Complete genomics

A Perspective on Using Genomic Information in Medicine

Clifford A. Reid, Ph.D.

April 2011

Overview

Topic

- Challenges facing more widespread and effective use of genomic information in electronic health information technology systems

Discussion

- Models of Genome Data Interpretation
- Genome Data Challenges
- Conclusions

Two Co-existing Models of Genomic Medicine

- TGen, Scripps, HMS, ...
- Medical imaging model
- Certified expert (e.g. cancer pathologist) analyzes data
- Low volume
- Custom

Research Medicine Model

- Myriad, Genomic Health, ...
- Lab diagnostic model
- Validated process (algorithm) analyzes data
- High volume
- Standard

Clinical Diagnostics Model

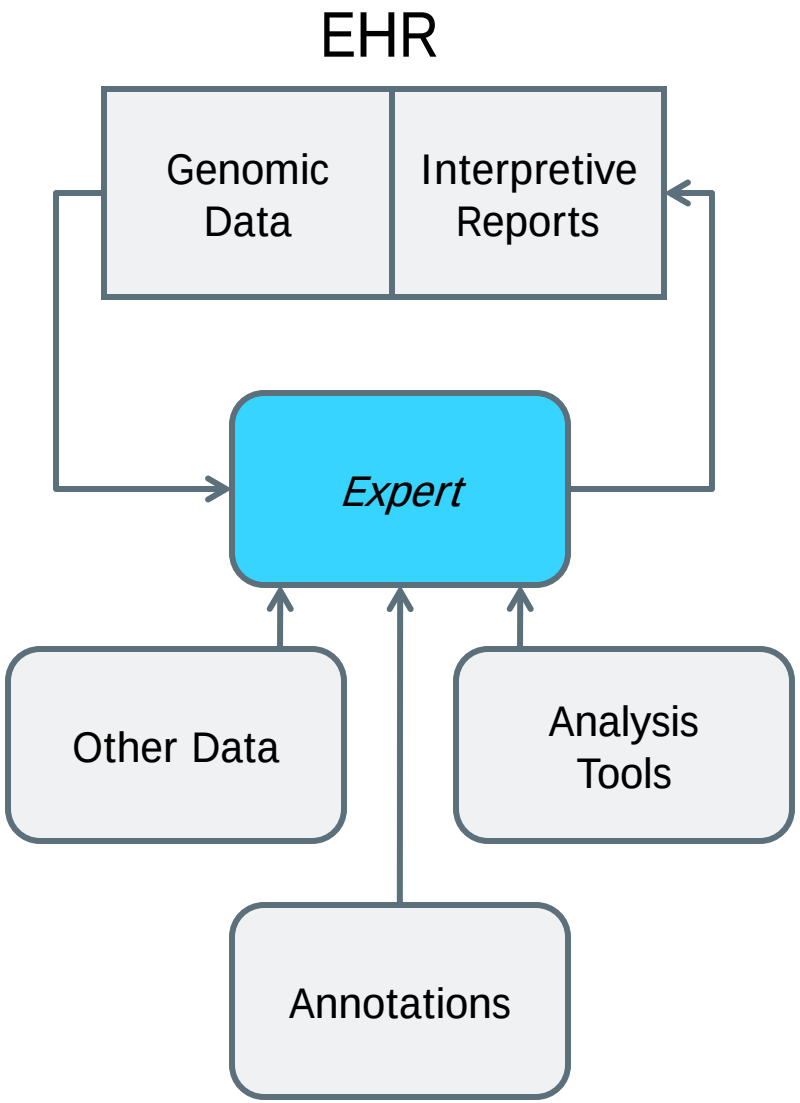
Compete Human Genome Data

summary-XXX.tsv	Summary statistics for genome sequence
var-XXX.tsv	Called sequence with respect to the reference genome, indicating variant and non-variant regions
gene-XXX.tsv	Annotations of variations in known protein coding gene sequences
geneVarSummary-XXX.tsv	Summary of variations in known protein coding gene sequences
dbSNPAnnotated-XXX.tsv	Calls at dbSNP loci
coverageRefScore-XXX.tsv	Base-level coverage and scores
evidenceIntervals-XXX.tsv	Contains the assembled sequence for each variant allele
evidenceDnbs-XXX.tsv	Contains the supporting reads for each assembled sequence
correlation-XXX.tsv	Correlations between assemblies that share supporting reads, for example duplicated regions
reads-XXX.tsv	Reads and base-level quality scores
mapping-XXX.tsv	Initial (pre-assembly) mappings of reads
lib_* _XXXX.tsv	Library file

