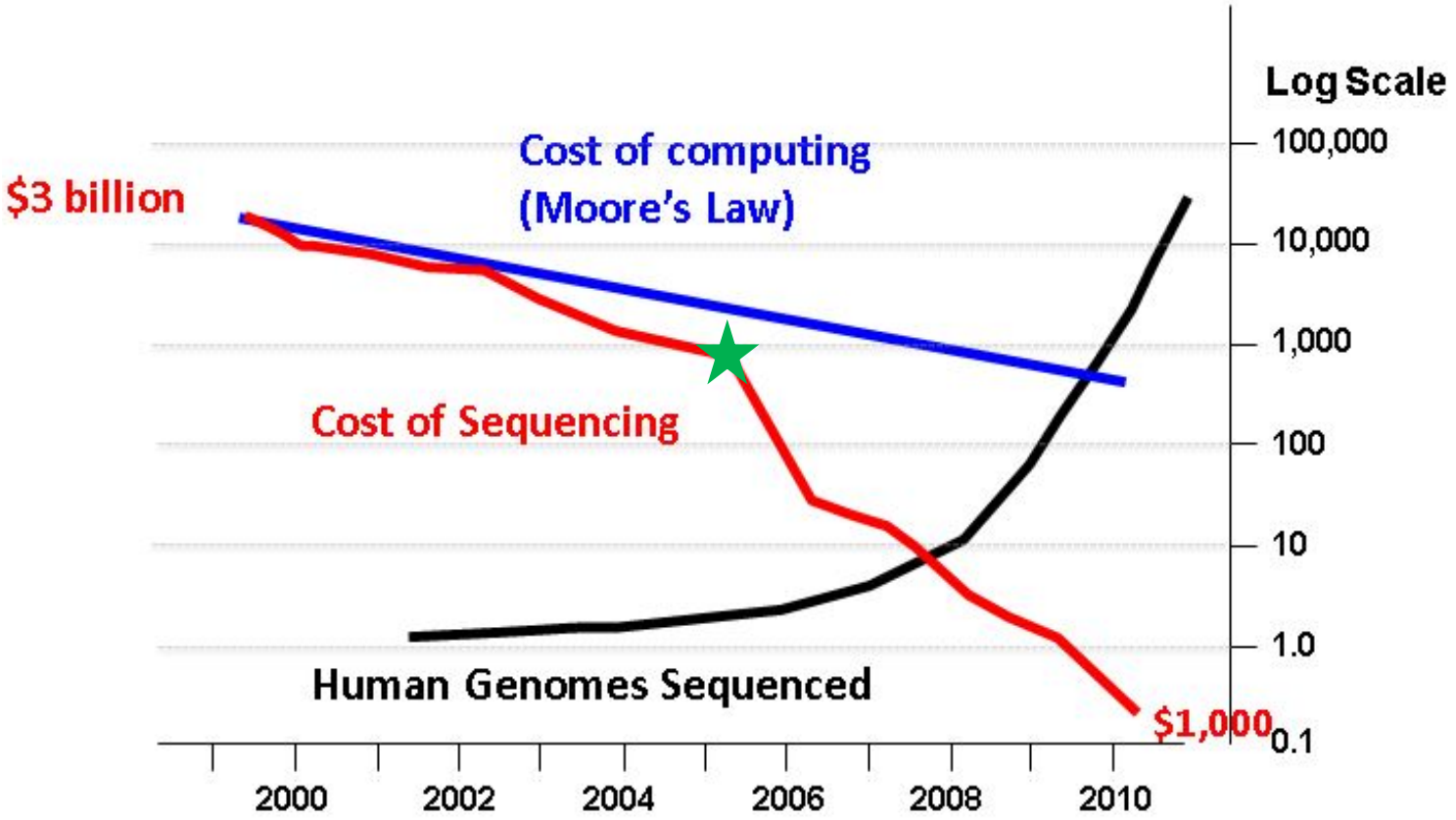
Search for Missing Disease Variants in Large Sequencing Projects

