Supplementary material

A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions

Jonas Paulsen [1,*], Einar A Rødland [2] Lars Holden[3], Marit Holden[3], Eivind Hovig[1,2]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital. PO Box 4950, Nydalen. 0424 Oslo, Norway [2]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway [3]Statistics for Innovation, Norwegian Computing Center, 0314 Oslo, Norway

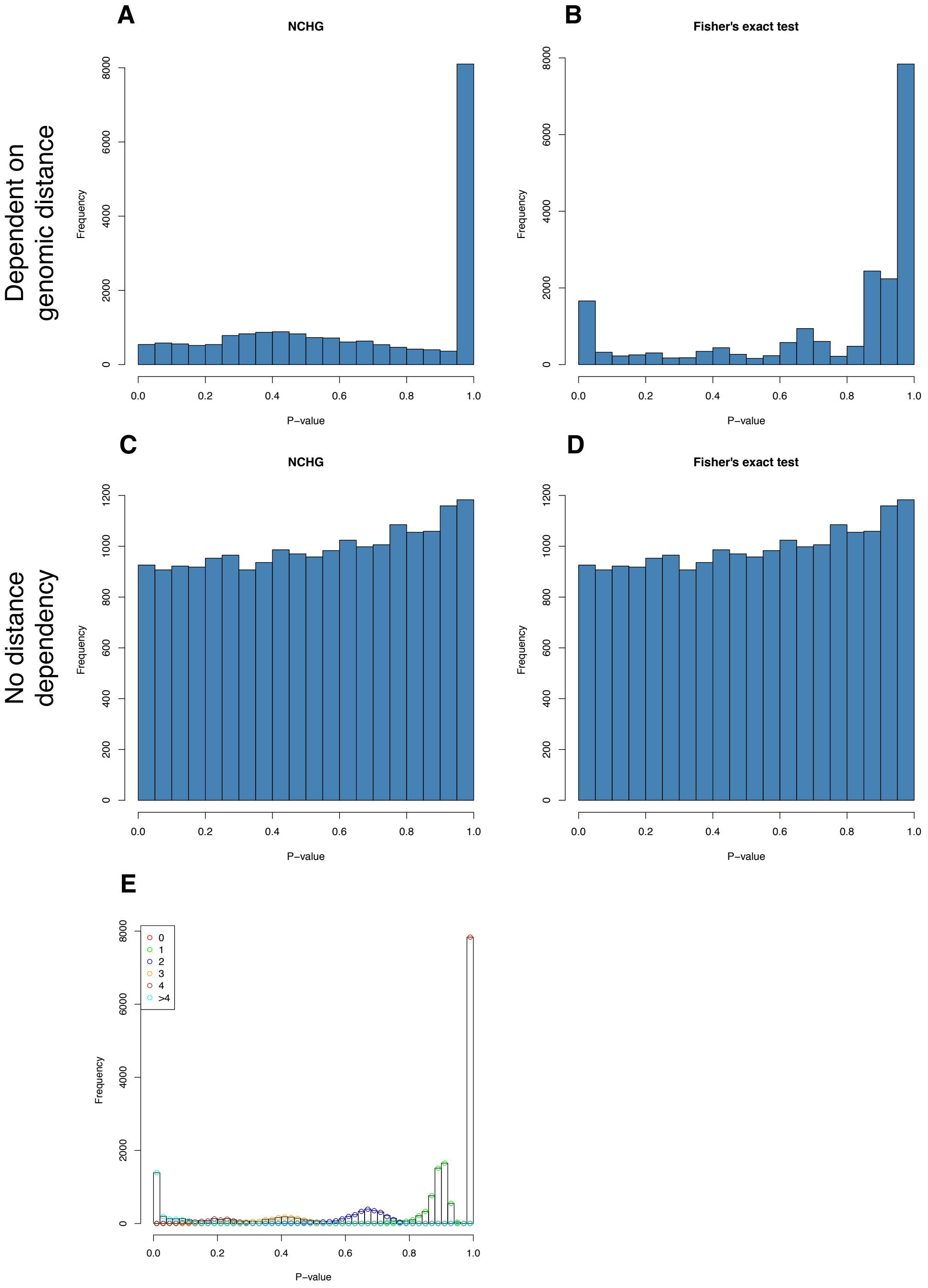

Figure S1: Clustering of P-values caused by discrete numbers for n_{ij}. (A) Raw P-values obtained using the NCHG test on data sampled with strong dependency on genomic distance. (B) Raw P-values obtained using Fisher's exact test with strong dependency on genomic distance. (C) Raw P-values obtained using the NCHG test on data sampled without any dependency on genomic distance. (D) Raw P-values obtained using Fisher's exact test on data sampled without any dependency on genomic distance. (E) Example of P-value clustering due to discrete numbers for n_{ij}. The histogram is based on the same data as shown in B, but with higher number of bins and color-coding according to n_{ij}, to visualize the cause of the clustering of the P-values.

