EN-TEx / Personal Genomes subgroup

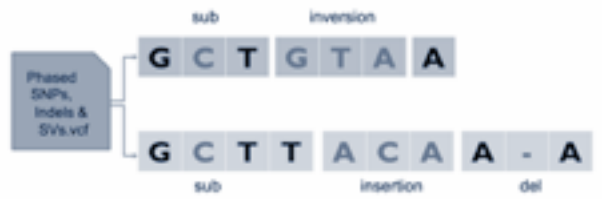
M Gerstein & E Aiden
'19 ENCODE meeting

Goals of EN-TEx

Collaboration between **EN**CODE & G**TEx**

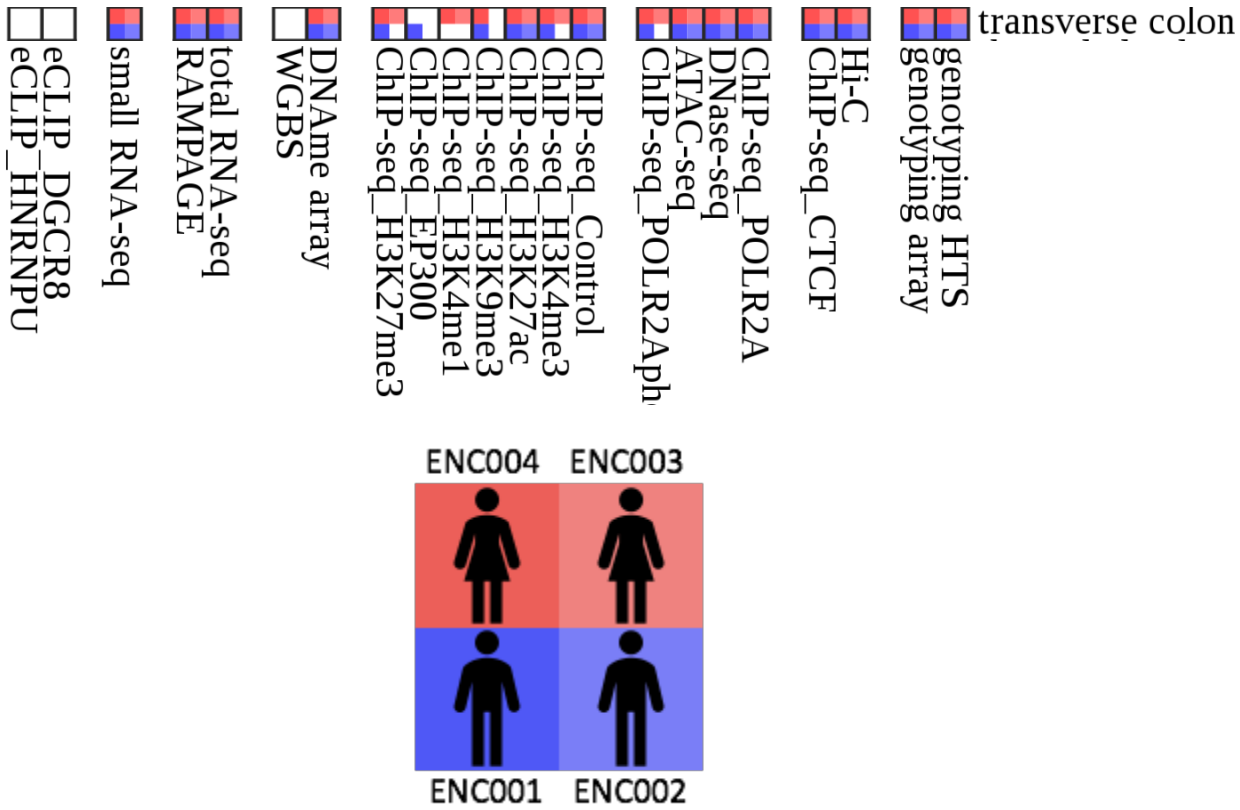
Cataloguing genomic elements in personal genomes with matched ENCODE datasets & the full breadth of assays on many tissues

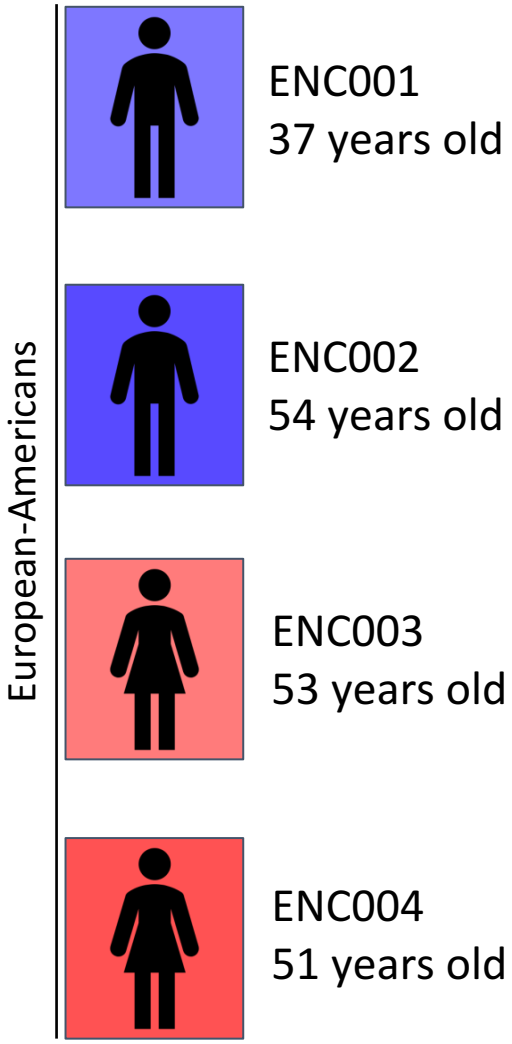
1. Collect matched WGS & functional genomic data across ~20 tissues for 4 individuals.
2. Catalogue genomic elements across individuals & tissues and study their variation
3. Study the utility of phased personal diploid genomes for analyzing functional genomics data



1. The Matched Data Matrix & Personal Genome

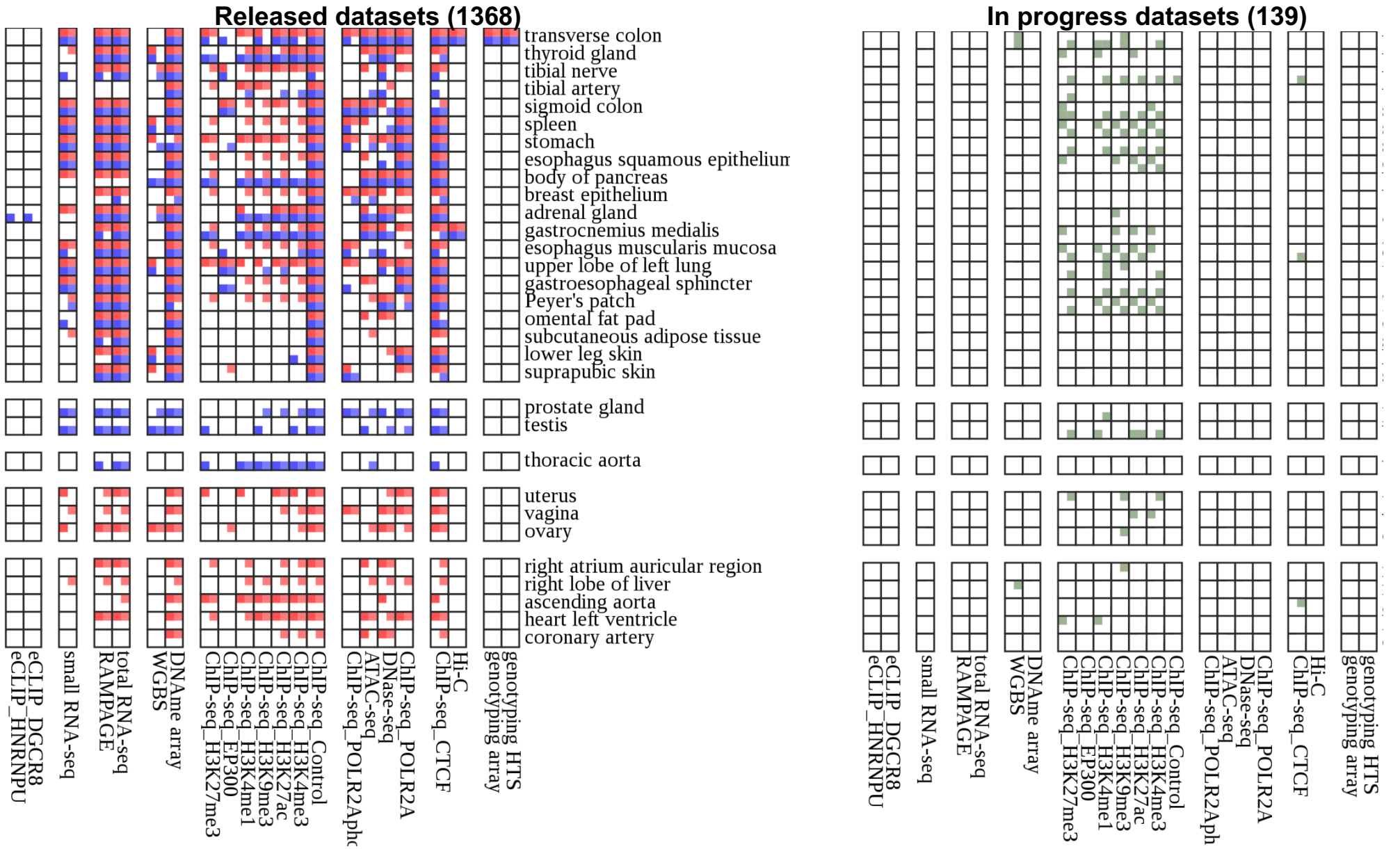
EN-TEx
Assays, Individuals & Tissue Types





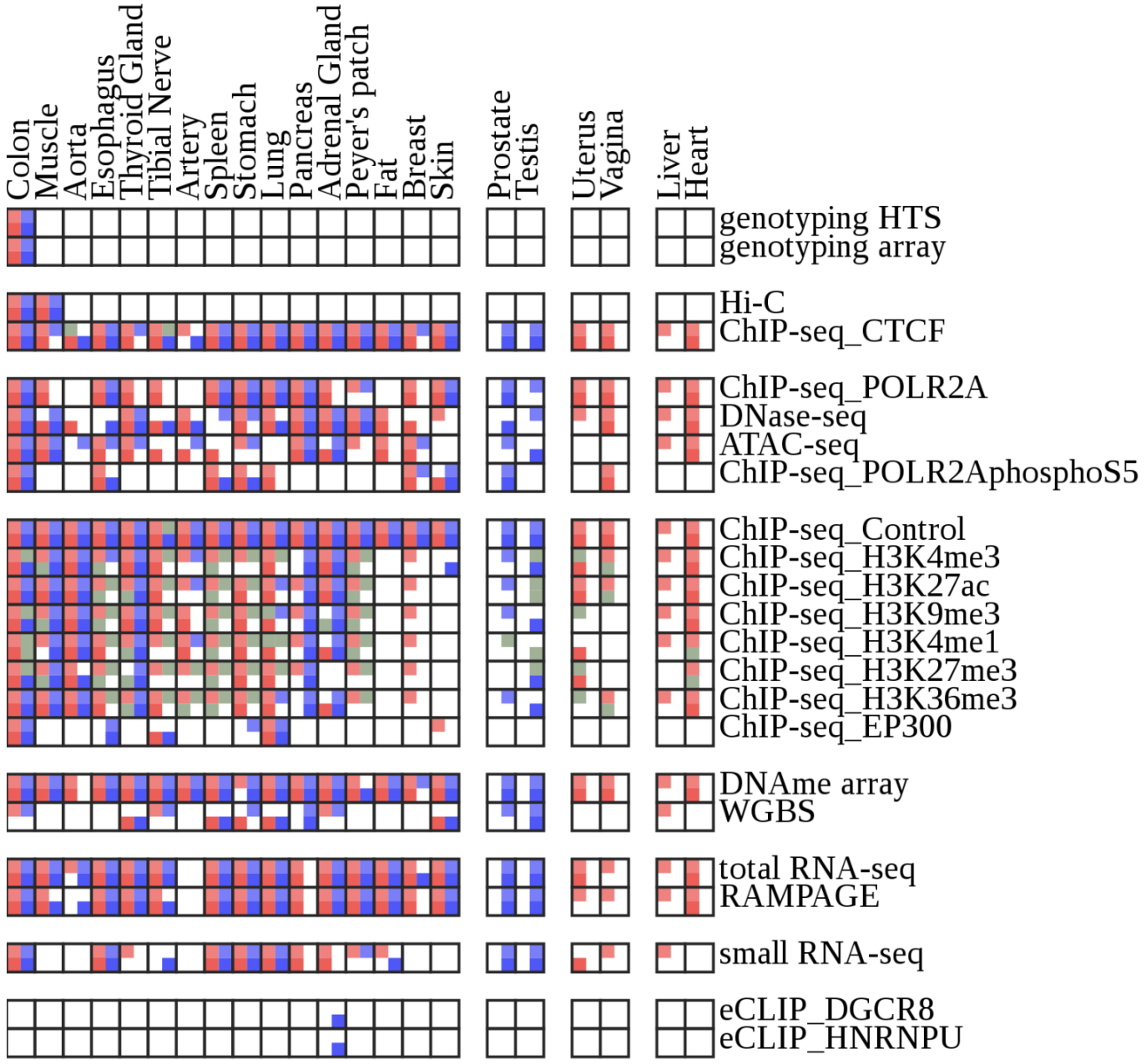
- Gastrocnemius medialis
- Stomach
- Transverse colon
- Peyer's patch
- Sigmoid colon
- Right lobe of liver
- Body of pancreas
- Esophagus squamous epithelium
- Gastroesophageal sphincter
- Esophagus muscularis mucosa
- Testis
- Prostate gland
- Uterus
- Vagina
- Ovary
- Thyroid gland
- Adrenal gland
- Spleen
- Right atrium auricular region
- Heart left ventricle
- Thoracic aorta
- Ascending aorta
- Omental fat pad
- Subcutaneous adipose tissue
- Tibial nerve
- Suprapubic skin
- Lower leg skin
- Breast epithelium
- Upper lobe of left lung

