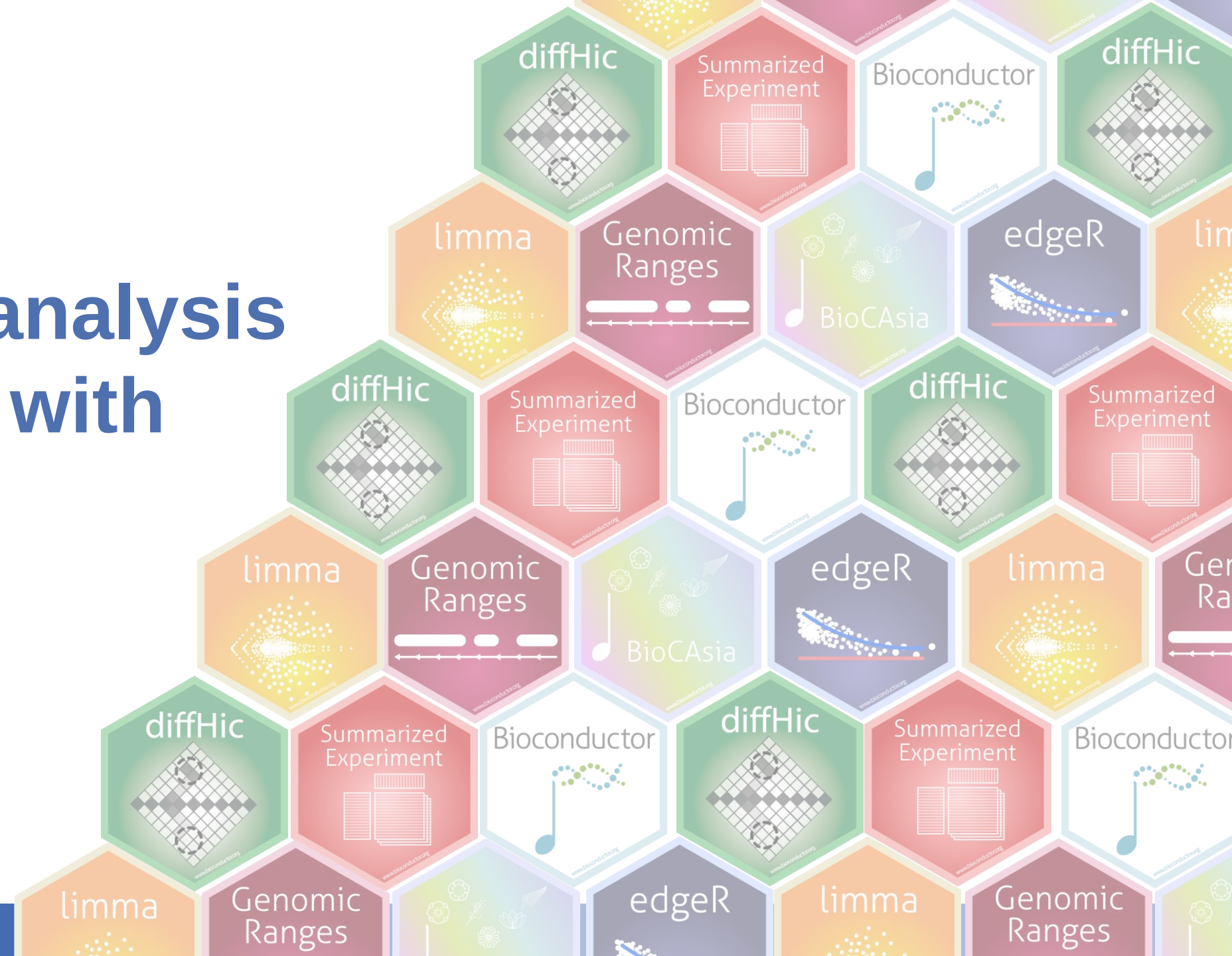
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Bioconductor
Hands-on Training Day:

Differential analysis of Hi-C data with diffHic



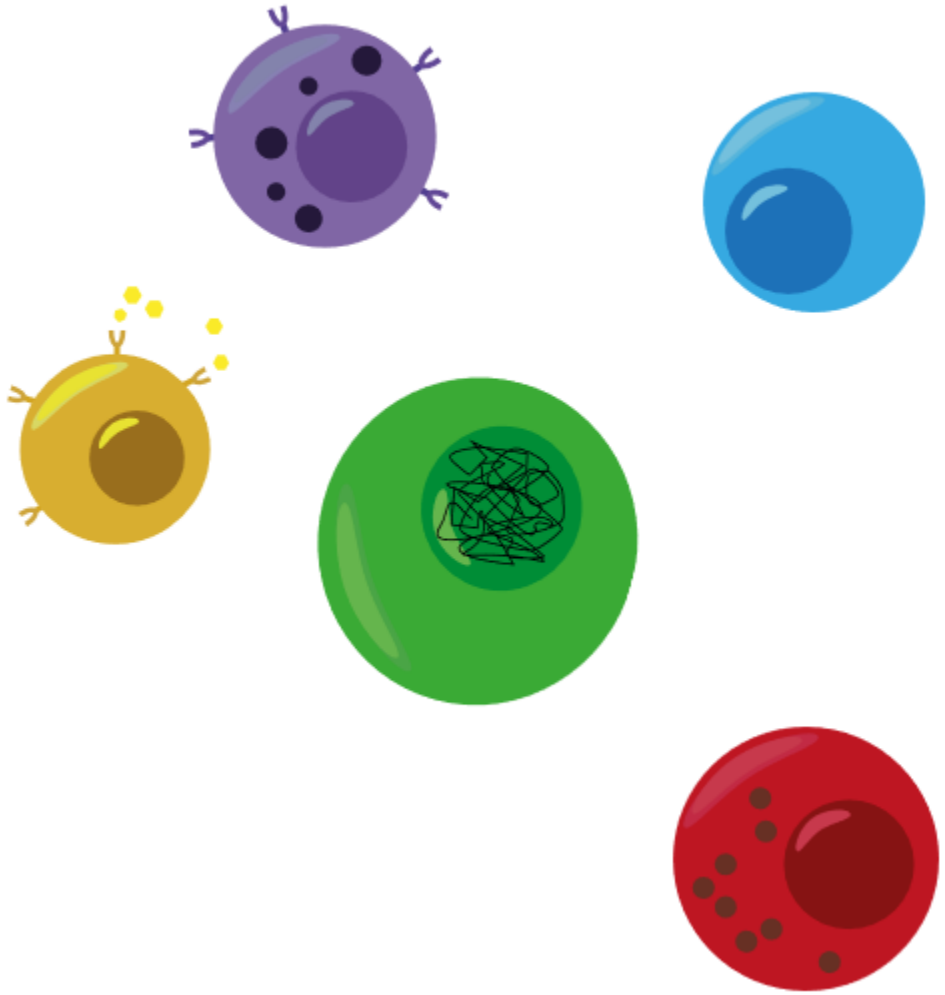
Hannah Coughlan
Hannah.Coughlan@wehi.edu.au
29th of November, 2018

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Outline

1. Introduction
    - Chromatin structure
    - HiC library construction
    - Analysis of HiC data

2. Tutorial: *diffHic* analysis of immune cell types



Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Resources

Lun and Smyth *BMC Bioinformatics* (2015) 16:258
DOI 10.1186/s12859-015-0683-0

BMC Bioinformatics

SOFTWARE Open Access

diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data

Aaron T.L. Lun[1,2] and Gordon K. Smyth[1,3*]

User: Aaron Lun

Reputation:	**21,160**
Status:	Trusted
Location:	Cambridge, United Kingdom
Scholar ID:	Google Scholar Page
Last seen:	27 minutes ago
Joined:	4 years, 2 months ago
Email:	i**********@gmail.com

I am a research associate in the field of computational biology at the Cancer Research UK Cambridge Institute in the United Kingdom. I am the author and maintainer of the csaw, diffHic, InteractionSet, scran, cydar, beachmat, DropletUtils, chipseqDB and simpleSingleCell packages; a co-author and co-maintainer of the scater, SingleCellExperiment and iSEE packages; a co-maintainer of the edgeR package; a co-author of the TENxBrainData package; and an occasional contributor to the limma package.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Bioconductor 3.9 » Software Packages » diffHic (development version)

diffHic

platforms all | rank 529 / 1636 | posts 2 / 1 / 0.5 / 0 | in Bioc 3.5 years
build warnings | updated before release

DOI: 10.18129/B9.bioc.diffHic
This is the **development** version of diffHic; for the stable release version, see diffHic.

Differential Analyis of Hi-C Data

Bioconductor version: Development (3.9)

Detects differential interactions across biological conditions in a Hi-C experiment. Methods are provided for read alignment and data pre-processing into interaction counts. Statistical analysis is based on edgeR and supports normalization and filtering. Several visualization options are also available.

Author: Aaron Lun [aut, cre], Gordon Smyth [aut]

Maintainer: Aaron Lun <infinite.monkeys.with.keyboards at gmail.com>

Citation (from within R, enter `citation("diffHic")`):

> Lun ATL, Smyth GK (2015). "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data." *BMC Bioinformatics*, **16**, 258.

