Challenges and strategies in the analysis and quantification of mRNA-Seq datasets.

School of Systems Biology Research Day

Parsa Hosseini
Bioinformatics/Computational Biology Ph.D.

Topics Covered

- RNA-Seq Overview
  - RNA-Seq vs. Affymetrix GeneChips
- RNA-Seq dataset organization
- Normalization, Differential Expression

Topics Covered

- RNA-Seq Overview
  - RNA-Seq vs. Affymetrix GeneChips
- RNA-Seq dataset organization
- Normalization, Differential Expression

RNA-Seq Overview

- New transcriptomics tool for rapid parallel sequencing, referred to as ‘high throughout sequencing’.
- RNA-Seq is part of an umbrella of technologies known as ‘next-generation sequencing’
  - Faster sequencing than Sanger
  - Safer than Maxam-Gilbert

RNA-Seq Overview

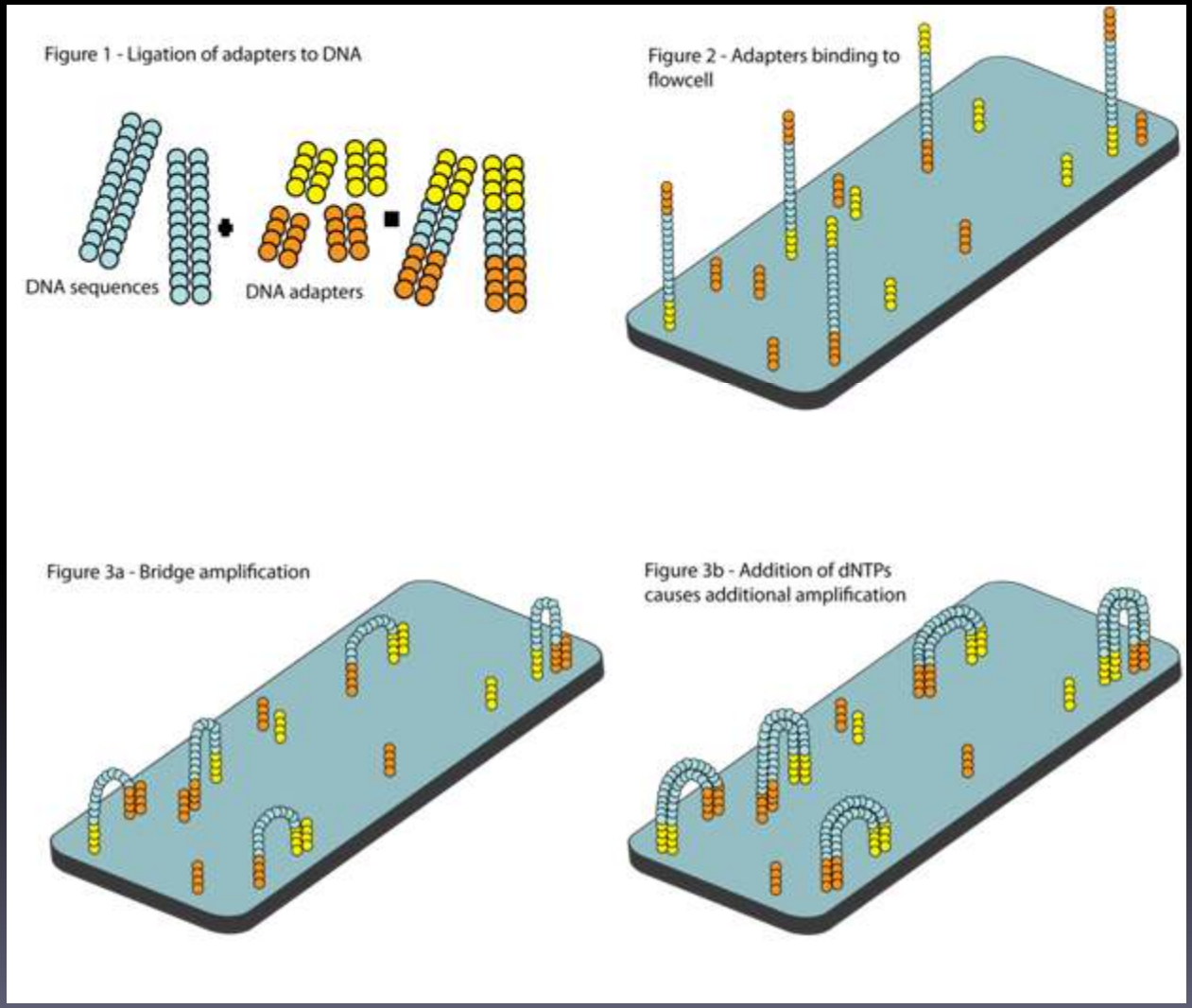
Technology	Read (bp; approx)	$ / Mb	Error %	Throughput (Mb/day)
illumina GAIIx	75 - 100	2	0.1 - 1	~400
Roche 454 Titanium	340	55	4 - 5	~300
SOLiD v3	50	1	< 1	~600
Helicos HeliScope	25 - 50	1	< 1	>1,100
Polonator	13 - 26	1	< 1	4,000 – 5,000

- Potential Uses:
  - Re-sequencing, de-novo sequencing.
  - Alternative splice analysis, SNPs.
  - Transcriptome profiling.
  - Multiplexed sequencing.
  - miRNA, gene discovery.
  - Drive treatment of cancer patients (Sci. Trans. Med, 2011)

RNA-Seq vs. Affymetrix GeneChips

- RNA-Seq
  - Low amount of RNA required.
  - Quantifying less-abundant transcripts.
  - Throughput orders or magnitude higher.
  - Low background noise.
  - Single-base pair resolution.
  - Alternative splicing / novel isoforms.
- Though cost is expensive, costs are dropping rapidly; differential against microarrays will be minimal.

Illumina sequencing – 1/3



Illumina sequencing – 2/3

Figure 4 - Denaturation of DNA bridges

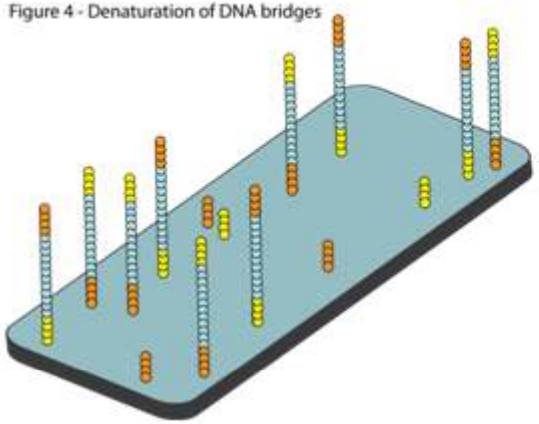


Figure 5 - Repeating steps in Figure 3a - 4. Resultant are generated clusters

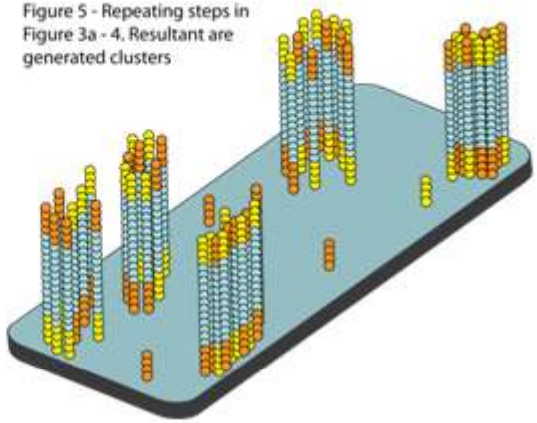


Figure 6a - Addition of reversible terminators binding to the various clusters

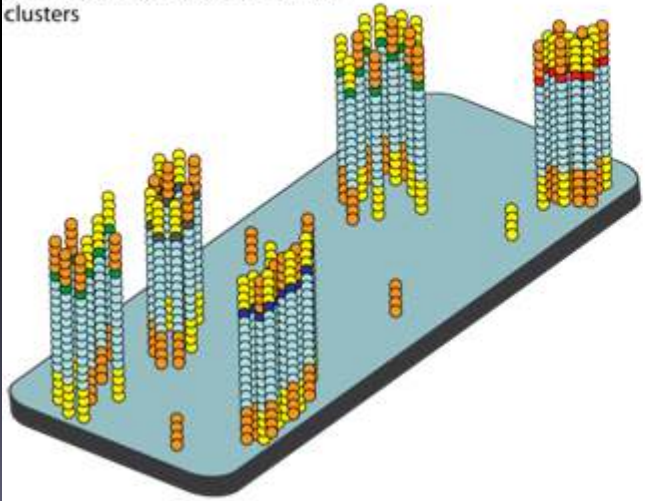
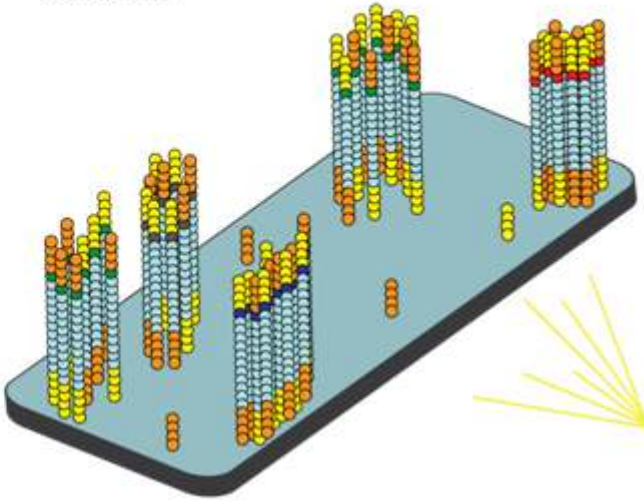
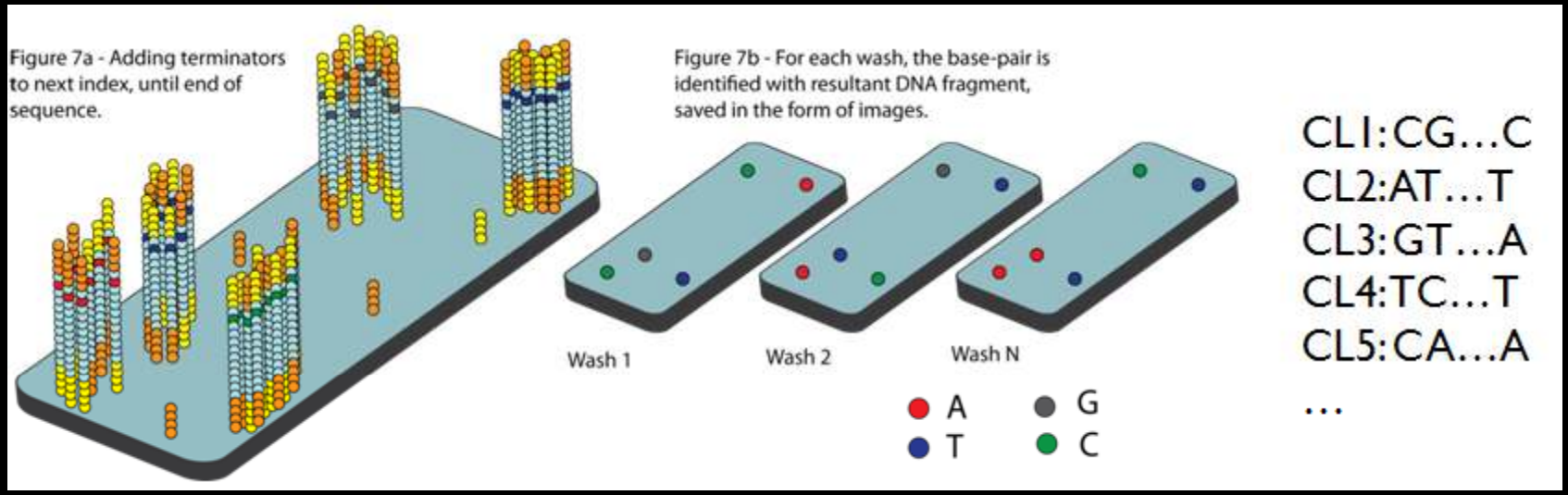


Figure 6b - Laser excitation and base identification



Illumina sequencing – 3/3



- Multiplex sequencing (upto 96 samples / flowcell)
  - 7bp tag identifies each sample.

Topics Covered

- RNA-Seq
  - RNA-Seq vs. Affymetrix GeneChips
- RNA-Seq dataset organization
- Normalization, Differential Expression

RNA-Seq dataset organization

- From 2010 on, TIFF replaced with thumbnails.
  - Flow-cell = 100 – 120 tiles.
  - After laser-emission, an thumbnail is captured for each nucleotide for each tile (hence 4x per tile; 0.04 MB each).
  - => 7 (#/lanes) * 80 (#/cycles) * 4 (bases) * 0.04 (#/MB per image) * 120 (#/tiles)
  - **~ 11 GB (storage drops 2x order-of-magnitude)**
- But as costs drop, throughput increases, longer reads.

RNA-Seq dataset organization

- 2x popular NGS repositories:
  1. Sequence Read Archive (SRA)
  2. European Nucleotide Archive (ENA)

- February 2011: SRA set for closure
  - October 2011 **(1.34 PB)**
  - September 2010 **(100 TB)**
  - Budgetary constraints, storage/management issues
    - A centralized DB for all terabyte-sized NGS datasets?

RNA-Seq dataset organization

- Reads are represented with 'quality (Q) scores'

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Q-scores determine P(base=incorrectly called)
  - Fastq quality ASCII -> integer determines Q-score

Q-value	% incorrect	% correct
1	79.433	20.567
10	10	90
15	3.162	96.838
20	1	99
25	0.316	99.684
30	0.1	99.9
35	0.032	99.968
40	0.01	99.99
45	0.003	99.997

$$Q = -10\log(P) \quad (1)$$

$$\frac{Q}{-10} = \log(P) \quad (2)$$

$$P = 10^{\frac{-Q}{10}} \quad (3)$$

Topics Covered

- RNA-Seq Overview
  - RNA-Seq vs. Affymetrix GeneChips
- RNA-Seq dataset organization
- Normalization, Differential Expression

Normalization, Differential Expression

- Early normalization = Divide read-counts by #/ sequence reads. Drawback = Two datasets may have differing depth, coverage.
- Popular normalization approaches:
  - RPKM (reads per kilobase per million of mapped reads); (Mortazavi, 2008)
  - TMM (trimmed mean of M scores); (Oshlack, 2010)
  - Log-transform read counts and treat like microarray dataset (t-test, etc)

Normalization, Differential Expression

- Utilizing statistical testing and models will help infer differential expressed genes (DEGs).
- Poisson distribution can be used to find DEGs
  - (Marioni, 2008), (Wang, 2010)
- Caveat = Poisson has been shown to predict smaller variance from read-count datasets.
  - Result = extra variance from replicates is often goes unappreciated, hence 'over-dispersion' (more variance in dataset than from model).
  - Conclusion = not harnessing variance makes it hard to find false-positive DEGs.
- Modeling count data using Negative Binomial Distribution (NB) can address the over-dispersion.
  - (Robinson, 2010), (Anders, 2010); tools such as edgeR, DESeq.

Normalization, Differential Expression