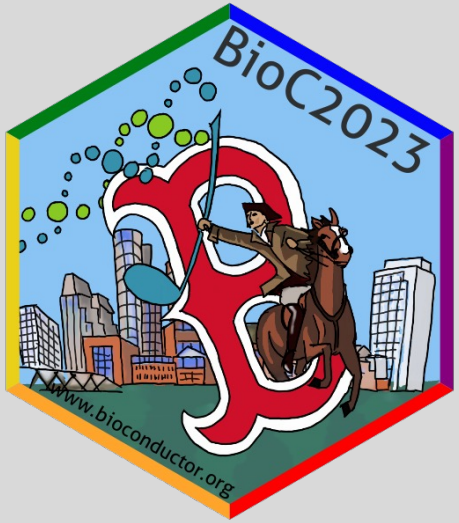
Orchestrating Hi-C analysis with Bioconductor

Package demonstration

Jacques Serizay

Bioconductor Conference 2023



Resources

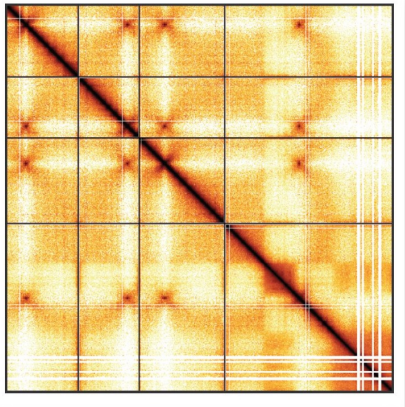
- Book "*Orchestrating Hi-C analysis with Bioconductor*": https://js2264.github.io/OHCA/
- Package demo (Bioc2023) walkthrough: https://js2264.github.io/OHCA.Bioc2023/
- HiCExperiment & HiContacts repositories (& others): https://github.com/js2264/OHCA/

Orchestrating Hi-C analysis with Bioconductor

Welcome

This is the landing page of the **"Orchestrating Hi-C analysis with Bioconductor"** book. **The primary aim of this book is to introduce the `R` user to Hi-C analysis**. This book starts with key concepts important for the analysis of chromatin conformation capture and then presents `Bioconductor` tools that can be leveraged to process, analyze, explore and visualize Hi-C data.

Authors: Jacques Serizay [aut, cre]
Version: 1.1.0
Modified: 2023-04-14
Compiled: 2023-07-24
Environment: R version 4.3.1 (2023-06-16), Bioconductor 3.18
License: MIT + file LICENSE
Copyright: J. Serizay



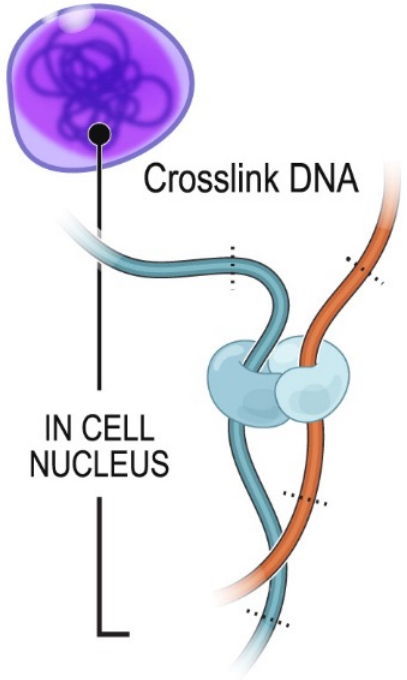
Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ Introduction to the `OHCA` ecosystem
- ❑ Importing Hi-C data with `HiCExperiment`
- ❑ Manipulating and visualizing Hi-C data with `HiContacts`
- ❑ Inter-operability with existing Hi-C packages
- ❑ Wrapping-up

Outline

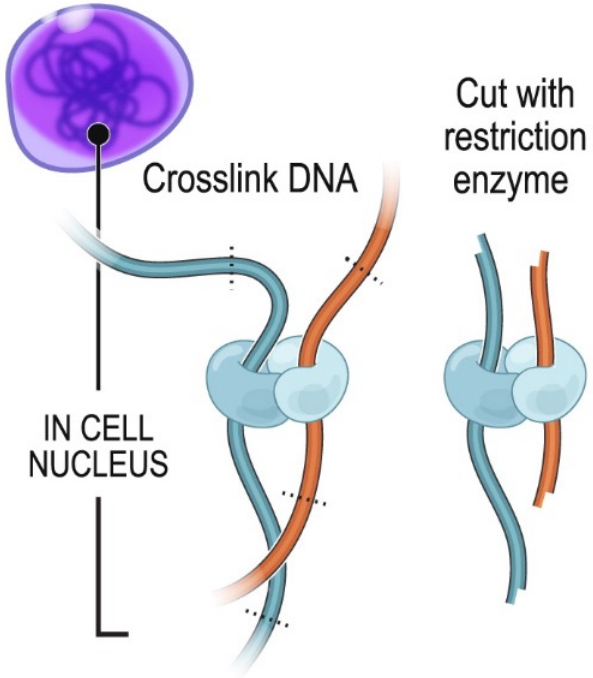
- [x] **Overview of Chromosome Conformation Capture technical aspects**
- [ ] Introduction to the `OHCA` ecosystem
- [ ] Importing Hi-C data with `HiCExperiment`
- [ ] Manipulating and visualizing Hi-C data with `HiContacts`
- [ ] Inter-operability with existing Hi-C packages
- [ ] Wrapping-up

