Supplementary Information

Master regulators of FGFR2 signalling and breast cancer risk.

Michael NC Fletcher[1], Mauro AA Castro[1], Xin Wang, Ines de Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce AJ Ponder, Florian Markowetz, Kerstin B Meyer.

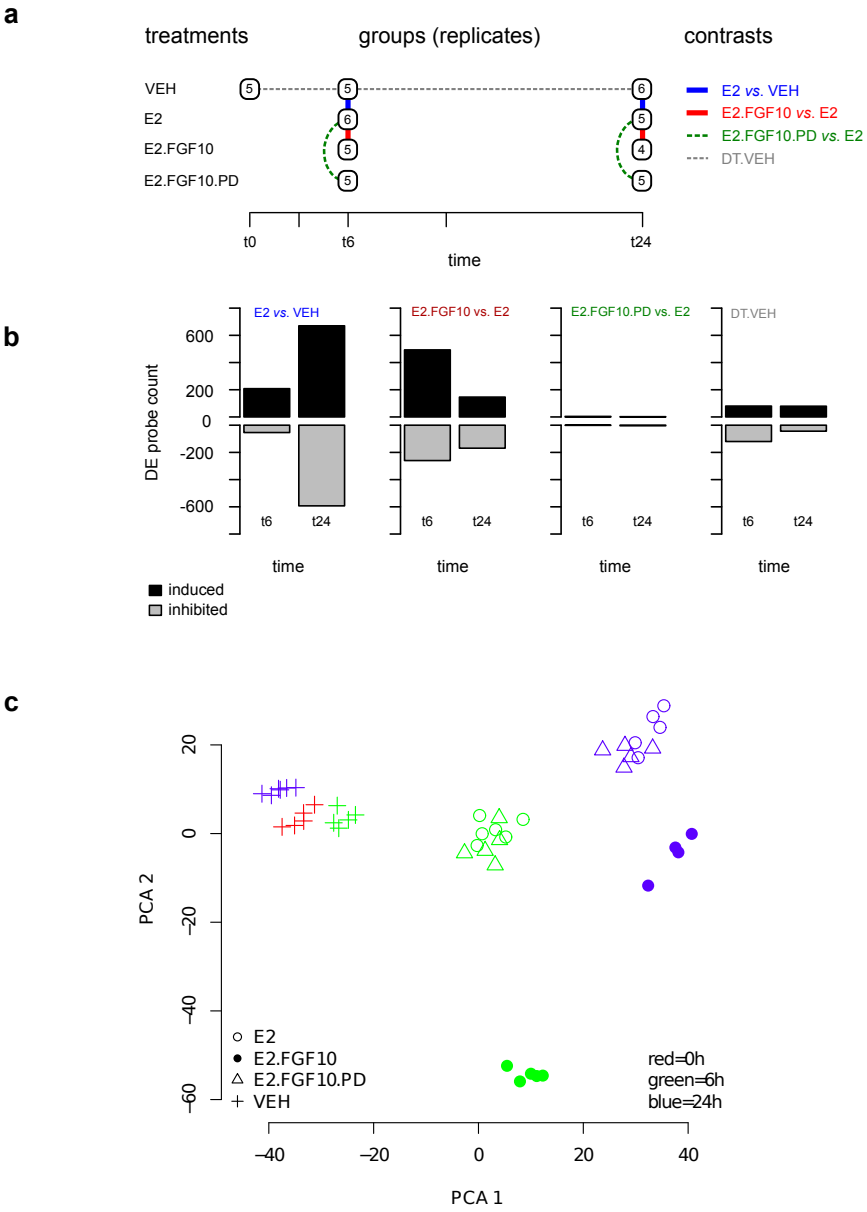
Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way Cambridge, CB2 0RE, UK. Email: kerstin.meyer@cancer.org.uk.

