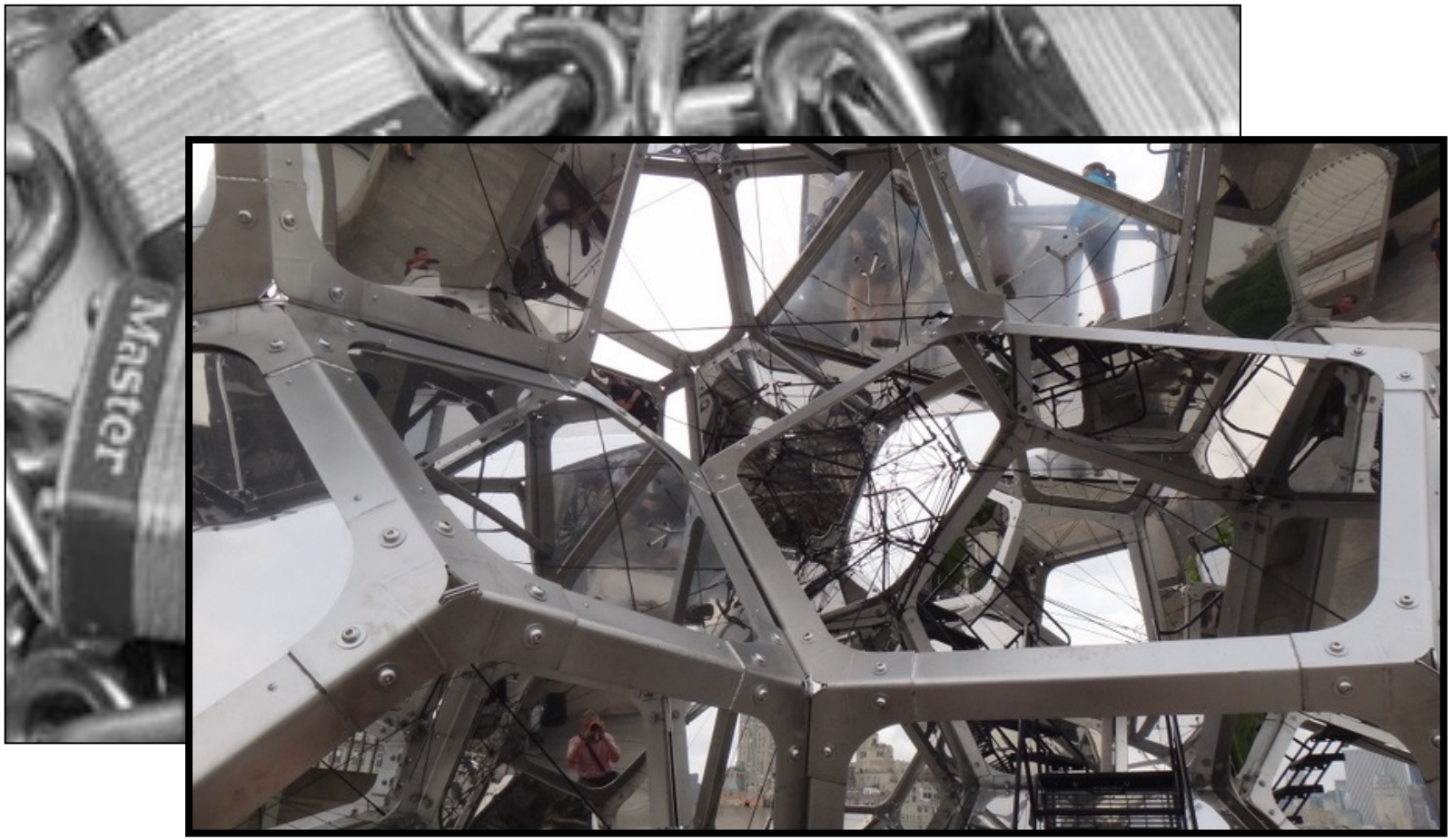
Genomic Privacy:
Intertwined Social & Technical Aspects

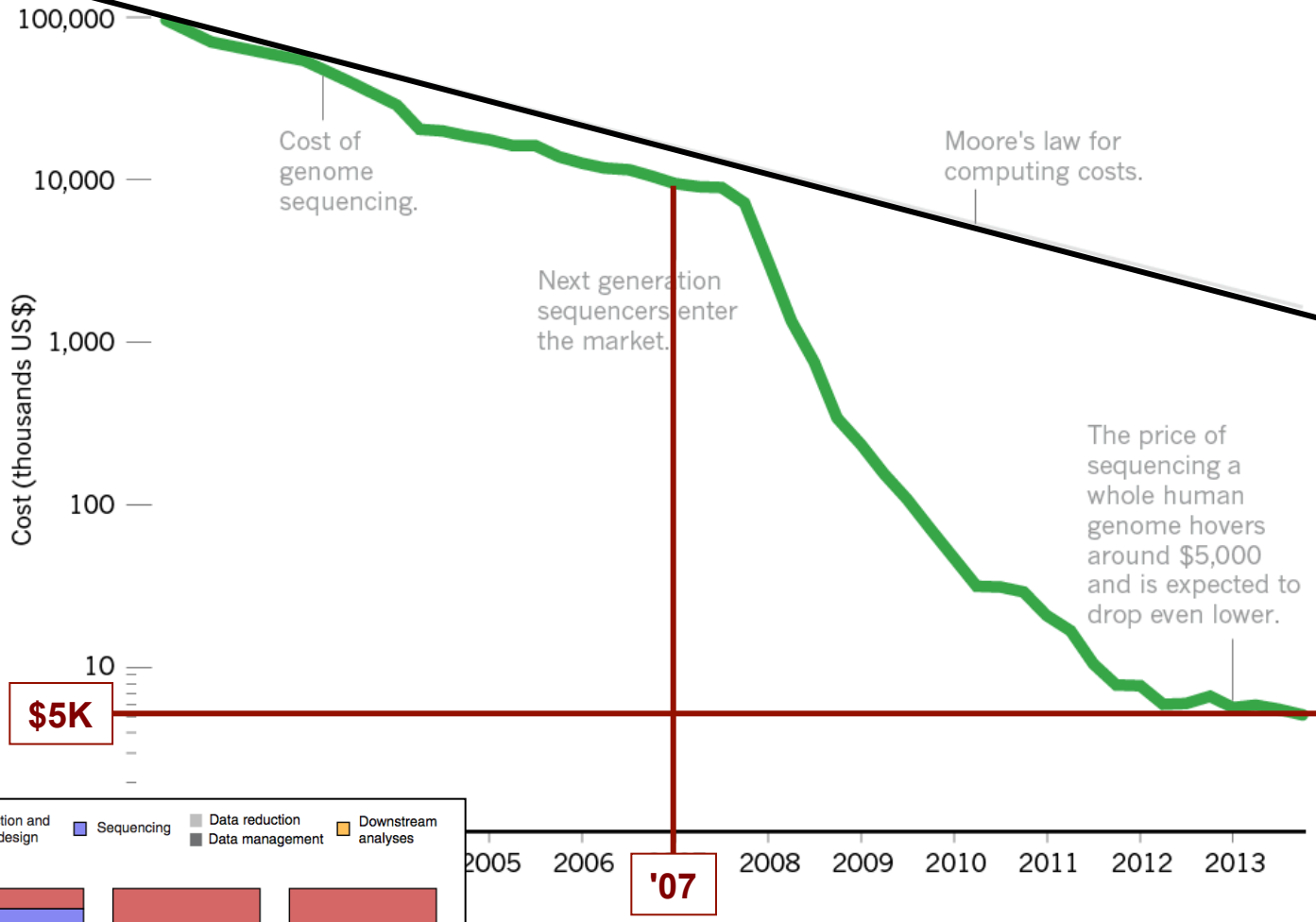


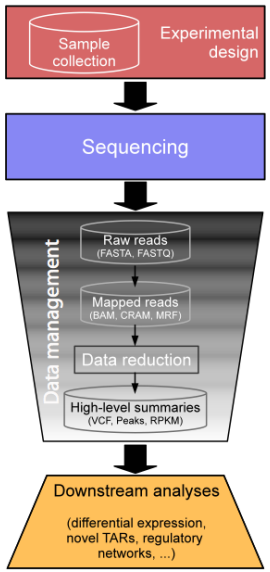
Mark Gerstein, Yale
Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein). See last slide for more info.

