HUMAN LONGEVITY, INC.

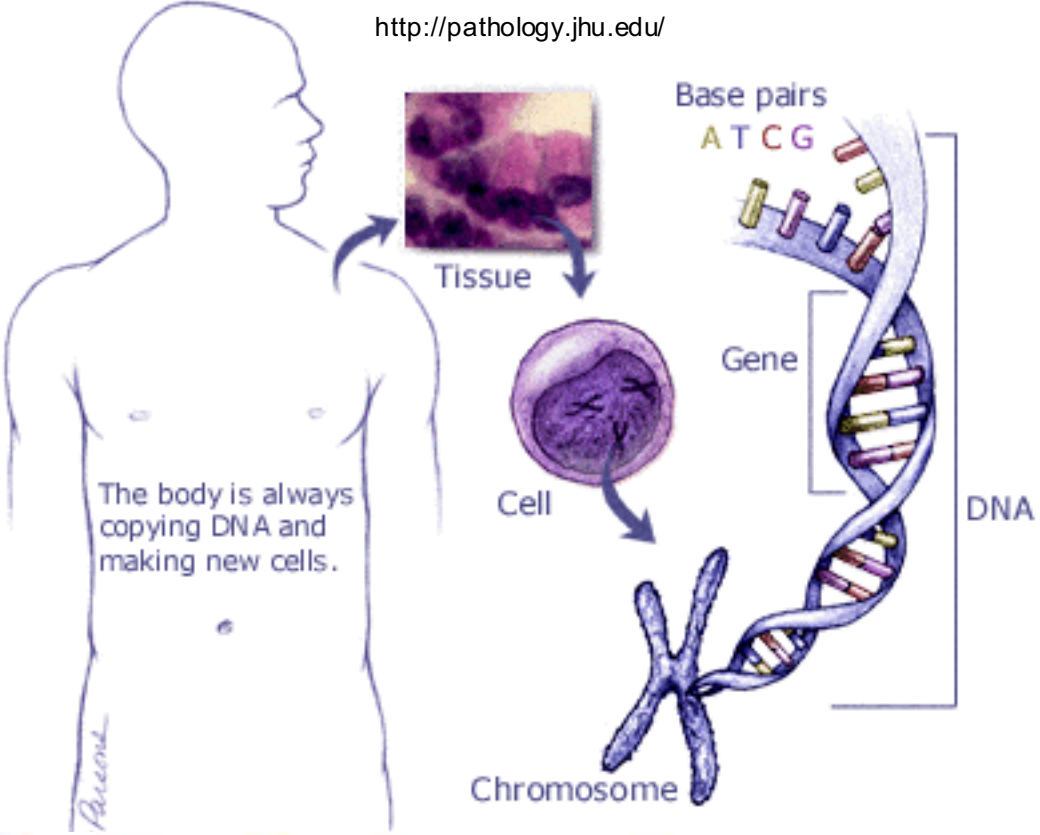
What we learn from "clinically"-deep 10,000 genomes?

Ahmed Moustafa

Human Longevity, Inc. (HLI)
American University in Cairo (AUC)

BioVision Alexandria 2016
Bibliotheca Alexandrina

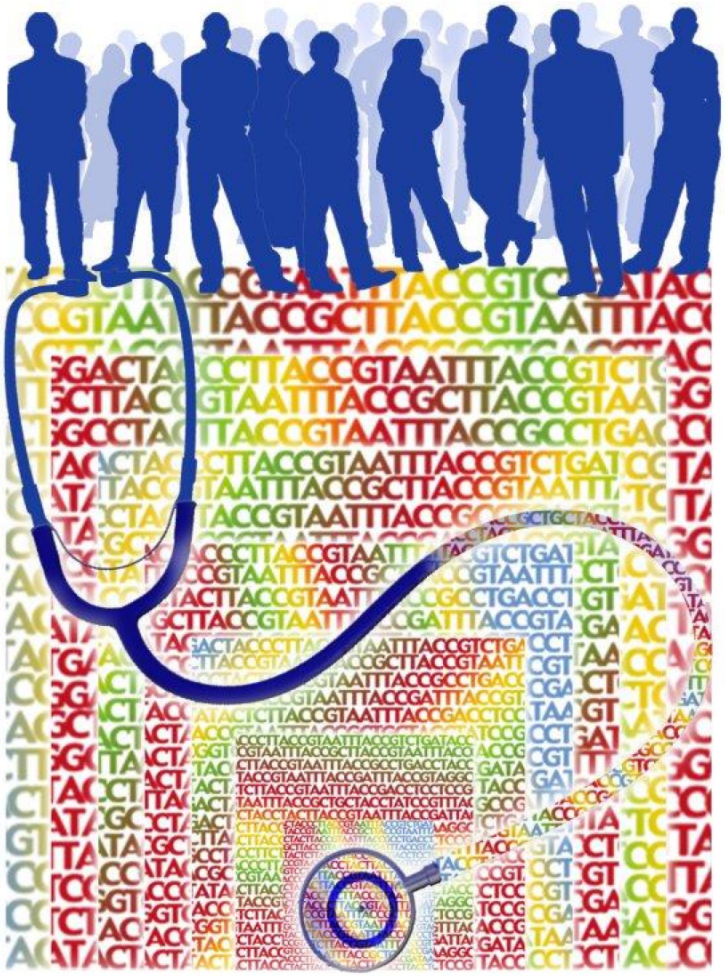
http://pathology.jhu.edu/



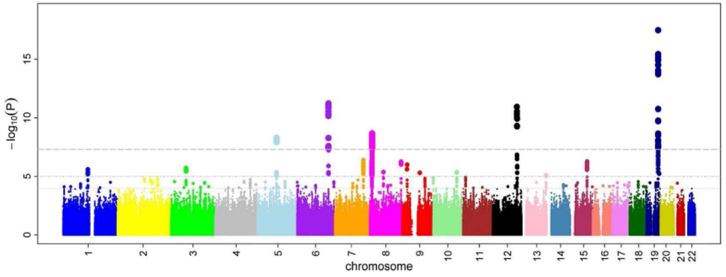
https://www.genome.gov/

To identify such genetic changes (mutations, variants, polymorphism) that affect our "**phenotype**", we need to "**genotype**".

To associate **w/ statistical power** the genotype to the phenotype, we need to genotype "**populations**".



https://www.genome.gov/



https://en.wikipedia.org/wiki/Manhattan_plot

GWAS standard:
p-value < 5e-8

Genotyping Approaches

Illumina
HumanOmni

- **SNP Array** – with ~2.5 million markers (e.g., Illumina) covers only < 0.01% of the human genome. SNP arrays are designed based on known markers from published specific population studies → SNP bias
- **Whole-Exome Sequencing (WES)** – covers only the coding component of the genome (~2%). However, most disease associated SNPs from genome-wide association studies (GWAS) are non-coding
- **Whole-Genome Sequencing (WGS)**

Public Genome Projects

Study	No.	Average Coverage	No. Deep Coverage
1000 Genomes 2015	2,504	7.4x	453
Japan Genome 2015	1,070	32.4x	1,070
Iceland Genome 2015	3,545	20x	909
UK Genome 2015	3,781	7x	0
African Genome 2015	320	4x	0
Sardinian Genome 2015	2,120	4x	0
Netherland Genome 2014	750	13x	0
Total	**14,090**		**2,432**

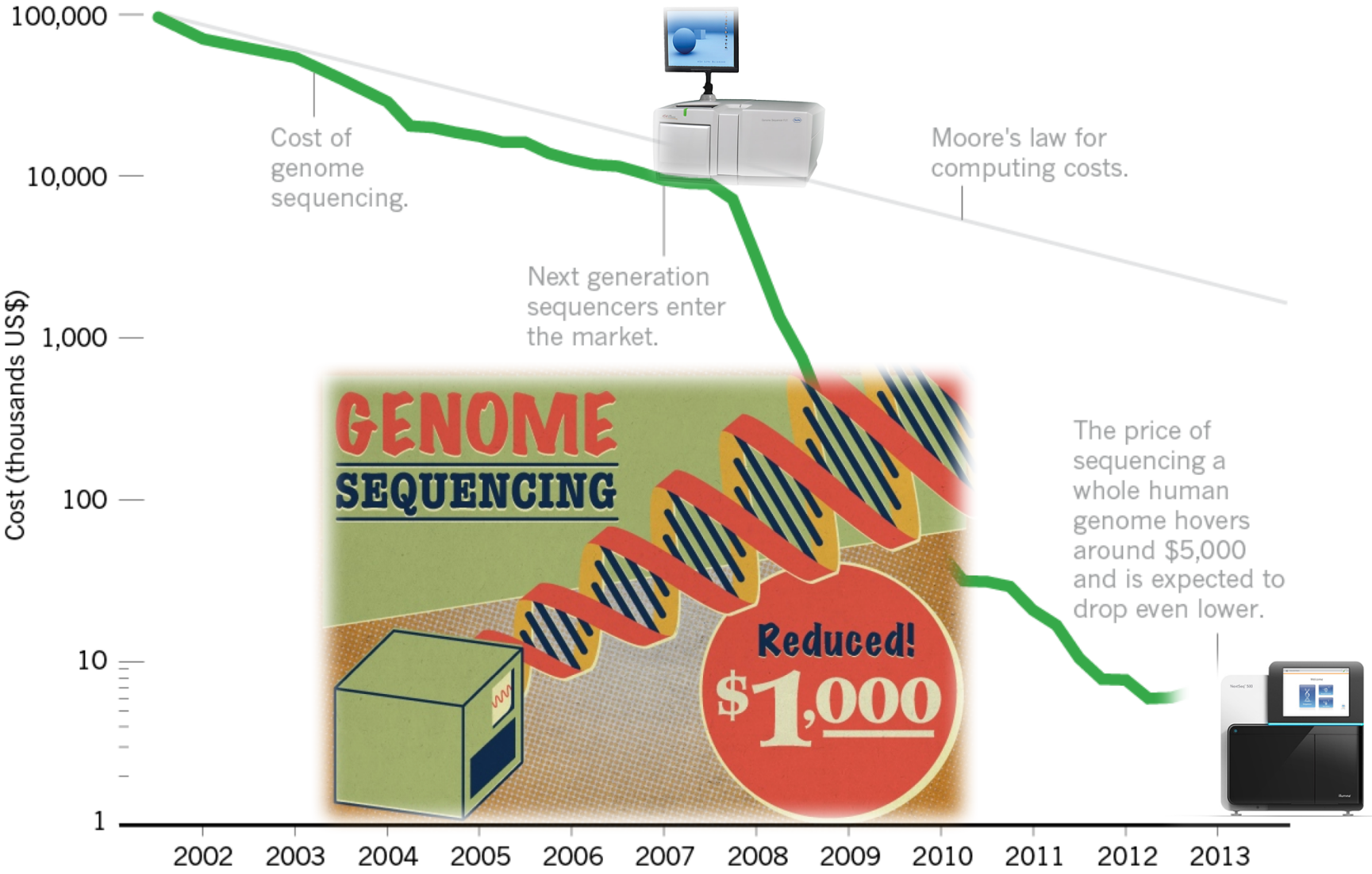
The Clinical Genome

- In Depth Coverage
- In Breadth Coverage
- Quality and Reproducibility

http://dx.doi.org/10.1038/507294a

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.



