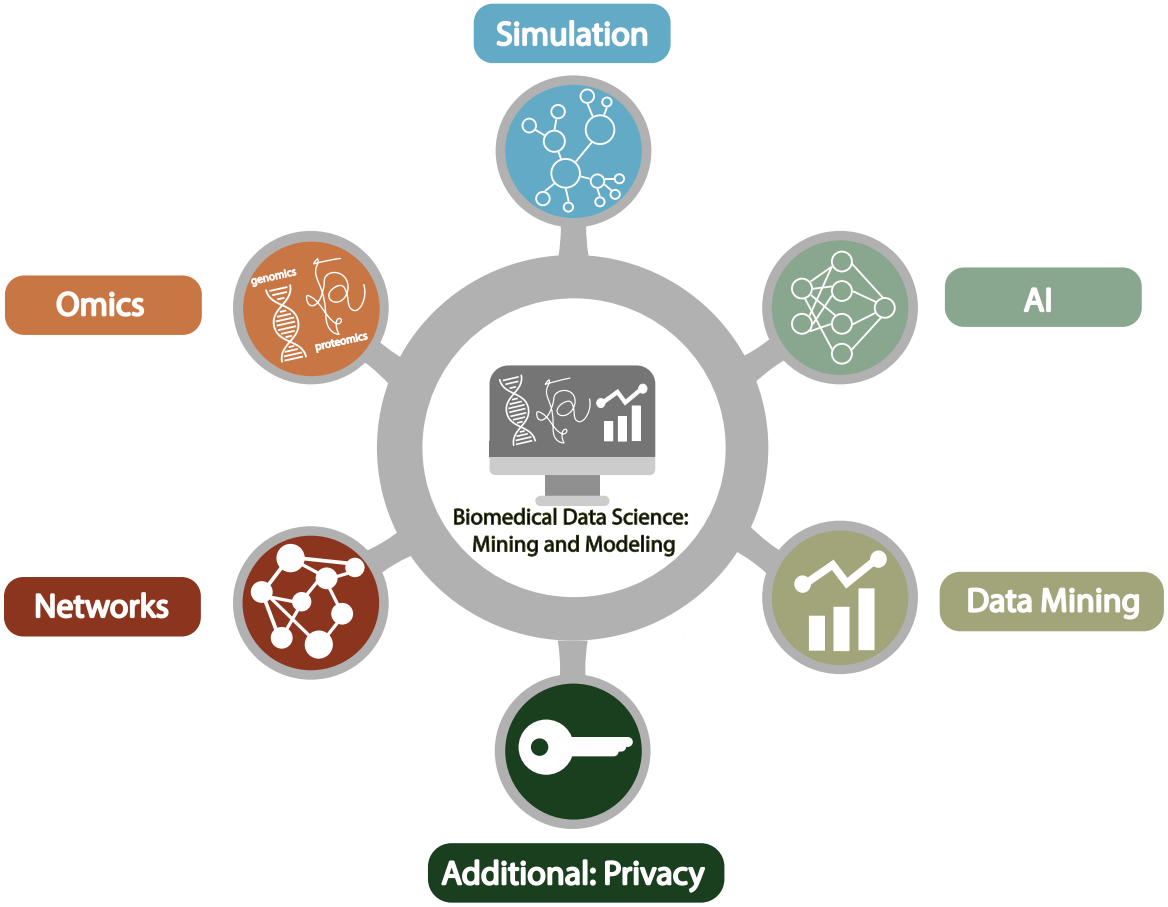
Biomedical Data Science (GersteinLab.org/courses/452)

Summary of Human Germline & Somatic Variation (25m6b)



Last edit in spring '25. Now includes gnomAD. Mostly unchanged (but with slight edits) from 22m6b & 2021's M6b [which has a video].

