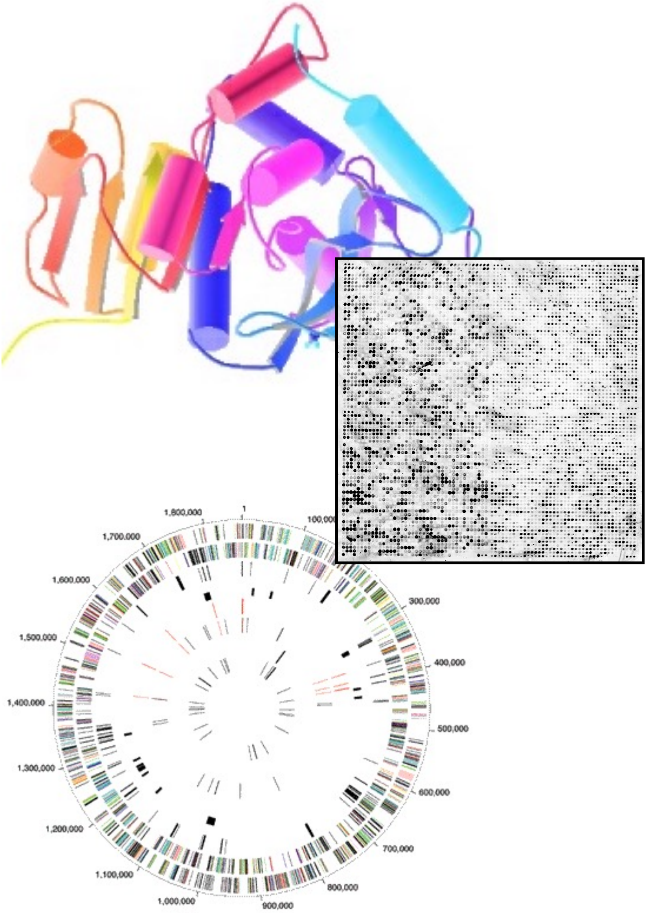
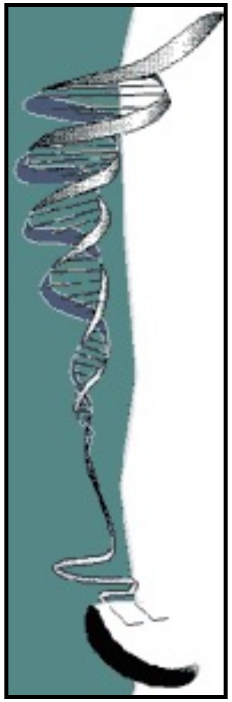
Biomed. Data Science:

Basic Multi-omic Analyses



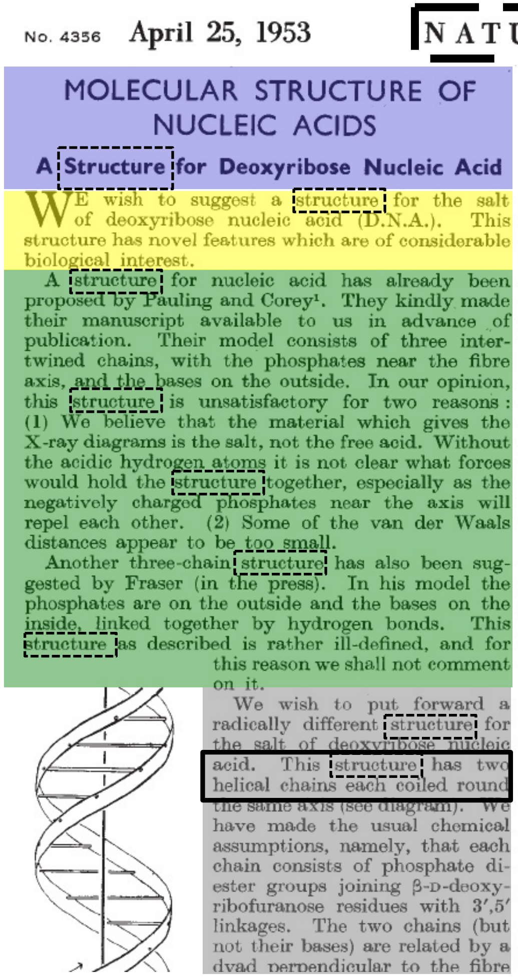


Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’21, final)

What is Annotation? (For Written Texts?)



No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

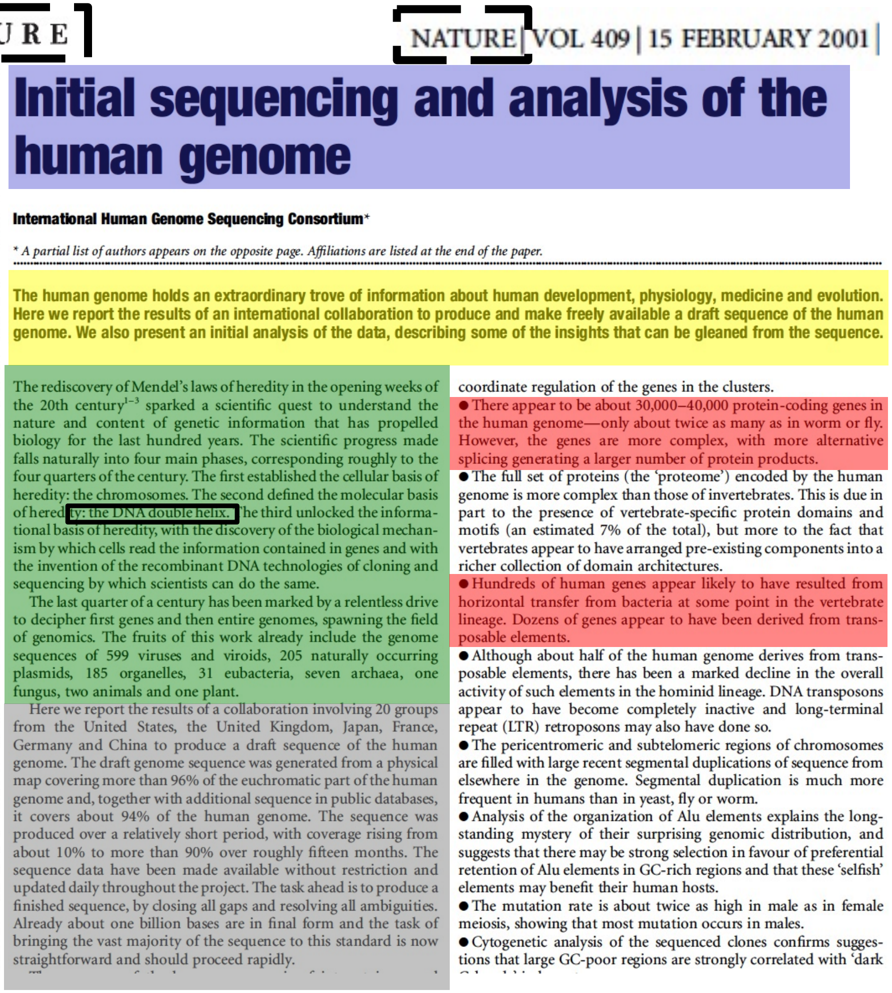
A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre



NATURE | VOL 409 | 15 FEBRUARY 2001 |

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

coordinate regulation of the genes in the clusters.

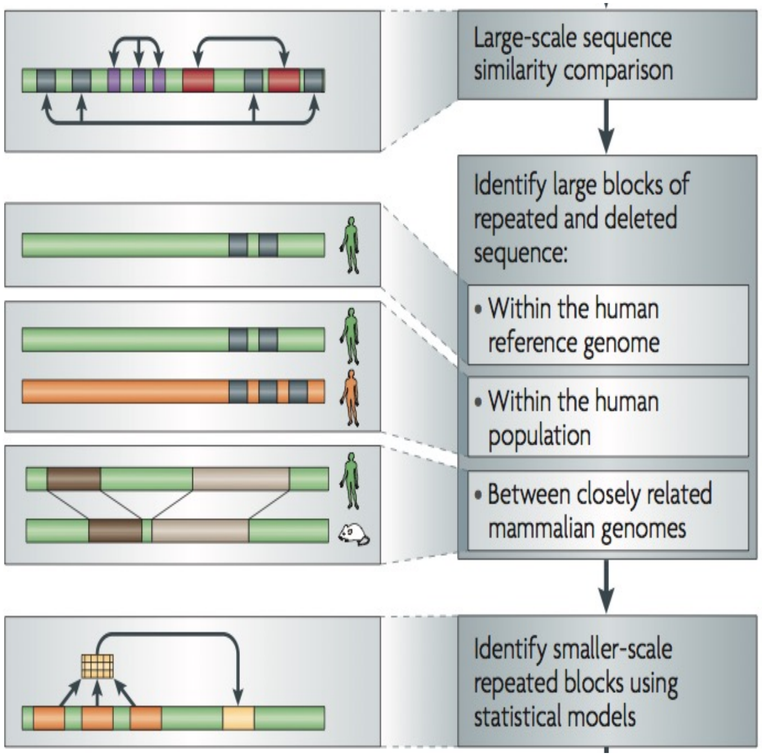
- ● There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- ● The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- ● Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- ● Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- ● The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- ● Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- ● The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- ● Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark

Non-coding Annotations: Overview

Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**

Functional Genomics
Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription

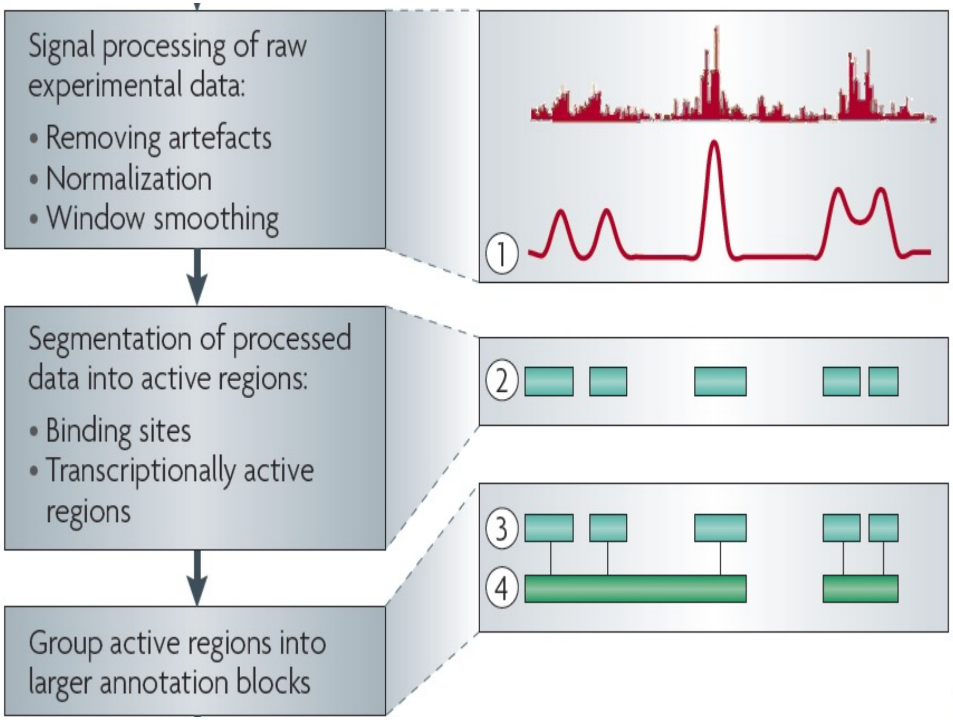


Large-scale sequence similarity comparison

Identify large blocks of repeated and deleted sequence:

- Within the human reference genome
- Within the human population
- Between closely related mammalian genomes

Identify smaller-scale repeated blocks using statistical models



Signal processing of raw experimental data:

- Removing artefacts
- Normalization
- Window smoothing

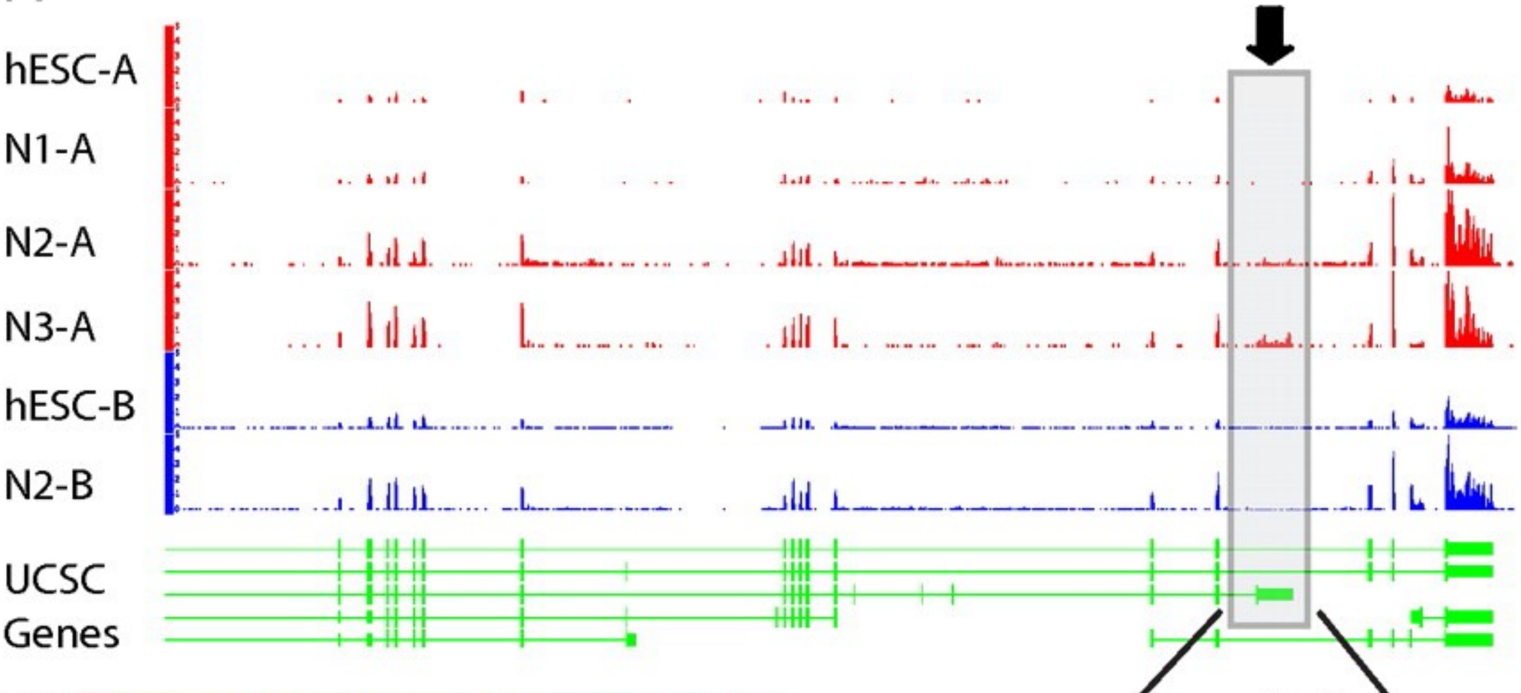
Segmentation of processed data into active regions:

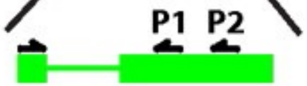
- Binding sites
- Transcriptionally active regions

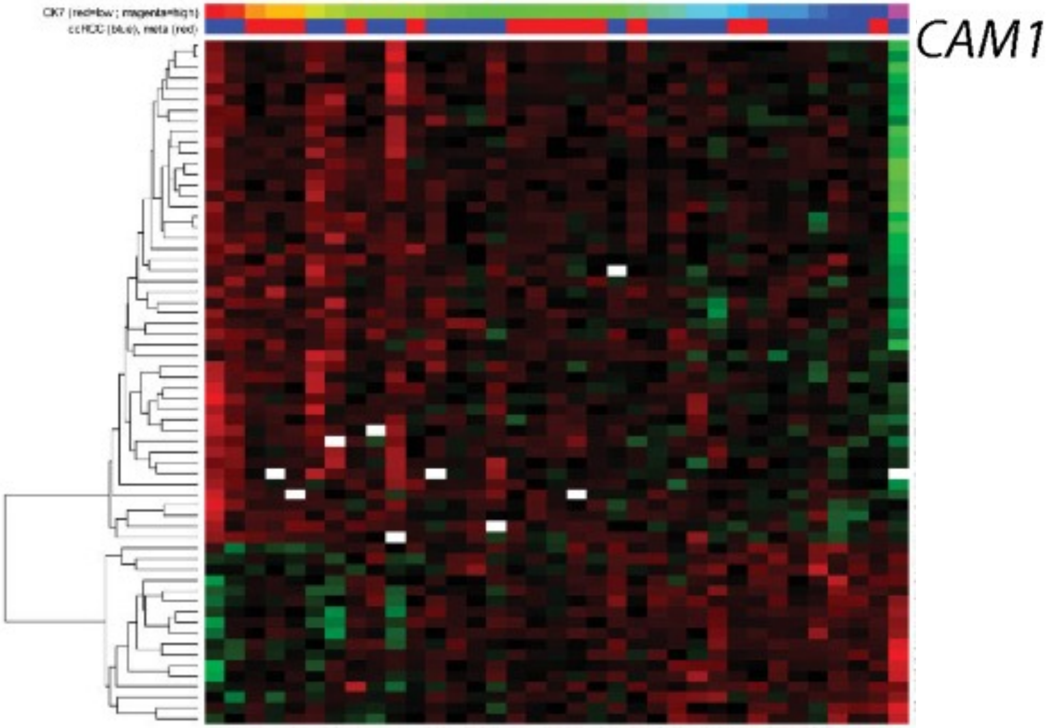
Group active regions into larger annotation blocks

[*Nat. Rev. Genet.* (2010) 11: 559]

RNA-seq



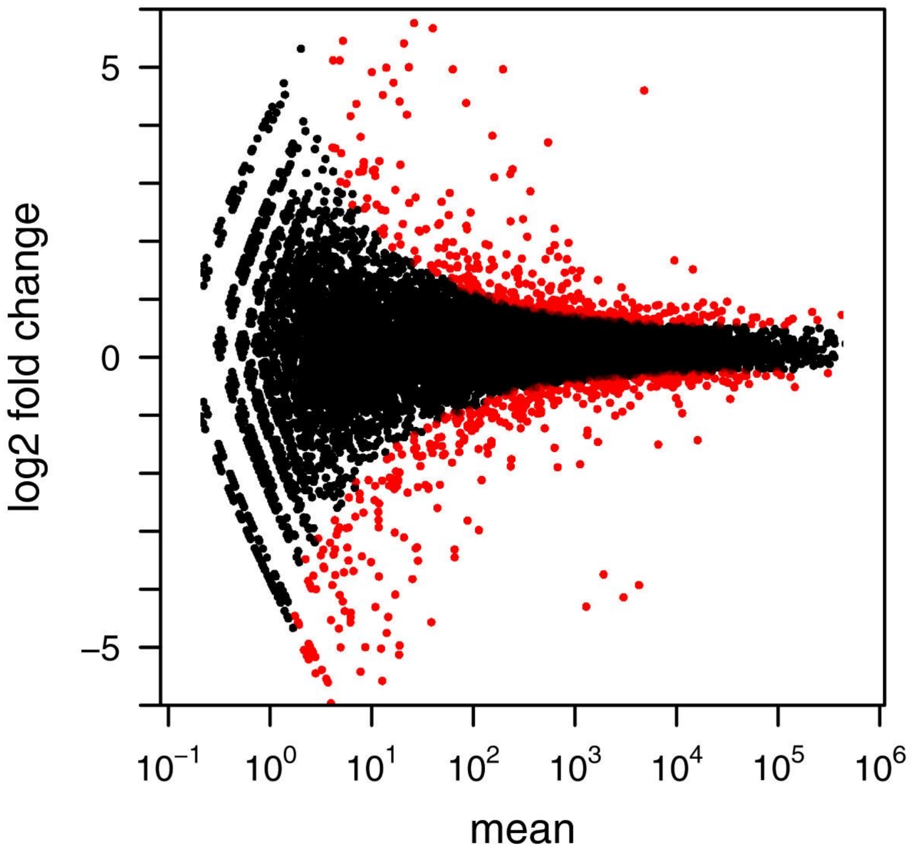




Information from RNA-seq: Avg. signal at exons & TARs (RPKMs)

[*PNAS* 4:107: 5254 ; *IJC* 123:569]

