



Analysis of Sequence Variation

GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2025-05-07

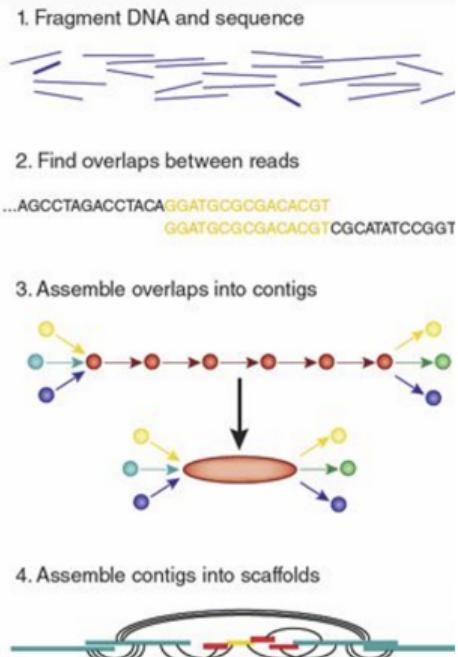
Whole Genome/Exome Sequencing

Whole Genome Sequencing

- ▶ **Whole Genome Sequencing (WGS)**
- ▶ Generate enough reads to attain:
 - ▶ More than 95% coverage of source genome
 - ▶ More than 30x average depth
- ▶ Two strategies:
 - ▶ De novo: assemble reads into a new sequence
 - ▶ Re-sequence: refine an existing reference sequence

De novo assembly

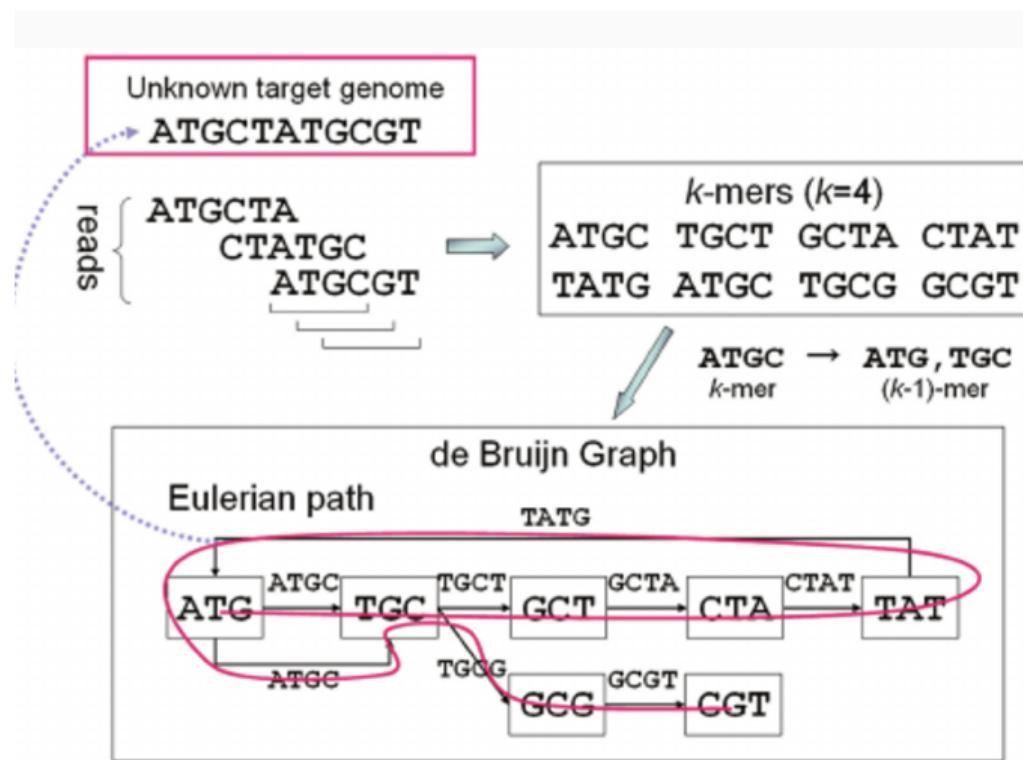
- **Genome Assembly** - Create new reference 'from scratch'
- Examine reads for overlapping sequence
- **Contig** - longer assembled sequence from short reads
- **Scaffold** - assembled contigs
- **Chromosome** - assembled scaffolds
- Assembly from short reads is hard



De novo assembly: 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
- ▶ de Brujin Graphs and Overlap Layout Consensus

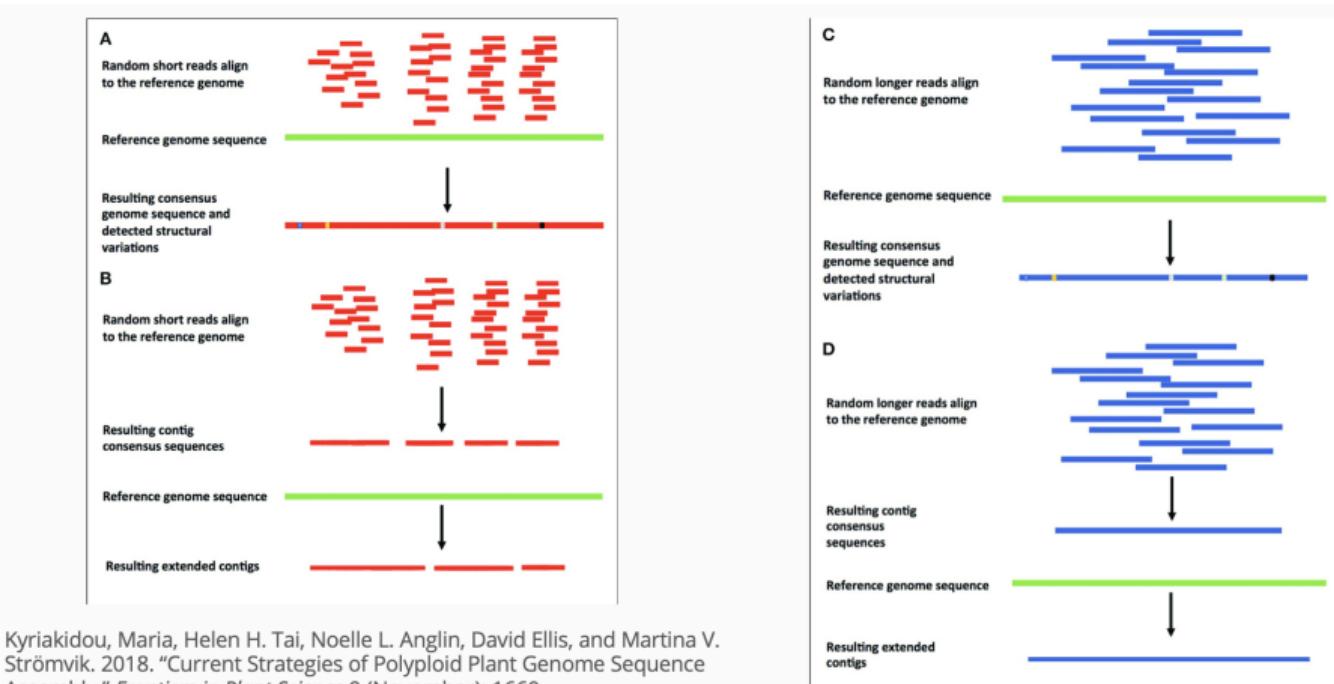
De novo assembly: graph-based



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



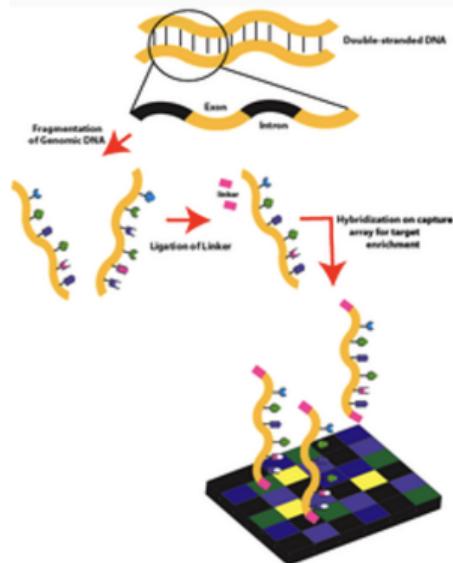
Kyriakidou, Maria, Helen H. Tai, Noelle L. Anglin, David Ellis, and Martina V. Strömvik. 2018. "Current Strategies of Polyploid Plant Genome Sequence Assembly." *Frontiers in Plant Science* 9 (November): 1660.

Whole Exome Sequencing

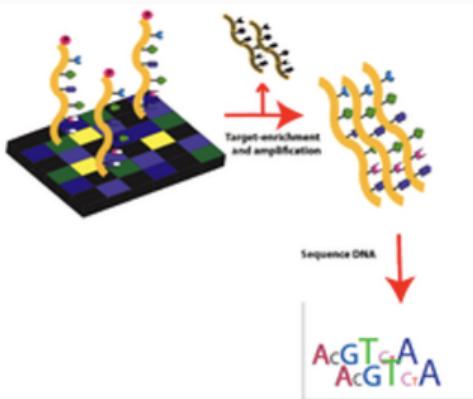
- ▶ **Whole Exome Sequencing (WES)**
- ▶ Exons 1%-2% of human genome sequence
- ▶ Pre-select reads that map to exons
- ▶ Sequence to much greater depth than WGS
- ▶ Identify coding variants

Exome Sequence Selection

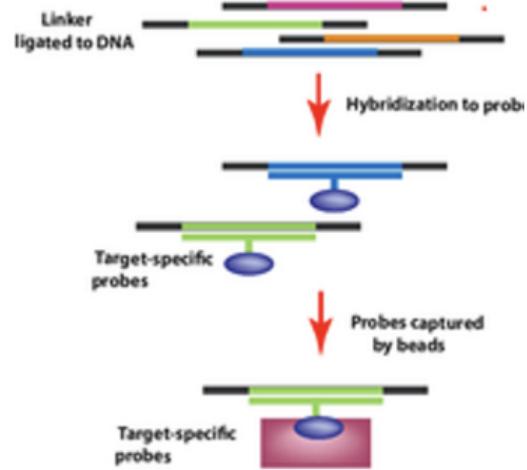
Array-based capture



In-solution



In-Solution Capture

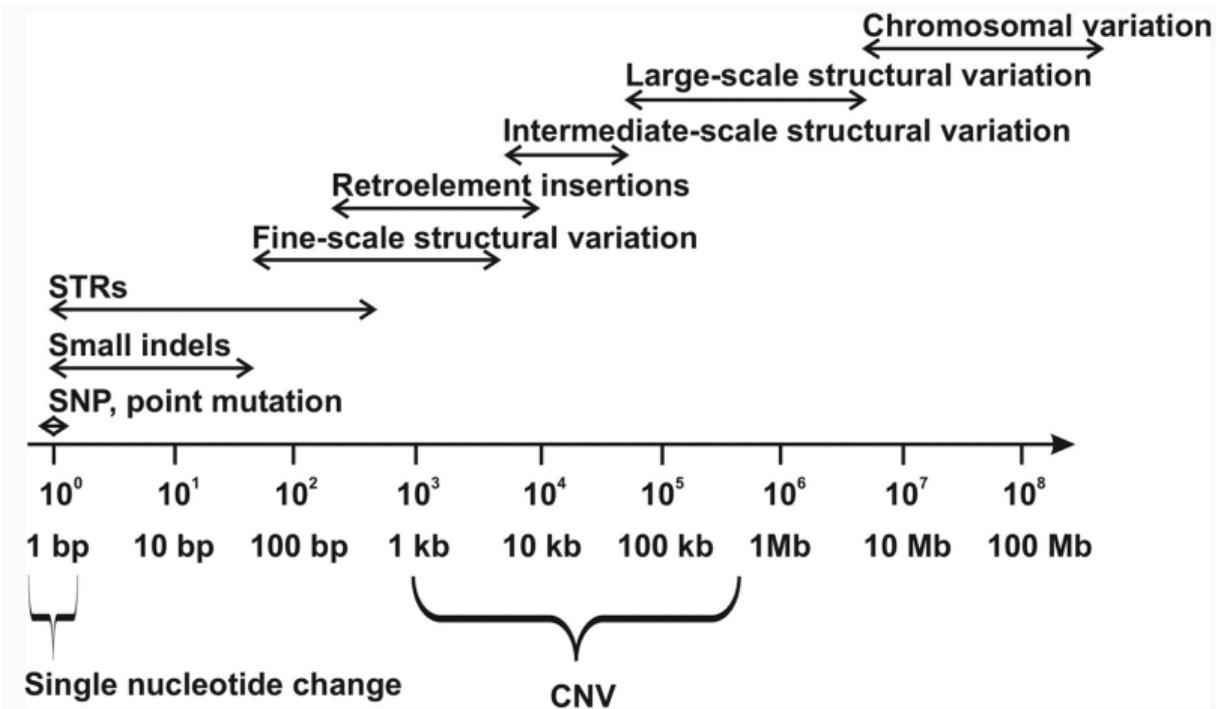


SNP and Variant Calling

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

reference:

AA-TACGG**A**CGGACTTTA

read1:

AA**C**TACGG-

read2:

AA**C**TACGG-

read3:

AA**C**TACGG-

CGG**C**CTTTA

INsertion

SUBstitution

DEletion

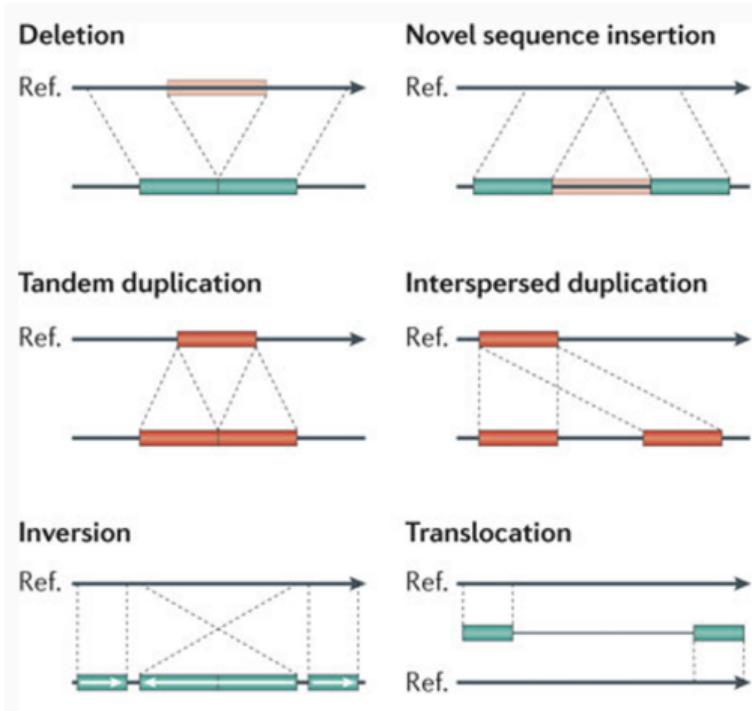
Single Nucleotide Polymorphisms (SNPs)

- ▶ Most commonly studied type of variant
- ▶ Types of single nucleotide alterations:
 - ▶ **Variant** (SNV) - any single mutation
 - ▶ **Polymorphism** (SNP) - SNV observed in significant frequency in population (e.g. >1%)
- ▶ Typically base changes, e.g. A to 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

Structural variation



Genotyping

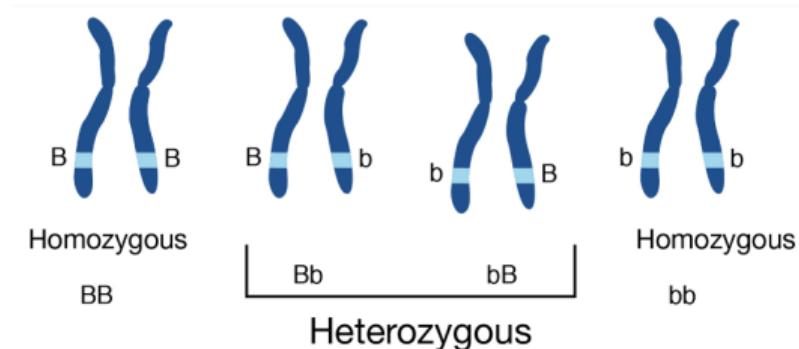
- ▶ **Genotype:** an individual's variant(s)
- ▶ **Phenotype:** an individual's physical form
- ▶ **De novo** variant calling/detection: given genomic reads and a reference, find all the variants
- ▶ **Genotyping:** examine individual for *a priori* variants at known locations
 - ▶ Can use arrays (i.e.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

Human Genomic Variants

- ▶ Humans have diploid genome
 - ▶ i.e. 2 copies of each gene
- ▶ dbSNP - NCBI repository for SNPs
 - ▶ As of June 2021 (Build 155)
 - ▶ 3.3B submitted variants
 - ▶ 1B annotated human SNPs Every human has on average:
 - ▶ A variant every 1kb
 - ▶ 2-3 million SNPs

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



<https://www.genome.gov/genetics-glossary/heterozygous>

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

dbSNP - NCBI SNP Database

National Library of Medicine
National Center for Biotechnology Information

Log in

dbSNP Short Genetic Variations

Search for terms Search Examples: rs268, BRCA1 and more Advanced search

i Welcome to the Reference SNP (rs) Report

All alleles are reported in the Forward orientation. Click on the Variant Details tab for details on Genomic Placement, Gene, and Amino Acid changes. HGVS names are in the HGVS tab.

Reference SNP (rs) Report

 Download    

 Switch to classic site

rs429358

Current Build 155
Released April 9, 2021

Organism

Homo sapiens

Clinical Significance Reported in ClinVar

Position

chr19:44908684 (GRCh38.p13) 

Gene : Consequence APOE : Missense Variant

Alleles

T>C

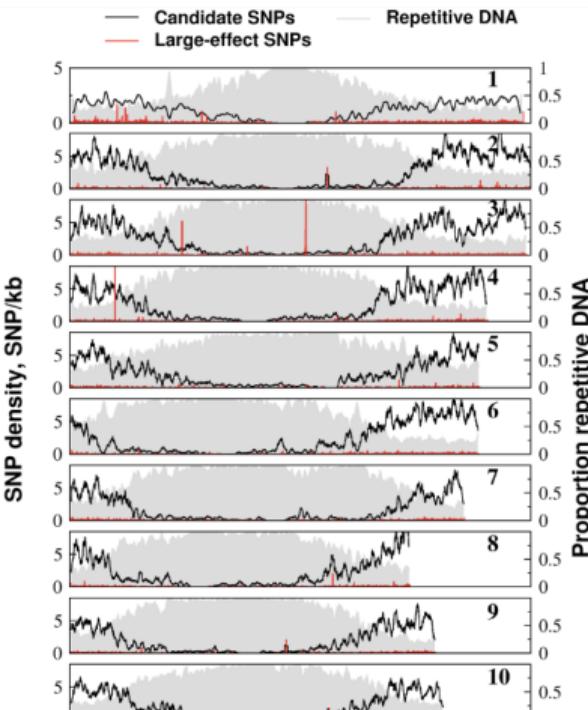
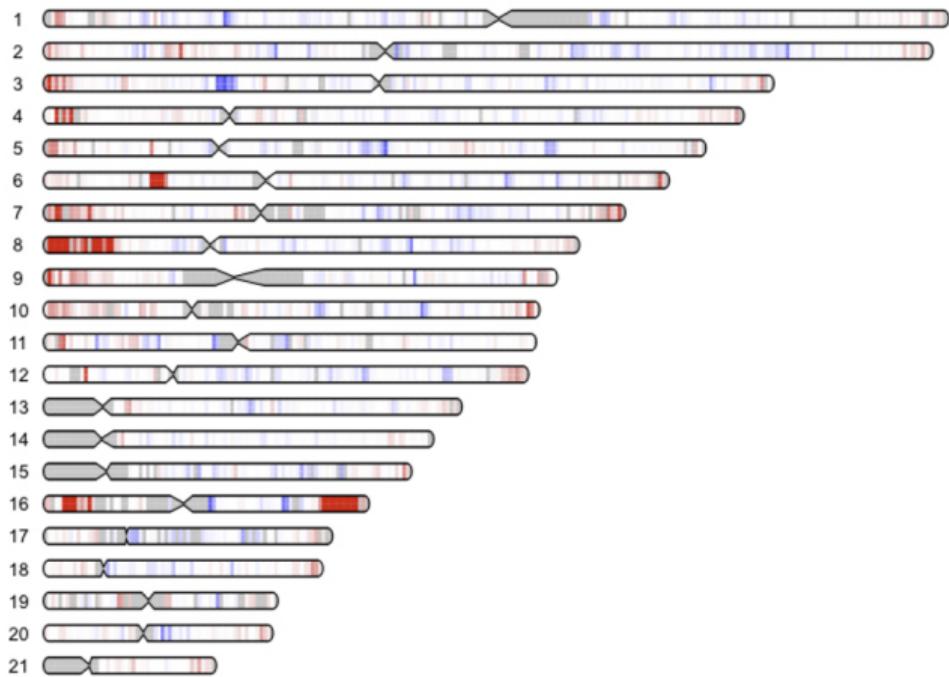
Publications

416 citations

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



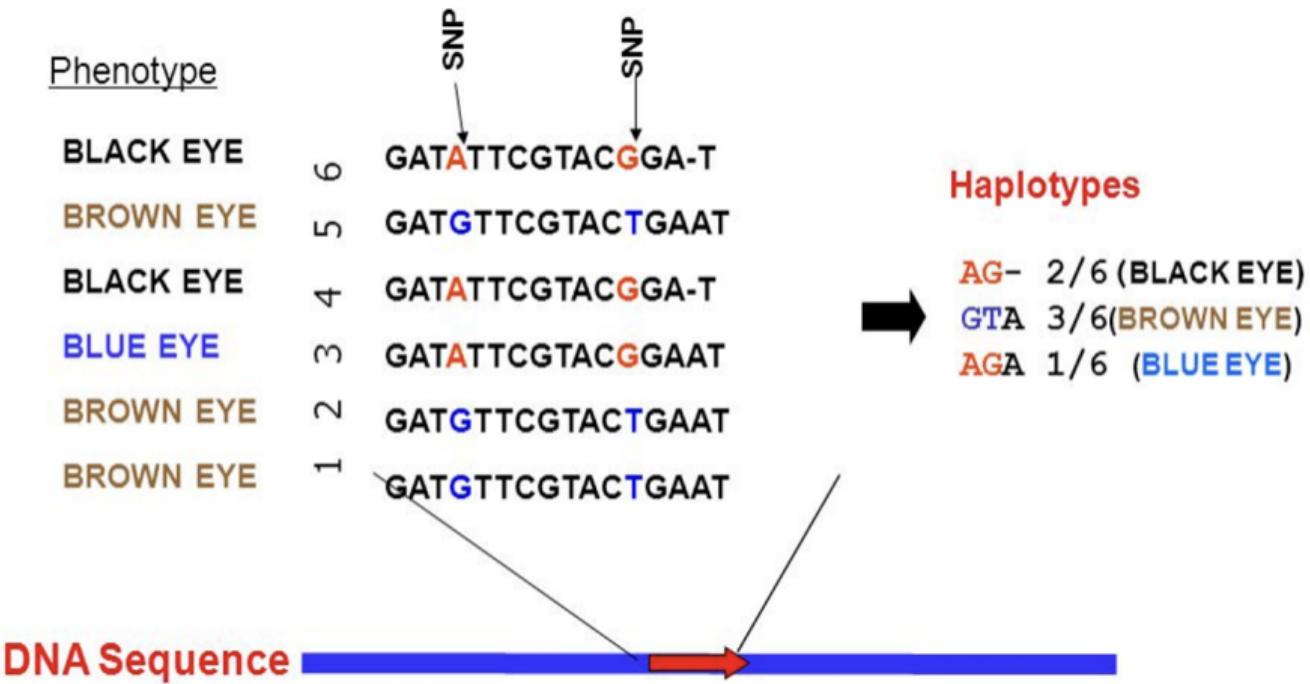
Variants Larger Than A Read

- ▶ **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



Methods for SNP Calling

1. GATK (Genome Analysis Toolkit):

- ▶ Developed by the Broad Institute, GATK is a widely used toolkit for variant discovery in high-throughput sequencing data.
- ▶ It employs a best practices pipeline for variant calling, including base quality score recalibration, indel realignment, and variant quality score recalibration.
- ▶ GATK offers various tools such as HaplotypeCaller and UnifiedGenotyper for variant calling from both exome and whole genome sequencing data.

2. Mutect2:

- ▶ Mutect2 is part of the GATK toolkit and is specifically designed for somatic mutation calling, particularly in cancer genomes.
- ▶ It utilizes a probabilistic model to differentiate true somatic mutations from sequencing artifacts.
- ▶ Mutect2 can be applied to both exome and whole genome sequencing data to identify somatic variants.

Methods for SNP Calling

3. Samtools:

- ▶ Samtools is a suite of programs for interacting with high-throughput sequencing data in the SAM/BAM format.
- ▶ It includes the mpileup command for generating pileup data, and bcftools for variant calling from the pileup data.
- ▶ Samtools is efficient and widely used for variant calling in both exome and whole genome sequencing datasets.

4. FreeBayes:

- ▶ FreeBayes is a Bayesian genetic variant detector designed to detect SNPs, indels, and complex polymorphisms in high-throughput sequencing data.
- ▶ It utilizes a haplotype-based approach to increase sensitivity and specificity, particularly in the context of population-scale sequencing data.
- ▶ FreeBayes can be used for variant calling in both exome and whole genome sequencing studies.

Methods for SNP Calling

5. VarScan:

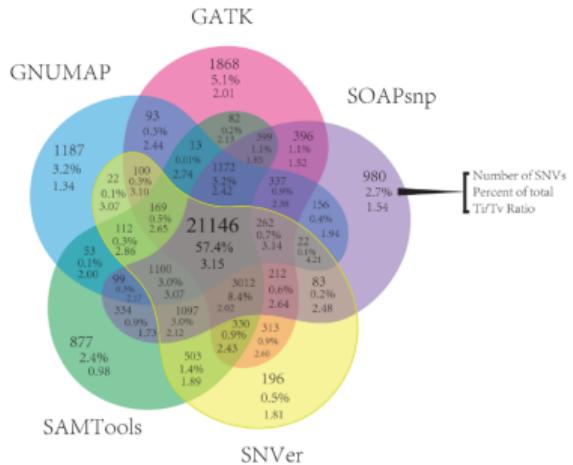
- ▶ VarScan is a platform-independent tool for variant detection in massively parallel sequencing data.
- ▶ It is optimized for calling germline variants, somatic mutations, and copy number alterations.
- ▶ VarScan supports both exome and whole genome sequencing data and provides a range of options for variant calling and filtering.

6. DeepVariant:

- ▶ DeepVariant is an end-to-end deep learning-based variant caller developed by Google.
- ▶ It employs a convolutional neural network (CNN) architecture to call SNPs and indels with high accuracy.
- ▶ DeepVariant is particularly effective in identifying complex variants and can be applied to both exome and whole genome sequencing data.

Inconsistencies Among SNP Callers

Low concordance of variant-calling pipelines (O'Rawe, *Genome Med*, 2013)



SAMtools Example

Multiple sample SNP calling:

```
## Load samtools
module load samtools

## mpileup
samtools mpileup -f genomefile.fa \
myalignments.sorted.bam > myalignments.vcf
```

GATK Example

```
module load samtools
# Index reference
samtools faidx chrX_5MB.fa

module load gatk
# make sequence dictionary
gatk CreateSequenceDictionary -R chrX_5MB.fasta
# make .vcf file!
gatk HaplotypeCaller \
    -R chrX_5MB.fa -I proband_short_bwa.sorted.bam \
    -O proband_short.vcf
```

VCF files

8 fixed columns: #CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT
 Additional columns, one for each sample, with sample ID

(a) VCF example

Header	##fileformat=VCFv4.1 ##fileDate=20110413 ##source=VCFtools ##reference=file:///refs/human_NCBI36.fasta ##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens"> ##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens"> ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##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"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">																																																							
Body	<table border="1"> <thead> <tr> <th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th><th>FORMAT</th><th>SAMPLE1</th><th>SAMPLE2</th></tr> </thead> <tbody> <tr> <td>1</td><td>1</td><td>.</td><td>ACG</td><td>A,AT</td><td>40</td><td>PASS</td><td>.</td><td>GT:DP</td><td>1/1:13</td><td>2/2:29</td></tr> <tr> <td>1</td><td>2</td><td>.</td><td>C</td><td>T,CT</td><td>.</td><td>PASS</td><td>H2;AA=T</td><td>GT</td><td>0 1</td><td>2/2</td></tr> <tr> <td>1</td><td>5</td><td>rs12</td><td>A</td><td>G</td><td>67</td><td>PASS</td><td>.</td><td>GT:DP</td><td>1 0:16</td><td>2/2:20</td></tr> <tr> <td>X</td><td>100</td><td></td><td>T</td><td></td><td></td><td>PASS</td><td>SVTYPE=DEL;END=200</td><td>GT:GQ:DP</td><td>1:12:</td><td>0/0:20:36</td></tr> </tbody> </table>	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29	1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0 1	2/2	1	5	rs12	A	G	67	PASS	.	GT:DP	1 0:16	2/2:20	X	100		T			PASS	SVTYPE=DEL;END=200	GT:GQ:DP	1:12:	0/0:20:36
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2																																														
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29																																														
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0 1	2/2																																														
1	5	rs12	A	G	67	PASS	.	GT:DP	1 0:16	2/2:20																																														
X	100		T			PASS	SVTYPE=DEL;END=200	GT:GQ:DP	1:12:	0/0:20:36																																														

Complete Variant Calling Pipeline (Outdated!)

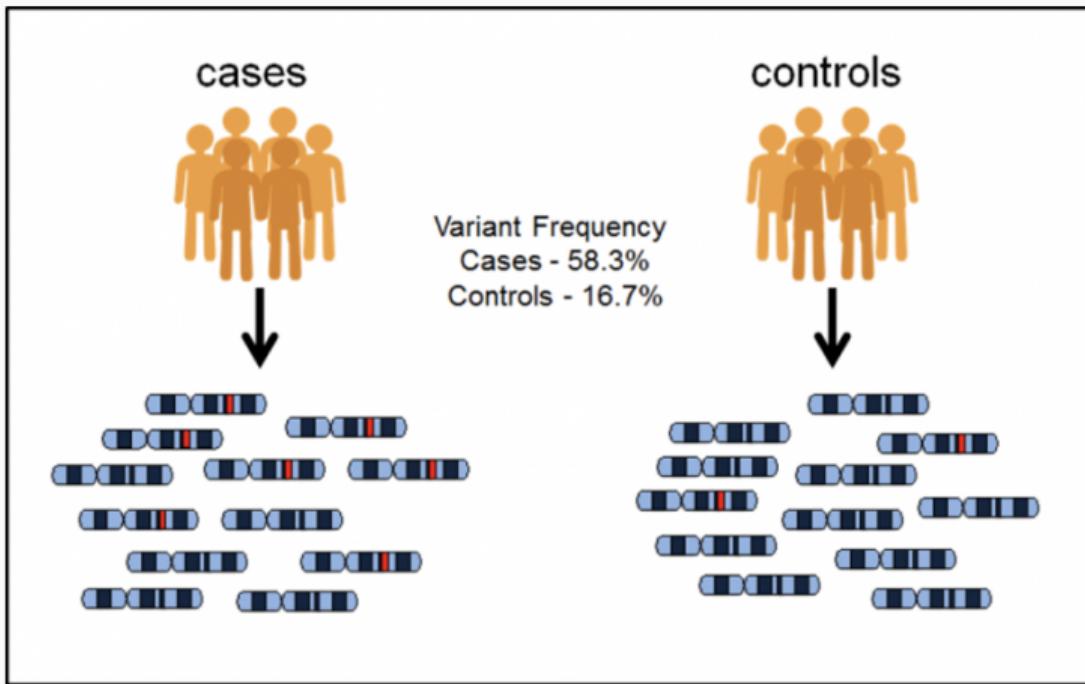
Our analysis pipeline consisted of the following:

- ▶ Align the FASTQ files to genome
- ▶ Convert SAM file to BAM, add read group info
- ▶ Filter the reads based on quality (BAMTools)
- ▶ Samtools to sort and index, and use Picard to mark duplicates
- ▶ GATK calibration, realignment, variant calling (HaplotypeCaller, Mutect2)
- ▶ Filter the called variants (GATK filtersnps and filterindels).
- ▶ Annotation of SNPs (snpEff, condel)
- ▶ Filter by frequency (thousand genomes, TCGA, etc.)
- ▶ Downstream analysis (rare variants, pedigree, pathway level, etc)

Genome Wide Association Studies (GWAS)

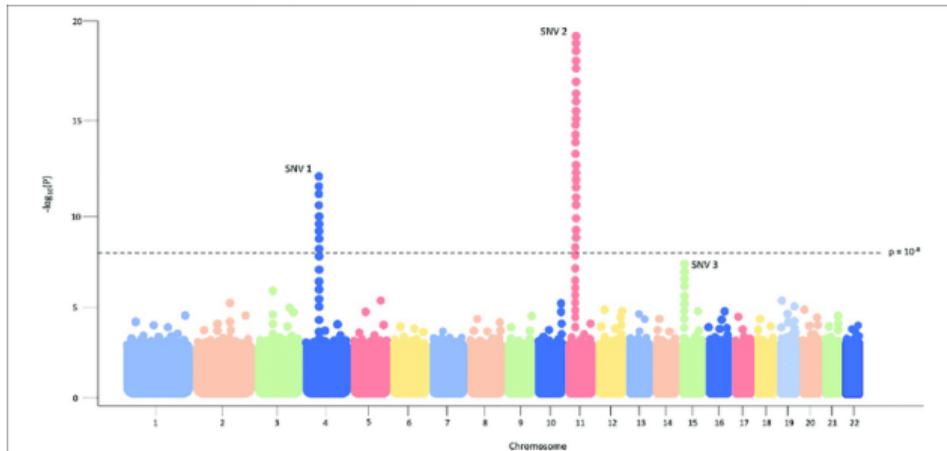
- ▶ Which SNPs are associated with a variable of interest?
 - ▶ e.g. disease, height Does the frequency of any SNP differ between groups? Associated SNPs have:
- ▶ Effect size - e.g. amount of increased risk
- ▶ p-value - precision of effect
- ▶ **Risk allele**: allele associated with increased or decreased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

Genome Wide Association Studies (GWAS)



Manhattan Plot

- ▶ In a **Manhattan plot**, each dot represents a genetic marker (SNP) on the x-axis, and its $-\log_{10}(p\text{-value})$ on the y-axis.
- ▶ The horizontal lines represent the significance threshold.
- ▶ Peaks represent regions where genetic variants are significantly associated with the trait or disease.



Genome-Wide Significance Threshold

- ▶ The genome-wide significance threshold is a threshold used to determine if an association between a genetic variant and a trait is statistically significant.
- ▶ It takes into account the multiple comparisons problem arising from testing thousands or millions of genetic markers across the genome.

Bonferroni Correction

- ▶ One method to establish the genome-wide significance threshold is the Bonferroni correction.
- ▶ The Bonferroni-corrected threshold is calculated by dividing the desired significance level (e.g., 0.05) by the number of tests performed.
- ▶ The most commonly accepted threshold is

$$p < 5 \times 10^{-8},$$

based on a Bonferroni correction for all independent common SNPs across the human genome

False Discovery Rate (FDR)

- ▶ Another method to control for multiple comparisons is the False Discovery Rate (FDR).
- ▶ FDR controls the proportion of false positives among all significant results.
- ▶ It is less conservative than the Bonferroni correction and allows for a higher number of false positives while still controlling the overall error rate.

Downstream Annotation and Analysis (Outdated!)

Downstream Annotation Tools (old list):

- ▶ snpEff (<http://snpeff.sourceforge.net/>)
- ▶ Condel (<http://bg.upf.edu/condel/home>)
- ▶ SIFT <http://sift.jcvi.org/>
- ▶ Polyphen 2 <http://genetics.bwh.harvard.edu/pph2/>
- ▶ <http://mutationassessor.org/>
- ▶ Ensembl variant effect predictor
(<http://www.ensembl.org/info/docs/variation/vep/index.html>)
- ▶ Thousand Genomes variant frequency (e.g. 1% threshold) and Exome Sequence Project variant frequency (e.g. 1%).

Session info

```
sessionInfo()

## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.4.1
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices  utils       datasets   methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.4.2    fastmap_1.2.0    cli_3.6.5      tools_4.4.2
## [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10    rmarkdown_2.29
## [9] knitr_1.50        xfun_0.52       digest_0.6.37   rlang_1.1.6
## [13] evaluate_1.0.3
```