Statistical methods for analyzing the Hi-C contact matrices, a brief overview

KKY

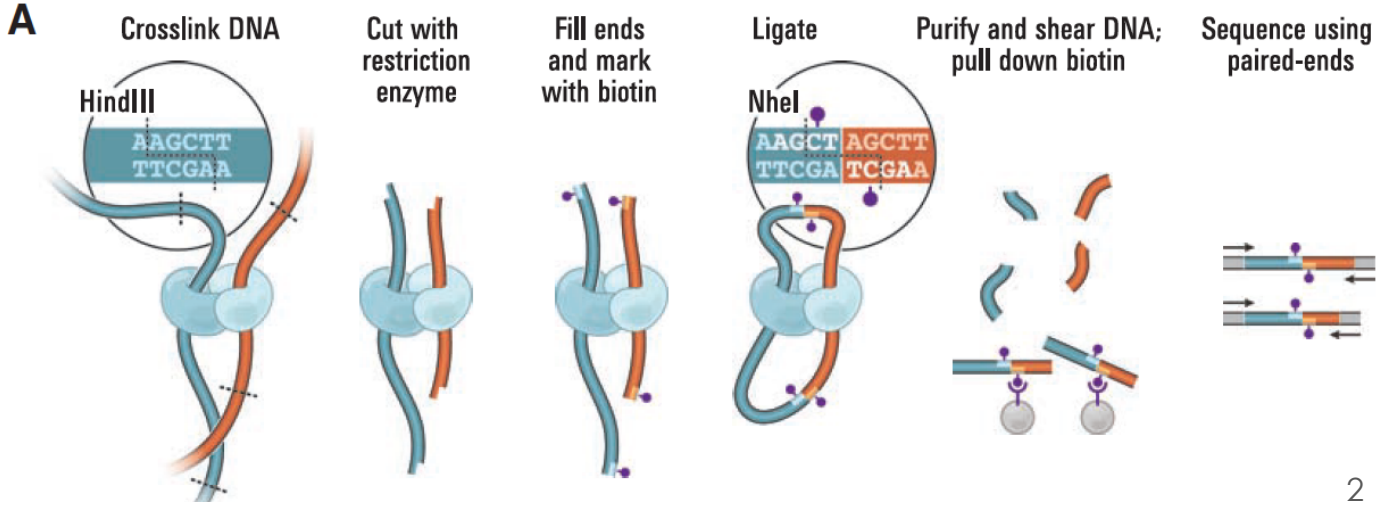
Gerstein Lab JC, July 2013

Hi-C experiment

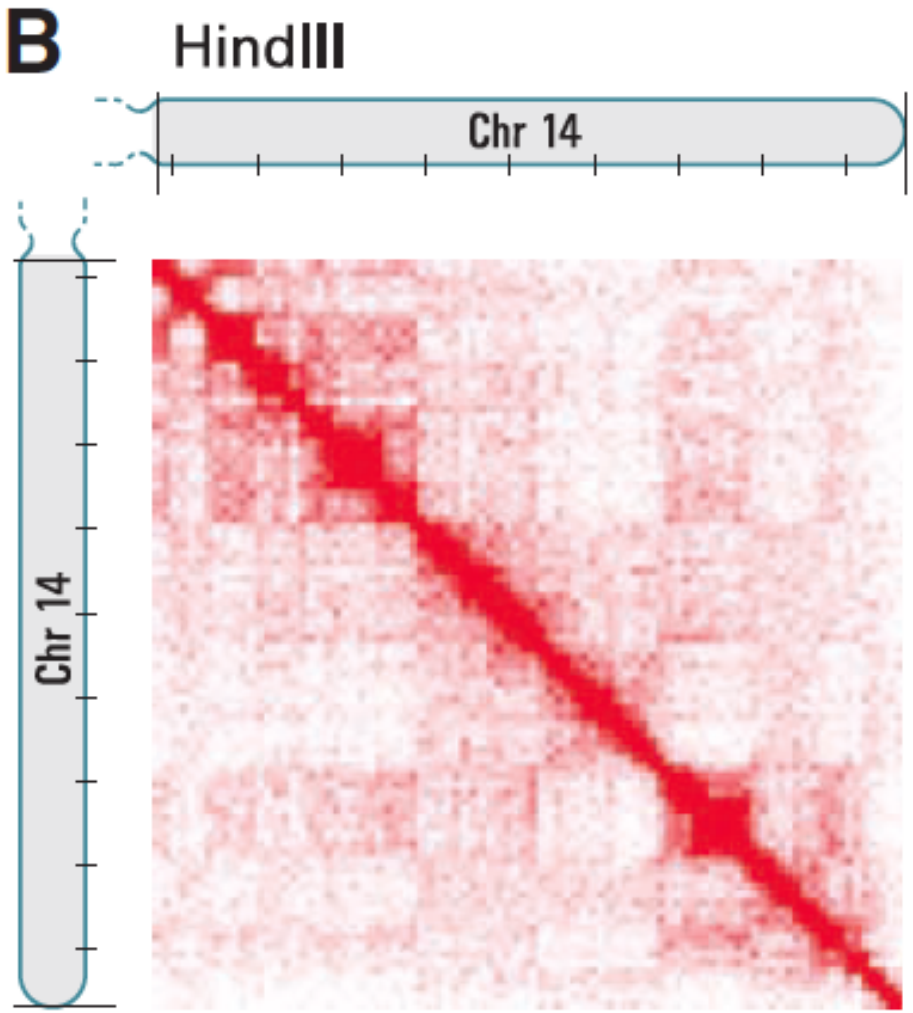
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity



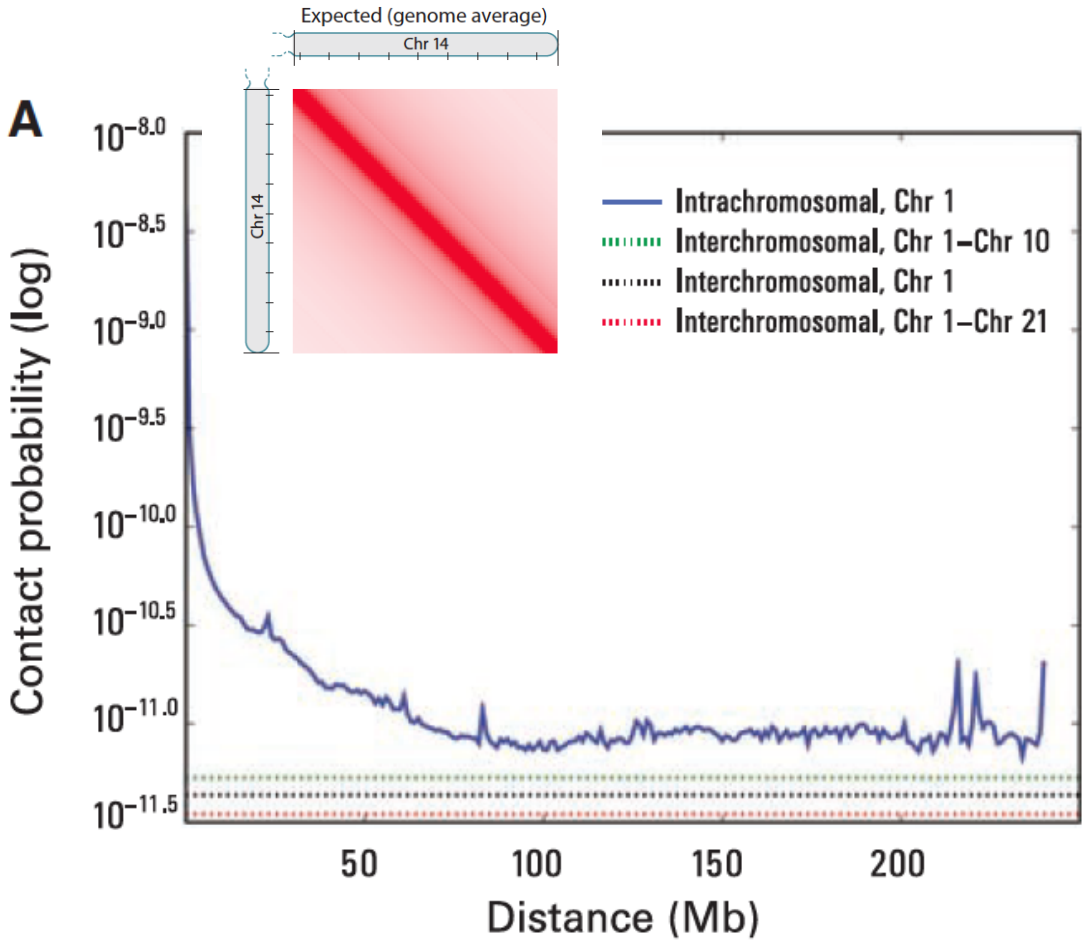
Contact Matrix



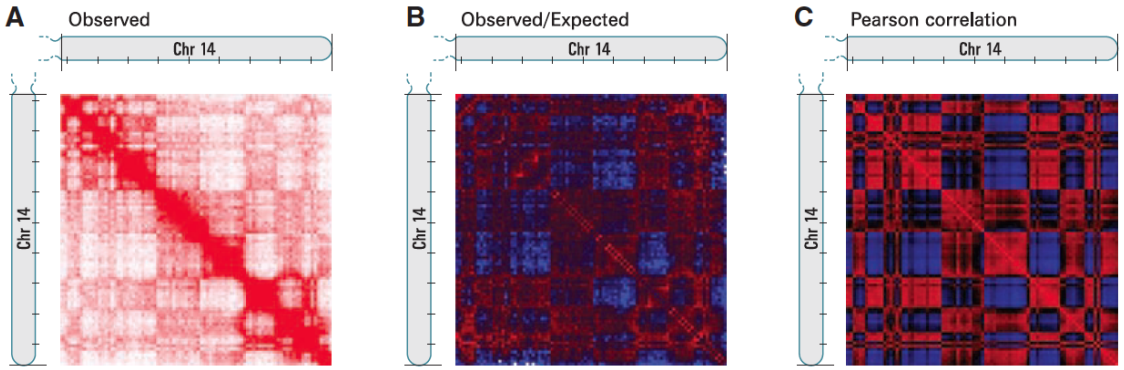
9 OCTOBER 2009 VOL 326 SCIENCE www.sciencemag.org

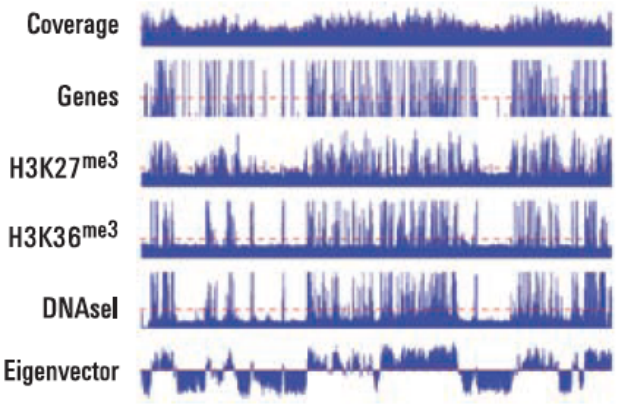
Normalization of the matrix

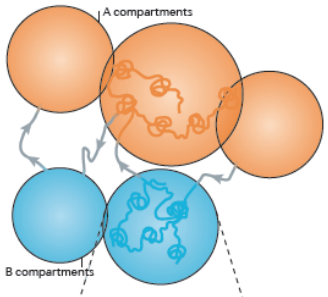
- Artificial separation of intrachromosomal and interchromosomal interactions
  - Intrachromosomal expectation as a function of distance: Obs/Exp
  - Interchromosomal expectation –
    number of reads*f_i*f_j
    Obs/Exp

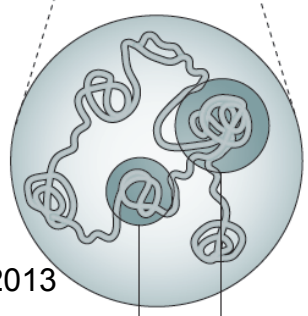


