RNA-seq Approach to Study Gene Expression Profiles in Non-Model Organisms

Asela Wijeratne (wijeratne.1@osu.edu)
Saranga Wijeratne (wijeratne.3@osu.edu)
Tea Meulia (Meulia.1@osu.edu)

THE OHIO STATE UNIVERSITY

Why we sequence transcriptomes?

- Show repertoire of expressed sequences, including rare transcripts
- Gene expression
- SNPs
- Alternative splicing
- Structural variation
- Practical alternative to genome sequencing for non-model organisms

Genomic model vs. non-model organisms

- **Model organism** is a non-human species that is extensively studied to understand particular biological phenomena
- Genomic model organisms:
  - Occupy a pivotal position in the evolutionary tree
  - Some quality of their genome makes them ideal to study
- Non-Genomic model organisms
  - Rest of them
  - Important for many reasons:
    - Human pathogens
    - Agricultural pathogens and pests

RNA-seq for non-model organisms

- No sequenced genomes most of the time
- Most analytical tools are designed for model organisms
- Present unique challenges for quality control for data analyses

What are we going to learn

- Basics of initial designing the experiment
- Analyses pipeline
  - Primary analyses
    - Read preprocessing
    - Transcript assembly
    - Assessing quality of assembly
    - Mapping reads back assembled transcripts
  - Secondary analyses
    - Transcript characterization and annotation
    - Comparative gene expression

Sequencing

Platform of choice

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.

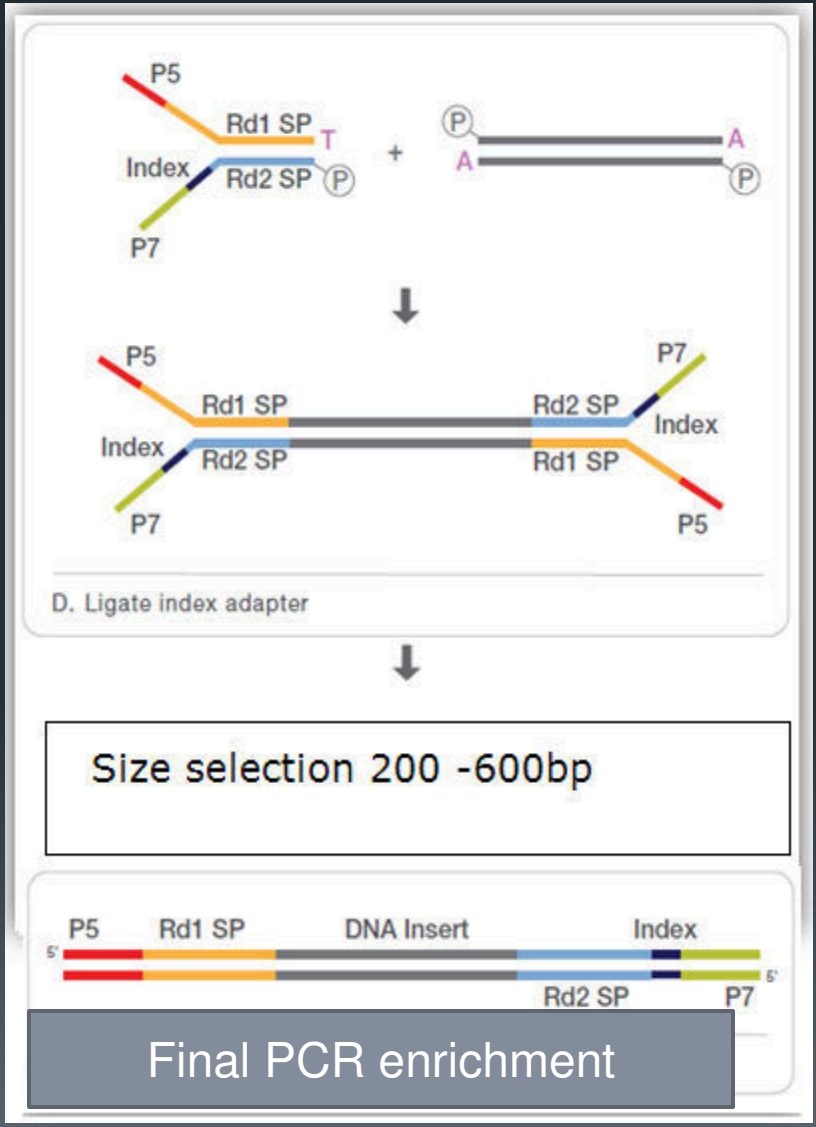
From Michael Metzker, http://view.ncbi.nlm.nih.gov/pubmed/19997069

Different read types for Illumina sequencing

Read type

Paired end reads

Single end reads



Replication and sequencing depth

- **Replication:**
  - At least two biological replicates for expression analyses
- **Sequencing depth**:
  - Depends:
    - Goals of the experiments
    - Samples and conditions; sample preparation
    - Sequencing type
    - Number and length of the genes

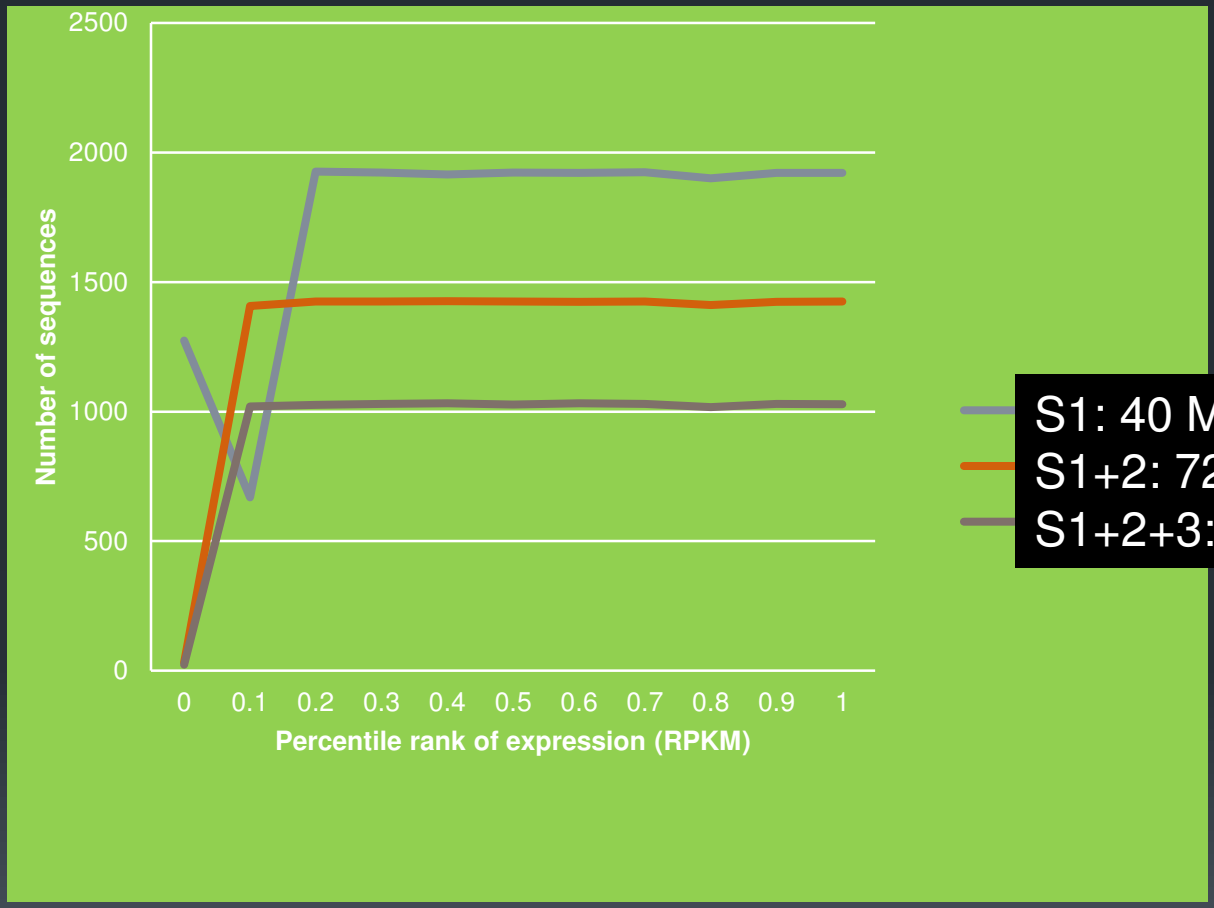
No of transcripts assembled increases as the no of reads increase

Combination	Reads (millons)	Expressed genes $	Number of reference cDNA with matches to assembled sequences*	
			Redundant	Non-redundant
Sample1	40	13416	10500	10060
Sample 1+2	72	15646	14250	13607
Sample 1+2+3	112	16916	19223	14370

$These is based on the alignments to reference cDNA; Number of sequences with more than 200 Illumina reads aligned

*number of reference cDNA that showed sequence similarity to assembled transcripts that showed at least 80% coverage and had at least 95% sequence identity

Number of sequences in different expression groups

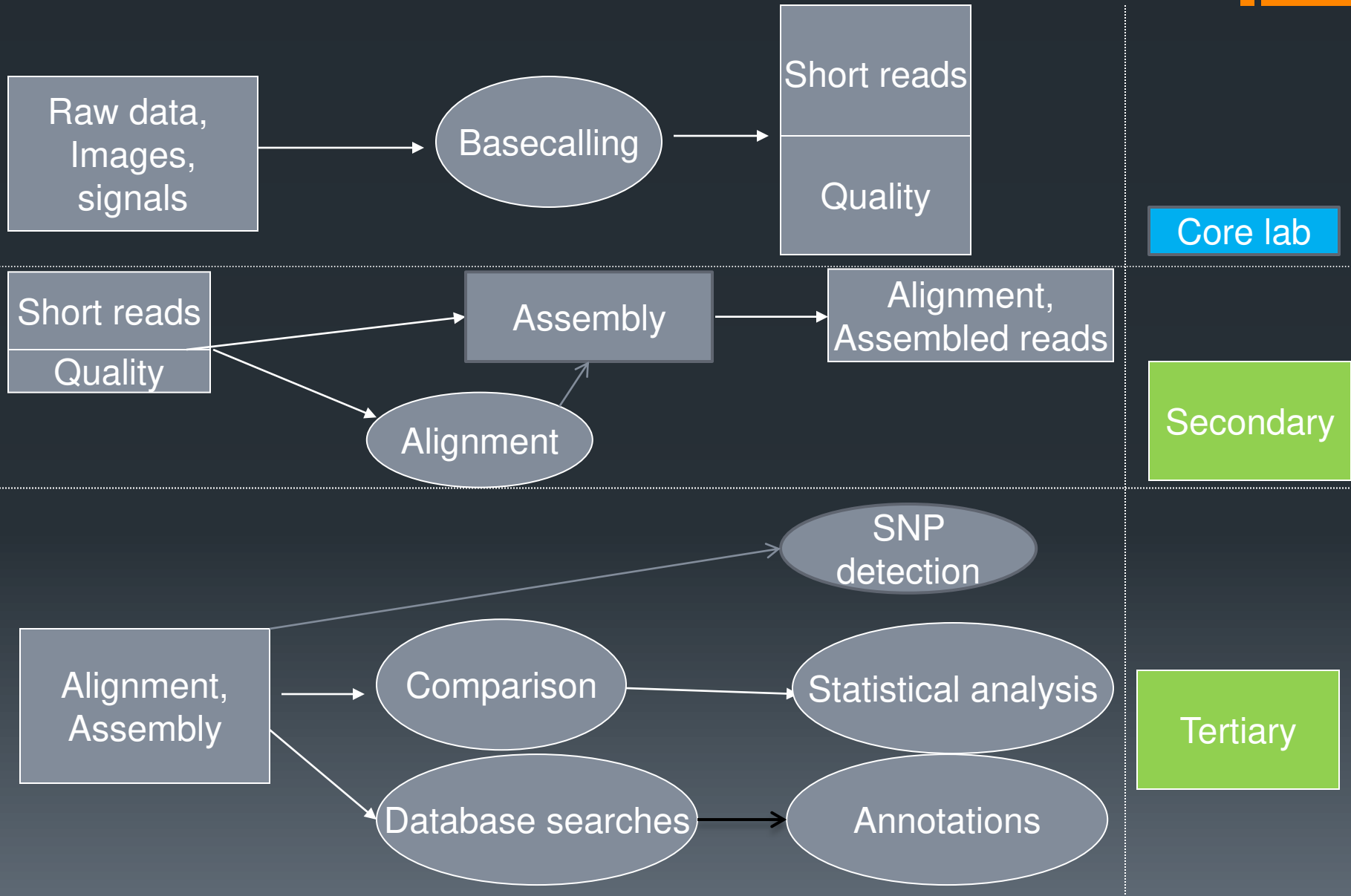


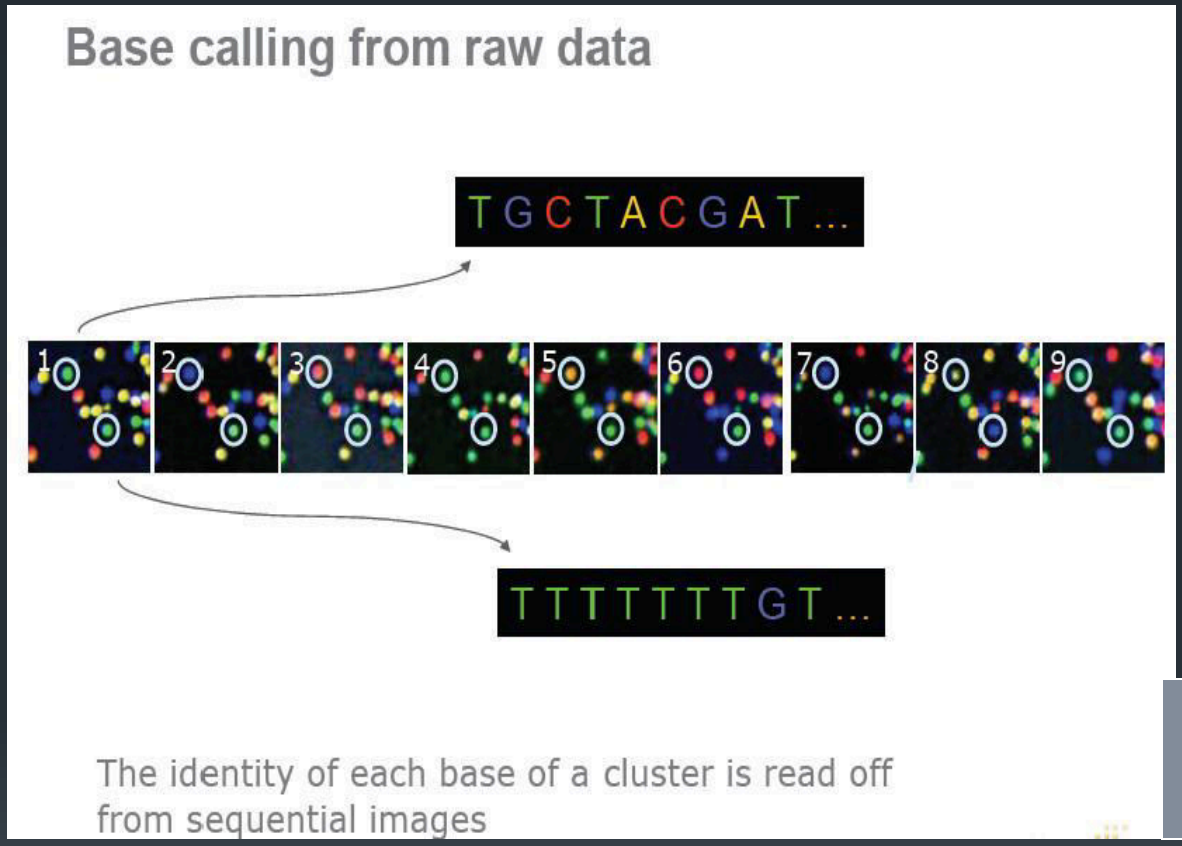
This shows number of reference cDNA that showed sequence similarity to assembled transcripts that showed at least 80% coverage and had at least 95% sequence identity. X axis is percentile ranking of the expression and Y axis is the number of sequences

Analyzing the data

- Manually using UNIX terminal prompt
- Automated using PERL or Python scripts
- Using Makefile
- Reproducibly Workflow environment
  - Galaxy
  - Taverna

Steps of data analyses





CASAVA

Fastq file

From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4

FASTQ

```
@BILLIEHOLIDAY:1:1:6:768#0/1
CATGATGGCAGAGGCAGAGGACAGGTTGCCAAAGCTCTCGCTTCTGGAACGTCTGAGGTTAT
CAATAAGCTC
+BILLIEHOLIDAY:1:1:6:768#0/1
abbbababbbbaa_`a`_aa`^_aaa_]_aa^`]^a`abbbbbbbbbbbbbabbbba`abbb
bbbbbbbbbb
```

Whatever_name	the unique instrument name
1	flowcell lane
1	tile number within the flowcell lane
6	'x'-coordinate of the cluster within the tile
768	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Sources of error

- De-phasing
  - Lagging strand de-phasing from incomplete extension
  - Leading strand de-phasing from over-extension
- Polymerase errors (10^{-5} to 10^{-7})
- More likely to have an error after G
- PCR induced errors (AT or GC rich regions)
- Cross-cluster bridge formation

BER=base error rate

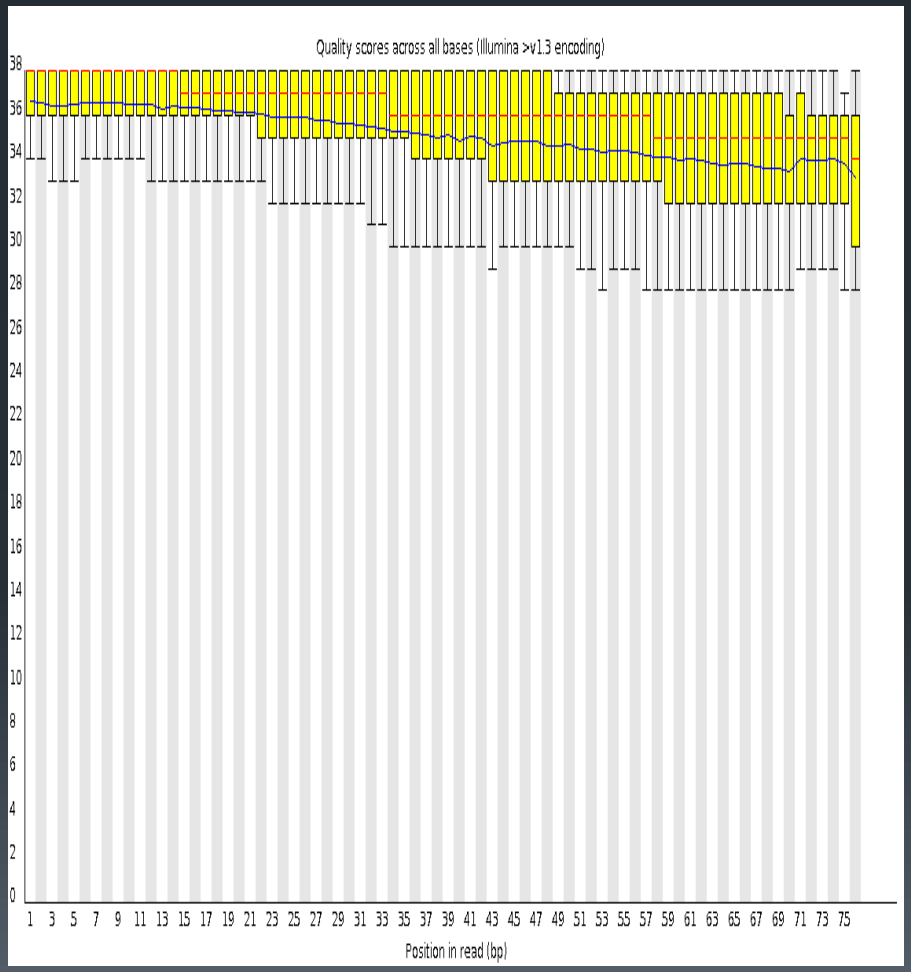
- BER: Estimated probability of a base being wrong
- Phred quality score (as of Illumina pipeline 1.3):

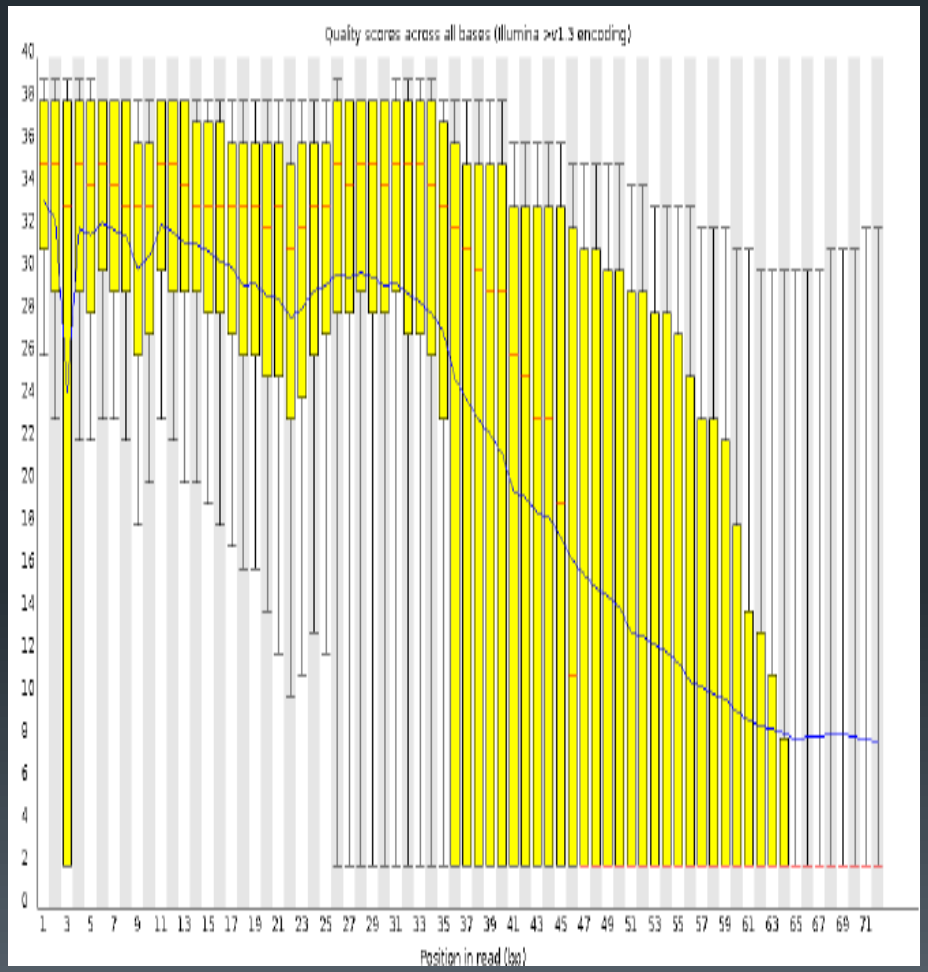
$$Q = -10 \cdot \log_{10}(BER)$$

- In a FASTQ file they are encoded as ASCII: Q+64
- May be used to filter out poor quality reads, and to improve alignments

Phred Quality Score	Probability of Incorrect Based Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

Quality assessment - FASTQC





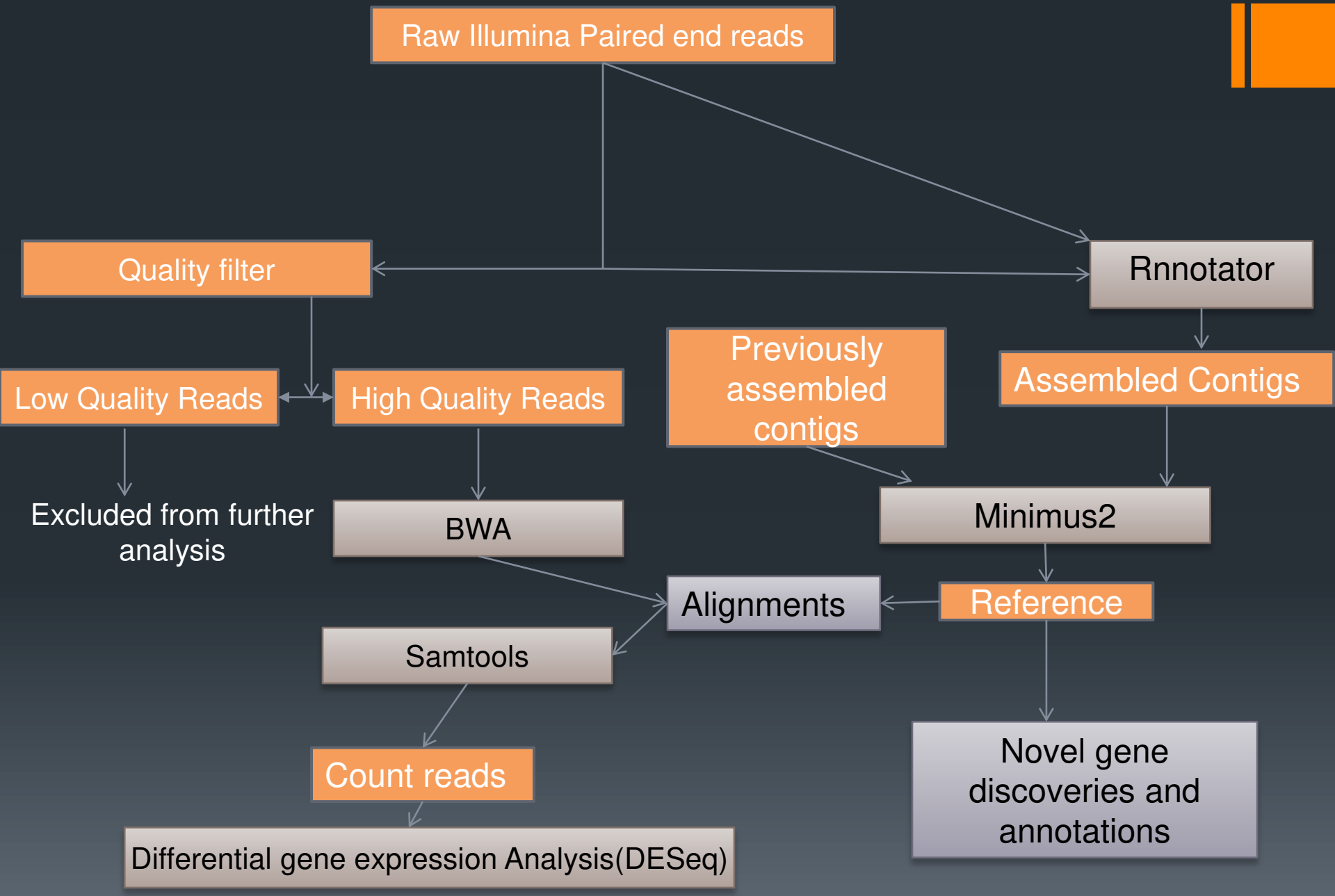
Preprocessing

- Why
  - Get rid of artifacts from library preparations (PCR) and sequencing steps (errors/adapters)
    - Resulted in fragmented assemblies or mis-assemblies
    - Inflate dataset
- Remove
  - Adapter sequences
  - Low quality bases
    - FASTX-toolkit
    - Cutadapt
    - Custom perl script from UC Davis for paired-end reads
- Remove contaminates
  - DeconSeq
    - Web based
    - Standalone

Assembling sequence reads into transcripts

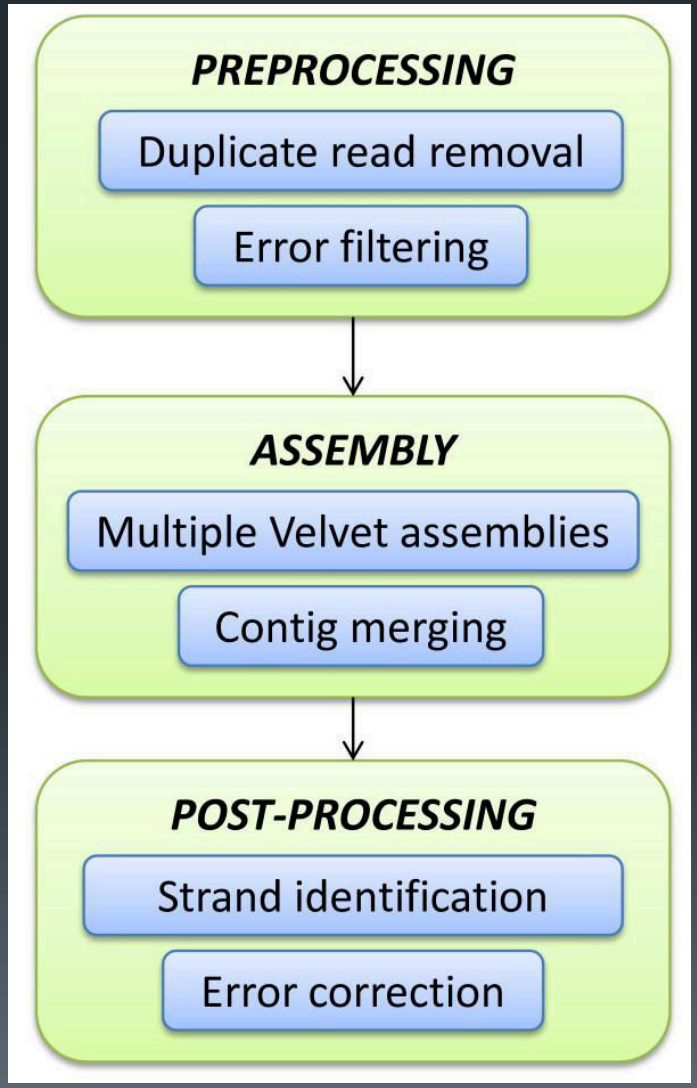
Transcript Assembly of Illumina reads

- Challenges
  - Short reads (76-150bp)
  - Large datasets
  - Needs a large number of reads
  - Transcript expression is not uniform
- *De novo* assembly
  - No reference
    - Rnnotator, Trinity
  - Highbred assembly
    - Uses Longer reads (previous cDNA or 454 data) with Illumina short reads (Best approach)
- Reference based assembly
  - TopHat/Cufflinks, ERANGE, and Scripture


Raw Illumina Paired end reads
Quality filter
Rnnotator
Low Quality Reads
High Quality Reads
Previously assembled contigs
Assembled Contigs
Excluded from further analysis
BWA
Minimus2
Alignments
Reference
Samtools
Count reads
Novel gene discoveries and annotations
Differential gene expression Analysis(DESeq)

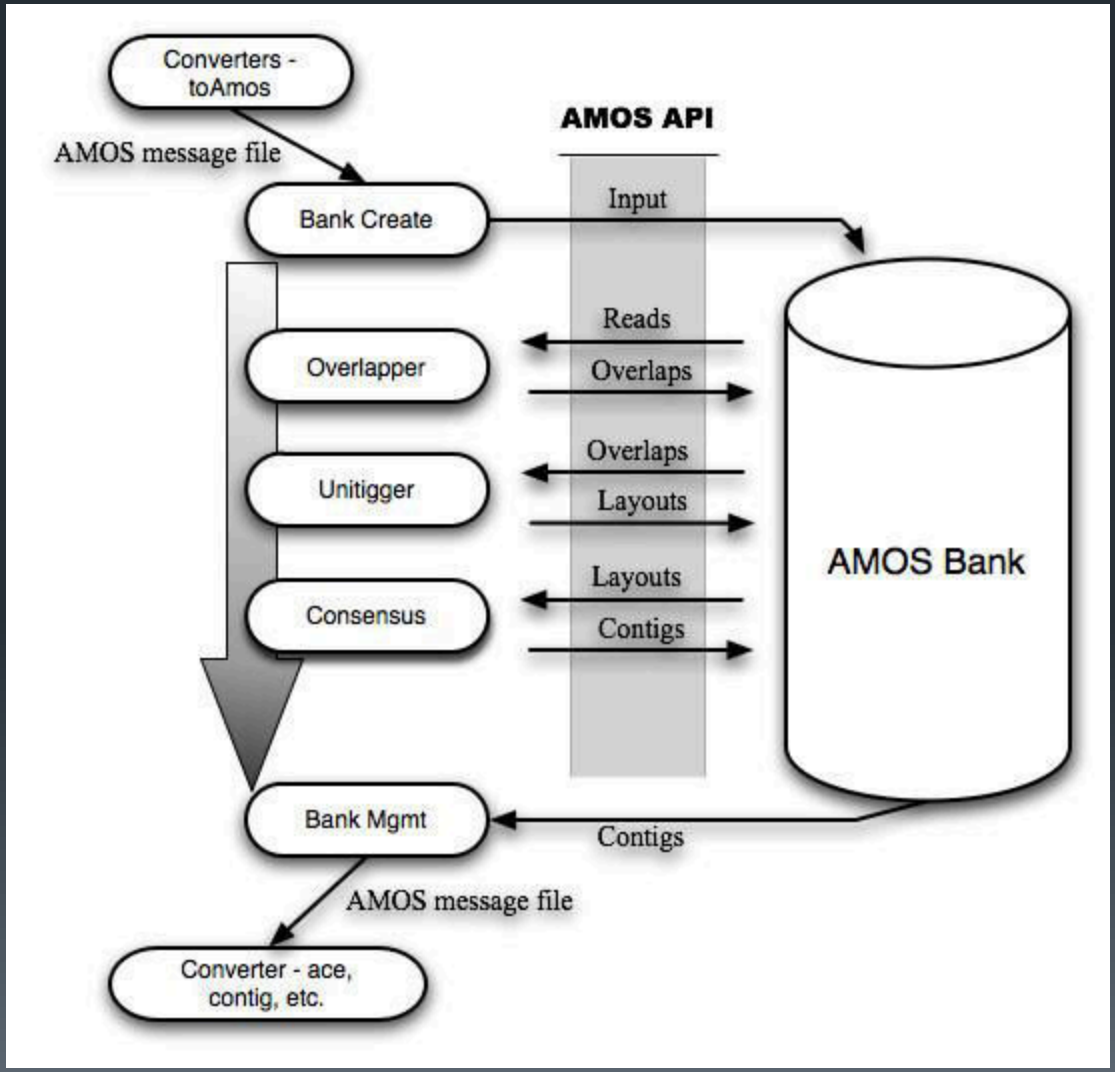
The Rnnotator assembly pipeline

- Input preprocessed data
- Strand information
- Remove duplicates
- K-mer based filtering



Martin et al. *BMC Genomics* 2010 11:663 doi:10.1186/1471-2164-11-663

Minumus2 in Amos package



Assessing the quality of the assembly

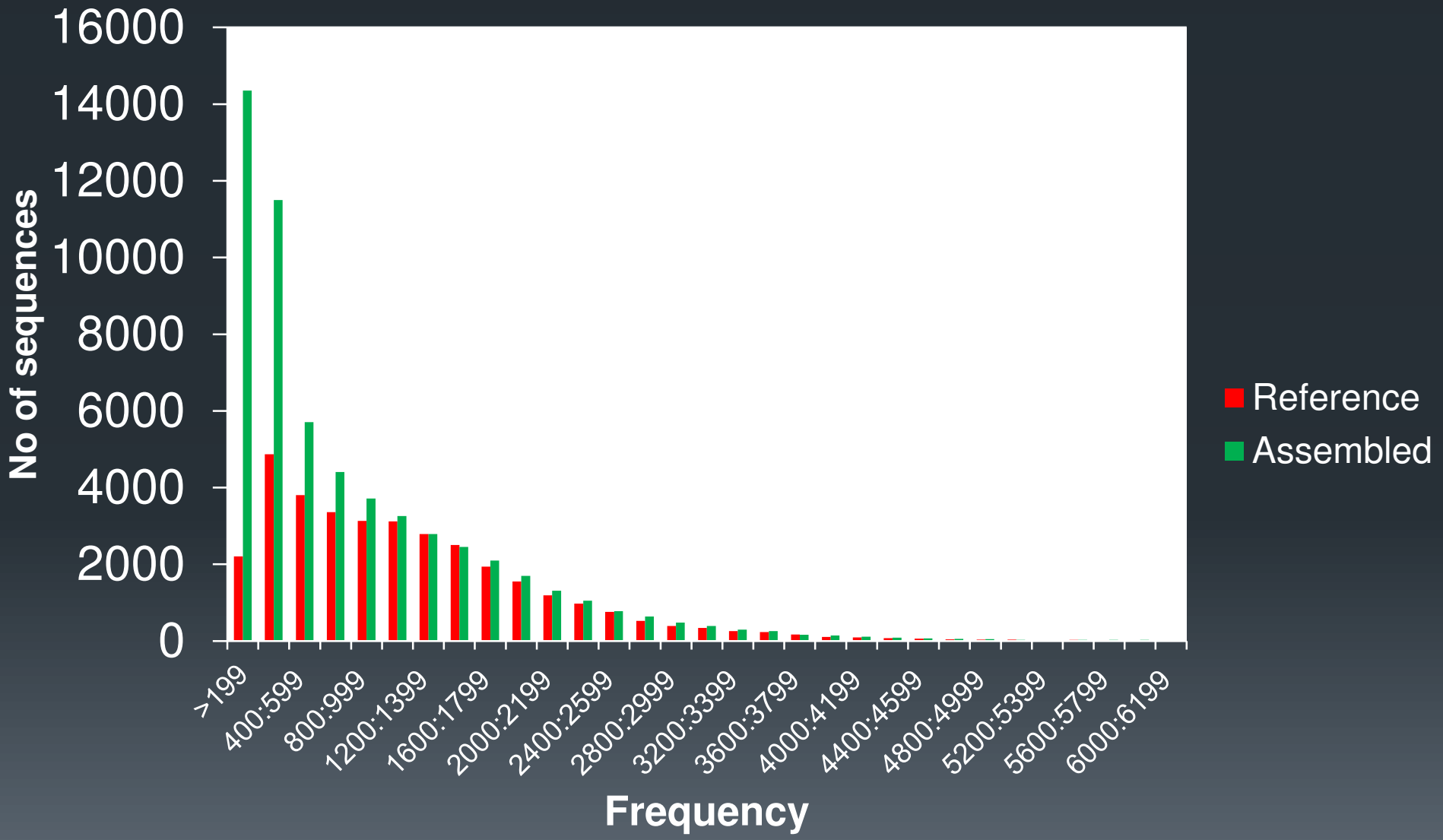
A comparison of the performance between the Rnnotator assembly and a single Velvet assembly.

	Rnnotator (non-stranded)	Rnnotator	Velvet	Oases	Multiple-*k*
***C. albicans* SC5314**					
Accuracy[1]	94.0	95.0	97.4	92.3	96.6
Completeness[2]	81.9	80.4	66.7	79.9	85.9
▪ Contiguity[3]	58.4	58.0	46.6	47.9	37.3
▪ Gene fusions[4]	1.73	0.26	1.18	1.31	0.20

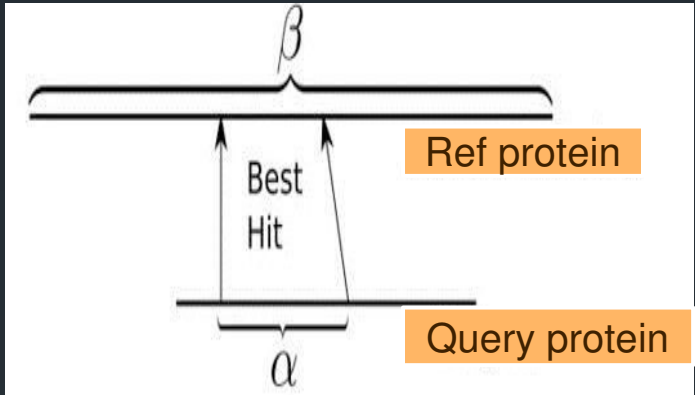
[1]Accuracy: the percentage of contigs that share at least 95% identity with the reference genome;
[2]Completeness: percentage of known genes covered by the contigs to at least 80% of the gene length;
[3]Contiguity: percentage of complete genes covered by a *single* contig over at least 80% of the gene length.
[4]Gene fusions: the percentage of contigs that contain more than 50% of two or more annotated genes.

Martin *et al. BMC Genomics* 2010 **11**:663 doi:10.1186/1471-2164-11-663

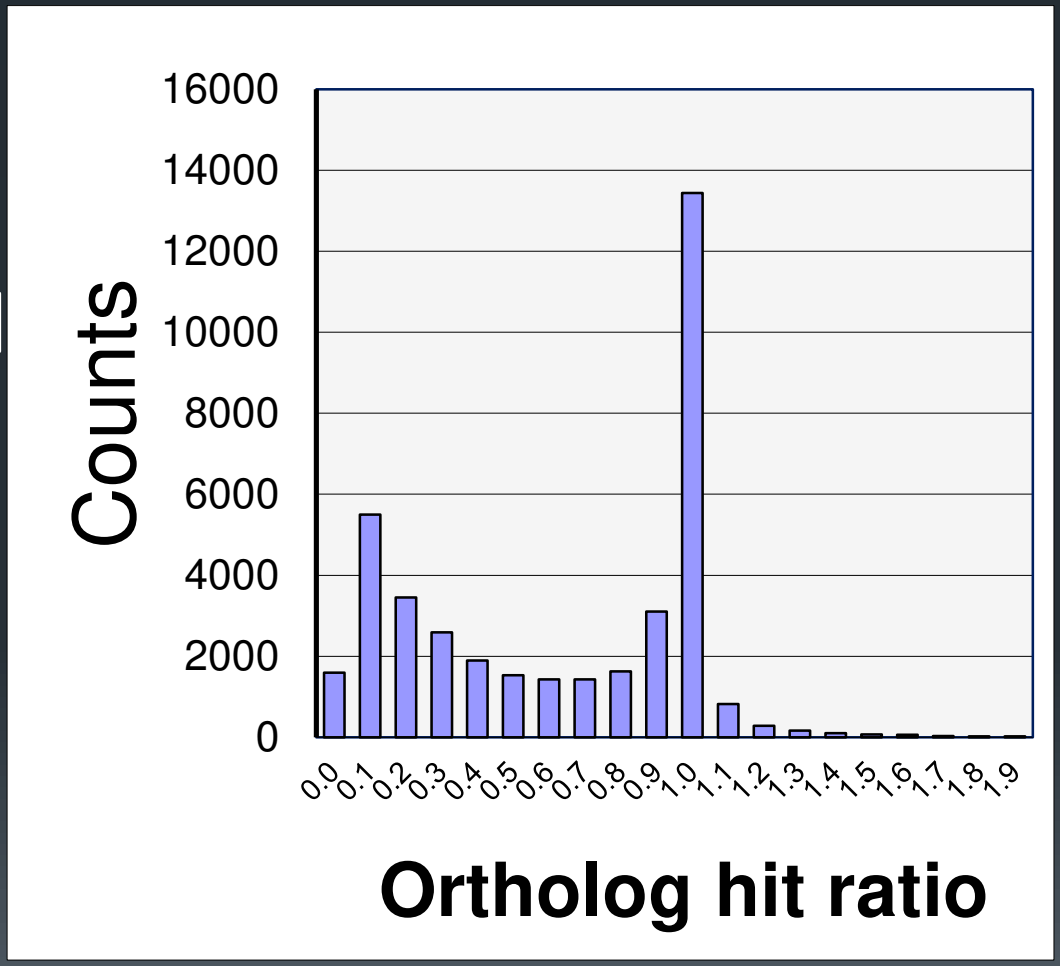
Length distribution



Ortholog Hit Ratio indicates the completeness of assembly



Total expressed	20721
More than 200 reads	17068



About 20,000 assembled ESTs had more than 0.8 OHR

O'Neil, 2010; Ewen-Campen, 2011 Python script

Depth of sequencing and assembly can be assessed using Ultra Conserved Othologs (UCO)

- Compare with eukaryotic ultraconserved genes
- http://compgenomics.ucdavis.edu/compositae_reference.php

Annotation of assembled transcripts

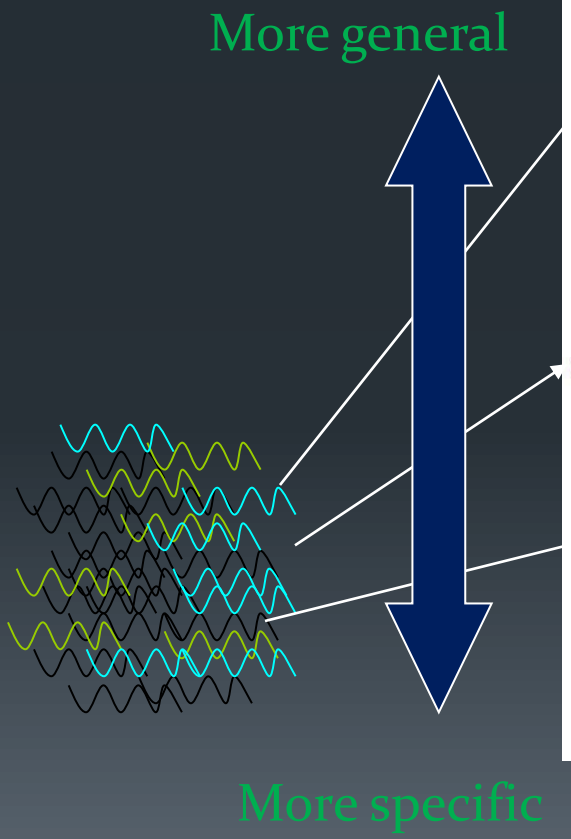
The Gene Ontology

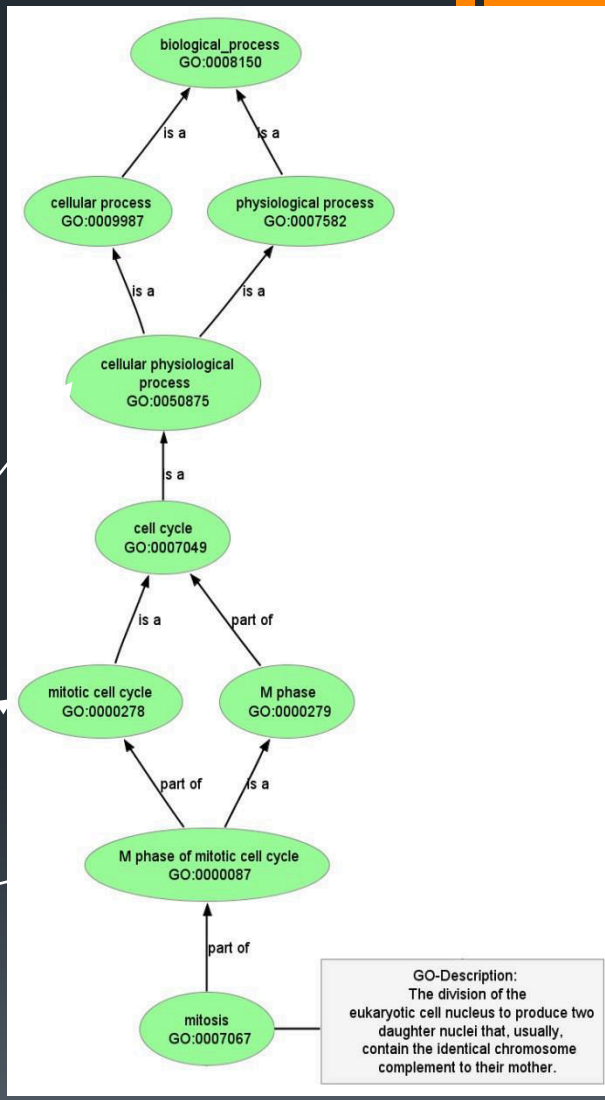
- Gene Ontology Consortium
- Provide a controlled vocabulary to describe gene and gene product attributes in any organism
- Includes both the development of the Ontology and the maintenance of a Database of annotations

Adapted from Blast2Go presentation

The Ontology

- Annotations are given to the most specific (low) level.
- True path rule: annotation at a term implies annotation to all its parent terms
- Annotation is given with an Evidence Code:
  - **IDA**: inferred by direct assay
  - **TAS**: traceable author statement
  - **ISS**: infered by sequence similarity
  - **IEA**: electronic annotation
  - ....





Adapted from Blast2Go presentation

Blast2GO

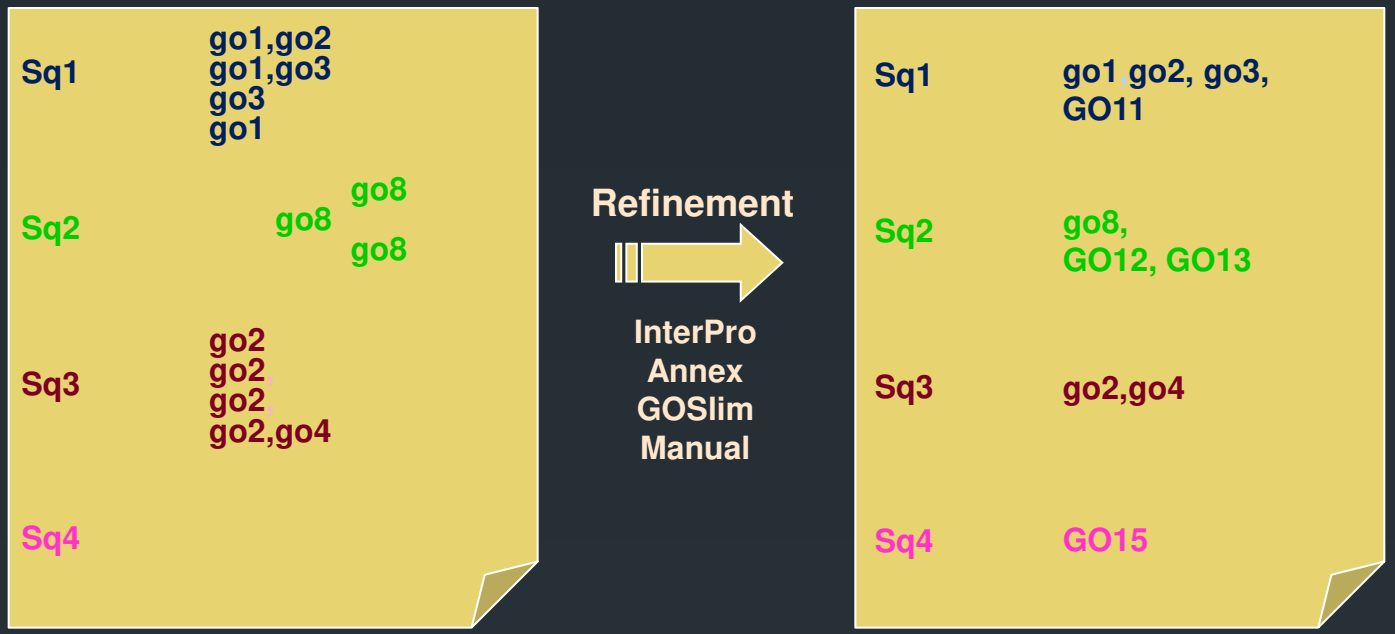
- Suite for functional annotation and data mining on functional data
  - Considerations for **annotation**
    - Simlarity
    - Length of the overlap
    - Percentage of hit sequence spanned by the overlap
    - Evidence original annotation
    - Blast hits and motif hits
    - Refinement by additional methods
  - Visualization:
    - Annotation charts
    - **Knowledge discovery on the DAG**
- Desktop Java application
- web interface @ Babelomics: **Babelomics for non-model**

Adapted from Blast2Go presentation

Blast2GO Annotation strategy

Adapted from Blast2Go presentation

Blast2GO Annotation Strategy

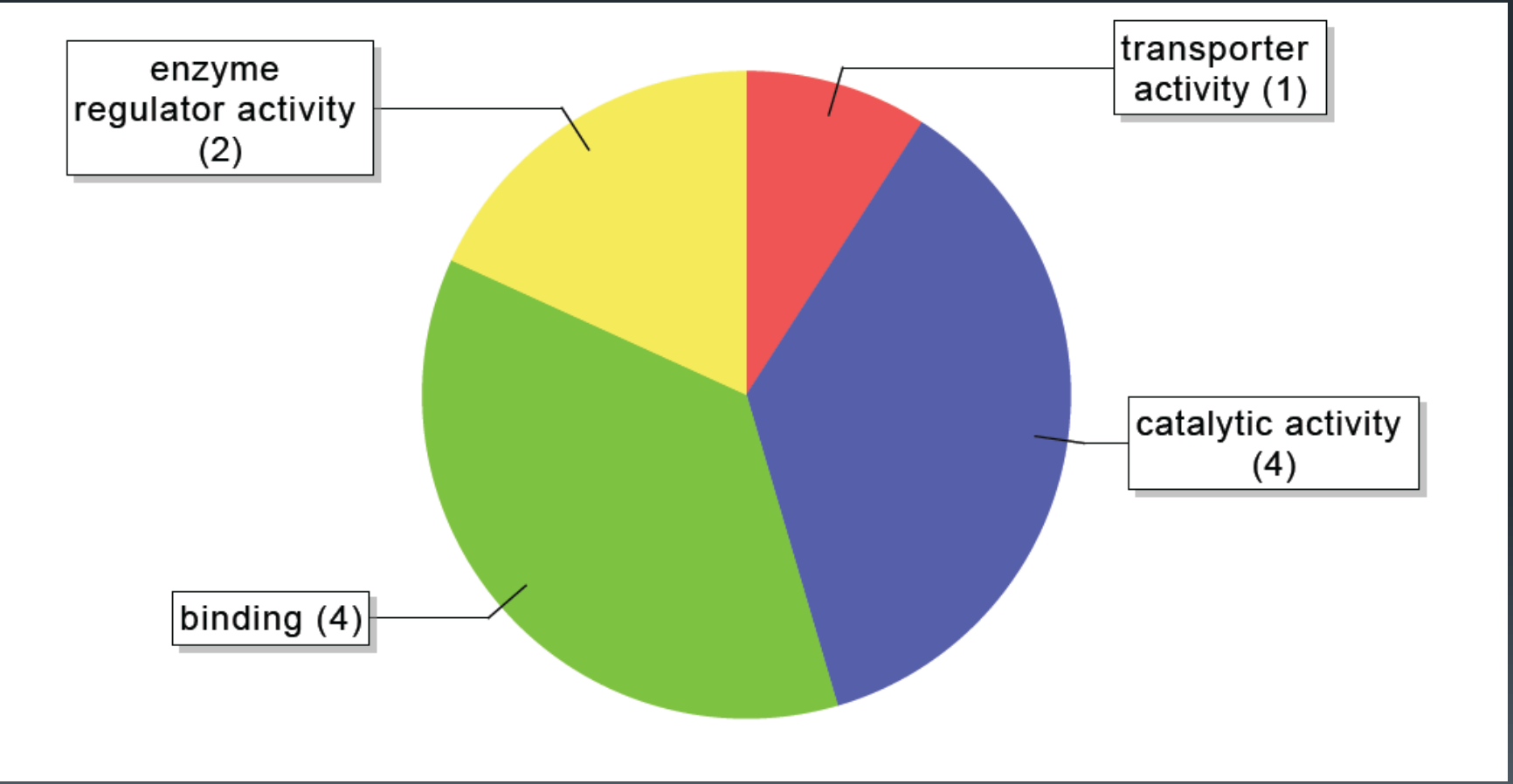


More information:
http://www.blast2go.com/data/blast2go/b2g_tutorial_23062009.pdf

Blast2go output

Seq. Name	Seq. Description	Seq. Length	#Hits	min. eValue	mean Similarity	#GOs	GOs
Gene 1	---NA---	165	0			0	
Gene 2	pathogenesis-related protein 1	480	20	9.57E-94	89.95%	3	C:extracellular region; P:defense response to fungus; P:killing of cells of another organism

Molecular function



Mapping

Mapping reads to assembled transcripts

- Aligners
  - Burrow Wheeler Transform (BWT)
    - Fast
    - Need good quality data
    - Bowtie, BWA, SOAP2
  - Hash tables
    - Slower
    - More sensitive – better for SNP finding
    - PerM, SHRIMP, BFAST, ELAND
- Mapping back to a reference genome
  - Splice aware aligners:
    - TOPHAT, MapSplice
- Mapping back to assembled transcripts
  - No splice junctions

mRNA seq reads mapped to cDNA using BWA

Sample	*Total Reads*	*Reads Mapped*	*% Reads Mapped*	
Sample 1	36,856,702	29,479,287	80%	Treatment 1
Sample 2	38,967,190	34,688,060	89%	
Sample 3	50,018,335	37,275,875	75%	
Sample 1	39,584,939	30,702,758	78%	Treatment 2
Sample 2	32,622,057	27,450,258	84%	
Sample 3	39,060,469	31,627,675	81%	

Differential gene expression analyses

Estimate the number of reads in each transcript

- Short reads aligned to a transcript are counted
- Count each read at most once
- Reads are discarded
  - Not uniquely mapped
  - Aligned to several genes
  - Poor alignment quality score
  - (for paired-end reads) the mates matches to different genes

Simon Anders

Normalization

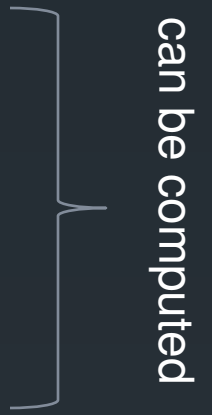
- Number of reads (coverage) vary between samples (Sequencing depth
- Other technical effects
- **RPKM** (Reads Per KB per Million mapped reads)
  - divide counts per million reads and by gene length
- RPKM assumes:
  - Total amount of RNA per cell is constant
  - Most genes do not change expression
- RPKM is invalid if there are a few very highly expressed genes that have dramatic change in expression (dominate the pool of reads)
- Quantile normalization (Bullard, 2010)
- Produces non-integer counts, not good for Poisson or Negative Binomial model based methods

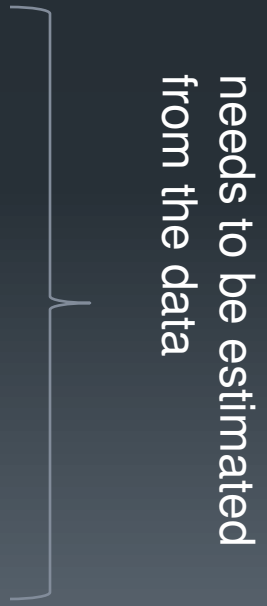
Scaling factor normalization as in DEseq

- Reference sample
  - The geometric mean of the counts in all samples for each gene
- Get the sequencing depth of a sample relative to the reference
  - Calculate for each gene the quotient of the counts in the test sample divided by the counts the of reference sample
- Median of all the quotients is the depth of the library

Noise

- Shot noise
  - The variance in counts that persists even if everything is exactly equal
    - unavoidable, appears even with perfect replication
    - dominant noise for weakly expressed genes
- Technical noise
  - from sample preparation and sequencing
- Biological noise
  - Dominant noise for strongly expressed genes

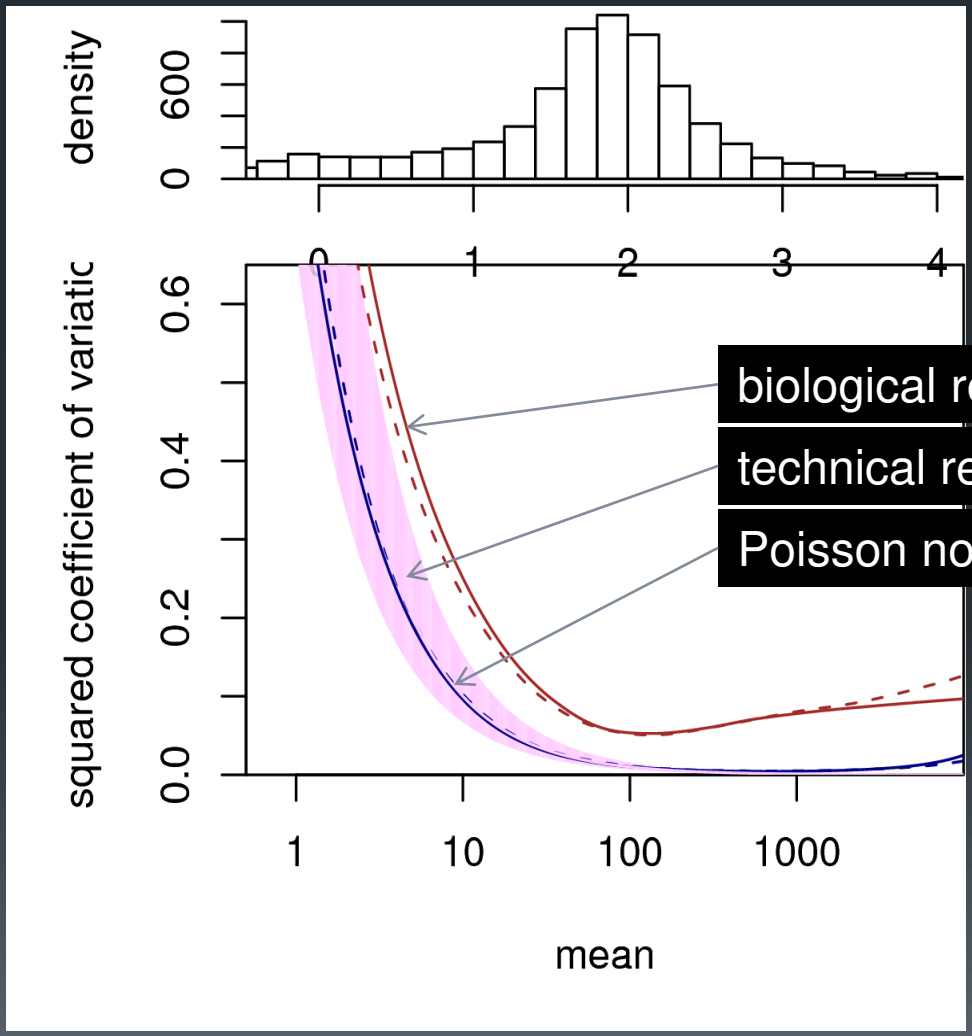




Statistical methods for DEG analyses

- Mathematically shown:
  - If
    - number of reads is large
    - Probability of a read mapped to a gene is small
  - Binomial distribution is well approximated by Poisson distribution
- Poisson distribution: mean = variance
- counts for the same gene from different technical replicates have a variance equal to the mean (Poisson)
- counts for the same gene from different biological replicates have a variance exceeding the mean (overdispersion)
- The negative-binomial distribution
  - A commonly used generalization of the Poisson distribution with two parameters
- Estimate a scaling factor to use with the statistical model

Biological vs technical replicates



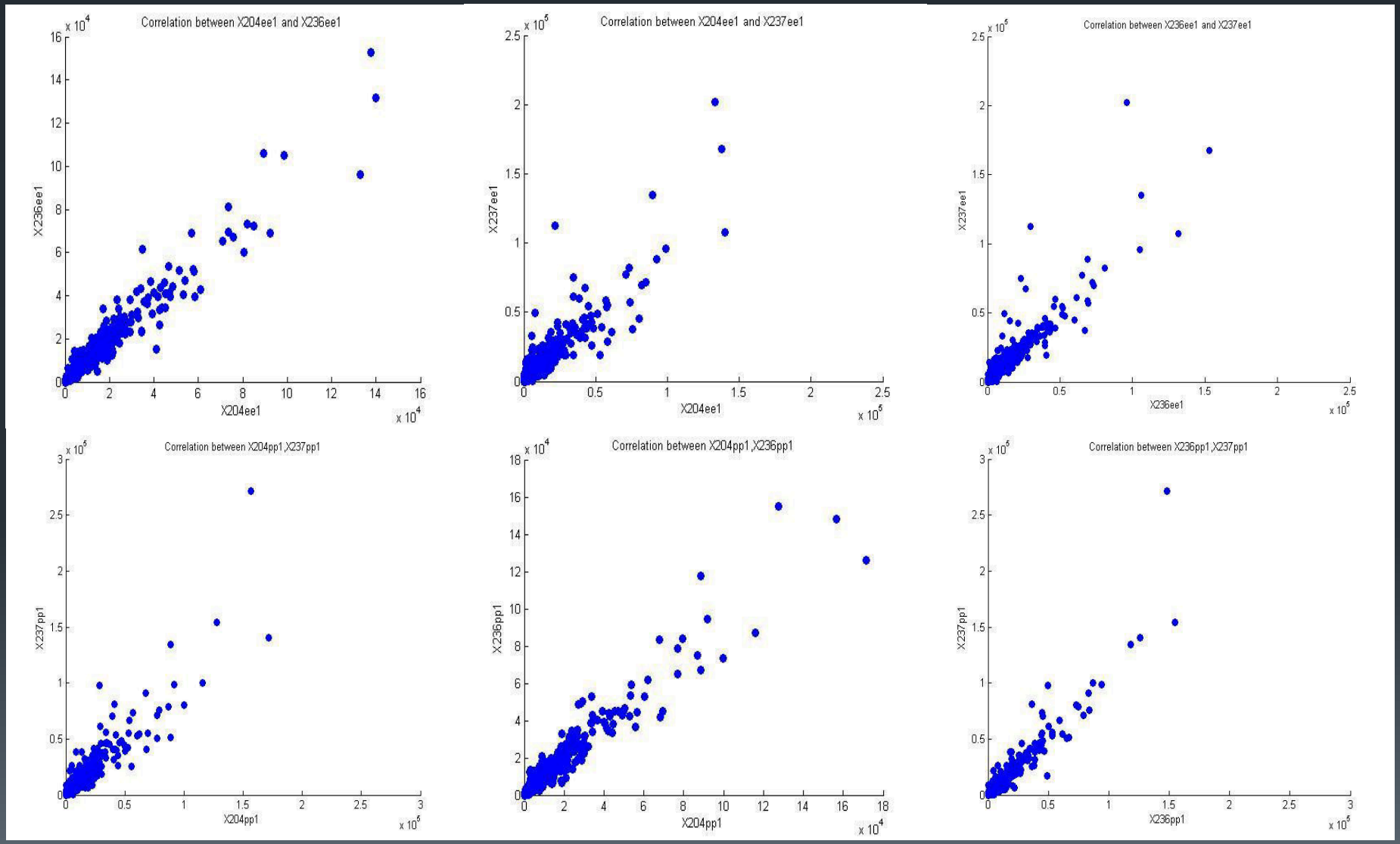
RNA-Seq of yeast [Nagalakshmi et al, 2008]

Packages for testing for differential

- Parametric
  - Based on negative-binomial distribution:
    - edgeR (Robinson, Mcarthy, Smyth)
    - DESeq (Anders, Huber)
    - BaySeq (Hardcastle, Kelly)
  - Based on Binomial distribution:
    - DEGSeq (Wang et al.)
- Non parametric
  - Cuffdiff
  - NOIseq

Performance was compared by Kvam, 2011
BaySeq was slightly better

Correlation between replicates



DEseq output

id	Base Mean	Base MeanA	Base MeanB	Fold Change	log2FoldChange	pval	padj
Gene 1	110	22	198	9	3	0	0
Gene 2	544	860	227	0	-2	0	0

Base Mean	Mean of Base MeanA and B
Base MeanA	Reads from sample A divided by the size factor
Base MeanB	Reads from sample A divided by the size factor
padj	adjusted P value for multiple testing with the Benjamini-Hochberg procedure

Tutorial

- http://galaxy.oardc.ohio-state.edu
- email : mcic@gmail.com
- pwd: glbiouse
- **Shared data:**
  - **Published Pages:** GLBIO RNA-seq Analysis Exercise
  - **Data Libraries**
    - DESeq Sample Data: to use with DEseq
    - Rnnotator - final_contigs.fa: For mapping
    - Rnnotator_Contigs_blastx_with_nr: For annotation

GLBIO_WorkFlow

MCIC Galaxy | Analyze Data | Workflow | Shared Data | Admin | Help | User

Shared Workflow | GLBIO_WorkFlow

Step	Annotation
Step 1: Input dataset **sample_2_1.fastq** *select at runtime*	sample_2_1.fastq
Step 2: Input dataset **sample_2_2.fastq** *select at runtime*	sample_2_2.fastq
Step 3: Input dataset **sample_1_1.fastq** *select at runtime*	sample_1_1.fastq
Step 4: Input dataset **sample_1_2.fastq** *select at runtime*	sample_1_2.fastq
Step 5: Input dataset **blastx_input** *select at runtime*	Rnnotator Contigs blast against nt

MCIC Galaxy

Analyze Data | Workflow | Shared Data | Admin | Help | User

Shared Workflow | GLBIO_WorkFlow

Step 6: Rnnotator

Library

Libraries 1

Library
non strand-specific paired-end library

Insert Length
300

Filename
select at runtime

Libraries 2

Library
non strand-specific paired-end library

Insert Length
300

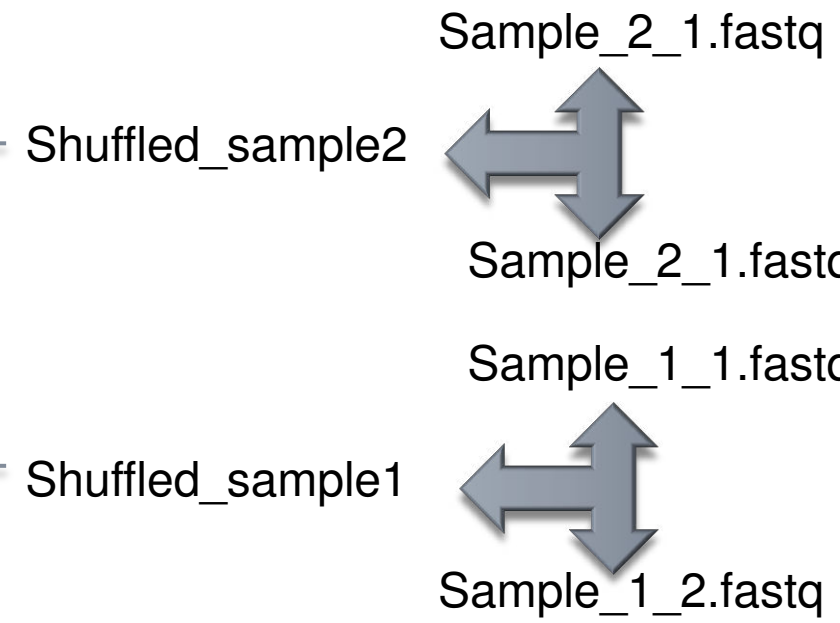
Filename
select at runtime

Use Default General Options
Yes

Use Default Read Pre-processing Options
Yes

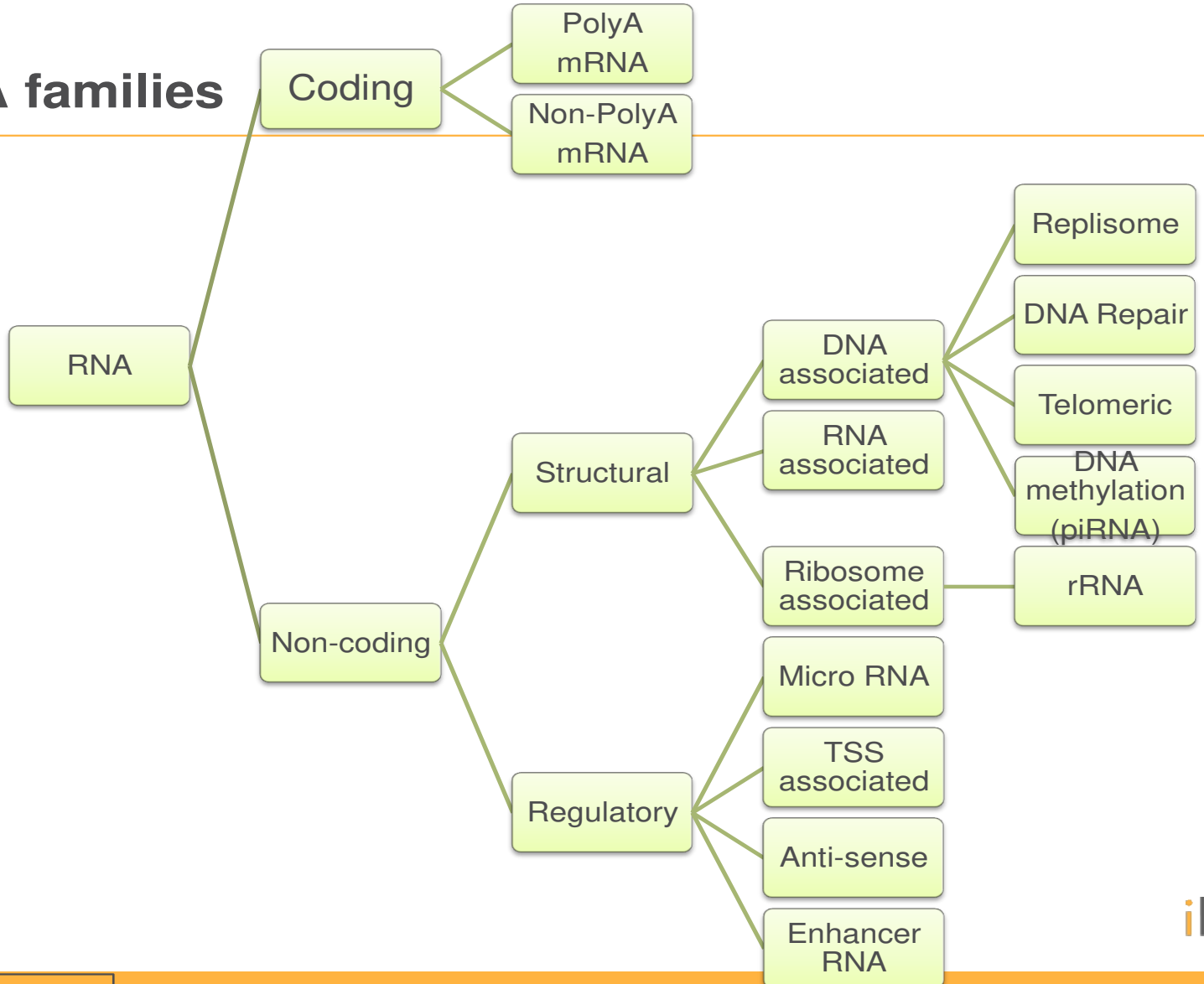
Use Default Assembly Options
No

[-a assembler] Assembler to use (velvet, oases) (default: velvet)



Thank you !!!!

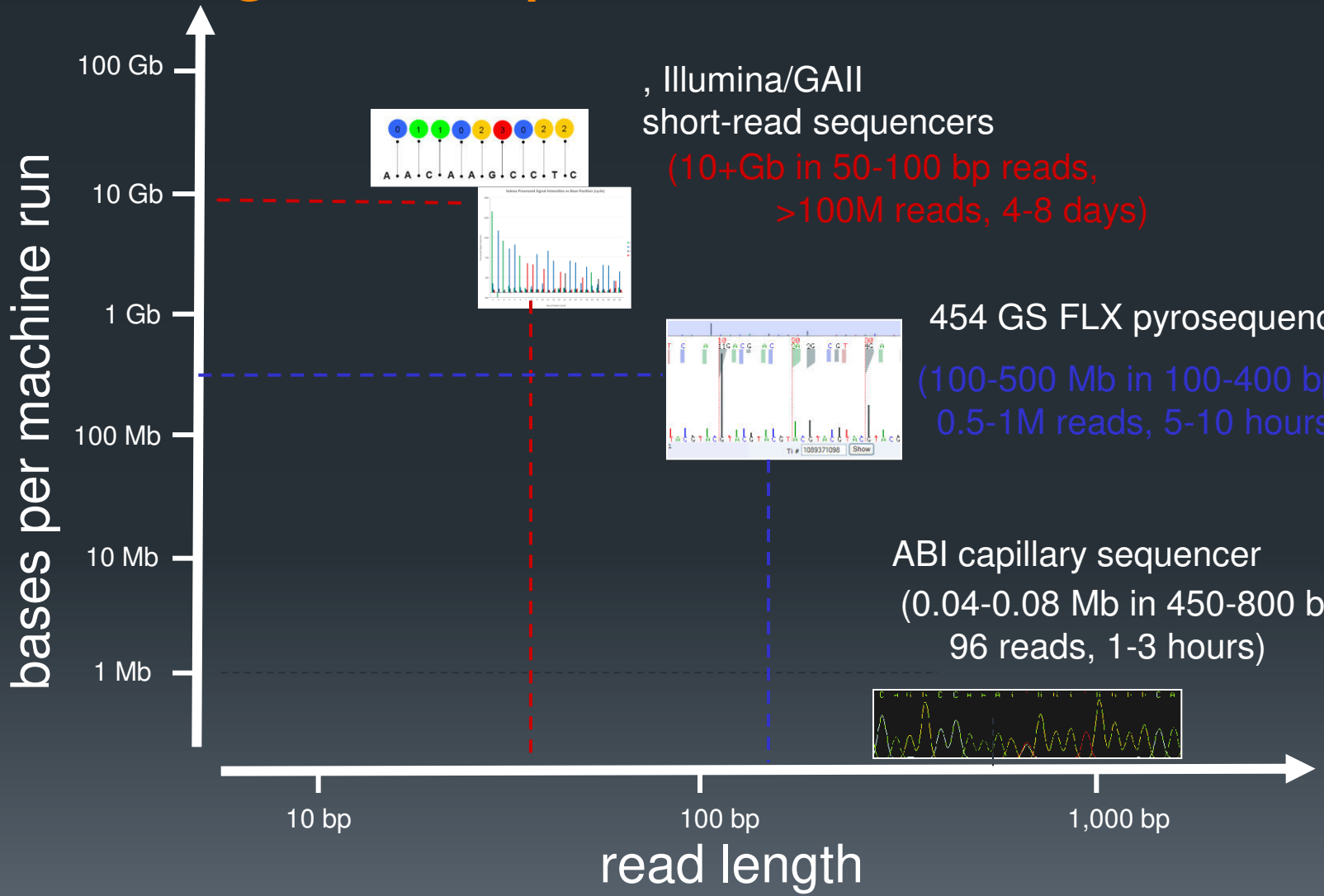
RNA families



illumina

From John McPherson, OICR

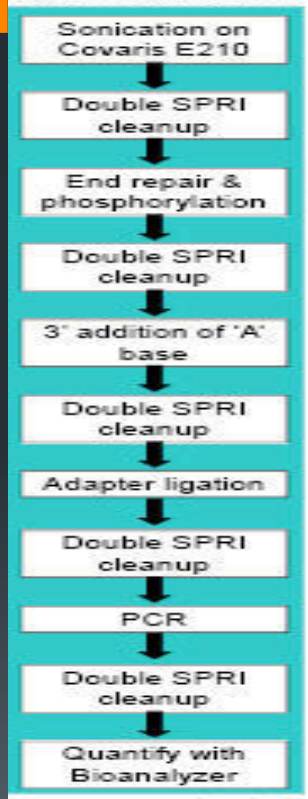
Second-gen sequencers



Au

Developments

Automated Library Construction In response to the increasing demand for constructing Illumina libraries, a semi-automated process which enables us to construct 96 Illumina libraries in approximately 6-8 hours has been developed. With a few simple modifications, the library production efficiency has doubled. The modifications include the shearing of DNA with a Covaris E210, and the cleaning of enzymatic reactions and fragment size selection with SPRI beads and a magnetic plate holder. Recently, a BioMek FX robot has been programmed to carry out the library construction process.



Workflow for automated library construction

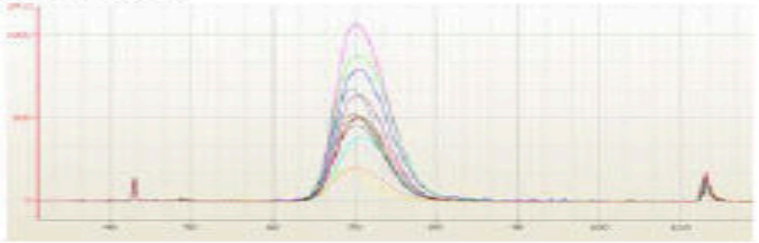
BioMek FX Robot The Beckman-Coulter Biomek FX robot is used to construct 96 Illumina libraries in parallel. This process automates the repetitive pipetting involved in library construction process and enables a single operator to construct 96 Illumina libraries in 3 days with minimal ergonomic risk.


Deck Layout of the BioMek Robot

Illumina Library Quality The efficiency, quality and reproducibility of libraries created on BioMek are currently being optimized.

Library Quantification

Using the Agilent Bioanalyzer High Sensitivity DNA chip, the quantity of libraries was assessed. A sample is required to have a concentration of at least 10 nM. The graph below shows the Bioanalyzer traces of 13 libraries constructed by the BioMek FX robot.



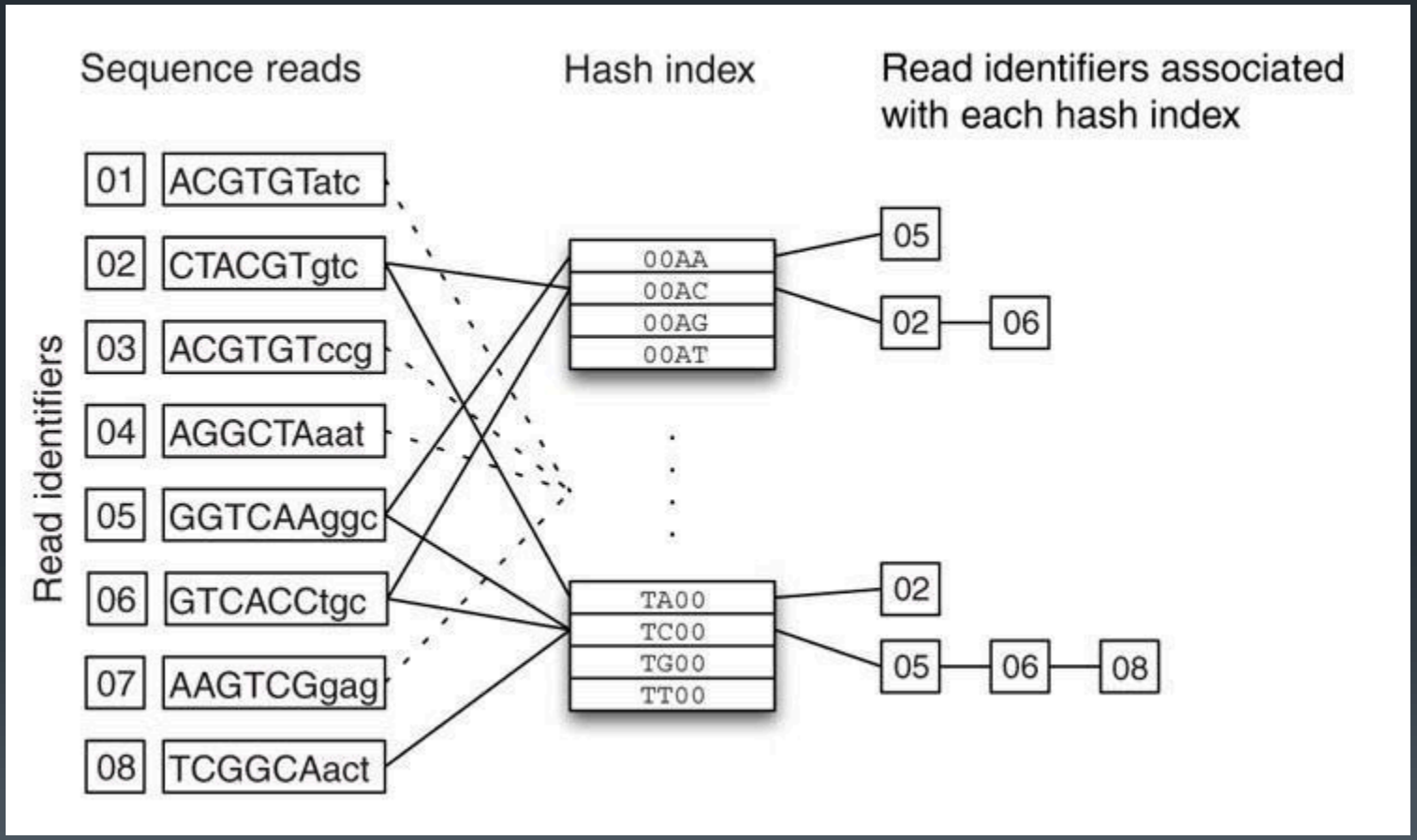
Different types of platforms

- ABI capillary sequencer (First generation)
- Current generation (Second generation)
  - Illumina
  - 454
  - AB/SOLiDv3
  - Ion torrent
- Next generation (Third generation)
  - PacBio

BWT aligners

Transformation				
Input	All Rotations	Sorting All Rows in Alphabetical Order by their first letters	Taking Last Column	Output Last Column
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	**A**NANA|^B **A**NA|^BAN **A**|^BANAN **B**ANANA|^ **N**ANA|^BA **N**A|^BANA **^**BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|**^** NANA|^B**A** NA|^BAN**A** ^BANANA| |^BANAN**A**	BNN^AA|A

http://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform



Flicek and Birney, Nature Methods **6**, S6 - S12 (2009)

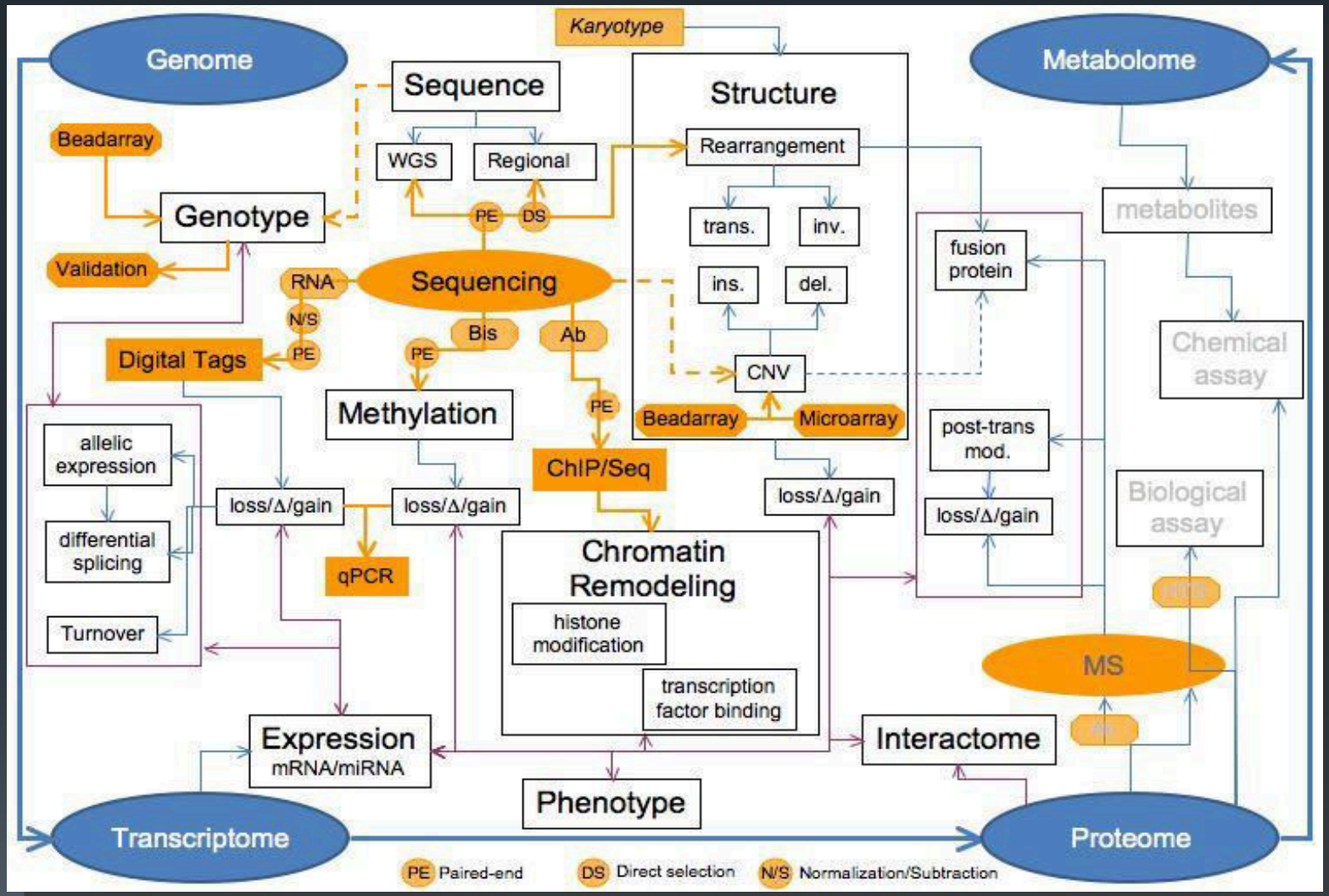
Concepts of functional annotation

- Gene/Protein function
  - Referes to the molecular function of a gene or a protein:

    Tyrosine kinase
- Functional annotation
  - More general, referes to the characterization of functional aspect of the protein.

    Stress-related, cytoplasm, ABC transporter
  - Also referes to the process of assingment of a function label
  - Habitually, standard vocabularies are used to assign function

“Omics” wolrd


Karyotype
Genome
Metabolome
Sequence
Structure
Beadarray
WGS
Regional
Rearrangement
Genotype
PE
DS
trans.
inv.
metabolites
fusion protein
Validation
RNA
Sequencing
ins.
del.
N/S
Bis
Ab
Digital Tags
PE
CNV
Chemical assay
Methylation
Beadarray
Microarray
allelic expression
ChIP/Seq
post-trans mod.
loss/Δ/gain
loss/Δ/gain
loss/Δ/gain
Biological assay
differential splicing
qPCR
Chromatin Remodeling
histone modification
Turnover
MS
transcription factor binding
Expression
mRNA/miRNA
Interactome
Phenotype
Transcriptome
Proteome
PE Paired-end
DS Direct selection
N/S Normalization/Subtraction

- v of count values is modelled as $v = s\mu + \alpha s2\mu2$,
- where μ is the expected normalized count value (estimated by the average normalized count value), s is the size factor for the sample under consideration, and α is the dispersion value for the gene under consideration.