R / *Bioconductor* for Epigenomic Analysis

Martin Morgan mtmorgan@fhcrc.org
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

25 August 2014

R and *Bioconductor*

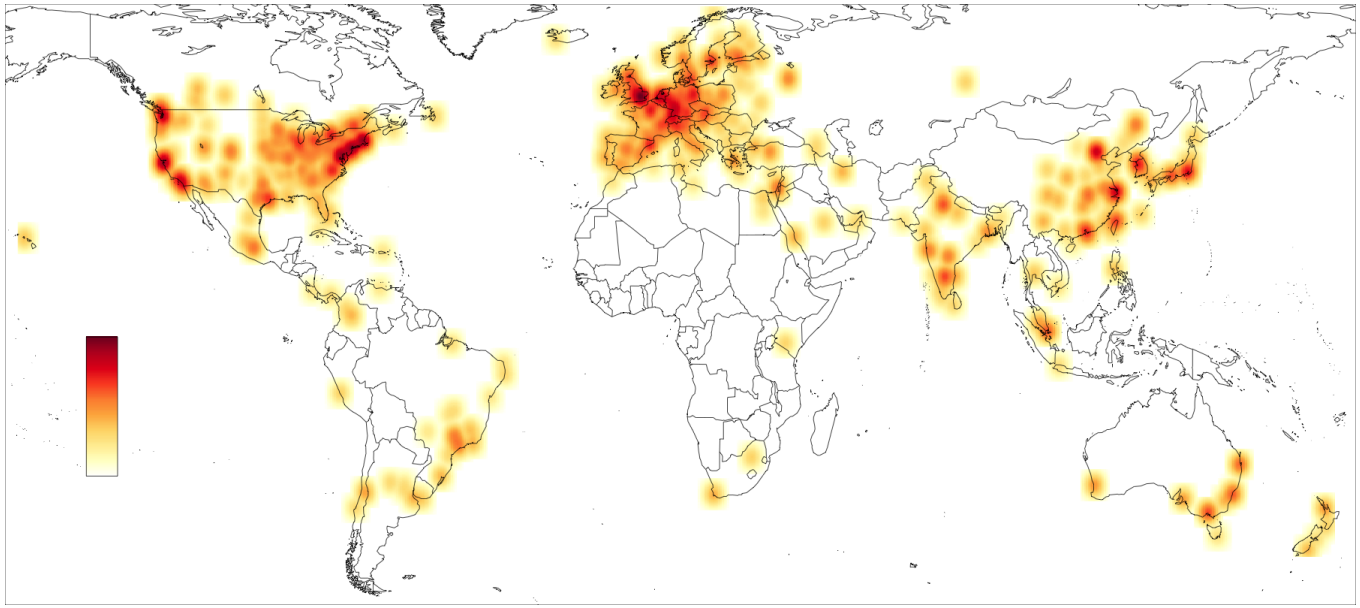
R, `http://r-project.org`

- Open-source, statistical programming language; widely used in academia, finance, pharma, ...
- Core language, 'base' and $>$ 5000 contributed packages
- Interactive sessions, scripts, packages

Bioconductor, `http://bioconductor.org`

- Analysis and comprehension of high-throughput genomic data
- Themes: rigorous statistical analysis; reproducible work flows; integrative analysis
- $>$ 12 years old, 825 packages
- Courses, conferences (package developers: travel scholarships available!), mailing list, ...

Bioconductor



Trailing 12 month statistics

- 1702 PubMedCentral full-text citations
- 9.5M package downloads to 242,000 distinct IP addresses
- 1M sessions from 400k visitors to web site
- Annual conferences; courses; active mailing list; ...

Epigenomics

1. Methylation
   - Illumina 450k arrays
   - Whole genome and restricted representation bisulfite sequencing
2. ChIP-Seq
3. Integration & visualization
   - Data types, e.g., expression
   - Consortium resources *via* *AnnotationHub*
4. A key data structure: *GRanges*

Find relevant packages using BiocViews, in addition to standard scholarly approaches.

