Raw and processed data

Jeffrey Leek, Assistant Professor of Biostatistics

May 17, 2016

Definition of data

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

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

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

Definition of data

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

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

Variables: A measurement or characteristic of an item.

Definition of data

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

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

Qualitative: Country of origin, sex, treatment

Quantitative: Height, weight, blood pressure

Raw versus processed data

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

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

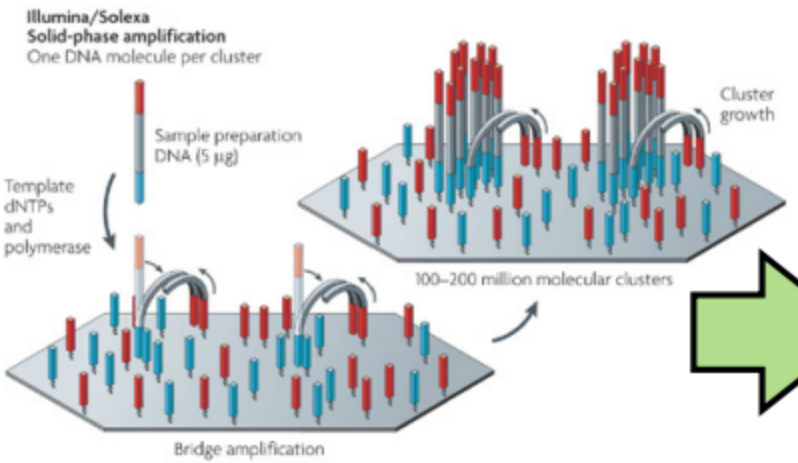
Processed data * Data that is ready for analysis * Processing can include merging, subsetting, transforming, etc. * There may be standards for processing * All steps should be recorded

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

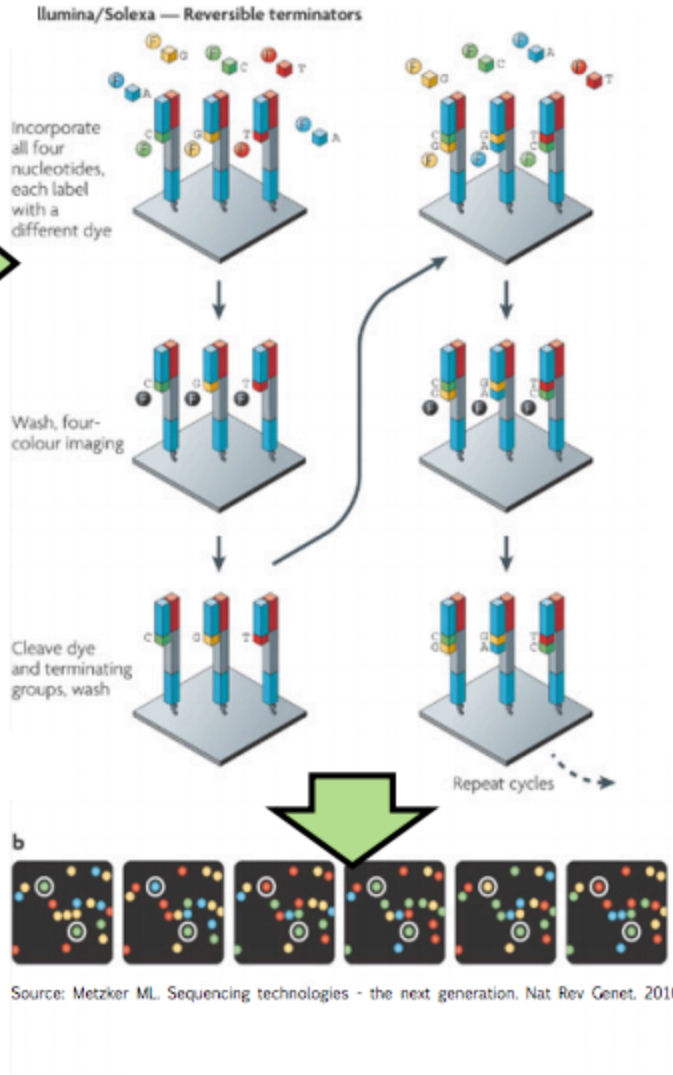
An example of a processing pipeline

http://www.illumina.com.cn/support/sequencing/

An example of a processing pipeline



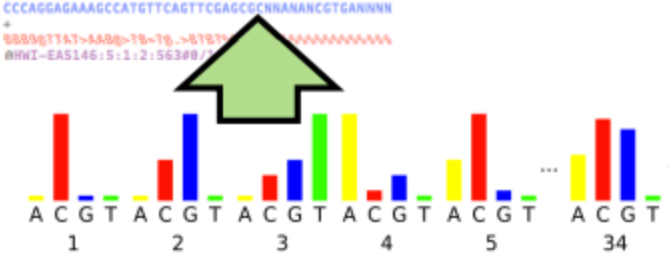
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

name
sequence
quality scores

x 100s of millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009

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