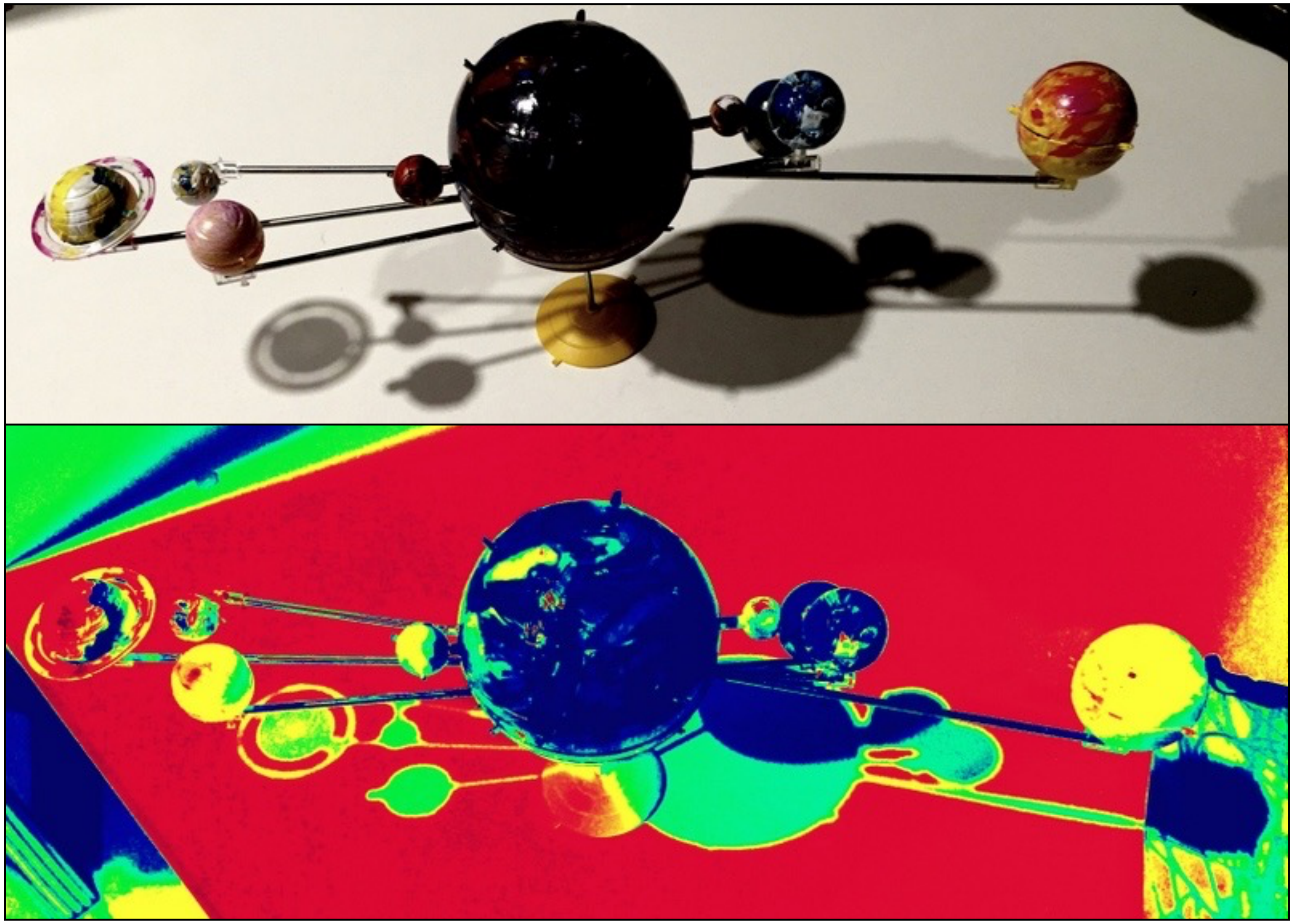
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

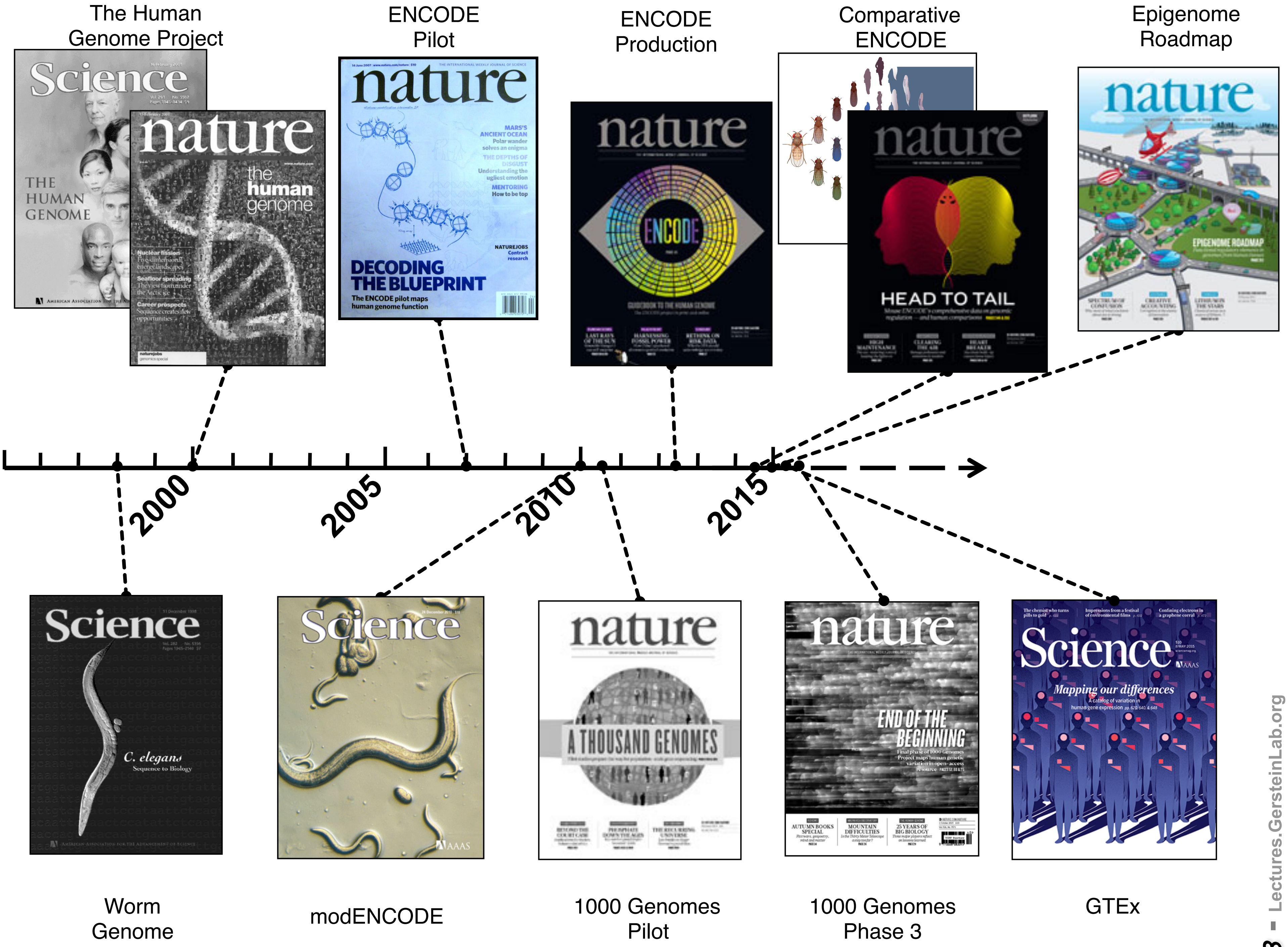


Mark Gerstein, Yale

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

Large-scale RNA

- Recent advent of many consortia & group producing large scale RNA-seq following on DNA sequencing
- Often this is of human subjects (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

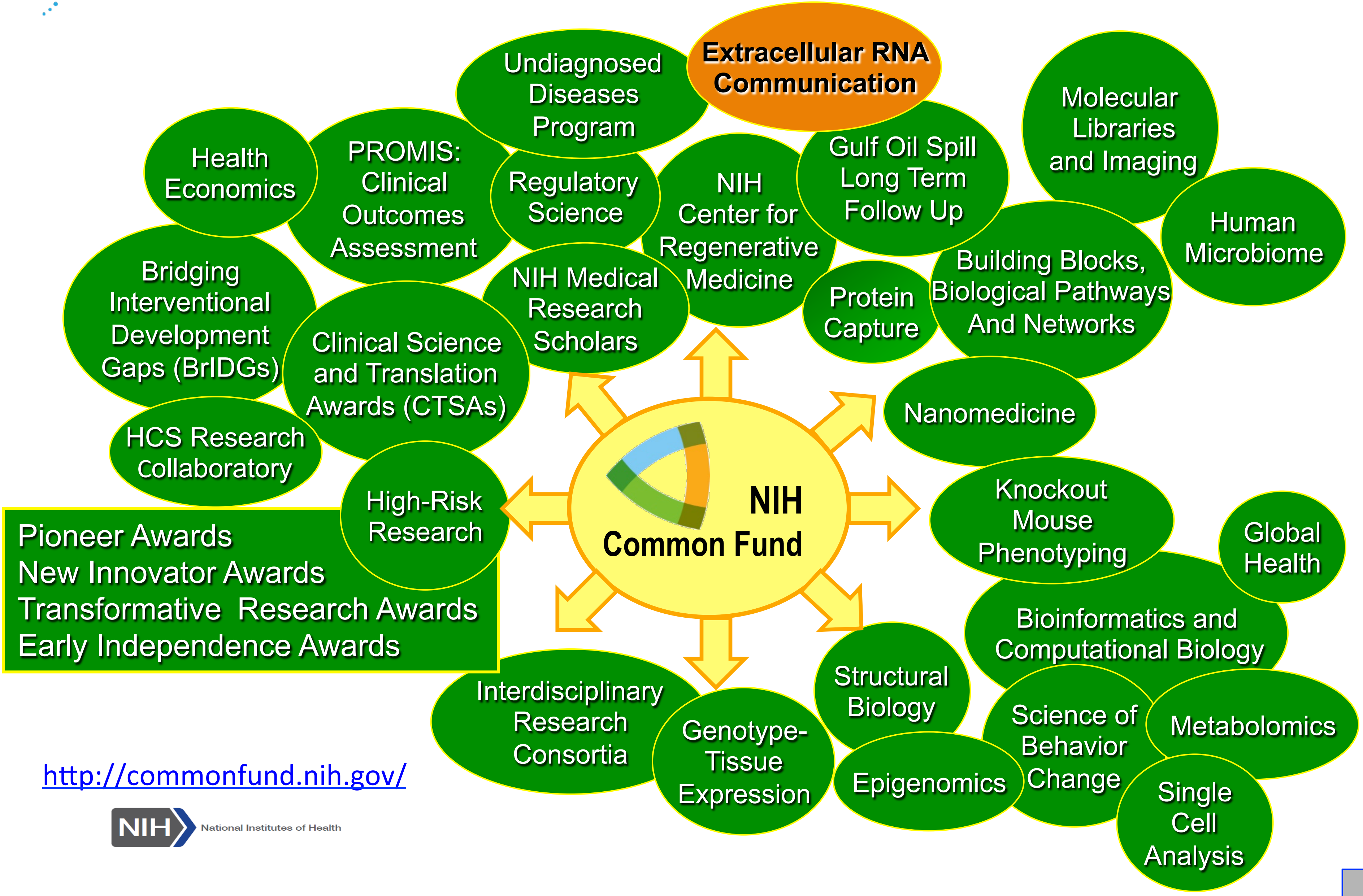
modENCODE

1000 Genomes Pilot

1000 Genomes Phase 3

GTEx

Common Fund Programs



NIH Common Fund

Undiagnosed Diseases Program

Extracellular RNA Communication

Molecular Libraries and Imaging

Health Economics

PROMIS: Clinical Outcomes Assessment

Regulatory Science

NIH Center for Regenerative Medicine

Gulf Oil Spill Long Term Follow Up

Human Microbiome

Bridging Interventional Development Gaps (BrIDGs)

NIH Medical Research Scholars

Protein Capture

Building Blocks, Biological Pathways And Networks

Clinical Science and Translation Awards (CTSAs)

HCS Research Collaboratory

Nanomedicine

High-Risk Research

Pioneer Awards
New Innovator Awards
Transformative Research Awards
Early Independence Awards

Knockout Mouse Phenotyping

Global Health

Bioinformatics and Computational Biology

Interdisciplinary Research Consortia

Genotype-Tissue Expression

Structural Biology

Epigenomics

Science of Behavior Change

Metabolomics

Single Cell Analysis

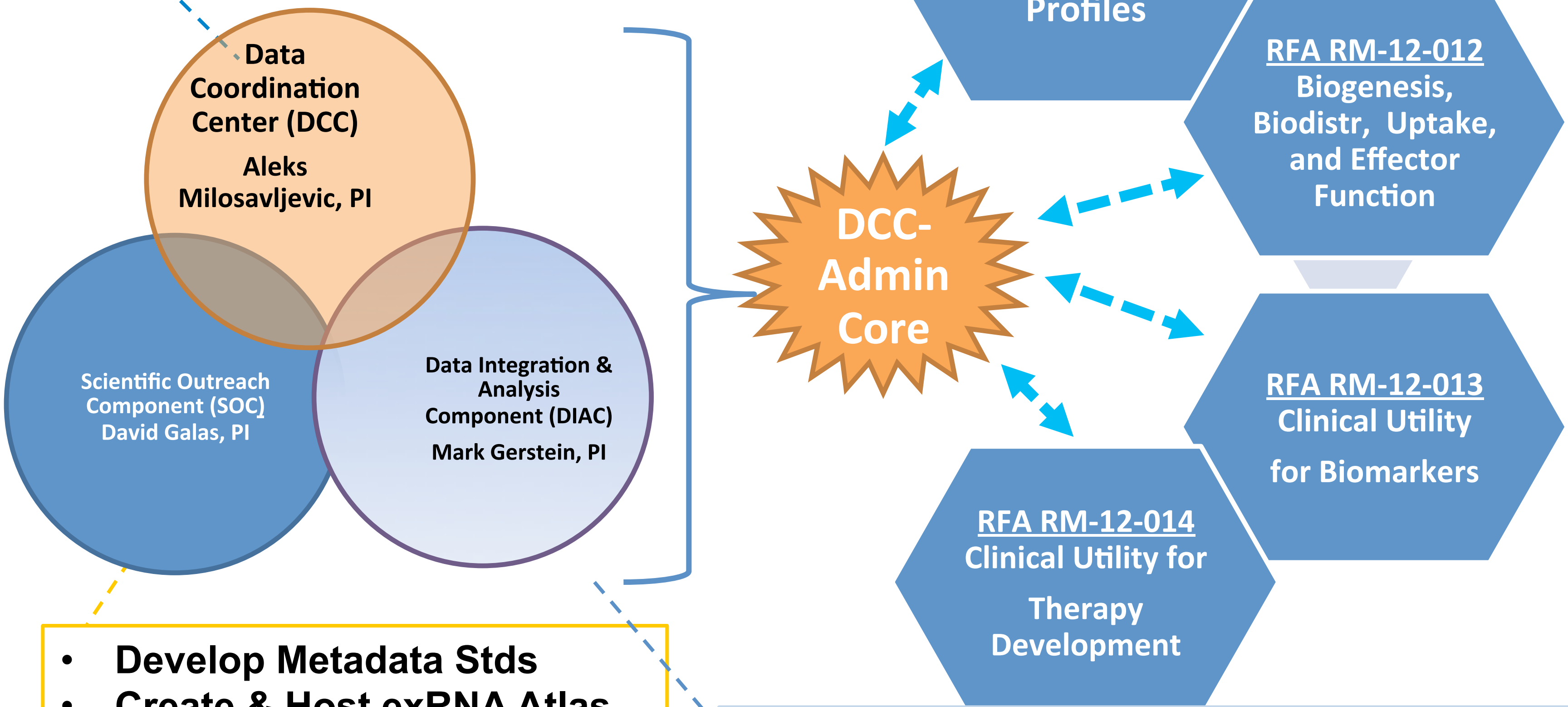
http://commonfund.nih.gov/

NIH National Institutes of Health

ERCC Organization

Data Management & Resource Repository (DMRR)

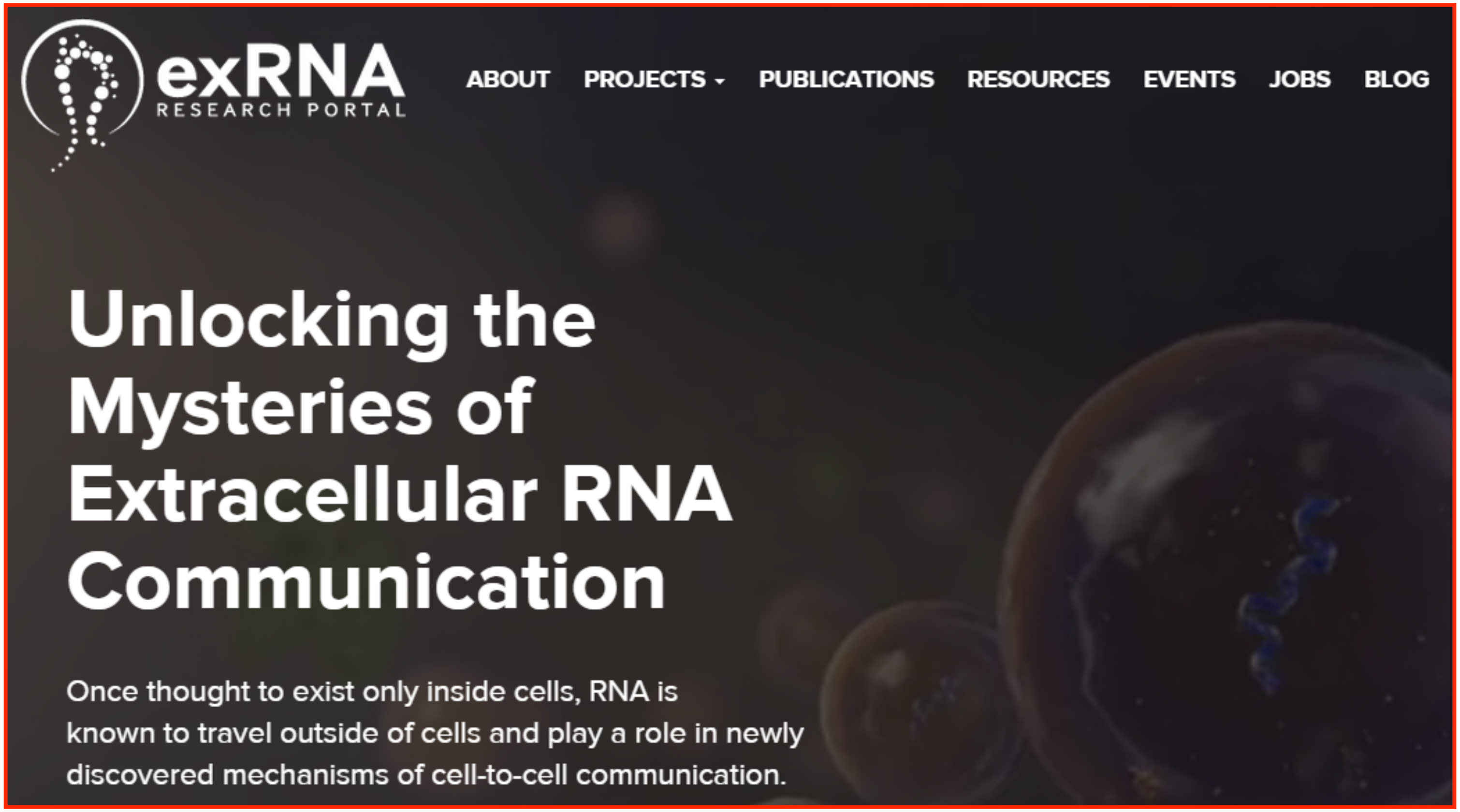
- Curated Gene Context, Network Modules, Pathways



- Develop Metadata Stds
- Create & Host exRNA Atlas
- Data Analysis & Visualizaton

- Develop Analysis Pipelines/Tools
- Integrative Analysis
- Profile/Gene/Network Modules

eXRNA Portal

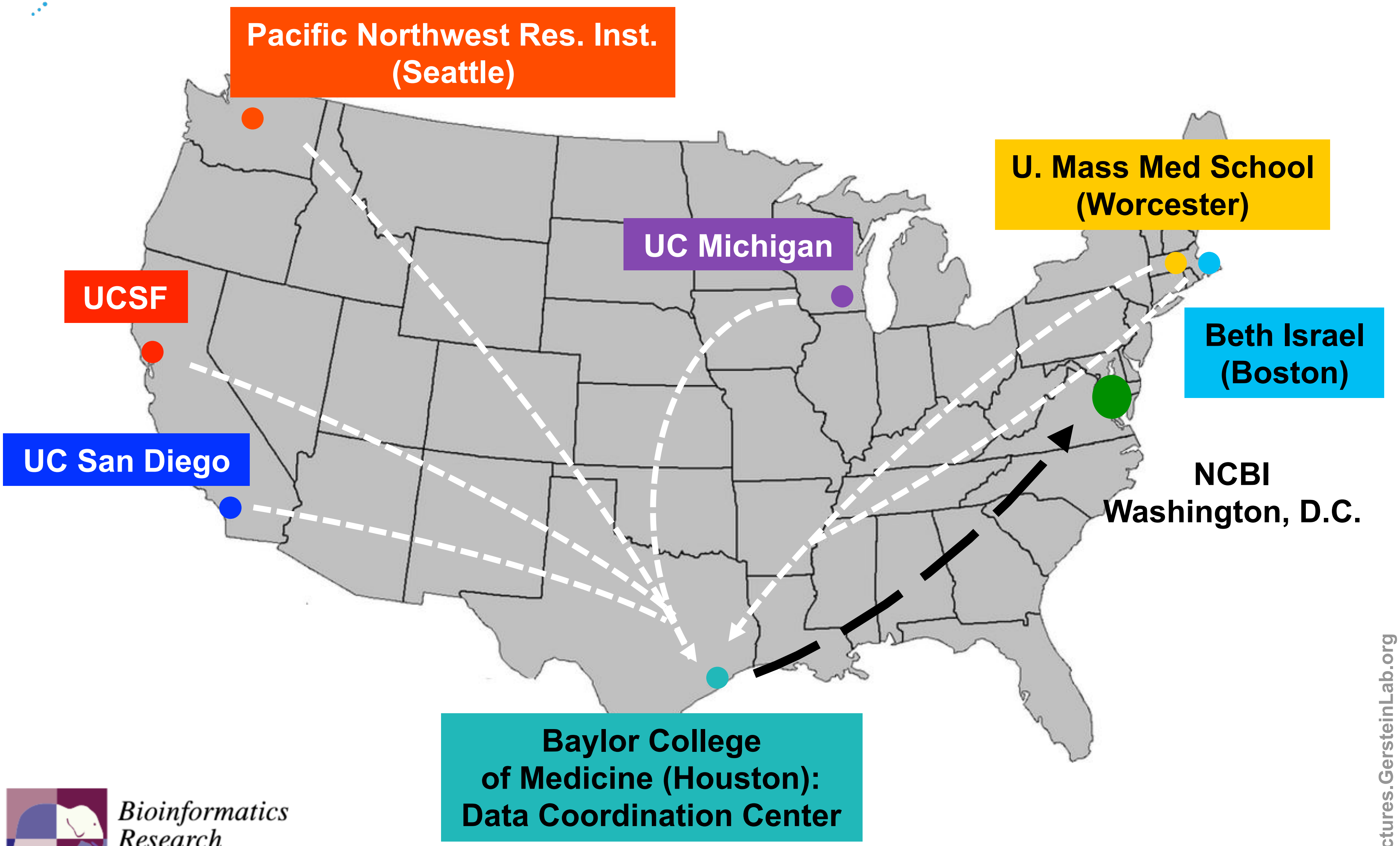


Unlocking the Mysteries of Extracellular RNA Communication

Once thought to exist only inside cells, RNA is known to travel outside of cells and play a role in newly discovered mechanisms of cell-to-cell communication.

www.exrna.org

Regional exRNA Mapping Centers and Data Coordination Center


Pacific Northwest Res. Inst. (Seattle)
U. Mass Med School (Worcester)
UC Michigan
UCSF
Beth Israel (Boston)
UC San Diego
NCBI Washington, D.C.
Baylor College of Medicine (Houston): Data Coordination Center

Bioinformatics Research Laboratory

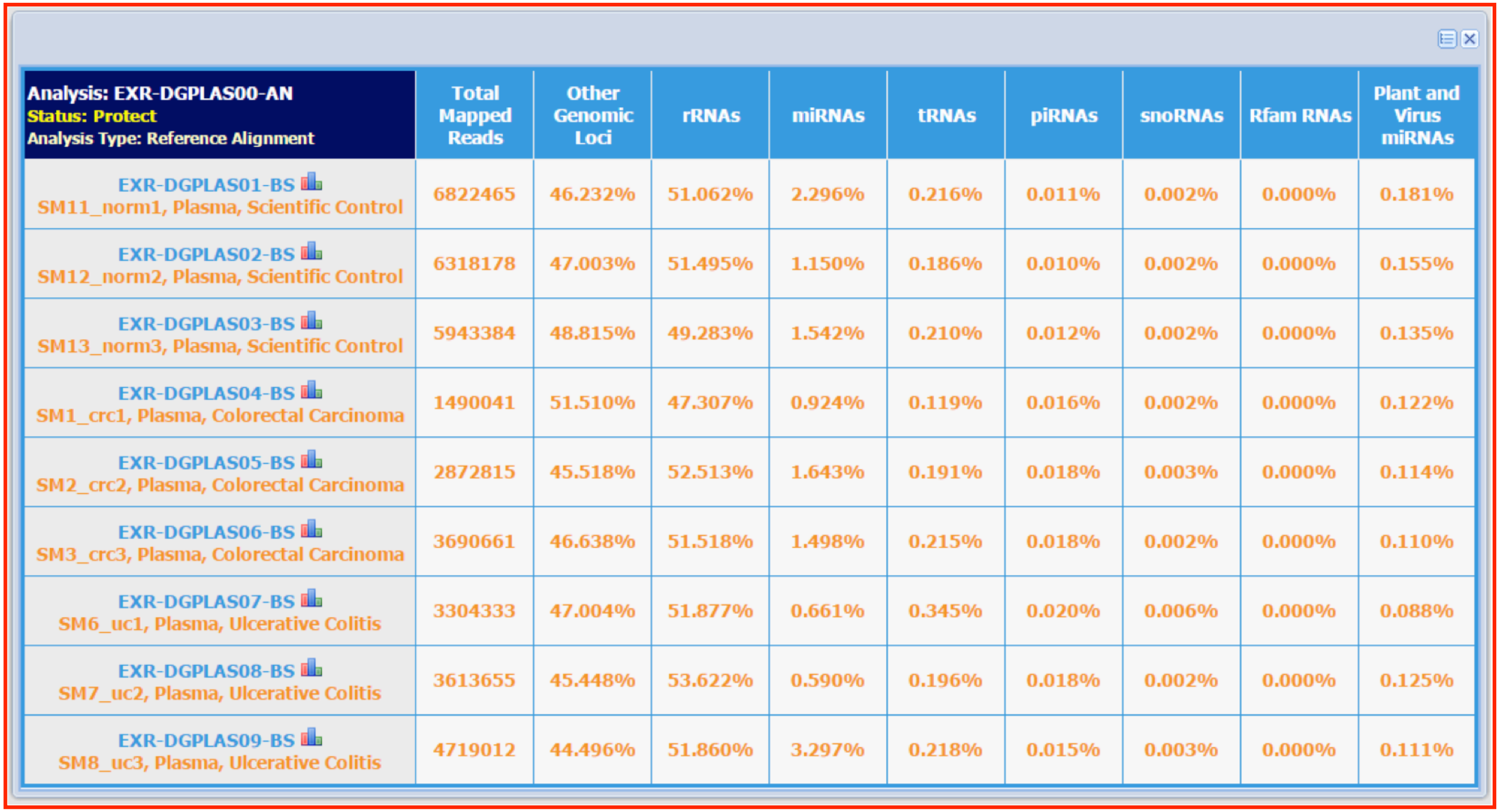
RNA-Seq Profiling of Human exRNAs – Multiple Biological Fluids

- Short & long non-coding, non-coding, circular coding

- Plasma
- Serum
- Urine
- Saliva
- Cerebrospinal fluid
- Cord blood
- Seminal fluid
- Bronchoaveolar fluid
- Placenta

- Endogenous vs exogenous (environment/diet)

ExRNA Atlas



Analysis: EXR-DGPLAS00-AN Status: Protect Analysis Type: Reference Alignment	Total Mapped Reads	Other Genomic Loci	rRNAs	miRNAs	tRNAs	piRNAs	snoRNAs	Rfam RNAs	Plant and Virus miRNAs
EXR-DGPLAS01-BS SM11_norm1, Plasma, Scientific Control	6822465	46.232%	51.062%	2.296%	0.216%	0.011%	0.002%	0.000%	0.181%
EXR-DGPLAS02-BS SM12_norm2, Plasma, Scientific Control	6318178	47.003%	51.495%	1.150%	0.186%	0.010%	0.002%	0.000%	0.155%
EXR-DGPLAS03-BS SM13_norm3, Plasma, Scientific Control	5943384	48.815%	49.283%	1.542%	0.210%	0.012%	0.002%	0.000%	0.135%
EXR-DGPLAS04-BS SM1_crc1, Plasma, Colorectal Carcinoma	1490041	51.510%	47.307%	0.924%	0.119%	0.016%	0.002%	0.000%	0.122%
EXR-DGPLAS05-BS SM2_crc2, Plasma, Colorectal Carcinoma	2872815	45.518%	52.513%	1.643%	0.191%	0.018%	0.003%	0.000%	0.114%
EXR-DGPLAS06-BS SM3_crc3, Plasma, Colorectal Carcinoma	3690661	46.638%	51.518%	1.498%	0.215%	0.018%	0.002%	0.000%	0.110%
EXR-DGPLAS07-BS SM6_uc1, Plasma, Ulcerative Colitis	3304333	47.004%	51.877%	0.661%	0.345%	0.020%	0.006%	0.000%	0.088%
EXR-DGPLAS08-BS SM7_uc2, Plasma, Ulcerative Colitis	3613655	45.448%	53.622%	0.590%	0.196%	0.018%	0.002%	0.000%	0.125%
EXR-DGPLAS09-BS SM8_uc3, Plasma, Ulcerative Colitis	4719012	44.496%	51.860%	3.297%	0.218%	0.015%	0.003%	0.000%	0.111%

Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

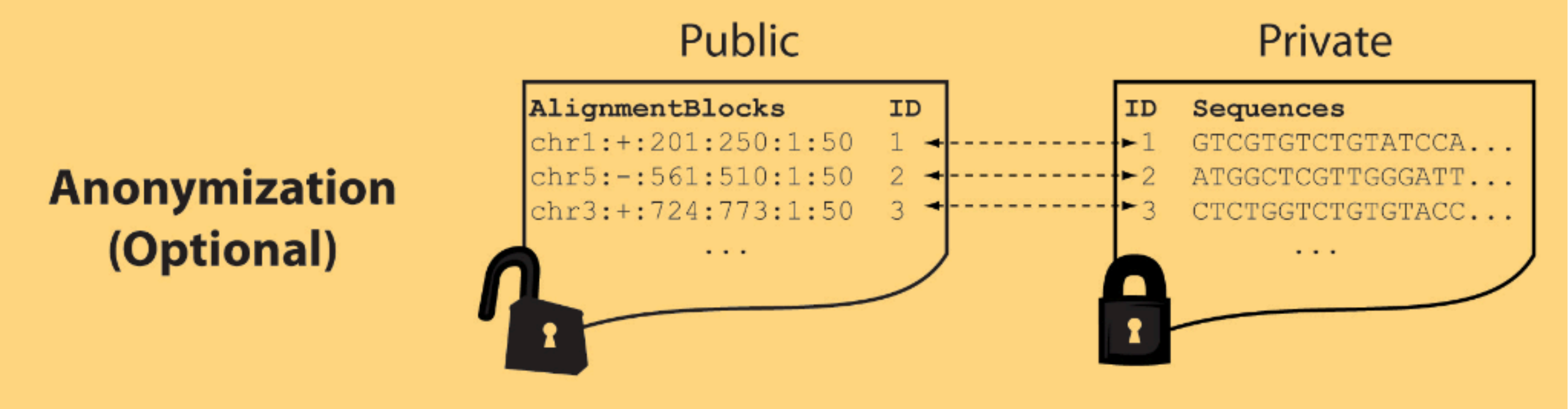
- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

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



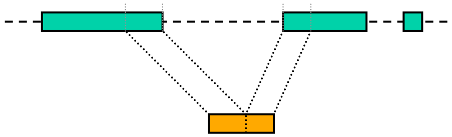
Anonymization (Optional)

Public

AlignmentBlocks	ID
chr1:+:201:250:1:50	1
chr5:-:561:510:1:50	2
chr3:+:724:773:1:50	3
...	

Private

ID	Sequences
1	GTCGTGTCTGTATCCA...
2	ATGGCTCGTTGGGATT...
3	CTCTGGTCTGTGTACC...
	...



Mapping coordinates without variants (MRF)

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

[*Bioinformatics* 27: 281]

MRF Examples

10X Compression Ex.

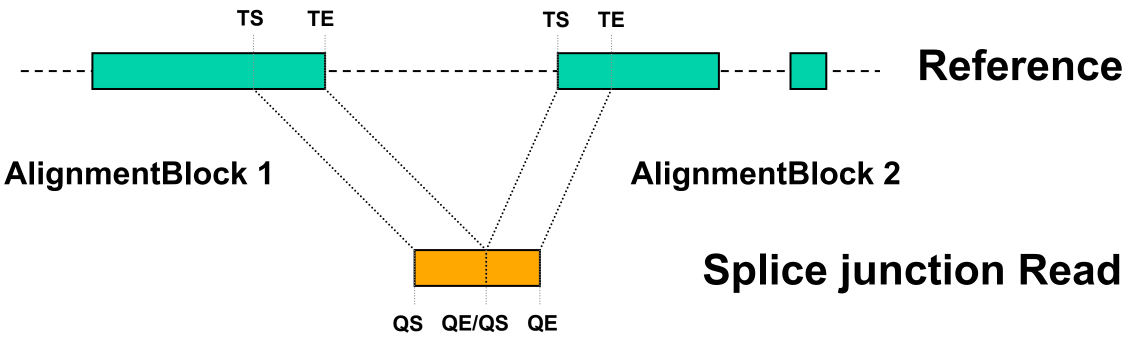
Raw ELAND export file has uncompressed file size: ~4 GB; total number of reads: ~20 million; number of mapped reads: ~12 million .

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

BAM file
has a size of ~1.2 GB.

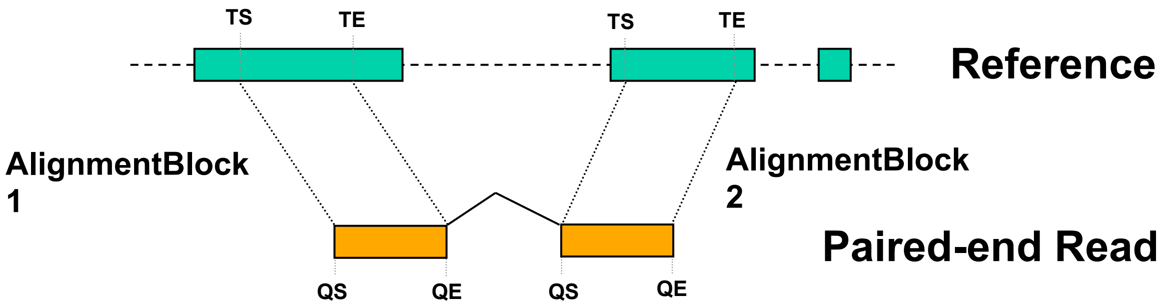
Reference based compression (ie CRAM) is similar but it stores actual variant beyond just position of alignment block

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



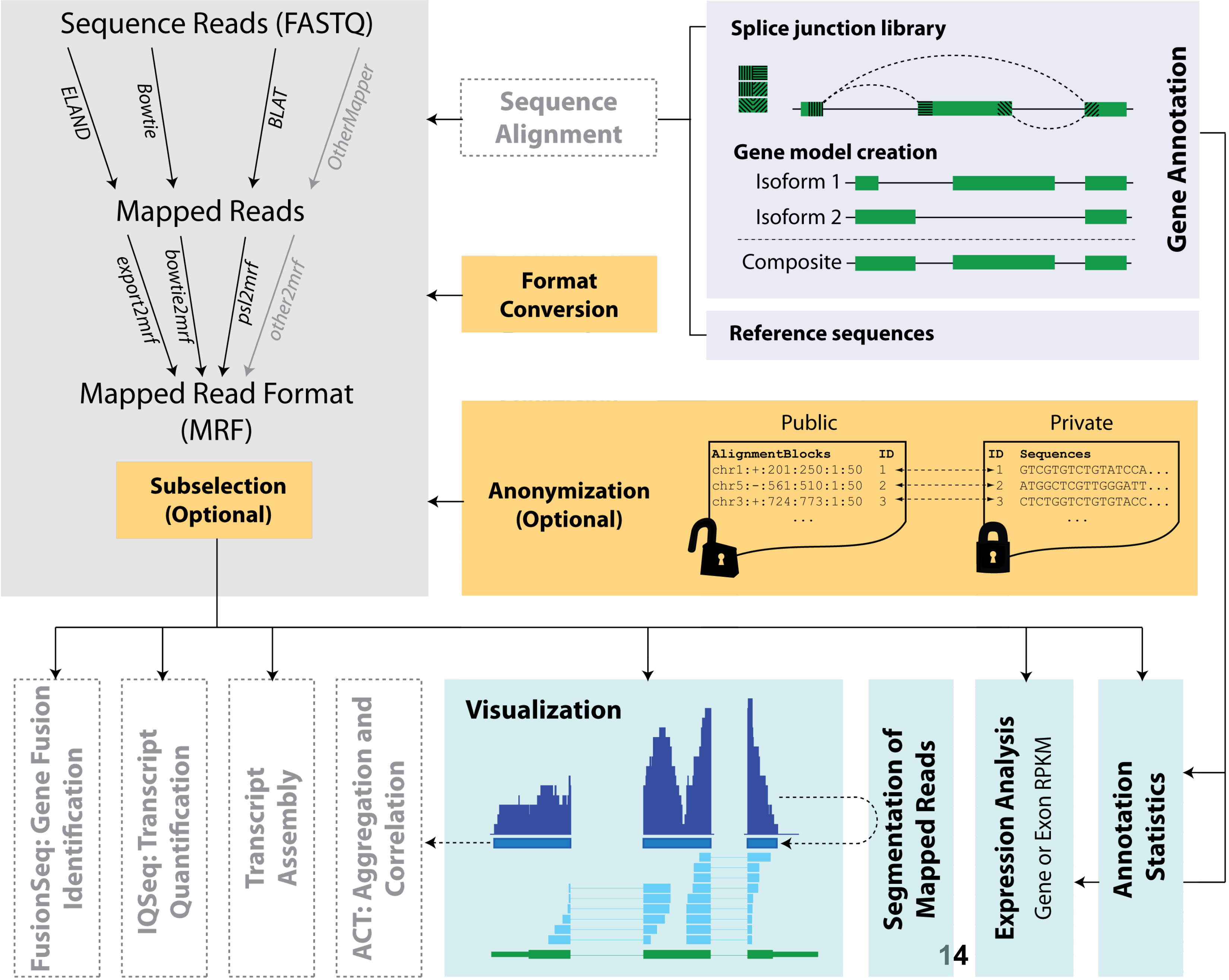
Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

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



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

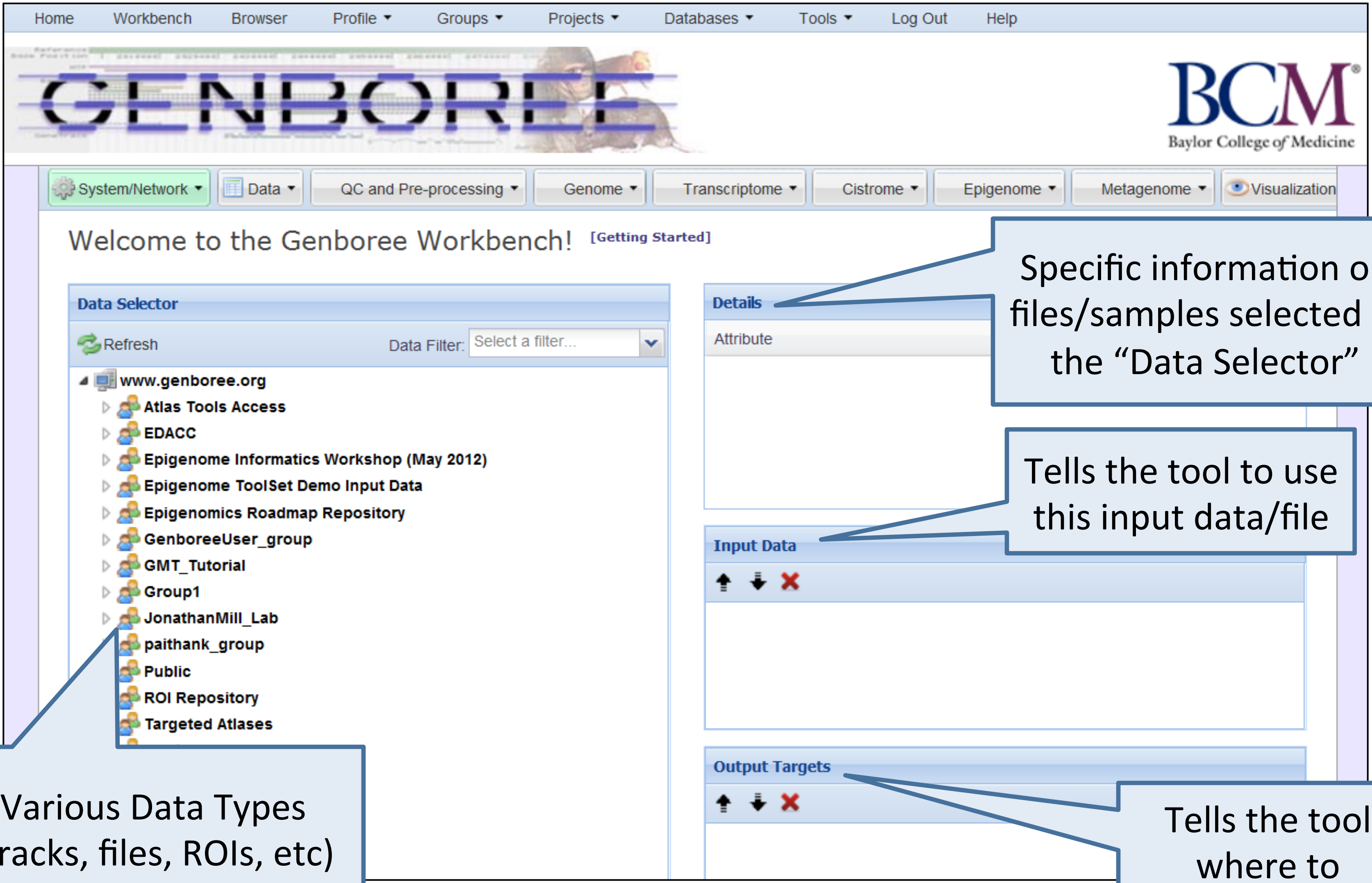
[Habegger et al., Bioinformatics ('11)]



AlignmentBlocks	ID
chr1:+:201:250:1:50	1
chr5:-:561:510:1:50	2
chr3:+:724:773:1:50	3
...	

ID	Sequences
1	GTCGTGTCTGTATCCA...
2	ATGGCTCGTTGGGATT...
3	CTCTGGTCTGTGTACC...
	...

The Genboree Workbench: Web-based Data Management & Analysis



Welcome to the Genboree Workbench! [Getting Started]

- www.genboree.org
  - Atlas Tools Access
  - EDACC
  - Epigenome Informatics Workshop (May 2012)
  - Epigenome ToolSet Demo Input Data
  - Epigenomics Roadmap Repository
  - GenboreeUser_group
  - GMT_Tutorial
  - Group1
  - JonathanMill_Lab
  - paithank_group
  - Public
  - ROI Repository
  - Targeted Atlases

Specific information on files/samples selected in the "Data Selector"

Tells the tool to use this input data/file

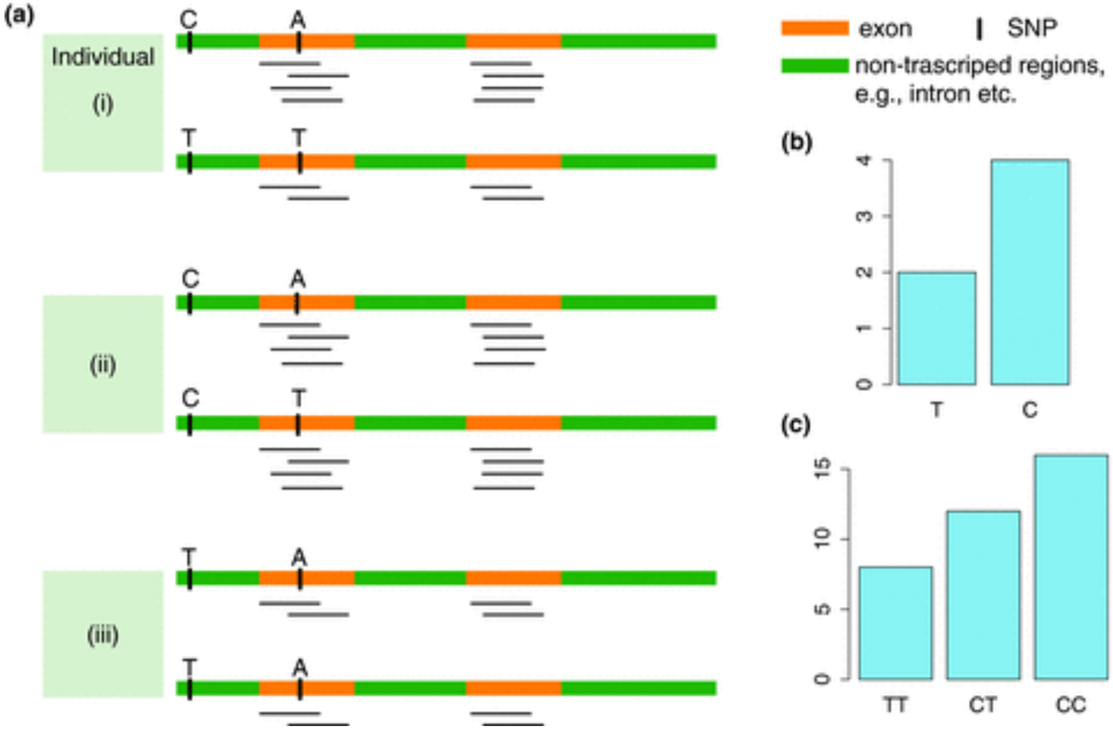
Tells the tool where to deposit results

Various Data Types (tracks, files, ROIs, etc)

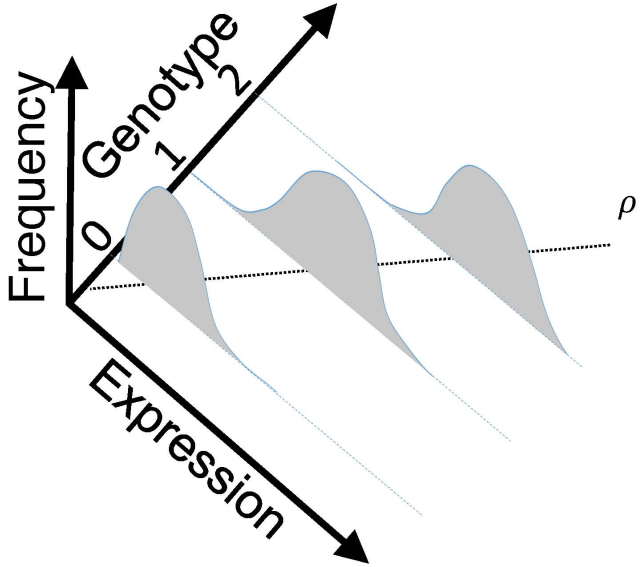
Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

eQTL Mapping Using RNA-Seq Data



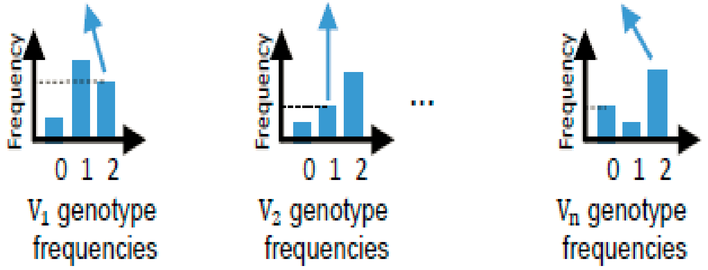
[Biometrics 68(1) 1–11]

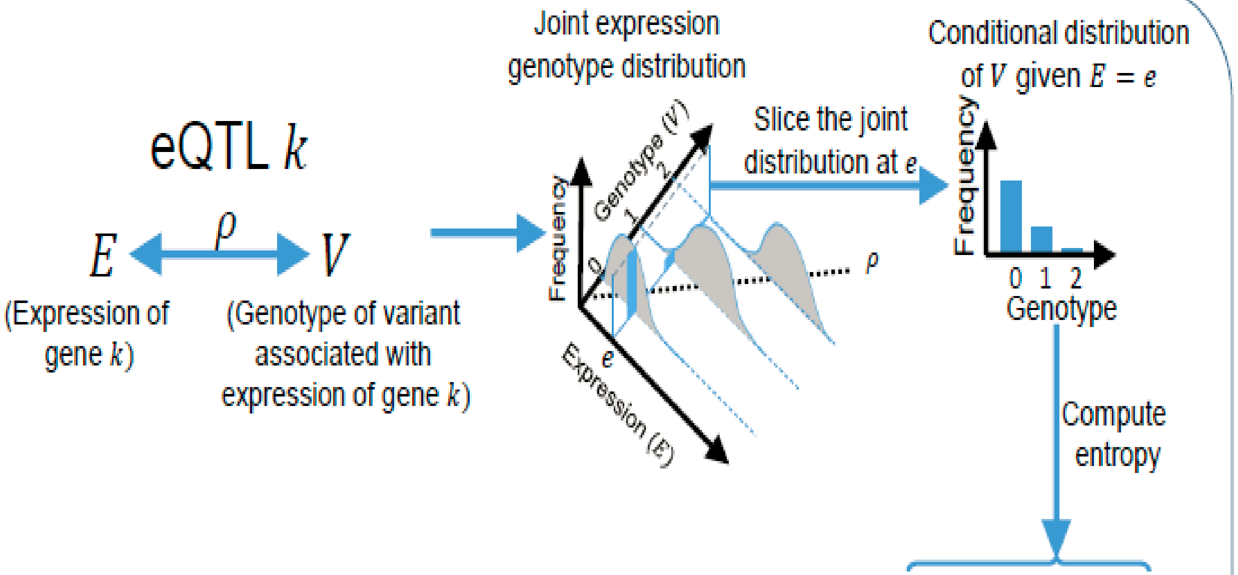


- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Information Content and Predictability

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




eQTL k

E ρ V

(Expression of gene k)

(Genotype of variant associated with expression of gene k)

Joint expression genotype distribution

Slice the joint distribution at e

Conditional distribution of V given $E = e$

Compute entropy

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

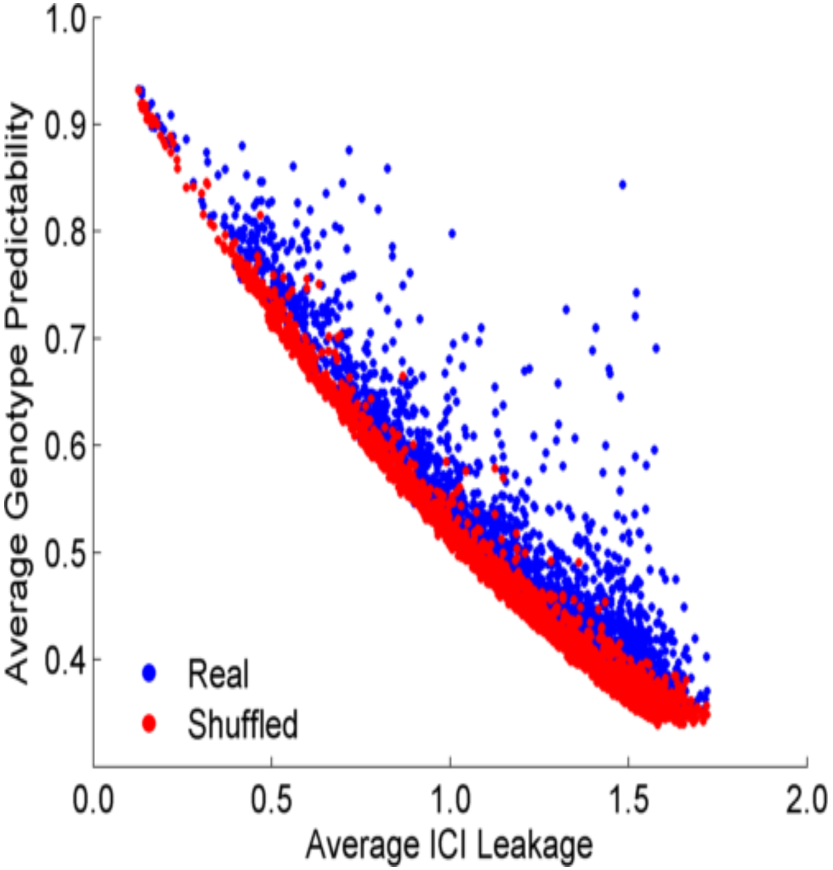
[Harmanci et al. Nat. Meth. ('16)]

Representative Expression, Genotype, eQTL Datasets

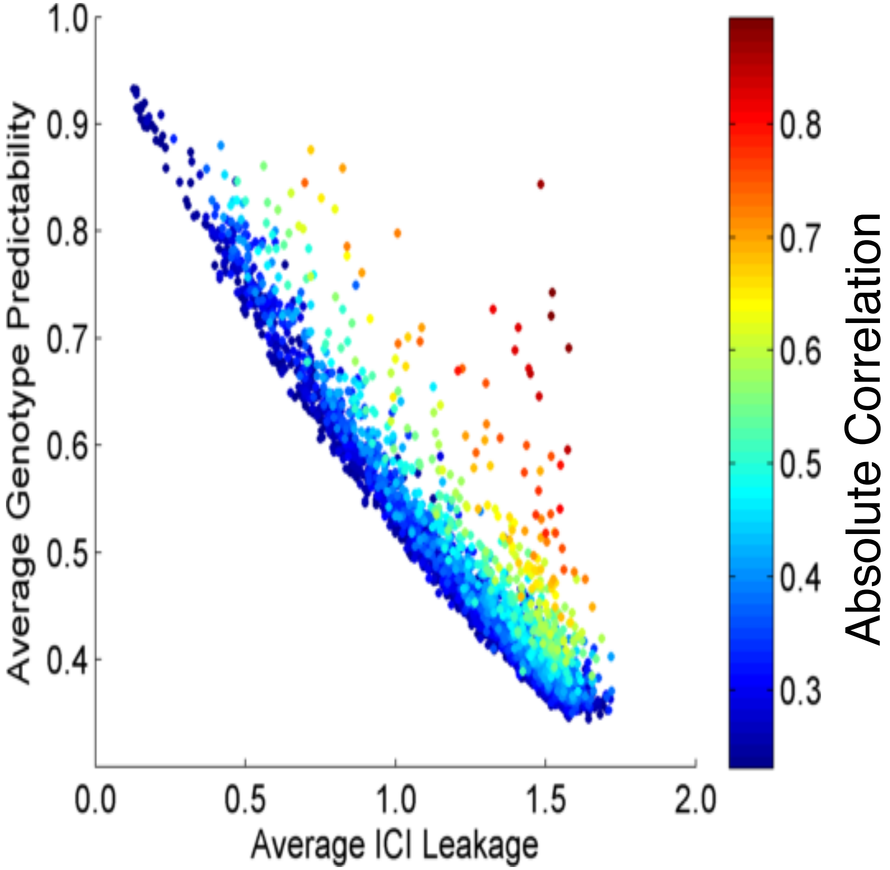
- mRNA sequencing for 462 individuals
  - Publicly availableQuantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)
- Genotypes are available from the 1000 Genomes Project

GEUVADIS
CONSORTIUM

Per eQTL and ICI Cumulative Leakage versus Genotype Predictability

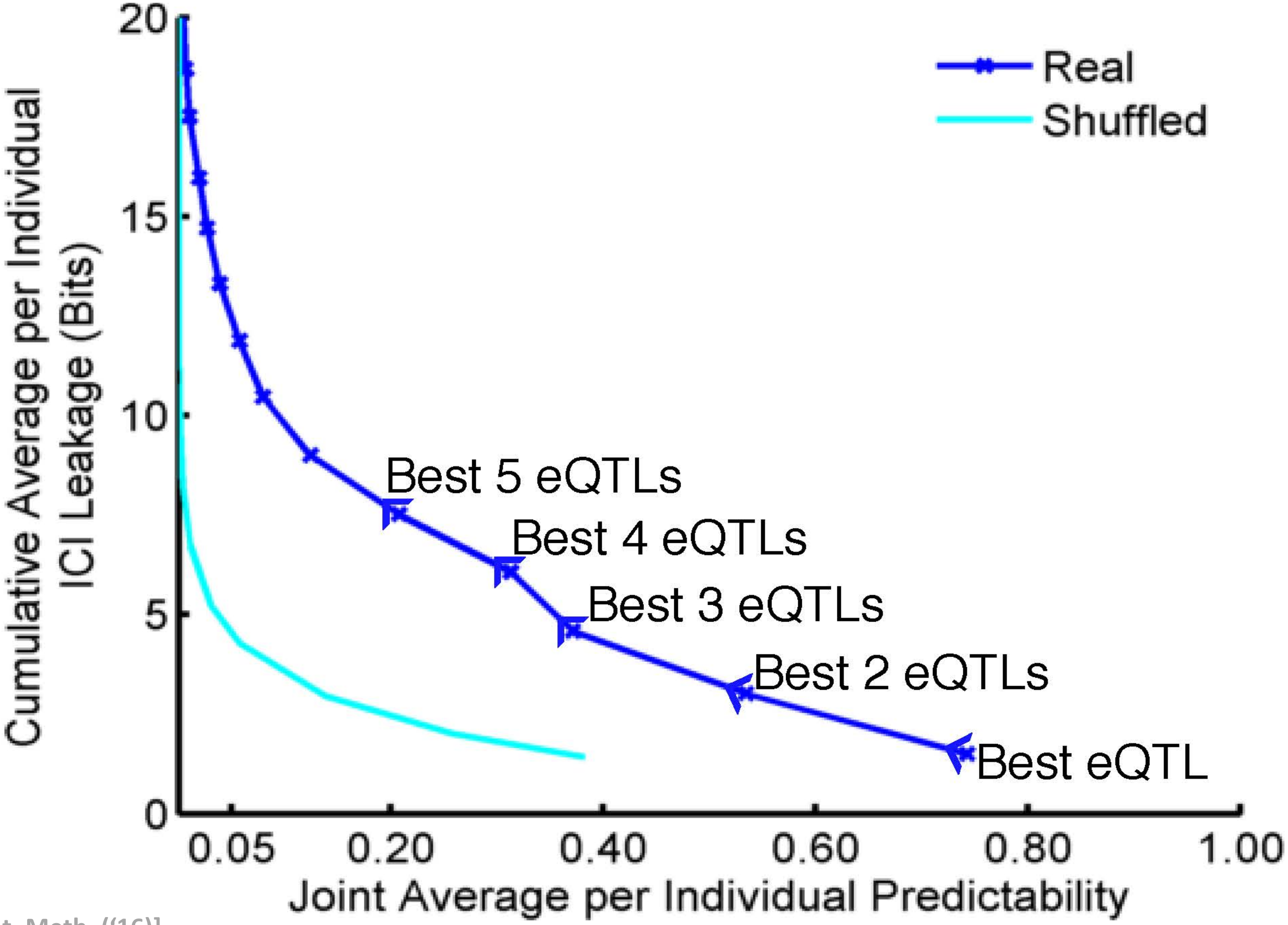


Colors by absolute correlation



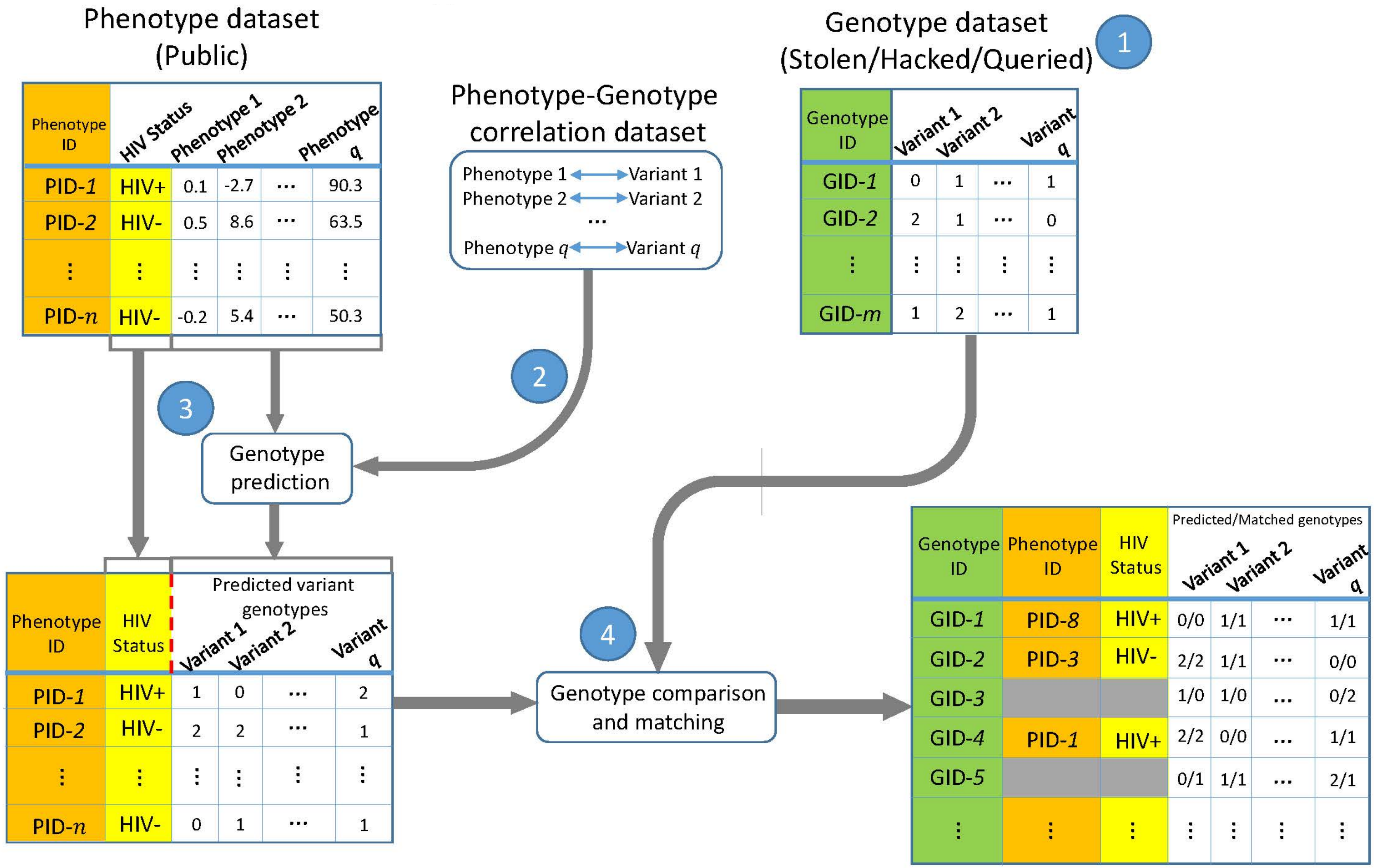
[Harmanci et al. Nat. Meth. ('16)]

Cumulative Leakage versus Joint Predictability



[Harmanci et al. Nat. Meth. ('16)]

Linking Attack Scenario



Phenotype dataset (Public)

Phenotype ID	HIV Status	Phenotype 1	Phenotype 2		Phenotype q
PID-*1*	HIV+	0.1	-2.7	···	90.3
PID-*2*	HIV-	0.5	8.6	···	63.5
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	-0.2	5.4	···	50.3

Phenotype-Genotype correlation dataset

Phenotype 1 ↔ Variant 1
Phenotype 2 ↔ Variant 2
···
Phenotype *q* ↔ Variant *q*

Genotype dataset (Stolen/Hacked/Queried)

Genotype ID	Variant 1	Variant 2		Variant q
GID-*1*	0	1	···	1
GID-*2*	2	1	···	0
⋮	⋮	⋮	⋮	⋮
GID-*m*	1	2	···	1

Genotype prediction

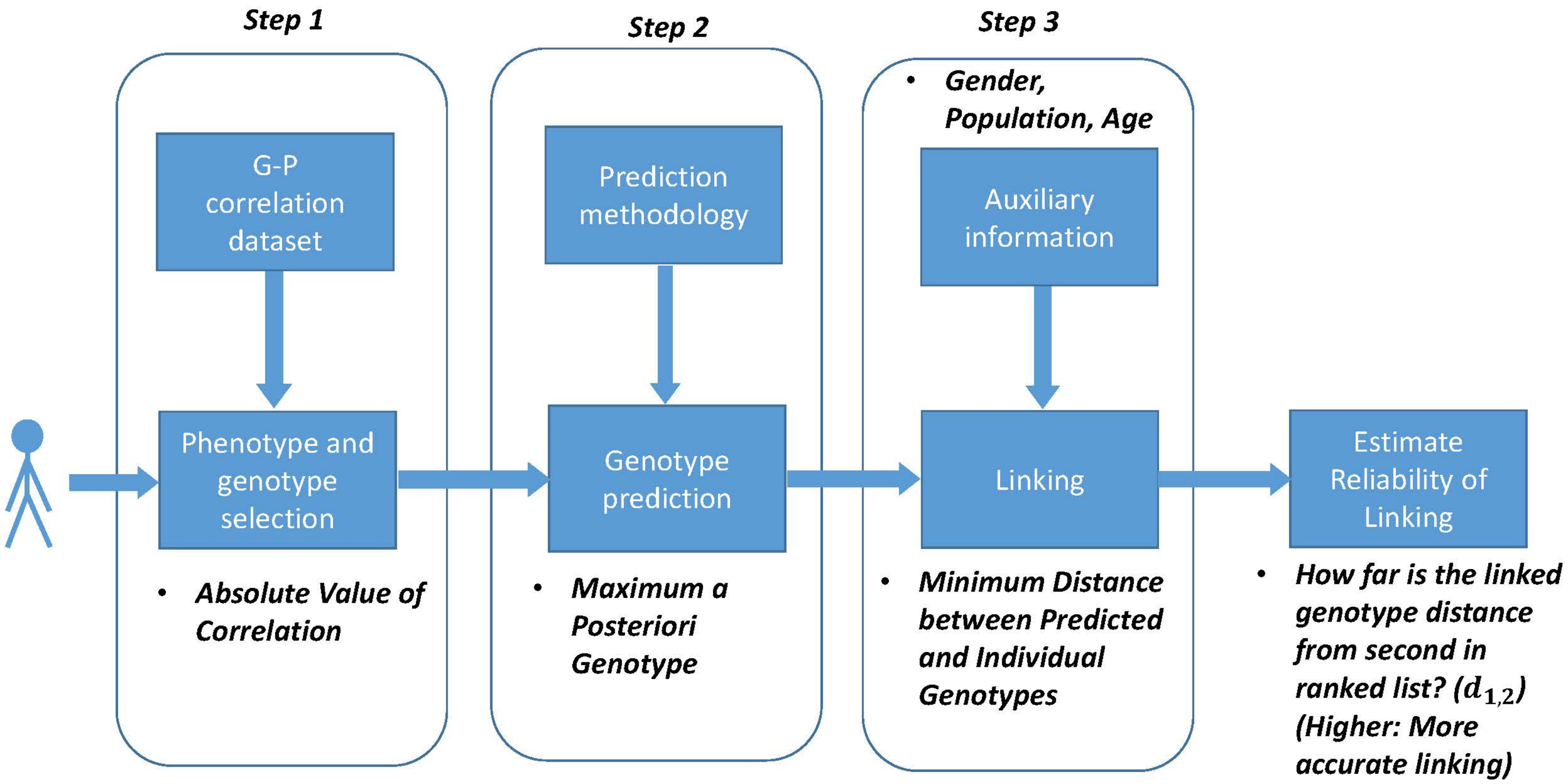
Phenotype ID	HIV Status	Predicted variant genotypes			
		Variant 1	Variant 2		Variant q
PID-*1*	HIV+	1	0	···	2
PID-*2*	HIV-	2	2	···	1
⋮	⋮	⋮	⋮	⋮	⋮
PID-*n*	HIV-	0	1	···	1

Genotype comparison and matching

Genotype ID	Phenotype ID	HIV Status	Predicted/Matched genotypes			
			Variant 1	Variant 2		Variant q
GID-*1*	PID-*8*	HIV+	0/0	1/1	···	1/1
GID-*2*	PID-*3*	HIV-	2/2	1/1	···	0/0
GID-*3*			1/0	1/0	···	0/2
GID-*4*	PID-*1*	HIV+	2/2	0/0	···	1/1
GID-*5*			0/1	1/1	···	2/1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[Harmanci et al. Nat. Meth. ('16)]

Steps in Instantiation of a (Mock) Linking Attack



Step 1

G-P correlation dataset

Phenotype and genotype selection

- ***Absolute Value of Correlation***

Step 2

Prediction methodology

Genotype prediction

- ***Maximum a Posteriori Genotype***

Step 3

- ***Gender, Population, Age***

Auxiliary information

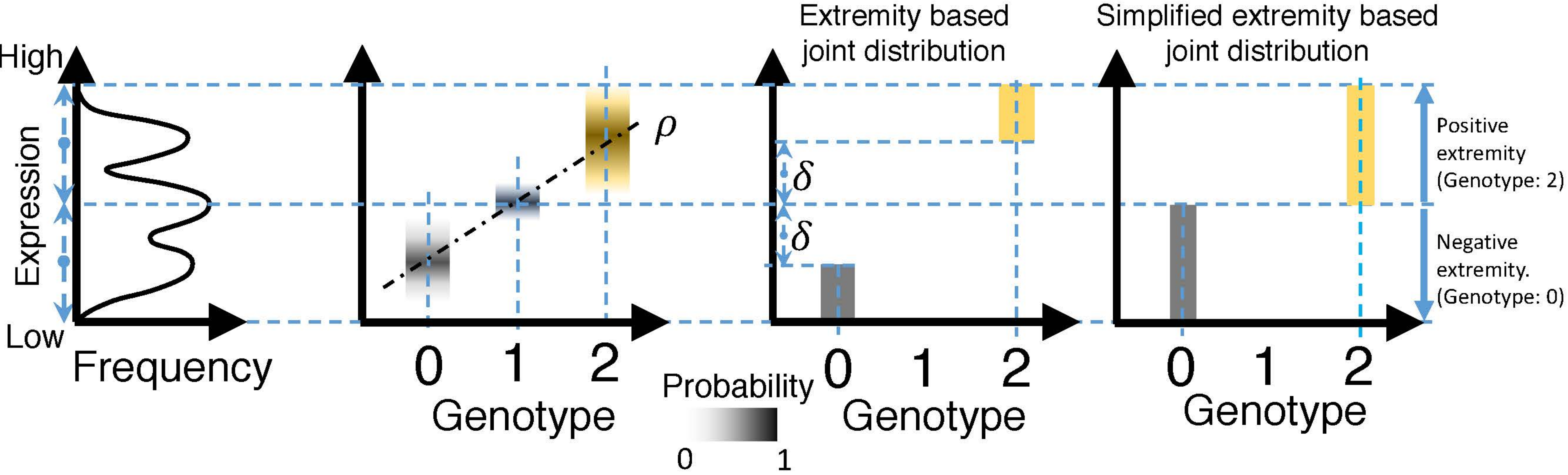
Linking

- ***Minimum Distance between Predicted and Individual Genotypes***

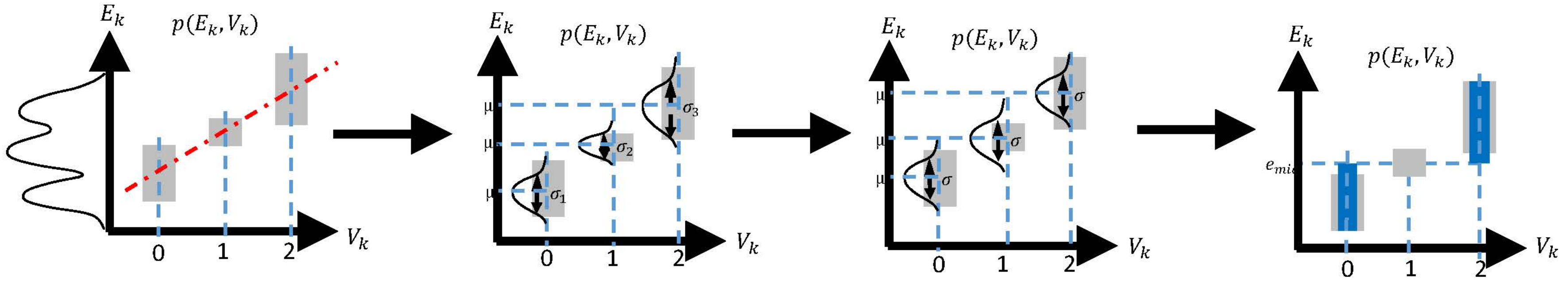
Estimate Reliability of Linking

- ***How far is the linked genotype distance from second in ranked list? ($d_{1,2}$) (Higher: More accurate linking)***

[Harmanci et al. Nat. Meth. ('16)]

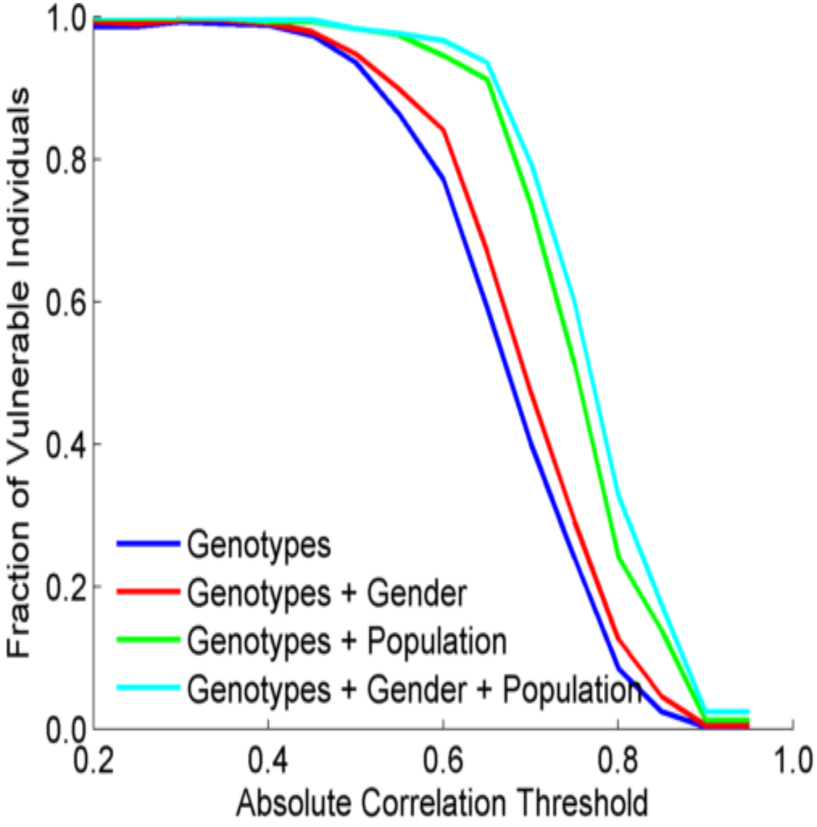

High
Expression
Low
Frequency
ρ
0
1
2
Genotype
Probability
0
1
Extremity based joint distribution
δ
δ
0
1
2
Genotype
Simplified extremity based joint distribution
Positive extremity (Genotype: 2)
Negative extremity. (Genotype: 0)
0
1
2
Genotype

Levels of Expression-Genotype Model Simplifications:

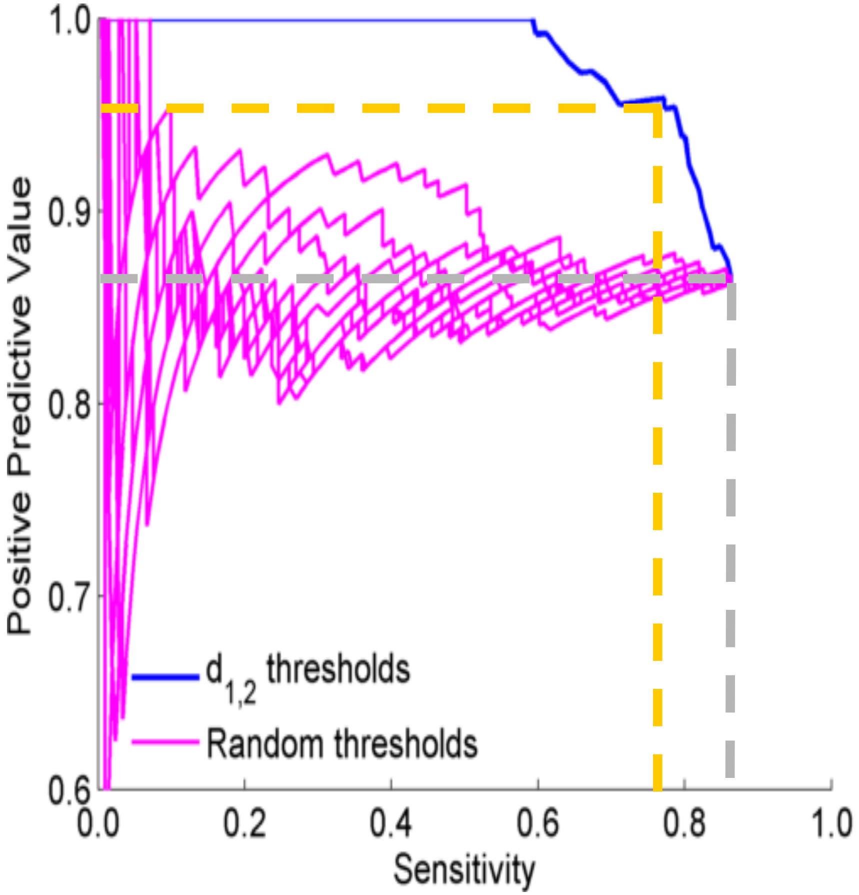


[Harmanci et al. Nat. Meth. ('16)]

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)]

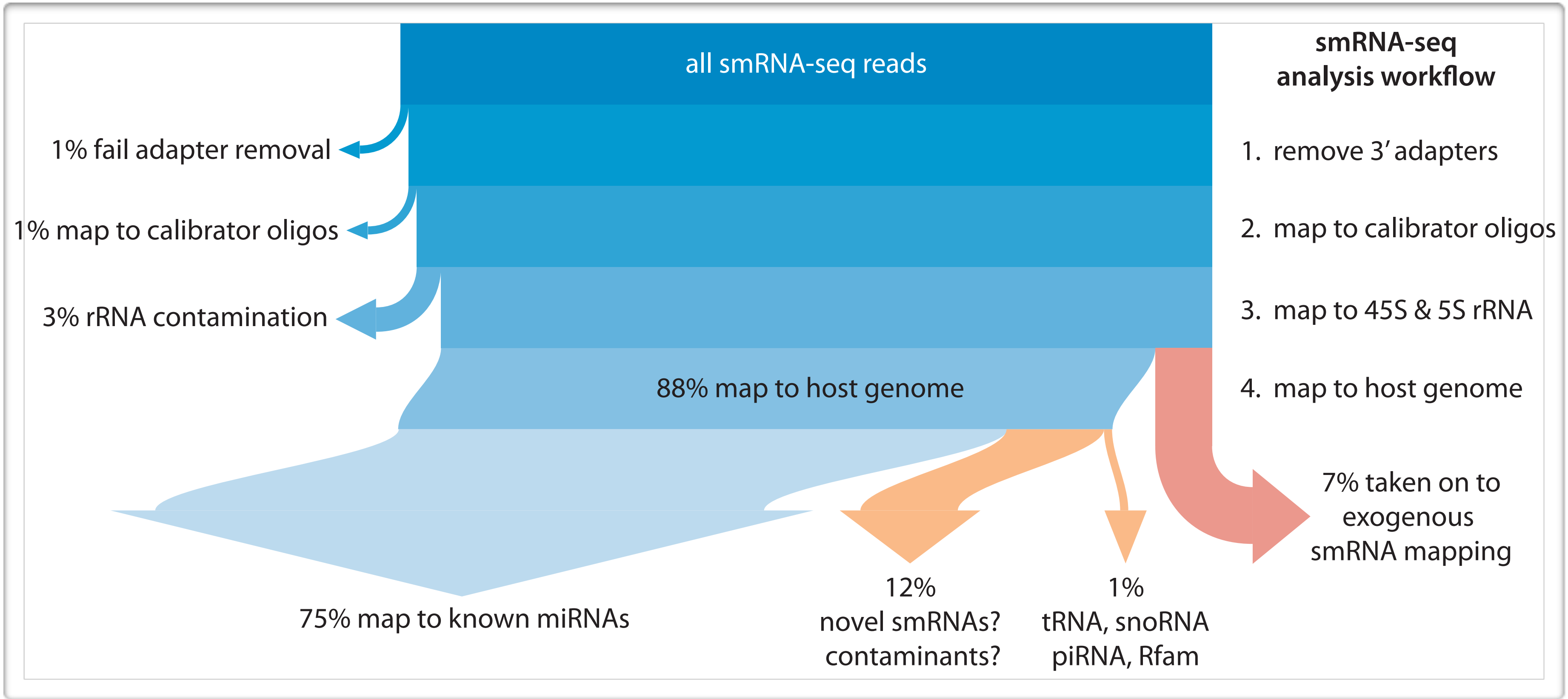
Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

for a typical cellular sample…


all smRNA-seq reads
smRNA-seq analysis workflow
1% fail adapter removal
1. remove 3′ adapters
1% map to calibrator oligos
2. map to calibrator oligos
3% rRNA contamination
3. map to 45S & 5S rRNA
88% map to host genome
4. map to host genome
7% taken on to exogenous smRNA mapping
75% map to known miRNAs
12% novel smRNAs? contaminants?
1% tRNA, snoRNA piRNA, Rfam

- exRNA samples typically much noisier
- cascade of read-alignment steps mitigates contamination

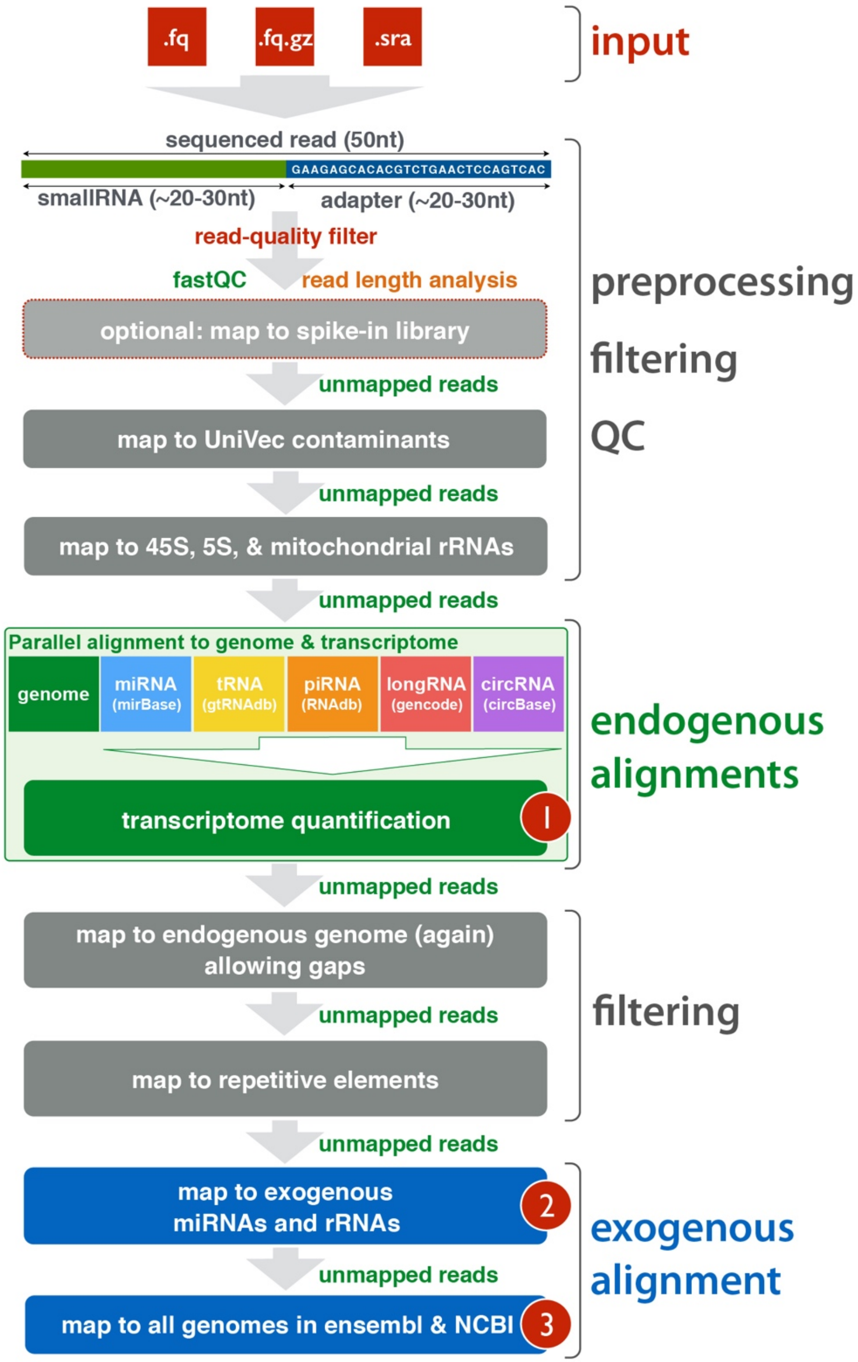
extra-**ce**llular **R**NA **p**rocessing **t**oolkit

- automatic pre-processing and QC of sequence reads
- explicit filtering of contaminants & rRNA
- quantification of spike-in sequences and many different smallRNA biotypes
- support for random barcodes (Bioo)
- choice of 3 end-points:

  1 endogenous only

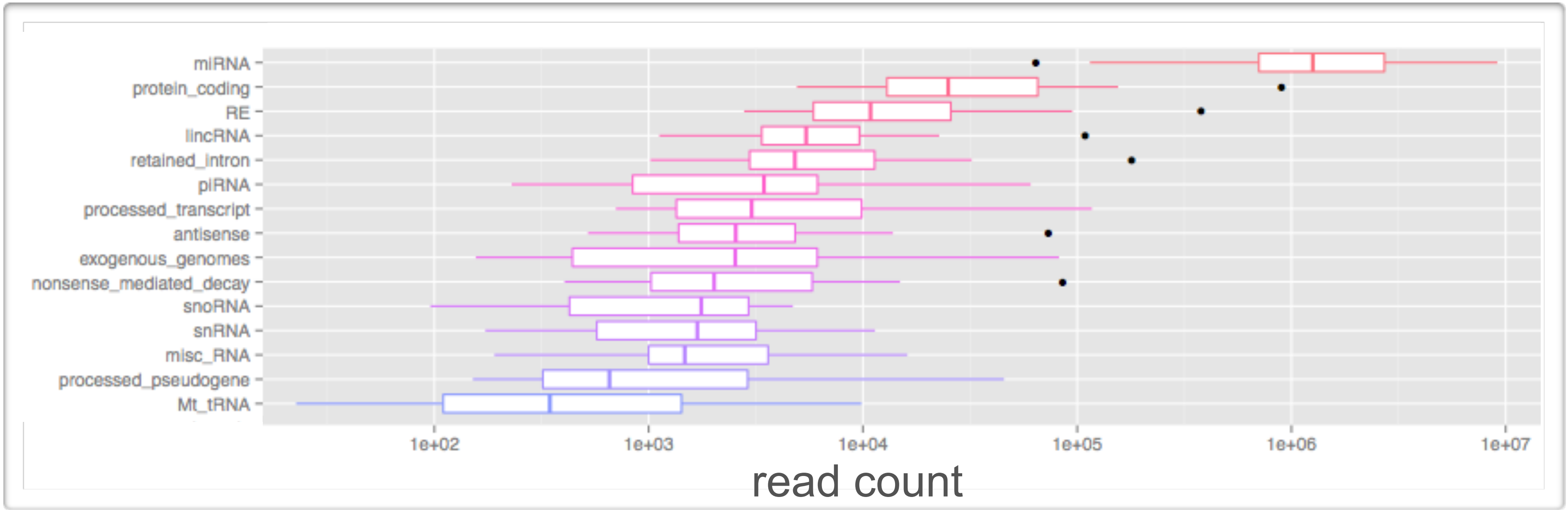
  2 exogenous miRNA +rRNA

  3 exogenous genomes

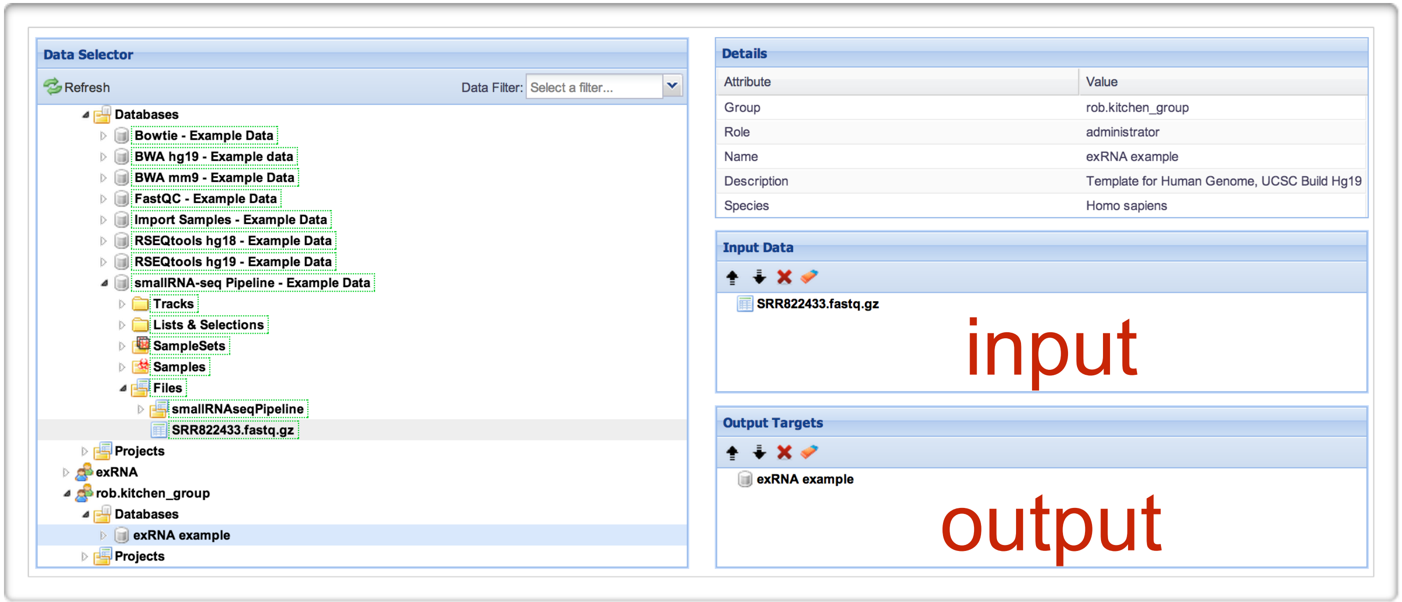


total reads by biotype

- large contribution from miRNA and mRNA
- also some signal from exogenous sequences



exceRpt @ Genboree.org



Data Selector

Refresh — Data Filter: Select a filter...

- Databases
  - Bowtie - Example Data
  - BWA hg19 - Example data
  - BWA mm9 - Example Data
  - FastQC - Example Data
  - Import Samples - Example Data
  - RSEQtools hg18 - Example Data
  - RSEQtools hg19 - Example Data
  - smallRNA-seq Pipeline - Example Data
    - Tracks
    - Lists & Selections
    - SampleSets
    - Samples
    - Files
      - smallRNAseqPipeline
      - SRR822433.fastq.gz
- Projects
- exRNA
- rob.kitchen_group
  - Databases
    - exRNA example
  - Projects

Details

Attribute	Value
Group	rob.kitchen_group
Role	administrator
Name	exRNA example
Description	Template for Human Genome, UCSC Build Hg19
Species	Homo sapiens

Input Data

SRR822433.fastq.gz

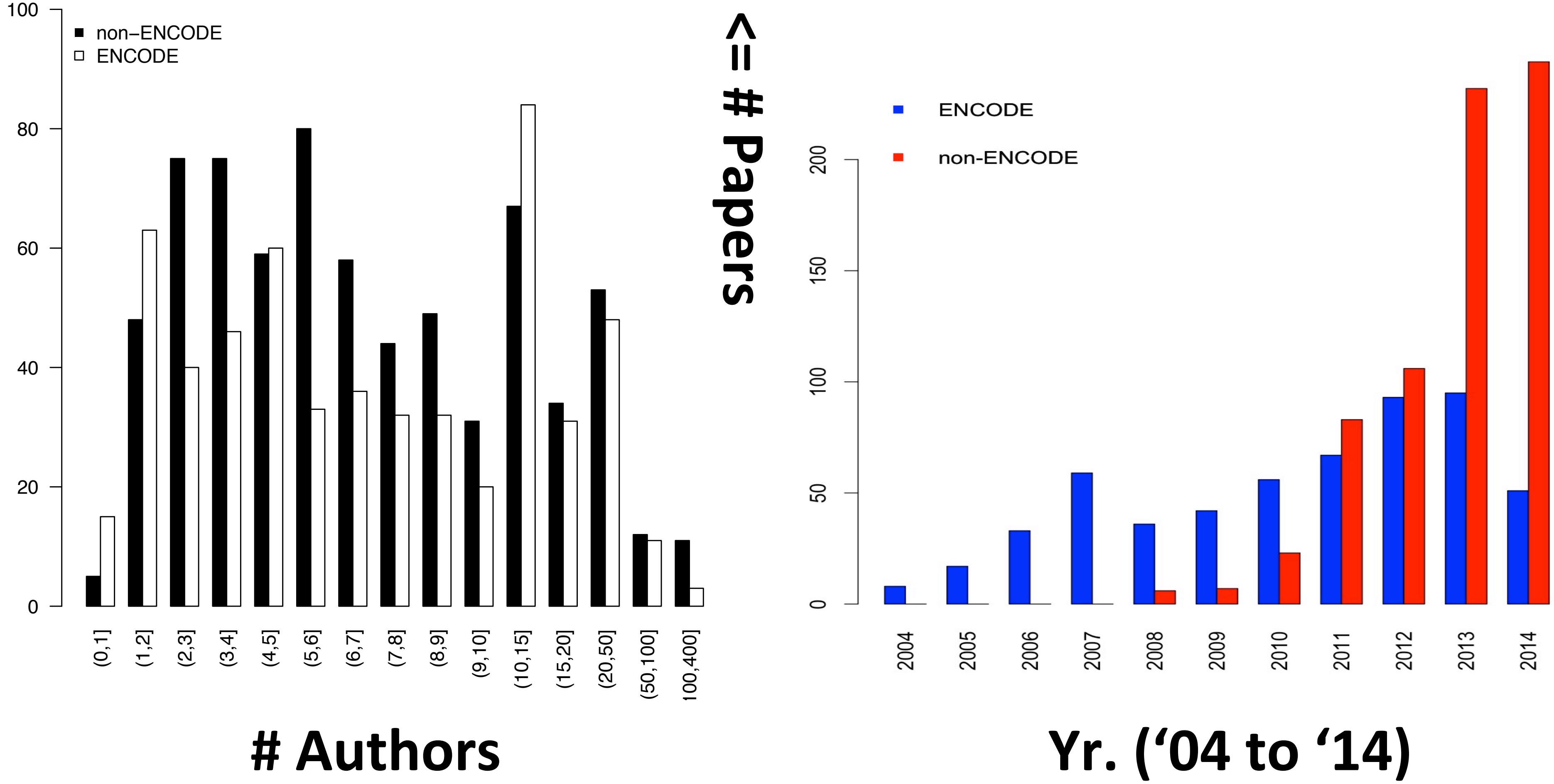
Output Targets

exRNA example

- extremely simple to use (1 input, 1 output)
- can process multiple samples in parallel
- very customisable (choice of smRNA libs, calibrators, etc)

Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

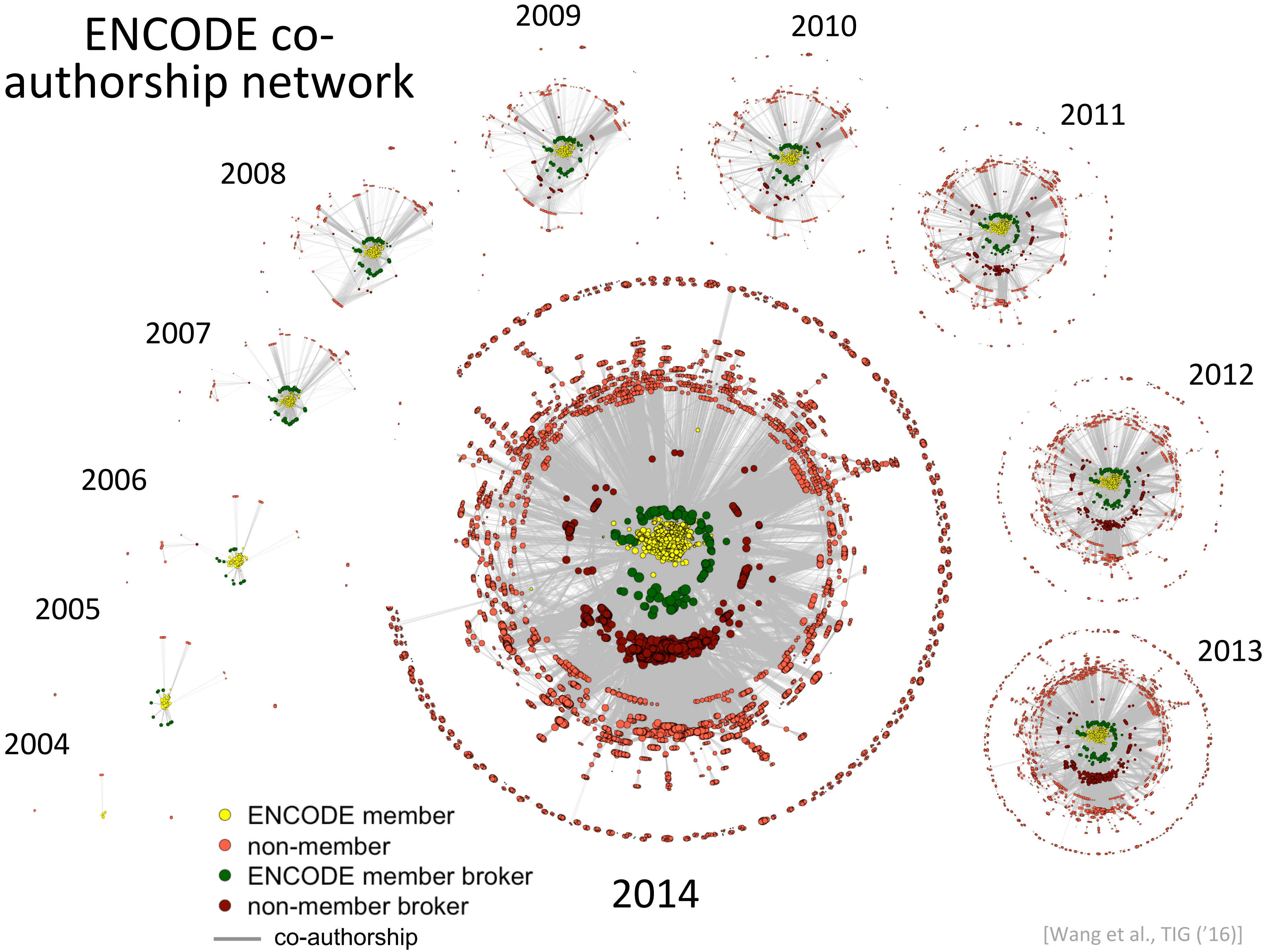
- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals



Papers authored by ENCODE consortium members vs. those that use ENCODE data but were not funded by ENCODE

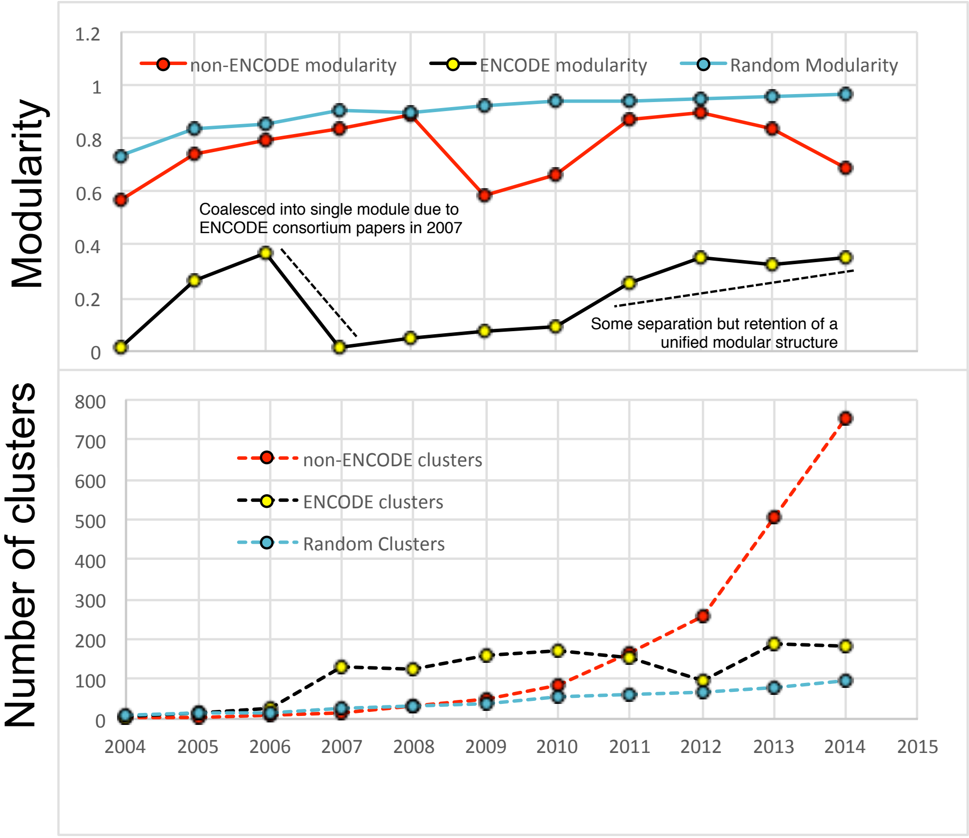
[Wang et al., TIG ('16)]

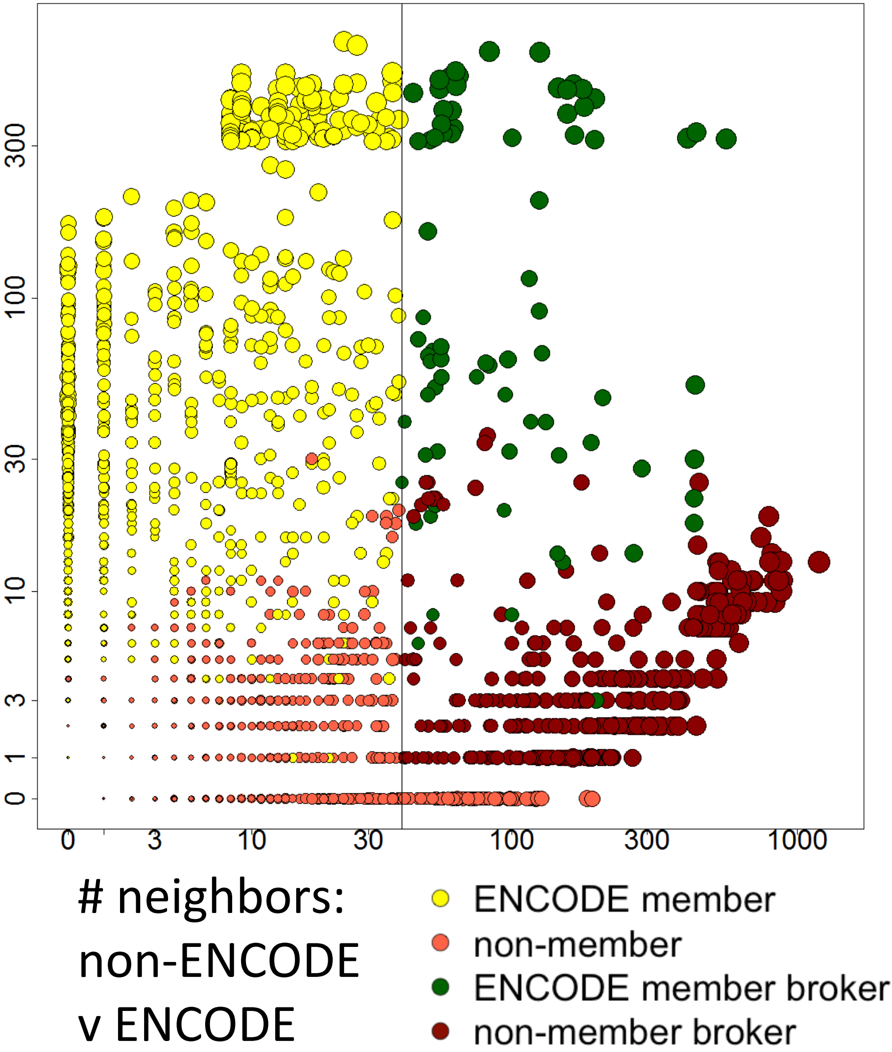
ENCODE co-authorship network



[Wang et al., TIG ('16)]

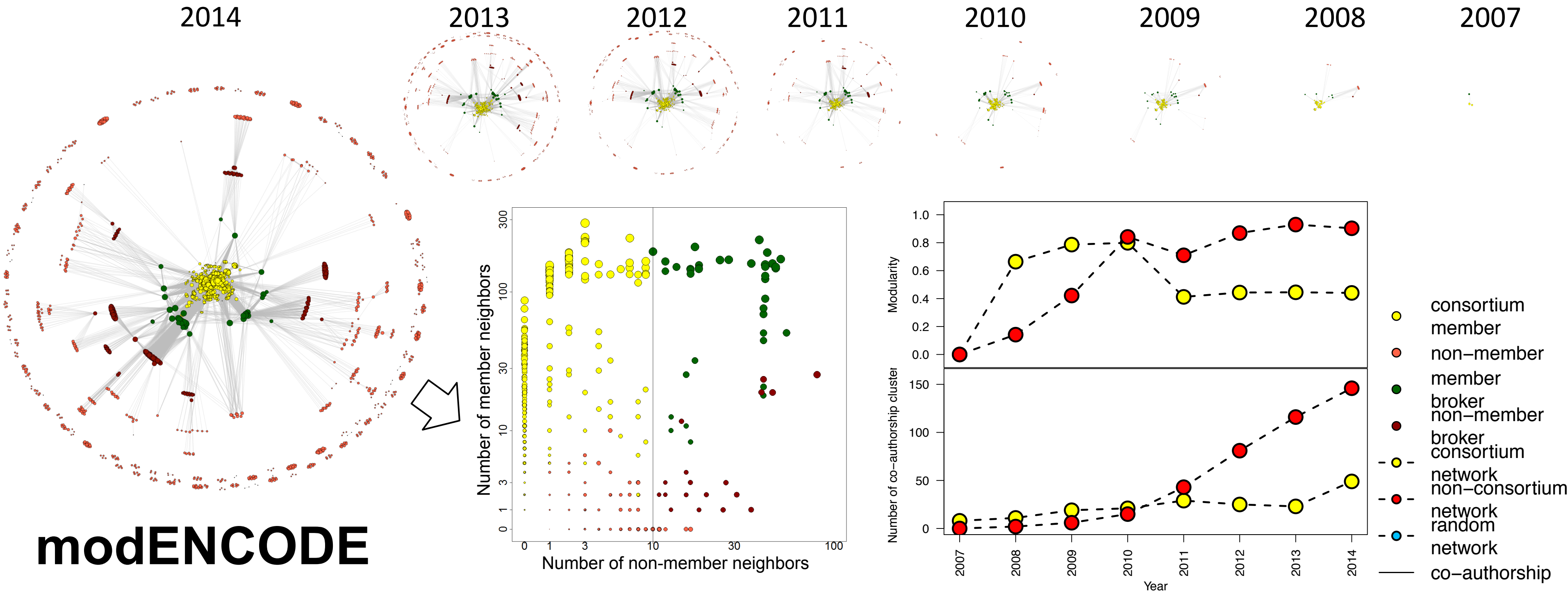
Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)





[Wang et al., TIG ('16)]

modENCODE


2014
2013
2012
2011
2010
2009
2008
2007
Number of member neighbors
300
100
30
10
3
1
0
Number of non-member neighbors
0
1
3
10
30
100
Modularity
1.0
0.8
0.6
0.4
0.2
0.0
Number of co-authorship cluster
150
100
50
0
Year
2007
2008
2009
2010
2011
2012
2013
2014
consortium member
non–member
member broker
non–member broker
consortium network
non–consortium network
random network
co–authorship

Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

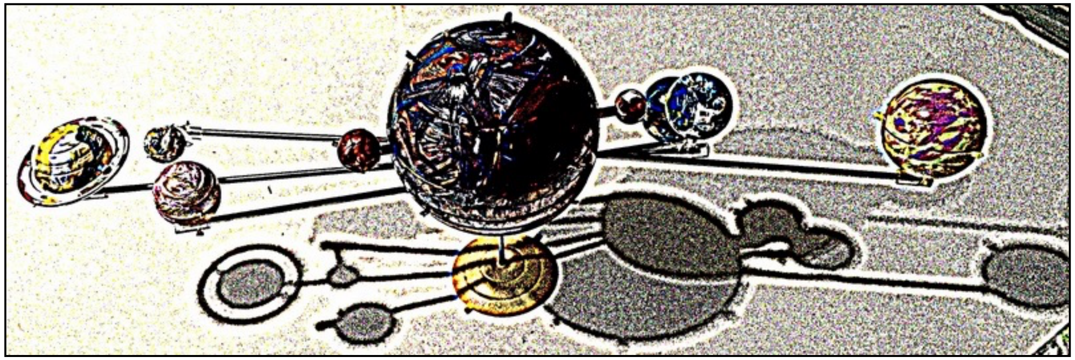
- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals

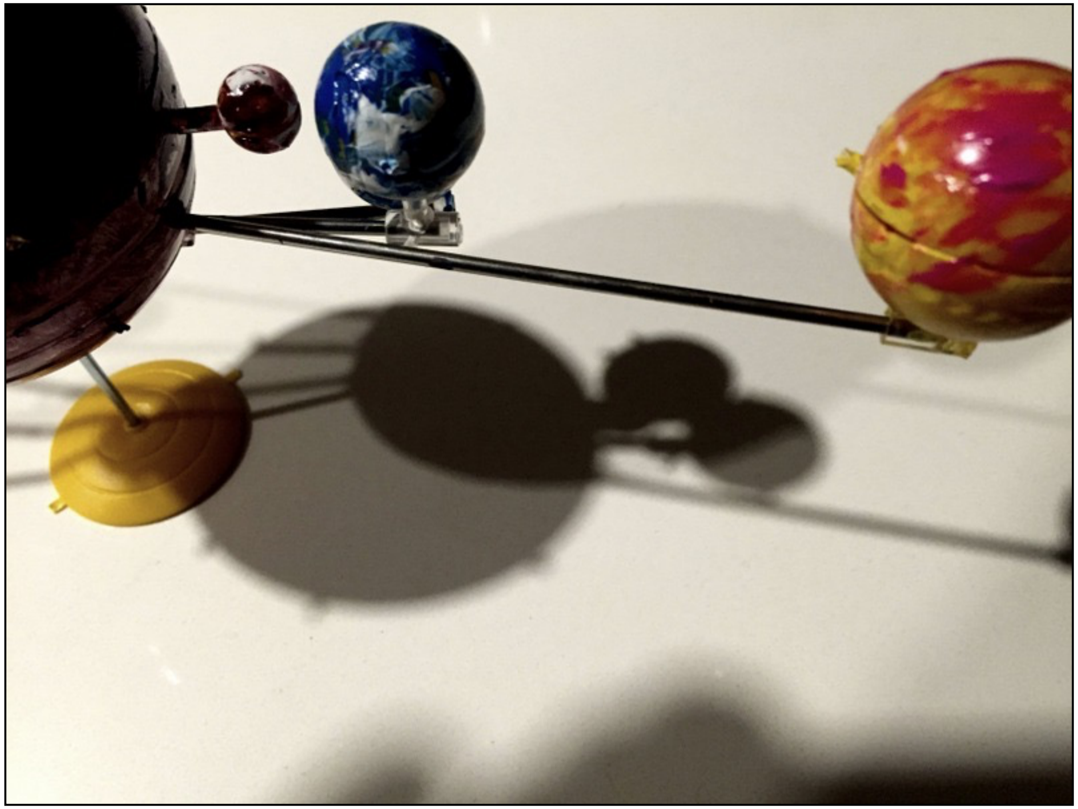
PrivaSeq.gersteinlab.org
A Harmanci

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

"Encode authors"
D Wang, KK Yan,
J Rozowsky, E Pan

exRNA.org & **exceRpt**
R Kitchen
J Rozowsky
A Milosavljevic
M Roth
S Subramanian





Acknowledgments

Hiring Postdocs. See gersteinlab.org/jobs

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - 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