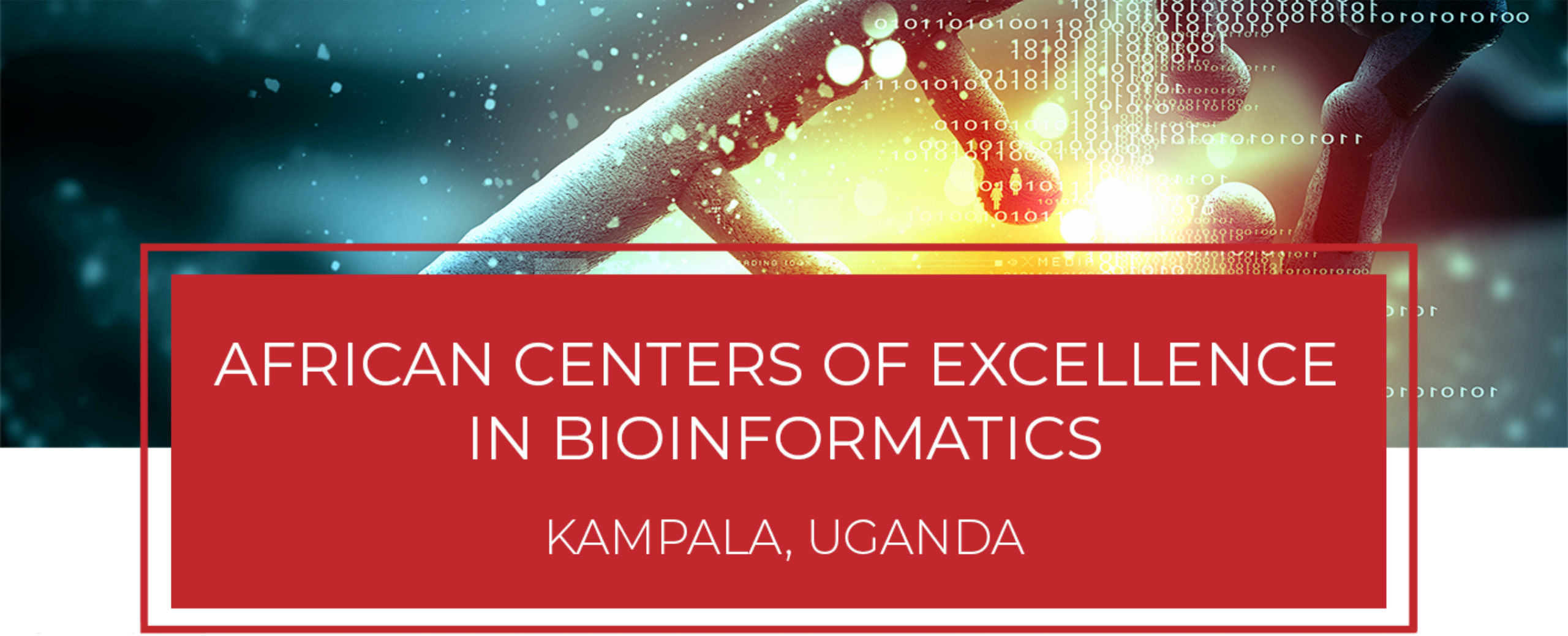


Public projects based on human NGS data

Today's Instructor

Eric Karlins, MS

Computational Genomics Specialist

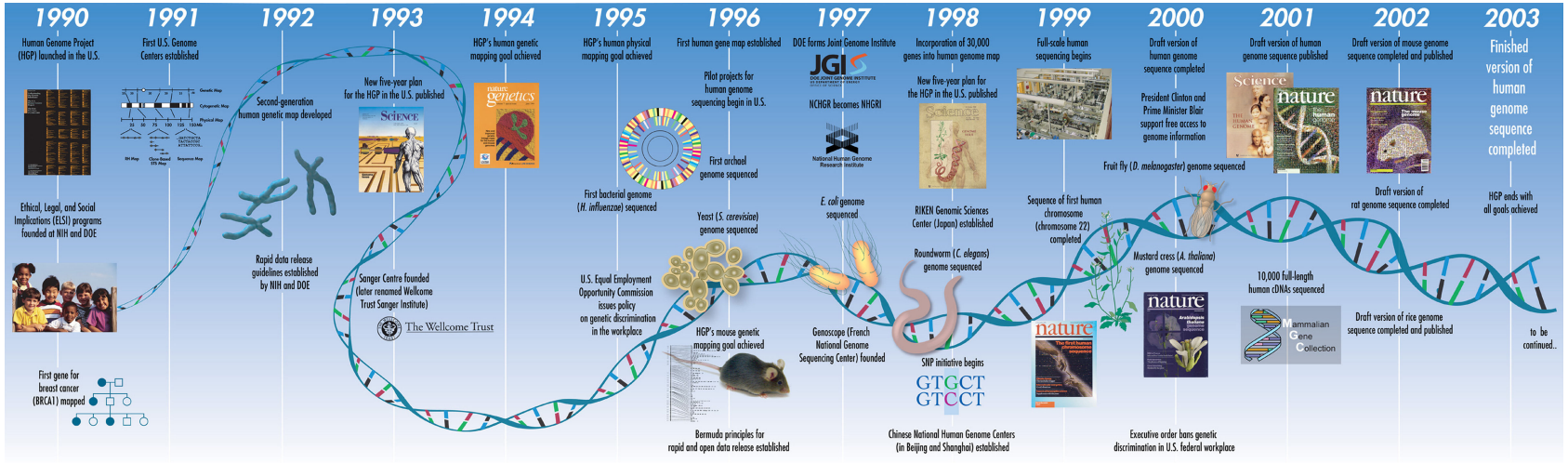
- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov

  Instructor:
    - karlinser@mail.nih.gov

Topics

- Public projects based on human NGS data
- Using population frequency data
- VCF file format
- Other variant annotations

The Human Genome Project

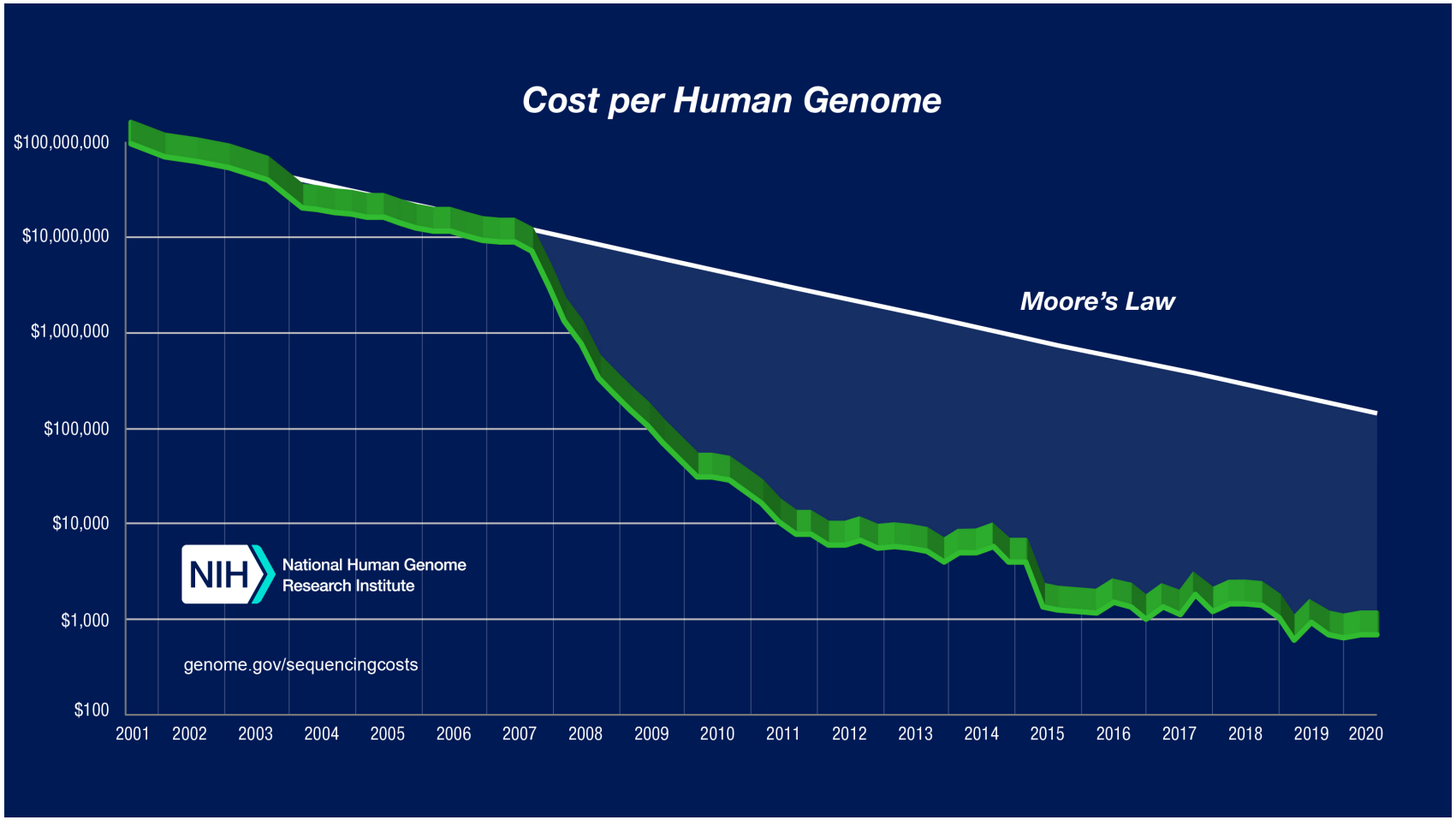


https://www.genome.gov/human-genome-project/Timeline-of-Events

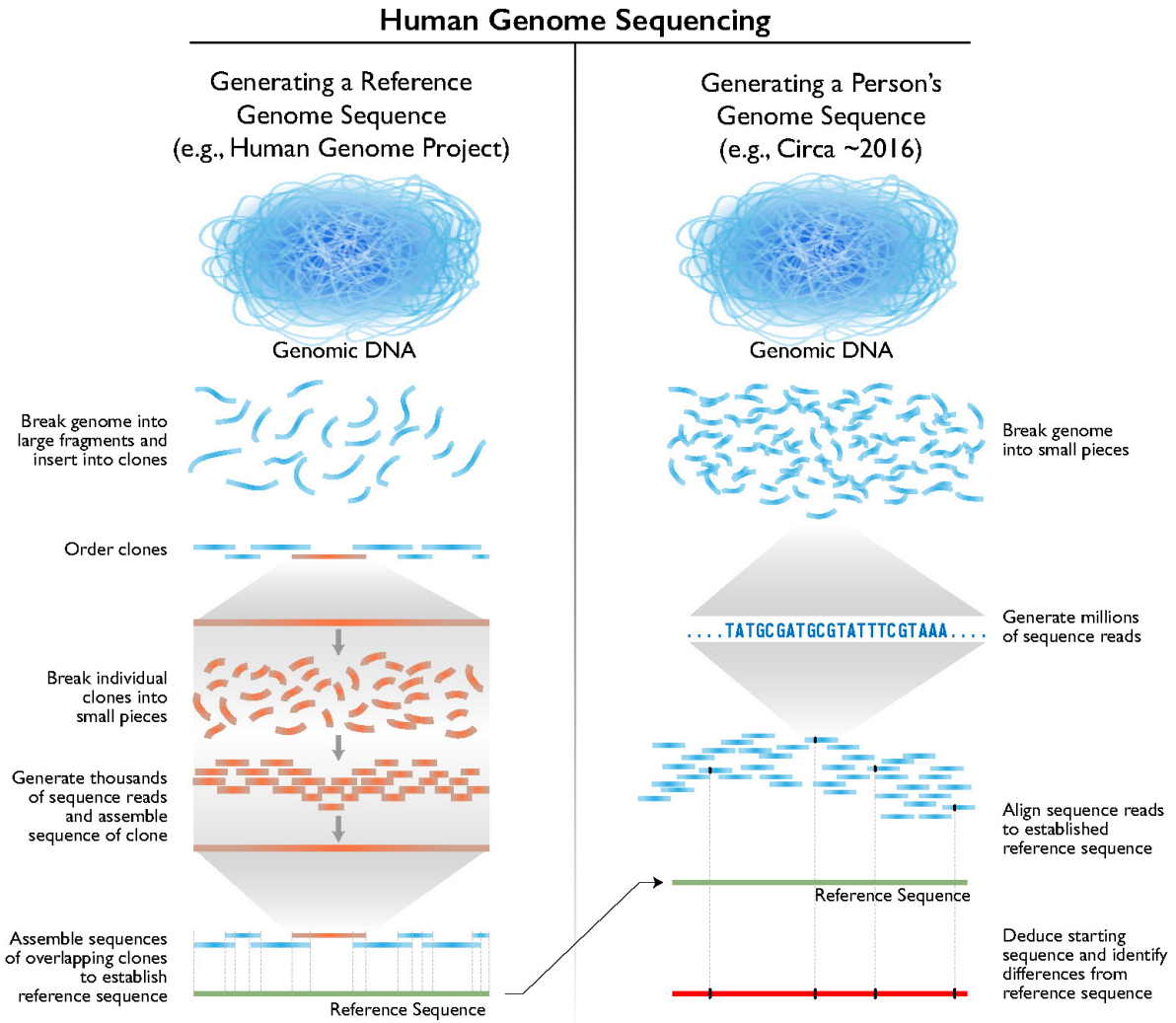
The Human Genome Project

- 13 years to complete
- Cost $3 Billion to complete
- Now a human genome can be sequenced in days at a cost of about $1000
- https://www.genome.gov/human-genome-project

Cost of Whole Genome Sequencing



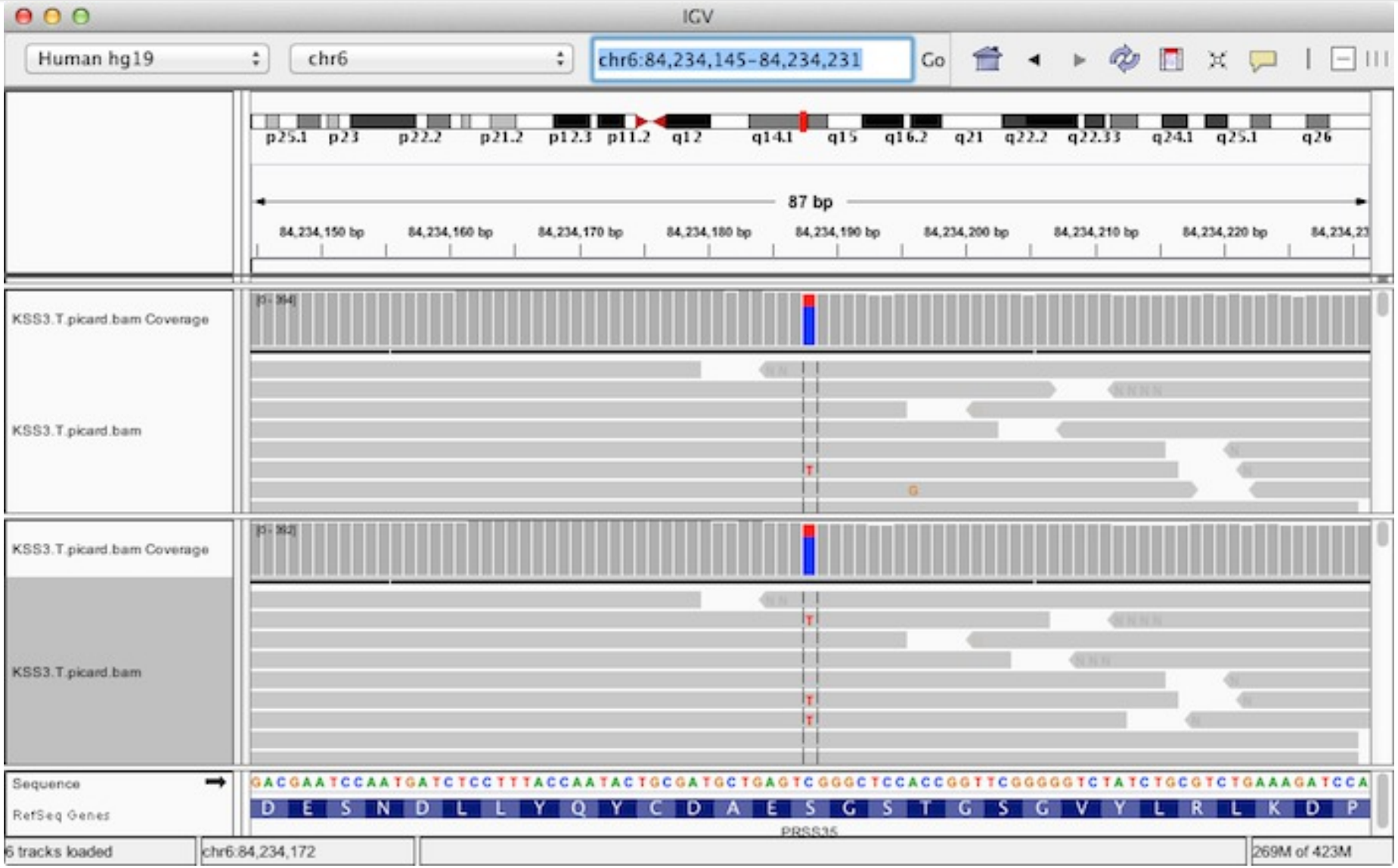
Sequencing Schematic



NGS Sequencing Basics

- Raw sequencing reads come off the sequencer as ″fastq″ files.
- Fastq files just contain the nucleotide sequence and some quality information.
- Fastq files are aligned to a reference genome to make BAM files.
- Most common experiments are Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS)

Aligned Reads are saved as BAM files



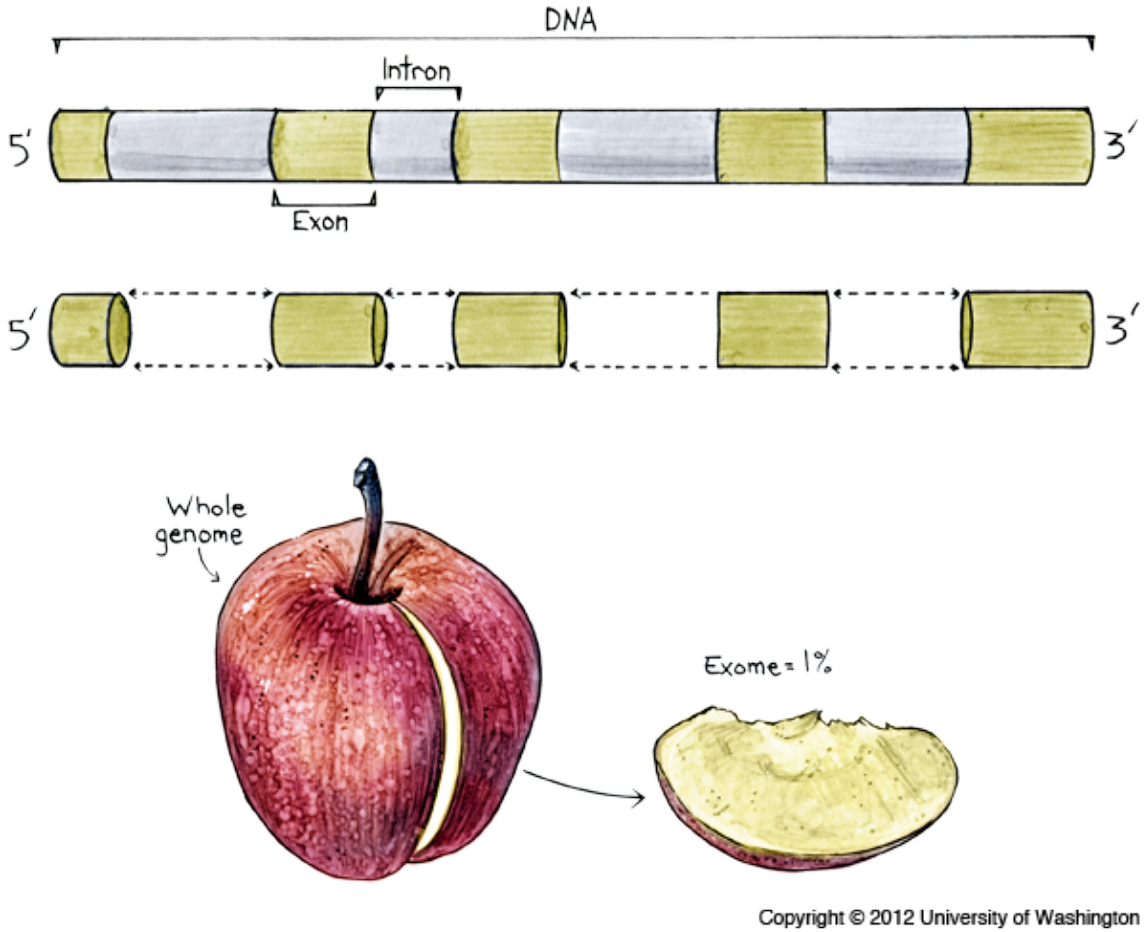
BAM file

- SAM stands for Sequence Alignment Map
- BAM is a binary SAM file
- Meta-data is stored in the BAM header
- Sequencing data is stored as reads in the BAM file
- BAM file is indexed for quick retrieval of a region

VCF file format

- Variant calls from NGS experiments stored in VCF file
- BAM files are used as input to variant caller to create VCF
- Many samples and many variants can be stored in one VCF file
- Usually Single Nucleotide Variants (SNVs) or small indels, but can also be larger Structural Variants (SVs)
- The Variant Call Format (VCF) Version 4.2 Specification
- https://samtools.github.io/hts-specs/VCFv4.2.pdf

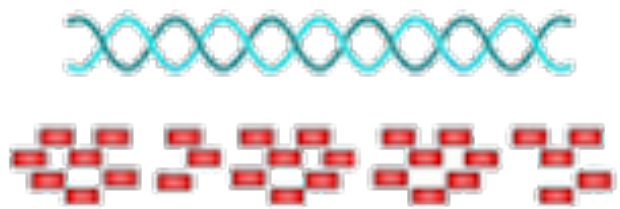
WGS vs WES



https://www.my46.org/intro/whole-genome-and-exome-sequencing

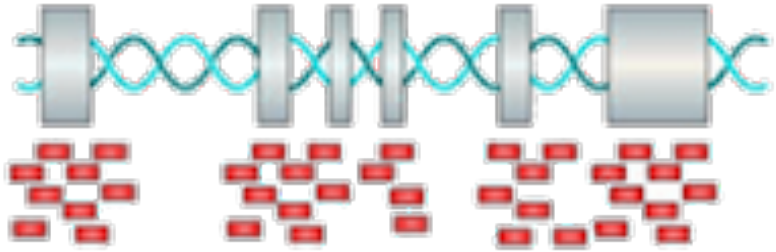
Common NGS methods

Whole genome sequencing



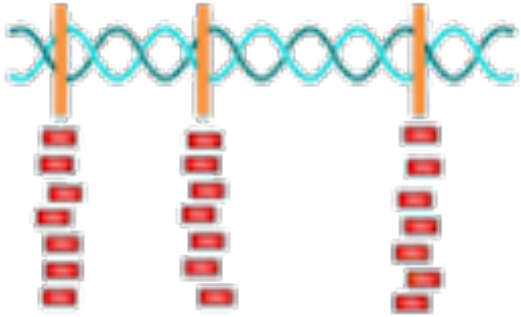
- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

Whole exome sequencing



- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

Targeted sequencing



- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

http://www.genomesop.com/somatic-mutations/

International HapMap Project

The DNA sequence of any two people is 99.5 percent identical. The variations, however, may greatly affect an individual's disease risk. Sites in the DNA sequence where individuals differ at a single DNA base are called single nucleotide polymorphisms (SNPs). Sets of nearby SNPs on the same chromosome are inherited in blocks. This pattern of SNPs on a block is a haplotype. Blocks may contain a large number of SNPs, but a few SNPs are enough to uniquely identify the haplotypes in a block. The HapMap is a map of these haplotype blocks and the specific SNPs that identify the haplotypes are called tag SNPs.

https://www.genome.gov/10001688/international-hapmap-project

1000 Genomes Project

https://www.internationalgenome.org/

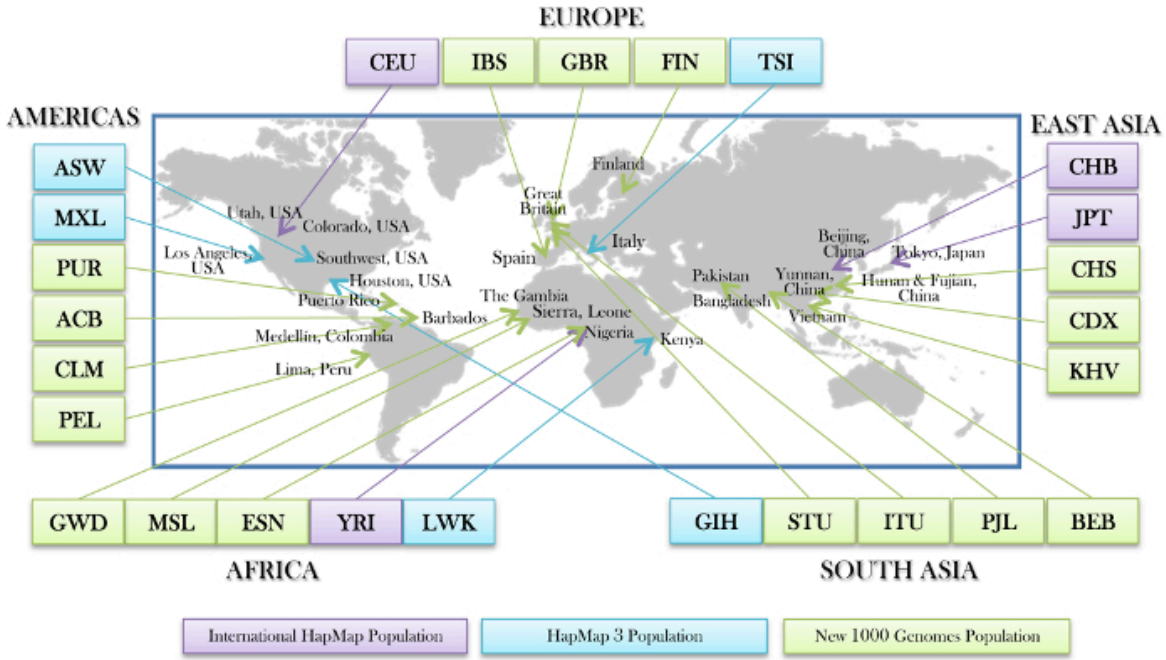
1000 Genomes Project

- 2504 samples from distinct populations around the globe
- Now all have high coverage WGS
- Raw data and variant calls available for download
- Frequencies available through Ensembl browser

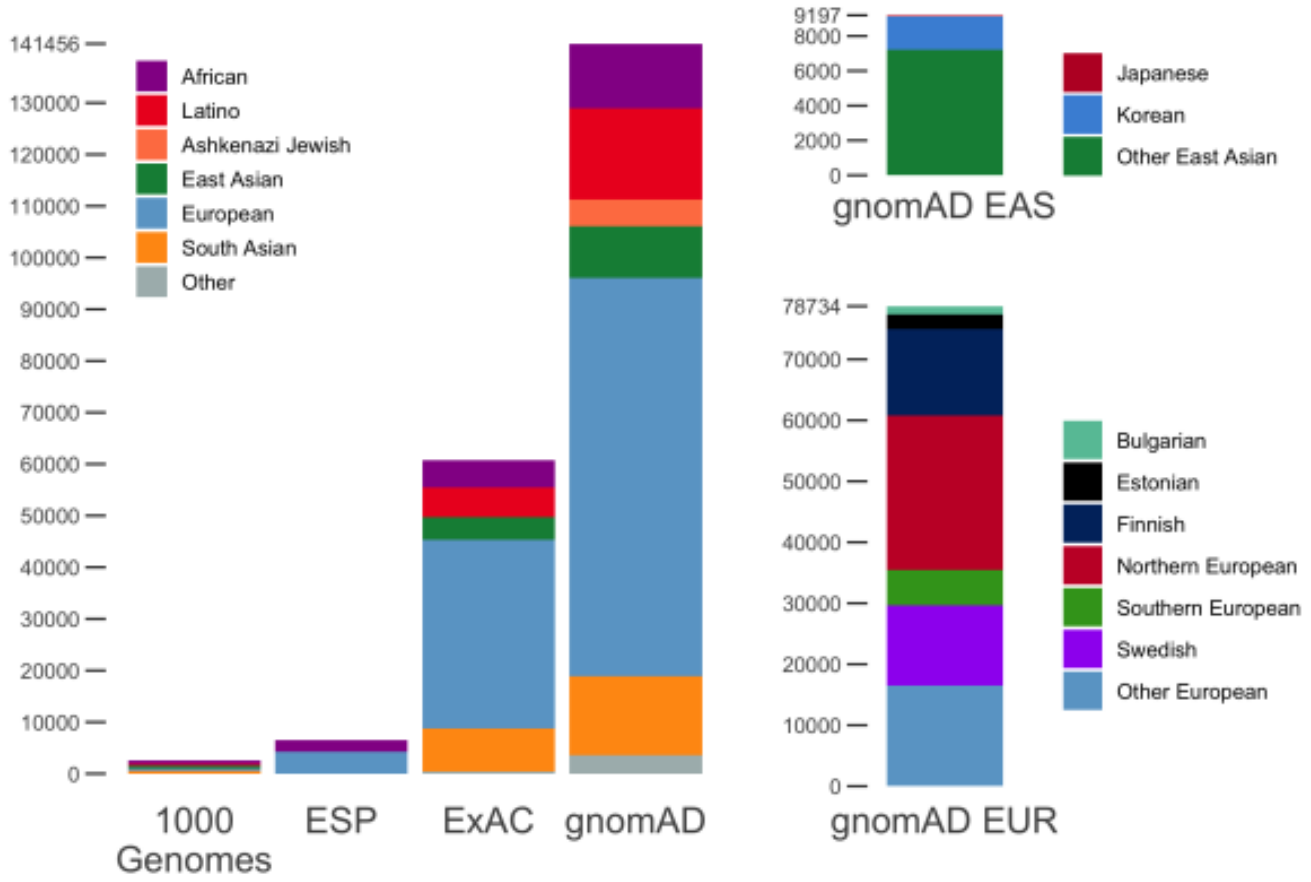
DNA and cell lines available to order

Cell lines and DNA from Coriell for 1000 genomes samples

All the samples from the 1000 Genomes Project are available as lymphoblastoid cell lines (LCLs) and LCL derived DNA from the Coriell Cell Repository as part of the NHGRI Catalog. In addition, Standard Population DNA Panels for the 1000 Genomes and HapMap projects are available at $1000 or less each (see panel identifiers below).



WES and WGS databases

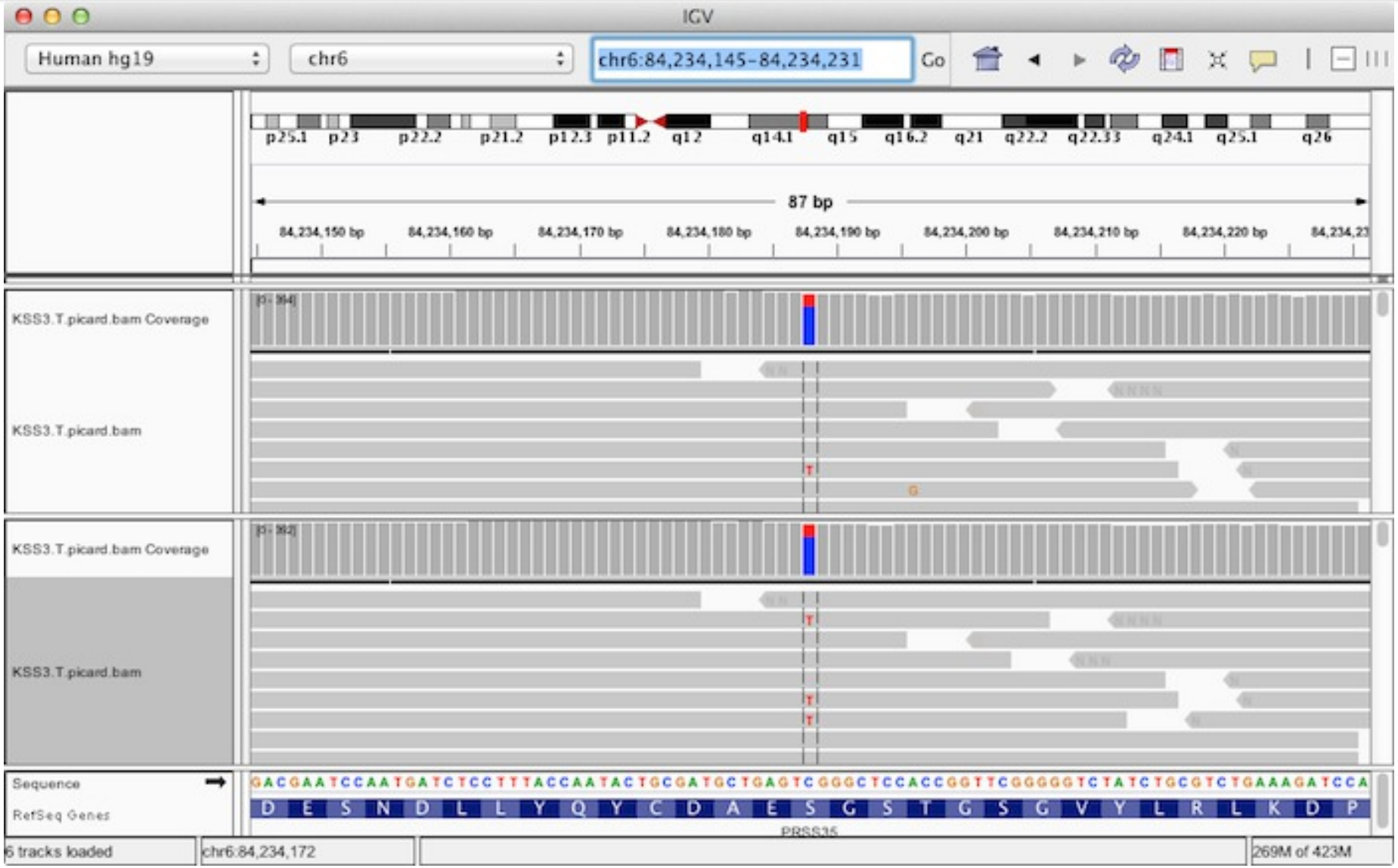


https://macarthurlab.org/2018/10/17/gnomad-v2-1/

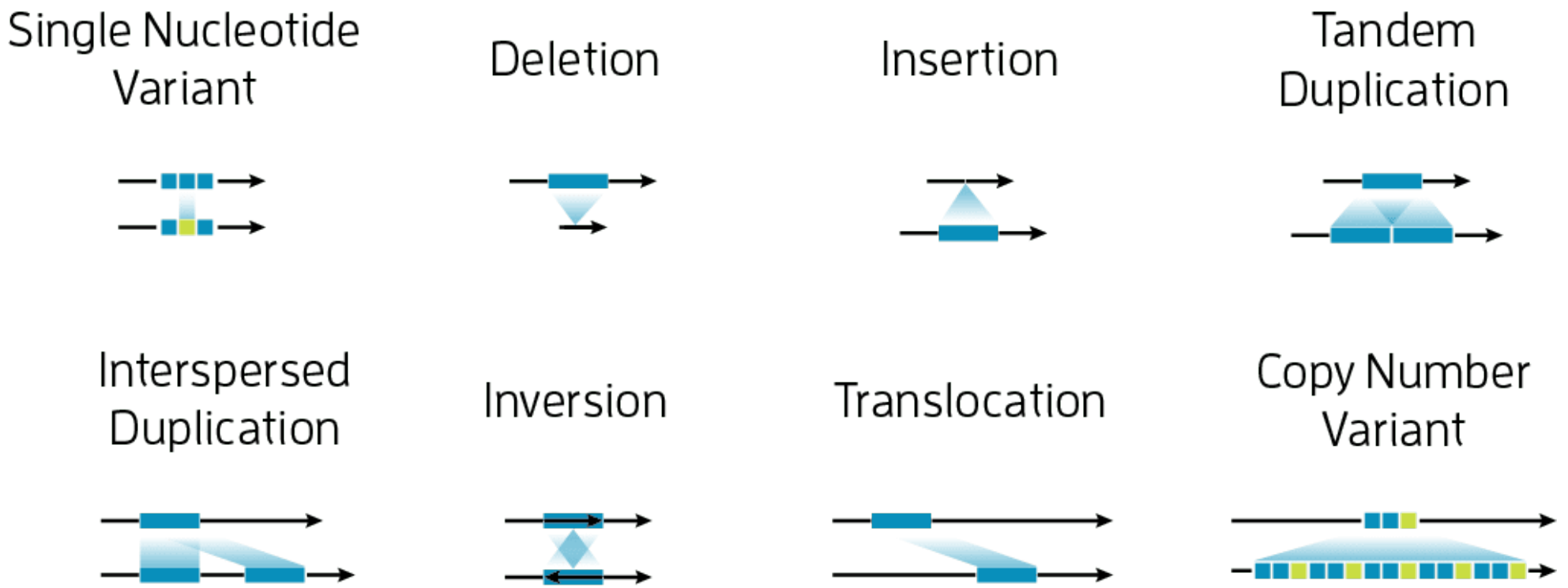
gnomAD

- https://gnomad.broadinstitute.org/
- The v2 data set (GRCh37/hg19) includes 125,748 exome sequences and 15,708 whole-genome sequences
- The v3.1 data set (GRCh38) spans 76,156 genomes
- Samples from many populations, though most are European
- Only summary statistics available, no individual level data.

Aligned Reads are saved as BAM files

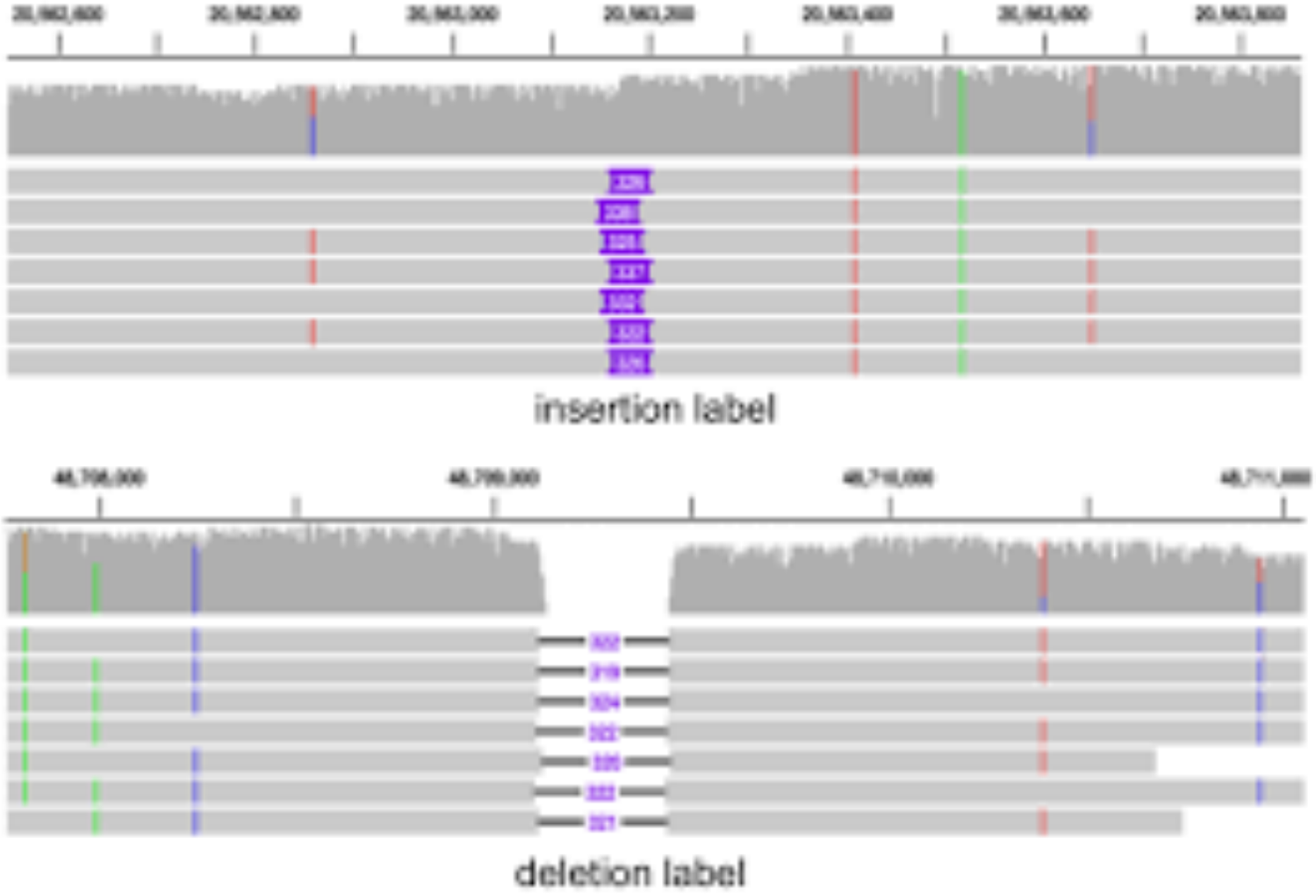


Variants that can be discovered from NGS



Types of Variants

Insertion/deletion (indel)



VCF file format

- Variant calls from NGS experiments stored in VCF file
- BAM files are used as input to variant caller to create VCF
- Many samples and many variants can be stored in one VCF file
- Usually Single Nucleotide Variants (SNVs) or small indels, but can also be larger Structural Variants (SVs)
- The Variant Call Format (VCF) Version 4.2 Specification
- https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF Example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Variant Annotation Tools

- Variant Effect Predictor (VEP)
- https://useast.ensembl.org/Tools/VEP
- SnpEff
- https://pcingola.github.io/SnpEff/
- ANNOVAR
- https://annovar.openbioinformatics.org/en/latest/