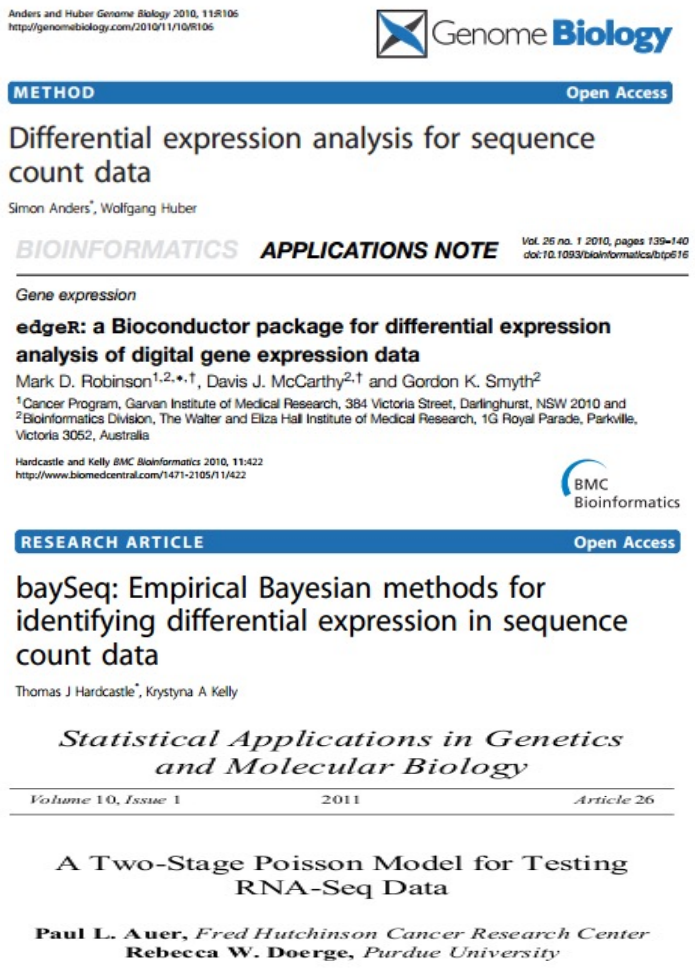
Differential expression analysis



Genome Biology, 2010 11:R106

Differential expression analysis: Count-based

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage poisson model



Anders and Huber *Genome Biology* 2010, 11:R106
http://genomebiology.com/2010/11/10/R106

Genome **Biology**

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 26 no. 1 2010, pages 139–140*
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1] Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Hardcastle and Kelly *BMC Bioinformatics* 2010, 11:422
http://www.biomedcentral.com/1471-2105/11/422

BMC Bioinformatics

RESEARCH ARTICLE Open Access

baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Thomas J Hardcastle*, Krystyna A Kelly

Statistical Applications in Genetics and Molecular Biology

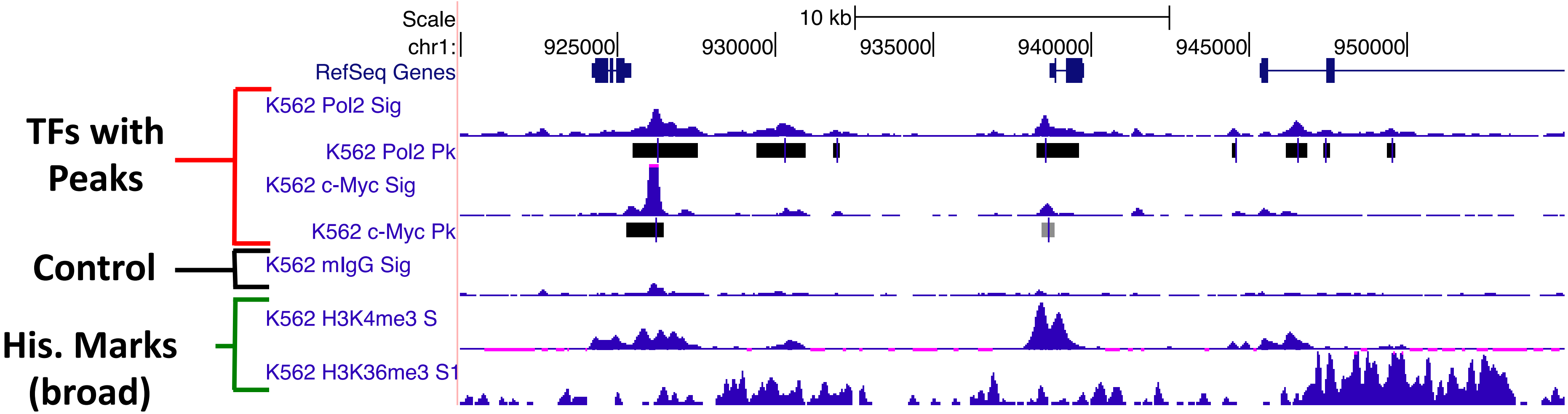
Volume 10, *Issue* 1 2011 *Article* 26

A Two-Stage Poisson Model for Testing RNA-Seq Data

Paul L. Auer, *Fred Hutchinson Cancer Research Center*
Rebecca W. Doerge, *Purdue University*

Chip-seq

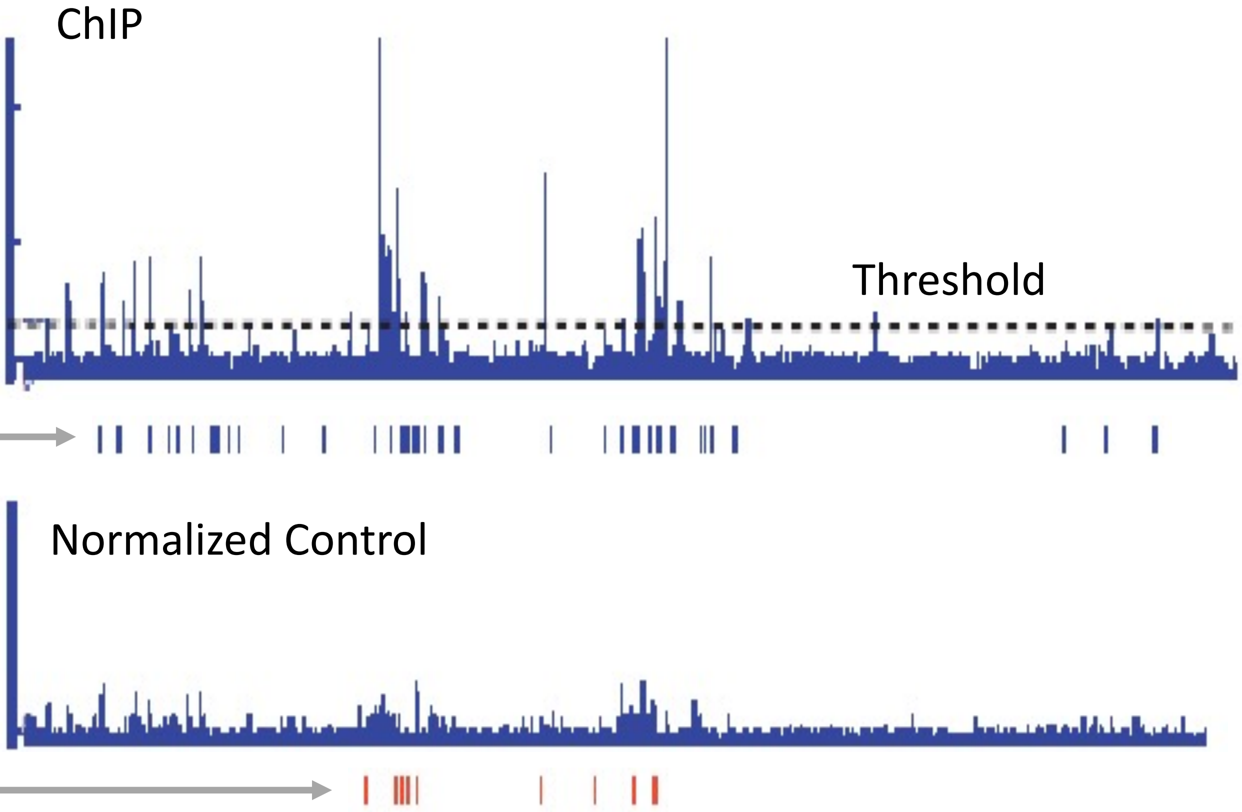
Information from Chip-seq



[*Science* 330: 1775
+ ENCODE Data
Sources
TFs & Control: Yale
HMs: UW & Broad]

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window-based Poisson (MACS),
  - Fold change statistics (SPP)



ChIP

Threshold

Potential Targets

Normalized Control

- Score against the control

Significantly Enriched targets

[Rozowsky et al. ('09) *Nat Biotech*]

Data Flow: Chip-seq expts. to co-associating peaks

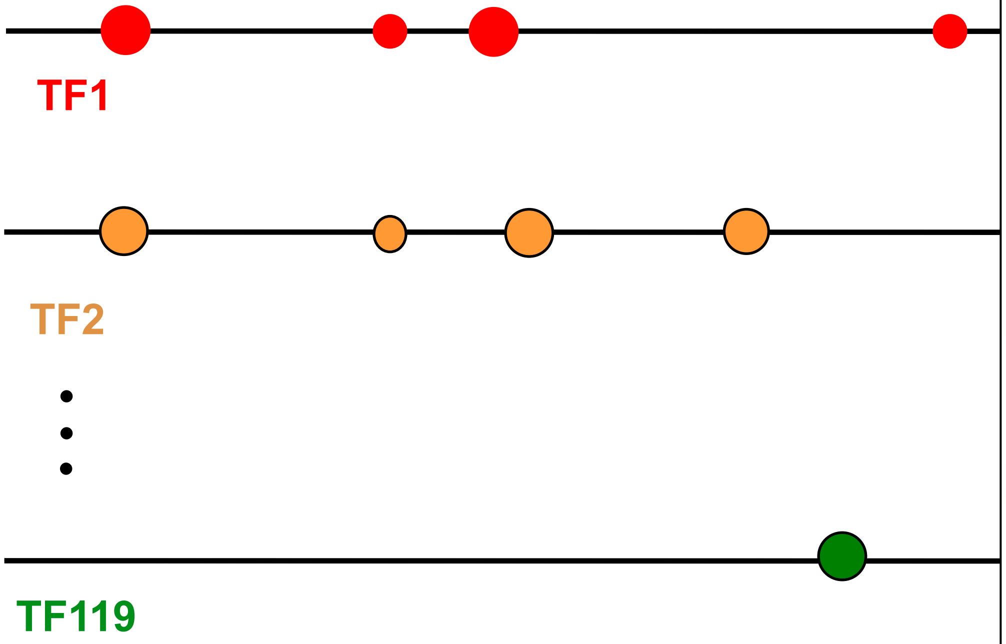
119 TFs from 458 ChIP-Seq experiments (2 Tb tot.)

