Functionally Coherent Transcription Factor Target Networks Illuminate Control of Epithelial Remodelling and Oncogenic *Notch*

Ian M. Overton[1,2,3,4*], Andrew H. Sims[1], Jeremy A. Owen[2,5], Bret S. E. Heale[1,6], Matthew J. Ford[1,7], Alexander L. R. Lubbock[1,8], Erola Pairo-Castineira[1], Abdelkader Essafi[1,9]

1. MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK
2. Department of Systems Biology, Harvard University, Boston MA 02115, USA
3. Centre for Synthetic and Systems Biology (SynthSys), University of Edinburgh, Edinburgh EH9 3BF, UK
4. Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast BT9 7AE, UK
5. Department of Physics, Massachusetts Institute of Technology, Cambridge MA 02139, USA
6. Current address: Intermountain Healthcare, 3930 River Walk, West Valley City UT 84120, USA
7. Current address: Rosalind & Morris Goodman Cancer Research Centre, McGill University, Montreal H3A 0G4, Canada
8. Current address: Department of Biochemistry, Vanderbilt University, Nashville TN 37232, USA
9. Current address: School of Cellular and Molecular Medicine, University of Bristol, Bristol BS8 1TD, UK

*Corresponding author: i.overton@qub.ac.uk

Co-author emails:

Andrew H. Sims: andrew.sims@ed.ac.uk; Jeremy A. Owen: jeremyandrewowen@gmail.com; Bret S. E. Heale: bheale@gmail.com; Matthew J. Ford: matthew.ford@mail.mcgill.ca; Alexander L. R. Lubbock: alex.lubbock@vanderbilt.edu; Erola Pairo-Castineira: erola.pairo-castineira@igmm.ed.ac.uk; Abdelkader Essafi: a.essafi@bristol.ac.uk

Abstract

Background

Cell identity is governed by gene expression, regulated by Transcription Factor (TF) binding at cis-regulatory modules. Decoding the relationship TF binding patterns and the regulation of cognate target genes is nontrivial, remaining a fundamental limitation in understanding cell decision-making mechanisms. Identification of TF physical binding that is biologically 'neutral' is a current challenge. We studied cell identity in the context of Epithelial to Mesenchymal Transition (EMT), a cell programme fundamental for normal embryonic development that contributes to tumour progression and fibrosis.

Results

We developed the NetNC software for discovery of functionally coherent TF targets. NetNC was applied to analyse gene regulation by the EMT TFs Snail, Twist in early embryogenesis and also to modENCODE 'HOT' regions. Predicted neutral binding accounted for 50% to ≥80% of candidate target genes assigned from significant binding peaks. Novel gene functions and network modules were identified, including regulation of chromatin organisation and crosstalk with notch signalling. Orthologues of predicted TF targets discriminated breast cancer molecular subtypes and NetNC analysis predicted new gene functions; for example, evidencing networks that reshape Waddington's landscape during EMT-like phenotype switching. Predicted invasion roles for *SNX29*, *ATG3*, *UNK* and *IRX4* were validated using a tractable cell model.

Conclusions

We found extensive neutral TF binding across the nine datasets examined and showed that NetNC performs well in identifying functionally coherent targets. HOT regions had comparatively high functional coherence. Our results illuminate conserved molecular networks that regulate epithelial remodelling in development and disease, with potential implications for precision medicine.

Keywords

Transcription factor; gene regulation; network biology; mesoderm; EMT; epithelial remodelling; breast cancer; ChIP-seq; bioinformatics; regulatory genomics.

1 Background

Transcriptional regulatory factors (TFs) govern gene expression, which is a crucial determinant of phenotype. Therefore, mapping transcriptional regulatory networks is an attractive approach to gain understanding of the molecular mechanisms underpinning both normal biology and disease [1–3]. TF action is controlled in multiple ways; including protein-protein interactions, DNA sequence affinity, 3D chromatin conformation, post-translational modifications and the processes required for TF delivery to the nucleus [3–5]. The interplay of mechanisms influencing TF specificity across different biological contexts encompasses considerable complexity and genome-scale assignment of TFs to individual genes is challenging [1,5,6]. Indeed, much remains to be learned about the regulation of gene expression. For example, the relationship between enhancer sequences and the transcriptional activity of cognate promoters is only beginning to be understood [4,5]. Prediction of TF occupancy from DNA sequence composition alone has had only limited success, likely because protein interactions influence TF binding specificity [5,7].

TF binding sites may be determined experimentally using chromatin immunoprecipitation followed by sequencing (ChIP-seq) or microarray (ChIP-chip). These and related methods (e.g. ChIP-exo, DamID) have revealed a substantial proportion of statistically significant ‘neutral’ TF binding, that has apparently no effect on transcription from the promoters of assigned target genes [1,8–10]. Evidence suggests that neutral binding can arise from TF association with euchromatin; for example, the binding of randomly-selected TFs and genome-wide transcription levels are correlated [11–13]. Genomic regions that bind large numbers of TFs have been termed Highly Occupied Target (HOT) regions [14]. HOT regions are enriched for disease SNPs and can function

as developmental enhancers [15,16]. However, a considerable proportion of individual TF binding events at HOT regions may have little effect on gene expression and association with chromatin accessibility suggests non-canonical regulatory function such as sequestration of TFs or in 3D genome organisation [17,18] as well as possible technical artefacts [19]. A proportion of apparently neutral binding sites may also have more subtle functions; for example in combinatorial context-specific regulation and in buffering transcriptional noise [2,20]. Furthermore, enhancers may control the expression of genes that are sequence-distant but spatially close due to the 3D chromatin conformation [17,18]. Current approaches to match bound TFs to candidate target genes may miss these distant regulatory relationships. Identification of *bona fide*, functional TF target genes remains a major obstacle in understanding the regulatory networks that control cell behaviour [2,5,10,13,21].

The set of genes regulated by an individual TF typically have overlapping expression patterns and coherent biological function [22–24]. Indeed, gene regulatory networks are organised in a hierarchical, modular structure and TFs frequently act upon multiple nodes of a given module [25,26]. Therefore, we hypothesised that functional TF targets collectively share network properties that may differentiate them from neutrally bound sites. Graph theoretic analysis can reveal biologically meaningful gene modules, including cross-talk between canonical pathways [27–29] and conversely may enable elimination of neutrally bound candidate TF targets derived from statistically significant ChIP-seq or ChIP-chip peaks. For this purpose, we have developed a novel algorithm (NetNC) for functional TF target discovery and therefore to help illuminate mechanisms controlling cell phenotype, for example to inform causality in regulatory network inference [1,6]. NetNC analyses the connectivity between candidate TF target genes in the context of a functional gene network (FGN), in order to discover biologically coherent TF targets. Network approaches afford significant advantages for handling biological complexity, enable genome-scale analysis of gene function [30,31], and are not restricted to predefined gene groupings used by standard

functional annotation tools (e.g. GSEA, DAVID) [27,32,33]. FGNs seek to comprehensively represent gene function and provide a useful framework for analysis of noisy real-world data [34,35]. Clustering is frequently applied to a FGN in order to define a fixed network decomposition, as basis for identification of biological modules [36,37]. Modules with a high proportion of genes associated with a given experimental condition, such as drug treatment, may define the network response and so illuminate the underlying biology. Predefined, fixed network modules may miss important features of the condition-specific set of genes; for example, gene products with corresponding nodes in the FGN may be absent from the biological condition(s) analysed. Indeed, it is typical for any given cell type to express only a subset of the genes encoded in its genome, hence clusters derived from analysis of the whole genome network may not accurately capture the biological interactions that occur in the context of a particular cell type or environment. Additionally, context-specific interactions are a common feature of biological networks, for example the varied repertoire of biophysical interactions in different cell types or between cell states, such as in the stages of the cell cycle [38]. Therefore, modules are defined dynamically *in vivo* and there is benefit in developing analysis approaches that can discover condition-specific communities of interacting genes. The NetNC algorithm satisfies this remit, enabling identification of coherent genes and modules according to the context represented by the gene list and a FGN.

We applied NetNC and a novel FGN (DroFN) to predict functional targets for multiple datasets that measured the binding of the Snail and Twist TFs, as well for modENCODE HOT regions [14]. Snail and Twist have important roles in Epithelial to Mesenchymal Transition (EMT), a multi-staged morphogenetic programme fundamental for normal embryonic development that contributes to tumour progression and fibrosis [39–42]. Integrative analysis of the predicted functional Snail, Twist targets, Notch screens and human breast cancer transcriptomes gave insights into both developmental and cancer biology. Predicted functional TF targets from NetNC analysis with no previously described role in invasion were validated *in vitro*.

2 Results

In the subsections below we first describe a *D. melanogaster* functional gene network (DroFN) and an algorithm developed for functional transcription factor target prediction (NetNC). NetNC performed well against other approaches in discrimination of biologically related genes from synthetic neutrally bound targets. Using DroFN, NetNC and our synthetic benchmark, we estimated the proportion of neutral binding for nine Chromatin Immunoprecipitation (ChIP) microarray (ChIP-chip) or pyrosequencing (ChIP-seq) datasets, drawn from five different studies [9,14,24,43,44]. These nine datasets are referred to as 'TF_ALL'; please see Methods section 4.3 for important details about the TF_ALL datasets. NetNC predicted Snail and Twist functional targets in early embryogenesis, revealing clusters of regulation for multiple genes in key developmental processes, including chromatin remodelling, transcriptional regulation and neural development. Predicted functional targets were enriched for *Notch* signalling modifiers and captured important aspects of human breast cancer biology. The DroFN network and NetNC software are made freely available as Additional Files associated with this manuscript.

2.1 A comprehensive *D. melanogaster* functional gene network (DroFN)

We developed a functional gene network (DroFN; 11,432 nodes, 787,825 edges) to provide a systems-wide map of *D. melanogaster* signalling and metabolism (Additional File 1). Evaluation of DroFN with time-separated blind test data derived from KEGG (TEST-NET) found good performance compared with the DroID [45] and GeneMania [46] networks (Table 1, Supplementary Figure S1). The DroFN network was more highly connected than DroID, and had 2.6-fold higher average degree. GeneMania predicts shared Gene Ontology terms rather than KEGG pathway comembership, which may account for some of the performance gap found with GeneMania when compared to DroFN and DroID. However GeneMania performance on TEST-NET is similar to published values for 'Biological Process' terms [46]. The overlap between DroFN and the

Drosophila proteome interaction map (DPiM [47]) was highly significant (FET $p<10^{-308}$). DroFN and DPiM had 999 genes in common and 37.8% (2175/5747) of DroFN edges for these genes were also found in DPiM. The False Positive Rate for DroFN (0.047) was close to the prior for functional interaction estimated from KEGG (0.044); a proportion of these estimated false positives may represent *bona fide* interactions that were not annotated in KEGG. Overall, DroFN provides a useful genome-scale map of pathway comembership in *D. melanogaster*.

Table 1. Evaluation of DroFN on Time Separated Blind Test Data (TEST-NET).

Network	MCC	FPR	TPR	AUC
DroFN	0.448	0.047	0.475	0.773
DROID	0.383 (max 0.385)	0.0046	0.199	0.598
GeneMania	0.133 (max 0.243)	0.121	0.274	0.582

Table 1. Column headings: Matthews correlation coefficient (MCC), false positive rate (FPR), true positive rate (TPR), area under the Receiver Operator Characteristic curve (AUC). DroFN performed best on the data examined and had FPR close to the functional interaction prior estimated from the training data (0.044). Values of AUC for DroFN were significantly better than DROID ($p=2.13\text{x}10^{-11}$) or GeneMania ($p=3.19\text{x}10^{-22}$).

2.2 A novel algorithm for discovery of functional transcription factor binding (NetNC)

Large numbers of statistically significant TF binding sites appear to be neutral (non-functional) [8,10,24]. We developed the NetNC algorithm for genome-scale prediction of functional TF target genes (Figure 1). In broad terms, NetNC seeks to discover the biological functions common to a list of genes, therefore defining groups of genes with common function and revealing biologically defining characteristics. This general paradigm has been applied widely, for example in network-based approaches [27,28,48,49] and in enrichment analysis [32,33,50].

NetNC builds upon observations that TFs coordinately regulate multiple functionally related targets [22–24] and has been calibrated for discovery of biologically coherent genes in noisy data. The first stage in NetNC calculates hypergeometric mutual clustering (HMC) p-values [51] for each pair of candidate TF targets (H_1) that are connected in the functional gene network (FGN). Empirical estimation of positive False Discovery Rate (pFDR) [52] across H_1 is enabled by deriving HMC p-values from resampled genes (H_0). Resampling to generate H_0 controls for the number of candidate TF target genes analysed and the FGN structure. Iterative minimum cut is then computed on the pFDR thresholded network with a graph density stopping criterion [53]. Connected components of the resulting graph consisting of less that three nodes are discarded. The approach described above is edge-centric and is termed 'Functional Target Identification' (FTI), seeking to distinguish all biologically coherent gene pairs from functionally unrelated targets (e.g. arising from neutral TF binding). Additionally, NetNC has a node-centric 'Functional Binding Target' (FBT) mode that employs regularised Gaussian mixture modelling for unsupervised clustering with automatic cardinality selection [54]. NetNC-FBT analyses degree-normalised Node Functional Coherence Scores (NFCS); examples of NFCS profiles and the fitted mixture models are visualised in Supplementary Figure S2. NetNC-FBT is parameter-free and so did not require calibration on training data.

The gold-standard data for NetNC development and validation took KEGG pathways to represent biologically coherent relationships, combined with 'Synthetic Neutral Target Genes' (SNTGs) derived by resampling from the DroFN network. A total of 17,600 datasets (Additional File 2, Additional File 3) were developed to contain between 5% and 80% SNTGs; therefore, the gold-standard data covered a wide range of possible values for the proportion of neutrally bound candidate TF target genes. NetNC was robust to variation in the input dataset size and %SNTGs, outperforming HC-PIN [37] and MCL [36] on blind test data (Figure 2, Supplementary Table S1). Previous work that evaluated nine clustering algorithms, including MCL, found that HC-PIN had

strong performance in functional module identification and was robust against false positives [37]; therefore we selected HC-PIN for extensive comparison against NetNC. In general, NetNC was more stringent, with lower False Positive Rate (FPR) and higher Matthews Correlation Coefficient (MCC) than HC-PIN. MCC provides a balanced measure of predictive power across the positive (KEGG pathway) and negative (SNTG) classes of genes in the gold standard; therefore MCC is an attractive approach for assessment of overall performance. NetNC-FBT typically had lowest FPR and performed well on larger datasets. We saw a spread of performance values across resamples with identical number of pathways and %SNTG (Figure 2), which arose from expected differences between resamples. For example, differences in the network density of the resampled SNTG genes may impact upon the power of NetNC to discriminate between SNTGs and KEGG pathway nodes. NetNC's performance advantages were most prominent on blind test data with ≥50% SNTGs (Figure 2) and all nine of the TF_ALL datasets were predicted to contain ≥50% neutrally bound targets (Figure 3, see subsection 2.3, below). Therefore, given the performance advantage on blind test data with ≥50% STNGs (Figure 2), NetNC appears as the method of choice for identification of functional TF targets from genome-scale binding data.

2.3 Estimating neutral binding for EMT transcription factors and Highly Occupied Target (HOT) regions

We predicted functional target genes for the Snail and Twist TFs for developmental stages around gastrulation in *D. melanogaster*. Fly embryos perform rapid nuclear divisions and transcription, leading the formation of the syncytial blastoderm at about 2 hours. Nuclear divisions slow during cellularisation of the blastoderm after 2 hours and gastrulation occurs around 3 hours [55–57]. Using NetNC and DroFN, we analysed Chromatin ImmunoPrecipitation (ChIP) microarray (ChIP-chip) or sequencing (ChIP-seq) data for overlapping time periods in early embryogenesis produced by four different laboratories and also the modENCODE Highly Occupied Target (HOT) regions

[9,14,24,43,44]. Nine datasets in total were studied (TF_ALL, Table 2), enabling investigation of multiple factors that are commonly applied in discovery of candidate TF targets - including: peak intensity threshold; multiple developmental time periods, multiple antibodies, different analytical platforms, and using transcribed genes for peak assignment. Further details of the TF_ALL datasets are given in Methods subsection 4.3. The proportion of neutrally bound candidate target genes was estimated using a novel approach (NetNC-lcFDR) that calculated local FDR (lcFDR) from NetNC pFDR values, with calibration against the known SNTG fraction in gold standard data. Local FDR estimates the false discovery rate at a specific score value (or range of values) in contrast to global FDR which is calculated using all of the values above a score threshold. We note that global pFDR was unsuitable for estimating the total fraction of neutral binding. For example, every TF_ALL dataset had pFDR=1 at the NetNC score threshold that included all candidate target genes; hence, a naïve approach based on global pFDR would always give a global neutral binding estimate of 100%. Furthermore, lcFDR may capture differences in score profiles that are missed by global pFDR, illustrated in Supplementary Figure S3.

NetNC-lcFDR estimates of neutral binding across TF_ALL ranged from 50% to ≥80% (Figure 3A, Table 2). Reassuringly, the dataset with the most stringent peak calling (twi_1-3h_hiConf [9]) had the highest (NetNC-lcFDR) or second highest (NetNC-FTI) predicted functional binding proportion. Target genes for regions bound during two consecutive developmental time periods (twi_2-6h_intersect [44]) also ranked highly, followed by HOT regions (Figure 3A, Table 2). Indeed, twi_2-6h_intersect had a significantly greater percentage of predicted functional targets (binomial $p<4.0\text{x}10^{-15}$) with stronger pFDR and lcFDR profiles than either the twi_2-4h_intersect or twi_4-6h_intersect datasets from the same study, but where binding was during a single time period [44] (Figure 3). Therefore, predicted functional binding was enriched for regions occupied at >1 time period or by multiple TFs - including HOT regions, which had high functional coherence relative to the other datasets examined. Interestingly, a very similar proportion

of functional targets was predicted by NetNC-lcFDR for binding sites derived from either the union or intersection of two Twist antibodies (NetNC-lcFDR=25-30%) from the same study [24], although the NetNC-FTI value was higher for input data representing the intersection of antibodies (30.5% (116/334) vs 23% (424/1848)). Substantial numbers of candidate target genes in all nine TF_ALL datasets passed a global FDR (pFDR) or lcFDR threshold value of 0.05 (Figure 3B, 3D). Even datasets with high predicted total neutral binding included candidate targets that met stringent NetNC FDR thresholds. For example, despite having a relatively low proportion of predicted total functional binding (Figure 3A) the datasets sna_2-3h_union, twi_2-3h_union respectively had the highest and second-highest proportion of genes passing lcFDR<0.05 (Figure 3B); these datasets were also highly ranked at pFDR<0.05 (Figure 3D).