Chromosome Conformation Capture



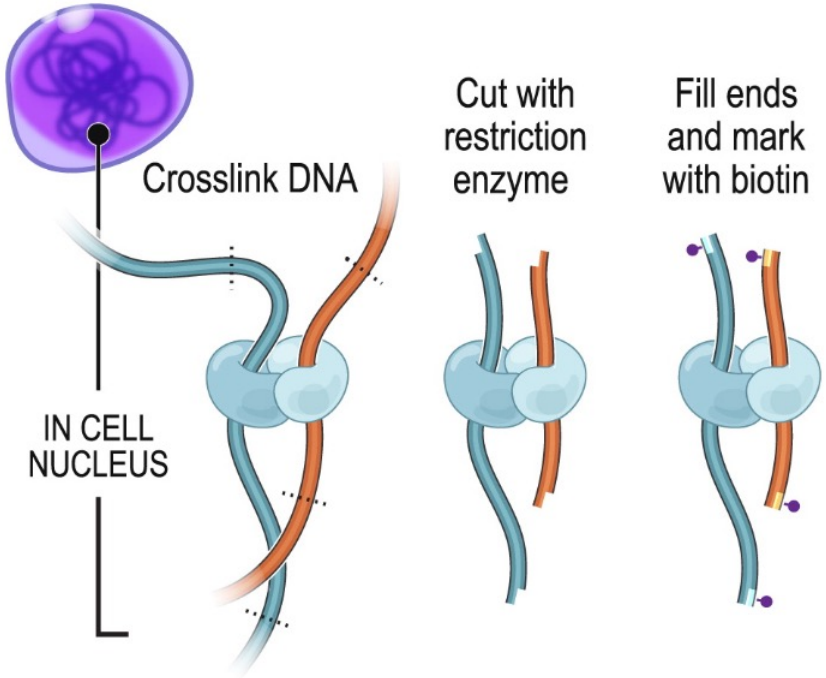
Rao et al., Cell 2014

Chromosome Conformation Capture



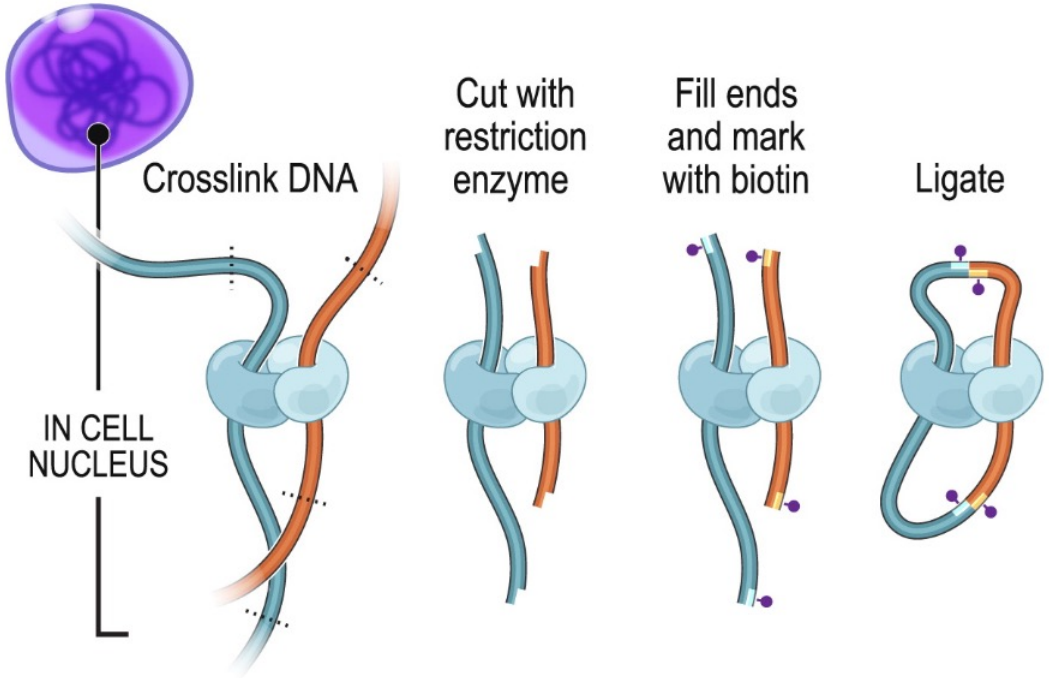
Rao et al., Cell 2014

Chromosome Conformation Capture



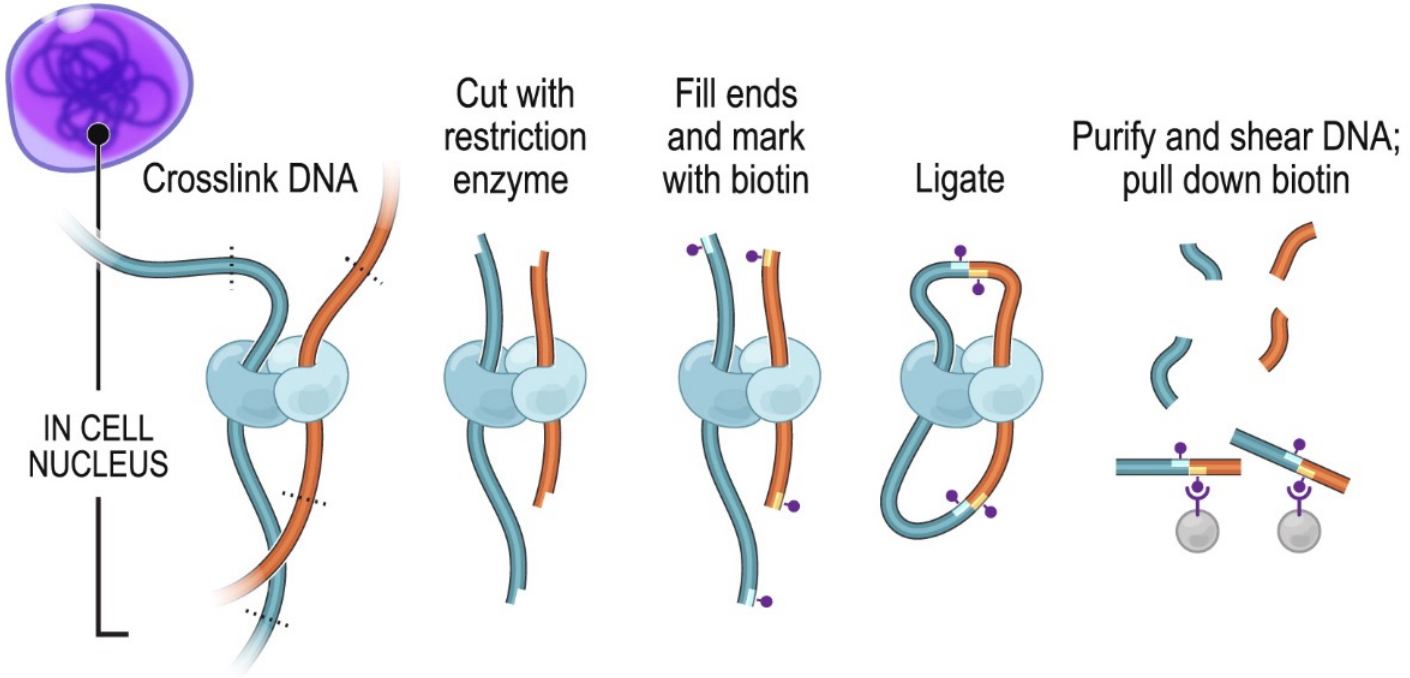
Rao et al., Cell 2014

Chromosome Conformation Capture



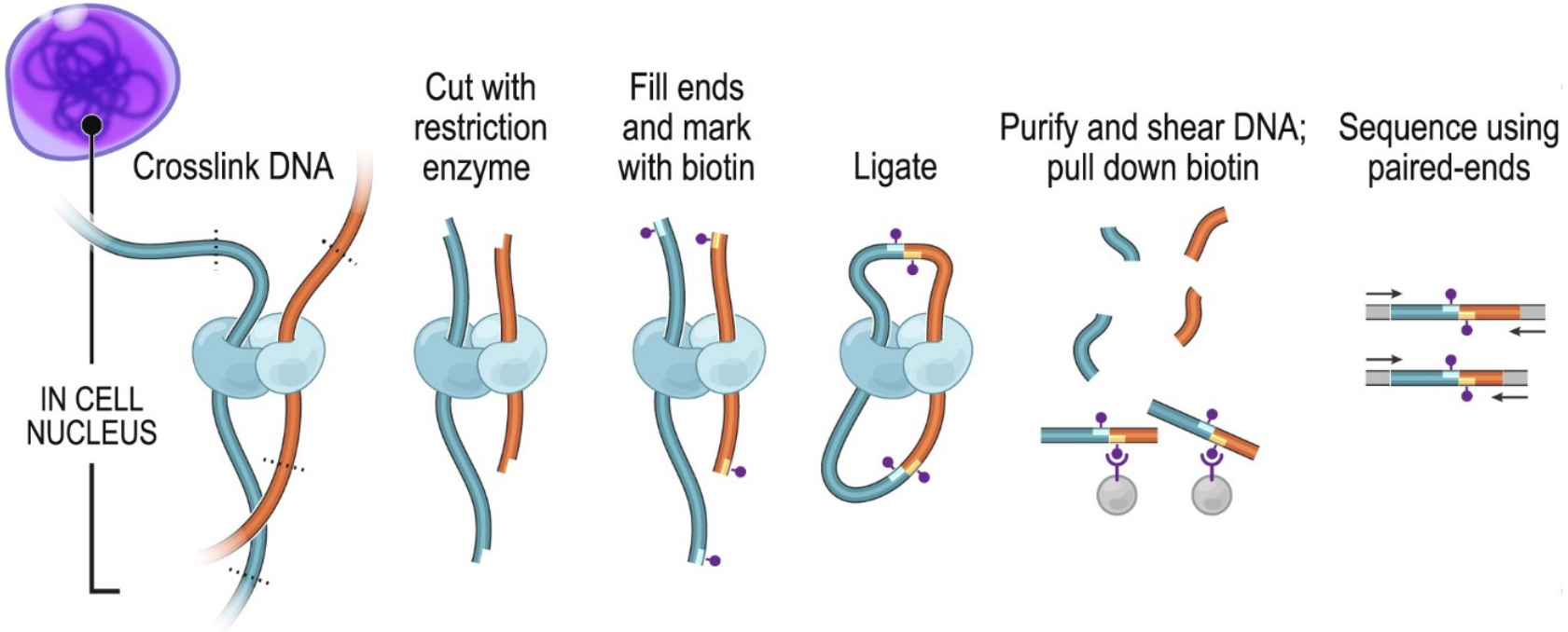
Rao et al., Cell 2014

Chromosome Conformation Capture



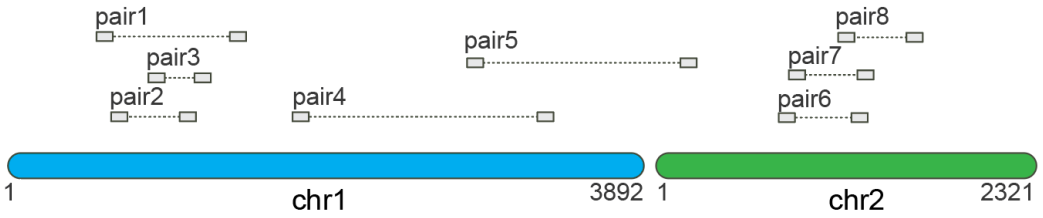
Rao et al., Cell 2014

Chromosome Conformation Capture

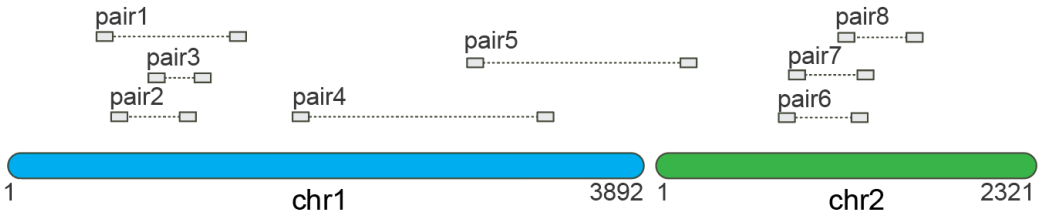


Rao et al., Cell 2014

Chromosome Conformation Capture processing



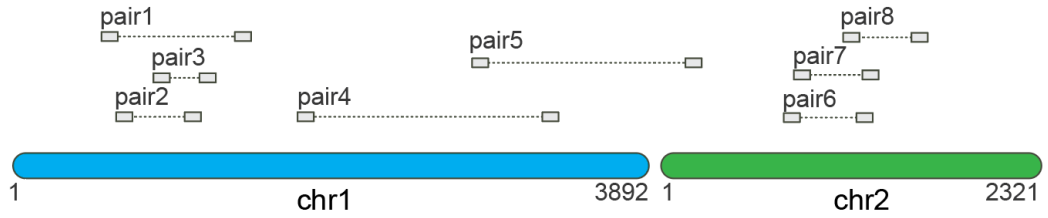
Chromosome Conformation Capture processing



pairs file

pairID	chr1	start1	end1	chr2	start2	end2
pair1	chr1	520	620	chr1	1312	1412
pair2	chr1	681	781	chr1	1124	1224
pair3	chr1	912	1012	chr1	1076	1176
pair4	chr1	1743	1843	chr1	3356	3456
pair5	chr1	2875	2975	chr2	243	343
pair6	chr2	743	843	chr2	1182	1282
pair7	chr2	798	898	chr2	1213	1313
pair8	chr2	1112	1212	chr2	1524	1624

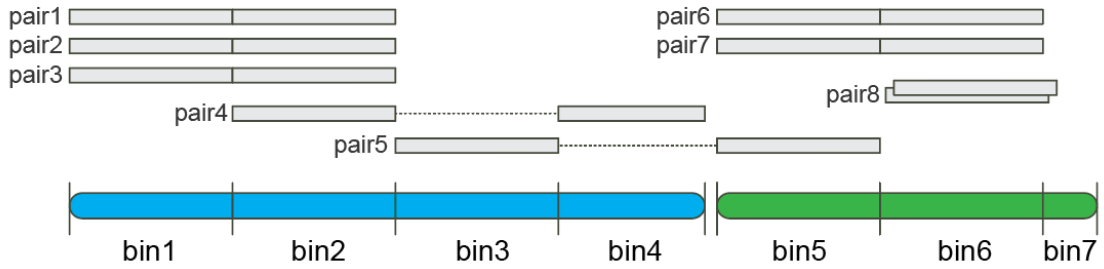
Chromosome Conformation Capture processing



pairs file

pairID	chr1	start1	end1	chr2	start2	end2
pair1	chr1	520	620	chr1	1312	1412
pair2	chr1	681	781	chr1	1124	1224
pair3	chr1	912	1012	chr1	1076	1176
pair4	chr1	1743	1843	chr1	3356	3456
pair5	chr1	2875	2975	chr2	243	343
pair6	chr2	743	843	chr2	1182	1282
pair7	chr2	798	898	chr2	1213	1313
pair8	chr2	1112	1212	chr2	1524	1624

Binning



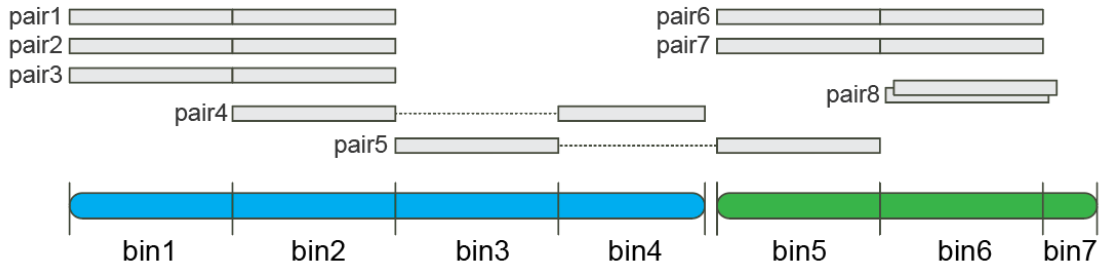
binned pairs file (temp.)

pairID	binID_1	binID_2
pair1	bin1	bin2
pair2	bin1	bin2
pair3	bin1	bin2
pair4	bin2	bin4
pair5	bin3	bin5
pair6	bin5	bin6
pair7	bin5	bin6
pair8	bin6	bin6

regions file

binID	chr	start	end
bin1	chr1	1	1000
bin2	chr1	1001	2000
bin3	chr1	2001	3000
bin4	chr1	3001	3892
bin5	chr2	1	1000
bin6	chr2	1001	2000
bin7	chr2	2001	2321

Chromosome Conformation Capture processing



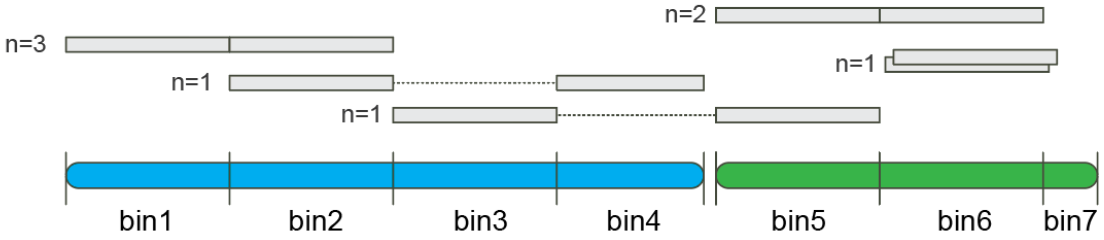
binned pairs file (temp.)

pairID	binID_1	binID_2
pair1	bin1	bin2
pair2	bin1	bin2
pair3	bin1	bin2
pair4	bin2	bin4
pair5	bin3	bin5
pair6	bin5	bin6
pair7	bin5	bin6
pair8	bin6	bin6

regions file

binID	chr	start	end
bin1	chr1	1	1000
bin2	chr1	1001	2000
bin3	chr1	2001	3000
bin4	chr1	3001	3892
bin5	chr2	1	1000
bin6	chr2	1001	2000
bin7	chr2	2001	2321

Summarizing



binned counts

binID_1	binID_2	counts
bin1	bin2	3
bin2	bin4	1
bin3	bin5	1
bin5	bin6	2
bin6	bin6	1

regions file

binID	chr	start	end
bin1	chr1	1	1000
bin2	chr1	1001	2000
bin3	chr1	2001	3000
bin4	chr1	3001	3892
bin5	chr2	1	1000
bin6	chr2	1001	2000
bin7	chr2	2001	2321

Chromosome Conformation Capture processing

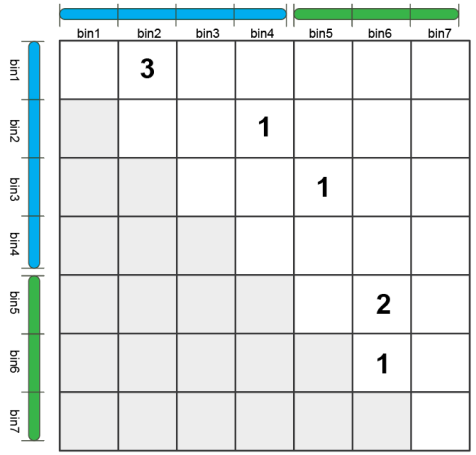
binned counts

binID_1	binID_2	counts
bin1	bin2	3
bin2	bin4	1
bin3	bin5	1
bin5	bin6	2
bin6	bin6	1

regions file

binID	chr	start	end	norm.
bin1	chr1	1	1000	<dbl>
bin2	chr1	1001	2000	...
bin3	chr1	2001	3000	
bin4	chr1	3001	3892	
bin5	chr2	1	1000	
bin6	chr2	1001	2000	
bin7	chr2	2001	2321	

=



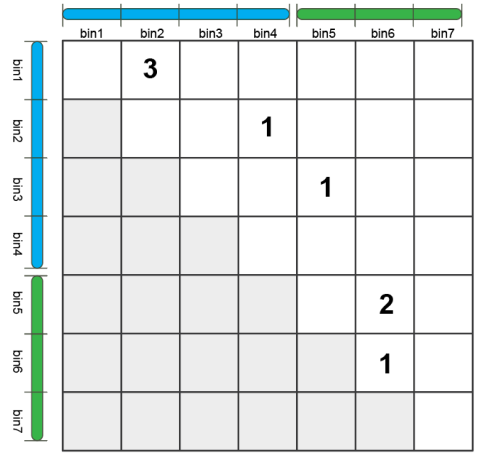
Chromosome Conformation Capture processing

