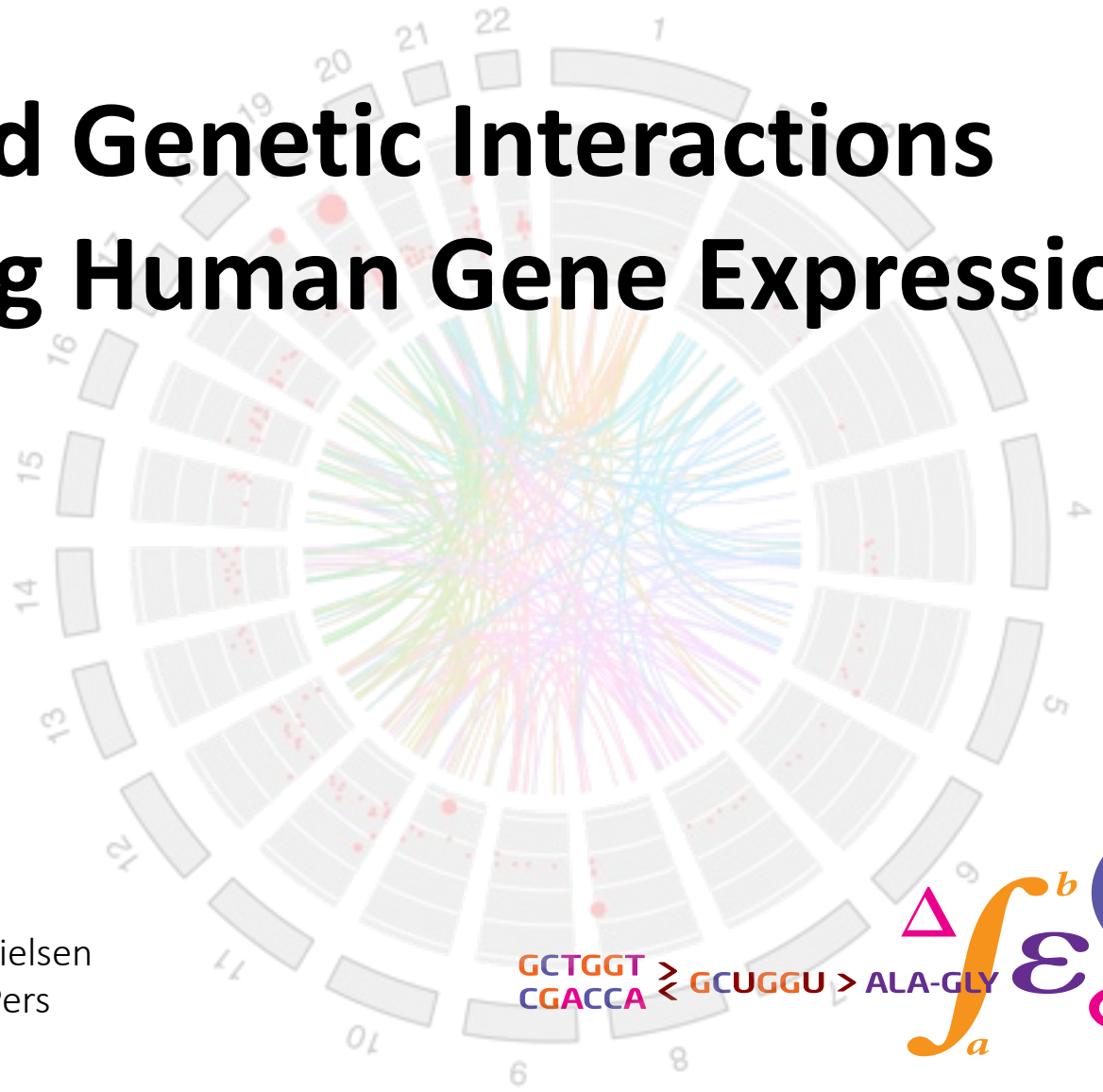
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

DTU

Spatial and Genetic Interactions Influencing Human Gene Expression

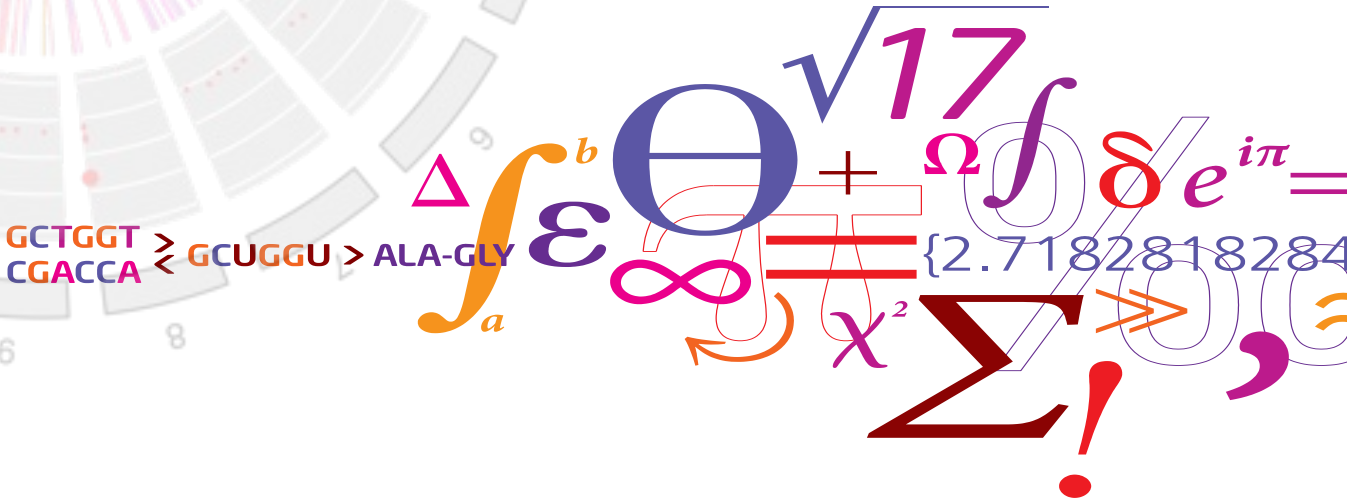
Pascal Timshel



26th November 2015

Master Thesis Defense
35 ECTS points

Supervisors: Henrik Bjørn Nielsen
Tune Hannes Pers



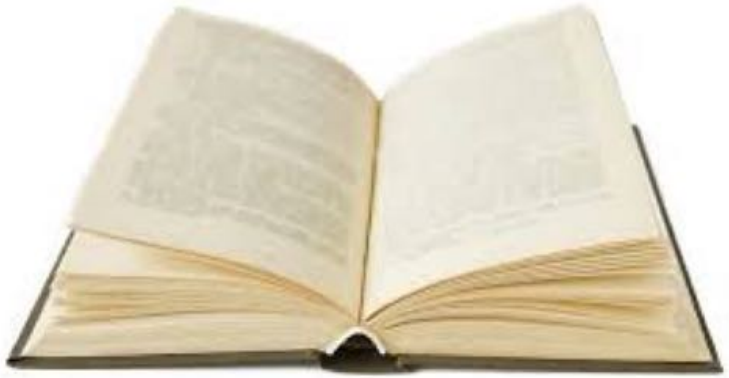
Center for Biological Sequence Analysis
Department of Systems Biology

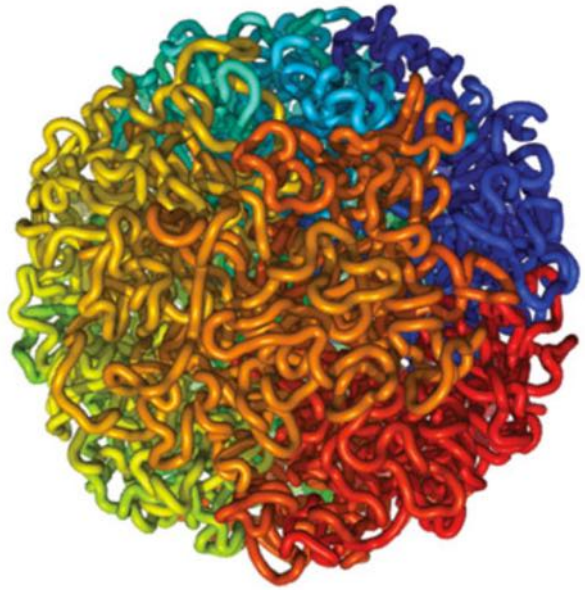
Outline

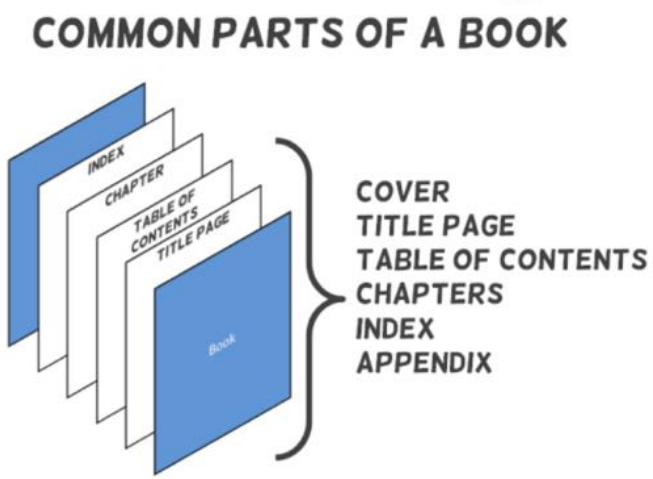
- Motivation
- Spatial Interactions (Genome Organization)
- Genetic Interactions (Epistasis)
- Spatial Epistasis Hypothesis
- Methods and Computational Framework
- Epistatic Signals
- “Meta-analysis”

The Genome has a 3rd Dimension

- The genome is traditionally “read” 1D linear trac
- The genomic information might be interpreted differently in a 3D context



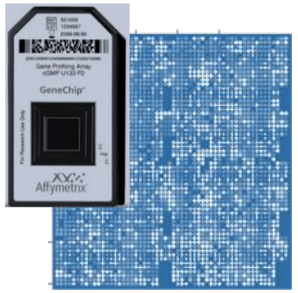


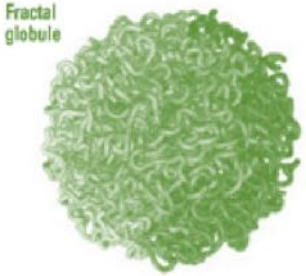


Motivation and overview

Connecting gene expression and genetic interactions

and chromosomal organization





How do spatial and genetic interactions influence gene expression?

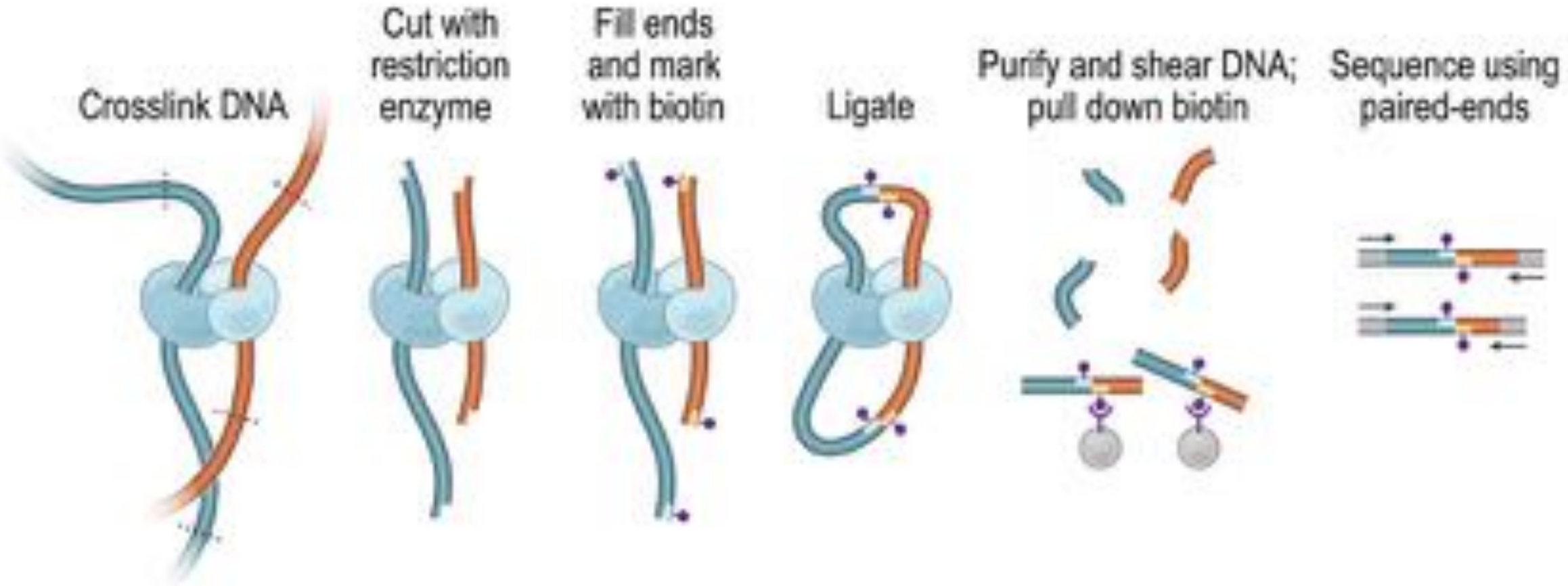
Research questions

1. What are the regulatory **molecular mechanisms** that **cause epistasis**?
2. Can regulatory **molecular mechanisms facilitate identification** of **epistatic** effects?
3. To which **extent** does **epistasis influences human gene expression**?

Outline

- Motivation
- **Spatial Interactions (Genome Organization)**
- Genetic Interactions (Epistasis)
- Spatial Epistasis Hypothesis
- Methods and Computational Framework
- Epistatic Signals
- “Meta-analysis”

Hi-C protocol (Rao et al., 2014, *Cell*)



