Functional Transcription Factor Target Networks Illuminate Control of Epithelial Remodelling

Citation: Cancers 12 (10): 2823 (2020)

As Published: http://dx.doi.org/10.3390/cancers12102823

Publisher: Multidisciplinary Digital Publishing Institute

Persistent URL: https://hdl.handle.net/1721.1/131303

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

cancers

MDPI

Article

Functional 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,†], Matthew J. Ford [1,‡], Alexander L. R. Lubbock [1,§], Erola Pairo-Castineira [1] and Abdelkader Essafi [1,‖]

1 MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK; andrew.sims@ed.ac.uk (A.H.S.); bheale@gmail.com (B.S.E.H.); matthew.ford@mail.mcgill.ca (M.J.F.); alex.lubbock@vanderbilt.edu (A.L.R.L.); erola.pairo-castineira@igmm.ed.ac.uk (E.P.-C.); a.essafi@bristol.ac.uk (A.E.)
2 Department of Systems Biology, Harvard University, Boston, MA 02115, USA; jaowen@mit.edu
3 Centre for Synthetic and Systems Biology (SynthSys), University of Edinburgh, Edinburgh EH9 3BF, UK
4 Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast BT9 7AE, UK
5 Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
* Correspondence: i.overton@qub.ac.uk
† Current address: Intermountain Healthcare, 3930 River Walk, West Valley City, UT 84120, USA.
‡ Current address: Rosalind & Morris Goodman Cancer Research Centre, McGill University, Montreal, QC H3A 0G4, Canada.
§ Current address: Department of Biochemistry, Vanderbilt University, Nashville, TN 37232, USA.
‖ Current address: School of Cellular and Molecular Medicine, University of Bristol, Bristol BS8 1TD, UK.

Received: 9 August 2020; Accepted: 24 September 2020; Published: 30 September 2020

Abstract: Cell identity is governed by gene expression, regulated by transcription factor (TF) binding at cis-regulatory modules. Decoding the relationship between TF binding patterns and gene regulation is nontrivial, remaining a fundamental limitation in understanding cell decision-making. We developed the NetNC software to predict functionally active regulation of TF targets; demonstrated on nine datasets for the TFs Snail, Twist, and modENCODE Highly Occupied Target (HOT) regions. Snail and Twist are canonical drivers of epithelial to mesenchymal transition (EMT), a cell programme important in development, tumour progression and fibrosis. Predicted "neutral" (non-functional) TF binding always accounted for the majority (50% to 95%) of candidate target genes from statistically significant peaks and HOT regions had higher functional binding than most of the Snail and Twist datasets examined. Our results illuminated conserved gene networks that control epithelial plasticity in development and disease. We identified new gene functions and network modules including crosstalk with notch signalling and regulation of chromatin organisation, evidencing networks that reshape Waddington's epigenetic landscape during epithelial remodelling. Expression of orthologous functional TF targets discriminated breast cancer molecular subtypes and predicted novel tumour biology, with implications for precision medicine. Predicted invasion roles were validated using a tractable cell model, supporting our approach.

Keywords: network biology; ChIP-seq; breast cancer; transcription factors; EMT; functional gene network; mesoderm; *Drosophila melanogaster*; gene regulation; epithelial remodelling

1. 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 [1–3]. TF action is controlled

in multiple ways; including protein–protein interactions, DNA sequence affinity, 3D chromatin conformation, post-translational modifications and the processes required for TF delivery to the nucleus [3–5]. A complex interplay of mechanisms influences TF specificity across different biological contexts and genome-scale assignment of TFs to individual genes is challenging [1,5,6].

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 [1,7–9]. Genomic regions that bind large numbers of TFs, termed Highly Occupied Target (HOT) regions [10], are enriched for disease SNPs and can function as developmental enhancers [11,12]. 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 organization [13,14], as well as possible technical artefacts [15]. Apparent neutral binding events may also have subtle functions; for example, in combinatorial context-specific regulation or in buffering noise [2,16]. While recent integrative work enhances context-independent TF target prediction [17], identification of bona fide functional TF target genes remains a major obstacle in understanding the regulatory networks that control cell behaviour [2,5,9,18,19].

Genes regulated by an individual TF typically have overlapping expression patterns and coherent biological function [20–22]. Indeed, gene regulatory networks are organised in a hierarchical, modular structure and TFs frequently act upon multiple nodes of a given module [23,24]. Therefore, we hypothesised that the network properties of functional TF targets are different to those of neutrally bound sites. Network analysis can reveal biologically meaningful gene modules, including cross-talk between canonical pathways [25–27] 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 [28,29], and are not restricted to predefined gene groupings used by standard functional annotation tools (e.g., GSEA, DAVID) [25,30,31]. Clustering is frequently applied to define biological modules [32,33]. However, pre-defined modules may miss condition-specific features; for example, gene products may be absent in the biological condition(s) analysed but included in pre-defined network modules. 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 [34]. We developed an algorithm (NetNC) with capability for context-specific functional TF target discovery and applied this to study the epithelial to mesenchymal transition (EMT) TFs Snail and Twist. EMT is a multi-staged morphogenetic programme fundamental for normal embryonic development that contributes to tumour progression and fibrosis [35–38]. We predicted Snail and Twist functional targets, integrating these predictions with results from genetic screens and breast cancer transcriptomes; in order to study epithelial remodelling in development and disease.

2. Results and Discussion

2.1. A Comprehensive Drosophila melanogaster Functional Gene Network (DroFN)

Our approach requires a genome-scale map of gene function; for this purpose we developed the DroFN network (11,432 genes; 787,825 interactions). DroFN models *Drosophila melanogaster* signalling and metabolism, integrating the Gene Ontology (GO) [39] and STRING [40] databases, calibrated against KEGG pathways using Bayesian Logistic Regression [41]. DroFN performed well on blind test data (TEST-NET) compared with other *D. melanogaster* gene networks (DroID [42], GeneMania [43]) (Table S1, Figure S1). The positive class in TEST-NET was formed from gene pairs within the same KEGG pathway and the negative class was derived from gene pairs with no evidence for a pathway

interaction. The overlap between DroFN and the *Drosophila* proteome interaction map (DPiM [44]) was highly significant (Fisher Exact Test $p < 10^{-308}$). DroFN and DPiM had 999 genes in common and 37.8% (2175/5747) of DroFN interactions for these genes were also found in DPiM. The DroFN false positive rate (0.043) was close to the prior expectation for functional gene interactions (0.044); thus, a proportion of these apparent false positives might represent bona fide interactions missing from the gold standard KEGG pathways. Overall, DroFN provides a comprehensive, high-quality representation of pathway comembership in *D. melanogaster*.

2.2. Prediction of Functional Transcription Factor Targets

We present NetNC, an algorithm for genome-scale prediction of functional TF target genes (Figure 1). NetNC built upon observations that TFs coordinately regulate multiple functionally related targets [20–22] and was calibrated for discovery of biologically coherent genes in noisy data, according to the structure of a gene network. This approach required optimisation for elimination of noise in biological data, rather than for community detection. Statistical evaluation of network coherence for an input gene list, including false discovery rate (FDR) estimation, is applied within NetNC as basis for numerical thresholding. Therefore, NetNC can analyse single-subject datasets, which is an important emerging area for precision medicine [45]; for example, to derive networks from genes with high or low relative expression according to ranked expression values from a single sample. Application of statistical and graph theoretic methods in NetNC for quantitative evaluation of relationships between genes offers an alternative to the classical emphasis on individual genes in studying the relationship between genotype and phenotype.

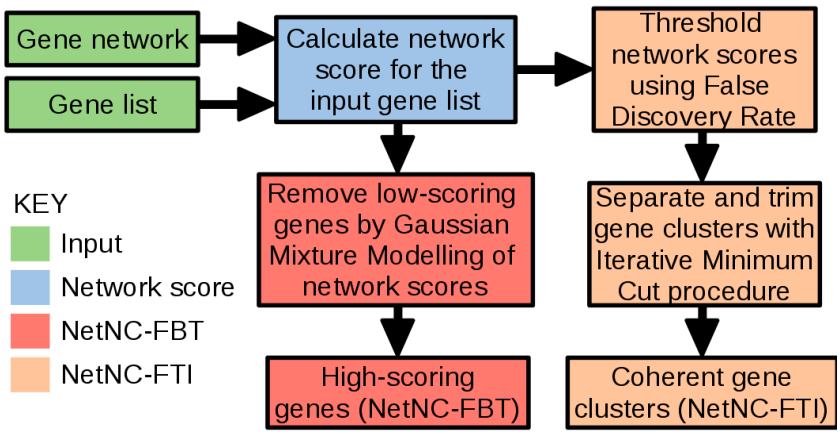

Figure 1. Overview of the NetNC algorithm. A gene network and a gene list are required by NetNC as input (green). The current study analyses candidate TF targets; however, NetNC could be applied to any gene list and network. NetNC calculates a network score (blue) using Hypergeometric Mutual Clustering (HMC) for each gene pair in the input gene list according to connections in the network. Two analysis settings are (a) NetNC-FBT (red), where Gaussian Mixture Modelling identifies high-scoring genes; (b) NetNC-FTI (orange), which produces coherent gene clusters by thresholding network scores according to False Discovery Rate followed by Iterative Minimum Cut.

NetNC has two different analysis settings, NetNC-FTI ("Functional Target Identification") and NetNC-FBT ("Functional Binding Target"), shown in Figure 1. Biological similarity between gene pairs is represented in NetNC using shared network neighbours, formalised by the Hypergeometric Mutual Clustering coefficient [46]; further analysis steps then enable prediction of functional TF targets (Figure 1). The current study reports results from NetNC with DroFN as a reference network, however NetNC may be used to analyse any network and node list. We chose the DroFN network because of its favourable performance in comparisons against DroID and GeneMania (Table S1, Figure S1). In order to assess NetNC performance and to calibrate algorithm parameters, we developed gold standard data

using KEGG pathways and "Synthetic Neutral Target Genes" (SNTGs). Clustering coefficient (CC) values predicted by models trained on the synthetic benchmark data (methods Section 3.7) matched the CC values calculated directly on the nine TF_ALL datasets, which are described in methods Section 3.11. For 8/9 datasets the difference between predicted and actual CC was <0.1 (median CC difference = 0.051, 95% CI 0.007–0.136). This similarity in CC values for the synthetic and biologically-derived candidate TF target genes supports the application of our benchmark in the context of network-based functional TF target prediction. NetNC was robust to variation in input dataset size and %SNTGs, outperforming the clustering algorithms HC-PIN [33] and MCL [32] (Figure 2, Table S2). In general, NetNC was more stringent, with lower false positive rate (FPR) and higher Matthews Correlation Coefficient (MCC). MCC provided a measure of overall performance in correctly separating functional targets from neutral binding (SNTGs). FPR indicated the proportion of predicted functional targets that were SNTGs and therefore classified incorrectly. At the highest %SNTG, NetNC-FTI overall performance (MCC) was around 50% to 67% better than HC-PIN, and NetNC-FBT typically had lowest FPR. The task of separating SNTGs from all of the genes that form pathways is subtly different to cluster identification; thus neither HC-PIN nor MCL were developed for the precise application evaluated here. A number of network-based analysis method such as HotNet2 [47] and PrixFixe [48] were developed for very different application areas, for example requiring mutational frequency data, and so were not suitable for application to functional TF target discovery. While NetNC-FTI performed best overall, NetNC-FBT is parameter-free, did not require training and therefore may be more robust for analysis of diverse input datasets and reference networks other than DroFN. NetNC's performance advantages were most prominent on data with ≥50% SNTGs (Figure 2) and predicted neutral binding for TF_ALL was ≥61% (Figure 3, Table 1) or ≥50% using an alternative calibration method (Figure S2).

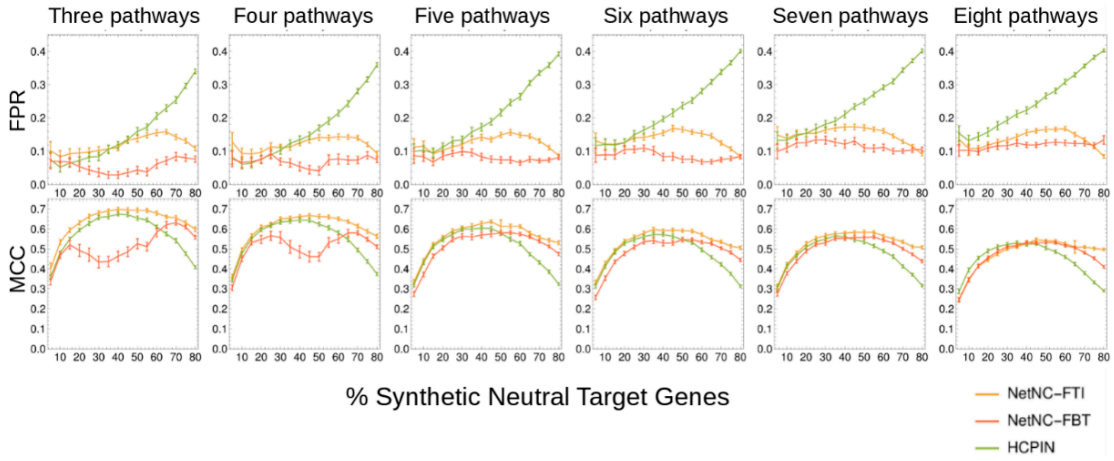

Figure 2. Evaluation of NetNC and HC-PIN on blind test data. Performance values reflect discrimination of KEGG pathway nodes from Synthetic Neutral Target Genes (STNGs), shown for NetNC-FTI (orange), NetNC-FBT (red) and HC-PIN (green). False positive rate (FPR, top row) and Matthews Correlation Coefficient (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 false positive rate and higher MCC values. Error bars reflect 95% confidence intervals calculated from quantiles of SNTG resamples. Also see Table S2.

