Excitement, panic, and genomic data

Sarah Wheelan
What is DNA?
Where is DNA?

- nearly every cell in a human body contains a copy of the entire genome
- length of one human genome: 2 meters
- cheek swab gets lots of DNA
Why sequence DNA?

- Predict disease (risk, presence, stage)
- Determine best treatment for an individual
- Explore population sequence variation
  - Single base changes
  - Insertions, deletions
  - Large-scale rearrangements
Types of sequencing in human genetics

- whole genome
- transcriptome (expression)
- exome (only the parts of the genes that encode protein)
- targeted by protein binding
- resequencing
goal of sequencing

• find variation!
  • disease-causing
  • disease-preventing
  • disease-predicting
  • interesting things that make us all different
the promise of individualized medicine

• detect disease, predict most effective treatment

• follow disease progression over the course of treatment

• predict and prevent disease

  • determine the most effective preventative measure for each person

• predict disease risk
problems in individualized medicine

- incomplete information
  - understanding is still limited
  - major mechanisms (epigenetics) not easily testable, currently
- treatment options are limited
- current clinical records resources insufficient
- doctors & patients ready?
problems in individualized medicine

- major ethical concerns
  - family genetics revealed? twins?
  - prenatal diagnosis
  - testing for untreatable diseases
  - privacy! (insurance, social pressures)
Current sequencing platforms (table obsolete by 4/1/2012)

<table>
<thead>
<tr>
<th>Platform</th>
<th>read length</th>
<th># reads (approx)</th>
<th>run time</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>~400</td>
<td>1 x 10^6</td>
<td>10 hrs</td>
<td>1%</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>100</td>
<td>3-4 human genomes</td>
<td>11 days</td>
<td>0.5%</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>2x250bp</td>
<td>12 x 10^6</td>
<td>1-2 days</td>
<td>?</td>
</tr>
<tr>
<td>SOLiD</td>
<td>75</td>
<td>2-3 human genomes</td>
<td>11 days</td>
<td>0.05%</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>100-140</td>
<td>100s of millions</td>
<td>2-3 hrs</td>
<td>~1%</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>100</td>
<td>?</td>
<td>hours</td>
<td>?</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>thousands</td>
<td>60,000</td>
<td>15 min</td>
<td>15%</td>
</tr>
<tr>
<td>Oxford Nanopore GridION</td>
<td>100kb?</td>
<td>?</td>
<td>variable</td>
<td>?</td>
</tr>
<tr>
<td>Oxford Nanopore MinION</td>
<td>10s of kb</td>
<td>?</td>
<td>variable</td>
<td>?</td>
</tr>
</tbody>
</table>
but computers grow fast . . . right?

- 1956: first hard disk (IBM)
  - size of two refrigerators
  - 5 megabyte capacity
  - $10,000 per megabyte
Data issues

- Single human genome is 3 billion bases, haploid (one base is about 1 byte)
  - At least 5-10GB for sequence data, 50-100GB for medically accurate data
  - 1-5 GB more for analysis data
- World capacity is now 13 quadrillion DNA bases a year
- World record for data transfer over LAN: 23GB/s (65.4 days)
- Satellite transfer: 2000 days
genomes are getting cheap but . . .

- sequence production outstrips advances in storage, computing
- analysis costs are going up, not down
- CPU time
- data transfer
- need more personnel to handle more complex experimental designs

**MUSINGS**

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*
my prediction

• suboptimal use of genomic data

  • >150 genome databases published in 2011 . . . formats not interchangeable

  • continued difficulty in reproducing results and accessing data

• why?

  • good data standards are hard to follow

  • hard to make a database that’s flexible
needs

• genetic counselors: July 2011 — only 13 states license genetic counselors

• fast, reliable databases

• query minimally precomputed data on the fly
conservative goals

- clinical records linked to precomputed variants
- database of precomputed variants not static
- information about genetic variation makes a difference in clinical care
Commentary

Data reporting standards: making the things we use better
John Quackenbush

Address: Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Biostatistics 44 Binney Street, Sm222, Boston, MA 02115, USA. Email: johnq@immy.harvard.edu

Data Standards for Omics Data: The Basis of Data Sharing and Reuse