Table 2. Predicted Functional binding for Snail, Twist and HOT candidate target genes.

Dataset			Predicted functional targets	
Name	Developmental time period(s)	Candidate target genes*	NetNC-FTI	NetNC-lcFDR
twi_1-3h_hiConf	1-3h	664	202 (30%)	45-50%
twi_2-6h_intersect	2-4h and 4-6h	615	241 (39%)	40-45%
twi_2-4h_intersect	2-4h only (not 4-6h)	801	182 (23%)	25-30%
twi_4-6h_intersect	4-6h only (not 2-4h)	818	126 (15%)	25-30%
HOT	0-12h^{+}	677	174 (26%)	35-40%
twi_2-3h_union	2-3h	1848	424 (23%)	25-30%
sna_2-3h_union	2-3h	1158	226 (20%)	25%
twi_2-4h_Toll^{10b}	2-4h	1238	279 (23%)	30-35%
sna_2-4h_Toll^{10b}	2-4h	1488	211 (14%)	≤20%

*number of candidate targets that mapped to DroFN nodes. +multiple time periods and 41 different TFs [14].

Table 2. Results for NetNC are given based on 'Functional Target Identification' (NetNC-FTI) and mean local FDR (NetNC-lcFDR) calibrated against datasets with a known proportion of resampled Synthetic Neutral Target Genes (SNTG) described in Methods section 4.2.3. The above datasets correspond to the following developmental stages: 2-4h stages 4-9 (except '2-4h_intersect datasets which were stages 5-7 [44]); 2-3h stages 4-6; 1-3h stages 2-6; 4-6h stages 8-9 [44]; 0-12h stages 1-15; gastrulation occurs at stage 6 [57].

ChIP peak intensity putatively correlates with functional binding, although some weak binding sites have been shown to be functional [10,58]. We found a significant correlation between genes' NetNC NFCS values and ChIP peak enrichment scores in 6/8 datasets (q<0.05, HOT regions not analysed). The two datasets with no significant correlation (twi_1-3h_hiConf, twi_2-6h_intersect) were derived from protocols that enrich for functional targets and had the lowest predicted neutral binding proportion (Figure 3A). Indeed, the median peak score for twi_2-6h_intersect was significantly higher than data from the same study that was restricted to a single time period (twi_2-4h_intersect, $q<5.0\text{x}10^{-56}$; twi_4-6h_intersect, $q<4.8\text{x}10^{-58}$). Therefore the relationship of peak intensity with functional binding in twi_1-3h_hiConf, twi_2-6h_intersect appears to have been eliminated by the application of protocols that enriched for functional targets. Functional TF targets identified by NetNC were also enriched for human orthologues, defined by InParanoid [59]. For example, human orthologues were assigned to 72% (453/628) of the NetNC-FBT predicted functional target genes for twi_2-3h_union, significantly higher than the value (50%, 616/1220) for the full dataset ($p<3\text{x}10^{-28}$ binomial test). Genome-wide expectation for human-fly orthology was 46%, calculated with reference to the fly genome, which was significantly lower than the value of 72% for the twi_2-3h_union predicted functional targets ($p<5\text{x}10^{-40}$). The enrichment for evolutionary conservation of NetNC results aligns with the fundamental developmental processes captured by the datasets analysed (i.e. gastrulation, mesoderm development) and is consistent with the predicted functional target genes playing roles in these processes.

NetNC-lcFDR estimates of neutral binding agreed well with the Functional Target Identification results (NetNC-FTI, Table 2). Indeed, neutral binding estimates from these two methods had median difference of only 5.5% and were significantly correlated across TF_ALL, despite considerable methodological differences (r=0.85, p=0.008, Supplementary Figure S4). This concordance supports the results from both NetNC-FTI and NetNC-lcFDR.

2.4 Genome-scale functional transcription factor target networks

NetNC results offer a global representation of the mechanisms by which Snail and Twist exert tissue-specific regulation in early *D. melanogaster* embryogenesis (Figure 4, Supplementary Figure S5, Additional File 4). NetNC-FTI results for the nine TF_ALL datasets overlapped and clusters were manually annotated into biologically similar groups, with reference to Gene Ontology enrichment and FlyBase annotations [33,60–62]. Eleven biological groupings were identified in at least 4/9 TF_ALL datasets, including developmental regulation (9/9), chromatin organisation (6/9), ion transport (6/9), mushroom body development (6/9), phosphatases (6/9), splicing (5/9) and regulation of translation (5/9) (Supplementary Table S2). Very few clusters were composed entirely from genes identified only in a single dataset, examples included: snoRNAs/nucleolar proteins (twi_2-3h_union), transferases (HOT), defense response/immune response (twi_2-4h_Toll^{10b}) and chitin metabolism (twi_2-4h_intersect) (Figure 4, Supplementary Figure S5). We investigated the robustness of NetNC-FTI to subsampled input using TF_ALL (Supplementary Tables S3, S4). The median overlap of genes returned in analysis of an individual complete TF binding dataset for 95%, 80% and 50% subsamples, averaged across TF_ALL, was respectively 89%, 81%, 75% (respective median 95% CI 72%-97.2%, 66%-92%, 58%-97%). The median overlap for network edges with node subsampling at 95%, 80%, 50%, across TF_ALL, was 91%, 84% and 77% (median 95% CI 83-96%, 74-94%, 37-92%). Overall, subsampling had a moderate effect on NetNC predictions and greater sensitivity was observed at lower subsampling rates, as expected. Some subsamples taken as input to NetNC had low overlap with the NetNC-FTI reference output (reference_net) for any given complete input dataset. Indeed, the reference_net represented between 14% to 39% of the total input gene list across the nine TF_ALL datasets. Subsamples that excluded a high proportion of the nodes in reference_net would be expected to result in weaker hypergeometric mutual clustering values for the remaining nodes found in reference_net because these nodes would be expected to have relatively few common neighbours. Therefore, subsampling of the input gene list is expected to

produce NetNC results that have reduced overlap with reference_net; this effect is also a source of variation in overlap across subsamples, reflected in the 95% CI values. Also, the probability of sampling nodes in reference_net is lower when a smaller fraction of the complete input TF_ALL gene list is covered by reference_net, leading to a greater subsampling-associated loss of nodes and edges. Consistent with this interpretation, TF_ALL datasets with the highest NetNC-FTI functional binding proportion (Table 2) (twi_1-3h_hiConf, twi_2-6h_intersect, HOT) were less sensitive to subsampling than datasets with relatively low predicted functional binding such as sna_2-4h_Toll[10b] and twi_4-6h_intersect (Supplementary Tables S3, S4).

The developmental regulation cluster (DRC) encompassed key conserved morphogenetic pathways, for example: notch, wnt and fibroblast growth factor (FGF). *Notch* signalling modifiers from public data [63] overlapped significantly with NetNC-FTI results for each TF_ALL dataset (q <0.05), including the DRC, chromatin organisation and mediator complex clusters (Figure 4, Supplementary Figure S5). *Notch* was identified as an important control node across TF_ALL where it had highest betweenness centrality in the DRC for three datasets and ranked (by betweenness) among the top ten DRC genes for 8/9 datasets. The activation of *Notch* can result in diverse, context-specific transcriptional outputs and the mechanisms regulating this pleiotropy are not well understood [63–66]. NetNC predicted functional Snail and Twist binding to many regulatory genes in the *Notch* network neighbourhood, therefore providing evidence for novel factors controlling the transcriptional consequences of *Notch* activation in cell fate decisions controlled by these TFs. This is consistent with previous demonstration of signalling crosstalk for *Notch* with *twist* and *snail* in multiple systems; for example in adult myogenic progenitors [67] and hypoxia-induced EMT [68]. *Wingless* also frequently had high betweenness, ranking within the top ten DRC genes in six datasets and was highest ranked in two instances. Thirteen genes were present in the DRC for at least seven of the TF_ALL datasets (DRC-13, Supplementary Table S5), and these genes had established functions in the development of mesodermal derivatives such as

muscle, the nervous system and heart [65,67,69–73]. Public *in situ* hybridisation (ISH) data for the DRC-13 genes indicated their earliest expression in (presumptive) mesoderm at: stages 4-6 (*wg*, *en*, *twi*, *N*, *htl*, *how*), stages 7-8 (*rib*, *pyd*, *mbc*, *abd-A*) and stages 9-10 (*pnt*) [74–77]. The remaining two DRC-13 genes had no evidence for mesodermal expression (*fkh*) or no data available (*jar*). However, other studies had shown that *fkh* is essential for caudal visceral mesoderm development [78] and had demonstrated *jar* expression in the midgut mesoderm [79]. The above data are consistent with direct regulation of DRC-13 by Twist and Snail in (presumptive) mesoderm, as predicted by NetNC-FTI.

Chromatin organisation clusters included polycomb-group (PcG) and trithorax-group (TrxG) genes; the most frequently identified were the Polycomb Repressive Complex 1 (PRC1) genes *ph-d*, *psc* [80] and *su(var)3-9*, a histone methyltransferase that functions in gene silencing [81,82] (Supplementary Table S6). Other NetNC-FTI coherent genes with function related to PcG/TrxG included: the PRC1 subunit *ph-p* [80]; c*orto* which physically interacts with PcG and TrxG proteins [83,84]; the TrxG-related gene *lolal* that is required for silencing at polycomb response elements [85,86]; *taranis* which has genetic interactions with TrxG and PcG [87–89]; TrxG genes *trithorax*, *moira* [90–93]. The gene silencing factor s*u(var)205* was also returned by NetNC-FTI in four TF_ALL datasets *[94,95]*. Therefore, NetNC found direct regulation by Snail and Twist of a) PRC1 core components and other gene silencing factors, b) TrxG genes, c) modifiers of PcG, TrxG activity.

Brain development clusters were found for six TF_ALL datasets, as well as members of the proneural *achaete-scute* complex and *Notch* signalling components [96]. Snail regulation of neural clusters is consistent with its well characterised roles in repression of ectodermal (neural) genes in the prospective mesoderm [97–99]. Additionally, Snail is important for neurogenesis in fly development and also in mammals [100,101]. Therefore, binding to these neural functional modules could reflect potentiation of transcription to enable rapid activation in combination with other

transcription factors as and when required within specific neural developmental trajectories [44,102]. The mushroom body is a prominent structure in the fly brain that is important for olfactory learning and memory [103]. Twist is typically a transcriptional activator [99] although appears to contribute to Snail's repressive activity [104] and Twist-related protein 1 was shown to directly repress Cadherin-1 in breast cancers [105]. Our NetNC results predict novel Twist functions, for example in regulation of mushroom body neuroblast proliferation factors such as *retinal homeobox*, *slender lobes*, and *taranis* [106–108].

2.5 Breast cancer subtype is characterised by differential expression of orthologous Snail and Twist functional targets

Genes that participate in EMT have roles in metastasis and drug resistance across multiple cancers [39,109,110]. Indeed, the NetNC-FTI Snail and Twist targets included known drivers of tumour biology and also predicted novel cancer driver genes (Figure 4, Supplementary Figure S5, Supplementary Tables S2, S5, S6). Breast cancer intrinsic molecular subtypes with distinct clinical trajectories have been extensively validated and complement clinico-pathological parameters [111,112]. These subtypes are known as luminal-A, luminal-B, HER2-overexpressing, normal-like and basal-like [111]. All of the NetNC-FTI networks for the nine TF_ALL datasets overlapped with known cancer pathways, including significant enrichment for *Notch* modifiers (q<0.05). We hypothesised that orthologous genes from NetNC clusters for Snail and Twist would stratify breast cancers by intrinsic molecular subtype. Indeed, aberrant activation of *Notch* orthologues in breast cancers had been demonstrated and was linked with EMT-like signalling, particularly for the basal-like and claudin-low subtypes [113–117].

2.5.1 Unsupervised clustering with predicted functional targets recovers breast cancer intrinsic subtypes

We identified 57 human orthologues (ORTHO-57) that were NetNC-FTI functional targets in ≥4 TF_ALL datasets and were also represented within integrated gene expression microarray data for 2999 breast tumours (BrC_2999) [118]. Unsupervised clustering with ORTHO-57 stratified BrC_2999 by intrinsic molecular subtype (Figure 5). Clustering with NetNC results for individual Twist and Snail datasets also recovered the intrinsic breast cancer subtypes (Supplementary Figure S6). Features within the heatmap were marked according to the dendrogram structure and gene expression values (Figure 5). Basal-like tumours were characterised by *EN1* and *NOTCH1*, aligning with previous work (feature_Bas; Figure 5) [113,114,119]. Interestingly, elevated *ETV6* expression was also largely restricted to the basal-like subtype. Others had reported *ETV6* copy number amplifications in 21% of basal-like tumours and identified recurrent gene fusions with *ETV6* in several cancers [120–123]. The Luminal A subtype (feature_LumA), shared gene expression characteristics with luminal B (feature_LumB$_2$, ERBB3, MYO6) and normal-like (DOCK1, ERBB3, MYO6) tumours. High *BMPR1B* expression was a clear defining feature of the luminal A subtype, in agreement with previous results demonstrating oncogenic BMP signalling in luminal epithelia [124]. Others had previously shown that the *BMP2* ligand may be pleiotropic in breast cancers and development, promoting EMT characteristics in some contexts [125–127]. Tumours with high relative *BMP2* expression were typically basal-like while luminal cancers had low *BMP2*; therefore, our data align with *BMP2* upregulation as a feature of the EMT programme in basal-like cancers. The luminal B subtype had been established to have worse prognosis than luminal A, but more favourable prognosis than *ESR1* negative cancers [111,128]. Several genes were highly expressed in both feature_LumB$_1$ and in *ESR1* negative subtypes (feature_ERneg), including *ECT2*, *SNRPD1*, *SRSF2* and *CBX3*; our data suggest that these genes might contribute to worse survival outcomes for luminal B relative to luminal A cancers. Indeed, the luminal A as well as normal-like

tumour subtypes had low expression of these genes and *CBX3, ECT2* had previously been correlated with poor prognosis [129,130]. Furthermore, *SNRPD1* is a component of core splicesomal small nuclear ribonucleoproteins (snRNPs) and *SRSF2* is a splicing factor [131]; RNA splicing was shown to be a survival factor in siRNA screening across multiple basal-like cancer cell lines and was suggested to have potential therapeutic value [132]. Feature_LoExp broadly represents genes with low detection rates (indicated by the %P column in Figure 5) and the tumours populating feature_LoExp are a mixture of subtypes, but largely from a single study [133]. Notably, key EMT genes (*SNAI2, TWIST1, QKI*) had highest relative expression in normal-like tumours (feature_NL, Figure 5). Indeed, *SNAI2* and *TWIST1* were both assigned to the normal-like centroid. Feature_NL also included homeobox transcription factors (*HOXA9*, *MEIS2*) and a secreted cell migration guidance gene (*SLIT2*) [134–136]. Some genes had high expression in both normal-like (feature_NL) and basal-like cancers, including: the *QKI* RNA-binding protein that regulates circRNA formation in EMT [137] and the *FZD1* wnt/β-catenin receptor. Indeed, genes in feature_Bas and feature_NL clustered together in the gene dendogram, reflecting greater gene expression similarity to each other than to genes within features for the other breast cancer subtypes (Figure 5). Therefore, these data revealed concordance in gene expression between the normal-like and basal-like subtypes, including known EMT-related genes.

2.5.2 Integrating NetNC functional target networks and breast cancer transcriptome profiling

We visualised basal-like and normal-like gene annotations for orthologues in the NetNC-FTI networks, offering a new perspective on the molecular circuits controlling these different subtypes (Figure 4, Supplementary Figure S5). We focussed on basal-like and normal-like cancers because they accounted for the large majority of genes in the datasets examined and were prominent in results from the centroid and heatmap analysis (Figure 5, Supplementary Figure S6). Additionally,

EMT had been shown to be important for basal-like breast cancer biology [138,139] and key EMT genes were annotated to the normal-like subtype in our analysis. NetNC-FTI clusters that contained splicing factors and components of the ribosome were associated with the normal-like subtype in results for three datasets (twi_2-4h_intersect, twi-2-6h_intersect, twi_2-3h_union); twi_2-3h_union also had communities for the proteosome and proteosome regulatory subunits where a high proportion of genes were annotated to the normal-like subtype. Orthologues in the sna_2-4h_Toll^{10b} 'RNA degradation and transcriptional regulation' cluster were annotated to the basal-like subtype and never to the normal-like subtype; this cluster included *HECA*, which had been reported to function as both a tumour suppressor [140,141] and an oncogene [142]. *HECA* was also identified in NetNC-FTI analysis of twi_2-4h_intersect and twi_4-6h_intersect; these two datasets had Twist binding at different, non-contiguous sites that were both assigned to *hdc*, the *D. melanogaster* orthologue of *HECA*. *Hdc* was a *Notch* signalling modifier with roles in cell survival [143,144], differentiation of imaginal primodia [145], RNA interference [146], *Notch* signalling [63] and tracheal branching morphogenesis - upregulated by the *snail* gene family member *escargot* [147]. *HECA* was upregulated in basal-like relative to normal-like tumours ($p<3.3x10^{-23}$). Taken together, these data support participation of *HECA* in an EMT-like gene expression programme in basal-like breast cancers. An 'ion antiporter and GPCR' cluster for the sna_2-4h_Toll^{10b} dataset (Figure 4) included the Na^+/H^+ antiporter *SLC9A6* that also belonged to the twi_2-4h_Toll^{10b} 'transmembrane transport' cluster (Supplementary Figure S5). Alterations in pH by Na^+/H^+ exchangers, particularly *SLC9A1*, had been shown to drive basal-like breast cancer progression and chemoresistance [148–150]. *SLC9A6* was 1.6-fold upregulated in basal-like relative to normal-like tumours ($p<8.4x10^{-71}$) and may drive pH dysregulation as part of an EMT-like programme in basal-like breast cancers. The sna_2-4h_Toll^{10b} NetNC-FTI clusters were depleted in orthologues annotated to the NL subtype, compared with results for Twist. For example, sna_2-4h_Toll^{10b} had 4/18 clusters with two or more normal-like orthologous genes, significantly fewer than twi_2-3h_union (12/27, Figure 4 panel A;

binomial $p<0.01$) and twi_ 2-4h_Toll^{10b} (8/17, Supplementary Figure S3; $p<0.035$). A further cluster that was specific to basal-like cancers in the twi_2-3h_union dataset was annotated to 'mitochondrial translation', an emerging area of interest for cancer therapy [151,152]. Orthologues annotated to the basal-like subtype were frequently located in NetNC-FTI chromatin organisation clusters. For example, the twi_2-3h_union 'chromatin organisation and transcriptional regulation' cluster had six genes annotated to the basal-like subtype, including three Notch signalling modifiers (*ash1*, *tara*, *Bap111)* that were respectively orthologous to *ASH1L, SERTAD2* and *SMARCE1*. The *ASH1L* histone methyltransferase was a candidate poor prognosis factor with copy number amplifications in basal-like tumours [153]; *SERTAD2* was a known bromodomain interacting oncogene and E2F1 activator [154,155]; *SMARCE1*, a core subunit of the SWI/SNF chromatin remodelling complex, had been shown to regulate *ESR1* function and to potentiate breast cancer metastasis [156,157]. Therefore our integrative analysis predicted specific chromatin organisation factors downstream of Snail and Twist, identifying orthologous genes that may control *Notch* output and basal-like breast cancer progression.

