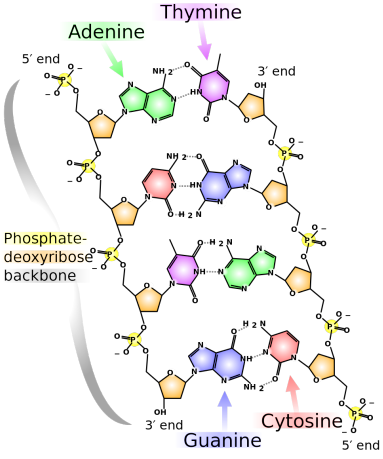
Introduction to High-Throughput Sequencing and RNA-seq

L Collado-Torres

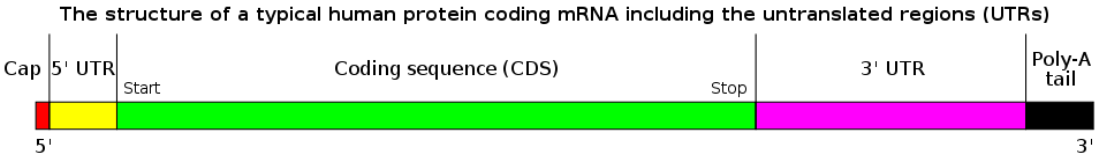
March 6th, 2013

DNA[1]



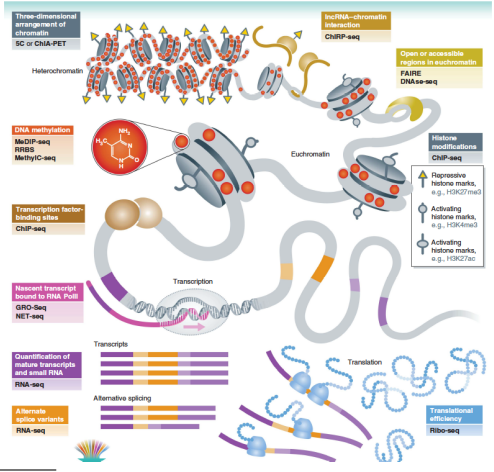
[1] Wikipedia. *DNA*. URL: http://en.wikipedia.org/wiki/DNA (visited on 03/05/2013).

Human mRNA[2]



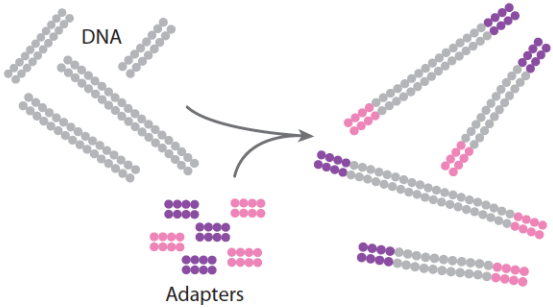
[2] Wikipedia. *Messanger RNA*. URL: http://en.wikipedia.org/wiki/Messenger_RNA (visited on 03/05/2013).

Panorama[3]



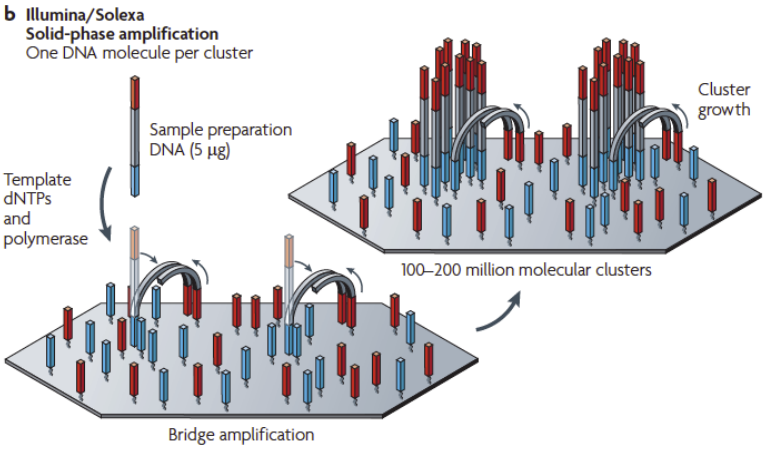
[3] Wendy Weijia Soon, Manoj Hariharan, and Michael P. Snyder. "High-throughput sequencing for biology and medicine". In: *Molecular Systems Biology* 9.1 (). URL: http://www.nature.com/msb/journal/v9/n1/fig_tab/msb201261_F2.html (visited on 03/05/2013).