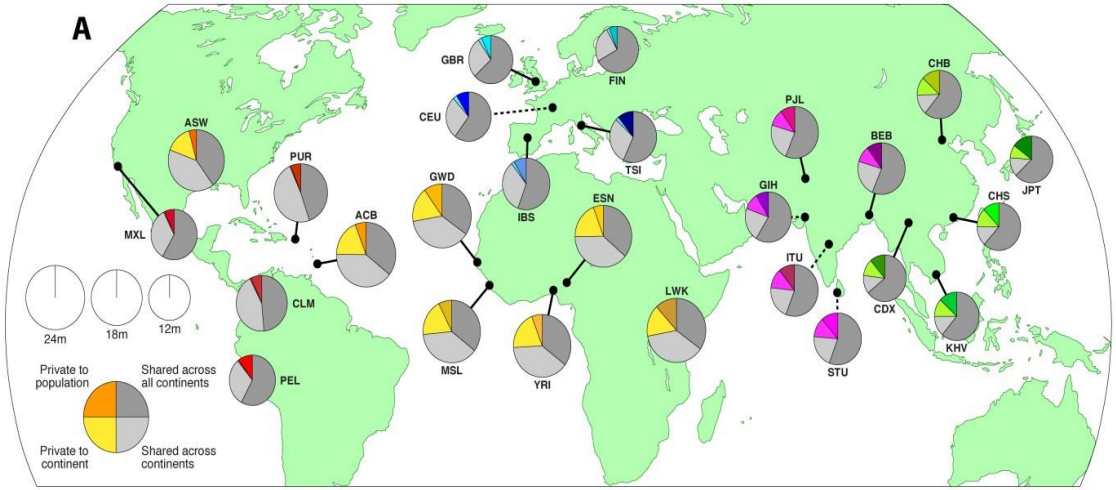
Mark Gerstein
Yale U.

1000G SV (Pilot, Phase I & III)

gnomAD (ExAC, v2, v3, v4)

[1000 Genomes Consortium, Nature (2010, 2012); Mills et al., Nature (2011)]
[gnomAD v3 paper: Konrad et al., Nature (2020)]

