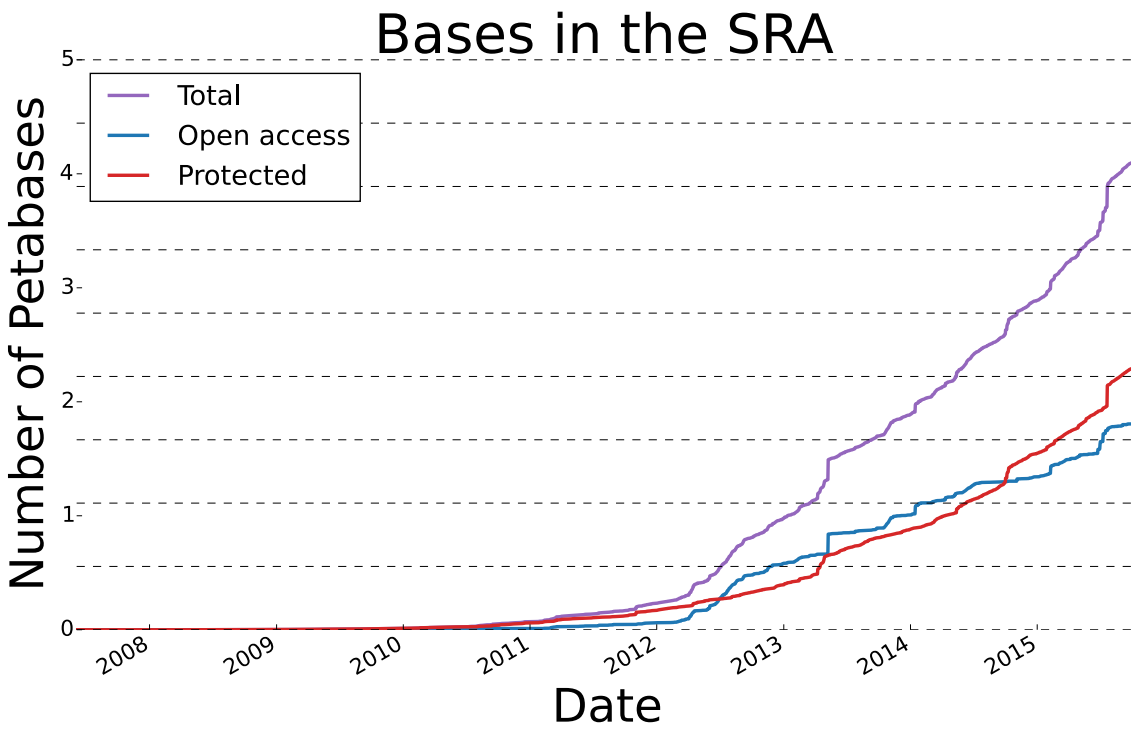
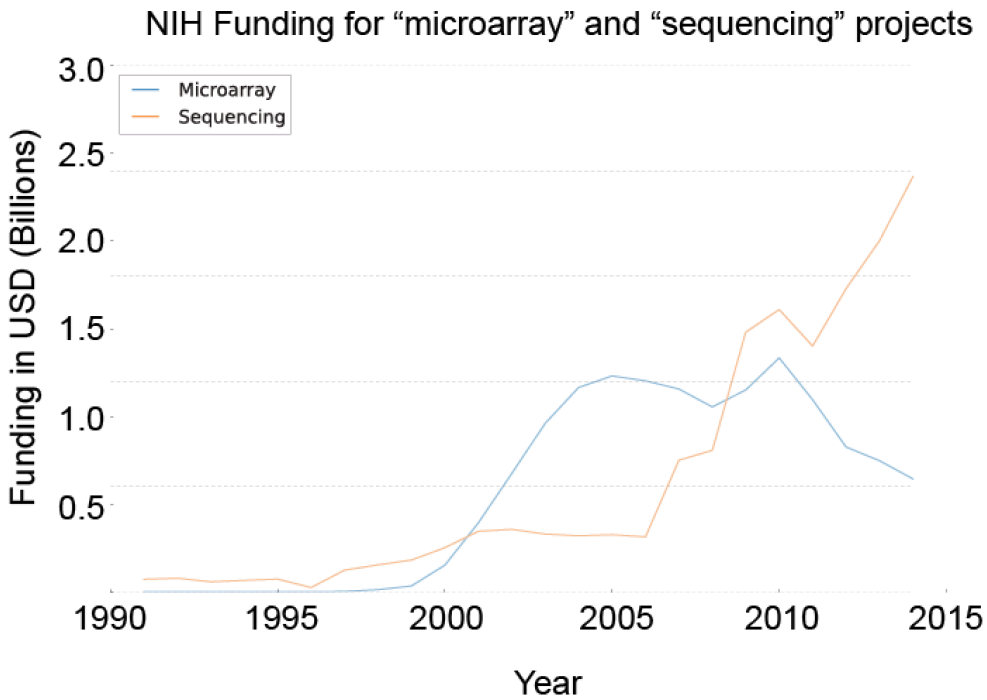
Sequencing cost reductions have resulted in an explosion of data

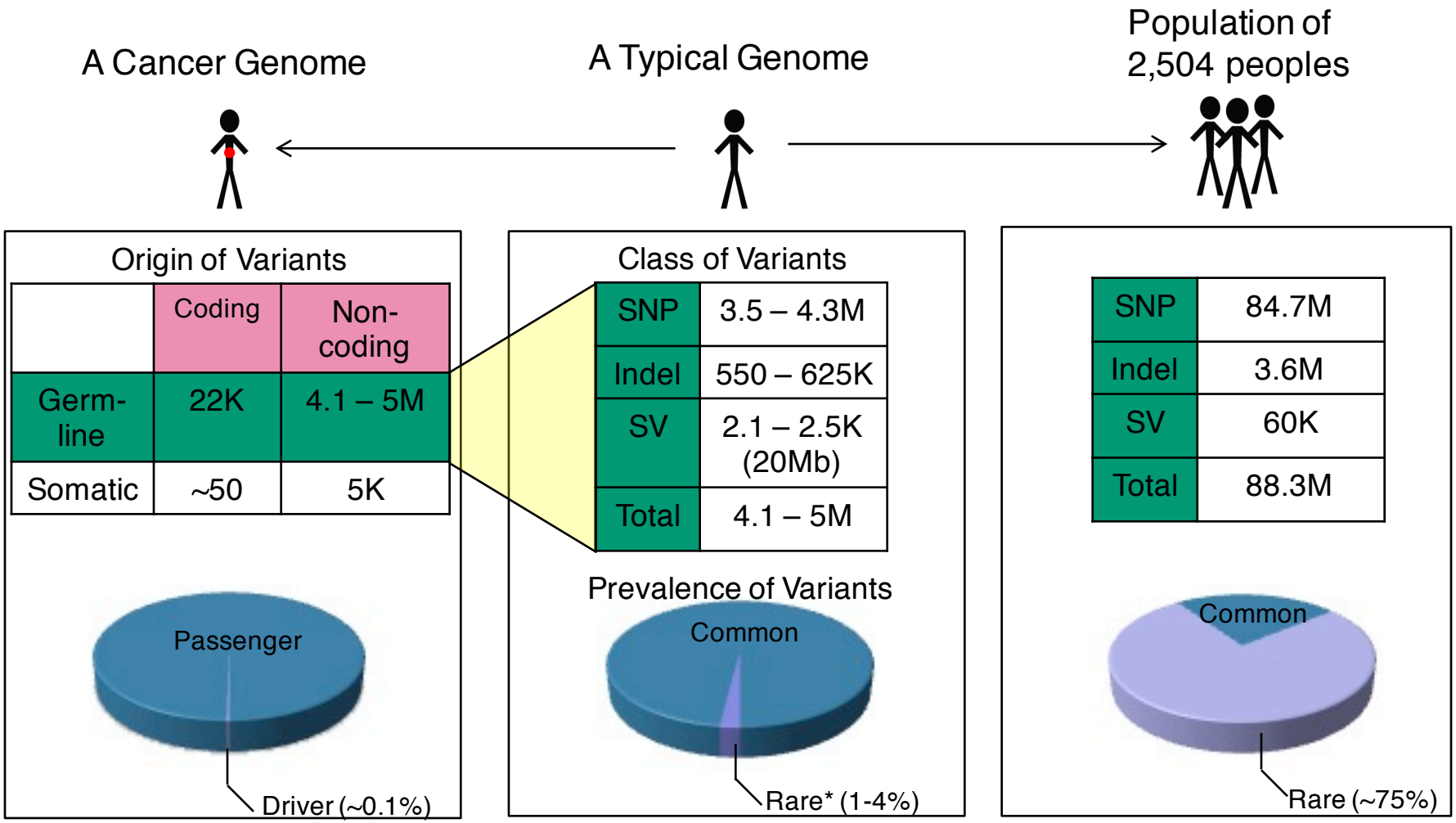
- The type of sequence data deposited has changed as well.
  - Protected data represents an increasing fraction of all submitted sequences.
  - Data from techniques utilizing NGS machines has replaced that generated via microarray.





