Basic utilities in NGS-based research

KAWAJI, Hideya
kawaji AT gsc.riken.jp

RIKEN OSC シーケンサ利用技術講習会, 5th

(for Illumina Genome Analyzer)
http://www.osc.riken.jp/event/101216/

Goal of this talk

- Introduction of basic utilities, with some concrete steps/commands

- Go through a set of computation flow

Read the original articles/ documents to understand the principles

Might be outdated

Far from comprehensive

Target audience

No instruction of installation

- UNIX and R users, with
- Basic understanding of gene expression and epigenome
- Conceptual understanding of NGS analysis

An example of analysis flow and tools

- Mapping to the reference genome

  BWA, SAMtools

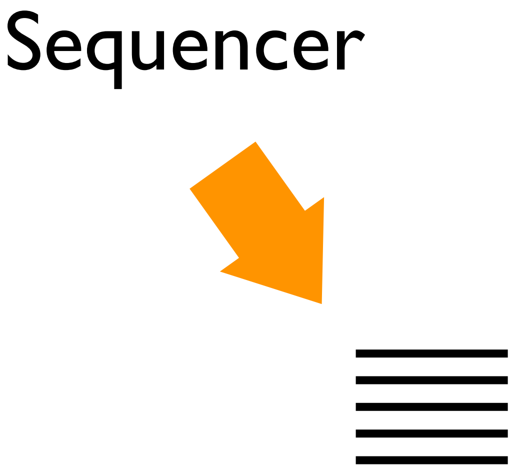
- Work on the genomic coordinates

  SAMtools, BEDTools, UCSC Tools

- Expression analysis / peak detection

  edgeR / MACS

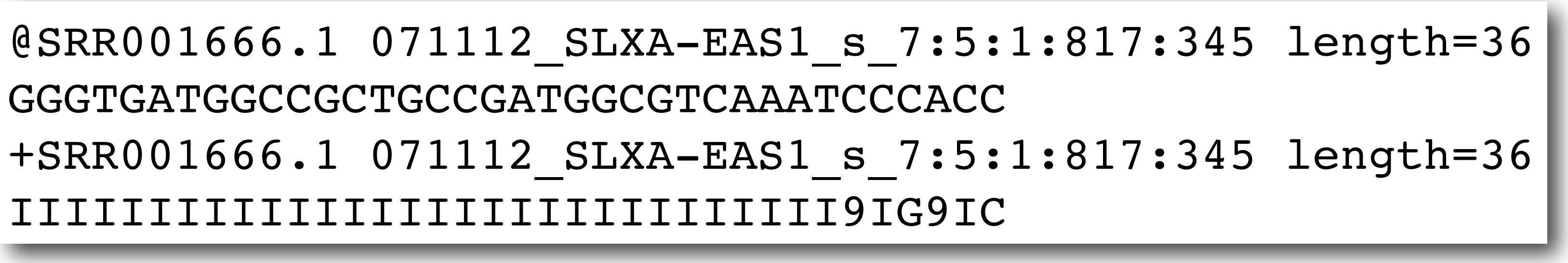
Sequencer output



Sequencer

Sequence (base) quality

- Encoded in FASTQ (PMID: 20015970)



```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$$