Installation

To install this package, start R and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("diffHic", version = "3.9")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("diffHic")
```

PDF		diffHic Vignette
PDF		diffHicUsersGuide.pdf
PDF		Reference Manual
Text		NEWS

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

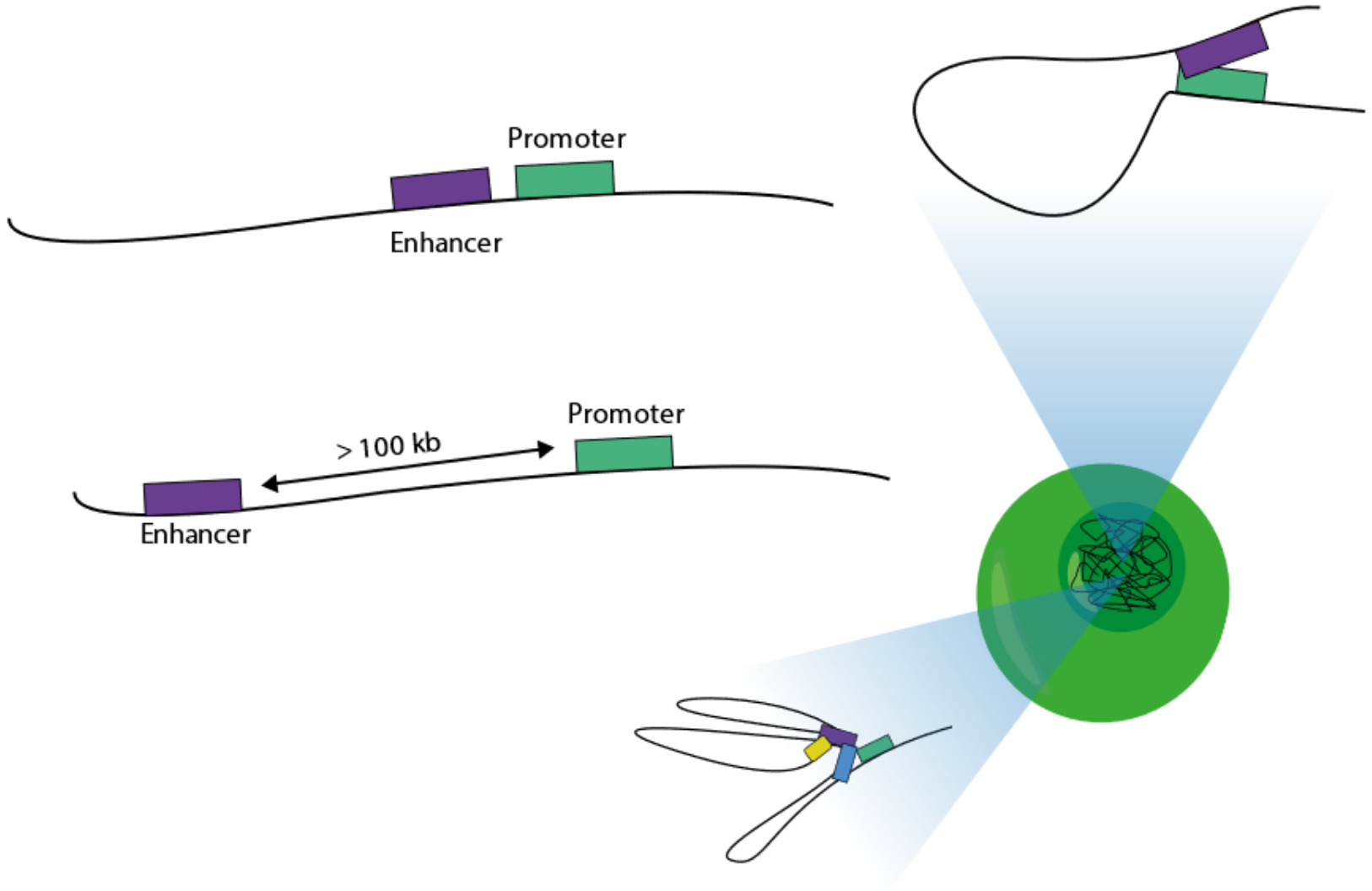
Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

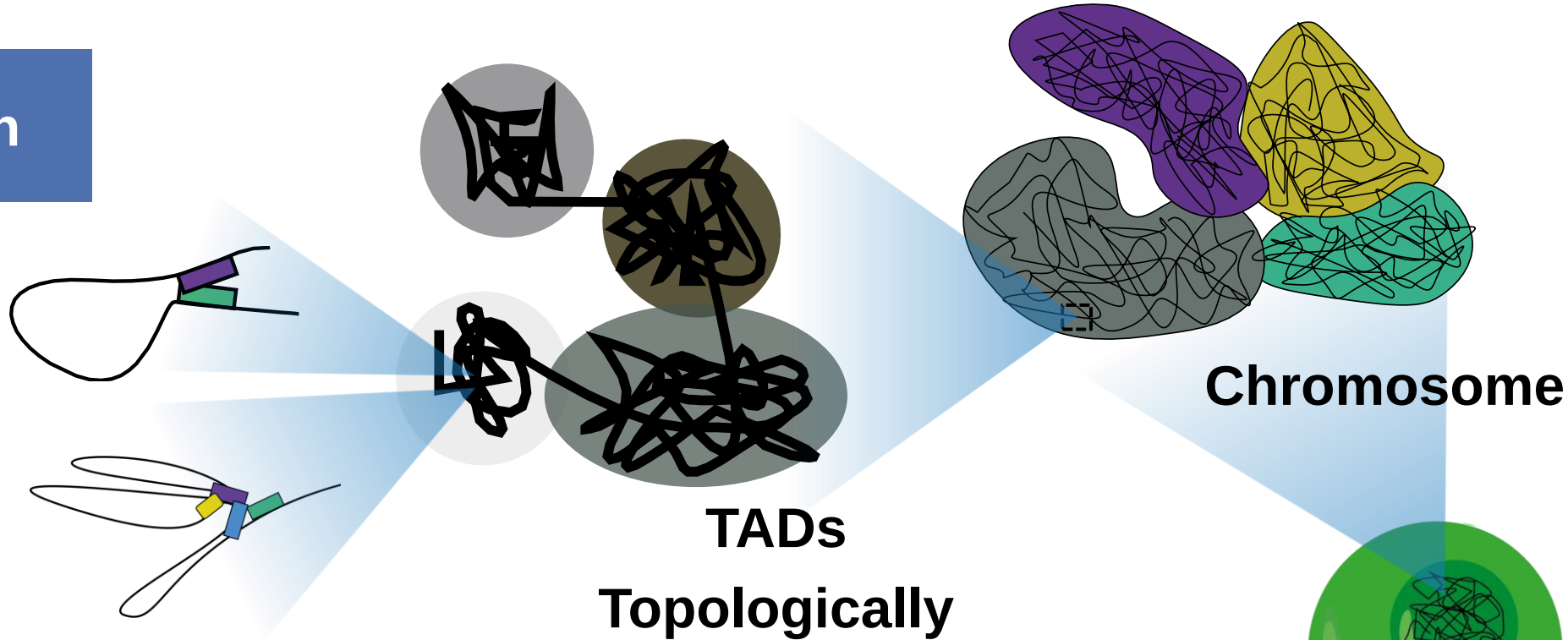
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Introduction



Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Introduction



Loops
Enhancer-promoter

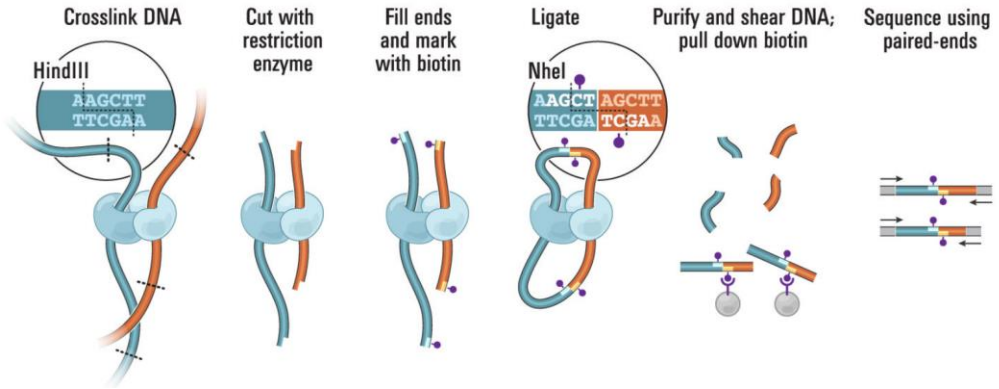
TADs
Topologically
associating domains

Chromosome

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

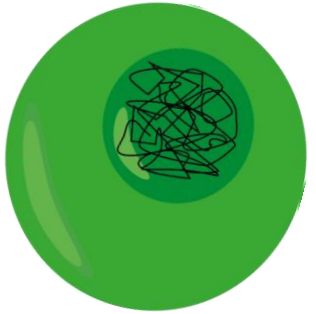
Hi-C

***Hi*gh-throughput *c*hromosome conformation capture**



- **The probability of contact between two loci should be governed by chance**

- **However, certain contacts occur far more often than expected by chance**

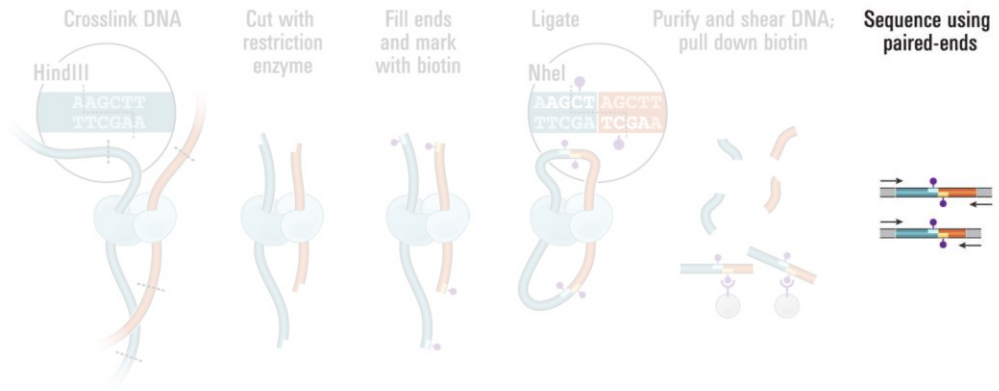


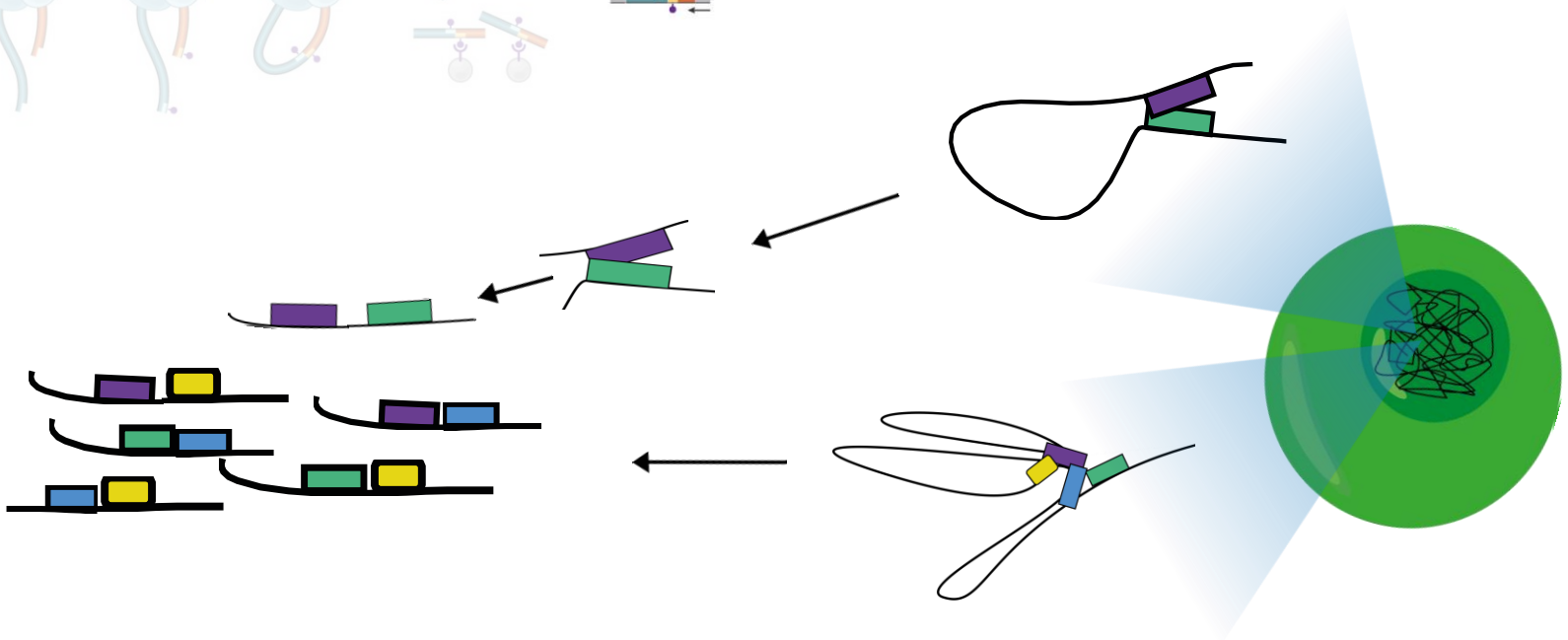
(Lieberman-Aiden et al, Science, 2009)

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Hi-C

***Hi*gh-throughput *c*hromosome conformation capture**





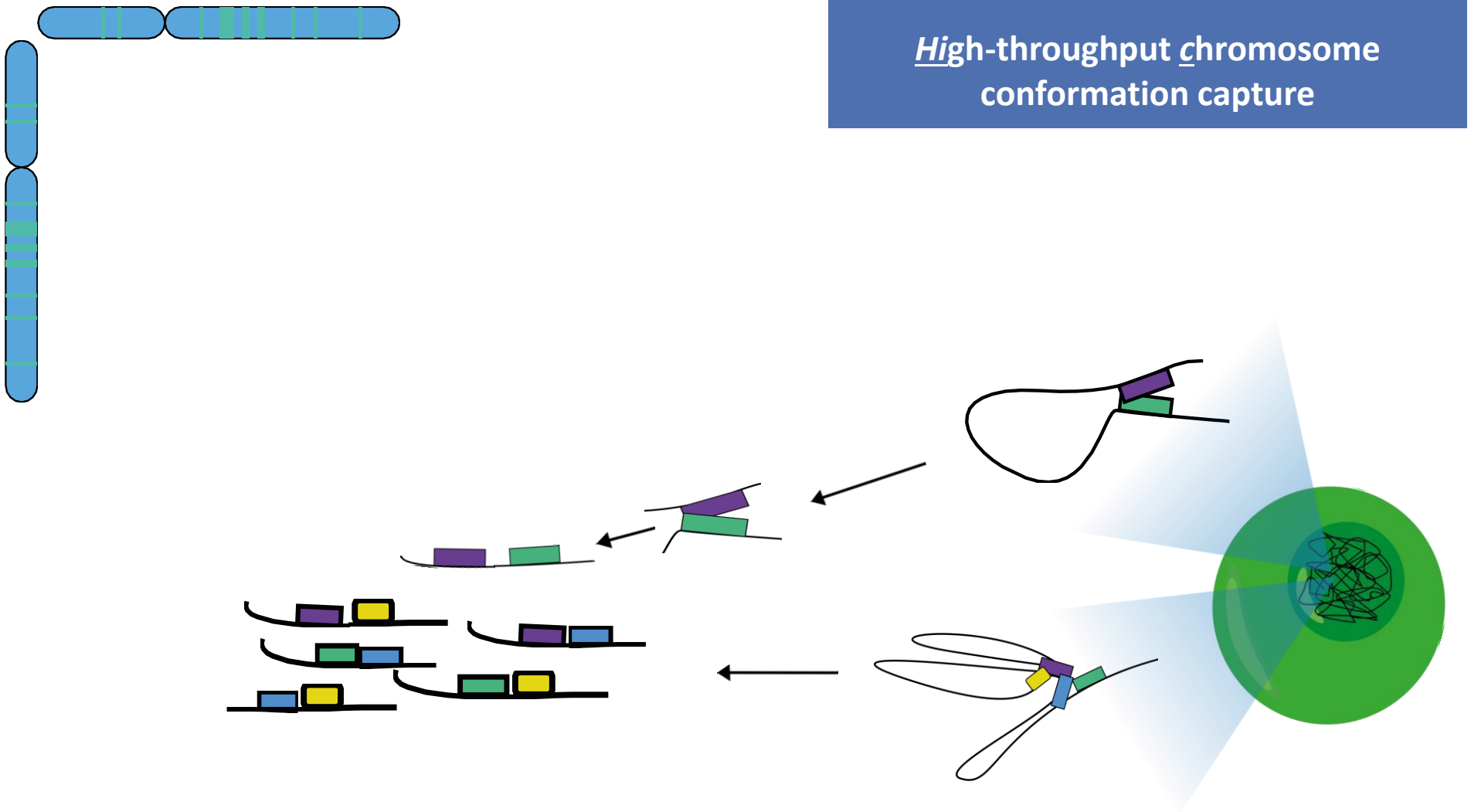
diffHic pipeline starts here with fastq

(Lieberman-Aiden et al, Science, 2009)

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Hi-C

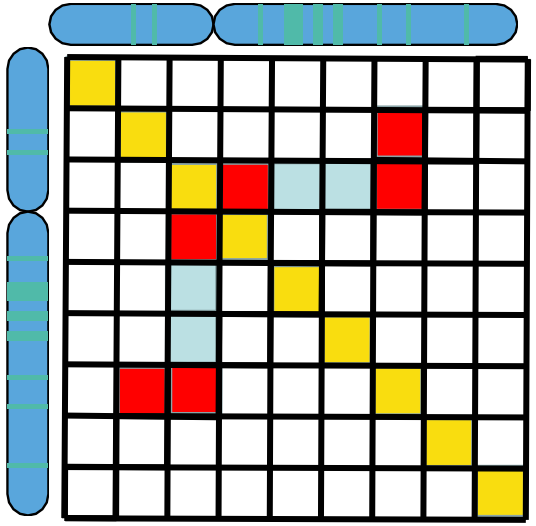
***Hi*gh-throughput *c*hromosome conformation capture**

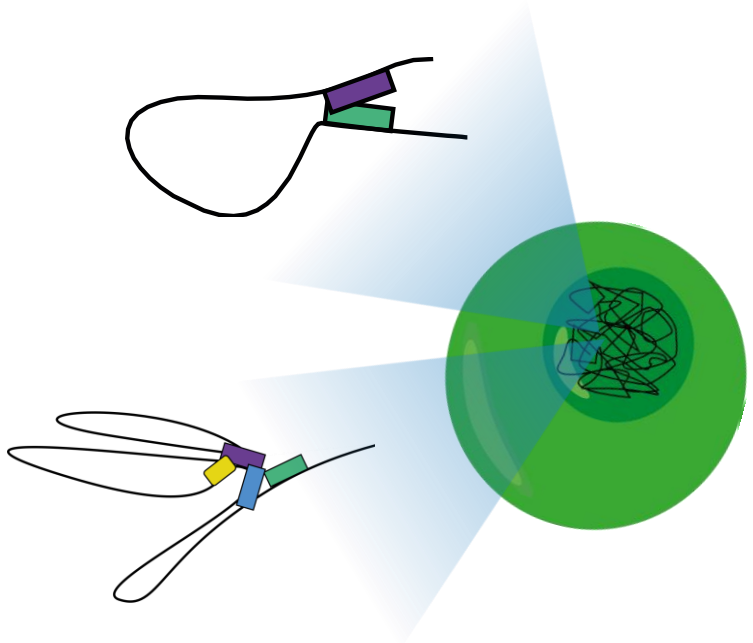


Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Hi-C

***Hi*gh-throughput *c*hromosome conformation capture**

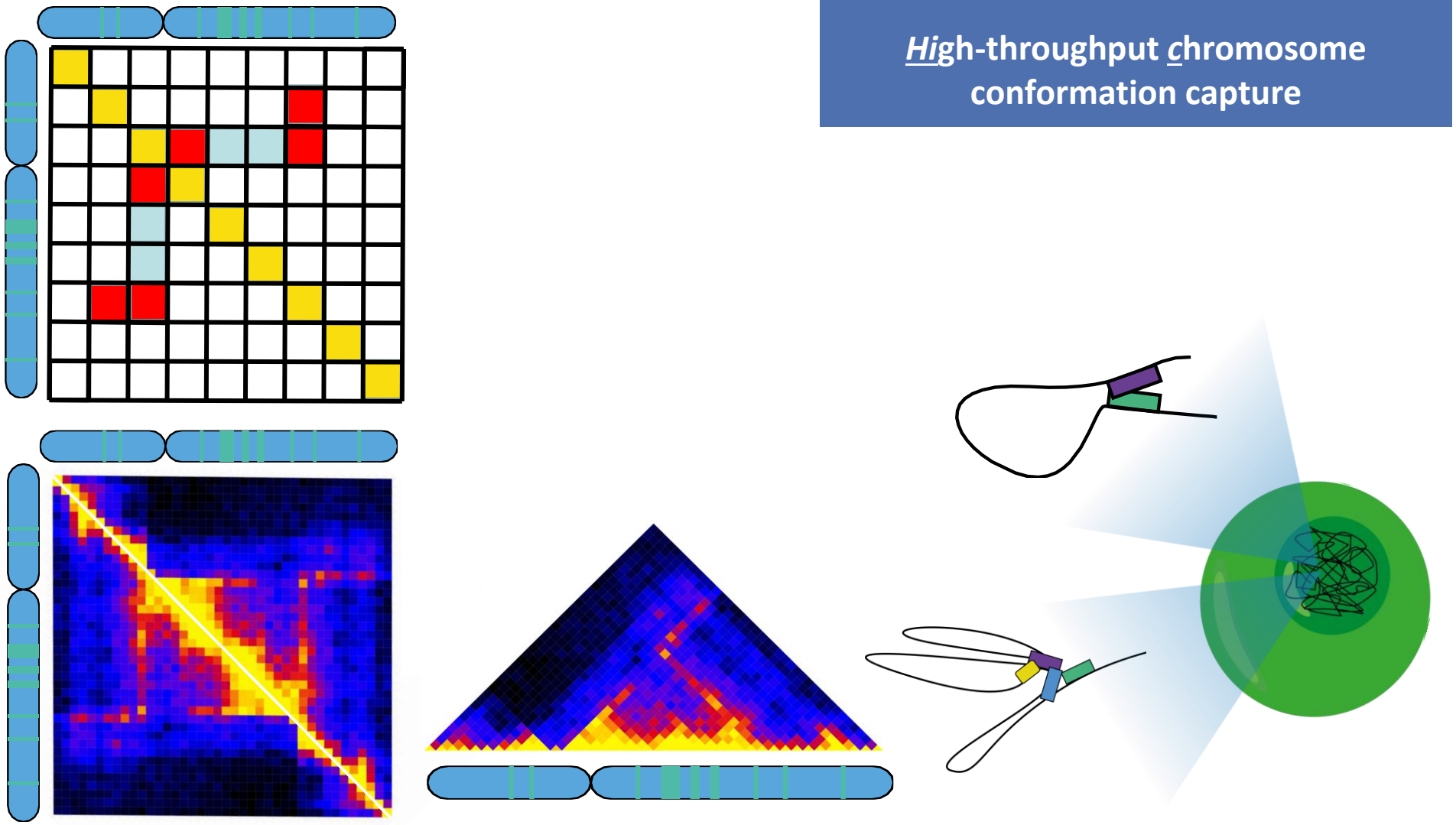




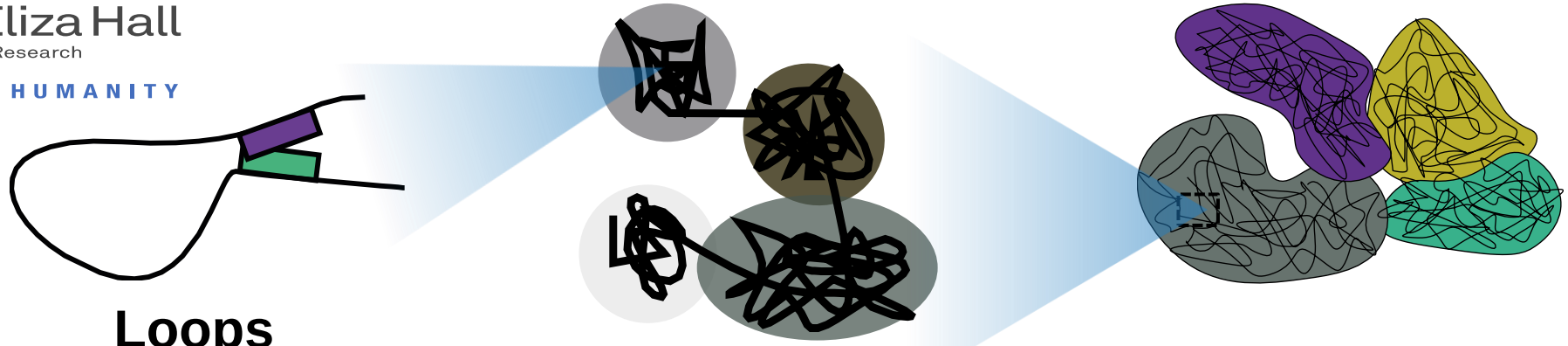
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Hi-C

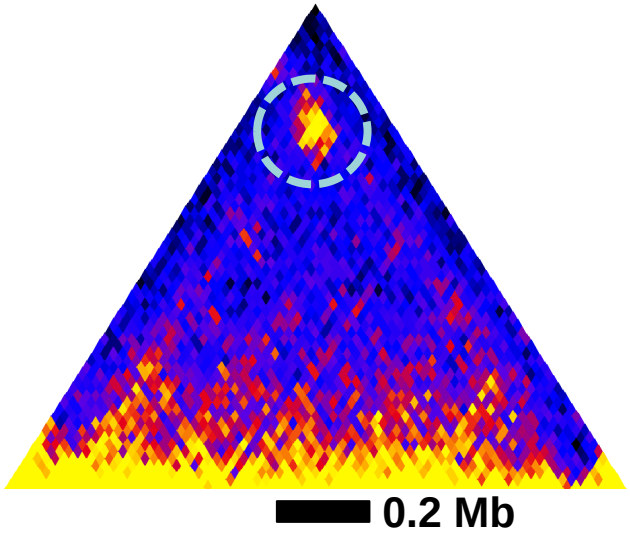
***Hi*gh-throughput *c*hromosome conformation capture**



Phanstiel. DH, (2015). *Sushi: Tools for visualizing genomics data*. R package version 1.16.0.



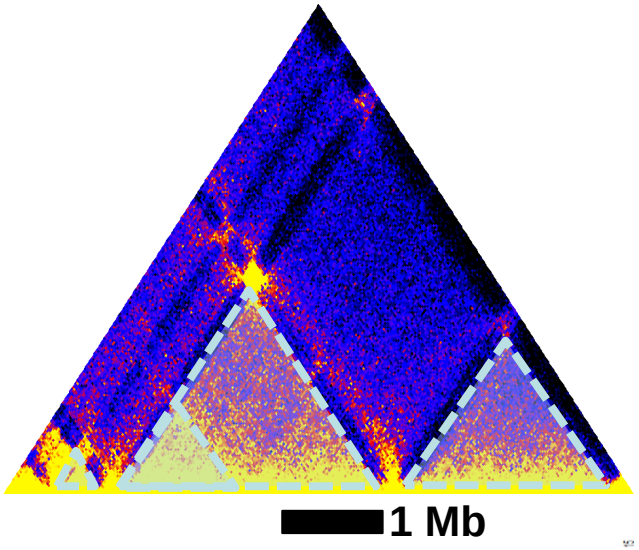
Loops
Enhancer-promoter



Peak calling/interaction methods

- HiCUPPS: compares to local background (Rao et al, Cell, 2014)
- HOMER: expected/observed (Heinz et al, Mol. Cell, 2010)

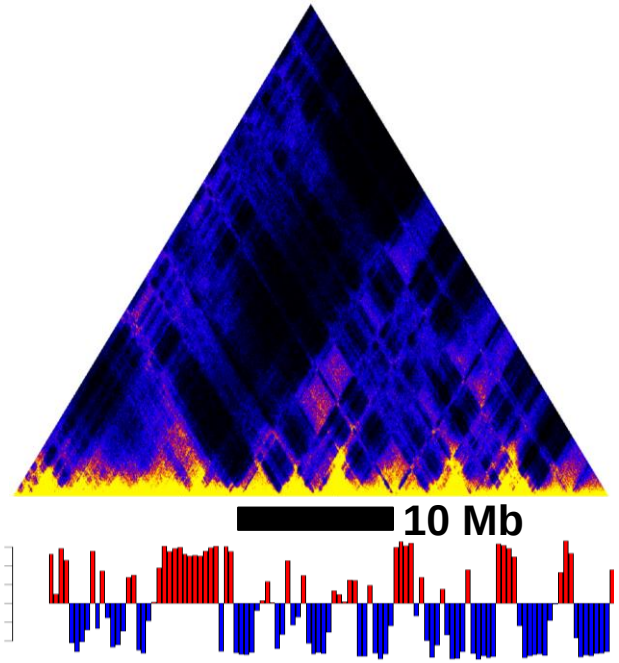
TADs



TAD finding methods

- DomainCaller (HMM): Models upstream/downstream interaction bias (Dixon et al, Nature, 2012)
- TADbit (Serra et al, bioRXiv, 2015)

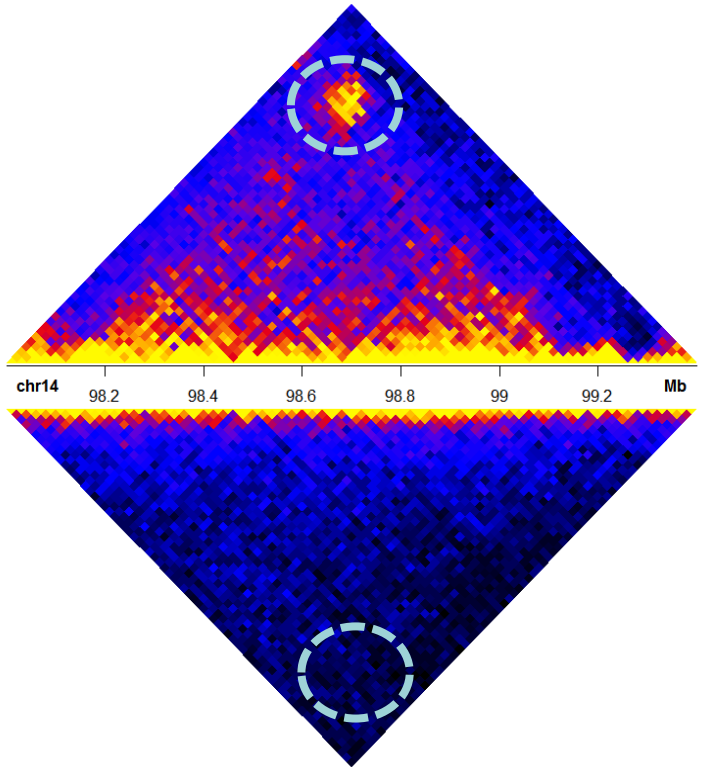
Chromosomes



A/B compartment

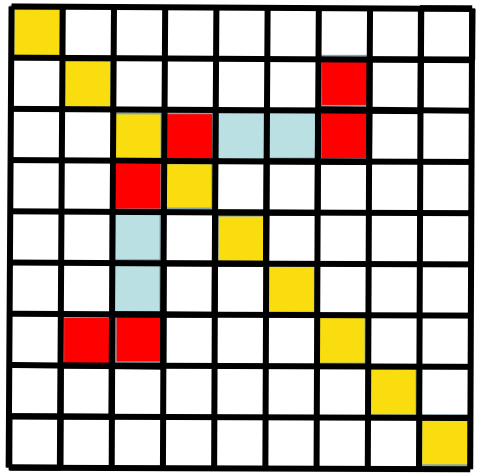
- Partition chromosome by first PCA indicates A/B compartments (Lieberman-Aiden et al, 2009)

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY



diffHic *(Lun & Smyth, BMC Bioinformatics, 2015)*

- Detects ***differential*** interactions (DIs) across biological conditions in a ***Hi-C*** experiment
- Statistical analysis uses *edgeR* *(Robinson et al, Bioinformatics, 2010) (McCarthy et al, Nucleic Acids Research, 2012)*
- Available on Bioconductor



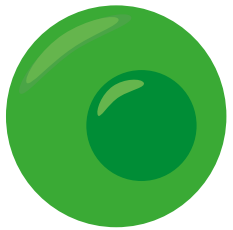
Anchor1	Anchor2	Sample 1	Sample 2	Sample 3	Sample4
chr1:1-50kb	chr1:50-100kb	1	1	2	3
chr1:50-100kb	chr1:100-150kb	1	0	3	2
chr1:100-150kb	chr1:150-200kb	56	59	65	62
chr1:150-200kb	chr1:200-250kb	10	13	16	17
