Reverse engineering of genotype-phenotype map using individual genetic variation

Vitaly V. Gursky[1,2], Konstantin N. Kozlov[2], Ivan V. Kulakovskiy[3-5], Sergey V. Nuzhdin[6], Maria G. Samsonova[2]

[1] Ioffe Institute, St. Petersburg, Russia

[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

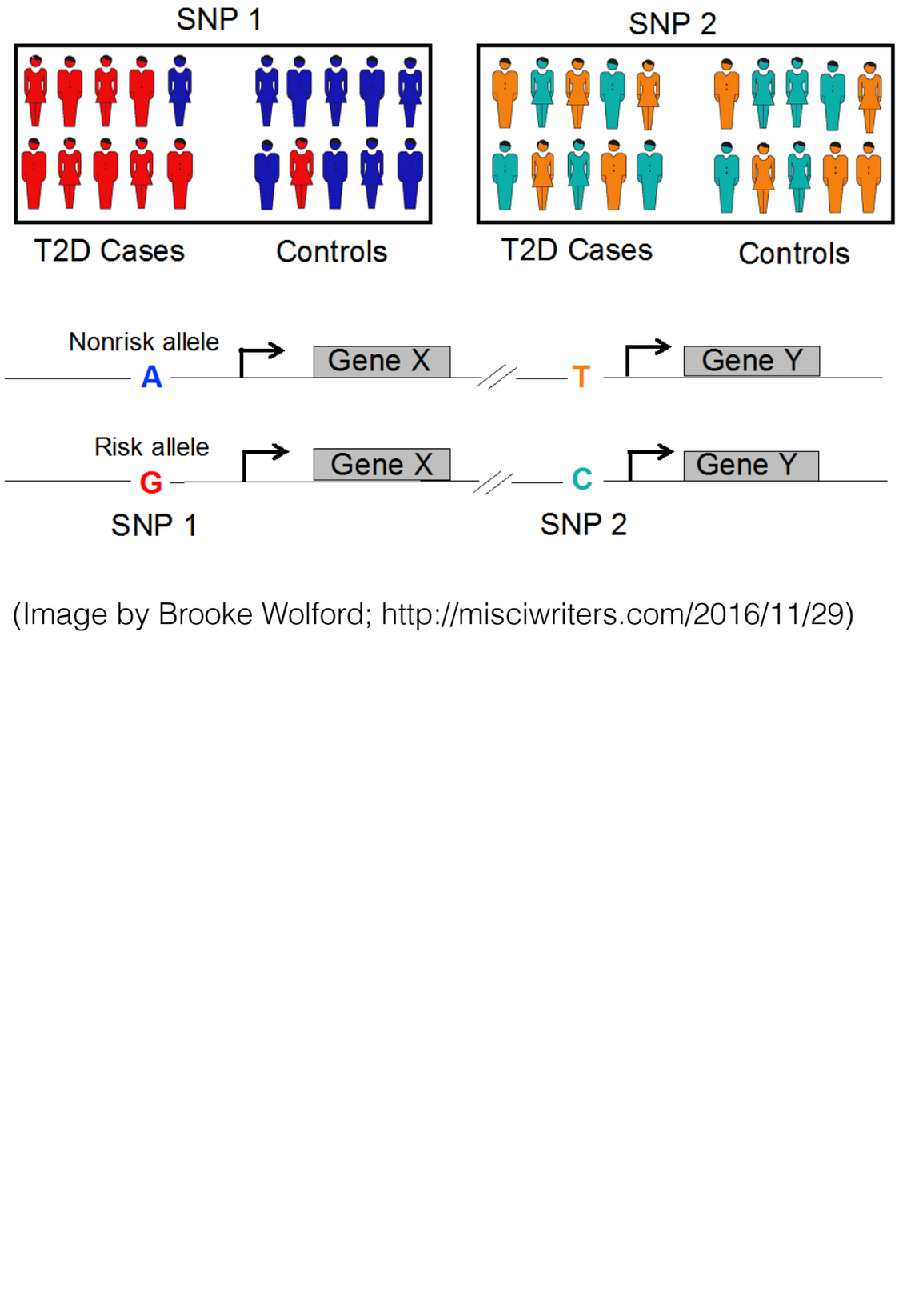
[3] Engelhardt Institute of Molecular Biology, Moscow, Russia

[4] Vavilov Institute of General Genetics, Moscow, Russia

[5] Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, Moscow, Russia

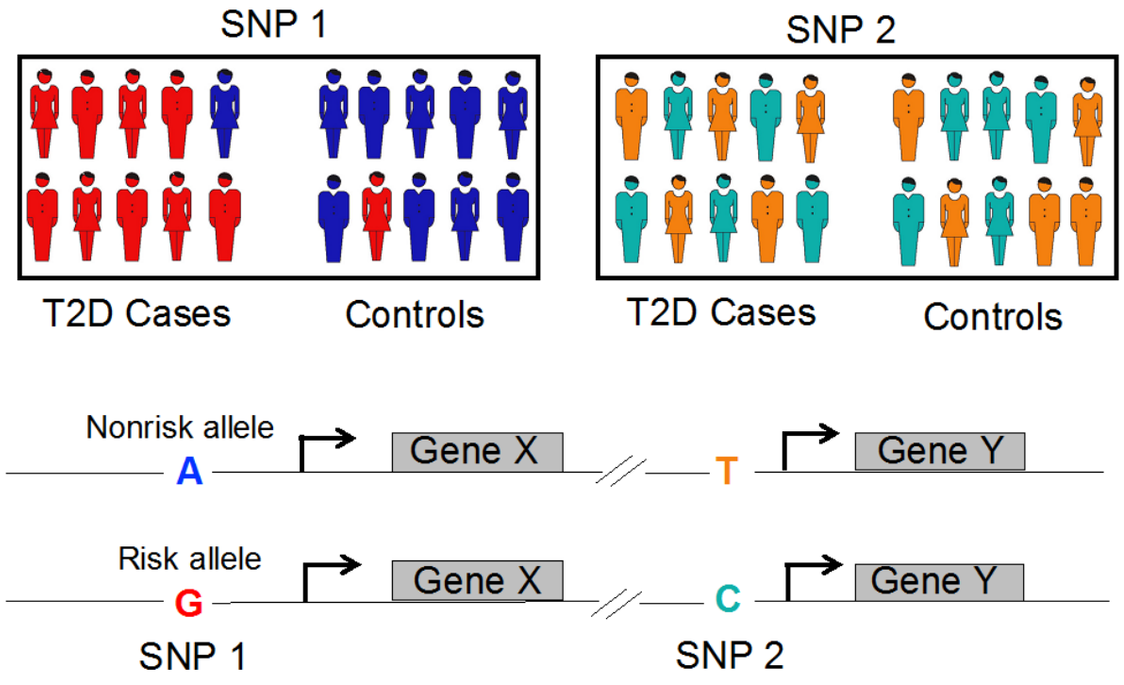
[6] University of Southern California, Los Angeles, U.S.A.

Genotype-phenotype maps



(Image by Brooke Wolford; http://misciwriters.com/2016/11/29)

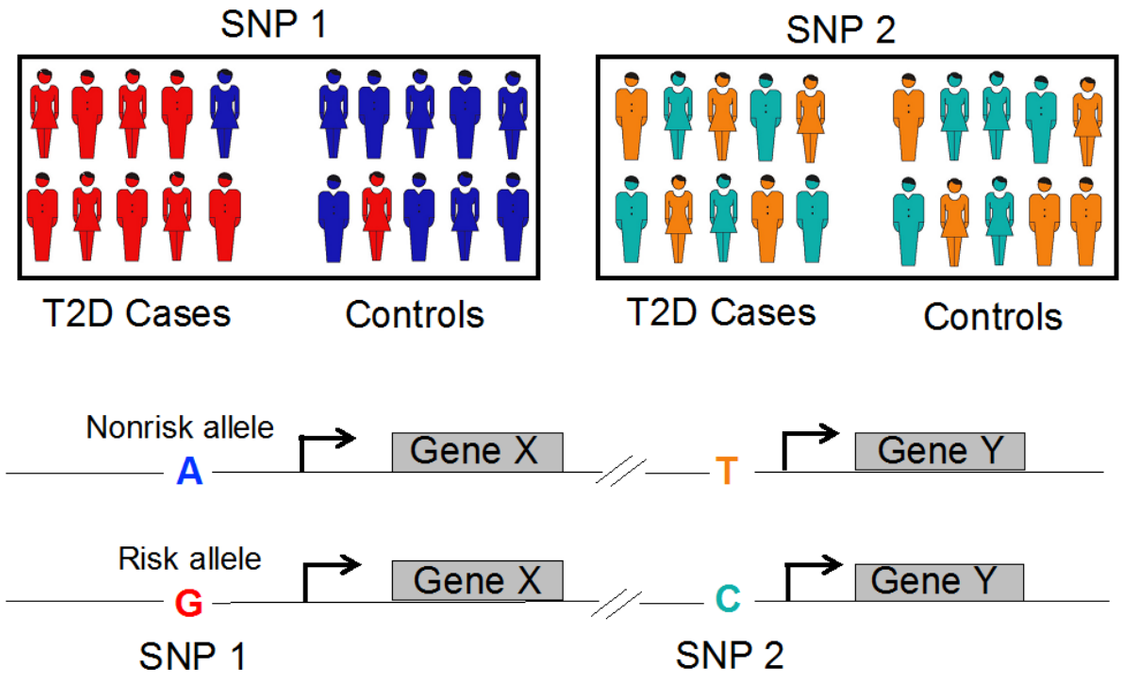
Genotype-phenotype maps



Genome
Wide
Association
Studies

(Image by Brooke Wolford; http://misciwriters.com/2016/11/29)

Genotype-phenotype maps



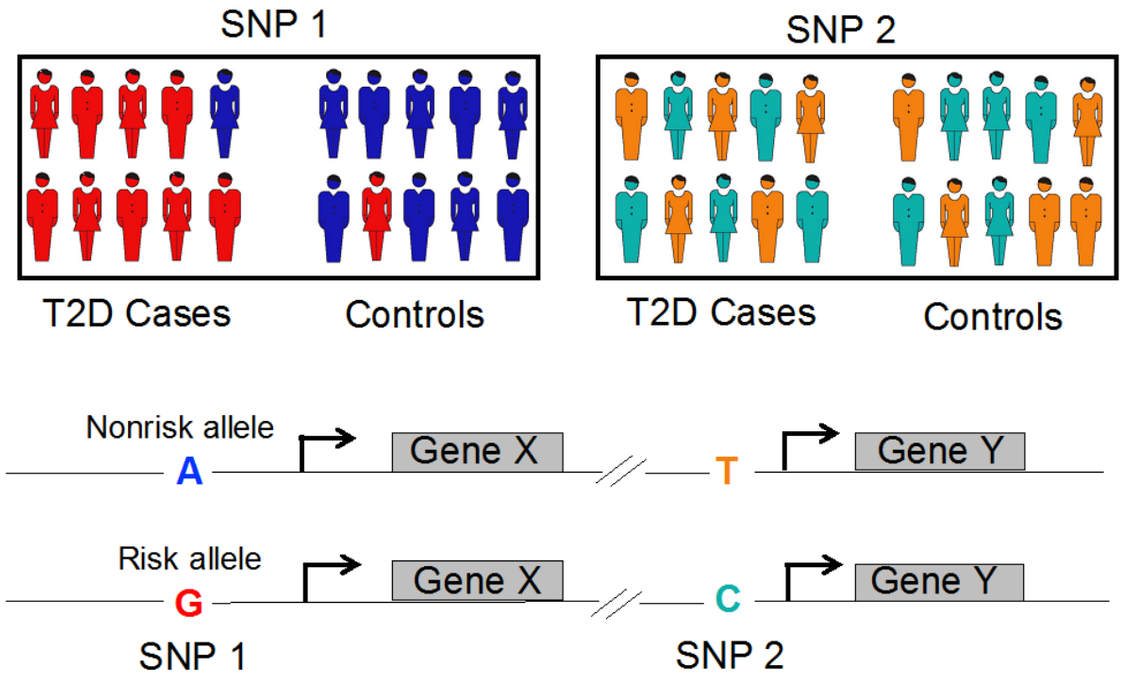
Genome
Wide
Association
Studies

(Image by Brooke Wolford; http://misciwriters.com/2016/11/29)

Alternative approach:

Modeling gene regulatory networks on an ‘average’ genome and **predicting** how genetic variation effects **molecular phenotype** (gene expression patterns)

Genotype-phenotype maps



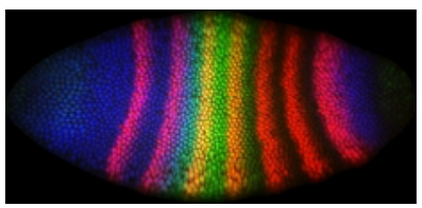
Genome
Wide
Association
Studies

(Image by Brooke Wolford; http://misciwriters.com/2016/11/29)

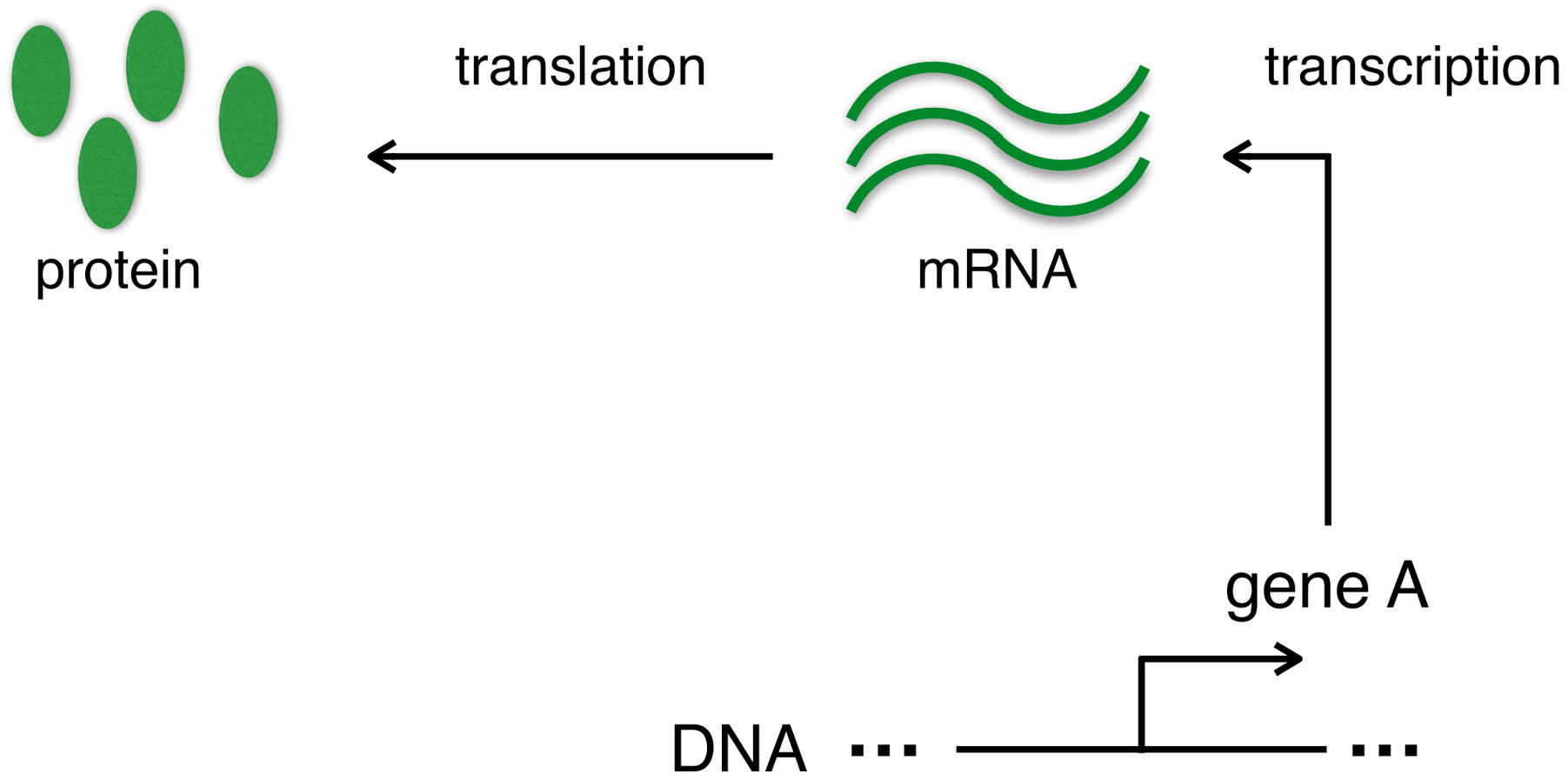
Alternative approach:

Modeling gene regulatory networks on an ‘average’ genome and **predicting** how genetic variation effects **molecular phenotype** (gene expression patterns)

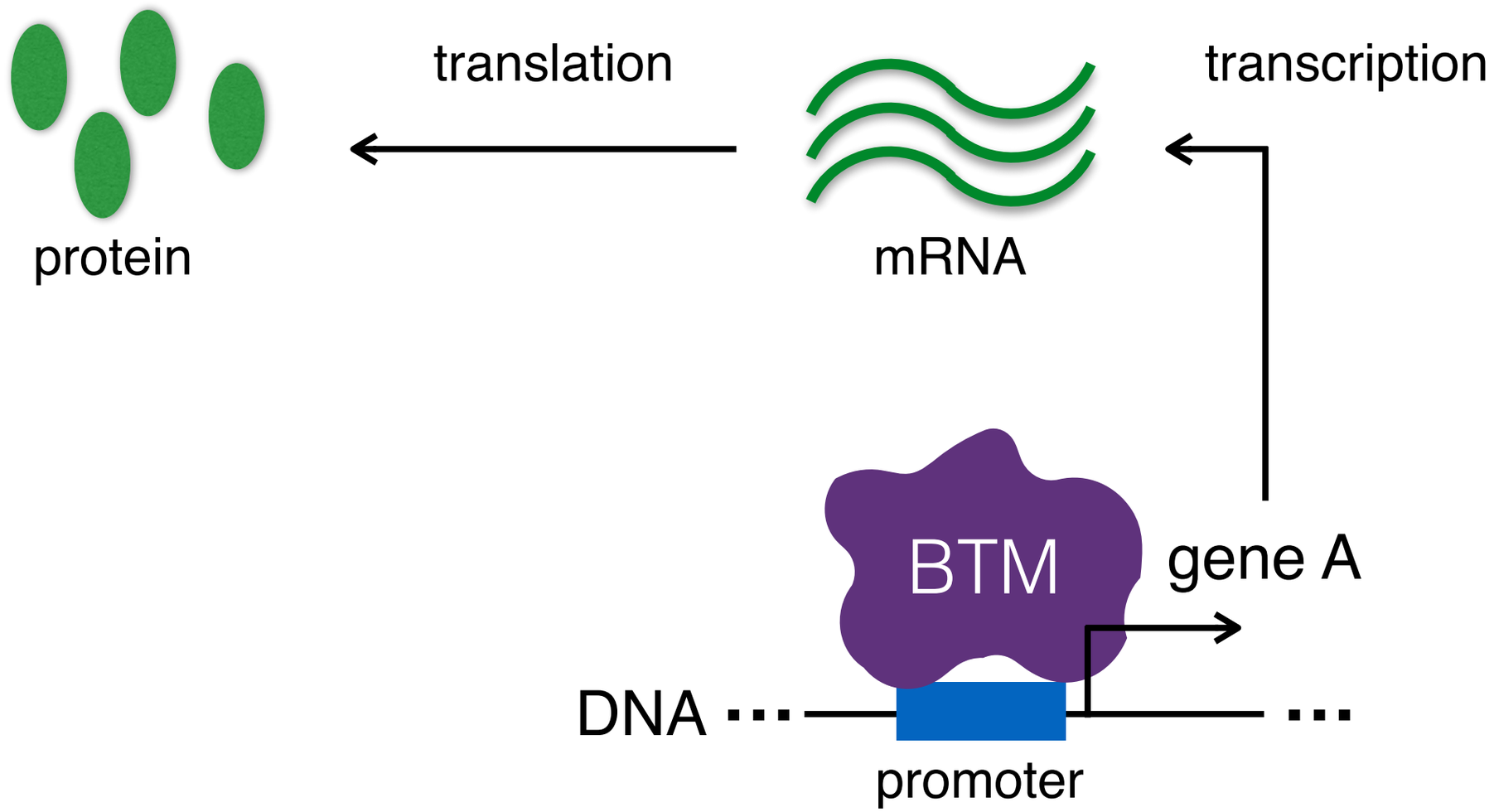
Model organisms and processes are helpful as a training ground



