Genomic tools

Libor Mořkovský, Václav Janoušek

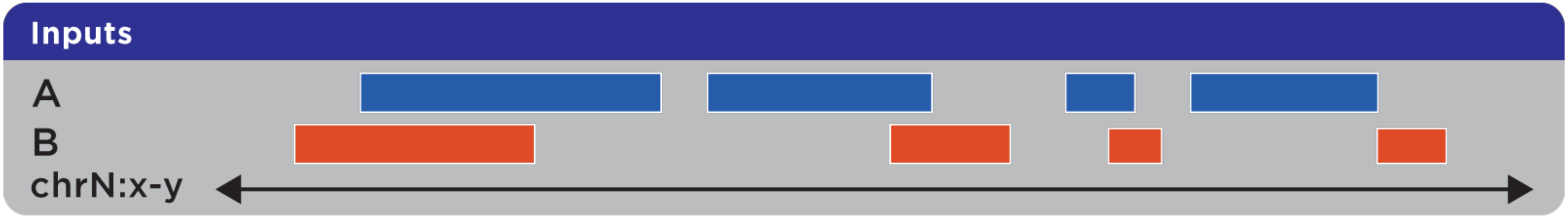
https://ngs-course.readthedocs.io/en/praha-february-2019/

Genome arithmetics: bedtools/bedops

- Operations with genomic data based on their physical position in genome
- Variables:
  - chromosome
  - feature start, feature end
  - id
  - strand
- Basic data format: BED

Genome arithmetics: Examples

- Two sets of features (BED files):



http://bedops.readthedocs.org

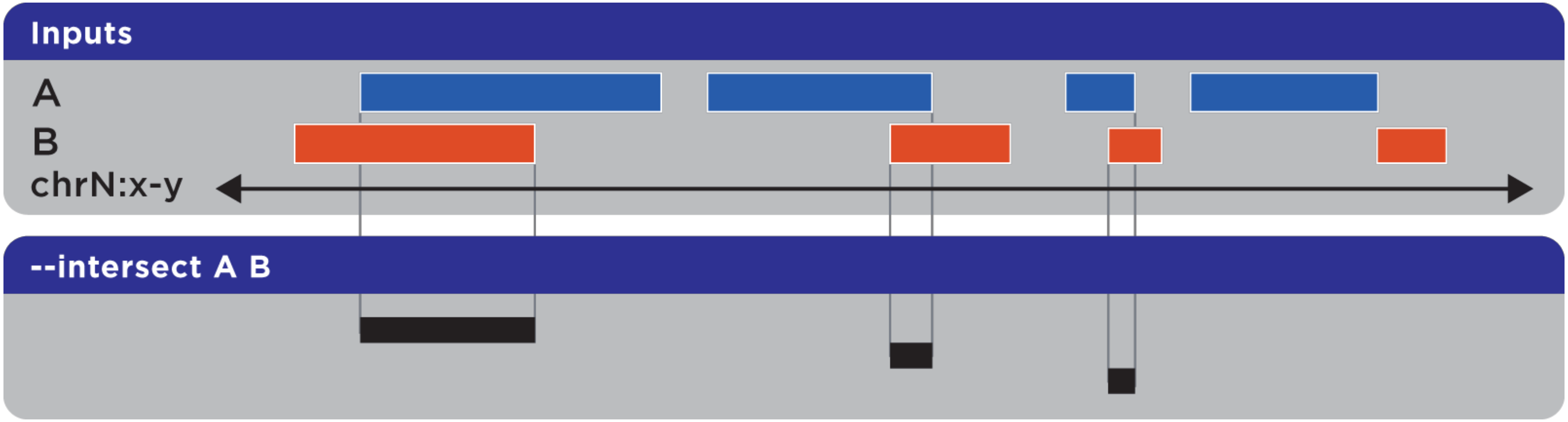
```
chr1    1000    1200
chr1    1700    2100
chr2    1100    1500
```

```
chr1    700     1100
chr1    1400     1500
chr1    1600     1900
```

New set of features based on combination of the previous sets using a specific rule