Prepare DNA[4]



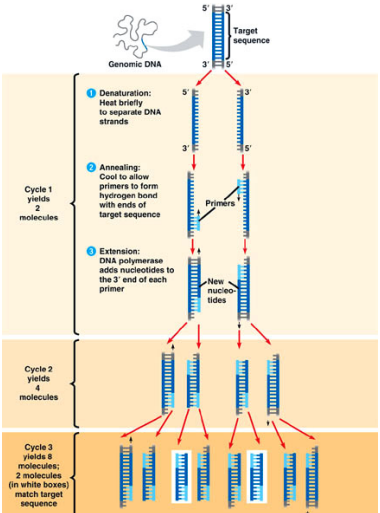
[4] Elaine R Mardis. "Next-generation DNA sequencing methods". In: *Annual Review of Genomics and Human Genetics* 9 (2008). PMID: 18576944.

Amplify[5]



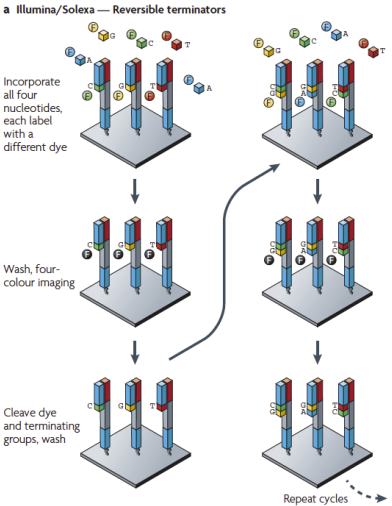
[5] Michael L. Metzker. "Sequencing technologies — the next generation". In: *Nat Rev Genet* 11.1 (2010).

PCR[6]



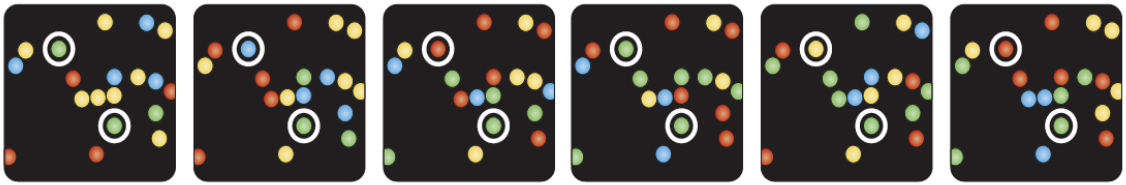
[6] SCHOOLWORKHELPER. *PCR: Uses, Steps, Purpose*. URL: http://schoolworkhelper.net/pcr-uses-steps-purpose/ (visited on 03/05/2013).

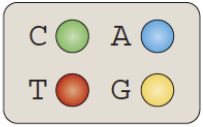
Sequencing by synthesis[7]



[7] Michael L. Metzker. "Sequencing technologies — the next generation". In: *Nat Rev Genet* 11.1 (2010).

Analyze cluster images[8]



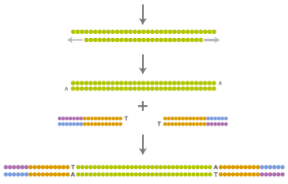


Top: CATCGT
Bottom: CCCCCC

[8] Michael L. Metzker. "Sequencing technologies — the next generation". In: *Nat Rev Genet* 11.1 (2010).

HiSeq 2000[9]

Figure 3: Next-Generation Sequencing Simplified



Library Preparation
<6 h (<3 h hands-on)

Cluster Generation
<4 h (<10 min hands-on)

Sequencing by Synthesis
1.5-8 days (<10 min hands-on)

RTA v1.7, CASAVA v1.7
2 days (30 min hands-on)

From simplified sample preparation kits, to automated cluster generation, to streamlined sequencing by synthesis, to complete data analysis, Illumina's HiSeq 2000 sequencing system offers the industry's simplest next-generation sequencing workflow.