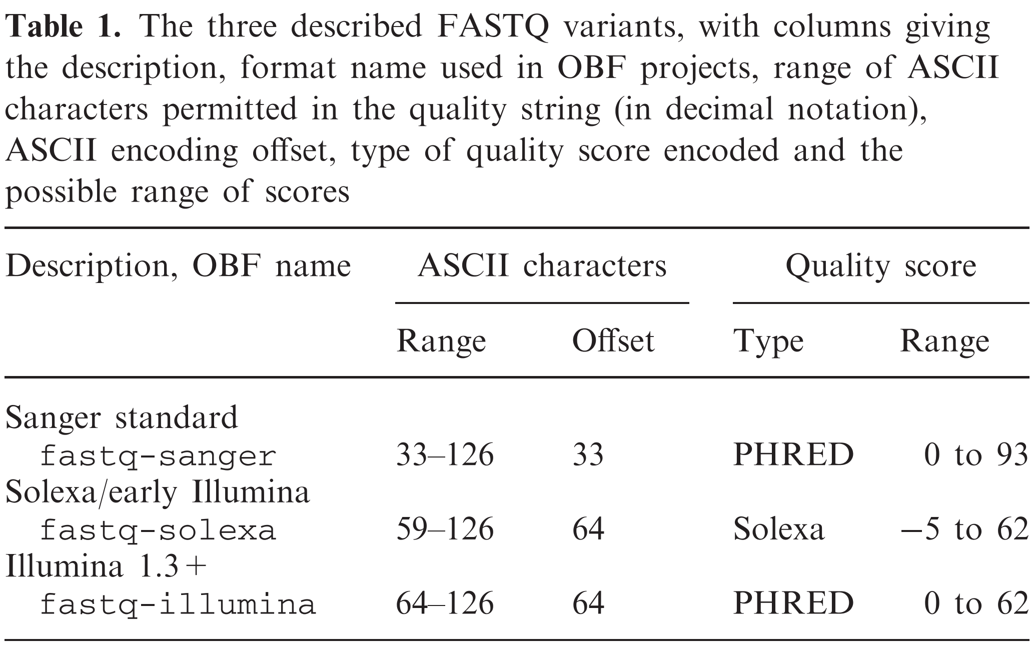
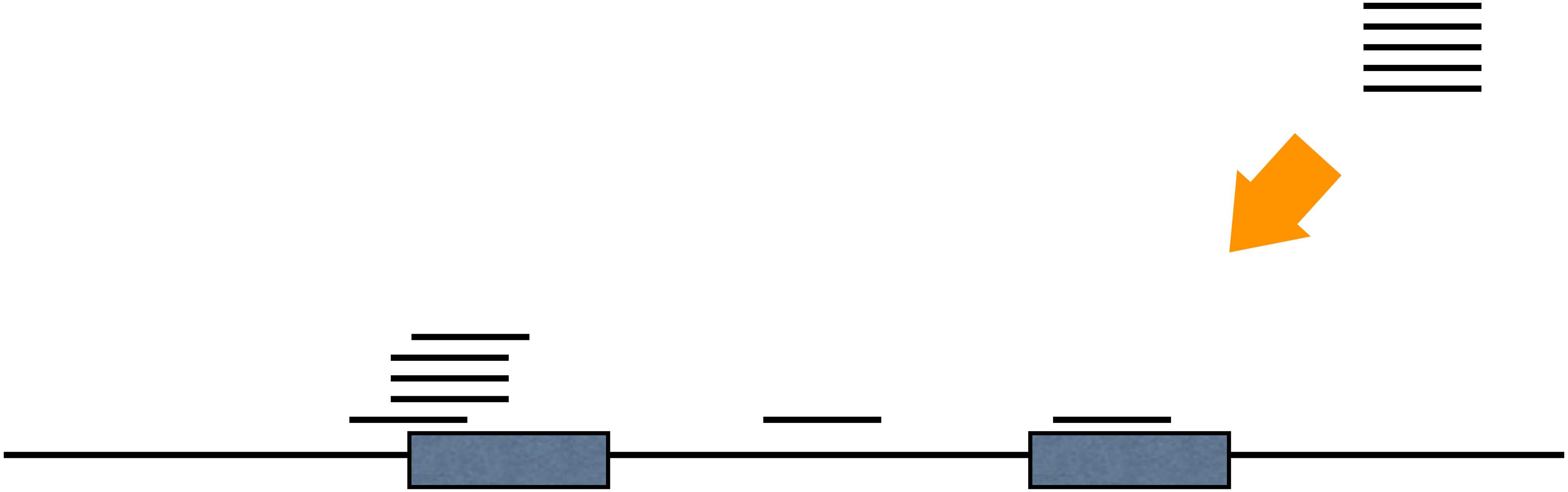


Table 1. The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

sequence quality	Pe (error probability)	1 - Pe	ascii code in SAM
40	1.00E-04	99.99%	I
30	1.00E-03	99.9%	?
20	1.00E-02	99%	5
10	1.00E-01	90%	+
0	1.00E+00	0%	!

Mapping



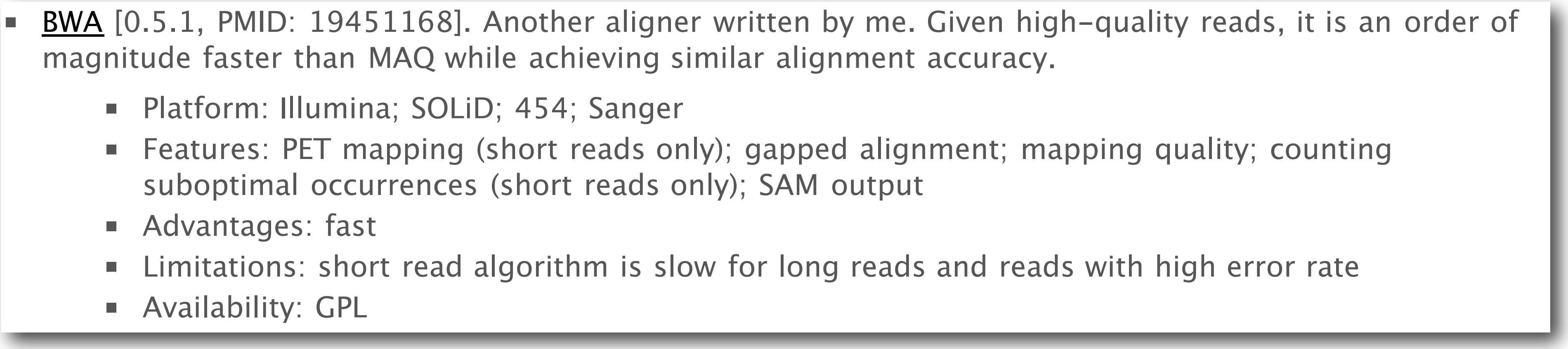
- Many aligners perform alignment of the reads to the reference genome
- Overview of NGS aligners by Heng Li:

  http://lh3lh3.users.sourceforge.net/NGSalign.shtml

- Alignment is not just genomic coordinates - results needs to be stored in a standard way

BWA

- Burrows-Wheeler Alignment Tool
- PMID: 19451168 (for short read), PMID: 20080505 (for long read)
- work fast reasonably, consider sequence/ mapping quality, and output the results in a standard format (SAM)



- BWA [0.5.1, PMID: 19451168]. Another aligner written by me. Given high-quality reads, it is an order of magnitude faster than MAQ while achieving similar alignment accuracy.