Genome arithmetics: Examples

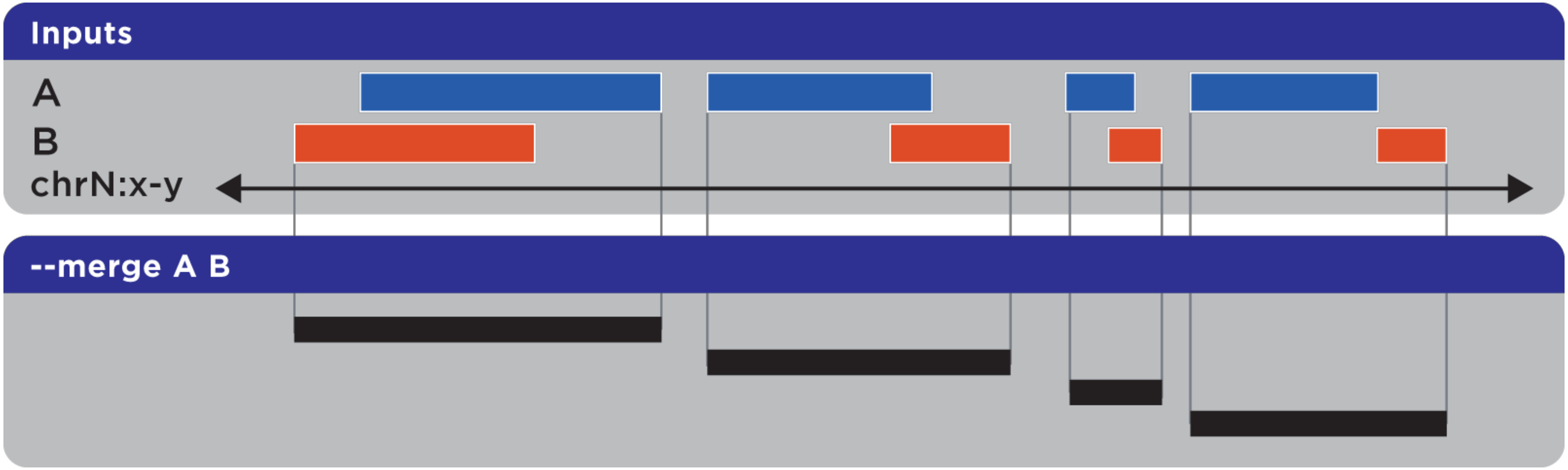
- The rule: Get parts of features that overlap



http://bedops.readthedocs.org

Genome arithmetics: Examples

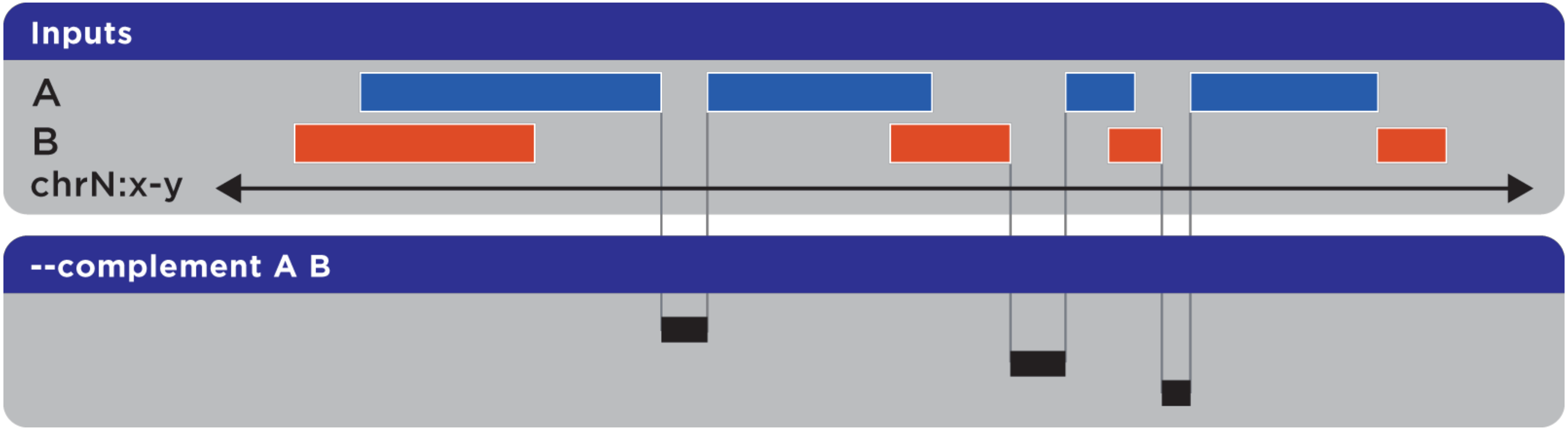
- The rule: Merge entire features



http://bedops.readthedocs.org

Genome arithmetics: Examples

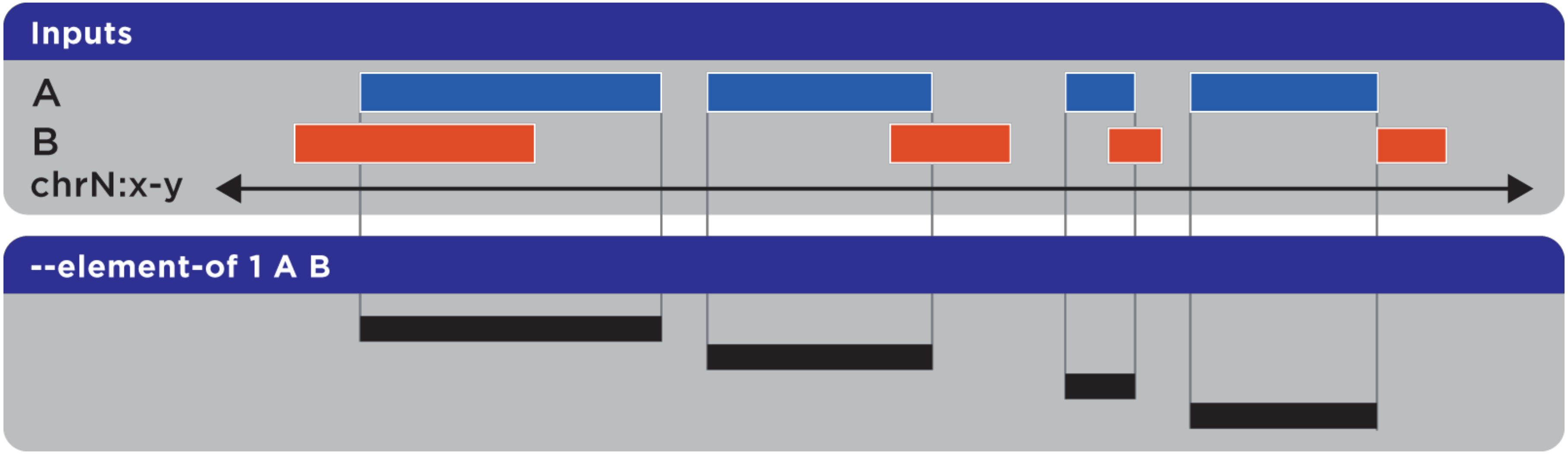
- The rule: Get complement features



http://bedops.readthedocs.org

Genome arithmetics: Examples

- The rule: Report A which overlaps B



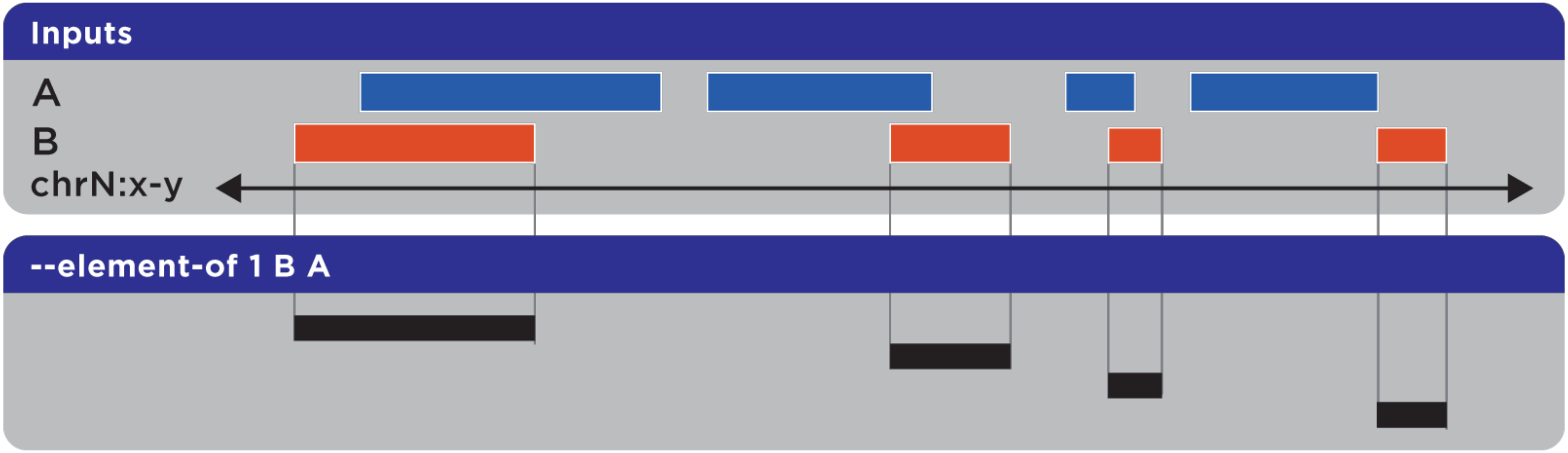
http://bedops.readthedocs.org

bedtools

BEDOPS

Genome arithmetics: Examples

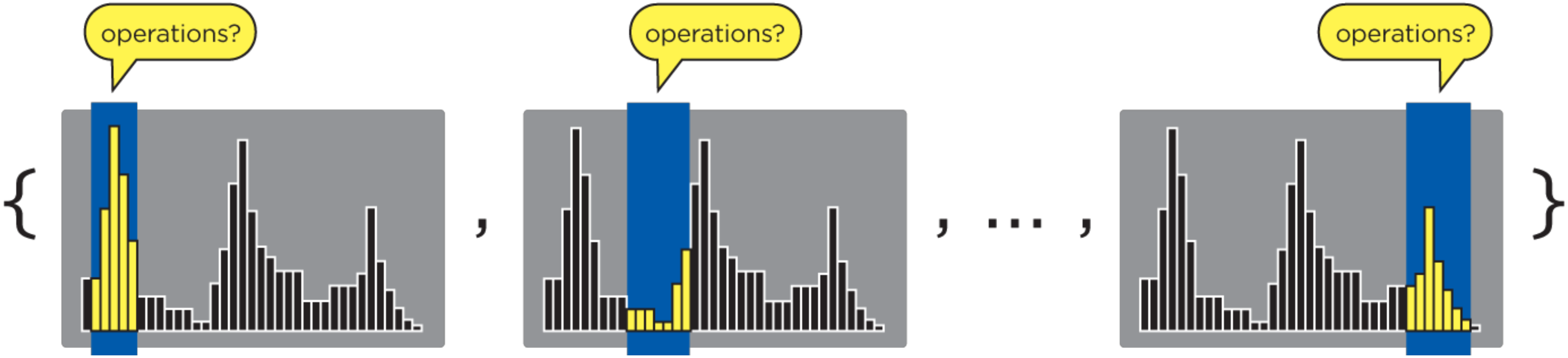
- The rule: Report B which overlaps A



http://bedops.readthedocs.org

Genome feature summary

- Statistics, summary
- bedmap, bedtools (coverageBed, groupBy)
- e.g. depth coverage, base pair coverage, etc.



Other tools in bedtools

- makewindows
- cluster
- shuffle
- random
- … *explore the bedtools website for further tools and the documentation*

Exercise

1. Count the number of open chromatin regions overlapping with genes or are within 1000 bp window on each side of a gene

(Use Ensembl.NCBIM37.67.bed and encode.bed files)

Exercise

2. Make three sets of sliding windows across mouse genome (1 Mb, 5 Mb) with the step size 0.2 by the size of the window and obtain gene density within these sliding windows.

Variation data: vcftools

- Efficient working with VCF data
- Quality control
- Basic evolutionary genetics measures/statistics
  - transition/transversion
  - heterozygosity, relatedness
  - Hardy-Weinberg
  - Weir & Cockerham's Fst
  - Nucleotide diversity
  - Linkage Disequilibrium

vcftools: starting

- Opening and viewing a vcf file:

```
vcftools \
--gzvcf /data-shared/mus_mda/00-popdata/popdata_mda.vcf.gz \
--recode --stdout | less -S
```

- Creating a new vcf file:

```
vcftools \
--gzvcf /data-shared/mus_mda/00-popdata/popdata_mda.vcf.gz \
--recode --out new_vcf
```

vcftools: data filtering

- Sample/Variant retrieval by name:
  - Individual/Variant names to keep/remove have to be specified in a separate file

```
--keep ind.txt # Keep these individuals
--remove ind.txt # Remove these individuals
--snps snps.txt # Keep these SNPs
--snps snps.txt --exclude # Remove these SNPs
```

```
vcftools \
--gzvcf /data-shared/mus_mda/00-popdata/popdata_mda.vcf.gz \
--keep /data-shared/mus_mda/00-popdata/euro_samples.txt \
--recode --stdout |
less -S
```

vcftools: data filtering

- Variant filtering based on physical location

```
--chr 11 # Keep just this chromosome
--not-chr 11 # Remove this chromosome
--not-chr 11 -not-chr 2 # Remove these two chromosomes
--from-bp 20000000 # Keep SNPs from this position
--to-bp 22000000 # Keep SNPs to this position
--bed keep.bed # Keep only SNPs overlapping with locations
listed in a file
--exclude-bed remove.bed # The opposite of the previous
```

```
vcftools \
--gzvcf /data-shared/mus_mda/00-popdata/popdata_mda.vcf.gz \
--keep /data-shared/mus_mda/00-popdata/euro_samples.txt \
--chr 11 \
--from-bp 22000000 \
--to-bp 23000000 \
--recode \
--stdout |
less -S
```

vcftools: data filtering

- Variant filtering based on other features

```
--maf 0.2 # Keep just variants with Minor Allele Freq higher
than 0.2
--hwe 0.05 # Keep just variants which do not deviate from HW
equilibrium (p-value = 0.05)
--max-missing (0-1) # Remove SNPs with given proportion of
missing data (0 = allowed completely missing, 1 = no missing
data allowed)
--minQ 20 # Minimal quality allowed (Phred score)
```

```
vcftools \
--gzvcf /data-shared/mus_mda/00-popdata/popdata_mda.vcf.gz \
--keep /data-shared/mus_mda/00-popdata/euro_samples.txt \
--recode --stdout |
vcftools --vcf - \ ← stdin
--max-missing 1 \
--maf 0.2 \
--recode --stdout | less -S
```

vcftools: summary/statistics

- molecular evolution/population genetic

```
--site-pi # Calculates per-site nucleotide diversity (π)
--window-pi 1000000 --window-pi-step 250000 # Calculates per-
site nucleotide diversity for windows of 1Mb with 250Kb step
--weir-fst-pop pop1.txt --weir-fst-pop pop2.txt # Calculates
Weir & Cockerham's Fst
--fst-window-size 1000000 --fst-window-step 250000 #
Calculates Fst for windows of 1Mb with 250Kb step
```

```
vcftools \
--vcf /data-shared/mus_mda/00-popdata/popdata_mda_euro.vcf \
--weir-fst-pop /data-shared/mus_mda/00-popdata/musculus_samps.txt \
--weir-fst-pop /data-shared/mus_mda/00-popdata/domesticus_samps.txt \
--stdout |
less -S
```