1000GP SV Phase3 (2015) and gnomAD V4 (2023) summary Stats

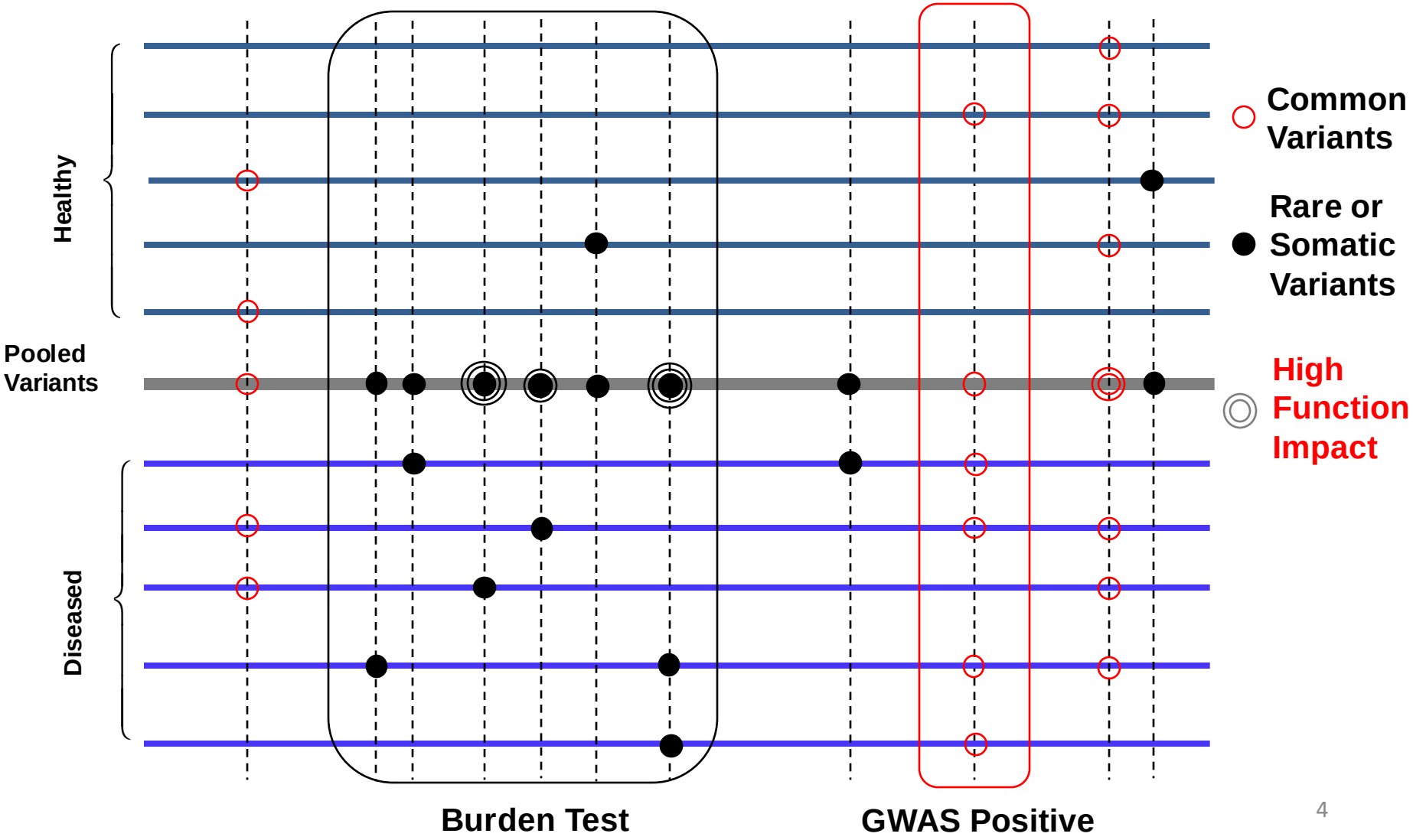


	gnomAD v4*		
	Sample count	%	Increase from v2
Admixed American	30,019	3.72%	1.7x
African/African American	37,545	4.65%	3x
Ashkenazi Jewish	14,804	1.83%	2.9x
East Asian	22,448	2.78%	2.3x
European^	622,057	77.07%	8.1x
Middle Eastern	3,031	0.38%	19.2x
Remaining^	31,712	3.93%	8.8x
South Asian	45,546	5.64%	3x
Total	**807,162**	-	-

- 68,818 SVs → **Increase by ~ 17.5 times** → 1,199,117 SVs
- 2,504 unrelated individuals → **~ 322 times** → 807,162 individuals
- 26 populations
- 37,250 SVs with resolved breakpoints

[2] 1000GP Phase3 SV paper. Submi ed to Nature, 2015.
[3] 1000GP ConsorSu m. Submi ed to Nature, 2015.
[4] gnomAD project https://gnomad.broadinstitute.org/
[5] gnomAD v3 paper: Konrad et al., Nature (2020)