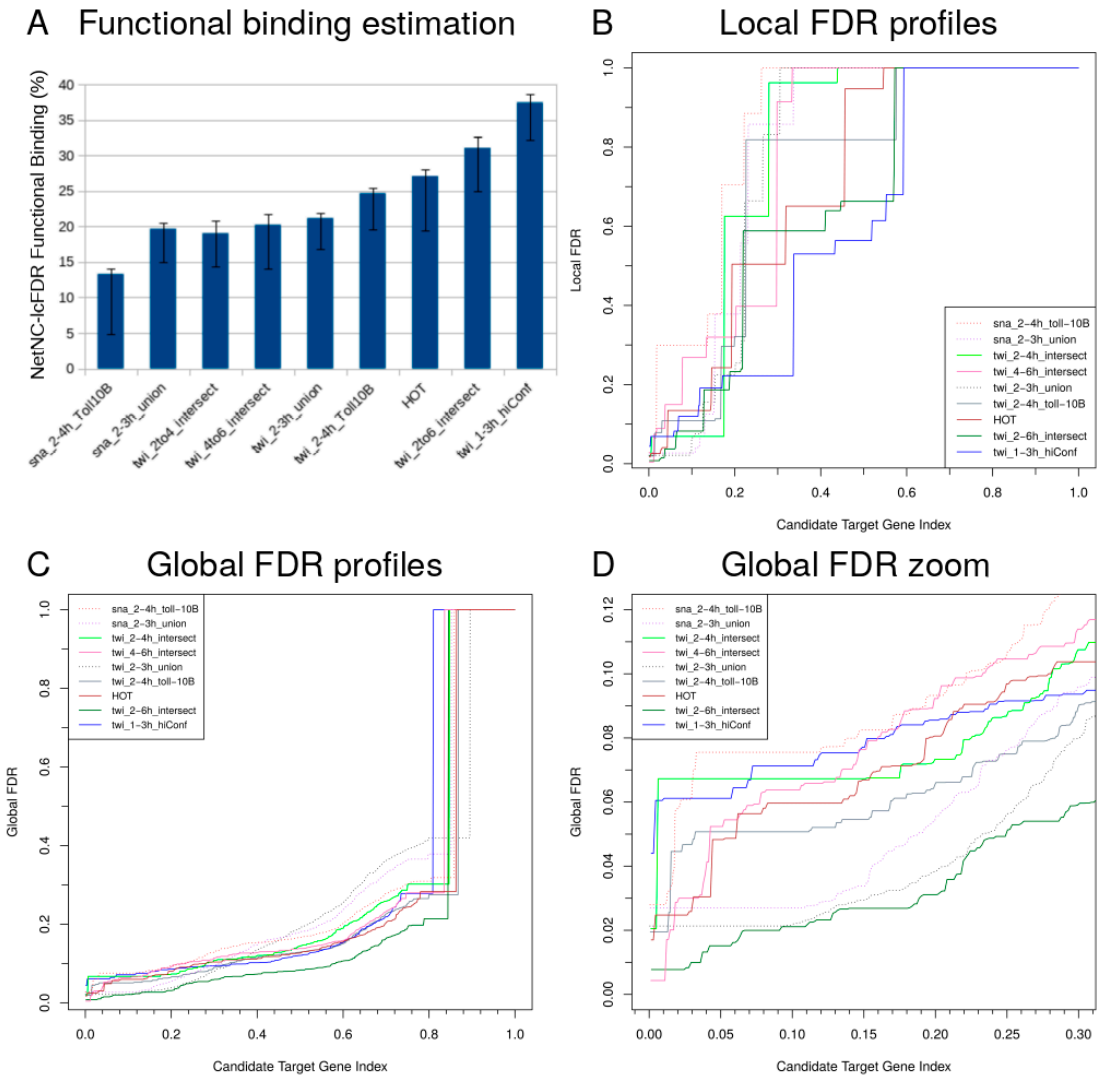

Figure 3. Functional transcription factor binding and false discovery rate (FDR) profiles. (**A**) Estimation of total functional binding. Median values are shown for NetNC-lcFDR, calibrated against resampled genes. Error bars represent 95% CI calculated using quantiles of results for the resampled datasets, which gave predicted functional binding ranging from 5% (sna_2–3h lower CI) to 39% (twi_1–3h_hiConf upper CI) across TF_ALL. Calibration based on synthetic data resulted in slightly higher functional binding estimates, up to 50% (Figure S2). (**B**–**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.

Given the performance advantage at ≥50% STNGs, NetNC appears the method of choice for analysis of genome-scale TF occupancy data. However, NetNC may also be applied to analyse various data-types, including in: 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 [49]; and discovery of functional coherence in chromosome conformation capture data (4-C, 5-C), for example in enhancer regulatory relationships [50,51]. We also compared NetNC against the NEST algorithm [52] and against node degree (Table S3). The NEST output required filtering and did not provide a threshold for separation of predicted neutral binding from functional targets, therefore we compared Area Under the Receiver Operator Characteristic curve (AUC). There was a trend towards NetNC having better performance than the filtered NEST results; predictions that took either node degree or the unfiltered NEST output

had substantially lower AUC values than NetNC. NetNC results were robust to subsampled input gene lists (Table S4). Gold standard datasets and DroFN are available from the BioStudies database: www.ebi.ac.uk/biostudies/studies/S-BSST460. NetNC is available from https://github.com/overton-group/NetNC.

Table 1. Predicted functional binding for Snail, Twist and Highly Occupied Target (HOT) candidate target genes. The developmental time periods correspond to the following developmental stages: 2–4 h stages 4–9 (except 2–4 h_intersect datasets which were stages 5–7 [53]); 2–3 h stages 4–6; 1–3 h stages 2–6; 4–6 h stages 8–9 [53]; 0–12 h stages 1–15. Also see Figure S3.

Dataset				Predicted Functional Targets ‡	
Name	**Developmental Time Period(s)**	**Total Candidate Target Genes ***	**Candidate Target Genes in DroFN**	**NetNC-FTI**	**NetNC-lcFDR (95% CI)**
twi_1–3h_hiConf	1–3 h	755	664	202 (30%)	37% (32–39%)
twi_2–6h_intersect	2–4 h and 4–6 h	743	615	241 (39%)	31% (25–33%)
twi_2–4h_intersect	2–4 h only (not 4–6 h)	1028	801	182 (23%)	19% (14–21%)
twi_4–6h_intersect	4–6 h only (not 2–4 h)	1026	818	126 (15%)	20% (14–22%)
HOT	0–12 h $^{+}$	1648	677	174 (26%)	27% (19–28%)
twi_2–3h_union	2–3 h	2285	1848	424 (23%)	21% (17–22%)
sna_2–3h_union	2–3 h	1424	1158	226 (20%)	20% (15–21%)
twi_2–4h_Toll^{10b}	2–4 h	1578	1238	279 (23%)	25% (20–25%)
sna_2–4h_Toll^{10b}	2–4 h	1822	1488	211 (14%)	13% (5–14%)

* mapped to FlyBase; ‡ for candidate target genes in DroFN; $^{+}$ multiple time periods and 41 different transcription factors (TFs).

2.3. Analysis of EMT Transcription Factors and Highly Occupied Target (HOT) Regions

We predicted functional target genes for the Snail and Twist TFs in developmental stages around *D. melanogaster* gastrulation. Chromatin ImmunoPrecipitation (ChIP) microarray (ChIP-chip) or sequencing (ChIP-seq) data from four different laboratories were analysed for overlapping time periods in early embryogenesis [8,22,53,54]. Two of these datasets were from studies examining the regulatory networks that control dorsoventral patterning, including mesoderm development [53,54]. Further Snail and Twist binding data was available from a comparatively large investigation of patterning in early fly development by the Berkley Drosophila Transcriptional Network Project, which analysed 21 TFs including dorsoventral factors [22]. We also obtained data from more recent work that compared ChIP-seq to other technologies for genome-scale analysis of gene regulation, including identification of new Twist-occupied regions and a consensus motif analysis [8]. An additional dataset examined TF binding hotspots termed "Highly Occupied Target" (HOT) regions that were annotated by the modENCODE project [10]. Overall, nine main datasets were studied (TF_ALL, Table 1; please see Methods Section 3.11 for further details), the proportion of predicted functional TF binding for these datasets ranged from 5% to 39% (Figure 3A, Table 1). At the time of writing, the above datasets remain leading sources of information for Snail and Twist binding in early mesoderm development. NetNC does not perform peak calling—but is intended for downstream analysis of candidate target genes from statistically significant peaks, in order to enable separation of predicted functional targets from neutral binding. In addition to the predictions from NetNC-FTI, we developed a complementary approach to estimate the total functional binding in each TF_ALL dataset; this method was based on local FDR (NetNC-lcFDR; Methods, Section 3.4).

Reassuringly, candidate TF targets from the most stringent peak calling approach (twi_1–3h_hiConf [8]) had comparatively high predicted functional binding (PFB). Despite having high PFB, twi_1–3h_hiConf had the smallest proportion of genes passing pFDR < 0.05 (Figure 3D). High PFB was also found for targets bound during two consecutive developmental time periods (twi_2–6h_intersect [53]), as well as for HOT regions despite their lack of known TF motifs [11,55,56] (Figure 3A, Table 1). Indeed, twi_2–6h_intersect had significantly higher PFB (binomial $p < 4.0 \times 10^{-15}$)

than datasets from the same study that represented a single time period (twi_2–4h_intersect, twi_4–6h_intersect) [53] (Figure 3). Therefore, PFB 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 [2,5,57–59]. PFB 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 [53]. However, taking the union of candidate binding sites could eliminate false negatives arising from epitope steric occlusion due to protein interactions. The similarity in PFB 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 [22,53]. 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 recovered at lcFDR < 0.05 or pFDR < 0.05 (Figure 3). Even datasets with low PFB 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 have contributed to the low PFB for sna_2–4h_Toll^{10b}. We note that our analysis 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. Predicted neutral targets for twi_2-4_intersect, twi_2-6_intersect and twi_4-6_intersect were overwhelmingly unchanged in *Twist* loss-of-function gene expression data from the same study [53] (respectively 96–97%, 93–94%, 90–91%, and 93–95% were unchanged at the 4–5 h, 5–6 h, 6–7 h, and 7–8 h time points; $q < 0.05$, 1.5 FC). We also note that NetNC-lcFDR and NetNC-FTI neutral binding estimates showed good agreement (Table 1, Figure S3).

ChIP peak intensity putatively correlates with functional binding, although some weak binding sites are functional [9,60]. The NetNC Node Functional Coherence Score (NFCS) and ChIP peak enrichment scores were significantly, although weakly, correlated in 6/8 datasets ($q < 0.05$, HOT regions not analysed; median rho = 0.11). Datasets with no significant correlation (twi_1–3h_hiConf, twi_2–6h_intersect) derived from protocols that enriched for functional targets and had highest PFB (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, $q < 5.0 \times 10^{-56}$; twi_4–6h_intersect, $q < 4.8 \times 10^{-58}$). The number of orthologues for each dataset correlated strongly with the number of predicted functional targets (r = 0.973). However, sna_2–3h_union and twi_2–3h_union functional targets had significantly higher proportion of orthologues (>80%) than the next highest dataset (twi_2–4h_intersect, 67%; respective binomial test $q < 3.9 \times 10^{-4}$, $q < 1.2 \times 10^{-11}$), which might be explained by the use of RNA polymerase binding data in assigning candidate target genes. NetNC-FTI predictions were enriched for human orthologues relative to respective candidate target genes in DroFN (Table S5), for example predicted twi_2–3h_union functional targets had 82% (347/424) human orthology vs. 61% (1135/1848) for the pool of candidate targets (binomial $q < 8.7 \times 10^{-19}$). Annotation bias might contribute something to this significant enrichment, because conserved genes are more deeply studied and may associated with a higher degree in the DroFN network; high degree is also expected for conserved genes because of their functional importance [61]. We note that enrichment for evolutionary conservation in Snail and Twist functional targets aligned with their regulation of fundamental developmental processes [35,36].

2.4. Genome-Scale Functional Transcription Factor Target Networks

NetNC results offered a global representation of tissue-specific regulation by Snail and Twist in early *D. melanogaster* embryogenesis (Figure S4, Data File S1). Results revealed 11 biological groupings common to ≥4/9 TF_ALL datasets (Table S6). We found Snail and Twist regulation of multiple core cell processes that govern the global composition of the transcriptome and proteome, including: transcription, chromatin organisation, RNA splicing, translation and protein turnover. These predicted regulatory events may contribute to either repression or activation of individual genes in the (presumptive) mesoderm. A "Developmental Regulation Cluster" (DRC) was identified in every TF_ALL dataset and contained members of multiple key conserved morphogenetic pathways, including notch and wnt. We examined predicted functional targets that were found in the DRC and chromatin organisation clusters for multiple TF_ALL datasets (Figure 4A); *Wingless* had the highest degree and strongest edges in the combined network; *Notch* and *forkhead* had joint second highest frequency, represented in 8/9 NetNC-FTI results for TF_ALL. Many of the DRC genes were previously reported to be important for mesoderm development [53,62] and their interactions suggest functional relationships. For instance, genes that interact with *T48* might contribute to *fog*-independent ventral furrow formation [63]. The edge between *Notch* and *wingless* was identified most frequently in the combined DRC network. *Notch* signalling modifiers identified in at least two public datasets [64] overlapped significantly with the overall NetNC results for each TF_ALL dataset (q <0.05), including members of the DRC, chromatin organisation and mediator complex clusters (Figure 4, Figure S4). Activation of *Notch* can result in diverse, context-specific transcriptional outputs and the mechanisms regulating this pleiotropy are not well understood [64–67]. 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* or *snail* was previously shown in multiple systems, for example in adult myogenic progenitors [68] and hypoxia-induced EMT [69]. Consistent with previous studies [64,65], our results predicted 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 (Figure S4). For example, the mediator complex and transcription initiation subcluster for twi_2–3h_union had 13 nodes, of which five were *Notch* modifiers including orthologues of *MED7*, *MED8*, and *MED31* (Data File S1). These results highlight key control points regulated by Snail, Twist in fly mesoderm specification; including *wingless*, *forkhead*, and *Notch*. Thirteen DRC genes were present in ≥7 TF_ALL datasets (DRC-13, Table S7), and had established functions in the development of mesodermal derivatives such as muscle, the nervous system and heart [66,68,70–74]. Supporting NetNC predictions, in situ hybridisation for DRC-13 genes indicated 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*) [75–77]. The remaining two DRC-13 genes had no evidence for mesodermal expression (*fkh*) or no data available (*jar*). However, *fkh* was essential for caudal visceral mesoderm development [78] and *jar* was expressed in midgut mesoderm [79].

