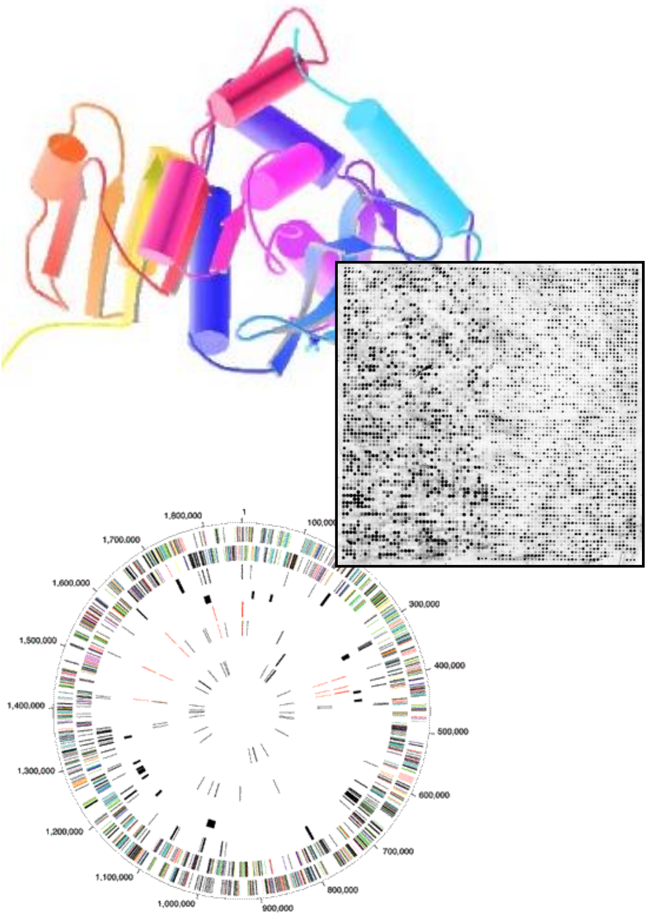
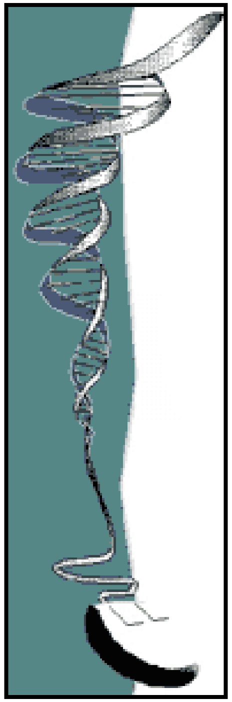
Biomed. Data Sci:

1000G+PCAWG Summary





Mark Gerstein, Yale University

gersteinlab.org/courses/452

(Last edit in spring ’22, final.
This is 22m6b which has new slide 8 relative to last year’s M6b.)