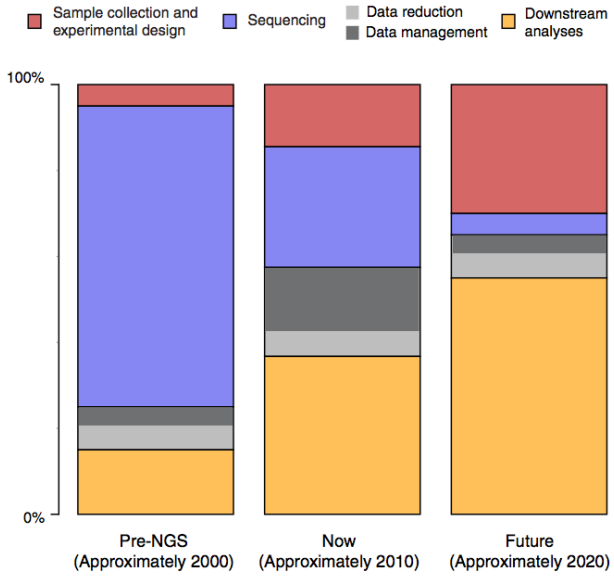
Setting the Stage: the Advent of Personal Genomics

- Human Genome sequence in 2000 for >$2 billion
- A Human Genome can be sequenced today for ~$1000
- Hundreds of thousands of SNPs can be interrogated for ~$99

The Explosion of Data in Genomics: the Numbers







From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Nature 507, 294; Sboner et al. ('11) GenomeBiology]

DTC Genomics

- Industry spurred by falling prices of sequencing and computation
- Major players were Navigenics, DeCode & 23andMe.

- 23andMe
  - has >1M Customers
  - $99 per analysis
  - Promotes sharing of Data
  - Currently in trouble with the FDA & limited to only recreational (e.g., ancestry related) analysis
  - Millions in VC funding

Genomic Privacy: Intertwined Social & Technical Issues

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

Privacy

Privacy is a personal and fundamental right guaranteed by the US Constitution

Privacy Act 1974

Including:

- **Inherent in the limits on the First Amendment is a constitutional right to privacy.**
- **Fourth Amendment against search and seizure *US v Amerson* 483 F. 3d 73 (2d Cir. 2007);**
- **Due Process Clauses of the Fifth and Fifteenth Amendments.**

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)]

Tricky Privacy Considerations in Personal Genomics

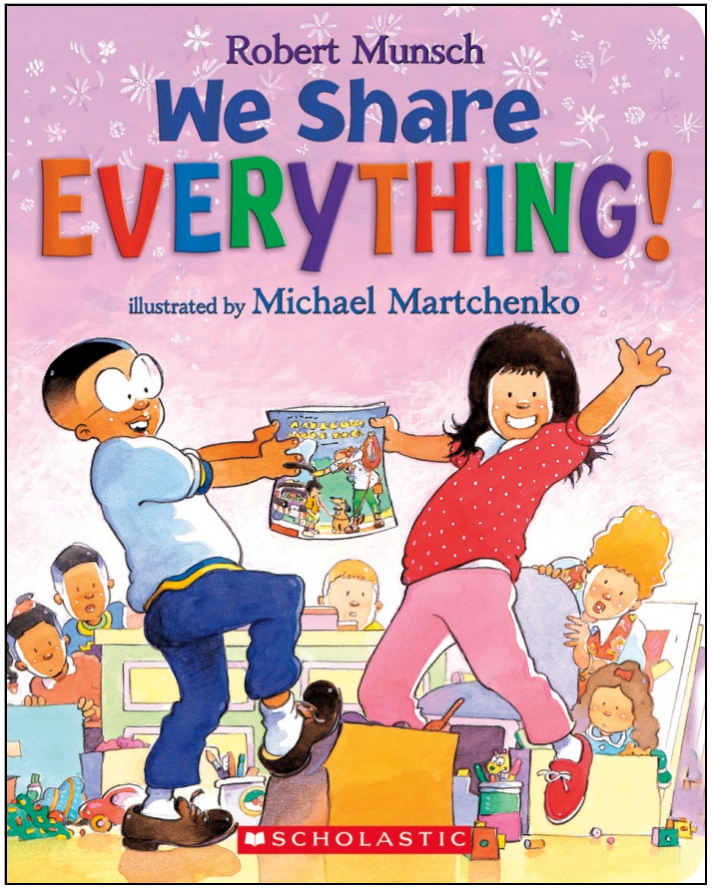
- Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
  - Genomic sequence very revealing about one's children. Is true consent possible?
  - Once put on the web it can't be taken back
- Culture Clash: Genomics historically has been a proponent of "open data" but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

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

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)]

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

Genetic Information Nondisclosure Act of 2008

- Title I relating to Health Insurance
- Title II relating to Employment Discrimination
- GINA Prohibits:
  - group and individual health insurers from using genetic data for determining eligibility or premiums
  - insurers from requesting that the insured undergo genetic testing
  - employers from using genetic data to may employment decisions
  - Employers from requesting genetic data about an employee or their family

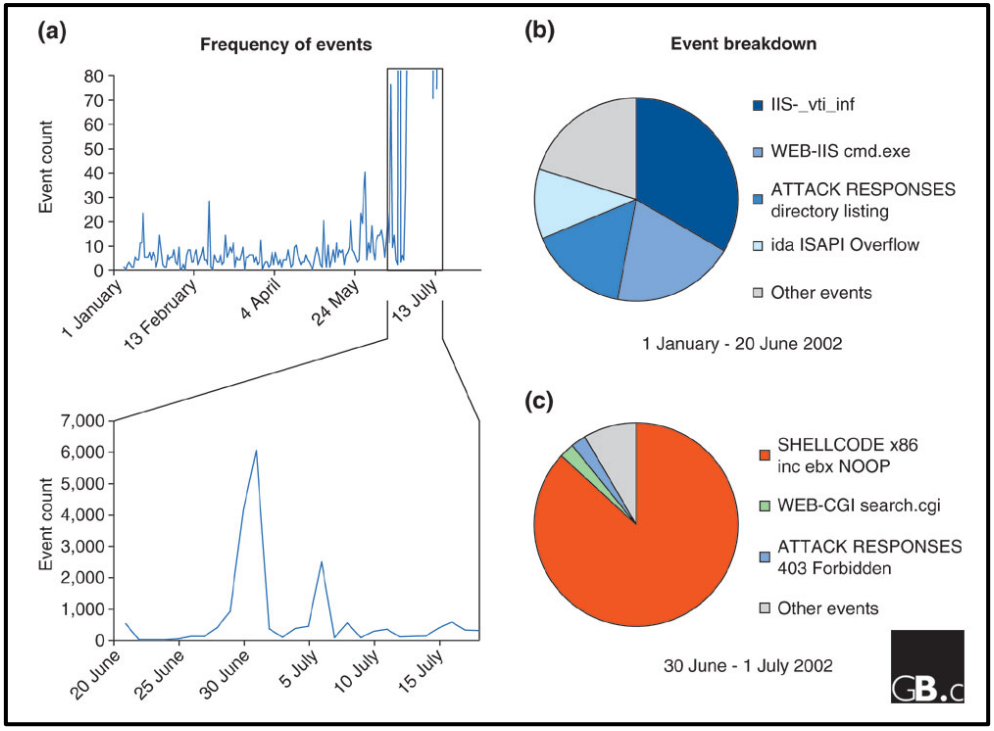
Current Social & Technical Solutions

- 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]

POLICYFORUM

ETHICS

Identifiability in Genomic Research

Genomic data are unique to the individual and must be managed with care to maintain public trust.