2.6 Novel Twist and Snail functional targets influence invasion in a breast cancer model of EMT

Our analysis underlined the functional relevance of novel regulators of EMT and cell invasion, including *SNX29* (also known as *RUNDC2A*), *ATG3*, *IRX4* and *UNK*. Therefore, we investigated the functional and instructive role of these genes in an established cell model of invasion by overexpressing *SNAI1* in MCF7 cells [158]. MCF7 cells are weakly invasive [159], thus the *SNAI1*-inducible MCF7 cell line was well suited to study alteration in expression of the selected genes in terms of their influence on invasion in conjunction with *SNAI1* induction, knockdown or independently. This was achieved by the co-transfection of cDNAs of these genes alongside a doxycycline-inducible vector (pGoldiLox, [160]) that expressed either *SNAI1* cDNA or validated

shRNAs against *SNAI1 [161]*. To test for the instructive role of these genes, we ectopically expressed the selected NetNC functional targets in a transwell invasion assay that contained MCF7 with or without *SNAI1* cDNA, *SNAI1* shRNAs, mCherry control or scrambled control shRNA (Figure 6).

Over-expression of *IRX4* significantly increased invasion relative to controls in all conditions examined and *IRX4* had high relative expression in a subset of basal-like breast cancers (Figures 5, 6). *IRX4* is a homeobox transcription factor involved in cardiogenesis, marking a ventricular-specific progenitor cell [162] and is also associated with prostate cancer risk [163]. *SNX29* belongs to the sorting nexin protein family that function in endosomal sorting and signalling [164,165]. *SNX29* is poorly characterised and ectopic expression significantly reduced invasion in a *SNAI1*-dependent manner (Figure 6). Since we obtained these results, *SNX29* downregulation has been associated with metastasis and chemoresistance in ovarian carcinoma [166], consistent with *SNX29* inhibition of invasion driven by Snail. *ATG3* is an E2-like enzyme required for autophagy and mitochondrial homeostasis [167,168], we found that *ATG3* overexpression significantly increased invasion. Consistent with our results, knockdown of *ATG3* has been reported to reduce invasion in hepatocellular carcinoma [169]. *UNK* is a RING finger protein homologous to the fly unkempt protein which binds mRNA, functions in ubiquitination and was upregulated in cells undergoing gastrulation [170]. Others have reported that *UNK* mRNA binding controls neuronal morphology and can induce spindle-like cell shape in fibroblasts [171,172]. We found that *UNK* significantly increased MCF7 cell invasion in a manner that was additive with and independent of Snail, supporting a potential role in breast cancer progression. Indeed, *UNK* was overexpressed in cancers relative to controls in the ArrayExpress GeneAtlas [173].

3 Discussion

Our novel Network Neighbourhood Clustering (NetNC) algorithm and *D. melanogaster* functional gene network (DroFN) were applied to predict functional transcription factor binding targets from statistically significant ChIP-seq and ChIP-chip peak assignments during early fly development (TF_ALL). Seven of the nine TF_ALL datasets included developmental time periods encompassing stage four (syncytial blastoderm, 80-130 minutes), cellularisation of the blastoderm (stage five, 130-170 minutes) and initiation of gastrulation (stage 6, 170-180 minutes) [9,24,43,44,57]. The datasets twi_2-4h_intersect, sna_2-4h_intersect, twi_2-4h_Toll^{10b} and sna_2-4h_Toll^{10b} additionally included initial germ band elongation (stage seven, 180-190 minutes) [43,44,57]; twi_2-4h_Toll^{10b} and sna_2-4h_Toll^{10b} may have also included stages eight (190-220 minutes) and nine (220-260 minutes) [43,57]. Twi_2-4h_intersect and sna_2-4h_intersect were tightly staged between stages 5-7 [44]. Additional to stages four, five and six, twi_1-3h_hiConf may have included the latter part of stage two (preblastoderm, 25-65 minutes) and stage three (pole bud formation, 65-80 minutes) [57]. The twi_4-6h_intersect dataset was restricted to stages eight to nine which included germ band elongation and segmentation of neuroblasts [44,57]. The above differences in the biological material analysed could be an important factor underlying variation between datasets, although there was considerable overlap in the functional networks predicted for TF_ALL (Figure 4, Supplementary Table S2, Supplementary Figure S5).

NetNC analysis substantiated Snail and Twist function in regulating components of multiple core cell processes that govern the global composition of the transcriptome and proteome (Figure 4, Supplementary Figure S5). These processes included transcription, chromatin organisation, RNA splicing, translation and protein turnover (ubiquitination). We identified a ‘Developmental Regulation Cluster’ (DRC) which was the major transcriptional control module identified in all nine TF_ALL datasets. *Notch* and also *wingless* had consistently high betweenness centrality in the DRC, which is a measure of a node’s influence within a network [174]. In this context, high

betweenness centrality may highlight genes with key roles in determining the global network state, and so are important for controlling phenotype. Therefore *Notch*, *wingless* were predicted to be key control points regulated by Snail, Twist in the mesoderm specification network. *Notch* signalling putatively integrates with multiple canonical pathways [63] including interaction with the Wnt gene family which have many conserved roles across metazoan development, such as in axis specification and mesoderm patterning (reviewed in [175] and [176]). Our results are complementary to qualitative dynamic modelling where key control nodes may not necessarily have high betweenness [177]. Orthologues of both *Notch* and *wingless* were previously shown to be aberrantly regulated in breast cancers, (for example [113,178], and we found that unsupervised clustering using predicted Snail and Twist functional targets stratified five intrinsic breast cancer subtypes [111] (Figure 5). While more recent studies have classified greater numbers of breast cancer subtypes, for example identifying ten groups [179], the five subtypes employed in our analysis had been widely used, extensively validated, exhibited clear differences in prognosis, overlapped with subgroups defined using standard clinical markers (*ESR1*, *HER2*), and so were associated with distinct treatment pathways [111,112]. Analysis of the twi_2-3h_union dataset revealed a basal-like specific cluster for ‘mitochondrial translation’ (MT) (Figure 4). Inhibition of MT is a therapeutic strategy for AML and mitochondrial metabolism is currently being explored in the context of cancer therapy [151,152]. Our results highlight MT as a potentially attractive target in basal-like breast cancers, aligning with previous work linking MT upregulation with deletion of *RB1* and *p53*, which occurs in approximately 20% of triple negative breast cancers [180,181]. NetNC analysis provided functional context for many *Notch* modifiers and proposed mechanisms of signalling crosstalk by predicting regulation of modifiers by Twist, Snail (Figure 4, Supplementary Figure S5, Additional File 4). Clusters where multiple modifiers were identified may represent cell meso-scale units that are particularly important for *Notch* signalling in the context of mesoderm development and EMT (Additional File 4). For example, the mediator complex and transcription

initiation subcluster for twist_union (Figure 4) had 13 nodes, of which 5 were *Notch* modifiers including orthologues of *MED7*, *MED8*, *MED31*. Our results show regulation of *Notch* signalling by Snail and Twist targeting of *Notch* transcriptional regulators, trafficking proteins, post-translational modifiers (e.g. ubiquitinylation) and receptor recycling (non-canonical, ligand-independent signalling) as well as regulation of pathways that may attenuate or modify the *Notch* signal, consistent with previous studies [63,64]. *Taranis*, a *Notch* modifier in the chromatin organisation cluster, was orthologous to the *SERTAD2* bromodomain interacting oncogene [154] which had elevated expression in a basal-like breast cancer cluster that contained *NOTCH1* (Figure 4, Figure 5). Our integrative analysis suggests that *SERTAD2* could control the phenotypic consequences of *NOTCH1* activation in basal-like breast cancers through a chromatin remodelling mechanism. Notch signalling modulation has been applied in a clinical setting, for example in treatment of Alzheimer's disease, and is a promising area for cancer therapy [64,182–184]. Orthologues of *Notch* modifiers identified in our analysis provide a pool of candidates that could potentially inform development of companion diagnostics or combination therapies for agents targeting the notch pathway in basal-like breast cancers. In addition to *Notch* signalling, *taranis* also functions to stabilise the expression of *engrailed* in regenerating tissue [87]. The *engrailed* orthologue *EN1* is a survival factor in basal-like breast cancers [119]; *SERTAD2* and *EN1* were both located within the basal-like breast cancer cluster 'Bas' (Figure 5). Indeed, *EN1* was the clearest single basal-like cancer biomarker in the data examined. Therefore, we speculate that *SERTAD2* may cooperate with *EN1* in basal-like breast cancers, reflecting conservation of function between fly and human; indeed, our results evidence coordinated expression of these two genes as part of a gene expression programme controlled by EMT TFs. Regulation of *EN1*, *SERTAD2* within an EMT programme could harmonise previous reports of key roles for both neural-specific and EMT TFs in basal-like breast cancers [119,138]. The *taranis* chromatin organisation cluster also contained *Notch* modifiers *ash1*, *Bap111*, which were respectively orthologous to the *ASH1L*, *SMARCE1* breast

cancer poor prognosis factors [153,157]. The notch pathway had been shown to drive EMT-like characteristics as well as to mediate hypoxia-induced invasion in multiple cell lines [68]. Previous work had also shown that *SMARCE1*, a SWI/SNF complex member, interacted with Hypoxia Inducible Factor 1A (*HIF1A*) signalling and had significant effects on cell viability upon knockdown/ectopic expression alongside disruption of notch family signalling by gamma-secretase inhibition [157]. *SMARCE1* was recently shown to be important in early-stage cancer invasion [185]. Aligning with these studies, our results evidence conserved function for *SMARCE1* in (partial) EMT signalling in both mesoderm development and breast cancer progression, possibly in regulation of SWI/SNF targeting. SWI/SNF had been reported to regulate chromatin switching in oral cancer EMT [186]. NetNC results showing predicted regulation of chromatin organisation genes by Snail, Twist also included core polycomb group (PcG) and trithorax components, suggesting novel crosstalk with epigenetic regulation mechanisms in specifying mesodermal cell fates. PcG genes have long been considered to be crucial oncofetal regulators and have become the focus of significant cancer drug development efforts [187,188]. Our findings align with previous reports that gene silencing in EMT involves PcG, for example at *Cdh1*, *CDKN2A* [188–191] and support a model where EMT TFs control the expression of their own coregulators; for example, *Snai1* was shown to recruit polycomb repressive complex 2 members [189]. Overall, these NetNC results predicted components of feedback loops where the Snail, Twist EMT transcription factors regulate chromatin organisation genes that, in turn, may both reinforce and coordinate downstream stages in gene expression programmes for mesoderm development and cancer progression. Stages of the EMT programme had been described elsewhere, reviewed in [39]; our results map networks that may control the remodelling of Waddington's landscape - identifying crosstalk between Snail, Twist, epigenetic modifiers and regulation of key developmental pathways, including notch [192]. We speculate that dynamic interplay between successive cohorts of TFs and chromatin organisation

factors could be an attractive mechanism to determine progress through and the ordering of steps in (partial) EMTs, consistent with 'metastable' intermediate stages [39].

Our work integrates datasets from *D. melanogaster* and human breast cancers, offering insight into the biology of epithelial remodelling in both systems. Indeed, the fly genome is relatively small and hence more tractable for network studies, while the availability of data for analysis (e.g. ChIP-chip, ChIP-seq, genetic screens) is enhanced by both considerable community resources and the relative ease of experimental manipulation [193,194]. The datasets sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b} represent embryos formed entirely from mesodermal lineages [43] and, together, had significantly greater proportion of basal-like breast cancer genes than the combined sna_2-3h_union, twi_2-3h_union datasets ($p<8.0\text{x}10^{-4}$). This enrichment aligned with previous reports showing that basal-like breast cancers have EMT characteristics [138,139]. Indeed the NetNC results map mechanistic commonalities between mesoderm development and breast cancers, including evidence for molecular features of EMT in normal-like (NL) breast cancers. Multiple EMT factors, including *SNAI2* and *TWIST1*, had highest expression values in NL cancers and were assigned to the NL centroid. Previous work had shown enrichment of non-epithelial genes in the normal-like subtype [128]. EMT was known to confer stem-like cell properties [178,195,196] and our results were consistent with dedifferentiation or arrested differentiation due to activation of an EMT-like programme, forming a stem-like cell subpopulation in NL cancers. For example, *SNAI2* had been linked with a stem-like signature in breast cancer metastasis and was critical for maintenance of mammary stem cells [197,198]. NetNC predicted targets for Twist included the proteosome, splicing and ribosomal components; orthologous genes for these subnetworks were largely assigned to the NL subtype in multiple TF_ALL datasets, suggesting potential regulation of these cell systems by *TWIST1* in NL cancers. Some EMT genes were highly expressed in both basal-like and NL cancers, for example *QKI* (Figure 5); EMT-like signalling may therefore be a common thread connecting these two subtypes despite other important differences, such as hormone

receptor status [199]. Indeed, the majority of predicted Snail and Twist functional targets had orthologues that were assigned to either basal-like or NL cancers, providing further evidence that EMT-like signalling is important in both subtypes. We note that cell-compositional effects, associated with a previously reported high proportion of stromal tissue in NL tumours [200], could explain the observed enrichment of EMT molecular characteristics in this subtype. In addition to stromal compositional differences in the NL subtype, as noted above, an EMT signature might reflect inhibition of differentiation. Indeed, NL cancers were previously shown to have high expression of stem cell markers [128,201–203]. Our results demonstrated that NetNC functional targets from fly mesoderm development capture clinically relevant molecular features of breast cancers and revealed novel candidate drivers of tumour progression. Roles in control of invasion were found for four predicted functional targets (*UNK*, *SNX29*, *ATG3*, *IRX4*) in ectopic expression and shRNA knockdown experiments with a Snail inducible breast cancer cell line. Potential artefacts associated with changes in cell growth or proliferation are controlled within the transwell assays used, because values reflect the ratio of signal from cells located at either side of the matrigel barrier. These *in vitro* confirmatory results both support the novel analysis approach and evidence new function for the genes examined.

All nine of the TF_ALL datasets had high predicted NetNC-lcFDR neutral binding proportion (PNBP), ranging from 50% to ≥80%. These PNBP values may reflect an upper limit on neutral binding because some functional targets could be missed; for example due to errors in assigning enhancer binding to target genes and *bona fide* regulation of genes that have few DroFN edges with other candidate ChIP-seq or ChIP-chip targets. While neutral TF binding may arise partly from non-specific associations of TFs with euchromatin, alternative explanations include dormant binding, possibly reflecting developmental lineage [204] or enhancer priming [205]. Additionally, calibration of lcFDR values against synthetic data based on KEGG might influence neutral binding estimates, due to potential differences in network properties between TF targets and

KEGG pathways; such as clustering coefficient. Candidate target genes that were assigned to peaks according to RNA polymerase occupancy [24] had PNBP similar to or lower than datasets where RNA polymerase data was not used. Therefore, we found no evidence of benefit in using RNA polymerase binding data to guide peak matching. Candidate targets for the twi_2-4h_Toll^{10b}, sna_2-4h_Toll^{10b} datasets were defined using a relatively generous peak threshold (two-fold enrichment), which may explain the high PNBP found for sna_2-4h_Toll^{10b}. Twi_2-4h_Toll^{10b} had similar PNBP to the other Twist datasets analysed, although application of a higher peak enrichment threshold would likely lead to a lower PNBP value for this dataset. Indeed, twi_2-6h_intersect had the strongest peak intensity and lowest PNBP compared with other datasets from the same study (twi_2-4h_intersect, twi_4-6h_intersect). Candidate targets for twi_2-6h_intersect were continuously bound across two different time periods; the only other member of TF_ALL that represented binding at multiple time periods was the HOT dataset, which also had low PNBP. Indeed, the only dataset with lower PNBP than either HOT or twi_2-6h_intersect was the Twist ChIP-seq 'high-confidence' dataset (twi_1-3h_hiConf) where the most stringent peak filtering protocols had been applied [9]. Twi_1-3h_hiConf was the only ChIP-seq dataset analysed in this study, however this factor alone is unlikely to explain the high proportion of predicted functional binding. Indeed, overlap with ChIP-chip results informed classification of the 'high-confidence' ChIP-seq peaks taken for twi_1-3h_hiConf [9]. We found similar NetNC PNBP values for datasets produced by taking either the intersection or the union of two independent Twist antibodies. Hits identified by multiple antibodies may be technically more robust due to reduced off-target binding [44]. However, taking the union of candidate binding sites could eliminate false negatives arising from epitope steric occlusion, for example due to context-specific protein interactions. The similarity of PNBP values for either the intersection or the union of Twist antibodies suggests that, despite the higher expected technical specificity, the intersection of candidate targets may not enrich for functional binding sites at the 1% peak-calling FDR threshold applied [24,44]. In general, fewer

false negatives implies recovery of numerically more functional TF targets that therefore may produce denser clusters in DroFN which, in turn, could facilitate NetNC discovery of functional targets. Indeed, datasets representing the union of two antibodies ranked highly in terms of both the total number and proportion of genes recovered at lcFDR<0.05 or pFDR<0.05 (Figure 3).