[Muir et al. (‘15) GenomeBiol.]

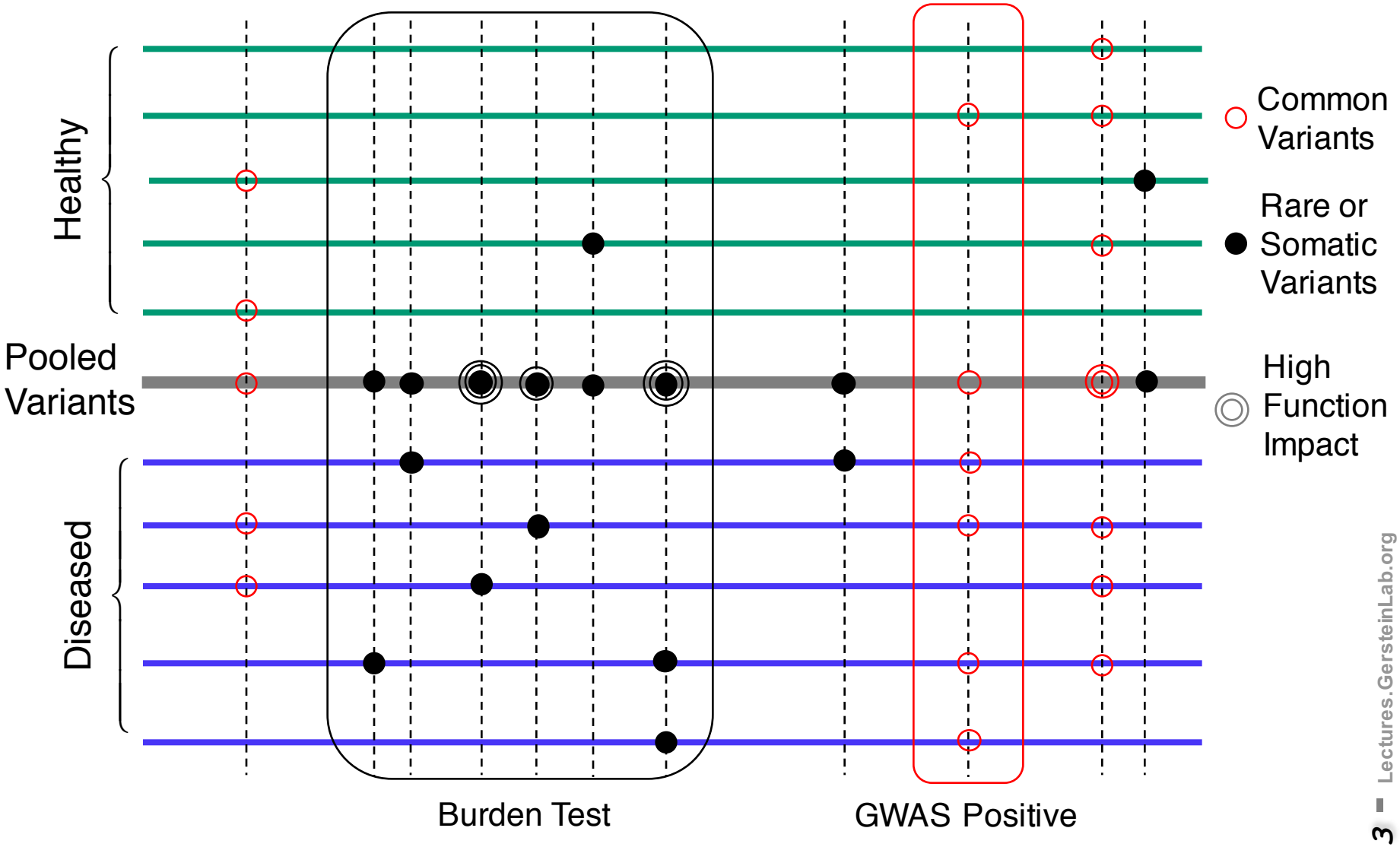
Human Genetic Variation



* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

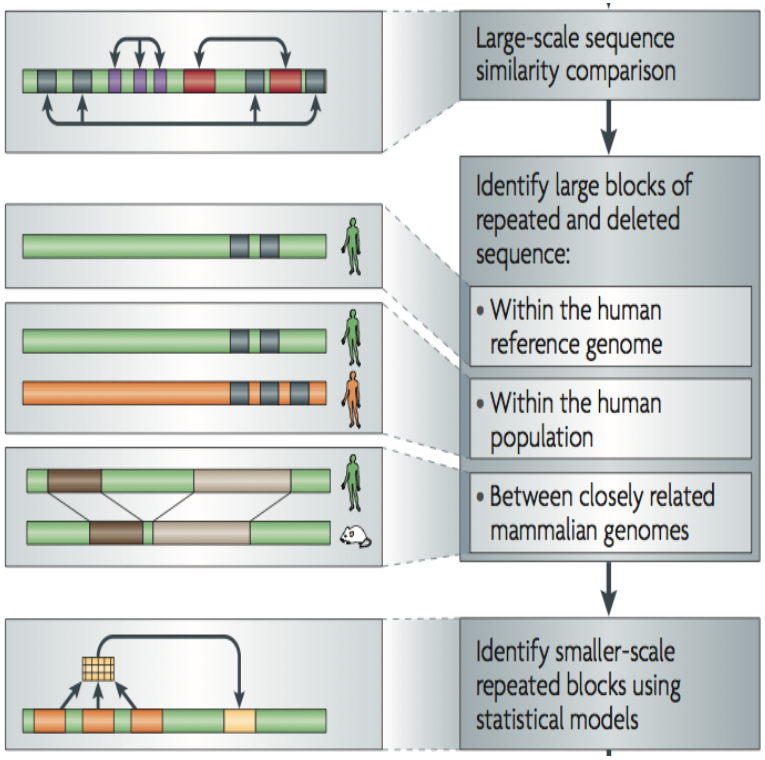
Association of Variants with Diseases