William W. Lowrance and Francis S. Collins

Genomic research can now readily generate data that cover significant portions of the human genome at levels of detail unique to individuals. Data can now be categorized with respect to disease-related genes and linked to clinical, family, and social data. Identifiability, the potential for such data to be associated with specific individuals, is therefore a pivotal concern. Research, health of privacy was among the issues examined by the National Institutes of Health (NIH) in a recent public consultation (*6*).

New Modes of Data Flow

Until recently, most genomic research used data and biospecimens obtained fairly directly, from the data subjects themselves or clinical repositories or specialized research Wellcome Trust Case Control Consortium do and U.K. Biobank will) (*7*). Among the design and governance issues are whether and how to de-identify the data and at what stages to conduct scientific and ethics review.

These new data flows, genomewide analyses, and novel arrangements such as the Informed Cohort scheme recently proposed by Kohane *et al.* (*8*) are relatively uncharted

Matching against reference genotype. The number of DNA markers such as single-nucleotide polymorphisms (SNPs) that are needed to uniquely identify a single person is small; Lin *et al.* estimate that only 30 to 80 SNPs could be sufficient

Linking to nongenetic databases. A second route to identifying genotyped subjects is deduction by linking and then matching geno-type- plus-associated data (such as gender, age, or disease being studied) with data in health-care, administrative, criminal, disaster response, or other databases … If the nongenetic data are overtly identified, the task is straightforward

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1*]

Identifying anonymized 1000G individuals through DB xref

OPEN ACCESS Freely available online PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1*]

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

A framework for accurately and robustly resolving whether individuals are in a complex genomic DNA mixture using high-density single nucleotide polymorphism (SNP) genotyping microarrays.

We demonstrate an approach for rapidly and sensitively determining whether a trace amount (<1%) of genomic DNA from an individual is present within a complex DNA mixture

European Journal of Human Genetics

Search

Journal home > Archive > Letters > Full text

Journal home
Advance online publication
About AOP
Current issue
Archive
Practical Genetics
Gene Cards
Focuses
News

Online submission
For authors
For referees

Letter

European Journal of Human Genetics (2009) **17,** 147–149; doi:10.1038/ejhg.2008.198; published online 22 October 2008

On Jim Watson's *APOE* status: genetic information is hard to hide

Dale R Nyholt[1], Chang-En Yu[2] and Peter M Visscher[1]

[1]Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, QLD, Australia
[2]Division of Gerontology and Geriatric Medicine, Department of Medicine, Geriatric Research, Education, and Clinical Center, Veteran Affairs Puget Sound Health Care System, University of Washington School of Medicine, Seattle, WA, USA

Correspondence: Dr DR Nyholt, Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, Queensland QLD 4006, Australia. Tel: 61 7 3362 0258; Fax: 61 7 3362 0101; E-mail: daleN@qimr.edu.au

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.

Cross correlated small set of identifiable IMDB movie database rating records with large set of "anonymized" Netflix customer ratings

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 (eg photos of eye color)**
- **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)]

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

RNA-seq

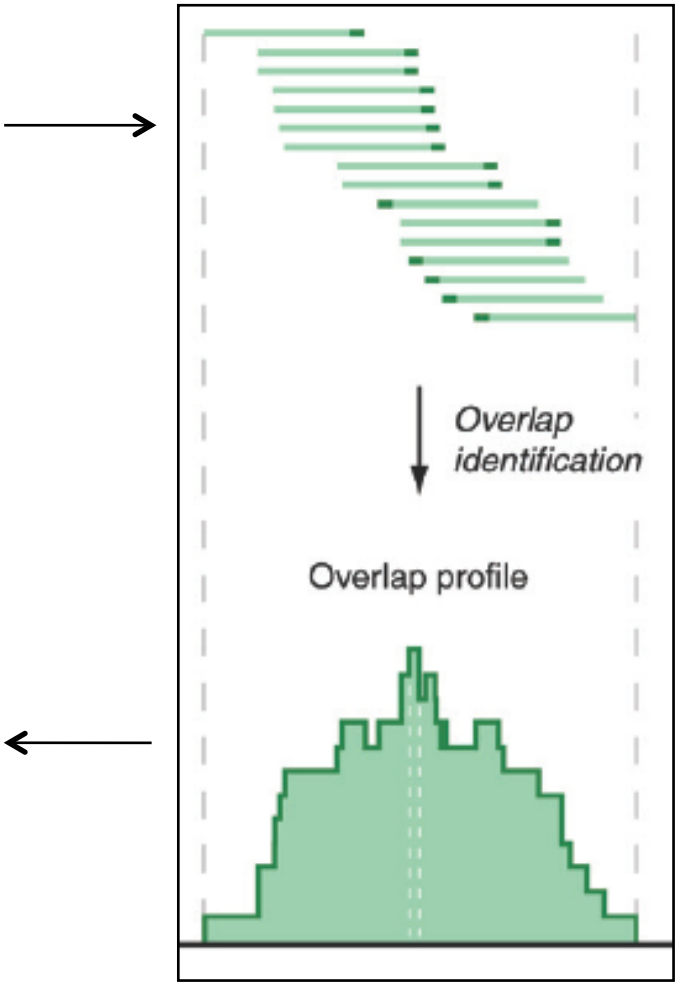
- Genome-wide transcription
  - Gene transcription and non-canonical transcription
  - Fusion transcripts
  - Alternative splicing
  - Allele specific expression
- Large magnitude of RNA-seq data generated
  - ENCODE, modENCODE, TCGA, GTEx, Roadmap, psychENCODE, etc.
  - Mostly the data is about the phenotype (e.g., cancer gene expression), but the individual information often comes along as collateral
  - Maybe we can separate private info but couple it with the public presentation?

RNA-seq

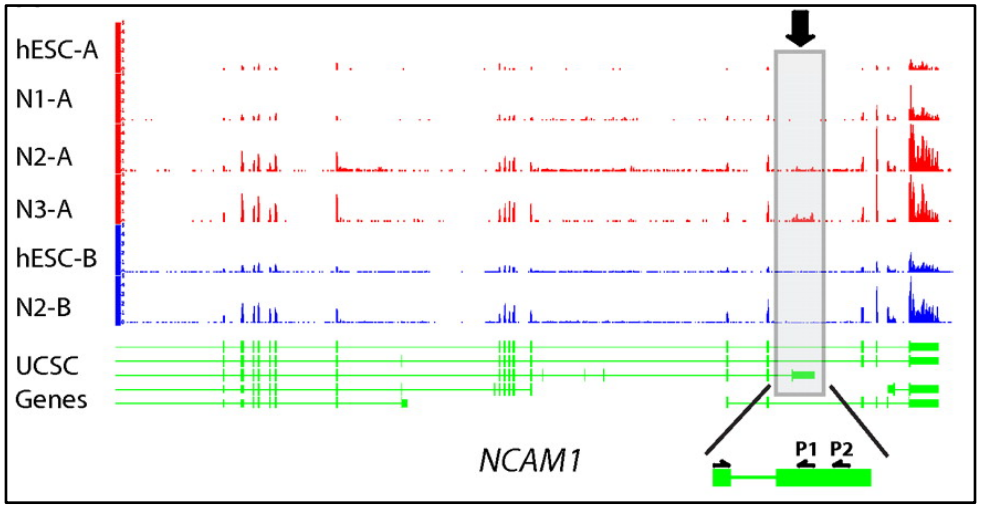
RNA-seq uses next-generation sequencing technologies to reveal RNA presence and quantity within a biological sample.

