“Bigomics”: Challenges and promises in large scale sequencing projects

Theodore M. Wong
Ken Yocum

MSST 2014

illumina®

The promise of the Illumina $1000 genome

- Less interesting: sequence populations to do anthropological studies
- Much more interesting: sequence populations of cancer patients to discover the causes and cures for cancer

Curing cancer – the grand challenge of genomics

- Occurs in ~1.6 million new patients each year in the United States
- Kills ~0.6 million patients each year
- Affects ~13 million people in any given year
- **Caused by genetic mutations**

The goal of genomic oncology

Create a one million whole genome + phenotype record repository of cancer patient genomes, to facilitate research into the genetic causes of cancer and individualized treatments for cancer patients

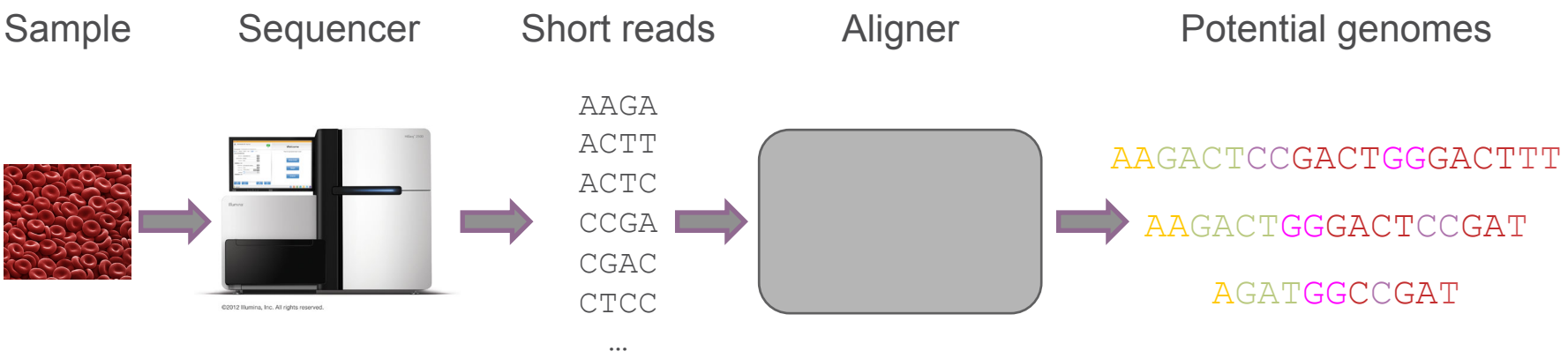
- Why the whole genome?
  - May enable treatments based on a combination of mutated and non-mutated genes
  - Enables reanalysis of newly identified functional regions (e.g., ENCODE project)

- Why a million?
  - Enables statistically valid discovery of cancer-causing mutations

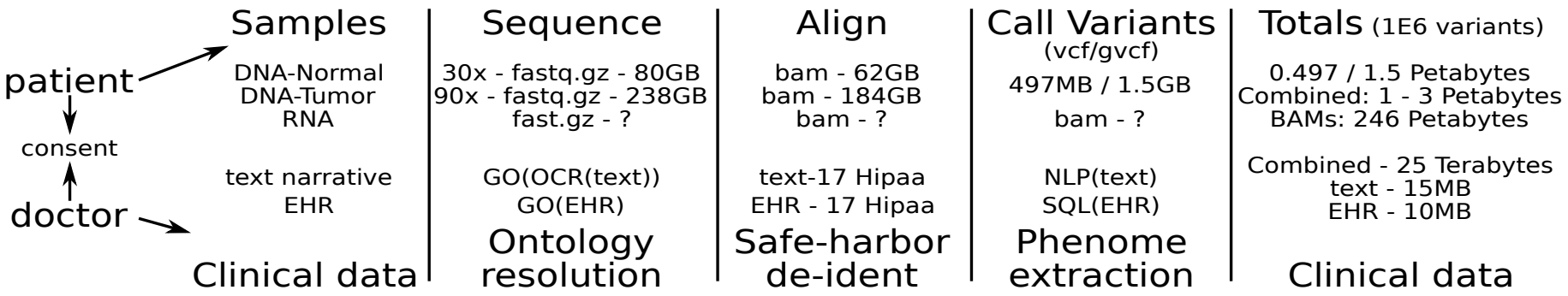
Challenges facing genomic oncology today

- Extracting useful, structured genome + phenotype (G+P) records from raw sequencer data and electronic health records
- Getting consented records for a meaningful-size cohort of study participants
- Protecting the privacy of G+P records for participants / patients

Genome sequencing in one slide



Phenotype extraction in one slide



The state of genomics

- Illumina sequencers can sequence a genome for $1000
- Aligners for producing binary alignment maps (BAMs) are increasingly accurate
- A modern 1 PB storage rack holds up to 10000 whole-genome BAMs
- Variant call format (VCF) files provide more compact representation than BAMs for storage and genome analysis – 10 GB vs. 100-200 GB

Moving from genomes to genomes + phenotypes (G+P)

- Genomic records track mutations
- Phenotypic records track patient vital statistics and disease symptoms
- Need both to find links between mutation and disease, or to screen for disease

The challenge of extracting G+P records

- Different variant callers produce differing output from the same input BAMs
  - Talwalkar *et al*: SMaSH toolkit for benchmarking variant caller accuracy
- Variant caller algorithms still need improvement when identifying large structural variants: insertions/deletions, copy number variations, translocations…
  - Paten *et al*: Cactus graphs for variant representation
- The VCF format is not well suited to representing structural variants
  - Massie *et al*: ADAM genomic formats for cloud-scale genomic computing
- Electronic health records (EMRs) are stored in inconsistent / incompatible schemas
  - Natural language and medical billing code post-processing to extract EMRs

Getting consent to us G+P records

- Institutional review boards / state and Federal laws require patient consent before commencing research / treatment using patient genomic records

- HIPAA requires patient consent for release of protected health information

The challenges of getting consents

- Incompatible consents hamper aggregating small corpuses into a large corpus
  - Follow the principal investigator / clinician, not the patient
  - Cover narrow use of genomic records / samples
  - Vary by country, state, and institution

- Lunshof *et al*: Open consent model in the Personal Genome Project

Protecting the privacy of G+P records

- Genomes are *not* “protected health information” as yet, but will be soon
- In general, “de-identification” of genomic data sufficient to keep it private

The challenges of protecting record privacy

- Collecting a million genomes is a nationwide/worldwide effort, involving often contradictory privacy and consent laws from several jurisdictions
- Combining genomic and phenotypic records is necessary to explore the genetic basis of disease, but inherently combines sensitive health information
- De-identification is not foolproof
  - Gymrek *et al*: Re-identified 1000 Genome Project donors by cross-referencing genomes with Y-chromosome and surname records at a social genealogy website
- Common consensus holds that technological privacy solutions are fallible

Related work

- Global Alliance: Exploring worldwide sharing of G+P records
- Broad Institute (Harvard / MIT): Exploring the genetic causes of disease
- UC Berkeley: Million cancer genome warehouse; genomic storage formats; genome pipeline processing algorithms
- UC Santa Cruz: Applying computational / “big data” techniques to genomics
- Illumina, Google, and other companies

Thank you!

twong@illumina.com

References (clickable links)

- Gymrek, M. et al., “Identifying Personal Genomes by Surname Inference”. Science 339(6117):321-324, Jan 2013
- Haussler, D. et al., “A Million Cancer Genome Warehouse”. UC Berkeley TR UCB/EECS-2012-211, Nov 2012
- Massie, M. et al., “ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing”. UC Berkeley TR UCB/EECS-2013-207, Dec 2013
- Paten, B. et al., “Cactus graphs for genome comparisons”. J Comput Biol. 18(3): 469-81 Mar 2011
- Talwalkar, A. et al., “Smash: A benchmarking toolkit for variant calling”, arXiv: 1310.8420 [q-bio.GN], 2013