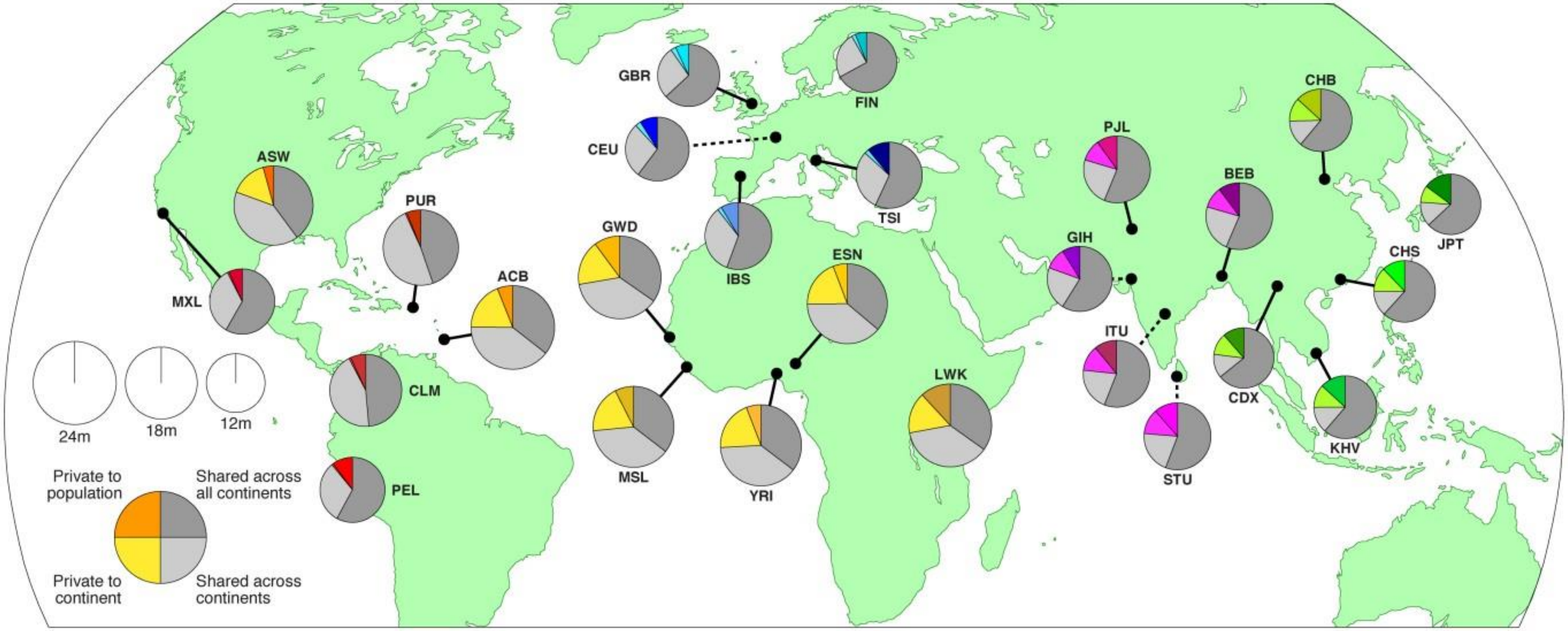
1000G SV
(Pilot, **Phase I** & III)

- **Many** different SV callers compared & used
  - including SRiC & CNVnator but also VariationHunter, Cortex, NovelSeq, PEMer, BreakDancer, Mosaik, Pindel, GenomeSTRiP, mrFast….
- Merging
- Genotyping
- Breakpoint assembly
- Mechanism Classification

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061
HUMAN STEM CELLS
BEYOND THE COURT CASE
Implications for the law, industry and ethics
PAGE 1031
OCEAN PRODUCTIVITY
PHOSPHATE DOWN THE AGES
Key nutrient plentiful after 'snowball' Earth
PAGES 1052 & 1088
AUTUMN BOOKS
THE RECURRING UNIVERSE
Lee Smolin on Roger Penrose's grand idea
PAGE 1034
NATURE.COM/NATURE
28 October 2010 £10
Vol. 467, No. 7319

[1000 Genomes Consortium, Nature (2010, 2012); Mills et al., Nature (2011)]

Summary Stats of 1000GP SV Phase3



- 68,818 SVs
- 2,504 unrelated individuals
- 26 populations
- 37,250 SVs with resolved breakpoints

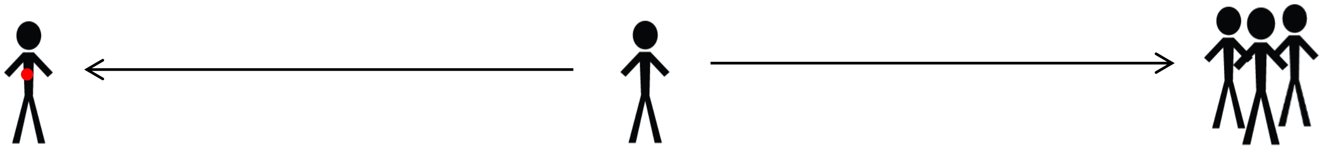
[2] 1000GP Phase3 SV paper. Submid ed to Nature, 2015.
[3] 1000GP ConsorS u m. Submid ed to Nature, 2015.