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

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



Reads => Signal

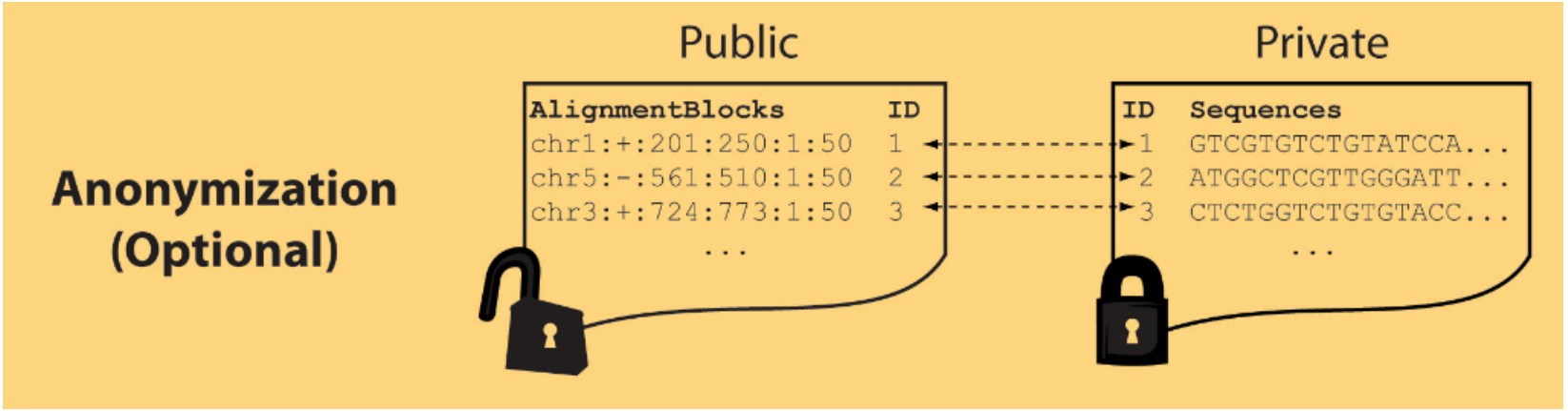


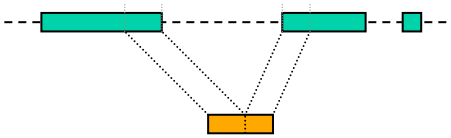
Quantitative information from RNA-seq signal: average signals at exon level (RPKMs)

[*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

10X Compression Ex.

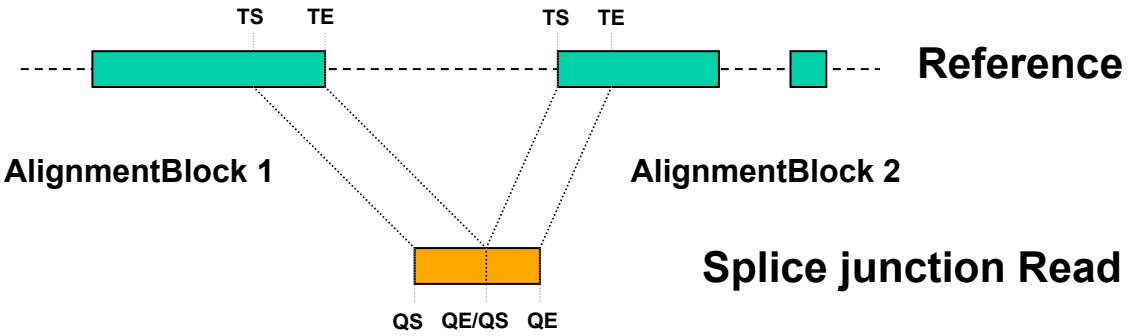
Raw ELAND export file has uncompressed file size: ~4 GB; total number of reads: ~20 million; number of mapped reads: ~12 million .

MRF file is significantly smaller (~400 MB uncompressed, ~130 MB compressed with gzip).

BAM file
has a size of ~1.2 GB.

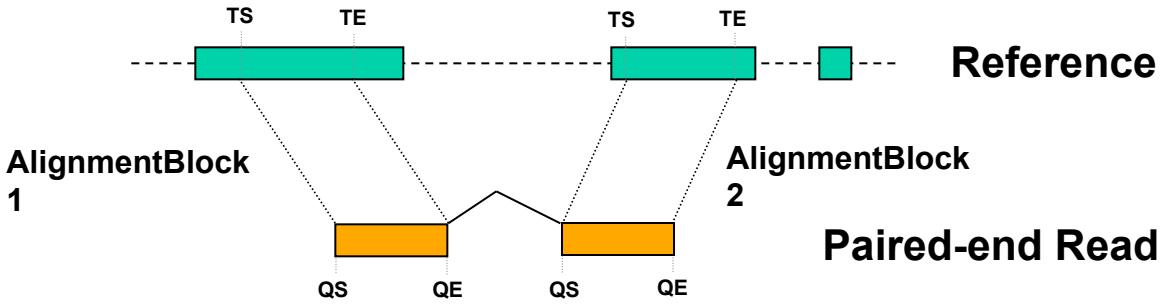
Reference based compression (ie CRAM) is similar but it stores actual variant beyond just position of alignment block

`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

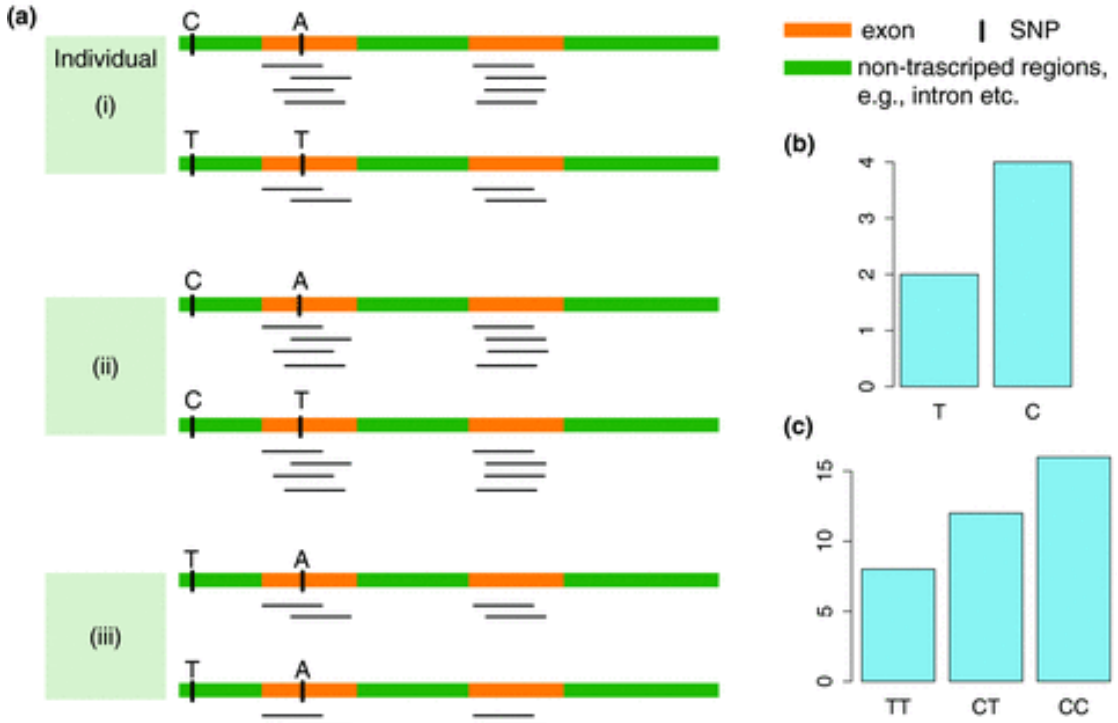
[Habegger et al., Bioinformatics ('11)]