  - Platform: Illumina; SOLiD; 454; Sanger
  - Features: PET mapping (short reads only); gapped alignment; mapping quality; counting suboptimal occurrences (short reads only); SAM output
  - Advantages: fast
  - Limitations: short read algorithm is slow for long reads and reads with high error rate
  - Availability: GPL

Mapping quality

- PMID: 18714091

- The same scaling to base quality

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

- Pe = 1 - [Ps, correct mapping probability]

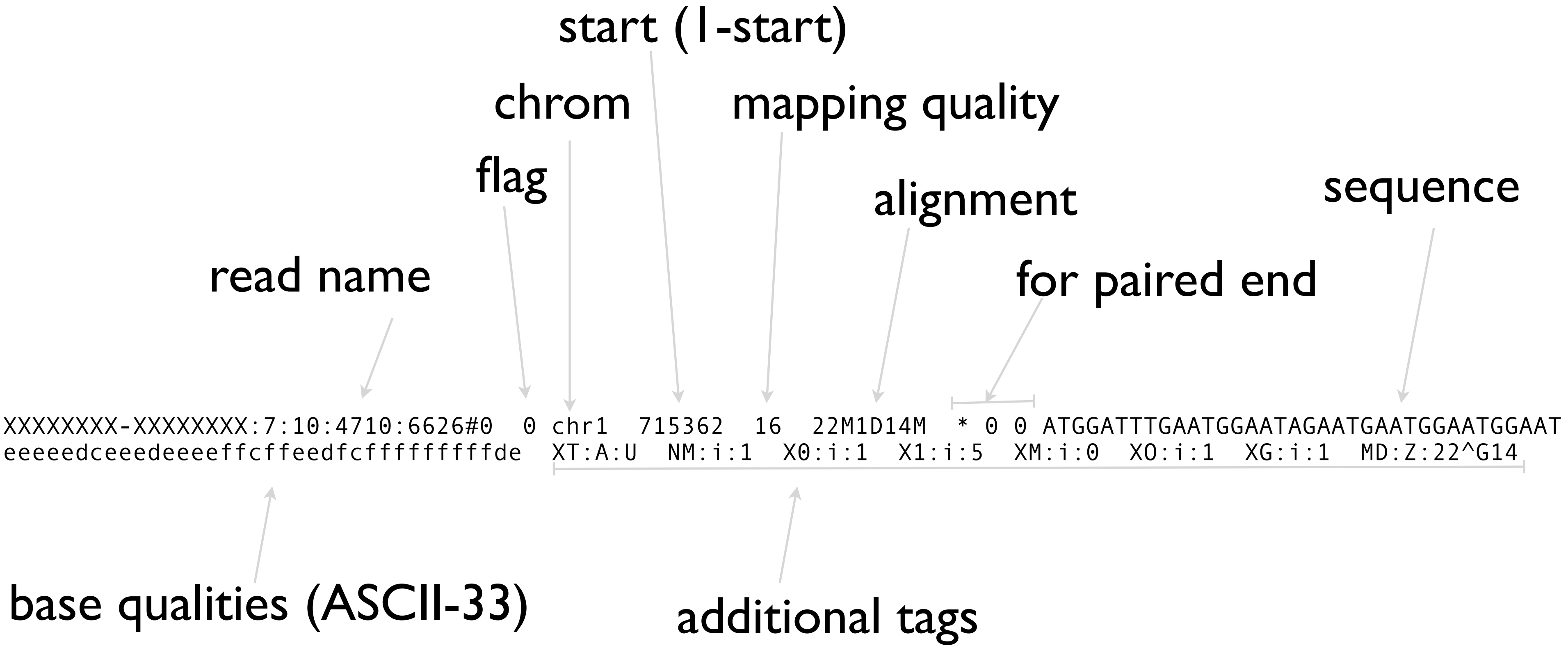
$$p_s(u|x,z) = \frac{p(z|x,u)}{\sum_{v=1}^{L-l+1} p(z|x,v)},$$

z: read
x: reference (genome)
u: position on the reference
P: probability that z arise from
the genomic coordinate x, u

SAM (Sequence Alignment/Map) format

- PMID: 19505943
- container of alignment (as well as sequence, sequence quality, and mapping quality)
- specification and utility (SAMtools) http://samtools.sourceforge.net

SAM example:



start (1-start)

chrom

mapping quality

flag

alignment

sequence

read name

for paired end

