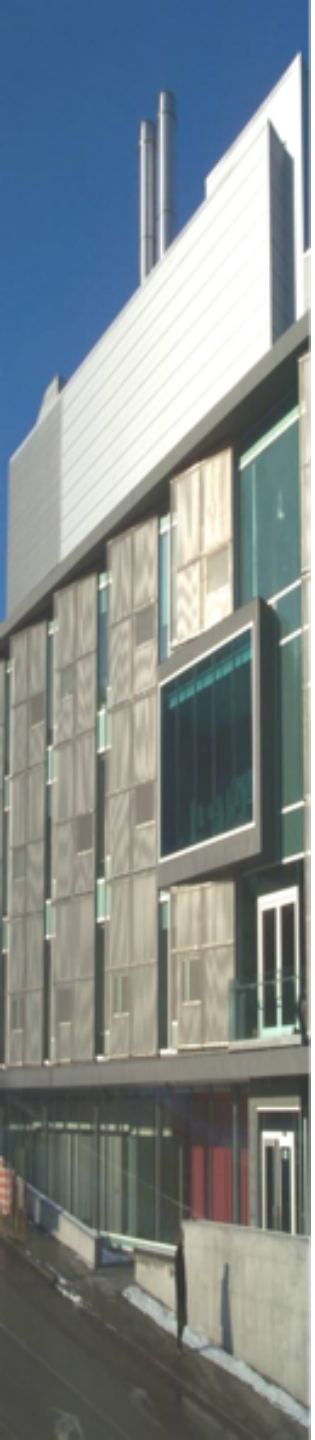
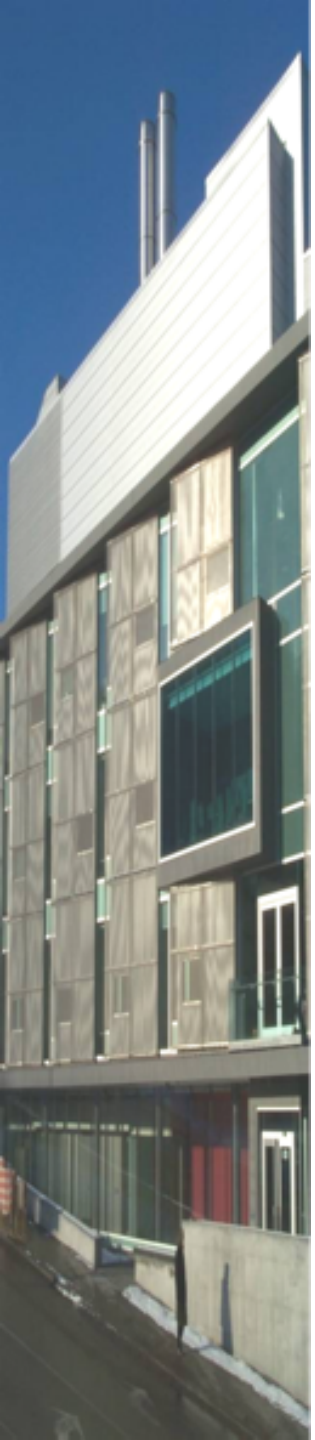


GenomeQuébec

ChIPseq analysis

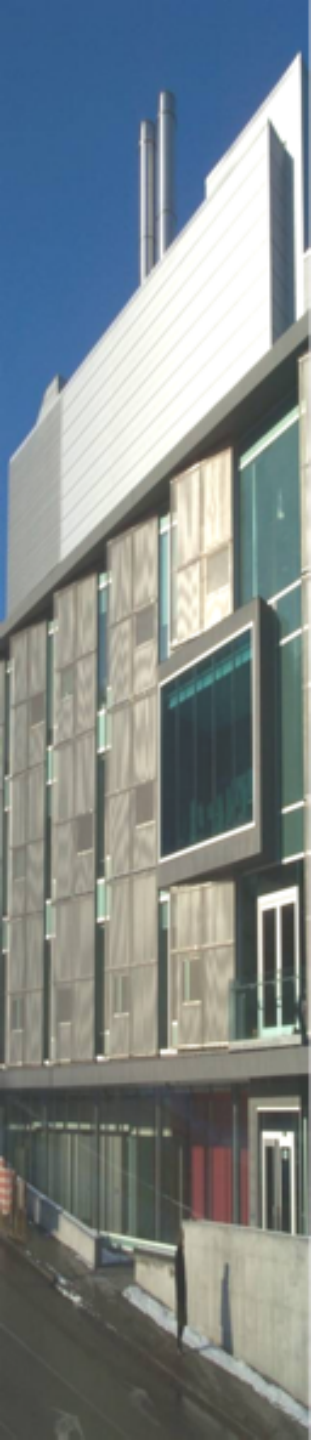
Bioinformatics Analysis Team
McGill University and Genome Quebec Innovation Center
bioinformatics.service@mail.mcgill.ca



What is ChIP-Sequencing?

- Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing
- Allow mapping of protein–DNA interactions *in vivo* on a genome scale

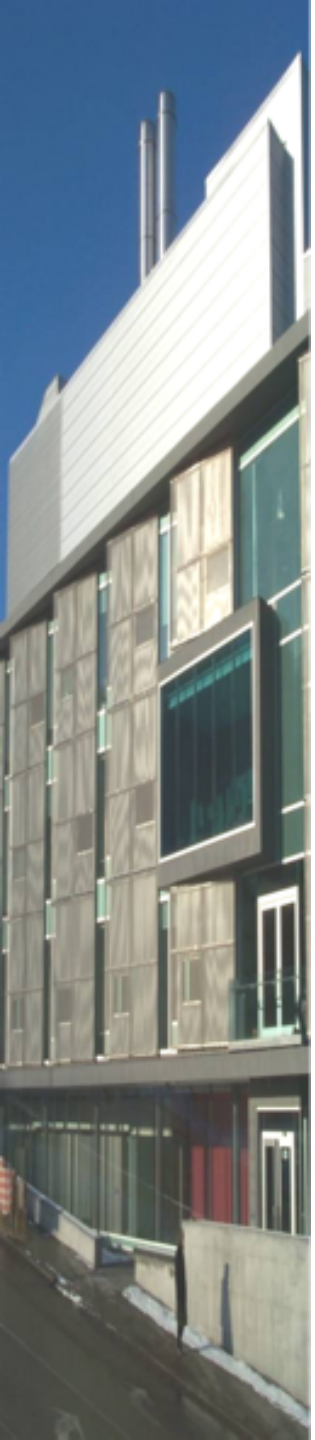
Modified from Bionformatics.ca

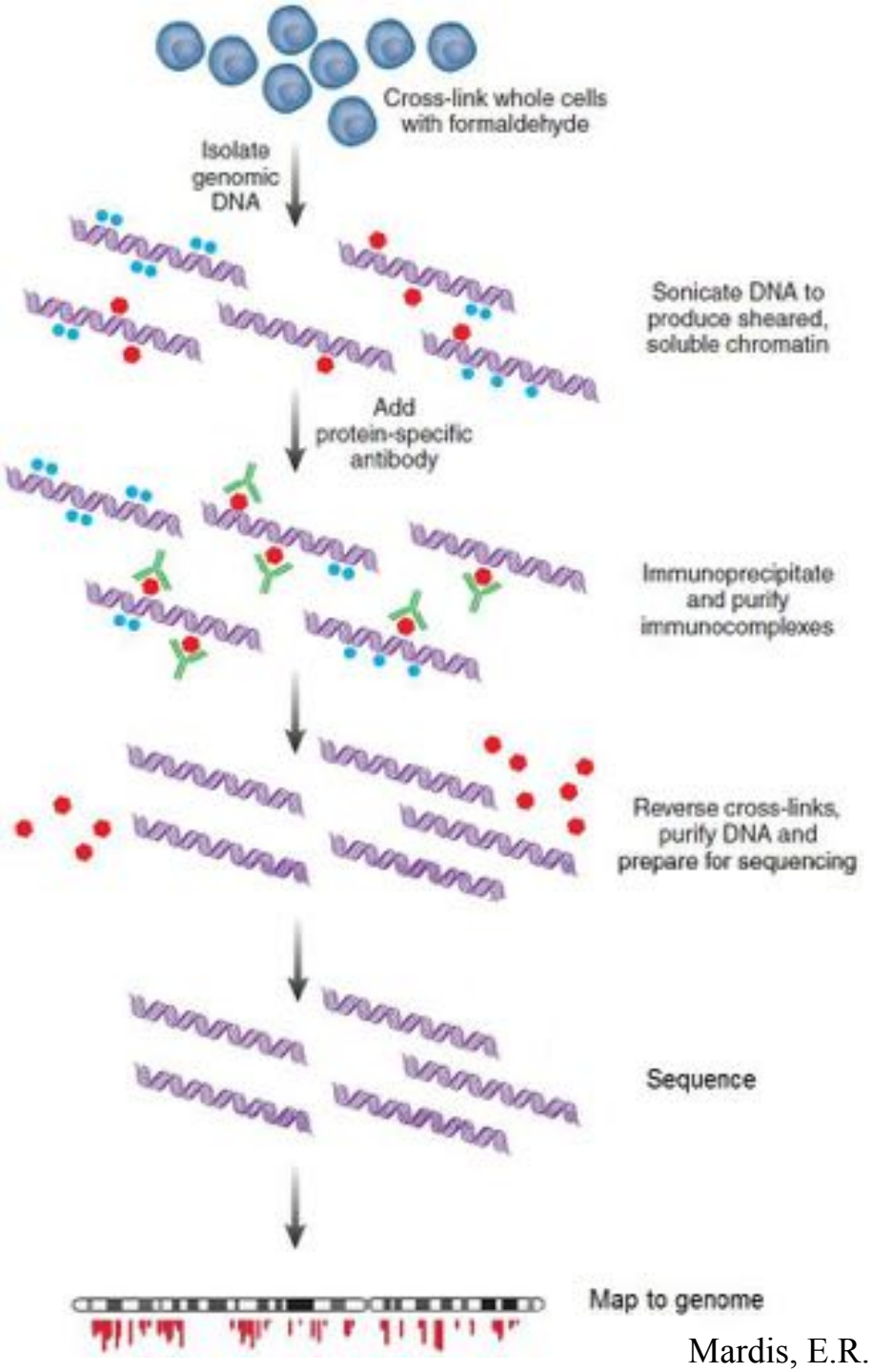


Why ChIP-Sequencing?

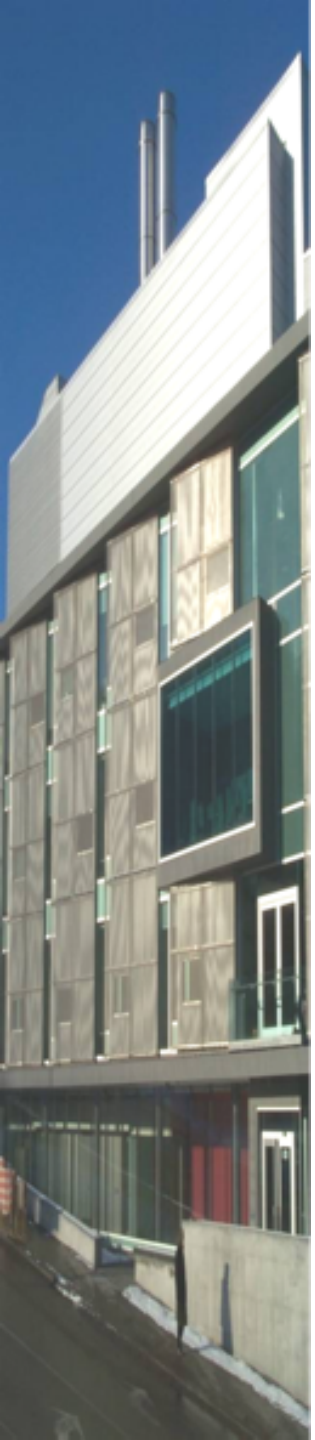
- Current microarray and ChIP-ChIP designs require knowing sequence of interest such as a promoter, enhancer, or RNA-coding domain.
- Higher accuracy
- Alterations in transcription-factor binding in response to environmental stimuli can be evaluated for the entire genome in a single experiment.

Modified from Bionformatics.ca



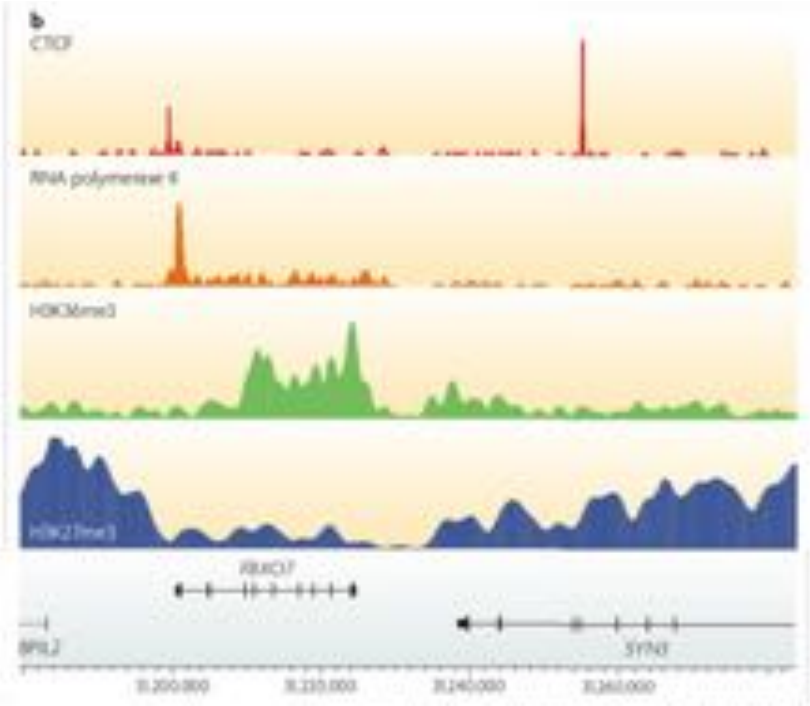


Mardis, E.R. *Nat. Methods* **4**, 613-614 (2007)

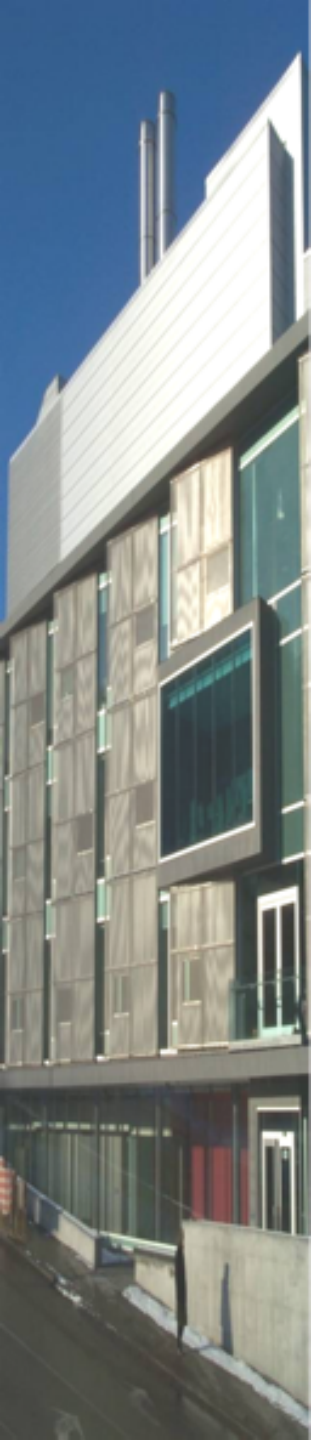


Chip-seqChallenges

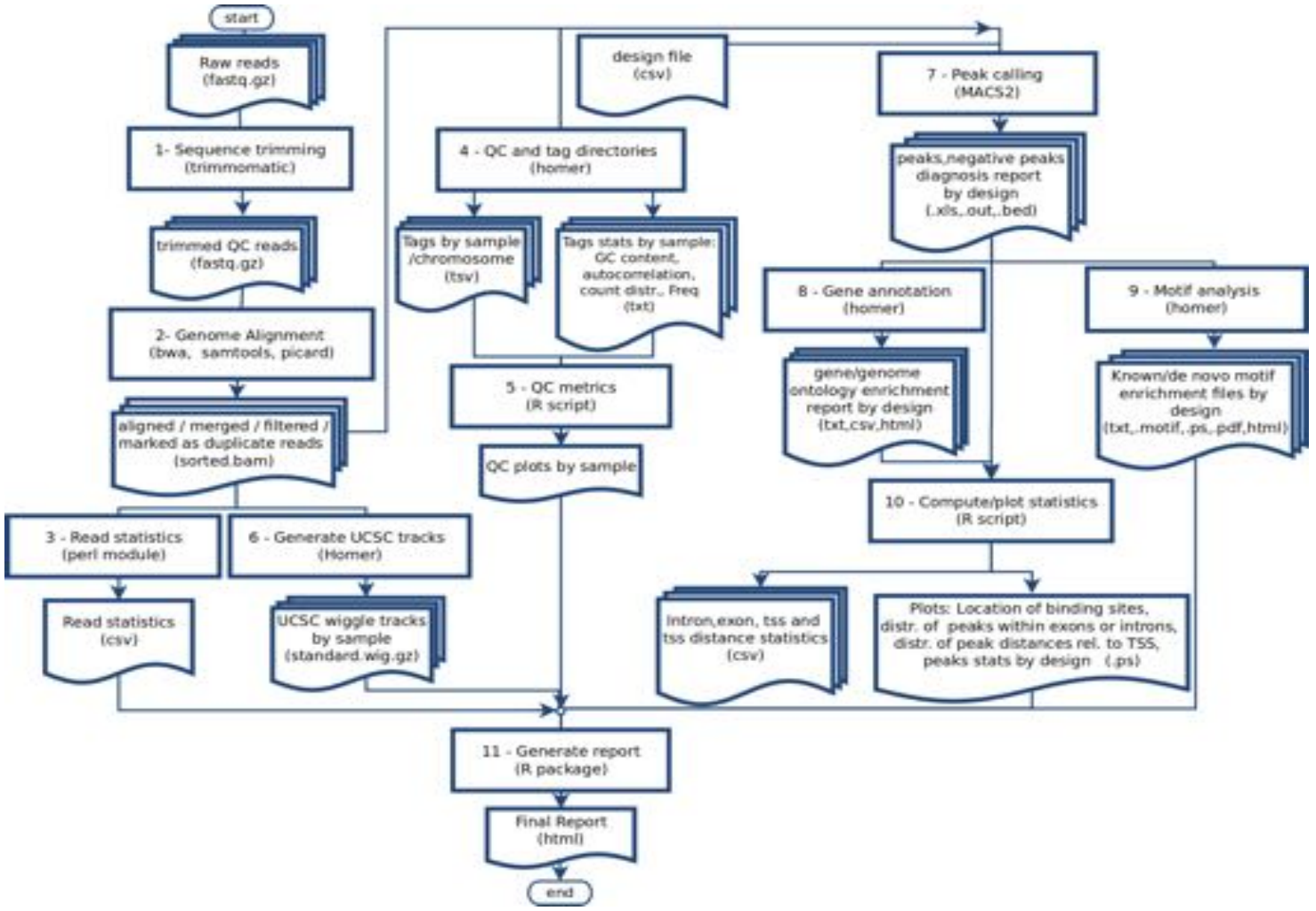
- Peak analysis
  - Peak detection
  - Finding exact binding sites
- Comparing results of different experiments
  - Normalization
  - Statistical tests

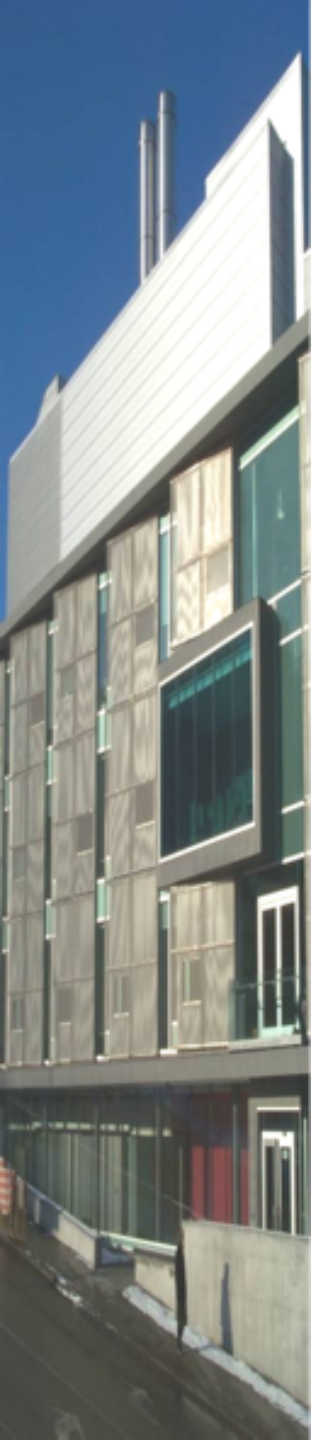


Peter J Park, Nature, 2009

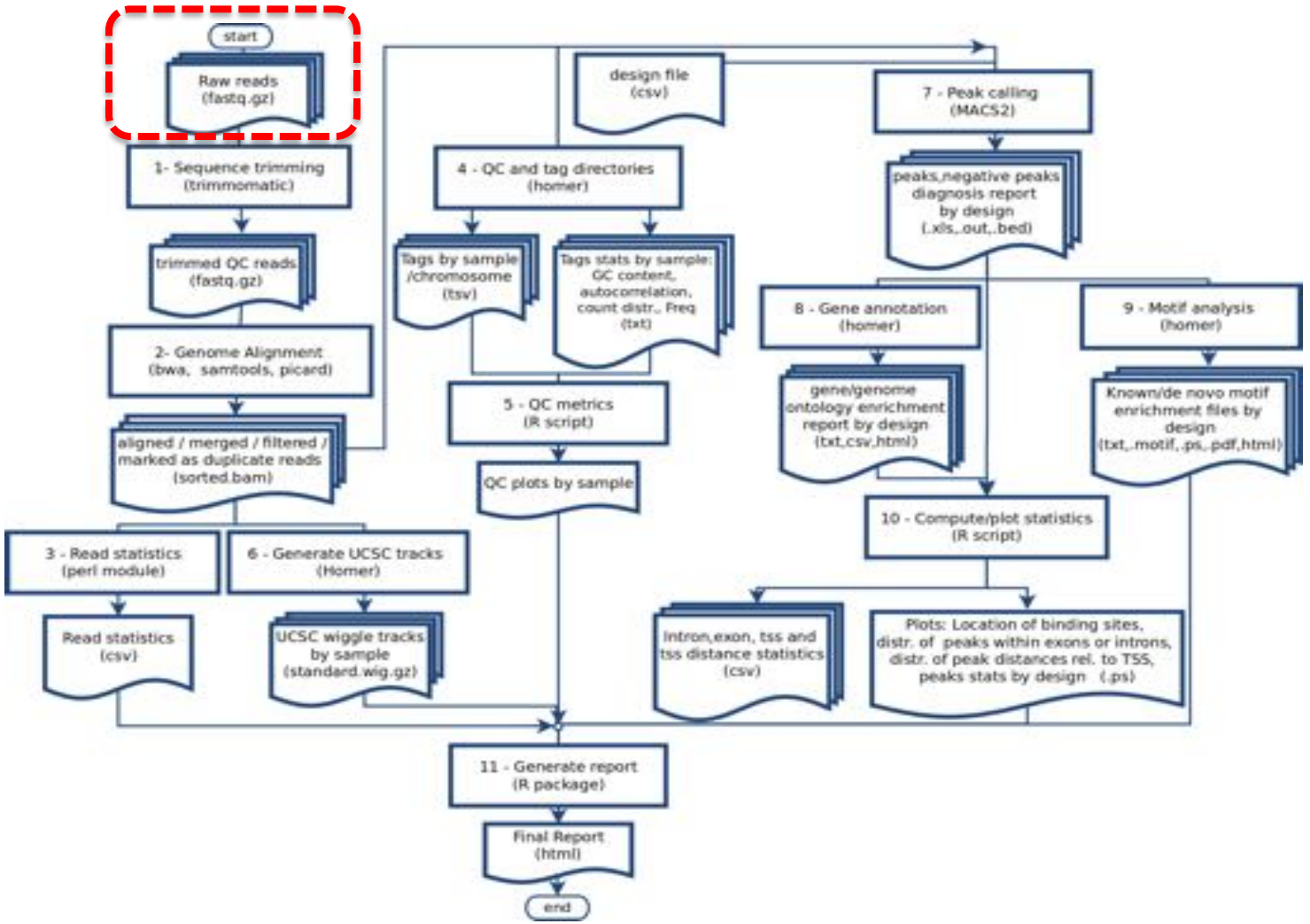


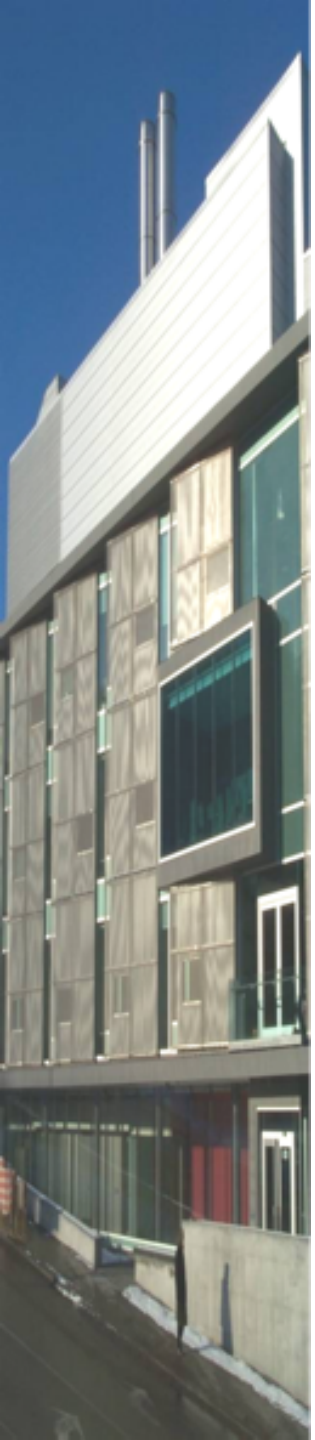
ChIPseq overview





ChIPseq: Input Data





Input Data: FASTQ

End 1

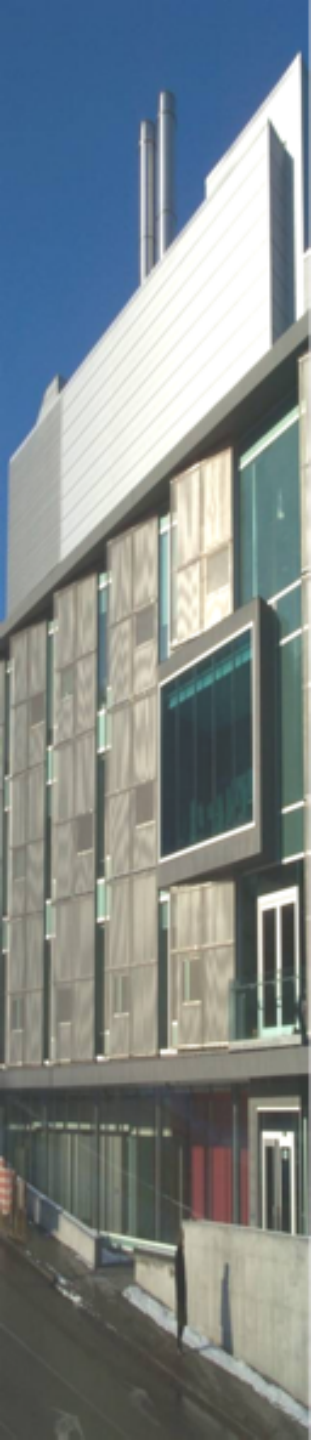
- SKMC_Input_R1.fastq.gz
- AG09309_Input_R1.fastq.gz
- SKMC_H3K4me3_R1.fastq.gz
- AG09309_H3K4me3_R1.fastq.gz

End 2

- SKMC_Input_R2.fastq.gz
- AG09309_Input_R2.fastq.gz
- SKMC_H3K4me3_R2.fastq.gz
- AG09309_H3K4me3_R2.fastq.gz

~ 10Gb each sample

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

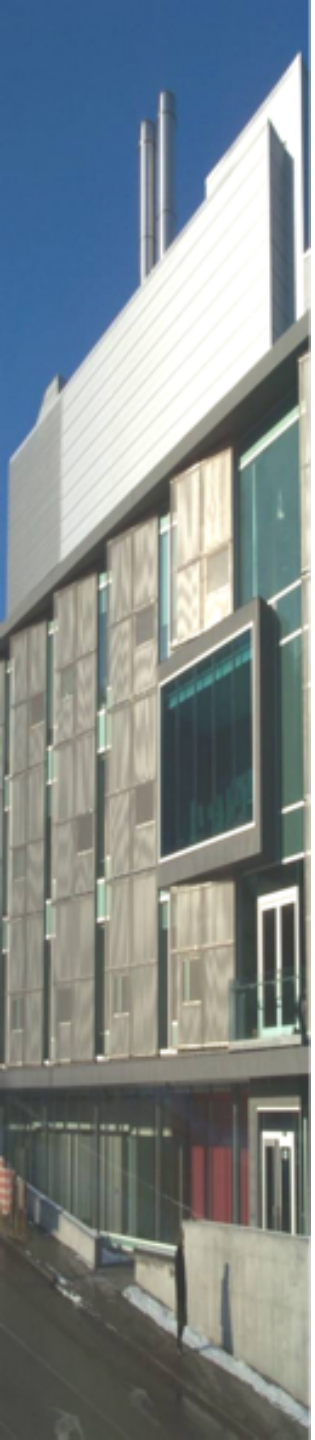


$$Q = -10 \log\_10 (p)$$

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 %



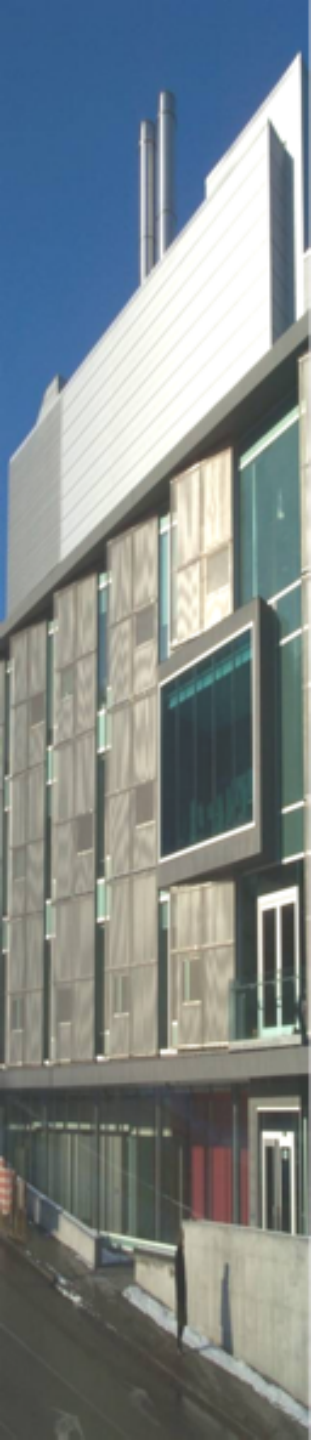
QC of raw sequences

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

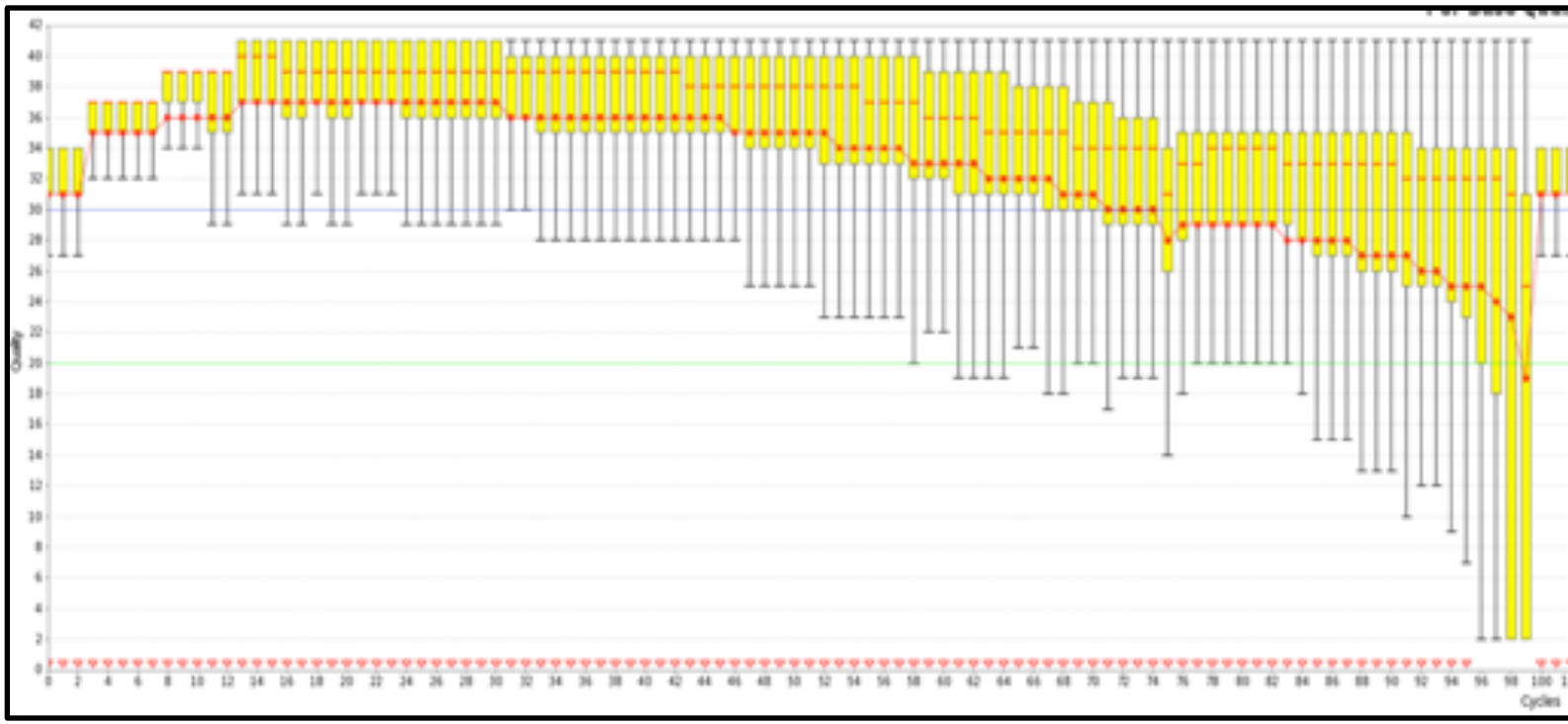
CSV | View/Set Filter | Download Read Files | Help with icons

Read Sets (64 elements) — Add/Remove Column

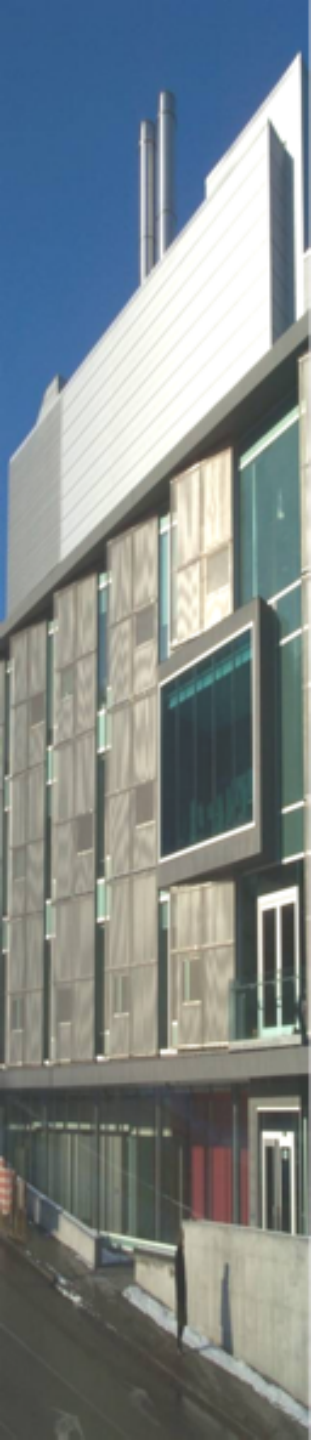
Name	Multiplex Key	Run	Region	QC	Status	Number of reads	Number of Bases	Average Quality	% Duplicate	% Passed Filter	Reads Fastq R1	Reads Fastq R2
W24P	Index_7	1177	4	QC		45,373,280	9,074,656,000	33	21.674	100	(4562MB)	(4546MB)
W25P	Index_8	1177	4	QC		45,066,800	9,013,360,000	33	17.943	100	(4527MB)	(4513MB)
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

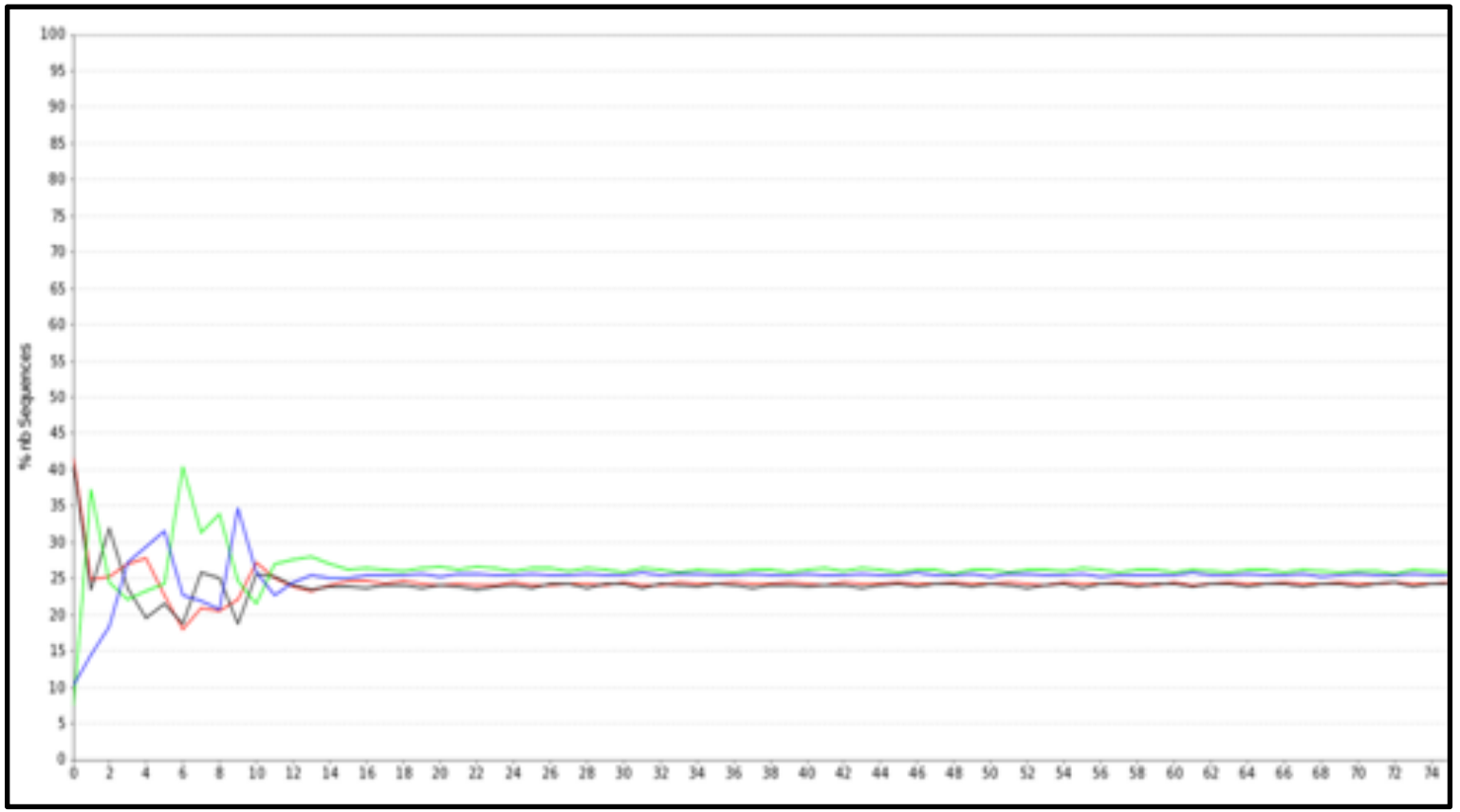


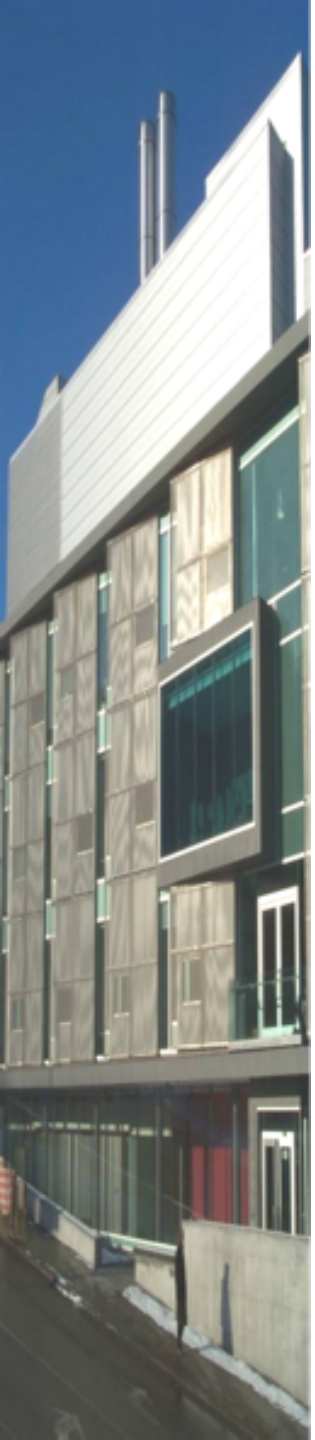
low qualtity bases can bias subsequent anlaysis
(i.e, SNP and SV calling, …)



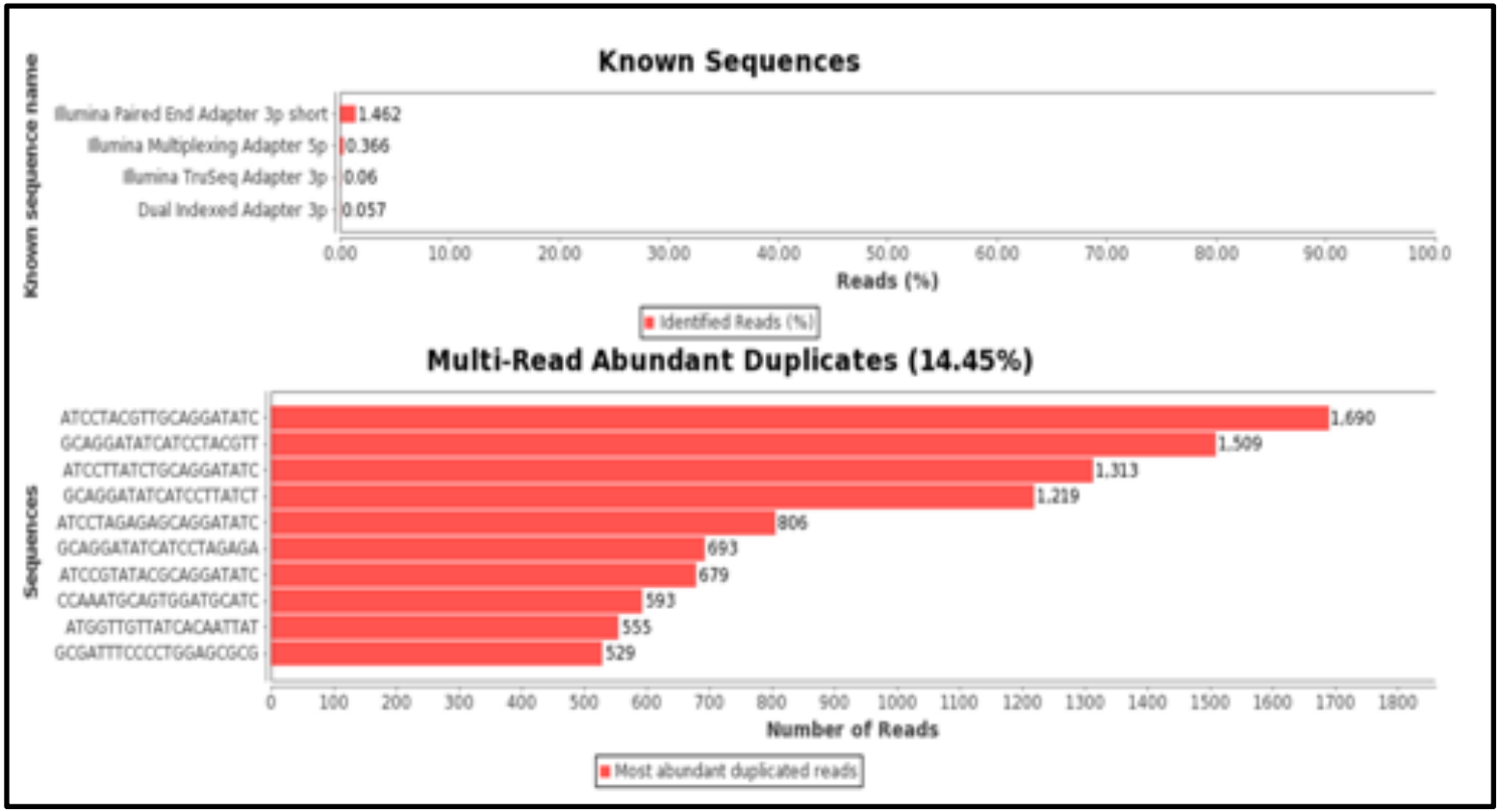
QC of raw sequences

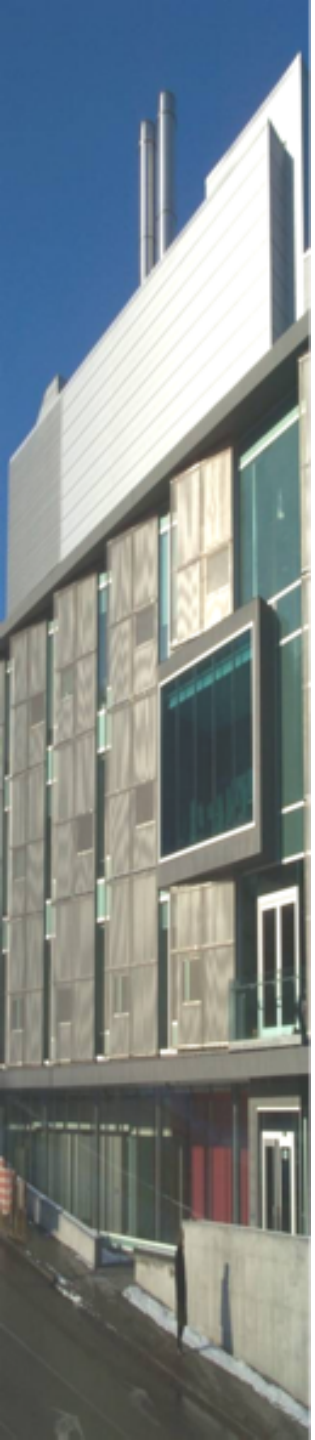
Positional Base-Content





QC of raw sequences



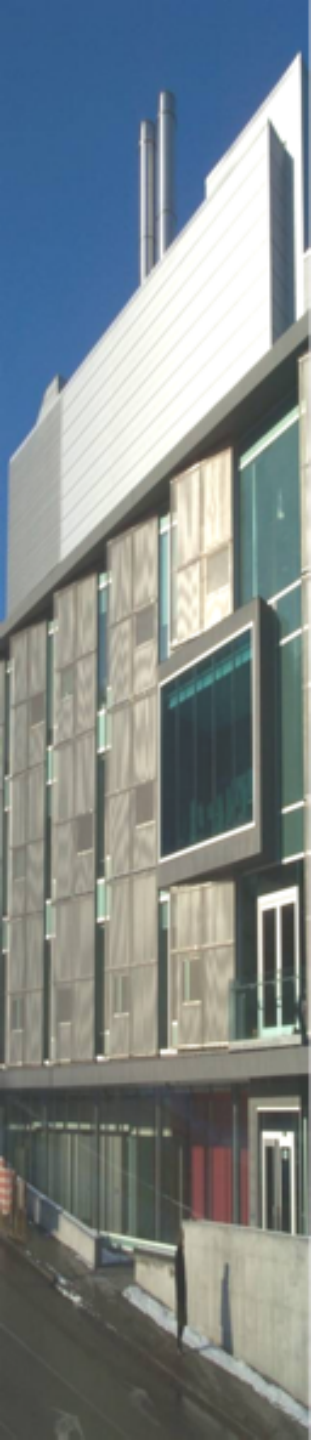


QC of raw sequences

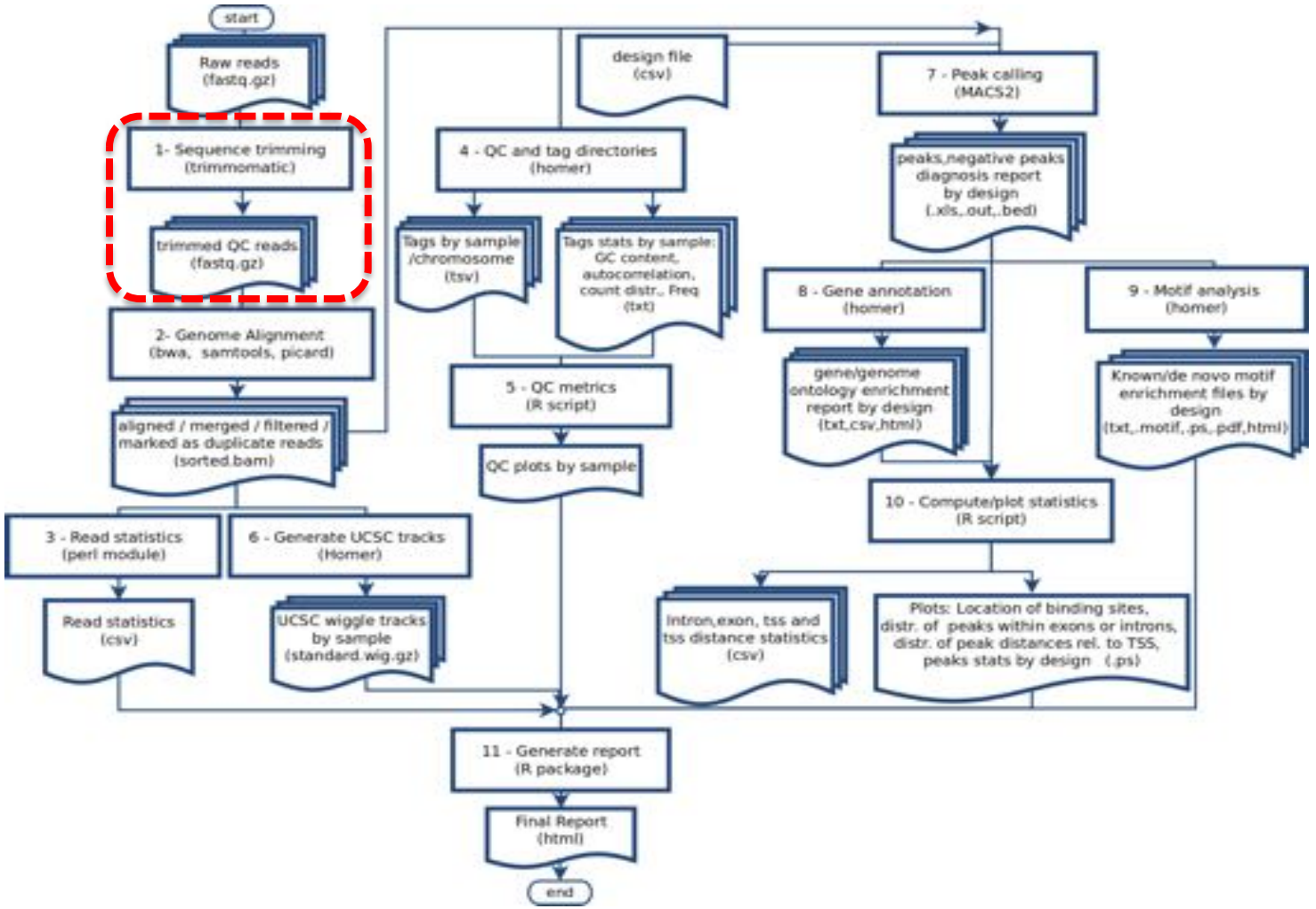
Species composition (via BLAST)

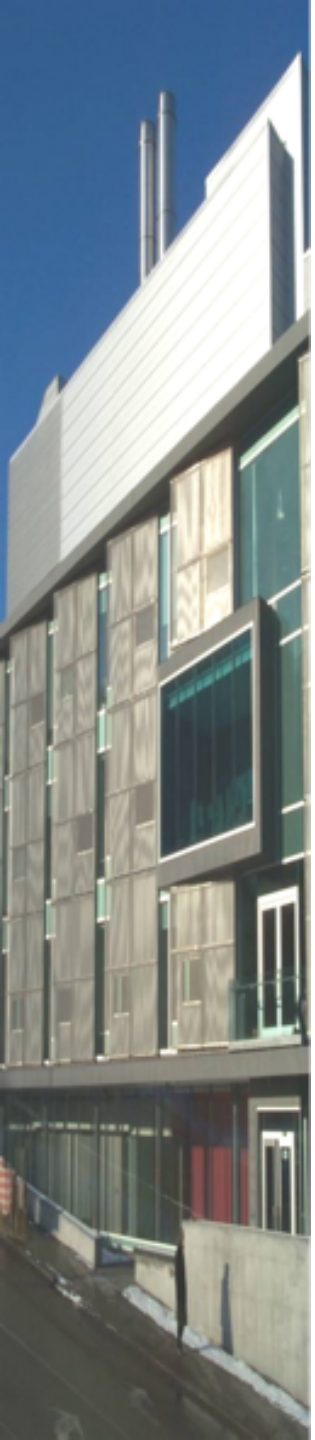
Blast Results (20 elements)

	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



ChIPseq: trimming and filtering

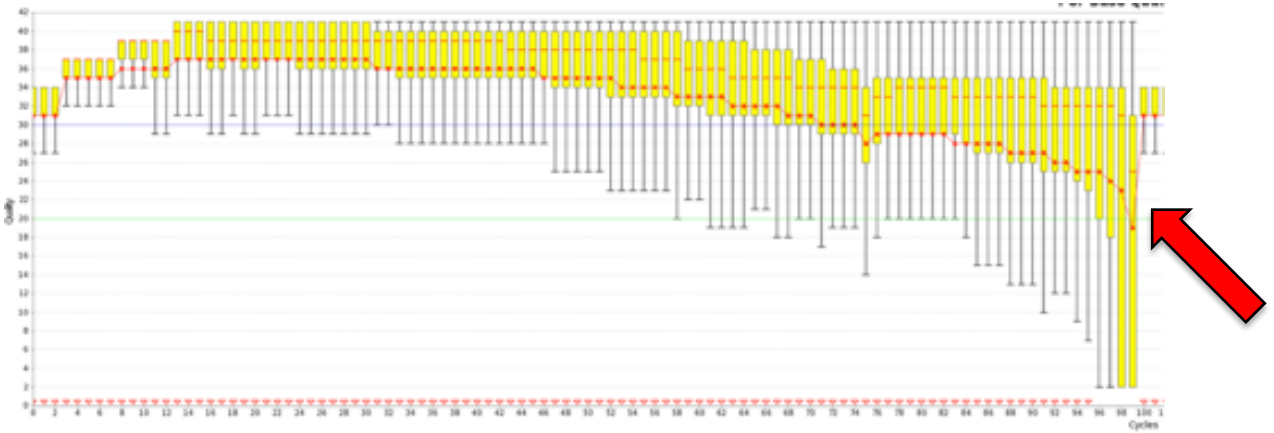




Read Filtering

- Clip Illumina **adapters**:

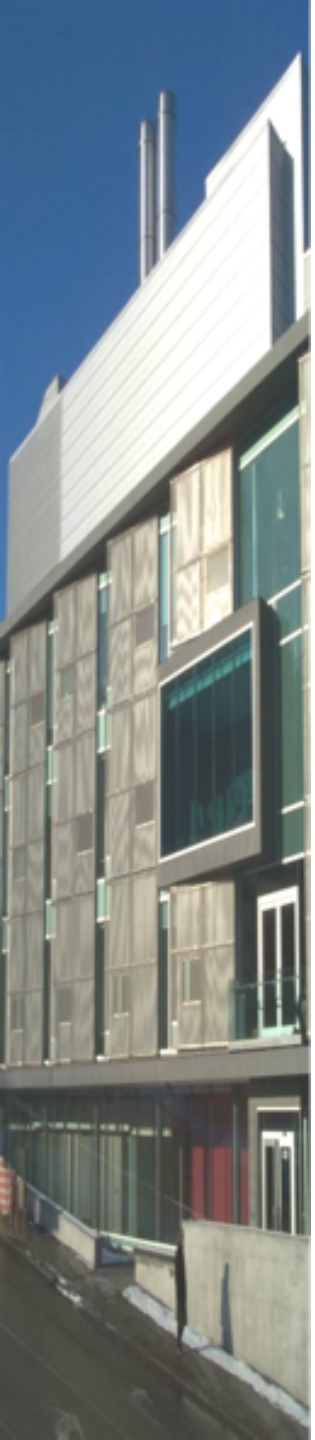
- Trim trailing **quality** < 30



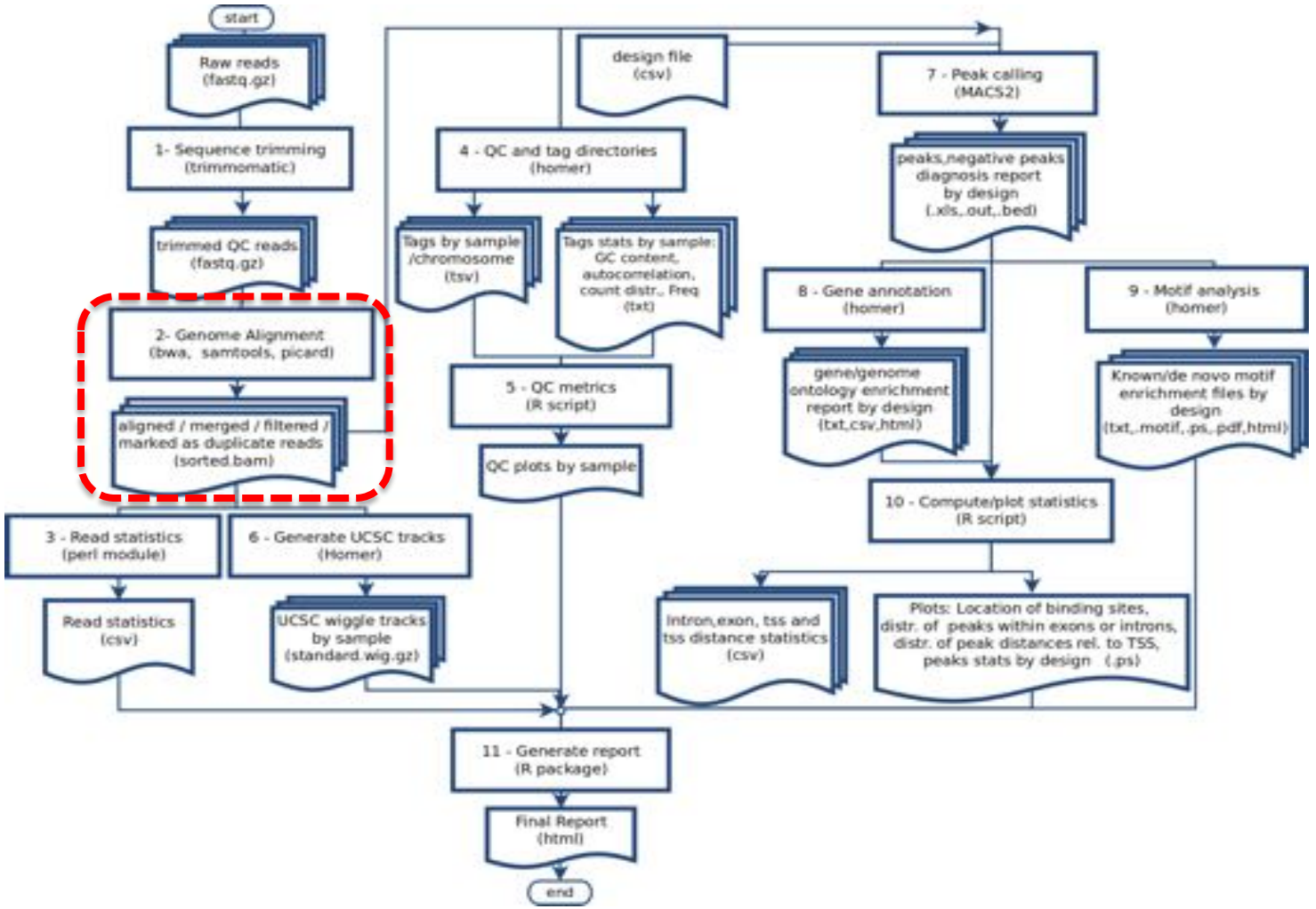
- Filter for read **length** ≥ 32 bp

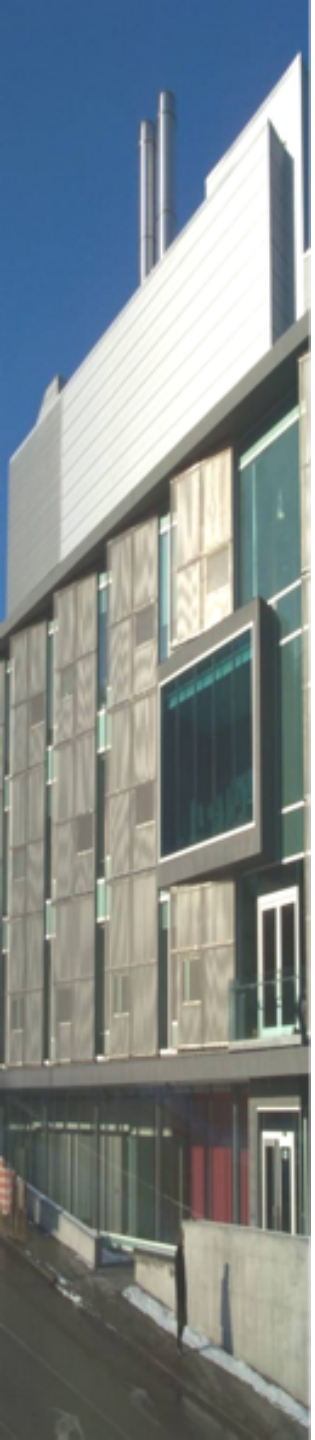
Trimmomatic

usadellab.org



ChIPseq: mapping





Read Mapping

- 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
- Used mapper: BWA
- Other mappers: Bowtie, STAR, GEM, etc.

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

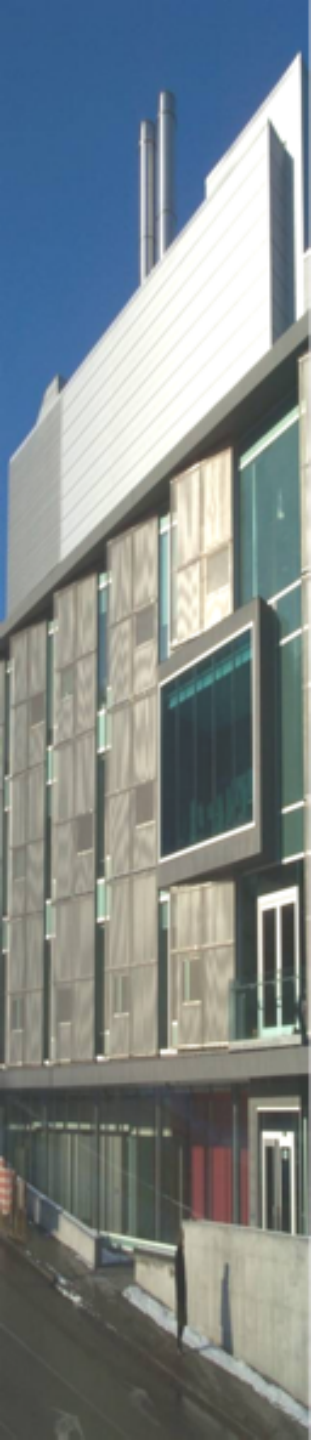
Read name | Flag | Reference Position | CIGAR | Mate Position

```
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

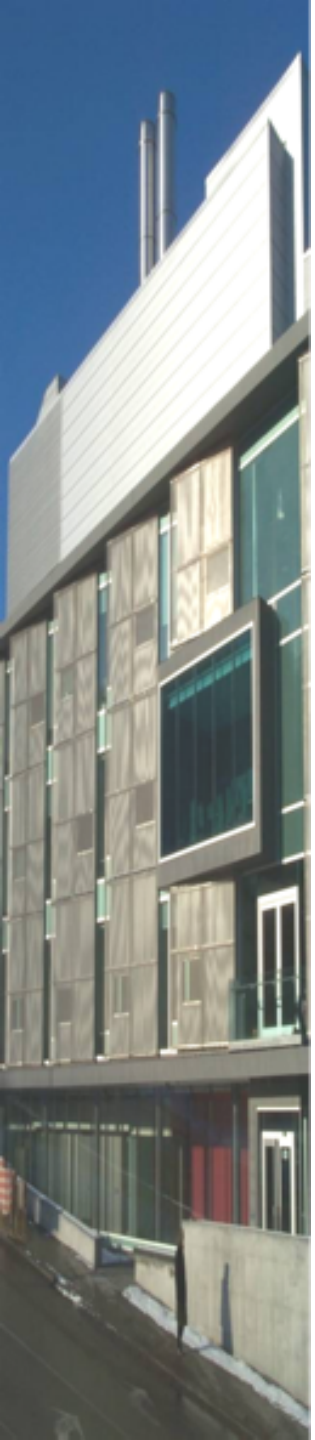
samtools.sourceforge.net

Picard

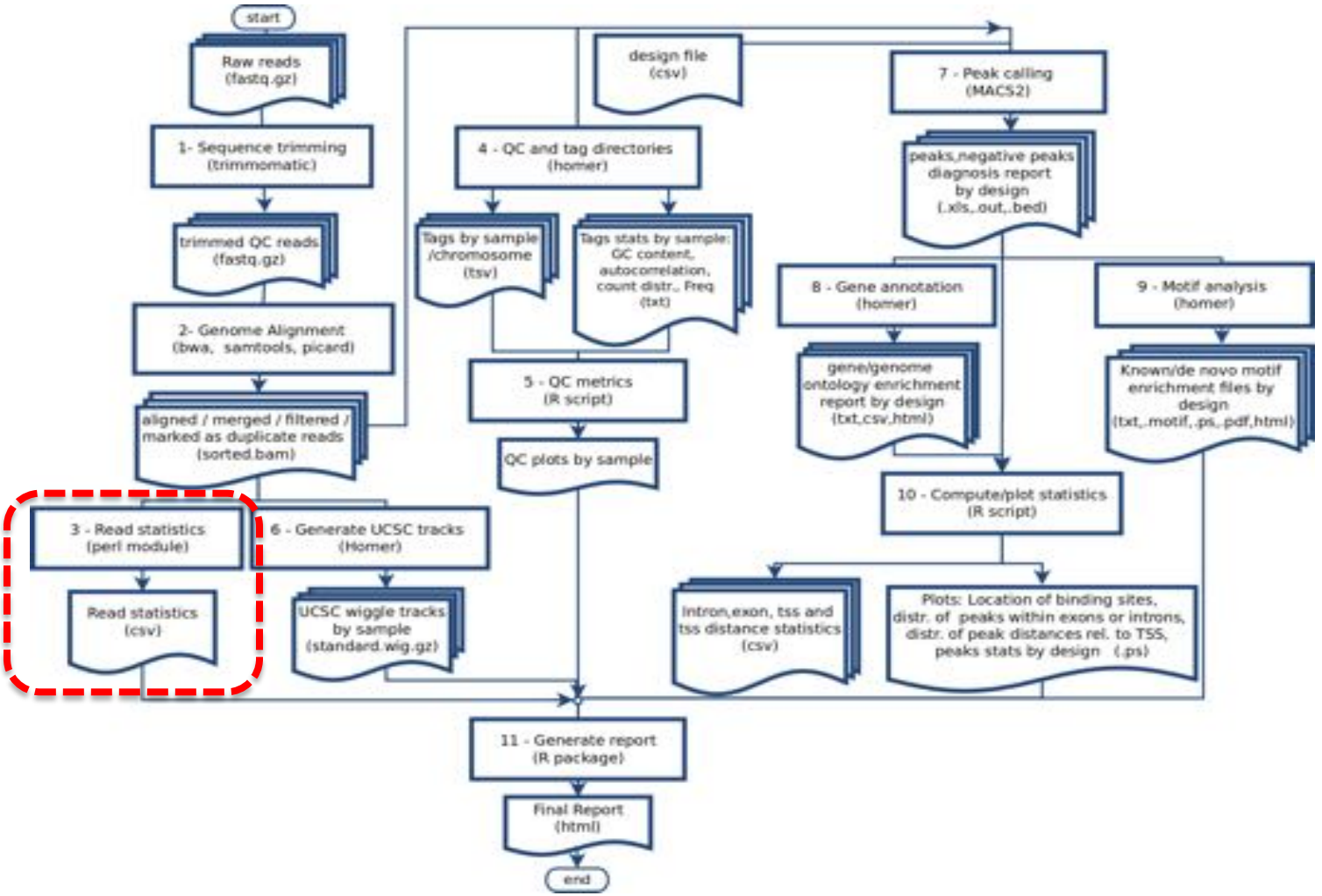
picard.sourceforge.net

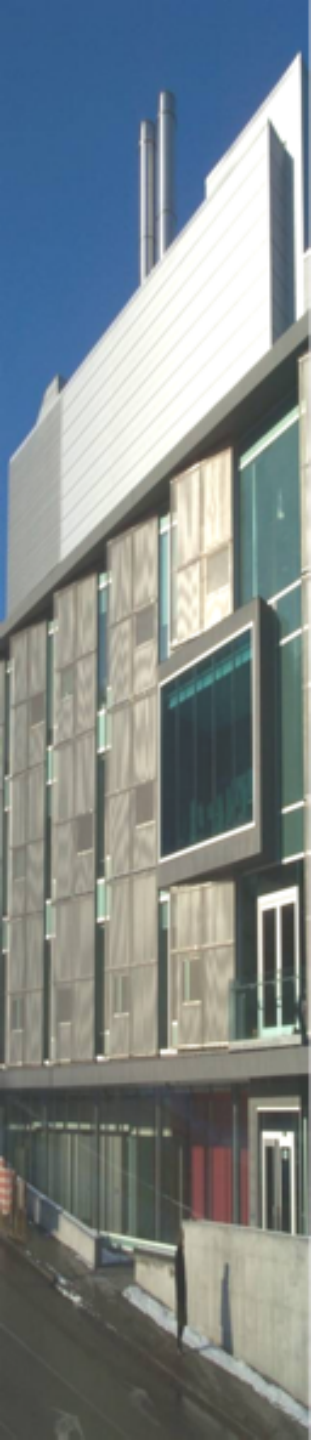
Sort, View, Index, Statistics, Etc.

```
$ 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)
$
```



ChIPseq: metrics





Metrics

- We implemented a small perl library that collects the trimming metrics (from trimmomatic) and the alignment metrics (samtools flagstats)

GET FULL TABLE

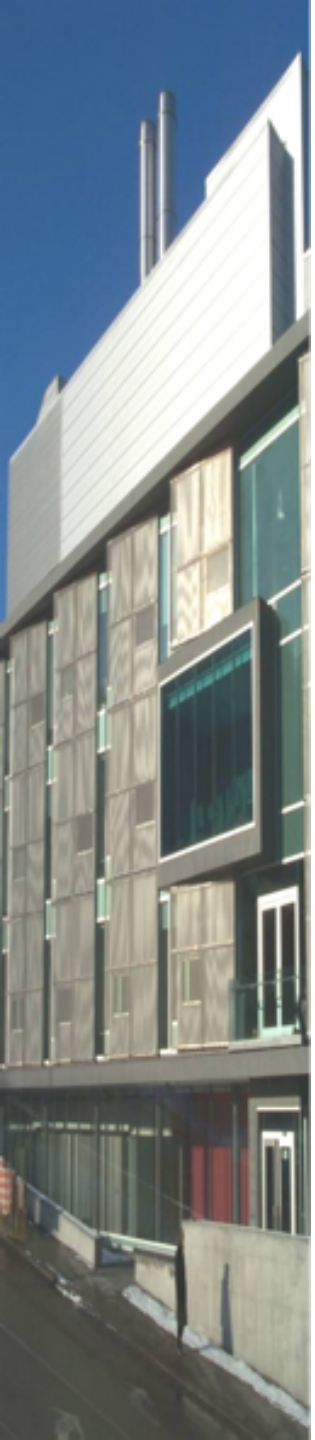
Table 2. Per sample trimming and alignment statistics

SampleName	Nb.QC_Passed.Reads	Nb.Aligned.Reads	Nb.Duplicate.Reads	pct.Aligned	pct.Duplicate
UW_ChipSeq_SKMC_H3K4me3	50137661	47490634	0	94.7204816754415	0
UW_ChipSeq_SKMC_Input	26058656	25384350	0	97.4123531159857	0

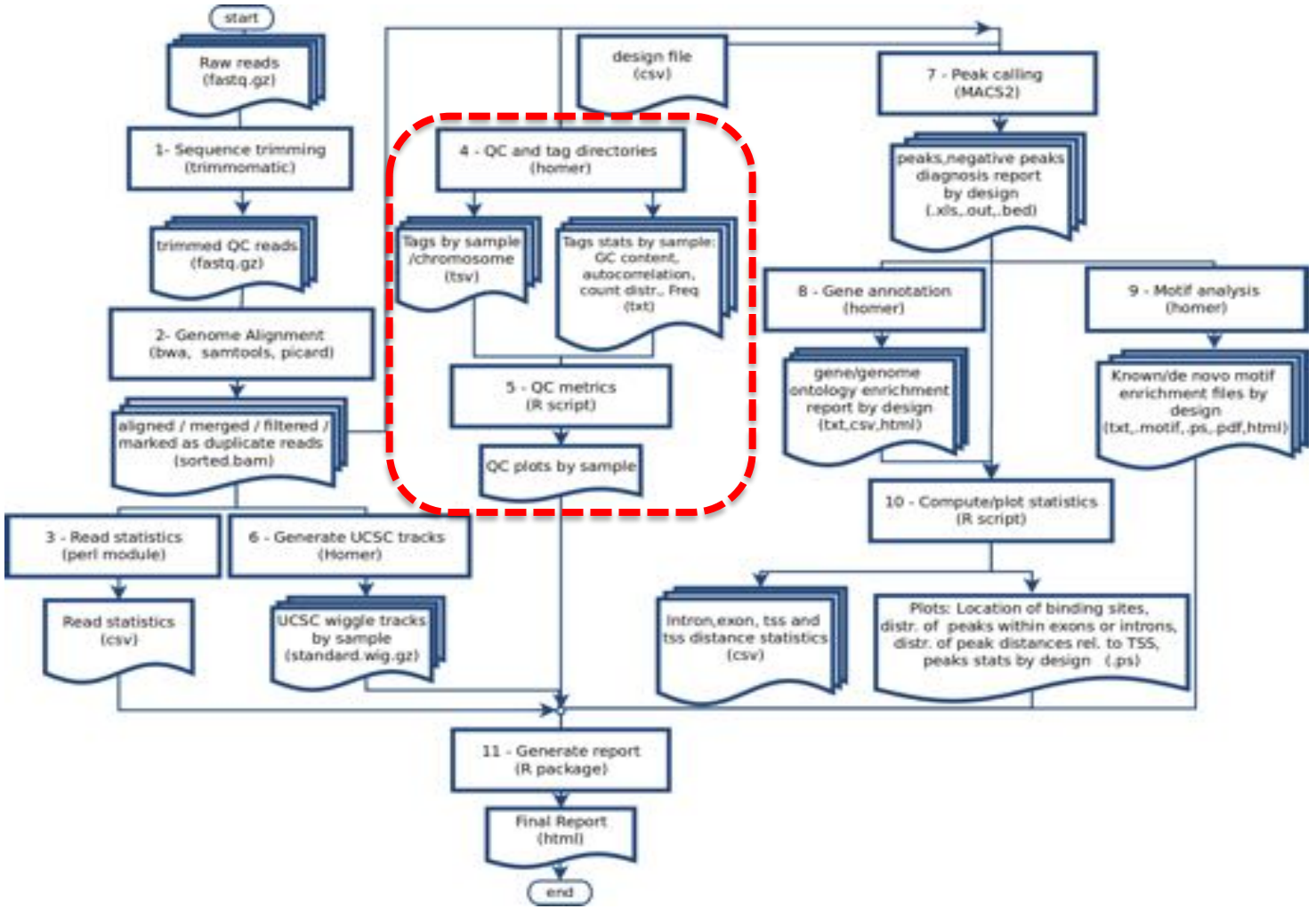
GET FULL TABLE

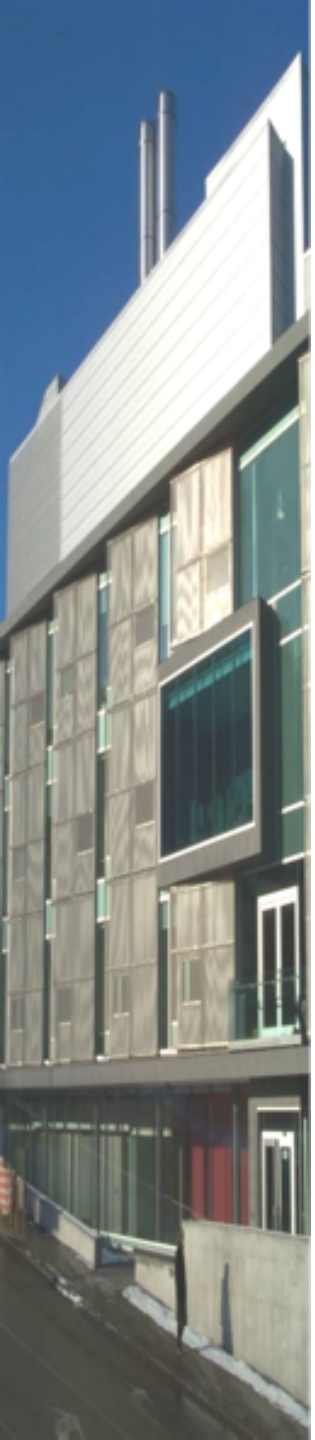
Table 3. Per lane trimming and alignment statistics

SampleName	Nb.QC_Passed.Reads	Nb.Aligned.Reads	Nb.Duplicate.Reads	pct.Aligned	pct.Duplicate
UW_ChipSeq_SKMC_H3K4me3 X_X GSM945214	23611562	22893919	0	96.9606288647909	0
UW_ChipSeq_SKMC_H3K4me3 X_Y GSM945214	26526099	24596715	0	92.7264691276316	0
UW_ChipSeq_SKMC_Input X_X GSM945161	26058656	25384350	0	97.4123531159857	0



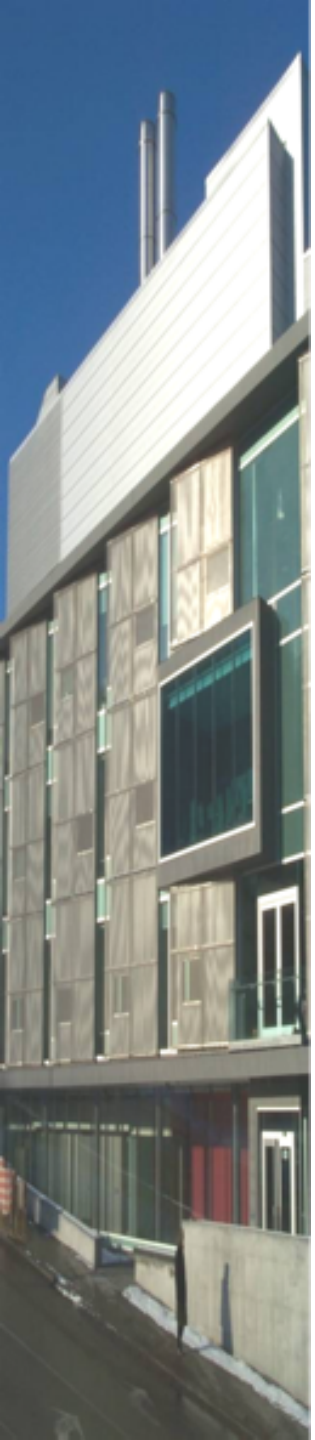
ChIPseq: QC and tag directory



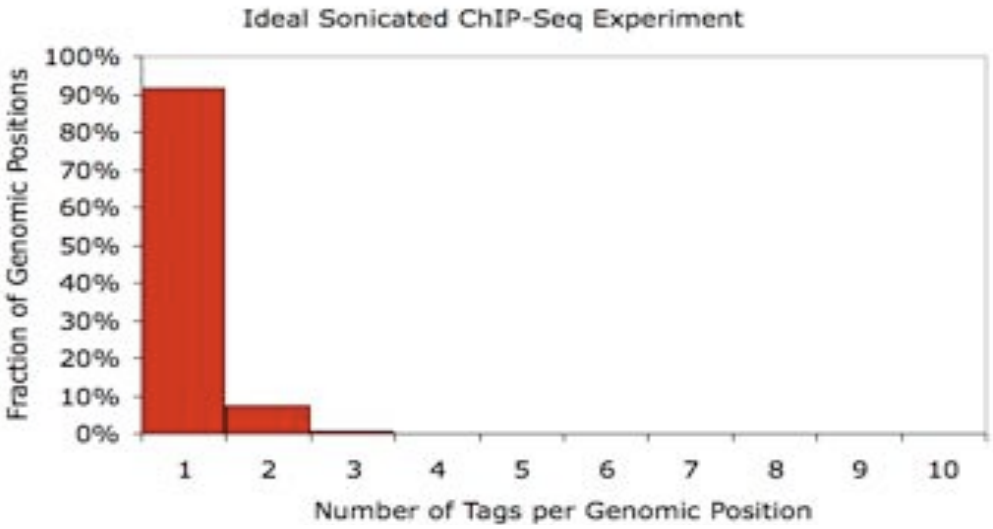


Homer - QC and tags

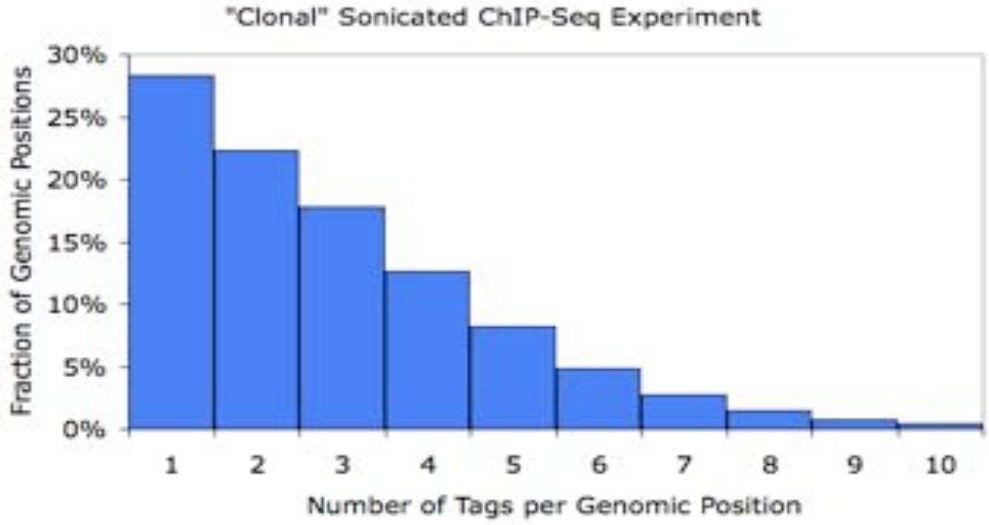
- During this phase several important parameters are estimated that are later used for downstream analysis, such as the estimated length of ChIP-Seq fragments
- Homer transforms the sequence alignment into platform independent data structure representing the experiment.
  - Clonal Tag Counts
  - Sequencing Fragment Length Estimation (tag autocorrelation)



HOMER – Clonal tag count

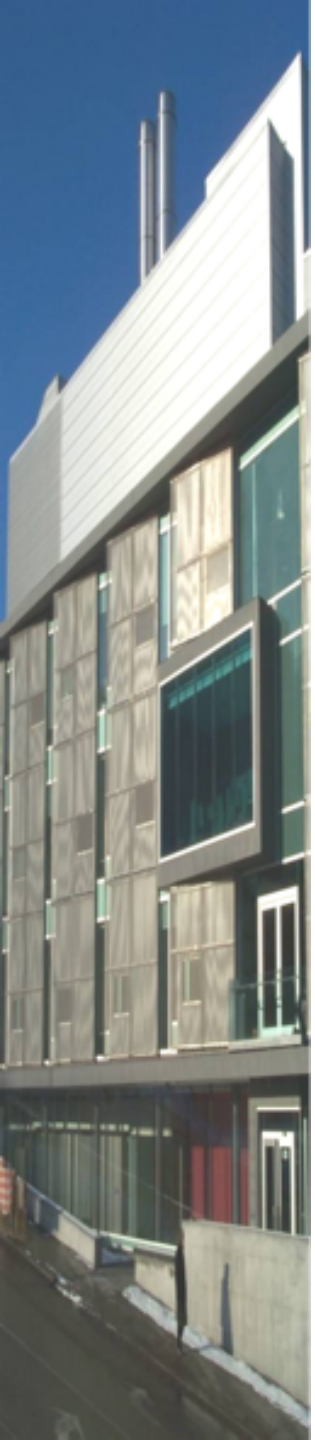


GO for subsequent analysis



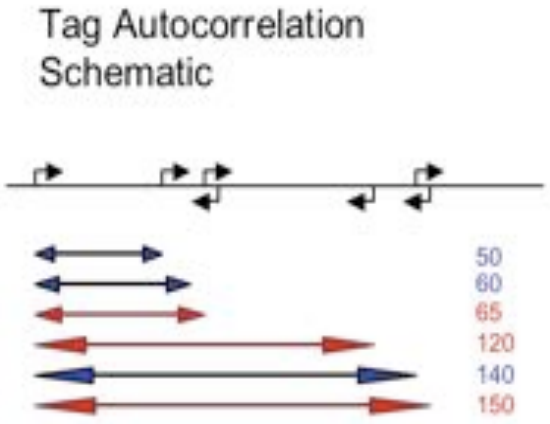
redo the ChIP and re-prep the sample for sequencing

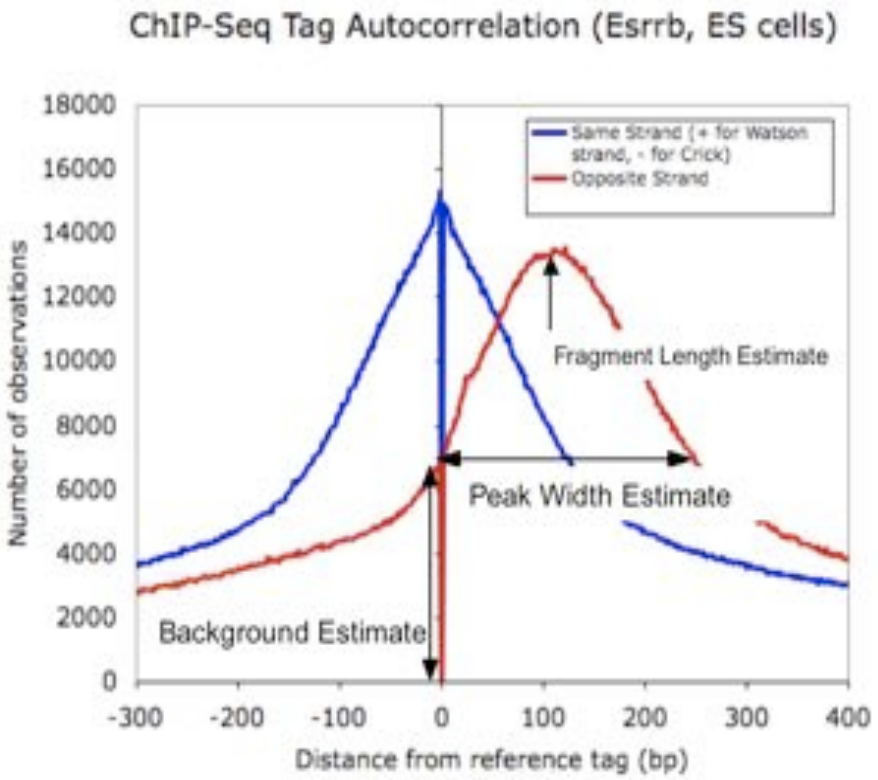
Modified from http://biowhat.ucsd.edu/homer/chipseq/qc.html



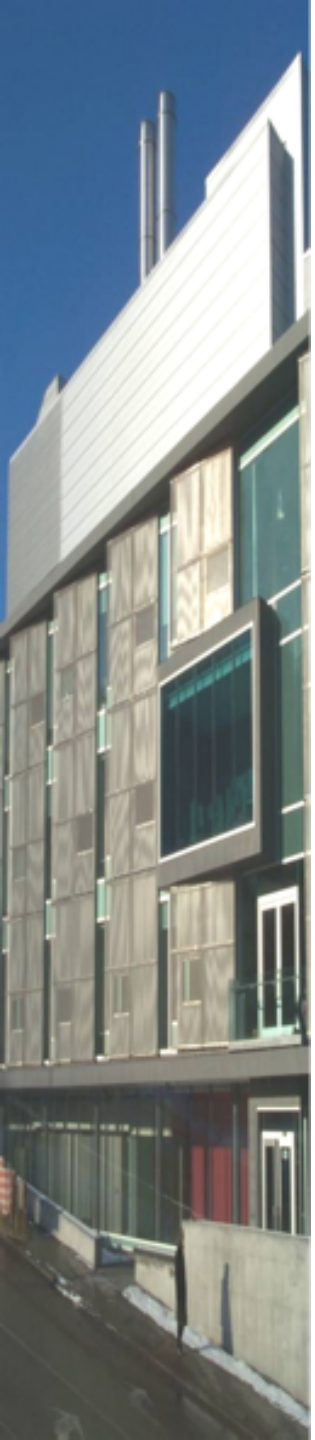
HOMER - Sequencing Fragment Length Estimation

- The specific size of fragments sequenced for a given experiment can be very important in extracting meaningful data and precisely determining the location of binding sites.

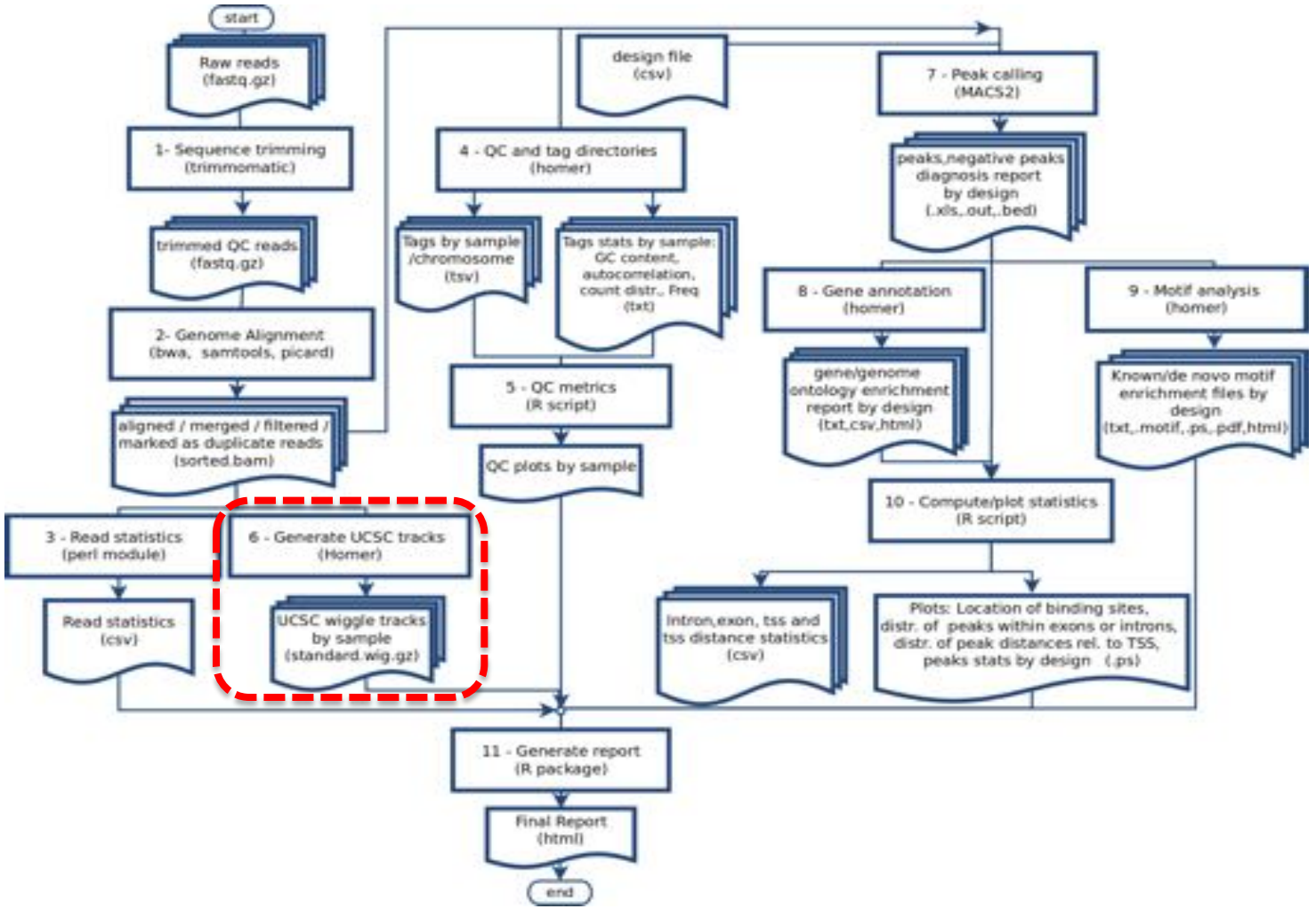


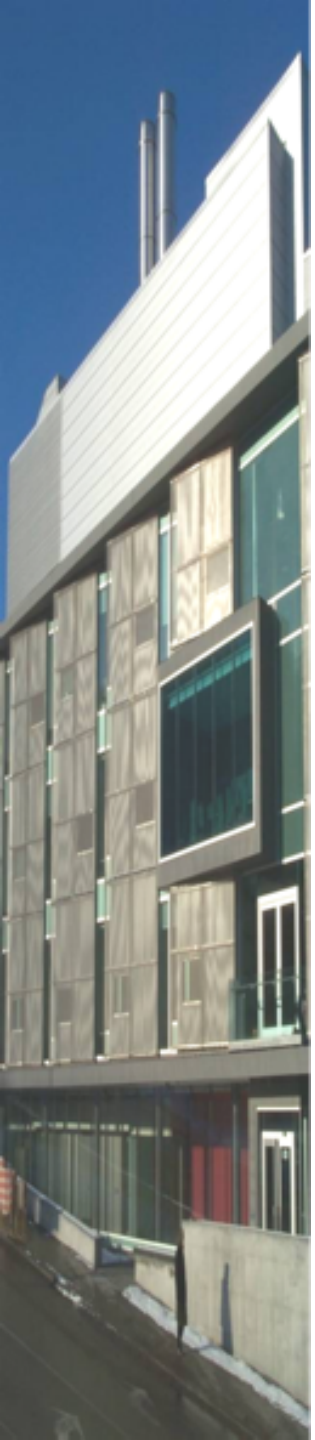


Modified from http://biowhat.ucsd.edu/homer/chipseq/qc.html



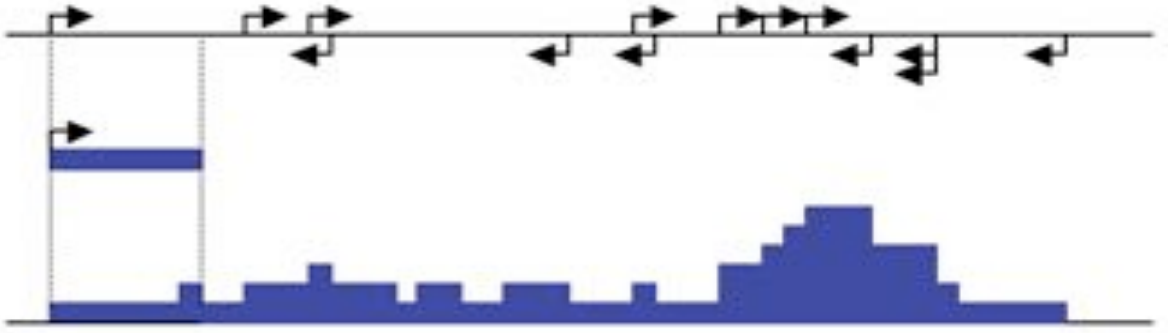
ChIPseq: Generate UCSC tracks



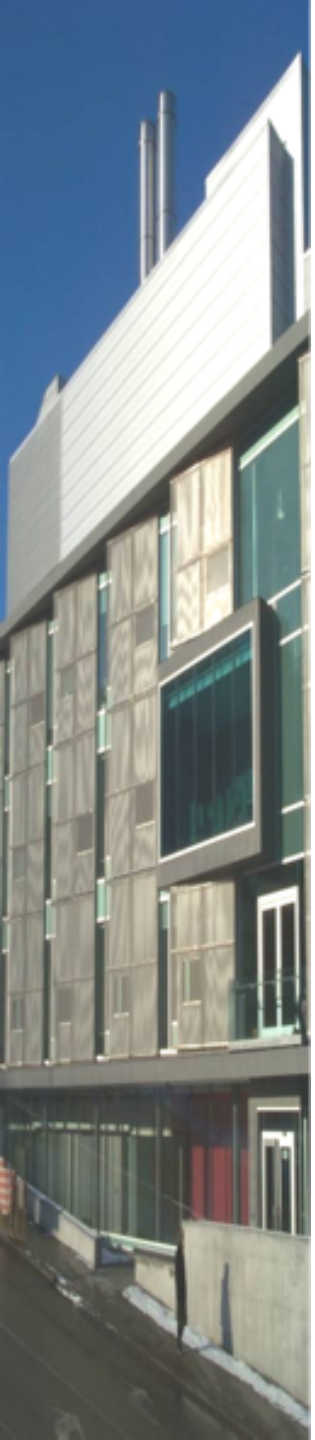


HOMER – UCSC visualisation

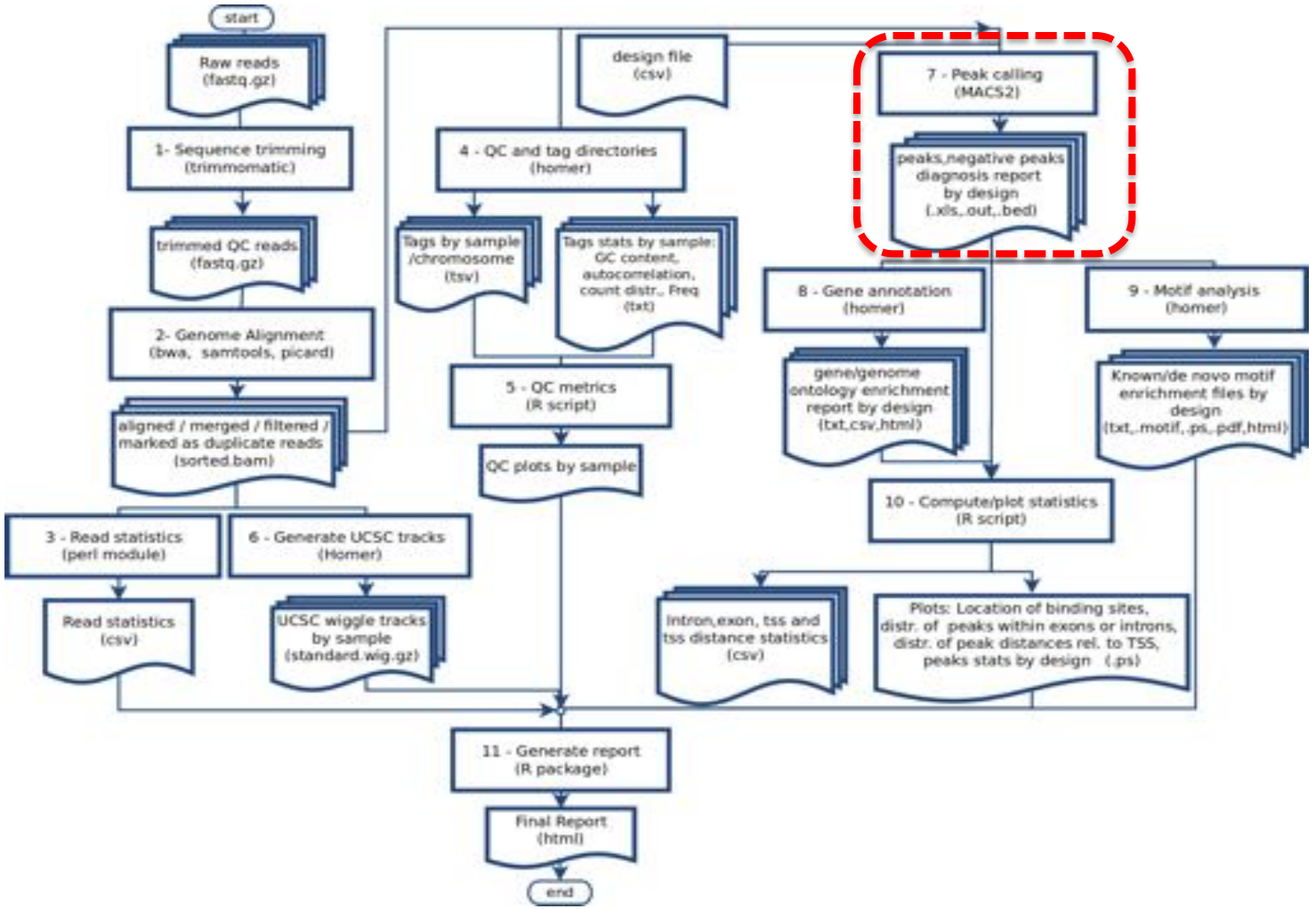
- It approximates the ChIP-fragment density at each position in the genome. This is done by starting with each tag and extending it by the estimated fragment length.
- The ChIP-fragment density is then defined as the total number of overlapping fragments at each position in the genome

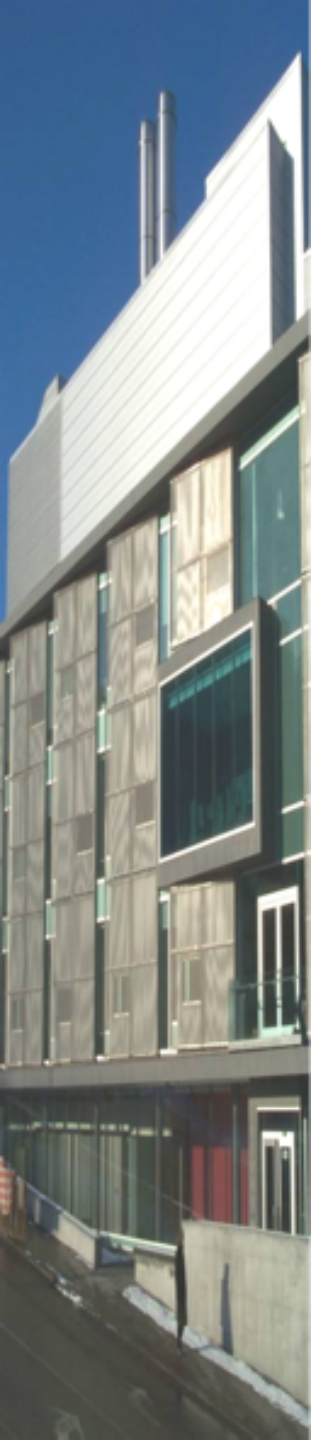


Modified from http://biowhat.ucsd.edu/homer/chipseq/ucsc.html



ChIPseq: Peak calling





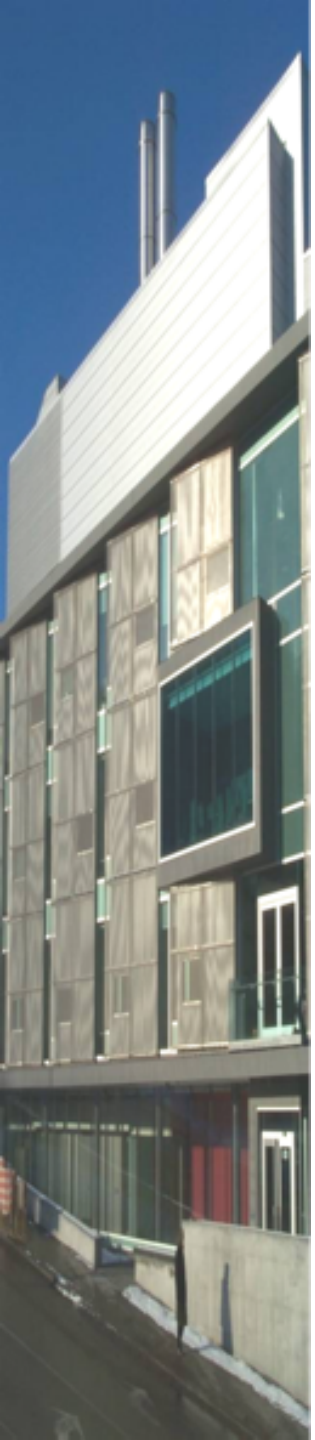
MACs2

MACS2:

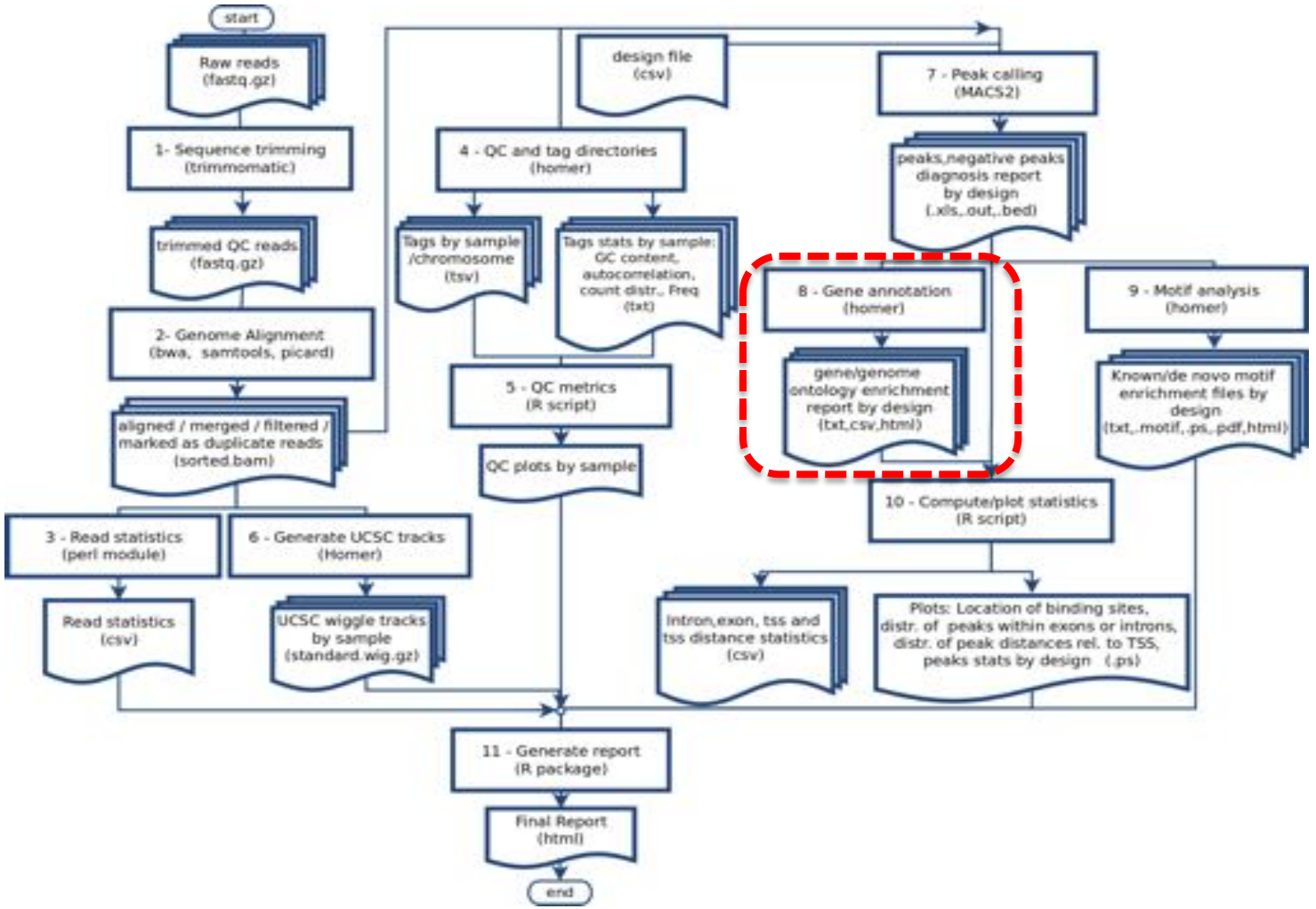
- Negative peaks file is not generated in MACS2, since MACS use q-value to replace empirical FDR (MACS1.4).
  - eFDR is calculated by calling negative peaks as false positives
  - Thus to generate a negative peak list, an additional design with the group indicators inversed must be added

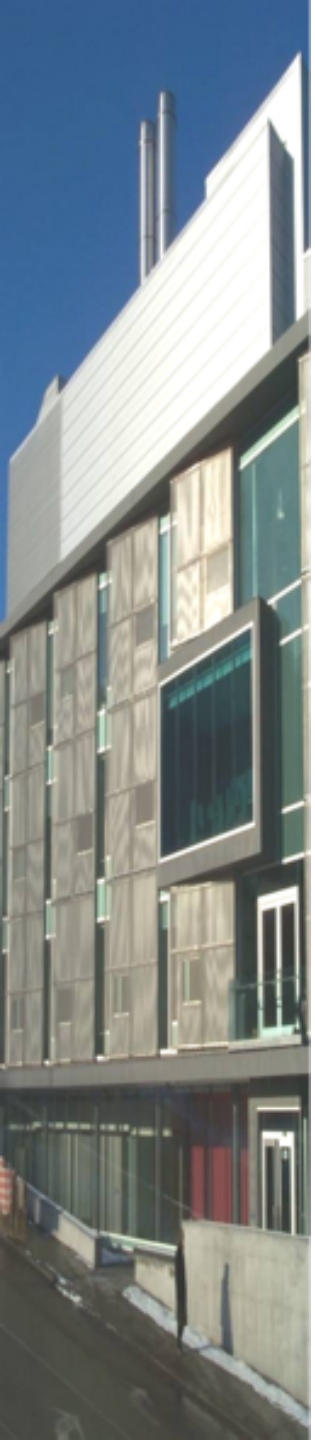
Files generated with MACS2:

- designName.diag.macs.out
- designName_model.r
- designName_peaks.bed
- designName_peaks.encodePeak
- designName_peaks.xls,
- designName_summits.bed



ChIPseq: Gene annotation

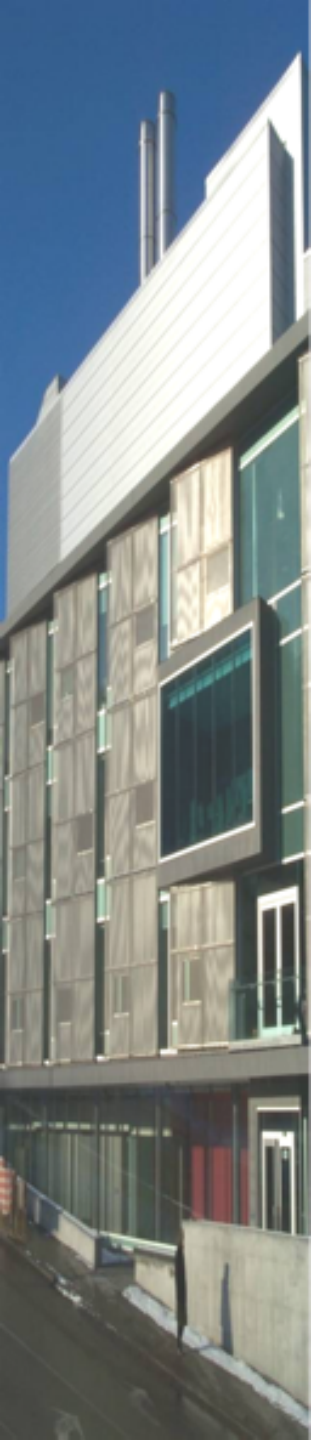




HOMER - annotation

- It efficiently assigns peaks to one of millions of possible annotations genome wide (refSeq):
  - TSS (by default defined from -1kb to +100bp)
  - TTS (by default defined from -100 bp to +1kb)
  - CDS Exons
  - 5' UTR Exons
  - 3' UTR Exons
  - Introns
  - Intergenic
- In addition HOMER can perform Gene Ontology Analysis

Modified from http://biowhat.ucsd.edu/homer/ngs/annotation.html



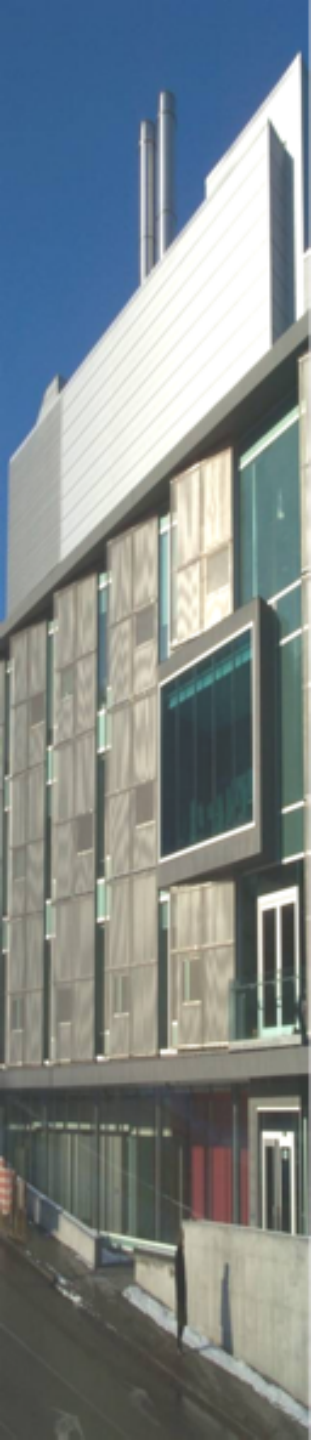
HOMER – annotation outputs

Files generated for each design:

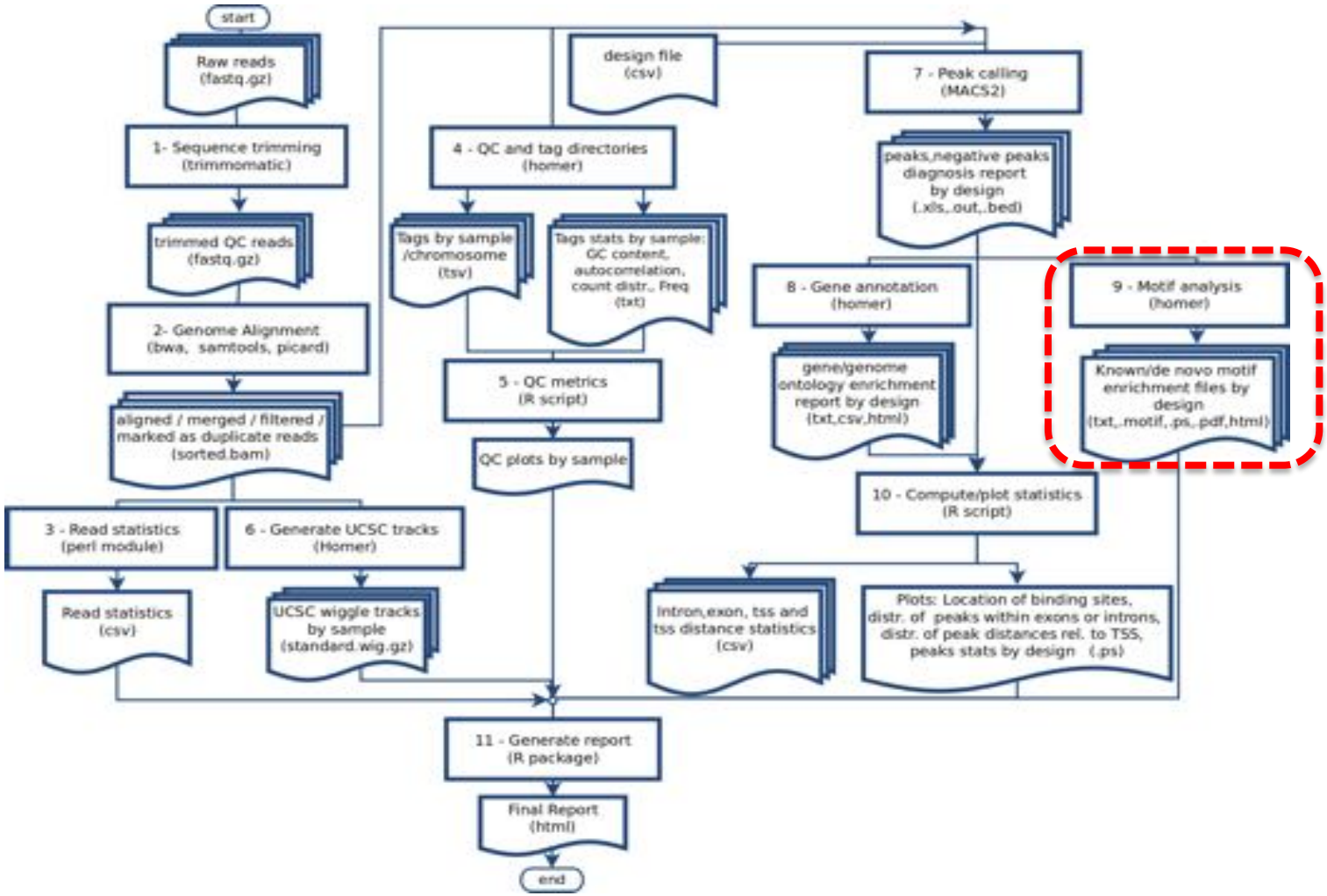
- designName.annotated.csv

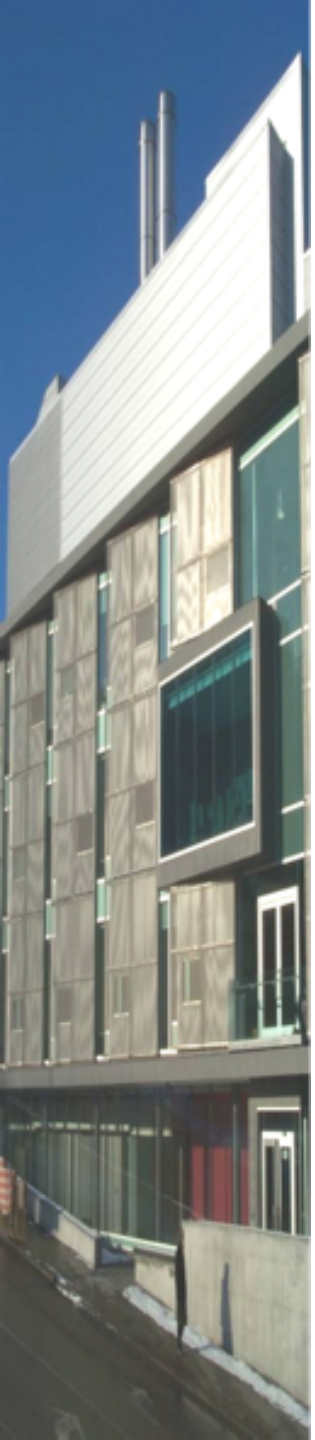
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	intergenic	intergenic	-50894	NM_001185	79670	Hs.597057	NM_001185	ENSG000000	ZCCHC6	DKFZp66681	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	intergenic	intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	intergenic	intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_17	intron (NM_17	81017	NM_001195	145282	Hs.660396	NM_001195	ENSG000001	MIPOL1	DKFZp313M	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembr
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	intergenic	intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	intergenic	intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CK	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	intergenic	intergenic	-8849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

- geneOntology.html
- GenomeOntology.html



ChIPseq: Motif analysis

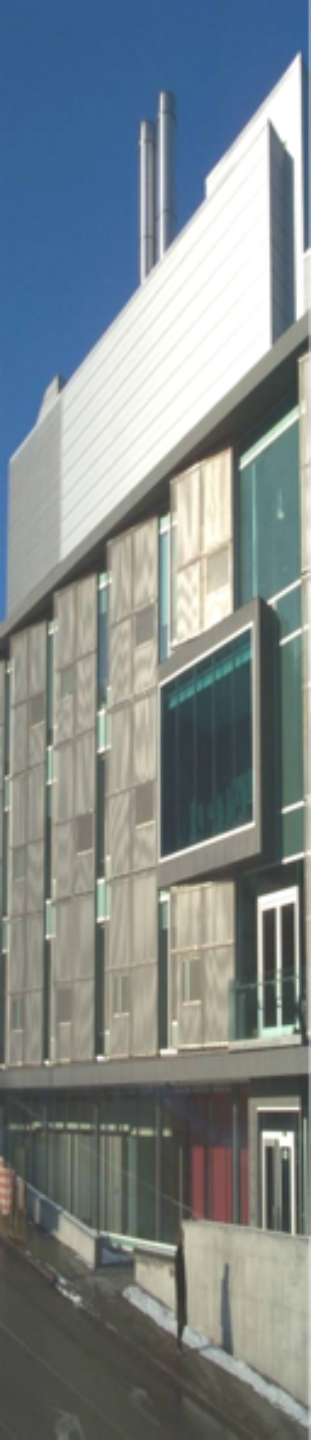




HOMER - Motifs

- De Novo and Known motif analysis:
  - It tries to identify the regulatory elements that are specifically enriched in on set relative to the other.
  - It uses ZOOPS scoring (zero or one occurrence per sequence) coupled with the hypergeometric enrichment calculations (or binomial) to determine motif enrichment.
  - It also tries to account for sequenced bias in the dataset

Modified from http://biowhat.ucsd.edu/homer/motif/index.html

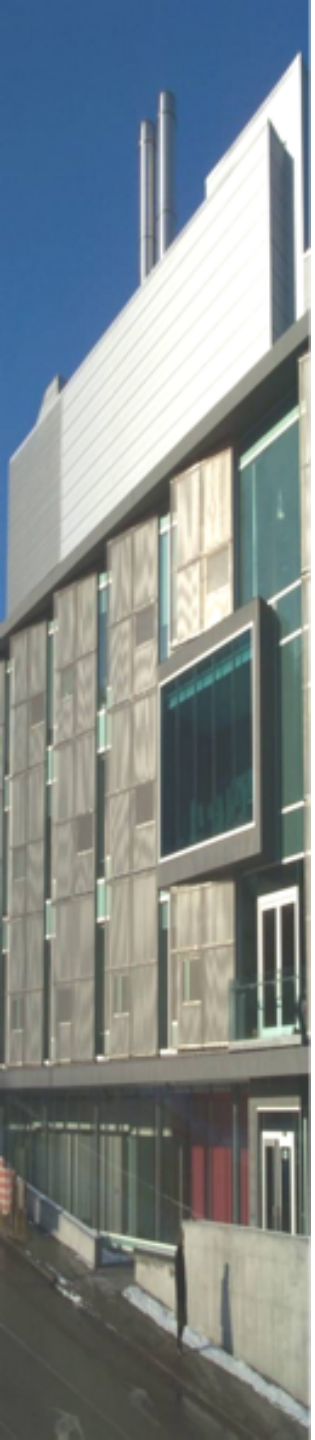


HOMER – Motifs output

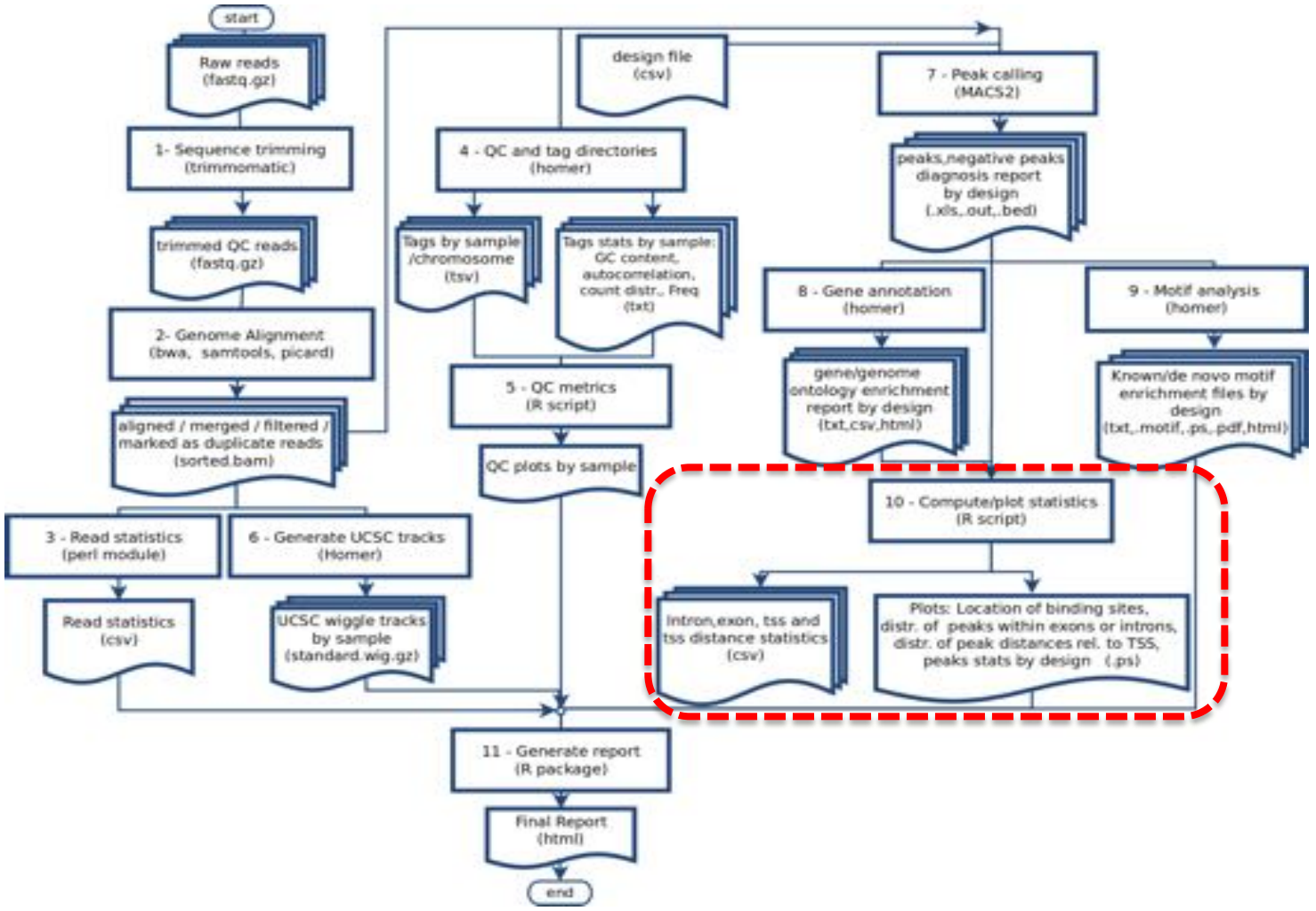
- File generated for each design:
  - homerResults.html
  - knownResults.html

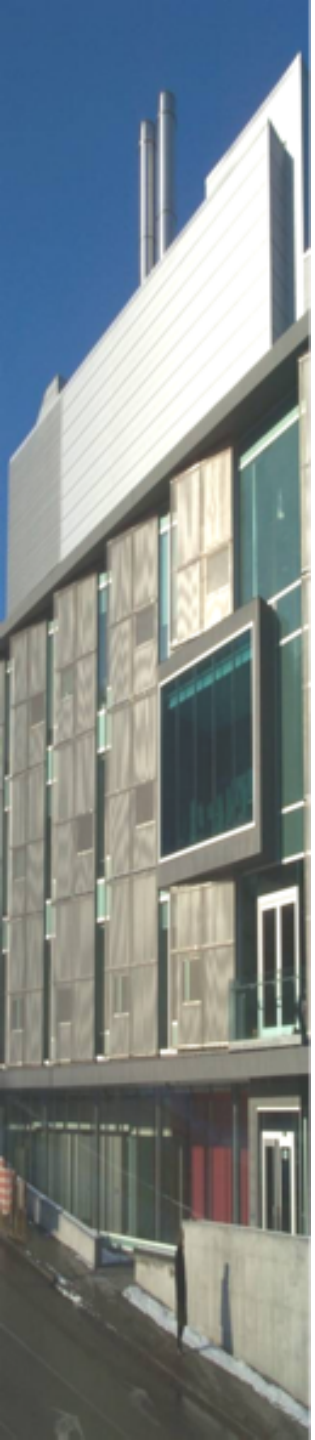
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-12661	-2.915e+04	70.91%	15.19%	40.5bp (65.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2		1e-578	-1.332e+03	27.14%	16.52%	54.0bp (65.5bp)	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
3		1e-384	-8.860e+02	17.77%	10.53%	53.9bp (62.1bp)	Unknown/Homeobox /Limb-p300-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
4		1e-164	-3.783e+02	3.17%	1.28%	52.2bp (62.9bp)	PH0048.1_Hoxa13 More Information | Similar Motifs Found	motif file (matrix)
5		1e-151	-3.485e+02	3.38%	1.47%	50.2bp (65.4bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
6		1e-107	-2.485e+02	1.21%	0.35%	56.3bp (69.7bp)	CTCF(Zf)/CD4+-CTCF-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
7		1e-72	-1.671e+02	2.10%	1.02%	55.1bp (58.5bp)	MA0029.1_Evi1 More Information | Similar Motifs Found	motif file (matrix)

Modified from http://biowhat.ucsd.edu/homer/ngs/peakMotifs.html



ChIPseq: Plots

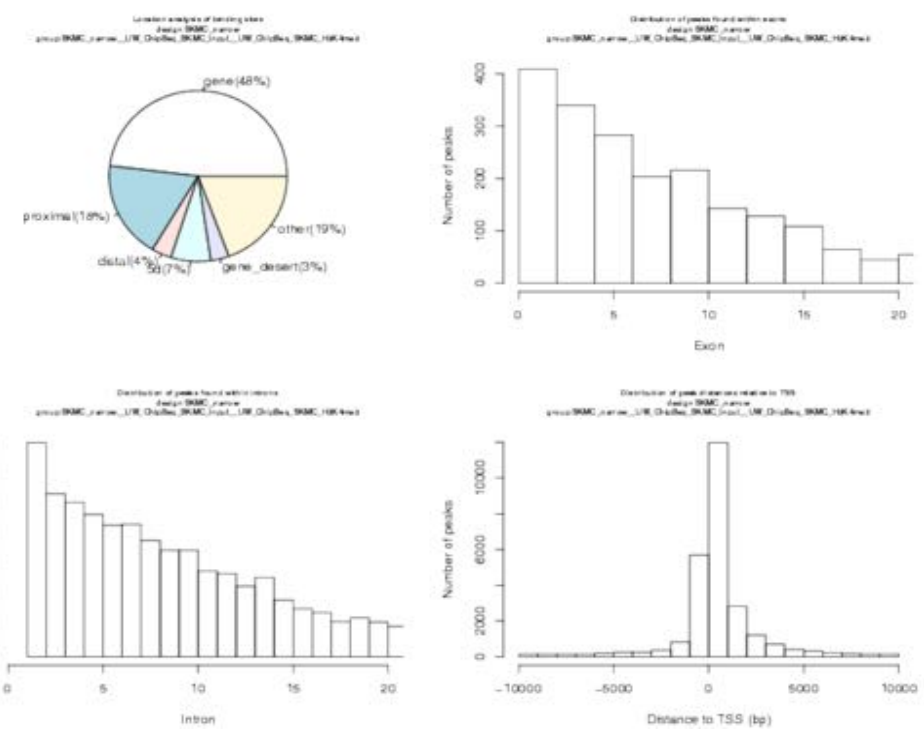


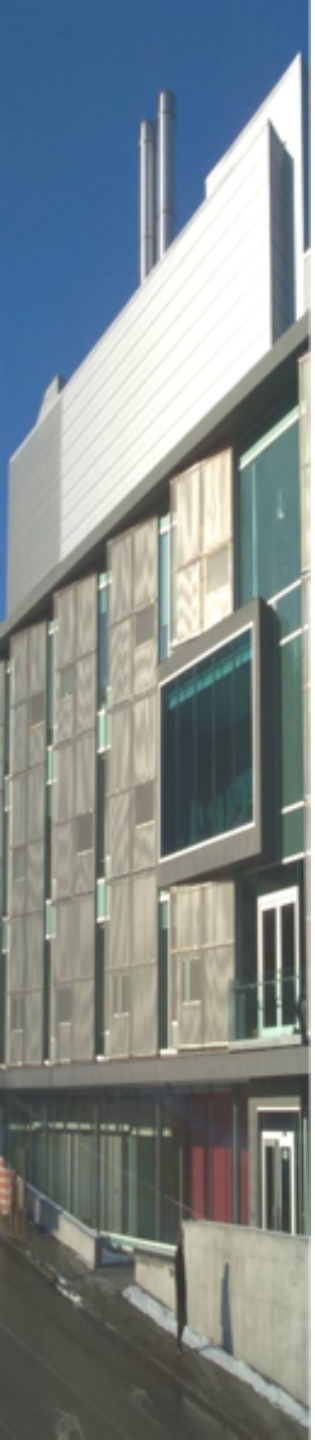


Home-made Rscript

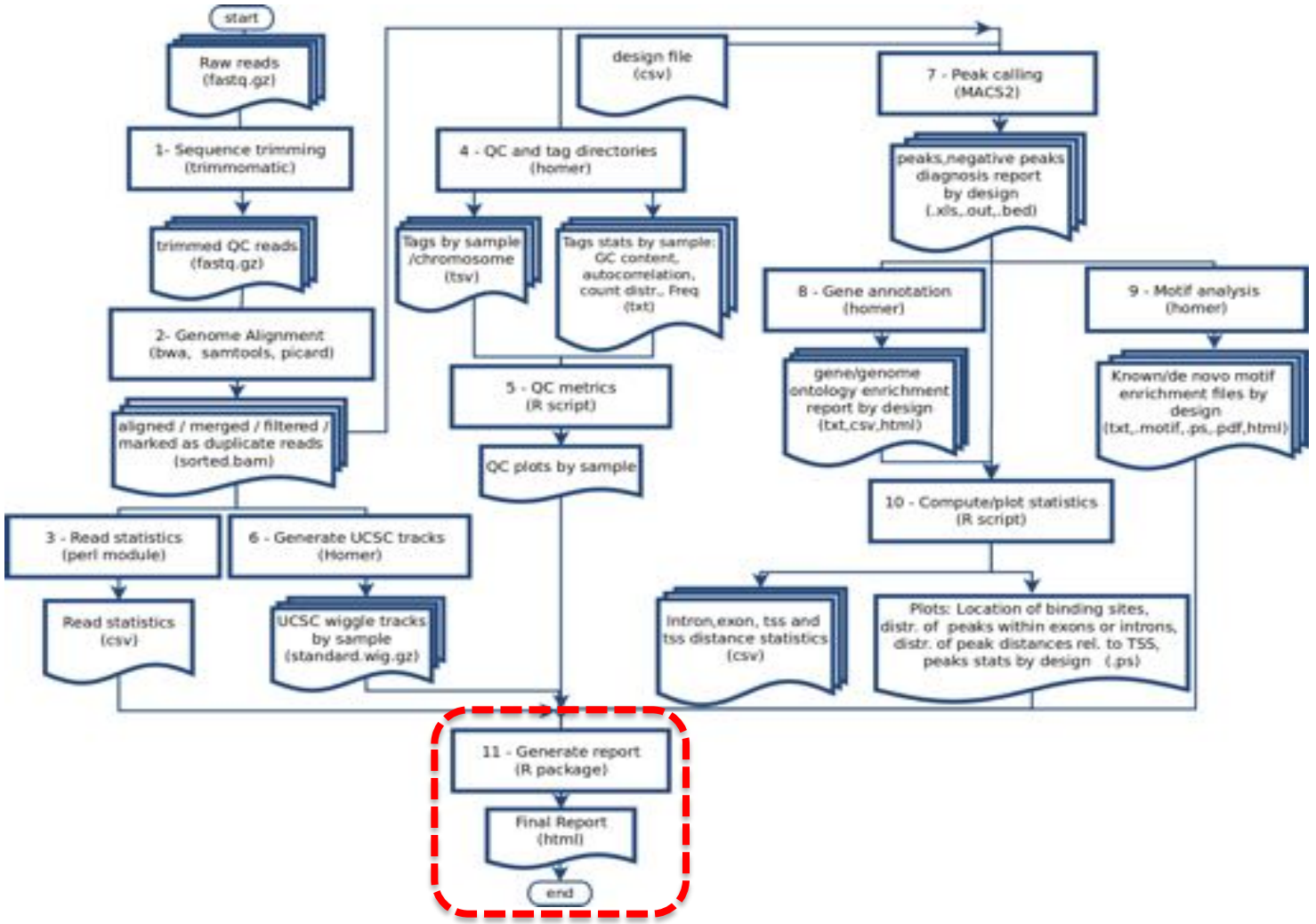
Plot the Following Statistics:

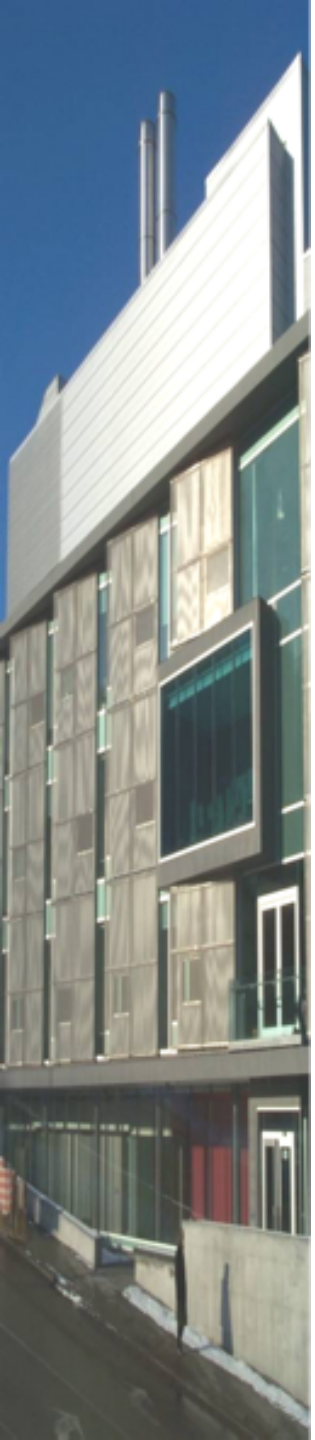
- –Location of binding sites
- –Distribution of peaks within introns
- –Distribution of peaks within exons
- –Distribution of peaks distances relative to TSS





ChIPseq: Generate report





Home-made Rscript

Generate report

- Noozle-based html report that contains description of the analysis as well as various QC summary statistics, main references of the software and methods used during the analysis and the list of processing parameters

Files generated:

- FinalReport.html, links to peaks, annotation, motifs, qcstats

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