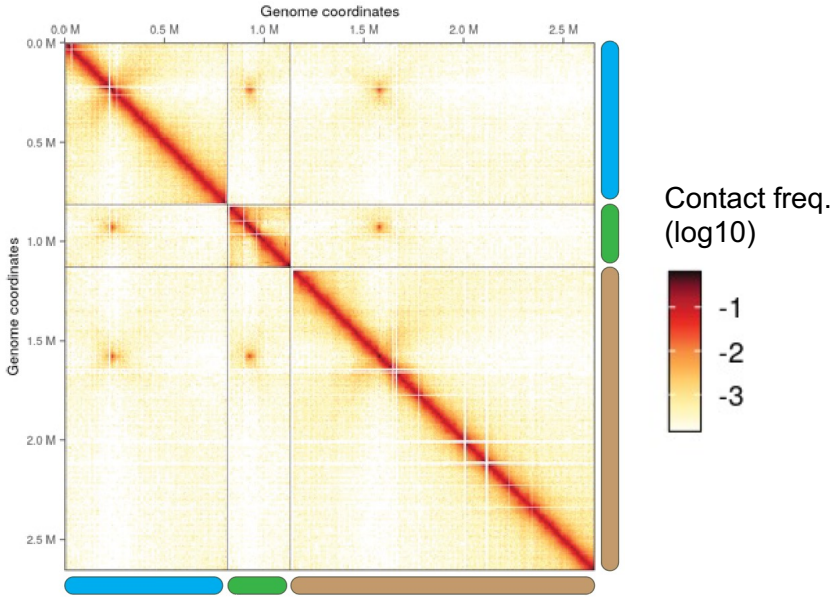
binned counts

binID_1	binID_2	counts
bin1	bin2	3
bin2	bin4	1
bin3	bin5	1
bin5	bin6	2
bin6	bin6	1

regions file

binID	chr	start	end	norm.
bin1	chr1	1	1000	<dbl>
bin2	chr1	1001	2000	...
bin3	chr1	2001	3000	
bin4	chr1	3001	3892	
bin5	chr2	1	1000	
bin6	chr2	1001	2000	
bin7	chr2	2001	2321	





Contact freq. (log10)

-1
-2
-3

2500 bins of 1kb

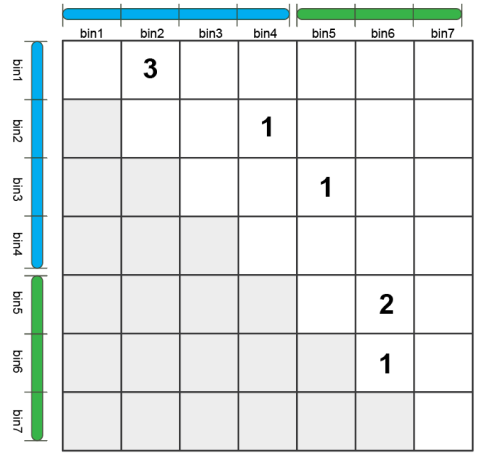
Chromosome Conformation Capture processing

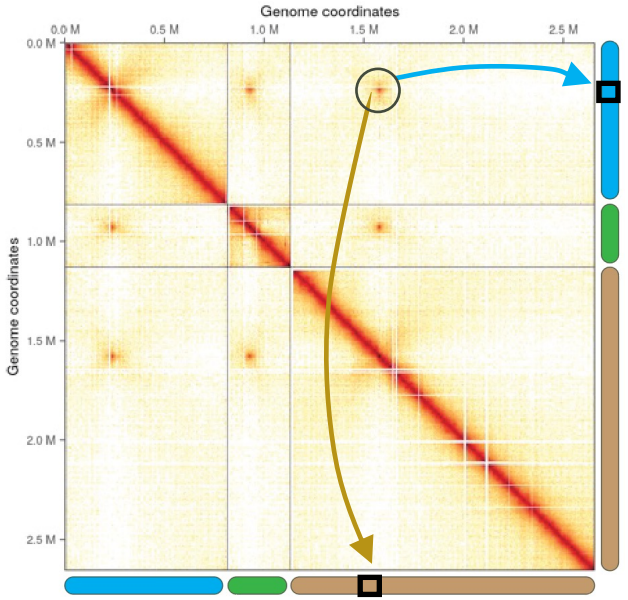
binned counts

binID_1	binID_2	counts
bin1	bin2	3
bin2	bin4	1
bin3	bin5	1
bin5	bin6	2
bin6	bin6	1

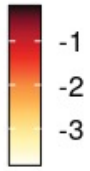
regions file

binID	chr	start	end	norm.
bin1	chr1	1	1000	<dbl>
bin2	chr1	1001	2000	...
bin3	chr1	2001	3000	
bin4	chr1	3001	3892	
bin5	chr2	1	1000	
bin6	chr2	1001	2000	
bin7	chr2	2001	2321	





Contact freq. (log10)



2500 bins of 1kb

Hi-C data file formats (1): human-readable files

- `.pairs`:
  - Stores interactions in (compressed) plain text
  - Lossless format
  - Human-readable
  - No random access supported by Bioconductor (could be implemented by the 4DN-provided `Rpairix` package)

- `HiC-Pro files`:
  - Stores binned matrices in two text files: a `regions.txt` and a `counts.txt`
  - Human-readable
  - Can take up significant space
  - Does not support normalization along with raw counts
  - No natively supported random access

Hi-C data file formats (2): binarized files

- `.cool`:
  - Stores binned matrices in a **binarized, HDF5-based** archive
  - Supports random access (`rhdf5lib`)
  - Supports multi-resolution matrices
  - Handles a **single** normalization vector (can be tweaked)
  - Well-documented, robust `cooler` API, **compatible** with `pairtools`, `cooltools` & `coolpuppy`

- `.hic`:
  - Stores binned matrices in a **binarized proprietary** format
  - Supports random access (`strawr`)
  - Supports multi-resolution matrices
  - Handles **multiple** normalization vectors
  - Somewhat documented, monolith `juicer tools` API.

Statement of need

- 3 independent formats for Hi-C binned matrices

  → **Lack of parsing methods in R, let alone a unified one**

- Many Hi-C packages already developed in R

  → **No dedicated Hi-C data structure**

- No basic arithmetic for Hi-C matrices

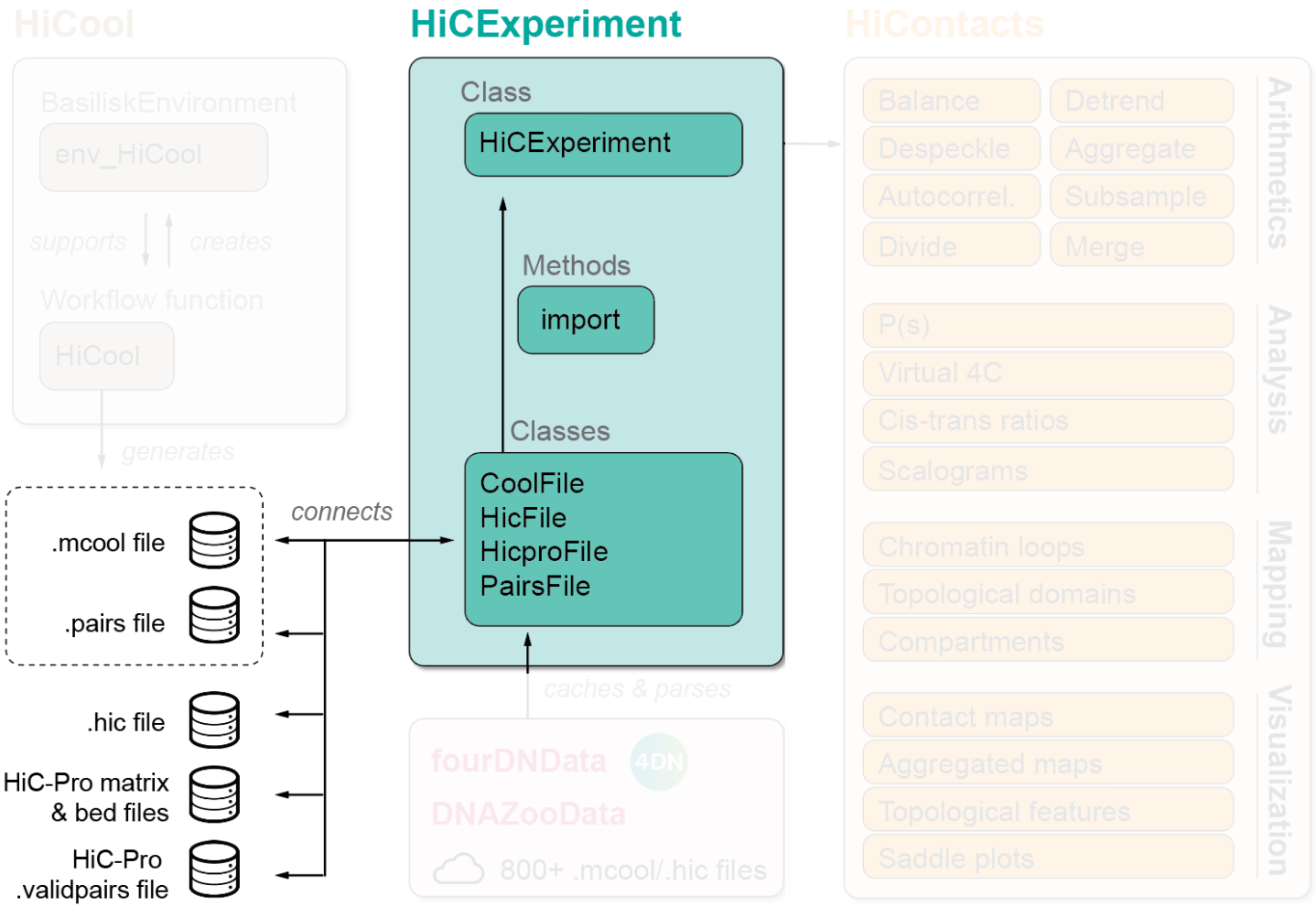
Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ **Introduction to the `OHCA` ecosystem**
- ❑ Importing Hi-C data with `HiCExperiment`
- ❑ Manipulating and visualizing Hi-C data with `HiContacts`
- ❑ Inter-operability with existing Hi-C packages
- ❑ Wrapping-up

The OHCA ecosystem

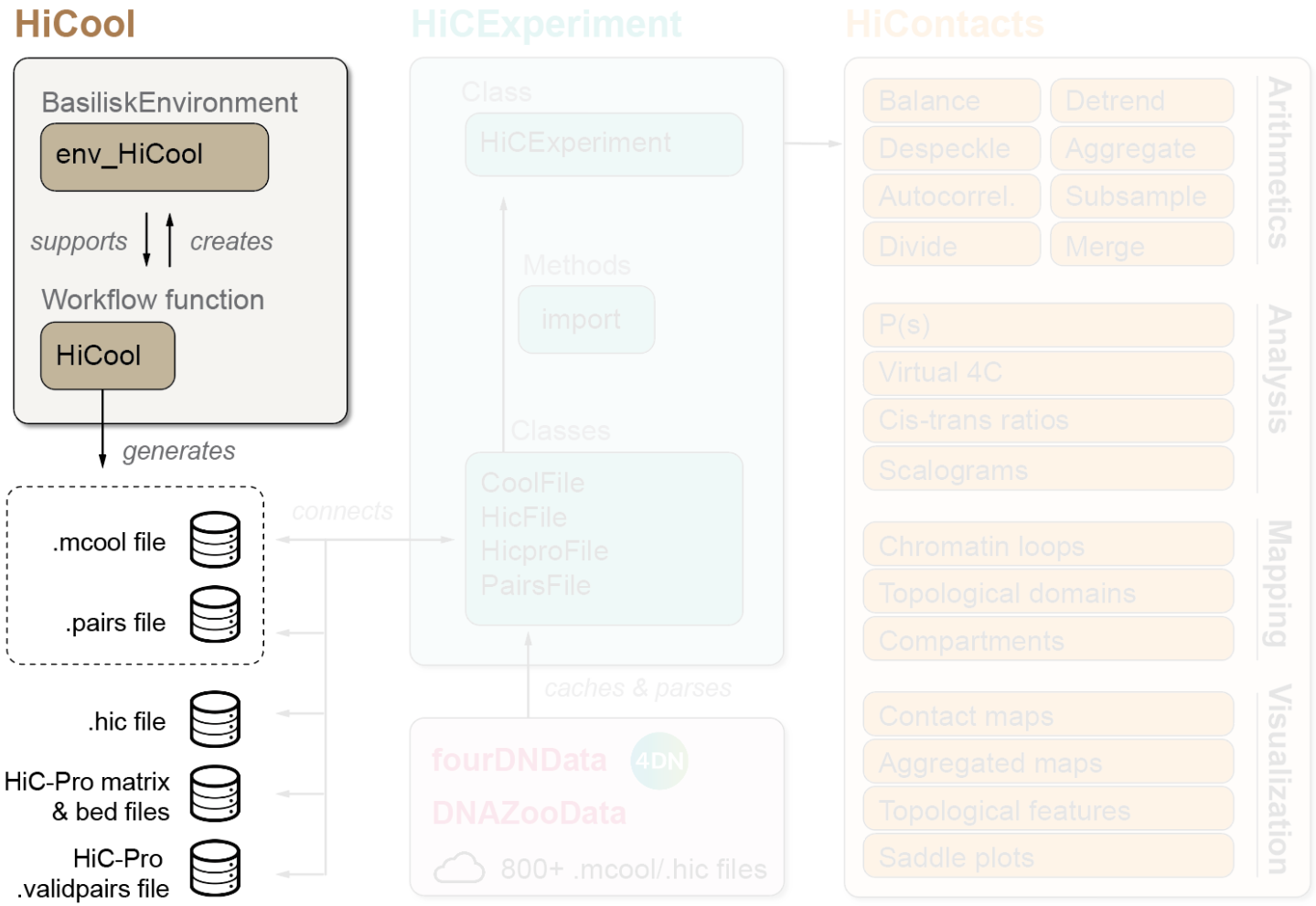
❖ HiCExperiment

- Defines the `HiCExperiment` class of object and associated methods, including subsetting of in memory Hi-C data
- Based on `Ginteractions` and `BiocFile` classes
- Import methods for all Hi-C data file formats
- Supports random access for `.hic` and `.cool` files