Supplementary Figures

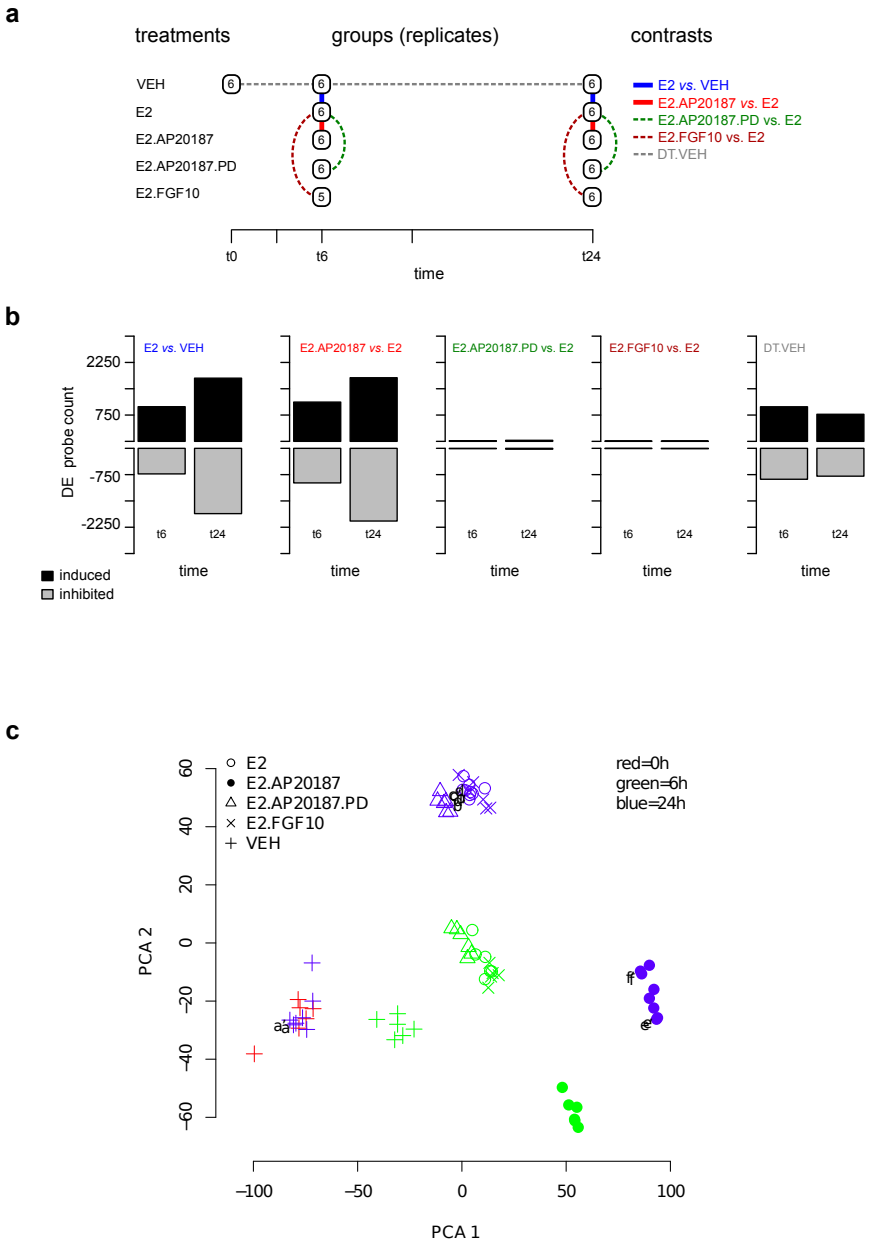
Supplementary Tables

Supplementary Methods

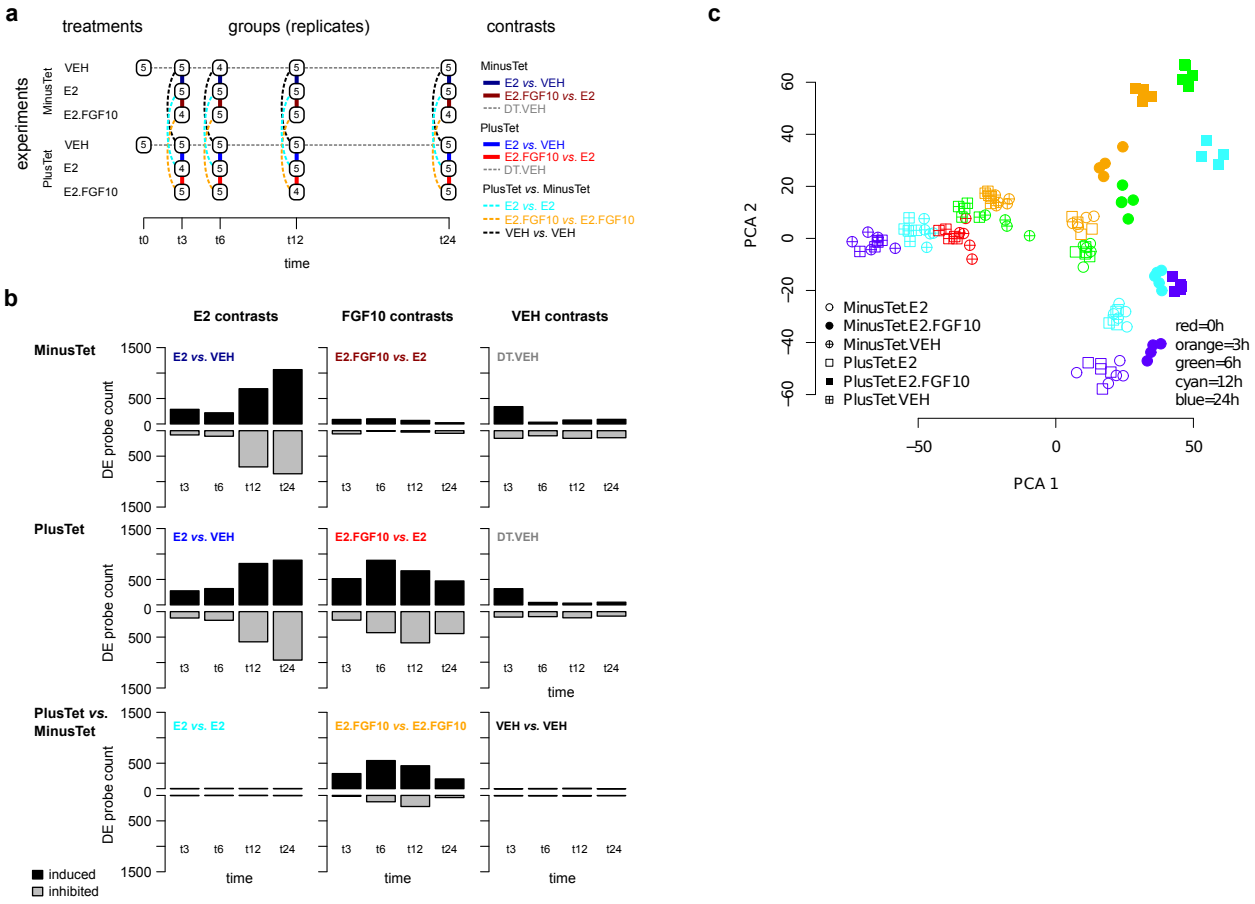
[1]joint first authors



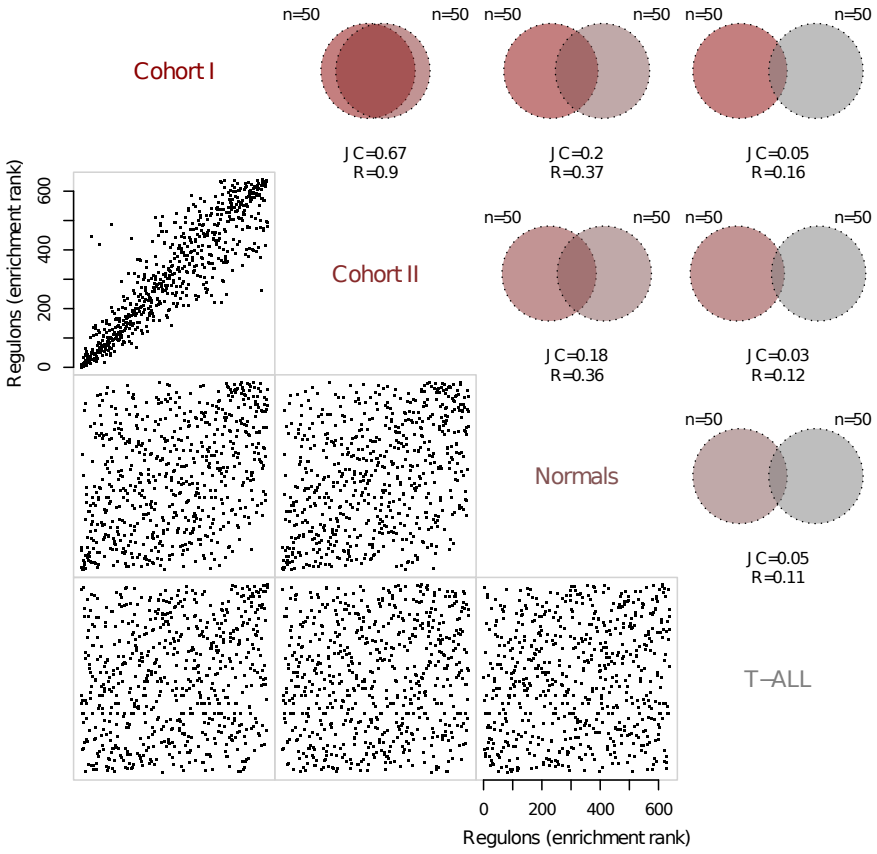
Supplementary Figure S1: **Experimental design of endogenous FGFRs perturbation experiments and summary of the differential expression analysis.** (**a**) The number of biological replicates analysed per condition (VEH: vehicle; E2: estradiol; FGF10; PD: FGFR kinase inhibitor PD173074) and time point. The limma contrasts that were calculated are shown by coloured lines. The results are shown in (**b**) by bar charts depicting the number of probes significantly deregulated at each time point after stimulation ($P<0.01$, limma moderated t-statistics). (**c**) PCA analysis of variation observed for the differentially expressed genes (n=2141) in the microarray analysis.



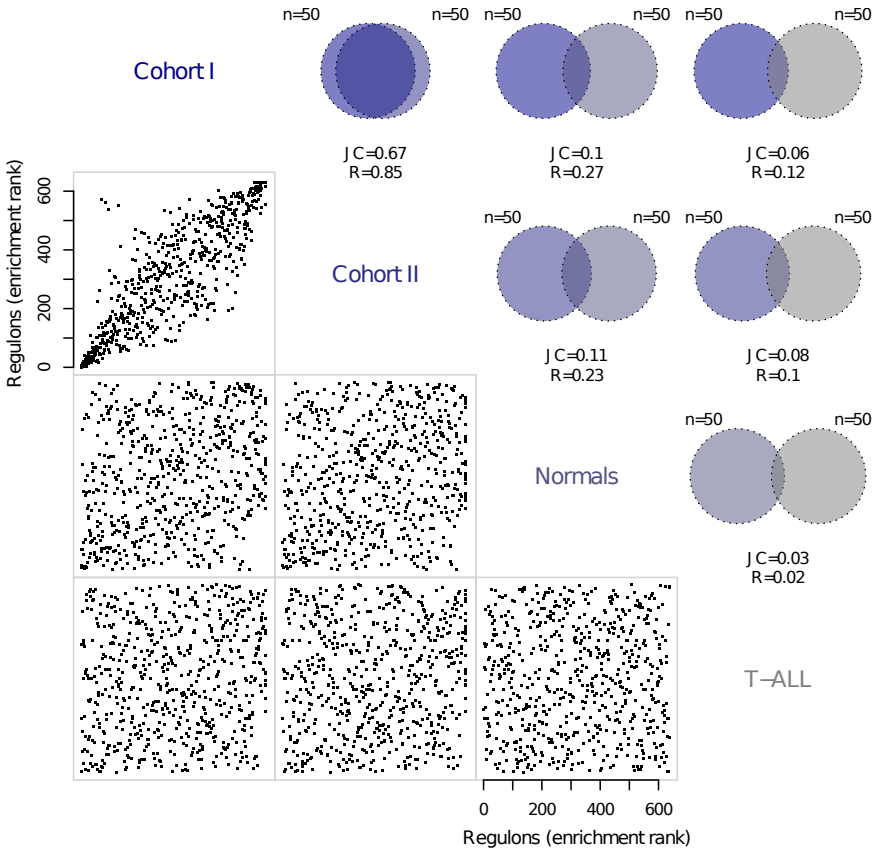
Supplementary Figure S2: **Experimental design of iF2 construct perturbation experiments and summary of the differential expression analysis.** (**a**) The number of biological replicates analysed per condition (VEH: vehicle; E2: estradiol; FGF10: FGFR; AP20187) and time point. The limma contrasts that were calculated are shown by coloured lines and results given in (**b**) with bar charts depicting the number of probes significantly deregulated at each time point after stimulation (P<0.01, limma moderated t-statistics). (**c**) PCA analysis of variation observed for the DE genes (n= 7647) in the microarray analysis. The letters a-f and a'-f' indicate technical repeats included in the microarray experiments.



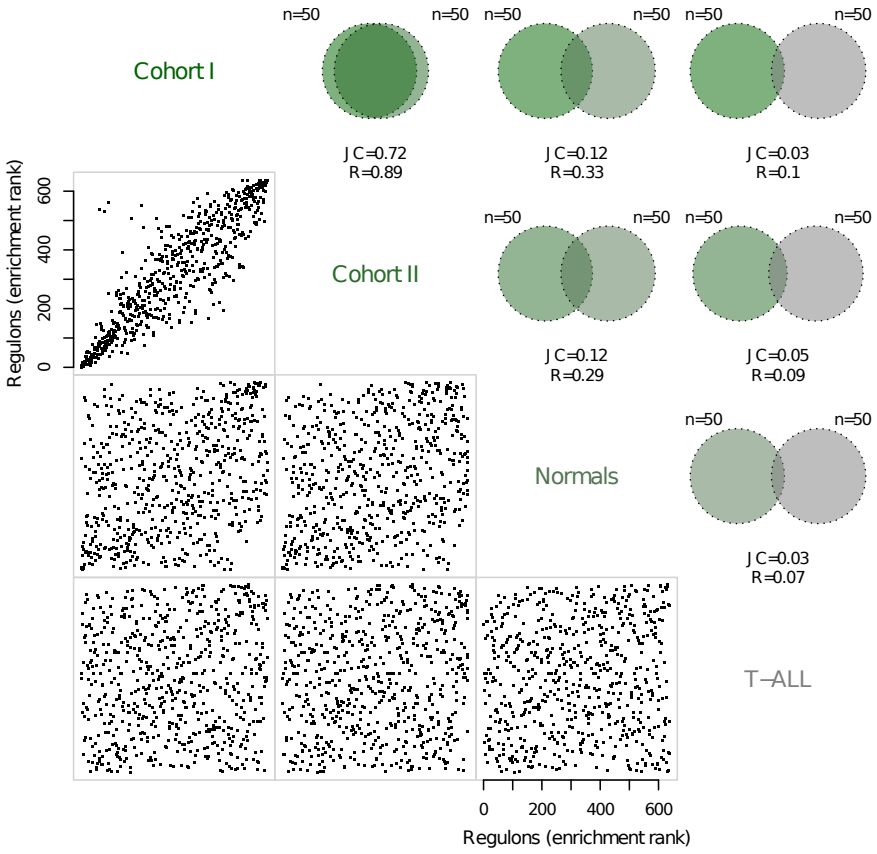
Supplementary Figure S3: **Experimental design of FGFR2b perturbation experiments and summary of the differential expression analysis.** (**a**) The number of biological replicates analysed per condition (VEH: vehicle; E2: estradiol; FGF10: FGFR; Tet: tetracycline) and time point. The limma contrasts that were calculated are shown by coloured lines and results given in (**b**) with bar charts depicting the number of probes significantly deregulated at each time point after stimulation ($P<0.01$, limma moderated t-statistics). (**c**) PCA analysis of variation observed for the DE genes (n=2519) in the microarray analysis.