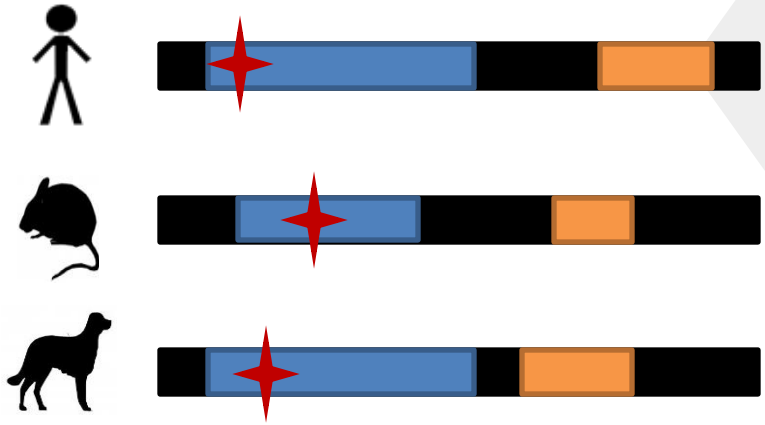
Rare & Common Variants

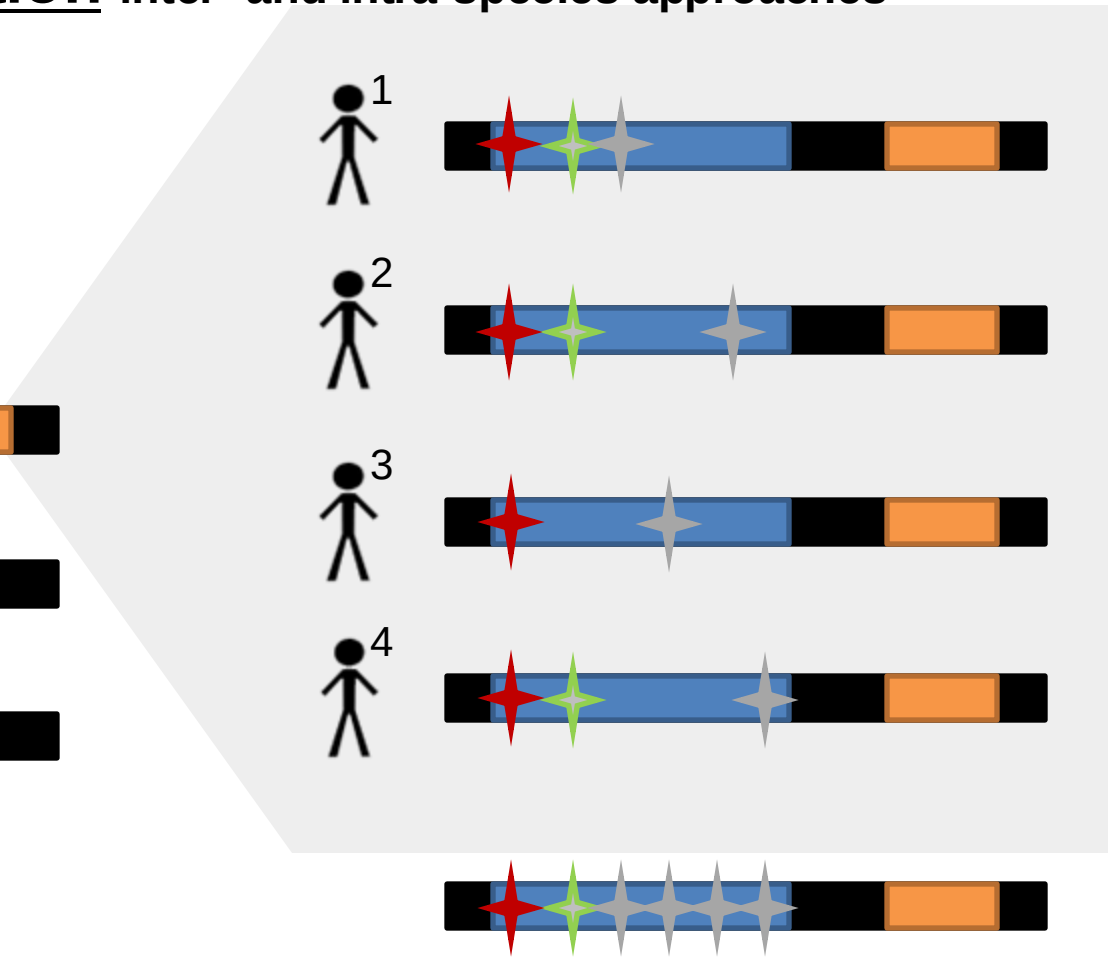


Quantifying Selection inter- and intra-species approaches

'Conservation'

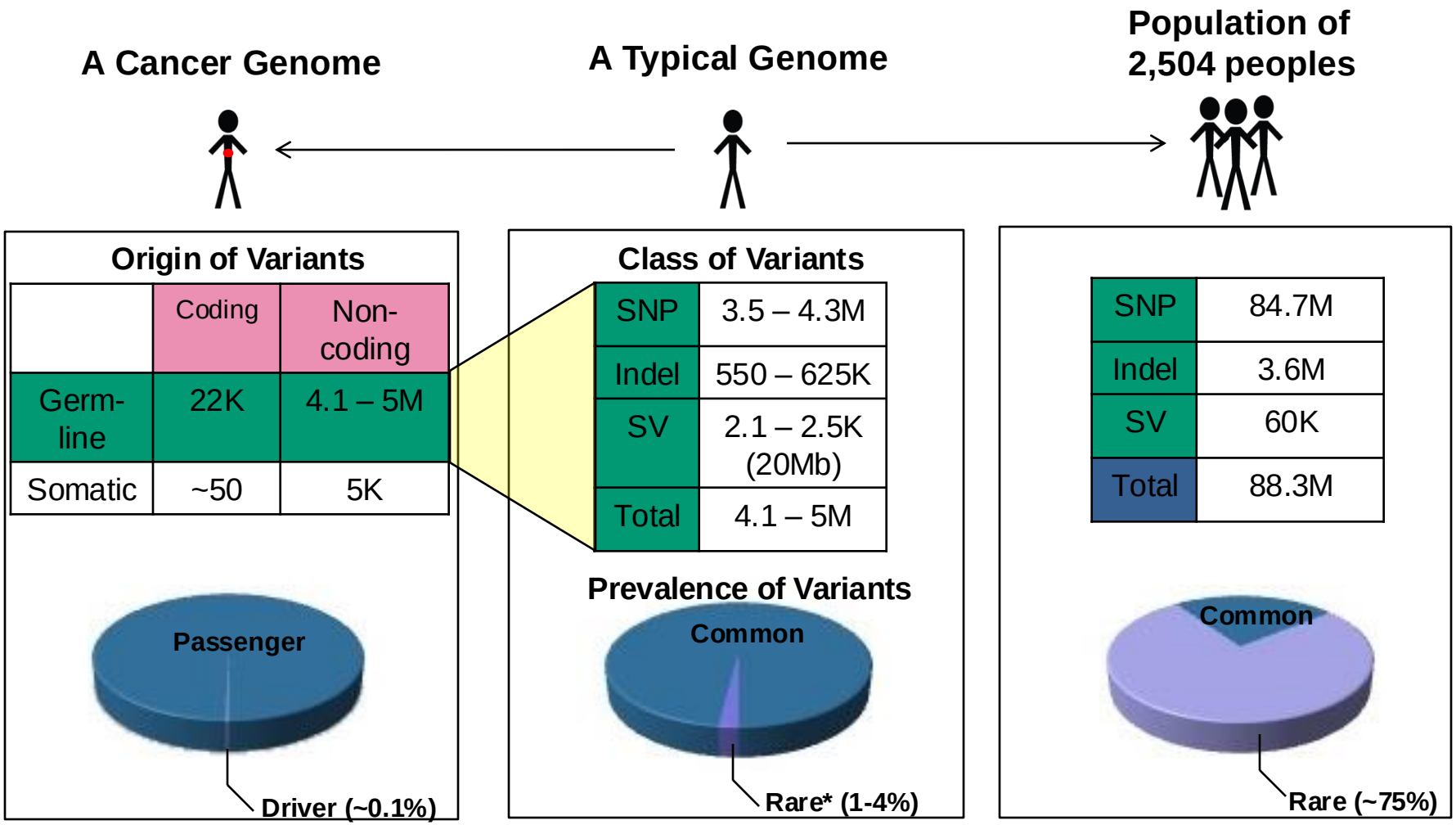
- Typically defined by comparison across species
- dN/dS in coding regions
- GERP noncoding





