It's not 'just' about BIGDATA

How to get to actionable clinical knowledge from BIGDATA

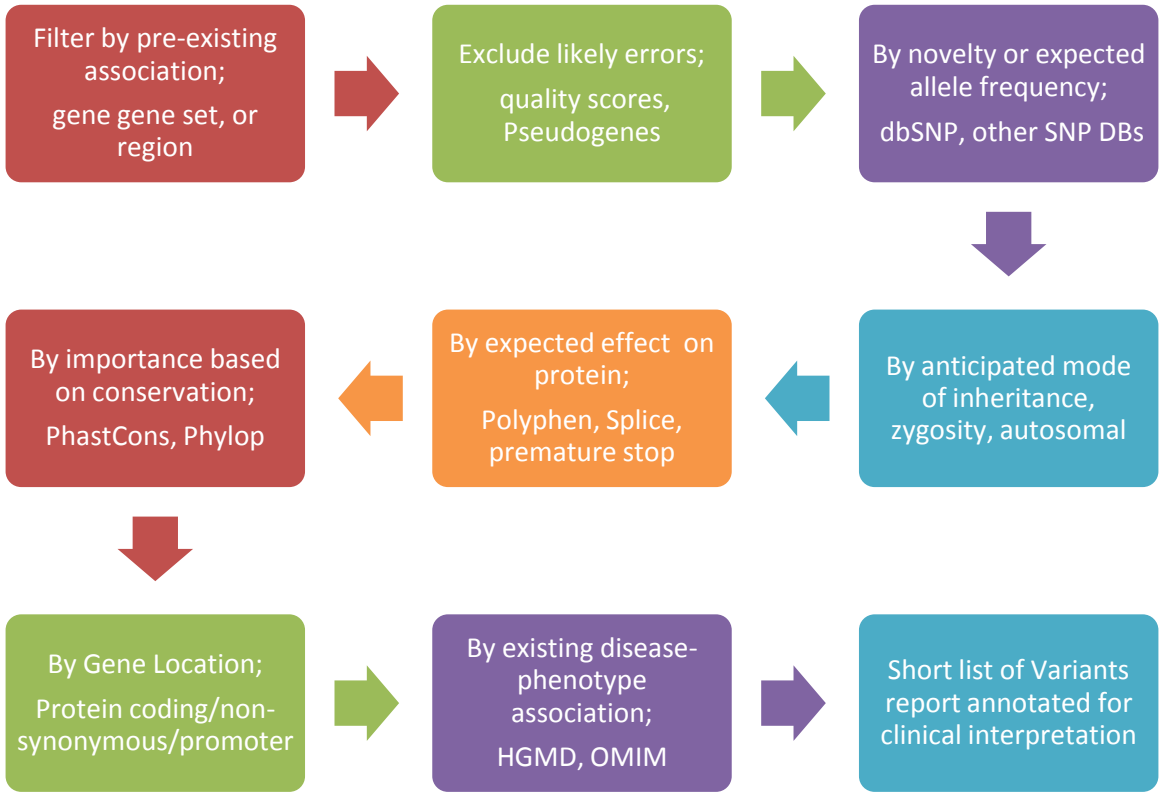
Subha Madhavan, Ph.D.
Innovation Center for Biomedical Informatics
Georgetown University

Scale

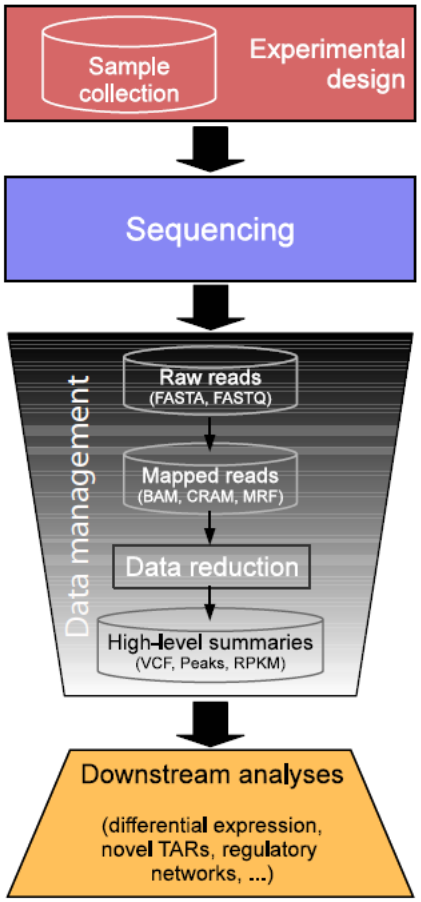
- The world's current sequencing capacity is estimated to be 13 quadrillion DNA bases a year requiring 3 or more exabytes of storage
- The NIH-funded 1000 Genomes Project deposited 200 terabytes of data into the GenBank archive during the project's first 6 months of operation, twice as much as had been deposited into all of GenBank for the entire 30 years preceding
- Million Genome projects??

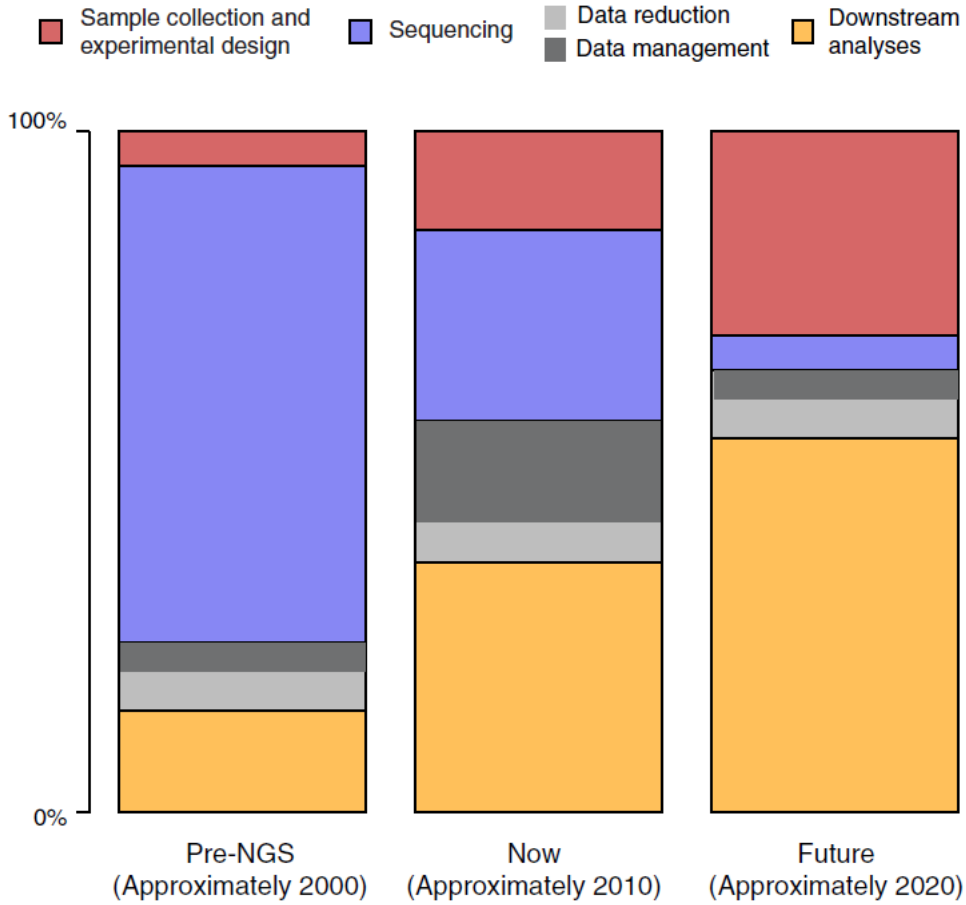
Motivating Scenarios

- BIGDATA for the Clinician
  - Personalized Genomic Medicine



Real Cost of Sequencing





Sboner *et al*. *Genome Biology* 2011, **12**:125

Motivating Scenarios (2)

- BIGDATA for the Epidemiologist – Public Health
  - Monitoring disease incidence and pathogenic outbreaks
  - Real time microbial detection for controlling infectious diseases

Levels of data
MegaGenomes Data Hierarchy

Level 0

- Unaligned raw reads
  - Single
  - Paired end/mate pair
  - Complex sub-reads
- Quality scores
  - Per-base position
  - Homopolymer flow intensity
- Access pattern:
  - Sequential over reads (parallel over sub-reads)
  - Write-once, read rarely
- Storage:
  - Not necessary, if level 1 is a strict superset

Level 1

- Aligned reads
  - Primary and secondary alignments
  - Alignment quality scoring
- Access patterns:
  - Write-once, read infrequently
  - Interval queries, sequential
  - Mate/subread lookup, quasi-local random access
- Storage:
  - Highly compressed application-specific binary, 100 GB/genome
    - BAM, H5BAM
  - Can also be re-aligned to new references if unaligned sequences are stored, removing need to preserve level 0 data

Level 2

- Called consensus sequence
  - One or more nucleotide sequences that represent the consensus of the read data
  - Aligned to reference coordinates, variable ploidy
  - Allele-specific ploidy estimates
  - Missing and reference nucleotides represented
  - Consensus quality scores
- Structure graph
  - Based on chimeric reads and clustered sub-reads
- Access pattern:
  - Write-once, read-infrequently
  - Interval queries
- Storage
  - Highly compressed application-specific binary

Level 3

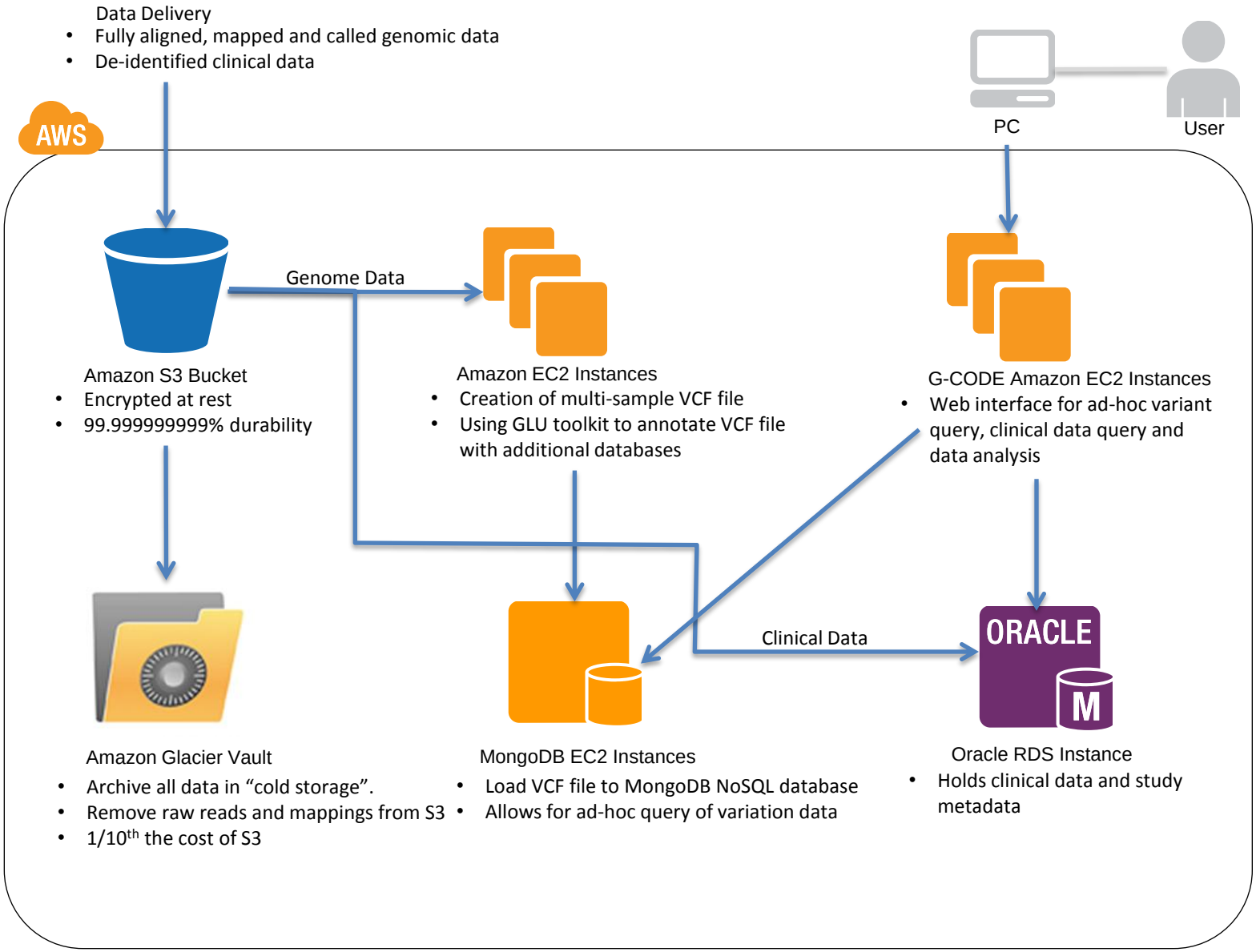
- Extensible, annotated, multi-subject “variant file”
  - VCF-like, but variable-ploidy, preserving missing and reference calls and quality scores
- Access pattern
  - Monolithic offline updates, read-often
  - Interval and subject subset queries
  - Filters by annotation

Level 4 (actionable data)

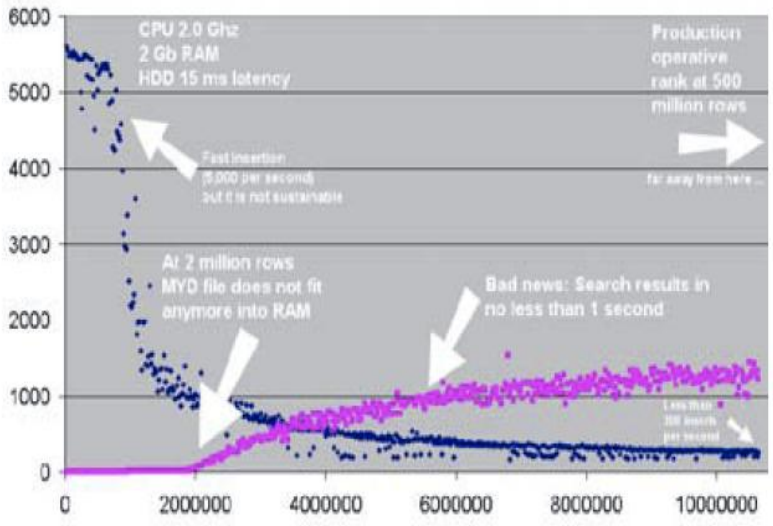
- Reduced, hypothesis-specific, integrated dataset suitable for mining
  - Genomic data orders of magnitude smaller than levels 0-2

Some practical solutions we are leveraging

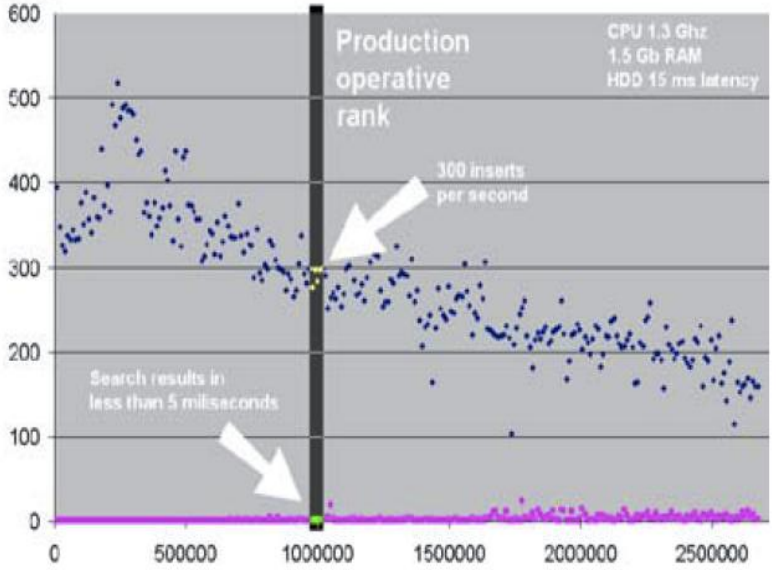
Elastic storage and computes on the cloud



NoSQL Databases

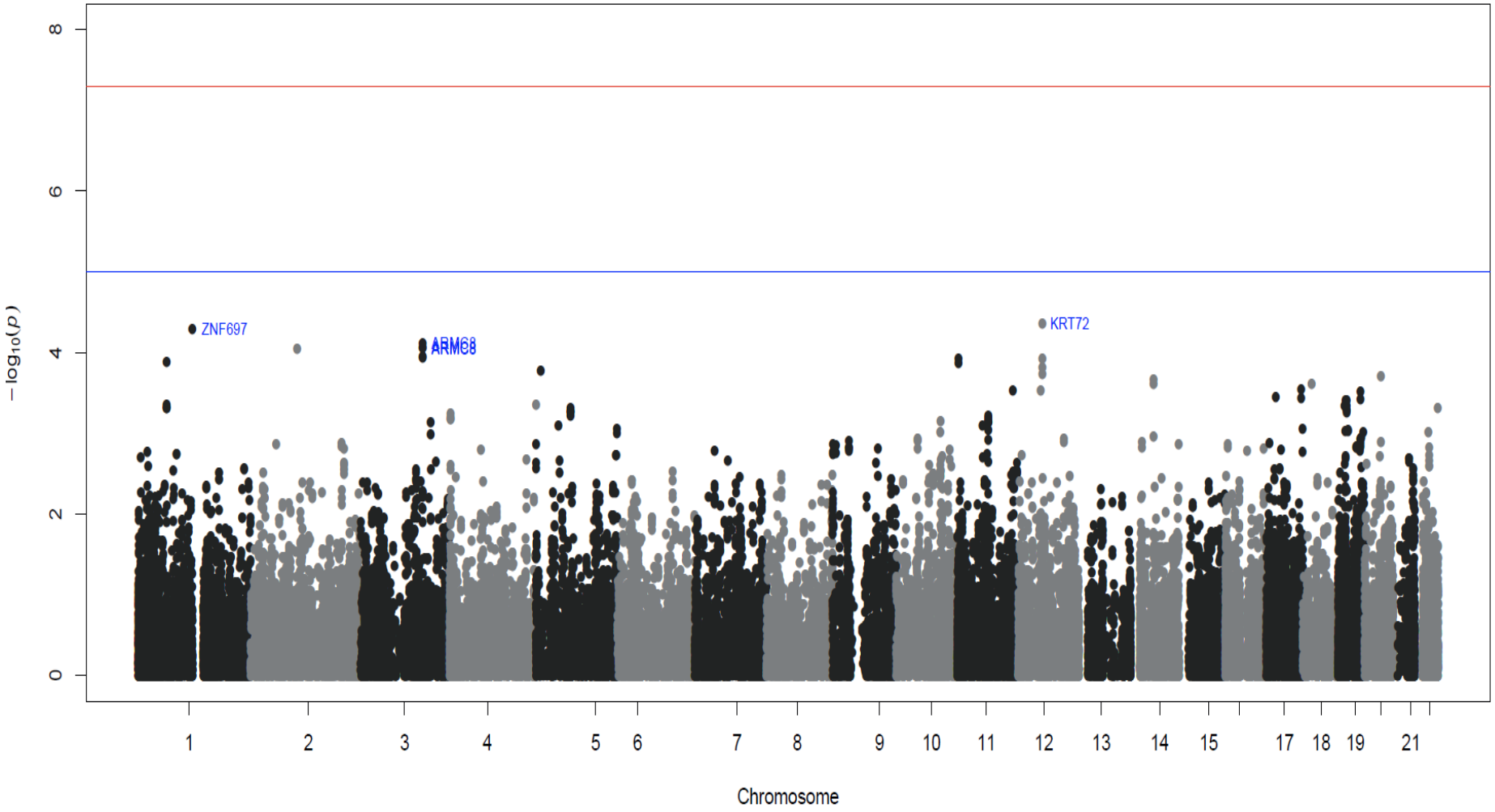


(a) Performance of SQL for a web-search query



(b) Performance of NoSQL for a web-search query

Smart Data Visualization



It’s not “just” about BIGDATA

BIGDATA → Actionable Knowledge

- Need better visualization techniques
- Need smart/parallel processing
- Policy needs to catch up with technology
- Cost is an issue

Innovation Center for Biomedical Informatics (ICBI)

http://informatics.georgetown.edu/