chr1:200-250kb	chr1:250-300kb	2	5	7	8
chr2:100-150kb	chr2:150-200kb	3	2	4	6
chr2:200-250kb	chr2:250-300kb	4	2	1	2
chr2:300-350kb	chr2:350-400kb	21	19	25	27
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

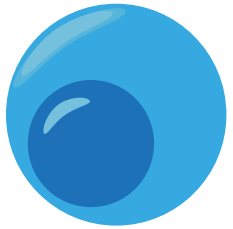
Outline

1. Introduction
    - Chromatin structure
    - HiC library construction
    - Analysis of HiC data

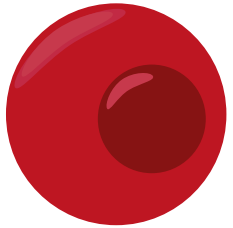
2. ***diffHic*** **analysis of immune cell types**



T cell



B cell



Granulocyte

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

diffHic pipeline

1. Alignment to genome with bowtie2
   - SORT bam files
   - Synchronize pair information
   - Mark duplicate reads
   - SORT bam files again

BAM files

2. Create the fragment library with genome and MboI

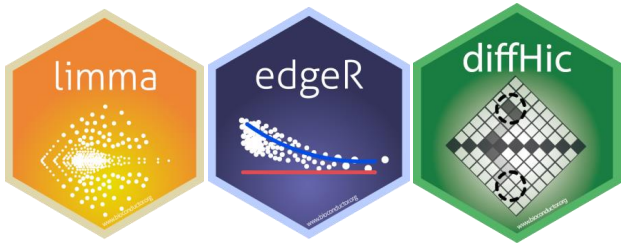
3. Map reads to fragment library and quality control

4. Remove artefacts

HDF5 libraries

5. Counting libraries into bins

Anchor1	Anchor2	Sample 1	Sample 2	Sample 3	Sample4
chr1:1-50kb	chr1:50-100kb	1	1	2	3
chr1:50-100kb	chr1:100-150kb	1	0	3	2
chr1:100-150kb	chr1:150-200kb	56	59	65	62
chr1:150-200kb	chr1:200-250kb	10	13	16	17
chr1:200-250kb	chr1:250-300kb	2	5	7	8
chr2:100-150kb	chr2:150-200kb	3	2	4	6
chr2:200-250kb	chr2:250-300kb	4	2	1	2
chr2:300-350kb	chr2:350-400kb	21	19	25	27
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…



6. Filtering of low count bins

7. Normalisation

8. Sample clustering

9. Variation between replicates

10. Differential analysis with *edgeR*

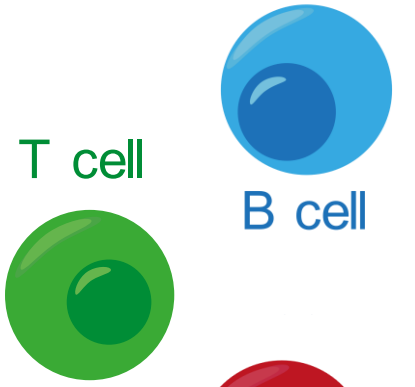
DI list

11. Visualisation

Plots

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

diffHic pipeline



Fastq files

UNIX

1. Alignment to genome with bowtie2
   - SORT bam files
   - Synchronize pair information
   - Mark duplicate reads
   - SORT bam files again

BAM files

2. Create the fragment library with genome and MboI

3. Map reads to fragment library and quality control

4. Remove artefacts

HDF5 libraries

5. Counting libraries into bins

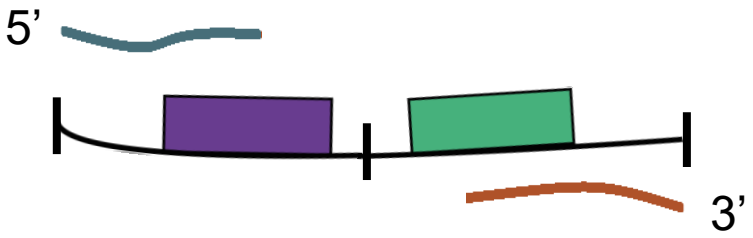
Anchor1	Anchor2	Sample 1	Sample 2	Sample 3	Sample4
chr1:1-50kb	chr1:50-100kb	1	1	2	3
chr1:50-100kb	chr1:100-150kb	1	0	3	2
chr1:100-150kb	chr1:150-200kb	56	59	65	62
chr1:150-200kb	chr1:200-250kb	10	13	16	17
chr1:200-250kb	chr1:250-300kb	2	5	7	8
chr2:100-150kb	chr2:150-200kb	3	2	4	6
chr2:200-250kb	chr2:250-300kb	4	2	1	2
chr2:300-350kb	chr2:350-400kb	21	19	25	27
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

1. Alignment to genome with bowtie: Pre-splitting alignment

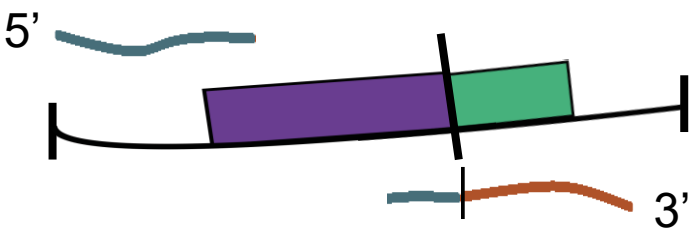
Bam files need to be (PICARD/SAMtools):

1. Sorted by name
2. Fix mate information
3. Mark duplicate reads



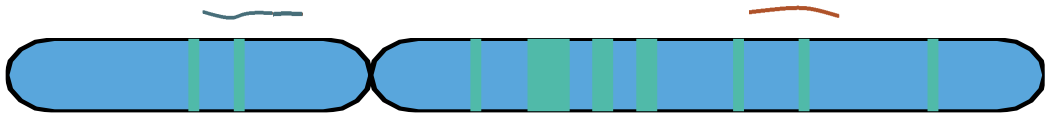
No restriction site:

Independently align each read with bowtie



With restriction site = chimeric read:

Before alignment split the read at ligation junction (Cutadapt) then
Independently align each read with bowtie



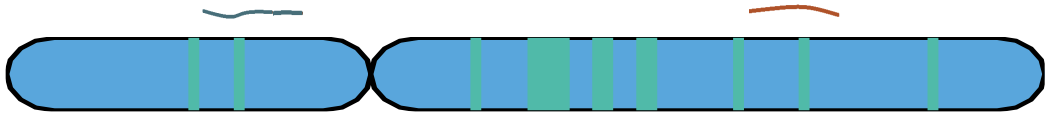
Walter+Eliza Hall Institute of Medical Research

DISCOVERIES FOR HUMANITY

2. Create the fragment library with genome and MboI

- The resolution of Hi-C data is limited by the restriction sites
- Report the read alignment location in terms of the restriction fragment.
- Create library of fragments and map reads