Workflow

HiSeq X Ten

72hrs

Sequencing (40 flowcells)

4 Packaging (9 TB/day)

4 Upload (650 GB/sample)

amazon web services™

12 BCL to FASTQ

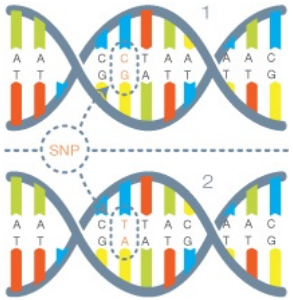
25hrs

FASTQ to BAM (aligner)

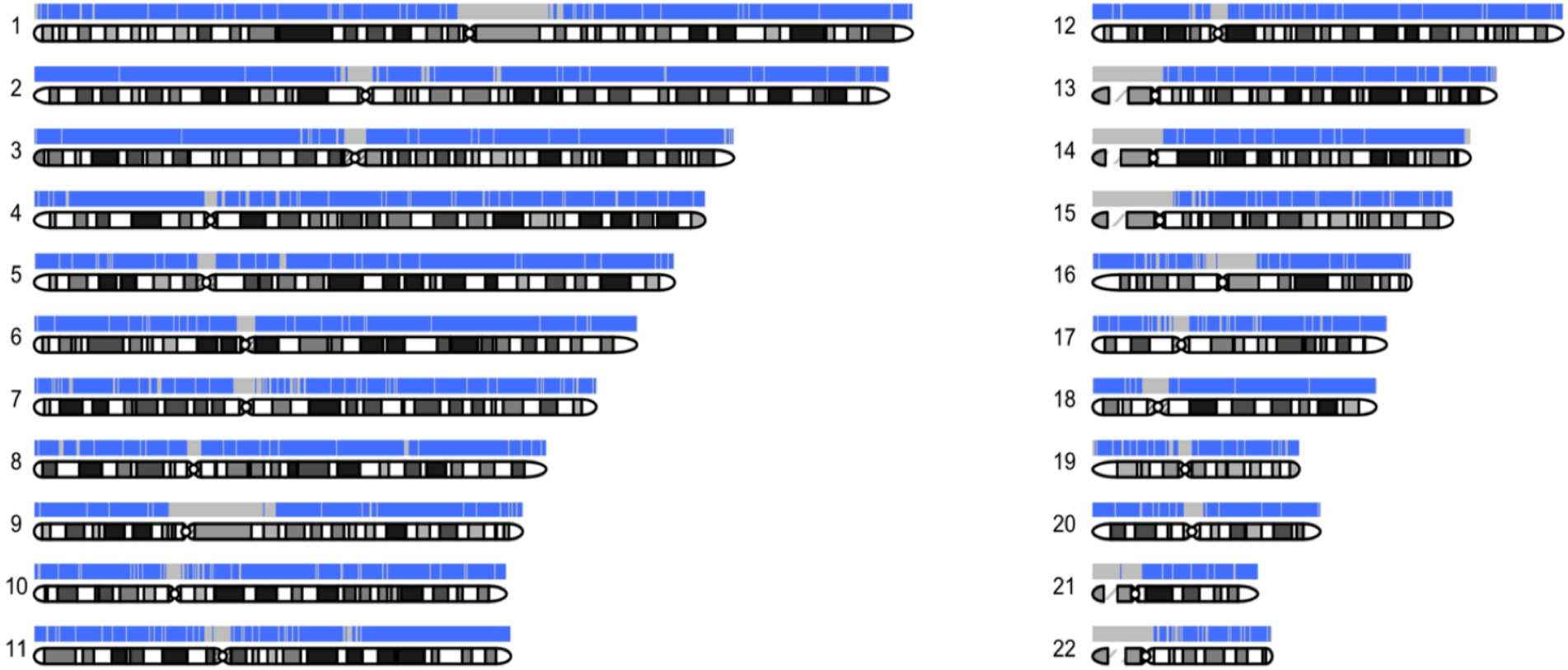
8 BAM to (g)VCF (variant caller)

Permanent storage (S3)