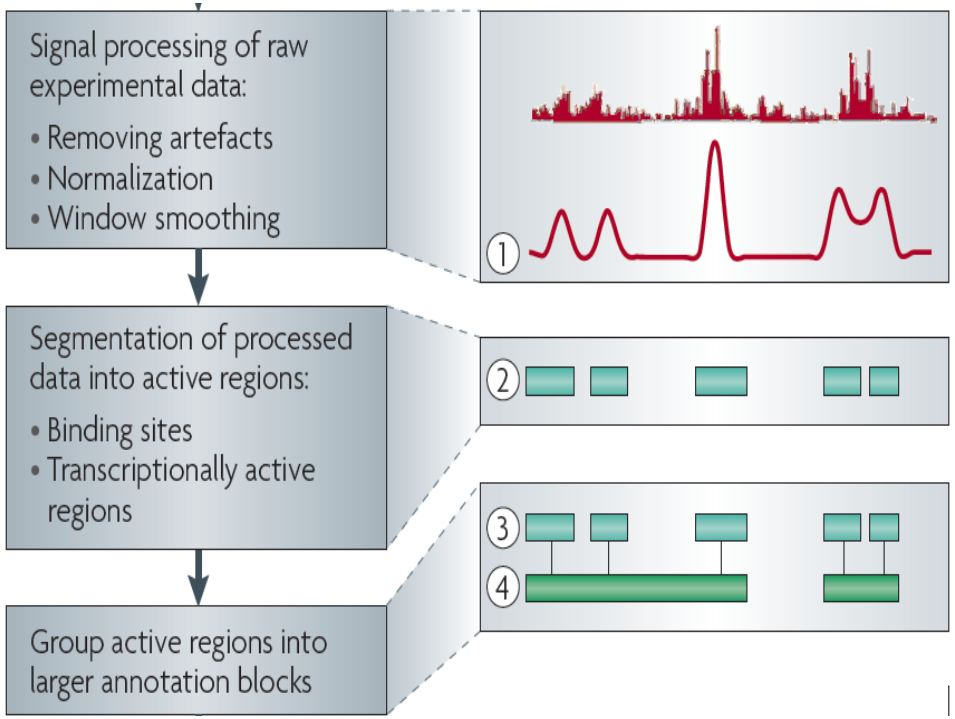


Non-coding Annotations: Overview

Sequence features, incl. **Conservation**

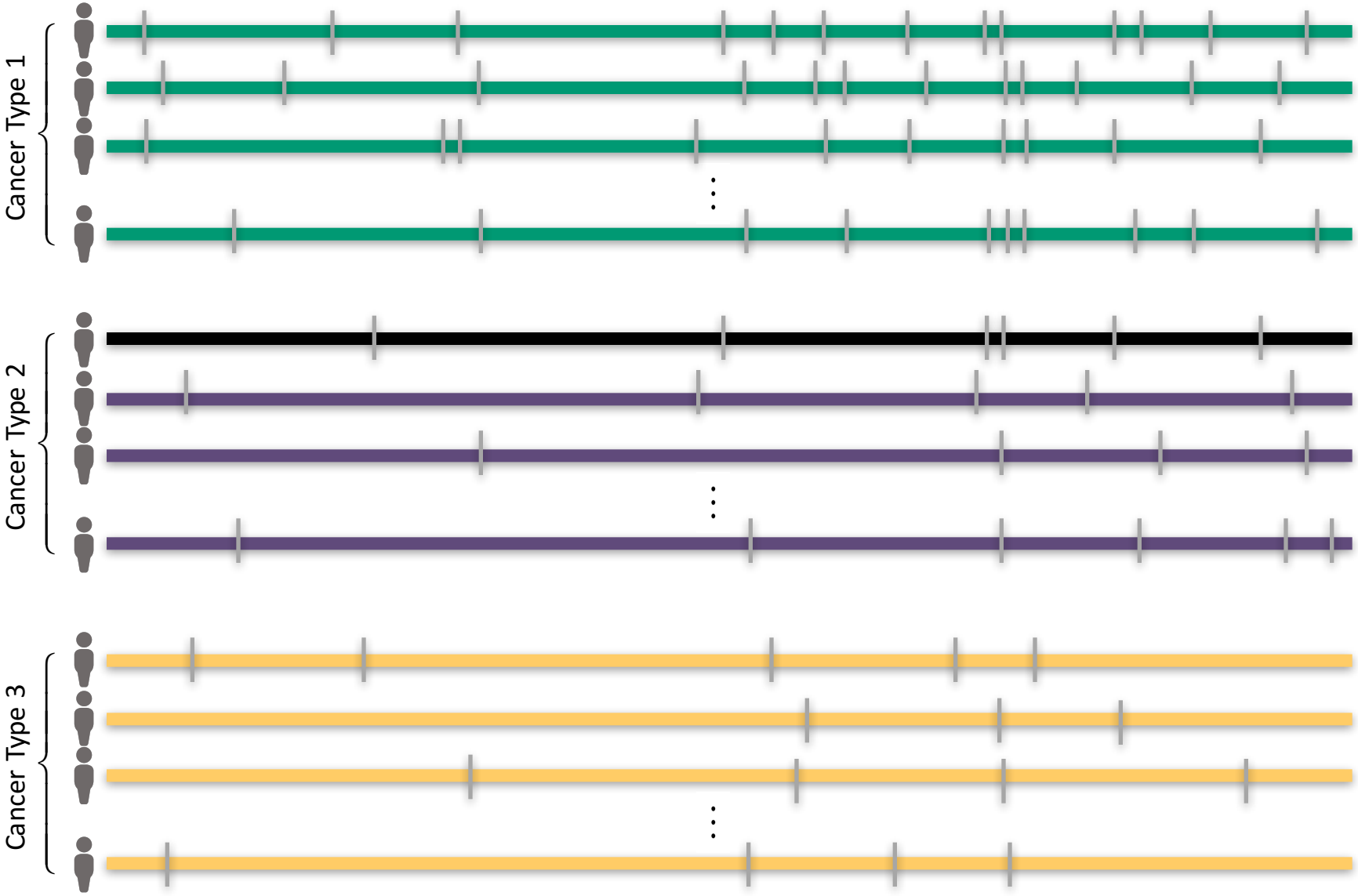
Functional Genomics
Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription



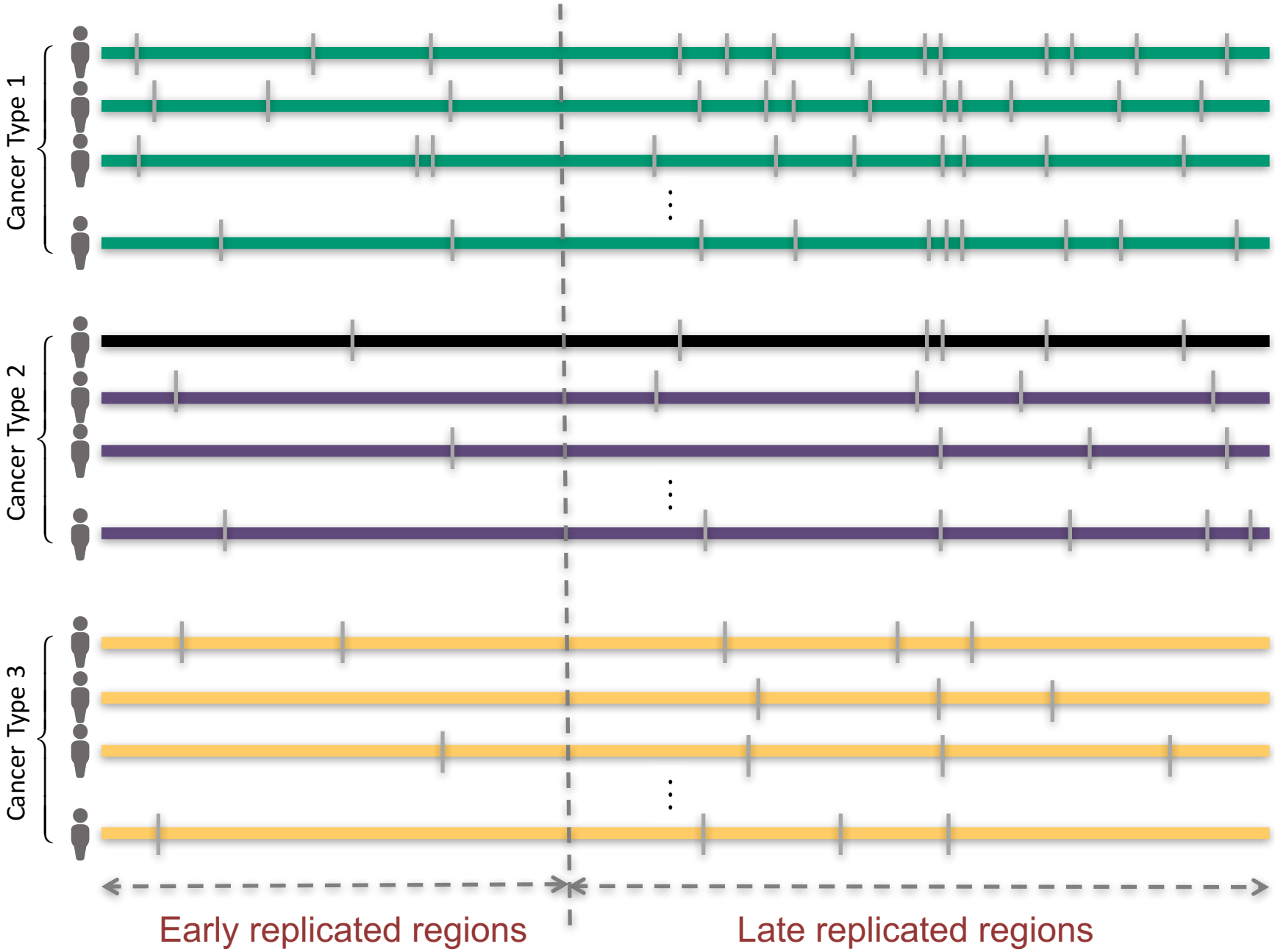


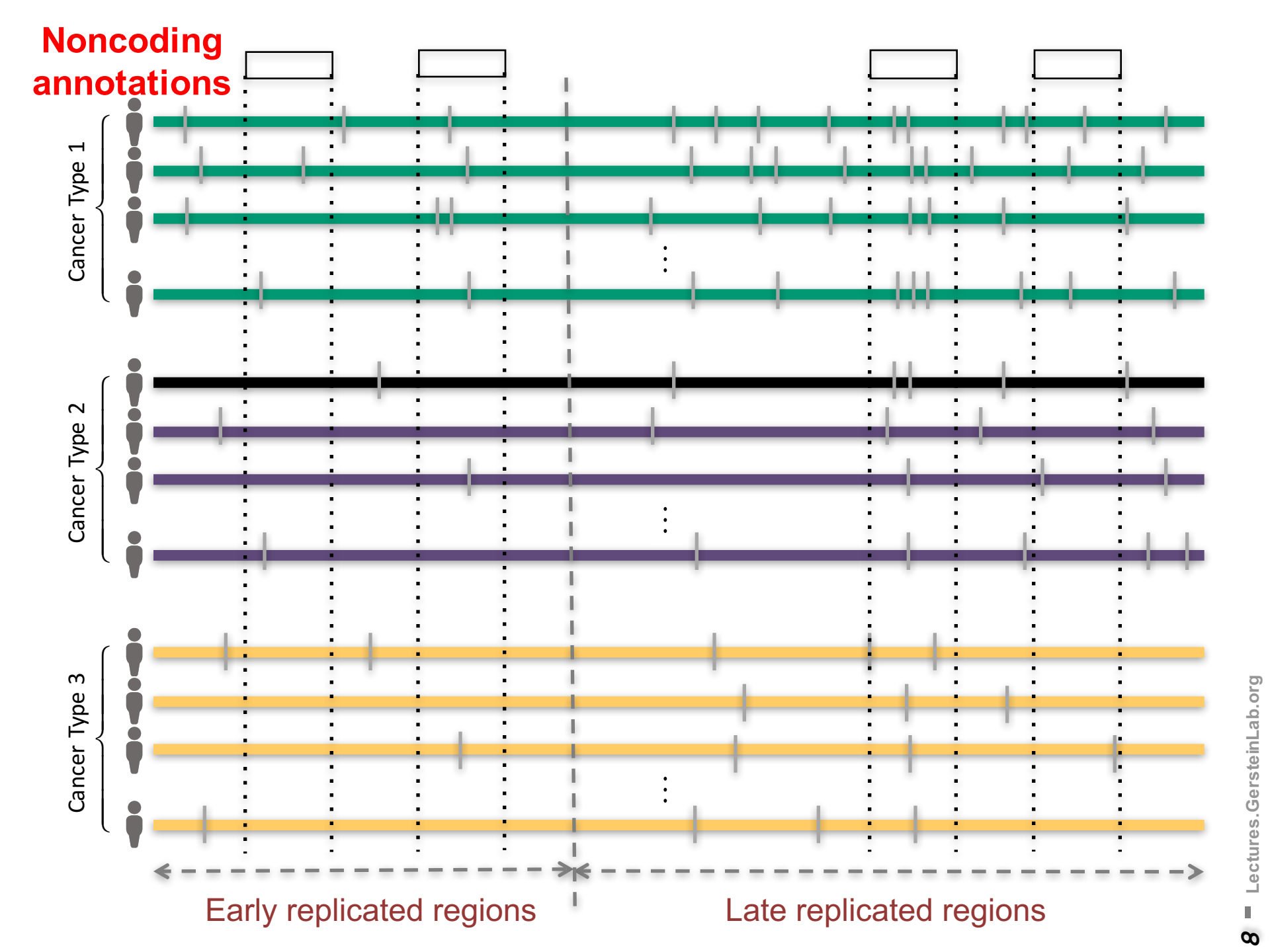
[Alexander et al., *Nat. Rev. Genet.* ('10)]