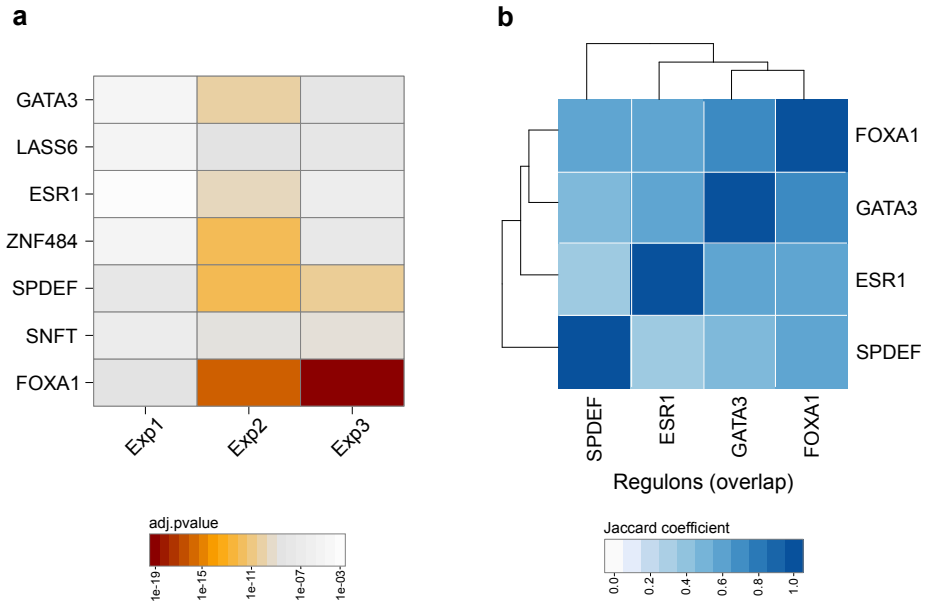
Supplementary Figure S4: **MRA agreement among regulons derived from different transcriptional networks (TN).** Regulons are ranked by the enrichment p-value estimated for E2.FGF10 signature (*Exp1*) and the graphs show the comparisons of regulon rank for cohort I, cohort II, normal breast tissue and T-ALL for all regulons. The correlation coefficient R is given for each comparison. The Venn diagrams depict the same comparison, but showing only the overlap obtained for the top 50 ranks, quantified by the Jaccard Coefficient (JC).



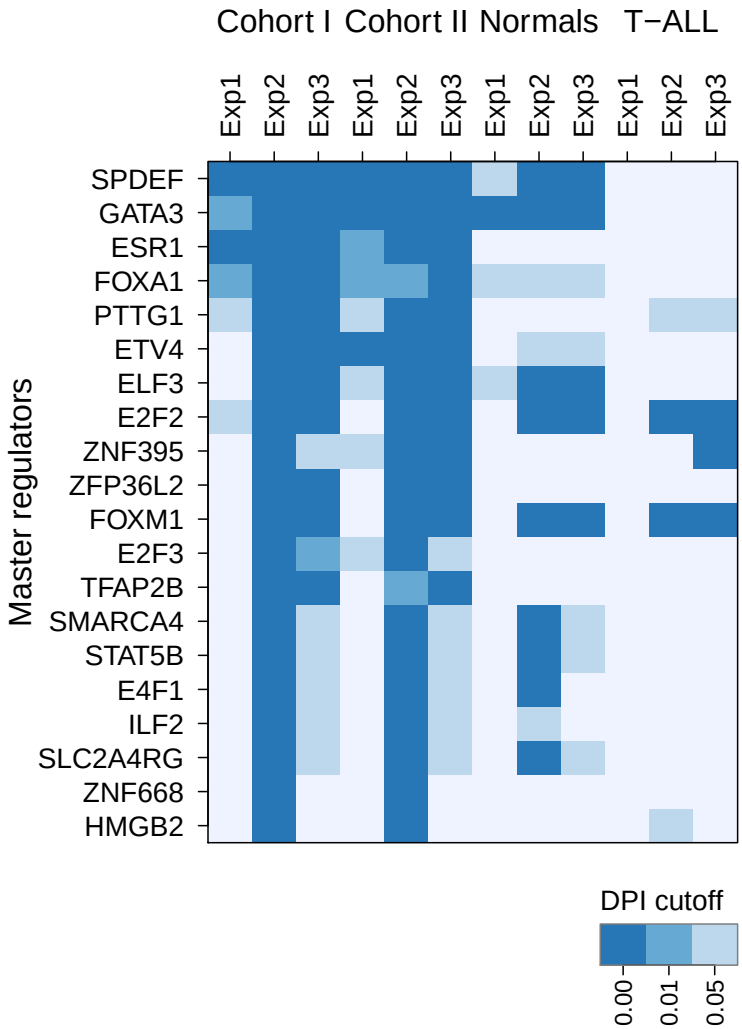
Supplementary Figure S5: **MRA agreement among regulons derived from different transcriptional networks (TN).** Regulons are ranked by the enrichment p-value estimated for E2.AP20187 signature (iF2 construct perturbation experiments) (*Exp2*) and the graphs show the comparisons of regulon rank for cohort I, cohort II, normal breast tissue and T-ALL for all regulons. The correlation coefficient R is given for each comparison. The Venn diagrams depict the same comparison, but showing only the overlap obtained for the top 50 ranks, quantified by the Jaccard Coefficient (JC).



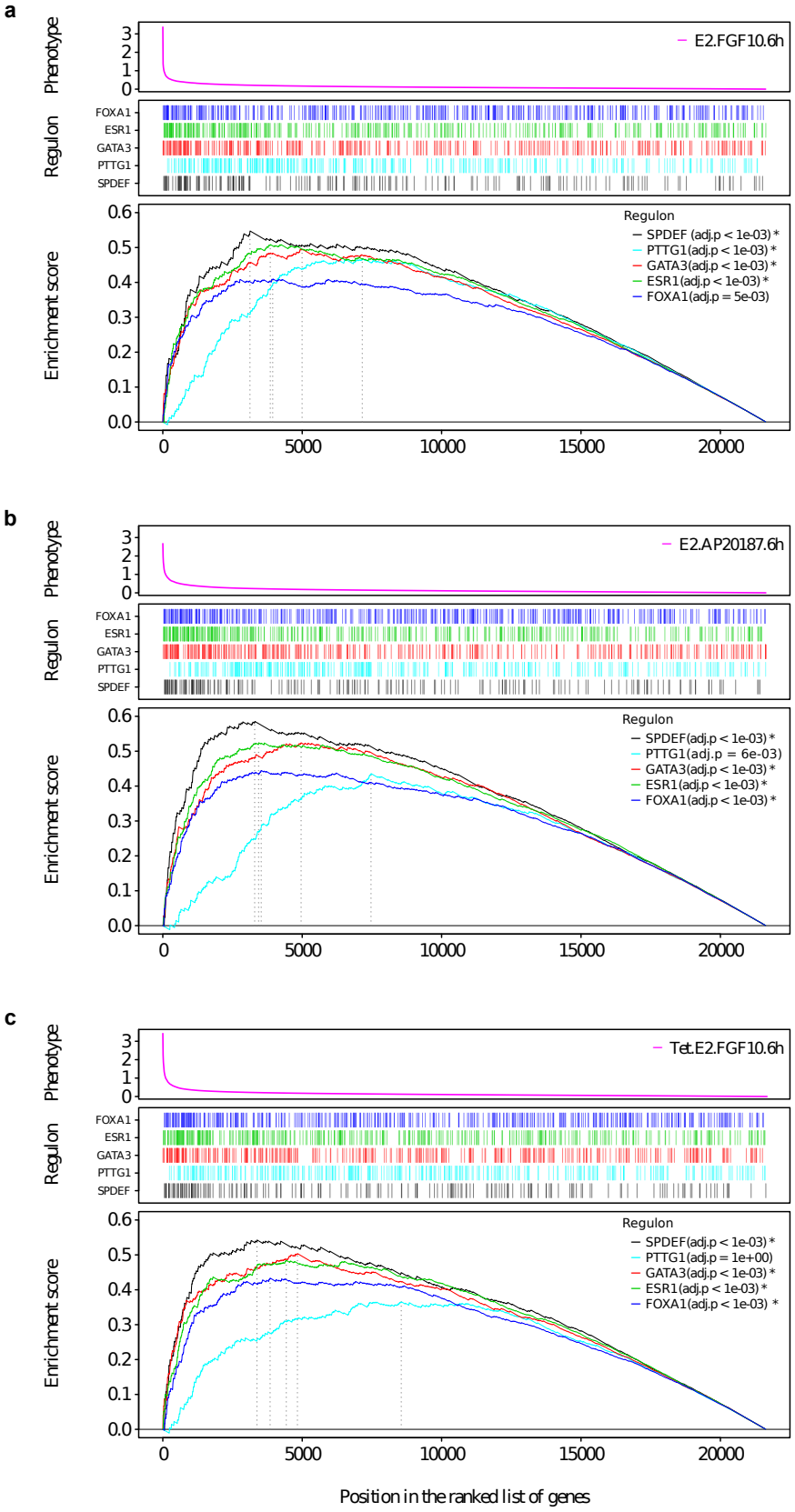
Supplementary Figure S6: **MRA agreement among regulons derived from different transcriptional networks (TN).** Regulons are ranked by the enrichment p-value estimated for PT.E2.FGF10 signature (FGFR2b perturbation experiments) (*Exp3*) and the graphs show the comparisons of regulon rank for cohort I, cohort II, normal breast tissue and T-ALL for all regulons. The correlation coefficient R is given for each comparison. The Venn diagrams depict the same comparison, but showing only the overlap obtained for the top 50 ranks, quantified by the Jaccard Coefficient (JC).