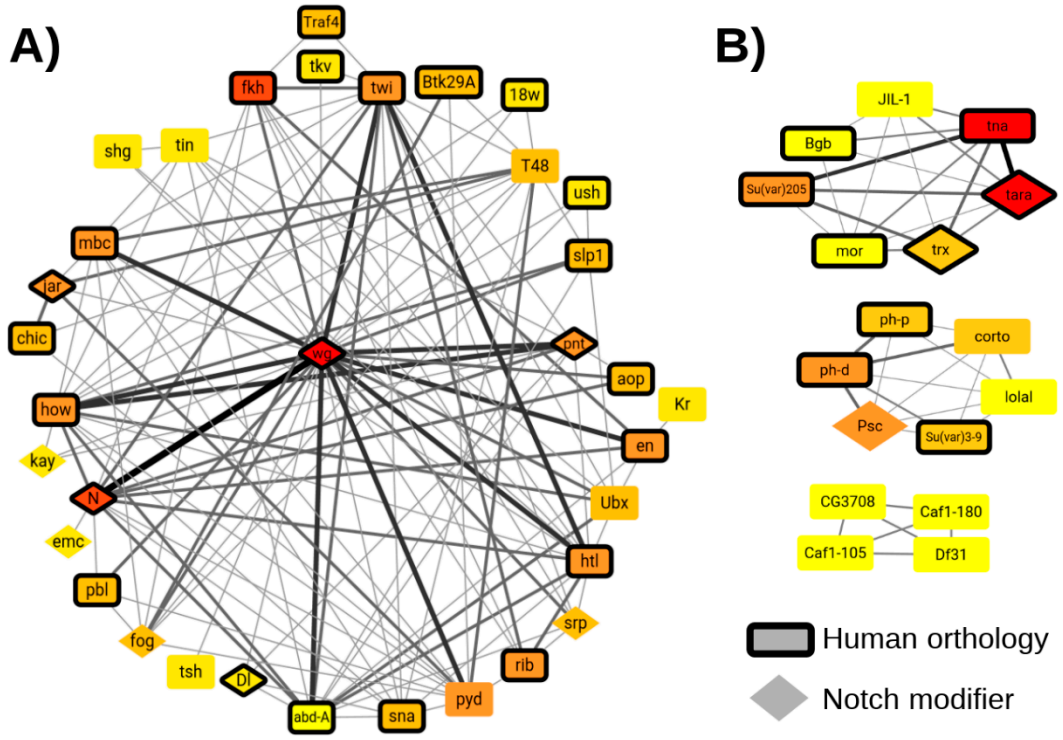

Figure 4. Clusters of developmental regulation and chromatin organisation genes identified by NetNC in multiple TF_ALL datasets. These clusters visualise the combined NetNC-FTI results across the nine TF_ALL datasets; node fill colour, edge width and edge colour indicate frequency of occurrence. Notch modifiers (diamonds) were reported in at least two screens and genes with InParanoid human orthology are shown with black borders. NetNC-FTI clusters for the individual TF_ALL datasets are shown in Figure S4 and are available in Cytoscape format (Data File S1). (**A**) Developmental regulation cluster genes in at least five (yellow), up to nine (red) datasets. Edges shown were in ≥5/9 datasets, up to a maximum of 8/9 for *N* and *wg*. Thirteen genes (DRC_13) were present in ≥7/9 datasets, including *wg* (9/9) which had highest degree, *N* (8/9) and *fkh* (8/9). (**B**) Genes in chromatin organisation clusters from two (yellow) up to six (red) TF_ALL datasets. The three clusters were associated with trithorax-group (top), polycomb group (middle), and chromatin assembly factors (bottom). These results predict components regulated by Snail and Twist in establishing the chromatin blueprint for mesodermal lineages.

Networks produced by NetNC-FTI for each of the nine TF_ALL datasets frequently included chromatin organisation clusters (Figure S4 and Table S6); recurrently identified nodes from these clusters corresponded to trithorax-group (TrxG) and polycomb-group (PcG) genes which exert dynamic, opposing gene-regulatory activity [80] (Figure 4B). The PcG cluster had Polycomb Repressive Complex 1 (PRC1) genes *ph-d*, *ph-p*, and *Psc* [81], the gene silencing factor S*u(var)3–9* [82], as well as *corto* [83] and *lolal* [84] (Figure 4B, Table S8). These results predict that *corto* and *lolal* function in concert with core PRC1 members in *D. melanogaster* mesoderm development under the control of Twist and Snail. Indeed, interaction with accessory proteins enables context-specific PRC1 function, and would merit further study in *Drosophila* [80]. The TrxG cluster contained *tara* [85], *trx* [86], *tna* [87], *mor* [88], s*u(var)205* [89], *Bgb* [90], and *JIL-1* [91]. *Tara* and *tna* were predicted functional targets in 6/9 datasets, interact with each other and with the *brahma* chromatin remodelling complex [87]. A third cluster was formed from chromatin assembly factors, including the poorly characterised gene CG3708 that is orthologous to the nucleosome assembly protein NAP1L1. PcG genes are crucial oncofetal regulators and the focus of significant cancer drug development efforts [92,93]. These results align with reports that gene silencing in EMT involved PcG [93,94] and with *Snai1* recruitment of Polycomb Repressive Complex 2 members [94], supporting a model where EMT TFs control the expression of their own coregulators.

Snail regulation of neural genes (Table S6, Figure S4) was consistent with its repression of ectodermal (neural) genes in the prospective mesoderm [62,95,96]. Indeed, clusters relating to brain development were found in six TF_ALL datasets. Additionally, Snail is important for neurogenesis in fly and mammals [97,98]. Therefore, binding to neural functional modules might reflect potentiation for rapid activation in combination with other transcription factors within neural developmental trajectories [53,99]. NetNC results predicted novel Twist functions, for example in activation or repression of mushroom body neuroblast proliferation factors Rx, sle, and tara. The mushroom body is a prominent structure in the fly brain, important for olfactory learning and memory [100], identified in analysis of six TF_ALL datasets (Table S6). Twist is typically a transcriptional activator [96] although may contribute to Snail's repressive activity [101]. Indeed, TWIST1 repressed Cadherin-1 in breast cancers [102].

2.5. Breast Cancer Subtypes are Recovered by Unsupervised Clustering with Orthologous Snail and Twist Functional Targets

We analysed the conserved molecular networks that orchestrate epithelial remodelling in development and cancers by combining NetNC results for TF_ALL with the results of *Notch* screens and breast cancer transcriptomes; data integration was based on identifier matching and orthology mapping (methods). Predicted Snail and Twist targets included known cancer genes and also suggested novel drivers (Figure S4, Table S5, Table S7 and Table S8). 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 [103]. Many developmentally patterned fly genes are orthologues of established cancer drivers. Breast cancer intrinsic molecular subtypes with distinct clinical trajectories were extensively validated and complement clinico-pathological parameters [104,105]. These subtypes are known as luminal-A, luminal-B, HER2-overexpressing, normal-like, and basal-like. While more recent studies have classified further subtypes, for example identifying ten groups [106], 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 [104,105]. The NetNC-FTI networks for all nine TF_ALL datasets overlapped with known cancer pathways, including significant enrichment for *Notch* modifiers ($q < 0.05$). We hypothesised that orthologous genes from NetNC clusters for Snail and Twist would stratify breast cancers by intrinsic molecular subtype. Indeed, aberrant activation of *Notch* orthologues in breast cancers had been demonstrated, and linked with EMT-like signalling, particularly for basal-like and claudin-low subtypes [107–109]. One might expect the predicted Snail and Twist functional targets to be prognostic in multiple different cancers due to the representation of established cancer cell processes, for example DNA replication and repair as well as developmental regulation (Figure S4, Table S6).

We hypothesised that orthologous genes from Snail and Twist functional targets would stratify breast cancers into clinically meaningful groups. Sixty-eight DroFN genes were predicted functional targets in four or more of the nine TF_ALL datasets and also had human orthology. Fifty-seven of these sixty-eight genes (ORTHO-57) were represented in integrated gene expression microarray data for 2999 breast tumours (BrC_2999) [110]. Taking a threshold of at least four datasets reduced the gene list size used for clustering in order to help prevent stratification effects arising from autocorrelation [111]. Unsupervised clustering using ORTHO-57 and also NetNC results for individual Twist, Snail datasets stratified BrC_2999 by intrinsic molecular subtype (Figure 5, Figure S5). Predicted functional targets had significantly higher centroid values than equivalent resampled orthologues from DroFN, demonstrating that the observed stratification of breast cancer subtypes was significantly greater than expected by chance (Figure S5). Previous work suggested roles in tumour biology for many ORTHO-57 genes; however, few were linked to an intrinsic breast cancer subtype. Searching PubMed with the gene name(s) and "breast cancer" found little evidence of function in breast cancer invasion for 13/57 genes (ADSS, CREG1, ATP5A1, SRSF2, SNRPD1, RNPS1, TEC, HIVEP3, SERTAD2,

NACC2, GULP1, IRX4, and TRIB2). Heatmap features were annotated as dashed black boxes according to the dendrogram structure and gene expression intensity (Figure 5). The datasets sna_2–4h_Toll^{10b}, twi_2–4h_Toll^{10b} represented embryos formed entirely from mesodermal lineages [54] and, together, had significantly greater proportion of basal-like breast cancer genes than the combined sna_2–3h_union, twi_2–3h_union datasets ($p < 8.0 \times 10^{-4}$); consistent with EMT characteristics of basal-like breast cancers [112]. As expected, Basal-like tumours were characterised by *EN1* and *NOTCH1* [107,108,113]. Notch signalling modulation is a promising area for cancer therapy [65] and *Notch* modifier orthologues from our analysis could potentially inform development of companion diagnostics or combination therapies 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 [114,115]. 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 the major defining feature for luminal A tumours, aligning with oncogenic BMP signalling in luminal epithelia [116]. However, BMP2 expression was highest in basal-like cancers, where it may drive an EMT programme [117]. Several genes were highly expressed in both Luminal B and *ESR1* negative subtypes (feature_LumB$_1$, feature_ERneg) including *ECT2*, *SNRPD1*, *SRSF2*, *CBX3*; these genes might contribute to worse survival outcomes for luminal B relative to luminal A cancers [104,118]. Indeed, analysis with the GEPIA2 resource [119] revealed that these four genes stratified breast cancer patients by overall survival in an independent cohort from The Cancer Genome Atlas [120], where high expression conferred worse prognosis (Figure S6). Feature_LoExp represented genes with low detection rates across a mixture of subtypes, largely from a single study [121]. Key EMT genes (*SNAI2*, *TWIST1*, and v*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*). Genes with high expression in both normal-like and basal-like cancers included *QKI*, which regulates circRNA formation in EMT [122], and the *FZD1* wnt/β-catenin receptor. Moreover, genes in feature_Bas and feature_NL clustered together, identifying similarities between normal-like and basal-like subtypes. EMT may confer stem-like cell properties [123–125] 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 NL cancers [118,126], indeed *SNAI2* was important in both mammary and breast cancer stem cells [127,128]. However, high stromal content in NL tumours [129] might also contribute to an EMT-like gene expression signature. In summary, the predicted functional TF targets from fly mesoderm development captured clinically relevant molecular features of breast cancers and proposed candidate subtype-specific drivers of tumour progression.

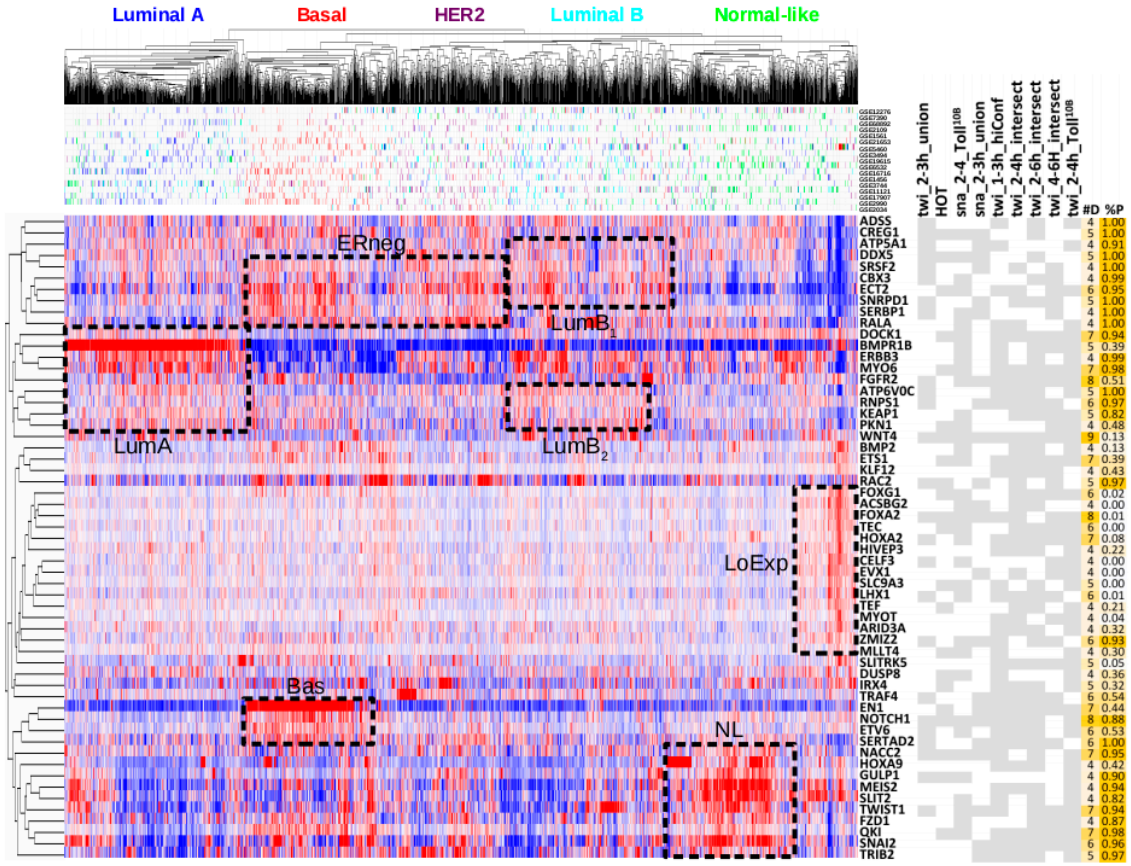

Figure 5. Predicted functional transcription factor targets capture human breast cancer biology. The heatmap shows gene expression in 2999 primary breast tumours for ORTHO-57 genes (red = high, white = mean, blue = low). The mosaic above the heatmap indicates intrinsic molecular subtype: Luminal A (blue), basal-like (red), HER2-overexpressing (purple), luminal B (light blue), and normal-like (green). Annotated heatmap features (black dashed lines) identified genes upregulated in one or more intrinsic subtype; "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) across TF_ALL; the number of datasets where the gene was identified by NetNC-FTI (#D) and the percentage of present calls across the 2999 tumours (%*p*) are shown. The LoExp feature corresponded overwhelmingly to genes with low %*p* values and to samples from a single study [121]. Some genes were annotated to more than one feature and reciprocal patterns of gene expression were found. For example, *BMPR1B*, *ERBB3*, and *MYO6* were strongly upregulated in feature LumA but downregulated in basal-like and *HER2*-overexpressing cancers. Unexpectedly, feature NL (normal-like) had high expression of canonical EMT drivers, including *SNAI2*, *TWIST*, and *QKI*. Some of the EMT genes in feature NL were upregulated in many basal-like tumours, while genes in feature Bas (*NOTCH*, *SERTAD2*) had relatively high expression in normal-like tumours. Also see Figure S5.