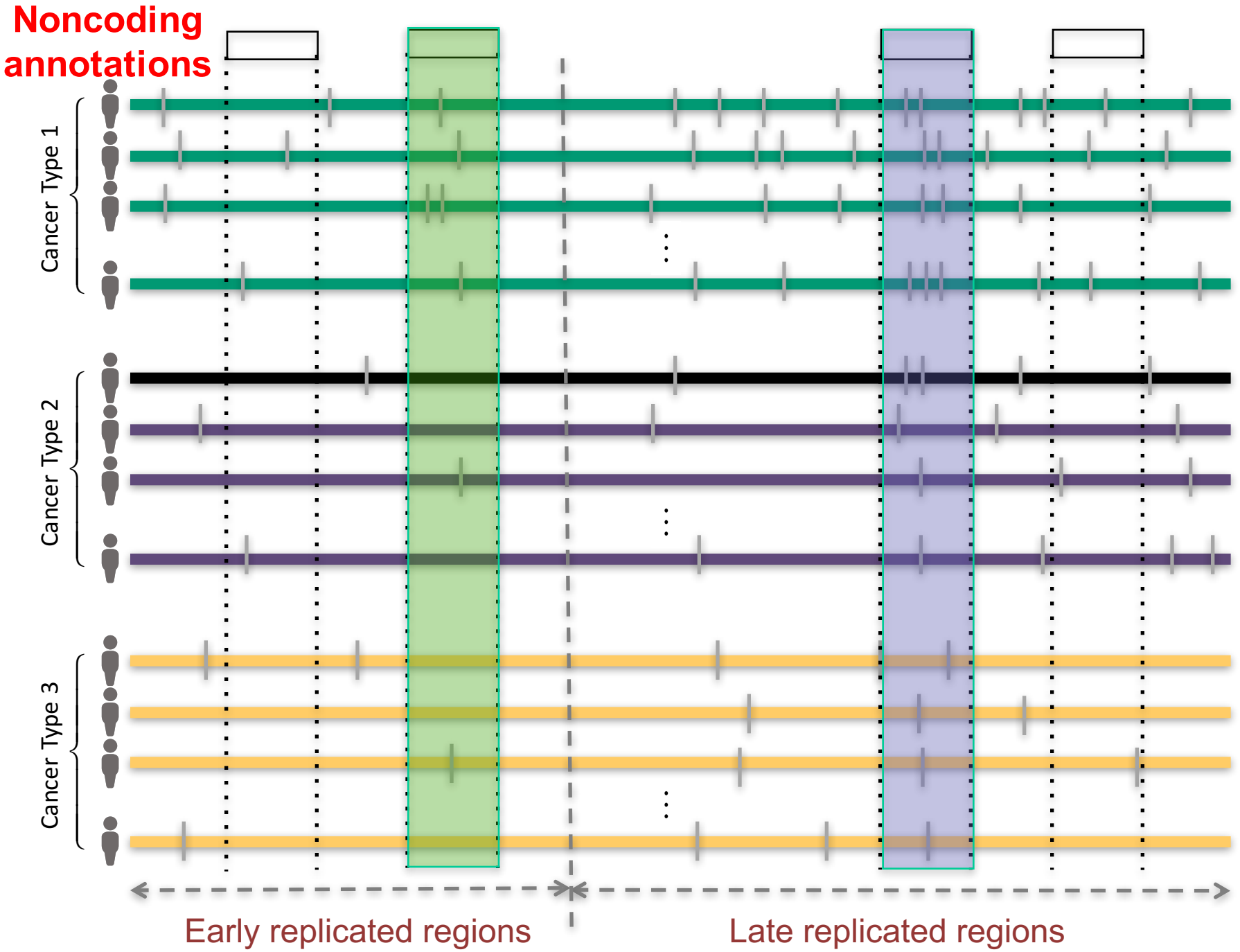
Mutation recurrence



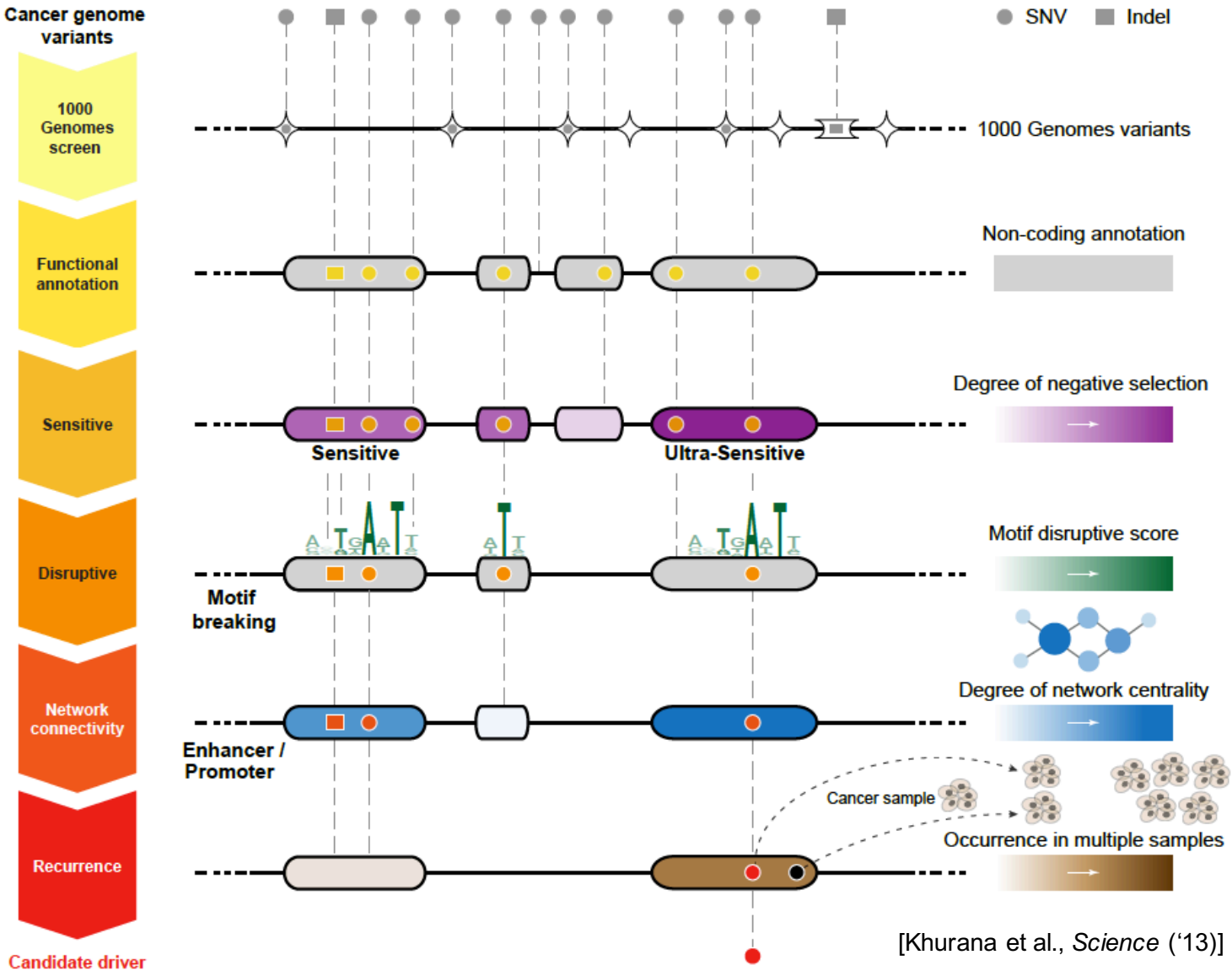
Mutation recurrence




Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions

Identification of non-coding candidate drivers amongst somatic variants: Scheme



