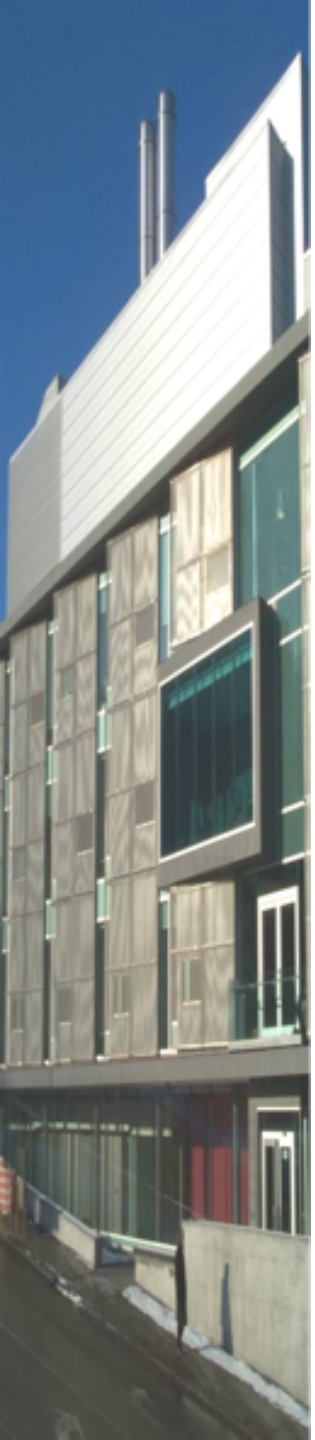
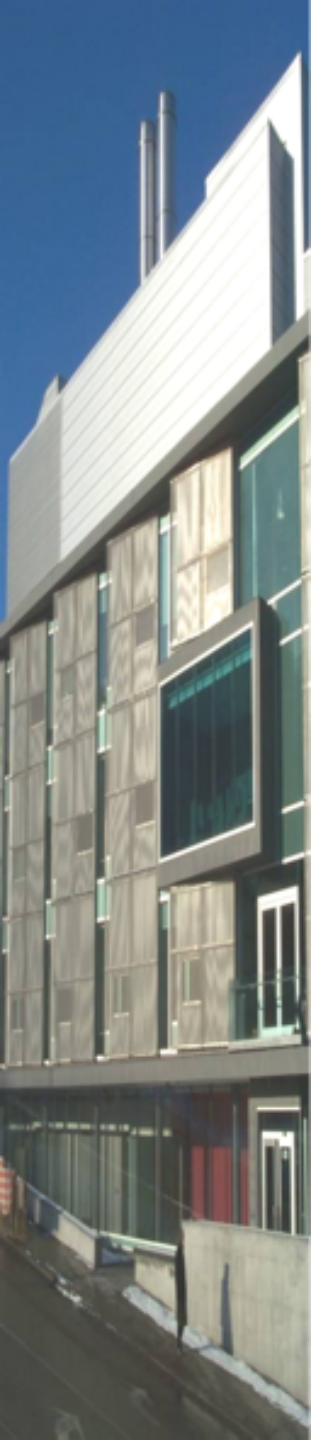


GenomeQuébec

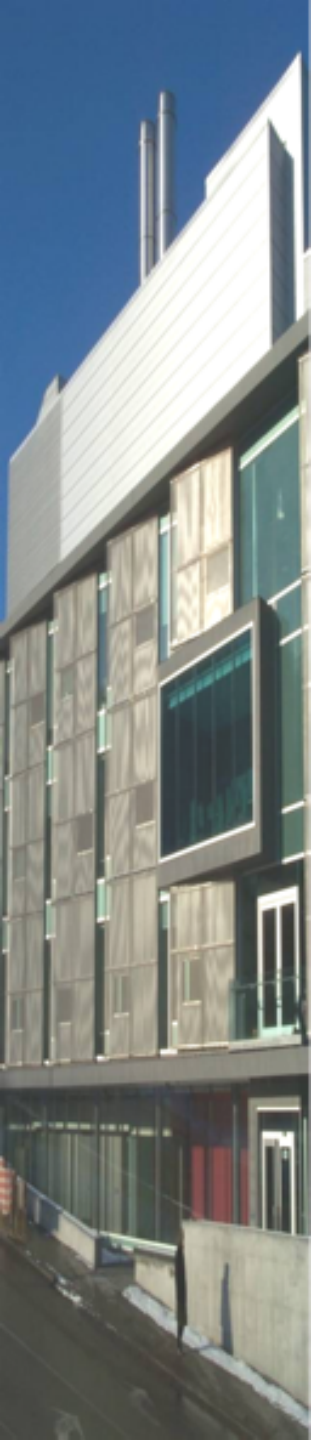
DNAseq analysis

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



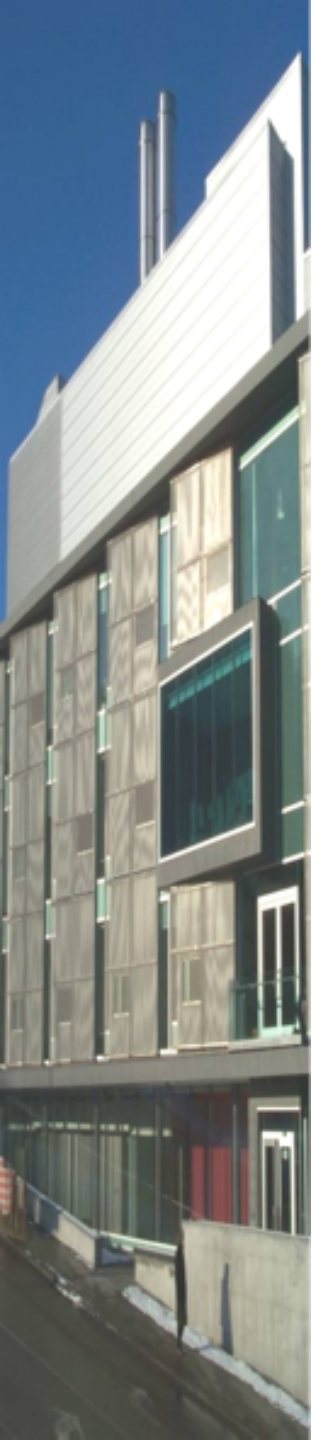
What is DNAseq ?

- DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule.
- The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.



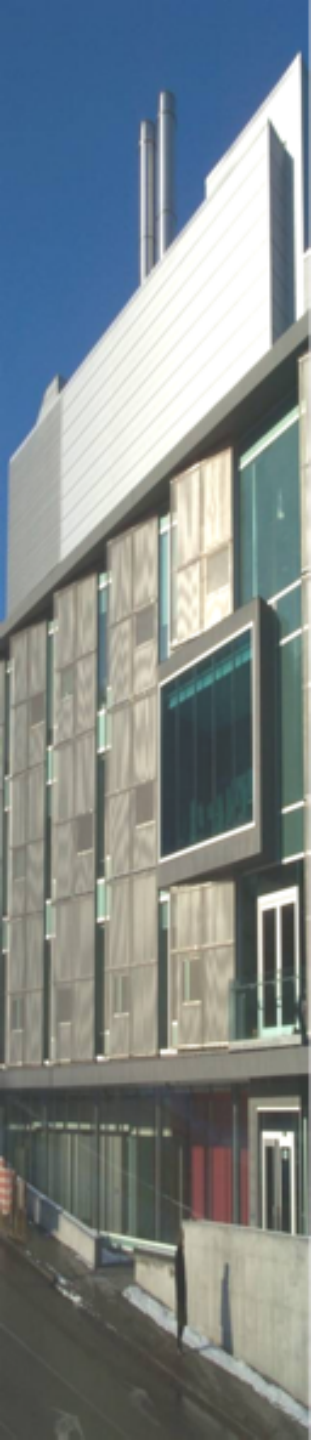
Why dnaSeq ?

- Whole genome sequencing:
  - Whole genome SNV detection
  - Structural variant
  - Capture the regulatory region information
  - Cancer analysis
  - De novo genome assembly
- Whole exome sequencing:
  - Cheaper
  - Capture the coding region information
  - Rare diseases analysis

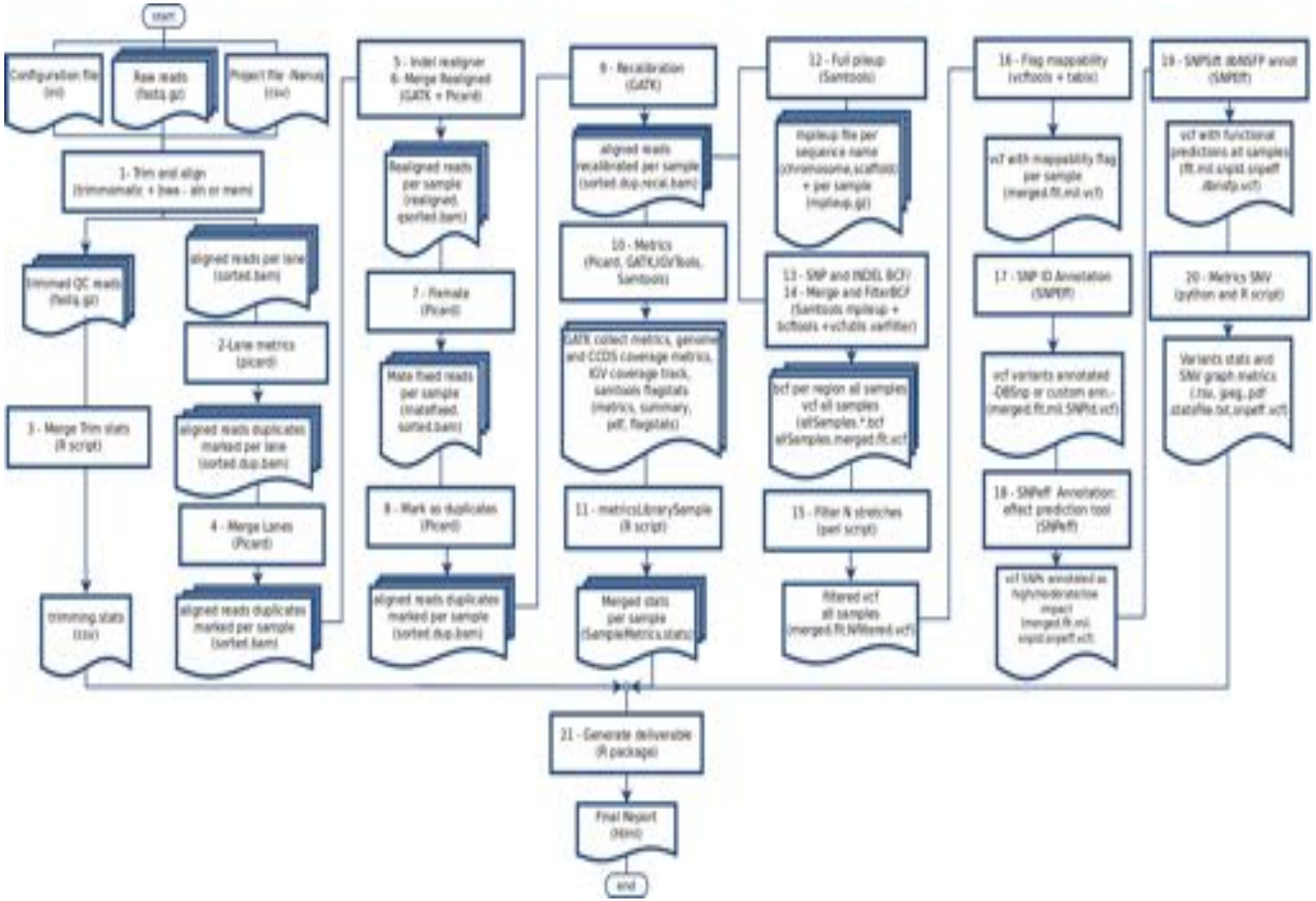


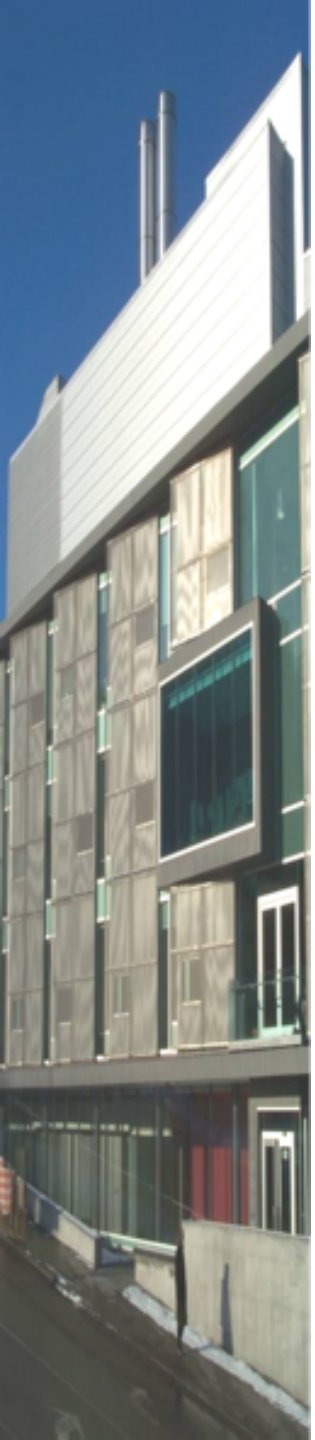
What the DNAseq problem is about ?

- Strings of 100 to ≈1kb letters
- Puzzle of 3,000,000,000 letters
- Usually have 120,000,000,000 letters you need to fit
- Many pieces don't fit :
  - sequencing error/SNP/Structural variant
- Many pieces fit in many places:
  - Low complexity region/microsatellite/repeat

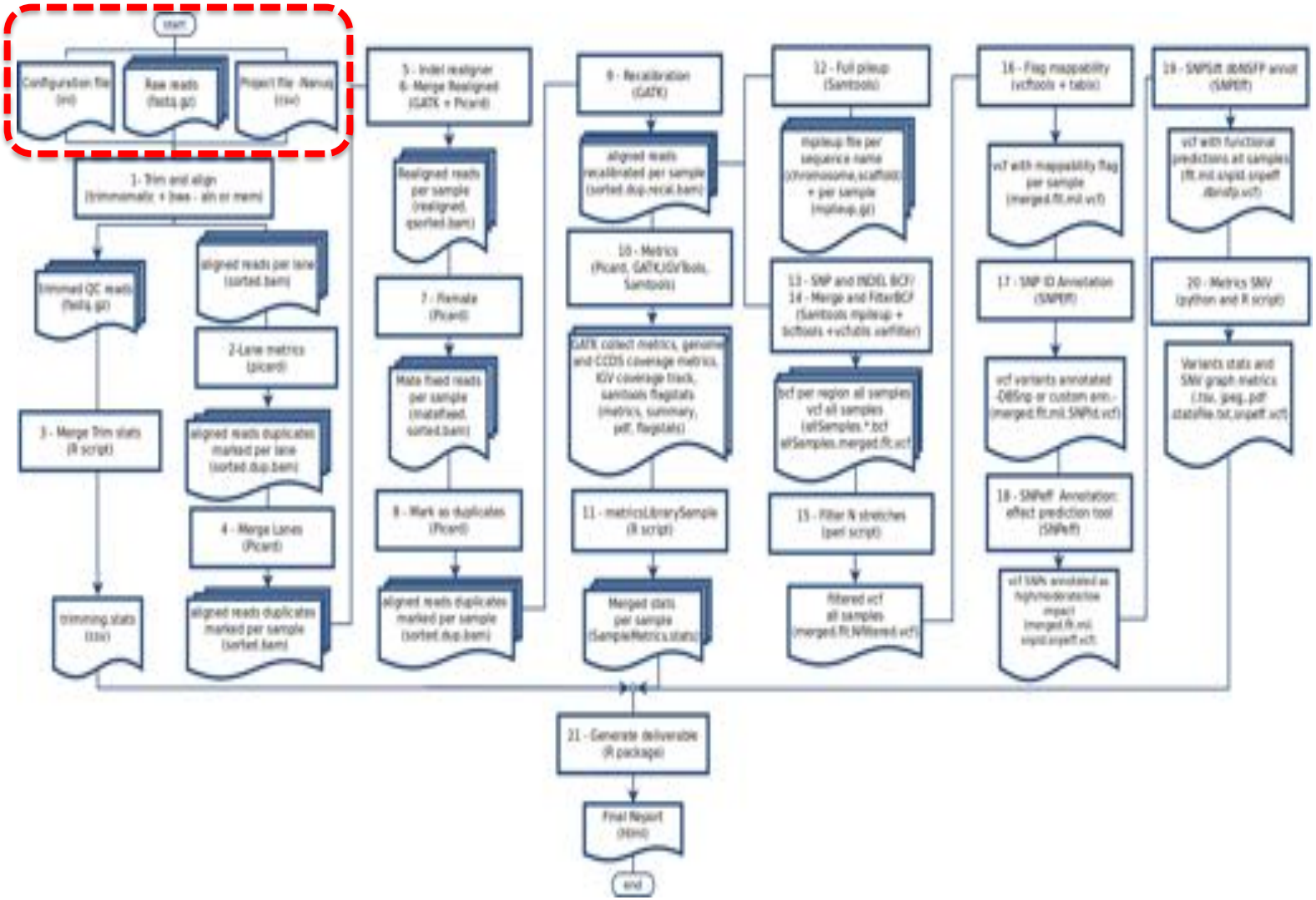


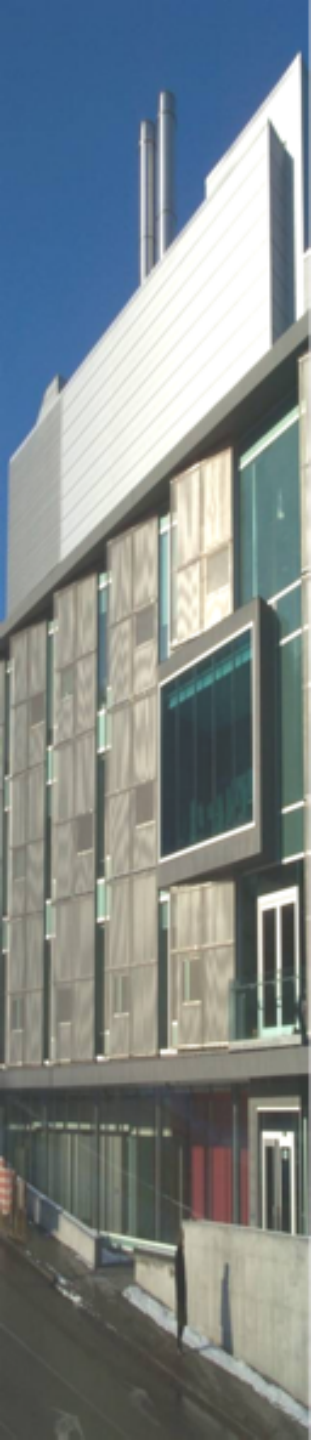
DNAseq overview





DNAseq: Input Data





Input Data: FASTQ

End 1

Sample1_R1.fastq.gz

Sample2_R1.fastq.gz

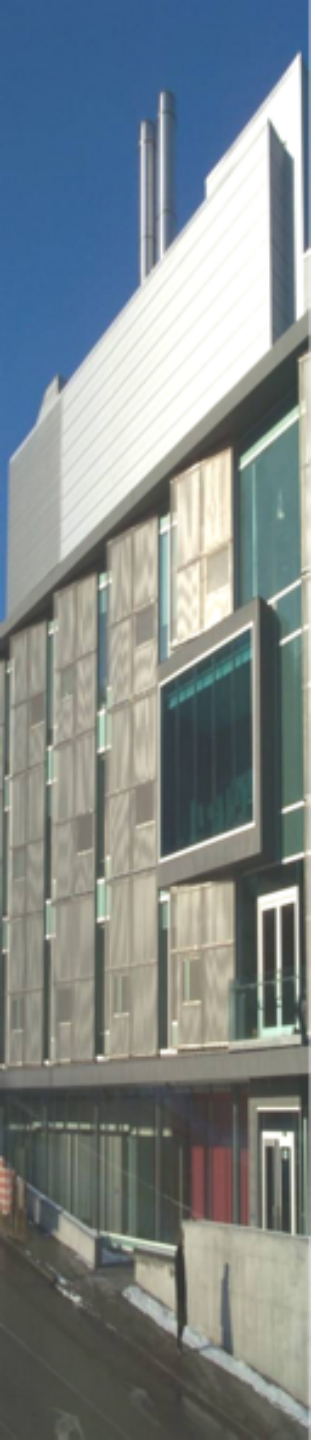
End 2

Sample1_R2.fastq.gz

Sample_R2.fastq.gz

Each sample will generate between 5Gb (100x WES) to 300Gb (100x WGS) of data

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

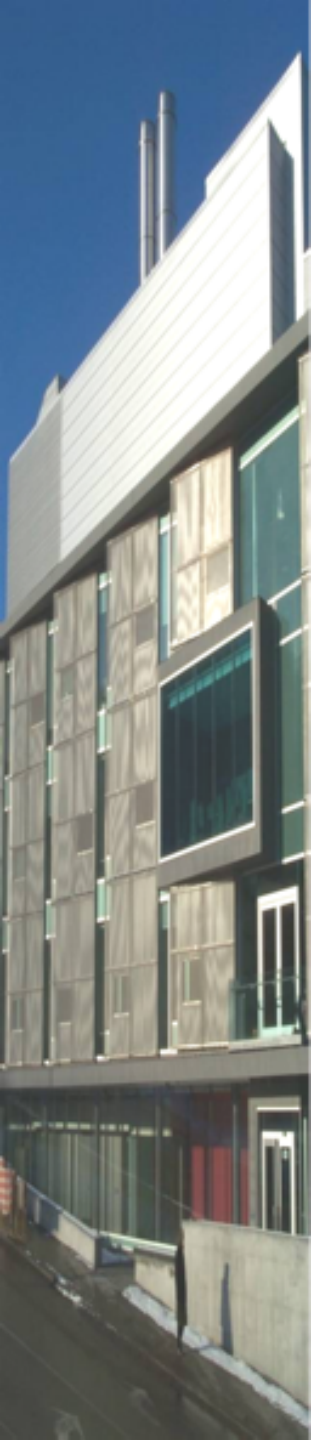


$$Q = -10 \log\_10 (p)$$

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 %



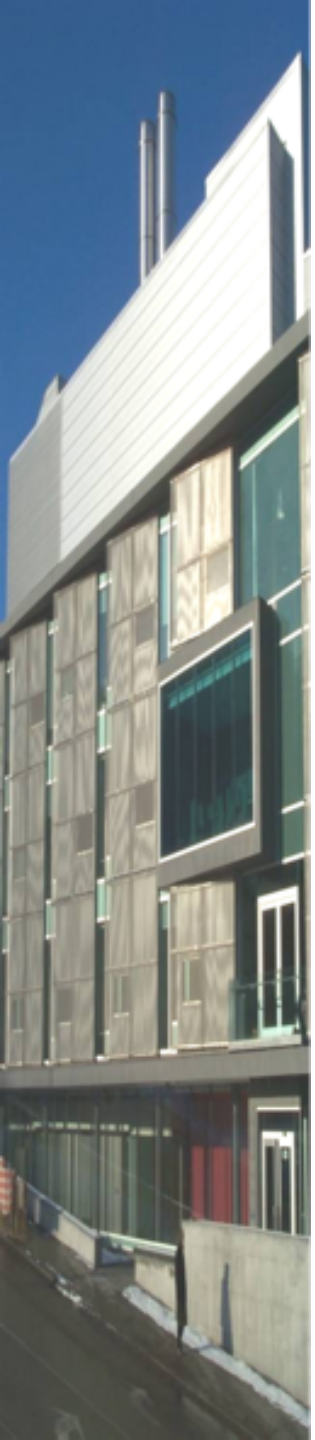
QC of raw sequences

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

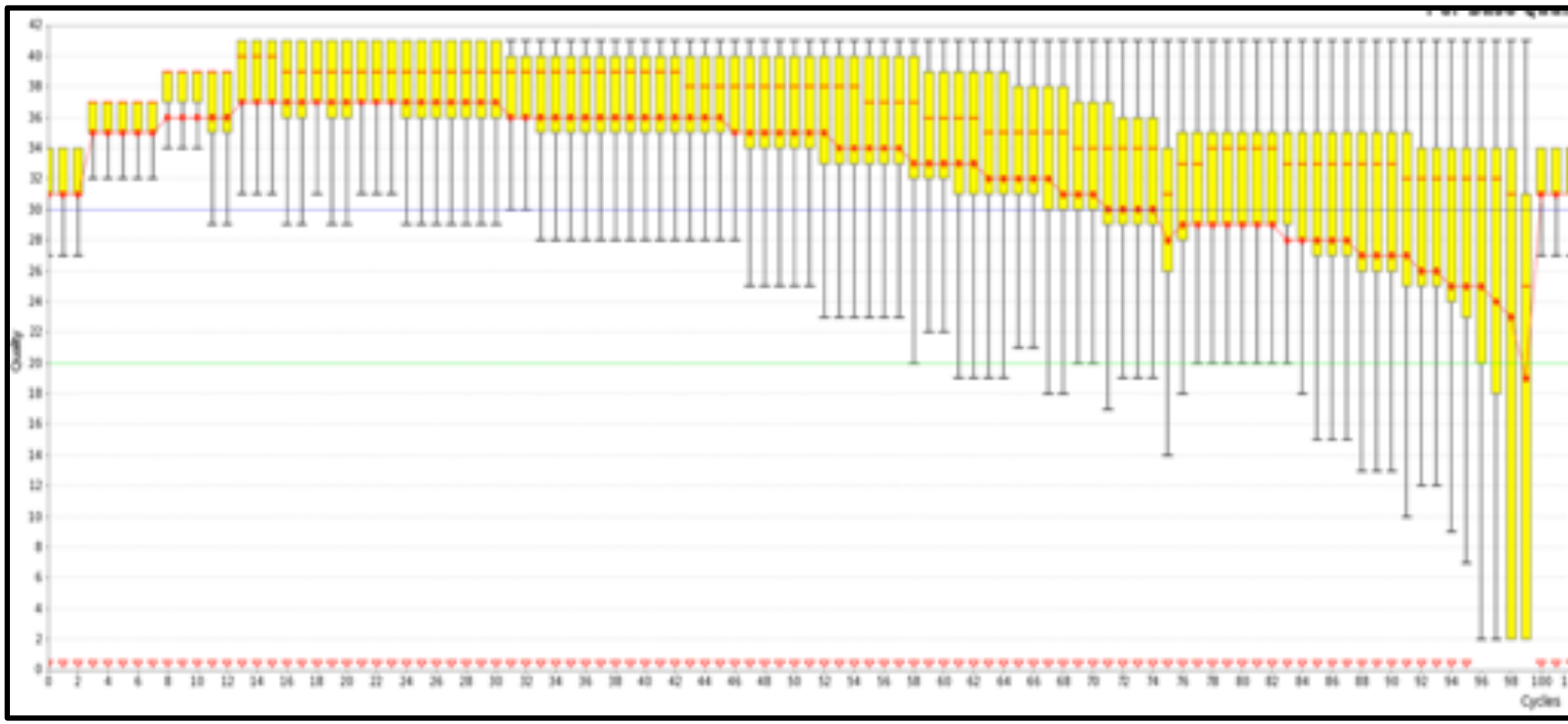
CSV | View/Set Filter | Download Read Files | Help with icons

Read Sets (64 elements) — Add/Remove Column

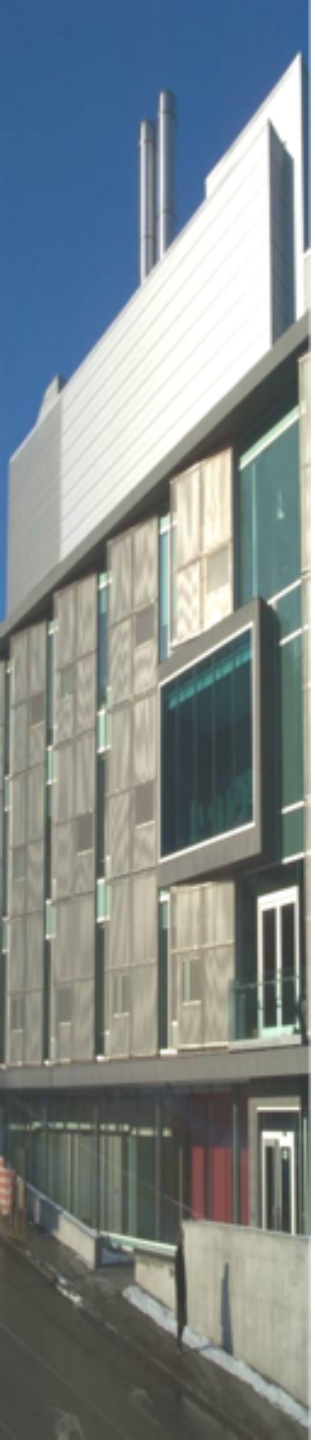
Name	Multiplex Key	Run	Region	QC	Status	Number of reads	Number of Bases	Average Quality	% Duplicate	% Passed Filter	Reads Fastq R1	Reads Fastq R2
W24P	Index_7	1177	4	QC		45,373,280	9,074,656,000	33	21.674	100	(4562MB)	(4546MB)
W25P	Index_8	1177	4	QC		45,066,800	9,013,360,000	33	17.943	100	(4527MB)	(4513MB)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