↓

Signal Tracks

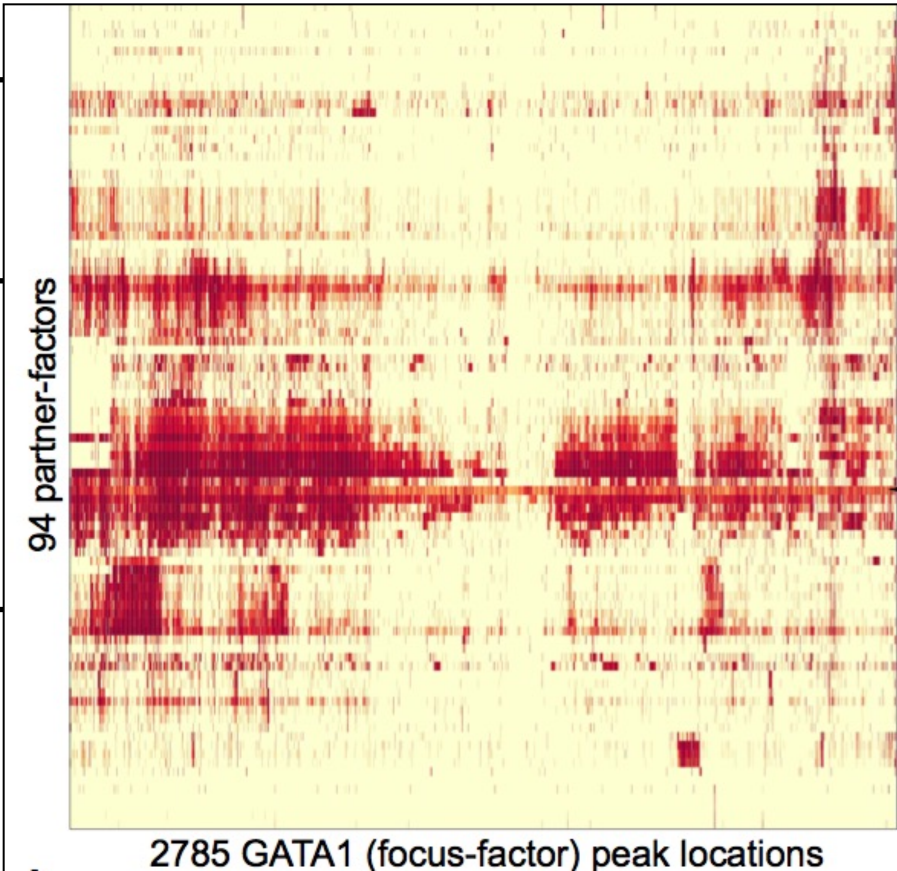
↓

7M Peaks from Uniform Peak Calling



- Mostly in Tier 1 cell lines
  - K562, GM12878, H1h-ESC…
- Matching RNA-Seq data in all cell-lines

- SPP & PeakSeq
- thresholding w. IDR (replicas)

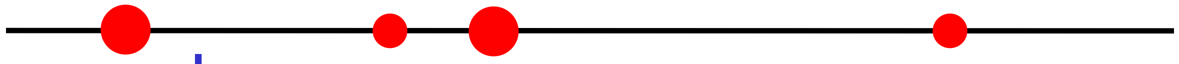


[Gerstein et al. Nature (in press, '12)]

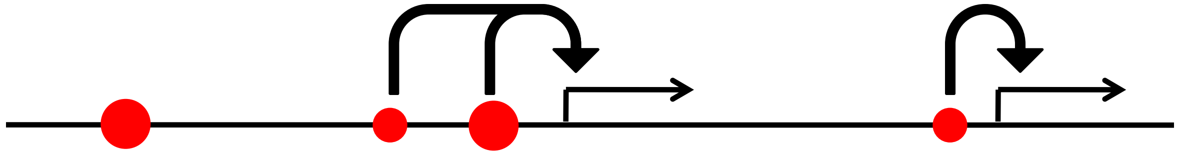
Data Flow: peaks to proximal & distal networks

[Cheng et al., *Bioinfo.* ('11);
Gerstein et al. *Nature* (in press, '12) ;
Yip et al., *GenomeBiology* (in press, '12)]

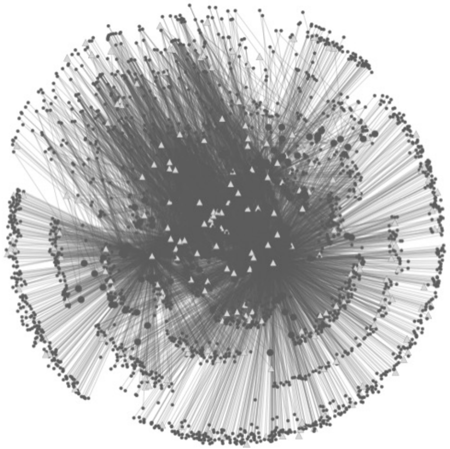
Peak Calling



Assigning TF binding sites to targets

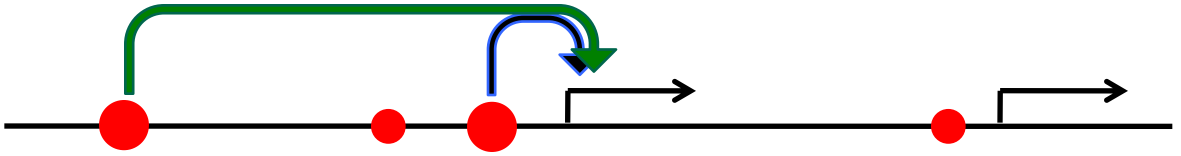


~500K
Edges

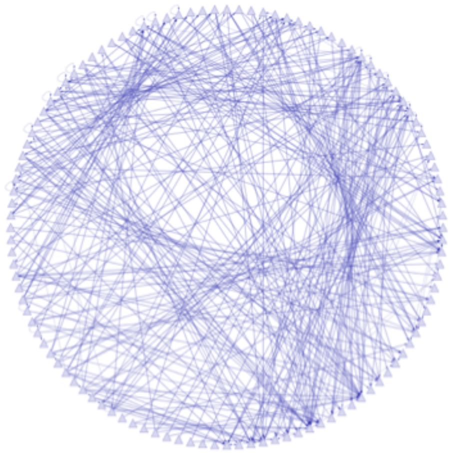


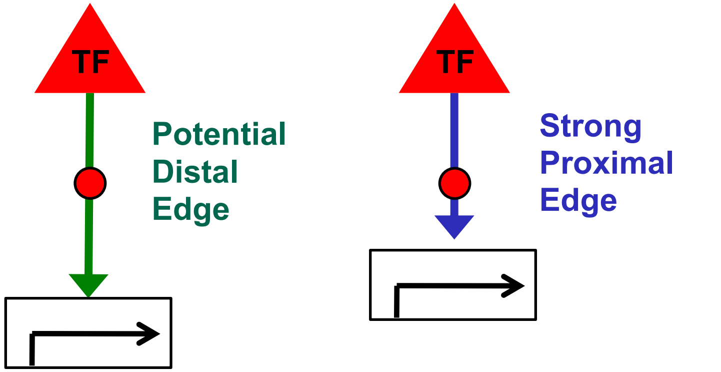
Filtering high confidence edges & distal regulation

Based on stat. model combining
signal strength & location relative to typical binding



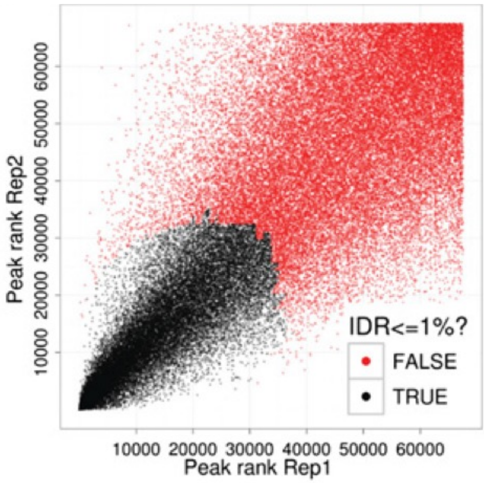
~26K
Edges

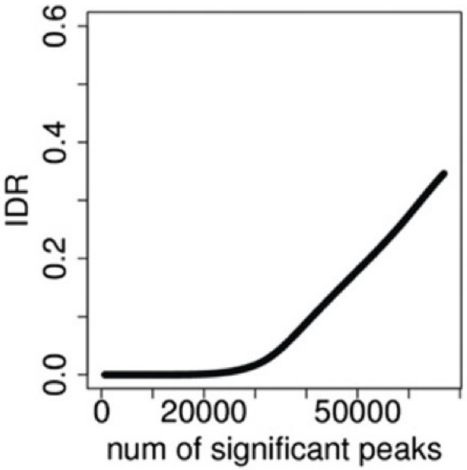




The irreproducible discovery rate (IDR)

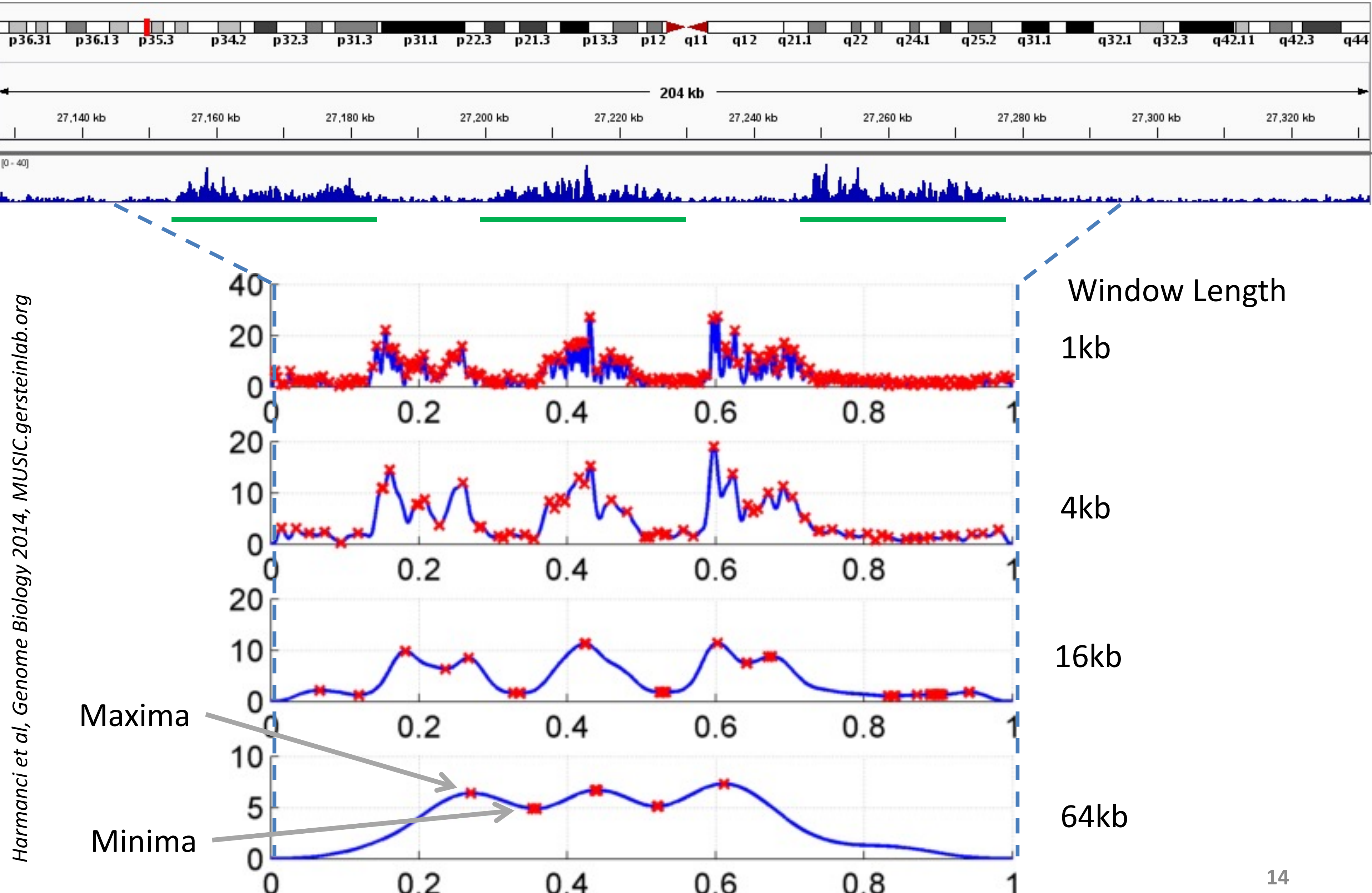
- Unified approach to measure the reproducibility of findings identified from replicate high-throughput experiments.
- Idea : call peaks with low cutoff and classify peaks as reproducible or not (bivariate rank distributions) based on overlap of ranked peaks (consistency)