[9] Illumina. *HiSeq 2000 Sequencing System*. URL: http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf (visited on 03/05/20

HiSeq 2000

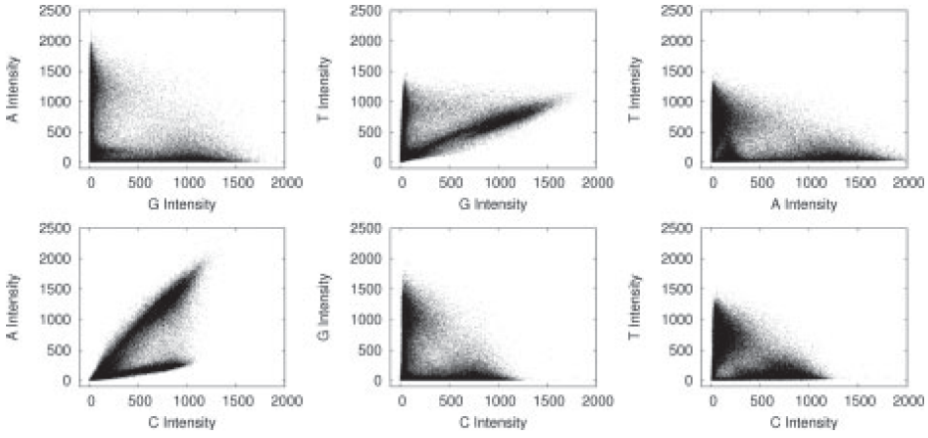
More info on this blog post `http://www.politigenomics.com/2010/01/hiseq-2000.html`

Other 2nd generation sequencers[10]

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

[10] Lin Liu et al. "Comparison of next-generation sequencing systems". In: *Journal of biomedicine & biotechnology* (2012). PMID: 22829749.

Cross-talk[11]



[11] Nava Whiteford et al. "Swift: primary data analysis for the Illumina Solexa sequencing platform". In: *Bioinformatics (Oxford, England)* 25.17 (2009). PMID: 19549

Phasing and pre-phasing[12]

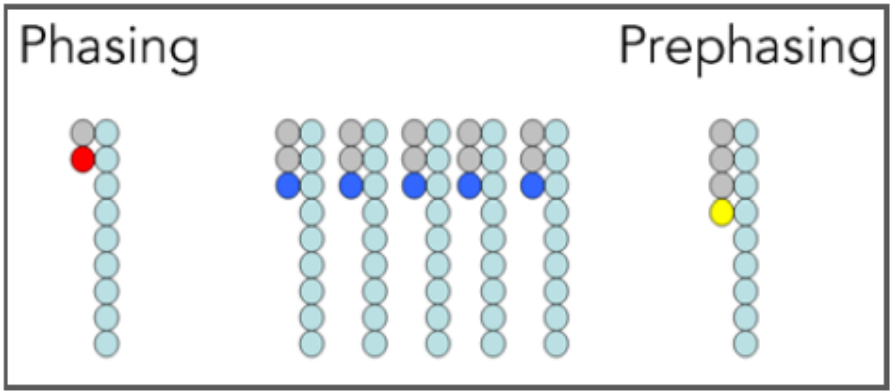
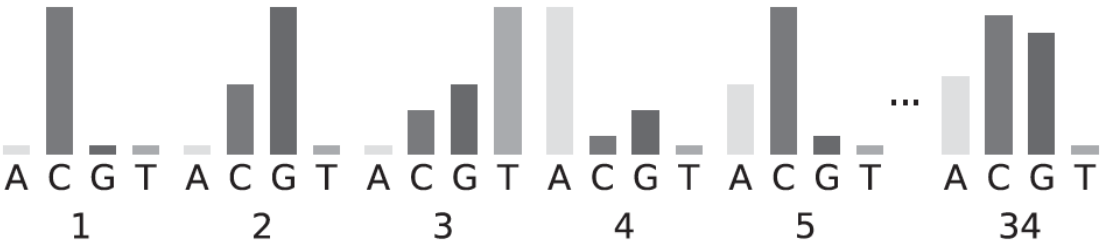


Figure 3 Phasing and Prephasing

[12] Illumina. *Pipeline CASAVA User Guide 15003807 (Pipeline V. 1.4 and Casava V.1.0).*