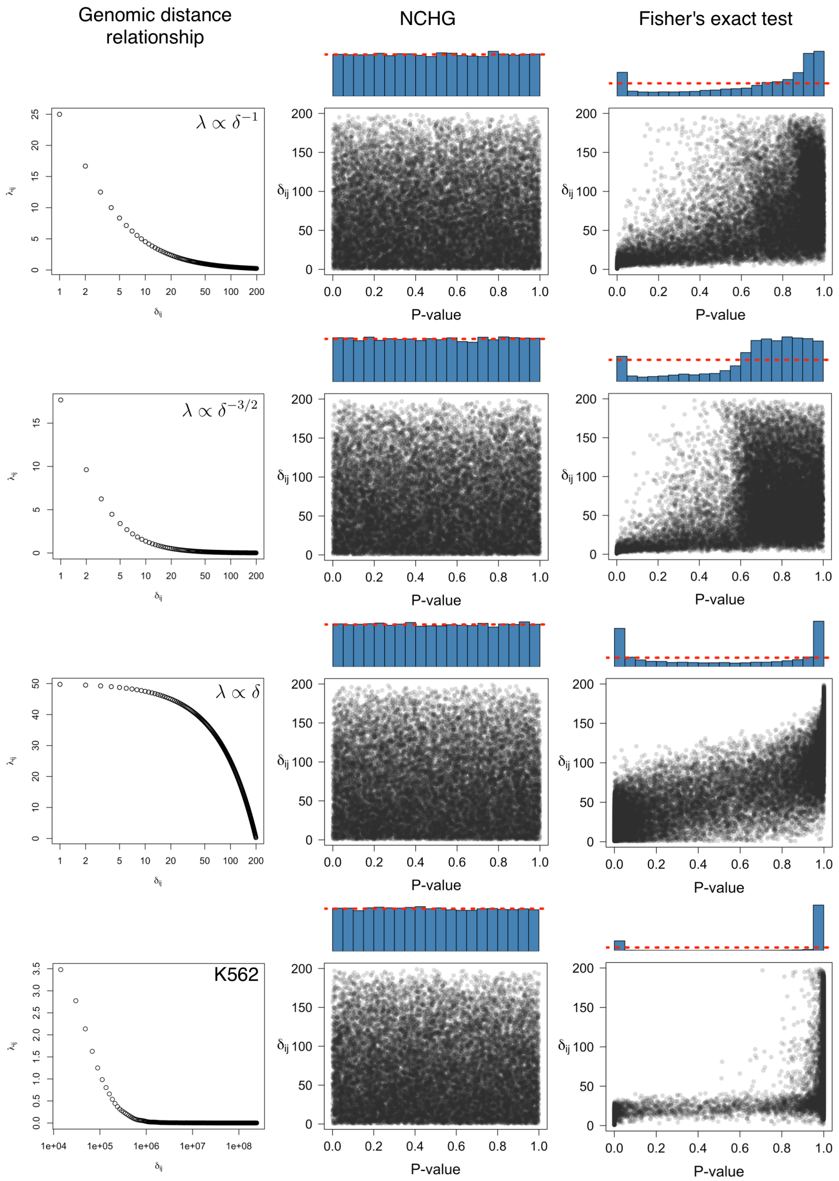

Figure S2: Comparison of Fisher's exact test and the NCHG test for various choices of the relationship between genomic distance (δ_{ij}) and expectancy (λ_{ij}). P-values are plotted against the genomic distance, and blue histograms indicate the distribution of P-values, calculated as explained in the Methods section. Red dashed lines indicate the expected fraction for a uniform distribution of P-values. Left panel: relationship between genomic distance (δ_{ij}) and expectancy (λ_{ij}) used for sampling of data. Middle panel: P-value distribution using the NCHG method. Right panel: P-value distribution using Fisher's exact test. All datasets, except the one based on K562, were produced using randomly sampled interaction frequencies between 200 "anchors", according to a Poisson model with expectation (λ_{ij}) dependent on genomic distance as given in the upper right corner of each plot in the left panel. For sampled data based on K562 (lower row), we sampled interaction frequencies using a Poisson model dependent on the actual genomic distance relationship (left) for this cell line. Genomic positions were chosen based on the first 200 anchors on chromosome 22, and data were sampled for all possible pairs of anchors.

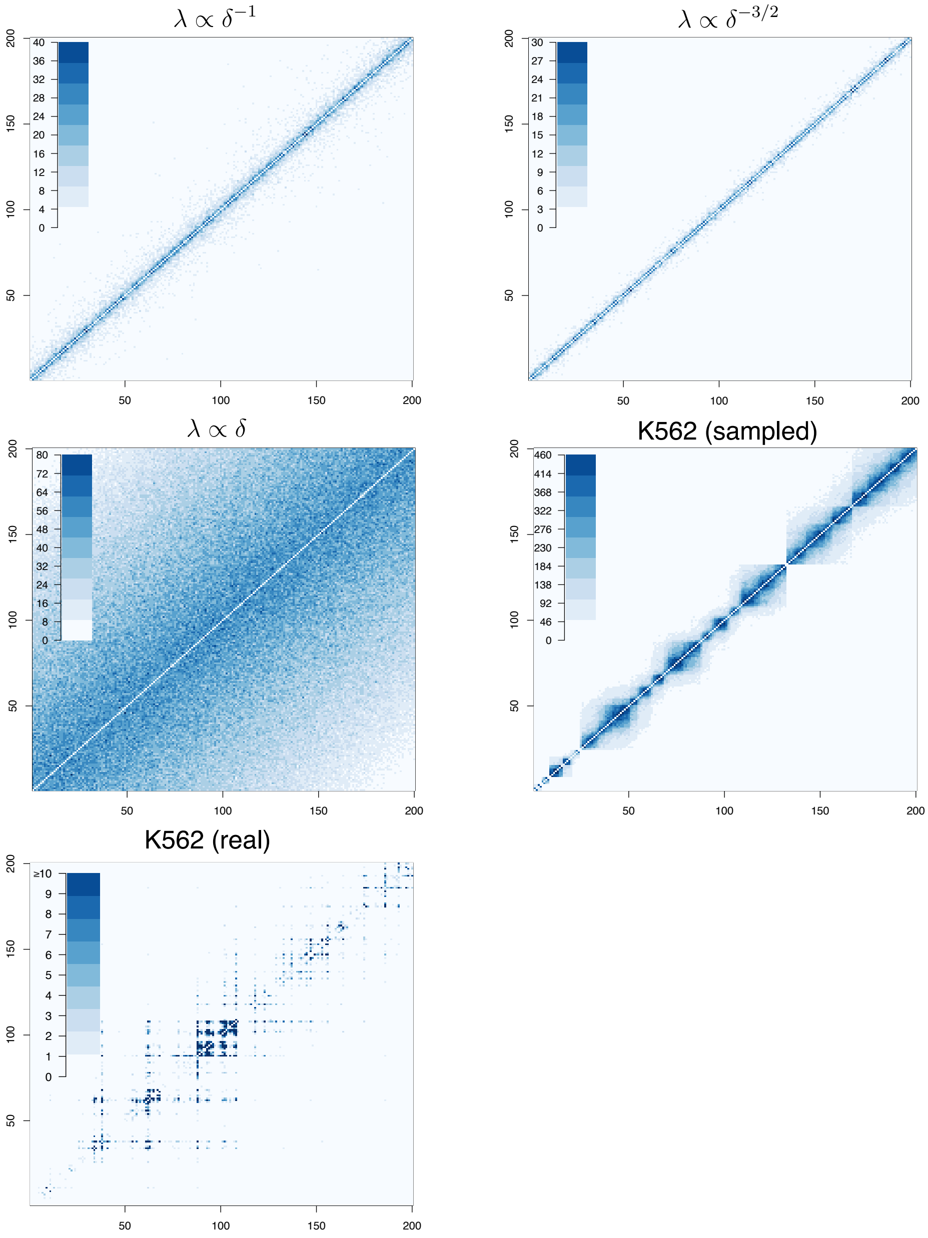

Figure S3: Heat maps showing the data sampled based on various choices of the relationship between genomic distance (δ_{ij}) and expectancy (λ_{ij}), as explained for Figure S1. Sampled interaction frequencies are visualized for all-versus-all of the 200 anchors. For data sampled based on the genomic distance relationship in the K562 cell-line, only the 200 first anchors on chromosome 22 are shown. The bottom heat map shows observed interaction frequencies between the first 200 anchors on chromosome 22, from the K562 cell line.

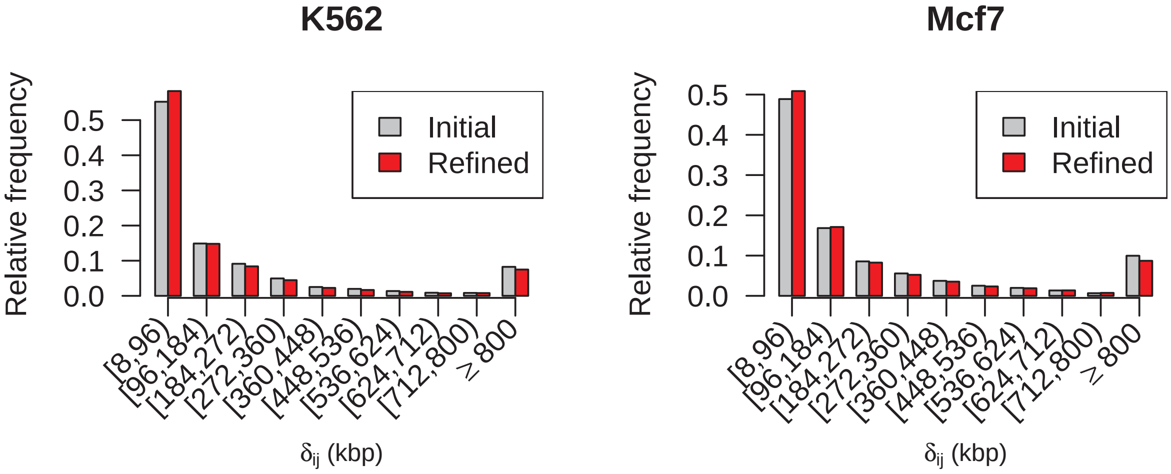

Figure S4: Effect of refinement on the significant interactions. Bar-plots showing the relative frequencies of genomic distances (δ_{ij}) for the significant interactions using the NCHG test, before (gray) and after (red) refinement. Genomic distances are divided into 10 groups as indicated. Results for K562 (left) and Mcf7 (right) are shown.