```
> library(diffHic)
> library(BSgenome.Mmusculus.UCSC.mm10)
> hg.frag<-cutGenome(BSgenome.Mmusculus.UCSC.mm10,"GATC",4)
> hs.frag
> GRanges object with 6684545 ranges and 0 metadata columns:
seqnames ranges strand <Rle> <IRanges> <Rle>
[1] chr1 1-3000194 *
[2] chr1 3000191-3000816 *
[3] chr1 3000813-3001051 *
[4] chr1 3001048-3001122 *
[5] chr1 3001119-3001798 *
... ... ... ...
[6684541] chrUn_JH584304 92982-97154 *
[6684542] chrUn_JH584304 97151-108839 *
[6684543] chrUn_JH584304 108836-109113 *
[6684544] chrUn_JH584304 109110-114452 *
[6684545] chrUn_JH584304 114449-114452 *
------- seqinfo: 66 sequences from mm10 genome
```



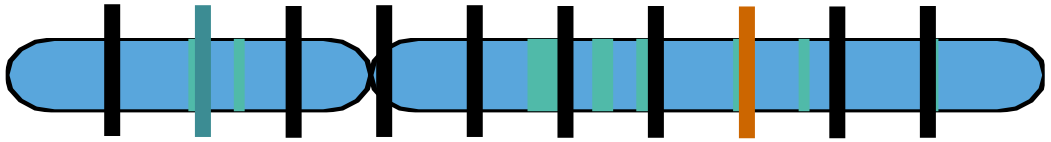
Walter+Eliza Hall Institute of Medical Research

DISCOVERIES FOR HUMANITY

2. Create the fragment library with genome and MboI

- The resolution of Hi-C data is limited by the restriction sites
- Report the read alignment location in terms of the restriction fragment.
- Create library of fragments and map reads