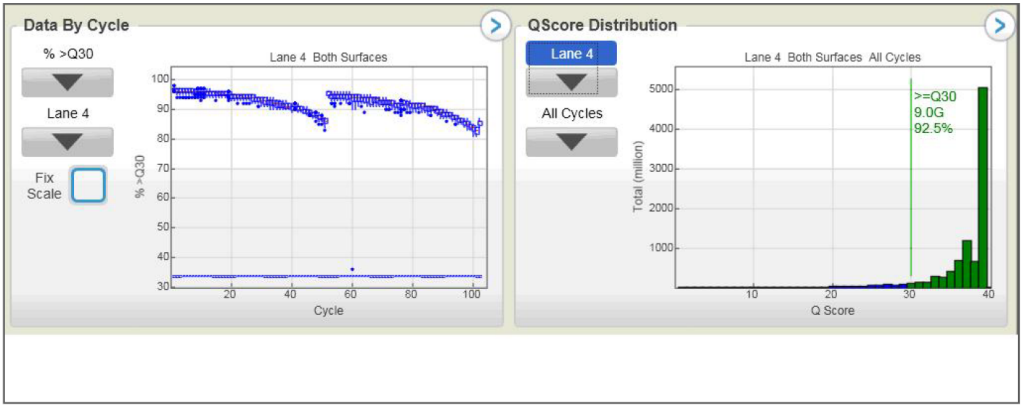
Phasing example[13]



[13] Nava Whiteford et al. "Swift: primary data analysis for the Illumina Solexa sequencing platform". In: *Bioinformatics (Oxford, England)* 25.17 (2009). PMID: 19549

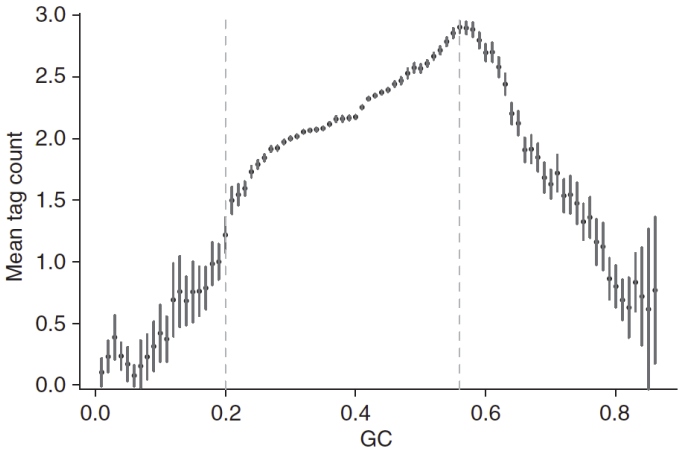
Sequence quality[14]

Figure 3 SAV Screenshot Showing Excellent Quality Metrics



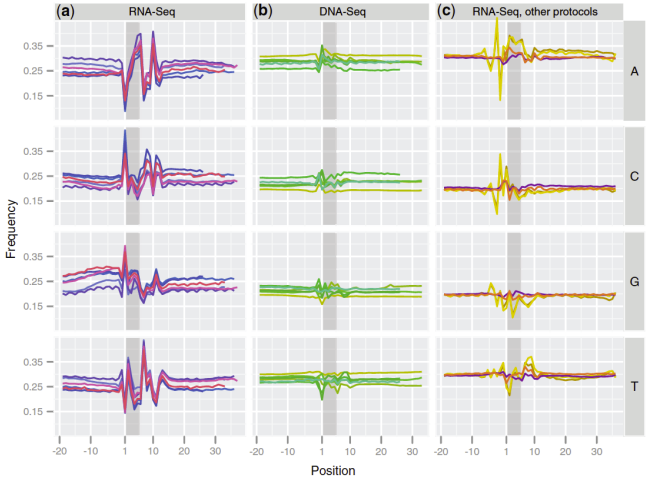
[14] Illumina. *CASAVA User Guide (15011196 D)*. URL: http://support.illumina.com/downloads/casava_user_guide_15011196.ilmn (visited on 03/05/2013).