- Metrics for selection within population
  - SNP density (confounded by mutation rate)
- Depletion of common polymorphisms for regions under selection (also an enrichment of rare variants)

Human Genetic Variation (1000G)



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

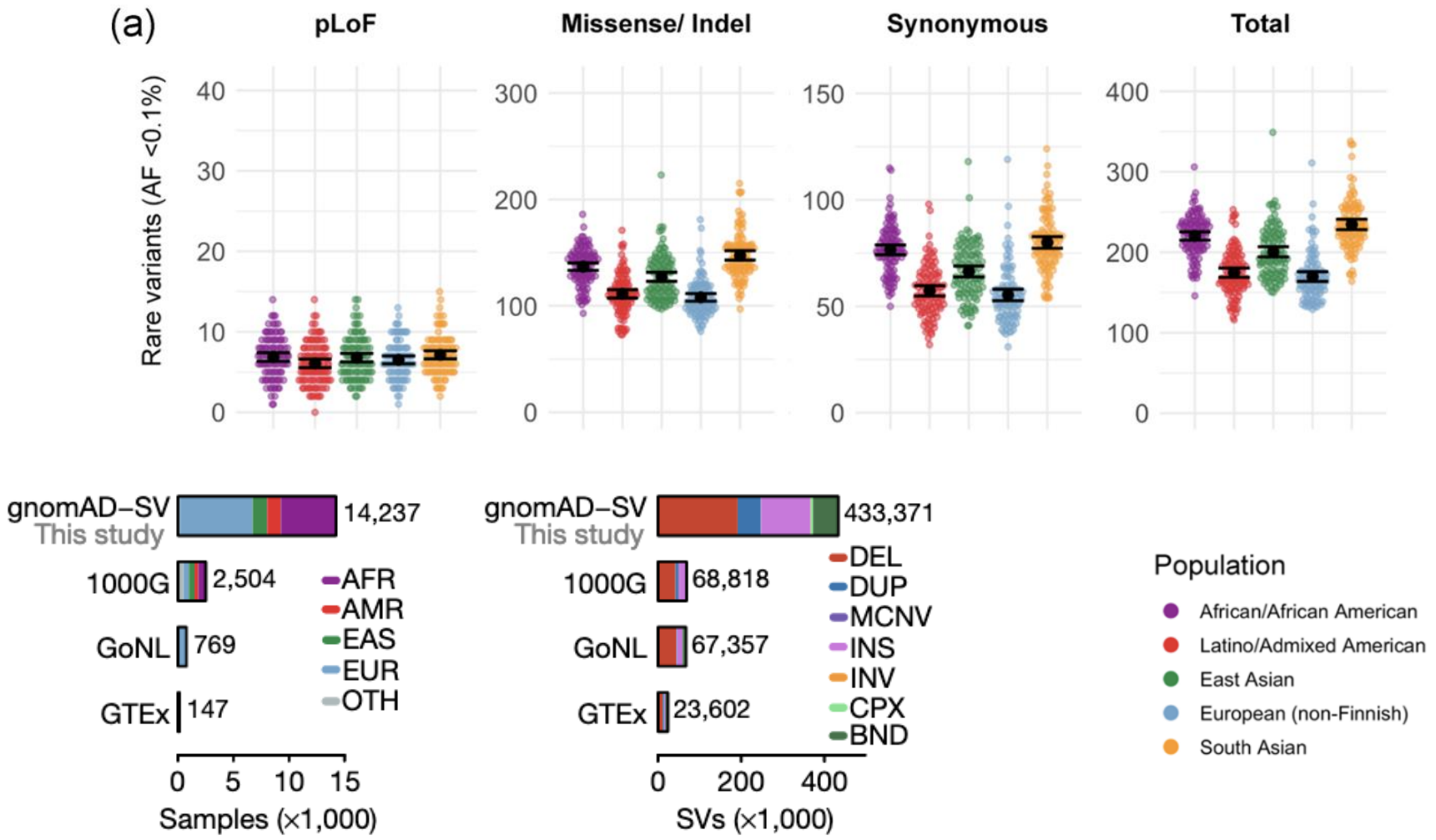
[6] The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Phase 3 (‘15): Median Autosomal Variant Sites Per Genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean Coverage	8.2		7.6		7.7		7.4		8.0	
	Var. Sites	Singletons	Var. Sites	Singletons	Var. Sites	Singletons	Var. Sites	Singletons	Var. Sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large Deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (LINE1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
NonSynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBS	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

