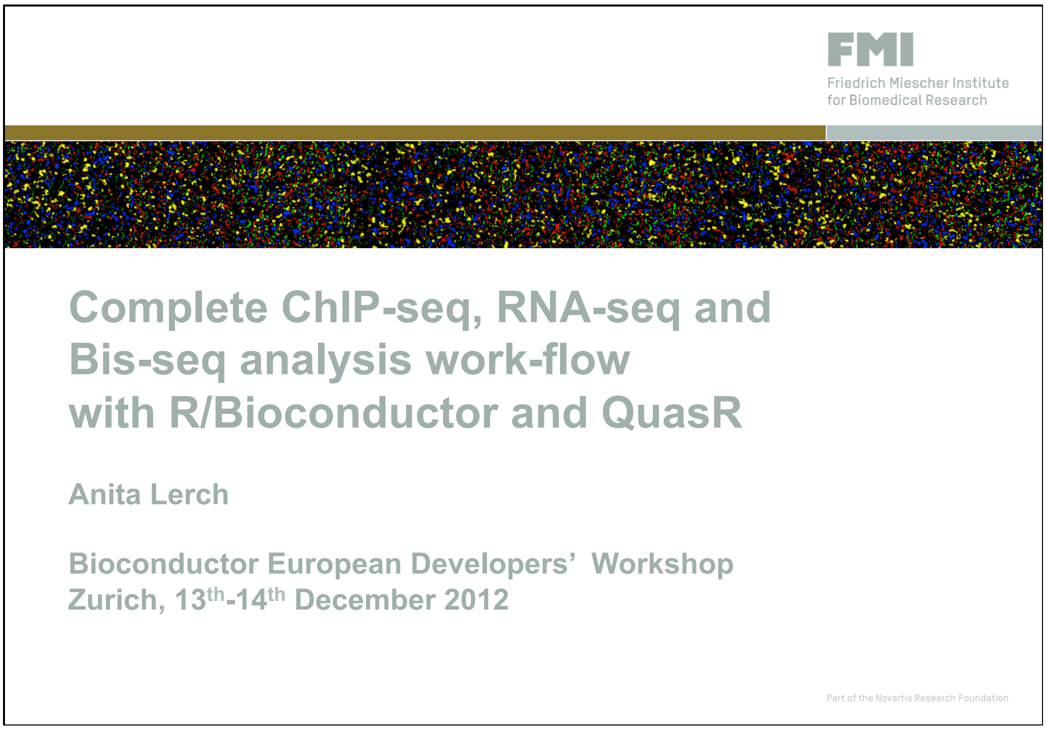


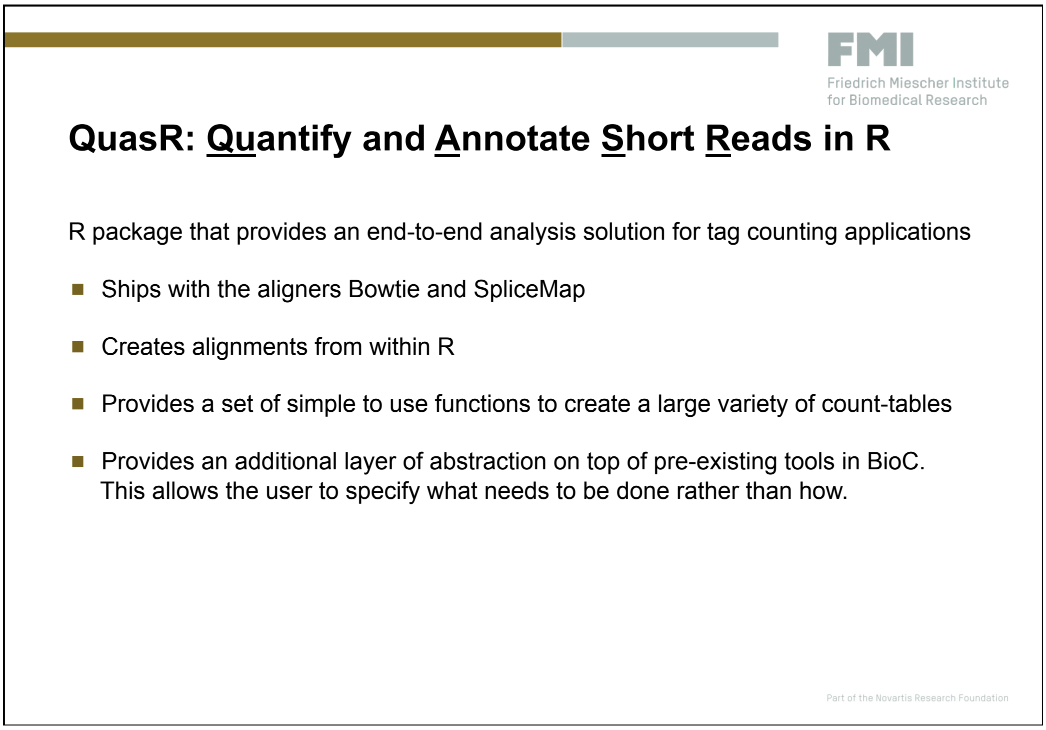
FMI
Friedrich Miescher Institute
for Biomedical Research

Complete ChIP-seq, RNA-seq and Bis-seq analysis work-flow with R/Bioconductor and QuasR

Anita Lerch

Bioconductor European Developers' Workshop
Zurich, 13th-14th December 2012

Part of the Novartis Research Foundation



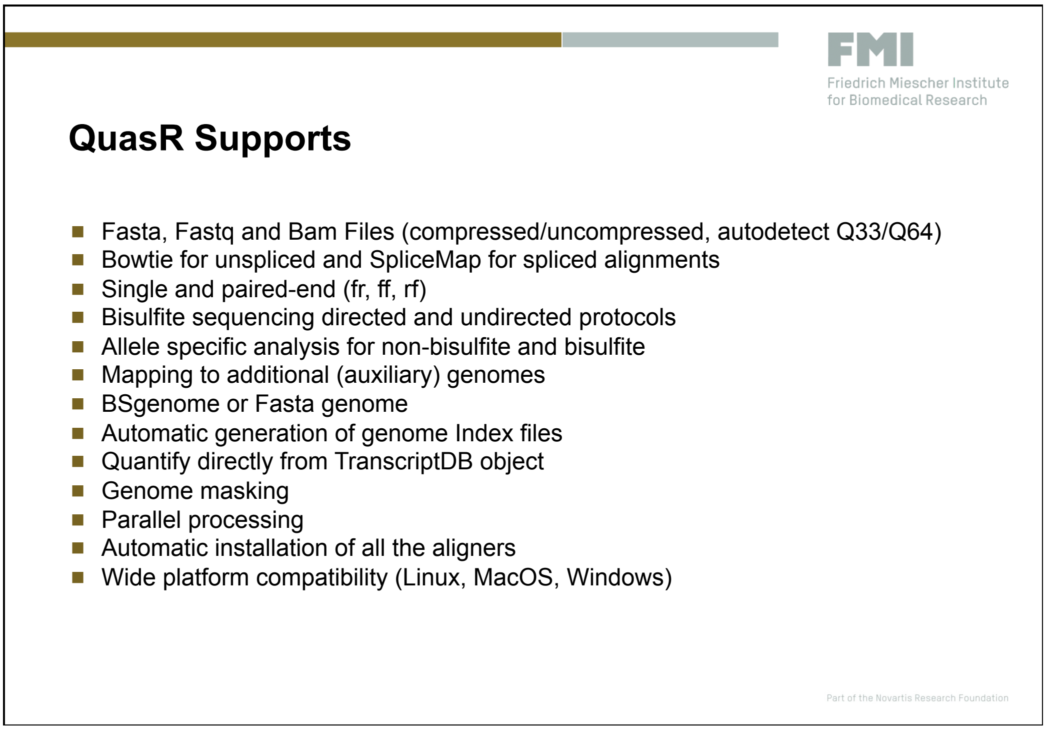
FMI
Friedrich Miescher Institute
for Biomedical Research

QuasR: Quantify and Annotate Short Reads in R

R package that provides an end-to-end analysis solution for tag counting applications

- Ships with the aligners Bowtie and SpliceMap
- Creates alignments from within R
- Provides a set of simple to use functions to create a large variety of count-tables
- Provides an additional layer of abstraction on top of pre-existing tools in BioC. This allows the user to specify what needs to be done rather than how.

Part of the Novartis Research Foundation

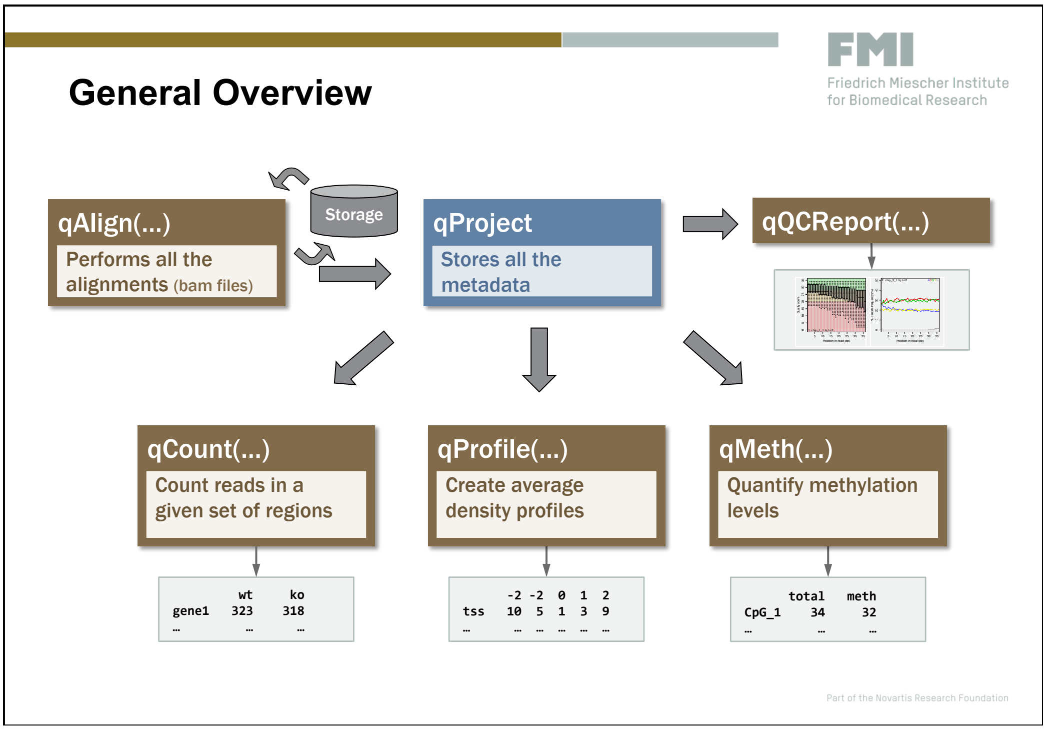


FMI
Friedrich Miescher Institute
for Biomedical Research

QuasR Supports

- Fasta, Fastq and Bam Files (compressed/uncompressed, autodetect Q33/Q64)
- Bowtie for unspliced and SpliceMap for spliced alignments
- Single and paired-end (fr, ff, rf)
- Bisulfite sequencing directed and undirected protocols
- Allele specific analysis for non-bisulfite and bisulfite
- Mapping to additional (auxiliary) genomes
- BSgenome or Fasta genome
- Automatic generation of genome Index files
- Quantify directly from TranscriptDB object
- Genome masking
- Parallel processing
- Automatic installation of all the aligners
- Wide platform compatibility (Linux, MacOS, Windows)

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

General Overview

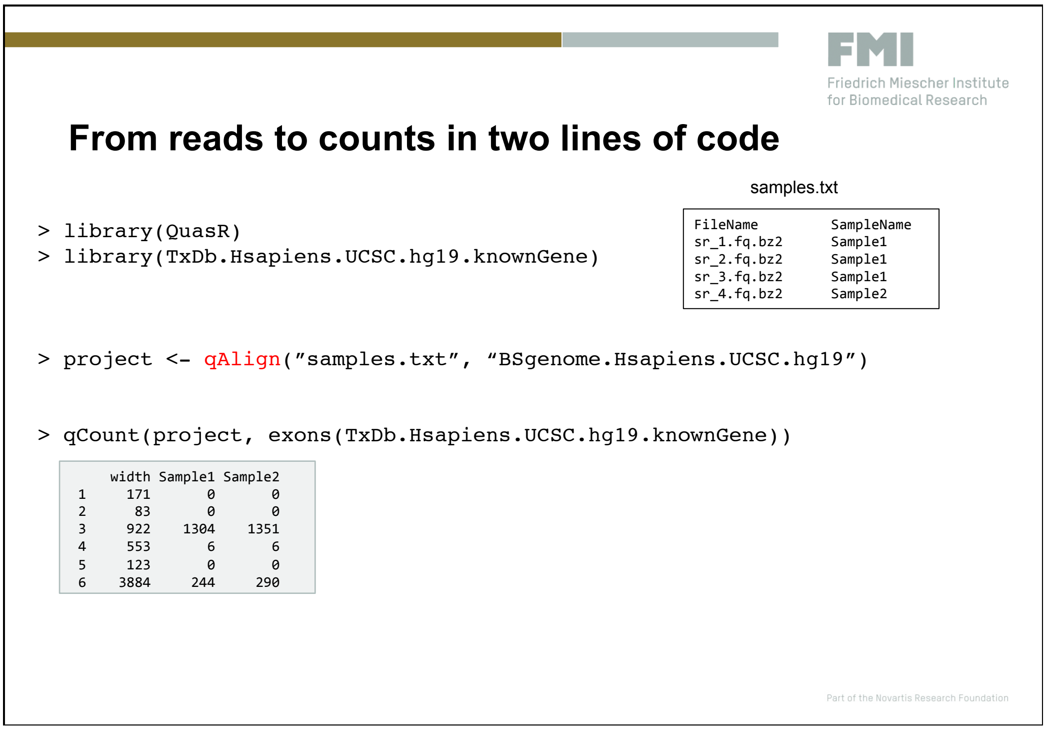
qAlign(...)
Performs all the alignments (bam files)
Storage
qProject
Stores all the metadata
qQCReport(...)
qCount(...)
Count reads in a given set of regions
qProfile(...)
Create average density profiles
qMeth(...)
Quantify methylation levels

```
        wt    ko
gene1  323   318
…        …     …
```

```
     -2 -2  0  1  2
tss  10  5  1  3  9
…     …  …  …  …  …
```

```
       total  meth
CpG_1    34    32
…         …     …
```

Part of the Novartis Research Foundation



Friedrich Miescher Institute
for Biomedical Research

From reads to counts in two lines of code

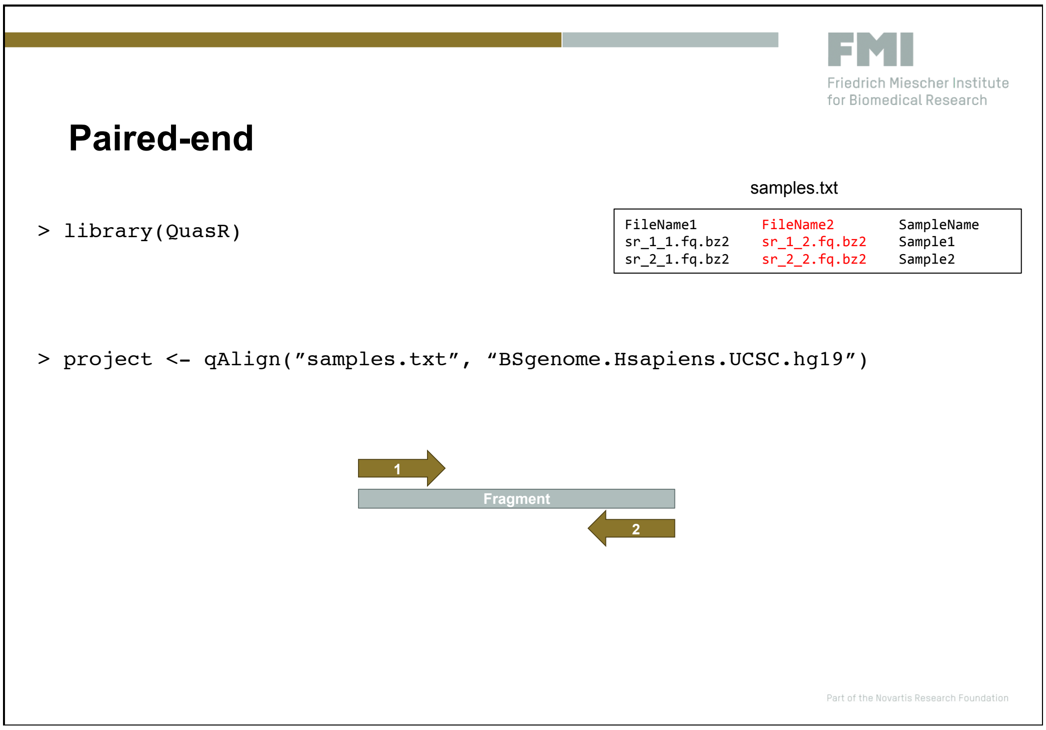
samples.txt

```
FileName        SampleName
sr_1.fq.bz2     Sample1
sr_2.fq.bz2     Sample1
sr_3.fq.bz2     Sample1
sr_4.fq.bz2     Sample2
```

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)

> qCount(project, exons(TxDb.Hsapiens.UCSC.hg19.knownGene))
```

```
   width Sample1 Sample2
1    171       0       0
2     83       0       0
3    922    1304    1351
4    553       6       6
5    123       0       0
6   3884     244     290
```

Part of the Novartis Research Foundation



Friedrich Miescher Institute
for Biomedical Research

Paired-end

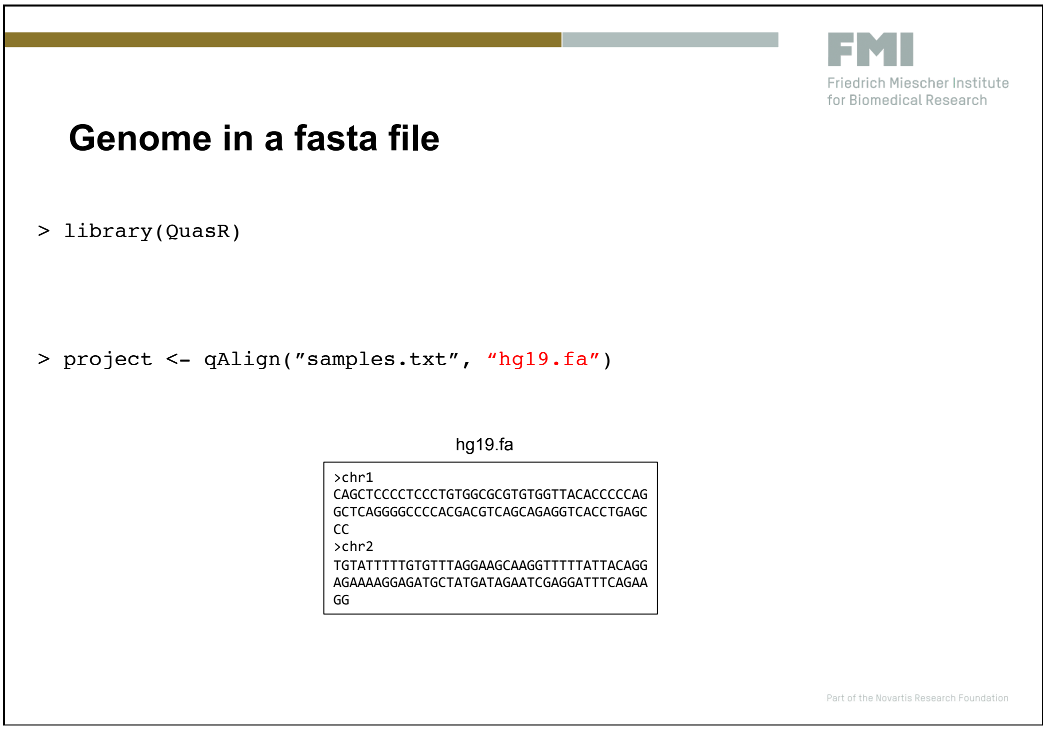
samples.txt

```
FileName1       FileName2       SampleName
sr_1_1.fq.bz2   sr_1_2.fq.bz2   Sample1
sr_2_1.fq.bz2   sr_2_2.fq.bz2   Sample2
```

```
> library(QuasR)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)
```

1
Fragment
2

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

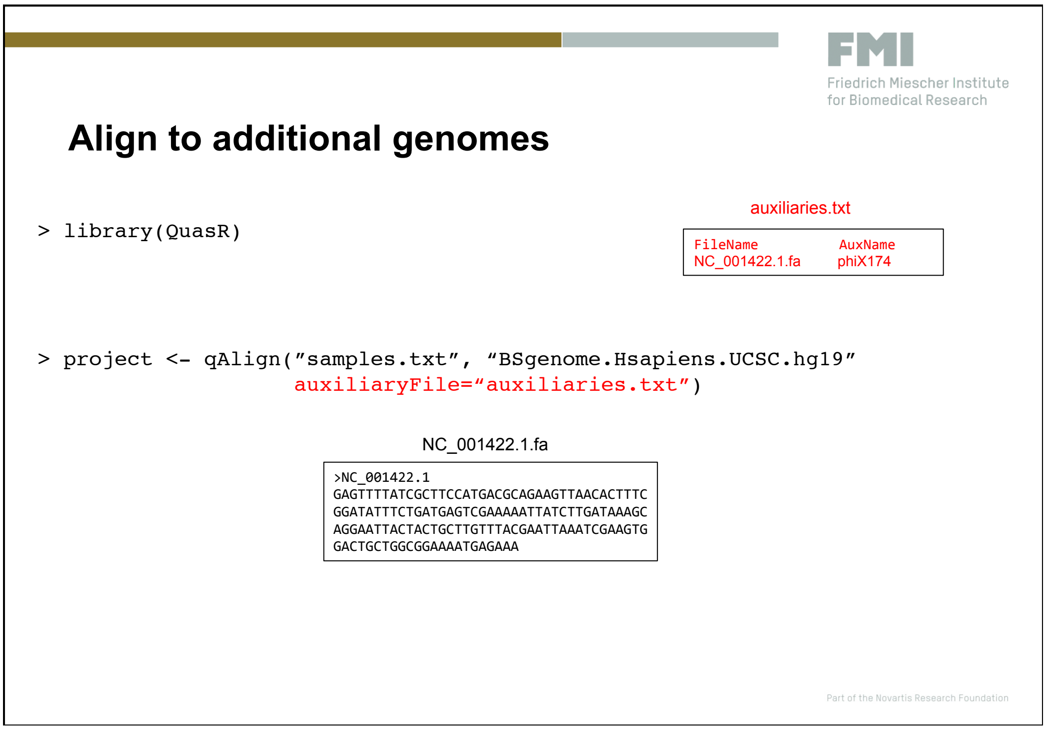
Genome in a fasta file

```
> library(QuasR)

> project <- qAlign(”samples.txt”, “hg19.fa”)
```

hg19.fa

```
>chr1
CAGCTCCCCTCCCTGTGGCGCGTGTGGTTACACCCCCAG
GCTCAGGGGCCCCACGACGTCAGCAGAGGTCACCTGAGC
CC
>chr2
TGTATTTTTGTGTTTAGGAAGCAAGGTTTTTATTACAGG
AGAAAAGGAGATGCTATGATAGAATCGAGGATTTCAGAA
GG
```

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

Align to additional genomes

```
> library(QuasR)
```

auxiliaries.txt

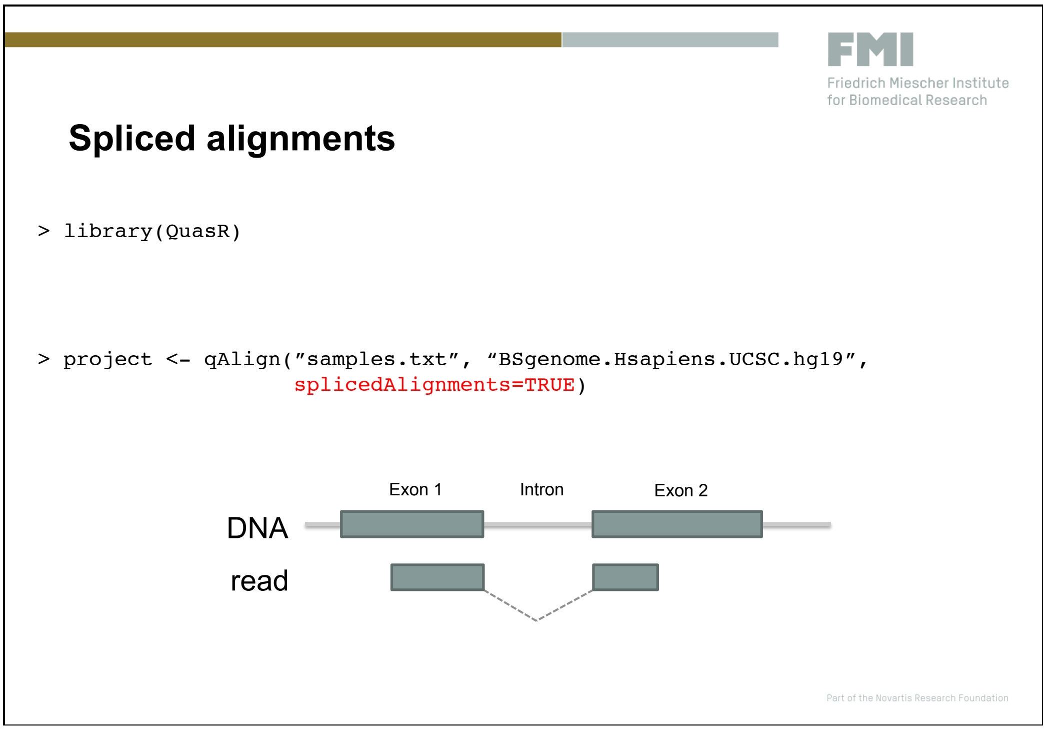
```
FileName          AuxName
NC_001422.1.fa    phiX174
```

```
> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”
                    auxiliaryFile=“auxiliaries.txt”)
```

NC_001422.1.fa

```
>NC_001422.1
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC
GGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGC
AGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTG
GACTGCTGGCGGAAAATGAGAAA
```

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

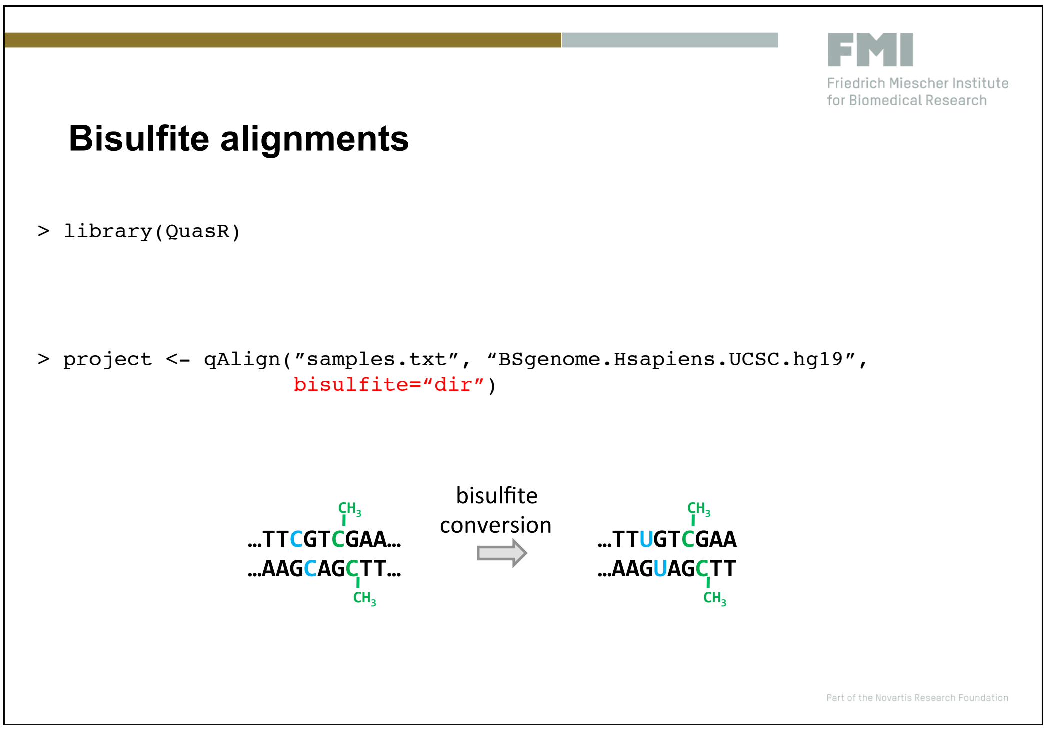
Spliced alignments

```
> library(QuasR)
```

```
> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”,
                    splicedAlignments=TRUE)
```

Exon 1
Intron
Exon 2
DNA
read

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

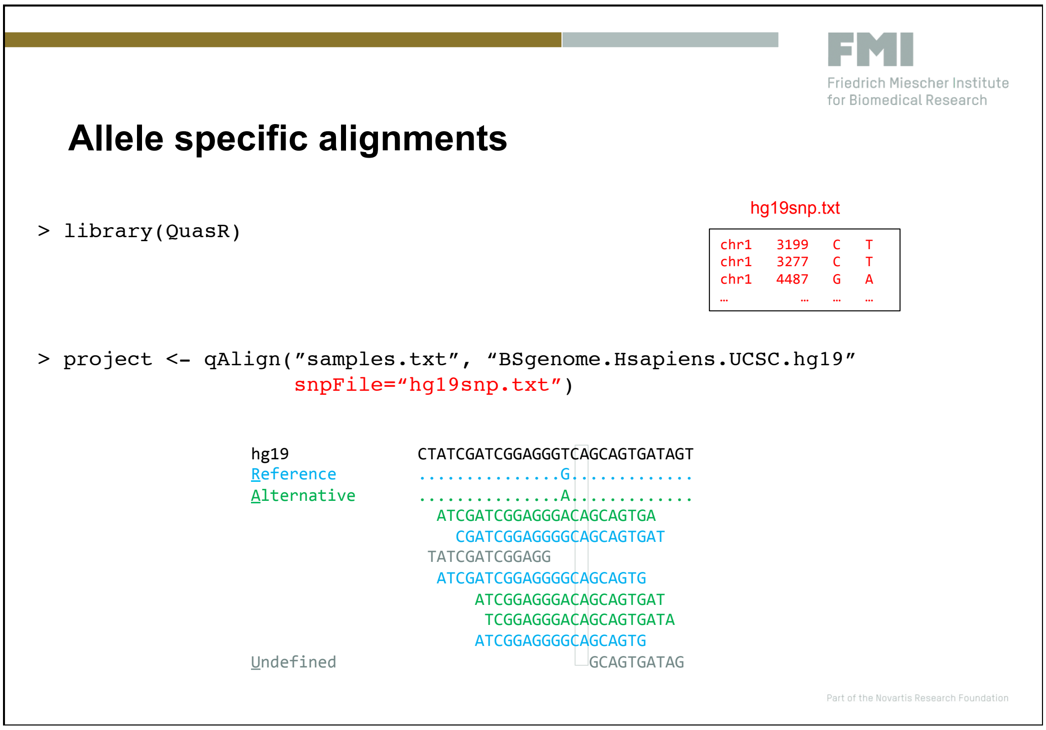
Bisulfite alignments

```
> library(QuasR)
```

```
> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”,
                    bisulfite=“dir”)
```

CH_3
...TTCGTCGAA...
...AAGCAGCTT...
CH_3
bisulfite
conversion
CH_3
...TTUGTCGAA
...AAGUAGCTT
CH_3

Part of the Novartis Research Foundation



Allele specific alignments

```
> library(QuasR)
```

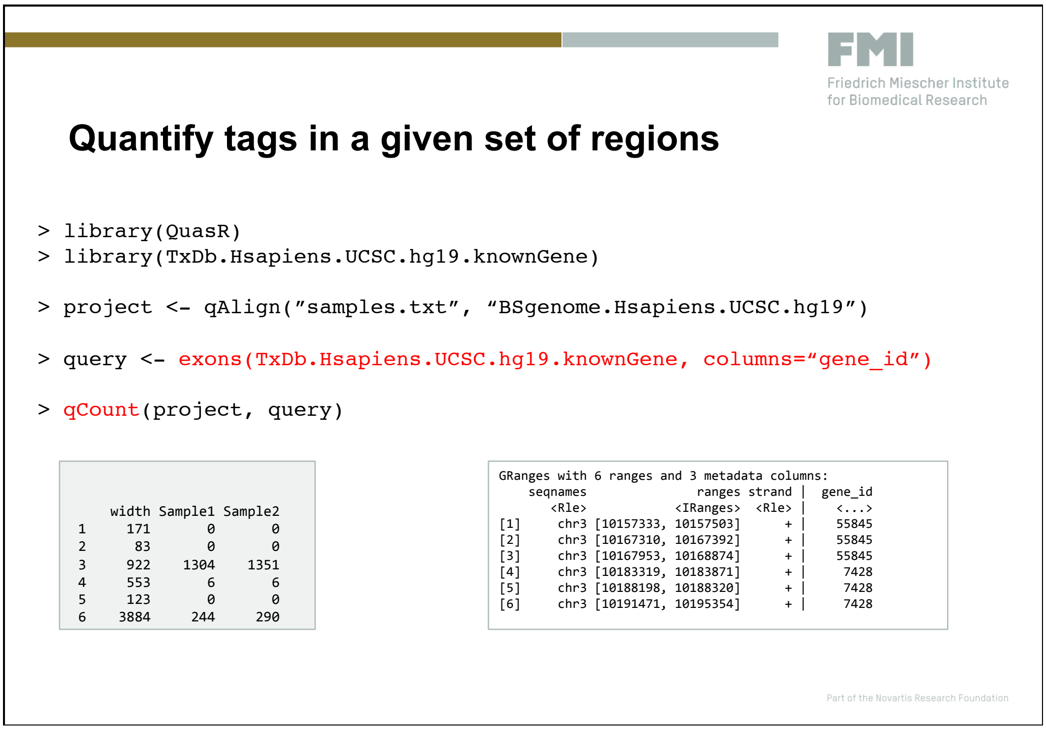
hg19snp.txt

```
chr1   3199   C   T
chr1   3277   C   T
chr1   4487   G   A
…         …   …   …
```

```
> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”
                   snpFile=“hg19snp.txt”)
```

```
hg19              CTATCGATCGGAGGGTCAGCAGTGATAGT
Reference         ................G.............
Alternative       ................A.............
                    ATCGATCGGAGGGACAGCAGTGA
                      CGATCGGAGGGGCAGCAGTGAT
                   TATCGATCGGAGG
                    ATCGATCGGAGGGGCAGCAGTG
                        ATCGGAGGGACAGCAGTGAT
                         TCGGAGGGACAGCAGTGATA
                        ATCGGAGGGGCAGCAGTG
Undefined                           GCAGTGATAG
```

Part of the Novartis Research Foundation



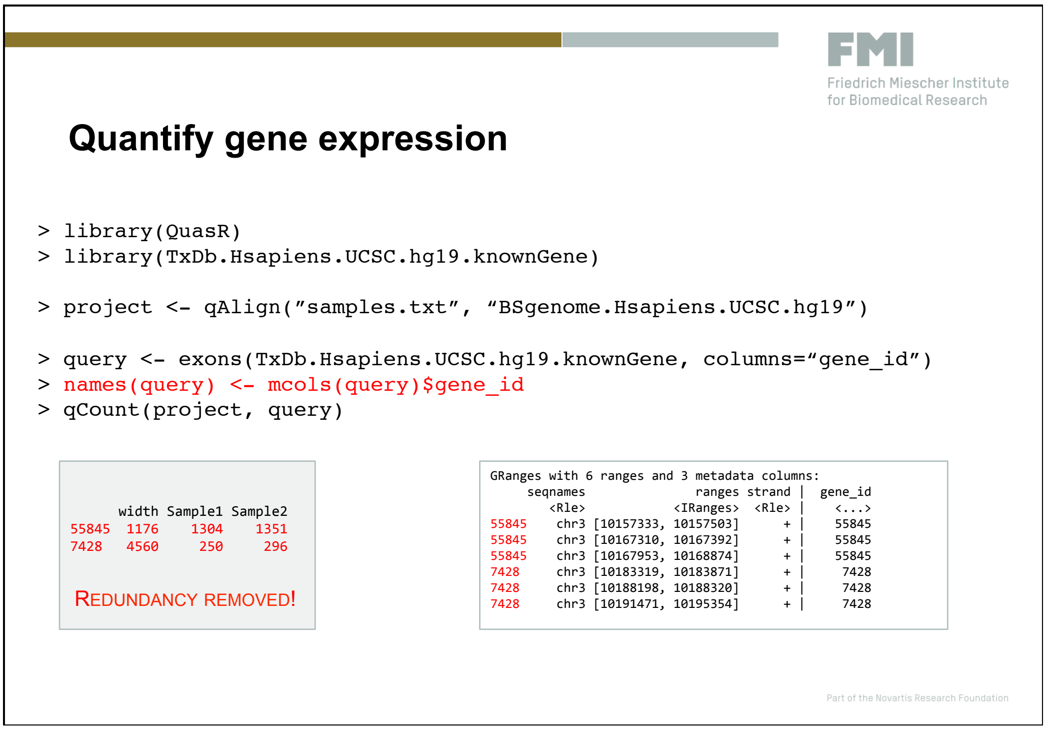
FMI
Friedrich Miescher Institute
for Biomedical Research

Quantify tags in a given set of regions

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)

> query <- exons(TxDb.Hsapiens.UCSC.hg19.knownGene, columns=“gene_id”)

> qCount(project, query)
```

```
   width Sample1 Sample2
1    171       0       0
2     83       0       0
3    922    1304    1351
4    553       6       6
5    123       0       0
6   3884     244     290
```

```
GRanges with 6 ranges and 3 metadata columns:
    seqnames                 ranges strand |   gene_id
       <Rle>              <IRanges>  <Rle> |     <...>
[1]     chr3 [10157333, 10157503]      + |     55845
[2]     chr3 [10167310, 10167392]      + |     55845
[3]     chr3 [10167953, 10168874]      + |     55845
[4]     chr3 [10183319, 10183871]      + |      7428
[5]     chr3 [10188198, 10188320]      + |      7428
[6]     chr3 [10191471, 10195354]      + |      7428
```



FMI
Friedrich Miescher Institute
for Biomedical Research

Quantify gene expression

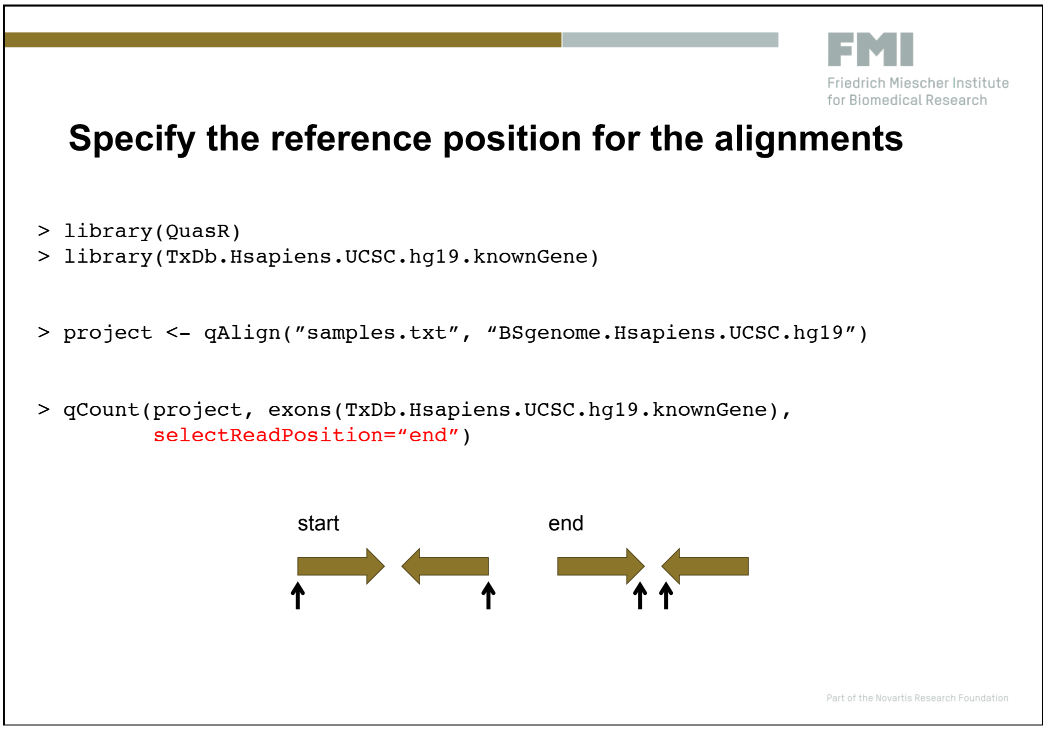
```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)

> query <- exons(TxDb.Hsapiens.UCSC.hg19.knownGene, columns=“gene_id”)
> names(query) <- mcols(query)$gene_id
> qCount(project, query)
```

```
      width Sample1 Sample2
55845  1176    1304    1351
7428   4560     250     296
```

REDUNDANCY REMOVED!

```
GRanges with 6 ranges and 3 metadata columns:
     seqnames                 ranges strand |  gene_id
        <Rle>              <IRanges>  <Rle> |    <...>
55845    chr3 [10157333, 10157503]      + |    55845
55845    chr3 [10167310, 10167392]      + |    55845
55845    chr3 [10167953, 10168874]      + |    55845
7428     chr3 [10183319, 10183871]      + |     7428
7428     chr3 [10188198, 10188320]      + |     7428
7428     chr3 [10191471, 10195354]      + |     7428
```

Part of the Novartis Research Foundation

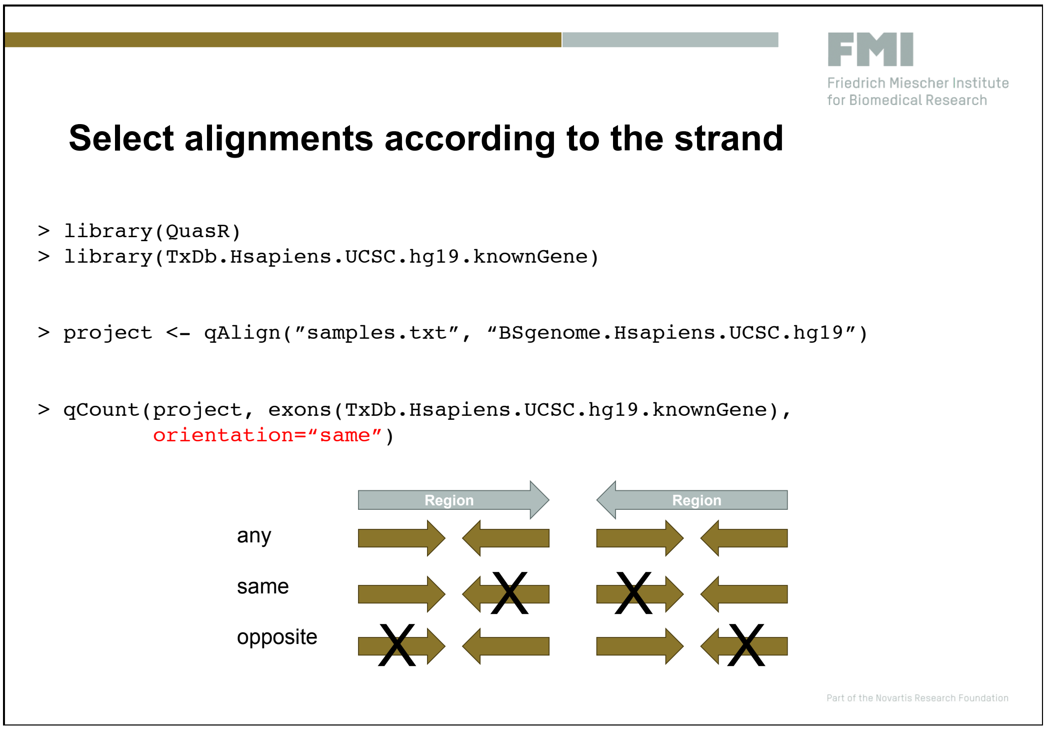


FMI
Friedrich Miescher Institute
for Biomedical Research

Specify the reference position for the alignments

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)


> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)


> qCount(project, exons(TxDb.Hsapiens.UCSC.hg19.knownGene),
         selectReadPosition=“end”)
```

Part of the Novartis Research Foundation



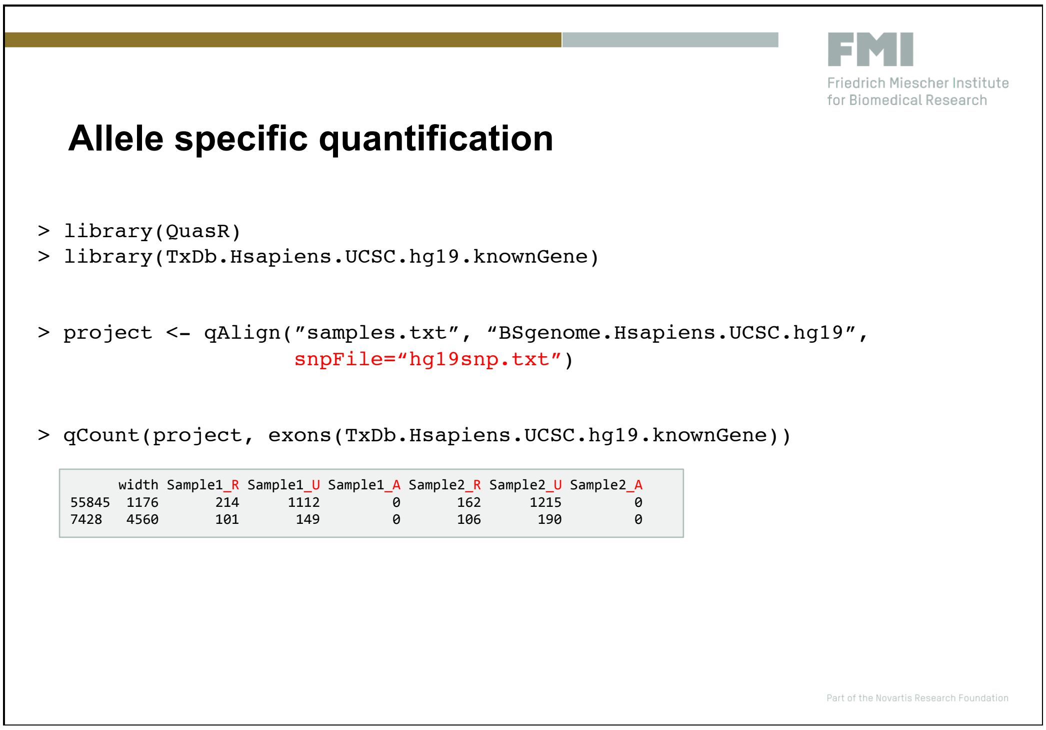
Friedrich Miescher Institute
for Biomedical Research

Select alignments according to the strand

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)


> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)


> qCount(project, exons(TxDb.Hsapiens.UCSC.hg19.knownGene),
         orientation=“same”)
```

Region
Region
any
same
opposite
X
X
X
X

Part of the Novartis Research Foundation



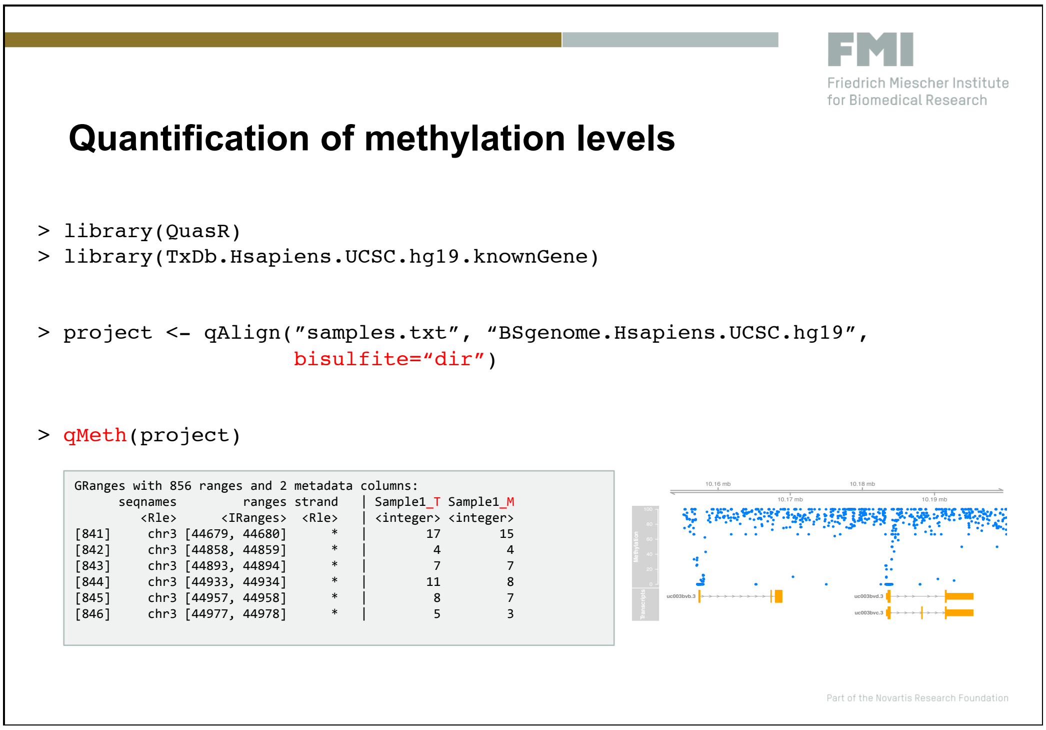
Friedrich Miescher Institute
for Biomedical Research

Allele specific quantification

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)


> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”,
                    snpFile=“hg19snp.txt”)


> qCount(project, exons(TxDb.Hsapiens.UCSC.hg19.knownGene))
```

```
      width Sample1_R Sample1_U Sample1_A Sample2_R Sample2_U Sample2_A
55845  1176       214      1112         0       162      1215         0
7428   4560       101       149         0       106       190         0
```

Part of the Novartis Research Foundation



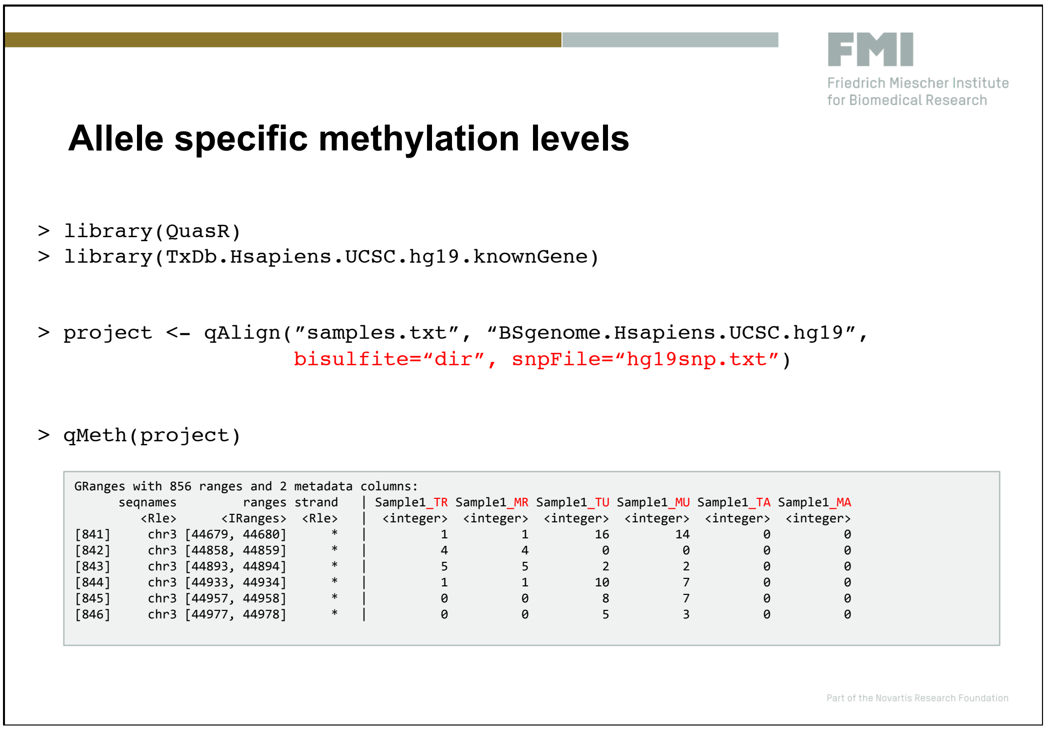
Friedrich Miescher Institute
for Biomedical Research

Quantification of methylation levels

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign("samples.txt", "BSgenome.Hsapiens.UCSC.hg19",
                    bisulfite="dir")

> qMeth(project)
```

```
GRanges with 856 ranges and 2 metadata columns:
      seqnames          ranges strand   | Sample1_T Sample1_M
         <Rle>       <IRanges>  <Rle>   | <integer> <integer>
[841]     chr3 [44679, 44680]       *   |        17        15
[842]     chr3 [44858, 44859]       *   |         4         4
[843]     chr3 [44893, 44894]       *   |         7         7
[844]     chr3 [44933, 44934]       *   |        11         8
[845]     chr3 [44957, 44958]       *   |         8         7
[846]     chr3 [44977, 44978]       *   |         5         3
```

Part of the Novartis Research Foundation



Friedrich Miescher Institute
for Biomedical Research

Allele specific methylation levels

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign("samples.txt", "BSgenome.Hsapiens.UCSC.hg19",
                    bisulfite="dir", snpFile="hg19snp.txt")

> qMeth(project)
```

```
GRanges with 856 ranges and 2 metadata columns:
      seqnames          ranges strand   | Sample1_TR Sample1_MR Sample1_TU Sample1_MU Sample1_TA Sample1_MA
         <Rle>       <IRanges>  <Rle>   |  <integer>  <integer>  <integer>  <integer>  <integer>  <integer>
[841]     chr3 [44679, 44680]       *   |          1          1         16         14          0          0
[842]     chr3 [44858, 44859]       *   |          4          4          0          0          0          0
[843]     chr3 [44893, 44894]       *   |          5          5          2          2          0          0
[844]     chr3 [44933, 44934]       *   |          1          1         10          7          0          0
[845]     chr3 [44957, 44958]       *   |          0          0          8          7          0          0
[846]     chr3 [44977, 44978]       *   |          0          0          5          3          0          0
```

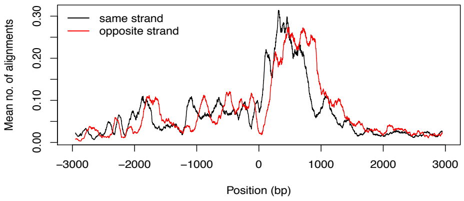
Part of the Novartis Research Foundation

FMI
Friedrich Miescher Institute
for Biomedical Research

Genomic profiles

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)

> query <- cds(TxDb.Hsapiens.UCSC.hg19.knownGene, columns=“gene_id”)

> qProfile(project, query, upstream=3000, downstream=3000)
```

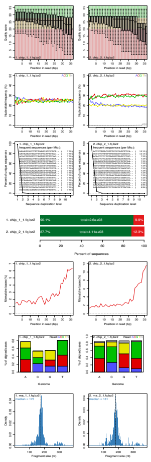
```
$coverage
      -3000 -2999 -2998 -2997 -2996 ...
query     8     8     8     8     8 ...

$Sample1
      -3000 -2999 -2998 -2997 -2996 ...
query     1     0     0     0     0 ...

$Sample2
      -3000 -2999 -2998 -2997 -2996 ...
query     0     0     0     2     0 ...
```

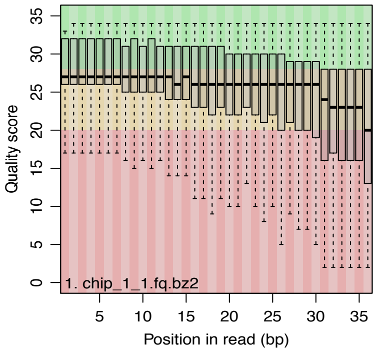


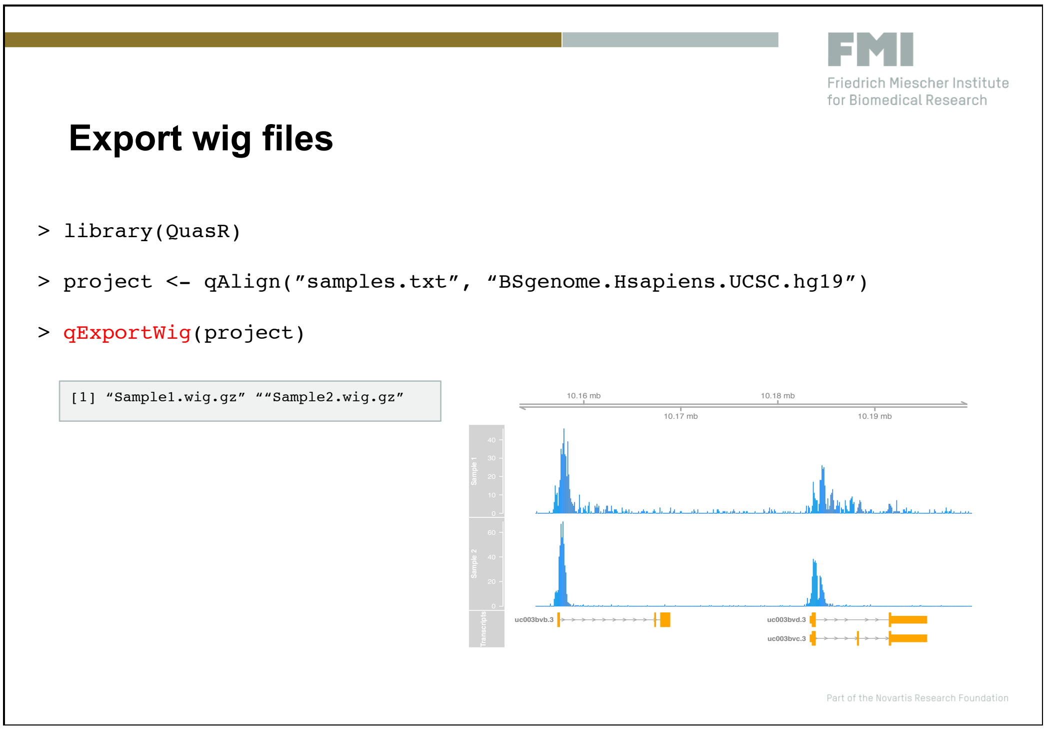
FMI
Friedrich Miescher Institute
for Biomedical Research

Quality control plots

```
> qQCReport(project, "qc_plots.pdf”)
```







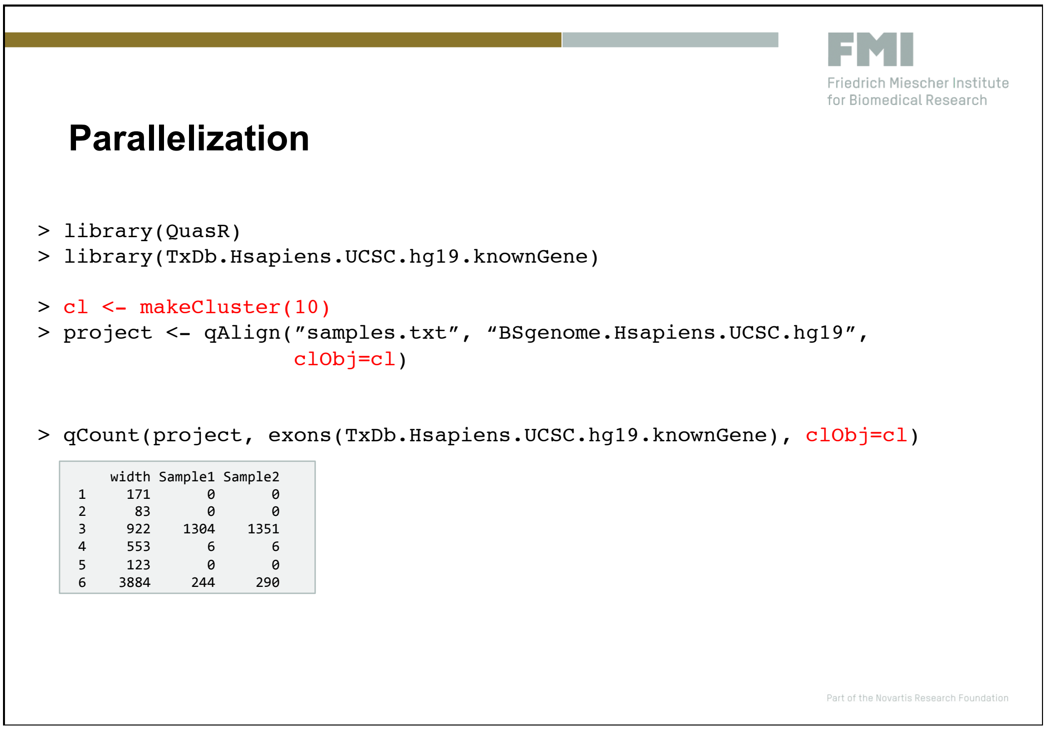
FMI
Friedrich Miescher Institute
for Biomedical Research

Export wig files

```
> library(QuasR)

> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”)

> qExportWig(project)
```

```
[1] “Sample1.wig.gz” ““Sample2.wig.gz”
```

Part of the Novartis Research Foundation



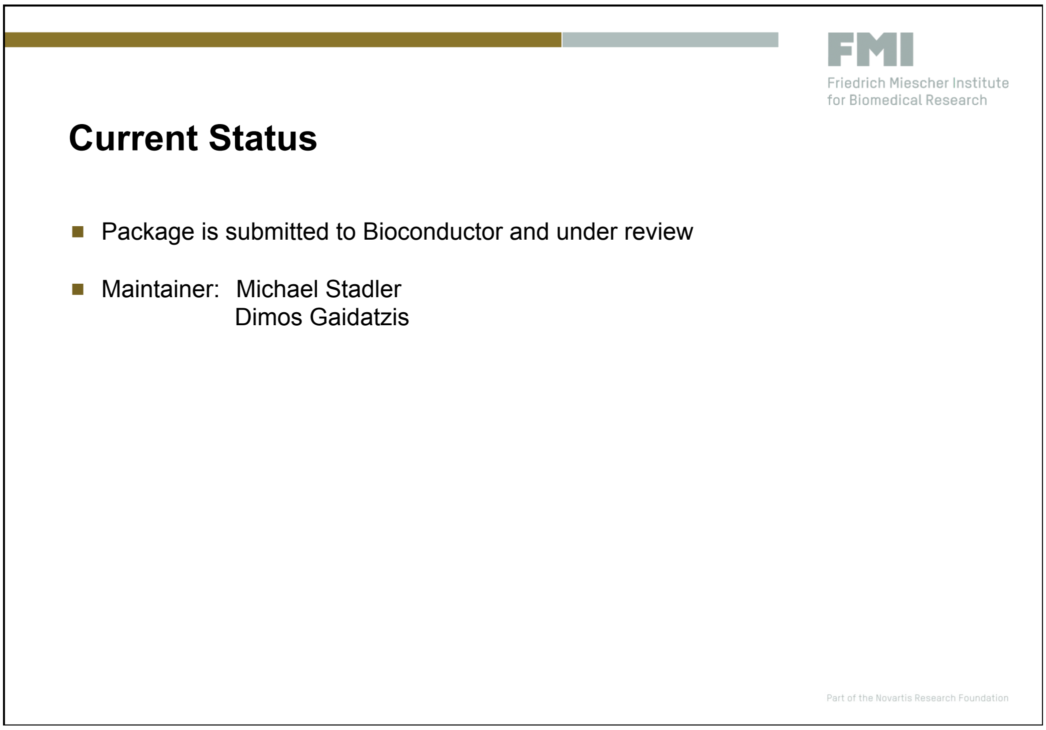
FMI
Friedrich Miescher Institute
for Biomedical Research

Parallelization

```
> library(QuasR)
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)

> cl <- makeCluster(10)
> project <- qAlign(”samples.txt”, “BSgenome.Hsapiens.UCSC.hg19”,
                    clObj=cl)


> qCount(project, exons(TxDb.Hsapiens.UCSC.hg19.knownGene), clObj=cl)
```

```
   width Sample1 Sample2
1    171       0       0
2     83       0       0
3    922    1304    1351
4    553       6       6
5    123       0       0
6   3884     244     290
```

Part of the Novartis Research Foundation

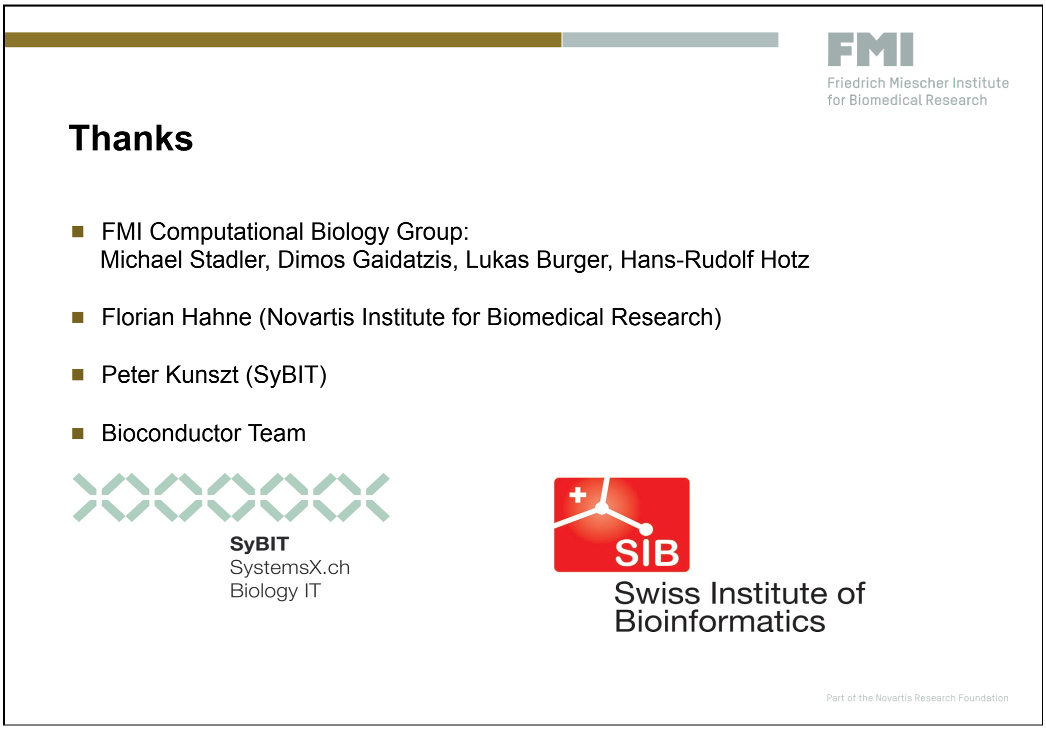


FMI
Friedrich Miescher Institute
for Biomedical Research

Current Status

- Package is submitted to Bioconductor and under review
- Maintainer: Michael Stadler
  Dimos Gaidatzis

Part of the Novartis Research Foundation



FMI
Friedrich Miescher Institute
for Biomedical Research

Thanks

- FMI Computational Biology Group:
  Michael Stadler, Dimos Gaidatzis, Lukas Burger, Hans-Rudolf Hotz
- Florian Hahne (Novartis Institute for Biomedical Research)
- Peter Kunszt (SyBIT)
- Bioconductor Team

SyBIT
SystemsX.ch
Biology IT

SIB
Swiss Institute of
Bioinformatics

Part of the Novartis Research Foundation