```
XXXXXXXX-XXXXXXXX:7:10:4710:6626#0  0 chr1  715362  16  22M1D14M  * 0 0 ATGGATTTGAATGGAATAGAATGAATGGAATGGAAT
eeeeedceeedeeeeffcffeedfcfffffffffde  XT:A:U  NM:i:1  X0:i:1  X1:i:5  XM:i:0  XO:i:1  XG:i:1  MD:Z:22^G14
```

base qualities (ASCII-33)

additional tags

BAM format

- compressed version of SAM file
- fast access to alignment when indexed
- SAMtools provide native support

Align FASTQ file with BWA

```
bwa aln ${genome} ${fastq} \
| bwa samse ${genome} - ${fastq}
| samtools view -bT ${genome} - > ${outfile}
```

Sort and index BAM

```
samtools sort ${outfile} ${outfile}
mv -f ${outfile}.bam ${outfile}
samtools index ${outfile}
```

An example of analysis flow and tools

- Mapping to the reference genome

  BWA, SAMtools

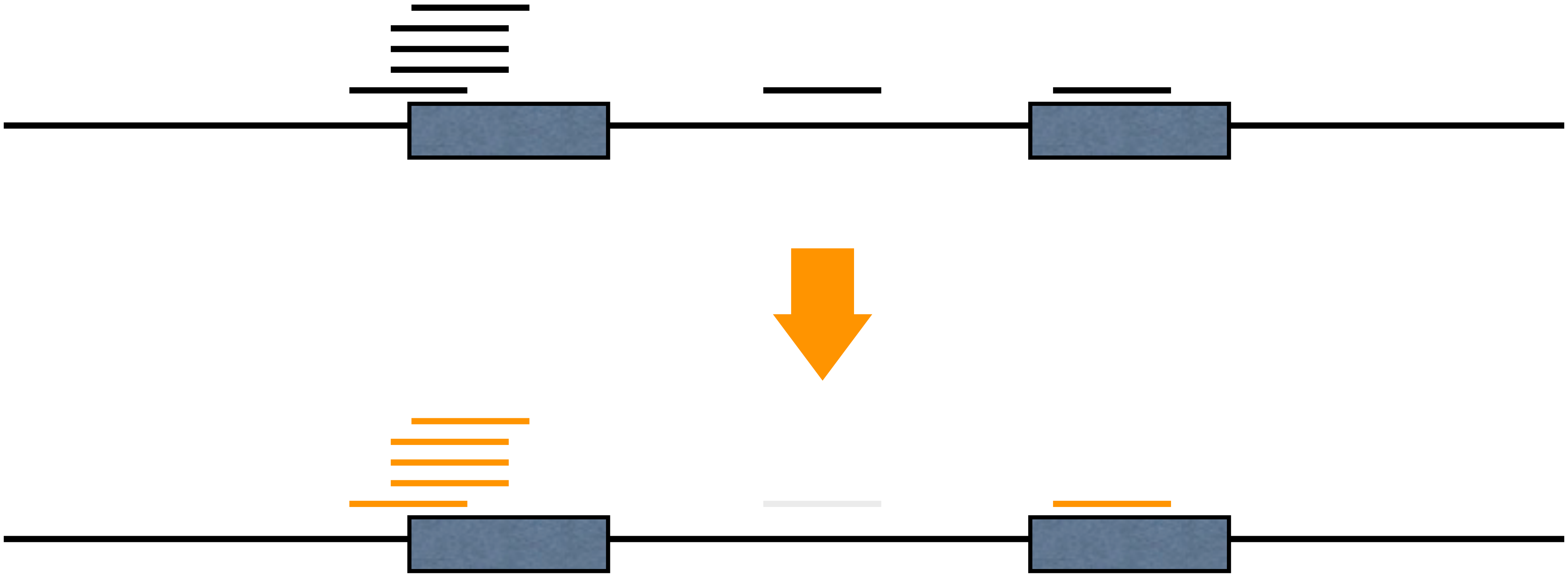
- Work on the genomic coordinates

  SAMtools, BEDTools, UCSC Tools

- Expression analysis / peak detection

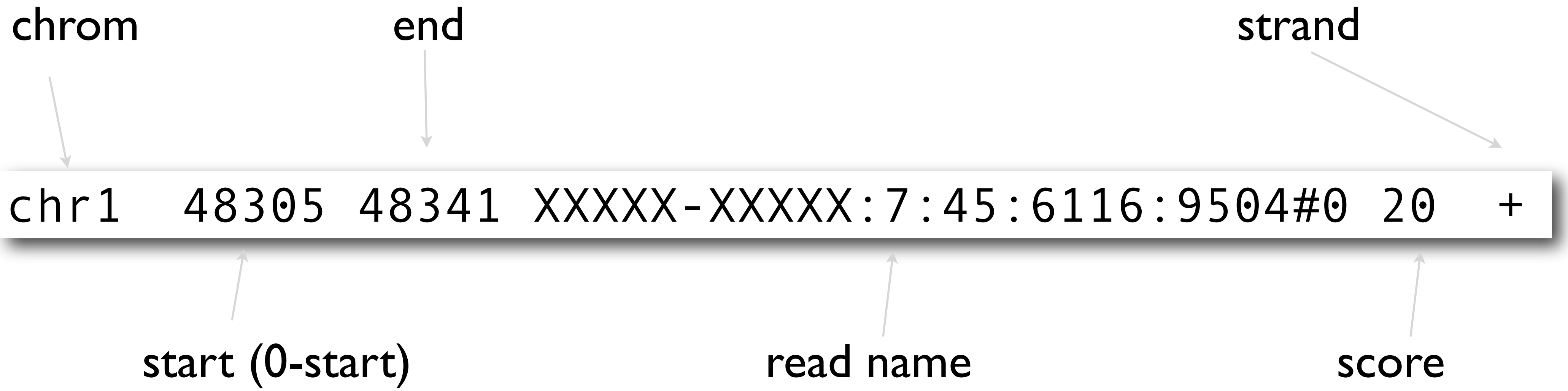
  edgeR / MACS

Select/count alignments



BED (Browser Extensible Data) format

- http://genome.ucsc.edu/FAQ/FAQformat.html



chrom

end

strand