Genome Research 2012, 22:1813-1831

Multiscale Analysis, Minima/Maxima based Coarse Segmentation



Window Length

1kb

4kb

16kb

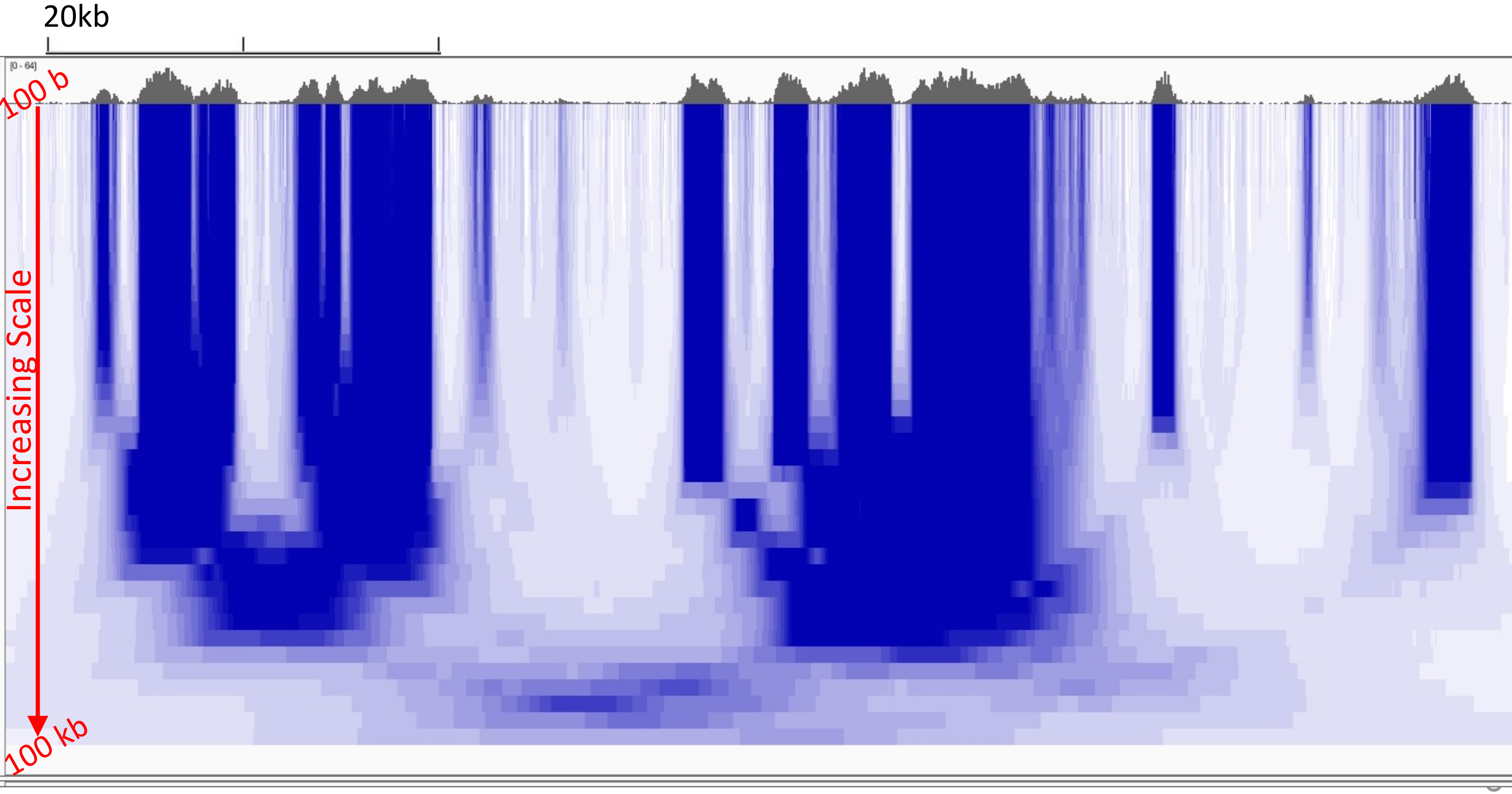
64kb

Maxima

Minima

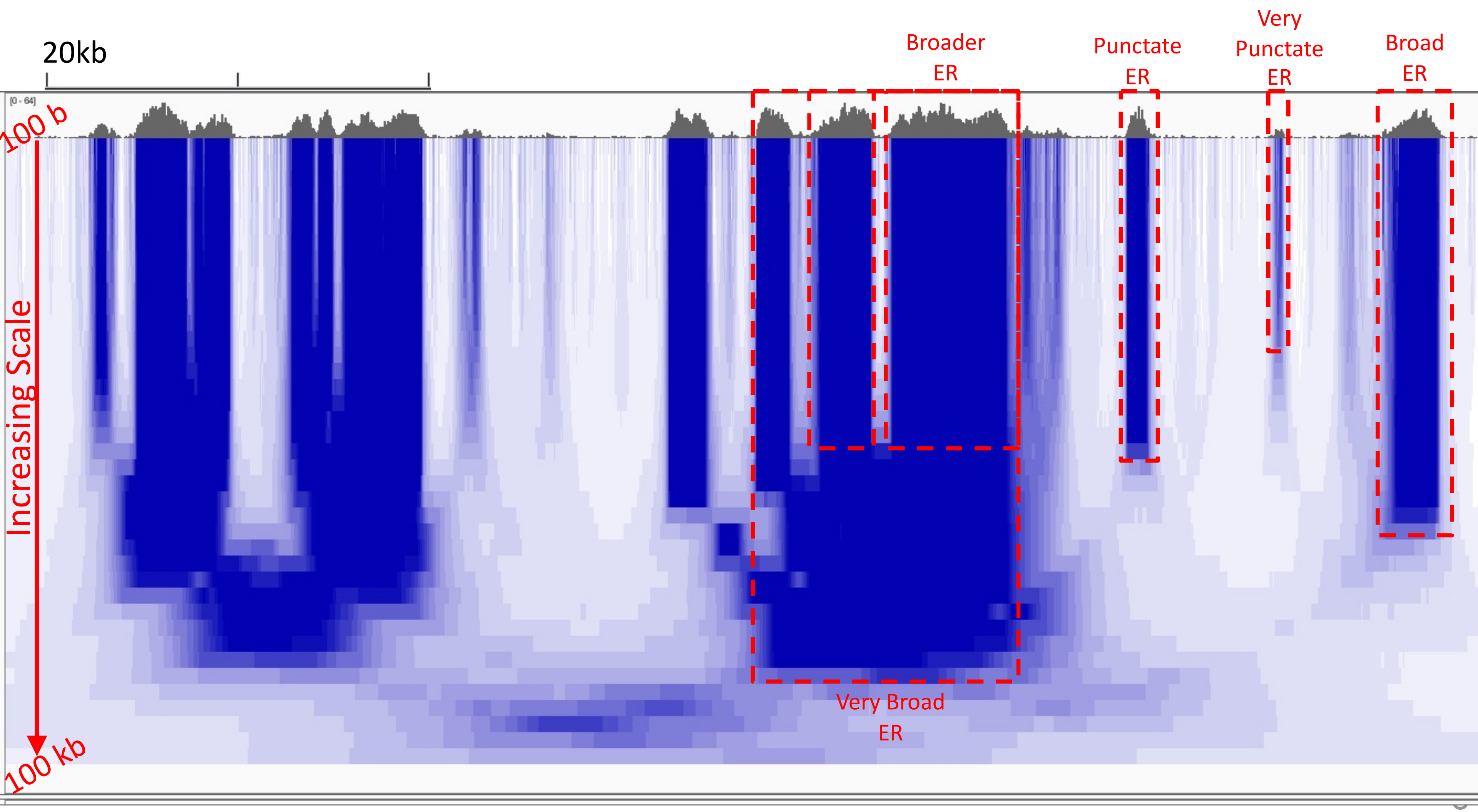
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



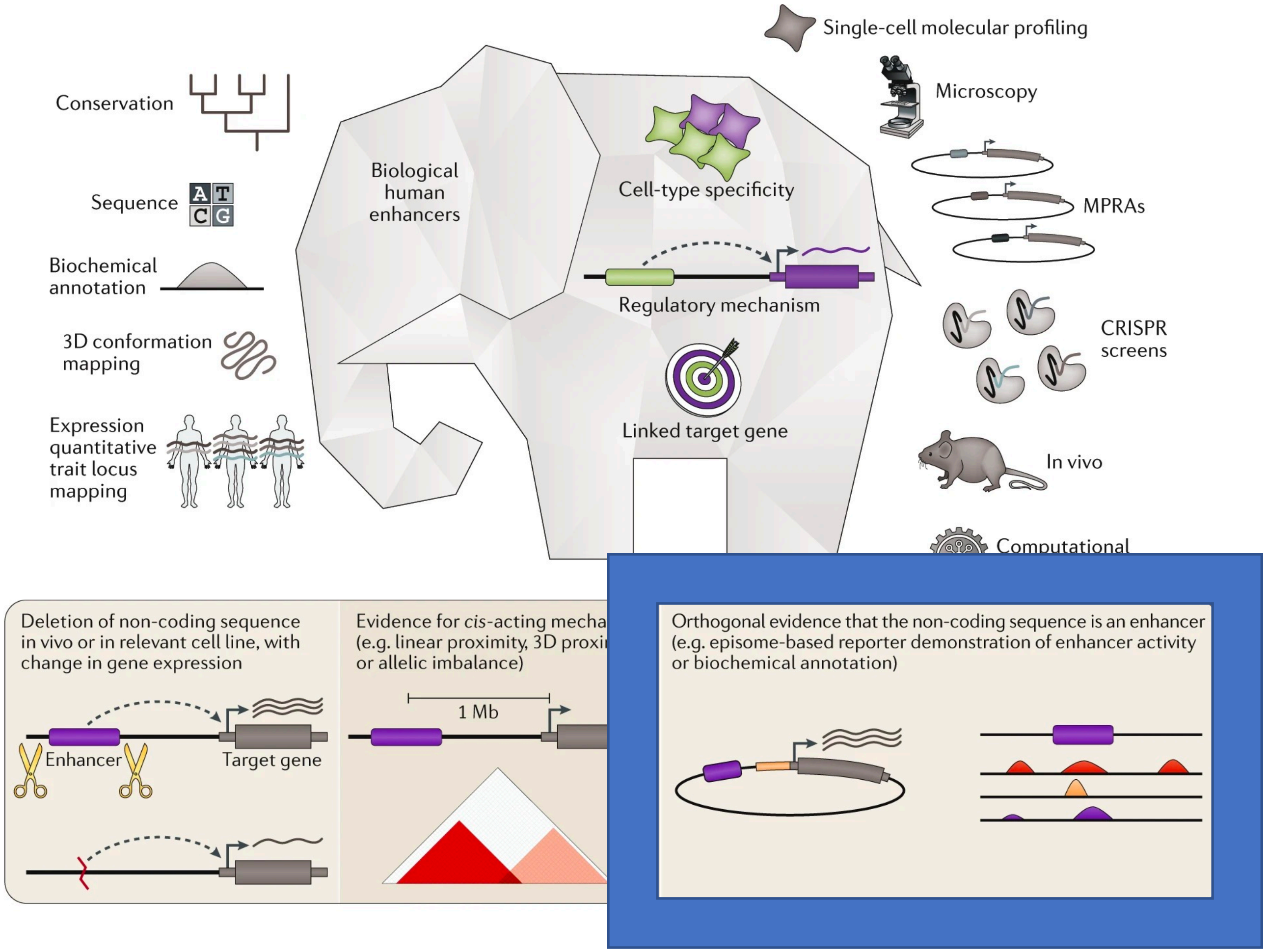
[Harmanci *et al, Genome Biol.* ('14)]

Multiscale Decomposition



[Harmanci *et al, Genome Biol.* ('14)]

Simple Integration for Elements



Deletion of non-coding sequence in vivo or in relevant cell line, with change in gene expression

Evidence for *cis*-acting mecha
(e.g. linear proximity, 3D proxi
or allelic imbalance)

Orthogonal evidence that the non-coding sequence is an enhancer (e.g. episome-based reporter demonstration of enhancer activity or biochemical annotation)

Background on computational annotation for non-coding regions

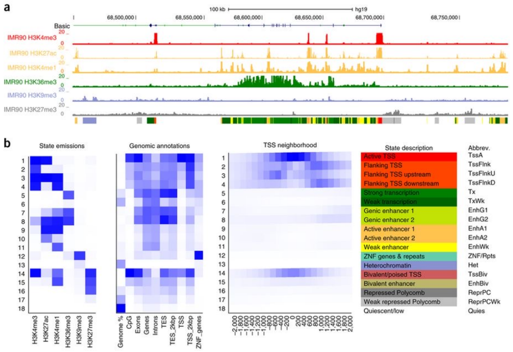
- **Peak calling**:
  - ✓PeakSeq, SPP, MACS2, Hotspot …
  - ✓ENCODE Encyclopedia

- **Genome segmentation**: partition the genome into regions (states) with distinct epigenomic profiles, then assign each state a functional label.
  - ✓ChromHMM: Multivariate Hidden Markov Model
  - ✓Segway: Dynamic Bayesian Network Model

- Supervised regulatory prediction: learn predictive models from labeled dataset of regulatory elements.
  - ✓ CSI-ANN: Time-Delay Neural Network
  - ✓ RFECS: Random Forest
  - ✓ DEEP: Ensemble SVM + Artificial Neural Network
  - ✓ REPTILE: Random Forest
  - ✓ gkm-SVM: Gapped k-mer

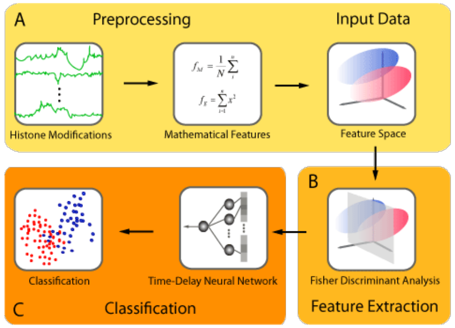
- **Target finding**
  - ✓ Ripple, TargetFinder, JEME, PreSTIGE, IM-PET

ChromHMM



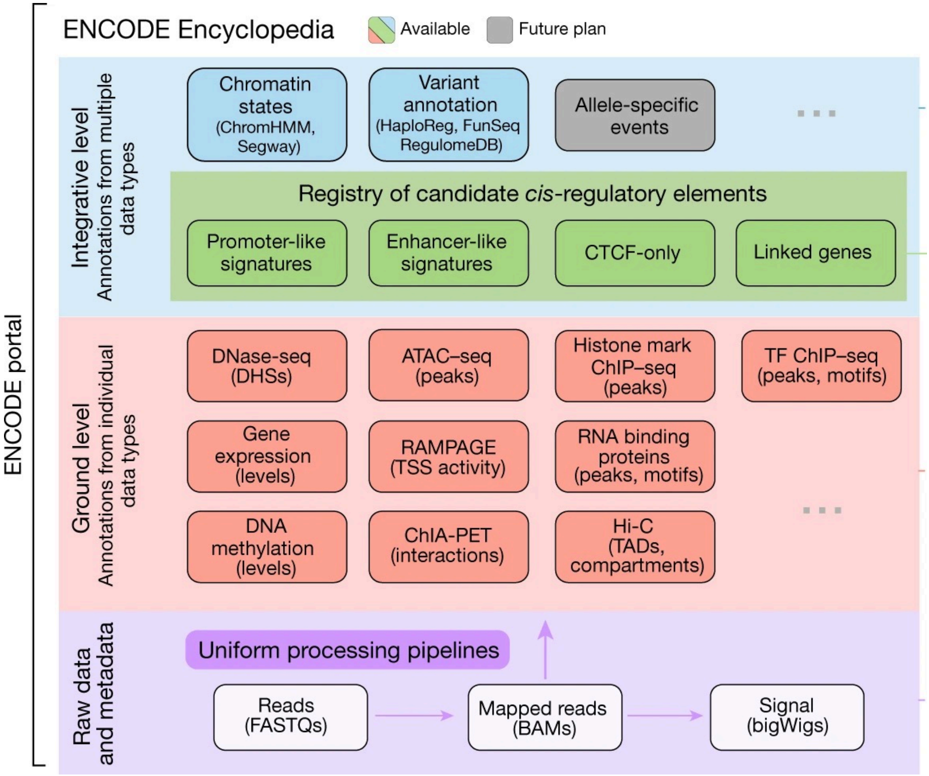
J. Ernst, M. Kellis. *Nat. Protoc., 2017*

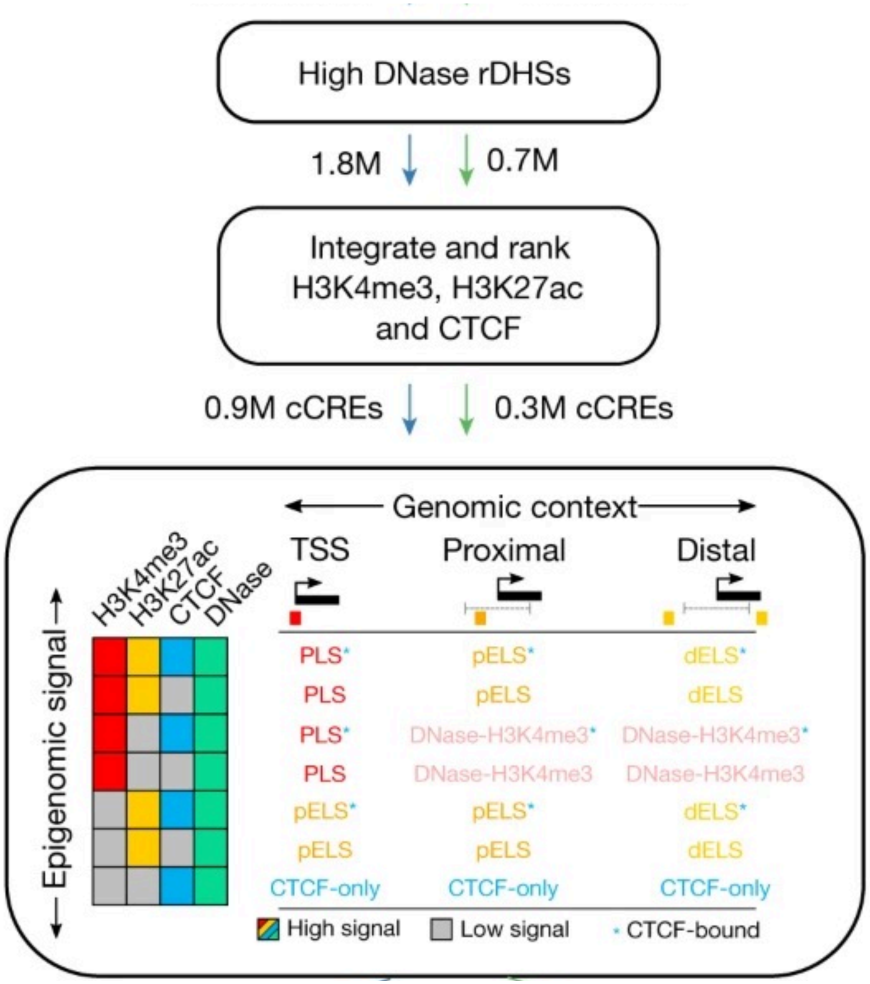
CSI-ANN



H.A. Firpi, D. Ucar, K. Tian. Bioinformatics, 2010

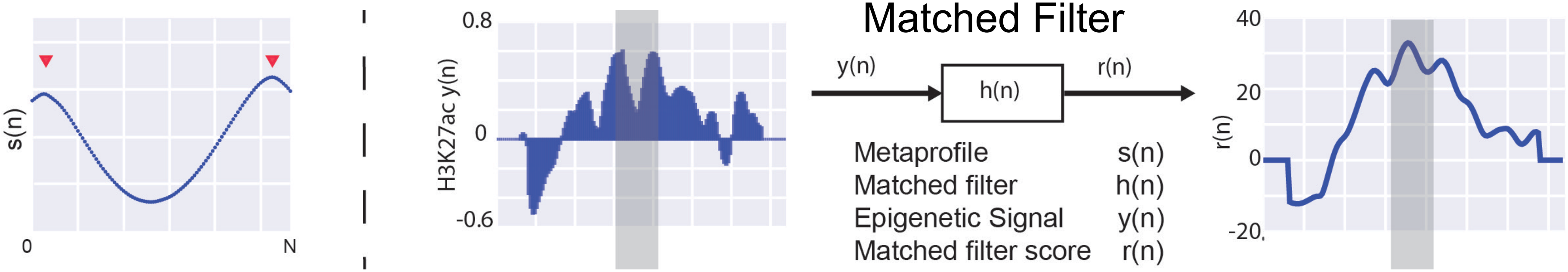
Broad ENCODE Annotation





[Enocode Consortium et al. Nature ('20)]

Matched Filter recognize shape patterns

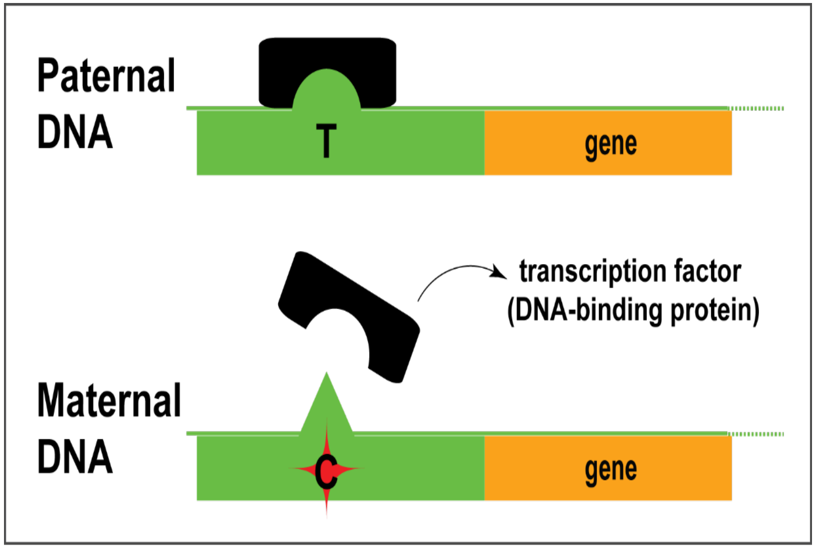


Metaprofile	s(n)
Matched filter	h(n)
Epigenetic Signal	y(n)
Matched filter score	r(n)

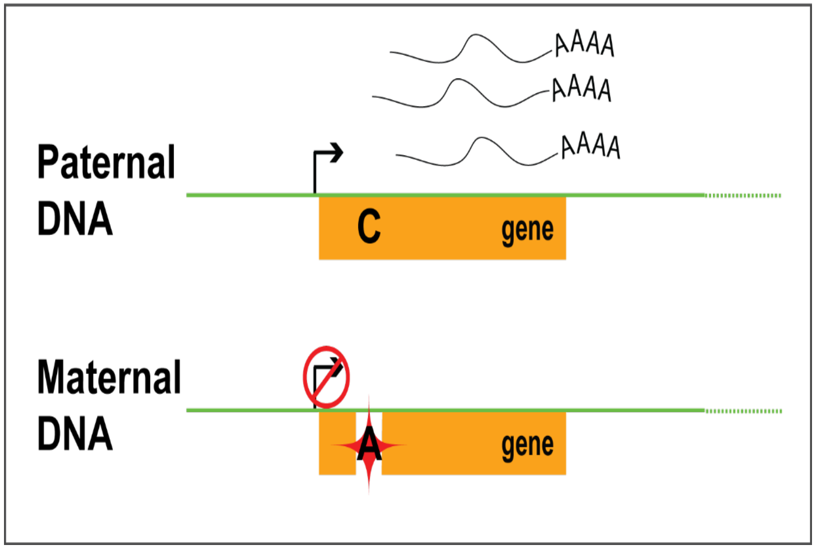
[biorxiv.org/content/early/2018/08/05/385237; Sethi et al. ('20) *Nat. Meth.*]

ASB/ASE & eQTL

Allele-specific binding and expression



Genomic variants
affecting allele-specific behavior
e.g. allele-specific binding
(ASB)



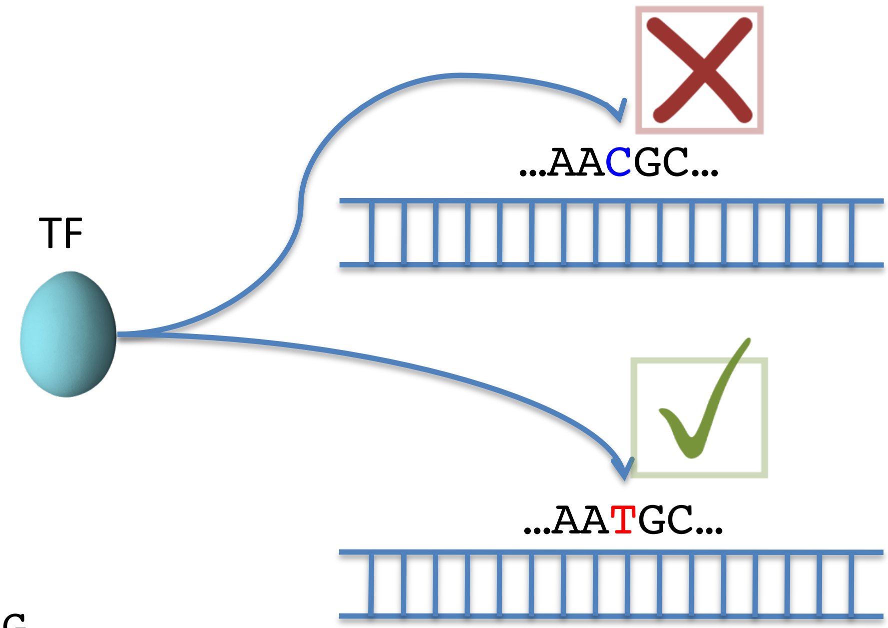
e.g. allele-specific expression
(ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                TGCACGTTGGGAGTTGGC

           10 x T
            2 x C
```



Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants

Many Technical Issues in Determining ASE/ASB: Reference Bias (naïve alignment against reference)

ASE/ASB Example:

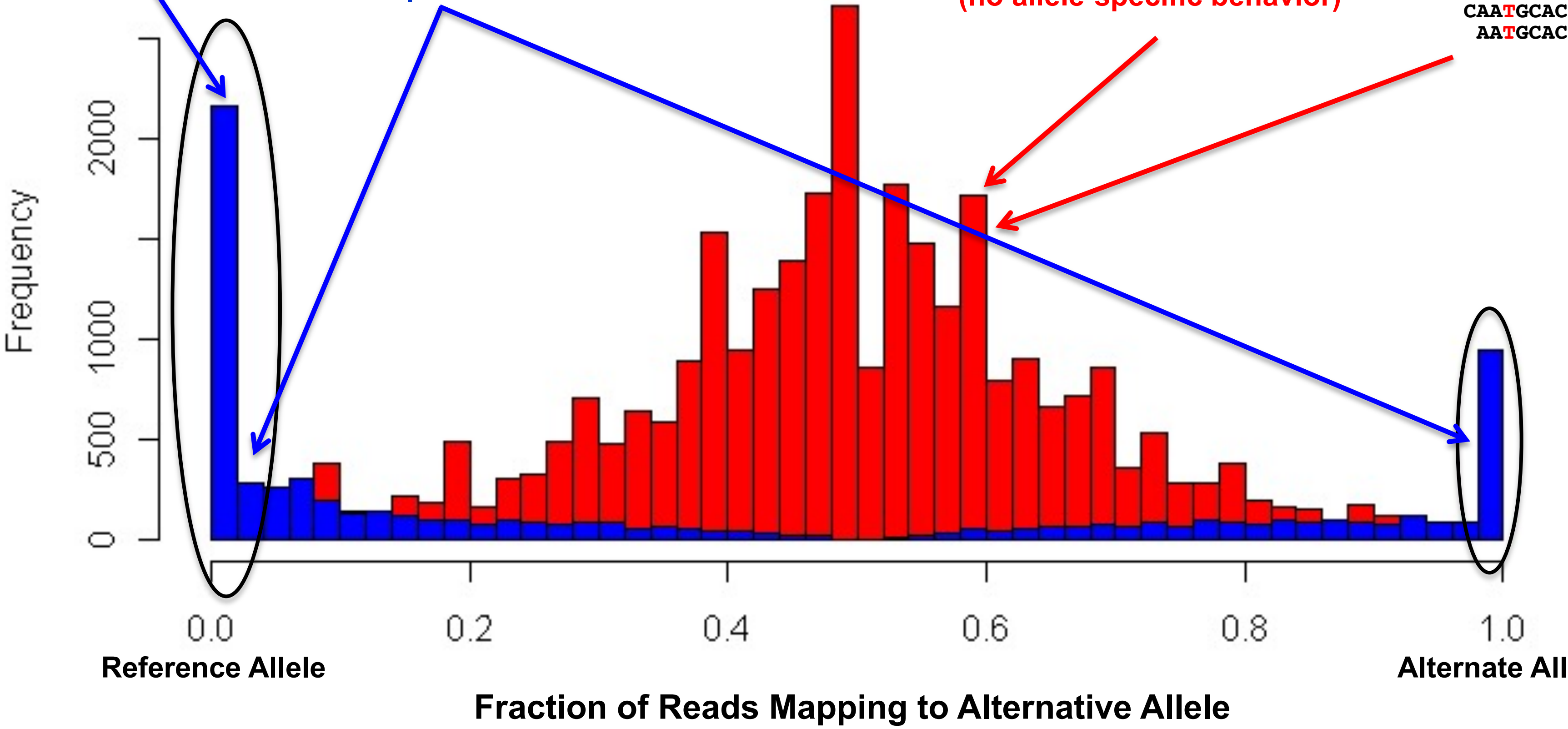
```
…GTCAATGCAC
…GTCAATGCACG
…GTCAATGCACGTC
…GTCAATGCACGTCG
…GTCAACGCACGTCGGGA
 GTCAATGCACGTCGAGAG
    CAATGCACGTCGGGAGTT
      AATGCACGTCGGGAGTTG
```

Null Example:

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
      ATAGCGTCAATGCACGT…
       TAGCGTCAACGCACGT…
          CGTCAACGCACGT…
             CAATGCACGT…
              AATGCACGT…
```

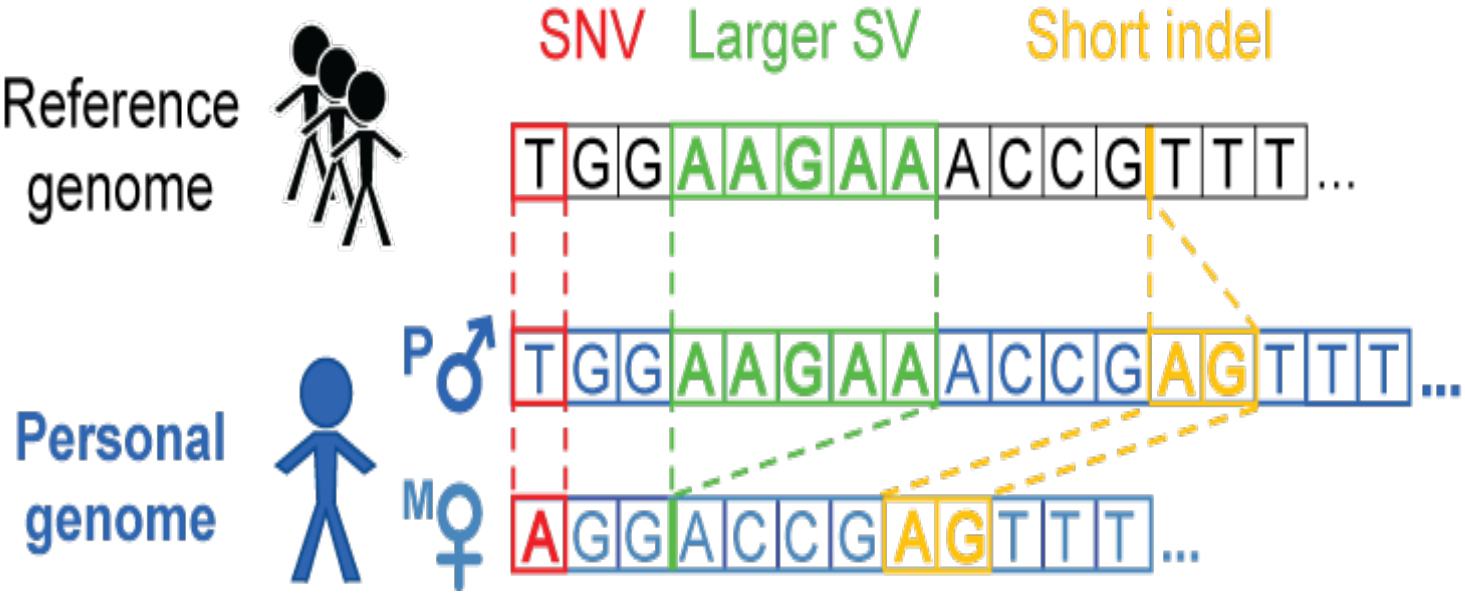
Allele-Specific SNPs

Binomial Null Distribution (no allele-specific behavior)



[Rozowsky et al., MSB ('11)]

How to build a personal genome



(fasta; reference)

+

(vcf, variants
phased or unphased)

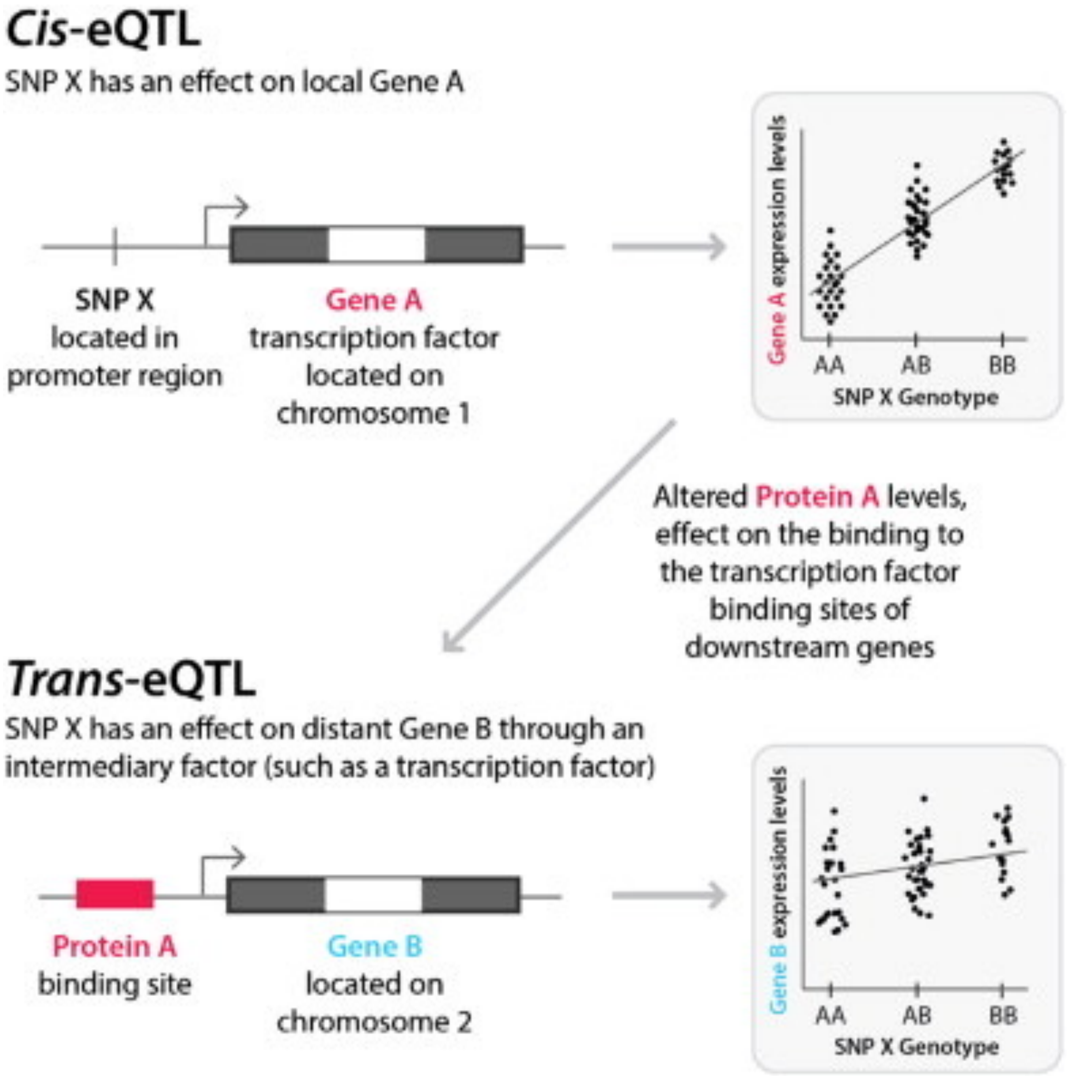
↓

(fasta; for each
haplotype)

alleleseq.gersteinlab.org

Rozowsky *et al. Mol Syst Biol* (2011)

Expression quantitative trait



Cis-eQTL

SNP X has an effect on local Gene A

SNP X located in promoter region

Gene A transcription factor located on chromosome 1

Altered Protein A levels, effect on the binding to the transcription factor binding sites of downstream genes

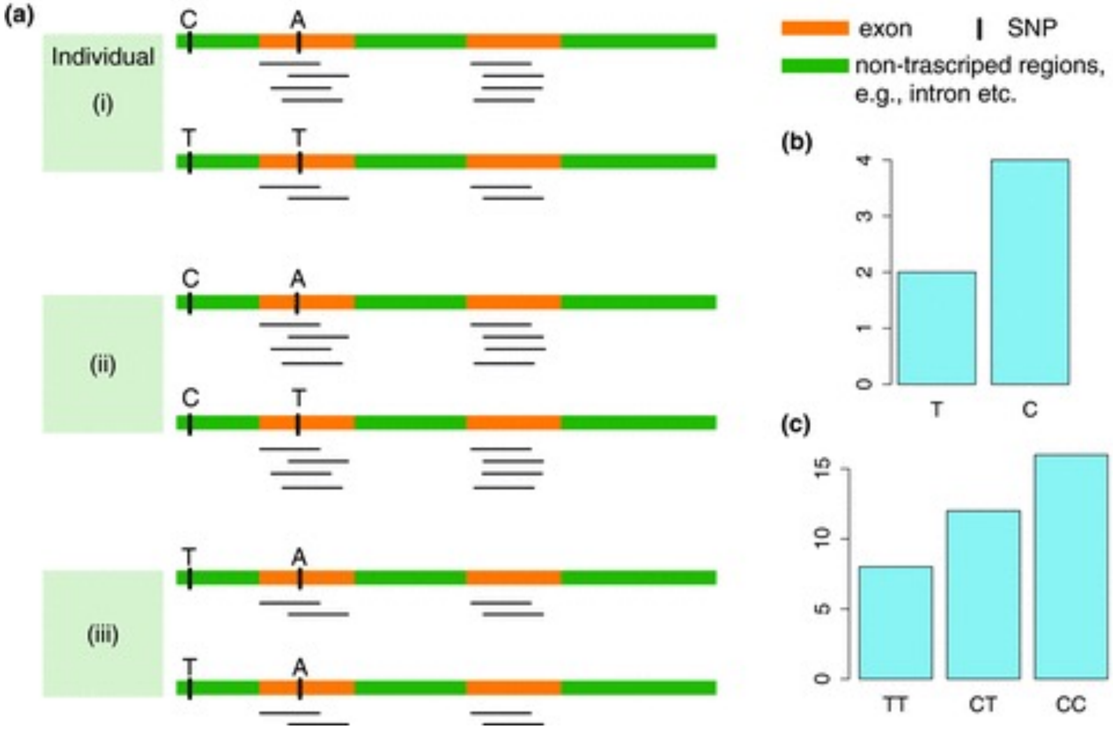
Trans-eQTL

SNP X has an effect on distant Gene B through an intermediary factor (such as a transcription factor)

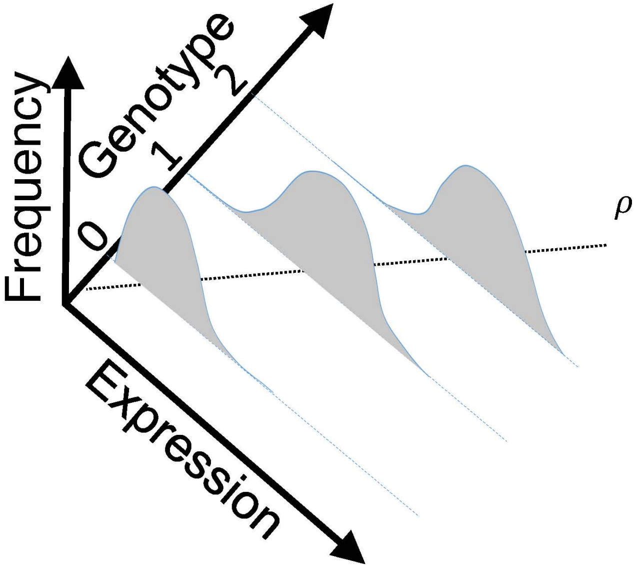
Protein A binding site

Gene B located on chromosome 2

Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2014, 10:1896-1902



[*Biometrics 68(1) 1–11*]



eQTL Mapping Using RNA-Seq Data

- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data