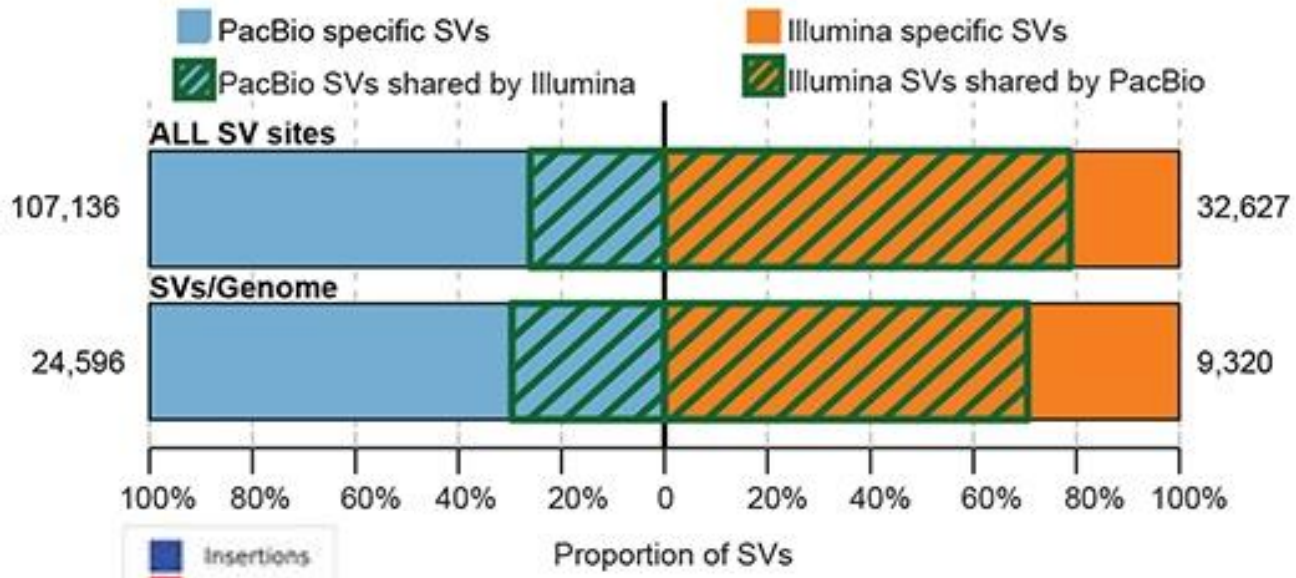
[9] 1000GP Consortium. Nature, 2015.