The OHCA ecosystem

- **HiCExperiment**
- **HiCool**
  - Processing fastq files into `.pairs` files and binned `.hic/.cool` files
  - Wraps `hicstuff`, a lightweight python Hi-C processing library
  - Based on `basilisk`

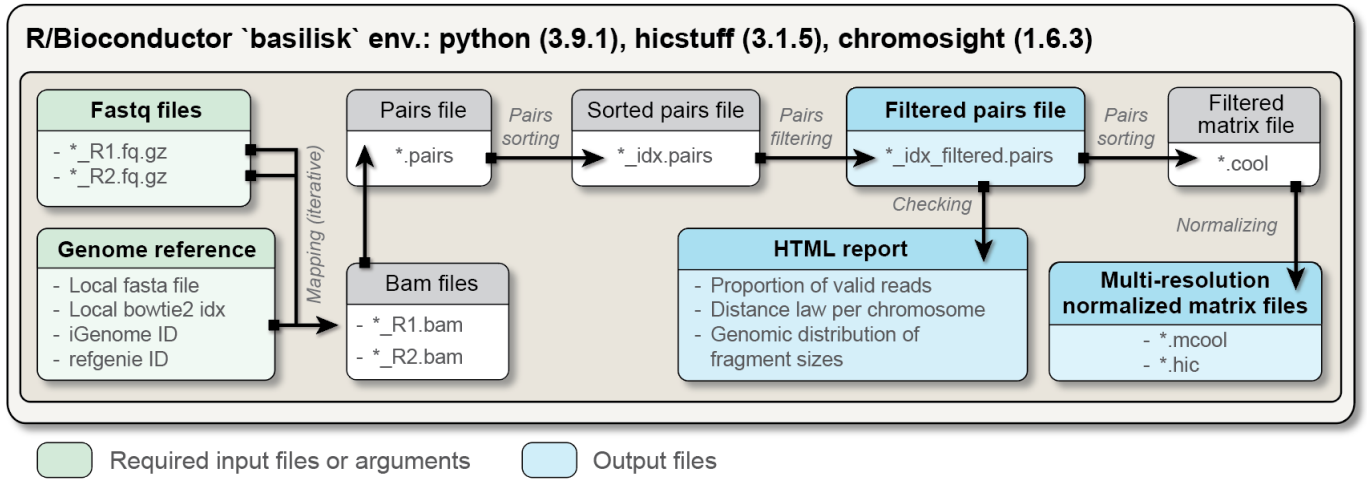


Matthey-Doret et al., Hi-C Data Analysis Ed. 2022

Aaron Lun, JOSS 2022

The OHCA ecosystem

- **HiCExperiment**
- **HiCool**
  - Processing fastq files into `.pairs` files and binned `.hic/.cool` files
  - Wraps `hicstuff`, a lightweight python Hi-C processing library
  - Based on `basilisk`

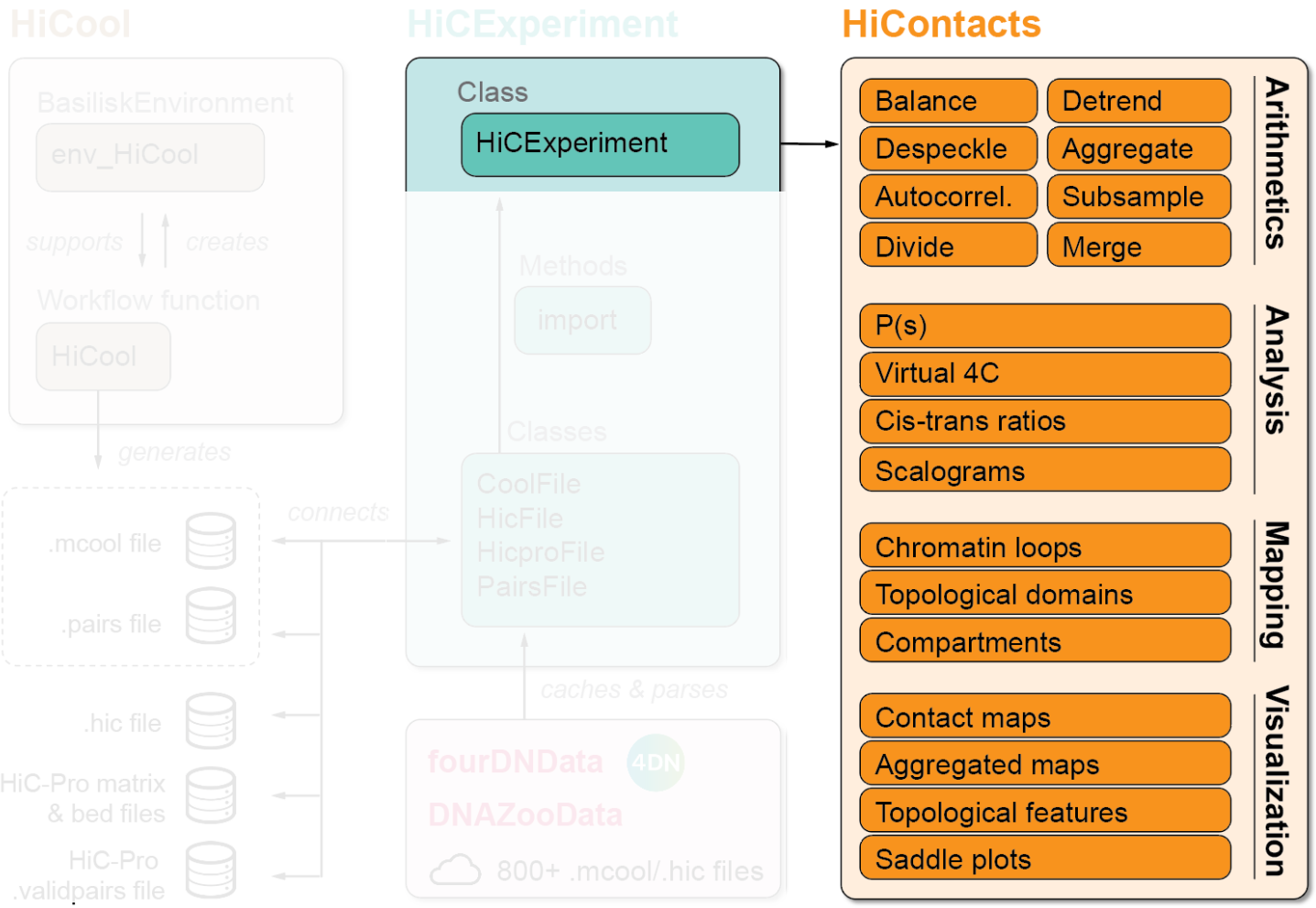


Matthey-Doret et al., Hi-C Data Analysis Ed. 2022

Aaron Lun, JOSS 2022

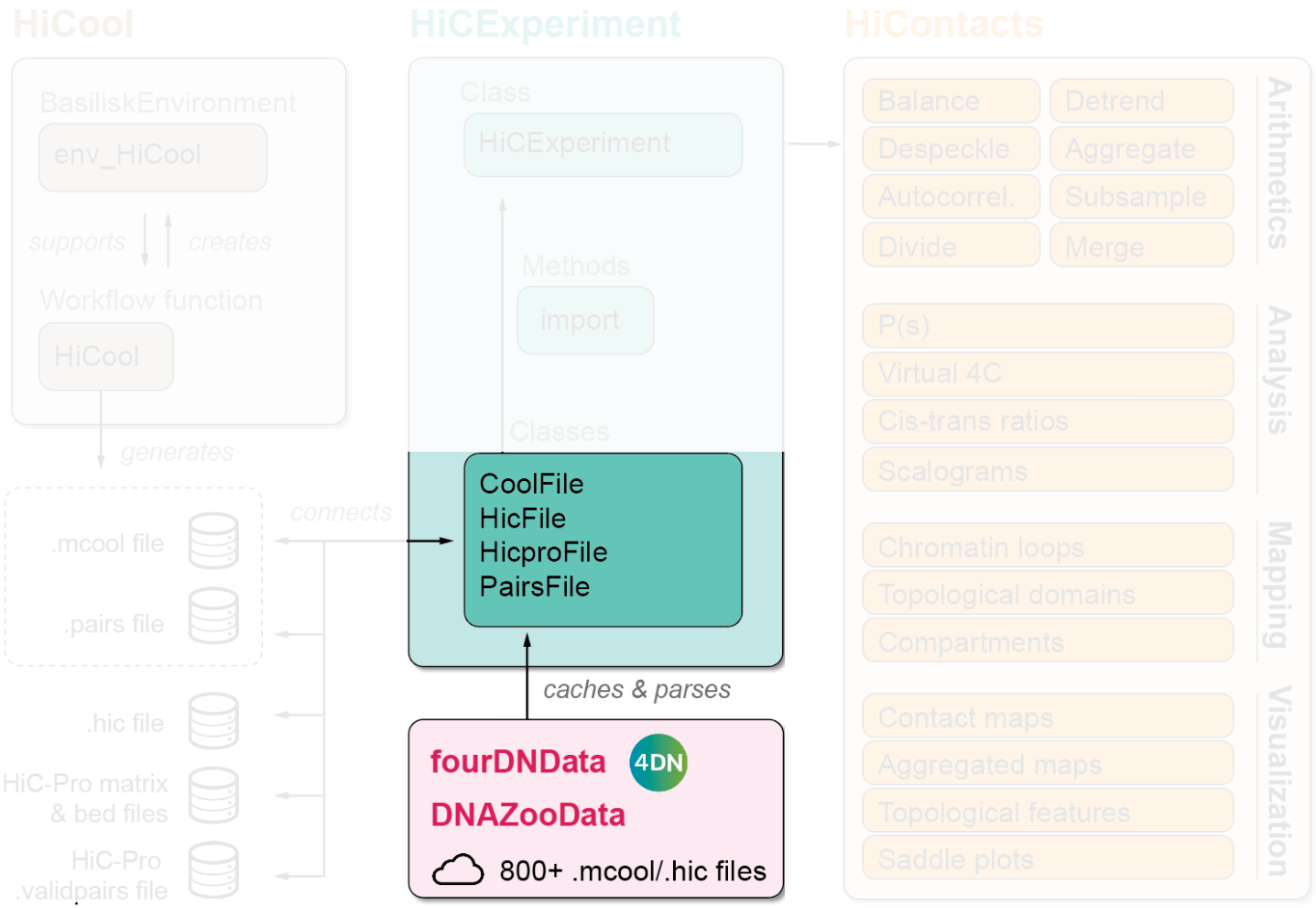
The OHCA ecosystem

- **HiCExperiment**
- **HiCool**
- **HiContacts**
  - Genomic arithmetic for Hi-C maps
  - Genomic interactions analytical tools
  - Implementations of algorithms to annotate 3D features (compartments, TADs, loops)
  - Extensive ggplot2-based visualization methods



The OHCA ecosystem

- **HiCool**
- **HiCExperiment**
- **HiContacts**
- **fourDNData / DNAZooData**
  - Data packages
  - Gateways to 2 major Hi-C consortia
  - `ExperimentHub`-based file accession
  - Support `BiocFileCache`