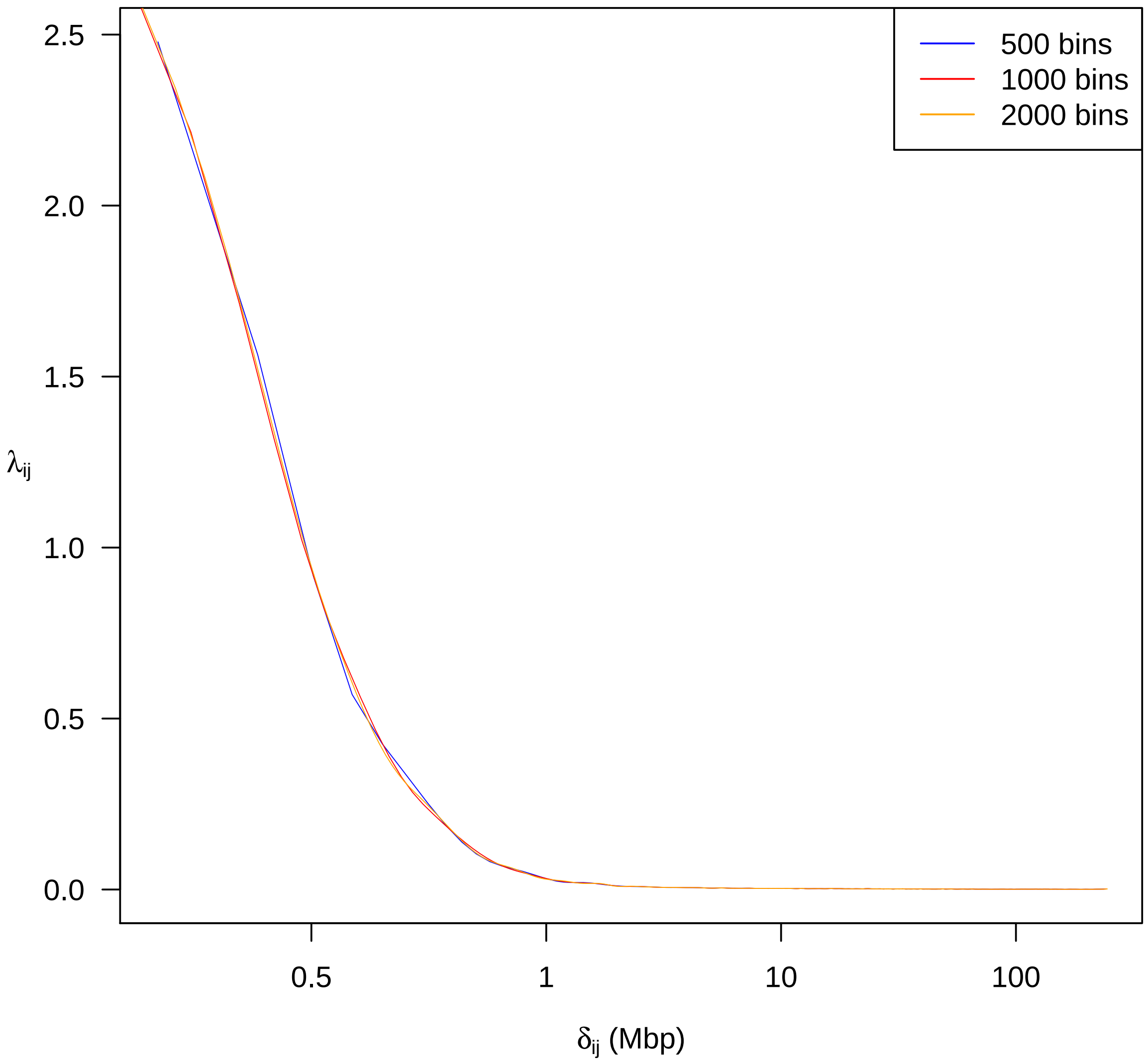

Figure S5: Dependency of the estimate of λ_{ij} for different choices of number of quantiles used. Genomic distance (δ_{ij}), defined as the distance between pairs of anchors, plotted against the smoothed average number of interactions (λ_{ij}), for 500 quantiles (blue), 1000 quantiles (red) and 2000 quantiles (orange). Results are based on the K562 cell line.

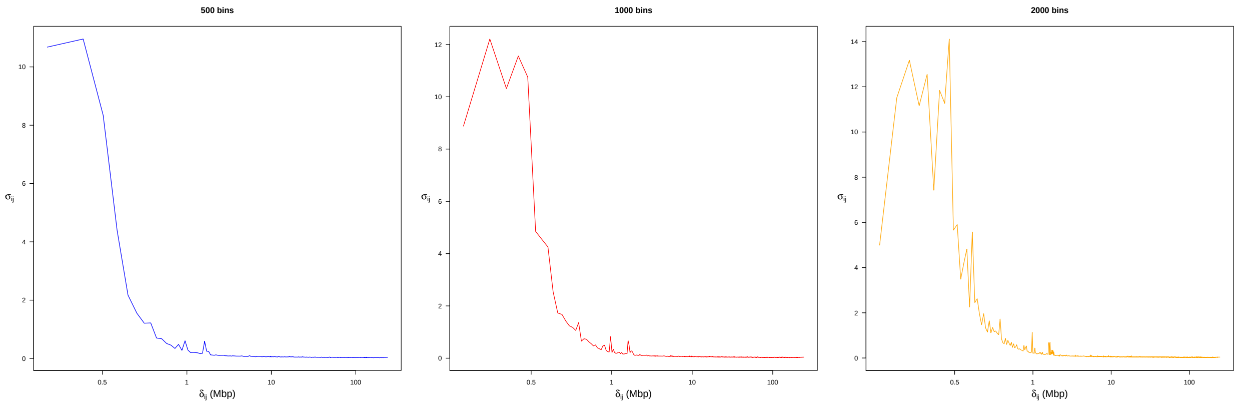

Figure S6: Dependency of the standard deviation (σ_{ij}) for different choices of number of quantiles used. Genomic distance (δ_{ij}), defined as the distance between pairs of anchors, plotted against the standard deviation for each of the bins (σ_{ij}), for 500 quantiles (left), 1000 quantiles (middle) and 2000 quantiles (right). Results are based on the K562 cell line.

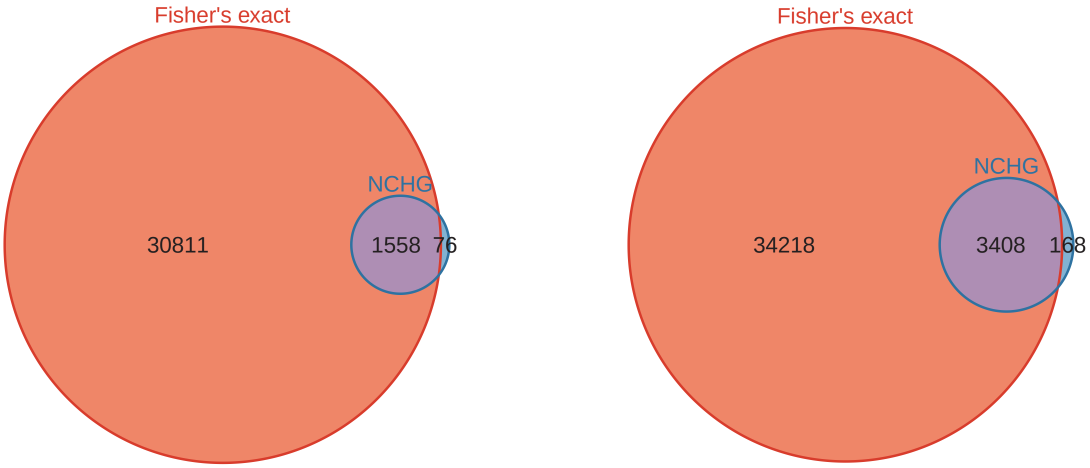

Figure S7: Venn-diagram comparisons of significant interactions from the NCHG test and Fisher's exact test, when no cutoff on n_{ij} is used. Blue circles indicate the significant interactions using the NCHG test, while red circles indicate significant interactions using Fisher's exact test. Numbers indicate the number of significant interactions within each subset. Results for K562 (left) and Mcf7 (right) are shown.

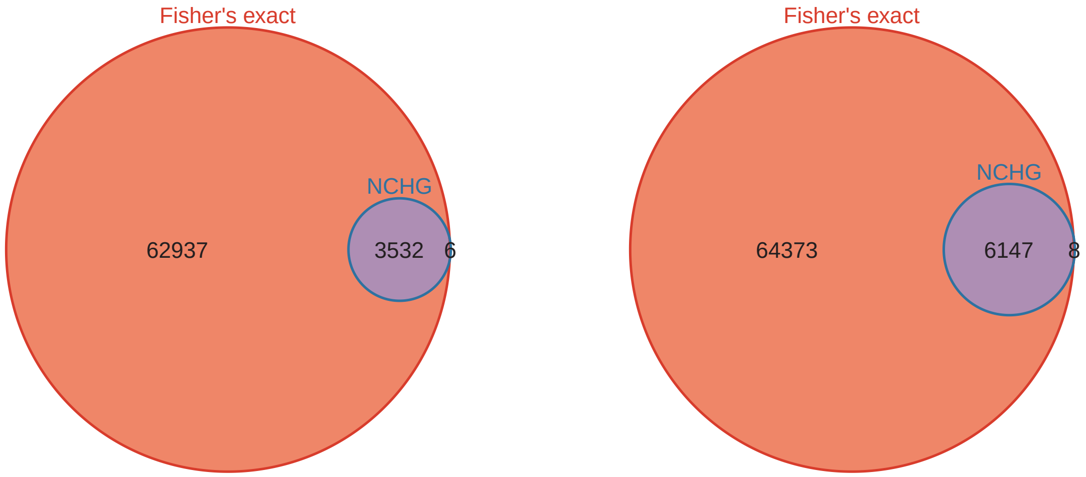

Figure S8: Venn-diagram comparisons of significant interactions from the NCHG test and Fisher's exact test, when no cutoff on n_{ij} is used, but where only interactions with genomic distances $\leq$1Mb are considered. Blue circles indicate the significant interactions using the NCHG test, while red circles indicate significant interactions using Fisher's exact test. Numbers indicate the number of significant interactions within each subset. Results for K562 (left) and Mcf7 (right) are shown.

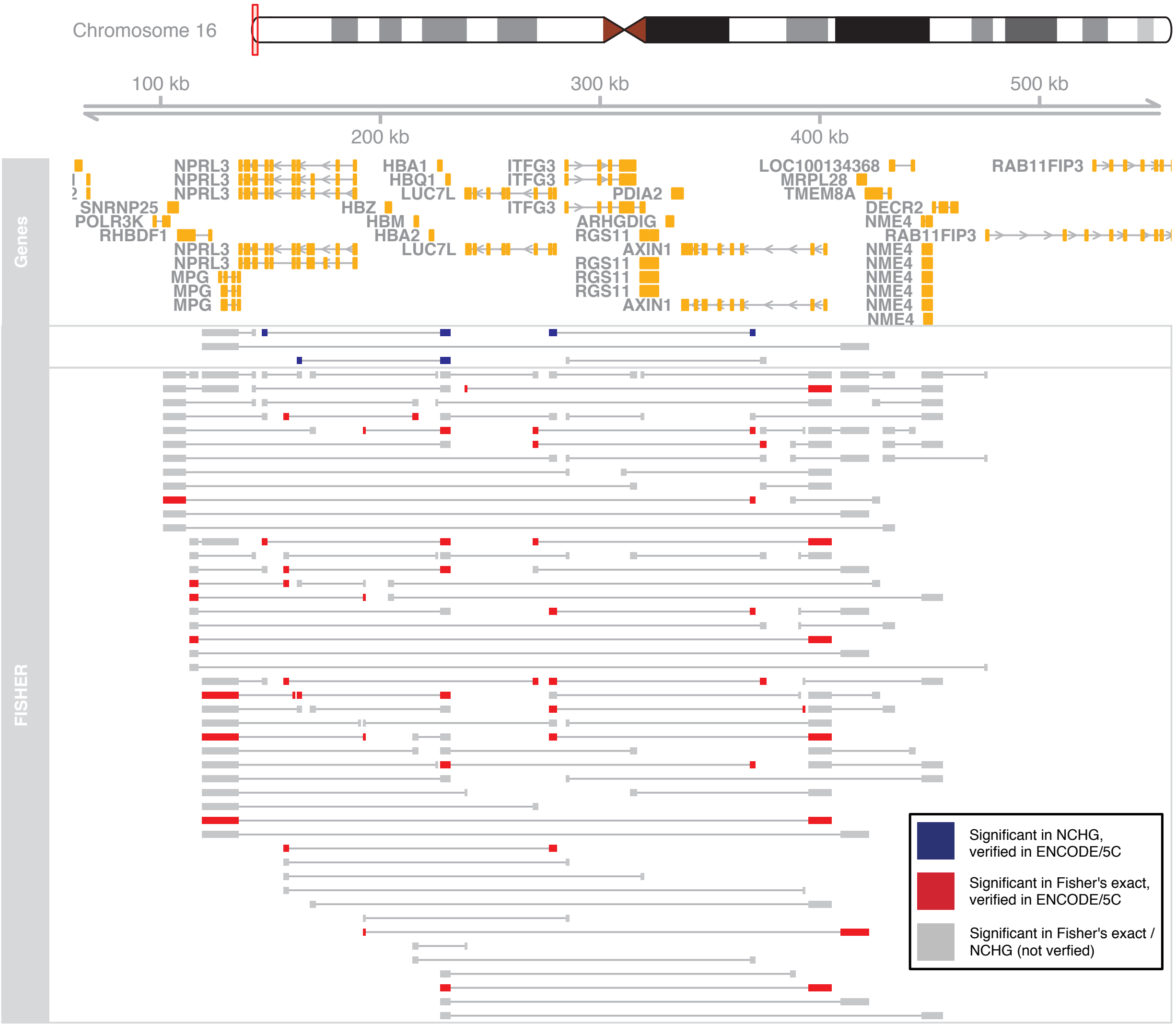

Figure S9: Significant interactions involving the α-globin cluster for the K562 cell-line, from the NCHG test and Fisher's exact test, when no cutoff on n_{ij} is used, but where only interactions with genomic distances $\leq$1Mb are considered. Chromosomal position is indicated on the top, with annotated genes shown in orange. Significant interactions from the NCHG test and Fisher's exact test are shown as connected segments. Significant interactions colored blue and red, for the NCHG test and Fisher's exact test, respectively, are verified using 5C data from Sanyal et al. [1]. Gray segments indicate interactions that were not found in the 5C data set.