NetNC may be widely useful for discovery of highly connected gene groups across multiple different data types. Further possible applications include: identification of differentially expressed pathways and macromolecular complexes from functional genomics data; illuminating common biology among CRISPR screen hits in order to inform prioritisation of candidates for follow-up work [206]; and discovery of functional coherence in chromosome conformation capture data (4-C, 5-C), for example in enhancer regulatory relationships [207,208]. NetNC may be applied to any undirected network; including protein-protein or genetic interactions, telecommunications, climate and social networks. Indeed, context-specific effects are important for many disciplines; for example a given social event is unlikely to involve everyone in the social network, and regulatory changes may only apply to a subset of businesses in an economic model. The multiple complementary analysis modes in NetNC provide adaptability to extract value from real-world datasets. A parameter-free mode, NetNC-FBT, provides resilience to enable discovery of coherent genes with graph properties different to those of the KEGG pathways used in calibration of the 'Functional Target Identification' analysis mode (NetNC-FTI). NetNC-FBT employs unsupervised clustering, and analyses the shape of the NFCS score distribution rather than absolute score values. Therefore, NetNC-FBT can separate high-scoring arbitrary subgraphs from disconnected or sparsely connected nodes in the input data. We note that NetNC-FBT had a low false positive rate on blind test data (Figure 2). On the other hand, the NetNC-FTI approach does not assume that the input gene list contains a large proportion of low-scoring genes and therefore has clear advantages for analysis of datasets largely composed of functionally coherent genes (nodes). Also, NetNC-FTI gave the best overall performance for discrimination between biological pathways and Synthetic

Neutral Target Genes (SNTGs). The NetNC software distribution includes a conservative, empirical method for estimation of local False Discovery Rate (lcFDR) from global FDR values, which could be useful in a wide range of applications. For example, FDR estimation is fundamental for mass spectrometry proteomics [209,210] where target-decoy searching approaches typically utilise a single 'decoy' search as the basis for fitting a null (H_0) score distribution in order to estimate lcFDR [209–211]. However, NetNC generates H_0 by resampling, which would be equivalent to having multiple decoy searches, which therefore enables estimation of local FDR by stepping through global FDR values. There might be merit in further investigation of the NetNC local FDR estimation strategy in the context of proteomics database searching. Evaluation on blind test data alongside leading clustering algorithms (MCL [36], HC-PIN [37]) showed that NetNC performed well overall, with particular advantages for analysis of datasets that had substantial (>50%) synthetic neutral TF binding. Indeed, the nine TF_ALL datasets examined were predicted to have at least 50% neutral binding, aligning well with application of NetNC for discovery of functional targets in ChIP-chip and ChIP-seq data. TF binding focus networks derived from NetNC may also be useful in prioritising components for inclusion within regulatory network modelling. Software and datasets are made freely available as Additional Files associated with this publication.

NetNC does not require *a priori* definition of gene groupings, but instead dynamically defines clusters within the subnetwork induced in DroFN by the input gene list. Therefore, NetNC is complementary to techniques that employ static, predefined gene groupings such as GSEA [32], DAVID [33] and GGEA [50]). For example, NetNC discovered functional groups for poorly characterised genes (Figure 4A, bottom right). Additionally, NetNC may be used for dimensionality reduction in gene-wise multiple hypothesis testing. One example application could be analysis of a gene list defined using a differential expression fold-change threshold, providing a hypothesis-generating step prior to evaluation of statistical significance performed on individual coherent genes or on gene clusters. The NetNC output would therefore identify a subset of genes, based on network

coherence, for input into significance testing. Benjamini-Yekutieli false discovery rate control [212] would be appropriate due to the expected dependency of expression values from genes within NetNC clusters. This approach appears attractive for analysis of high-dimensional data, such as transcriptome profiling, where statistical power is diluted by the large number of hypotheses (genes) tested relative to the small number of biological samples that are typically available for analysis. Indeed, established functional genomics data processing workflows involve filtering to reduce dimensionality; for example to eliminate genes with expression values indistinguishable from the assay 'background' [213,214]. NetNC could be deployed as a filter to select coherent genes according to the prior knowledge encoded by a functional gene network (FGN); NetNC would therefore generate a hypothesis for candidate differentially expressed genes based on the biological context represented by the FGN and the assumption that gene expression changes occur coherently, forming network communities. Statistical evaluation of this network coherence property, including estimation of FDR, is available within NetNC for numerical thresholding. Therefore, NetNC has novel applications in distillation of knowledge from high-dimensional data, including single-subject datasets which is an important emerging area for precision medicine [215]. Application of statistical and graph theoretic methods for quantitative evaluation of relationships between genes (nodes) in NetNC offers an alternative to the classical emphais on individual genes in studying the relationship between genotype and phenotype [216].

4 Conclusions

We demonstrated a novel approach to functional TF target discovery using the NetNC algorithm, which was developed and calibrated to separate the signal for functionally coherent target genes from the 'noise' of neutral binding in a ChIP-seq or ChIP-chip experiment. NetNC compared well to application of standard network community detection approaches in this context. Indeed, all nine TF datasets studied had a high level of neutral binding, corresponding closely with the benchmark

datasets where NetNC had the greatest performance advantages. Therefore, NetNC appears the method of choice for elimination of neutral binding and offers an unbiased, systematic approach to help maximise the value of genome-scale TF occupancy data. We investigated current experimental strategies designed to enrich for functional TF binding and, reassuringly, found that NetNC predicted highest coherence for the most stringent peak filtering approach [9]. TF binding at overlapping time points and higher peak intensity were associated with functional coherence. However, taking the intersection of binding from multiple antibodies for a single TF did not demonstrate clear benefit over taking the union of binding sites in our analysis, possibly due to NetNC actively identifying the larger pool of functional TF binding sites within longer candidate target gene lists. Our results align with evidence that HOT regions function in gene regulation, despite depletion for known TF motifs [15,217,218], and supported the emerging picture of widespread combinatorial control involving TF-TF interactions, cooperativity and TF redundancy [2,5,7,219,220].

We presented genome-scale maps of genes downstream of Snail and Twist in *D. melanogaster* early development, finding considerable overlap in results across multiple datasets. Integration of NetNC networks with *Notch* screens and the expression of orthologous human breast cancer genes provided for deep analysis of the conserved molecular networks that orchestrate epithelial remodelling in development and tumour progression. For example, we evidenced regulation of major epigenetic regulators that impact upon polycomb, trithorax; and proposed new TF functions, such as regulation of mushroom body proliferation factors by Twist. Surprisingly, orthologous functional TF targets discriminated between intrinsic breast cancer subtypes. We revealed subtype-specific molecular features as well as commonalities between individual subtypes. For example, subtype-specific features included differential regulation of BMP signalling components (BMPR1, BMP2) between luminal and basal-like cancers. We identified upregulation of basal-like features in the luminal B but not luminal A cancers that may contribute to the relatively

worse prognosis of luminal B subtype. The normal-like subtype had a distinctive EMT signature and multiple genes were highly expressed in both normal-like and basal-like cancers. Differences between the biology of normal-like and basal-like subtypes were captured in clusters of orthologous genes within the NetNC functional TF binding networks. Cell processes specific to basal-like cancers included RNA degradation, transcriptional regulation, ion antiport, mitochondrial translation and chromatin organisation; genes in these processes were *Notch* signalling modifiers and may control the consequences of notch activation in basal-like tumours. Therefore, our work crystalises information from multiple datasets in order to predict novel molecular characteristics for clinically important breast cancer subtypes. Our approach is supported by results validating invasion roles for four functional targets predicted from NetNC analysis of candidate Snail and Twist targets.

5 Methods

5.1 A Comprehensive *D. melanogaster* functional gene network (DroFN)

A high-confidence, comprehensive *Drosophila melanogaster* functional network (DroFN) was developed using a previously described inference approach [49]. Functional interaction probabilities, corresponding to pathway co-membership, were estimated by logistic regression of Bayesian probabilities from STRING v8.0 scores [221] and Gene Ontology (GO) coannotations [60], taking KEGG [222] pathways as gold standard.

Gene pair co-annotations were derived from the GO database of March 25th 2010. The GO Biological Process (BP) and Cellular Component (CC) branches were read as a directed graph and genes added as leaf terms. The deepest term in the GO tree was selected for each gene pair, and BP was given precedence over CC. Training data were taken from KEGG v47, comprising 110 pathways (TRAIN-NET). Bayesian probabilities for STRING and GO coannotation frequencies were derived from TRAIN-NET [49]. Selection of negative pairs from TRAIN-NET using the *perl*

rand() function was used to generate training data with equal numbers of positive and negative pairs (TRAIN-BAL), which was input for logistic regression, to derive a model of gene pair functional interaction probability:

$$p(I \mid GO, STRING) = \frac{1}{1+\left(e^{-6.75+1.03\,pGO+1.12\,pSTRING}\right)} \quad (1)$$

Where:

pGO is the Bayesian probability derived from Gene Ontology coannotation frequency

pSTRING is the Bayesian probability derived from the STRING score frequency

The above model was applied to TRAIN-NET and the resulting score distribution thresholded by seeking a value that maximised the F-measure [223] and True Positive Rate (TPR), while also minimising the False Positive Rate (FPR). The selected threshold value ($p \geq 0.779$) was applied to functional interaction probabilities for all possible gene pairs to generate the high-confidence network, DroFN.

For evaluation of the DroFN network, time separated test data (TEST-TS) were taken from KEGG v62 on 13/6/12, consisting of 14 pathways that were not in TRAIN-NET. TEST-TS was screened against TRAIN-NET, eliminating 34 positive and 218 negative gene pairs to generate the blind test dataset TEST-NET (4599 pairs). GeneMania (version of 10th August 2011) [46] and DROID (v2011_08) [45] were assessed against TEST-NET.

5.2 Network neighbourhood clustering (NetNC) algorithm

NetNC identifies functionally coherent nodes in a subgraph *S* of functional gene network *G* (an undirected graph), induced by some set of nodes of interest *D*; for example, candidate transcription factor target genes assigned from analysis of ChIP-seq data. Intuitively, we consider the proportion of common neighbours for nodes in *S* to define coherence; for example, nodes that share neighbours

have greater coherence than nodes that do not share neighbours. The NetNC workflow is summarised in Figure 1 and described in detail below. Two analysis modes are available a) node-centric (parameter-free) and b) edge-centric, with two parameters. Both modes begin by assigning a p-value to each edge (S_{ij}) from Hypergeometric Mutual Clustering (HMC) [51], described in points one and two, below.

1. A two times two contingency table is derived for each edge S_{ij} by conditioning on the Boolean connectivity of nodes in S to S_i and S_j. Nodes S_i and S_j are not counted in the contingency table.
2. Exact hypergeometric p-values [51] for enrichment of the nodes in S that have edges to the nodes S_i and S_j are calculated using Fisher's Exact Test from the contingency table. Therefore, a distribution of p-values (H_1) is generated for all edges S_{ij}.
3. The NetNC edge-centric mode employs positive false discovery rate [52] and an iterative minimum cut procedure [53] to derive clusters as follows:
    a) Subgraphs with the same number of nodes as S are resampled from G, application of steps 1 and 2 to these subgraphs generates an empirical null distribution of neighbourhood clustering p-values (H_0). This H_0 accounts for the effect of the sample size and the structure of G on the S_{ij} hypergeometric p-values (p_{ij}). Each NetNC run on TF_ALL in this study resampled 1000 subgraphs to derive H_0.
    b) Each edge in S is associated with a positive false discovery rate (q) estimated over p_{ij} using H_1 and H_0. The neighbourhood clustering subgraph C is induced by edges where the associated $q \leq Q$.
    c) An iterative minimum cut procedure [53] is applied to C until all components have density greater than or equal to a threshold Z. Edge weights in this procedure are taken as the negative log p-values from H_1.

d) As described in section 4.2.3, thresholds Q and Z were chosen to optimise the performance of NetNC on the 'Functional Target Identification' task using training data taken from KEGG. Connected components with less than three nodes are discarded, in line with common definitions of a 'cluster'. Remaining nodes are classified as functionally coherent.

4. The node-centric, parameter-free mode proceeds by calculating degree-normalised node functional coherence scores (NFCS) from H_1, then identifies modes of the NFCS distribution using Gaussian Mixture Modelling (GMM) [54]:

a) The node functional coherence score (NFCS) is calculated by summation of S_{ij} p-values in H_1 (p_{ij}) for fixed S_i, normalised by the S_i degree value in S (di):

$$NFCS_i = -\frac{1}{d_i}\sum_j \log(p_{ij}) \qquad (2)$$

b) GMM is applied to identify structure in the NFCS distribution. Expectation-maximization fits a mixture of Gaussians to the distribution using independent mean and standard deviation parameters for each Gaussian [54,224]. Models with 1..9 Gaussians are fitted and the final model selected using the Bayesian Information Criterion (BIC).

c) Nodes in high-scoring mode(s) are predicted to be ‘Functionally Bound Targets’ (FBTs) and retained. Firstly, any mode at NFCS<0.05 is excluded because this typically represents nodes with no edges in S (where NFCS=0). A second step eliminates the lowest scoring mode if >1 mode remains. Very rarely a unimodal model is returned, which may be due to a large non-Gaussian peak at NFCS=0 confounding model fitting; if necessary this is addressed by introducing a tiny Gaussian noise component (SD=0.01) to the NFCS=0 nodes to produce NFCS_GN0. GMM is performed on NFCS_GN0 and nodes eliminated according to the above procedure on the resulting model. This

procedure was developed following manual inspection of results on training data from KEGG pathways with 'synthetic neutral target genes' (STNGs) as nodes resampled from *G* (TRAIN-CL, described in section 2.2.1).

Therefore, NetNC can be applied to predict functional coherence using either edge-centric or node-centric analysis modes. The edge-centric mode automatically produces a network, whereas the node-centric analysis does not output edges; therefore to generate networks from predicted FBT nodes an edge pFDR threshold may be applied, pFDR≤0.1 was selected as the default value. The statistical approach to estimate pFDR and local FDR are described in the sections below.

5.2.1 Estimating positive false discovery rate for hypergeometric mutual clustering p-values

The following procedure is employed to estimate positive False Discovery Rate (pFDR) [52] in the NetNC edge-centric mode. Subgraphs with number of nodes identical to *S* are resampled from *G* to derive a null distribution of HMC *p*-values (H_0) (section 4.2, above). The resampling approach for pFDR calculation in NetNC-FTI controls for the structure of the network *G*, including degree distribution, but does not control for the degree distribution or other network properties of the subgraph *S* induced by the input nodelist (*D*). In scale free and hierarchical networks, degree correlates with clustering coefficient; indeed, this property is typical of biological networks [225]. Part of the rationale for NetNC assumes that differences between the properties of *G* and *S* (for example; degree, clustering coefficient distributions) may enable identification of clusters within *S*. Therefore, it would be undesirable to control for the degree distribution of *S* during the resampling procedure for pFDR calculation because this would also partially control for clustering coefficient. Indeed clustering coefficient is a node-centric parameter that has similarity with the edge-centric Hypergeometric Clustering Coefficient (HMC) calculation [51] used in the NetNC algorithm to

analyse *S*. Hence, the resampling procedure does not model the degree distribution of *S*, although the degree distribution of *G* is controlled for. Positive false discovery rate is estimated over the *p*-values in H_1 (p_{ij}) according to Storey [52]:

$$pFDR = E\left(\frac{V}{R}\right), R>0 \quad (3)$$

Where:

R denotes hypotheses (edges) taken as significant

V are the number of false positive results (type I error)

NetNC steps through threshold values (p_α) in p_{ij} estimating *V* using edges in H_0 with $p \leq p_\alpha$. H_0 represents *Y* resamples, therefore *V* is calculated at each step:

$$V = \frac{H_0}{Y}, p \leq p_\alpha \quad (4)$$

The H_1 *p*-value distribution is assumed to include both true positives and false positives (FP); H_0 is taken to be representative of the FP present in H_1. This approach has been successfully applied to peptide spectrum matching [226,227]. The value of *R* is estimated by:

$$R = \sum_{p \in H_1} \begin{cases} 1 & p_{ij} \leq p_\alpha \\ 0 & otherwise \end{cases} \quad (5)$$

Additionally, there is a requirement for monotonicity:

$$pFDR_{x+1} \geq pFDR_x \ , \quad p_x < p_{x+1} \quad (6)$$

Equation (6) represents a conservative procedure to prevent inconsistent scaling of pFDR due to sampling effects. For example consider the scaling of pFDR for $pFDR_{x+1}$ at a p_{ij} value with additional edges from H_1 but where no more resampled edges (i.e. from H_0) were observed in the interval between p_x and p_{x+1}; before application of equation (6), the value of $pFDR_{x+1}$ would be lower than $pFDR_x$. The approach also requires setting a maximum on estimated pFDR, considering that there may be values of p_α where R is less than V. We set the maximum to 1, which would correspond to a prediction that all edges at p_{ij} are FPs. The assumption that H_1 includes false positives is expected to hold in the context of candidate transcription factor target genes and also generally across biomedical data due to the stochastic nature of biological systems [228–230]. We note that an alternative method to calculate R using both H_1 and H_0 would be less conservative than the approach presented here.