2.6. 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 S4). Interestingly, key EMT genes were assigned to the normal-like subtype, which was also associated with splicing factors, the ribosome, the proteasome and proteasome regulatory subunits. The sna_2–4h_$Toll^{10b}$ "RNA degradation and transcriptional regulation" cluster was exclusively annotated to the basal-like subtype, including *HECA*, which was upregulated in basal-like relative to normal-like tumours ($p < 3.3 \times 10^{-23}$). NetNC also identified the fly orthologue of *HECA*, *hdc*, in both twi_2–4h_intersect and

twi_4–6h_intersect, bound at non-contiguous sites. *Hdc* was a multifunctional *Notch* signalling modifier, including in cell survival [130] and tracheal branching morphogenesis, activated by *escargot* [131]. 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 [132]. *SLC9A6* was upregulated in basal-like relative to normal-like tumours ($p < 8.4 \times 10^{-71}$) and might cause pH dysregulation as part of an EMT-like programme in basal-like cancers.

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 that had copy number amplifications in basal-like tumours [133]; the *SERTAD2* bromodomain interacting oncogene and E2F activator [134]; and *SMARCE1*, a core subunit of the SWI/SNF chromatin remodelling complex that regulated *ESR1*, interacted with *HIF1A* signalling and potentiated breast cancer metastasis [135–137]. Notch can promote EMT-like characteristics and mediated hypoxia-induced invasion in multiple cell lines [69]. Consistent with these studies, our work supported 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 controlled chromatin switching in oral cancer EMT [138]. *Taranis*, orthologous to *SERTAD2*, also functioned to stabilise the expression of *engrailed* in regenerating tissue [85]. The *engrailed* orthologue *EN1* was the clearest single basal-like biomarker in the data examined (Figure 5) and acted as a survival factor [113]. *SERTAD2* and *EN1* expression values correlated in the basal-like tumours ($n = 573$, $p < 2.0 \times 10^{-9}$, rho = 0.25) but not across the entire cohort ($n = 2999$, $p = 0.44$, rho = −0.03). Our results suggest that *SERTAD2* could cooperate with *EN1* in a subset of basal-like cancers, where coordinated expression of these two genes may form 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 [112,113]. Therefore, our results highlight 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 predicted 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 [35]; our results mapped networks that may control the remodelling of Waddington's landscape—identifying crosstalk between Snail, Twist, epigenetic modifiers and regulation of key developmental pathways [139]. Dynamic interplay between successive cohorts of TFs and chromatin organisation factors is an attractive potential mechanism to determine progress through and the ordering of steps in (partial) EMTs, consistent with "metastable" intermediate stages.

2.7. 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. We investigated the functional and instructive role of four genes in an established invasion model [140]; *SNX29* (also known as *RUNDC2A*), *ATG3*, *IRX4*, and *UNK*. These genes were selected for experimental follow-up from the large pool of NetNC results for TF_ALL according to novelty in the context of cell invasion and EMT. The orthologues of *IRX4* and *SNX29* were predicted to be regulated by both Snail and Twist, the orthologue of *ATG3* was predicted to be a functional target of Twist; and the orthologue of *UNK* was predicted to be regulated by Snail. MCF7 cells were weakly invasive [141], 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).

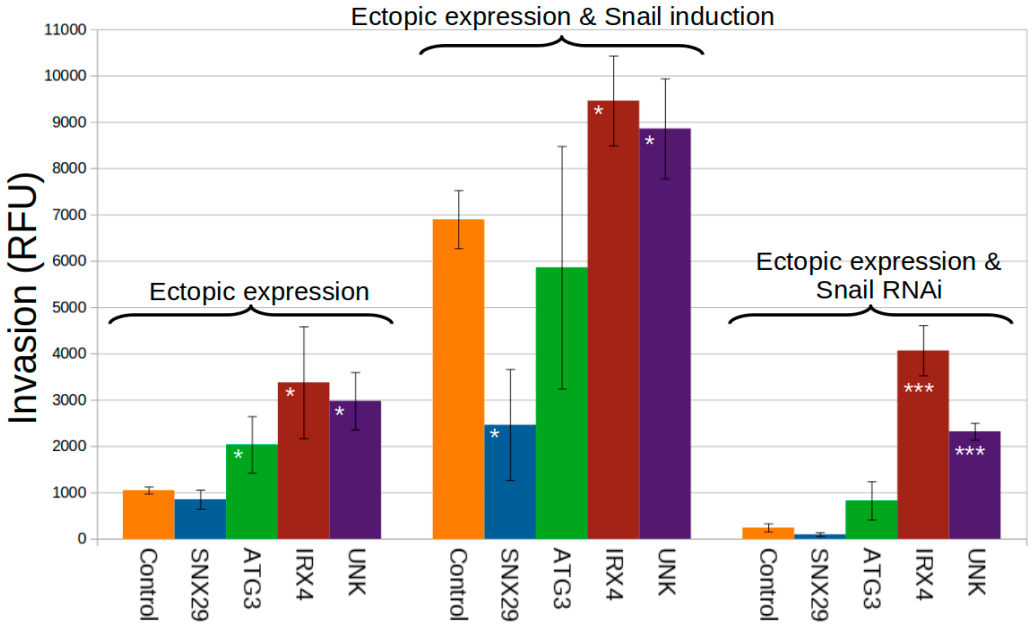

Figure 6. Validation of candidate invasion genes in breast cancer cells. The fluorescence signal from invasive MCF7 cells is shown. Induction of every gene examined significantly changed invasion in least one of three conditions: (a) Ectopic expression; (b) ectopic expression and *SNAI1* induction; (c) 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* (green) had higher invasion at background levels of *SNAI1* (without induction or knockdown). Mean values are shown, error bars indicate 95% CI, $n = 3$; * $q < 0.05$; *** $q < 5.0 \times 10^{-4}$.

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 and 6). *IRX4* was a homeobox transcription factor involved in cardiogenesis, marking a ventricular-specific progenitor cell [142] and was also associated with prostate cancer risk [143]. *SNX29* was poorly characterised, belonged to the sorting nexin protein family that function in endosomal sorting and signalling [144], 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 [145], consistent with *SNX29* inhibition of invasion driven by Snail. *ATG3* was an E2-like enzyme required for autophagy and mitochondrial homeostasis [146], *ATG3* overexpression significantly increased MCF7 invasion. Knockdown of *ATG3* reduced invasion in hepatocellular carcinoma [147]. *UNK* was a RING finger protein homologous to *unkempt* which bound mRNA, functioned in ubiquitination and was upregulated in gastrulation [148]. Others reported that UNK mRNA binding controlled neuronal morphology and induced spindle-like cell shape in fibroblasts [149,150]. 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 [151]. These in vitro confirmatory results both supported the novel analysis approach and evidenced new function for the genes examined.

3. Methods

3.1. A Comprehensive D. Melanogaster Functional Gene Network (DroFN)

A high-confidence, comprehensive *Drosophila melanogaster* functional network (DroFN) was developed using a previously described supervised learning approach in order to model global gene function [152]. Genes were network nodes and their interactions represented associations

within biological pathways. Gene interactions were quantified by functional interaction probabilities, reflecting pathway co-membership, estimated by logistic regression of Bayesian probabilities from STRING v8.0 scores [40] and Gene Ontology (GO) coannotations [39]; KEGG [41] pathways were taken as gold standard.

Gene pair co-annotations were derived from the GO database of 25th March 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). Positive gold standard gene pairs were derived from genes found within the same pathway, the remaining gene pairs had no evidence for a pathway comembership interaction and were therefore assigned to the negative class. Bayesian probabilities for STRING and GO coannotation frequencies were derived from TRAIN-NET [152]. Selection of non-interacting "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). This approach to selection of negative pairs by random sampling helps to avoid bias [153]. TRAIN-BAL which was input for logistic regression, to derive a model of gene pair functional interaction probability (Equation (1)):

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

where: pGO is the Bayesian probability derived from Gene Ontology coannotation frequency, pSTRING is the Bayesian probability derived from the STRING score frequency.

The above model was applied to TRAIN-NET and the resulting score distribution thresholded by seeking a value that maximised the F-measure [154] and true positive rate (TPR), while also minimising the false positive rate (FPR). The selected threshold value ($p \geq 0.779$) was applied to functional interaction probabilities for all possible gene pairs to generate the high-confidence network, DroFN.

For evaluation of the DroFN network, time separated test data (TEST-TS) were taken from KEGG v62 on 13/6/12, consisting of 14 pathways that were not in TRAIN-NET. The pathways in the time separated test dataset were not present in KEGG at the time when the training data was downloaded, supporting stringent independent evaluation; indeed, this principle is employed in community critical assessment work [155]. Gene pairs were eliminated from TEST-TS if either gene was found in TRAIN-NET, removing 76 positive and 1294 negative gene pairs to generate the blind test dataset TEST-NET (3481 pairs). Therefore, all gene pairs in TEST-NET corresponded to the most stringent evaluation class, termed "C3" by Park and Marcotte [156]. The most up-to-date versions of GeneMania (v2017-03-14) [43] and DroID (v2018_08) [42] downloaded April 2020 were assessed against TEST-NET (Table S1, Figure S1).

Enrichment of DroFN edges in DPiM [44] 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 25,797 edges in DroFN, of which 2175 were common to both networks. A 2×2 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, 23,622, 469,132. The enrichment *p*-value was calculated by Fisher's Exact Test.

3.2. Network Neighbourhood Clustering (NetNC) Algorithm

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

have greater coherence than nodes that do not share neighbours. The NetNC workflow is summarised in Figure 1 and described in detail below. Two analysis approaches are available (a) node-centric, parameter-free (NetNC-FBT) and (b) edge-centric, with two parameters (NetNC-FTI). Both approaches begin by assigning a *p*-value to each edge (S_{ij}) from Hypergeometric Mutual Clustering (HMC) [46], 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 [46] 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 analysis setting (NetNC-FTI) employs positive false discovery rate [157] and an iterative minimum cut procedure [158] 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 $q \leq Q$. Therefore, Q is the NetNC-FTI threshold for false discovery rate (q).

    (c) An iterative minimum cut procedure [158] 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. Therefore, Z is the threshold for the density of network components output by NetNC-FTI.

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

4. The node-centric, parameter-free approach (NetNC-FBT) proceeds by calculating degree-normalised node functional coherence scores (NFCS) from H_1, then identifies statistical modes of the NFCS distribution using Gaussian Mixture Modelling (GMM) [159].

    (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) (Equation (2)):

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

    (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 [159,160]. Models with 1..9 Gaussians are fitted and the final model selected using the Bayesian Information Criterion (BIC).

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

eliminated according to the above procedure on the resulting model. This procedure was developed following manual inspection of results on training data from KEGG pathways with "synthetic neutral target genes" (STNGs) as nodes resampled from *G* (TRAIN-CL).

Therefore, NetNC can be applied to predict functional coherence using either edge-centric or node-centric analysis settings. The NetNC-FTI (edge-centric) approach automatically produces a network, whereas the NetNC-FBT (node-centric) analysis does not output edges; therefore, to generate networks from predicted NetNC-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 is described in the sections below.

3.3. Estimating Positive False Discovery Rate for Hypergeometric Mutual Clustering p-Values

The following procedure is employed to estimate positive False Discovery Rate (pFDR) [157] in the NetNC-FTI approach (edge-centric). Subgraphs with number of nodes identical to *S* are resampled from *G* to derive a null distribution of HMC *p*-values (H_0) (described above). The resampling approach for pFDR calculation in NetNC-FTI controls for the structure of the network *G*, including degree distribution, (because G is fixed) 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 [161]. 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 [46] used in the NetNC algorithm to analyse *S*. Hence, the resampling procedure does not model the degree distribution of *S*, although the degree distribution of *G* is controlled for. Positive false discovery rate is estimated over the *p*-values in H_1 (p_{ij}) according to Storey [157] (Equation (3)):

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

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 $p \le p_\alpha$. H_0 represents *Y* resamples, therefore *V* is calculated at each step (Equation (4)):

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

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 [162,163]. The value of *R* is estimated by (Equation (5)):

$$R = \sum\nolimits_{p \in H_1} \begin{cases} 1 & p_{ij} \le p_\alpha \\ 0 & otherwise \end{cases}. \tag{5}$$

Additionally, there is a requirement for monotonicity (Equation (6)):

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

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 [164–167].

3.4. Estimating Local False Discovery Rate from Global False Discovery Rate

We developed an approach to estimate local false discovery rate (lcFDR) [167], 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 (Equation (7)):

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

where l denotes the $pFDR_{min}$ closest to and smaller than k, and where at least one node has $pFDR_{min} \equiv pFDR_l$. Therefore, our approach can be conceptualised as operating on ordered $pFDR_{min}$ values. n indicates the nodes in D with $pFDR_{min}$ values meeting threshold k. X represents the number of nodes at $p_\alpha \equiv k$. The number of FPs for nodes with $p_\alpha \equiv k$ (FP_k) is estimated by subtracting the FP for threshold l from the FP at threshold k. Thus, division of FP_k by X gives local false discovery rate bounded by k and l (Figure S7). If we define the difference between $pFDR_k$ and $pFDR_l$ (Equation (8)):

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

Substituting $pFDR_k$ for ($pFDR_l + pFDR_\Delta$) into Equation (7) and then simplifying (Equation (9)):

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

Equations (7) and (9) do not apply to the node(s) in D at the smallest possible value of $pFDR_{min}$ because $pFDR_l$ would be undefined; instead, the value of $lcFDR_k$ is calculated as the (global) $pFDR_{min}$ value. Indeed, global FDR and local FDR are equivalent when H_1 consists of objects at a single $pFDR_{min}$ value. Taking the mean $lcFDR_k$ across D provided an estimate of neutral binding in the TF_ALL datasets and was calibrated against mean lcFDR values from analysis of "faux" candidate targets resampled from DroFN—where the number of resampled targets was identical to the number of candidate target genes in the TF_ALL dataset analysed. For comparison, we also calibrated mean lcFDR for TF_ALL against values from synthetic data with known %SNTGs (Figure S2). Estimation of the total proportion of neutral binding in ChIP-chip or ChIP-seq data required lcFDR rather than (global) pFDR and, for example, accounts for the shape of the H_1 distribution. In the context of NetNC analysis of TF_ALL, mean lcFDR may be interpreted as the probability that any candidate target gene is neutrally bound in the dataset analysed; therefore, providing estimation of the total neutral binding proportion. Computer code for calculation of lcFDR is provided within the NetNC distribution. Estimates of SNTGs by the NetNC-FBT approach were not taken forward due to large 95% CI values (Figure S8).

