bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functionally Coherent Transcription Factor Target Networks Illuminate Control of Epithelial Remodelling

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Summary

Cell identity is governed by gene expression, regulated by Transcription Factor (TF) binding at cis-regulatory modules. We developed the NetNC software to decode the relationship between TF binding and the regulation of cognate target genes in cell decision-making; demonstrated on nine datasets for the Snail and Twist TFs, and also modENCODE 'HOT' regions. Results illuminated conserved molecular networks controlling development and disease, with implications for precision medicine. Predicted 'neutral' TF binding accounted for the majority (50% to ≥80%) of candidate target genes from statistically significant peaks and HOT regions had high functional coherence. Expression of orthologous functional TF targets discriminated breast cancer molecular subtypes and predicted novel tumour biology. We identified new gene functions and network modules including crosstalk with notch signalling and regulation of chromatin organisation, evidencing networks that reshape Waddington's landscape during epithelial remodelling. Predicted invasion roles were validated using a tractable cell model, supporting our computational approach.

Keywords

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Transcriptional regulatory factors (TFs) govern gene expression, which is a crucial determinant of phenotype. Mapping transcriptional regulatory networks is an attractive approach to understand the molecular mechanisms underpinning both normal biology and disease (Rhee et al., 2014; Shlyueva et al., 2014; Stampfel et al., 2015). 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 (Khoueiry et al., 2017; Rhee et al., 2014; Zabidi and Stark, 2016). A complex interplay of mechanisms influences TF specificity across different biological contexts and genome-scale assignment of TFs to individual genes is challenging (Khoueiry et al., 2017; Shlyueva et al., 2014; Wilczynski and Furlong, 2010).

TF binding sites may be determined 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 assigned target genes (Biggin, 2011; Li et al., 2008; Ozdemir et al., 2011; Shlyueva et al., 2014). Genomic regions that bind large numbers of TFs are termed Highly Occupied Target (HOT) regions (Roy et al., 2010), are enriched for disease SNPs and can function as developmental enhancers (Kvon et al., 2012; Li et al., 2015). 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 (Montavon et al., 2011; Moorman et al., 2006) as well as possible technical artefacts (Teytelman et al., 2013). Apparent neutral binding events may also have subtle functions; for example in combinatorial context-specific regulation or in buffering noise (Cannavò et al., 2016; Stampfel et al., 2015). Identification of *bona fide*, functional TF target genes remains a major obstacle in understanding the regulatory

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

networks that control cell behaviour (Biggin, 2011; Brown and Celniker, 2015; Keung et al., 2014; Khoueiry et al., 2017; Stampfel et al., 2015).

Genes regulated by an individual TF typically have overlapping expression patterns and coherent biological function (Igual et al., 1996; Karczewski et al., 2014; MacArthur et al., 2009). Indeed, gene regulatory networks are organised in a hierarchical, modular structure and TFs frequently act upon multiple nodes of a given module (Hartwell et al., 1999; Hooper et al., 2007). Therefore, we hypothesised that functional TF targets have network properties that are different to neutrally bound sites. Network analysis can reveal biologically meaningful gene modules, including cross-talk between canonical pathways (Ideker et al., 2002; Jaeger et al., 2017; Vidal et al., 2011) and so may enable elimination of neutrally bound candidate TF targets derived from statistically significant ChIP-seq or ChIP-chip peaks. Network approaches afford significant advantages for handling biological complexity, enable genome-scale analysis of gene function (Greene et al., 2015; Hu et al., 2016), and are not restricted to predefined gene groupings used by standard functional annotation tools (e.g. GSEA, DAVID) (Huang et al., 2009; Ideker et al., 2002; Subramanian et al., 2005). Clustering is frequently applied to define biological modules (Enright et al., 2002; Wang et al., 2011). However, pre-defined modules may miss condition-specific features; for example, gene products may be absent in the biological condition(s) analysed but present in the base network. Hence, clusters derived from a whole-genome network may not accurately capture biological interactions that occur in a particular context. Context-specific interactions are common, 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 (Pawson and Nash, 2003). We developed a functional TF target discovery algorithm (NetNC) with capability for context-specific module discovery. The Snail and Twist TFs 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 (Giampieri et al., 2009; Lim and Thiery, 2012; Nieto et al., 2016;

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Yu et al., 2013). We analysed Snail and Twist binding datasets, integrating with genetic screens and breast cancer transcriptomes to gain insight into epithelial remodelling in development and disease.

RESULTS AND DISCUSSION

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

We developed a systems-wide map of *D. melanogaster* signalling and metabolism (DroFN; 11,432 nodes, 787,825 edges). DroFN performed well on time-separated blind test data (TEST-NET) compared with DroID (Yu et al., 2008) and GeneMania (Warde-Farley et al., 2010) (Table S1, Figure S1). The overlap between DroFN and the *Drosophila* proteome interaction map (DPiM (Guruharsha et al., 2011)) was highly significant (Fisher Exact Test $p<10^{-308}$). 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 DroFN False Positive Rate (0.047) was close to the prior for functional interaction estimated from KEGG (0.044); thus, some DroFN false positives may represent *bona fide* interactions not annotated in KEGG. Overall, DroFN provides a useful genome-scale map of pathway comembership in *D. melanogaster*.

Prediction of functional transcription factor binding

We developed the NetNC algorithm for genome-scale prediction of functional TF target genes (Figure 1). NetNC builds upon observations that TFs coordinately regulate multiple functionally related targets (Igual et al., 1996; Karczewski et al., 2014; MacArthur et al., 2009) and was calibrated for discovery of biologically coherent genes in noisy data. An edge-centric 'Functional Target Identification' (FTI) mode seeks to distinguish all biologically coherent gene pairs from functionally unrelated targets. The node-centric 'Functional Binding Target' (FBT) mode analyses degree-normalised Node Functional Coherence Scores (NFCS). Gold-standard data for NetNC development and validation took KEGG pathways to represent biologically coherent relationships,

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

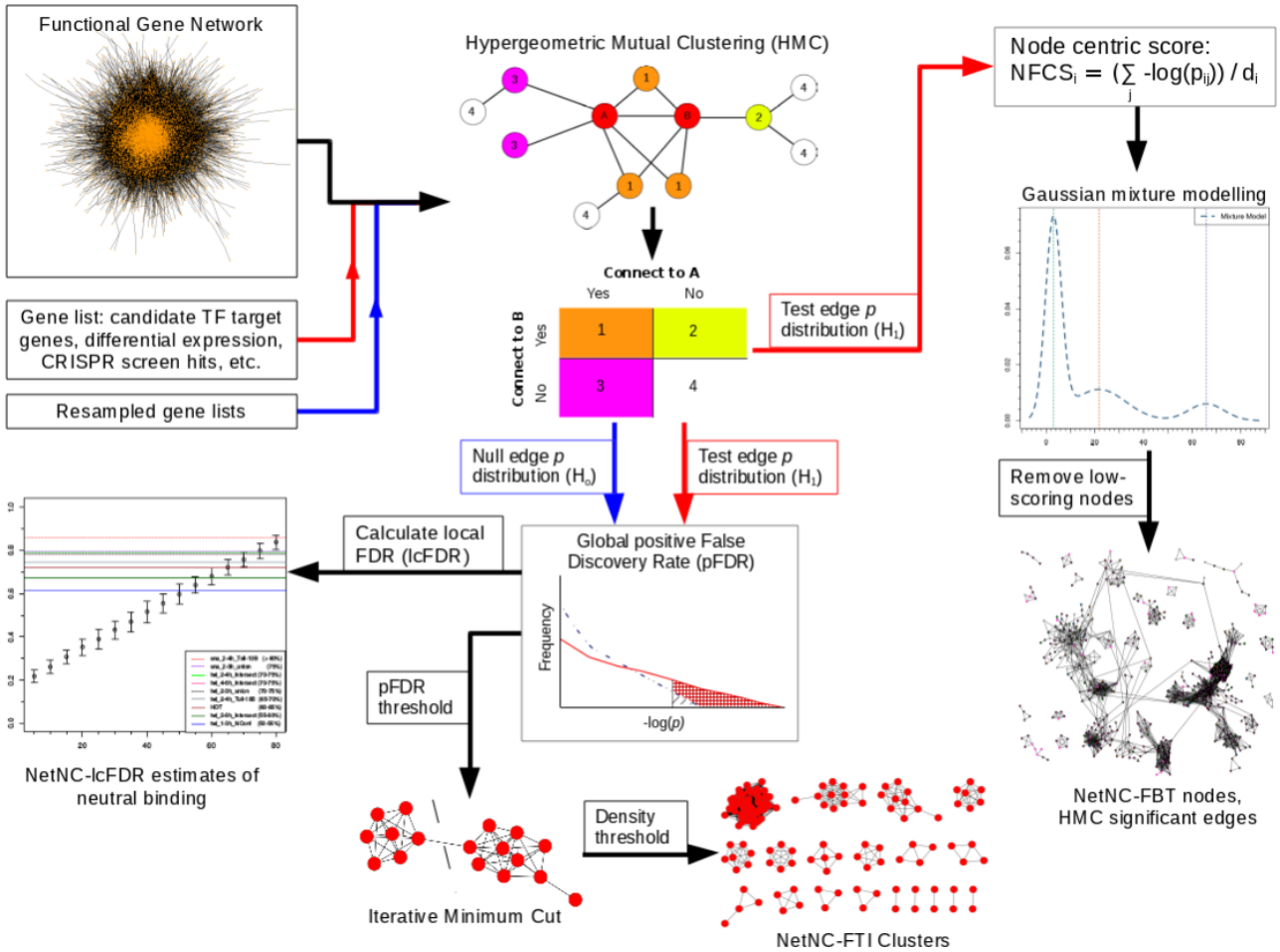


Figure 1. Overview of the NetNC algorithm. NetNC input data may be candidate TF target genes and a functional gene network (top, left). However, NetNC may be applied to any gene list and reference network. Hypergeometric Mutual Clustering (HMC) p-values are calculated for candidate TF target genes (top, middle); the node numbers and colours in the HMC graph correspond to those in the contingency table. HMC p-values are then employed in either i) a node-centric analysis mode (NetNC-FBT, right top) or ii) an edge-centric mode (NetNC-FTI) using global False Discovery Rate (pFDR, middle) followed by iterative minimum cut (bottom). NetNC can estimate local FDR (lcFDR) to predict the proportion of neutrally bound candidate target genes (left). 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).

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

combined with 'Synthetic Neutral Target Genes' (SNTGs). NetNC was robust to variation in input dataset size and %SNTGs, outperforming HC-PIN (Wang et al., 2011) and MCL (Enright et al., 2002) (Figure 2, Table S2). In general, NetNC was more stringent, with lower False Positive Rate (FPR) and higher Matthews Correlation Coefficient (MCC) than HC-PIN. At the highest %SNTG, MCC values for NetNC-FTI were around 50% to 67% higher than those for HC-PIN and NetNC-FBT typically had lowest FPR. NetNC's performance advantages were most prominent on data with ≥50% SNTGs (Figure 2) and all nine TF_ALL datasets were predicted to contain ≥50% neutrally bound targets (Figure 3). Given the performance advantage at ≥50% STNGs, NetNC appears the method of choice for analysis of genome-scale TF occupancy data. However, NetNC may be applied to multiple different tasks, including: 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 (Shalem et al., 2014); and discovery of functional coherence in chromosome conformation capture data (4-C, 5-C), for example in enhancer regulatory relationships (Dostie et al., 2006; Simonis et al., 2006).

NetNC identifies coherent genes according to the functional gene network (FGN) structure. Statistical evaluation of network coherence, including FDR estimation, is available within NetNC for numerical thresholding. Therefore, NetNC is generally applicable to high-dimensional data including analysis of single-subject datasets, which is an important emerging area for precision medicine (Vitali et al., 2017). Application of statistical and graph theoretic methods for quantitative evaluation of relationships between genes (nodes) in NetNC offers an alternative to the classical emphasis on individual genes in studying the relationship between genotype and phenotype. Gold standard datasets and DroFN are available from the Biostudies database, NetNC is available from https://github.com/IanOverton/NetNC.git.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

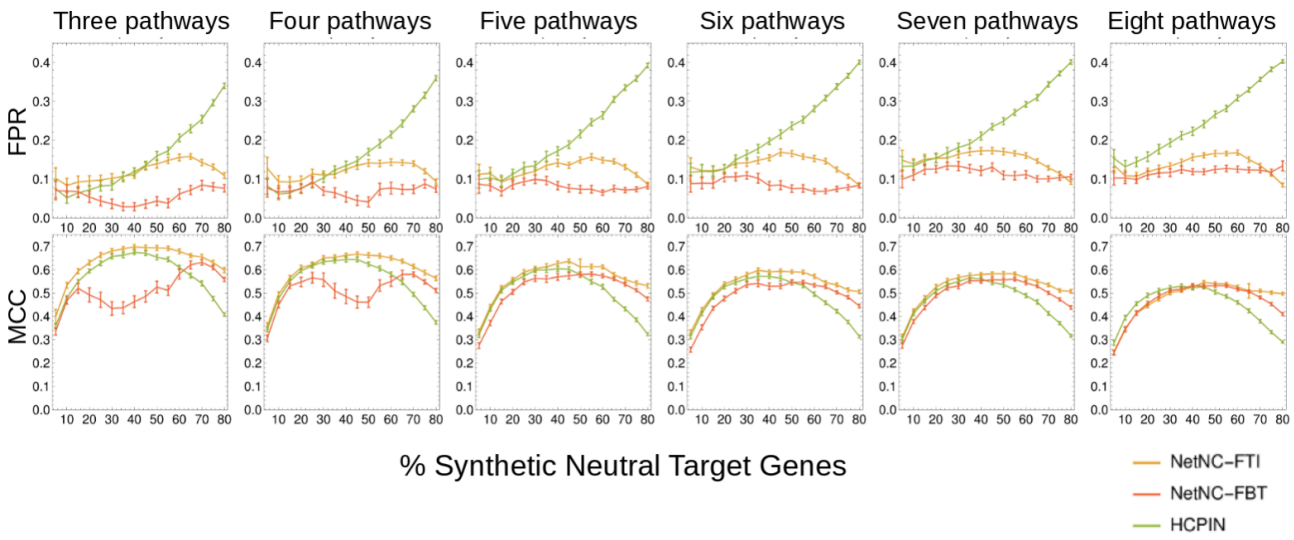


Figure 2. Evaluation of NetNC and HC-PIN on blind test data. Performance values reflect discrimination of KEGG pathway nodes from Synthetic Neutral Target Genes (STNGs), shown for NetNC-FTI (orange), NetNC-FBT (red) and HC-PIN (green). False Positive Rate (FPR, top row) and Matthews Correlation Coefficient (MCC, bottom row) values are given. 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 SNTG resamples. Also see Table S2.

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*; analysing 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 modENCODE Highly Occupied Target (HOT) regions (MacArthur et al., 2009; Ozdemir et al., 2011; Roy et al., 2010; Sandmann et al., 2007; Zeitlinger et al., 2007). Nine datasets were studied (TF_ALL, Table 1; please see Methods for details). NetNC-lcFDR predicted neutral binding proportion (PNBP) across TF_ALL ranged from 50% to ≥80% (Figure 3A, Table 1). Reassuringly, the most stringent peak calling approach (twi_1-

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3h_hiConf (Ozdemir et al., 2011) had the lowest (NetNC-lcFDR) or second lowest (NetNC-FTI) PNBP. Despite having lowest overall PNBP, twi_1-3h_hiConf also had the smallest proportion of genes passing pFDR<0.05 (Figure 3D). Targets bound during two consecutive developmental time periods (twi_2-6h_intersect (Sandmann et al., 2007)) also ranked highly, and so did HOT regions despite depletion for known TF motifs (Boyle et al., 2014; Chen et al., 2014; Kvon et al., 2012) (Figure 3A, Table 1). Indeed, twi_2-6h_intersect had significantly greater PNBP (binomial $p<4.0\text{x}10^{-15}$) than datasets from the same study that represented a single time period (twi_2-4h_intersect, twi_4-6h_intersect) (Sandmann et al., 2007) (Figure 3). Therefore, predicted functional binding was enriched for regions occupied at >1 time period or by multiple TFs and results supported the emerging picture of widespread combinatorial control involving TF-TF interactions, cooperativity and TF redundancy (Jolma et al., 2015; Khoueiry et al., 2017; Long et al., 2016; Spitz and Furlong, 2012; Stampfel et al., 2015).