Methylation: selected packages

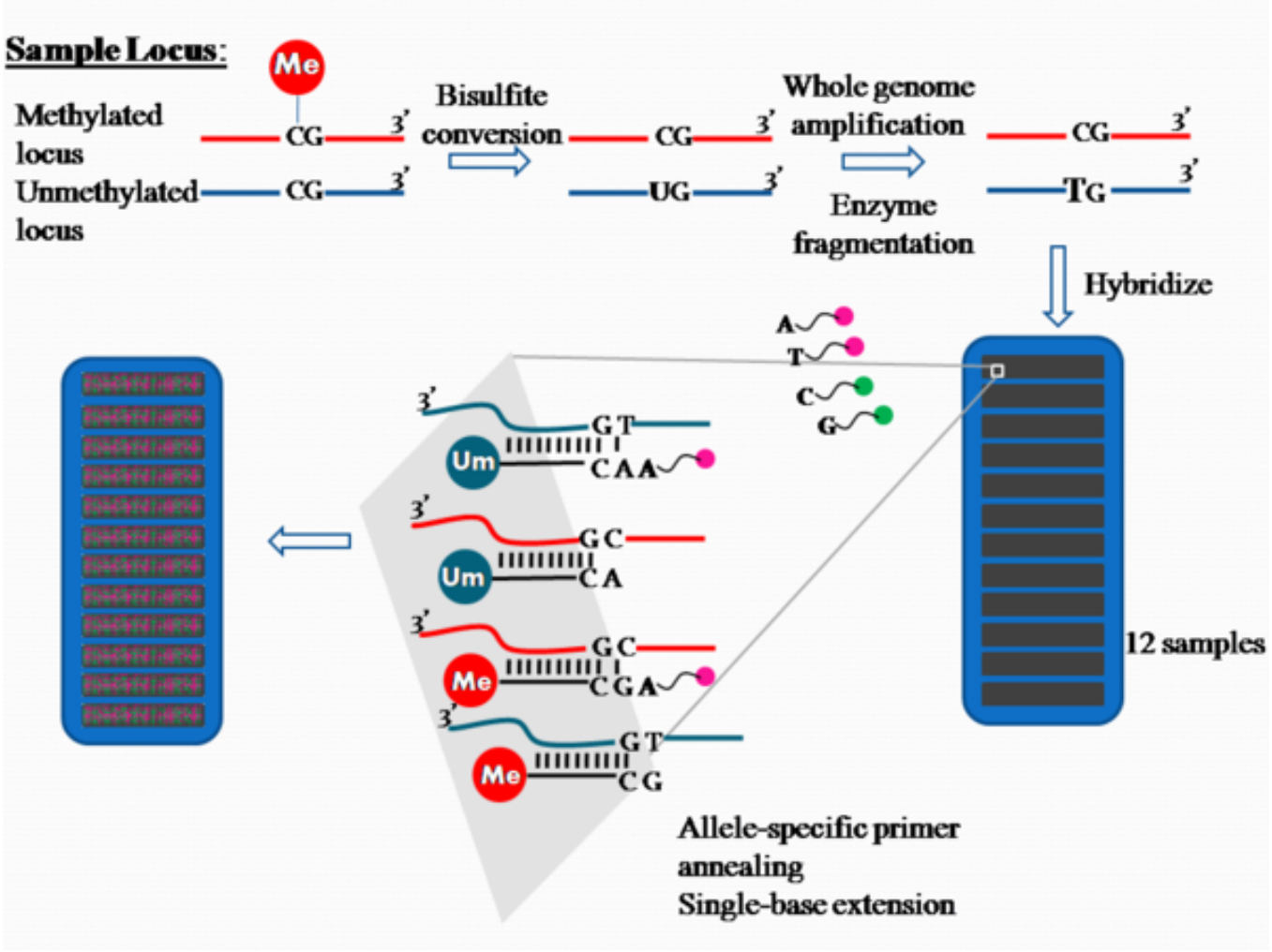
450k arrays

- *minfi* – pre-processing, differential methylation
- *ChAMP* – comprehensive work flow