3.5. Median Difference and Correlation between Estimates of Functional Binding from NetNC Functional Target Identification and Local False Discovery Rate

Candidate target genes that passed NetNC-FTI thresholds were considered functional targets (FTI_FT). The proportion of FTI_FT genes was compared to the proportion of functional binding estimated by lcFDR (lcFDR_FB, Figure 3A). The modulus of the difference between FTI_FT and

lcFDR_FB for each dataset gave a distribution of differences in predicted functional binding and the median of this distribution was quoted in the text above.

3.6. NetNC Benchmarking Data

There are significant challenges in separating functional TF targets from neutral binding in experimentally determined candidate TF target gene lists [1,2,5]. For example, it is not straightforward to define neutral TF binding by examining target genes that do not change steady-state expression upon knockout of the TF in question: TF knockout might not affect expression of *bona fide* TF targets where an additional TF (or TFs) are partially redundant with the lost TF [2,5]; and loss of TF binding may alter gene expression dynamics, such as a change in oscillation or stochasticity, without changing steady-state gene expression measured across a large population of cells [164,165]. Gene expression changes following TF knockout can also incorrectly propose "functional" targets that change expression via an indirect mechanism. For example, manipulation of gene expression by knockout or overexpression causes systemic changes in gene expression, including regulation by feedback loops with complicated logical outputs [6]. Therefore, we developed synthetic benchmark data for the purpose of development and evaluation of NetNC—where biologically coherent genes were taken to represent functional TF targets, and randomly sampled genes taken to represent neutral binding. We consider the relevance of our synthetic benchmark data to the nine TF_ALL datasets in the section below; firstly we outline the construction of the synthetic benchmark. Gold standard data for NetNC benchmarking and parameterisation took pathways from KEGG to represent biologically coherent gene groups (v62, downloaded 13/6/12) [41]. 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 S9. 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 [168,169]. TEST-CL also included the related pathways dme04745 "Phototransduction" and dme00600 "Sphingolipid metabolism", for example where ceramide kinase regulates photoreceptor homeostasis [170–172].

Gold standard datasets were also developed in order to investigate the effect of dataset size and noise on NetNC performance. The inclusion of noise as resampled network nodes into the gold-standard data was taken to model neutral TF binding [1,7] and matches expectations on data taken from biological systems in general [164,167]. Therefore, gold standard datasets were generated by combining TRAIN-CL with nodes resampled from the network (G). The final proportion of resampled nodes (Synthetic Neutral Target Genes, SNTGs) ranged from 5% through to 80% in 5% increments. SNTGs were drawn by uniform resampling from the DroFN network using the rand() function in perl excluding the genes in TRAIN-CL. Since we expected variability in the network proximity of SNTGs to pathway nodes, 100 resampled datasets were generated per %SNTG increment. Additional 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 [32] utilised a single subsample for 10%, 25%, 50%, and 75% SNTGs (TRAIN-CL-SR, TEST-CL-SR).

3.7. Comparison of Synthetic NetNC Benchmark to Experimentally Determined TF Binding Data

Transcription factors act to coordinately regulate multiple functionally related targets [20–22]. Accordingly, we reasoned that biological pathways could be taken as synthetic functional TF targets.

We modelled neutral binding by random resampling across all of the genes represented in the DroFN network, with the rationale that the creation and disappearance of neutral binding sites would not be driven by evolutionary selection. In order to further explore the correspondence between the synthetic benchmark and the TF binding datasets analysed (TF_ALL), we compared the global clustering coefficient (CC) of the network induced in DroFN by the nine TF_ALL datasets and the synthetic benchmark (TEST-CL_ALL). CC provides a single graph-theoretic measure of clustering in each dataset, which is a key property used by NetNC to identify functional TF targets. Therefore, CC is an appropriate measure to use for comparison of the synthetic and experimentally determined datasets. Accordingly, we developed linear models using TEST-CL_ALL in order to predict the CC of the biological TF_ALL data. Each dataset in TF_ALL had been assigned to a proportion of neutral binding, matched to a %SNTG value in the synthetic benchmark (Figure S2). For model training, we took the mean CC across the 100 repeats per %SNTG level for the six different dataset sizes in the synthetic data (from three up to eight pathways in each). This generated six values of CC and six values for the number of nodes (#nodes) in the dataset per %SNTG, which were used for the model:

$$\text{CC} = \text{intercept} + (\text{coefficient} \times \log(\#\text{nodes})). \quad (10)$$

As an illustrative example, the sna_2–3h_union dataset was matched to the benchmark dataset with 75% SNTGs. Therefore, a model was fitted using the TEST-CL_75% dataset with the six values for the number of nodes (for each of the six sub-datasets from three to eight pathways) and their six corresponding CC values. The model based on TEST-CL_75% was used to predict CC for sna_2–3h_union; the predicted CC value was compared to the CC calculated for the graph induced by sna_2–3h_union in DroFN. The fitted models therefore accounted for the expected influence of dataset size on the CC of the induced subnetwork. The "glm" function in R was used for model fitting. The regularised fit, determined by Akaike Information Criterion, was always superior for models where the logarithm of the number of nodes was taken, rather than taking raw values. The calculated CC values were for subnetworks induced in DroFN by the relevant TF binding dataset and included disconnected nodes.

3.8. NetNC-FTI Parameter Optimisation

NetNC-FTI analysed the TRAIN-CL_ALL datasets 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-FTI 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 positives and SNTGs were negatives for the FTI MCC calculation. The NetNC-FTI approach therefore tests discrimination of pathway nodes from SNTGs, which is particularly relevant to identification of functionally coherent candidate TF targets from ChIP-chip or ChIP-seq peaks.

Parameter selection for NetNC-FTI 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-FTI parameters (Q, Z) are available from the BioStudies database (www.ebi.ac.uk/biostudies/studies/S-BSST460) and contour plots showing mean MCC across Q, Z values per %SNTG are provided in Figure S9. A "SNTG specified" parameter set was developed for situations where an estimate of the input data noise component is available, for example from the NetNC-FBT approach. In this parameterisation, for each of the 16 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 S10. Parameters were also determined for analysis without any prior information about the %SNTG in the input data. For this purpose, a contour plot was produced to represent the proportion of datasets where NetNC-FTI performed better than 75% of the maximum performance across TRAIN-CL_ALL for the FTI task in the Q, Z parameter space. The maximum circle approach described above was applied to the contour plot in order to derive "robust" parameter values (Q, Z), which were respectively 0.120, 0.306 (NetNC-FTI).

3.9. Performance on Blind Test Data

We compared NetNC-FTI and NetNC-FBT against leading methods, HC-PIN [33] and MCL [32] on blind test data (Figure 2, Table S2). HC-PIN was obtained from the developers and is currently available within the cytocluster Cytoscape app (https://apps.cytoscape.org/apps/cytocluster); MCL is available from https://micans.org. 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 [33]; 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 BioStudies 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 [152]. 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 S11). Comparison to NEST [52] and baseline node degree was performed on TEST-CL-SR (Table S3). NEST required expression values, therefore a uniform expression value was added to the NEST input for all TEST-CL-SR nodes. The NEST output included genes that were not present in the input data from TEST-CL-SR and these additional genes were removed in order to produce the "Filtered NEST" dataset. The NEST scores or node degree were analysed separately against the labelling of TEST-CL-SR nodes as KEGG pathways (positives) or SNTGs (negatives), enabling calculation of area under the Receiver Operator Characteristic curve for each method examined (Table S3).

3.10. 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 above and described previously in detail [8,10,22,53,54]. A hundred subsamples of each of these datasets were taken at rates of 95%, 80%, and 50%, thereby producing a total of 2700 datasets (TF_SAMPL). NetNC-FTI results across TF_SAMPL were used as input for calculation of median and 95% confidence intervals for the edge and gene overlap per subsampling rate for each transcription factor dataset analysed. The NetNC resampling parameter (Y) was set at 100, the default value. The edge overlap was calculated as the proportion of edges returned by NetNC-FTI for the subsampled dataset that were also present in NetNC-FTI results for the full dataset (i.e., at 100%). Therefore, nine values for median overlap and 95% CI were produced per subsampling rate for both edge and gene overlap, corresponding to the nine transcription factor binding datasets (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 Supplementary Materials.

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 [173] unless otherwise specified in the text. Calculation of pFDR and local FDR values by NetNC is described in the sections above.

3.11. 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) [22,53,54] and ChIP followed by solexa pyrosequencing (ChIP-seq) [8]. Additionally "highly occupied target" regions, reflecting multiple and complex transcription factor occupancy profiles, were obtained from ModEncode [10]. Nine datasets were analysed in total (TF_ALL) and are summarised below.

The "union" datasets (WT embryos 2–3 h, mostly late stage four or early stage five) combined ChIP-chip peaks significant at 1% FDR for two different antibodies targeted at the same TF and these were assigned to the closest transcribed gene according to RNA Polymerase II binding data [22]. 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 [54,174]; 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 [54]. The two-fold enrichment threshold selected for this study reflects "weak" binding, although was expected to include functional TF targets [9]. 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 38,562 regions, of which 1855 had complexity score ≥8 and had been mapped to 1648 FlyBase genes according to the nearest transcription start site [10]; 677 of these genes were matched to a DroFN node (HOT). The "HighConf" data took Twist ChIP-seq binding peaks in WT embryos (1–3 h) 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 [8] of which 755 were mapped to FlyBase. 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 [53] of which 1444 mapped to DroFN (Intersect_ALL). Subsets of Intersect_ALL identified regions bound only at 2–4 h (twi_2–4h_intersect, 801 genes), or only at 4–6 h (twi_4–6h_intersect, 818 genes), or "continuously bound" regions identified at both 2–4 and 4–6 h (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 min), cellularisation of the blastoderm (stage five, 130–170 min) and initiation of gastrulation (stage 6, 170–180 min) [8,22,53,54,175]. 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 min) [53,54,175]; twi_2–4h_Toll^{10b} and sna_2–4h_Toll^{10b} may have also included stages eight (190–220 min) and nine (220–260 min) [54,175]. Twi_2–4h_intersect and sna_2–4h_intersect were tightly staged between stages 5–7 [53]. Additional to stages four, five and six, twi_1–3h_hiConf may have included the latter part of stage two (preblastoderm, 25–65 min) and stage three (pole bud formation, 65–80 min) [175]. The twi_4–6h_intersect dataset was restricted to stages eight to nine which included germ band elongation and segmentation of neuroblasts [53,175]. Therefore, there were differences in the biological material used across TF_ALL.

The Notch signalling modifiers analysed in this study were selected based on identification in at least two of the screens reported in [64]. Networks were annotated using GO and FlyBase [31,39,176,177].

3.12. 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, and GSE68892 (formerly geral-00143 from caBIG)). All datasets were Affymetrix U133A/plus 2 chips and were summarised with Ensembl alternative CDF [178]. RMA normalization [179] and ComBat batch correction [180] were applied to remove dataset-specific bias as previously described [110,181]. Intrinsic molecular subtypes were assigned based upon the highest correlation to Sorlie centroids [104], applied to each dataset separately. Centred average linkage clustering was performed using the Cluster and TreeView programs [182]. Centroids were calculated for each gene based upon the mean expression across each of the Sorlie intrinsic subtypes [104]. 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 [183]. Differential expression was calculated by t-test comparing normalised (unsquared) expression values in normal-like and basal-like tumours with false discovery rate correction [173].

3.13. Invasion Assays for Validation of Genes Selected from NetNC Results

MCF-7 Tet-On cells were purchased from Clontech and maintained as previously described [184]. 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 allows the cells to invade through a matrigel barrier utilising basement membrane-coated inserts according to the manufacturer's protocol. Briefly, MCF7 cells transfected with the constructs (Doxycycline-inducible *SNAI1* cDNA or *SNAI1 shRNA* with or without candidate gene cDNA) were suspended in serum-free medium. *SNAI1* cDNA or *SNAI1 shRNA* were cloned in our doxycyline-inducible pGoldiLox plasmid (pGoldilox-Tet-ON for cDNA and pGolidlox-tTS for shRNA expression) using validated shRNAs against *SNAI1* (NM_005985 at position 150 of the transcript [184]). pGoldilox has been used previously to induce and knock down the expression of *Ets* genes [185]. 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 h. 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 [186]; please see Data File S2 for gene primers. The knockdown efficiency for Snai1 was >81% (5.4-fold knockdown), Snai1 induction produced 2.0-fold overexpression. 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.

3.14. Data and Software Availability

NetNC is available at https://github.com/overton-group/NetNC. The following data are available from the European Bioinformatics Institute BioStudies database (https://www.ebi.ac.uk/biostudies/studies/S-BSST460): The DroFN network; all gold standard datasets; HCPIN input, output and performance summary files. The all-vs-all connectivity matrix before application of the DroFN edge weight threshold is available upon request.