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

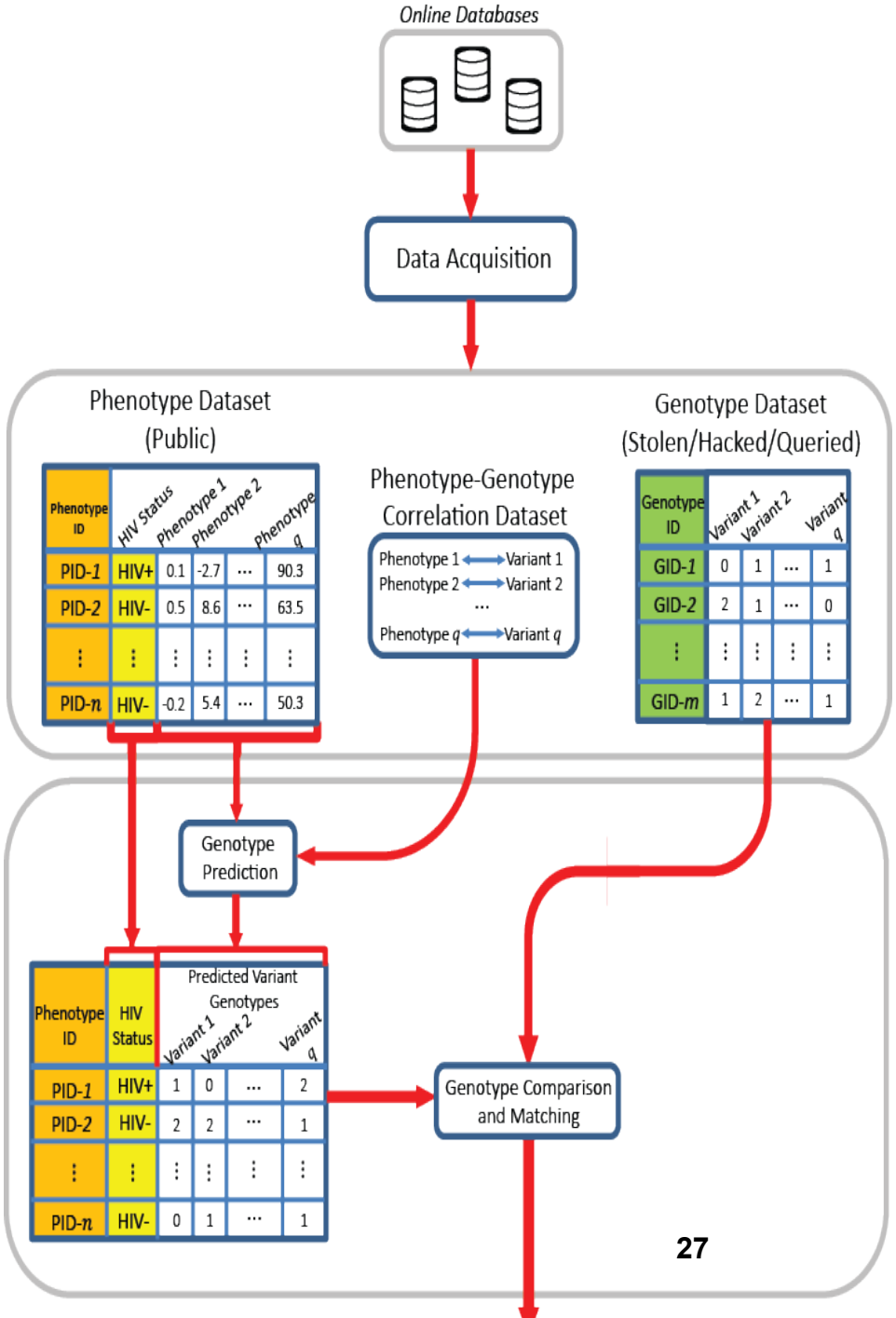
eQTL Mapping Using RNA-Seq Data



- 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

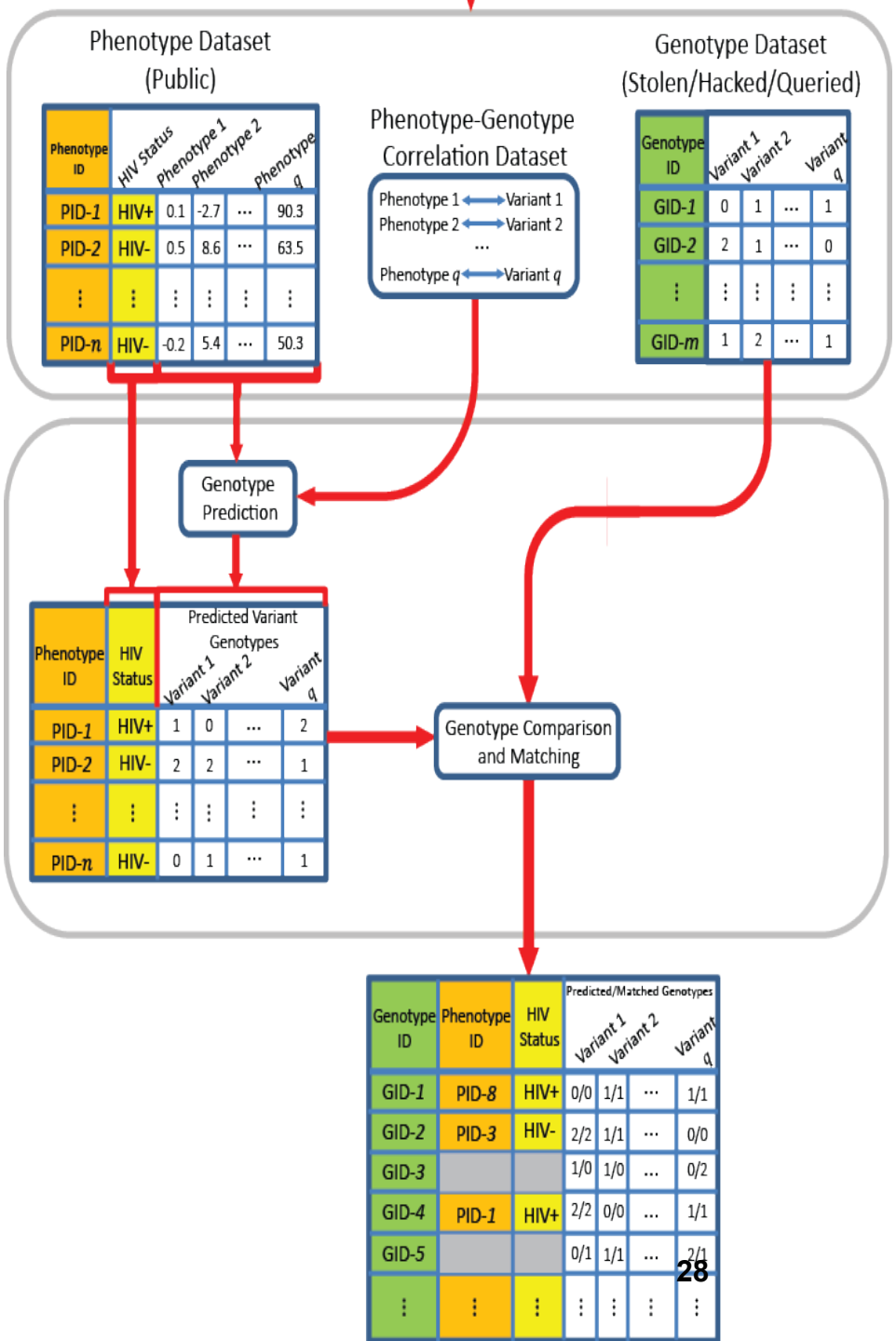
[*Biometrics 68(1) 1–11*]

