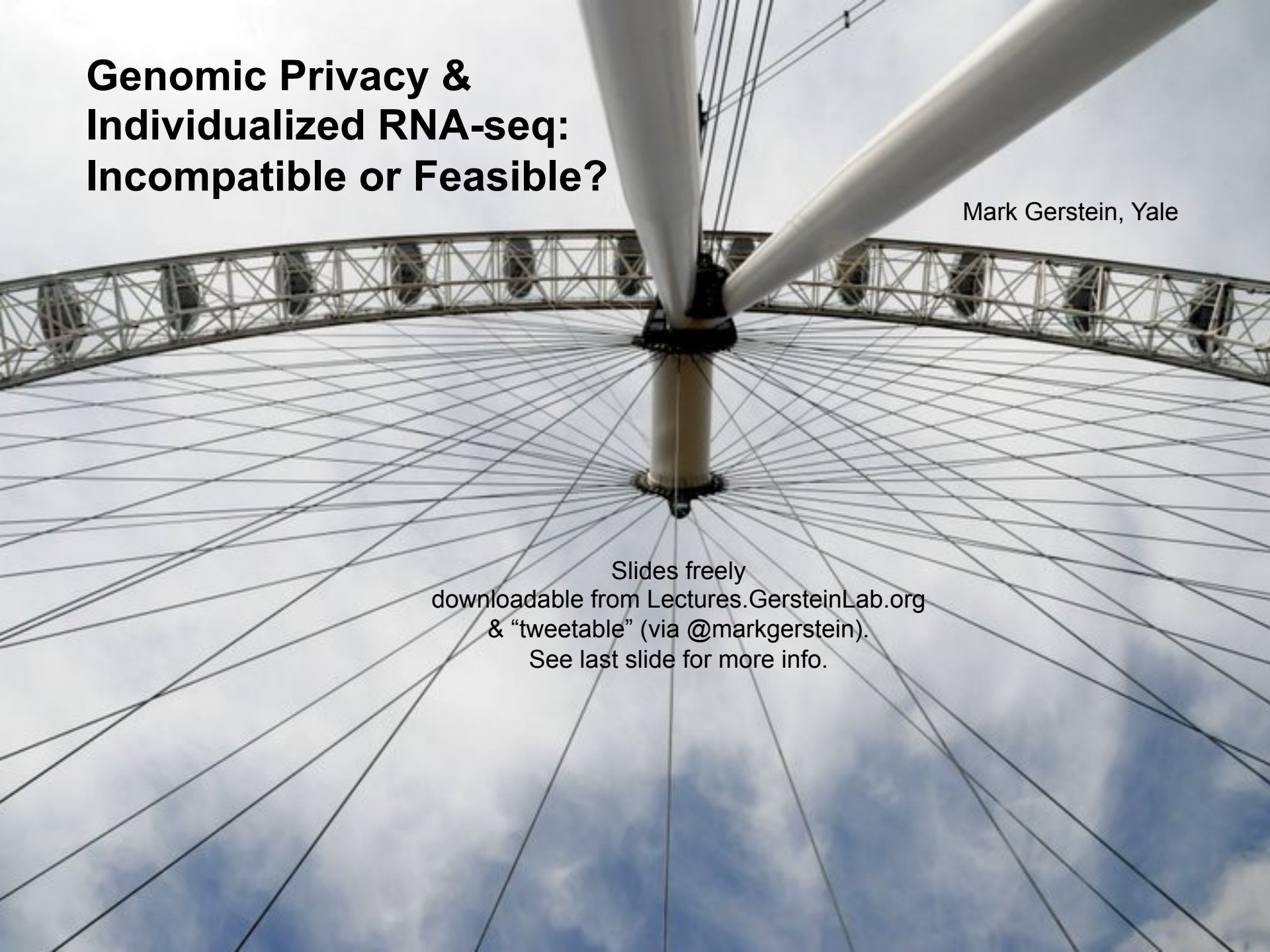
Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

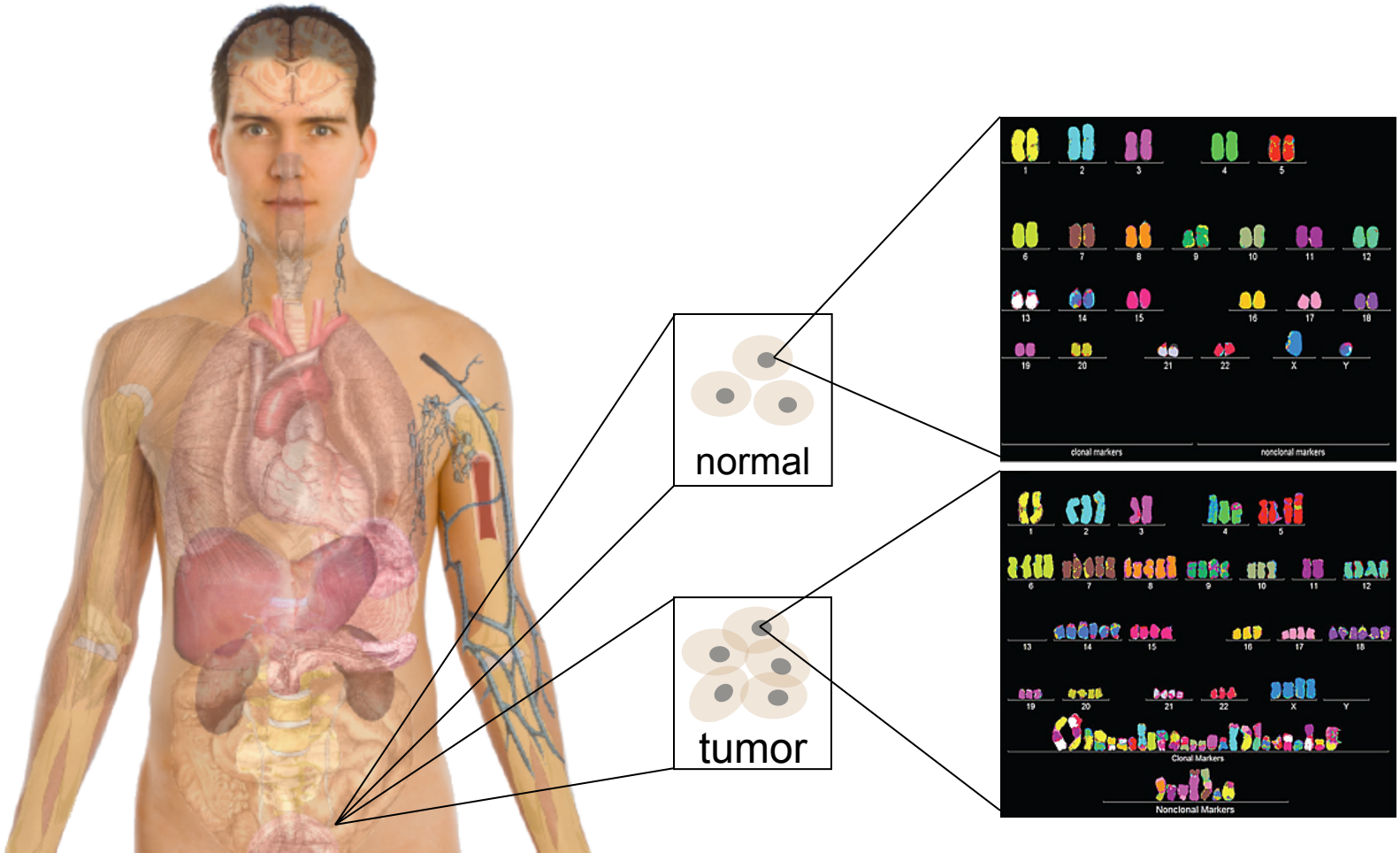
Mark Gerstein, Yale

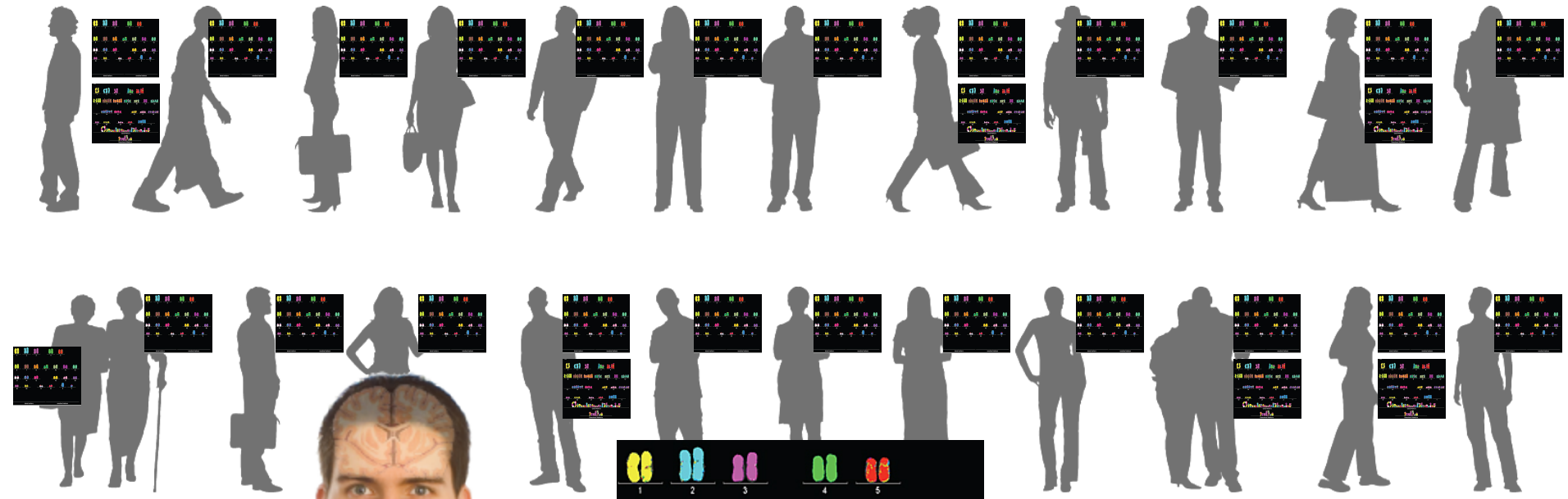


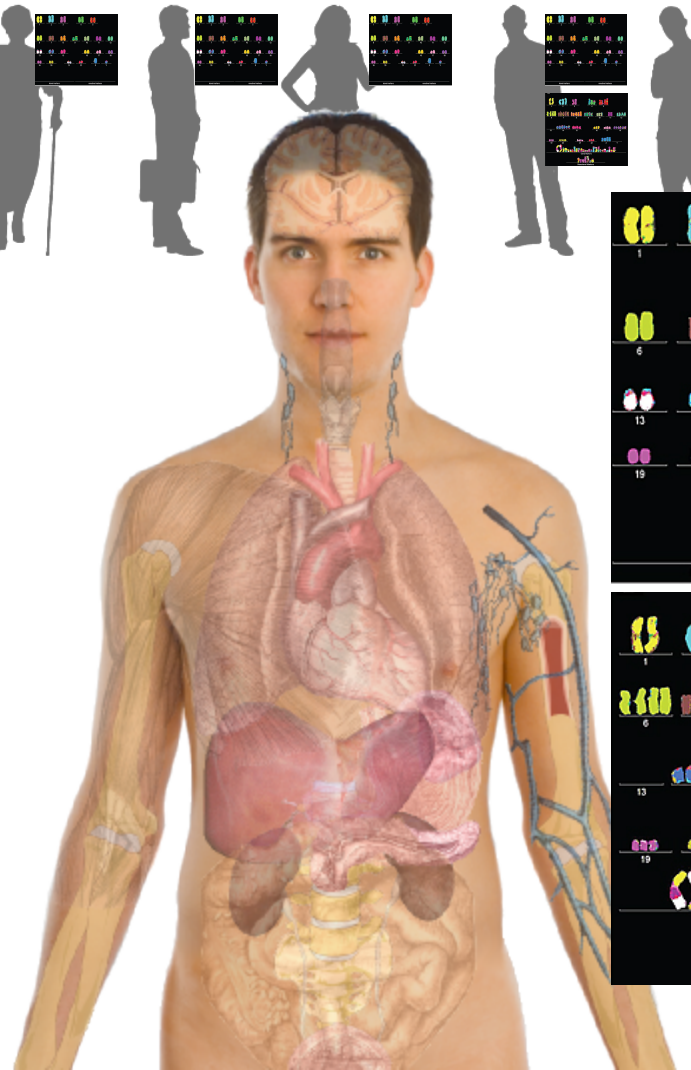
Slides freely
downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).
See last slide for more info.

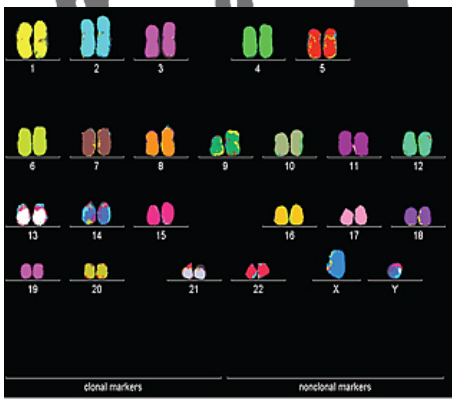
Personal Genomics & Transcriptomics as a Gateway into Biology

Personal genomes (& Transcriptomes) soon will become a commonplace part of medical research & eventually treatment (esp. for cancer).
They will provide a primary connection for biological science to the general public.









Placing the individual into the context of the population & using the population to build a interpretative model

Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

The Conundrum of Genomic Privacy: Is it a Problem?

Yes

Genetic Exceptionalism :
genome is potentially very revealing about one's identity & characteristics

- Most discussion of Identification Risk but what about Characterization Risk?
  - Finding you were in study X vs identifying that you have trait Y from studying your identified genome

No

Shifting societal foci

No one really cares about *your* genes

You might not care

[Klitzman & Sweeney ('11), J Genet Couns 20:98l; Greenbaum & Gerstein ('09), New Sci. (Sep 23)]

Genomics has similar "Big Data" Dilemma in the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
- We confront privacy risks every day we access the internet
- (...or is the genome more exceptional & fundamental?)

