First steps with PCHi-C data

Mónica Cabrera-Pasadas

What is PCHi-C?

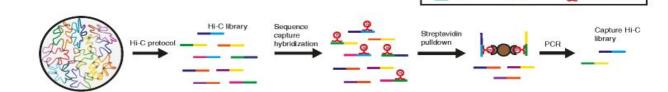
https://pubmed.ncbi.nlm.nih.gov/30010637/

> J Vis Exp. 2018 Jun 28;(136):57320. doi: 10.3791/57320.

Promoter Capture Hi-C: High-resolution, Genomewide Profiling of Promoter Interactions

Stefan Schoenfelder 1 , Biola-Maria Javierre 2 , Mayra Furlan-Magaril 3 , Steven W Wingett 4 , Peter Fraser 5

- Technique to enable the **genome-wide detection of distal promoter-interacting regions** (PIRs), for all promoters in a single experiment.
- In PCHi-C, highly complex **Hi-C libraries are specifically enriched for promoter sequences** through in-solution hybrid selection with thousands of biotinylated RNA baits complementary to the ends of all promoter-containing restriction fragments. **The aim is to pull-down promoter sequences and their frequent interaction partners** such as enhancers and other potential regulatory elements. **After high-throughput paired-end sequencing, a statistical test is applied** to each promoter-ligated restriction fragment to identify significant PIRs at the restriction fragment level



Biotinylated RNA pmb

Data processing and analysis: HiCUP

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4706059/

Version 1. F1000Res. 2015; 4: 1310.

PMCID: PMC4706059

Published online 2015 Nov 20. doi: 10.12688/f1000research.7334.1

PMID: 26835000

HiCUP: pipeline for mapping and processing Hi-C data

Steven Wingett, a,1,2 Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews

https://www.youtube.com/watch?v=i6imVs66aew

1) **Create Aligner Indices** HiCUP uses the aligner Bowtie or Bowtie2 to map sequences to a reference genome, requiring the construction of genome index files. These indices MUST be constructed from the same reference genome files as used by the HiCUP Digester script.

Execute the script:

/apps/BOWTIE/1.2.2/GCC/bowtie-build /genomes/human/GRCh37/release_92/Homo_sapiens.GRCh37.92.dna.chromosome.*.fa /genomes/human/GRCh37/release_92/indices/ebwt/Homo_sapiens_GRCh37_92_index

2) **Create a digested reference genome.** To filter out common experimental artefacts, HiCUP requires the positions at which the restriction enzyme(s) used in the protocol cut the genome. The script HiCUP Digester creates this reference genome digest file.

Execute the script:

hicup_digester --genome Human_GRCh37 --re1 A^AGCTT,HindIII *.fa

Data processing and analysis: HiCUP

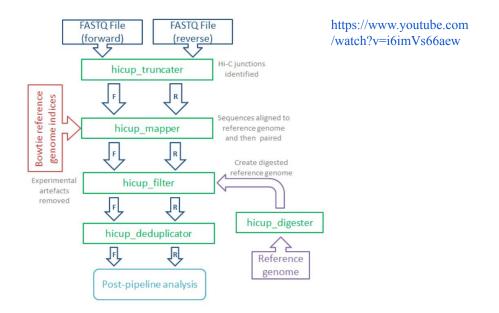
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4706059/

Version 1. <u>F1000Res.</u> 2015; 4: 1310. Published online 2015 Nov 20, doi: 10.12688/f1000research.7334.1 PMCID: PMC4706059

PMID: 26835000

HiCUP: pipeline for mapping and processing Hi-C data

Steven Wingett, a,1,2 Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews



- 3) Run the HiCUP Pipeline to run the following scripts:
 - HiCUP Truncater
 - HiCUP Mapper
 - HiCUP Filter
 - HiCUP Deduplicator

Execute the script:

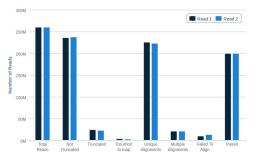
hicup --config [Configuration Filename]

```
Outdir: /gpfs/projects/bsc08/bsc08471/PCHiC/p53/HiCUP/MOCK1K0
Threads: 10
Quiet:0
Keep:0
Zip:1
Bowtie: /apps/BOWTIE/1.2.2/GCC/bowtie
R: /apps/R/3.5.1/INTEL/bin/R
Index: /gpfs/projects/bsc08/shared projects/IJC 3Dchromatin/genomes/human/
GRCh37/current/indices/ebwt/Homo sapiens GRCh37 92 index
Digest: /gpfs/projects/bsc08/shared projects/IJC 3Dchromatin/genomes/human/
GRCh37/current/digest genome/
Digest Human GRCh37 92 HindIII None 10-51-28 27-03-2020.txt.gz
Format:
Longest: 800
Shortest: 150
/gpfs/scratch/bsc08/bsc08471/data/Original Sierra Babraham/MOCK1KO/
lane5 NoIndex L005 R1.fastq.qz
/gpfs/scratch/bsc08/bsc08471/data/Original Sierra Babraham/MOCK1KO/
lane5 NoIndex L005 R2.fastq.qz
```

	Read 1	Read 2
Total Reads	260,220,163	260,220,163
Not Truncated	235,680,822	237,180,471
Truncated	24,539,341	23,039,692
Too short to map	3,636,518	3,312,742
Average length of truncated sequence	30.48	30.68

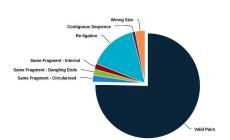
	Read 1	Read 2
Unique Alignments	225,269,321	222,599,191
Multiple Alignments	21,219,285	20,944,830
Failed To Align	10,095,039	13,363,400
Paired	199,380,409	199,380,409

MOCK1KO



Filtering

	Di-Tag Count
Valid Pairs	150,356,474
Invalid Pairs	49,023,935
Same Circularised	4,032,537
Same Fragment Dangling Ends	3,869,234
Same Fragment Internal	3,409,417
Re-ligation	29,674,629
Contiguous Sequence	1,964,153
Wrong Size	6,073,965
Total Pairs	199,380,409

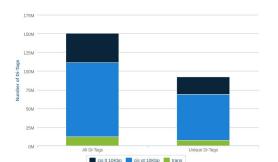


De-duplication

Percentage uniques: 61.31

	All Di-Tags	Unique Di-Tags
Read Pairs	150,356,474	92,188,914
Cis-close (< 10Kbp)	38,953,686	23,513,623
Cis-far (> 10Kbp)	99,348,757	61,327,198
Trans	12,054,031	7,348,093

60.64% captured

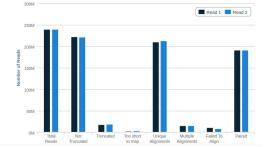


Truncating and Mapping

	Read 1	Read 2
otal Reads	238,997,995	238,997,995
ot Truncated	221,594,521	220,935,772
runcated	17,403,474	18,062,223
oo short to map	2,521,312	2,703,612
verage length of truncated sequence	30.62	30.46

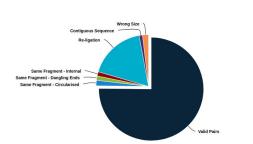
	Read 1	Read 2
Unique Alignments	210,092,040	212,427,796
Multiple Alignments	15,228,891	15,377,475
Failed To Align	11,155,752	8,489,112
Paired	190,543,734	190,543,734

MOCK2KO



Filtering

	Di-Tag Coun
Valid Pairs	143,277,38
Invalid Pairs	47,266,35
Same Circularised	3,247,97
Same Fragment Dangling Ends	2,575,49
Same Fragment Internal	2,529,25
Re-ligation	33,418,56
Contiguous Sequence	1,960,73
Wrong Size	3,534,32
Total Pairs	190,543,73

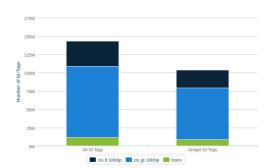


De-duplication

Percentage uniques: 72.25

	All Di-Tags	Unique Di-Tags
Read Pairs	143,277,381	103,520,686
Cis-close (< 10Kbp)	34,667,708	24,536,025
Cis-far (> 10Kbp)	96,962,386	70,512,261
Trans	11.647.287	8,472,400

84.07% captured



	Read 1	Read 2
Total Reads	237,206,273	237,206,273
Not Truncated	215,855,743	217,284,462
Truncated	21,350,530	19,921,811
Too short to map	3,168,487	2,869,826

	Read 1	Read 2
Unique Alignments	202,774,198	201,074,276
Multiple Alignments	21,448,640	21,245,398
Failed To Align	9,814,948	12,016,773
Paired	179,267,942	179,267,942

Average length of truncated sequence

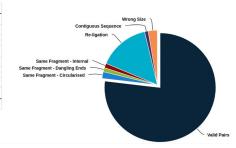
							Read 1	Read 2
200M -								_
150M -	Н	╫			Н			Н
100M -	H	Ш			Н			
50M -		-11-			╫			
OM =	Total	Not	Truncated	Too short	Unique	Multiple	Failed To	Paired

MOCK1WT

30.63

Filtering

	Di-Tag Count
Valid Pairs	137,856,713
Invalid Pairs	41,411,229
Same Circularised	3,505,164
Same Fragment Dangling Ends	2,228,245
Same Fragment Internal	2,361,660
Re-ligation	26,707,387
Contiguous Sequence	1,861,146
Wrong Size	4,747,627
Total Pairs	179,267,942

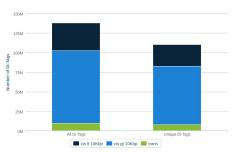


De-duplication

Percentage uniques: 79.98

	All Di-Tags	Unique Di-Tags
Read Pairs	137,856,713	110,258,841
Cis-close (< 10Kbp)	35,372,353	27,768,184
Cis-far (> 10Kbp)	92,857,185	74,762,080
Trans	9,627,175	7,728,577

60.54% captured

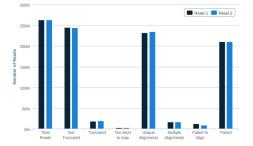


Truncating and Mapping

	Read 1	Read 2
Total Reads	262,846,987	262,846,987
Not Truncated	244,168,124	243,349,278
Truncated	18,678,863	19,497,709
Too short to map	2,622,881	2,824,189
Average length of truncated sequence	30.75	30.59

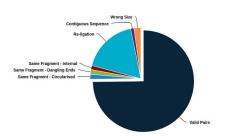
	Read 1	Read 2
Unique Alignments	231,472,455	234,266,415
Multiple Alignments	16,726,334	16,961,649
Failed To Align	12,025,317	8,794,734
Paired	210,134,874	210,134,874





Filtering

	Di-Tag Coun
Valid Pairs	157,219,58
Invalid Pairs	52,915,29
Same Circularised	3,231,70
Same Fragment Dangling Ends	3,136,114
Same Fragment Internal	2,683,92
Re-ligation	37,065,52
Contiguous Sequence	2,182,90
Wrong Size	4,615,12
Total Pairs	210,134,874

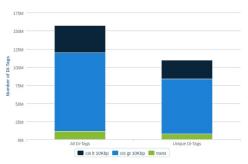


De-duplication

Percentage uniques: 69.50

	All Di-Tags	Unique Di-Tags
Read Pairs	157,219,580	109,269,204
Cis-close (< 10Kbp)	37,519,662	25,455,372
Cis-far (> 10Kbp)	108,180,145	75,737,034
Trans	11,519,773	8,076,798

83.92% captured



	Read 1	Read 2
Total Reads	254,246,572	254,246,572
Not Truncated	232,542,012	232,114,002
Truncated	21,704,560	22,132,570
Too short to map	3,227,600	3,348,599
Average length of truncated sequence	30.40	30.31

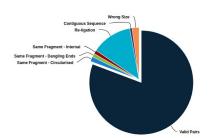
	Read 1	Read 2
Unique Alignments	220,277,091	222,959,303
Multiple Alignments	18,119,934	18,282,403
Failed To Align	12,621,947	9,656,267
Paired	197,626,650	197,626,650

250M 250M 200M 200M

NUT1KO

Filtering

	Di-Tag Count
Valid Pairs	159,589,869
Invalid Pairs	38,036,781
Same Circularised	2,982,392
Same Fragment Dangling Ends	2,160,120
Same Fragment Internal	2,200,455
Re-ligation	24,942,953
Contiguous Sequence	1,442,178
Wrong Size	4,308,683
Total Pairs	197,626,650

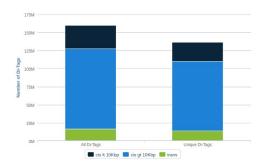


De-duplication

Percentage uniques: 85.54

	All Di-Tags	Unique Di-Tags
Read Pairs	159,589,869	136,509,044
Cis-close (< 10Kbp)	32,097,743	26,817,855
Cis-far (> 10Kbp)	111,322,605	95,761,764
Trans	16,169,521	13,929,425

81.47% captured

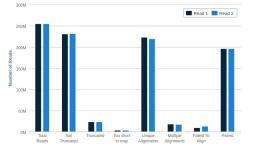


Truncating and Mapping

	Read 1	Read 2
otal Reads	254,321,968	254,321,968
ot Truncated	230,545,257	231,022,910
runcated	23,776,711	23,299,058
oo short to map	3,770,422	3,618,186
verage length of truncated sequence	30.17	30.27

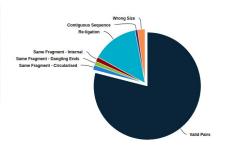
	Read 1	Read 2
nique Alignments	222,383,852	219,040,487
Multiple Alignments	18,258,963	18,010,077
ailed To Align	9,908,731	13,653,218
aired	196,262,561	196,262,561

NUT2KO



Filtering

	Di-Tag Count
Valid Pairs	154,978,896
Invalid Pairs	41,283,665
Same Circularised	2,896,538
Same Fragment Dangling Ends	2,339,752
Same Fragment Internal	2,652,751
Re-ligation	27,498,222
Contiguous Sequence	1,545,528
Wrong Size	4,350,874
Total Pairs	196,262,561

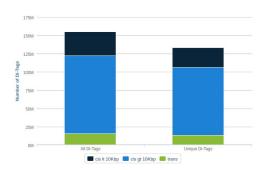


De-duplication

Percentage uniques: 86.02

	All Di-Tags	Unique Di-Tags
Read Pairs	154,978,896	133,305,826
Cis-close (< 10Kbp)	32,615,575	27,359,854
Cis-far (> 10Kbp)	107,050,681	92,673,429
Trans	15,312,640	13,272,543

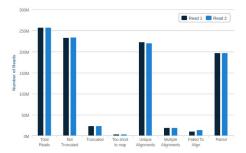
80.57% captured



	Read 1	Read :
Total Reads	256,945,663	256,945,66
Not Truncated	233,091,580	233,703,84
Truncated	23,854,083	23,241,81
Too short to map	3,674,019	3,502,93
Average length of truncated sequence	30.25	30.3

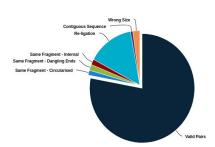
	Read 1	Read 2
Unique Alignments	223,212,597	219,883,055
Multiple Alignments	19,514,372	19,211,373
Failed To Align	10,544,675	14,348,301
Paired	196,570,229	196,570,229

NUT1WT



Filtering

	Di-Tag Count
Valid Pairs	153,286,821
Invalid Pairs	43,283,408
Same Circularised	2,861,675
Same Fragment Dangling Ends	3,431,353
Same Fragment Internal	3,271,191
Re-ligation	28,028,813
Contiguous Sequence	1,435,159
Wrong Size	4,255,217
Total Pairs	196,570,229

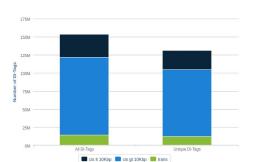


De-duplication

Percentage uniques: 85.44

	All Di-Tags	Unique Di-Tags
Read Pairs	153,286,821	130,971,918
Cis-close (< 10Kbp)	31,366,204	26,223,310
Cis-far (> 10Kbp)	107,687,407	92,515,211
Trans	14,233,210	12,233,397

75.42% captured

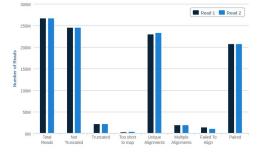


Truncating and Mapping

	Read 1	Read 2
Total Reads	267,094,677	267,094,677
Not Truncated	245,451,408	244,838,689
Truncated	21,643,269	22,255,988
Too short to map	3,283,318	3,462,112
Average length of truncated sequence	30.35	30.22