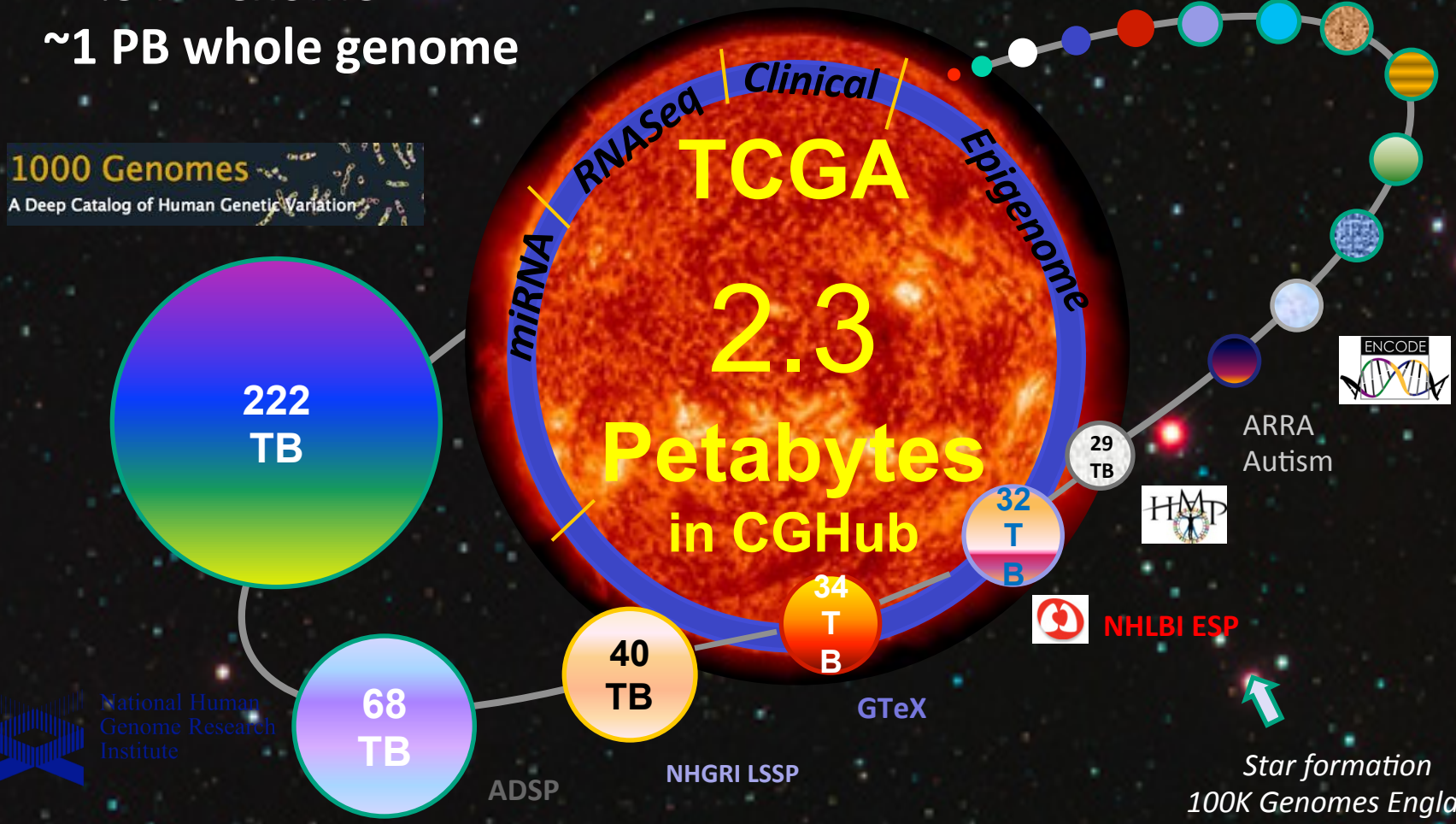
[Khurana et al., *Science* ('13)]

Sequence Universe

SRA ~1 petabyte

TCGA endpoint: ~2.5 Petabytes

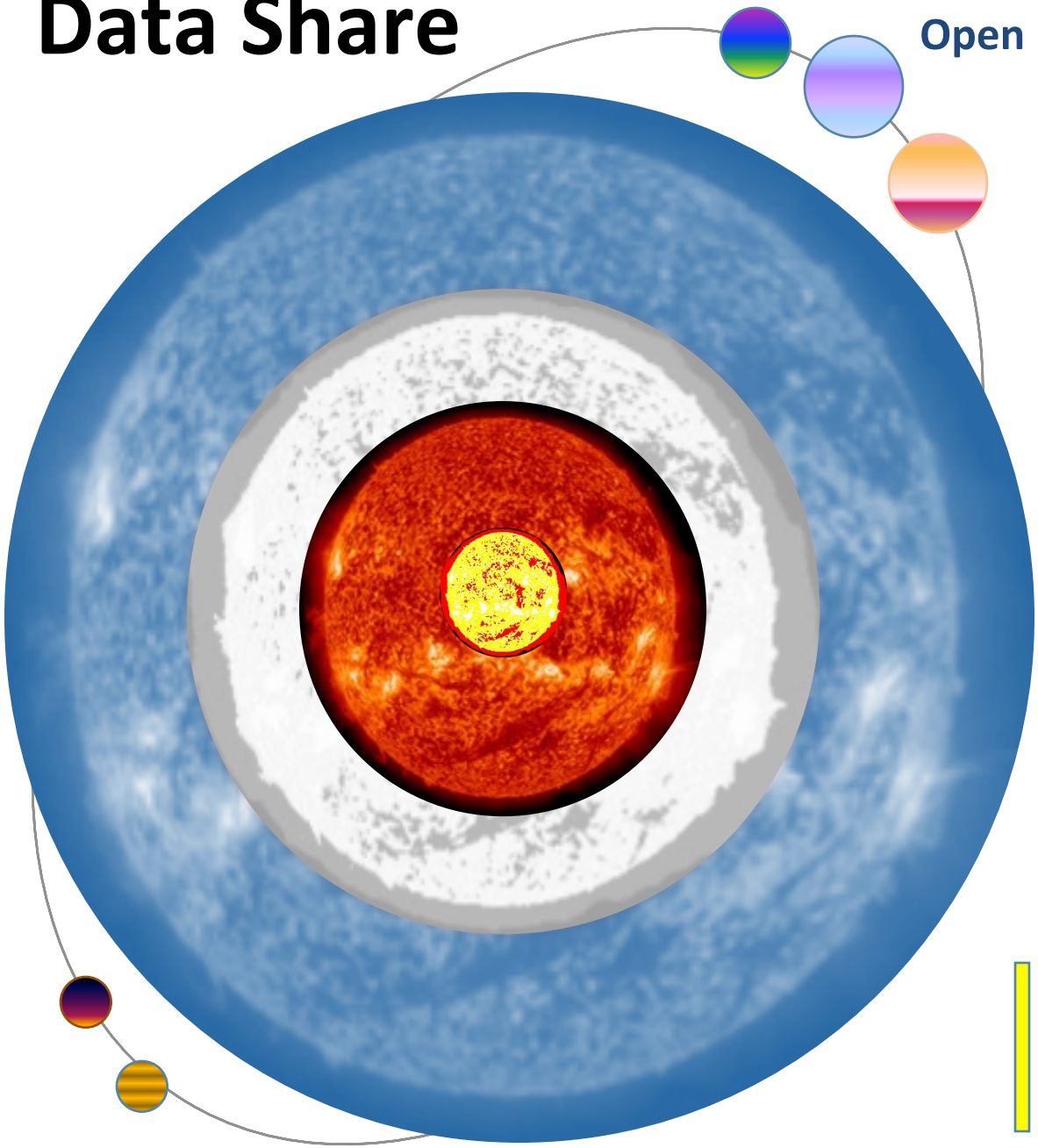
- **~1.5 PB exome**
- **~1 PB whole genome**