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct)]

Tricky Privacy Considerations in Personal Genomics

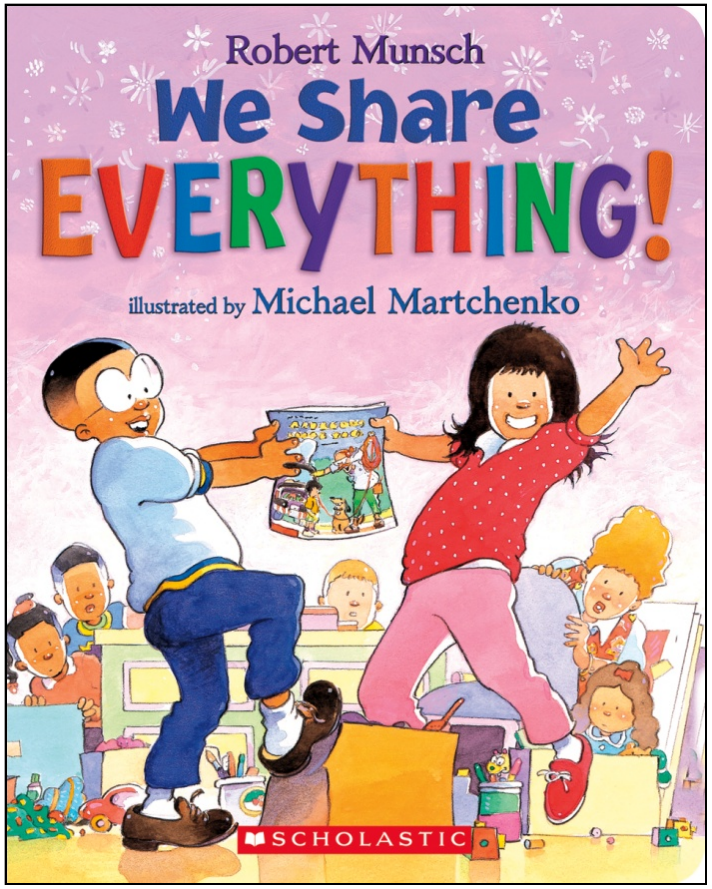
- Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
  - Genomic sequence very revealing about one's children. Is true consent possible?
  - Once put on the web it can't be taken back
- Culture Clash: Genomics historically has been a proponent of "open data" but not clear personal genomics fits this
- Ethically challenged history of genetics

- Ownership of the data & what consent means (Hela)
  - Could your genetic data give rise to a product line?

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]

The Other Side of the Coin: Why we should share

- Sharing helps speed research
  - Large-scale mining of this information is important for medical research
  - Privacy is cumbersome, particularly for big data
- Sharing is important for reproducible research
- Sharing is useful for education



[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

[Economist, 15 Aug '15]

The Dilemma

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- Quantification
  - What is acceptable risk? What is acceptable data leakage? Can we quantify leakage?
    - Ex: photos of eye color
  - Cost Benefit Analysis: how helpful is identifiable data in genomic research v. potential harm from a breach?
- Maybe a we need a few "test pilots" (ala PGP)?
  - Sports stars & celebrities?

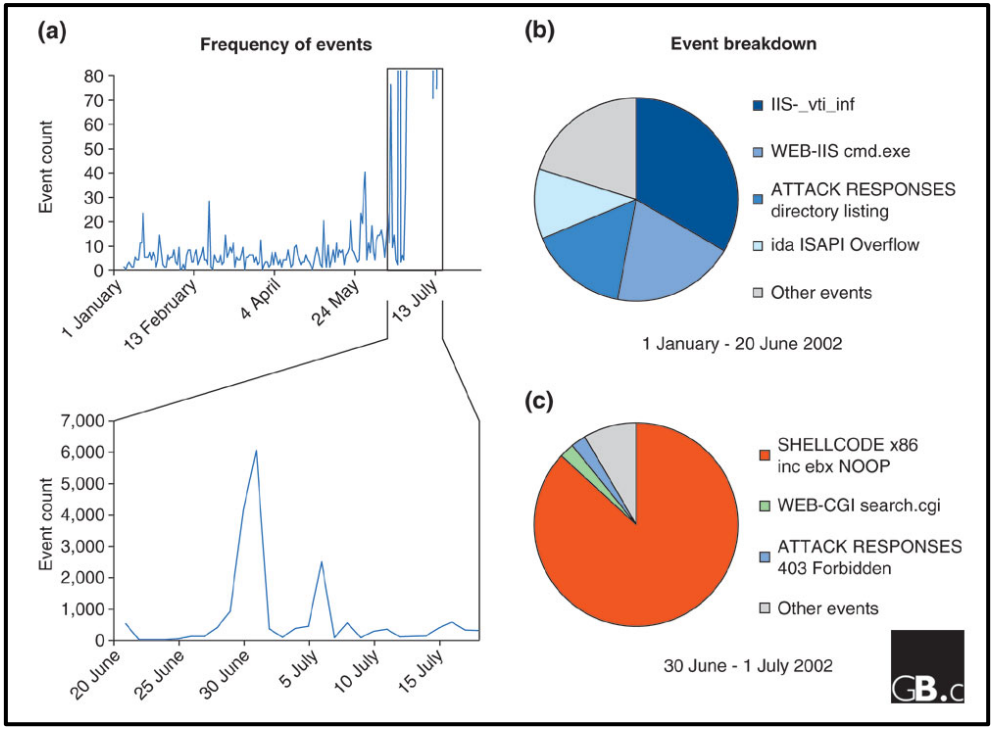
Current Social & Technical Solutions

- Consents
- “Protected” distribution of data (dbGAP)
- Local computes on secure computer

- Issues
  - Non-uniformity of consents & paperwork
    - Different international norms, leading to confusion
  - Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
  - Many schemes get “hacked”

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]

Difficulty in Securing Computers & Data



[Smith et al ('05), Genome Bio]

Genomic Privacy Hacks, Mostly Focusing on Identification

- Early genomic studies were based on small cohorts
  - Individuals give consent to participate but request anonymity
    - HAPMAP, PGP, 1000 Genomes…
  - Focus on hiding the participation of individuals
  - Attacks aimed at detecting whether an individual with known genotypes participated a study
    - "Detection of genomes in a mixture" (Homer et al 2008, Im et al 2012)
- As more people are genotyped, more individuals are in large private genomic databases
  - Detection of an individual is irrelevant, as their participation is already known
    - Current EX: "An individual's genomic/phenotypic data is most certainly stored in their hospital"
    - Future: >1M people's health information is part of a NIH/PMI or NHS databases
- Identification attacks now focus on pinpointing individuals by cross-referencing large seemingly independent datasets
  - Illustrates that a leaked/hacker/stolen dataset, even when anonymized, can leak information
  - Sweeney et al 2013, Gymrek et al 2013

Gymrek et al, "Identifying Personal Genomes by Surname Inference" (2013)
Homer et al, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays." (2008)
Im et al, "On Sharing Quantitative Trait GWAS Results in an Era of Multiple-omics Data and the Limits of Genomic Privacy" (2012)
Sweeney et al, "Identifying Participants in the Personal Genome Project by Name" (2013)

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

NetFlix challenge as an example of a "Linking Attack", characterizing already identified individuals in IMDB, with their (previously hidden) movie viewing habits

Cross correlated small set of identifiable IMDB rating records with large set of "anonymized" Netflix customer ratings, which were being used for a Machine Learning competition

Strawman Hybrid Social & Tech Proposed Solution?

- Fundamentally, researchers have to keep genetic secrets
  - Genetic Licensure & training for individuals (similar to medical license, drivers license)
- Technology to make things easier
  - Cloud computing & enclaves (eg solution of Genomics England)
- Technological barriers shouldn't create a social incentive for “hacking”

- **Quantifying Leakage & allowing a small amounts of it**
- **Careful separation & coupling of private & public data**
  - **Lightweight, freely accessible secondary datasets coupled to underlying variants**
  - Selection of stub & "test pilot" datasets for benchmarking
  - Develop programs on public stubs on your laptop, then move the program to the cloud for private production run

[D Greenbaum, M Gerstein (‘11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

Large-scale RNA

- Recent advent of much large scale RNA-seq (& other functional genomics data) following on DNA sequencing
- Often this is of human subjects & produced by large consortia (eg TCGA, PCAWG, GTEx) and needs to be protected
- Useful to build tools & approaches that interact with these data

The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE

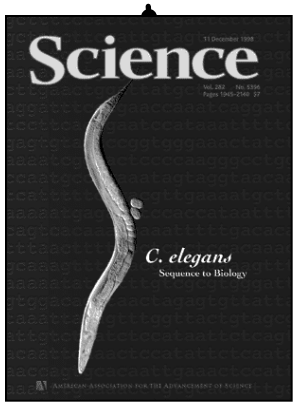
Epigenome Roadmap

2000

2005

2010

2015



Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Phase 3

GTEx

2-side nature of functional genomics data: Analysis can be very General/Public or Individual/Private

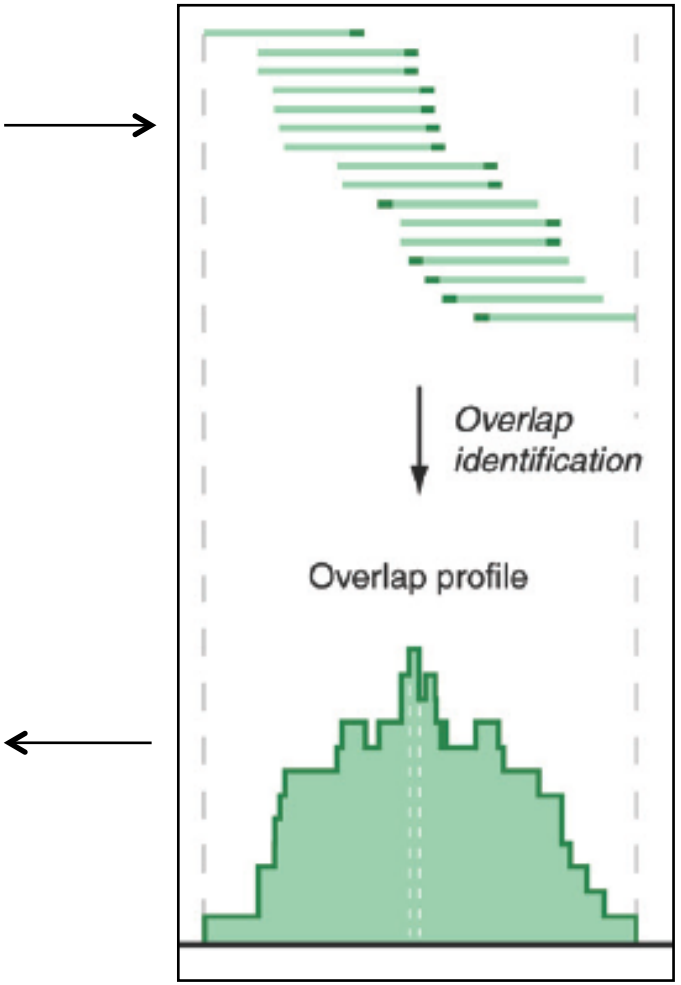
- General quantifications related to overall aspects of a condition & are not tied to an individual's genotype - ie what genes go up in cancer
  - However, data is derived from an individual & tagged with an individual's genotype
- Other calculations aim to use genotype & specific aspects of the quantification to derive general relations related to sequence variation & gene expression
- Some calculations and data derive finding very specific to the variants in a particular individual

RNA-seq

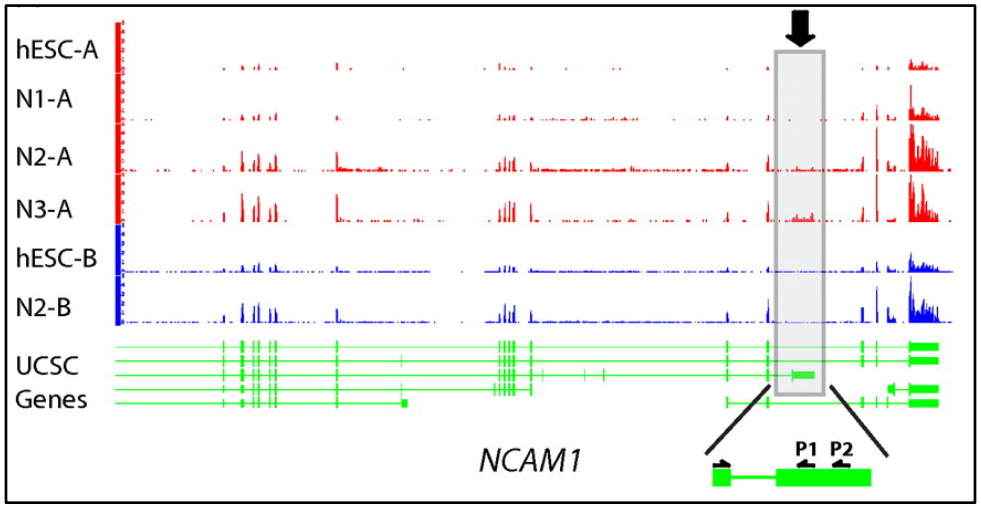
RNA-seq uses next-generation sequencing technologies to reveal RNA presence and quantity within a biological sample.

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Reads (fasta)
- Quality scores (fastq)
- Mapping (BAM)
- Contain variant information in transcribed regions



Reads => Signal

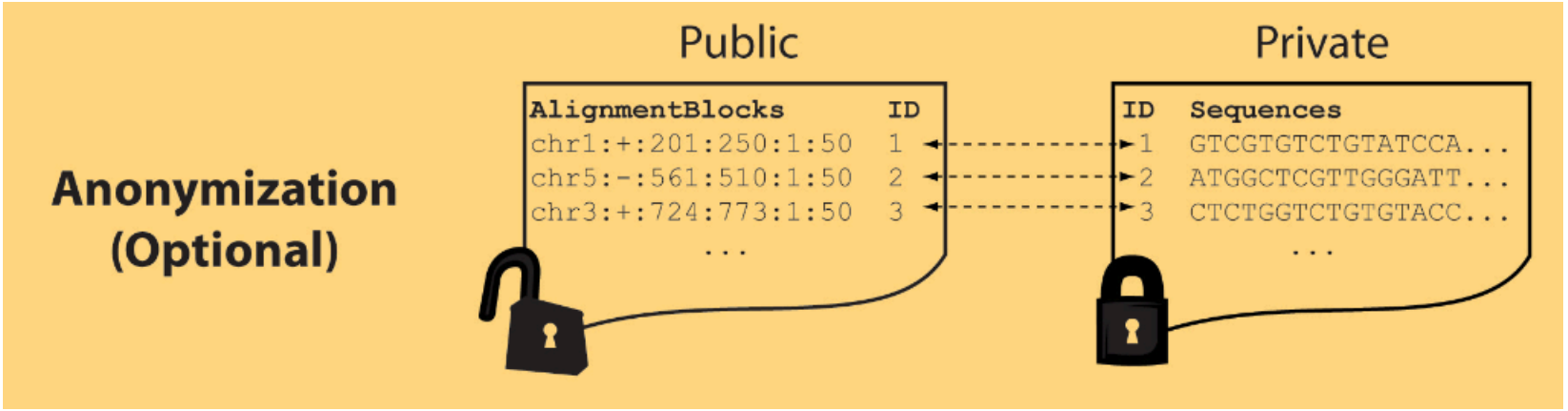


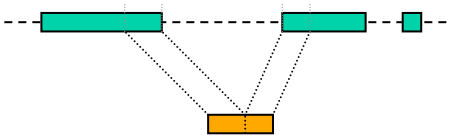
Quantitative information from RNA-seq signal: average signals at exon level (RPKMs)

[*PLOS CB* 4:e1000158; *PNAS* 4:107: 5254 ; *IJC* 123:569]

Light-weight formats

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data





Mapping coordinates without variants (MRF)

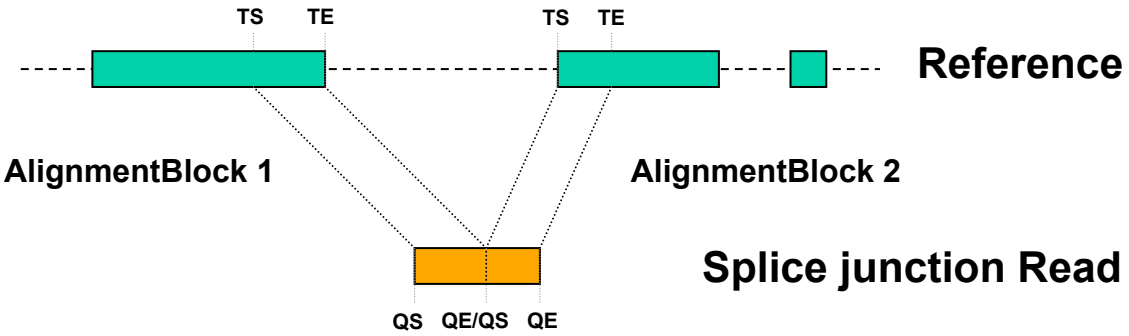
Reads (linked via ID, 10X larger than mapping coord.)