Table 1 – Reports and files provided, Green files only provided with ST001RM product

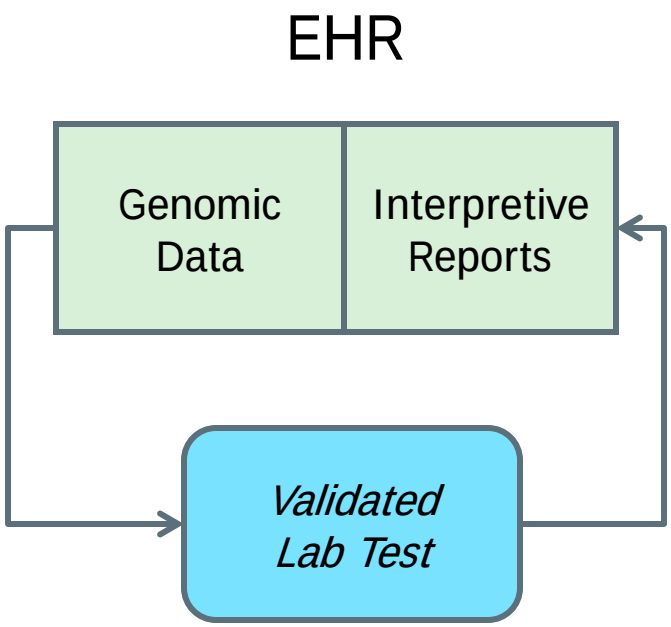
- 6B bases (mostly identical)
- 4M variants
  - 10% complex: indels, SVs, CNVs, ...
- Quality metrics
- Annotations from literature
- Supporting (raw) data
  - 400 Gb (compressed)
  - 800 Gb (BAM standard)

Research Medicine Information Systems



- Start with genome from EHR, or resequence it (update)
- Integrate with other data sources (expression, epigenomic, ...)
- Look up most recent (not curated) annotations
- Analyze (radiologist-physician paradigm)
- Expert writes custom report for the EHR (signed/dated)
- Physician takes action

Clinical Diagnostics Information Systems



- Lab creates validated mappings from genomic data to medical meaning (FDA needs to weigh in)
- Physician sends genome (data) and question, lab report goes into EHR, informs physician's decision
- New/updated mappings require revalidation (long process)

Genome Data Trends: Complexity and Size

- Complexity is growing
  - Broad Nature paper on 38 multiple myeloma genomes: half of key mutations are structural variations (not SNP/indel markers)
  - EHR needs to store all variants in standard representation (very tough to get right)
- Size is shrinking
  - Much of research community still stores raw data (reads)
  - Expect clinical community to use variants only
  - Size/cost comparison:

	Size	Cost (AWS)
Raw Data	400-800 Gb	$600-1,200/yr
Finished Genome	20 Gb	$30/yr
Variants Only	1 Gb	$1.50/yr

Conclusions

- Need to support both models
  - Research medicine: leading edge, custom analysis
  - Clinical diagnostics: validated, standard analysis
- Shared requirements
  - EHR: basic genome data (variants) and signed/dated reports
- Research medicine unique requirements
  - Additional rich/flexible genome data, annotations, etc.
  - Tools for expert analysis
- Clinical diagnostics unique requirements
  - Validated systems (labs) for generating reports (data → analysis)
  - Teach docs how to interpret reports, take actions (analysis → action)

Complete genomics

Questions