Downstream analysis



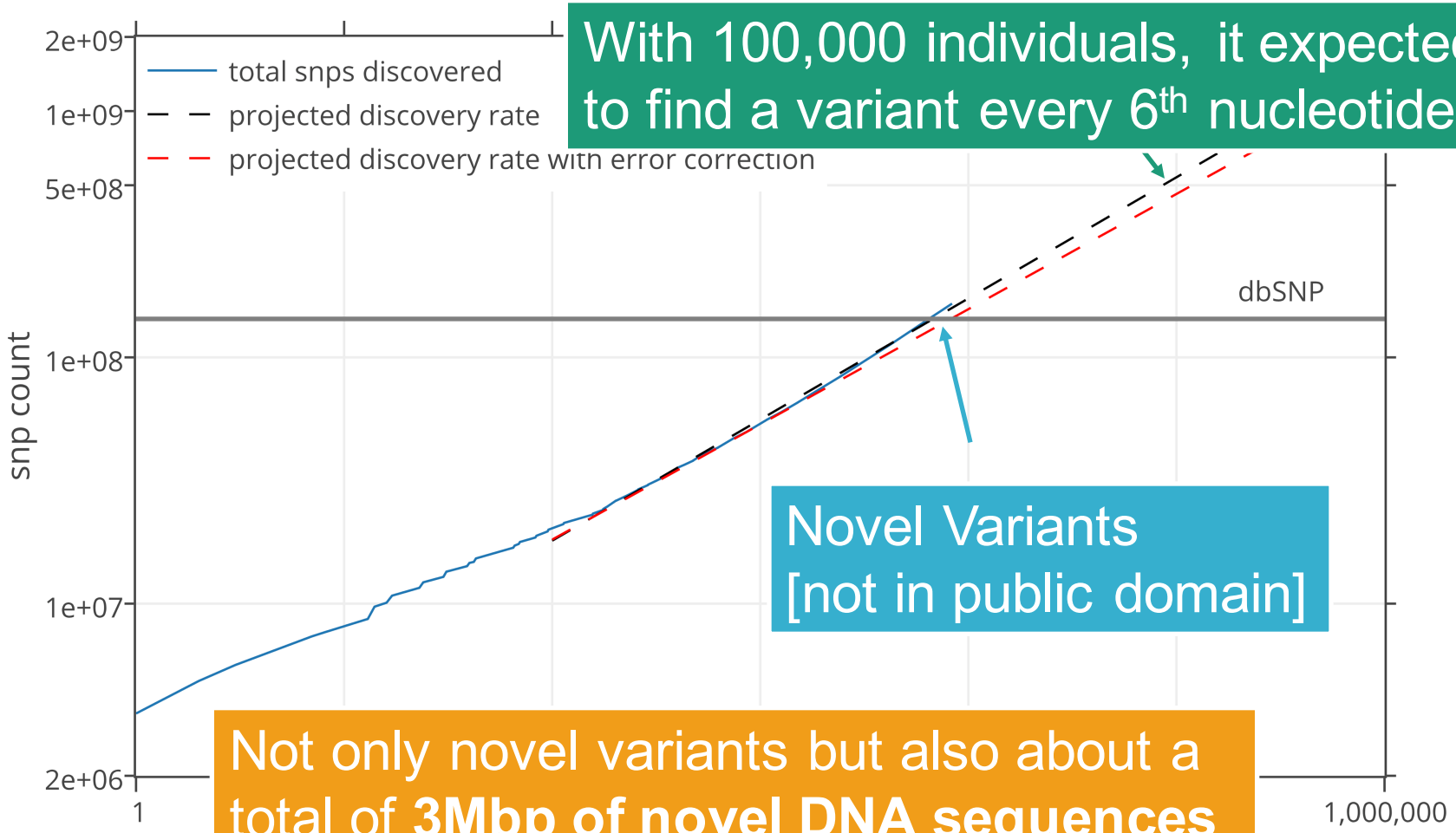
Overall Quality Metrics



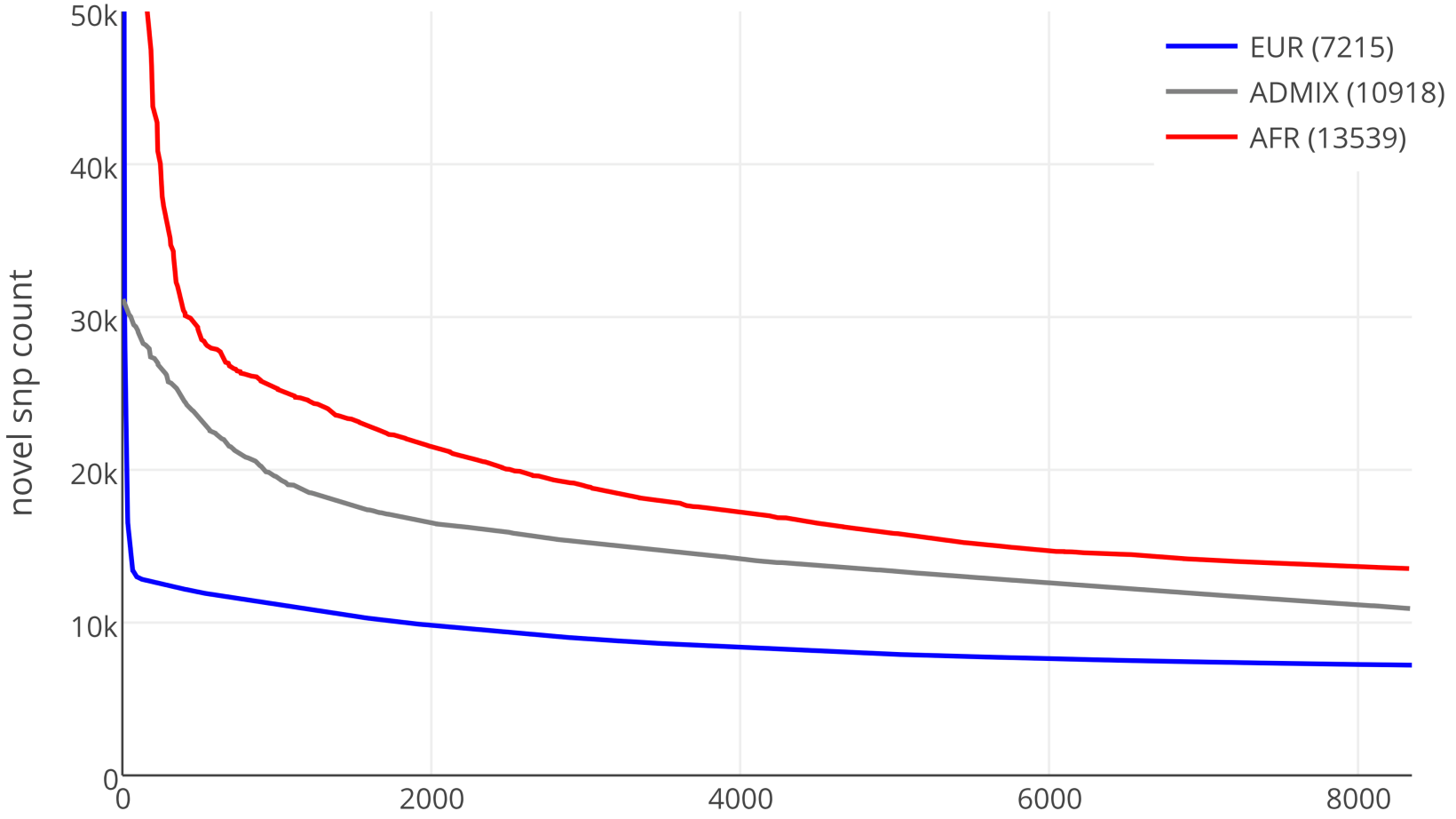
30x-40x Depth | 85% Coverage (Blue) |
92% of Exome | 96% of Clinical Variants

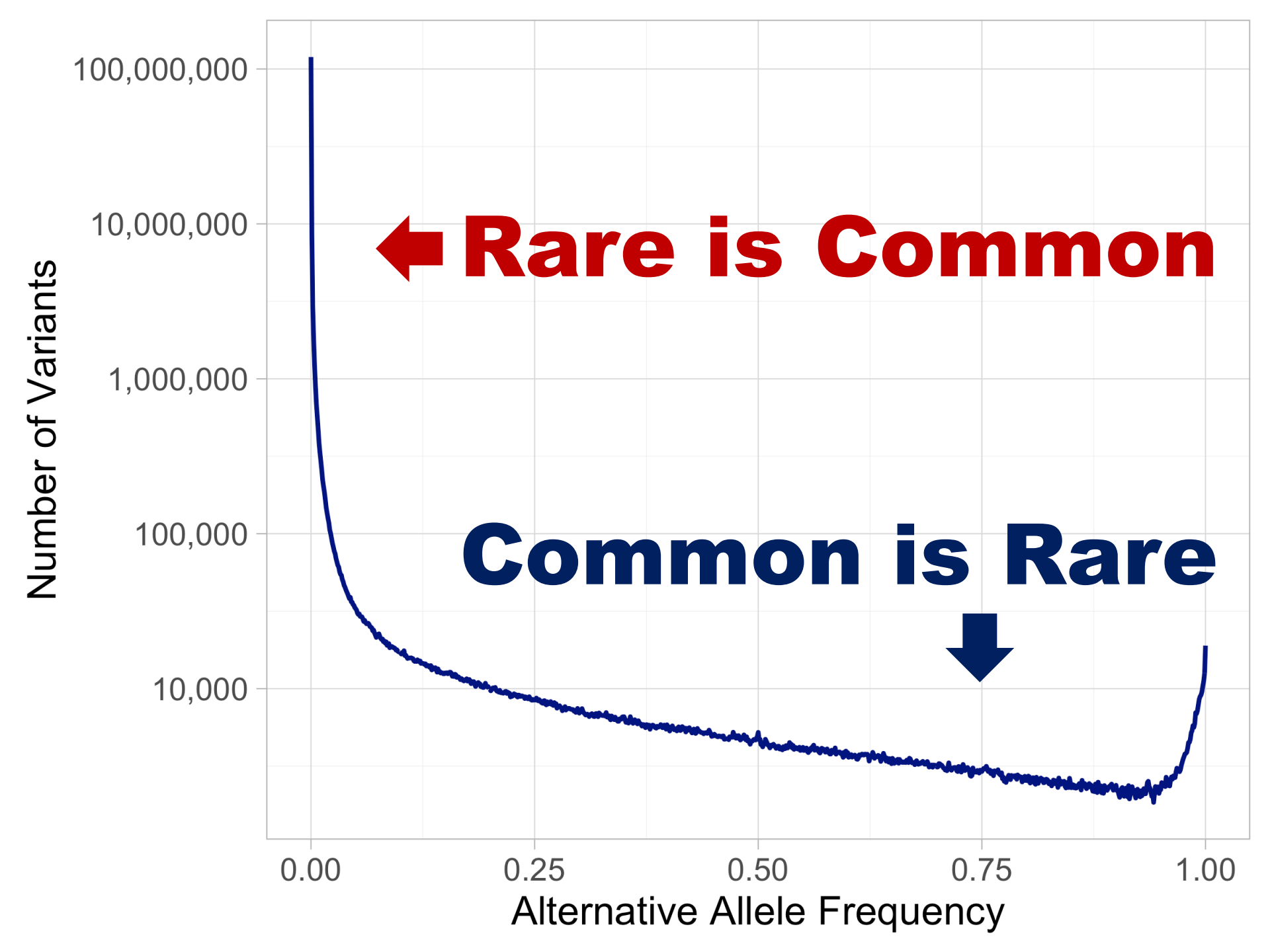

Whole-Genome Sequencing
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
WGS → 85%
Whole-Exome Sequencing
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
WES → 2%

150 Million Variants

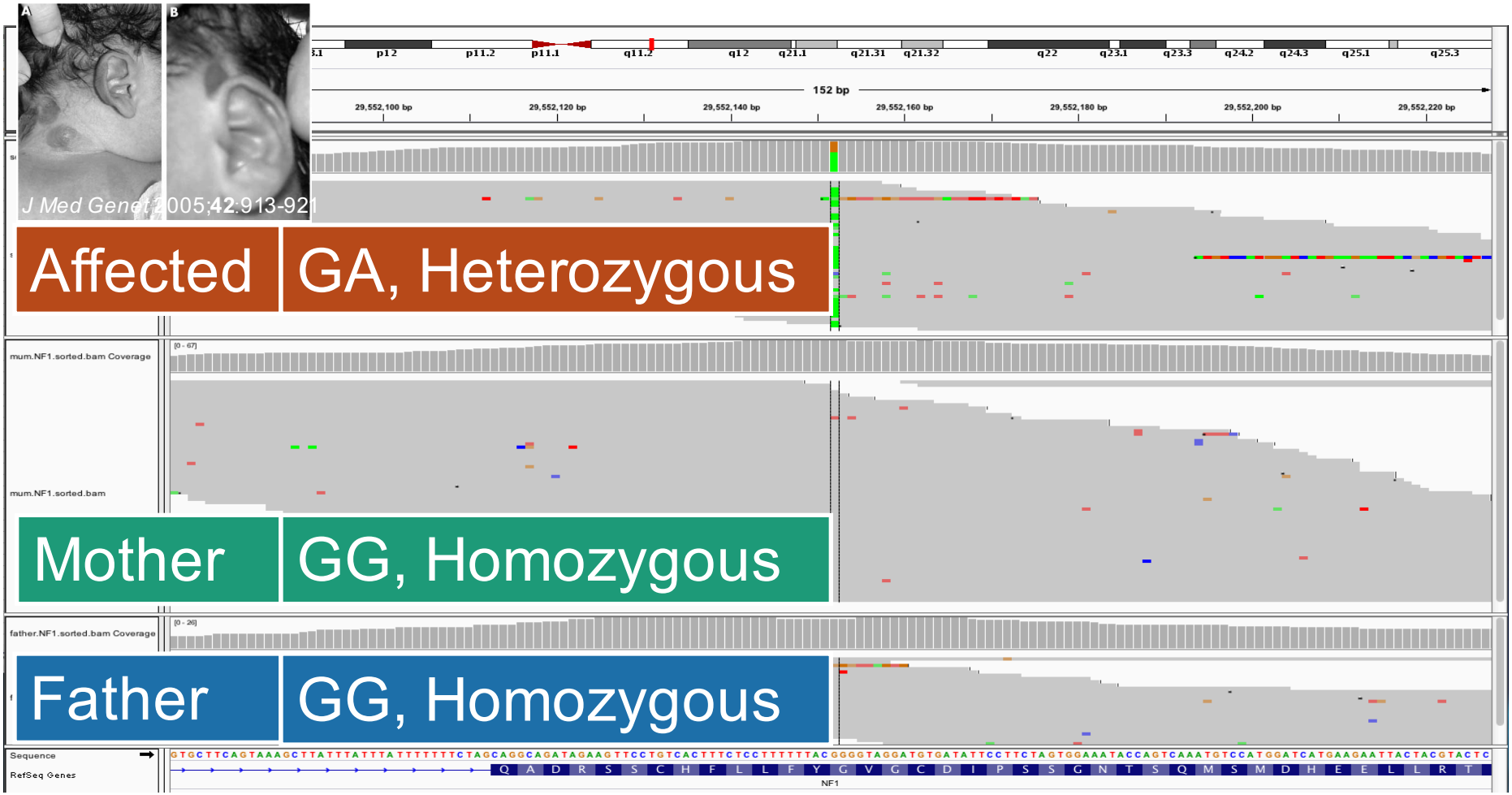


Novel Variants by Individual




100,000,000
10,000,000
1,000,000
100,000
10,000
Number of Variants
Rare is Common
Common is Rare
0.00
0.25
0.50
0.75
1.00
Alternative Allele Frequency

De novo mutation in *NF1*



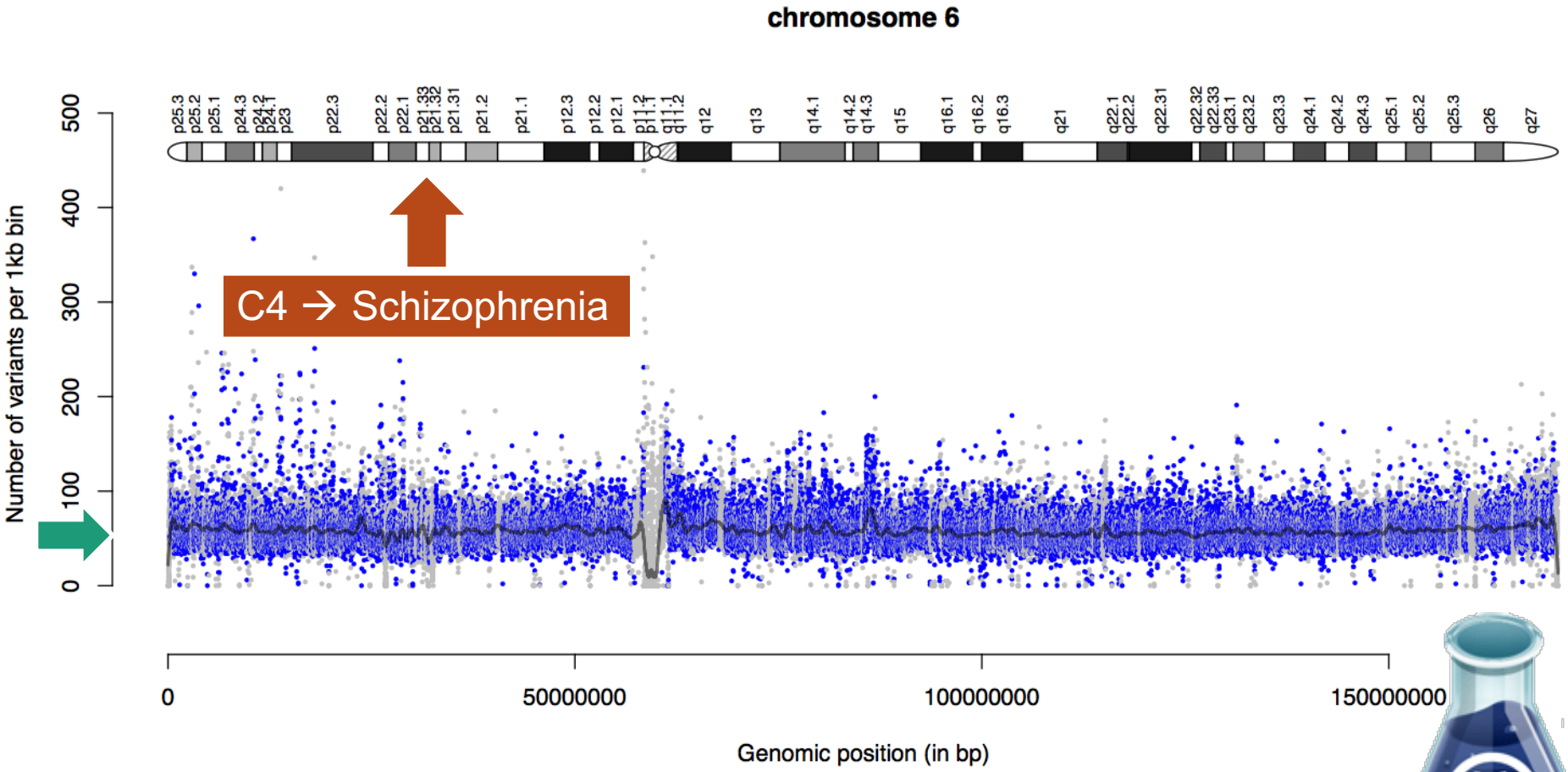
NF1 → neurofibromin (tumor suppressor)

Delleman syndrome (congenital)

Mutation from G → A, only in the son, introduces a splice acceptor leading to skipping the beginning of an exon.

Rates of Variation

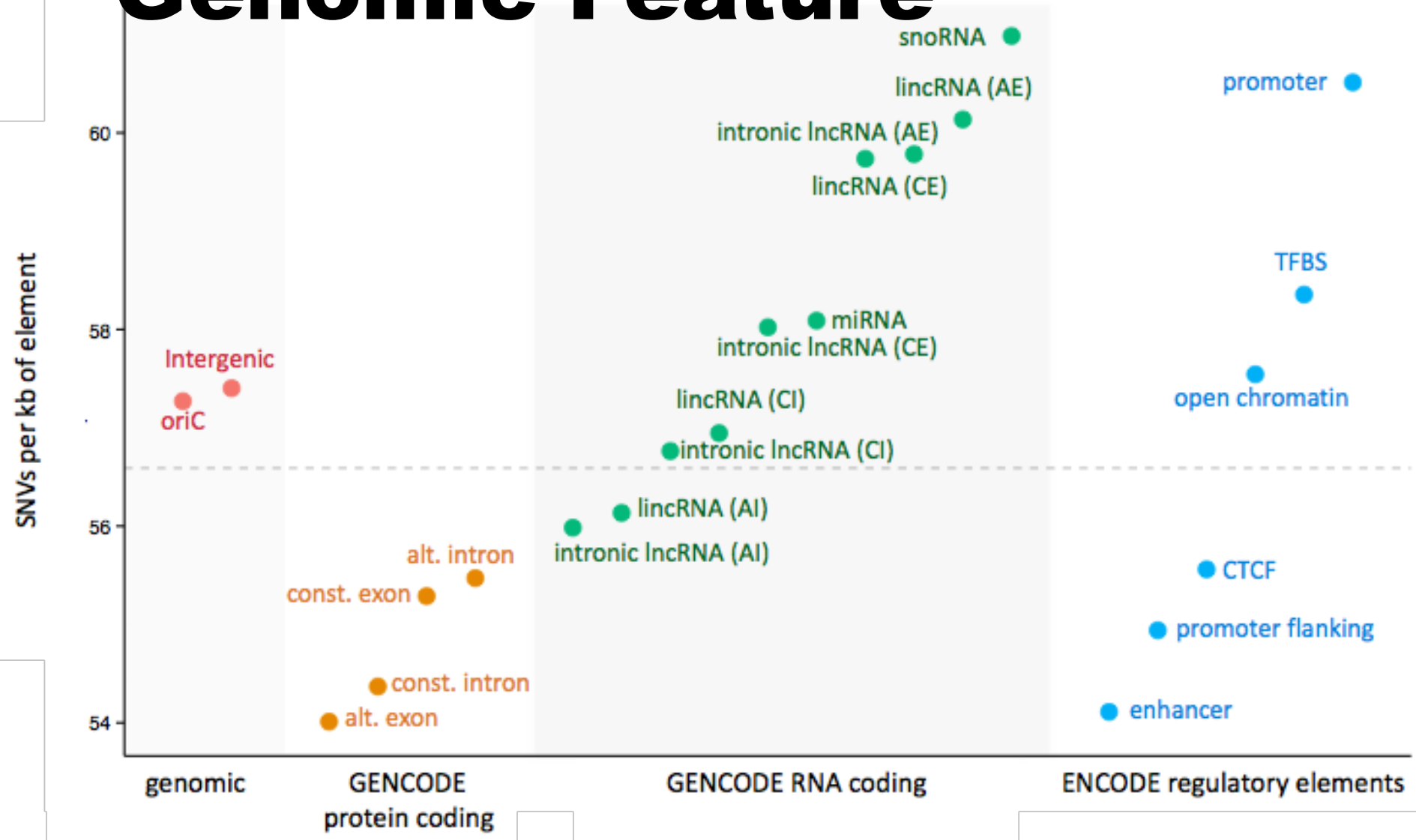
Rates of Variation in 10,000 Genomes



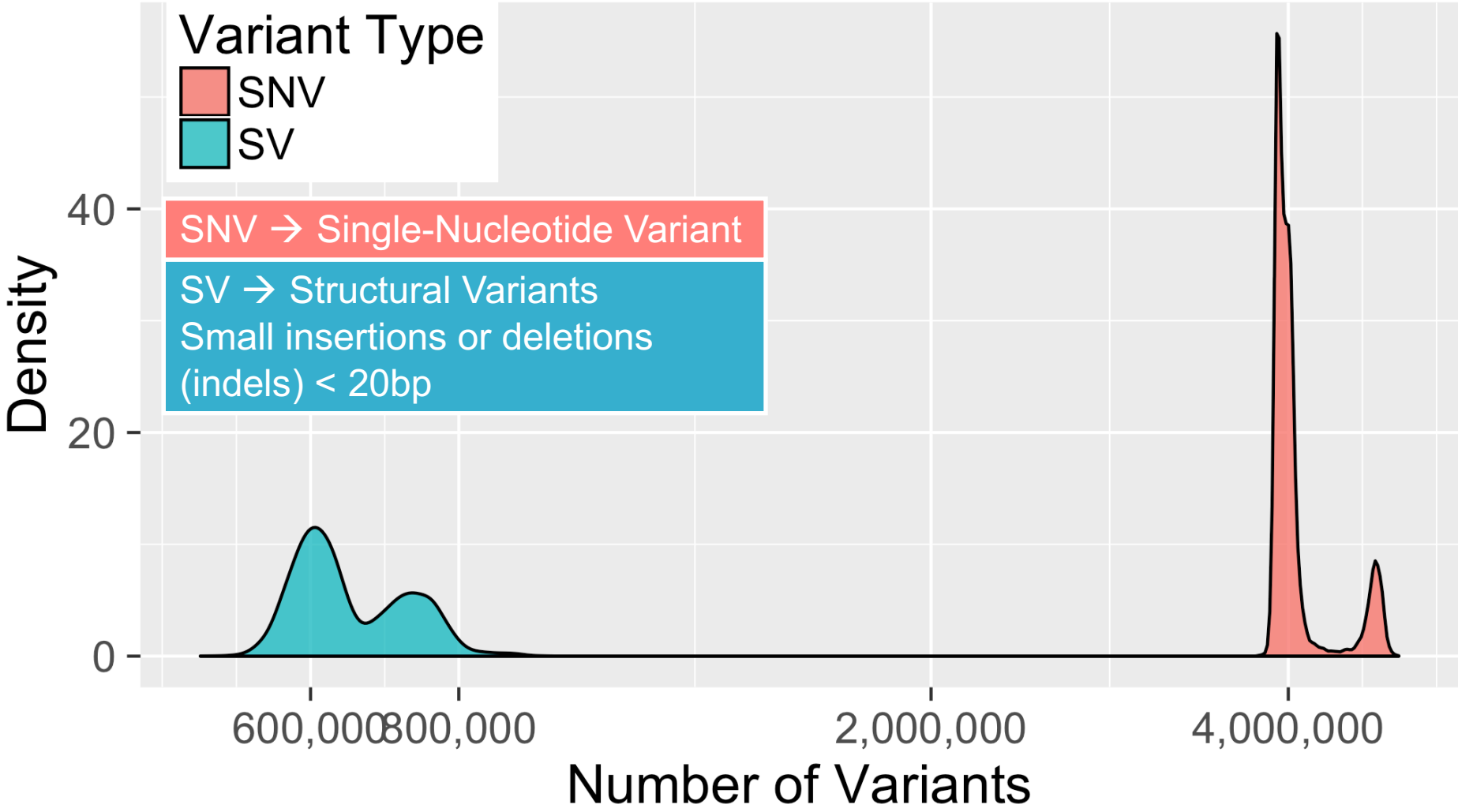
Blue → Overlap w/ Genome-in-a-Bottle (GIAB) high-confidence
Grey → Outside of GIAB

https://sites.stanford.edu/abms/giab

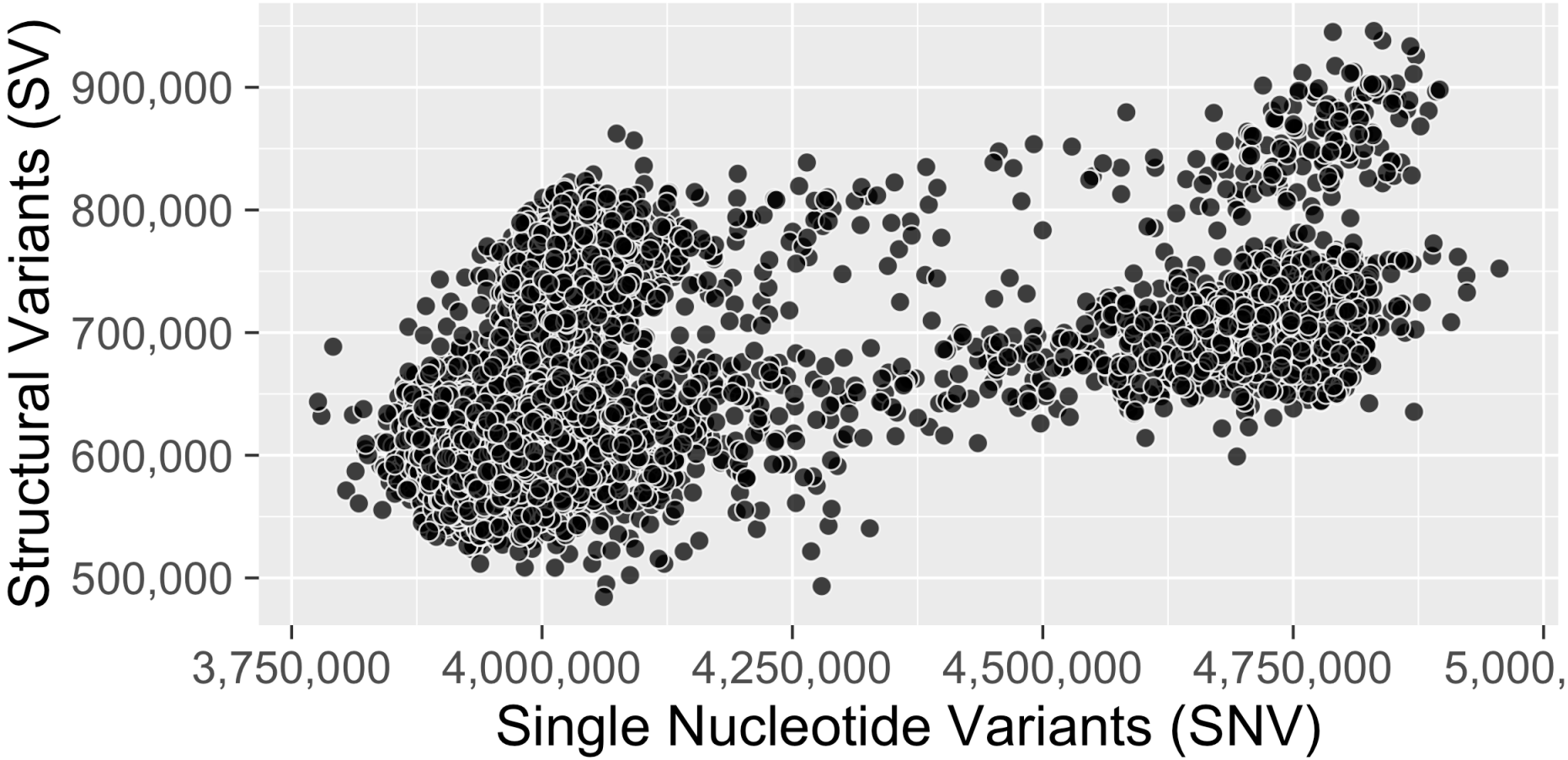
Rates of Variation by Genomic Feature



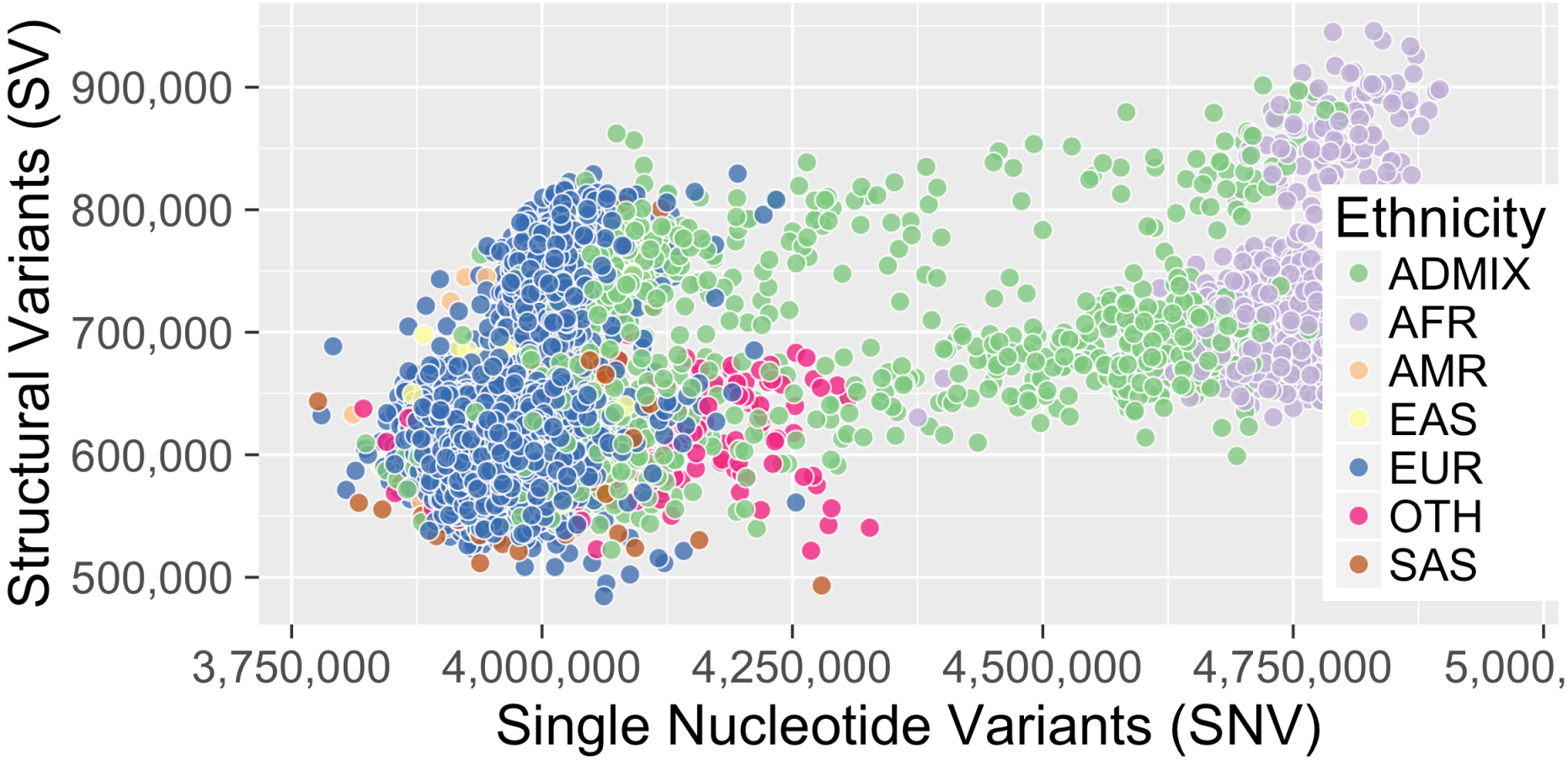
Variants per Individual



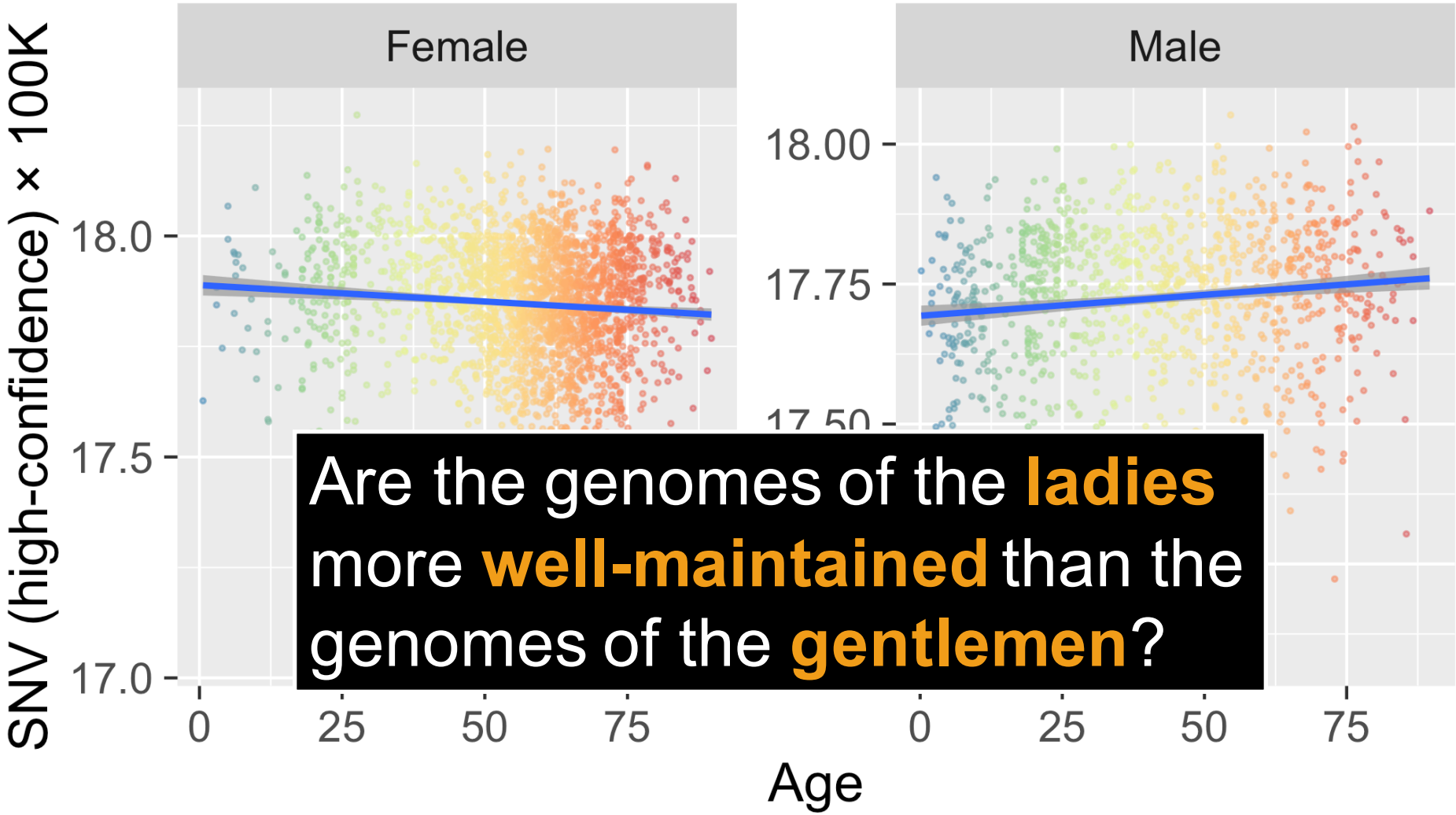
Variants by Ancestry



Variants by Ancestry

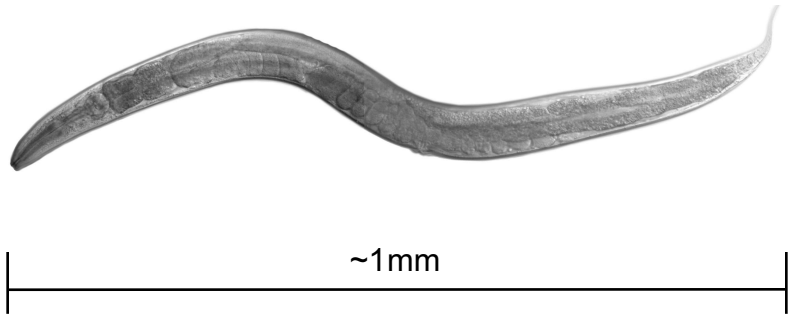


Variants by Gender EUR



The 3D [Non-Coding] Genome

Gene Number ≠ Organismal Complexity



Protein-coding genes ~ 19,000	Protein-coding genes ~ 19,000

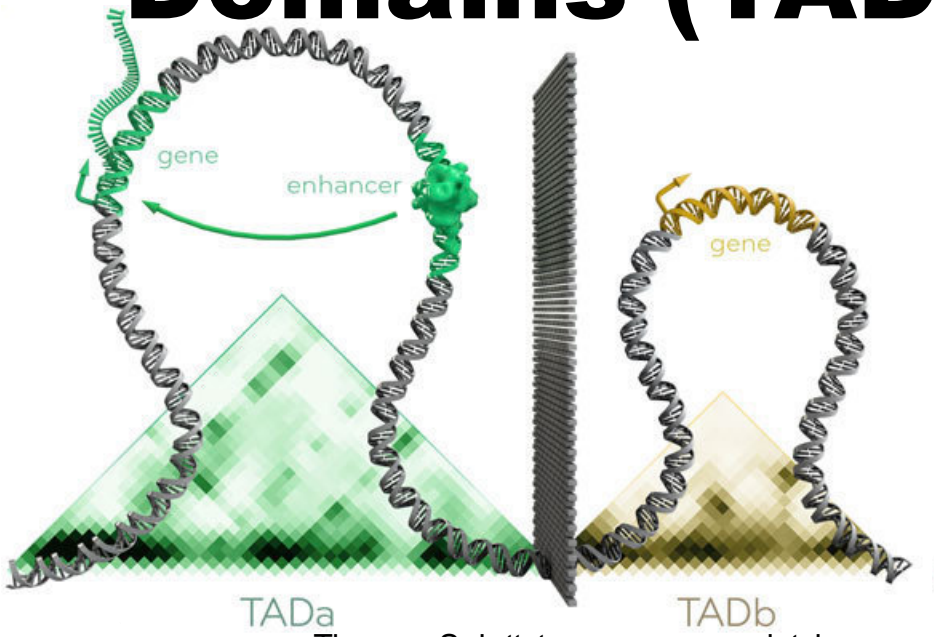
Where does the complexity come from?

“Gene regulation”

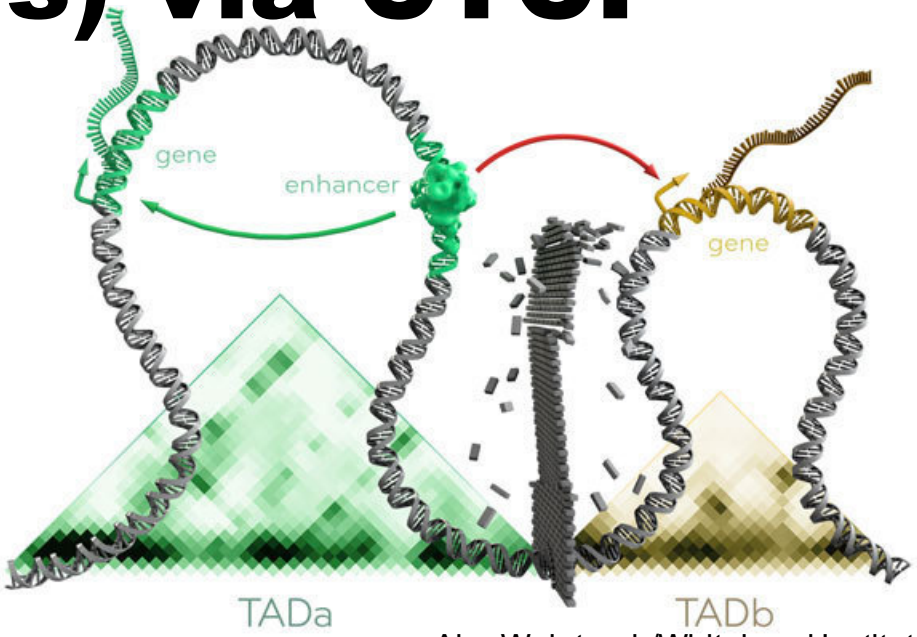
by region (e.g., breast versus kidney)
in response to environmental signals
in development (e.g., embryo versus adult)

28 weeks	1 hour	8 months	3 years	4 years	5 years

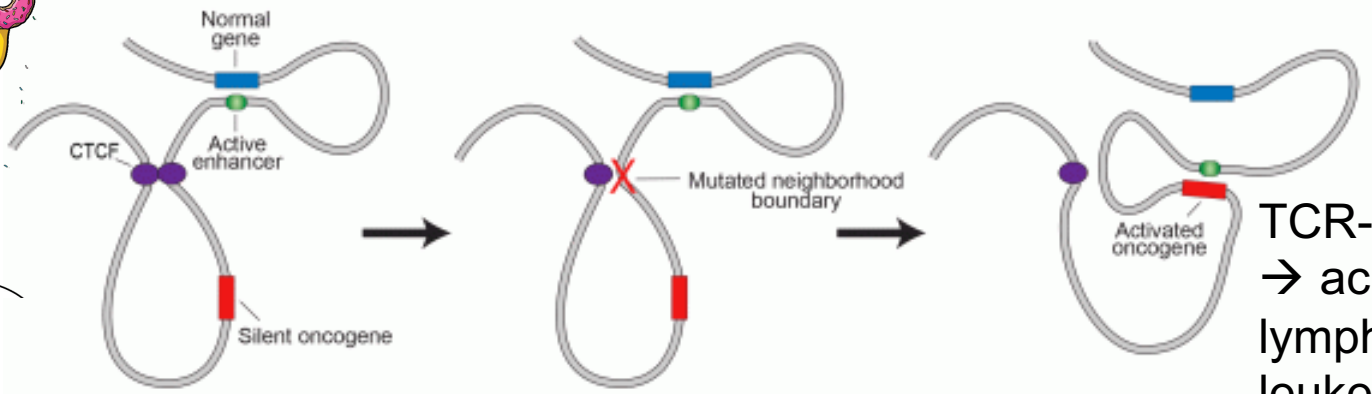
Topologically Associating Domains (TADs) via CTCF



Thomas Splettstoesser, www.scistyle.com

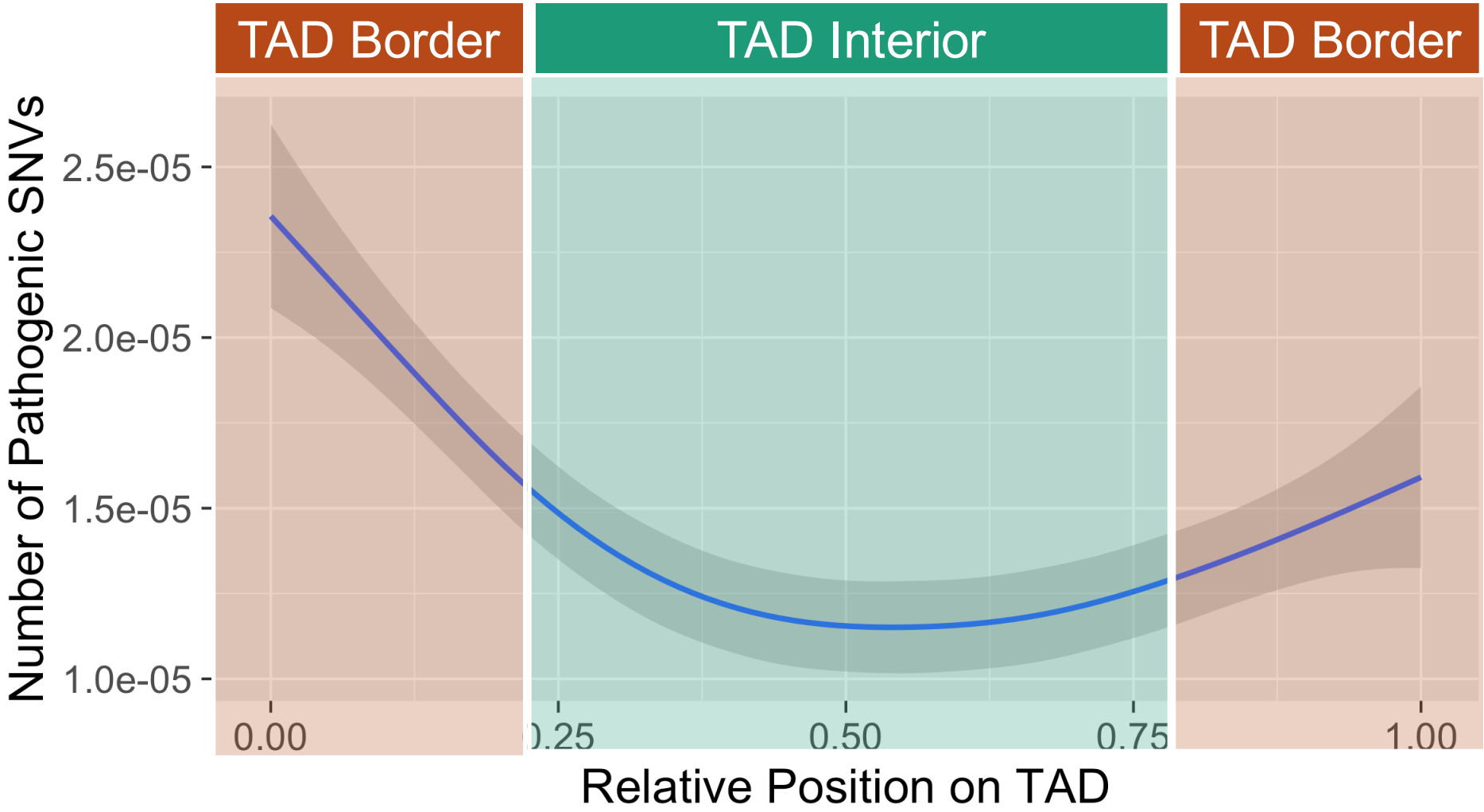


Abe Weintraub/Whitehead Institute

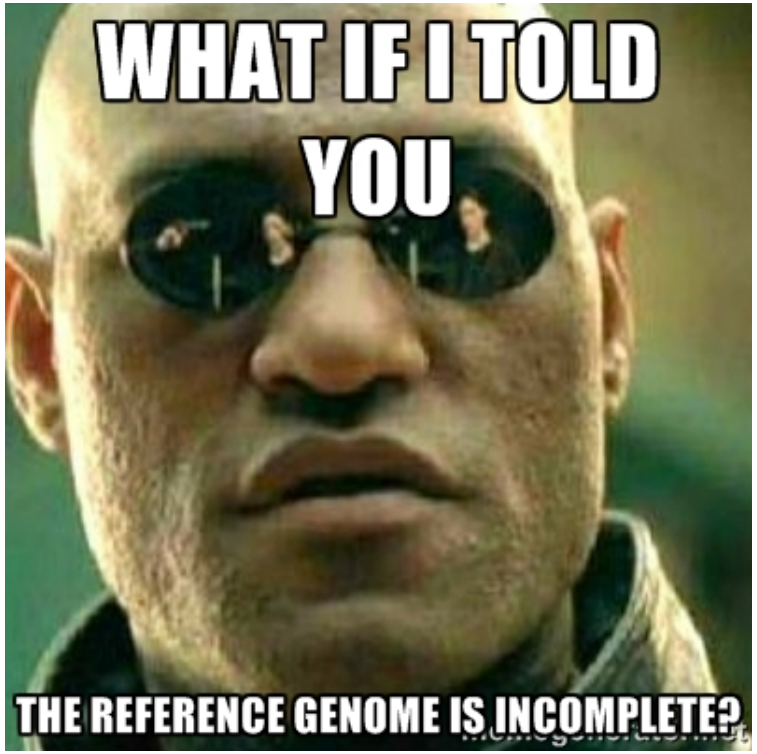


TCR-*LMO2*
→ acute lymphoblastic leukemia (ALL)

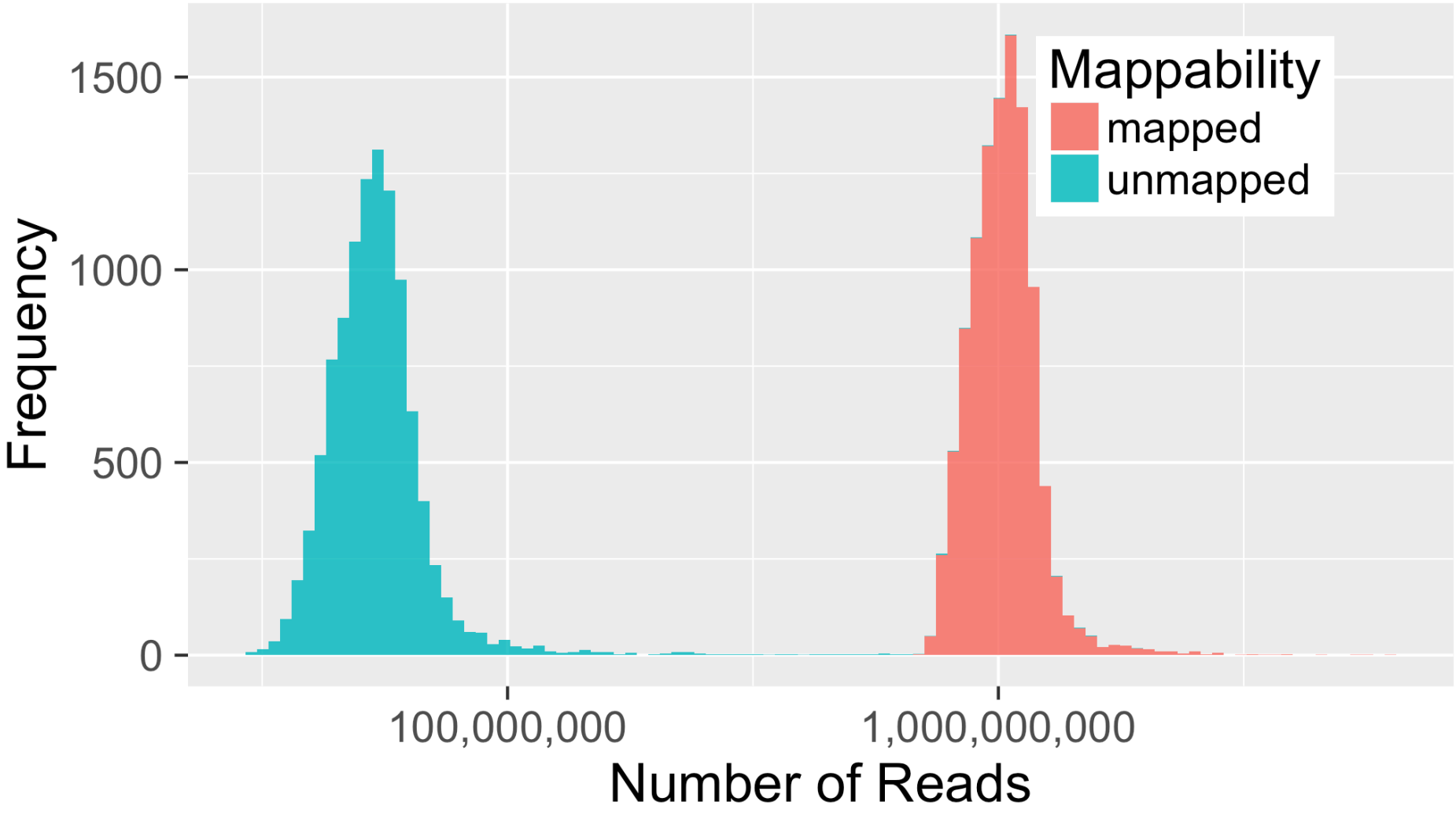
Pathogenicity over TADs



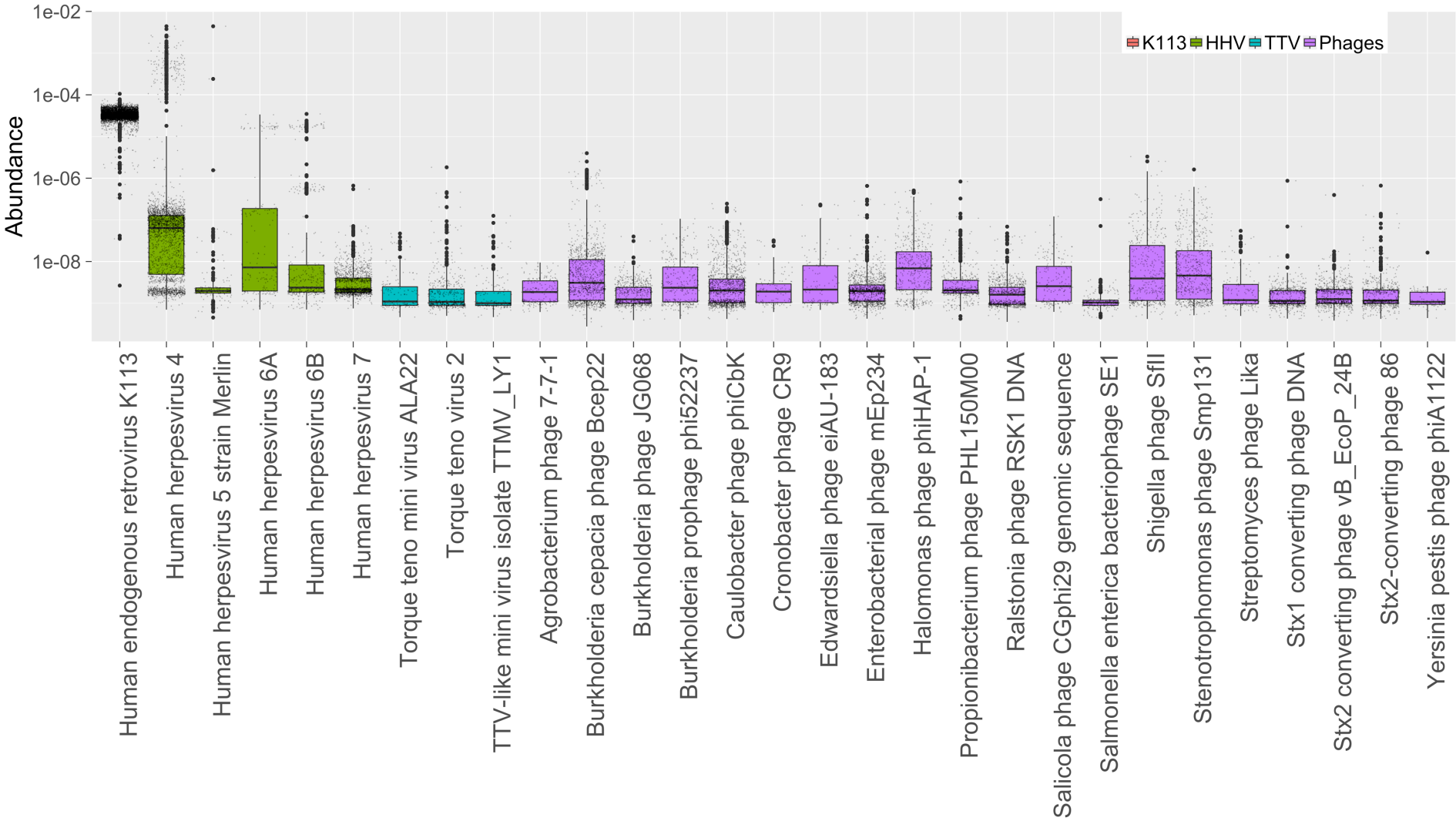
The Unmapped Genome



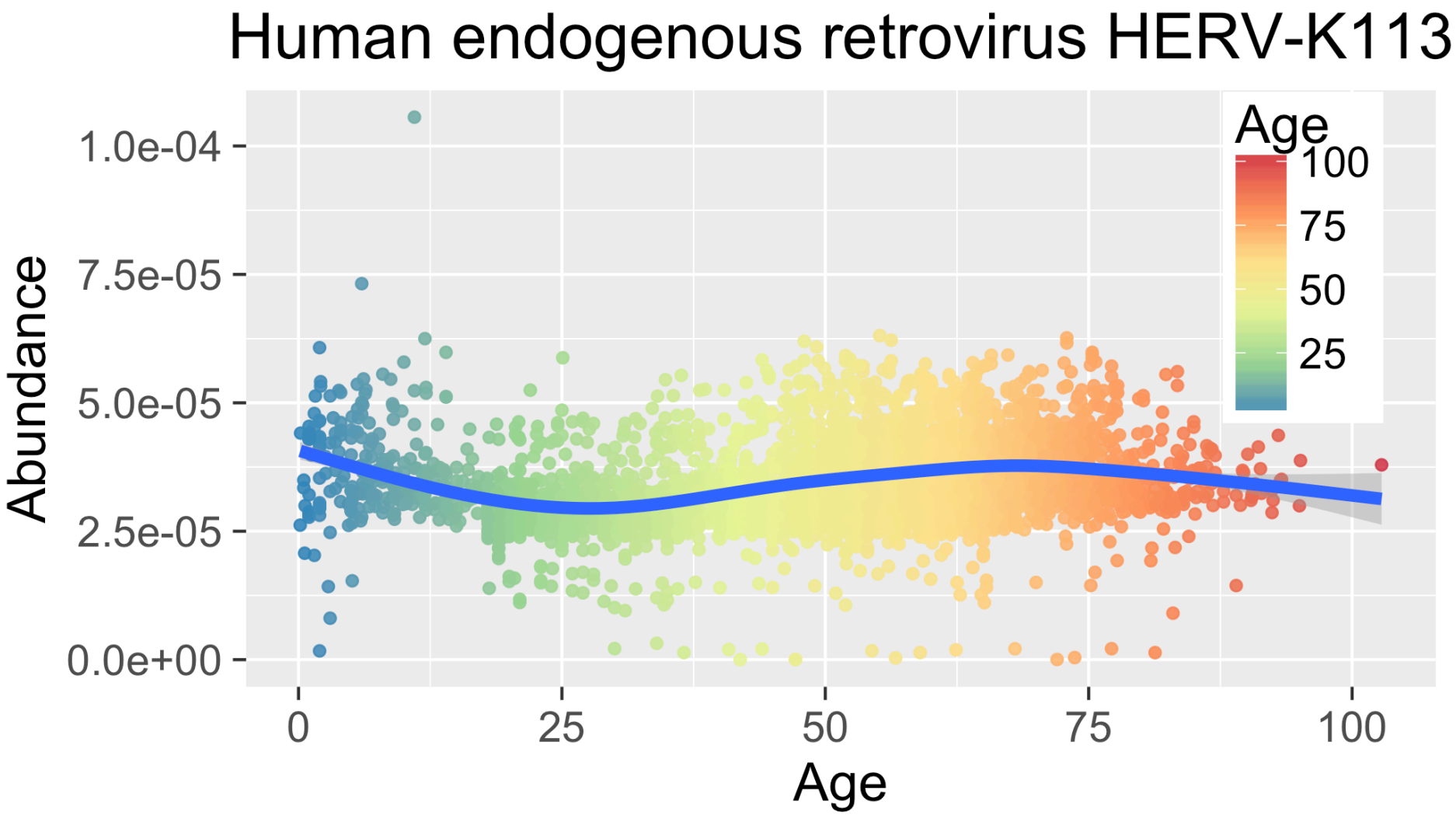
Mapped & Unmapped (per Individual)



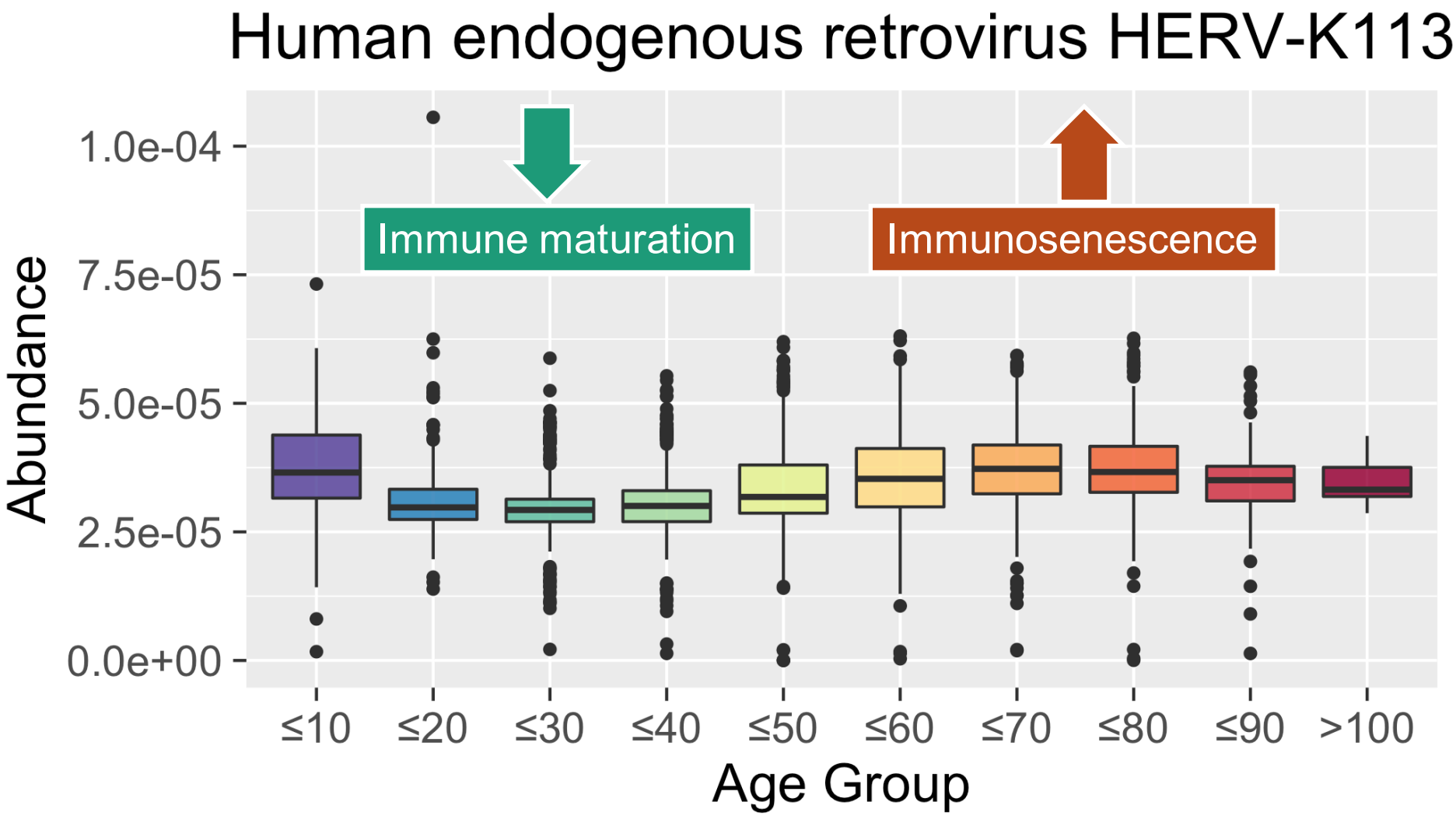
Viral Load in Unmapped



Viral Load in Unmapped



Viral Load in Unmapped



Summary

- Generated first 10,000 deep coverage genomes with clinical standards (coverage, quality, reproducibility)
- Identified 150 million SNVs
- Identified an individual contribution of an average of 8,000 novel variants

Summary cont'd

- Formulated high-resolution profiles in coding sites with tolerance score
- Identified association between pathogenicity of SNVs and and TADs
- Detected dynamics in abundance of Human Endogenous Retrovirus K113 associated with age


Ischaemic heart disease?
Stroke?
Hypertension ?
Egyptian Genome!
Cirrhosis?
Cancer?
Kidney Disease?
AFP