GC bias[15]



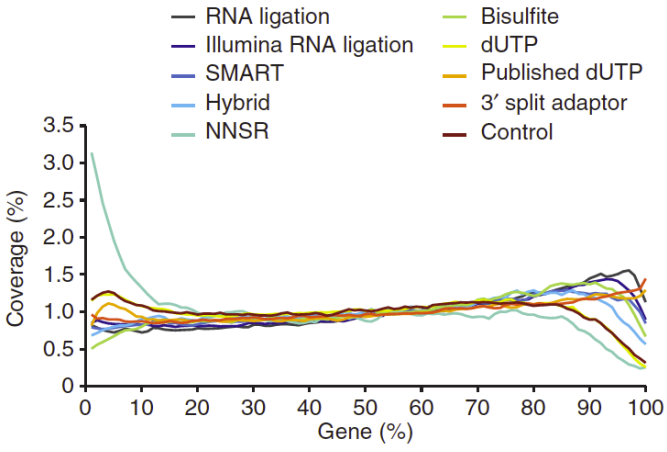
[15] Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. "Overcoming bias and systematic errors in next generation sequencing data". In: *Genome Medici* 2.12 (2010). PMID: 21144010.

Random primers bias[16]



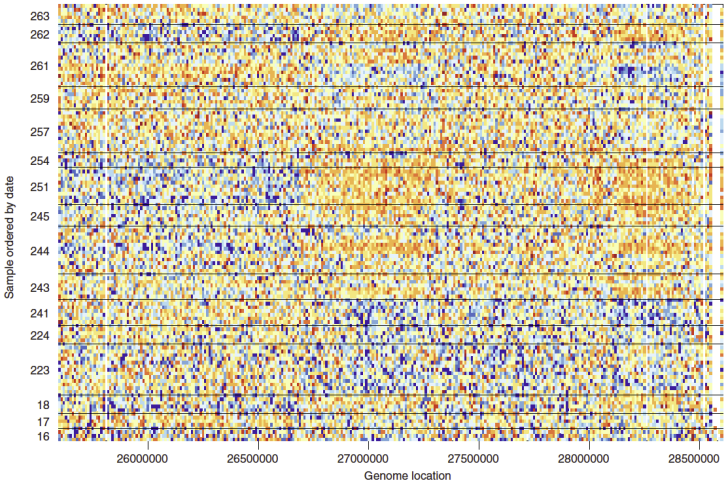
[16] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. "Biases in Illumina transcriptome sequencing caused by random hexamer priming". In: *Nucleic Acids Research* (2010). PMID: 20395217.

Library type[17]



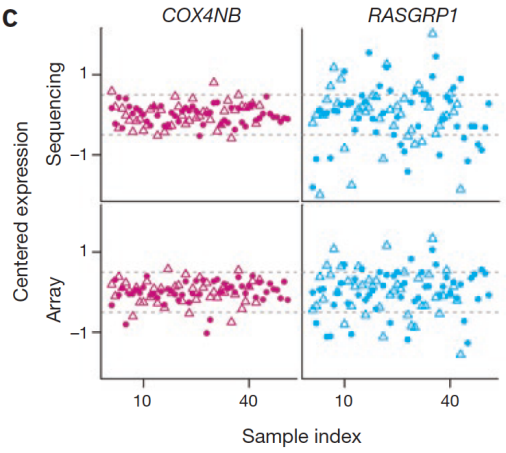
[17] Joshua Z Levin et al. "Comprehensive comparative analysis of strand-specific RNA sequencing methods". In: *Nature Methods* 7.9 (2010). PMID: 20711195.

Batch effects[18]



[18] Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. "Overcoming bias and systematic errors in next generation sequencing data". In: *Genome Medici* 2.12 (2010). PMID: 21144010.

Biological variability[19]



[19] Kasper D Hansen et al. "Sequencing technology does not eliminate biological variability". In: *Nature biotechnology* 29.7 (2011). PMID: 21747377.

The future

- Further improvements in library preparation
- Single cell sequencing
- Third generation sequencers like Pacific Biosciences

And biostatistical methods =)

Thanks!

- Google Calendar
  https://www.google.com/calendar/embed?src=7hprep991i5prd5l5ftksbsfb8%40group.calendar.google.com&ctz=America/New_York
- Slides at http://www.biostat.jhsph.edu/~lcollado/misc/HTSintro.pdf