```
chr1  48305 48341 XXXXX-XXXXX:7:45:6116:9504#0 20  +
```

start (0-start)

read name

score

BEDtools

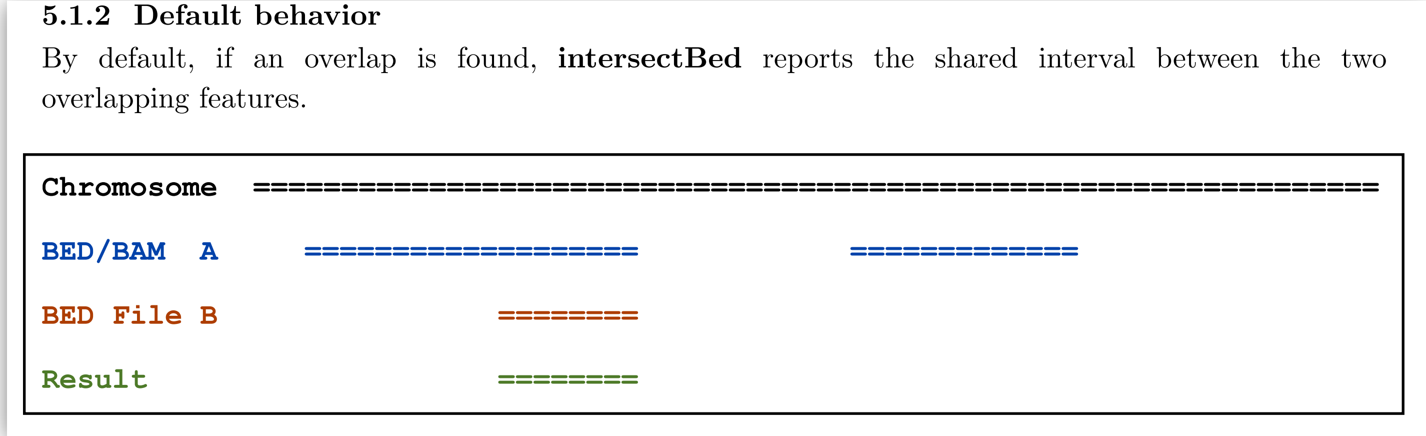
- PMID: 20110278
- A set of tools, which enables us a wide range of operation on the genomic coordinates.
- Well documented

example.

```
bamToBed -i ${bamfile} > ${bedfile}
```

```
bamToBed -i ${bamfile} \
| intersectBed -a stdin -b genes.bed > ${bedfile}
```

intersectBed



5.1.2 Default behavior

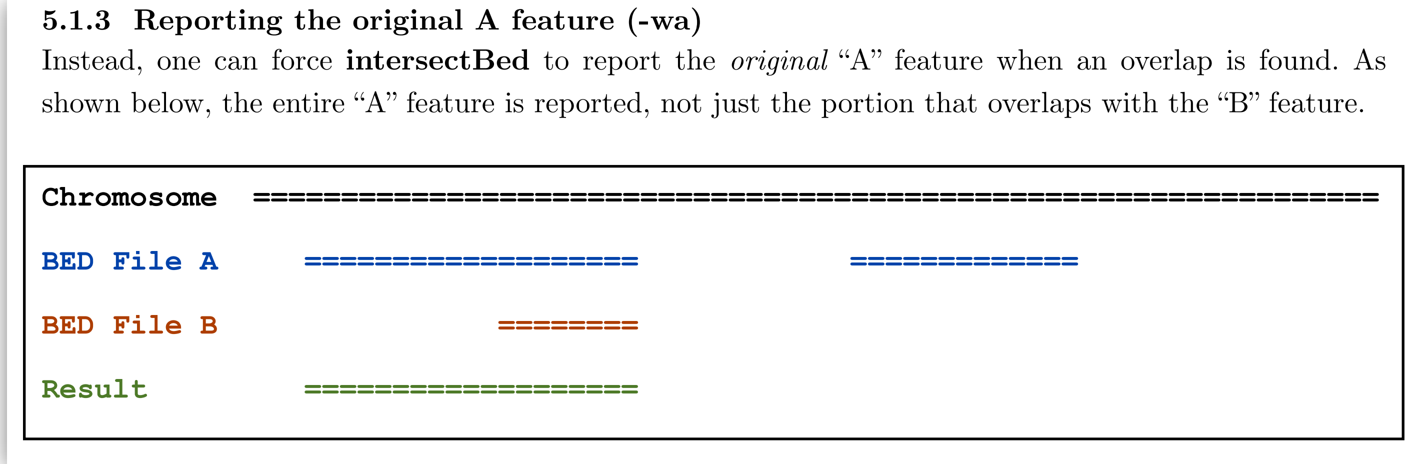
By default, if an overlap is found, **intersectBed** reports the shared interval between the two overlapping features.