Heidi Sofia, 7-16-15

JESS3

Data Share



Open resources interface with API

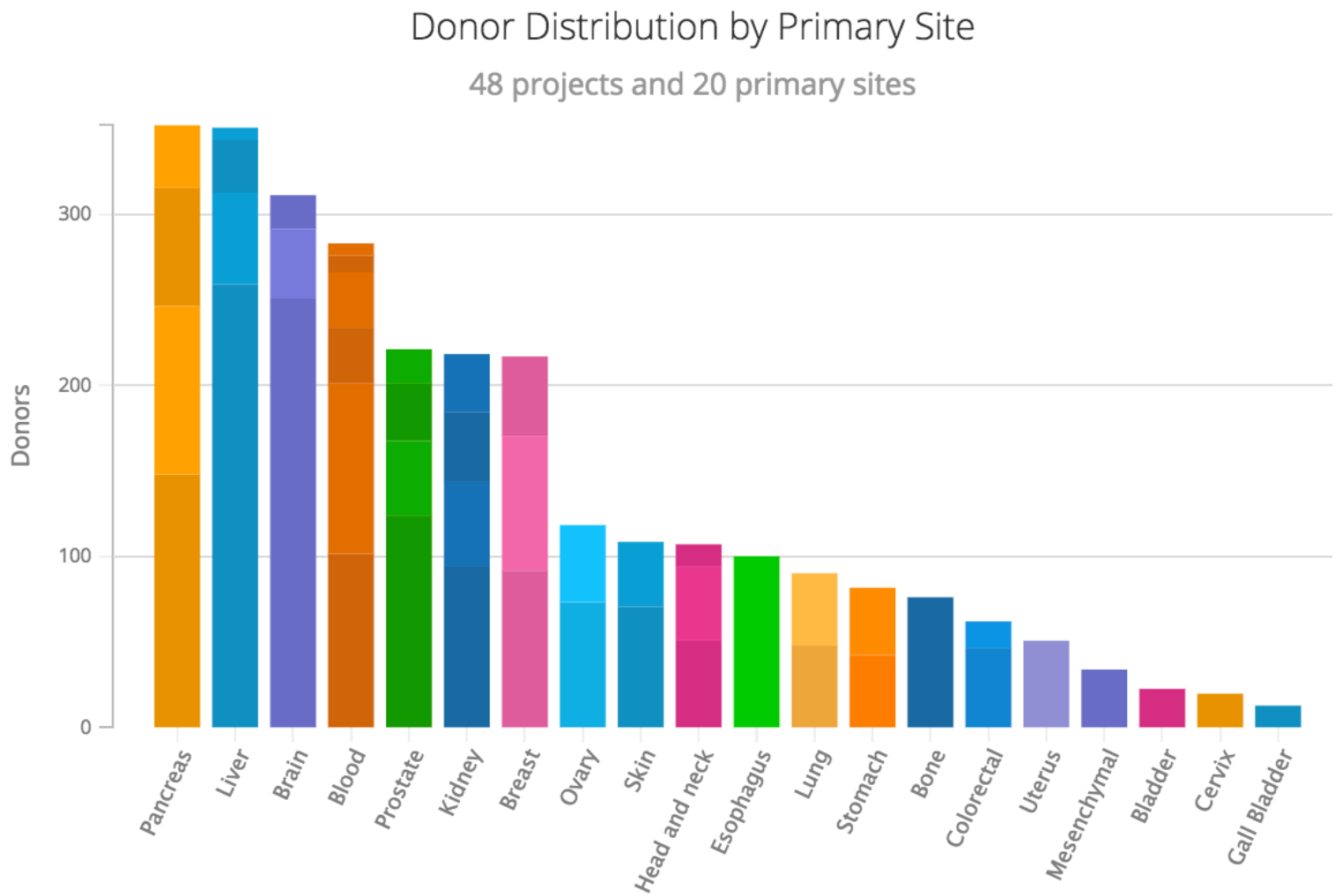
Privacy Belt
Cutting-edge cryptographic technology to ensure privacy for results returned outside of dbGaP authorization

Secure Resource
Must use internal tools
Requires user registration

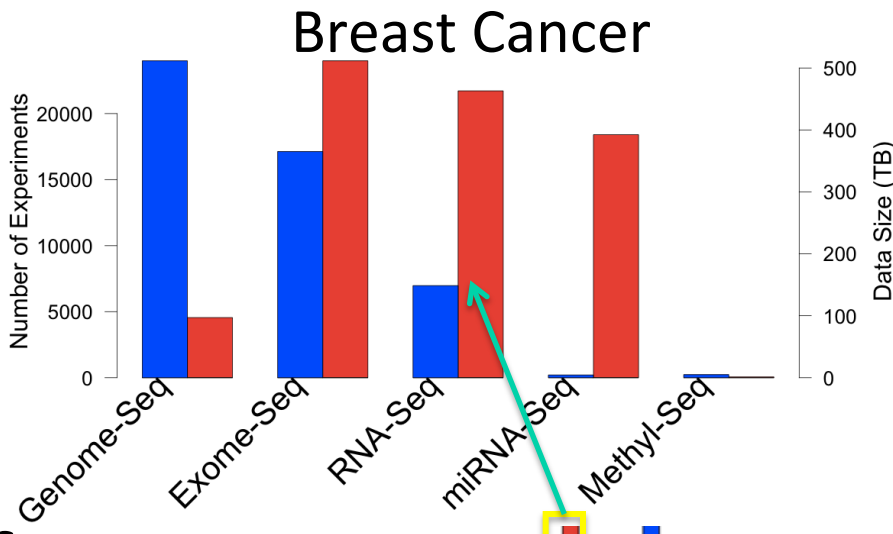
Limited Partner Grant
Bring outside tools to data
Download results only
Requires dbGaP authorization

Trusted Partner Contract
Allows data download
Requires dbGaP authorization

PCAWG: PANCANCER ANALYSIS OF WHOLE GENOMES



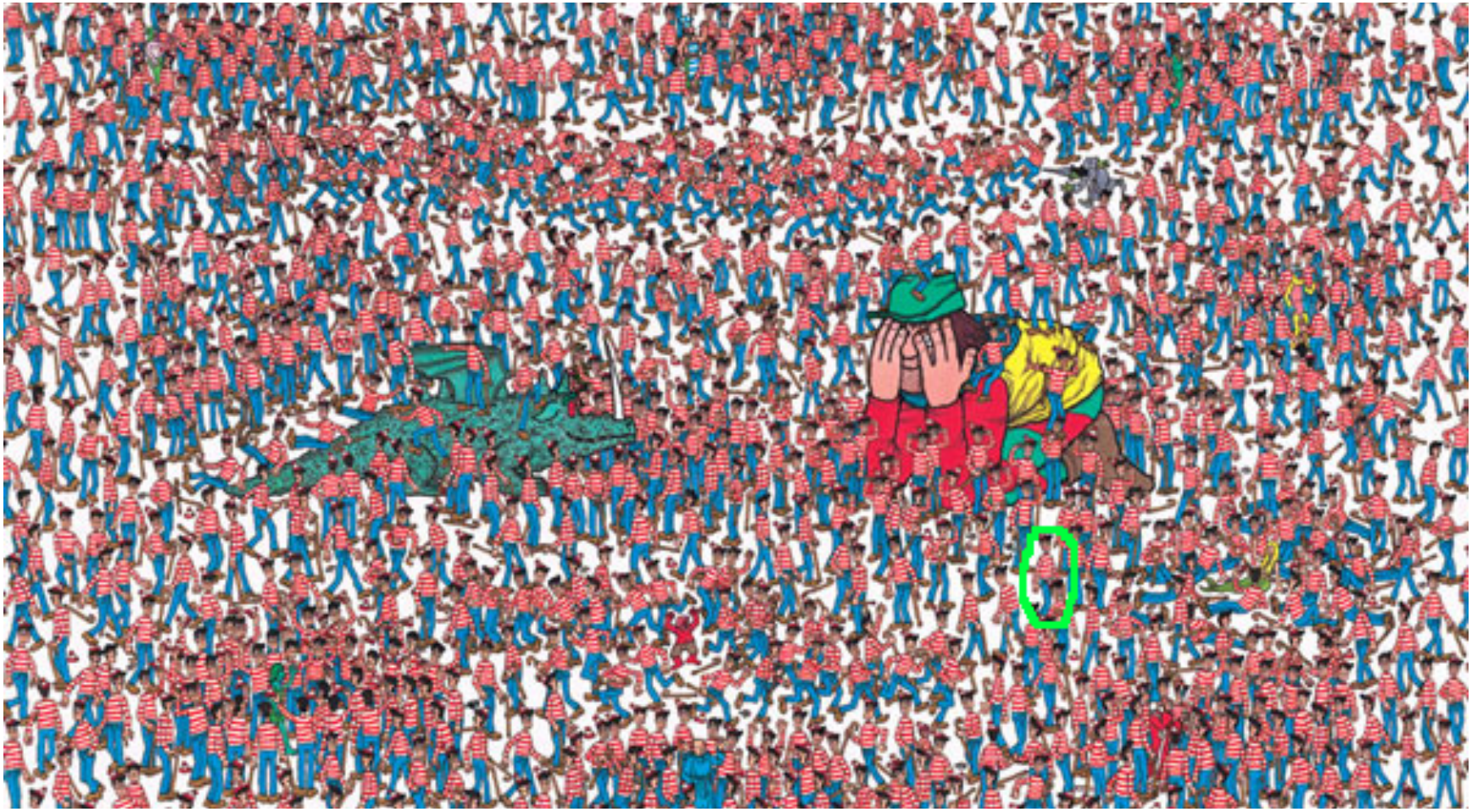
TCGA: What's in a petabyte?



>30 TCGA Cancer Types
>73K Experiments
>11K Patients

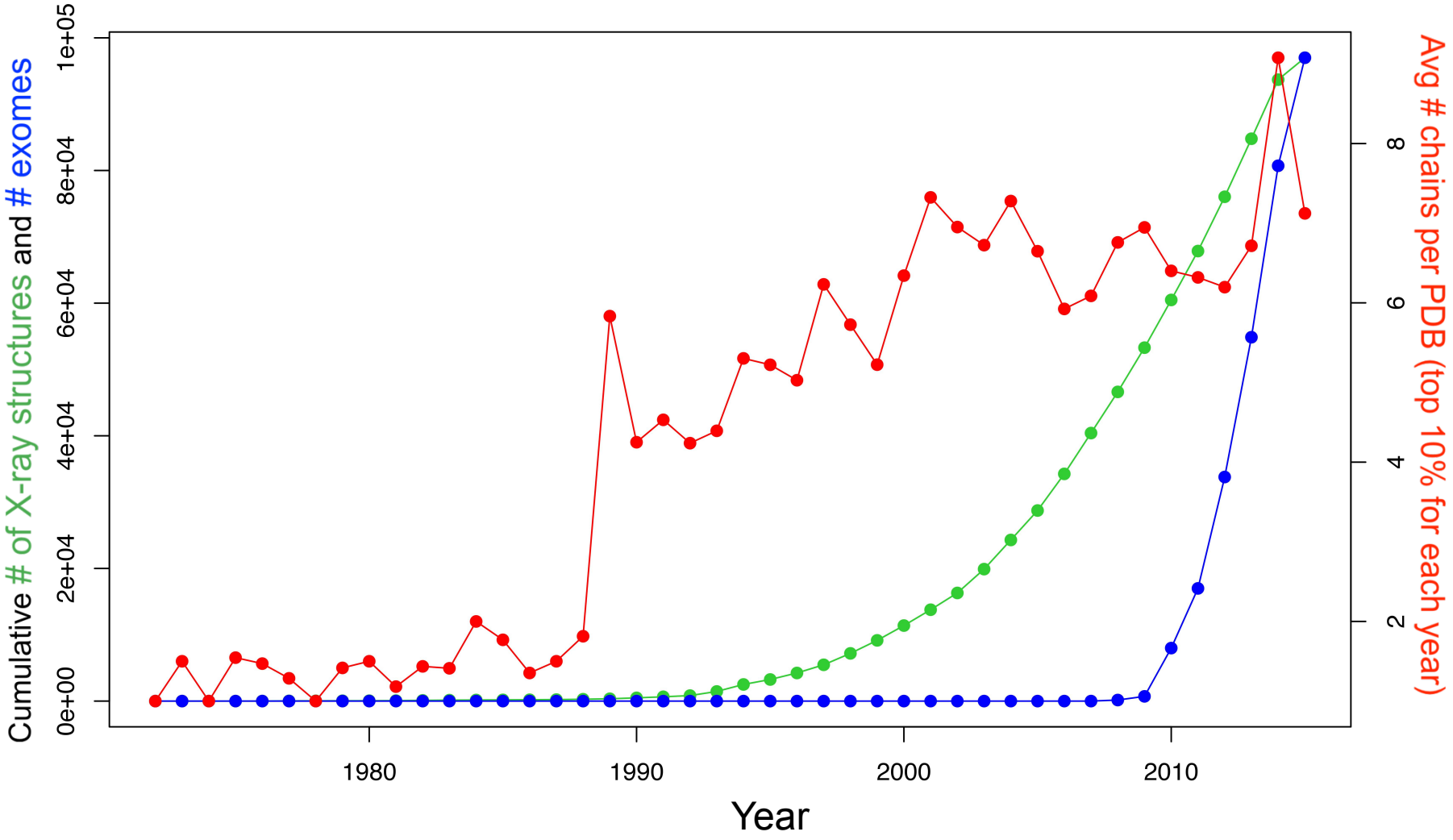
https://cghub.ucsc.edu/

Where is Waldo?
(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)

The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.



Exome data hosted on NCBI Sequence Read Archive (SRA)

[Sethi et al. COSB ('15)]

Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes

[Sethi et al. COSB ('15)]

PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)