Linking Attack Scenario



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

Linking Attack Scenario



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

Quantifying ICI Leaked and Predictability

- Amount of individual characterizing Information (ICI) in a set of *n* variants:
  - How many rare genotypes there are in the set?

$$ICI(\{V_1 = g_1, V_2 = g_2, \ldots, V_n = g_n\}) = \overbrace{\sum_{k=1}^{n} \underbrace{-\log\left(p(V_k = g_k)\right)}_{\text{Convert the genotype frequency to number of bits that can be used to characterize individual}}}^{\text{Sum individual characterizing information from all variants}}$$

Variant 1 Genotype | Variant 2 Genotype | Variant *n* Genotype

$(g_i \in \{0,1,2\})$

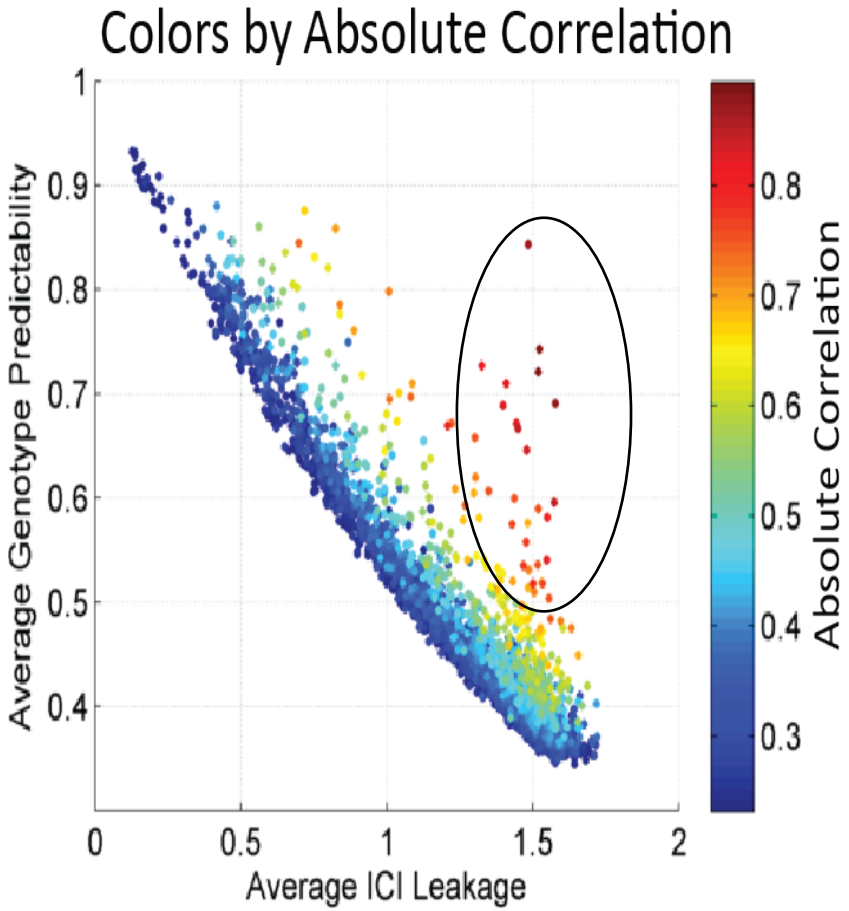
- Predictability of genotypes given expression levels (π): Given that the k_{th} gene's expression level is $e_{k,j}$, how much randomness is left in the genotype?

$$\pi\left(V_k | E_k = e_{k,j}\right) = \underbrace{\exp\left(-1 \times \overbrace{H(V_k | E_k = e_{k,j})}^{\text{Randomness left in } V_k \text{ given } E_k = e_{k,j}}\right)}_{\text{Convert the entropy to average probability}}$$

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

Relating Predictability to ICI Leakage – A few key SNPs with good predictability & Great Leakage.

Perhaps these could removed?



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

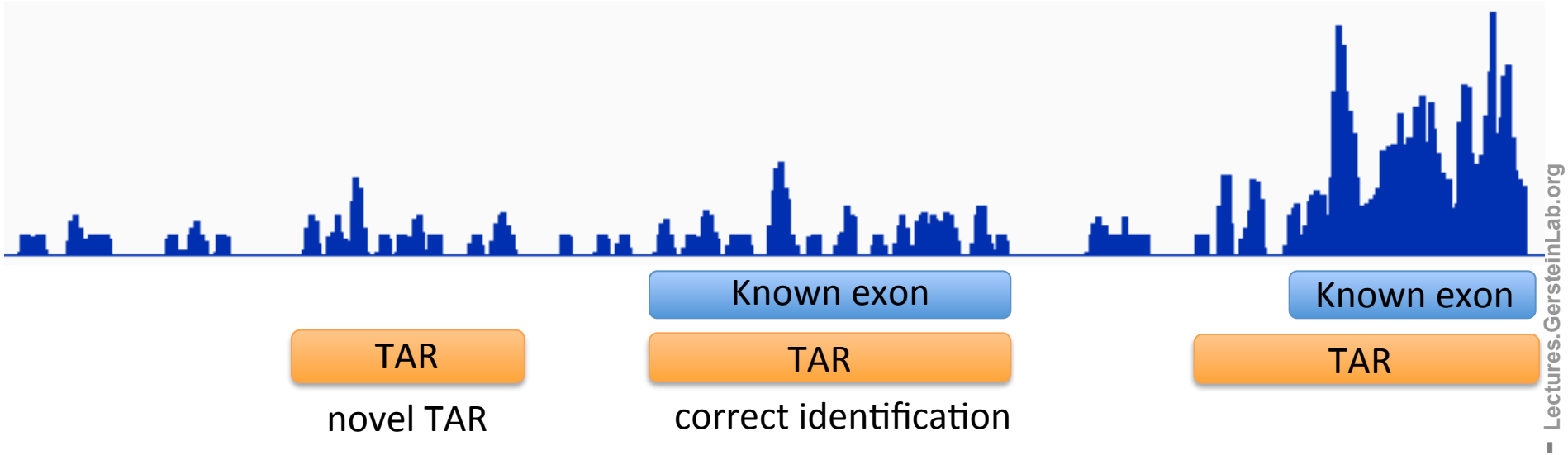
- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

Discovering Transcriptionally Active Regions (novel RNA contigs)

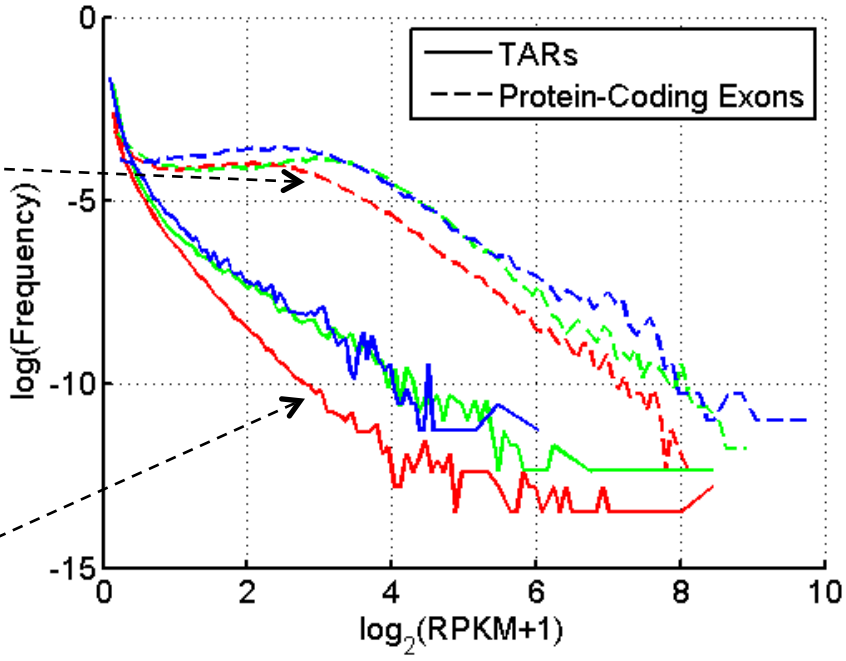
- Cluster reads setting minimum-run and maximum gap parameters for newly identified transcribed regions (TARs)
- Assess exon discovery rates for known genes and noncoding RNAs



ENCODE RNA-Seq Data [*Nature* 512:445 ('14)]

Non-Canonical Transcription

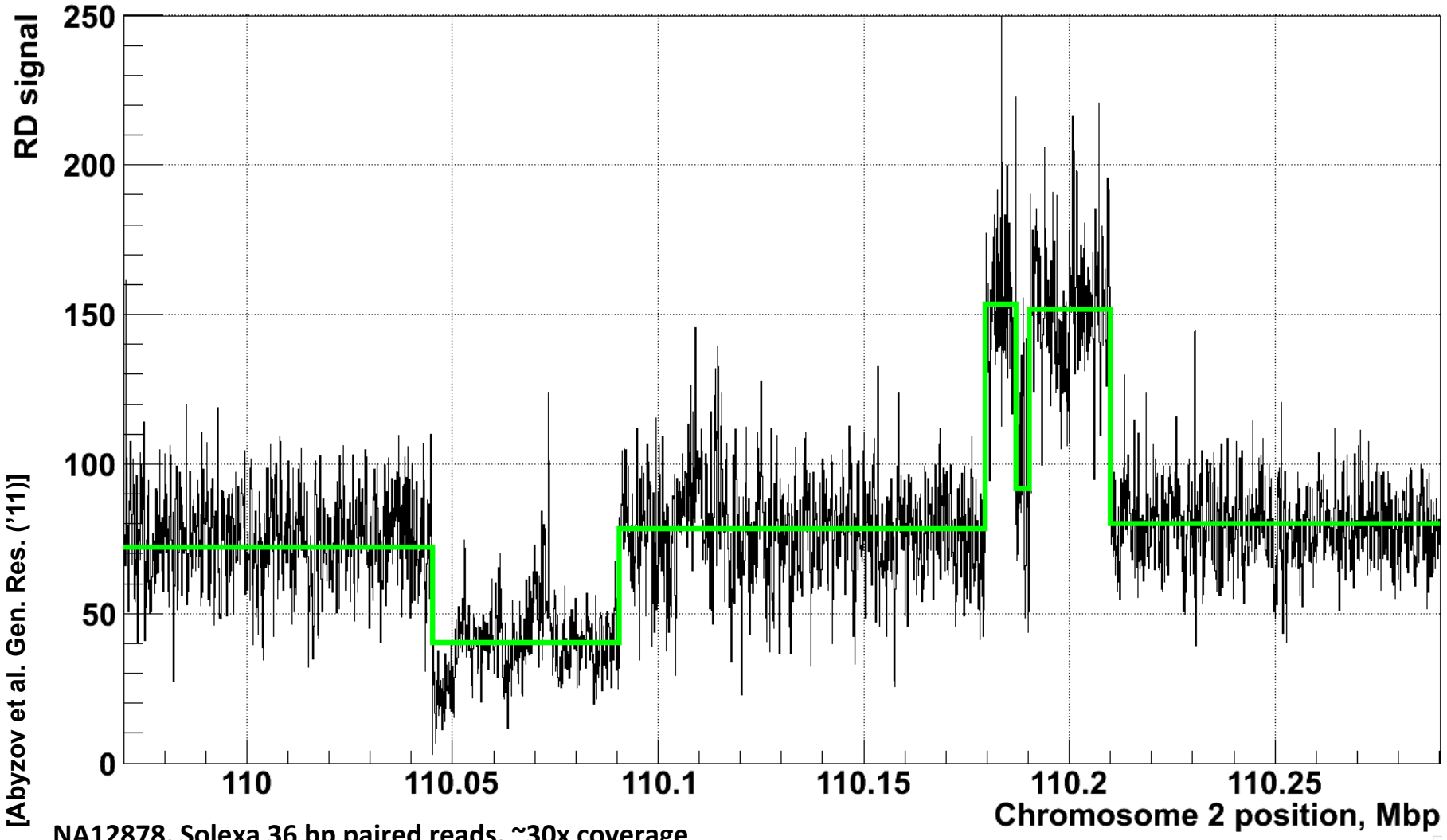
	Human		
	Elements	Genome Coverage	
		Kb	%
mRNAs (exons)	20,007	86,560	3.0
Pseudogenes	11,216	27,089	0.95
Total ncRNAs	22,154	17,770	0.62
Regions Excluding mRNAs, Pseudogenes or Annotated ncRNAs	283,816	2,731,811	95.5
Transcription Detected (TARs)	708,253	916,401	32.0



- 4.5% of human genome are transcribed and associated with standard annotations;
- 32% of human genome give rise to TARs or non-canonical transcription, i.e., transcription from genomic regions not associated with standard annotations;
- Non-canonical transcripts show lower transcription level compared to protein coding transcripts

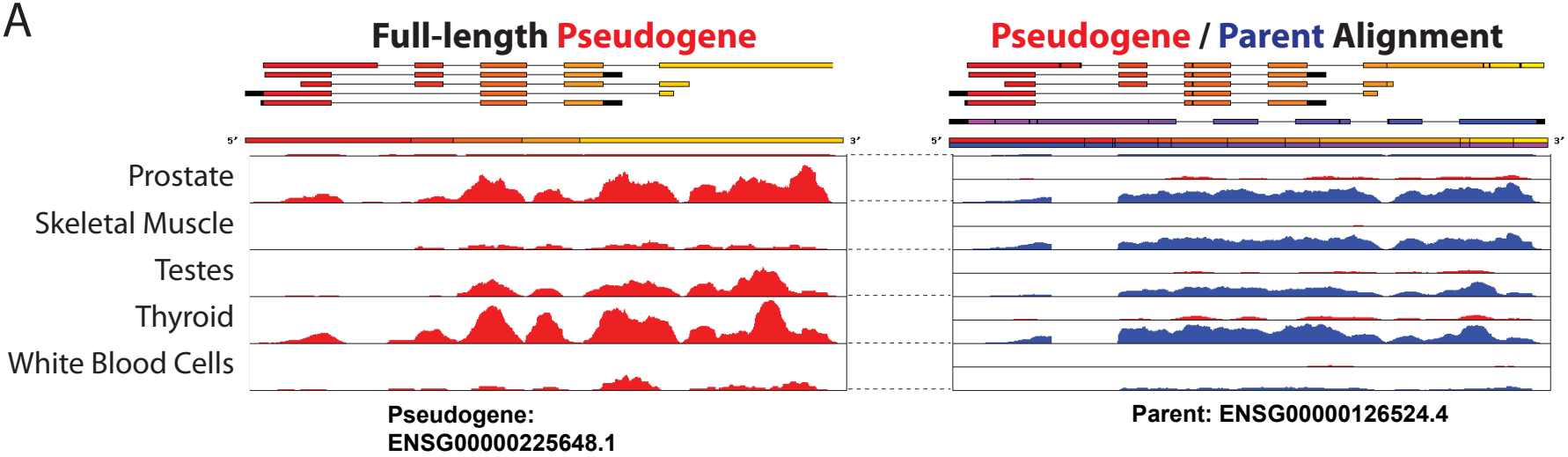
[*Nature 512(7515):445-8*]