[*Bioinformatics* 27: 281]

MRF Examples

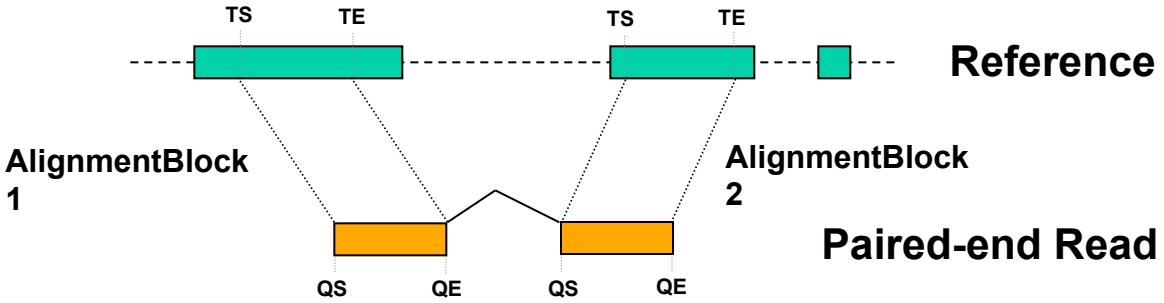
Reference based compression **(ie CRAM)** is similar but it stores actual variant beyond just position of alignment block

`chr2:+:601:630:1:30,chr2:+:921:940:31:50`



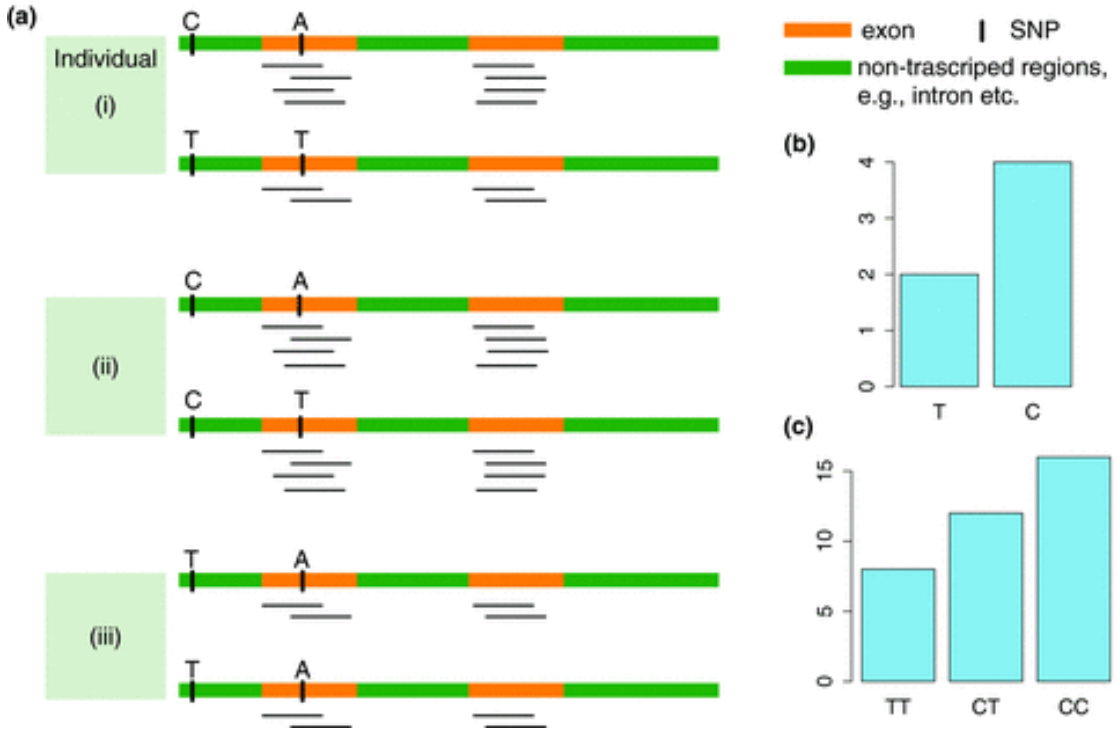
Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

`chr9:+:431:480:1:50|chr9:+:945:994:1:50`

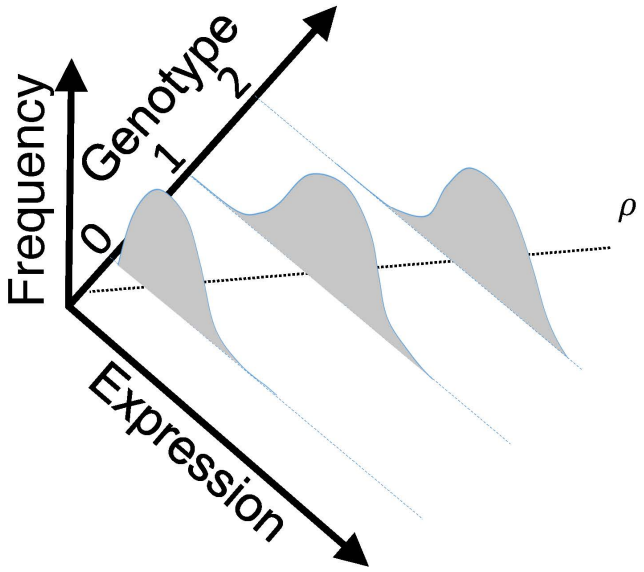


Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

[Habegger et al., Bioinformatics ('11)]



[*Biometrics 68(1) 1–11*]

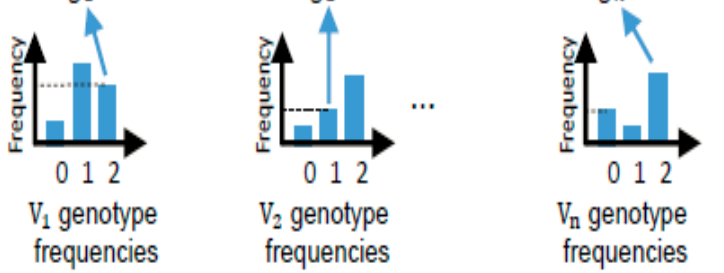


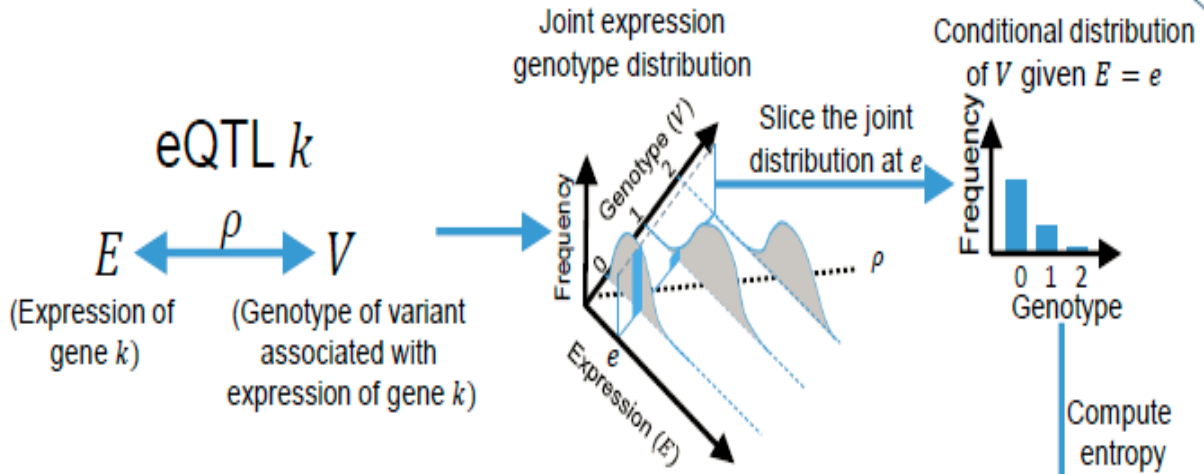
eQTL Mapping Using RNA-Seq Data

- 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

Information Content and Predictability

$$ICI\left(\begin{array}{c}\text{Individual has variant}\\ \text{genotypes } g_1, g_2, \ldots, g_n\\ \text{for variants } V_1, V_1, \ldots, V_n\end{array}\right) = \log\left(\frac{1}{\text{Frequency of } V_1 \text{ genotype } g_1 = 2}\right) + \log\left(\frac{1}{\text{Frequency of } V_2 \text{ genotype } g_2 = 1}\right) + \ldots + \log\left(\frac{1}{\text{Frequency of } V_n \text{ genotype } g_n = 2}\right)$$



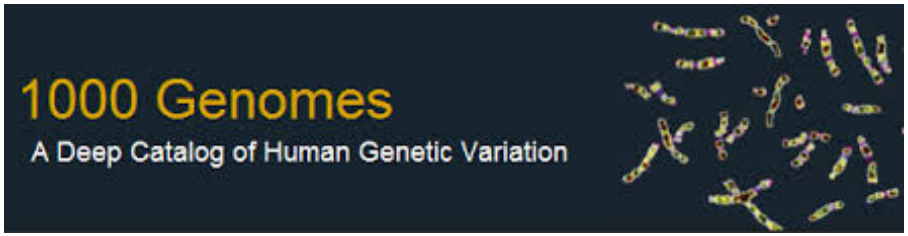


$$\pi\left(V \mid E = e\right) = e^{-H\left(V \mid E = e\right)}$$

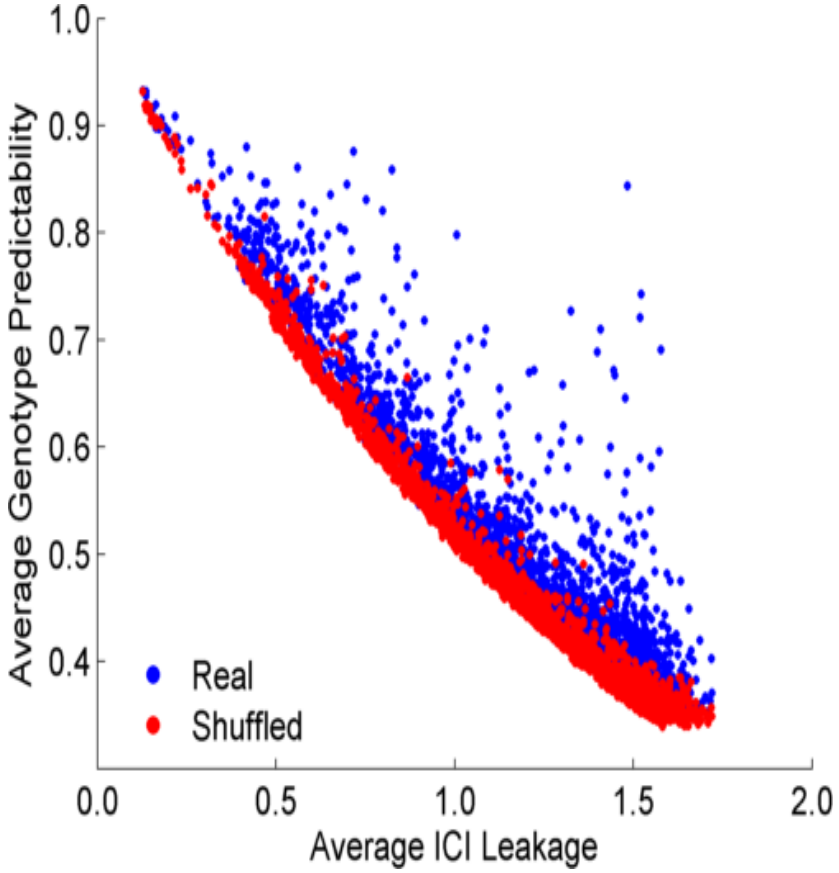
[Harmanciet al. Nat. Meth. (in revision)]

Representative Expression, Genotype, eQTL Datasets

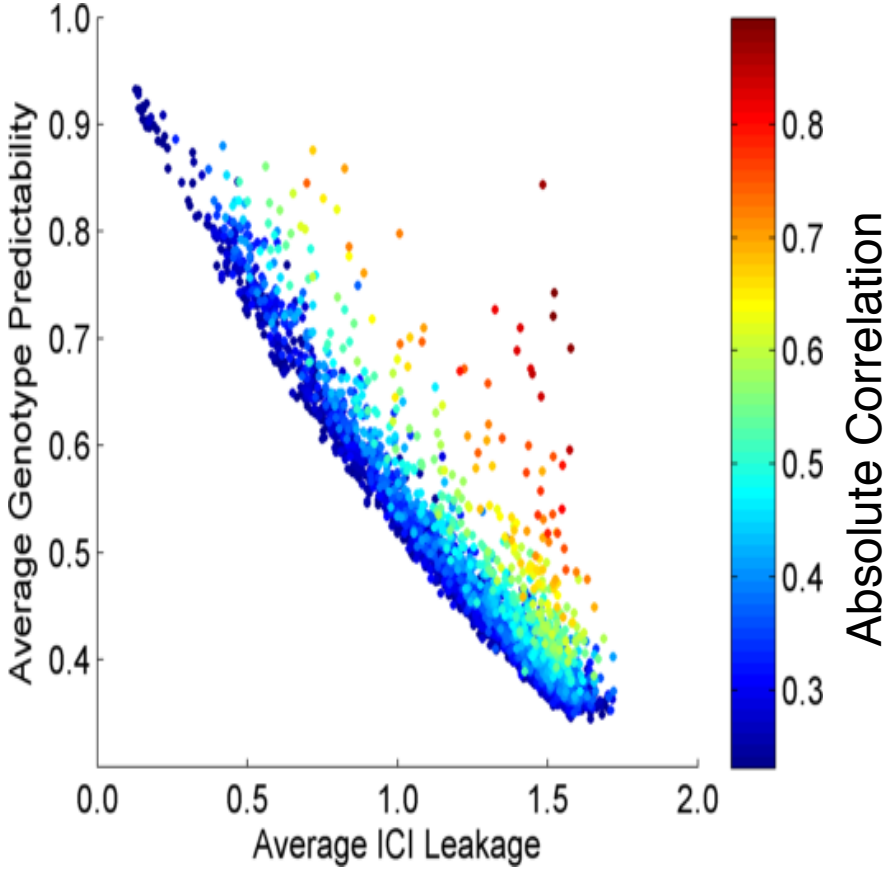
- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals
  - Publicly available quantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)