5.2.2 Estimating local false discovery rate from global false discovery rate

We developed an approach to estimate local false discovery rate (lcFDR) [231], being the probability that an object at a threshold (p_α) is a false positive (FP). Our approach takes global pFDR values as basis for lcFDR estimation. In the context of NetNC analysis using the DroFN network, a FP is defined as a gene (node) without a pathway comembership relationship to any other nodes in the nodelist D. The most significant pFDR value ($pFDR_{min}$) from NetNC was determined for each node S_i across the edge set S_{ij}. Therefore, $pFDR_{min}$ is the pFDR value at which node S_i would be included in a thresholded network. We formulated lcFDR for the nodes with $pFDR_{min}$ meeting a given p_α (k) as follows:

$$lcFDR_k = \frac{((n \times pFDR_k) - ((n - X) \times pFDR_l))}{X} \quad (7)$$

Where l denotes the $pFDR_{min}$ closest to and smaller than k, and where at least one node has $pFDR_{min} \equiv pFDR_l$. Therefore, our approach can be conceptualised as operating on ordered $pFDR_{min}$ values. n indicates the nodes in D with $pFDR_{min}$ values meeting threshold k. X represents the number of nodes at $p_\alpha \equiv k$. The number of FPs for nodes with $p_\alpha \equiv k$ (FP_k) is estimated by subtracting the FP for threshold l from the FP at threshold k. Thus, division of FP_k by X gives local false discovery rate bounded by k and l (Supplementary Figure S7). If we define the difference between $pFDR_k$ and $pFDR_l$:

$$pFDR_\Delta = pFDR_k - pFDR_l \quad (8)$$

Substituting $pFDR_k$ for ($pFDR_l + pFDR_\Delta$) into equation (7) and then simplifying gives:

$$lcFDR_k = ((n \times pFDR_\Delta) / X) + pFDR_l \quad (9)$$

Equations (7) and (9) do not apply to the node(s) in D at the smallest possible value of $pFDR_{min}$ because $pFDR_l$ would be undefined; instead, the value of $lcFDR_k$ is calculated as the (global) $pFDR_{min}$ value. Indeed, global FDR and local FDR are equivalent when H_1 consists of objects at a single $pFDR_{min}$ value. Taking the mean $lcFDR_k$ across D provided an estimate of neutral binding in the studied ChIP-chip, ChIP-seq datasets and was calibrated against mean lcFDR values from datasets that had a known proportion of Synthetic Neutral Target Genes (SNTGs). Estimation of the total proportion of neutral binding in ChIP-chip or ChIP-seq data required lcFDR rather than (global) pFDR and, for example, accounts for the shape of the H_1 distribution. In the context of NetNC analysis of TF_ALL, mean lcFDR may be interpreted as the probability that any candidate target gene is neutrally bound in the dataset analysed; therefore providing estimation of the total neutral binding proportion. Computer code for calculation of lcFDR is provided within the NetNC

distribution (Additional File 5). Estimates of SNTGs by the NetNC-FBT approach were not taken forward due to large 95% CI values (Supplementary Figure S8).

5.2.3 NetNC benchmarking and parameter optimisation

Gold standard data for NetNC benchmarking and parameterisation were taken as pathways from KEGG (v62, downloaded 13/6/12) [222]. Training data were selected as seven pathways (TRAIN-CL, 184 genes) and a further eight pathways were selected as a blind test dataset (TEST-CL, 186 genes) summarised in Supplementary Table S7. For both TRAIN-CL and TEST-CL, pathways were selected to be disjoint and to cover a range of different biological functions. However, pathways with shared biology were present within each group; for example TRAIN-CL included the pathways dme04330 'Notch signaling' and dme04914 'Progesterone-mediated oocyte maturation', which are related by notch involvement in oogenesis [232,233]. TEST-CL also included the related pathways dme04745 'Phototransduction' and dme00600 'Sphingolipid metabolism', for example where ceramide kinase regulates photoreceptor homeostasis [234–236].

Gold standard datasets were also developed in order to investigate the effect of dataset size and noise on NetNC performance. The inclusion of noise as resampled network nodes into the gold-standard data was taken to model neutral TF binding [1,8] and matches expectations on data taken from biological systems in general [228,230]. Therefore, gold standard datasets were generated by combining TRAIN-CL with nodes resampled from the network (G) and combining these with TRAIN-CL. The final proportion of resampled nodes (Synthetic Neutral Target Genes, SNTGs) ranged from 5% through to 80% in 5% increments. Since we expected variability in the network proximity of SNTGs to pathway nodes (S), 100 resampled datasets were generated per %SNTG increment. Further gold-standard datasets were generated by taking five subsets of TRAIN-CL, from three through seven pathways. Resampling was applied for these datasets as described above

to generate node lists representing five pathway sets in TRAIN-CL by sixteen %SNTG levels by l00 repeats (TRAIN_CL_ALL, 8000 node lists; Additional File 2). A similar procedure was applied to TEST-CL, taking from three through eight pathways to generate data representing six pathway subsets by sixteen noise levels by 100 repeats (TEST-CL_ALL, 9600 node lists, Additional File 3). Data based on eight pathways (TEST-CL_8PW, 1600 node lists) were used for calibration of lcFDR estimates. Preliminary training and testing against the MCL algorithm [36] utilised a single subsample for 10%, 25%, 50% and 75% SNTGs (TRAIN-CL-SR, TEST-CL-SR; Additional File 6).

NetNC analysed the TRAIN-CL_ALL datasets in edge-centric mode, across a range of FDR (*Q*) and density (*Z*) threshold values. Performance was benchmarked on the Functional Target Identification (FTI) task which assessed the recovery of biological pathways and exclusion of SNTGs. Matthews correlation coefficient (MCC) was computed as a function of NetNC parameters (Q, Z). MCC is attractive because it is captures predictive power in both the positive and negative classes. FTI was a binary classification task for discrimination of pathway nodes from noise, therefore all pathway nodes were taken as as positives and SNTGs were negatives for the FTI MCC calculation. The FTI approach therefore tests discrimination of pathway nodes from SNTGs, which is particularly relevant to identification of functionally coherent candidate TF targets from ChIP-chip or ChIP-seq peaks.

Parameter selection for NetNC on the FTI task analysed MCC values for the 100 SNTG resamples across five pathway subsets by sixteen SNTG levels in TRAIN-CL_ALL over the Q, Z values examined, respectively ranging from up to 10^{-7} to 0.8 and from up to 0.05 to 0.9. Data used for optimisation of NetNC parameters (Q, Z) are given in Additional File 7 and contour plots showing mean MCC across Q, Z values per %SNTG are provided in Supplementary Figure S9. A 'SNTG specified' parameter set was developed for situations where an estimate of the input data noise component is available, for example from the node-centric mode of NetNC. In this parameterisation, for each of the sixteen datasets with different proportions of SNTG (5% .. 80%),

MCC values were normalized across the five pathway subsets of TRAIN-CL (from three through seven pathways), by setting the maximum MCC value to 1 and scaling all other MCC values accordingly. The normalised MCC values <0.75 were set to zero and then a mean value was calculated for each %SNTG value across five pathway subsets by 100 resamples in TRAIN-CL_ALL (500 datasets per noise proportion). This approach therefore only included parameter values corresponding to MCC performance ≥75% of the maximum across the five TRAIN-CL pathway subsets. The high performing regions of these 'summary' contour plots sometimes had narrow projections or small fragments, which could lead to parameter estimates that do not generalise well on unseen data. Therefore, parameter values were selected as the point at the centre of the largest circle (in (Q, Z) space) completely contained in a region where the normalised MCC value was ≥0.95. This procedure yielded a parameter map: (SNTG Estimate) → (Q, Z), given in Supplementary Table S8. NetNC parameters were also determined for analysis without any prior belief about the %SNTG in the input data - and therefore generalise across a wide range of %SNTG and dataset sizes. For this purpose, a contour plot was produced to represent the proportion of datasets where NetNC performed better than 75% of the maximum performance across TRAIN-CL_ALL for the FTI task in the Q, Z parameter space. The maximum circle approach described above was applied to the contour plot in order to derive 'robust' parameter values (Q, Z), which were respectively 0.120, 0.306 (NetNC-FTI).

5.2.4 Performance on blind test data

We compared NetNC against leading methods, HC-PIN [37] and MCL [36] on blind test data (Figure 2, Supplementary Table S1). Input, output and performance summary files for HC-PIN on TEST-CL are given in Additional File 8. HC-PIN was run on the weighted graphs induced in DroFN by TEST-CL with default parameters (lambda = 1.0, threshold size = 3). MCL clusters in DroFN significantly enriched for query nodes from TEST-CL-SR were identified by resampling to

generate a null distribution [49]. Clusters with $q<0.05$ were taken as significant. MCL performance was optimised for the Functional Target Identification (FTI) task over the TRAIN-CL-SR datasets for MCL inflation values from 2 to 5 incrementing by 0.2. The best-performing MCL inflation value overall was 3.6 (Supplementary Table S9).

5.2.5 Subsampling of transcription factor binding datasets and statistical testing

Robustness of NetNC performance was studied by taking 95%, 80% and 50% resamples from nine public transcription factor binding datasets, summarised in section 4.3 and described previously in detail [9,14,24,43,44]. A hundred subsamples of each of these datasets were taken at rates of 95%, 80% and 50%, thereby producing a total of 2700 datasets (TF_SAMPL). NetNC-FTI results across TF_SAMPL were used as input for calculation of median and 95% confidence intervals for the edge and gene overlap per subsampling rate for each transcription factor dataset analysed. The NetNC resampling parameter (Y) was set at 100, the default value. The edge overlap was calculated as the proportion of edges returned by NetNC-FTI for the subsampled dataset that were also present in NetNC-FTI results for the full dataset (i.e. at 100%). Therefore, nine values for median overlap and 95% CI were produced per subsampling rate for both edge and gene overlap, corresponding to the nine transcription factor binding datasets (Supplementary Table S3). The average (median) value of these nine median overlap values, and of the 95% CI, was calculated per subsampling rate; these average values are quoted in Results section 2.4.

False discovery rate (FDR) correction of *p*-values was applied where appropriate and is indicated in this manuscript by the commonly used notation '*q*' Benjamini-Hochberg correction was applied [237] unless otherwise specified in the text. The pFDR and local FDR values calculated by NetNC are described in Methods sections 4.2, 4.2.1 and 4.2.2 (above).

5.3 Transcription factor binding and Notch modifier datasets

We analysed public Chromatin Immunoprecipitation (ChIP) data for the transcription factors *twist* and *snail* in early *Drosophila melanogaster* embryos. These datasets were derived using ChIP followed by microarray (ChIP-chip) [24,43,44] and ChIP followed by solexa pyrosequencing (ChIP-seq) [9]. Additionally 'highly occupied target' regions, reflecting multiple and complex transcription factor occupancy profiles, were obtained from ModEncode [14]. Nine datasets were analysed in total (TF_ALL) and are summarised below.

The 'union' datasets (WT embryos 2-3h, mostly late stage four or early stage five) combined ChIP-chip peaks significant at 1% FDR for two different antibodies targeted at the same TF and these were assigned to the closest transcribed gene according to PolII binding data [24]. Additionally, where the closest transcribed gene was absent from the DroFN network then the nearest gene was included if it was contained in DroFN. This approach generated the datasets sna_2-3h_union (1158 genes) and twi_2-3h_union (1848 genes). The union of peaks derived from two separate antibodies maximised sensitivity and may have reduced potential false negatives arising from epitope steric occlusion. For the 'Toll^{10b}' datasets, significant peaks with at least two-fold enrichment for Twist or Snail binding were taken from ChIP-chip data on Toll^{10b} mutant embryos (2-4h), which had constitutively activated Toll receptor [43,238]; mapping to DroFN generated the datasets twi_2-4h_Toll^{10b} (1238 genes), sna_2-4h_Toll^{10b} (1488 genes). Toll^{10b} embryos had high expression of Snail and Twist, which drove all cells to mesodermal fate trajectories [43]. The two-fold enrichment threshold selected for this study reflects ‘weak’ binding, although was expected to include functional TF targets [10]. Therefore the candidate target genes for twi_2-4h_Toll^{10b} and sna_2-4h_Toll^{10b} were expected to contain a significant proportion of false positives. The Highly Occupied Target dataset included 38562 regions, of which 1855 had complexity score ≥8 and had been mapped to 1648 FlyBase genes according to the nearest transcription start site [14]; 677 of these genes were matched to a DroFN node (HOT). The ‘HighConf’ data took Twist

ChIP-seq binding peaks in WT embryos (1-3h) that had been reported to be 'high confidence' assignments; high confidence filtering was based on overlap with ChIP-chip regions, identification by two peak-calling algorithms and calibration against peak intensities for known Twist targets, corresponding to 832 genes [9]. A total of 664 of these genes were found in DroFN (twi_1-3h_hiConf) and represented the most stringent approach to peak calling of all the nine TF_ALL datasets. The intersection of ChIP-chip binding for two different Twist antibodies in WT embryos spanning two time periods (2-4h and 4-6h) identified a total of 1842 target genes [44] of which 1444 mapped to DroFN (Intersect_ALL). Subsets of Intersect_ALL identified regions bound only at 2-4 hours (twi_2-4h_intersect, 801 genes), or only at 4-6 hours (twi_4-6h_intersect, 818 genes), or 'continuously bound' regions identified at both 2-4 and 4-6 hours (twi_2-6h_intersect, 615 genes). Assigned gene targets may belong to more than one subset of Intersect_ALL because time-restricted binding was assessed for putative enhancer regions prior to gene mapping; overlap of the Intersect_ALL subsets ranged between 30.2% and 55.4%. The Intersect_ALL datasets therefore enabled assessment of functional enhancer binding according to occupancy at differing time intervals and also to examine the effect of intersecting ChIPs for two different antibodies upon the proportion of predicted functional targets recovered.

The Notch signalling modifiers analysed in this study were selected based on identification in at least two of the screens reported in [63].

5.4 Breast cancer transcriptome datasets and molecular subtypes

Primary breast tumour gene expression data were downloaded from NCBI GEO (GSE12276, GSE21653, GSE3744, GSE5460, GSE2109, GSE1561, GSE17907, GSE2990, GSE7390, GSE11121, GSE16716, GSE2034, GSE1456, GSE6532, GSE3494, GSE68892 (formerly geral-00143 from caBIG)). All datasets were Affymetrix U133A/plus 2 chips and were summarised with Ensembl alternative CDF [239]. RMA normalisation [240] and ComBat batch correction [241]

were applied to remove dataset-specific bias as previously described [118,242]. Intrinsic molecular subtypes were assigned based upon the highest correlation to Sorlie centroids [111], applied to each dataset separately. Centred average linkage clustering was performed using the Cluster and TreeView programs [243]. Centroids were calculated for each gene based upon the mean expression across each of the Sorlie intrinsic subtypes [111]. These expression values were squared to consider up and down regulated genes in a single analysis. Orthology to the DroFN network was defined using Inparanoid [59]. Differential expression was calculated by t-test comparing normalised (unsquared) expression values in normal-like and basal-like tumours with false discovery rate correction [237].

5.5 Invasion assays for validation of genes selected from NetNC results

MCF-7 Tet-On cells were purchased from Clontech and maintained as previously described [161].To analyse the ability of transfected MCF7 breast cancer cells to degrade and invade surrounding extracellular matrix, we performed an invasion assay using the CytoSelect™ 24-Well Cell Adhesion Assay kit. This transwell invasion assay allow the cells to invade through a matrigel barrier utilising basement membrane-coated inserts according to the manufacturer's protocol. Briefly, MCF7 cells transfected with the constructs (Doxycycline-inducible *SNAI1* cDNA or *SNAI1 shRNA* with or without candidate gene cDNA) were suspended in serum-free medium. *SNAI1* cDNA or *SNAI1 shRNA* were cloned in our doxycyline-inducible pGoldiLox plasmid (pGoldilox-Tet-ON for cDNA and pGolidlox-tTS for shRNA expression) using validated shRNAs against *SNAI1* (NM_005985 at position 150 of the transcript [161]). pGoldilox has been used previously to induce and knock down the expression of *Ets* genes [160]. Following overnight incubation, the cells were seeded at 3.0×10^5 cells/well in the upper chamber and incubated with medium containing serum with or without doxycyline in the lower chamber for 48 hours. Concurrently, 10^6 cells were treated in the same manner and grown in a six well plate to confirm over-expression and

knockdown. mRNA was extracted from these cells and quantitative real-time PCR (RT-qPCR) was performed as previously described [244]; please see Additional File 9 for gene primers. The transwell invasion assay evaluated the ratio of CyQuant dye signal at 480/520 nm in a plate reader of cells from the two wells and therefore controlled for potential proliferation effects associated with ectopic expression. We used empty vector (mCherry) and scrambled shRNA as controls and to control for the non-specific signal. At least three experimental replicates were performed for each reading.

6 List of abbreviations

AML: Acute Myeloid Leukemia

AUC: Area Under the Receiver Operator Characteristic Curve

BIC: Bayesian Information Criterion

BP: Biological Process

CC: Cellular Component

ChIP: Chromatin Immunoprecipitation

ChIP-chip: Chromatin Immunoprecipitation microarray

ChIP-exo: Chromatin Immunoprecipitation with lambda exonuclease digestion (and sequencing)

ChIP-seq: Chromatin Immunoprecipitation sequencing

CI: Confidence Interval

circRNA: circular RNA

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

DamID: DNA Adenine Methyltransferase Identification

DAVID: Database for Annotation Visualisation and Integrated Discovery

DroFN: Drosophila Functional Network

DroID: Drosophila Interactions Database

DPiM: Drosophila Protein Interaction Map

DRC: Developmental Regulation Cluster

EMT: Epithelial to Mesenchymal Transition

FDR: False Discovery Rate

FGF: Fibroblast Growth Factor

GMM: Gaussian Mixture Modelling

GO: Gene Ontology

GPCR: G-protein Coupled Receptor

pFDR: positive False Discovery Rate

lcFDR: local False Discovery Rate

FGN: Functional Gene Network

FP: False Positive

FPR: False Positive Rate

GGEA: Gene Graph Enrichment Analysis

GSEA: Gene Set Enrichment Analysis

HMC: Hypergeometric Mutual Clustering

HOT: Highly Occupied Target

KEGG: Kyoto Encyclopaedia of Genes and Genomes

MCC: Matthews Correlation Coefficient

MT: Mitochondrial Translation

modENCODE: Model Organism Encyclopedia of DNA Elements

NetNC: Network Neighbourhood Clustering

NetNC-FTI: NetNC Functional Target Identification mode

NetNC-FBT: NetNC Functional Binding Target mode

NetNC-lcFDR: NetNC local False Discovery Rate

NFCS: Node Functional Coherence Score

NL: Normal-Like

PcG: Polycomb Group

pFDR: Positive False Discovery Rate

PNBP: NetNC Predicted Neutral Binding Proportion

RT-qPCR: quantitative real-time PCR

siRNA: small interfering RNA

shRNA: short hairpin RNA

SNTG: Synthetic Neutral Target Gene

snoRNA: small nucleolar RNA

snRNP: small nuclear ribonucleoprotein

TF: Transcription Factor

TF_ALL: The nine TF binding datasets studied in this work.

TPR: True Positive Rate

TrxG: Trithorax Group

3D: Three dimensional

4-C: Circular chormosome conformation capture

5-C: Chromosome Conformation Capture Carbon Copy

7 Declarations

Data and software availability

Software and key datasets are made freely available as Additional Files associated with this publication as follows:

Additional File 1: DroFN network and gold standard datasets for network inference.

Additional File 2: TRAIN_CL_ALL (NetNC training data).

Additional File 3: TEST_CL_ALL (NetNC test data).

Additional File 4: Cytoscape sessions with NetNC-FTI results for TF_ALL.

Additional File 5: NetNC software distribution.

Additional File 6: TRAIN-CL-SR and TEST-CL-SR (used for comparison with MCL algorithm).

Additional File 7: NetNC results on training data used for parameter optimisation (Q, Z).

Additional File 8: HCPIN input, output and performance summary files on TEST-CL.

Additional File 9: Primers for RT-qPCR.

Other previously published datasets are available from appropriate repositories and/or files associated with the relevant publications. Please see Methods for specific details and references.

Ethics approval and consent to participate

All tumour data were previously published and obtained from publicly available databases.

Consent for publication

Not applicable

Competing interests

None declared

Funding

Medical Research Council (MC_UU_12018/25; IMO), Royal Society of Edinburgh Scottish Government Fellowship cofunded by Marie Curie Actions (IMO), Marie Curie Fellowship (BH), Breast Cancer Now (AHS). AE was supported by a Wellcome Trust Beit Memorial Fellowship (AE) and by funding from Prof. Nick Hastie's laboratory (MC_PC_U127527180).

Authors' contributions

IMO conceived the overall project, obtained funding, designed the computational and statistical aspects, implemented and benchmarked the NetNC algorithm, performed analysis of all TF datasets and the validation data, interpreted results, produced Figures 1, 3, 4, 6, produced all Tables except as noted below, performed orthology mapping, annotated the heatmap features in Figure 5 and supervised JO, BH, ALRL, MJF, EP-C. JO implemented the iterative minimum cut, co-designed and implemented the NetNC parameter optimisation, assisted with NetNC benchmarking and produced Figures 2, S9, Table S8. BH obtained funding, co-designed and implemented the DroFN network inference, benchmarking and produced Figure S1. MJF co-designed and implemented the comparison of NetNC against the MCL algorithm, produced Table S9. ALRL co-designed and implemented the Gaussian Mixture Modelling aspects of NetNC and co-designed Equation 9. IO, JO and ALRL wrote the NetNC software distribution. AHS obtained funding, co-designed and implemented the breast cancer transcriptome analysis, interpreted results, produced Figures 5 and S6. AE obtained funding, interpreted results, designed and performed all bench laboratory experiments including tissue culture, transfection and transwell assays. EP-C assisted with annotation, visualisation and interpretation of the NetNC-FTI networks, including production of Figure S5. IO led the writing of the manuscript and revised it for important intellectual content with input from JO, AHS, AE, BH, EP-C, ALRL. All authors read and approved the submitted manuscript.

Acknowledgements

IMO is grateful to Prof Jeremy Gunawardena and Prof Peter Sorger for hosting him at HMS and for helpful discussions. Thanks to Prof PS Thiagarajan, Prof Andrew Millar, Prof Wendy Bickmore, Prof Nick Hastie, Prof Mike Levine, Prof Ben Lehner and Prof Julian Dow for invaluable

comments. Mr Nick Moir and Dr Seanna McTaggart assisted with testing the NetNC software distribution.

8 References

1. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15:272–86.

2. Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, Stark A. Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature. 2015;528:147–51.

3. Rhee DY, Cho D-Y, Zhai B, Slattery M, Ma L, Mintseris J, et al. Transcription Factor Networks in Drosophila melanogaster. Cell Reports. 2014;8:2031–43.

4. Zabidi MA, Stark A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. Trends in Genetics. 2016;32:801–14.

5. Khoueiry P, Girardot C, Ciglar L, Peng PC, Gustafson EH, Sinha S, et al. Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. Elife [Internet]. 2017 [cited 2017 Oct 9];6. Available from: http://europepmc.org/abstract/MED/28792889

6. Wilczynski B, Furlong EEM. Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. Developmental Biology. 2010;340:161–9.

7. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527:384–8.

8. Li X, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, et al. Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS Biol. 2008;6:e27.

9. Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, et al. High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. Genome Res. 2011;21:566–77.

10. Biggin MD. Animal Transcription Networks as Highly Connected, Quantitative Continua. Developmental Cell. 2011;21:611–26.

11. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 2012;22:1658–67.

12. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

13. Brown JB, Celniker SE. Lessons from modENCODE. Annual Review of Genomics and Human Genetics. 2015;16:31–53.

14. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science. 2010;330:1787–97.

15. Kvon EZ, Stampfel G, Yáñez-Cuna JO, Dickson BJ, Stark A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev. 2012;26:908–13.

16. Li H, Chen H, Liu F, Ren C, Wang S, Bo X, et al. Functional annotation of HOT regions in the human genome: implications for human disease and cancer. Scientific Reports. 2015;5:11633.

17. Moorman C, Sun LV, Wang J, Wit E de, Talhout W, Ward LD, et al. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. PNAS. 2006;103:12027–32.

18. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, et al. A Regulatory Archipelago Controls Hox Genes Transcription in Digits. Cell. 2011;147:1132–45.

19. Teytelman L, Thurtle DM, Rine J, Oudenaarden A van. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. PNAS. 2013;110:18602–7.

20. Cannavò E, Khoueiry P, Garfield D, Geeleher G, Zichner T, Gustafson E, et al. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr Biol. 2016;26:38–51.

21. Keung AJ, Bashor CJ, Kiriakov S, Collins JJ, Khalil AS. Using Targeted Chromatin Regulators to Engineer Combinatorial and Spatial Transcriptional Regulation. Cell. 2014;158:110–20.

22. Igual JC, Johnson AL, Johnston LH. Coordinated regulation of gene expression by the cell cycle transcription factor Swi4 and the protein kinase C MAP kinase pathway for yeast cell integrity. The EMBO Journal. 1996;15:5001–13.

23. Karczewski KJ, Snyder M, Altman RB, Tatonetti NP. Coherent Functional Modules Improve Transcription Factor Target Identification, Cooperativity Prediction, and Disease Association. PLOS Genetics. 2014;10:e1004122.

24. MacArthur S, Li X-Y, Li J, Brown JB, Chu HC, Zeng L, et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biology. 2009;10:R80.

25. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52.

26. Hooper SD, Boué S, Krause R, Jensen LJ, Mason CE, Ghanim M, et al. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Molecular Systems Biology. 2007;3:72.

27. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18:S233–40.

28. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. Cell. 2011;144:986–98.

29. Jaeger S, Igea A, Arroyo R, Alcalde V, Canovas B, Orozco M, et al. Quantification of Pathway Cross-talk Reveals Novel Synergistic Drug Combinations for Breast Cancer. Cancer Res. 2017;77:459–69.

30. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nat Rev Genet. 2016;17:615–29.

31. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47:569–76.

32. 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. Proc Natl Acad Sci USA. 2005;102:15545–50.

33. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.

34. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature. 1999;402:83.

35. Pe'er D, Hacohen N. Principles and Strategies for Developing Network Models in Cancer. Cell. 2011;144:864–73.

36. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.

37. Wang J, Li M, Chen J, Pan Y. A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2011;8:607–20.

38. Pawson T, Nash P. Assembly of Cell Regulatory Systems Through Protein Interaction Domains. Science. 2003;300:445–52.

39. Nieto MA, Huang RY-J, Jackson RA, Thiery JP. EMT: 2016. Cell. 2016;166:21–45.

40. Lim J, Thiery JP. Epithelial-mesenchymal transitions: insights from development. Development. 2012;139:3471–86.

41. Giampieri S, Manning C, Hooper S, Jones L, Hill CS, Sahai E. Localized and reversible TGFbeta signalling switches breast cancer cells from cohesive to single cell motility. Nat Cell Biol. 2009;11:1287–96.

42. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, et al. Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. Science. 2013;339:580–4.

43. Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, et al. Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev. 2007;21:385–90.

44. Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EEM. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev. 2007;21:436–49.

45. Yu J, Pacifico S, Liu G, Finley Jr. RL. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics. 2008;9:461.

46. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 2010;38:W214–20.

47. Guruharsha KG, Rual J-F, Zhai B, Mintseris J, Vaidya P, Vaidya N, et al. A Protein Complex Network of Drosophila melanogaster. Cell. 2011;147:690.

48. Schramm G, Wiesberg S, Diessl N, Kranz A-L, Sagulenko V, Oswald M, et al. PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. Bioinformatics. 2010;26:1225–31.

49. Overton IM, Graham S, Gould KA, Hinds J, Botting CH, Shirran S, et al. Global network analysis of drug tolerance, mode of action and virulence in methicillin-resistant S. aureus. BMC Syst Biol. 2011;5:68.

50. Geistlinger L, Csaba G, Küffner R, Mulder N, Zimmer R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. Bioinformatics. 2011;27:i366–73.

51. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. PNAS. 2003;100:4372–6.

52. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64:479–98.

53. Ford LR, Fulkerson DR. Maximal flow through a network. Canadian Journal of Mathematics. 1956;8:399–404.

54. Lubbock ALR, Katz E, Harrison DJ, Overton IM. TMA Navigator: network inference, patient stratification and survival analysis with tissue microarray data. Nucleic Acids Research. 2013;41:W562–8.

55. Edgar BA, Schubiger G. Parameters controlling transcriptional activation during early drosophila development. Cell. 1986;44:871–7.

56. Leptin M. Drosophila Gastrulation: From Pattern Formation to Morphogenesis. Annual Review of Cell and Developmental Biology. 1995;11:189–212.

57. Campos-Ortega JA, Hartenstein V. The Embryonic Development of Drosophila melanogaster. 2nd ed. Berlin, Heidelberg: Springer Science & Business Media; 1997.

58. Chen J, Hu Z, Phatak M, Reichard J, Freudenberg JM, Sivaganesan S, et al. Genome-Wide Signatures of Transcription Factor Activity: Connecting Transcription Factors, Disease, and Small Molecules. PLOS Computational Biology. 2013;9:e1003198.

59. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research. 2009;38:D196–203.

60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

61. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005;21:3448–9.

62. Gramates LS, Marygold SJ, Santos G dos, Urbano J-M, Antonazzo G, Matthews BB, et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 2017;45:D663–71.

63. Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet. 2012;13:654–66.

64. Ntziachristos P, Lim JS, Sage J, Aifantis I. From Fly Wings to Targeted Cancer Therapies: A Centennial for Notch Signaling. Cancer Cell. 2014;25:318–34.

65. Bray SJ. Notch signalling in context. Nat Rev Mol Cell Biol. 2016;17:722–35.

66. Nowell CS, Radtke F. Notch as a tumour suppressor. Nature Reviews Cancer. 2017;17:145.

67. Bernard F, Krejci A, Housden B, Adryan B, Bray SJ. Specificity of Notch pathway activation: Twist controls the transcriptional output in adult muscle progenitors. Development. 2010;137:2633–42.

68. Sahlgren C, Gustafsson MV, Jin S, Poellinger L, Lendahl U. Notch signaling mediates hypoxia-induced tumor cell migration and invasion. PNAS. 2008;105:6392–7.

69. Baylies MK, Bate M. twist: a myogenic switch in Drosophila. Science. 1996;272:1481–4.

70. Xie Y, Li X, Deng X, Hou Y, O'Hara K, Urso A, et al. The Ets protein Pointed prevents both premature differentiation and dedifferentiation of *Drosophila* intermediate neural progenitors. Development. 2016;143:3109–18.

71. Chen C-M, Freedman JA, Bettler DR, Manning SD, Giep SN, Steiner J, et al. polychaetoid is required to restrict segregation of sensory organ precursors from proneural clusters in Drosophila. Mechanisms of Development. 1996;57:215–27.

72. Lo PCH, Skeath JB, Gajewski K, Schulz RA, Frasch M. Homeotic Genes Autonomously Specify the Anteroposterior Subdivision of the Drosophila Dorsal Vessel into Aorta and Heart. Developmental Biology. 2002;251:307–19.

73. Trujillo GV, Nodal DH, Lovato CV, Hendren JD, Helander LA, Lovato TL, et al. The canonical Wingless signaling pathway is required but not sufficient for inflow tract formation in the Drosophila melanogaster heart. Developmental Biology. 2016;413:16–25.

74. Hammonds AS, Bristow CA, Fisher WW, Weiszmann R, Wu S, Hartenstein V, et al. Spatial expression of transcription factors in Drosophila embryonic organ development. Genome Biol. 2013;14:R140.

75. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, et al. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. 2002;3:research0088.

76. Hartley D, Xu T, Artavanis-Tsakonas S. The embryonic expression of the Notch locus of Drosophila melanogaster and the implications of point mutations in the extracellular EGF-like domain of the predicted protein., The embryonic expression of the Notch locus of Drosophila melanogaster and the implications of point mutations in the extracellular EGF-like domain of the predicted protein. EMBO J. 1987;6, 6:3407, 3407–17.

77. BDGP. Berkley Drosophila Genome Project in situ homepage [Internet]. Available from: http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl

78. Kusch T, Reuter R. Functions for Drosophila brachyenteron and forkhead in mesoderm specification and cell signalling. Development. 1999;126:3991–4003.

79. Millo H, Bownes M. The expression pattern and cellular localisation of Myosin VI during the Drosophila melanogaster life cycle. Gene Expression Patterns. 2007;7:501–10.

80. Shao Z, Raible F, Mollaaghababa R, Guyon JR, Wu C, Bender W, et al. Stabilization of Chromatin Structure by PRC1, a Polycomb Complex. Cell. 1999;98:37–46.

81. Czermin B, Schotta G, Hülsmann BB, Brehm A, Becker PB, Reuter G, et al. Physical and functional association of SU(VAR)3-9 and HDAC1 in Drosophila. EMBO Reports. 2001;2:915.

82. Schotta G, Ebert A, Krauss V, Fischer A, Hoffmann J, Rea S, et al. Central role of *Drosophila* SU(VAR)3–9 in histone H3-K9 methylation and heterochromatic gene silencing. The EMBO Journal. 2002;21:1121–31.

83. Salvaing J, Lopez A, Boivin A, Deutsch JS, Peronnet F. The Drosophila Corto protein interacts with Polycomb-group proteins and the GAGA factor. Nucleic Acids Res. 2003;31:2873–82.

84. Lopez A, Higuet D, Rosset R, Deutsch J, Peronnet F. corto genetically interacts with Pc-G and trx-G genes and maintains the anterior boundary of Ultrabithorax expression in Drosophila larvae. Mol Gen Genomics. 2001;266:572–83.

85. Mishra K, Chopra VS, Srinivasan A, Mishra RK. Trl-GAGA directly interacts with lola like and both are part of the repressive complex of Polycomb group of genes. Mechanisms of Development. 2003;120:681–9.

86. Quijano JC, Wisotzkey RG, Tran NL, Huang Y, Stinchfield MJ, Haerry TE, et al. lolal Is an Evolutionarily New Epigenetic Regulator of dpp Transcription during Dorsal–Ventral Axis Formation. Molecular Biology and Evolution. 2016;33:2621.

87. Schuster KJ, Smith-Bolton RK. Taranis Protects Regenerating Tissue from Fate Changes Induced by the Wound Response in Drosophila. Developmental Cell. 2015;34:119–28.

88. Calgaro S, Boube M, Cribbs DL, Bourbon H-M. The Drosophila gene taranis encodes a novel trithorax group member potentially linked to the cell cycle regulatory apparatus. Genetics. 2002;160:547.

89. Fauvarque MO, Laurenti P, Boivin A, Bloyer S, Griffin-Shea R, Bourbon HM, et al. Dominant modifiers of the polyhomeotic extra-sex-combs phenotype induced by marked P element insertional mutagenesis in Drosophila. Genet Res. 2001;78:137–48.

90. Tie F, Banerjee R, Saiakhova AR, Howard B, Monteith KE, Scacheri PC, et al. Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. Development (Cambridge, England). 2014;141:1129.

91. Ingham P, Whittle R. Trithorax: A new homoeotic mutation of Drosophila melanogaster causing transformations of abdominal and thoracic imaginal segments. Molec Gen Genet. 1980;179:607–14.

92. Hong S-T, Choi K-W. Antagonistic roles of Drosophila Tctp and Brahma in chromatin remodelling and stabilizing repeated sequences. Nature Communications. 2016;7:12988.

93. Crosby MA, Miller C, Alon T, Watson KL, Verrijzer CP, Goldman-Levi R, et al. The trithorax Group Gene moira Encodes a Brahma-Associated Putative Chromatin-Remodeling Factor in Drosophila melanogaster. Molecular and Cellular Biology. 1999;19:1159.

94. Fanti L, Dorer DR, Berloco M, Henikoff S, Pimpinelli S. Heterochromatin protein 1 binds transgene arrays. Chromosoma. 1998;107:286–92.

95. Fanti L, Pimpinelli S. HP1: a functionally multifaceted protein. Current Opinion in Genetics & Development. 2008;18:169–74.