Example of Application of CNVnator to RD data

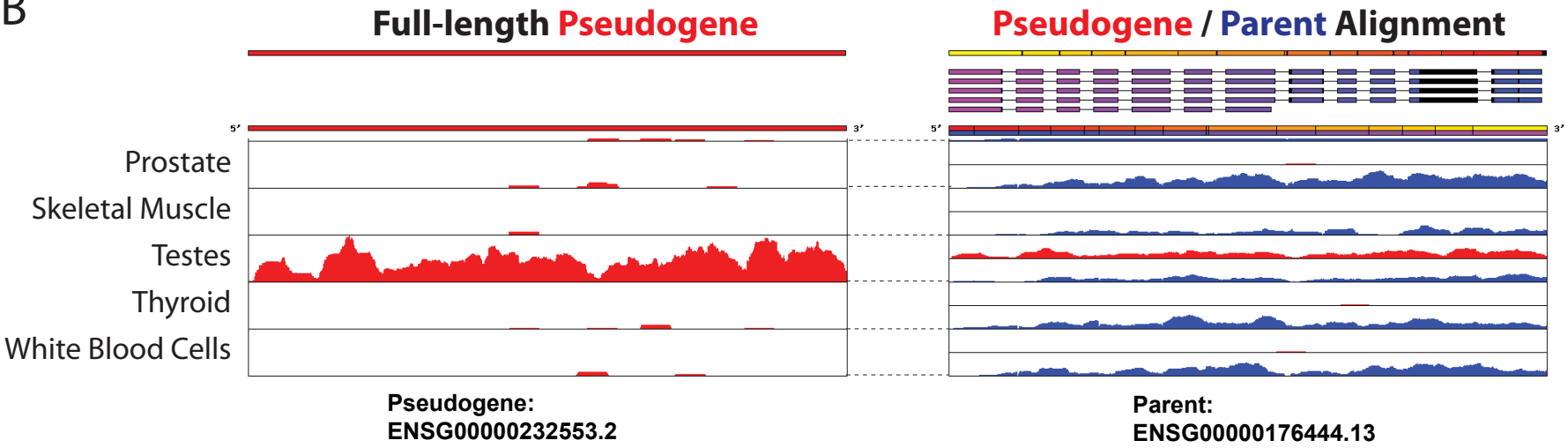


Many non-canonical transcripts are real but some potentially reflecting mis-mapping from genes

A



B



- Non-coding transcription may correlate with coding transcription
- Potential mapping artifacts: reads from coding regions mapped to non-coding elements or vice versa due to sequence similarity

[*Genome Biol. 13: R51*]

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

Genomic Privacy: Intertwined Social & Technical Issues

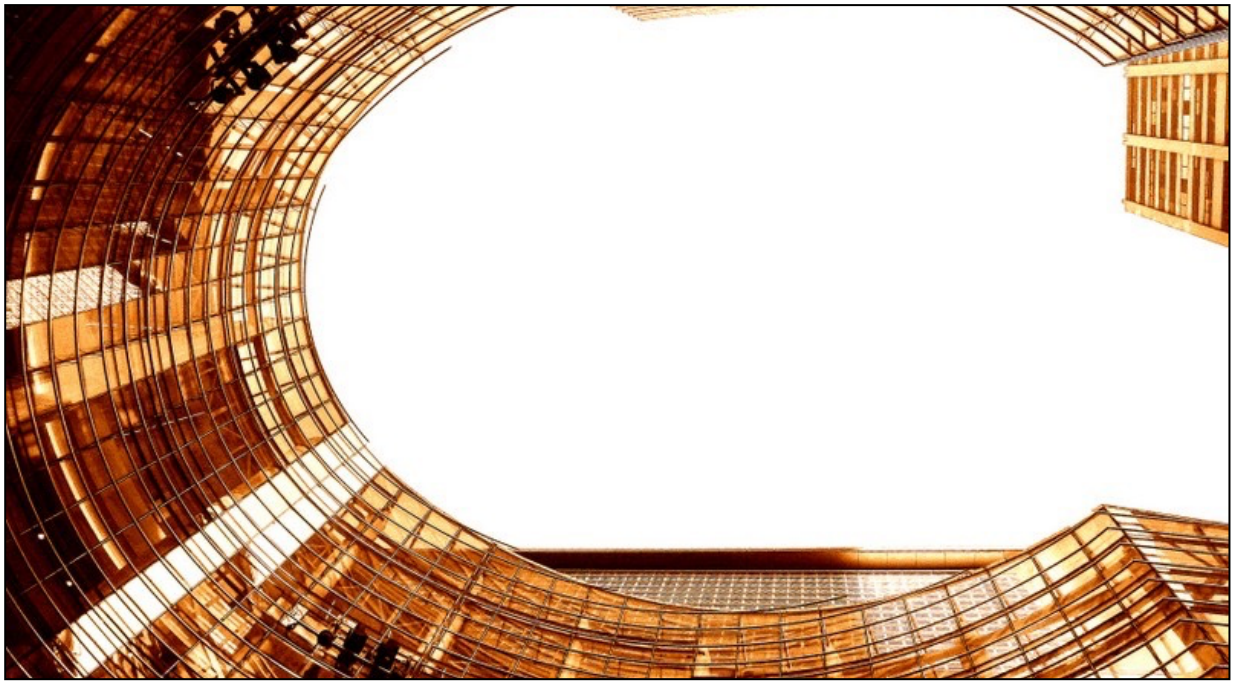
- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

Genomic Privacy: Intertwined Social & Technical Issues

- Setting the Stage: the Advent of Personal Genomics
- **The Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private vs the need for large-scale data-sharing to enable med. research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Issues: burdensome security, inconsistencies + ways the solutions have been partially **"hacked"**
  - Strawman **Hybrid Soc-Tech Proposal** (Licensure, Cloud Enclaves. **Quantifying Leaks, & Closely Coupled priv.-public data**)

- **RNA-seq – practical problem of publicly sharing some of the info**
  - Removing SNVs in reads using **MRF**
  - Quantifying & removing variant info from expression levels + **eQTLs**
  - Further complications: SVs in pervasive transcription?

Acknowledgements



CNVnator
A Abyzov

Cost of sequencing
A Sboner, XJ Mu

Hiring Postdocs. See gersteinlab.org/jobs !

Data privacy

[papers.gersteinlab.org/subject/privacy]

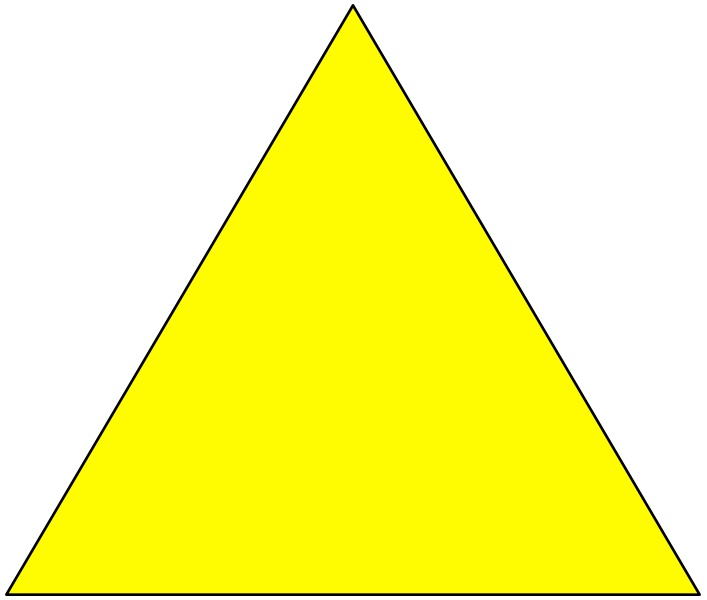
A Harmanci, D Greenbaum

ENCODE pseudogenes & transcriptome
B Pei, L Habegger, J Rozowsky, A Harmanci, KK Yan

RSEQtools
L Habegger, A Sboner, TA Gianoulis, J Rozowsky, A Agarwal, M Snyder

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 .