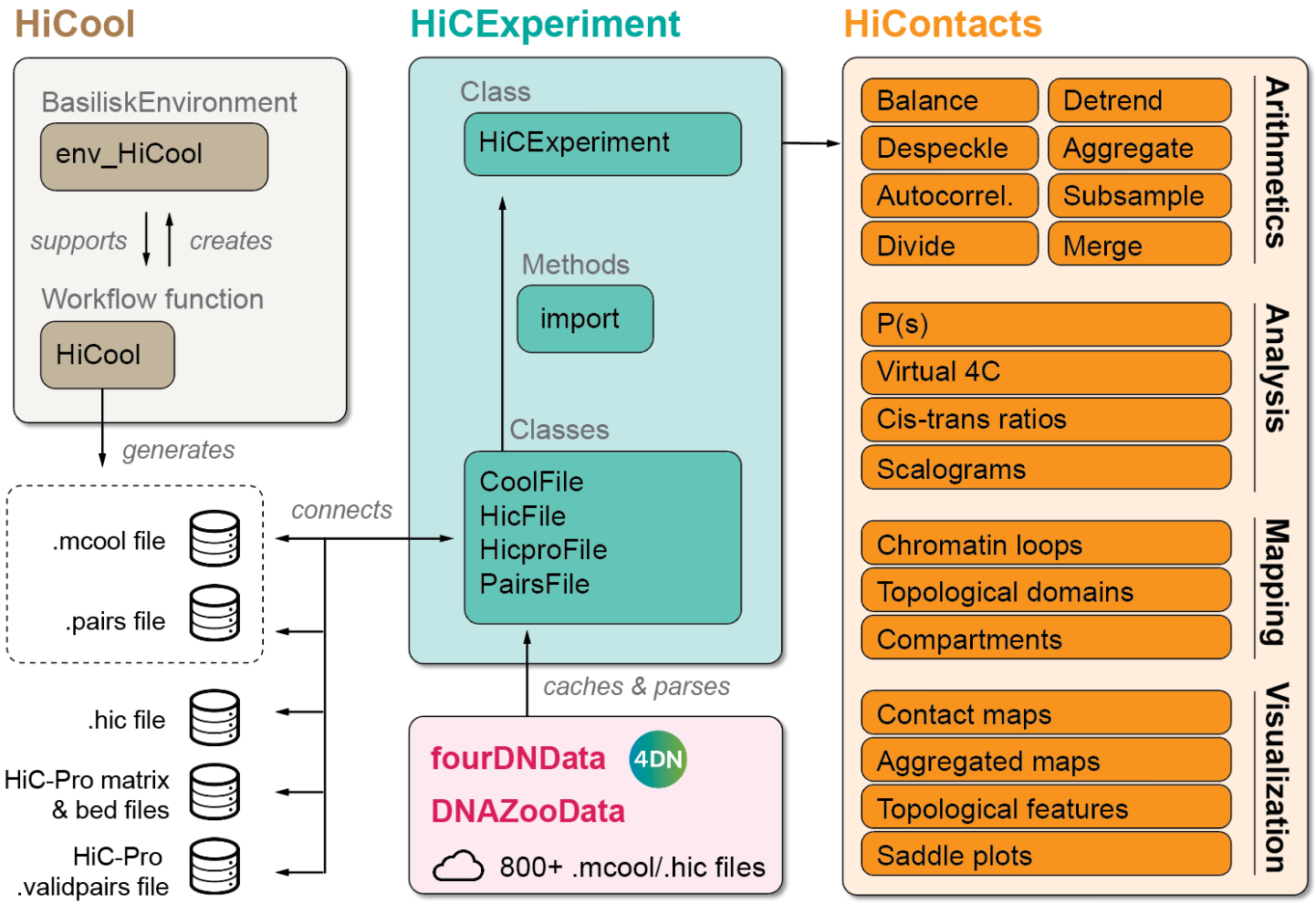


Reiff et al., Nat. Comm. 2022

Dudchenko et al., Science 2017

The OHCA ecosystem

- HiCool
- HiCExperiment
- HiContacts
- fourDNData / DNAZooData



Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ Introduction to the `OHCA` ecosystem
- ❑ **Importing Hi-C data with `HiCExperiment`**
- ❑ Manipulating and visualizing Hi-C data with `HiContacts`
- ❑ Inter-operability with existing Hi-C packages
- ❑ Wrapping-up

ContactFile et PairsFile: connection to disk-stored Hi-C data

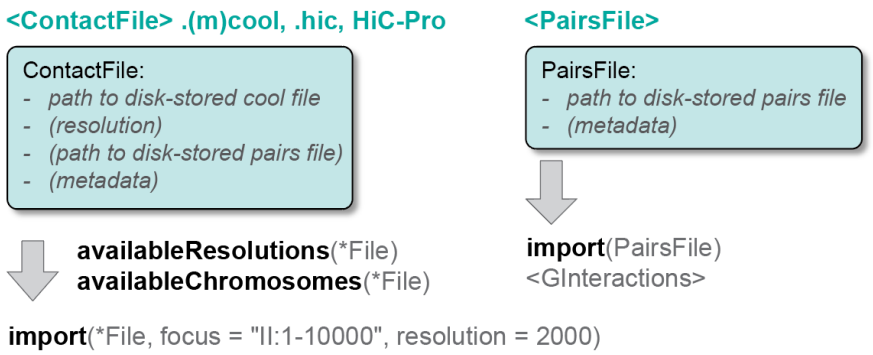
<ContactFile> .(m)cool, .hic, HiC-Pro

ContactFile:
- *path to disk-stored cool file*
- *(resolution)*
- *(path to disk-stored pairs file)*
- *(metadata)*

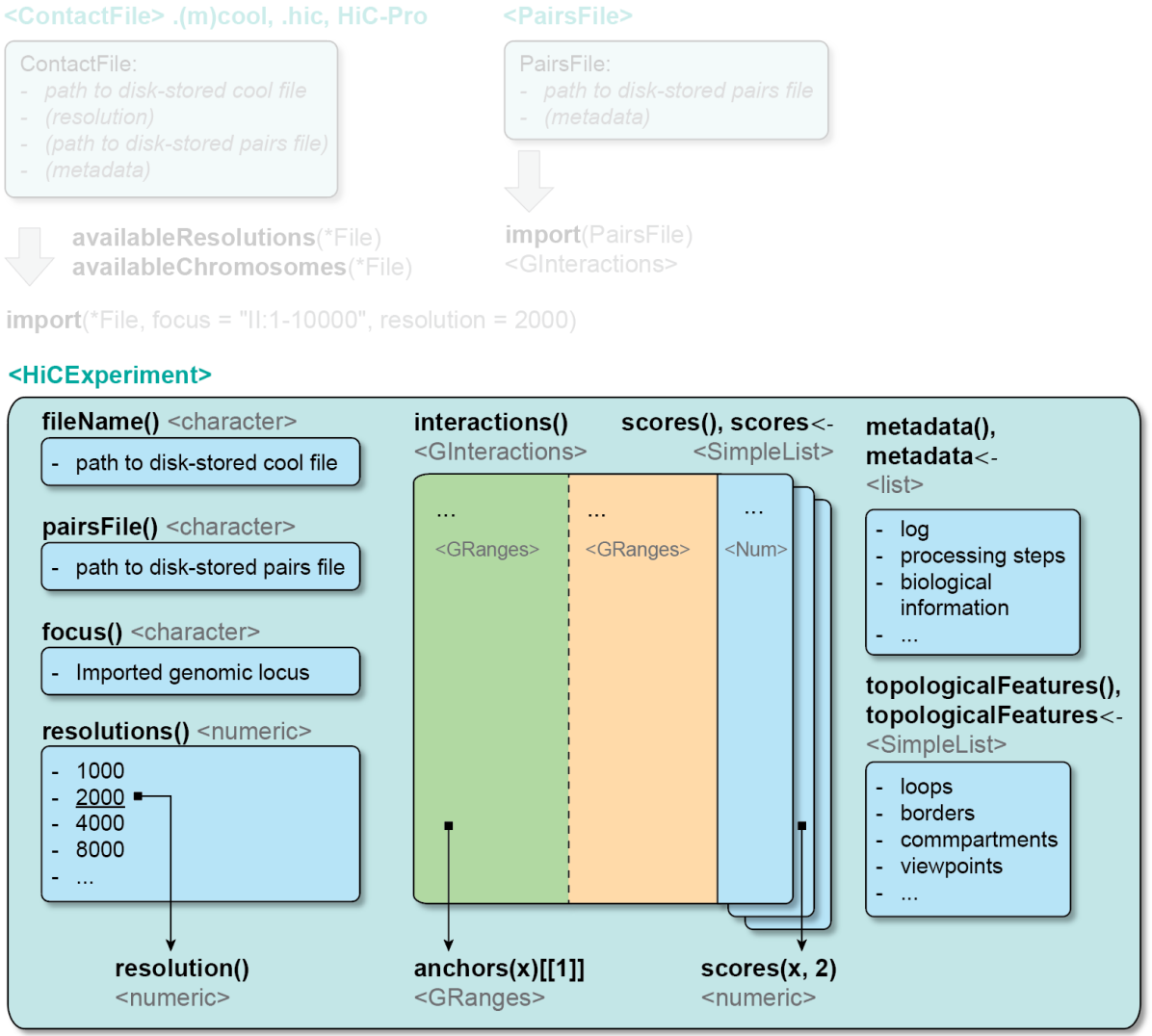
<PairsFile>

PairsFile:
- *path to disk-stored pairs file*
- *(metadata)*

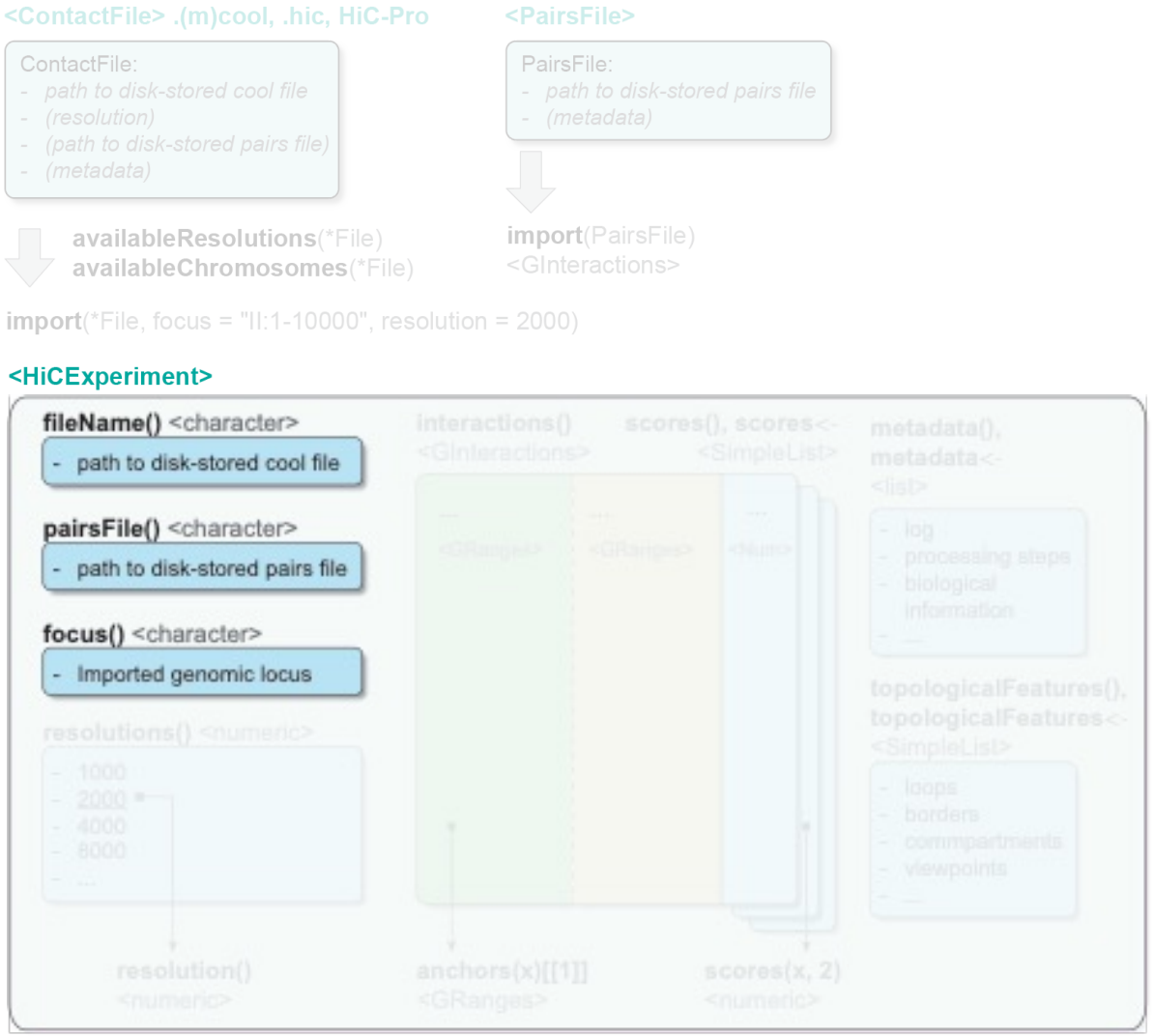
ContactFile et PairsFile: connection to disk-stored Hi-C data