```
> library(diffHic)
> library(BSgenome.Mmusculus.UCSC.mm10)
> hg.frag<-cutGenome(BSgenome.Mmusculus.UCSC.mm10,"GATC",4)
> hs.frag
> GRanges object with 6684545 ranges and 0 metadata columns:
seqnames ranges strand <Rle> <IRanges> <Rle>
[1] chr1 1-3000194 *
[2] chr1 3000191-3000816 *
[3] chr1 3000813-3001051 *
[4] chr1 3001048-3001122 *
[5] chr1 3001119-3001798 *
... ... ... ...
[6684541] chrUn_JH584304 92982-97154 *
[6684542] chrUn_JH584304 97151-108839 *
[6684543] chrUn_JH584304 108836-109113 *
[6684544] chrUn_JH584304 109110-114452 *
[6684545] chrUn_JH584304 114449-114452 *
------- seqinfo: 66 sequences from mm10 genome
> diagnostics <- preparePairs("aligned.bam", hs.param,
file="hdf5_file.h5", dedup=TRUE, minq=10, chim.dist=800)
```



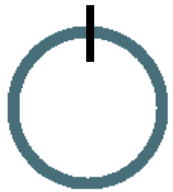
Walter+Eliza Hall Institute of Medical Research

DISCOVERIES FOR HUMANITY

3. Map reads to fragment library and quality control

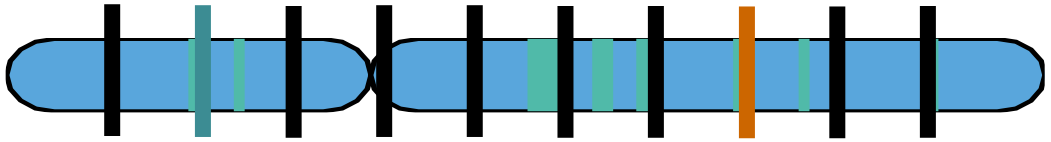
- *preparePairs*: converts the read position into an index (x2), pointing to the matching restriction fragment in hg.frag.
- The fragments to which the reads mapped are referred to as “anchors”

```
> diagnostics <- preparePairs("aligned.bam", hs.param,
file="hdf5_file.h5", dedup=TRUE, minq=10, chim.dist=800)
> diagnostics
$pairs
        total   marked filtered mapped
        7068675 103594 1532760  5460120
$same.id
        dangling self.circle
        423612   138248
$singles [1]
        0
$chimeras
        total   mapped  multi   invalid
        2495159 1725927 1040908 68231
```



Self circling

Dangling ends

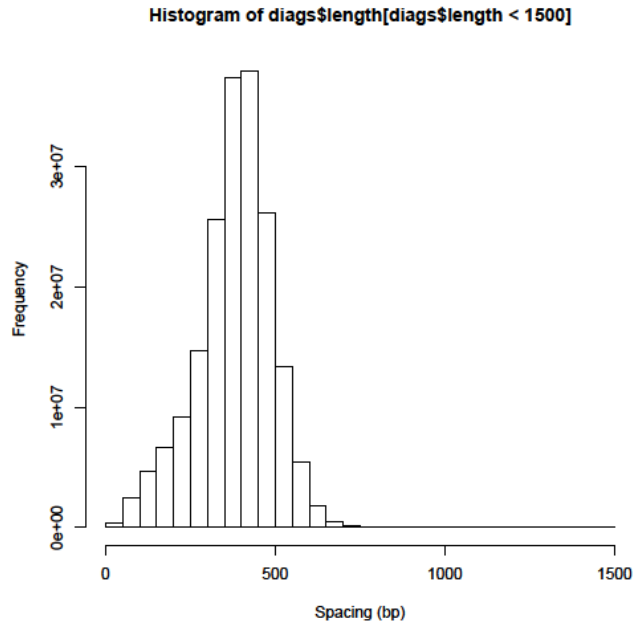


Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

4. Remove artefacts

- prunePairs: removes read pairs that we suspect are additional artefacts.
- Max.frag = offsite cleavage
- min.inward/min.outward = remove read pairs based on insert size and on the strand orientation

```
> prunePairs("hdf5_file.h5", hs.param,
        file.out="hdf5_file_trimmed.h5",
        max.frag=700,
        min.inward=1000,
        min.outward=16000)

total length inward outward retained
4896653 870339 94644 82964 3860024

diags <- getPairData("hdf5_file.h5", hs.param)

hist(diags$length[diags$length < 1500], ylab="Frequency",
xlab="Spacing (bp)", main="", col="grey80")
```

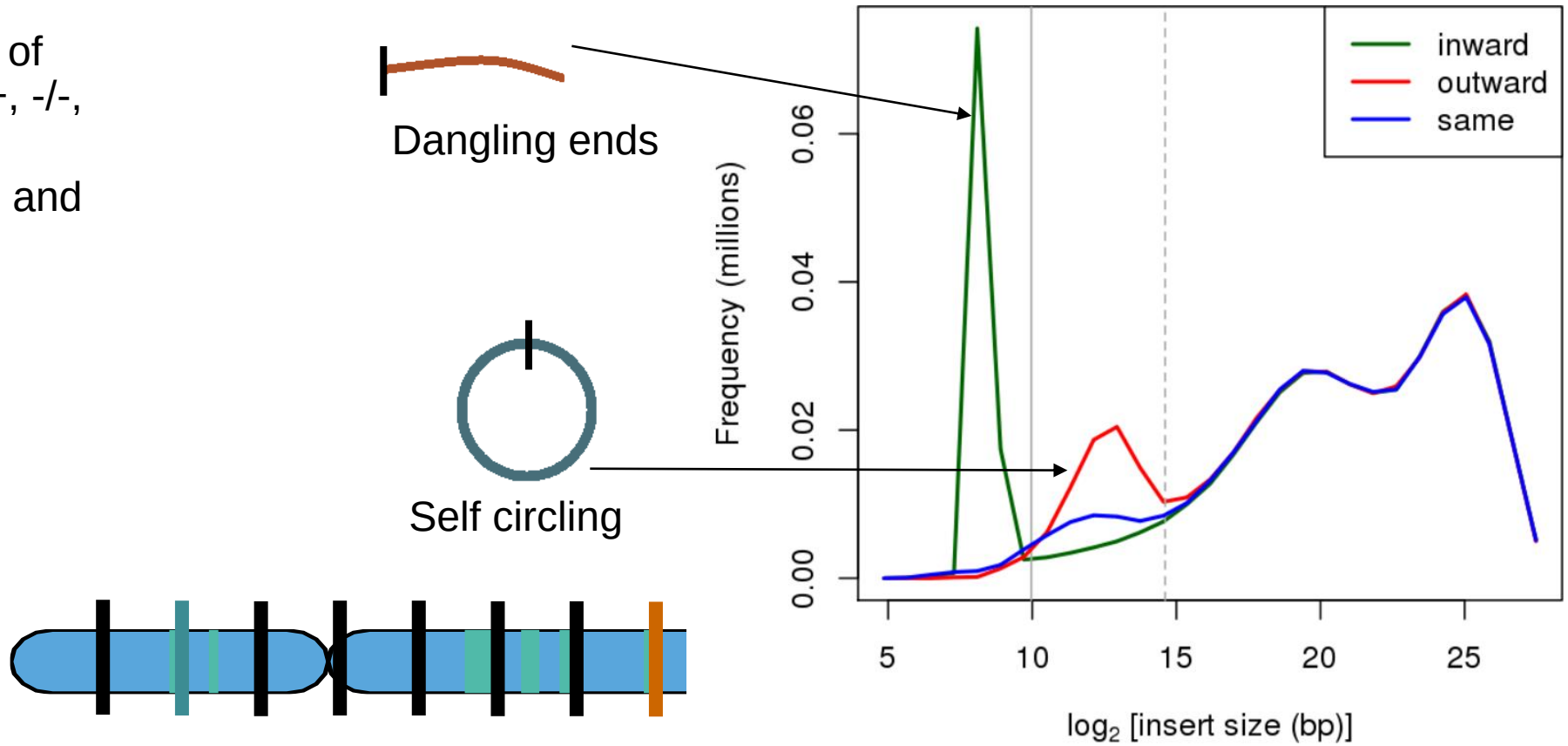


Walter+Eliza Hall Institute of Medical Research
DISCOVERIES FOR HUMANITY

4. Remove artefacts

- If different pieces of DNA were DNA randomly ligated together would expect to observe equal proportions of all strand orientations (+/+, -/-, +/- and -/+)
- Spikes indicate self circles and dangling ends

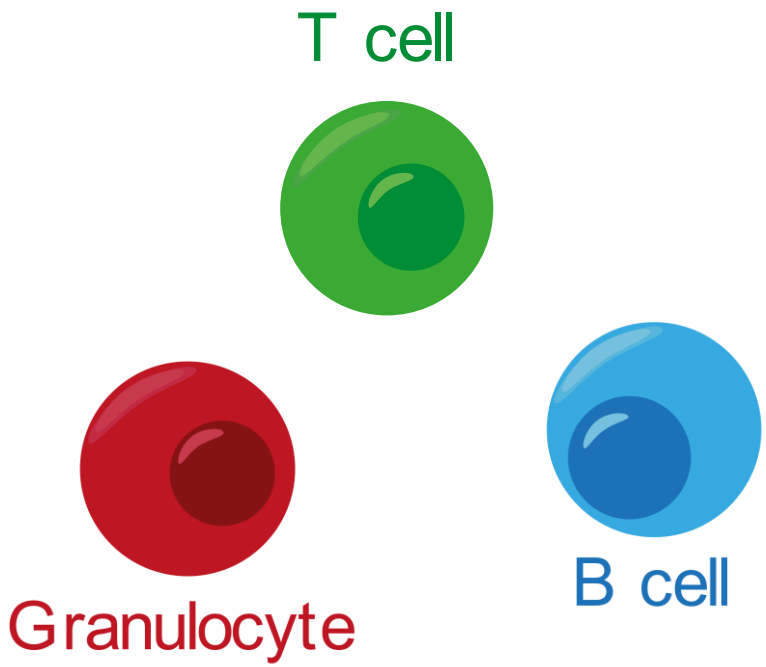
```
min.inward<- 1000
min.outward<- 16000

llinsert <- log2(diags$insert + 1L)
intra <- !is.na(llinsert)
……

plot(0,0,type="n", xlim=c(xmin, xmax), ylim=c(0,
ymax),xlab=expression(log[2]~"[insert size (bp)]"), ylab="Frequency
(millions)")
```

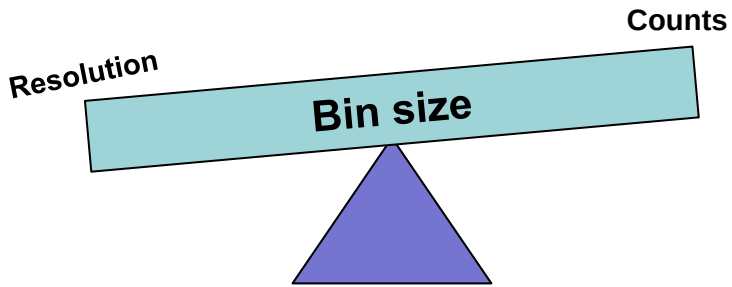


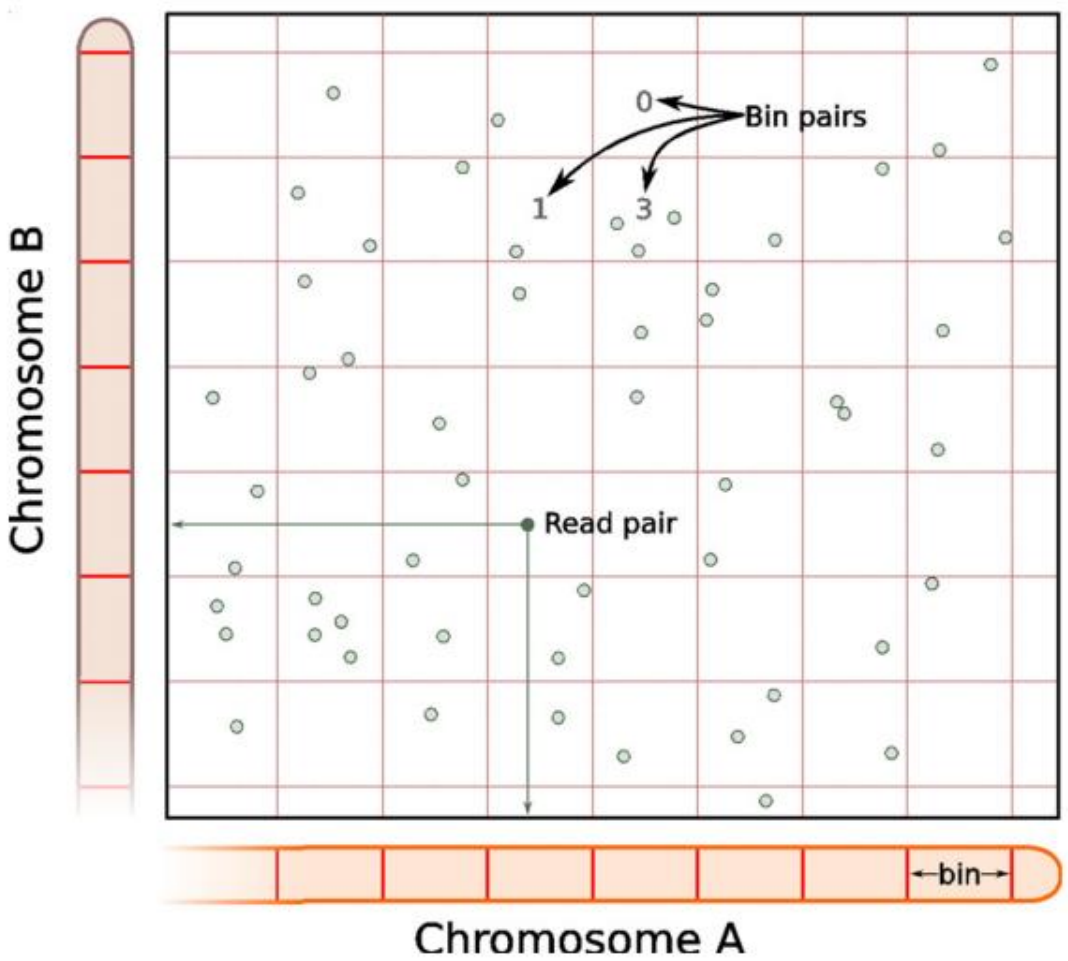
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

5. Counting libraries into bins

```
bin.size <- 100e3
data<-squareCounts(files, hs.param,
                   width=bin.size, filter=10,
                   restrict.regions = TRUE)
```

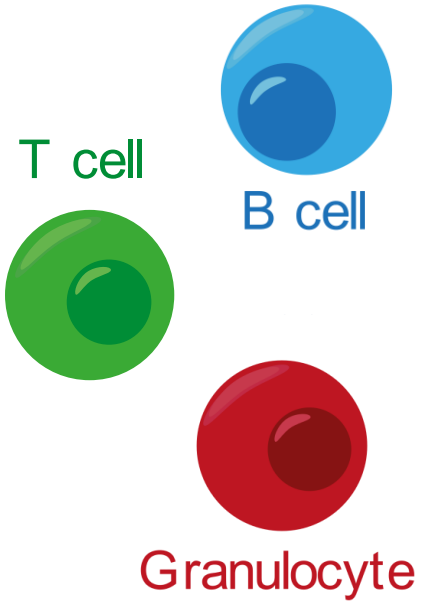






Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

diffHic pipeline



Anchor1	Anchor2	Sample 1	Sample 2	Sample 3	Sample4
chr1:1-50kb	chr1:50-100kb	1	1	2	3
chr1:50-100kb	chr1:100-150kb	1	0	3	2
chr1:100-150kb	chr1:150-200kb	56	59	65	62
chr1:150-200kb	chr1:200-250kb	10	13	16	17
chr1:200-250kb	chr1:250-300kb	2	5	7	8
chr2:100-150kb	chr2:150-200kb	3	2	4	6
chr2:200-250kb	chr2:250-300kb	4	2	1	2
chr2:300-350kb	chr2:350-400kb	21	19	25	27
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

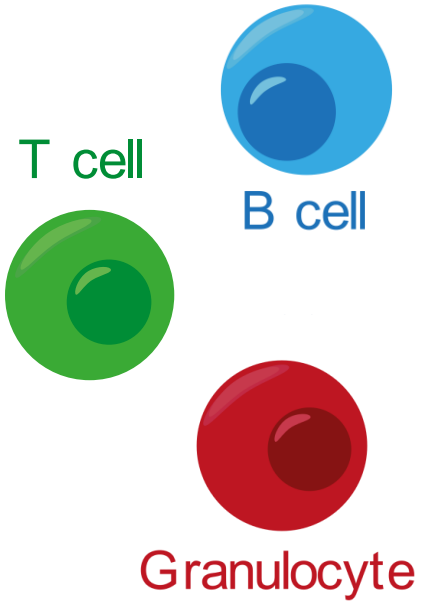
6. Filtering of low interactions → 7. Normalisation → 8. Sample clustering → 9. Variation between replicates → 10. Differential analysis with *edgeR* → 11. Visualisation

10. Differential analysis with *edgeR* → DI list

11. Visualisation → Plots

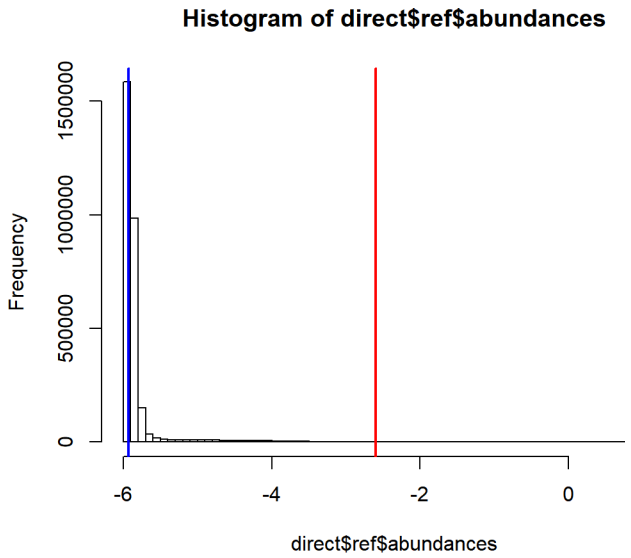
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

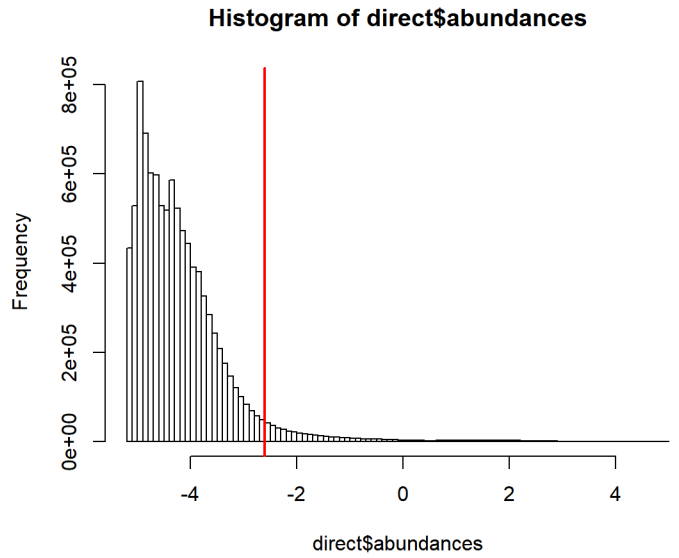
6. Filtering of low interactions



- Many different strategies to remove uninteresting interactions
- We will use filterDirect function
- Threshold: median abundance across inter-chromosomal bin pairs

```
background <- squareCounts(files, hs.param,
                           width=1e6)
direct <- filterDirect(data, reference=background)

high.ab <- direct$abundances
                           > direct$threshold + log2(10)