Per eQTL and ICI Cumulative Leakage versus Genotype Predictability

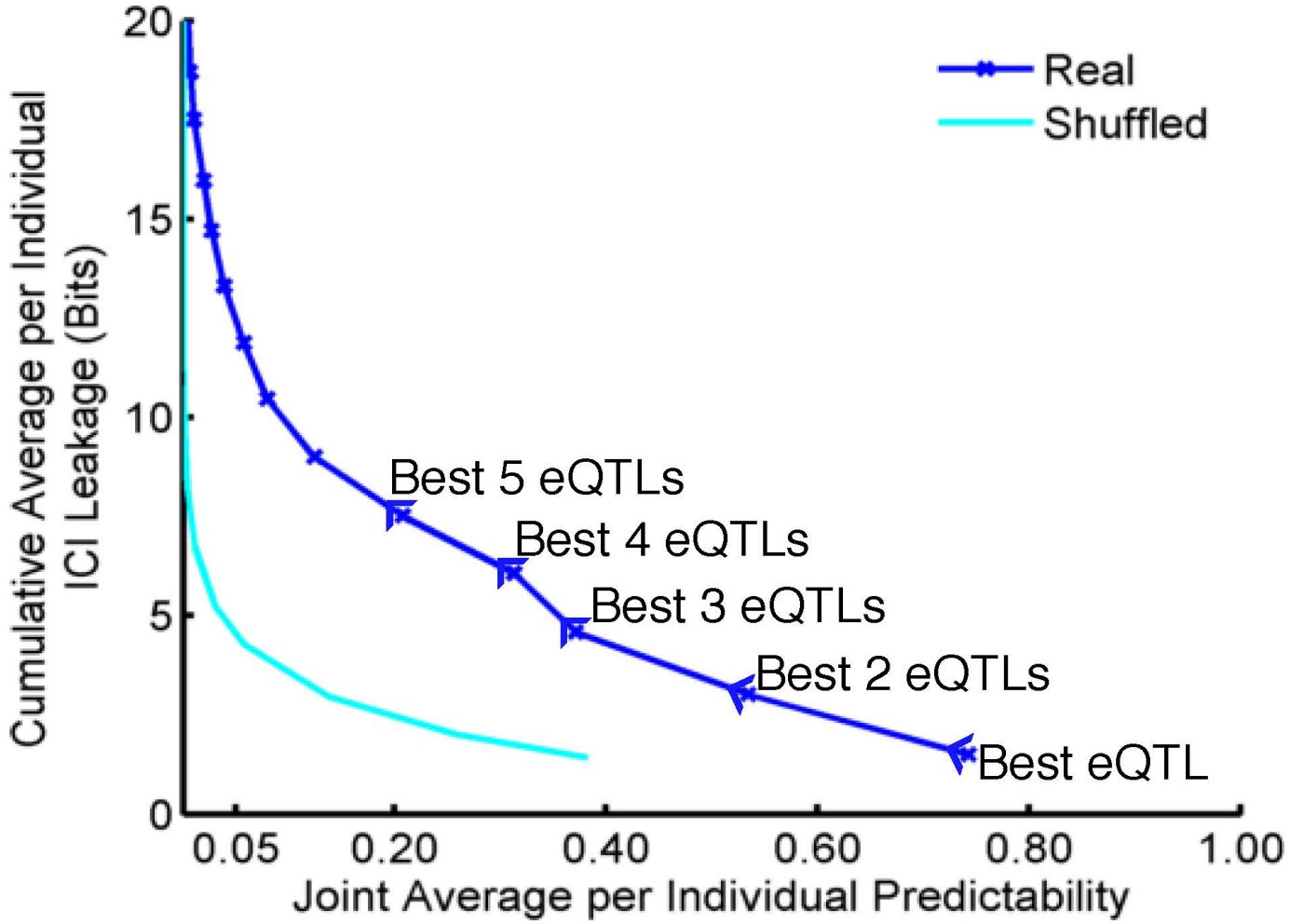


Colors by absolute correlation



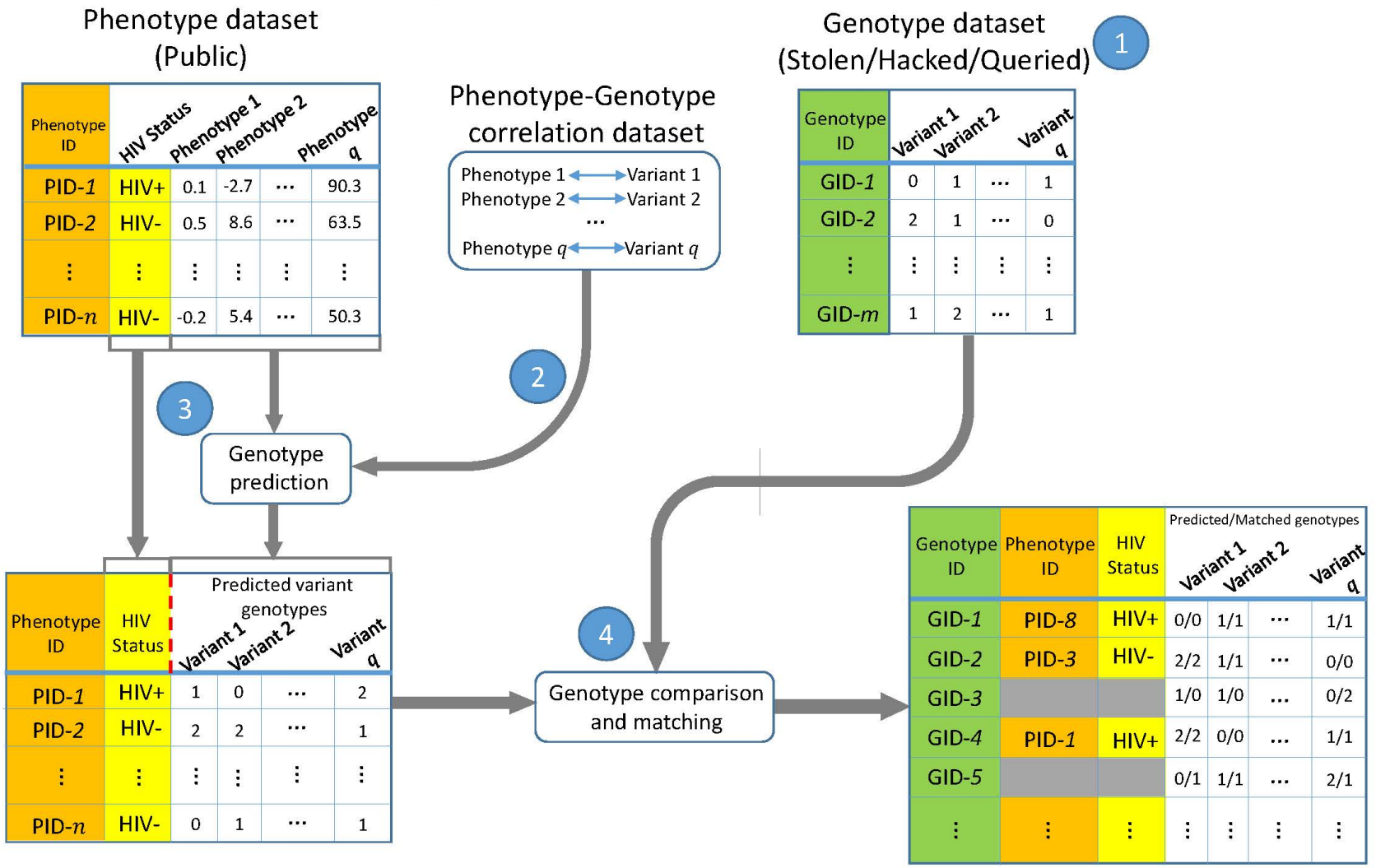
[Harmanciet al. Nat. Meth. (in revision)]

Cumulative Leakage versus Joint Predictability



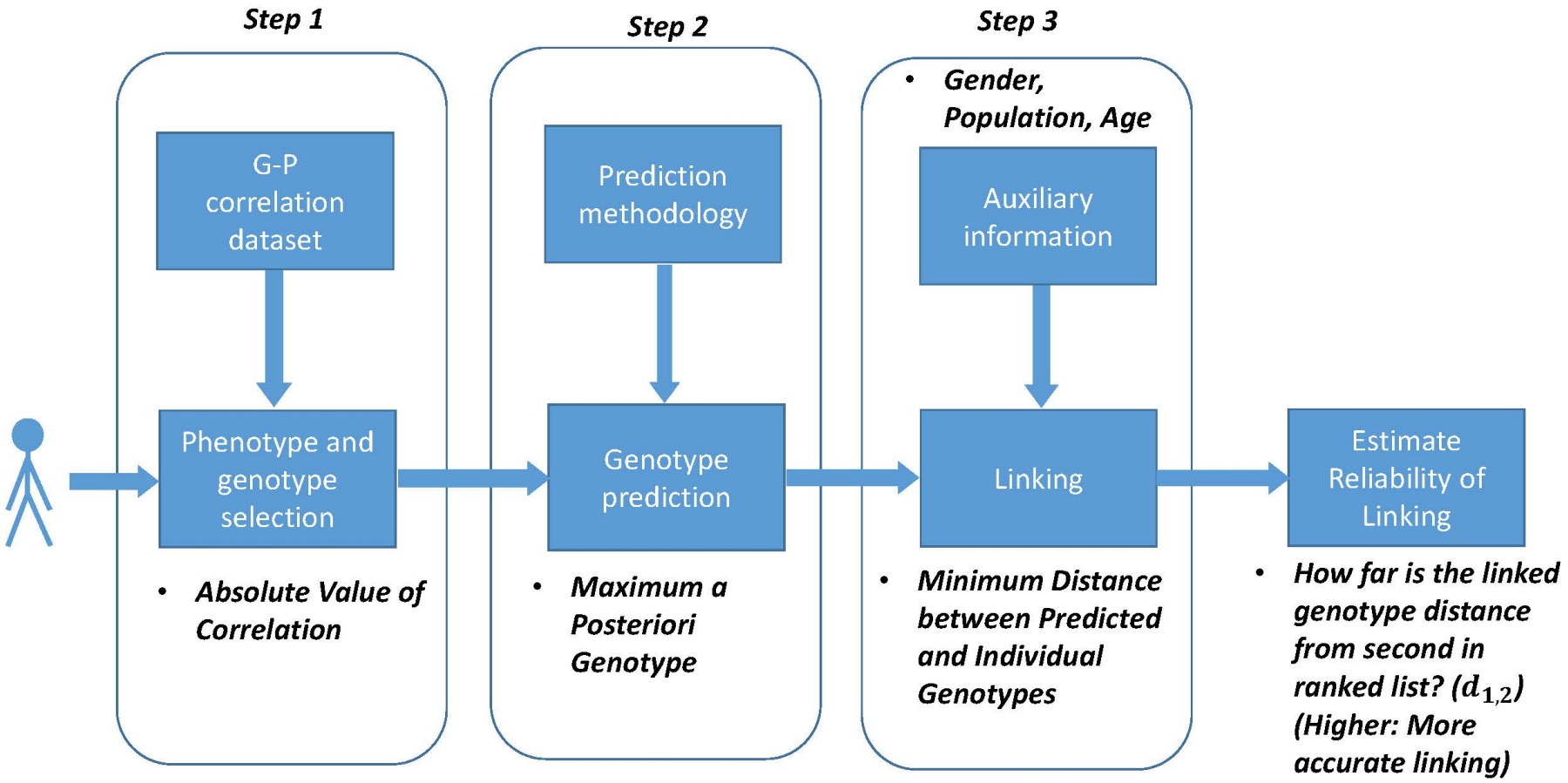
[Harmanciet al. Nat. Meth. (in revision)]

Linking Attack Scenario



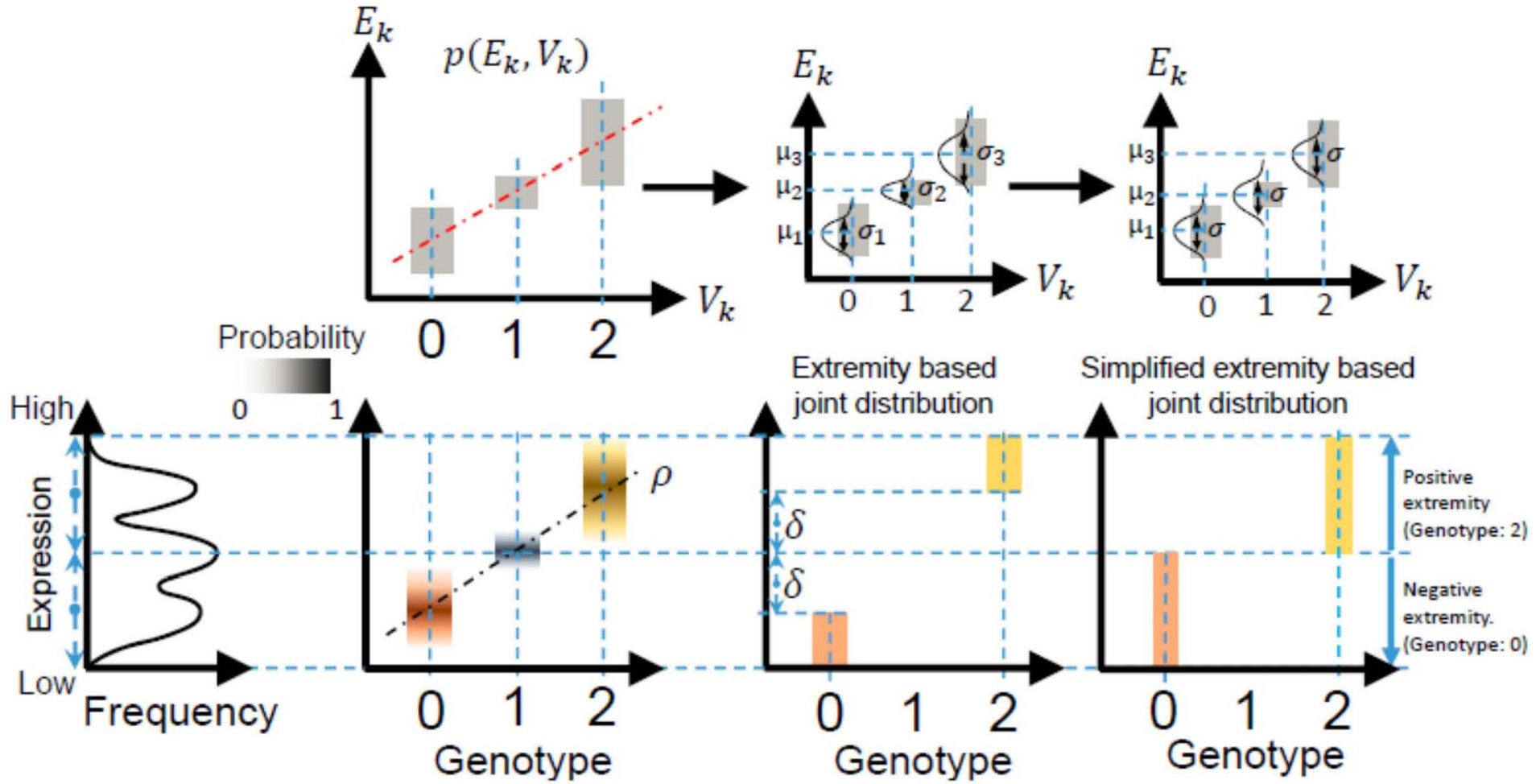
[Harmanciet al. Nat. Meth. (in revision)]