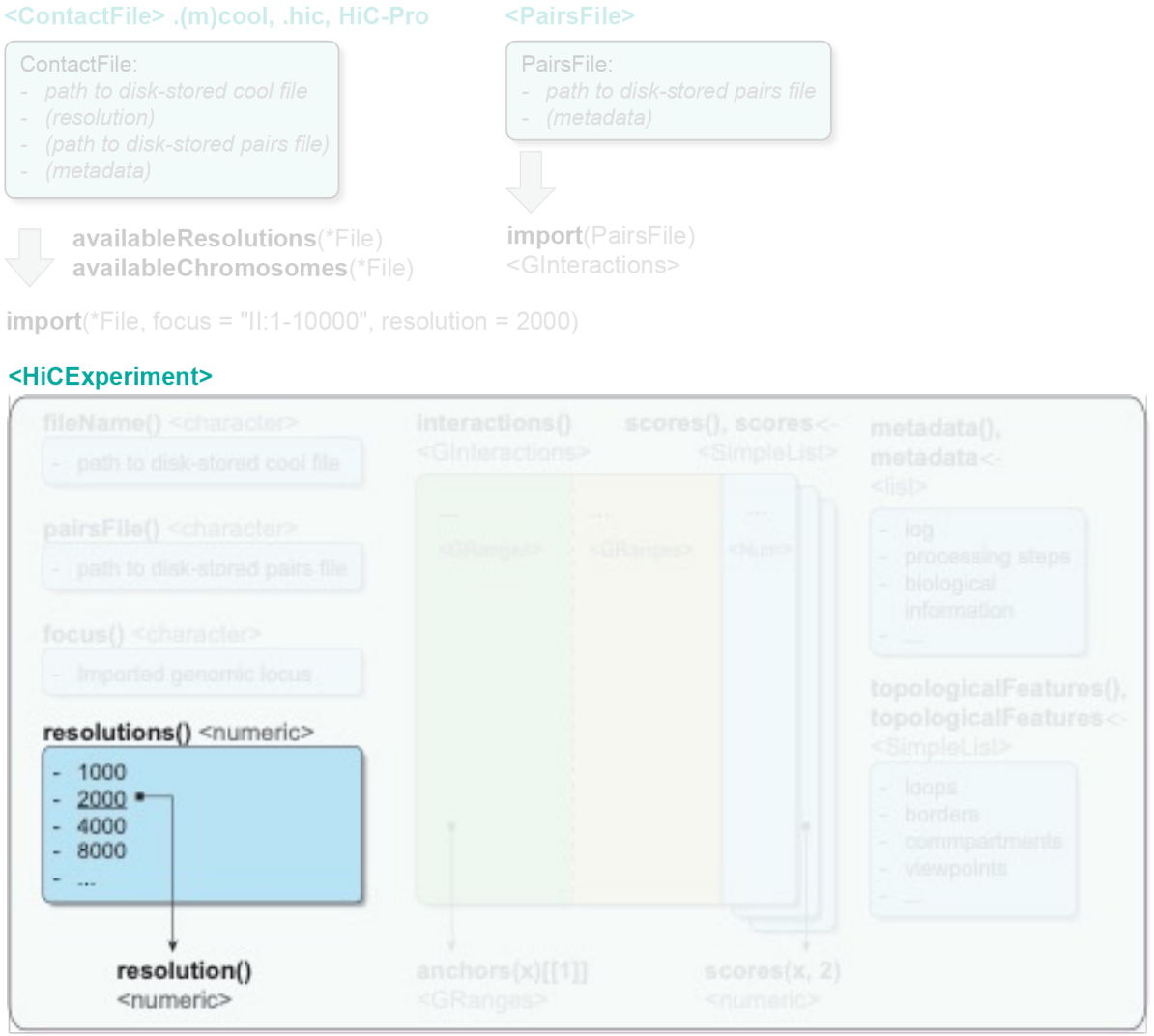
HiCExperiment class



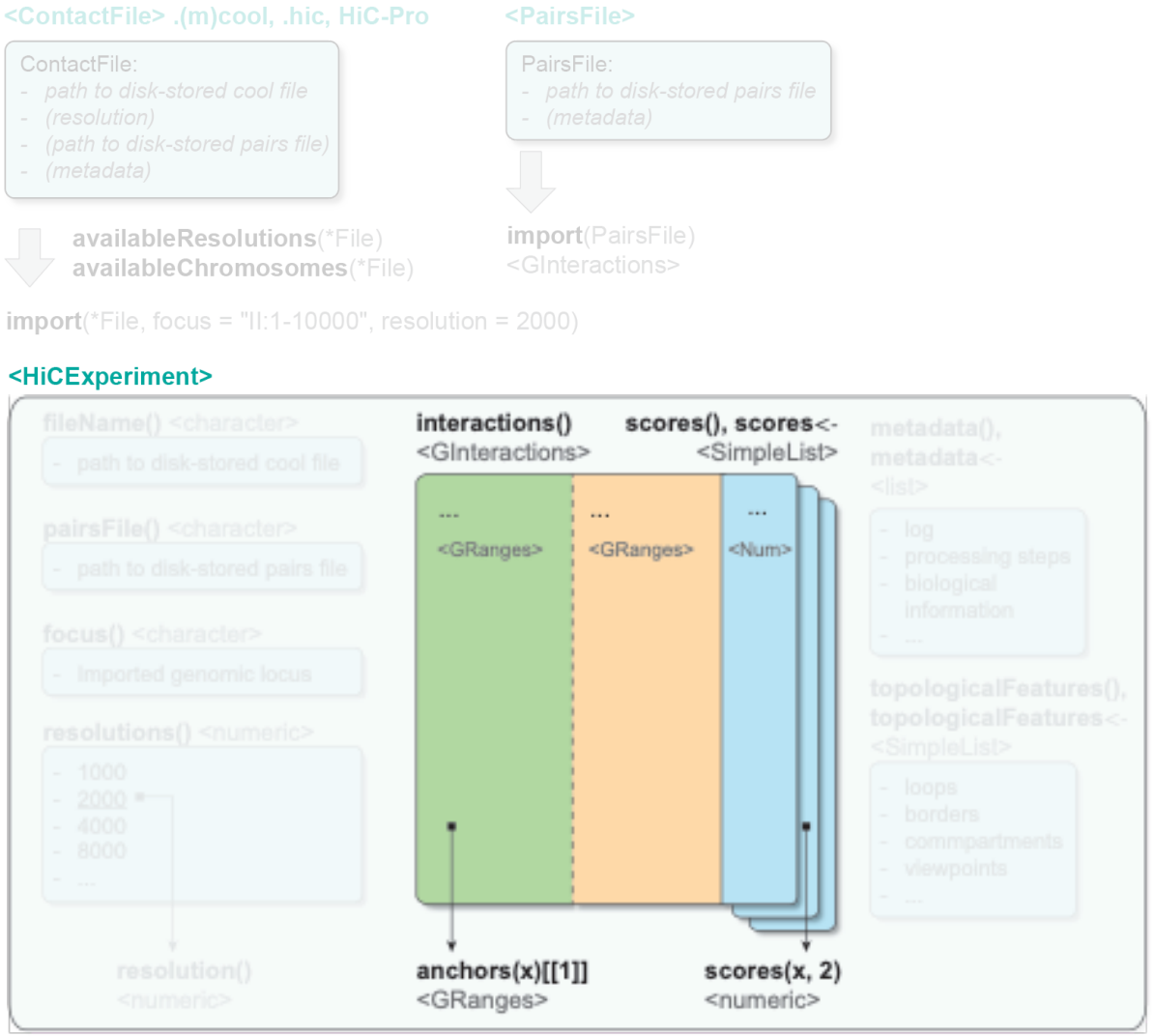
HiCExperiment class



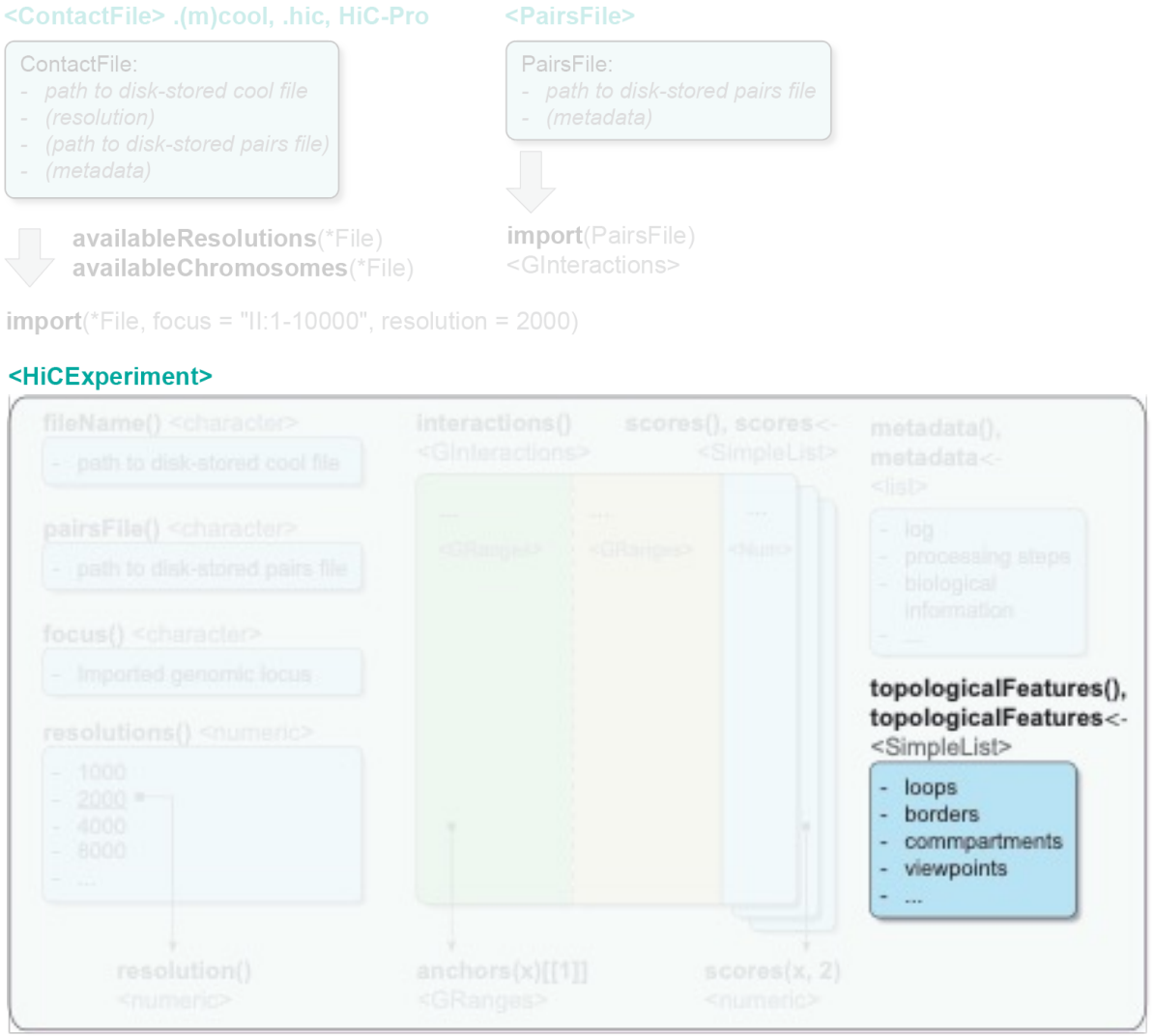
HiCExperiment class



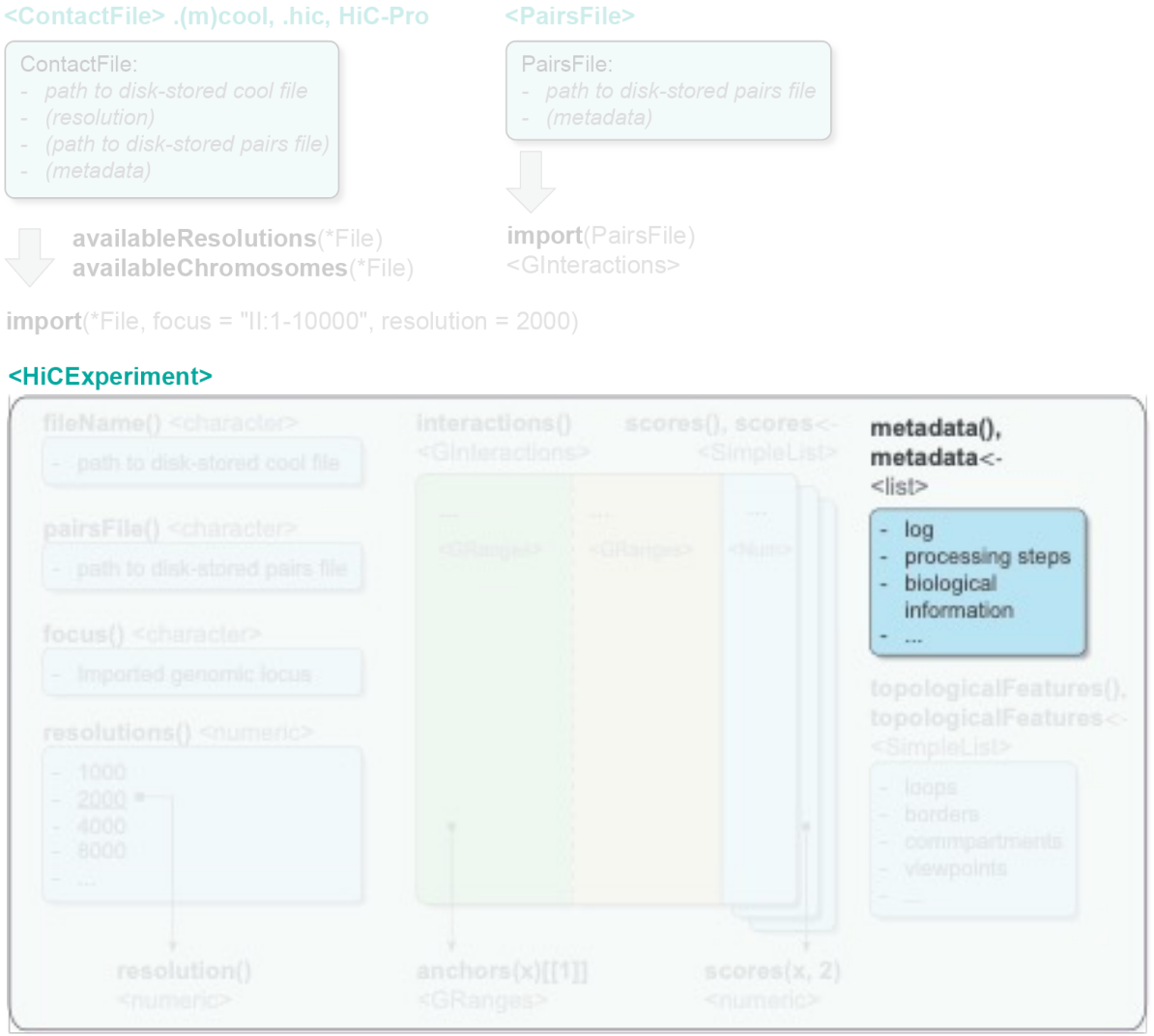
HiCExperiment class



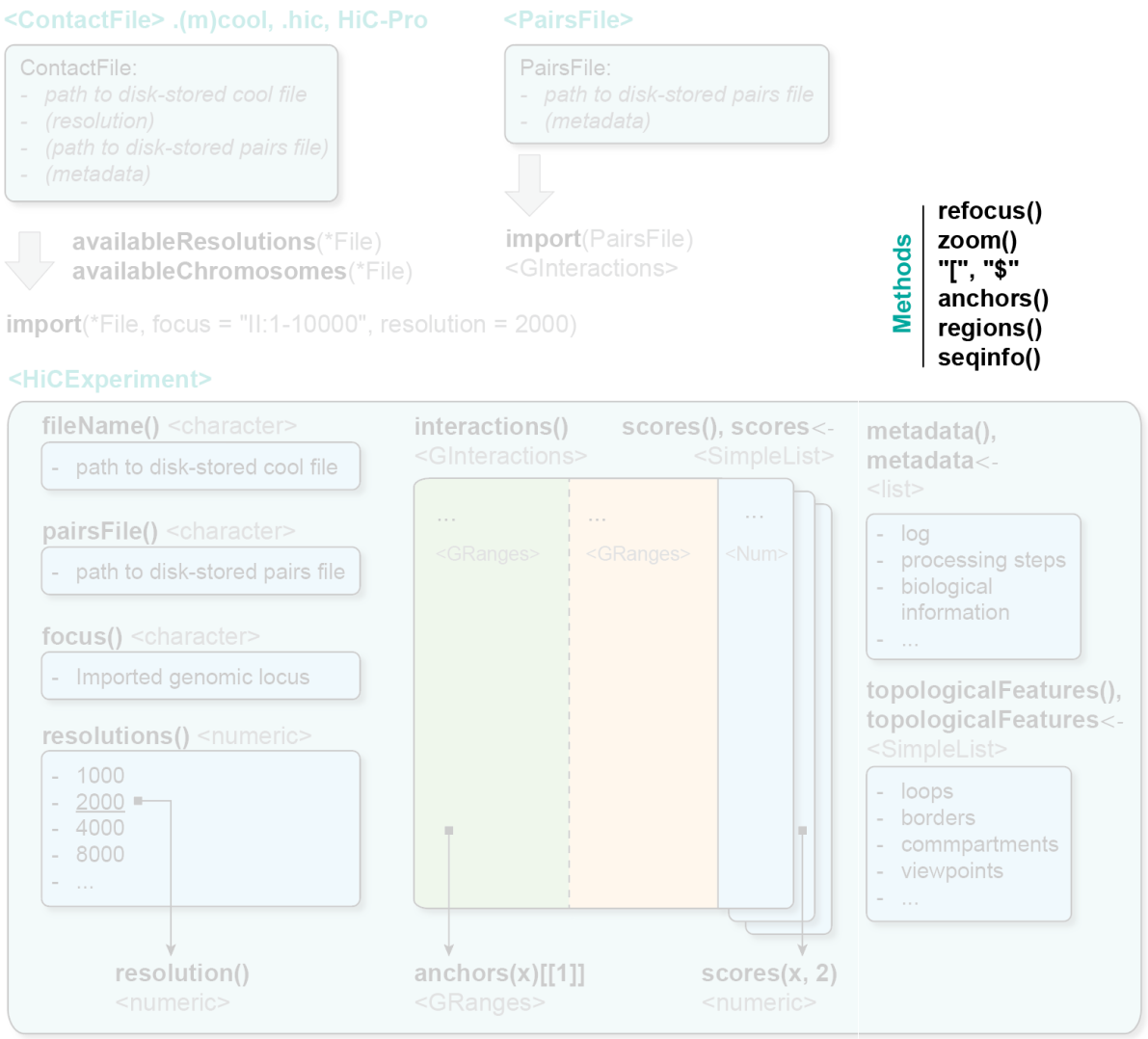
HiCExperiment class



HiCExperiment class



HiCExperiment class



Parsing Hi-C data in R

Subsetting before importing

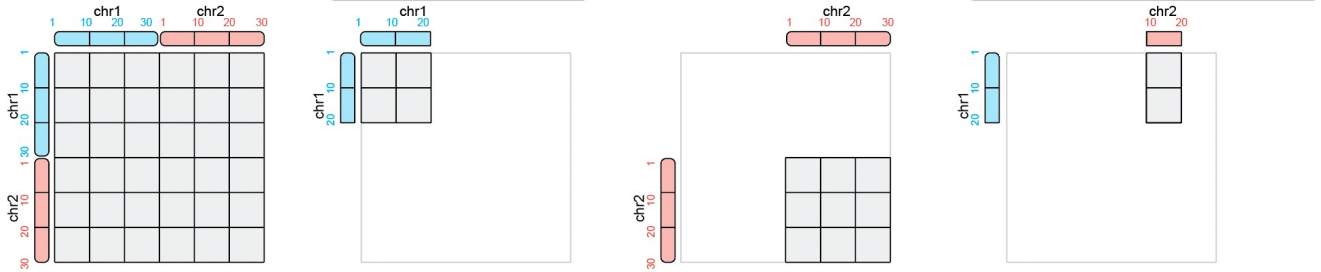
```
cool_file <- CoolFile("....mcool")
```

```
hic <- import(cool_file)
```

```
hic <- import(
  cool_file,
  focus = "chr1:1-20"
)
```

```
hic <- import(
  cool_file,
  focus = "chr2"
)
```

```
hic <- import(
  cool_file,
  focus = chr1:1-20|chr2:11-20"
)
```



Subsetting after importing

```
hic <- import(cool_file)
```

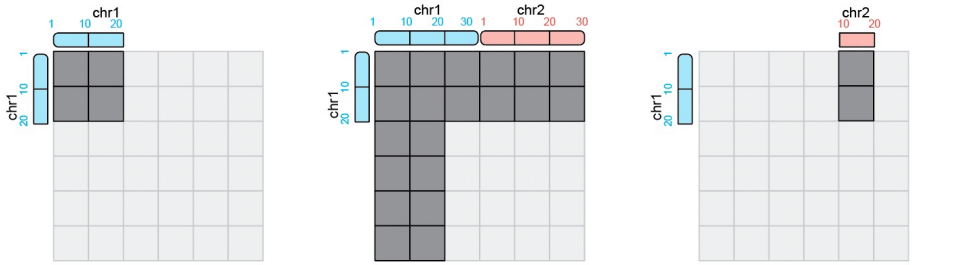
```
subsetByOverlaps(
  hic,
  GRanges("chr1:1-20"),
  type = "within"
)
```

```
hic["chr1:1-20"]
```

```
subsetByOverlaps(
  hic,
  GRanges("chr1:1-20"),
  type = "any"
)
```

```
subsetByOverlaps(
  hic,
  GInteractions(
    GRanges("chr1:1-20"),
    GRanges("chr2:11-20"))
)
```

```
hic["chr1:1-20|chr2:11-20"]
```



Demonstration: Parsing Hi-C data in R

Notebook:

https://js2264.github.io/OHCA.Bioc2023/parsing.html

Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ Introduction to the `OHCA` ecosystem
- ❑ Importing Hi-C data with `HiCExperiment`
- ❑ **Manipulating and visualizing Hi-C data with `HiContacts`**
- ❑ Inter-operability with existing Hi-C packages
- ❑ Wrapping-up

Visualizing Hi-C data in R

- `HiContacts::plotMatrix()`: a generic method to plot contact map from `(Aggr)HiCExperiment`, `GInteractions` and `matrix` objects
  - Argument `` `maxDistance = <int>` `` : to plot a horizontal Hi-C map, up to a certain distance from the diagonal
  - Argument `` `scale = c('linear', 'log10', 'exp0.2')` `` : to rescale the counts
  - Argument `` `limits = c(..., ...)` `` : to clamp the color scale (re-scaled counts) to specific minimum/maximum values
  - Argument `` `cmap = c(..., ..., ...)` `` : to specify colors used to create a gradient color scale
