Coverage Analysis for NGS Data Experiments

Marcel Caraciolo, Bioinformatician and CTO
marcel@genomika.com.br

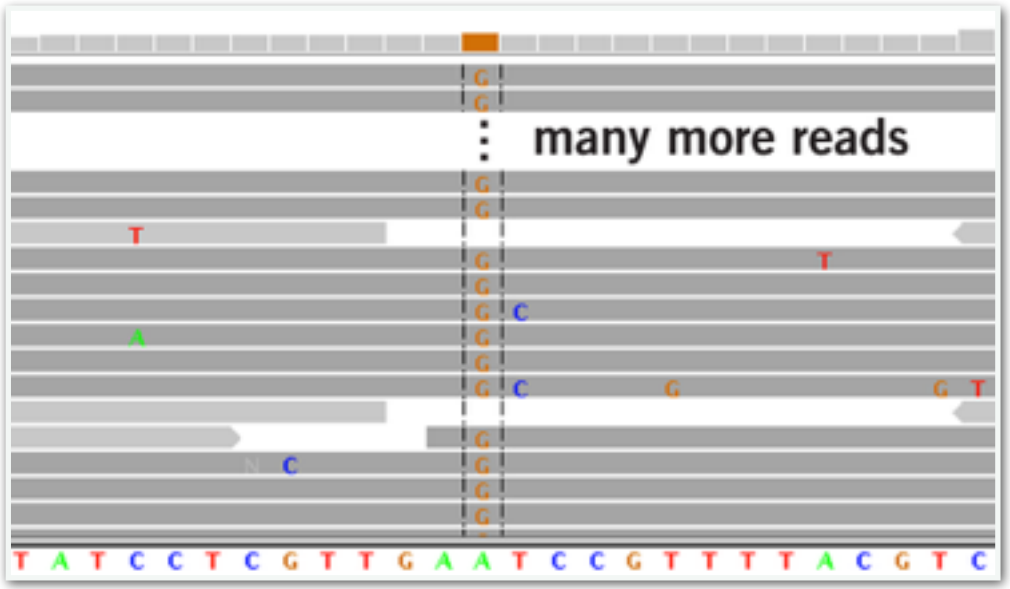
Outline

- Introduction to Coverage Analysis
- Main metrics and evaluation methods
- Hands-on (GATK, BedTools, Chanjo)
- Conclusions and further steps

What's Coverage ?

Coverage

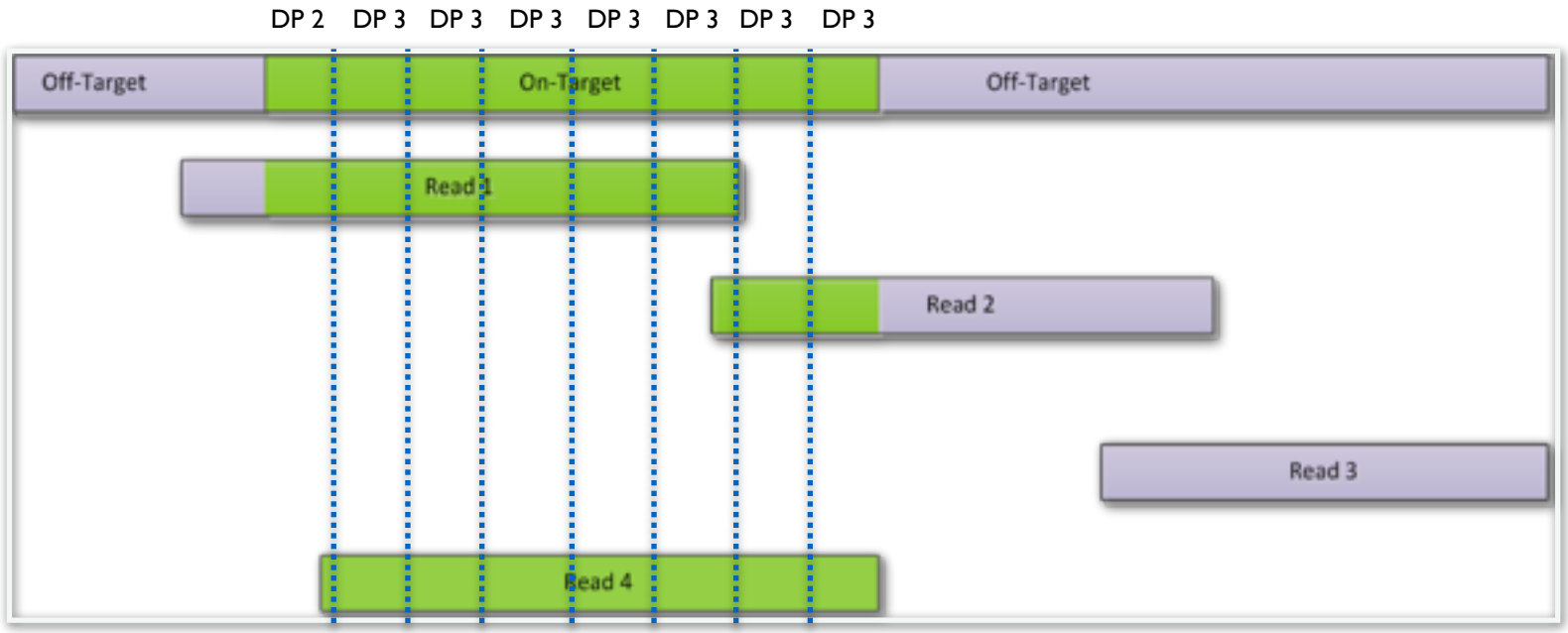
The read depth in a single position or how many times the base has been uniquely interrogated.



What's Coverage ?

As for a set of bases, such as a target region, it represents the average read depth across the given interval.

