean J, Riely GJ, Balak M, Marks JL, Ladanyi M, Miller VA, et al. Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. Clin Cancer Res. 2008; 14(22):7519–7525. https://doi.org/10.1158/1078-0432.CCR-08-0151 PMID: 19010870

40. Landi L, Cappuzzo F. HER2 and lung cancer. Expert Review of Anticancer Therapy. 2013; 13 (10):1219–28. https://doi.org/10.1586/14737140.2013.846830 PMID: 24134423

41. Luo M, Fu LW. Redundant kinase activation and resistance of EGFR-tyrosine kinase inhibitors. Am J Cancer Res. 2014; 4(6):608–628. PMID: 25520855

42. Bertotti A, Burbridge MF, Gastaldi S, Galimi F, Torti D, Medico E, et al. Only a subset of Met-activated pathways are required to sustain oncogene addiction. Sci Signal. 2009; 2(102):er11. PMID: 20039471

43. Canfield K, Li J, Wilkins OM, Morrison MM, Ung M, Wells W, et al. Receptor tyrosine kinase ERBB4 mediates acquired resistance to ERBB2 inhibitors in breast cancer cells. Cell Cycle. 2015; 14(4):648–655. https://doi.org/10.4161/15384101.2014.994966 PMID: 25590338

44. Cheung HW, Du J, Boehm JS, He F, Weir BA, Wang X, et al. Amplification of CRKL induces transformation and EGFR inhibitor resistance in human non small cell lung cancers. Cancer Discovery. 2011;.

45. Azuaje F, Tiemann K, Niclou SP. Therapeutic control and resistance of the EGFR-driven signaling network in glioblastoma. Cell Commun Signal. 2015; 13:23. https://doi.org/10.1186/s12964-015-0098-6 PMID: 25885672

46. Guix M, Faber AC, Wang SE, Olivares MG, Song Y, Qu S, et al. Acquired resistance to EGFR tyrosine kinase inhibitors in cancer cells is mediated by loss of IGF-binding proteins. J Clin Invest. 2008; 118 (7):2609–2619. https://doi.org/10.1172/JCI34588 PMID: 18568074

47. Peled N, Wynes MW, Ikeda N, Ohira T, Yoshida K, Qian J, et al. Insulin-like growth factor-1 receptor (IGF-1R) as a biomarker for resistance to the tyrosine kinase inhibitor gefitinib in non-small cell lung cancer. Cell Oncol (Dordr). 2013; 36(4):277–288. https://doi.org/10.1007/s13402-013-0133-9

48. Baker AT, Zlobin A, Osipo C. Notch-EGFR/HER2 Bidirectional Crosstalk in Breast Cancer. Front Oncol. 2014; 4:360. https://doi.org/10.3389/fonc.2014.00360 PMID: 25566499

49. Wang Z, Li Y, Ahmad A, Azmi AS, Banerjee S, Kong D, et al. Targeting Notch signaling pathway to overcome drug resistance for cancer therapy. Biochim Biophys Acta. 2010; 1806(2):258–267. https://doi.org/10.1016/j.bbcan.2010.06.001 PMID: 20600632

50. Al-Hussaini H, Subramanyam D, Reedijk M, Sridhar SS. Notch signaling pathway as a therapeutic target in breast cancer. Mol Cancer Ther. 2011; 10(1):9–15. https://doi.org/10.1158/1535-7163.MCT-10-0677 PMID: 20971825

51. Loh YN, Hedditch EL, Baker LA, Jary E, Ward RL, Ford CE. The Wnt signalling pathway is upregulated in an in vitro model of acquired tamoxifen resistant breast cancer. BMC Cancer. 2013; 13:174. https://doi.org/10.1186/1471-2407-13-174 PMID: 23547709

52. Chikazawa N, Tanaka H, Tasaka T, Nakamura M, Tanaka M, Onishi H, et al. Inhibition of Wnt signaling pathway decreases chemotherapy-resistant side-population colon cancer cells. Anticancer Res. 2010; 30(6):2041–2048. PMID: 20651349

53. Ballas MS, Chachoua A. Rationale for targeting VEGF, FGF, and PDGF for the treatment of NSCLC. Onco Targets Ther. 2011; 4:43–58. https://doi.org/10.2147/OTT.S18155 PMID: 21691577

54. Chatterjee S, Heukamp LC, Siobal M, Schottle J, Wieczorek C, Peifer M, et al. Tumor VEGF:VEGFR2 autocrine feed-forward loop triggers angiogenesis in lung cancer. The Journal of Clinical Investigation. 2013; 123(4):1732–1740. https://doi.org/10.1172/JCI65385 PMID: 23454747

55. Bianco R, Rosa R, Damiano V, Daniele G, Gelardi T, Garofalo S, et al. Vascular Endothelial Growth Factor Receptor-1 Contributes to Resistance to Anti-Epidermal Growth Factor Receptor Drugs in Human Cancer Cells. Clinical Cancer Research. 2008; 14(16):5069–5080. https://doi.org/10.1158/1078-0432.CCR-07-4905 PMID: 18694994

56. Azuma K, Kawahara A, Sonoda K, Nakashima K, Tashiro K, Watari K, et al. FGFR1 activation is an escape mechanism in human lung cancer cells resistant to afatinib, a pan-EGFR family kinase inhibitor. Oncotarget. 2014; 5(15). https://doi.org/10.18632/oncotarget.1866 PMID: 25115383

57. Zhang J, Ji JY, Yu M, Overholtzer M, Smolen GA, Wang R, et al. YAP-dependent induction of amphiregulin identifies a non-cell-autonomous component of the Hippo pathway. Nat Cell Biol. 2009; 11 (12):1444–1450. https://doi.org/10.1038/ncb1993 PMID: 19935651

58. Zhao Y, Yang X. The Hippo pathway in chemotherapeutic drug resistance. Int J Cancer. 2015; 137 (12):2767–2773. https://doi.org/10.1002/ijc.29293 PMID: 25348697

59. Huang JM, Nagatomo I, Suzuki E, Mizuno T, Kumagai T, Berezov A, et al. YAP modifies cancer cell sensitivity to EGFR and survivin inhibitors and is negatively regulated by the non-receptor type protein tyrosine phosphatase 14. Oncogene. 2013; 32(17):2220–2229. https://doi.org/10.1038/onc.2012.231 PMID: 22689061

60. Serizawa M, Takahashi T, Yamamoto N, Koh Y. Combined treatment with erlotinib and a transforming growth factor-beta type I receptor inhibitor effectively suppresses the enhanced motility of erlotinib-resistant non-small-cell lung cancer cells. J Thorac Oncol. 2013; 8(3):259–269. https://doi.org/10.1097/JTO.0b013e318279e942 PMID: 23334091

61. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim Js, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009; 25(1):75–82. https://doi.org/10.1093/bioinformatics/btn577 PMID: 18990722

62. Huang daW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956

63. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics. 2006; 22(23):2926–2933. https://doi.org/10.1093/bioinformatics/btl483 PMID: 17000751

64. Han J, Shi X, Zhang Y, Xu Y, Jiang Y, Zhang C, et al. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. Sci Rep. 2015; 5:13044. https://doi.org/10.1038/srep13044 PMID: 26267116

65. Han J, Li C, Yang H, Xu Y, Zhang C, Ma J, et al. A novel dysregulated pathway-identification analysis based on global influence of within-pathway effects and crosstalk between pathways. J R Soc Interface. 2015; 12(102):20140937. https://doi.org/10.1098/rsif.2014.0937 PMID: 25551156

66. Liu L, Greger J, Shi H, Liu Y, Greshock J, Annan R, et al. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. Cancer Res. 2009; 69(17):6871–6878. https://doi.org/10.1158/0008-5472.CAN-08-4490 PMID: 19671800

67. Karakashev SV, Reginato MJ. Hypoxia/HIF-alpha induces lapatinib resistance in ERBB2-positive breast cancer cells via regulation of DUSP2. Oncotarget. 2015; 6(4):1967–1980. https://doi.org/10.18632/oncotarget.2806 PMID: 25596742

68. Xia W, Bacus S, Husain I, Liu L, Zhao S, Liu Z, et al. Resistance to ErbB2 tyrosine kinase inhibitors in breast cancer is mediated by calcium-dependent activation of RelA. Mol Cancer Ther. 2010; 9(2):292–299. https://doi.org/10.1158/1535-7163.MCT-09-1041 PMID: 20124457

69. Lynch MA, Petrel TA, Song H, Knobloch TJ, Casto BC, Ramljak D, et al. Responsiveness to transforming growth factor-beta (TGF-beta)-mediated growth inhibition is a function of membrane-bound TGF-beta type II receptor in human breast cancer cells. Gene Expr. 2001; 9(4-5):157–171. https://doi.org/10.3727/000000001783992560 PMID: 11444526

70. Serra V, Eichhorn PJ, Garcia-Garcia C, Ibrahim YH, Prudkin L, Sanchez G, et al. RSK3/4 mediate resistance to PI3K pathway inhibitors in breast cancer. J Clin Invest. 2013; 123(6):2551–2563. https://doi.org/10.1172/JCI66343 PMID: 23635776

71. Hegde PS, Rusnak D, Bertiaux M, Alligood K, Strum J, Gagnon R, et al. Delineation of molecular mechanisms of sensitivity to lapatinib in breast cancer cell lines using global gene expression profiles. Mol Cancer Ther. 2007; 6(5):1629–1640. https://doi.org/10.1158/1535-7163.MCT-05-0399 PMID: 17513611

72. Liu NN, Xi Y, Callaghan MU, Fribley A, Moore-Smith L, Zimmerman JW, et al. SMAD4 is a potential prognostic marker in human breast carcinomas. Tumour Biol. 2014; 35(1):641–650. https://doi.org/10.1007/s13277-013-1088-1 PMID: 23975369

73. Deckers M, van Dinther M, Buijs J, Que I, Lowik C, van der Pluijm G, et al. The tumor suppressor Smad4 is required for transforming growth factor beta-induced epithelial to mesenchymal transition and bone metastasis of breast cancer cells. Cancer Res. 2006; 66(4):2202–2209. https://doi.org/10.1158/0008-5472.CAN-05-3560 PMID: 16489022

74. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013; 41(Database issue):D955–961. https://doi.org/10.1093/nar/gks1111 PMID: 23180760

75. Hu J, Rho HS, Newman RH, Zhang J, Zhu H, Qian J. PhosphoNetworks: a database for human phosphorylation networks. Bioinformatics. 2014; 30(1):141–142. https://doi.org/10.1093/bioinformatics/btt627 PMID: 24227675

76. Khalil S, Tan GA, Giri DD, Zhou XK, Howe LR. Activation status of Wnt/beta-catenin signaling in normal and neoplastic breast tissues: relationship to HER2/neu expression in human and mouse. PLoS ONE. 2012; 7(3):e33421. https://doi.org/10.1371/journal.pone.0033421 PMID: 22457761

77. Lamb R, Ablett MP, Spence K, Landberg G, Sims AH, Clarke RB. Wnt pathway activity in breast cancer sub-types and stem-like cells. PLoS ONE. 2013; 8(7):e67811. https://doi.org/10.1371/journal.pone.0067811 PMID: 23861811

78. Sasaki T, Suzuki H, Yagi K, Furuhashi M, Yao R, Susa S, et al. Lymphoid enhancer factor 1 makes cells resistant to transforming growth factor beta-induced repression of c-myc. Cancer Res. 2003; 63(4):801–806. PMID: 12591729

79. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res. 2000; 28(1):126–128. https://doi.org/10.1093/nar/28.1.126 PMID: 10592200

80. Takano Y, Kato Y, Masuda M, Ohshima Y, Okayasu I. Cyclin D2, but not cyclin D1, overexpression closely correlates with gastric cancer progression and prognosis. J Pathol. 1999; 189(2):194–200. https://doi.org/10.1002/(SICI)1096-9896(199910)189:2%3C194::AID-PATH426%3E3.0.CO;2-P PMID: 10547574

81. Schmidt M, Fernandez de Mattos S, van der Horst A, Klompmaker R, Kops GJ, Lam EW, et al. Cell cycle inhibition by FoxO forkhead transcription factors involves downregulation of cyclin D. Mol Cell Biol. 2002; 22(22):7842–7852. https://doi.org/10.1128/MCB.22.22.7842-7852.2002 PMID: 12391153

82. Qin H, Chan MW, Liyanarachchi S, Balch C, Potter D, Souriraj IJ, et al. An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. BMC Syst Biol. 2009; 3:73. https://doi.org/10.1186/1752-0509-3-73 PMID: 19615063

83. Kanehisa M. The KEGG database. Novartis Found Symp. 2002; 247:91–101. https://doi.org/10.1002/0470857897.ch8 PMID: 12539951

84. Kechagioglou P, Papi RM, Provatopoulou X, Kalogera E, Papadimitriou E, Grigoropoulos P, et al. Tumor suppressor PTEN in breast cancer: heterozygosity, mutations and protein expression. Anticancer Res. 2014; 34(3):1387–1400. PMID: 24596386

85. Eichhorn PJ, Gili M, Scaltriti M, Serra V, Guzman M, Nijkamp W, et al. Phosphatidylinositol 3-kinase hyperactivation results in lapatinib resistance that is reversed by the mTOR/phosphatidylinositol 3-kinase inhibitor NVP-BEZ235. Cancer Res. 2008; 68(22):9221–9230. https://doi.org/10.1158/0008-5472.CAN-08-1740 PMID: 19010894

86. Bouchalova K, Cizkova M, Cwiertka K, Trojanec R, Friedecky D, Hajduch M. Lapatinib in breast cancer—the predictive significance of HER1 (EGFR), HER2, PTEN and PIK3CA genes and lapatinib plasma level assessment. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub. 2010; 154 (4):281–288. https://doi.org/10.5507/bp.2010.043 PMID: 21293538

87. Du G, Bian L, Wang T, Xu X, Zhang S, Guo Y, et al. [PTEN loss correlates withthe clinical efficacy of lapatinib in HER2 positive metastatic breast cancer with trastuzumab-resistance]. Zhonghua Yi Xue Za Zhi. 2015; 95(28):2264–2267. PMID: 26710948

88. Carpenter RL, Lo HW. Regulation of Apoptosis by HER2 in Breast Cancer. Journal of Carcinogenesis & Mutagenesis. 2013; 0(0):1–7. https://doi.org/10.4172/2157-2518.S7-003 PMID: 27088047

89. Huang S, Benavente S, Armstrong EA, Li C, Wheeler DL, Harari PM. p53 modulates acquired resistance to EGFR inhibitors and radiation. Cancer Res. 2011; 71(22):7071–7079. https://doi.org/10.1158/0008-5472.CAN-11-0128 PMID: 22068033

90. Boeckx C, Baay M, Wouters A, Specenier P, Vermorken JB, Peeters M, et al. Anti-epidermal growth factor receptor therapy in head and neck squamous cell carcinoma: focus on potential molecular mechanisms of drug resistance. Oncologist. 2013; 18(7):850–864. https://doi.org/10.1634/theoncologist.2013-0013 PMID: 23821327

91. Han C, Zhao R, Liu X, Srivastava A, Gong L, Mao H, et al. DDB2 suppresses tumorigenicity by limiting the cancer stem cell population in ovarian cancer. Mol Cancer Res. 2014; 12(5):784–794. https://doi.org/10.1158/1541-7786.MCR-13-0638 PMID: 24574518

92. Ennen M, Klotz R, Touche N, Pinel S, Barbieux C, Besancenot V, et al. DDB2: a novel regulator of NF-kB and breast tumor invasion. Cancer Res. 2013; 73(16):5040–5052. https://doi.org/10.1158/0008-5472.CAN-12-3655 PMID: 23774208

93. Kattan Z, Marchal S, Brunner E, Ramacci C, Leroux A, Merlin JL, et al. Damaged DNA binding protein 2 plays a role in breast cancer cell growth. PLoS ONE. 2008; 3(4):e2002. https://doi.org/10.1371/journal.pone.0002002 PMID: 18431487

94. He YH. Involvement of DDB2 Down-regulation in Lapatinib-induced Cross-resistance to Chemotherapy. Ph.D. Thesis, National Digital Library of Theses and Dissertations in Taiwan; 2012. Available from: http://handle.ncl.edu.tw/11296/ndltd/06833793361974888604

95. Nieto-Sampedro M, Valle-Argos B, Gomez-Nicola D, Fernandez-Mayoralas A, Nieto-Diaz M. Inhibitors of Glioma Growth that Reveal the Tumour to the Immune System. Clin Med Insights Oncol. 2011; 5:265–314. https://doi.org/10.4137/CMO.S7685 PMID: 22084619

96. Wee KB, Surana U, Aguda BD. Oscillations of the p53-Akt network: implications on cell survival and death. PLoS ONE. 2009; 4(2):e4407. https://doi.org/10.1371/journal.pone.0004407 PMID: 19197384

97. Chen Z, Trotman LC, Shaffer D, Lin HK, Dotan ZA, Niki M, et al. Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. Nature. 2005; 436(7051):725–730. https://doi.org/10.1038/nature03918 PMID: 16079851

98. Sun NK, Sun CL, Lin CH, Pai LM, Chao CC. Damaged DNA-binding protein 2 (DDB2) protects against UV irradiation in human cells and Drosophila. J Biomed Sci. 2010; 17:27. https://doi.org/10.1186/1423-0127-17-27 PMID: 20398405

99. Carson C, Omolo B, Chu H, Zhou Y, Sambade MJ, Peters EC, et al. A prognostic signature of defective p53-dependent G1 checkpoint function in melanoma cell lines. Pigment Cell Melanoma Res. 2012; 25(4):514–526. https://doi.org/10.1111/j.1755-148X.2012.01010.x PMID: 22540896

100. Kim HK, Choi IJ, Kim CG, Kim HS, Oshima A, Michalowski A, et al. A gene expression signature of acquired chemoresistance to cisplatin and fluorouracil combination chemotherapy in gastric cancer

patients. PLoS ONE. 2011; 6(2):e16694. https://doi.org/10.1371/journal.pone.0016694 PMID: 21364753

101. Awan A, Bari H, Yan F, Moksong S, Yang S, Chowdhury S, et al. Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. IET Syst Biol. 2007; 1(5):292–297. https://doi.org/10.1049/iet-syb:20060068 PMID: 17907678

102. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140. https://doi.org/10.1038/msb4100180 PMID: 17940530

103. Guney E, Sanz-Pamplona R, Sierra A, Oliva B. In: Azmi AS, editor. Understanding Cancer Progression Using Protein Interaction Networks. Dordrecht: Springer Netherlands; 2012. p. 167–195.

104. Wang E. COSMIC: Cancer Gene census; 2017. Available from: http://cancer.sanger.ac.uk/census/

105. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4(3):177–183. https://doi.org/10.1038/nrc1299 PMID: 14993899

106. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS ONE. 2010; 5(2):e8918. https://doi.org/10.1371/journal.pone.0008918 PMID: 20169195

107. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42(Database issue):D472–477. https://doi.org/10.1093/nar/gkt1102 PMID: 24243840

108. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

109. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012; 40(Database issue):D1301–1307. https://doi.org/10.1093/nar/gkr1074 PMID: 22096230

110. Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika. 1996; 61(3):401–425. https://doi.org/10.1007/BF02294547

111. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. Statistics and Computing. 2000; 10(4):325–337. https://doi.org/10.1023/A:1008929526011

112. Adams S, Carter N, Hadlock C, Haughton D, Sirbu G. Change in connectivity in a social network over time: A bayesian perspective. Connections. 2008; 28:17–27.

113. Han J, Shi X, Zhang Y, Xu Y, Jiang Y, Zhang C, et al. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. Sci Rep. 2015; 5:13044. https://doi.org/10.1038/srep13044 PMID: 26267116

Chapter 6

Inferring Network Structures

6.1 Introduction

During my PhD candidature, I was also involved with some additional projects in collaboration with another PhD student (*Salem A. Alyami*) in our research group, which are not directly related with my main research hypothesis, but have enormous potential to be used in relevant research problems. These auxiliary works of mine yielded several publications, some of which have already been published, and others are in preparation. In this chapter, I list all of these publications with corresponding abstracts. Full texts of these published articles are included as Appendix D.

6.2 Relevance to my primary research focus

Although these additional projects do not directly fit in the scope of this thesis, they share with my thesis a common but key component in their frameworks: '*Inferring Network Structure*'. All the collaborative projects with *Salem A. Alyami* involve the structure and parameter inference of Bayesian networks (BNs) by MCMC (Markov Chain Monte Carlo) sampling methods. BNs are a widely used tool for modelling cell

signalling networks [1], and MCMC methods are an important statistical technique for inferring BNs. My book chapter reviews some methods which infer gene-gene relationship networks [2, 3] by combining multiple heterogenous datasets, such as gene expression, copy number aberration, methylation, and protein-protein interaction which may provide some insights into how integrated approaches can be used in inferring biological networks.

6.3 Articles Published (total: 3)

1) *Title*: Uniform Sampling of Directed and Undirected Graphs Conditional on Vertex Connectivity [4]

- *Article nature:* Journal article (peer-reviewed)
- *Percentage of contribution:* 30% (Implementing the model, analysing the results and proofreading the manuscript)
- *Abstract:*

  Many applications in graph analysis require a space of graphs or networks to be sampled uniformly at random. For example, one may need to efficiently draw a small representative sample of graphs from a particular large target space. We assume that a uniform distribution $f(N, E) = 1/|X|$ has been defined, where N is a set of nodes, E is a set of edges, (N, E) is a graph in the target space X and $|X|$ is the (finite) total number of graphs in the target space. We propose a new approach to sample graphs at random from such a distribution. The new approach uses a Markov chain Monte Carlo method called the Neighbourhood Sampler. We validate the new sampling technique by simulating from feasible spaces of directed or undirected graphs, and compare its computational efficiency with the conventional Metropolis-Hastings Sampler. The simulation results indicate efficient uniform sampling of the target spaces, and more rapid rate of convergence than Metropolis-Hastings Sampler.

- *Reference*:

  Alyami, S., **Azad, A.K.M.**, Keith, JM. "Uniform Sampling of Directed and Undirected Graphs Conditional on Vertex Connectivity." *Electronic Notes in Discrete Mathematics*, 2016, 53:43-55. DOI: 10.1016/j.endm.2016.05.005

2) *Title*: The Neighborhood MCMC sampler for learning Bayesian networks [5]

- *Article nature:* Conference article (peer-reviewed)

- *Percentage of contribution:* 30% (Implementing the model, analysing the results and proofreading the manuscript)

- *Abstract:*

  Getting stuck in local maxima is a problem that arises while learning Bayesian networks (BNs) structures. In this paper, we studied a recently proposed Markov chain Monte Carlo (MCMC) sampler, called the Neighbourhood sampler (NS), and examined how efficiently it can sample BNs when local maxima are present. We assume that a posterior distribution $f(N, E|D)$ has been defined, where D represents data relevant to the inference, N and E are the sets of nodes and directed edges, respectively. We illustrate the new approach by sampling from such a distribution, and inferring BNs. The simulations conducted in this paper show that the new learning approach substantially avoids getting stuck in local modes of the distribution, and achieves a more rapid rate of convergence, compared to other common algorithms e.g. the MCMC Metropolis-Hastings sampler.

- *Reference*:

  Alyami, S., **Azad, A.K.M.**, Keith, JM. "The Neighborhood MCMC sampler for learning Bayesian networks." *Proc. SPIE 10011, First International Workshop on Pattern Recognition*, 2016, 100111K. DOI: 10.1117/12.2242708

3) *Title*: Integrating heterogeneous datasets for cancer module identification. [6]

- *Article nature:* Book chapter
- *Percentage of contribution:* 95% (Concept, collecting data and writing manuscript)
- *Abstract:*

  The availability of multiple heterogeneous high-throughput datasets provides an enabling resource for cancer systems biology. Types of data include: Gene Expression (GE), Copy Number Aberration (CNA), miRNA expression, Methylation, and Protein-Protein Interactions (PPI). One important problem that can potentially be solved using such data is to determine which of the possible pair-wise interactions among genes contribute to a range of cancer-related events, from tumorigenesis to metastasis. It has been shown by various studies that applying integrated knowledge from multi-omics datasets elucidates such complex phenomena with higher statistical significance than using a single type of dataset individually. However, computational methods for processing multiple data types simultaneously are needed. This chapter reviews some of the computational methods that use integrated approaches to find cancer-related modules.

- *Reference*:

  Azad, A. K. M.. "Integrating Heterogeneous Datasets for Cancer Module Identification", *Bioinformatics Volume II: Structure, Function, and Applications*, pages 119-137. Springer New York, New York, NY, 2017.

6.4 Articles in preparation (total: 3)

1) *Title*: BN-MCMC: a software for inferring Bayesian Network using MCMC methods

- *Article nature:* Journal article
- *Target journal:* Bioinformatics (peer-reviewed)
- *Percentage of contribution:* 50% (Contributing to the software modelling, implementing the complete software, and contributing to the result analysis)
- *Abstract:*

  This software article presents a graphical user interface (GUI): BN-MCMC for sampling Bayesian networks (BNs) using *three* MCMC sampling methods: *Neighbourhood sampler*, *Hit-and-Run sampler*, and *Metropolis-Hastings sampler*. Each of the sampling methods use adaptive techniques of both adjacent graph selection and function scoring that enables the inference of large-scale BNs. This interface provides a user-friendly environment with intuitive software design. For each of the samplers, a set of numerical outputs are saved in local files, and a set of graphical outputs are depicted in the result panel. All the input parameters including method and output settings are separated from the result panel. Given the enormous importance of Bayesian network inference in various fields of research from social network to systems biology, we hope *BN-MCMC* can be significantly useful to a vast community of researchers.
- *Reference*:

  Azad, A.K.M. and Alyami, SA. and Keith, JM. "BN-MCMC: a software for inferring Bayesian Network using MCMC methods."