Analogous hierarchical organization of protein and genome structure

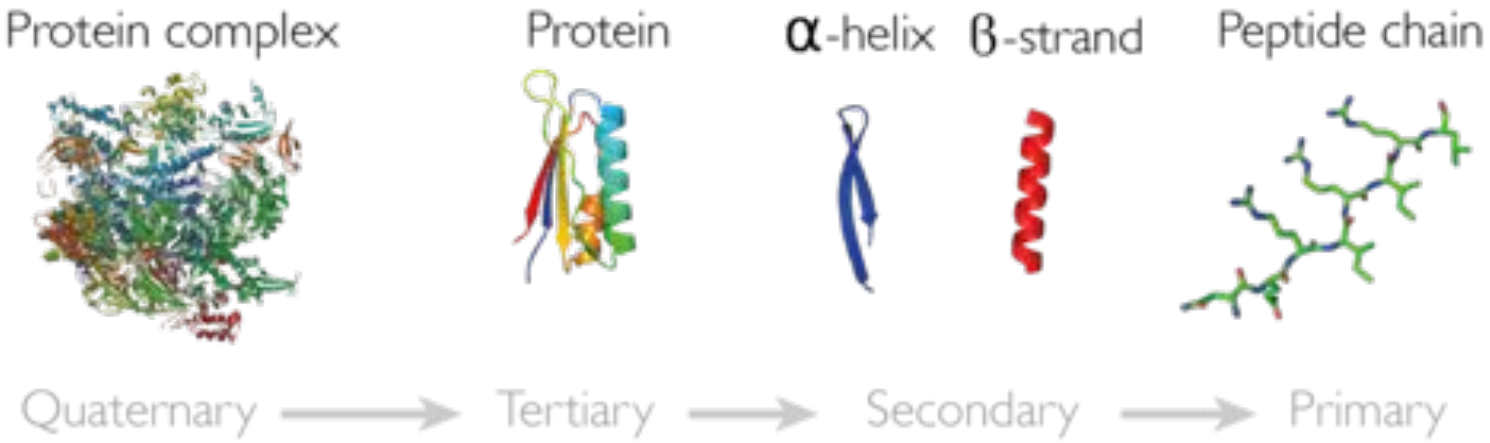
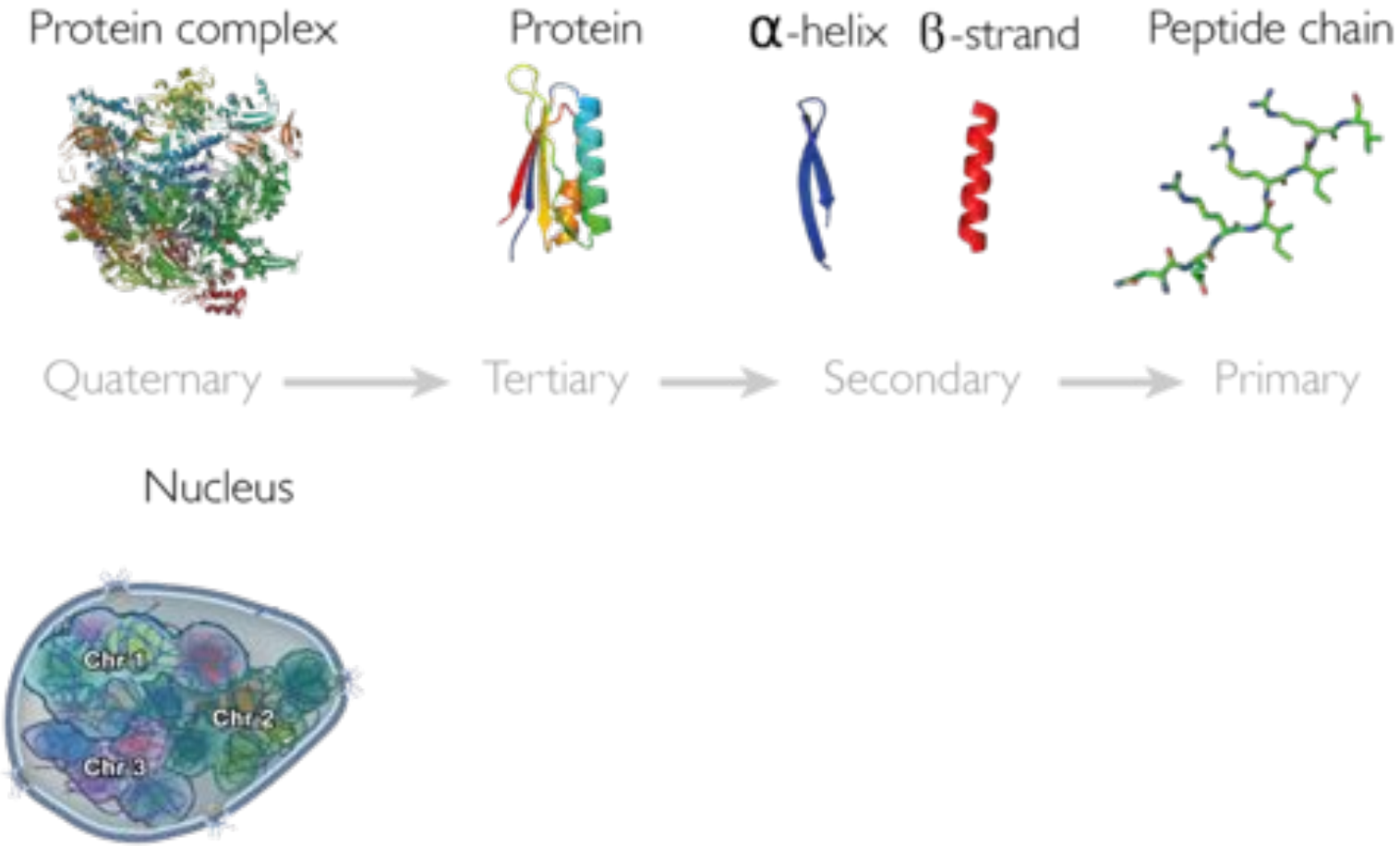
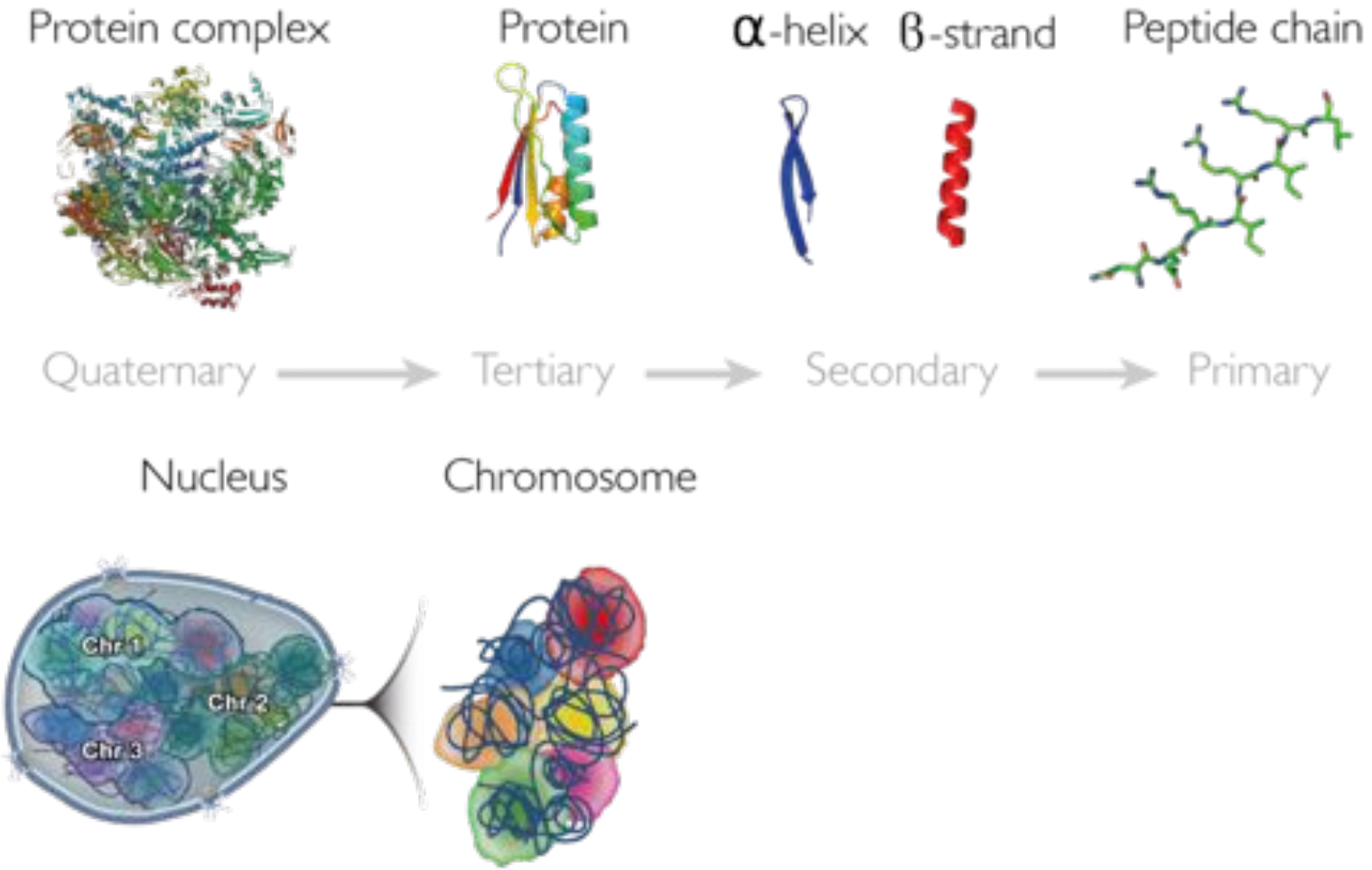


Figure inspired by Sexton, 2015, *Cell*

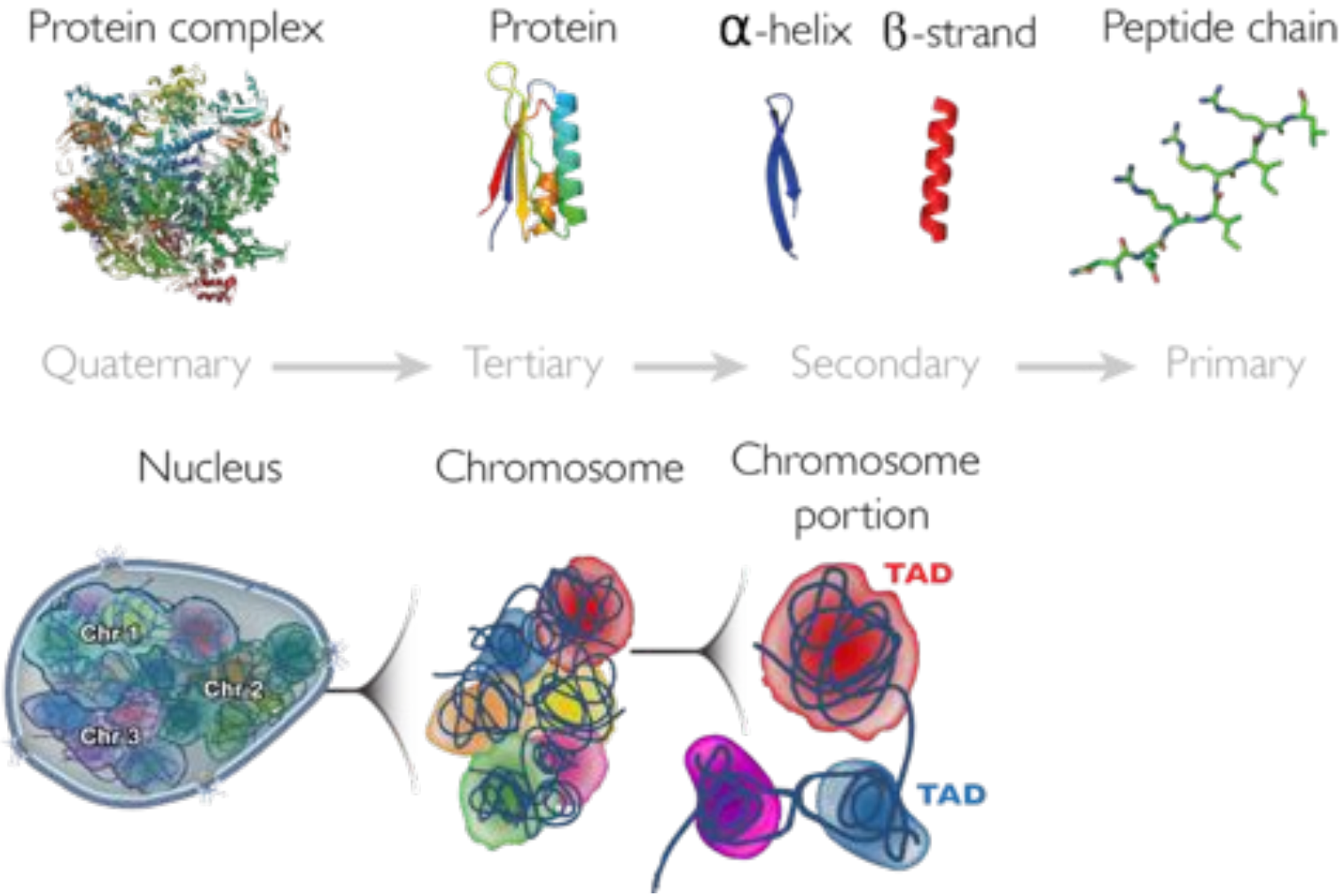
Analogous hierarchical organization of protein and genome structure



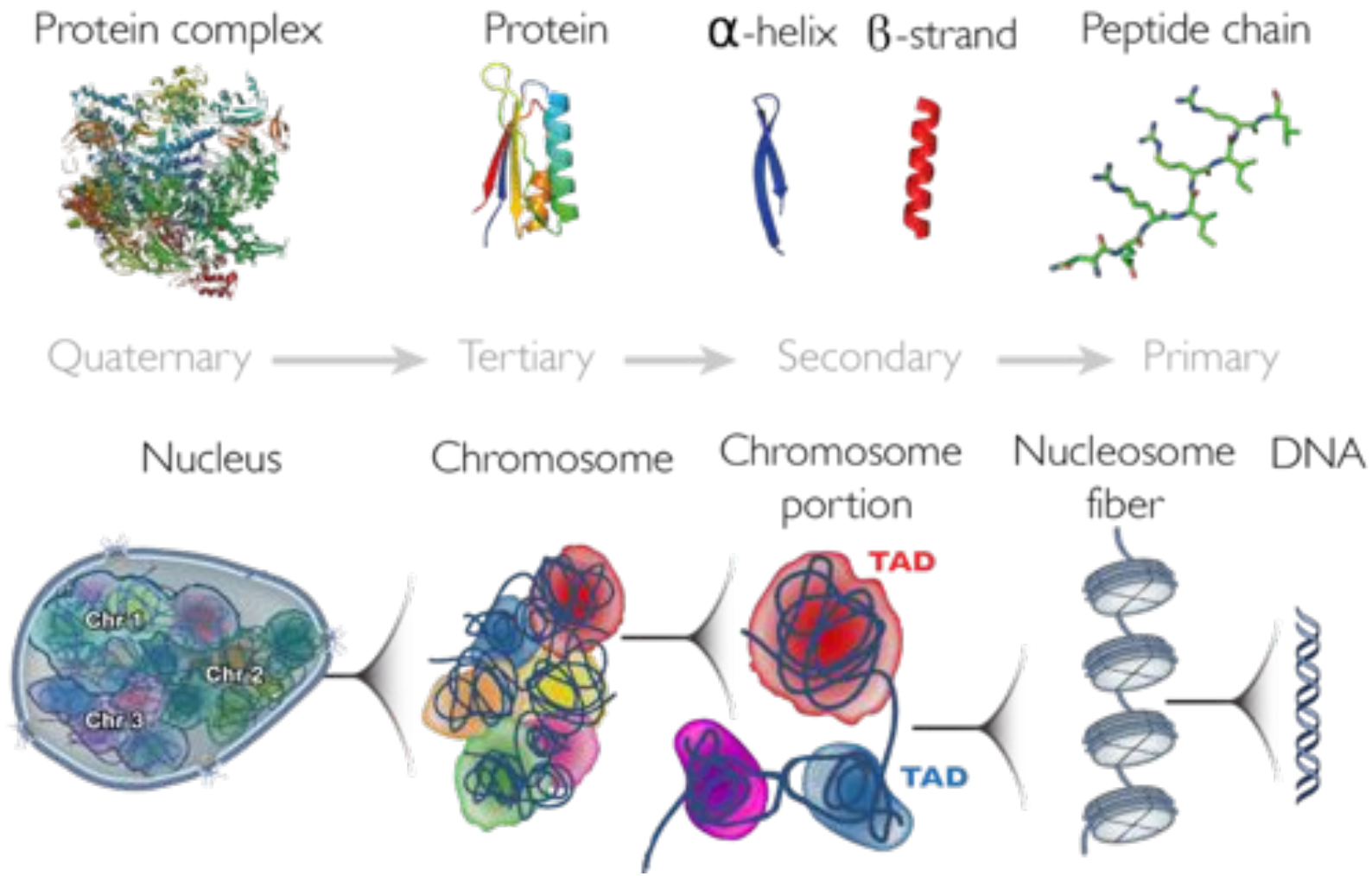
Analogous hierarchical organization of protein and genome structure