Steps in Instantiation of a (Mock) Linking Attack



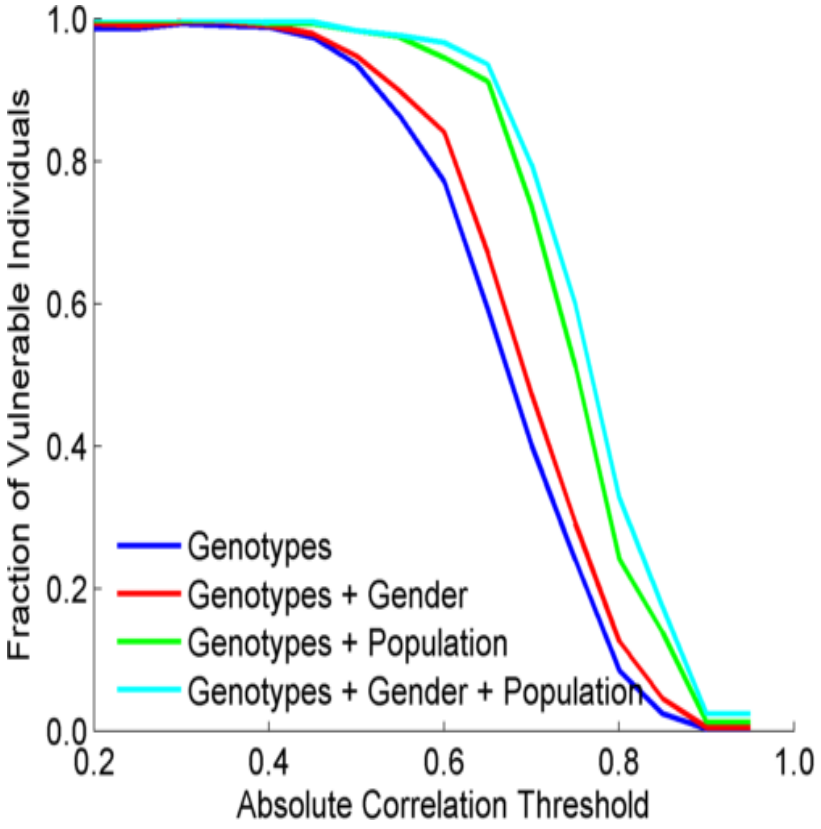
[Harmanciet al. Nat. Meth. (in revision)]

Levels of Expression-Genotype Model Simplifications

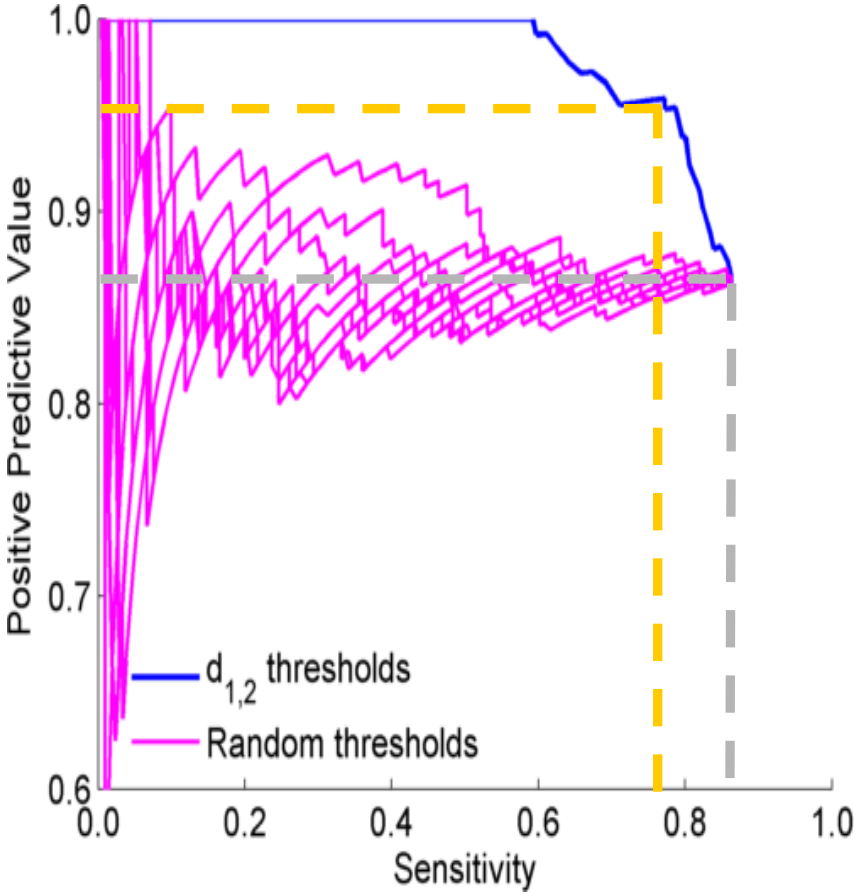


[Harmanciet al. Nat. Meth. (in revision)]

Extremity based linking with homozygous genotypes



Attacker can estimate the reliability of linkings



Sensitivity: Fraction of correctly linked Individuals among all individuals

PPV: Fraction of correctly linked individuals among selected individuals

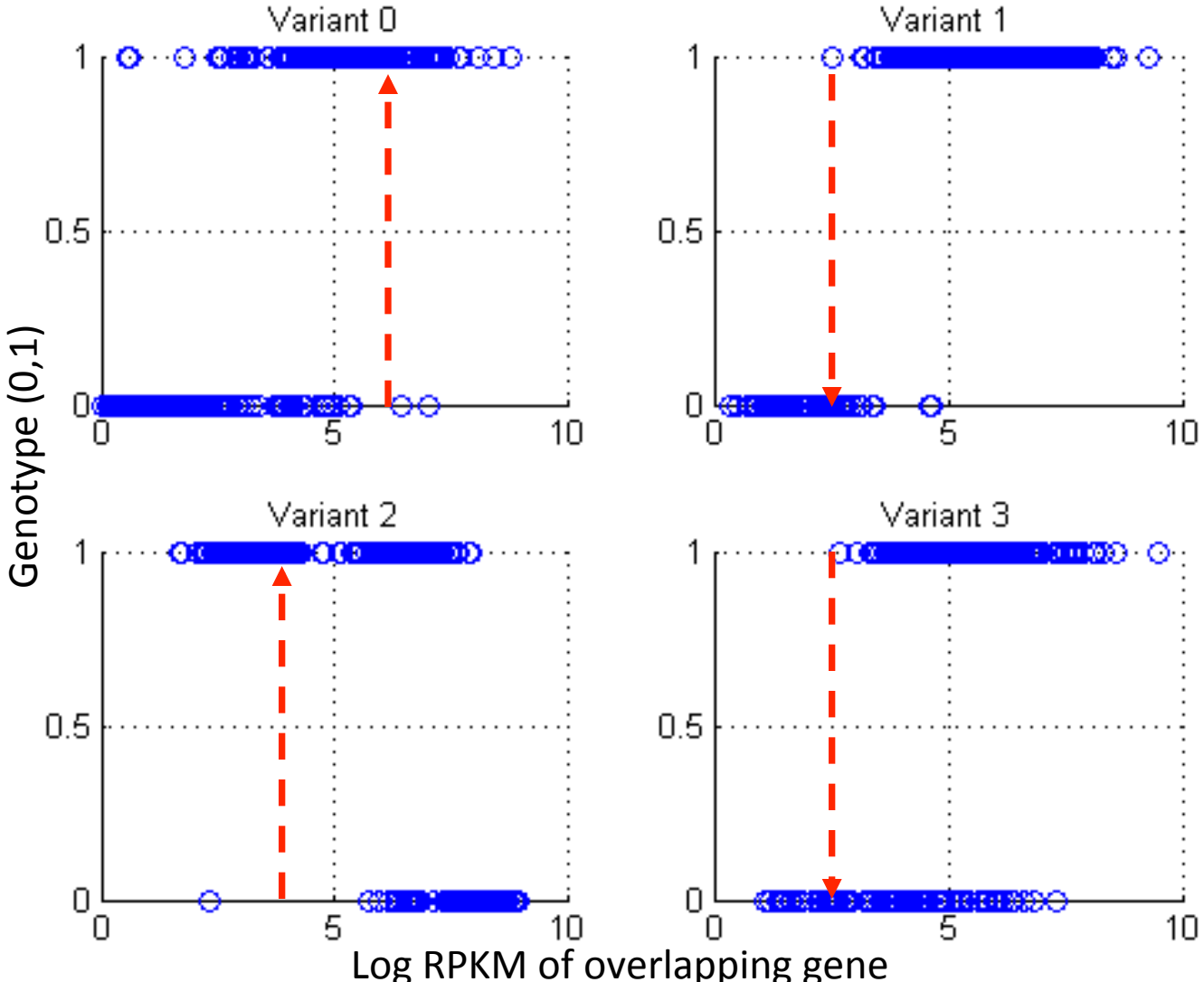
[Harmanciet al. Nat. Meth. (in revision)]

Small Data Leakage from just Gene Expression Data: 4 eQTL-SNP genotypes

Example: Vulnerable sample variants, expressions

- Variant 0 (1, 6)
- Variant 1 (0, 2)
- Variant 2 (1, 3)
- Variant 3 (0, 2)

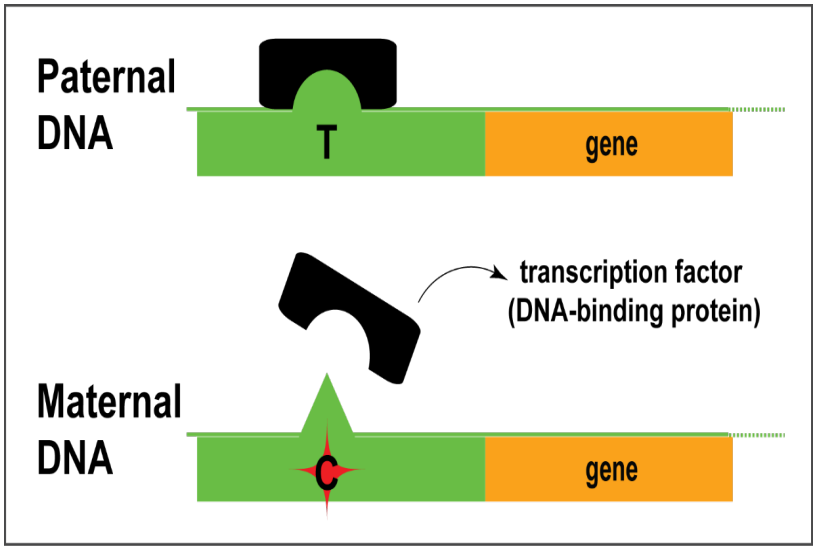
Expression levels are outliers and are predictive of the genotype!



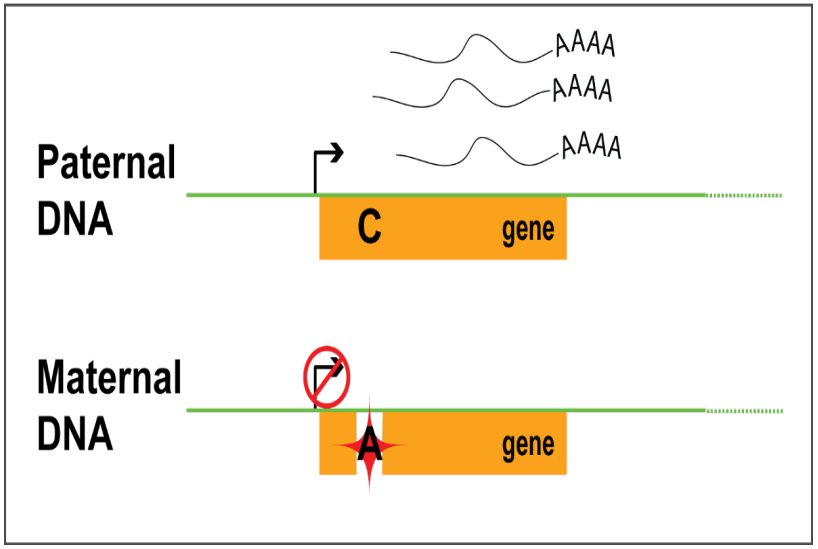
Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

Allele-specific binding and expression



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



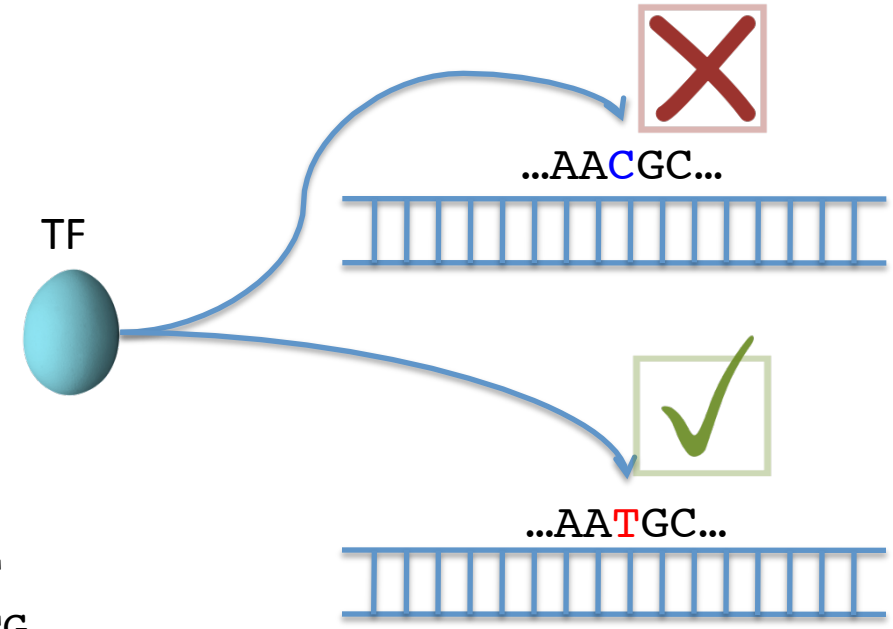
e.g. allele-specific expression (ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