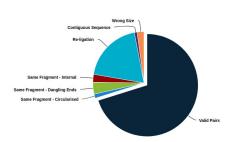
	Read 1	Read 2
Unique Alignments	229,943,623	233,208,196
Multiple Alignments	19,385,856	19,690,134
Failed To Align	14,481,880	10,734,235
Paired	206,896,517	206,896,517

NUT2WT



Filtering

	Di-Tag Count
Valid Pairs	145,245,887
Invalid Pairs	61,650,630
Same Circularised	2,690,781
Same Fragment Dangling Ends	7,740,238
Same Fragment Internal	5,099,724
Re-ligation	39,827,041
Contiguous Sequence	1,816,617
Wrong Size	4,476,229
Total Pairs	206,896,517

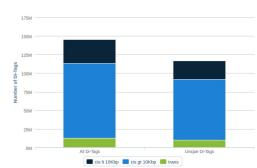


De-duplication

Percentage uniques: 80.40

	All Di-Tags	Unique Di-Tags
Read Pairs	145,245,887	116,773,849
Cis-close (< 10Kbp)	31,757,682	24,991,710
Cis-far (> 10Kbp)	101,009,029	81,700,166
Trans	12,479,176	10,081,973

73.3 % captured



Summary table QC

			HiC	Captured			
Prefix	Biological Replicates	Total reads processed	Total read-pairs	Valid read-pairs	% Cis	% Total Captured	Number of captured reads
моск_ко	2						
	MOCK_1_KO	260.220.163	199.380.409	150.356.474	92,03%	60.64%	55.904.480
	MOCK_2_KO	238.997.995	190.543.734	143.277.381	91,82%	84.07%	87.025.198
MOCK_WT	2						
	MOCK_1_WT	237.206.273	179.267.942	137.856.713	92,99%	60.54%	66.753.587
	MOCK_2_WT	262.846.987	210.134.874	157.219.580	92,61%	83.92%	91.703.030
NUT_KO	2						
	NUTLIN_1_KO	254.246.572	197.626.650	159.589.869	89,90%	81.47%	111.212.666
	NUTLIN_2_KO	254.321.968	196.262.561	154.978.896	90,04%	80.57%	107.409.204
NUT_WT	2						
	NUTLIN_1_WT	256.945.663	196.570.229	153.286.821	90,66%	75.42%	98.779.621
	NUTLIN_2WT	267.094.677	206.896.517	145.245.887	91,37%	73.3%	85.591.406
	MOCK_WT NUT_KO	MOCK_KO 2 MOCK_1_KO MOCK_2_KO MOCK_2_KO MOCK_1_WT 2 MOCK_2_WT MOCK_2_WT NUT_KO 2 NUTLIN_1_KO NUTLIN_2_KO NUT_WT 2 NUTLIN_1_WT	Prefix Replicates processed MOCK_KO 2	Prefix Biological Replicates Total reads processed Total read-pairs MOCK_KO 2	MOCK_KO 2 read-pairs read-pairs MOCK_KO 2 199.380.409 150.356.474 MOCK_1_KO 260.220.163 199.380.409 150.356.474 MOCK_2_KO 238.997.995 190.543.734 143.277.381 MOCK_WT 2 197.267.942 137.856.713 MOCK_2_WT 262.846.987 210.134.874 157.219.580 NUT_KO 2 197.626.650 159.589.869 NUTLIN_1_KO 254.246.572 197.626.650 159.589.869 NUT_WT 2 196.262.561 154.978.896 NUT_WT 2 196.570.229 153.286.821	Prefix Biological Replicates Total reads processed Total read-pairs Valid read-pairs % Cis MOCK_KO 2 4 4 4 4 4 4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Prefix Biological Replicates Total reads processed Total read-pairs Valid read-pairs % Cis % Total Captured MOCK_KO 2

Different number of reads

Downsampled bams

Downsampling bams according to MOCK1KO (sample with less reads → **55.904.480 captured reads)**

Pre-CHICAGO correlation matrix

NUT2WT		0.970	0.961	0.958	0.954	0.956	0.950	0.930	Pearson Correlation 0.94 0.96 0.98 1.00
NUT1WT	0.970		0.964	0.963	0.953	0.954	0.942	0.925	Fig 1. Correlation matrix of the biological replicates based on their
NUT2KO	0.961	0.964		0.972	0.959	0.963	0.951	0.941	number of reads per interaction using the CHiCAGO input (chinput) was performed using Absolute
NUT1KO	0.958	0.963	0.972		0.958	0.962	0.949	0.938	Pearson Correlation-based distance. This matrix shows the correlation
MOCK2WT	0.954	0.953	0.959	0.958		0.976	0.962	0.948	among samples analysed at the same period experimentally. We can clearly observe a separation
MOCK2KO	0.956	0.954	0.963	0.962	0.976		0.966	0.954	among the samples treated with nutlin3a and the control.
MOCK1WT	0.950	0.942	0.951	0.949	0.962	0.966		0.969	Nevertheless, we can also see that the correlations goes together with the experimental analysis order
MOCK1KO	0.930	0.925	0.941	0.938	0.948	0.954	0.969		(MOCK2WT, MOCK2KO / MOCK1WT, MOCK1KO /
	NUT2WT	NUT1WT	NUT2KO	NUT1KO	MOCK2WT	MOCK2KO	MOCK1WT	MOCK1KO	NUT1WT, NUT1KO, NUT2WT, NUT2KO.

Data processing and analysis: CHiCAGO

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0992-2

Method | Open Access | Published: 15 June 2016

CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data

CHiCAGO ("Capture Hi-C Analysis of Genomic Organisation")

<u>Jonathan Cairns, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser & Mikhail Spivakov</u> ⊡

Genome Biology 17, Article number: 127 (2016) Cite this article

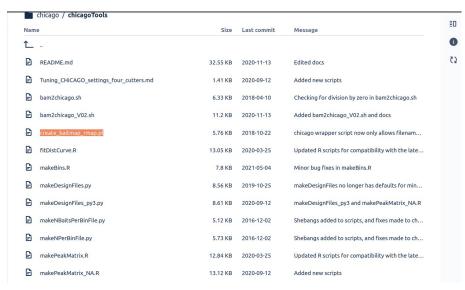
https://bitbucket.org/chicagoTeam/chicago/src/master/

Assumption that "significant" interactions emerge as outliers on a distance-dependent local background profile. The background levels in CHi-C decrease as the genomic distance between the bait and other end increases.

CHiCAGO features a novel background correction procedure and a **two-component convolution background model accounting for both real, but expected, interactions** as well as assay and sequencing artefacts.

Before start, you will need:

- 1. Five restriction map information files ("design files"):
- **Restriction map file** (.rmap) a bed file containing coordinates of the restriction fragments. By default, 4 columns: chr, start, end, fragmentID.
- **Bait map file** (.baitmap) a bed file containing coordinates of the baited restriction fragments, and their associated annotations. By default, 5 columns: chr, start, end, fragmentID, baitAnnotation. The regions specified in this file, including their fragmentIDs, must be an exact subset of those in the .rmap file. The baitAnnotation is a text field that is used only to annotate the output and plots.
- *nperbin* file (.npb), *nbaitsperbin* file (.nbpb), *proxOE* file (.poe) Precompute these tables from the .rmap and .baitmap files, using the Python script makeDesignFiles.py from *chicagoTools* at our <u>Bitbucket repository</u>. Refer to the chicagoTools README file for more details.



2) You will also need input data files, which should be in CHiCAGO input format, .chinput. You can obtain .chinput files from aligned Capture Hi-C BAM files by running bam2chicago.sh, available as part of chicagoTools. (To obtain BAM files from raw fastq files, use a Hi-C alignment & QC pipeline such as <u>HiCUP</u>.

```
3) library(Chicago)
cd <- setExperiment(designDir = testDesignDir, settingsFile = settingsFile)
cd <- readAndMerge(files=files, cd=cd)
cd <- chicagoPipeline(cd)
```

*Default calibration: monocytes. → Understand the data sparsity for your samples and re-calibrate.

*seed is needed

4) Outputs

CHiCAGO outputs:

- ibed format
- seqmonk format
- washU_text format

```
## 1 20 119103 138049 DEFB126 11 5.12

## 2 20 161620 170741 DEFB128 11 5.12

## 3 20 119103 138049 DEFB126 6 6.81

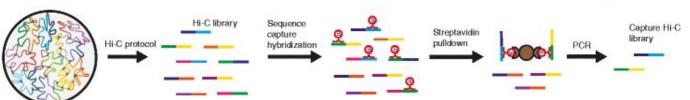
## 4 20 523682 536237 CSNK2A1 6 6.81

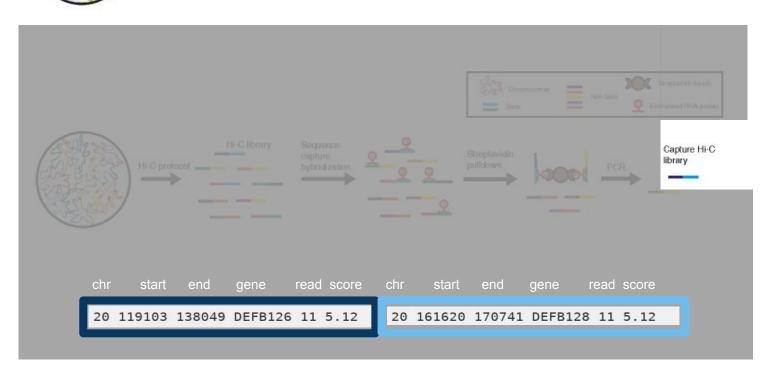
## 5 20 161620 170741 DEFB128 16 5.17

## 6 20 73978 76092 . 16 5.17
```





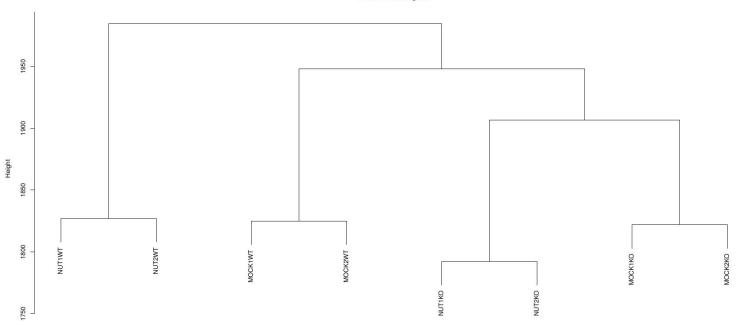




How to compare samples? **PeakMatrix**



Cluster Dendrogram



Summary table QC

https://docs.google.com/spreadsheets/d/1g1ELBSpffGPcdRFwoWKK7xzxGI3OMXXZrWNw-yQwSss/edit#gid=0

			HICUP					otured	Downsampling			
Samples	Prefix	Biological Replicates	Total reads processed	Total read-pairs	Valid read-pairs	% Cis	% Total Captured	Number of captured reads	Number of captured reads (after downsampling)	Significant interactions (individual replicates after downsampling)	Significant cis	
PCHiC HCT116-/- MOCK	моск_ко	2										
2		MOCK_1_KO	260.220.163	199.380.409	150.356.474	92,03%	60.64%	55.904.480	55.904.480	101.280	99.189	
		MOCK_2_KO	238.997.995	190.543.734	143.277.381	91,82%	84.07%	87.025.198	55.861.268	103.793	102.126	
PCHiC HCT116+/+ MOCK	MOCK_WT	2										
		MOCK_1_WT	237.206.273	179.267.942	137.856.713	92,99%	60.54%	66.753.587	55.872.622	108.175	106.724	
100		MOCK_2_WT	262.846.987	210.134.874	157.219.580	92,61%	83.92%	91.703.030	55.937.362	109.413	107.866	
PCHiC HCT116-/- Nutlin3a	NUT_KO	2										
0		NUTLIN_1_KO	254.246.572	197.626.650	159.589.869	89,90%	81.47%	111.212.666	55.946.714	104.016	102.530	
		NUTLIN_2_KO	254.321.968	196.262.561	154.978.896	90,04%	80.57%	107.409.204	55.857.922	104.858	103.430	
PCHiC HCT116+/+ Nutlin3a	NUT_WT	2										
70 EV		NUTLIN_1_WT	256.945.663	196.570.229	153.286.821	90,66%	75.42%	98.779.621	55.904.975	111.934	110.401	
		NUTLIN_2WT	267.094.677	206.896.517	145.245.887	91,37%	73.3%	85.591.406	55.886.848	111.390	109.815	