```
Chromosome  ================================================================

BED/BAM  A     ===================            =============

BED File B                ========

Result                    ========
```



5.1.3 Reporting the original A feature (-wa)

Instead, one can force **intersectBed** to report the *original* "A" feature when an overlap is found. As shown below, the entire "A" feature is reported, not just the portion that overlaps with the "B" feature.

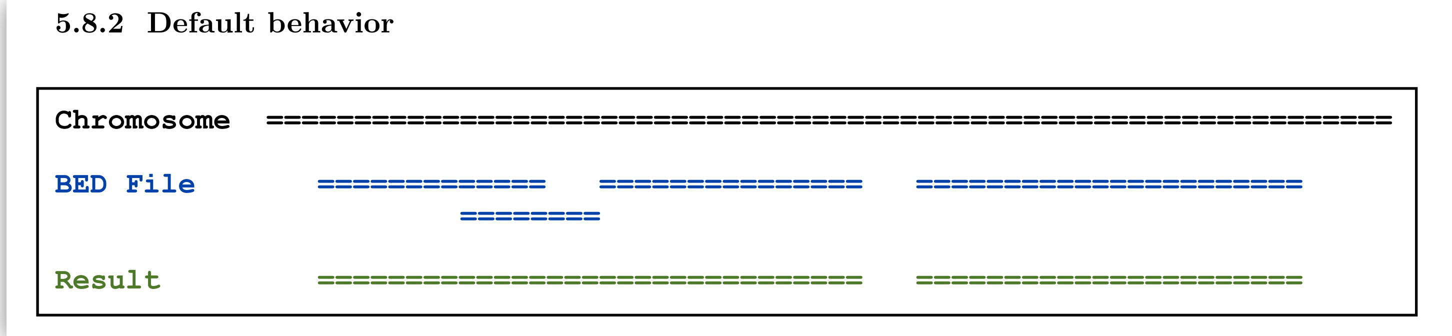
```
Chromosome  ================================================================

BED File A     ===================            =============

BED File B                ========

Result         ===================
```

mergeBed



5.8.2 Default behavior

```
Chromosome  ================================================================

BED File       =============   ===============   ======================
                       ========

Result         ===============================   ======================
```

from bedtools manual

Jim kent source tree

- http://genomewiki.ucsc.edu/index.php/Genome_Browser_Software_Features
- http://genome.ucsc.edu/admin/git.html
- A huge source tree including UCSC Genome Browser, BLAT, etc.
- Also includes utilities to get annotation and create custom tracks

```
genePredToGtf -utr ${DB} refGene /dev/stdout \
| grep --perl-regexp "\texon\t"
```

Filtering alignment with mapping quality

by samtools

```
samtools view -bq 10 ${bamfile} > ${result_bam}
```

Discard redundant reads (for single-end)

by samtools

```
samtools rmdup -s ${bamfile} ${result_bam}
```

Convert BAM file to BED

by bedtools

```
bamToBed -i ${bamfile} > ${bedfile}
```

Obtain refseq transcript coordinates

by jim kent source tree

```
genePredToFakePsl hg18 refGene /dev/stdout t.cds\
| pslToBed /dev/stdin /dev/stdout > refgene.bed
```

Obtain refseq TSS proximal regions

by jim kent source tree

```
genePredToFakePsl hg18 refGene /dev/stdout t.cds\
| pslToBed /dev/stdin /dev/stdout \
| awk '
  BEGIN{OFS="\t"}
  {
    if ($6 == "+"){$3 = $2+1}
    if ($6 == "-"){$2 = $3-1}
    print $1,$2-500,$3+500,$4,$5,$6
  }
'
```

Select the reads within the region of interests

by bedtools

```
bamToBed -i ${bamfile} \
| intersectBed -s -wa -a stdin -b ann.bed
```

Count the reads within the region of interests

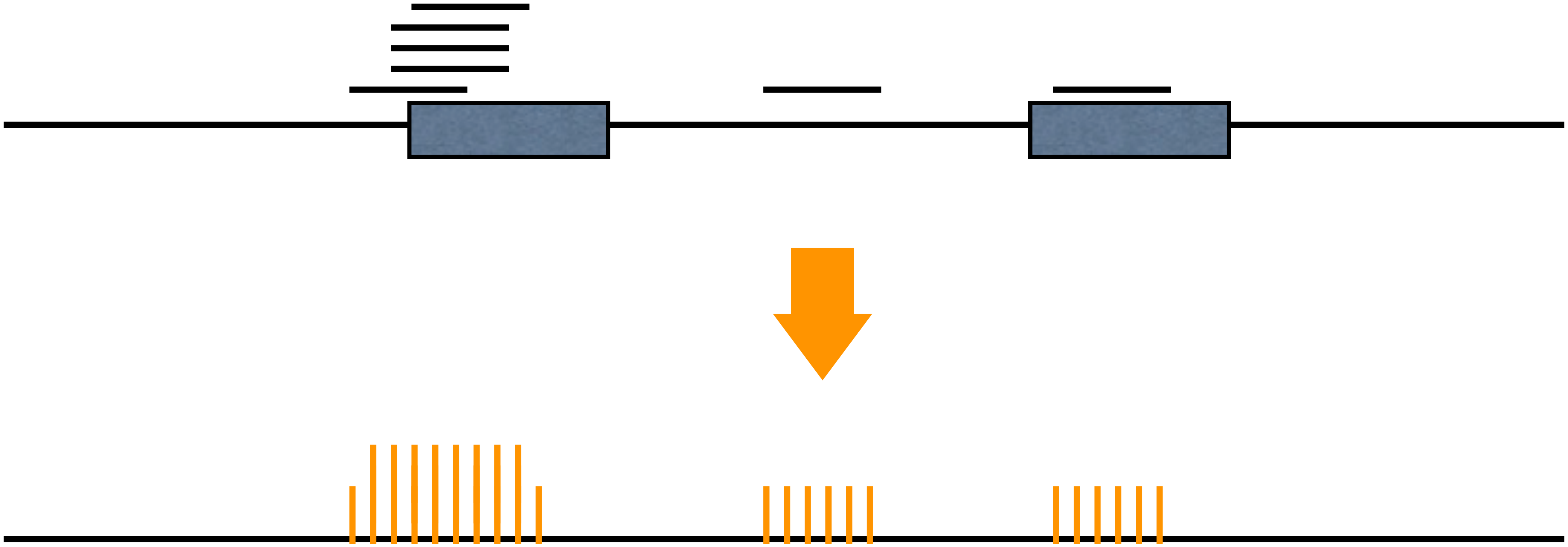
by bedtools