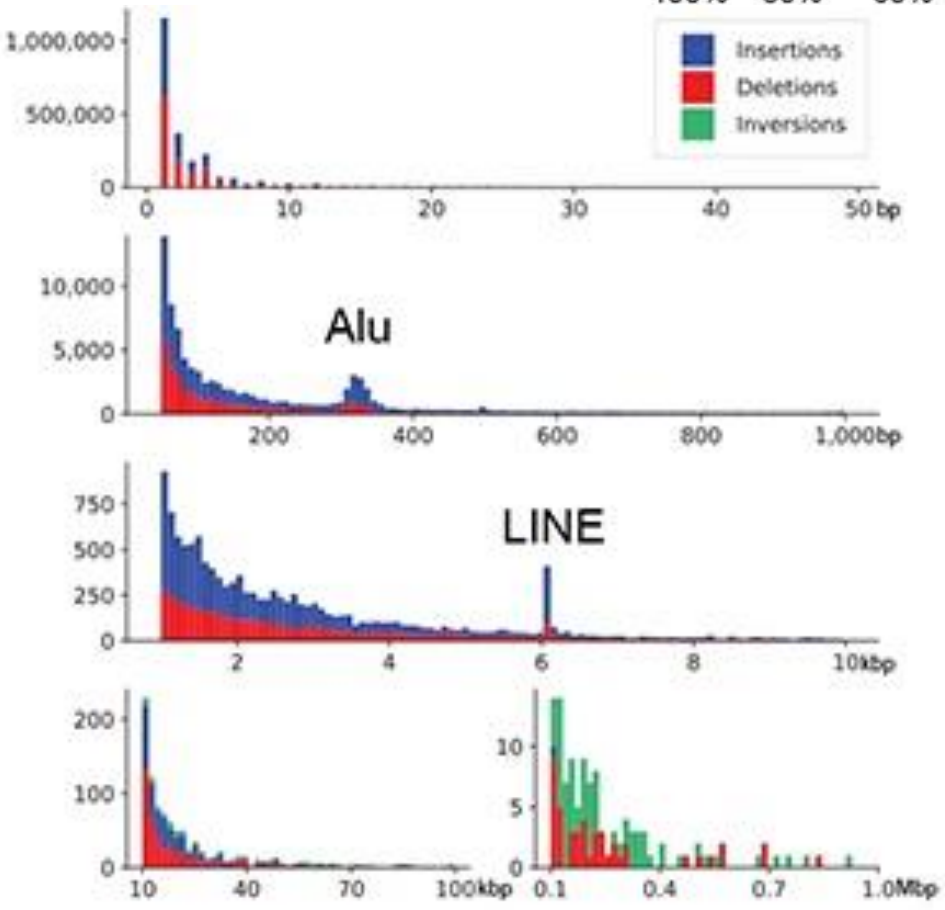
gnomAD: Mean Counts of Rare and Unique Coding Variants Across Populations



[7] Gudmundsson et al., Human Mutation 1-19 (2021)
[8] gnomAD-SV: Ryan et at., Nature (2020)

Updating the SV Numbers with Most Current Technology (PacBio HiFi)



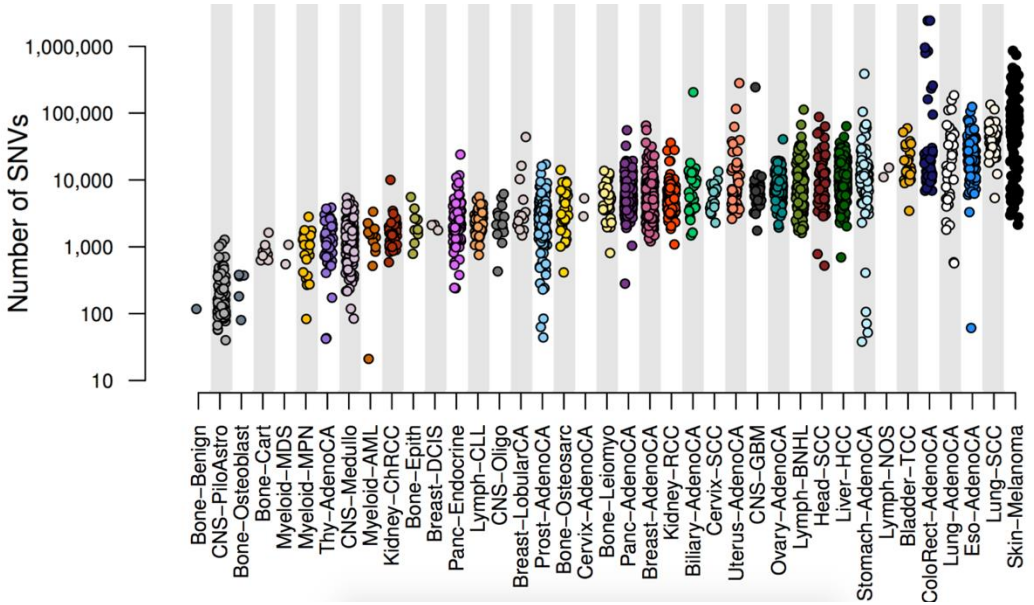


- On average, detected: 24,653 SVs, 794,406 indels, and 3,895,274 SNVs per diploid human genome

[10] Ebert et al. Science ('21)

PCAWG summary (somatic variants)