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

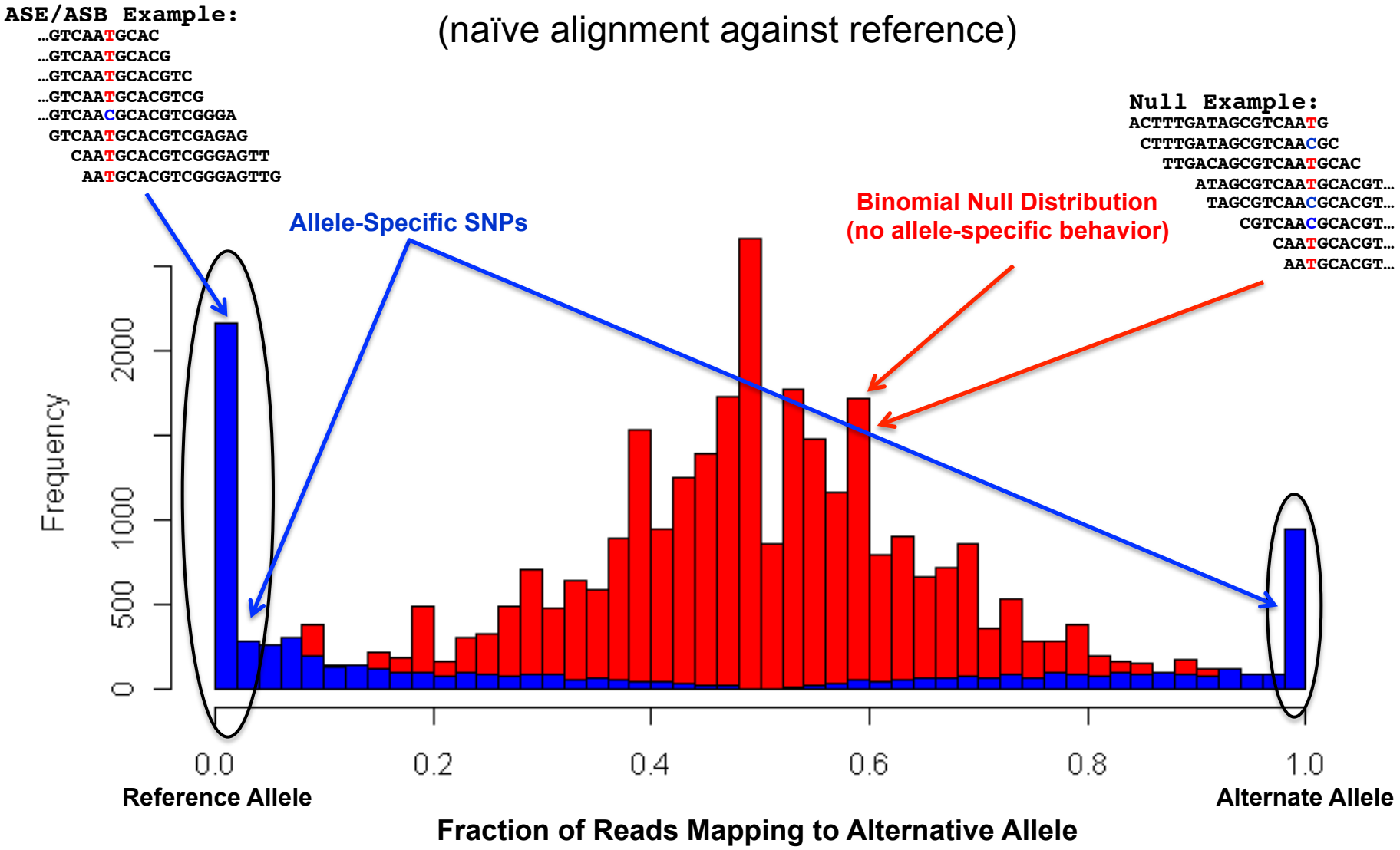
```
10 x T
 2 x C
```



Haplotypes with a Heterozygous Polymorphism

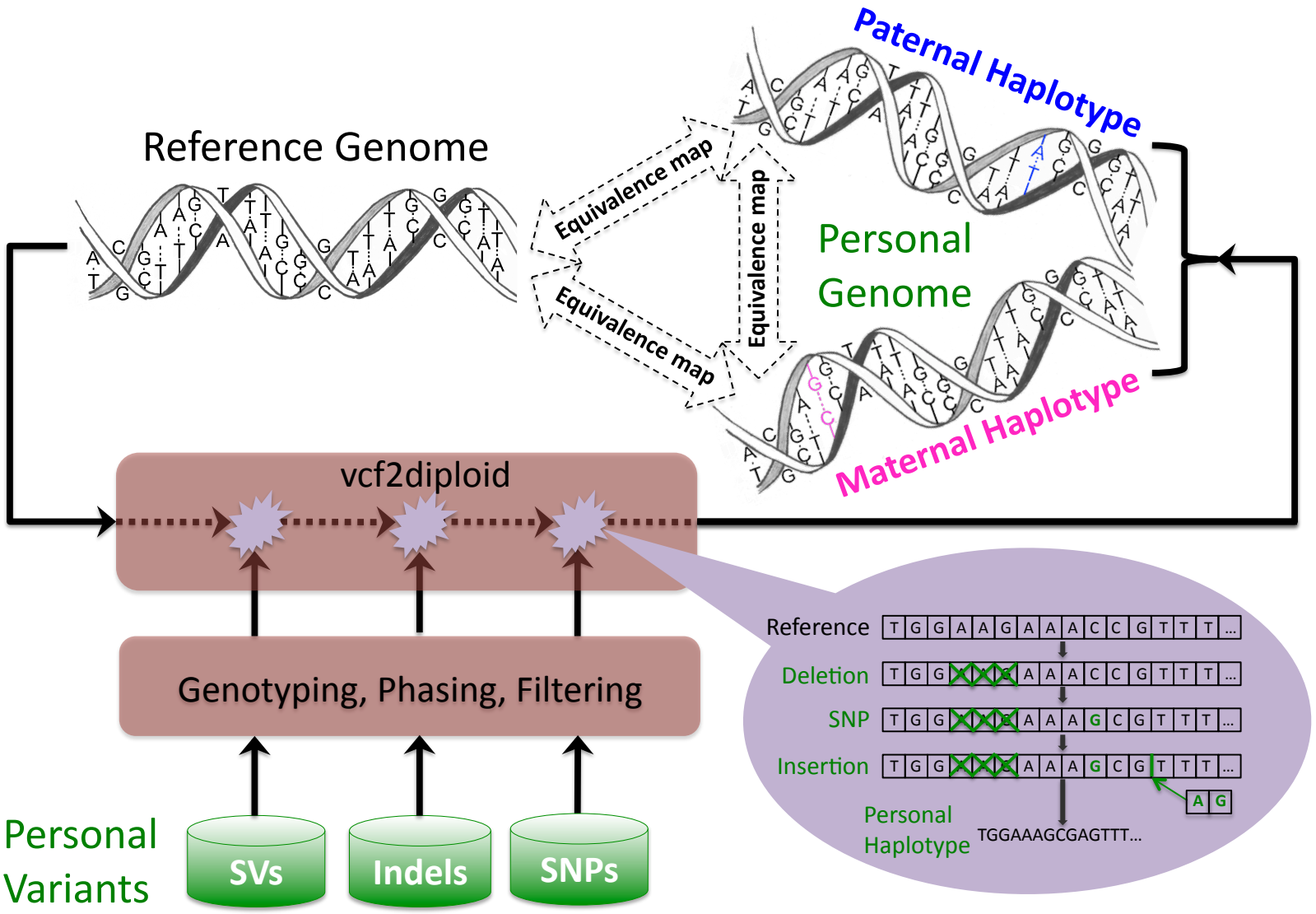
Interplay of the annotation and individual sequence variants

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



[Rozowsky et al., MSB (‘11)]

Construction of a Personal Diploid Genome & Transcriptome

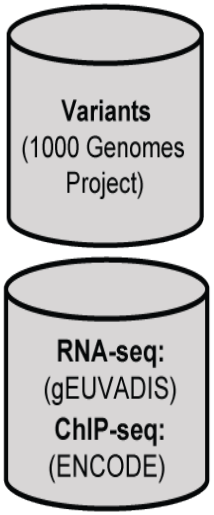


[alleleseq.gersteinlab.org]

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

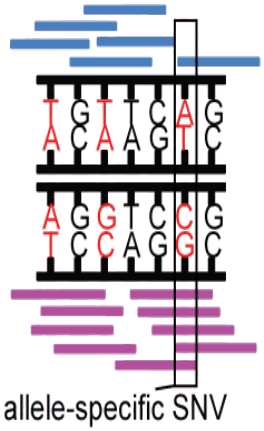
AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes



2. Align ChIP-seq & RNA-seq reads

3. Detect allele-specific variants via a series of filters and tests

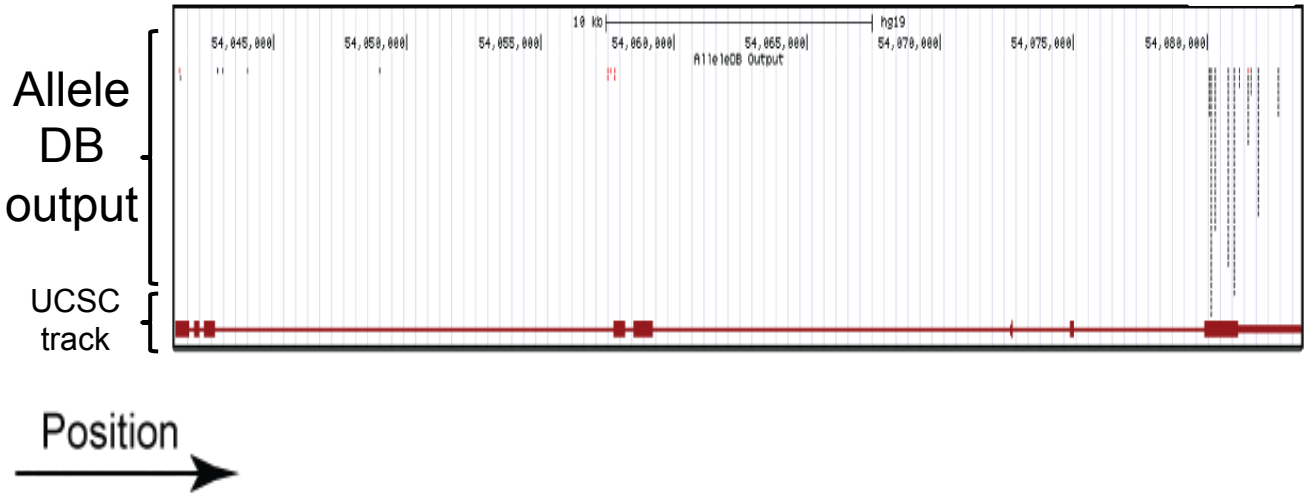


Many Technical Issues:
Reference bias, Ambiguous mapping bias, Over-dispersed (non binomial null)

alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*, '16)

AlleleDB: Annotating rare & common allele-specific variants over a population

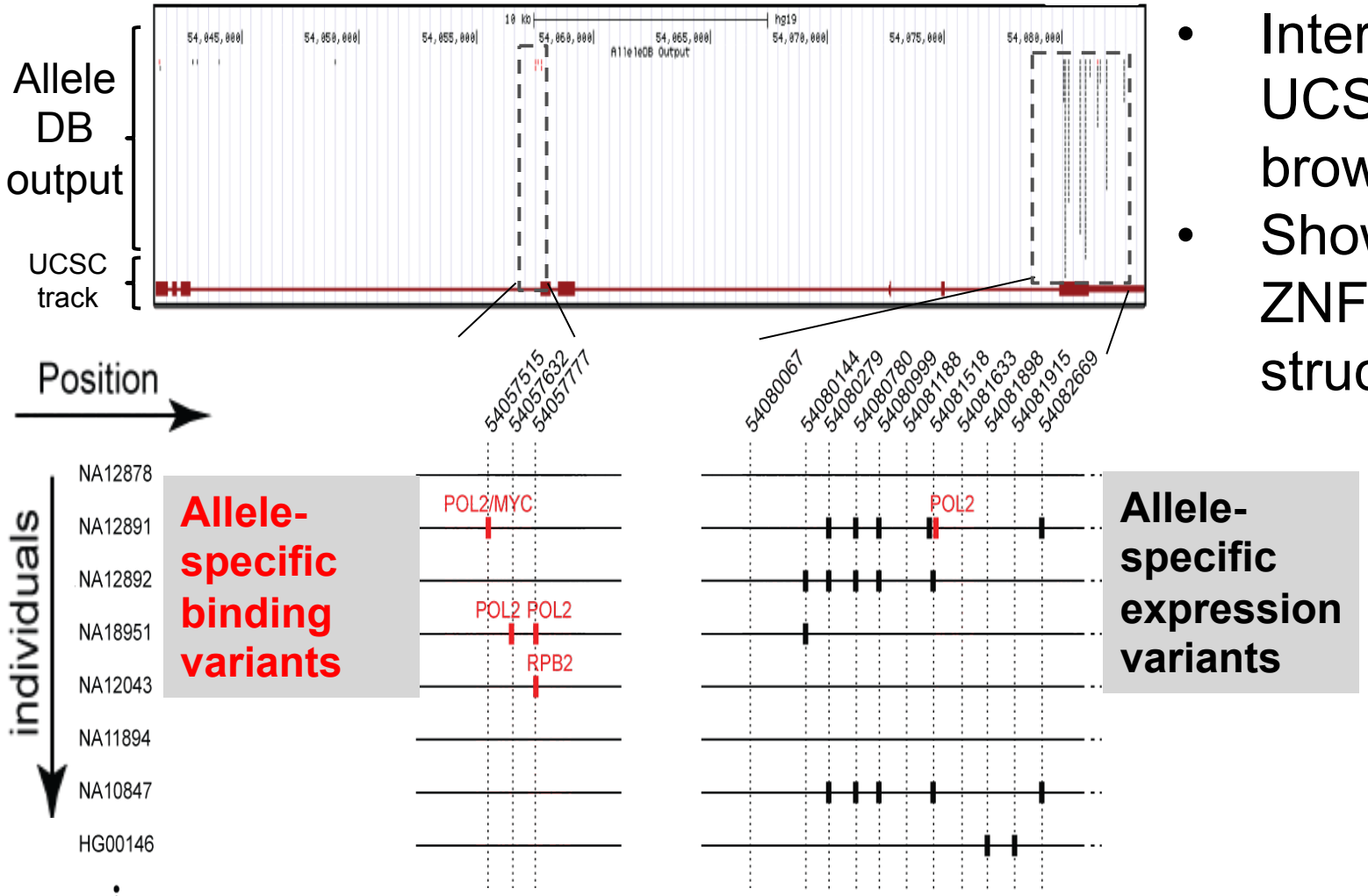


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*, in press)