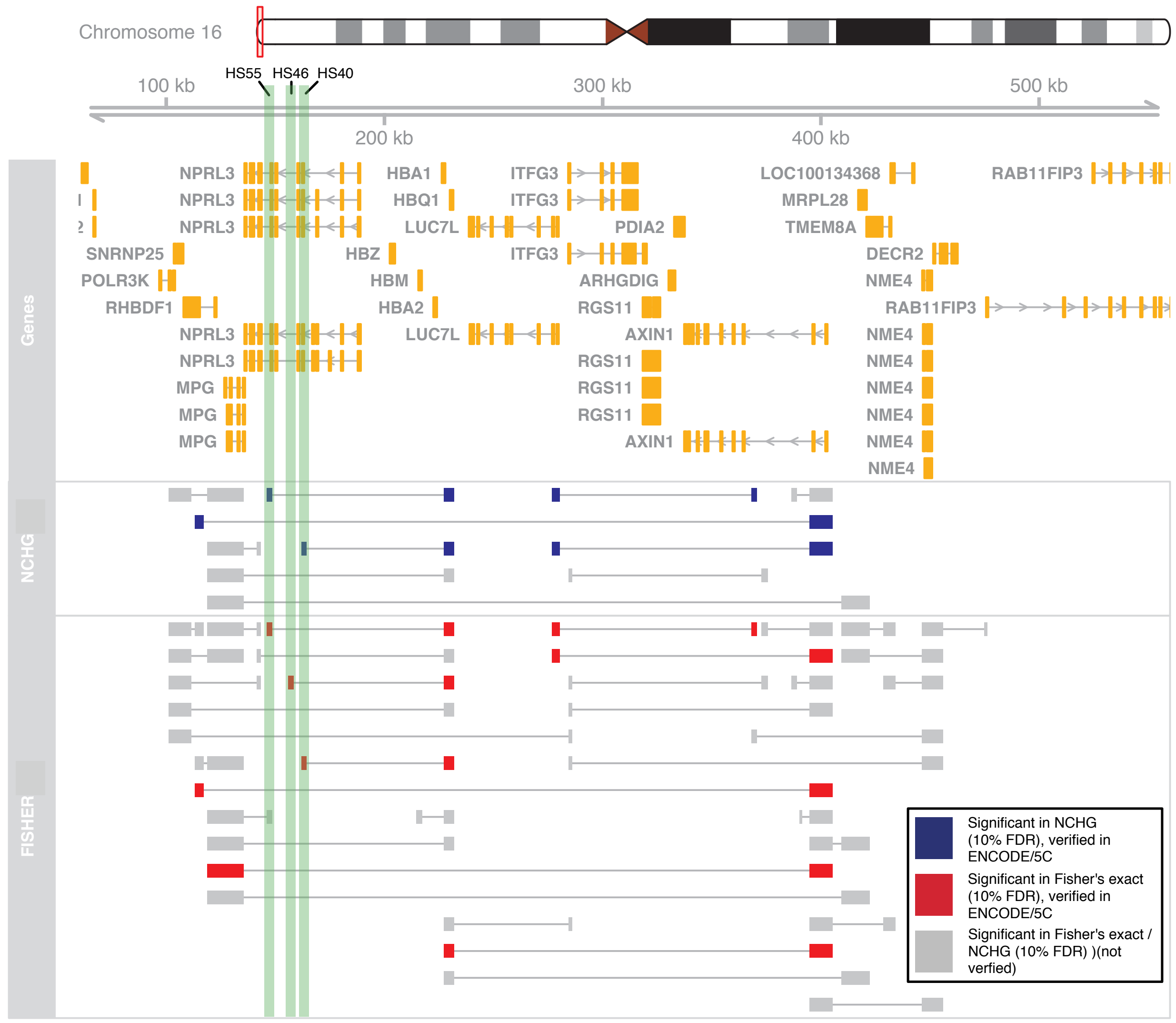

Figure S10: Significant interactions selected at 10% FDR, involving the α-globin cluster for the K562 cell-line, from the NCHG test and Fisher's exact test. Chromosomal position is indicated on the top, with annotated genes shown in orange. Significant interactions from the NCHG test and Fisher's exact test (at 10% FDR) are shown as connected segments. Significant interactions colored blue and red, for the NCHG test and Fisher's exact test, respectively, are verified using 5C data from Sanyal et al. [1]. Gray segments indicate interactions that were not found in the 5C data set.