96. Campos-Ortega JA. Mechanisms of early neurogenesis in Drosophila melanogaster. J Neurobiol. 1993;24:1305–27.

97. Leptin M. twist and snail as positive and negative regulators during Drosophila mesoderm development. Genes Dev. 1991;5:1568–76.

98. Wieschaus E, Nüsslein-Volhard C. The Heidelberg Screen for Pattern Mutants of Drosophila: A Personal Account. Annual Review of Cell and Developmental Biology. 2016;32:1–46.

99. Gilmour D, Rembold M, Leptin M. From morphogen to morphogenesis and back. Nature. 2017;541:311–20.

100. Ashraf SI, Ip YT. The Snail protein family regulates neuroblast expression of inscuteable and string, genes involved in asymmetry and cell division in Drosophila. Development. 2001;128:4757–67.

101. Zander MA, Burns SE, Yang G, Kaplan DR, Miller FD. Snail Coordinately Regulates Downstream Pathways to Control Multiple Aspects of Mammalian Neural Precursor Development. J Neurosci. 2014;34:5164–75.

102. Nevil M, Bondra ER, Schulz KN, Kaplan T, Harrison MM. Stable Binding of the Conserved Transcription Factor Grainy Head to its Target Genes Throughout Drosophila melanogaster Development. Genetics. 2017;205:605–20.

103. Caron SJC, Ruta V, Abbott LF, Axel R. Random convergence of olfactory inputs in the Drosophila mushroom body. Nature. 2013;497:113–7.

104. Lin S, Ewen-Campen B, Ni X, Housden BE, Perrimon N. In Vivo Transcriptional Activation Using CRISPR/Cas9 in Drosophila. Genetics. 2015;201:433–42.

105. Vesuna F, van Diest P, Chen JH, Raman V. Twist is a transcriptional repressor of E-cadherin gene expression in breast cancer. Biochemical and Biophysical Research Communications. 2008;367:235–41.

106. Kraft KF, Massey EM, Kolb D, Walldorf U, Urbach R. Retinal homeobox promotes cell growth, proliferation and survival of mushroom body neuroblasts in the Drosophila brain. Mechanisms of Development. 2016;142:50–61.

107. Orihara-Ono M, Suzuki E, Saito M, Yoda Y, Aigaki T, Hama C. The slender lobes gene, identified by retarded mushroom body development, is required for proper nucleolar organization in Drosophila. Developmental Biology. 2005;281:121–33.

108. Manansala MC, Min S, Cleary MD. The Drosophila SERTAD protein Taranis determines lineage-specific neural progenitor proliferation patterns. Developmental Biology. 2013;376:150–62.

109. Creighton CJ, Chang JC, Rosen JM. Epithelial-Mesenchymal Transition (EMT) in Tumor-Initiating Cells and Its Clinical Implications in Breast Cancer. J Mammary Gland Biol Neoplasia. 2010;15:253–60.

110. Wang Z, Li Y, Kong D, Banerjee S, Ahmad A, Azmi AS, et al. Acquisition of Epithelial-Mesenchymal Transition Phenotype of Gemcitabine-Resistant Pancreatic Cancer Cells Is Linked with Activation of the Notch Signaling Pathway. Cancer Res. 2009;69:2400–7.

111. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. PNAS. 2003;100:8418–23.

112. Cejalvo JM, Dueñas EM de, Galvan P, García-Recio S, Gasión OB, Paré L, et al. Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. Cancer Res. 2017;canres.2717.2016.

113. Stylianou S, Clarke RB, Brennan K. Aberrant Activation of Notch Signaling in Human Breast Cancer. Cancer Res. 2006;66:1517–25.

114. Barnawi R, Al-Khaldi S, Majed Sleiman G, Sarkar A, Al-Dhfyan A, Al-Mohanna F, et al. Fascin Is Critical for the Maintenance of Breast Cancer Stem Cell Pool Predominantly via the Activation of the Notch Self-Renewal Pathway. Stem Cells. 2016;34:2799–813.

115. Ingthorsson S, Briem E, Bergthorsson JT, Gudjonsson T. Epithelial Plasticity During Human Breast Morphogenesis and Cancer Progression. J Mammary Gland Biol Neoplasia. 2016;21:139–48.

116. Zhang M, Lee AV, Rosen JM. The Cellular Origin and Evolution of Breast Cancer. Cold Spring Harb Perspect Med. 2017;7:a027128.

117. Chen J, Imanaka N, Chen J, Griffin JD. Hypoxia potentiates Notch signaling in breast cancer leading to decreased E-cadherin expression and increased cell migration and invasion. Br J Cancer. 2009;102:351–60.

118. Moleirinho S, Chang N, Sims AH, Tilston-Lünel AM, Angus L, Steele A, et al. KIBRA exhibits MST-independent functional regulation of the Hippo signaling pathway in mammals. Oncogene. 2013;32:1821–30.

119. Beltran AS, Graves LM, Blancafort P. Novel role of Engrailed 1 as a prosurvival transcription factor in basal-like breast cancer and engineering of interference peptides block its oncogenic function. Oncogene. 2014;33:4767–77.

120. Adélaïde J, Finetti P, Bekhouche I, Repellini L, Geneix J, Sircoulomb F, et al. Integrated Profiling of Basal and Luminal Breast Cancers. Cancer Res. 2007;67:11565–75.

121. Letessier A, Ginestier C, Charafe-Jauffret E, Cervera N, Adélaïde J, Gelsi-Boyer V, et al. ETV6 gene rearrangements in invasive breast carcinoma. Genes Chromosomes Cancer. 2005;44:103–8.

122. Golub TR, Barker GF, Bohlander SK, Hiebert SW, Ward DC, Bray-Ward P, et al. Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proc Natl Acad Sci USA. 1995;92:4917–21.

123. Buijs A, Sherr S, van Baal S, van Bezouw S, van der Plas D, Geurts van Kessel A, et al. Translocation (12;22) (p13;q11) in myeloproliferative disorders results in fusion of the ETS-like TEL gene on 12p13 to the MN1 gene on 22q11. Oncogene. 1995;10:1511–9.

124. Chapellier M, Bachelard-Cascales E, Schmidt X, Clément F, Treilleux I, Delay E, et al. Disequilibrium of BMP2 Levels in the Breast Stem Cell Niche Launches Epithelial Transformation by Overamplifying BMPR1B Cell Response. Stem Cell Reports. 2015;4:239–54.

125. Ma L, Lu M-F, Schwartz RJ, Martin JF. Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. Development. 2005;132:5601–11.

126. Ren J, Dijke P ten. Bone Morphogenetic Proteins in the Initiation and Progression of Breast Cancer. SpringerLink. 2017;409–33.

127. Katsuno Y, Hanyu A, Kanda H, Ishikawa Y, Akiyama F, Iwase T, et al. Bone morphogenetic protein signaling enhances invasion and bone metastasis of breast cancer cells through Smad pathway. Oncogene. 2008;27:6322–33.

128. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS. 2001;98:10869–74.

129. Liang Y-K, Lin H-Y, Chen C-F, Zeng D. Prognostic values of distinct CBX family members in breast cancer. Oncotarget. 2017;8:92375–87.

130. Wang H, Liang J, Zheng H, Xiao H. Expression and prognostic significance of ECT2 in invasive breast cancer. Journal of Clinical Pathology. 2018;71:442–5.

131. Bermingham JR, Arden KC, Naumova AK, Sapienza C, Viars CS, Fu XD, et al. Chromosomal localization of mouse and human genes encoding the splicing factors ASF/SF2 (SFRS1) and SC-35 (SFRS2). Genomics. 1995;29:70–9.

132. Chan S, Sridhar P, Kirchner R, Lock YJ, Herbert Z, Buonamici S, et al. Basal-A Triple Negative Breast Cancer Cells Selectively Rely on RNA Splicing for Survival. Mol Cancer Ther. 2017;molcanther.0461.2017.

133. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res. 2010;12:R5.

134. Schmid BC, Rezniczek GA, Fabjani G, Yoneda T, Leodolter S, Zeillinger R. The neuronal guidance cue Slit2 induces targeted migration and may play a role in brain metastasis of breast cancer cells. Breast Cancer Res Treat. 2007;106:333–42.

135. Oulad-Abdelghani M, Chazaud C, Bouillet P, Sapin V, Chambon P, Dollé P. Meis2, a novel mouse Pbx-related homeobox gene induced by retinoic acid during differentiation of P19 embryonal carcinoma cells. Developmental Dynamics. 1997;210:173–83.

136. Borrow J, Shearman AM, Stanton VP, Becher R, Collins T, Williams AJ, et al. The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9. Nat Genet. 1996;12:159–67.

137. Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, et al. The RNA Binding Protein Quaking Regulates Formation of circRNAs. Cell. 2015;160:1125–34.

138. Sarrió D, Rodriguez-Pinilla SM, Hardisson D, Cano A, Moreno-Bueno G, Palacios J. Epithelial-Mesenchymal Transition in Breast Cancer Relates to the Basal-like Phenotype. Cancer Research. 2008;68:989–97.

139. Guen VJ, Chavarria TE, Kröger C, Ye X, Weinberg RA, Lees JA. EMT programs promote basal mammary stem cell and tumor-initiating cell stemness by inducing primary ciliogenesis and Hedgehog signaling. PNAS. 2017;201711534.

140. Makino N, Yamato T, Inoue H, Furukawa T, Abe T, Yokoyama T, et al. Isolation and Characterization of the Human Gene Homologous to the Drosophila Headcase (hdc) Gene in Chromosome Bands 6q23-q24, a Region of Common Deletion in Human Pancreatic Cancer. DNA Sequence. 2001;11:547–53.

141. Lin X, Li J, Yin G, Zhao Q, Elias D, Lykkesfeldt AE, et al. Integrative analyses of gene expression and DNA methylation profiles in breast cancer cell line models of tamoxifen-resistance indicate a potential role of cells with stem-like properties. Breast Cancer Research. 2013;15:R119.

142. Chien C-C, Chang C-C, Yang S-H, Chen S-H, Huang C-J. A homologue of the Drosophila headcase protein is a novel tumor marker for early-stage colorectal cancer. Oncol Rep. 2006;15:919–26.

143. Resende LPF, Boyle M, Tran D, Fellner T, Jones DL. Headcase Promotes Cell Survival and Niche Maintenance in the Drosophila Testis. PLOS ONE. 2013;8:e68026.

144. Resende LPF, Truong ME, Gomez A, Jones DL. Intestinal stem cell ablation reveals differential requirements for survival in response to chemical challenge. Developmental Biology. 2017;424:10–7.

145. Weaver TA, White RA. headcase, an imaginal specific gene required for adult morphogenesis in Drosophila melanogaster. Development. 1995;121:4149–60.

146. Dorner S, Lum L, Kim M, Paro R, Beachy PA, Green R. A genomewide screen for components of the RNAi pathway in Drosophila cultured cells. PNAS. 2006;103:11880–5.

147. Steneberg P, Englund C, Kronhamn J, Weaver TA, Samakovlis C. Translational readthrough in the hdc mRNA generates a novel branching inhibitor in theDrosophila trachea. Genes Dev. 1998;12:956–67.

148. Cardone RA, Casavola V, Reshkin SJ. The role of disturbed pH dynamics and the Na+/H+ exchanger in metastasis. Nat Rev Cancer. 2005;5:786–95.

149. Amith SR, Fliegel L. Na+/H+ exchanger-mediated hydrogen ion extrusion as a carcinogenic signal in triple-negative breast cancer etiopathogenesis and prospects for its inhibition in therapeutics. Seminars in Cancer Biology. 2017;43:35–41.

150. Stock C, Cardone RA, Busco G, Krähling H, Schwab A, Reshkin SJ. Protons extruded by NHE1: Digestive or glue? European Journal of Cell Biology. 2008;87:591–9.

151. Škrtić M, Sriskanthadevan S, Jhas B, Gebbia M, Wang X, Wang Z, et al. Inhibition of Mitochondrial Translation as a Therapeutic Strategy for Human Acute Myeloid Leukemia. Cancer Cell. 2011;20:674–88.

152. Weinberg SE, Chandel NS. Targeting mitochondria metabolism for cancer therapy. Nat Chem Biol. 2015;11:9–15.

153. Liu L, Kimball S, Liu H, Holowatyj A, Yang Z-Q, Liu L, et al. Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. Oncotarget. 2014;6:2466–82.

154. Hsu SI-H, Yang CM, Sim KG, Hentschel DM, O'Leary E, Bonventre JV. TRIP-Br: a novel family of PHD zinc finger- and bromodomain-interacting proteins that regulate the transcriptional activity of E2F-1/DP-1. The EMBO Journal. 2001;20:2273.

155. Cheong JK, Gunaratnam L, Zang ZJ, Yang CM, Sun X, Nasr SL, et al. TRIP-Br2 promotes oncogenesis in nude mice and is frequently overexpressed in multiple human tumors. Journal of Translational Medicine. 2009;7:8.

156. García-Pedrero JM, Kiskinis E, Parker MG, Belandia B. The SWI/SNF Chromatin Remodeling Subunit BAF57 Is a Critical Regulator of Estrogen Receptor Function in Breast Cancer Cells. J Biol Chem. 2006;281:22656–64.

157. Sethuraman A, Brown M, Seagroves TN, Wu Z-H, Pfeffer LM, Fan M. SMARCE1 regulates metastatic potential of breast cancer cells through the HIF1A/PTK2 pathway. Breast Cancer Research. 2016;18:81.

158. Dhasarathy A, Kajita M, Wade PA. The Transcription Factor Snail Mediates Epithelial to Mesenchymal Transitions by Repression of Estrogen Receptor Alpha. Mol Endocrinol. 2007;21:2907–18.

159. Lacroix M, Leclercq G. Relevance of Breast Cancer Cell Lines as Models for Breast Tumours: An Update. Breast Cancer Res Treat. 2004;83:249–89.

160. Peluso S, Douglas A, Hill A, Angelis CD, Moore BL, Grimes G, et al. Fibroblast growth factors (FGFs) prime the limb specific Shh enhancer for chromatin changes that balance histone acetylation mediated by E26 transformation-specific (ETS) factors. eLife. 2017;6:e28590.

161. Liu J, Jiang G, Liu S, Liu Z, Pan H, Yao R, et al. Lentivirus-delivered short hairpin RNA targeting SNAIL inhibits HepG2 cell growth. Oncology Reports. 2013;30:1483–7.

162. Nelson DO, Lalit PA, Biermann M, Markandeya YS, Capes DL, Addesso L, et al. Irx4 Marks a Multipotent, Ventricular-Specific Progenitor Cell. Stem Cells. 2016;34:2875–88.

163. Xu X, Hussain WM, Vijai J, Offit K, Rubin MA, Demichelis F, et al. Variants at IRX4 as prostate cancer expression quantitative trait loci. Eur J Hum Genet. 2014;22:558–63.

164. Cullen PJ. Endosomal sorting and signalling: an emerging role for sorting nexins. Nat Rev Mol Cell Biol. 2008;9:574–82.

165. Marat AL, Haucke V. Phosphatidylinositol 3-phosphates—at the interface between cell signalling and membrane traffic. The EMBO Journal. 2016;35:561–79.

166. Zhu L, Hu Z, Liu J, Gao J, Lin B. Gene expression profile analysis identifies metastasis and chemoresistance-associated genes in epithelial ovarian carcinoma cells. Med Oncol. 2015;32:426.

167. Oral O, Oz-Arslan D, Itah Z, Naghavi A, Deveci R, Karacali S, et al. Cleavage of Atg3 protein by caspase-8 regulates autophagy during receptor-activated cell death. Apoptosis. 2012;17:810–20.

168. Radoshevich L, Murrow L, Chen N, Fernandez E, Roy S, Fung C, et al. ATG12 conjugation to ATG3 regulates mitochondrial homeostasis and cell death. Cell. 2010;142:590–600.

169. Li J, Yang B, Zhou Q, Wu Y, Shang D, Guo Y, et al. Autophagy promotes hepatocellular carcinoma cell invasion through activation of epithelial–mesenchymal transition. Carcinogenesis. 2013;34:1343–51.

170. Mohler J, Weiss N, Murli S, Mohammadi S, Vani K, Vasilakis G, et al. The embryonically active gene, unkempt, of Drosophila encodes a Cys3His finger protein. Genetics. 1992;131:377–88.

171. Murn J, Zarnack K, Yang YJ, Durak O, Murphy EA, Cheloufi S, et al. Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes Dev. 2015;29:501–12.

172. Murn J, Teplova M, Zarnack K, Shi Y, Patel DJ. Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nat Struct Mol Biol. 2016;23:16–23.

173. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, et al. ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Research. 2009;37:D868–72.

174. Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977;40:35–41.

175. Nusse R, Clevers H. Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell. 2017;169:985–99.

176. Schubert M, Holland LZ. The Wnt Gene Family and the Evolutionary Conservation of Wnt Expression [Internet]. Landes Bioscience; 2013 [cited 2017 Dec 6]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6212/

177. Mbodj A, Gustafson EH, Ciglar L, Junion G, Gonzalez A, Girardot C, et al. Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes. PLoS Comput Biol. 2016;12:e1005073–e1005073.

178. DiMeo TA, Anderson K, Phadke P, Feng C, Perou CM, Naber S, et al. A Novel Lung Metastasis Signature Links Wnt Signaling with Cancer Cell Self-Renewal and Epithelial-Mesenchymal Transition in Basal-like Breast Cancer. Cancer Res. 2009;69:5364–73.

179. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346.

180. Jones RA, Robinson TJ, Liu JC, Shrestha M, Voisin V, Ju Y, et al. RB1 deficiency in triple-negative breast cancer induces mitochondrial protein translation. J Clin Invest. 126:3739–57.

181. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole genome sequences. Nature. 2016;534:47–54.

182. Shih I-M, Wang T-L. Notch Signaling, γ-Secretase Inhibitors, and Cancer Therapy. Cancer Res. 2007;67:1879–82.

183. Messersmith WA, Shapiro GI, Cleary JM, Jimeno A, Dasari A, Huang B, et al. A Phase I, Dose-Finding Study in Patients with Advanced Solid Malignancies of the Oral γ-Secretase Inhibitor PF-03084014. Clin Cancer Res. 2015;21:60–7.