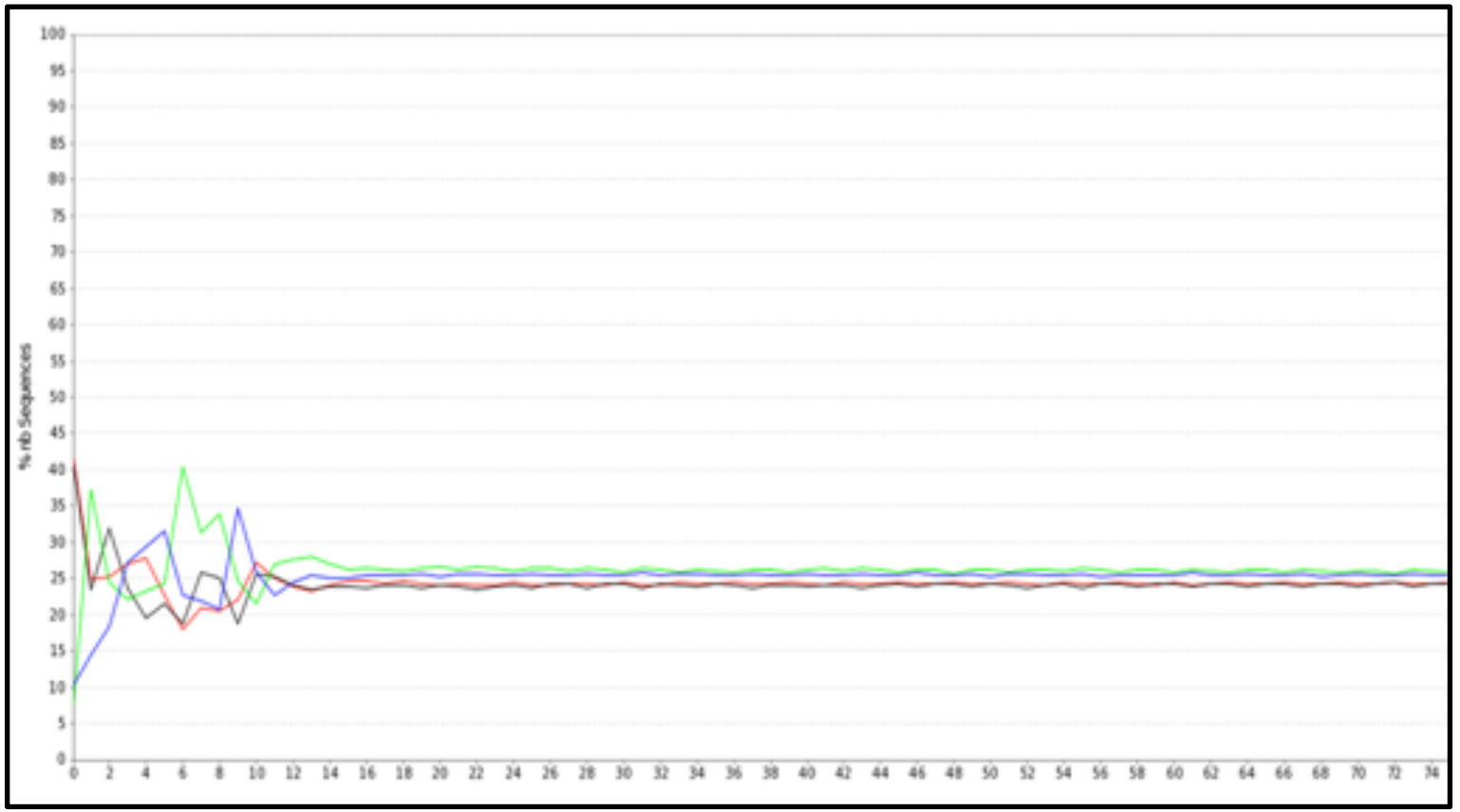


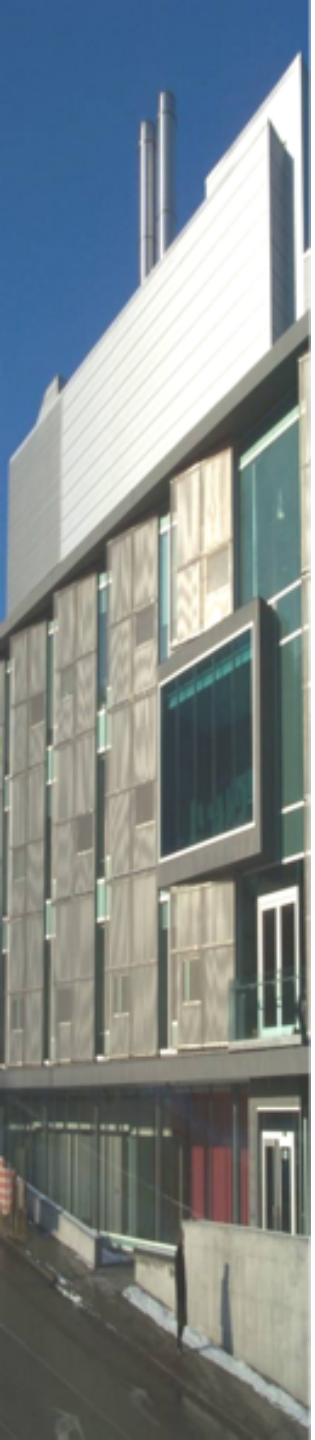
low qualtity bases can bias subsequent anlaysis
(i.e, SNP and SV calling, …)



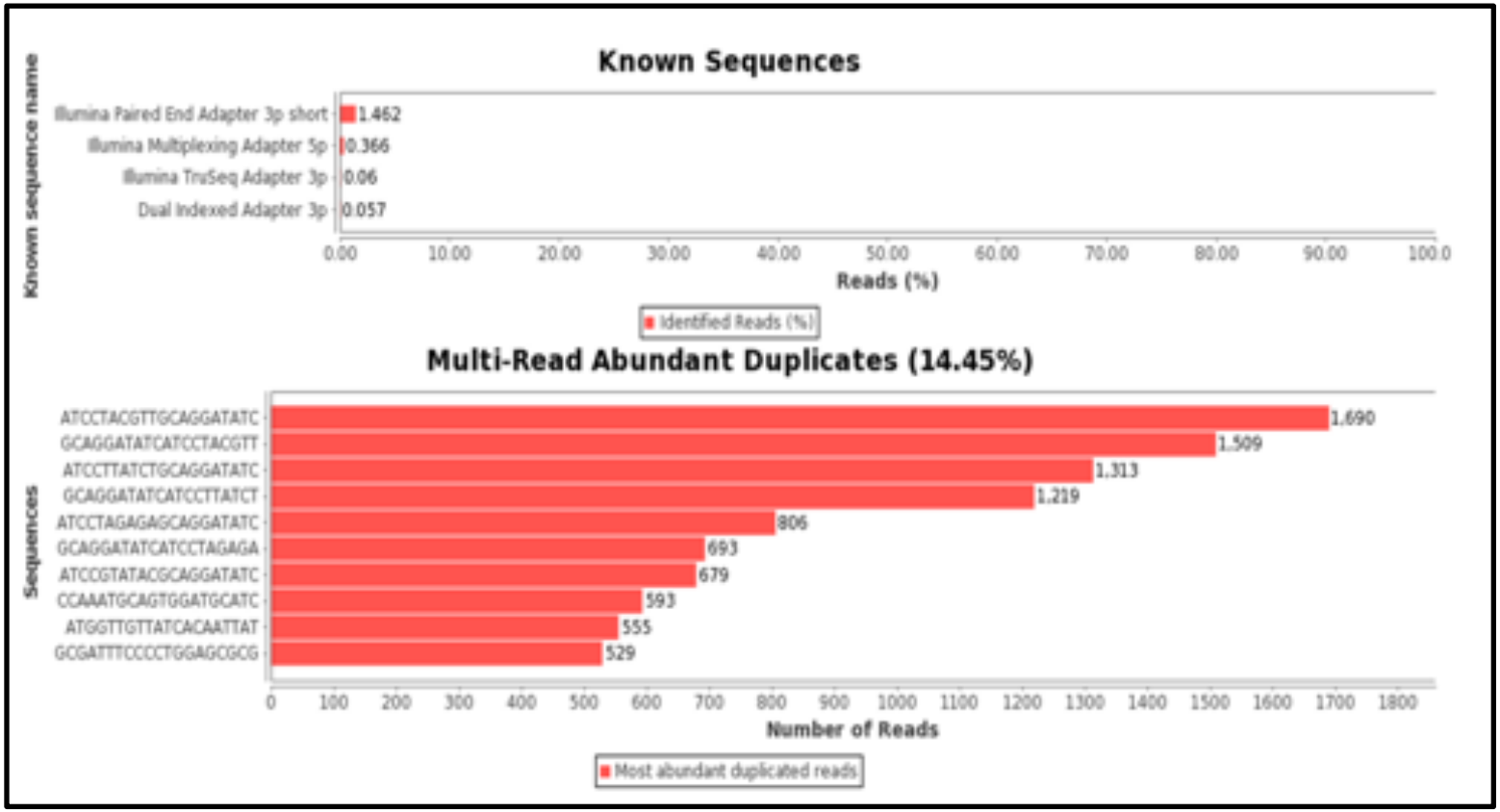
QC of raw sequences

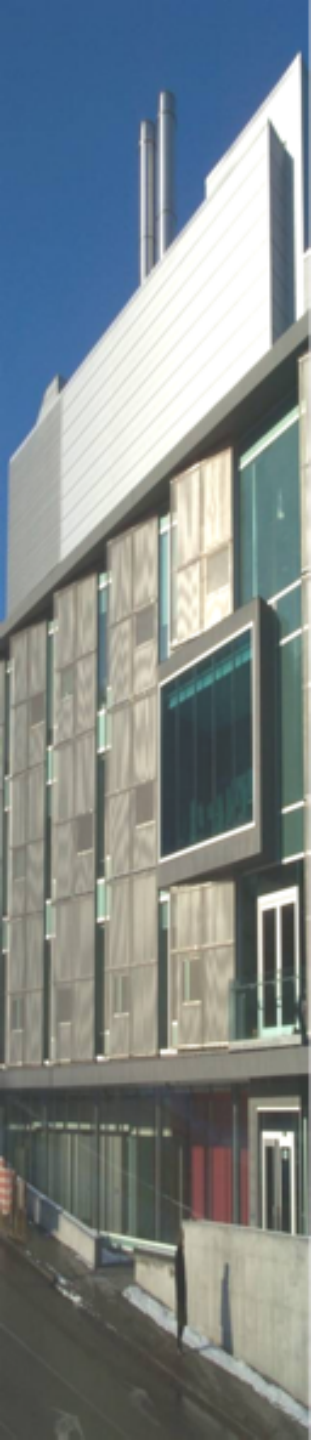
Positional Base-Content





QC of raw sequences



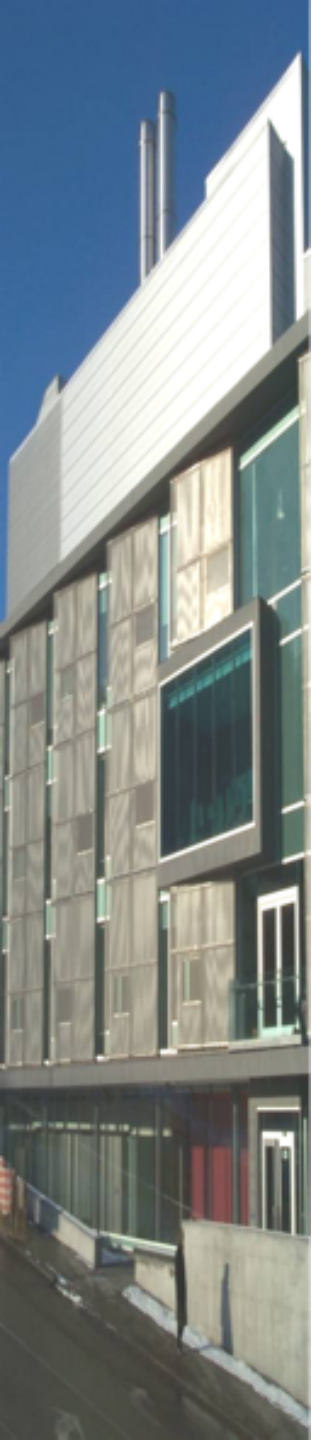


QC of raw sequences

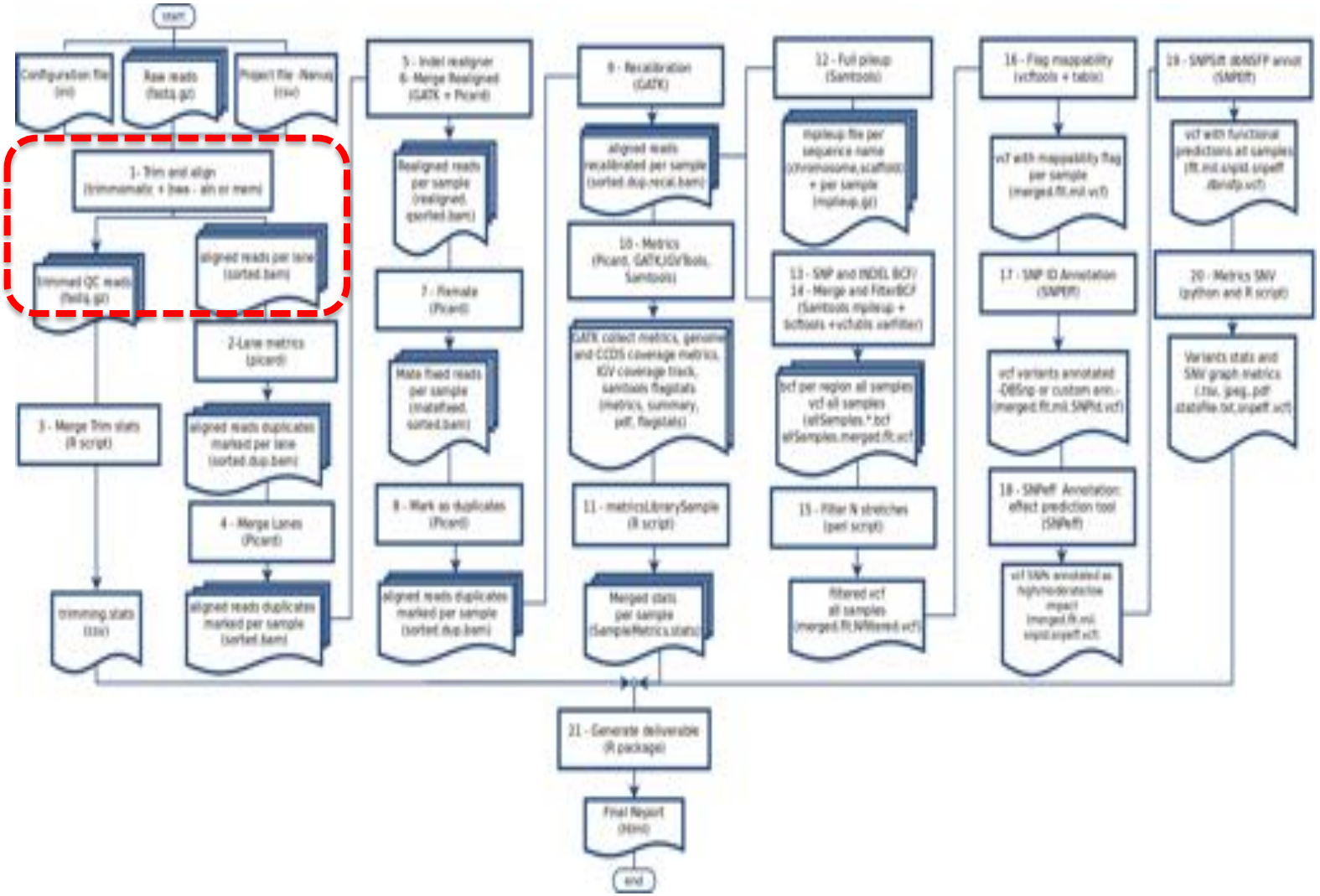
Species composition (via BLAST)

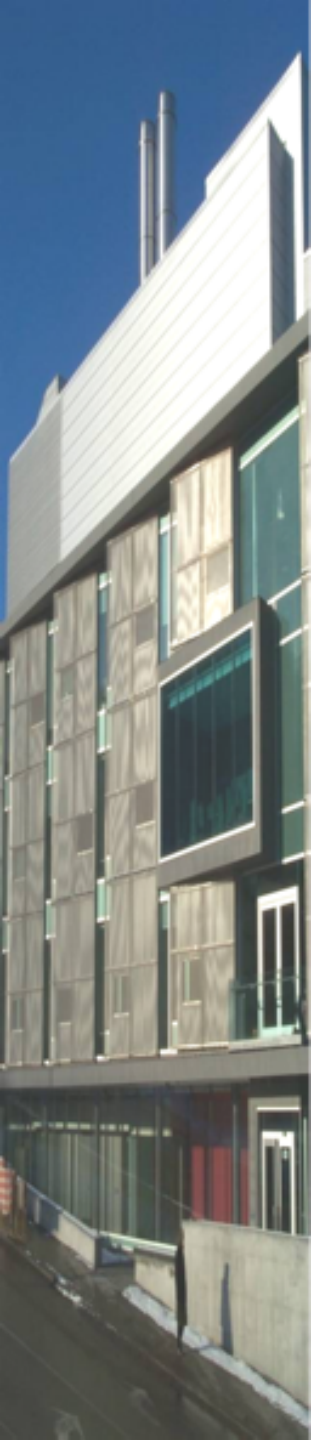
Blast Results (20 elements)

	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



DNA-Seq: Trimming and aligning

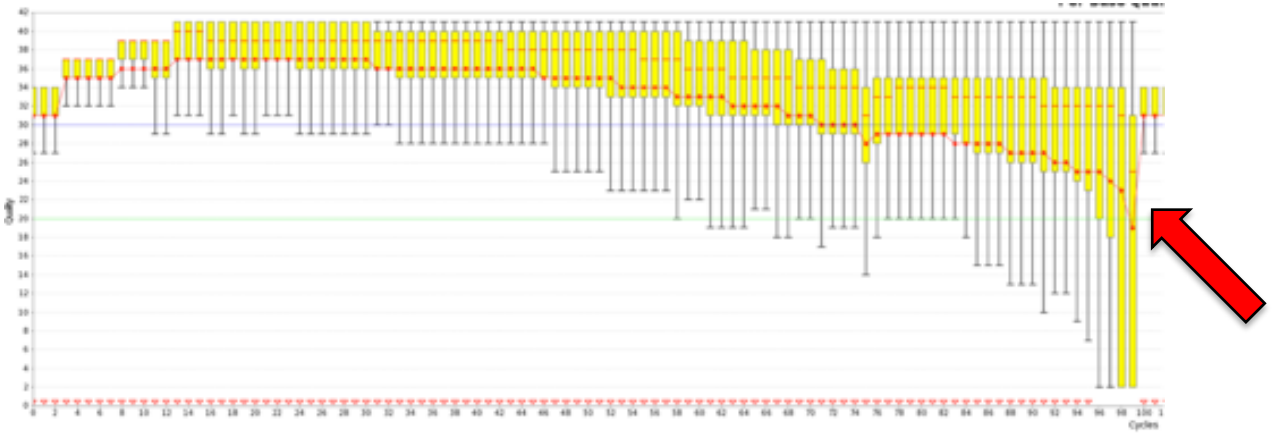




Read Filtering

- Clip Illumina **adapters**:

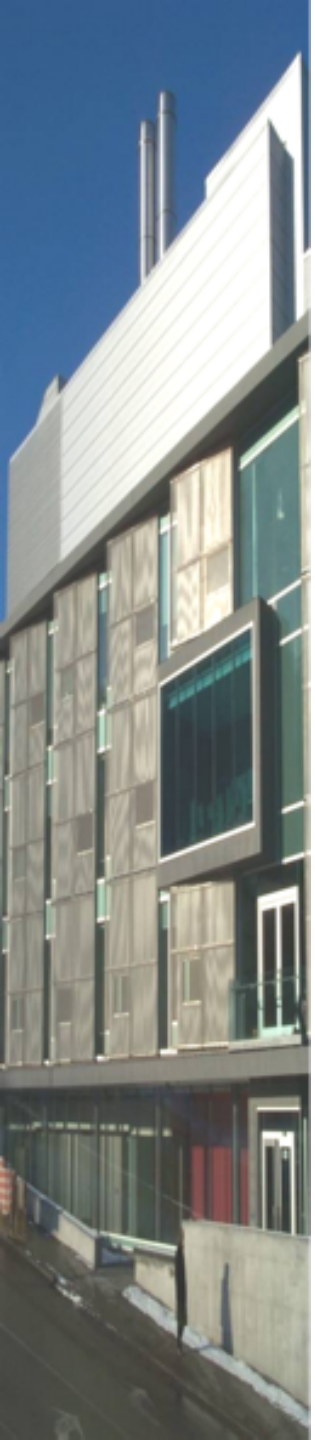
- Trim trailing **quality** < 30



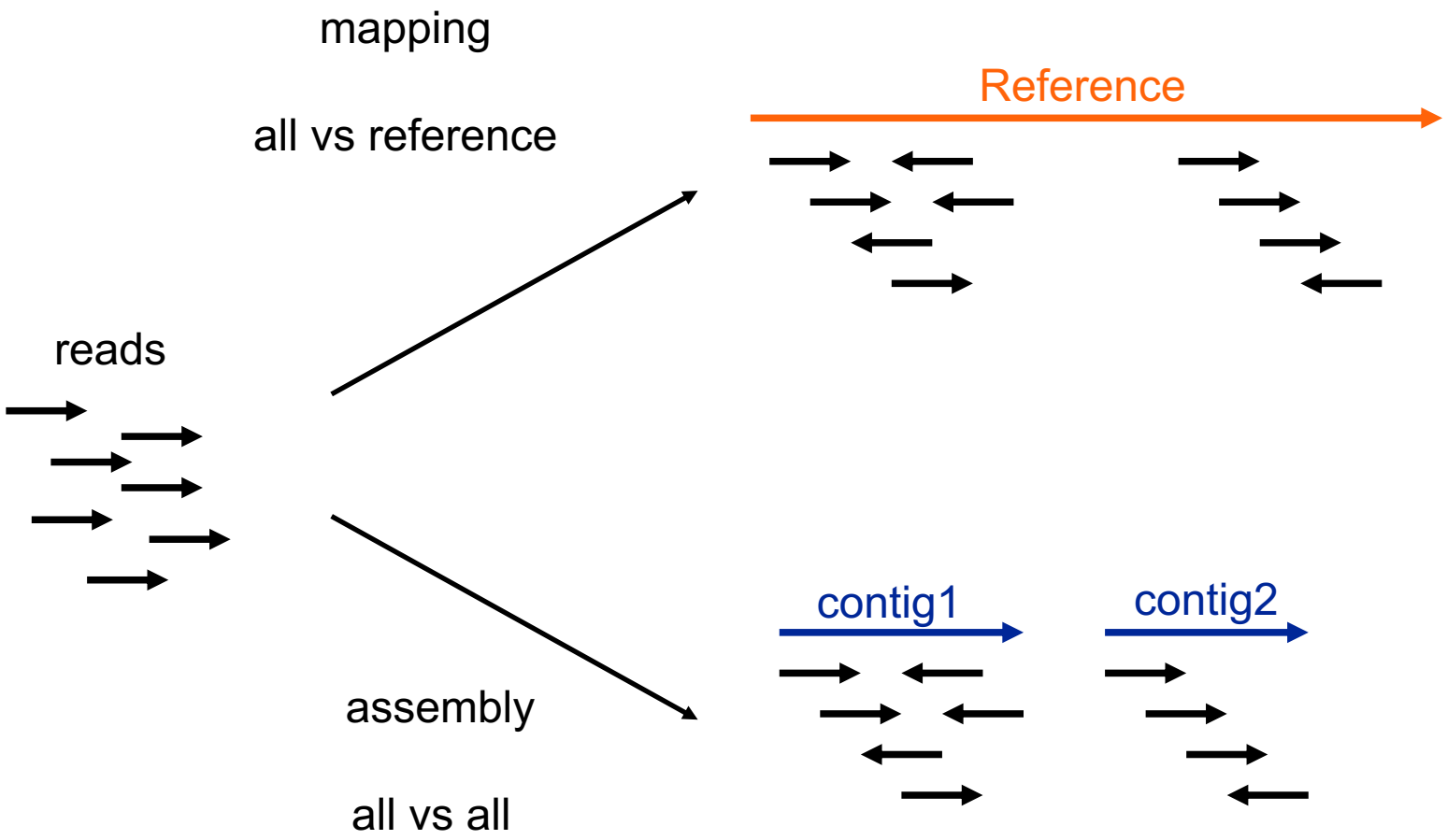
- Filter for read **length** ≥ 32 bp

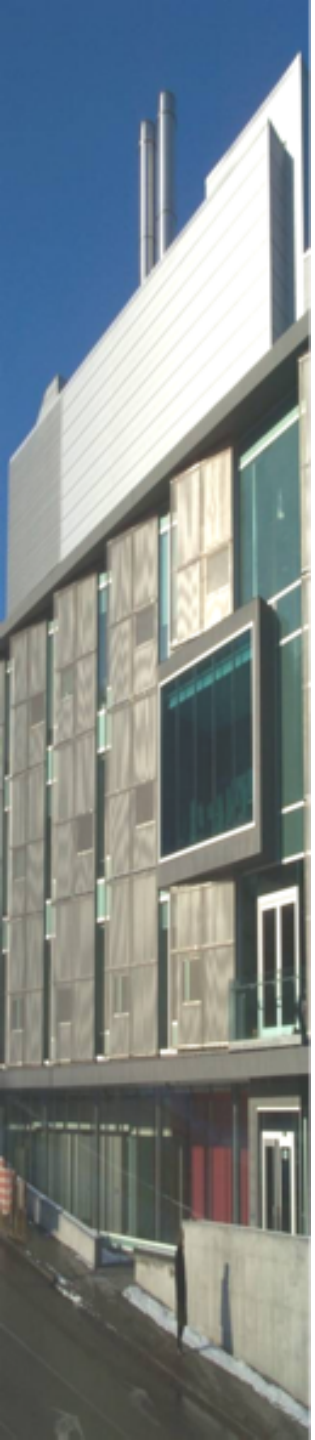
Trimmomatic

usadellab.org

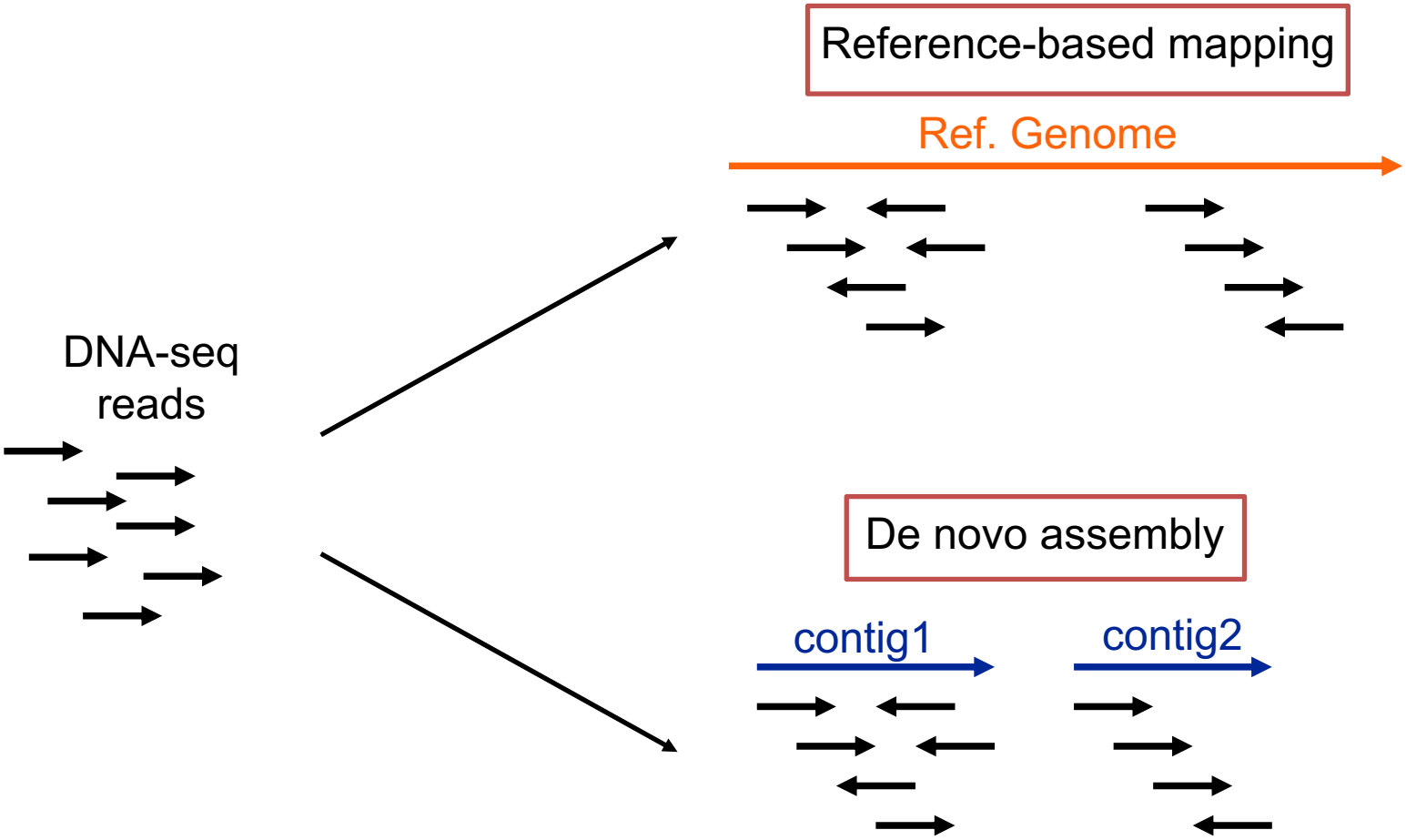


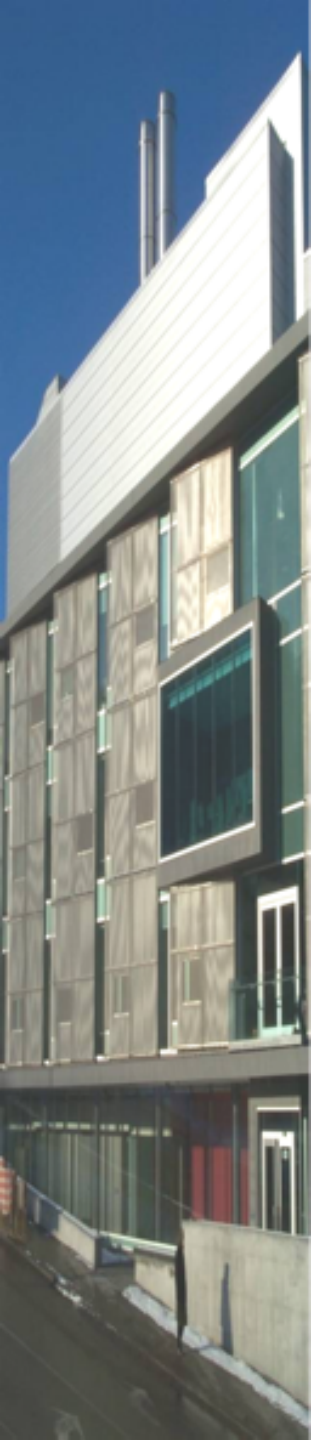
Assembly vs. Mapping





RNA-seq: Assembly vs 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

Sample1.bam

Sample2.bam

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

between 10Gb ot 500Gb each bam

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

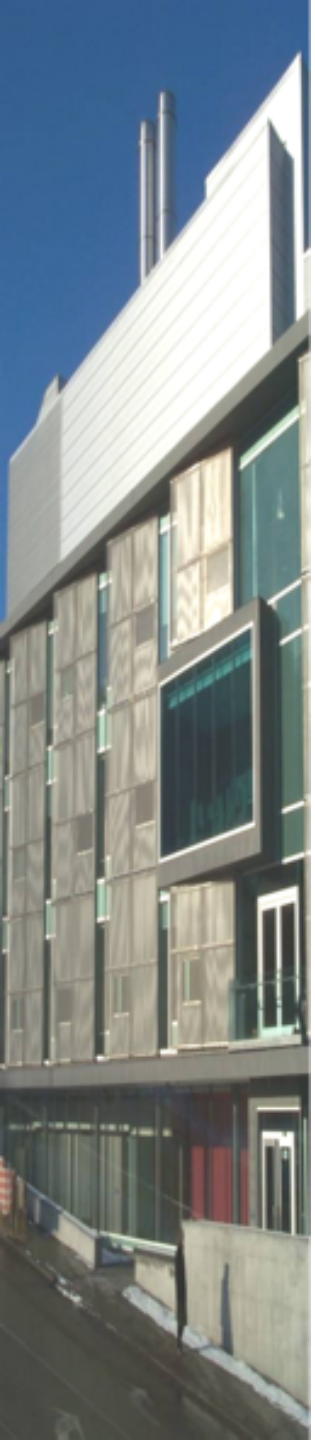
Read name | Flag | Reference Position | CIGAR | Mate Position

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

Bases

Base Qualities

SAM: **S**equence **A**lignment/**M**ap format



The BAM/SAM format

SAMtools

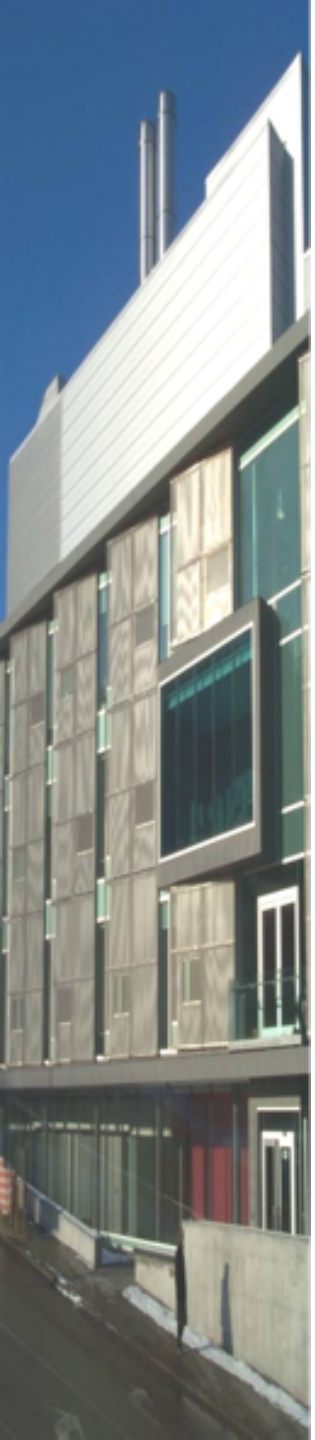
samtools.sourceforge.net

Picard

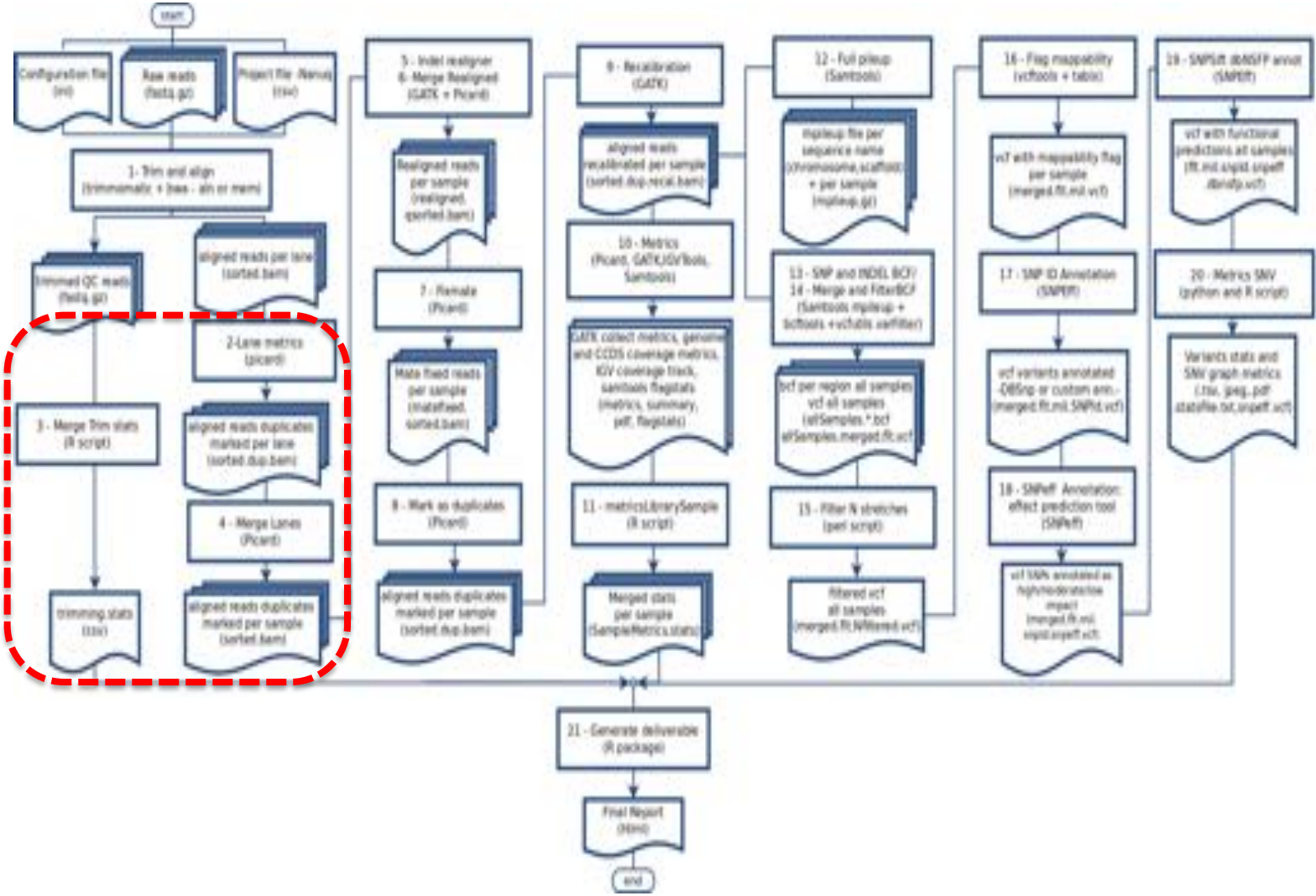
picard.sourceforge.net

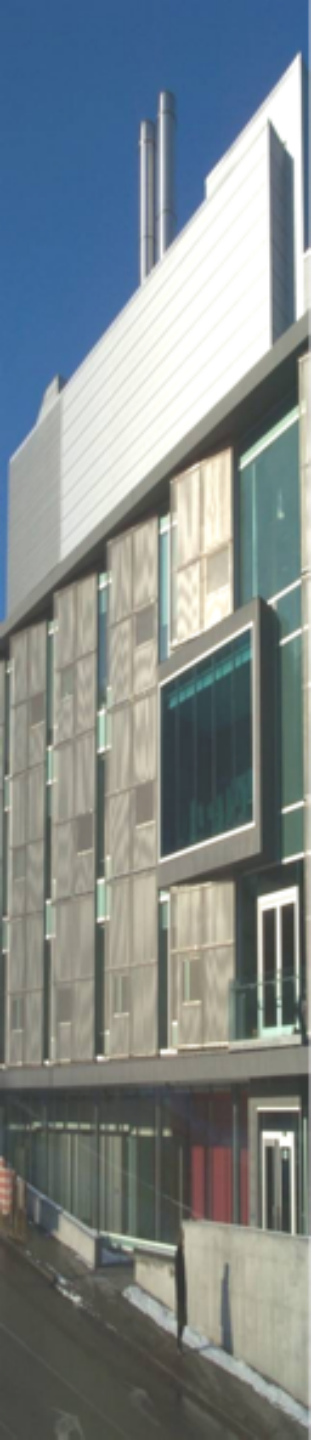
Sort, View, Index, Statistics, Etc.

```
$ 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)
$
```



DNA-Seq: metrics





Included metrics

- Metrics are collected form the output of Trimmomatic, Samtools and Picard softwares

Step 4: By sample sequence and alignment metrics

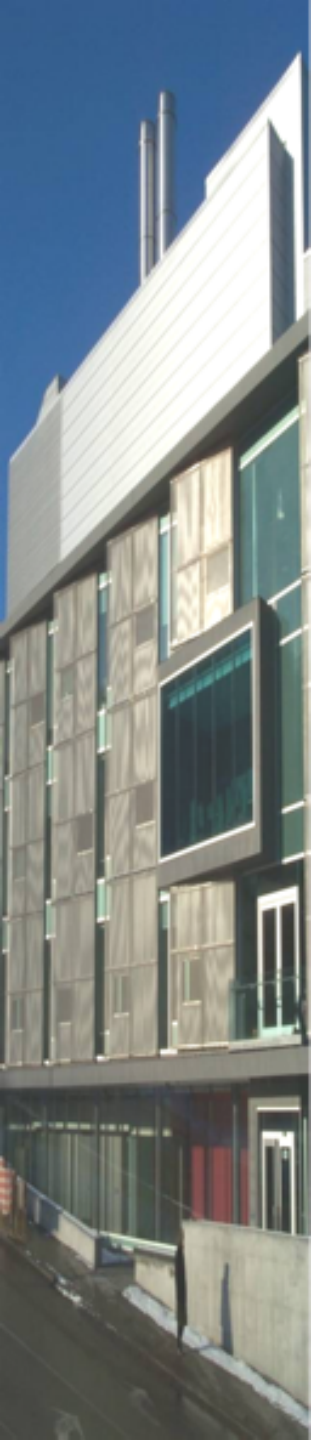
General summary statistics are provided for each sample. Sample lanes have been merge together for clarity.

GET FULL TABLE

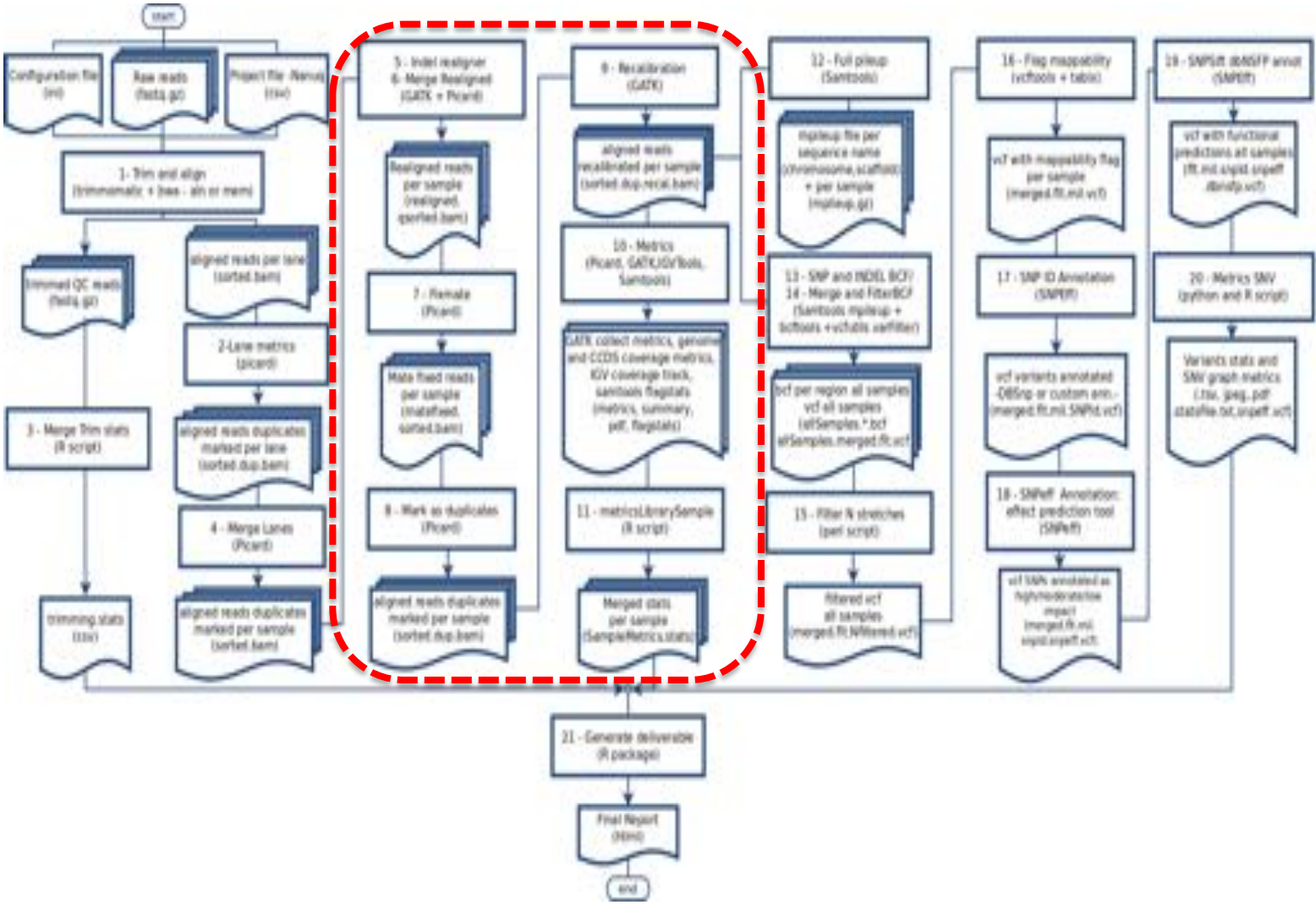
Table 2. By sample sequencing and alignment statistics

Sample	Raw reads	Surviving reads	%	Mapped reads	%	Not Duplicate	Duplicate	%.1	Pair Orientation
NA12383	4095378	4092086	100	3949621	97	3922873	26748	0.65	FR

- Raw reads: the total number of read obtained from the sequencer
- Surviving reads: the number of remaining reads after the trimming step
- %: Surviving reads / Raw reads
- Mapped reads: the number of read aligned
- %: Mapped reads / Surviving reads
- Not Duplicate: the number of not duplicated read entries
- Duplicate: the number of duplicate read entries providing alternative coordinates
- %: Duplicate / Mapped reads
- Pair Orientation: the library pared-end read design
- Mean Insert Size: the mean distance between the left most base position of the read1 and the right most base position of the read 2
- Standard Deviation: the standard deviation of distance between the left most base position of the read1 and the right most base position of the read 2
- WG Mean Coverage: total number of aligned read / size of the genome
- CCDS Mean Coverage: total number of aligned read in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_10: total number of bases with a coverage >= 10x in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_25: total number of bases with a coverage >= 25x in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_50: total number of bases with a coverage >= 50x in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_75: total number of bases with a coverage >= 75x in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_100: total number of bases with a coverage >= 100x in the CCDS/capture region / size of the CCDS/capture region
- CCDS %_bases_above_500: total number of bases with a coverage >= 500x in the CCDS/capture region / size of the CCDS/capture region



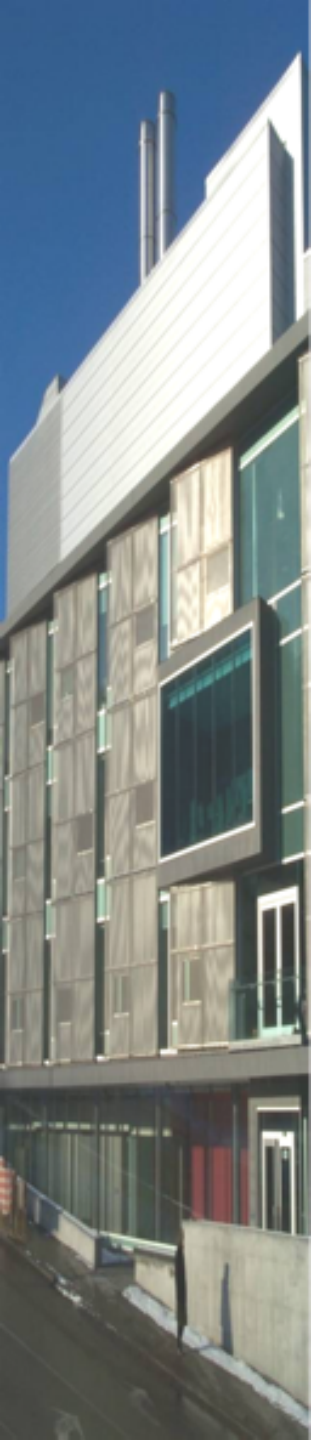
DNA-Seq: Alignment refinement



Local indel realignment

- Primary alignment with *BWA* bio-bwa.sourceforge.net
- Local re-alignment around indels with *GATK*
- Possible mate inconsistency are fixed using *Fixmate*

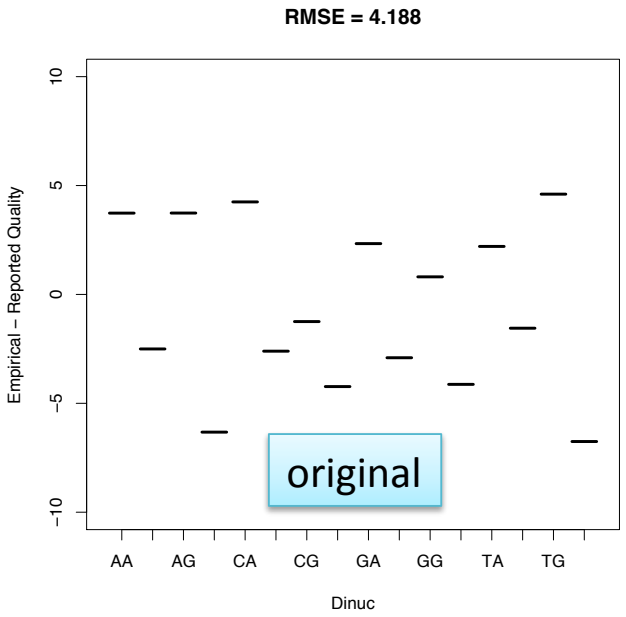
image: broadinstitute.org/gatk



Duplicates and recalibration

- Mark duplicates with *Picard*
- Base Quality Score Recalibration *GATK*

Example Bias in the qualities reported depending of the nucleotide context



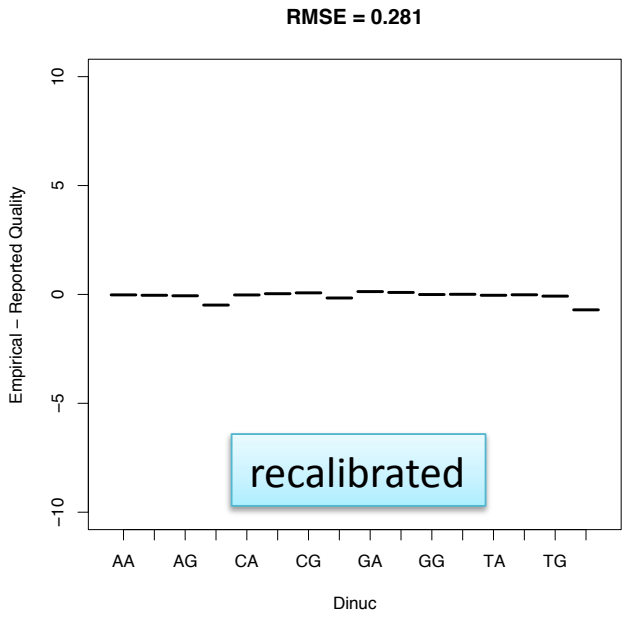
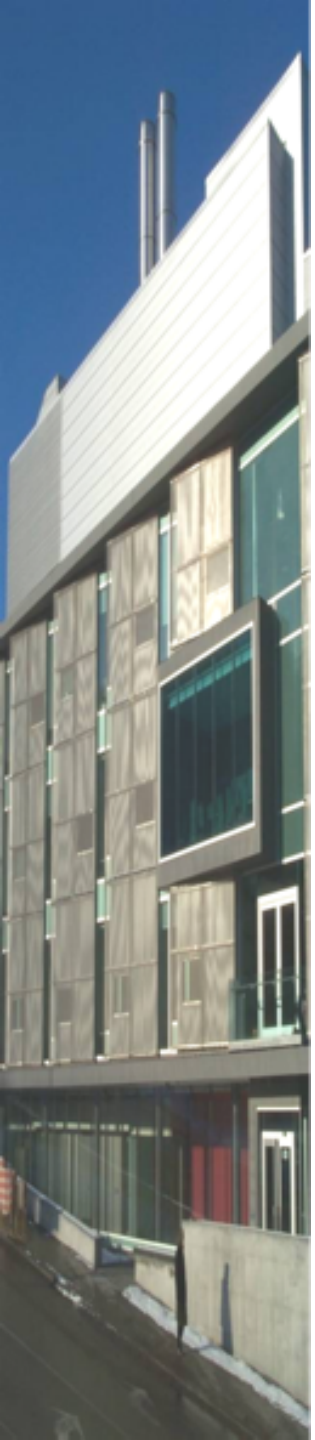
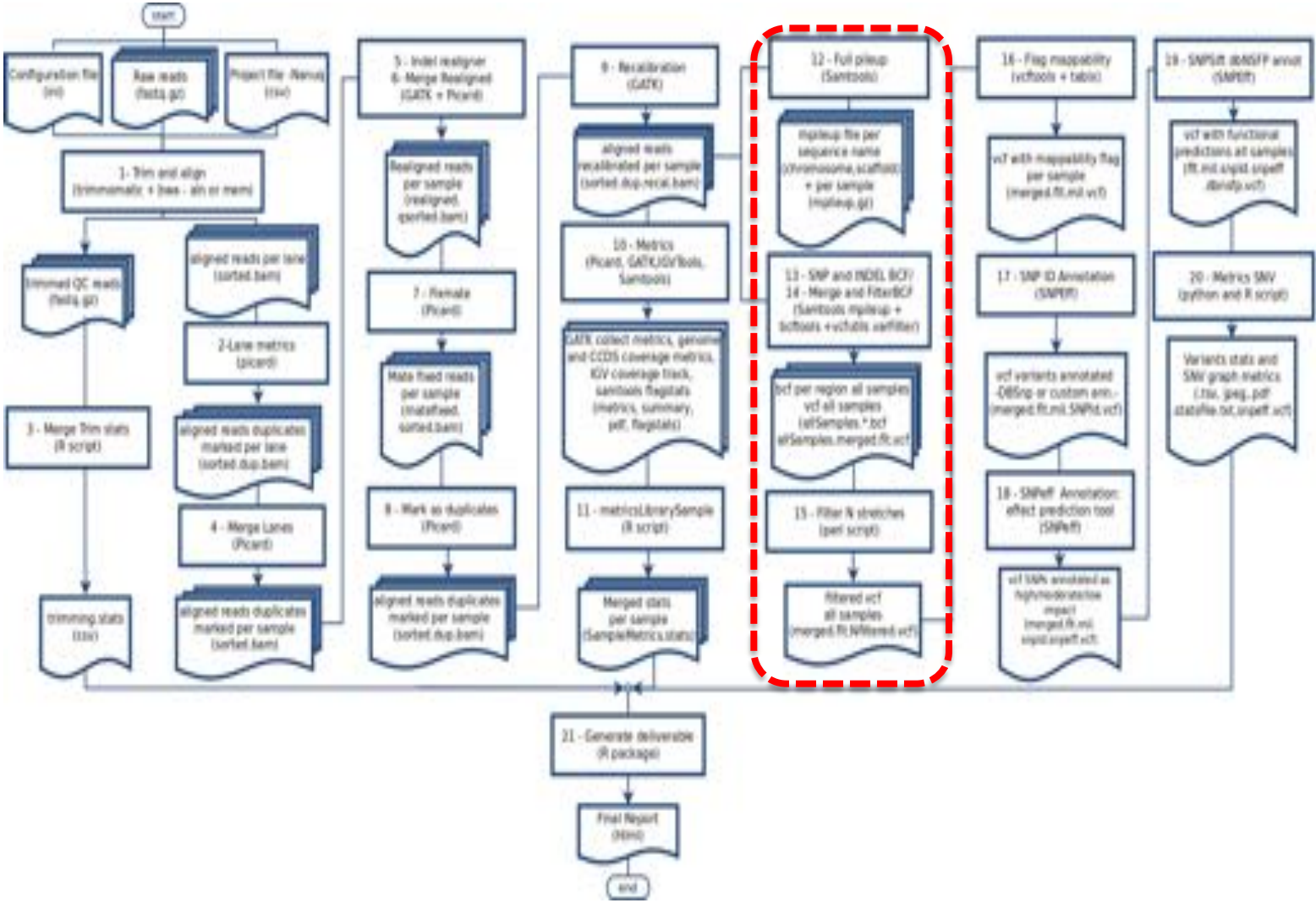
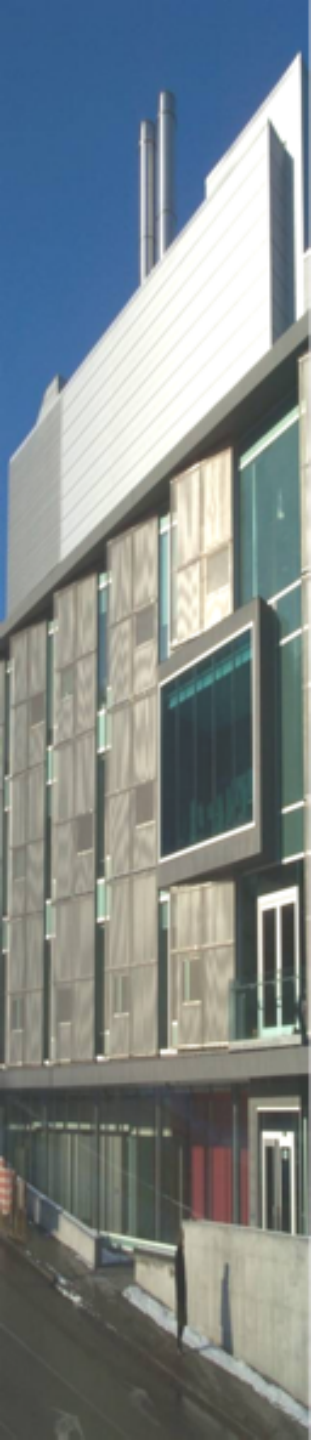


image: broadinstitute.org/gatk



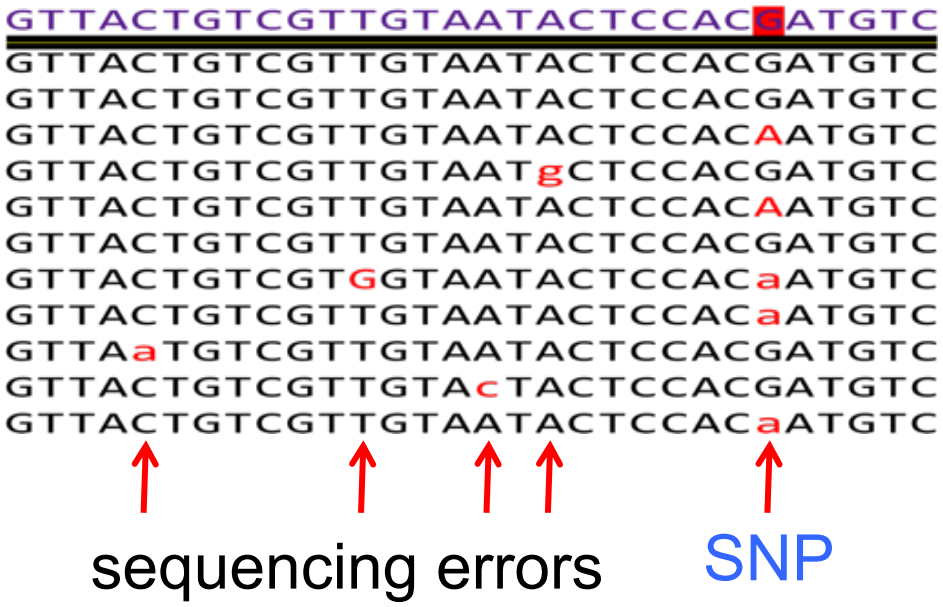
DNA-Seq: SNV calling



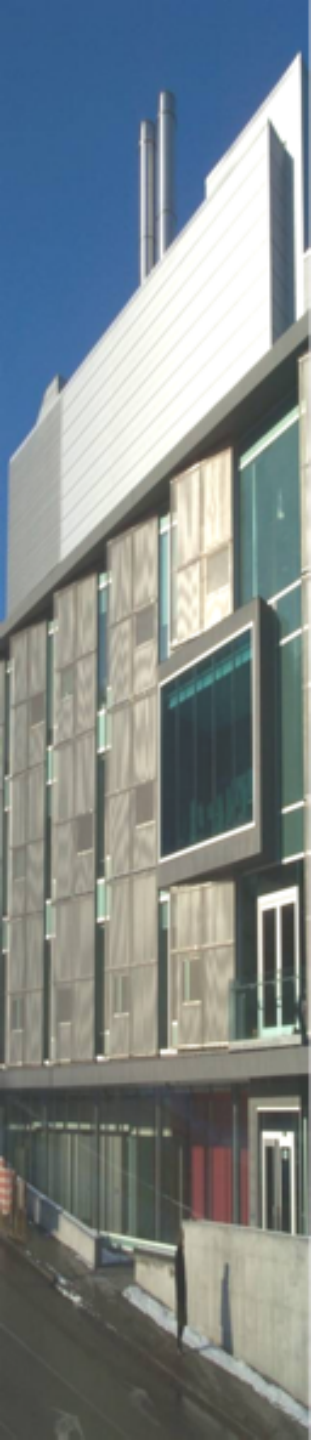


Single Nucleotide Variant calling

- Aim: differentiate real SNPs from sequencing errors

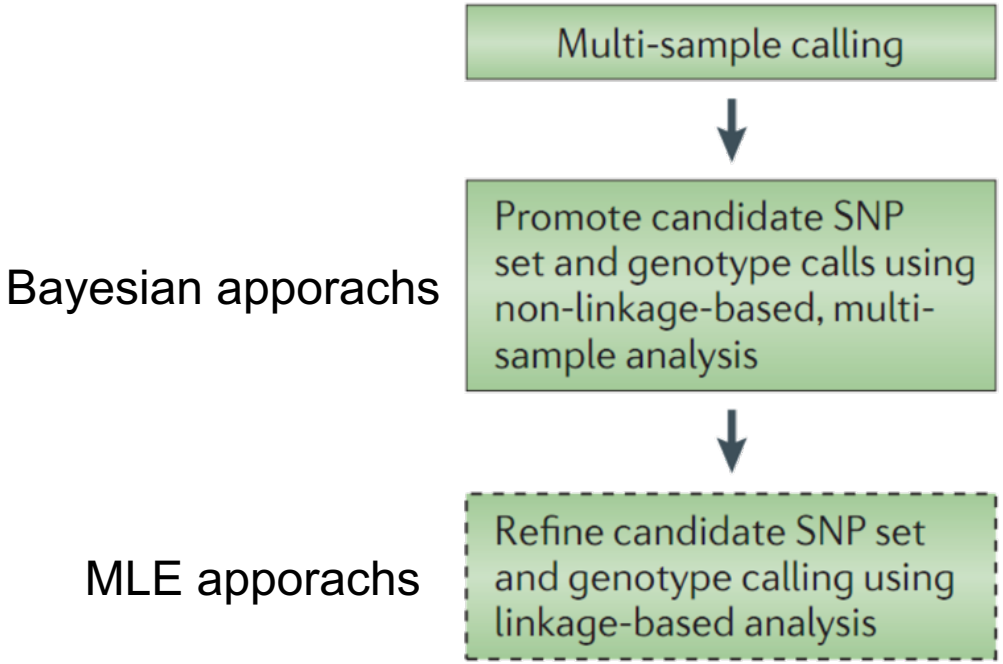


- An accurate SNP discovery is closely linked with a good base quality and a sufficient depth of coverage



SNP and genotype calling workflow

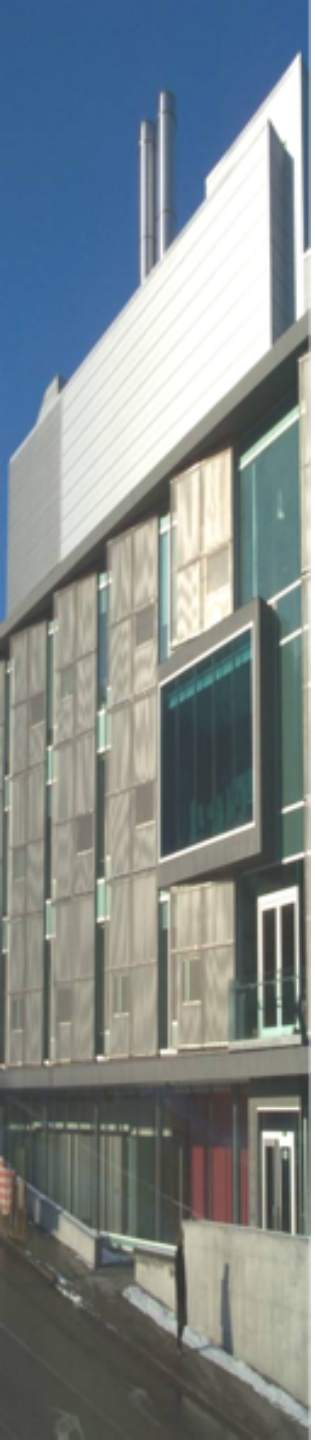
Variants from multiple samples are called simultaneously using the mpileUp method from samtools and quality filtered using bcftools



Bayesian apporachs

MLE apporachs

Nielsen et al June 2011



The variant format : vcf

- **V**ariant **C**all **F**ormat

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	D000FYW	D000G08
9	130216120	rs2244331	G	T	129	.	<...>	GT:PL:DP:SP:GQ	1/1:90,12,0:4:0:13	0/1:76,0,69:8:0:68
9	130216951	rs2244218	G	T	999	.	<...>	GT:PL:DP:SP:GQ	0/1:219,0,151:24:0:99	0/1:191,0,255:57:1:99
9	130217050	rs2243509	C	G	999	.	<...>	GT:PL:DP:SP:GQ	0/1:255,0,128:37:8:99	0/1:212,0,231:41:7:99
9	130219669	rs2243906	C	T	999	.	<...>	GT:PL:DP:SP:GQ	0/1:255,0,255:68:2:99	0/1:255,0,252:65:2:99
9	130219743	rs2243903	T	C	999	.	<...>	GT:PL:DP:SP:GQ	0/1:255,0,255:80:3:99	0/1:244,0,255:51:1:99
9	130219990	rs2243898	G	C	999	.	<...>	GT:PL:DP:SP:GQ	0/1:255,0,224:69:0:99	0/1:255,0,227:48:11:99
9	130220661	rs2265685	T	C	999	.	<...>	GT:PL:DP:SP:GQ	1/1:255,105,0:35:0:99	1/1:245,63,0:21:0:99
9	130220663	rs35636470	G	A	999	.	<...>	GT:PL:DP:SP:GQ	0/1:220,0,136:36:0:99	0/1:140,0,142:22:2:99
9	130220673	rs7874732	C	A	120	.	<...>	GT:PL:DP:SP:GQ	0/1:67,0,234:35:0:70	0/1:88,0,141:15:0:91
9	130220678	rs28654608	C	A	105	.	<...>	GT:PL:DP:SP:GQ	0/1:54,0,229:36:0:57	0/1:86,0,143:16:0:89

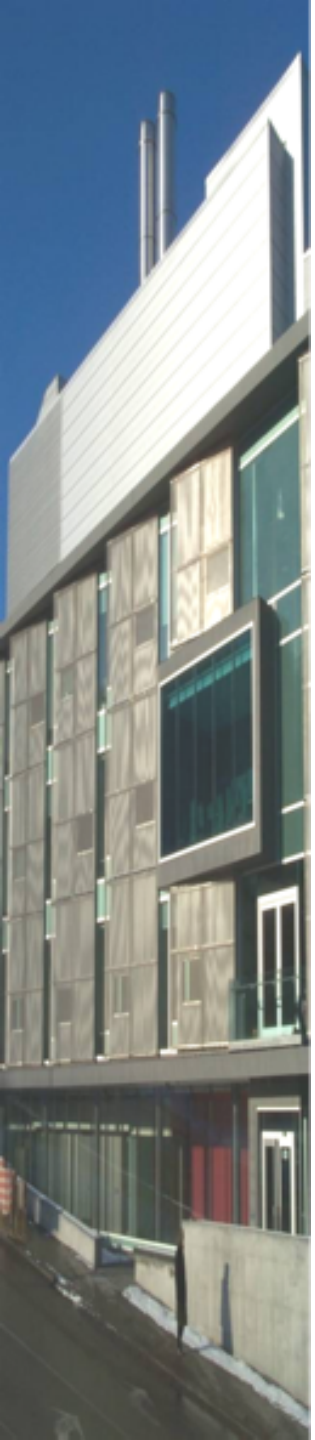
Column FORMAT defines ":" separated values

GT = Genotype

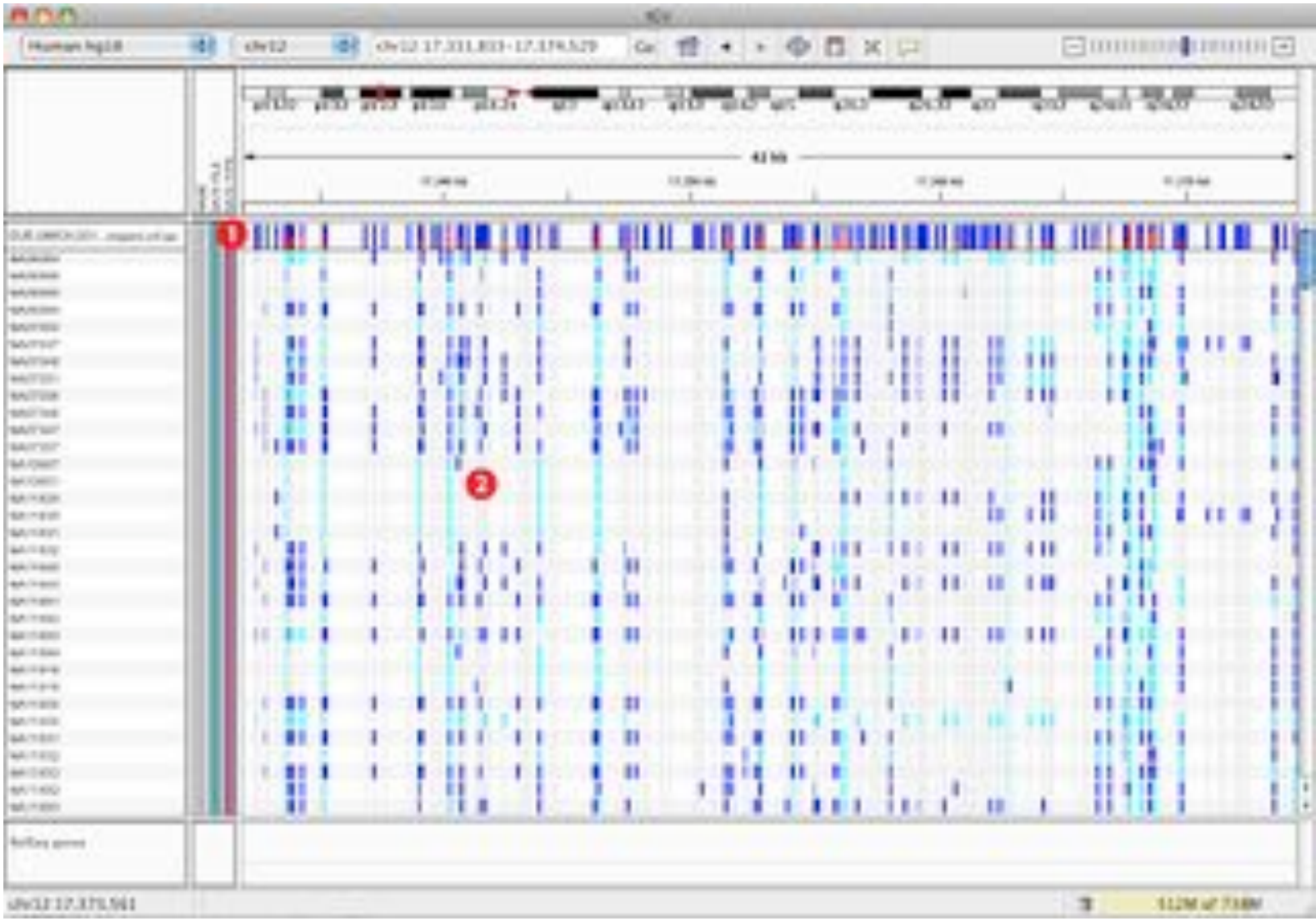
DP = depth

...

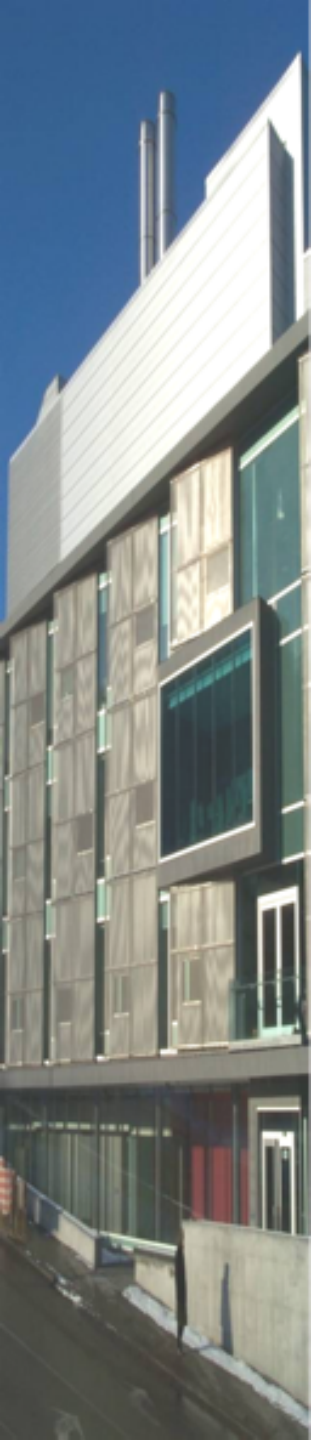
vcftools.sourceforge.net/specs.html



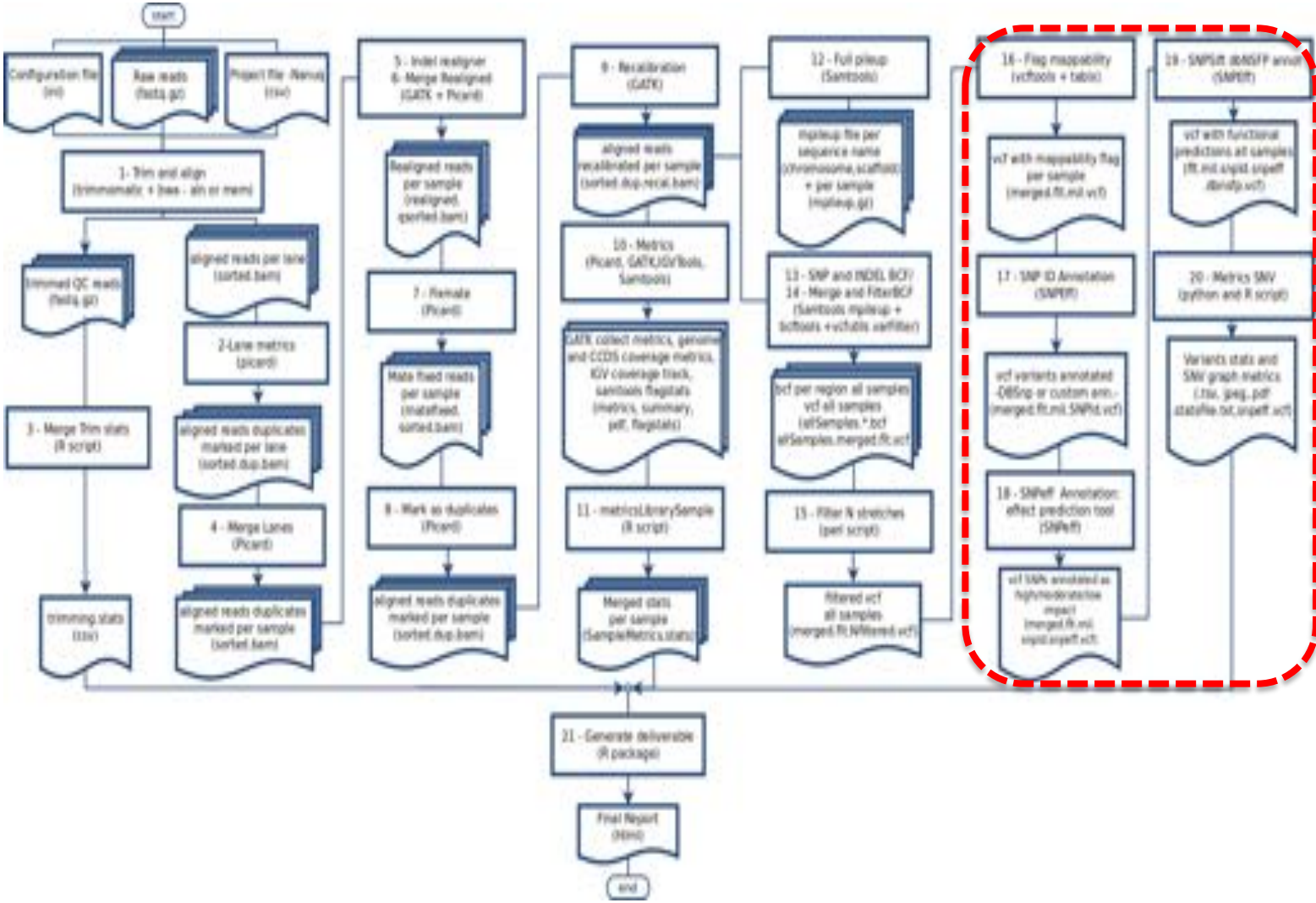
VCF visualization in IGV

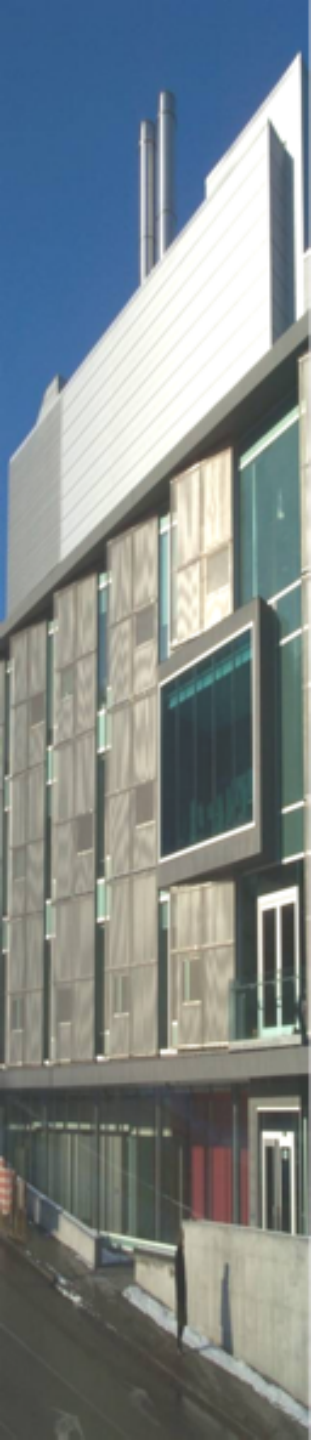


broadinstitute.org/igv/viewing_vcf_files



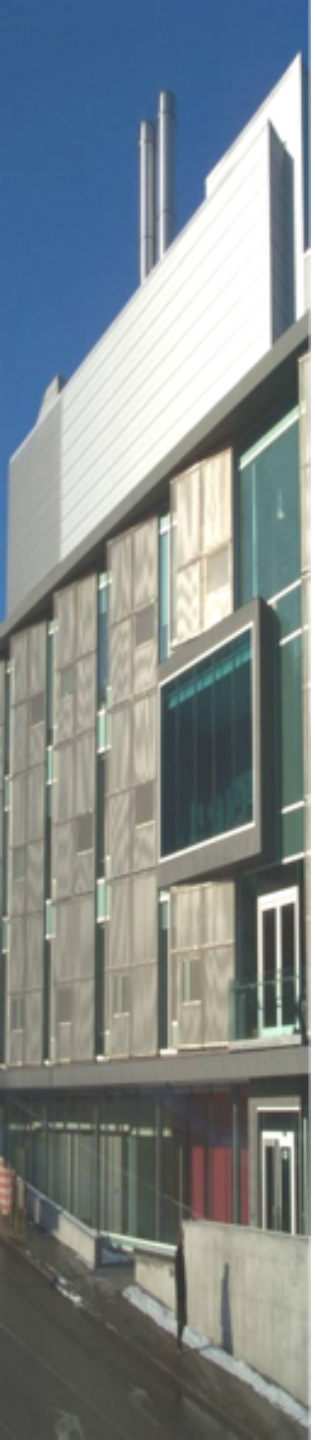
DNA-Seq: SNV annotation and metrics





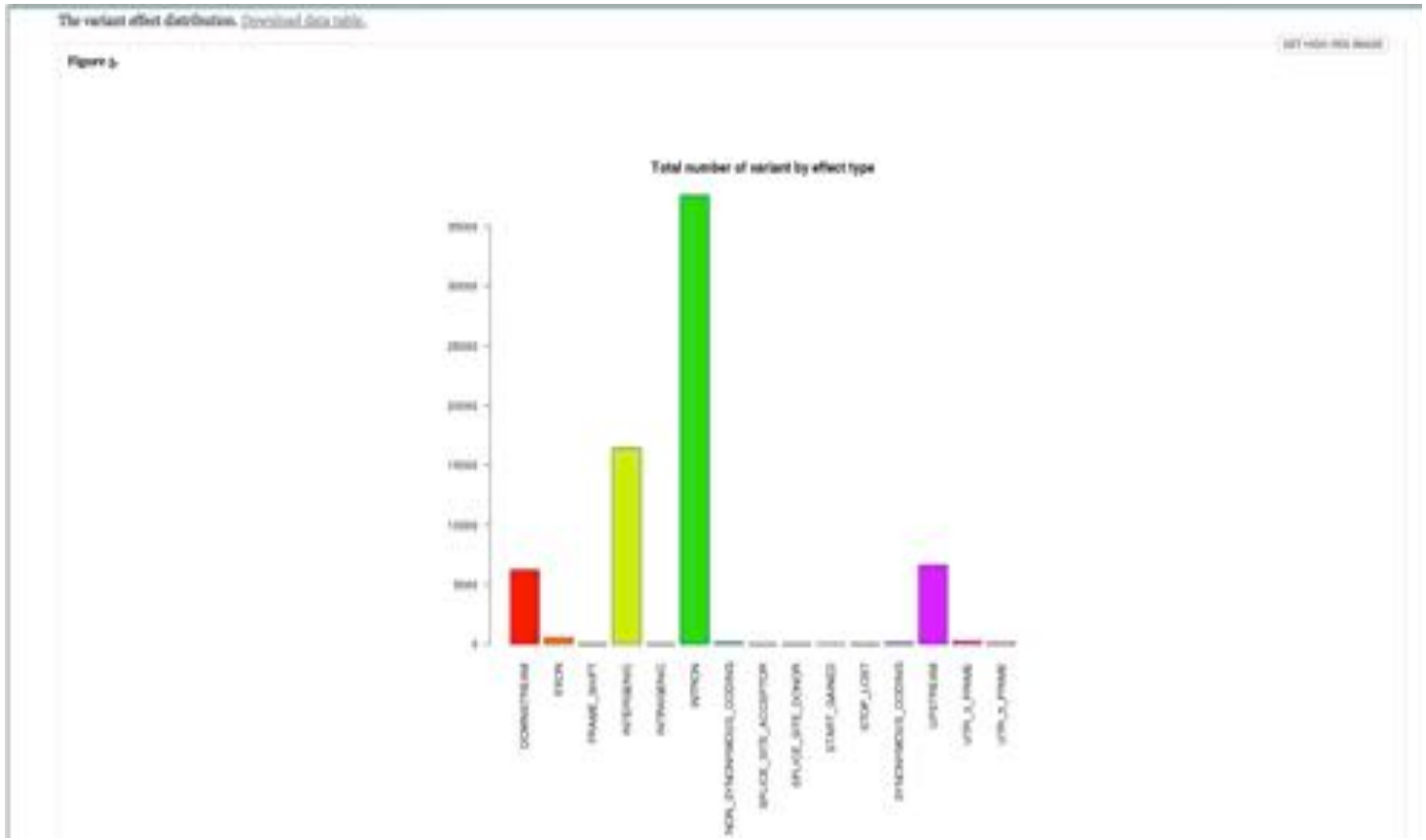
Variant annotation

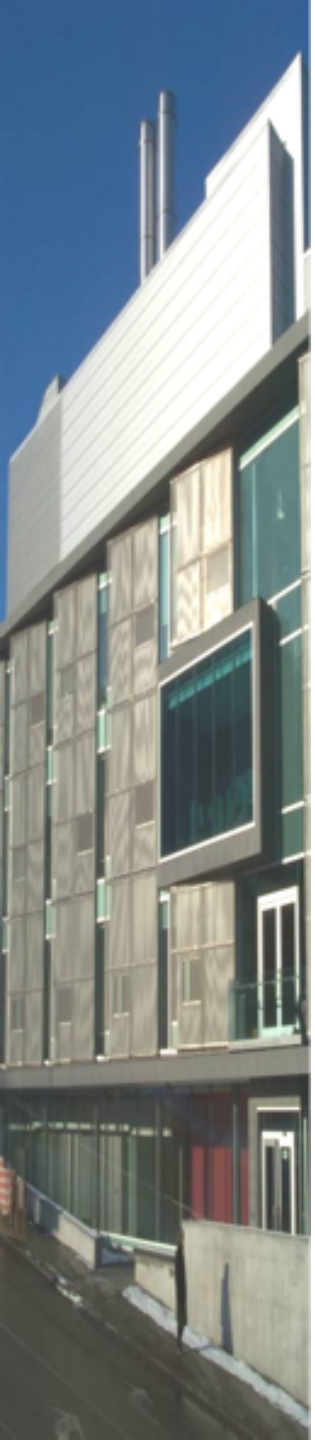
- Hypo- or hyper-mappabilty flag
  - Mark SNV in low confidence regions
- dbSNP [*SnpSift*]
  - Mark already known variant
- Variant effects [*SnpEff*]
  - predict the effects of variants on genes (such as amino acid changes)
- *dbNSFP* [*SnpSift*]
  - Functional annotations of the change
- *Cosmic*[*SnpSift*]
  - Known somatic mutations



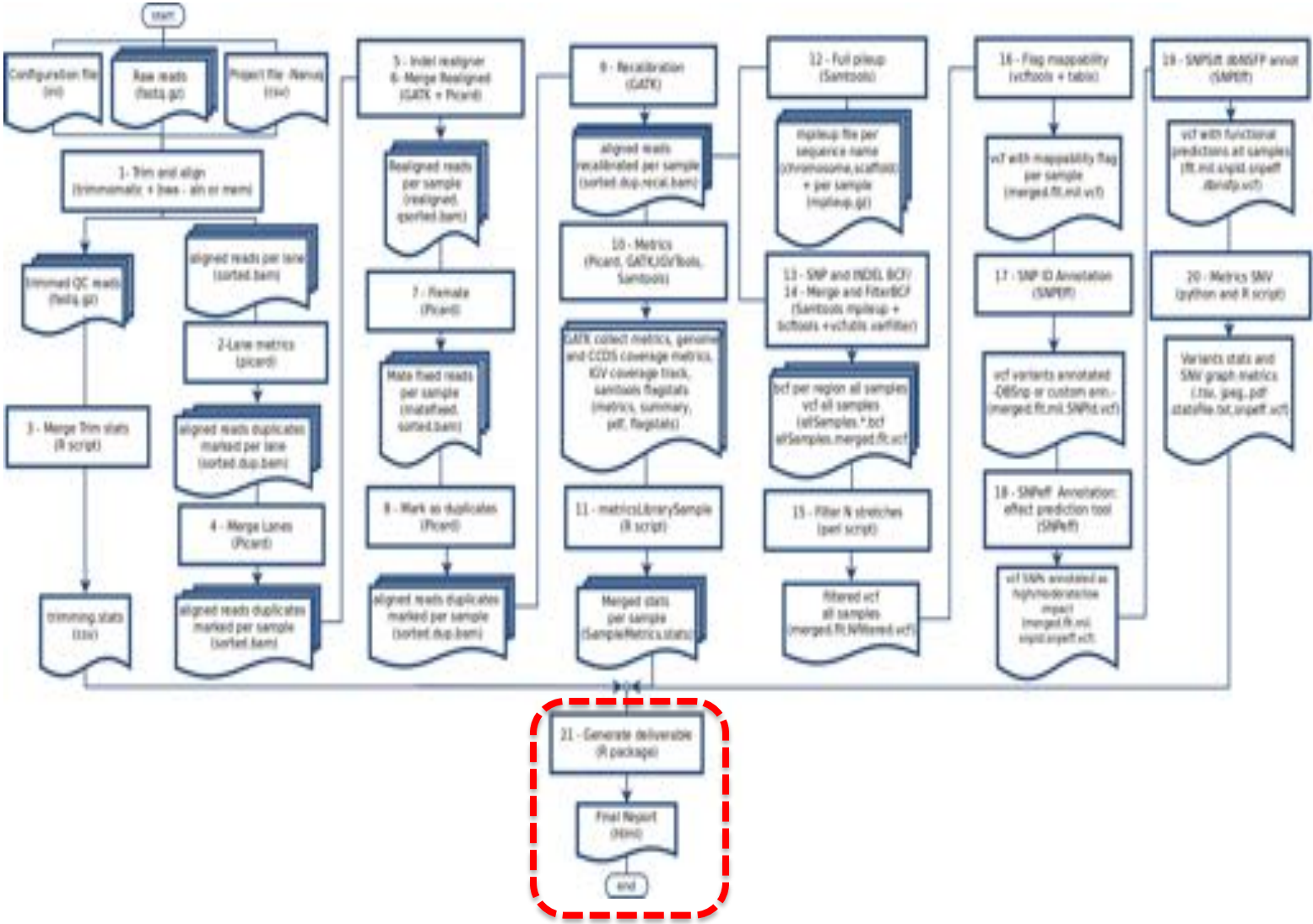
SNV statistics

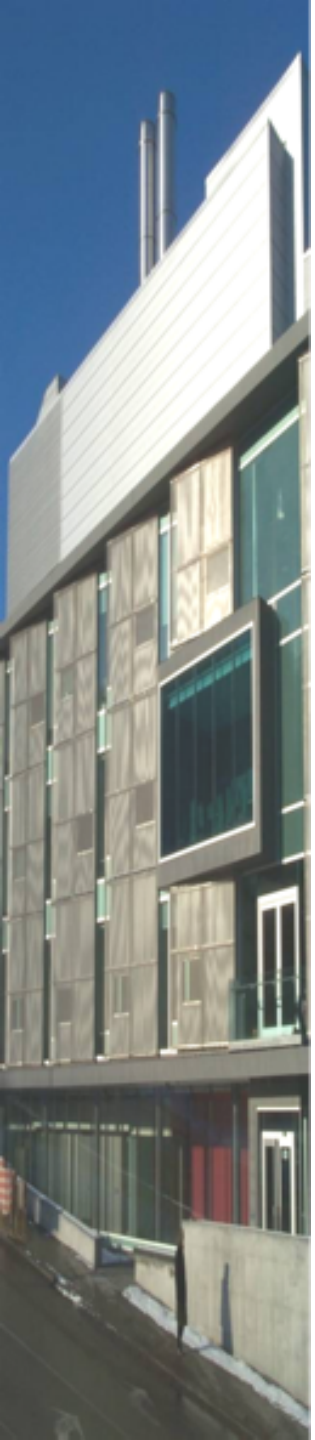
- Statistics are generated from the SNPeff stats outputs
- Example of one of the SNv metrics graph





DNA-Seq: Generate report





Home-made Rscript

Generate report

- Noozle-based html report which describe the entire analysis and provide QC, 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