```
bamToBed -i ${bamfile} \
| intersectBed -s -c -a stdin -b ann.bed
```

BedGraph (Wiggle) file for genome browser

by bedtools

```
bamToBed -i ${bamfile} \
| genomeCoverageBed -bg -i stdin -g hg18.genome
```



An example of analysis flow and tools

- Mapping to the reference genome

  BWA, SAMtools

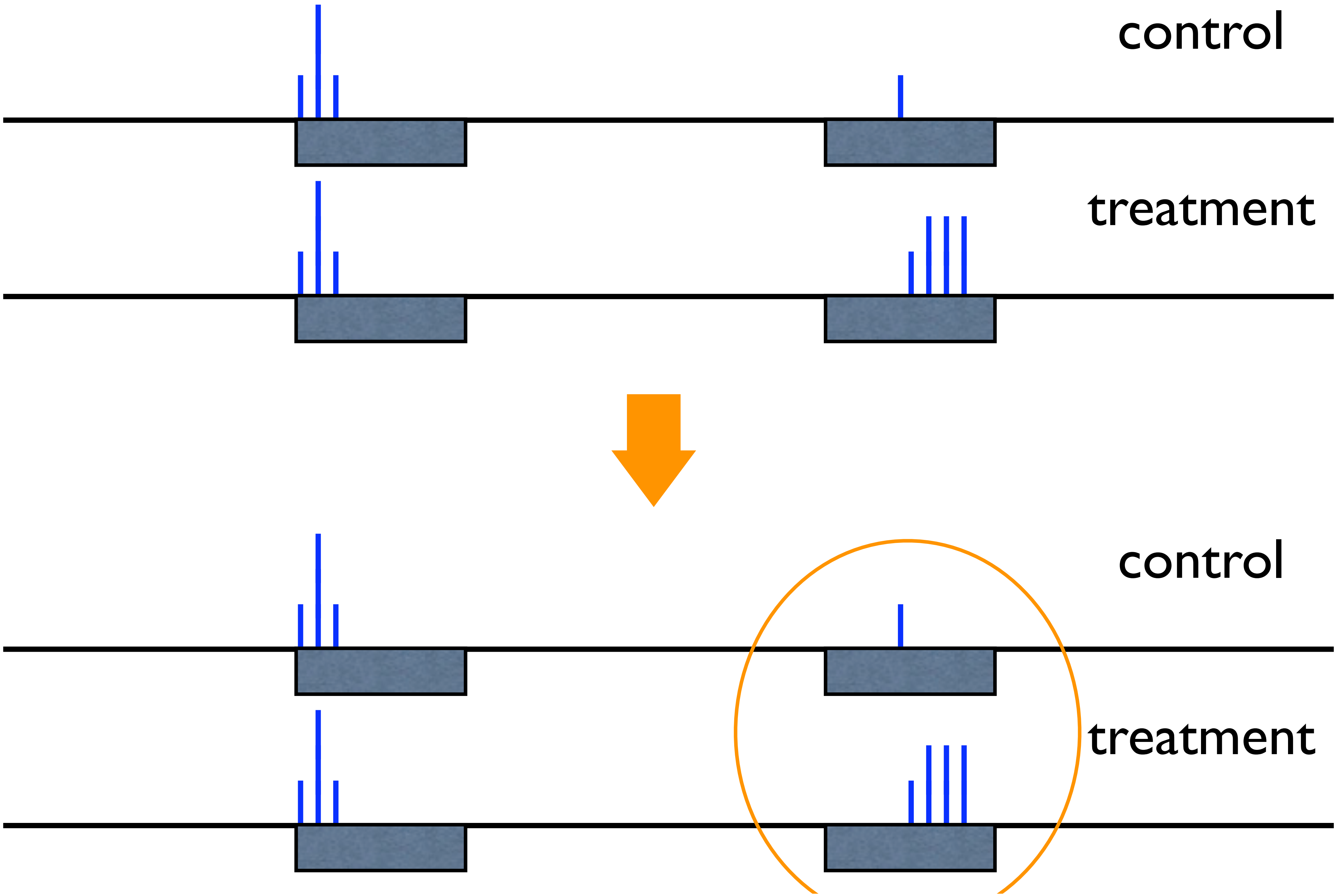
- Work on the genomic coordinates

  SAMtools, BEDTools, UCSC Tools

- Expression analysis / peak detection

  edgeR / MACS

Find differentially expressed regions



Negative Binomial Distribution

- a.k.a Gamma-Poisson mixture
- Theoretical random sampling should follow Poisson distribution
- Variance between replicates are modeled in Gamma distribution (over dispersion)

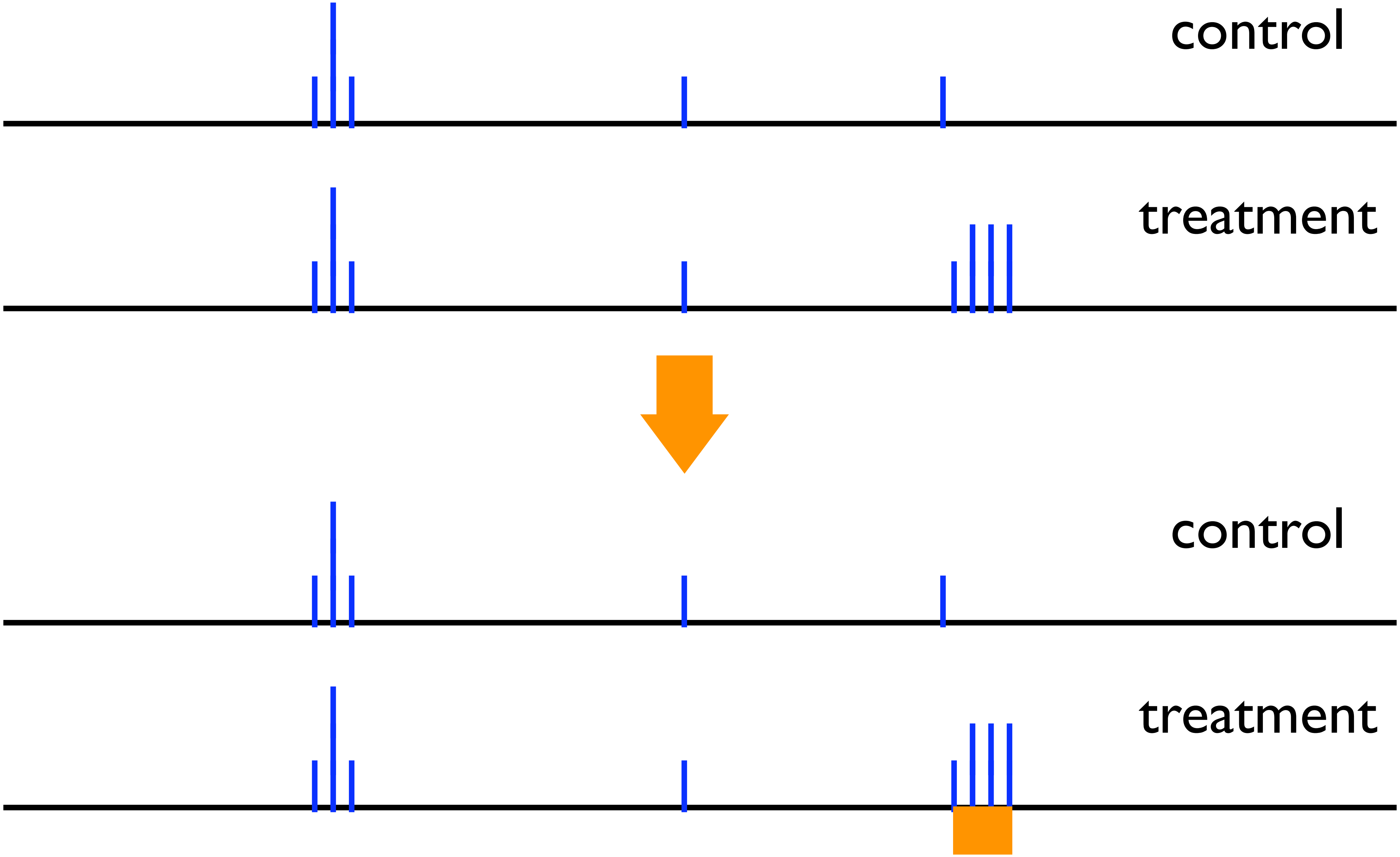
gene	ctl1	ctl2	ctl3	kd1	kd2	kd3
A	8	3	2	7	5	9
B	129	50	78	143	152	99
C	523	670	428	18	23	8
...						

edgeR (in R/bioconductor)

- PMID: 19910308
- Estimate over-dispersion of negative binomial model
- simple differential analysis