184. Takebe N, Miele L, Harris PJ, Jeong W, Bando H, Kahn M, et al. Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. Nature Reviews Clinical Oncology. 2015;12:445–64.

185. Sokol ES, Feng Y-X, Jin DX, Tizabi MD, Miller DH, Cohen MA, et al. SMARCE1 is required for the invasive progression of in situ cancers. PNAS. 2017;201703931.

186. Mohd-Sarip A, Teeuwssen M, Bot AG, De Herdt MJ, Willems SM, Baatenburg de Jong RJ, et al. DOC1-Dependent Recruitment of NURD Reveals Antagonism with SWI/SNF during Epithelial-Mesenchymal Transition in Oral Cancer Cells. Cell Reports. 2017;20:61–75.

187. Sparmann A, Lohuizen M van. Polycomb silencers control cell fate, development and cancer. Nature Reviews Cancer. 2006;6:846–56.

188. Koppens M, Lohuizen M van. Context-dependent actions of Polycomb repressors in cancer. Oncogene. 2016;35:1341–52.

189. Herranz N, Pasini D, Díaz VM, Francí C, Gutierrez A, Dave N, et al. Polycomb Complex 2 Is Required for E-cadherin Repression by the Snail1 Transcription Factor. Molecular and Cellular Biology. 2008;28:4772–81.

190. Yang M-H, Hsu DS-S, Wang H-W, Wang H-J, Lan H-Y, Yang W-H, et al. Bmi1 is essential in Twist1-induced epithelial–mesenchymal transition. Nature Cell Biology. 2010;12:982–92.

191. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial–mesenchymal transition. Nat Rev Mol Cell Biol. 2014;15:178–96.

192. Hemberger M, Dean W, Reik W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nature Reviews Molecular Cell Biology. 2009;10:526–37.

193. Wangler MF, Hu Y, Shulman JM. Drosophila and genome-wide association studies: a review and resource for the functional dissection of human complex traits. Disease Models & Mechanisms. 2017;10:77–88.

194. Mohr SE, Hu Y, Kim K, Housden BE, Perrimon N. Resources for Functional Genomics Studies in Drosophila melanogaster. Genetics. 2014;genetics.113.154344.

195. Mani SA, Guo W, Liao M-J, Eaton EN, Ayyanan A, Zhou AY, et al. The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. Cell. 2008;133:704–15.

196. Schmidt JM, Panzilius E, Bartsch HS, Irmler M, Beckers J, Kari V, et al. Stem-Cell-like Properties and Epithelial Plasticity Arise as Stable Traits after Transient Twist1 Activation. Cell Reports. 2015;10:131–9.

197. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015;526:131–5.

198. Guo W, Keckesova Z, Donaher JL, Shibue T, Tischler V, Reinhardt F, et al. Slug and Sox9 Cooperatively Determine the Mammary Stem Cell State. Cell. 2012;148:1015–28.

199. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015;5:2929–43.

200. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Molecular Oncology. 2011;5:5–23.

201. Marcato P, Dean CA, Pan D, Araslanova R, Gillis M, Joshi M, et al. Aldehyde Dehydrogenase Activity of Breast Cancer Stem Cells Is Primarily Due To Isoform ALDH1A3 and Its Expression Is Predictive of Metastasis. STEM CELLS. 2011;29:32–45.

202. Raha D, Wilson TR, Peng J, Peterson D, Yue P, Evangelista M, et al. The Cancer Stem Cell Marker Aldehyde Dehydrogenase Is Required to Maintain a Drug-Tolerant Tumor Cell Subpopulation. Cancer Res. 2014;74:3579–90.

203. Sieuwerts AM, Kraan J, Bolt J, van der Spoel P, Elstrodt F, Schutte M, et al. Anti-Epithelial Cell Adhesion Molecule Antibodies and the Detection of Circulating Normal-Like Breast Tumor Cells. J Natl Cancer Inst. 2009;101:61–6.

204. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, et al. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. Cell. 2012;148:473–86.

205. Factor DC, Corradin O, Zentner GE, Saiakhova A, Song L, Chenoweth JG, et al. Epigenomic Comparison Reveals Activation of "Seed" Enhancers during Transition from Naive to Primed Pluripotency. Cell Stem Cell. 2014;14:854–63.

206. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science. 2014;343:84–7.

207. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, Wit E de, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics. 2006;38:1348.

208. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.

209. Käll L, Storey JD, Noble WS. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics. 2008;24:i42–8.

210. Blakeley P, Overton IM, Hubbard SJ. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. J Proteome Res. 2012;11:5221–34.

211. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods. 2007;4:207–14.

212. Benjamini Y. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2001;29:1165–88.

213. Quackenbush J. Microarray data normalization and transformation. Nature Genetics. 2002;32:496–501.

214. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012;7:562–78.

215. Vitali F, Li Q, Schissler AG, Berghout J, Kenost C, Lussier YA. Developing a ‘personalome’ for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. Brief Bioinform. 2017;bbx149.

216. Baliga NS, Björkegren JLM, Boeke JD, Boutros M, Crawford NPS, Dudley AM, et al. The State of Systems Genetics in 2017. Cell Systems. 2017;4:7–15.

217. Chen RA-J, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J. Extreme HOT regions are CpG-dense promoters in C. elegans and humans. Genome Res. 2014;24:1138–46.

218. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512:453.

219. Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell. 2016;167:1170–87.

220. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.

221. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37:D412-416.

222. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38:D355-360.

223. van Rijsbergen CJ. Information Retrieval. 1979; Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.2325

224. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J Roy Stat Soc Ser B. 1977;39:1–38.

225. Yamada T, Bork P. Evolution of biomolecular networks — lessons from metabolic and protein interactions. Nature Reviews Molecular Cell Biology. 2009;10:791–803.

226. Fitzgibbon M, Li Q, McIntosh M. Modes of Inference for Evaluating the Confidence of Peptide Identifications. J Proteome Res. 2008;7:35–9.

227. Sennels L, Bukowski-Wills J-C, Rappsilber J. Improved results in proteomics by use of local and peptide-class specific false discovery rates. BMC Bioinformatics. 2009;10:179.

228. Raj A, van Oudenaarden A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell. 2008;135:216–26.

229. Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression underlies incomplete penetrance. Nature. 2010;463:913–8.

230. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12:323–34.

231. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association. 2001;96:1151–60.

232. López-Schier H, St Johnston D. Delta signaling from the germ line controls the proliferation and differentiation of the somatic follicle cells during Drosophila oogenesis. Genes Dev. 2001;15:1393–405.

233. Schmitt A, Nebreda AR. Signalling pathways in oocyte meiotic maturation. J Cell Sci. 2002;115:2457–9.

234. Acharya U, Patel S, Koundakjian E, Nagashima K, Han X, Acharya JK. Modulating Sphingolipid Biosynthetic Pathway Rescues Photoreceptor Degeneration. Science. 2003;299:1740–3.

235. Dasgupta U, Bamba T, Chiantia S, Karim P, Tayoun ANA, Yonamine I, et al. Ceramide kinase regulates phospholipase C and phosphatidylinositol 4, 5, bisphosphate in phototransduction. PNAS. 2009;106:20063–8.

236. Yonamine I, Bamba T, Nirala NK, Jesmin N, Kosakowska-Cholody T, Nagashima K, et al. Sphingosine kinases and their metabolites modulate endolysosomal trafficking in photoreceptors. J Cell Biol. 2011;192:557–67.

237. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc Ser B. 1995;57:289–300.

238. Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M. Whole-Genome Analysis of Dorsal-Ventral Patterning in the Drosophila Embryo. Cell. 2002;111:687–701.

239. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33:e175–e175.

240. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31:e15–e15.

241. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.

242. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. BMC Medical Genomics. 2008;1:42.

243. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. PNAS. 1998;95:14863–8.

244. Essafi A, Webb A, Berry RL, Slight J, Burn SF, Spraggon L, et al. A Wt1-Controlled Chromatin Switching Mechanism Underpins Tissue-Specific Wnt4 Activation and Repression. Developmental Cell. 2011;21:559–74.

245. Cano A, Pérez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, Barrio MG del, et al. The transcription factor Snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. Nature Cell Biology. 2000;2:76–83.

246. Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-Mesenchymal Transitions in Development and Disease. Cell. 2009;139:871–90.

9 Figure legends

Figure 1 Overview of the NetNC algorithm. NetNC input data may be a list of candidate TF target genes and a reference network such as a functional gene network (top, left). However, NetNC may be applied to analyse any gene list, for example derived from CRISPR-Cas9 screens or differential expression analysis. Hypergeometric Mutual Clustering (HMC) p-values are calculated for candidate TF target genes (top, middle); the node numbers and colours in the HMC graph correspond directly to those given in the contingency table cells. HMC p-values are then employed in either i) a node-centric analysis mode (NetNC-FBT) with Gaussian Mixture Modelling (right top) or ii) an edge-centric mode (NetNC-FTI) that involves empirical estimation of global False Discovery Rate (pFDR, middle) followed by iterative minimum cut with a graph density stopping criterion (bottom). We also developed an approach to calculate local FDR (lcFDR) in order to predict the proportion of neutrally bound candidate target genes for the TF_ALL datasets (left). NetNC-FTI takes thresholds for pFDR and graph density from calibration against synthetic data based on KEGG pathways. NetNC-FBT is parameter-free and therefore offers flexibility for analysis of datasets with network properties that may differ to the synthetic data used for calibration. NetNC can produce pathway-like clusters and also biologically coherent node lists for which edges may be taken using a standard FDR or Family Wise Error Rate (FWER) threshold on the HMC p-values (right).

Figure 2 Evaluation of NetNC and HC-PIN on blind test data. Performance values reflect discrimination of KEGG pathway nodes from Synthetic Neutral Target Genes (STNGs), shown for NetNC-FTI (orange), NetNC-FBT (red) and HC-PIN (green). False Positive Rate (FPR, top row) and Matthews Correlation Coefficient (MCC, bottom row) values are given. The data shown represents analysis of TEST-CL_ALL, which included subsets of three to eight pathways, shown in columns, and sixteen %STNG values were analysed (5% to 80%, x-axis). NetNC performed best on the data examined with typically lower FPR and higher MCC values. Error bars reflect 95% confidence intervals calculated from quantiles of the SNTG resamples (per datapoint: n=100 for NetNC, n=99 for HC-PIN). The NetNC-FBT analysis mode was the most stringent and had lowest FPR across the datasets examined - but also had lower MCC, particularly on the three or four pathway datasets. In general, MCC for NetNC and HC-PIN rose with increasing SNTG percentage, up to around 40%. HC-PIN performance declined at SNTG values >40% wheras NetNC performance remained high. At the highest %SNTG, MCC values for NetNC-FTI were around 50% to 67% higher than those for HC-PIN. The performance advantage for NetNC was also apparent upon inspection of the HC-PIN FPR profiles which rose to around 0.4 at 80% SNTGs; HC-PIN typically had significantly higher FPR than NetNC. There was a trend towards worse overall performance for all methods as the number of pathways in the dataset (and hence dataset size) increased. Indeed, NetNC-FTI maximal MCC values were respectively around 0.7, 0.55 for the three, eight pathway datasets. Performance advantages for NetNC were particularly apparent on data with ≥50% SNTGs.

Figure 3 Neutral transcription factor binding and false discovery rate (FDR) profiles.

Panel A: Estimation of total neutral binding. Black circles show NetNC mean lcFDR values for the TEST-CL_8PW data, ranging from 5% to 80% SNTGs; error bars represent 95% CI calculated from quantiles of the SNTG resamples (n=100 per datapoint). Coloured horizontal lines show mean NetNC-lcFDR values for the TF_ALL datasets. Comparison of the known TEST-CL_8PW %SNTG values with estimated total neutral binding values from mean NetNC-lcFDR showed systematic overestimation of neutral binding. Cross-referencing mean NetNC-lcFDR values for TF_ALL with those for TEST-CL_8PW gave estimates of neutral binding between 50% and ≥80% (see panel key).

Panels B (Local FDR profiles), C (Global FDR profiles) and D (Global FDR zoom). Line type and colour indicates dataset identity (see key). Candidate target gene index values were normalised from zero to one in order to enable comparison across the TF_All datasets. **Panel B:** Profiles of lcFDR are shown. Although sna_2-3h_union and twi_2-3h_union had high mean lcFDR (panel A, above), they also had the highest proportion and largest numbers of genes with lcFDR<0.05. **Panel C:** Profiles of global FDR (pFDR) are shown. Profiles of pFDR and lcFDR were similar. For example, sna_2-3h_union and twi_2-3h_union both had relatively high proportion of genes passing lcFDR<0.05 and pFDR<0.05. However differences were observed, for example twi_2-6h_intersect had the greatest proportion of genes passing pFDR threshold values between 0.01 and 0.2, in contrast to equivalent lcFDR values (panel B) where no single dataset dominates. pFDR was smoother than lcFDR (panel B) because of the procedure to prevent inconsistent lcFDR scaling (equation (6)).

Panel D: pFDR values visualised around commonly applied threshold values, including those selected in NetNC parameter optimisation. Interestingly, the high-confidence dataset twi_1-3h_hiConf, which had the lowest predicted overall proportion of neutral binding (panel A), also had proportionally very few genes passing a threshold of pFDR<0.05.

Figure 4 NetNC-FTI functional target networks for Snail and Twist. The key (bottom right) indicates annotations for human orthology (bold node border) and *Notch* screen hits (triangular nodes). Many orthologues were assigned to either basal-like (BL, red) or normal-like centroids (NL, green); otherwise, node colour indicates upregulated gene expression in NL (blue) compared to BL (orange) subtypes (q<0.05) or no annotation (grey). Clusters with at least four members are shown; cytoscape sessions with full NetNC-FTI results are given in Additional File 4. In general, NetNC-FTI clusters formed recognised groupings of gene function, including previously characterised protein complexes.

Panel A: twi_2-3h_union. Predicted functional targets cover several areas of fundamental biochemistry including splicing, DNA replication, energy metabolism, translation and chromatin organisation. Regulation of multiple conserved processes by Twist is consistent with the extensive cell changes required during mesoderm development. Clusters annotated predominantly to either NL or BL subtypes include mitochondrial translation (BL) and the proteosome (NL). These results predict novel functions for Twist, for example in regulation of mushroom body neuroblast proliferation factors.

Panel B: sna_2-4h_Toll^{10b}. Multiple clusters of transcription factors were identified, aligning with previous studies that identified Snail as a master transcriptional regulator [39,245,246]. These clusters included the *achaete-scute* complex (bottom right) and polycomb group members (bottom left). Direct targeting of *achaete-scute* by Snail in prospective mesoderm is consistent with repression of neurectodermal fates [97–99]. Orthologues in the clusters ‘RNA degradation, transcriptional regulation’; ‘axis specification’ and ‘phosphatases’ were only annotated to the basal-like subtype.

Panel C: twi_2-6h_intersect. A large proportion of predicted functional targets for twi_2-6h_intersect belonged to the ‘developmental regulation’ NetNC-FTI cluster; regulatory factors may be enriched in this dataset due to the criterion for continuous binding across two developmental time windows. The developmental regulation cluster contained *mrr,* the orthologue of *IRX4*, which was BL upregulated (orange) and was investigated in follow-up experiments (Figure 6).

Figure 5 Predicted functional transcription factor targets capture human breast cancer biology. The heatmap shows results of unsupervised clustering with gene expression data for 2999 primary breast tumours and 57 orthologues of NetNC-FTI functional targets that were identified in at least four of nine TF_ALL datasets (ORTHO-57). Expression values were log2 transformed and mean-centred to give relative values across tumours (red=high, white=mean, blue=low). Intrinsic molecular subtype for each tumour is shown by the mosaic above the heatmap and below the dendogram, from left to right : luminal A (blue), basal-like (red), HER2-overexpressing (purple), luminal B (light blue) and normal-like (green). Source data identifiers are given to the right of the subtype mosaic. Features annotated onto the heatmap as black dashed lines identified genes upregulated in one or more intrinsic subtype; these features were termed 'Bas' (basal-like), 'NL'(normal-like), 'ERneg' (basal-like and HER2-overexpressing), 'LumB$_1$'(luminal B), 'LumB$_2$'(luminal B), 'LumA' (luminal A) and 'LoExp' (low expression). The table to the right of the heatmap indicates inclusion (grey) or absence (white) of genes in NetNC-FTI results across the TF_ALL datasets. The column '#D' gives the number of TF_ALL datasets where the gene was returned by NetNC-FTI and '%P' column details the percentage of present calls for gene expression across the 2999 tumours. The LoExp feature corresponded overwhelmingly to genes with low %P values and to samples from a single dataset [133]. Some genes were annotated to more than one feature and reciprocal patterns of gene expression were found. For example, *BMPR1B*, *ERBB3* and *MYO6* were strongly upregulated in feature LumA but downregulated in basal-like and *HER2*-overexpressing cancers. Unexpectedly, feature NL (normal-like) had high expression of canonical EMT drivers, including *SNAI2*, *TWIST* and *QKI*. Some of the EMT genes in feature NL were also highly expressed in many basal-like tumours, while genes in feature Bas (*NOTCH*, *SERTAD2*) were upregulated in normal-like tumours.

Figure 6 Validation of candidate invasion genes in breast cancer cells. The fluorescence CyQuant dye signal from invading MCF7 cells is shown (RFU) for the transwell assay. Induction of each of the four genes examined significantly changed MCF7 invasion when compared to controls (orange) in least one of three conditions: a) ectopic expression; b) ectopic expression and *SNAI1* induction; c) ectopic expression with shRNA knockdown of *SNAI1*. The orthologous genes studied were: *SNX29* (blue), which showed a significant reduction in invasion compared with the *SNAI1* induction control; *UNK* (purple) and *IRX4* (dark red) where invasion was significantly increased all three conditions examined; *ATG3* which had significantly higher invasion at background levels of *SNAI1* (without induction or knockdown). All datapoints are n=3. Statistical significance in comparisons against the appropriate control experiment is indicated as follows: * $q<0.05$; *** $q<5.0\text{x}10^{-4}$