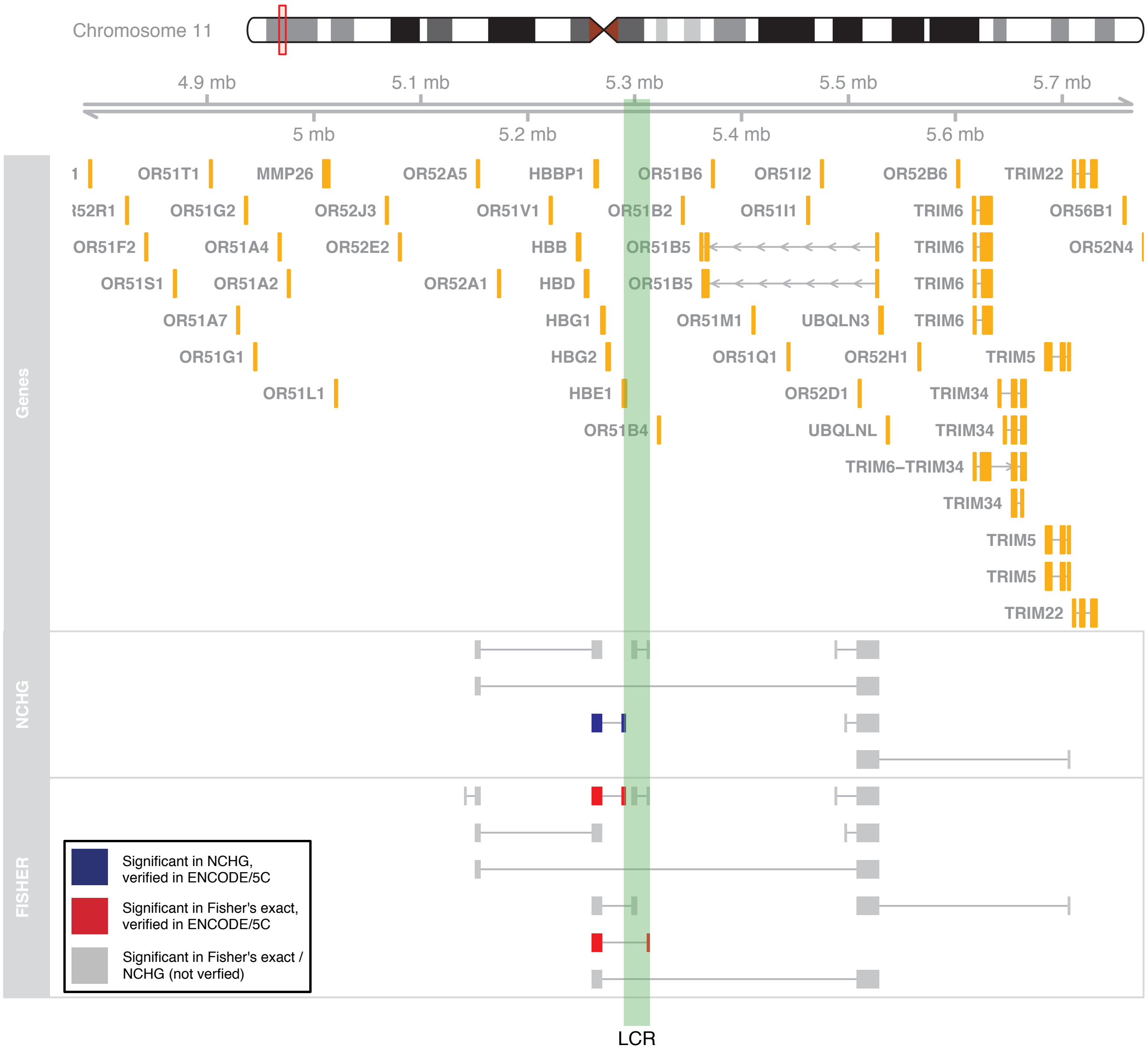

Figure S11: Significant interactions involving the β-globin cluster (ENCODE target region ENm009) for the K562 cell-line. Significant interactions from the NCHG test and Fisher's exact test are shown. Significant interactions colored red and blue, for the NCHG test and Fisher's exact test respectively, are verified using 5C in Sanyal et al. [1]. The position of the locus control region (LCR) is indicated using green shading.

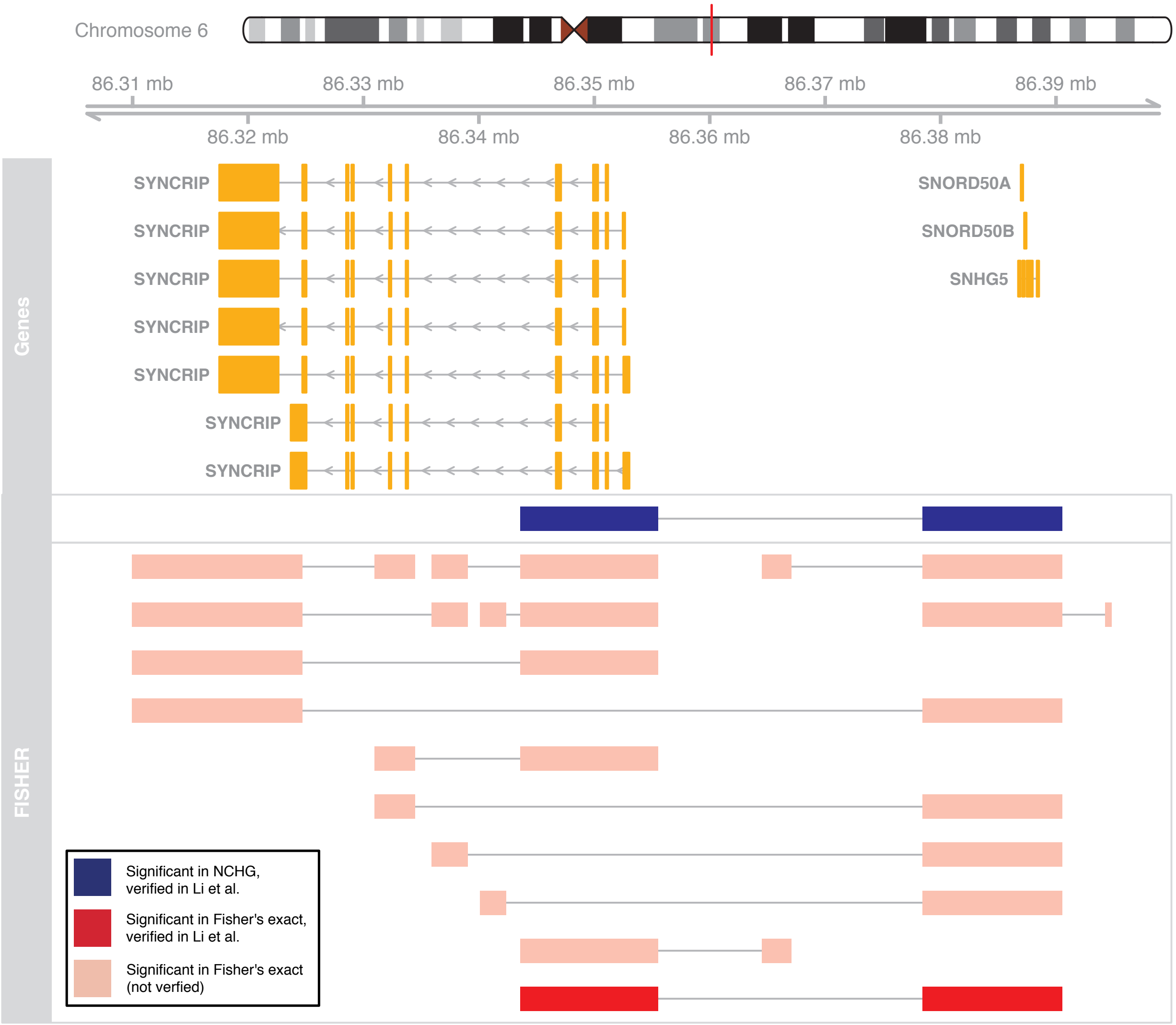

Figure S12: Significant interactions involving the SYNCRIP gene for the K562 cell-line. Significant interactions from the NCHG test and Fisher's exact test are shown. Significant interactions colored dark red and dark blue, are verified (using 3C) in Li et al. [2]. Pink interactions indicate interactions found using Fisher's exact test that have not been verified.