Human Genetic Variation

A Cancer Genome

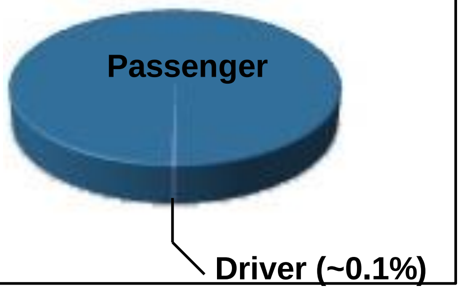
A Typical Genome

Population of 2,504 peoples



Origin of Variants

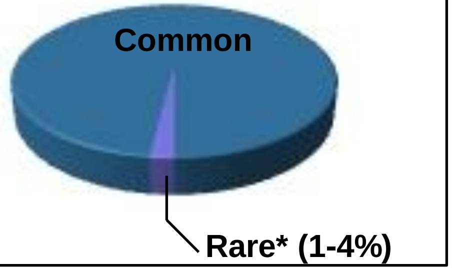
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



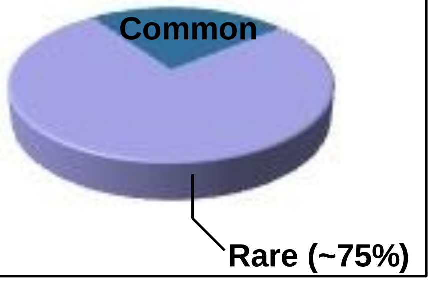
Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants



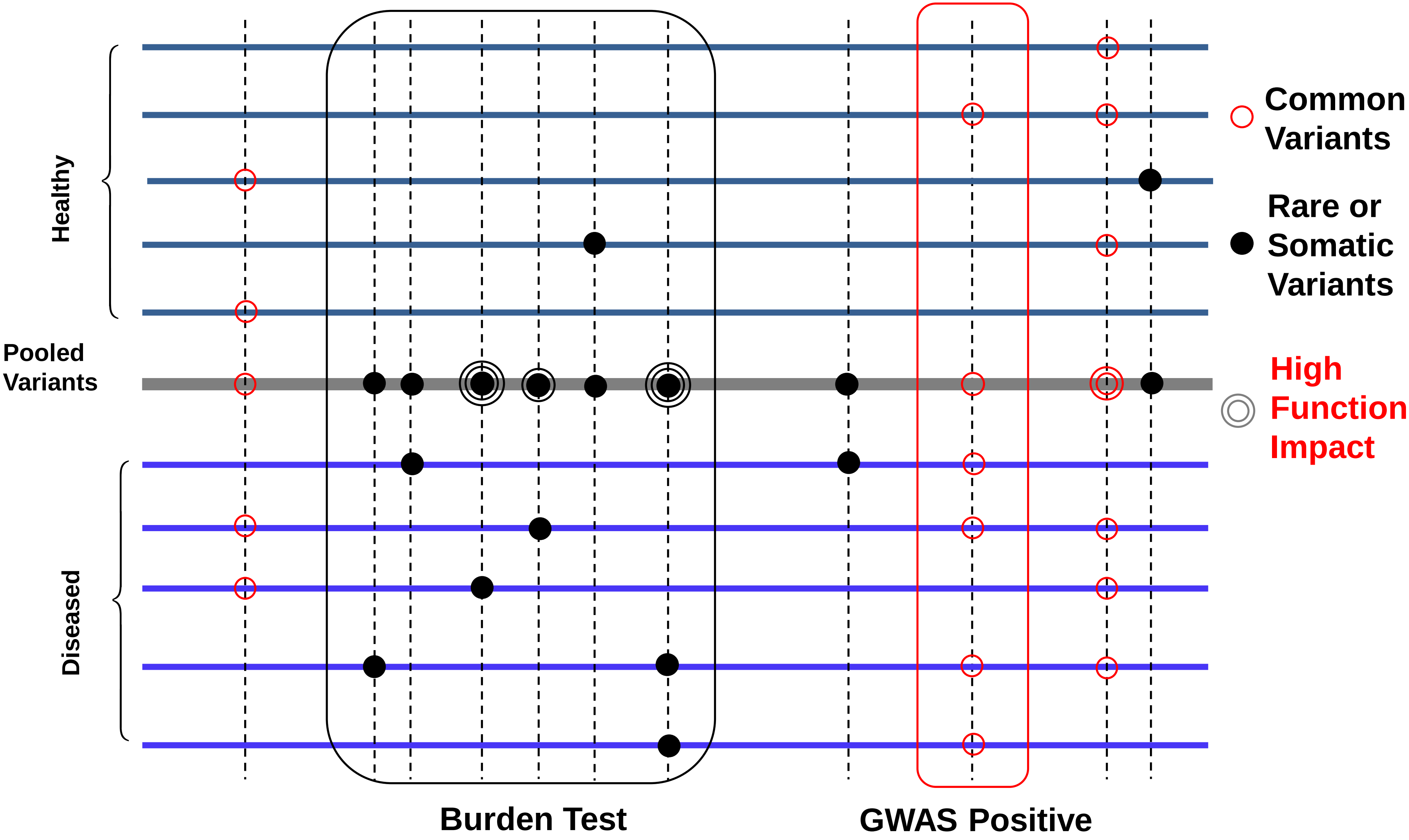
SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M