EN-TEx Full Data Matrix as of Dec-2018

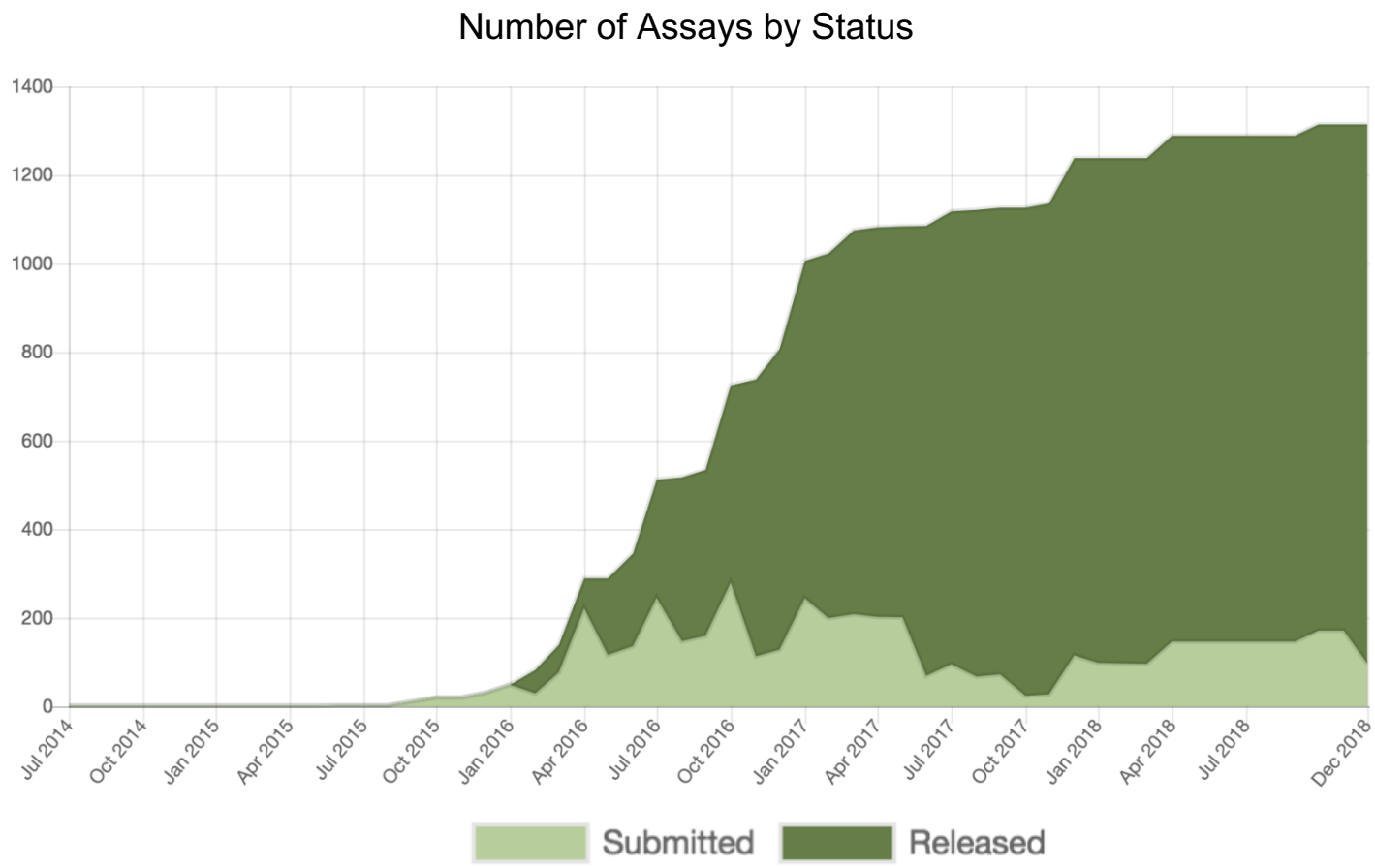


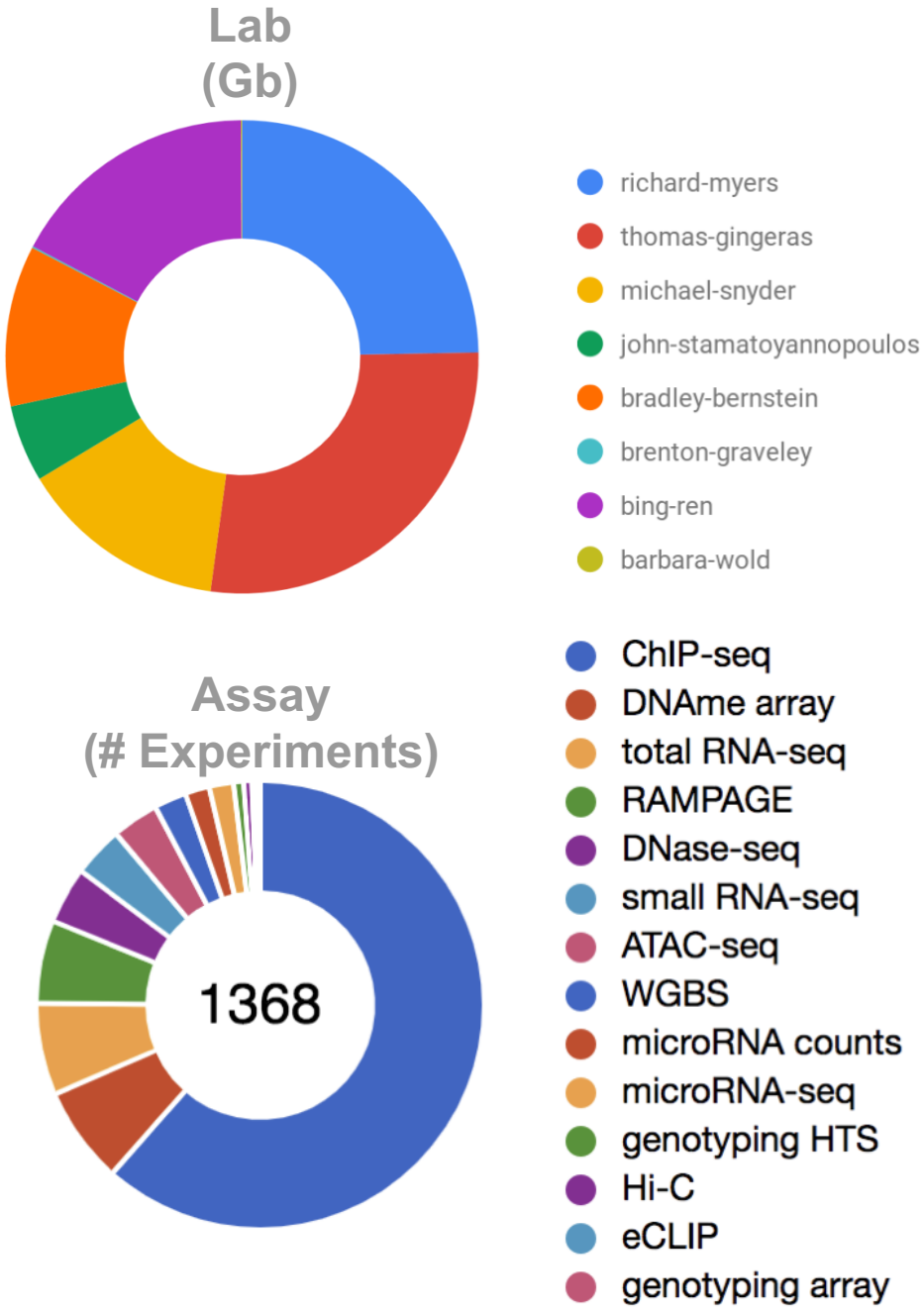
Additional data releases (beyond 1368+139) - expected in Jan & April

Collapsed & simplified analysis matrix

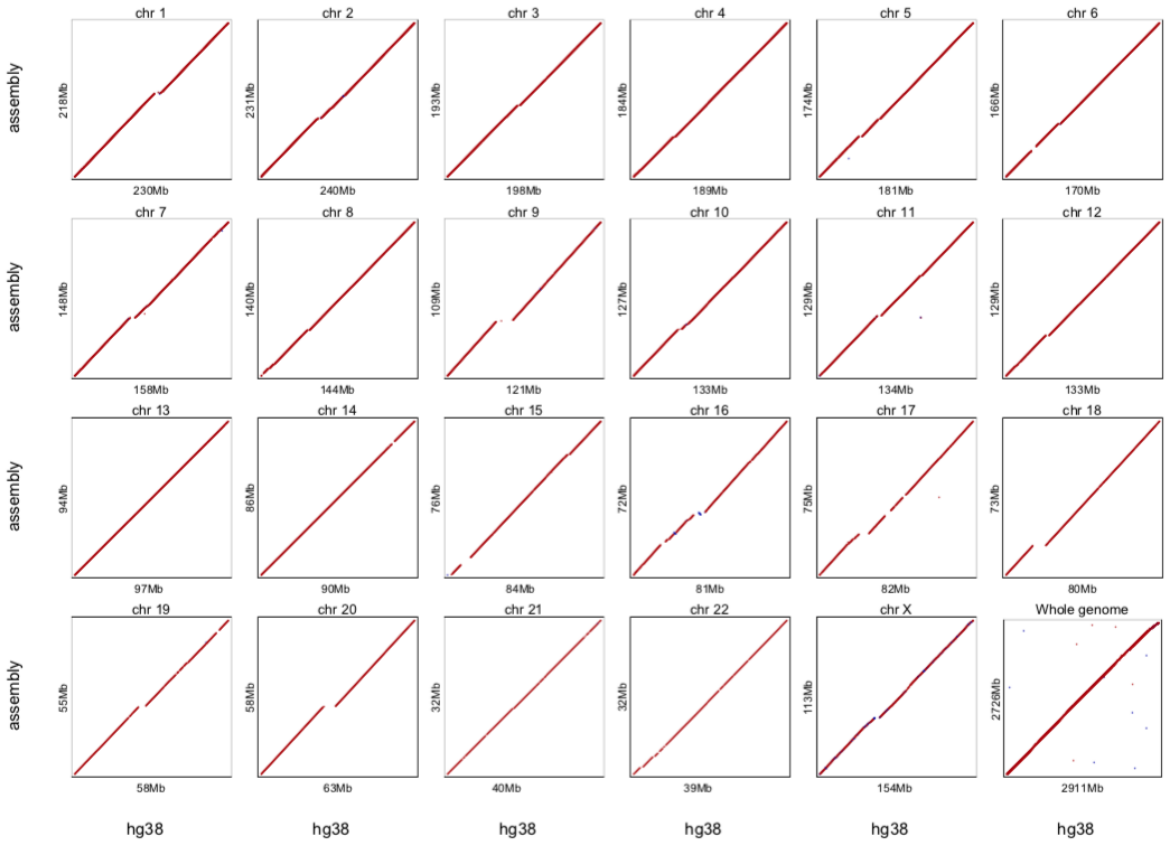


EN-TEx dataset (Number of assays)





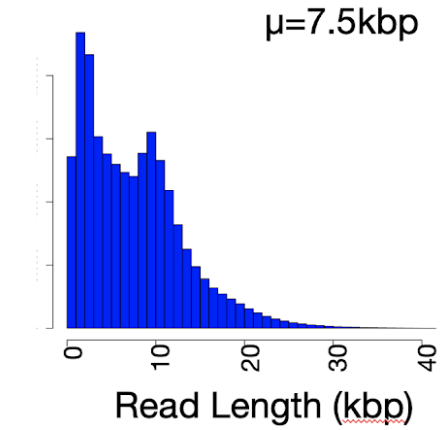
EN-TEx Hi-C Assembly with chromosomal length scaffolds



Comparison with reference

PacBio

Long SVs (~50bp- > 25000kb)



55x Long Reads
****Only for ENC-002 & ENC-003***

HiC Analysis

Illumina m=350bp;
> 120x Paired End
All 4 samples
Full chromosome phasing

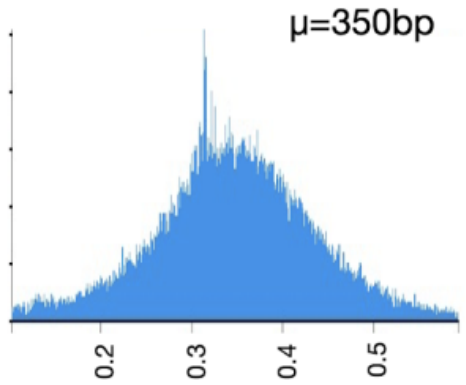
Ren Lab Hi-C data
Aiden Lab assembly
(more on tools later)

Aiden lab & Olga Dudchenko

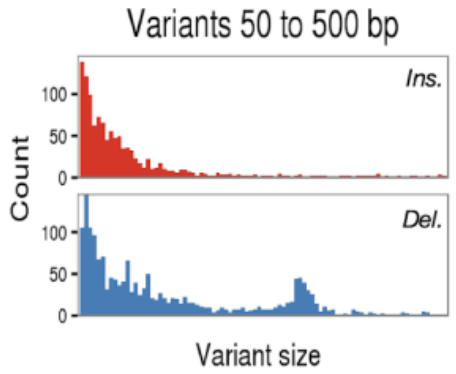
EN-TEx WGS with whole chromosome phasing

Illumina

SNPs

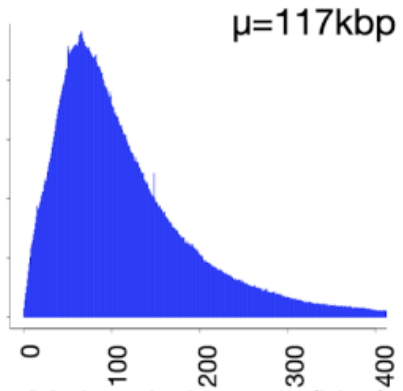


60x Paired End
All 4 samples

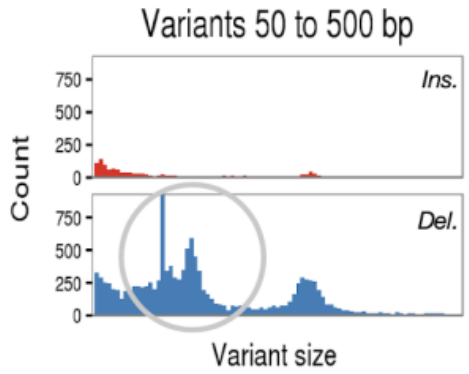


10X Genomics

Phasing and long SV

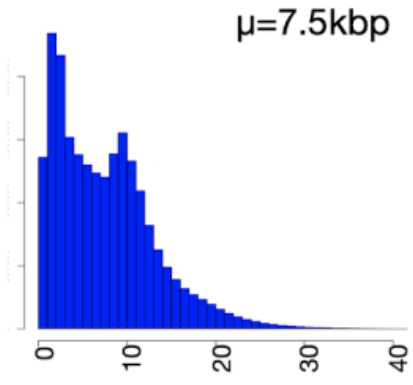


35x Linked Reads
All 4 samples

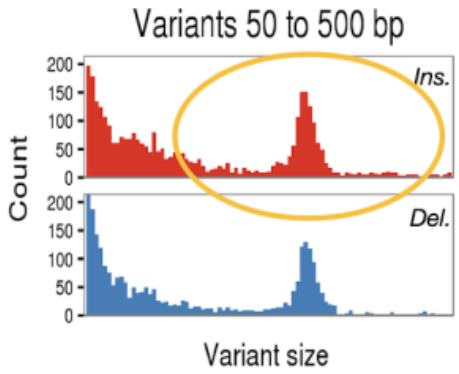


PacBio

Long SVs (~50bp- > 25000kb)



55x Long Reads
****Only for ENC-002 & ENC-003***



HiC Analysis

Illumina m=350bp;
> 120x Paired End
All 4 samples

Full chromosome phasing

Michael Schatz, Bing Ren, Thomas Gingeras

Variant calls

	Sub	Ins (<50)	Del (<50)	SV (>50)
ENC-001	3,895,883	255,857	281,757	0
ENC-002	3,877,699	272,322	296,040	24,696
ENC-003	4,024,169	288,027	314,366	26,181
ENC-004	3,947,657	263,092	293,823	0

SV count by type

Sample	SV-Ins	SV-Del	Inv	Dup	Trans
ENC-002*	11,745	10,460	281	1,850	360
ENC-003*	13,077	10,813	248	1,746	297
GIAB son*	13,750	11,487	300	1,647	0
NA12878*	11,127	9,385	230	1,057	394
1000 Genomes‡	168	42,279	786	6,025	0

* PacBio Long Reads w/NGMLR & Sniffles
‡ Illumina short reads across 2,504 humans

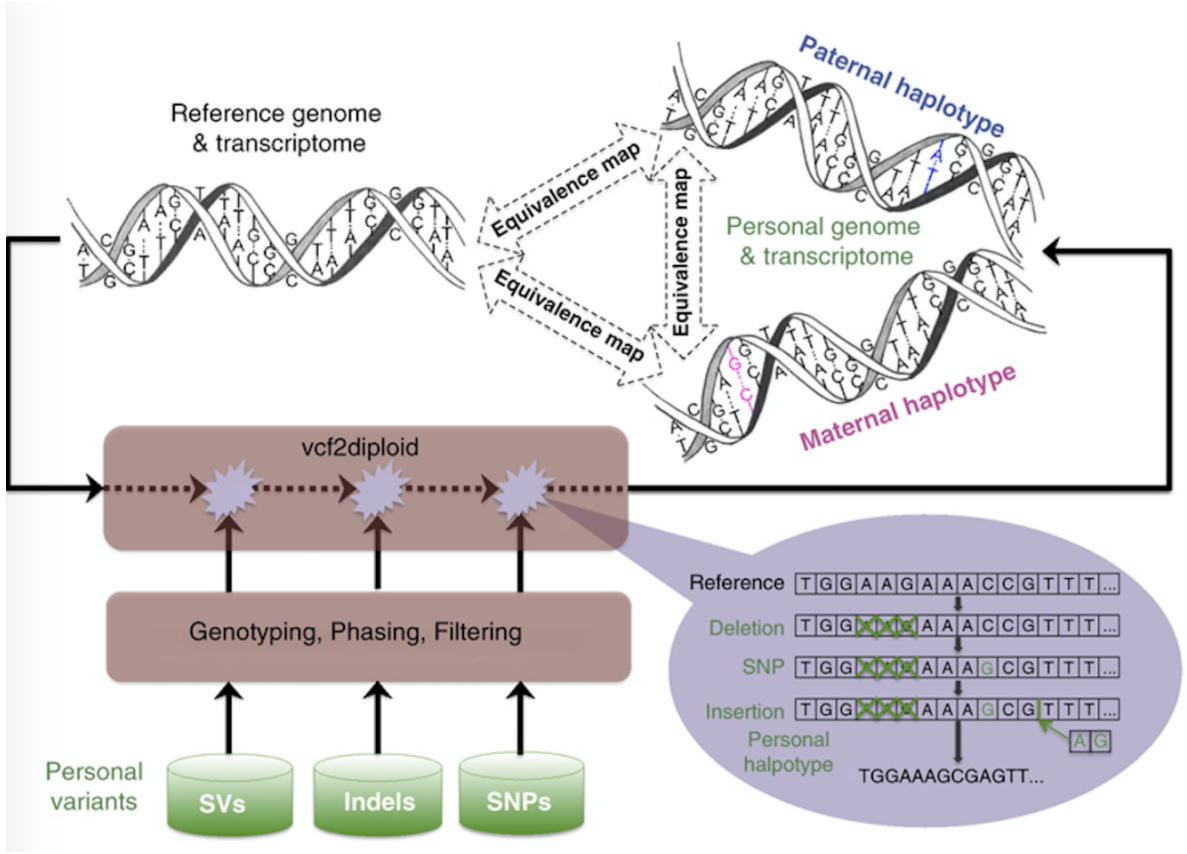
EN-TEX SV calls: very developed & accurate

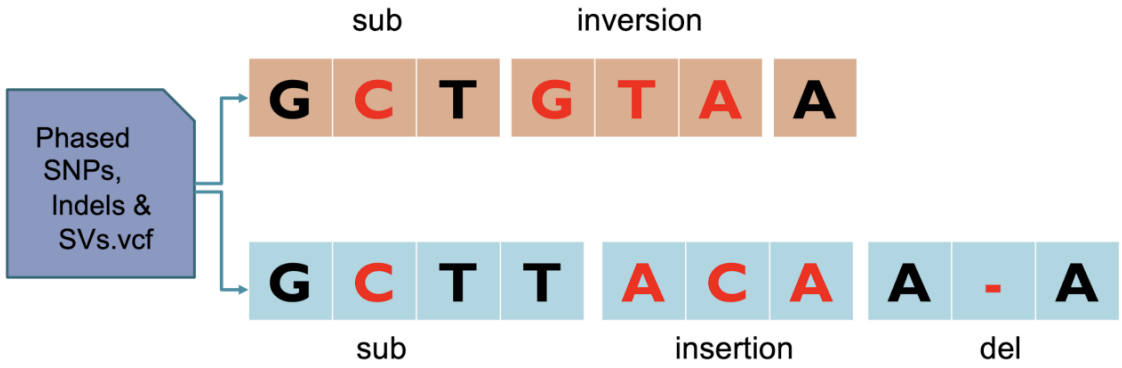
SV concordance among genomes

	ENC002	ENC003	NA12878	GIAB
ENC002	19725	12297	10746	12573
ENC003	12297	20693	10771	12551
NA12878	10746	10771	16966	11237
GIAB	12573	12551	11237	21036

Michael Schatz

Assembling a phased personal diploid genome from all the variant calls





Stitching based on AlleleSeq pipeline enhanced for SVs
• Maintains a mapping from reference to personal genome coordinates for liftover

Using 10X + HiC + PacBio, assemble nearly perfect diploid human genomes
• Phased diploid genome can be aligned or aligned against

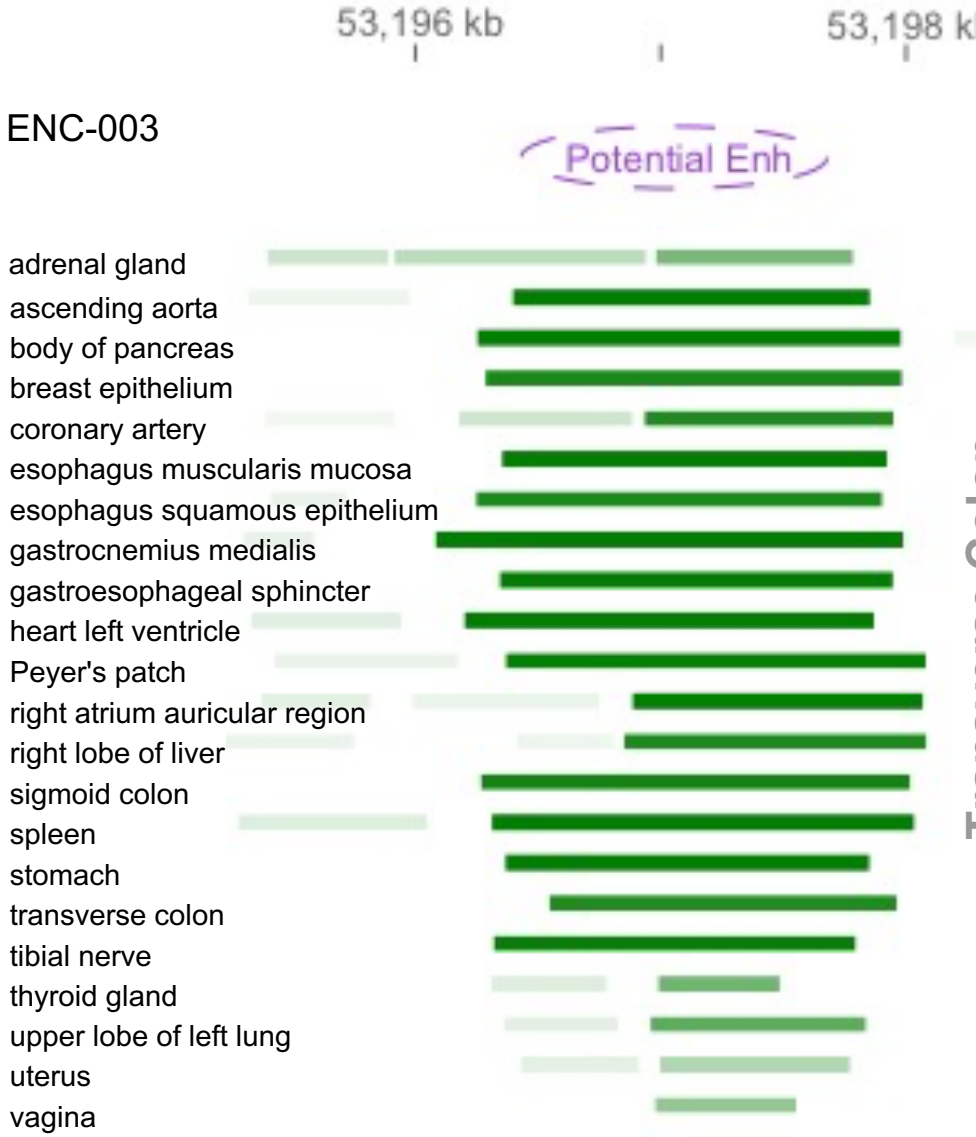
Rozowsky et al. 2011
Michael Schatz

2. Variation in Transcriptome & Epigenome across Individuals & Tissues

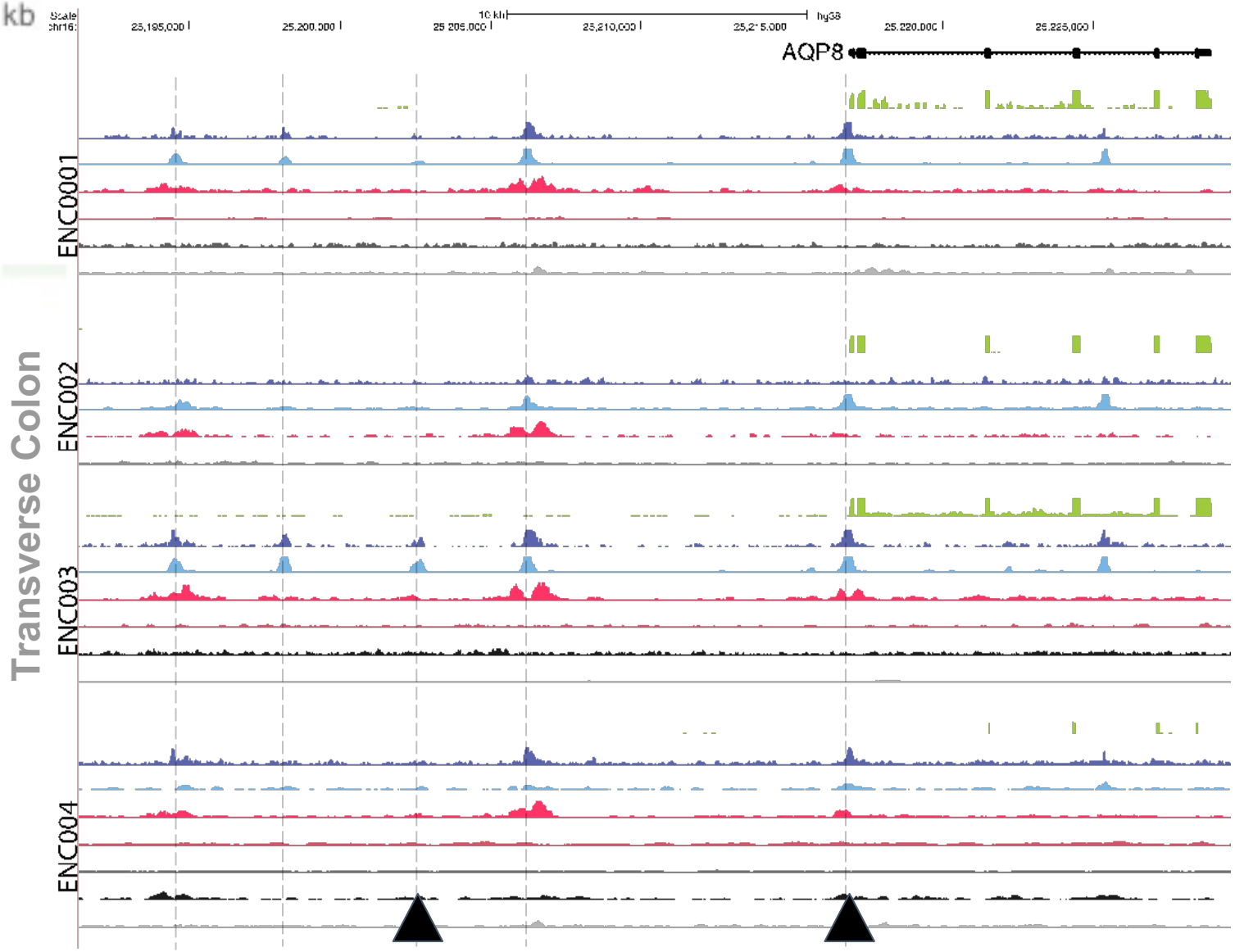
Var. across tissues & individuals (single locus)

RNA-seq ATAC-seq DNAse H3K27ac
H3K27me3 H3K36me3 H3K4me1 H3K4me3

Peak variation across tissues



Signal variation across individuals



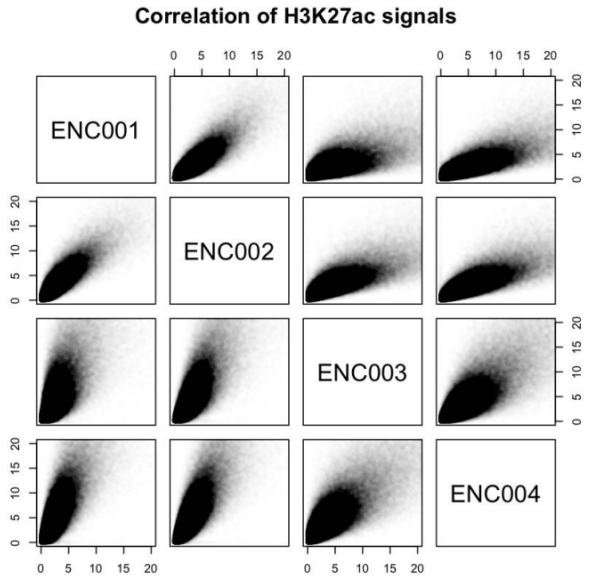
Chromatin variation across individuals (aggregated): Integration vs Single assay

Transverse Colon

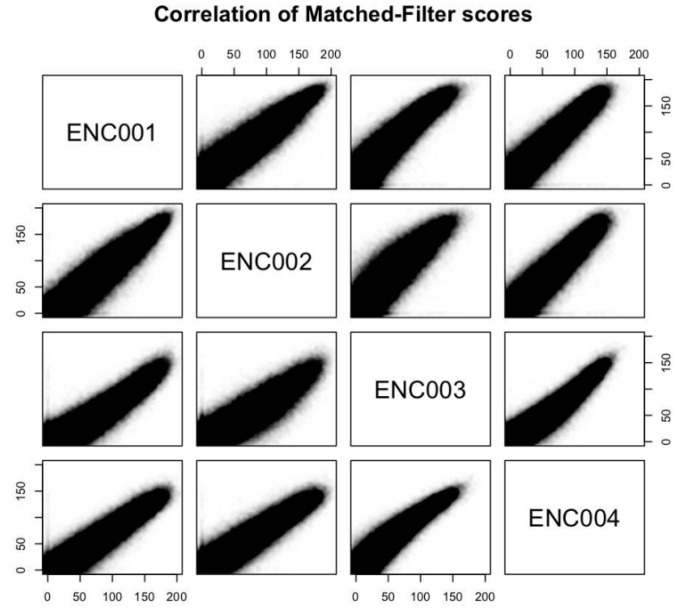
Distal DHS

	ENC-001	ENC-002	ENC-003	ENC-004
ENC-001	161005	108157 (67%)	116780 (73%)	57606 (36%)
ENC-002	108493 (69%)	156610	104568 (67%)	62256 (40%)
ENC-003	116976 (75%)	104364 (66%)	156961	57176 (36%)
ENC-004	58353 (37%)	63135 (40%)	57818 (36%)	159323

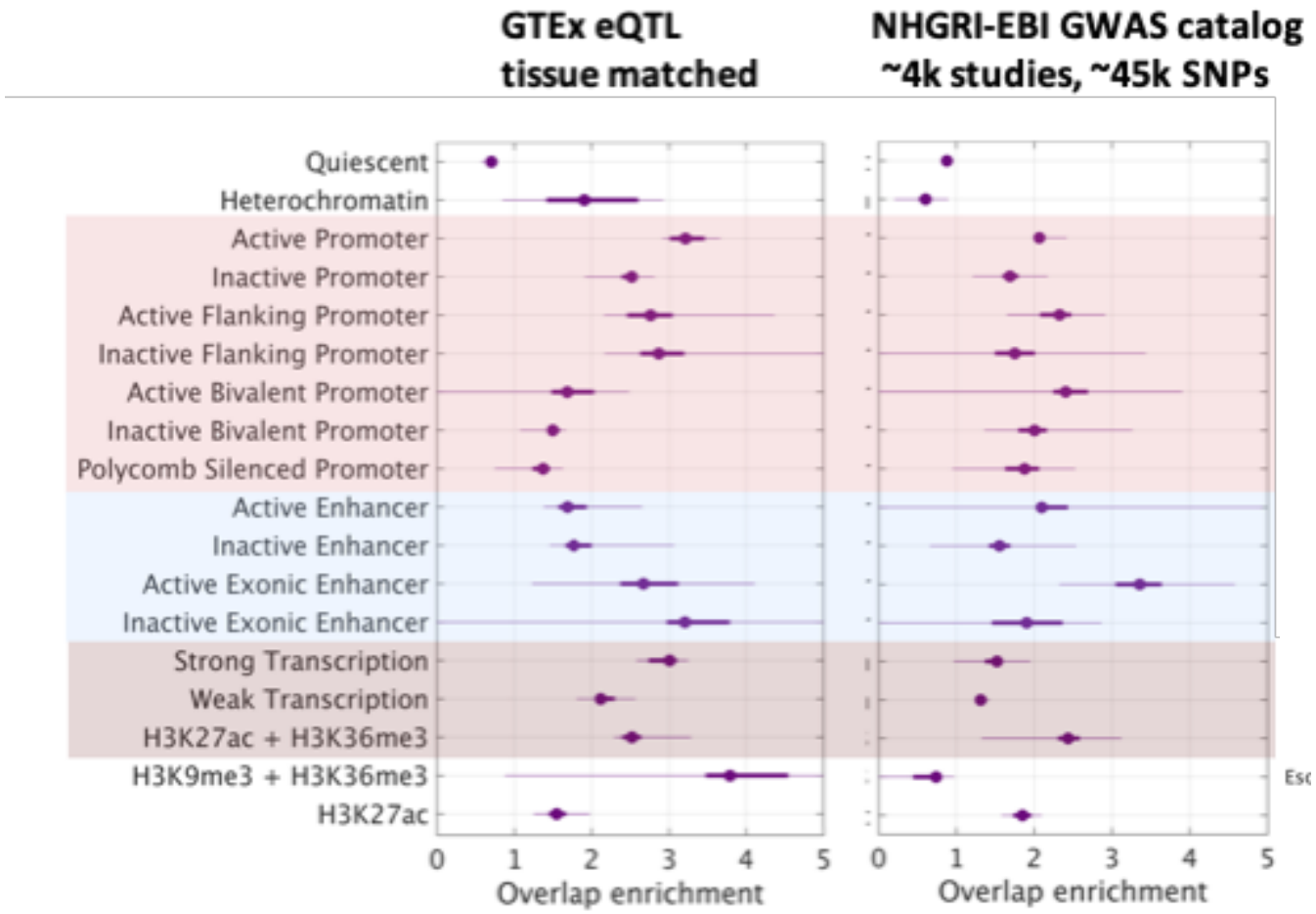
Correlation of K27 Signal



Corr. of Integrated Cross-signal Score (MF Lin. Combo)



Methods of integration (chromHMM)



GWAS variants in "**cholesterol**" studies

83 studies; 1,183 SNPs; 680 unique SNPs

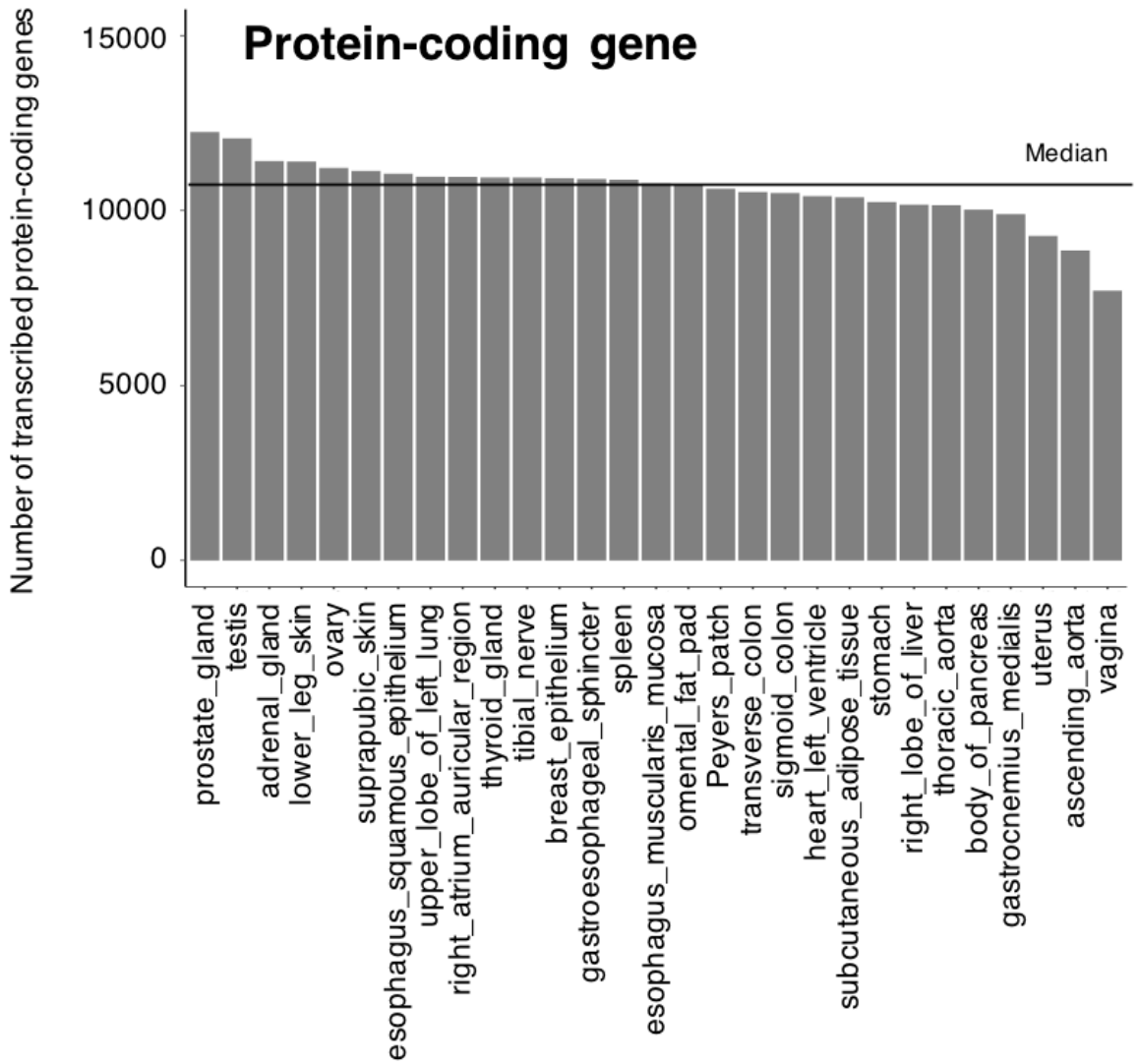
EN-TEx ChromHMM segmentation: one donor, 10 tissues

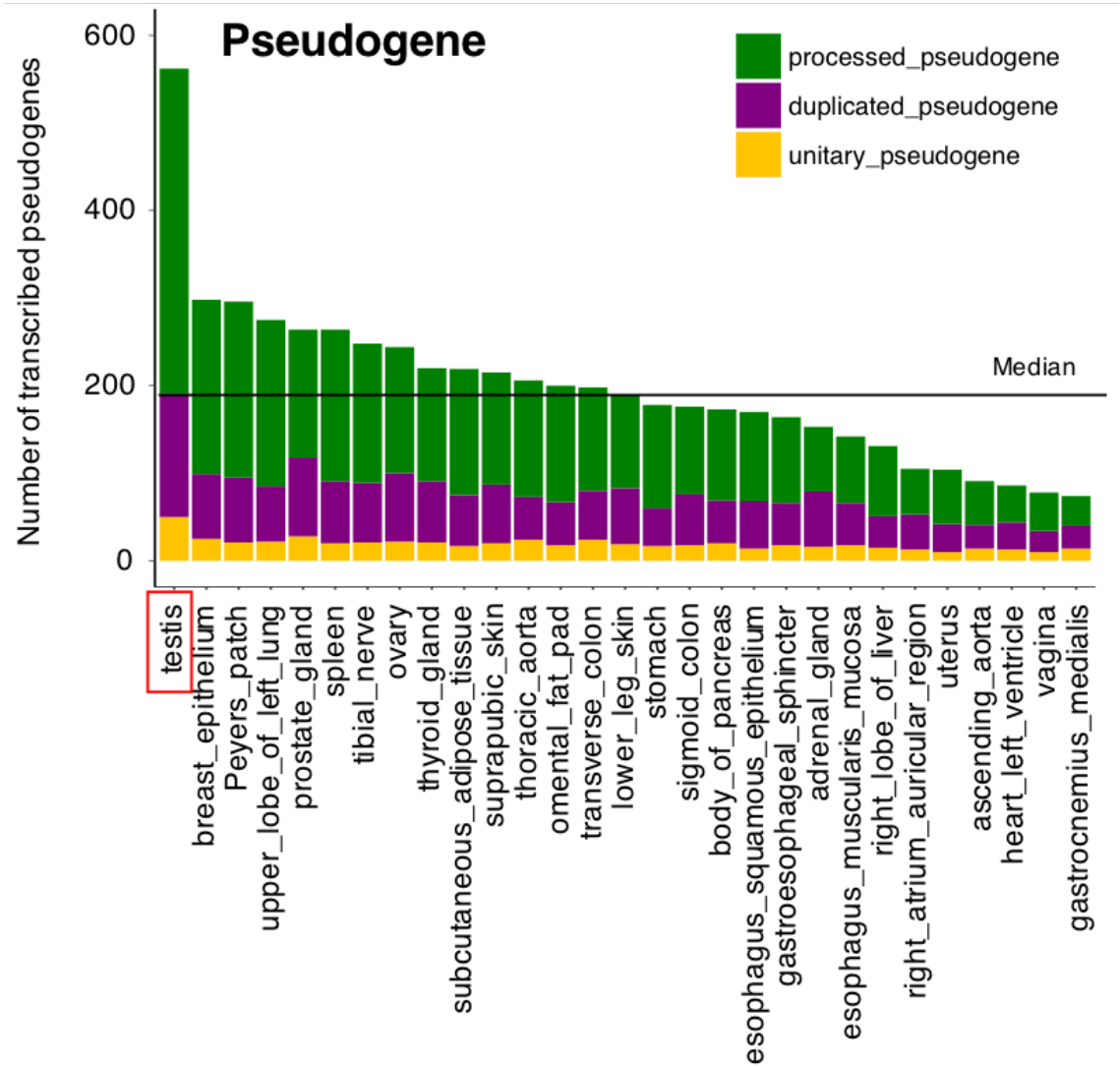
H3 K4 me3	H3 K4 me1	H3 K27 ac	H3 K36 me3	H3 K27 me3	H3 K9 me3	Genome coverage %	Exons	TSS +/- 2kb	Chromatin State
0	0	0	0	0	0	75.5	0.5	0.5	Quiescent
0	0	0	0	1	45	1.6	0.7	0.4	Heterochromatin
98	4	98	0	1	0	0.7	9.0	23.5	Active Promoter
79	7	9	0	2	0	0.3	5.1	16.5	Inactive Promoter
97	87	99	0	1	0	0.2	4.2	16.9	Active Flanking Promoter
90	82	7	0	2	0	0.1	5.0	19.6	Inactive Flanking Promoter
87	31	88	1	73	4	0.03	8.5	19.4	Active Bivalent Promoter
62	41	1	1	84	5	0.11	8.5	16.6	Inactive Bivalent Promoter
0	0	0	0	33	0	0.8	2.6	5.7	Polycomb Silenced Promoter
3	74	93	1	0	0	1.0	1.6	2.5	Active Enhancer
0	48	6	0	0	0	1.0	1.3	3.2	Inactive Enhancer
37	55	90	80	0	0	0.04	11.7	5.4	Active Exonic Enhancer
4	49	9	68	0	0	0.03	8.8	3.5	Inactive Exonic Enhancer
0	0	0	76	0	0	2.5	5.9	0.7	Strong Transcription
0	0	0	3	0	0	12.0	2.1	1.0	Weak Transcription
1	3	58	75	0	0	0.14	8.1	1.2	H3K27ac + H3K36me3
4	0	1	70	0	80	0.12	7.6	0.9	H3K9me3 + H3K36me3
1	6	65	0	0	0	1.9	1.3	2.4	H3K27ac

One of the most comprehensive datasets of consistent collection of histone modifications for multiple individuals and tissues allows better chromatin segmentation

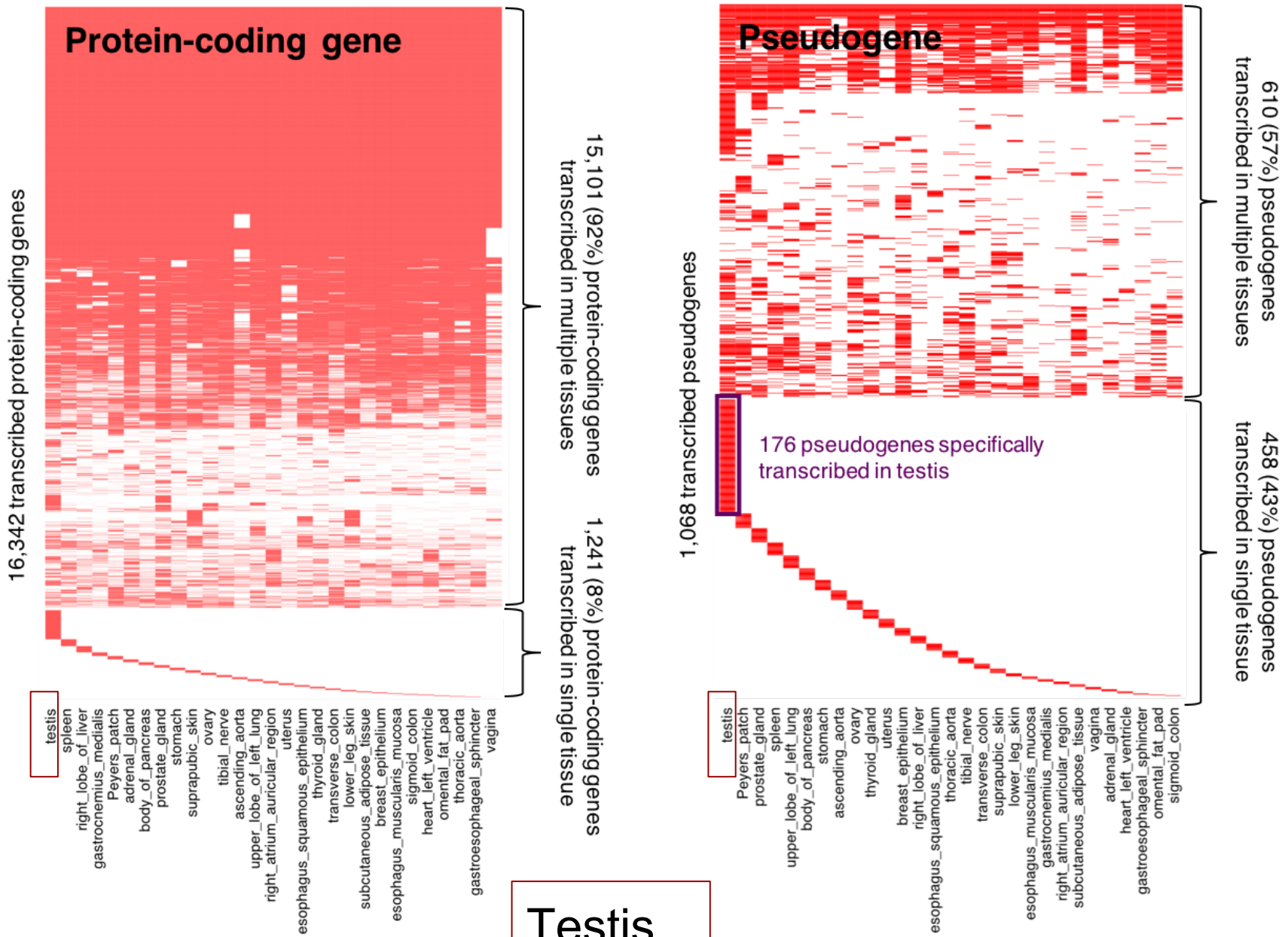
[Alex Dobin]

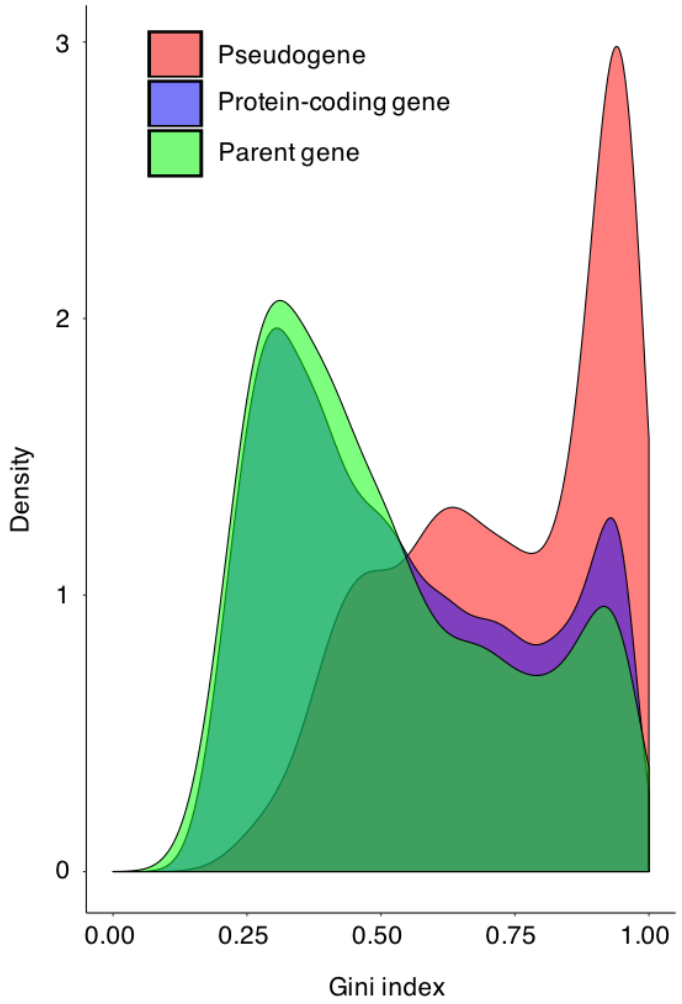
Landscape of EN-TEx transcription





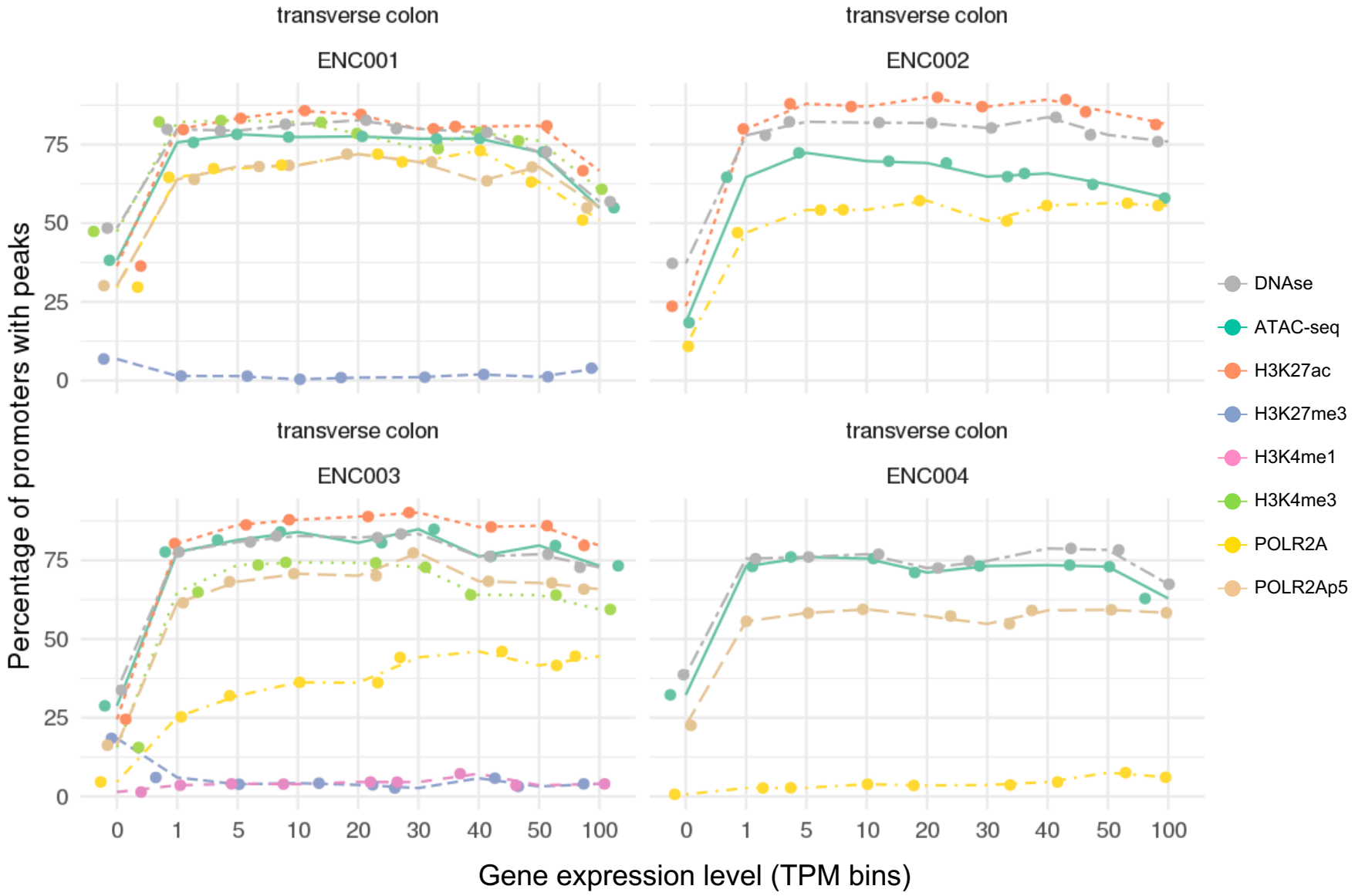
Tissue-specificity of gene vs pseudogene expression





Testis

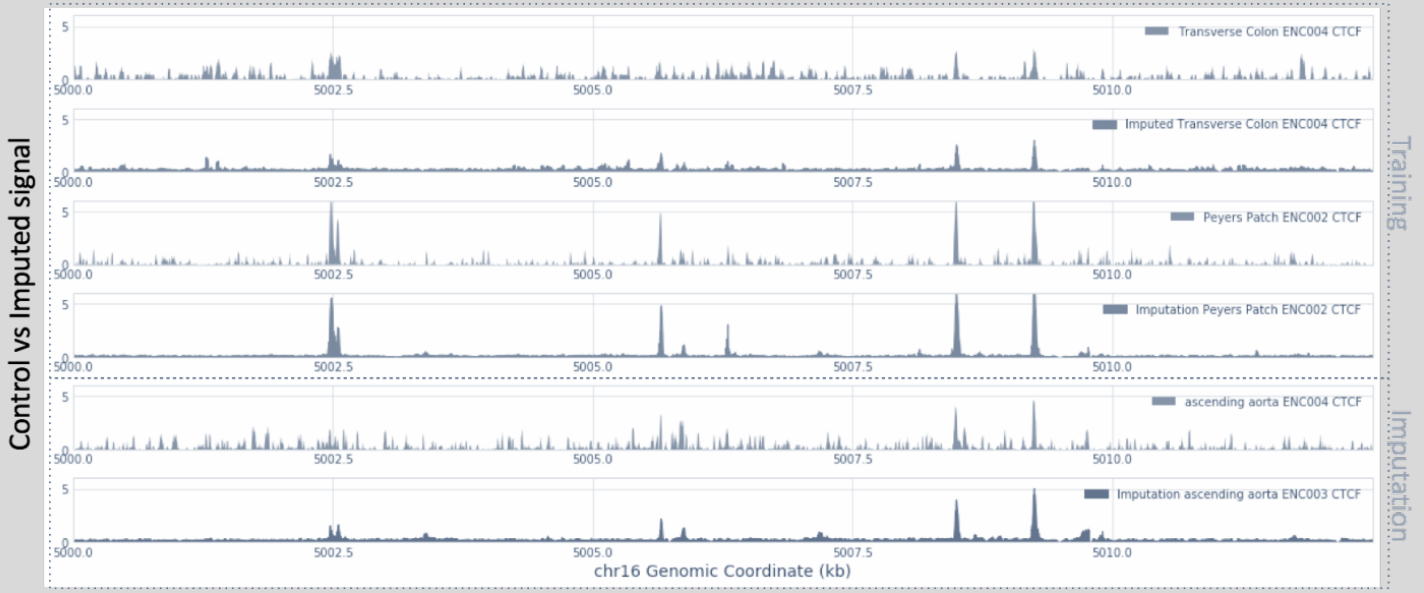
Comparing chromatin activity with gene expression



Imputation & RCA visualization

Application of Avocado to the ENTEx Dataset

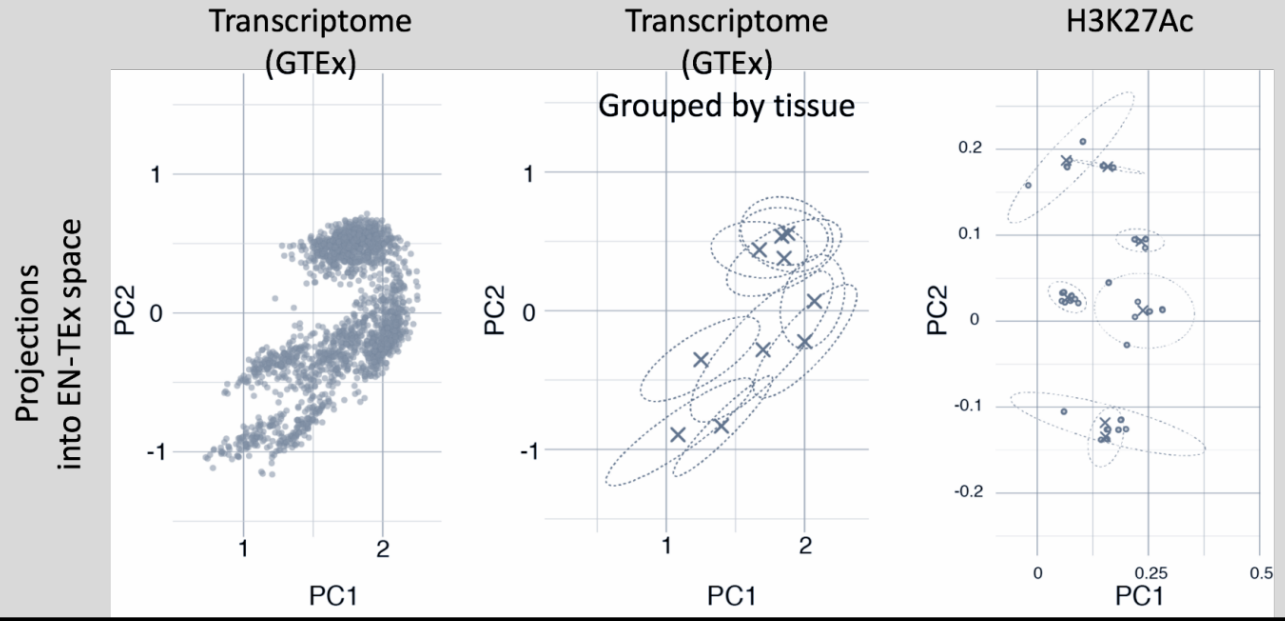
910 sample imputations from Avocado (Deep learning mod
Histone modifications + CTCF + RNApol II



Jacob Schreiber, Bill Noble

Reference Component Analysis (RCA)

Applied to EN-TEx data, giving consistent transcriptome v epigenome comparison

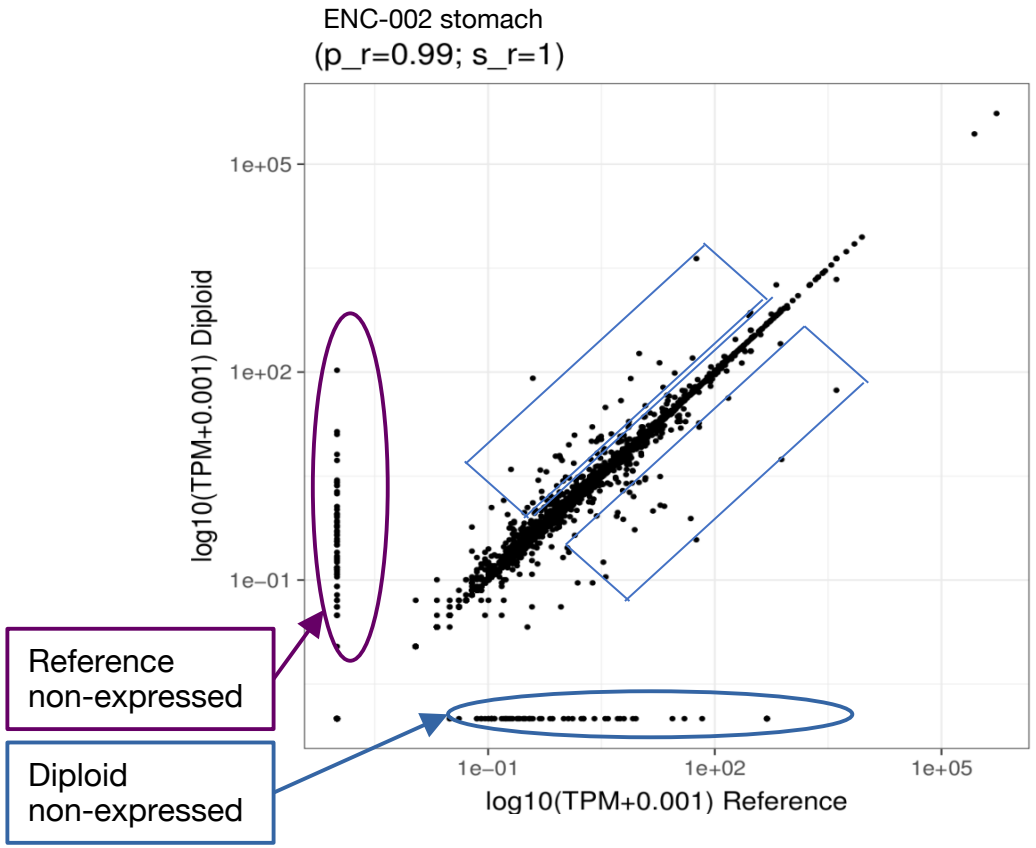


Already discussed in DAC presentation

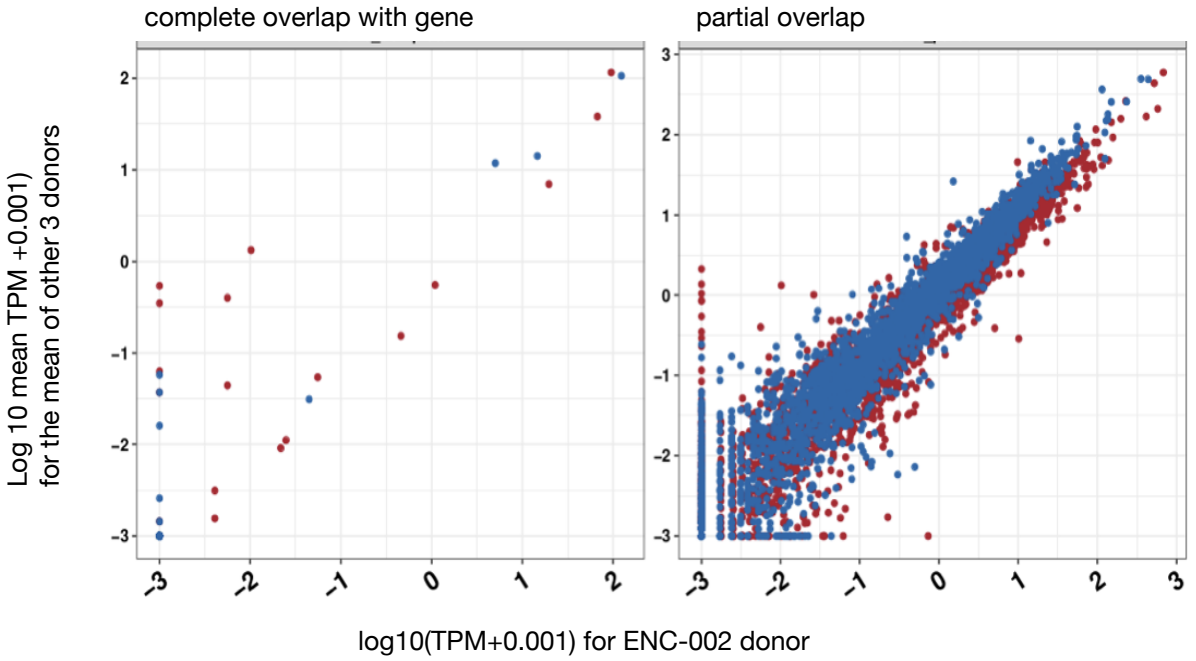
3. Detailed Effect of Specific Genomic Variants from the Personal Genome

Diploid personal genomes: Effect of variants on Gene Expression

Genes quantified differently when mapped to diploid personal genome compared to the reference genome



Effect of deletions on gene expression



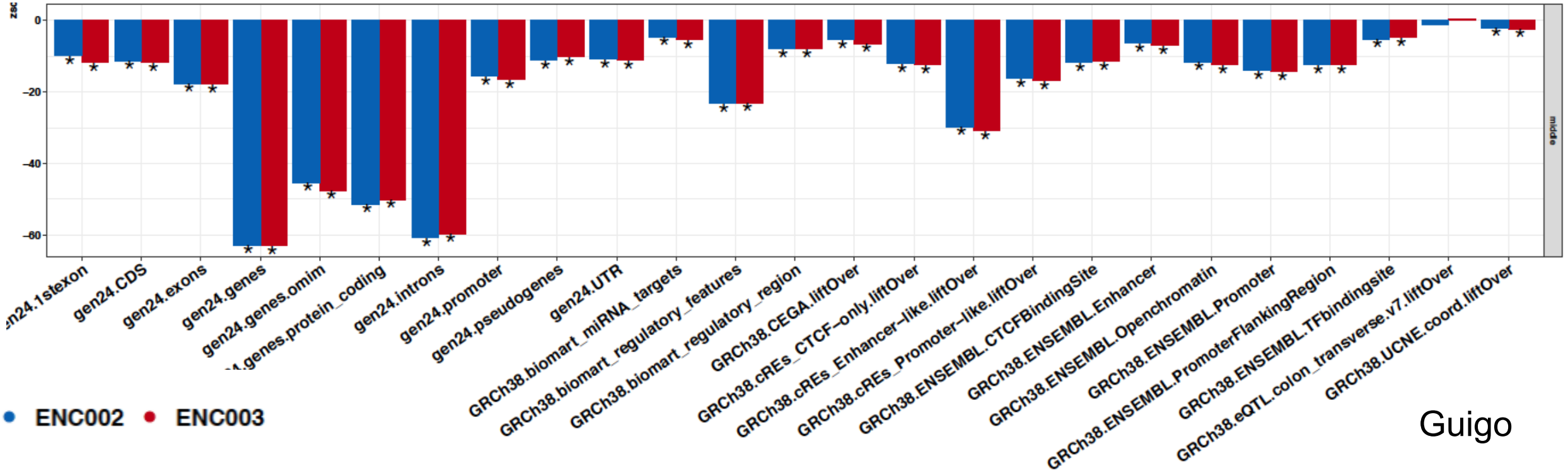
- 728 (252 protein-coding) genes quantified differently
- Examples of protein-coding genes:
  - with significantly higher expression across multiple tissues in personal genome: HLA-DRB1, HLA-DQA1, HLA-DQB1, IGHV4-31
  - in the reference genome: NBPF26, UGT2B15, HLA-DRB5, AC073333.1, RP11-514P8.6, FOXD4L3, NANOG, CTD-3126B10.5, RIMBP3

Anna Vlasova, Roderic Guigo, Thomas Gingeras

Functional elements are depleted in SVs

Calculated **partial** (1bp) and **complete** overlap between SVs and genomic elements

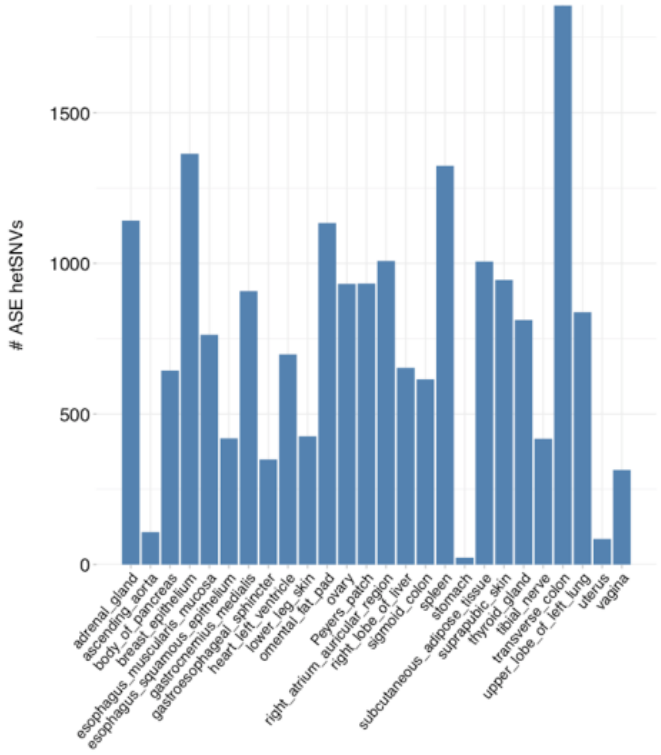
Complete overlap

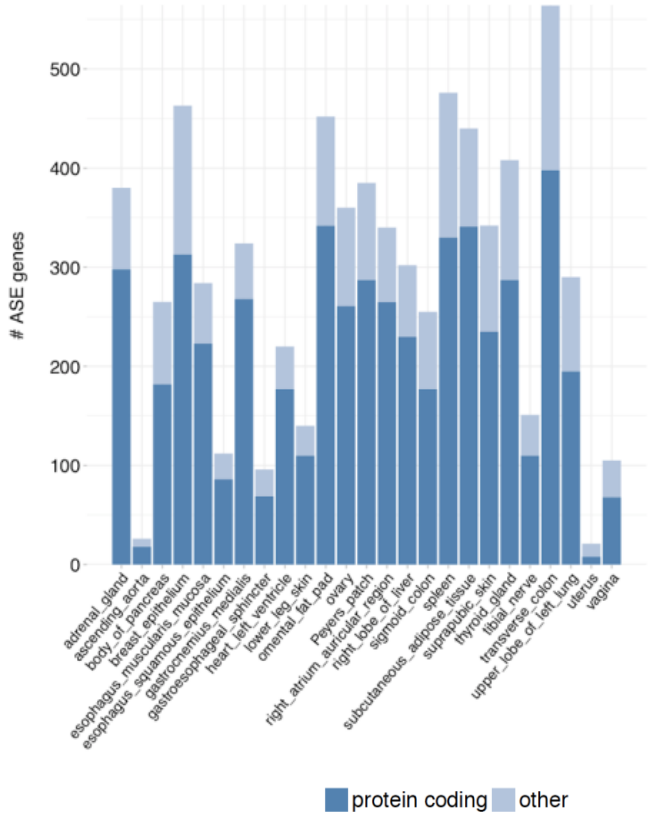


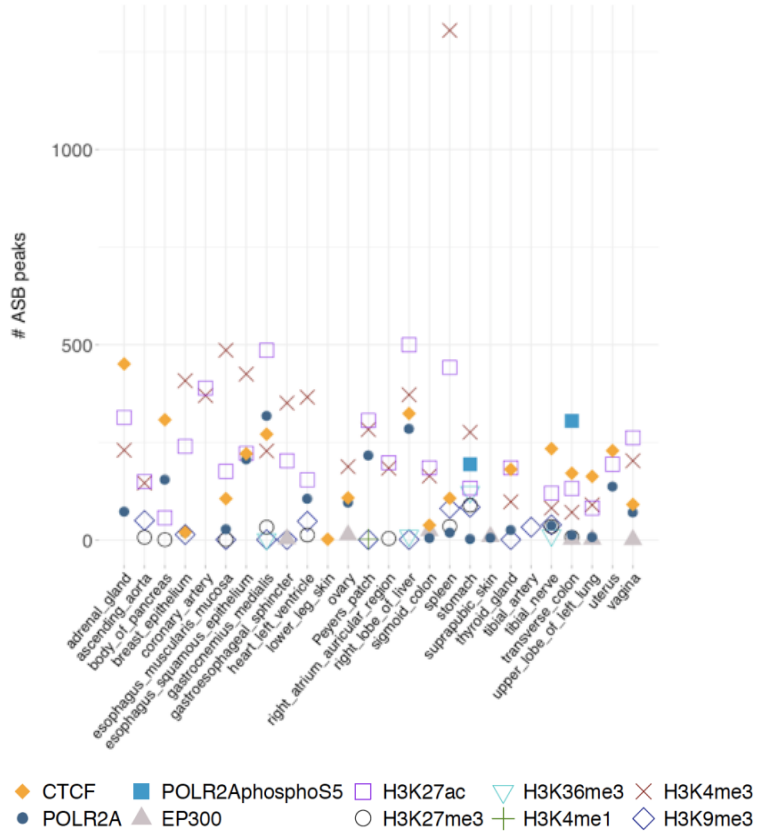
Guigo

Generating a catalogue of SNVs, genes, & binding sites associated with allelic activity across different tissues

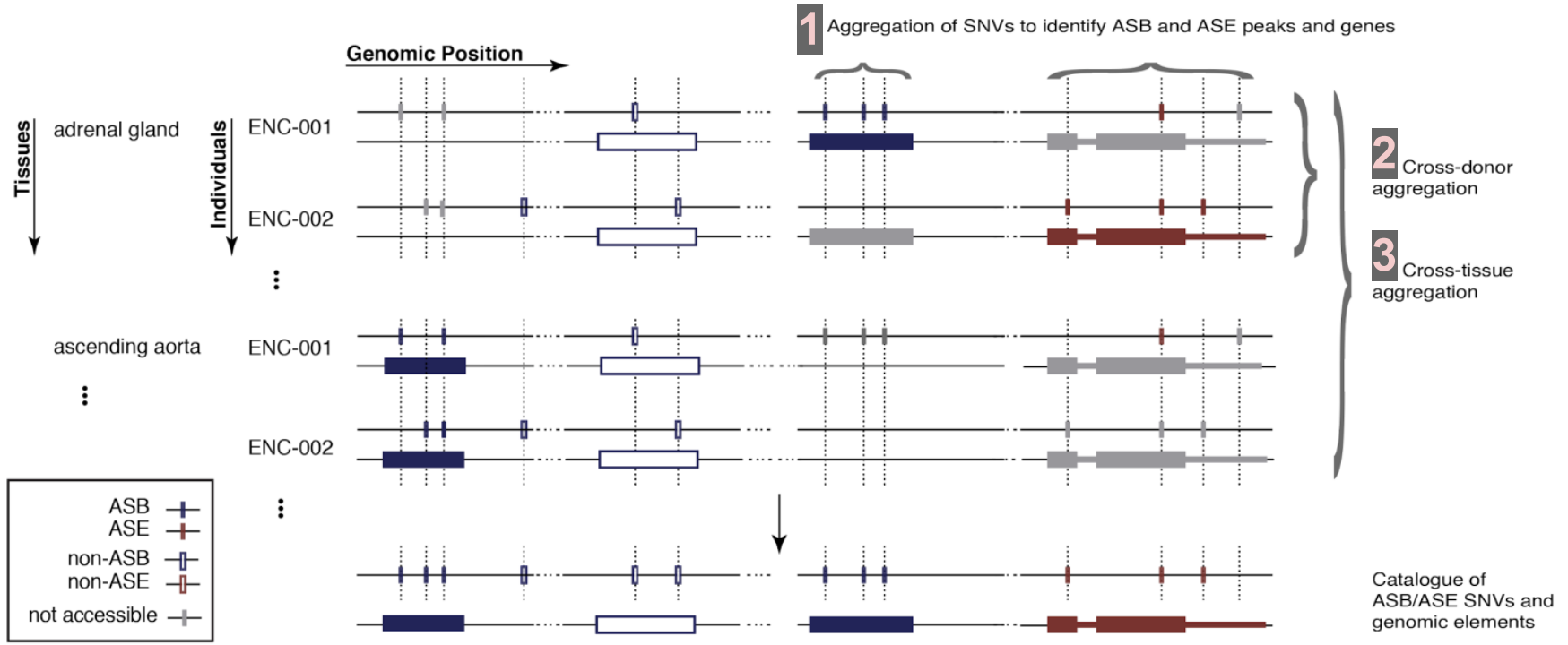
Counts of ASE hetSNVs and genes and ASB peaks; all tissues ENC-003



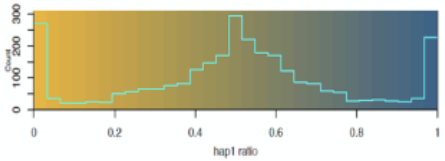


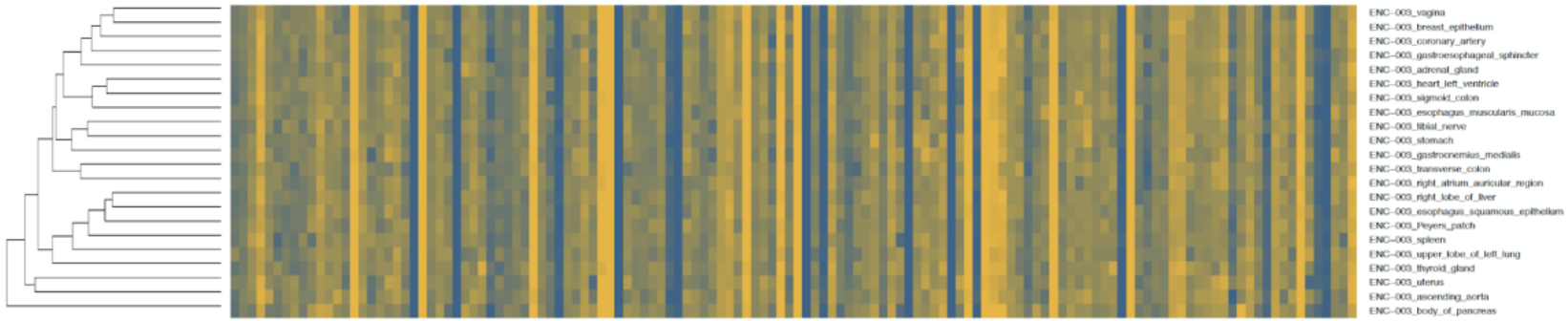


Developing an integrated cross element (1), individual (2) & tissue (3) ASE and ASB annotation for genomic features (genes & ccREs)

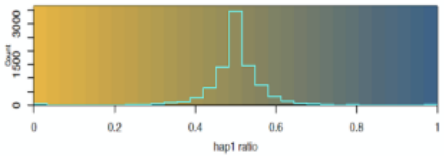


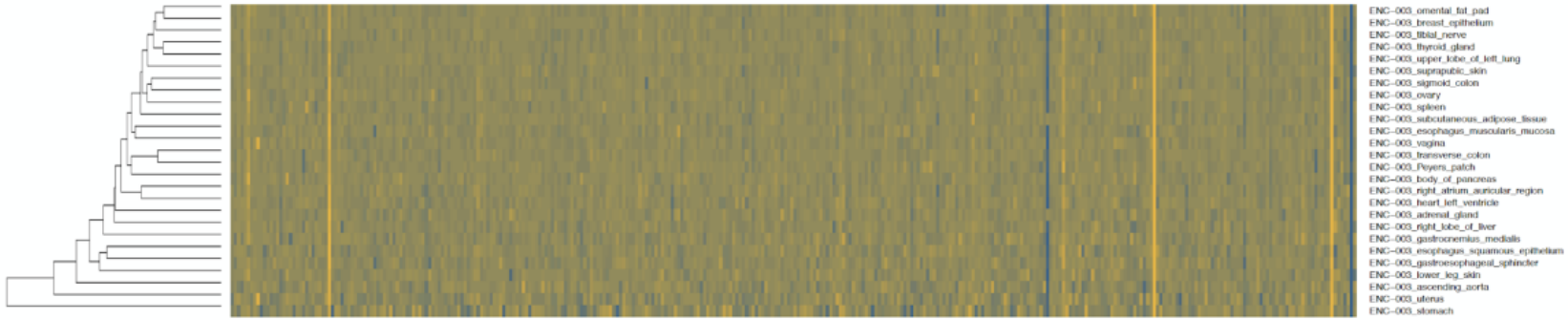
ASB (H3K27ac) in ccREs across ENC-003 tissues





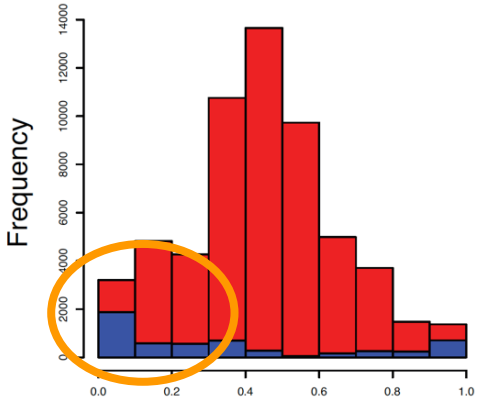
ASE in protein-coding genes across ENC-003 tissues



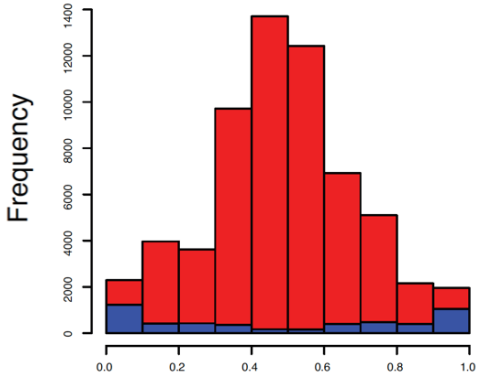


Using personal genomes alleviates reference mapping bias

Reference Genome



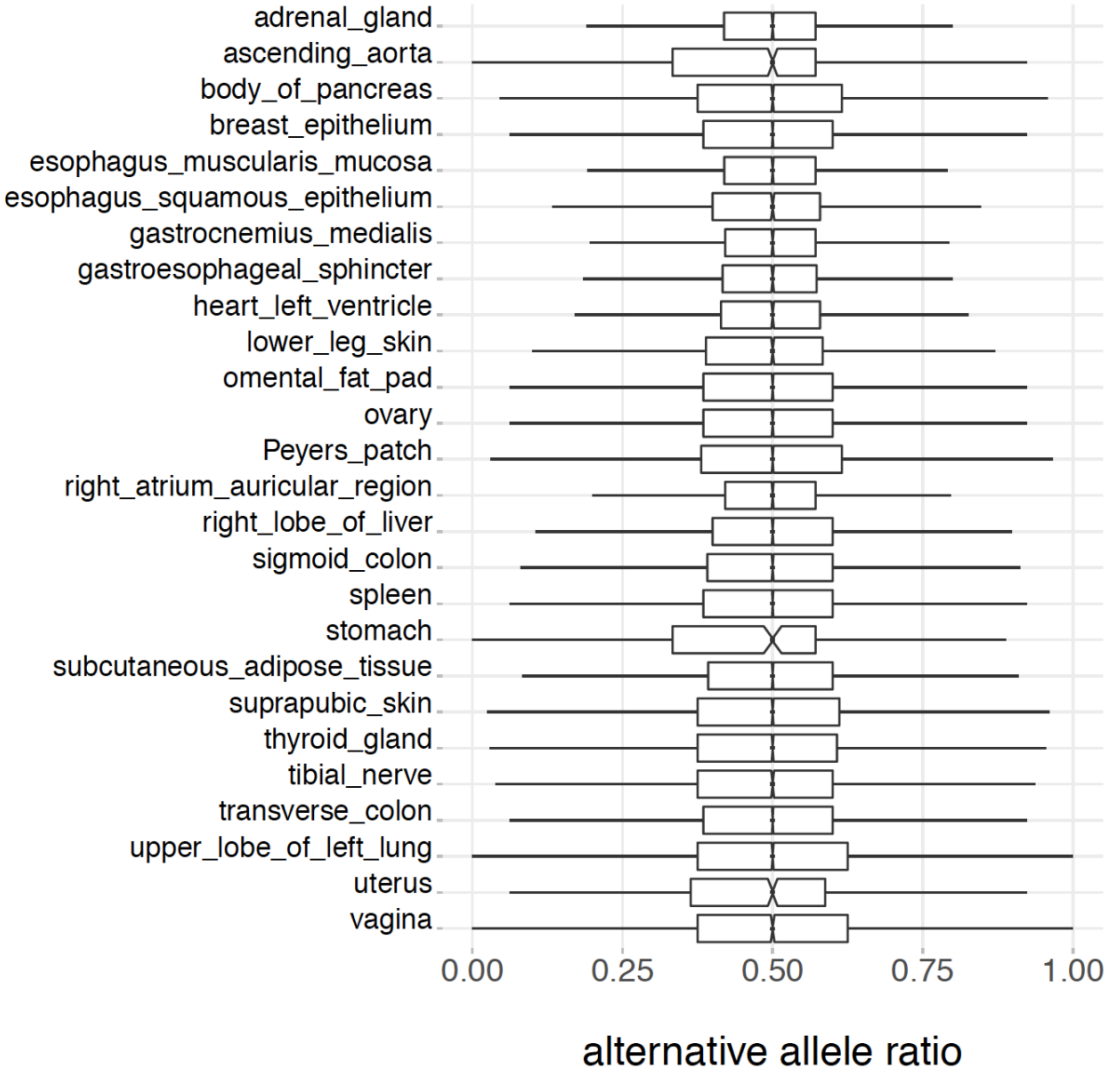
Personal Genome NA12878



alternative allele ratio

(fraction of RNA-Seq reads mapping to alternative allele per heterozygous SNP)

Personal Genome ENC-003; all RNA-seq samples



alternative allele ratio

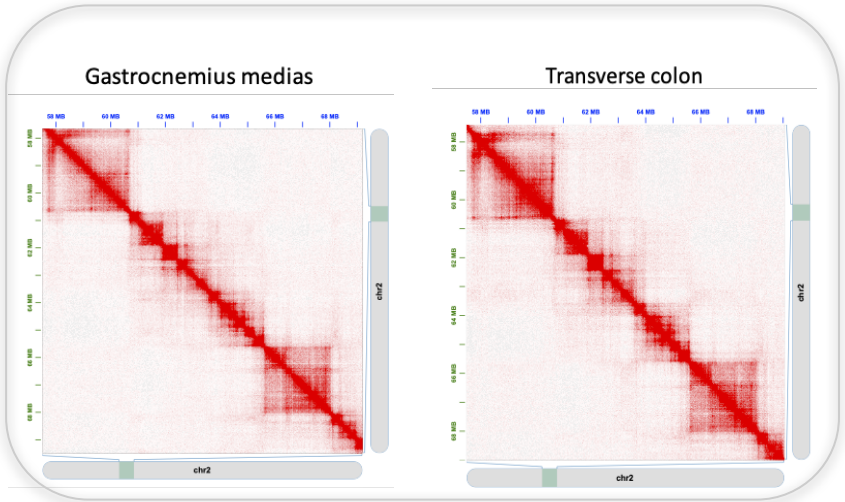
Rozowsky et al. 2011

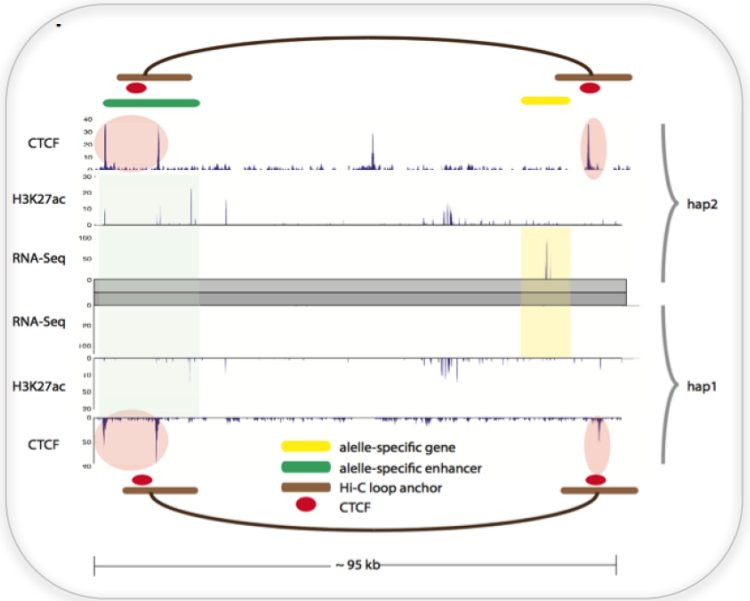
EN-TEx Summary

1) High-quality assembly of personal genomes from integration of long & short read sequencing
   - Determination of many types of SVs
   - Matched data matrix almost filled in
2) Great variation of raw transcriptome & epigenome data across individuals and tissues
   - Assay integration gives a more stable picture for epigenetics across individuals
   - Pseudogenes are more tissue specific than genes
3) Personal genomic variants have clear impact on transcription
   - Generating a catalogue of allelic elements and genes, by integrating across individuals and tissues

Future Directions for the EN-TEx Analysis

- Variation in 3D organization of genome across tissues and individuals
  - loops
  - domains
  - compartments
- Allele-specific Hi-C
- Variation in methylation across tissues and individuals
- Allele-specific methylation





EN-TEX

HudsonAlpha Institute
Richard M. Myers Lab
Stanford University
J. M Cherry Lab
I. Gabdank DCC
M. Snyder Lab

Broad Institute
Bradley E. Bernstein
Charles B. Epstein

Baylor College of Medicine
Erez L. Aiden Lab

Salk Institute
Joseph R. Ecker Lab

Yale University
Mark B. Gerstein
Joel Rozowsky
Fábio C. P. Navarro
Gamze Gursoy
Timur R. Galeev
Mengting Gu
Yucheng T. Yang
Chengfei Yan

Berkeley National Laboratory
Len A Pennacchio Lab

California Institute of Technology
Barbara Wold Lab

Pennsylvania State University
Ross Hardison Lab

University of California San Diego
Bing Ren Lab

Cold Spring Harbor Laboratory
Thomas Gingeras
Alexander Dobin
Jorg Drenkow
Cassidy Danyko
Jesse Gillis
Carrie A. Davis
Chris Zaleski
Alex Scavelli

NHGRI (funding agency)
Mike Pazin
Elise Feingold
Dan Gilchrist
Michael Pagan
Eileen Cahill

Center for Genomic Regulation
Roderic Guigo
Julien Lagarde
Anna Vlasova
Alessandra Breschi
Sarah Djebali
Dimitri Pervouchine

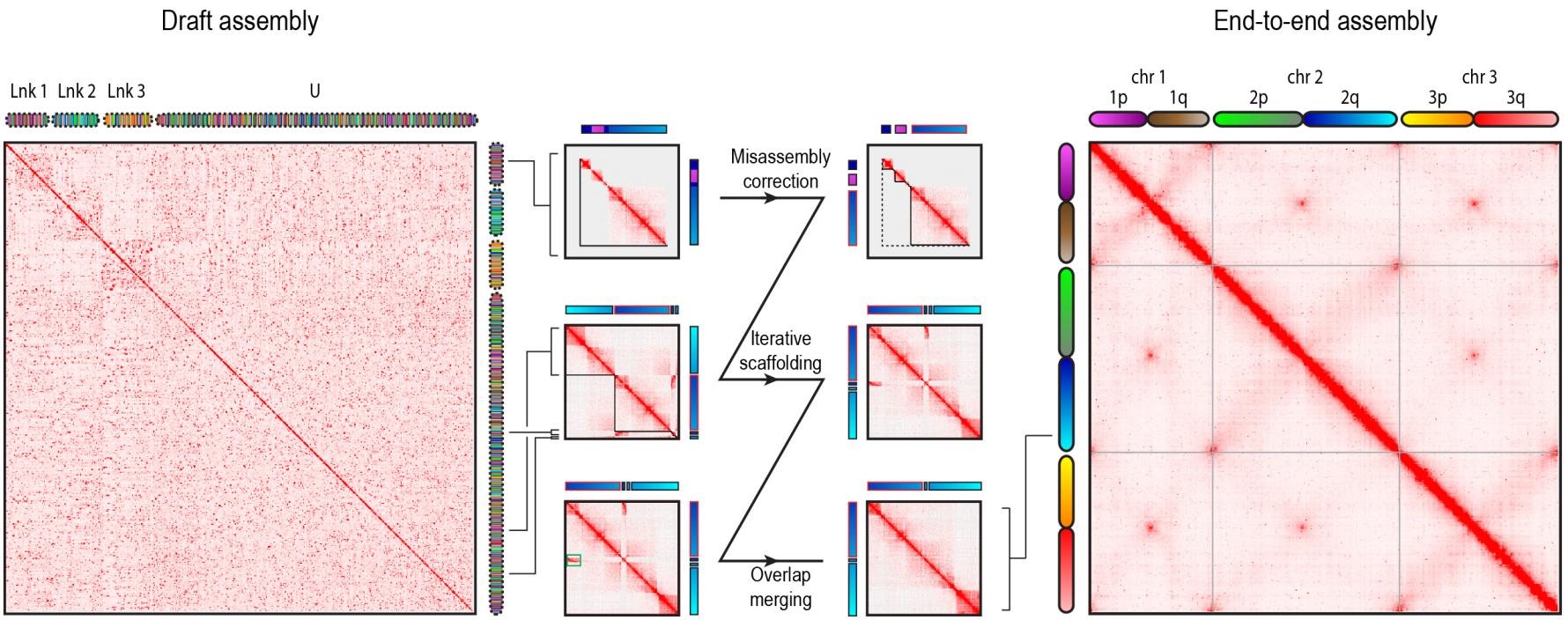
University of Connecticut
Brenton Graveley Lab

University of Washington
John A. Stamatoyannopoulos
Rajinder Kaul
Jacob Schreiber
Bill Noble

University of Massachusetts
Zhiping Weng
Jill Moore
Henry Pratt
University of California, Irvine
Ali Mortazavi
Rabi Murad
Johns Hopkins University
Michael Schatz
Fritz Sedlazeck
Han Fang
Maria Nattestad

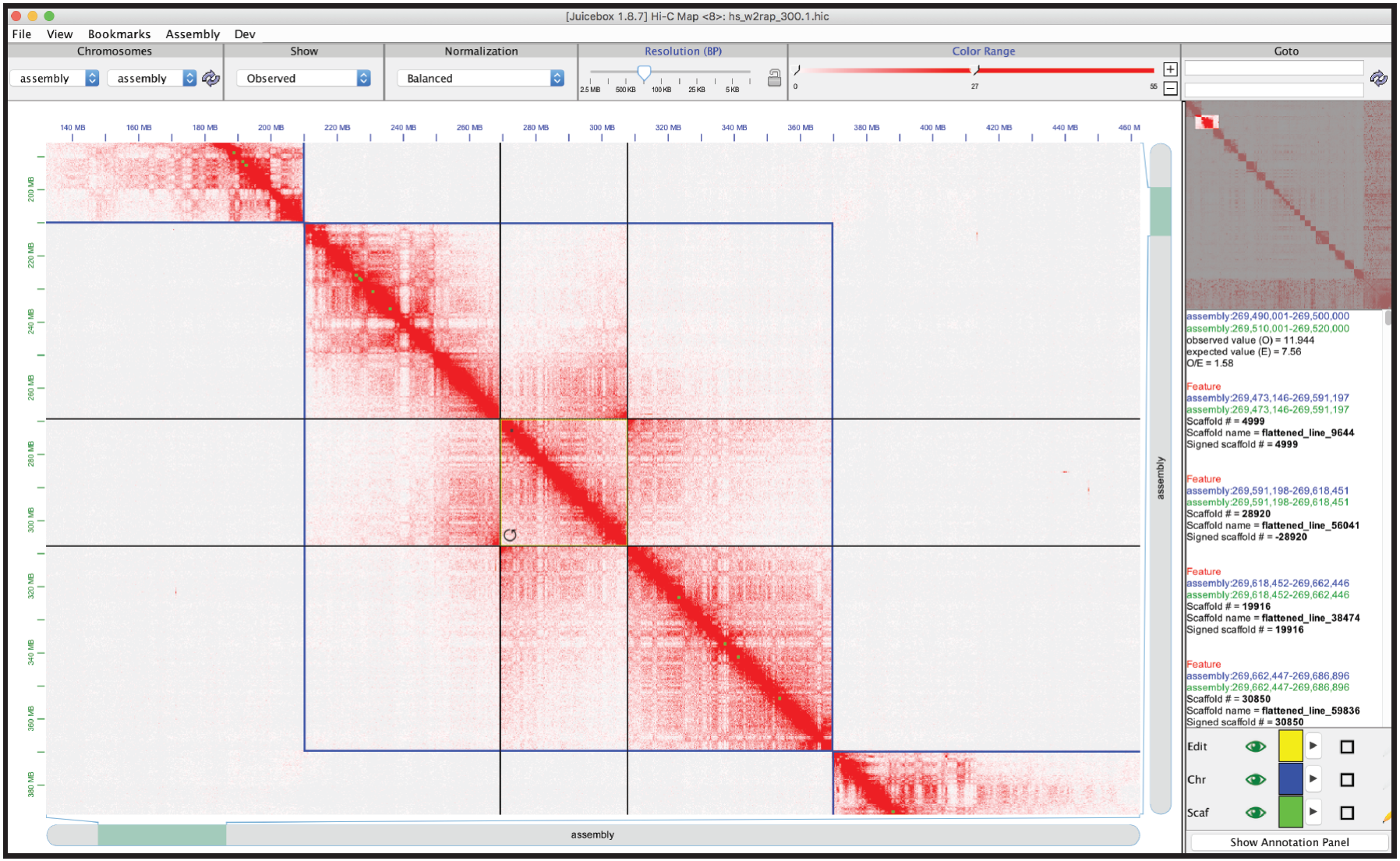
PERSONALIZED GENOMES

3D-DNA: AUTOMATIC 3D DE NOVO GENOME ASSEMBLY



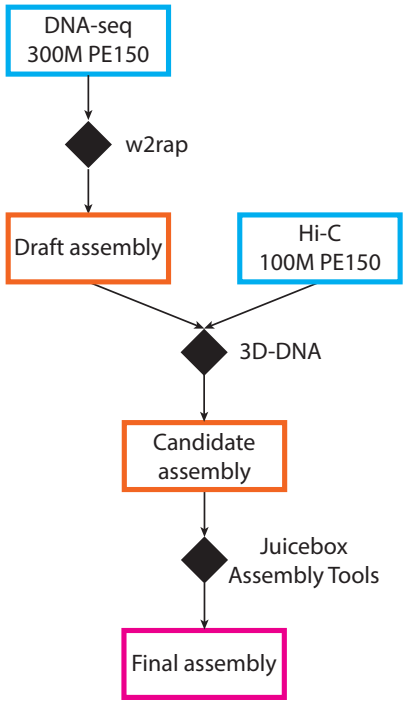
Dudchenko et al., *Science, 2017*

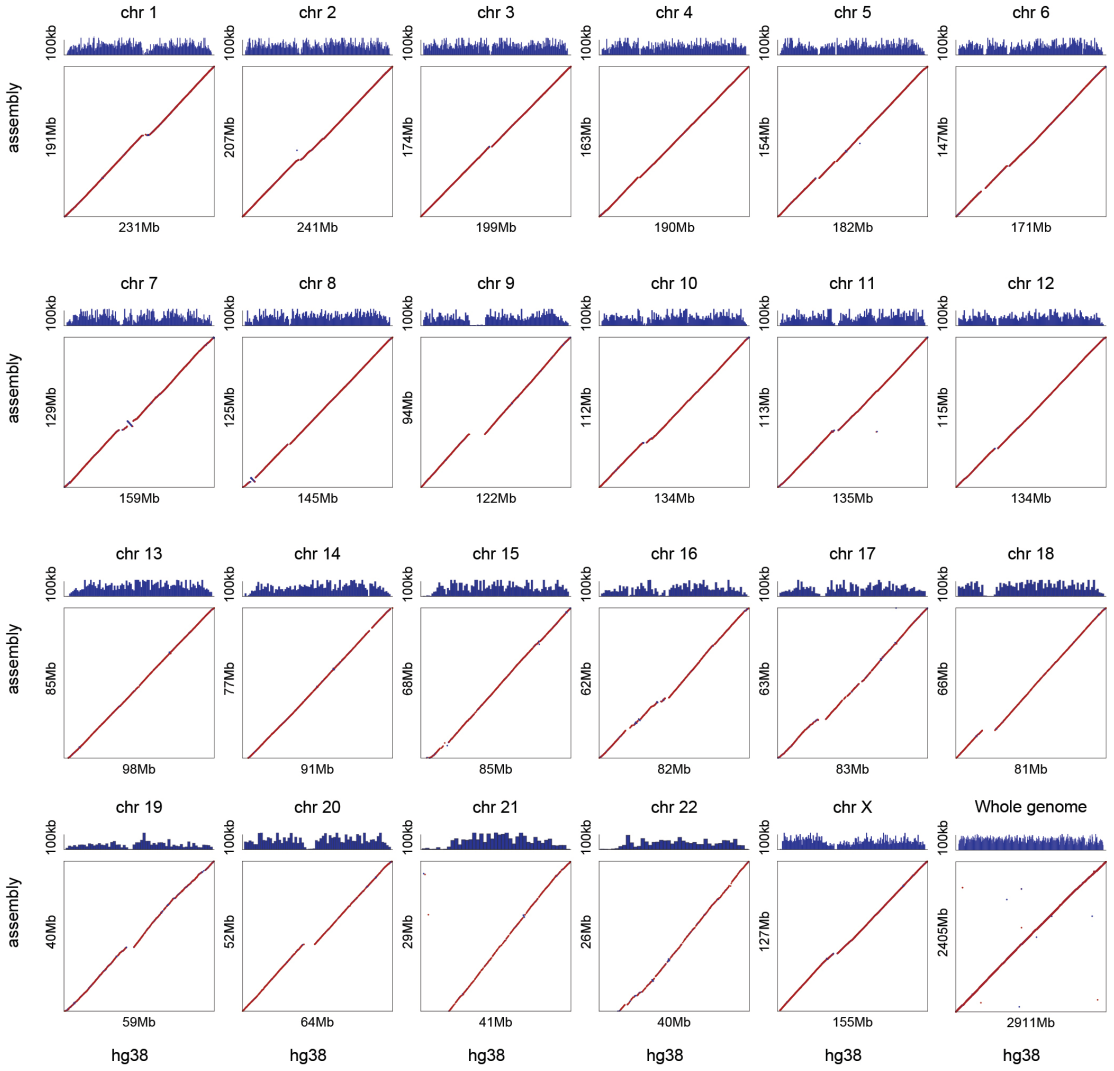
JUICEBOX ASSEMBLY TOOLS



Dudchenko et al., *Biorxiv, 2018*

HERE IS A HUMAN GENOME WE MADE FOR $1000





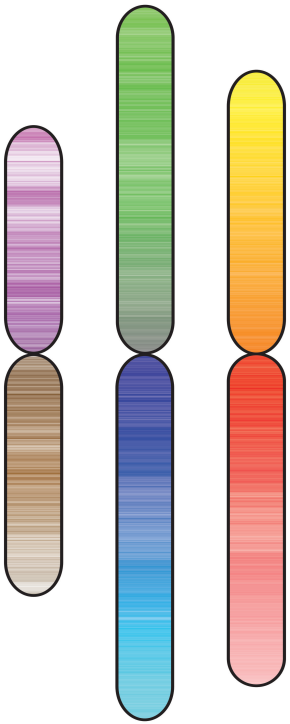
Dudchenko et al., *bioRxiv* 2018

END-TO-END (GAPPY) ASSEMBLY: $1000
PLATINUM ASSEMBLY: ~$20K

AaegL4:

Sanger DNA-Seq + Illumina Hi-C

Contig NG50: 82 Kb



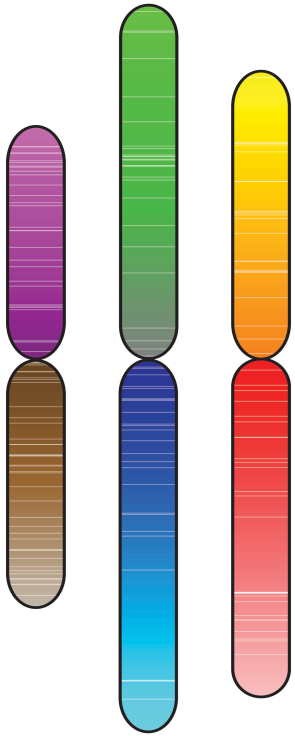
Number of gaps: 25,582

Dudchenko et al., *Science*, 2017

AaegL5:

PacBio DNA-Seq + Illumina Hi-C

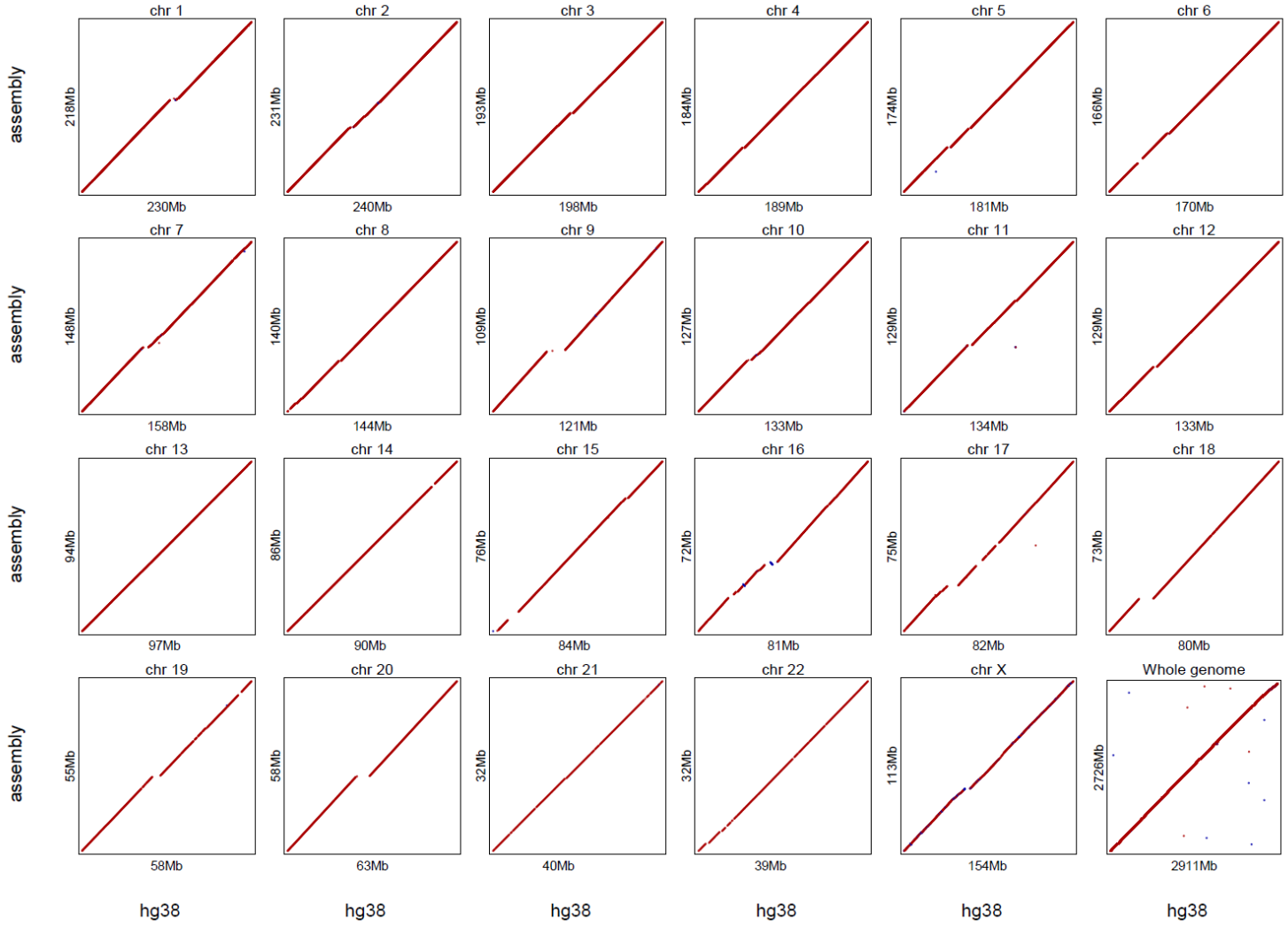
Contig NG50: 11.8 Mb



Number of gaps: 173

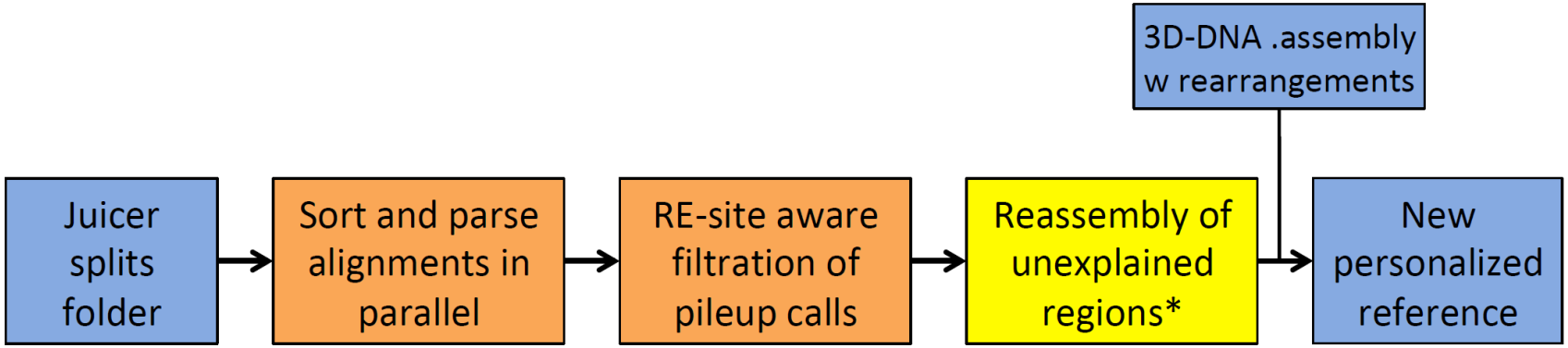
Matthews, Dudchenko, Kingan et al., *Nature*, 2018

OPPORTUNISTIC PROJECT ($0!): ENTEX 002 & ENTEX 003

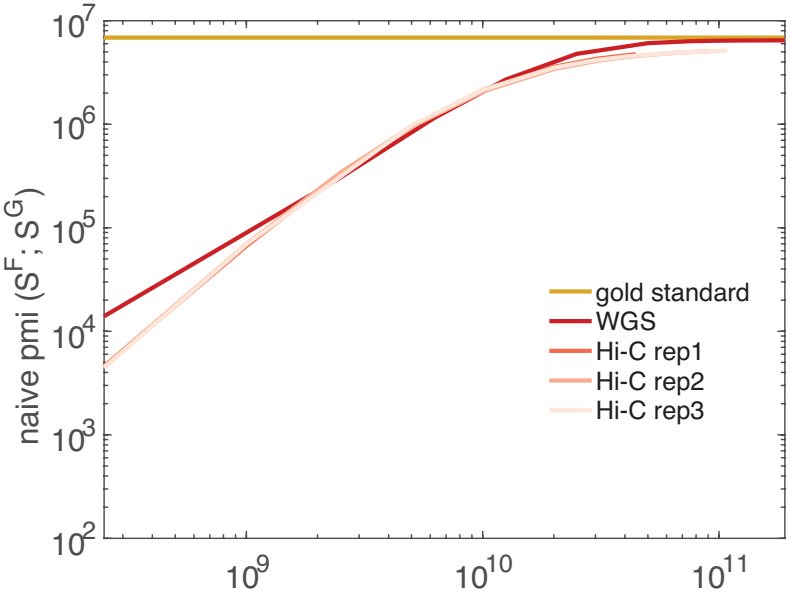


Olga Dudchenko, Jan 2018

OPPORTUNISTIC PROJECT ($0!): HI-C DATA YIELDS GENOME-WIDE SNP CALLING

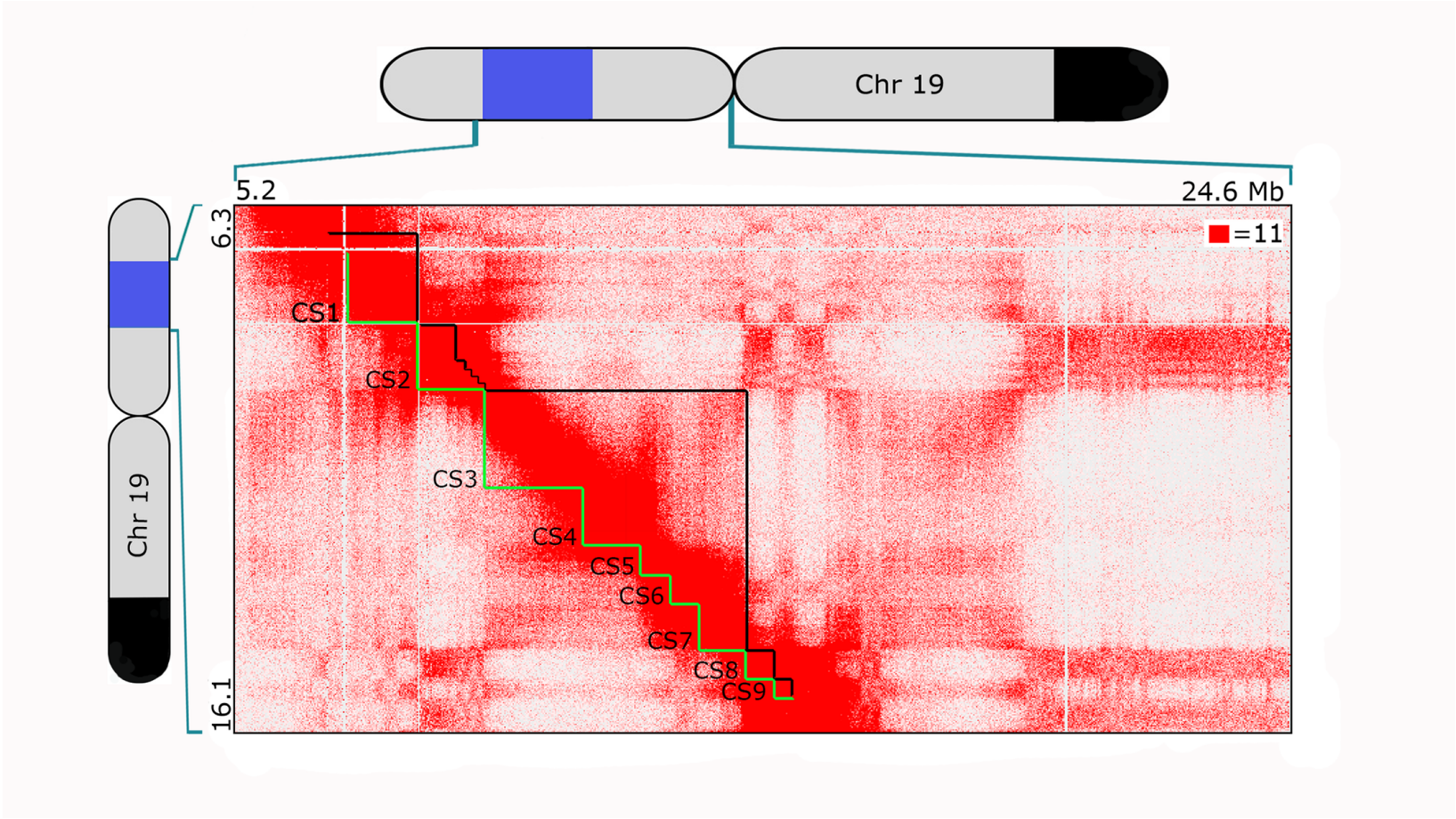


- Available personalized genome modules for Hi-C data include:
  - variant calling
  - phasing
  - assembly



Olga Dudchenko (see also work by Gamze Gursoy, Mark Gerstein)

ENCODE4 PILOT PROJECT ($1000): PGP1 & HEPG2



Data generation in progress

Nir et al., *Nature Genetics* 2018

QUESTIONS FOR DISCUSSION

- What is the value of the personal genome
  & having functional genomic assays matched to it?

- How to think about genomic element calls
  relative to substantial epigenetic variation across individuals?

- How best to fill out the EN-TEx data matrix in the coming months?
  - What types of integrated calculations
    are best done over the EN-TEx matrix?

- How to best include opportunistic Hi-C personalized genome data into ENCODE?

- Should we prioritize additional reference cell lines & samples for *de novo* genome assembly efforts?
  - Backfilling? (K562, IMR90, HCT-116, HMEC, NHEK etc.)
  - EN-TEx 1&4?