BF528 - Whole Genome Sequencing and Genomic Variation

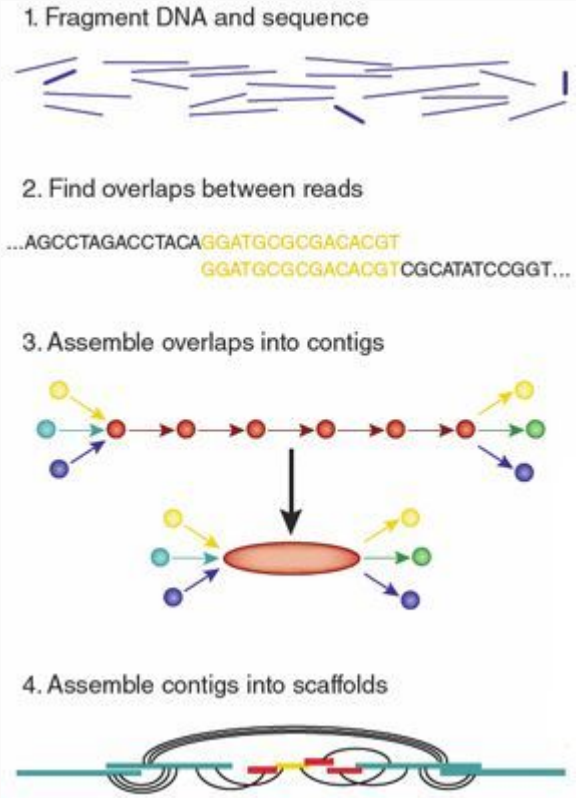
Whole Genome/Exome Sequencing

Whole Genome Sequencing

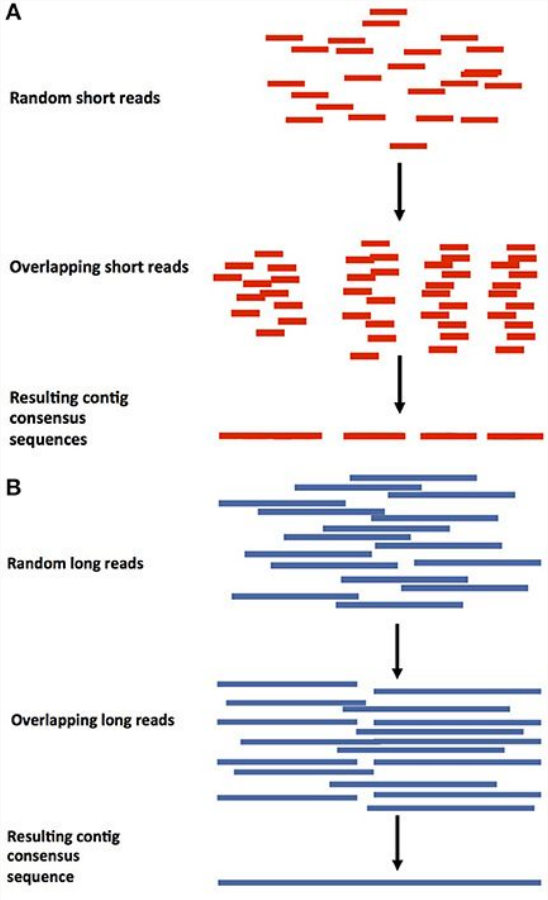
- **W**hole **G**enome **S**equencing (WGS)
- Generate enough reads to attain:
  - >95% coverage of source genome
  - >30x average depth
- Two strategies:
  - **De novo**: assemble reads into a new sequence
  - **Re-sequence**: refine an existing reference sequence

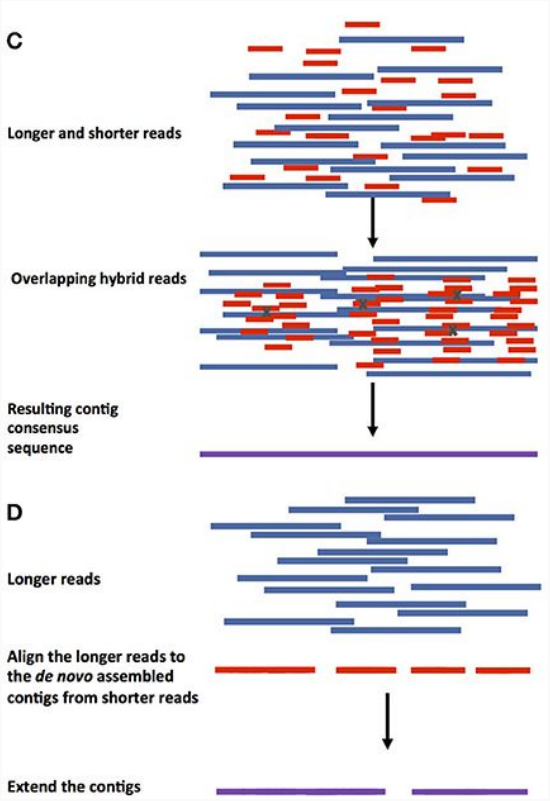
De novo assembly

- **Genome Assembly** - Create new reference ‘from scratch’
- Examine reads for overlapping sequence
- **Contig** - longer assembled sequence from short reads
- **Scaffold** - assembled contigs
- **Chromosome** - assembled scaffolds
- Assembly from short reads is hard