Yan Asmann, PhD
Associate Professor of BioMedical Informatics
Mayo College of Medicine
Mayo Clinic

mayo

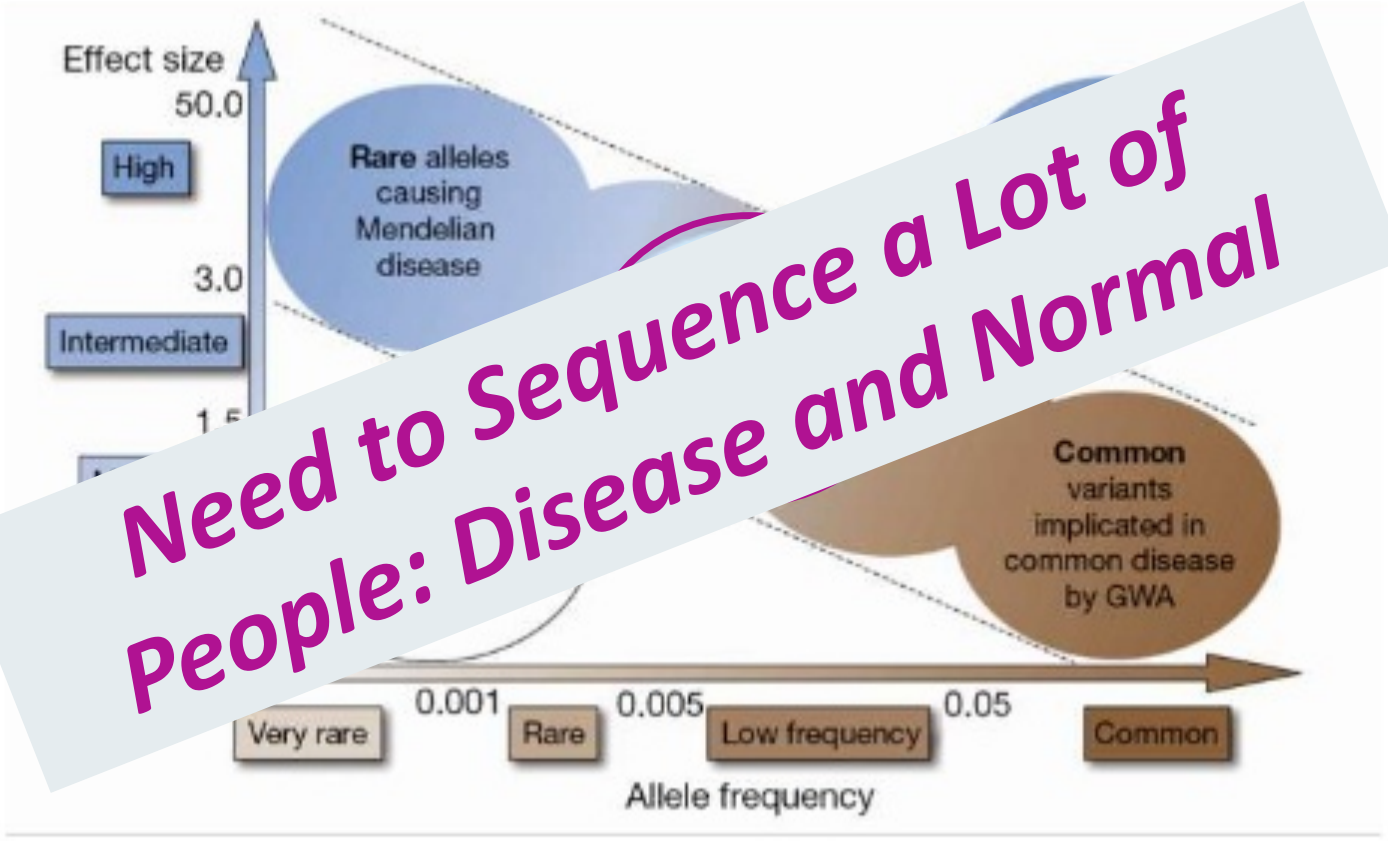
The Sequencing Explosion:



- Human Genome Project: 10 years, $ 3b
- Today: One day, $ 1K
- ★ NextGen Sequencing Machines

mayo

Landscape of Human Disease Genetics: Relationship Between Gene Mutations/Variations and Diseases



"Rare variants could be the primary drivers of common diseases."

