


CHiCAGO: Statistical methodology for signal detection in Capture Hi-C data

Jonathan Cairns

jonathan.cairns@babraham.ac.uk

 @jonathancairns

Fraser/Spivakov labs, Babraham Institute

4th October 2016



Table of Contents

1 Introduction

2 The CHiCAGO model

3 Results

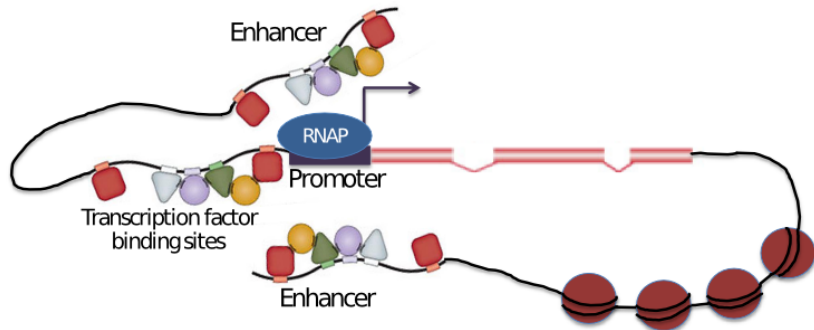
Table of Contents

1 Introduction

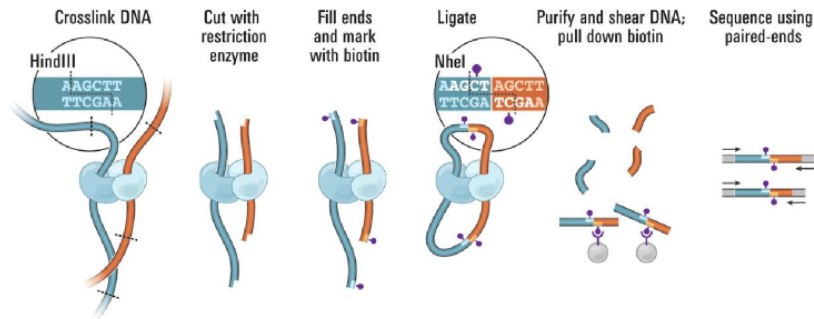
2 The CHiCAGO model

3 Results

Motivation

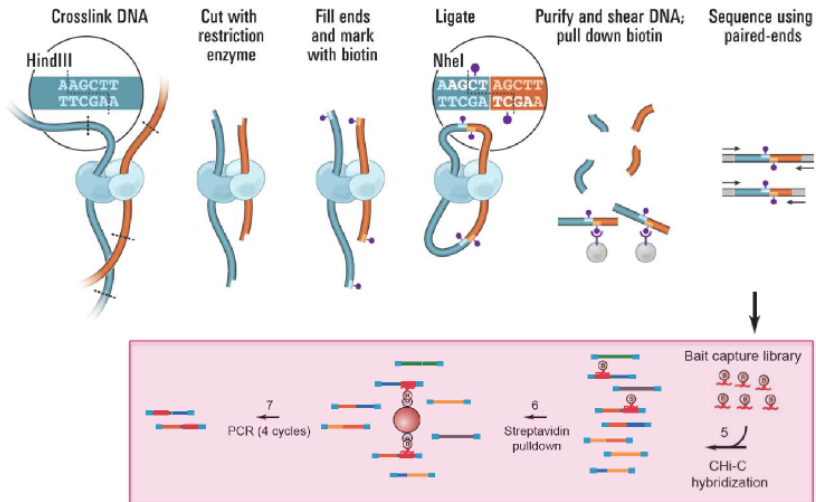


Chi-C: improved resolution at promoters, over Hi-C



Lieberman-Aiden *et al* (2009)

CHi-C: improved resolution at promoters, over Hi-C



- Approx. 12-fold increase in read coverage

Schönfelder *et al* (2015), Mifsud *et al* (2015), Sahlén *et al* (2015)

- Align reads & filter out artefacts with HiCUP

The data

- Align reads & filter out artefacts with HiCUP
- Obtain counts X_{ij} :