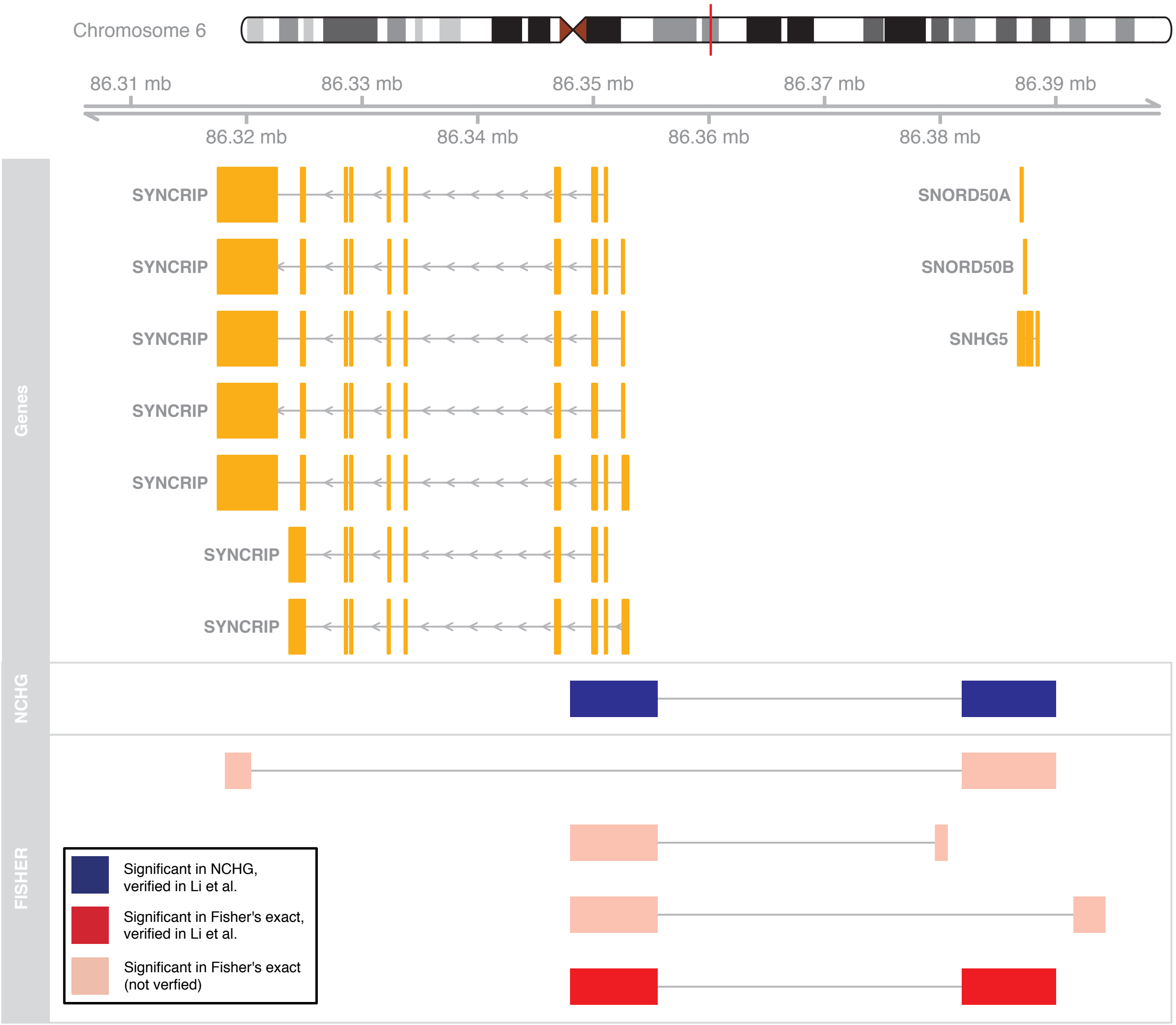

Figure S13: Significant interactions involving the SYNCRIP gene for the Mcf7 cell-line. Significant interactions from the NCHG test and Fisher's exact test are shown. Significant interactions colored dark red and dark blue, are verified (using 3C) in Li et al. [2]. Pink interactions indicate interactions found using Fisher's exact test that have not been verified.

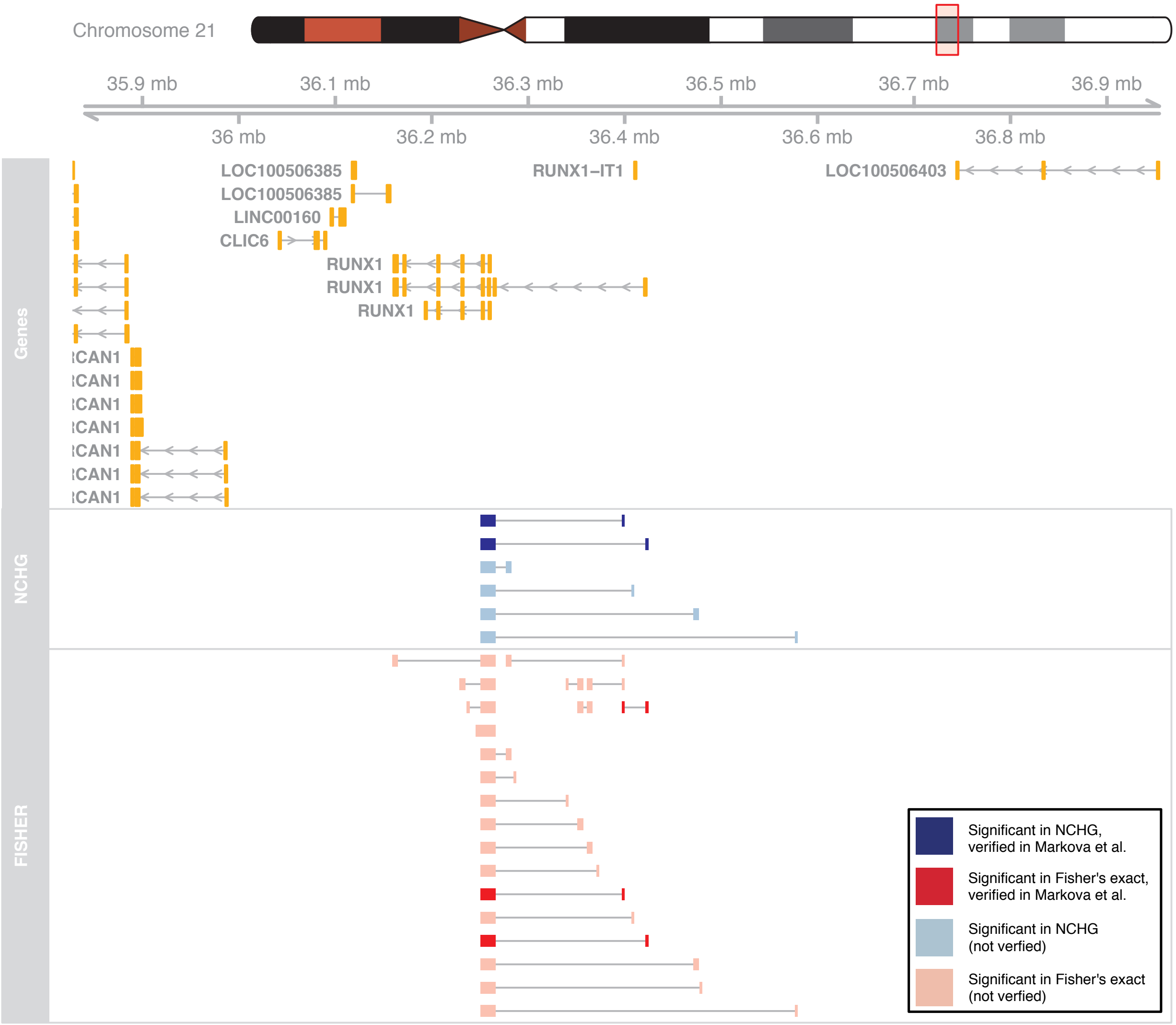

Figure S14: Significant interactions involving the RUNX1 gene for the K562 cell-line. Significant interactions from the NCHG test and Fisher's exact test are shown. Significant interactions colored dark red and dark blue, are verified (using 3C) in Markova et al. [3]. Pink and light blue interactions indicate interactions that have not been verified previously.

