Raw and processed data

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

- What is considered "raw" may vary from person to person

Definition of data

> "Data are values of qualitative or quantitative variables, belonging to a set of items."

http://en.wikipedia.org/wiki/Data

Set of items: Sometimes called the population; the set of objects you are interested in

Variables: A measurement or characteristic of an item.

Qualitative: Country of origin, sex, treatment

Quantitative: Height, weight, blood pressure

- Qual/quant are ways of measuring variables

Raw versus processed data

Raw data

- The original source of the data
- Often hard to use for data analyses
- Data analysis *includes* processing /cleaning
- Raw data may only need to be processed once

http://en.wikipedia.org/wiki/Raw_data

Processed data

- Data that is ready for analysis
- Processing can include merging, subsetting, transforming, etc. ⌐ extract image, text, etc.
- There may be standards for processing
- All steps should be recorded

http://en.wikipedia.org/wiki/Computer_data_processing

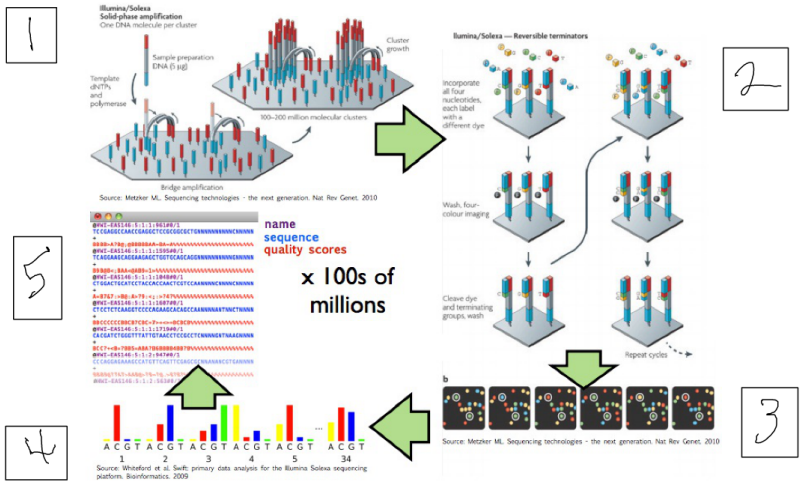
└ Hard to use
  └ Complicated, hard to parse, difficult to analyze, etc.
└ Need to keep a record of steps used in data processing
  └ Impacts downstream analyses
  └ VERY IMPORTANT

An example of a processing pipeline

Illumina Hi-Seq DNA Sequencer

http://www.illumina.com.cn/support/sequencing/sequencing_instruments/hiseq_1000.asp

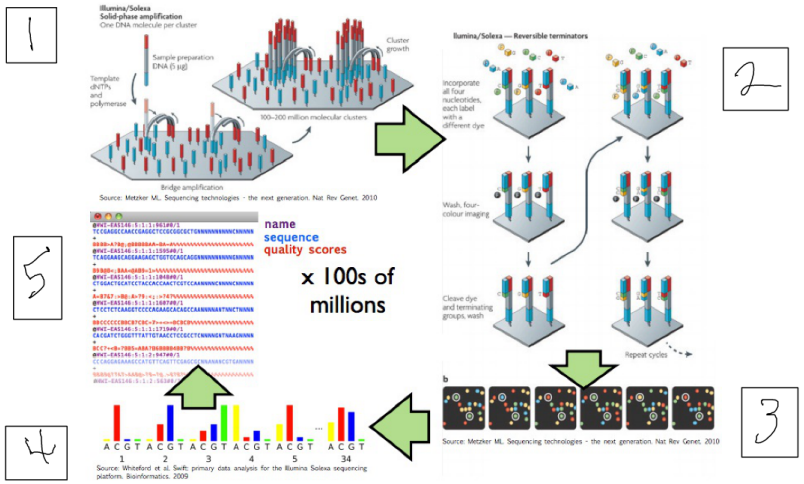
An example of a processing pipeline



http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/seqIntro.pdf

- Rough Illumina Hi-Seq processing pipeline
  1. Start w/ DNA frags bound to slide (~500 bp)
     - Make copies
  2. Sequencing by synthesis
     - Different color for each nucleotide
  3. Create a series of images
     - Images are sequential w/ each "dot" being one strand of DNA

An example of a processing pipeline

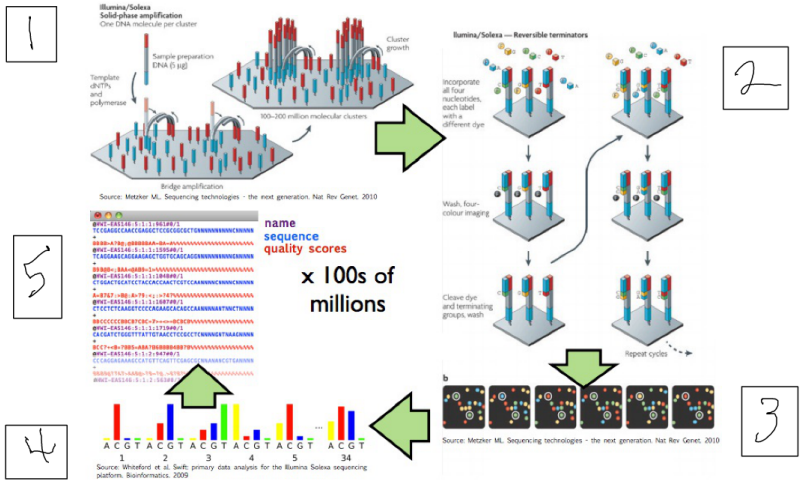


http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/seqIntro.pdf

(4) Use consensus summation in each image to assign nucleotide identity

(5) FASTQ file of sequence

An example of a processing pipeline



http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/seqIntro.pdf

- What can be considered the "raw" data varies
  - i.e. images, which must be processed into profiles, profiles which must be processed into sequences, or sequence which can be analyzed to learn more about the seq.
- Need to be aware of all processing steps b/c they can have huge impact on down stream analysis