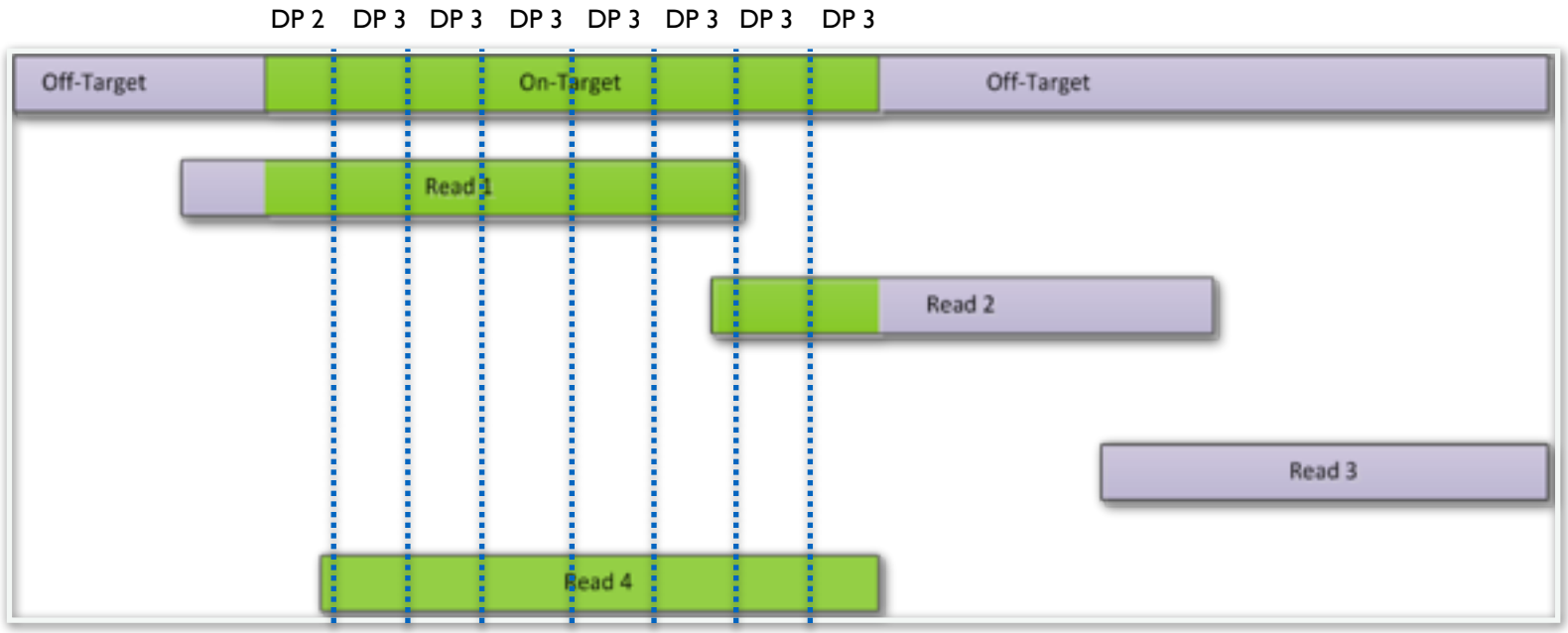
Average DP: 5.66

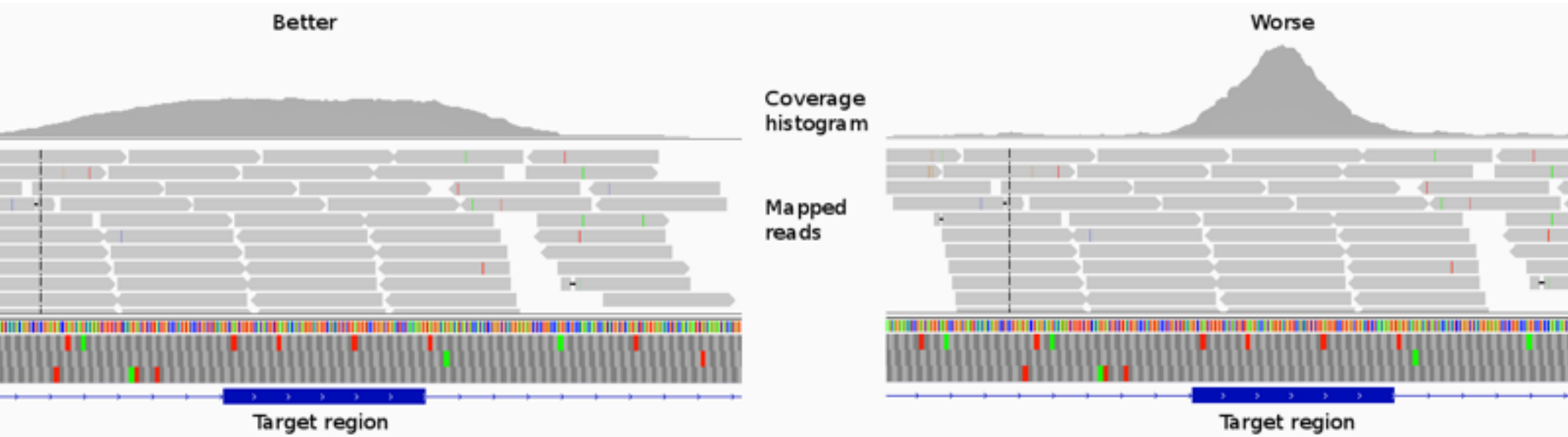


What's Coverage ?

As for a set of bases, such as a target region, it represents the average read depth across the given interval.

Average DP: 5.66




Better
Worse
Coverage histogram
Mapped reads
Target region
Target region

Coverage Analysis

Targeted enrichment sequencing by NGS is a common approach to interrogate specific loci or the whole exome, for instance, in the human exome.

Main problems:

- The efficiency and the lack of bias in the enrichment process needs to be assessed as a quality control step prior to downstream analysis of sequencing data.
- It can be used to distinguish and discard random sequencing and alignment errors. Disease-causing variants can likewise be missed due to poor coverage.
- Save your time, and of course to avoid incorrect conclusion from the analysis.

Coverage Analysis

Illustrated motivation and example:

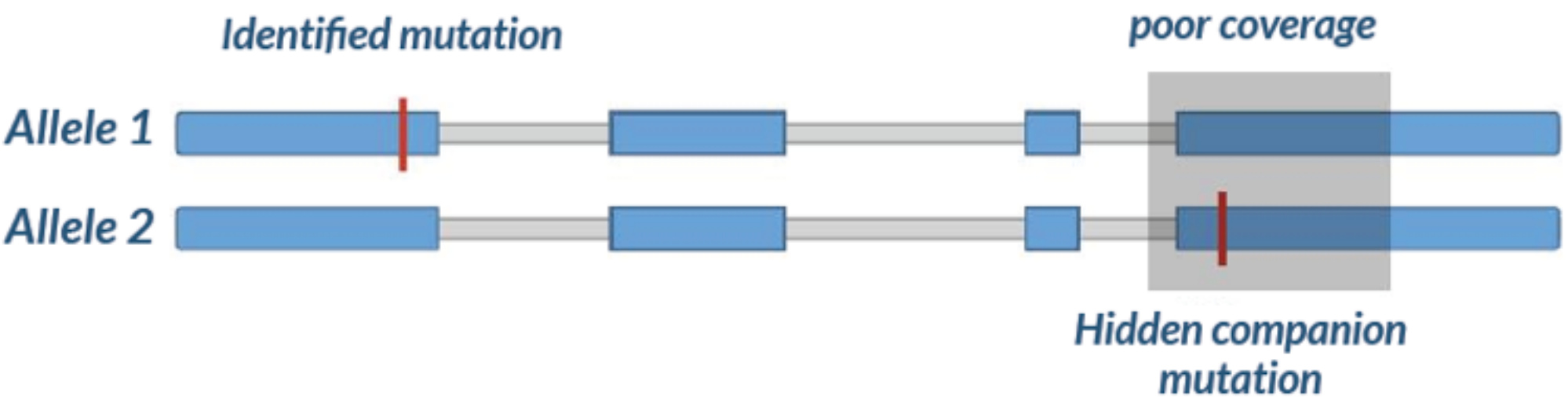


Illustration of a genomic region in a heterozygote case. The figure shows patient's two alleles of the investigated gene. The companion of the identified mutation is missed in an area of poor coverage.

Coverage Analysis

Rna-Seq and transcript analysis

We don't know the size/length of the transcriptome.

The best approach is to perform a pseudo coverage *(count of alignments spanning a genomic position)*

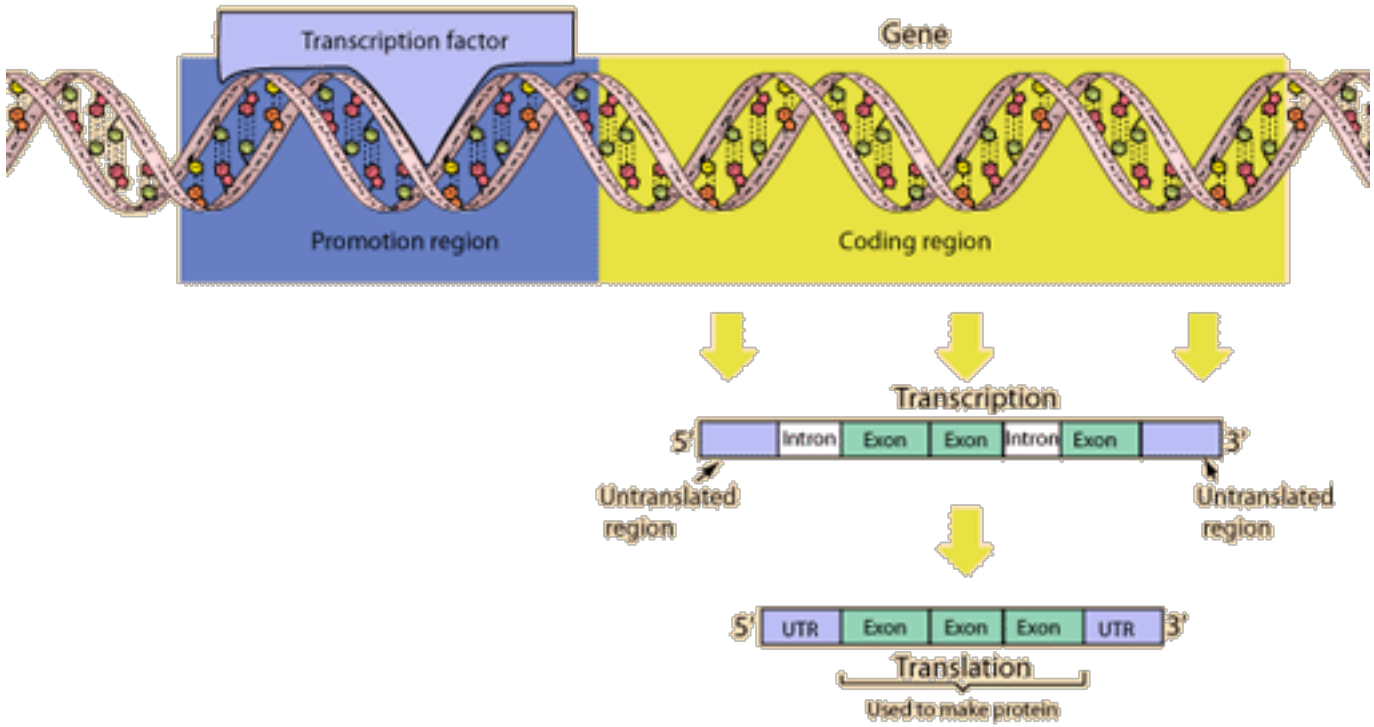
Metrics: RPKM (Reads per KB per million reads)