keep <- high.ab & filterDiag(data, by.diag=1L)
data <- data[keep,]
```

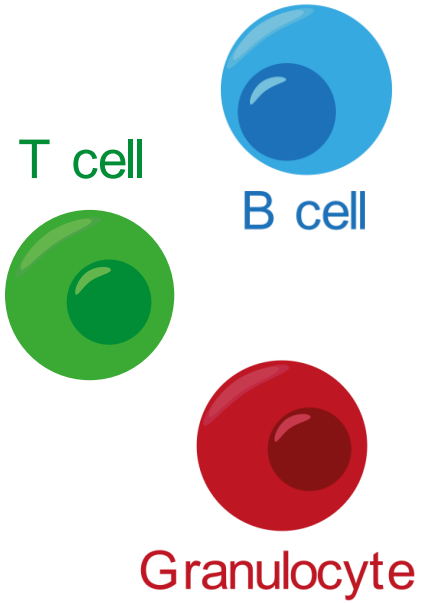




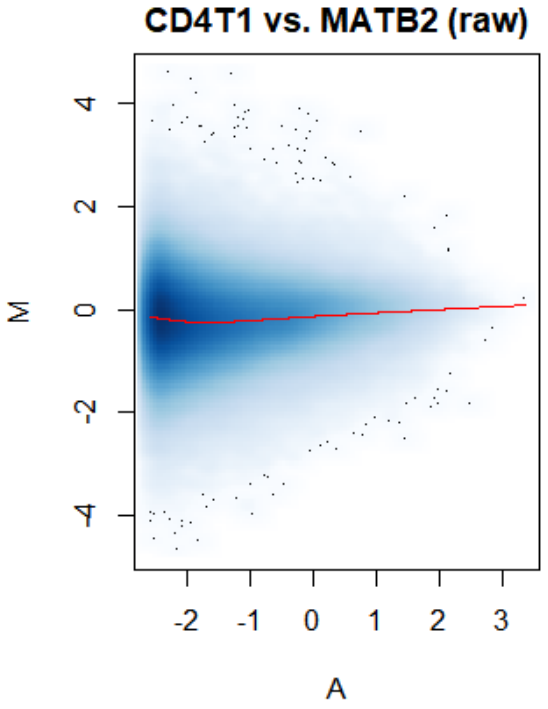
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

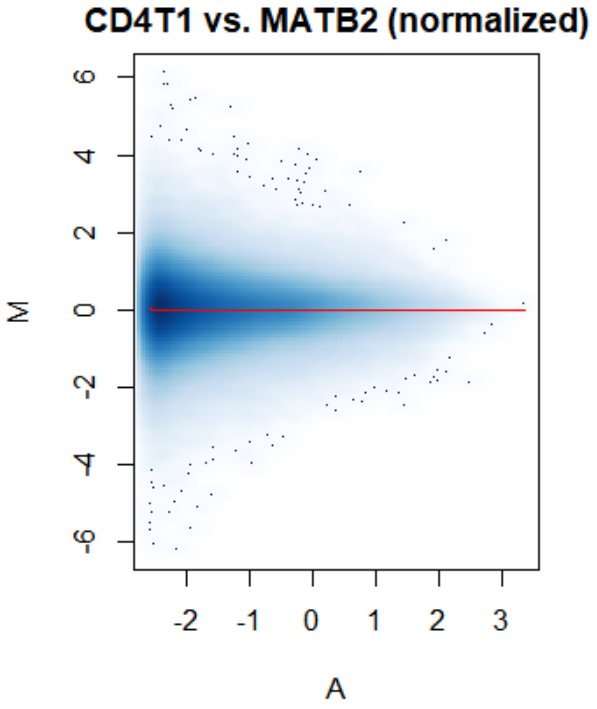
```
library(csaw)
data.offs <- normOffsets(data, type="loess")
head(assay(data.offs,2))
```

7. Normalisation



- Hi-C data contains complex and systematic biases:
  - Fragment length
  - GC content
  - Mappability
- However, these cancel out for a differential analysis
- But we may observe abundance-dependent trended biases
- Non-linear normalisation

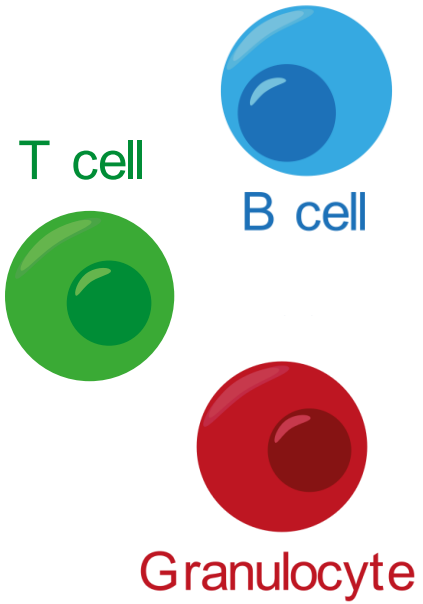


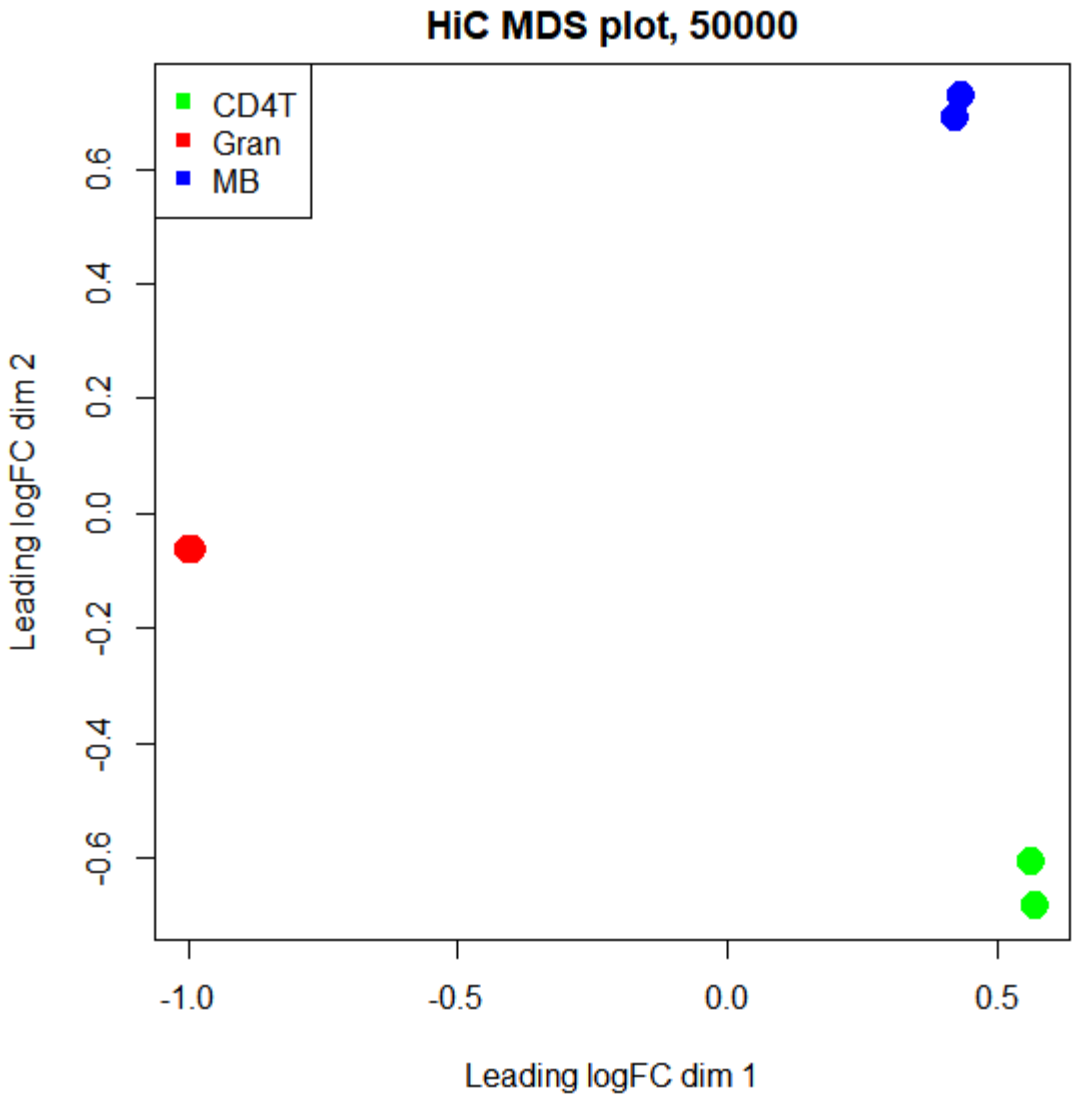


Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

8. Sample clustering

```
group.col <-c("red","red","green","green","blue","blue")
plotMDS(normc, col=group.col,pch = as.numeric(19), top=50000, cex=2,
        dim.plot = c(1,2), main = "HiC MDS plot, 50000")
legend("topleft", legend = unique(group), col =unique(group.col), pch =
        as.numeric(15), cex=1.0, title.col = "black")
```



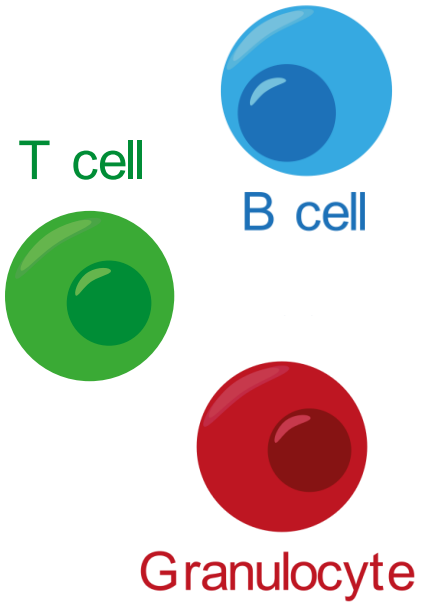


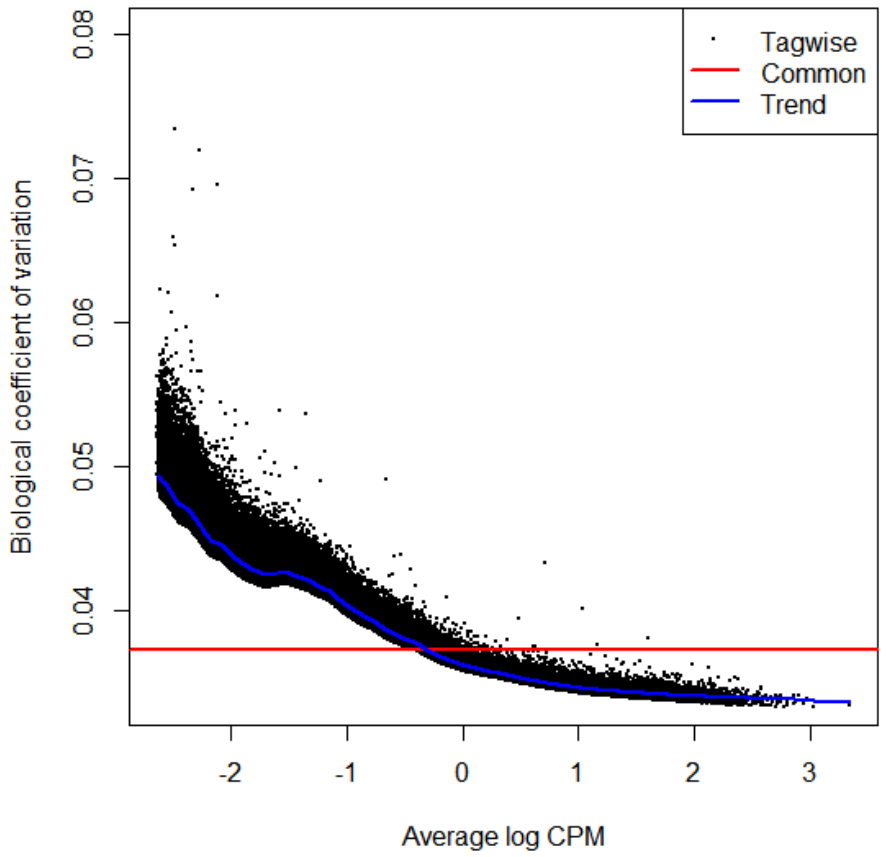
Walter+Eliza Hall Institute of Medical Research

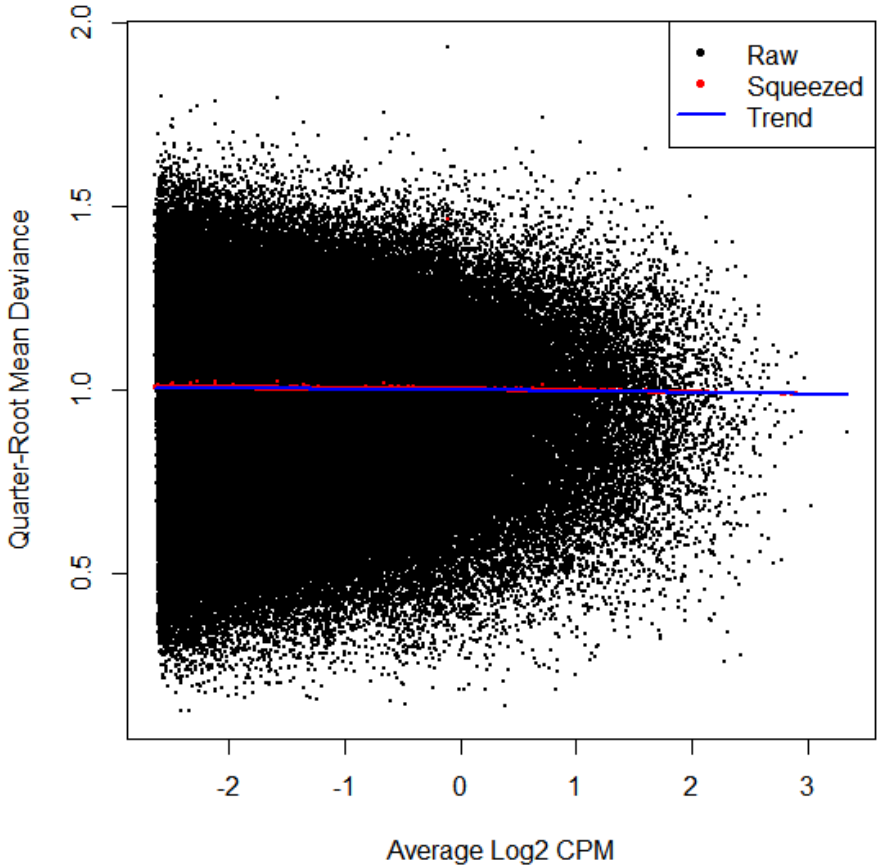
DISCOVERIES FOR HUMANITY

9. Variation between replicates

