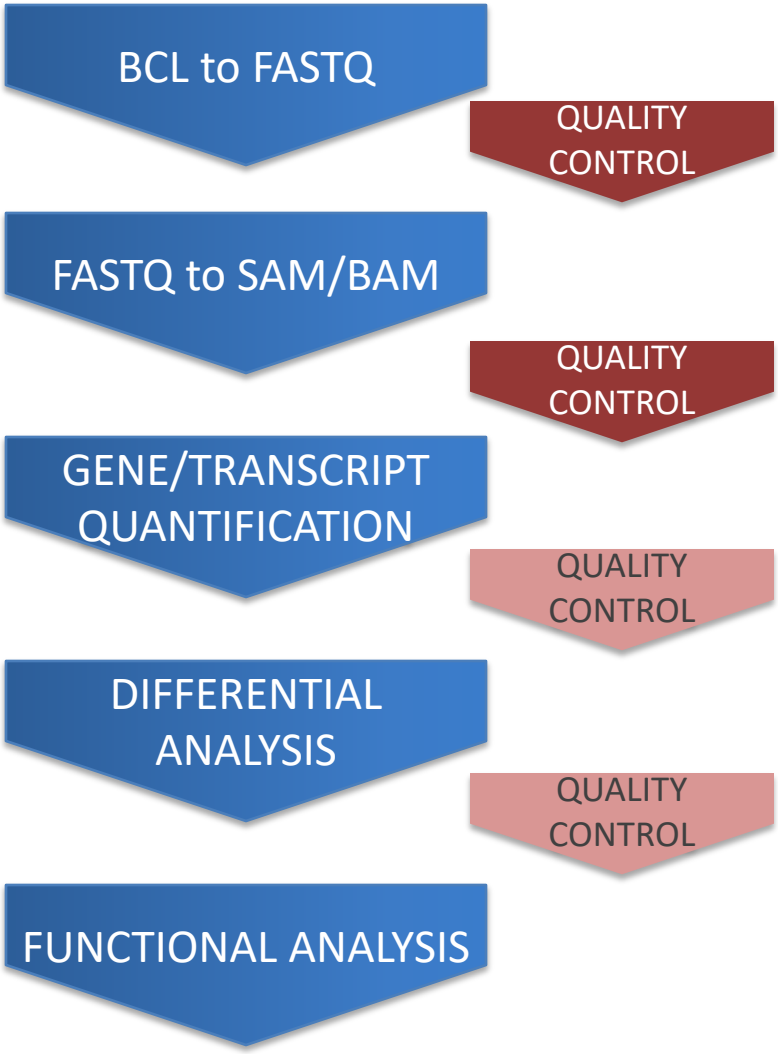
RNA-seq
Introduction to the analysis pipeline

David Gomez-Cabrero

RNA-seq: Introduction to the analysis pipeline



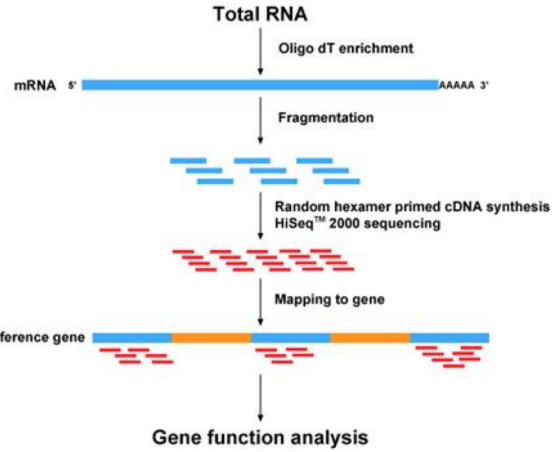
RNA-seq: Introduction to the analysis pipeline

BCL to FASTQ

QC

FASTQ to SAM/BAM

Aim: Mapping every sequence



READ
CCTTCTTTAATA

READ 1
CACAACCTTAA

READ 2
AGATGTCAGG

CCTTCTTTAATA CACAACCTTAA AGATGTCAGG

REFERENCE GENOME

TTTAAATCAATACCATTTAAGGTTCTGACCTTCTCTTTTAATATTTTGACACAACCTTAACTTGCAGATGTCAGGGTCTTC

Bowtie(2), BWA works by:

- Creating an index over the genome: FM-index
- Mapping each read using the index

Insertions

Deletions

Mismatches

…

Langmead, B., & Salzberg, S. L. (2012). Nature Methods

Lindner, B., & Friedel, S. L. (2012). PloS One

RNA-seq: Introduction to the analysis pipeline

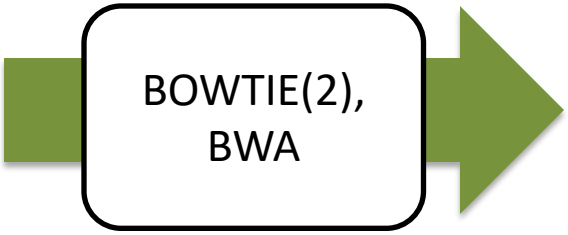
BCL to FASTQ

QC

FASTQ to SAM/BAM

Aim: Mapping every sequence

SEQUENCE



MAPPED: SAM

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                       CAGCGGCAT
```

The corresponding SAM format is:[1]

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

RNA-seq: Introduction to the analysis pipeline

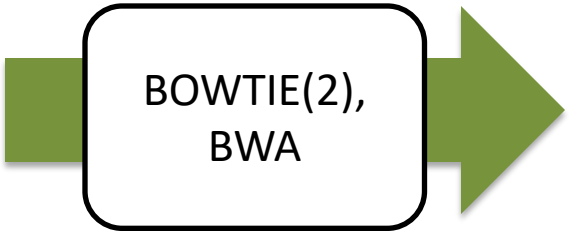
BCL to FASTQ

QC

FASTQ to SAM/BAM

Aim: Mapping every sequence

SEQUENCE



MAPPED: SAM

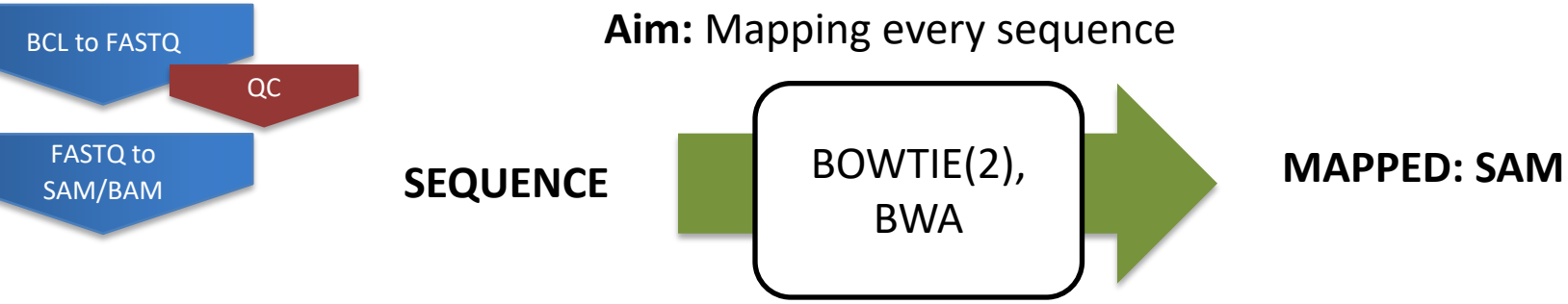
```
@HD   VN:1.0                                                    Format version, sort order,...
@SQ   SN:chr20 LN:62435964                                      Sequence name
@RG   ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891                       Platform, library, sample,...
@RG   ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
      AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
      NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
      ACCTATATCTTGGCCT GCCGATGCGGCCT A <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
      MF:i:18 RG:Z:L2
```

FLAG

QUALITY

CHECK
http://genome.sph.umich.edu/wiki/SAM

RNA-seq: Introduction to the analysis pipeline

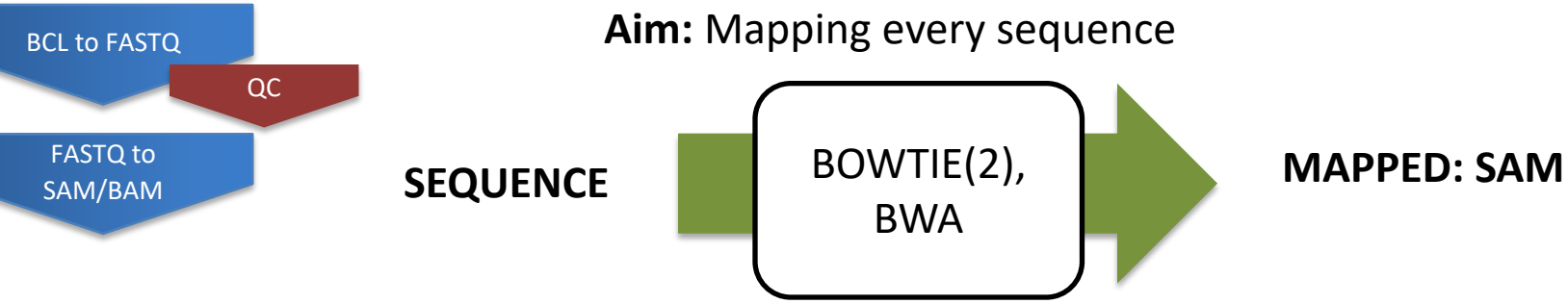


Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

FLAG: Combination of bitwise FLAGs.[4] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

RNA-seq: Introduction to the analysis pipeline



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	`[0,`2^{16}`-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,`2^{31}`-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,`2^{8}`-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	`[0,`2^{31}`-1]`	Position of the mate/next read
9	TLEN	Int	`[-`2^{31}`+1,`2^{31}`-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

3. RNAME: Reference sequence NAME of the alignment. If `@SQ` header lines are present, RNAME (if not '*') must be present in one of the `SQ-SN` tag. An unmapped segment without coordinate has a '*' at this field. However, an unmapped segment may also have an ordinary coordinate such that it can be placed at a desired position after sorting. If RNAME is '*', no assumptions can be made about POS and CIGAR.

4. POS: 1-based leftmost mapping POSition of the first matching base. The first base in a reference sequence has coordinate 1. POS is set as 0 for an unmapped read without coordinate. If POS is 0, no assumptions can be made about RNAME and CIGAR.

5. MAPQ: MAPping Quality. It equals $-10\log_{10}\Pr\{\text{mapping position is wrong}\}$, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

RNA-seq: Introduction to the analysis pipeline

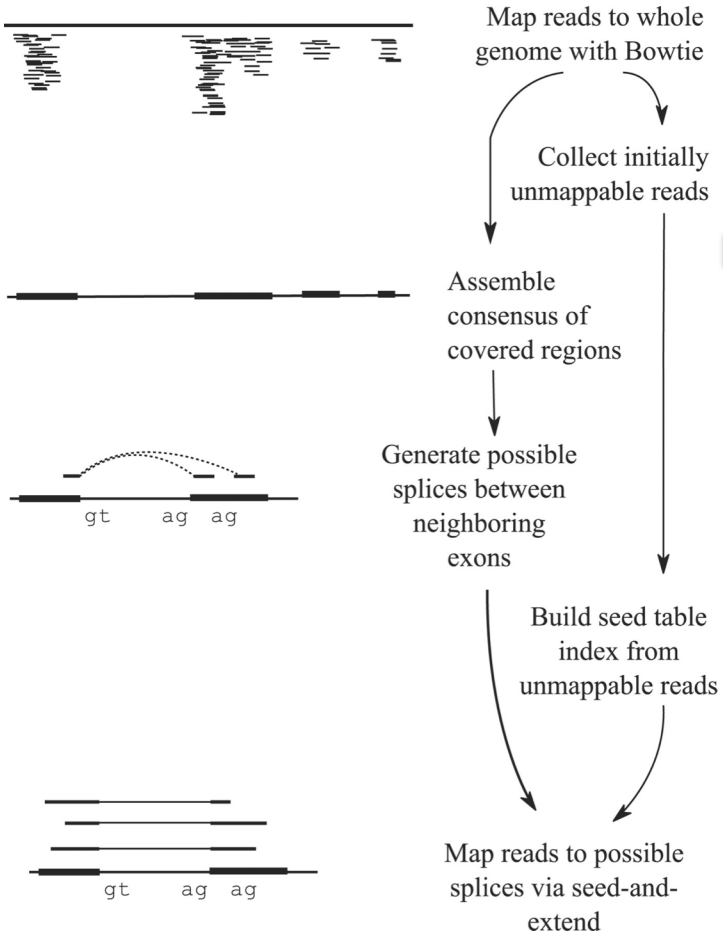
BCL to FASTQ

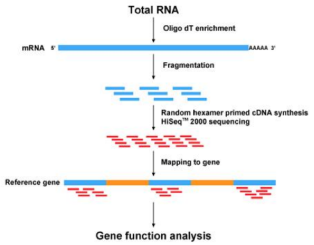
QC

FASTQ to SAM/BAM

Aim: Mapping every sequence

TOPHAT





QC

- % mapped
- % mapped 1 read
- % chimeric
- % duplicate mapping

PICCARD TOOLS

SAMTOOLS

Kim, D., et al. (2013) Genome Biology, Trapnell et al. (2009). Bioinformatics

RNA-seq: Introduction to the analysis pipeline

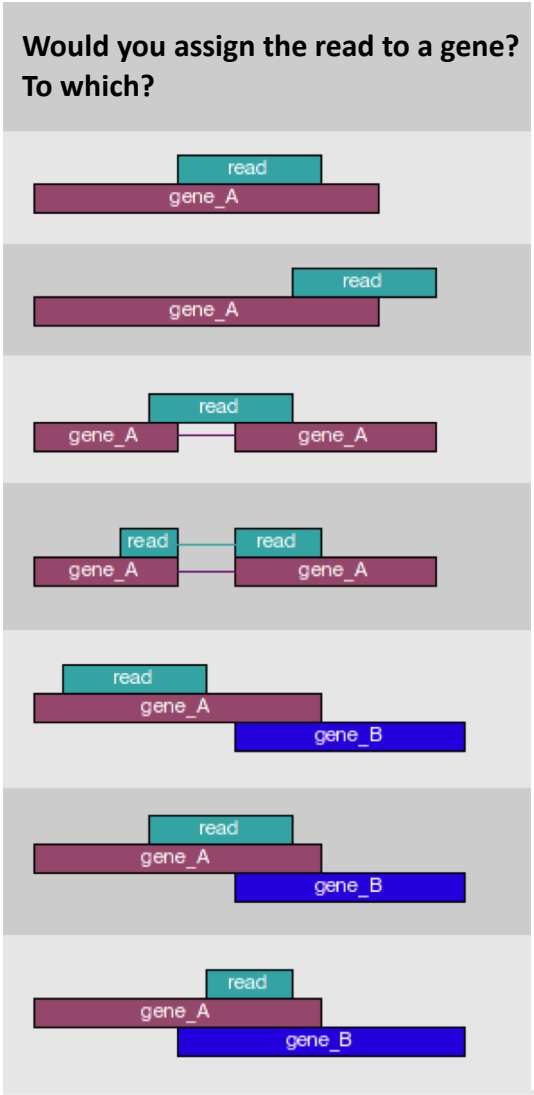
BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION



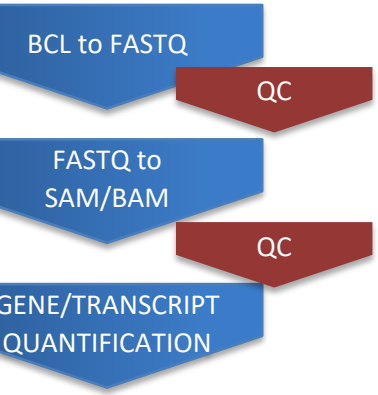
BIOINFORMATICS ORIGINAL PAPER
Vol. 31 no. 2 2015, pages 166–169
doi:10.1093/bioinformatics/btu638
Genome analysis
Advance Access publication September 25, 2014

HTSeq—a Python framework to work with high-throughput sequencing data

Simon Anders*, Paul Theodor Pyl and Wolfgang Huber
Genome Biology Unit, European Molecular Biology Laboratory, 69111 Heidelberg, Germany
Associate Editor: Michael Brudno

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

RNA-seq: Introduction to the analysis pipeline



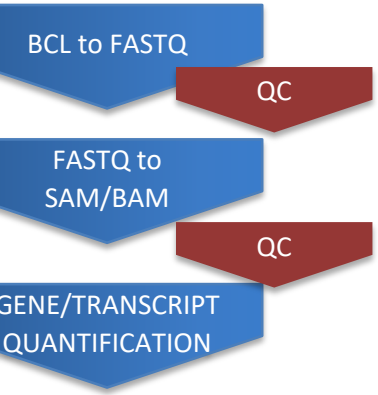
	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H	Batch_4_Ctr_2H	Batch_1_Ctr_6H	Batch_3_Ctr_6H	Batch_4_Ctr_6H	Batch_1_Ctr_12H	Batch_3_Ctr_12H	Batch_4_Ctr_12H	Batch_1_Ctr_18H	Batch_3_Ctr_18H	Batch_4_Ctr_18H
ENSMUSG00000026385	5936	6017	6871	4859	7371	6728	4263	8122	5783	4892	6158	8162	5382	7272	6208
ENSMUSG00000062248	7313	8675	11390	8861	9538	9036	23518	12993	7804	8455	9843	8211	10525	9397	12363
ENSMUSG00000030695	224723	183474	144567	126522	245484	190956	324866	231070	162338	124592	209620	195721	176776	160089	244650
ENSMUSG00000078908	2716	2534	4211	2314	3899	1719	4199	3073	1021	3078	2709	2301	3119	2856	2264
ENSMUSG00000087412	4994	6755	5235	6856	3237	3319	506	2996	980	6809	5004	7987	2507	2513	1819
ENSMUSG00000005204	7064	8326	13616	7934	9873	6937	10174	10424	6257	8452	9439	8228	8698	9012	10317
ENSMUSG00000036676	437	423	664	441	559	229	252	398	229	417	417	325	458	693	268
ENSMUSG00000079671	25	17	16	27	26	30	0	13	45	15	15	12	3	18	0
ENSMUSG00000037965	3917	4181	4550	3859	4014	3697	3028	4095	3551	3123	3184	3915	3749	4863	2129
ENSMUSG00000032417	36	38	43	19	33	27	78	22	9	17	23	50	26	29	26
ENSMUSG00000091086	189	216	256	181	116	78	31	108	34	200	82	289	57	162	3
ENSMUSG00000032329	2380	2339	3210	1963	2548	1360	2554	2341	1125	2371	2408	2013	2499	2468	2051
ENSMUSG00000047417	9121	10004	16617	10104	13254	9922	16748	13349	6408	10926	13303	11792	11453	11084	12958
ENSMUSG00000038895	2004	2301	2665	1908	2487	3064	4578	3160	2378	1797	2019	2958	1974	1998	2609
ENSMUSG00000018379	17661	17368	31568	18021	23241	12498	21906	20805	9416	19729	20616	17836	19529	25797	13958
ENSMUSG00000081476	408	479	775	554	489	421	850	653	257	469	626	608	386	438	539
ENSMUSG00000024853	25667	27645	47371	26006	34340	18356	34582	34666	15448	29569	34022	24351	33152	36010	26125
ENSMUSG00000004768	649	653	1055	359	875	383	803	681	287	463	529	557	507	626	420
ENSMUSG00000027746	1609	2060	2838	2129	2103	1903	2096	2618	2006	1789	2130	2038	2315	2787	2176
ENSMUSG00000090357	10	18	5	21	18	35	0	36	37	18	21	18	29	41	15
ENSMUSG00000079426	21126	22451	35646	17063	31168	19101	42215	27395	14054	24376	27651	27013	22451	26030	28291
ENSMUSG00000049044	78	63	45	20	71	63	79	39	52	44	45	23	57	54	0
ENSMUSG00000043162	974	989	1557	1074	1343	1369	2874	1250	1001	918	1111	1443	1009	1445	1513
ENSMUSG00000097573	70	73	106	74	83	120	55	134	65	60	54	53	57	45	3
ENSMUSG00000044813	1314	1313	1546	1233	1682	1553	1657	1779	1077	1327	1854	1903	1525	1697	2117
ENSMUSG00000050965	7805	7480	15792	7196	12528	6208	8247	11490	4369	7826	10073	7807	8651	11724	6577
ENSMUSG00000037513	5605	5787	9907	4686	7239	3931	8797	6130	3185	6365	6499	5592	6125	5965	6098
ENSMUSG00000040029	2805	2720	5077	2181	3370	1502	1517	2831	1125	3174	2566	1958	2996	3765	1350
ENSMUSG00000026426	2467	2064	3018	1730	3673	1810	3361	2606	1238	2071	2317	2051	2254	2376	1766
ENSMUSG00000020794	1951	1980	2797	2013	2202	1324	1980	2001	1029	1853	2290	1395	2416	2673	1209

Each cell: count per gene per sample

1) Are they comparable as they are now?

2) What is the probabilistic model to consider?

RNA-seq: Introduction to the analysis pipeline



	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

For a given gene what can be the distribution?

Poisson

Negative Binomial

...

PROBABILISTIC MODEL

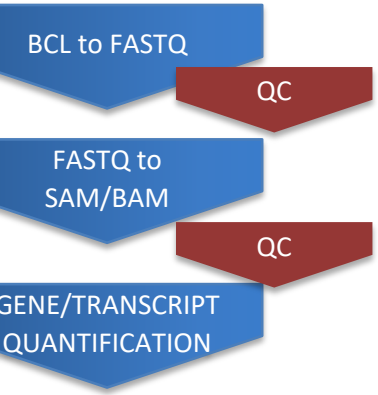
Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [2]Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [4]Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; [5]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA

RNA-seq: Introduction to the analysis pipeline



	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

For a given gene what can be the distribution?

Poisson

Negative Binomial

…

Poisson distribution

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

- ***Discrete*** *probability distribution*
- *The* ***probability of a given number of events*** *occurring in* ***a fixed interval of time and/or space***
- *If these events occur with a* ***known average rate and independently of the time since the last event***
- ***In Poisson mean equals variance***

RNA-seq: Introduction to the analysis pipeline

BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION

PROBABILISTIC MODEL

	Batch_1_Ctr_ZH	Batch_2_Ctr_ZH	Batch_4_Ctr_ZH	Batch_1_Ctr_ZH	Batch_3_Ctr_ZH
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7113	8675	11360	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4984	6755	5233	6856	3237
ENSMUSG00000003204	7064	8326	13016	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4592	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

Poisson

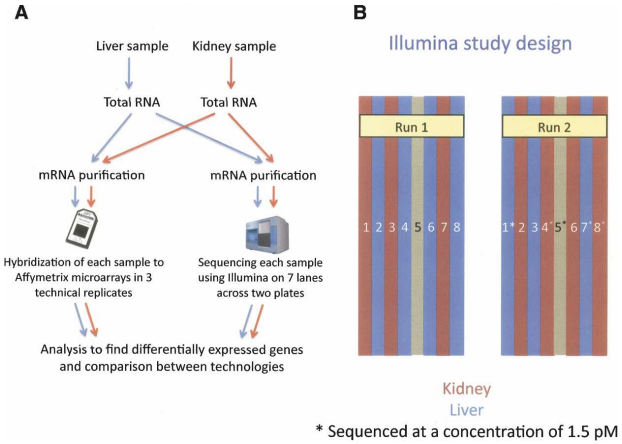


Figure 1. Graphical representation of the study design. (*A*) Summary of the experimental design. (*B*) The lanes in which each sample was sequenced across the two runs. In each run, the control sample was sequenced in lane *5*. Samples were sequenced at two concentrations: 1.5 pM (indicated by an asterisk) and 3 pM (no asterisk).

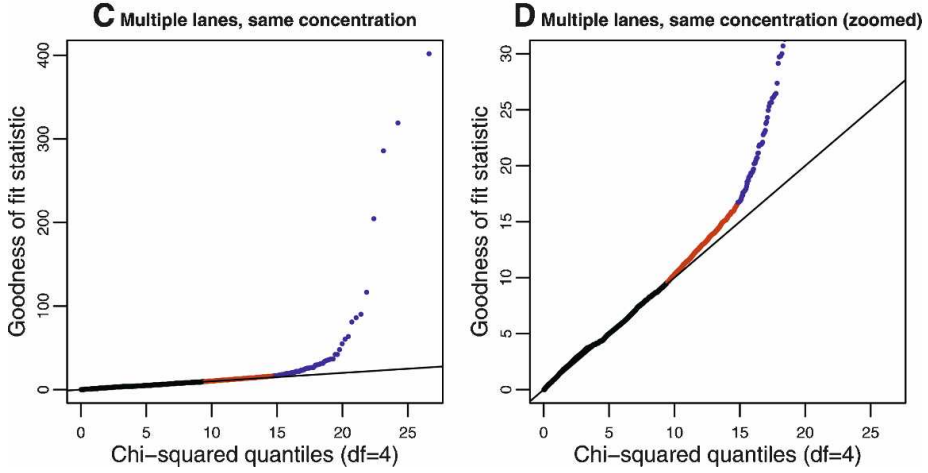
Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

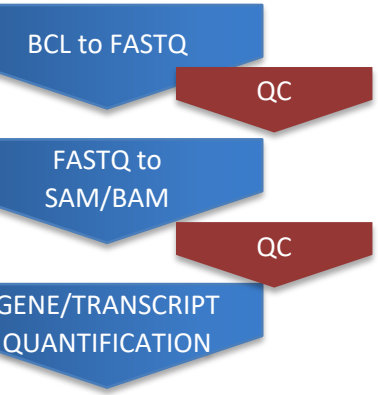
John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

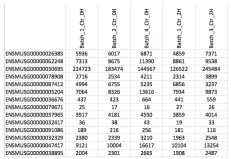
[1]*Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA;* [2]*Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA;* [3]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA;* [4]*Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA;* [5]*Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA*

To compare multiple lanes for a lane effect, we took a closely related approach based on the following Poisson model. If x_{ijk} represents the number of reads mapped to gene j for the kth lane of data from sample i, x_{ijk} can be modeled as independent Poisson random variables with mean $\mu_{ijk} = c_{ik}\lambda_{ijk}$, where the λ_{ijk} are constrained to sum to 1 across genes j. The parameter c_{ik} represents the total rate at which lane k of sample i produces reads, and the parameter λ_{ijk} represents the rate at which reads map to gene j (in lane k of sample i) relative to other genes. The hypothesis of no lane effect corresponds to λ_{ijk} being constant across lanes k. For each gene, we compute a goodness-of-fit statistic across L lanes to test this hypothesis: if there is no lane effect, then this statistic should be χ^2 distributed on $L - 1$ degrees of freedom. A *qq*-plot of these values (Fig. 2C,D; Supplemental Fig. 6) shows that, in each case, only a small proportion of genes (~0.5%) show strong evidence for a lane effect (i.e., extra-Poisson variation).



RNA-seq: Introduction to the analysis pipeline





Poisson?

Negative Binomial

Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [2]Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [4]Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; [5]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 21 2007, pages 2881–2887
doi:10.1093/bioinformatics/btm453

Gene expression

Moderated statistical tests for assessing differences in tag abundance

Mark D. Robinson[1,2] and Gordon K. Smyth[2,*]

PROBABILISTIC MODEL

Assuming an NB distribution for the tag counts Y_{ij}, we have:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi)$$

where ϕ is the dispersion. We choose the parameterization such that $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi)$, making $\phi = 0$ the Poisson distribution.

Let λ_i be the true relative abundance of this tag in RNA of class i. Then $\mu_{ij} = m_{ij}\lambda_i$ where m_{ij} is the library size for sample j. To assess differences in relative abundance, the null hypothesis H_0: $\lambda_1 = \lambda_2$ is tested against the two-sided alternative, and this is repeated for each tag.

RNA-seq: Introduction to the analysis pipeline

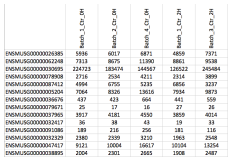
BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION



Poisson?

Negative Binomial

Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [2]Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA; [4]Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; [5]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 21 2007, pages 2881–2887
doi:10.1093/bioinformatics/btm453

Gene expression

Moderated statistical tests for assessing differences in tag abundance

Mark D. Robinson[1,2] and Gordon K. Smyth[2,*]

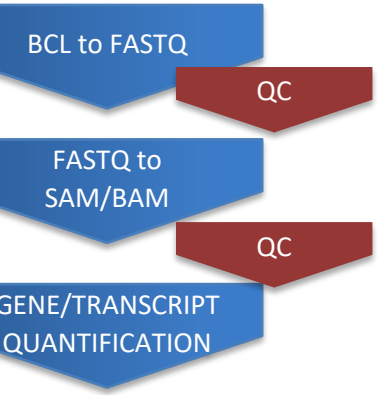
PROBABILISTIC MODEL

Negative Binomial NB(r,p)

$$f(k;r,p) \equiv \Pr(X=k) = \binom{k+r-1}{k} p^k (1-p)^r \quad \text{for } k = 0,1,2,\ldots$$

- ***Discrete** probability distribution*
- Number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted r) occurs
- ***In NB variance=*** $\dfrac{pr}{(1-p)^2}$
- ***And*** $\text{Poisson}(\lambda) = \lim_{r\to\infty} \text{NB}\left(r, \dfrac{\lambda}{\lambda+r}\right).$

RNA-seq: Introduction to the analysis pipeline



	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

Negative Binomial

DESeq2 and edgeR

PROBABILISTIC MODEL

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

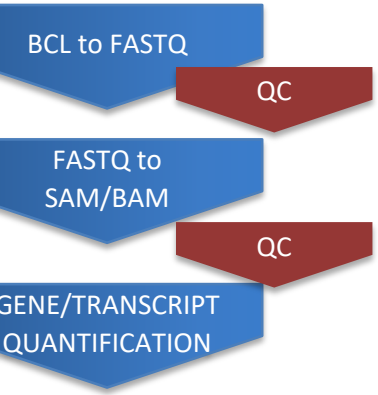
Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Functional Genomics Center UNI ETH, Zurich, Switzerland. [7]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.D.R. (mark.robinson@imls.uzh.ch) or W.H. (whuber@embl.de).

Box 2 | Differences between DESeq and edgeR

The two packages described in this protocol, DESeq and edgeR, have similar strategies to perform differential analysis for count data. However, they differ in a few important areas. First, their look and feel differs. For users of the widely used limma package[60] (for analysis of microarray data), the data structures and steps in edgeR follow analogously. The packages differ in their default normalization: edgeR uses the trimmed mean of M values[56], whereas DESeq uses a relative log expression approach by creating a virtual library that every sample is compared against; in practice, the normalization factors are often similar. Perhaps most crucially, the tools differ in the choices made to estimate the dispersion. edgeR moderates feature-level dispersion estimates toward a trended mean according to the dispersion-mean relationship. In contrast, DESeq takes the maximum of the individual dispersion estimates and the dispersion-mean trend. In practice, this means DESeq is less powerful, whereas edgeR is more sensitive to outliers. Recent comparison studies have highlighted that no single method dominates another across all settings[27,61,62].

RNA-seq: Introduction to the analysis pipeline

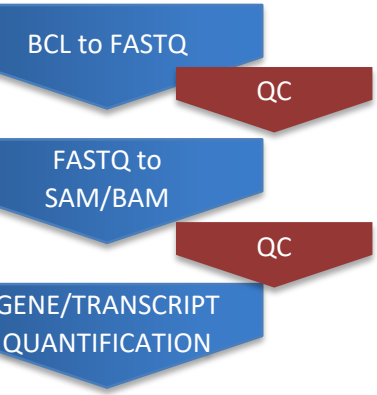


NORMALIZATION

	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487
	270855	235300	213911	175848	302439

How we compare genes between samples when we have different different sequencing depth?

RNA-seq: Introduction to the analysis pipeline



	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

How we compare genes between samples

Sequencing depth
Length of the gene

NORMALIZATION

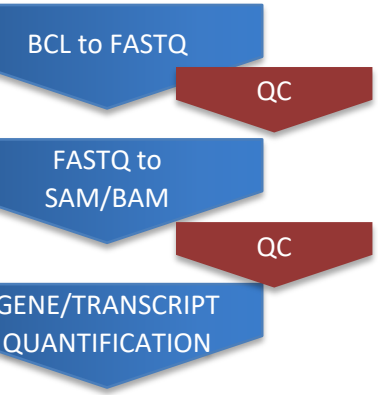
Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

RPKM = reads per kilobase per million
= [# of mapped reads]/[length of transcript in kilo base]/[million mapped reads]
= [# of mapped reads]/([length of transcript]/1000)/([total reads]/10^6)

FPKM = fragments per kilobase per million
= [# of fragments]/[length of transcript in kilo base]/[million mapped reads]
= [# of fragments]/([length of transcript]/1000)/([total reads]/10^6)

RNA-seq: Introduction to the analysis pipeline



	Batch_1_Ctr_0H	Batch_2_Ctr_0H	Batch_4_Ctr_0H	Batch_1_Ctr_2H	Batch_3_Ctr_2H
ENSMUSG00000026385	5936	6017	6871	4859	7371
ENSMUSG00000062248	7313	8675	11390	8861	9538
ENSMUSG00000030695	224723	183474	144567	126522	245484
ENSMUSG00000078908	2716	2534	4211	2314	3899
ENSMUSG00000087412	4994	6755	5235	6856	3237
ENSMUSG00000005204	7064	8326	13616	7934	9873
ENSMUSG00000036676	437	423	664	441	559
ENSMUSG00000079671	25	17	16	27	26
ENSMUSG00000037965	3917	4181	4550	3859	4014
ENSMUSG00000032417	36	38	43	19	33
ENSMUSG00000091086	189	216	256	181	116
ENSMUSG00000032329	2380	2339	3210	1963	2548
ENSMUSG00000047417	9121	10004	16617	10104	13254
ENSMUSG00000038895	2004	2301	2665	1908	2487

How we compare genes between samples

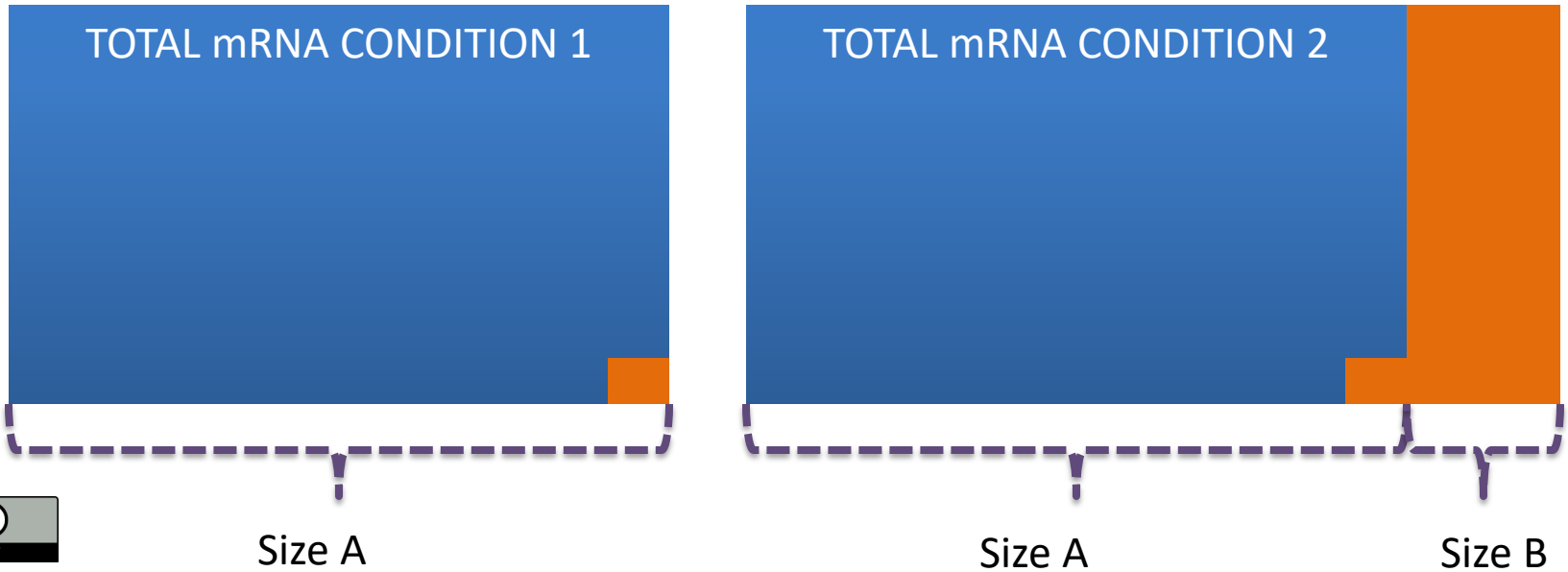
Sequencing depth
Length of the gene

NORMALIZATION

Problems with RPKM/FPKM

Depicted mRNA of gene X

In condition 2 all genes will be considered as lower expressed compared to Condition 1 expect orange gene



RNA-seq: Introduction to the analysis pipeline

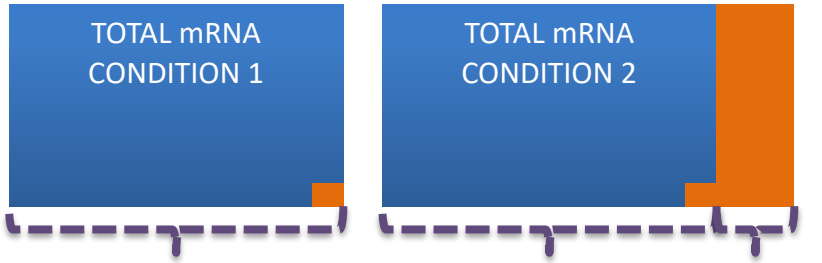
BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION



$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

$$\text{where } S_k = \sum_{g=1}^{G} \mu_{gk} L_g;$$

Robinson and Oshlack *Genome Biology* 2010, **11**:R25
http://genomebiology.com/2010/11/3/R25

Genome Biology

METHOD Open Access

A scaling normalization method for differential expression analysis of RNA-seq data

Mark D Robinson[1,2*], Alicia Oshlack[1*]

NORMALIZATION

Proportions of reads to gene g in sample k

$$M_g = \log_2 \frac{Y_{gk} / N_k}{Y_{gk'} / N_{k'}}$$

Proportions of reads to gene g in sample k'

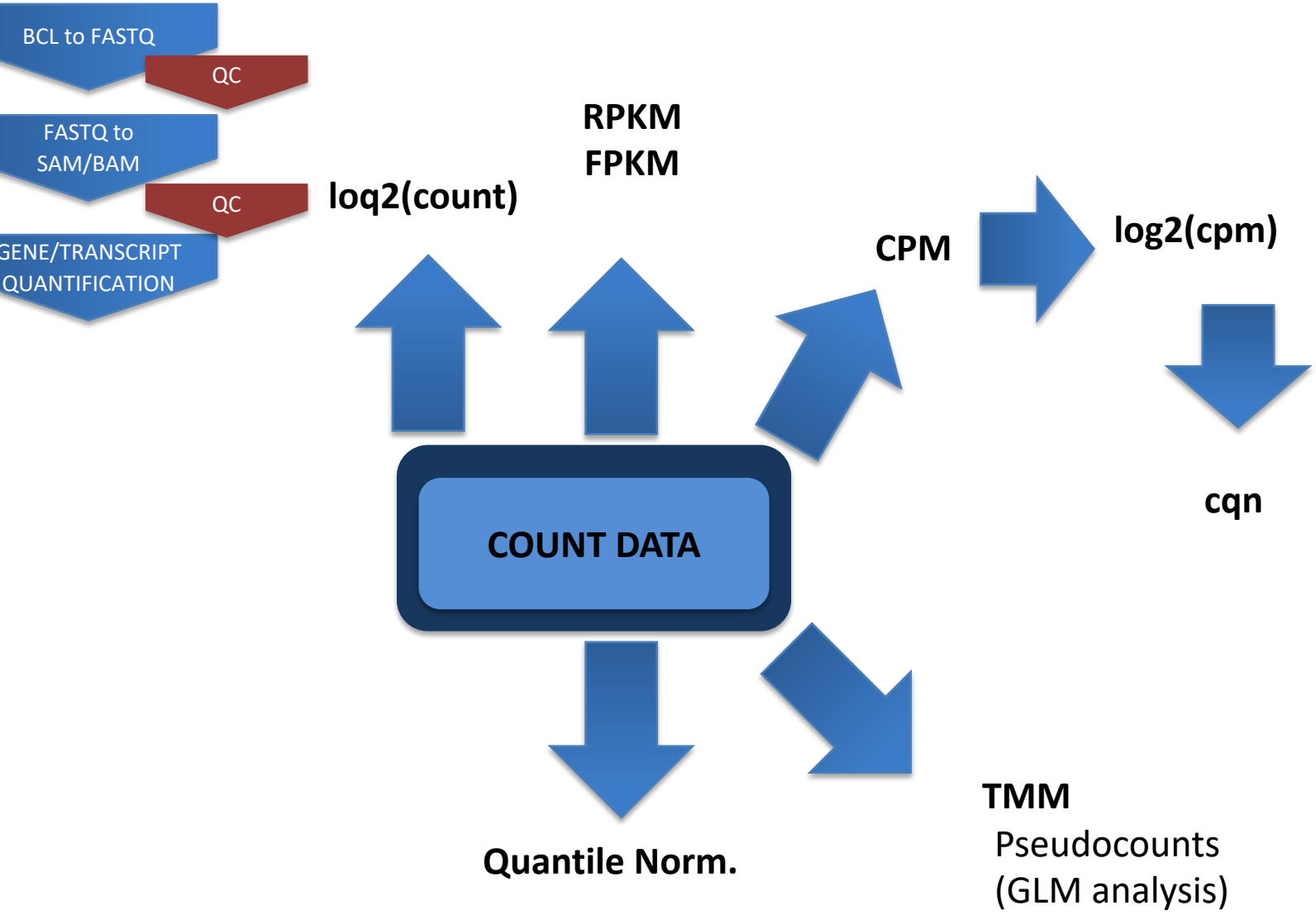
1) Trim % of top values

2) Compute weighted average of M values

3) Correction factor (offset) PSEUDO-COUNTS

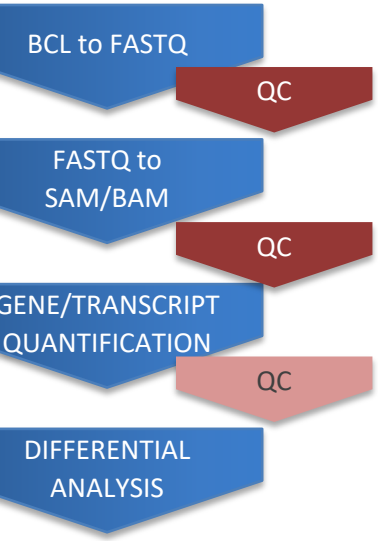
UNDER THE ASSUMPTION THAT MOST OF THEM ARE NOT DIFFERENTIALLY EXPRESSED!!

RNA-seq: Introduction to the analysis pipeline

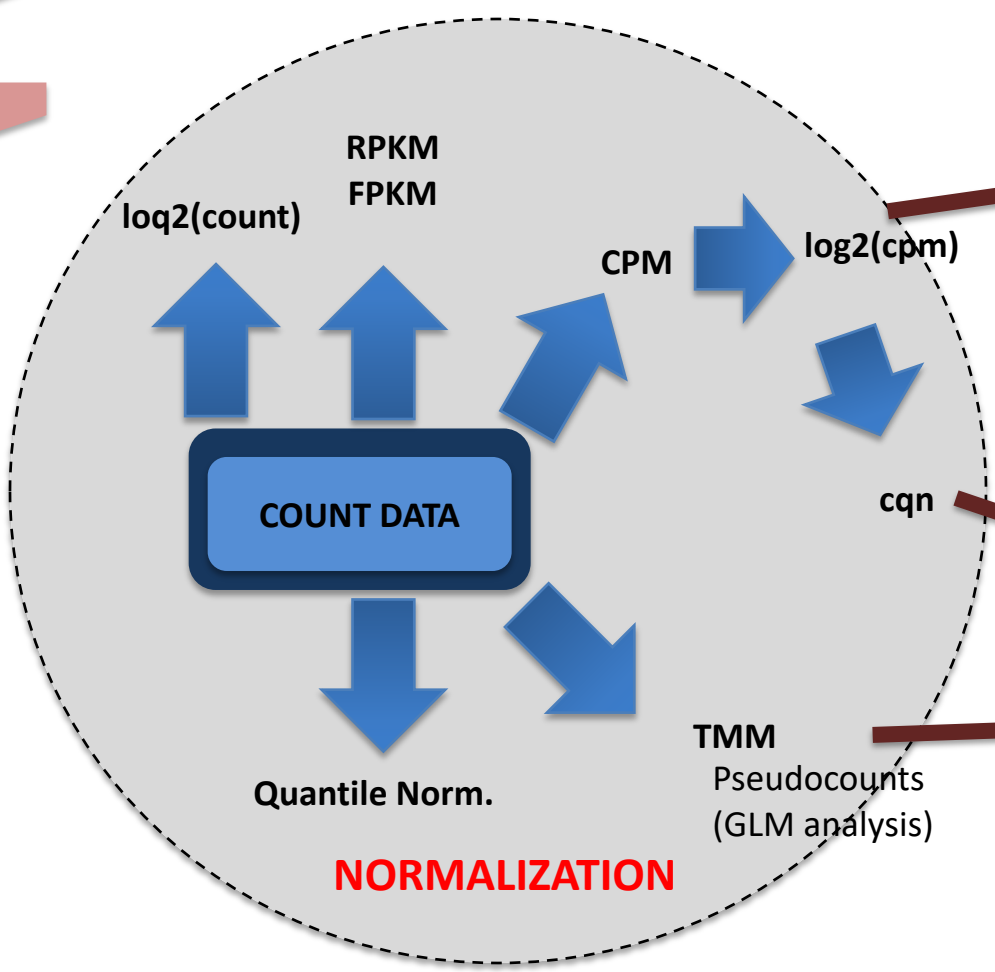


NORMALIZATION

RNA-seq: Introduction to the analysis pipeline



DIFFERENTIAL EXPRESSION



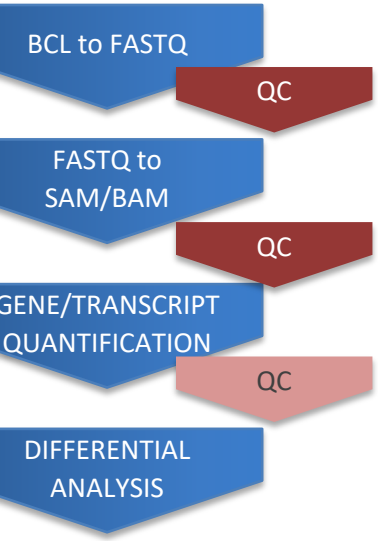
Linear Models

limma

Generalized Linear Models

edgeR
DESeq(2)

RNA-seq: Introduction to the analysis pipeline



TYPE OF MODELING

Linear Models: limma,…

Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

Preprint January 2004; with corrections 30 June 2009

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Generalized Linear Models: edgeR, DESeq,…

A generalized linear model (or GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor—that is a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

RNA-seq: Introduction to the analysis pipeline

BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION

QC

DIFFERENTIAL ANALYSIS

TYPE OF MODELING

Linear Models: limma,…

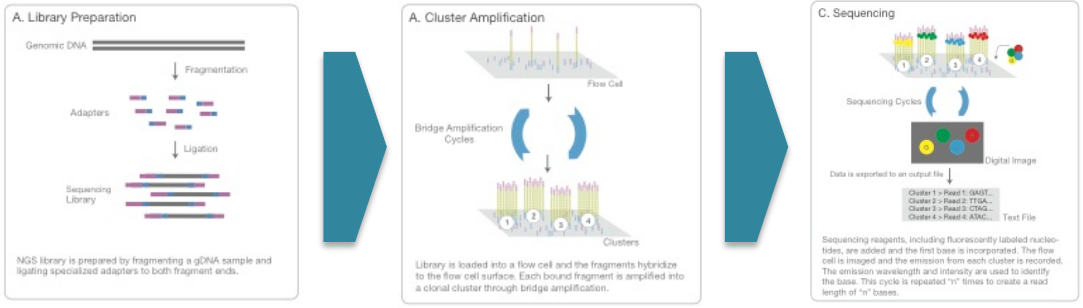
Generalized Linear Models: edgeR, DESeq,…

DESIGN

Gender, Age, …

Batch, Library Preparation,…

AT THE VERY BEGINNING



RNA-seq: Introduction to the analysis pipeline

BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION

QC

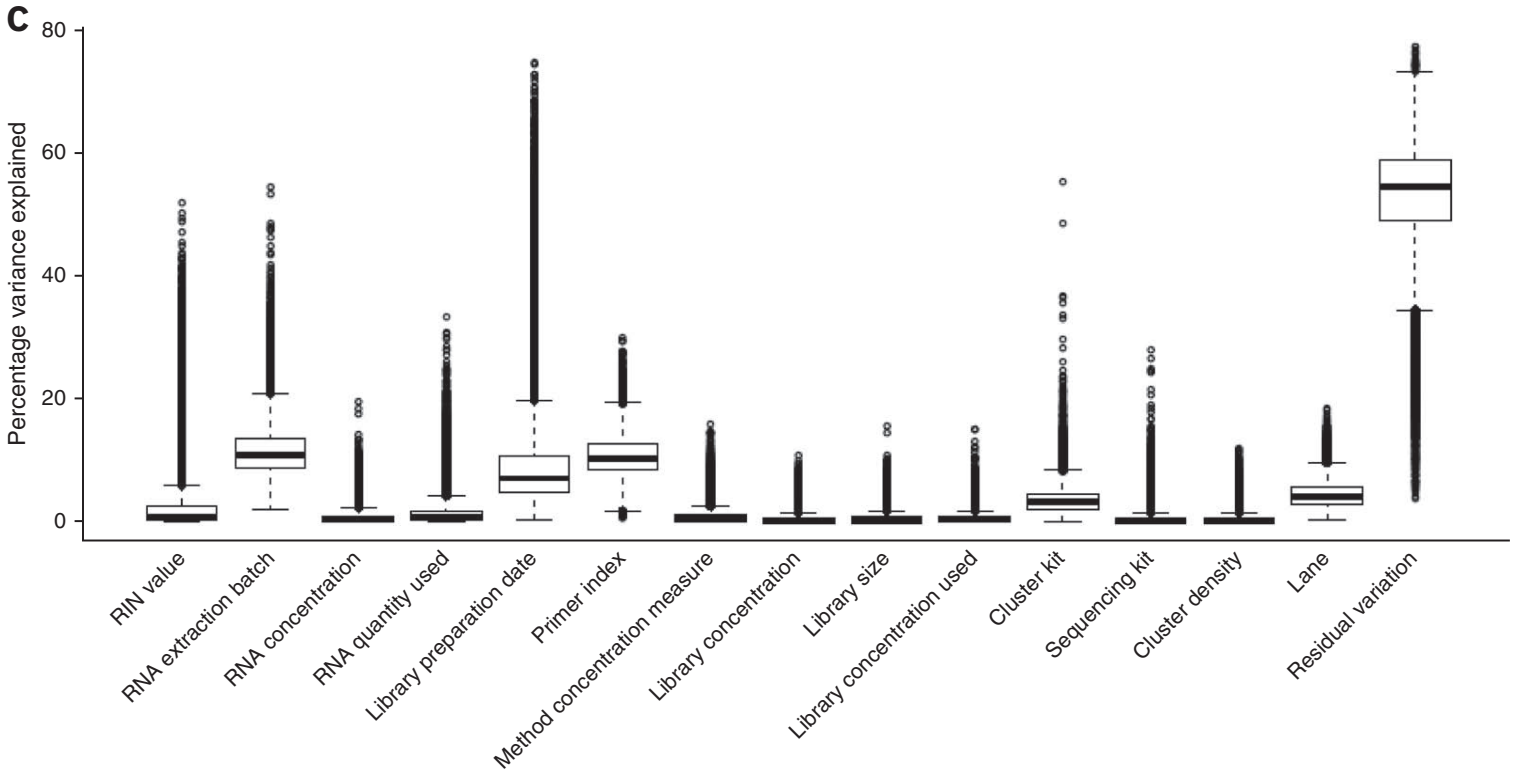
DIFFERENTIAL ANALYSIS

ARTICLES

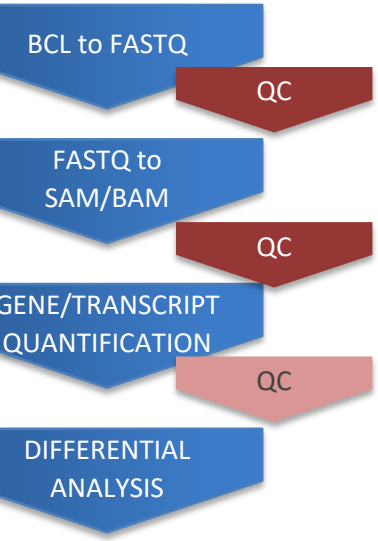
nature biotechnology

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Peter A C 't Hoen[1,2], Marc R Friedländer[3–6,15], Jonas Almlöf[7,15], Michael Sammeth[3–5,8,14], Irina Pulyakhina[1], Seyed Yahya Anvar[1,9], Jeroen F J Laros[1,2,9], Henk P J Buermans[1,9], Olof Karlberg[7], Mathias Brännvall[7], The GEUVADIS Consortium[10], Johan T den Dunnen[1,2,9], Gert-Jan B van Ommen[1], Ivo G Gut[8], Roderic Guigó[3–5], Xavier Estivill[3–6], Ann-Christine Syvänen[7], Emmanouil T Dermitzakis[11–13] & Tuuli Lappalainen[11–13]



RNA-seq: Introduction to the analysis pipeline



OUTPUT OF DIFFERENTIAL EXPRESSION

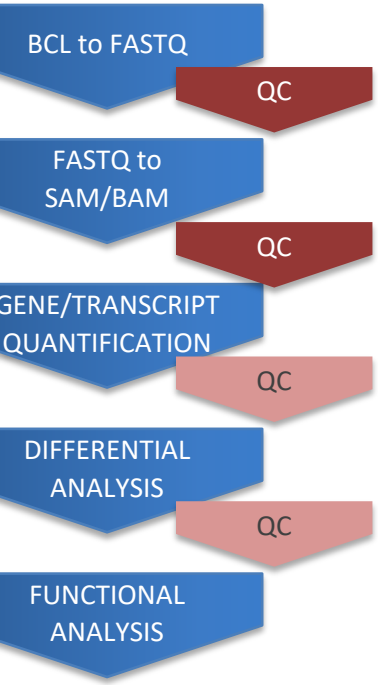
	logFC	logCPM	LR	PValue	FDR
FBgn0039155	-4.61	5.87	902	3.96e-198	2.85e-194
FBgn0025111	2.87	6.86	641	2.17e-141	7.81e-138
FBgn0039827	-4.05	4.40	457	2.11e-101	5.07e-98
FBgn0035085	-2.58	5.59	408	9.31e-91	1.68e-87
FBgn0000071	2.65	4.73	365	2.46e-81	3.54e-78
FBgn0003360	-3.12	8.42	359	3.62e-80	4.34e-77

LogFC: what do we mean by logFC?

P-Value:

FDR: why is this important?

RNA-seq: Introduction to the analysis pipeline



	logFC	logCPM	LR	PValue	FDR
FBgn0039155	-4.61	5.87	902	3.96e-198	2.85e-194
FBgn0025111	2.87	6.86	641	2.17e-141	7.81e-138
FBgn0039827	-4.05	4.40	457	2.11e-101	5.07e-98
FBgn0035085	-2.58	5.59	408	9.31e-91	1.68e-87
FBgn0000071	2.65	4.73	365	2.46e-81	3.54e-78
FBgn0003360	-3.12	8.42	359	3.62e-80	4.34e-77

TO GENE-SETS...

GSEA
Gene Set Enrichment Analysis

Fisher / Hypergeometric

Network based

MSigDB
Molecular Signatures Database

GO
the Gene Ontology

KEGG
Kyoto Encyclopedia of Genes and Genomes

METHODS

GENE SETS

RNA-seq: Introduction to the analysis pipeline

BCL to FASTQ

QC

FASTQ to SAM/BAM

QC

GENE/TRANSCRIPT QUANTIFICATION

QC

DIFFERENTIAL ANALYSIS

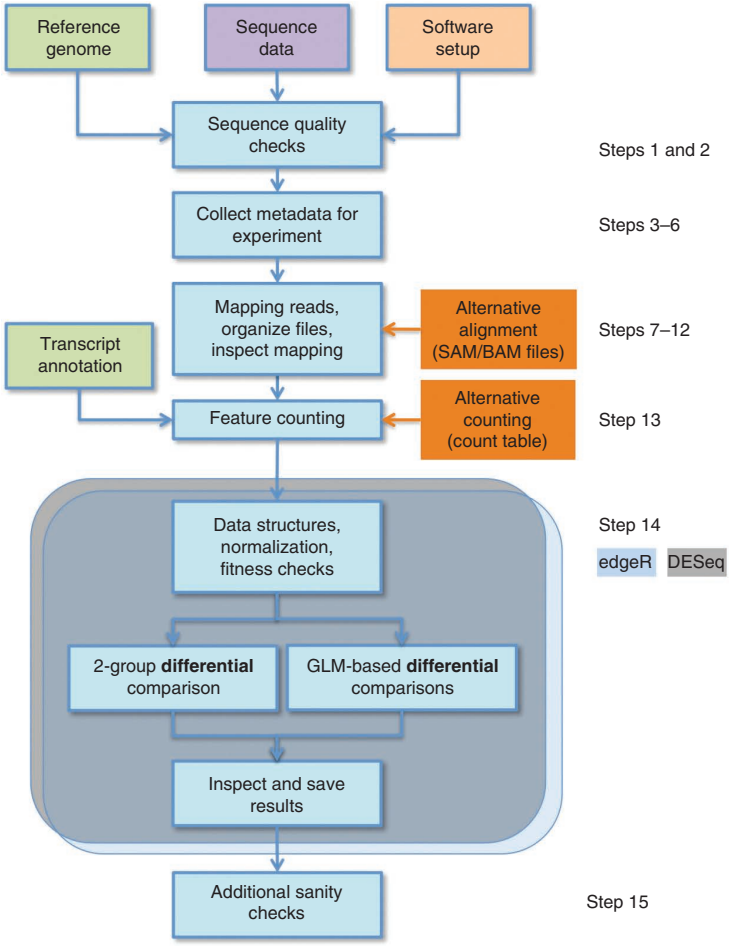
QC

FUNCTIONAL ANALYSIS

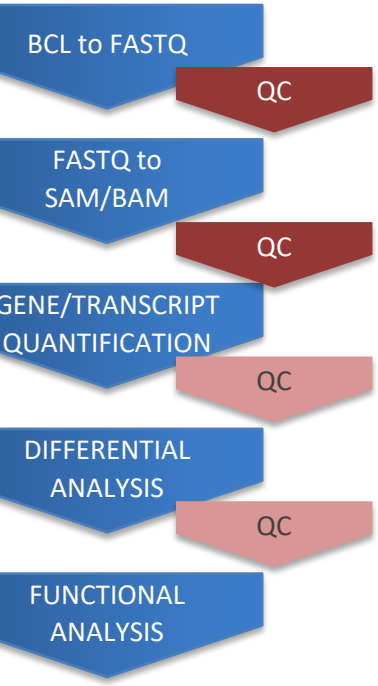
PROTOCOL

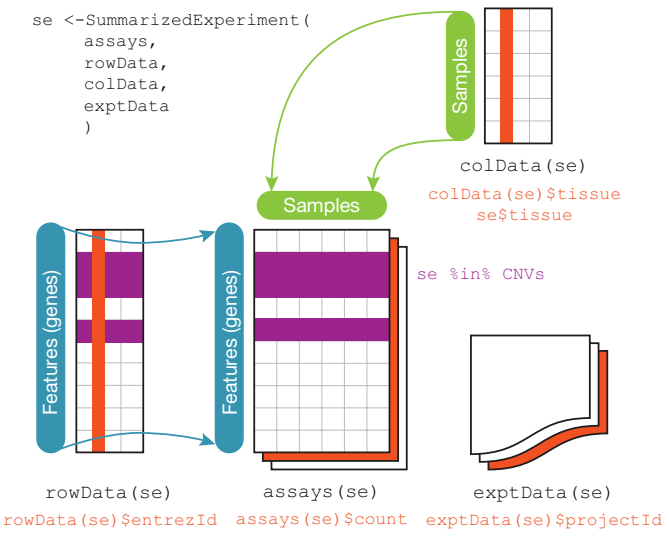
Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]



RNA-seq: Introduction to the analysis pipeline





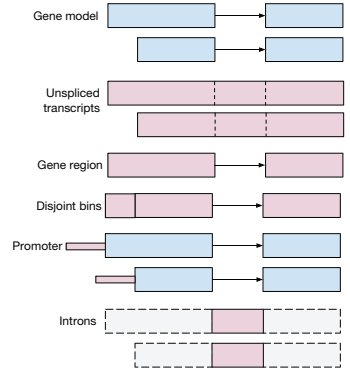


Figure 1 | Example uses of the Ranges algebra. A *GRanges* object, `g` (top), represents two transcript isoforms of a gene, each with two exons. The coordinates of unspliced transcripts are identified with the function `range(g)`. Calculating the gene region involves flattening the gene model into its constituent exons and reducing these to nonoverlapping ranges, `reduce(unlist(g))`. Ranges defining disjoint bins, `disjoin(unlist(g))`, are useful in counting operations, e.g., in RNA-seq analysis. Putative promoter ranges are found using strand-aware range extension, `flank(range(g), width = 100)`. Elementary operations can be composed to succinctly execute queries such as `psetdiff(range(g), g)` for computing the intron ranges.

Orchestrating high-throughput genomic analysis with Bioconductor

Wolfgang Huber[1], Vincent J Carey[2,3], Robert Gentleman[4], Simon Anders[1], Marc Carlson[5], Benilton S Carvalho[6], Hector Corrada Bravo[7], Sean Davis[8], Laurent Gatto[9], Thomas Girke[10], Raphael Gottardo[11], Florian Hahne[12], Kasper D Hansen[13,14], Rafael A Irizarry[3,15], Michael Lawrence[4], Michael I Love[3,15], James MacDonald[16], Valerie Obenchain[5], Andrzej K Oleś[1], Hervé Pagès[5], Alejandro Reyes[1], Paul Shannon[5], Gordon K Smyth[17,18], Dan Tenenbaum[5], Levi Waldron[19] & Martin Morgan[5]