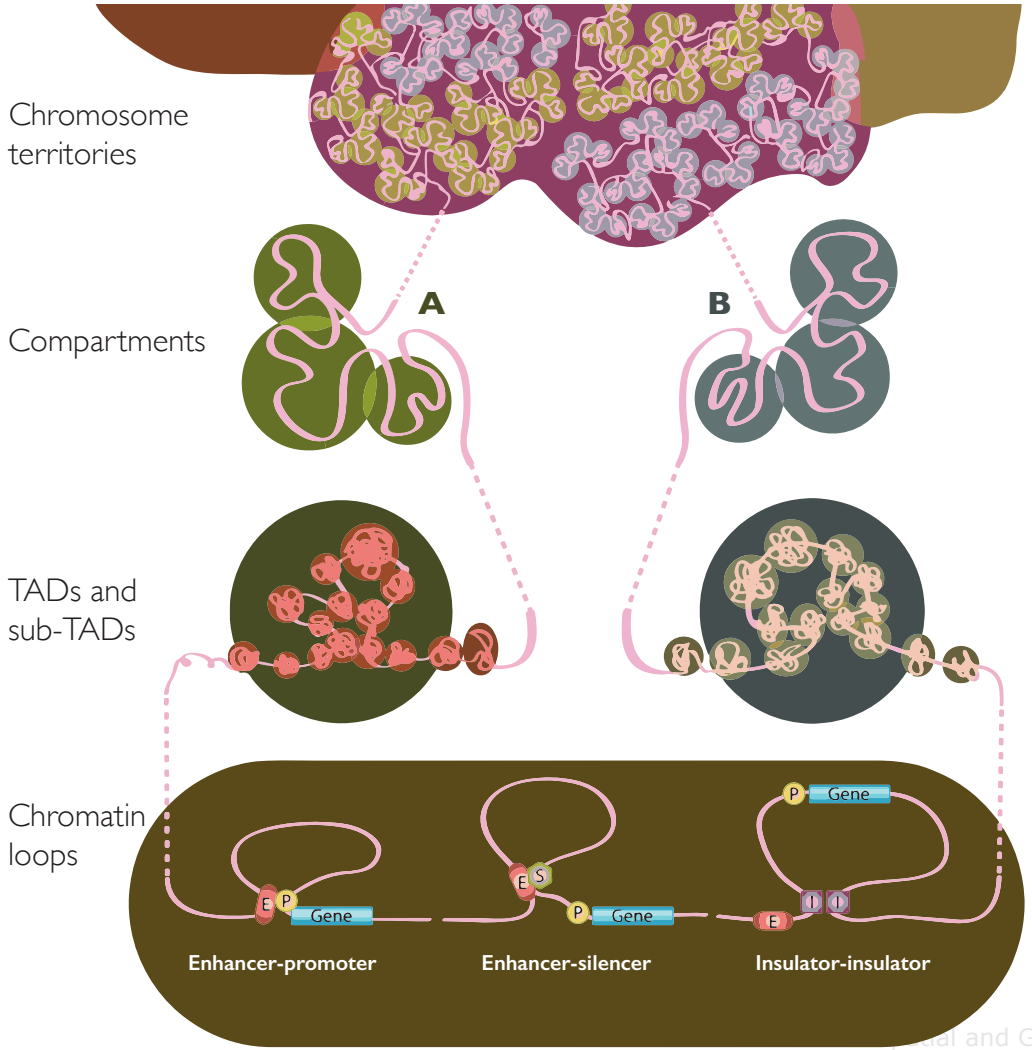
Analogous hierarchical organization of protein and genome structure



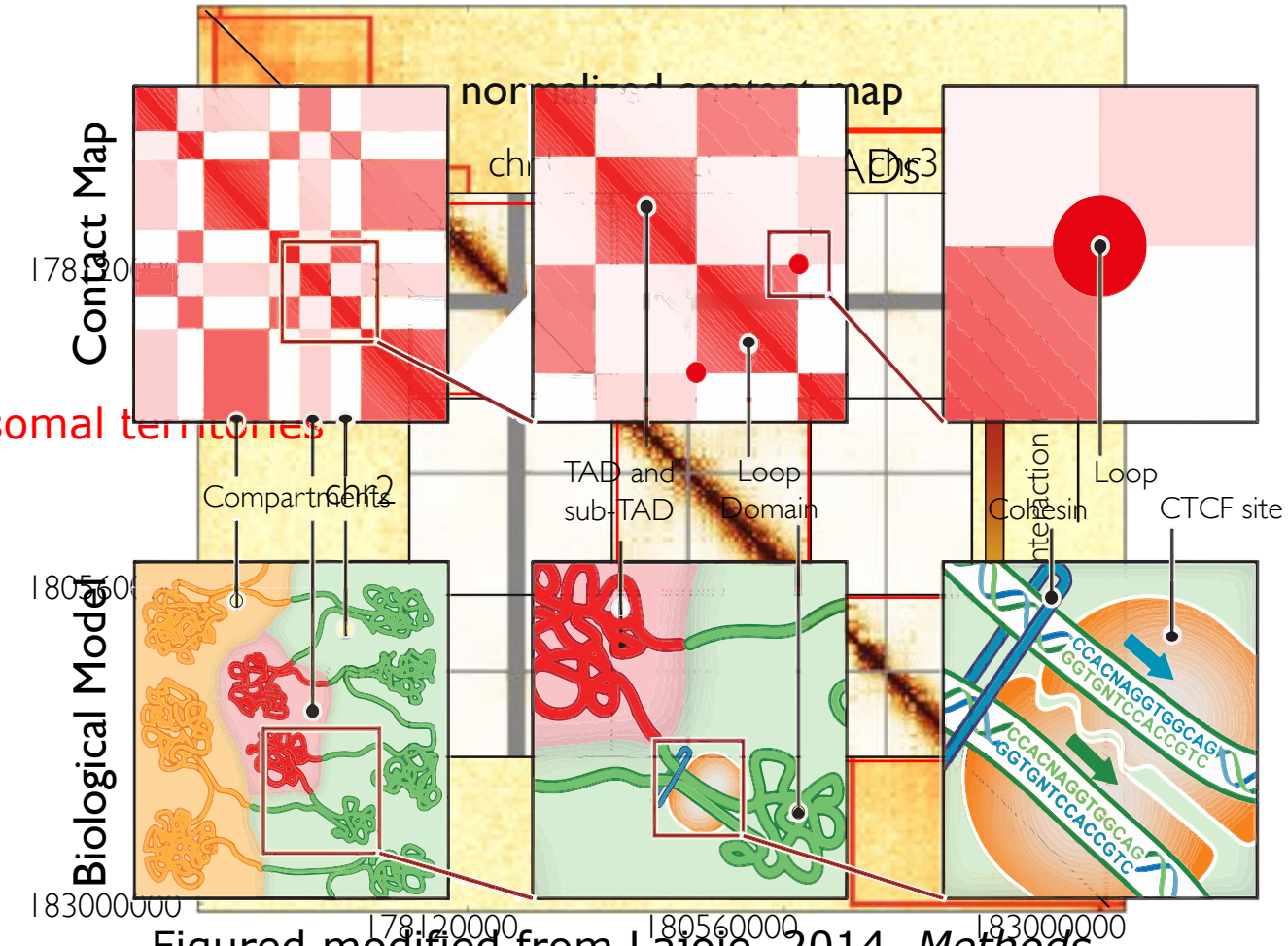
Analogous hierarchical organization of protein and genome structure



Hierarchical Organization of the Human Genome



Contact Maps Reveals Genome Structures



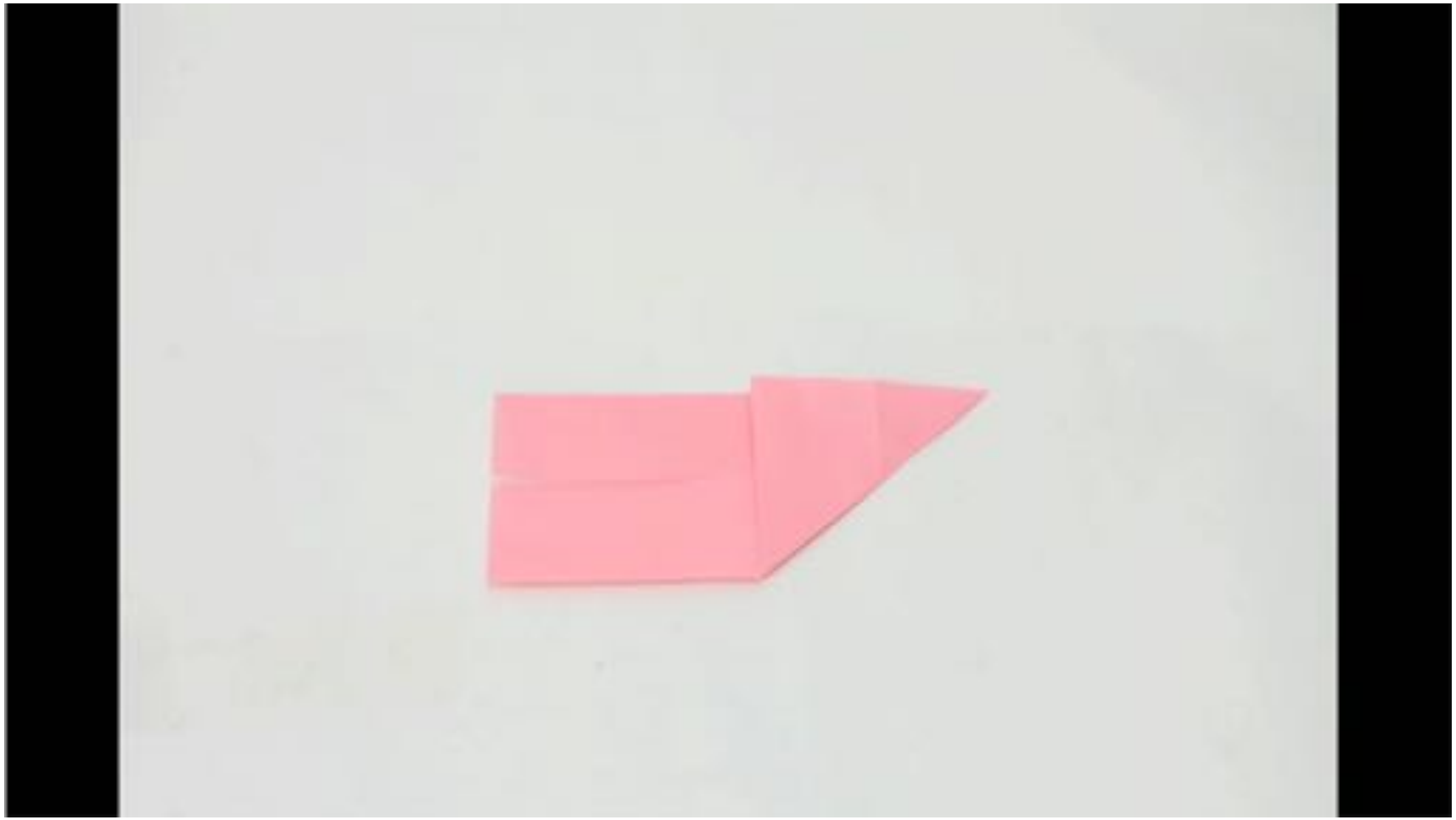
Figured modified from Lajoie, 2014, *Methods*

IMR90 Fibroblast, Chromosome 1

Rao, 2014, *Cell*

Filippova, 2013, *Algorithms Bioinforma*.

"Epigenetics is Genome Origami"



Visualization - circos plot

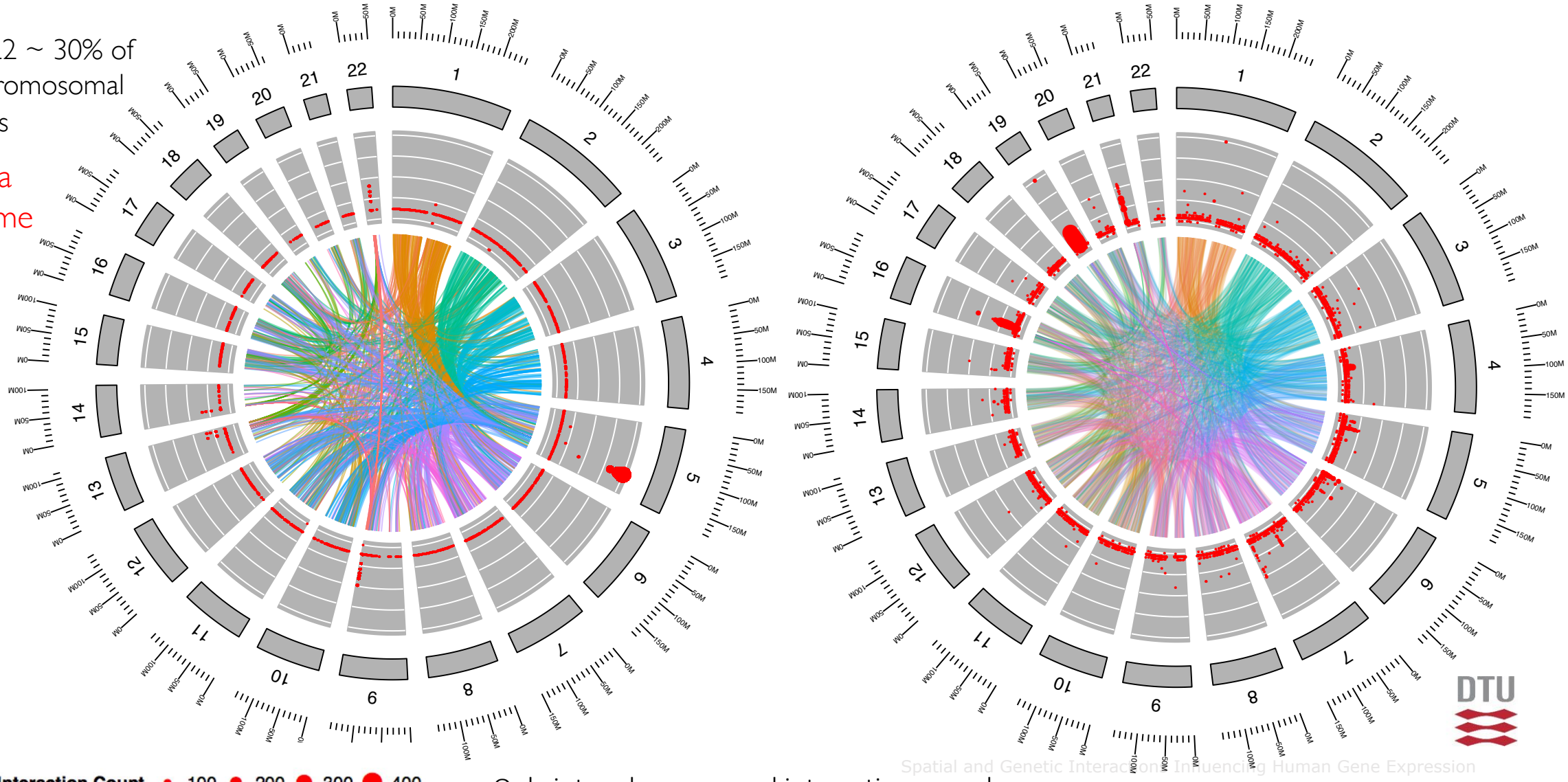
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

K652 cells Leukemia cancer cell line

hESC cells

Chr9-Chr22 ~ 30% of all inter-chromosomal interactions

Philadelphia chromosome



Only inter-chromosomal interactions are shown

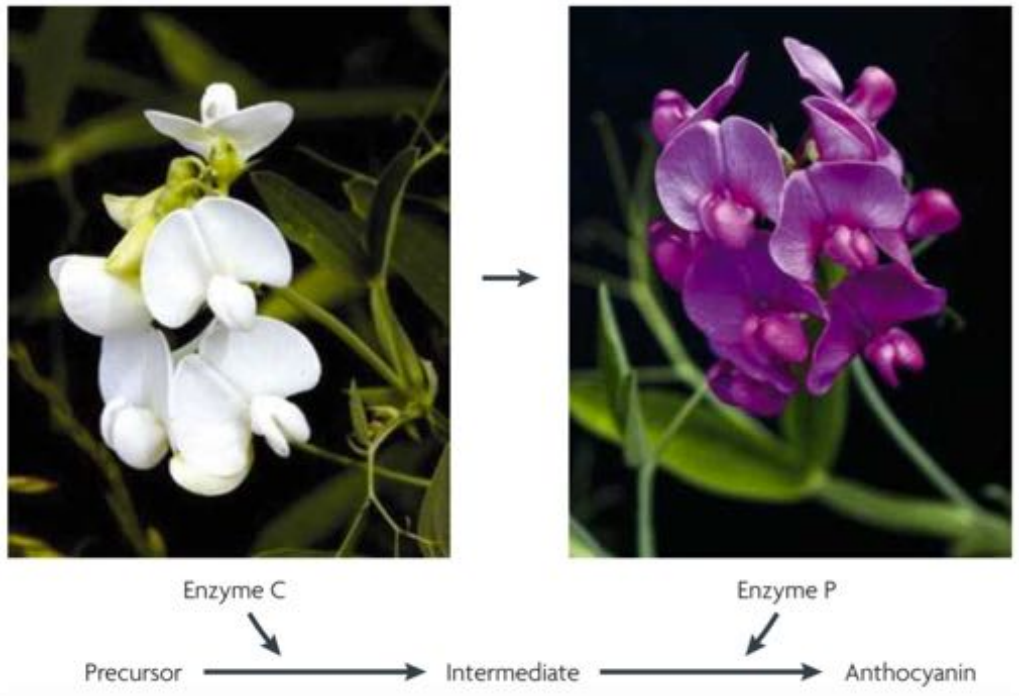
DTU

Outline

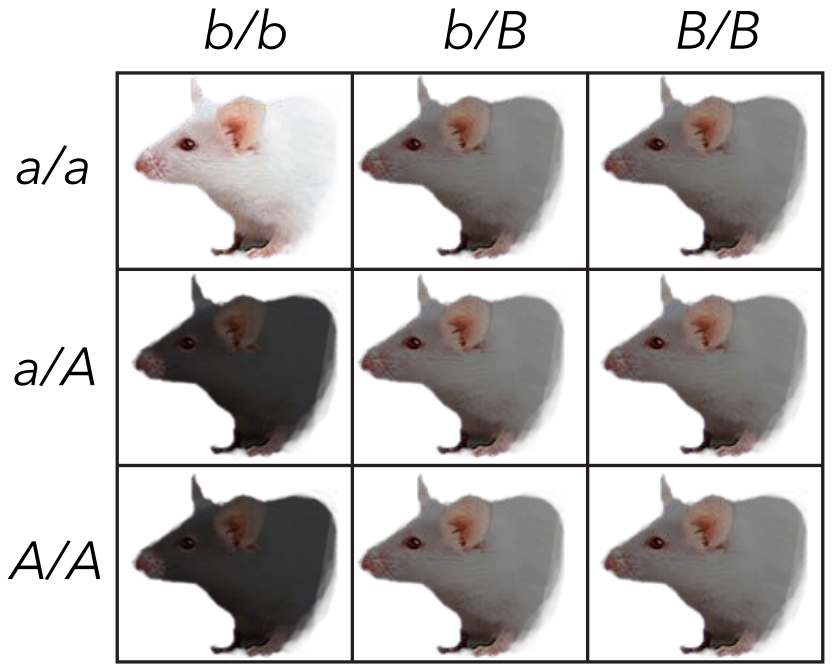
- Motivation
- Spatial Interactions (Genome Organization)
- **Genetic Interactions (Epistasis)**
- Spatial Epistasis Hypothesis
- Methods and Computational Framework
- Epistatic Signals
- “Meta-analysis”

What is Epistasis?