4. Conclusions

We developed and validated the novel NetNC algorithm for identification of biologically coherent transcription factor (TF) target genes, and a comprehensive *D. melanogaster* functional gene network (DroFN). While NetNC was developed for functional TF target discovery, the approach may be widely useful for recovery of functionally coherent nodes in noisy data, for example in analysis of differential gene expression or CRISPR screen data. The network-based statistical framework in NetNC is applicable to single sample datasets and includes a novel method for estimation of local false discovery rate (FDR) from global FDR values. Analysis of Snail, Twist and modENCODE highly occupied target (HOT) regions found from 50% to 95% of candidate target genes were neutrally bound across the nine datasets analysed (TF_ALL). Correlation of the predicted neutral binding proportion with experimental and analytical factors across TF_ALL suggested consideration of strategies to enrich for functional TF targets. Datasets representing > 1 time period or that were derived from multiple TFs (HOT regions) had a relatively high proportion of functional binding, aligning with the emerging picture of widespread combinatorial control involving TF–TF interactions, cooperativity and TF redundancy [2,5,57–59]. The NetNC functional target networks provide a map of genome-scale regulation by Snail and Twist in early *D. melanogaster* embryogenesis. Each of the networks for the nine TF_ALL datasets was significantly enriched in Notch signalling modifiers, and we predicted genes involved in signalling cross-talk—where Snail and Twist may act to control the pleiotropic consequences of Notch activation. Eleven biological functions were annotated to at least four of the nine TF_ALL networks, including developmental regulation, chromatin organisation and mushroom body development. Predicted Snail and Twist regulation of chromatin structure, including PRC1 core components and other gene silencing factors, provides evidence for the action of EMT TFs in controlling the expression of their own coregulators. Unsupervised clustering with orthologues of the NetNC functional targets stratified 2999 breast cancer transcriptomes into the five intrinsic subtypes [104]; demonstrating that the regulation by Snail and Twist in fly mesoderm development captures important features of breast cancer biology. We identified breast cancer subtype-specific genes and network modules. Results in the basal-like subtype suggest a role for *HECA* in an EMT-like gene expression programme, and predict orthologous Snail, Twist functional targets that may control the consequences of *Notch* activation through chromatin remodelling. Our integrative analysis revealed subtype-specific genes that may prove useful for precision medicine. For example, potentially informing development of companion diagnostics or combination therapies targeting the notch pathway in basal-like tumours. We validated predicted roles in invasion for four NetNC functional targets in a breast cancer model, supporting our approach.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/12/10/2823/s1, Figure S1: Reciever Operator Characteristic curves for Functional Gene Networks on time separated blind test data, Figure S2: Predicted Neutral Binding Proportion for TF_ALL with calibration against synthetic data, Figure S3: Agreement between NetNC-lcFDR and NetNC-FTI estimates of functional binding, Figure S4: Networks of functionally coherent candidate target genes identified by NetNC-FTI, Figure S5: Unsupervised clustering of 2999 primary breast cancers with NetNC results for individual Snail and Twist datasets, Figure S6: Survival analysis of four genes associated with Luminal B but not Luminal A subtypes, Figure S7: Estimation of local False Discovery Rate from positive False Discovery Rate, Figure S8: Proportion of neutral binding predicted by NetNC-FBT and NetNC-lcFDR, Figure S9: NetNC Functional Target Identification Performance on TRAIN-CL_ALL, Table S1: Evaluation of DroFN on Time Separated Blind Test Data, Table S2: Results from preliminary study of NetNC and MCL performance on TEST-CL-SR, Table S3: Overlap for NetNC-FTI subsamples of the TF_ALL datasets, Table S4: Enrichment of evolutionary conservation in functional targets of Snail and Twist, Table S5: Overall summary of

clusters identified by NetNC-FTI across the nine TF_ALL datasets, Table S6: Frequently identified genes within the Developmental Regulation Clusters, Table S7: NetNC functionally coherent TF targets found in chromatin organisation clusters across multiple datasets, Table S8: Summary of pathways selected for gold standard data TRAIN-CL and TEST-CL, Table S9: NetNC-FTI parameter values optimised for percentage Synthetic Neutral Target Genes, Table S10: NetNC-FTI parameter values optimised for percentage Synthetic Neutral Target Genes (%SNTG), Table S11: Optimisation of MCL Inflation value for the Functional Target Identification task, Data file S1: Cytoscape sessions with NetNC-FTI results for TF_ALL, Data File S2: Primers for RT-qPCR, Supplementary text: Subsampling results, related to Table S4.

Author Contributions: Conceptualization, I.M.O.; Data curation, I.M.O.; Formal analysis, I.M.O., A.H.S., J.A.O., B.S.E.H., and M.J.F.; Funding acquisition, I.M.O., A.H.S., B.S.E.H., and A.E.; Investigation, I.M.O., A.H.S., J.A.O., B.S.E.H., M.J.F., E.P.-C., and A.E.; Methodology, I.M.O., A.H.S., J.A.O., B.S.E.H., M.J.F., A.L.R.L., and A.E.; Project administration, I.M.O.; Resources, I.M.O., A.H.S., and A.E.; Software, I.M.O., J.A.O. and A.L.R.L.; Supervision, I.M.O.; Validation, I.M.O. and A.E.; Visualization, I.M.O., A.H.S., J.A.O., and E.P.-C.; Writing—original draft, I.M.O., J.A.O., and A.E.; Writing—review and editing, I.M.O., A.H.S., J.A.O., B.S.E.H., M.J.F., A.L.R.L., E.P.-C., and A.E. All authors have read and agreed to the published version of the manuscript.

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

Acknowledgments: Ian M. Overton is grateful to Jeremy Gunawardena and Peter Sorger for hosting his visit to HMS and for helpful discussions. Thanks to P.S. Thiagarajan, Andrew Millar, Wendy Bickmore, Nick Hastie, Ben Lehner and Julian Dow for invaluable comments. Nick Moir and Seanna McTaggart assisted with testing the NetNC software distribution.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shlyueva, D.; Stampfel, G.; Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.* **2014**, *15*, 272–286. [CrossRef] [PubMed]
2. Stampfel, G.; Kazmar, T.; Frank, O.; Wienerroither, S.; Reiter, F.; Stark, A. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **2015**, *528*, 147–151. [CrossRef] [PubMed]
3. 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. Transcription factor networks in drosophila melanogaster. *Cell Rep.* **2014**, *8*, 2031–2043. [CrossRef] [PubMed]
4. Zabidi, M.A.; Stark, A. Regulatory enhancer–core-promoter communication via transcription factors and cofactors. *Trends Genet.* **2016**, *32*, 801–814. [CrossRef] [PubMed]
5. Khoueiry, P.; Girardot, C.; Ciglar, L.; Peng, P.C.; Gustafson, E.H.; Sinha, S.; Furlong, E.E. Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *eLife* **2017**, 6. [CrossRef]
6. Wilczynski, B.; Furlong, E.E.M. Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. *Dev. Biol.* **2010**, *340*, 161–169. [CrossRef]
7. 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. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol.* **2008**, *6*, e27. [CrossRef]
8. Ozdemir, A.; Fisher-Aylor, K.I.; Pepke, S.; Samanta, M.; Dunipace, L.; McCue, K.; Zeng, L.; Ogawa, N.; Wold, B.J.; Stathopoulos, A. High resolution mapping of twist to DNA in drosophila embryos: Efficient functional analysis and evolutionary conservation. *Genome Res.* **2011**, *21*, 566–577. [CrossRef]
9. Biggin, M.D. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **2011**, *21*, 611–626. [CrossRef]
10. 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. Identification of functional elements and regulatory circuits by drosophila modENCODE. *Science* **2010**, *330*, 1787–1797. [CrossRef]
11. Kvon, E.Z.; Stampfel, G.; Yáñez-Cuna, J.O.; Dickson, B.J.; Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.* **2012**, *26*, 908–913. [CrossRef] [PubMed]

12. Li, H.; Chen, H.; Liu, F.; Ren, C.; Wang, S.; Bo, X.; Shu, W. Functional annotation of hot regions in the human genome: Implications for human disease and cancer. *Sci. Rep.* **2015**, *5*, 11633. [CrossRef] [PubMed]
13. Moorman, C.; Sun, L.V.; Wang, J.; de Wit, E.; Talhout, W.; Ward, L.D.; Greil, F.; Lu, X.-J.; White, K.P.; Bussemaker, H.J.; et al. Hotspots of transcription factor colocalization in the genome of drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 12027–12032. [CrossRef]
14. Montavon, T.; Soshnikova, N.; Mascrez, B.; Joye, E.; Thevenet, L.; Splinter, E.; de Laat, W.; Spitz, F.; Duboule, D. A regulatory archipelago controls hox genes transcription in digits. *Cell* **2011**, *147*, 1132–1145. [CrossRef]
15. Teytelman, L.; Thurtle, D.M.; Rine, J.; van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18602–18607. [CrossRef]
16. Cannavò, E.; Khoueiry, P.; Garfield, D.; Geeleher, G.; Zichner, T.; Gustafson, E.; Ciglar, L.; Korbel, J.; Furlong, E. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr. Biol.* **2016**, *26*, 38–51. [CrossRef] [PubMed]
17. Garcia-Alonso, L.; Holland, C.H.; Ibrahim, M.M.; Turei, D.; Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **2019**, *29*, 1363–1375. [CrossRef]
18. Keung, A.J.; Bashor, C.J.; Kiriakov, S.; Collins, J.J.; Khalil, A.S. Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell* **2014**, *158*, 110–120. [CrossRef]
19. Brown, J.B.; Celniker, S.E. Lessons from modENCODE. *Annu. Rev. Genom. Hum. Genet.* **2015**, *16*, 31–53. [CrossRef]
20. Igual, J.C.; Johnson, A.L.; Johnston, L.H. Coordinated regulation of gene expression by the cell cycle transcription factor Swi4 and the protein kinase C MAP kinase pathway for yeast cell integrity. *EMBO J.* **1996**, *15*, 5001–5013. [CrossRef]
21. Karczewski, K.J.; Snyder, M.; Altman, R.B.; Tatonetti, N.P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. *PLoS Genet.* **2014**, *10*, e1004122. [CrossRef] [PubMed]
22. 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. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **2009**, *10*, R80. [CrossRef] [PubMed]
23. Hartwell, L.H.; Hopfield, J.J.; Leibler, S.; Murray, A.W. From molecular to modular cell biology. *Nature* **1999**, *402*, C47–C52. [CrossRef] [PubMed]
24. Hooper, S.D.; Boué, S.; Krause, R.; Jensen, L.J.; Mason, C.E.; Ghanim, M.; White, K.P.; Furlong, E.E.; Bork, P. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Mol. Syst. Biol.* **2007**, *3*, 72. [CrossRef] [PubMed]
25. Ideker, T.; Ozier, O.; Schwikowski, B.; Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **2002**, *18*, S233–S240. [CrossRef]
26. Vidal, M.; Cusick, M.E.; Barabási, A.-L. Interactome networks and human disease. *Cell* **2011**, *144*, 986–998. [CrossRef] [PubMed]
27. Jaeger, S.; Igea, A.; Arroyo, R.; Alcalde, V.; Canovas, B.; Orozco, M.; Nebreda, A.R.; Aloy, P. Quantification of pathway cross-talk reveals novel synergistic drug combinations for breast cancer. *Cancer Res.* **2017**, *77*, 459–469. [CrossRef] [PubMed]
28. Hu, J.X.; Thomas, C.E.; Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **2016**, *17*, 615–629. [CrossRef] [PubMed]
29. 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. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **2015**, *47*, 569–576. [CrossRef]
30. 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. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 15545–15550. [CrossRef]
31. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13. [CrossRef] [PubMed]

32. Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584. [CrossRef] [PubMed]
33. Wang, J.; Li, M.; Chen, J.; Pan, Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 607–620. [CrossRef] [PubMed]
34. Pawson, T.; Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **2003**, *300*, 445–452. [CrossRef] [PubMed]
35. Nieto, M.A.; Huang, R.Y.-J.; Jackson, R.A.; Thiery, J.P. EMT: 2016. *Cell* **2016**, *166*, 21–45. [CrossRef]
36. Lim, J.; Thiery, J.P. Epithelial-mesenchymal transitions: Insights from development. *Development* **2012**, *139*, 3471–3486. [CrossRef]
37. Giampieri, S.; Manning, C.; Hooper, S.; Jones, L.; Hill, C.S.; Sahai, E. Localized and reversible TGFbeta signalling switches breast cancer cells from cohesive to single cell motility. *Nat. Cell Biol.* **2009**, *11*, 1287–1296. [CrossRef]
38. 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. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **2013**, *339*, 580–584. [CrossRef]
39. 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. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
40. Jensen, L.J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; et al. STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **2009**, *37*, D412–D416. [CrossRef]
41. Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2010**, *38*, D355–D360. [CrossRef] [PubMed]
42. Yu, J.; Pacifico, S.; Liu, G.; Finley, R.L., Jr. DroID: The drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC Genom.* **2008**, *9*, 461. [CrossRef] [PubMed]
43. 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. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*, W214–W220. [CrossRef] [PubMed]
44. 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. A protein complex network of drosophila melanogaster. *Cell* **2011**, *147*, 690. [CrossRef] [PubMed]
45. Vitali, F.; Li, Q.; Schissler, A.G.; Berghout, J.; Kenost, C.; Lussier, Y.A. Developing a 'personalome' for precision medicine: Emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Brief. Bioinform.* **2017**, bbx149. [CrossRef]
46. Goldberg, D.S.; Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 4372–4376. [CrossRef]
47. Leiserson, M.D.M.; Vandin, F.; Wu, H.-T.; Dobson, J.R.; Eldridge, J.V.; Thomas, J.L.; Papoutsaki, A.; Kim, Y.; Niu, B.; McLellan, M.; et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **2015**, *47*, 106–114. [CrossRef]
48. Taşan, M.; Musso, G.; Hao, T.; Vidal, M.; MacRae, C.A.; Roth, F.P. Selecting causal genes from genome-wide association studies via functionally-coherent subnetworks. *Nat. Methods* **2015**, *12*, 154–159. [CrossRef]
49. 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. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **2014**, *343*, 84–87. [CrossRef]
50. Simonis, M.; Klous, P.; Splinter, E.; Moshkin, Y.; Willemsen, R.; de Wit, E.; van Steensel, B.; de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **2006**, *38*, 1348. [CrossRef]
51. 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. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **2006**, *16*, 1299–1309. [CrossRef] [PubMed]

52. Jiang, P.; Wang, H.; Li, W.; Zang, C.; Li, B.; Wong, Y.J.; Meyer, C.; Liu, J.S.; Aster, J.C.; Liu, S. Network analysis of gene essentiality in functional genomics experiments. *Genome Biol.* **2015**, *16*, 239. [CrossRef] [PubMed]
53. Sandmann, T.; Girardot, C.; Brehme, M.; Tongprasit, W.; Stolc, V.; Furlong, E.E.M. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev.* **2007**, *21*, 436–449. [CrossRef] [PubMed]
54. Zeitlinger, J.; Zinzen, R.P.; Stark, A.; Kellis, M.; Zhang, H.; Young, R.A.; Levine, M. Whole-genome ChIP-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. *Genes Dev.* **2007**, *21*, 385–390. [CrossRef]
55. Chen, R.A.-J.; Stempor, P.; Down, T.A.; Zeiser, E.; Feuer, S.K.; Ahringer, J. Extreme HOT regions are CpG-dense promoters in C. elegans and humans. *Genome Res.* **2014**, *24*, 1138–1146. [CrossRef]
56. Boyle, A.P.; Araya, C.L.; Brdlik, C.; Cayting, P.; Cheng, C.; Cheng, Y.; Gardner, K.; Hillier, L.; Janette, J.; Jiang, L.; et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **2014**, *512*, 453. [CrossRef]
57. Jolma, A.; Yin, Y.; Nitta, K.R.; Dave, K.; Popov, A.; Taipale, M.; Enge, M.; Kivioja, T.; Morgunova, E.; Taipale, J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **2015**, *527*, 384–388. [CrossRef]
58. Long, H.K.; Prescott, S.L.; Wysocka, J. Ever-changing landscapes: Transcriptional enhancers in development and evolution. *Cell* **2016**, *167*, 1170–1187. [CrossRef]
59. Spitz, F.; Furlong, E.E.M. Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.* **2012**, *13*, 613–626. [CrossRef]
60. Chen, J.; Hu, Z.; Phatak, M.; Reichard, J.; Freudenberg, J.M.; Sivaganesan, S.; Medvedovic, M. Genome-wide signatures of transcription factor activity: Connecting transcription factors, disease, and small molecules. *PLoS Comput. Biol.* **2013**, *9*, e1003198. [CrossRef]
61. Aris-Brosou, S. Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. *Mol. Biol. Evol.* **2005**, *22*, 200–209. [CrossRef] [PubMed]
62. Wieschaus, E.; Nüsslein-Volhard, C. The heidelberg screen for pattern mutants of drosophila: A personal account. *Annu. Rev. Cell Dev. Biol.* **2016**, *32*, 1–46. [CrossRef]
63. Gheisari, E.; Aakhte, M.; Müller, H.-A.J. Gastrulation in drosophila melanogaster: Genetic control, cellular basis and biomechanics. *Mech. Dev.* **2020**, 103629. [CrossRef] [PubMed]
64. Guruharsha, K.G.; Kankel, M.W.; Artavanis-Tsakonas, S. The Notch signalling system: Recent insights into the complexity of a conserved pathway. *Nat. Rev. Genet.* **2012**, *13*, 654–666. [CrossRef] [PubMed]
65. Ntziachristos, P.; Lim, J.S.; Sage, J.; Aifantis, I. From fly wings to targeted cancer therapies: A centennial for notch signaling. *Cancer Cell* **2014**, *25*, 318–334. [CrossRef] [PubMed]
66. Bray, S.J. Notch signalling in context. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 722–735. [CrossRef]
67. Nowell, C.S.; Radtke, F. Notch as a tumour suppressor. *Nat. Rev. Cancer* **2017**, *17*, 145. [CrossRef]
68. Bernard, F.; Krejci, A.; Housden, B.; Adryan, B.; Bray, S.J. Specificity of notch pathway activation: Twist controls the transcriptional output in adult muscle progenitors. *Development* **2010**, *137*, 2633–2642. [CrossRef]
69. Sahlgren, C.; Gustafsson, M.V.; Jin, S.; Poellinger, L.; Lendahl, U. Notch signaling mediates hypoxia-induced tumor cell migration and invasion. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6392–6397. [CrossRef]
70. Baylies, M.K.; Bate, M. Twist: A myogenic switch in drosophila. *Science* **1996**, *272*, 1481–1484. [CrossRef]
71. Xie, Y.; Li, X.; Deng, X.; Hou, Y.; O'Hara, K.; Urso, A.; Peng, Y.; Chen, L.; Zhu, S. The ets protein pointed prevents both premature differentiation and dedifferentiation of *drosophila* intermediate neural progenitors. *Development* **2016**, *143*, 3109–3118. [CrossRef] [PubMed]
72. Chen, C.-M.; Freedman, J.A.; Bettler, D.R.; Manning, S.D.; Giep, S.N.; Steiner, J.; Ellis, H.M. Polychaetoid is required to restrict segregation of sensory organ precursors from proneural clusters in drosophila. *Mech. Dev.* **1996**, *57*, 215–227. [CrossRef]
73. Lo, P.C.H.; Skeath, J.B.; Gajewski, K.; Schulz, R.A.; Frasch, M. Homeotic genes autonomously specify the anteroposterior subdivision of the drosophila dorsal vessel into aorta and heart. *Dev. Biol.* **2002**, *251*, 307–319. [CrossRef] [PubMed]
74. Trujillo, G.V.; Nodal, D.H.; Lovato, C.V.; Hendren, J.D.; Helander, L.A.; Lovato, T.L.; Bodmer, R.; Cripps, R.M. The canonical wingless signaling pathway is required but not sufficient for inflow tract formation in the drosophila melanogaster heart. *Dev. Biol.* **2016**, *413*, 16–25. [CrossRef]

75. Hammonds, A.S.; Bristow, C.A.; Fisher, W.W.; Weiszmann, R.; Wu, S.; Hartenstein, V.; Kellis, M.; Yu, B.; Frise, E.; Celniker, S.E. Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol.* **2013**, *14*, R140. [CrossRef]
76. Tomancak, P.; Beaton, A.; Weiszmann, R.; Kwan, E.; Shu, S.; Lewis, S.E.; Richards, S.; Ashburner, M.; Hartenstein, V.; Celniker, S.E.; et al. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biol.* **2002**, *3*, research0088. [CrossRef]
77. Hartley, D.; Xu, T.; Artavanis-Tsakonas, S. The embryonic expression of the notch locus of drosophila melanogaster and the implications of point mutations in the extracellular EGF-like domain of the predicted protein., the embryonic expression of the notch locus of drosophila melanogaster and the implications of point mutations in the extracellular EGF-like domain of the predicted protein. *EMBO J.* **1987**, *6*, 3407–3417.
78. Kusch, T.; Reuter, R. Functions for drosophila brachyenteron and forkhead in mesoderm specification and cell signalling. *Development* **1999**, *126*, 3991–4003.
79. Millo, H.; Bownes, M. The expression pattern and cellular localisation of myosin VI during the drosophila melanogaster life cycle. *Gene Expr. Patterns* **2007**, *7*, 501–510. [CrossRef]
80. Kuroda, M.I.; Kang, H.; De, S.; Kassis, J.A. Dynamic competition of polycomb and trithorax in transcriptional programming. *Annu. Rev. Biochem.* **2020**, *89*, 235–253. [CrossRef]
81. Shao, Z.; Raible, F.; Mollaaghababa, R.; Guyon, J.R.; Wu, C.; Bender, W.; Kingston, R.E. Stabilization of chromatin structure by PRC1, a polycomb complex. *Cell* **1999**, *98*, 37–46. [CrossRef]
82. Schotta, G.; Ebert, A.; Krauss, V.; Fischer, A.; Hoffmann, J.; Rea, S.; Jenuwein, T.; Dorn, R.; Reuter, G. Central role of *Drosophila* SU(VAR)3–9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J.* **2002**, *21*, 1121–1131. [CrossRef] [PubMed]
83. Lopez, A.; Higuet, D.; Rosset, R.; Deutsch, J.; Peronnet, F. Corto genetically interacts with Pc-G and trx-G genes and maintains the anterior boundary of Ultrabithorax expression in drosophila larvae. *Mol. Gen. Genom.* **2001**, *266*, 572–583. [CrossRef] [PubMed]
84. Mishra, K.; Chopra, V.S.; Srinivasan, A.; Mishra, R.K. Trl-GAGA directly interacts with lola like and both are part of the repressive complex of polycomb group of genes. *Mech. Dev.* **2003**, *120*, 681–689. [CrossRef]
85. Schuster, K.J.; Smith-Bolton, R.K. Taranis protects regenerating tissue from fate changes induced by the wound response in drosophila. *Dev. Cell* **2015**, *34*, 119–128. [CrossRef]
86. Tie, F.; Banerjee, R.; Saiakhova, A.R.; Howard, B.; Monteith, K.E.; Scacheri, P.C.; Cosgrove, M.S.; Harte, P.J. Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. *Development (Cambrige. Engl.)* **2014**, *141*, 1129. [CrossRef]
87. Gutierrez, L. The drosophila trithorax group gene tonalli(tna) interacts genetically with the Brahma remodeling complex and encodes an SP-RING finger protein. *Development* **2003**, *130*, 343–354. [CrossRef]
88. Crosby, M.A.; Miller, C.; Alon, T.; Watson, K.L.; Verrijzer, C.P.; Goldman-Levi, R.; Zak, N.B. The trithorax group gene moira encodes a brahma-associated putative chromatin-remodeling factor in drosophila melanogaster. *Mol. Cell. Biol.* **1999**, *19*, 1159–1170. [CrossRef]
89. Fanti, L.; Dorer, D.R.; Berloco, M.; Henikoff, S.; Pimpinelli, S. Heterochromatin protein 1 binds transgene arrays. *Chromosoma* **1998**, *107*, 286–292. [CrossRef]
90. Pulikkan, J.A.; Hegde, M.; Ahmad, H.M.; Belaghzal, H.; Illendula, A.; Yu, J.; O'Hagan, K.; Ou, J.; Muller-Tidow, C.; Wolfe, S.A.; et al. CBFβ-SMMHC inhibition triggers apoptosis by disrupting MYC chromatin dynamics in acute myeloid leukemia. *Cell* **2018**, *174*, 172–186.e21. [CrossRef]
91. Bao, X.; Deng, H.; Johansen, J.; Girton, J.; Johansen, K.M. Loss-of-function alleles of the JIL-1 histone H3S10 kinase enhance position-effect variegation at pericentric sites in drosophila heterochromatin. *Genetics* **2007**, *176*, 1355–1358. [CrossRef] [PubMed]
92. Sparmann, A.; van Lohuizen, M. Polycomb silencers control cell fate, development and cancer. *Nat. Rev. Cancer* **2006**, *6*, 846–856. [CrossRef] [PubMed]
93. Koppens, M.; van Lohuizen, M. Context-dependent actions of polycomb repressors in cancer. *Oncogene* **2016**, *35*, 1341–1352. [CrossRef] [PubMed]
94. 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. Polycomb complex 2 is required for E-cadherin repression by the snail1 transcription factor. *Mol. Cell. Biol.* **2008**, *28*, 4772–4781. [CrossRef] [PubMed]
95. Leptin, M. Twist and snail as positive and negative regulators during drosophila mesoderm development. *Genes Dev.* **1991**, *5*, 1568–1576. [CrossRef]

96. Gilmour, D.; Rembold, M.; Leptin, M. From morphogen to morphogenesis and back. *Nature* **2017**, *541*, 311–320. [CrossRef]
97. Ashraf, S.I.; Ip, Y.T. The snail protein family regulates neuroblast expression of inscuteable and string, genes involved in asymmetry and cell division in drosophila. *Development* **2001**, *128*, 4757–4767.
98. Zander, M.A.; Burns, S.E.; Yang, G.; Kaplan, D.R.; Miller, F.D. Snail coordinately regulates downstream pathways to control multiple aspects of mammalian neural precursor development. *J. Neurosci.* **2014**, *34*, 5164–5175. [CrossRef]
99. Nevil, M.; Bondra, E.R.; Schulz, K.N.; Kaplan, T.; Harrison, M.M. Stable binding of the conserved transcription factor grainy head to its target genes throughout drosophila melanogaster development. *Genetics* **2017**, *205*, 605–620. [CrossRef]
100. Caron, S.J.C.; Ruta, V.; Abbott, L.F.; Axel, R. Random convergence of olfactory inputs in the drosophila mushroom body. *Nature* **2013**, *497*, 113–117. [CrossRef]
101. Lin, S.; Ewen-Campen, B.; Ni, X.; Housden, B.E.; Perrimon, N. In vivo transcriptional activation using CRISPR/Cas9 in drosophila. *Genetics* **2015**, *201*, 433–442. [CrossRef] [PubMed]
102. Vesuna, F.; van Diest, P.; Chen, J.H.; Raman, V. Twist is a transcriptional repressor of E-cadherin gene expression in breast cancer. *Biochem. Biophys. Res. Commun.* **2008**, *367*, 235–241. [CrossRef] [PubMed]
103. Mohr, S.E.; Hu, Y.; Kim, K.; Housden, B.E.; Perrimon, N. Resources for functional genomics studies in drosophila melanogaster. *Genetics* **2014**, *197*, 1–18. [CrossRef] [PubMed]
104. Sørlie, T.; Tibshirani, R.; Parker, J.; Hastie, T.; Marron, J.S.; Nobel, A.; Deng, S.; Johnsen, H.; Pesich, R.; Geisler, S.; et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8418–8423. [CrossRef]
105. Cejalvo, J.M.; de Dueñas, E.M.; Galvan, P.; García-Recio, S.; Gasión, O.B.; Paré, L.; Antolin, S.; Martinello, R.; Blancas, I.; Adamo, B.; et al. Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. *Cancer Res.* **2017**, *77*, 2213–2221. [CrossRef]
106. 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. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346. [CrossRef]
107. Stylianou, S.; Clarke, R.B.; Brennan, K. Aberrant activation of notch signaling in human breast cancer. *Cancer Res.* **2006**, *66*, 1517–1525. [CrossRef]
108. Barnawi, R.; Al-Khaldi, S.; Majed Sleiman, G.; Sarkar, A.; Al-Dhfyan, A.; Al-Mohanna, F.; Ghebeh, H.; Al-Alwan, M. Fascin is critical for the maintenance of breast cancer stem cell pool predominantly via the activation of the notch self-renewal pathway. *Stem Cells* **2016**, *34*, 2799–2813. [CrossRef]
109. Ingthorsson, S.; Briem, E.; Bergthorsson, J.T.; Gudjonsson, T. Epithelial plasticity during human breast morphogenesis and cancer progression. *J. Mammary Gland. Biol. Neoplasia* **2016**, *21*, 139–148. [CrossRef]
110. 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. KIBRA exhibits MST-independent functional regulation of the hippo signaling pathway in mammals. *Oncogene* **2013**, *32*, 1821–1830. [CrossRef]
111. Venet, D.; Dumont, J.E.; Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **2011**, *7*, e1002240. [CrossRef]
112. Sarrió, D.; Rodriguez-Pinilla, S.M.; Hardisson, D.; Cano, A.; Moreno-Bueno, G.; Palacios, J. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **2008**, *68*, 989–997. [CrossRef] [PubMed]
113. Beltran, A.S.; Graves, L.M.; Blancafort, P. Novel role of engrailed 1 as a prosurvival transcription factor in basal-like breast cancer and engineering of interference peptides block its oncogenic function. *Oncogene* **2014**, *33*, 4767–4777. [CrossRef]
114. Adélaïde, J.; Finetti, P.; Bekhouche, I.; Repellini, L.; Geneix, J.; Sircoulomb, F.; Charafe-Jauffret, E.; Cervera, N.; Desplans, J.; Parzy, D.; et al. Integrated profiling of basal and luminal breast cancers. *Cancer Res.* **2007**, *67*, 11565–11575. [CrossRef] [PubMed]
115. 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. ETV6 gene rearrangements in invasive breast carcinoma. *Genes Chromosomes Cancer* **2005**, *44*, 103–108. [CrossRef]
116. Chapellier, M.; Bachelard-Cascales, E.; Schmidt, X.; Clément, F.; Treilleux, I.; Delay, E.; Jammot, A.; Ménétrier-Caux, C.; Pochon, G.; Besançon, R.; et al. Disequilibrium of BMP2 levels in the breast stem cell