  - Argument `` `compare.to = <HiCExperiment>` `` : to plot 2 different `HiCExperiment` objects on each side of the diagonal

Demonstration: Visualizing Hi-C data in R

Notebook:

https://js2264.github.io/OHCA.Bioc2023/visualizing.html

Investigating Hi-C data in R (1)

- Arithmetic operations:
  - `HiContacts::normalize()`: Normalize the contact matrix for coverage
  - `HiContacts::detrend()`: Compute the observed/expected contact matrix
  - `HiContacts::autocorrelate()`: Compute the auto-correlation matrix
  - `HiContacts::despeckle()`: Smooth a contact matrix using a Gaussian blur
  - `HiContacts::divide()`: Divide two `HiCExperiment` objects
  - `HiContacts::merge()`: Merge multiple `HiCExperiment` objects together
  - `HiContacts::aggregate()`: Extract “snippets” from a `HiCExperiment` object and compute average signal

Demonstration: Investigating Hi-C data in R (1)

Notebook:

https://js2264.github.io/OHCA.Bioc2023/investigating.html

Investigating Hi-C data in R (2)

- Genomic interactions analysis:
  - `HiContacts::getPs()`: Compute the distance-dependent interaction frequency
  - `HiContacts::v4C()`: Compute the interaction profile of a genomic locus with the rest of the genome
  - `HiContacts::cisTransRatio()`: Compute the ratio of intra-chromosomal / inter-chromosomal interactions
  - `HiContacts::scalogram()`: Compute the distance-dependent interaction frequency for sliding windows along the genome

Demonstration: Investigating Hi-C data in R (2)

Notebook:

https://jserizay.com/OHCA.Bioc2023/investigating.html - interactions-analysis

Investigating Hi-C data in R (3)

- Annotation of topological features:
  - `HiContacts::getCompartments()`: Compute and phase the eigenvector E1 from a Hi-C contact matrix, and extract A/B compartments
  - `HiContacts::getDiamondInsulation()`: Compute the diamond insulation score along the diagonal of a Hi-C contact matrix, and extract significant borders
  - `HiContacts::getLoops()`: Annotate focal contact enrichment (a.k.a. loops) using the computer vision-based `chromosight` algorithm

These functions implement standard methods used to annotate compartments, domains and chromatin loops, but more advanced approaches exist and are available in R.

