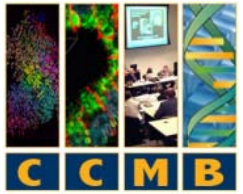
National Center for Integrative Biomedical Informatics



RNA-Seq and Differentially Expressed Gene Analysis Pipeline

NCIBI/RCMI –Workshop on Translational Bioinformatics

U-M, Ann Arbor, Michigan

July 29-30, 2010

RNA-Seq Workshop

Jim Cavalcoli, Yongsheng Bai,
Xiao-Wei Chen, Rich McEachin

Roadmap

- Next Generation Sequencing (Jim)
  - Methods and platforms
- Differential Gene Expression (Xiao-Wei)
  - SEC-24a
- RNA-Seq (Rich)
- Analysis pipeline (Yongsheng)
  - Candidate Gene Selection (hands on)
  - ConceptGen (hands on)
  - Biological relevance

Next-Generation Sequencing

Overview of methods and platforms available

Outline

- Technology Description and examination of different platforms
- Biological Applications of the Methods
- Informatics applications

Next-Generation Sequencing

- Should be called “Now-generation”…. It’s Here!

- Enhanced sequencing capabilities
  - Increased throughput (huge increase in # of reads, decrease in time to produce)
  - Decreased costs per base
  - Ability to sequence from individual samples

Platform Information

Company	Roche	Applied Biosystems	Illumina	Dover Systems	Helicos	Pacific Biosciences
Platform	FLX	SOLiD 3	GA II, HiSeq2000	Polonator	Heliscope	-
Method of Sequencing	Emulsion PCR on beads	Emulsion PCR on beads	Bridge PCR Amplification	Emulsion PCR on beads	Single molecule sequencing	Single molecule SMRT
Chemistry	Pyrosequencing with polymerase	Ligation (dual-base encoding)	Reversible terminator with polymerase	Ligation (single base encoding)	Asynchronous extension using polymerase	Single molecule SMRT
Machine cost	~500K	~600K	~600K	~170K	~999K	~600K (anticipated)
Reagents per Run (cost per MB)	5K ($60)	3K ($2)	4-6K ($2)	1K ($1)	18K ($2)	$ 99 (?)
Capacity per Run	0.5 GB	60+ GB	25-35K (much more with HiSeq)	4-10 GB	28 GB	72 Mb (predicted)
Read Length	400-600 nt	50 nt	36-100 nt	13 nt	25-50 nt	580-2,800 nt
Input DNA required	500 ng-1ug	10 ng (200-10kb input fragment)	0.1-10ug (200-500 bp fragment)	-	As low as 50 pg	-
Advantages	Long read length	Accuracy of dual base calls, high output	No emulsion PCR	Open source model	No clonal amplification	No clonal amplification
Disadvantages	Unreliable for homopolymer regions	Short read length	Medium read length, long run times	Short read length	Cost of machine	?
Primary Error Type	Indel	substitution	substitution	substitution	deletion	

Table created jointly from: Lerner and Fleisher, 2010. The Auk 127(1):4-15; and Shendon and Ji, 2008. Nat. Biotech. 26(10):1135-1145.

More about major platforms

- Illumina Solexa
  - http://www.youtube.com/watch?v=77r5p8IBwJk
- Roche 454
  - http://www.youtube.com/watch?v=bFNjxKHP8Jc
- ABI SOLiD
  - http://www.youtube.com/watch?v=nlvyF8bFDwM

- Choice of technology is dependent on the experimental method and hypothesis

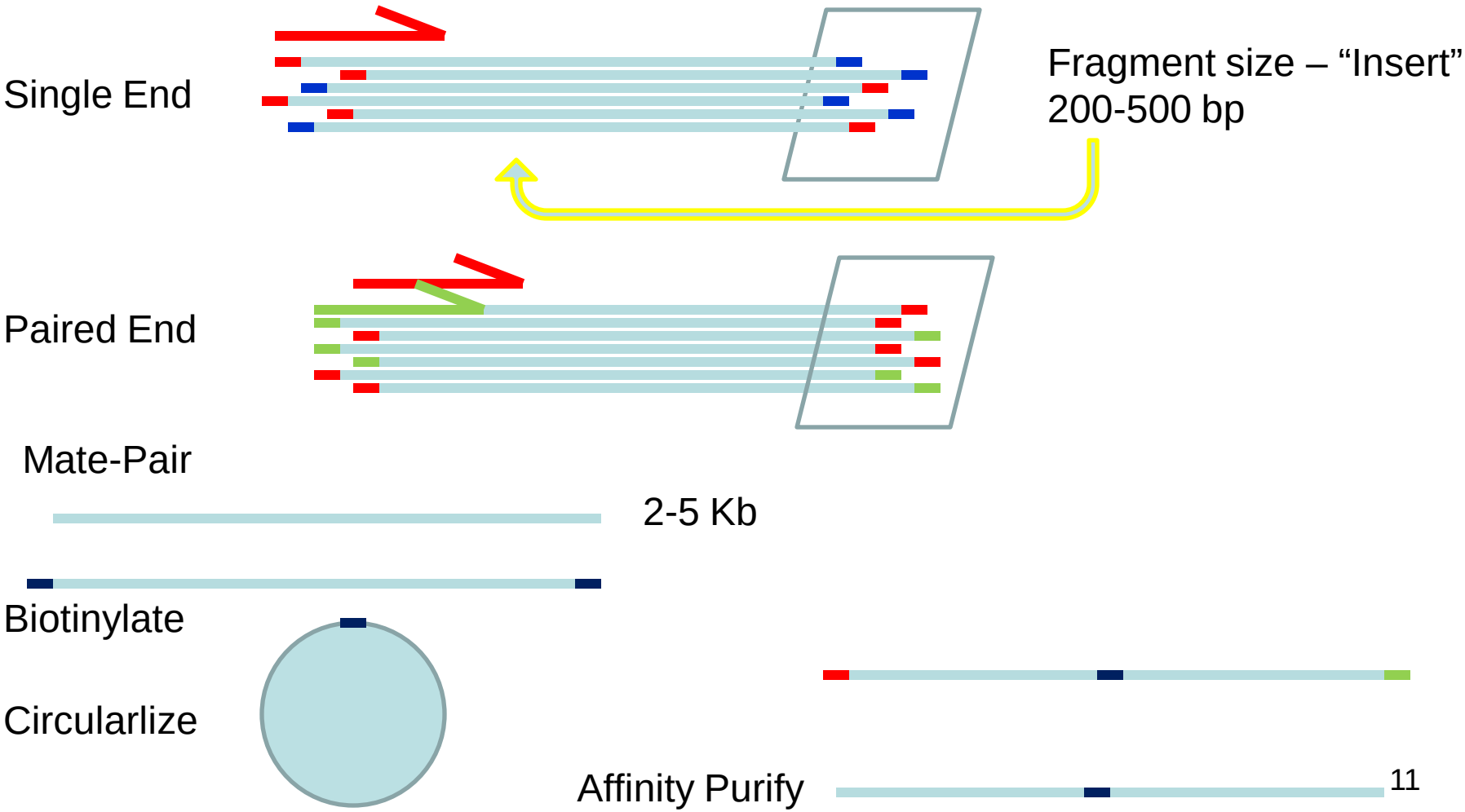
Types of Applications to Biology

- Genomic Resequencing
  - Sequencing select genomic regions, and comparing to a reference genome
- De novo assembly of novel genomes
  - Needs lots of depth of coverage
  - Works best for small (bacterial) genomes
  - Paired ends and different size libraries
- RNA Expression (RNA-Seq)
  - Expressed genes and level of expression – more detail to follow
- Protein binding to DNA (ChIP-Seq)
  - Immunoprecipitation of Protein bound to DNA (chromatin)
- Primer-specific sequencing (16S RNA)
  - Identifies communities of 16S RNA in microbe / samples
- Metagenomic sequencing
  - shotgun sequencing from a community of DNA
- Methylation of DNA (Bisulfite sequencing)

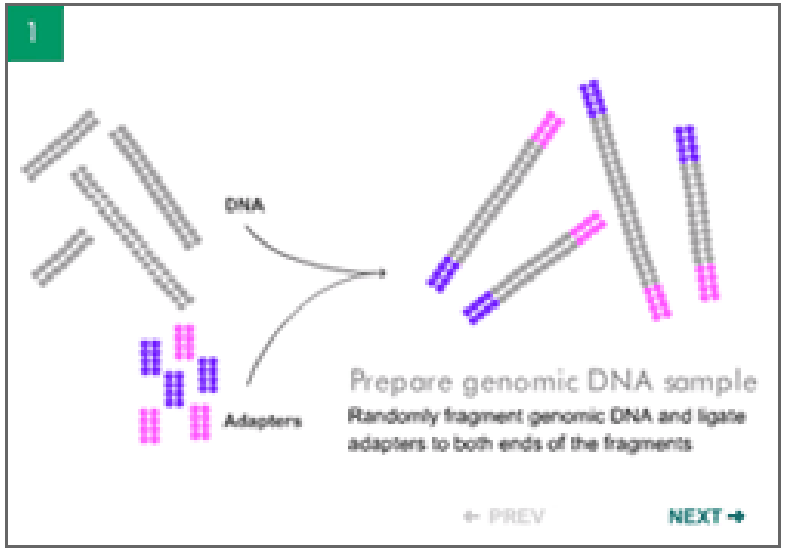
Library type

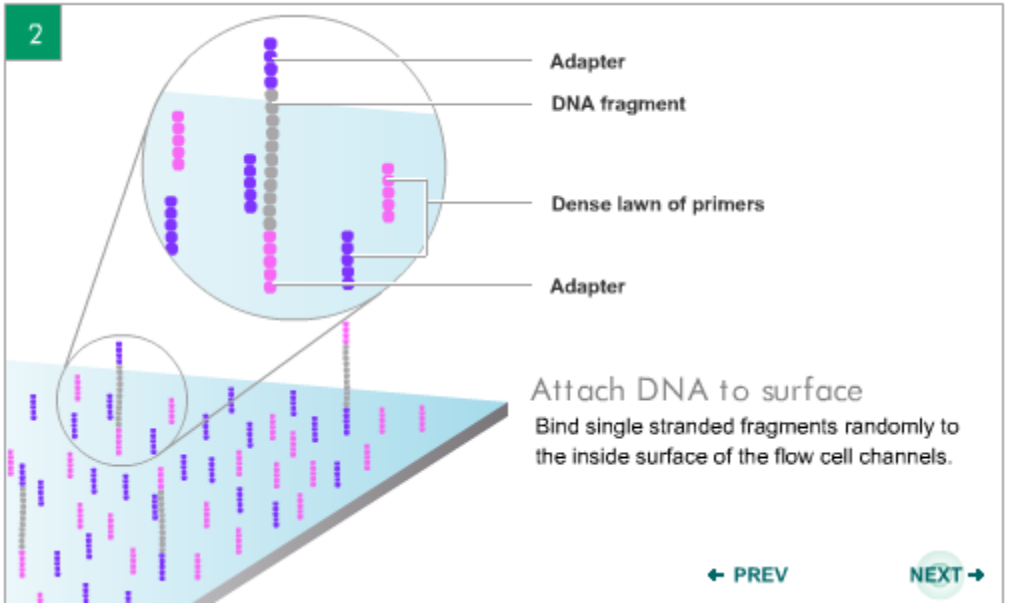
- Three major types of sequencing can be planned for:
  - Single End reads
  - Paired End reads
  - Mate Pair reads

Library Methods



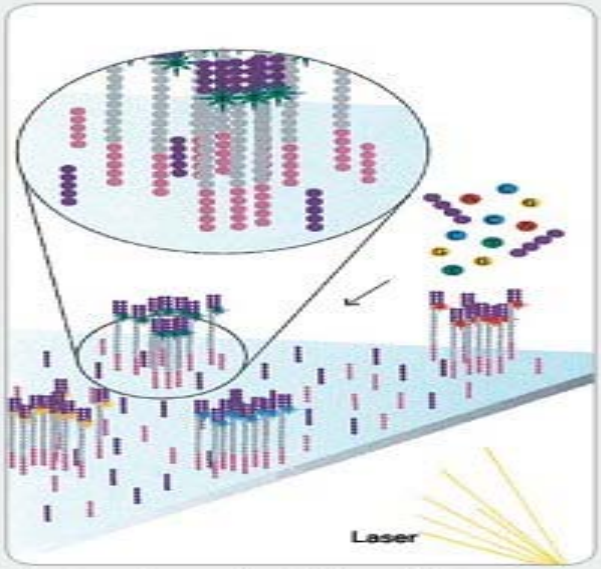
Illumina: Solexa





http://www.illumina.com/pages.ilmn?ID=203

7. DETERMINE FIRST BASE

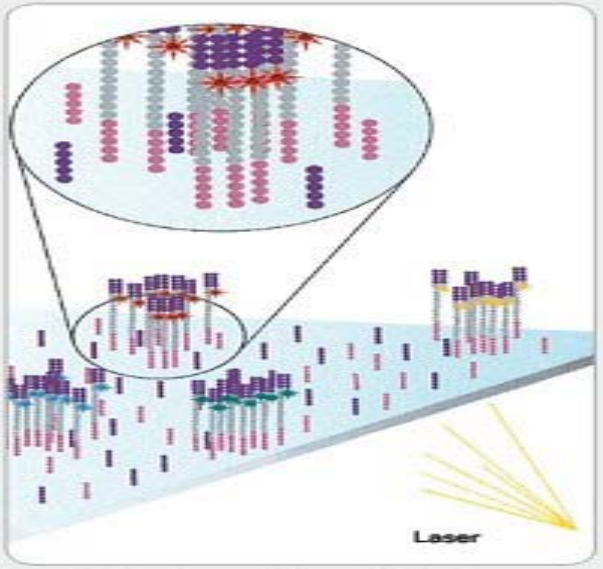


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

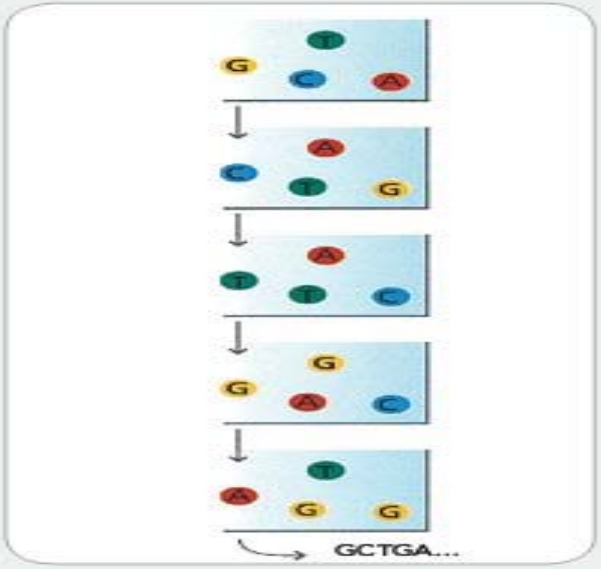


Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE

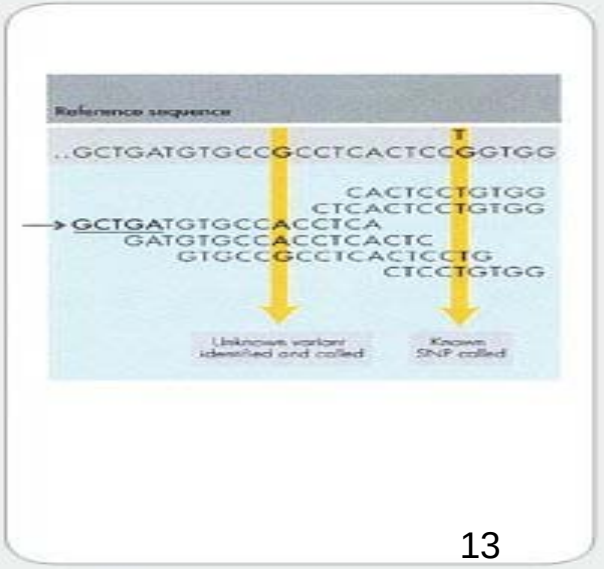
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



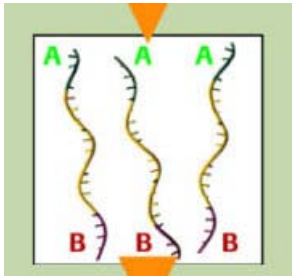
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



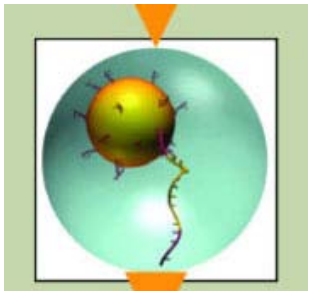
Align data, compare to a reference, and identify sequence differences.

454 Process



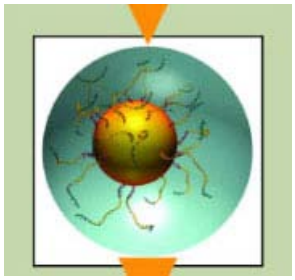
Library Preparation

Using a series of standard molecular biology techniques, short adaptors (A and B) - specific for both the 3' and 5' ends - are added to each fragment. The adaptors are used for purification, amplification, and sequencing steps. Single-stranded fragments with A and B adaptors compose the sample library used for subsequent workflow steps.



One Fragment = One Bead

The single-stranded DNA library is immobilized onto specifically designed DNA Capture Beads. Each bead carries a unique single-stranded DNA library fragment. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture resulting in microreactors containing just one bead with one unique sample-library fragment.

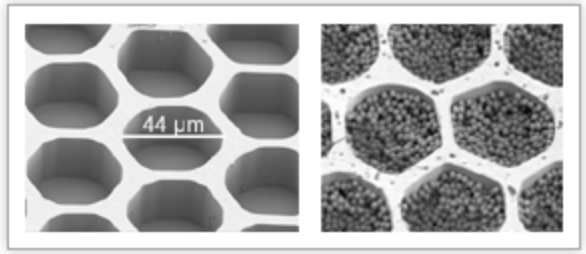


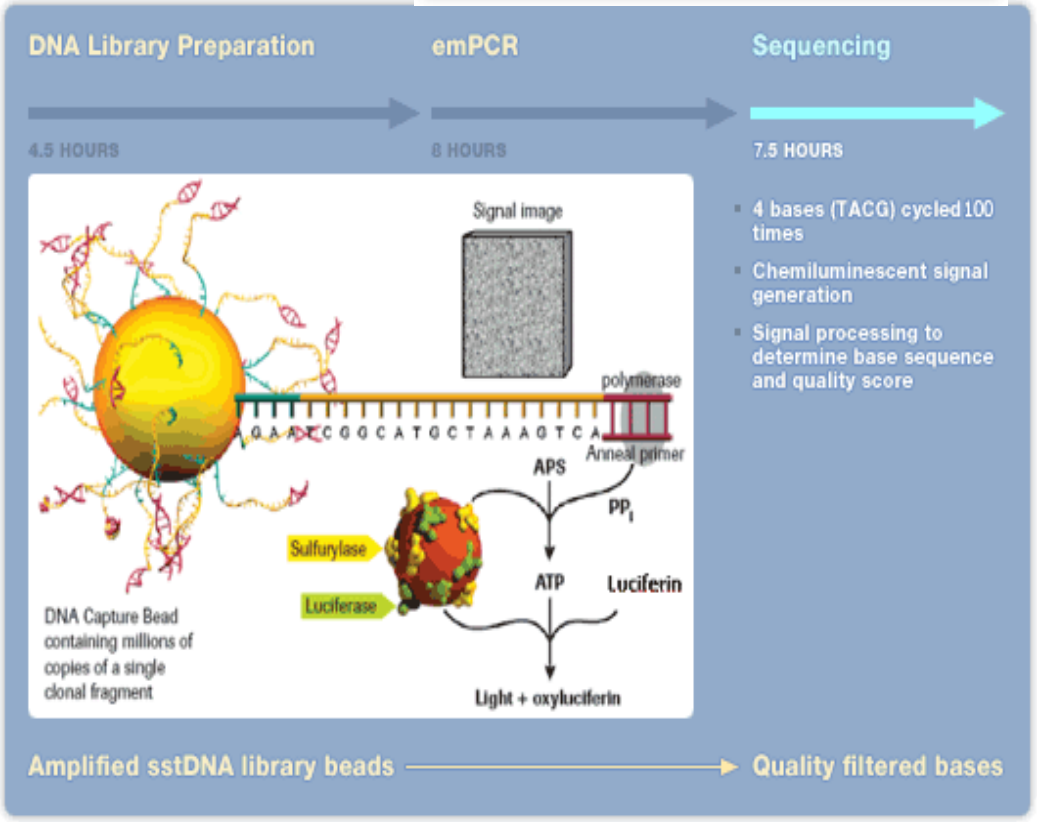
emPCR (Emulsion PCR) Amplification

Each unique sample library fragment is amplified within its own microreactor, excluding competing or contaminating sequences. Amplification of the entire fragment collection is done in parallel; for each fragment, this results in a copy number of several million per bead. Subsequently, the emulsion PCR is broken while the amplified fragments remain bound to their specific beads.

454 (Roche)

- Beads with millions of copies of DNA are sequenced in parallel.
- Polymerase extends the existing DNA strand by adding nucleotide(s). If a nucleotide complementary to the template strand is flowed into a well,
- The Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal that is recorded by the CCD camera.
- The signal strength is proportional to the number of nucleotides, for example, homopolymer stretches, incorporated in a single nucleotide flow

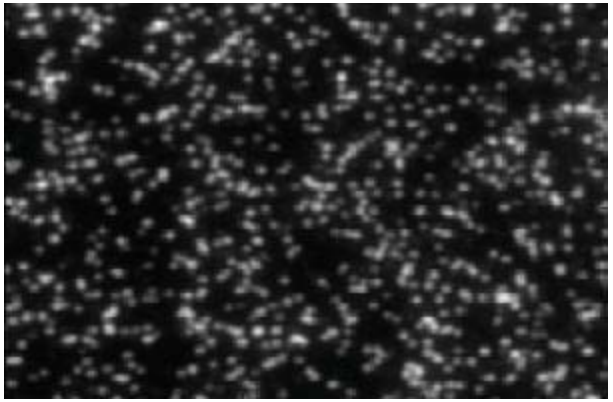




Output from GenomeAnalyzer II (Illumina – Solexa)

- Read length 36, 75, 100 nt
  - 150 nt and more, soon
  - Single Read, Paired-End reads as well as Mate-Pair reads
- 8 lanes per flow cell, 15-20 million reads per lane
- ~ 30 Gbases per flow cell
  - PE, 100 nt
- Accuracy is ~99 - 99.5%
  - Primary type of error: Substitution
  - 150 million errors per flow cell!

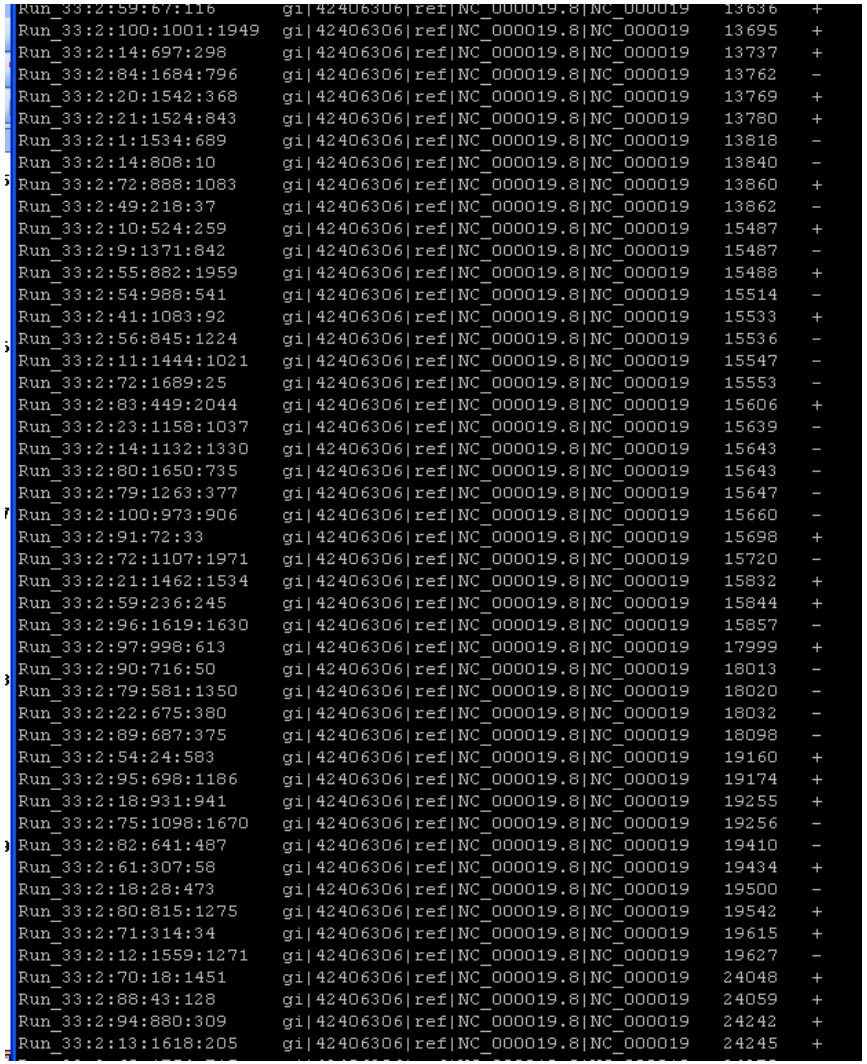
Data output and processing



- Image data output (tiff files)
  - 100 tiles per lane, 8 lanes per flow cell, 100 cycles.
  - 4 images (A,G,C,T) per tile per cycle = 320,000 images
  - Each tiff image is ~ 7 MB = 2,240,000 MB of data (2.24 TB !)
  - 4.5 TB for 100 nt Paired-end read

- Illumina Pipeline:
  - Firecrest (image analysis)
    - Locates clusters and calculates intensity and noise
  - Bustard (base calling)
    - Deconvolutes signal and corrects for cross-talk, phasing
  - GERALD – generation of recursive analyses linked by dependency
    - ELAND – (Efficient large-scale alignment of nucleotide databases)

Sequence text output



Run_33:2:100:1001:1949	gi|42406306|ref|NC_000019.8|NC_000019	13695	+
Run_33:2:14:697:298	gi|42406306|ref|NC_000019.8|NC_000019	13737	+
Run_33:2:84:1684:796	gi|42406306|ref|NC_000019.8|NC_000019	13762	-
Run_33:2:20:1542:368	gi|42406306|ref|NC_000019.8|NC_000019	13769	+
Run_33:2:21:1524:843	gi|42406306|ref|NC_000019.8|NC_000019	13780	+
Run_33:2:1:1534:689	gi|42406306|ref|NC_000019.8|NC_000019	13818	-
Run_33:2:14:808:10	gi|42406306|ref|NC_000019.8|NC_000019	13840	-
Run_33:2:72:888:1083	gi|42406306|ref|NC_000019.8|NC_000019	13860	+
Run_33:2:49:218:37	gi|42406306|ref|NC_000019.8|NC_000019	13862	-
Run_33:2:10:524:259	gi|42406306|ref|NC_000019.8|NC_000019	15487	+
Run_33:2:9:1371:842	gi|42406306|ref|NC_000019.8|NC_000019	15487	-
Run_33:2:55:882:1959	gi|42406306|ref|NC_000019.8|NC_000019	15488	+
Run_33:2:54:988:541	gi|42406306|ref|NC_000019.8|NC_000019	15514	-
Run_33:2:41:1083:92	gi|42406306|ref|NC_000019.8|NC_000019	15533	+
Run_33:2:56:845:1224	gi|42406306|ref|NC_000019.8|NC_000019	15536	-
Run_33:2:11:1444:1021	gi|42406306|ref|NC_000019.8|NC_000019	15547	-
Run_33:2:72:1689:25	gi|42406306|ref|NC_000019.8|NC_000019	15553	-
Run_33:2:83:449:2044	gi|42406306|ref|NC_000019.8|NC_000019	15606	+
Run_33:2:23:1158:1037	gi|42406306|ref|NC_000019.8|NC_000019	15639	-
Run_33:2:14:1132:1330	gi|42406306|ref|NC_000019.8|NC_000019	15643	-
Run_33:2:80:1650:735	gi|42406306|ref|NC_000019.8|NC_000019	15643	-
Run_33:2:79:1263:377	gi|42406306|ref|NC_000019.8|NC_000019	15647	-
Run_33:2:100:973:906	gi|42406306|ref|NC_000019.8|NC_000019	15660	-
Run_33:2:91:72:33	gi|42406306|ref|NC_000019.8|NC_000019	15698	+
Run_33:2:72:1107:1971	gi|42406306|ref|NC_000019.8|NC_000019	15720	-
Run_33:2:21:1462:1534	gi|42406306|ref|NC_000019.8|NC_000019	15832	+
Run_33:2:59:236:245	gi|42406306|ref|NC_000019.8|NC_000019	15844	+
Run_33:2:96:1619:1630	gi|42406306|ref|NC_000019.8|NC_000019	15857	-
Run_33:2:97:998:613	gi|42406306|ref|NC_000019.8|NC_000019	17999	+
Run_33:2:90:716:50	gi|42406306|ref|NC_000019.8|NC_000019	18013	-
Run_33:2:79:581:1350	gi|42406306|ref|NC_000019.8|NC_000019	18020	-
Run_33:2:22:675:380	gi|42406306|ref|NC_000019.8|NC_000019	18032	-
Run_33:2:89:687:375	gi|42406306|ref|NC_000019.8|NC_000019	18098	-
Run_33:2:54:24:583	gi|42406306|ref|NC_000019.8|NC_000019	19160	+
Run_33:2:95:698:1186	gi|42406306|ref|NC_000019.8|NC_000019	19174	+
Run_33:2:18:931:941	gi|42406306|ref|NC_000019.8|NC_000019	19255	+
Run_33:2:75:1098:1670	gi|42406306|ref|NC_000019.8|NC_000019	19256	-
Run_33:2:82:641:487	gi|42406306|ref|NC_000019.8|NC_000019	19410	-
Run_33:2:61:307:58	gi|42406306|ref|NC_000019.8|NC_000019	19434	+
Run_33:2:18:28:473	gi|42406306|ref|NC_000019.8|NC_000019	19500	-
Run_33:2:80:815:1275	gi|42406306|ref|NC_000019.8|NC_000019	19542	+
Run_33:2:71:314:34	gi|42406306|ref|NC_000019.8|NC_000019	19615	+
Run_33:2:12:1559:1271	gi|42406306|ref|NC_000019.8|NC_000019	19627	-
Run_33:2:70:18:1451	gi|42406306|ref|NC_000019.8|NC_000019	24048	+
Run_33:2:88:43:128	gi|42406306|ref|NC_000019.8|NC_000019	24059	+
Run_33:2:94:880:309	gi|42406306|ref|NC_000019.8|NC_000019	24242	+
Run_33:2:13:1618:205	gi|42406306|ref|NC_000019.8|NC_000019	24245	+

Other software applications for assembly and alignment

Align/Assemble to a reference

* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hou workstation with 2 gigabytes of memory. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony Solexa 1G machine.
* EULER - Short read assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research).
* Exonerate - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Bi EMBL. C for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C,
* MOSAIK - Reference guided aligner/assembler. Written by Michael Strömberg at Boston College.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has prelimina handle ABI SOLiD data. Written by Heng Li from the Sanger Centre.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficier library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, A Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. P(required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available free for evaluatio use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* RMAP - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinforma OS required.
* SeqMap - Works like ELand, can do 3 or more bp mismatches and also INDELs. Written by Hui Jiang from the Wong lab at Stanford. Builds available for n
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Bru Stephen Rumble at the University of Toronto.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a sequence or a set of reference sequences.. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto referenc Author is Ruiqiang Li at the Beijing Genomics Institute. C++ for Unix.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases usi Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq genome builds for a number of organisms. Web Portal. OS independent.

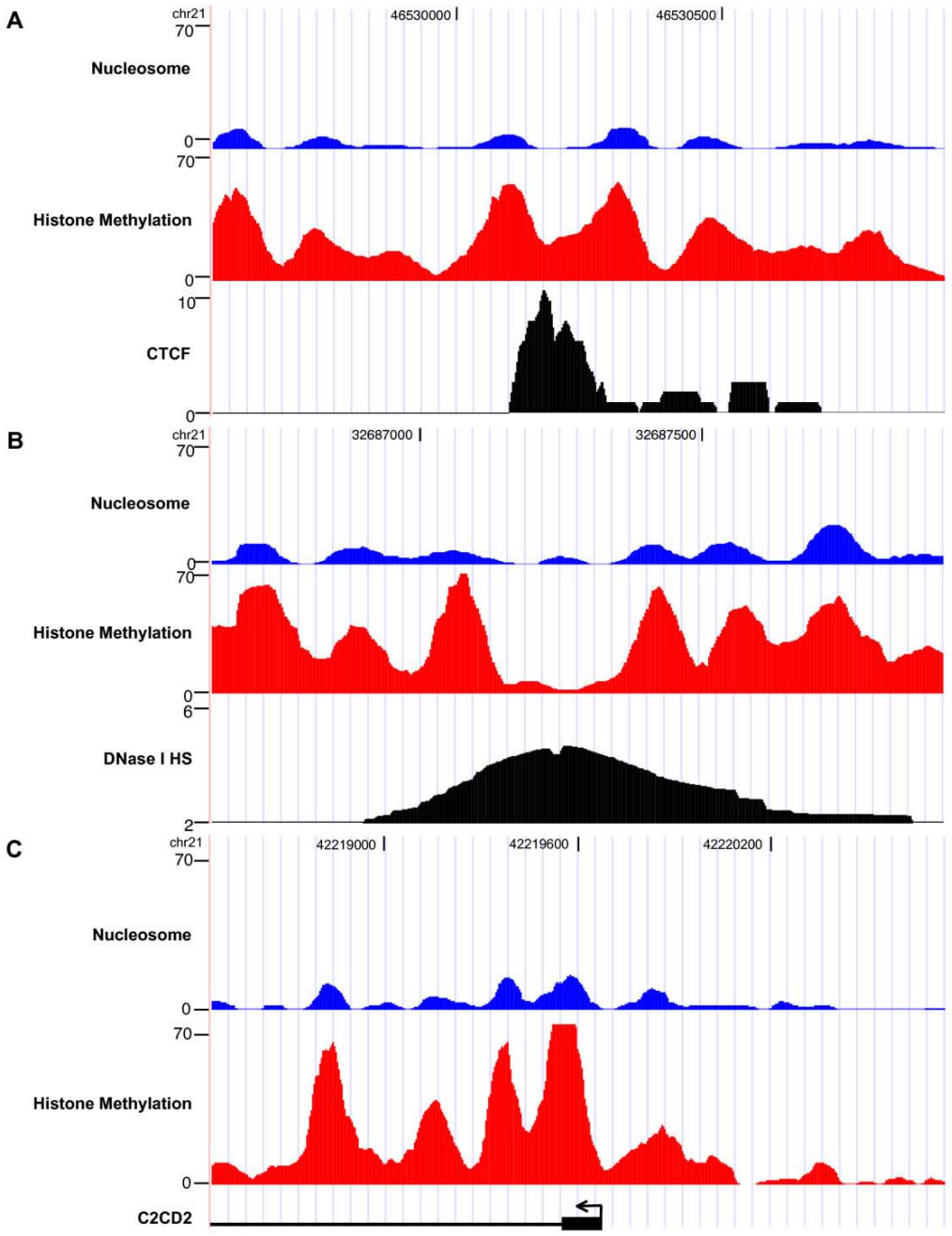
***de novo* Align/Assemble**

* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing te or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecul
* SSAKE - Version 2.0 of SSAKE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X co paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

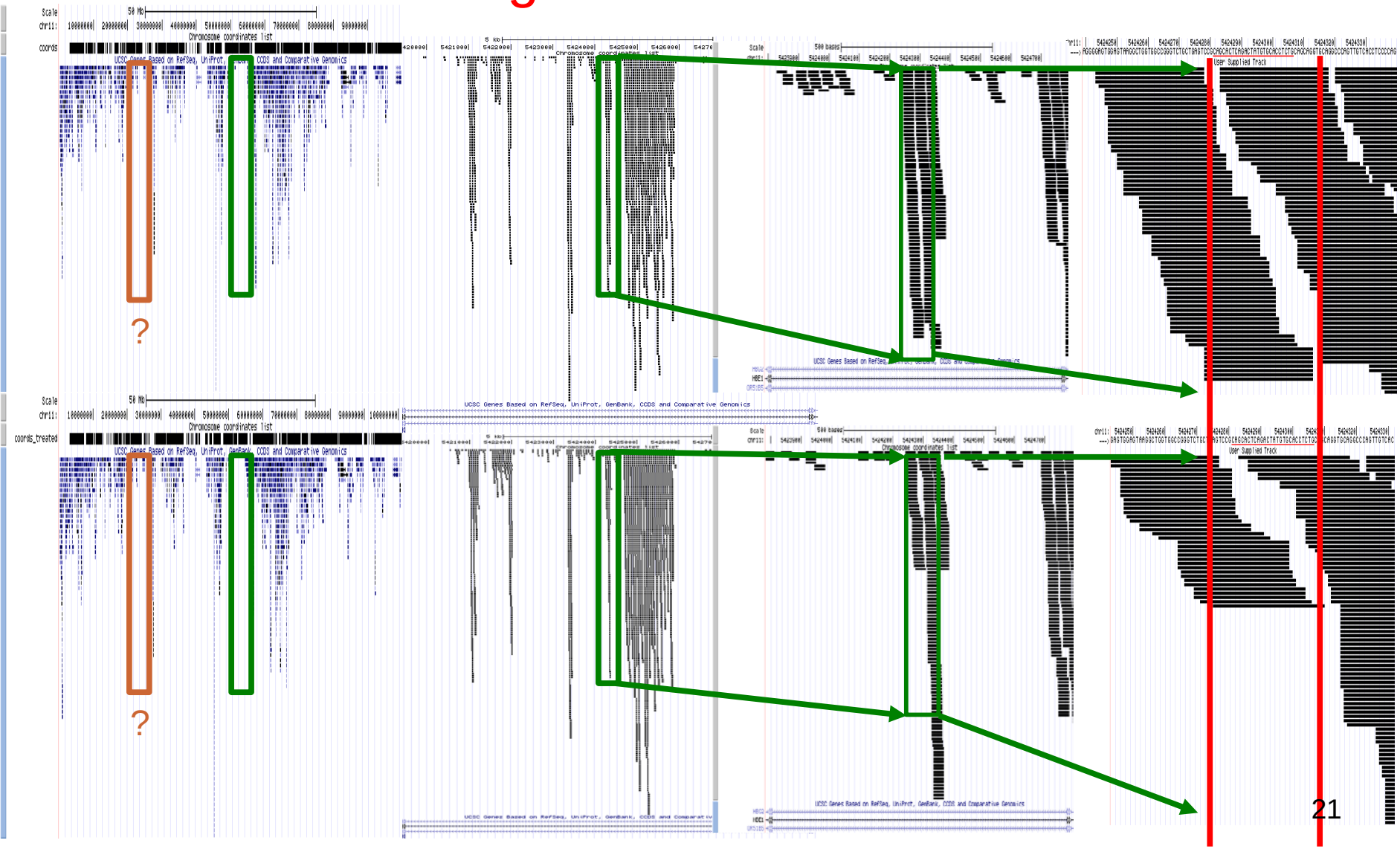
SNP/Indel Discovery:

http://seqanswers.com/

Data from ChIP Seq experiments



Xbp1 transcript Hit Maps global view and Ire1a cleavage sites zoomed in on.



Bioinformatics workflow

- Image extraction
- Base Calling, quality scoring
- Align reads to known sequence OR each other
- Assemble Reads
- Analysis of genes, regions
- Coverage, quantification
- Annotation

Summary

- Technology is available to rapidly use DNA sequencing to address biology questions without needing to be a sequencing center.

- Bioinformatics is challenged to keep up and develop robust methods as the technology is rapidly changing and improving.

- *3rd-Gen sequencing is just about here.*

Xiao-Wei Slides

- Not available for printing

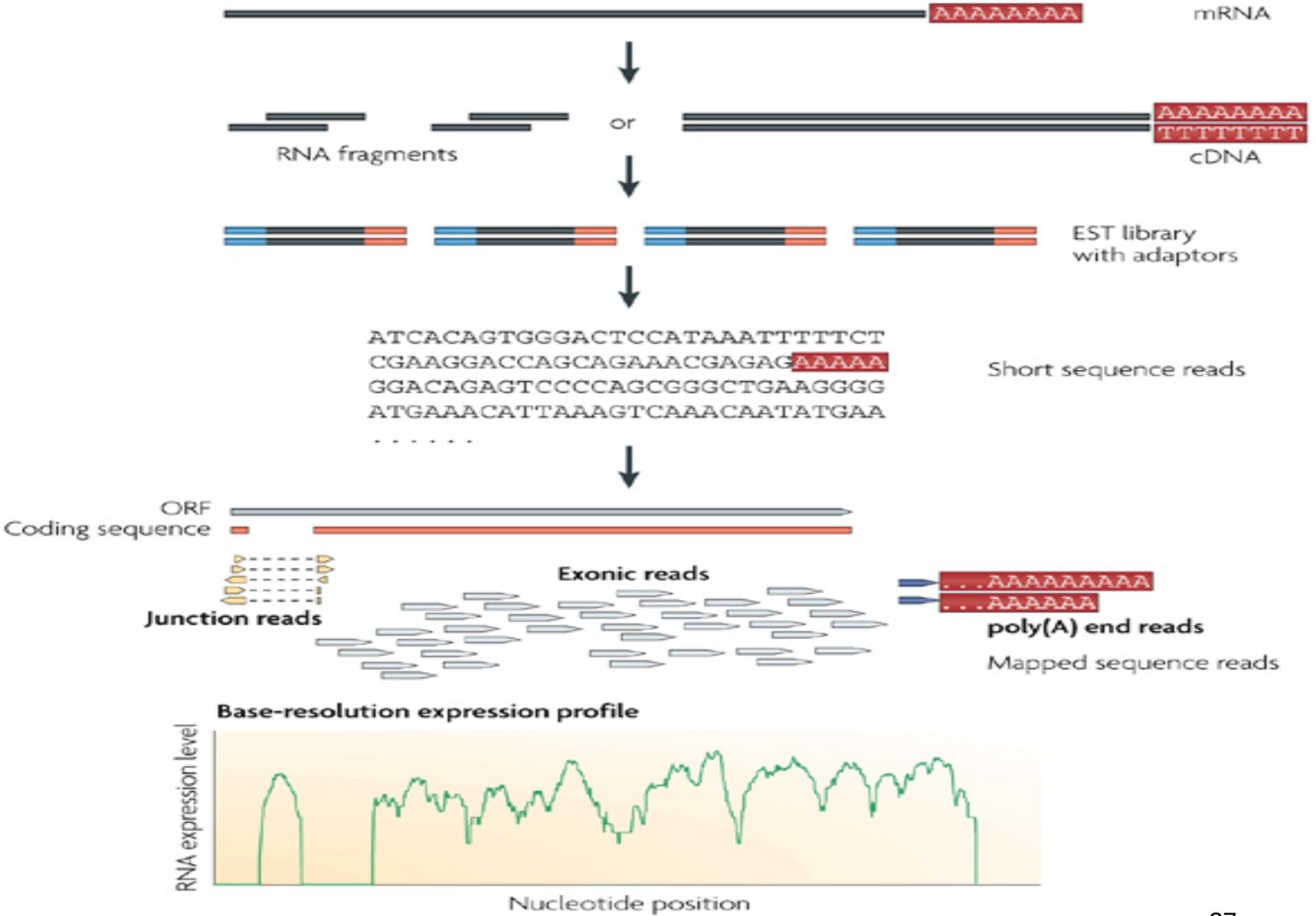
RNA-Seq

- Transcriptome analysis comparable to microarray analysis
  - Complementary DNA (cDNA) is generated from mRNA
- Rather than hybidizing to array, cDNA “reads” are sequenced using next-gen technologies
- Reads are aligned to a reference genome and a transcriptome map is constructed

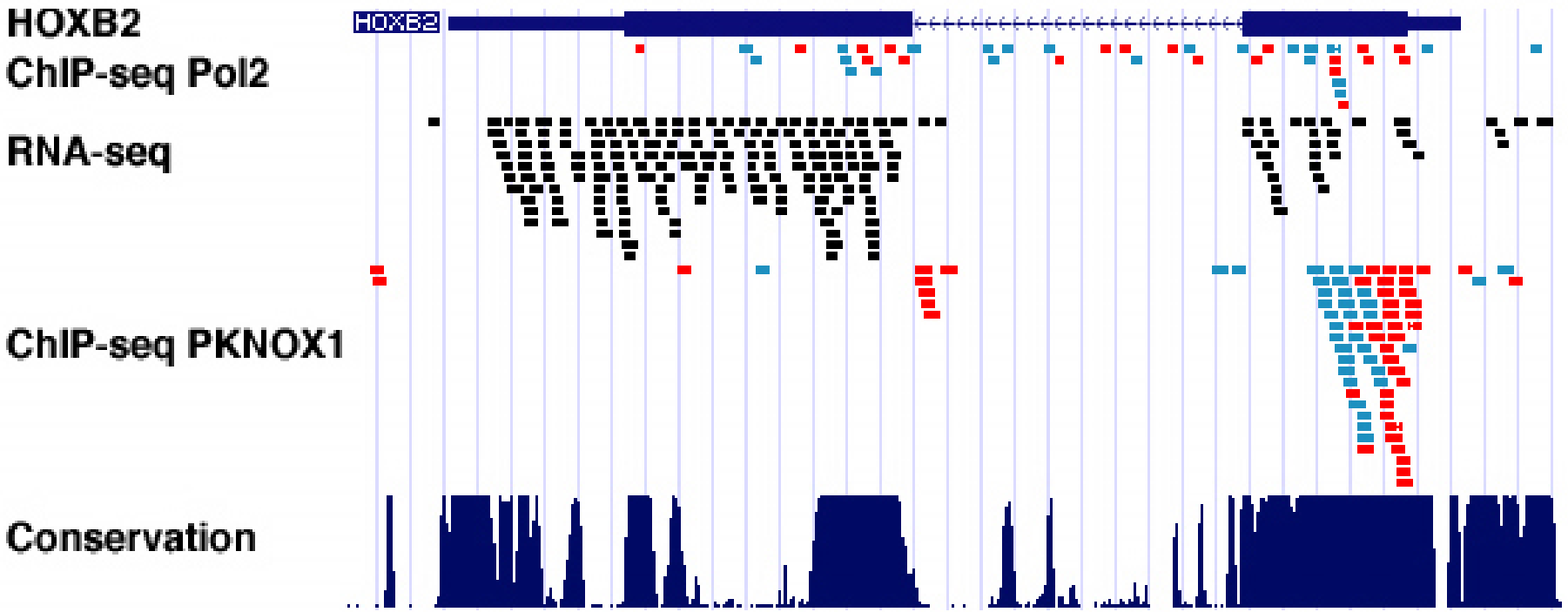
RNA-Seq: a revolutionary tool for transcriptomics Nat Rev Genet. 2009 Jan;10(1):57-63

Aims of RNA-Seq

- Quantify mRNA abundance
- Determine the transcriptional structure of genes: start sites, 5′ and 3′ ends, splicing patterns
- Quantify changing expression levels under comparable conditions
  - Sec24a wild type versus mutant


AAAAAAAA
mRNA
or
RNA fragments
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
Short sequence reads
ORF
Coding sequence
Exonic reads
Junction reads
. . .AAAAAAAAA
. . .AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

Visualizing RNA Seq Results



ChIP-seq and RNA-seq data exemplified at the HOXB2 gene

Alignment Issues

- Identify Intron/Exon Boundaries
  - Known splice sites based on gene models
  - Transcripts consistent with novel splice sites
- Multiple matches of read to sequence
  - Paralogous sequences
  - Repetitive sequence

RNA-Seq Strengths

- High-throughput quantitative measurement of transcript abundance
- Expression levels correlate well with qPCR
- Costs continue to fall due to multiplexing
- Expected to replace microarrays for transcriptomic studies
- Automated pipeline

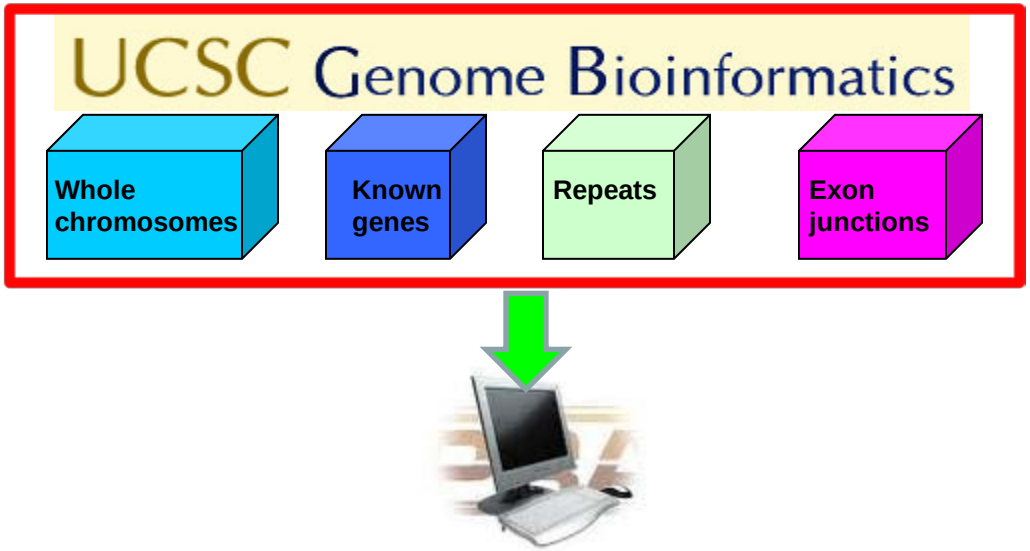
ERANGE (Enhanced Read Analysis of Gene Expression)

- Developed by Mortazavi *et al.* in Wold Lab at Caltech
- Open-source, *nix platforms, python 2.5+
- NOT a “point-and-click”, turn-key package
- Memory and computation intensive
- Dual-use for CHIP-Seq and RNA-Seq analysis
- Require Cistematic version of the genomes
  - A platform for *cis*-regulatory element analysis within and across multiple genomes
  - Available at http://cistematic.caltech.edu
- Need genome sequences and gene models from UCSC
- Details available at http://woldlab.caltech.edu/rnaseq/

RNA-Seq Pipeline Workflow

Step 1. Setup the path and prepare necessary files

- Set up access paths for ERANGE (Enhanced Read Analysis of Gene Expression), Cistematic, and Python
- Download and prepare input data locally from UCSC: chromosomes, gene models, repeatMask sequences (and other necessary annotation files) genome.ucsc.edu



- Create splice files and build expanded genome & repeatMask database

RNA-Seq Pipeline Workflow (Cont.)

Step 2. Map reads to Mouse genome

- Obtain NGS read sequence files (i.e. Solexa read data file s_*_sequence.txt)

```
@unknown_0001:6:1:1156:2319#0/1
GATAATCCATCACNCGTTAAAAATTTGCNTACTACCA
+unknown_0001:6:1:1156:2319#0/1
Za^^`aaaaa^Z^ZEZ[[[[aaaaaaaaa^E^`^^`aaa
```

- Align reads to the reference genome
  - ELAND (for read less then 30 bp)
  - BOWTIE (read length > 32 bp)
  - BLAT (read length > 50 bp)

RNA-Seq Pipeline Workflow (Cont.)

Step 3. ERANGE counting pipeline (http://woldlab.caltech.edu/rnaseq)

- Count reads falling on gene models
  - Filter out reads that overlap repeats
  - Map reads with a certain radius of genes
- Identify novel transcripts
  - Reads from all samples that did not fall within known exons were aggregated into CANDIDATE exons by requiring regions with at least 15 reads whose starts are not separated by more than 30bp
  - CANDIDATE exons that fell within 20 kb of one another but further than 20 kb from any other gene were aggregated into predicted “FAR” loci
- Weight multi-reads

RNA-Seq Pipeline Workflow (Cont.)

Step 3. ERANGE counting pipeline (http://woldlab.caltech.edu/rnaseq)

- Reads that fell onto exons and candidate exons as well as splices were summed up for each locus and normalized by the predicted mRNA length into expanded exonic read density
  - Expressed as reads per KB per million reads (RPKM) using the formula $10^9 C/NL$
    - Where C is the number of mappable reads falling on exons
    - N is the total number of mappable reads
    - L is the sum of the exon lengths in bps
  - *.final.rpkm (uniques + spliced + multireads + RNAFAR) is the most comprehensive result among all of the output files
  - RPKM can be converted into absolute transcript numbers

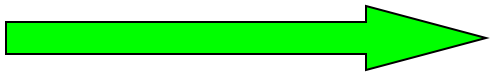
RNA-Seq Pipeline Workflow (Cont.)

Step 4. Gene level differential expression (DE) analysis based on ERANGE output

ERANGE output file (*.final.rpkm)

```
gene      len_kb    RPKM
Alb1      2.028     31420.21
Apoe      1.089     13931.63
Apoa2     0.477     11242.40
…
```

R scripts to detect DE genes



A list of differentially expressed genes

```
Entrez Gene ID
77371
13119
30939
76574
13117
14373
…
```

RNA-Seq Pipeline Workflow (Cont.)

Step 5. Biological concepts enrichment analysis for differentially expressed (down-regulated) genes

A list of differentially expressed genes

Entrez Gene ID
77371
13119
30939
76574
13117
14373
…

Top over-represented biological concepts

(Metabolic Processes)
fatty acid synthesis
monocarboxylic acid
carboxylic acid
organic acid
…

References

- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 2008*,* 5, 621-628.
- Sartor M, Mahavisno V, Keshamouni V, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T, Athey B, Omenn G: ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* 2010*,* 26(4):456-463.