

Implications of Affordable Whole Genome Sequencing

Quality and Management of WGS Data



Clifford A. Reid, Ph.D.
President & CEO

June 15, 2010

Agenda: Two Topics

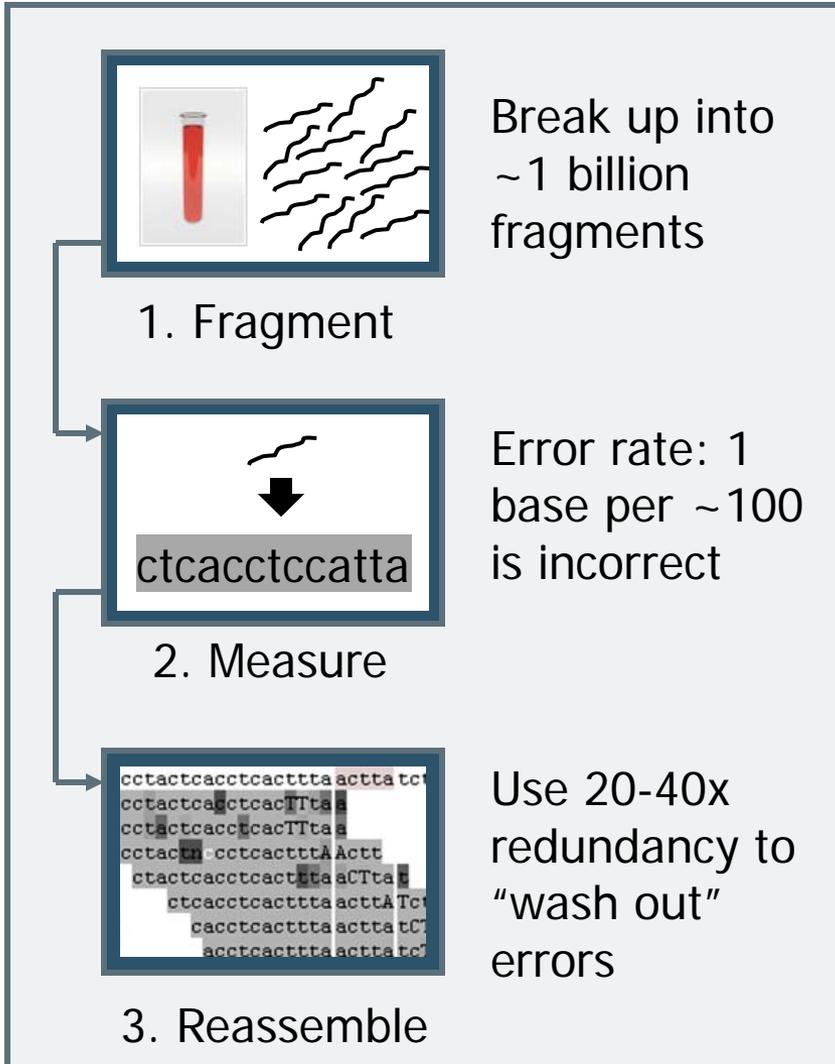
WGS Data Quality

- Simple observations and calculations
- Implications for discovery research
- Implications for clinical applications

WGS Data Management

- Simple observations and calculations
- Implications for discovery research
- Implications for clinical applications

WGS Data Quality



- Markers are **measured**, genomes are **calculated**
- Quality of calculated result depends on:
 - a. quality of individual measurements,
 - b. amount of redundancy, and
 - c. quality of heuristic assembly algorithms
- Software that assembles genomes is complex
 - 1.75 million lines of code
 - Parameters (heuristics) that trade off types of errors (mistakes vs. misses)

WGS Data Quality

1 error per 100,000 bases ...

← 3 billion base positions →

... ATCGATCGGATGACATCACGATTCATCA...

30,000 errors per genome

1% of genome codes for proteins

... AATCA — TCGATTCA — TCAA...

300 errors in coding regions

... an error in ~300 of ~30,000 genes (1%)

- Great for research
 - 100 samples → probability 5 of 100 samples have (random) error in same genes = 1 in 10 billion
 - 95 correct genes overwhelm the 5 incorrect, lead to correct research result
- Unacceptable for clinical use
 - Single genome per patient
 - if this was your genome, 1% of genes “wrong” too high
 - Need to independently validate every decision variant

How Much Accuracy is Needed?

Proposal:

1 error per 10,000,000 bases ...

← 3 billion base positions →

... ATCGATCGCTGACATCA~~X~~CGATTCATCA...

300 errors per genome

1% of genome codes for proteins

... AATCA — TCGATTCA — T~~X~~AA...

3 errors in coding regions

... an error in ~3 of ~30,000 genes =
99.99% accurate

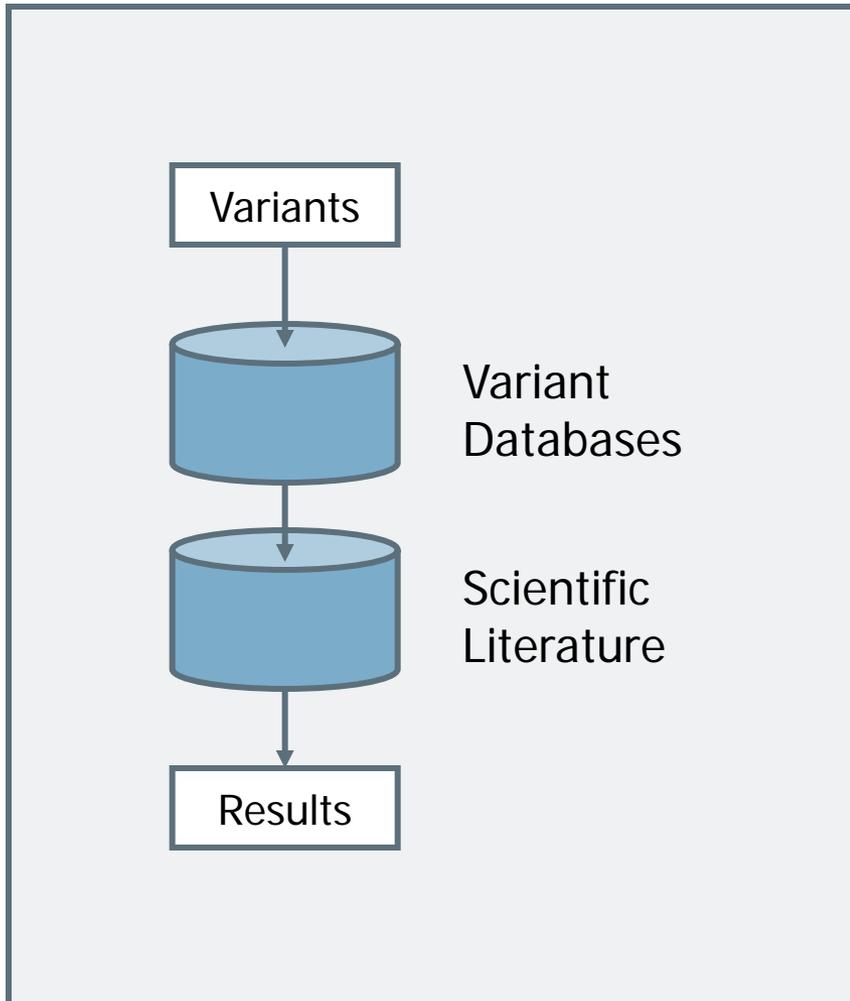
- How accurate does WGS have to be for clinical applications?
 - Judgment call, but here is a proposal
 - 100x more accurate: $10^{-5} \rightarrow 10^{-7}$
 - Diagnostic quality for many (not all) applications
- How long to achieve and deploy clinical quality?
 - 2-4 years in reference labs
 - 6-? years in point-of-care instruments
 - Challenge: certify 1+ million lines of software, for any genome

WGS Data: Size of a Genome

1 Whole Human Genome

	<u>Size (Gb)</u>	<u>\$/Yr (AWS)</u>
Images	3,000	\$5,000
↓		
Reads	300	\$500
↓		
Genome	30	\$50
↓		
Variants	0.3	\$0.50
↓		
Annotations	<1.0	\$1.50

- How much data is generated?
 - Images (raw data) large, but immediately discarded
 - Reads (used to create variants) 1/10th of images
 - Variants (differences from reference genome) very small
- How much data is kept?
 - Research: moving to variants
 - Clinical: until errors reduced to near zero, read data is important
 - Soon to cost as much to store genome as to resequence it: storing molecules cheaper than storing bits



- Variant databases
 - dbSNP: 23,653,729 entries in the variation table
 - Databases growing rapidly with new discoveries (rare variants, structural variations, ...)
- What do variants mean?
 - Mostly unknown – next great WGS research challenge
 - Requires a knowledge base of discoveries (scientific literature)
 - Huge challenge in medical education

WGS Quality and Data: Other Issues

- Validation
 - Different WGS methods give different results
 - Which is right? Each makes a different set of mistakes
 - Normal genomes have “ground truth” – independent validation works
 - Cancer genomes much harder
- Quality Control
 - Even simple procedures can fail (e.g. 23andMe)
 - WGS is much more complex than genotyping
- WGS Knowledge Bases
 - Curation
 - Medical education
- Electronic Health Records (EHR)
 - Format and interchange
 - Privacy and security

Conclusion: Implications of Affordable WGS

- Research: revolution is underway
 - WGS quality and cost supports powerful research studies
 - Data management: variants and outsourcing (cloud computing)
- Clinical markers: discovery is underway
 - WGS studies will discover new markers with clinical importance
 - Markers easy to measure, easy to look up meaning and take action
- Clinical WGS: two future applications
 - Cancer genomes: too variable in structure to rely on markers
 - Universal genetic panel: part of EHR, look up when needed

Questions