- Nat Rev Genet. 2010

mayo

The Parallel Effort of Developing Bioinformatics Pipeline for Sequencing Data Analyses

Read Alignment

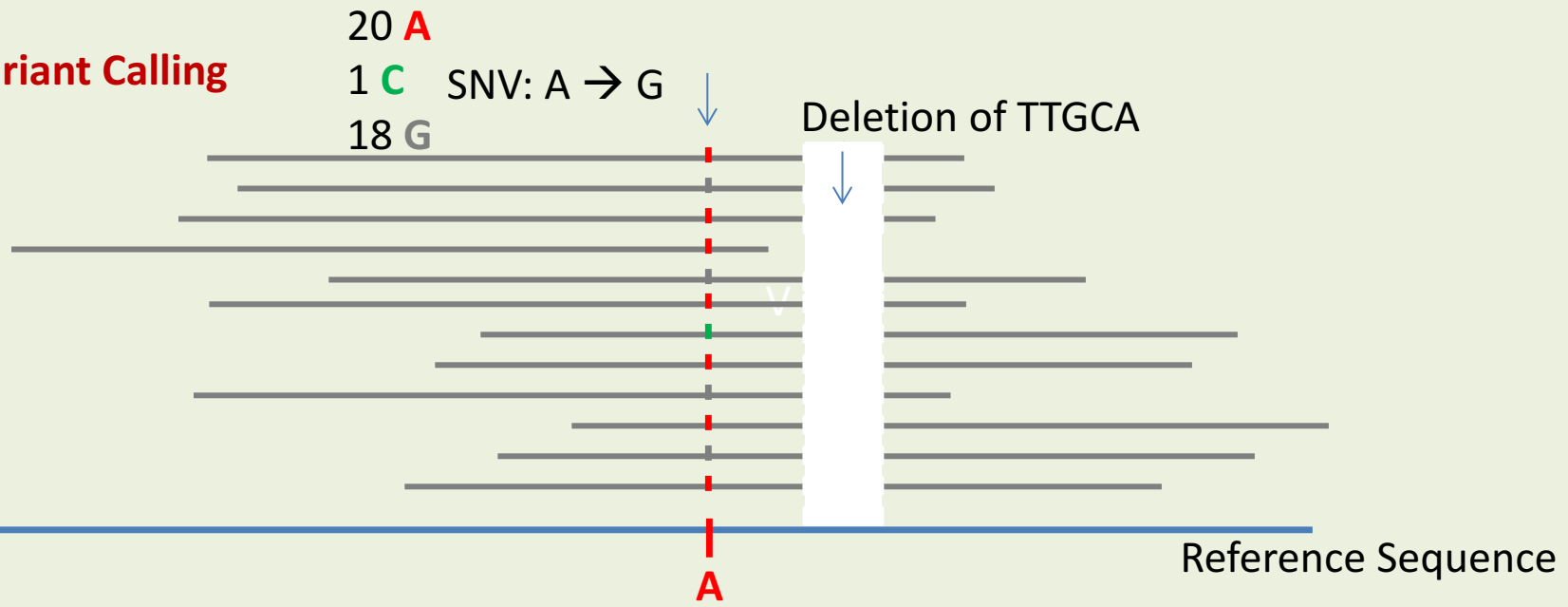
Hundreds of millions of reads per sample

Read Mapped to Reference Genome

↓

Fine tuning of the Mapping: local realignment, local assembly, etc.

Variant Calling



Read Alignment

Read Alignment

Hundreds of millions of reads per sample

Read Mapped to Reference Genome

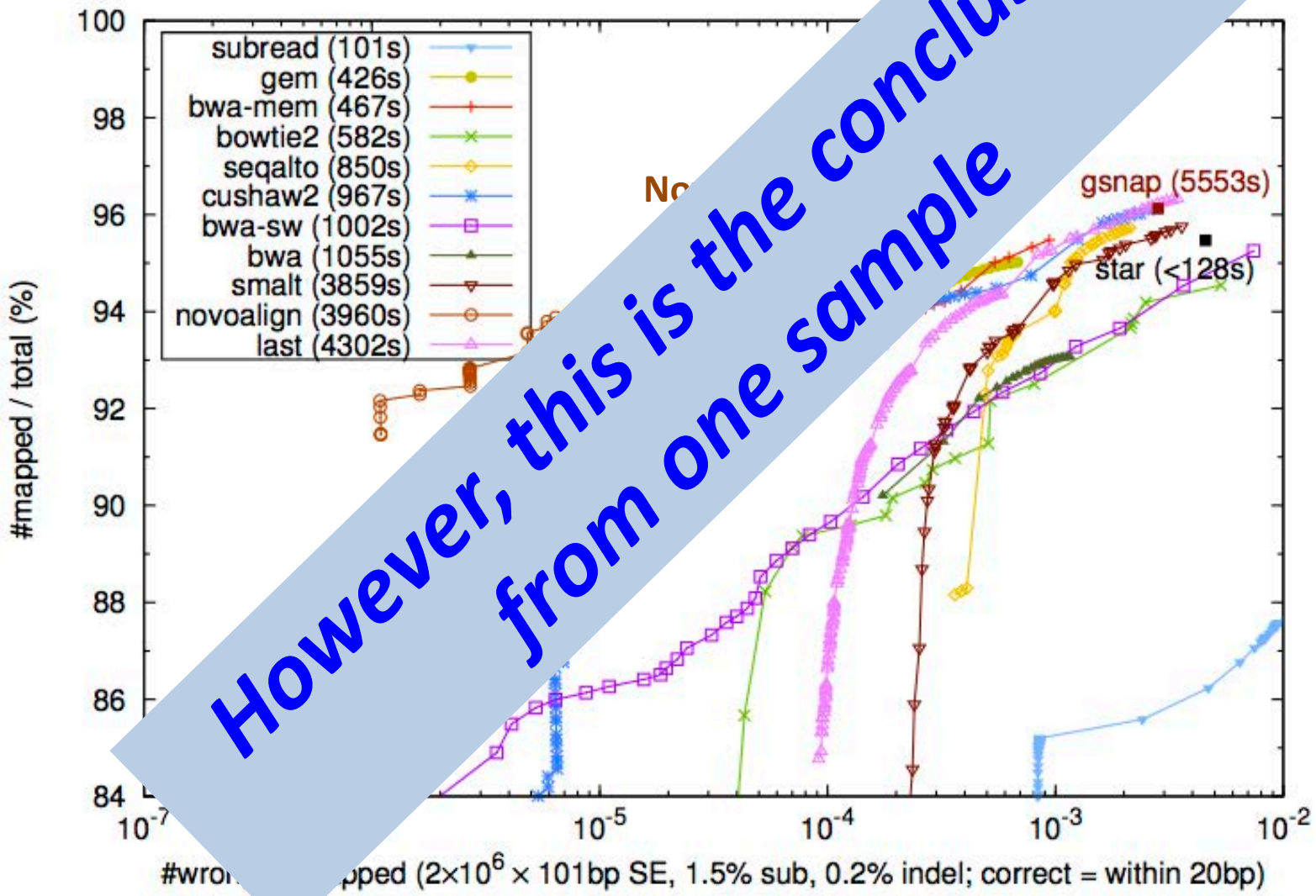
↓

Fine tuning of the Mapping: local realignment, local assembly, etc.

- CPU and RAM Intensive and Slow
- Different Aligner: designed specifically for NextGen Sequencing, balancing speed and accuracy, and dealing with genomic complexity
- Parameters of Aligners: biased towards different types of variants (INDEL[gap], Single Nucleotide Substitution)
- Current Practice: **One single aligner**; **Default parameters**

mayo

Rationale of using one single aligner: benchmark the performance of aligners using reads

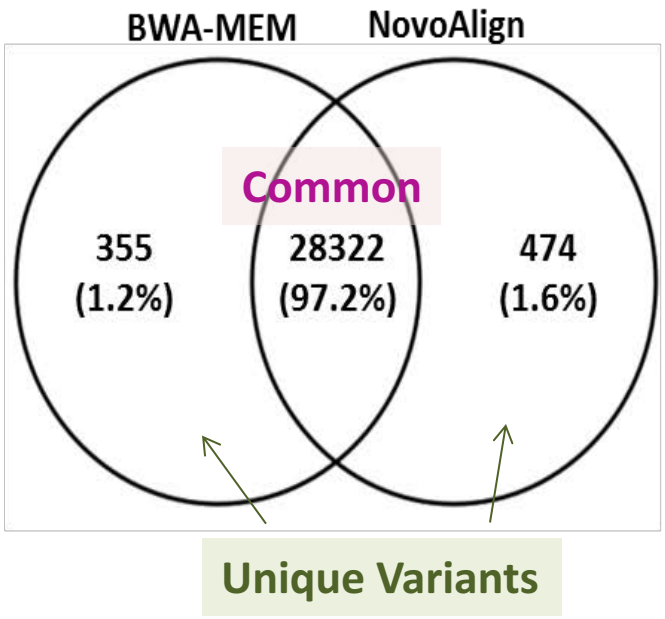


However, this is the conclusion from one sample

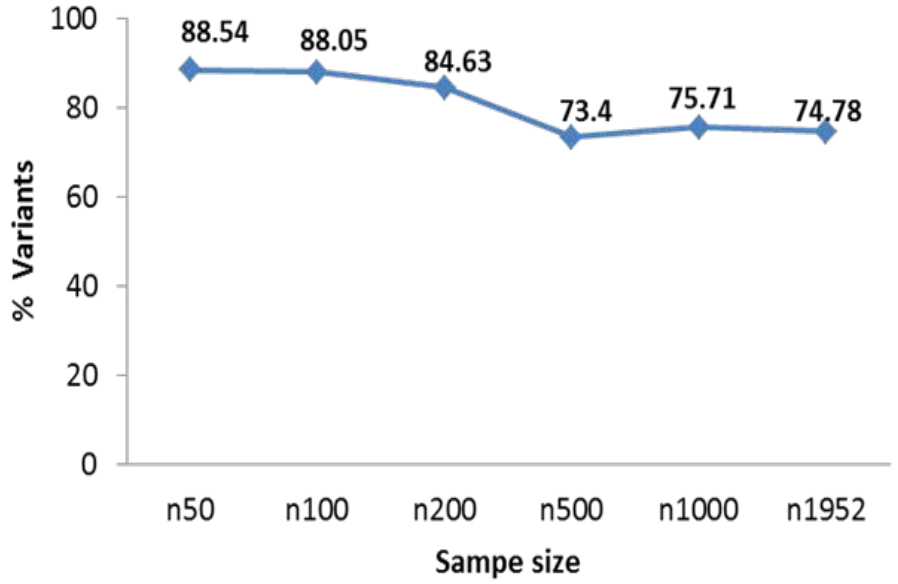
mayo

Hunt for Rare Variants: Increasingly Large Sample Size

a Variant Calling: One Sample



b Percent Common Variants vs. Sample Size



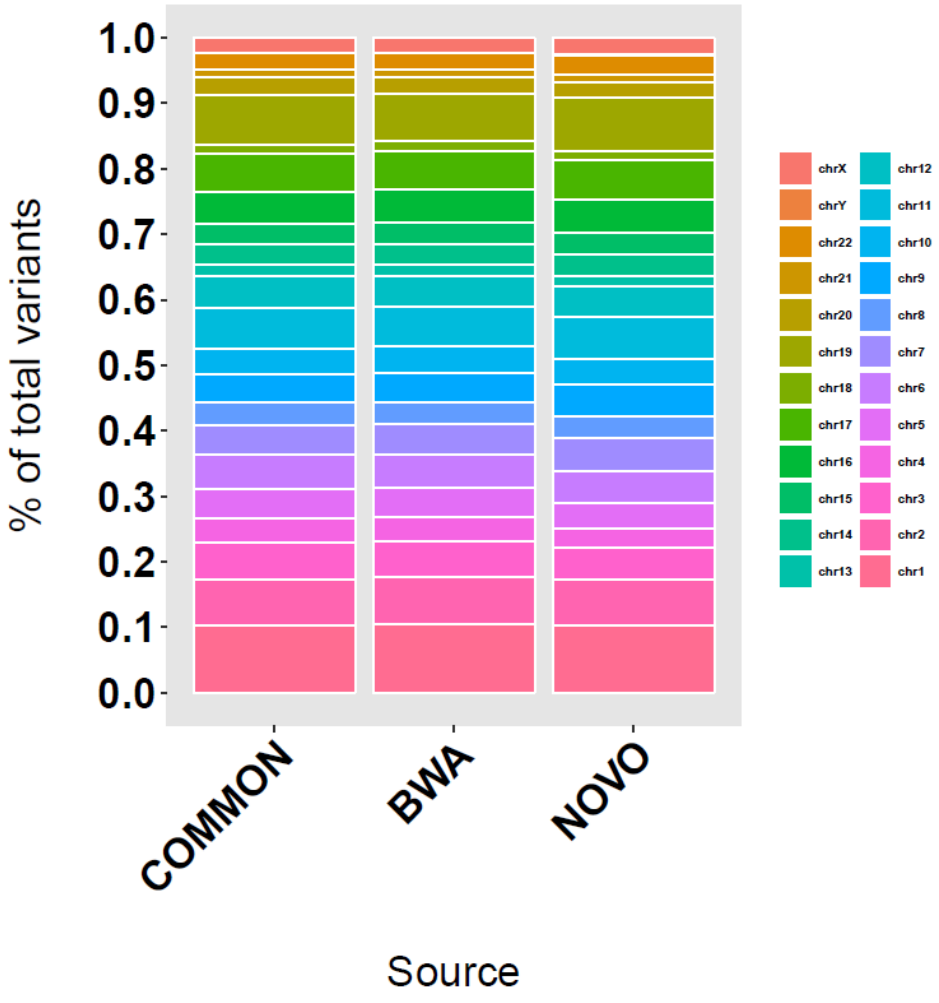
mayo

Compare the Characteristics and Quality of the Unique and Common Variants

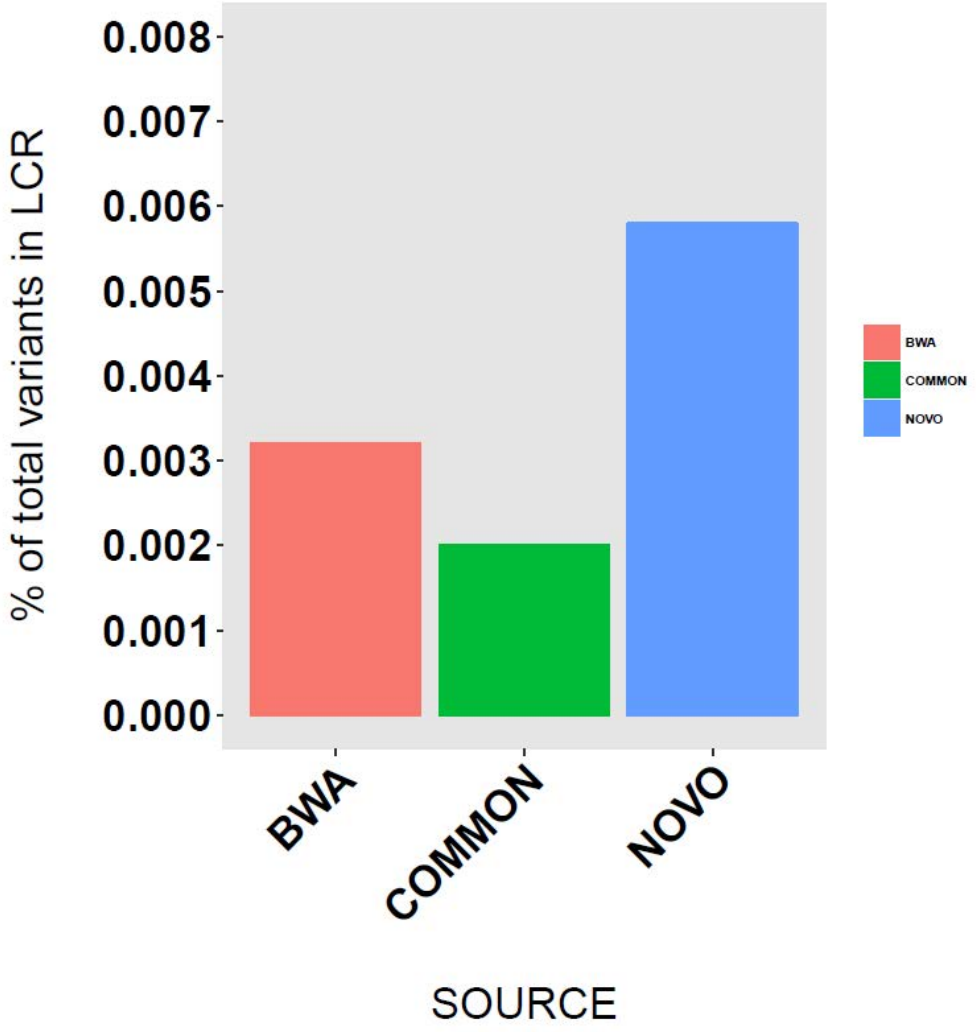
- Exome Sequencing of ~10,000 individuals with and without Alzheimer's Disease
- Sample Size: 2000
- Total Variants: ~20 million
- % Unique Variants: 25.22% of total variants
- % Common Variants: 74.78% of total variants

mayo

Chromosome distribution of unique and common variants



More unique variants are located in Low Complexity Regions than common variants



Most unique variants are known variants already reported in public databases

	BWA-unique	Novo-unique	Common
% known	75.48	75.12	87.04
% novel	24.52	24.88	12.96
% CADD Score >20 in novel variants	39.49	36.93	40.41

mayo

Similar Functional Categories of Unique vs. Common Variants

	BWA unique	Novo unique	Common	Fisher's P-value
Tier1	1.75	1.77	1.34	0.7751
Tier2	61.72	59.36	56.89	
Tier3	36.52	38.87	41.77	

- Tier 1: Protein truncation
- Tier 2: Amino Acid Substitution, Protein Translation Codon Frame Shift
- Tier 3: Other

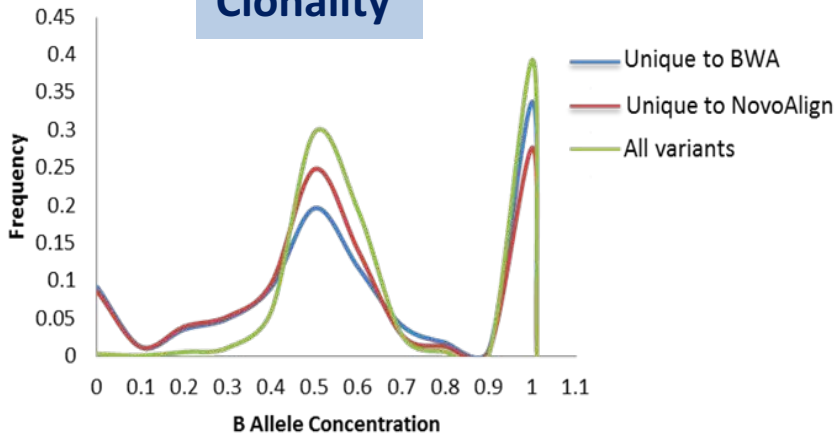
mayo

Unique Variants are more rare

Unique Variants: Biased in rare variants of 1-2% in population.

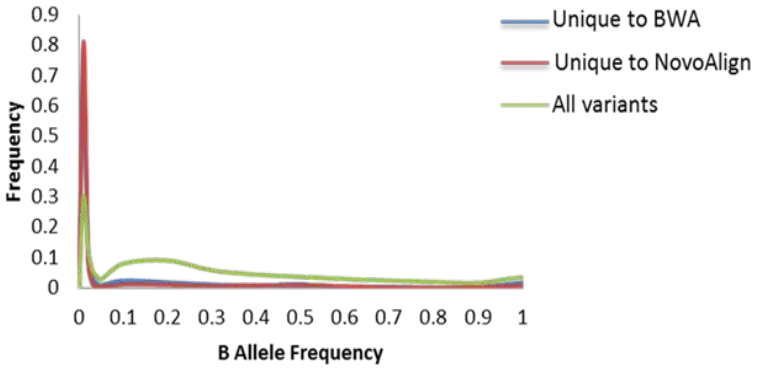
a

Clonality



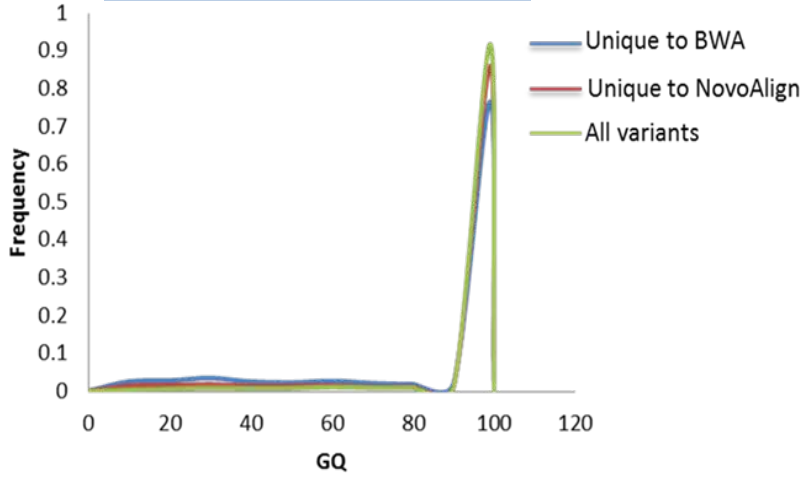
b

Population Frequency



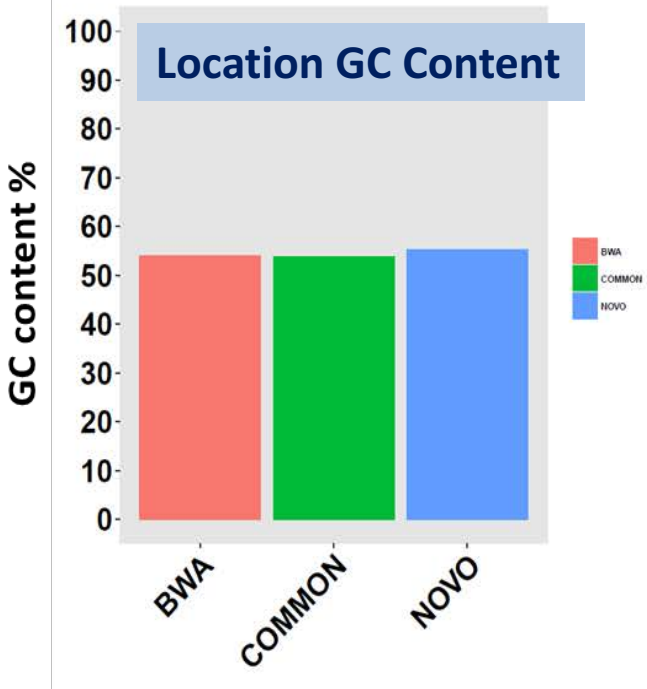
c

Sequencing Quality



d

Location GC Content



Summary:

- Use multiple aligners rescued a significant % of good variants (sample size dependent)
- Different parameter settings of the same aligner also rescued substantial % of good quality variants (data not shown)

- Similar observations when testing different variant calling strategies

- **Testing our observations further on Blue Water**
  - ✓ Larger sample size (up to 10,000)
  - ✓ Include a more expensive list of tools and parameters
  - ✓ Estimate the computational resources necessary
  - ✓ Project on-going
  - ✓ Project co-PI: Dr. Liudmila Sergeevna Mainzer of UIUC

mayo

Acknowledgement

Mayo Clinic:

Asmann Lab:
Yingxue Ren, PhD
Vivek Sarangi, MS
Shulan Tian, PhD

ADSP PIs:
Rosa Rademakers, PhD
Nilufer Taner, MD/PhD
Owen Ross, PhD
Steven Younkin, MD/PhD

UIUC/NCSA:

Liudmila Sergeevna Mainzer, PhD
Matthew Hudson, PhD
Jacob Heldenbrand