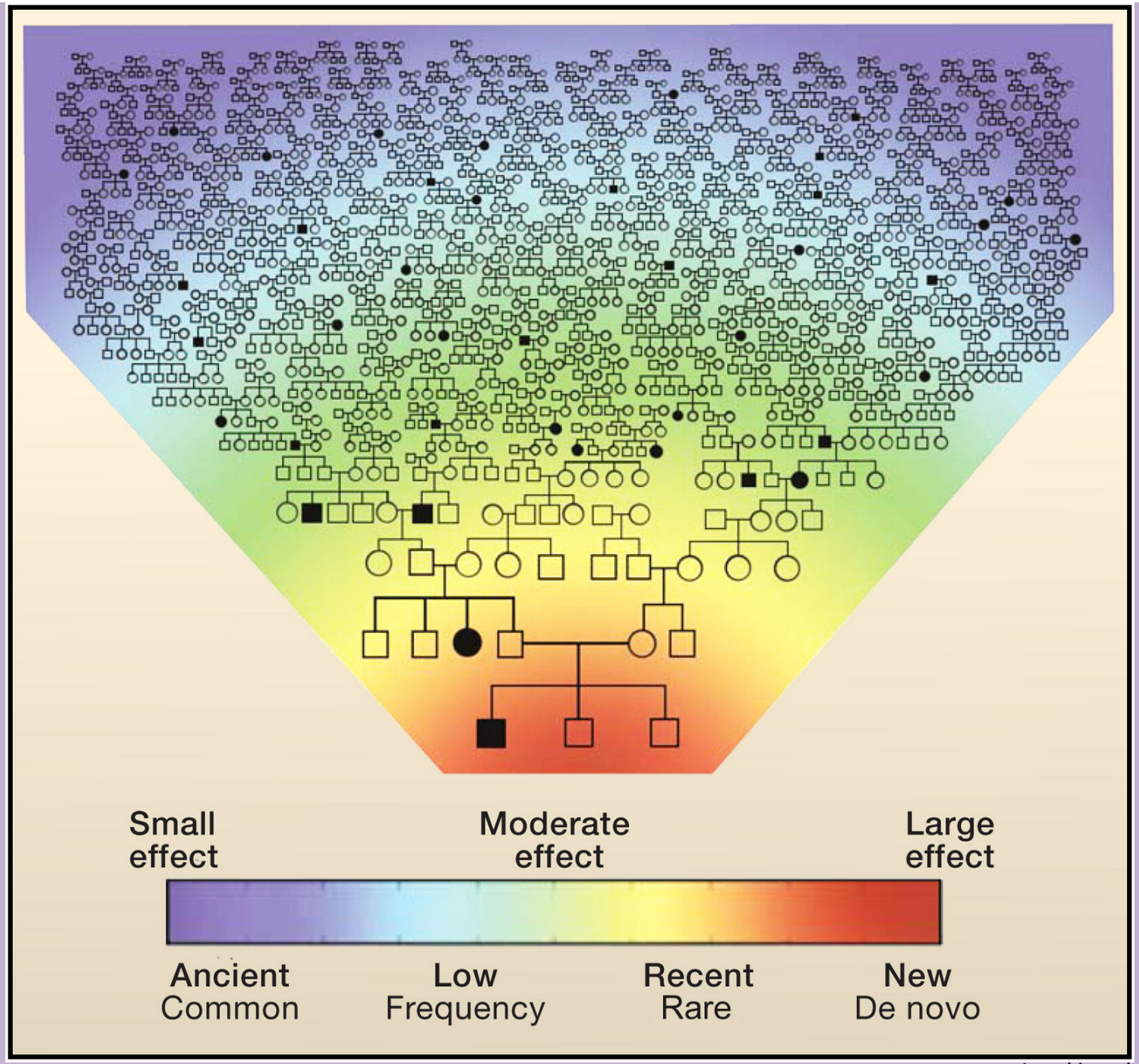
Parallelization of whole genome analysis on a Cray XE6

THE UNIVERSITY OF CHICAGO

NORTHWESTERN UNIVERSITY

Computation Institute

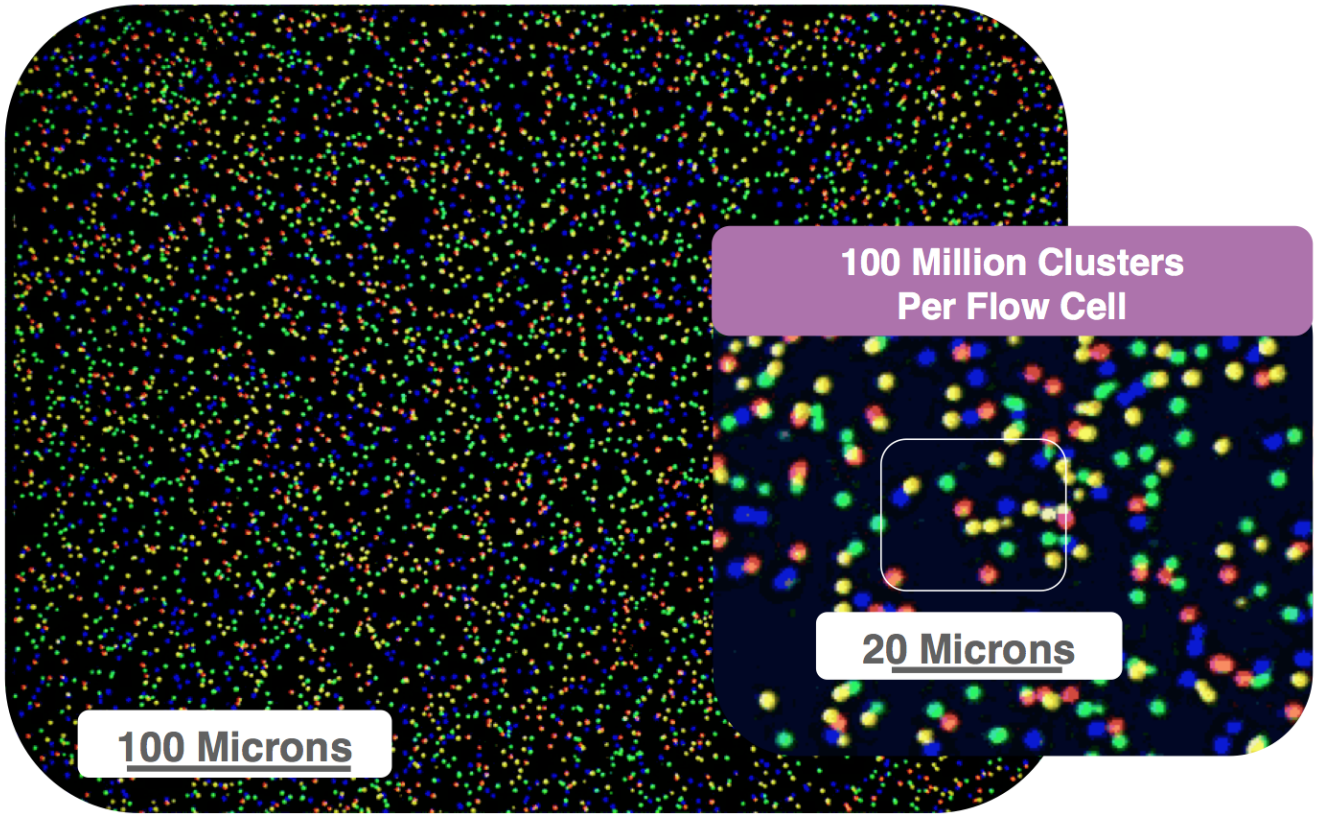
CENTER FOR GENETIC MEDICINE



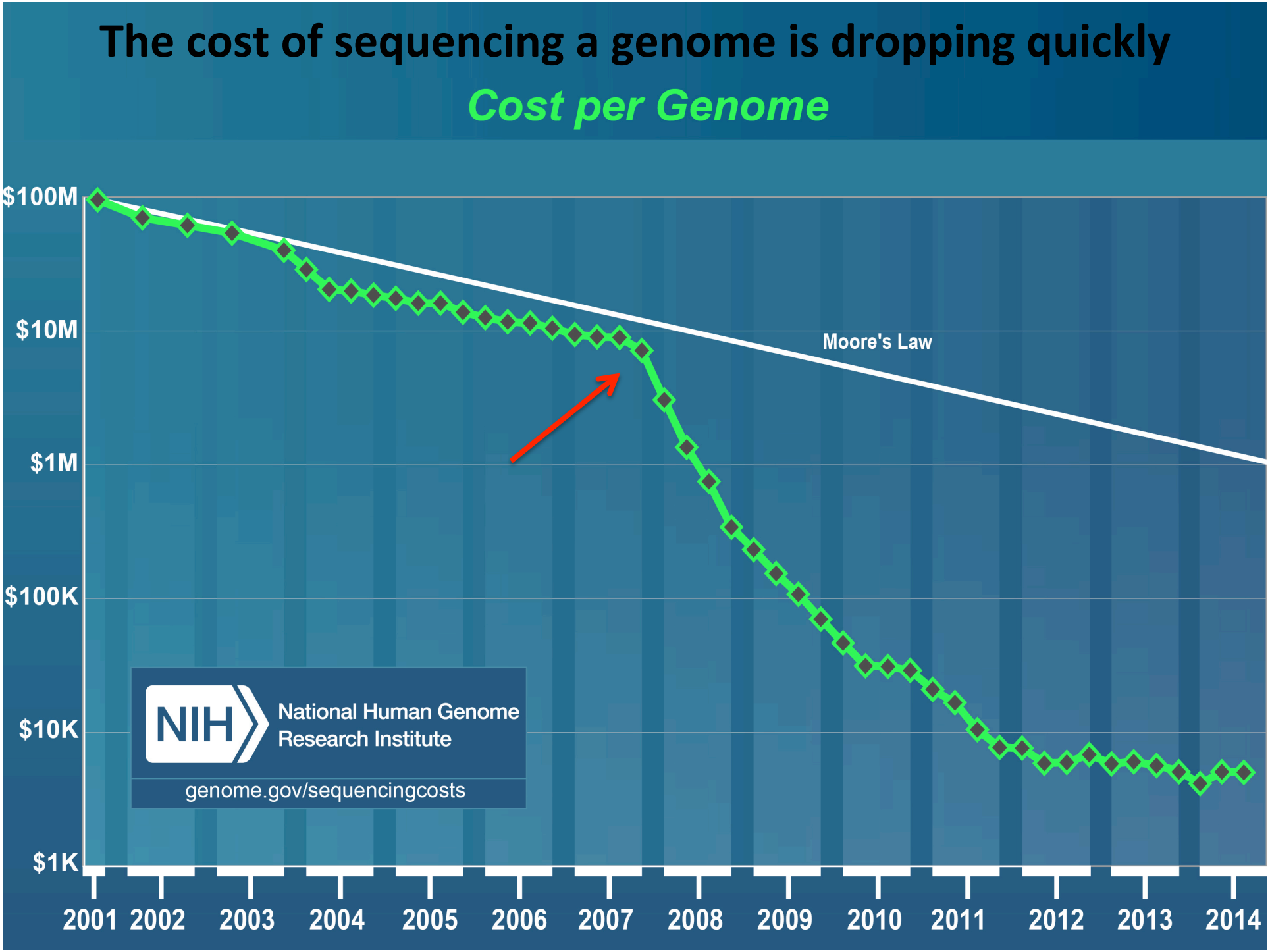
Lupski et al. Cell 147 (2011)

Next Generation sequencing is Massively parallel: fast

Flow cell

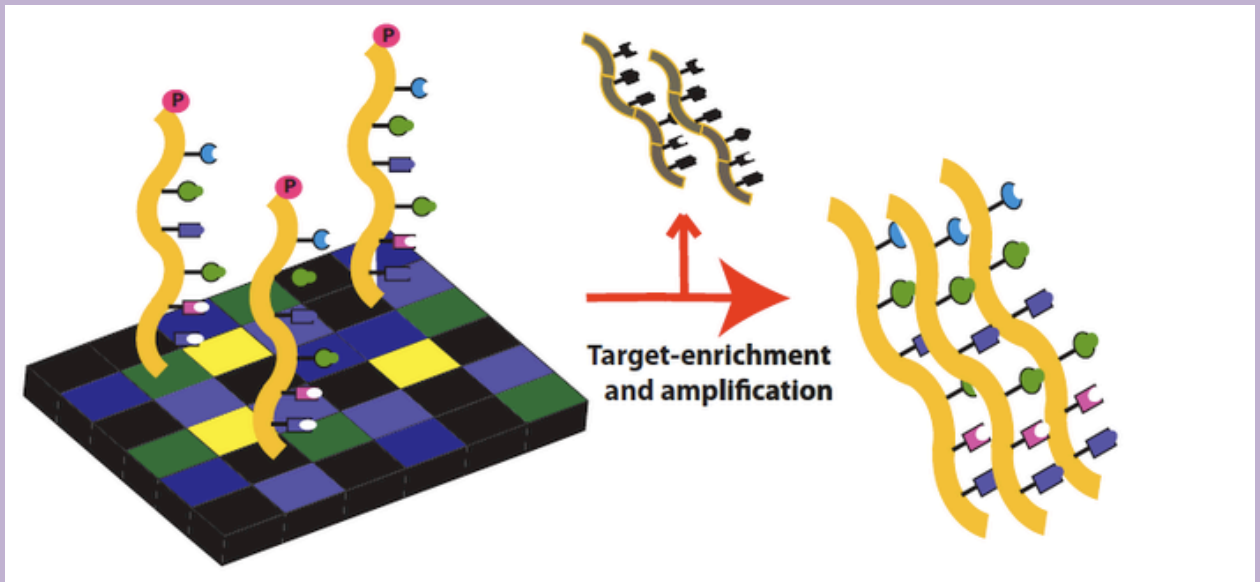


Red = A Blue = G Yellow = C Green = T


The cost of sequencing a genome is dropping quickly
Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

Targeted gene sequencing

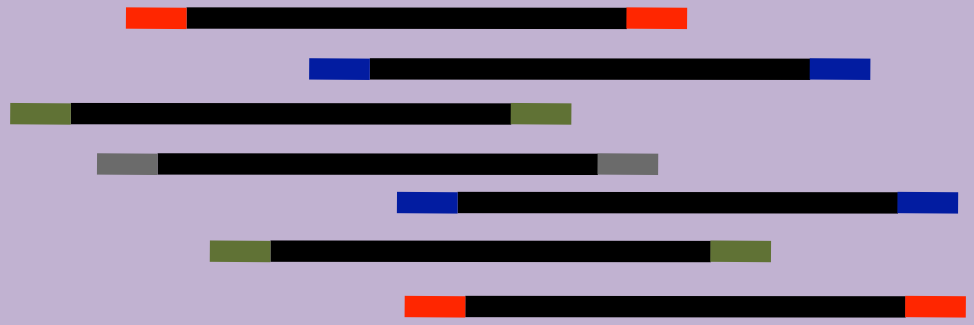
- Clinical and research applications
- Analyzes 1-100's of genes
- Identifies variation in genes previously established as disease causing
- Analyzes coding region of genes
- Costly (several $1000s)



Whole exome sequencing (WES)

- Relies on predetermined exon identification.
- High coverage 50-100X
- Only includes ~1-2% of genome
- Does not include regulatory regions
- Approximately $1000 (research setting)

Whole Genome Sequencing



150 base pair sequences X 3 billion base pairs
X 40 fold coverage

= 800,000,000 sequences to align per genome

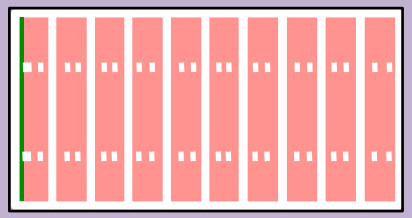
Whole genome sequencing

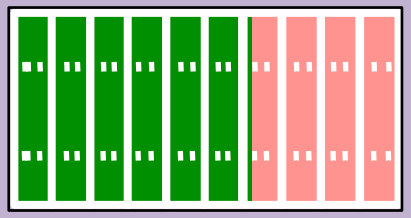
- **Comprehensive**
  - **Single nucleotide polymorphisms (SNPs), insertion/ deletion (indels) polymorphisms, splice site variants, structural variation**
- **Potential to identify new genes**
- **Potential to identify multiple pathologic variants as modifiers**
- **Cost ~ $ 3,000**

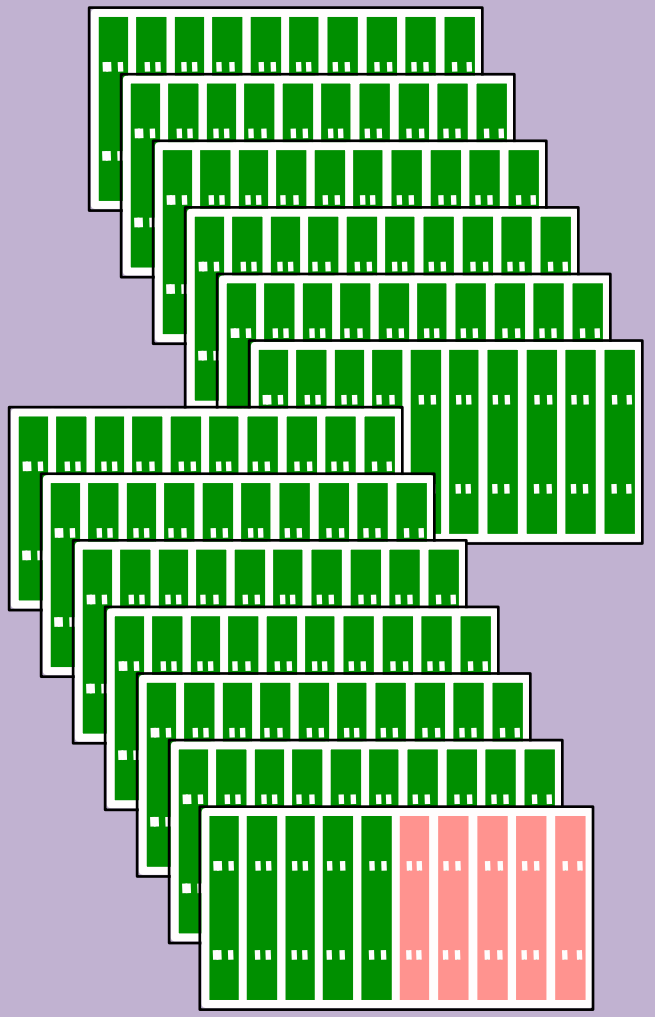
Gene Panel
50X coverage
1,581,742 bp
0.16Gb

Exome
50X coverage
62Mbase genome
6.2Gb

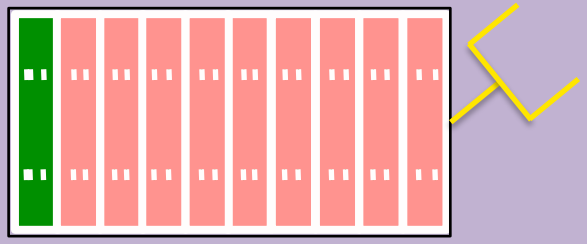
Genome
35X coverage
2.8 Gbase genome
125Gb





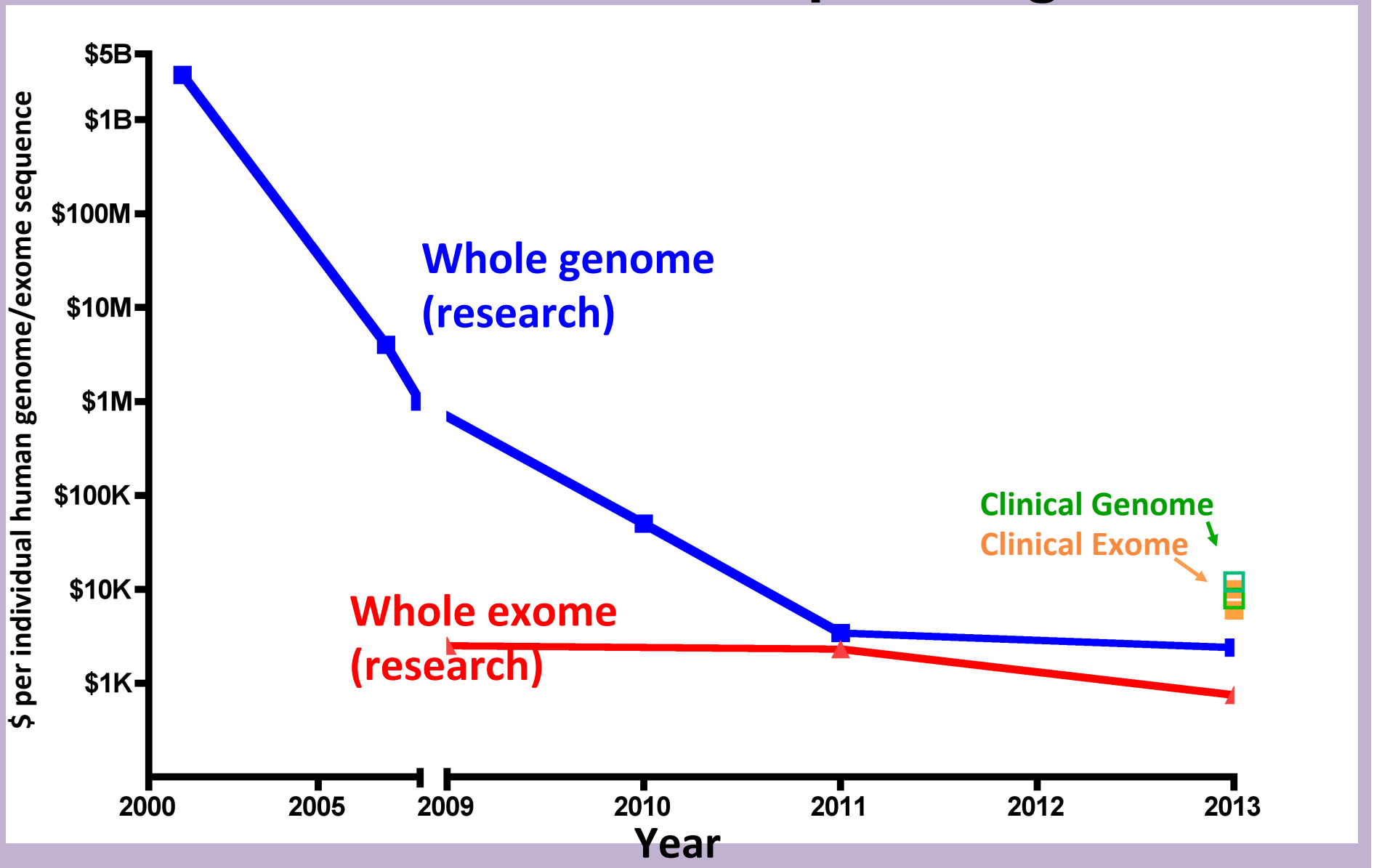


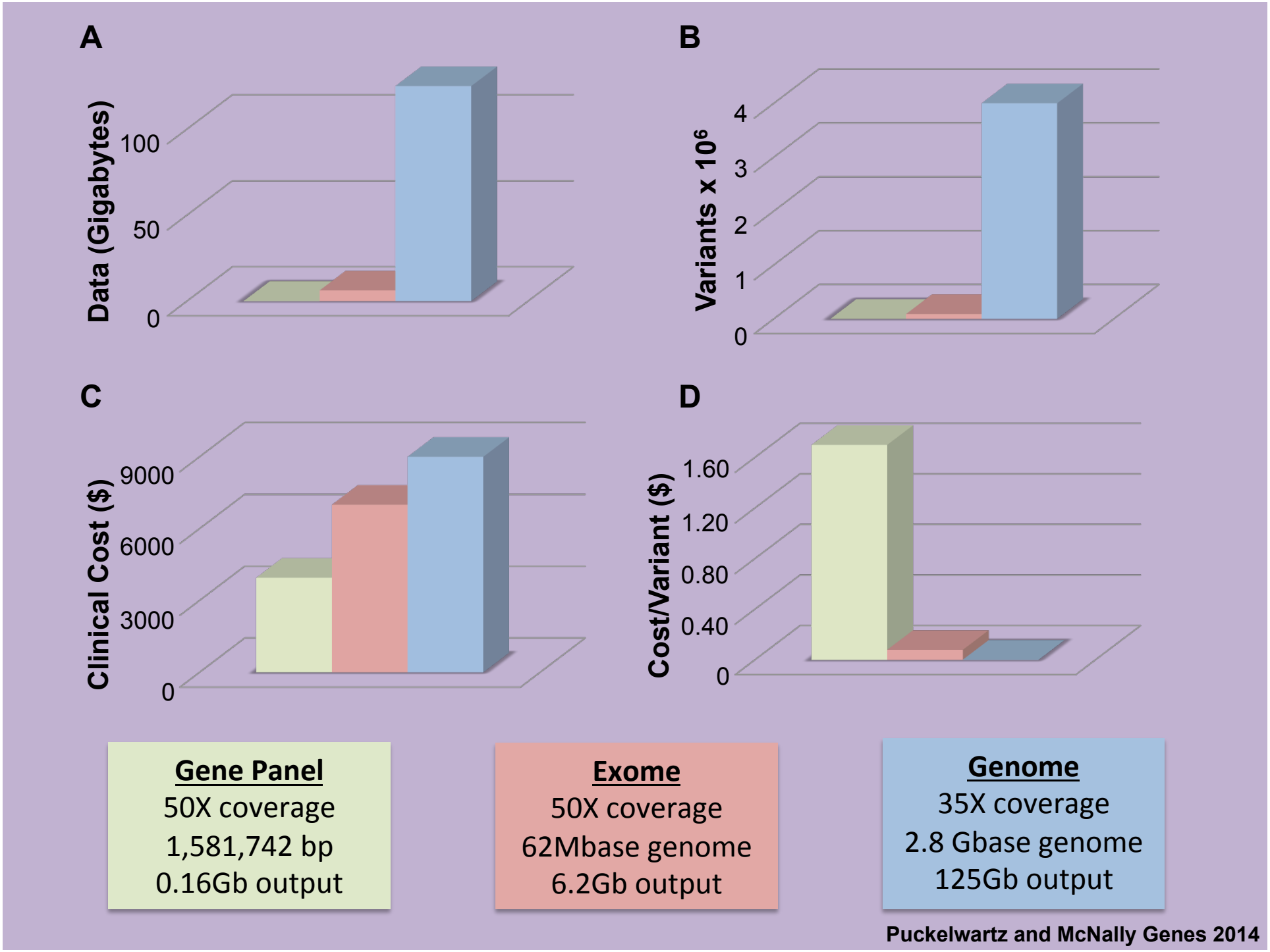
1 Gigabyte=

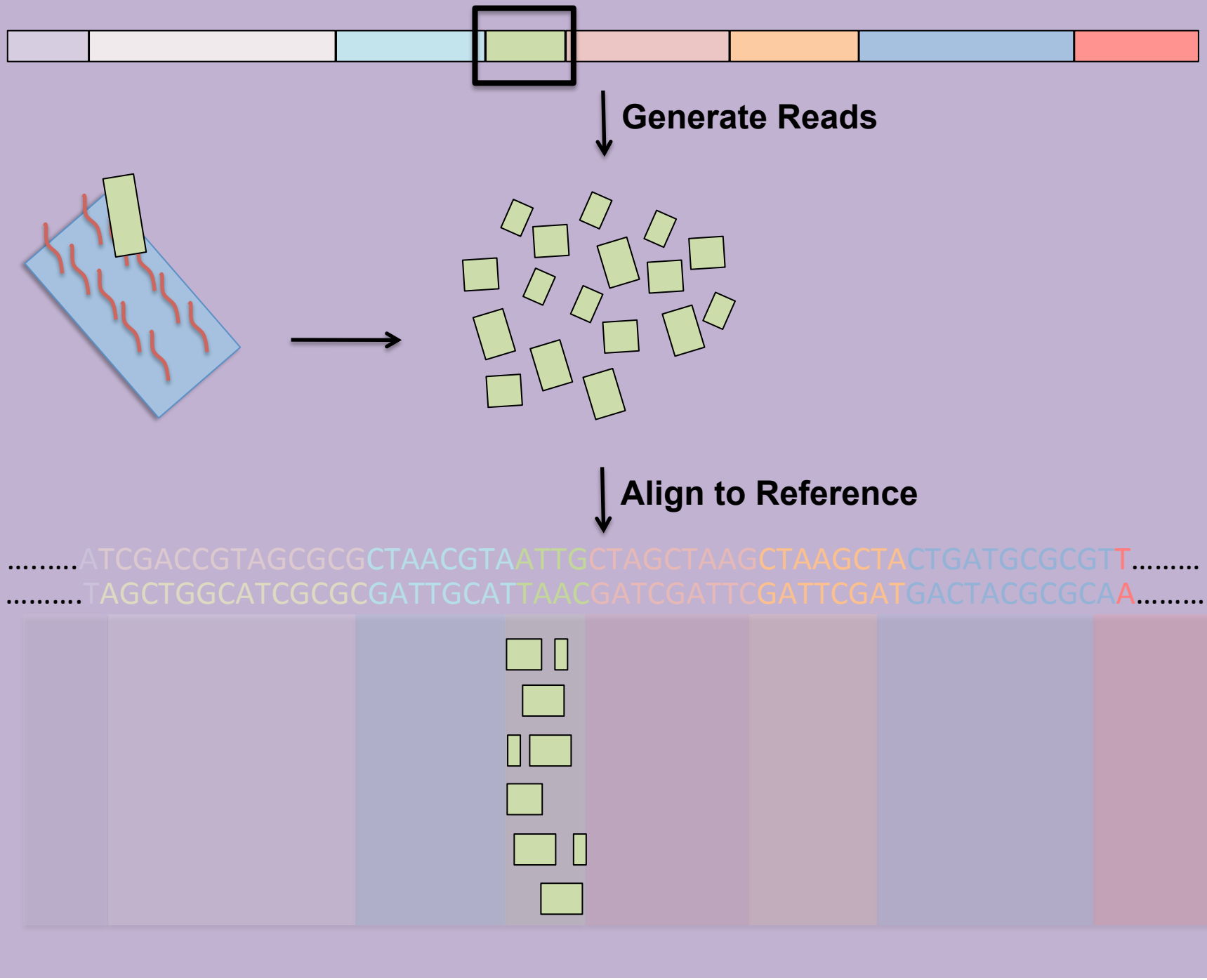


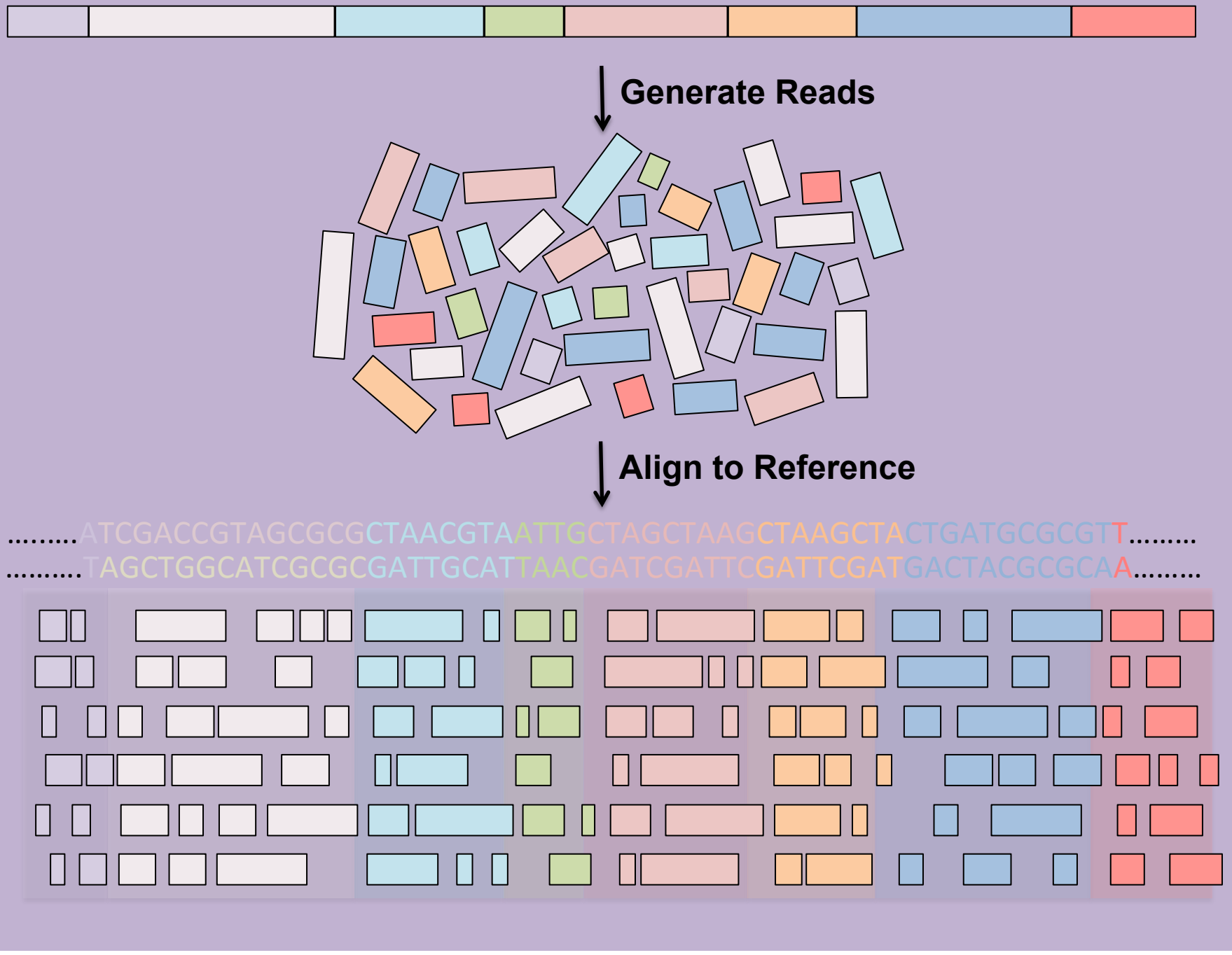
10 yards of books

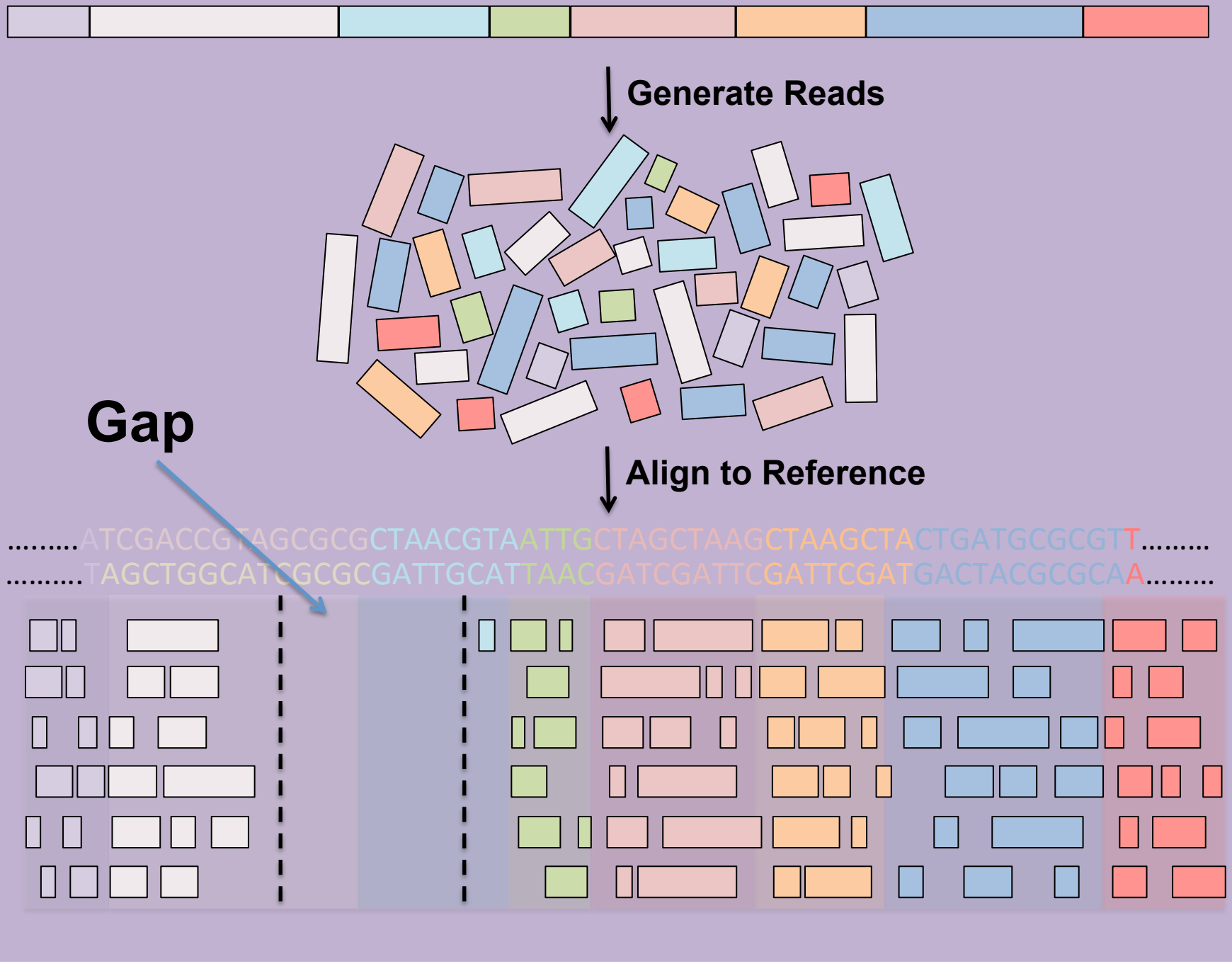
Costs for Whole Genome Sequencing and Whole Exome Sequencing




A
Data (Gigabytes)
100
50
0
B
Variants x 10⁶
4
3
2
1
0
C
Clinical Cost ($)
9000
6000
3000
0
D
Cost/Variant ($)
1.60
1.20
0.80
0.40
0
Gene Panel
50X coverage
1,581,742 bp
0.16Gb output
Exome
50X coverage
62Mbase genome
6.2Gb output
Genome
35X coverage
2.8 Gbase genome
125Gb output
Puckelwartz and McNally Genes 2014


Generate Reads
Align to Reference
.....ATCGACCGTAGCGCGCTAACGTAATTGCTAGCTAAGCTAAGCTACTGATGCGCGTT.....
.....TAGCTGGCATCGCGCGATTGCATTAACGATCGATTCGATTCGATGACTACGCGCAA.....


Generate Reads
Align to Reference
.....ATCGACCGTAGCGCGCTAACGTAATTGCTAGCTAAGCTAAGCTACTGATGCGCGTT.....
.....TAGCTGGCATCGCGCGATTGCATTAACGATCGATTCGATTCGATGACTACGCGCAA.....


Generate Reads
Gap
Align to Reference
.....ATCGACCGTAGCGCGCTAACGTAATTGCTAGCTAAGCTAAGCTACTGATGCGCGTT.....
.....TAGCTGGCATCGCGCGATTGCATTAACGATCGATTCGATTCGATGACTACGCGCAA.....

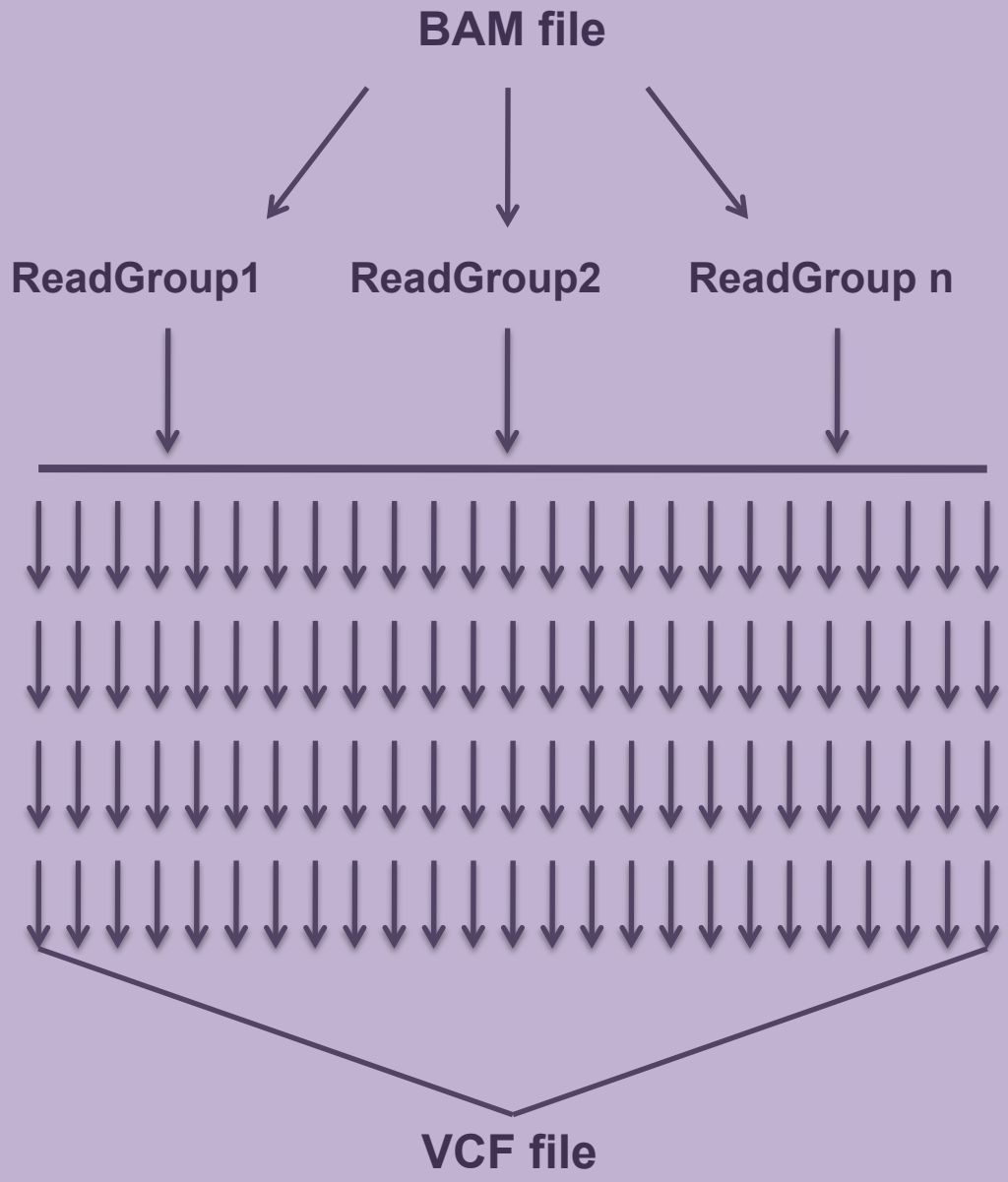
Why use a supercomputer?

- Whole genome analysis is severely limited by time constraints (many steps, many of which are computationally expensive)
- A parallel machine's size and speed allows for efficient analysis of multiple genomes
- A parallel machine allows for testing of new methods, algorithms, parameters and for comparing with old ones

A Parallel workflow: MegaSeq

- Uses a Cray XE6 to achieve the parallelization required for multiple genome analysis
- Relies on open source software (free, at least to academia), BWA, SAMTOOLS, BAMTOOLS, PICARD, GATK
- Employs a MapReduce approach and multithreading to take advantage of the distributed nodes.

Aligning and calling variants with Megaseq



Step

Sam2Fastq
by Readgroup

Align
Compress
Sort
(BWA mem)

Merge ReadGroups
Split by Chromosome

Mark Duplicates (Picard)
IndelRealigner (GATK)
Recalibrate Scores (GATK)

Call Variants (Haplotype Caller)
Generate VCF files
Merge VCF files

Puckelwartz et al. Bioinformatics 2014;30:1508-1513

Raw data extraction phase

- Extract raw sequence from bam files, only necessary when fastq not provided
  - Picard Suite SamtoFastq
- Split patient sequences by readgroup
  - each ~150bp sequence has a unique identifier based on machine, sample, library, lane and flow cell location
  - provides easy “data packet” for downstream analysis
- ~12 hours
- Other approaches are possible (bamutil, biobambam)

Alignment-BWA

- Burrows-Wheeler Aligner (BWA) uses gapped alignment
- One node per read group
- Alignment scales perfectly (linear speedup with the number of cores used)
- Aln/sampe:
  - Trimmed all short sequence reads to a quality of 30
  - Convert alignment files to readable format using BWA-tpx. Conversion does not scale perfectly
  - ~10 hours
- mem:
  - no need for trimming
  - No need for conversion
  - no scaling issues
  - <<~ 3 hours, depending on number of readgroups

Cleaning Computation Requirements

- After alignment, readgroups per genome are merged, then each genome is split by chromosome
- For cleaning, each step was performed on 25 cores concurrently
  - (3 nodes, plus one core – 24 chromosomes + mitochondria)
- Threading was used, where available
- 58GB memory/number of jobs per node for java programs
- Java programs were also given 2 threads for GC which better managed memory issues allowing us to pack more jobs per node

Cleaning alignments

- **Picard & Samtools process the aligned reads to prepare for variant calling**
  - **Mark Duplicates (Picard) identifies and flags duplicates that can be produced during library preparation**
  - **Megaseq1a: Samtools does all the splitting & sorting business – very fast, neat**
  - **Megaseq1b: splitting is done directly after bwa mem, without any step to disk; bamtools**
- **Megaseq1a: ~9 hours; Megaseq1b: ~1-3 hours?**

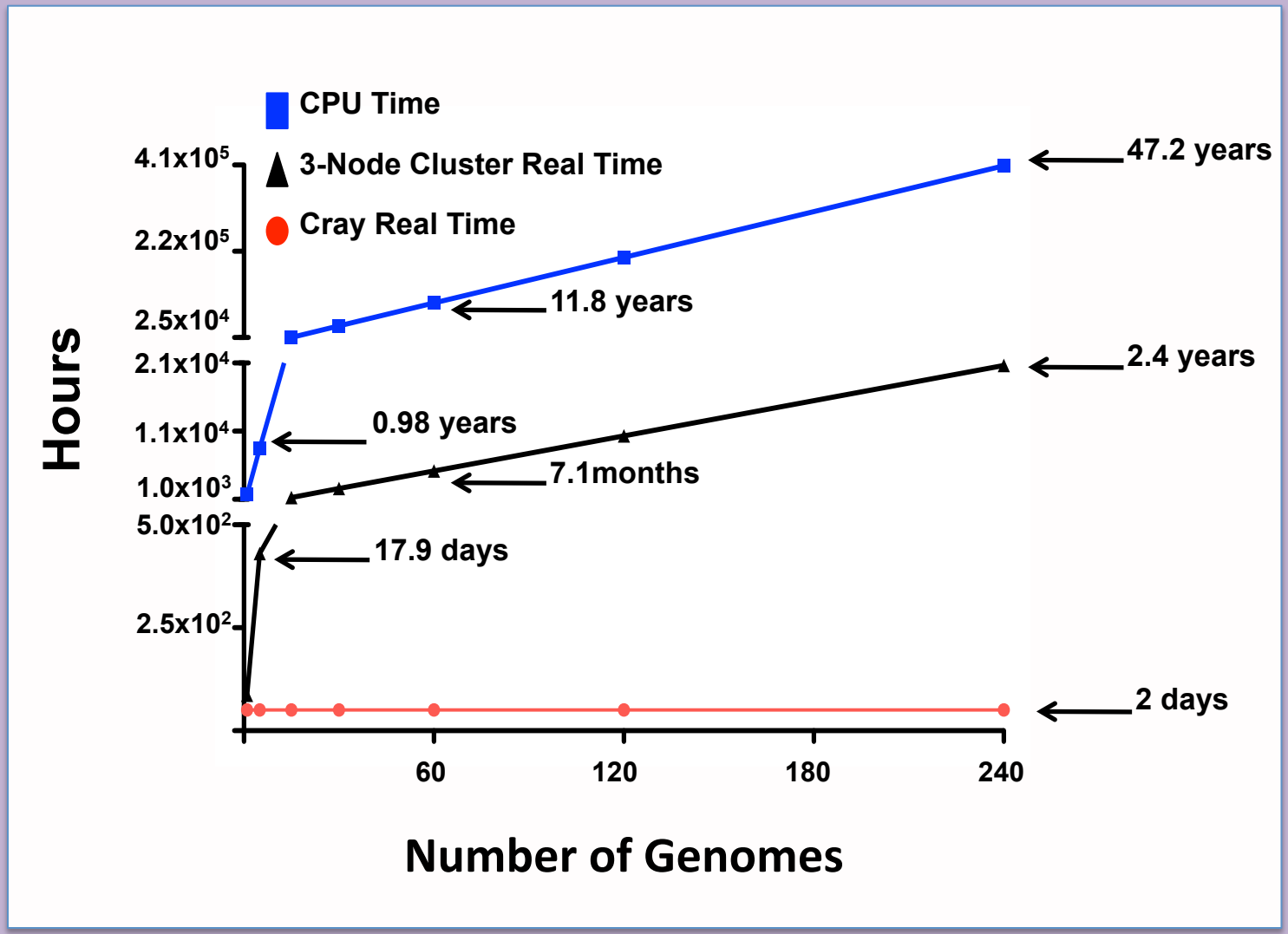
Genome Analysis ToolKit (GATK)

- **Broad Institute**
- **Local realignment around indels:**
  - **alignment is performed using each sequence read individually**
  - **uses multiple alignments at the suspected indel to identify mismatches**
- **Base quality score recalibration:**
  - **more closely matches the actual probability of mismatching the referent genome**
  - **corrects for any variation in quality between machine cycle and sequence context**
- **~7-12 hours**

Variant Calling

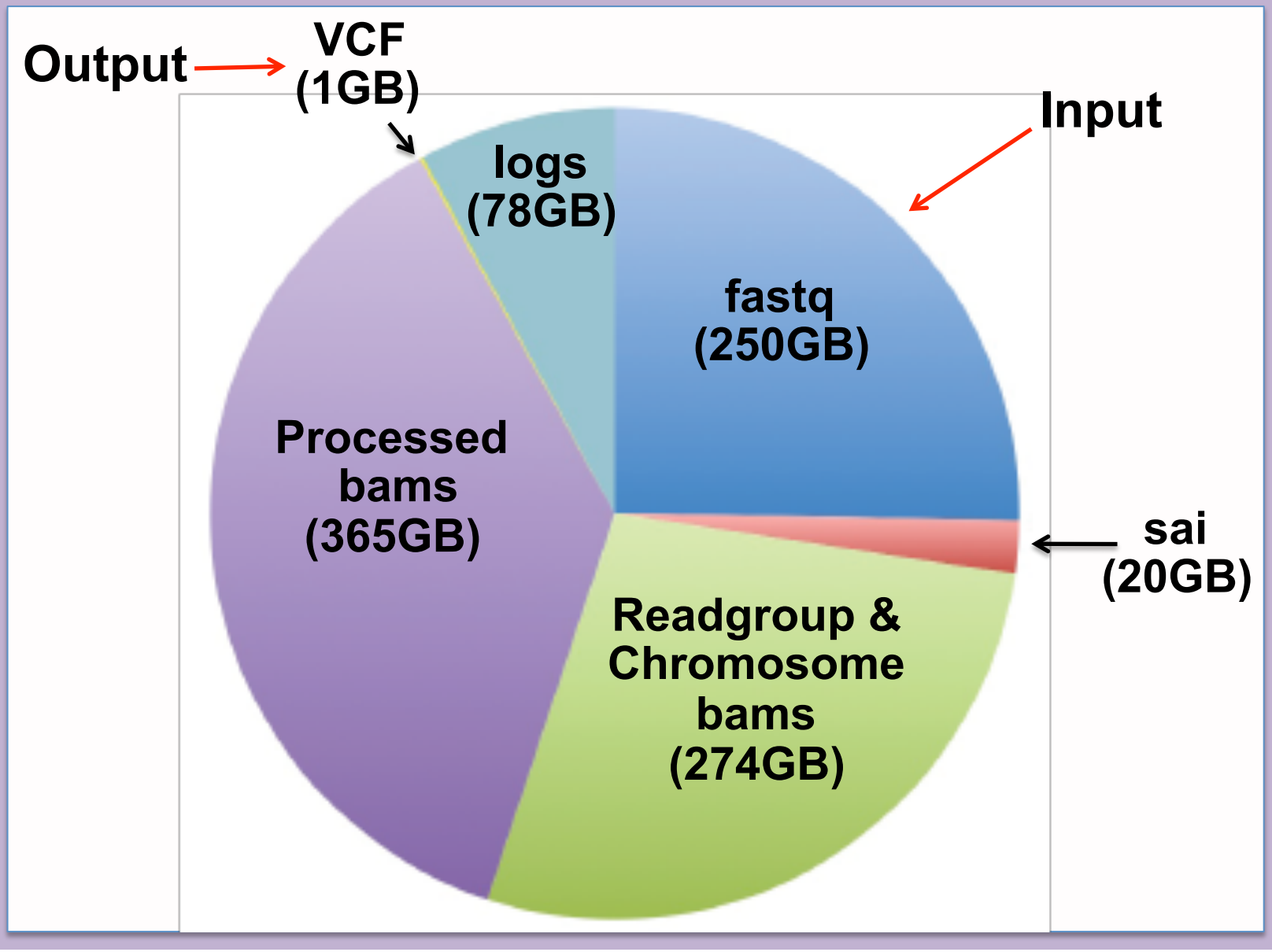
- Haplotype Caller (GATK) calls SNVs (single variants) and insertion/deletion variants simultaneously
- ~3 nodes with 25 X concurrency per genome
- Variants were filtered based on quality metrics including quality score, depth and others
- ~1-4 hours

CPU and real time constraints of Whole Genome Analysis (WGA)

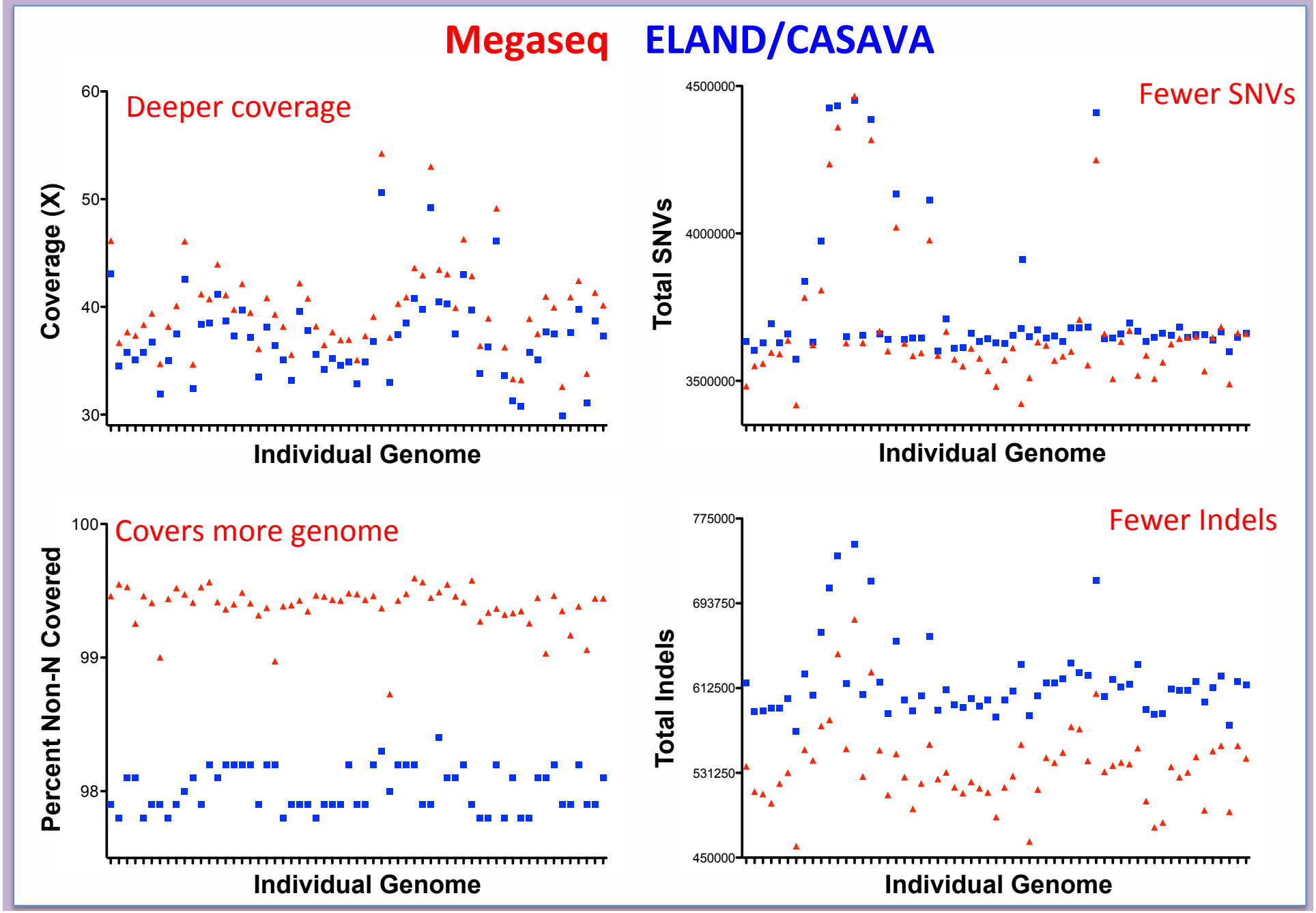


Puckelwartz et al. Bioinformatics 2014;30:1508-1513

Disk space constraints of Genome Analysis (1TB/genome)



Megaseq covers more of the genome with fewer variants



Individual Genomes

- ~3-4 million SNPs differ from the reference per genome
- 130-400 rare non-synonymous variants per genome
- 10-20 Loss of function
- 0-8 variants per genome are predicted “highly damaging”

Acknowledgements

Elizabeth McNally, PI
Megan Roy-Puckelwartz
Alexis Demonbreun
Dave Barefield
Eugene Wyatt
Ellis Kim
Joshua DeJong
Maddie Allen
Brandon Gardner
Quan Gao
Bridget Biersmith
Andy Vo
Kay Marie Lamar
Michele Hadhazy
Judy Earley
Will Montag
Jessie Golbus

Argonne Nat'l Labs/CI
Ian Foster

Gerald Dorn - Wash U
Euan Ashley - Stanford
Rick Dewey - Stanford
Sharlene Day - Michigan
Tom Cappola - Penn