Lieberman-Aiden et al., Science 2009

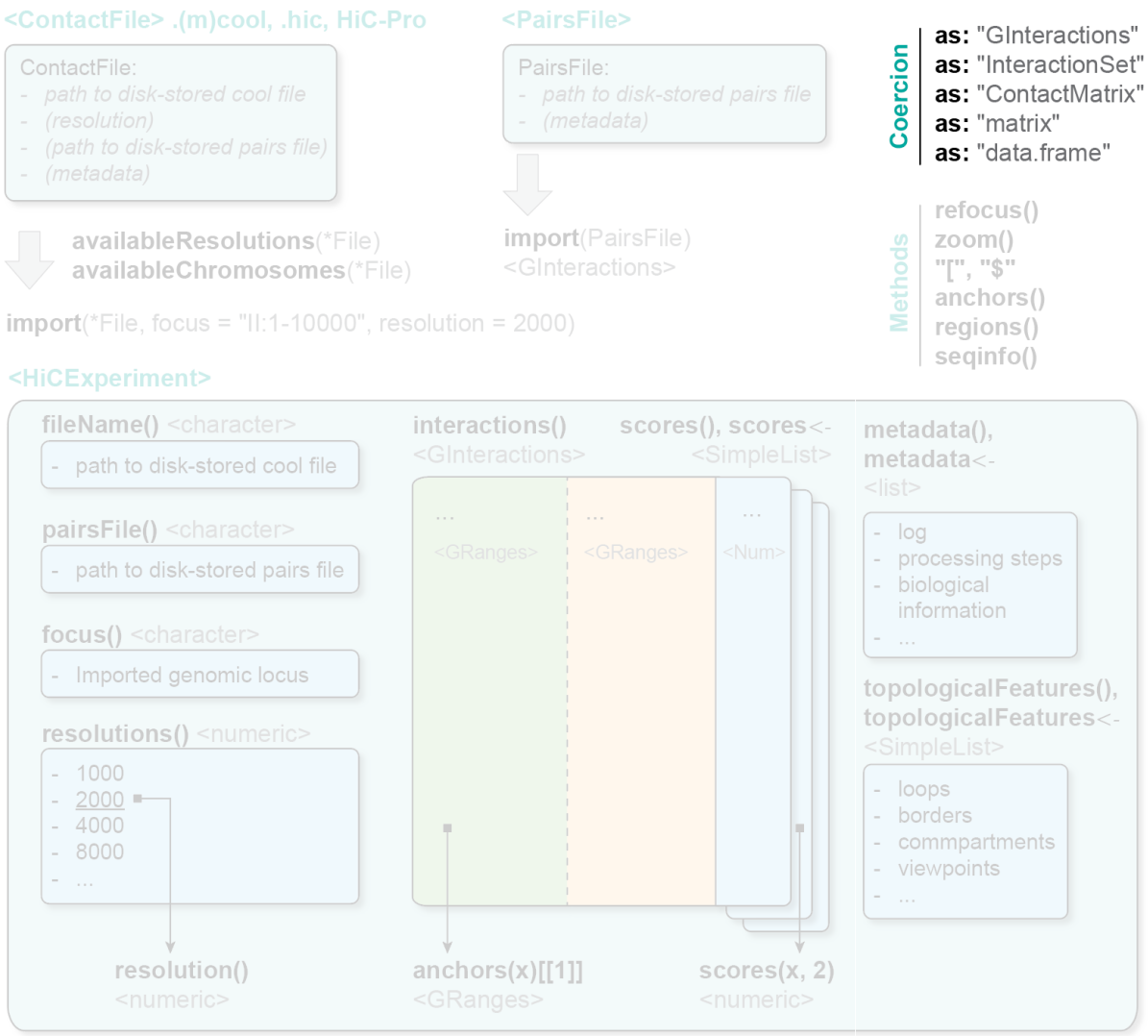
Crane et al., Nature 2015

Matthey-Doret et al., Nat. Comm. 2020

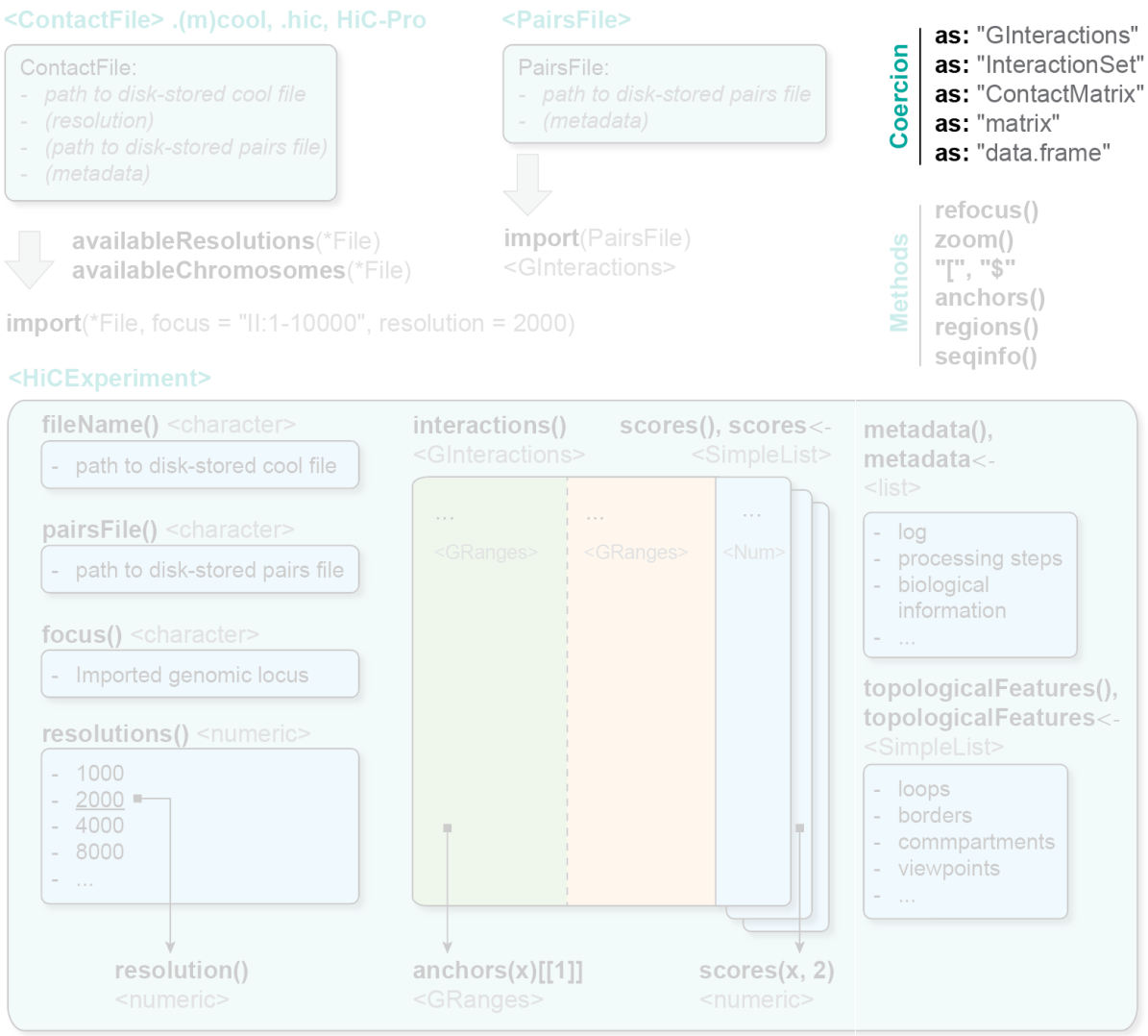
Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ Introduction to the `OHCA` ecosystem
- ❑ Importing Hi-C data with `HiCExperiment`
- ❑ Manipulating and visualizing Hi-C data with `HiContacts`
- ❑ **Inter-operability with existing Hi-C packages**
- ❑ Wrapping-up

Coercing HiCExperiment objects



Coercing HiCExperiment objects




```
hicrep
(multi)HiCcompare
TopDom
GOTHiC
HiCDCPlus
HiCDOC
…
```

Demonstration: Inter-operability with other packages

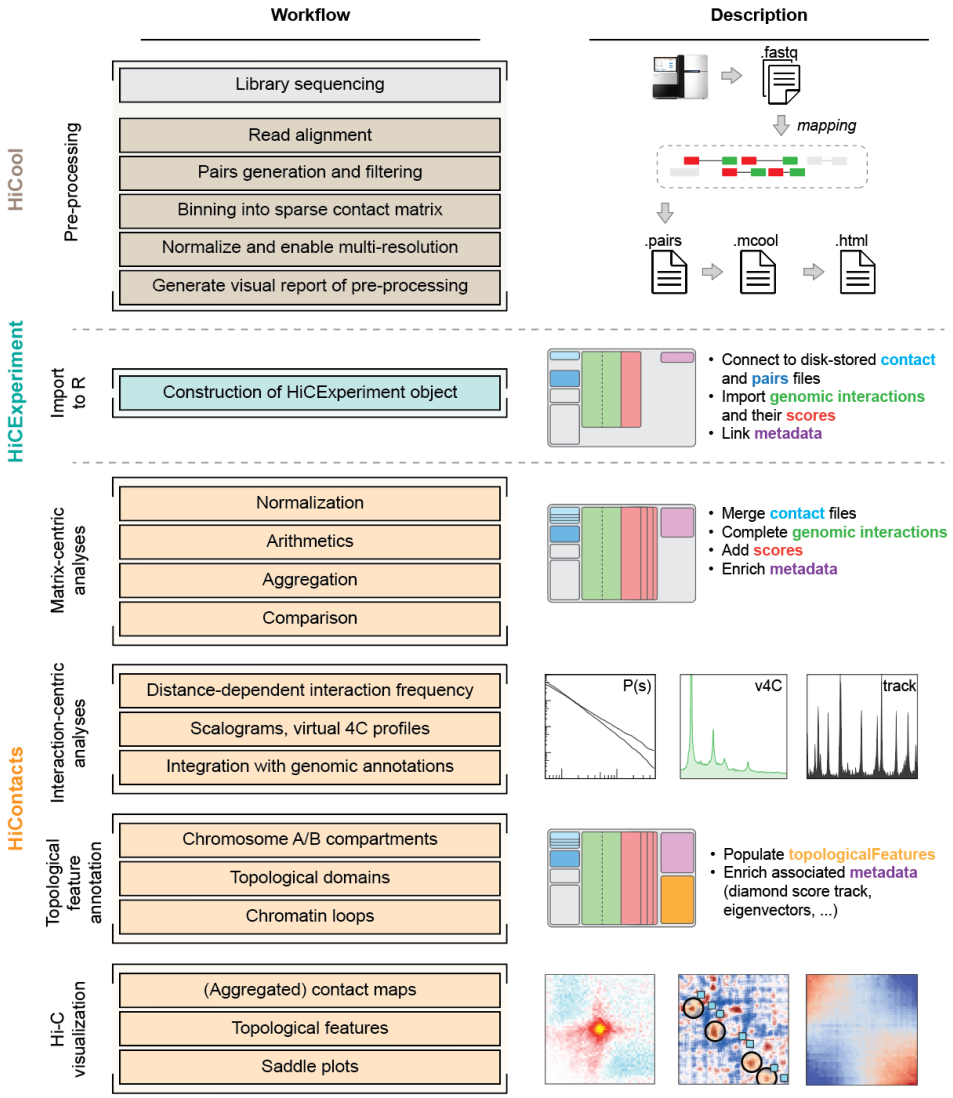
Notebook:

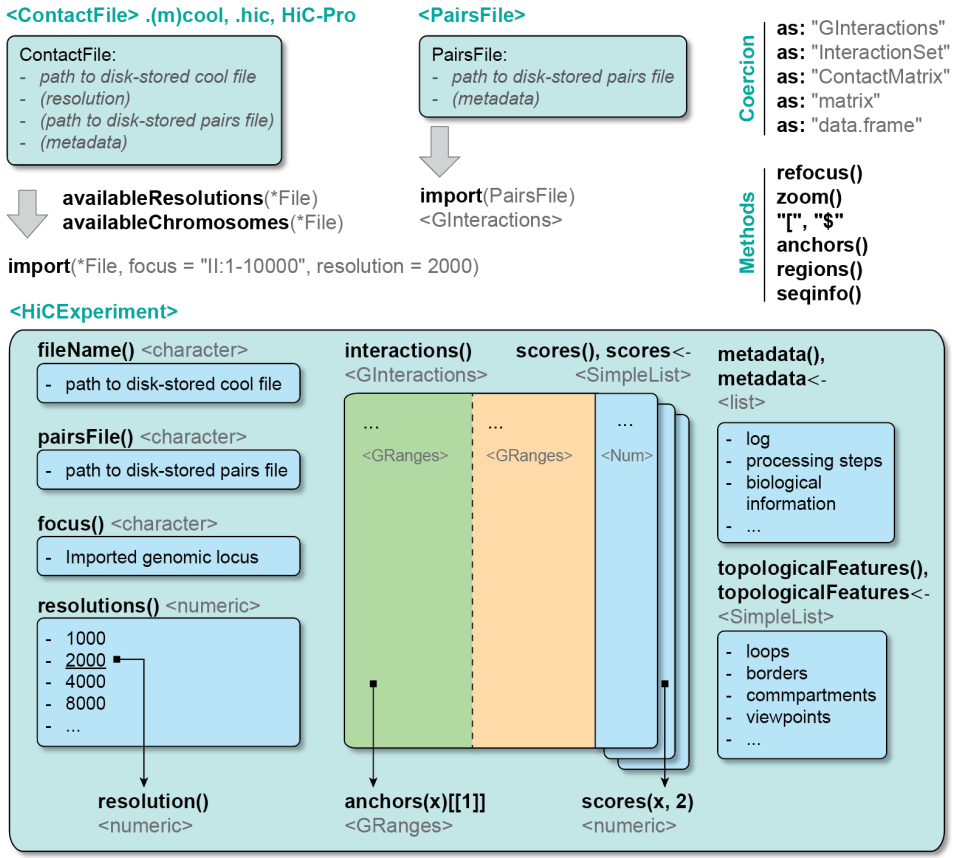
https://js2264.github.io/OHCA.Bioc2023/interoperability.html

Outline

- ❑ Overview of Chromosome Conformation Capture technical aspects
- ❑ Introduction to the `OHCA` ecosystem
- ❑ Importing Hi-C data with `HiCExperiment`
- ❑ Manipulating and visualizing Hi-C data with `HiContacts`
- ❑ Inter-operability with existing Hi-C packages
- ❑ **Wrapping-up**

Orchestrating Hi-C analysis with Bioconductor





Acknowledgments

- All core Bioconductor packages and infrastructure 👷👷

And more specifically:

- InteractionSet

Lun ATL, Perry M, Ing-Simmons E (2016). “Infrastructure for genomic interactions: Bioconductor classes for Hi-C, ChIA-PET and related experiments.” F1000Res., 5, 950.

- BiocIO

Morgan M, Lawrence M, Van Twisk D (2023). BiocIO: Standard Input and Output for Bioconductor Packages. R package version 1.10.0

Orchestrating Hi-C analysis with Bioconductor

- Book "*Orchestrating Hi-C analysis with Bioconductor*": https://js2264.github.io/OHCA/
- Package demo (Bioc2023) walkthrough: https://js2264.github.io/OHCA.Bioc2023/
- HiCExperiment & HiContacts repositories (& others): https://github.com/js2264/OHCA/