Before starting…

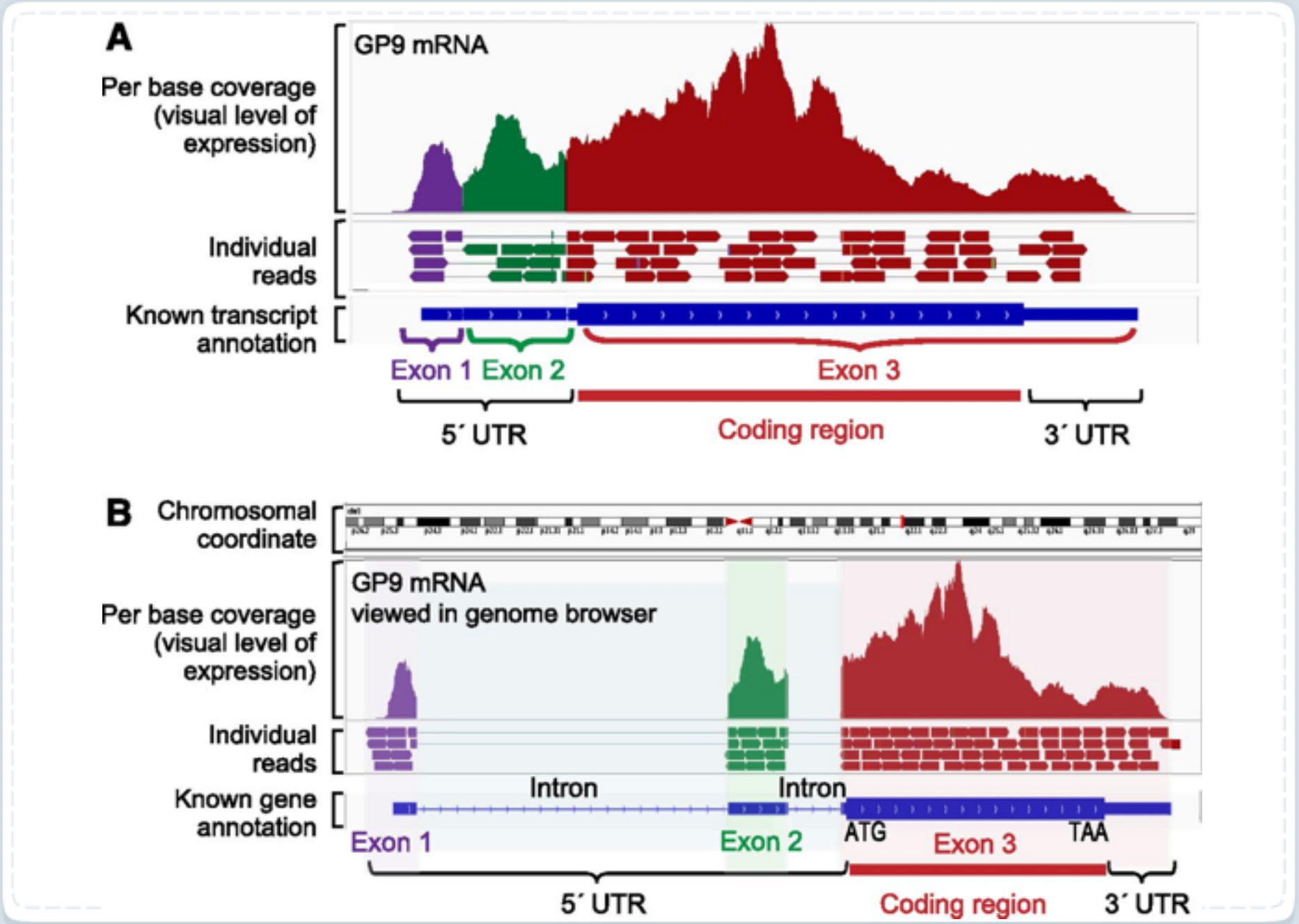
Some important definitions:

CDS (coding regions of gene)

Portion of a gene's DNA or RNA,composed of exons, that codes for protein.



Before starting . . .

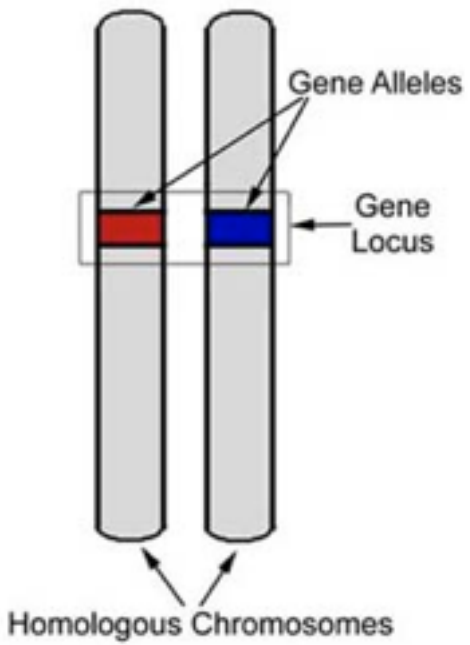


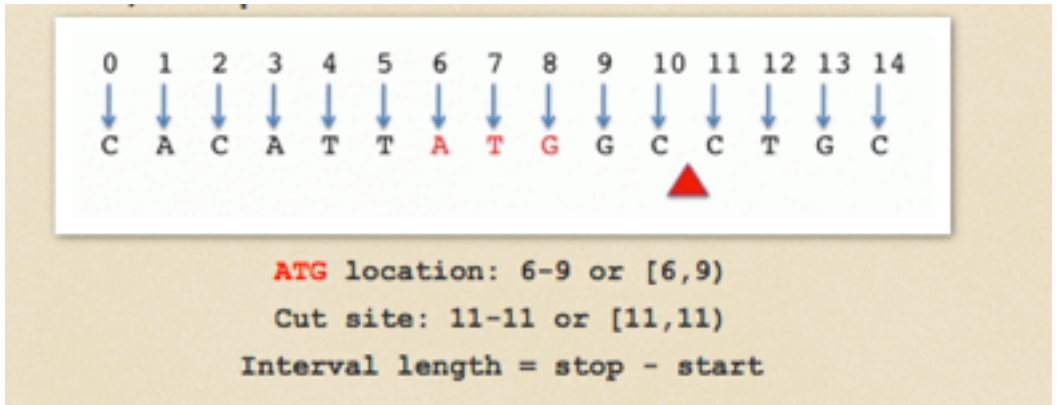
Before starting . . .

Some important definitions:

Loci coordinates (genomic coordinates)