De novo assembly: greedy algorithm

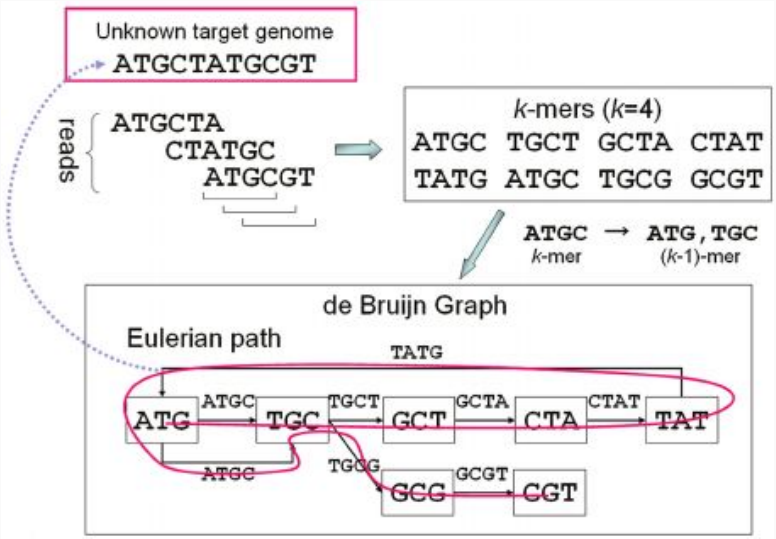




Kyriakidou, Maria, Helen H. Tai, Noelle L. Anglin, David Ellis, and Martina V. Strömvik. 2018. "Current Strategies of Polyploid Plant Genome Sequence Assembly." *Frontiers in Plant Science* 9 (November): 1660.

De novo assembly: graph-based

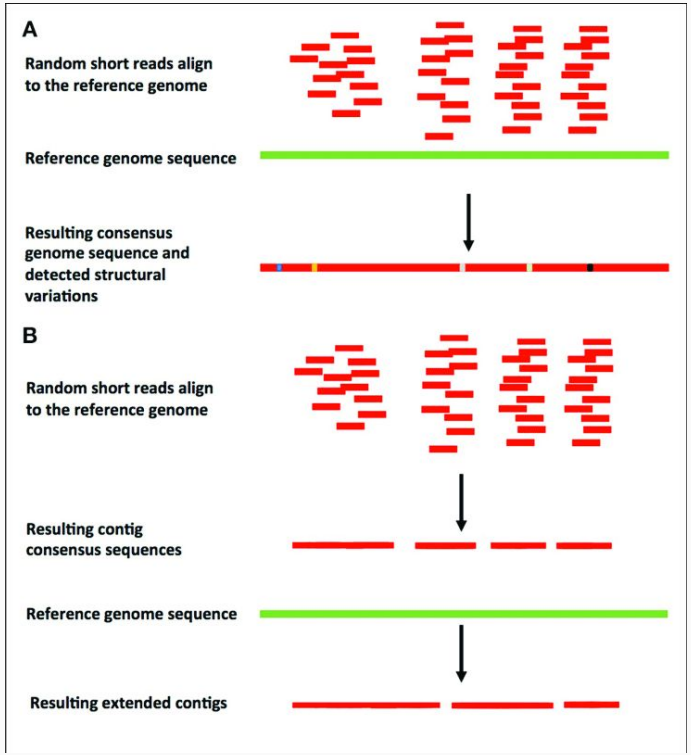
- Greedy assembly creates linear sequences
- Vulnerable to finding local optima
- Graph representation considers all sequence content simultaneously
- Graph data structure can encode variability (e.g. insertions, SNPs)
- Computationally much more expensive