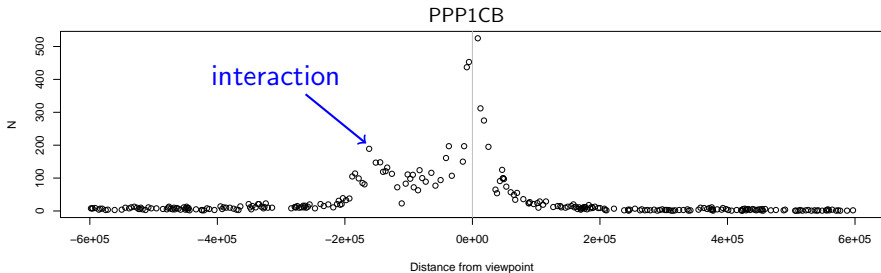
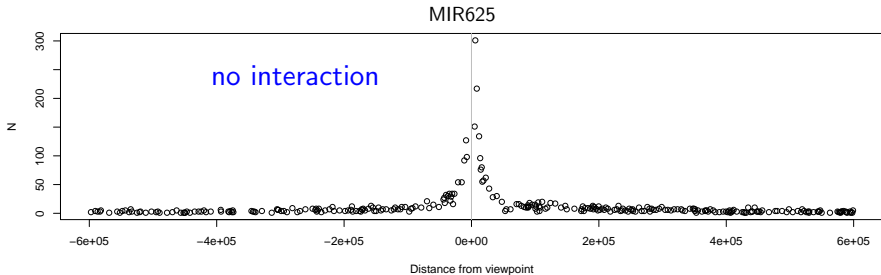


Table of Contents

1 Introduction

2 The CHiCAGO model

3 Results

CHiCAGO – Capture Hi-C Analysis of Genomic Organization.

Genome Biology

[ABOUT](#)[ARTICLES](#)[SUBMISSION GUIDELINES](#)[METHOD](#) | [OPEN ACCESS](#)

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

[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](#) and [Mikhail Spivakov](#) 

[†] Contributed equally

Genome Biology 2016 17:127 | DOI: 10.1186/s13059-016-0992-2 | © The Author(s). 2016

Received: 1 April 2016 | Accepted: 25 May 2016 | Published: 15 June 2016

[Download PDF](#)[Export citations >](#)[Table of Contents ^](#)[Abstract](#)[Background](#)[Results](#)[Discussion](#)

Background comes from two sources:

Background comes from two sources:

	Brownian	Technical
Source	Random collisions	Sequencing artefacts

Background comes from two sources:

	Brownian	Technical
Source	Random collisions	Sequencing artefacts
Depends on distance?	Yes (decreasing)	No

Background comes from two sources:

	Brownian	Technical
Source	Random collisions	Sequencing artefacts
Depends on distance?	Yes (decreasing)	No
Dominates	Close to bait	Far from bait

Background comes from two sources:

	Brownian	Technical
Source	Random collisions	Sequencing artefacts
Depends on distance?	Yes (decreasing)	No
Dominates	Close to bait	Far from bait

Under H_0 (no interaction), counts are sum of the two components:

$$X_{ij} = B_{ij} + T_{ij}$$

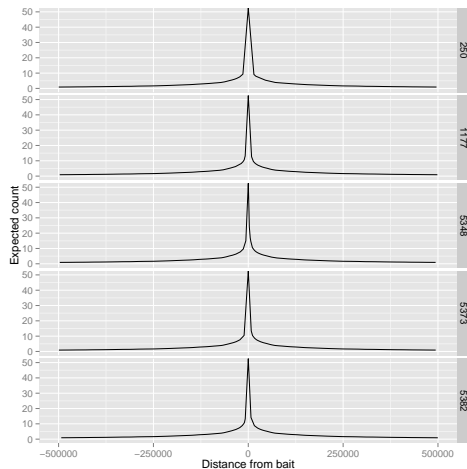
Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