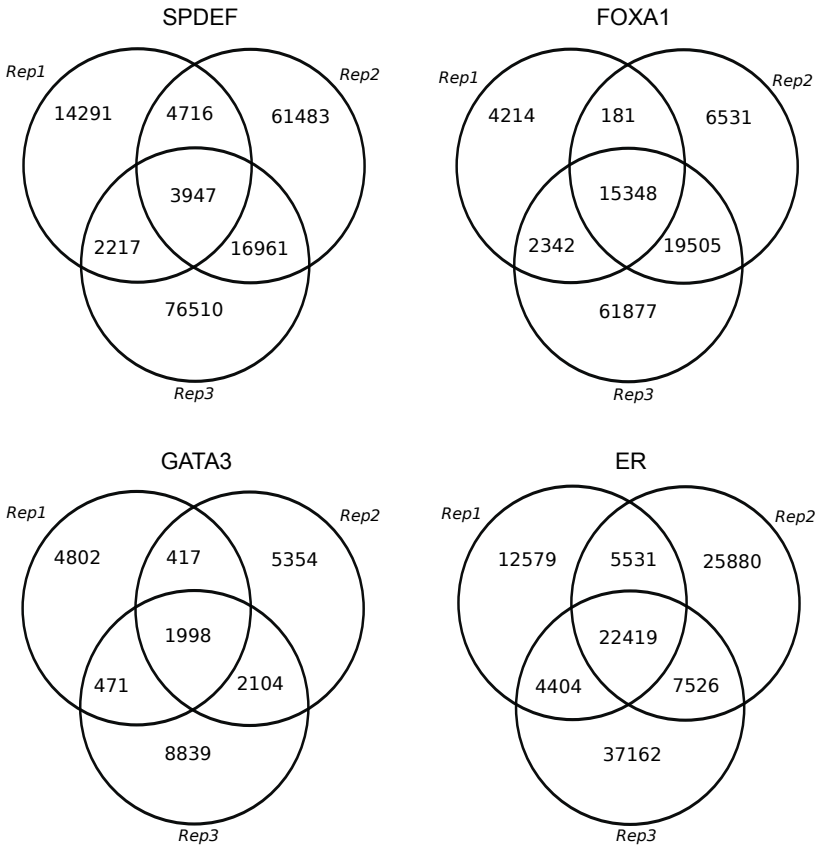
Supplementary Figure S7: **MRA analysis using the TCGA derived transcriptional network.** (**a**) A transcriptional network was calculated on expression profiles from the TCGA data set and master regulators (MRs) of FGFR2 signaling were determined by MRA using the FGFR2 signatures obtained in the three experiments. The heatmap shows the MRs found in all three experiments (*Exp1-3*), and their enrichment (in shades of orange) estimated from the MRA analysis. (**b**) The heatmap shows the hierarchical clustering on the Jaccard similarity coefficient (in shades of blue) focused on the overlap between the regulons of FOXA1, GATA3, ESR1 and SPDEF.



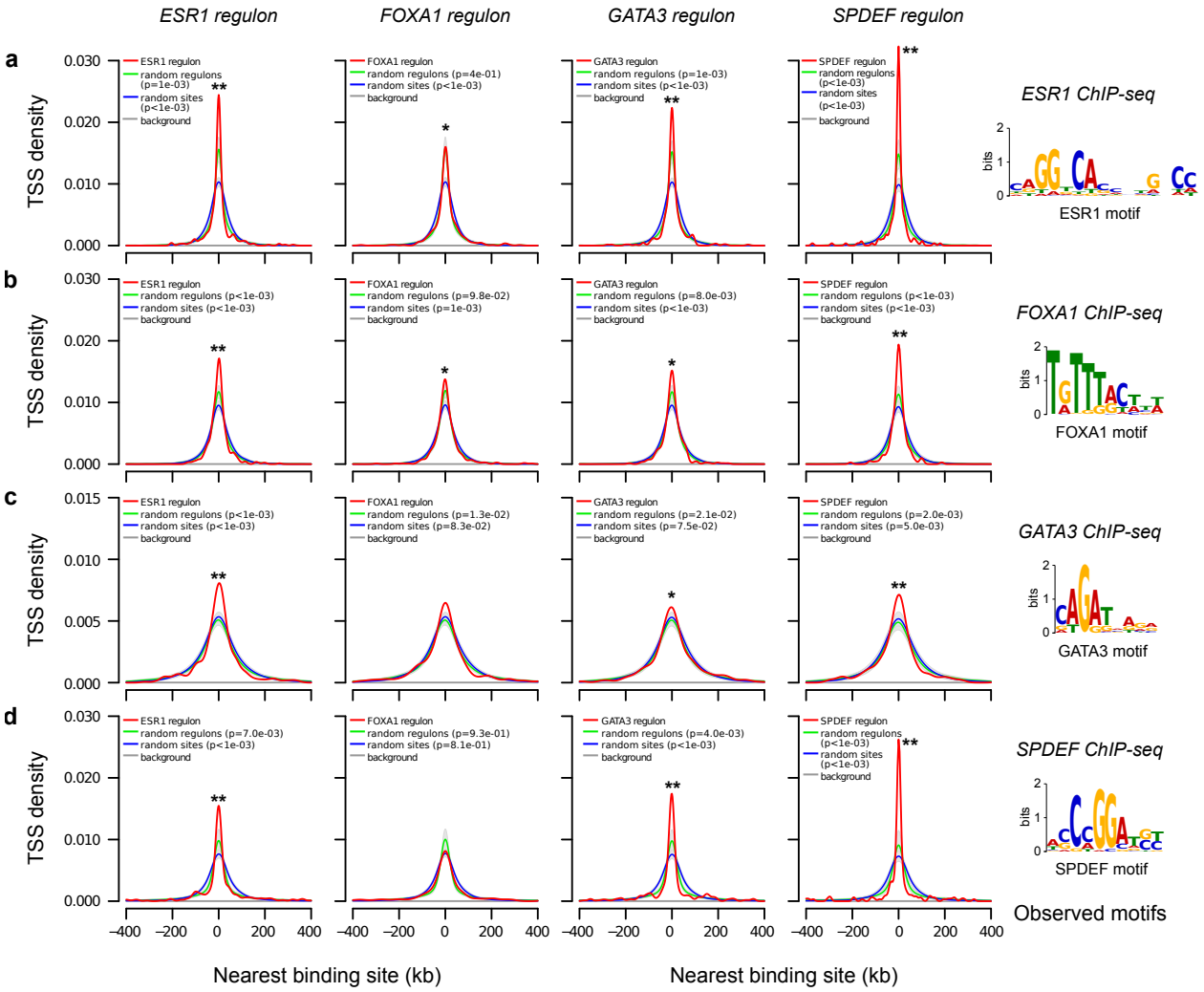
Supplementary Figure S8: **MRA analysis using different DPI thresholds.** Enrichment of regulons in either one or two cohorts in at least two experimental systems. Shading indicates significant enrichment in the transcriptional network computed with a DPI threshold of 0.00, 0.01 or 0.05.



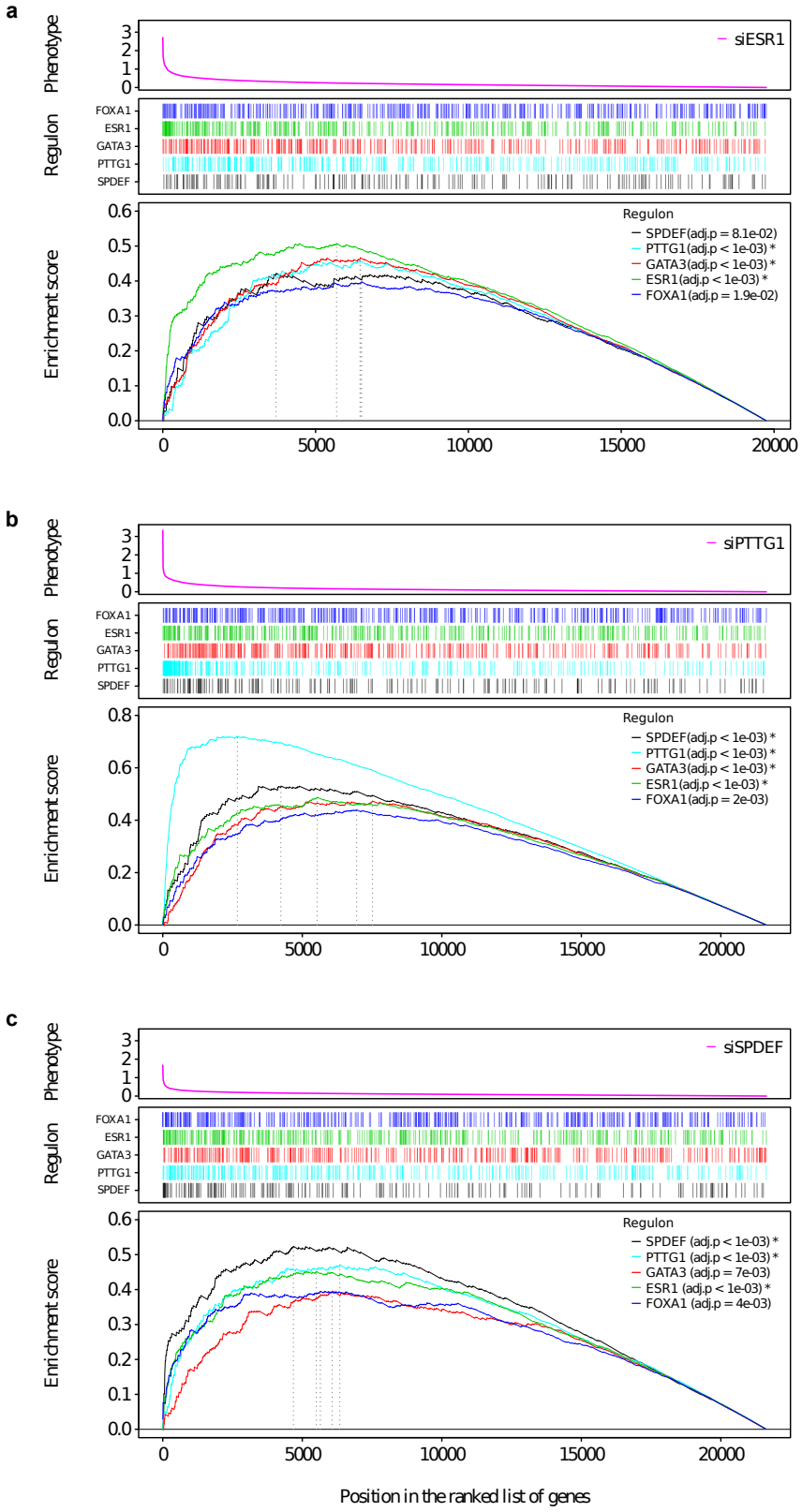
Supplementary Figure S9: **GSEA of the genes in each of the 5 master regulators.** MR regulons (from the DPI-filtered TN) are ranked by their response to FGFR2 signalling using the expression signatures *Exp1* (**a**), *Exp2* (**b**) and *Exp3* (**c**). *: $P < 0.001$, weighted Kolmogorov-Smirnov test.



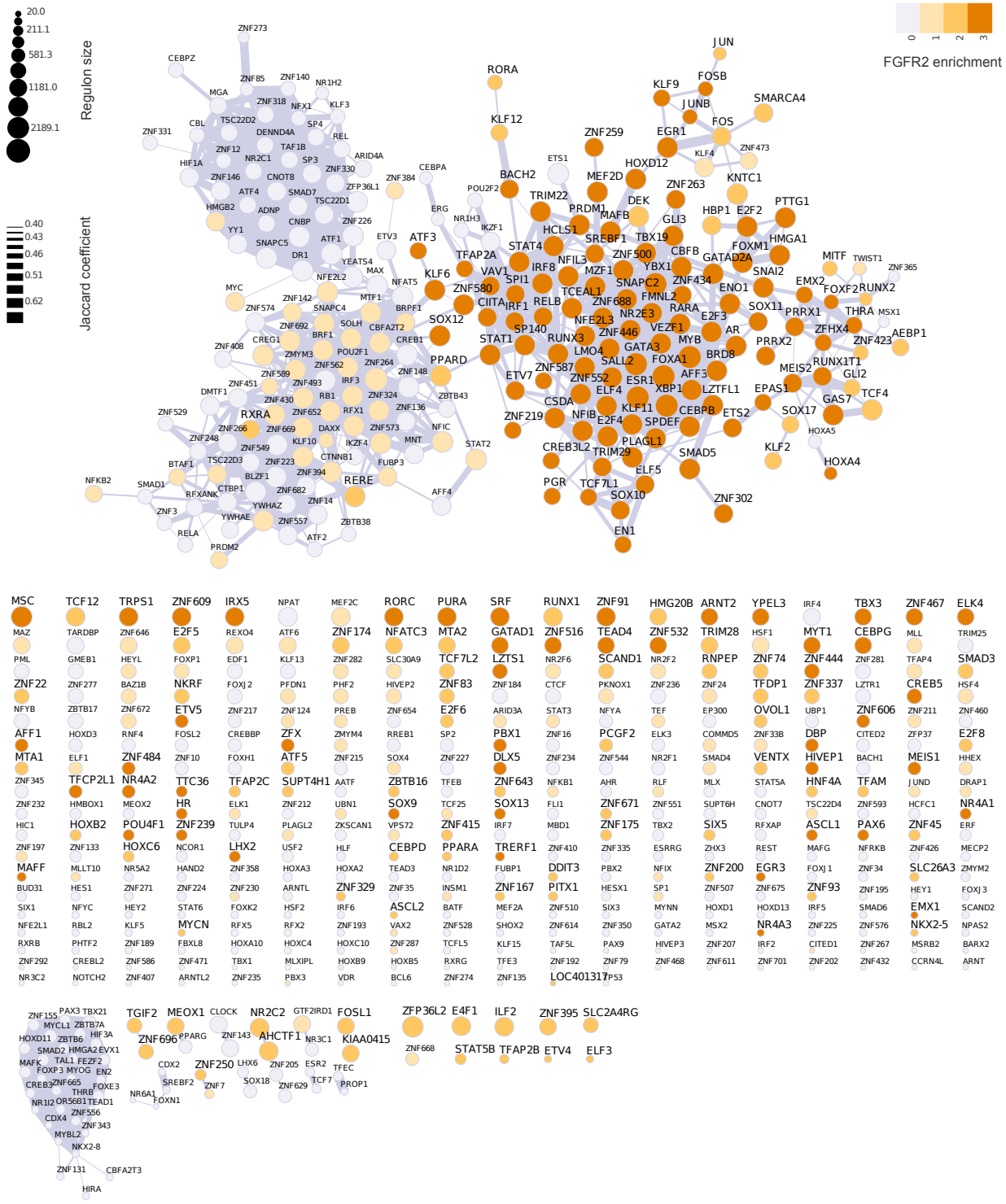
Supplementary Figure S10: **Overlapping peaks between triplicates of SPDEF, FOXA1, GATA3 and ER ChIP-seq data.** Only binding events that occurred in at least two out of three biological replicates were considered for further downstream network analysis. Venn Diagrams were obtained using the ChIPpeakAnno R package [60].



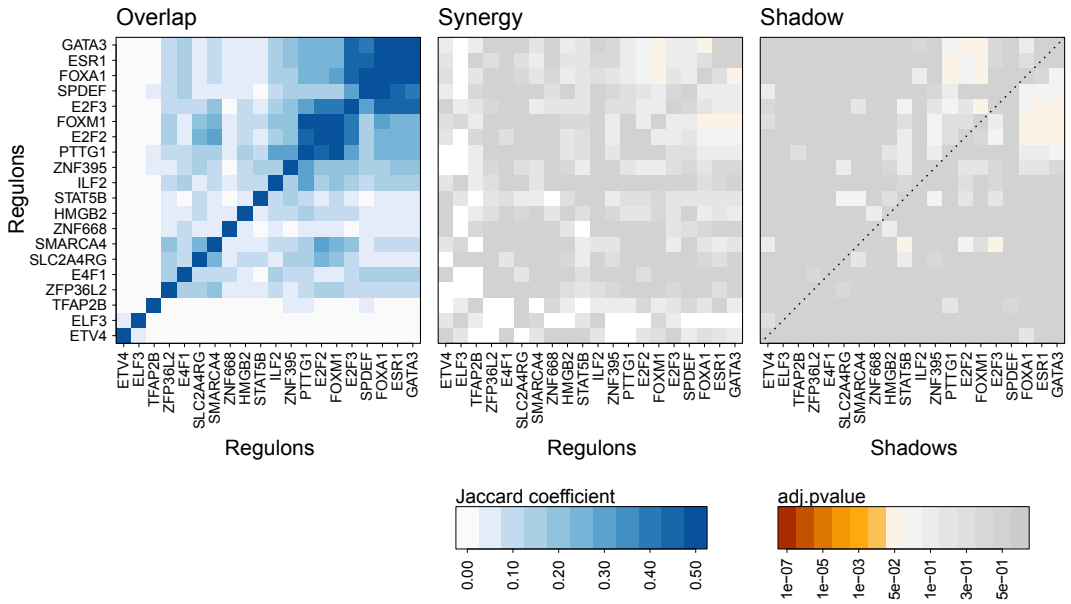
Supplementary Figure S11: **Regulons are enriched in experimentally determined binding sites.** Enrichment of binding sites of the ESR1, FOXA1, GATA3 and SPDEF regulons in ChIP-seq data obtained in MCF-7 cells for ESR1 (**a**) FOXA1 (**b**) GATA3 (**c**), and SPDEF (**d**). A background distribution is shown as a reference line (grey line) and represents the distance between the TSS and a random peak placed in the same chromosome. Observed distances (red line) were compared to random regulons (green line; mean ±SD; n=1000) and random sites (blue line; mean ±SD; n=1000) by permutation analyses. *: observed regulon/sites are statistically significantly different (P<0.05, permutation analysis) to random sites, but not random regulons. **: observed regulons/sites significantly different (P<0.05, permutation analysis) to both random sites and random regulons (regulons from the DPI-filtered TN).



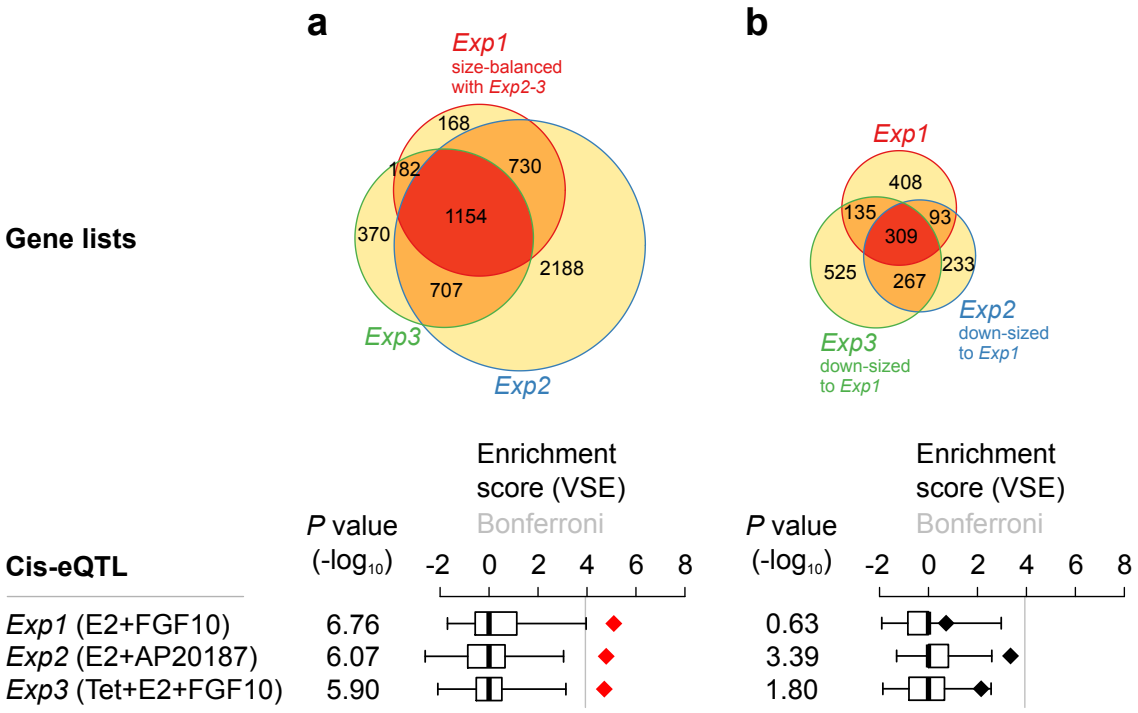
Supplementary Figure S12: **Enrichment of MR regulons with differentially expressed genes from knock-down experiments.** The GSEA plots show the MR regulons ranked by their response to (**a**) ESR1, (**b**) PTTG1 and (**c**) SPDEF siRNAs. These GSEA plots are complementary to the results presented in the **Supplementary Table S4**, which shows an enrichment analysis using hypergeometric test (MRs from the DPI-filtered TN). *: $P < 0.001$, weighted Kolmogorov-Smirnov test.



Supplementary Figure S13: **Fully annotated enrichment map derived from the transcriptional network (TN) in breast cancer.** Edges depicts the overlap of regulons and shades of orange indicate degree of enrichment of a regulon in at least one of the three FGFR2 gene signatures.



Supplementary Figure S14: **Statistical analysis of the overlap of regulons computed for the transcriptional network (TN).** The overlap, synergy and shadowing are depicted. Shadowing is only computed for those regulons whose overlap is significant.



Supplementary Figure S15: **Enrichment of the breast cancer AVS in FGFR2-related gene loci using size-balanced gene lists.** (**a**) Venn diagram showing the overlap between the genes deregulated after FGFR2 signalling in the experimental systems *Exp1-3* using the size-balanced *Exp1* and the original *Exp2* and *Exp3* gene lists. (**b**) Venn diagram showing the original *Exp1* gene list with the downsized *Exp2* and *Exp3* gene lists. The box plots show the normalized null distribution of the EVSE analysis using the breast cancer AVS and the gene lists in each panel.

Supplementary Table S1. MRA analysis[1] using the meta-PCNA signature on the filtered transcriptional network calculated for Metabric cohort I.

MRA rank[2]	Regulon	Universe size[3]	Regulon size[4]	Total hits[5]	Expected hits[6]	Observed hits[7]	P-value	Adjusted pvalue
1	PTTG1	19747	352	128	2.28	49	3.36E-54	1.81E-51
2	FOXM1	19747	398	128	2.58	41	1.65E-39	4.43E-37
3	LHX2	19747	45	128	0.29	13	1.57E-20	2.81E-18
4	STAT5B	19747	112	128	0.73	14	5.97E-16	8.03E-14
5	E2F2	19747	216	128	1.40	16	5.15E-14	5.54E-12
6	HMGB2	19747	161	128	1.04	14	1.44E-13	1.29E-11
7	ZNF395	19747	137	128	0.89	11	9.13E-11	7.02E-09
8	ZNF167	19747	95	128	0.62	9	5.82E-10	3.48E-08
9	TGIF2	19747	46	128	0.30	7	5.30E-10	3.48E-08
10	ZNF423	19747	53	128	0.34	7	1.73E-09	9.33E-08
11	ILF2	19747	328	128	2.13	13	2.92E-08	1.43E-06
12	KLF9	19747	117	128	0.76	8	6.97E-08	3.12E-06
13	HOXD13	19747	86	128	0.56	7	8.68E-08	3.59E-06
14	PURA	19747	98	128	0.64	7	2.41E-07	9.28E-06
15	DEK	19747	149	128	0.97	8	5.53E-07	1.98E-05
16	KNTC1	19747	129	128	0.84	7	1.98E-06	6.65E-05
17	NR2C2	19747	38	128	0.25	4	4.47E-06	1.42E-04
18	ZNF93	19747	20	128	0.13	3	7.53E-06	2.03E-04
19	ELK4	19747	42	128	0.27	4	7.43E-06	2.03E-04
20	E2F6	19747	42	128	0.27	4	7.43E-06	2.03E-04
21	E2F5	19747	112	128	0.73	6	8.40E-06	2.15E-04
22	RUNX1T1	19747	272	128	1.76	9	1.13E-05	2.77E-04
23	E2F3	19747	93	128	0.60	5	3.17E-05	7.41E-04
//								
446	GATA3	19747	221	128	1.43	1	4.21E-01	5.07E-01
518	FOXA1	19747	331	128	2.15	1	6.35E-01	6.60E-01
533	SPDEF	19747	226	128	1.46	0	7.72E-01	7.79E-01
537	ESR1	19747	352	128	2.28	0	9.01E-01	9.02E-01

Supplementary Table S1 (*continued*). MRA analysis using the meta-PCNA signature on the filtered transcriptional network calculated for Metabric cohort II.

MRA rank[2]	Regulon	Universe size[3]	Regulon size[4]	Total hits[5]	Expected hits[6]	Observed hits[7]	P-value	Adjusted pvalue
1	PTTG1	19747	322	128	2.09	46	2.49E-51	1.32E-48
2	FOXM1	19747	453	128	2.94	40	8.90E-36	2.36E-33
3	E2F8	19747	23	128	0.15	13	8.65E-26	1.53E-23
4	E2F2	19747	225	128	1.46	25	2.47E-25	3.28E-23
5	HMGB2	19747	127	128	0.82	16	6.19E-18	6.57E-16
6	ILF2	19747	249	128	1.61	17	3.87E-14	3.43E-12
7	VENTX	19747	75	128	0.49	10	1.89E-12	1.43E-10
8	TGIF2	19747	57	128	0.37	9	3.06E-12	2.03E-10
9	ZNF395	19747	137	128	0.89	12	5.20E-12	3.07E-10
10	PURA	19747	128	128	0.83	11	4.09E-11	2.17E-09
11	ZNF219	19747	47	128	0.30	7	6.36E-10	3.07E-08
12	ZNF643	19747	30	128	0.19	6	7.33E-10	3.24E-08
13	KLF9	19747	105	128	0.68	9	1.57E-09	6.43E-08
14	DEK	19747	136	128	0.88	9	1.96E-08	7.42E-07
15	E2F5	19747	111	128	0.72	8	4.40E-08	1.56E-06
16	TFDP1	19747	81	128	0.53	7	5.41E-08	1.80E-06
17	VAX2	19747	34	128	0.22	5	7.64E-08	2.39E-06
18	NR2C2	19747	69	128	0.45	6	3.15E-07	9.29E-06
19	GAS7	19747	249	128	1.61	10	6.83E-07	1.91E-05
20	STAT5B	19747	82	128	0.53	6	1.04E-06	2.75E-05
21	BAZ1B	19747	57	128	0.37	5	1.83E-06	4.61E-05
22	TFAP2C	19747	68	128	0.44	5	5.19E-06	1.25E-04
23	E2F3	19747	203	128	1.32	8	7.13E-06	1.65E-04
24	ENO1	19747	77	128	0.50	5	1.07E-05	2.37E-04
25	ZBTB16	19747	22	128	0.14	3	1.13E-05	2.39E-04
26	ZNF423	19747	23	128	0.15	3	1.36E-05	2.77E-04
27	KNTC1	19747	172	128	1.11	7	1.66E-05	3.26E-04
28	ZNF671	19747	89	128	0.58	5	2.47E-05	4.68E-04
//								
473	FOXA1	19747	427	128	2.77	2	5.25E-01	5.90E-01
508	GATA3	19747	343	128	2.22	1	6.55E-01	6.84E-01
517	ESR1	19747	367	128	2.38	1	6.91E-01	7.09E-01
520	SPDEF	19747	187	128	1.21	0	7.05E-01	7.20E-01

[1] Significant regulons for $P < 0.001$ (hypergeometric test) are shown. MRs from the extended MR list are highlighted in green.

[2] In addition, the rank and adjusted p-value of all five MRs (relevant in all three experiments) are shown highlighted in red.

[3] Number of genes in the transcriptional network.

[4] Number of genes in a given regulon.

[5] Number of genes in the meta-PCNA signature.

[6] Expected overlap between "*total hits*" and "*regulon size*".

[7] Observed overlap between "*total hits*" and "*regulon size*".

Supplementary Table S2. Alignment rate and number of peaks per sample.

Experiment ID	PF reads	Aligned reads	Aligned reads	Number of peaks
* mf016_MCF7_SPDEF_Input	29,278,946	28,352,386	96.8%	NA
mf025_MCF7_SPDEF_CRI01	31,578,289	30,716,655	97.3%	25,171
mf026_MCF7_SPDEF_CRI01	38,520,935	37,397,617	97.1%	99,635
mf027_MCF7_SPDEF_CRI01	32,416,721	31,586,232	97.4%	87,107
jc189_FOXA1_ChIP_MCF7_full_CRI01	13,406,298	12,406,981	92.5%	22085
jc193_FOXA1_ChIP_MCF7_full_CRI01	26,940,662	26,475,802	98.3%	99072
jc368_FoxA1_FM_veh_3h_CRI01	31,976,286	31,306,658	97.9%	41565
jc485_MCF7_GATA3_E2_SAN01	35,745,318	34,800,750	97.4%	13412
jc556_MCF7-E2-GATA3_CRI01	31,992,134	30,459,215	95.2%	9873
jc633_MCF7-GATA3-E2_CRI01	26,415,088	24,449,926	92.6%	7688
* jc379_MCF_input_rep3_CRI01	28,226,691	27,858,051	98.7%	NA
jc780_MCF7_ER_E2_rep1_CRI01	58,609,743	55,231,539	94.2%	71511
mf002_MCF7_ERa_E2_45min_CRI01	36,978,860	35,673,329	96.5%	44933
mf006_MCF7_ERa_E2_45min_CRI01	27,258,888	24,846,928	91.2%	61356

*Input sequence

Supplementary Table S3. Consistency of ChIP-seq signal between three biological replicates of SPDEF, FOXA1, GATA3 and ESR1 data sets. For each set of triplicates, read coverage was used to score all non-overlapping genomic regions that were called in at least two out of the three replicates. The strength of the correlation between pairs of replicates is described by the Spearman's and the Pearson's correlation coefficients.

Transcription Factor	Replicates	Spearman's correlation coefficient	Pearson's correlation coefficient
SPDEF	*Replicate1 vs Replicate3*	0.615	0.979
	Replicate1 vs Replicate2	0.739	0.986
	Replicate2 vs Replicate3	0.726	0.980
FOXA1	*Replicate1 vs Replicate3*	0.673	0.808
	Replicate1 vs Replicate2	0.638	0.865
	Replicate2 vs Replicate3	0.638	0.623
GATA3	*Replicate1 vs Replicate3*	0.592	0.941
	Replicate1 vs Replicate2	0.592	0.963
	Replicate2 vs Replicate3	0.515	0.976
ESR1	*Replicate1 vs Replicate3*	0.679	0.797
	Replicate1 vs Replicate2	0.662	0.775
	Replicate2 vs Replicate3	0.712	0.853

Supplementary Table S4. Enrichment of regulons with differentially expressed genes after knock-down SPDEF, ESR1 and PTTG1 using siRNA.

	Regulon[1]	Universe size[2]	Regulon size[3]	Total hits[4]	Expected hits[5]	Observed hits[6]	P-value	Adjusted pvalue	
siESR1	ESR1	19747	367	1016	18.88	52	9.53E-12	1.91E-10	*
	FOXM1	19747	453	1016	23.31	50	1.30E-07	2.47E-06	*
	ZFP36L2	19747	177	1016	9.11	24	4.37E-06	7.86E-05	*
	TFAP2B	19747	107	1016	5.51	16	3.16E-05	5.38E-04	*
	GATA3	19747	343	1016	17.65	33	2.02E-04	3.24E-03	
	E2F2	19747	225	1016	11.58	24	2.49E-04	3.73E-03	
	PTTG1	19747	322	1016	16.57	31	2.94E-04	4.11E-03	
	SPDEF	19747	187	1016	9.62	19	1.59E-03	2.06E-02	
	ELF3	19747	77	1016	3.96	10	1.88E-03	2.25E-02	
	ZNF668	19747	106	1016	5.45	12	3.09E-03	3.40E-02	
	ETV4	19747	31	1016	1.59	5	4.45E-03	4.45E-02	
	E2F3	19747	203	1016	10.44	18	8.74E-03	7.14E-02	
	FOXA1	19747	427	1016	21.97	33	7.93E-03	7.14E-02	
	ZNF395	19747	137	1016	7.05	13	1.10E-02	7.72E-02	
	E4F1	19747	146	1016	7.51	13	1.84E-02	1.11E-01	
	SLC2A4RG	19747	171	1016	8.80	12	1.03E-01	5.17E-01	
	SMARCA4	19747	143	1016	7.36	10	1.19E-01	5.17E-01	
	HMGB2	19747	127	1016	6.53	8	2.07E-01	6.22E-01	
	ILF2	19747	249	1016	12.81	11	6.34E-01	1.00E+00	
	STAT5B	19747	82	1016	4.22	3	6.14E-01	1.00E+00	

	Regulon[1]	Universe size[2]	Regulon size[3]	Total hits[4]	Expected hits[5]	Observed hits[6]	P-value	Adjusted pvalue	
siPTTG1	PTTG1	19747	322	1051	17.14	111	2.32E-61	4.63E-60	*
	FOXM1	19747	453	1051	24.11	115	8.54E-48	1.62E-46	*
	HMGB2	19747	127	1051	6.76	48	1.04E-29	1.88E-28	*
	E2F2	19747	225	1051	11.98	52	1.79E-20	3.05E-19	*
	ILF2	19747	249	1051	13.25	48	1.23E-15	1.96E-14	*
	E2F3	19747	203	1051	10.80	42	4.50E-15	6.75E-14	*
	ZNF395	19747	137	1051	7.29	30	4.84E-12	6.78E-11	*
	STAT5B	19747	82	1051	4.36	20	1.19E-09	1.55E-08	*
	SPDEF	19747	187	1051	9.95	28	2.08E-07	2.49E-06	*
	SLC2A4RG	19747	171	1051	9.10	21	1.13E-04	1.24E-03	
	ESR1	19747	367	1051	19.53	35	3.14E-04	3.14E-03	
	FOXA1	19747	427	1051	22.73	39	3.87E-04	3.48E-03	
	GATA3	19747	343	1051	18.26	32	7.66E-04	6.13E-03	
	ZFP36L2	19747	177	1051	9.42	17	6.41E-03	4.49E-02	
	ELF3	19747	77	1051	4.10	8	2.08E-02	1.25E-01	
	TFAP2B	19747	107	1051	5.69	10	2.71E-02	1.35E-01	
	ETV4	19747	31	1051	1.65	3	8.03E-02	3.21E-01	
	ZNF668	19747	106	1051	5.64	4	6.72E-01	1.00E+00	
	E4F1	19747	146	1051	7.77	4	8.94E-01	1.00E+00	
	SMARCA4	19747	143	1051	7.61	6	6.44E-01	1.00E+00	

	Regulon[1]	Universe size[2]	Regulon size[3]	Total hits[4]	Expected hits[5]	Observed hits[6]	P-value	Adjusted pvalue	
siSPDEF	SPDEF	19747	187	201	1.90	13	7.75E-09	1.55E-07	*
	ESR1	19747	367	201	3.74	12	1.04E-04	1.98E-03	
	TFAP2B	19747	107	201	1.09	5	7.95E-04	1.43E-02	
	FOXA1	19747	427	201	4.35	11	1.48E-03	2.52E-02	
	ZNF395	19747	137	201	1.39	4	1.32E-02	2.11E-01	
	STAT5B	19747	82	201	0.83	2	5.13E-02	7.69E-01	
	SMARCA4	19747	143	201	1.46	3	5.86E-02	8.20E-01	
	SLC2A4RG	19747	171	201	1.74	2	2.53E-01	1.00E+00	
	PTTG1	19747	322	201	3.28	4	2.32E-01	1.00E+00	
	FOXM1	19747	453	201	4.61	3	6.80E-01	1.00E+00	
	E2F3	19747	203	201	2.07	3	1.53E-01	1.00E+00	
	E2F2	19747	225	201	2.29	3	1.97E-01	1.00E+00	
	ZFP36L2	19747	177	201	1.80	2	2.69E-01	1.00E+00	
	ILF2	19747	249	201	2.53	1	7.23E-01	1.00E+00	
	HMGB2	19747	127	201	1.29	2	1.40E-01	1.00E+00	
	ZNF668	19747	106	201	1.08	1	2.93E-01	1.00E+00	
	ELF3	19747	77	201	0.78	1	1.85E-01	1.00E+00	
	E4F1	19747	146	201	1.49	1	4.39E-01	1.00E+00	
	ETV4	19747	31	201	0.32	0	2.72E-01	1.00E+00	
	GATA3	19747	343	201	3.49	5	1.38E-01	1.00E+00	

[1] Only regulons of the extended MR list are given; MRs from the DPI-filtered TN.

[2] Number of genes in the transcriptional network.

[3] Number of genes in a given regulon.

[4] Number of genes in the transcriptional network which show significantly differential expression upon corresponding siRNA treatment.

[5] Expected overlap between "*total hits*" and "*regulon size*".

[6] Observed overlap between "*total hits*" and "*regulon size*".

* $P < 0.001$, hypergeometric test.

Supplementary Table S5. Overlap* of genome-wide TF binding sites from ChIP-seq data from MCF-7 cells.

	ESR1 (% of ESR1 sites)	**FOXA1** (% of FOXA1sites)	**GATA3** (% of GATA3 sites)	**SPDEF** (% of SPDEF sites)
ESR1	100%	38%	62%	17%
FOXA1	25%	100%	47%	6%
GATA3	13%	16%	100%	3%
SPDEF	4%	2 %	3%	100%

* We considered that there was an overlap of binding if peak summits mapped within a +/- 500bp window, which corresponds to the approximate average peak length in our data sets.

Supplementary Table S6. Oligonucleotide primer sequences used in this study (5' -> 3').

Gene expression	
DGUOK-for DGUOK-rev	GCTGGTGTTGGATGTCAATG GCCTGAACTTCATGGTATTGG
IL8-for IL8-rev	AAAGCTTTCTGATGGAAGAGAG CCAGGAATCTTGTATTGCATC
UBC-for UBC-rev	CAGAGGTGGGATGCAAATCT TTGCTTTGACGTTCTCGATG
ChIP-RT-PCR	
MYC enh for MYC enh rev	GCTCTGGGCACACACATTGG GGCTCACCCTTGCTGATGCT
GREB1 enh 3 for GREB1 enh 3 rev	GAAGGGCAGAGCTGATAACG GACCCAGTTGCCACACTTTT
EGR2 enh for EGR2 enh rev	AAAGGCCATCTCATCTGTGTTCC GAGGACATTTGGGAACAGATGG
TOX3 for TOX3 rev	GCCTGAAAGAGAATGTGATCTAAGATT CCCCAAAGAGTTGGCTGTTAAA
CCND1 for CCND1 rev	TGCCACACACCAGTGACTTT ACAGCCAGAAGCTCCAAAAA

Supplementary Methods

Microarray and principal component analysis

Gene expression was examined as described in the *Methods* section. **Supplementary Figures S1-S3** depict the experimental layout and the number of deregulated probes that were detected under each experimental condition. In *Exp1* estradiol induced and repressed approximately the same number of genes, while the response to FGF10 included more up- than down-regulated genes. This response was ablated by the broad-spectrum FGFR kinase inhibitor PD173074, demonstrating that the observed gene expression changes were FGFR-specific. Similar observations were made in *Exp2* with the dimerisable iF2 construct (**Supplementary Fig. S2b**), although here the number of FGFR regulated genes continued to increase over the time course of the experiment with up to 5000 probes showing differential expression. The kinetics of gene regulation after stimulation of overexpressed FGFR2 (**Supplementary Fig. S3b**) mimicked that of endogenous FGFR, with approximately 1500 differentially expressed genes (DEG). As a quality control principal component analysis was carried out. When using the whole probe universe (n=47231) the samples were distributed randomly. However, when the analysis was carried out with differentially regulated genes (significant gene counts for $P<0.01$, limma moderated t-statistics) the samples clustered according to the experimental conditions used (**Supplementary Figs. S1c, S2c and S3c**). Importantly the estradiol only stimulated samples cluster with the sample stimulated by the combination of estradiol, FGF10 and the FGFR kinase inhibitor (PD173074) demonstrating FGFR specificity of the signalling (**Supplementary Fig. S1c**). Similarly in *Exp2* E2 only treated samples cluster with those stimulated by E2, AP20187 and the kinase inhibitor. In this experiment FGF10 signalling alone has little effect, presumably because signalling molecules required by endogenous FGFR1b/2b are sequestered by the overexpressed iF2 signalling construct (**Supplementary Fig. S1b**). Technical replicates denoted a/a'-f/f' show little variation. In *Exp3* minusTet and plusTet.E2 samples cluster together, indicating that overexpression of FGFR2 without FGF10 stimulation has no effect on the DEG list, and FGF10 treatment causes a stronger shift of the cluster in FGFR2 overexpressing cells (plus Tet) than in un-induced cells (minus Tet) (**Supplementary Fig. S3c**). Overall the PCA analysis therefore demonstrates that experimental variation between repeats is low and validates the specificity of the FGFR2 expression response.

Master regulators of FGFR2 signalling

The correlation of the regulon ranks in each network was calculated to show the agreement between the different underlying regulatory networks. The enrichment statistic for each regulon was used to rank MRs identified in each network and the correlation statistic (R) between these lists was compared between networks (**Supplementary Figs. S4-S6**). For each gene expression signature, there was very good correlation in the enrichment rank between breast cancer cohort I and II (R=0.85-0.9). A somewhat lower level of correlation was observed in the ranks determined for normal breast (R= 0.29-0.39), while virtually no overlap was seen with the

T-ALL network. An even higher agreement between cohorts was found when only considering the 50 top ranked regulons, as measured by the Jaccard Coefficient (JC). The rank of identified breast cancer MRs is therefore highly reproducible between the test and validation data sets and is related to that in the normal breast tissue data set, but is unrelated to the ranks obtained in a different cancer context (T-ALL).

The MRA identified 5 MRs when the overlap of the three FGFR2 gene signatures was considered. However, *Exp1* resulted in a smaller number of differentially expressed genes than seen in *Exp2* and *Exp3* and consequently fewer MRs are identified. When applying the less stringent criterion of enrichment in only two of the three experiments for the selection of MRs, an extended list of MRs was defined (**Fig. 5a** and **Supplementary Fig. S8**), which was further considered in the synergy and shadowing analysis (see below).

Once identified, we further tested the responsiveness of each of the defined regulons to FGFR2 signalling using gene set enrichment analysis (GSEA). In contrast to the MRA that considers only the top differentially expressed genes, the GSEA uses the complete rank information in order to test the association between a known set of genes and a given phenotypic difference [54]. Here regulons are treated as gene sets and the FGFR2 perturbation experiments as phenotypes, an extension of the GSEA as previously described [34]. **Supplementary Figure S9** shows that the SPDEF, ESR1, GATA3 and FOXA1 regulons are consistently FGFR2-responsive in *Exp1-3*. Enrichment of the PTTG1 regulon depended on the gene expression signature used.

Regulon validation

For ESR1 and SPDEF, ChIP-seq experiments were performed in MCF-7 cells, while existing data was analysed for FOXA1 [28] and GATA3 [29]. ChIP-seq experiments for SPDEF, FOXA1, GATA3 and ER transcription factors were performed on three biological replicates per each transcription factor. In addition, two input DNA samples were sequenced as a control of SPDEF and ER ChIP-seq, respectively, giving a total of 14 sequenced samples. Peak regions were identified in all ChIP-seq TF data sets using the peak caller algorithm MACS [56] with the default parameters, and using the correspondent control data set when appropriate. For each sample, 36bp single-end reads were obtained. Quality control of the raw sequenced data was conducted using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and good quality scores across all bases were observed. **Supplementary Table S2** summarizes the number of sequenced reads and peaks found per sample.

We obtained the overlapping peaks across all three replicates for a given transcription factor (**Supplementary Fig. S10**). Only binding events that occurred in at least two out of three biological replicates were considered for further downstream network analysis. To evaluate the consistency of signal between replicates we measured the read coverage over all genomic regions covering peak regions that were called in at least two of the three replicates. There is a good correlation between pairs of replicates, with Spearman's correlation coefficients ranging from 0.51 to 0.74 (**Supplementary Table S3**).

The ChIP-seq data was further used to examine the distribution of TF binding sites. **Sup-**

plementary Figure S11 shows that ESR1, FOXA1, GATA3 and SPDEF binding was significantly enriched in each of their own regulons when compared to random sites in the same regulons. Enrichment of both ESR1 and FOXA1 binding was detected across all four regulons, further supporting the idea that these transcription factors co-operate closely [28]. In addition, binding sites for all of four of these TFs are enriched in the SPDEF regulon, supporting the finding that SPDEF cooperates with the ER-network (**Supplementary Fig. S11a**). The overlap of binding sites was also analysed on a genome-wide basis and we find that 20% of SPDEF binding sites map near ESR1 binding sites (**Supplementary Table S5**). The overlap of binding by SPDEF and the other MRs is lower. These data support our conclusion that a subset of ERα regulated genes is co-regulated by SPDEF and may also point to the possibility that even if target genes are co-regulated by two or more TFs, the relevant TFs may bind to distinct regulatory regions, for example enhancer versus promoter elements.

The co-regulatory nature of the MRs was further ascertained in siRNA experiments. Hypergeometric tests were used to identify the regulons most responsive to siRNA transfections for ESR1, PTTG and SPDEF (**Supplementary Table S4**) and found that all MRs tested were significantly enriched. The effect of the siRNAs phenotypes were further assessed on gene expression of the regulons of our five MRs using GSEA, confirming that all five MR-regulons were significantly perturbed in each experiment, compared to a random gene list (**Supplementary Fig. S12**). GSEA is far more sensitive in detecting change than a hypergeometric test since the whole probe universe contributes to the analysis. Of particular interest with respect to the ER-network was the finding that siSPDEF was sufficient to modulate expression of the ESR1 regulons ($P<0.001$, weighted Kolmogorov-Smirnov test), confirming its place in a network of regulators rather than just as a downstream effector.

Overlap, synergy and shadowing the extended MR list

To better understand the relationship between the identified regulons, we used the extended list of MRs enriched in at least two of our three FGFR gene signatures and examined their overlap, 'shadow' and 'synergy' as previously described [34]. Regulon shadowing has been described as a potential confounding factor when assessing master regulators [34]. If two enriched regulons overlap significantly, one of them may appear enriched because of the common enriched targets.

The overlap simply compares the number of genes present in the intersect of two regulons compared to the total number (union) of regulated genes. **Supplementary Figure S14** confirms the ER-related cluster seen for the five MRs. However in this extended list a second cluster was apparent that included E2F3, FOXM1, E2F2 and PTTG1. E2F3, E2F2 and FOXM1 are well-studied cell cycle regulators [33,34] while PTTG1, also known as securin, functions to maintain chromosome stability, regulate cell cycle progression and appropriate cell division [59]. Interestingly, these four TFs and a total of nine out of the additional 15 MRs of the extended list are also significantly enriched when applying the proliferation-related meta-PCNA signature [26] to cohort I using MRA (**Supplementary Table S1**), further supporting the idea that the regulons of these factors are enriched due to the highly proliferative state of cancer cells. The

clustering of E2F2 and FOXM1 with PTTG1 is also apparent in the network visualisation (**Fig. 7c**) and is fully consistent with their overlapping function in control of the cell cycle and proliferation. In addition, the extended list of MRs contains a number of regulators such as ETV4, ELF3 and SMARCA4 that may be of interest in terms of mediating the FGFR2 response more directly and further study of these would be warranted.

The synergy analysis [34] examines if the enrichment of the applied gene expression signature is greater in the intersect of two regulons than the enrichment found in the union of two regulons. This analysis investigates whether two MRs act independently or on a set of overlapping genes. Within the highly connected clusters we find that the intersect contains the most strongly FGFR2-responsive genes. However, when a large number of genes are found in the intersect the difference between the intersect and the union is negligible and no synergy is detected. In contrast, when comparing MRs from cluster 1 (ESR1, FOXA1, GATA3 and SPDEF) with cluster 2 (PTTG, FOXM1, E2F3 and E2F2) there is significantly greater FGFR2- responsiveness in the intersect of regulons, than in the union (**Supplementary Fig. S14**). As demonstrated above, the two different clusters of regulons do not have a very large overlap and their MRs map to different parts of the TF network. The synergy analysis therefore suggests that MRs from two different clusters can co-operate to regulate a small, but common set of FGFR2-responsive genes.

We next carried out a 'shadow' analysis, which is performed for significantly overlapping regulons and contrasts enrichment between the unique part of a regulon in a pair with the enrichment in each regulon individually. Therefore a comparison of A versus B (A-unique versus A-total) will yield different results from a comparison of B versus A (B-unique versus B-total) and the results are not symmetrical. To define A as a shadow of B, for example, B should not be a shadow of A. **Supplementary Figure S14** indicates that there is little significant shadowing in either cluster. In other words, there is less FGFR2-reponsiveness in the unique part of the regulon than in the whole regulon, confirming that the TFs within each cluster act cooperatively to drive the FGFR2 response.

Use of size-balanced gene lists in the EVSE

Exp1-3 were designed as three distinct models of activating FGFR2 signalling. Their differences reflect both the biology underlying each experimental system and the strength of the FGFR2 signal transmitted in each case. Consequently the identified DEG lists vary significantly in length (**Fig. 1**). Importantly, the length of these gene lists can influence the outcome of downstream statistical analysis, such as EVSE. Our comparison of the three experimental systems revealed that, despite the difference in the number of deregulated genes, the rank order of FGFR2-responsive genes remains very similar between experiments, as revealed by MRA analysis (**Fig. 3**). We used this feature in order to generated similar sized gene lists. In *Exp1* we ranked all genes on their response (log fold change) to FGFR2 singalling and extended the DEG list to the average length of *Exp2* and *Exp3*, but only included those genes that were FGFR2-responsive in at least one other experiment. This places a constraint on choosing 'ran-

dom' non-significant genes and for this reason the extended gene list for *Exp1* is slightly shorter than that for *Exp3*. The size-balanced gene lists were tested in the EVSE and each experiment yielded significant enrichment scores when tested in the cis-eQTL analysis using breast cancer AVS (**Supplementary Fig. S15a** and **Fig. 2b**). As a control we downsized the gene lists obtained for *Exp2* and *Exp3* to that of *Exp1* using the same rank information approach. The ranking was performed separately for each time point analysed and non-overlapping hits were combined into the final gene list, resulting in similar sized lists for each of the three experiments (**Supplementary Fig. S15b**). Again we tested the derived gene lists in the EVSE analysis, but the short gene lists did not generate statistically significant associations. This suggests that genes in the rank below approximately 950 significantly contribute to the signal and that short genes lists lack the statistical power to reveal an association between FGFR2-responsive genes and breast cancer AVS.

Source code

The source code developed in this study is publicly available from the Bioconductor (http://www.bioconductor.org/), separated in three packages for ease of use and distribution:

- R data package *Fletcher2013a*
  http://bioconductor.org/packages/devel/data/experiment/html/Fletcher2013a.html
- R data package *Fletcher2013b*
  http://bioconductor.org/packages/devel/data/experiment/html/Fletcher2013b.html
- R software package *RTN*
  http://bioconductor.org/packages/devel/bioc/html/RTN.html

The R data package *Fletcher2013a* contains the time-course gene expression data from MCF-7 cells treated under different experimental systems in order to perturb FGFR2 signalling, while the R data package *Fletcher2013b* contains a set of pre-computed transcriptional networks and related datasets used in the network analysis. The source code for transcriptional network reconstruction and master regulator analysis is distributed in the R software package *RTN*.

Supplementary References

59 Wang Z, Yu R, Melmed S. Mice lacking pituitary tumor transforming gene show testicular and splenic hypoplasia, thymic hyperplasia, thrombocytopenia, aberrant cell cycle progression, and premature centromere division. *Mol Endocrinol*, 15(11):1870-9, 2001

60 Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, Green MR. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11;11:237, 2010.