niche launches epithelial transformation by overamplifying BMPR1B cell response. *Stem Cell Rep.* **2015**, *4*, 239–254. [CrossRef] [PubMed]
117. Ma, L.; Lu, M.-F.; Schwartz, R.J.; Martin, J.F. Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. *Development* **2005**, *132*, 5601–5611. [CrossRef] [PubMed]
118. Sørlie, T.; Perou, C.M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10869–10874. [CrossRef] [PubMed]
119. Tang, Z.; Kand, B.; Li, C.; Chen, T.; Zhang, Z. GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **2019**, *47*, W556–W560. [CrossRef]
120. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61–70. [CrossRef]
121. Popovici, V.; Chen, W.; Gallas, B.G.; Hatzis, C.; Shi, W.; Samuelson, F.W.; Nikolsky, Y.; Tsyganova, M.; Ishkin, A.; Nikolskaya, T.; et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* **2010**, *12*, R5. [CrossRef] [PubMed]
122. Conn, S.J.; Pillman, K.A.; Toubia, J.; Conn, V.M.; Salmanidis, M.; Phillips, C.A.; Roslan, S.; Schreiber, A.W.; Gregory, P.A.; Goodall, G.J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **2015**, *160*, 1125–1134. [CrossRef] [PubMed]
123. 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. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **2008**, *133*, 704–715. [CrossRef]
124. DiMeo, T.A.; Anderson, K.; Phadke, P.; Feng, C.; Perou, C.M.; Naber, S.; Kuperwasser, C. A novel lung metastasis signature links wnt signaling with cancer cell self-renewal and epithelial-mesenchymal transition in basal-like breast cancer. *Cancer Res.* **2009**, *69*, 5364–5373. [CrossRef] [PubMed]
125. 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. Stem-cell-like properties and epithelial plasticity arise as stable traits after transient twist1 activation. *Cell Rep.* **2015**, *10*, 131–139. [CrossRef]
126. Sieuwerts, A.M.; Kraan, J.; Bolt, J.; van der Spoel, P.; Elstrodt, F.; Schutte, M.; Martens, J.W.M.; Gratama, J.-W.; Sleijfer, S.; Foekens, J.A. Anti-epithelial cell adhesion molecule antibodies and the detection of circulating normal-like breast tumor cells. *J. Natl. Cancer Inst.* **2009**, *101*, 61–66. [CrossRef]
127. 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. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **2015**, *526*, 131–135. [CrossRef]
128. Guo, W.; Keckesova, Z.; Donaher, J.L.; Shibue, T.; Tischler, V.; Reinhardt, F.; Itzkovitz, S.; Noske, A.; Zürrer-Härdi, U.; Bell, G.; et al. Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **2012**, *148*, 1015–1028. [CrossRef]
129. Prat, A.; Perou, C.M. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* **2011**, *5*, 5–23. [CrossRef]
130. Resende, L.P.F.; Truong, M.E.; Gomez, A.; Jones, D.L. Intestinal stem cell ablation reveals differential requirements for survival in response to chemical challenge. *Dev. Biol.* **2017**, *424*, 10–17. [CrossRef]
131. Steneberg, P.; Englund, C.; Kronhamn, J.; Weaver, T.A.; Samakovlis, C. Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the drosophila trachea. *Genes Dev.* **1998**, *12*, 956–967. [CrossRef] [PubMed]
132. Amith, S.R.; Fliegel, L. Na+/H+ exchanger-mediated hydrogen ion extrusion as a carcinogenic signal in triple-negative breast cancer etiopathogenesis and prospects for its inhibition in therapeutics. *Semin. Cancer Biol.* **2017**, *43*, 35–41. [CrossRef] [PubMed]
133. Liu, L.; Kimball, S.; Liu, H.; Holowatyj, A.; Yang, Z.-Q.; Liu, L.; Kimball, S.; Liu, H.; Holowatyj, A.; Yang, Z.-Q. Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. *Oncotarget* **2014**, *6*, 2466–2482. [CrossRef] [PubMed]
134. 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. TRIP-Br2 promotes oncogenesis in nude mice and is frequently overexpressed in multiple human tumors. *J. Transl. Med.* **2009**, *7*, 8. [CrossRef]

135. García-Pedrero, J.M.; Kiskinis, E.; Parker, M.G.; Belandia, B. The SWI/SNF chromatin remodeling subunit BAF57 is a critical regulator of estrogen receptor function in breast cancer cells. *J. Biol. Chem.* **2006**, *281*, 22656–22664. [CrossRef]
136. Sethuraman, A.; Brown, M.; Seagroves, T.N.; Wu, Z.-H.; Pfeffer, L.M.; Fan, M. SMARCE1 regulates metastatic potential of breast cancer cells through the HIF1A/PTK2 pathway. *Breast Cancer Res.* **2016**, *18*, 81. [CrossRef]
137. 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. SMARCE1 is required for the invasive progression of in situ cancers. *Proc. Natl. Acad. Sci. USA* **2017**, 201703931. [CrossRef]
138. 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. DOC1-dependent recruitment of NURD reveals antagonism with SWI/SNF during epithelial-mesenchymal transition in oral cancer cells. *Cell Rep.* **2017**, *20*, 61–75. [CrossRef]
139. Hemberger, M.; Dean, W.; Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: Digging waddington's canal. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 526–537. [CrossRef]
140. Dhasarathy, A.; Kajita, M.; Wade, P.A. The transcription factor snail mediates epithelial to mesenchymal transitions by repression of estrogen receptor alpha. *Mol. Endocrinol.* **2007**, *21*, 2907–2918. [CrossRef]
141. Lacroix, M.; Leclercq, G. Relevance of breast cancer cell lines as models for breast tumours: An update. *Breast Cancer Res. Treat.* **2004**, *83*, 249–289. [CrossRef] [PubMed]
142. 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. Irx4 marks a multipotent, ventricular-specific progenitor cell. *Stem Cells* **2016**, *34*, 2875–2888. [CrossRef] [PubMed]
143. Xu, X.; Hussain, W.M.; Vijai, J.; Offit, K.; Rubin, M.A.; Demichelis, F.; Klein, R.J. Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur. J. Hum. Genet.* **2014**, *22*, 558–563. [CrossRef]
144. Marat, A.L.; Haucke, V. Phosphatidylinositol 3-phosphates-at the interface between cell signalling and membrane traffic. *EMBO J.* **2016**, *35*, 561–579. [CrossRef] [PubMed]
145. Zhu, L.; Hu, Z.; Liu, J.; Gao, J.; Lin, B. Gene expression profile analysis identifies metastasis and chemoresistance-associated genes in epithelial ovarian carcinoma cells. *Med. Oncol* **2015**, *32*, 426. [CrossRef]
146. Doherty, J.; Baehrecke, E.H. Life, death and autophagy. *Nat. Cell Biol.* **2018**, *20*, 1110. [CrossRef]
147. Li, J.; Yang, B.; Zhou, Q.; Wu, Y.; Shang, D.; Guo, Y.; Song, Z.; Zheng, Q.; Xiong, J. Autophagy promotes hepatocellular carcinoma cell invasion through activation of epithelial–mesenchymal transition. *Carcinogenesis* **2013**, *34*, 1343–1351. [CrossRef]
148. Mohler, J.; Weiss, N.; Murli, S.; Mohammadi, S.; Vani, K.; Vasilakis, G.; Song, C.H.; Epstein, A.; Kuang, T.; English, J. The embryonically active gene, unkempt, of Drosophila encodes a Cys3His finger protein. *Genetics* **1992**, *131*, 377–388.
149. 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. Control of a neuronal morphology program by an RNA-binding zinc finger protein, unkempt. *Genes Dev.* **2015**, *29*, 501–512. [CrossRef]
150. Murn, J.; Teplova, M.; Zarnack, K.; Shi, Y.; Patel, D.J. Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein unkempt. *Nat. Struct. Mol. Biol.* **2016**, *23*, 16–23. [CrossRef]
151. Parkinson, H.; Kapushesky, M.; Kolesnikov, N.; Rustici, G.; Shojatalab, M.; Abeygunawardena, N.; Berube, H.; Dylag, M.; Emam, I.; Farne, A.; et al. Array express update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* **2009**, *37*, D868–D872. [CrossRef] [PubMed]
152. Overton, I.M.; Graham, S.; Gould, K.A.; Hinds, J.; Botting, C.H.; Shirran, S.; Barton, G.J.; Coote, P.J. Global network analysis of drug tolerance, mode of action and virulence in methicillin-resistant S. aureus. *BMC Syst. Biol.* **2011**, *5*, 68. [CrossRef] [PubMed]
153. Ben-Hur, A.; Noble, W.S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **2005**, *21*, i38–i46. [CrossRef]
154. Van Rijsbergen, C.J. *Information Retrieval*; Butterworths: London, UK, 1979.

155. Zhou, N.; Jiang, Y.; Bergquist, T.R.; Lee, A.J.; Kacsoh, B.Z.; Crocker, A.W.; Lewis, K.A.; Georghiou, G.; Nguyen, H.N.; Hamid, M.N.; et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **2019**, *20*, 244. [CrossRef]
156. Park, Y.; Marcotte, E.M. Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* **2012**, *9*, 1134–1136. [CrossRef] [PubMed]
157. Storey, J.D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2002**, *64*, 479–498. [CrossRef]
158. Ford, L.R.; Fulkerson, D.R. Maximal flow through a network. *Can. J. Math.* **1956**, *8*, 399–404. [CrossRef]
159. Lubbock, A.L.R.; Katz, E.; Harrison, D.J.; Overton, I.M. TMA navigator: Network inference, patient stratification and survival analysis with tissue microarray data. *Nucleic Acids Res.* **2013**, *41*, W562–W568. [CrossRef]
160. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38. [CrossRef]
161. Yamada, T.; Bork, P. Evolution of biomolecular networks-lessons from metabolic and protein interactions. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 791–803. [CrossRef]
162. Fitzgibbon, M.; Li, Q.; McIntosh, M. Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.* **2008**, *7*, 35–39. [CrossRef] [PubMed]
163. Sennels, L.; Bukowski-Wills, J.-C.; Rappsilber, J. Improved results in proteomics by use of local and peptide-class specific false discovery rates. *BMC Bioinform.* **2009**, *10*, 179. [CrossRef] [PubMed]
164. Raj, A.; van Oudenaarden, A. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **2008**, *135*, 216–226. [CrossRef] [PubMed]
165. Raj, A.; Rifkin, S.A.; Andersen, E.; van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **2010**, *463*, 913–918. [CrossRef]
166. Marusyk, A.; Almendro, V.; Polyak, K. Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **2012**, *12*, 323–334. [CrossRef]
167. Efron, B.; Tibshirani, R.; Storey, J.D.; Tusher, V. Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1160. [CrossRef]
168. López-Schier, H.; St Johnston, D. Delta signaling from the germ line controls the proliferation and differentiation of the somatic follicle cells during Drosophila oogenesis. *Genes Dev.* **2001**, *15*, 1393–1405. [CrossRef]
169. Schmitt, A.; Nebreda, A.R. Signalling pathways in oocyte meiotic maturation. *J. Cell Sci* **2002**, *115*, 2457–2459.
170. Acharya, U.; Patel, S.; Koundakjian, E.; Nagashima, K.; Han, X.; Acharya, J.K. Modulating sphingolipid biosynthetic pathway rescues photoreceptor degeneration. *Science* **2003**, *299*, 1740–1743. [CrossRef]
171. 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. Ceramide kinase regulates phospholipase C and phosphatidylinositol 4, 5, bisphosphate in phototransduction. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 20063–20068. [CrossRef]
172. Yonamine, I.; Bamba, T.; Nirala, N.K.; Jesmin, N.; Kosakowska-Cholody, T.; Nagashima, K.; Fukusaki, E.; Acharya, J.K.; Acharya, U. Sphingosine kinases and their metabolites modulate endolysosomal trafficking in photoreceptors. *J. Cell Biol* **2011**, *192*, 557–567. [CrossRef] [PubMed]
173. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
174. Stathopoulos, A.; Van Drenth, M.; Erives, A.; Markstein, M.; Levine, M. Whole-genome analysis of dorsal-ventral patterning in the drosophila embryo. *Cell* **2002**, *111*, 687–701. [CrossRef]
175. Campos-Ortega, J.A.; Hartenstein, V. *The Embryonic Development of Drosophila Melanogaster*, 2nd ed.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1997; ISBN 978-3-662-22489-2.
176. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21*, 3448–3449. [CrossRef] [PubMed]
177. Gramates, L.S.; Marygold, S.J.; dos Santos, G.; Urbano, J.-M.; Antonazzo, G.; Matthews, B.B.; Rey, A.J.; Tabone, C.J.; Crosby, M.A.; Emmert, D.B.; et al. Fly base at 25: Looking to the future. *Nucleic Acids Res.* **2017**, *45*, D663–D671. [CrossRef] [PubMed]

178. 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. Evolving gene/transcript definitions significantly alter the interpretation of gene chip data. *Nucleic Acids Res.* **2005**, *33*, e175. [CrossRef]
179. Irizarry, R.A.; Bolstad, B.M.; Collin, F.; Cope, L.M.; Hobbs, B.; Speed, T.P. Summaries of affymetrix gene chip probe level data. *Nucleic Acids Res.* **2003**, *31*, e15. [CrossRef]
180. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef]
181. Sims, A.H.; Smethurst, G.J.; Hey, Y.; Okoniewski, M.J.; Pepper, S.D.; Howell, A.; Miller, C.J.; Clarke, R.B. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets-improving meta-analysis and prediction of prognosis. *BMC Med. Genom.* **2008**, *1*, 42. [CrossRef]
182. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef]
183. Östlund, G.; Schmitt, T.; Forslund, K.; Köstler, T.; Messina, D.N.; Roopra, S.; Frings, O.; Sonnhammer, E.L.L. In paranoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **2009**, *38*, D196–D203. [CrossRef] [PubMed]
184. Liu, J.; Jiang, G.; Liu, S.; Liu, Z.; Pan, H.; Yao, R.; Liang, J. Lentivirus-delivered short hairpin RNA targeting SNAIL inhibits HepG2 cell growth. *Oncol. Rep.* **2013**, *30*, 1483–1487. [CrossRef] [PubMed]
185. Peluso, S.; Douglas, A.; Hill, A.; Angelis, C.D.; Moore, B.L.; Grimes, G.; Petrovich, G.; Essafi, A.; Hill, R.E. Fibroblast growth factors (FGFs) prime the limb specific Shh enhancer for chromatin changes that balance histone acetylation mediated by E26 transformation-specific (ETS) factors. *eLife* **2017**, *6*, e28590. [CrossRef] [PubMed]
186. 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. A Wt1-controlled chromatin switching mechanism underpins tissue-specific Wnt4 activation and repression. *Dev. Cell* **2011**, *21*, 559–574. [CrossRef]