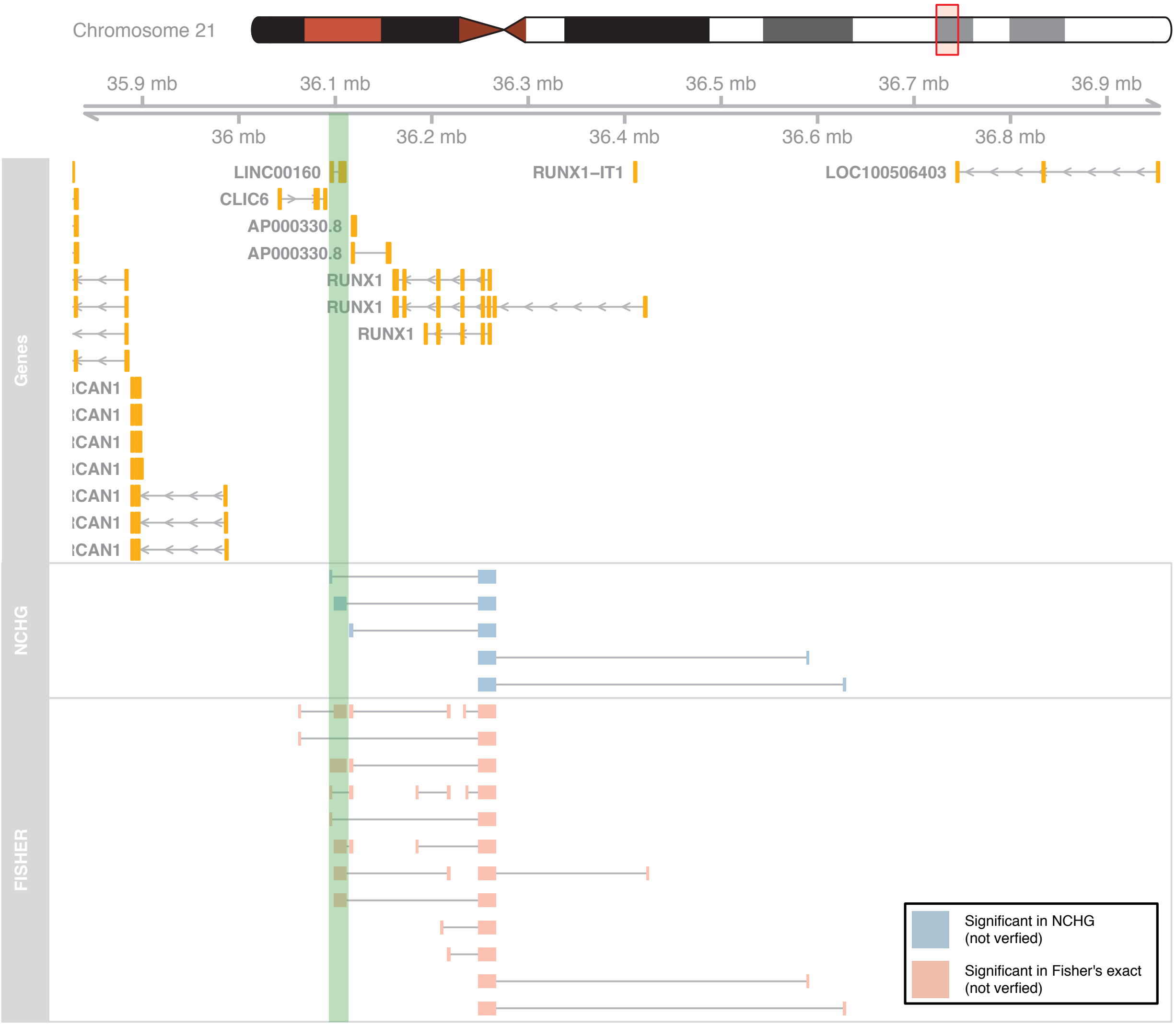

Figure S15: Significant interactions involving the RUNX1 gene for the Mcf7 cell-line. Significant interactions from the NCHG test and Fisher's exact test are shown. Significant interactions colored light blue correspond to interactions identified using the NCHG test, while interactions in pink correspond to interactions identified using the Fisher's exact test. The lncRNA (LINC00160) is highlighted using green shading. No 3C verifications are available for this cell-line.

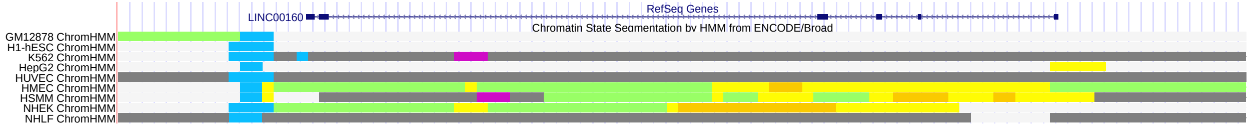

Figure S16: Cell-type specific enhancer activity of LINC00160. Chromatin states from Ernst et al. [4] are visualized for nine different cell-lines. Yellow and orange regions indicate weak and strong enhancers, respectively. Green color indicates weak transcriptions, gray color indicates Polycomb repressed states, purple color indicates inactive/poised promoter, blue color indicates insulator (CTCF).

Table S1: Significant interactions using no cutoff on n_{ij}. The table shows the median genomic distance and the actual number of significant interactions (N) for different number of observed interactions n_{ij}.

	K562				Mcf7			
	NCHG		Fisher's exact		NCHG		Fisher's exact	
n_{ij}	median δ_{ij}	N	median δ_{ij}	N	median δ_{ij}	N	median δ_{ij}	N
1	66687150.0	26	106627.0	1984	31894920.0	130	346293.0	1740
2	14571312.0	60	67625.5	7851	24281000.0	195	79462.0	8396
3	1314249.5	32	63896.75	5468	614163.5	120	72548.5	5987
4	1154050.0	43	64215.0	3678	365046.0	119	62993.75	4126
5	402776.0	53	62493.0	2495	236002.5	144	62287.5	2971
6	193900.0	51	56785.75	1826	168000.0	159	57691.5	2254
7	130718.0	51	57237.0	1383	140927.0	165	58286.5	1599
8	98394.75	48	52657.0	1027	128956.0	153	56884.75	1314
9	74439.25	50	49878.0	827	75838.0	135	52305.0	979
≥10	49403.50	1220	46618.75	5830	64838.0	2256	47202.5	8260

Table S2: Significant interactions using no cutoff on n_{ij}, but with a cutoff on the genomic distance at $\delta_{ij} \leq 1Mb$. The table shows the median genomic distance and the actual number of significant interactions (N) for different number of observed interactions n_{ij}.

	K562				Mcf7			
	NCHG		Fisher's exact		NCHG		Fisher's exact	
n_{ij}	median δ_{ij}	N	median δ_{ij}	N	median δ_{ij}	N	median δ_{ij}	N
1	724554.0	31	126716.0	29207	648839.0	127	145235.0	27171
2	279408.0	184	96182.25	12808	320772.0	415	102955.0	13547
3	135973.0	225	78147.5	6666	235325.5	464	88654.75	7344
4	121923.5	204	72782.5	4062	157509.0	437	71977.0	4627
5	102692.0	227	67701.0	2655	145370.0	417	66715.5	3155
6	84045.5	212	60529.0	1895	100065.0	399	61195.5	2359
7	62484.25	196	59560.75	1426	91292.25	366	61226.0	1661
8	71717.5	162	54113.25	1048	90510.5	301	59000.5	1349
9	52993.25	154	51331.75	842	68912.0	258	54064.0	1003
≥10	45416.0	1943	46872.75	5860	53007.0	2971	47511.25	8304

References

[1] Sanyal,A., Lajoie,B.R., Jain,G., and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature,* **489**, 109–113.

[2] Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J., et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell,* **148**, 84–98.

[3] Markova,E.N., Kantidze,O.L., and Razin,S.V. (2011) Transcriptional regulation and spatial organisation of the human AML1/RUNX1 gene. *J. Cell. Biochem.,* **112**, 1997–2005.

[4] Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M., et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* **473**, 43–49.