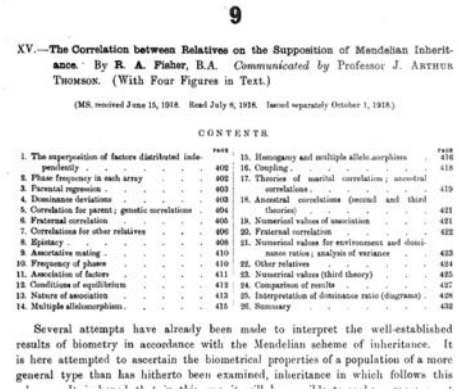
- Biological definition (Bateson, 1909, *Cambridge University Press*)
  - "Standing upon". Allele at one locus masks the effects of alleles at one or more other loci.



Bateson and Punnett classical sweet pea experiment
Adopted from Phillips, 2008, *Nat. Rev. Genet*



What is Epistasis?


9

XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. By R. A. Fisher, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS

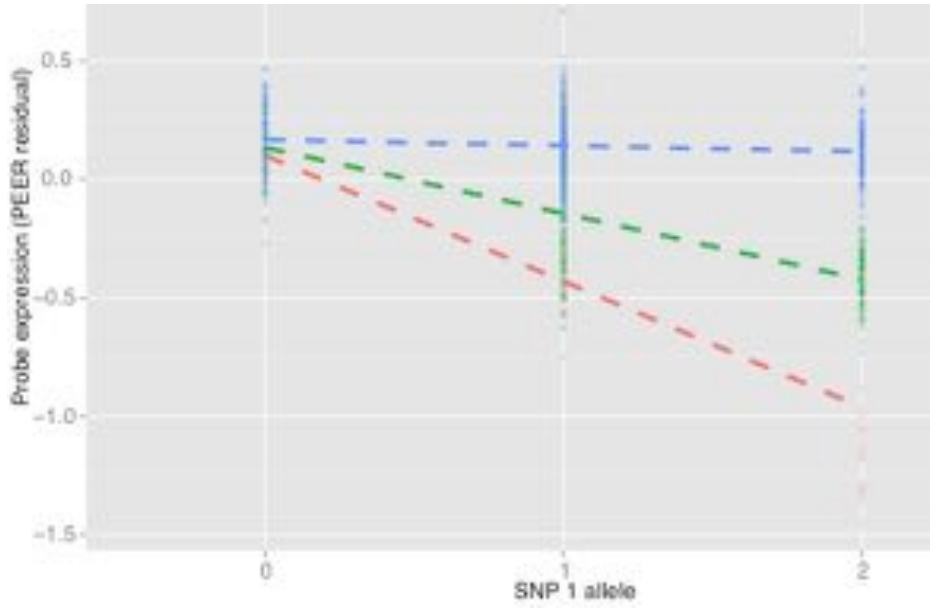
1. The superposition of factors distributed independently . . . . . page 402
2. Phase frequency in each array 402
3. Parental regression 403
4. Dominance deviations 403
5. Correlation for parent; genetic correlations 404
6. Fraternal correlation 405
7. Correlations for other relatives 406
8. Epistacy 408
9. Assortative mating 410
10. Frequency of phases 410
11. Association of factors 411
12. Conditions of equilibrium 413
13. Nature of association 413
14. Multiple allelomorphism 415
15. Homogamy and multiple allelomorphism . . . . . page 416
16. Coupling 418
17. Theories of marital correlation; ancestral correlations 419
18. Ancestral correlations (second and third theories) 421
19. Numerical values of association 421
20. Fraternal correlation 422
21. Numerical values for environment and dominance ratios; analysis of variance 423
22. Other relatives 424
23. Numerical values (third theory) 425
24. Comparison of results 427
25. Interpretation of dominance ratio (diagram) 428
26. Summary 432

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this

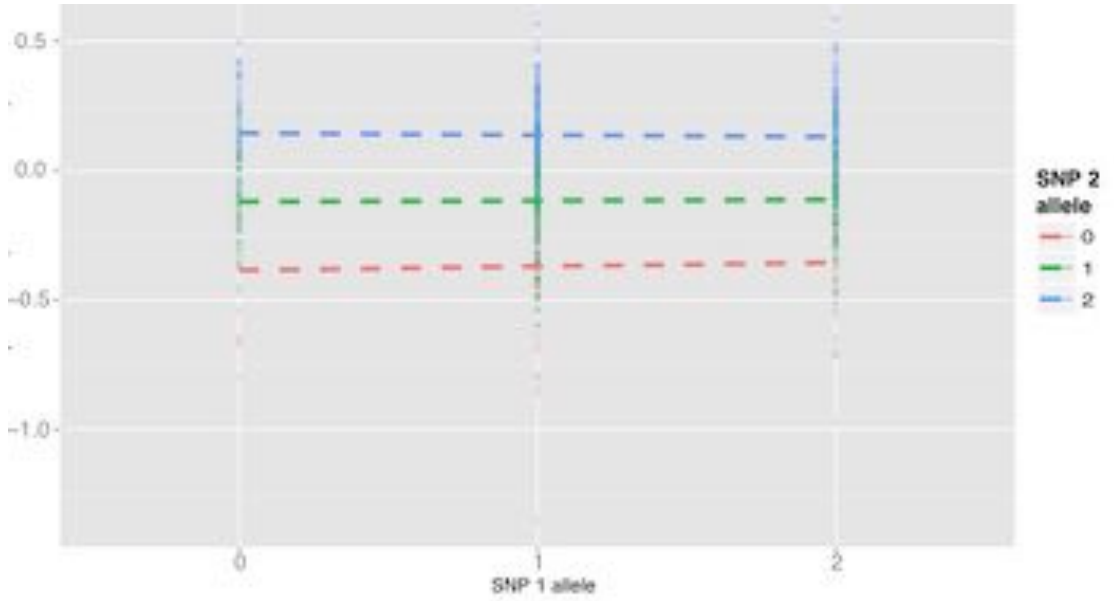
- Statistical definition (Fisher, 1919, *Trans. R. Soc.*)
  - –deviations from additivity in a linear statistical model

$$Expression(probe) = \alpha + \beta_1 \cdot SNP_1 + \beta_2 \cdot SNP_2 + \beta_{12} \cdot SNP_1 \cdot SNP_2$$

$P < 7 \times 10^{-45}$



$P = 0.5$



Outline

- Motivation
- Spatial Interactions (Genome Organization)
- Genetic Interactions (Epistasis)
- **Spatial Epistasis Hypothesis**
- Methods and Computational Framework
- Epistatic Signals
- “Meta-analysis”

Spatial Epistasis Hypothesis

Spatial genomic interactions are enriched for genetic interactions

Spatial interactions via chromatin crosstalk

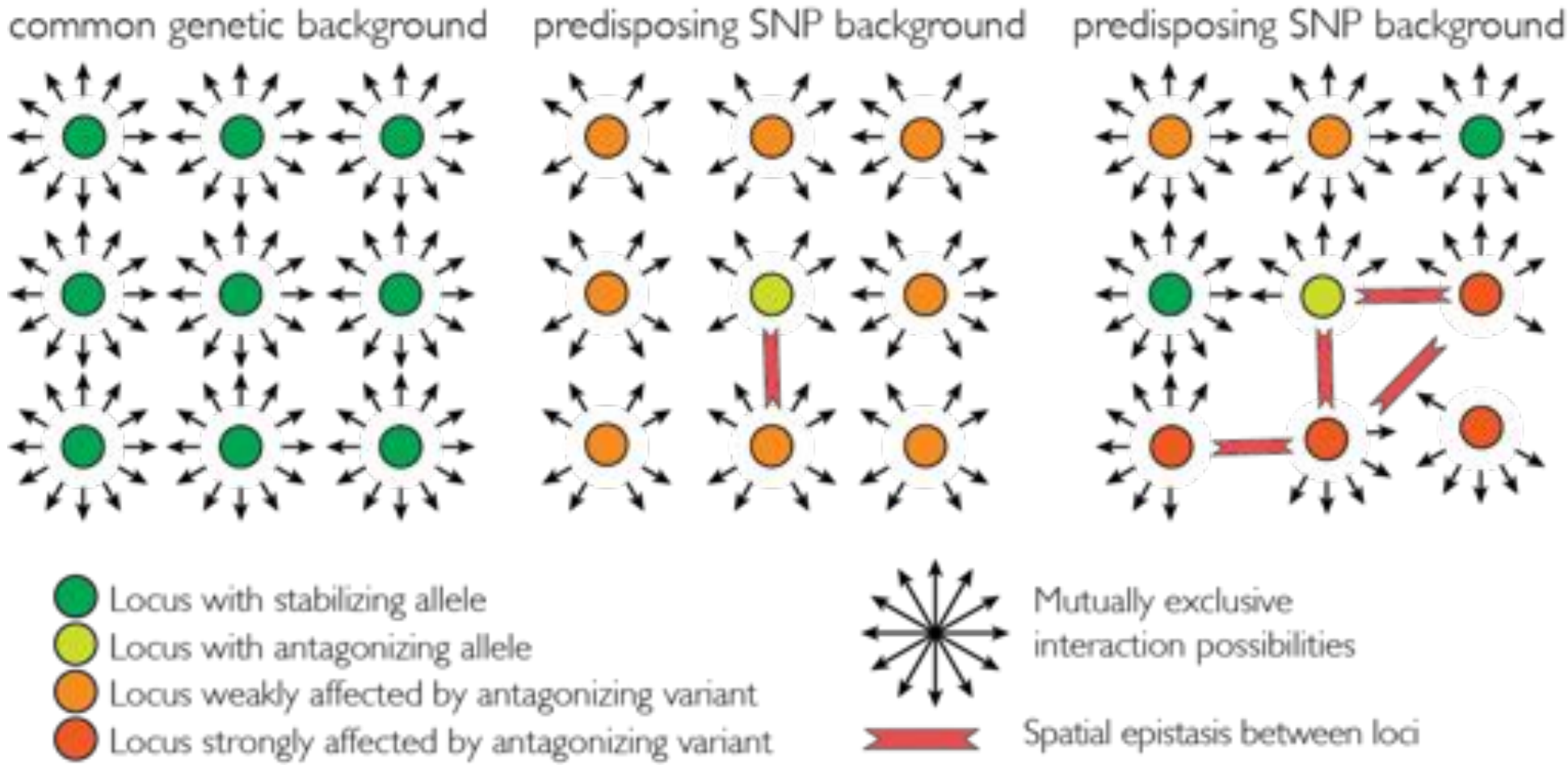
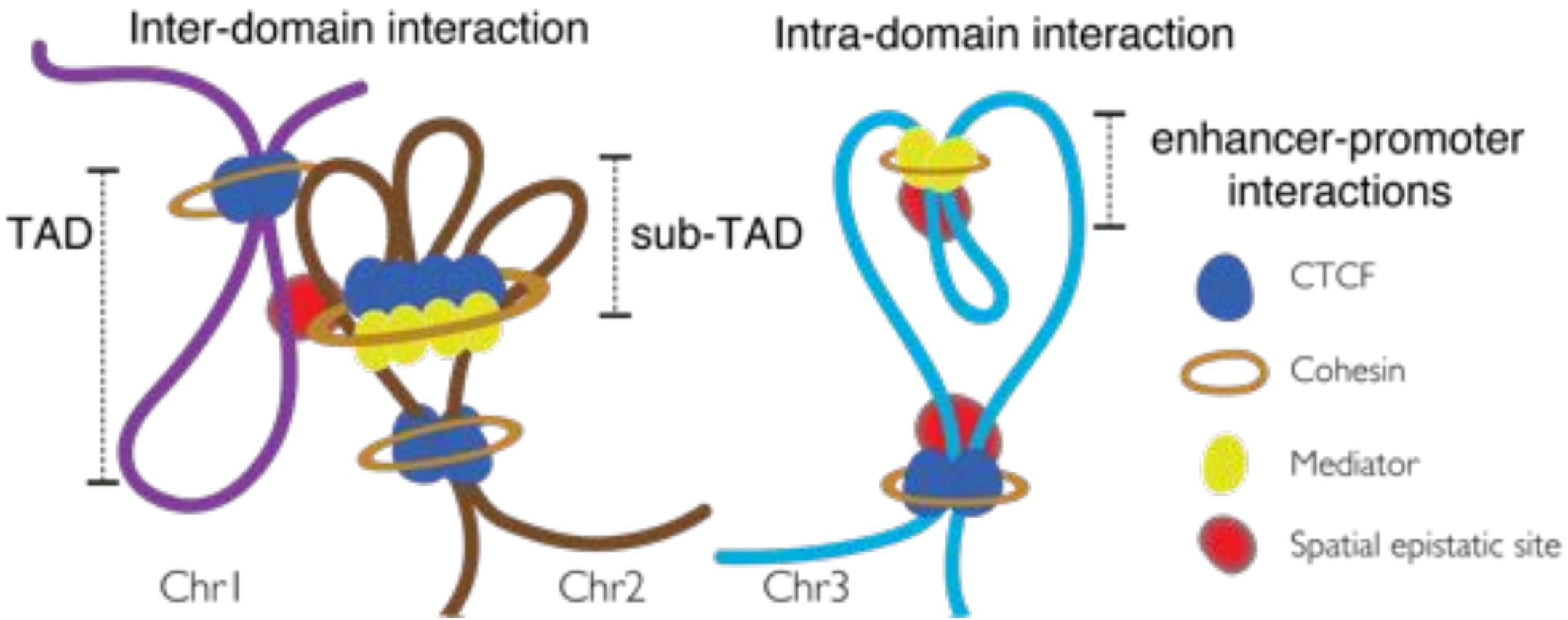
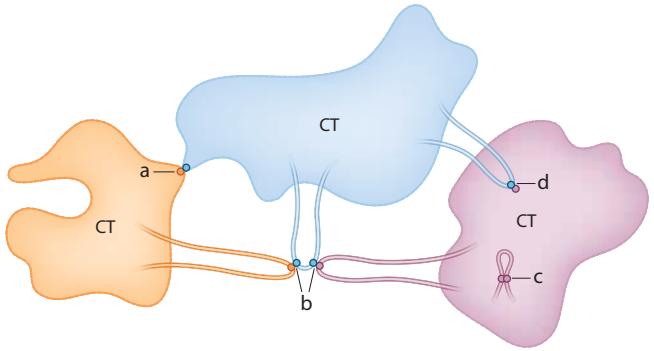