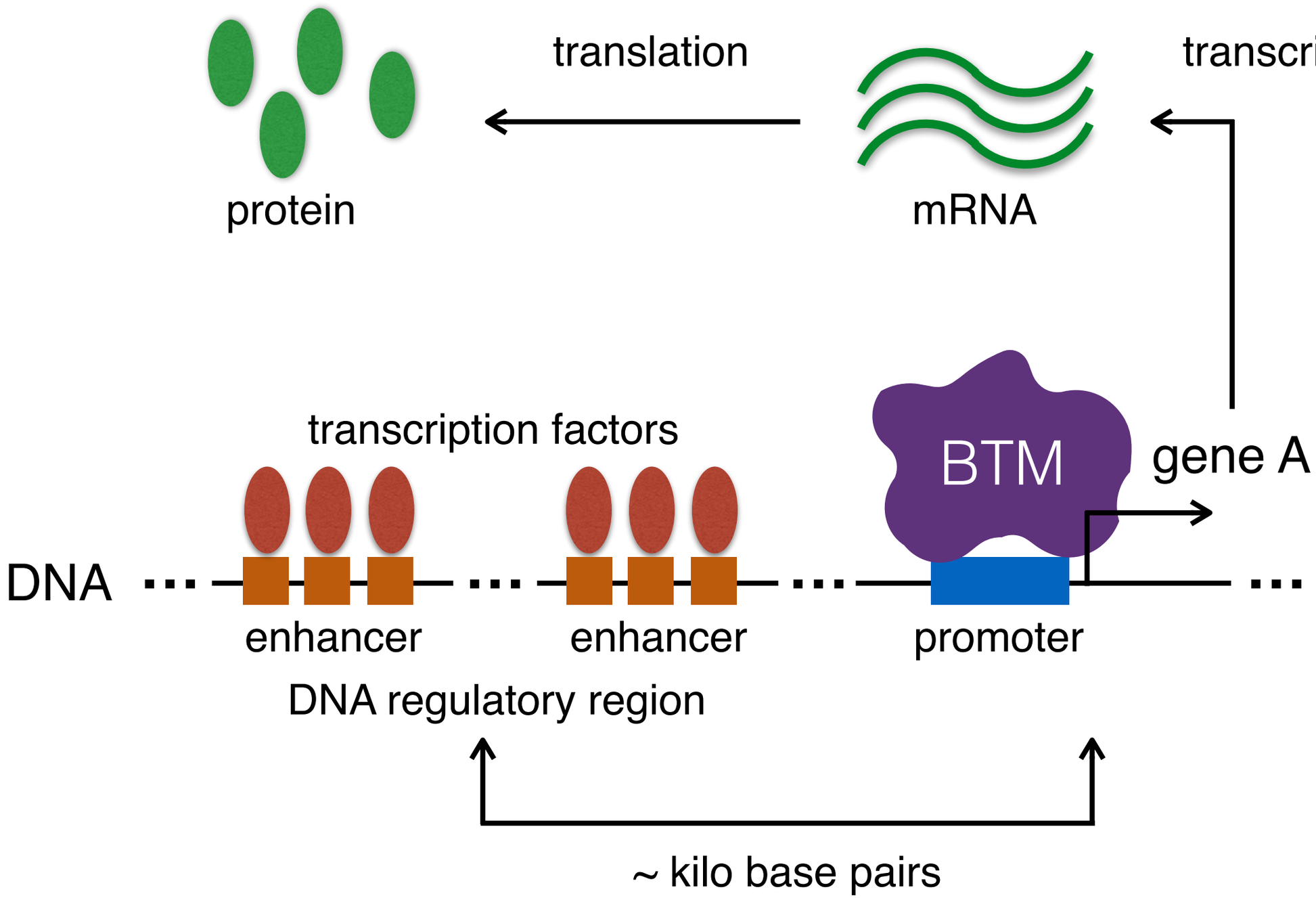
Gene expression



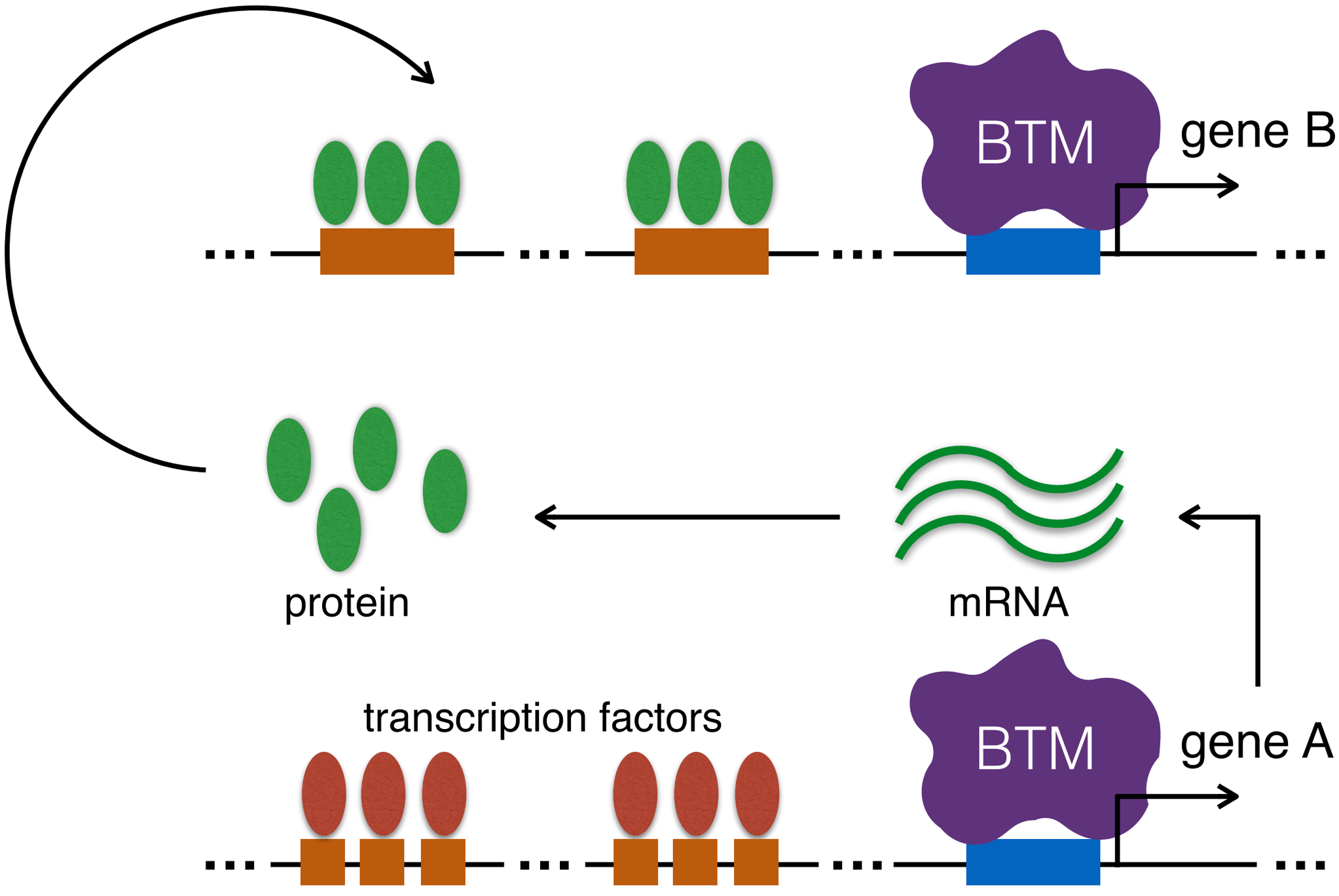
Gene expression



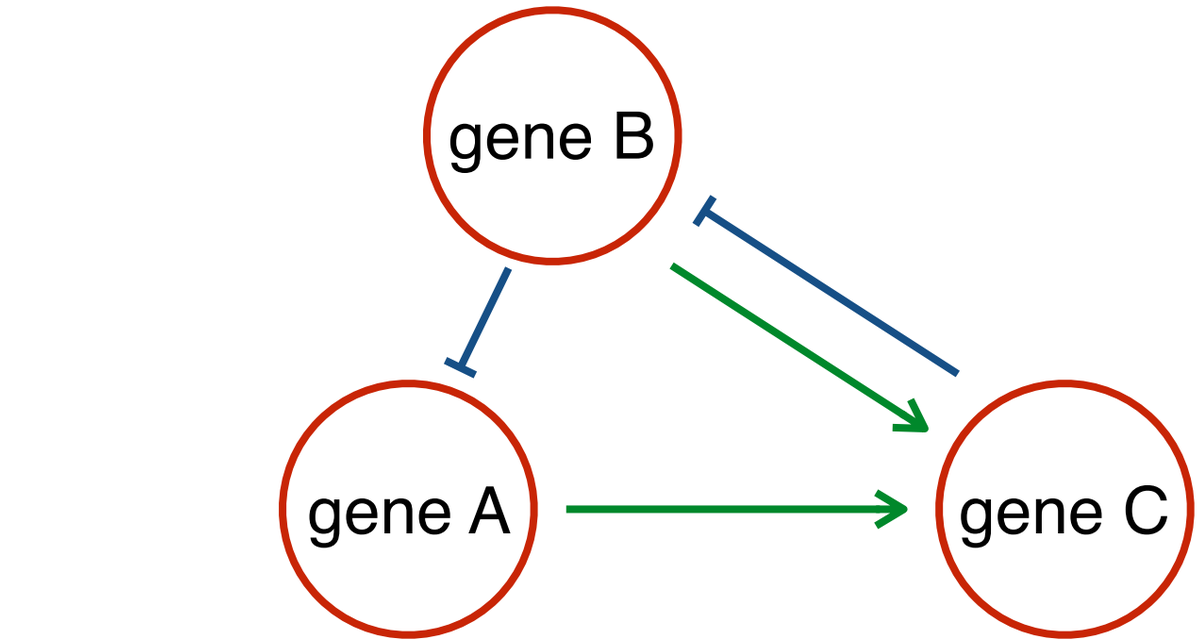
Gene expression



Transcriptional regulation



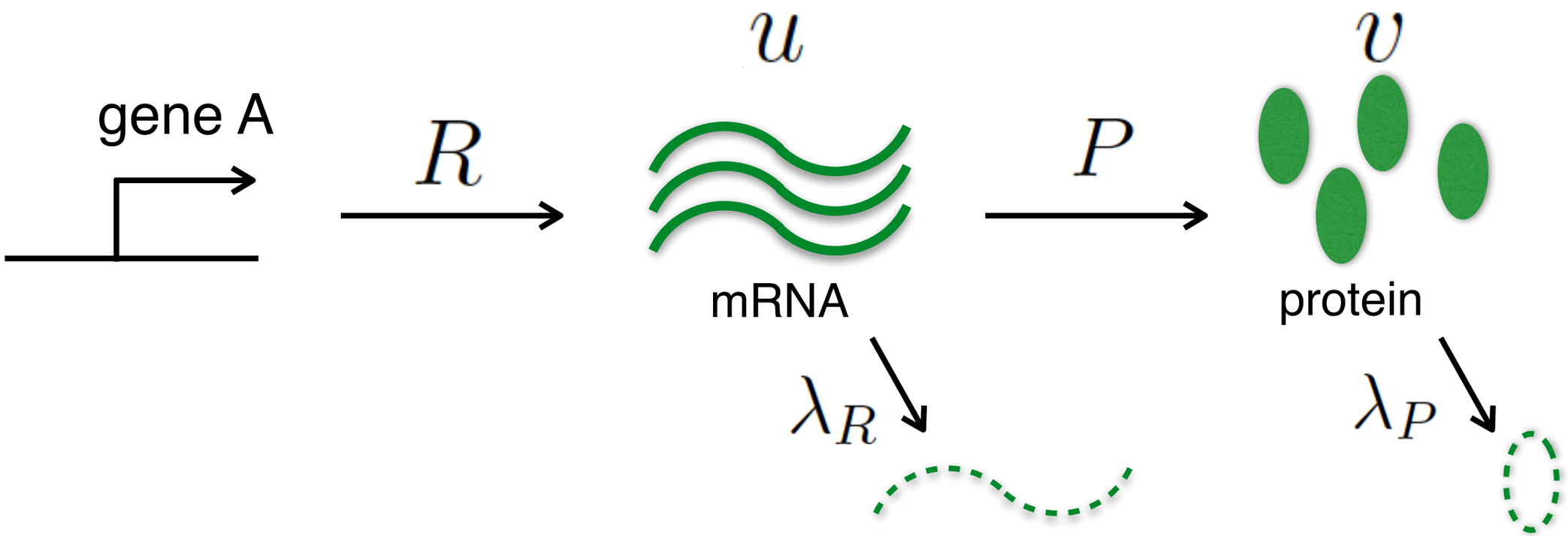
Gene regulatory network



transcriptional activation

transcriptional repression

Reaction-diffusion equations



$$\frac{\partial u}{\partial t} = R\left(x,t\right) - \lambda_R u(x,t) + D_R \Delta u(x,t)$$

$$\frac{\partial v}{\partial t} = P\left(x,t\right) - \lambda_P v(x,t) + D_P \Delta v(x,t)$$

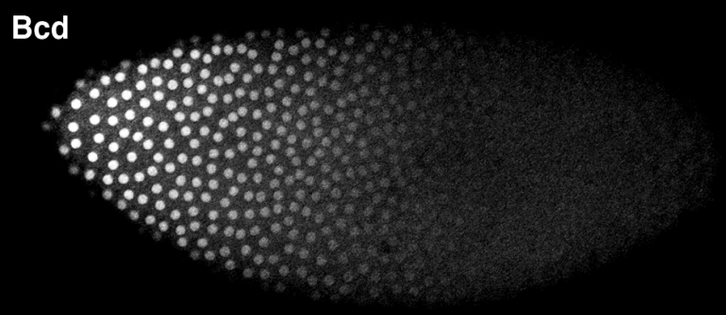
Segmentation genes in *Drosophila* development

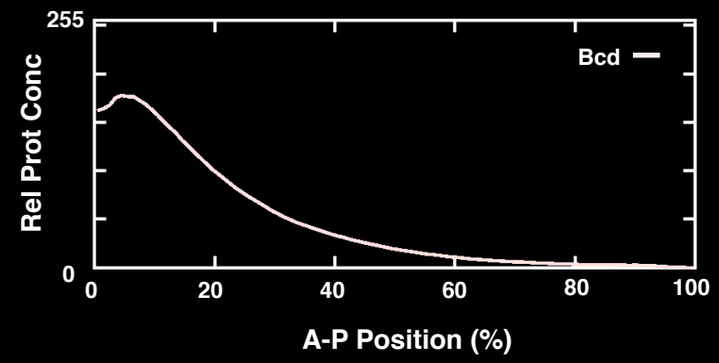
Maternal coordinate genes

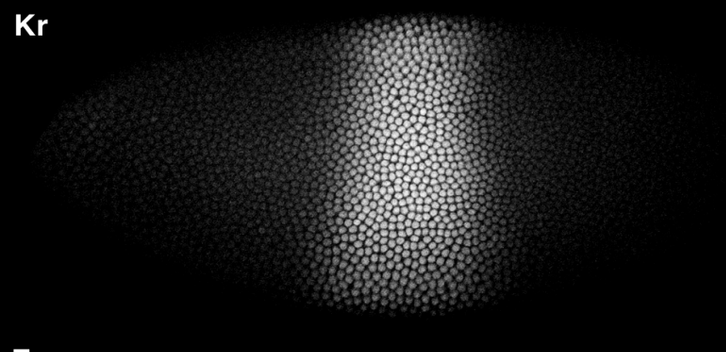
Gap genes

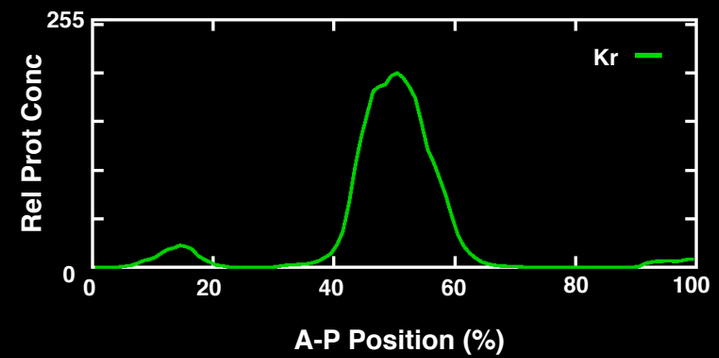
Pair-rule genes

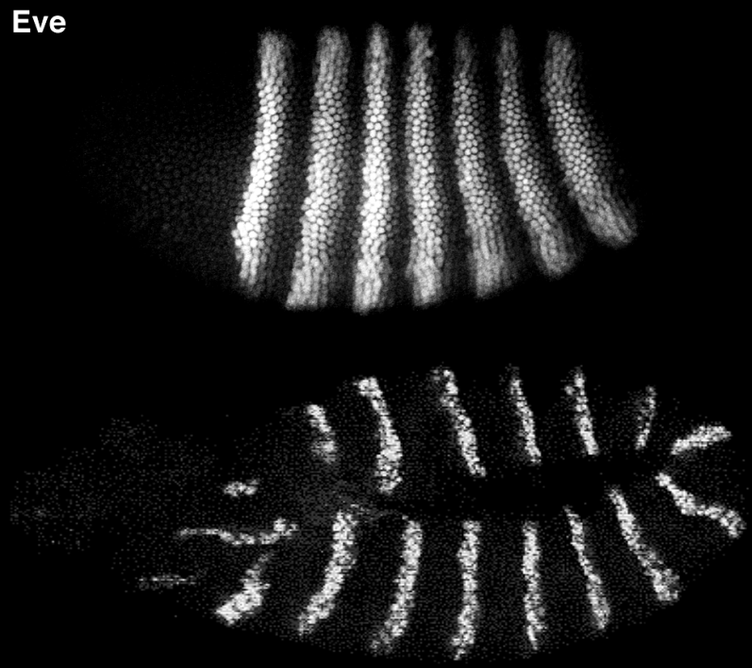
Segment polarity genes

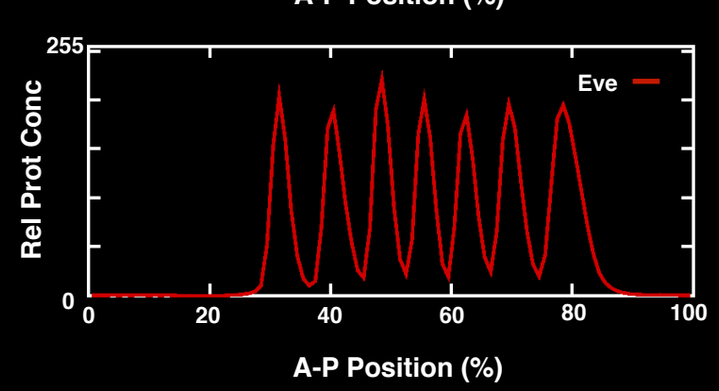






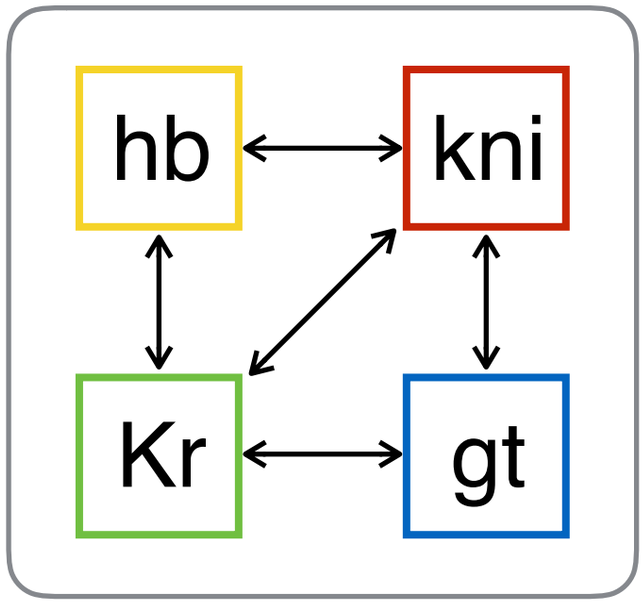


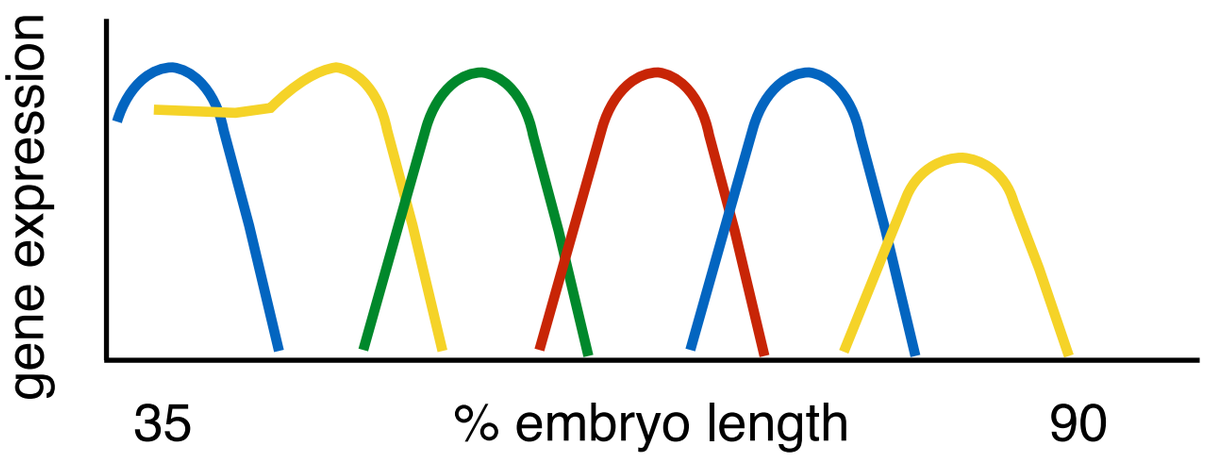




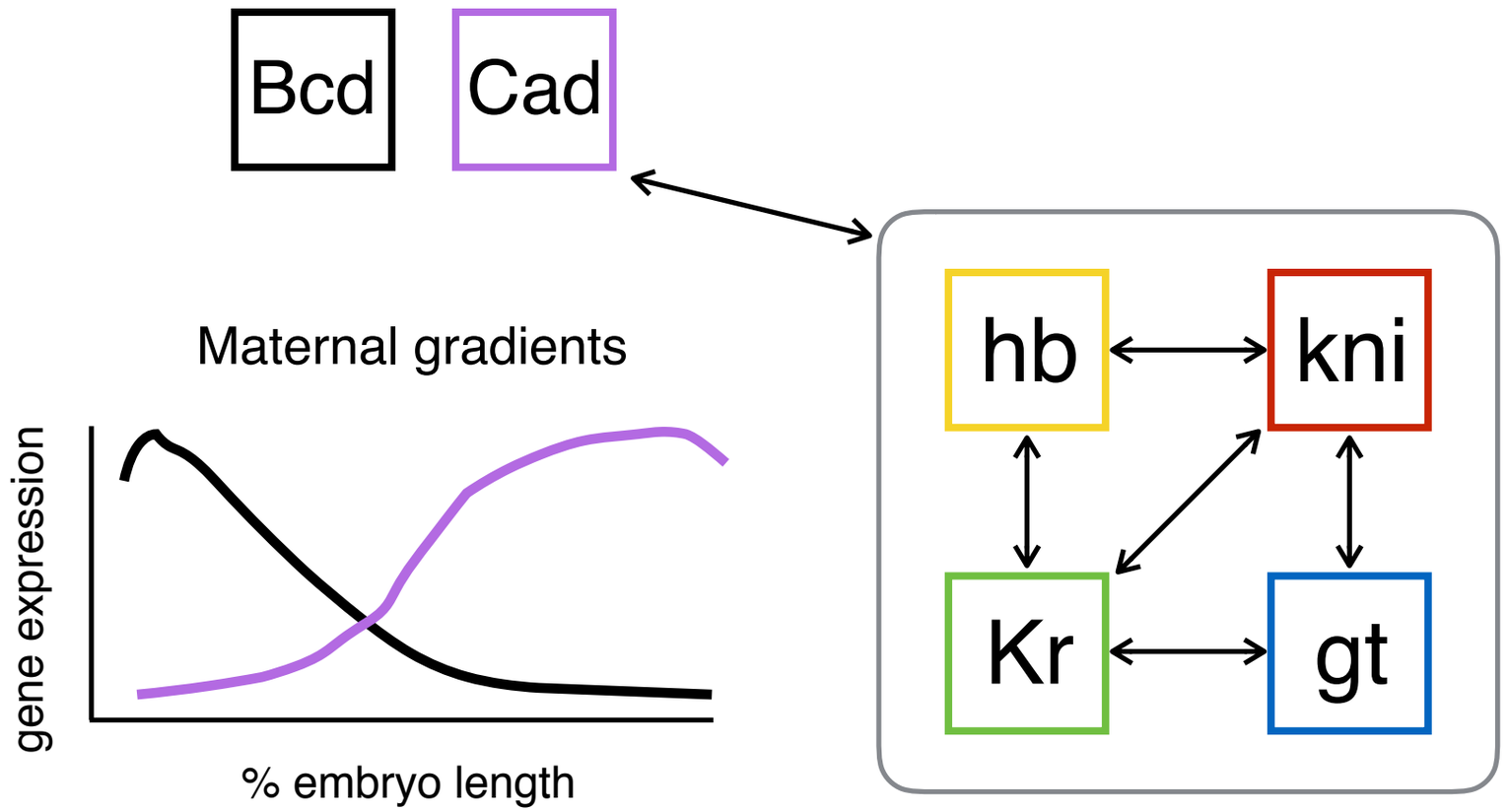
mRNA data: SuperFly
Protein data: FlyEx

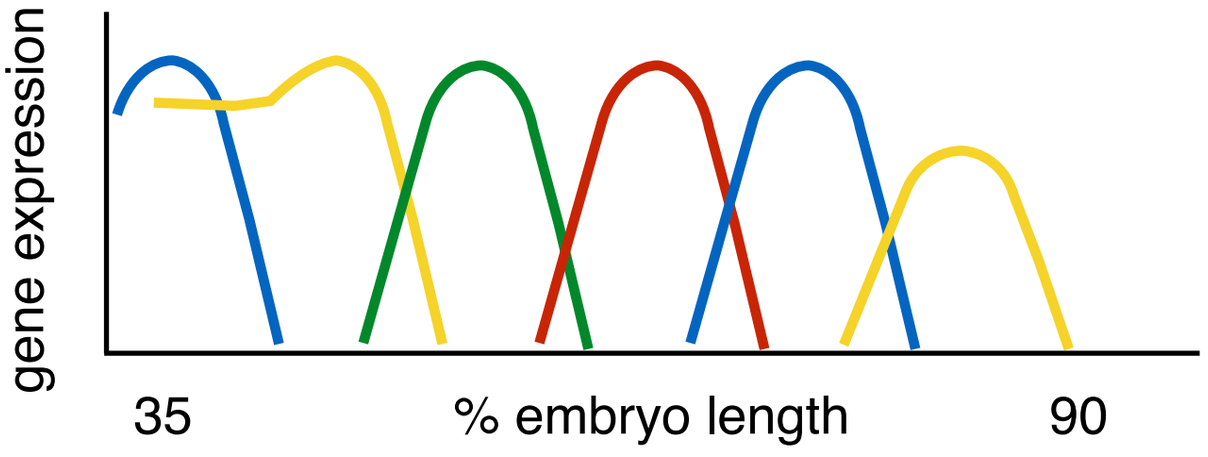
Gap gene network



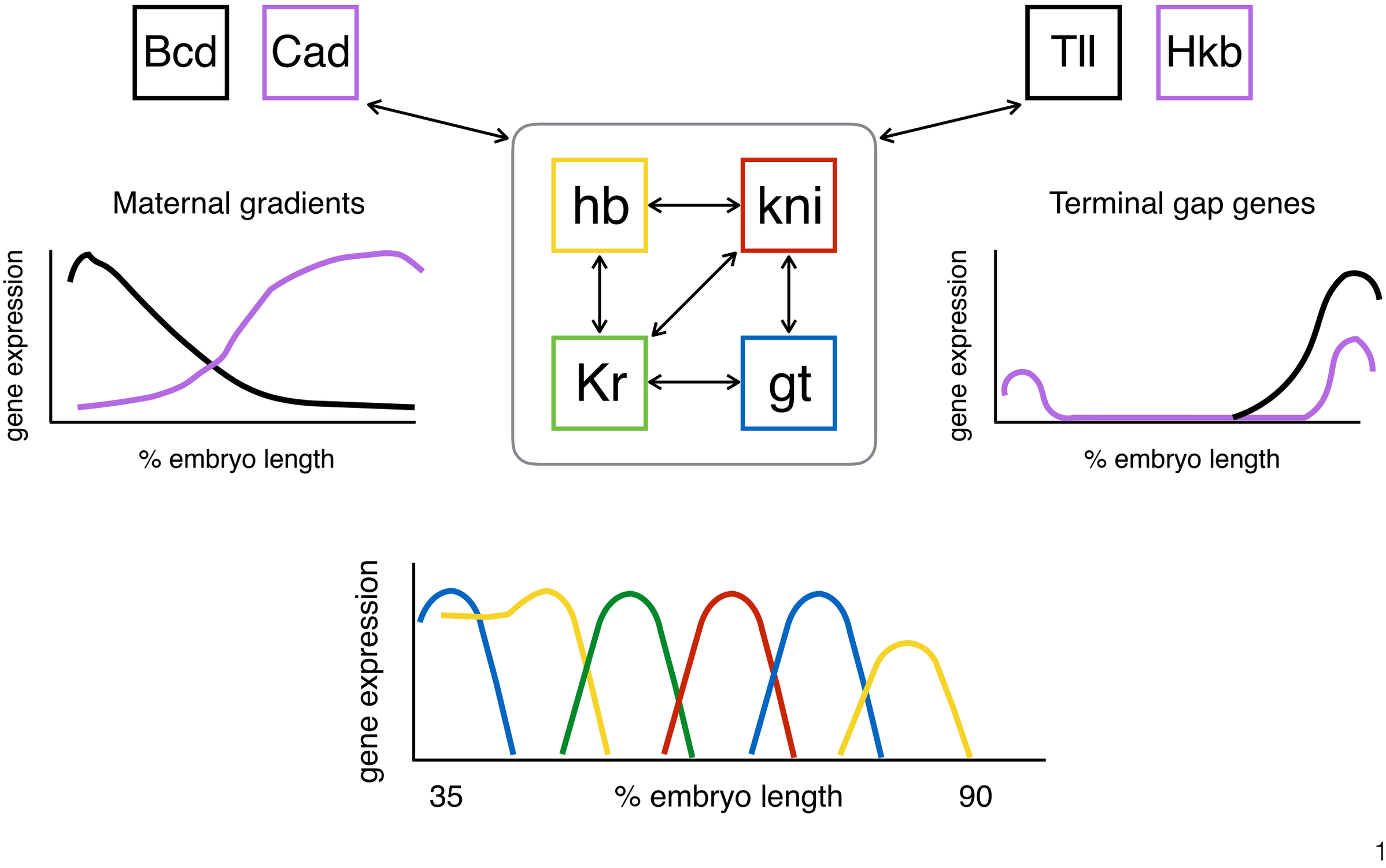


Gap gene network



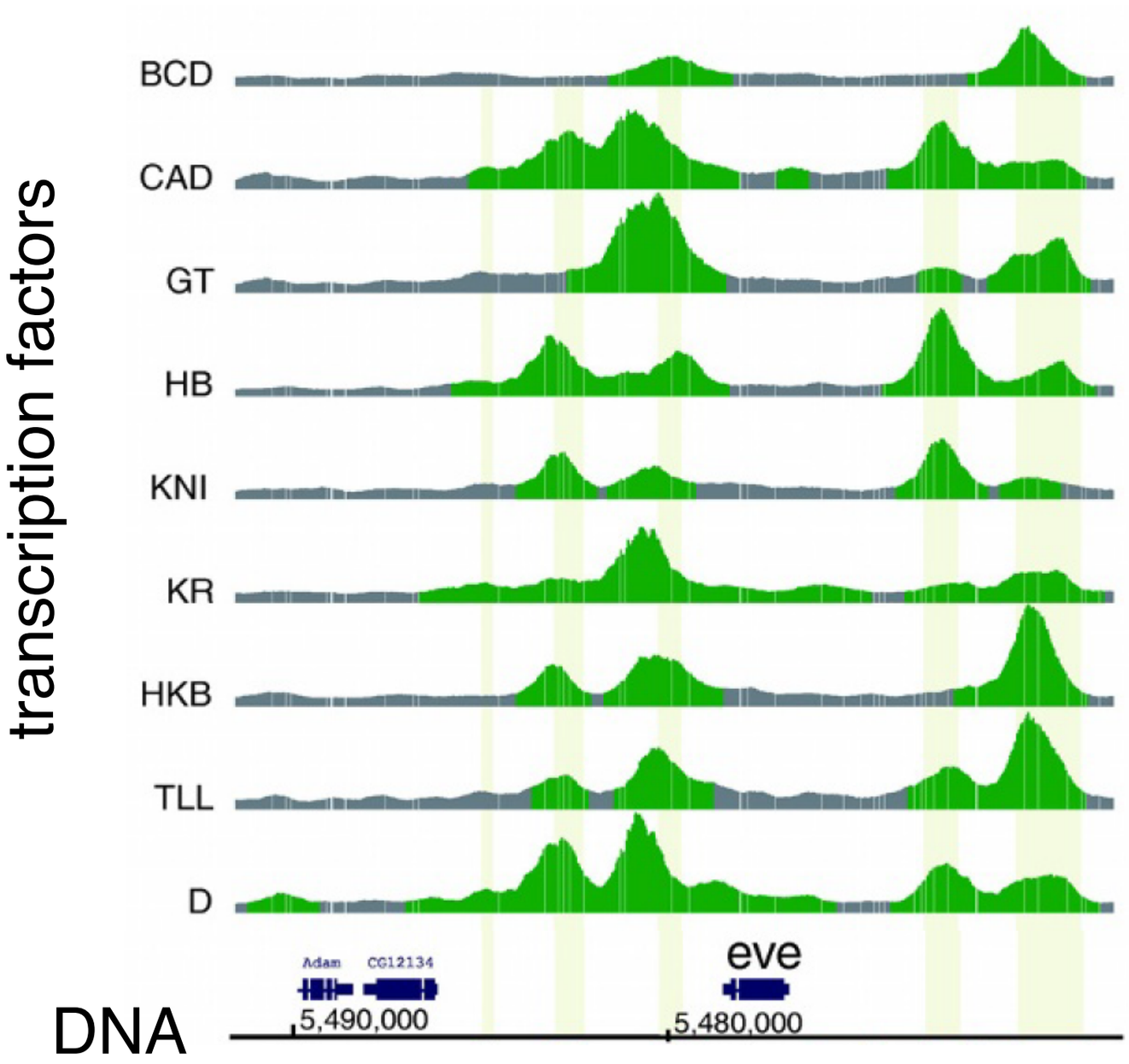


Gap gene network



Information about binding sites

Probability of TF binding to DNA according to ChIP-Seq data for *D. melanogaster*:



Binding motif

Processing data:

TCCCTGAACGG
TCCGAGAACCT
TTGCTCCGCA–
TTCCTGAGCTG
TTCGTAAGGAG

Aligned sequences

↓

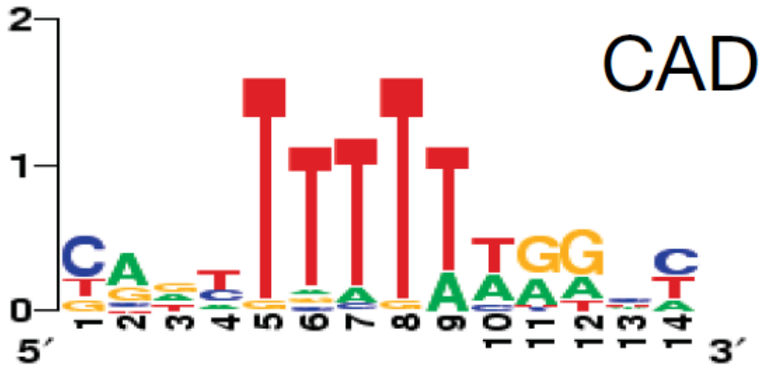
Positional weight matrix (PWM)

A: 00001142020
C: 02430110410
G: 00120303113
T: 53004000011

↓

TYCSTGARCNG
(IUPAC format)

Consensus site sequence
(motif)



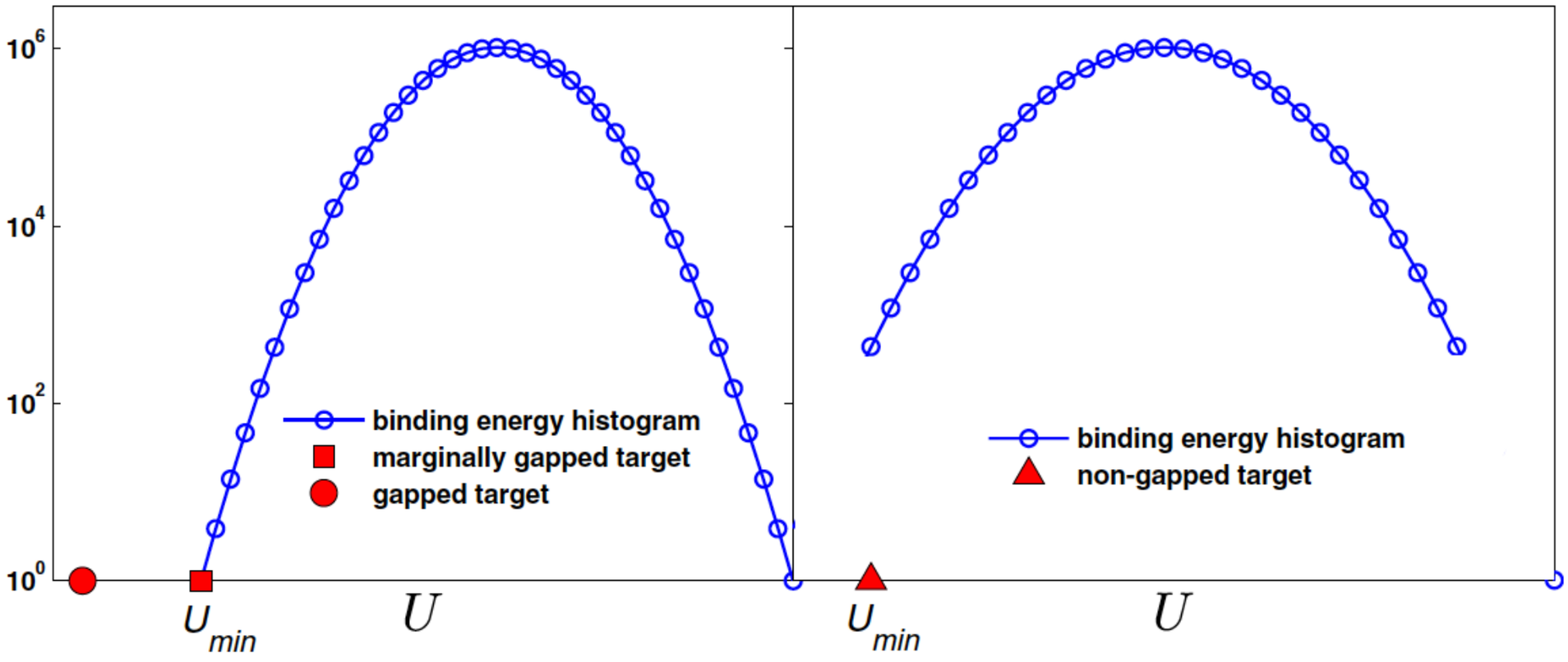
How many binding sites can we expect?

Connection of **PWM-scores** with **binding energy**:

Berg and von Hippel PH (1987) J Mol Biol, 193;

Stormo and Fields (1998) Trends Biochem Sci 23.

Typical binding energy histograms (in log scale):



From Sheinman et al. (2012) Rep. Progr. Phys., 75(2)

Known enhancers are not always enough

nature genetics

Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster even skipped* gene

Hilde Janssens[1], Shuling Hou[2], Johannes Jaeger[1], Ah-Ram Kim[1], Ekaterina Myasnikova[3], David Sharp[4] & John Reinitz[1]

OPEN ACCESS Freely available online

PLOS | GENETICS

Rearrangements of 2.5 Kilobases of Noncoding DNA from the *Drosophila even-skipped* Locus Define Predictive Rules of Genomic *cis*-Regulatory Logic

Ah-Ram Kim[1,2], Carlos Martinez[1], John Ionides[3], Alexandre F. Ramos[4], Michael Z. Ludwig[1], Nobuo Ogawa[5], David H. Sharp[6], John Reinitz[1,7]*

Open chromatin regions match binding patterns

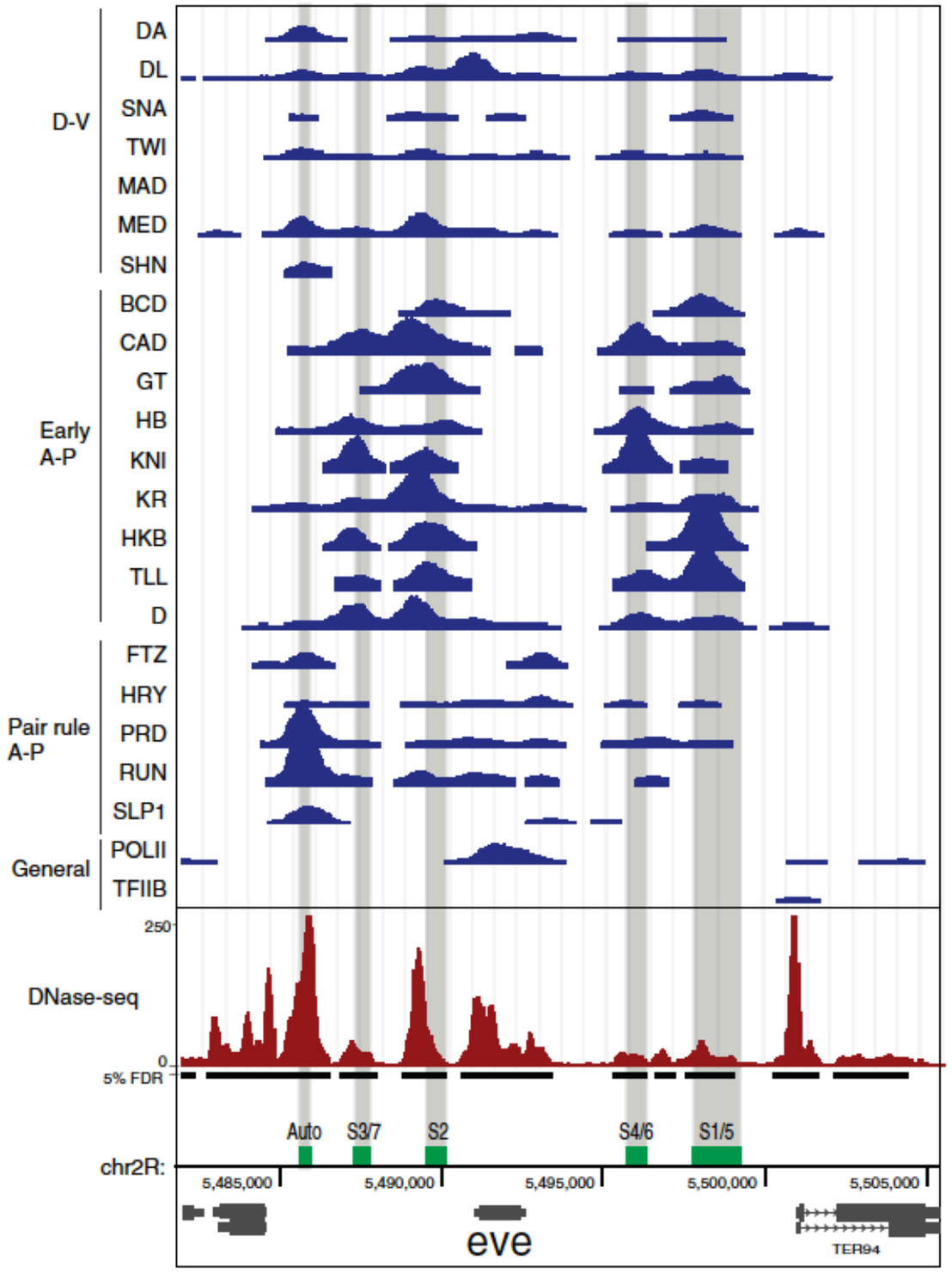


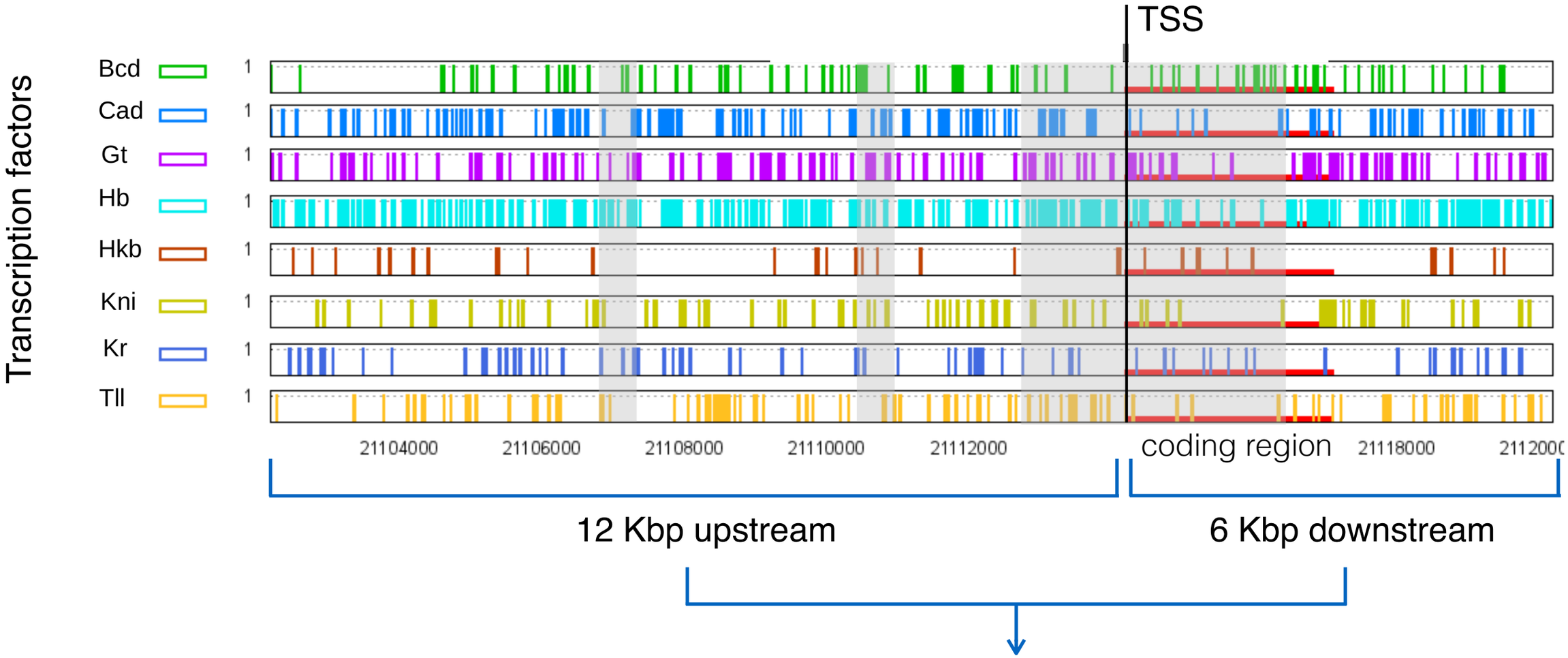
Figure 1 DNaseI accessibility and *in vivo* DNA binding by transcription factors across the *eve* locus.

From Li et al. (2011) Genome Biology, 12(4)

Analysis of gap gene regulatory regions

Improved D. melanogaster major **position matrix motifs by ChiPMunk** software (http://autosome.ru/iDMMPMM/) (Kulakovskiy et al. 2009, 2010).

Example of information obtained by the software for Kruppel's regulatory region:



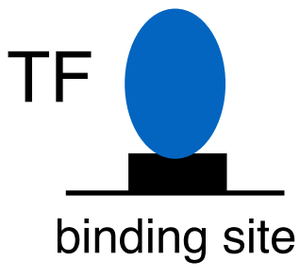
18 Kbp region comprising all **classic developmental enhancers** from the **REDfly** database.

Select binding sites from the **open chromatin regions** (according to DNaseI accessibility data).

Total number of sites: from 889 (28 per TF per target gene) to 1419 (44 per TF per target gene).

Molecular configurations via thermodynamic approach

Configurations:

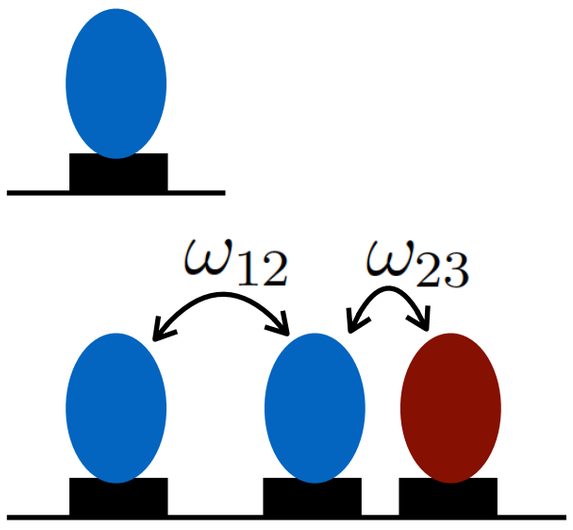


Statistical weights:

$$q_i = K v_i \exp\left(-\Delta G_i\right) = K v_i \exp\left(P(S_i) - P(S_{\max})\right)$$

Molecular configurations via thermodynamic approach

Configurations:



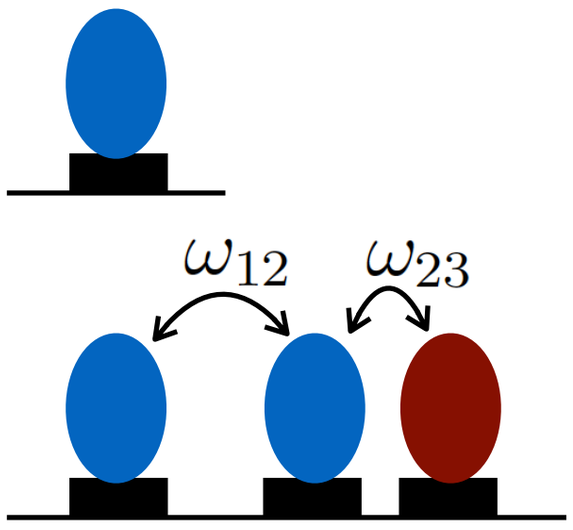
Statistical weights:

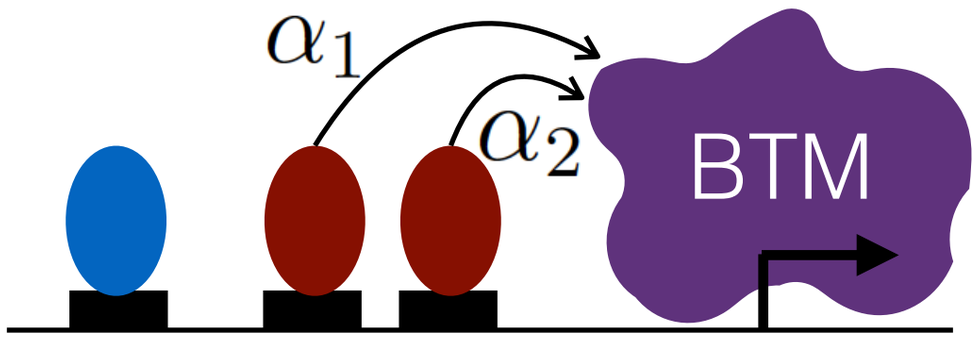
$$q_i = K v_i \exp\left(-\Delta G_i\right) = K v_i \exp\left(P(S_i) - P(S_{\max})\right)$$

$$W(\sigma) = \prod_i q_i \prod_j \omega_{ij}$$

Molecular configurations via thermodynamic approach

Configurations:





Statistical weights:

$$q_i = K v_i \exp\left(-\Delta G_i\right) = K v_i \exp\left(P(S_i) - P(S_{\max})\right)$$

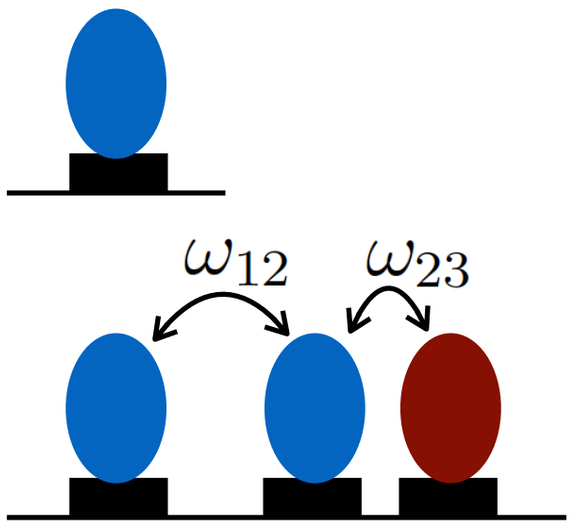
$$W(\sigma) = \prod_i q_i \prod_j \omega_{ij}$$

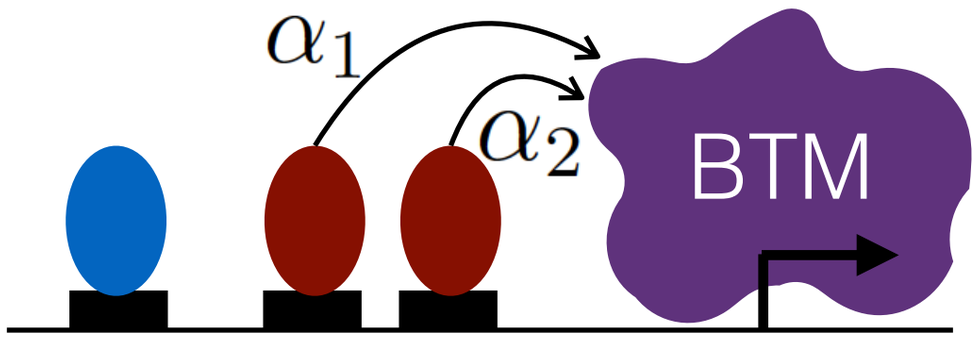
$$W(\sigma)Q(\sigma),\ Q(\sigma) = \prod_i \alpha_i$$

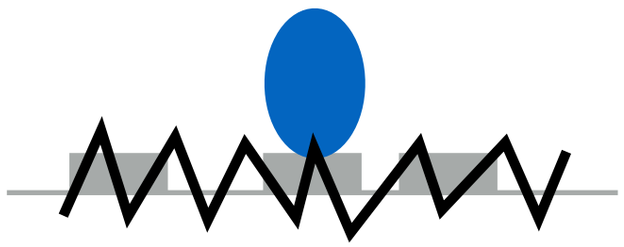
α_i — activation strengths.

Molecular configurations via thermodynamic approach

Configurations:







Statistical weights:

$$q_i = K v_i \exp\left(-\Delta G_i\right) = K v_i \exp\left(P(S_i) - P(S_{\max})\right)$$

$$W(\sigma) = \prod_i q_i \prod_j \omega_{ij}$$

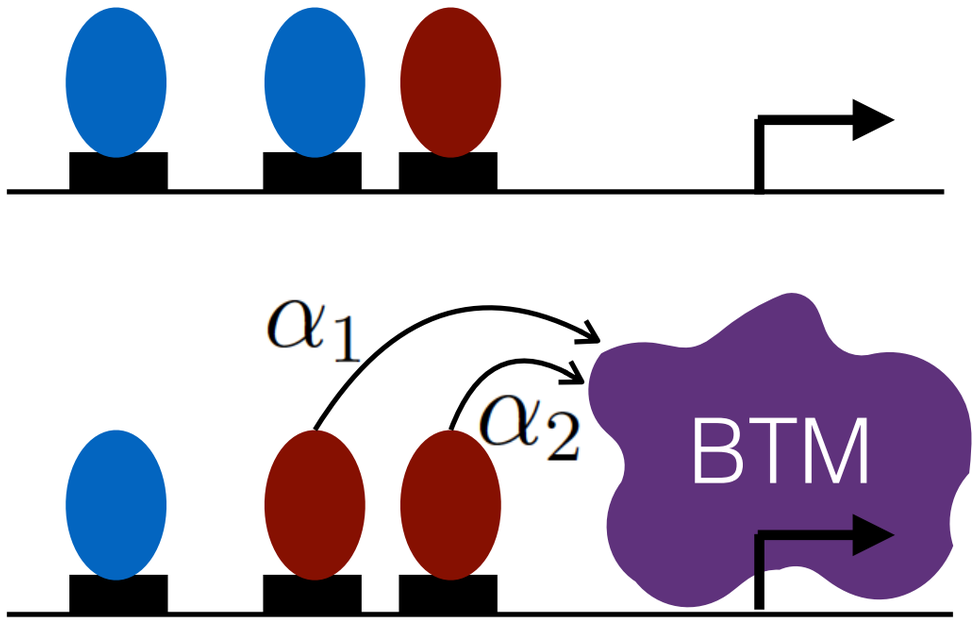
$$W(\sigma)Q(\sigma),\ Q(\sigma) = \prod_i \alpha_i$$

α_i — activation strengths.

Short-range repression: $q_i \to \beta q_i$

β — repression strengths.

Transcriptional activation via thermodynamic approach



$$W(\sigma) = \prod_i q_i \prod_j \omega_{ij}$$

$$W(\sigma)Q(\sigma),\ Q(\sigma) = \prod_i \alpha_i$$

α_i — activation strengths.

Probability of transcriptional activation of gene '*a*':

$$E_a = \frac{Z_{ON}}{Z_{ON} + Z_{OFF}} = \frac{\sum_\sigma W(\sigma)Q(\sigma)}{\sum_\sigma W(\sigma)Q(\sigma) + \sum_\sigma W(\sigma)}$$

He et al. PLoS Comput. Biol., 2010; Kozlov et al. BMC Genomics, 2014, 2015

Hybrid dynamical model for gap gene network

Reaction-diffusion equations for mRNA concentrations (u_a) and protein concentrations (v_a):

$$E_a = \frac{Z_{ON}}{Z_{ON} + Z_{OFF}} = \frac{\sum_\sigma W(\sigma)Q(\sigma)}{\sum_\sigma W(\sigma)Q(\sigma) + \sum_\sigma W(\sigma)}$$

$$\frac{\partial u_a}{\partial t} = r_a E_a \left(\mathbf{v}(x,t), \mathbf{V}(x,t)\right) - \lambda_a u_a(x,t) + d_a \frac{\partial^2 u_a}{\partial x^2},$$

$$\frac{\partial v_a}{\partial t} = R_a u_a(t-\tau) - \Lambda_a v_a(x,t) + D_a \frac{\partial^2 v_a}{\partial x^2}$$

$\mathbf{v}(x,t)$ — vector of gap protein concentrations (Hb, Kr, Gt, Kni).

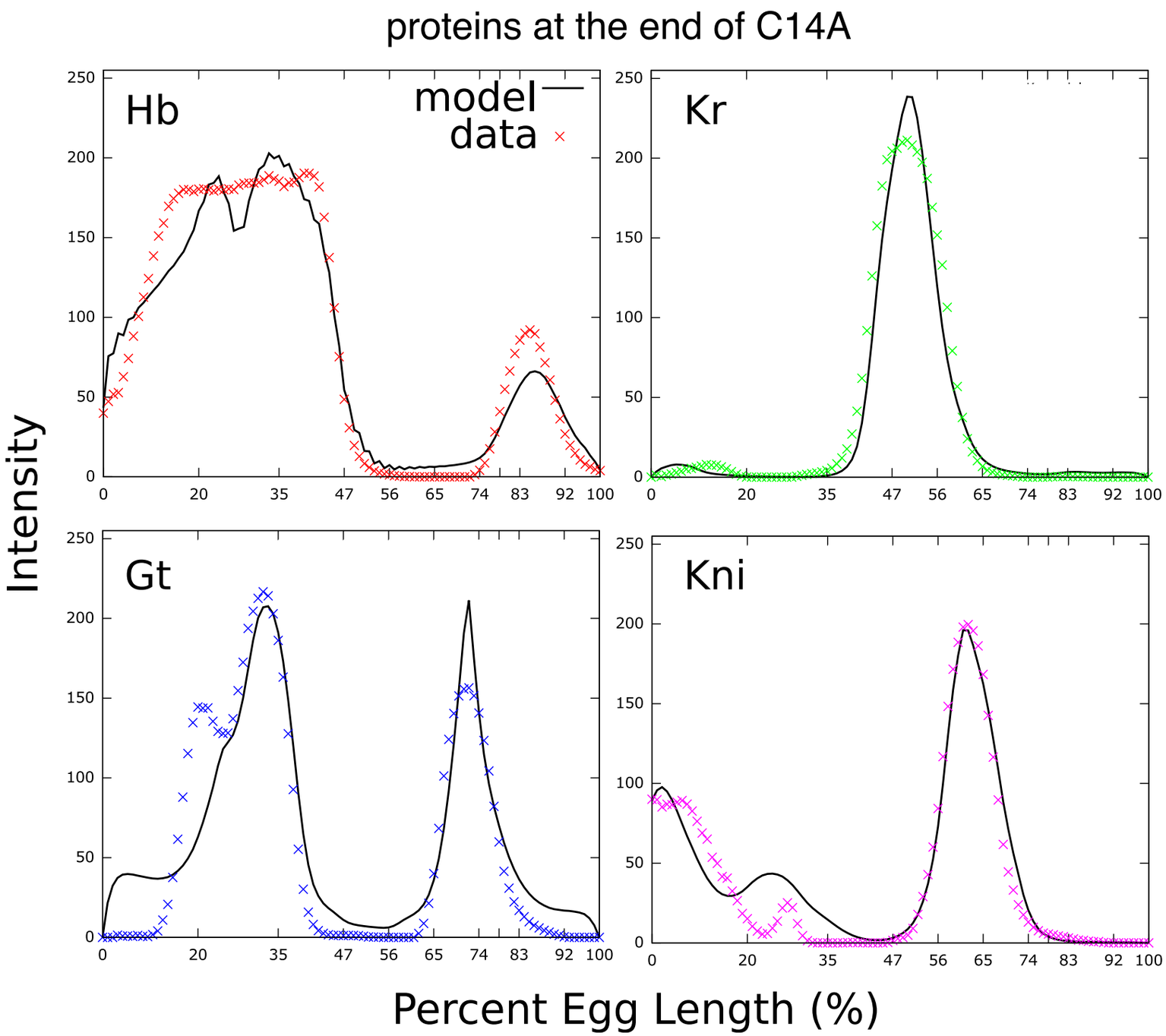
$\mathbf{V}(x,t)$ — vector of protein concentrations for external regulators (Bcd, Cad, Tll, Hkb).

τ — delay time (transcription+translation times).

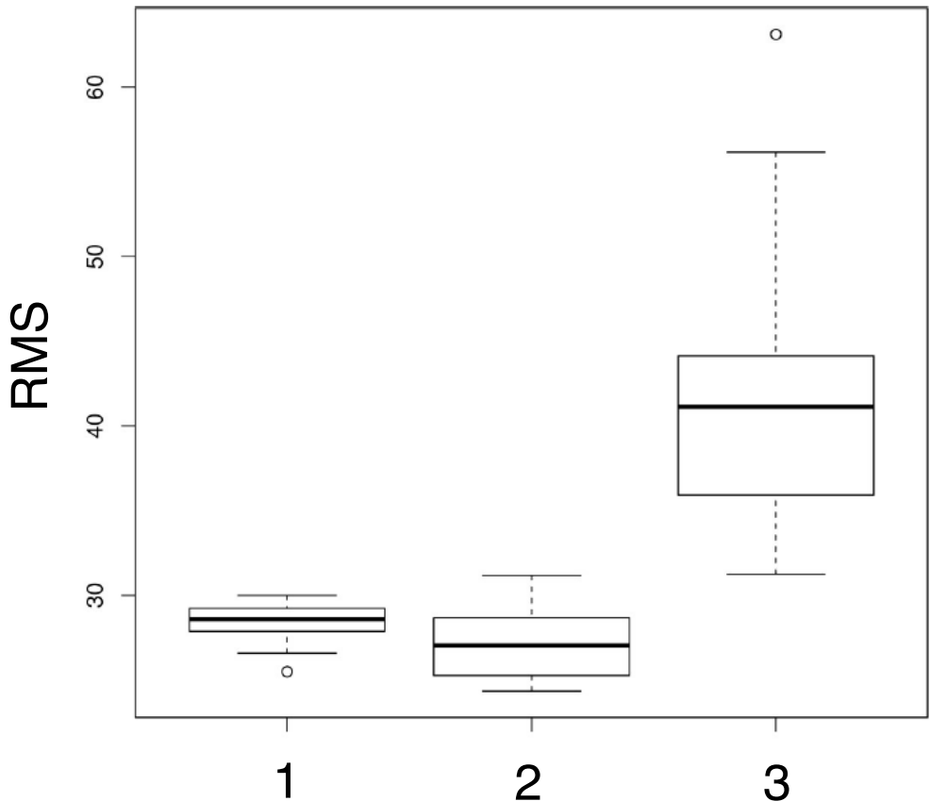
Kozlov et al. (2014) BMC Genomics 15(Suppl 12): S6; Kozlov et al. (2015) BMC Genomics 16(Suppl 13): S7

Fitting to wild-type expression patterns

The model was fitted to the gap gene expression data for cleavage cycles 13-14A.



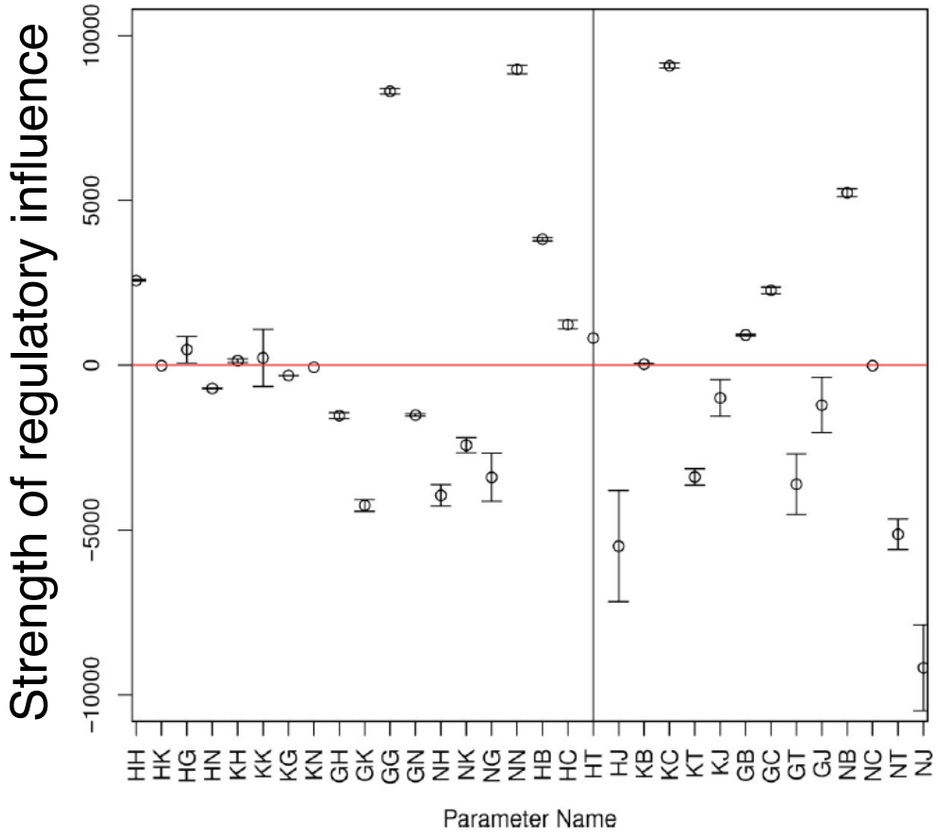
Cross-validation, negative control, local identifiability analysis



1: cross-validation

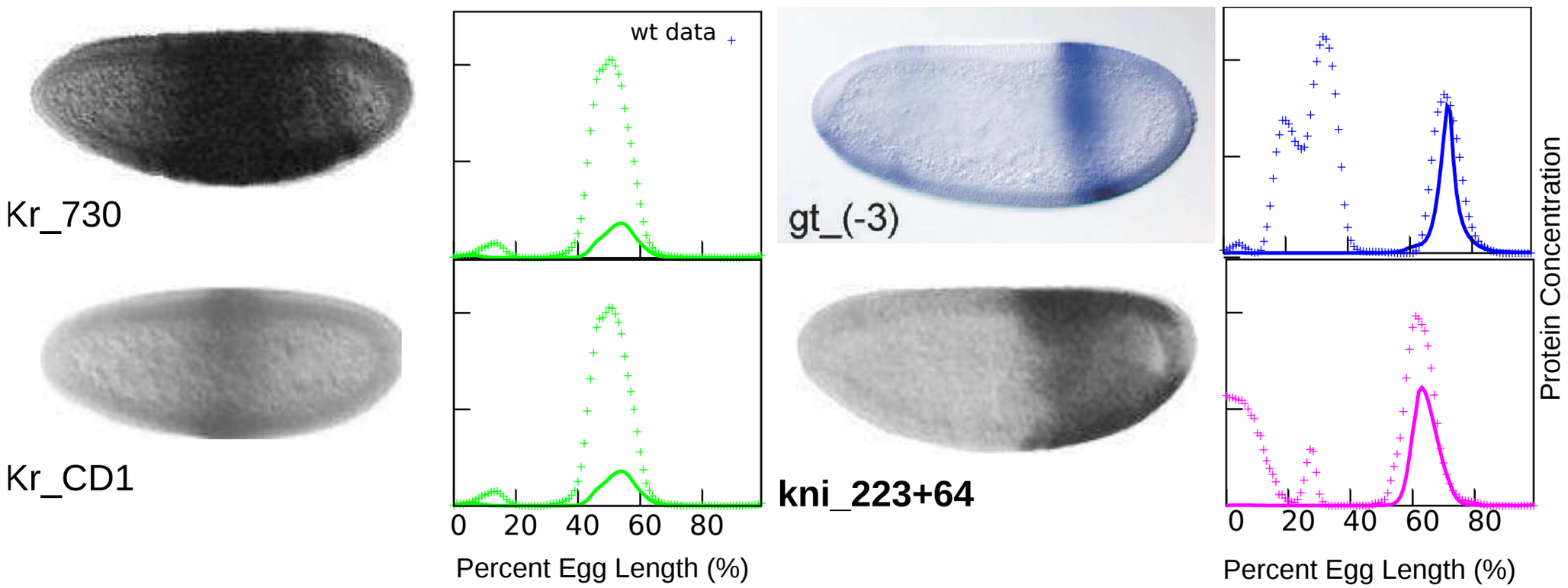
2: original fitting

3: fitting to random data



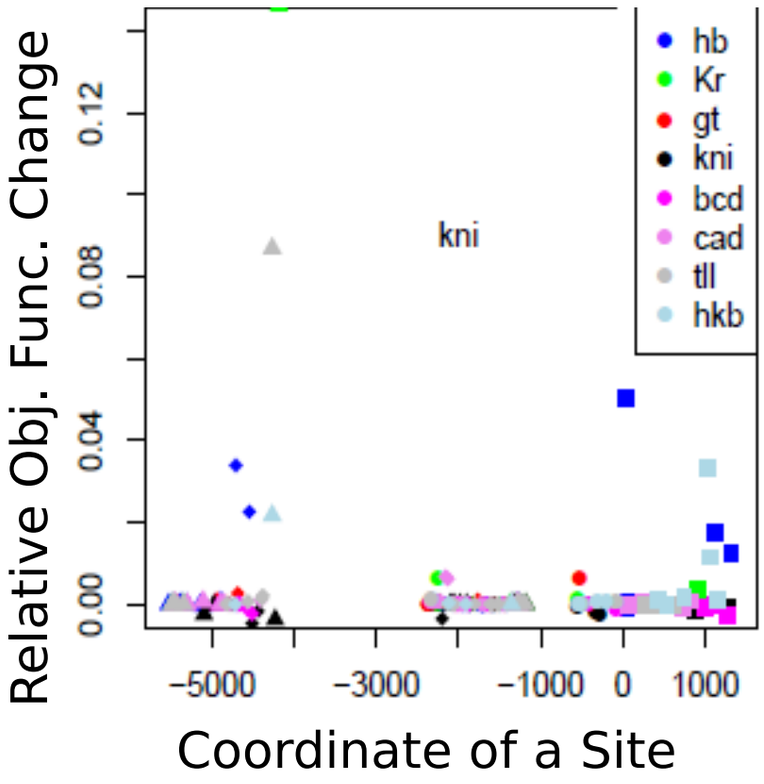
Testing on expression in reporter constructs

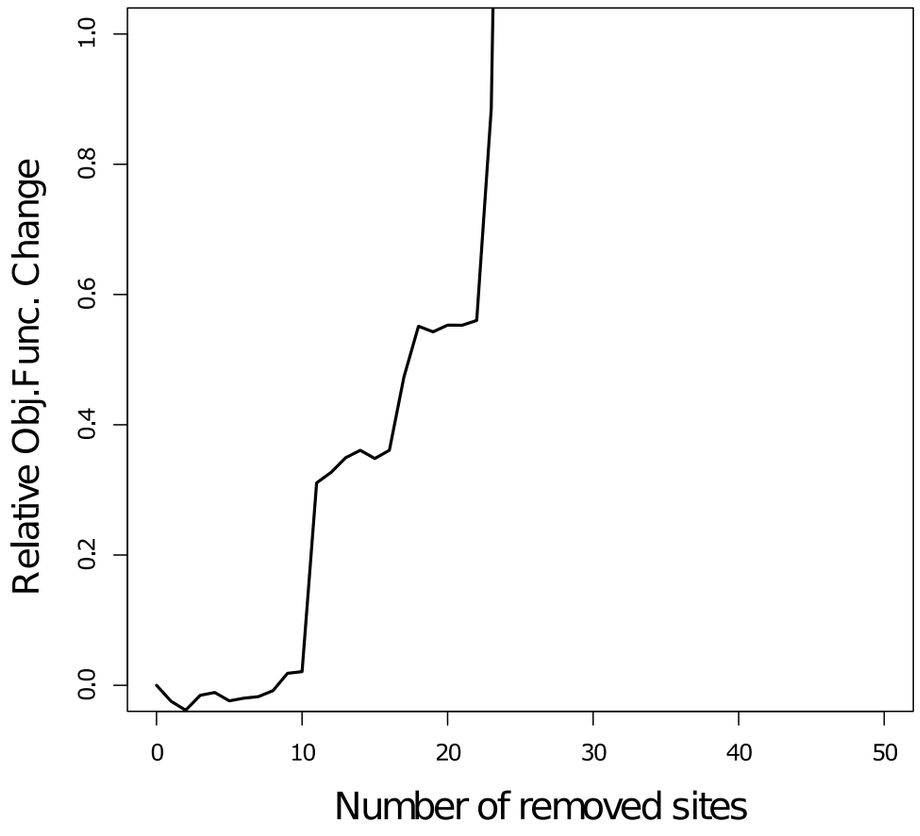
In silico predictions of expression patterns for various reporter constructs (Schroeder et al., 2004; Gallo et al., 2010):



Predictions for the regulatory landscape:
Many weak binding sites working in concert

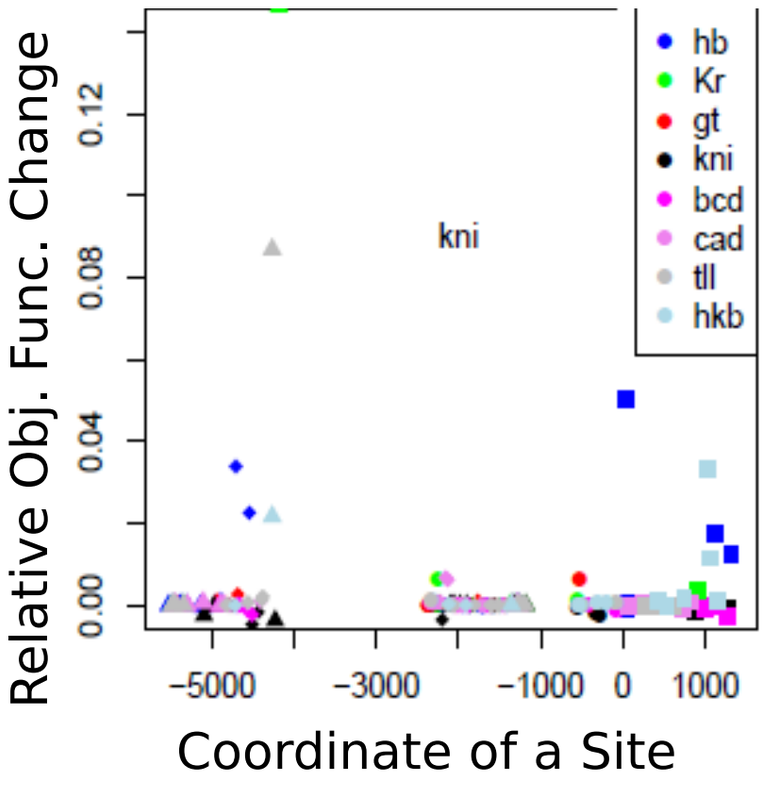
We calculate the **regulatory weight** of each binding site as the RMS-difference between the model outputs with and without this site.

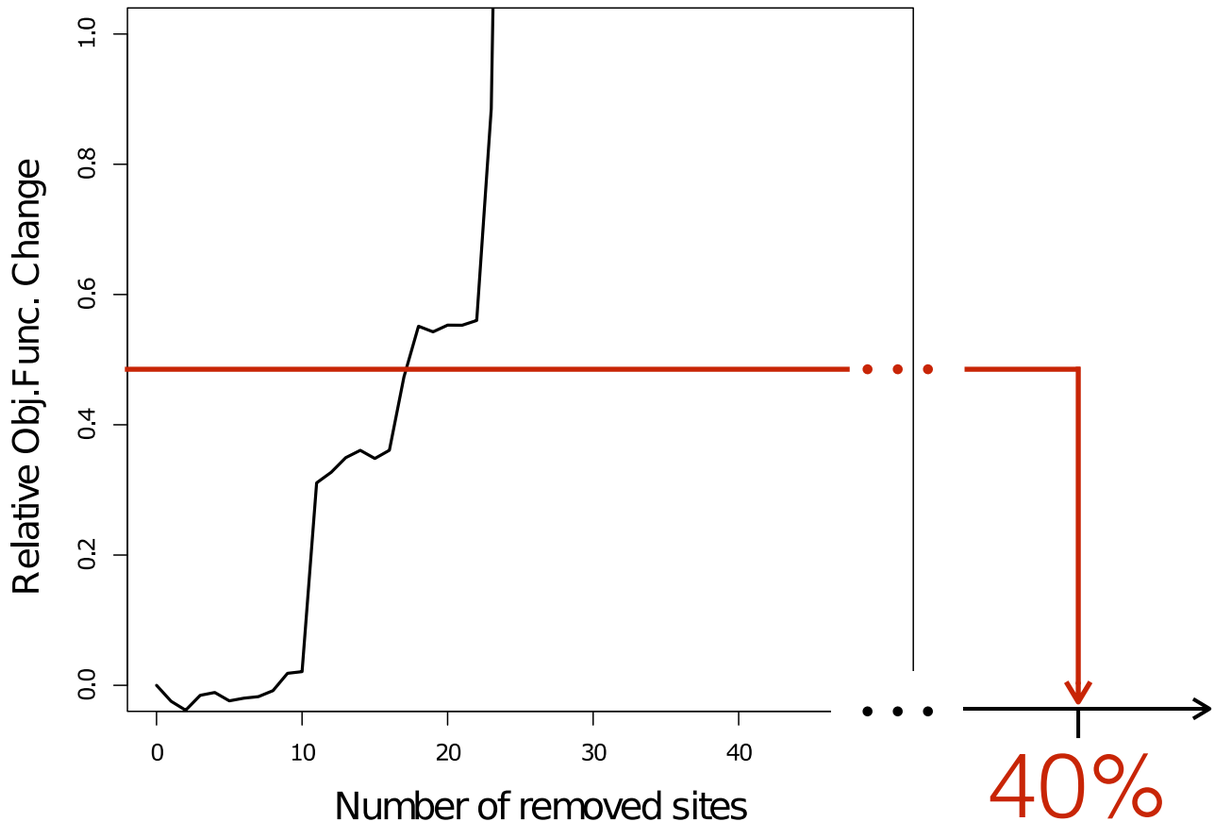




Predictions for the regulatory landscape: Many weak binding sites working in concert

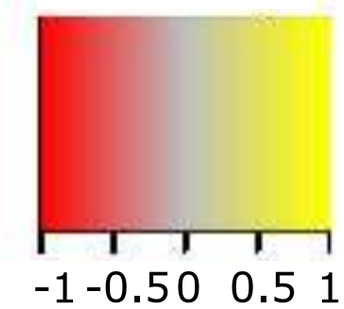
We calculate the **regulatory weight** of each binding site as the RMS-difference between the model outputs with and without this site.



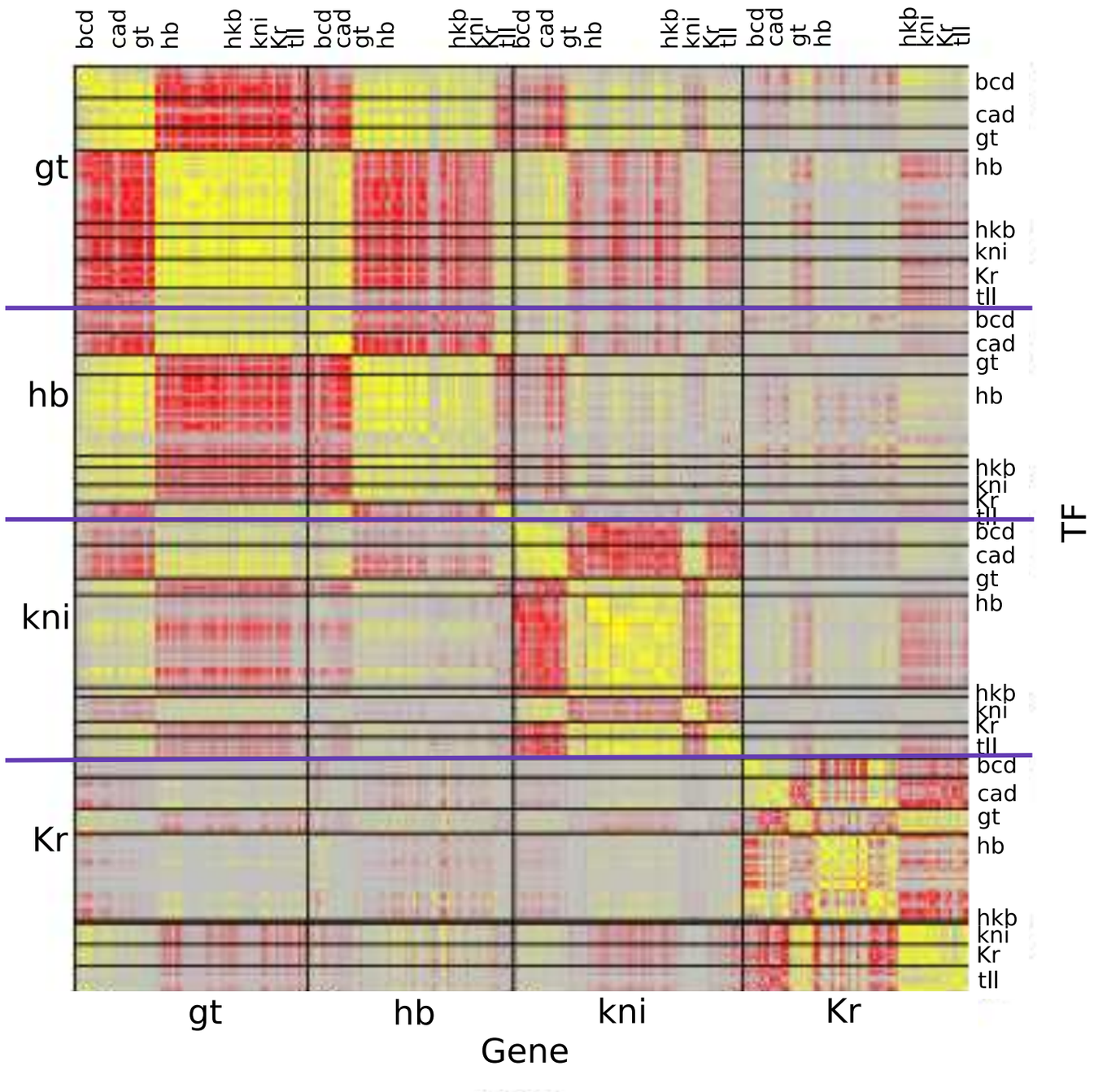


The removing of binding sites from the outside of known enhancers (approximately 40% of sites) did not change the patterns qualitatively. This shows that proximate sites demonstrate extensive **compensatory actions**.

Direct calculation of compensatory effects

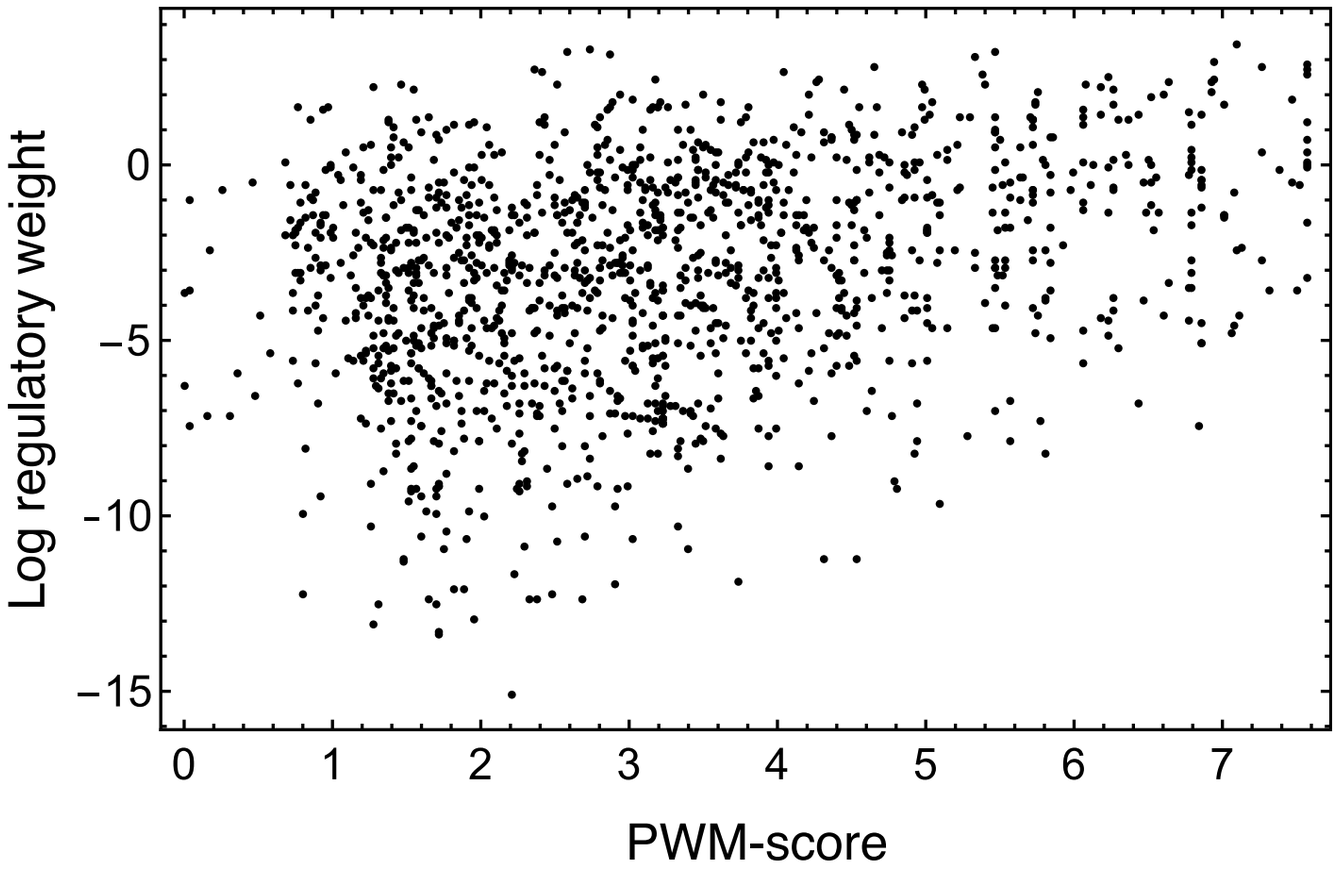


Heatmap of correlations between binding sites:



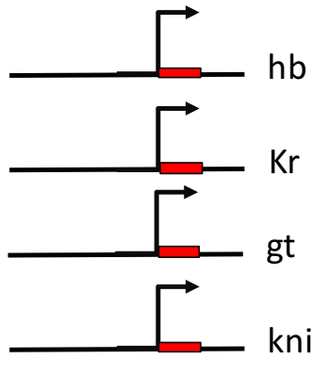
Binding affinity vs. influence on expression

Another type of ‘weakness’ in the concept of ‘weak sites working in concert’: Only **weak correlation** between the strength of influence on expression and the binding affinity for binding sites:



Simulating evolution of regulatory regions under elevated mutational pressure

Individual =



Genotype = DNase accessible regulatory regions of 4 gap genes

Phenotype = Expression patterns of the gap gene network

Population = 100 individuals

↓

Random mutations (rate=0.001 per bp per generation)

↓

Calculate ‘phenotypes’

↓

Negative selection = select 20 individuals with the least rms-difference between their phenotypes and wild-type expression patterns

↓

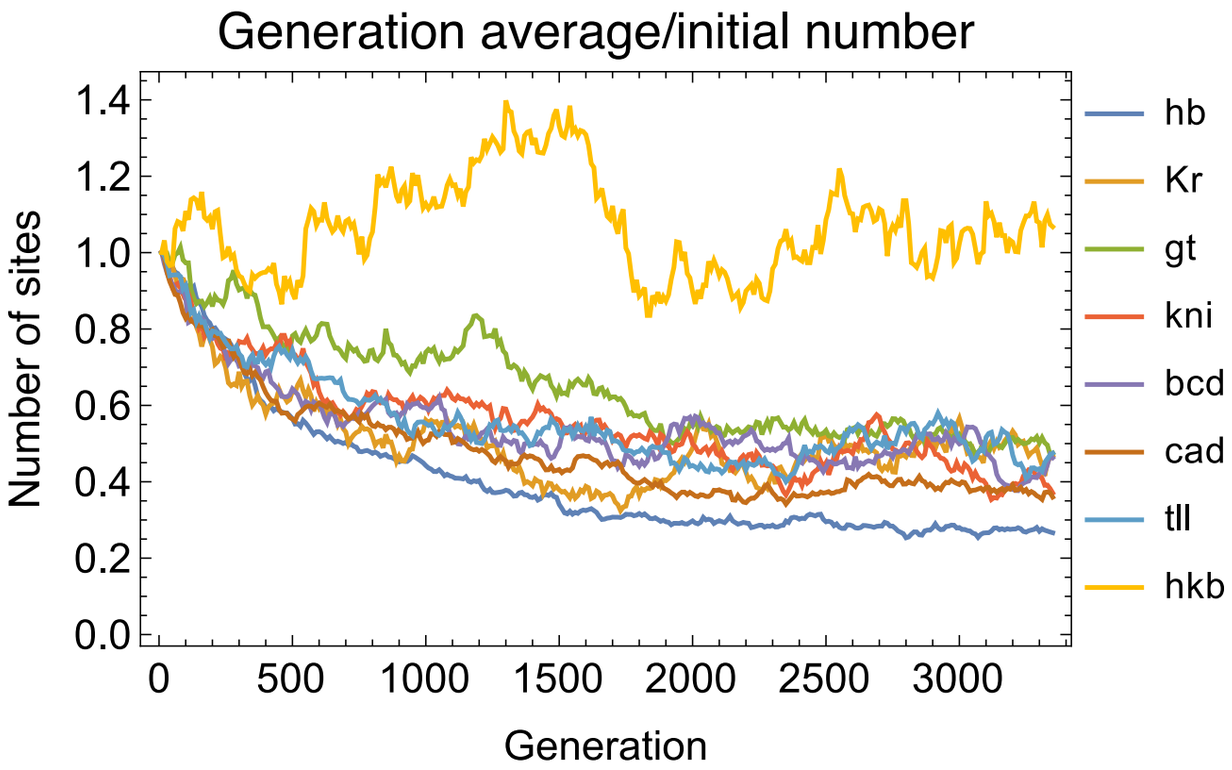
New generation (100 individuals) = recombination of genes from parents randomly selected from the 20 ‘best’ individuals

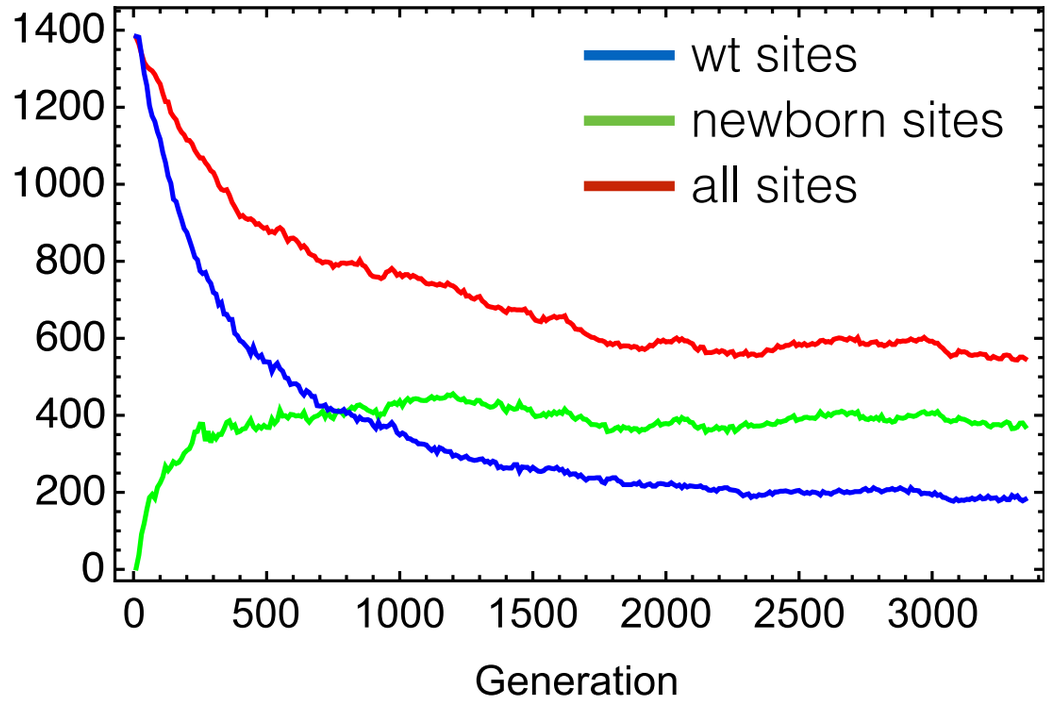
↓

Repeat for 3500 generations

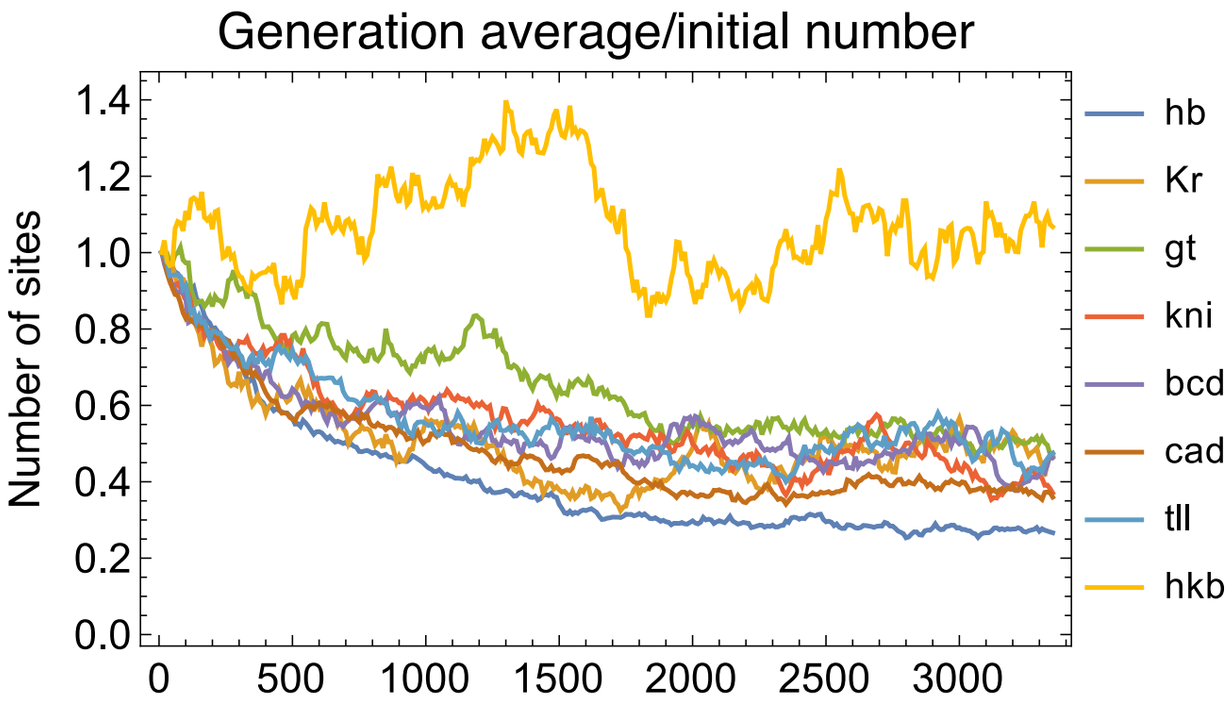
Chertkova et al. (2017) BMC Evolutionary Biology, 17(Suppl 1)

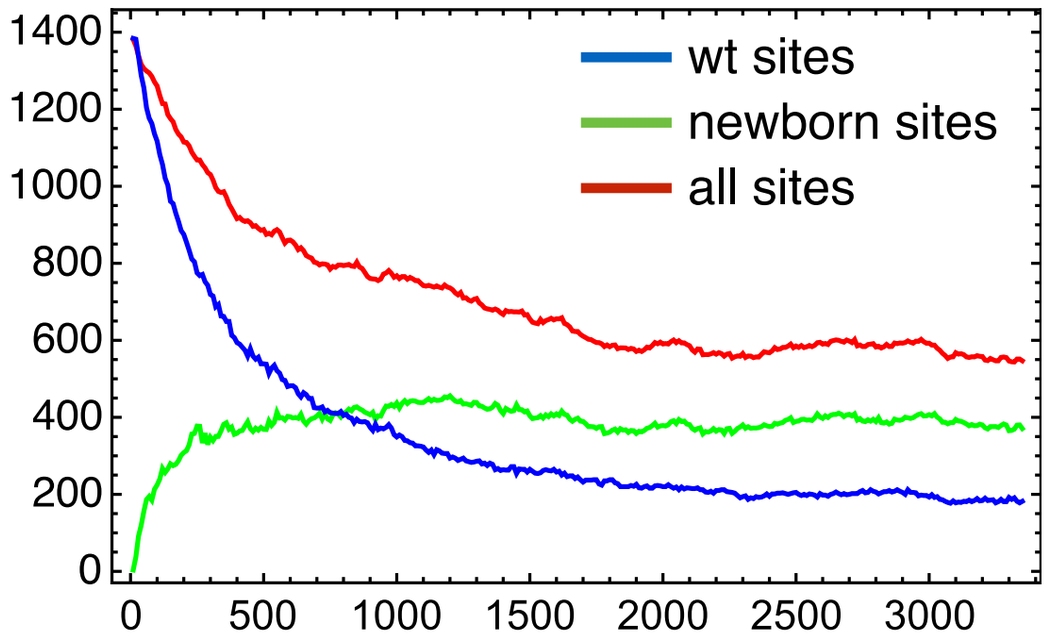
Dynamics of the number of binding sites



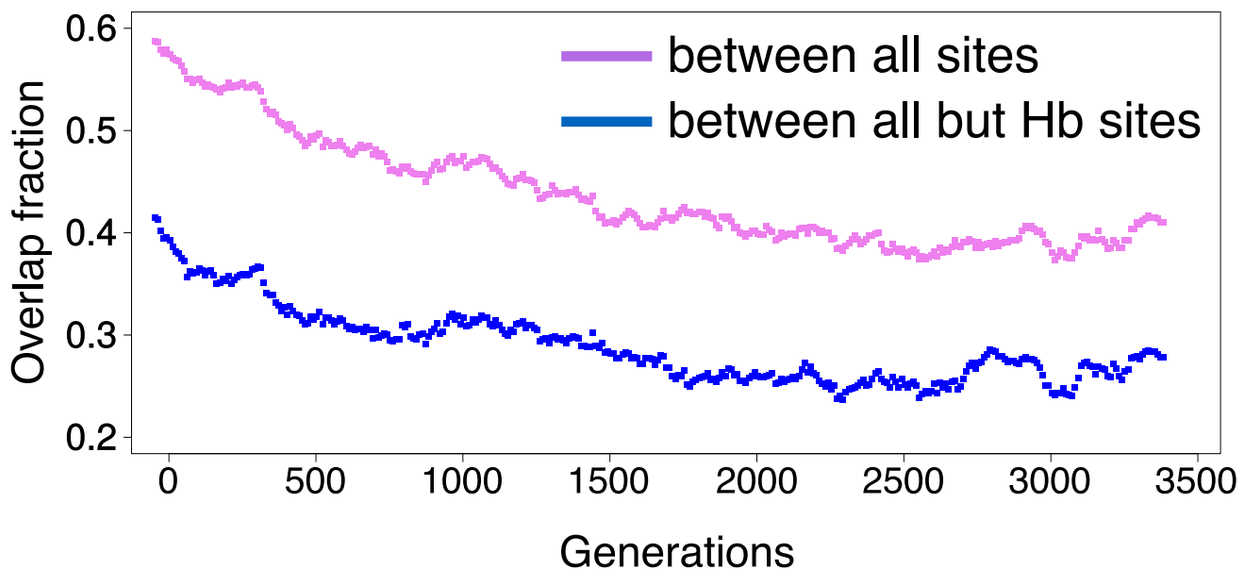


Dynamics of the number of binding sites

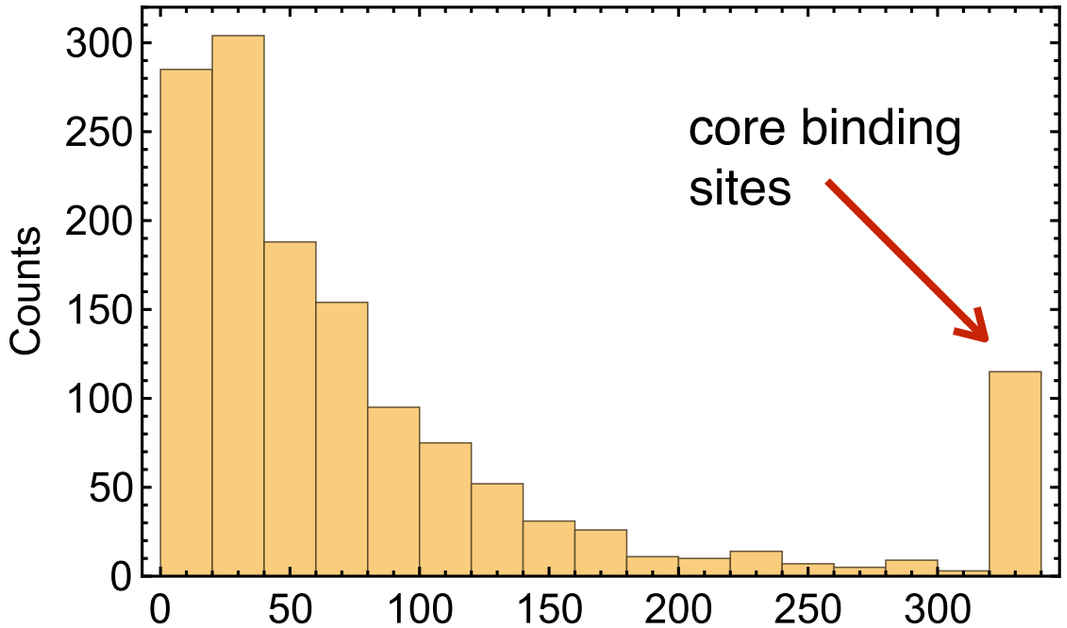




The dynamics of the population average number of overlapping events

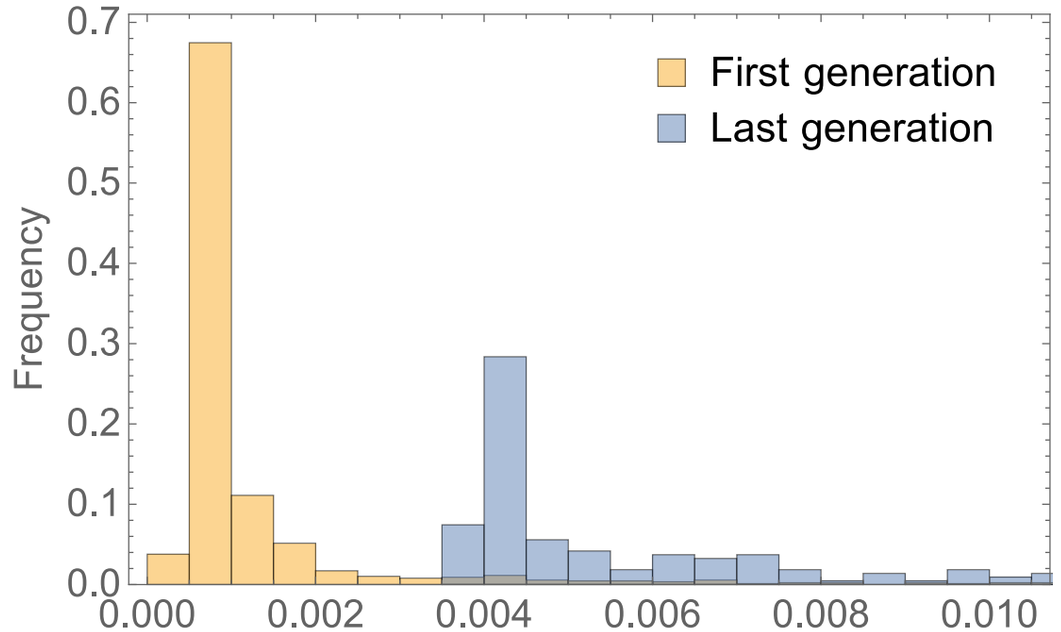


Core binding sites



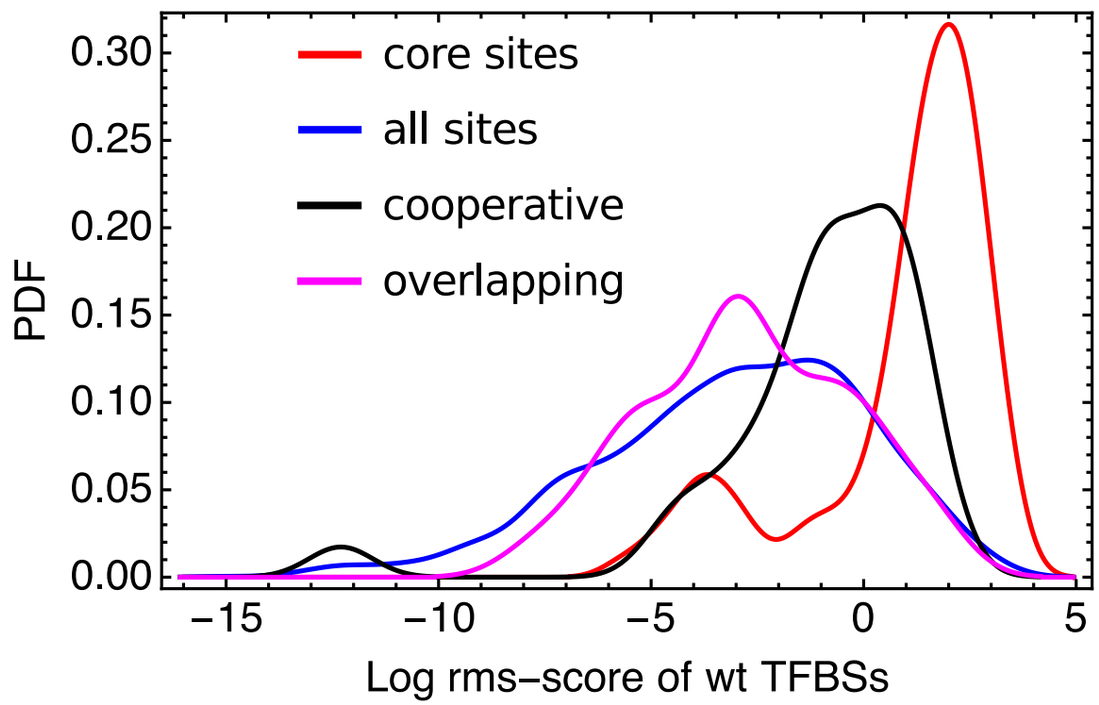
TFBS lifetime (10 x gen.)

Site score = rms difference between expression patterns with and without this site



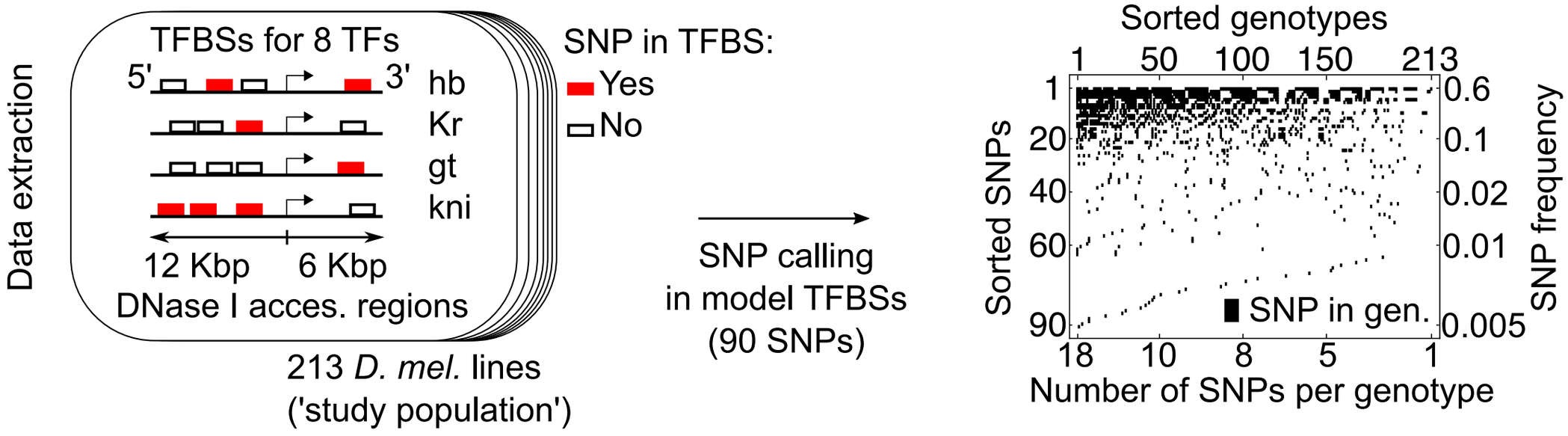
Site score

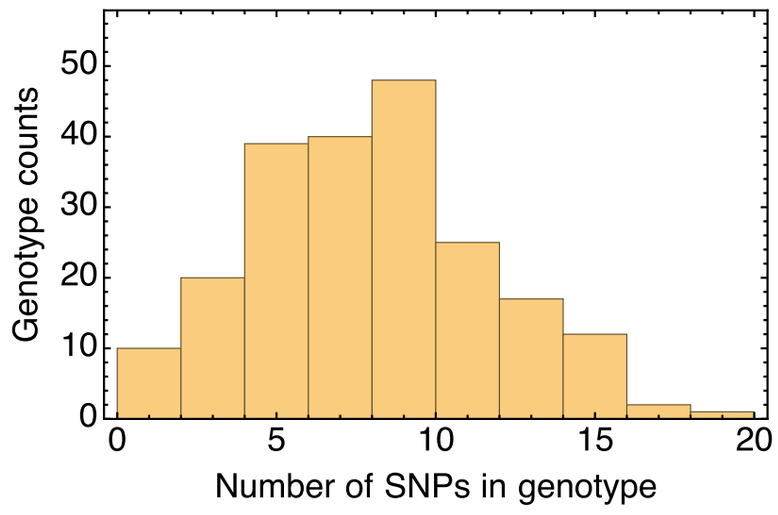
Distributions of site scores for wt conditions (zero generation):



Analysis of genetic variation in *Drosophila* lines

We apply the gene expression model to analyze single nucleotide polymorphisms (**SNPs**) in the regulatory regions of gap genes in a panel of **213 natural sequenced *D. melanogaster* lines** from two populations (Raleigh, NC, and Winters, CA) (Campo et al. 2013).

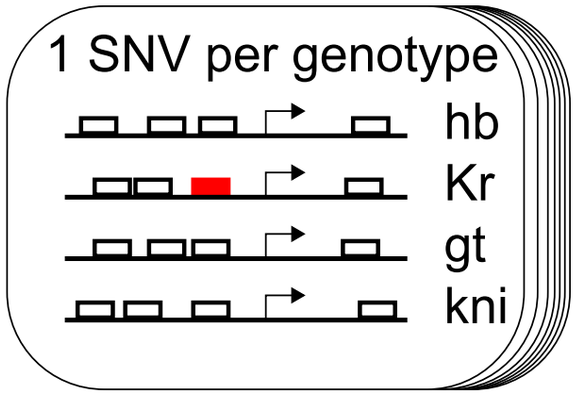




Contrasting natural genetic variation with simulated SNPs

Simulation of random point mutations:

1) **One** point mutation per genotype:



16 656 SNVs for 889 TFBSs

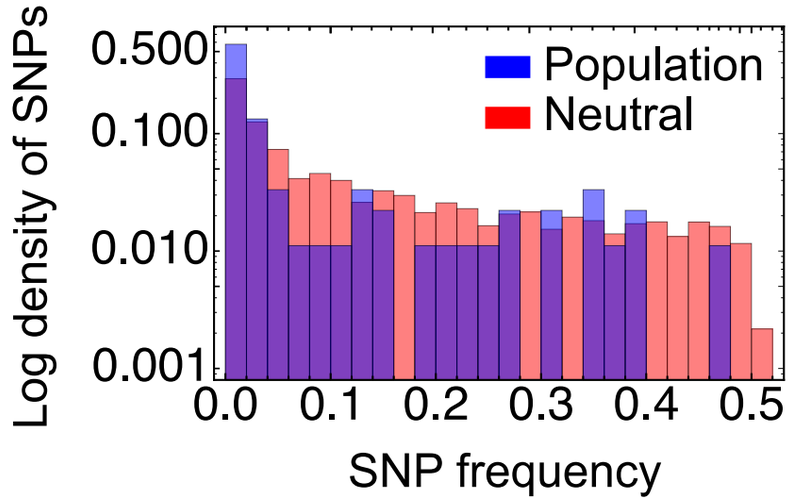
Contrasting natural genetic variation with simulated SNPs

Simulation of random point mutations:

2) **Sets of SNPs** per genotype, considering 2 frequency spectra:

2.1) **Neutral spectrum**: the frequency distribution of SNPs extracted from short intron positions of the D. mel genome.

2.2) **Population-derived spectrum**



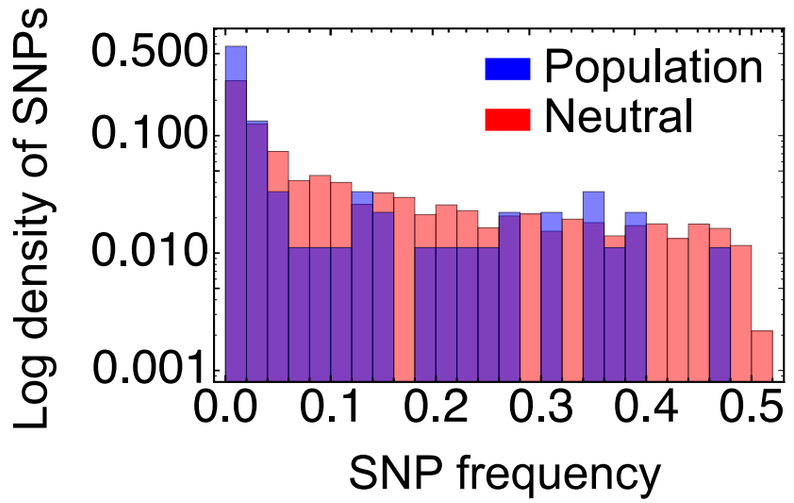
Contrasting natural genetic variation with simulated SNPs

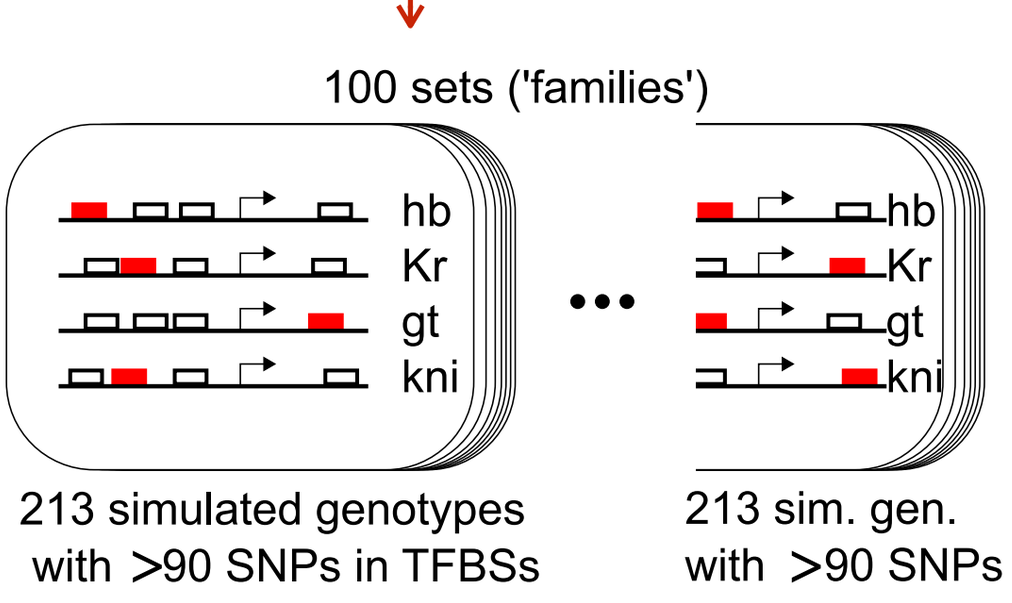
Simulation of random point mutations:

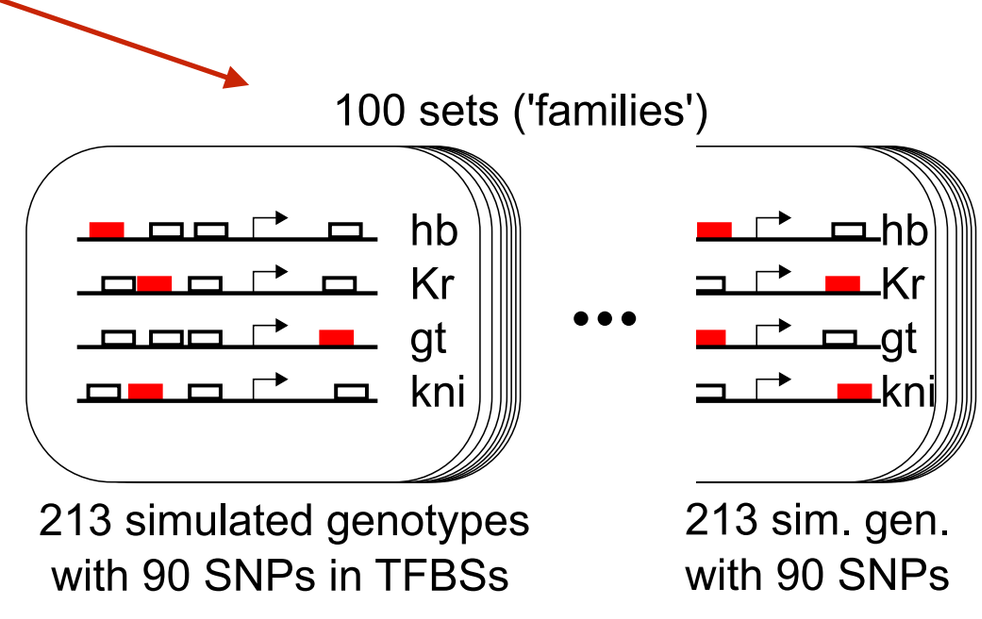
2) **Sets of SNPs** per genotype, considering 2 frequency spectra:

2.1) **Neutral spectrum**: the frequency distribution of SNPs extracted from short intron positions of the D. mel genome.

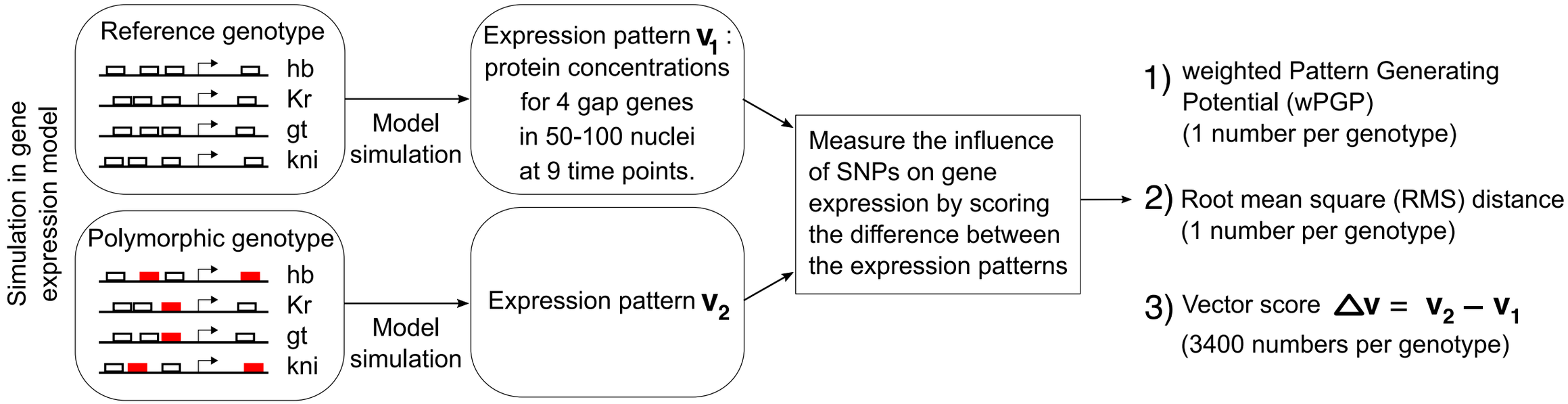
2.2) **Population-derived spectrum**



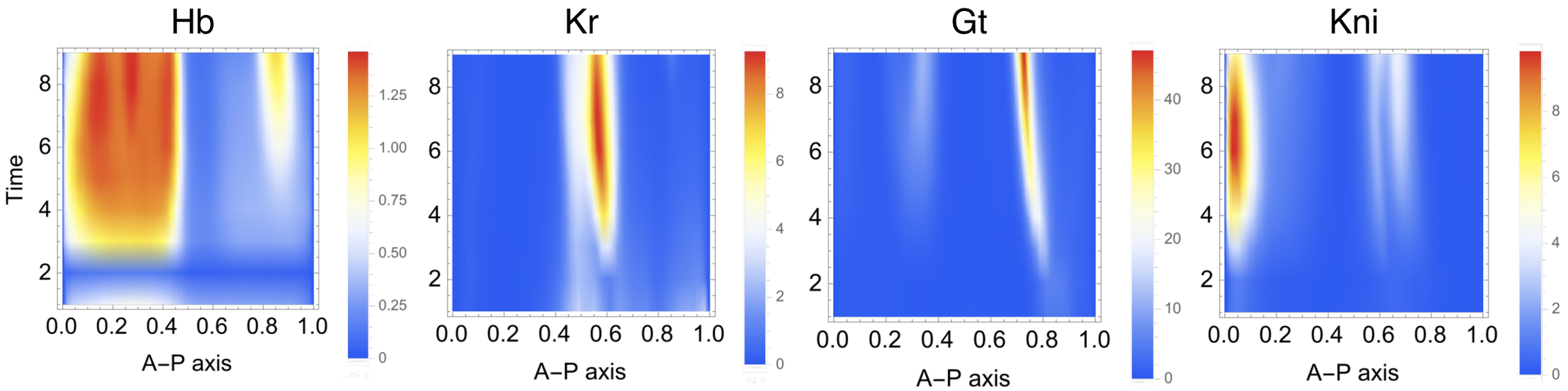


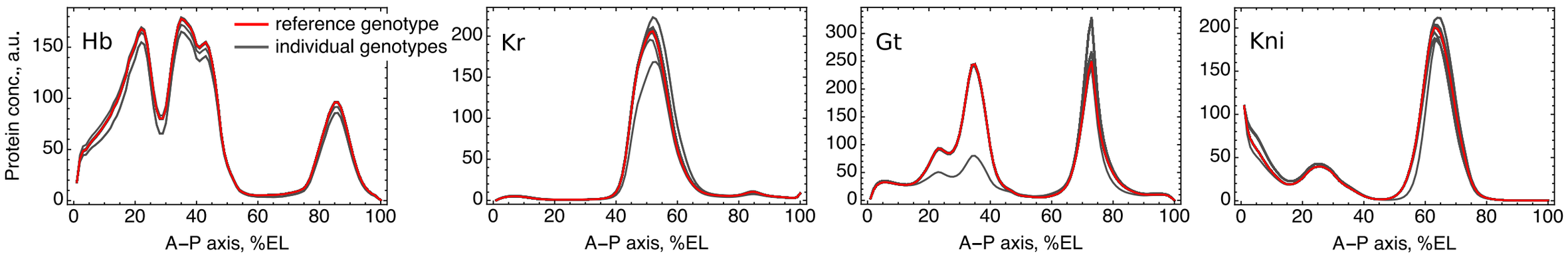


Translating natural genetic variation to gene expression



Model simulation of the population genotypes





Weighted pattern generating potential (wPGP) is better than RMS difference

$$w = \sum_{t,a} f^a(t), \quad f^a = 0.5 - 0.5 * (\text{reward} - \text{penalty}),$$

$$\text{reward} = \frac{\sum_i V_i^a \min(V_i^a, v_i^a)}{\sum_i (V_i^a)^2}, \quad \text{penalty} = \frac{\sum_i \max(0, v_i^a - V_i^a)(V_{\max}^a - V_i^a)}{\sum_i (V_{\max}^a - V_i^a)^2},$$

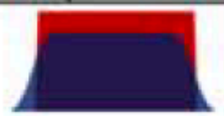
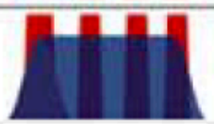
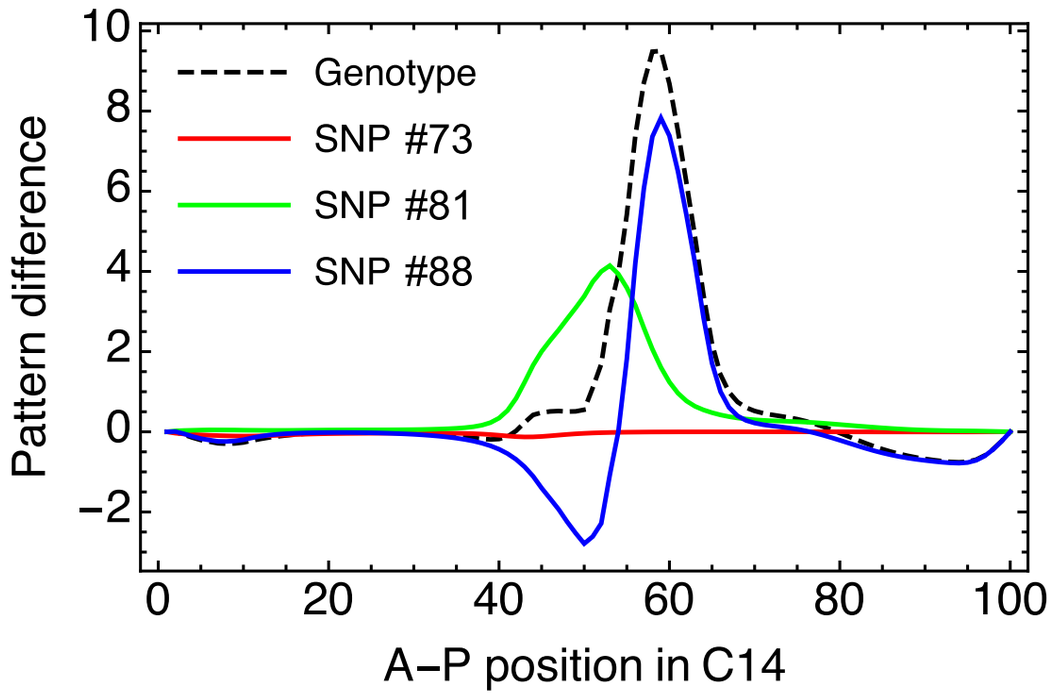
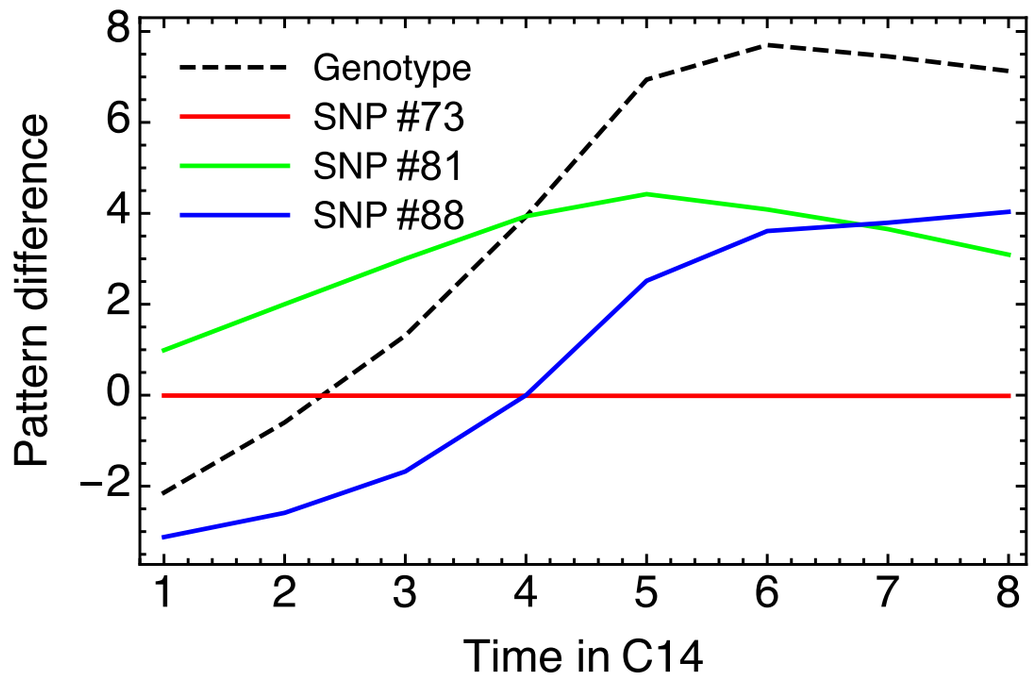
Characteristic	Expression	PGP	CC	1-RMSE
Sensitive to scaling		0.81	0.96	0.81
		0.58	0.96	0.43
Sensitive to shift in basal expression		0.69	0.96	0.62
		0.39	0.96	0.71
Normalization for length of expression domain		0.69	0.96	0.62
		0.69	0.94	0.74
Sensitive to partial pattern		0.56	0.40	0.53
		0.51	0.63	0.62

Figure 3B from Kazemian et al. (2010) Plos Biol. 8: e1000456

Sign of SNP influence on expression

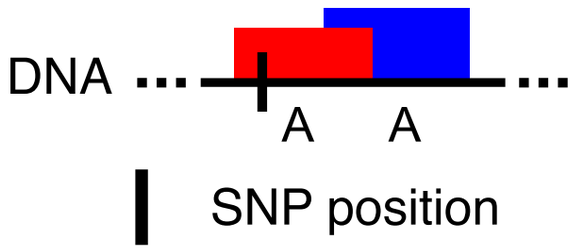
SNP's influence on expression can be **sign-alternating**:





Specific examples of sign analysis

SNP influence sign = sign of $\dfrac{\partial E}{\partial q}$

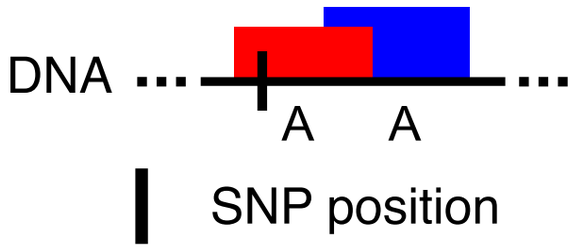


$$\frac{\partial E}{\partial q} > 0 \quad <=> \quad \alpha > \frac{1+\alpha_1 q_1}{1+q_1}$$

A/R: activating/repressing binding site

Specific examples of sign analysis

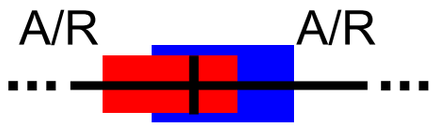
SNP influence sign = sign of $\dfrac{\partial E}{\partial q}$

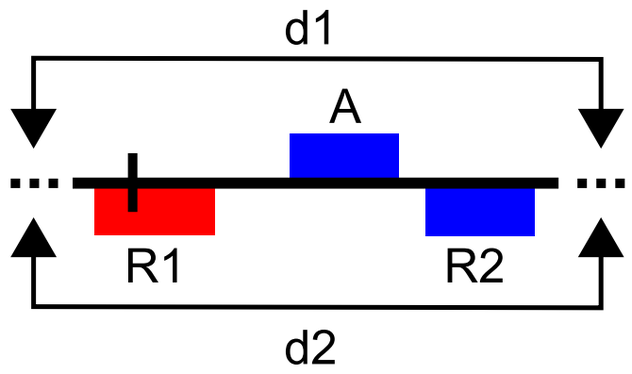


$$\frac{\partial E}{\partial q} > 0 \quad <=> \quad \alpha > \frac{1+\alpha_1 q_1}{1+q_1}$$

A/R: activating/repressing binding site

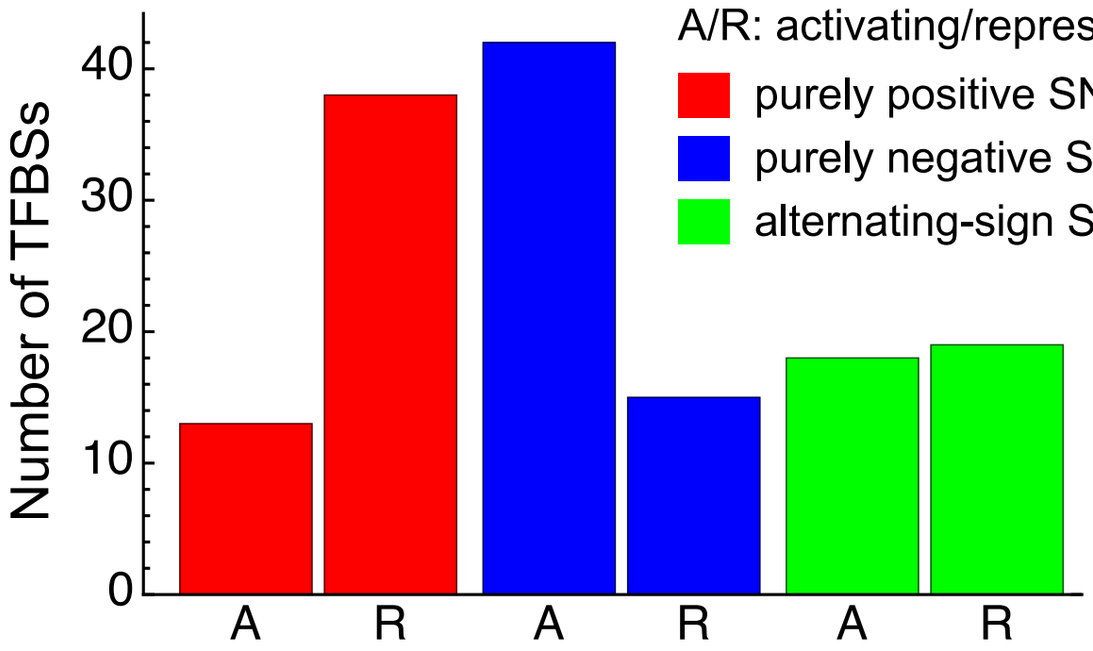
Other examples of alternating sign of SNP influence on expression:

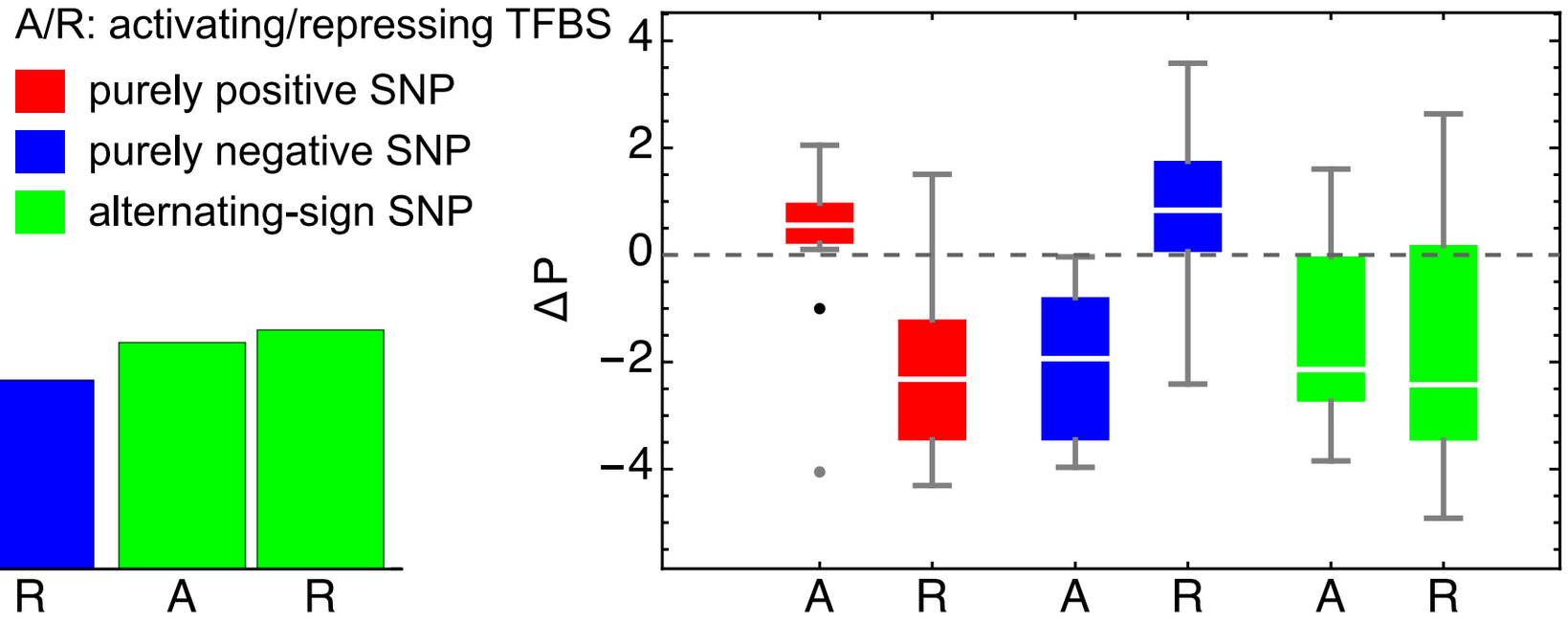




SNPs activate (repress) expression via repressing (activating) sites

Distribution of SNPs from the population among activating and repressing binding sites:





Combination of influences from multiple SNPs

Mutating a set of *n* binding sites in the regulatory region of a gene:

$$S_i \to \tilde{S}_i, \quad i = 1, \ldots, n$$

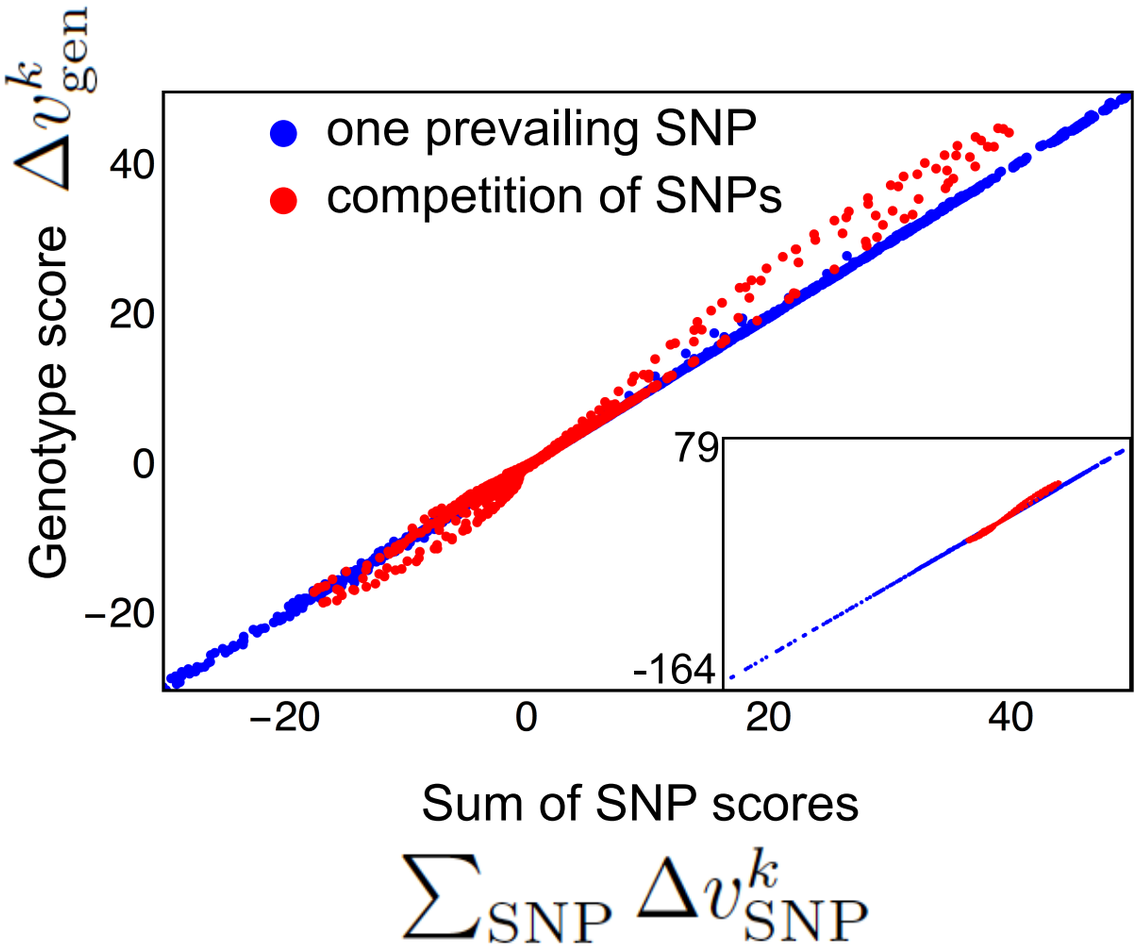
δ_i — scaled difference between a PCM matrix element due to *i* th SNP.

Probability of transcriptional activation for the mutated regulatory region:

$$\tilde{E} = \frac{Z_{ON} + P_n(\delta_1, \ldots, \delta_n)}{Z_{ON} + Z_{OFF} + Q_n(\delta_1, \ldots, \delta_n)}$$

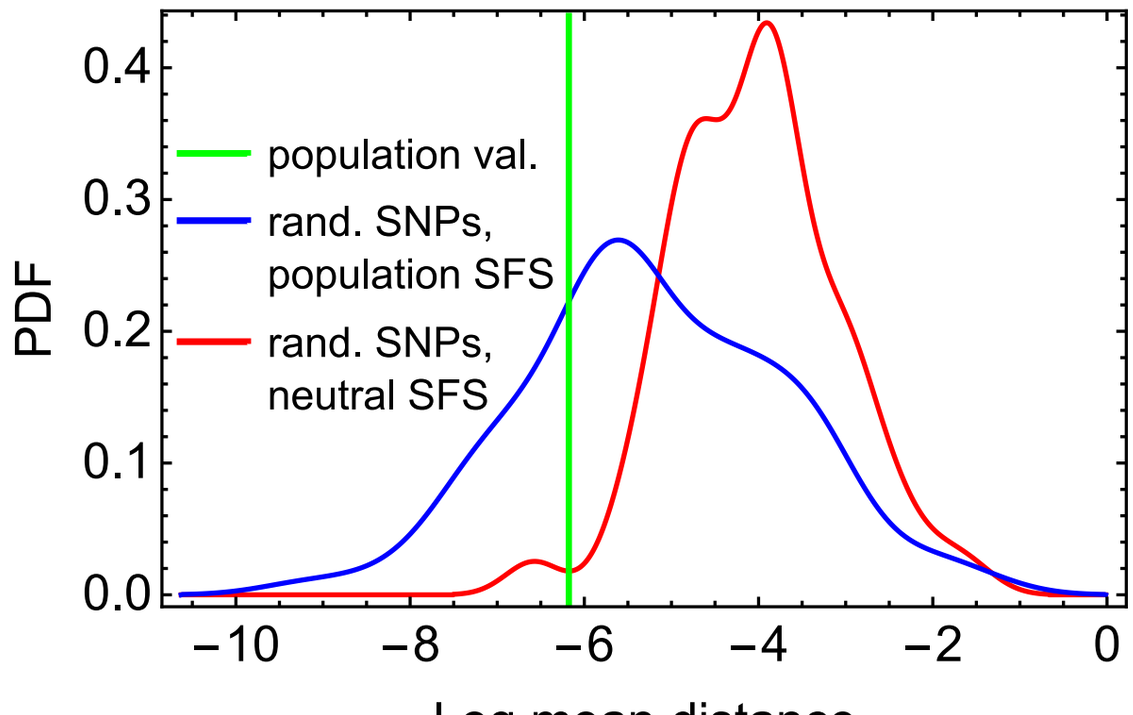
$\mathbf{\Delta v}_{\mathrm{gen}}$ = nonlinear function of $\sum_{\mathrm{SNP}} \mathbf{\Delta v}_{\mathrm{SNP}}$

SNP influence combination in the population genotypes



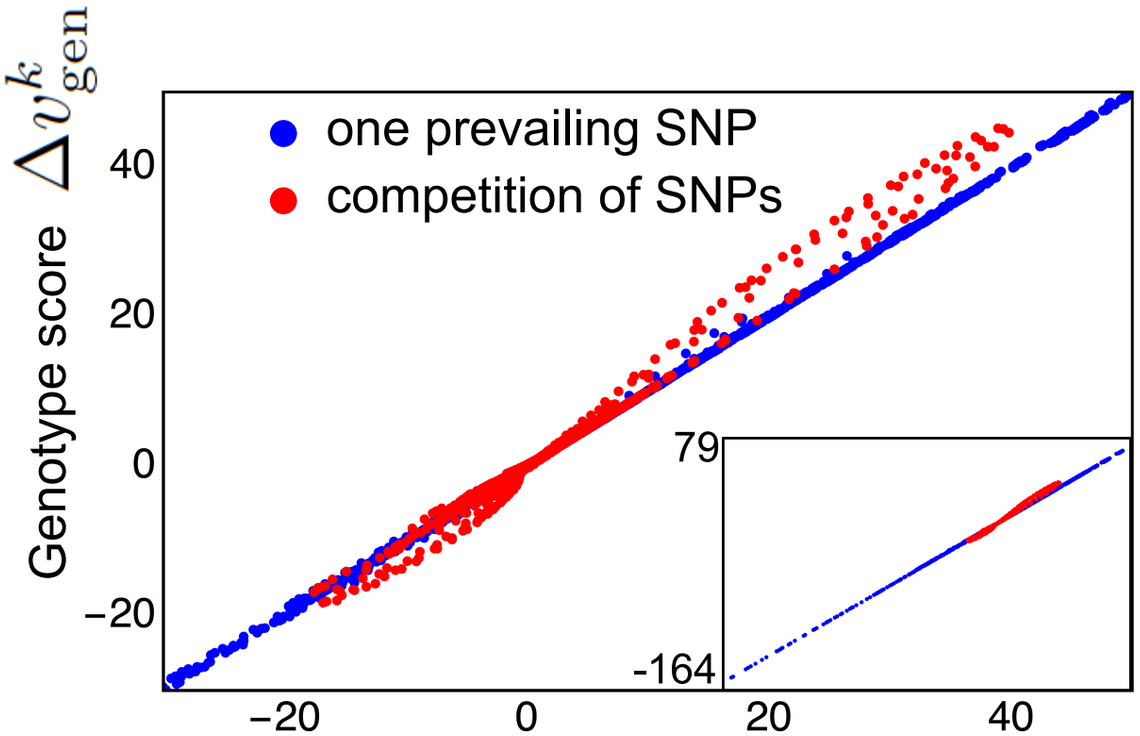
Sum of SNP scores

$\sum_{\rm SNP} \Delta v^k_{\rm SNP}$



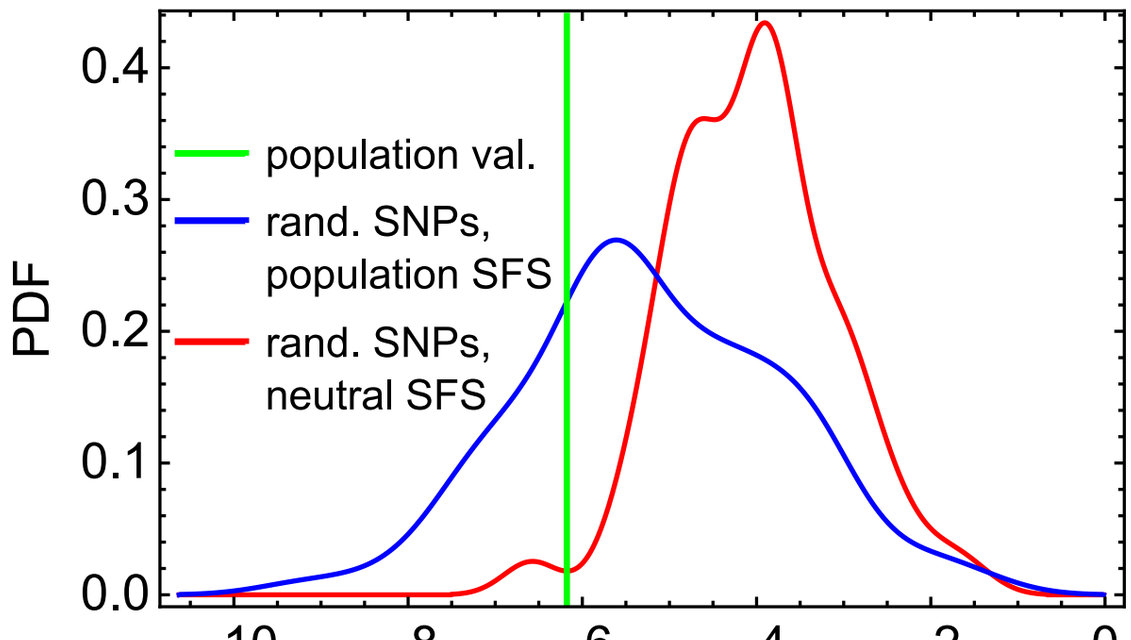
Log mean distance

SNP influence combination in the population genotypes



Sum of SNP scores

$$\sum_{\rm SNP} \Delta v^k_{\rm SNP}$$



Log mean distance

OPEN ACCESS Freely available online

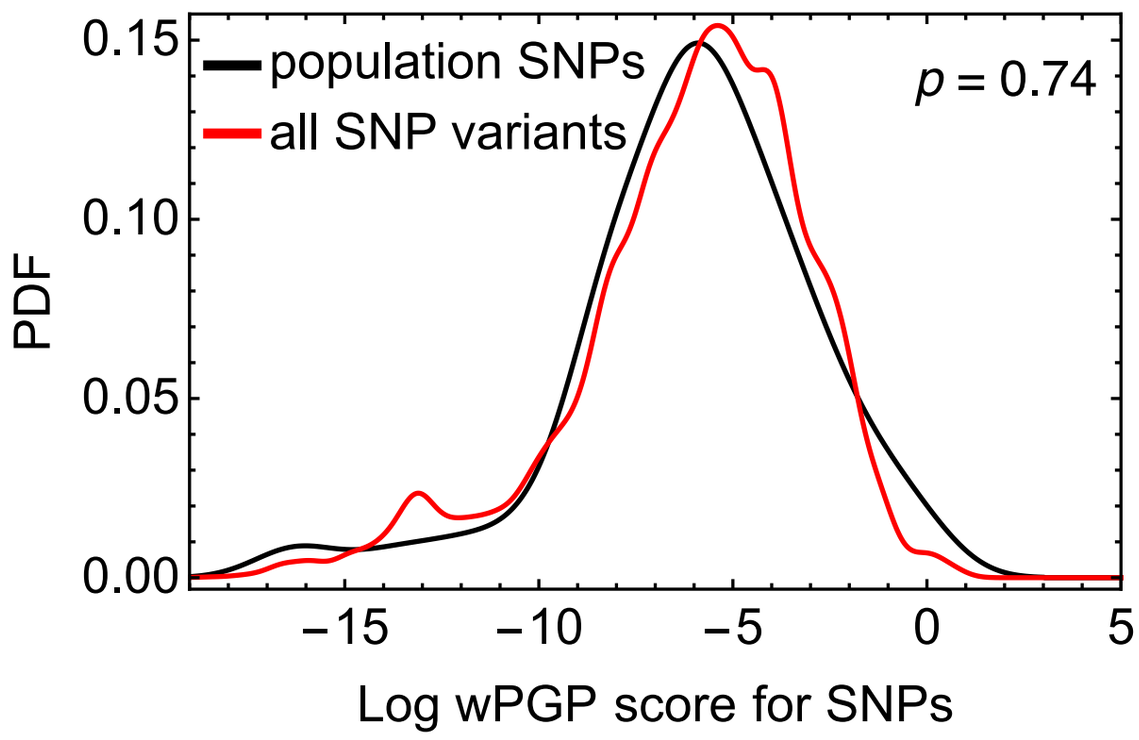
PLoS GENETICS

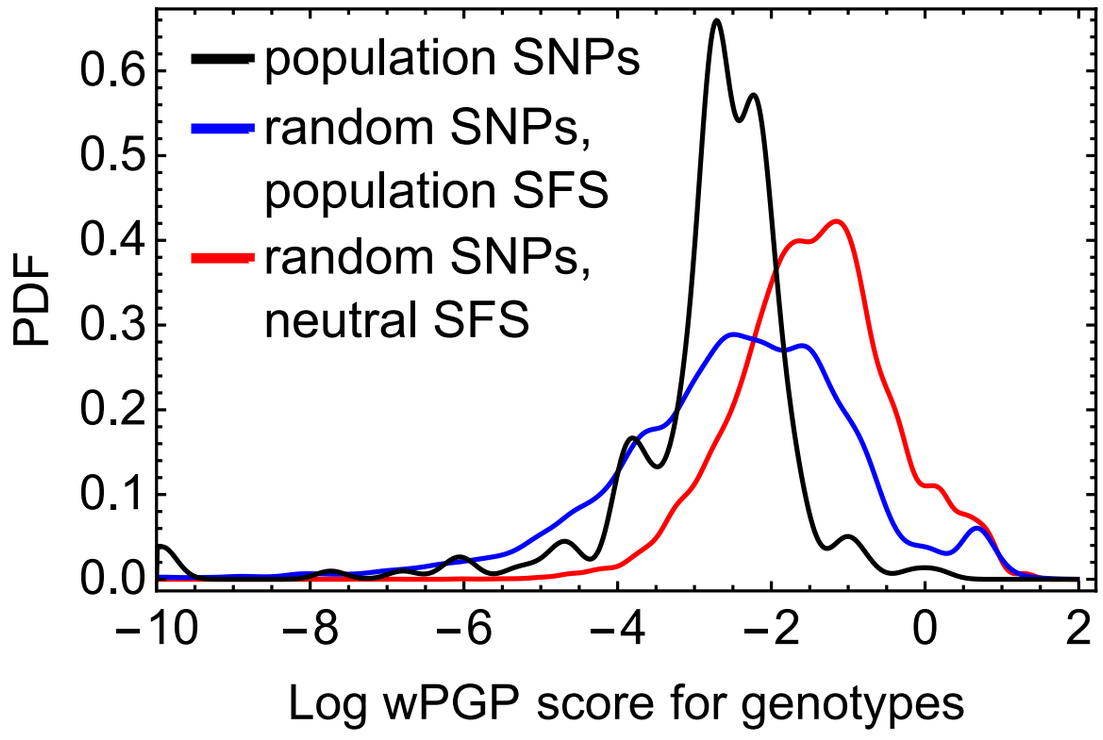
Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits

William G. Hill[1]*, Michael E. Goddard[2,3], Peter M. Visscher[4]

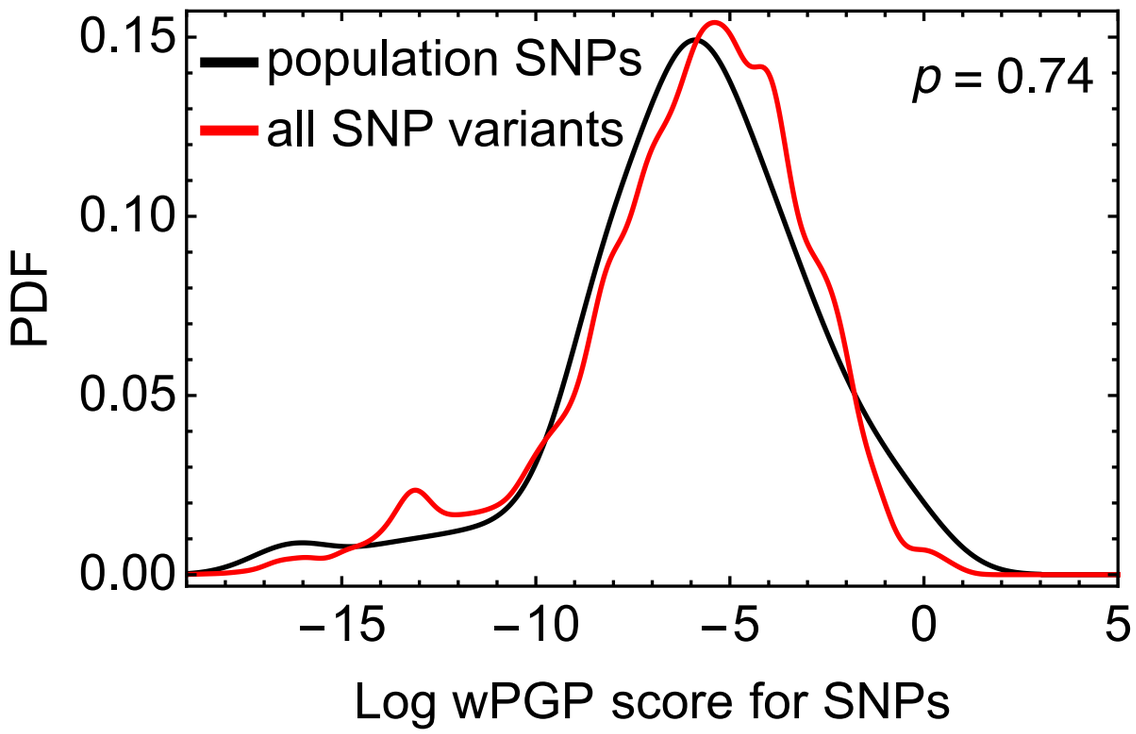
1 Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **2** Faculty of Land and Food Resources, University of Melbourne, Victoria, Australia, **3** Department of Primary Industries, Victoria, Australia, **4** Queensland Institute of Medical Research, Brisbane, Australia

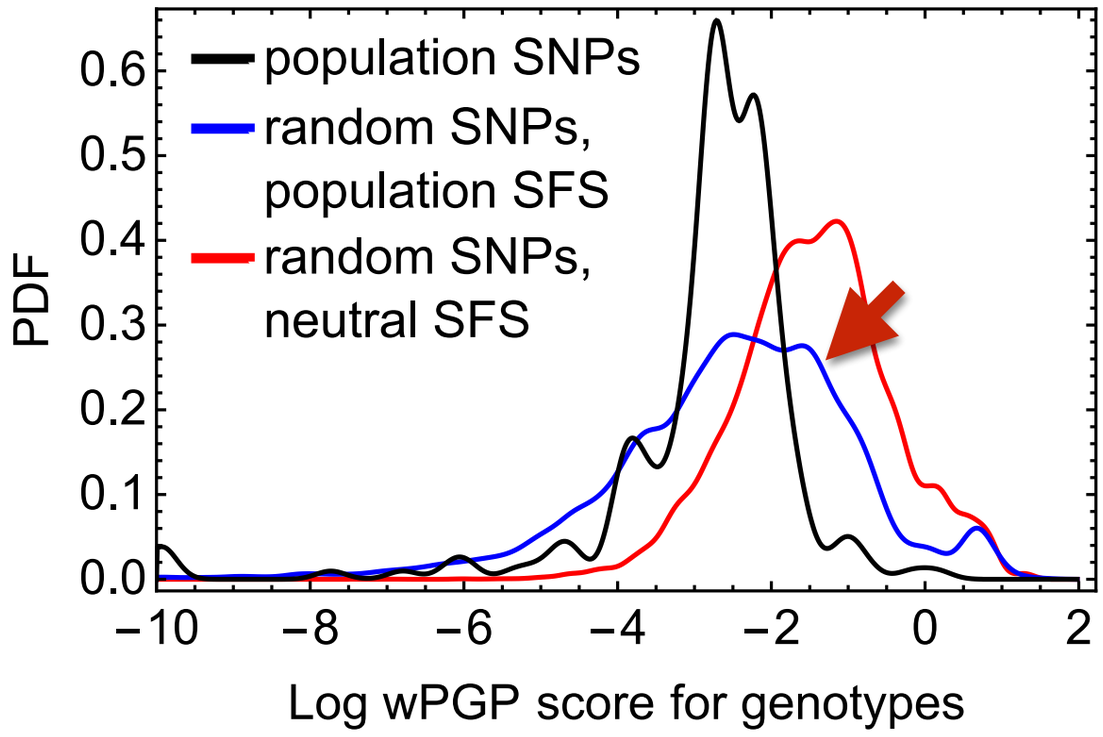
Individual SNPs do not seem to be under purifying selection, but combinations of SNPs do.





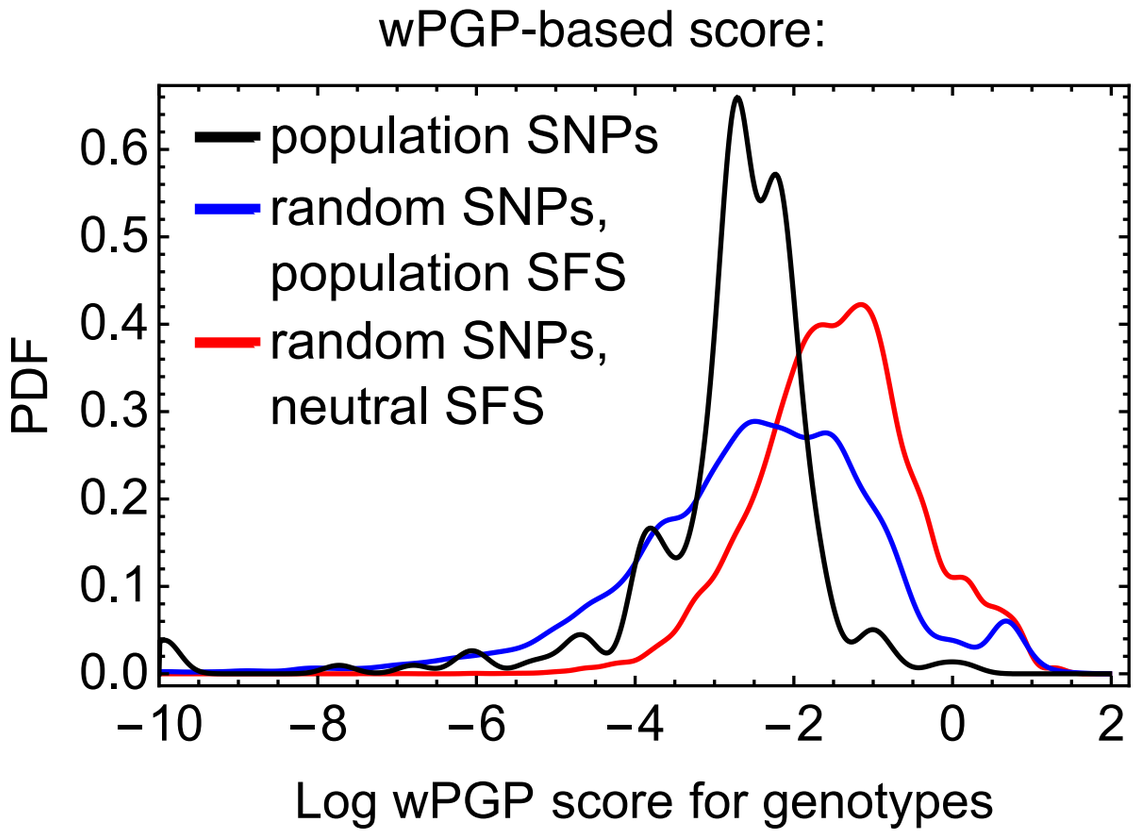
Individual SNPs do not seem to be under purifying selection, but combinations of SNPs do.

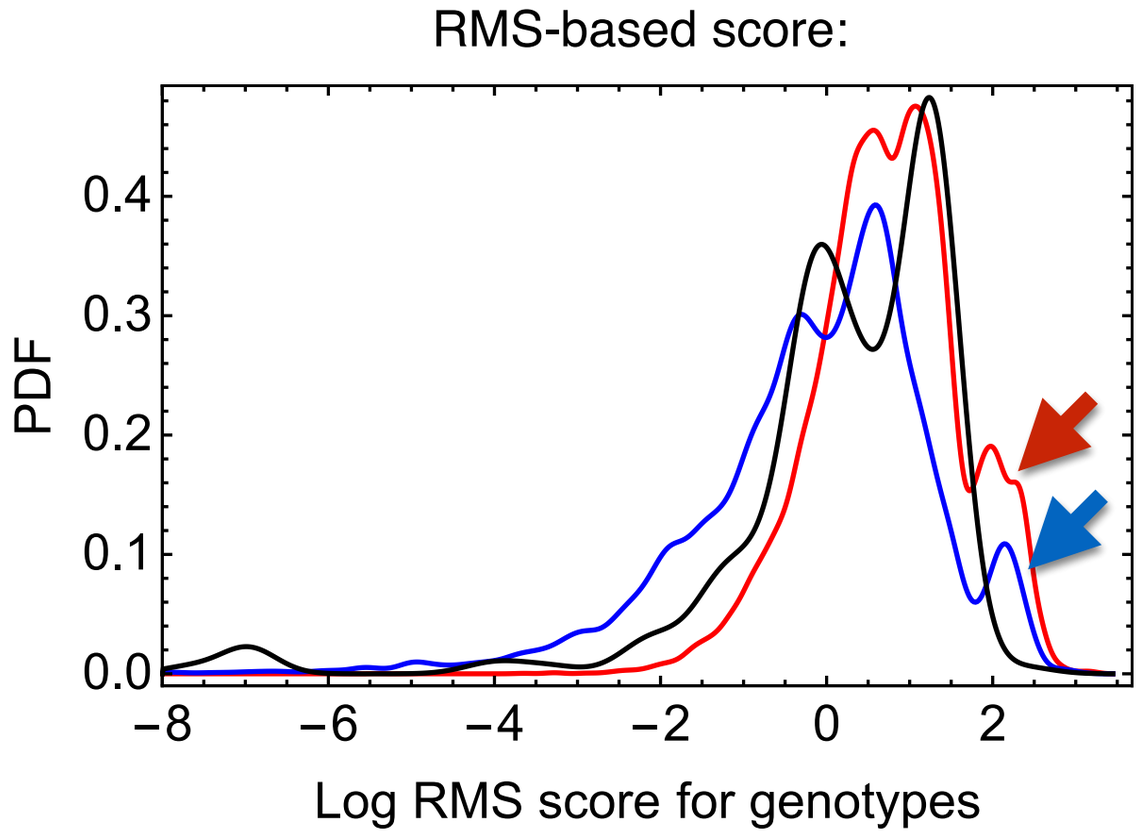




Only SNP **combinations with strong influence** can be distinguished from random SNPs under the **population-derived spectrum**.

RMS and wPGP measures lead to different evolutionary predictions





Thank you

Systems biology and bioinformatics Lab, Peter the Great Polytechnic University

Maria Samsonova
Sergey Nuzhdin
Konstantin Kozlov
Svetlana Surkova
Alyona Sokolkova
Alexandra Chertkova
Anna Igolkina

Ioffe Institute

Alexander Samsonov
Sergey Rukolaine

University of Southern California

Sergey Nuzhdin
Paul Marjoram

Engelhardt Institute of Molecular Biology,

Ivan Kulakovskiy