AlleleDB: Annotating rare & common allele-specific variants over a population

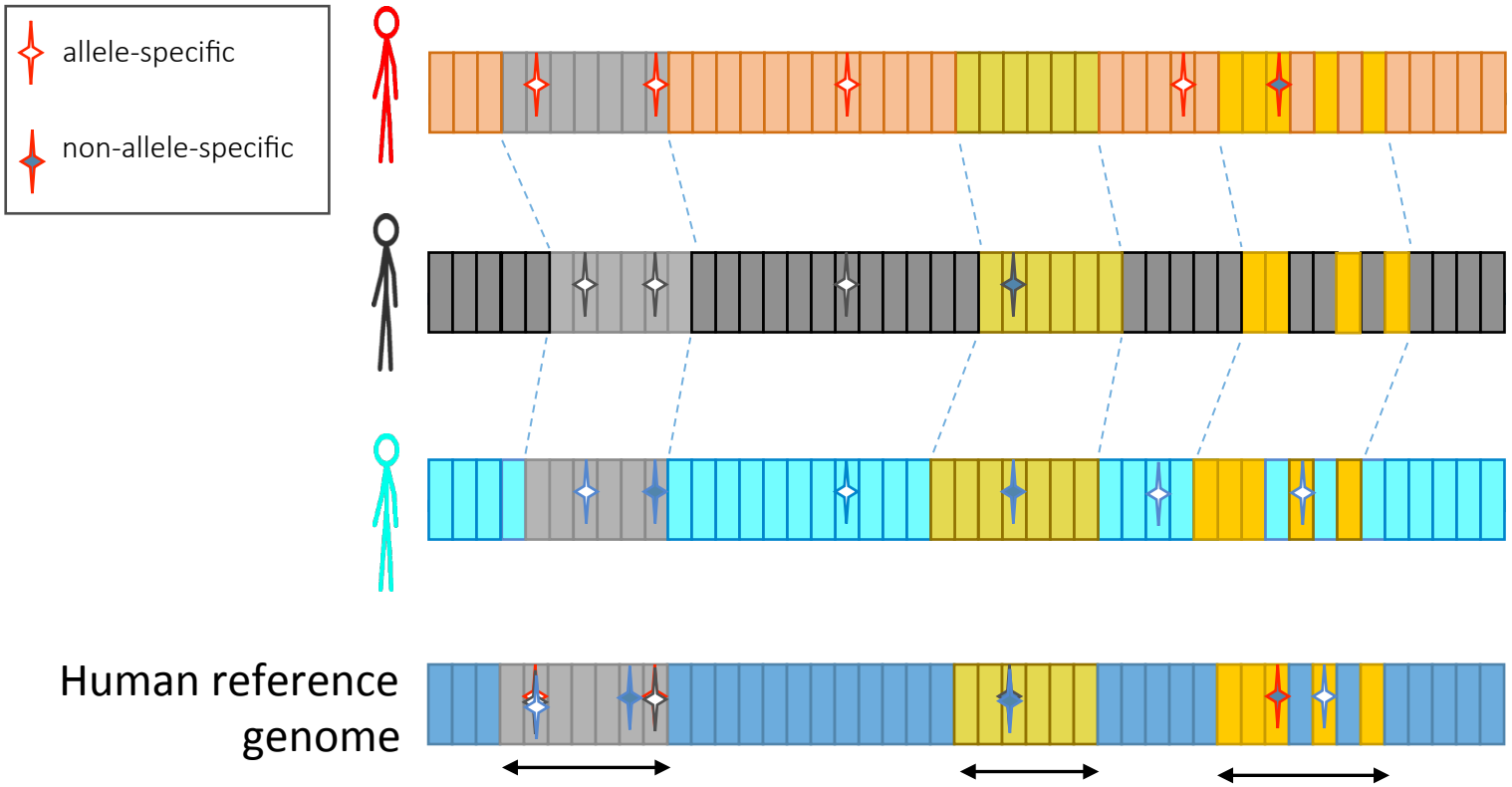


- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

Chen J. *et al.* (*Nature Commun*, '16)

Collecting ASE/ASB variants into allele-specific genomic regions

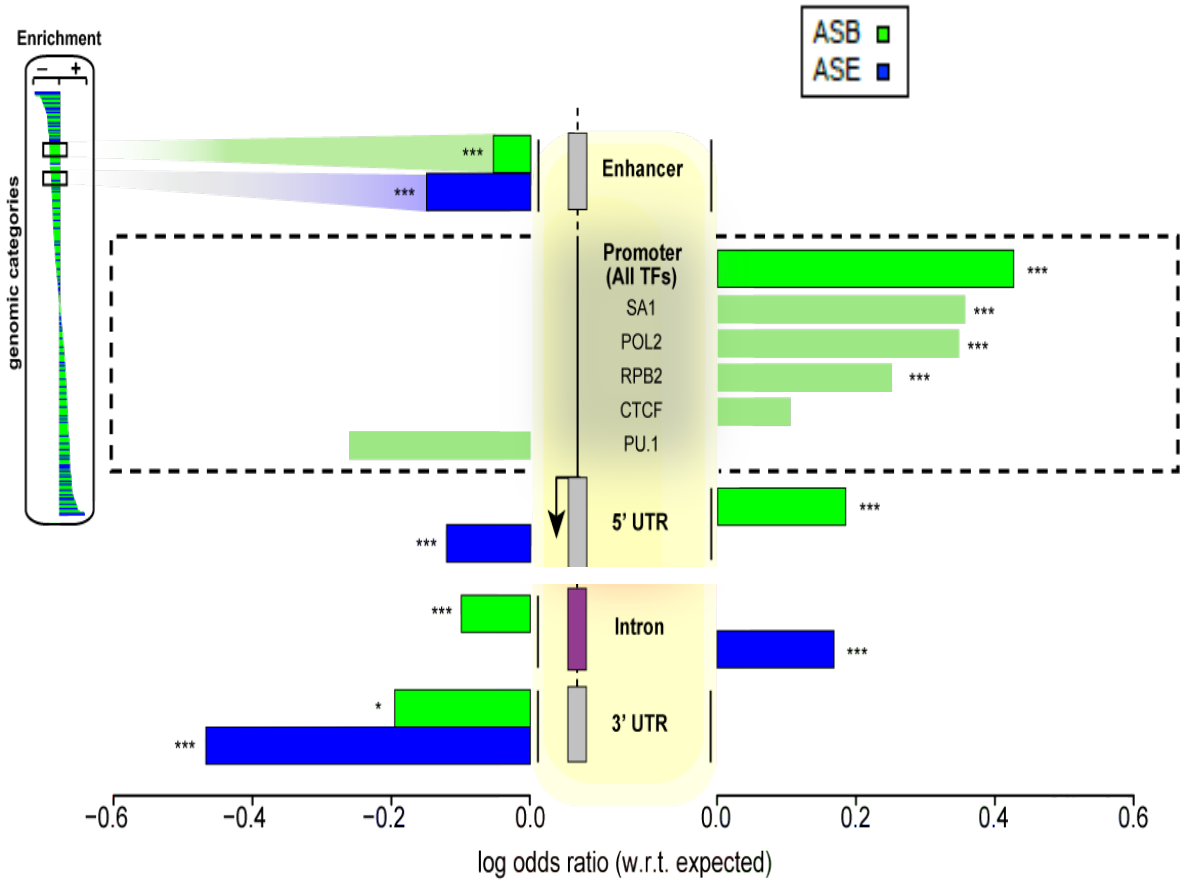
Does a particular genomic element have a higher tendency to be allele-specific? Fisher's exact test, for the **enrichment** of allele-specific variants in the element (with respect to non-allele-specific variants that could potentially be called as allelic)



alleledb.gersteinlab.org

Chen J. *et al.* (*Nature Commun*,'16)

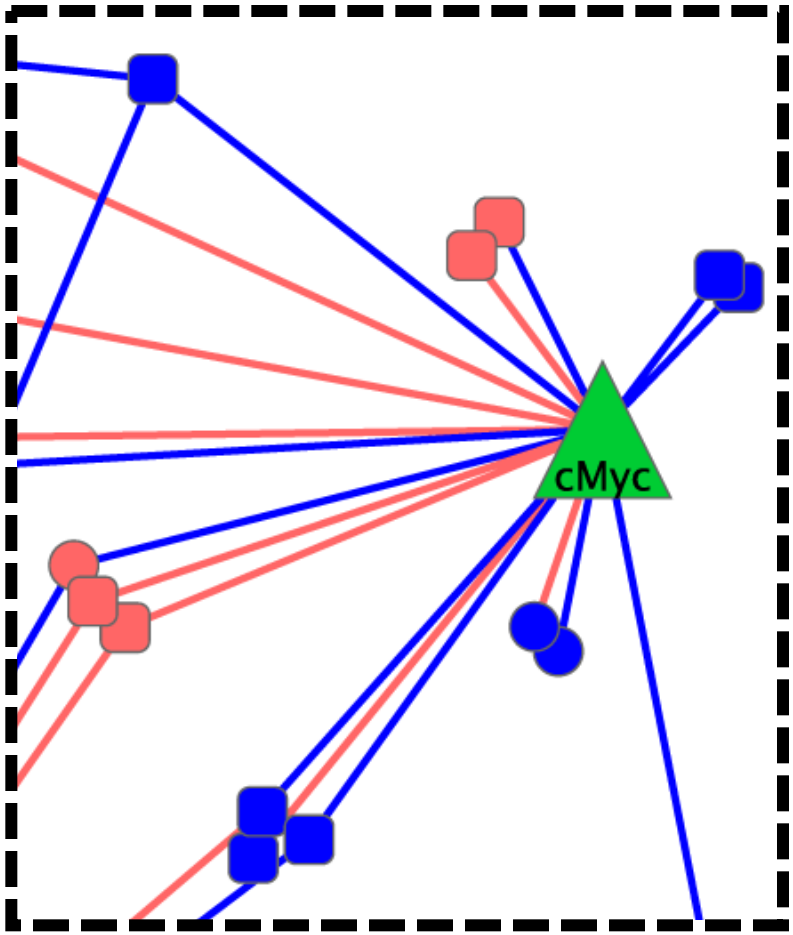
Groups of elements that are enriched or depleted in allelic activity

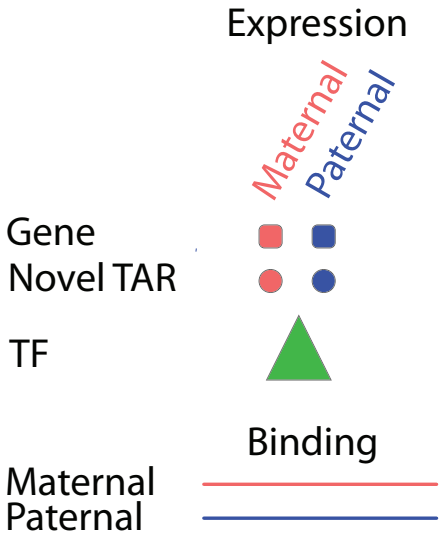


Chen J. *et al.* (*Nature Commun*, '16)

Allele-Specific Behavior in the Regulatory Network

- In GM12878, determine ASB for ~50 TFs & ASE using RNA-Seq
  - ~20% of expressed genes show ASE
  - ~10% of binding sites show ASB
- GM12878 Allele-Specific "Difference" Network
  - Just proximal edges with ASB
  - Just target nodes with ASE



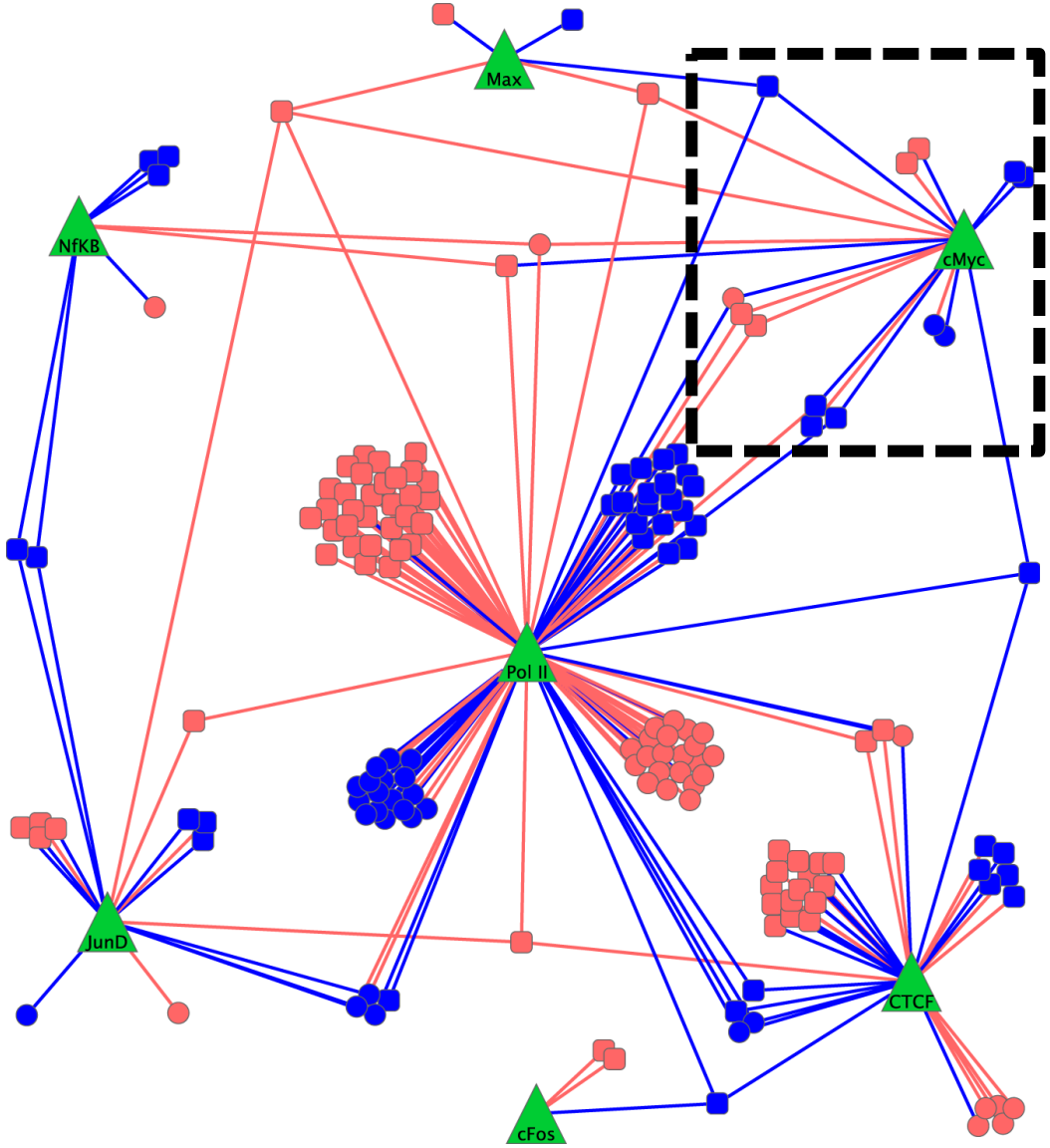


[Rozowsky et al. *MSB* ('11)]

Maternal & Paternal Personal Regulatory Networks:

combinatorial coordination

of ASE & ASB

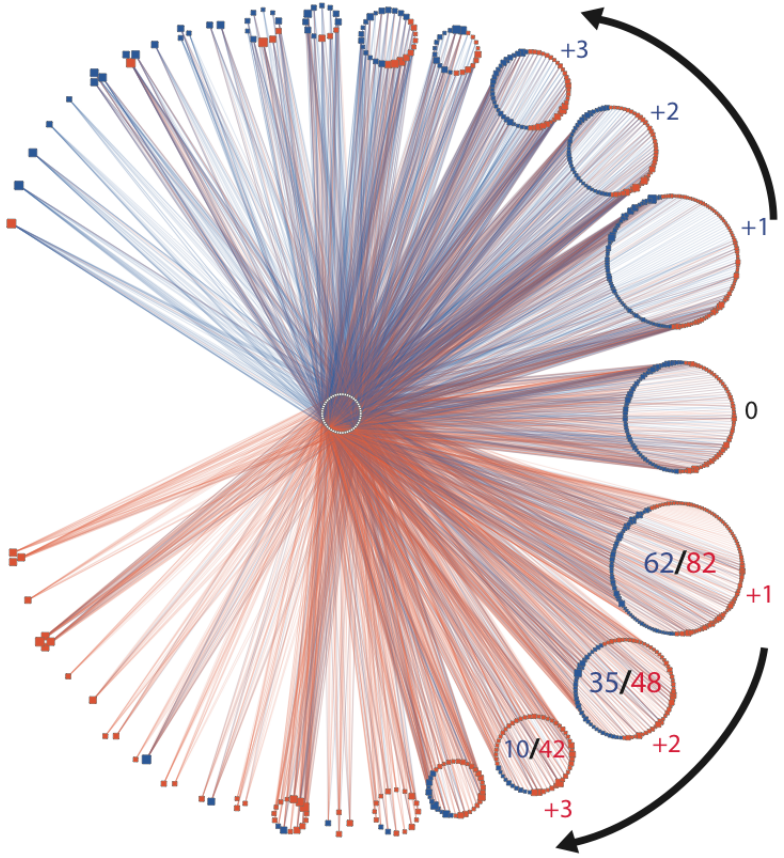


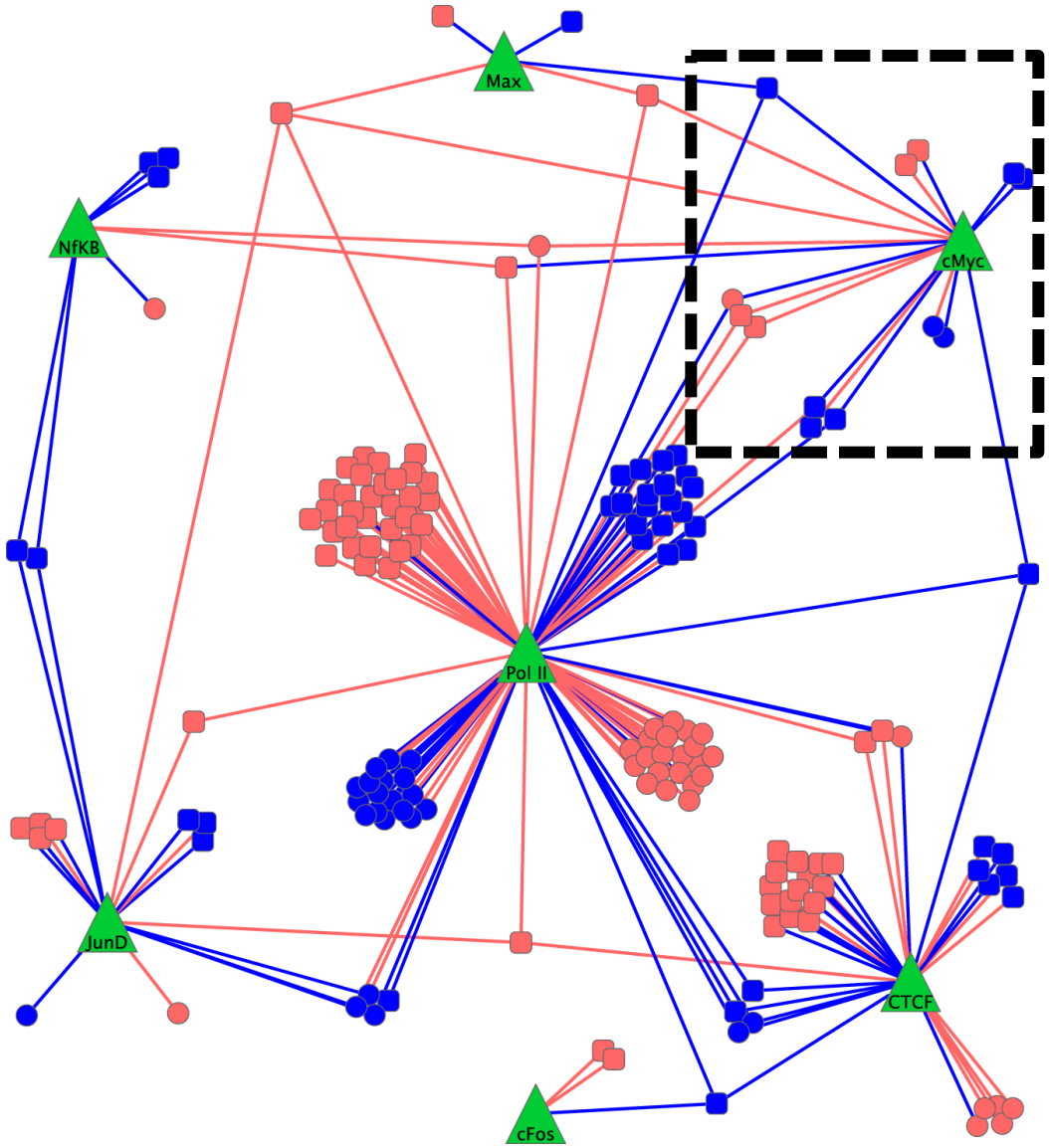
[Rozowsky et al. *MSB* ('11)]

Maternal & Paternal Personal Regulatory Networks:

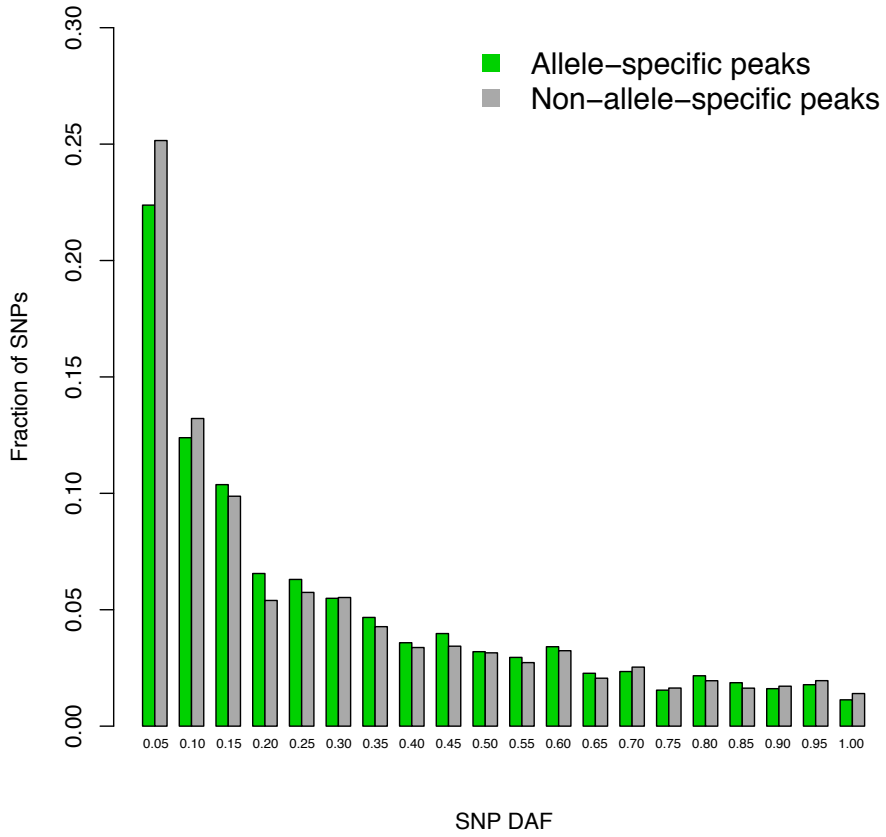
combinatorial coordination

of ASE & ASB





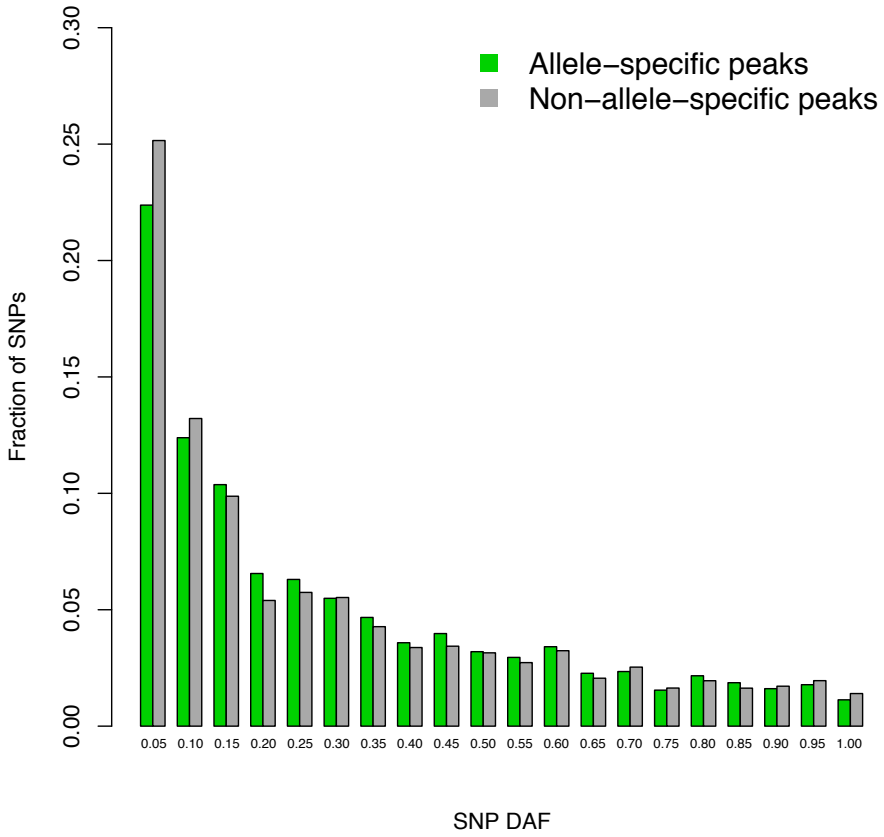
[Rozowsky et al. *MSB* ('11)]



(from 1000G pilot & phase I)

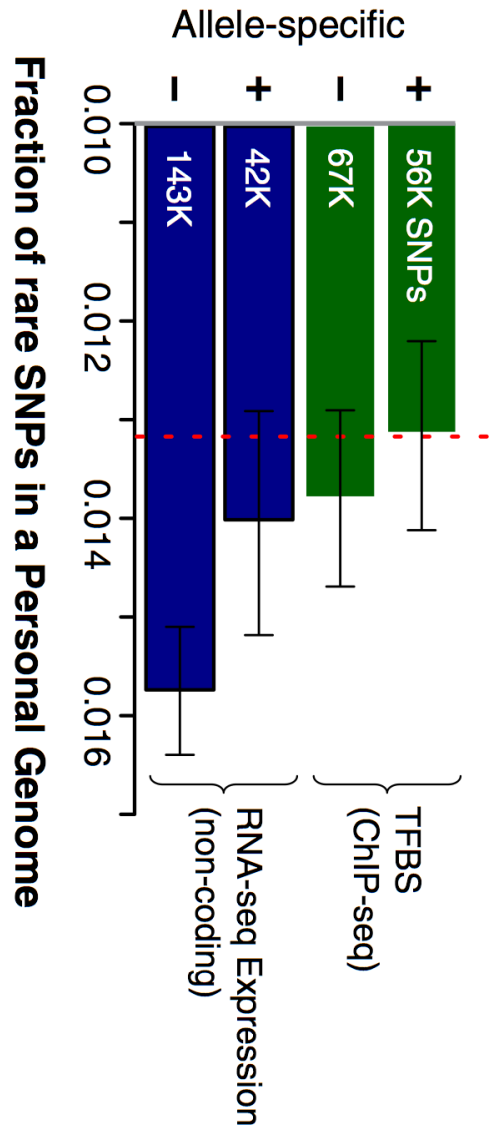
More "allelic" components under weaker selection

[Gerstein et al. *Nature* ('12)]



(from 1000G pilot & phase I)

More "allelic" components under weaker selection



[Khurana et al. *Science* ('13)]
[Gerstein et al. *Nature* ('12)]

Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

Genomic Privacy & Individualized RNA-seq: Incompatible or Feasible?

- The General Dilemma of Genomic Privacy
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies, burdensome security, various "hacks"
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
- RNA-seq: How to Publicly Share Some of it
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- Allelic Expression & Binding Activity
  - Difference in molecular activity between specific alleles
  - RNA-seq calculations that fundamentally involve individual variants
  - Aggregating results with AlleleDB to define allelic elements & subnetworks
  - Allelic elements tend to be under weaker selection

papers.gersteinlab.org/subject/**privacy - D Greenbaum**

PrivaSeq.gersteinlab.org - **A Harmanci**

RSEQtools.gersteinlab.org [MRF]
L Habegger, A Sboner,
TA Gianoulis, J Rozowsky, A Agarwal, M Snyder

AlleleDB.gersteinlab.org
J Chen, J Rozowsky,
T Galeev, A Harmanci,
R Kitchen, J Bedford,
A Abyzov, Y Kong, L Regan

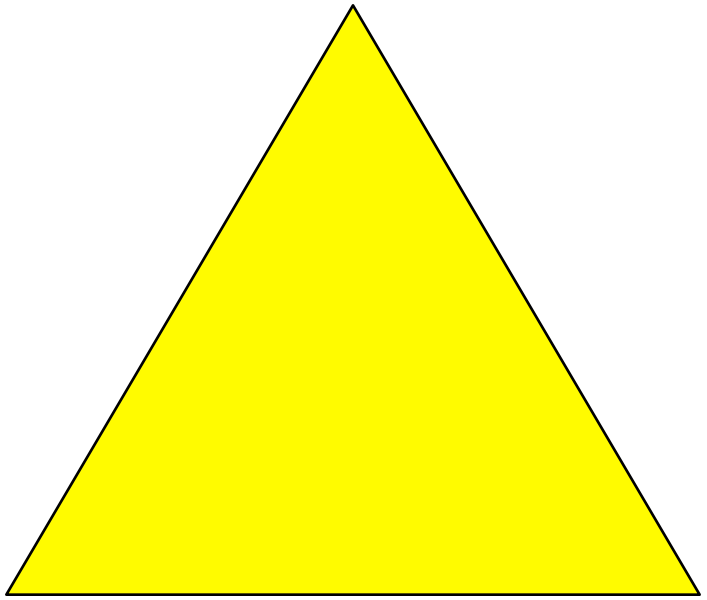
AlleleSeq.gersteinlab.org
J Rozowsky,
A Abyzov,
J Wang, P Alves, D Raha,
A Harmanci, J Leng,
R Bjornson,
Y Kong,
N Kitabayashi,
N Bhardwaj,
M Rubin,
M Snyder

Acknowledgements

Hiring Postdocs. See gersteinlab.org/jobs !

Default Theme

- Default Outline Level 1
  - Level 2



More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt .