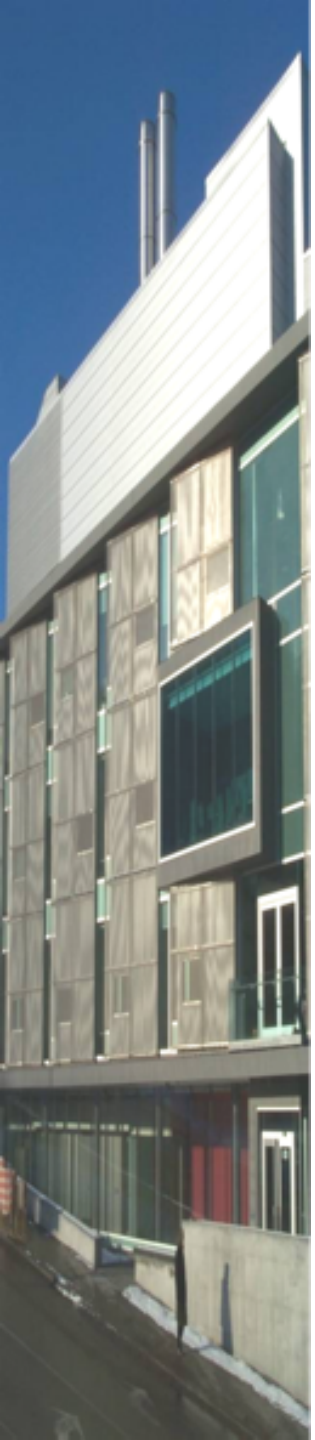
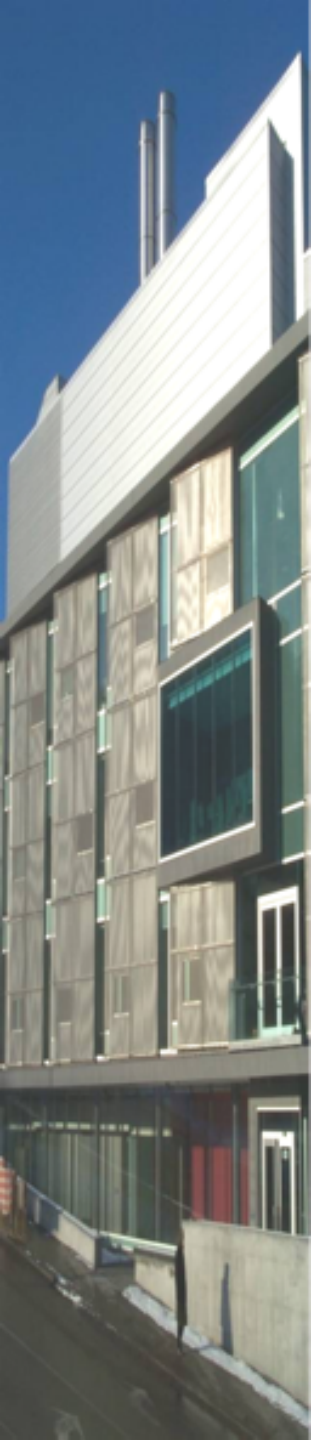


GenomeQuébec

RNAseq analysis

Bioinformatics Analysis Team
McGill University and Genome Quebec Innovation Center
bioinformatics.service@mail.mcgill.ca



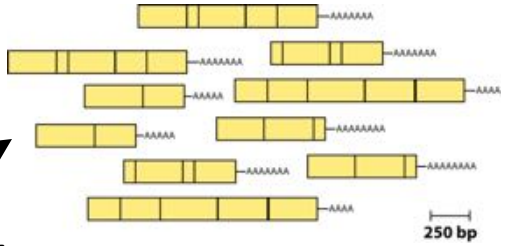
Why sequence RNA?

- Functional studies
  - Genome may be constant but experimental conditions have pronounced effects on gene expression
- Some molecular features can only be observed at the RNA level
  - Alternative isoforms, fusion transcripts, RNA editing
- Interpreting mutations that do not have an obvious effect on protein sequence
  - 'Regulatory' mutations
- Prioritizing protein coding somatic mutations (often heterozygous)

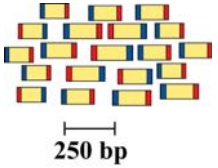
Modified from Bionformatics.ca

RNA-seq

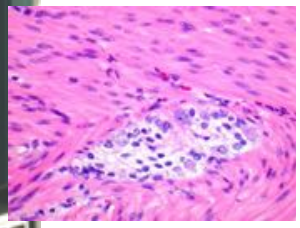
Isolate RNAs



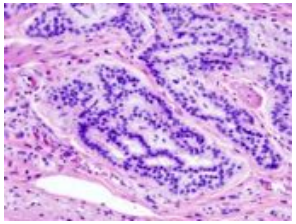
Generate cDNA, fragment, size select, add linkers



Samples of interest

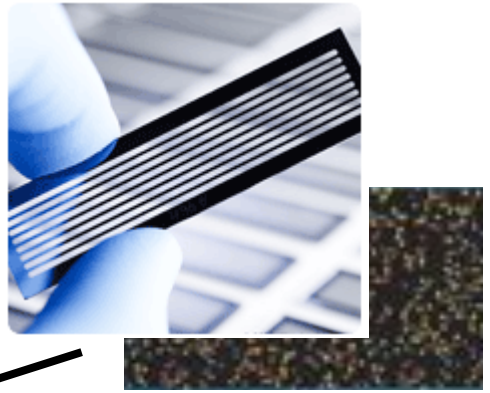


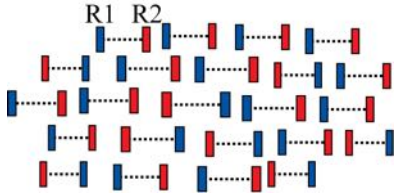
Condition 1
(normal colon)



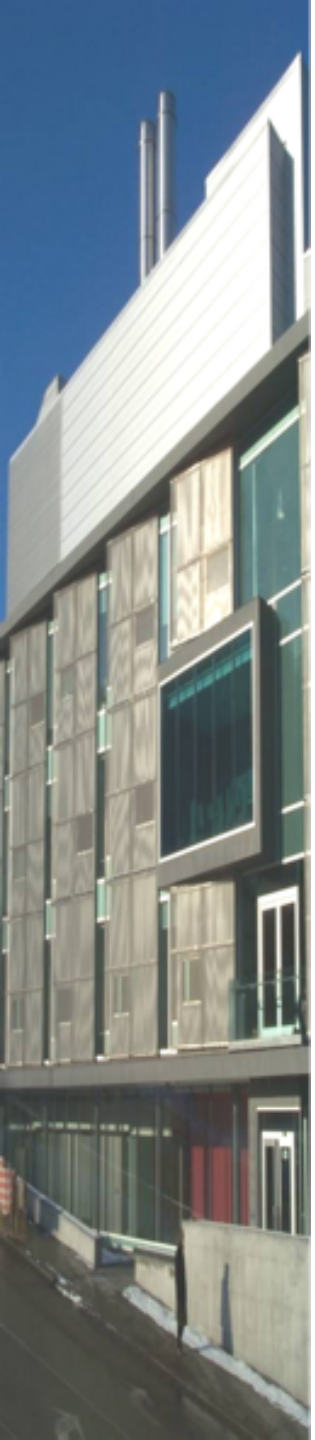
Condition 2
(colon tumor)

Sequence ends



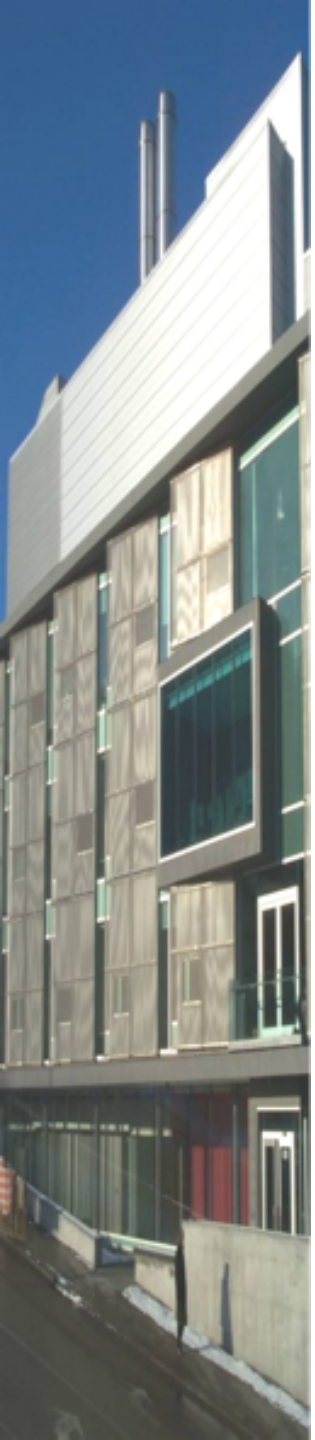


100s of millions of paired reads
10s of billions bases of sequence



RNA-seq – Applications

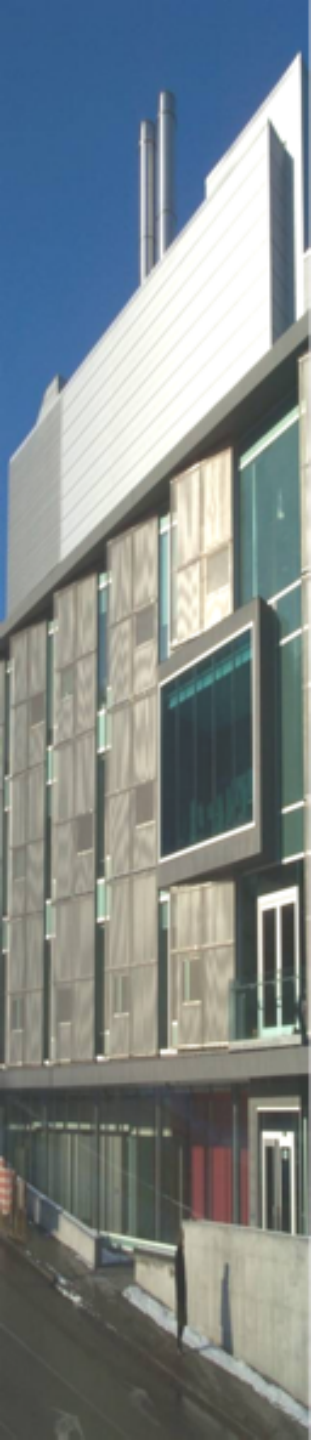
- Gene expression and differential expression
- Transcript discovery
- SNV, RNA-editing events, variant validation
- Allele specific expression
- Gene fusion events detection
- Genome annotation and assembly
- etc …



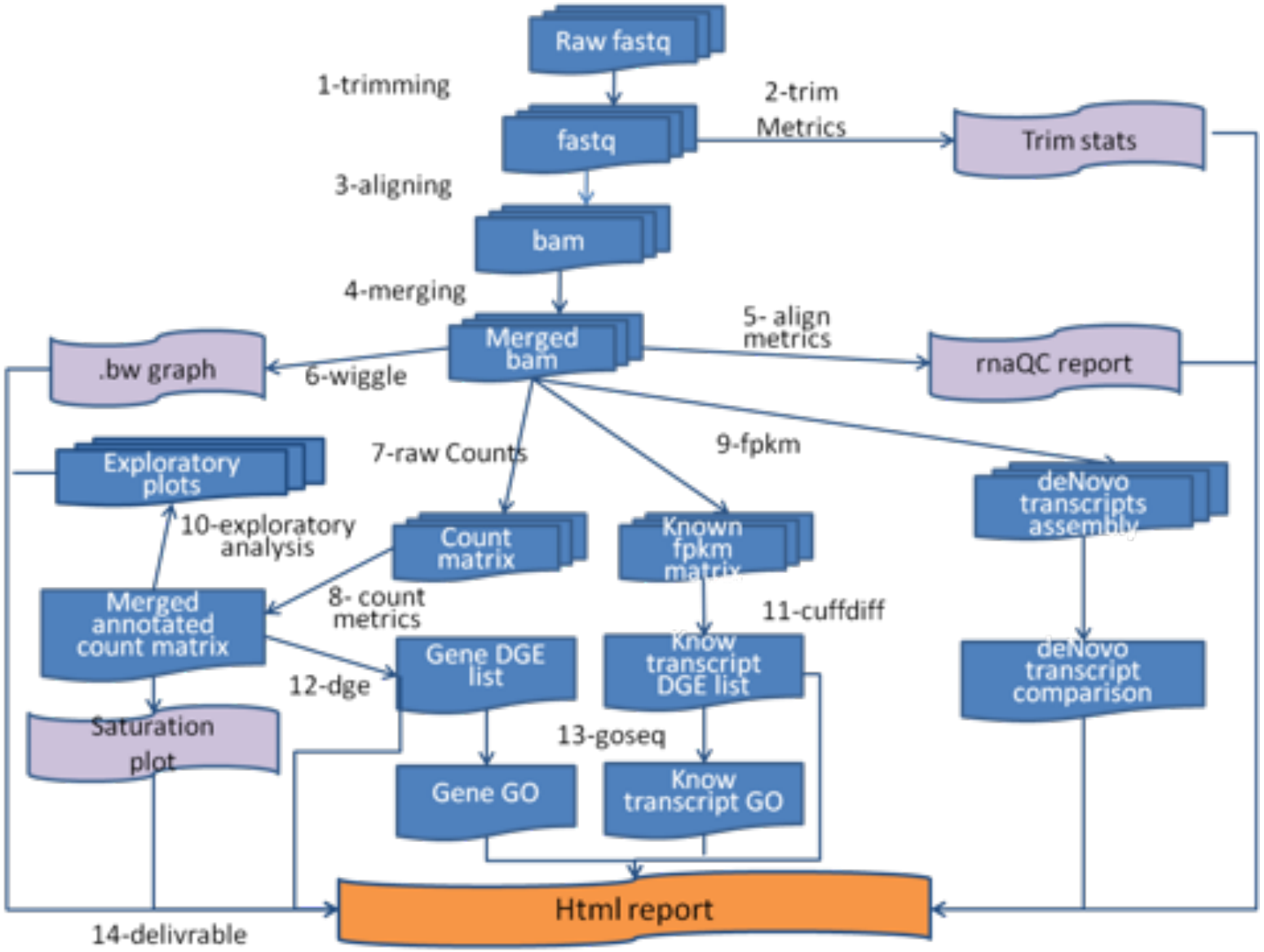
RNAseq Challenges

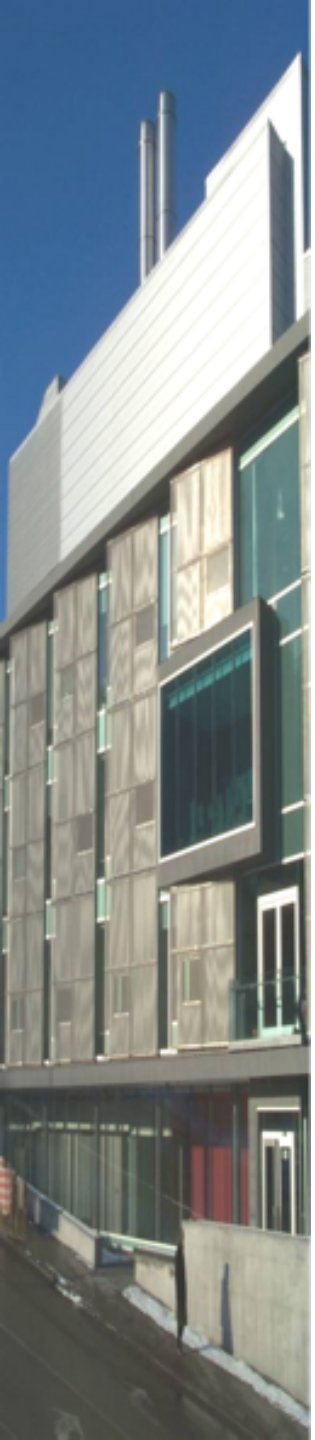
- RNAs consist of small exons that may be separated by large introns
  - Mapping splice-reads to the genome is challenging
  - Ribosomal and mitochondrial genes are misleading
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
- RNA is fragile and easily degraded
  - Low quality material can bias the data

Modified from Bionformatics.ca

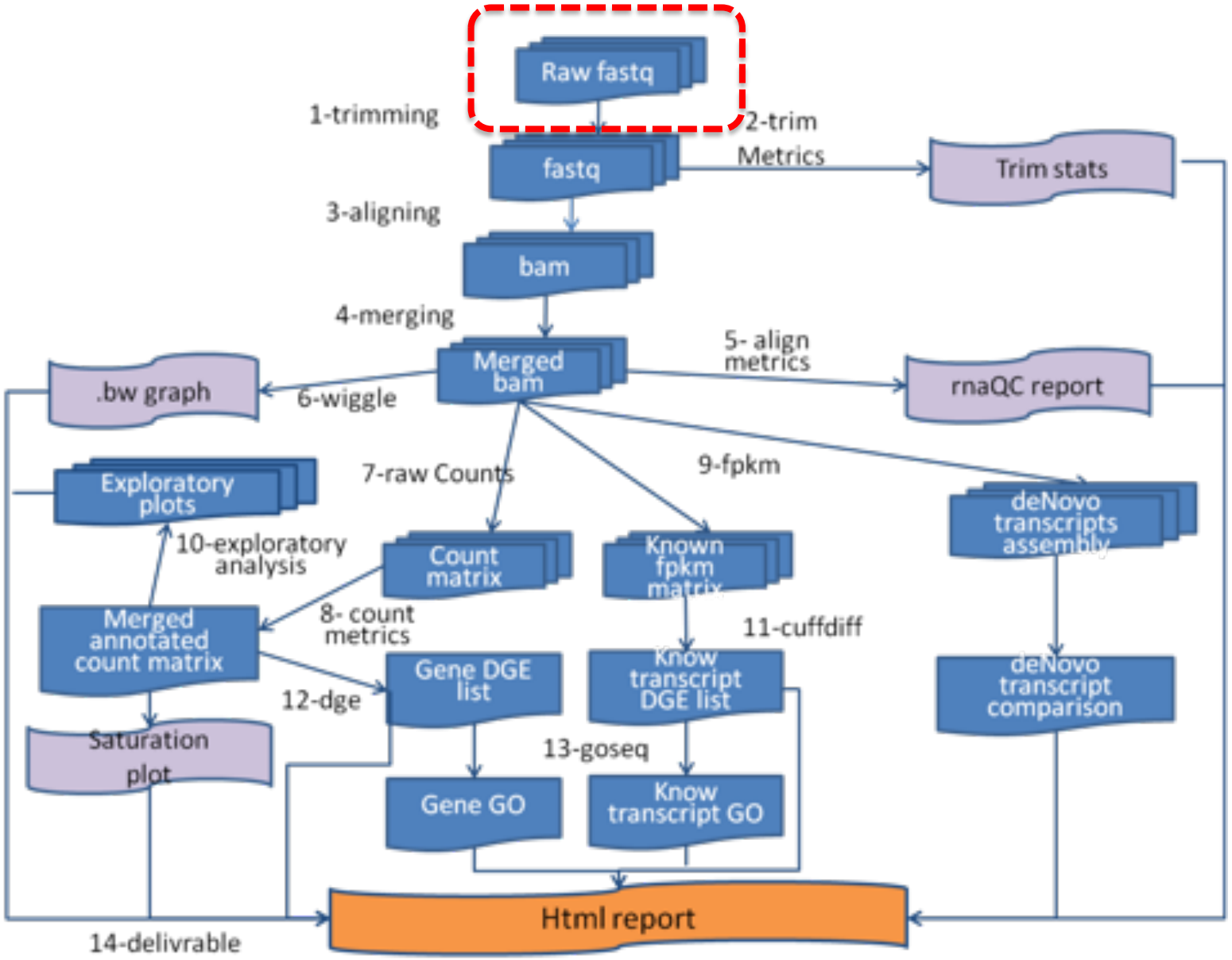


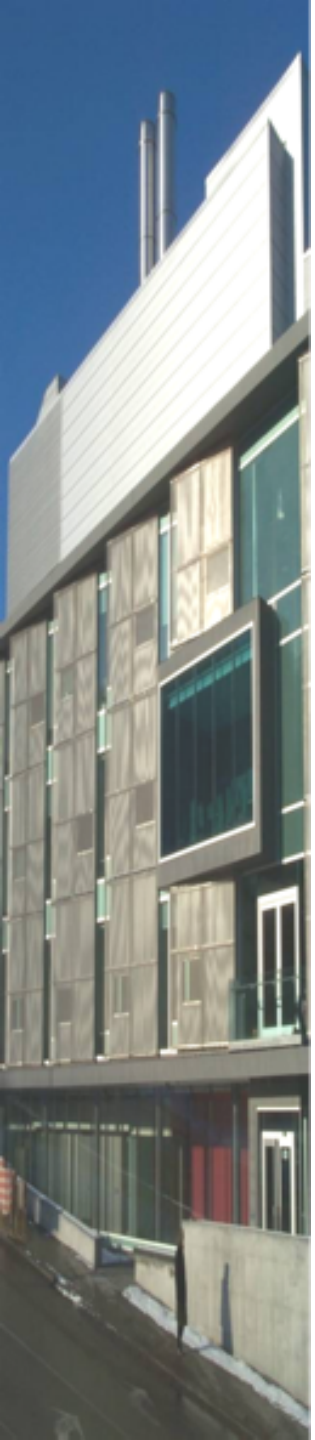
RNA-Seq: Overview





RNA-Seq: Input Data





Input Data: FASTQ

End 1

Control1_R1.fastq.gz

Control2_R1.fastq.gz

KnockDown1_R1.fastq.gz

KnockDown2_R1.fastq.gz

End 2

Control1_R2.fastq.gz

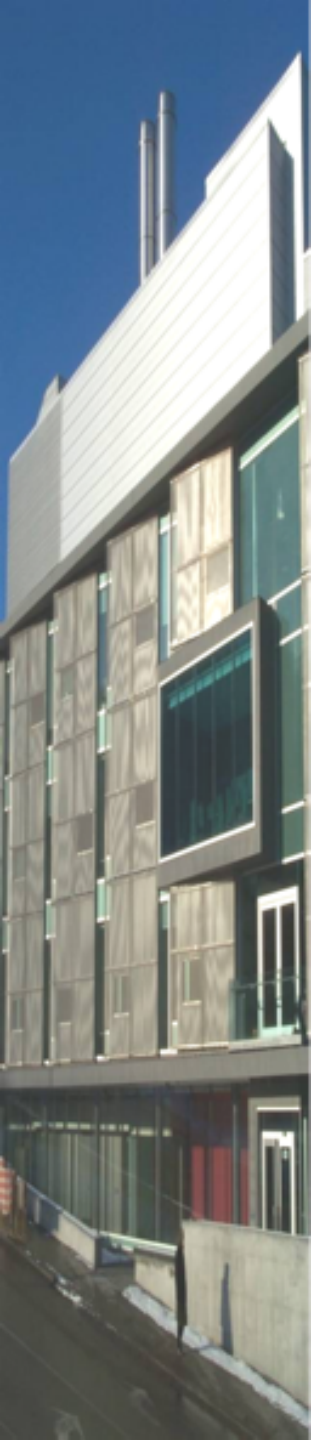
Control2_R2.fastq.gz

KnockDown1_R2.fastq.gz

KnockDown2_R2.fastq.gz

~ 10Gb each sample

```
@ERR127302.1 HWI-EAS350_0441:1:1:1055:4898#0/1
GGCTCATCTTGAACTGGGTGGCGACCGTCCCTGGCCCCTTCTTGACACCCA
+
4=B@D99BDDDDDDD:DD?B<<=?>6B##########################
```

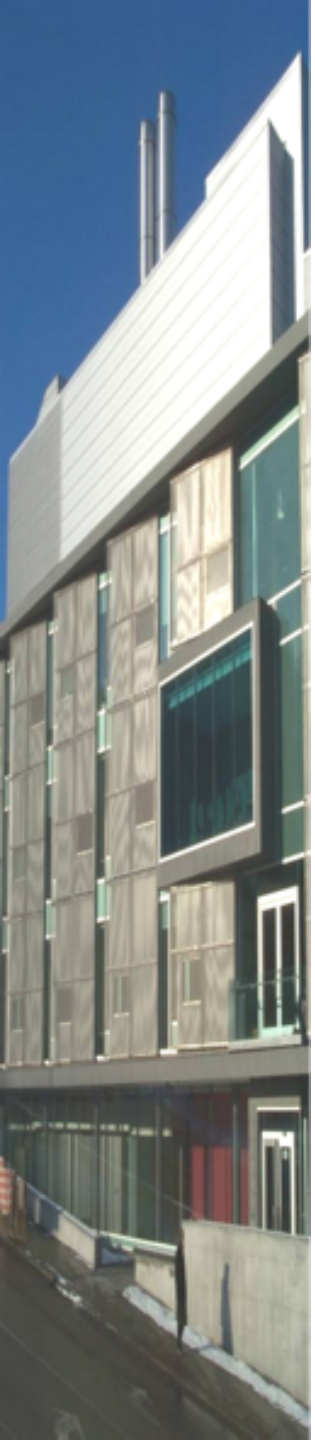


$$Q = -10 \log\_10 (p)$$

Where Q is the quality and p is the probability of the base being incorrect.

What is a base quality?

Base Quality	P_{error}(obs. base)
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %



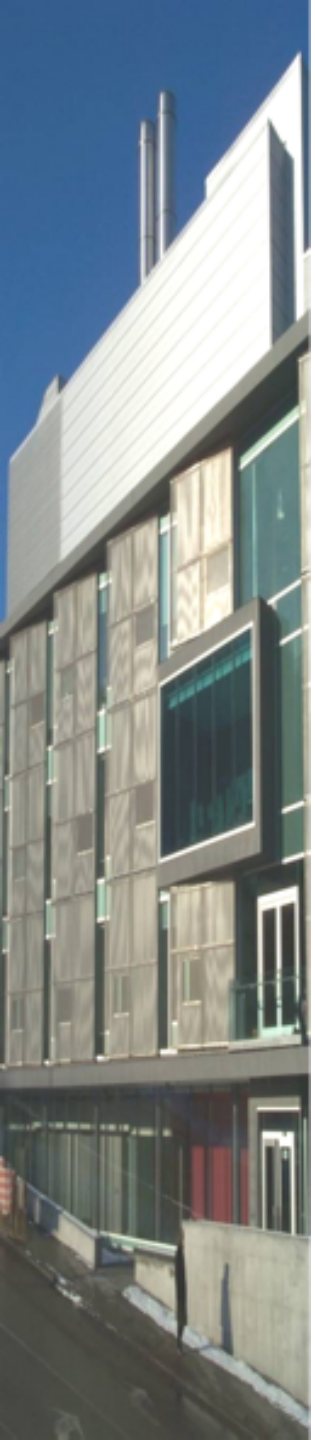
QC of raw sequences

Project Details | Samples (41) | Libraries (32) | HiSeq Read Sets (64) | Read Sets Search | Documents (0) | Assemblies (0) | Uploaded Analyses (0)

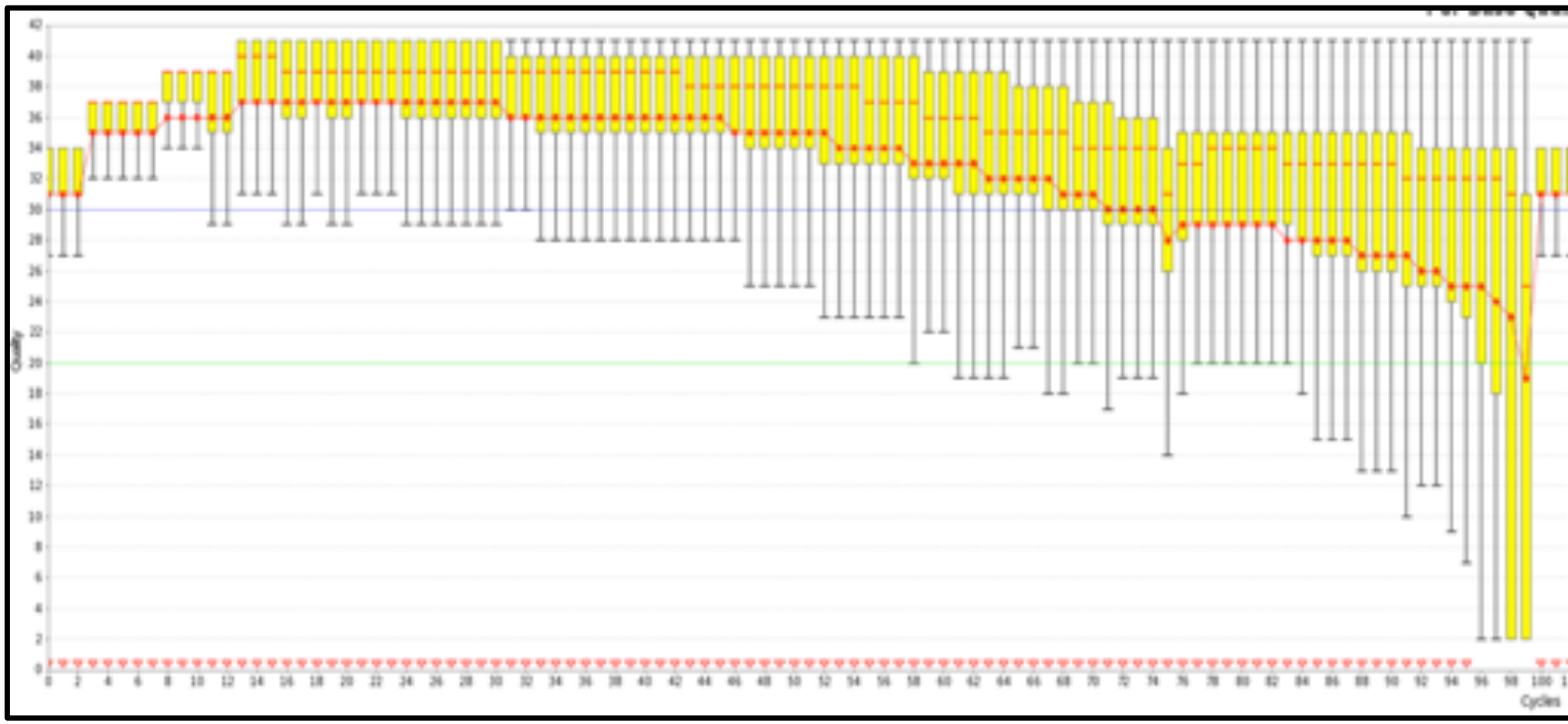
CSV | View/Set Filter | Download Read Files | Help with icons

Read Sets (64 elements) — Add/Remove Column

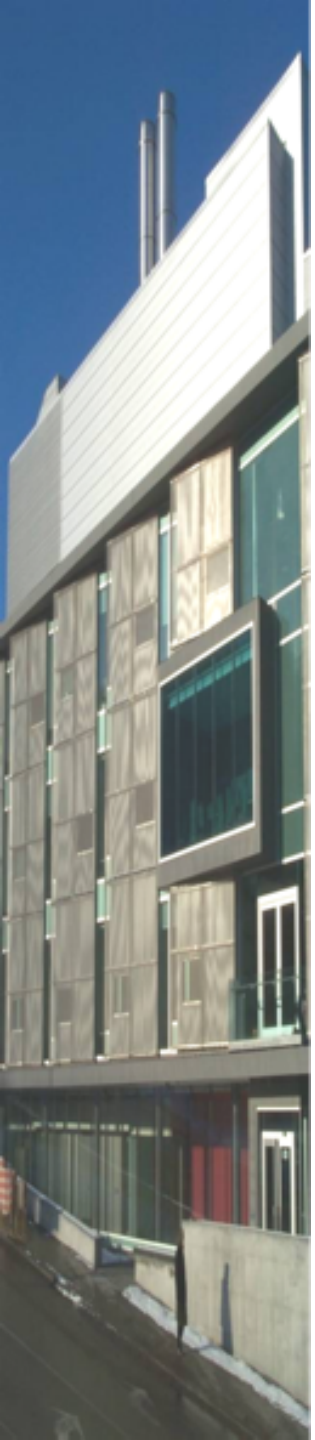
Name	Multiplex Key	Run	Region	QC	Status	Number of reads	Number of Bases	Average Quality	% Duplicate	% Passed Filter	Reads Fastq R1	Reads Fastq R2
W24P	Index_7	1177	4	QC		45,373,280	9,074,656,000	33	21.674	100	(4562MB)	(4546MB)
W25P	Index_8	1177	4	QC		45,066,800	9,013,360,000	33	17.943	100	(4527MB)	(4513MB)
W29P1	Index_9	1177	4	QC		70,319,214	14,063,842,800	33	17.51	100	(7061MB)	(7038MB)
W16P1	Index_6	1177	4	QC		55,160,915	11,032,183,000	33	14.447	100	(5553MB)	(5529MB)
W29P1	Index_9	1177	3	QC		70,276,618	14,055,323,600	33	17.58	100	(7029MB)	(7012MB)
W25P	Index_8	1177	3	QC		45,097,360	9,019,472,000	33	18.036	100	(4512MB)	(4503MB)
W24P	Index_7	1177	3	QC		45,502,426	9,100,485,200	33	21.815	100	(4557MB)	(4545MB)
W16P1	Index_6	1177	3	QC		55,290,201	11,058,040,200	33	14.542	100	(5545MB)	(5527MB)



QC of raw sequences

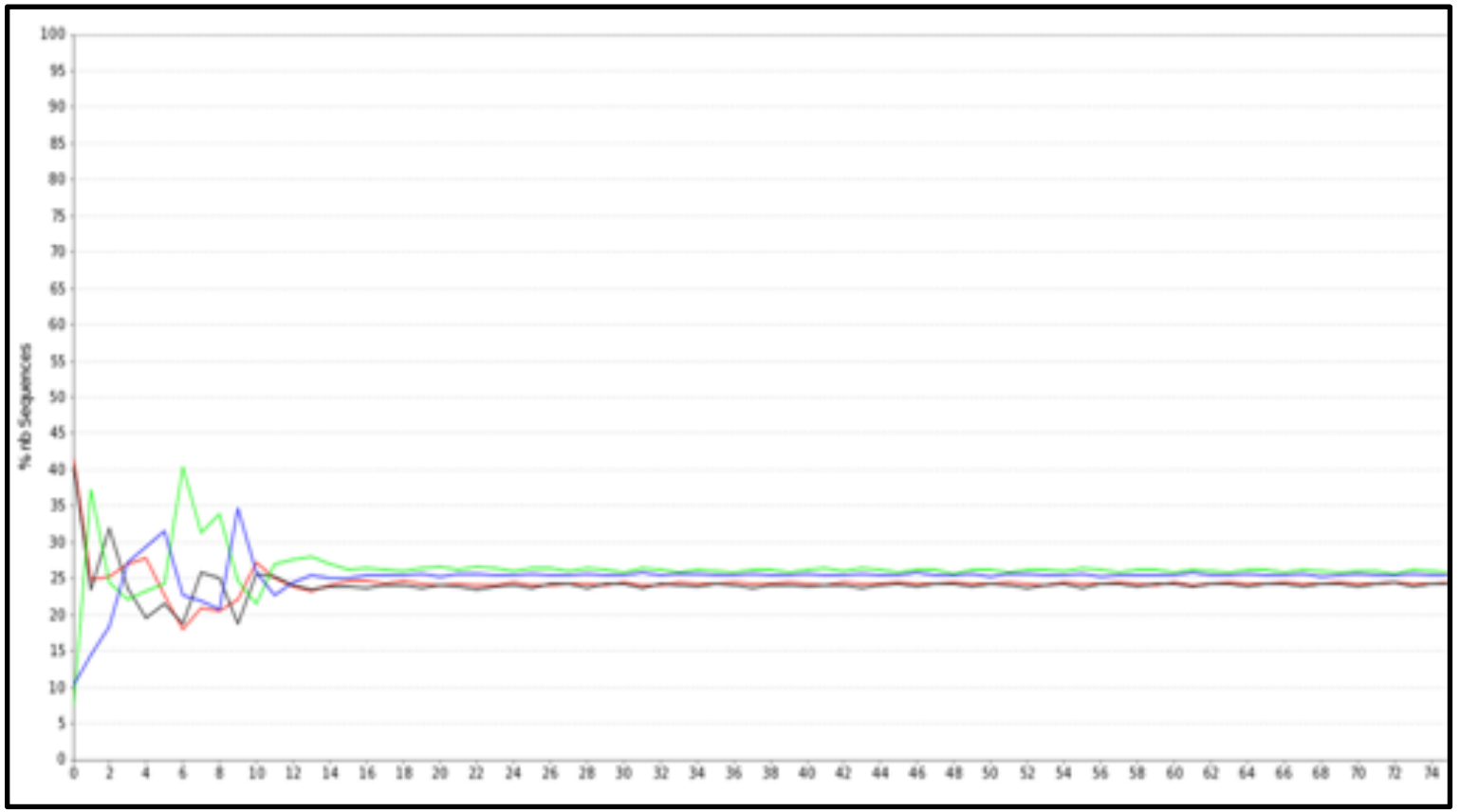


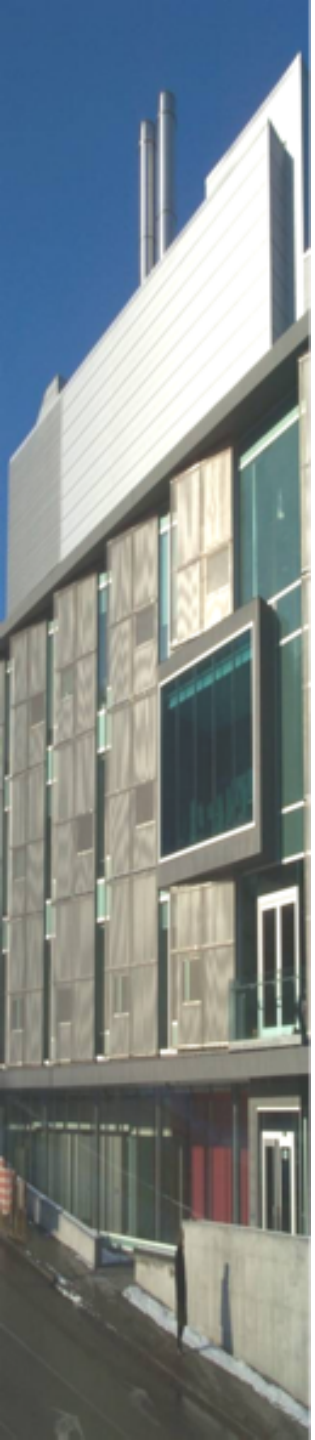
low qualtity bases can bias subsequent anlaysis
(i.e, SNP and SV calling, …)



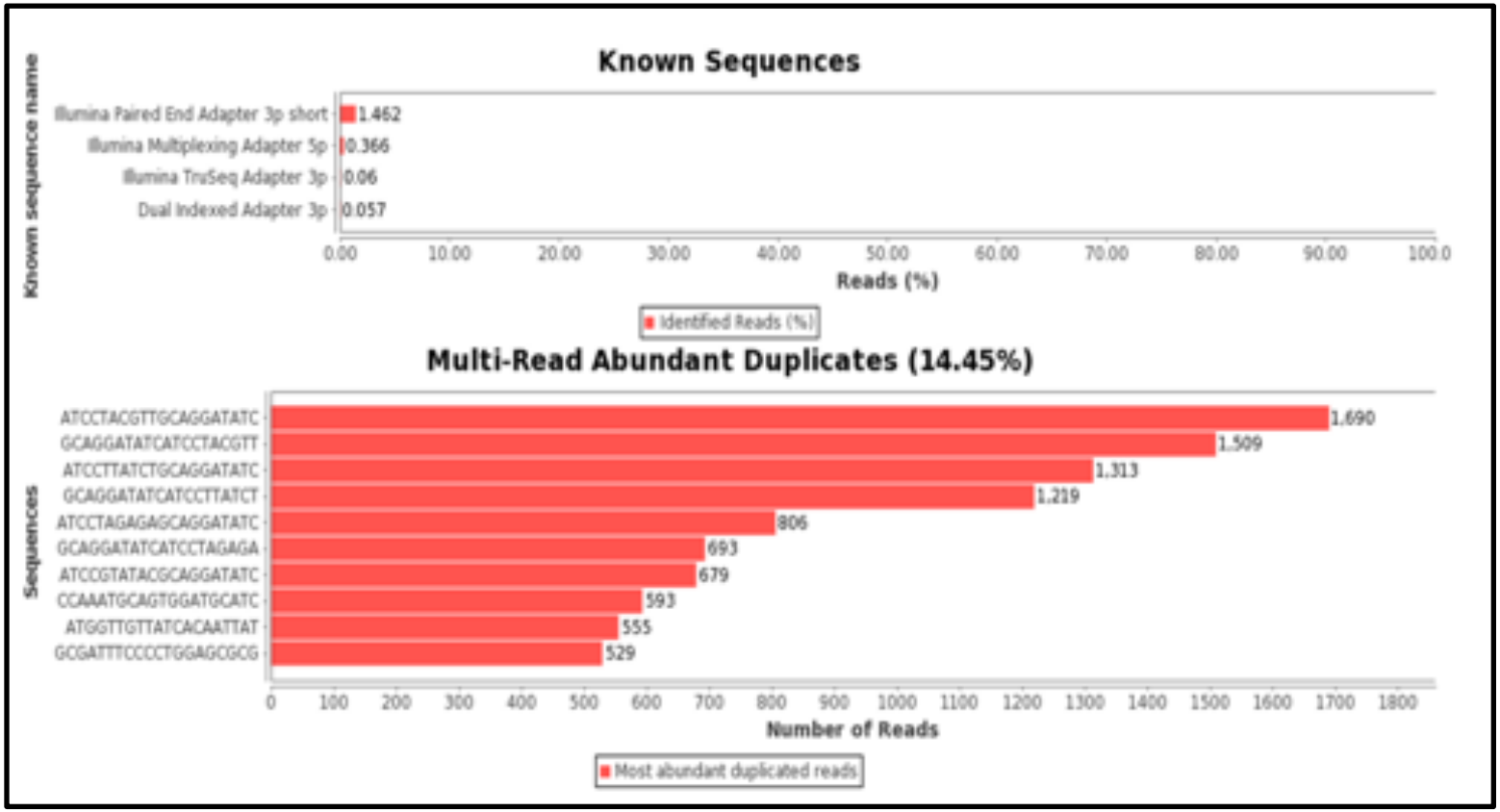
QC of raw sequences

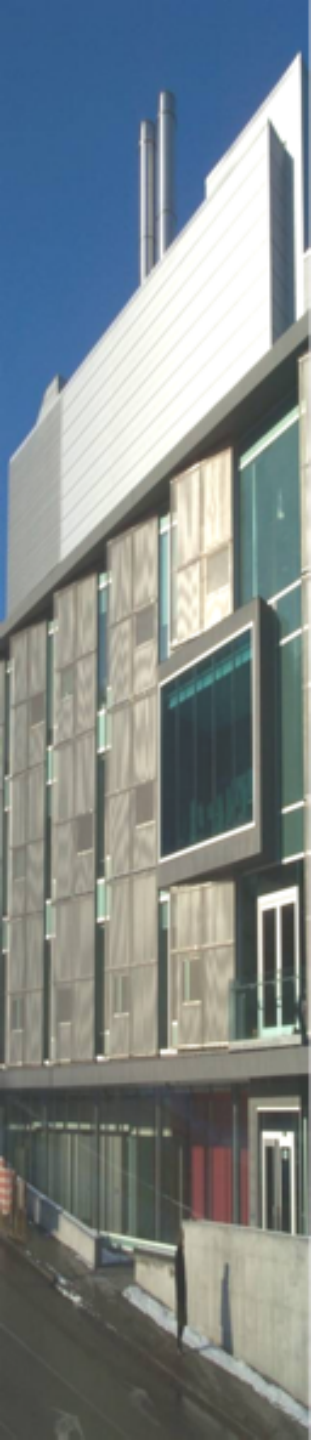
Positional Base-Content





QC of raw sequences



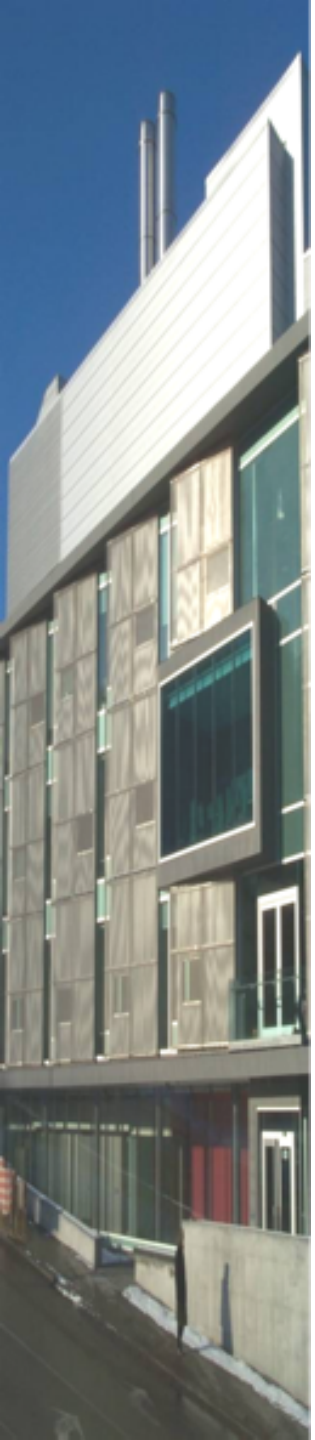


QC of raw sequences

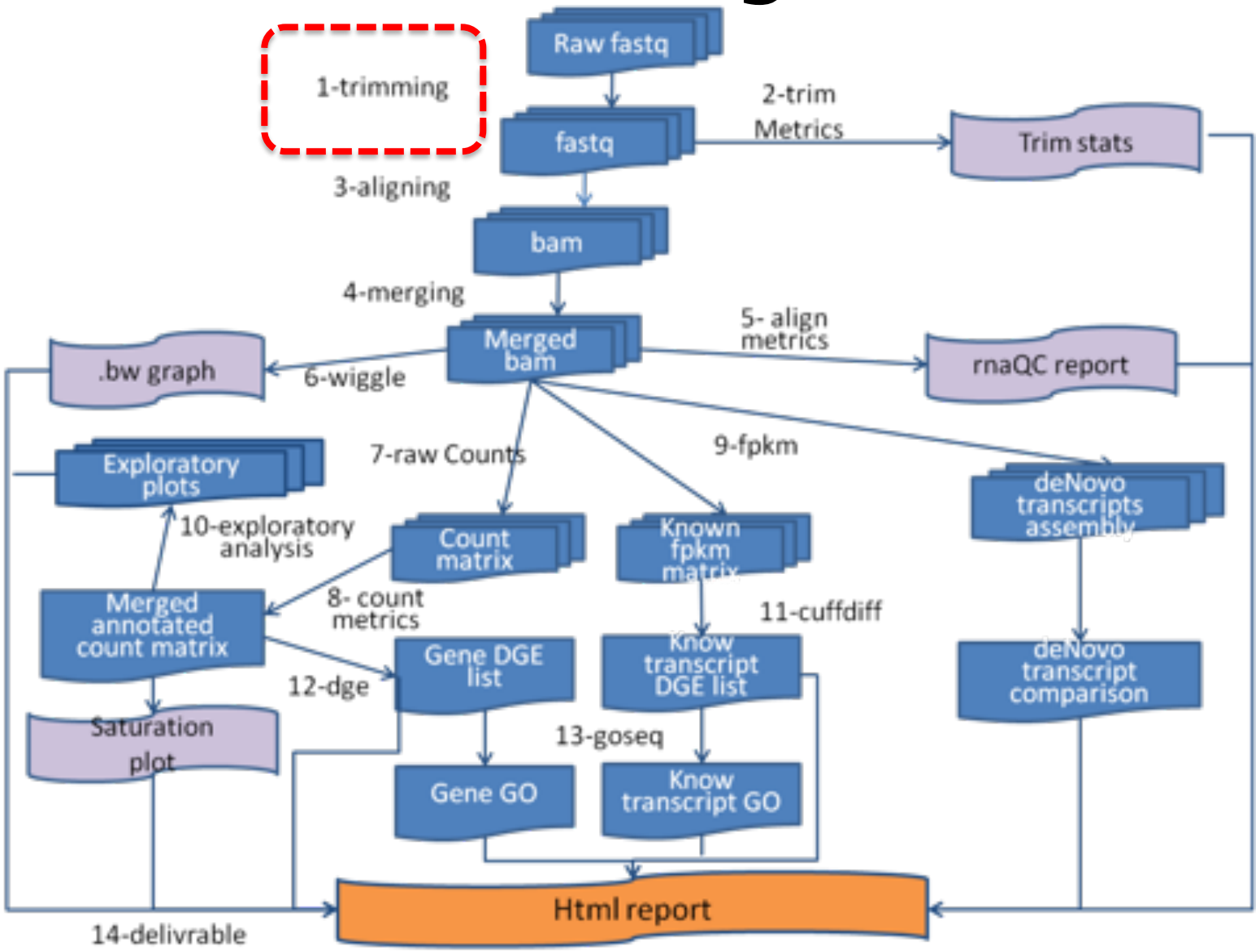
Species composition (via BLAST)

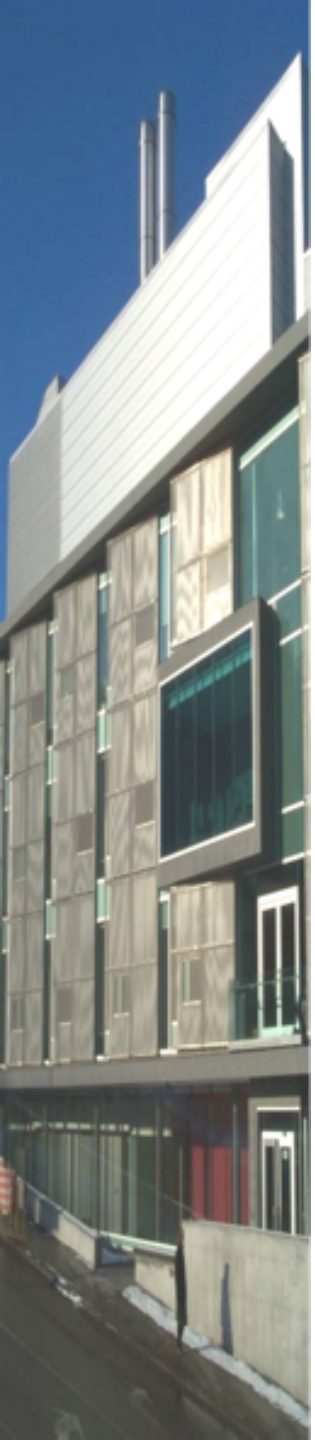
Blast Results (20 elements)

	Species	Hit Count
1	Mus_musculus	89,696
2	PREDICTED:_Mus	2,898
3	Mouse_DNA	1,579
4	TSA:_Anolis	1,217
5	Synthetic_construct	1,202
6	Rattus_norvegicus	571
7	PREDICTED:_Rattus	463
8	PREDICTED:_Dasypus	245
9	PREDICTED:_Cricetulus	238
10	PREDICTED:_Ceratotherium	140
11	Xenopus_laevis	97
12	TSA:_Nannochloropsis	74
13	Human_DNA	65
14	Trachemys_scripta	61
15	Chain_2,	55
16	TSA:_Nothobranchius	54
17	PREDICTED:_Odobenus	40
18	PREDICTED:_Nomascus	38
19	Chain_5,	37
20	Mus_musculus,	31



RNA-Seq: Trimming and Filtering

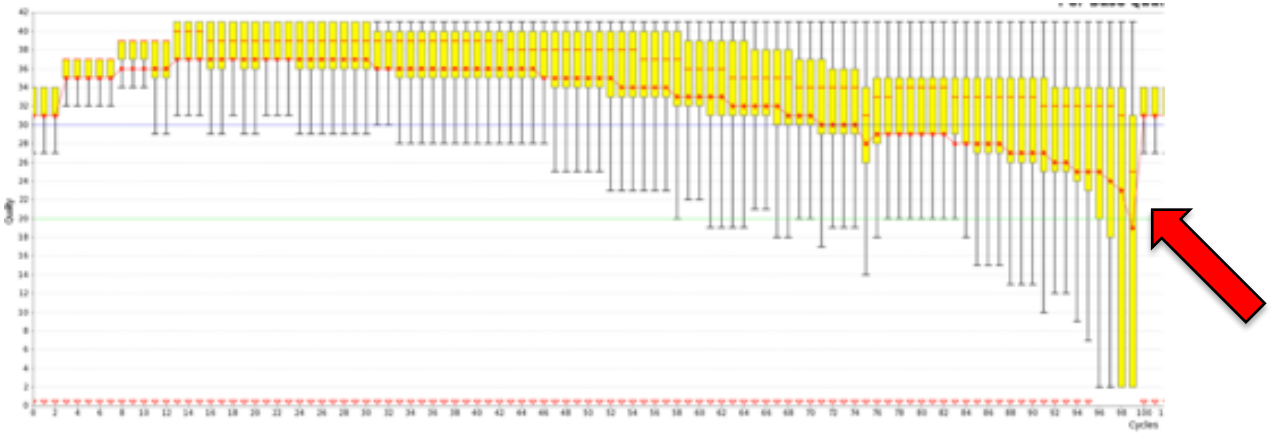




Read Filtering

- Clip Illumina **adapters**:

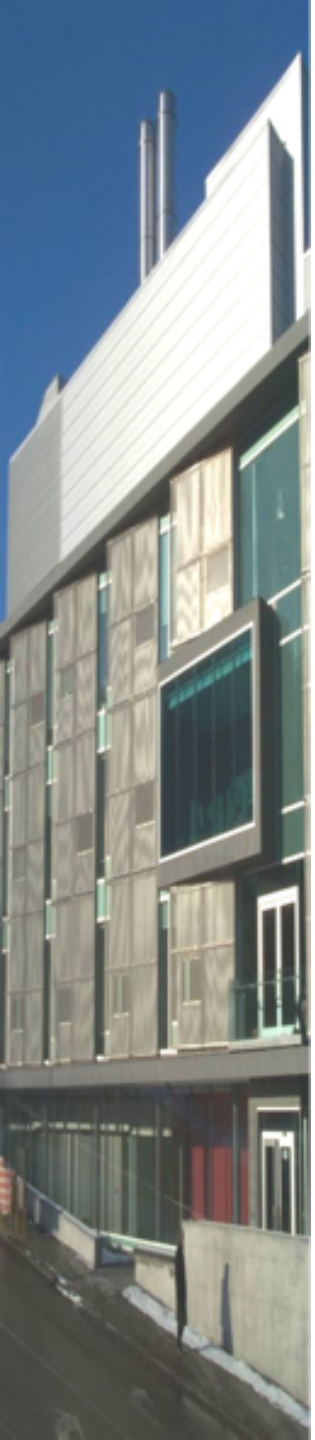
- Trim trailing **quality** < 30



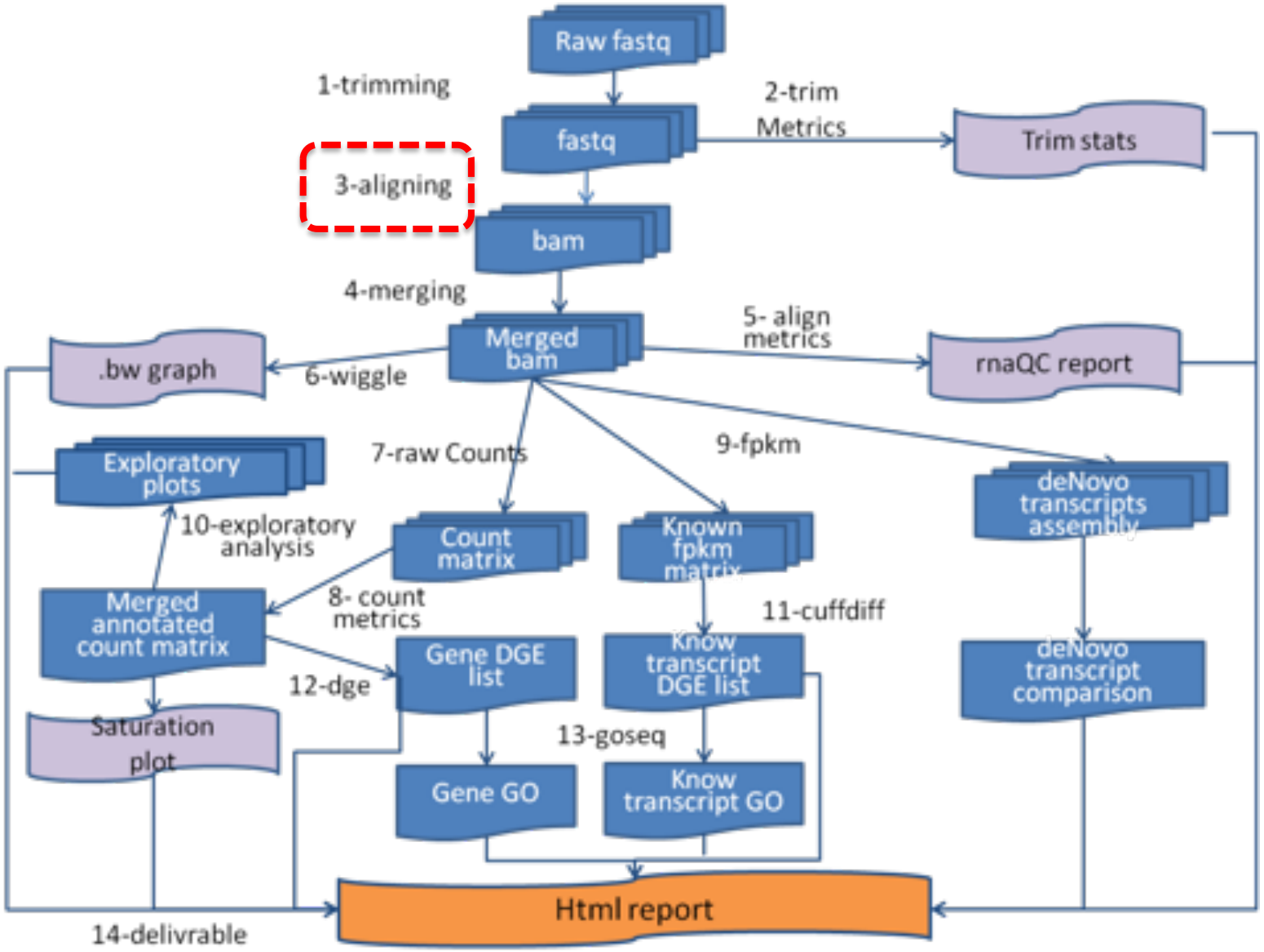
- Filter for read **length** ≥ 32 bp

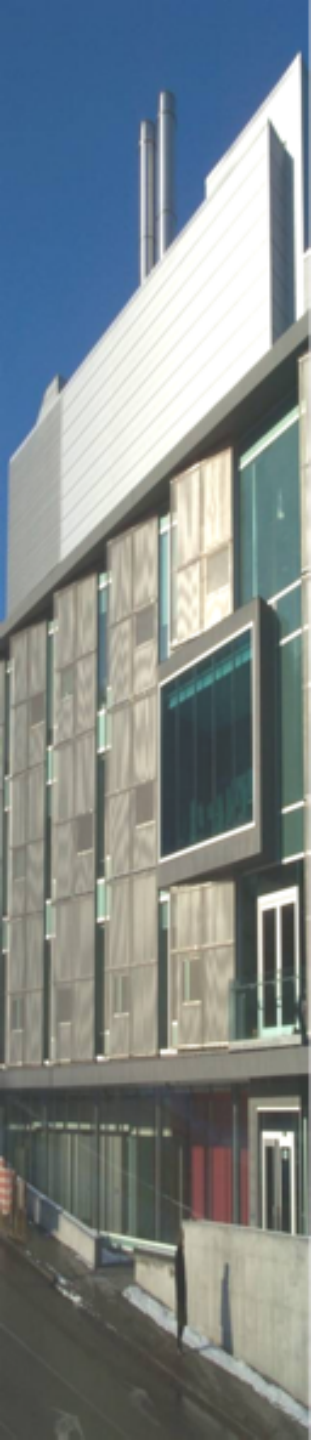
Trimmomatic

usadellab.org

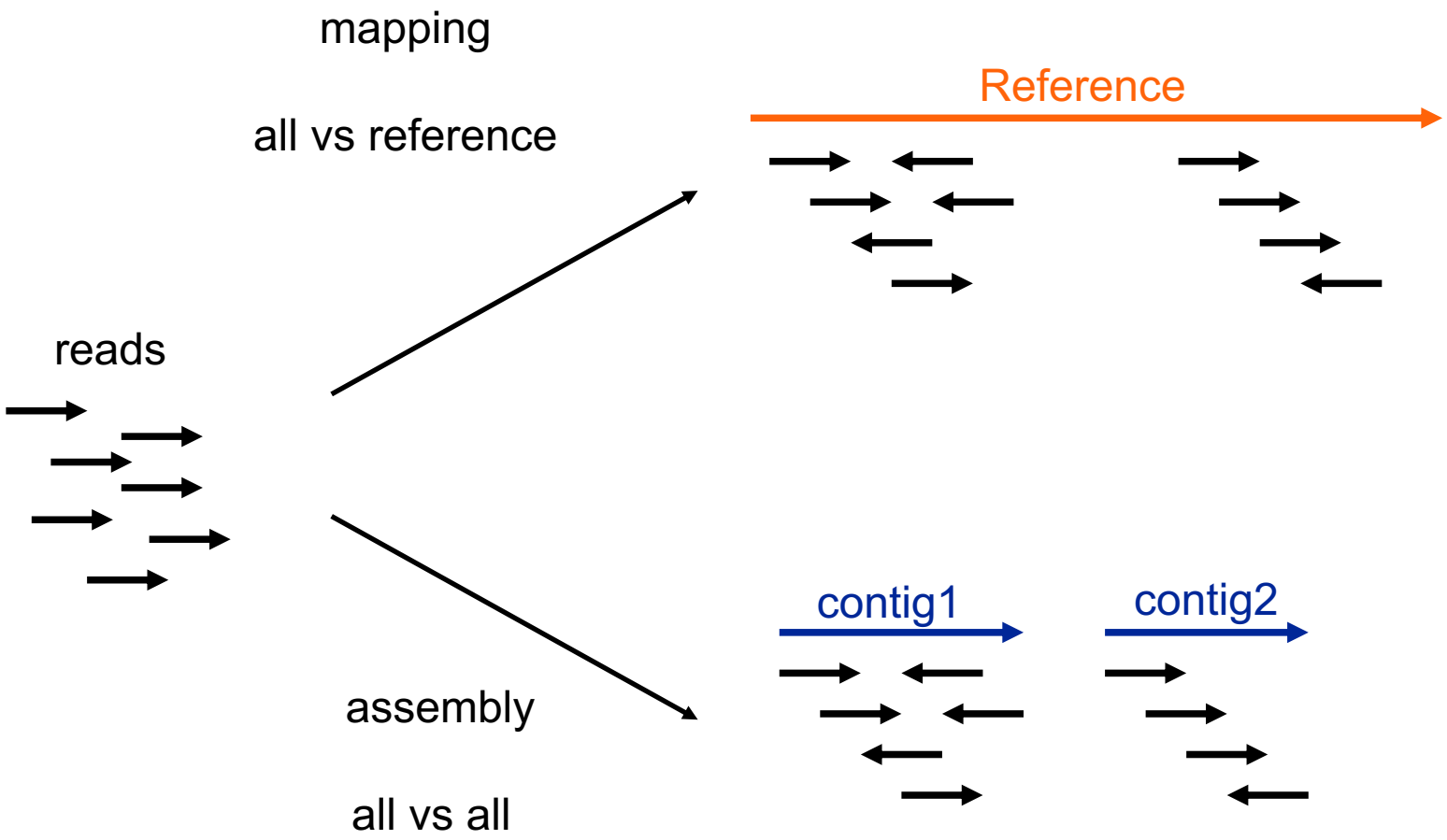


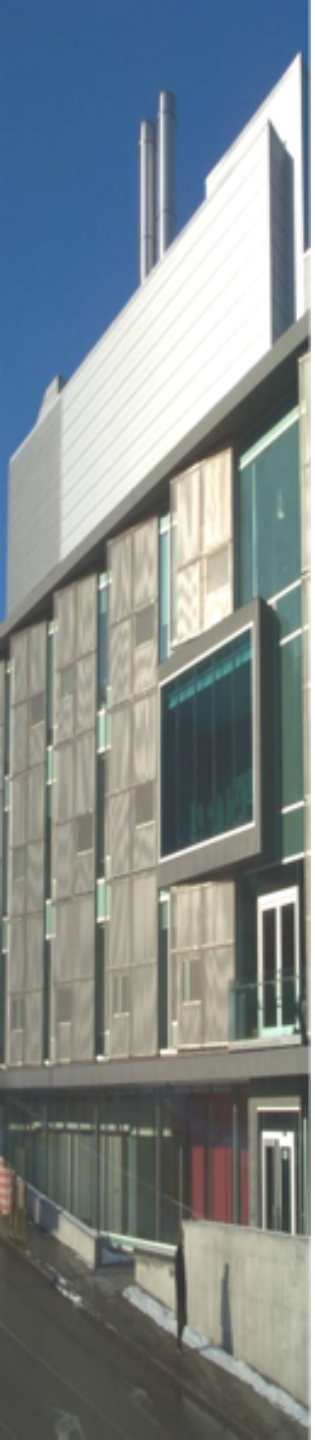
RNA-Seq: Mapping



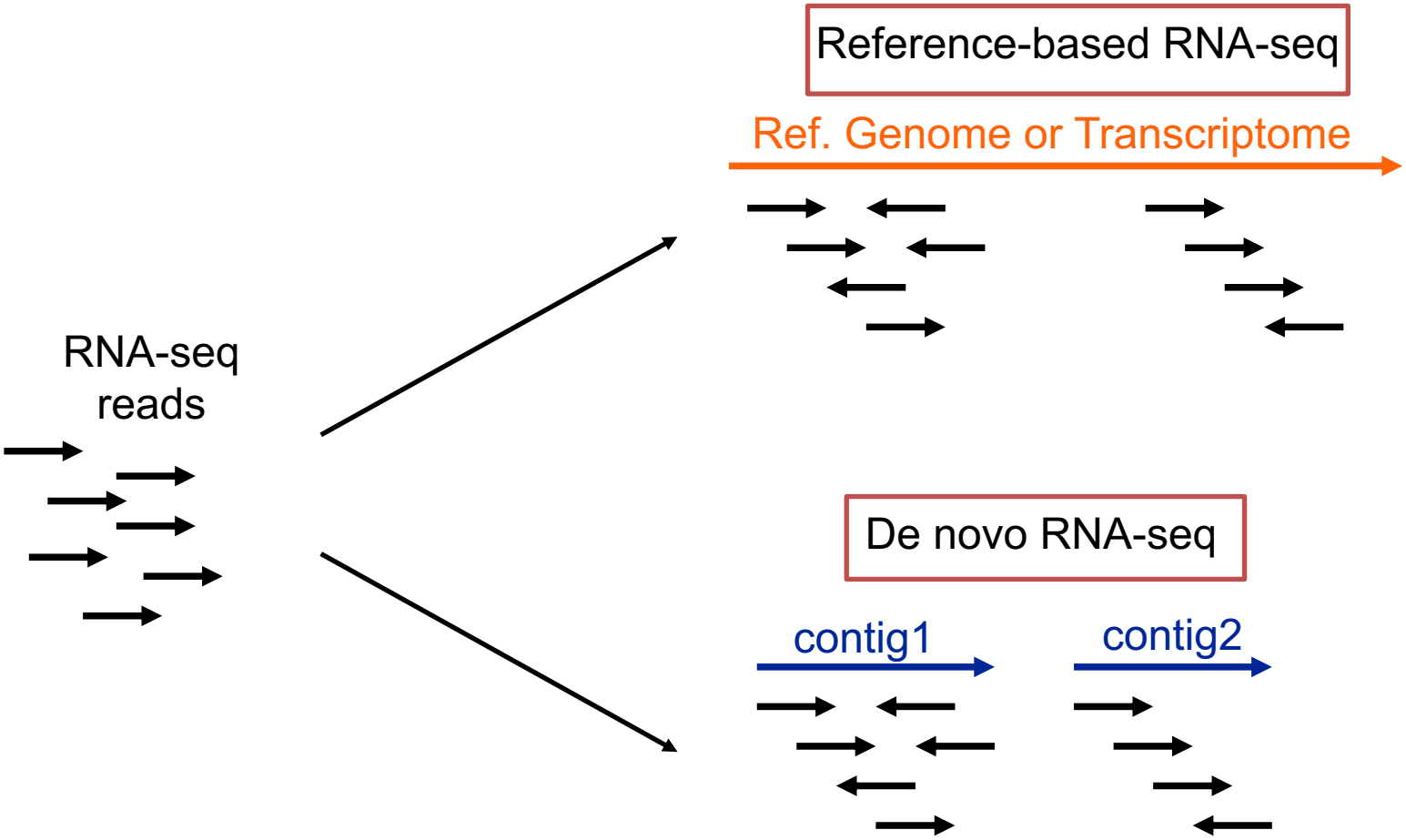


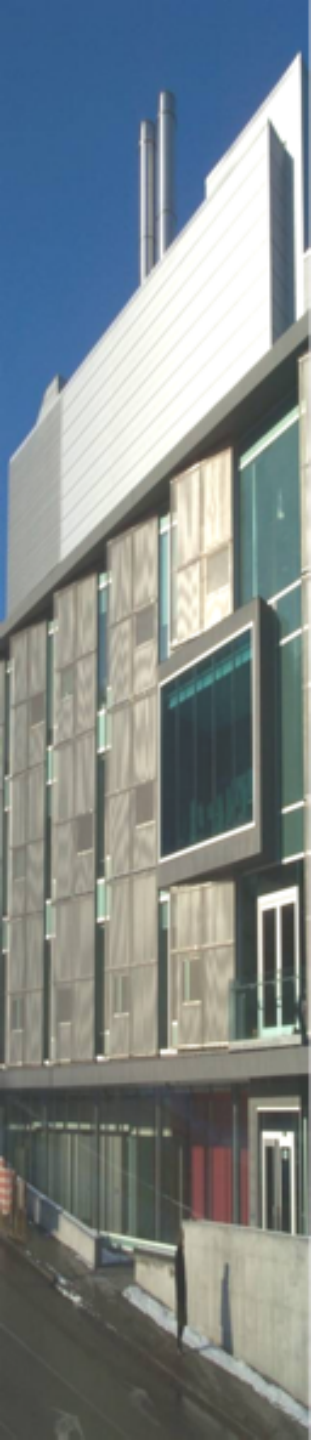
Assembly vs. Mapping





RNA-seq: Assembly vs Mapping





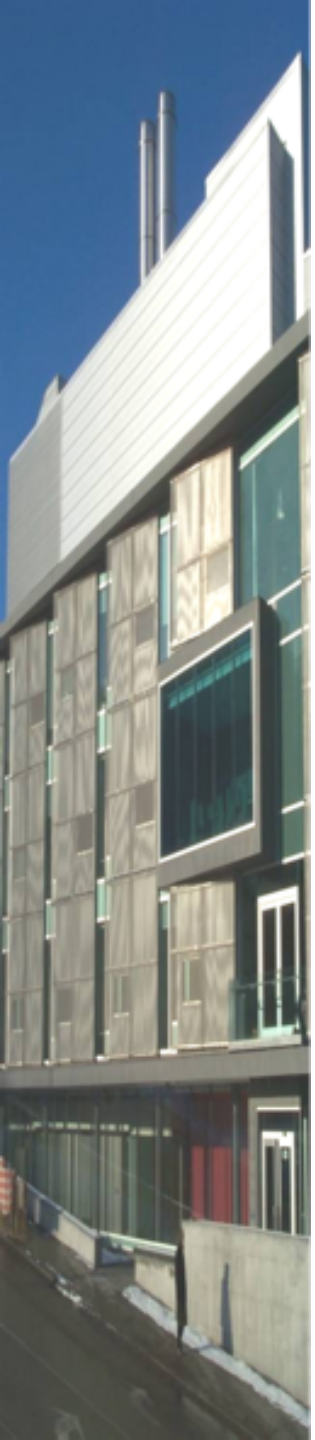
Read Mapping

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome
Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

- Mapping problem is challenging:
  - Need to map millions of short reads to a genome
  - Genome = text with billons of letters
  - Many mapping locations possible
  - NOT exact matching: sequencing errors and biological variants (substitutions, insertions, deletions, splicing)
- Clever use of the **Burrows-Wheeler Transform** increases speed and reduces memory footprint
- Other mappers: BWA, Bowtie, STAR, GEM, etc.

Bowtie alignment performance versus SOAP and Maq

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
Bowtie	PC	16 m 41 s	17 m 57 s	29.5	1,353	-	71.9
Maq		17 h 46 m 35 s	17 h 53 m 7 s	0.49	804	59.8×	74.7



TopHat: Spliced Reads

- Bowtie-based
- TopHat: finds/maps to possible splicing junctions.
- Important to assemble transcripts later (cufflinks)


Kim et al. Genome Biology 2013, 14:R36
http://genomebiology.com/2013/14/4/R36
Genome Biology
METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

SAM/BAM

Control1.bam

Control2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

KnockDown1.bam

KnockDown2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

~ 10Gb each bam

- Used to store alignments
- SAM = text, BAM = binary

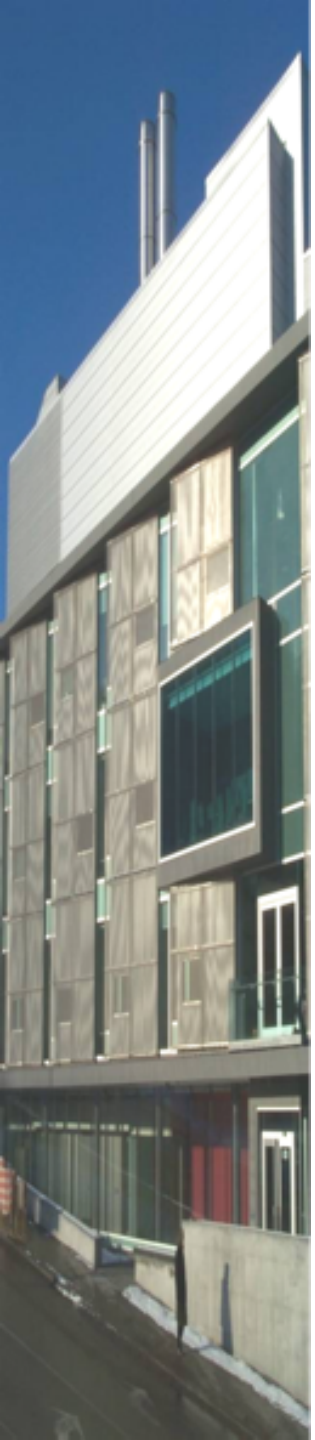
Read name | Flag | Reference Position | CIGAR | Mate Position

```
SRR013667.1 99 19 8882171 60 76M = 8882214 119
NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAGCAGA
#>A@BABAAAAADDEGCEFDHDEDBCFDBCDBCBDCEACB>AC@CDB@>
...
```

Bases

Base Qualities

SAM: Sequence Alignment/Map format



The BAM/SAM format

SAMtools

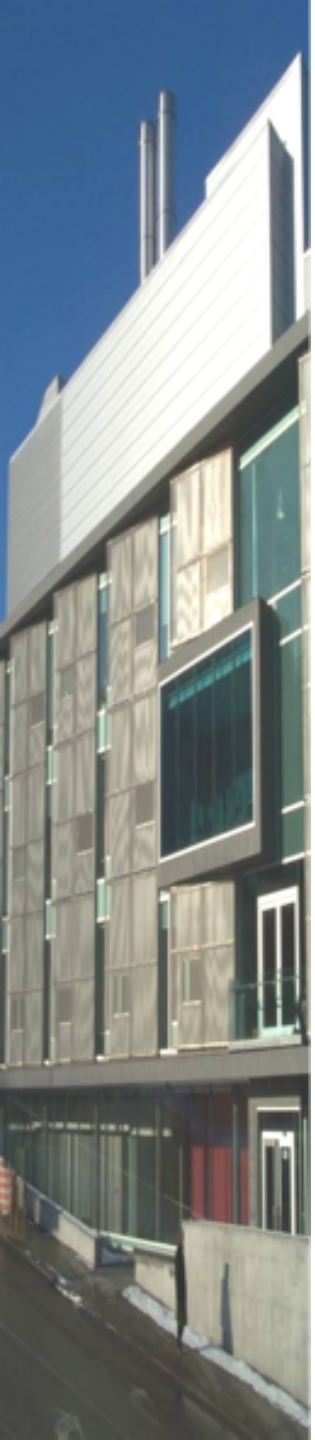
samtools.sourceforge.net

Picard

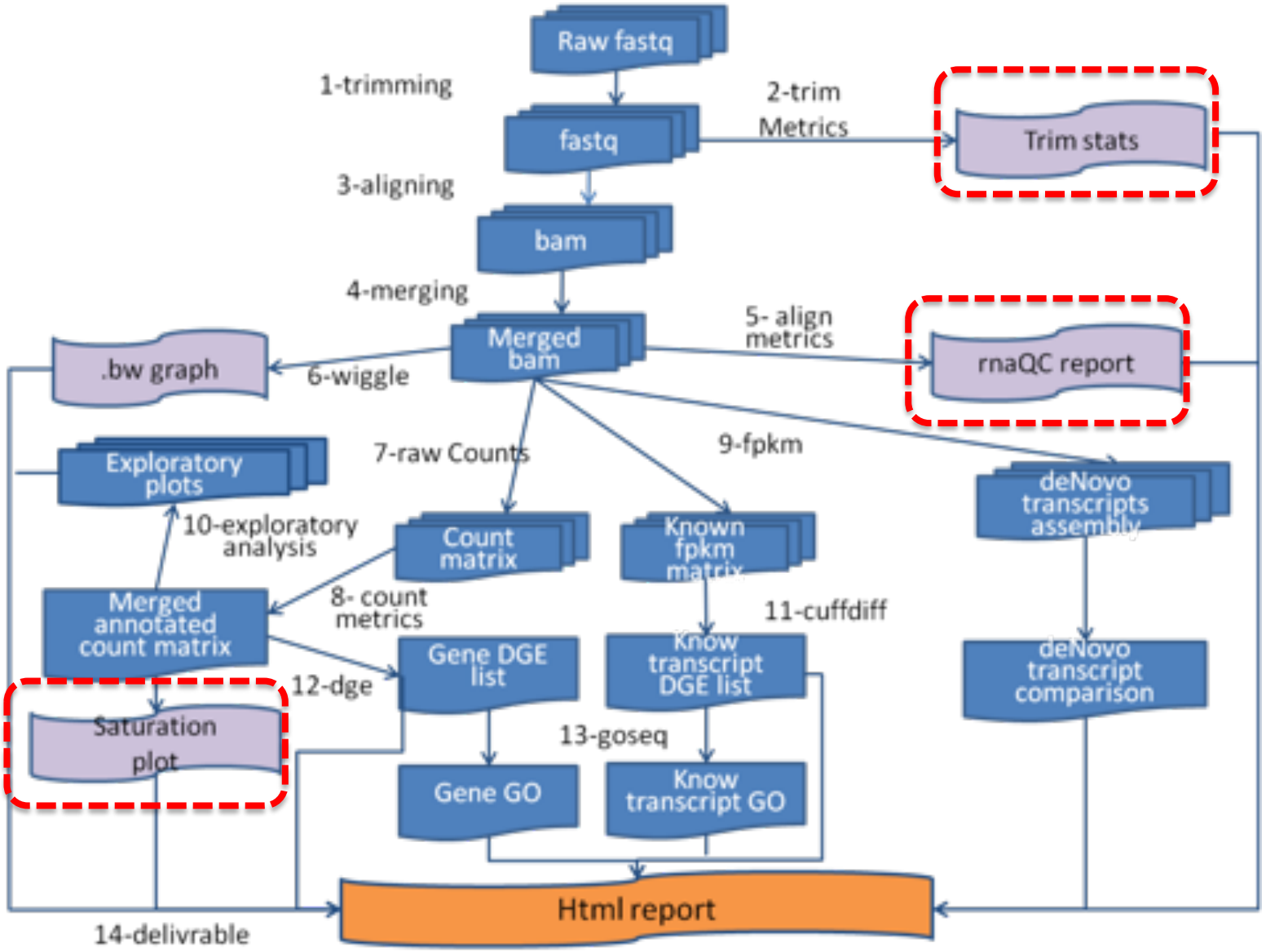
picard.sourceforge.net

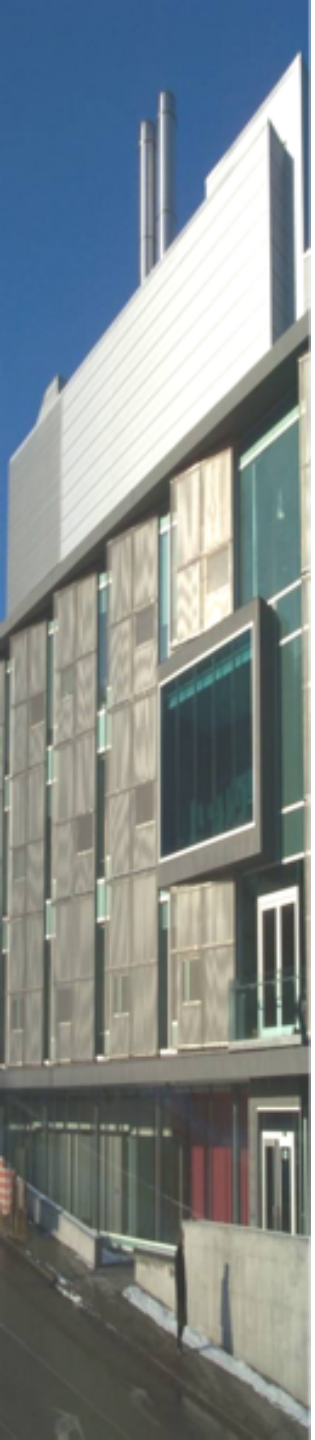
Sort, View, Index, Statistics, Etc.

```
$ samtools flagstat C1.bam
110247820 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
110247820 + 0 mapped (100.00%:nan%)
110247820 + 0 paired in sequencing
55137592 + 0 read1
55110228 + 0 read2
93772158 + 0 properly paired (85.06%:nan%)
106460688 + 0 with itself and mate mapped
3787132 + 0 singletons (3.44%:nan%)
1962254 + 0 with mate mapped to a different chr
738766 + 0 with mate mapped to a different chr (mapQ>=5)
$
```



RNA-Seq: Alignment QC





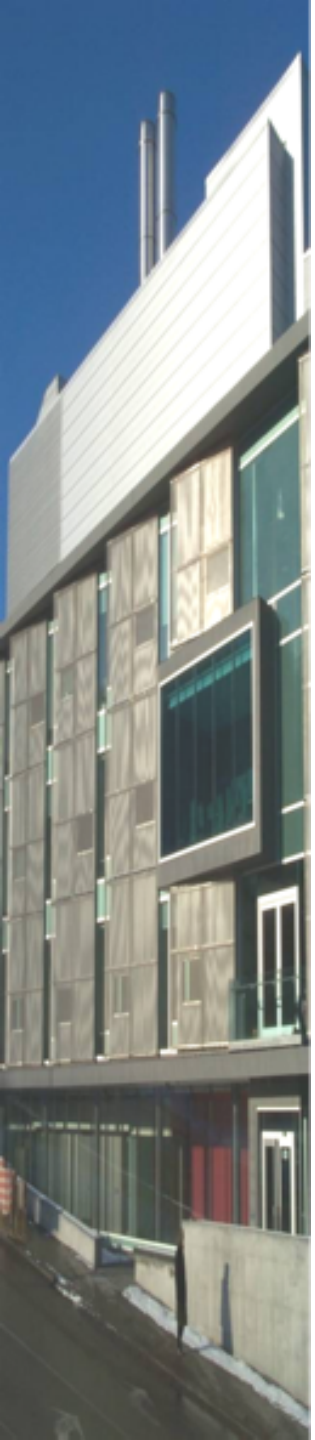
RNA-seQc summary statistics

Sample	Raw reads	Surviving reads	%	Aligned read	%	Alternative aligments	%	rRNA reads	%	Coverage	Exonic Rate	Genes
SRR032239	75863506	73901226	97	32623314	44	9551452	29	0	0	10	0.89	23659
SRR032271	90133262	86099068	96	36723524	43	11732251	32	0	0	11	0.85	24614
SRR032238	76849376	69978578	91	31360242	45	12406875	40	0	0	8	0.79	25595
SRR032272	95302396	92029240	97	38476617	42	12749003	33	0	0	10	0.83	25618
SRR032270	66809402	63033594	94	27645755	44	7250171	26	0	0	10	0.92	23395
SRR032240	75268300	67112328	89	31345651	47	9268748	30	0	0	10	0.88	24825

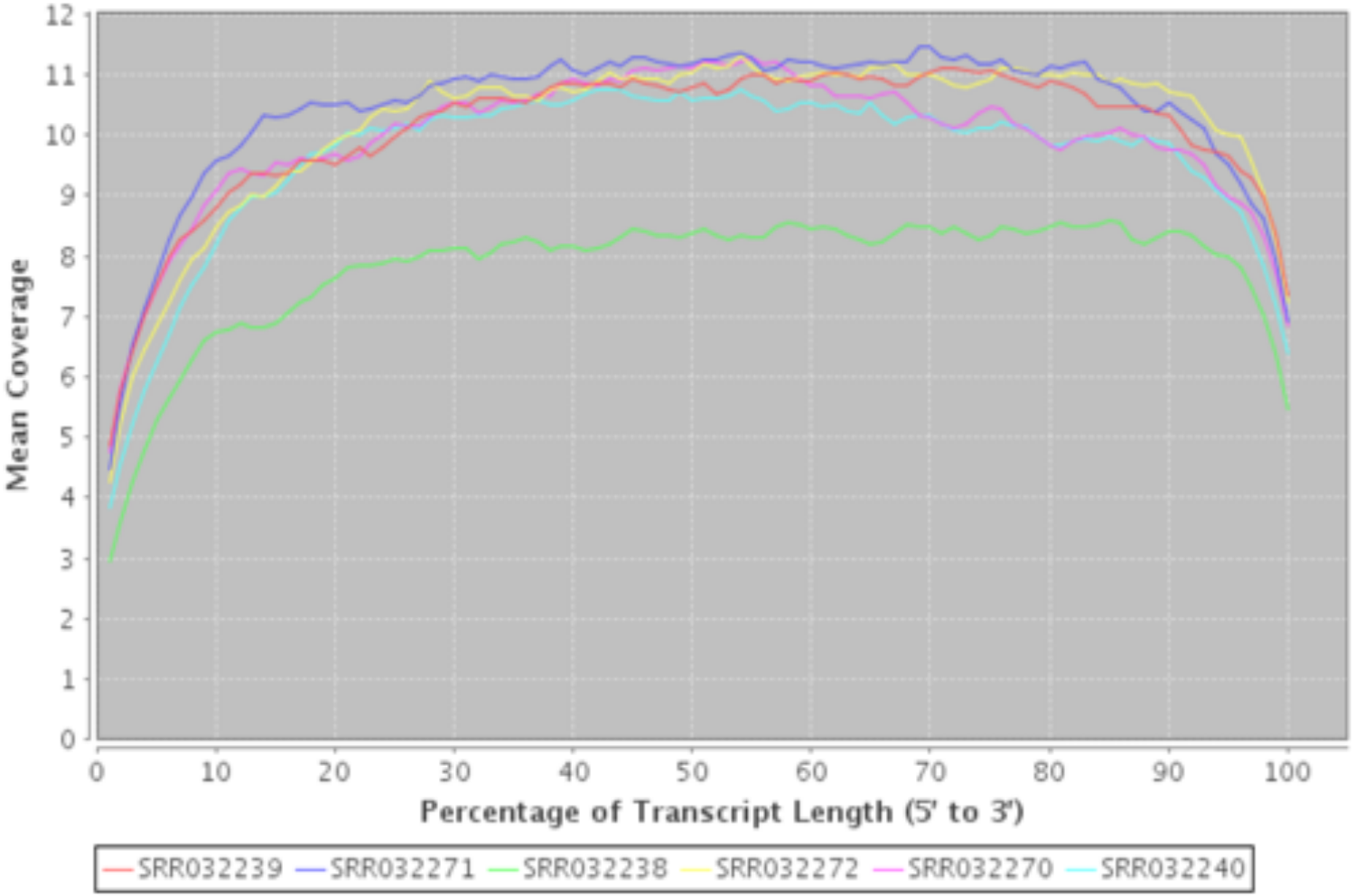
RNA-SeQC: RNA-seq metrics for quality control and process optimization

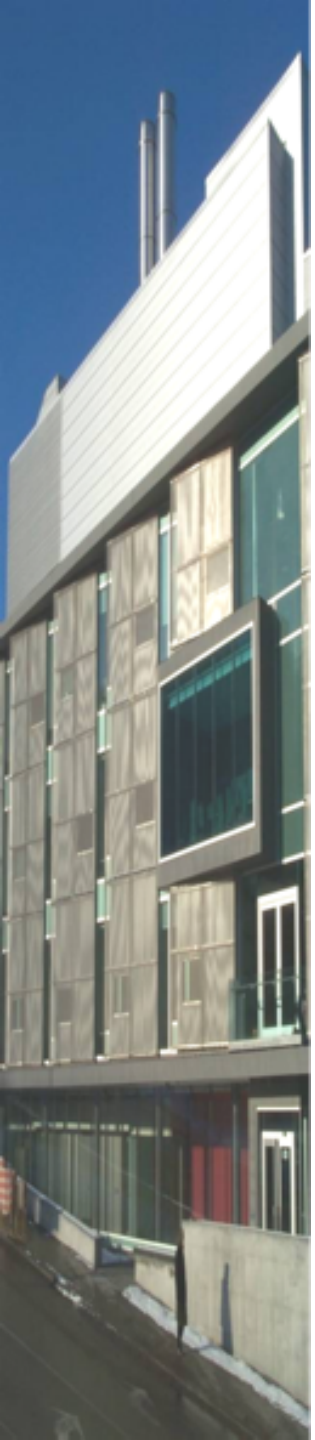
David S. DeLuca*, Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler and Gad Getz*
The Broad Institute of MIT and Harvard, Cambridge, MA, USA

broadinstitute.org/cancer/cga/rna-seqc



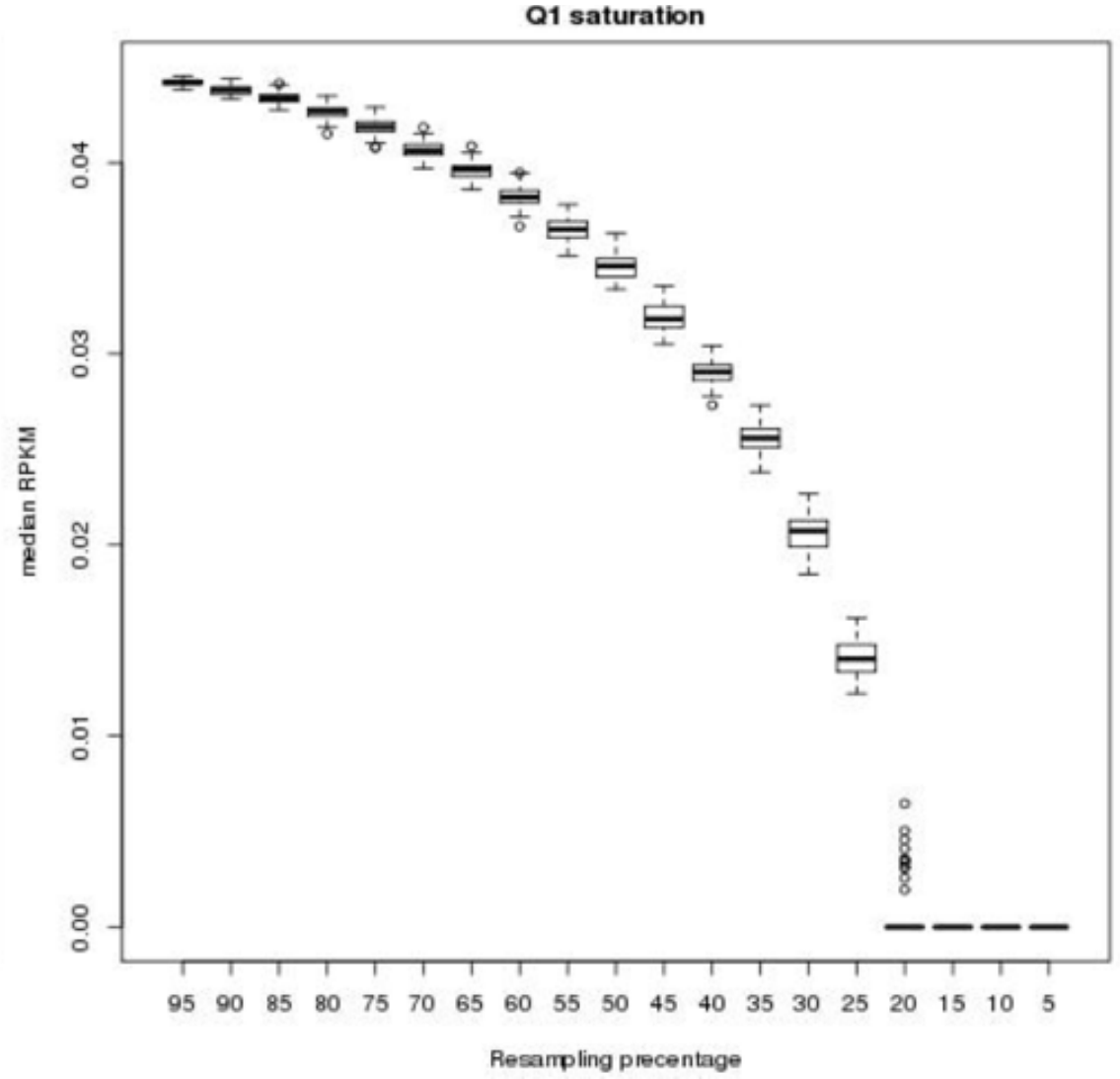
RNA-seQc covergae graph

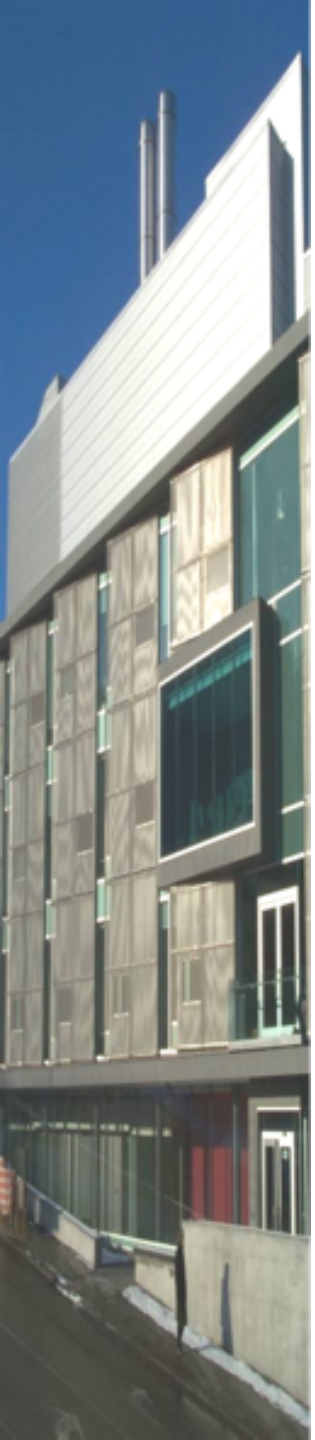




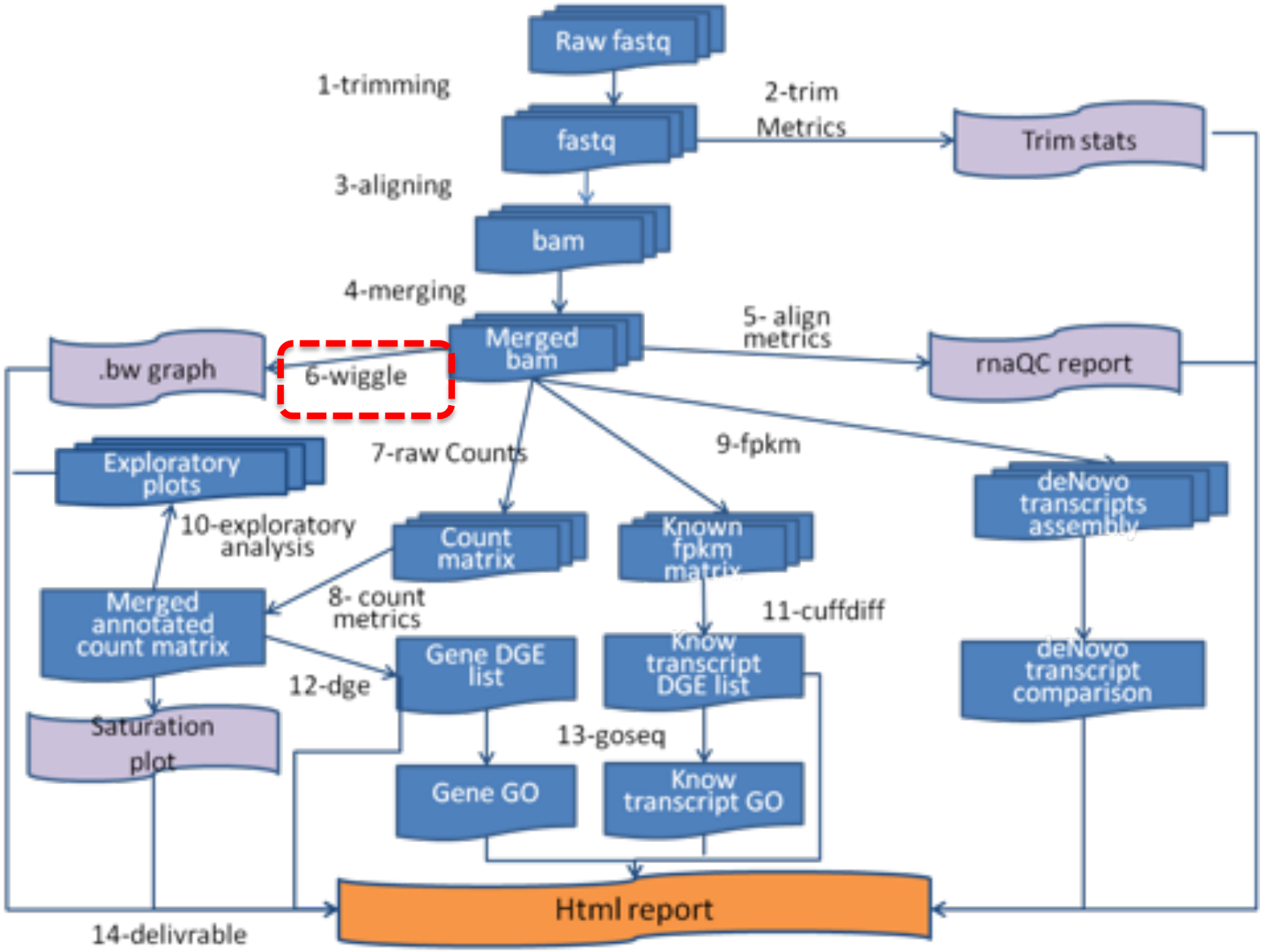
Home-made Rscript: saturation

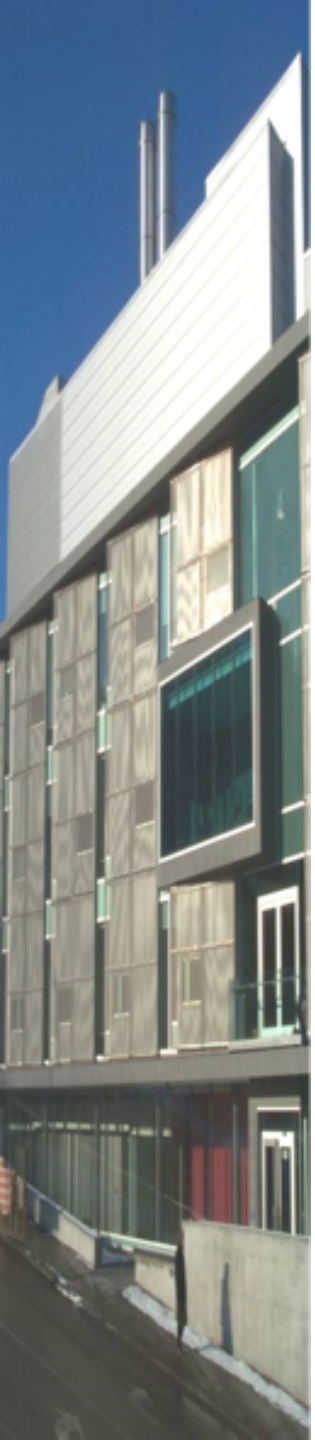
RPKM Saturation Analysis





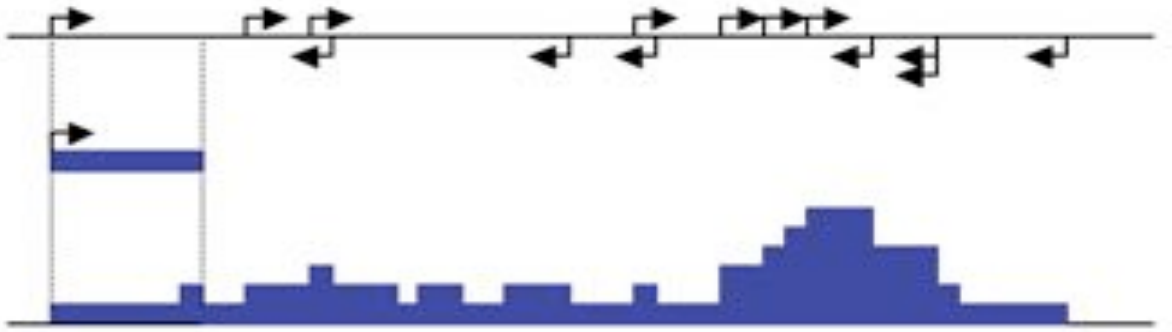
RNA-Seq:Wiggle



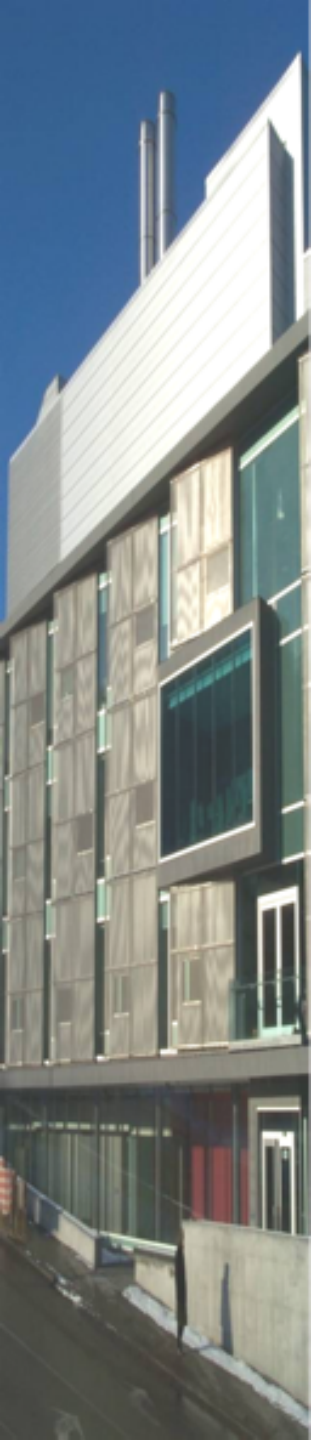


UCSC: bigWig Track Format

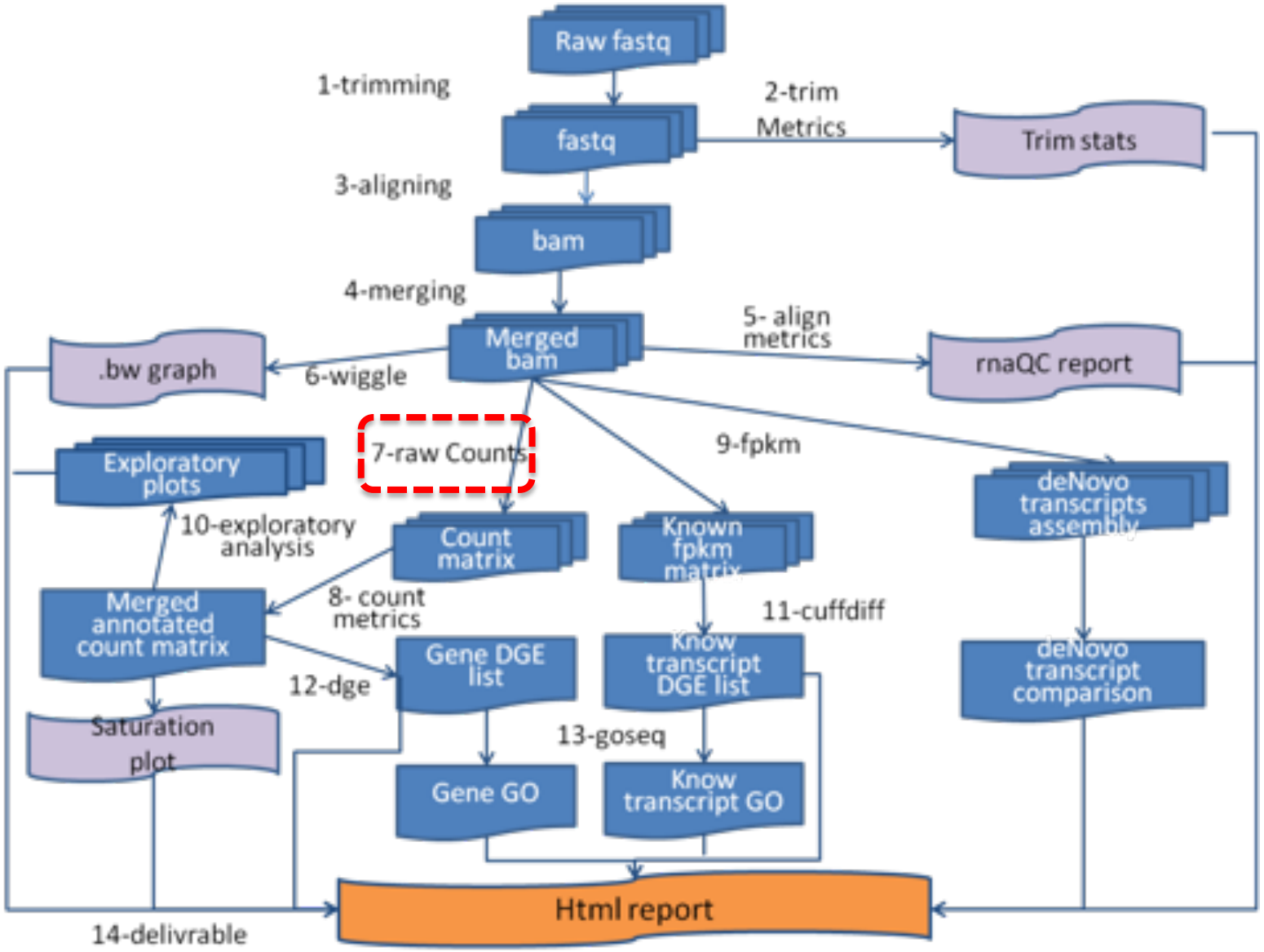
- The bigWig format is for display of dense, continuous data that will be displayed in the Genome Browser as a graph.
- Count the number of read (coverage at each genomic position:

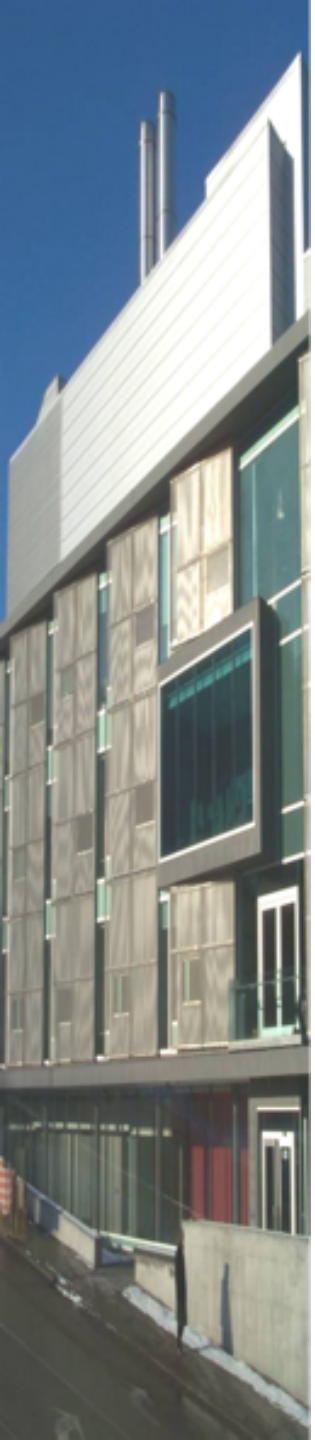


Modified from http://biowhat.ucsd.edu/homer/chipseq/ucsc.html

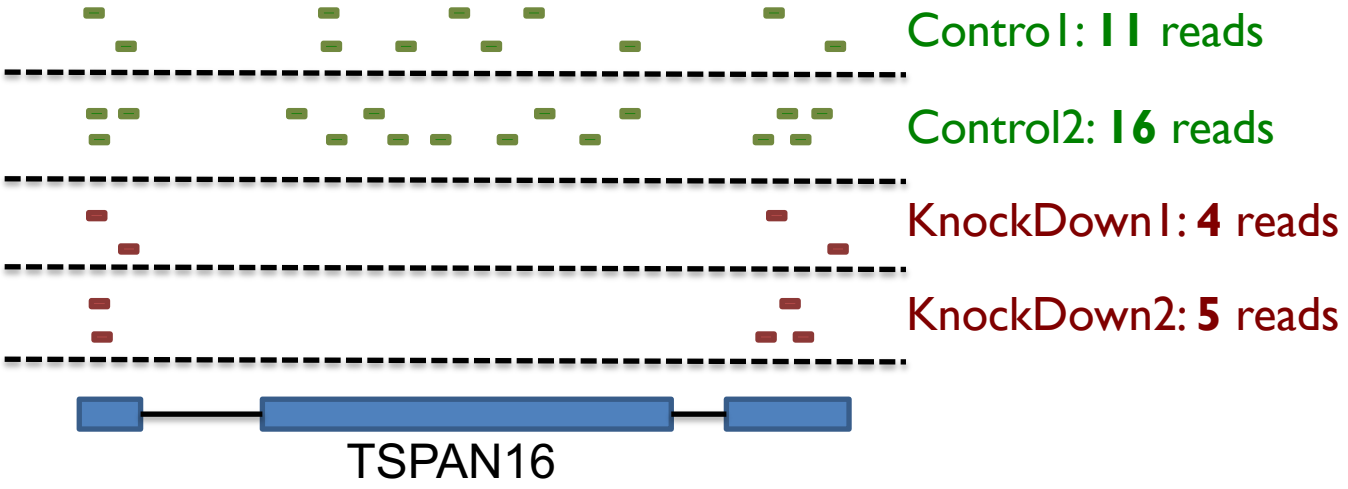


RNA-Seq:Gene-level counts





HTseq: Gene-level counts

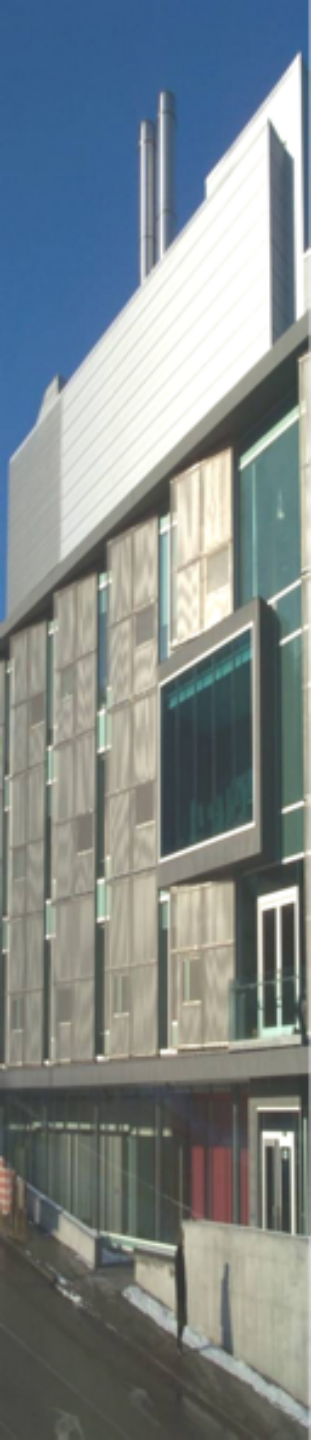


- Reads (BAM file) are counted for each gene model (gtf file) using *HTSeq-count*:

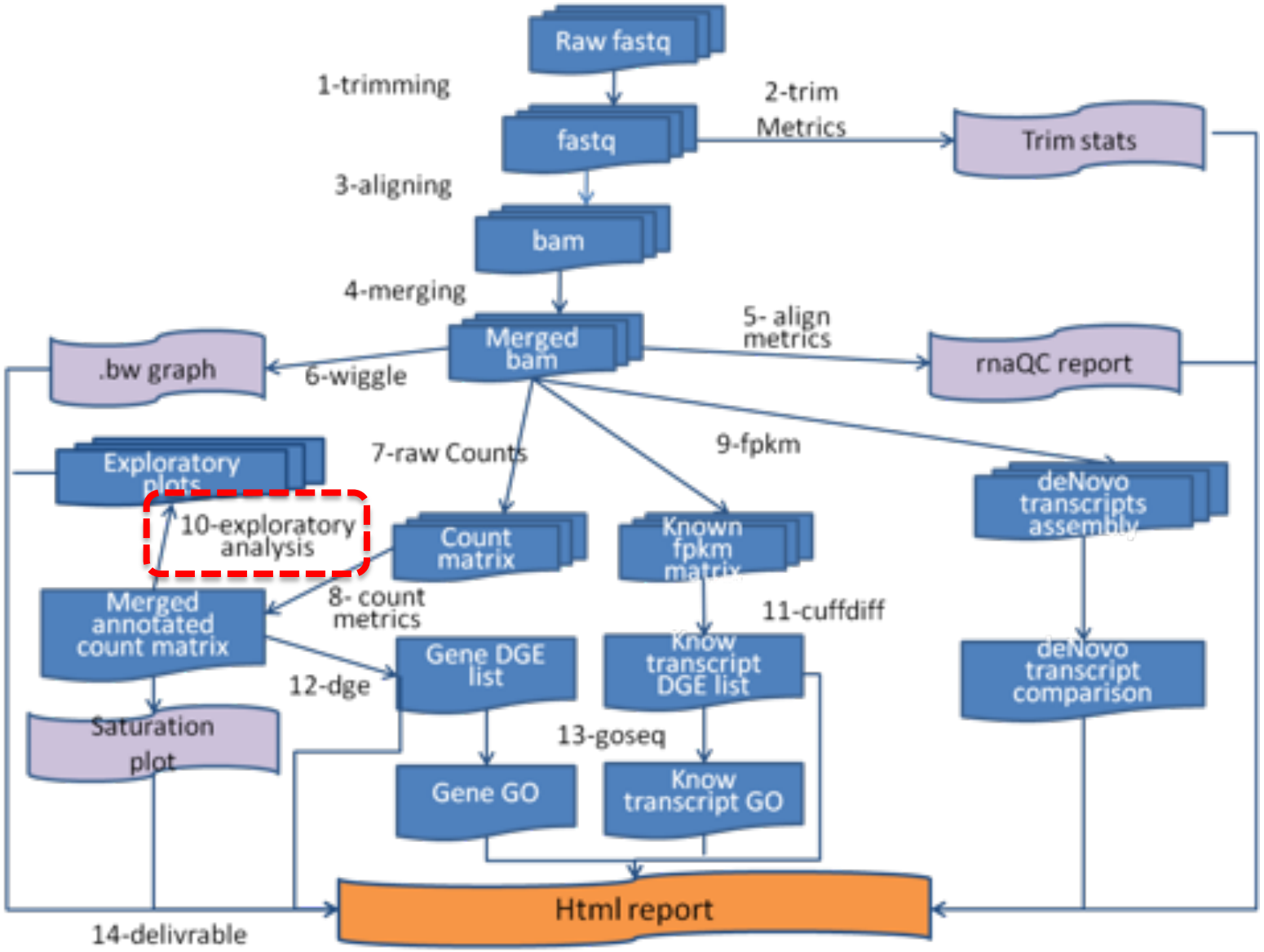
	Control1	Control2	KnockDown1	KnockDown2
TSPAN6	11	16	4	5
TNMD	1	0	0	0
DPM1	435	743	836	739
SCYL3	203	218	416	352
C1orf112	216	643	714	704
FGR	2365	5011	2828	2294
CFH	6	1	4	0
FUCA2	380	865	431	523
...	...	...	...	...
NFYA	888	827	1674	1580

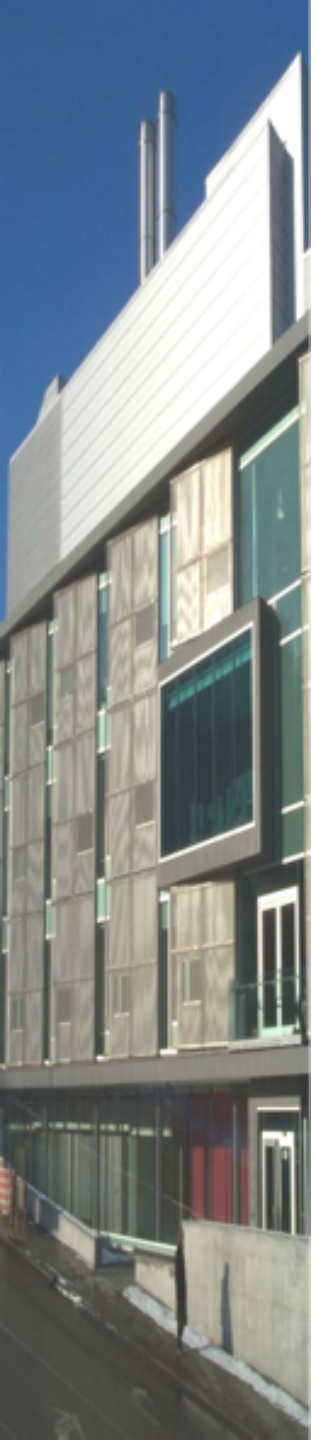
HTSeq

www-huber.embl.de/users/anders/HTSeq

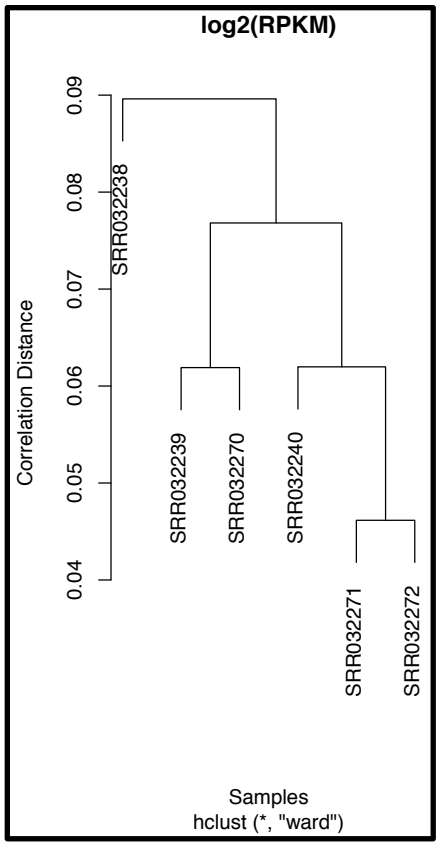


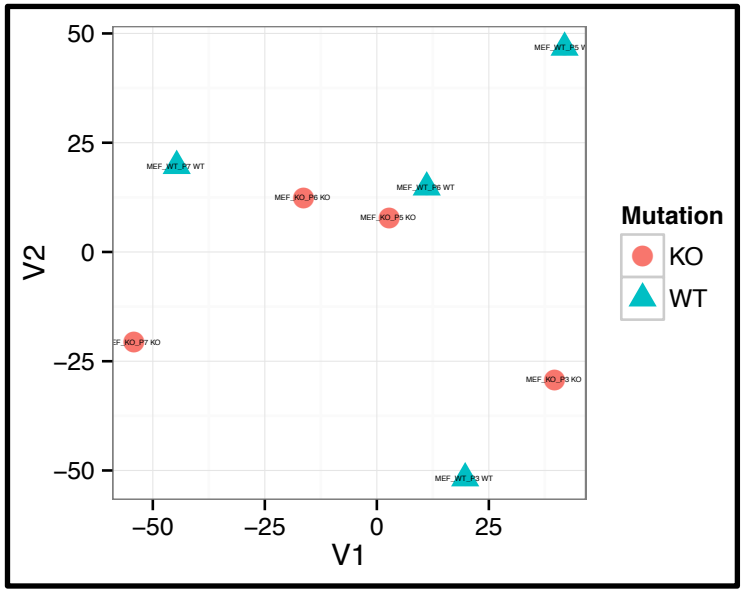
RNA-seq: EDA

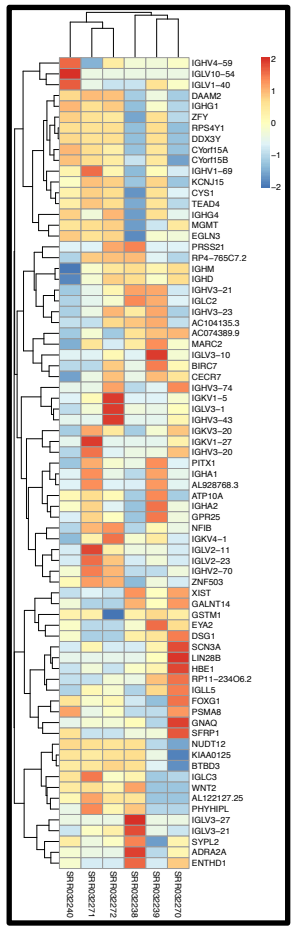

Raw fastq
1-trimming
fastq
2-trim
Metrics
Trim stats
3-aligning
bam
4-merging
Merged bam
5- align metrics
.bw graph
6-wiggle
rnaQC report
Exploratory plots
7-raw Counts
9-fpkm
deNovo transcripts assembly
10-exploratory analysis
Count matrix
Known fpkm matrix
Merged annotated count matrix
8- count metrics
11-cuffdiff
Gene DGE list
Know transcript DGE list
deNovo transcript comparison
12-dge
Saturation plot
13-goseq
Gene GO
Know transcript GO
Html report
14-delivrable

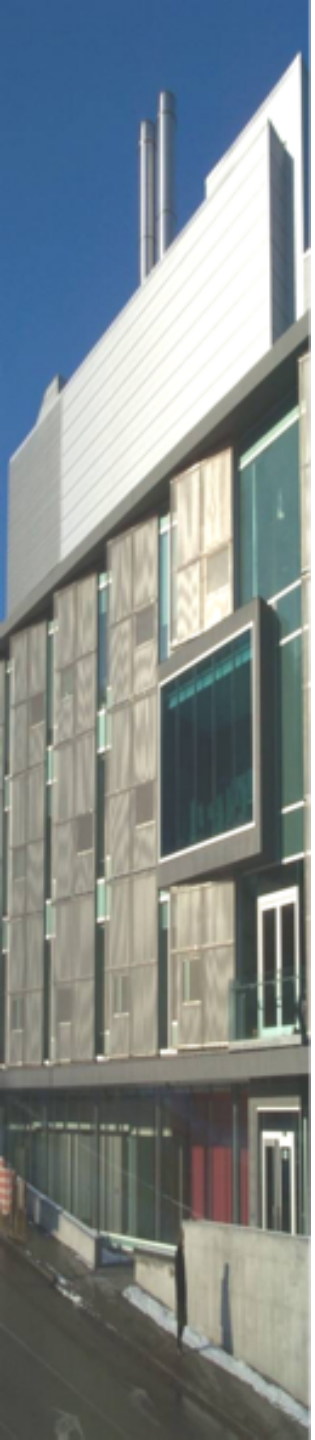


gqSeqUtils R package: Exploratory Data Analysis

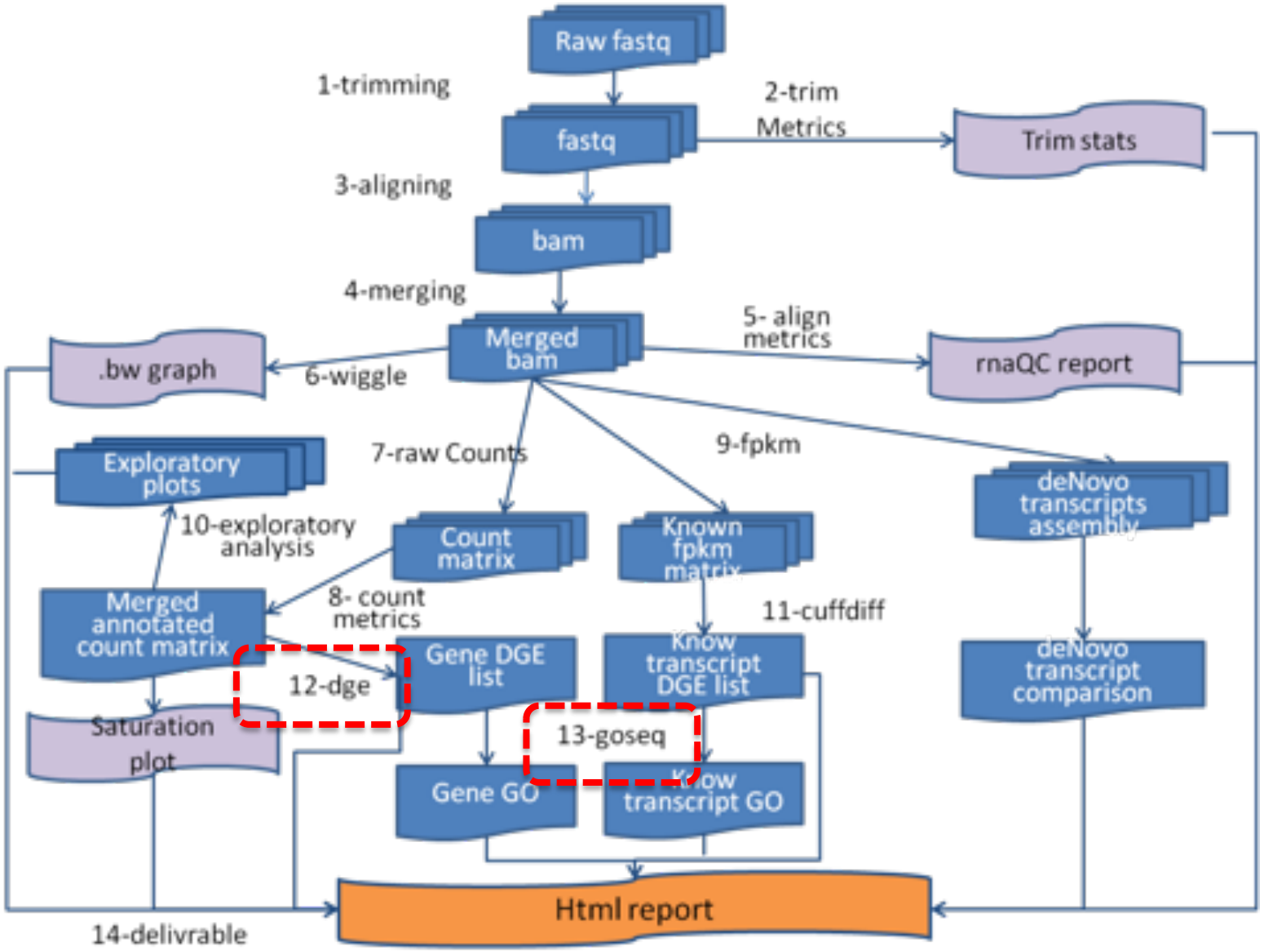


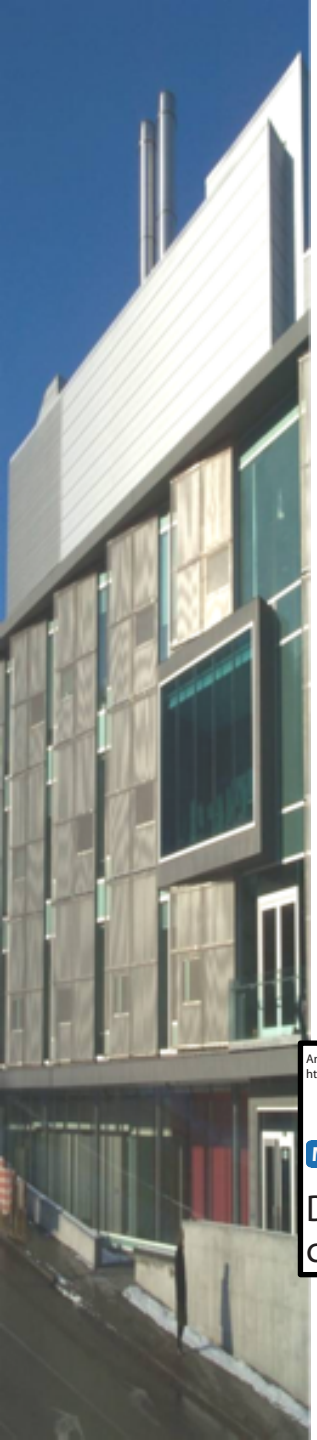






RNA-Seq:Gene-level DGE





Home-made Rscript: Gene-level DGE

- *edgeR* and *DESeq* : Test the effect of exp. variables on gene-level read counts
- GLM with negative binomial distribution to account for biological variability (not Poisson!!)

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

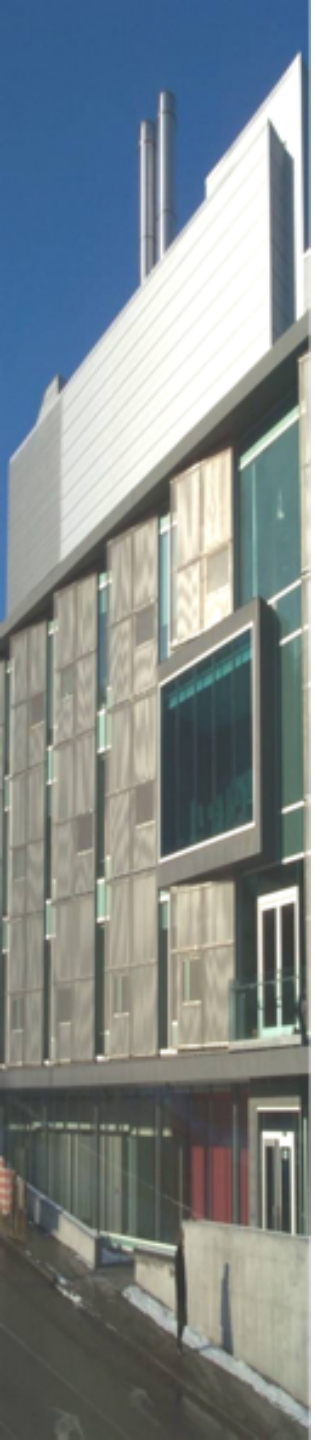
DEseq

4288–4297 Nucleic Acids Research, 2012, Vol. 40, No. 10 Published online 28 January 2012
doi:10.1093/nar/gks042

Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation

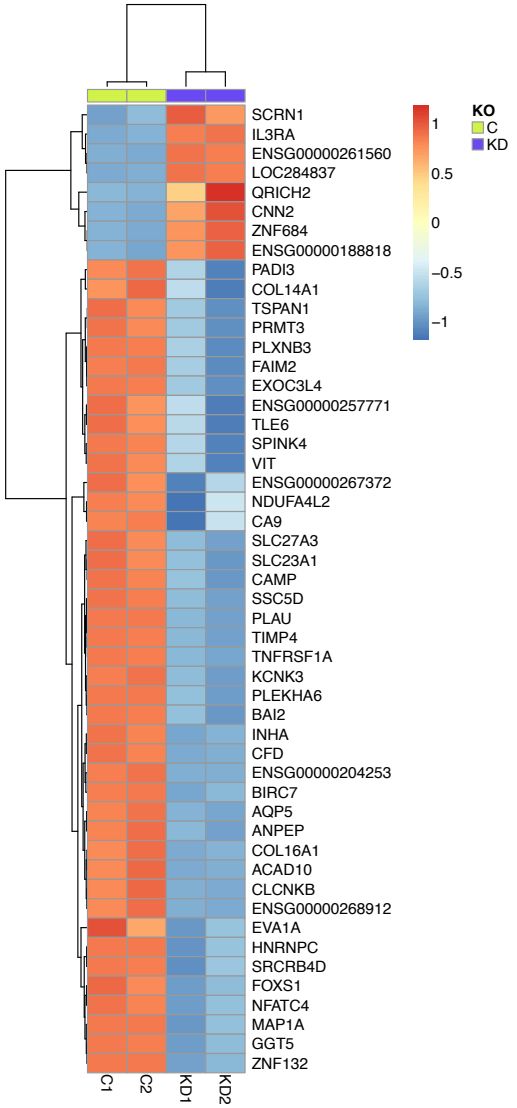
Davis J. McCarthy[1], Yunshun Chen[1,2] and Gordon K. Smyth[1,3,*]

edgeR



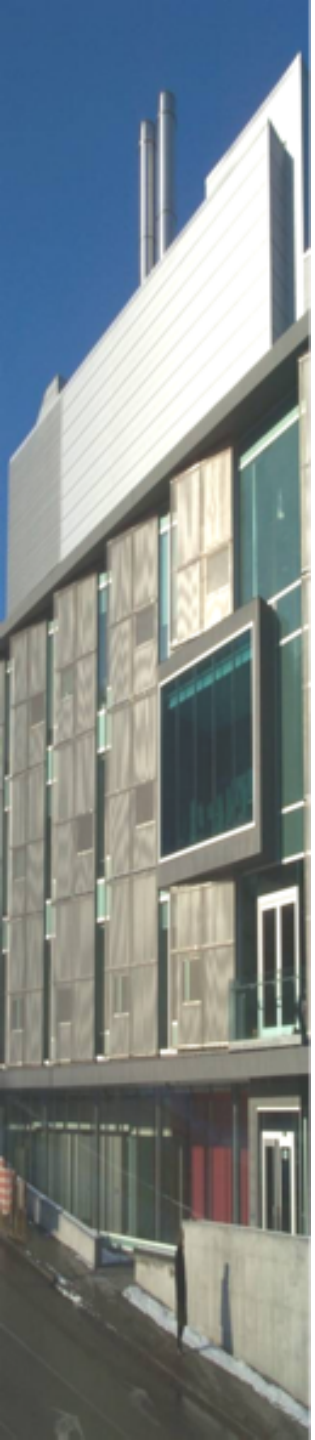
Differential Gene Expression

SYMBOL	logFC	PValue	FDR	counts.C1	counts.C2	counts.KD1	counts.KD2
HNRNPC	-5.26	9.19E-55	5.71E-50	12611	12404	244	443
FAIM2	-4.82	8.02E-29	2.49E-24	191	194	11	3
AC019178	-6.57	2.14E-28	4.42E-24	100	104	1	1
SSC5D	-2.95	2.39E-27	3.71E-23	2274	2123	318	276
GGT5	-3.03	1.03E-26	1.28E-22	838	803	93	117
EXOC3L4	-3.07	9.19E-21	9.51E-17	359	344	53	34
FOXS1	-4.02	1.69E-19	1.49E-15	113	92	5	8
AQP5	-3.73	2.82E-19	2.18E-15	106	113	9	8
SLC27A3	-2.39	6.97E-18	4.81E-14	736	637	144	129
TIMP4	-3.29	1.21E-17	7.52E-14	126	120	14	12

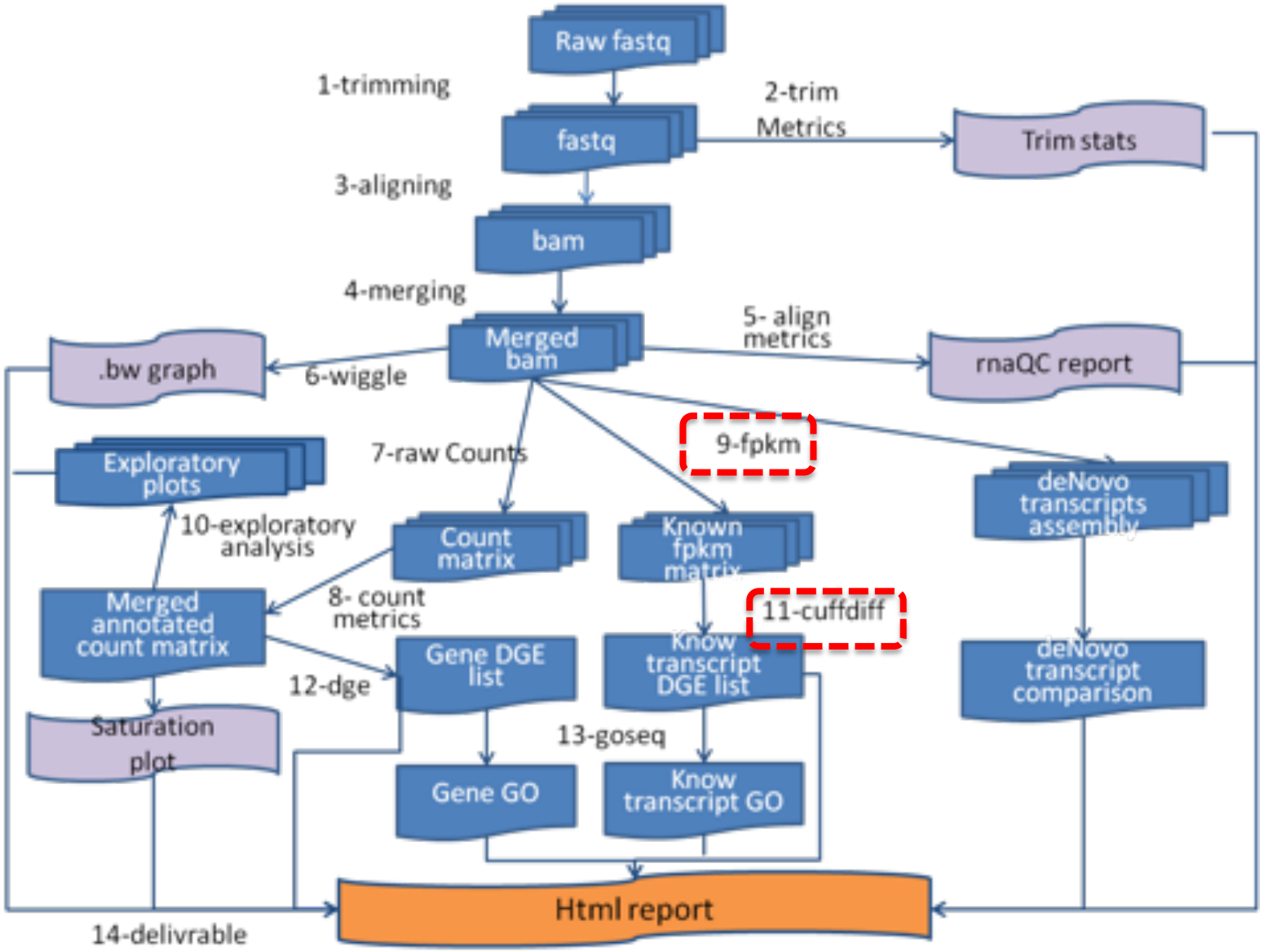


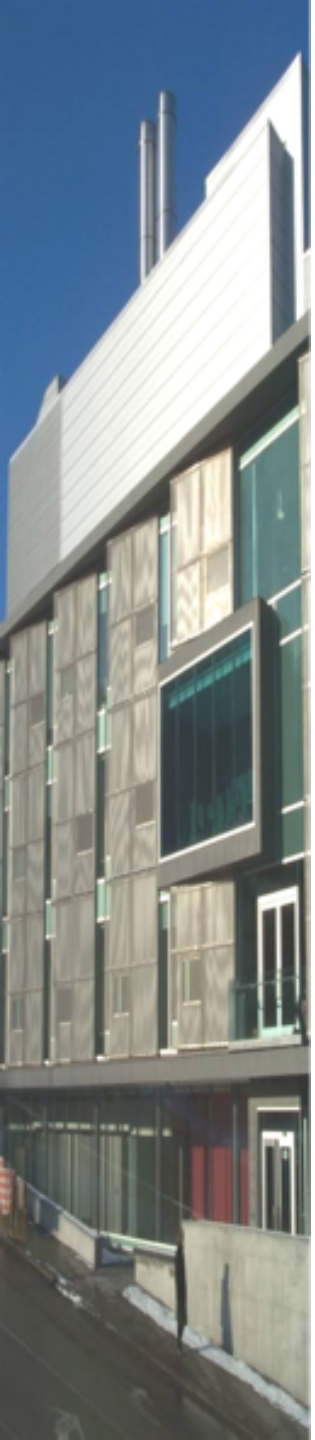
Downstream Analyses

- Pathways/Gene Set (e.g. **GOSeq**)
- Regulatory Networks
- Machine Learning / Classifiers



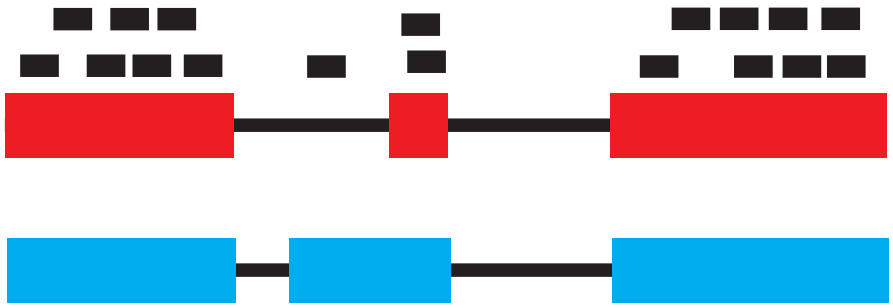
RNA-Seq:Transcript-level DGE





Cufflinks: transcipt assembly

- **Assembly**: Reports the most parsimonious set of transcripts (*transfrags*) that explain splicing junctions found by *TopHat*

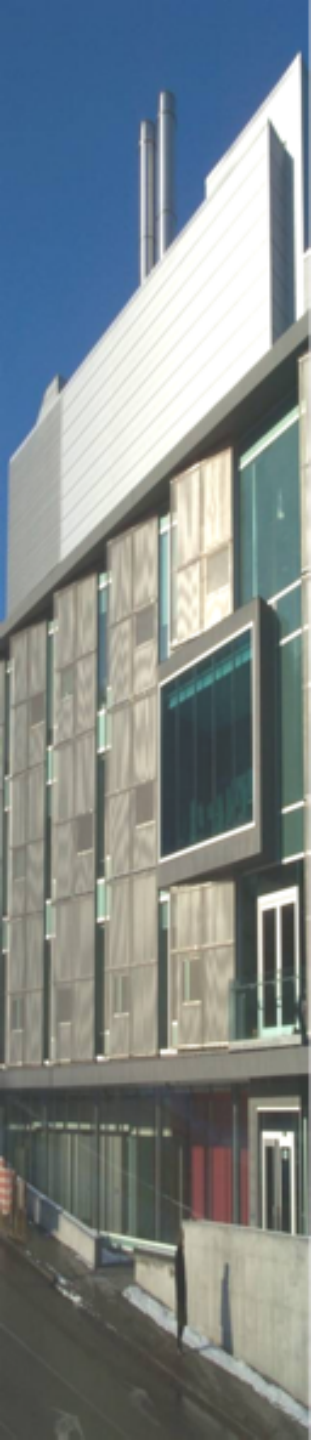


Isoform A

Isoform B

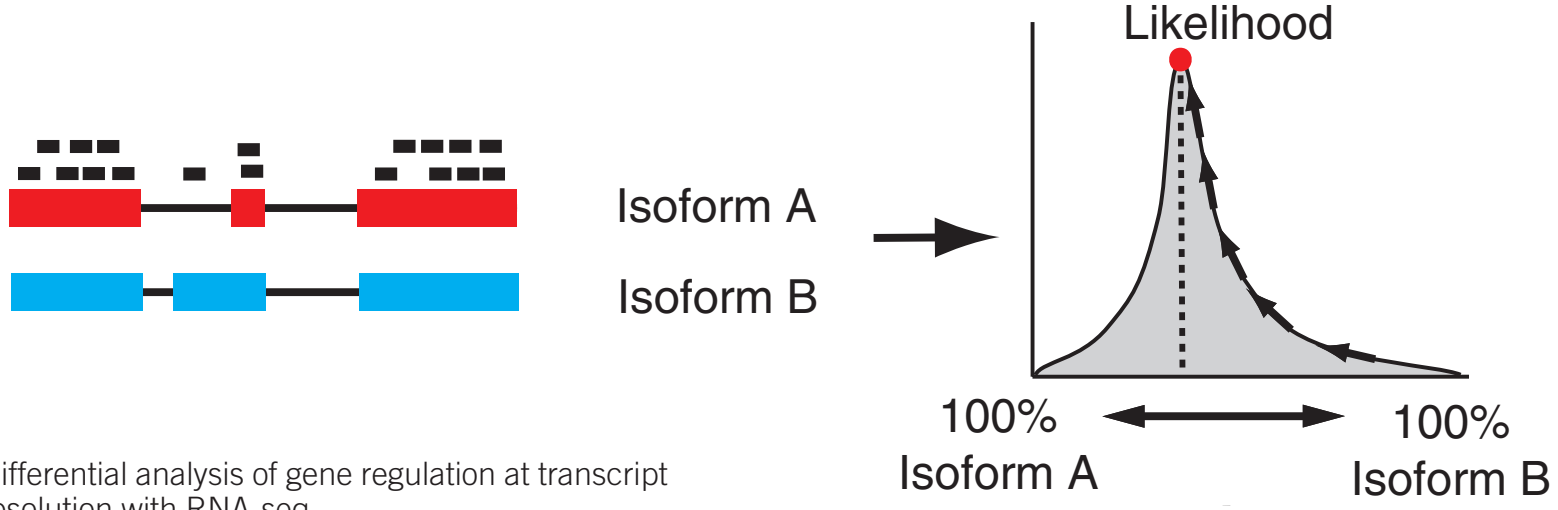
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]



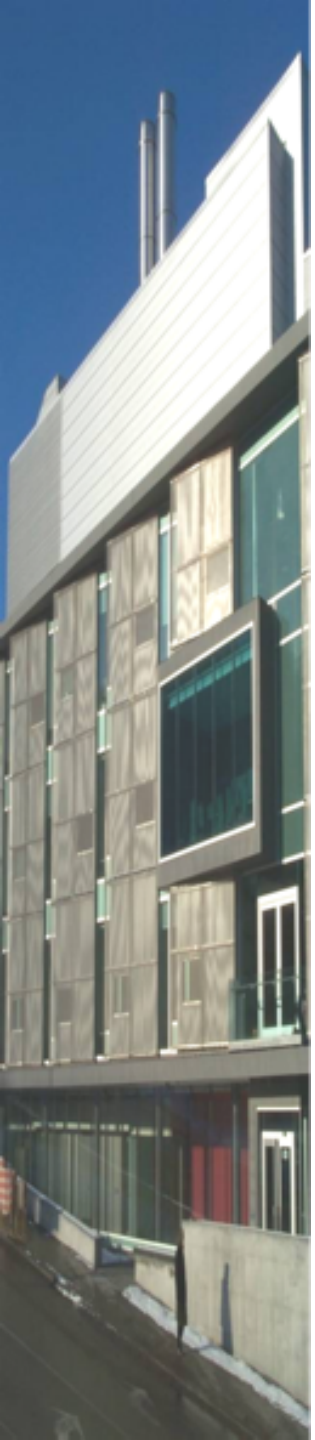
Cufflinks: transcript abundance

- **Quantification**: Cufflinks implements a linear statistical model to estimate an assignment of abundance to each transcript that explains the observed reads with maximum likelihood.



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

Cole Trapnell[1,2,6], David G Hendrickson[1,2,6], Martin Sauvageau[1,2], Loyal Goff[1–3], John L Rinn[1,2,7] & Lior Pachter[4,5,7]



Cufflinks: abundance output

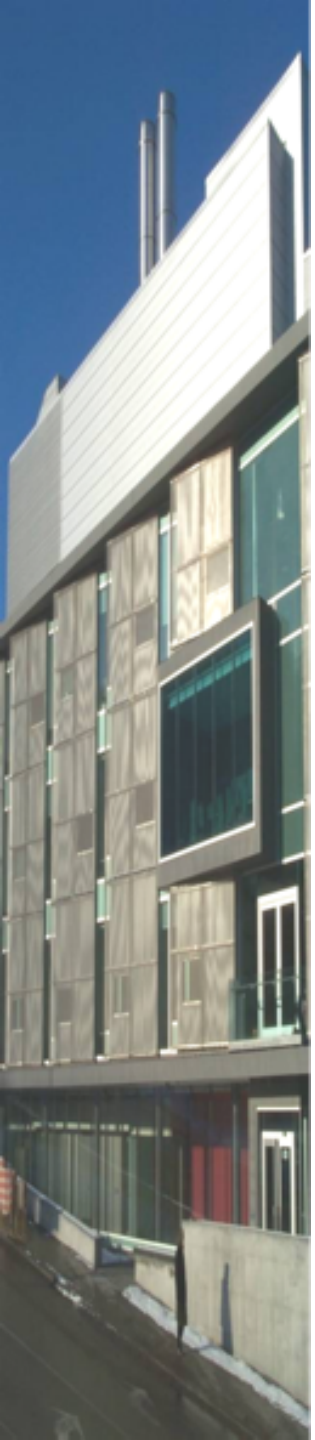
- Cufflinks reports abundances as **F**ragments **P**er **K**ilobase of exon model per **M**illion mapped fragments (FPKM)

	SRR032239	SRR032271	SRR032238	SRR032272
ENST00000379389	145.60	503.78	34.49	259.24
ENST00000433695	6.15	2.63	5.38	14.00
ENST00000379198	8.34	6.89	4.53	4.21
ENST00000343938	10.62	6.40	6.14	7.76
ENST00000378344	7.58	15.03	7.47	7.55
ENST00000377648	8.61	5.78	2.72	7.32
ENST00000302692	9.26	8.80	5.69	5.14

$$\mathrm{FPKM} = 10^9 \frac{C}{NL}$$

C: Number of read pairs (fragments) from transcript
N: Total number of mapped read pairs in library
L: number of exonic bases for transcript

- Normalizes for transcript length and lib. size



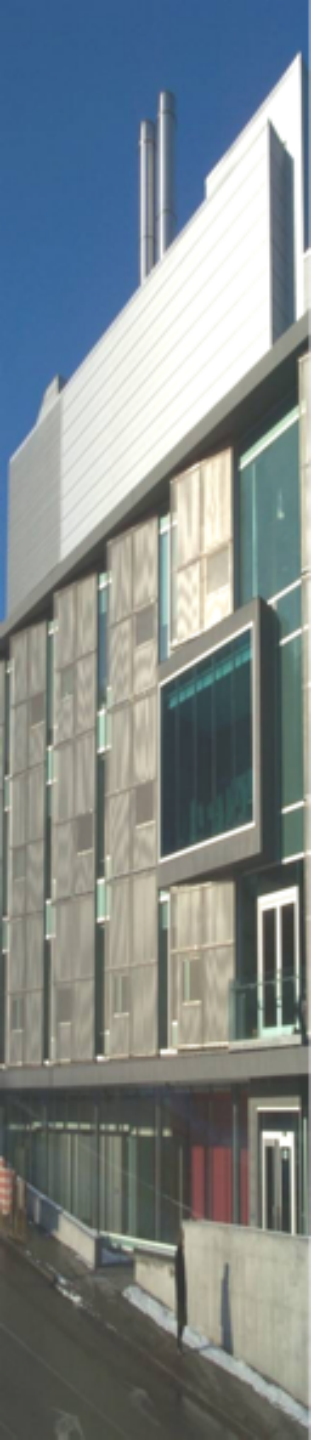
Cuffdiff: differential transcript expression

- **Cudiff**
  - Tests for differential expression of a cufflinks assembly

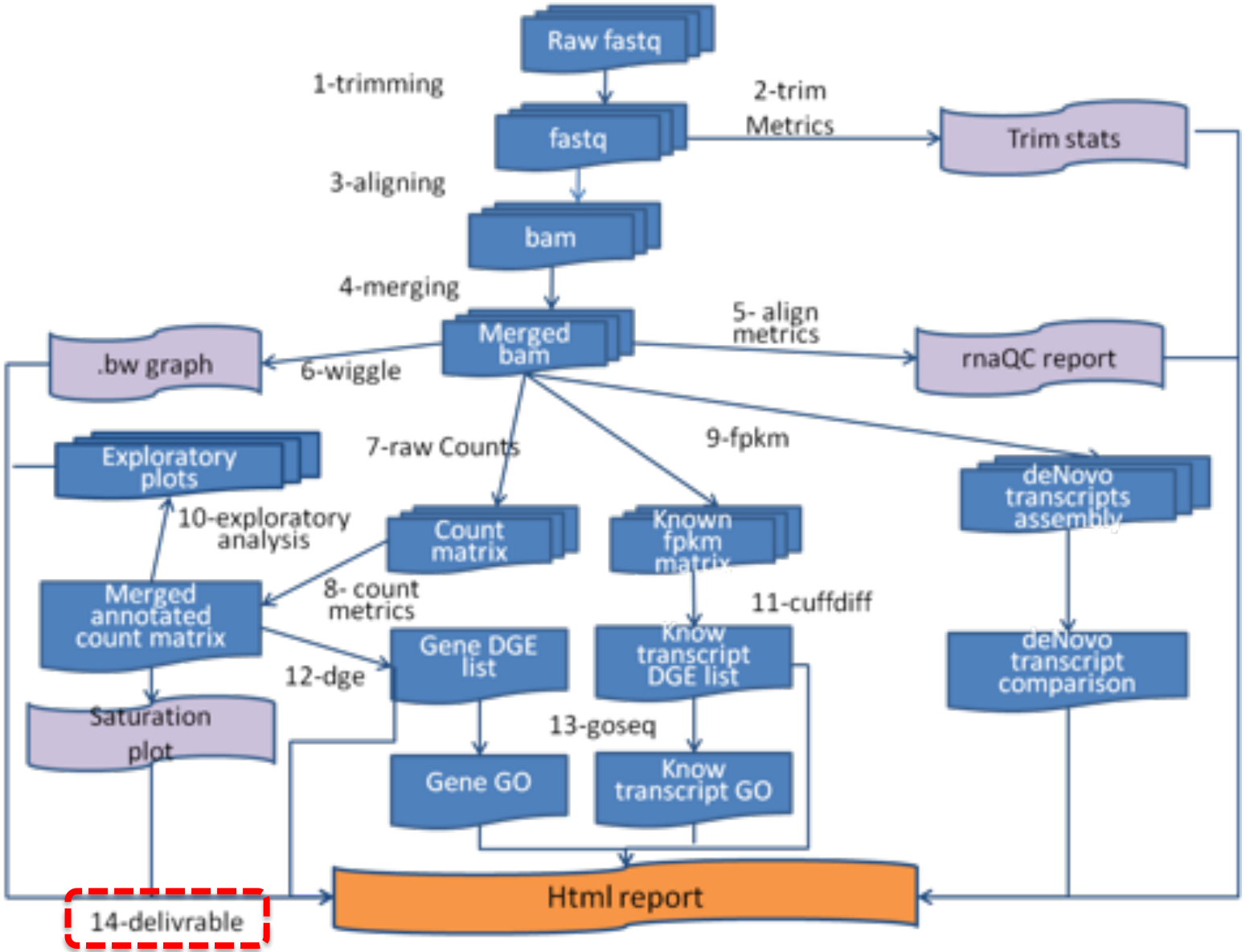
test_id	gene	log2.fold_change	p_value	q_value	fpkm.SRR032238	fpkm.SRR032239	...
ENST00000177694	TBX21	3.80433	5.00E-05	0.07989	0.255737557	0.09057553	...
ENST00000239461	PRRX1	-5.91726	5.00E-05	0.07989	19.62584291	0.018224123	...
ENST00000252971	MNX1	3.45374	5.00E-05	0.07989	0.376127203	0.407829904	...
ENST00000260227	MMP7	3.62719	5.00E-05	0.07989	1.106081955	0.365472353	...
ENST00000261192	BCAT1	-1.87185	5.00E-05	0.07989	14.26418416	13.46141175	...
ENST00000261978	LTBP2	-3.60277	5.00E-05	0.07989	1.285677603	0.021996299	...
...	...	...	...	...	...	...	...

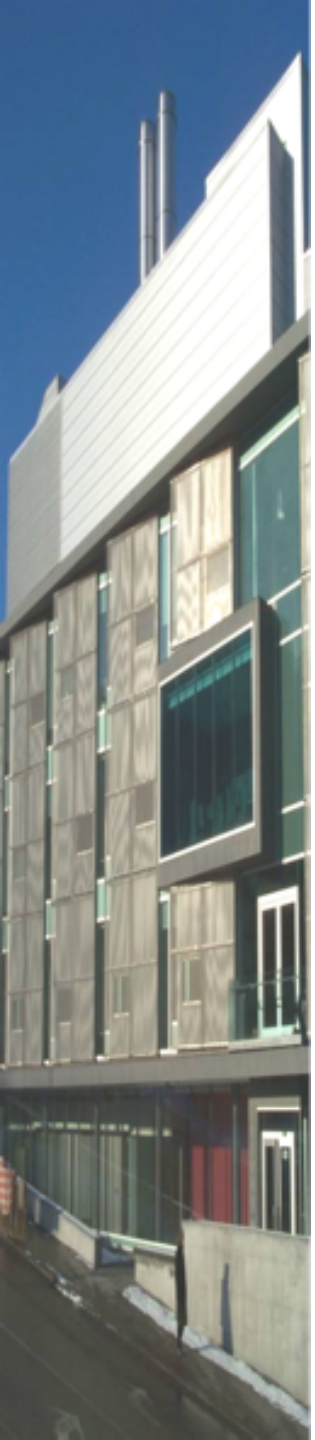
Differential analysis of gene regulation at transcript resolution with RNA-seq

Cole Trapnell[1,2,6], David G Hendrickson[1,2,6], Martin Sauvageau[1,2], Loyal Goff[1–3], John L Rinn[1,2,7] & Lior Pachter[4,5,7]



RNA-Seq:Generate report





Home-made Rscript

Generate report

- Noozle-based html report which describe the entire analysis and provide a general set of summary statistics as well as the entire set of results

Files generated:

- index.html, links to detailed statistics and plots

For examples of report generated while using our pipeline please visit our website