PNBP was similar for sites derived from either the union or intersection of two Twist antibodies, although the NetNC-FTI method found a higher number of functional targets for the intersection of antibodies (30.5% (116/334) vs 23% (424/1848)). Hits identified by multiple antibodies may be technically more robust due to reduced off-target binding (Sandmann et al., 2007). However, taking the union of candidate binding sites could eliminate false negatives arising from epitope steric occlusion due to protein interactions. Similar PNBP for either the intersection or the union of Twist antibodies suggests that, despite expected higher technical specificity, the intersection of candidate targets may not enrich for functional binding sites at the 1% peak-calling FDR threshold applied in (MacArthur et al., 2009; Sandmann et al., 2007). Fewer false negatives implies recovery of numerically more functional TF targets, likely producing denser clusters in DroFN, which could facilitate functional target detection by NetNC. Indeed, datasets representing the union of two antibodies ranked highly in terms of both the total number and proportion of genes

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Table 1. Predicted Functional binding for Snail, Twist and HOT candidate target genes. 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 (Sandmann et al., 2007)); 2-3h stages 4-6; 1-3h stages 2-6; 4-6h stages 8-9 (Sandmann et al., 2007); 0-12h stages 1-15; gastrulation occurs at stage 6 (Campos-Ortega and Hartenstein, 1997). Also see Figure S2.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

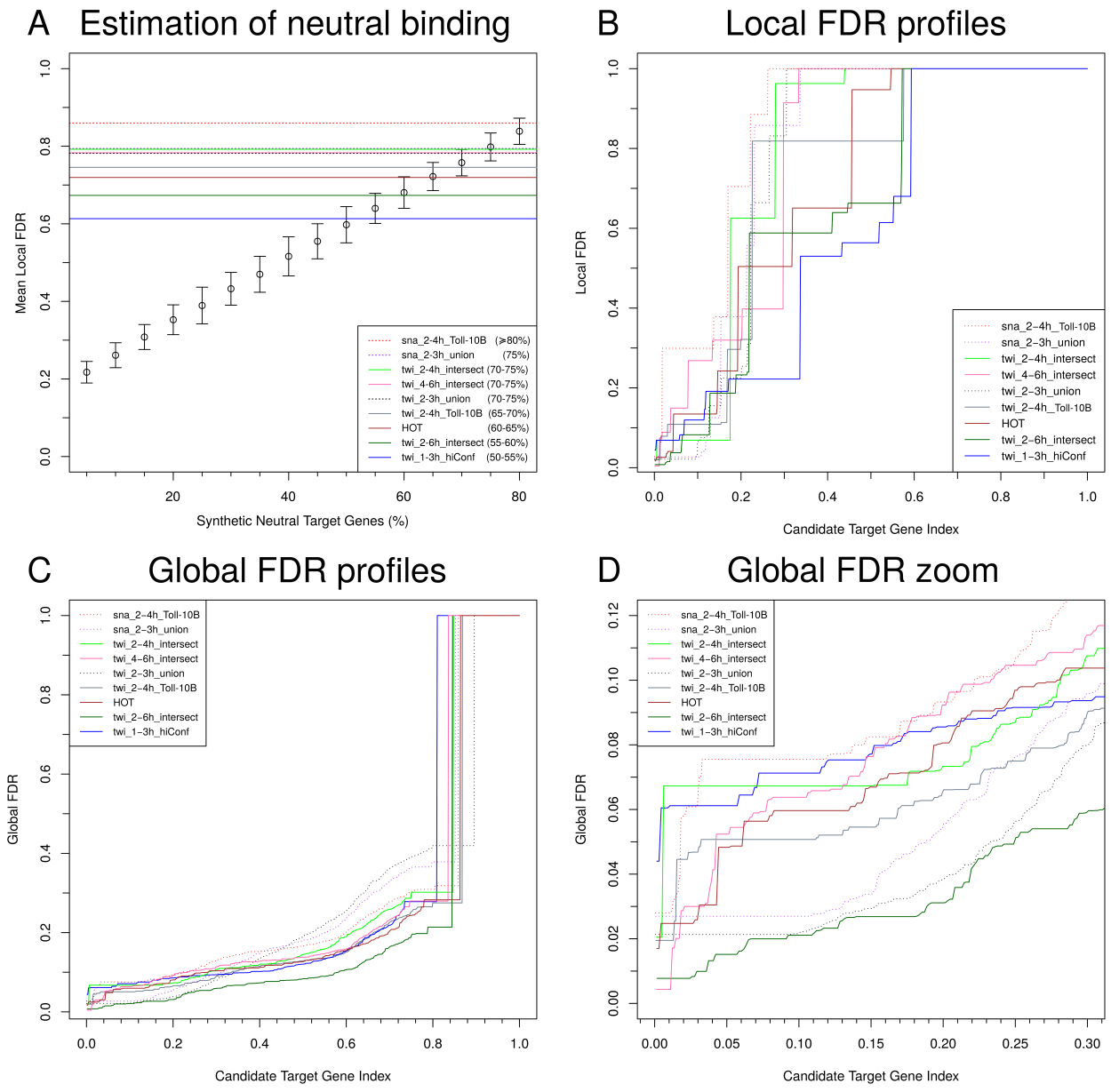


Figure 3. Neutral transcription factor binding and false discovery rate (FDR) profiles.

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. 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, C and D. 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.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recovered at lcFDR<0.05 or pFDR<0.05 (Figure 3). Even datasets with high PNBP had candidate target genes that passed stringent NetNC FDR thresholds; for example, sna_2-3h_union, twi_2-3h_union respectively had the highest and second-highest proportion of candidate targets at lcFDR<0.05 (Figure 3B). We found no evidence for benefit in using RNA polymerase binding data to guide allocation of peaks to candidate target genes (datasets sna_2-3h_union, twi_2-3h_union). The twi_2-4h_Toll^{10b}, sna_2-4h_Toll^{10b} datasets had a relatively low peak threshold (two-fold enrichment), which may contribute to the high PNBP for sna_2-4h_Toll^{10b}. We note that PNBP values might systematically overestimate neutral binding because some functional targets could be missed; for example due to errors in assigning enhancer binding to target genes and in *bona fide* regulation of genes that have few DroFN edges with other candidate targets. Additionally, calibration of lcFDR values against KEGG synthetic data might influence neutral binding estimates, due to potential differences in network properties between TF targets and KEGG pathways. Reasurringly, NetNC-lcFDR and NetNC-FTI neutral binding estimates showed good agreement (Table 1, Figure S2).

ChIP peak intensity putatively correlates with functional binding, although some weak binding sites are functional (Biggin, 2011; Chen et al., 2013). NetNC NFCS values and ChIP peak enrichment scores were significantly correlated in 6/8 datasets (q<0.05, HOT regions not analysed). Datasets with no significant correlation (twi_1-3h_hiConf, twi_2-6h_intersect) derived from protocols that enrich for functional targets and had lowest PNBP values (Figure 3A). Indeed, the median peak score for twi_2-6h_intersect was significantly higher than datasets taken from a single time period in same study (twi_2-4h_intersect, $q<5.0\text{x}10^{-56}$; twi_4-6h_intersect, $q<4.8\text{x}10^{-58}$). Predicted targets were enriched for human orthologues, for example 72% (453/628) of twi_2-3h_union NetNC-FBT functional targets had human orthology, significantly higher than the full twi_2-3h_union dataset (50%, 616/1220; $p<3\text{x}10^{-28}$ binomial test). Enrichment for evolutionary

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conservation in Snail and Twist functional targets aligns with their regulation of fundamental developmental processes.

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, Figure S3, Supplemental File 1). Networks were annotated using GO and FlyBase (Ashburner et al., 2000; Gramates et al., 2017; Huang et al., 2009; Maere et al., 2005), revealing eleven biological groupings common to ≥4/9 TF_ALL datasets (Table S3). NetNC results were robust to subsampled input and, as expected, higher variance was found at lower subsampling rates (Table S4).

NetNC analysis found Snail and Twist regulation of multiple core cell processes that govern the global composition of the transcriptome and proteome (Figure 4, Figure S3). These processes included transcription, chromatin organisation, RNA splicing, translation and protein turnover. A 'Developmental Regulation Cluster' (DRC) was identified in every TF_ALL dataset and contained members of multiple key conserved morphogenetic pathways, including notch, wnt and fibroblast growth factor (FGF). *Notch* signalling modifiers from public data (Guruharsha et al., 2012) 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, Figure S3). *Notch* was an important control node across TF_ALL, with highest betweenness centrality in the DRC for three datasets and ranked among the top ten DRC genes for 8/9 datasets. Activation of *Notch* can result in diverse, context-specific transcriptional outputs and the mechanisms regulating this pleiotropy are not well understood (Bray, 2016; Guruharsha et al., 2012; Nowell and Radtke, 2017; Ntziachristos et al., 2014). Our results provided functional context for many *Notch* modifiers and proposed signalling crosstalk mechanisms in cell fate decisions driven by Snail and Twist, where regulation of modifiers may control the consequences of *Notch* activation. Crosstalk between *Notch* and *twist*

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

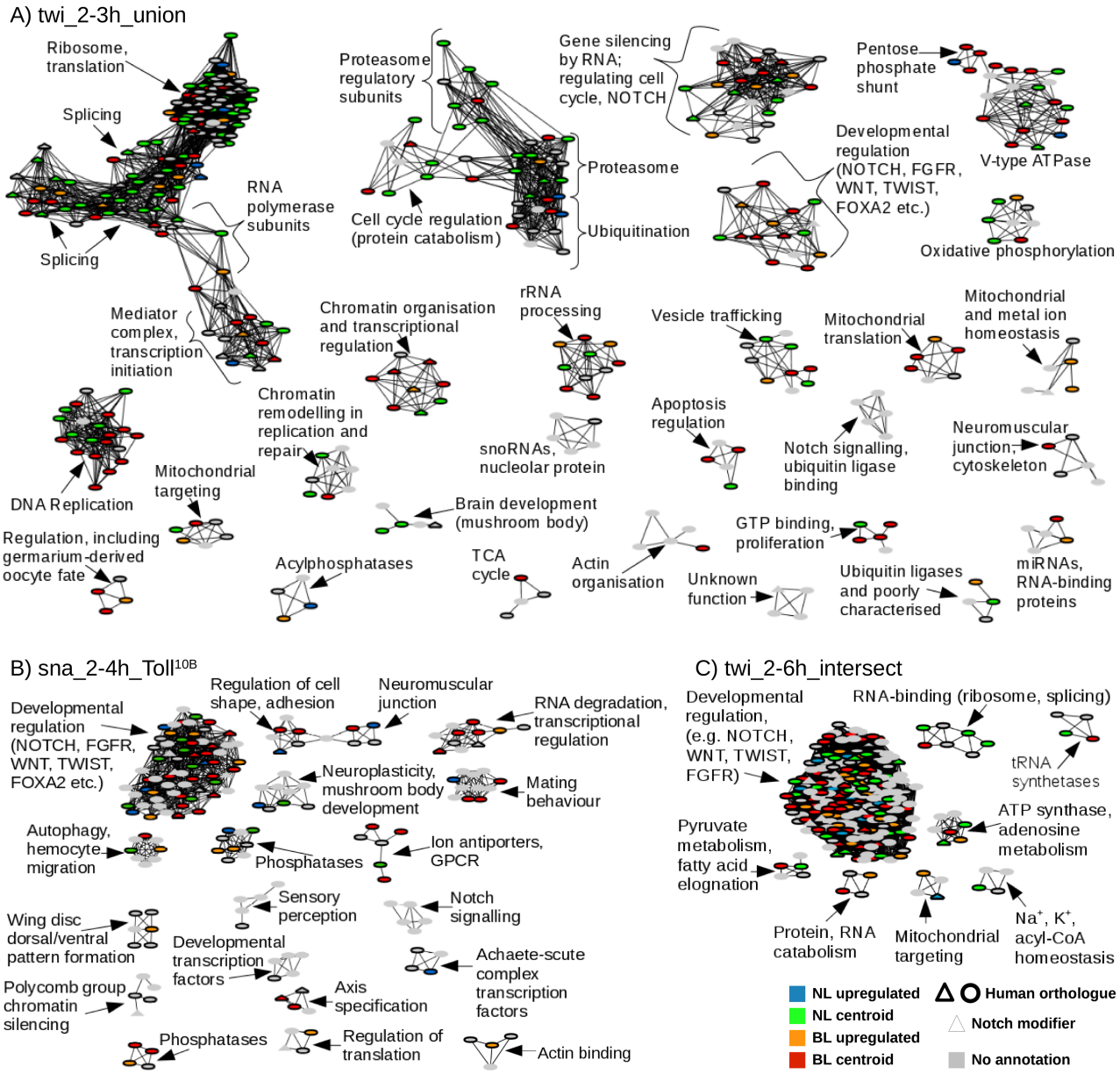


Figure 4. NetNC-FTI functional target networks for Snail and Twist. The key (bottom right) indicates annotations for human orthology (bold node border) and *Notch* screen hits (triangular nodes). Orthologues assigned basal-like (BL) or normal-like centroids (NL) are shown in red, green respectively; otherwise, node colour indicates differential expression NL (blue) vs BL (orange) subtypes (q<0.05) or no annotation (grey). Also see Figure S3, Tables S3 to S6 and Supplemental File 1. **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 proteasome (NL). These results predict novel Twist functions, for example in regulation of mushroom body neuroblast proliferation factors.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Panel B: sna_2-4h_Toll[10b]. Multiple clusters of transcription factors were identified, aligning with Snail's role as a master transcriptional regulator. Orthologues in the clusters 'RNA degradation, transcriptional regulation'; 'axis specification' and 'phosphatases' were annotated exclusively to the basal-like subtype. **Panel C: twi_2-6h_intersect.** Most predicted functional targets belonged to the 'developmental regulation' cluster; dominance of regulatory factors might arise from the criterion for continuous binding across two developmental time windows. The developmental regulation cluster contained *mrr*, the orthologue of *IRX4*, which was upregulated in BL tumours.

or *snail* was previously shown in multiple systems, for example in adult myogenic progenitors (Bernard et al., 2010) and hypoxia-induced EMT (Sahlgren et al., 2008). Consistent with previous studies (Guruharsha et al., 2012; Ntziachristos et al., 2014), our results predict targeting of *Notch* transcriptional regulators, trafficking proteins, post-translational modifiers, receptor recycling and regulation of pathways that may attenuate or modify *Notch* signalling. Clusters where multiple modifiers were identified may represent cell meso-scale units important for *Notch* in the context of mesoderm development and EMT. For example, the mediator complex and transcription initiation subcluster for twi_2-3h_union (Figure 4) had thirteen nodes, of which five were *Notch* modifiers including orthologues of *MED7*, *MED8*, *MED31*. *Wingless* also frequently had high betweenness, ranking within the top ten DRC genes in six TF_ALL datasets and was twice top-ranked. Therefore *Notch* and *wingless* were key control points regulated by Snail, Twist in the mesoderm specification network. Thirteen DRC genes were present in ≥7 TF_ALL datasets (DRC-13, Table S5), and had established functions in the development of mesodermal derivatives such as muscle, the nervous system and heart (Baylies and Bate, 1996; Bernard et al., 2010; Bray, 2016; Chen et al., 1996; Lo et al., 2002; Trujillo et al., 2016; Xie et al., 2016). Supporting NetNC-FTI predictions, *in situ* hybridisation for DRC-13 indicated 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*) (BDGP; Hammonds et al., 2013; Hartley et al., 1987; Tomancak et al., 2002). The remaining two DRC-13 genes had no evidence for mesodermal expression (*fkh*) or no data available (*jar*). However, *fkh* is

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

essential for caudal visceral mesoderm development (Kusch and Reuter, 1999) and *jar* is expressed in midgut mesoderm (Millo and Bownes, 2007).

Chromatin organisation clusters included several polycomb-group (PcG) and trithorax-group (TrxG) genes. Polycomb Repressive Complex 1 (PRC1) genes *ph-d*, *psc* (Shao et al., 1999) and *su(var)3-9* (Schotta et al., 2002) were identified most frequently in TF_ALL (Table S6). Further PcG/TrxG-related functional targets were: *ph-p* (Shao et al., 1999); c*orto* (Lopez et al., 2001); the TrxG-related gene *lolal* (Mishra et al., 2003); *taranis* (Schuster and Smith-Bolton, 2015); and TrxG genes *trithorax*, *moira* (Crosby et al., 1999; Tie et al., 2014). The gene silencing factor s*u(var)205* was also found in four TF_ALL datasets (Fanti et al., 1998). Therefore, NetNC predicted Snail and Twist regulation of: PRC1 core components and other gene silencing factors; TrxG genes; modifiers of PcG, TrxG activity. PcG genes are crucial oncofetal regulators and the focus of significant cancer drug development efforts (Koppens and Lohuizen, 2016; Sparmann and Lohuizen, 2006). Results align with previous reports that gene silencing in EMT involves PcG (Herranz et al., 2008; Koppens and Lohuizen, 2016), *Snai1* recruitment of Polycomb Repressive Complex 2 members (Herranz et al., 2008), and support a model where EMT TFs control the expression of their own coregulators.

Six TF_ALL datasets had brain development clusters. Snail regulation of neural genes is consistent with its repression of ectodermal (neural) genes in the prospective mesoderm (Gilmour et al., 2017; Leptin, 1991; Wieschaus and Nüsslein-Volhard, 2016). Additionally, Snail is important for neurogenesis in fly and mammals (Ashraf and Ip, 2001; Zander et al., 2014). Therefore, binding to neural functional modules might reflect potentiation for rapid activation in combination with other transcription factors within neural developmental trajectories (Nevil et al., 2017; Sandmann et al., 2007). NetNC results predict novel Twist functions, for example in activation or repression of mushroom body neuroblast proliferation factors *retinal homeobox*, *slender lobes*, and *taranis*. The mushroom body is a prominent structure in the fly brain, important for olfactory learning and memory (Caron et al., 2013). Twist is typically a transcriptional activator (Gilmour et al., 2017)

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

although may contribute to Snail's repressive activity (Lin et al., 2015). Indeed, Twist-related protein 1 repressed Cadherin-1 in breast cancers (Vesuna et al., 2008)

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

Integration of NetNC predictions, *Notch* screens and breast cancer transcriptomes enabled analysis of conserved molecular networks that orchestrate epithelial remodelling in development and cancers. NetNC-FTI Snail and Twist targets included known cancer genes and also predicted novel drivers (Figure 4, Figure S3, Tables S3, S5, S6). The fly genome is relatively tractable for network studies, while data availability (e.g. ChIP-chip, ChIP-seq, genetic screens) is enhanced by both considerable community resources and the relative ease of experimental manipulation (Mohr et al., 2014). Breast cancer intrinsic molecular subtypes with distinct clinical trajectories have been extensively validated and complement clinico-pathological parameters (Cejalvo et al., 2017; Sørlie et al., 2003). These subtypes are known as luminal-A, luminal-B, HER2-overexpressing, normal-like and basal-like. While more recent studies have classified more subtypes, for example identifying ten groups (Curtis et al., 2012), 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 aligned with distinct treatment pathways (Cejalvo et al., 2017; Sørlie et al., 2003). NetNC-FTI networks for all nine TF_ALL datasets overlapped with known cancer pathways, including significant enrichment for *Notch* modifiers ($q<0.05$). Aberrant activation of *Notch* orthologues in breast cancers had been demonstrated, and linked with EMT-like signalling, particularly for basal-like and claudin-low subtypes (Barnawi et al., 2016; Ingthorsson et al., 2016; Stylianou et al., 2006). We hypothesised that orthologous genes from Snail and Twist functional targets would stratify breast cancers into clinically meaningful groups.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unsupervised clustering with predicted functional targets recovers breast cancer intrinsic subtypes

Fifty-seven human orthologues (ORTHO-57) were NetNC-FTI functional targets in ≥4 TF_ALL datasets and represented within integrated gene expression microarray data for 2999 breast tumours (BrC_2999) (Moleirinho et al., 2013). Unsupervised clustering using 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 (Figure S4). Heatmap features were annotated according to the dendrogram structure and gene expression intensity (Figure 5). The datasets sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b} represent embryos formed entirely from mesodermal lineages (Zeitlinger et al., 2007) 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<8.0x10^{-4}); consistent with EMT characteristics of basal-like breast cancers (Sarrió et al., 2008). Basal-like tumours were characterised by *EN1* and *NOTCH1*, aligning with previous findings (Barnawi et al., 2016; Beltran et al., 2014; Stylianou et al., 2006). Notch signalling modulation is a promising area for cancer therapy (Ntziachristos et al., 2014) and 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. Elevated *ETV6* expression was also a feature of basal-like cancers, where copy number amplifications and recurrent gene fusions were previously reported (Adélaïde et al., 2007; Letessier et al., 2005). The Luminal A subtype (feature_LumA) had similarities with luminal B (feature_LumB$_2$, ERBB3, MYO6) and normal-like (DOCK1, ERBB3, MYO6) tumours. High *BMPR1B* expression was a defining feature of the luminal A subtype, in agreement with oncogenic BMP signalling in luminal epithelia (Chapellier et al., 2015). *BMP2* may be pleiotropic, promoting EMT characteristics in some contexts (Ma et al., 2005; Ren and Dijke, 2017). *BMP2* expression was typically high in basal-like tumours and low in luminal cancers, aligning with *BMP2*

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

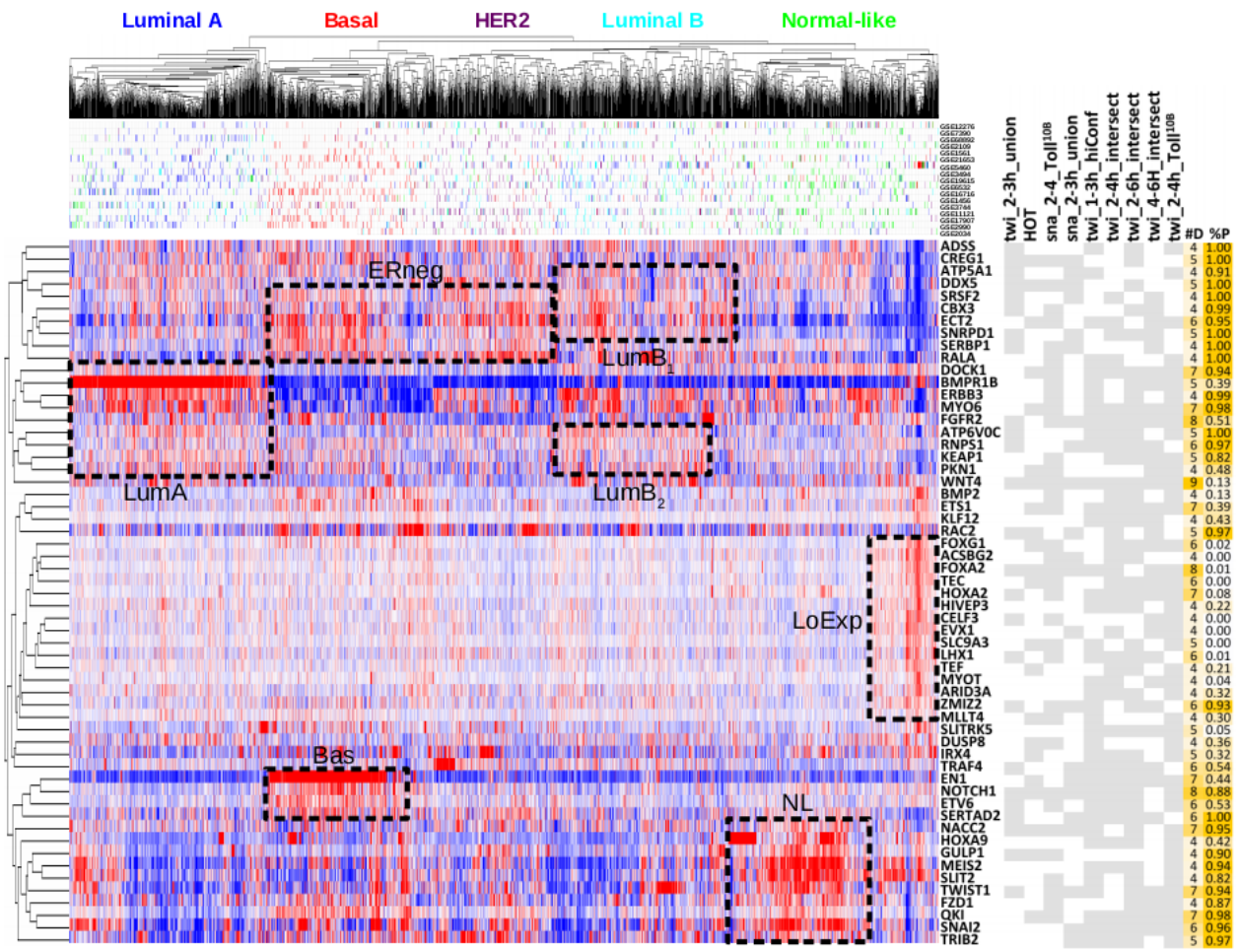


Figure 5. Predicted functional transcription factor targets capture human breast cancer biology. The heatmap shows results of unsupervised clustering with gene expression data for 2999 primary breast tumours and 57 orthologues of NetNC-FTI functional targets that were identified in at least four of nine 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 (Popovici et al., 2010). Some genes were annotated to more

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Also see Figure S4.

upregulation as a feature of the EMT programme in basal-like cancers. Several genes were highly expressed in both feature_LumB$_1$ and *ESR1* negative subtypes (feature_ERneg), including *ECT2*, *SNRPD1*, *SRSF2*, *CBX3*; our data suggest that these genes might contribute to the worse survival outcomes for luminal B relative to luminal A cancers. (Sørlie et al., 2001, 2003). *SNRPD1* and *SRSF2* function in splicing (Bermingham et al., 1995) which was linked to survival of multiple basal-like breast cancer cell lines (Chan et al., 2017). *CBX3* and *ECT2* were correlated with poor prognosis (Liang et al., 2017; Wang et al., 2018). These genes typically had low expression in Luminal A and normal-like subtypes. Feature_LoExp represents genes with low detection rates and tumours populating feature_LoExp are a mixture of subtypes, largely from a single study (Popovici et al., 2010). Key EMT genes (*SNAI2*, *TWIST1*, *QKI*) were assigned to the NL centroid and had highest relative expression in normal-like tumours (feature_NL, Figure 5). Feature_NL also included homeobox transcription factors (*HOXA9*, *MEIS2*) and a secreted cell migration guidance gene (*SLIT2*). Some genes had high expression in both normal-like and basal-like cancers, including *QKI*, which regulates circRNA formation in EMT (Conn et al., 2015), and the *FZD1* wnt/β-catenin receptor. Moreover, genes in feature_Bas and feature_NL clustered together, reflecting expression similarities between normal-like and basal-like subtypes. EMT may confer stem-like cell properties (DiMeo et al., 2009; Mani et al., 2008; Schmidt et al., 2015) and our results were consistent with dedifferentiation or arrested differentiation due to activation of an EMT-like programme in NL cancers. Previous work found stem cell markers in normal-like cancers (Sieuwerts et al., 2009; Sørlie et al., 2001) and *SNAI2* was critical for maintenance of mammary stem cells and linked with

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a stem-like signature in breast cancers (Guo et al., 2012; Lawson et al., 2015). Cell-compositional effects, associated with high stromal content in NL tumours (Prat and Perou, 2011), could also explain EMT-like molecular characteristics. We also found many differences between normal-like and basal-like cancer subtypes (Figure 4, Figure S3). Our results demonstrated that NetNC functional targets from fly mesoderm development capture clinically relevant molecular features of breast cancers and proposed novel candidate drivers of tumour progression.

Integrating NetNC functional target networks and breast cancer transcriptome profiling

Orthologous basal-like and normal-like genes were annotated onto NetNC-FTI networks, offering a new perspective on the molecular circuits controlling these different subtypes (Figure 4, Figure S3). We focussed on basal-like and normal-like cancers, which accounted for the large majority of networks studied and were prominent in centroid and heatmap analysis (Figure 5, Figure S4). Furthermore, EMT had been shown to be important for basal-like breast cancer biology (Sarrió et al., 2008) and, interestingly, key EMT genes were assigned to the normal-like subtype. Splicing factors and components of the ribosome were associated with the normal-like subtype in three networks (twi_2-4h_intersect, twi-2-6h_intersect, twi_2-3h_union); additionally, the proteasome and proteasome regulatory subunits in the twi_2-3h_union network had many genes annotated to the normal-like subtype.

The sna_2-4h_Toll[10b] 'RNA degradation and transcriptional regulation' cluster was exclusively annotated to the basal-like subtype and included *HECA*, which may function as a tumour suppressor (Lin et al., 2013) and an oncogene (Chien et al., 2006). *HECA* was also identified in NetNC-FTI analysis of twi_2-4h_intersect and twi_4-6h_intersect; these datasets had Twist binding at different, non-contiguous sites that were both assigned to *hdc*, the fly HECA orthologue. *Hdc* was a multifunctional *Notch* signalling modifier, including in cell survival

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Resende et al., 2017) and tracheal branching morphogenesis, activated by *escargot* (Steneberg et al., 1998). *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.

The *SLC9A6* Na^+/H^+ antiporter was found in NetNC-FTI ion transport clusters for sna_2-4h_Toll^{10b} and twi_2-4h_Toll^{10b}. Alterations in pH by Na^+/H^+ exchangers, particularly *SLC9A1*, drive basal-like breast cancer progression and chemoresistance (Amith and Fliegel, 2017). *SLC9A6* was 1.6-fold upregulated in basal-like relative to normal-like tumours ($p<8.4x10^{-71}$) and might drive pH dysregulation as part of an EMT-like programme in basal-like cancers. Sna_2-4h_Toll^{10b} clusters were depleted in normal-like annotations, compared with Twist; 4/18 clusters had two or more normal-like orthologues, significantly fewer than twi_ 2-4h_Toll^{10b} (8/17, $p<0.035$) and twi_2-3h_union (12/27, binomial $p<0.01$). A further basal-like cluster for twi_2-3h_union was annotated to 'mitochondrial translation' (MT), an emerging area of interest for cancer therapy (Weinberg and Chandel, 2015). Our results highlight MT as a potentially attractive target in basal-like breast cancers, aligning with work linking MT upregulation with loss of *RB1* and *p53* (Jones et al., 2016).

Chromatin organisation clusters frequently associated with basal-like annotations. For example, the twi_2-3h_union 'chromatin organisation and transcriptional regulation' cluster had six basal-like genes, including three Notch modifiers (*ash1*, *tara*, *Bap111)*. These were orthologous to the *ASH1L* histone methyltransferase and candidate poor prognosis factor with copy number amplifications in basal-like tumours (Liu et al., 2014); the *SERTAD2* bromodomain interacting oncogene and E2F activator (Cheong et al., 2009); and *SMARCE1*, a core subunit of the SWI/SNF chromatin remodelling complex that regulated *ESR1*, interacted with *HIF1A* signalling and potentiated breast cancer metastasis (García-Pedrero et al., 2006; Sethuraman et al., 2016; Sokol et al., 2017). Notch can promote EMT-like characteristics and mediated hypoxia-induced invasion in multiple cell lines (Sahlgren et al., 2008). Consistent with these studies, our work supported

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conserved function for *SMARCE1* in EMT-like signalling, both in mesoderm development and basal-like breast cancers, possibly downstream of *NOTCH1* and through regulation of SWI/SNF targeting. Indeed, SWI/SNF controled chromatin switching in oral cancer EMT (Mohd-Sarip et al., 2017). *Taranis*, orthologous to *SERTAD2*, also functions to stabilise the expression of *engrailed* in regenerating tissue (Schuster and Smith-Bolton, 2015). The *engrailed* orthologue *EN1* was the clearest single basal-like biomarker in the data examined (Figure 5) and acts as a survival factor (Beltran et al., 2014). *SERTAD2* could cooperate with *EN1* in basal-like cancers; 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 results demonstrating key roles for both neural-specific and EMT TFs in basal-like breast cancers (Beltran et al., 2014; Sarrió et al., 2008). Therefore, we identify chromatin organisation factors downstream of Snail and Twist with orthologues that may control *Notch* output and breast cancer progression through a chromatin remodelling mechanism. Indeed, NetNC results predict components of feedback loops where EMT TFs regulate chromatin organisation genes that, in turn, may both reinforce and coordinate downstream stages of gene expression programmes for mesoderm development and cancer progression. Stages of the EMT programme were described elsewhere, reviewed in (Nieto et al., 2016); 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 (Hemberger et al., 2009). 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 (Nieto et al., 2016).

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

NetNC results predicted new gene functions in EMT and cell invasion, including *SNX29* (also known as *RUNDC2A*), *ATG3*, *IRX4* and *UNK*. We investigated the functional and instructive role of these genes in an established invasion model (Dhasarathy et al., 2007). MCF7 cells are weakly invasive (Lacroix and Leclercq, 2004), 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 (Figure 6). 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.

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 (Nelson et al., 2016) and is also associated with prostate cancer risk (Xu et al., 2014). *SNX29* belongs to the sorting nexin protein family that function in endosomal sorting and signalling (Marat and Haucke, 2016). *SNX29* is poorly characterised and ectopic expression significantly reduced invasion in a *SNAI1*-dependent manner (Figure 6). Since we obtained these results, *SNX29* downregulation was associated with metastasis and chemoresistance in ovarian carcinoma (Zhu et al., 2015), consistent with *SNX29* inhibition of invasion driven by Snail. *ATG3* is an E2-like enzyme required for autophagy and mitochondrial homeostasis (Doherty and Baehrecke, 2018), *ATG3* overexpression significantly increased MCF7 invasion. Knockdown of *ATG3* reduced invasion in hepatocellular carcinoma (Li et al., 2013). *UNK* is a RING finger protein homologous to *unkempt* which binds mRNA, functions in ubiquitination and was upregulated in gastrulation (Mohler et al., 1992). Others reported that *UNK* mRNA binding controls neuronal

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

morphology and can induce spindle-like cell shape in fibroblasts (Murn et al., 2015, 2016). *UNK* significantly increased MCF7 invasion both independently of and additively with Snail; supporting a potential role in breast cancer progression. Indeed, *UNK* was overexpressed in cancers relative to controls in ArrayExpress (Parkinson et al., 2009). These *in vitro* confirmatory results both support the novel analysis approach and evidence new function for the genes examined.

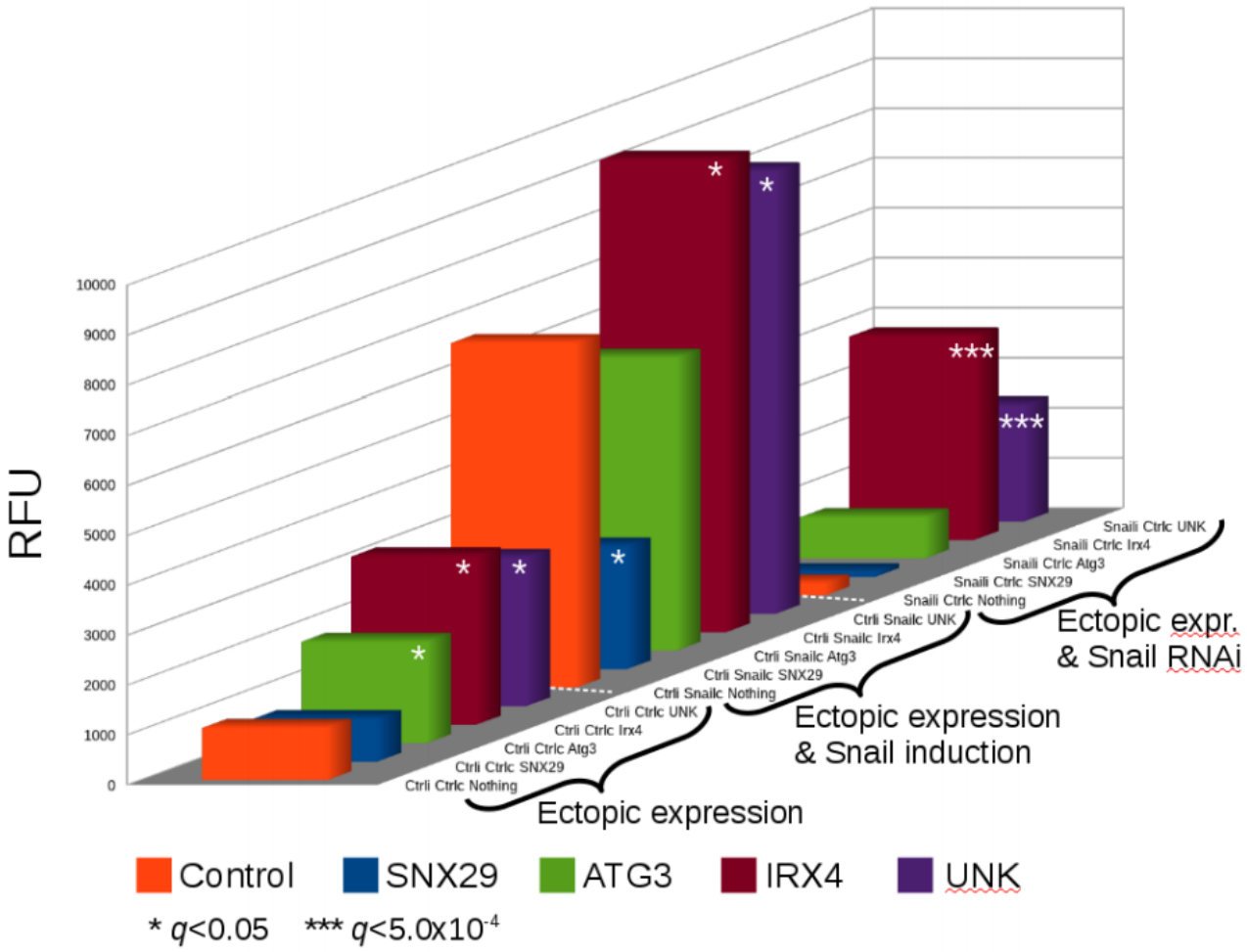


Figure 6. Validation of candidate invasion genes in breast cancer cells. Fluorescence from invasive MCF7 cells is shown. Induction of every gene examined significantly changed invasion in least one of three conditions: a) ectopic expression; b) ectopic expression and *SNAI1* induction; c) ectopic expression with shRNA knockdown of *SNAI1*. *SNX29* (blue) had reduced invasion compared with the *SNAI1* induction control (orange); *UNK* (purple) and *IRX4* (dark red) had increased invasion in all three conditions examined; *ATG3* had higher invasion at background levels of *SNAI1* (without induction or knockdown). Mean values are shown, n=3 datapoint; statistical significance is shown by * $q<0.05$; *** $q<5.0\times10^{-4}$. Also see Supplemental File 2.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

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 (Overton et al., 2011). Functional interaction probabilities, corresponding to pathway co-membership, were estimated by logistic regression of Bayesian probabilities from STRING v8.0 scores (Jensen et al., 2009) and Gene Ontology (GO) coannotations (Ashburner et al., 2000), taking KEGG (Kanehisa et al., 2010) 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 (Overton et al., 2011). 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:

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

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 (van Rijsbergen, 1979) and True Positive Rate

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(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) (Warde-Farley et al., 2010) and DROID (v2011_08) (Yu et al., 2008) were assessed against TEST-NET (Table S1, Figure S1).

Enrichment of DroFN edges in DroPIM was estimated as follows: a total of 999 genes were found in both DroFN and the DroPIM network thresholded at FDR 0.05 (DroPIM_FDR). These 999 genes had 5747 edges in DroPIM_FDR and 25797 edges in DroFN, of which 2175 were common to both networks. A 2x2 contingency table was constructed conditioning on the presence of edges for these 999 genes in the DroFN and DroPIM_FDR networks. The contingency table cell corresponding to edges not found in DroFN or DroPIM was populated by the number of possible edges for the 999 genes ($(n^2 - n) / 2$), subtracting the values from the other cells. Therefore the contingency table cell values were: 2175, 3572, 23622, 469132. The enrichment *p*-value was calculated by Fisher's Exact Test.

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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) (Goldberg and Roth, 2003), 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 (Goldberg and Roth, 2003) 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 (Storey, 2002) and an iterative minimum cut procedure (Ford and Fulkerson, 1956) 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 (Ford and Fulkerson, 1956) 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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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) (Lubbock et al., 2013):

   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 (Dempster et al., 1977; Lubbock et al., 2013). 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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

The following procedure is employed to estimate positive False Discovery Rate (pFDR) (Storey, 2002) 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 (Yamada and Bork, 2009). 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 (Goldberg and Roth, 2003) used in the NetNC algorithm to analyse *S*. Hence, the

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Storey, 2002):

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

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:

$$V = \frac{H_0}{Y}, p \leq p_\alpha \tag{4}$$

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 (Fitzgibbon et al., 2008; Sennels et al., 2009). The value of *R* is estimated by:

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

Additionally, there is a requirement for monotonicity:

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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 (Marusyk et al., 2012; Raj and van Oudenaarden, 2008; Raj et al., 2010). 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.

Estimating local false discovery rate from global false discovery rate

We developed an approach to estimate local false discovery rate (lcFDR) (Efron et al., 2001), 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:

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Figure S5). If we define the difference between $pFDR_k$ and $pFDR_l$:

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Median difference and correlation between estimates of neutral binding from NetNC Functional Target Identification and local False Discovery Rate

Candidate target genes that did not pass NetNC-FTI thresholds were labelled neutrally bound (FTI_NB). The proportion of FTI_NB genes was compared to the proportion of neutral binding estimated by lcFDR (lcFDR_NB, Figure 3A). The modulus of the difference between FTI_NB and lcFDR_NB for each dataset gave a distribution of differences in neutral binding proportion and the median of this distribution was quoted in the text above.

NetNC benchmarking and parameter optimisation

Gold standard data for NetNC benchmarking and parameterisation were taken as pathways from KEGG (v62, downloaded 13/6/12) (Kanehisa et al., 2010). 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 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 (López-Schier and St Johnston, 2001; Schmitt and Nebreda, 2002). TEST-CL also included the related pathways dme04745 'Phototransduction' and dme00600 'Sphingolipid metabolism', for example where ceramide kinase regulates photoreceptor homeostasis (Acharya et al., 2003; Dasgupta et al., 2009; Yonamine et al., 2011).

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-

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

standard data was taken to model neutral TF binding (Li et al., 2008; Shlyueva et al., 2014) and matches expectations on data taken from biological systems in general (Marusyk et al., 2012; Raj and van Oudenaarden, 2008). 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). 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). 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 (Enright et al., 2002) utilised a single subsample for 10%, 25%, 50% and 75% SNTGs (TRAIN-CL-SR, TEST-CL-SR).

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 available from the Biomodels database and contour plots showing mean MCC across Q, Z values per %SNTG are provided in Figure S7. 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 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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

Performance on blind test data

We compared NetNC against leading methods, HC-PIN (Wang et al., 2011) and MCL (Enright et al., 2002) on blind test data (Figure 2, Table S2). 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 (Wang et al., 2011); therefore HC-PIN was selected for extensive comparison against NetNC. Input, output and performance summary files for HC-PIN on TEST-CL are available from the Biomodels database (per datapoint, n=100 for NetNC, n=99 for HC-PIN). 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 (Overton et al., 2011). 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 (Table S9).

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 (MacArthur et al., 2009; Ozdemir et al., 2011; Roy et al., 2010; Sandmann et al., 2007; Zeitlinger et al., 2007). 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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Table S4). 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 Supplemental Material.

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 (Benjamini and Hochberg, 1995) unless otherwise specified in the text. Calculation of pFDR and local FDR values by NetNC is described in the sections above.

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) (MacArthur et al., 2009; Sandmann et al., 2007; Zeitlinger et al., 2007) and ChIP followed by solexa pyrosequencing (ChIP-seq) (Ozdemir et al., 2011). Additionally 'highly occupied target' regions, reflecting multiple and complex transcription factor occupancy profiles, were obtained from ModEncode (Roy et al., 2010). 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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these were assigned to the closest transcribed gene according to RNA Polymerase II binding data (MacArthur et al., 2009). 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 (Stathopoulos et al., 2002; Zeitlinger et al., 2007); 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 (Zeitlinger et al., 2007). The two-fold enrichment threshold selected for this study reflects 'weak' binding, although was expected to include functional TF targets (Biggin, 2011). 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 (Roy et al., 2010); 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 (Ozdemir et al., 2011). 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 (Sandmann et al., 2007) of which 1444 mapped to DroFN (Intersect_ALL). Subsets of

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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) (Campos-Ortega and Hartenstein, 1997; MacArthur et al., 2009; Ozdemir et al., 2011; Sandmann et al., 2007; Zeitlinger et al., 2007). 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) (Campos-Ortega and Hartenstein, 1997; Sandmann et al., 2007; Zeitlinger et al., 2007); 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) (Campos-Ortega and Hartenstein, 1997; Zeitlinger et al., 2007). Twi_2-4h_intersect and sna_2-4h_intersect were tightly staged between stages 5-7 (Sandmann et al., 2007). 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) (Campos-Ortega and Hartenstein, 1997). The twi_4-6h_intersect dataset was restricted to stages eight to nine which included germ band elongation and segmentation of neuroblasts (Campos-Ortega and Hartenstein, 1997; Sandmann et al., 2007). Therefore, there were differences in the biological material used across TF_ALL.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Notch signalling modifiers analysed in this study were selected based on identification in at least two of the screens reported in (Guruharsha et al., 2012).

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 (Dai et al., 2005). RMA normalisation (Irizarry et al., 2003) and ComBat batch correction (Johnson et al., 2007) were applied to remove dataset-specific bias as previously described (Moleirinho et al., 2013; Sims et al., 2008). Intrinsic molecular subtypes were assigned based upon the highest correlation to Sorlie centroids (Sørlie et al., 2003), applied to each dataset separately. Centred average linkage clustering was performed using the Cluster and TreeView programs (Eisen et al., 1998). Centroids were calculated for each gene based upon the mean expression across each of the Sorlie intrinsic subtypes (Sørlie et al., 2003). 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 (Östlund et al., 2009). Differential expression was calculated by t-test comparing normalised (unsquared) expression values in normal-like and basal-like tumours with false discovery rate correction (Benjamini and Hochberg, 1995).

Invasion assays for validation of genes selected from NetNC results

MCF-7 Tet-On cells were purchased from Clontech and maintained as previously described (Liu et al., 2013).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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Liu et al., 2013)). pGoldilox has been used previously to induce and knock down the expression of *Ets* genes (Peluso et al., 2017). 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 (Essafi et al., 2011); please see Supplemental File 2 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.

ACKNOWLEDGEMENTS

IMO is grateful to Prof Jeremy Gunawardena and Prof Peter Sorger for hosting his visit to HMS and for helpful discussions. Thanks to Prof PS Thiagarajan, Prof Andrew Millar, Prof Wendy Bickmore, Prof Nick Hastie, Prof Ben Lehner and Prof Julian Dow for invaluable comments. Mr Nick Moir and Dr Seanna McTaggart assisted with testing the NetNC software distribution. We are grateful for funding from Medical Research Council (MC_UU_12018/25; IMO), Royal Society of Edinburgh Scottish Government Fellowship cofunded by Marie Curie Actions (IMO), Marie Curie

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

AUTHOR 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 invasion assay 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.

DECLARATION OF INTERESTS

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Acharya, U., Patel, S., Koundakjian, E., Nagashima, K., Han, X., and Acharya, J.K. (2003). Modulating Sphingolipid Biosynthetic Pathway Rescues Photoreceptor Degeneration. Science *299*, 1740–1743.

Adélaïde, J., Finetti, P., Bekhouche, I., Repellini, L., Geneix, J., Sircoulomb, F., Charafe-Jauffret, E., Cervera, N., Desplans, J., Parzy, D., et al. (2007). Integrated Profiling of Basal and Luminal Breast Cancers. Cancer Res *67*, 11565–11575.

Amith, S.R., and Fliegel, L. (2017). 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 *43*, 35–41.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet *25*, 25–29.

Ashraf, S.I., and Ip, Y.T. (2001). The Snail protein family regulates neuroblast expression of inscuteable and string, genes involved in asymmetry and cell division in Drosophila. Development *128*, 4757–4767.

Barnawi, R., Al-Khaldi, S., Majed Sleiman, G., Sarkar, A., Al-Dhfyan, A., Al-Mohanna, F., Ghebeh, H., and Al-Alwan, M. (2016). Fascin Is Critical for the Maintenance of Breast Cancer Stem Cell Pool Predominantly via the Activation of the Notch Self-Renewal Pathway. Stem Cells *34*, 2799–2813.

Baylies, M.K., and Bate, M. (1996). twist: a myogenic switch in Drosophila. Science *272*, 1481–1484.

BDGP Berkley Drosophila Genome Project in situ homepage. http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl

Beltran, A.S., Graves, L.M., and Blancafort, P. (2014). 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 *33*, 4767–4777.

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

Bermingham, J.R., Arden, K.C., Naumova, A.K., Sapienza, C., Viars, C.S., Fu, X.D., Khotz, J., Manley, J.L., and Rosenfeld, M.G. (1995). Chromosomal localization of mouse and human genes encoding the splicing factors ASF/SF2 (SFRS1) and SC-35 (SFRS2). Genomics *29*, 70–79.

Bernard, F., Krejci, A., Housden, B., Adryan, B., and Bray, S.J. (2010). Specificity of Notch pathway activation: Twist controls the transcriptional output in adult muscle progenitors. Development *137*, 2633–2642.

Biggin, M.D. (2011). Animal Transcription Networks as Highly Connected, Quantitative Continua. Developmental Cell *21*, 611–626.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Boyle, A.P., Araya, C.L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L., Janette, J., Jiang, L., et al. (2014). Comparative analysis of regulatory information and circuits across distant species. Nature *512*, 453.

Bray, S.J. (2016). Notch signalling in context. Nat Rev Mol Cell Biol *17*, 722–735.

Brown, J.B., and Celniker, S.E. (2015). Lessons from modENCODE. Annual Review of Genomics and Human Genetics *16*, 31–53.

Campos-Ortega, J.A., and Hartenstein, V. (1997). The Embryonic Development of Drosophila melanogaster (Berlin, Heidelberg: Springer Science & Business Media).

Cannavò, E., Khoueiry, P., Garfield, D., Geeleher, G., Zichner, T., Gustafson, E., Ciglar, L., Korbel, J., and Furlong, E. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr Biol *26*, 38–51.

Caron, S.J.C., Ruta, V., Abbott, L.F., and Axel, R. (2013). Random convergence of olfactory inputs in the Drosophila mushroom body. Nature *497*, 113–117.

Cejalvo, J.M., Dueñas, E.M. de, Galvan, P., García-Recio, S., Gasión, O.B., Paré, L., Antolin, S., Martinello, R., Blancas, I., Adamo, B., et al. (2017). Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. Cancer Res *77*, 2213-2221.

Chan, S., Sridhar, P., Kirchner, R., Lock, Y.J., Herbert, Z., Buonamici, S., Smith, P., Lieberman, J., and Petrocca, F. (2017). Basal-A Triple Negative Breast Cancer Cells Selectively Rely on RNA Splicing for Survival. Mol Cancer Ther *16*, 2849-2861.

Chapellier, M., Bachelard-Cascales, E., Schmidt, X., Clément, F., Treilleux, I., Delay, E., Jammot, A., Ménétrier-Caux, C., Pochon, G., Besançon, R., et al. (2015). Disequilibrium of BMP2 Levels in the Breast Stem Cell Niche Launches Epithelial Transformation by Overamplifying BMPR1B Cell Response. Stem Cell Reports *4*, 239–254.

Chen, C.-M., Freedman, J.A., Bettler, D.R., Manning, S.D., Giep, S.N., Steiner, J., and Ellis, H.M. (1996). polychaetoid is required to restrict segregation of sensory organ precursors from proneural clusters in Drosophila. Mechanisms of Development *57*, 215–227.

Chen, J., Hu, Z., Phatak, M., Reichard, J., Freudenberg, J.M., Sivaganesan, S., and Medvedovic, M. (2013). Genome-Wide Signatures of Transcription Factor Activity: Connecting Transcription Factors, Disease, and Small Molecules. PLOS Computational Biology *9*, e1003198.

Chen, R.A.-J., Stempor, P., Down, T.A., Zeiser, E., Feuer, S.K., and Ahringer, J. (2014). Extreme HOT regions are CpG-dense promoters in C. elegans and humans. Genome Res. *24*, 1138–1146.

Cheong, J.K., Gunaratnam, L., Zang, Z.J., Yang, C.M., Sun, X., Nasr, S.L., Sim, K.G., Peh, B.K., Rashid, S.B.A., Bonventre, J.V., et al. (2009). TRIP-Br2 promotes oncogenesis in nude mice and is frequently overexpressed in multiple human tumors. Journal of Translational Medicine *7*, 8.

Chien, C.-C., Chang, C.-C., Yang, S.-H., Chen, S.-H., and Huang, C.-J. (2006). A homologue of the Drosophila headcase protein is a novel tumor marker for early-stage colorectal cancer. Oncol. Rep. *15*, 919–926.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conn, S.J., Pillman, K.A., Toubia, J., Conn, V.M., Salmanidis, M., Phillips, C.A., Roslan, S., Schreiber, A.W., Gregory, P.A., and Goodall, G.J. (2015). The RNA Binding Protein Quaking Regulates Formation of circRNAs. Cell *160*, 1125–1134.

Crosby, M.A., Miller, C., Alon, T., Watson, K.L., Verrijzer, C.P., Goldman-Levi, R., and Zak, N.B. (1999). The trithorax Group Gene moira Encodes a Brahma-Associated Putative Chromatin-Remodeling Factor in Drosophila melanogaster. Molecular and Cellular Biology *19*, 1159.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res *33*, e175–e175.

Dasgupta, U., Bamba, T., Chiantia, S., Karim, P., Tayoun, A.N.A., Yonamine, I., Rawat, S.S., Rao, R.P., Nagashima, K., Fukusaki, E., et al. (2009). Ceramide kinase regulates phospholipase C and phosphatidylinositol 4, 5, bisphosphate in phototransduction. PNAS *106*, 20063–20068.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. J Roy Stat Soc Ser B *39*, 1–38.

Dhasarathy, A., Kajita, M., and Wade, P.A. (2007). The Transcription Factor Snail Mediates Epithelial to Mesenchymal Transitions by Repression of Estrogen Receptor Alpha. Mol Endocrinol *21*, 2907–2918.

DiMeo, T.A., Anderson, K., Phadke, P., Feng, C., Perou, C.M., Naber, S., and Kuperwasser, C. (2009). A Novel Lung Metastasis Signature Links Wnt Signaling with Cancer Cell Self-Renewal and Epithelial-Mesenchymal Transition in Basal-like Breast Cancer. Cancer Res *69*, 5364–5373.

Doherty, J., and Baehrecke, E.H. (2018). Life, death and autophagy. Nature Cell Biology *20*, 1110.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association *96*, 1151–1160.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. PNAS *95*, 14863–14868.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.

Essafi, A., Webb, A., Berry, R.L., Slight, J., Burn, S.F., Spraggon, L., Velecela, V., Martinez-Estrada, O.M., Wiltshire, J.H., Roberts, S.G.E., et al. (2011). A Wt1-Controlled Chromatin Switching Mechanism Underpins Tissue-Specific Wnt4 Activation and Repression. Developmental Cell *21*, 559–574.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fanti, L., Dorer, D.R., Berloco, M., Henikoff, S., and Pimpinelli, S. (1998). Heterochromatin protein 1 binds transgene arrays. Chromosoma *107*, 286–292.

Fitzgibbon, M., Li, Q., and McIntosh, M. (2008). Modes of Inference for Evaluating the Confidence of Peptide Identifications. J. Proteome Res. *7*, 35–39.

Ford, L.R., and Fulkerson, D.R. (1956). Maximal flow through a network. Canadian Journal of Mathematics *8*, 399–404.

García-Pedrero, J.M., Kiskinis, E., Parker, M.G., and Belandia, B. (2006). The SWI/SNF Chromatin Remodeling Subunit BAF57 Is a Critical Regulator of Estrogen Receptor Function in Breast Cancer Cells. J. Biol. Chem. *281*, 22656–22664.

Giampieri, S., Manning, C., Hooper, S., Jones, L., Hill, C.S., and Sahai, E. (2009). Localized and reversible TGFbeta signalling switches breast cancer cells from cohesive to single cell motility. Nat. Cell Biol. *11*, 1287–1296.

Gilmour, D., Rembold, M., and Leptin, M. (2017). From morphogen to morphogenesis and back. Nature *541*, 311–320.

Goldberg, D.S., and Roth, F.P. (2003). Assessing experimentally derived interactions in a small world. PNAS *100*, 4372–4376.

Gramates, L.S., Marygold, S.J., Santos, G. dos, Urbano, J.-M., Antonazzo, G., Matthews, B.B., Rey, A.J., Tabone, C.J., Crosby, M.A., Emmert, D.B., et al. (2017). FlyBase at 25: looking to the future. Nucleic Acids Res *45*, D663–D671.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat Genet *47*, 569–576.

Guo, W., Keckesova, Z., Donaher, J.L., Shibue, T., Tischler, V., Reinhardt, F., Itzkovitz, S., Noske, A., Zürrer-Härdi, U., Bell, G., et al. (2012). Slug and Sox9 Cooperatively Determine the Mammary Stem Cell State. Cell *148*, 1015–1028.

Guruharsha, K.G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al. (2011). A Protein Complex Network of Drosophila melanogaster. Cell *147*, 690.

Guruharsha, K.G., Kankel, M.W., and Artavanis-Tsakonas, S. (2012). The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet *13*, 654–666.

Hammonds, A.S., Bristow, C.A., Fisher, W.W., Weiszmann, R., Wu, S., Hartenstein, V., Kellis, M., Yu, B., Frise, E., and Celniker, S.E. (2013). Spatial expression of transcription factors in Drosophila embryonic organ development. Genome Biol. *14*, R140.

Hartley, D., Xu, T., and Artavanis-Tsakonas, S. (1987). 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 *6*, 3407–3417.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47–C52.

Hemberger, M., Dean, W., and Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nature Reviews Molecular Cell Biology *10*, 526–537.

Herranz, N., Pasini, D., Díaz, V.M., Francí, C., Gutierrez, A., Dave, N., Escrivà, M., Hernandez-Muñoz, I., Croce, L.D., Helin, K., et al. (2008). Polycomb Complex 2 Is Required for E-cadherin Repression by the Snail1 Transcription Factor. Molecular and Cellular Biology *28*, 4772–4781.

Hooper, S.D., Boué, S., Krause, R., Jensen, L.J., Mason, C.E., Ghanim, M., White, K.P., Furlong, E.E., and Bork, P. (2007). Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Molecular Systems Biology *3*, 72.

Hu, J.X., Thomas, C.E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. Nat Rev Genet *17*, 615–629.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1–13.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18*, S233–S240.

Igual, J.C., Johnson, A.L., and Johnston, L.H. (1996). 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 *15*, 5001–5013.

Ingthorsson, S., Briem, E., Bergthorsson, J.T., and Gudjonsson, T. (2016). Epithelial Plasticity During Human Breast Morphogenesis and Cancer Progression. J Mammary Gland Biol Neoplasia *21*, 139–148.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

Jaeger, S., Igea, A., Arroyo, R., Alcalde, V., Canovas, B., Orozco, M., Nebreda, A.R., and Aloy, P. (2017). Quantification of Pathway Cross-talk Reveals Novel Synergistic Drug Combinations for Breast Cancer. Cancer Res *77*, 459–469.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res *37*, D412-416.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Jones, R.A., Robinson, T.J., Liu, J.C., Shrestha, M., Voisin, V., Ju, Y., Chung, P.E.D., Pellecchia, G., Fell, V.L., Bae, S., et al. (2016). RB1 deficiency in triple-negative breast cancer induces mitochondrial protein translation. J Clin Invest *126*, 3739–3757.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Karczewski, K.J., Snyder, M., Altman, R.B., and Tatonetti, N.P. (2014). Coherent Functional Modules Improve Transcription Factor Target Identification, Cooperativity Prediction, and Disease Association. PLOS Genetics *10*, e1004122.

Keung, A.J., Bashor, C.J., Kiriakov, S., Collins, J.J., and Khalil, A.S. (2014). Using Targeted Chromatin Regulators to Engineer Combinatorial and Spatial Transcriptional Regulation. Cell *158*, 110–120.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.C., Gustafson, E.H., Sinha, S., and Furlong, E.E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. Elife *6*, e28440.

Koppens, M., and Lohuizen, M. van (2016). Context-dependent actions of Polycomb repressors in cancer. Oncogene *35*, 1341–1352.

Kusch, T., and Reuter, R. (1999). Functions for Drosophila brachyenteron and forkhead in mesoderm specification and cell signalling. Development *126*, 3991–4003.

Kvon, E.Z., Stampfel, G., Yáñez-Cuna, J.O., Dickson, B.J., and Stark, A. (2012). HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev. *26*, 908–913.

Lacroix, M., and Leclercq, G. (2004). Relevance of Breast Cancer Cell Lines as Models for Breast Tumours: An Update. Breast Cancer Res Treat *83*, 249–289.

Lawson, D.A., Bhakta, N.R., Kessenbrock, K., Prummel, K.D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.-Y., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature *526*, 131–135.

Leptin, M. (1991). twist and snail as positive and negative regulators during Drosophila mesoderm development. Genes Dev. *5*, 1568–1576.

Letessier, A., Ginestier, C., Charafe-Jauffret, E., Cervera, N., Adélaïde, J., Gelsi-Boyer, V., Ahomadegbe, J.-C., Benard, J., Jacquemier, J., Birnbaum, D., et al. (2005). ETV6 gene rearrangements in invasive breast carcinoma. Genes Chromosomes Cancer *44*, 103–108.

Li, H., Chen, H., Liu, F., Ren, C., Wang, S., Bo, X., and Shu, W. (2015). Functional annotation of HOT regions in the human genome: implications for human disease and cancer. Scientific Reports *5*, 11633.

Li, J., Yang, B., Zhou, Q., Wu, Y., Shang, D., Guo, Y., Song, Z., Zheng, Q., and Xiong, J. (2013). Autophagy promotes hepatocellular carcinoma cell invasion through activation of epithelial–mesenchymal transition. Carcinogenesis *34*, 1343–1351.

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L.L., et al. (2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS Biol *6*, e27.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Liang, Y.-K., Lin, H.-Y., Chen, C.-F., and Zeng, D. (2017). Prognostic values of distinct CBX family members in breast cancer. Oncotarget *8*, 92375–92387.

Lim, J., and Thiery, J.P. (2012). Epithelial-mesenchymal transitions: insights from development. Development *139*, 3471–3486.

Lin, S., Ewen-Campen, B., Ni, X., Housden, B.E., and Perrimon, N. (2015). In Vivo Transcriptional Activation Using CRISPR/Cas9 in Drosophila. Genetics *201*, 433–442.

Lin, X., Li, J., Yin, G., Zhao, Q., Elias, D., Lykkesfeldt, A.E., Stenvang, J., Brünner, N., Wang, J., Yang, H., et al. (2013). 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 *15*, R119.

Liu, J., Jiang, G., Liu, S., Liu, Z., Pan, H., Yao, R., and Liang, J. (2013). Lentivirus-delivered short hairpin RNA targeting SNAIL inhibits HepG2 cell growth. Oncology Reports *30*, 1483–1487.

Liu, L., Kimball, S., Liu, H., Holowatyj, A., Yang, Z.-Q., Liu, L., Kimball, S., Liu, H., Holowatyj, A., and Yang, Z.-Q. (2014). Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. Oncotarget *6*, 2466–2482.

Lo, P.C.H., Skeath, J.B., Gajewski, K., Schulz, R.A., and Frasch, M. (2002). Homeotic Genes Autonomously Specify the Anteroposterior Subdivision of the Drosophila Dorsal Vessel into Aorta and Heart. Developmental Biology *251*, 307–319.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Lopez, A., Higuet, D., Rosset, R., Deutsch, J., and Peronnet, F. (2001). corto genetically interacts with Pc-G and trx-G genes and maintains the anterior boundary of Ultrabithorax expression in Drosophila larvae. Mol Gen Genomics *266*, 572–583.

López-Schier, H., and St Johnston, D. (2001). Delta signaling from the germ line controls the proliferation and differentiation of the somatic follicle cells during Drosophila oogenesis. Genes Dev. *15*, 1393–1405.

Lubbock, A.L.R., Katz, E., Harrison, D.J., and Overton, I.M. (2013). TMA Navigator: network inference, patient stratification and survival analysis with tissue microarray data. Nucleic Acids Research *41*, W562–W568.

Ma, L., Lu, M.-F., Schwartz, R.J., and Martin, J.F. (2005). Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. Development *132*, 5601–5611.

MacArthur, S., Li, X.-Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keränen, S.V., et al. (2009). 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 *10*, R80.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics *21*, 3448–3449.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mani, S.A., Guo, W., Liao, M.-J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., et al. (2008). The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. Cell *133*, 704–715.

Marat, A.L., and Haucke, V. (2016). Phosphatidylinositol 3-phosphates—at the interface between cell signalling and membrane traffic. The EMBO Journal *35*, 561–579.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer *12*, 323–334.

Millo, H., and Bownes, M. (2007). The expression pattern and cellular localisation of Myosin VI during the Drosophila melanogaster life cycle. Gene Expression Patterns *7*, 501–510.

Mishra, K., Chopra, V.S., Srinivasan, A., and Mishra, R.K. (2003). Trl-GAGA directly interacts with lola like and both are part of the repressive complex of Polycomb group of genes. Mechanisms of Development *120*, 681–689.

Mohd-Sarip, A., Teeuwssen, M., Bot, A.G., De Herdt, M.J., Willems, S.M., Baatenburg de Jong, R.J., Looijenga, L.H.J., Zatreanu, D., Bezstarosti, K., van Riet, J., et al. (2017). DOC1-Dependent Recruitment of NURD Reveals Antagonism with SWI/SNF during Epithelial-Mesenchymal Transition in Oral Cancer Cells. Cell Reports *20*, 61–75.

Mohler, J., Weiss, N., Murli, S., Mohammadi, S., Vani, K., Vasilakis, G., Song, C.H., Epstein, A., Kuang, T., and English, J. (1992). The embryonically active gene, unkempt, of Drosophila encodes a Cys3His finger protein. Genetics *131*, 377–388.

Mohr, S.E., Hu, Y., Kim, K., Housden, B.E., and Perrimon, N. (2014). Resources for Functional Genomics Studies in Drosophila melanogaster. Genetics *197*, 1-18.

Moleirinho, S., Chang, N., Sims, A.H., Tilston-Lünel, A.M., Angus, L., Steele, A., Boswell, V., Barnett, S.C., Ormandy, C., Faratian, D., et al. (2013). KIBRA exhibits MST-independent functional regulation of the Hippo signaling pathway in mammals. Oncogene *32*, 1821–1830.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A Regulatory Archipelago Controls Hox Genes Transcription in Digits. Cell *147*, 1132–1145.

Moorman, C., Sun, L.V., Wang, J., Wit, E. de, Talhout, W., Ward, L.D., Greil, F., Lu, X.-J., White, K.P., Bussemaker, H.J., et al. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. PNAS *103*, 12027–12032.

Murn, J., Zarnack, K., Yang, Y.J., Durak, O., Murphy, E.A., Cheloufi, S., Gonzalez, D.M., Teplova, M., Curk, T., Zuber, J., et al. (2015). Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes Dev *29*, 501–512.

Murn, J., Teplova, M., Zarnack, K., Shi, Y., and Patel, D.J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nat Struct Mol Biol *23*, 16–23.

Nelson, D.O., Lalit, P.A., Biermann, M., Markandeya, Y.S., Capes, D.L., Addesso, L., Patel, G., Han, T., John, M.C., Powers, P.A., et al. (2016). Irx4 Marks a Multipotent, Ventricular-Specific Progenitor Cell. Stem Cells *34*, 2875–2888.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nevil, M., Bondra, E.R., Schulz, K.N., Kaplan, T., and Harrison, M.M. (2017). Stable Binding of the Conserved Transcription Factor Grainy Head to its Target Genes Throughout Drosophila melanogaster Development. Genetics *205*, 605–620.

Nieto, M.A., Huang, R.Y.-J., Jackson, R.A., and Thiery, J.P. (2016). EMT: 2016. Cell *166*, 21–45.

Nowell, C.S., and Radtke, F. (2017). Notch as a tumour suppressor. Nature Reviews Cancer *17*, 145.

Ntziachristos, P., Lim, J.S., Sage, J., and Aifantis, I. (2014). From Fly Wings to Targeted Cancer Therapies: A Centennial for Notch Signaling. Cancer Cell *25*, 318–334.

Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2009). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research *38*, D196–D203.

Overton, I.M., Graham, S., Gould, K.A., Hinds, J., Botting, C.H., Shirran, S., Barton, G.J., and Coote, P.J. (2011). Global network analysis of drug tolerance, mode of action and virulence in methicillin-resistant S. aureus. BMC Syst Biol *5*, 68.

Ozdemir, A., Fisher-Aylor, K.I., Pepke, S., Samanta, M., Dunipace, L., McCue, K., Zeng, L., Ogawa, N., Wold, B.J., and Stathopoulos, A. (2011). High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. Genome Res. *21*, 566–577.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., et al. (2009). ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Research *37*, D868–D872.

Pawson, T., and Nash, P. (2003). Assembly of Cell Regulatory Systems Through Protein Interaction Domains. Science *300*, 445–452.

Peluso, S., Douglas, A., Hill, A., Angelis, C.D., Moore, B.L., Grimes, G., Petrovich, G., Essafi, A., and Hill, R.E. (2017). 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 *6*, e28590.

Popovici, V., Chen, W., Gallas, B.G., Hatzis, C., Shi, W., Samuelson, F.W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res *12*, R5.

Prat, A., and Perou, C.M. (2011). Deconstructing the molecular portraits of breast cancer. Molecular Oncology *5*, 5–23.

Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913–918.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ren, J., and Dijke, P. ten (2017). Bone Morphogenetic Proteins in the Initiation and Progression of Breast Cancer. In Vukicevic S., Sampath K.(eds) Bone Morphogenetic Proteins: Systems Biology Regulators pp409–433. Progress in Inflammation Research. Springer.

Resende, L.P.F., Truong, M.E., Gomez, A., and Jones, D.L. (2017). Intestinal stem cell ablation reveals differential requirements for survival in response to chemical challenge. Developmental Biology *424*, 10–17.

Rhee, D.Y., Cho, D.-Y., Zhai, B., Slattery, M., Ma, L., Mintseris, J., Wong, C.Y., White, K.P., Celniker, S.E., Przytycka, T.M., et al. (2014). Transcription Factor Networks in Drosophila melanogaster. Cell Reports *8*, 2031–2043.

van Rijsbergen, C.J. (1979). Information Retrieval. Butterworth-Heinemann.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., et al. (2010). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Sahlgren, C., Gustafsson, M.V., Jin, S., Poellinger, L., and Lendahl, U. (2008). Notch signaling mediates hypoxia-induced tumor cell migration and invasion. PNAS *105*, 6392–6397.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E.M. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev. *21*, 436–449.

Sarrió, D., Rodriguez-Pinilla, S.M., Hardisson, D., Cano, A., Moreno-Bueno, G., and Palacios, J. (2008). Epithelial-Mesenchymal Transition in Breast Cancer Relates to the Basal-like Phenotype. Cancer Research *68*, 989–997.

Schmidt, J.M., Panzilius, E., Bartsch, H.S., Irmler, M., Beckers, J., Kari, V., Linnemann, J.R., Dragoi, D., Hirschi, B., Kloos, U.J., et al. (2015). Stem-Cell-like Properties and Epithelial Plasticity Arise as Stable Traits after Transient Twist1 Activation. Cell Reports *10*, 131–139.

Schmitt, A., and Nebreda, A.R. (2002). Signalling pathways in oocyte meiotic maturation. J Cell Sci *115*, 2457–2459.

Schotta, G., Ebert, A., Krauss, V., Fischer, A., Hoffmann, J., Rea, S., Jenuwein, T., Dorn, R., and Reuter, G. (2002). Central role of *Drosophila* SU(VAR)3–9 in histone H3-K9 methylation and heterochromatic gene silencing. The EMBO Journal *21*, 1121–1131.

Schuster, K.J., and Smith-Bolton, R.K. (2015). Taranis Protects Regenerating Tissue from Fate Changes Induced by the Wound Response in Drosophila. Developmental Cell *34*, 119–128.

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

Sethuraman, A., Brown, M., Seagroves, T.N., Wu, Z.-H., Pfeffer, L.M., and Fan, M. (2016). SMARCE1 regulates metastatic potential of breast cancer cells through the HIF1A/PTK2 pathway. Breast Cancer Research *18*, 81.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science *343*, 84–87.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C., Bender, W., and Kingston, R.E. (1999). Stabilization of Chromatin Structure by PRC1, a Polycomb Complex. Cell *98*, 37–46.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272–286.

Sieuwerts, A.M., Kraan, J., Bolt, J., van der Spoel, P., Elstrodt, F., Schutte, M., Martens, J.W.M., Gratama, J.-W., Sleijfer, S., and Foekens, J.A. (2009). Anti-Epithelial Cell Adhesion Molecule Antibodies and the Detection of Circulating Normal-Like Breast Tumor Cells. J Natl Cancer Inst *101*, 61–66.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., Wit, E. de, Steensel, B. van, and Laat, W. de (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics *38*, 1348.

Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J., and Clarke, R.B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. BMC Medical Genomics *1*, 42.

Sokol, E.S., Feng, Y.-X., Jin, D.X., Tizabi, M.D., Miller, D.H., Cohen, M.A., Sanduja, S., Reinhardt, F., Pandey, J., Superville, D.A., et al. (2017). SMARCE1 is required for the invasive progression of in situ cancers. PNAS *114*, 4153-4158.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Rijn, M. van de, Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS *98*, 10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. PNAS *100*, 8418–8423.

Sparmann, A., and Lohuizen, M. van (2006). Polycomb silencers control cell fate, development and cancer. Nature Reviews Cancer *6*, 846–856.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613–626.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147–151.

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

Steneberg, P., Englund, C., Kronhamn, J., Weaver, T.A., and Samakovlis, C. (1998). Translational readthrough in the hdc mRNA generates a novel branching inhibitor in theDrosophila trachea. Genes Dev. *12*, 956–967.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 479–498.

Stylianou, S., Clarke, R.B., and Brennan, K. (2006). Aberrant Activation of Notch Signaling in Human Breast Cancer. Cancer Res *66*, 1517–1525.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. *102*, 15545–15550.

Teytelman, L., Thurtle, D.M., Rine, J., and Oudenaarden, A. van (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. PNAS *110*, 18602–18607.

Tie, F., Banerjee, R., Saiakhova, A.R., Howard, B., Monteith, K.E., Scacheri, P.C., Cosgrove, M.S., and Harte, P.J. (2014). Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. Development (Cambridge, England) *141*, 1129.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. *3*, research0088.

Trujillo, G.V., Nodal, D.H., Lovato, C.V., Hendren, J.D., Helander, L.A., Lovato, T.L., Bodmer, R., and Cripps, R.M. (2016). The canonical Wingless signaling pathway is required but not sufficient for inflow tract formation in the Drosophila melanogaster heart. Developmental Biology *413*, 16–25.

Vesuna, F., van Diest, P., Chen, J.H., and Raman, V. (2008). Twist is a transcriptional repressor of E-cadherin gene expression in breast cancer. Biochemical and Biophysical Research Communications *367*, 235–241.

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell *144*, 986–998.

Vitali, F., Li, Q., Schissler, A.G., Berghout, J., Kenost, C., and Lussier, Y.A. (2017). Developing a ‘personalome’ for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. Brief Bioinform bbx149.

Wang, H., Liang, J., Zheng, H., and Xiao, H. (2018). Expression and prognostic significance of ECT2 in invasive breast cancer. Journal of Clinical Pathology *71*, 442–445.

Wang, J., Li, M., Chen, J., and Pan, Y. (2011). A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics *8*, 607–620.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research *38*, W214–W220.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Weinberg, S.E., and Chandel, N.S. (2015). Targeting mitochondria metabolism for cancer therapy. Nat Chem Biol *11*, 9–15.

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

Wilczynski, B., and Furlong, E.E.M. (2010). Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. Developmental Biology *340*, 161–169.

Xie, Y., Li, X., Deng, X., Hou, Y., O'Hara, K., Urso, A., Peng, Y., Chen, L., and Zhu, S. (2016). The Ets protein Pointed prevents both premature differentiation and dedifferentiation of *Drosophila* intermediate neural progenitors. Development *143*, 3109–3118.

Xu, X., Hussain, W.M., Vijai, J., Offit, K., Rubin, M.A., Demichelis, F., and Klein, R.J. (2014). Variants at IRX4 as prostate cancer expression quantitative trait loci. Eur J Hum Genet *22*, 558–563.

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

Yonamine, I., Bamba, T., Nirala, N.K., Jesmin, N., Kosakowska-Cholody, T., Nagashima, K., Fukusaki, E., Acharya, J.K., and Acharya, U. (2011). Sphingosine kinases and their metabolites modulate endolysosomal trafficking in photoreceptors. J Cell Biol *192*, 557–567.

Yu, J., Pacifico, S., Liu, G., and Finley Jr., R.L. (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics *9*, 461.

Yu, M., Bardia, A., Wittner, B.S., Stott, S.L., Smas, M.E., Ting, D.T., Isakoff, S.J., Ciciliano, J.C., Wells, M.N., Shah, A.M., et al. (2013). Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. Science *339*, 580–584.

Zabidi, M.A., and Stark, A. (2016). Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. Trends in Genetics *32*, 801–814.

Zander, M.A., Burns, S.E., Yang, G., Kaplan, D.R., and Miller, F.D. (2014). Snail Coordinately Regulates Downstream Pathways to Control Multiple Aspects of Mammalian Neural Precursor Development. J. Neurosci. *34*, 5164–5175.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. (2007). Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev. *21*, 385–390.

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

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Material for Overton *et al.* 'Coherent Transcription Factor Target Networks Illuminate Epithelial Remodelling'

CONTENTS

Page

Supplemental Figures:

Figure S1 ……..……..……..

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

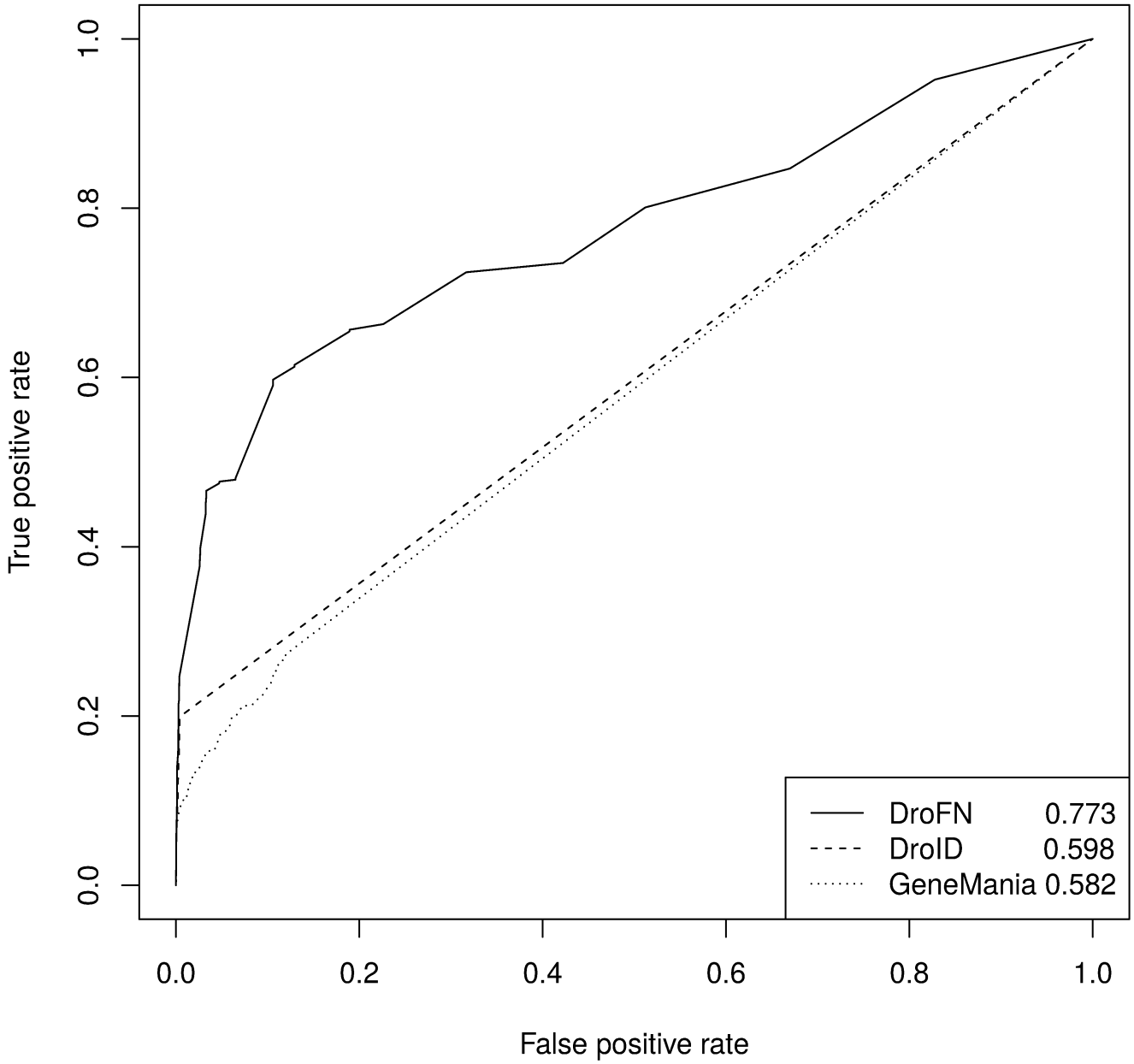


Figure S1. Related to Methods. Reciever Operator Characteristic (ROC) curves for Functional Gene Networks on time separated blind test data (TEST-NET). ROC curves for DroFN, DroID and GeneMania networks are shown with respective AUC values of 0.773, 0.598 and 0.582. DroFN has significantly larger AUC than DroID ($p<2.13\text{x}10^{-11}$) and GeneMania ($p<3.2\text{x}10^{-22}$) on TEST-NET.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

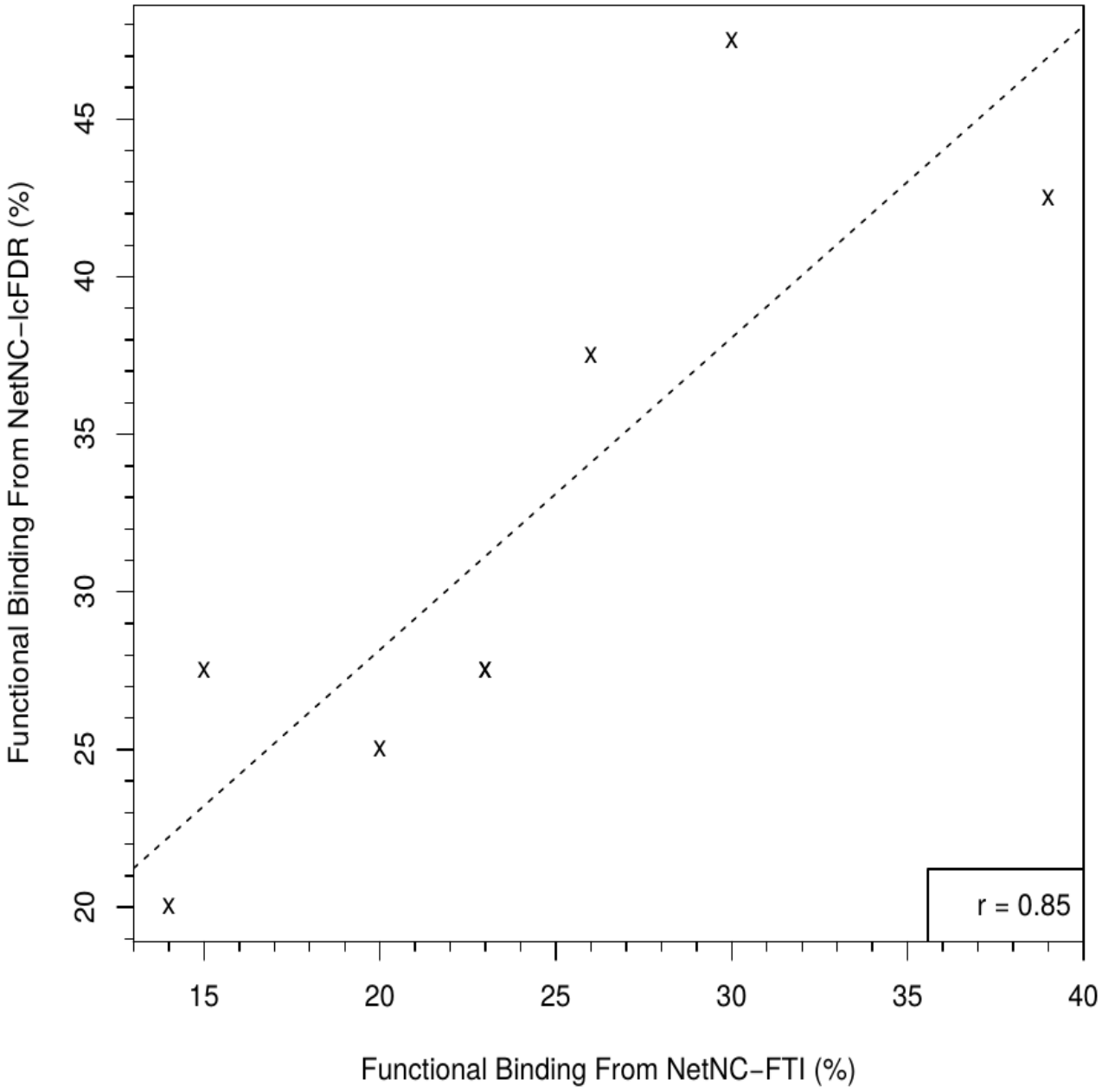


Figure S2. Related to Table 1. Agreement between NetNC-lcFDR and NetNC-FTI estimates of functional binding. The percentage of predicted functional binding from NetNC-lcFDR and NetNC-FTI for each of the nine TF_ALL datasets is shown as crosses ('X'). Results from the two different methods correlated well (r=0.9, *p*=0.008) and had median difference of only 5.5%. This concordance supports the results from both NetNC-FTI and NetNC-lcFDR.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

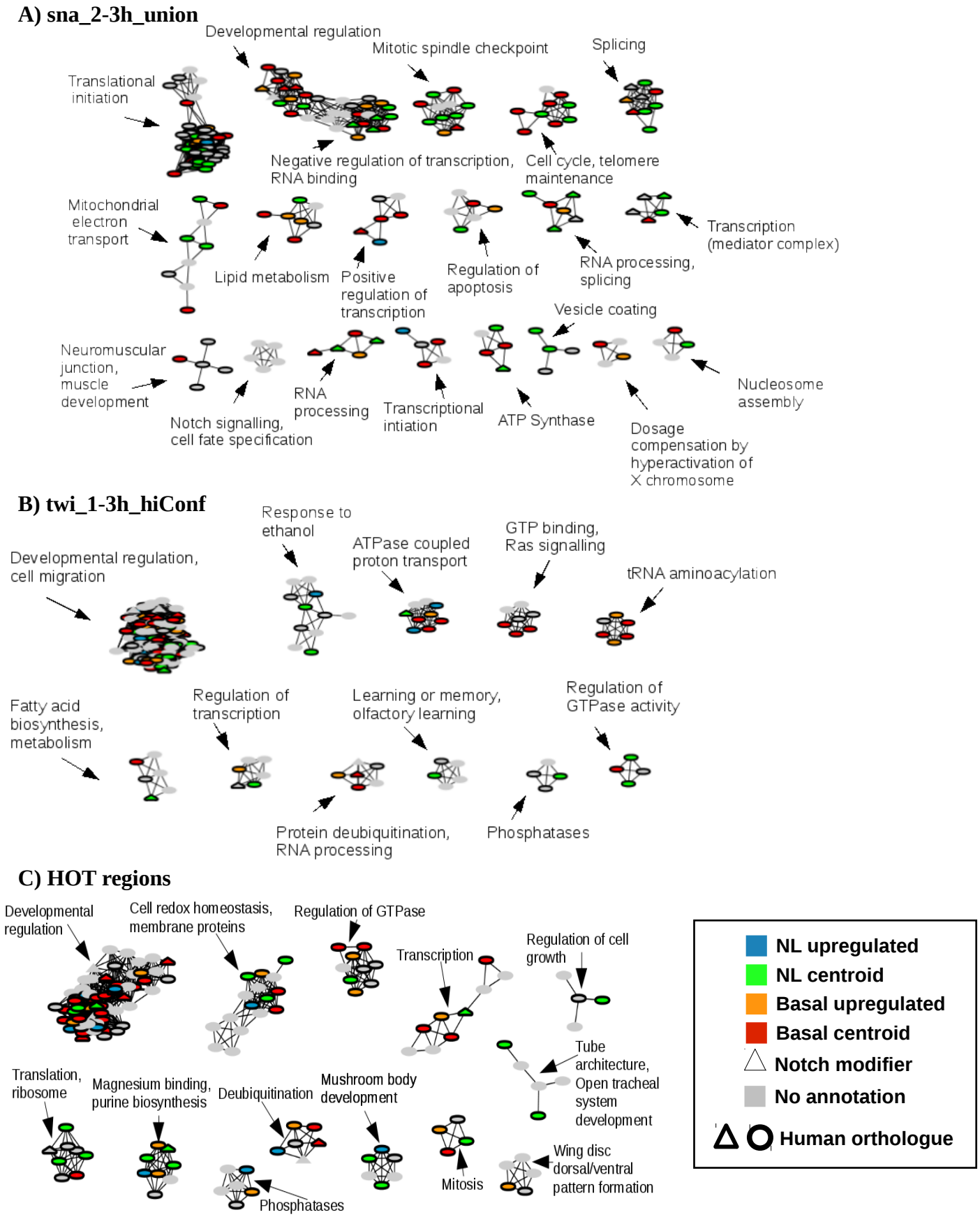


Figure S3. Networks of functionally coherent candidate target genes identified by NetNC-FTI.

Continued on next page.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

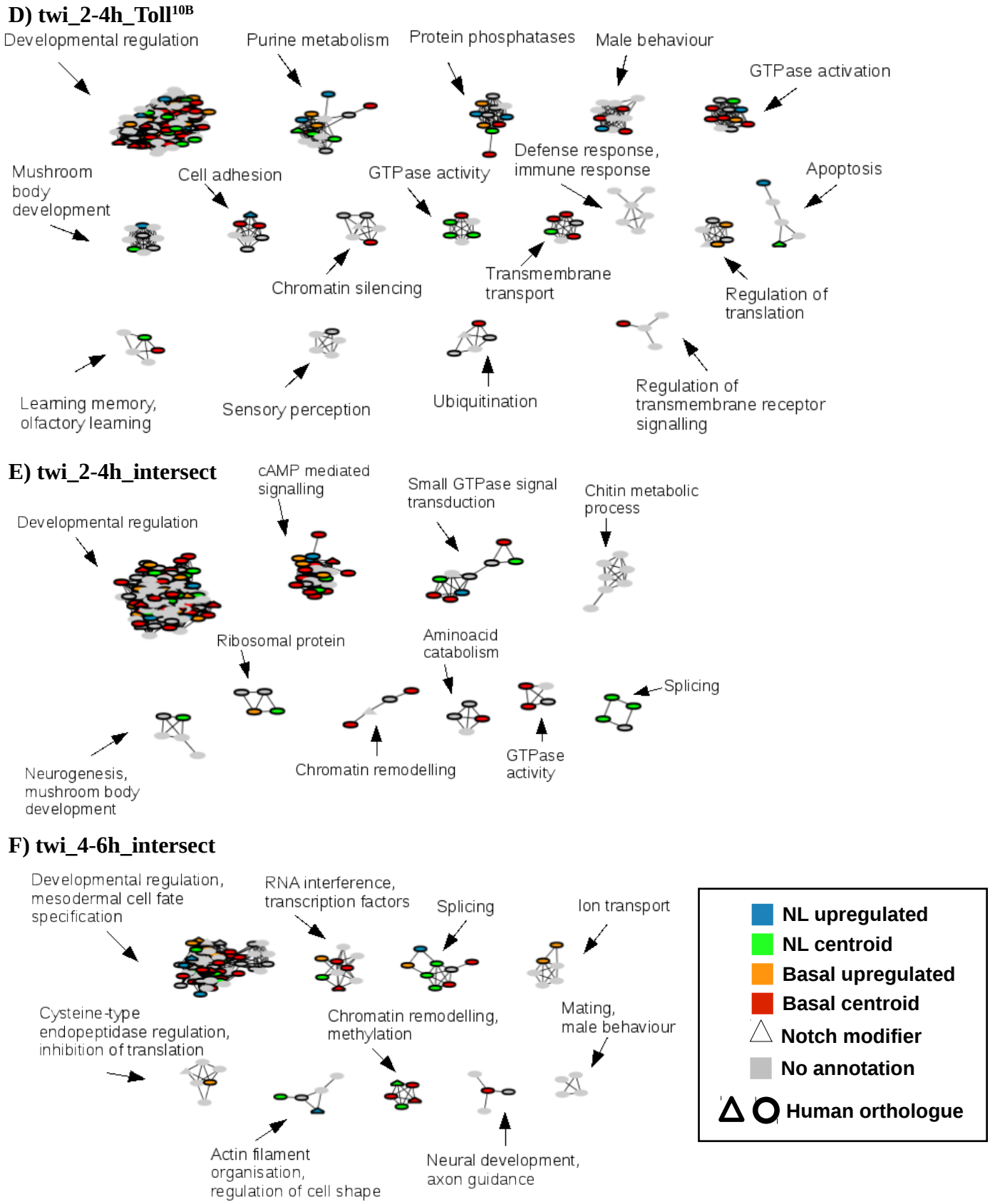


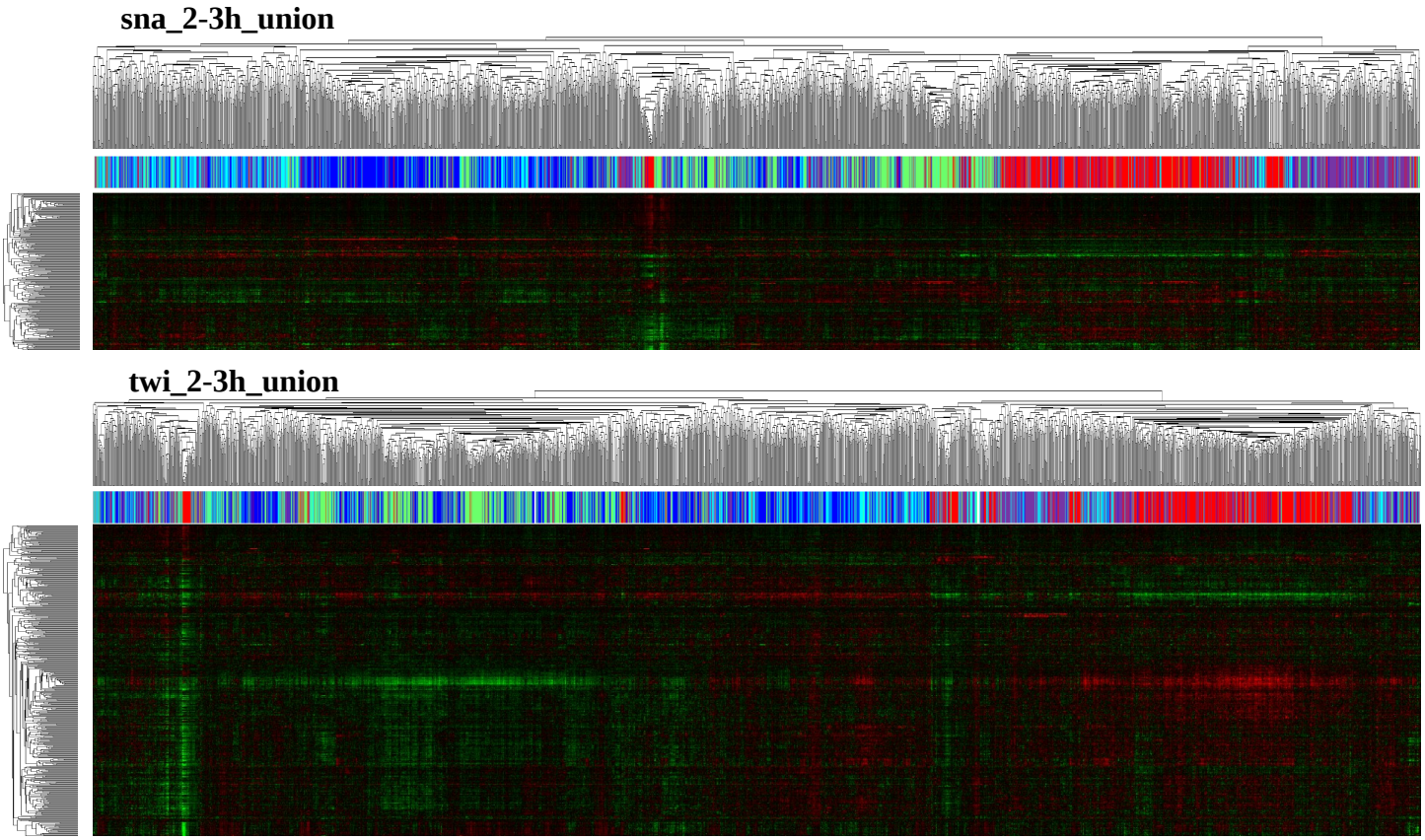
Figure S3. Related to Figure 4 and Supplemental File 1. Networks of functionally coherent candidate target genes identified by NetNC-FTI. The key (bottom right) indicates annotations for human orthology (bold node border) and Notch screen hits (triangular nodes). Many orthologues

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 to the BL or NL subtypes (grey). Clusters with at least four members are shown, cytoscape sessions with full NetNC-FTI results are given in Additional File 4. 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).

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(A)



(B)

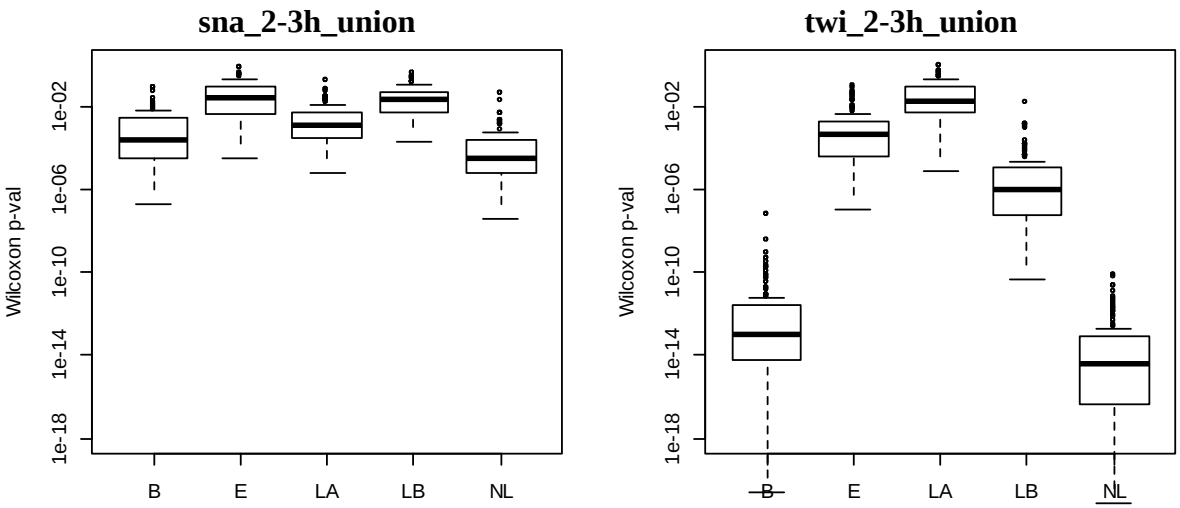


Figure S4. Related to Figure 5. Unsupervised clustering of 2999 primary breast cancers with NetNC results for individual Snail and Twist datasets.

(A) Data shown represent results clustering of primary breast tumour gene expression data (detailed in Methods, section 4.4) with orthologous genes from NetNC-FBT analysis of sna_2-3h_union (top, n=256) and twi_2-3h_union (bottom, n=467). Intrinsic subtypes are shown along the top of each

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

heatmap, colours correspond to: HER2-overexpressing (purple), basal-like (red), normal-like (green), luminal A (dark blue) and luminal B (light blue). Clustering based on sna_2-3h_union shows better separation of clusters than twi_2-3h_union.

(B) Significantly higher squared centroid values for multiple breast cancer subtypes (Wilcoxon $p<0.05$) were obtained when compared with resampled orthologues for NetNC-FBT analysis of sna_2-3h_union (left, n=256) and twi_2-3h_union (right, n=467); 100 resamples were performed. Breast cancer subtypes are indicated on the x-axis of each plot as follows: B (basal-like), E (HER2-overexpressing), LA (luminal A), LB (luminal B), NL (normal-like). Similar patterns were observed between these two datasets, with strongest enrichment for basal-like and normal-like centroids; sna_2-3h_union also showed relatively stronger enrichment for the luminal A centroid compared with results for twi_2-3h_union. Absolute differences in *p* between sna_2-3h_union and twi_2-3h_union may reflect the difference in dataset sizes (n=256 vs n=467), indeed stronger p-values are typically associated with comparisions across larger datasets.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

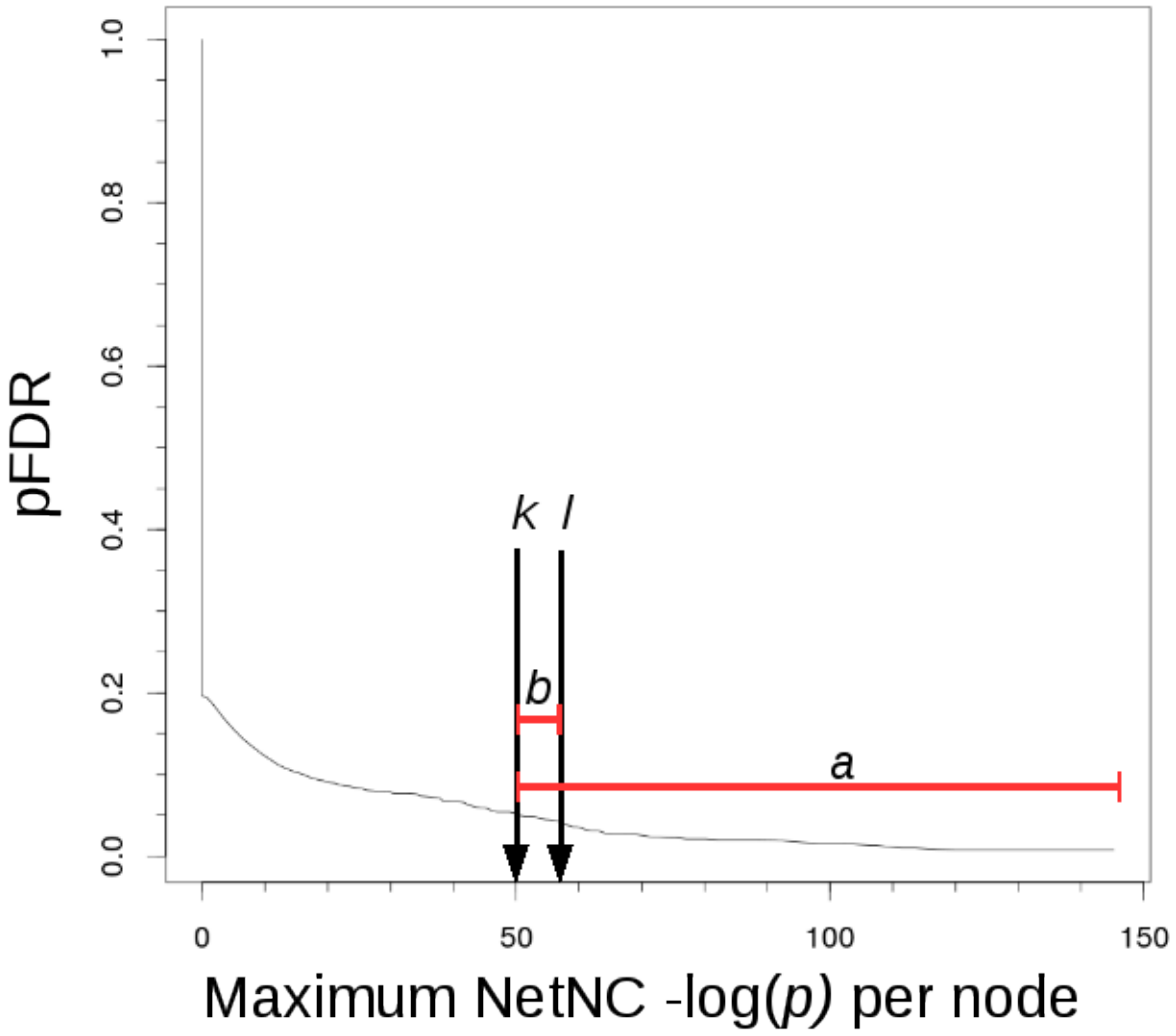


Figure S5. Related to Methods. Estimation of local False Discovery Rate (lcFDR) from positive False Discovery Rate. To estimate lcFDR the expected number of false positives (FP) is first estimated for the edges (X) under b (red bounding line). The pFDR at k ($pFDR_k$) multiplied by the number of edges (n) under a (red bounding line) gives the FP above threshold k (FP_k); the pFDR at l ($pFDR_l$) multiplied by ($n - X$) gives the expected number of false positives for threshold j (FP_l). Therefore, lcFDR bounded by k and l is estimated as ($FP_k - FP_l$) / X. We took values of k and l such that the interval (k, l) contained only edges that had $p \equiv k$. Values for the twi_2-6h_intersect dataset are shown.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

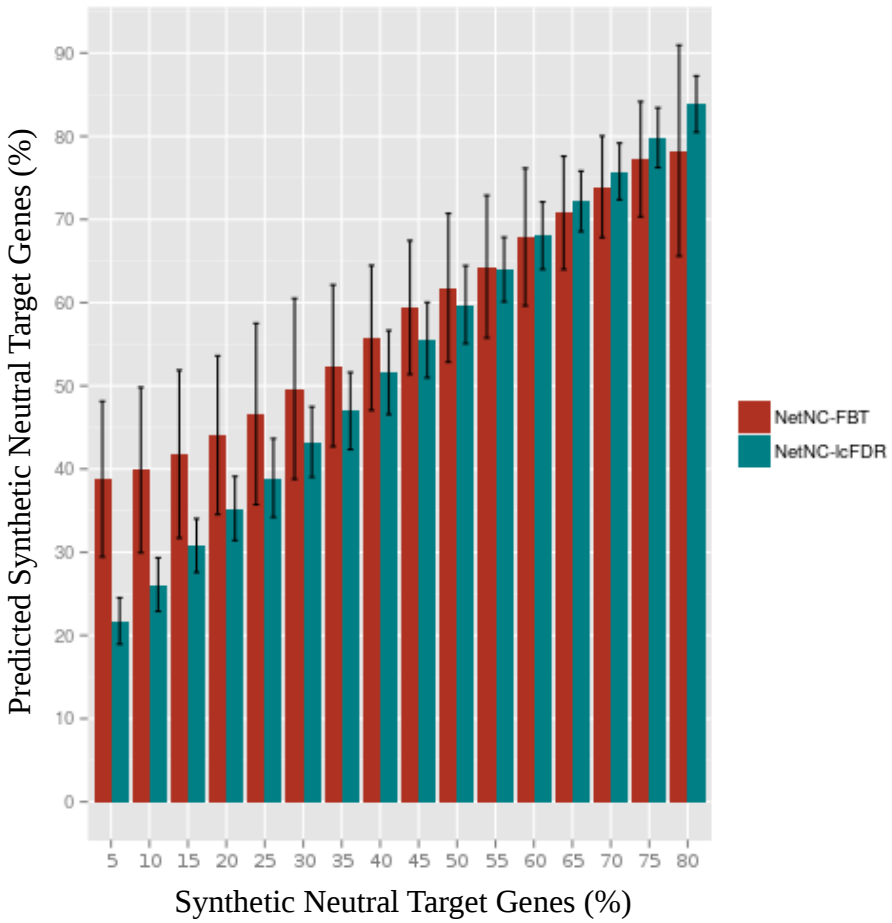


Figure S6. Related to Methods. Proportion of neutral binding predicted by NetNC-FBT and NetNC-lcFDR. The proportion of Synthetic Neutral Target Genes (SNTGs) in the TEST-CL_8PW datasets analysed is shown on the x-axis. The predicted proportion of SNTGs by NetNC-FBT (red) and NetNC-lcFDR (blue-green) is shown on the y-axis. Error bars represent 95% confidence intervals calculated over 100 resamples. Systematic overestimation of %SNTG was greatest for the NetNC-FBT approach which also had large 95% confidence intervals relative to NetNC-lcFDR.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

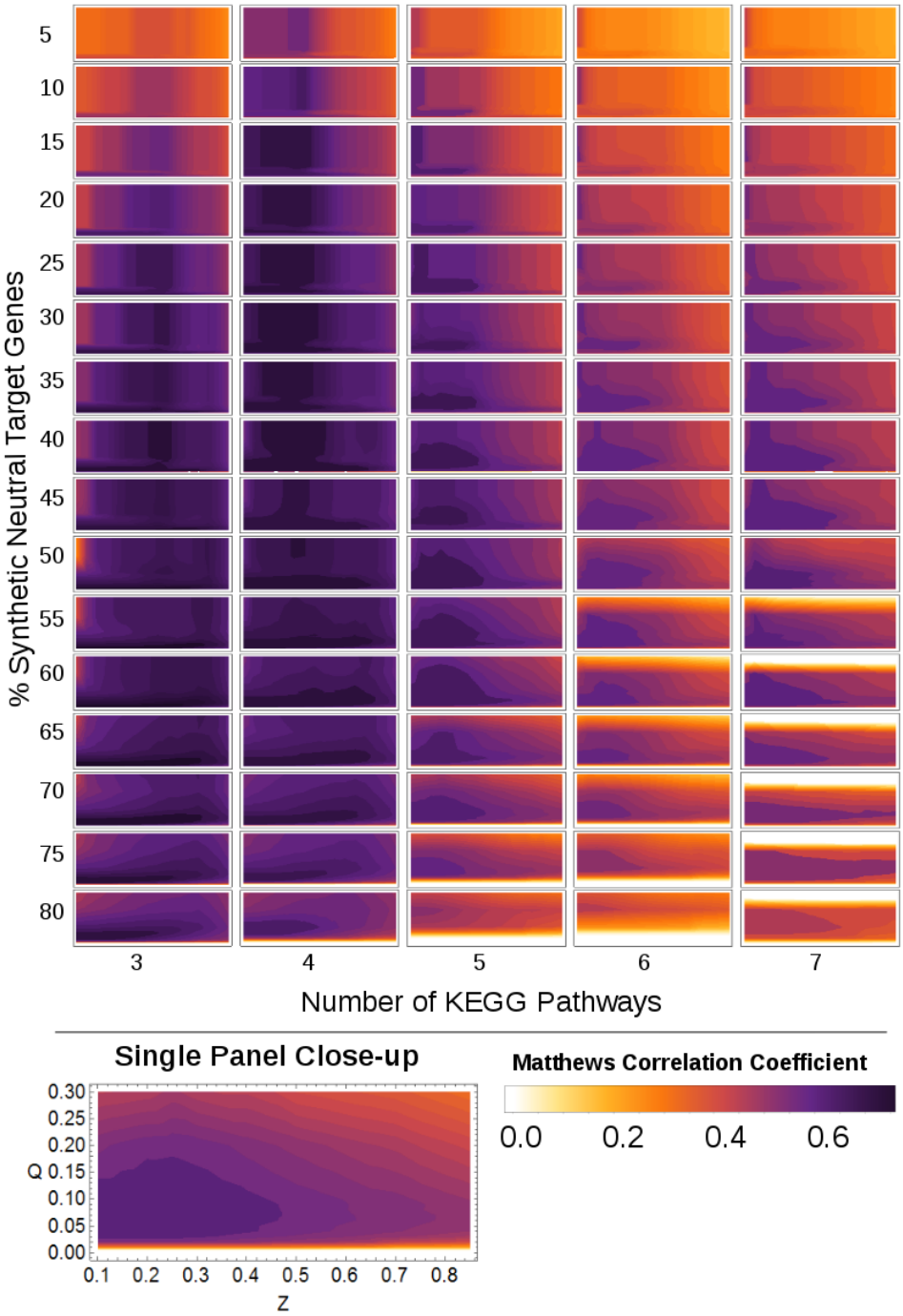


Figure S7. Related to Methods. NetNC Functional Target Identification Performance on TRAIN-CL_ALL. A range of values for the number of KEGG pathways (columns) and %Synthetic Neutral Target Genes (rows) were examined. Each invividual panel shows NetNC-FTI performance across values of the Q (FDR) and Z (graph density) parameters. Contours show the mean Matthews Correlation Coefficient value for the 100 resamples per individual panel. The NetNC-FTI parameters (Q, Z), were optimised on these data as described in the main manuscript Methods section 4.2.3.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 S1. Related to Methods. Evaluation of DroFN on Time Separated Blind Test Data (TEST-NET). 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=2.13x10^{-11}) or GeneMania (p=3.19x10^{-22}).

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Functional Target Identification (FTI)	
%SNTG	**NetNC MCC[+]**	**MCL MCC**
10	0.407	0.171 (0.258)
25	0.494	0.252 (0.309)
50	0.573	0.251 (0.365)
75	0.532	0.220 (0.220)

[+] NetNC-FTI parameters used (Z=0.1, Q=0.1) were preliminary. These parameters were determined empirically, with performance assessment on TRAIN-CL-SR.

Table S2. Related to Figure 2. Results from preliminary study of NetNC and MCL performance on TEST-CL-SR. MCC, Matthews Correlation Coefficient. MCL performance for overall best-performing inflation value (I=3.6) is given at FDR $p<0.05$. Performance values in brackets represent taking inflation values that were optimal at the value of %SNTG on TRAIN-CL-SR. Results represent performance on TEST-CL-SR, which comprised a single resample at each %SNTG value examined. MCC values for NetNC performance were considerably higher than those obtained using the MCL algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Descriptor for cluster or cluster grouping	twi_union	sna_union	sna_Toll^{10B}	twi_Toll^{10B}	twi_HiConf	twi_2-6h	twi_2-4h	twi_4-6h	HOT
Developmental Regulation									
Chromatin organisation									
Splicing									
Regulation of translation									
Olfactory learining, mushroom body development									
Ion transport									
Phosphatases									
Mitochondrial energy metabolism									
Cell shape, cytoskeleton, adhesion									
Ubiquitination									
Ribosome									
Apoptosis									
Neuromuscular junction									
Gene silencing by RNA									
Lipid metabolism									
Mating behaviour									
Mitochondrial targeting									
Cell division									
Mediator complex									
Vesicle trafficking									
Autophagy									

Table S3. Related to Figure 4 and Figure S3. Overall summary of clusters identified by NetNC-FTI across the nine TF_ALL datasets. Clusters and cluster groupings are listed on the left of the table, orange table cells indicates that the biological grouping was found in NetNC-FTI results for the corresponding dataset, blue table cells indicate that the grouping was not identified.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Dataset	Subsampling	Median percentage edge overlap (95% CI)	Median percentage gene overlap (95% CI)
sna_2-3h_union	95	90.7 (83.3-95.8)	89.0 (83.4-94.2)
sna_2-3h_union	80	83.9 (74.2-92.0)	80.3 (71.2-88.3)
sna_2-3h_union	50	77.5 (37.5-93.9)	73.1 (58.5-88.1)
twi_2-3h_union	95	94.9 (88.5-98.4)	92.4 (87.8- 97.0)
twi_2-3h_union	80	91.3 (81.8-95.3)	85.0 (80.0-90.5)
twi_2-3h_union	50	81.8 (67.1-94.5)	77.4 (69.6-87.6)
sna_2-4h_Toll10b	95	88.7 (54.2-96.7)	90.4 (72.1-97.1)
sna_2-4h_Toll10b	80	79.8 (28.0-92.5)	81.3 (55.9-91.3)
sna_2-4h_Toll10b	50	72.2 (15.4-99.5)	74.5 (46.7-99.5)
twi_2-4h_Toll10b	95	83.6 (42.5-98.5)	85.0 (63.4-96.6)
twi_2-4h_Toll10b	80	82.3 (33.6-95.9)	76.1 (54.6-91.6)
twi_2-4h_Toll10b	50	73.9 (24.5-95.3)	71.7 (49.4-92.2)
HOT	95	83.0 (61.0-97.1)	85.8 (65.4-97.6)
HOT	80	70.3 (41.1-90.4)	76.4 (59.9-91.9)
HOT	50	52.5 (32.1-88.9)	68.0 (52.8-92.3)
twi_4-6h_intersect	95	81.3 (29.4-96.5)	75.9 (47.7-95.9)
twi_4-6h_intersect	80	80.5 (16.1-94.9)	73.9 (38.5-91.6)
twi_4-6h_intersect	50	63.5 (12.3-100.0)	68.3 (36.8-100.0)
twi_2-4h_intersect	95	86.0 (61.2-98.8)	87.4 (72.2-97.3)
twi_2-4h_intersect	80	88.4 (51.0-99.2)	83.1 (66.0-97.2)
twi_2-4h_intersect	50	75.5 (41.9-100.0)	78.9 (57.9-100.0)
twi_2-6h_intersect	95	95.8 (89.7-99.6)	95.6 (91.0-98.9)
twi_2-6h_intersect	80	93.7 (81.1-98.9)	91.3 (85.0-96.8)
twi_2-6h_intersect	50	88.0 (75.2-99.2)	86.5 (77.9-96.6)
twi_1-3h_hiConf	95	96.4 (74.9-99.7)	93.9 (74.1-99.2)
twi_1-3h_hiConf	80	88.4 (67.0-98.8)	88.8 (77.2-95.6)
twi_1-3h_hiConf	50	79.1 (56.0-100.0)	83.0 (67.5-100.0)

Table S4. Related to Figure 4 and Figure S3. Overlap for NetNC-FTI subsamples of the TF_ALL datasets. Data represents the proportion of edges and genes returned by NetNC-FTI for the subsampled dataset that were also present in NetNC-FTI results for the full dataset (i.e. at 100%). At subsampling rates 95%, 80%, 50%, the average edge overlap values across TF_ALL were respectively 91%, 84% and 77%; derived as the median value (average) of the median overlap per dataset (respective median 95% CI 83-96%, 74-94%, 37-92%). At subsampling rates 95%,

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

80%, 50%, the average overlap values across TF_ALL were respectively 89%, 81% and 75%; derived as the median value (average) of the median overlap per dataset. (respective median 95% CI 72%-97.2%, 66%-92%, 58%-97%). 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 (twi_1-3h_hiConf, twi_2-6h_intersect, HOT; see Table 1) 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.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identifier	Name (count in TF_ALL)	Uniprot Function when available, otherwise the Flybase Summary is given (indicated by *)
FBgn0004009 Current ID: FBgn0284084	Wingless (9/9)	Binds as a ligand to a family of frizzled seven-transmembrane receptors and acts through a cascade of genes on the nucleus. Segment polarity protein. May be a growth factor. Acts on neighboring cells to regulate at least one gene, the homeobox segmentation gene engrailed. Wg signal represses arm phosphorylation. Wg signaling operates by inactivating the sgg repression of engrailed autoactivation. Wg and Wnt2 have a role in the developing trachea and together are responsible for all dorsal trunk formation. Wg also acts in the developing epidermis. Acts as a morphogen, and diffuses long distances despite its lipidation. Lipophorin is required for diffusion, probably by acting as vehicle for its movement, explaining how it can spread over long distances despite its lipidation. In non-neuronal cells, wls directs wg secretion via clathrin-mediated endocytosis and the retromer complex (a conserved protein complex consisting of Vps26 and Vps35) to sustain a wls traffic loop encompassing the Golgi, the cell surface, an endocytic compartment and a retrograde route leading back to the Golgi. In neuronal cells (the larval motorneuron NMJ), wg signal moves across the synapse through the release of wls-containing exosome-like vesicles.
FBgn0004647	Notch (8/9)	Signaling protein, which regulates, with both positive and negative signals, the differentiation of at least central and peripheral nervous system and eye, wing disk, oogenesis, segmental appendages such as antennae and legs, and muscles, through lateral inhibition or induction. Functions as a receptor for membrane-bound ligands Delta and Serrate to regulate cell-fate determination. Upon ligand activation, and releasing from the cell membrane, the Notch intracellular domain (NICD) forms a transcriptional activator complex with Su(H) (Suppressor of hairless) and activates genes of the E(spl) complex. Essential for proper differentiation of ectoderm. Fringe (fng) acts in the Golgi to determine the type of O-linked fucose on the EGF modules in N, altering the ability of N to bind

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		with Delta (Dl). O-fut1 also has a role in modulating the interaction. Rumi acts in the endoplasmic reticulum to glucosylate the EGF modules in N, this is required for correct folding and cleavage of N.
FBgn0000659	fork head (8/9)	Fkh promotes terminal as opposed to segmental development. In the absence of fkh, this developmental switch does not occur. The nuclear localization of the fkh protein suggest that fkh regulates the transcription of other, subordinate, genes.
FBgn0010389	heartless (7/9)	May be required for patterning of muscle precursor cells. May be essential for generation of mesodermal and endodermal layers, invaginations of various types of cells and CNS formation.
FBgn0011225	jaguar (7/9)	Myosin is a protein that binds to actin and has ATPase activity that is activated by actin. Together CLIP-190 and jar may coordinate the interaction between the actin and microtubule cytoskeleton. May link endocytic vesicles to microtubules and may be involved in transport in the early embryo and in the dynamic process of dorsal closure. It is believed that its function changes during the life cycle.
FBgn0000577	engrailed (7/9)	This protein specifies the body segmentation pattern. It is required for the development of the central nervous system. Transcriptional regulator that represses activated promoters. Wg signaling operates by inactivating the SGG repression of EN autoactivation.
FBgn0003118	pointed (7/9)	Required for glial-neuronal cell interactions at the ventral midline which are necessary for the proper elaboration of commissures in the embryonic CNS.
FBgn0017397 Current ID: FBgn0264491	held out wings (7/9)	Required for integrin-mediated cell-adhesion in wing blade. Vital role in steroid regulation of muscle development and to control heart rate. Required during embryogenesis, in late stages of somatic muscle development, for myotube migration and during metamorphosis for muscle reorganization.
FBgn0003254	ribbon (7/9)	*Ribbon is a nuclear BTB-domain protein, expressed in most embryonic cells. It is required for development of the salivary gland and trachea, as well as for dorsal closure. It regulates both growth and differentiation of salivary gland cells.
FBgn0000014	abdominal A (7/9)	Sequence-specific transcription factor which is part of a developmental regulatory system that provides cells with specific

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		positional identities on the anterior-posterior axis. Required for segmental identity of the second through eighth abdominal segments. Once a pattern of abd-A expression is turned on in a given parasegment, it remains on the more posterior parasegment, so that the complex pattern of expression is built up in the successive parasegments. Appears to repress expression of Ubx whenever they appear in the same cell, but abd-A is repressed by Abd-B only in the eight and ninth abdominal segments.
FBgn0003900	twist (7/9)	Involved in the establishment and dorsoventral patterning of germ layers in the embryo.
FBgn0003177 Current ID: FBgn0262614	polychaetoid (7/9)	*Polychaetoid is a broadly acting protein that is associated with multiple proteins at the surface and within the cytoskeleton, connecting events between the two.
FBgn0015513	myoblast city (7/9)	*Myoblast city is an unconventional bipartite GEF that, in combination with Ced-12, activates the Rac1 monomeric GTPase. It contributes to myoblast fusion, border cell migration and dorsal closure.

Table S5. Related to Figure 4 and Figure S3. Frequently identified genes within the Developmental Regulation Clusters. The thirteen genes in this table (DRC-13) were found in the developmental regulation cluster returned by NetNC-FTI in least seven of the nine TF_ALL datasets.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene symbol	Identifier	Datasets where NetNC-FTI results include the given gene	Brief summary of function
ph-d	FBgn0004860	sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b}, twi_2-4h_intersect	Polycomb Repressive Complex 1 (PRC1) [Shao 1999]
psc	FBgn0005624	sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b}, twi_2-4h_intersect	Polycomb Repressive Complex 1 (PRC1) [Shao 1999]
lolal	FBgn0022238	sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b}	TrxG-related gene, regulates dorso-ventral patterning and is required for silencing at polycomb response elements [Mishra 2003, Quijano 2016]
ph-p	FBgn0004861	sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b}	Polycomb Repressive Complex 1 (PRC1) [Shao 1999]
corto	FBgn0010313	sna_2-4h_Toll^{10b}, twi_2-4h_Toll^{10b}	physically interacts with PcG and TrxG proteins [Salvaing 2003, Lopez 2001]
trithorax	FBgn0003862	twi_2-3h_union, twi_4-6h_intersect	histone methlytransferase (TrxG gene) [Tie 2014, Ingham and Whittle 1980]
su(var)205	FBgn0003607	twi_2-3h_union, twi_4-6h_intersect	histone H3K9 binding, gene silencing factor [Fanti 1998, Fanti 2008]
moira	FBgn0002783	twi_2-3h_union, twi_4-6h_intersect	TrxG gene involved in nucleosome organisation [Hong & Choi 2016, Crosby 1999]
taranis	FBgn0040071	twi_2-3h_union, twi_4-6h_intersect	genetically interacts with TrxG and PcG genes [Schuster 2015, Calgaro 2002, Fauvarque 2001].
su(var)3-9	FBgn0263755 (DroFN ID: FBgn0003600)	twi_2-4h_intersect, twi_2-4h_Toll^{10b}, twi_2-3h_union	histone H3 methylase and gene silencing factor [Czermin 2001, Schotta 2002]

Table S6. Related to Figure 4 and Figure S3. NetNC functionally coherent TF targets found in chromatin organisation clusters across multiple datasets. The genes listed above were identified in NetNC-FTI chromatin organisation clusters for two or more of the TF_ALL datasets.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Dataset	KEGG Pathway ID	Pathway Name	Number of Genes
TRAIN-CL	dme03020	RNA polymerase	26
	dme03430	Mismatch repair	20
	dme04330	Notch signalling	22
	dme04914	Progesterone-mediated oocyte maturation	45
	dme00790	Folate biosynthesis	24
	dme00590	Arachidonic acid metabolism	15
	dme00280	Valine, leucine and isoleucine degradation	15
TEST-CL	dme00510	N-Glycan biosynthesis	36
	dme00760	Nicotinate and nicotinamide metabolism	11
	dme03030	DNA replication	34
	dme04140	Regulation of autophagy	14
	dme04350	TGF-beta signaling pathway	29
	dme04745	Phototransduction - fly	27
	dme04122	Sulfur relay system	9
	dme00600	Sphingolipid metabolism	9

Table S7. Related to Methods. Summary of pathways selected for gold standard data TRAIN-CL and TEST-CL.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

%SNTG	Optimised Z	Optimised Q
5	0.346	0.025
10	0.346	0.025
15	0.3735	0.0400171
20	0.3655	0.051006
25	0.3735	0.0612547
30	0.3615	0.0702808
35	0.3525	0.0834251
40	0.34	0.0920248
45	0.3325	0.0904339
50	0.3005	0.0681088
55	0.2945	0.0543454
60	0.25	0.0423662
65	0.2005	0.0546396
70	0.1355	0.0652494
75	0.2875	0.0991509
80	0.2535	0.2

Table S8. Related to Methods. NetNC-FTI parameter values optimised for percentage Synthetic Neutral Target Genes (%SNTG). This table shows parameter values for NetNC-FTI optimised on TRAIN_ALL. These data may be used to specify NetNC-FTI parameters where an estimate of %SNTG is available, for example derived from NetNC-lcFDR analysis. In general there is a trend toward less stringent values of the optimised parameters (higher Q, lower Z) as the %SNTG increases.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inflation Value	**Matthews Correlation Coefficient**			
	TRAIN-CL10%	TRAIN-CL25%	TRAIN-CL50%	TRAIN-CL75%
2	0.303*	0.318	0.376	0.271
2.2	0.241	0.348	0.269	0.253
2.4	0.217	0.357	0.418*	0.270
2.6	0.186	0.367*	0.402	0.359
2.8	0.174	0.340	0.400	0.365
3	0.161	0.300	0.341	0.343
3.2	0.201	0.323	0.370	0.355
3.4	0.204	0.355	0.384	0.368
3.6	0.196	0.344	0.406	0.393*
3.8	0.196	0.333	0.392	0.390
4	0.193	0.330	0.358	0.339
4.2	0.185	0.319	0.369	0.364
4.4	0.190	0.315	0.330	0.267
4.6	0.185	0.319	0.330	0.305
4.8	0.174	0.304	0.314	0.251
5	0.171	0.290	0.309	0.299

*highest (best-performing) value

Table S9. Related to Methods. Optimisation of MCL Inflation value for the Functional Target Identification task. Results shown represent performance on the TRAIN-CL-SR dataset. The overall best performing inflation value across all four TRAIN-CL datasets examined was 3.6.

bioRxiv preprint doi: https://doi.org/10.1101/455709; this version posted March 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of Supplemental Files

Supplemental_File_1.zip. Related to Figure 4 and Figure S3. Cytoscape sessions with NetNC-FTI results for TF_ALL.

Supplemental_File_2.zip. Related to Figure 6. Primers for RT-qPCR