Bisulfite sequencing

- *bsseq* – whole genome
- *BiSeq* – restricted representation

Methylation: Illumina 450k arrays



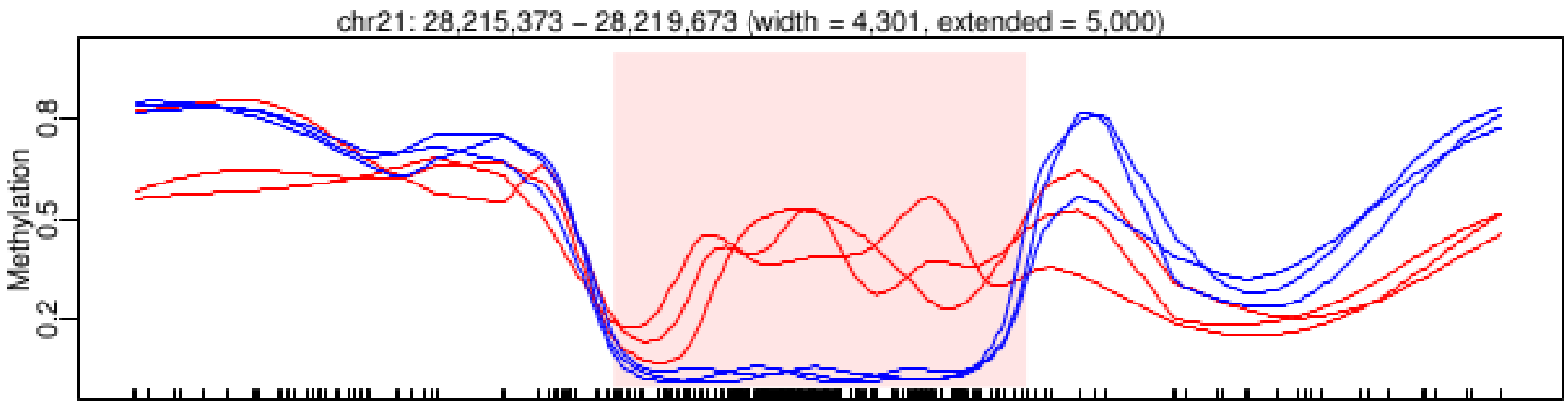
http://en.wikipedia.org/wiki/Illumina_Methylation_Assay

Methylation: *minfi*

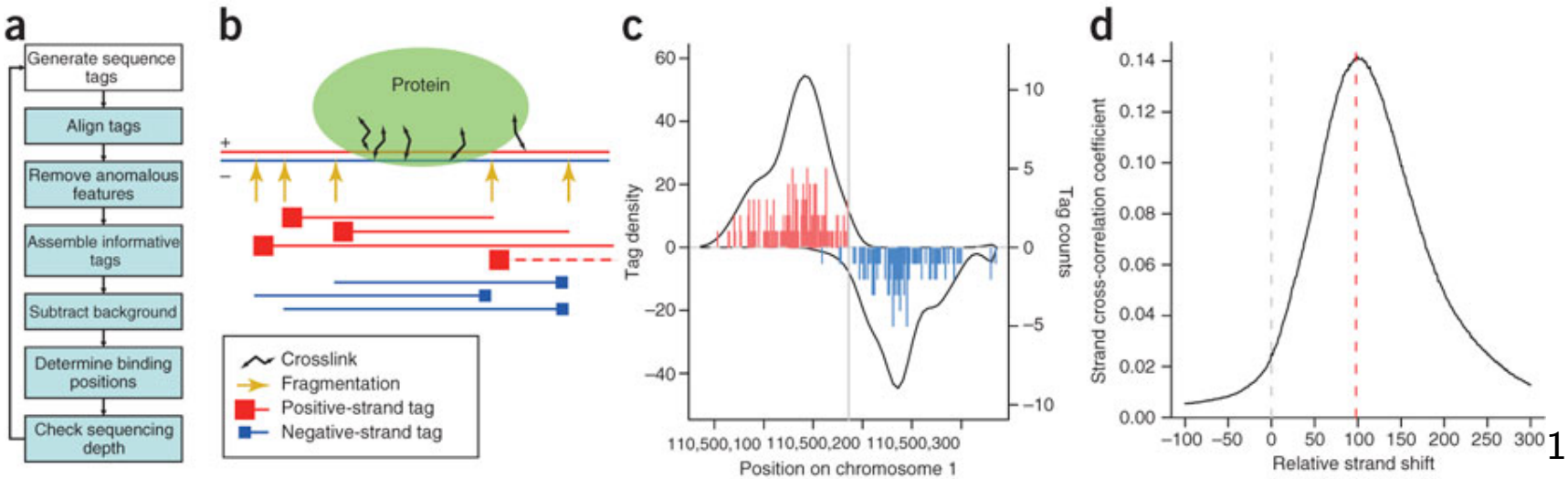
Work flow steps and representative functions