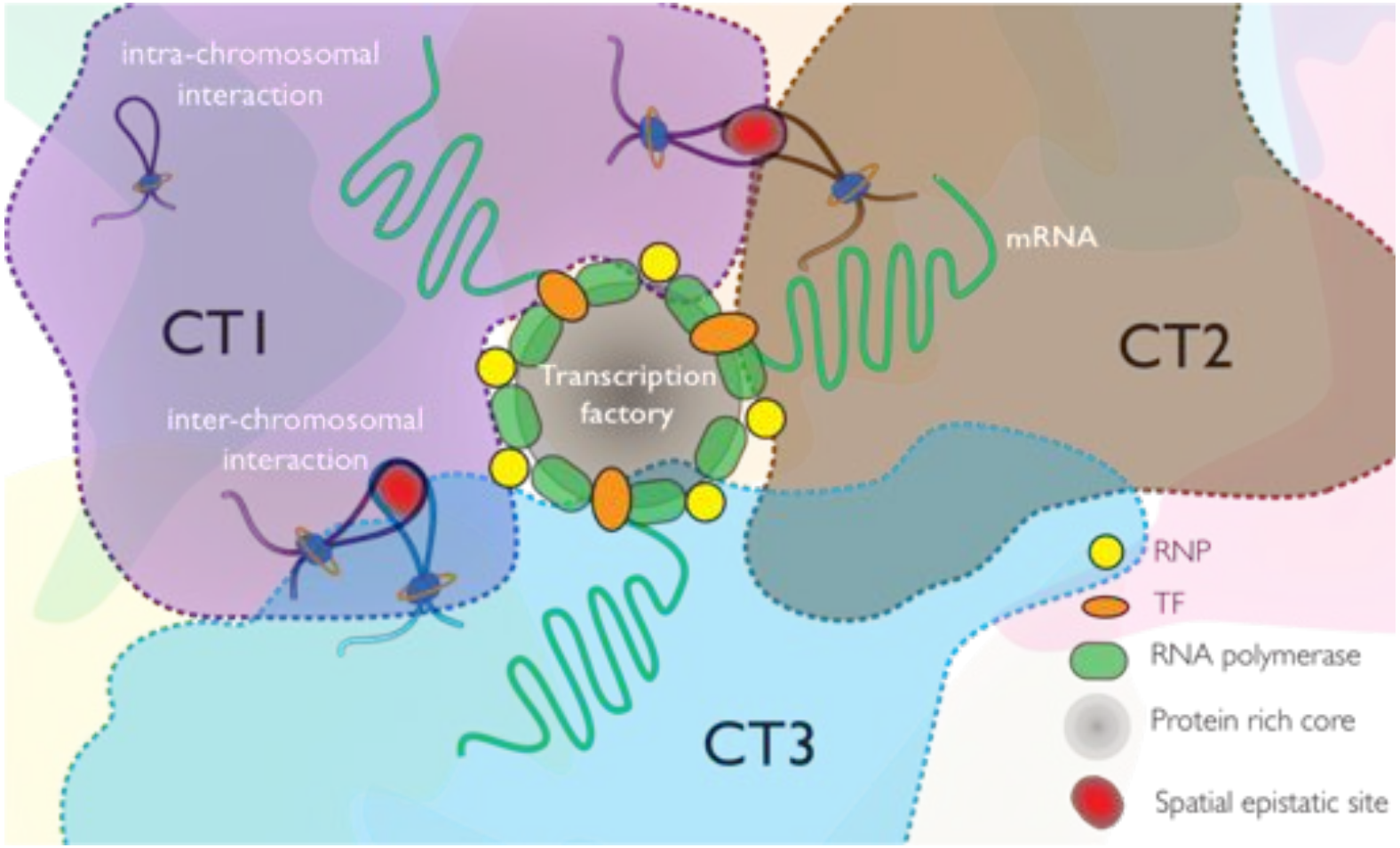
Figure inspired by Göndör, 2009, *Nature*

Biological Plausibility



Biological Plausibility





Outline

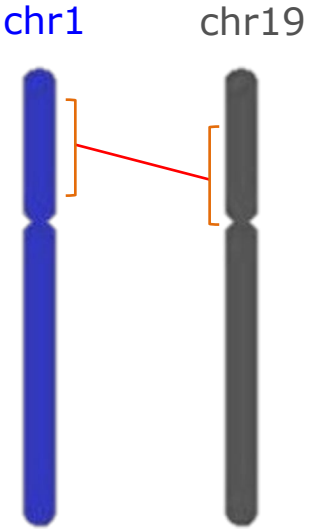
- Motivation
- Spatial Interactions (Genome Organization)
- Genetic Interactions (Epistasis)
- Spatial Epistasis Hypothesis
- **Methods and Computational Framework**
- Epistatic Signals
- “Meta-analysis”

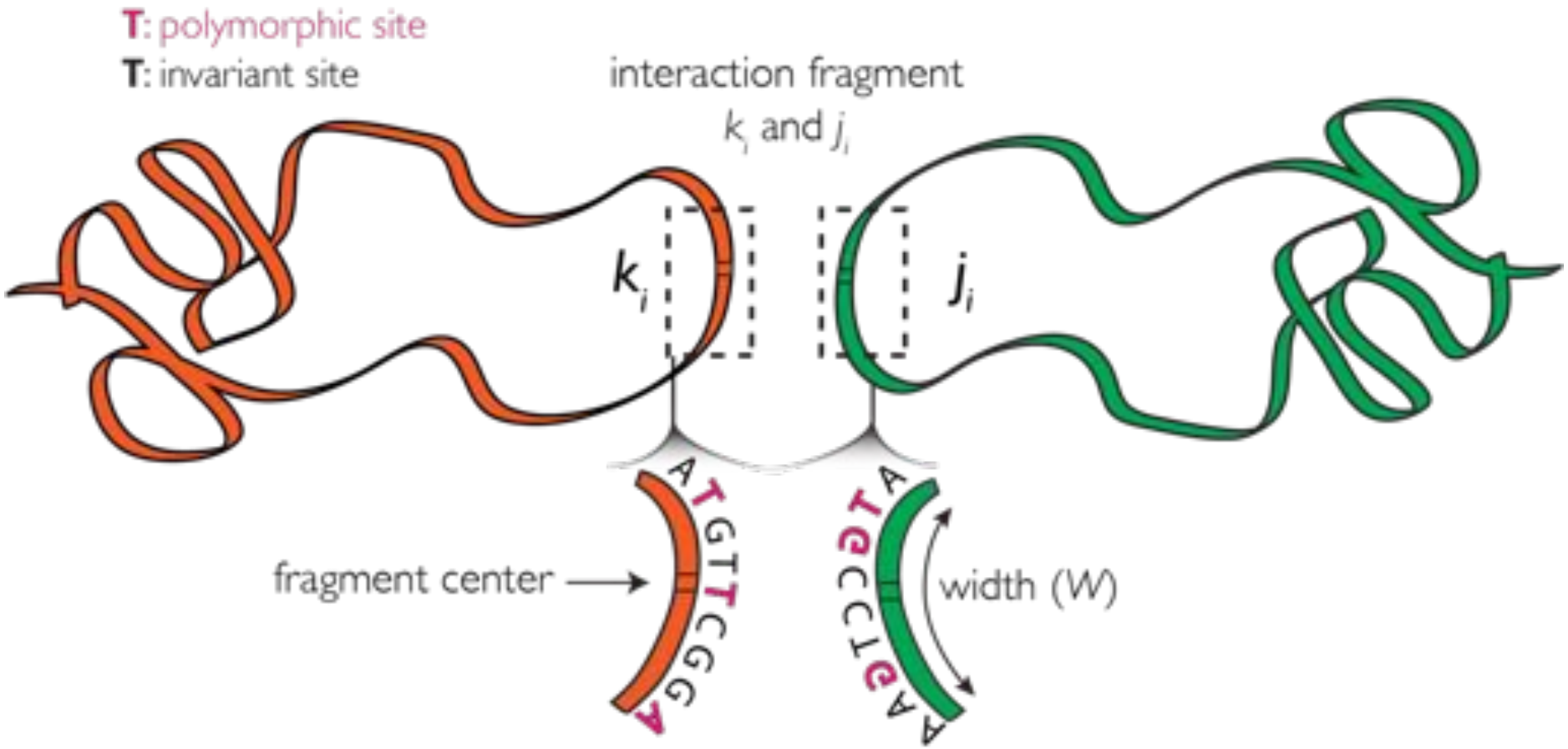
Methods: Hi-C data

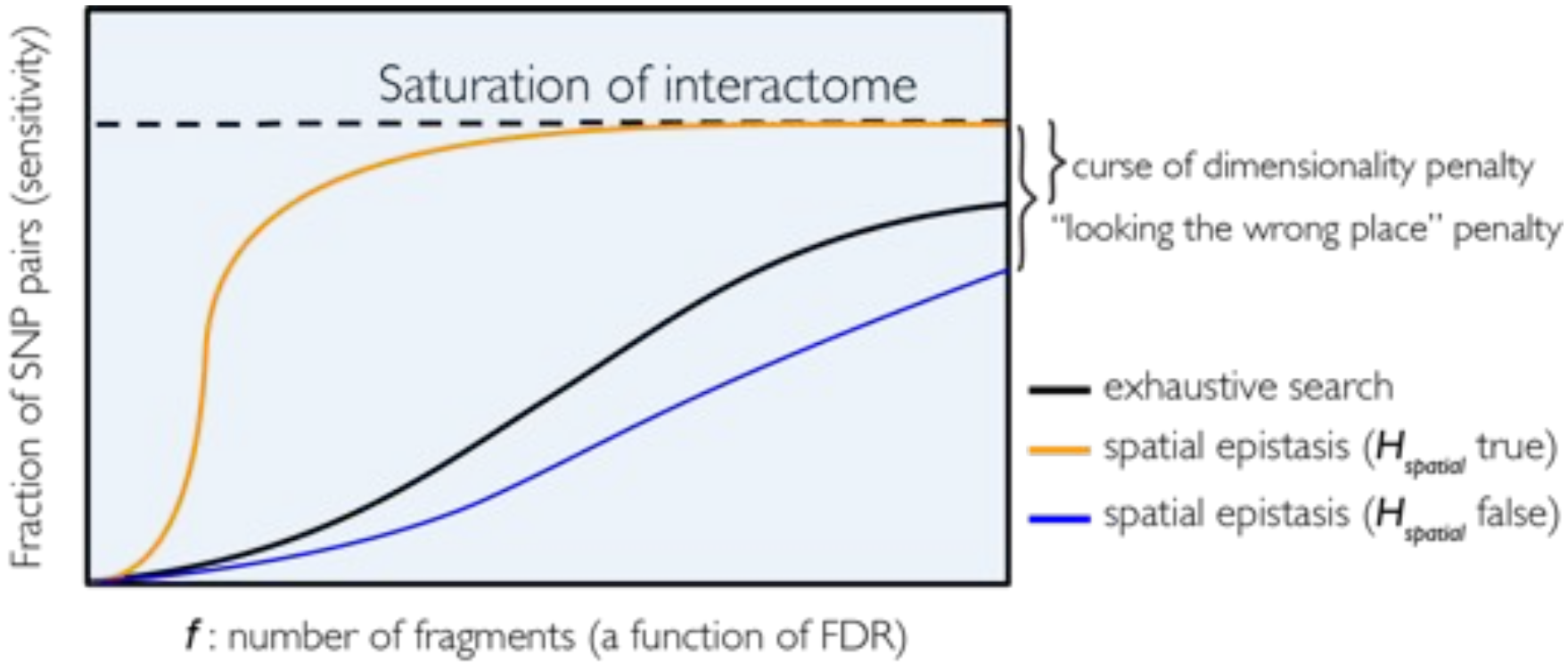
- Subset on inter-chromosomal interactions
  - Avoid problems with LD
  - Reduce the hypothesis space. Make calculations feasible.
- Select significance threshold (q-value) for interactions
- Select spatial interaction width (interaction width)

q-value cut-off

partnerA	partnerB	contactCount	p.value	q.value (FDR)
3:185305000	10:107435000	21	5.82184E-87	1.57872E-77
10:100105000	12:55885000	8	4.37324E-29	6.50331E-21
4:187965000	10:33385000	8	8.96398E-29	1.21182E-20
3:190745000	10:107585000	8	1.62105E-28	2.01971E-20
2:25205000	10:118945000	8	3.25578E-28	3.68768E-20
10:126375000	11:77885000	7	6.64959E-28	6.92406E-20
10:11495000	16:4115000	7	1.13398E-27	1.10286E-19
10:27665000	11:6745000	8	1.18998E-27	1.14763E-19

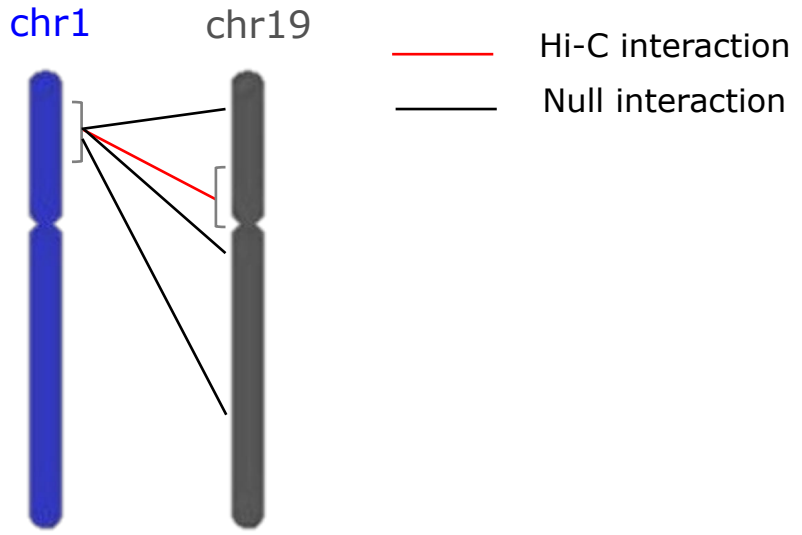



T: polymorphic site
T: invariant site
interaction fragment
k_i and j_i
k_i
j_i
fragment center
width (W)
ATGTTCGGG
AATCCCTTA


Saturation of interactome
Fraction of SNP pairs (sensitivity)
curse of dimensionality penalty
"looking the wrong place" penalty
exhaustive search
spatial epistasis ($H_{spatial}$ true)
spatial epistasis ($H_{spatial}$ false)
f : number of fragments (a function of FDR)