```
dispersion <- estimateDisp(y, design,robust=TRUE)
BCV <- sqrt(dispersion$common.dispersion)
BCV
plotBCV(dispersion, ylim=c(0.034,0.08))

fit <- glmQLFit(dispersion, design, robust=TRUE)
plotQLDisp(fit)
```



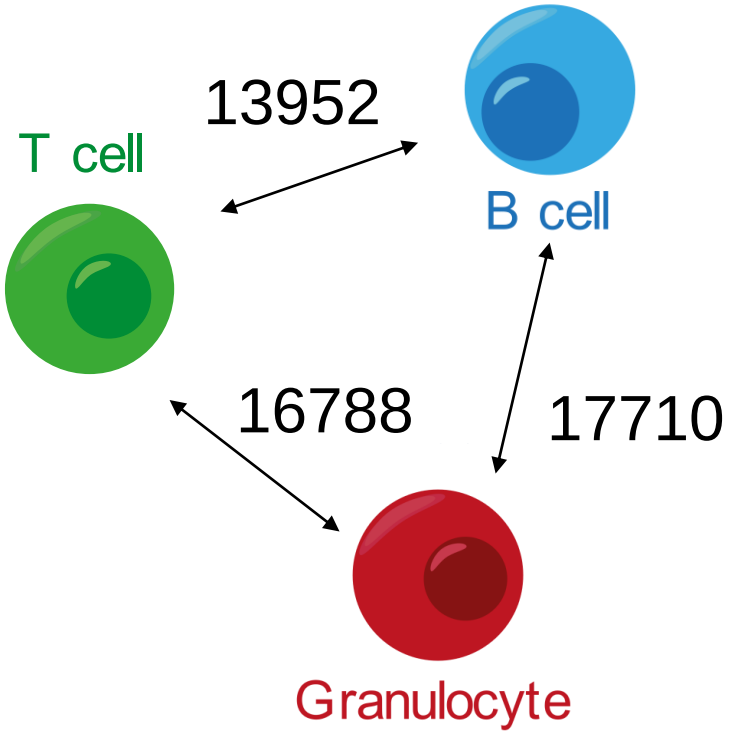




Walter+Eliza Hall Institute of Medical Research
DISCOVERIES FOR HUMANITY

10. Differential analysis with *edgeR*

- Define the contrasts
- Perform the test
- Cluster the adjacent DIs

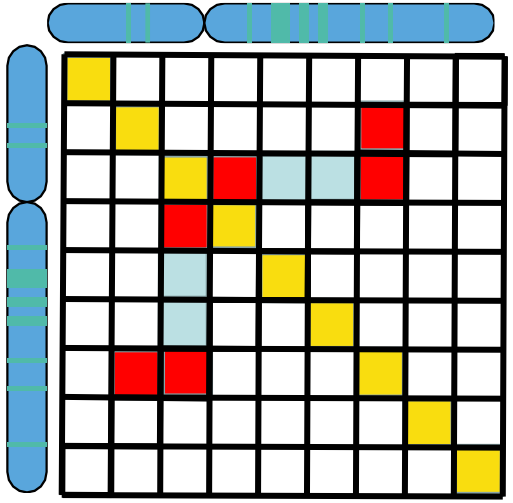


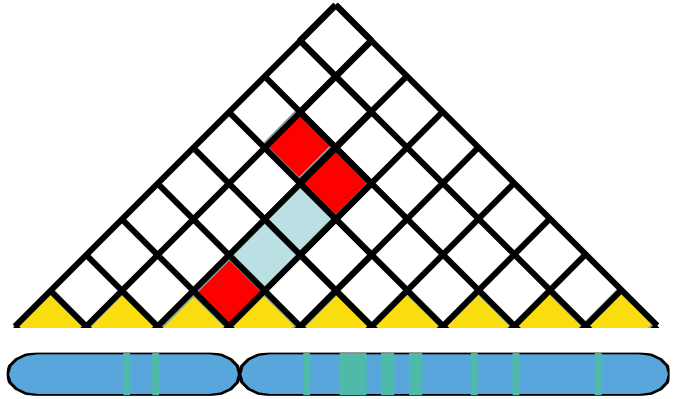
Exercise: Differential interactions between CD4+ T cells and granulocytes

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

10. Visualisation

- Plaid plots can be used to visualize read pairs in the interaction space (Lieberman-Aiden et al, Science, 2009).
- Rotated plaid plots can also be used BUT only for local interactions.
- Sushi package
  - Phanstiel. DH, (2015). *Sushi: Tools for visualizing genomics data*. R package version 1.16.0.
  - Tools for visualizing genomic data including Hi-C



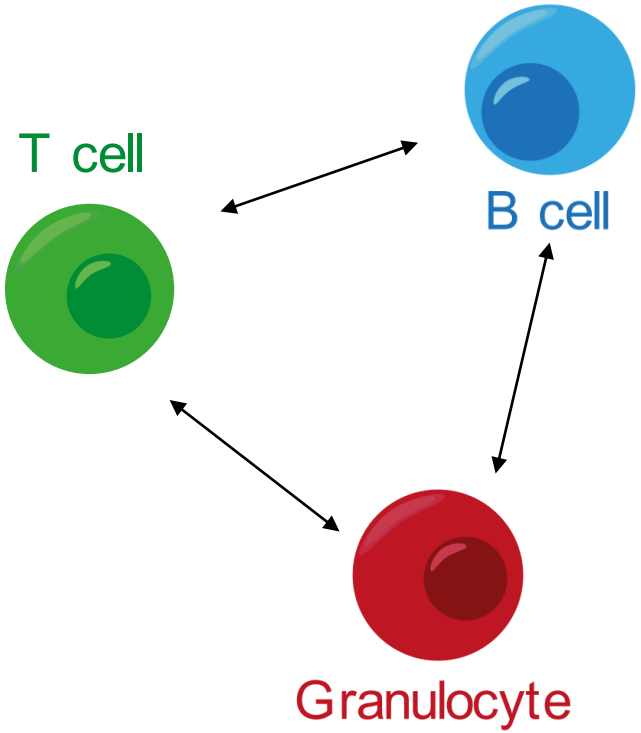


Exercise: Plot a similar DI (chr12) from the CD4+ T versus Grans

Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Analysis at 1 Mbp

Exercise: Perform the differential analysis with the bin.size=1 Mbp data



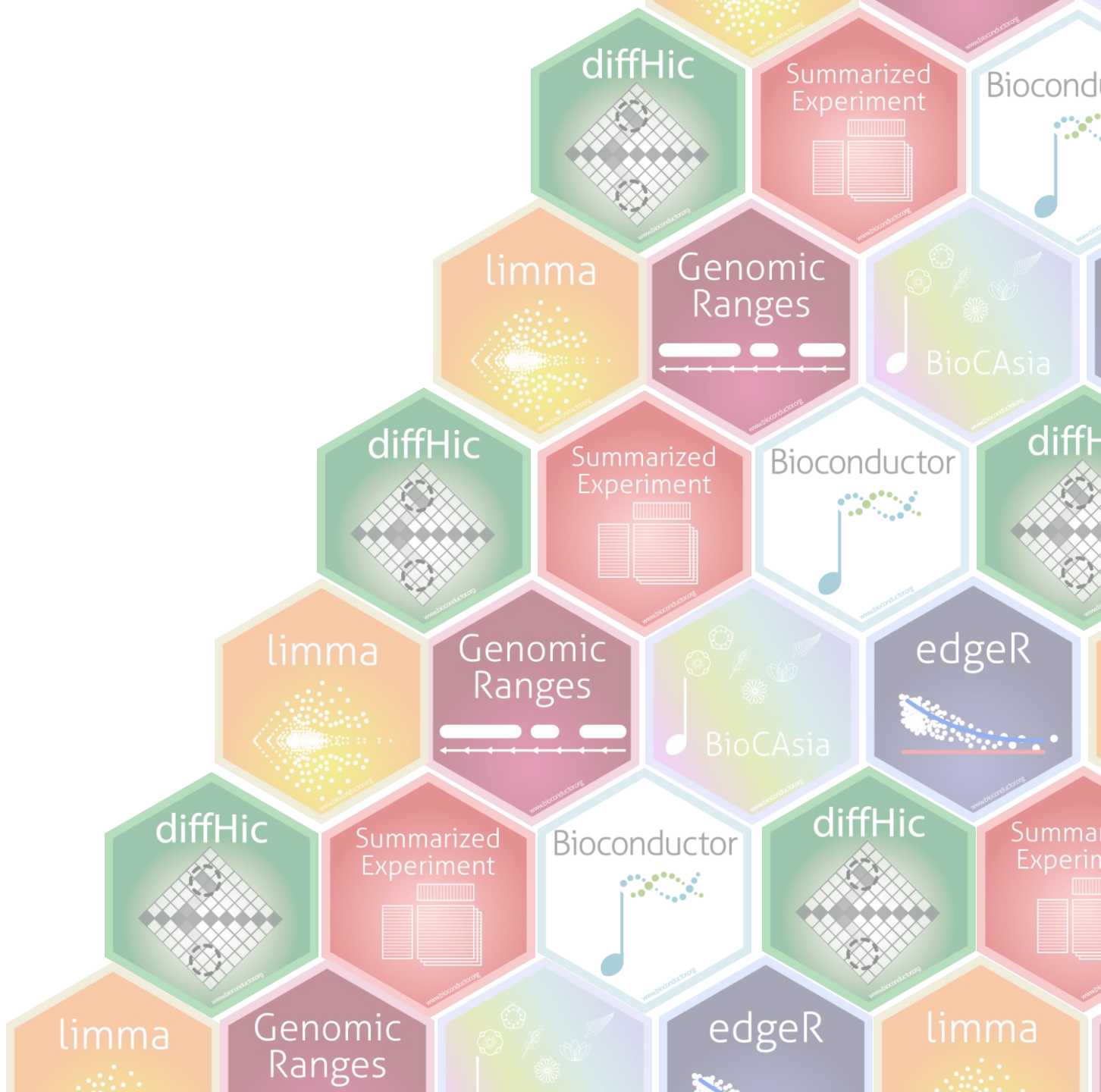
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Conclusions

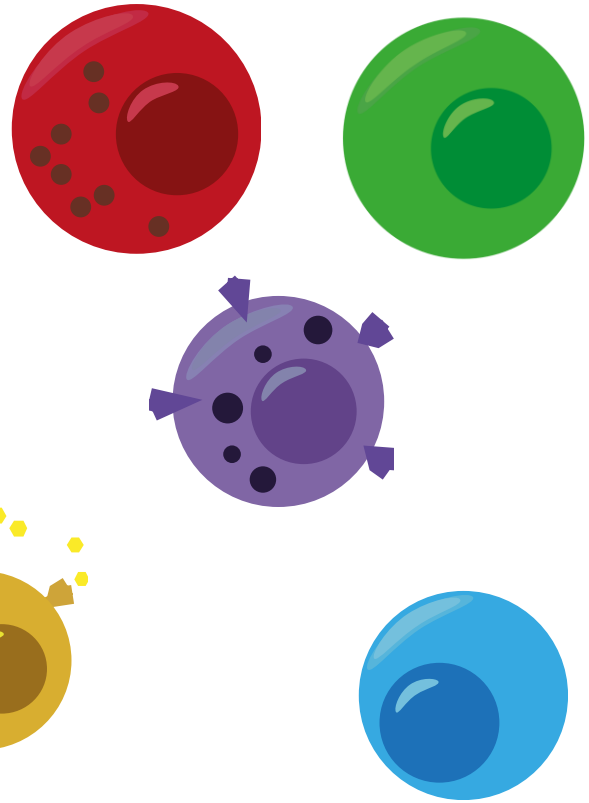
Differential analysis is powerful for analysis Hi-C data …

…..especially when combined with other types of genomic data

Use diffHic for HiC if you have replicates!



Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY



Acknowledgements

Smyth lab (Bioinformatics division)

Gordon Smyth

Aaron Lun (now Cambridge)

Alexandra Garnham

Connie Li Wai Suen

Allan lab (Molecular Immunology division)

Tim Johanson

Rhys Allan

Nadia Iannarella

Stephen Nutt