```
> library(edgeR)
> counts <- read.table(count_file)
> dge <- DGEList(
  counts = counts,
  group = c(“CTL”,”CTL”,”CTL”,”KD”,”KD”,”KD”)
)
> dge <- estimateCommonDisp(dge)
> de <- exactTest(dge)
```

Find significant peaks



MACS (A peak caller)

- PMID: 18798982
- Take a control experiment (genomic input or nonspecific antibody) into consideration

```
macs -t ChIP.bam -c Control.bam --format=BAM
```

Refer original papers/documents!

- BWA - PMID: 19451168

  Bioinformatics. 2009 Jul 15;25(14):1754-60.
  Fast and accurate short read alignment with Burrows-Wheeler transform.
  Li H, Durbin R.

- SAMtools - PMID:19505943

  Bioinformatics. 2009 Aug 15;25(16):2078-9.
  The Sequence Alignment/Map format and SAMtools.
  Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup.

- BEDtools - PMID: 20110278

  Bioinformatics. 2010 Mar 15;26(6):841-2.
  BEDTools: a flexible suite of utilities for comparing genomic features.
  Quinlan AR, Hall IM.

- Jim Kent Source Tree

  http://genome.ucsc.edu/admin/git.html
  http://genomewiki.ucsc.edu/index.php/Genome_Browser_Software_Features

- edgeR - PMID: 19910308

  Bioinformatics. 2010 Jan 1;26(1):139-40.
  edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.
  Robinson MD, McCarthy DJ, Smyth GK.

- MACS - PMID: 18798982

  Genome Biol. 2008;9(9):R137.
  Model-based analysis of ChIP-Seq (MACS).
  Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS.