UNSW AUSTRALIA

Never Stand Still

Introduction to RNA-Seq

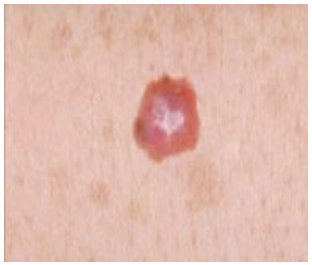
Susan Corley
Systems Biology Initiative
11 November 2014

SYSTEMS BIOLOGY INITIATIVE

RNA-Seq

RNA-Seq

Koala | Melanoma | Wheat | Soil Biota | Corals


Ramaciotti Center for Genomics

RNA Extracted

↓

Quantification

↓

Comparison between conditions

RNA-Seq advantages

- Dynamic gene expression captured in a snapshot (up or down-regulation in a particular condition/s)
- Information about alternative splicing – which isoforms are being used
- Can detect lowly expressed -> very highly expressed genes (large dynamic range)
- Prior knowledge not necessary – can predict novel transcripts

RNA-Seq experiment

Experimental design

Library preparation

Sequencing

Bioinformatic analysis

Validation

RNA-Seq experiment

Experimental design

Library preparation

Sequencing

Bioinformatic analysis

Validation

Experimental design

1. Apply the same design principles that you would to a laboratory experiment e.g. appropriate controls, and biological replicates
2. Make decisions about number of replicates and depth of coverage
3. If all processing cannot be done in one run design the split carefully so that you can adjust for batch effects
4. Get advice before committing to the sequencing

*In the absence of proper design, it is essentially impossible to partition biological variation from technical variation …No amount of statistical sophistication can separate confounded factors **after** data have been collected.*

Auer and Doerge "Statistical Design and Analysis of RNA Sequencing Data" *Genetics* (2010)

Replicates vs depth

20 M reads/sample	30 M reads/sample	40 M reads/sample	50 M reads/sample

300-400 M PE reads/lane
Fixed price per lane ~ $3200
Library prep $450/sample

Liu et al. (2014) Bioinformatics

MCF-7 cells:
treated with E2 (7) control (7)
Illumina HiSeq: > 30 M reads per sample, 50 bp
edgeR for differential expression

Analyzed using:
different number of replicates
different number of reads

DEGs increase with replicates and read depth
But diminishing return > 10 M reads

Replication adds significant power

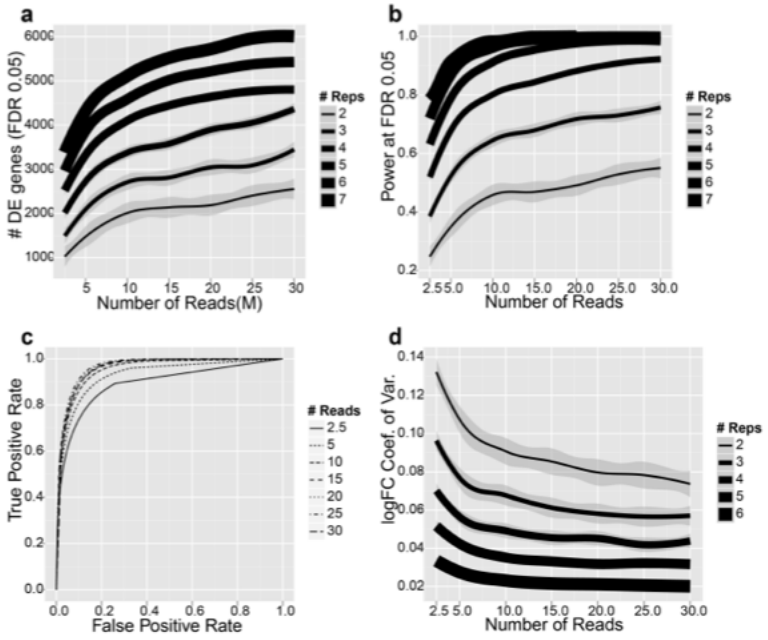
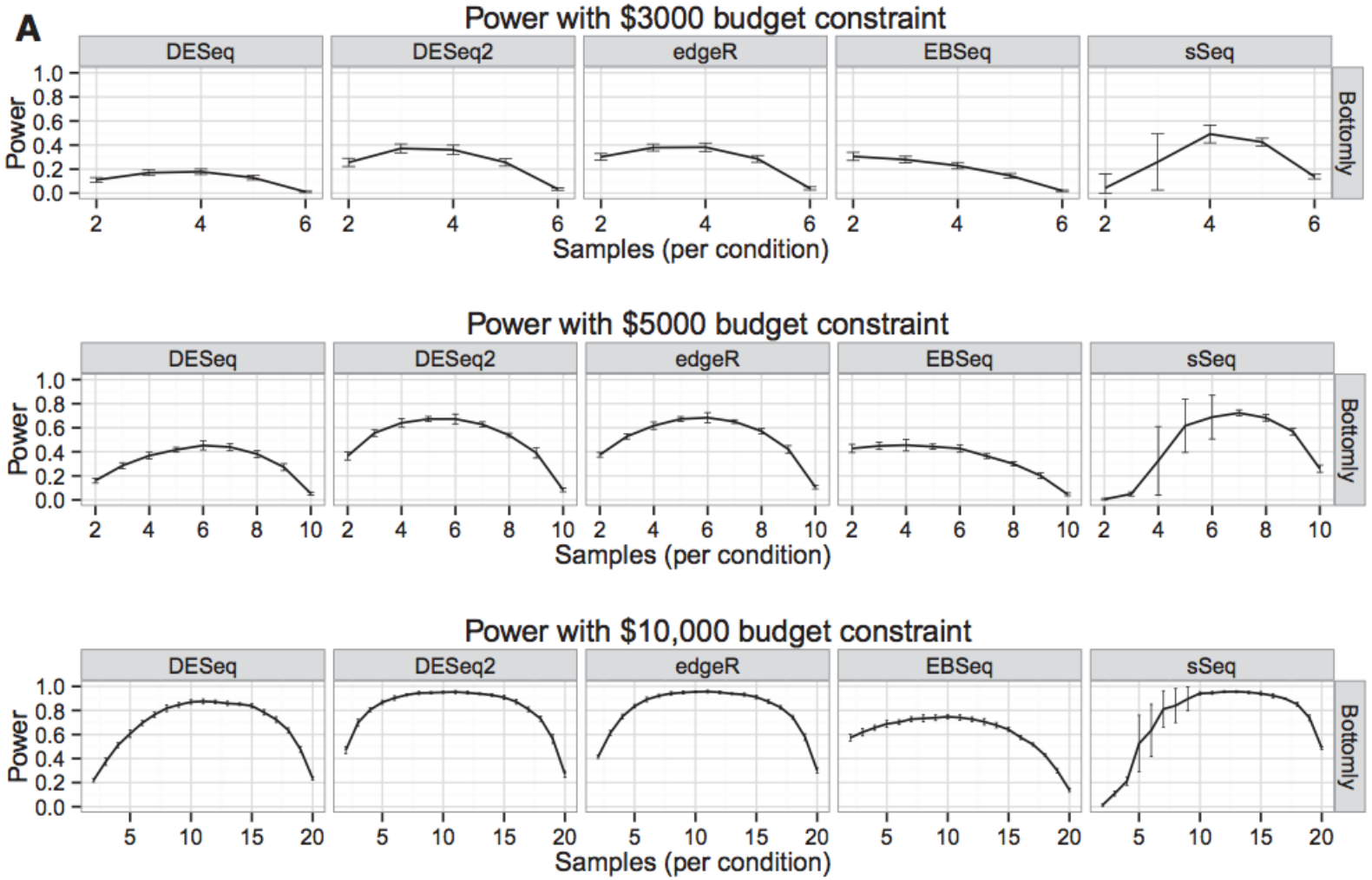


Fig. 1. (a) Increase in number of biological replication significantly increases the number of DE genes identified, while number of sequencing reads have diminishing return after 10M reads. Different color indicates different number of replication, with 2 replicate the darkest and 7 replicate lightest. The lines are smoothed average line of each replication level, with the shade corresponding to 95% confidence interval of the mean number of DE genes. (b) Power of detecting DE genes increases with both sequencing depth and biological replication level. Similar to the trends in (a), the power increases after 10M become smaller. (c) ROC curve for 3 biological replicates. Sequencing deeper than 10M reads does not significantly improve statistical power and precision for detecting DE genes. (d) The coefficient of variation (CV) of logFC for the top 100 differentially expressed genes. The CV of the logFC estimates decreases significantly as we add more biological replicates, while adding sequencing depth after 10M reads has much less effect.

Ching et al. (2014) RNA



Replicates vs Depth

- Answer depends on experiment being undertaken
  - Per-condition variance
  - Degree of difference between conditions
- For DE gene expression I would recommend 20 M reads per sample and $\geq$ 5 replicates
- For lowly expressed transcripts such as ncRNA greater read depth

RNA-Seq experiment

Experimental design

Library preparation

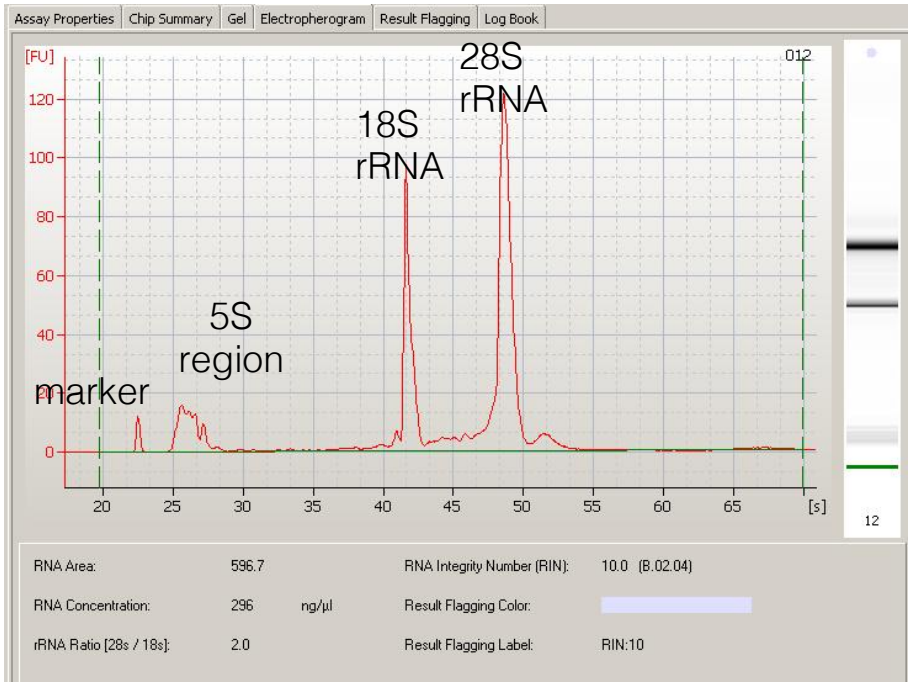
Sequencing

Bioinformatic analysis

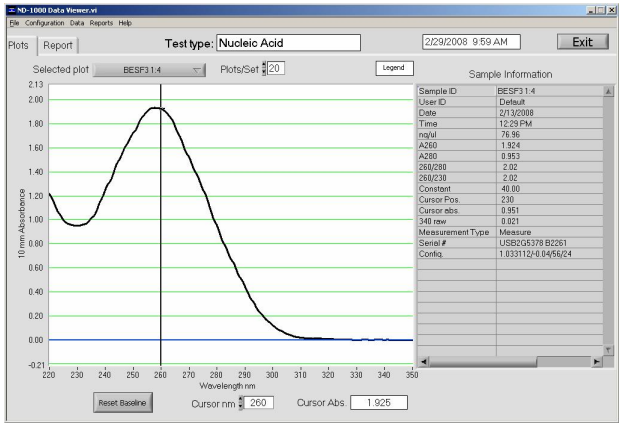
Validation

Library preparation

- Extraction of total RNA
- Tissue preparation – shearing, freezing etc. may result in degradation
- RNA quality check
- 28S/18S ~ 2
- RIN – RNA Integrity Number
- NanoDrop spectrophotometer - Concentration and Purity of RNA
- Advice from sequencing facility



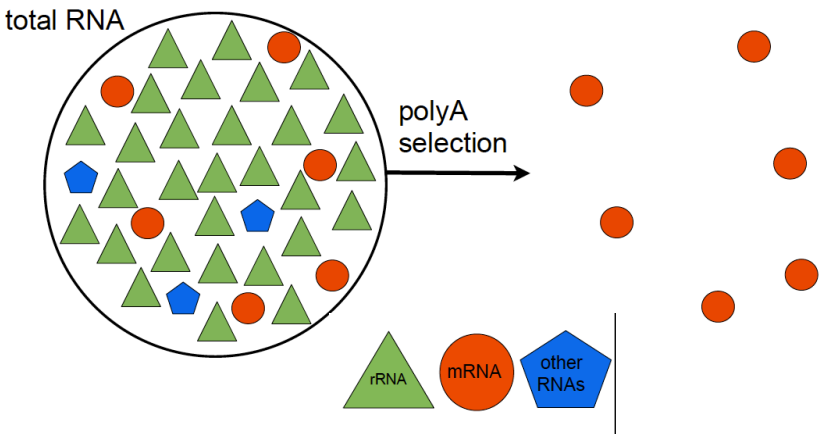
Bioanalyzer electropherogram



NanoDrop

Library preparation

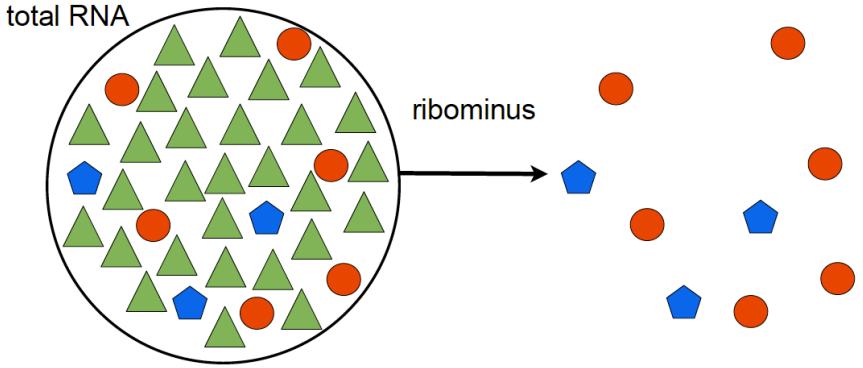
polyA selection



poly(A+)-transcripts:

- mRNAs,
- some lncRNAs
- immature microRNAs,
- snoRNAs

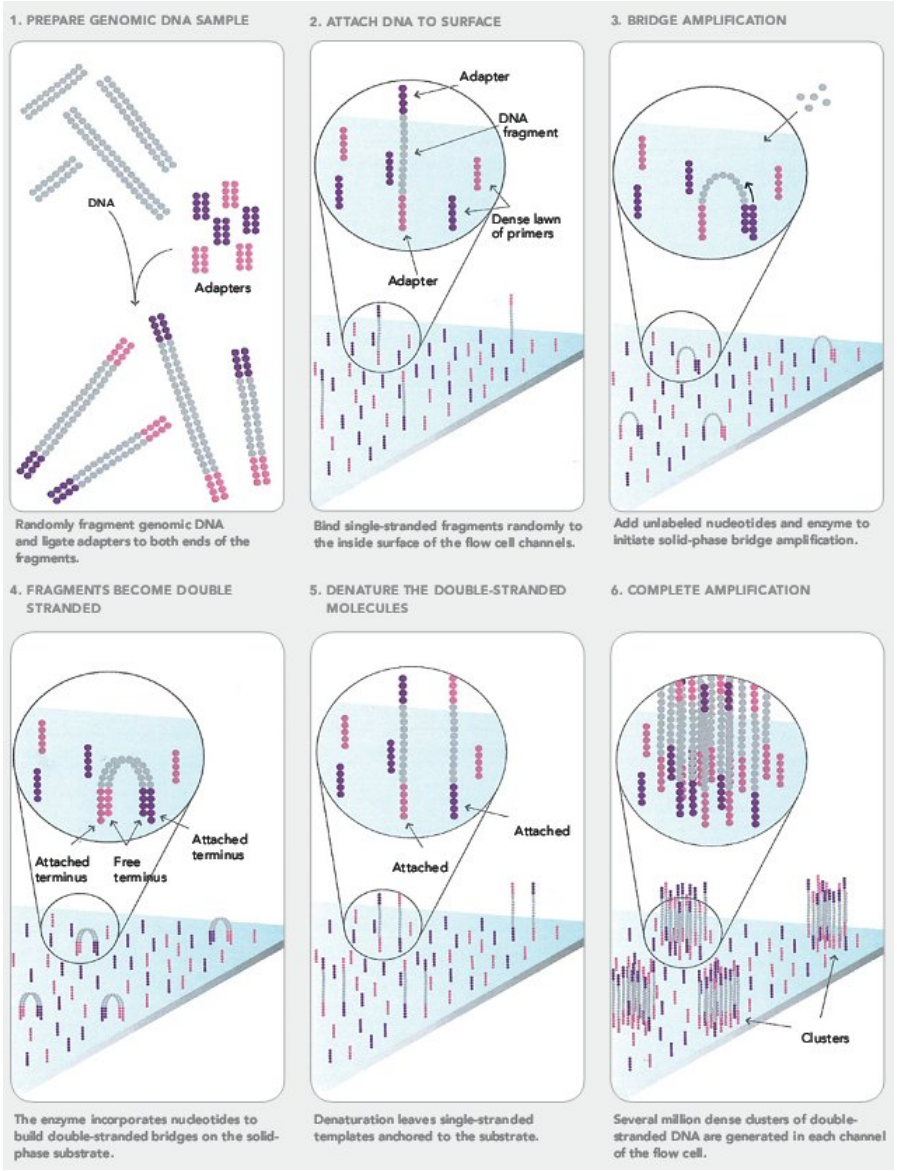
ribominus selection



non rRNA transcripts:

- Non- coding RNAs
- mRNAs,

Library preparation



Illumina

RNA-Seq experiment

Experimental design

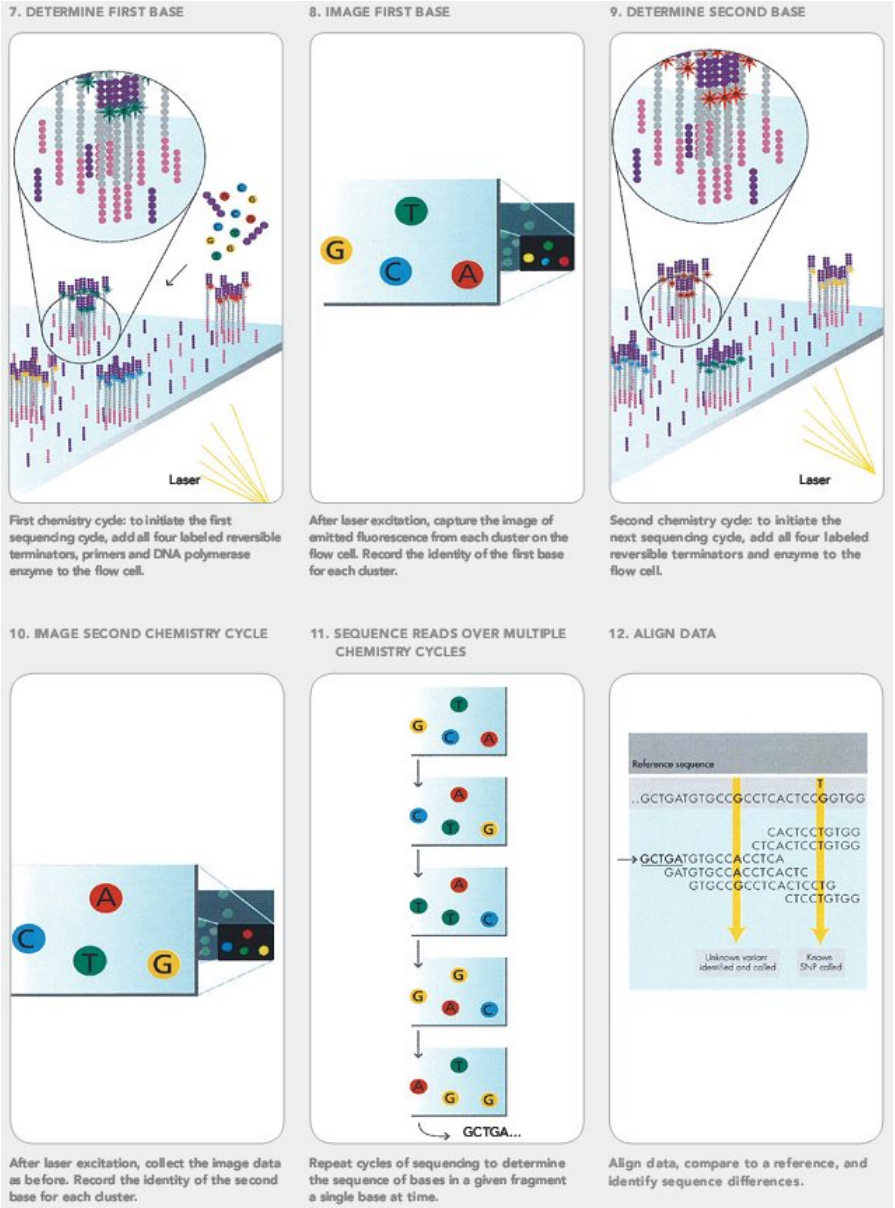
Library preparation

Sequencing

Bioinformatic analysis

Validation

Sequencing


7. DETERMINE FIRST BASE
8. IMAGE FIRST BASE
9. DETERMINE SECOND BASE
Laser
T
G
C
A
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
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.
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
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
12. ALIGN DATA
A
C
T
G
GCTGA...
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
Align data, compare to a reference, and identify sequence differences.

Sequencing

Fastq files

```
@HWI-ST1213:141:C17PWACXX:2:1101:1245:1968 1:N:0:ATGTCA
CTTGCATATATCCGGTCTTAATCCTCTGAGGTGATGAGACACTGCTCATGTGACCTTCTACCTTTGTGGATCCCTCTTGGTTTGTGACCTTCGTACCTTCC
+
@@@DDDA4C?FF>+12@CHHEHCEFEFI@BG1)?@FBAF@)?DFFIIDFI?B9F<F<CFC4=D=@@@F7;;.=?CACAD@.(6;@AA(>;;;=;?######
@HWI-ST1213:141:C17PWACXX:2:1101:1138:1997 1:N:0:ATGTCA
CTGAAAAGCAGCATGTCAGGGTATATCCCTAGTTACCTGGACAAAGACGAGCAGTGTGTCGTGTGTGGGGACAAGGCCACCGGTTATCACTACCGCTGTAT
+
CC  GHHGCFGFHIHJAFFJGGIJJJJHIIGIJIEGIJJIJJJJIIJJJBFGFHIICBBEHGFFD=BBDDBB@BDD?BB@>CCCCCCACBDDD@:>
@HW  :141:C17PWACXX:2:1101:1166:1997 1:N:0:ATGTCA
CCGAC  GTCTGCGCCGGGTGCTGGATGAGCTGACACTGGCCAGGACCGACCTGGAGATGCAGATTGAGCAGCTGAACGAGGAGCTGGCTTAACTG
+
@<?AD<L  <FFEFAGFFGF:FFDD94B8B9@>FE)=77=AEFFDD?B<?38?B959?>@BBB@@@(444:9?(:>3>?-095<@<9??#######
@HWI-ST  :C17PWACXX:2:1101:1321:1979 1:N:0:ATGTCA
TTTACTCC  ATGATGGCGAAGTGGGCTTTTGCTCATGTGTCATCTAGGGTGAAGCCTGAAAATAATGGGAATCAGTGAACAAATCCTGCTATGAT
+
CCCFFFFFH  JJJJJJJJJJIFHIIIIJJHHIJJJJJIJJJJJJIIJJGIHHGJGGDHEHHHHGFFFFDD@CBEED>CDDDDDDACCDDDDDEDC>
@HWI-ST  PWACXX:2:1101:1739:1993 1:N:0:ATGTCA
CGGATTCTA  AGATGCTGGGTAGACAGCAGGGTATACTCAGACTAGGACAGAGGAAAGCCTTAGCAGCGAGATCAGGGTGACTTGTTCCATCCTC
+
=?@DA1BDBDFBF+AEG@E<CHHB?FCE?DFFA?;C@A0?*???DAF<?FB;DEF>FFFCFFIBCE;>EFFFDDD>?35;>@3?,,(:>5:AAB:ABABB>
@HWI-ST1213:141:C17PWACXX:2:1101:1596:1998 1:N:0:ATGTCA
CTCCACTATGCTCATAGCAGCCTTATTTTTAATAGCCAGAAGCTGGAAAGAACCCAGATGCCCCTCAACAGAAGAATGGATACAGAAAATGAGGTATATTT
+
CCCFFFFFHHHHHJJIJJJIJJJJJJJJJJJJJJIGHHIJJJIJJJJIIJJJIJJJJJJJJIJJJIIIIIHHEHHFCDEFECDEECDDDDCDDDDCACCDEC
@HWI-ST1213:141:C17PWACXX:2:1101:1891:1948 1:N:0:ATGTCA
NACCAGACTGCCATAAAGAAGAATAACCCCAGGAAGTACCTTCGCAGTGTAGGCGATGGAGAGACTGTGGAGTTTGATGTTGTTGAAGGAGAAAAGGGTGC
+
#1=DFFFFHHHHHIJJJJJJJJJJJJJJIJJJJJIJJFHIJJJJJJJJHJIJJIIIIIJJJHHHHFFFFCEE@C6>AACCDDDDDBDCCCDDDAA?CDDD8?8
@HWI-ST1213:141:C17PWACXX:2:1101:1999:1954 1:N:0:ATGTCA
CTTTTCTGGGGTCTGATGAGCGTCGGCATCGGGCGCCTTAACCCGGCGTTCGGTTCATCCCGCAGCGCCAGTTCTGCTCACCAAAAGTGGCCCACTAGGCA
+
CCCFFFFFHHHHFHGIGJJJJJJJJJJJJJJJJJJIJFGFFHHHFFFDDD=BDDDBBDDEDDDBDDDB@DDDBCDDECCDCDDDDD?BDCAC@?A@B@?CA?@
@HWI-ST1213:141:C17PWACXX:2:1101:1846:1958 1:N:0:ATGTCA
CTTGTGCTACCAATGCTGTGACCTTCAATCTTAATCCCTCTCTCCTTGCACCACCTAAAAAGTTGACTCTCATCCTCATCTTCAAGGGCTCCTGAGCCTCT
+
CCCFFFFFHHHHHIJJJJIJJJJJJJIJJJJJJJJJJJJJJJJJJJJJJIIHEIJJGEJIJGIJJJJIIGHIJJJEHEEFGFFFFFFCDEDDDCCDDDCDC
@HWI-ST1213:141:C17PWACXX:2:1101:1993:1989 1:N:0:ATGTCA
CTCCACTCTCTGACCACTGGCCGCCTCCCGAGAGTTGTTACAGCTAATCGGATGCTGAAGCAGATGCTCTTCAGGTATCAAGGTTACATTGGTGCAGCCCT
+
CCCFFFFFHHHHHJJJJJIIJJJJJJJJJIJJIIGFHIJJIJIIJGIIIIJJIDHFHHFFFFFCECCCCCDD@AACCCDCCAA@CCDCCD>CA?ACCCA@C
@HWI-ST1213:141:C17PWACXX:2:1101:2110:1960 1:N:0:ATGTCA
CCACCCTGGGTGAACATCTGGCTGCTGGGTTCCATCTGCCTGTCCATGTCCCTCCACTTCCTCATCCTCTATGTCGACCCCCTGCCGATGATCTTCAAGCT
+
CCCFFFFFHHHHHDGIJJJJJIJJJJJJJJJGIGIIJIIIIJJJJJJJJIJJIJJJJIIJJJIJJJJJJJJJJJHHHHHHDFDDDDDBDDDDDDDEDECEDCDDD
@HWI-ST1213:141:C17PWACXX:2:1101:2227:1985 1:N:0:ATGTCA
TATAAGCCTCACGGTTGGGAAGGTCAAACTTGCTGACATTGGTCTTTGCCAGGATGGTCTTGAAAATATAGAATTTATCAGGATCTTCCACAATGTCCTTA
+
CCCFFFFFGHHHHIIJHJJGIIJFHIJJJIJJJJJJJJJJJJJGIJJIIHIJGIJJJIHIGIJGIIGEHFHHHDDF@DFFEEDEEDDDDD@CCBDCDDDDDC
@HWI-ST1213:141:C17PWACXX:2:1101:2191:1999 1:N:0:ATGTCA
CCCACCCCATATTAAACCCGAATGATATTTCCTATTTGCATACGCCATTCTACGCTCAATCCCCAATAAACTAGGAGGTGTCCTAGCCTTAATCTTATCTA
+
CCCFFFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJIJJJIJJIJIJJHHFFEFFFFFECECED?BDDDDCDCCDDDDEDDDCCCDD
@HWI-ST1213:141:C17PWACXX:2:1101:2462:1952 1:N:0:ATGTCA
```

RNA-Seq experiment

Experimental design

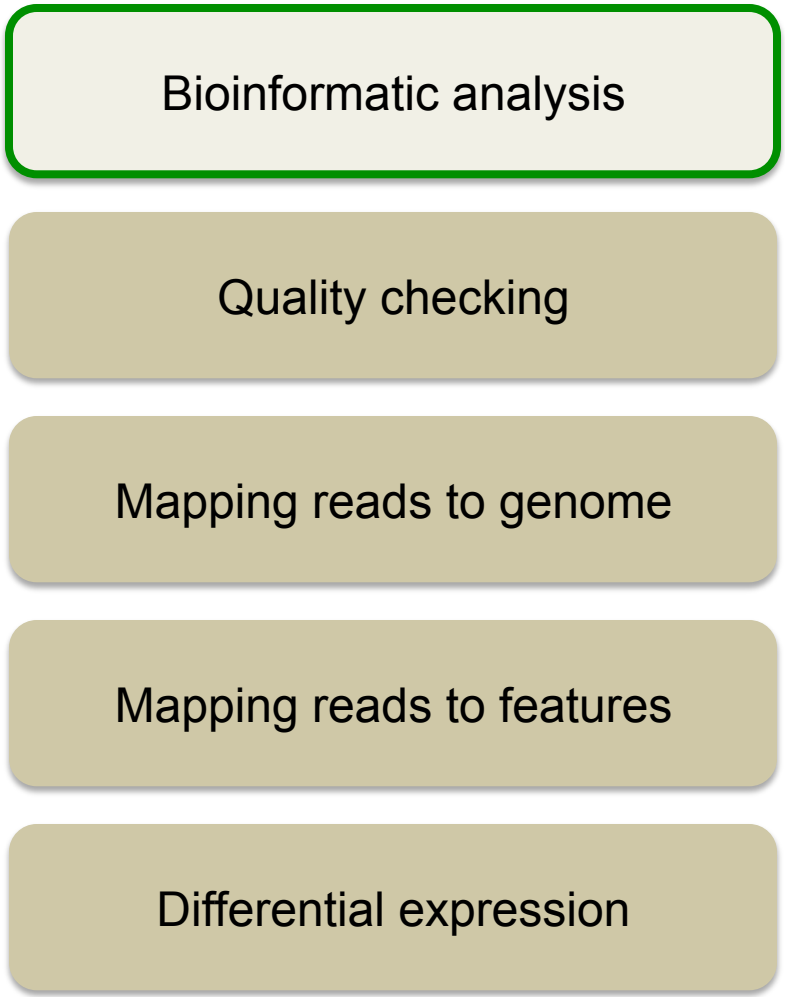
Library preparation

Sequencing

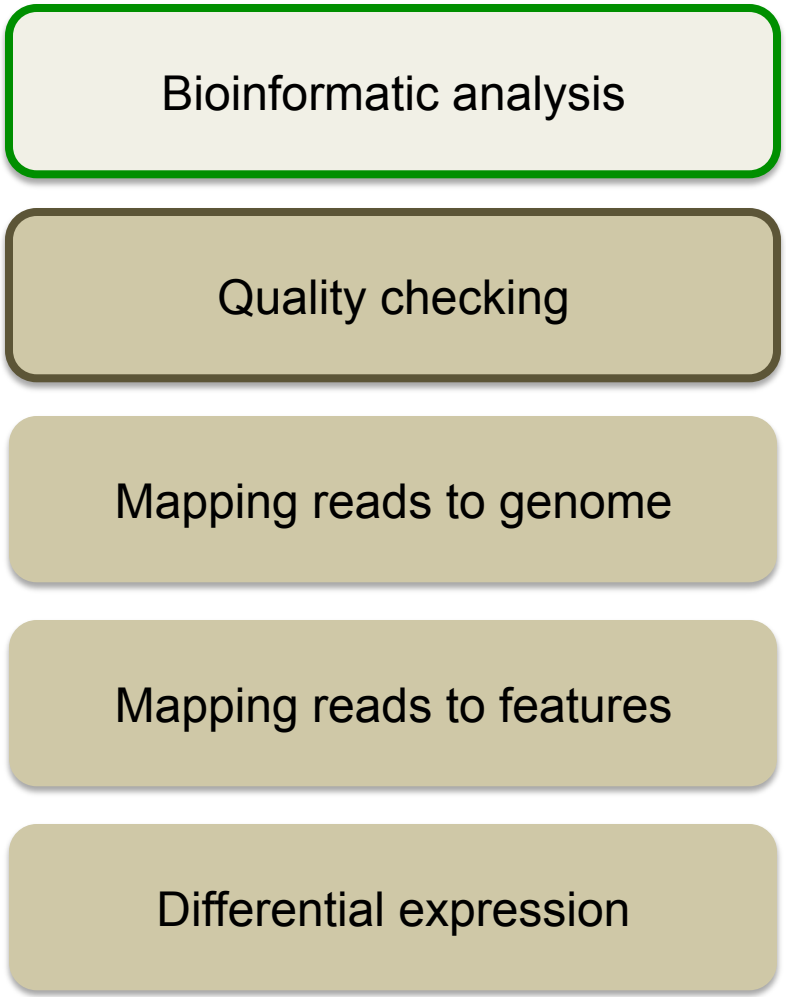
Bioinformatic analysis

Validation

RNA-Seq experiment

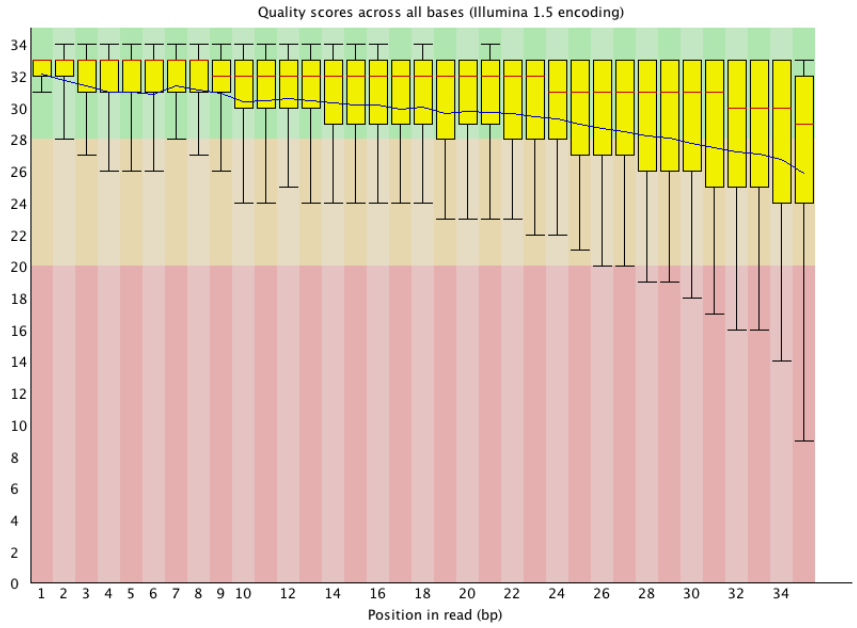


RNA-Seq experiment

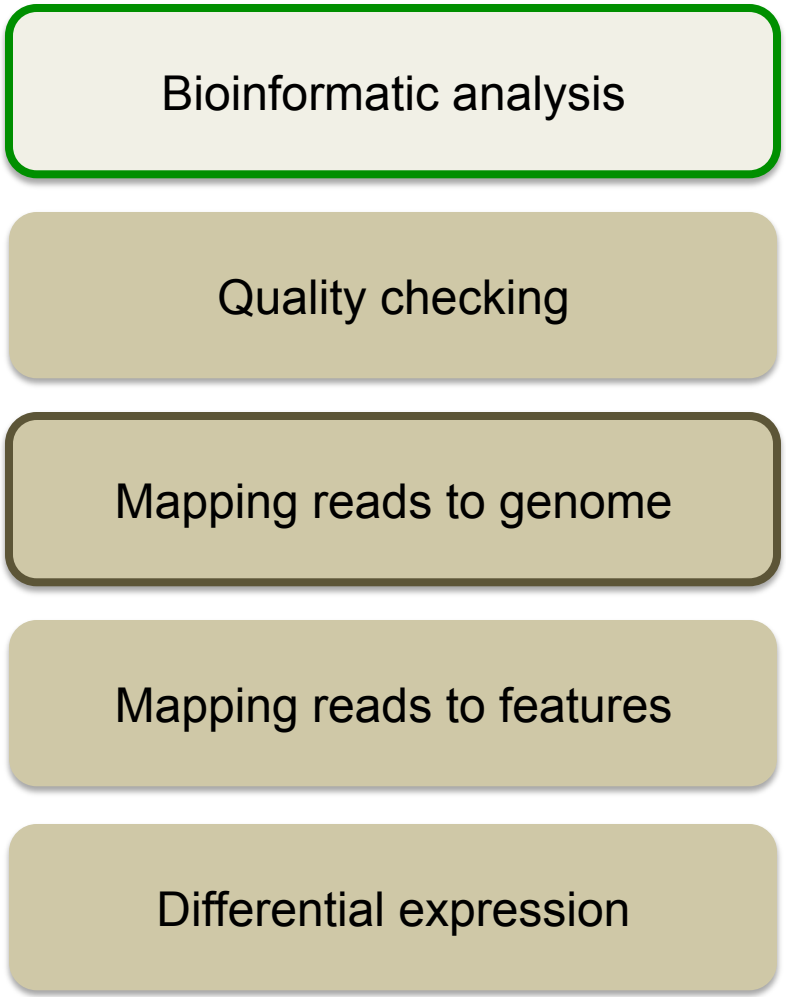


Quality checking

- Covered yesterday
- Fastqc tool
- Generally unnecessary to trim unless doing de novo assembly

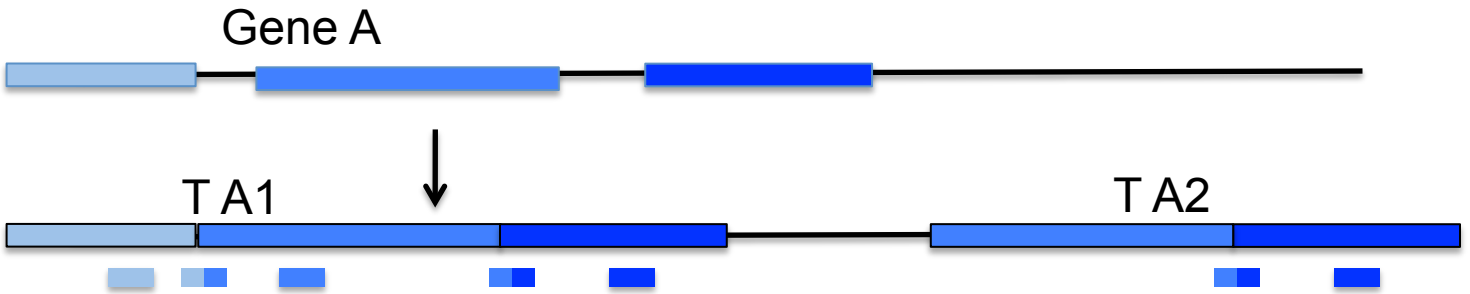


RNA-Seq experiment

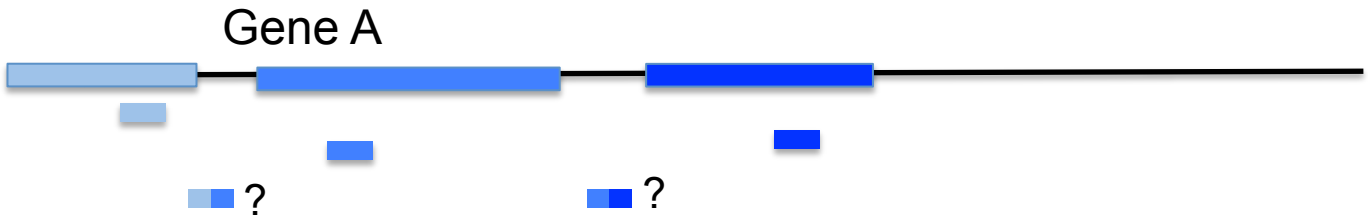


Alignment of reads to genome

- Mapping RNA reads back to a genome is tricky



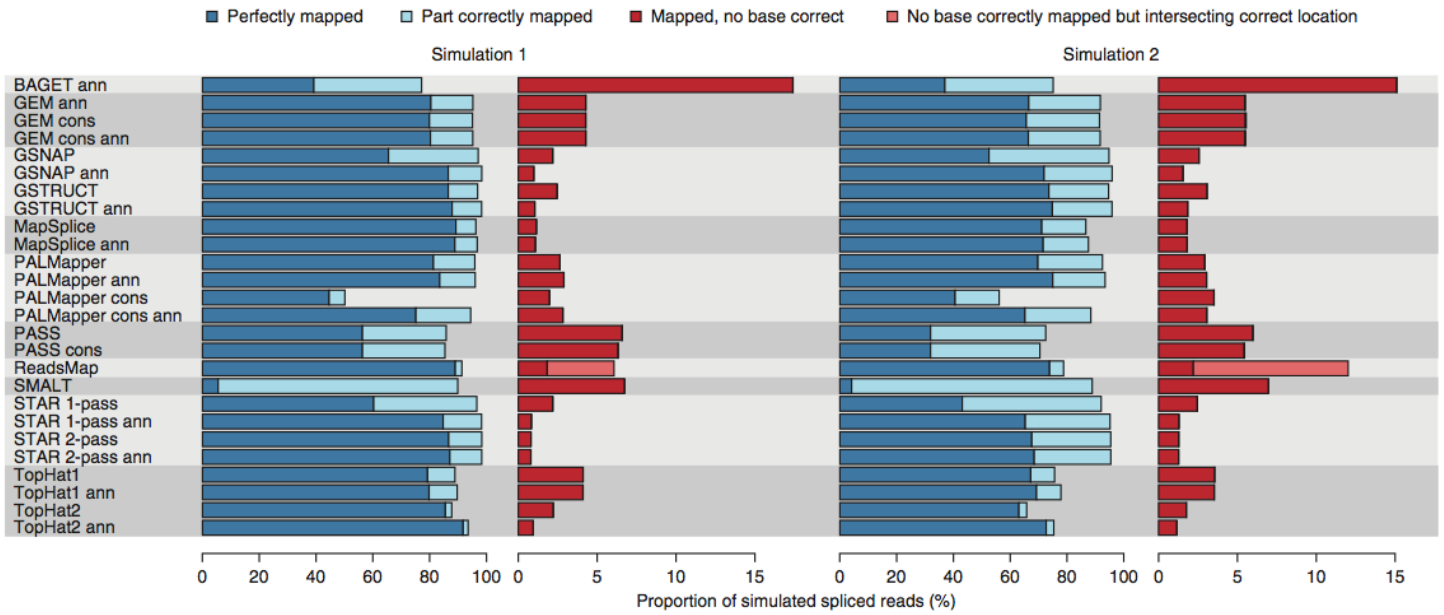
Mapping RNA-seq reads back to genome



Alignment of reads to genome

- Must use a splice aligner

- Tophat2
- GSNAP
- MapSplice
- PALMapper
- ReadsMap
- Star



Engstrom et al., *Nature Methods* (2013)

Alignment to genome

TopHat
A spliced read mapper for RNA-Seq

McKUSICK-NATHANS Institute of Genetic Medicine

Manual

- What is TopHat?
- Prerequisites
- Using TopHat

What is TopHat?

TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat runs on **Linux** and **OS X**.

Site Map

Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol

News and updates

New releases and related tools will be announced through the Bowtie **mailing**

```
Tophat –solexa-quals –g 2 –library-type fr-unstranded –j annotation/Danio_rerio.Zv9.66.spliceSites –o tophat/SV9_6h
      genome/ZV data/6h_1.fastq data/6h_2.fastq
```

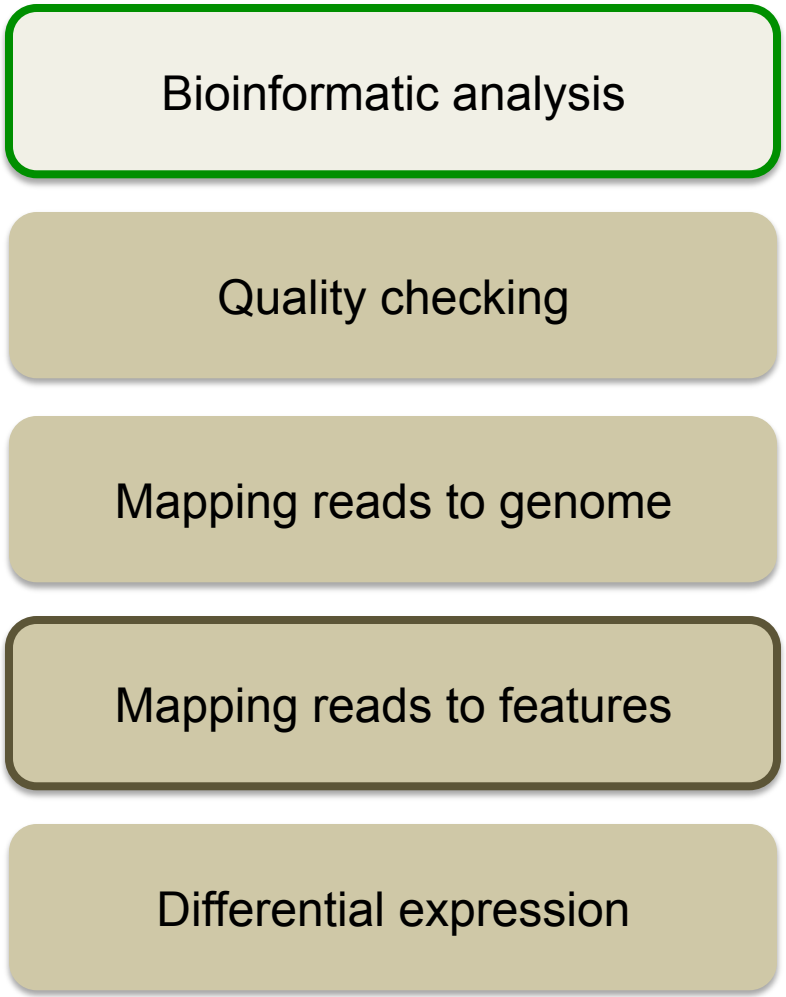
BAM/SAM file

```
HWI-ST1213:141:C17PWACXX:2:1101:1024:7546       99      chr7    132714204       50      101M    =       132714239       136
NTCAAGGTCTTTCTTTAGGAGAAAGGTAACAGCCATTTGGGCTCAGCCAGCAGACAGTGAAATCTCAGAGGGTAGGCACAGCAGAGCAACTACAAATGTGN   #1=DDDDBFDDHHIGIIICHIIGIII<FFHIIIIIIIIHAEFHIIEHEF?
DHHGGGGCGHIIGE@HGHE@AEHEED@>BECCECB2>>AAC:@ACCCC@@#     AS:i:-2 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:0A99C0     YT:Z:UU NH:i:1
HWI-ST1213:141:C17PWACXX:2:1101:1024:7546       147     chr7    132714239       50      101M    =       132714204       -136
NTTGGCCTCAGCCAGCAGACAGTGAAATCTCAGAGGGTAGGCACAGCAGAGCAACTACAAATGTGCCTAGATCACATGTCCGCAAACCCAAGAGAATGCAC   ##A?<(<B@DDCACADDEACEC@@DCDFFDCCHEHECHAGCGIHGEGIHFCGB?FB?
EIIIGHGJIGIJJIIHEJJJIGJJJJIHGGCFHFHHDFFDD@@B    AS:i:-3 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:0T4G95     YT:Z:UU NH:i:1
HWI-ST1213:141:C17PWACXX:2:1101:1024:34522      99      chr4    133601041       50      101M    =       133601148       208
NGGAGCCTCACCACCCTCTTCCCCAGCACTGTCGGCTGTCCACAGCGTCAGGTTGTCTCTCAGGAGCTGCATGATGAGGGTGCTGTCCTTGTAGGAGTCCN   #1=DDDDDDFD?DECGHGECEAEDCHGFHHI?FHGIGG?9??BBFGHS@F;D7=?=CEA?
DDE9@A;ACC?CC;A@>(:8(58@@5:>CCC(4:3:93?##      AS:i:-2 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:0C99T0     YT:Z:UU NH:i:1
HWI-ST1213:141:C17PWACXX:2:1101:1024:34522      147     chr4    133601148       50      101M    =       133601041       -208
NGGTGTGCAGGTCGGCCATGGCCTCGTCGAAGGTGGTCTTGGCCAGCGAGATGGCCTCCTCGGGGCTGTTGGCTATCTCGTAGTGGAAGACTGAAAAGTTC   ###@C?A?55->@A?>:>DB<505@DBAAC<BDDCCEDCAB?
551EC@GC=CCC:GDHEIHGHF9GJHG@HGGHGBA<HIIHGGG@AHDBFHFDFFDD@;B    AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0G100      YT:Z:UU NH:i:1
HWI-ST1213:141:C17PWACXX:2:1101:1025:18784      137     chrX    102186438       50      101M    *       0       0
AATCAATCTGAATGGTGTCGTTCACCTTGATGAGGGGATCAGGGTACCGAATAGTACGAGCATCATGGGTCACCAGGTGCGGGATTCCTTTTGTGCCTACN   @B@FFFFFH>FDF<B:A<FHIIDHE>FFEBFGCHJGGIIJI9?
BBFHGDBEAFH@FAA@'5:@@DE3;ACEECDAA?5;>BBD@<>:BCC>>AAC>ACC##     AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:100A0      YT:Z:UU NH:i:1
```

SYSTEMS BIOLOGY INITIATIVE

UNSW AUSTRALIA

RNA-Seq experiment



Mapping reads to features– transcript assembly

BAM file
Genomic location of read

Genomic Features GFF /GTF files
Gene, non-coding RNA etc

Homo_sapiens GRCh37.66_chr.gtf

chromosome	feature	start	end	score	strand	frame	attribute
chr10	miRNA	exon	77312216	77312311	.	+	.	gene_id "ENSG00000207583"; transcript_id "ENST00000384851"; exon_number "1"; gene_name "MIR606"; gene_biotype "miRNA"; transcript_name "MIR606-201";
chr10	sense_intronic	exon	77502498	77502641	.	+	.	gene_id "ENSG00000228280"; transcript_id "ENST00000445111"; exon_number "1"; gene_name "RP11-367B6.2.1"; gene_biotype "sense_intronic"; transcript_name "RP11-367B6.2.1-001";
chr10	sense_intronic	exon	77503224	77503513	.	+	.	gene_id "ENSG00000228280"; transcript_id "ENST00000445111"; exon_number "2"; gene_name "RP11-367B6.2.1"; gene_biotype "sense_intronic"; transcript_name "RP11-367B6.2.1-001";
chr10	miRNA	exon	77583444	77583536	.	-	.	gene_id "ENSG00000215921"; transcript_id "ENST00000401102"; exon_number "1"; gene_name "AL731568.1"; gene_biotype "miRNA"; transcript_name "AL731568.1-201";
chr10	miRNA	exon	77887009	77887069	.	-	.	gene_id "ENSG00000221232"; transcript_id "ENST00000408305"; exon_number "1"; gene_name "AC012047.1"; gene_biotype "miRNA"; transcript_name "AC012047.1-201";
chr10	snRNA	exon	78020558	78020661	.	-	.	gene_id "ENSG00000201954"; transcript_id "ENST00000365084"; exon_number "1"; gene_name "U6"; gene_biotype "snRNA"; transcript_name "U6.314-201";

Cufflinks gtf

chr1	Cufflinks	exon	32167722	32167828	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014371"; exon_number "66"; gene_name **"COL16A1"**; oId "CUFF.686.19"; nearest_ref "ENST00000271069"; **class_code "j"**; tss_id "TSS7101";
chr1	Cufflinks	exon	32169438	32169920	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014371"; exon_number "67"; gene_name **"COL16A1"**; oId "CUFF.686.19"; nearest_ref "ENST00000271069"; **class_code "j"**; tss_id "TSS7101";
chr1	Cufflinks	exon	32118101	32118454	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014375"; exon_number "1"; gene_name **"COL16A1"**; oId "ENST00000489280"; contained_in "TCONS_00014357"; nearest_ref "ENST00000489280"; **class_code "="**; tss_id "TSS7102";
chr1	Cufflinks	exon	32119200	32119277	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014375"; exon_number "2"; gene_name **"COL16A1"**; oId "ENST00000489280"; contained_in "TCONS_00014357"; nearest_ref "ENST00000489280"; **class_code "="**; tss_id "TSS7102";
chr1	Cufflinks	exon	32120401	32120459	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014375"; exon_number "3"; gene_name **"COL16A1"**; oId "ENST00000489280"; contained_in "TCONS_00014357"; nearest_ref "ENST00000489280"; **class_code "=";** tss_id "TSS7102";
chr1	Cufflinks	exon	32120915	32121031	.	-	.	gene_id "XLOC_002968"; transcript_id "TCONS_00014375"; exon_number "4"; gene_name **"COL16A1"**; oId "ENST00000489280"; contained_in "TCONS_00014357"; nearest_ref "ENST00000489280"; **class_code "="**; tss_id "TSS7102";

Quantifying reads

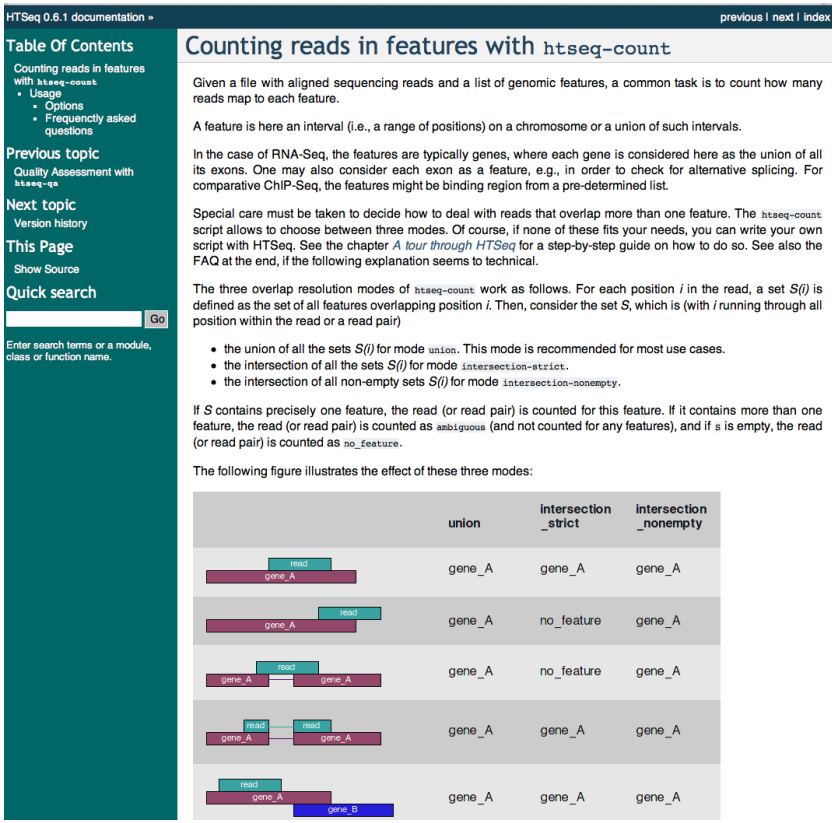
For example:

- Cufflinks – FPKM values for transcripts and genes and recently normalised counts
- Rsubread – R package
- HTSeq-count (python scripts)

Quantifying reads

BAM file

GTF file



HTSeq 0.6.1 documentation »

previous | next | index

Table Of Contents

Counting reads in features with `htseq-count`
- Usage
  - Options
  - Frequenctly asked questions

Previous topic

Quality Assessment with `htseq-qa`

Next topic

Version history

This Page

Show Source

Quick search

Go

Enter search terms or a module, class or function name.

Counting reads in features with `htseq-count`

Given a file with aligned sequencing reads and a list of genomic features, a common task is to count how many reads map to each feature.

A feature is here an interval (i.e., a range of positions) on a chromosome or a union of such intervals.

In the case of RNA-Seq, the features are typically genes, where each gene is considered here as the union of all its exons. One may also consider each exon as a feature, e.g., in order to check for alternative splicing. For comparative ChIP-Seq, the features might be binding region from a pre-determined list.

Special care must be taken to decide how to deal with reads that overlap more than one feature. The `htseq-count` script allows to choose between three modes. Of course, if none of these fits your needs, you can write your own script with HTSeq. See the chapter *A tour through HTSeq* for a step-by-step guide on how to do so. See also the FAQ at the end, if the following explanation seems to technical.

The three overlap resolution modes of `htseq-count` work as follows. For each position *i* in the read, a set *S(i)* is defined as the set of all features overlapping position *i*. Then, consider the set *S*, which is (with *i* running through all position within the read or a read pair)

- the union of all the sets *S(i)* for mode `union`. This mode is recommended for most use cases.
- the intersection of all the sets *S(i)* for mode `intersection-strict`.
- the intersection of all non-empty sets *S(i)* for mode `intersection-nonempty`.

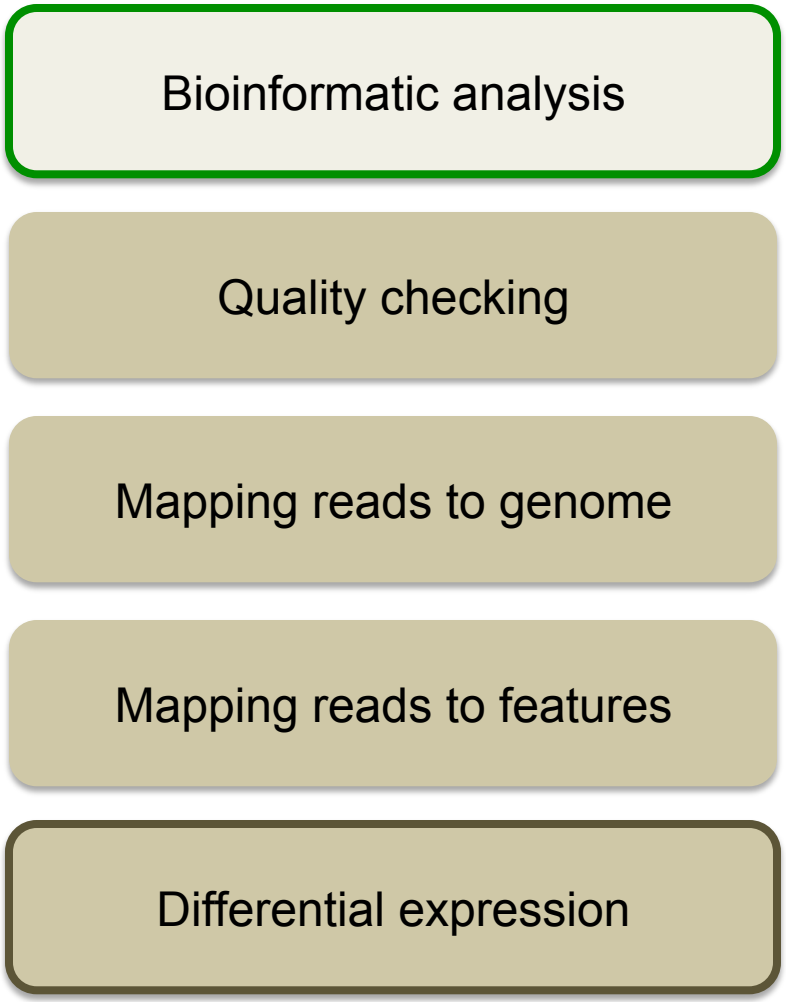
If *S* contains precisely one feature, the read (or read pair) is counted for this feature. If it contains more than one feature, the read (or read pair) is counted as `ambiguous` (and not counted for any features), and if `s` is empty, the read (or read pair) is counted as `no_feature`.

The following figure illustrates the effect of these three modes:

	union	intersection_strict	intersection_nonempty
read / gene_A	gene_A	gene_A	gene_A
read / gene_A	gene_A	no_feature	gene_A
read / gene_A gene_A	gene_A	no_feature	gene_A
read / gene_A read / gene_A	gene_A	gene_A	gene_A
read / gene_A gene_B	gene_A	gene_A	gene_A

	WT1	WT2	WT3	KO1	KO2	KO3
Mllt6	2798	2174	2303	1296	2320	1851
Zfp719	110	115	105	235	358	98
Mllt1	2557	1759	1970	1174	2307	1910
Mllt3	660	610	481	450	668	436
Rasl10a	106	92	78	49	63	62
Btg3	0	1	0	0	0	0
Btg2	2837	1642	2085	1491	2251	1707
Btg1	6639	4471	5002	4097	5659	3972
Nupl1	583	708	512	757	1341	726
Nupl2	234	197	216	198	244	214
Ttc30a2	1	1	0	2	3	1
Npat	349	423	364	533	862	324
Ttc30a1	111	109	84	71	111	97
Gga1	3154	2015	2318	1533	2717	2254
Gga3	908	765	700	622	1042	704
1700018C11Rik	0	0	2	1	0	2
Sorbs1	3343	2486	2345	1279	2100	1607
Sorbs3	1930	1540	1426	661	1059	1033
Sorbs2	631	516	436	345	500	310
Tshz2	584	593	443	304	483	390
Tshz3	354	390	298	181	244	194
Murc	636	358	441	192	446	346
2510012J08Rik	1100	831	842	738	1186	975
Speer4e	0	1	0	2	2	0
Speer4b	1	0	0	0	0	0
Arhgef25	711	671	545	311	559	378
Arhgef26	214	255	180	1257	2139	720
Efnb1	3766	2128	2989	1758	3547	2959
Mkl1	1645	1131	1306	741	1404	1170
Mkl2	1642	1417	1307	1199	1727	1210
Lmna	22181	15698	18439	16668	25554	22098
Fndc3c1	18	4	16	9	14	11
Sorcs1	42	68	34	43	48	29
Sorcs2	1071	804	788	452	764	514
Sorcs3	20	23	10	7	25	16
Mtap9	91	113	90	133	264	79
Mtap4	8247	6792	6856	4818	8099	6203
Mtap7	2841	2250	2106	2013	3406	1956
Mtap6	842	633	690	343	517	530
Trim34a	102	102	69	133	364	252
Trim34b	0	0	0	0	2	1
Mtap2	1384	1587	1181	1324	2114	869
Prepl	404	504	355	458	712	367
Cdhr1	333	178	162	317	499	395
Cdhr2	0	0	2	0	1	0
Cdhr3	13	4	7	9	8	4

RNA-Seq experiment



Differential expression analysis

Count based methods

Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2

Michael I Love, Wolfgang Huber and Simon Anders

bioRxiv posted online February 19, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/002832

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 26 no. 1 2010, pages 139–140*
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Probabilistic methods

nature biotechnology

Home | Current issue | News & comment | Research | Archive ▾ | Authors & referees ▾ | About the

home ▸ archive ▸ issue ▸ research ▸ article ▸ full text

NATURE BIOTECHNOLOGY | RESEARCH | ARTICLE

日本語要約

Differential analysis of gene regulation at transcript resolution with RNA-seq

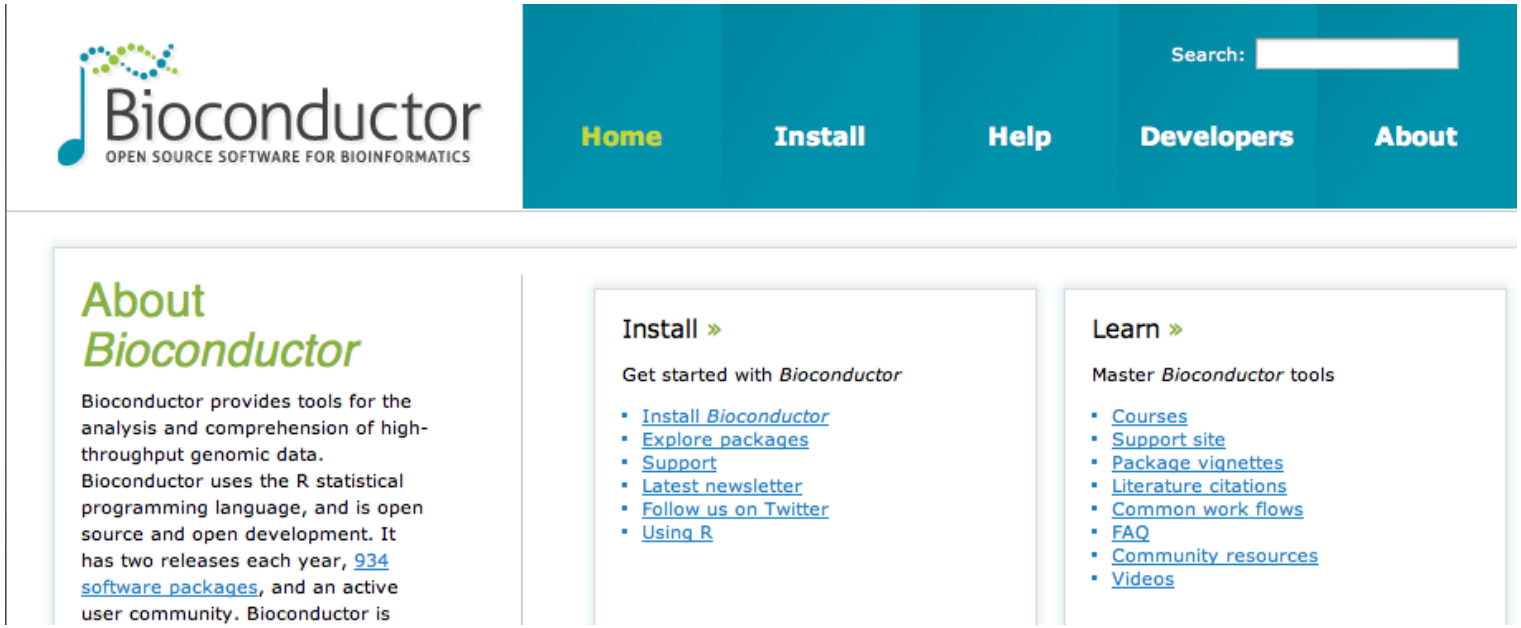
Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding authors

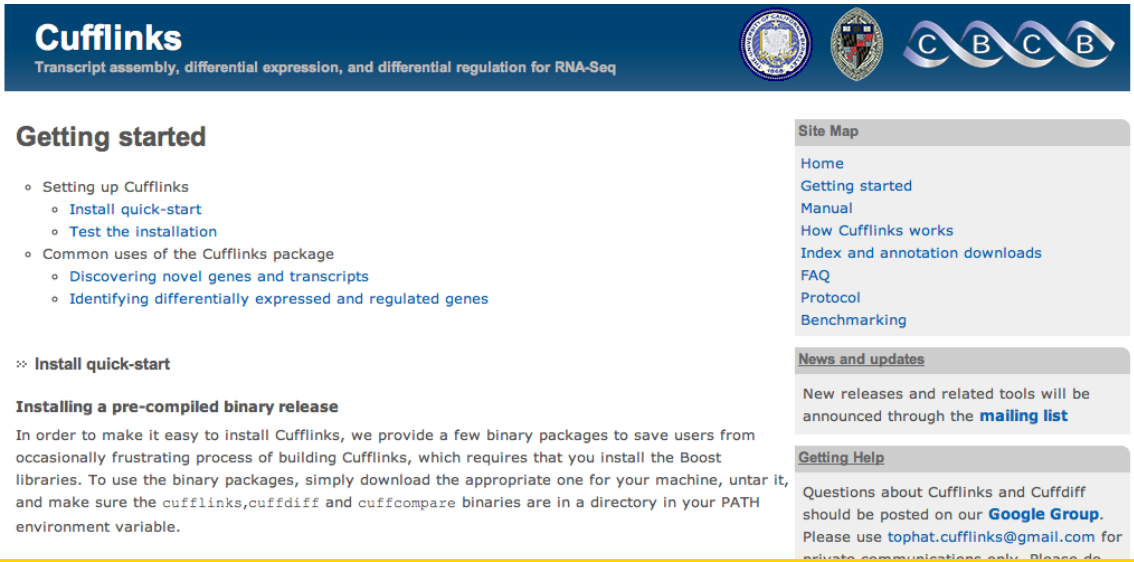
Nature Biotechnology **31**, 46–53 (2013) | doi:10.1038/nbt.2450
Received 04 May 2012 | Accepted 09 November 2012 | Published online 09 December 2012

SYSTEMS BIOLOGY INITIATIVE

UNSW AUSTRALIA



source("http://bioconductor.org/biocLite.R")
biocLite()

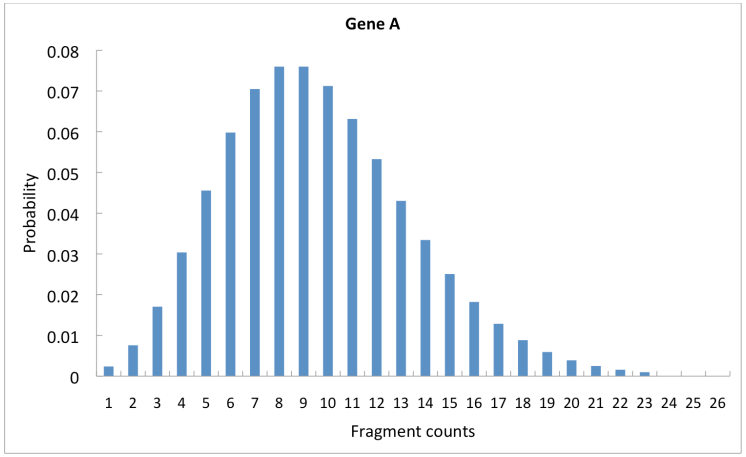


Differential expression

1. Intensive effort to develop the best statistical approach
2. RNA-Seq -> Count data - Discrete not Continuous (contrast microarray)
3. Small number of replicates per condition
4. Large number of features being tested (~ 20,000 genes)
5. Very large range of counts per gene (< 10 to 100,000s)
6. Different sequencing depth between samples

Differential expression

1. Read count per gene is a random variable
2. Counts modeled as a distribution e.g. Negative Binomial or Poisson Distribution
3. Parameters mean and variance estimated from the data
4. Null hypothesis – read count for controls and condition come from the same distribution
5. Statistical tests applied to see if Null Hypothesis should be rejected (p value)
6. Multiple testing adjustment (control false positive rate) –> **Adjusted p-value < 0.1**



Differential expression analysis tools

1. Direct count methods e.g. DESeq, edgeR, Voom, baySeq, TSPM
2. Transcript based methods e.g. Cuffdiff
3. Common steps
   - Normalisation of the counts for each sample
   - Calculating the within group variation (dispersion) for each gene
   - Determining whether the fold change between conditions is significant

Normalisation

1. Remove bias introduced by the experimental technique
2. Several approaches but still under development
3. Median-of-ratios (DESeq), TMM (edgeR), FPKM (Cuffdiff)

DESeq method:

(i) Geometric mean (GM) of gene expression across samples.

(ii) Ratio of gene count per sample to GM

(iii) Find median ratio

(iv) Use this ratio to weight the raw reads

FPKM (Fragments Per Kilobase per Million)

e.g. Sample1 30 M reads, Sample2 35 M reads

Gene A (2000 bases long), 100 reads

FPKM = 100 reads/(2 kilobase * 30 M) = 2

Sample 2 -Gene A 120 reads

FPKM = 120 reads/(2 kilobase * 35 M) =1.7

Normalisation

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012

doi:10.1093/bib/bbs046

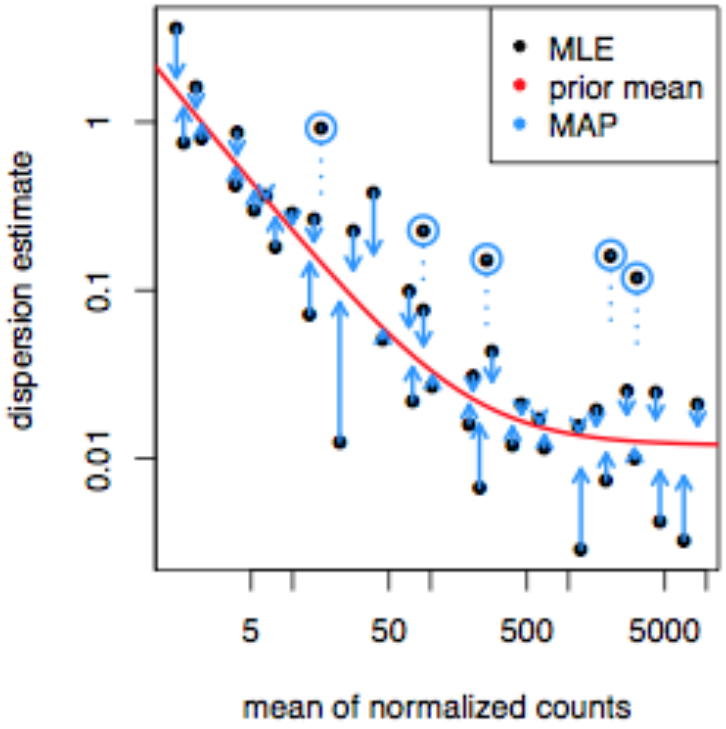
A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Dispersion

- Important step in calling differential expression
- Underestimation -> False positives
- Overestimation -> False negatives
- Small number of replicates makes it difficult
- Work around to borrow information from other genes



Love et al. 2014

Gene-wise dispersion from data
Fit curve relating expression and dispersion
Shrink gene-wise towards curve

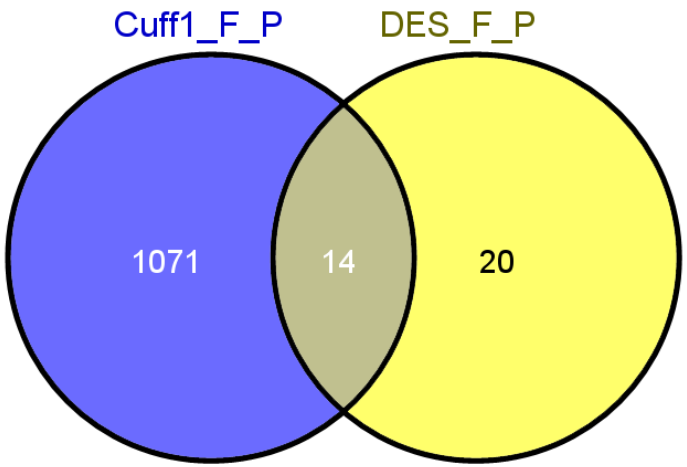
Statistical tests

1. Test probability that expression of Gene A in Condition 1 is different to expression in Condition 2
2. Also possible to test with more than one condition using Generalized Linear Models (GLMs) as implemented in edgeR and DESeq2 e.g. treatment over different time intervals, gender and treatment.
3. For GLMs you need to create a design matrix, but many examples are available such e.g. edgeR manual.

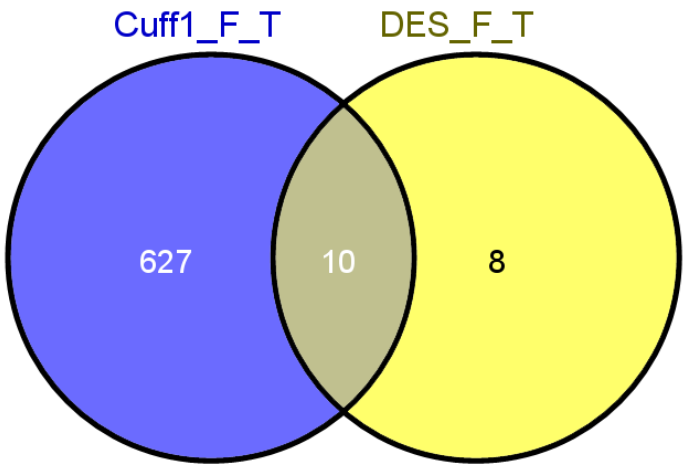
Evolution of tools

- Tophat – Cufflinks – Cuffdiff (1.3.0) pipeline for RNA-Seq
- RNA sourced from 4 different brain lobes – no independent biological replicates

Frontal lobe vs Parietal lobe



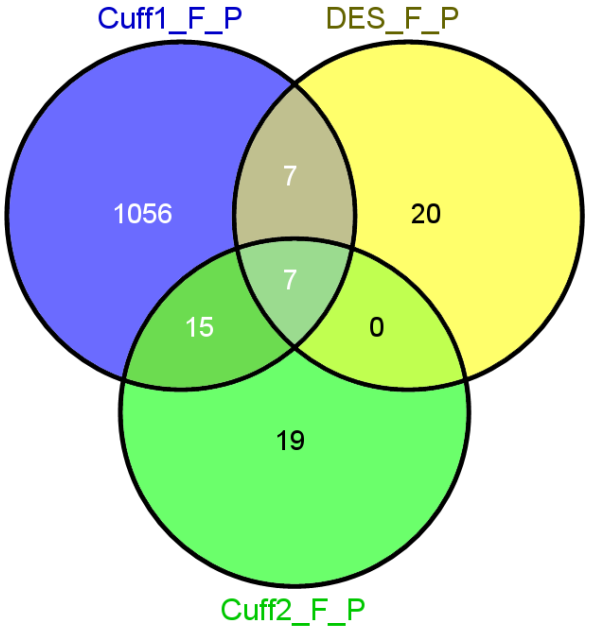
Frontal lobe vs Temporal lobe



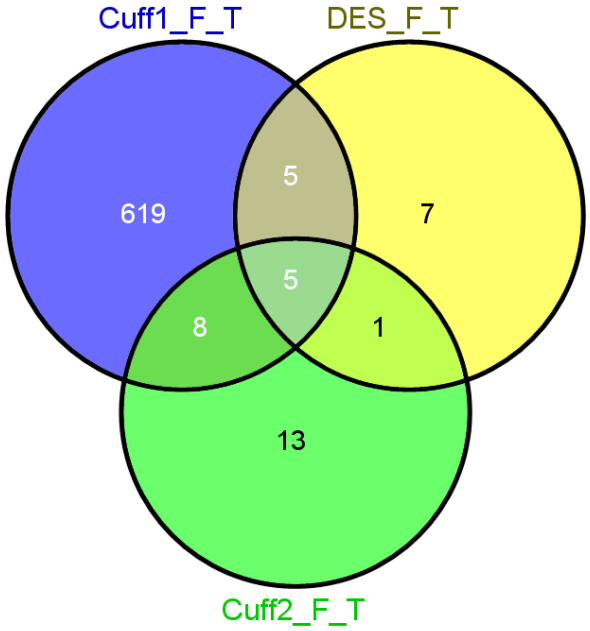
Evolution of tools

Comparing Cuffdiff, Cuffdiff2 and DESeq

Frontal lobe vs Parietal lobe

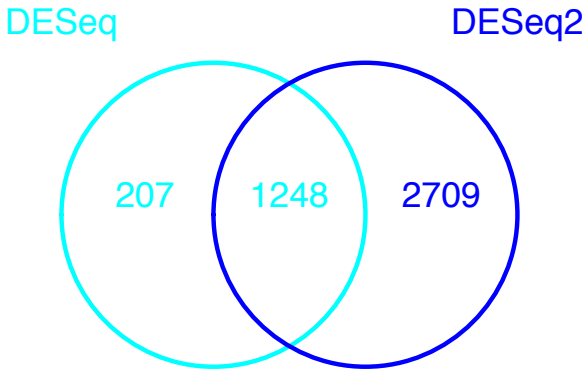


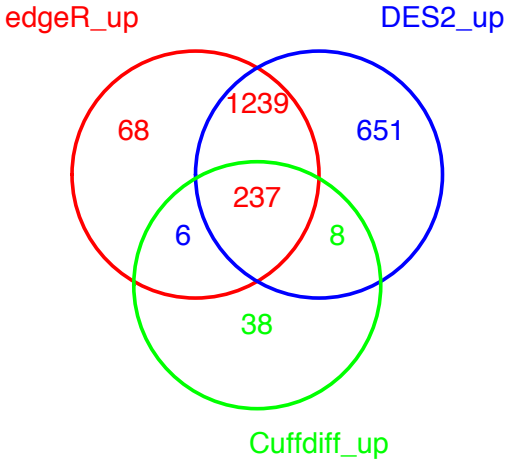
Frontal lobe vs Temporal lobe

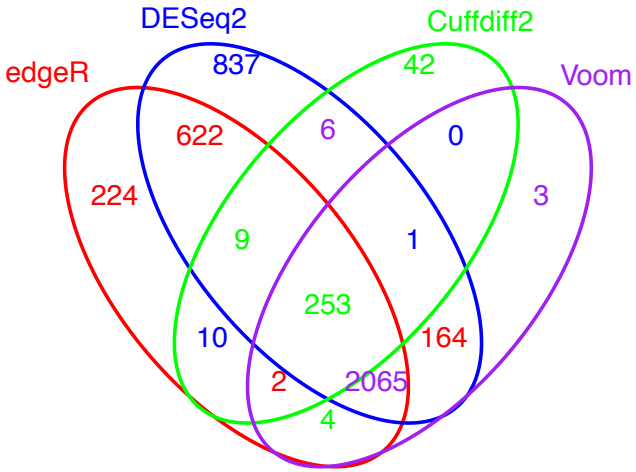


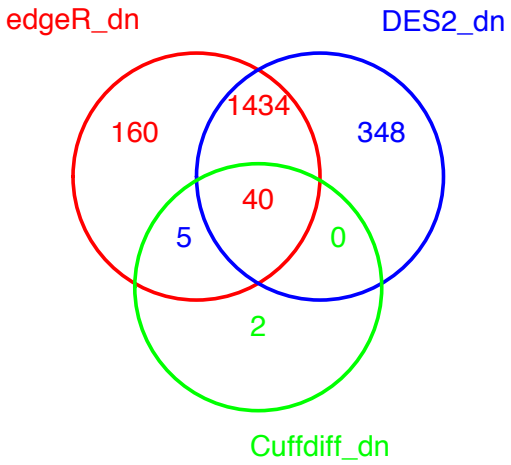
Cuffdiff (v 1.3.0), Cuffdiff2 (v 2.0.2), DESeq (v 1.10.1)

Performance of different analysis tools









Biological interpretation

- Assumption [RNA] -> [protein]
- Ideally other information should be incorporated e.g. proteomics, miRNA
- We generally rely on database annotations
  - But annotations per “gene”
  - Isoforms may have different functions
  - Functions may differ in different tissues/cells

- RNA-Seq takes a snapshot
  - but RNA is dynamic (expression, processing, degradation)
  - Single cell RNA-Seq might give a better idea of what is happening

- Genes/proteins operate in networks
  - What is the effect of change in expression in some genes in a pathway
  - compensation mechanisms/remodelling

RNA-Seq experiment

Experimental design

Library preparation

Sequencing

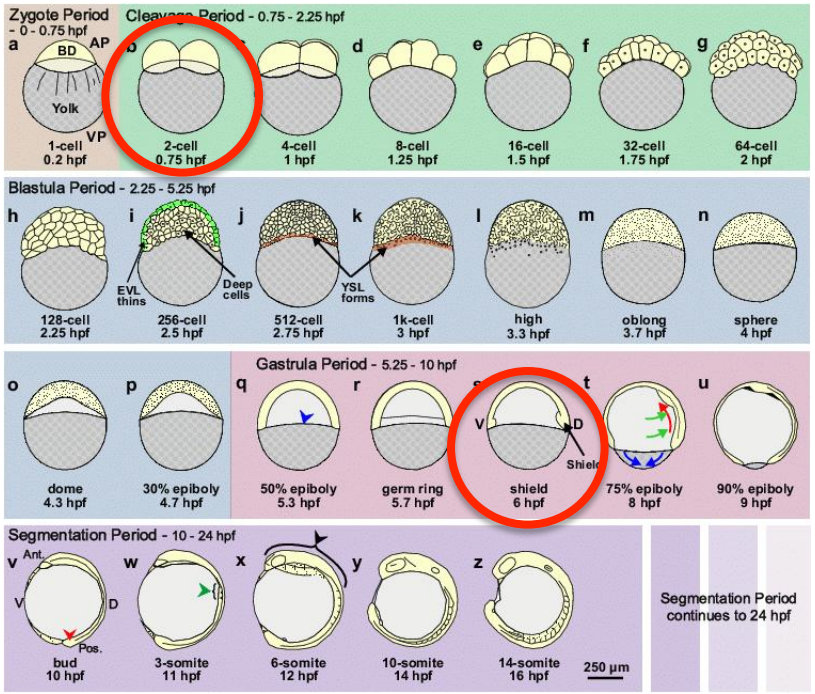
Bioinformatic analysis

Validation

Today's exercises

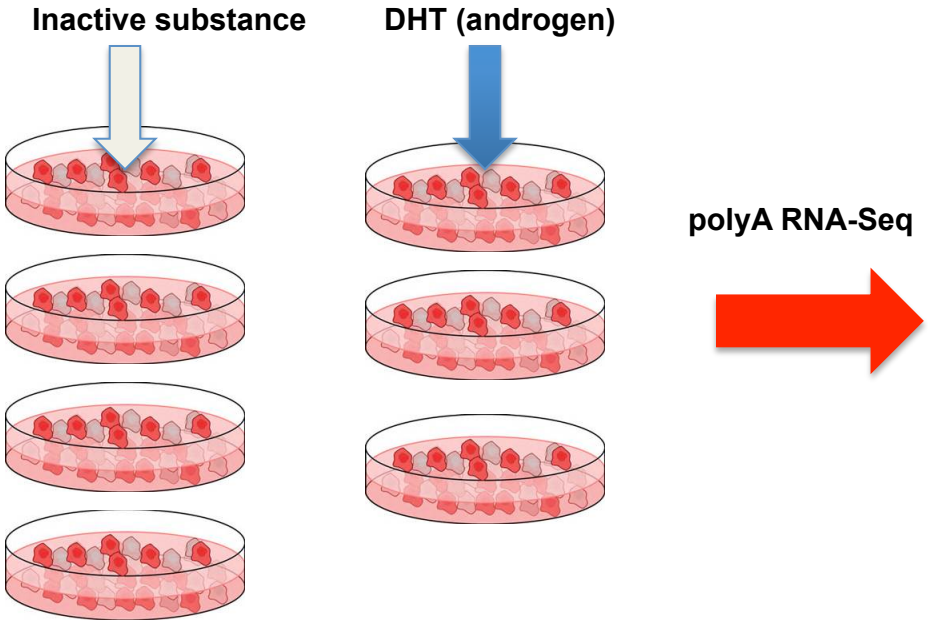
- Toy data set (small subset of data)
- Time and computer power required to map much greater than in exercise
- Tophat to map 6h sample
- Samtools to sort the bam
- Cufflinks to assemble transcriptome (with/out reference annotations)
- Cuffdiff to see which genes DE

Zebrafish (*Danio rerio*)

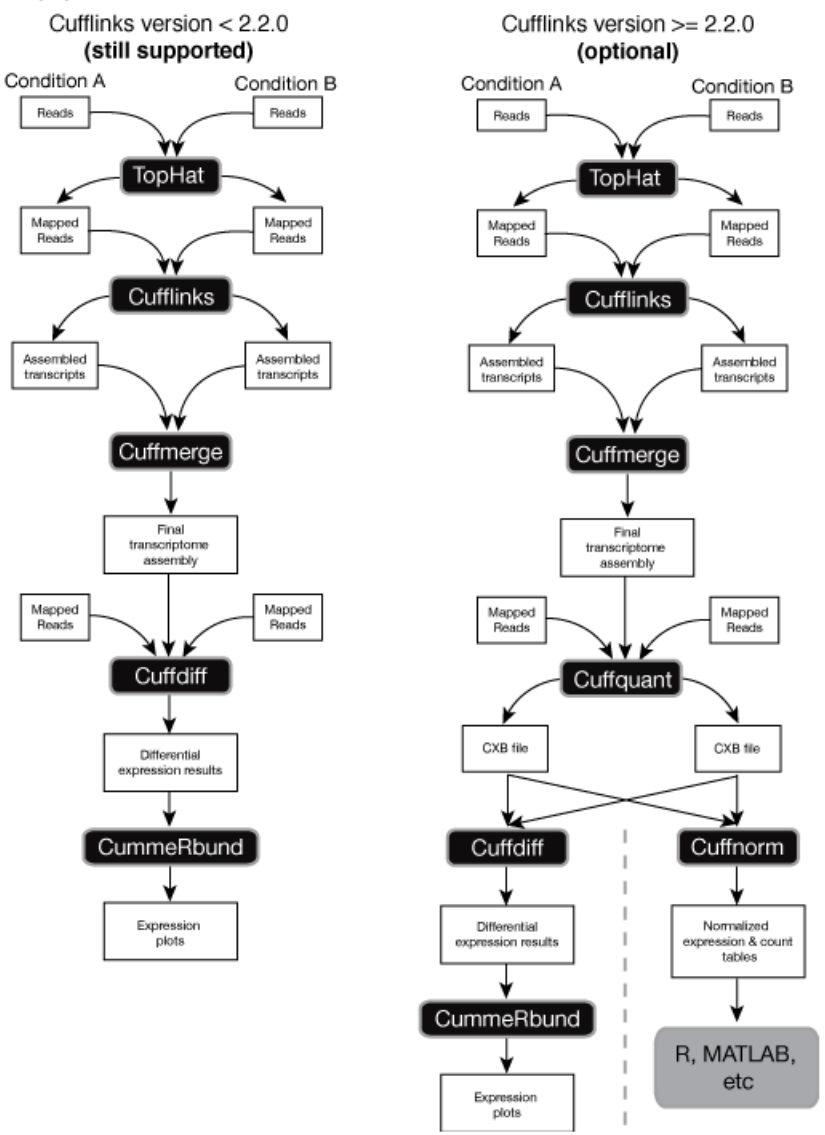


Today's exercises

- Case study in edgeR Users Guide (4.3)
- Experiment by Li et al. 2008
- LNCaP cells – androgen sensitive prostate cancer cell line



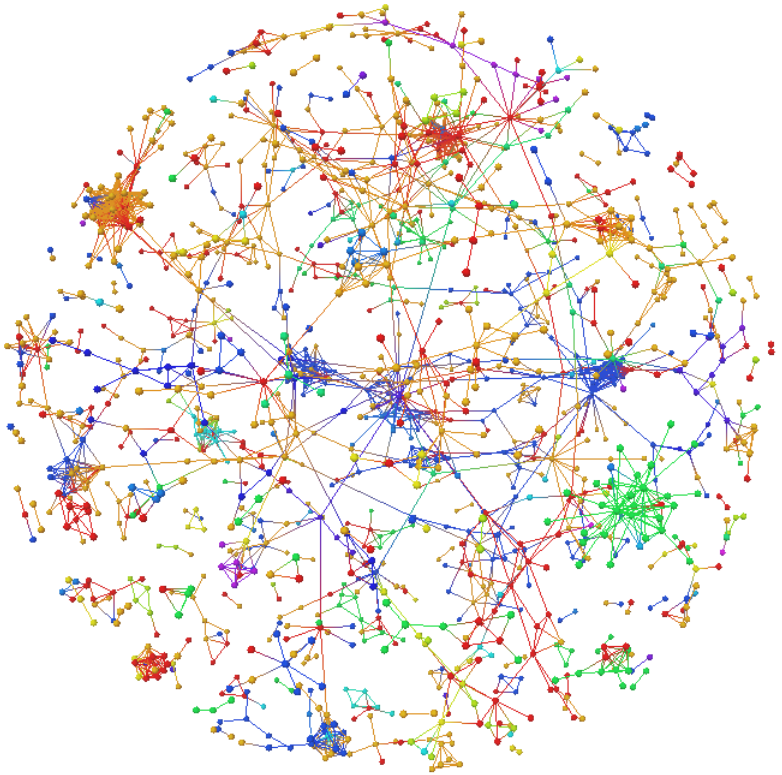
ensembl_ID	lane1	lane2	lane3	lane4	lane5	lane6	lane8
ENSG00000157214	5270	6208	7359	7521	9806	9786	4418
ENSG00000212875	4652	5409	6445	6668	5623	5727	2875
ENSG00000115053	4352	4612	5880	5892	4744	4948	1892
ENSG00000162669	4212	5000	5516	5812	4328	4172	1632
ENSG00000131051	4141	4926	5615	5548	8762	8479	3496
ENSG00000132570	4070	4704	5471	5548	6027	5665	2613
ENSG00000106070	4043	4513	6174	6385	1495	1507	420
ENSG00000122566	3775	4131	5427	5831	6959	7069	2505
ENSG00000008128	3467	4300	4514	5018	5800	5814	2079
ENSG00000096384	3354	3402	4147	4257	3240	3228	1108
ENSG00000142875	3166	3735	3920	4375	8123	7758	3880
ENSG00000087460	3055	3428	4292	4448	5301	5047	1913
ENSG00000169045	2903	3407	3901	4070	4114	4233	1635
ENSG00000151150	2845	3424	4021	4056	3263	3160	1270
ENSG00000081026	2808	3300	3708	3644	4252	4144	1896
ENSG00000212679	2727	3045	3369	3366	3497	3390	1709
ENSG00000099250	2700	2974	4605	4043	2640	2894	863
ENSG00000114867	2628	2862	4311	4500	5814	5715	1584
ENSG00000139220	2522	2789	3372	3484	3095	3242	1404
ENSG00000104067	2497	3117	3627	3900	4345	4519	1646
ENSG00000100201	2482	2996	3523	4118	4174	4045	1622
ENSG00000122786	2427	2763	2988	3063	1664	1626	585
ENSG00000159023	2273	2296	2667	2853	3202	3065	1194
ENSG00000101333	2256	2793	3456	3362	2702	2976	1320
ENSG00000140264	2248	2872	3594	3674	4355	5049	1472
ENSG00000154305	2246	2443	3129	3294	3701	4114	1572
ENSG00000217866	2168	2463	2883	2957	3291	3209	1682
ENSG00000170004	2162	2522	3240	3352	2948	3036	1100
ENSG00000167978	1935	2301	3155	3260	2698	2651	853
ENSG00000138326	1905	2007	2244	2286	2208	1986	975
ENSG00000196586	1902	2301	2726	2663	2753	2863	1086
ENSG00000184304	1884	2060	2504	2580	2072	2040	796
ENSG00000117523	1860	2164	2646	2754	2967	3154	1224
ENSG00000086205	1845	2077	2790	2682	1107	1072	375


Cufflinks version < 2.2.0
(still supported)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffdiff
Differential expression results
CummeRbund
Expression plots
Cufflinks version >= 2.2.0
(optional)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffquant
CXB file
CXB file
Cuffdiff
Cuffnorm
Differential expression results
Normalized expression & count tables
CummeRbund
R, MATLAB, etc
Expression plots

Biological interpretation

Genes/proteins operate in networks

- What is the effect of change in expression in some genes in a pathway –
- compensation mechanisms/remodelling?



Yeast protein interactions
Visualised with GEOMI
Ho et al. (2008)
J. Proteome Res. 7: 104-12

Biological interpretation

1. No single approach to downstream analysis – need to call upon background knowledge and intuition
2. GO analysis and database tools give hints about important functions and which genes may be involved in these functions
3. Sometimes a targeted approach may be better
   - Construct a tailored interaction map
   - Informed by the literature in the area
4. Human variation is significant
   - Will we move towards a personalised approach i.e. longitudinal studies for an individual rather than comparison with other individuals

Differential expression

1. Read count per gene is a random variable
2. Counts modeled as a distribution e.g. Negative Binomial or Poisson Distribution
3. Parameters mean and variance estimated from the data
4. Null hypothesis – read count for controls and condition come from the same distribution
5. Statistical tests applied to see if Null Hypothesis should be rejected (p value)
6. Multiple testing adjustment (control false positive rate) –> **Adjusted p-value < 0.1**