1. Data input: `read.450k.exp()`
2. Quality assessment: `densityPlot()`
3. Pre-processing, e.g., background correction, normalization: `preprocessIllumina()`
4. Differentially methylated probes: `dmpFinder()`
5. Differentially methylated regions: `bumphunter()`

See the vignette for additional detail.



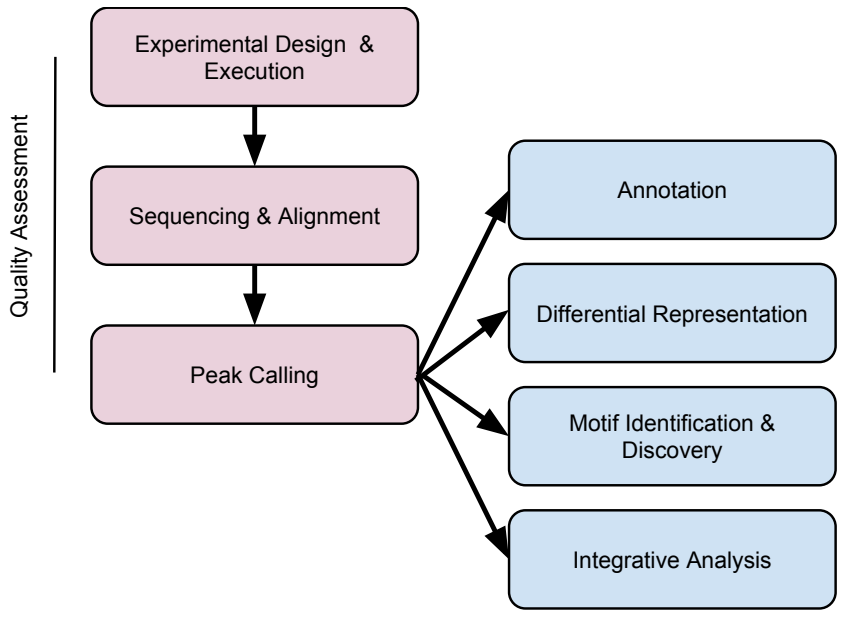
ChIP-seq: peak calling



- Chromatin immunoprecipitation, followed by sequencing to determine location of proteins bound to DNA
- Useful for locating transcription factor binding sites, histone modifications, ...

[1] Kharchenko et al. (2008)

ChIP-seq work flow



Analysis overview: Bailey et al. (2013)

- Annotation: what genes are my peaks near?
- Differential representation: which peaks are over- or under-represented in treatment 1, compared to treatment 2?
- Motif identification (peaks over known motifs?) and discovery
- Integrative analysis, e.g., association of regulatory elements and expression

ChIP-seq quality assessment: *ChIPQC*

Inputs: BAM files (raw data) and BED files (called peaks)

```
experiment <- ChIPQC(samples)
ChIPQCreport(experiment)
```

Output: HTML report — http://starkhome.com/ChIPQC/Reports/tamoxifen/ChIPQC.html

ChIP-seq annotation: *ChIPpeakAnno*, *ChIPseeker*

Inputs

- Peaks: e.g., from `rtracklayer::import()` BED files
- Annotation: gene boundaries or queries to *biomaRt*

```
library(ChIPpeakAnno)
## ...
annotated <- annotatePeakInBatch(peaks,
    AnnotationData=annotation)
```

Output: *RangedData* with annotations about near-by peaks.

ChIP-seq differential representation: *DiffBind*

Inputs: called peaks and raw BED or BAM files

```
library(DiffBind)
tamoxifen = dba(sampleSheet="tamoxifen.csv")
tamoxifen = dba.count(tamoxifen)
tamoxifen = dba.contrast(tamoxifen,
    categories=DBA_CONDITION)
tamoxifen = dba.analyze(tamoxifen)
tamoxifen.DB = dba.report(tamoxifen)
```

Outputs: diagnostics, visualizations, and 'top table' of differentially expressed regions.

Integration & visualization

- Combining multiple data types
  - *Rcade*, *Repitools*: ChIP / expression
- Import / export from common formats (BED, WIG, ...)
  - *rtracklayer* `import()`, `export()`
- *AnnotationHub*: accessing large-scale resources, e.g., ENCODE tracks
- Visualization

Integration & visualization: *AnnotationHub*

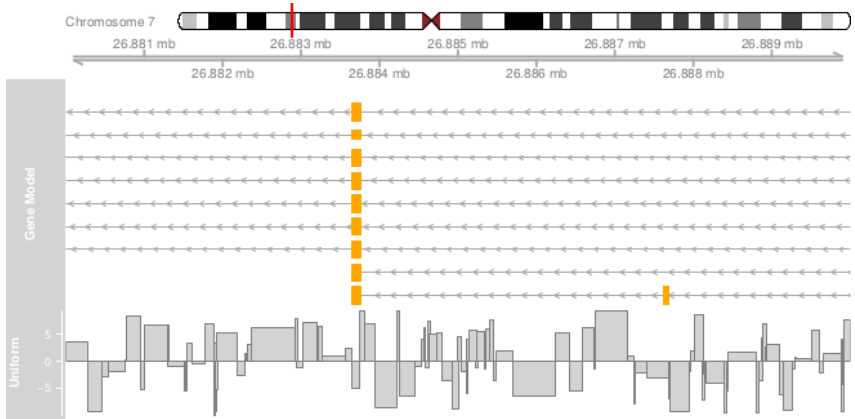
```
> library(AnnotationHub)
> hub = AnnotationHub()
> hub
class: AnnotationHub
length: 10780
filters: none
hubUrl: http://annotationhub.bioconductor.org/ah
snapshotVersion: 2.14/1.4.0; snapshotDate: 2014-05-15
hubCache: /home/mtmorgan/.AnnotationHub
> hub$<tab>
hub$dbSNP.organisms.human_9606.VCF. ... [302]
hub$haemcode.blood. ... [899]
hub$ensembl.release. ... [2611]
hub$inparanoid8.Orthologs.hom. ... [265]
hub$goldenpath. ... [6699]
hub$refnet. ... [4]
```

Integration & visualization

Gviz

- Static track-like visualizations
- Data panels

ggbio
epivizr

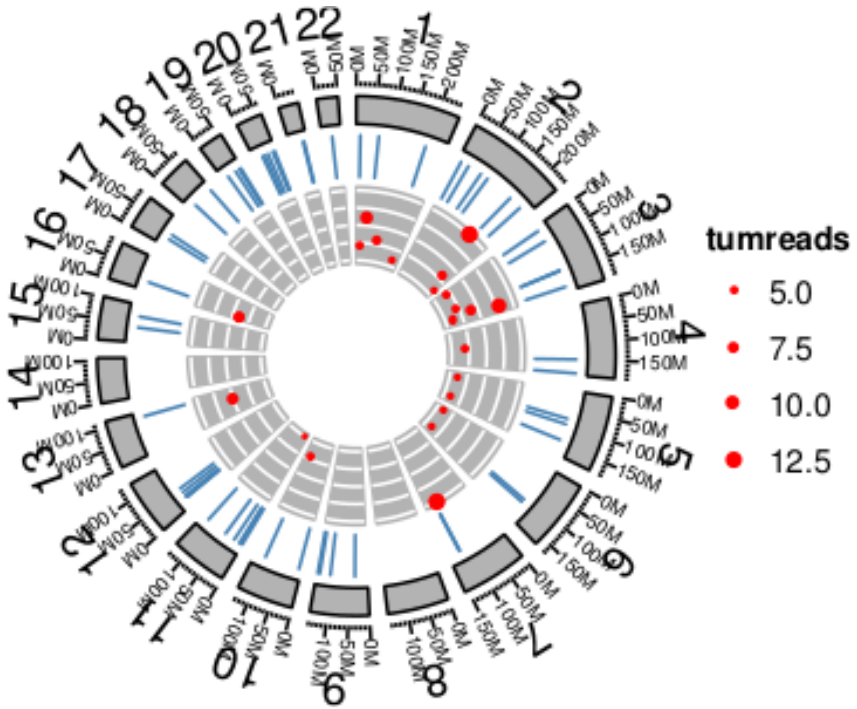


Integration & visualization

Gviz
ggbio

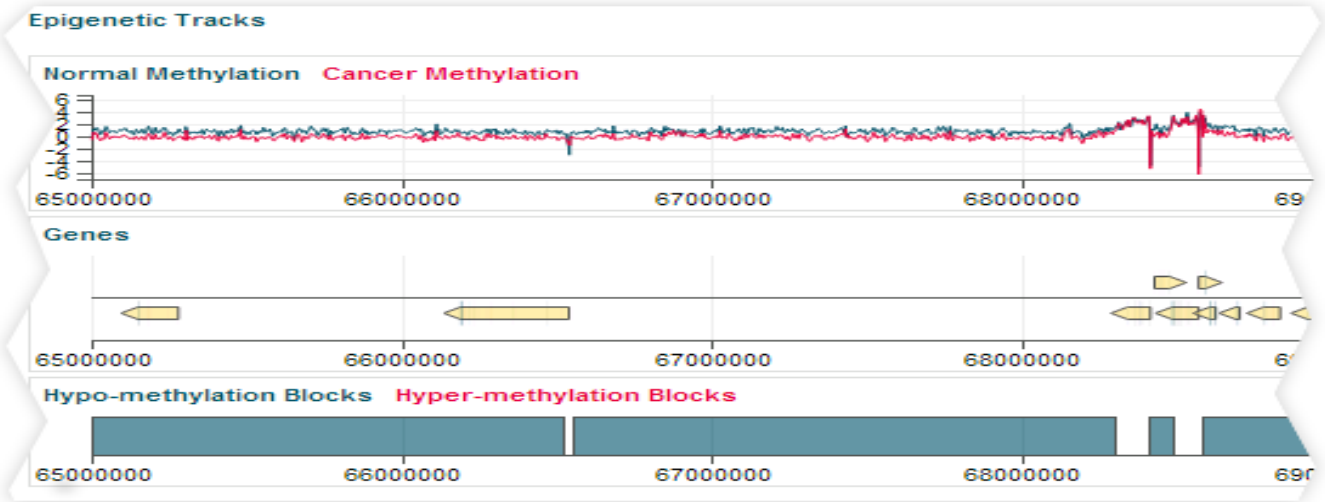
- Comprehensive visualizations
- `autoplot` file and data types

epivizr



Integration & visualization

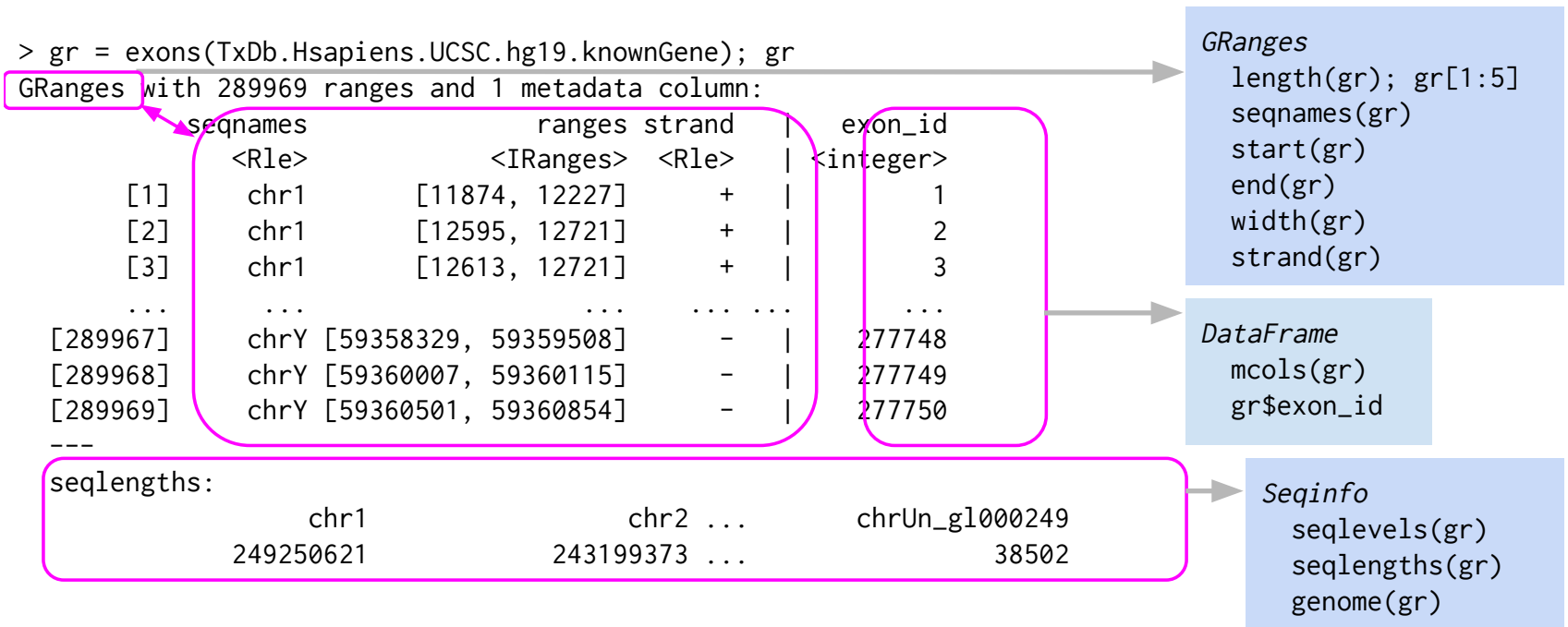
Gviz
ggbio
epivizr



- Genome browser with tight communication to *R* / *Bioconductor*
- Flexible *interactive*, *representation* and *computation*, e.g., 'brushing'

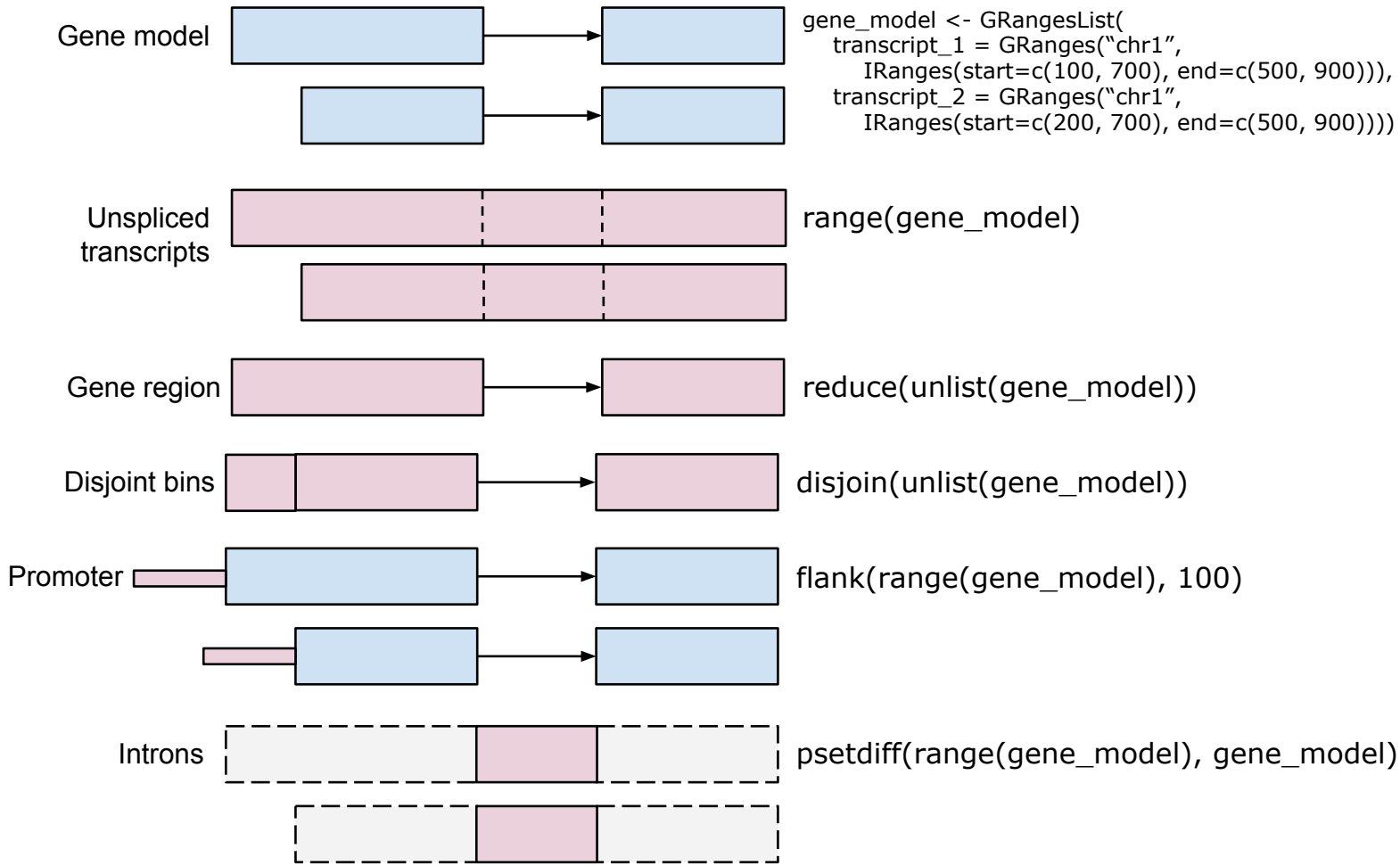
Genomic ranges for data integration

- Chromosome, start, end, strand define a *genomic range*
- Data (reads, CpG islands, peaks, . . .) are genomic ranges
- Annotations (exons, genes, binding sites, . . .) are genomic ranges



GenomicRanges, *GenomicAlignments* packages

Operating on genomic ranges



Funding

- US NIH / NHGRI 2U41HG004059; NSF 1247813

People

- Seattle Bioconductor team: Sonali Arora, Marc Carlson, Nate Hayden, Valerie Obenchain, Hervé Pagès, Dan Tenenbaum
- Vincent Carey, Robert Gentleman, Rafael Irizzary, Sean Davis, Kasper Hansen, Michael Lawrence, Levi Waldron
- International community of *Bioconductor* developers and users

References I

T. Bailey et al. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS Comput Biol*, 9(11):e1003326, 11 2013. doi: 10.1371/journal.pcbi.1003326.

P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26(12):1351–1359, Dec 2008.