Epigenomic enrichment

Mikhail Dozmorov

Spring 2018

Gene enrichment vs. genome enrichment

- **Gene set enrichment analysis** - summarizing many **genes** of interest, such as differentially expressed genes, with a few common **gene annotations** (molecular functions, canonical pathways)

- **Epigenomic enrichment analysis** - summarizing many **genomic regions** of interest, such as disease-associated genomic variants, with a few common **genome annotations** (chromatin states, transcription factor binding sites)

Genomic regions

- Gene/exon boundaries, promoters
- Single Nucleotide Polymorphisms (SNPs)
- Transcription Factor Binding Sites (TFBS)
- Differentially methylated regions
- CpG islands

Each genomic region has coordinates (unique IDs):

`Chromosome`, `Start`, `End`

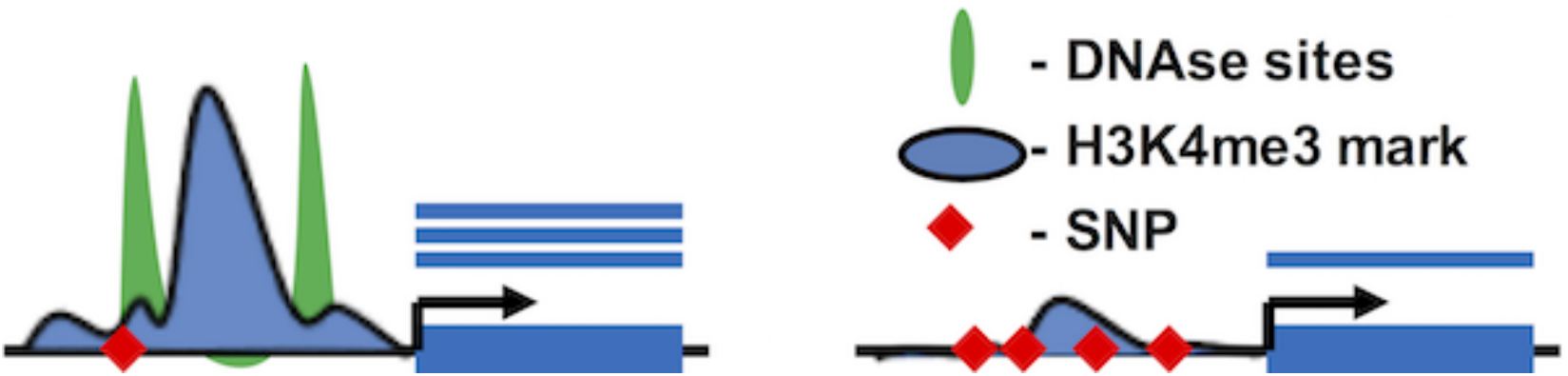
Annotations of genomic regions

- **Epigenomic (regulatory) regions** - genomic regions annotated as carrying functional and/or regulatory potential
- DNaseI hypersensitive sites
- Histone modification marks
- Transcription Factor Binding Sites
- DNA methylation
- Enhancers
- ...

Why "genomic region enrichment analysis"?

Enrichment = functional impact

- **Hypothesis**: SNPs in epigenomic regions may disrupt regulation
- More significant enrichment = more SNPs in epigenomic regions = more regulation is disrupted (SNP burden)

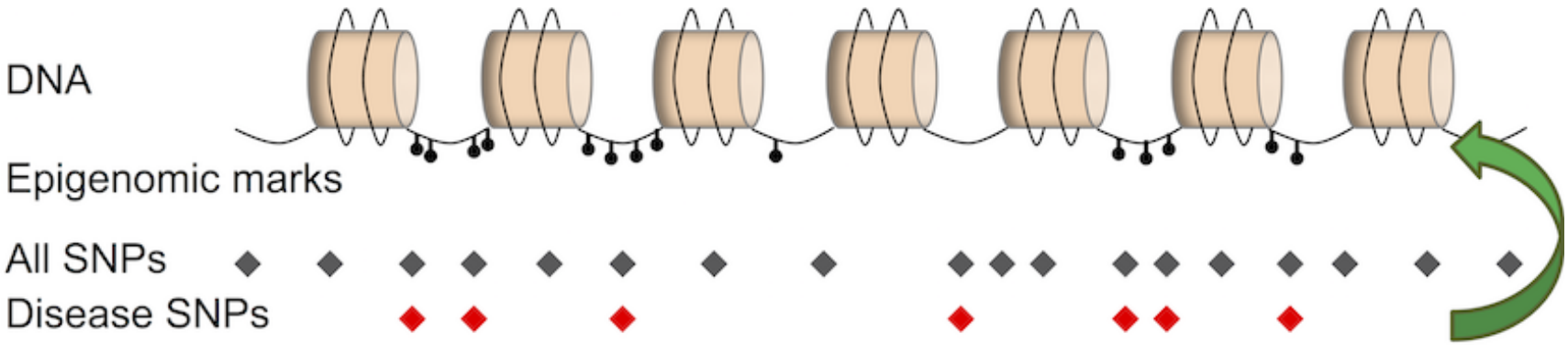


Regulatory marks are highly non-random

- Statistical analysis of pilot ENCODE regions showed highly non-random location of regulatory elements
- There are regulatory "hotspots" enriched in transcription factor binding sites, histone marks, as well as "deserts" of depleted regulatory marks
- Combinations of different types of regulatory marks matter

Zhang, Z. D., A. Paccanaro, Y. Fu, S. Weissman, Z. Weng, J. Chang, M. Snyder, and M. B. Gerstein. "Statistical Analysis of the Genomic Distribution and Correlation of Regulatory Elements in the ENCODE Regions." Genome Research 17, no. 6 (June 1, 2007): 787–97. https://doi.org/10.1101/gr.5573107.

Statistics of epigenomic enrichments



- 6 out of 7 disease-associated SNPs overlap with epigenomic marks
- How likely this to be observed by chance? (Chi-square test/Binomial test/Permutation test)

Basic concepts of epigenomic enrichments

TF1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
TF2	0	0	0	0	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1

- Pearson correlation coefficient r: this quantity gives equal weight to co-binding (1,1) and co-non-binding (0,0). Hence, high values may not necessarily imply high levels of co-occurrence. For the above example, $r = 0.36$.

Statistics of epigenomic enrichments

	TF2 no	TF2 yes	
TF1 no	n-k+t	m-t	m+n-k
TF1 yes	k-t	t	k
	n	m	m+n

- **Hypergeometric test**: it tests for co-occurrence based on the contingency table, which can be re-written using random variables
- Assume that the row and column sums (m, n, k) are fixed. The probability of observing t is hypergeometric. The p-value for the example is $p = Pr(T \geq 10 | H_0, m = 12, n = 8, k = 14) = 0.14$

Statistics of epigenomic enrichments

	TF2 no	TF2 yes	
TF1 no	n-k+t	m-t	m+n-k
TF1 yes	k-t	t	k
	n	m	m+n

- **Chi-square test**: it tests for dependence (not co-occurrence) between TF1 and TF2, and applies to contingency tables with very large counts
- The difference between observed and expected counts can be approximated by a chi-square distribution with one degree of freedom

$$D = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} are the observed counts, and E_{ij} are the expected counts under the null hypothesis, and are computed under the fixed row and column sums, e.g. $E_{22} = \frac{mk}{m+n}$

Statistics of epigenomic enrichments

- **Poisson distribution**: it can be used to compute how likely it is for a single TF to have, say, three binding events in 1 kb with 300 events in 1 Mb. The formula is

$$Pr(x = 3; L = 1kb; \rho = \frac{300}{1000kb}) = e^{L\rho}\frac{(L\rho)^x}{x!}$$

where ρ is the binding rate per bp.

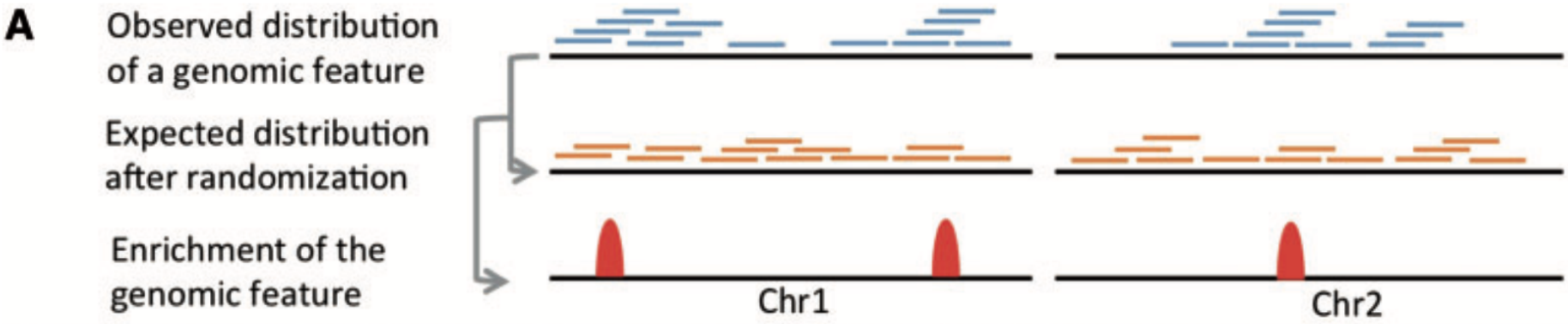
- **Fisher's method for combining p-values**: one can calculate a p-value for each TF in a genomic region to assess whether that TF has more binding sites than expected in this region. To assess whether both TFs bind to more sites than expected, p-values can be combined using Fisher's method

$$P = -2\sum_{i=1}^{n} logp_i$$

Permutation

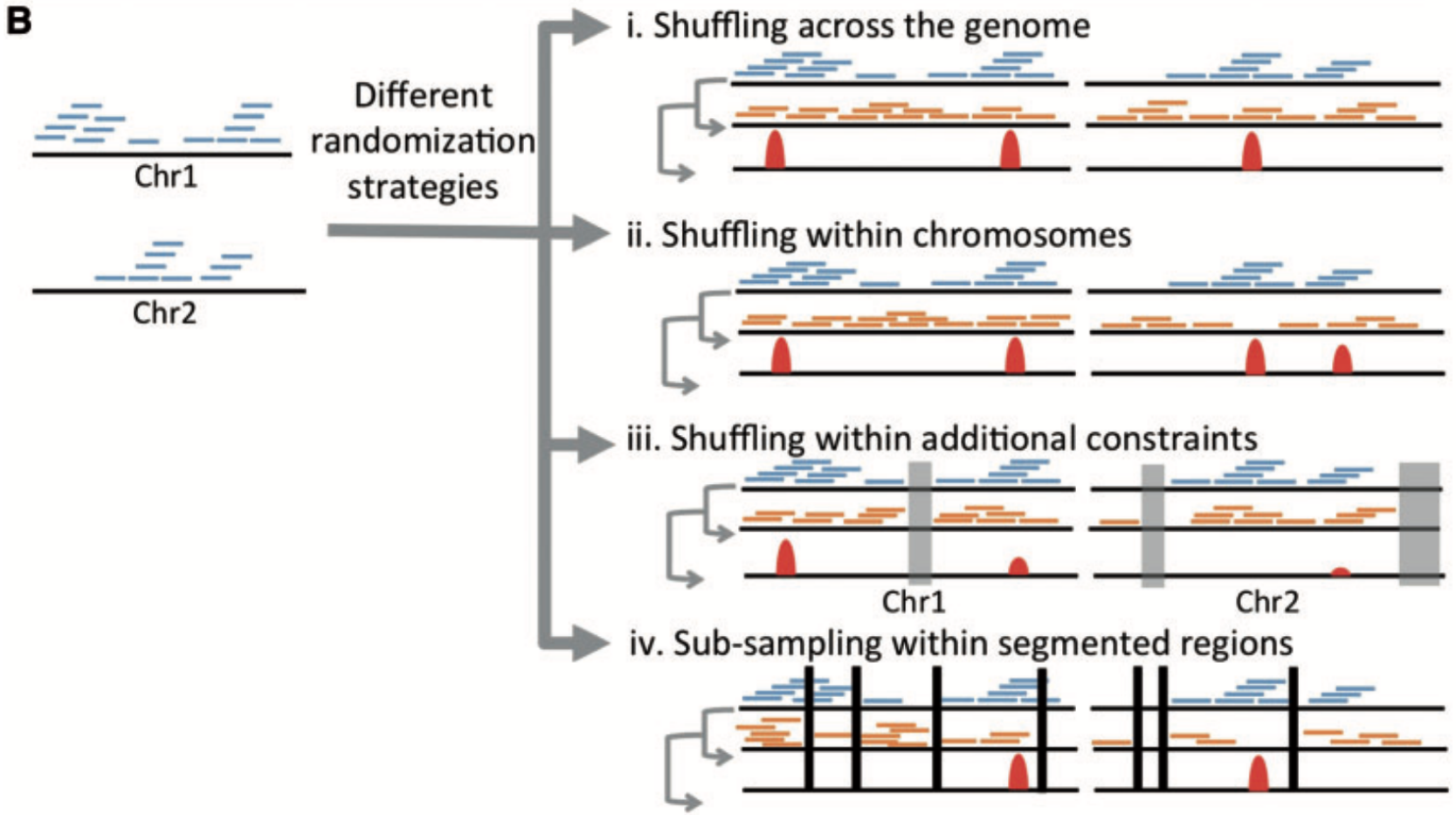
- Genomic features are nonrandomly distributed throughout the genome
- In permutation schemes, need to consider this to properly calculate observed and expected overlaps
- **Permutation test**: it tests for co-occurrence through repeatedly permuting observed enriched regions (or binding events) in one or both profiles many times
- A pre-defined co-occurrence score is calculated for each permutation
- Many permutations produce a null distribution of the co-occurrence score. One can then use this null distribution to compute a p-value for the observed co-occurrence score

Permutation



De, Subhajyoti, Brent S. Pedersen, and Katerina Kechris. "The Dilemma of Choosing the Ideal Permutation Strategy While Estimating Statistical Significance of Genome-Wide Enrichment." Briefings in Bioinformatics 15, no. 6 (November 2014): 919–28. https://doi.org/10.1093/bib/bbt053.

Permutation on steroids



Permutation strategies

Randomization method	Description	Advantage	Disadvantage
Genome-wide randomization	Shuffling one or more features unconstrained throughout the genome	Simple to implement. Assumes uniform distribution of features across the genome	Ignores chromosome-wide or local biases in the distribution
Chromosome-wide randomization	Shuffling one or more features unconstrained within respective chromosomes	Simple to implement. Accommodates chromosome-specific biases in the distribution	Ignores local or domain-level biases in the distribution
Randomization (dis)allowing overlaps	Overlap is allowed (or prohibited) among shuffled features on the genome	Biologically relevant in some scenarios (e.g. sites of amplification and deletions within a cancer genome cannot overlap)	Long run-time. Requires informed assumptions
Randomization with additional constraints	User-specific constraints are included in the model	Can accommodate case-specific biological or technical constraints	Long run-time. Requires informed assumptions
Randomization with fixed location model	Generating expected distribution by probabilistically sampling from the observed distribution	Biologically relevant in several scenarios (e.g. when analyzing transcription factor binding site co-occurrence)	Higher order organization of the features might be ignored.
Randomization with fixed locations fixed event type model	Shuffling location of the first feature, while keeping the location of the second feature unchanged	Preserves higher order structure of the second feature	The chromosome or domain-specific biases in the first feature are not considered
Randomization with sub-sampling accounting for genomic structure	Shuffling within respective segments	Highly powerful if correctly implemented. Segments can be generated based on sequence composition or biologically relevant assumptions	Potentially longer run time than others. Determining the segment boundaries is nontrivial

De, Subhajyoti, Brent S. Pedersen, and Katerina Kechris. "The Dilemma of Choosing the Ideal Permutation Strategy While Estimating Statistical Significance of Genome-Wide Enrichment." Briefings in Bioinformatics 15, no. 6 (November 2014):

Evaluating overlap between sets of genomic regions

Table 1 Methods for scoring overlapping and adjacent signals in two or more ChIP (or DamID) profiles. See text for details of these methods

Number of profiles under comparison	Accounting for spatial variability of events (Yes/No)	Method
Two	No	Simple counting[17,18,33,34]
		Pearson correlation coefficient[14,35–37]
		Hypothesis tests based on a single score
		Hypergeometric test[3,5,20,38]
		Chi-square test[36]
		Log-linear model[39]
		Permutation test[40–42]
	Yes	Poisson hierarchical model[43]
		Hidden Markov model[44]
		'Standard gene'[45]
Many	Yes	Overall assessment of co-occurrence
		Permutation test[4,46,47]
		Identification of 'co-localisation' hotspots:
		Multiple testing based on Poisson distribution[46]
		Clustering[14,37,49–51]
		Identification of *cis*-regulatory modules
		Factor regression[52]

Fu, Audrey Qiuyan, and Boris Adryan. "Scoring Overlapping and Adjacent Signals from Genome-Wide ChIP and DamID Assays." Molecular BioSystems 5, no. 12 (December 2009): 1429–38. https://doi.org/10.1039/B906880e.

Looking for significant GO enrichment

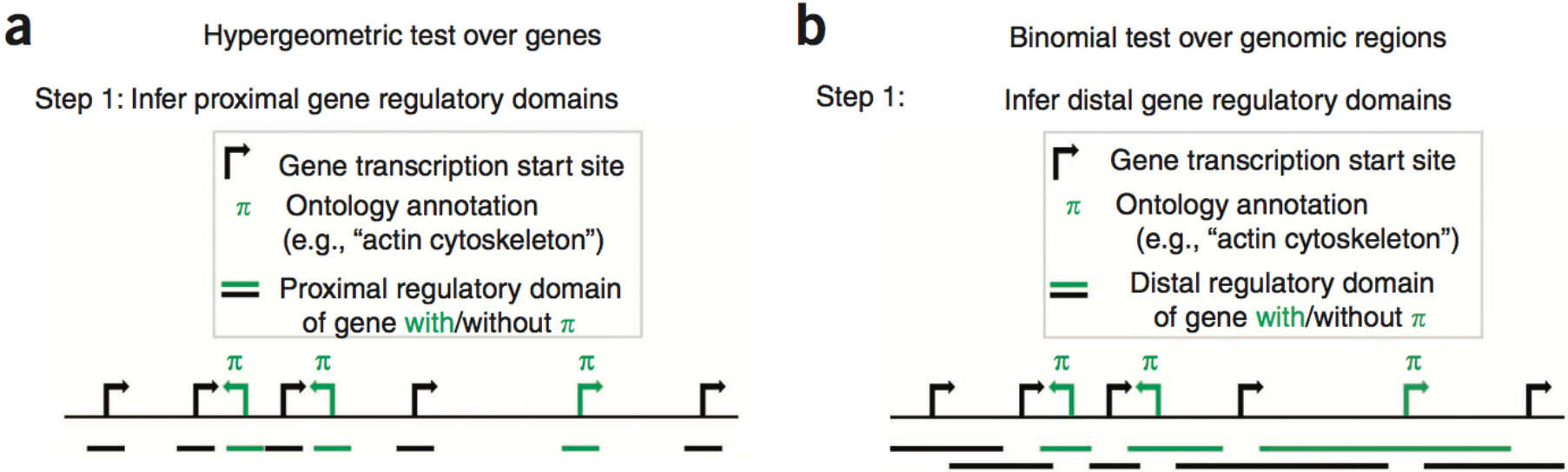
- We can look at biological significance of our peaks using Gene Ontologies (GO) terms genome annotations
  - GO: Set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences
- Popular tool: the Genomic Regions Enrichment of Annotations Tool (GREAT)

http://great.stanford.edu/public/html/

GREAT: Cis-regulatory regions functions prediction

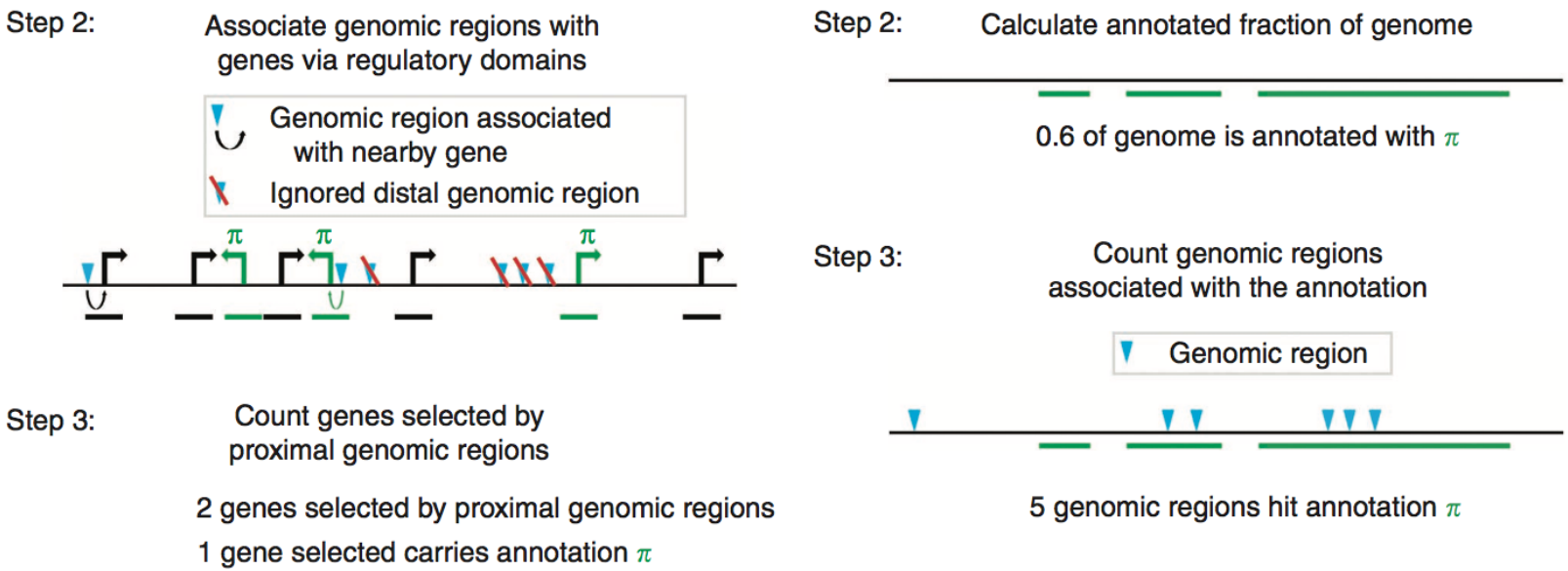
- Binding sites are often not located in the proximal region of the gene of interest
- GREAT looks beyond this proximal region
- Input: BED file with regions of interest
- Output: Matching GO terms for Molecular Functions, Biological Processes, Phenotypes, Diseases, etc.

GREAT: Cis-regulatory regions functions prediction



McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28, no. 5 (May 2010): 495–501. https://doi.org/10.1038/nbt.1630.

GREAT: Cis-regulatory regions functions prediction



McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28, no. 5 (May 2010): 495–501. https://doi.org/10.1038/nbt.1630.

GREAT: Cis-regulatory regions functions prediction

Step 4: Perform hypergeometric test over genes

$N = 8$ genes in genome

$K_\pi = 3$ genes in genome carry annotation π

$n = 2$ genes selected by proximal genomic regions

$k_\pi = 1$ gene selected carries annotation π

$P = \Pr_{hyper}(k \geq 1 \mid N = 8, K = 3, n = 2)$

Step 4: Perform binomial test over genomic regions

$n = 6$ total genomic regions

$p_\pi = 0.6$ fraction of genome annotated with π

$k_\pi = 5$ genomic regions hit annotation π

$P = \Pr_{binom}(k \geq 5 \mid n = 6, p = 0.6)$

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28, no. 5 (May 2010): 495–501. https://doi.org/10.1038/nbt.1630.