PCAWG : most comprehensive resource for cancer whole genome analysis



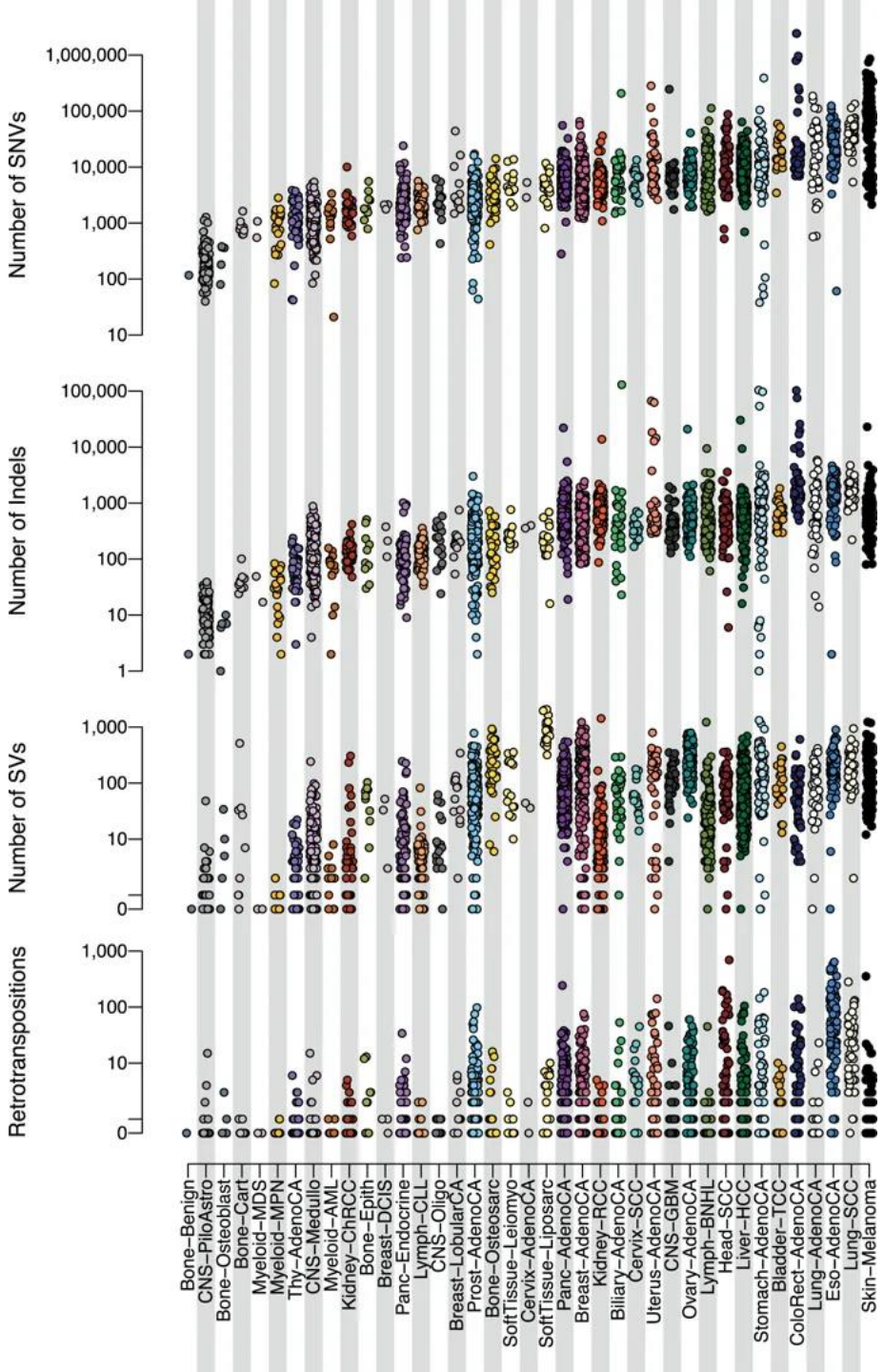
[11] Adapted from Campbell et. al., bioRxiv ('17). Now published as Nature 578: 82–93 (2020)

Project Goals:

- **To understand role of non-coding regions of cancer genomes in disease progression.**
- **Union of TCGA-ICGC efforts**

- **Jointly analyzing ~2800 whole genome tumor/normal pairs**
  - **> 580 researchers**
  - **16 thematic working groups**
  - **~30M total somatic SNVs**

PCAWG
PanCancer Analysis OF WHOLE GENOMES



PCAWG Summary Variant Totals by Cancer

Nature 578: 82–93 (2020)

References

- 1000G consortium. Nature, 526(7571), 68–74.
  A global reference for human genetic variation.
  https://doi.org/10.1038/nature15393
  (Focus on text associated with Table 1.)

- PCAWG consortium. Nature, 578(7793), 82–93.
  Pan-cancer analysis of whole genomes.
  https://doi.org/10.1038/s41586-020-1969-6
- (Focus on text associated with Extended Data Fig. 3)