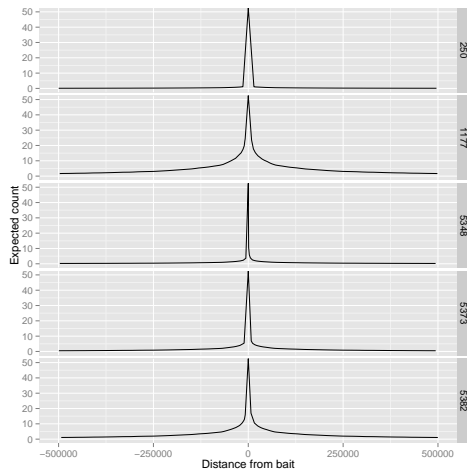
$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij})$$



Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

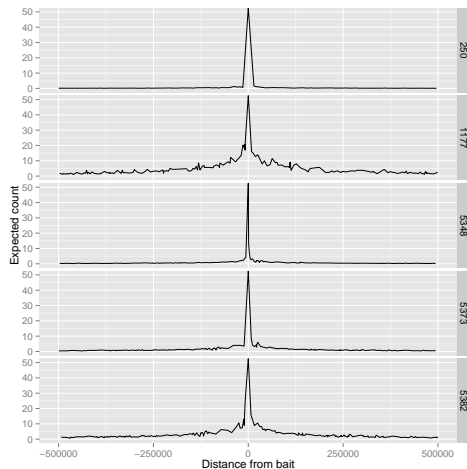
$B_{ij} \sim \text{NB}$, with $\mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j$



Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$B_{ij} \sim \text{NB}$, with $\mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$



Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

$f(d)$:

- estimated close to bait (< 1.5Mb) in 20kb bins.

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

$f(d)$:

- estimated close to bait ($< 1.5\text{Mb}$) in 20kb bins.
- bin-wise estimates $f(d_b)$ from geometric mean across baits

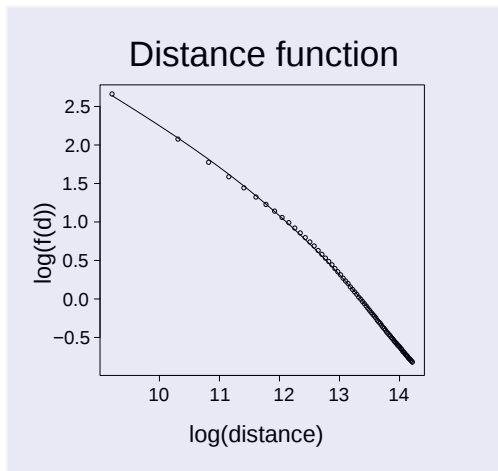
Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

$f(d)$:

- estimated close to bait (< 1.5Mb) in 20kb bins.
- bin-wise estimates $f(d_b)$ from geometric mean across baits
- interpolation: cubic fit on log-log scale



Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

- Bait-specific bias:

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

- Bait-specific bias:
 - Get bin-wise estimates for each bait.
 - Take median across bins – robust to interactions

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

- Bait-specific bias:
 - Get bin-wise estimates for each bait.
 - Take median across bins – robust to interactions
- Other-end-specific bias:

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$B_{ij} \sim \text{NB}, \text{ with } \mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$$

- Bait-specific bias:
 - Get bin-wise estimates for each bait.
 - Take median across bins – robust to interactions
- Other-end-specific bias:
 - Too sparse to estimate individually
 - Assume *trans*-chromosomal reads are mostly noise
 - Pool other-ends by *trans* counts
 - Estimate bias parameter, pool-wise
 - Bait-to-bait interactions treated separately

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$B_{ij} \sim \text{NB}$, with $\mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$

- Bait-specific bias:
 - Get bin-wise estimates for each bait.
 - Take median across bins – robust to interactions
- Other-end-specific bias:
 - Too sparse to estimate individually
 - Assume *trans*-chromosomal reads are mostly noise
 - Pool other-ends by *trans* counts
 - Estimate bias parameter, pool-wise
 - Bait-to-bait interactions treated separately
- Dispersion parameter

Brownian background estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$B_{ij} \sim \text{NB}$, with $\mathbb{E}(B_{ij}) = f(d_{ij}) \times (\text{bait bias})_j \times (\text{other end bias})_i$

- Bait-specific bias:
 - Get bin-wise estimates for each bait.
 - Take median across bins – robust to interactions
- Other-end-specific bias:
 - Too sparse to estimate individually
 - Assume *trans*-chromosomal reads are mostly noise
 - Pool other-ends by *trans* counts
 - Estimate bias parameter, pool-wise
 - Bait-to-bait interactions treated separately
- Dispersion parameter
 - Established maximum likelihood methods.

Technical noise estimation

$$X_{ij} = B_{ij} + T_{ij}$$

Technical noise estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$T_{ij} \sim \text{Pois}(\lambda_{ij})$$

Technical noise estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$T_{ij} \sim \text{Pois}(\lambda_{ij})$$

- Estimated entirely from *trans*-chromosomal reads

Technical noise estimation

$$X_{ij} = B_{ij} + T_{ij}$$

$$T_{ij} \sim \text{Pois}(\lambda_{ij})$$

- Estimated entirely from *trans*-chromosomal reads
- Pool baits and other-ends
- Pool-wise estimate: average number of reads per pair of *trans* fragments.

$$X_{ij} = B_{ij} + T_{ij}$$

- B is Negative Binomial, T is Poisson.
- $\Rightarrow X$ has Delaporte distribution.
- One-sided hypothesis test – Observed more than expected by chance?
- Get p -value

Statistical model - p -value weighting

- Simple p -value thresholding (even using Bonferroni/FDR)
→ many false positives (typically, at large distances, with only one read).

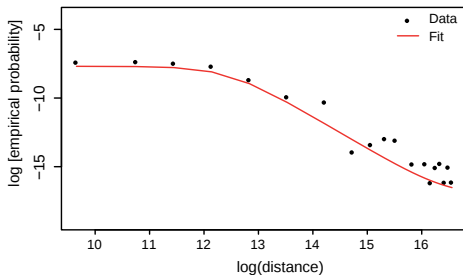
Statistical model - p -value weighting

- Simple p -value thresholding (even using Bonferroni/FDR)
 - many false positives (typically, at large distances, with only one read).

At large distances:

- far fewer reproducible interactions

Empirical probability of reproducible interaction



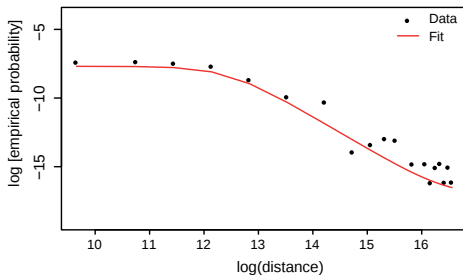
Statistical model - p -value weighting

- Simple p -value thresholding (even using Bonferroni/FDR)
→ many false positives (typically, at large distances, with only one read).

At large distances:

- far fewer reproducible interactions
- but vast majority of tests performed there

Empirical probability of reproducible interaction



Statistical model - p -value weighting

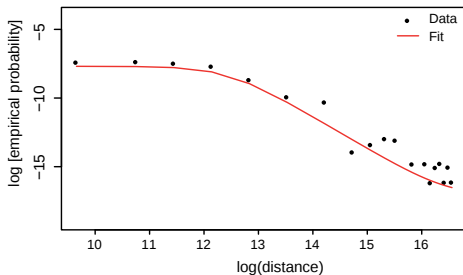
- Simple p -value thresholding (even using Bonferroni/FDR)
→ many false positives (typically, at large distances, with only one read).

At large distances:

- far fewer reproducible interactions
- but vast majority of tests performed there

So, large-distance false positives dominate.

Empirical probability of reproducible interaction



Statistical model - p -value weighting

- Simple p -value thresholding (even using Bonferroni/FDR)
 - many false positives (typically, at large distances, with only one read).

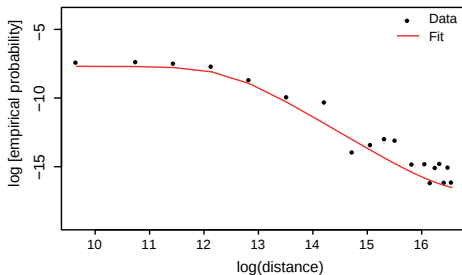
At large distances:

- far fewer reproducible interactions
- but vast majority of tests performed there

So, large-distance false positives dominate.

Solution: p -value weighting (Genovese *et al*, 2009) to downweight long-distance interactions

Empirical probability of reproducible interaction



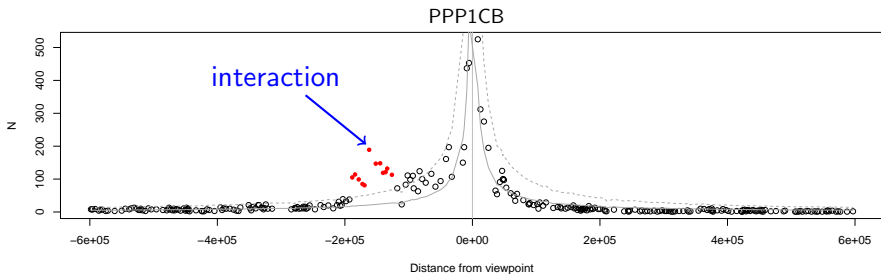
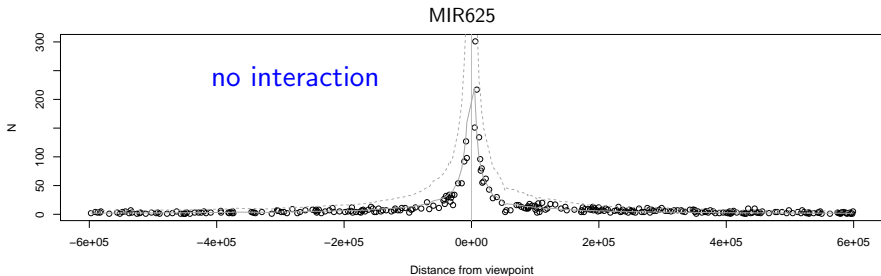


Table of Contents

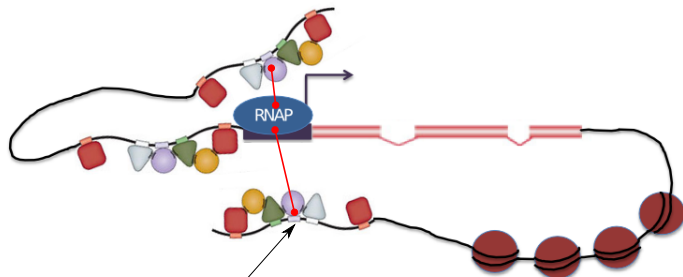
1 Introduction

2 The CHiCAGO model

3 Results

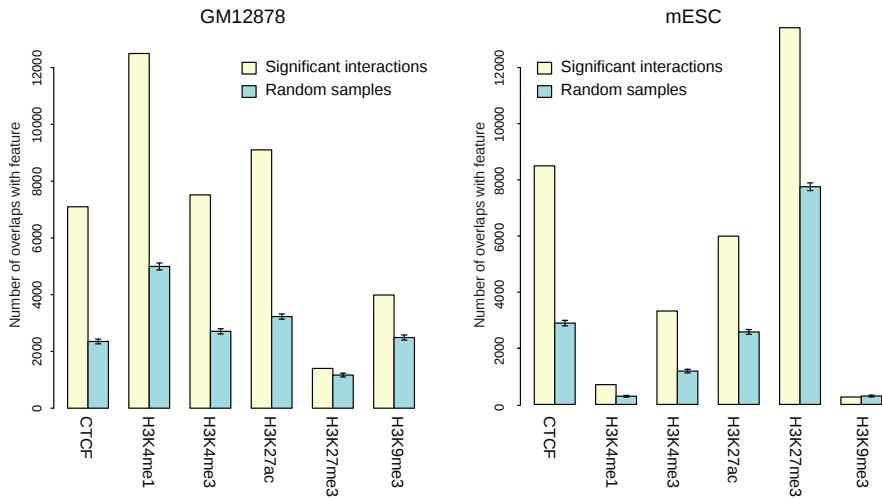
Downstream analysis

- *CHiCAGO*-derived interactions give us “Promoter-Interacting Regions” (PIRs).



Histone marks?
SNPs?
Other features?

Histone marks – significant enrichment at other ends

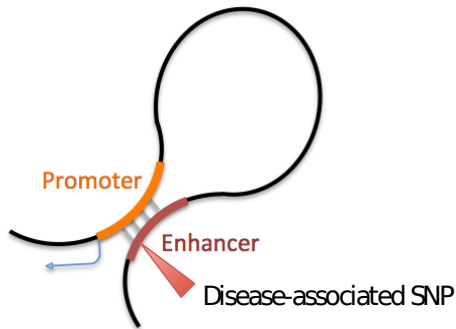


Paula Freire Pritchett

Interactions in blood cells

Javierre* / Burren* / Wilder* / Kreuzhuber* / Hill* et al. (in press)
Genomic regulatory architecture links disease variants to target genes.

- PChI-C in 17 blood cell types (primary cells)
- “Interactomes” found to be cell type-specific, matching lineage tree



- **CHiCAGO** finds interactions in Capture Hi-C data:
 - robustly
 - having normalised for various sources of bias
 - using p -value weighting (to account for variable true positive rate)
- Results provide biological understanding:
 - can detect cell type-specific interactions.
 - can show enrichment for histone marks.
 - can link disease-associated SNPs to their target genes.

Acknowledgements



CHiCAGO developers

- **Paula Freire Pritchett**
- **Steven W. Wingett**
- **Mikhail Spivakov**

Statistical Advice

- **Vincent Plagnol**
(UCL/Inivata)
- **Daniel Zerbino**
(EBI)

Additional Downstream Analysis

- **Csilla Várnai**
- **Andrew Dimond**

Data

- **Biola Javierre**
- **Stefan Schönfelder**
- **Cameron Osborne**
(KCL)
- **Peter Fraser**

<http://www.regulatorygenomicsgroup.org/chicago>



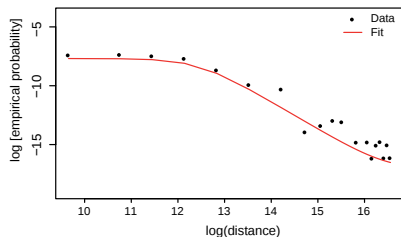
We make prior “guesses” U_{ij} . We allow U_{ij} to depend on d_{ij} , assuming that short-range interactions are more likely than long-range interactions, with a smooth transition between the two. The U_{ij} are transformed into weights W_{ij} by dividing through by the mean value, \bar{U} , ensuring that the average W_{ij} value is 1. Finally, weighted p -values are obtained by dividing the p -values by their respective weights:

$$Q_{ij} = \frac{p_{ij}}{W_{ij}}$$

We now specify the U_{ij} model in our particular context. (next slide)

p-value weighting

Empirical probability of reproducible interaction



Bounded logistic regression model: U_{ij} is assumed a function of both d_{ij} and a vector of parameters $\Theta = (\alpha, \beta, \gamma, \delta)$, according to

$$U_{ij} = \eta_{ij} U_{\max} + (1 - \eta_{ij}) U_{\min}$$

where

$$\eta_{ij} = \text{expit}(\alpha + \beta \log(d_{ij}))$$

$$U_{\min} = \text{expit}(\gamma)$$

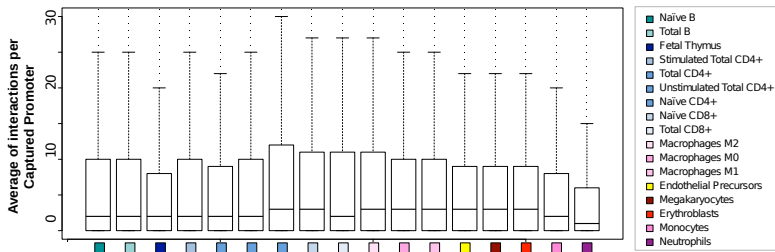
$$U_{\max} = \text{expit}(\delta)$$

Numbers of called interactions

- # interactions per sample: 130,000 – 190,000

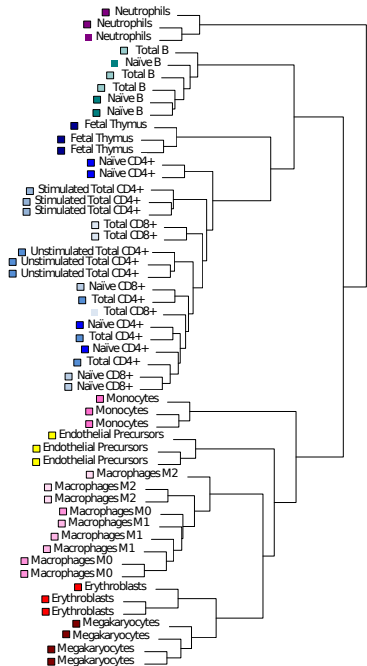
Numbers of called interactions

- # interactions per sample: 130,000 – 190,000
- # interactions per captured promoter:



Cell Type	Processed Reads	Capture Unique Valid Reads	Significant Interactions
■ Megakaryocytes	2,696,317,863	653,848,788	150,203
■ Erythroblasts	2,338,677,291	588,786,672	144,771
■ Neutrophils	2,241,977,639	736,055,569	131,609
■ Monocytes	1,942,858,536	572,357,387	151,389
■ Macrophages M0	2,125,716,849	668,675,248	163,791
■ Macrophages M1	2,067,485,594	497,683,496	163,399
■ Macrophages M2	2,055,090,022	523,561,551	173,449
■ Naïve B	2,127,262,739	629,928,642	171,439
■ Total B	1,874,130,921	702,533,922	183,119
■ Fetal Thymus	2,728,388,103	776,491,344	145,577
■ Naïve CD4+	2,797,861,611	844,697,853	192,048
■ Total CD4+	2,227,386,686	836,974,777	166,668
■ Unstimulated Total CD4+	2,034,344,692	721,030,702	177,371
■ Stimulated Total CD4+	1,971,143,855	749,720,649	188,714
■ Naïve CD8+	1,910,881,702	747,834,572	187,399
■ Total CD8+	1,849,225,803	628,771,947	183,964
■ Endothelial Precursors	2,308,749,174	420,536,621	141,382
	37,297,499,080	11,299,489,740	2,816,292
		* HICUP	*CHICAGO

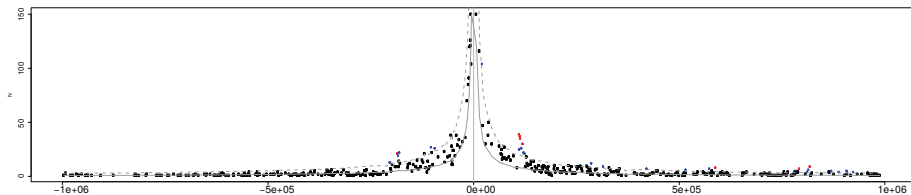
Distance
400 500 600 700



Sven Sewitz

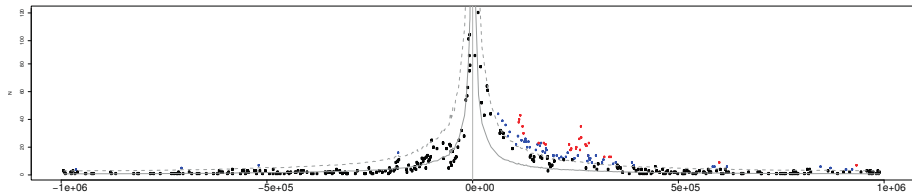
Megakaryocyte_D5_6_step2_chicago2

789407 - AP1S2



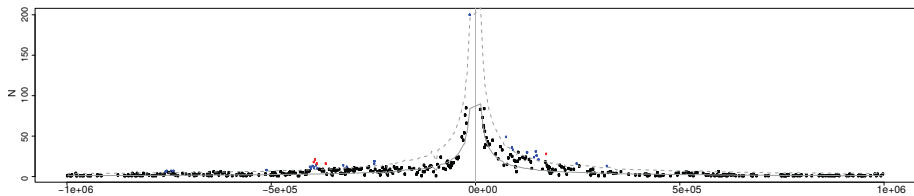
B1_Monocyte_D1_step2_chicago2

789407 - AP1S2



Megakaryocyte_D5_6_step2_chicago2

410124 - CD93-002,...



B1_Monocyte_D1_step2_chicago2

410124 - CD93-002,...