Null distribution

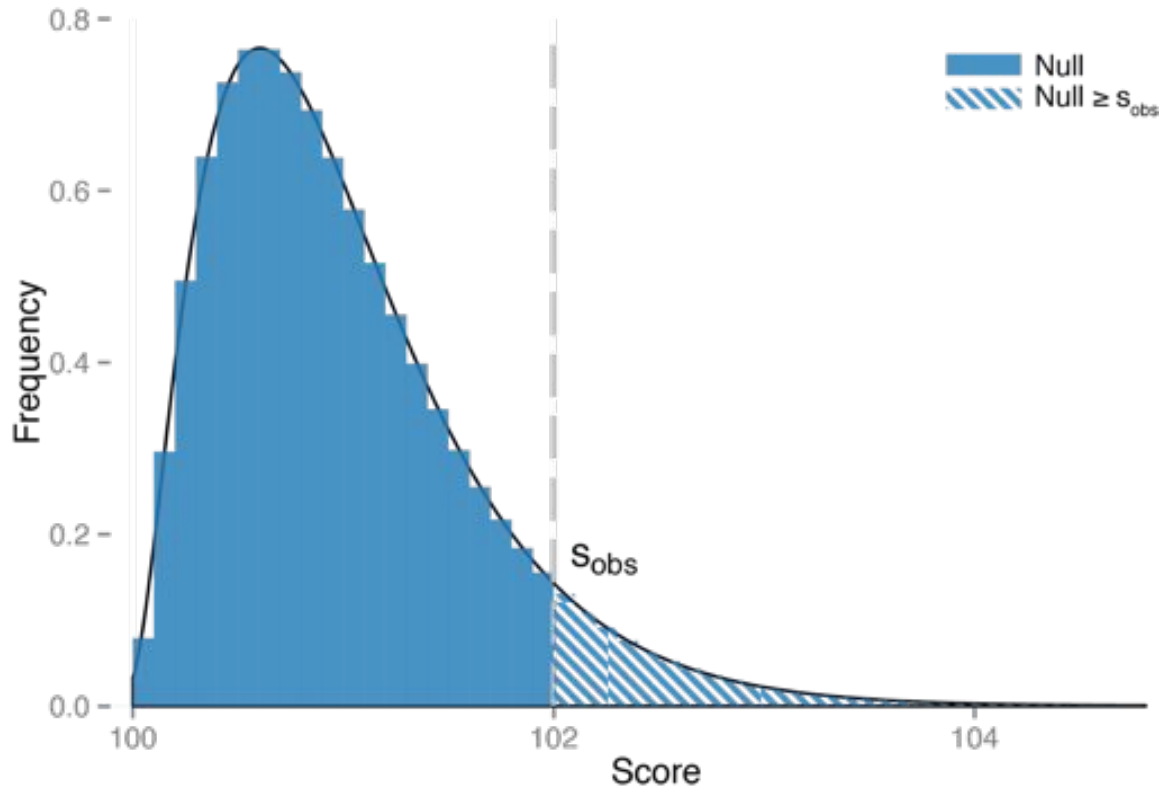
- Criteria for defining null distribution
  - –Only inter-chromosomal
  - –Distance criteria
- N=1000 permutations



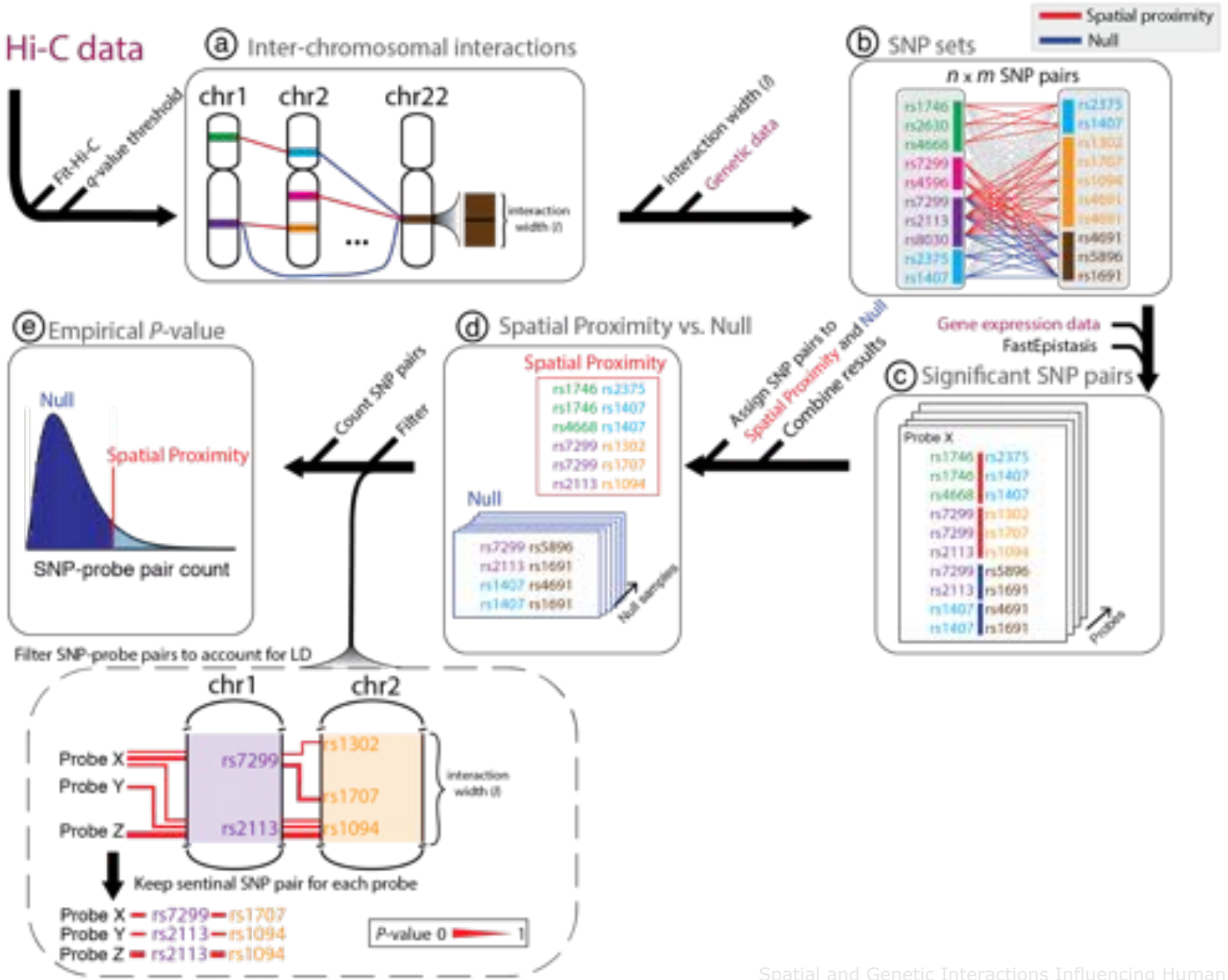
Empirical *P*-value

$$P_{empirical} = \frac{1 + \sum_{i=1}^{N} 1 \cdot [S_i \geq S_{obs}]}{N + 1}$$

N: Null samples (permutations)
S_{obs} : Observed score
S_i : i^{th} null sample score



Pipeline


Hi-C data
a Inter-chromosomal interactions
chr1
chr2
chr22
Fit-Hi-C
q-value threshold
interaction width (l)
Genetic data
b SNP sets
Spatial proximity
Null
n x m SNP pairs
c Significant SNP pairs
Gene expression data
FastEpistasis
Probe X
Assign SNP pairs to Spatial Proximity and Null
Combine results
d Spatial Proximity vs. Null
Spatial Proximity
Null
Null samples
Count SNP pairs
Filter
e Empirical P-value
Null
Spatial Proximity
SNP-probe pair count
Filter SNP-probe pairs to account for LD
chr1
chr2
Probe X
Probe Y
Probe Z
rs7299
rs2113
rs1302
rs1707
rs1094
interaction width (l)
Keep sentinal SNP pair for each probe
Probe X rs7299 rs1707
Probe Y rs2113 rs1094
Probe Z rs2113 rs1094
P-value 0 1

Key Data Sets

Genotype and Gene Expression Cohort

- Estonian Biobank (EGCUT); n=832
- Genotype data
  - Illumina HumanHap370CNV; Imputed to HapMap2
  - 2.5 million SNP
  - Standard QC thresholds (Anderson et al., 2010, *Nat. Protoc.*)
  - MAF >= 5 %
- Expression data
  - Illumina HT-12 V3.0 platform
  - ~48,000 probes
  - Removed 50 (hidden) confounding factors using PEER (Stegle et al., 2012 *Nat. Protoc.*)
  - Probe selection strategy to select “best” ~10,000 probes

Hi-C Data

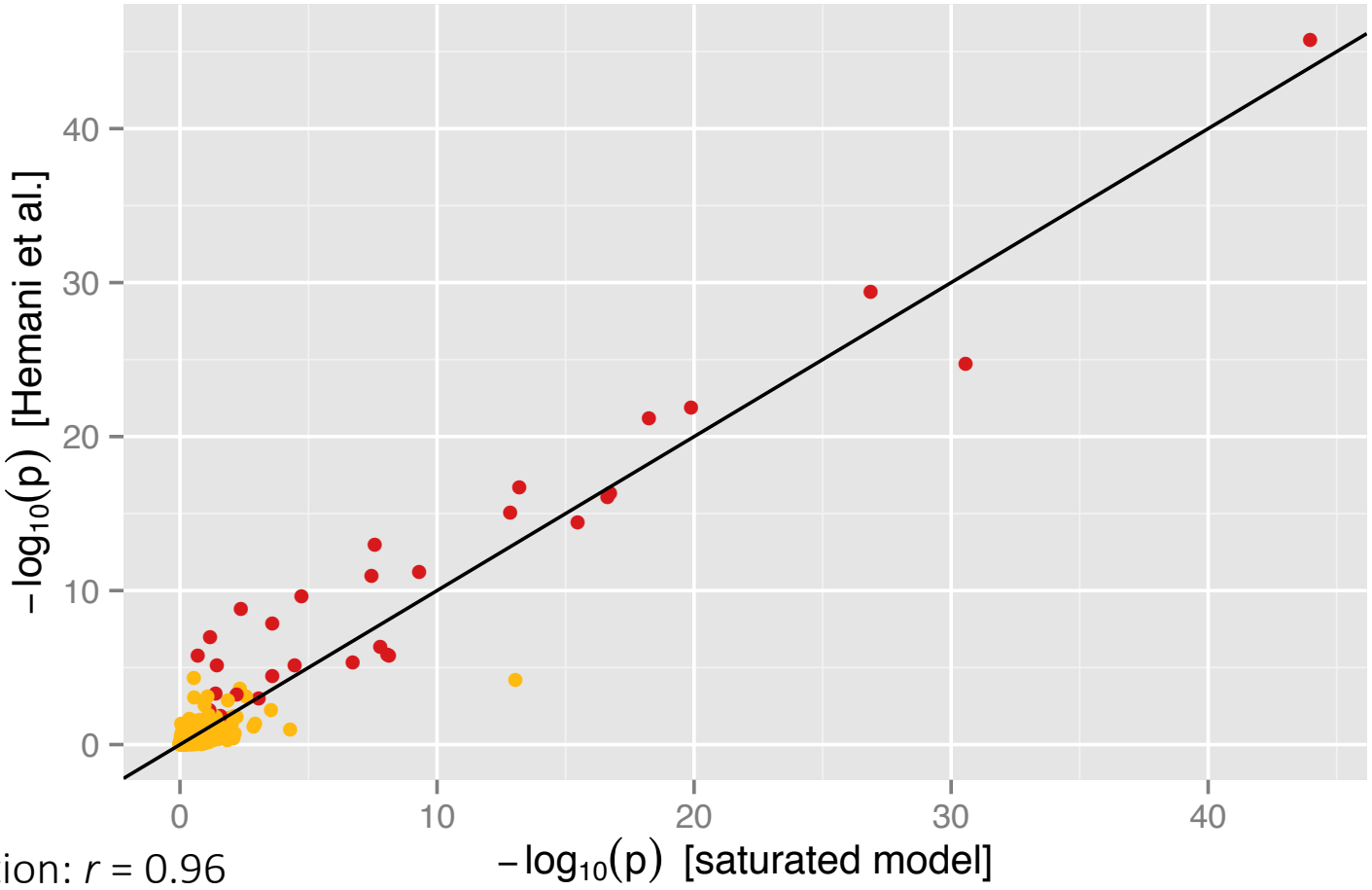
- Lieberman-Aiden, 2009, *Science*
  - K652 (bone marrow; erythroleukemia cell line) [~95 x 10^3 interactions]

  Poisson regression model. Available from Lan et at., 2012*, Nucleic Acids Res*
- Dixon, 2012, *Nature*
  - hESC (embryonic stem cells) [90 x 10^6 interactions]
  - IMR90 (lung fibroblast) [160 x 10^6 interactions]

  Fit-Hi-C method (Ay et al., 2014, *Genome Research*).

Epistatic Signals

Replication of Hemani (2014, *Nature*)



434 SNP-pairs

Pearson correlation: $r = 0.96$

Spatial Epistasis Discovery

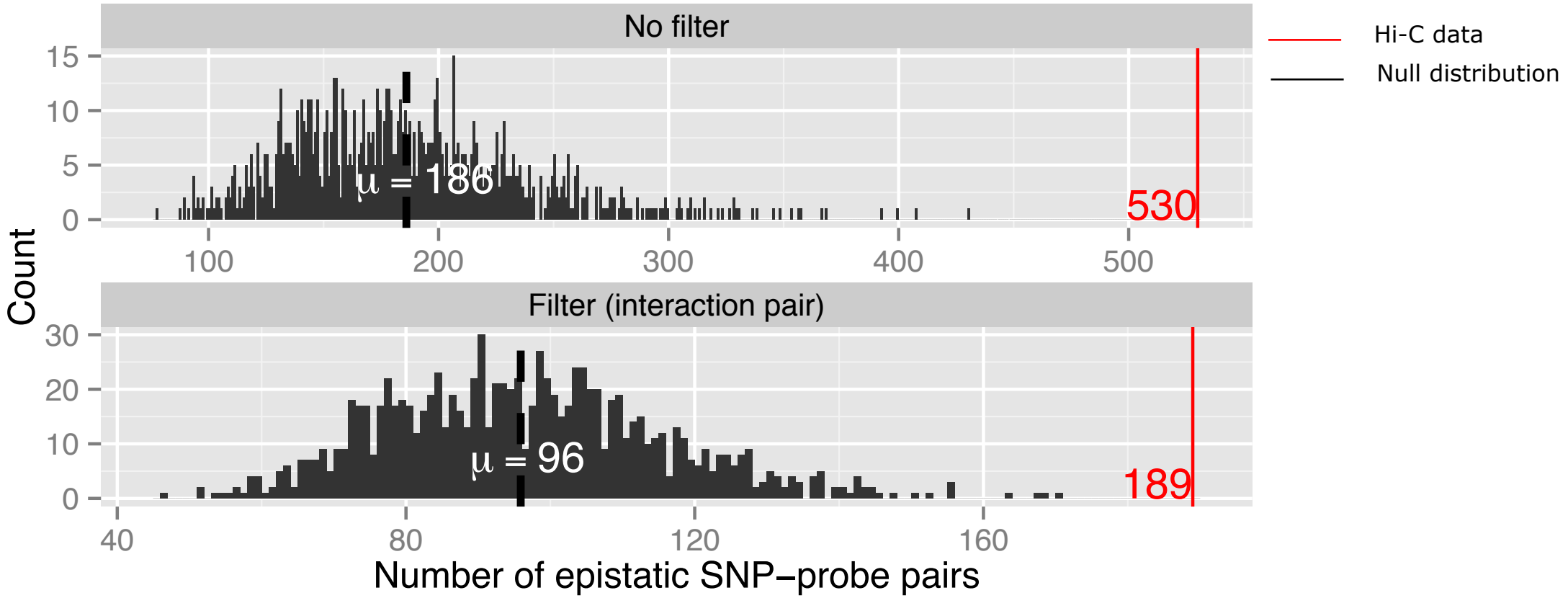
- Epistasis tests per probe: 1.33×10^{14}
- Single threaded CPU time: >12 years
- Only 0.35% of the possible two-locus SNP-pairs investigated
- Search space covered 6.62% of the genome

Spatial Epistasis Enrichment

Cell line	Interaction width, l	q-value ($n_{interactions}$)	No. epistasis tests *per probe*†	P-value‡
hIMR90	500 bp	10^{-6} (26,325)	260,273,820	0.078
	1,000 bp	10^{-6} (26,325)	1,042,826,407	0.001 *
	2,500 bp	10^{-7} (8,114)	733,442,624	0.020 *
	50,000 bp	10^{-9} (1,021)	4,284,422,018	0.927
hESC	500 bp	10^{-16} (2,665)	3,222,025	0.257
	500 bp	10^{-14} (11,919)	57,787,904	0.620
	1,000 bp	10^{-12} (39,966)	2,236,957,368	0.643
	2,500 bp	10^{-13} (24,084)	5,560,830,560	0.143
K562	1,000 bp	NA (1,263)	1,485,960	0.540
	5,000 bp	NA (1,263)	28,045,983	0.780
$hIMR90_{control}$	1,000 bp	NA (5,000)	45,668,788	0.169
$hESC_{control}$	1,000 bp	NA (5,000)	45,714,894	0.317

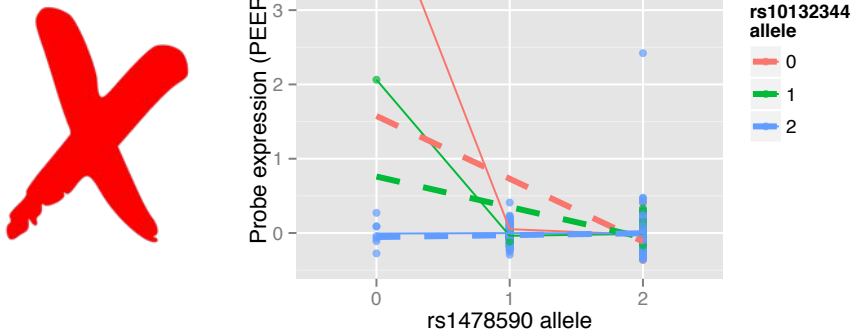
†total number of statistical tests between interacting and non-interacting loci *per probe*.
‡P-value derived from filtered SNP-probe pairs.

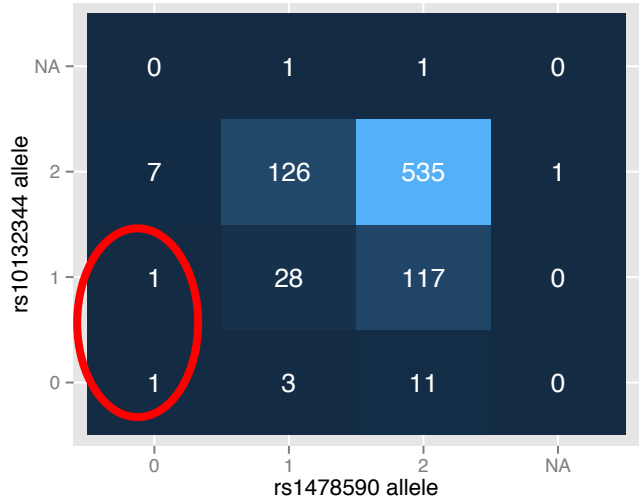
Strong Enrichment Signal for hIMR90 cells



$$\mathrm{hIMR90}[l = 1,000\mathrm{bp}; q = 10^{-6}]$$

Spurious Epistatic Effects for Rare Variants

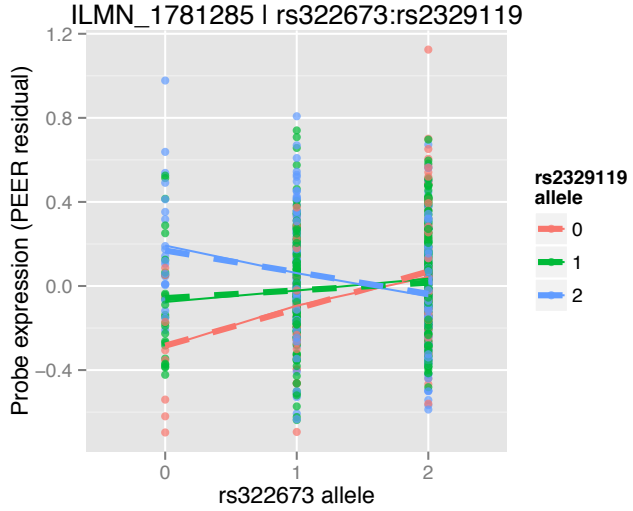


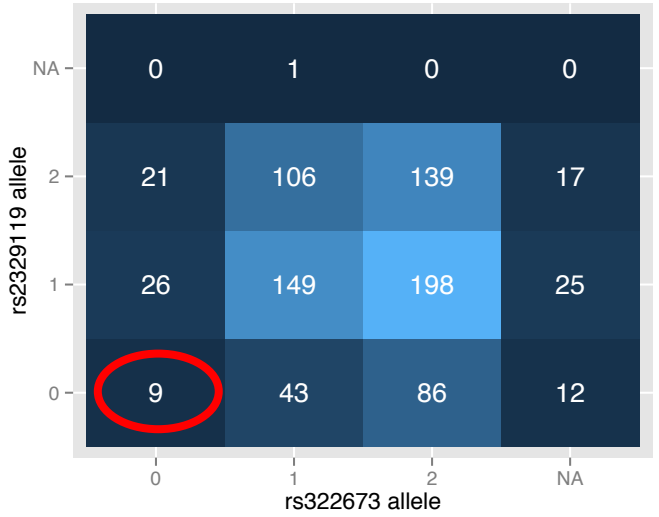


$\beta_{interaction} = -0.44$

P-value = $1.29 \cdot 10^{-39}$

minimal Genotype Class Count = 1



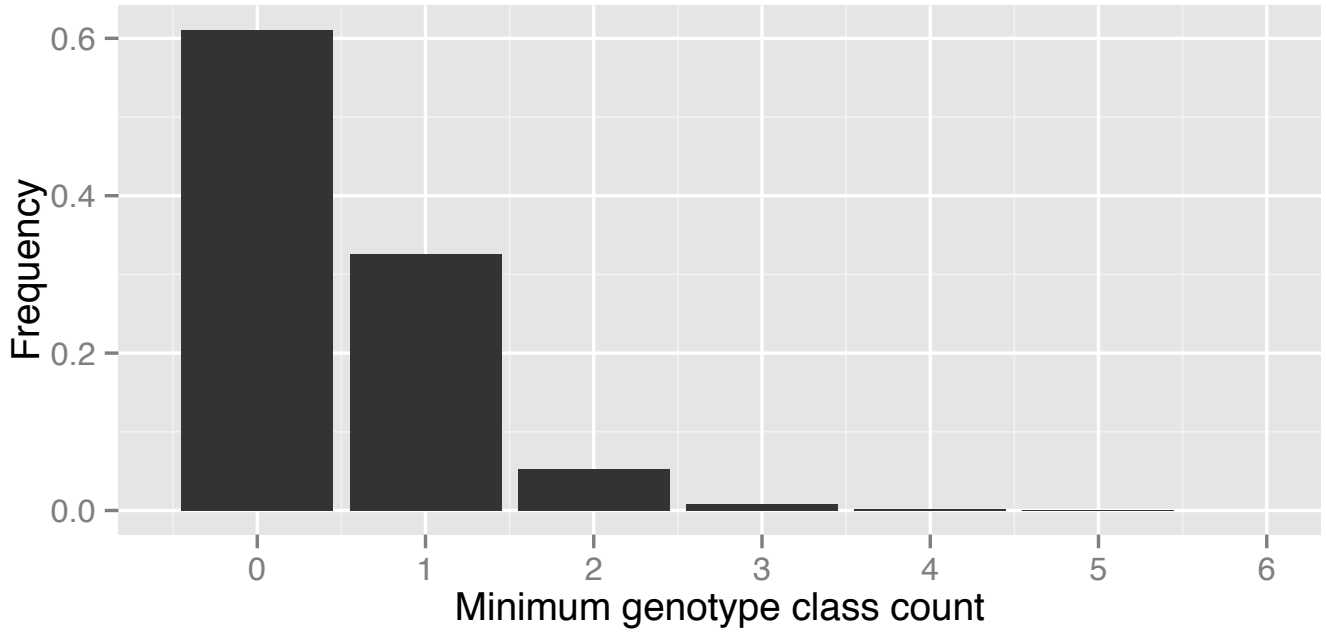


$\beta_{interaction} = -0.14$

P-value = $4.07 \cdot 10^{-11}$

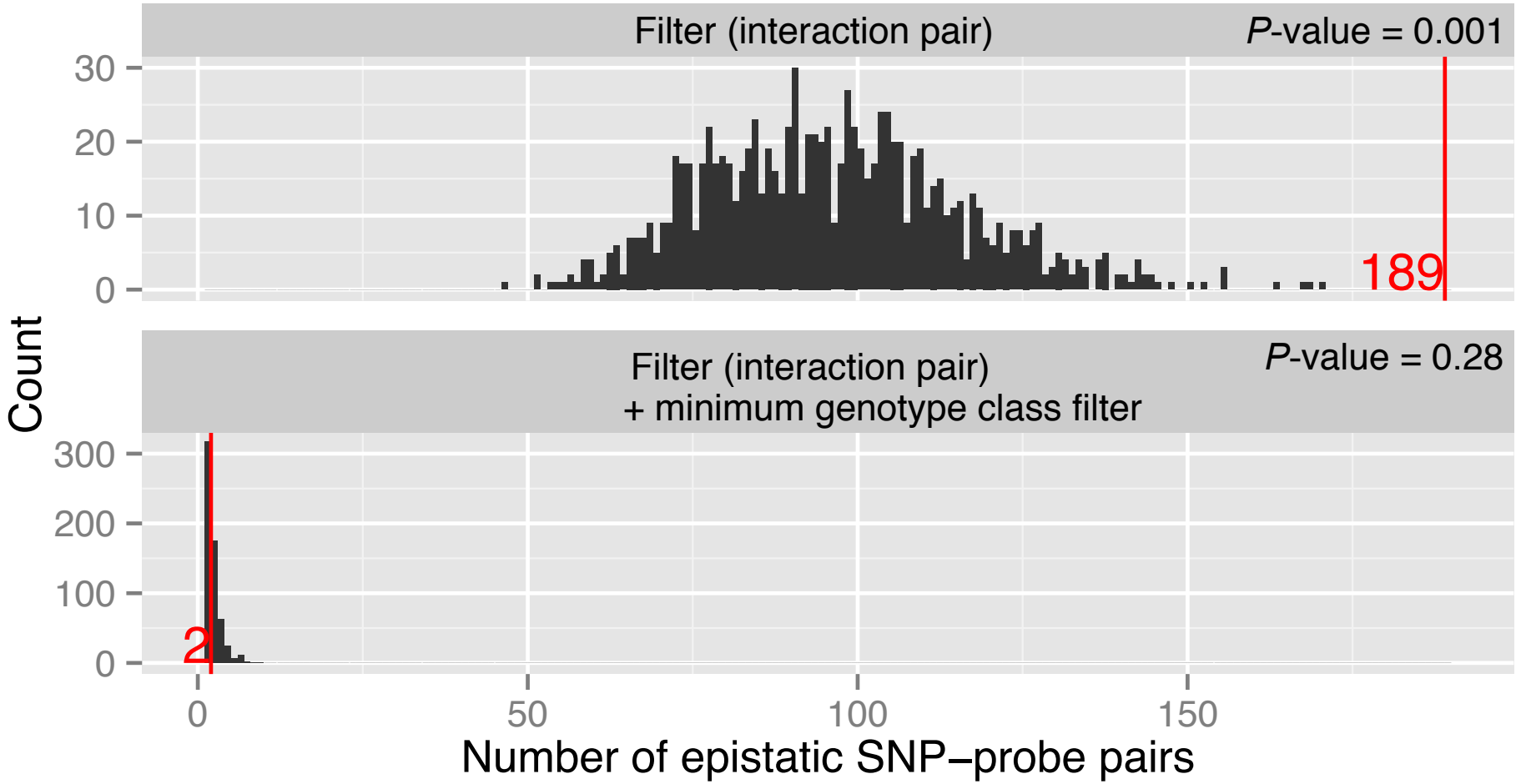
minimal Genotype Class Count = 9

Distribution of Minimum Genotype Class Count



Most apparent epistatic SNP-pairs have minimum genotype class count < 3

Deeper Insights into the hIMR90 Spatial Epistasis Enrichment



Conclusion

Conclusion

- Developed potential powerful computational framework for testing biological hypothesis related to mechanisms of epistasis.
- Insufficient power to detect epistasis
- Cannot reject or accept the spatial epistasis hypothesis

Future Work

- Test for spatial epistasis enrichment
  - –TAD boundaries
  - –CTCF motifs

Meta-analysis

Meta-analysis: datasets

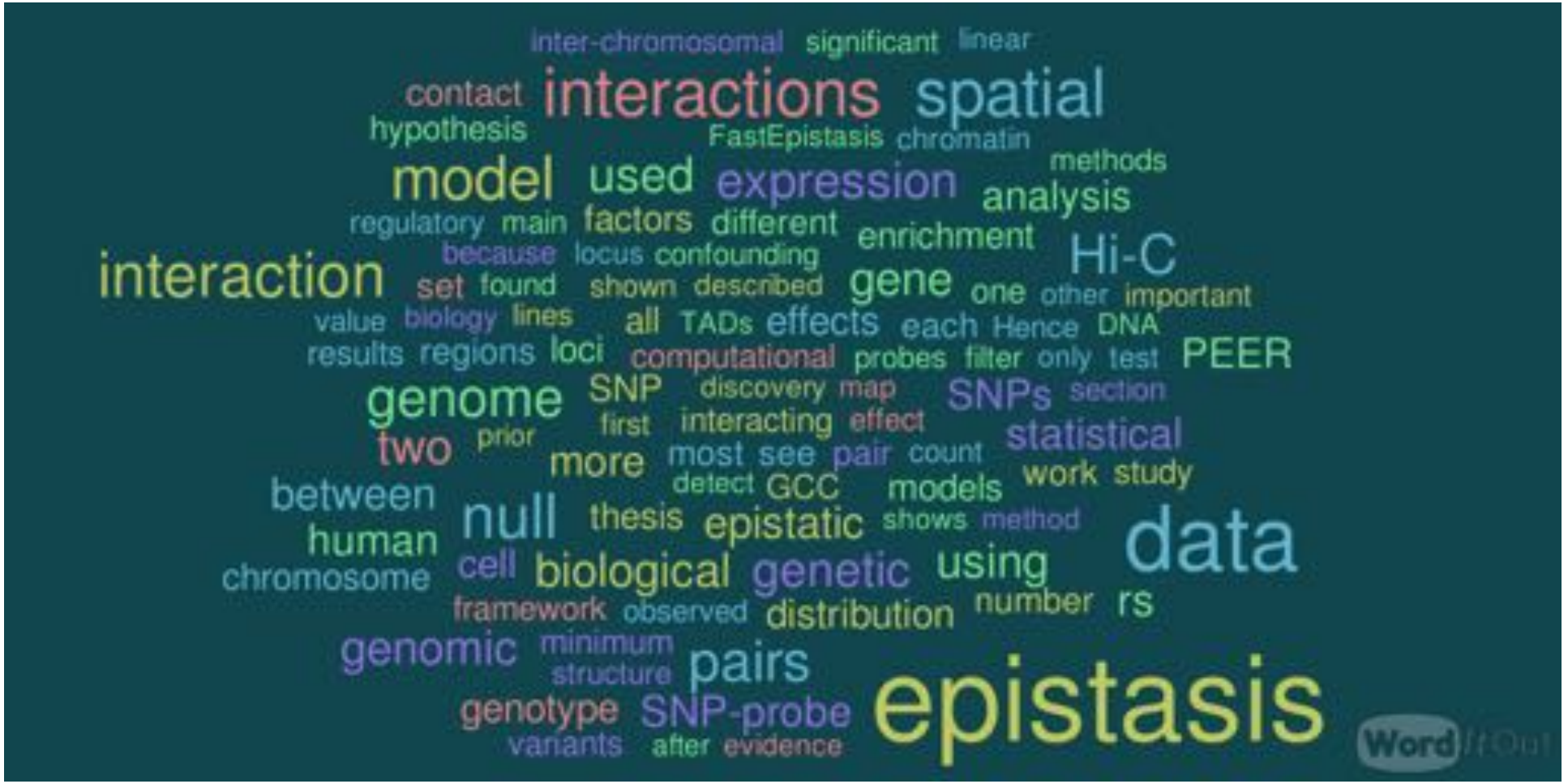
Exist.io: Understand your life.

Connected Apps

RescueTime

Thesis and Thoughts

Two sides of the same coin?



Thesis corpus

Thesis and Thoughts

Two sides of the same coin?