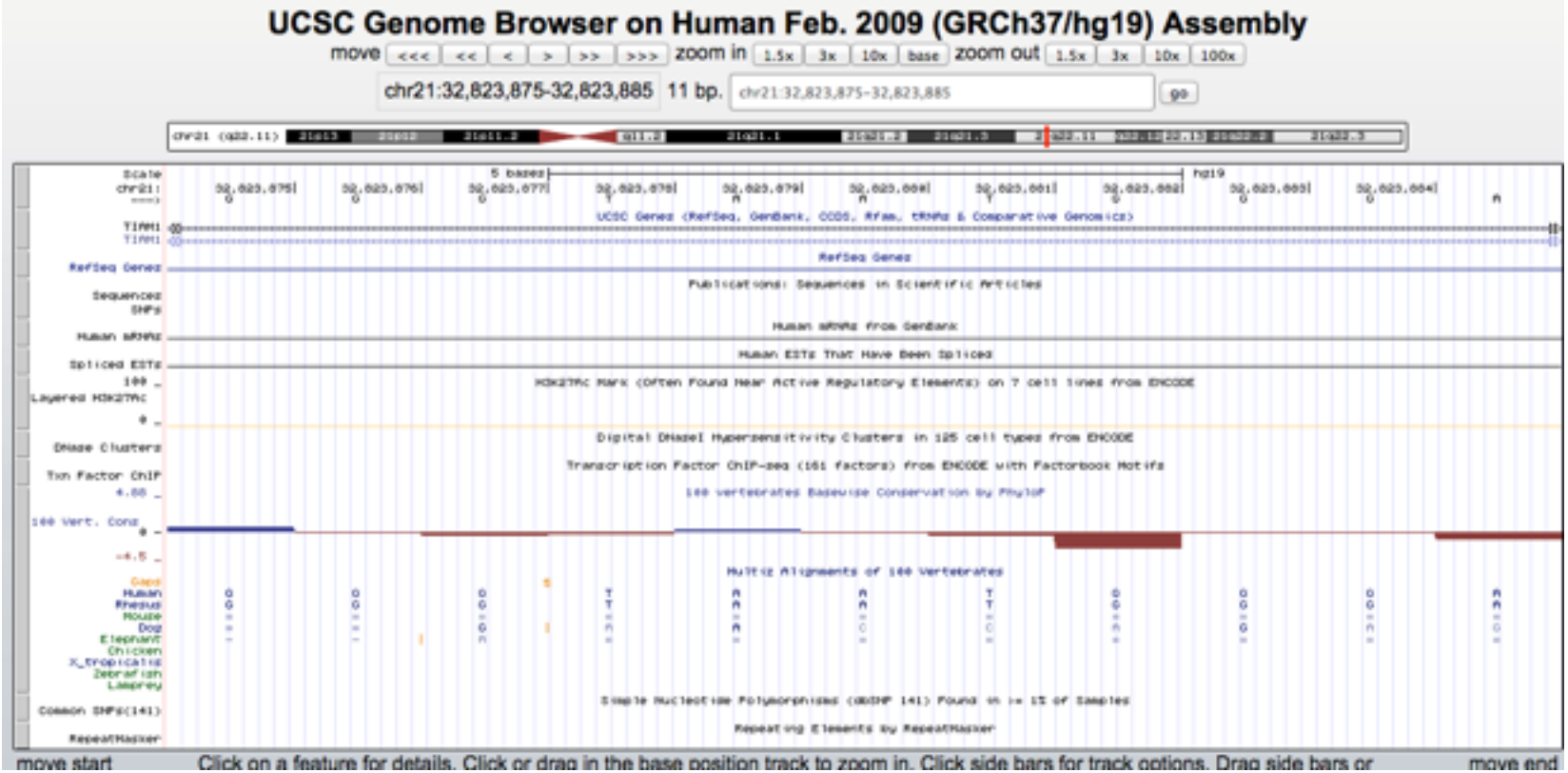
Specific location of a gene, DNA Sequence or position on a chromosome.





Before starting . . .

Some important definitions:

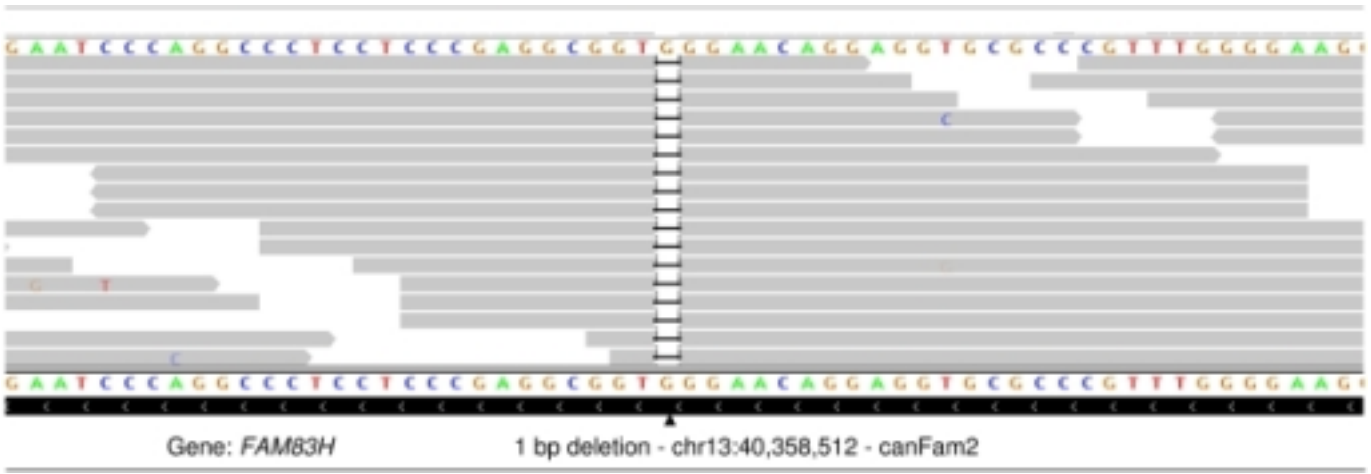


http://genome.ucsc.edu/cgi-bin/hgGateway

Before starting...

Some important definitions:

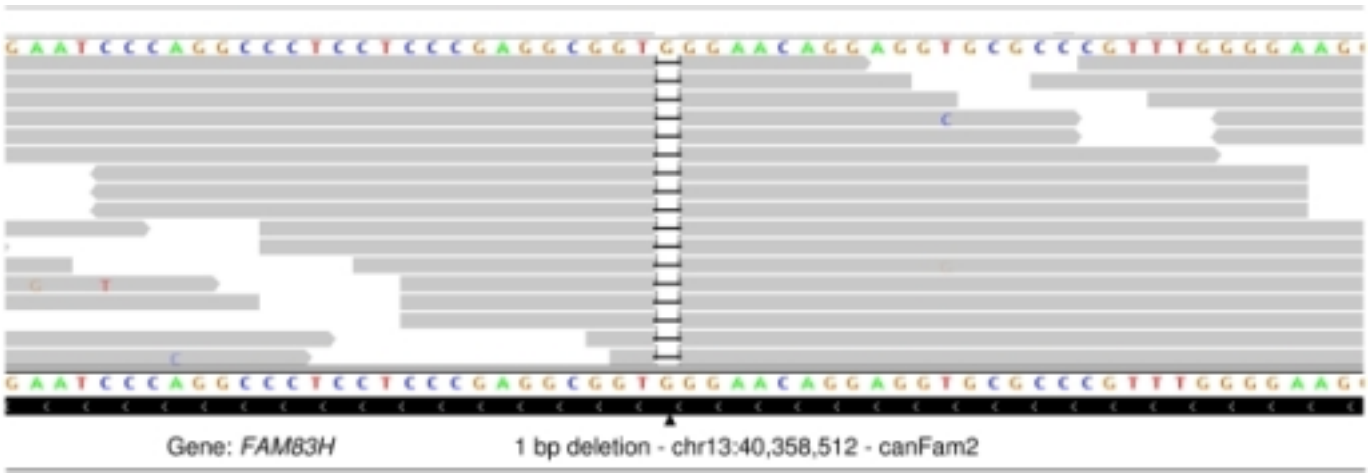
UCSC coordinates convention.



Before starting...

Some important definitions:

UCSC coordinates convention.



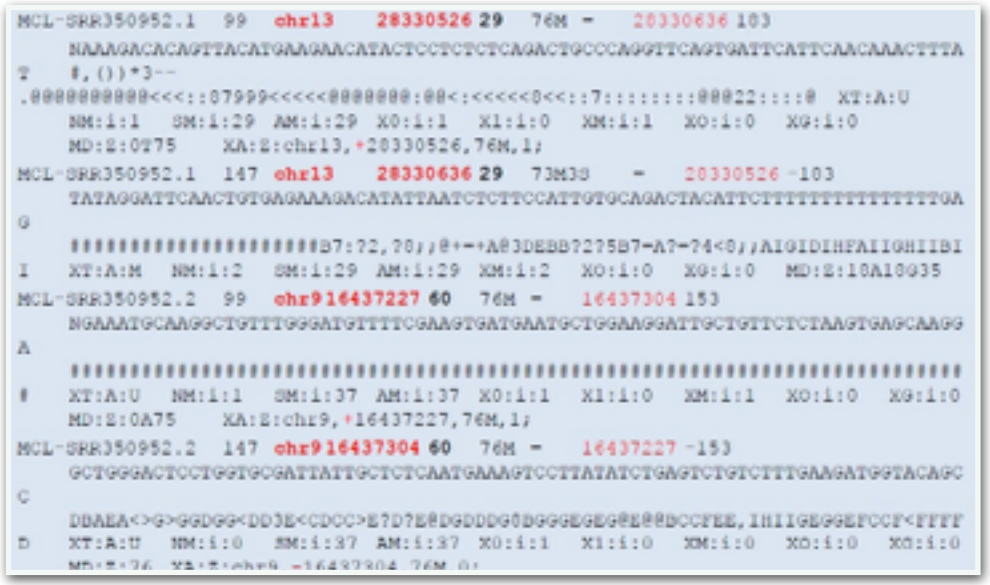
chr<chromsome>:<coordinate>

Before starting...

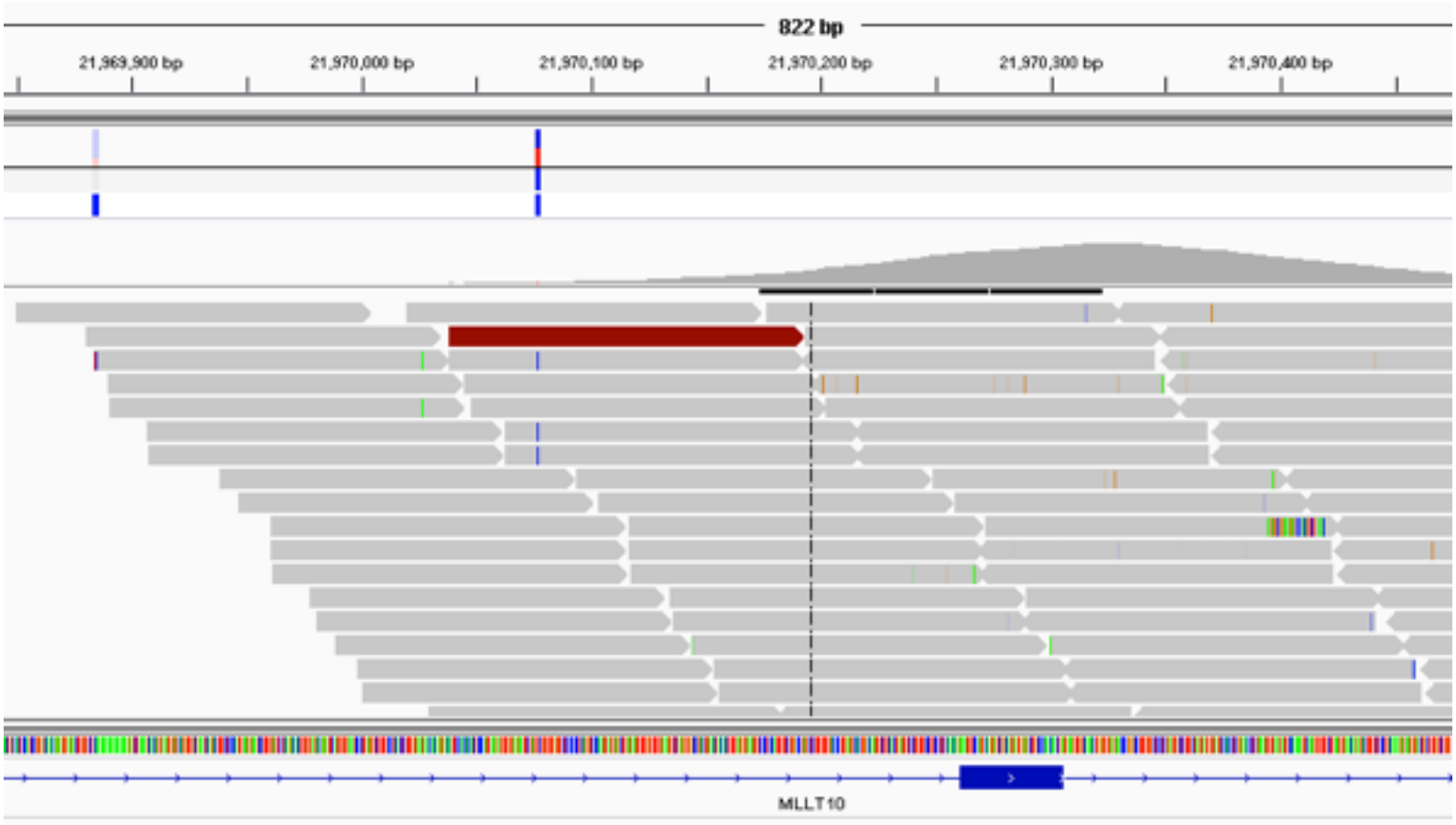
What files are important in the coverage analysis ?

BAM files

It's the binary version of a SAM file.SAM file is a tab-delimited text file that contains the sequence alignment data.



Before starting . . .



Integrative Genomics Viewer

http://www.broadinstitute.org/igv/

Hands-on: IGV

It requires Java Runtime at your machine.

http://www.broadinstitute.org/software/igv/download

Open at your IGV: http://www.bioinfomgp.org/_media/ngscat/download/example1.bam

Before starting...

What files are important in the coverage analysis ?

BED files

It's a tab-delimited text file that defines a feature track (target regions that we are interested to investigate).

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
chr3  127478198  127479365
chr3  127479365  127480532
chr3  127480532  127481699
```

Before starting . . .

What files are important in the coverage anal

B

chr7	55086970	55087058	EGFR_NM_005228_exon_1_chr7_f	0	+
chr7	55209978	55210130	EGFR_NM_005228_exon_2_chr7_f	0	+
chr7	55210997	55211181	EGFR_NM_005228_exon_3_chr7_f	0	+
chr7	55214298	55214433	EGFR_NM_005228_exon_4_chr7_f	0	+
chr7	55218986	55219055	EGFR_NM_005228_exon_5_chr7_f	0	+
chr7	55220238	55220357	EGFR_NM_005228_exon_6_chr7_f	0	+
chr7	55221703	55221845	EGFR_NM_005228_exon_7_chr7_f	0	+
chr7	55223522	55223639	EGFR_NM_005228_exon_8_chr7_f	0	+
chr7	55224225	55224352	EGFR_NM_005228_exon_9_chr7_f	0	+
chr7	55224451	55224525	EGFR_NM_005228_exon_10_chr7_f	0	+
chr7	55225355	55225446	EGFR_NM_005228_exon_11_chr7_f	0	+
chr7	55227831	55228031	EGFR_NM_005228_exon_12_chr7_f	0	+
chr7	55229191	55229324	EGFR_NM_005228_exon_13_chr7_f	0	+
chr7	55231425	55231516	EGFR_NM_005228_exon_14_chr7_f	0	+
chr7	55232972	55233130	EGFR_NM_005228_exon_15_chr7_f	0	+
chr7	55238867	55238906	EGFR_NM_005228_exon_16_chr7_f	0	+
chr7	55240675	55240817	EGFR_NM_005228_exon_17_chr7_f	0	+
chr7	55241613	55241736	EGFR_NM_005228_exon_18_chr7_f	0	+
chr7	55242414	55242513	EGFR_NM_005228_exon_19_chr7_f	0	+
chr7	55248985	55249171	EGFR_NM_005228_exon_20_chr7_f	0	+
chr7	55259411	55259567	EGFR_NM_005228_exon_21_chr7_f	0	+
chr7	55260458	55260534	EGFR_NM_005228_exon_22_chr7_f	0	+
chr7	55266409	55266556	EGFR_NM_005228_exon_23_chr7_f	0	+
chr7	55268008	55268106	EGFR_NM_005228_exon_24_chr7_f	0	+
chr7	55268880	55269048	EGFR_NM_005228_exon_25_chr7_f	0	+
chr7	55269427	55269475	EGFR_NM_005228_exon_26_chr7_f	0	+
chr7	55270209	55270318	EGFR_NM_005228_exon_27_chr7_f	0	+
chr7	55272948	55273310	EGFR_NM_005228_exon_28_chr7_f	0	+
chr1	115258671	115258798	NRAS_PROBE_2_chr1_r	0	-
chr1	115256484	115256570	NRAS_PROBE_3_chr1_r	0	-
chr1	115252220	115252348	NRAS_PROBE_4_1_chr1_r	0	-
chr1	115252190	115252239	NRAS_PROBE_4_2_chr1_r	0	-
chr12	25398252	25398315	KRAS_PROBE_2_chr12_r	0	-
chr12	25380220	25380317	KRAS_PROBE_3_chr12_r	0	-
chr12	25378600	25378721	KRAS_PROBE_4_1_chr12_r	0	-
chr12	25378548	25378615	KRAS_PROBE_4_2_chr12_r	0	-

63
30
97
64
31
98
65
32
99

Hands On Bed file

http://genome.ucsc.edu/cgi-bin/hgTables

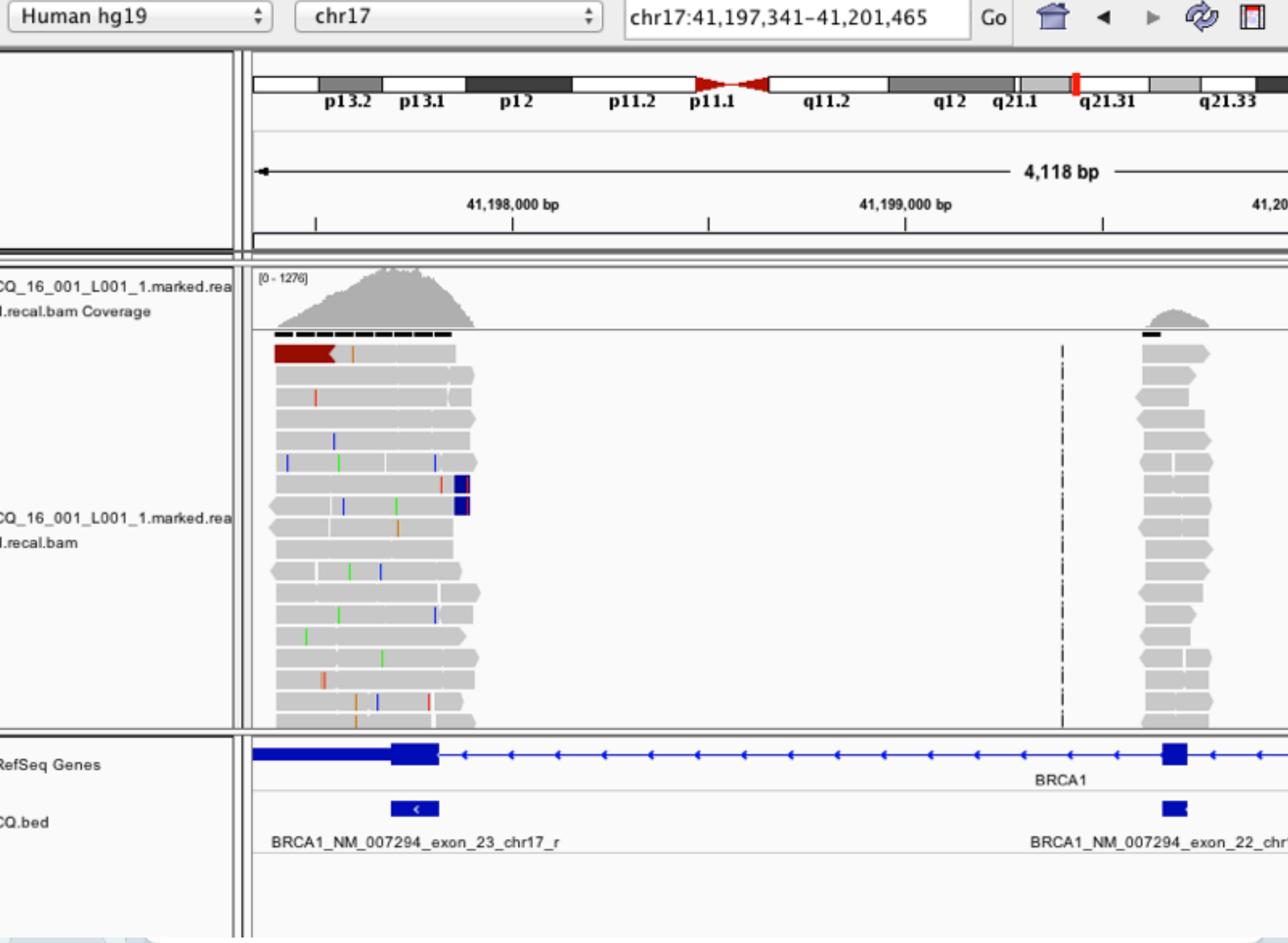
Hands-on: IGV

Open a bed file with your bam at your IGV.

http://www.broadinstitute.org/software/igv/download

Open at your IGV: http://www.bioinfomgp.org/_media/ngscat/download/example1.bam

http://www.bioinfomgp.org/_media/ngscat/download/seqcap.example1.bed


Human hg19
chr17
chr17:41,197,341-41,201,465
Go
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
4,118 bp
41,198,000 bp
41,199,000 bp
41,20
[0 - 1276]
CQ_16_001_L001_1.marked.rea
l.recal.bam Coverage
CQ_16_001_L001_1.marked.rea
l.recal.bam
RefSeq Genes
BRCA1
CQ.bed
BRCA1_NM_007294_exon_23_chr17_r
BRCA1_NM_007294_exon_22_chr

Main metrics

X coverage or depth coverage

Most common coverage metric. For a simple position it means the read depth. For a set of bases, such as target region, it represents the average read depth across the given interval.

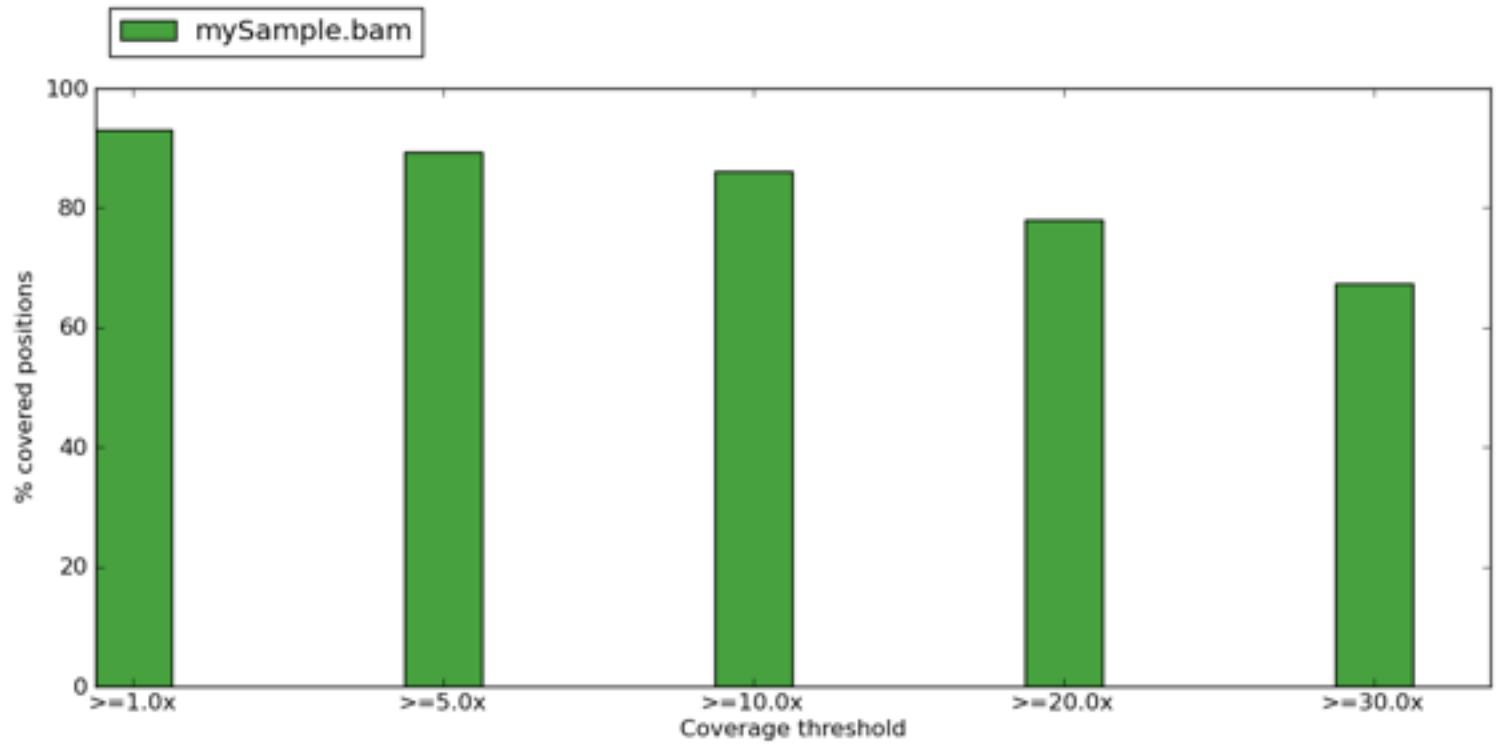
% coverage or breadth of coverage

It representes to what degree a given sequence has been covered at a given read depth. A cutoff for acceptable variant calling can be set at 30x read depth. % coverage means what ratio of sequence that passes our quality cutoff.

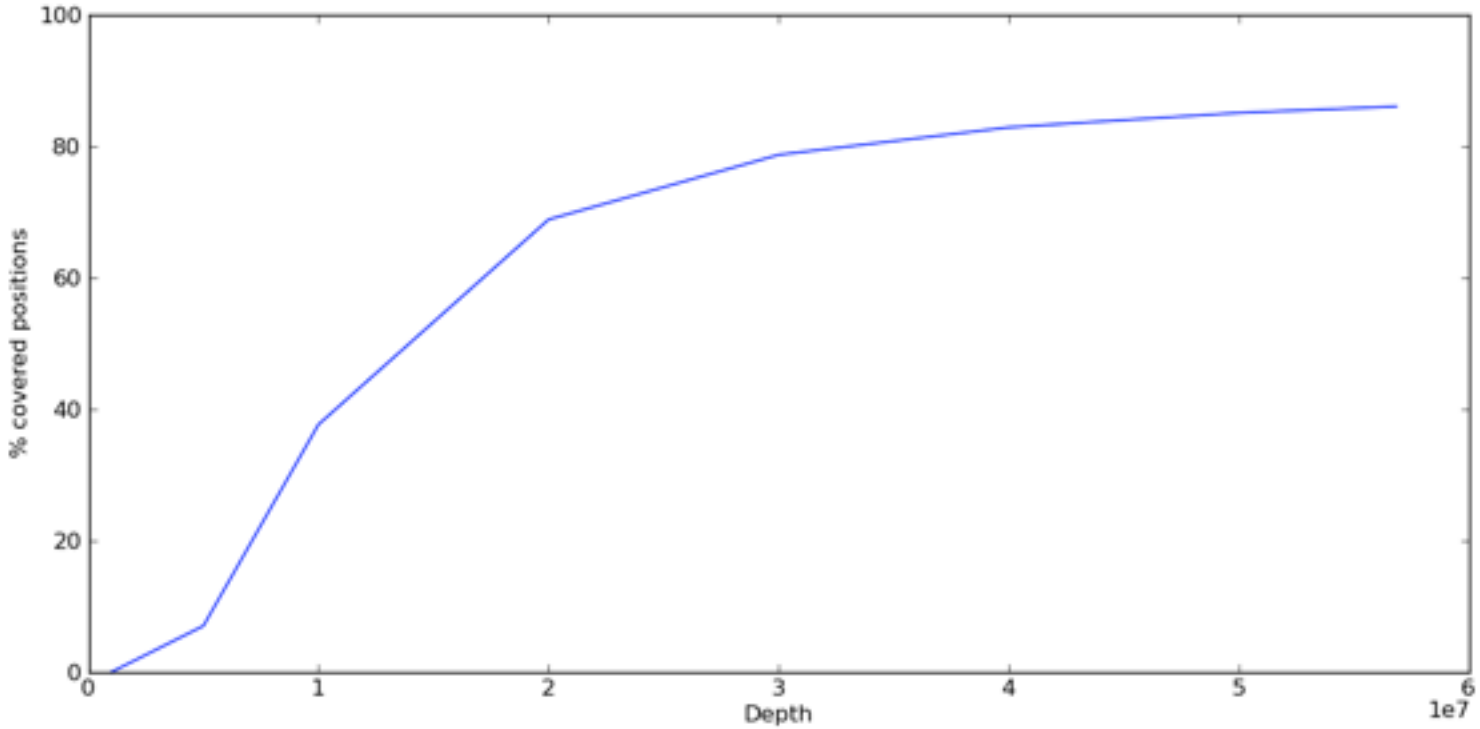
Physical coverage

It means the structural coverage for the whole chromosome. It is important for reliably calling structural variations.

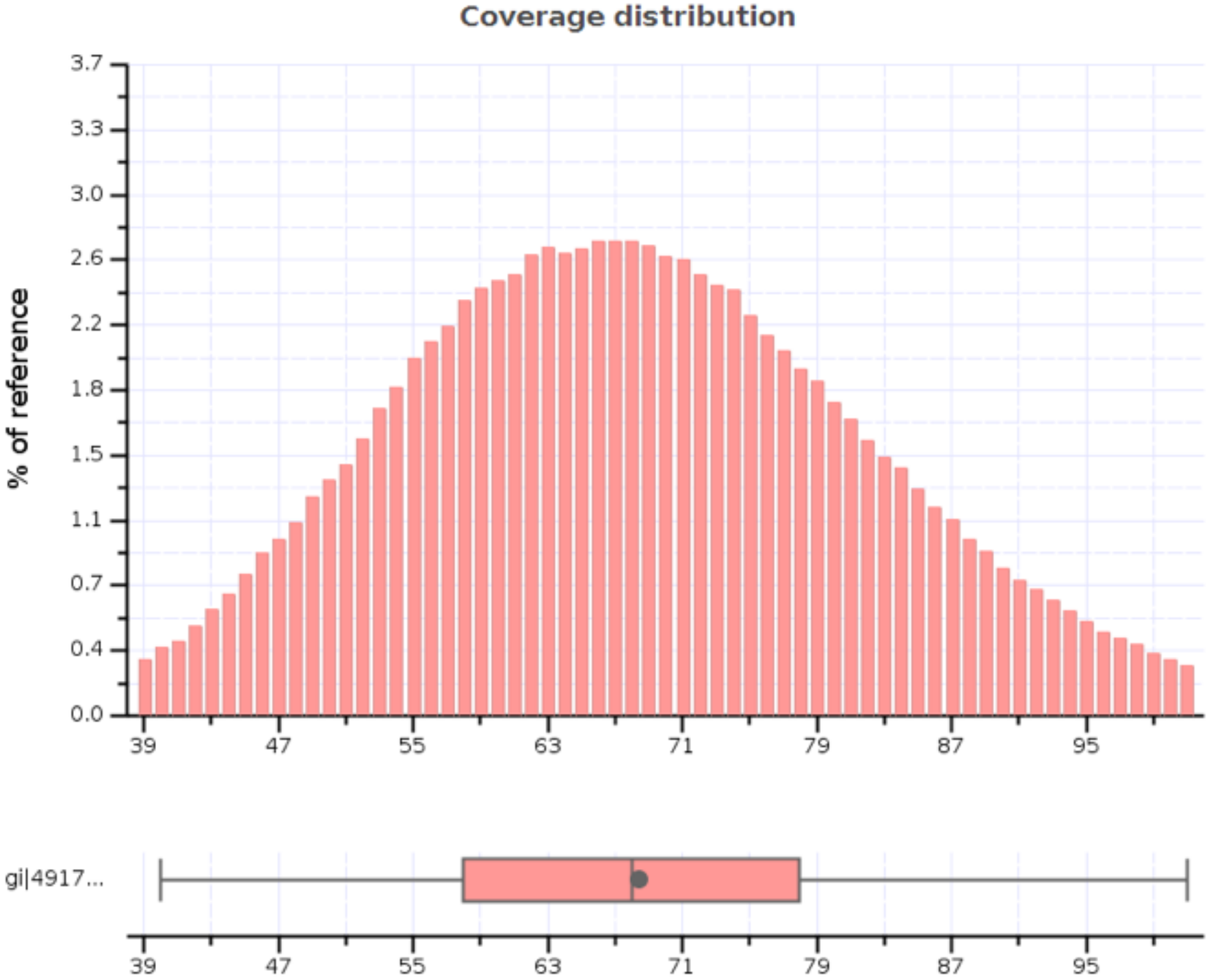
Main visualisations



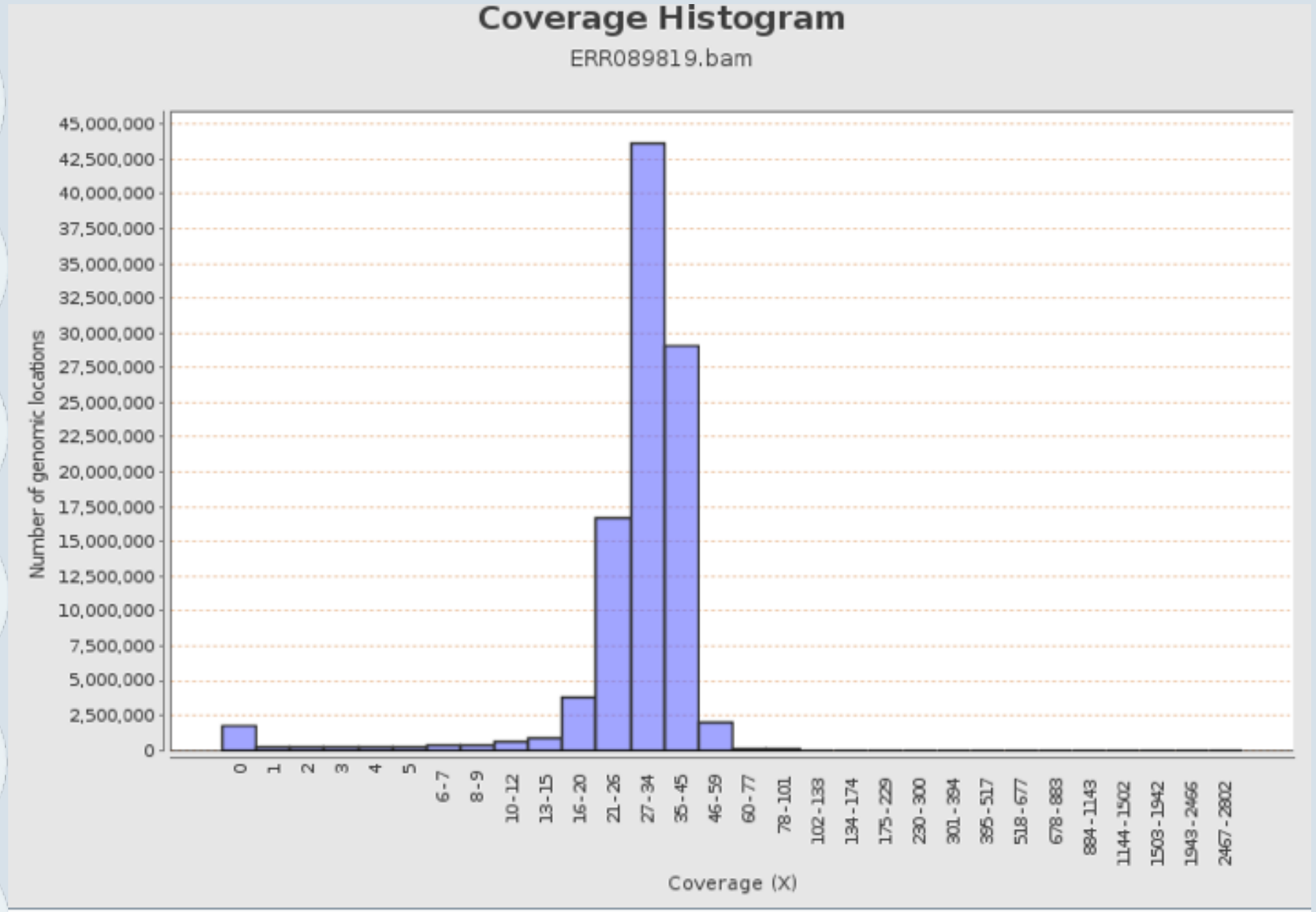
Main visualisations



Main visualisations



Main visualisations



Tools

Most used tools are:

Bedtools, open-source, Python and bash

GATK, open-source, Java and bash

Chanjo, open-source, Python and bash

BioConductor, open-source, R and bash

Bedtools

Fast, flexible toolset
for genomic arithmetic.

http://bedtools.readthedocs.org/

```
bedtools coverage -a reads.bed -b windows10kb.bed | head
chr1 0     10000 0  10000 0.00
chr1 10001 20000 33 10000 0.21
chr1 20001 30000 42 10000 0.29
chr1 30001 40000 71 10000 0.36
```

GATK

Genome Analysis Toolkit
to analyze high-throughput
sequencing
data.

gatk

```
java -Xmx2g -jar GenomeAnalysisTK.jar \
  -R ref.fasta \
  -T DepthOfCoverage \
  -o file_name_base \
  -I input_bams.list
  [-geneList refSeq.sorted.txt] \
  [-pt readgroup] \
  [-ct 4 -ct 6 -ct 10] \
  [-L my_capture_genes.interval_list]
```

https://www.broadinstitute.org/gatk/gatkdocs/
org_broadinstitute_gatk_tools_walkers_coverage_DepthOfCoverage.php

GATK

Genome Analysis Toolkit
to analyze high-throughput sequencing
data

if you supply the -geneList argument, DepthOfCoverage will output an additional summary file that looks as follows:

Gene_Name	Total_Cvg	Avg_Cvg	Sample_1_Total_Cvg	Sample_1_Avg_Cvg	Sample_1_Cvg_Q3	Sample_1_Cvg_Median	Sample_1_Cvg_Q1
SORT1	594710	238.27	594710	238.27	165	245	330
NOTCH2	3011542	357.84	3011542	357.84	222	399	>500
LMNA	563183	186.73	563183	186.73	116	187	262
NOS1AP	513031	203.50	513031	203.50	91	191	290

```
[-ct 4 -ct 6 -ct 10] \
[-L my_capture_genes.interval_list]
```

https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_coverage_DepthOfCoverage.php

Chanjo

Coverage analysis
for clinical sequencing.

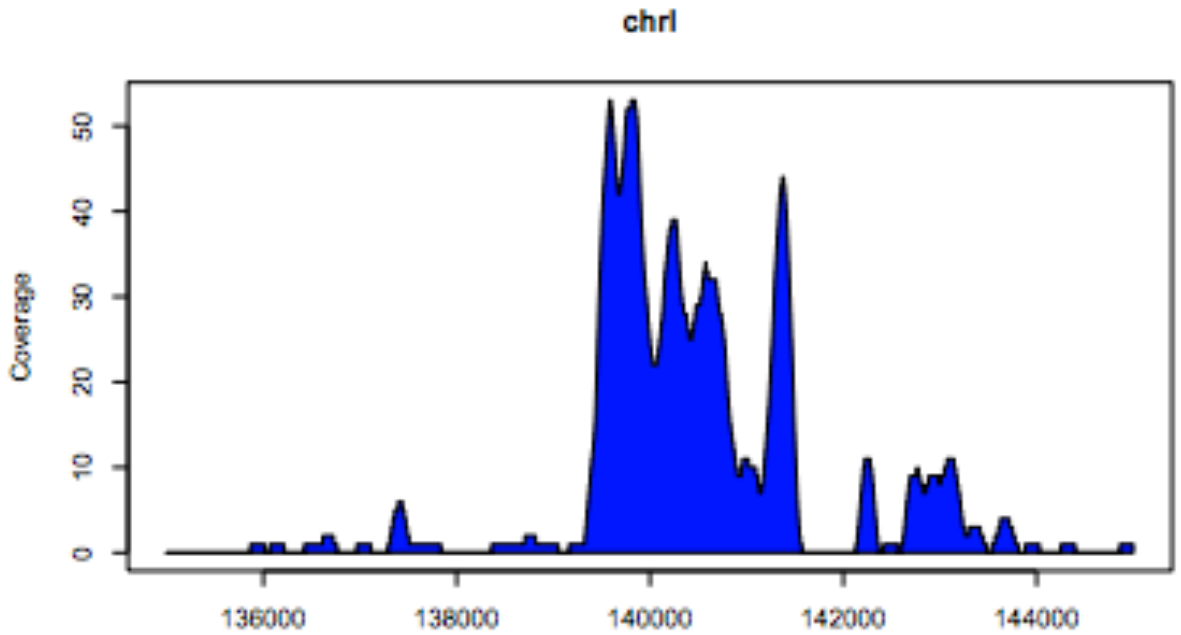
```
$ cat intervals.bed | chanjo annotate alignment.bam
#{"sample_id": "bavewira", ...}
1       10      15      interval-1      9.922   0.97231
2       45      55      interval-2      14.231  1.0
```

http://www.chanjo.co/en/latest/index.html

Bioconductor

http://master.bioconductor.org/help/course-materials/2010/SeattleJan10/day3/CoverageEDA.pdf

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



Conclusions

Simple introduction to Coverage Analysis

Main tools for downstream analysis

Further metrics: GC content, specificity, sensitivity, uniformity.

References

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., Wang, J., Wang, J., Yang, H., and Zhang, X. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biology, 12(9), R95.

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. Nature Biotechnology, 29(10), 908–914.

Frommolt, P., Abdallah, A. T., Altmller, J., Motameny, S. e., Thiele, H., Becker, C., Stemshorn, K., Fischer, M., Freil inger, T., and Nrnberg, P. (2012). Assessing the enrichment performance in targeted resequencing experiments. Human Mutation, 33(4), 635–641.

Hummel, M., Bonnin, S., Lowy, E., and Roma, G. (2011). TEQC: an r package for quality control in target capture experiments. Bioinformatics, 27(9), 1316–1317. Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842.

Sikkema-Raddatz, B., Johansson, L. F., de Boer, E. N., Almomani, R., Boven, L. G., van den Berg, M. P., van Spaendonck-Zwarts, K. Y., van Tintelen, J. P., Sijmons, R. H., Jongbloed, J. D. H., and Sinke, R. J. (2013). Targeted next-generation sequencing can replace sanger sequencing in clinical diagnostics. Human Mutation, 34(7), 1035–1042.

Tewhey, R., Nakano, M., Wang, X., Pabon-Pe ´ na, C., Novak, B., Giuffre, A., Lin, E., ¨Happe, S., Roberts, D. N., LeProust, E. M., et al. (2009). Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol, 10(10), R116

Coverage Analysis for NGS Data Experiments

Marcel Caraciolo, Bioinformatician and CTO
marcel@genomika.com.br