2) *Title*: Adaptive techniques for large-scale Bayesian network inference using MCMC methods

- *Article nature:* Journal article
- *Target journal:* Bioinformatics (peer-reviewed)
- *Percentage of contribution:* 40% (Contributing to the methodologies, implementing all the methods, and contributing to the result analysis)

- *Abstract:*

  Suppose G' is a neighbouring graph of G such that $G' \in \mathcal{N}_G$, where $\mathcal{N}_G$ is the set of all possible neighbouring graphs of graph G that can be obtained by adding, deleting, or reversing a directed edge. We propose two adaptive techniques for faster learning of Bayesian Networks. The first adaptive technique is developed to quickly define the next set of neighborhoods $\mathcal{N}_{G'}$, where $G' \in \mathcal{N}_G$. The technique assigns the set $\mathcal{N}_{G'}$ adaptively based on the obtained $\mathcal{N}_G$, not the entire graph G'. The core idea of the second adaptive technique is as follows. While using the Dirichlet-Multinomial (DM) model for learning parameters in Bayesian networks, the conditional probabilities table for each node becomes large when the allowed number of parents per node grows. Also, calculating the product in Equation 6.4.1 becomes a time-consuming process as the number of nodes increases.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \tag{6.4.1}$$

  where $Pa(X_i)$ is the parent configuration of variable X_i. Suppose G' is a neighbouring graph of G such that $G' \in \mathcal{N}_G$, where $\mathcal{N}_G$ is the set of all possible neighbouring graphs of graph G that can be obtained by adding, deleting, or reversing a directed edge. After we move from the current graph structure G to one of its neighborhood G' by modifying a directed edge $X_i \rightarrow X_j$, where $1 \leq i, j \leq n$, we update only the probability of the affected variable in the new structure, provided that the probabilities of other variables remain with no changes.

- *Reference*:

  Alyami, S., **Azad, A.K.M.**, Keith, JM. "Adaptive techniques for large-scale Bayesian network inference using MCMC methods."

3) *Title*: Discrete Hit-and-Run Markov Chain Algorithm to infer Bayesian Networks

- *Article nature:* Journal article
- *Target journal:* Journal of Statistical Computation and Simulation (peer-reviewed)
- *Percentage of contribution:* 50% (Contributing to the methodologies, implementing all the methods, and contributing to the result analysis)
- *Abstract:*

  We propose a new Markov Chain Monte Carlo (MCMC) approach to sample Bayesian Networks (BNs) from a discrete posterior distribution $f(N, E|D)$ defined on a finite graph space $\mathcal{X}$, where D represents data-points observed at discrete times, N is a set of vertices representing variables, and E is a set of directed edges describing the causal relationships between variables. The new sampler is related to the Hit-and-Run (HAR) sampler that has been shown to converge to a target continuous distributions with low probability of getting "stuck" at local maximum. In this paper, we modify the HAR sampler to generate BNs from discrete spaces. At iteration t, the next iterated graph g_{t+1} is defined by the current graph g_t. We use p_t and ℓ_t to indicate a random path representing a sequence of adjacent graphs and the length of that path, respectively. That is, $|p| \geq 1$. The notations $\ell_t.p_t$ facilitates large movements across the graph-space, which in principle should produce graphs that are less dependent. Our results demonstrate that the modified HAR sampler greatly alleviates the problems caused by local maxima, which in turn facilitates learning the structures of BNs better than the Metropolis-Hasting MCMC sampler.

- *Reference*:

  Alyami, S., **Azad, A.K.M.**, Keith, JM. "Discrete Hit-and-Run Markov Chain Algorithm to Sample Connected Bayesian Networks."

Bibliography

[1] E. Yoruk, M. F. Ochs, D. Geman, and L. Younes. A comprehensive statistical model for cell signaling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):592–606, 2011.

[2] Junhua Zhang, Shihua Zhang, Yong Wang, and Xiang-Sun Zhang. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Systems Biology*, 7(Suppl 2):S4, 2013.

[3] A. K. M. Azad and Hyunju Lee. Voting-based cancer module identification by combining topological and data-driven properties. *PLoS ONE*, 8(8):e70498, 08 2013.

[4] SA. Alyami, A.K.M. Azad, and JM. Keith. Uniform sampling of directed and undirected graphs conditional on vertex connectivity. *Electronic Notes in Discrete Mathematics*, 53:43 – 55, 2016.

[5] SA. Alyami, A.K.M. Azad, and JM. Keith. The neighborhood mcmc sampler for learning bayesian networks. *Proc. SPIE*, 10011:100111K–100111K–11, 2016.

[6] A. K. M. Azad. *Integrating Heterogeneous Datasets for Cancer Module Identification*, pages 119–137. Springer New York, New York, NY, 2017.

Chapter 7

Discussions, Conclusion & Future Works

The main objective of this thesis was to develop computational methods to model and identify cross-talk among signalling pathways, and to comprehensively characterise their roles as underlying mechanisms of acquired resistance. I hypothesised that signalling cross-talk that are potentially contributing to acquired resistance are more likely to be among 'aberrant signalling links' which were defined as signalling dependencies with differential probabilities of appearing in resistant-vs-parental conditions. Therefore, to infer aberrant signalling links in resistant-vs-parental conditions, I combined both computational and Bayesian statistical modelling (p_1-model, a special class of *exponential random graph models*) on data-driven signalling networks in resistant and parental conditions derived from gene-expression datasets of lapatinib-treated ErbB2-positive breast cancer cell-lines: SKBR3 and BT474. First, I proposed a novel approach for categorising signalling cross-talk in data-driven signalling networks and provided pseudocode for detecting those cross-talk among signalling pathways overlayed on those networks [Chapter 3]. I proposed a categorisation into Type-I and Type-II cross-talk that - as I showed - can map all types of cross-talk defined by other state-of-the-art approaches which rely on static topologies of cell signalling

networks and prior biological knowledge [Chapter 3]. Next, I employed a fully Bayesian approach exploiting the p_1-model to infer aberrant signalling links that have *high* posterior probabilities of appearing in the resistant network but *low* posterior probabilities of appearing in the parental network [Chapter 4]. Thereafter, I identified which of those aberrant signalling links form Type-I cross-talk between *EGFR/ErbB2* signalling and other signalling pathways from KEGG [1], Reactome [2], and WikiPathway [3] databases [Chapter 4]. The results suggested that in both SKBR3 and BT474 cell-lines, Notch, Wnt, GPCR, hedgehog, insulin receptor/IGF1R and TGF-β receptor signalling play roles as compensatory pathways [Chapter 4], activation of which can potentially cause up-regulation of EGFR/ErbB2 pathway genes via *signalling cross-talk* thereby attenuating the inhibitory effect of lapatinib (an EGFR/HER2 dual inhibitor). I also applied a similar Bayesian approach but with enhanced data-driven signalling network models of resistant and parental conditions [Chapter 5] using the same datasets: gene expression data from SKBR3 and BT474 cell-lines, and inferred potentially aberrant signalling links with differential posterior probabilities of appearing in resistant-vs-parental networks [Chapter 5]. Next, I conducted an edge set enrichment analysis of these aberrant signalling links comparing them to known signalling links [4] to find dysregulated pathways in acquired resistance [Chapter 5]. The results reported *many* but similar signalling pathways as dysregulated in acquired resistance in both SKBR3 and BT474 cell-lines including EGFR-related pathways and other receptor-related pathways [Chapter 5], many of which were previously reported as compensatory pathways [Chapter 4]. Moreover, analysing a novel pattern of aberrant signalling, called V-structures, I found that dysregulation of crucial genes in resistant-vs-parental conditions which are crucial for breast cancer metastasis and developing acquired resistance to EGFR-TKIs are involved in the switch of dependencies from '*targeted*' to '*bypass*' signalling events [Chapter 5]. These results from analysing putative aberrant signalling [Chapter 5] may provide further insights into the bypass mechanisms of targeted inhibition in cancer therapies.

Signalling rewiring is a significant barrier in maintaining sensitization to drug actions which is crucial for durable RTK-targeted therapies, and failure to which cause acquired drug resistance. As described in previous chapters [Chapter 4 and 5], the mechanisms of signalling rewiring involve transcriptional and post-translational up-regulations of RTKs, over-expression of ABC transporters, reactivation of targeted pathways, and gene over-expressions in effector pathways [Chapter 2]. In this thesis, by analysing co-ordinated differential expression in characterising cross-talk among signalling pathways, we shed light on above aspects of signal rewiring in acquired resistance. For example, in Chapter 4, I demonstrated that via *aberrant* signalling cross-talk, the activation of compensatory pathways (e.g. Notch, Wnt, GPCR, insulin receptor and TGF-β signalling pathways) potentially cause up-regulation of EGFR/ErbB pathway genes, for which the drug resistance occurs. Again, in Chapter 5, I showed that structures of many signalling pathways are rewired in acquired resistance. Again, by analysing a novel structure of aberrant signalling, called V-structures within/among those altered signalling pathways, I highlighted the bypass mechanism of targeted inhibition: *mRNA changes of many crucial genes in resistant-vs-parental conditions were related to the dependency switch from targeted signalling to bypass signalling links in order to avoid drug actions.*

In this thesis, I adapted simple correlation based approach (Pearson Correlation Coefficient) to reconstruct the gene-gene relationship (GGR) network which may only capture linear correlation among genes. Some other approaches such as mutual information and conditional mutual information can be used for the same purpose that can capture non-linear correlation as well. However, in this thesis, my main focus was to develop an approach which can study the statistical aspects of gene-gene relationship networks that can be reconstructed in any of the available tools.

In general, the levels of noise in the data, heterogeneity of different databases, and the extent of vagueness underlying different network inference methods, mean that the roles of genes or gene-gene interactions in drug escape can be very complicated. However, in

this thesis I aimed to address those issues by utilising the statistical modelling approach (i.e. p_1-model), which measures the probabilistic nature of pair-wise relationships in a gene-gene relationship network.

A key point about the p_1-model is that dyads in the given network structure are assumed to be *statistically independent* [5]. Since X is assumed to be a random matrix (matrix-valued random variable) defined on a network space $\mathcal{G}$, the consideration of dyadic independence helped to retain the tractability of the model for larger networks [5, 6]. More specifically, if dyads $D_{ij} = (X_{ij}, X_{ji})$ for $i < j$ are statistically independent then the distribution of X can be specified as their joint distribution [5], where we only have to specify the distribution of each dyad [see Appendix B of Chapter 4 for the detailed derivation of the p_1-model]. Hence, the p_1-model cannot deal with dyadic dependance such as transitivity, cliquing, and hierarchy [5]. But, biological networks often show dyadic dependencies: for example, Dougherty *et al.* [7] and Kolch *et al.* [8] reported that *transitive* signalling dependencies Raf→Akt, Raf→Erk, and Mek→Akt play important feedback roles in a well known signalling cascade Raf→Mek→Erk→Akt [9]. Therefore, direct application of the p_1-model may not be appropriate for such networks. However, in this thesis, before applying the p_1-models to data-driven models of signalling networks I defined them in such a way that they contain: 1) all *pair-wise* direct [Chapters 4 and 5] and 2) indirect relationships [Chapter 5] through which I may overcome the above limitation. I hypothesised that the use of data-driven modelling of networks by exploiting all pair-wise *direct* relationships (e.g. correlation coefficient and mutual information) and *indirect* relationships may implicitly capture the dyadic dependencies. More specifically, in Chapters 4 and 5 I used *all pair-wise* correlation calculations for modelling pair-wise *direct* relationships and thus I assume that dyadic dependencies can be implicitly captured in correlated node-pairs. For example, in the case of *transitive* dyadic dependencies, suppose A is correlated with B, and B is correlated with C. Now if A-B and B-C pairs are dependent upon each other (dyadic dependency [6]), then A and C should be correlated, and thus a dyad could be formed

in the network as a direct relationship since I explore *all pair-wise* correlations. Using the definition of indirect relationships given in Chapter 5, they can also model *transitive* dyadic dependencies which involve some intermediate *linker* nodes. For instance, in the above example if A and C are not directly correlated (i.e. their correlation value is less than the threshold [Chapters 4 and 5]), then the search for any indirect relationships between them by identifying statistically significant simple paths (in PPI network), will detect some transitive dyadic dependencies involving intermediate linker genes with the help of PPI information [Chapter 5]. Thus, I overcome a technical limitation of the p_1-model for larger networks with dyadic dependencies while exploiting its simplicity.

The analyses in this thesis open up some future research directions. First, many approaches including this study use only gene expression data as a proxy for signalling protein expression (i.e. protein activities) [10–12] in order to model the signalling pathway activities, whereas, gene expression only reflects the downstream effect of phosphoprotein (signalling protein) activities and may not directly correlate with the upstream protein expression [12]. Thus, to have better modelling of signalling cross-talk in a mechanistic way, protein expressions measuring the signalling activities of phosphoproteins of upstream signalling pathways need to be considered. In my future studies, I would like to develop a methodological framework to integrate phosphoproteomic datasets (i.e. expression of phosphoproteins) with transcriptomic datasets (i.e. gene expression) to investigate mechanistic details of acquired resistance. By using this framework, one of the key research questions that I aim to answer is: how do signalling cross-talk among upstream signalling pathways mediate downstream differential gene expression, and thereby elucidate novel mechanisms of acquired resistance to RTK inhibitors?

MCMC sampling approaches such as the Neighbourhood sampler [13], Hit-and-Run sampler [14], and Metropolis-Hastings sampler [15] can be used for better structural inference of data-driven signalling networks (i.e. gene-gene relationship networks [see Chapters 4 and 5]) prior to the application of the p_1-model. The approach in this

thesis constructs such networks using simple pair-wise correlations, considering the signalling networks as *undirected* networks. However, signalling activities are better explained when they are modelled using causal relationships, thus forming a *directed* network for which Bayesian Network models are very commonly used [16–18]. Some of the MCMC sampling methods that I introduced in Chapter 6 can be used in order to infer the data-driven signalling network structure considering them as Bayesian Network models. One of the primary challenges in using these techniques is sampling *large-scale* networks with high accuracy within *feasible* time and memory constraints [for details see original articles mentioned in Chapter 6].

Third, it may be possible to analyse additional *local* features that help to describe *global* network structure, and include them in the probability function (Equation 2.3.1) as additional *explanatory variables*, modelled by ERGMs (Exponential Random Graph Models). The rationale is: one of the useful property of ERGMs is that the probability of a given network structure depends on a set of locally determined explanatory variables. In these models new explanatory variables can be included, which can be any *graph statistic* such as number of triangles or other subgraphs [19]. For example, the results in Chapter 5 suggested that in both SKBR3 and BT474 cell-line, higher percentages of dysregulated genes were incident with V-structures (a *red* and a *green* pairs intersect at a shared node) compared to *non* V-structures (two *red* or two *green* pairs intersect at a shared node) in structural rewiring of signalling networks. Hence, this local feature (i.e. *the number of V-structures*) can be integrated into the probability function (Equation 2.3.1) as an explanatory variable to influence the global structure rewiring of a signalling network.

Finally, I would be keen to investigate more how to prioritise these putative findings from my framework so that it could be followed up by wet-lab experimental validations. The rationale is that one of the inherent features of data-driven approaches in systems biology is that they produce lots of putative hits which may require additional pruning before conducting any further experiments on them.

In this study, I developed computational frameworks that systematically predict all possible cross-talk among signalling pathways residing in data-driven signalling networks by analysing resistant-vs-parental datasets. The results uncovered some novel pathways as dysregulated and compensatory to targeted inhibition (e.g. EGFR/ErbB2 signalling pathway in cells treated with lapatinib) which may have the potential to be used as novel drug targets in combination with targeted pathways. Although the methodological framework is demonstrated here with two lapatinib-treated ErbB2-positive breast cancer cell-lines: SKBR3 and BT474, this approach is applicable to other cell-line datasets to elucidate novel mechanisms of acquired resistance to other RTK inhibitors.

Bibliography

[1] M. Kanehisa. The KEGG database. *Novartis Found. Symp.*, 247:91–101, 2002.

[2] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Res.*, 42(Database issue):D472–477, Jan 2014.

[3] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, 40(Database issue):D1301–1307, Jan 2012.

[4] Human signalling network, version 6. URL http://www.cancer-systemsbiology.org/dataandsoftware.htm.

[5] Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[6] Xiaolin Yang, Alessandro Rinaldo, and Stephen E. Fienberg. Estimation for dyadic-dependent exponential random graph models. *Journal of Algebraic Statistics*, 5 (1), apr 2014. doi: 10.18409/jas.v5i1.24.

[7] Michele K. Dougherty, Jürgen Müller, Daniel A. Ritt, Ming Zhou, Xiao Zhen Zhou, Terry D. Copeland, Thomas P. Conrads, Timothy D. Veenstra, Kun Ping Lu, and Deborah K. Morrison. Regulation of raf-1 by direct feedback phosphorylation. *Molecular Cell*, 17(2):215–224, jan 2005.

[8] W. Kolch. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. *Biochem. J.*, 351 Pt 2:289–305, Oct 2000.

[9] Mitchell Koch, Bradley M. Broom, and Devika Subramanian. Learning robust cell signalling models from high throughput proteomic data. *International Journal of Bioinformatics Research and Applications*, 5(3):241, 2009.

[10] A. Amadoz, P. Sebastian-Leon, E. Vidal, F. Salavert, and J. Dopazo. Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. *Sci Rep*, 5:18494, Dec 2015.

[11] P. Sebastian-Leon, J. Carbonell, F. Salavert, R. Sanchez, I. Medina, and J. Dopazo. Inferring the functional effect of gene expression changes in signaling pathways. *Nucleic Acids Res.*, 41(Web Server issue):W213–217, Jul 2013.

[12] R. Neapolitan, D. Xue, and X. Jiang. Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks. *Cancer Inform*, 13:77–84, 2014.

[13] Jonathan Keith, George Sofronov, and Dirk Kroese. *The Generalized Gibbs Sampler and the Neighborhood Sampler*, pages 537–547. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[14] Robert L. Smith. The hit-and-run sampler: A globally reaching markov chain sampler for generating arbitrary multivariate distributions. In *Proceedings of the*

28th Conference on Winter Simulation, WSC '96, pages 260–264. IEEE Computer Society, 1996.

[15] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[16] D. Pe'er. Bayesian network analysis of signaling networks: a primer. *Sci. STKE*, 2005(281):pl4, Apr 2005.

[17] S. M. Hill, Y. Lu, J. Molina, L. M. Heiser, P. T. Spellman, T. P. Speed, J. W. Gray, G. B. Mills, and S. Mukherjee. Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics*, 28(21):2804–2810, Nov 2012.

[18] K Sachs. *Bayesian network models of biological signaling pathways*. PhD thesis, Massachusetts Institute Of Technology, Biological Engineering Department, July 2006.

[19] Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, Oct 2007.

Appendix A

List of Abbreviations (most commonly used terms)

DAVID : Database for Annotation, Visualisation and Integrated Discovery

EGFR : Epidermal Growth Factor Receptor

ERGM : Exponential Random Graph Model

ESEA : Edge Set Enrichment Analysis

GATHER : Gene Annotation Tool to Help Explain Relationships

GGR : Gene-Gene Relationship

HER2 : Human Epidermal Growth Factor Receptor 2

MCMC : Markov Chain Monte Carlo

MPLE : Maximum Pseudolikelihood Estimation

PAGI : Pathway Analysis based on Global Influence

PPI : Protein-Protein interaction

RTK : Receptor Tyrosine Kinase

SCCA : Sparse Canonical Correlation Analysis

sGSCA : Signature-based Gene-Set Co-expression Analysis

SPIA : Signalling Pathway Impact Analysis

TF : Transcription Factor

TKI : Tyrosine Kinase Inhibitor

Appendix B

Appendix to Chapter 4

Additional File 1: Prediction of Signaling Cross-talks Contributing to Acquired Drug Resistance in Breast Cancer Cells by Bayesian Statistical Modeling

A. K. M. Azad[1,*], Alfons Lawen[2], Jonathan.Keith[1]
[1]School of Mathematical Science, Monash University
[2]Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University
[*]E-mail: aaza7@student.monash.edu

Appendix I: Derivation of p_1-model for directed network

Let $\mathbf{X}$ be a directed network with n nodes and a realization of of that network is represented as $\mathbf{X} = \mathbf{u}$. Let the binary outcome $u_{ij} = 1$ if $gene_i$ interacts with $gene_j$, or $u_{ij} = 0$ otherwise. Then $\mathbf{u}$ is a binary data matrix [1]. Let $Pr(\mathbf{u})$ be the probability function on G given by

$$Pr(u) = Pr(\mathbf{X} = \mathbf{u}) = \frac{1}{\kappa(\boldsymbol{\theta})} exp \sum_p \boldsymbol{\theta}_p z_p(\mathbf{u}) \tag{1}$$

where $z_p(\mathbf{u})$ is the network statistic of type p, $\boldsymbol{\theta}_p$ is the parameter associated with $z_p(\mathbf{u})$ and $\kappa(\boldsymbol{\theta})$ is the normalizing constant that ensures $Pr(\mathbf{u})$ is a proper probability distribution (sums to 1 over all $\mathbf{u}$ in G) [2]. The parameter $\boldsymbol{\theta}$ is a vector of model parameters associated with network statistics and needs to be estimated. See [3] for further details.

Derivation: The p_1-model considers the joint distribution of dyads $D_{ij} = (u_{ij}, u_{ji})$ with dyadic probabilities

$$m_{ij} = Pr(mutual \quad dyad) = Pr\{D_{ij} = (1,1)\}; \quad \forall (i < j), \tag{2}$$

$$a_{ij} = Pr(asymmetric \quad dyad) = Pr\{D_{ij} = (1,0)\}; \quad \forall (i \neq j), \tag{3}$$

$$n_{ij} = Pr(null \quad dyad) = Pr\{D_{ij} = (0,0)\}; \quad \forall (i < j), \tag{4}$$

and

$$m_{ij} + a_{ij} + a_{ji} + n_{ij} = 1; \quad \forall (i < j). \tag{5}$$

This model finds the probabilities of each type of dyadic relation for each pair of genes. Assuming all the dyads D_{ij} are statistically independent, the probability distribution of $\mathbf{X} = \mathbf{u}$ can be specified as the joint distribution of the dyads (such as, D_{12}, D_{13}, and so on) which may be expressed in the following way:

$$Pr(\boldsymbol{X} = \boldsymbol{u}) = \prod_{i<j} m_{ij}^{u_{ij}u_{ji}} \prod_{i \neq j} a_{ij}^{u_{ij}(1-u_{ji})} \prod_{i<j} n_{ij}^{(1-u_{ij})(1-u_{ji})} \tag{6}$$

In order to get an exponential form, the above equation can be reexpressed as follows:

$$Pr(\boldsymbol{X} = \boldsymbol{u}) = exp\left\{ \sum_{i<j} \rho_{ij} u_{ij} u_{ji} + \sum_{i \neq j} \theta_{ij} u_{ij} \right\} \prod_{i<j} n_{ij}, \tag{7}$$

where

$$\rho_{ij} = \log\left(\frac{m_{ij} n_{ij}}{a_{ij} a_{ji}} \right); \quad \forall (i < j) \tag{8}$$

and

$$\theta_{ij} = \log\left(\frac{a_{ij}}{n_{ij}}\right); \quad \forall\,(i \neq j) \tag{9}$$

Note that we interpret $n_{ij} = n_{ji}$ for i > j. The parameter ρ_{ij} is a log-odds ratio which measures the force of reciprocation of the edge between $gene_i$ and $gene_j$. By doing simple algebra as follows, it can be said that ρ_{ij} specifies the log of increase in the odds that $u_{ij} = 1$ given $u_{ji} = 1$

$$exp\,(\rho_{ij}) = \left\{\frac{Pr\,(u_{ij} = 1|u_{ji} = 1)}{Pr\,(u_{ij} = 0|u_{ji} = 1)}\right\} / \left\{\frac{Pr\,(u_{ij} = 1|u_{ji} = 0)}{Pr\,(u_{ij} = 0|u_{ji} = 0)}\right\} \tag{10}$$

The parameter θ_{ij}, is also a log-odds ratio which measures the probability of an asymmetric edge between $gene_i$ and $gene_j$ with $u_{ji} = 0$. This intuition can be explained by the following calculation:

$$exp\,(\theta_{ij}) = \frac{Pr\,(u_{ij} = 1|u_{ji} = 0)}{Pr\,(u_{ji} = 0|u_{ij} = 0)} \tag{11}$$

Equation (7) provides a more general family of distributions for $\mathbf{X}$ than Equation (1). However, Equation (7) contains too many parameters; therefore, restrictions were applied on the parameters, ρ_{ij} and θ_{ij} to obtain Equation (1) from Equation (7). Thus, the original p_1-model was postulated as following:

$$\rho_{ij} = \rho; \quad \forall\,(i < j)\,, \tag{12}$$

and

$$\theta_{ij} = \theta + \alpha_i + \beta_j; \quad \forall\,(i \neq j)\,, \tag{13}$$

Here, ρ indicates the global degree of reciprocity of the entire network; θ is the global density parameter; α_i is a local parameter measuring the *expansiveness* of $gene_i$ which is the propensity of $gene_i$ to send edges; and β_j represents the *attractiveness* of $gene_j$ which is the ability of $gene_j$ to attract edges. Based on the transformation of parameters in Equations (12) and (13), the distribution formula for the p_1-model can be rewritten as follows:

$$Pr\,(\boldsymbol{X} = \boldsymbol{u}) = exp\left\{\sum_{i<j} \rho M + \theta E + \sum_i \alpha_i \triangle_{out}\,(i) + \sum_j \beta_j \triangle_{in}\,(j)\right\} \prod_{i<j} n_{ij} \tag{14}$$

The above form of the p_1-model equation represents the exponential family of distributions with the following statistics: M - the number of reciprocated edges, E - total number of edges, and $\triangle_{out}\,(i)$ and $\triangle_{in}\,(i)$ - the in- and out-degree of $gene_i$. To facilitate Gibbs sampling, an equivalent log-linear formulation of the p_1-model was suggested by Fienberg and Wasserman [4]. In this formulation, a dyad (u_{ij},u_{ji}) is represented by four Bernoulli variables Y_{ij00}, Y_{ij10}, Y_{ij01} and Y_{ij11} as follows:

$$Y_{ijkl} = \begin{cases} 1 & if \quad u_{ij} = k, u_{ji} = l, \\ 0 & otherwise \end{cases}$$

Then, the p_1-model can be expressed with four log-linear equations:

$$log\,\{Pr\,(Y_{ij10} = 1)\} = \lambda_{ij} + \theta + \alpha_i + \beta_j \tag{15}$$

$$log\,\{Pr\,(Y_{ij01} = 1)\} = \lambda_{ij} + \theta + \alpha_j + \beta_i \tag{16}$$

$$log\,\{Pr\,(Y_{ij11} = 1)\} = \lambda_{ij} + 2\theta + \alpha_i + \alpha_j + \beta_i + \beta_j + \rho \tag{17}$$

$$log\,\{Pr\,(Y_{ij00} = 1)\} = \lambda_{ij} \tag{18}$$

for $i < j$. Here, $\lambda_{ij} = log\,(n_{ij})$ is the scaling parameter, which is fixed due to the constraint $\sum_{k,l} Y_{ijkl} = 1$.

References

1. Bulashevska, S., Bulashevska, A., Eils, R.: Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. BMC Bioinformatics **11**, 46 (2010)

2. Wasserman, S., Pattison, P.: Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. Psychometrika **61**(3), 401–425 (1996)

3. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association **76**(373), 33–50 (1981)

4. Fienberg, S., Wasserman, S.: Categorical data analysis of single sociometric relations. Sociological Methodology **12**, 156–192 (1981)

1 Supplementary Figures

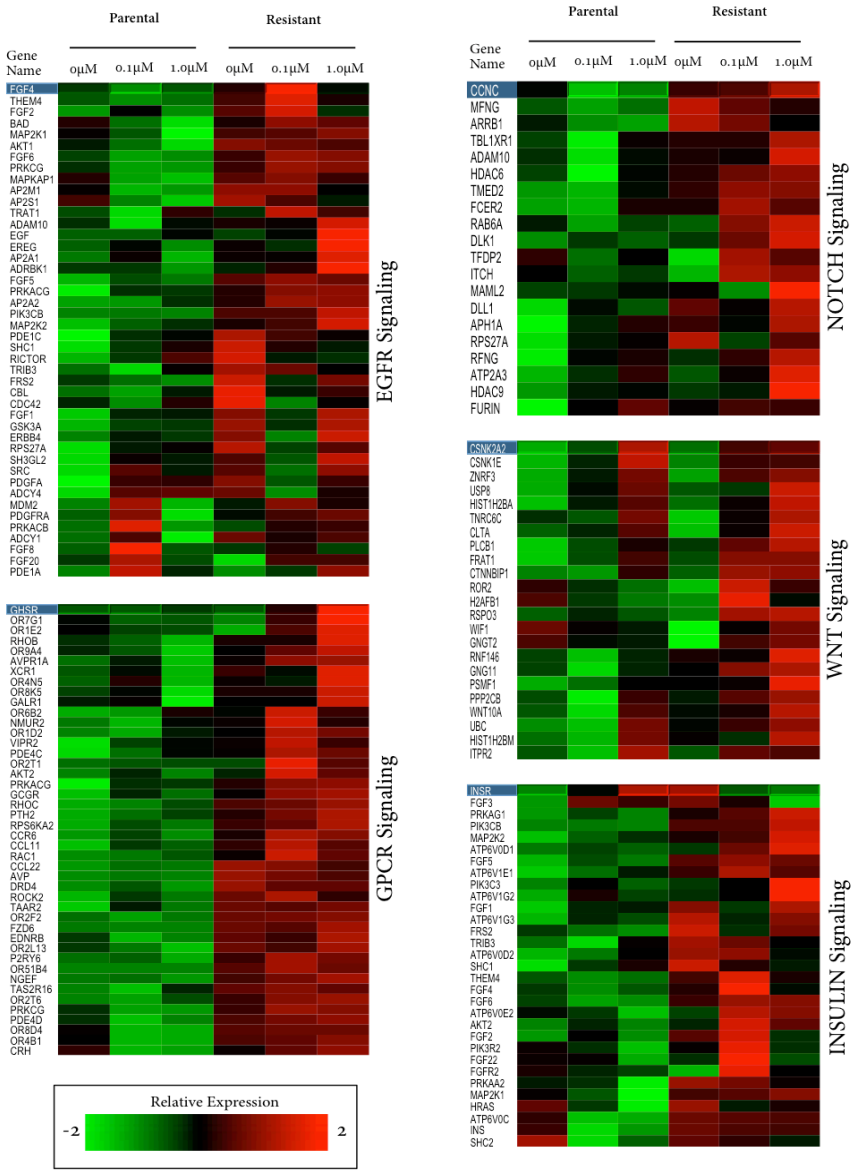


Figure 1. Comparative expression changes in parental and resistant conditions (SKBR3 cell-line, GSE38376) of some constituent genes of EGFR, GPCR, Notch, Wnt and insulin signaling. Expression of these genes in parental conditions is down-regulated but up-regulated in resistant conditions which signify the effect of drug resistance on those genes. Here, pathway annotations are from Reactome database.

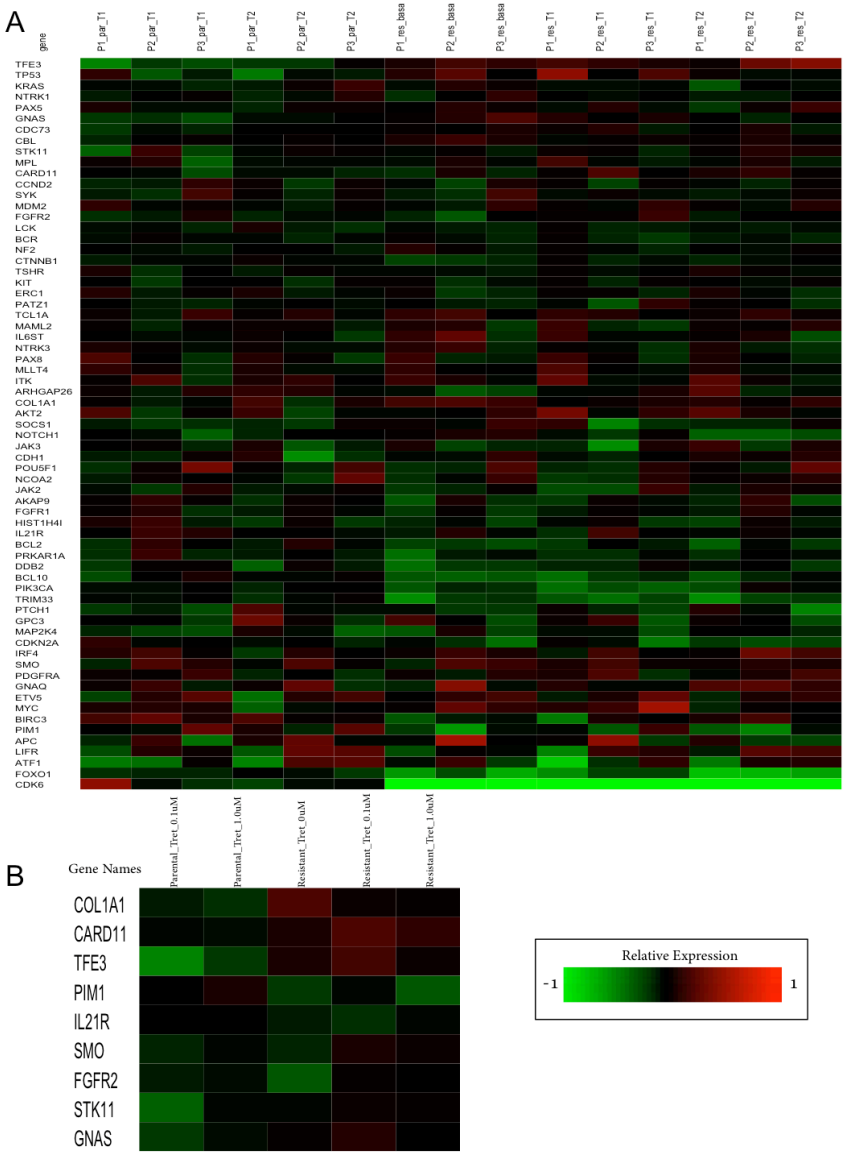


Figure 2. Fold change analysis of gene expressions in both parental and resistant conditions compared to parental basal condition (0 μM) in our primary dataset (SKBR3 cell-line, GSE38376). (A) Genes depicted here are from 104, 188 and 299 EGFR/ErbB cross-talks found using signaling pathways from Reactome, KEGG and WikiPathway databases, respectively **(B)** List of those genes that are dysregulated in parental treatment vs parental basal condition and reversely changed in resistant basal + resistant treatment vs parental basal condition. Each column here represents each condition's gene expression fold-change of either, one of the parental treatment conditions (0.1 μM, 1.0 μM), or one of the resistant conditions (0 μM, 0.1 μM, 1.0 μM) compared to parental basal treatment condition (0 μM).

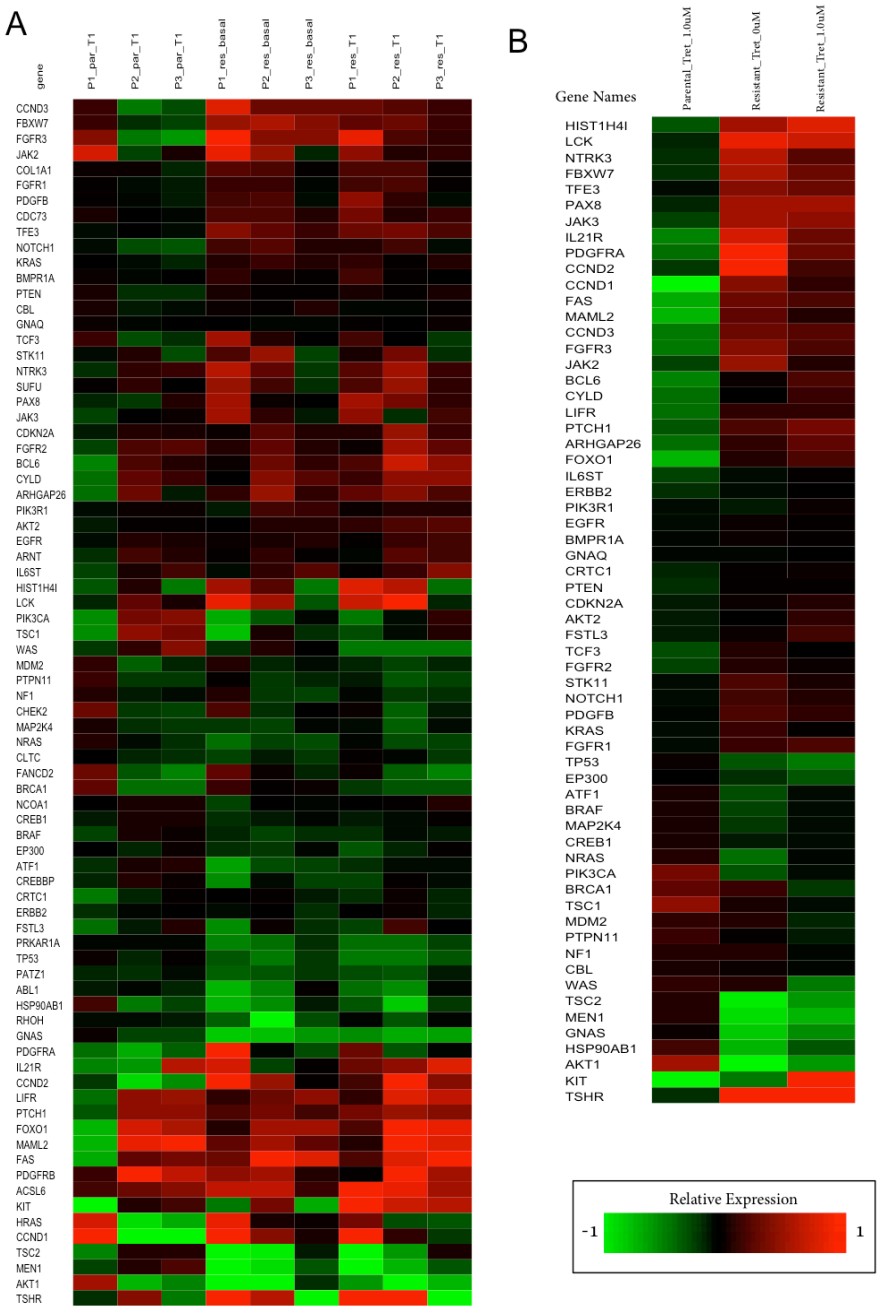


Figure 3. Fold change analysis of gene expressions in both parental and resistant conditions compared to parental basal condition (0 μM) in our validation dataset (BT474 cell-line, GSE16179). (A) Genes depicted here are from 83, 133 and 278 EGFR/ErbB cross-talks found using signaling pathways from Reactome, KEGG and WikiPathway databases, respectively **(B)** List of those genes that are dysregulated in parental treatment vs parental basal condition and reversely changed in resistant basal + resistant treatment vs parental basal condition. Each column here represents each condition's gene expression fold-change of either, one of the parental treatment conditions (0.1 μM, 1.0 μM), or one of the resistant conditions (0 μM, 0.1 μM, 1.0 μM) compared to parental basal treatment condition (0 μM).

Table S1: All 11,515 drug-resistant gene-pairs found in GSE38376.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM2_ESM.xlsx

Table S2: All 1,083 (841 distinct) cross-talks found between EGFR and other 23 signaling pathways from Reactome database.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM3_ESM.xlsx

Table S3: All 2,179 (1,050 distinct) cross-talks found between ErbB and other 34 signaling pathways from KEGG database.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM4_ESM.xlsx

Table S4: All 3,084 (876 distinct) cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM5_ESM.xlsx

Table S5: 104 drug-resistant cross-talks found between EGFR and other 23 signaling pathways from Reactome database [GSE38376].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM6_ESM.xlsx

Table S6: 188 drug-resistant cross-talks found between ErbB and other 34 signaling pathways from KEGG database [GSE38376].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM7_ESM.xlsx

Table S7: 299 drug-resistant cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database [GSE38376].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM8_ESM.xlsx

Table S8: 168 selected cross-talks which associated EGFR (or ErbB) signaling pathways with 6 other signaling pathways that were found in at least two different pathway analyses (i.e. Reactome and KEGG, or KEGG and WikiPathway, or Reactome and WikiPathway) [GSE38376].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM9_ESM.xlsx

Table S9: All 10,811 drug-resistant gene-pairs found in GSE16179.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM10_ESM.xlsx

Table S10: 83 drug-resistant cross-talks found between EGFR and other 23 signaling pathways from Reactome database [GSE16179]..

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM11_ESM.xlsx

Table S11: 133 drug-resistant cross-talks found between ErbB and other 34 signaling pathways from KEGG database [GSE16179].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM12_ESM.xlsx

Table S12: 278 drug-resistant cross-talks found between ErbB and other 62 signaling pathways from WikiPathway database [GSE16179].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM13_ESM.xlsx

Table S13: 86 drug-resistant cross-talks found in all Reactome, KEGG and WikiPathway analyses where both genes in a particular cross-talk was up-regulated in resistant conditions but down-regulated in parental conditions [GSE16179].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM14_ESM.xlsx

Table S14: 401 cross-talks from Reactome, KEGG and WikiPathway analyses where the genes are dysregulated in parental treatment vs parental basal condition, and reversely changed in resistant basal + resistant treatment vs parental basal condition [GSE16179].

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM2_ESM.xlsx

Table S15: All 11,515 drug-resistant gene-pairs found in GSE38376.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12918-014-0135-x/MediaObjects/12918_2014_135_MOESM15_ESM.xlsx

Appendix C

Appendix to Chapter 5

Supplementary Text: Bayesian Model of Signal Rewiring Reveals Mechanisms of Gene Dysregulation in Acquired Drug Resistance in Breast Cancer

A. K. M. Azad[1,*], Alfons Lawen[2], Jonathan.Keith[1]
[1]School of Mathematical Science, Monash University
[2]Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University
[*]E-mail: aaza7@student.monash.edu

Supplementary Methods

Parameter Selection

In this work, we used the following thresholds: *corrected* p-value thresholds for selecting differentially expressed genes, threshold for path length in the search of indirect relationships, cutoff thresholds for selecting direct relationships, odds ratio, and posteriori values. Detailed explanations and justifications behind these threshold selections are provided bellow. All other thresholds are p-value cutoffs for which a conventional value of 0.05 was used.

Selecting two different p-value thresholds for detecting differentially expressed genes in SKBR3 and BT474 cell-lines

For both SKBR3 and BT474 cell-lines, differentially expressed (DE) genes were selected based on Bonferroni-corrected p-values from a two-tailed pooled Students t-test. Genes that showed differential expression with corrected p-values $\leq$ threshold were selected as DE genes. By analysing gene expression datasets from parental and resistant conditions we identified 345 and 354 DE genes, for SKBR3 and BT474 cell-lines, respectively. The corresponding threshold values for corrected p-values were 0.01 and 0.05, respectively. It is true this implies SKBR3 DE gene selection was more stringent than that of BT474. This was done for two reasons: firstly, because the computational cost of using a conventional threshold of 0.05 with SKBR3 was prohibitive, and secondly, to ensure the numbers of DE genes in the two different cell-lines were comparable, and similarly for the sizes of the seed gene sets. Bayesian inference of model parameters (i.e. posterior probabilities of network edges) using MCMC sampling was the most time-consuming step in our whole framework. This step requires a longer period of time (considering the configurations of MCMC sampling) as the network size grows (i.e. as the number of seed genes increases). In this study, 0.01 and 0.05 p-value thresholds yielded 345 and 354 DE genes, and eventually 897 and 875 seed genes in total, for SKBR3 and BT474 cell-lines, respectively. We found that with these sizes of networks (both parental and resistant GGR networks) from SKBR3 and BT474 cell-lines, execution of the whole Bayesian inference procedure including MCMC sampling for 15,000 iterations and summarizing the monitored parameter (i.e. posterior probabilities of all node-pairs: of which there were 402,753 and 382,375 for SKBR3 and BT474 cell-lines, respectively) required more than one week, individually [data not shown]. Therefore, we feared that less stringent thresholds for SKBR3 (e.g. 0.05 yielding 135 additional DE genes) would result in an even larger GGR network for which the whole Bayesian inference step would run even longer. Therefore, we preferred to use the p-value threshold 0.01 for SKBR3 in order to produce a comparable number of DE genes, and thus a similar number of seed genes, so that we have the whole Bayesian inference procedure completed within a manageable time frame. Moreover, even after using a stringent threshold for SKBR3 cell-line our framework was able to identify high percentages of dysregulated pathways in acquired resistance: 75.56% (34 out of 45), 62.5% (15 out of 24) and 68.85% (42 out of 61) signalling pathways from KEGG, Reactome, and WikiPathways, respectively [S2 Table]. These results indicate that our framework demonstrated high performances in SKBR3 cell-line even with a more stringent p-value threshold. Therefore, we hypothesise that including a larger number of DE genes

by relaxing the threshold would yield little performance increase in terms of detecting aberrant signalling pathways in acquired resistance.

Selecting threshold for direct pair

Choosing a threshold for a co-expression network is a crucial step in network-based analysis since the representation of network structure, its functional relevance, and any network-based discoveries depend on the cutoff that is applied to all pair-wise co-expression values [1]. In this work, we applied a systematic approach proposed by Elo *et al.* [1] to identify an optimal cutoff threshold of such co-expression pairs (that is, *direct pairs*) by analysing the topological properties of a co-expression network. The approach is reported to achieve a balance between detecting as many biologically relevant co-expression links as possible, and controlling the *false-negative rates* [1]. It compares the clustering coefficient of the observed co-expression network and that of its randomized counterpart at a particular cutoff threshold point [1]. The clustering coefficient of a network is defined as follows:

$$C = \frac{1}{K} \sum_{k_i > 1} C_i \tag{1}$$

where $C_i = \frac{2E_i}{k_i(k_i-1)}$ is the clustering coefficient of the node (gene) i, E_i denotes the number of edges between $k_i(>1)$ first neighbours of node (gene) i [1]. Elo *et al.* hypothesised that the co-expression links omitted from the complete network by increasing the cutoff threshold value are more likely to be noise as long as the difference between the clustering coefficient of the observed network and its randomised counterpart is monotonically increasing [1]. Thus, a discrete optimization problem was formulated to find the optimal threshold C^* as follows:

$$C^* = \min_j \{r_j : C(r_j) - C_0(r_j) > C(r_{j+1}) - C_0(r_{j+1})\} \tag{2}$$

where the set of thresholds is: $r_0 < r_1 < ... < r_{J-1} < r_J$, $C(r)$ denotes the clustering coefficient of the co-expression network generated by applying the co-expression threshold r, and $C_0(r)$ is its randomised counterpart [1]. Here, $r_0 = 0$, $r_J = 1$, and $r_{j+1} = r_j + 1$ [1]. Elo *et al.* applied a configuration model in order to preserve the original degree distribution of the observed network [for details see Methods section of [1]]. Thus, the value of C_0 was formulated as follows:

$$C_0 = \frac{(\bar{k^2} - \bar{k})^2}{\bar{k}^3 N} \tag{3}$$

where $\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i$, $\bar{k^2} = \frac{1}{N}\sum_{i=1}^{N} k_i^2$, and N = total number of nodes in the network [1]. This procedure define C^* to be the first *local maxima* in the $C - C_0$ curve [1]. In our study, by modelling *signalling rewiring* in resistant-vs-parental conditions, we identify system-level perturbations in the resistant condition compared to the parental condition. Consequently, we choose the threshold value for the parental condition by comparing to a random reference network and use the same number of pairs for the resistant condition in order to make a fair comparison. For each of the parental gene expression data sets in SKBR3 and BT474 cell-lines, we applied the above approach to identify an appropriate value of C^*. The resulting values of C^* were 0.62 and 0.74, which demarcate approximately the top 20% of pairs from the respective distributions of co-expression values in both SKBR3 and BT474 cell-lines. Therefore, we selected the top 20% of pairs to be considered as direct pairs in our analysis for all the GGRs. This reasoning has been added into the supplementary text.

Selecting thresholds for odds and posteriori

For selecting the thresholds of odds ratios (of posterior probabilities [see Methods in the original text]) of gene-pairs in all four cases: SKBR3-Parental, SKRB3-Resistant, BT474-Parental, and BT474-Resistant,

we made frequency distributions of all odds ratio values, respectively, from which we chose the top 20% gene-pairs, with odds ratio values of 1.66, 2.53, 2.16, 12.03 as cutoff thresholds ('*th_odds*'), respectively [Supplementary Figure 2]. Next, we constructed distributions of posteriori values of those selected gene-pairs, respectively, and chose the top 50% of gene-pairs from them with posteriori values of 0.252, 0.212, 0.304, and 0.177 as cutoff thresholds, respectively ('*th_posteriori*') [Supplementary Figure 3]. Although these two types of thresholds were chosen *empirically* from their respective distributions, applying them yielded two important outcomes. Firstly, for both SKBR3 and BT474 cell-lines, two sets of selected pairs from the top 20% of *parental* and *resistant* distributions [Supplementary Figure 2], respectively were *completely disjoint*, which is consistent with our methodological requirements that *red* and *green* [for definitions see original texts] aberrant pairs should be *mutually exclusive*. Secondly, mixtures of two distributions (assumed to be two clusters: left-cluster shown in *green* color and right-cluster shown in *violet* color) were clearly apparent in all individual scatter plots of *ODDs* VS *posterior probabilities* of gene pairs after applying '*th_odds*' in all four cases: SKBR3-Parental, SKRB3-Resistant, BT474-Parental, and BT474-Resistant [Supplementary Figure 4], and the top 50% pairs (from their frequency distribution) were residing in the right cluster (violet color) [Supplementary Figure 4] which is assumed to contain more important pairs than the left cluster (green color) because the posterior probability values of pairs in the former distribution are greater than in those of the latter.

Path length threshold in indirect relationships

In the search for indirect pairs for which no direct relationships were found, we applied an approach which exploits statistically significant PPI paths [see Methods in the original text] for which we constrained the length of those paths to be 2. We hypothesise that increasing this length threshold would impose additional computational costs and increase the time-complexity of the framework. We therefore chose the value of such path lengths to be 2, which is the minimum possible value involving single *linker genes* [see Methods in the original text].

References

1. Elo LL, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics. 2007;23(16):2096–2103.

1 Supplementary Figures

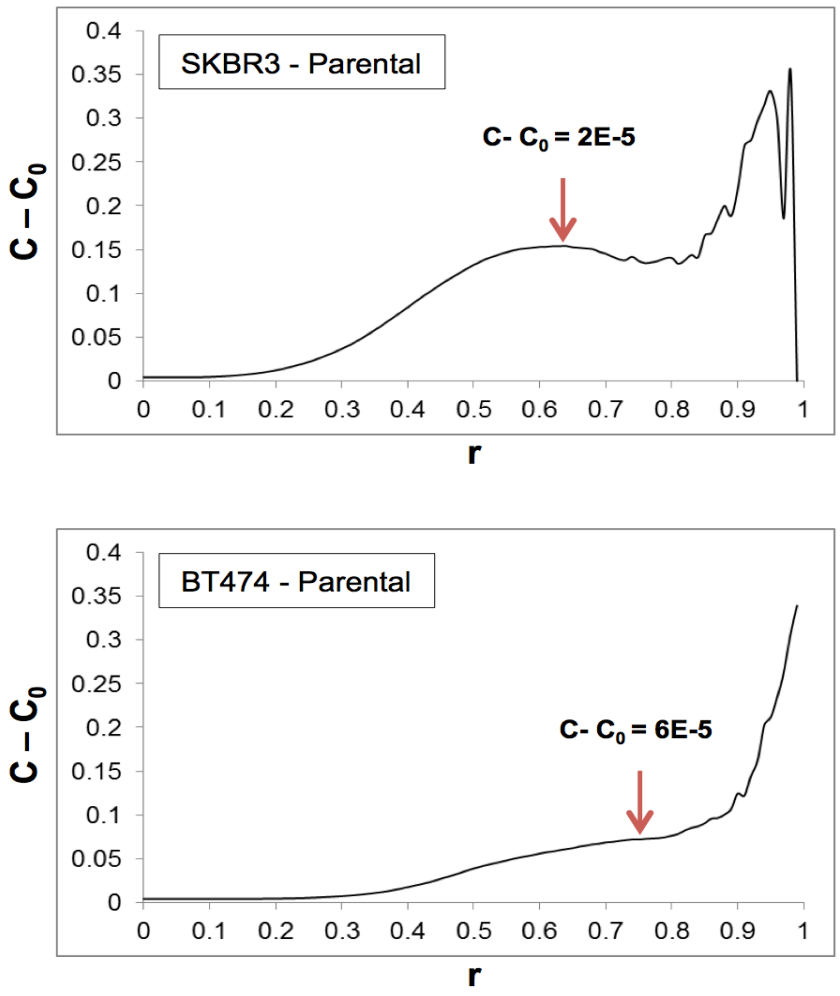


Figure 1. Behaviour of clustering coefficients (C) of observed co-expression networks against their randomized counterparts (C_0) at various co-expression value (absolute correlation), r. For each gene expression dataset, first *local maxima* in the $C - C_0$ is shown at the continuous case with red arrow.

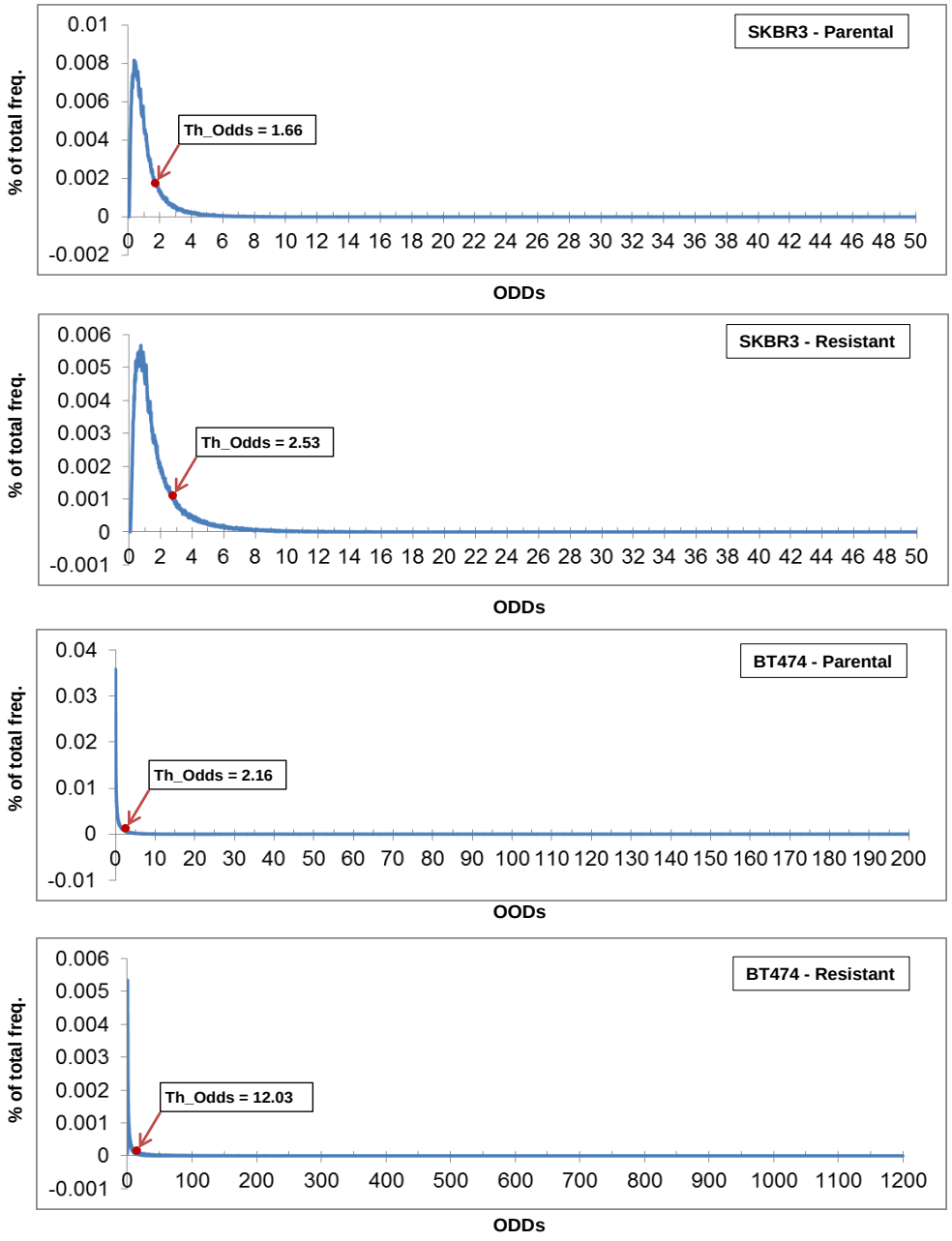


Figure 2. Frequency distributions of odds ratio values of gene-pairs in all four cases: SKBR3-Parental, SKBR3-Resistant, BT474-Parental, and BT474-Resistant after Bayesian analysis. In all four cases, selected thresholds of odds ratios, '*th_odds*' are shown with arrows which demarcated top 20% of pairs in their respective distributions.

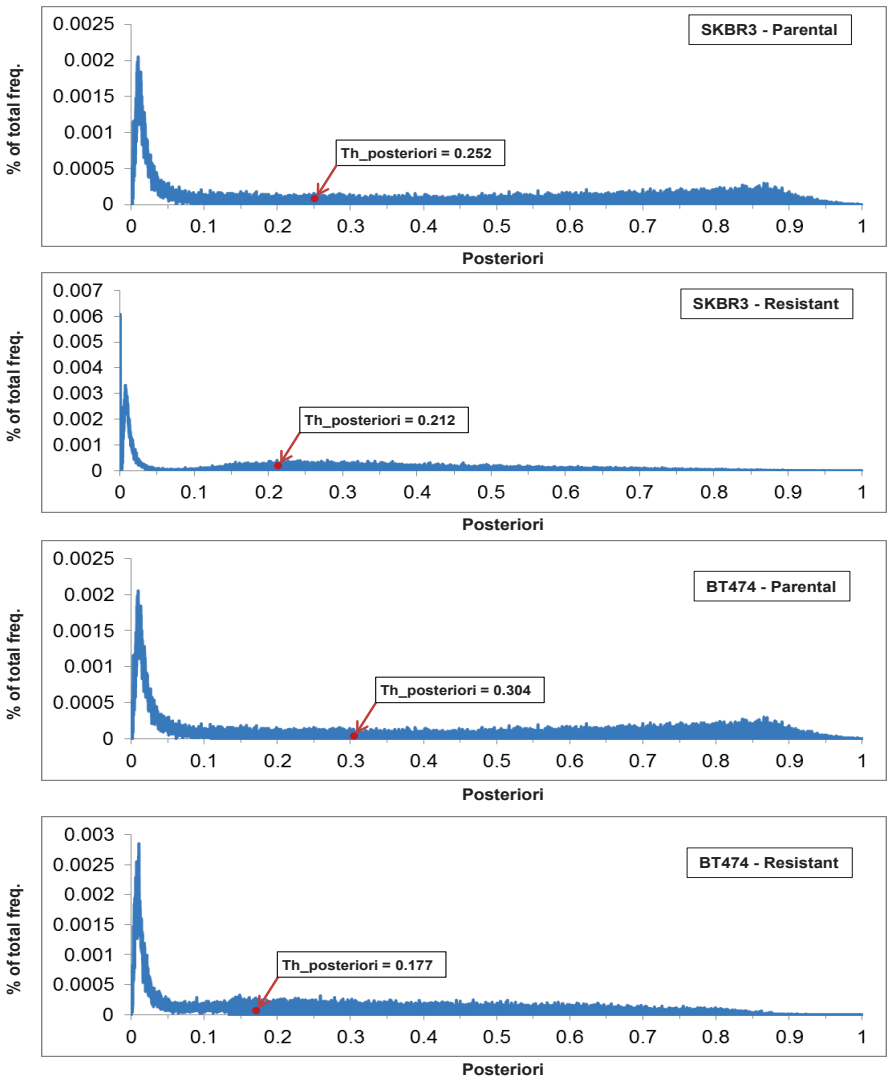


Figure 3. Frequency distributions of posteriori probabilities of gene-pairs that were selected after applying '*th_odds*' thresholds in all four cases: SKBR3-Parental, SKBR3-Resistant, BT474-Parental, and BT474-Resistant after Bayesian analysis. In all four cases, selected thresholds of posterior probabilities, '*th_posteriori*' are shown with arrows which demarcated top 50% of pairs in their respective distributions.

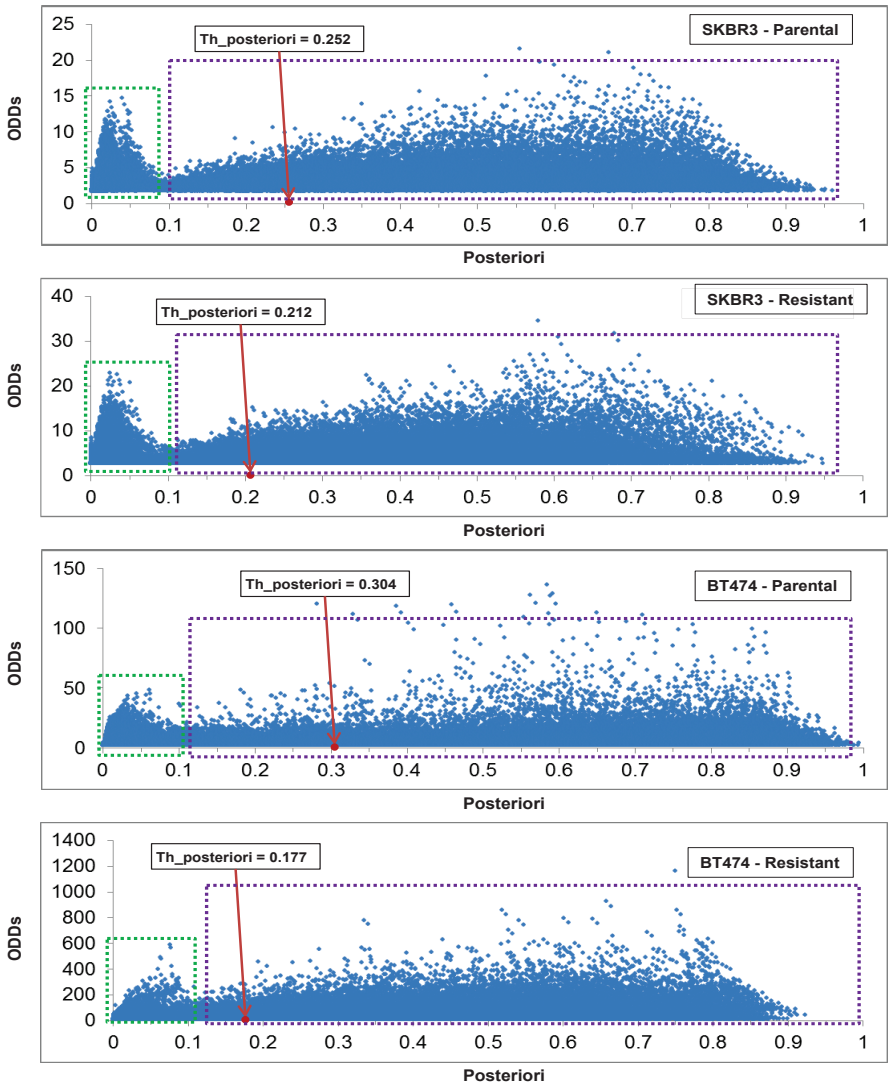


Figure 4. Scatter plots of *ODDs* VS *posterior probabilities* of gene-pairs that were selected after applying '*th_odds*' thresholds in all four cases: SKBR3-Parental, SKBR3-Resistant, BT474-Parental, and BT474-Resistant after Bayesian analysis. In all four cases: 1) two distinct distributions were shown within two boxes (in green and violet color), 2) selected thresholds of posterior probabilities, '*th_posteriori*' are shown with arrows which demarcated top 50% of pairs in their respective frequency distributions [See Supplementary Figure 3]

Table S1: List of identified putative aberrant gene-pairs (for both SKBR3 and BT474) cell-lines in acquired resistance.

Table S2: Full results of pathway enrichment tests of identified aberrant gene-pairs in acquired resistance from KEGG, Reactome, and WikiPathway databases for both SKBR3 and BT474 cell-lines.

Table S3: Comparing our current model with the previous model by observing the percentages of non-direct (indirect and PPI pair) enriched links (aberrant pairs as known signaling links) in the aberrant signaling pathways from KEGG, Reactome, and WikiPathway databases *that were detected by our current but not the previous model*, for both SKBR3 and BT474 cell-lines.

Table S4: Comparing our current model with the previous model by observing the percentages of non-direct (indirect and PPI pair) enriched links (aberrant pairs as known signaling links) in the aberrant signaling pathways from KEGG, Reactome, and WikiPathway databases *that were detected by both of our current and previous models, and were ranked (based on enrichment q-value) high in the current model but low in the previous model*, for both SKBR3 and BT474 cell-lines.

Table S5: Summary of Type-I, Type-II and Type-III enrichment of V-structures in KEGG, Reactome, and WikiPathway databases in *SKBR3* cell-line.

Table S6: Summary of Type-I, Type-II and Type-III enrichment of V-structures in KEGG, Reactome, and WikiPathway databases in *BT474* cell-line.

Table S7: CGC genes in all the Type-I, Type-II and Type-III V-structures in *SKBR3* and *BT474* cell-lines.

Appendix D

Appendix to Chapter 6

Uniform sampling of directed and un-directed graphs conditional to vertex connectivity

Salem A. Alyami, A K Azad and Jonathan M. Keith
Monash University (AUS)
Clayton Campus, Wellington Road, Clayton, Victoria 3800, Australia.
mailto: salem.alyami@monash.edu.

***Abstract*—Many applications in graph analysis require samplers to uniformly generate graphs at random over their spaces. One example task is how to efficiently create a small *sample* graphs from a particular huge target space. In this paper, we propose a new approach to sufficiently sample random graphs equally likely from their graph space. The new approach uses a Markov chain Monte Carlo (MCMC) method, called the Neighborhood Sampler (NS). We validate the new sampling technique by simulating from feasible spaces of directed and un-directed graphs. The simulation results show explicit and fast uniform recovery over the graph spaces we target.**

keyword: Graph space, Markov chain Monte Carlo approach, Neighborhood sampling.

I. Background

Random graphs (RGs) is a broad area of research that has been studied since the early works of [5] and [4]. RGs intend to combine two different theories, graph and probability. One area of interest within this broad area is the generation of random graphs according to some criteria. Two commonly used approaches to generate random graphs are either by using a probability distribution, or by using a random process [2]. One example model is the ErdösRńyi model [4]. It generates random graphs by referring to a uniform probability distribution, where the idea is to assign equal probability to all graphs given a particular number of edges. The latter model was extended to stochastically start with a particular number of nodes with no edges and then iteratively add one new edge sampled uniformly over the set of missing edges [1]. One aim of this paper is to find an efficient sampler that is capable to sample graphs that are uniformly distributed. In particular, we consider directed and un-directed graphs that their vertices are connected, in order to meet the wide-spread of their applications.

A graph $\mathcal{G}$ is expressed as a pair (N, E), where N is a set of nodes represents variables and E is a set of edges describes interactions or resoning between variables $X_1, X_2, \ldots, X_N$. Figure 1 illustrates some types of graph structures: Connected directed acyclic graph (CDAG), connected un-directed graph (CUDG) and connected directed cyclic graph (CDCG).

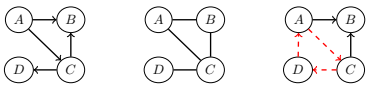


Fig. 1: From left to right: CDAG, CUDG and CDCG.

In the subsequent sections, we discuss two types of graph structures: CDAG in I-A and CUDG in I-B. In section II, we present the mathematical process of carrying out the general Neighborhood Sampler, followed by describing the technique of assigning local Neighborhoods for both CUDG structures in III-A and CDAG structures in III-B. In section IV, we introduce two graph algorithms used to detect connectivity in IV-A and cyclicity in IV-B of a particular type of graph. In section V, we show how to sample different graph structures uniformly from their graph spaces using the Neighborhood Sampler. In section VI, we demonstrate the simulations results at variant settings.

A. Directed acyclic graph

A directed edge between any pair of variables $X_i \rightarrow X_j$ indicates the dependency of X_j upon X_i. Typicaly, such cause and effect relationships are acyclic which contain no loops, where the starting node on a partcular path of edges is different from the last end node. In CDAGs, as the number of nodes grows, the space size would grow exponentially. Table I states the space sizes of CDAGs at certain numbers of nodes.

Nodes N.	CDAGs
3	18
4	446
5	26431

TABLE I: Space sizes at a certain number of nodes of CDAGs.

The space of a CDAG with only 3 nodes is small enough that the entire probability distribution and the sampler can be evaluated, see Figure 2.

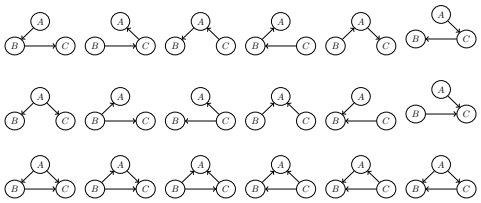


Fig. 2: Graph space of a CDAG with 3-node.

B. Undirected Graphs

CUDG is another broad class of graphs compacts representation of joint probability distributions. An un-directed edge between any pair of variables X_i and X_j within graph G indicates that they are interacted. The spaces of CUDGs are also dramatically increase as the number of nodes increase. Table II shows all possible CUDGs at some node numbers.

Nodes N.	CUDGs
4	38
5	728
6	26,704

TABLE II: Space sizes at a certain number of nodes of CUDGs.

From Table II, there are 38 possible CUDGs represent the entire space of a CUDG with 4 nodes which is tractable enough to be drawn as shown in Figure 3.

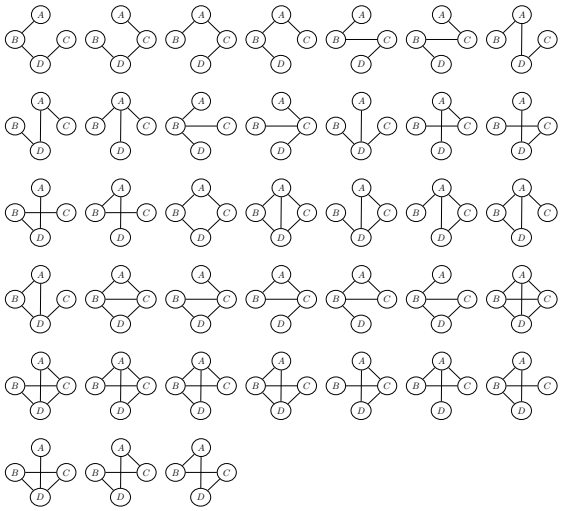


Fig. 3: The entire graph space of a CUDG with 4-node.

II. Neighborhood Sampler approach

Neighborhood Sampler (NS) is a recently introduced Markov chain Monte Carlo (MCMC) method by [6]. It cares not only about sampling over the local neighbourhoods $\mathcal{N}_x$ of a particular element x, but also reduces sampling from a complicated distribution to sample uniformly over these neighborhoods. It takes advantage from running the Markov chain process to ensure the convergence to a particular target distribution. For the purpose of this paper, we assume a Uniform target distribution function f is defined over a target space $\mathcal{X}$, with a counting measure μ on $\mathcal{X}$. The NS algorithmically can mitigate the effect of local modes. For every iterative simulation, the sampler traverses into space by transiting two elements: from element X to element Y which is sampled uniformly from $\mathcal{N}_X$ and then to element Z which is sampled uniformly from $\mathcal{N}_Y$. Then, a rejection step is used to reduce $\mathcal{N}_Y$ until we accept a particular Z. This means that every rejected element Z is excluded from being sampled again from the set $\mathcal{N}_Y$. This would speedily give a slight chance of moving onto another neighbourhood. To construct a Neighborhood Sampler, we must assign a unique neighborhood $\mathcal{N}_x$ to each element $x \in \mathcal{X}$, provides that $x \in \mathcal{N}_x$ for all $x \in \mathcal{X}$. Algorithm 1 describes sampling from an arbitrary distribution f with respect to μ using the NS [6].

Algorithm 1

Given the current state $\mathbf{X}_t = X$*:*

1) *Generate* $\mathbf{Y} \sim \mathsf{U}(\mathcal{N}_X)$ *where* $\mathsf{U}(\mathcal{N}_X)$ *is the uniform distribution (with respect to* μ*) on* $\mathcal{N}_X$*. Set* $H = \mathcal{N}_{\mathbf{Y}}$*.*
2) *Generate* $\mathbf{U} \sim \mathsf{U}(0, f(x)/\mu[\mathcal{N}_X])$*.*
3) *Generate* $\mathbf{Z}_1 \sim \mathsf{U}(H)$*.*
4) *Set* $k = 1$ *and iterate the following steps until* $f(\mathbf{Z}_k)/\mu[\mathcal{N}(\mathbf{Z}_k)] \geqslant \mathbf{U}$*:*
   a) *Reduce* H *by excluding* $\mathbf{Z}_k$ *while still contain* X*.*
   b) *Generate* $\mathbf{Z}_{k+1} \sim \mathsf{U}(H)$ *and set* $k := k + 1$*.*
5) *Set* $\mathbf{X}_{t+1} = \mathbf{Z}_k$*.*

III. Assigning local neighborhood graphs

A. With connected un-directed graphs

To construct the $\mathcal{N}_X$ for a particular CUDG X, the sampler considers all possible edges that can be added or deleted while preserving the condition of connectivity. All these valid graphs plus the original graph itself are defined as neighbourhood $\mathcal{N}_X$ of graph X, i.e. a one neighbourhood is identified by either adding an edge, deleting an edge, or do nothing, provides the graph remains connected. Figure 4 shows how all possible addable and deletable edges are identified to calculate the corresponding neighbourhoods after initialising a particular CUDG. The total number of the local neighbourhoods exist in $\mathcal{N}_X$ are referred by μ_X which is an integer number between 0 and ∞. From Figure 4, $\mu = 6$ because we have 2 addable edges, 3 deletable edges and the graph itself.

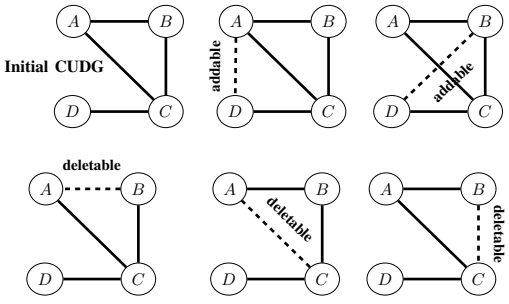


Fig. 4: From top to bottom: Initial CUDG, all possible addable edges, and all possible deletable edges.

B. With connected directed acyclic graphs

Given a CDAG, all possible edges that can be added or deleted provides that the CDAG remains connected and acyclic plus the given CDAG are defined as its neighbourhood $\mathcal{N}_X$. From Figure 5, $\mu = 7$ because we have 3 addable edges, 3 deletable edges and the graph itself.

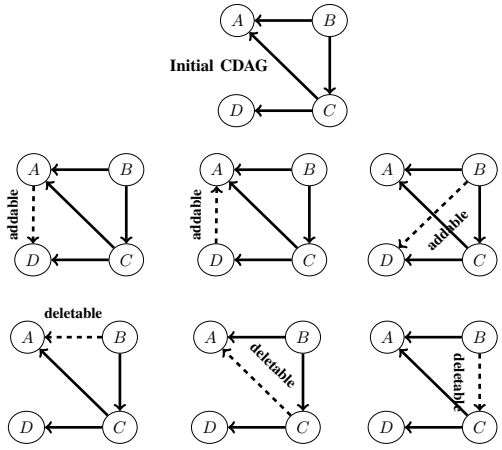


Fig. 5: From top to bottom: Initial CDAG, all possible addable edges, and all possible deletable edges.

IV. Graph algorithms

A. Detecting acyclicity and paths

Acyclicity is the required restriction to satisfy the criteria of directed acyclic graph. One possible technique to detect cycles is depth-first search (DFS) [3]. DFS is an algorithm for traversing tree or graph data structure using a process called in-traversal which allows visiting vertices of the graph. The in-traversal technique aims to navigate the graph to seek how its vertices are connected to each other and which vertices can reach from other vertices and so on. In the process of in-traversal every vertex and every edge is examined. One starts at arbitrary root and explores as far as possible along each branch before backtracking i.e. if there is no way to continue in deep the tree from a particular vertex, then mark this vertex as visited and backtrack to its parent. It typically takes time $O(|E|)$ linear in the size of the graph.

B. Detecting connectivity

The connectivity restriction requires all nodes to be connected to at least one other node in the network. This restriction ensures that every node would have at least one incident edge, so the resulting random network with n nodes must have at least $n-1$ edges. Removing an edge that disconnects the graph into two sub-graphs is called a bridge. One possible algorithm to find bridges is the Schmidt's Bridge Finding (SBF) algorithm [7]. SBF first converts a given graph into a spanning tree. A spanning tree is defined as a minimal set of connected un-directed edges that connect all nodes, in which adding one more edge to the spanning tree would create a cycle. Next, the algorithm will start to store the back-edges and then travels through the back-edges visited during the algorithm. The algorithm then would make a chain decomposition of the graph, where a chain is either a cycle or a path. All edges that are not in a chain are assigned as bridges.

V. Sampling graph space uniformly

Since we aim to sample RGs uniformly from their graph space, so we set the target function $f(x)$ in Algorithm 1 to a Uniform(a,b) distribution. Thus, we should expect from the sampler to return unbiased frequencies that are uniformly distributed. Optionally, we can set $f(x)=1$. The process of sampling CUDGs and CDAGs using the Neighbourhood sampler starts from initialising a candidate graph $X^{(0)}$ at random. The sampler must first assign local neighbourhoods $\mathcal{N}_X$ for the candidate graph $X^{(0)}=X$. Calculate μ_X and then generate a uniform value U from the interval $U(0,\frac{1}{\mu_X})$. Then, we sample another graph Y uniformly from $\mathcal{N}_X$, that is, $Y\in \mathsf{U}(\mathcal{N}_X)$. To sample a neighborhood graph uniformly, we assign $1/\mu$ probability to each neighborhood in $\mathcal{N}_Y$. Now, given $\mathcal{N}_Y$, we sample another graph Z uniformly, that is, $Z\in \mathsf{U}(\mathcal{N}_Y)$. Having calculated μ_Z, if $\frac{1}{\mu_Z}\geqslant U$, we accept the graph Z and set $X^{(t+1)}=Z$. Otherwise, we exclude the graph Z from $\mathcal{N}_Y$ and generate another Z until the acceptance ratio is satisfied. Below is pseudocode summarises sampling CDAGs and CUDGs spaces uniformly with the NS:

Algorithm 2 Sampling CDAGs and CUDGs uniformly

1: Initialise either a CUDG or CDAGs $X^{(0)}$ with $t=0$
2: **for all** $t=0,1,\ldots,n$ **do**
3: Given the current graph $\mathbf{X}_t=X$, find $\mathcal{N}_X$.
4: Calculate μ_X.
5: Generate $\mathbf{U}\in \mathsf{U}(0,\frac{1}{\mu_X})$.
6: Sample graph $\mathbf{Y}\in \mathsf{U}(\mathcal{N}_X)$ and find its $\mathcal{N}_Y$.
7: **for all** $k=1,2,\ldots,m$ **do**
8: Sample graph $\mathbf{Z}_k=Z\in \mathsf{U}(\mathcal{N}_Y)$, and find its $\mathcal{N}_Z$.
9: Calculate μ_Z
10: **if** $\frac{1}{\mu_Z}\geqslant \mathbf{U}$ **then**
11: set $\mathbf{X}^{t+1}=Z_k$
12: goto 2
13: **else**
14: Exclude Z_k form $\mathcal{N}_Y$
15: goto 7
16: **end if**
17: **end for**
18: **end for**

VI. Simulations

Simulations in this paper were written in C#.net on Acer (Aspire E1-570) computer with [3.40 GHz Intel i5-3337U]. We set our sampling to some feasible graph spaces of 4, 5 and 6 nodes of both CUDGs and CDACs. To evaluate convergence rates, we run different numbers of iterations from 100 up to 50 000 000. We test whether the sampler is able to explore the entire spaces of particular graph sizes or not. Also, whether the sampled graphs match the target uniform distribution we sample from or not.

With regard to the CDAGs, we start the simulation with 4-node graph. There are 446 CDAGs which takes about

2500 iterations from the sampler to go through. Two more simulations are run each time to compare how many graphs are visited by both simulations. Each simulation is run 50 times with 4 nodes and 100 iterations each. The maximum number of graphs visited in both simulations is 27, average is 14.30. This shows the ability of the sampler to traverse into the graph space and return a good enough number of graphs. Figure 6 states the frequencies of graphs sampled by the NS from a CDAG with 4 nodes at different number of iterations 5000, 50 000, and 500 000. The sampler with a short chain of 1000 iteration could explore 363 graphs out of 446 CDAGs which represent 81% of their graph space. The entire graph space of 4 nodes has being recovered uniformly with 5000 iteration as illustrated in Figure 6. Less than 5000 iteration i.e. 2500 iteration can also explore the whole graph space but not always. The uniform distribution we sample from becomes visible at 50 000 iteration and clear enough with 100 000 iterations.

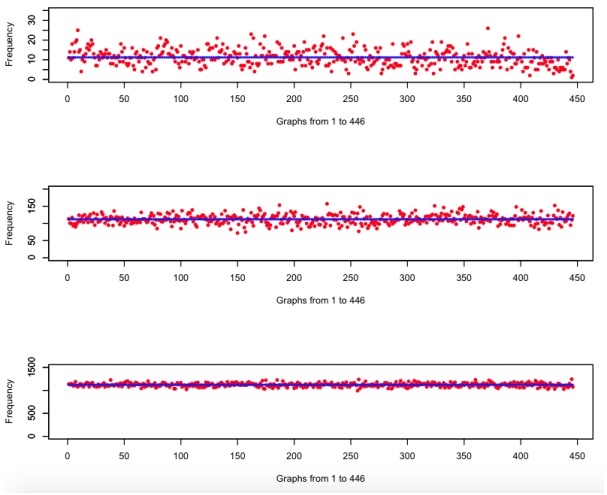


Fig. 6: From top to bottom: Neighbourhood Sampler (Red) vs True Distribution (Blue) with 5000, 50 000, and 500 000 iterations, respectively.

Considering a larger but feasible graph space with 5 nodes, there are 26790 connected DAGs. Running the simulation with 50 000 iterations, it gets through 25723 connected DAGs. Figure 7 shows that the sampler at iterations of 500 000, 5000 000, and 50 000 000 could successfully cover all the 26790 connected DAGs indexed on the horizontal axis. The graphs also have been shown to be uniformly distributed along as the number of iterations increases. The plots in Figure 7 compare the true frequencies produced by the uniform distribution and the sampled frequencies produced by the NS.

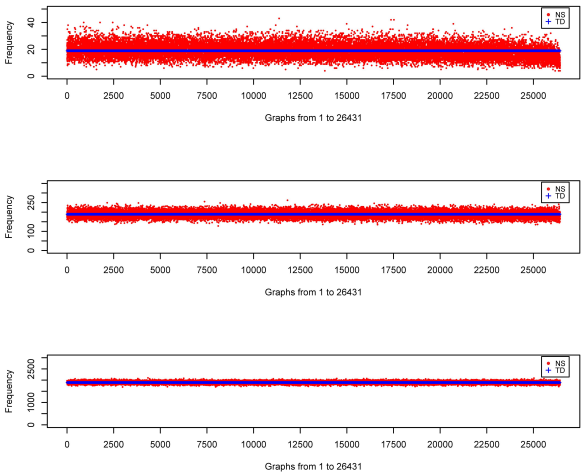


Fig. 7: From top to bottom: Neighbourhood Sampler (Red) vs True Distribution (Blue) with 500 000, 5000 000, and 50 000 000 iterations, respectively.

With CUDGs, we start also with a graph of 4 nodes where its graph space consists of 38 CUDGs. Consistently, the entire graph space of 4 nodes has being recovered uniformly at 250 iteration. Considering a larger but feasible graph space with 5 nodes, there are 728 CUDGs. Running the simulation with 1000 iterations, it gets through 489 CUDGs. The results also show that the entire graph space of 5-node is being always explored with only 5000 iterations at the execution time less than a second. The plots in Figure 8 compare true and sampled frequencies in which their sum of squared differences explicitly diminishes as the number of iteration grows.

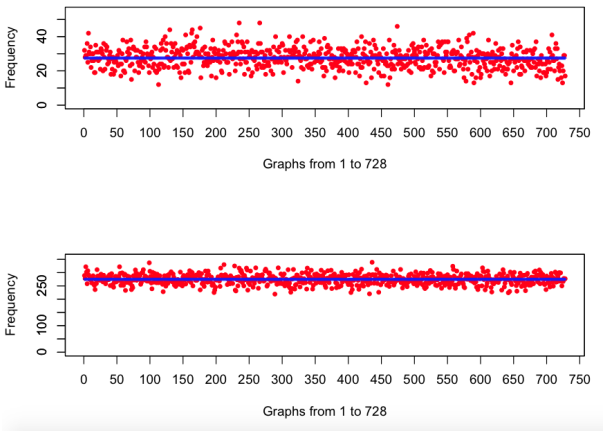


Fig. 8: From top to bottom: Neighbourhood Sampler (Red) vs True Distribution (Blue) with 20 000 iterations and 200 000 iterations, respectively.

For more assessing, we test the uniformity of sampling by exploring a CUDG with 6 nodes, where the size of its graph space is 26704 as shown in Table II. Figure 9 shows the exploration of the entire graph space. All sampled graphs are

indexed on axis X with their frequencies on axis Y, at 500 000 and 5000 000 iteration which demonstrates a faster recovery of uniform distribution.

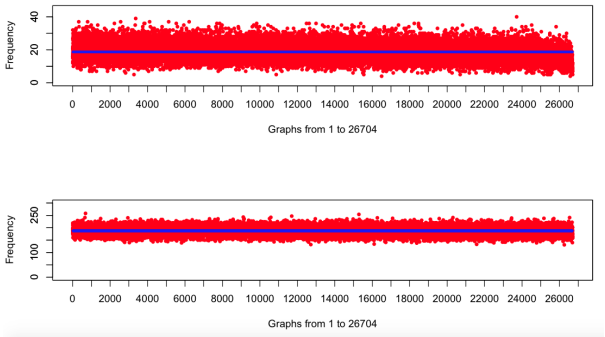


Fig. 9: From top to bottom: Neighbourhood Sampler (Red) vs True Distribution (Blue) with 500 000 iterations and 5 000 000 iterations, respectively.

VII. Conclusion

This paper presents a new approach to sampling uniformly over graph spaces. It covers simulating from different graph structures include un-directed graph and directed acyclic graphs. All these types of graph structures are conditional to vertex connectivity. We use a new Markov chain Monte Carlo method for carrying out the simulation. The new method is a new implementation of the Neighborhood sampler into the context of sampling graph space. One goal is to generate graphs that are likely equally over their graph spaces. The outputs demonstrate rapid exploring of the entire graph spaces for some feasible graph sizes as well as returning graphs' frequencies that are uniformly distributed.

References

[1] B. Bollobas. *Random Graphs*. Academic Press Inc., London Ltd., 1985. 1

[2] B. Bollobas. *Random Graphs*. Cambridge University Press, 2nd edition, 2001. 1

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press and McGraw-Hill, 2nd edition, 2001. 3

[4] P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959. 1

[5] E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959. 1

[6] J. M. Keith, G. Y. Sofronov, and D. P. Kroese. The generalised gibbs sampler and the neighborhood sampler. in *Monte Carlo and Quasi-Monte Carlo Methods 2006. Springer Berlin Heidelberg*, 31:537–547, 2008. 2

[7] J. M. Schmidt. A simple test on 2-vertex and 2-edge-connectivity. *Information Processing Letters*, 113(7):241–244, 2013. 3

The Neighborhood MCMC sampler for learning Bayesian networks

Salem A. Alyami[a], A. K. M. Azad[a], and Jonathan M. Keith[a]

[a]School of Mathematical Sciences, Monash University, Australia

ABSTRACT

Getting stuck in local maxima is a problem that arises while learning Bayesian network (BN) structures. In this paper, we studied a recently proposed Markov chain Monte Carlo (MCMC) sampler, called the Neighbourhood sampler (NS), and examined how efficiently it can sample BNs when local maxima are present. We assume that a posterior distribution $f(N, E|D)$ has been defined, where D represents data relevant to the inference, N and E are the set of nodes and directed edges, respectively. We illustrate the new approach by sampling from such a distribution, and inferring some BNs. The simulations conducted in this paper show that the new learning approach substantially avoids getting stuck in local modes of the distribution, and achieves a more rapid rate of convergence, compared to the MCMC Metropolis-Hastings sampler and other heuristic algorithms.

Keywords: Directed acyclic graph, structure inference, local maxima, graph space

1. INTRODUCTION

Bayesian Networks (BNs) are directed acyclic graphs (DAGs) that are used as a probabilistic method to visually represent directed causal relationships between variables, learned from a dataset. Nodes of the graph represent random variables, and directed edges represent causal relationships. Sampling algorithms in spaces of BNs are computationally intensive because the number of DAGs dramatically increases with the number of nodes. For example, there are 543 and 3 781 503 possible Bayesian networks in a graph space with 4 and 6 nodes, respectively. Learning a BN typically involves two conceptually different elements: structure learning and parameter learning. Structure learning involves inferring the variables that interact and the causal directions of those interactions; in other words it is inferring the set of edges connecting a set of candidate nodes. For a fixed structure, parameter learning involves quantitatively estimating probabilistic dependencies between variables. In practice, structure and parameter learning may be performed simultaneously. In this paper, both types of learning are explicitly considered while sampling BNs.

Bayesian network structures have been widely learned by score-based algorithms. This category of algorithms aim to maximise the pre-assigned score of each Bayesian network using a heuristic search algorithm. One of the most widely studied heuristic search methods is Greedy Algorithms (GAs).[1] GAs typically update a given Bayesian network by either adding, deleting or reversing a particular directed edge at each step. Among the most widely used special GAs are Hill-Climbing (HC) algorithm and Tabu Search (TS) algorithm.[2,3] The HC algorithm iteratively starts with an arbitrary Bayesian network, and then applies a local search to its neighbors in the hope to find a neighboring network with a better score. It repeats this process until no further improvements can be obtained. The TS algorithm also runs a local search similar to the HC, however, it intentionally enhances the performance of local search by relaxing its acceptance function i.e. when the search gets stuck at a local mimimum and no improving move is available, worsening moves can then be accepted. The TS algorithm also uses a memory structure that describes all visited solutions. If a particular Bayesian network has been previously visited but not improved the score, it is then marked as "tabu" and not considered again. Heuristic search is a problem when the immediate neighbours of a network do not provide any better solution. The category of constraint-based algorithms is another main class has been used to learn Bayesian networks. It aims to analyse the probabilistic relations entailed by the Markov property of Bayesian networks with conditional independence tests and then construct a Bayesian network that satisfies the corresponding d-separation statements. One

Corresponding author to Salem A. Alyami.
Salem A. Alyami.: E-mail: salem.alyami@monash.edu, Telephone: 1 505 123 1234

common algorithm attributed to this category is the Grow-Shrink (GS).[4] The GS approach constructs Bayesian networks by identifying the Markov blanket for each node, and then connect nodes. This is in order to avoid producing dense nets or incorrect causal relationships.

Markov chain Monte Carlo (MCMC)[5–7] is a subclass of stochastic sampling. It involves simulating a *Markov chain* process, which is constructed specifically to converge to the target distribution. Practically, MCMC is a highly flexible methodology for sampling from complicated target distributions. However, the efficiency of convergence of an MCMC sampler is often an issue, particularly for very high dimensional distributions. One technique used in structure learning of BNs is to sample from a posterior distribution over a space of BNs by using Markov chain Monte Carlo (MCMC) method. This presupposes that a prior distribution and likelihood model have been defined over graph space, and that Bayes' rule has been applied to obtain a posterior distribution. However, MCMC methods e.g. Metropolis-Hastings sampler can be slow to converge when the target distribution has local maxima.

The main goal of this paper is to examine the potential of a recently proposed MCMC method called the Neighbourhood MCMC sampler[8] to traverse the search space with reduced frequency of getting trapped in local modes. The general NS is outlined in Appendix A. The sampler possesses a number of unique features. The sampler has a reduction step in which rejected elements are excluded from being chosen a second time. Another feature is that each new element is chosen in two steps: starting from an initial element X, a neighbor Y is first selected, then a neighbor Z of Y is proposed. Another goal is to present a comparison study between the new approach and the MCMC Metropolis-Hastings (MH) sampler. Another difference is that our algorithm does not consider graphs obtained by reversing edges, unlike many heuristic algorithms. This is intentionally avoided in this study to reduce the number of neighbors of a particular graph.

The organisation of this paper is as follows. Section 2 describes the methods used to learn Bayesian Networks. This includes parameter learning using the Bayesian approach in 2.1, and structure learning using the new MCMC Neighborhood Sampler in 2.2. Using the MCMC MH is also outlined in 2.3. Section 2.4 discusses a range of constraints that may be imposed to reduce the size of a graph space. Section 3 reports all simulation results as follows: A comparison study between the new MCMC NS and MCMC MH is addressed in 3.1. Section 3.2 applies the Neighborhood Sampler to a popular application and compares the outputs with other common samplers.

2. METHODS

2.1 Parameter learning with Dirichlet-Multinomiall distribution

Learning the parameters of a BN corresponds to learning the local conditional probabilities among the variables encoded. Given data D and fixed graph G, we first define the probability distribution $P(X_i|Pa(X_i))$ for each variable X_i given its parents $Pa(X_i)$. We write $P(X_i = k|Pa(X_i) = j) = \theta_{ijk}$, where θ_{ijk} is the probability of each bin k within each variable X_i, given that its parents are in configuration j. The multinomial distribution is the likelihood function for relating these parameters to observational data.[9] To complete the Bayesian model, we also need to assign prior probabilities to parameter values. Here, we use the conjugate prior of the multinomial model, which is the Dirichlet distribution. The ultimate formula of the Dirichlet-Multinomial posterior model is expressed in Equation 1:[9]

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{1}$$

where N_{ijk} is the number of observations in bin k of node i corresponding to a parent configuration j, r_i is the number of possible state values (bins) for a particular variable X_i, q_i is the total number of configurations of parent state values of X_i, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, and α_{ijk} are the hyper-parameters (hyper-conditional probabilities), $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. Expression 1 is useful in that it gives the likelihood of a set of data in terms of the structure only, without reference to the parameters associated with each node. In this paper, the Dirichlet priors have been assigned as $\alpha_{ijk} = \alpha/q_i.r_i$, where α is the total imaginary counts for the Dirichlet prior. The posterior probability distribution of the graph G given data D can now be constructed as:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \tag{2}$$

To sample from 2 using MCMC, we need only consider the numerator, since the denominator does not depend on G and will cancel out. We also assume a uniform prior on the graph $P(G)$. Note that equivalent graphs have the same prior probabilities, likelihoods and posterior probabilities. Consequently, equivalence classes of graphs have prior and posterior probabilities proportional to the number of equivalent graphs in that class. This is possibly undesirable, as there is no obvious reason why larger equivalence classes should be preferred *a priori*. In principle, this effect could be counteracted by assigning a prior probability to each graph G inversely proportional to the size of its equivalence class. However, for simplicity we have retained the uniform prior in what follows. A practical issue that arises when working with 1 is the very high values that result from multiplying several gamma functions together. The solution is to work with the *log* of these values wherever possible.

2.2 Structure learning with the new MCMC Neighborhood Sampler

In this secion, we particularise the NS in Algorithm 2 in Appendix A to learn BN structures. In what follows, $G^{(t)}$ denotes a graph sampled at iteration t of the Neighborhood Sampler (NS). The process of learning BN structures using the NS begins by selecting an arbitrary graph $G^{(0)} = G$. Since we are interested in sampling Bayesian networks, every generated graph must be directed and acyclic. We also require that all sampled graphs must be connected. This is an appropriate requirement to reduce the size of graph space. Next, the sampler requires assigning a set of local neighborhoods denoted by $\mathcal{N}_G$ for each candidate graph G in the search space. The set $\mathcal{N}_G$ consists of the graph G itself, all graphs that can be obtained by adding a directed edge, provided the graph remains acyclic, and all graphs that can be obtained by deleting an edge, provided the graph remains connected. Let μ_G be the total number of graphs in $\mathcal{N}_G$ i.e. all possible addable and deletable edges, which may be any strictly positive integer. For example, in Figure 1, $\mu_G = 7$ because we have 3 addable edges, 3 deletable edges and the graph itself. Given the value of μ_G and the probability $f(G)$ calculated for a particular graph $G^{(t)} = G$, we generate a uniform value U by sampling uniformly U from the interval $(0, \frac{f(G)}{\mu_G})$. Then a new graph H_1 is sampled uniformly from the set $\mathcal{N}_G$ such that $H_1 \in \mathsf{U}(\mathcal{N}_G)$. Each graph in $\mathcal{N}_G$ has a probability of $1/\mu_G$ to be sampled. Find $\mathcal{N}_{H_1}$ and again sample graph H_2 uniformly, such that $H_2 \in \mathsf{U}(\mathcal{N}_{H_1})$. Having calculated μ_{H_2} and its probability $f(H_2)$, the graph H_2 is accepted if $\frac{f(H_2)}{\mu_{H_2}}) \geqslant U \in \mathsf{U}(0, \frac{f(G)}{\mu_G})$, then we set $G^{(t+1)} = H_2$ and go to the next iteration. Otherwise, we exclude the graph H_2 from $\mathcal{N}_{H_1}$ and sample another H_2 until the acceptance ratio is satisfied. The pseudocode in Algorithm 1 summarises the entire process of learning BNs using the NS.

Algorithm 1 Learning BNs with the NS

1: Initialise graph $G^{(0)}$.
2: **for all** $t = 0, 1, \ldots, n$ **do**
3: Given the current graph $G^{(t)} = G$, find $\mathcal{N}_G$.
4: Calculate μ_G and $f(G)$.
5: Generate $u \in \mathsf{U}(0, \frac{f(G)}{\mu_G})$.
6: Sample graph $H_1 \in \mathsf{U}(\mathcal{N}_{H_1})$ and identify neighborhood $\mathcal{N}_{H_1}$.
7: **for all** $k = 1, 2, \ldots, |\mathcal{N}_{H_1}|$ **do**
8: Sample graph $H_2 \in \mathsf{U}(\mathcal{N}_{H_1})$.
9: Calculate μ_{H_2} and $f(H_2)$
10: **if** $\frac{f(H_2)}{\mu_{H_2}} \geqslant u$ **then**
11: set $G^{(t+1)} = H_2$, goto 2
12: **else**
13: Exclude H_2 from $\mathcal{N}_{H_1}$, goto 7
14: **end if**
15: **end for**
16: **end for**

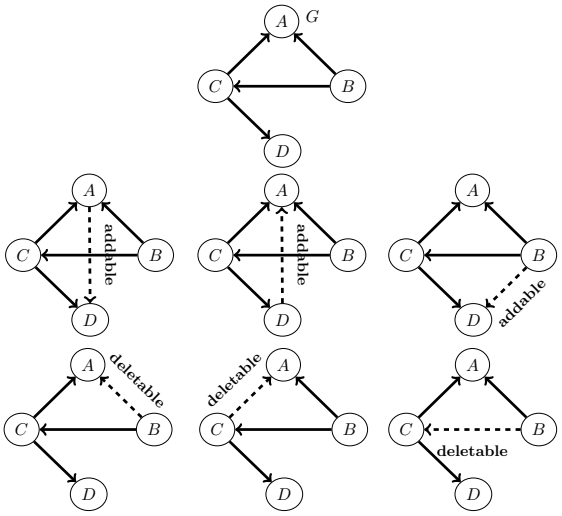


Figure 1: From top to bottom: Initial network, all possible addable edges, and all possible deletable edges.

2.3 Structure learning with the MCMC Metropolis-Hastings Sampler

The Metropolis-Hastings (MH) sampler was first presented in.[10] Like other MCMC methods, the MH algorithm generates a Markov process which *asymptotically* reaches a unique stationary distribution.[11] Several other

MCMC methods, including the Metropolis algorithm,[12] Metropolised independence sampler[10] and Gibbs sampling,[13] are special cases of MH.[14] To sample DAGs with MH, we must first define a proposal distribution. To ensure it is comparable in terms of computational effort with the NS, we set the proposal to the uniform distribution over the same neighbourhoods. Given the current graph $G^{(t)} = G$, where $f(G) > 0$, we draw a connected DAG $H \in \mathcal{N}_G$ in accordance with the uniform proposal $q(H|G) = 1/\mu_G$. Also, we find $q(G|H) = 1/\mu_H$. Then, we draw a Uniform (0,1) random value U. If $\frac{f(H)}{f(G)}.\frac{\mu_G}{\mu_H} \geqslant \mathbf{U}$, then set $G^{(t+1)} = H$, otherwise $G^{(t+1)} = G$.

2.4 Detecting structural constraints of BNs: Acyclicity, connectivity and node degree

Adding justifiable additional restrictions on a very large DAG space is a sound technique that we use here to reduce its cardinality. We use four restrictions that are appropriate in many applications: connectivity, acyclicity, and limiting in-degree and out-degree.

The connectivity restriction requires all the nodes to be connected to at least one other node in the network, so the resulting random network with n nodes must have at least $n - 1$ edges. For any edge (X_i, X_j) to be deleted in order to find a neighbour graph, we observe the connectivity of the resulting graph. In doing so, at first we delete that edge and apply the Breadth-First-Search (BFS) algorithm to detect if the graph becomes disconnected, or not. This indicates, one or more nodes loose connectivity with the remaining set of nodes in the graph when it becomes disconnected. Simply, after removing that particular edge (X_i, X_j), we checked if the exist any simple path from all other nodes in the graph to the node X_i (, or node X_j). Here, the BFS algorithm is exploited to find such simple path(s). BFS starts at any arbitrary node and explores all it's neighbourhood nodes before observing others. Thus, if there exist any such simple path(s) that ensures the resulting graph would be connected even if we remove the edge (X_i, X_j). Therefore, the edge (X_i, X_j) will be considered as a deletable edge. Once the connectivity checking is done, we restore back the edge (X_i, X_j) that we deleted previously. Note, for the of DAG connectivity, we do not consider the directions of edges as important here.

Acyclicity is a required restriction for a Bayesian network. We use the Depth-First-Search (DFS) algorithm to detect cycles for DAGs only. Before adding any edge $(X_i \to X_j)$ into a particular graph to find its neighbour, we observe if we can find any cycle in the resulting graph. That means, we first added that edge $(X_i \to X_j)$ in that graph and run DFS algorithm if any cycle evolves in the resulting graph or not. If not, we considered that edge to be an addable edge in the graph in finding it's neighbour, otherwise not. DFS is an algorithm for traversing tree or graph data structures. One starts at an arbitrary root and explores as far as possible along each branch before backtracking. Next, we remove back the edge $(X_i \to X_j)$ that we added before.

In-degree and out-degree are integer numbers that respectively represent the number of head and tail endpoints incident on a node, or equivalently the respective numbers of parents and children that a particular node possesses. Setting a maximum number of parents or children for each node can dramatically reduce the size of graph space where there is reliable prior knowledge about these parameters.

3. RESULTS

In this section, we use the Neighbourhood MCMC Sampler to infer Bayesian networks. We respectively describe the sampling processes of the NS from a feasible graph space subject to the posterior Dirichlet-Multinomial distribution. The efficiency of the sampler to learn structures and explore the entire posterior is also assessed by comparing it with other adopted samplers. All simulations in this paper were written in C#.net on an Acer (Aspire E1-570) computer with 3.40 GHz Intel i5-3337U and 8 GB RAM.

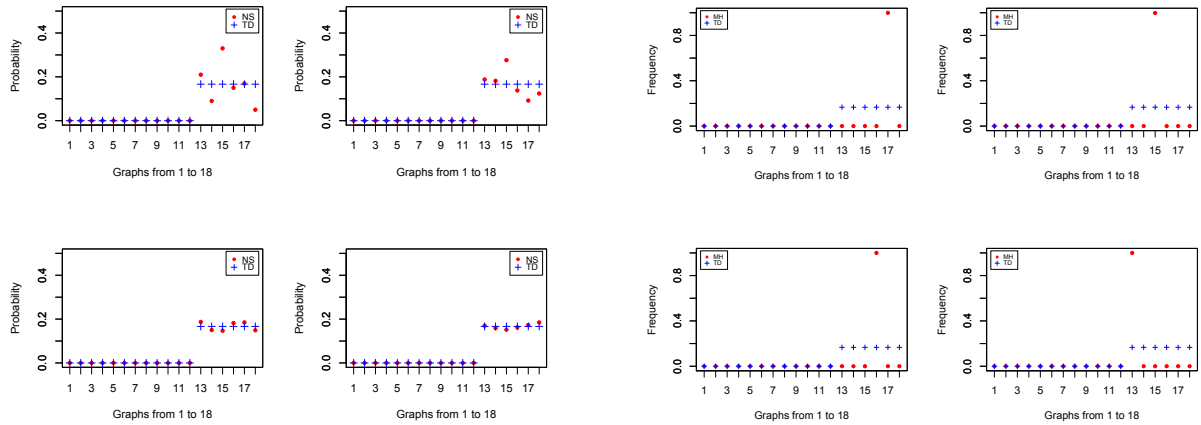
3.1 Neighborhood Sampler vs Metropolis-Hastings

It is notable that in principle a single iteration of MH requires less computation than an iteration of NS. We mention two aspects of comparison between the NS and MH. First, there are two nested for-loops involved in the NS: the first loop determines the number of iterations and the second selects a new graph $H_2 \in \mathcal{N}_{H_1}$ by rejection sampling. The inner nested loop produces one accepted graph per iteration of the outer loop, and may involve a number of rejected graphs (0 up to $(\mu_{H_2} - 1)$ rejections). The MH, on the other hand, has only one loop, which also returns one new graph per iteration but is more likely than NS to repeat the same graph. Second, at every iteration with the NS, we sample H_1 given G, then sample H_2 given H_1, whereas MH samples H given G only.

Figure 8 in Appendix B illustrates how the mediator graph H_1 in the NS enables a larger number of graphs to be reached within one iteration, providing a better chance to move to a new graph. Thus an advantageous property of the NS is that it is less likely than MH to get stuck in a local maxima for some number of iterations.

3.1.1 Application 1: Exploring posterior distribution

There are only 18 connected DAGs in the entire space of Bayesian networks with 3 nodes. This space is sufficiently small that the posterior probability can feasibly be calculated for every such graph, and it thus provides a good test environment for investigating the behaviour of MCMC methods. The series plots in Figure 2 show the proportions in which each graph has been sampled after 100, 500, 1000 and 5000 iterations of the NS and MH, compared to the actual posterior distribution. No burn-in samples have been discarded for these plots. The sampled proportions produced by the NS in Figure 2a are adequate estimates of the true posterior probabilities for most purposes after about 1000 iterations. In particular, note that the NS has sampled equivalent graphs with approximately equal probabilities, as it should. Figure 2b shows that the MH sampler does not explore the entire sample space. Instead, it repeatedly samples only one of the 6 equivalent graphs that possess the same highest probability 0.1667. It is effectively stuck at a single point in the space.

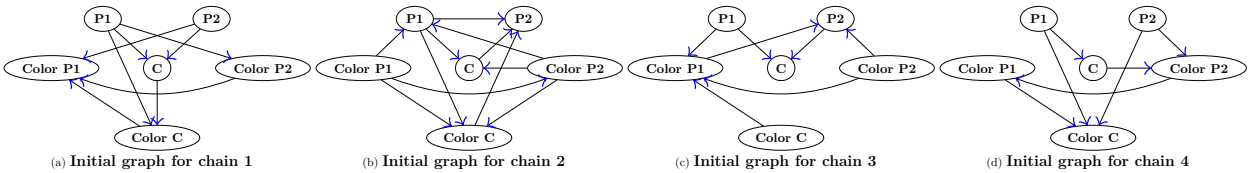


(a) NS (b) MH

Figure 2: From top to bottom, left to right: Exploring true distribution (TD) of Dirichlet-Multinomial with 3-nodes by the NS and MH at 100, 500, 1000, and 5000 iterations, respectively

3.1.2 Application 2: Mendel's Peas Network

This network was designed by Norsys Software Corp in 1998 and includes six variables. The two variables P1 and P2 are mated to produce another variable C. Each of these variables represents a plant genotype and has three possible state values RR, Rr and rr, where R is the allele for red and r is the allele for white. These three variables probabilistically determine an additional three variables: the observed colours of P1, P2 and C. Each of these colour variables has two possible state values: red and white. Using the conditional probabilities shown in Appendix C, 5000 data-points were simulated, with each data-point including values for all six variables. The graph space was reduced by imposing a maximum of three parents and three children for each node. We ran the NS and MH with four chains at the fixed random initial graphs shown in Figure 3 of 1000 iterations each.



(a) Initial graph for chain 1 (b) Initial graph for chain 2 (c) Initial graph for chain 3 (d) Initial graph for chain 4

Figure 3: Four initial graphs for running four Markov chains with lengths of 1000 iterations each using the NS and MH.

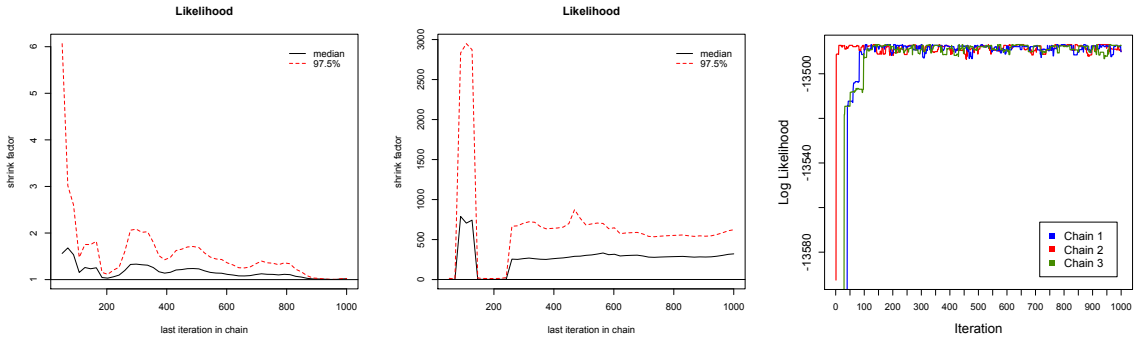
Table 1 summarises the outputs of these latter simulations produced by the NS and MH. Unlike the MH, reasonable results were obtained with the NS for 1000 iterations, with all four chains sampling the true network as shown in columns 5 and 6 of Table 1, where the ranking of graphs is according to sampling frequencies. The number of total sampled graphs in column 7 of Table 1 is more consistent and more fully explores the search space with the NS than the MH. Although, the MH sampled the true graph in the first chain, the last three chains have not even sampled the true graph. It is likely that the first chain was stuck in a local mode containing the true graph, and all other chains were stuck in other local modes.

MCMC Sampler	Chain Number	Initial Graph	Highest Frequency	True Graph Frequency	True Graph Ranking	Total Sampled Graphs
NS	1	Figure 3a	59	16	11	185
	2	Figure 3b	70	5	60	158
	3	Figure 3c	81	19	11	145
	4	Figure 3d	44	44	1	183
MH	1	Figure 3a	69	46	3	102
	2	Figure 3b	223	NA	NA	8
	3	Figure 3c	874	NA	NA	11
	4	Figure 3d	308	NA	NA	32

Table 1: Outputs of three Markov chains with 1000 iterations sampled by the NS and MH.

Gelman and Rubin diagnostic[15] is a reliable convergence test measures the difference between the within-chain variance and the between-chain variance using a value called the "scale reduction factor". It requires simulating multiple chains ($m \geqslant 2$) each of length $2t$, where t is the number of iterations, with overdisperse starting values. The first t samples in each chain are then discarded, and the within-chain and between-chain variances are evaluated. A weighted sum W of the within-chain and between-chain variances is then used to calculate the potential scale reduction factor $\widehat{R} = \frac{\widehat{Var}(x)}{W}$. The output consists of the 50% and 97.5% quantiles of the distributions of scale reduction factors. If these quantiles are both less than 1.2, the chains may be considered to be sampling from the same distribution, and then the number of iteration t needs no to be increased.

The scale reduction factors for the 50% and 97.5% quantiles produced by the four chains in Table 1 using NS are 1.02 and 1.02, respectively, and using MH are 321 and 623, respectively. This suggests that convergence has occurred after 1000 iterations with the NS as in Figure 4a where its reduction factors are less than 1.1, which was not the case with MH as shown in Figure 4b.



(a) Gelman and Rubin test with NS (b) Gelman and Rubin test with MH (c) Log Likelihoods with NS

Figure 4: MCMC diagnostic tests for chains of size 1000 iteration

One more graphical diagnostic we consider is the time-series plot of the log-likelihood at each iteration, which

can be used to judge the point at which burn-in has occurred. The log likelihood functions of three Markov chains generated by the NS are plotted with1000 iterations each as shown in Figure 4c which indicate that the sampler has a short burn-in period of approximately 100 iterations.

We generated a longer chain of 5000 iterations and discarded the first 100 iterations. Then, we report the estimated posterior probability for each individual edge by determining the proportion of sampled graphs in which that edge is present. This is a useful way to summarise an MCMC sample. The posterior probability of inclusion was then estimated for each edge after discarding the first 100 graph. For simplicity, we assigned the numbers $1, 2, 3, 4, 5$, and 6 to the variables P1, P2, Colour P1, Colour P2, C, and Colour C, respectively. We also let $p_{ij}, i, j = 1, \ldots, 6$ stand for the proportion of sampled graphs that contain a directed edge $i \to j$. The matrix of sampled proportions for each edge is illustrated in Figure 5 which also plots the edges that have posterior probabilities $> 50\%$, and they exactly correspond to the edges in the true network.

$$\begin{bmatrix} p_{11}=0 & p_{12}=0 & p_{13}=0.63 & p_{14}=0.06 & p_{15}=1 & p_{16}=0.24 \\ p_{21}=0 & p_{22}=0 & p_{23}=0.03 & p_{24}=0.76 & p_{25}=1 & p_{26}=0.18 \\ p_{31}=0.37 & p_{32}=0 & p_{33}=0 & p_{34}=0.13 & p_{35}=0.05 & p_{36}=0.31 \\ p_{41}=0 & p_{42}=0.24 & p_{43}=0.08 & p_{44}=0 & p_{45}=0.07 & p_{46}=0.23 \\ p_{51}=0 & p_{52}=0 & p_{53}=0.18 & p_{54}=0.20 & p_{55}=0 & p_{56}=0.99 \\ p_{61}=0 & p_{62}=0 & p_{63}=0.21 & p_{64}=0.26 & p_{65}=0.01 & p_{66}=0 \end{bmatrix}$$

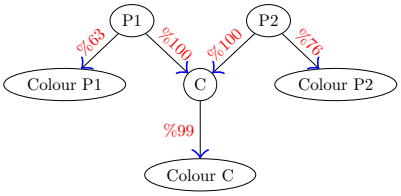


Figure 5: Mean posterior graph at a threshold $> 50\%$.

3.2 Comparing NS with non MCMC algorithms

The Diagnostic Chest Clinic Network is a popular Bayes net example in the medical domain.[16] The network aims to represent the risks of a patient having tuberculosis, lung cancer or bronchitis based on several factors, including whether or not a patient is a smoker or has traveled to Asia recently. Each variavle takes a binary values, either 0 or 1 to respectively indicate the absence or presence of a particular risk. We assume that the network is connected and that the number of parents of each node can never be greater than four. These restrictions reduce the number of graphs in the graph space and the computational time required to sample from the posterior distribution.

We simulated 10000 data points, each consisting of a value for each of the eight variables, based on the conditional probabilities presented in Appendix D. We investigated the short term behaviour of the NS for this data, by running simulations in which the number of iterations ranged from 50 to 5000, starting with random initial graphs. We found that the true network was rarely sampled in simulations shorter than 1000 iterations, which is thus a lower limit on the length of the burn-in phase. We also ran a long simulation using 30 0000 iterations, starting from a randomly selected initial graph. The matrix of the resulting posterior probability for each edge, and the highest eight proportions and their corresponding edges are illustrated and plotted in Figure 6.

$$\begin{bmatrix} 0 & 0.02 & 0.40 & 0.05 & 0.01 & 0.16 & 0.03 & 0 \\ 0.05 & 0 & 0.09 & 0.62 & 0.67 & 0.13 & 0 & 0 \\ 0.49 & 0.05 & 0 & 0.19 & 0.02 & 0.86 & 0.29 & 0 \\ 0.08 & 0.34 & 0.18 & 0 & 0 & 0.89 & 0.3 & 0 \\ 0.05 & 0.33 & 0.05 & 0 & 0 & 0.09 & 0 & 1 \\ 0.04 & 0.01 & 0.14 & 0.11 & 0 & 0 & 0.88 & 1 \\ 0.03 & 0 & 0.02 & 0.02 & 0 & 0.04 & 0 & 0 \\ 0.01 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

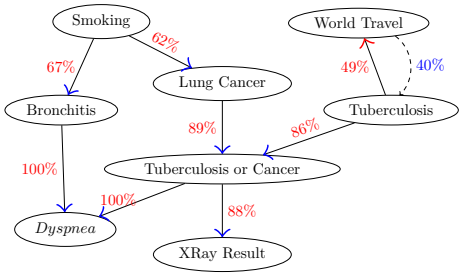
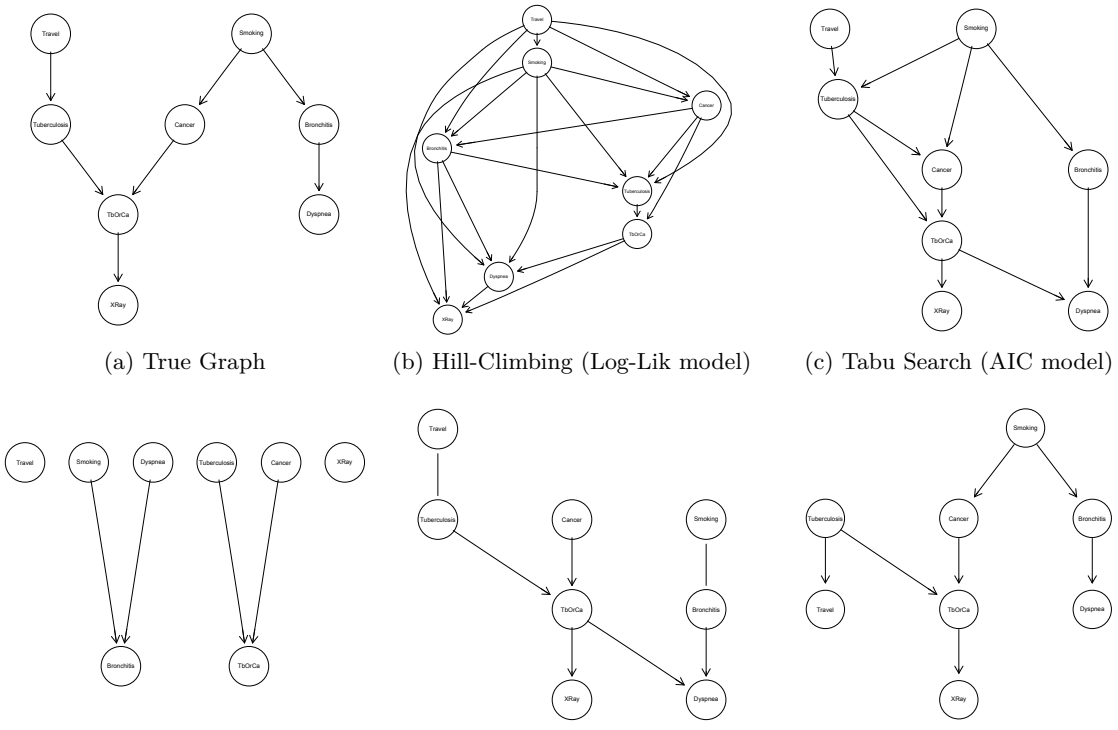


Figure 6: The highest eight proportions at a threshold > 40.

Most of the edges that belong to the true graph have higher proportions. For example, the two edges of "Bronchitis" → "Dyspnea" and "Tuberculosis or Cancer" → "Dyspnea" have been sampled with a posterior probability of 100%. It is noted that the direction between the pair of nodes "World Travel" and "Tuberculosis" is not sampled in the correct direction, and it returned a posterior probability of 49%. If we consider the highest nine proportions, we will obtain the correct direction between the "World Travel" and "Tuberculosis" with a posterior probability of 40%. There are two possible factors for this slight error that might explain this incorrect direction. First, the expected number of individuals who have both travelled and have Tuberculosis is only 5 out of 10000, and thus the actual number of such individuals simulated has high proportional variation. Second, this network is equivalent to the true network, and should in principle have the same posterior probability.

We compared the posterior structure obtained by the NS in Figure 6 (also introduced in Figure 7f) with the structures resulting from using GS, HC and TS algorithms. We used the *bnlearn* R package to apply these three algorithms to our 10 000 simulated data points generated using the conditional probabilities in Appendix D. To make them comparable with our simulation settings, initially all algorithms were run for 30 000 iterations using the multinomial log-likelihood score, with a maximum of four parents for each node. All learned structures are shown in Figure 7, and comparable with the true graph in Figure 7a. Even though we increased the number of iterations for the HC and TS algorithms up to 100 000 using different values of tabu at 10, 50 and 100, no solutions better than those shown in Figures 7b and 7c were obtained. It has been also noted that the structures learned by the HC and TS are identical whether one uses the *Akaike* (AIC) score model, or the *multinomial log-likelihood* (Log-Lik) score. For this reason it was sufficient to run the HC with the AIC score and the TS with the Log-Lik score. The GS with both the asymptotic χ^2 test and a Monte Carlo permutation test (mc-x2) has shown a better solution in Figure 7e compared to the GS nonparametric in Figure 7d. However, Figure 7e still has one missing edge and two undirected edges.



(a) True Graph (b) Hill-Climbing (Log-Lik model) (c) Tabu Search (AIC model)

(d) Grow-Shrink (nonparametric) (e) Grow-Shrink (parametric mc-x2) (f) MCMC NS

Figure 7: Diagnostic Chest Clinic Network learned by samplers using 5000 simulated data points

4. CONCLUSION

This paper presents practical instances of the Neighbourhood MCMC sampler, which is a new sophisticated Markov chain Monte Carlo algorithm. The sampler promisingly provides a new candidate MCMC approach to sampling Bayesian networks. This has been combined with using the Dirichlet-Multinomial distribution to learn the conditional probabilities among discrete variables. The primary simulations conducted in this paper have shown the effectiveness of the sampler to explore spaces of Bayesian networks and rapidly converge to the high probability density region. The correctness of our implementation of the sampler has been validated using some Bayesian networks. The computational efficiency of the sampler has been assessed by comparing it to the Metropolis-Hastings Markov chain Monte Carlo Sampler, and other widely used score-based and constraint-based algorithms. Unlike the Metropolis-Hastings, the Neighbourhood MCMC sampler substantially avoids the problem of getting stuck at a local maximum graph. In future work, we aim to apply the new approach to solve larger Bayesian network problems using a new developed adaptive technique for faster graph-neighborhood assigning to reduce the time complexity required by some graph algorithms to detect cycles and connectivity.

APPENDIX A. THE GENERAL MCMC NEIGHBORHOOD SAMPLER

NS assumes a density f has been defined on some measure space $(\mathcal{X}, \Sigma, \mu)$, where $\mathcal{X}$ is a target space, Σ is a σ-algebra and μ is a reference measure. NS is constructed by assigning a unique neighborhood $\mathcal{N}_x$ to each element $x \in \mathcal{X}$ that conditionally involves $x \in \mathcal{N}_x$ for all $x \in \mathcal{X}$. In relation to the counting measure μ over $\mathcal{N}_x$, it must take a positive real number, that is, $0 < \mu(\mathcal{N}_x) < \infty$ for all $x \in \mathcal{X}$. Algorithm 2 in Appendix A performs iteratively the general NS to sample from an arbitrary distribution having density f with respect to μ.[8] Step 4(a) in Algorithm 2 suggests reducing $\mathcal{N}_y$ by excluding elements from $H(x, y, z_1, z_2, \ldots, z_k)$ until satisfying the acceptance ratio $f(\mathbf{Z}_k)/\mu[\mathcal{N}(\mathbf{Z}_k)] \geqslant \mathbf{U}$. Definitely the acceptance ratio will be satisfied at least by its equality with the element x that still belongs to $\mathcal{N}_y$.

Algorithm 2 The Neighbourhood Sampler: *Given the current state* $\mathbf{X}_t = x$*:*

1. *Generate* $\mathbf{Y} \sim \mathsf{U}(\mathcal{N}_x)$ *where* $\mathsf{U}(\mathcal{N}_x)$ *is the uniform distribution (with respect to* μ*) on* $\mathcal{N}_x$*. Set* $H = \mathcal{N}_{\mathbf{Y}}$*.*
2. *Generate* $\mathbf{U} \sim \mathsf{U}(0, f(x)/\mu[\mathcal{N}_x])$*.*
3. *Generate* $\mathbf{Z}_1 \sim \mathsf{U}(H)$*.*
4. *Set* $k = 1$ *and iterate the following steps until* $f(\mathbf{Z}_k)/\mu[\mathcal{N}(\mathbf{Z}_k)] \geqslant \mathbf{U}$*:*
   (a) *Reduce* H *by excluding* $\mathbf{Z}_k$ *while still containing* x*.*
   (b) *Generate* $\mathbf{Z}_{k+1} \sim \mathbf{U}(H)$ *and set* $k := k + 1$*.*
5. *Set* $\mathbf{X}_{t+1} = \mathbf{Z}_k$*.*

APPENDIX B. ASSIGNING NEIGHBORING GRAPHS IN ONE ITERATION

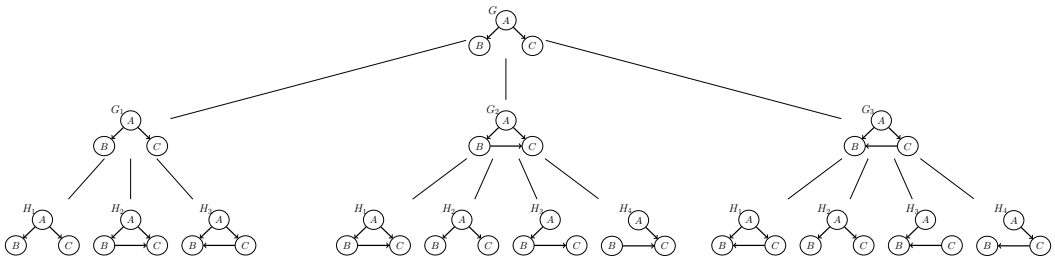


Figure 8: 3 possible graphs if we apply the MH and 7 possible graphs (after excluding similar graphs) if we apply the NS.

APPENDIX C. CONDITIONAL PROBABILITIES TO SIMULATE DATA FOR MENDEL'S PEAS NETWORK

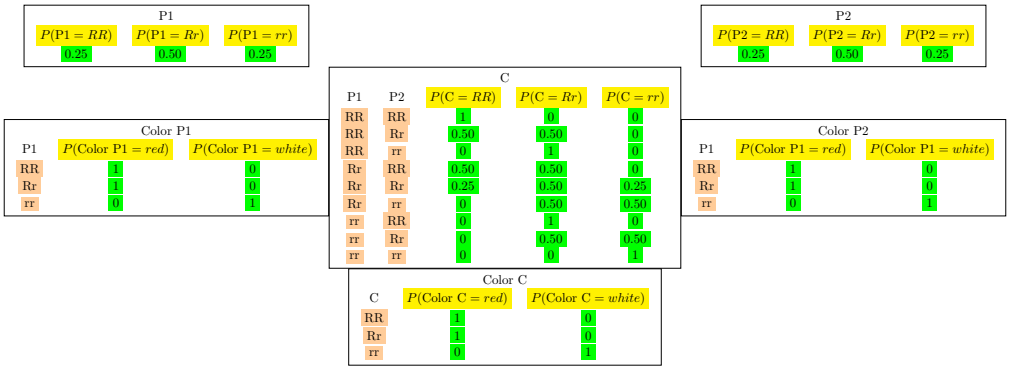


Table 2: The conditional probabilities used to simulate 5000 datapoints.

APPENDIX D. CONDITIONAL PROBABILITIES TO SIMULATE DATA FOR THE DIAGNOSTIC CHEST CLINIC NETWORK

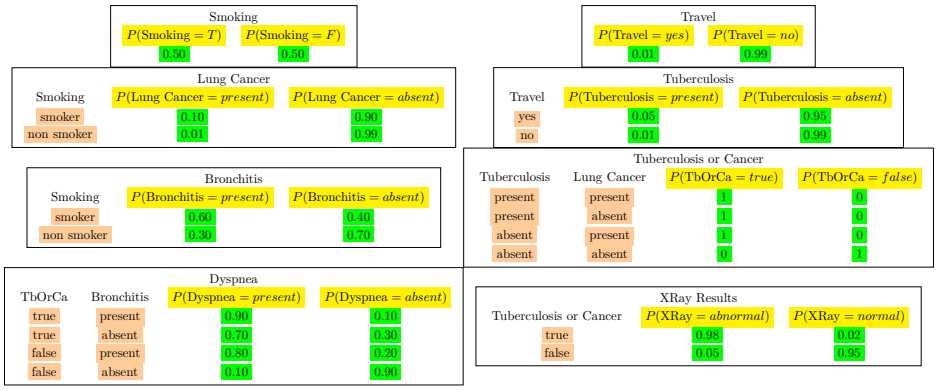


Table 3: The conditional probabilities used to simulate 10000 datapoints.

ACKNOWLEDGMENTS

We thank our colleagues from the School of Mathematical Sciences at Monash University who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] Chickering, D. M., Geiger, D., and Heckerman, D., "Learning Bayesian networks: search methods and experimental results. In *Learning from Data: Artificial Intelligence and Statistics V* (eds Fisher, D. and Lenz, H.-J.)," *Lecture Notes in Statistics* **112**, 112–128 (1996).

[2] Glover, F., "Tabu Search - Part 1," *ORSA Journal on Computing* **1**(2), 190–206 (1989).

[3] Glover, F., "Tabu Search - Part 2," *ORSA Journal on Computing* **2**(1), 4–32 (1990).

[4] Margaritis, D. and Thrun, S., "Bayesian network induction via local neighborhoods," technical report, DTIC Document (2000).

[5] Hrycej, T., "Gibbs sampling in Bayesian networks," *Artificial Intelligence* **46**(3), 351–364 (1990).

[6] Riggelsen, C., "MCMC learning of Bayesian network models by Markov blanket decomposition," in [*Proceedings of the 16th European Conference on Machine Learning*], *ECML'05*, 329–340, Springer-Verlag, Berlin, Heidelberg (2005).

[7] Ram, R. and Chetty, M., "MCMC based Bayesian inference for modeling gene networks," in [*Proceedings of the 4th IAPR International Conference on Pattern Recognition in Bioinformatics*], *PRIB '09*, 293–306, Springer-Verlag, Berlin, Heidelberg (2009).

[8] Keith, J. M., Sofronov, G. Y., and Kroese, D. P., "The generalised gibbs sampler and the neighborhood sampler. in *Monte Carlo and Quasi-Monte Carlo Methods 2006*," *Springer Berlin Heidelberg* **31**, 537–547 (2008).

[9] Heckerman, D., Geiger, D., and Chickering, D. M., "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning* **20**(3), 197–243 (1995).

[10] Hastings, W. K., "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika* **57**(1), 97–109 (1970).

[11] Robert, C. P. and Casella, G., [*Monte Carlo statistical methods*], Springer (2004).

[12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., "Equations of state calculations by fast computing machines," *Journal of Chemical Physics* **21**(6), 1087–1092 (1953).

[13] Casella, G. and George, E., "Explaining the gibbs sampler," *The American Statistician* **46**(3), 167–174 (1992).

[14] Gelman, A., "Iterative and non-iterative simulation algorithms," Technical Report 347, University of California, Dept. of Statistics (1992).

[15] Brooks, S. P. and Gelman, A., "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics* **7**, 434–455 (1997).

[16] Lauritzen, S. L. and Spiegelhalter, D. J., "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society, Series B (Methodological)* **50**(2), 15224 (1988).

Integrating heterogeneous datasets for cancer module identification

A. K. M. Azad
PhD student
Monash University
School of Mathematical Sciences
Clayton campus, Wellington Road, Clayton, Victoria 3800, Australia
mailto:a.azad@monash.edu

Abstract

The availability of multiple heterogeneous high-throughput datasets provides an enabling resource for cancer systems biology. Types of data include: Gene Expression (GE), Copy Number Aberration (CNA), miRNA expression, Methylation, and Protein-Protein Interactions (PPI). One important problem that can potentially be solved using such data is to determine which of the possible pairwise interactions among genes contribute to a range of cancer-related events, from tumorigenesis to metastasis. It has been shown by various studies that applying integrated knowledge from multi-omics datasets elucidates such complex phenomena with higher statistical significance than using a single type of dataset individually. However, computational methods for processing multiple data types simultaneously are needed. This chapter reviews some of the computational methods that use integrated approaches to find cancer-related modules.

1 Introduction

Cancer is a common genetic disease involving a range of factors. Genomic, epigenomic, and differential gene expression aberrations all play vital roles in a cancer's initiation, development and malignance [1]. It has been reported by various studies that cancer related activities including cell proliferation, angiogenesis, and metastasis are associated with abrupt changes in regulatory and signaling pathways [2–6]. Mutations involving somatic and copy number aberrations of some genes can either directly affect some key pathways, or have a cumulative effect when they occur across network modules representing common functional activities in cancer [7, 8]. Consequently, identifying cancer modules is of primary importance to the effective diagnosis and treatment of cancer patients.

One of the core steps of cancer module identification involves modeling gene-gene relationships in a network. Many algorithms have been developed for this purpose, but most apply only to homogeneous datasets, that is, data of only one type, usually GE data or PPI information [9–15]. Most of the methods relying only on GE data apply differential expression analysis but it is often hard to determine whether such variations in expression are causative or merely an effect of complex diseases [16]. Differential expression analysis can produce false negatives and false positives: some important genes in cancer related pathways may not be identified as differentially expressed, whereas some differentially expressed genes may not be relevant to cancer [17]. Typically CNA regions identified by some approaches [18–20] using only CNA datasets are spatially extensive, which makes it difficult to identify a specific gene causing genomic aberration [21]. PPI can provide important information in characterizing topological properties of the network involving cancer genes [7]. However, PPI information for multiple cell types and developmental stages is still incomplete, which limits its usefulness in developing methods for cancer module identification.

Recent studies have demonstrated the 'genomic footprint' of driver mutations on gene expression [21–23]. This happens when somatic mutations and copy number aberrations affect a genes' transcriptional changes directly or indirectly [24] and thus perturb some core pathways relevant to cancer growth and malignance [1]. Research carried out for The Cancer Genome Atlas on both glioblastoma [25] and ovarian carcinoma [26] demonstrated the simultaneous occurrences of mutations, copy number aberrations, and gene expression changes in a significant number of patients in the core components of some key pathways

[see Note 1]. In this chapter we discuss some methods that find cancer related modules by integrating multiple heterogeneous datasets.

This chapter is organized as follows. We first briefly introduce some of the main sources of data that can be used and the required preprocessing steps essential for subsequent integrated analysis. Then, we describe methods that integrate information from heterogeneous data sources to find cancer related modules/sub-networks [see Note 1]. Finally, we address some approaches for validating identified modules.

2 Data Sources

Gene Expression data from cancer samples can be primarily found in the database GEO (Gene Expression Omnibus) [27]. It is a database of gene expression values measured using high throughput hybridization arrays (also known as chips or microarrays). Sample values are reposited both in raw and normalized versions. Another comprehensive collection of gene expression data from various cancer samples is the The Cancer Genome Atlas (TCGA) [28]. There are three different levels of datasets available in TCGA: Level 1 consists of low-level (not normalized) data for a single sample probe, Level 2 consists of normalized single sample probe data, and Level 3 consists of aggregated gene-level data (grouped by mapped probes with gene symbols). **Mutation, Copy number aberration, DNA methylation,** and **miRNA expression** datasets can also be found in TCGA data portal.

Preprocessing is an important step in data integration, especially when paired samples are used [see Note 2]. Preprocessing of GE values includes scale transformation, imputing missing values, handling redundancies, pattern standardization (i.e. normalising to a zero mean and unit standard deviation), and other transformations [29]. Preprocessing of CNA data in microarray chips is typically more complex than that of GE data, and can include quantile normalization, imputing missing values, summarizing multiple probes at a single locus (with mean or median), segmentation of genomic regions, and mapping segmented CNA values in genomic regions into corresponding gene symbols [17, 30]. Probe level methylation data from CpG sites can be normalized between 0 and 1 by finding the following ratio [31] :

$$\beta_i = \frac{max(M_i, 0)}{(max(M_i, 0) + max(U_i, 0) + \alpha)} \tag{1}$$

where β_i is the Beta-value for an i^{th} interrogated CpG site, and M_i and U_i are the intensities measured by the i^{th} methylated and unmethylated probes. After background adjustment, intensities (M_i and U_i) may become negative, but in the above definition those negative values are reset to 0. Again, when both M_i and U_i intensities are very low, a constant offset α (default value = 100) is added to the denominator to regularize Beta-value, as suggested by Illumina [31].

3 Methods for integrating heterogeneous datasets

Figure 1 generalizes a possible approach that integrates multiple heterogenous datasets in order to find cancer related modules in a gene-gene network. The gene-gene network can be modeled either by exploiting combined knowledge from multiple datasets or by merging individual networks built upon corresponding datasets. In these networks, nodes represent genes and the edges can be modeled as the relationships (i.e. directed and/or

undirected) among them. PPI information can be useful at various stages of network-modeling. After modeling the integrated network various module detection techniques such as, optimization models, hierarchical clustering, etc. can be applied to find cancer related modules. The following sections describe some of the methods that use integrated approaches for cancer module identification.

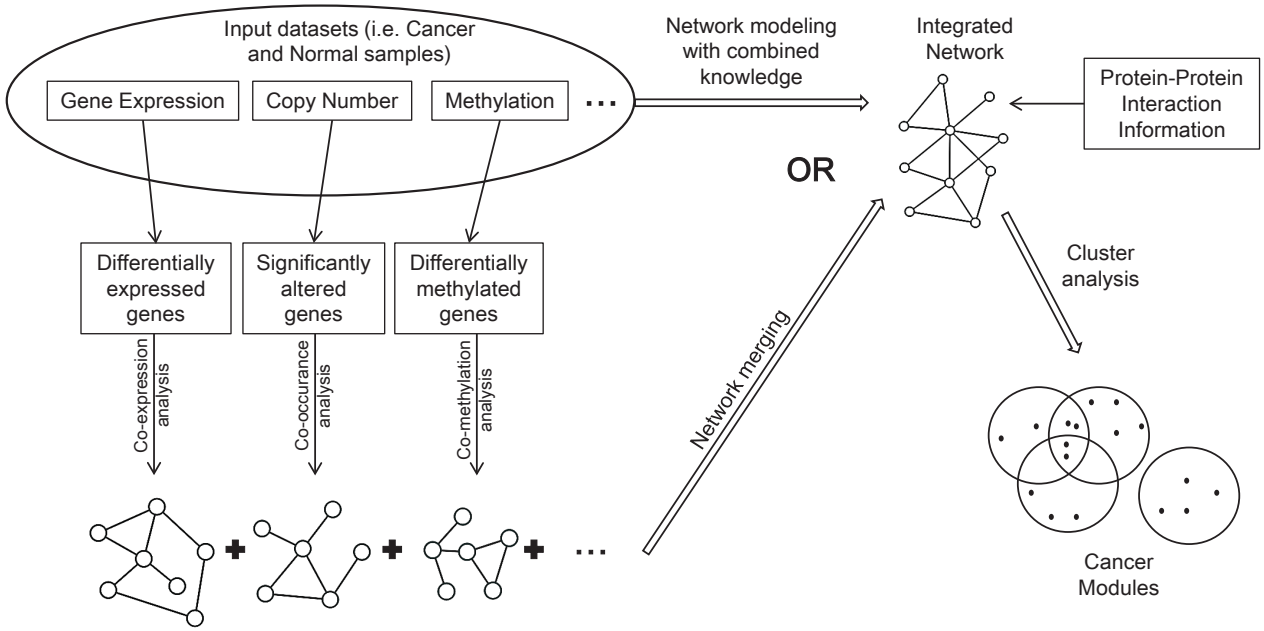


Figure 1. Schematic diagram of a possible integrated approach for cancer module identification. Each input dataset contains both caner and normal samples. In network modeling, genes are identified based on differential information in the two-conditional studies (cancer vs normal), and edges can be defined according to pair-wise correlation.

3.1 *iMCMC*

A method known as iMCMC (**i**dentify **M**utated **C**ore **M**odule in **C**ancer) [32] was developed for the simultaneous analysis of three heterogeneous datasets: Gene Expression (GE), Copy Number Aberration (CNA) and sequence mutations. These are combined to infer a network in which core cancer modules are identified [see Note 3]. The method involves an optimisation model followed by statistical significance tests. This method initially starts with building two different networks, one generated from GE data and the other by combining somatic mutations with CNAs over common samples. These two networks are then combined to construct an integrated network.

First, a binary matrix A_0 is constructed in which the columns represent the paired samples containing somatic mutations and CNAs, and the rows represent genes that the samples have in common. Each entry in A_0 is set to 1 if a mutation occurs in the corresponding gene and sample, or if there is a statistically significant copy number variation detected; otherwise the entry is set to 0. Genes that are mutated in the same samples in A_0 are combined into larger *metagenes*, and thus a new matrix A called the mutation matrix is obtained. Another data matrix, B is built from the expression values. Its entries are real values representing the relative expression of a given gene in a particular sample. The following two paragraphs explain the methodologies for constructing the *Expression Network* (**EN**) and *Mutation Network* (**MN**) from the data matrices B and A, respectively.

Constructing the Expression Network: The *Expression Network* is based on the gene expression dataset. In this network, both nodes and edges are weighted. Nodes represent genes and their corresponding weights reflect the extent to which a mutation in that gene affects the expression levels of other genes. Each edge weight is defined as the absolute correlation between the expression levels of the two corresponding genes.

The definition of nodes in the **EN** depends on both data matrices A and B. New sets of genes and samples are defined as: $G^{'} = G_A \cap G_B$ and $S^{'} = S_A \cap S_B$, where (G_A, S_A) and (G_B, S_B) are the sets of genes and samples in the two data matrices A and B, respectively. For each gene $g_i \in G^{'}$, the corresponding samples in $S^{'}$ are classified into two groups, based on that gene's mutation status in A. The numbers of samples in each group are denoted $n_i^{(1)}$ and $n_i^{(2)}$. Then, for each g_i, a mutation-correlated expression vector $e_i = \left(e_i^{(1)}, e_i^{(2)}\right)$ is constructed, where $e_i^{(1)}$ and $e_i^{(2)}$ are defined as follows:

$$\begin{aligned} e_i^{(1)} &= \left\{ b_{ki} : a_{ki} = 1, k \in S^{'} \right\}, \\ e_i^{(2)} &= \left\{ b_{ki} : a_{ki} = 0, k \in S^{'} \right\}. \end{aligned} \tag{2}$$

Here a_{ki} and b_{ki} denote the entries for the i-th gene and k-th sample in the data matrices A and B, respectively. To determine whether there are significant differences between the expression levels in $e_i^{(1)}$ and $e_i^{(2)}$, p-values are calculated using *mattest* in MATLAB. A small p-value indicates that mutations in the gene in question affect the expression levels of other genes. Since there should be a minimum of two samples in each group for conducting this test, the set of nodes G in the **EN** is defined as follows:

$$G = \left\{ g_i \in G^{'} : n_i^{(1)} \geq 2, n_i^{(2)} \geq 2 \right\}. \tag{3}$$

And the weight of each node in **EN** is defined as follows:

$$f_i = 1 - \frac{1}{d} \sum_{r=1}^{d} p_r, \quad \forall g_i \in G \tag{4}$$

where d is the total number of genes in G_B and p_r is the p-value calculated for gene g_r as described above. The weight u_{ij} of any edge in G is defined as the absolute Pearson correlation between two mutation-correlated expression vectors e_i and e_j, among the samples in $S^{'}$. In the case of *metagenes*, node and edge weights are defined as the averages of the corresponding values of their constituent genes.

Constructing the Mutation Network To build the *Mutation Network* (**MN**) from the mutation matrix A, the same gene set G is used as for the network **EN**. The weight of any node (or gene), $g_i \in G$ is defined as follows:

$$h_i = \frac{m_i}{m}, \tag{5}$$

where m is the total number of samples in A and m_i is the total number of mutations occurring in the samples of A for a particular gene g_i. The weight v_{ij} of any edge between genes (g_i, g_j) in **MN** is defined as the ratio of the number of samples in which *exactly* one of the gene pair is mutated to the number of samples in which *at least* one of the gene pair is mutated in A.

The Integrative Network: An integrative network $\mathcal{M}$ is constructed by combining the expression network **EN** with the mutation network **MN**. It is necessary to first adjust the weights of nodes and edges in **EN** and **MN** so that they become comparable. Two balancing terms, ξ and η, are defined for the networks **EN** and **MN** respectively as follows:

$$\xi = \tfrac{u}{f}, \eta = \tfrac{v}{h}, \tag{6}$$

where $f = max\,(f_i)$ and $u = max\,(u_{ij})$ in **EN**, and $h = max\,(h_i)$ and $v = max\,(v_{ij})$ in **MN**. Now, if $F = \{f_i\}$ and $U = \{u_{ij}\}$ then the edge weights U and node weights ξF are said to have *balanced values* in **EN**. Similarly, if $H = \{h_i\}$ and $V = \{v_{ij}\}$ then the edge weights V and node weights ηH have balanced values in **MN**. A relative importance term can also be introduced to modify the relative impact of the two networks **EN** and **MN** on the integrated network. Let k denote the relative importance of **MN** relative to **EN** and set $\delta \cdot \left(\frac{u}{v}\right) = k$, so $\delta = k \cdot \frac{v}{u}$. In the remainder of this description, we set $k = 1$. Thus, node weights c_i and edge weights w_{ij} can be defined as follows:

$$\begin{aligned} w_{ij} &= \delta \cdot u_{ij} + v_{ij}, \\ c_i &= \delta\xi \cdot f_i + h_i, \end{aligned} \tag{7}$$

where, $i, j = 1, ..., n$. Here, n is the total number of genes in G.

An optimization model for identifying core cancer pathways: The final step of this approach is to identify some core modules in the integrative network $\mathcal{M}$, where each such module contains genes with both high node-weights and high edge-weights. For this purpose, an optimization model (previously reported by Wang *et al.* [33]) is employed. The optimisation problem is stated as follows:

$$\begin{aligned} max \quad & \textstyle\sum_i \sum_j w_{ij} x_i x_j + \lambda c_i x_i, \\ s.t. \quad & x_1^\beta + x_2^\beta + ... + x_n^\beta = 1, \\ & x_i \geq 0, i = 1, ..., n, \end{aligned} \tag{8}$$

where the non-negative vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ contains the degrees of each node in a particular module (sub-network). The first term in the objective function states the inter-connectivity within the module, whereas the second term specifies the degree of association between the nodes and the module. The role of the positive parameter λ here is to balance these two terms [see Note 4]. In this model, the regularization constraint over the variable $\mathbf{x} = (x_1, x_2, ..., x_n)$ controls the number of nodes to be selected in the

module and the parameter β adjusts the strength of this regularization. Here we set β=1 to find small-sized core modules.

The following iterative algorithm [33] provides an easy solution of the above optimization model by finding a local maximum in the vicinity of a predetermined initial approximate solution:

$$x_i^{t+1} = \left(x_i^t \frac{2(WX)_i + \lambda c_i}{2X^T WX + \lambda \sum_i c_i x_i^t} \right)^{\frac{1}{\beta}}, \tag{9}$$

where $W = \{w_{ij}\}$ is the $n \times n$ edge weight matrix, and $X = (x_1^t, x_2^t, ..., x_n^t)^T$ is the solution vector at the t-th iteration. The non-zero entries in solution vector $\mathbf{x}$ define a particular module (sub-network) where in practice the entries are defined as zero if they are less than 0.1. Once a locally optimal solution is obtained, corresponding nodes are removed from the network and the whole process is repeated again to find additional modules.

3.2 *Wen et al.*

The method of Wen *et al.* integrates DNA methylation, gene expression and protein-protein interaction datasets to identify causal network modules in colorectal cancer [34]. The method starts with collecting a set of candidate causal genes. This collection is the union of a set of differentially methylated genes and a common subset of known cancer genes from DNA methylation chips, the Cancer Gene Census (CGC) [35], and tumor associated genes in the TAG database. Employing a minimum multi-set cover strategy due to Kim *et al.* [36], a gene is determined to be differentially methylated if its comparative β value (a measurement of DNA methylation level) between tumor and paired non-tumor samples is ≥ 0.2 [37, 38].

Next, a comprehensive protein-protein interaction (PPI) network is developed integrating five curated human PPI databases: HPRD [39], BioGrid [40], IntAct [41], MINT [42], and Reactome [43]. Only those interactions that are found in at least three of these databases are considered. The resulting network contained 7001 nodes and 19,188 edges, where each edge e is assigned a weight calculated as follows:

$$w(e) = 1 - |cor(x, y)| = 1 - \left| \frac{\sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{m} (y_i - \overline{y})^2}} \right|. \tag{10}$$

Here $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_m)$ are expression profiles of the two nodes in an edge e, and $\overline{x}$ and $\overline{y}$ are mean values of x and y, respectively. This PPI network is further decomposed into network modules by applying the Markov Clustering algorithm [44], but only those modules are selected which contain at least one candidate causal gene. The activities of each network module M_i in sample S_j are calculated as follows:

$$M_{ij} = \frac{\sum_{g_m \in \{CGC \cap M_i\}} \sum_{(g_m, g_n) \in E(g_m)} \frac{g_{mj} + g_{nk}}{2}}{\sqrt{\sum_{g_m \in \{CGC \cap M_i\}} \#(E(g_m))}}, \tag{11}$$

where $E(g_m)$ is the set of edges belongs to the candidate causal gene g_m in module M_i, $\#(E(g_m))$ represents the total number of edges in $E(g_m)$, and g_{mj} is the normalized

gene expression value of the gene g_m in sample S_j. Next, a classifier is built for selecting the causal modules as follows:

$$\begin{array}{c} \left\|S-\overline{S}_{control}\right\|_2^2-\left\|S-\overline{S}_{case}\right\|_2^2<0, \quad for \quad S\in S_{control} \quad , \\ \left\|S-\overline{S}_{case}\right\|_2^2-\left\|S-\overline{S}_{control}\right\|_2^2<0, \quad for \quad S\in S_{case} \quad , \end{array} \tag{12}$$

where S, $S_{control}$, S_{case}, $\overline{S}_{control}$, and $\overline{S}_{case}$ are the sample, the set of non-tumor samples, tumor samples, the center of non-tumor samples, and the center of tumor samples set, respectively. These classifier conditions can be furthur simplified as follows (for details, see supplementary texts of original article):

$$\mathbf{C}\cdot(x_1,x_2,...,x_k)^T\leq 0, \tag{13}$$

where x_i is an indicator variable having value 1 if module M_i is selected, and 0 otherwise; and $\mathbf{C}$ is a matrix that is defined as a function of M_{ij} as follows:

$$\mathbf{C}:=\left\langle\left(M_{1i}-\frac{M_{11}+...+M_{1n}}{n}\right)^2,...,\left(M_{ki}-\frac{M_{k1}+...+M_{kn}}{n}\right)^2\right\rangle \tag{14}$$

Here, any element C_{ij} of the above matrix $\mathbf{C}$ represents the contribution of the module M_j to the i^{th} sample condition. The objectives of this classifier are two-fold: 1) classifying tumor and non-tumor samples, 2) identifying a small number of modules. This module identification problem is modelled as a binary integer linear programming problem as follows:

$$\begin{array}{c} min_{x_1,x_2,...,x_k}\sum_{j=1}^k x_j+\lambda\sum_{i=1}^s\sum_{j=1}^k C_{ij}\cdot x_j \\ s.t. \quad \mathbf{C}\cdot(x_1,x_2,...,x_k)^T\leq 0 \\ \sum_{i=1}^k x_i\geq 1, \quad x_i=0,1, \quad i\in\{1,2,...,k\}\,, \end{array} \tag{15}$$

where s is the number of samples. In this objective function, the first term encourages a small number of modules to be found whereas the second term implies the maximization of the classification abilities of modules by minimizing $\mathbf{C}\cdot(x_1,x_2,...,x_k)^T$. λ is the controlling parameter which balances the trade off between those two terms. However, this binary integer linear programming model for module identification is computationally extensive. Therefore, this problem is further resolved by reformulating the model to a simple linear programming model where the binary variables $x_i\in\{0,1\}$ are relaxed to a continuous variables $x_i\in[0,1]$. For further detail, see Note 5.

3.3 *Cerami et al.*

The method of Cerami *et al.* [45] is an integrated approach for identifying core pathways altered in glioblastoma. It combines sequence mutation, Copy Number Aberration (CNA) and Protein-Protein Interaction (PPI) datasets. The first step of this method is to construct a global Human Interaction Network (HIN) from literature curated data sources only. To cover more interaction information, the HIN is constructed based on the union of a) interactions obtained from the HPRD website (http://www.hhprd.org/) and b) various signaling pathway databases, specifically Reactome, NCI/Nature Pathway Interaction DB, and MSKCC Cancer Cell Map from Pathway Common (http://www.pathwaycommons.org). Information from the last of these pathway sources was in BioPAX format, which is represented as subgraphs of biochemical networks. A set of rules was defined to map these

subgraphs into binary interaction data. After removing all redundancies and self-directed interactions, the HIN contained 9,264 genes and 68,111 interactions.

Sequence mutation and copy number datasets of Glioblastoma Multiforme (GBM) for paired samples were collected from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). Copy number aberration data was analyzed using the RAE algorithm [19] which discretizes all isoforms of autosomal genes into multiple putative aberration states, and finds statistically aberrant regions with q-values. Next, the statistical significance of each gene's aberration is defined as the minimum of the q-values of all the spanning regions over the corresponding gene's coding locus. A set of altered genes is identified, where a gene is defined as altered if it has a validated non-synonymous somatic nucleotide substitution, or a homozygous deletion, or a multi-copy amplification only.

Next, a GBM-specific network was constructed in which the node set is the union of the set of altered genes and a set of linker genes. For each gene in the altered gene set, the corresponding neighbour genes are identified in the HIN. Neighbour genes having degree one are trivially ignored, as they are connected to exactly one altered gene. The remaining neighbour genes with degree ≥ 2 have the potential to connect two or more altered genes, and are thus considered to be candidate linker genes. Only linker genes that are found to be statistically significant by a hypergeometric test among all other candidate linker genes are further assessed. The null hypothesis is: *the linker genes connect the observed number of altered genes in HIN only by chance.* P-values from the statistical assessment of this hypothesis are further corrected using the Benjamini-Hochberg procedure [46] giving corresponding q-values, and the genes having q-values ≤ 0.05 are selected as a final list of linker genes. The final network contained six linker genes connecting 66 GBM altered genes, and their corresponding PPI interactions in the HIN.

To find network modules in the resulting GBM-specific network, the *edge-betweenness algorithm* was applied. Originally proposed by Girvan and Newman [47], this algorithm applies a divisive approach where at each iteration an edge with the highest edge-betweenness score among all other edges is identified and removed from the network in order to reveal modular structure. The *edge-betweenness score* of a particular edge is defined as the number of shortest paths between pairs of nodes that traverse that edge [47]. More specifically, the shortest paths between all pairs of vertices are identified, and then for each edge the number of shortest paths that include that edge is counted and considered as the *edge-betweenness score* for that particular edge. After each edge removal, the *edge-betweenness scores* of the edges of the updated network are recalculated. (Only those edges which are affected by the particular edge removal require recalculation of this score.) To obtain a partition yielding the best modular structure, *network modularity* [48] is also calculated after each edge removal. This process continues until there are no remaining edges. The maximum network modularity score obtained during this process indicates the optimal number of edges to be removed. The *network modularity score* is defined as follows:

$$M = \sum_{s=1}^{N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right], \tag{16}$$

where N_M is the number of modules, l_s is the number of edges within module s, L is the total number of edges in the network, and d_s is the summation of the degrees of all the edges within module s. *Modularity* quantifies the fraction of network edges connecting the nodes within modules minus the expected number of network edges obtained by forming

random connections among the nodes within the module, subject to the same modular divisions. A value of M close to 0 indicates that the number of within-module edges is consistent with random formation, whereas a value close to 1 indicates stronger modular structure. This procedure results in a set of modules extracted from the GBM-specific network.

3.4 *VToD*

VToD [17] integrates Gene Expression (GE), Copy Number Aberration (CNA) and PPI (Protein-Protein Interaction) datasets in order to find cancer related modules in glioblastoma and ovarian cancer patients. The GE and CNA data matrices are obtained from TCGA data portal [28]; both are Level 3 datasets. The PPI dataset is obtained from Cerami *et al.* [45]. This method provides an integrated framework that infers pair-wise relationships between genes based on both *data-driven* and *topological properties* [see Note 3]. A data-driven property of a pair of genes is a correlation observed between the data obtained for those genes. These correlations may be of three types: GE-GE, GE-CNA, or CNA-CNA correlations. Data-driven properties also include the indirect relationships discussed below. Topological properties are connections observed in PPI networks.

Constructing a Gene-Gene Relationship Network: The method starts with a set of seed genes S, thought to be related to cancer progression and malignance. This set is a union of a set of differentially expressed and a set of significantly altered genes. Differential expression is identified using a two-tailed pooled t-test, and the corresponding p-values are corrected using the Bonferonni correction. A set of significantly altered genes is found by mapping gene symbols to the collected focal aberrant regions [25,26] identified by GISTIC [18] and RAE [19] algorithms. Next, the Gene-Gene Relationship Network (GGR), a weighted undirected network, is defined. Nodes of this network represent the seed genes and edges represent direct or indirect pair-wise relationships among genes. The absolute value of the Pearson correlational coefficient (PCC) is used to identify pair-wise relationships between genes, and as a weight on each edge.

For any gene-pair ($gene_i$,$gene_j$), all three types of absolute PCC value (GE-GE, GE-CNA and CNA-CNA) are calculated, depending on data availability. The maximum of these is defined as the data-driven property of that particular gene-pair and termed its r_value. For the gene-pairs ($gene_i, gene_i$) this r_value is considered to be 0. If an r_value is greater than some threshold then a direct relationship is defined for that particular gene-pair. The gene-pairs for which a direct relationship is not found may still be connected if an indirect relationship is identified. An indirect relationship between two particular genes is a statistically significant simple path joining those two genes in the PPI network [see Note 6]. To identify such statistically significant paths, the observed paths between particular gene-pairs are compared with the path in a random PPI network, which is generated in such a way that gene interactions are randomly assigned while the network topology and gene expression values are the same as those in the observed PPI network. In other words, the random PPI network has the same number of interactions (edges) as the observed one, but the genes (nodes) of the observed PPI network are shuffled in the random PPI network. The null hypothesis for this statistical significance test is: *the geometric mean of r_values of the simple path found in random PPI network is greater or equal to that of the observed path.* In order to reduce the time complexity, a heuristic search is applied only for those gene-pairs for which there is a connection in the PPI

network [see Note 6]. All the simple paths between two genes with a fixed path length are identified using a Breadth First Search (BFS) algorithm. Furthermore, only those simple paths are selected in which all the constituent genes have either GE, or CNA, or both datasets available. Since there can be multiple such paths found, a path P^* with maximum average PPI connectivity is selected:

$$P^* = \max_P \left\{ \frac{1}{n} \sum_{l=1}^{n} norm\_deg(gene_i) \right\} \tag{17}$$

where $norm\_deg(gene_i)$ is the degree of connectivity for $gene_i$ normalised by the global maximum connectivity in the PPI network, and n is the number of genes along the path. The statistical significance of the path P^* is measured as above, and is selected if its corresponding p-value is bellow 0.05. For the gene-pairs for which a statistically significant path is found, an edge is added to the GGR network, where the edge weight is the average of all the pair-wise r_values of gene-pairs along the path P^*.

Module Detection: Next, a **V**oting based module detection algorithm identifies overlapping modules in the GGR network by combining **To**pological and **D**ata-driven properties. The name of the method - VToD - is an acronym for this procedure. First, a pairwise score (vote) is calculated for every pair $\{g, m\} \in S$ using the following equation:

$$vote(g, m) = \frac{norm\_deg(m)}{SPL(g, m)} + r\_value(g, m) \tag{18}$$

where above $norm\_deg(m)$ is the degree of connectivity of m normalised by the global maximum PPI connectivity, $SPL(g, m)$ is the shortest path length between the two genes in the PPI network, and $r\_value(g, m)$ is the relationship value calculated for the constructed network GGR. This definition states how much vote-score a gene m can get from another gene g, for any pair $\{g, m\} \in S$. Note, the $vote(g, m)$ score in the above equation is not a symmetrical measure because of the definition of the topological property ($norm\_deg(m)$ in above equation). A high score indicates either i) a gene-pair $\{g, m\}$ has high data-driven relationship r_values, or ii) any gene g is interacting with a gene m with a high topological value in the PPI network. Note, the shortest path length SPL is constrained by a user-defined threshold to control the compactness of the module. If any of the shortest paths has length above that threshold, that path is ignored.

Next, for any gene $g \in S$, corresponding vote-scores with all the genes $m \in S$ (including g) are stored in a table. Here, $vote(g, g)$ is defined with the $norm\_deg(g)$ only, since $r\_value(g, g) = 0$, and $SPL(g, g)$ is not defined for the PPI network as it doesn't contain any self-loop. Next, the table for the gene g containing vote-scores of all the genes $m \in S$ is sorted in descending order of vote-score. In that sorted table, the ranking of each gene m is defined as its local rank. Then, in that sorted table, the cumulated vote-score from the top-ranked vote-scores of the m ($\in S$) genes is calculated. If the cumulated vote-score of the top-ranked m gene(s) is(are) within the top k% (a user-defined threshold) of total cumulative vote-score in that particular table (for gene g), then that(those) top-ranked m gene(s) are considered as candidate representative gene(s) of that particular gene g. Next, if the $vote(g, m)$ score(s) of this(these) top-ranked m gene(s) are within top $vote\_th$% (a user-defined threshold) of the distribution of all pair-wise vote-scores (considered as the global rank of the gene m), then this(these) m gene(s) are finally selected as a representative gene(s) of the particular gene g. Thus, this technique makes it possible to find

overlapping modules in the network by allowing multiple representative m genes to be selected for a particular gene g. More importantly, this method can select a gene $m \in S$ (i.e. a hub-gene in PPI network) as a representative gene for multiple $g \in S$ genes, thus revealing a modular structure. Next, these modular structures, called 'pre-modules', are formed, each with a representative gene m in the centre and aggregating all the genes g that chose m. A pre-module is defined as the initial state of a module before merging it with other pre-modules to get the final module. After removing redundancies and small pre-modules (typically with ≤ 3 genes), a module merging algorithm is conducted. Two pre-modules merge if their pair-wise members are closely connected in the PPI network (topological property) or highly related in GGR (data-driven property). For this purpose, a pair-wise merging value $MV(C_i, C_j)$ between any two pre-modules C_i and C_j is calculated as follows:

$$MV(C_i, C_j) = \frac{IC(C_i, C_j)}{n_i} + \frac{1}{n_i \times n_j} \sum_{g_k \in C_i} \sum_{g_l \in C_j} r\_value(g_k, g_l) \tag{19}$$

where n_i and n_j are the sizes of two pre-modules C_i and C_j, respectively, and $n_i \leq n_j$ (Note, here it is assumed that, C_i is bigger than C_j). Inter-connectivity $IC(C_i, C_j)$ is a kind of topological property relating C_i to C_j: it is the proportion of genes in C_i having at least one PPI partner in C_j. The second term in the above equation denotes the data-driven property for the pair C_i and C_j: it is the average of the gene-gene relationship values over all pairs of a gene in C_i with a gene in C_j. At each iteration of the module merging procedure, two pre-modules with the highest pair-wise merging value (calculated using the above equation) are merged together and replaced by the newly merged module. This merging process continues until the highest pair-wise merging value at some iteration becomes less than some threshold $merging\_th$ (for the details of this threshold selection, see supplementary method of original article).

4 Validating cancer sub-networks

There are several ways to validate cancer modules identified by the above procedures. Most of them involve statistical hypothesis testing and are specific to the methodology used to identify modules. However, there are a few general techniques that can be used to validate modules, as follows:

4.1 Topological validation

Ideally, a modular network is expected to have dense intra-module connections but sparse inter-module connections. Therefore, proposed networks can be assessed for both high density of connections within modules and high separability of component modules. Equation 16 states the modularity measurement [48] which compares the connection-density of a particular module with that of a module formed by making random connections among its constituent genes. Similarly, the following equation quantifies the *separability* of modules [48].

$$seperationScore = \sum_{s=1}^{N_M} \left[1 - \left(\frac{2l_s}{d_s} \right)^2 \right] \tag{20}$$

where N_M is the number of modules, l_s is the number of edges within module s, and d_s is the summation of the degrees of all the edges within module s. Both 'Modularity' and 'Separability' scores can be calculated using above equations (Equation 16, and Equation 20, respectively) where higher 'Modularity' value indicates stronger modular structure, and higher 'Separability' score indicates a particular module is more easily separable from the original network topology by deleting some edges, respectively.

4.2 Enrichment analysis

f-measure: Modules can also be validated using a quantity known as an f-measure [48]. This quantity evaluates the accuracy of identified modules by comparing them with known reference modules such as: GO functional categories, known biological pathways, and others. f-measure can be calculated using the following equation:

$$f - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{21}$$

where, $Precision = \frac{|M \cap F_i|}{|M|}$ and $Recall = \frac{|M \cap F_i|}{|F_i|}$ are the *true positive rate* and *positive predictive value*, respectively. Here, M is a particular module and F_i is a known functional module. For example, a Module M (typically, a set of genes) is mapped to a known functional category F_i: 'Cell Cycle', then the $Precision$ and the $Recall$ are the fractions of genes common to both M and F_i to the size of M, and to the size of F_i, respectively. Bigger modules will have higher $Recall$ values, whereas smaller modules will have higher $Precision$ values. Therefore, the accuracy of any identified module M can be measured by calculating the harmonic mean of these two values as $f - measure$.

Hypergeometric analysis: A hypergeometric test can also be used to assess modules statistically [48]. P-values can be calculated using the hypergeometric distribution to indicate the significance of correspondence between a module and a known functional category.

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\left(\binom{|X|}{i}\binom{|V| - |X|}{n-i}\right)}{\binom{|V|}{n}} \tag{22}$$

where $|V|$ is the total number of genes (i.e. all the genes in human genome), $|X|$ is the number of genes in a known functional category (such as a GO term or known pathway), n is the number of genes in an identified module, and k is the number of genes in the intersection of that particular module with the known functional category. Here, a low p-value indicates that the identified module is significantly enriched in known functions or pathways. For example, a 'dhyper' function in a built-in R-package called 'stats' can be used to calculate p-values of the hypergeometric test [49].

5 Notes

1. In general, most of the integrative approaches that aim to find cancer related modules are based on a common hypothesis: *tumors are characterised by aberrations in specific biological modules that are critical in terms of cancer initiation and malignance*. There are two major steps in such methods, 1) building the network model,

and 2) identifying modules (sub-networks). In defining gene dependencies in network models, some methods rely on PPI information only [15, 45], some on data-driven information only [32, 50–52] and some on both of those properties [13, 17].

2. In any integrated approach, higher statistical significance is achieved by using paired sample data rather than unpaired data. Moreover, pair-wise relationships between genes obtained by integrative approaches applied to unpaired sample data may produce false positive results [24]. Here, paired data indicates using various heterogeneous data types (eg. GE, CNA, methylation, miRNA) measured on the same samples. However, appropriate data normalization and standardization techniques are crucial to obtain correct inferences using paired data.

3. Integrating as many heterogeneous datasets as possible can improve characterizations of driver genes and cancer modules. Zhang *et al.* found that the integration of three heterogeneous datasets (GE + CNA + mutation) provides additional useful information and can produce statistically significant core modules in both glioblastoma and ovarian cancer compared to the integration of two heterogeneous datasets (GE + mutation, or CNA + mutation) [32]. Similarly, Azad *et al.* showed that modules found by combining topological and data-driven properties (PPI + GE + CNA) of gene-pairs result in better functional enrichment than those found by using only topological (PPI), or only data-driven (GE + CNA) properties [17]. Akavia *et al.* reported that combining CNA and GE provides greater sensitivity for identifying RAB27A as a novel driver gene in a melanoma dataset. They also showed that this gene would not be selected based on CNA alone [21].

4. The parameters of the iMCMC method for integrating somatic mutation, CNA and GE datasets are set in such a way that the method can balance the influence of different data sources on the network, and on the vertex and edge weights. [32].

5. The problem of module identification in Wen *et al* is formulated as a binary Integer Linear Programming (ILP) problem, which is NP-hard. To resolve this issue, the binary variables $x_i \in \{0, 1\}$ are relaxed to continuous variables $x_i \in [0, 1]$. The problem is then solved using a simple linear programming algorithm. To choose the penalty parameter λ, the classification ability of the identified modules is defined as follows:

$$CP = max\left(C \cdot (x_1, x_2, ..., x_k)^T\right) \tag{23}$$

where the term on the right-hand side is the maximum element of the vector. The ILP is then solved for each value of λ between 0 and 1, in increments of 0.01, and the value of λ that produces the smallest value of CP is selected. The justification for this is the observation that smaller values of the elements of $\mathbf{C} \cdot (x_1, x_2, ..., x_k)^T$ indicate a greater ability to distinguish between cancer and normal samples.

6. VToD combines GE, CNA and PPI information among gene pairs to find cancer related modules. In searching for indirect relationships among gene-pairs, VToD considers the sub-network (with the genes for which pair-wise direct relationships are not defined) as fully-connected. Therefore, to find a statistically significant indirect relationship considering a set of intermediate genes is an NP-hard problem.

This problem is solved heuristically by restricting pair-wise adjacency among gene-pairs employing PPI information only, and converting that problem into finding a statistically significant simple path between gene-pairs. However, a threshold for the length of a simple path is a crucial parameter for handling time-complexity in this regard.

References

1. Zhang S, Liu CC, Li W, Shen H, Laird PW, et al. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 40: 9379–9391.

2. Davies H, Bignell GR, Cox C, Stephens P, Edkins ea S (2002) Mutations of the BRAF gene in human cancer. Nature 417: 949–954.

3. Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz ea D (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell 116: 855–867.

4. Santarosa M, Ashworth A (2004) Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. Biochim Biophys Acta 1654: 105–122.

5. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. Nat Med 10: 789–799.

6. Hanahan D, Weinberg R (2011) Hallmarks of cancer: The next generation. Cell 144: 646 - 674.

7. Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291–2297.

8. Qiu YQ, Zhang S, Zhang XS, Chen L (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data. BMC Bioinformatics 11: 26.

9. de Lichtenberg U, Jensen LJ, Brunak S, Bork P (2005) Dynamic complex formation during the yeast cell cycle. Science 307: 724–727.

10. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176.

11. Subramanian A, Tamayo P, Mootha ea V K (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545–15550.

12. Liu X, Liu ZP, Zhao XM, Chen L (2012) Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc 19: 241–248.

13. Wen Z, Liu ZP, Yan Y, Piao G, Liu Z, et al. (2012) Identifying responsive modules by mathematical programming: an application to budding yeast cell cycle. PLoS ONE 7: e41854.

14. He D, Liu ZP, Honda M, Kaneko S, Chen L (2012) Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. J Mol Cell Biol 4: 140–152.

15. Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 9: 471–472.

16. Iorns E, Lord CJ, Turner N, Ashworth A (2007) Utilizing RNA interference to enhance cancer drug discovery. Nat Rev Drug Discov 6: 556–568.

17. Azad AKM, Lee H (2013) Voting-based cancer module identification by combining topological and data-driven properties. PLoS ONE 8: e70498.

18. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proceedings of the National Academy of Sciences 104: 20007–20012.

19. Taylor BS, Barretina J, Socci ND, DeCarolis P, Ladanyi M, et al. (2008) Functional Copy-Number Alterations in Cancer. PLoS ONE 3: e3179.

20. Hur Y, Lee H (2011) Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays. BMC Bioinformatics 12: 146.

21. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. (2010) An integrated approach to uncover drivers of cancer. Cell 143: 1005 - 1017.

22. Jornsten R, Abenius T, Kling T, Schmidt L, Johansson E, et al. (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. Mol Syst Biol 7: 486.

23. Schadt EE, Lamb J, Yang X (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37: 710–717.

24. Lee H, Kong SW, Park PJ (2008) Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. Bioinformatics 24: 889–896.

25. TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.

26. TCGA (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.

27. Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Research 30: 207-210.

28. The cancer genome atlas - data portal. URL `https://tcga-data.nci.nih.gov/tcga`.

29. Herrero J, Diaz-Uriarte R, Dopazo J (2003) Gene expression data preprocessing. Bioinformatics 19: 655–656.

30. van de Wiel MA, Picard F, van Wieringen WN, Ylstra B (2011) Preprocessing and downstream analysis of microarray DNA copy number profiles. Brief Bioinformatics 12: 10–21.

31. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, et al. (2010) Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 1–9.

32. Zhang J, Zhang S, Wang Y, Zhang XS (2013) Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. BMC Systems Biology 7: S4.

33. Wang Y, Xia Y (2008) Condition specific subnetwork identification using an optimization model .

34. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L (2013) An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc 20: 659–667.

35. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nat Rev Cancer 4: 177–183.

36. Kim YA, Wuchty S, Przytycka TM (2011) Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol 7: e1001095.

37. Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun eaT H M .

38. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, et al. (2009) Genome-wide DNA methylation profiling using Infinium assay. Epigenomics 1: 177–200.

39. Peri S, Navarro JD, Kristiansen TZ, Amanchy ea R (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32: 497–501.

40. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–539.

41. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien eaT S .

42. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, et al. (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 38: D532–539.

43. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. .

44. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.

45. Cerami E, Demir E, Schultz N, Taylor BS, Sander C (2010) Automated network analysis identifies core pathways in glioblastoma. PLoS ONE 5: e8918.

46. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: pp. 289-300.

47. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.

48. (2009) Modularity analysis of protein interaction networks. In: Zhang A, editor, Protein Interaction Networks: Computational Analysis, Cambridge: Cambridge University Press.

49. R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

50. Vandin F, Upfal E, Raphael BJ (2012) De novo discovery of mutated driver pathways in cancer. Genome Res 22: 375–385.

51. Zhao J, Zhang S, Wu LY, Zhang XS (2012) Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 28: 2940–2947.

52. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 4: 34.