Daily log corpus (10. November 2015 and onwards)

Similarity index: 0.3972 (cosine)

Thesis and Thoughts

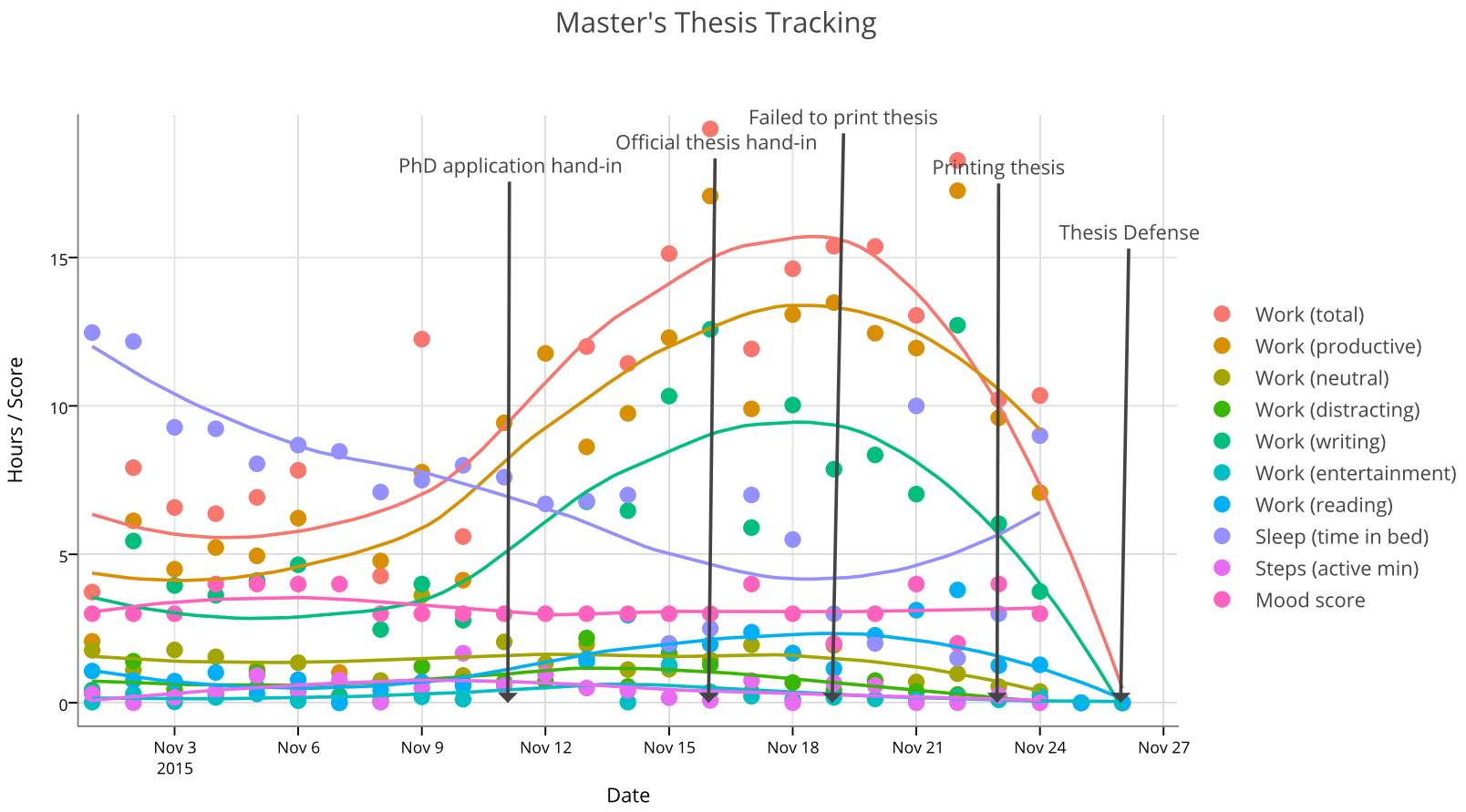
Two sides of the same coin?

Daily log corpus (all 2015)

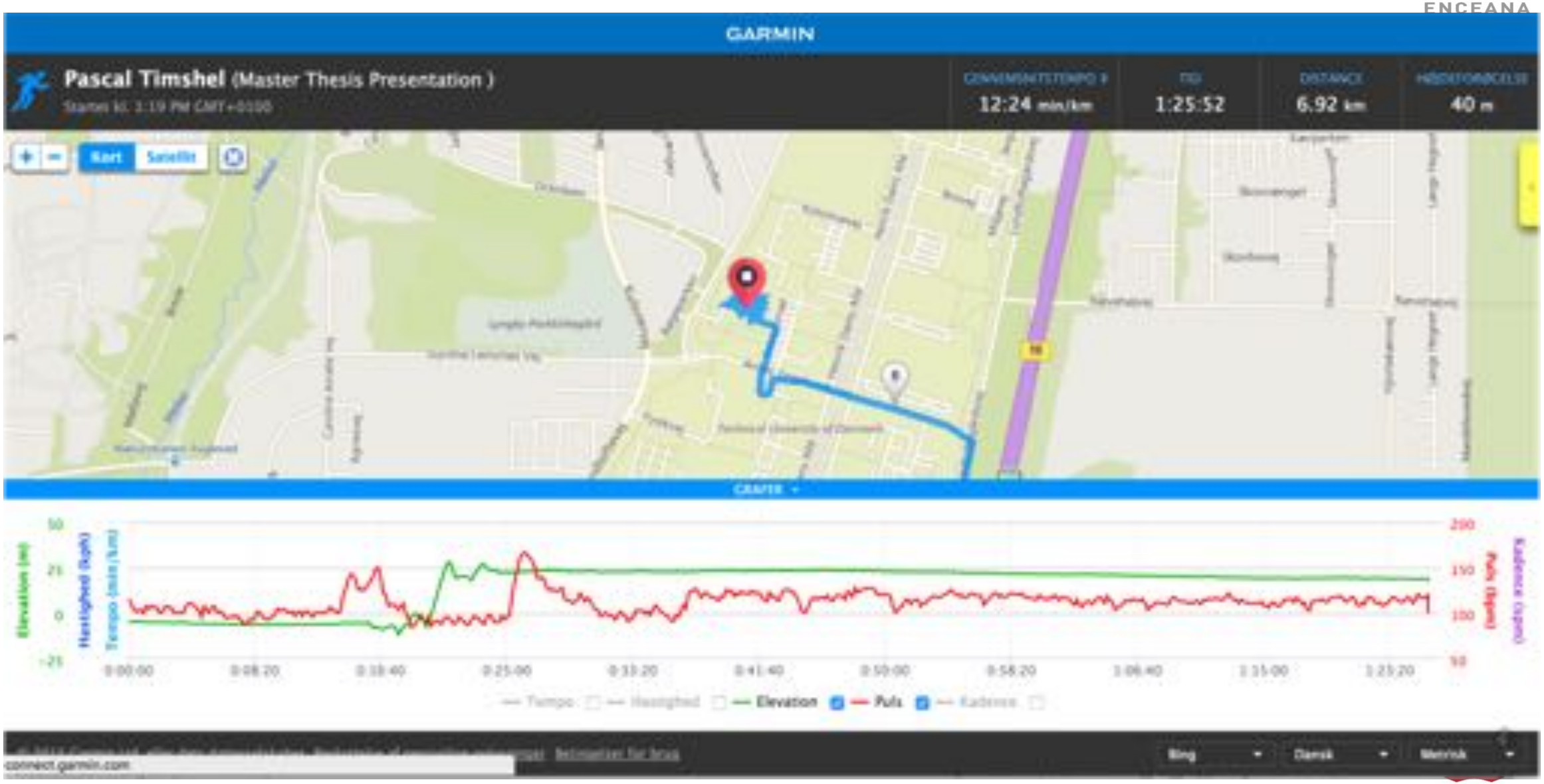
Thesis writing in numbers – a timeline

Your life before a deadline – or lack thereof

Thesis writing in numbers
Your life before a deadline



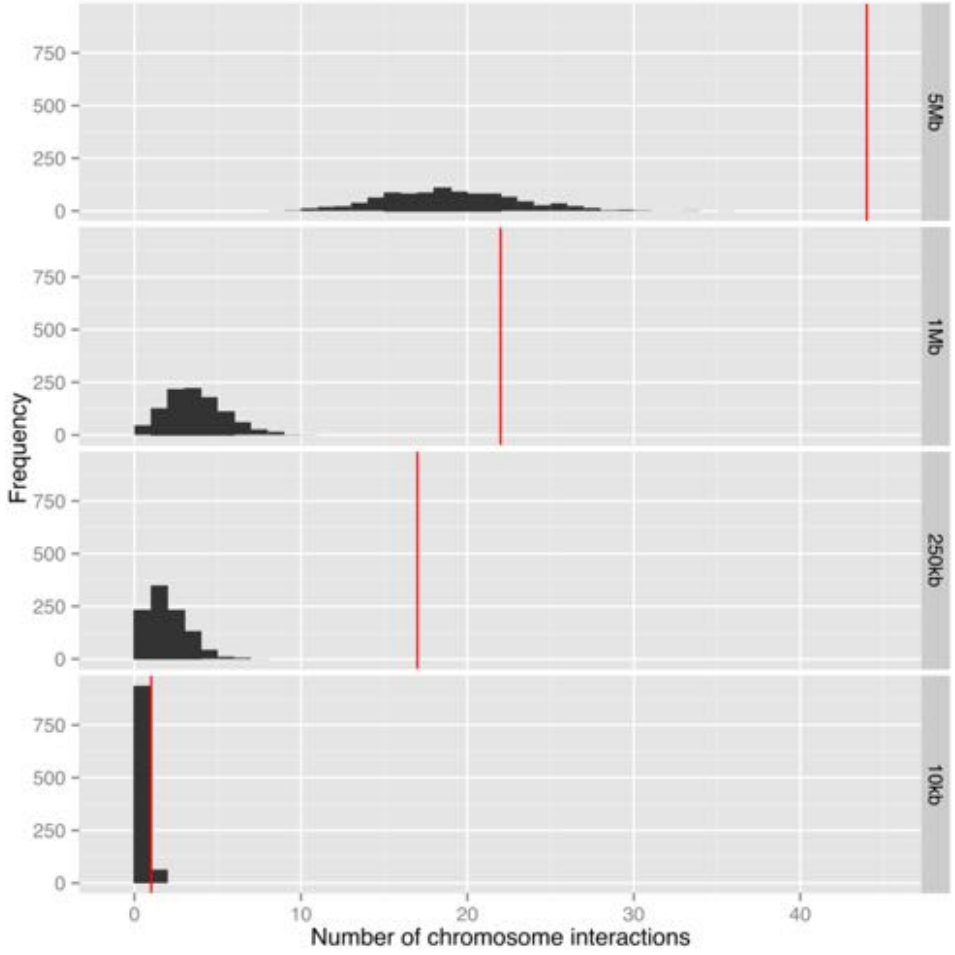

GARMIN
TM


GARMIN
Pascal Timshel (Master Thesis Presentation)
12:24 min/km
1:25:52
6.92 km
40 m
Kart
Satellit
Elevation
Puls
connect.garmin.com
Dansk
Metrisk

Appendix

Background: Hemani et al., 2014, *Nature*

- Models for biological interaction were examined
- Enrichment of epistatic SNP-pairs within 5 Mb of spatial interacting chromosomal regions.
- 44 out of 501 SNP-pairs co-locate ($P_{emperical}<1.8 \times 10^{-10}$)



Epistasis table

SNP1	rsID	MAF1	rsID	MAF2	SNP2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
8:18885665	rs10104492	0.28	rs12451549	0.10	17:46875212	1.72E-11	0.1731	3	CSHL1	17:59341967	hic_1_4837	hic_1
2:30967126	rs1862959	0.20	rs10503890	0.18	11:21075311	2.24E-13	0.18653	3	APOD	3:196777220	hic_1_1089	hic_1

Table B.4: Top confident **spatial epistastic SNP-probe pairs** identified in this study. That is, these SNP-pairs map to psychical interacting genomic loci (in hIMR90 cells). The SNP-probe pairs were selected using the criteria min GCC ≥ 3 (these two SNP-probe pairs appear on the right hand side of Table B.1). See Table B.3 for a description of the table headers.

SNP1	rsID	MAF1	SNP2	rsID	MAF2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
4:104098247	rs3074938	0.46	3:163183258	rs4299495	0.24	2.57E-11	0.066	10	LOC727752	19:51651	null_86_7175	null_86
3:57250044	rs9840663	0.32	22:47915370	rs80567	0.35	1.13E-11	0.086	12	ATP1F1	1:28436788	null_150_7035	null_150
9:81976039	rs7045654	0.32	16:12633764	rs7194011	0.42	2.64E-11	-0.078	19	SLC46A3	13:28172532	null_198_4075	null_198
1:176775553	rs6700266	0.35	7:82563554	rs6947662	0.4	3.10E-11	0.067	15	P1GR	1:205168794	null_420_1555	null_420
5:69859603	rs10036399	0.29	15:48836816	rs12911143	0.37	1.87E-11	0.064	11	KIAA1147	7:141003222	null_435_7721	null_435
2:135143088	rs4964158	0.39	16:63565417	rs35226	0.31	7.17E-12	-0.050	13	ABCC11	16:46823387	null_928_6011	null_928

Table B.5: Top confident **epistastic SNP-probe pairs** identified in this study. The SNP-probe pairs were selected using the criteria min GCC ≥ 10. These SNP-probe pairs are the strongest evidence of statistical epistasis found in this thesis. None of these SNP-probe pairs were reported by (Hemani et al., 2014). Notice that only two of the SNP-pairs co-localize with the probe. All the SNP-pairs map to non-interacting genomic loci ("null samples", see Figure 6.2). See Table B.3 for a description of the table headers.

Epistasis table

SNP1	rsID	MAF1	SNP2	rsID	MAF2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.63E-23	-0.5636	0	ITK	5:156614448	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.46E-26	0.5306	0	CREBL2	7:137210715	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	8.02E-18	0.5514	0	FAIM3	1:200144550	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.87E-31	0.5051	0	EGR2	10:64242090	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	7.56E-25	0.5474	0	LRMP	12:25152225	null_357_4752	null_357
5:152896308	**rs10041179**	**0.07**	**6:80265632**	**rs9294148**	**0.07**	**2.00E-110**	**1.2144**	**0**	**KIAA0672**	**17:12835495**	**null_357_4752**	**null_357**
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	8.45E-17	0.3753	0	KIAA1407	3:115160080	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	6.25E-13	0.4408	0	VIL2	6:159106859	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.63E-11	0.6132	0	FCRL2	1:155982443	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.35E-13	0.3674	0	SIAH1	16:46052618	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	4.57E-18	0.4974	0	COL9A2	1:40538963	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	7.09E-36	0.5992	0	FMOD	1:201576664	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.77E-30	0.5593	0	GPT2	16:45522597	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.60E-18	0.4203	0	STARD7	2:96214684	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.01E-25	0.6053	0	FCRL5	1:155750161	null_357_4752	null_357
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	9.17E-52	0.2780	1	ASS1	9:132310134	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.23E-14	0.1545	1	RHOHTB1	10:62373868	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	8.18E-21	0.1586	1	DZIP3	3:109896180	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	4.13E-15	0.1347	1	FAM60A2P	8:7104545	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.73E-24	0.2147	1	GLP1R	6:39163079	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.40E-30	0.1971	1	TXNRD2	22:18299943	null_339_6338	null_339
2:218046715	**rs1478590**	**0.11**	**14:69666638**	**rs10132344**	**0.11**	**1.29E-39**	**0.4352**	**1**	**CSHL1**	**17:58041967**	**null_359_6338**	**null_339**
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	8.16E-17	0.1293	1	NCOR2	12:123077955	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	4.96E-15	0.1320	1	NUDT4P1	1:143848976	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.52E-27	0.1737	1	C1orf200	1:9637055	null_339_6338	null_339
3:25307427	**rs322673**	**0.26**	**13:79204571**	**rs2329119**	**0.42**	**4.07E-11**	**-0.1306**	**9**	**DUSP1**	**5:172128082**	**null_974_3900**	**null_974**

Table B.3: Examples of epistastic effects with different levels of confidence (measured by minimum GCC). The SNP-probe pairs are sorted by their minimum GCC. The table presents data for three exemplary SNP-pairs. SNP-probe pairs marked in bold are plotted in Figure B.6 on page 87. This table serves to illustrate some of the spurious epistatic effects (minGCC < 3) underlying the hIMR90 spatial epistasis enrichment. `P-value`: 1-*df* test for the significance of the genetic interaction term; `Beta`: effect size estimate of the genetic interaction term; `Gene`: the gene-probe mapping listed in the Illumina manifest file; `Probe`: chromosomal coordinates of the probe; EID: Experiment Identifier; EIID: Experiment Interaction Identifier.

PEER Factor Correlation