* 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



Healthy

Pooled Variants

Diseased

Burden Test

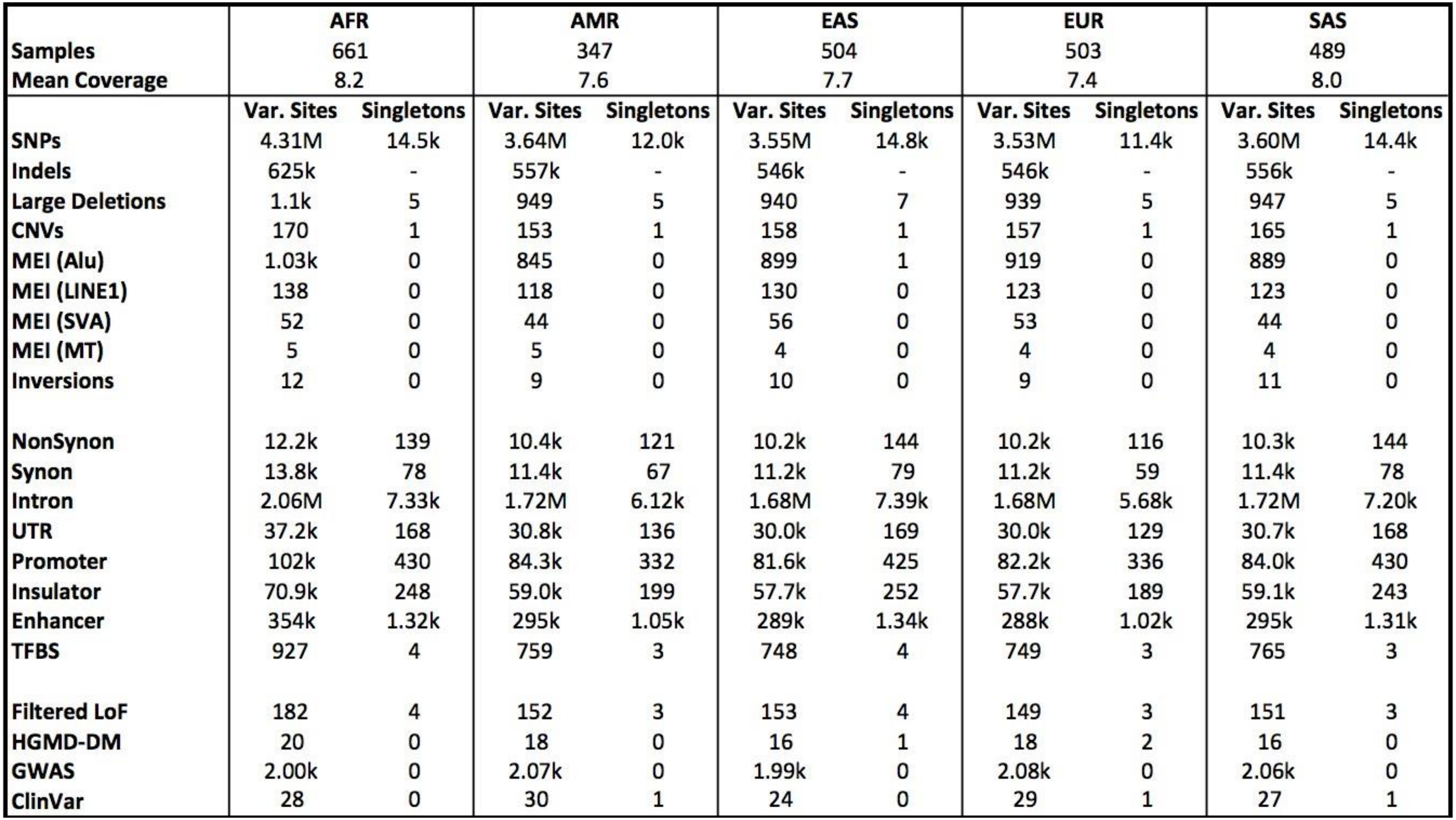
GWAS Positive

○ Common Variants

● Rare or Somatic Variants

◎ High Function Impact

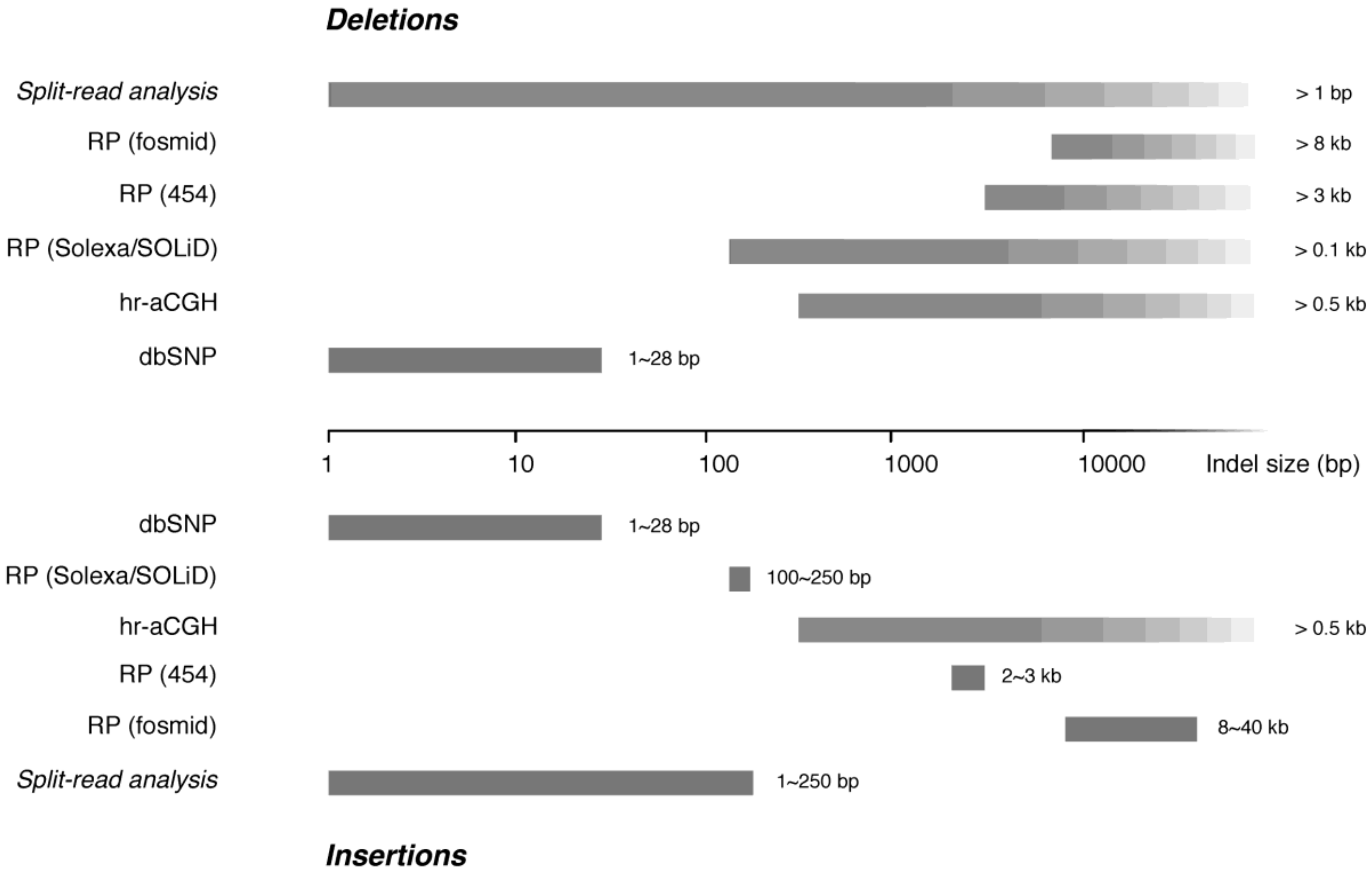
Phase 3: Median Autosomal Variant Sites Per Genome



	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

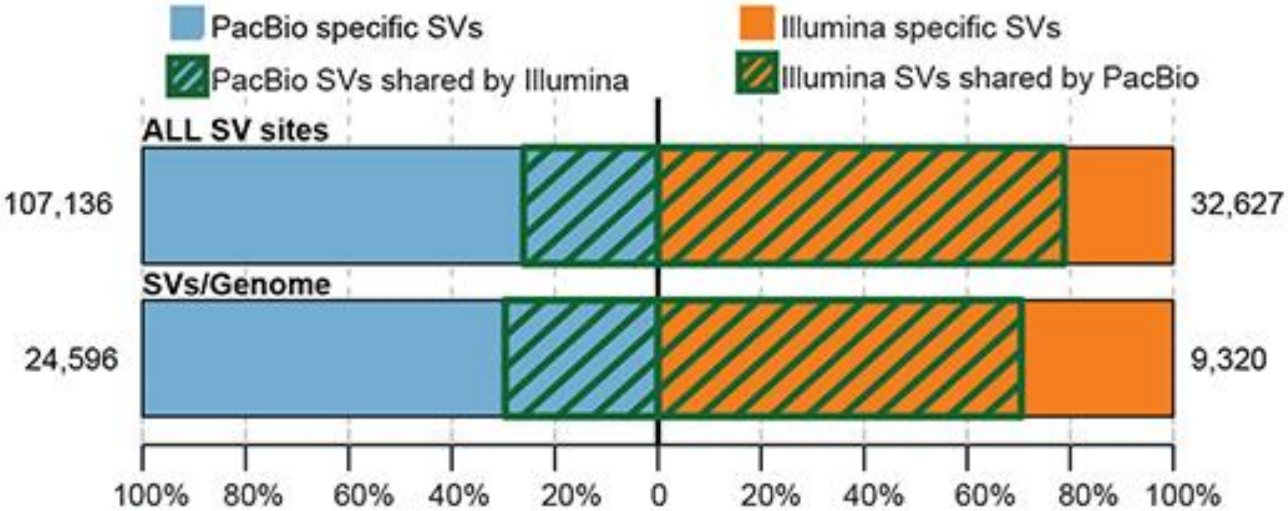
[3] 1000GP Consortium. Submitted to Nature, 2015.

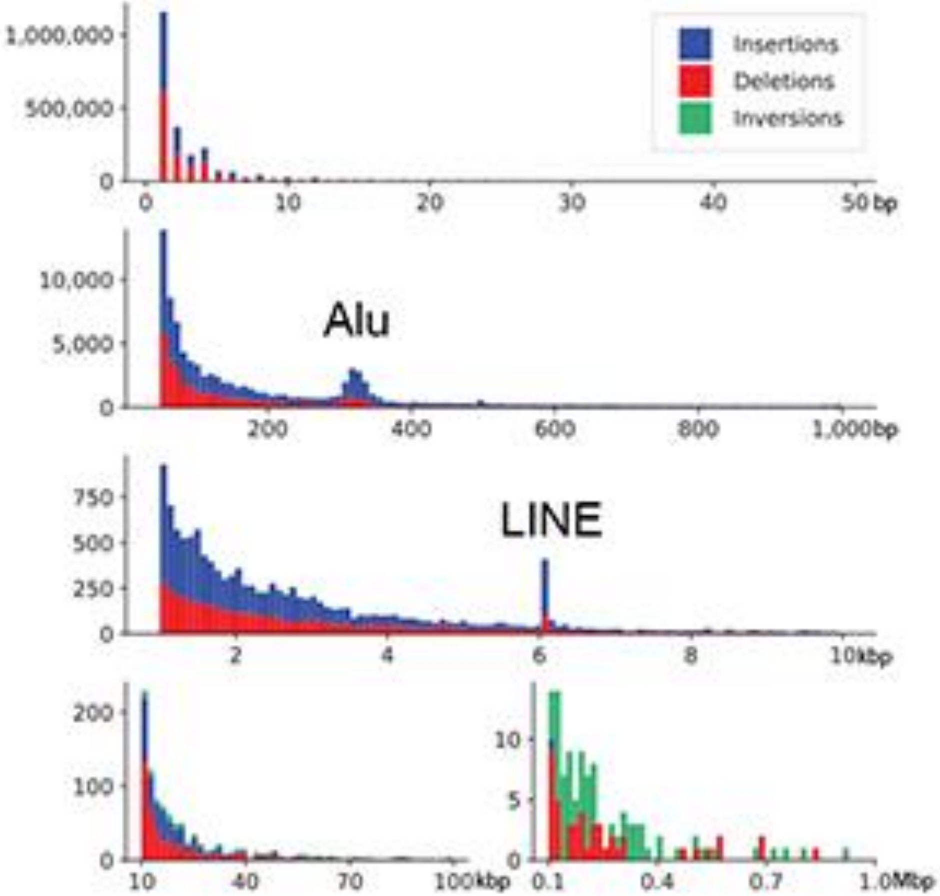
Different Approaches Work Differently on Different Events



[Zhang et al. ('11) *BMC Genomics*]

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



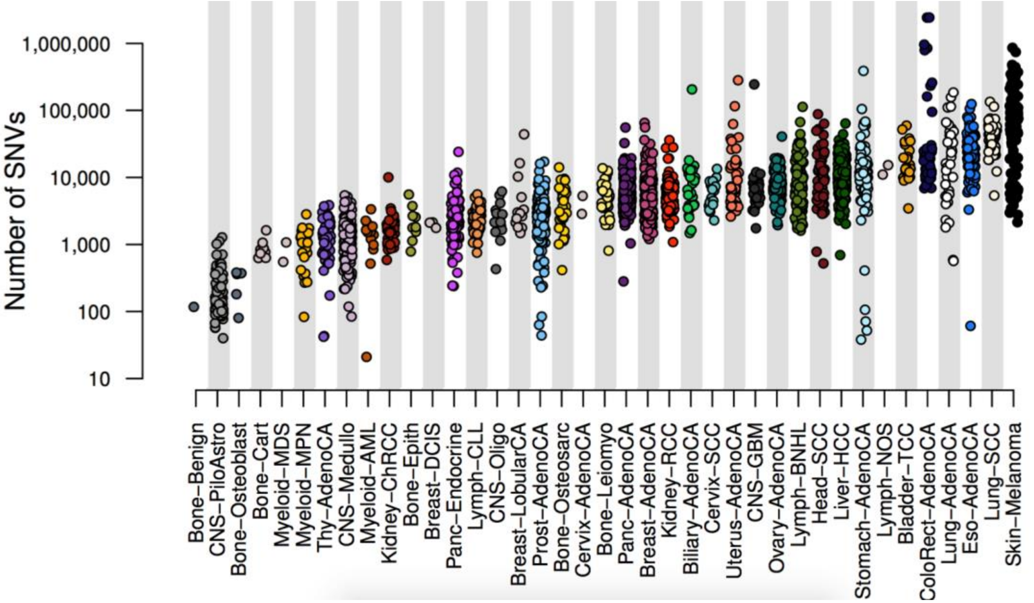


- On average, detected: 24,653 SVs, 794,406 indels, and 3,895,274 SNVs per diploid human genome

[Ebert et al. Science ('21)]

PCAWG summary

PCAWG : most comprehensive resource for cancer whole genome analysis



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

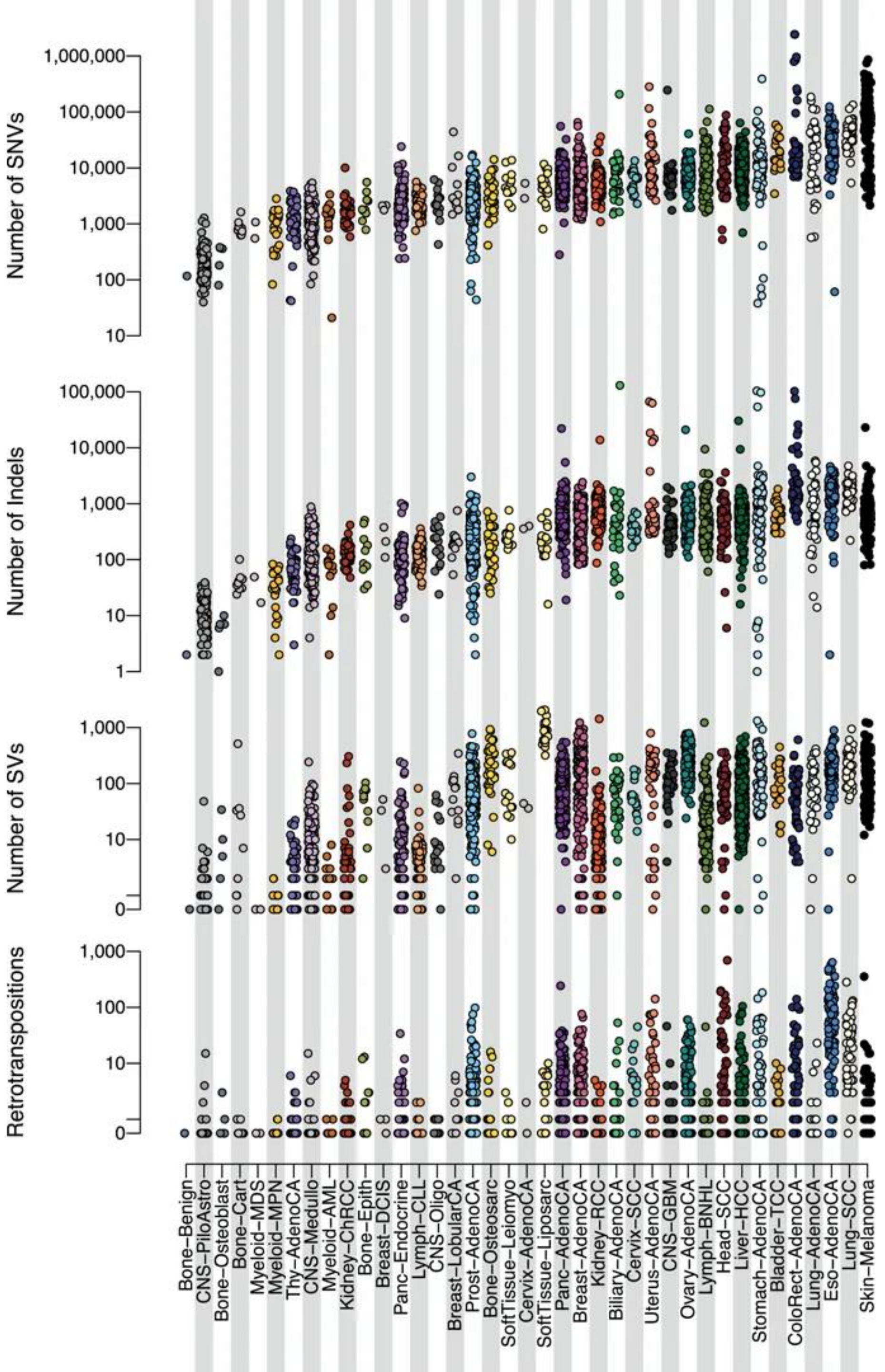
Project Goals:

- **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)