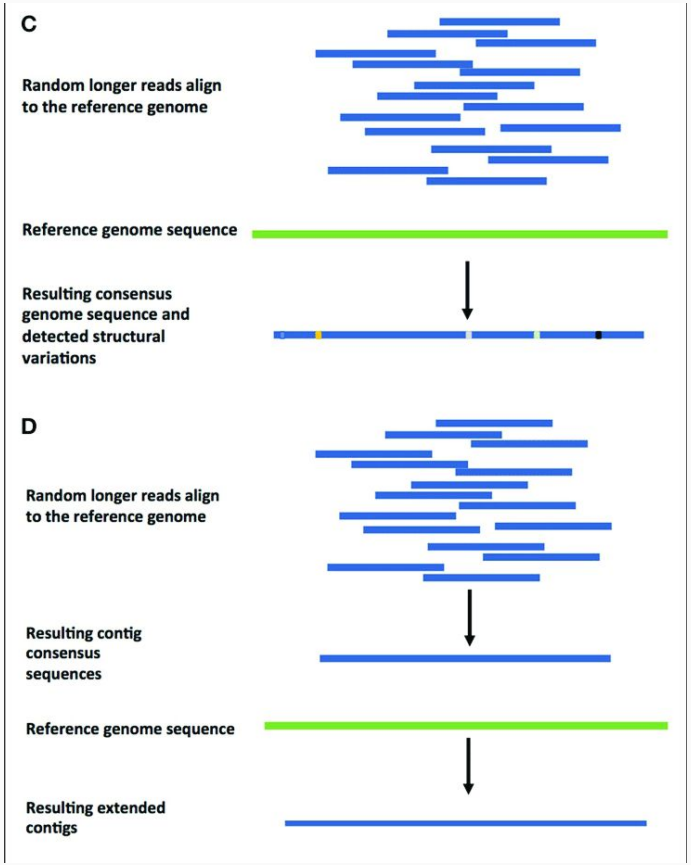


Reference guided genome assembly

- Refine existing reference with new sequence
- Can discover:
  - New structural variants
  - Novel insertions/alternate haplotypes or scaffolds
  - Polymorphisms
- Faster, easier than de novo assembly
- More sensitive to existing biases in reference

Reference guided genome assembly





Kyriakidou, Maria, Helen H. Tai, Noelle L. Anglin, David Ellis, and Martina V. Strömvik. 2018. "Current Strategies of Polyploid Plant Genome Sequence Assembly." *Frontiers in Plant Science* 9 (November): 1660.

Human genome reference

Global statistics

Number of regions with alternate loci or patches	182
Total sequence length	3,234,834,689
Total assembly gap length	243,146,473
Gaps between scaffolds	271
Number of scaffolds	463
Scaffold N50	44,983,201
Scaffold L50	22
Number of contigs	705
Contig N50	38,440,852
Contig L50	25
Total number of chromosomes and plasmids	25

Assembly GRCh37 patch 13 (hg19)

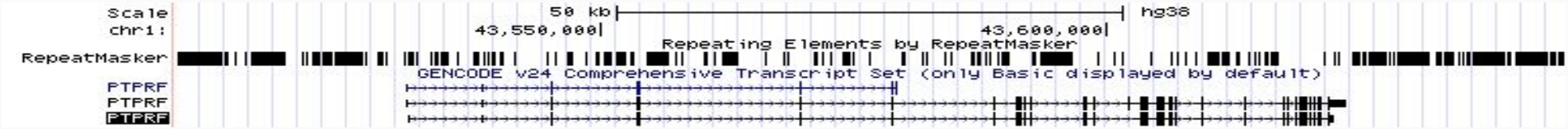
Global statistics

Number of regions with alternate loci or patches	317
Total sequence length	3,257,347,282
Total assembly gap length	161,368,351
Gaps between scaffolds	349
Number of scaffolds	875
Scaffold N50	59,364,414
Scaffold L50	17
Number of contigs	1,536
Contig N50	56,413,054
Contig L50	19
Total number of chromosomes and plasmids	25

Assembly GRCh38 patch 12 (hg38)

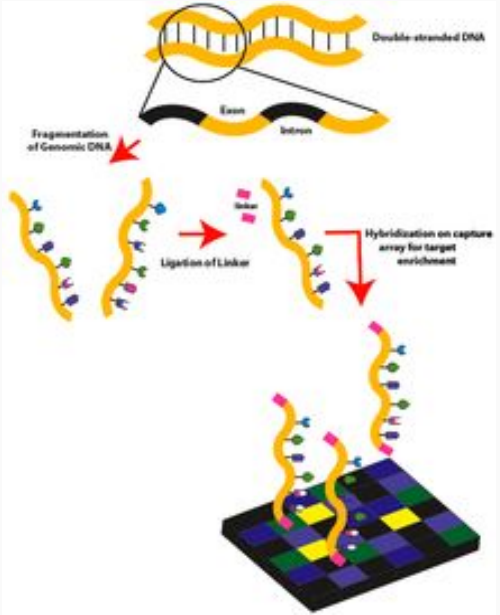
Whole Exome Sequencing

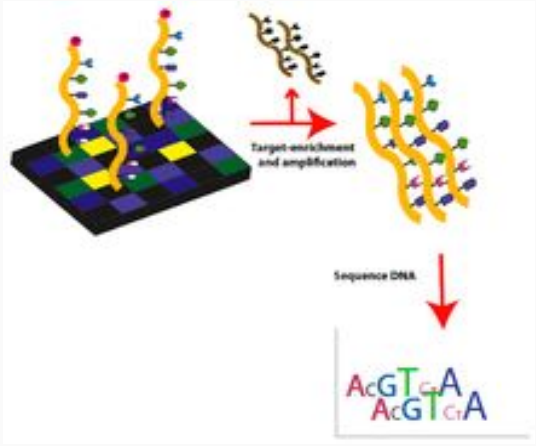
- **W**hole **E**xome **S**equencing (WES)
- Exons 1%-2% of human genome sequence
- Pre-select reads that map to exons
- Sequence to much greater depth than WGS
- Identify coding variants



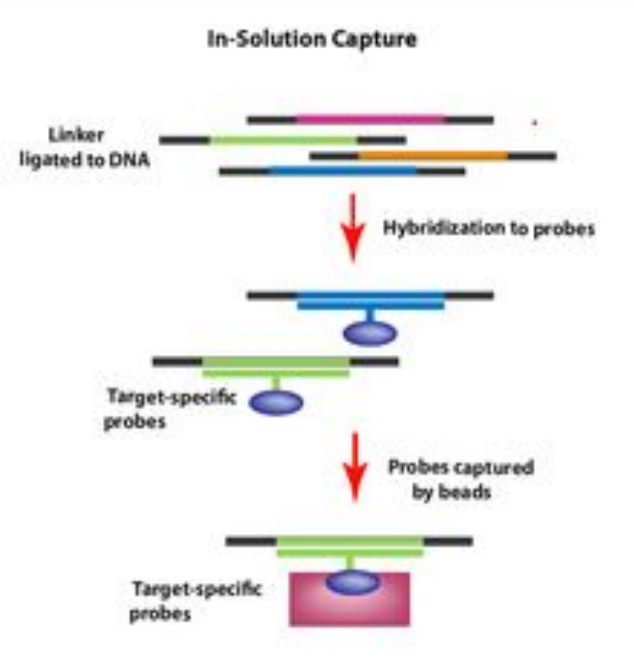
Exome Sequence Selection

Array-based capture





In-solution

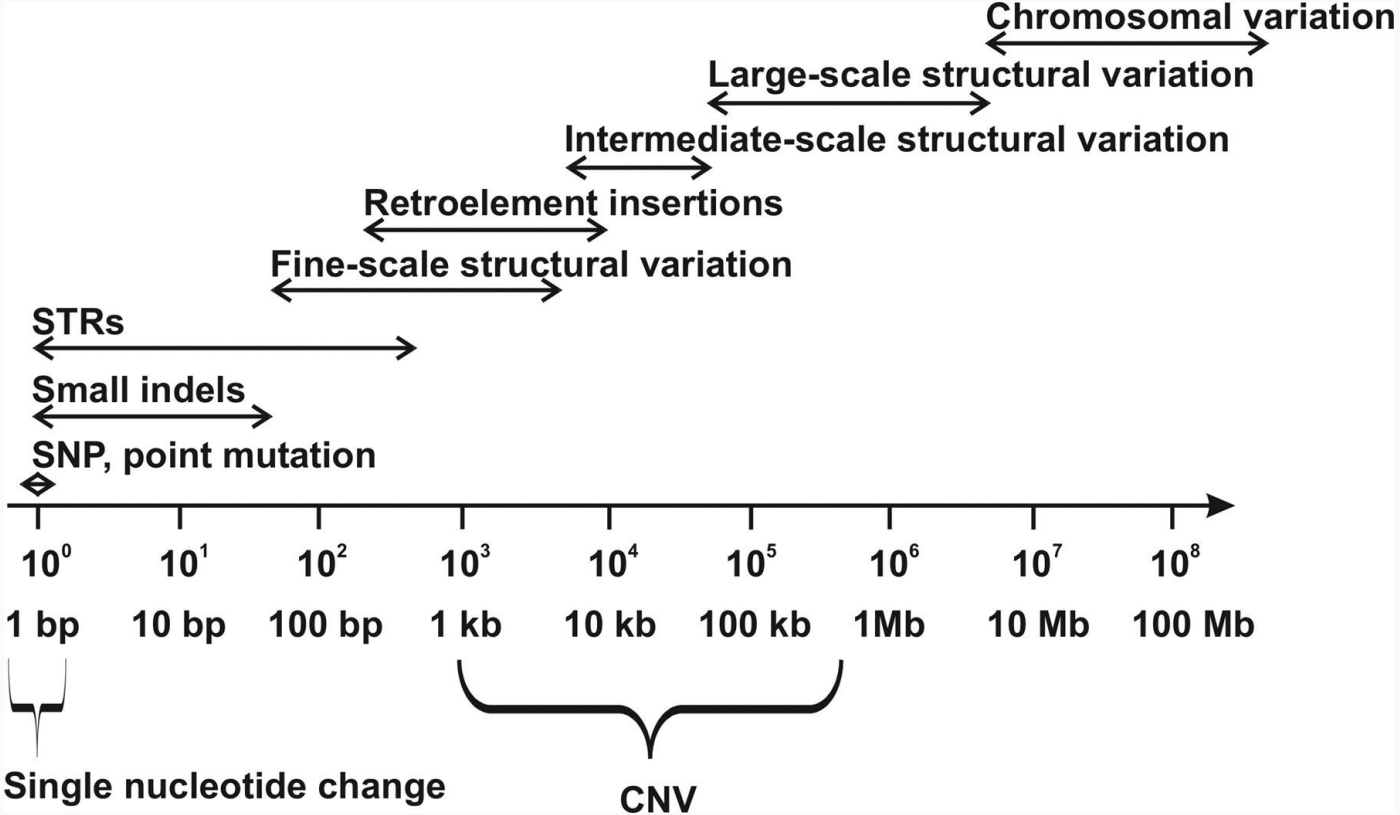


Genomic Variation

Genomic Variants

- Individual genomes from same species vary
- WGS/WES compared with reference can identify differences
- **Variant**: sequence that varies within species
- Two general types:
  - Small: <50 bp, single nucleotide polymorphisms (SNPs), indels
  - Large: >50 bp, copy number variations, duplications, deletions, translocations, inversions

Genomic variants



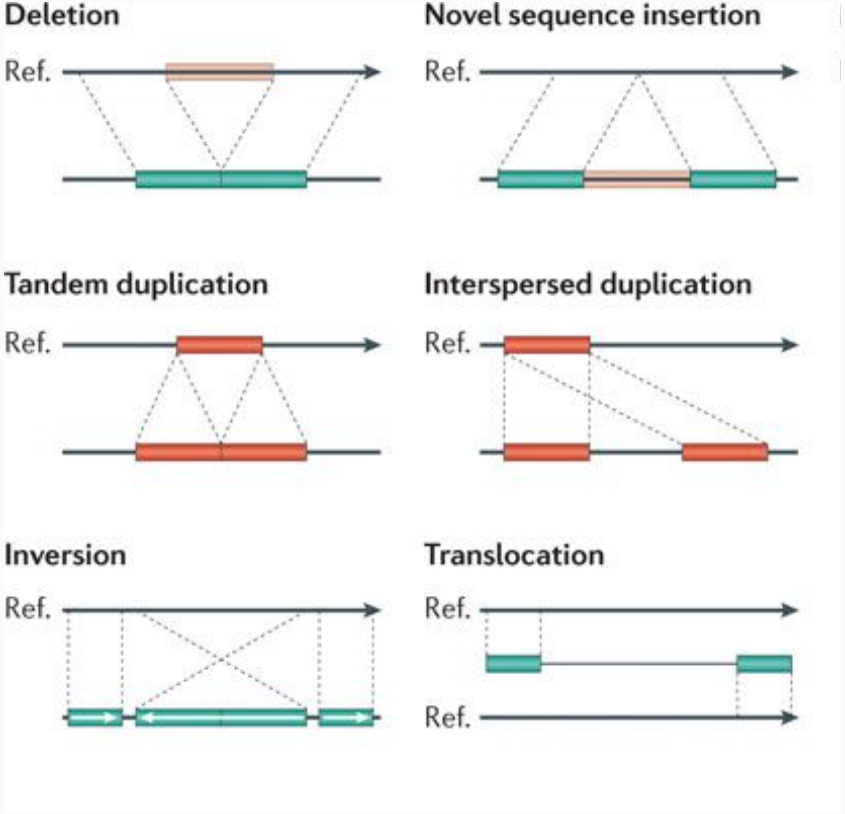
Point mutations

Single Nucleotide Polymorphisms (SNPs)

- Most commonly studied type of variant
- Types of single nucleotide alterations:
  - **Variant (SNV)** - any single mutation
  - **Polymorphism (SNP)** - SNV of appreciable frequency in population (e.g. >1%)
- Typically base changes, e.g. A → C
- SNPs (usually) indicate shared ancestry
- May suggest disease mechanism

Structural variation

- Deletions - sequence missing
- Insertions:
  - Novel - new sequence added
  - Mobile-element - copied/moved from elsewhere in genome
- Duplications:
  - Tandem - consecutive
  - Interspersed - non-consecutive
- Inversion - segment is reversed
- Translocation - segment moves



Genotyping

- **Genotype**: an individual's variant(s)
- **Phenotype**: an individual's physical form
- **De novo** variant calling/detection: given genomic reads and a reference, find all the variants
- **Genotyping**: examine individual for *a priori* variants at known locations
- Can use arrays (i.e. SNP Chip) or WGS/WES

Human Genomic Variants

- Humans have diploid genome
  - i.e. 2 copies of each gene
- 660 million annotated human SNPs
- 113 million of those have been validated
- Every human has on average:
  - A variant every 1kb
  - 2-3 million SNPs
- dbSNP - NCBI repository for SNP info

Human Genomic Terminology

- **Allele**: sequence containing a variant
- **Homozygous variant**: both same allele
  - i.e. either same variant or same as reference
- **Heterozygous variant**: different alleles
  - i.e. one is variant, one is reference/different variant

Human Genomic Terminology

- For coding (exonic) variants:
  - **Synonymous or sense**: variant does not change amino acid sequence
  - **Non-synonymous or mis-sense**: variant causes amino acid change
  - **Non-sense**: causes early termination of protein by introducing stop codon
  - **Frameshift**: insertion or deletion causes complete recoding of downstream proteins

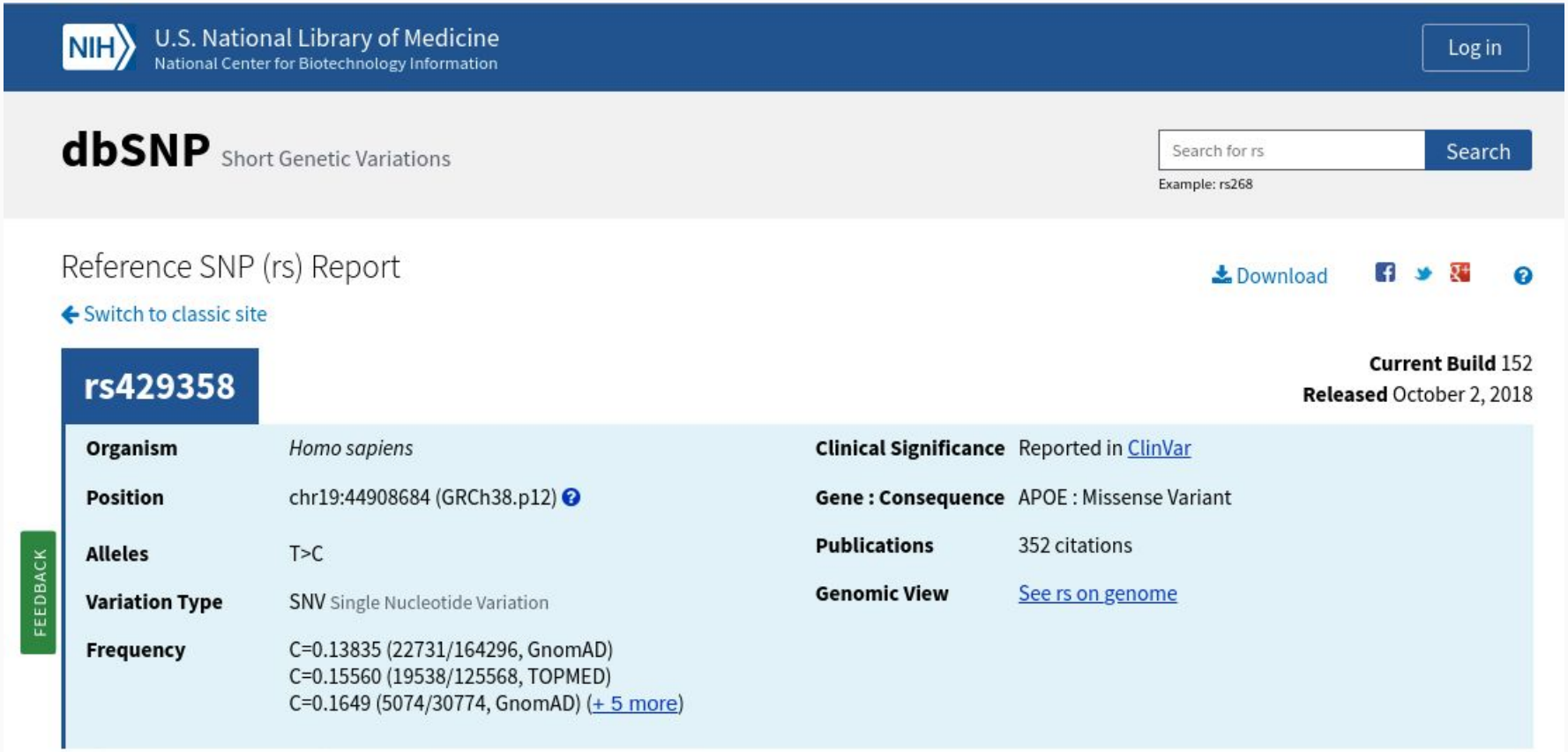
Genomic Variant Terminology

- **Germline**: Inherited from parents
  - e.g. blue eyes, familial disease risk
- **Somatic**: Acquired during life
  - e.g. tumor vs normal tissue
- **Allele frequency**: how common is a given variant in some population, e.g.:
  - 1% of human population
  - 30% within people with some disease

Genomic Variant Encoding

- **Reference allele** - reference sequence
- **Alternate allele** - variant sequence
- We show alleles as:
  - **0/0** both reference allele
  - **0/1** one reference allele, one alternate
  - **1/1** both non-reference allele, homozygous
  - **1/2** both non-reference, heterozygous

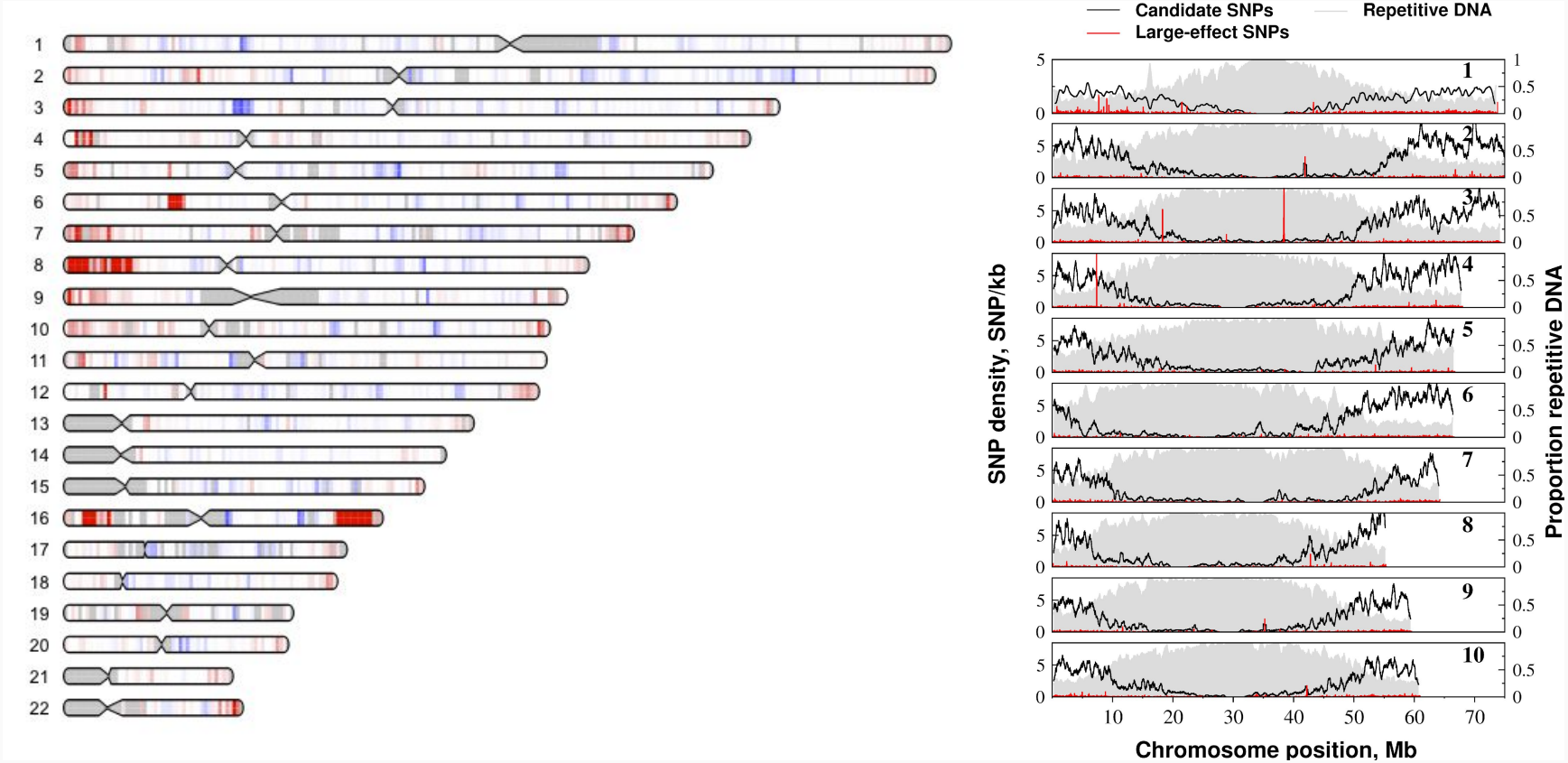
dbSNP - NCBI SNP Database



Variants Smaller Than A Read

- Finding SNPs, indels almost a solved problem
- SNPs called are 95% accurate
  - i.e. with sufficient coverage
- Structural variants cause false positives
  - Duplications, somatic mutations may cause 3 or more alleles to be observed

SNP and indel density is non-random



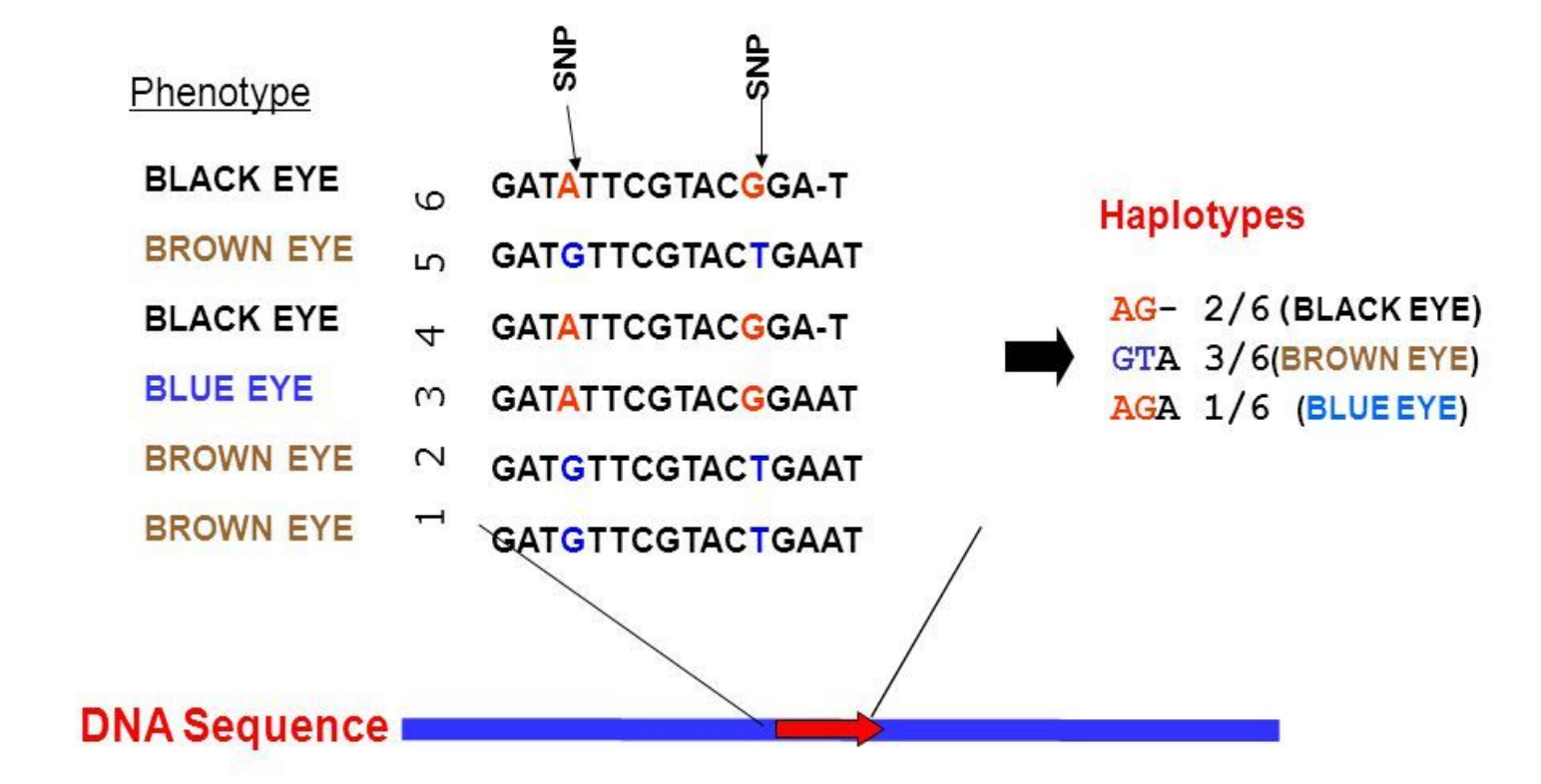
Variants Larger Than A Read

- **Structural Variation (SV)**
- Two types:
  - Balanced - Do not change amount of DNA
  - Copy Number Variants (CNV) - Change amount of DNA
- Scales:
  - Mini (hundreds of basepairs)
  - Macro (visible by a microscope) variants
- Much harder to find (especially balanced)
- Non-random: SV ‘hotspots’

Haplotyping

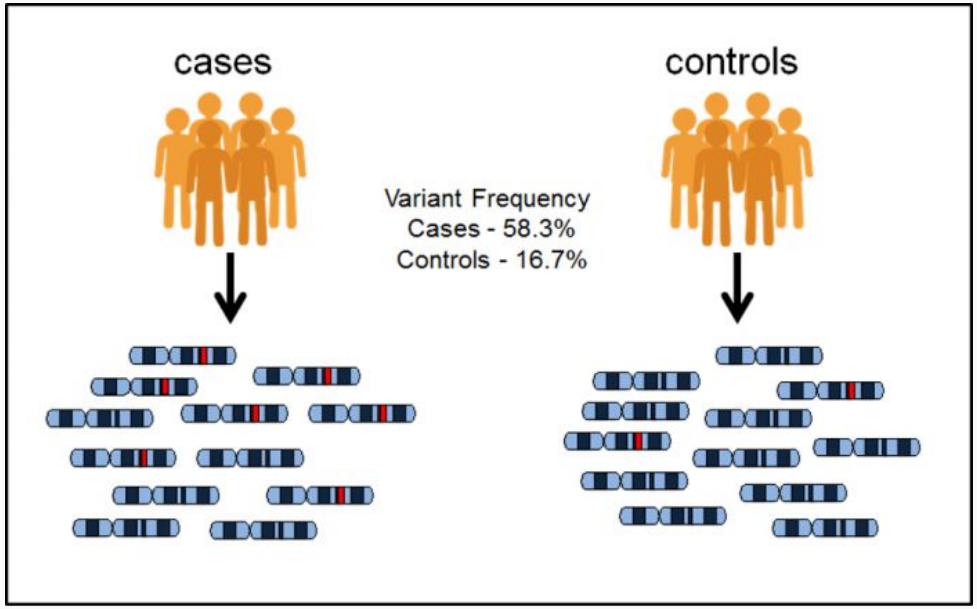
- DNA recombines in large blocks
- SNPs in a block move around together
- Looking at the common SNPs in a block, reveals the ancestry information
- **Linkage Disequilibrium (LD)**: adjacent SNPs co-occur more often than expected
  - i.e. 2 SNPs are *in LD* with each other

Haplotyping



Genome Wide Association Studies (GWAS)

- Which SNPs are associated with a variable of interest?
  - e.g. disease, height
- Does the frequency of any SNP differ between groups?
- Associated SNPs have:
  - Effect size - e.g. amount of increased risk
  - p-value - precision of effect
- **Risk allele**: allele associated with increased or decreased probability of having a disease



Calling SNPs - samtools

```
samtools mpileup -u -v -r chr22:29268316-29300343
-d 150 -f ../06/ref/chr22.fa
NA12878_phased_chr22.bam >
NA12878_chr22_samtools_EWSR1.vcf
```

Calling small variants - GATK

GATK - Genome Analysis Toolkit

```
gatk HaplotypeCaller \
     -L chr22:29268316-29300343 \
     -R ../06/ref/chr22.fa \
     -I NA12878_phased_chr22.bam \
     -O NA12878_chr22_gatk_EWSR1.vcf.gz \
     -ERC GVCF # BP_RESOLUTION
```