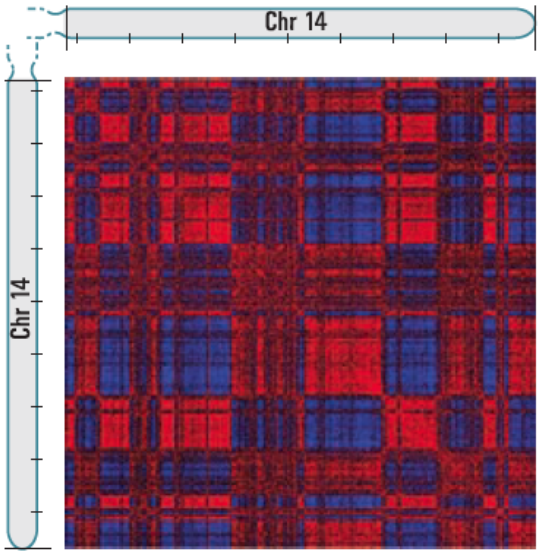
Eigenvectors: Genome compartments











Dekker et al. Nat. Rev. Genetics 2013

A more sophisticated error model

ANALYSIS

nature genetics

Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture

Eitan Yaffe & Amos Tanay

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Correspondence should be addressed to A.T. (amos.tanay@weizmann.ac.il).

Received 16 February; accepted 25 August; published online 16 October 2011; doi:10.1038/ng.947

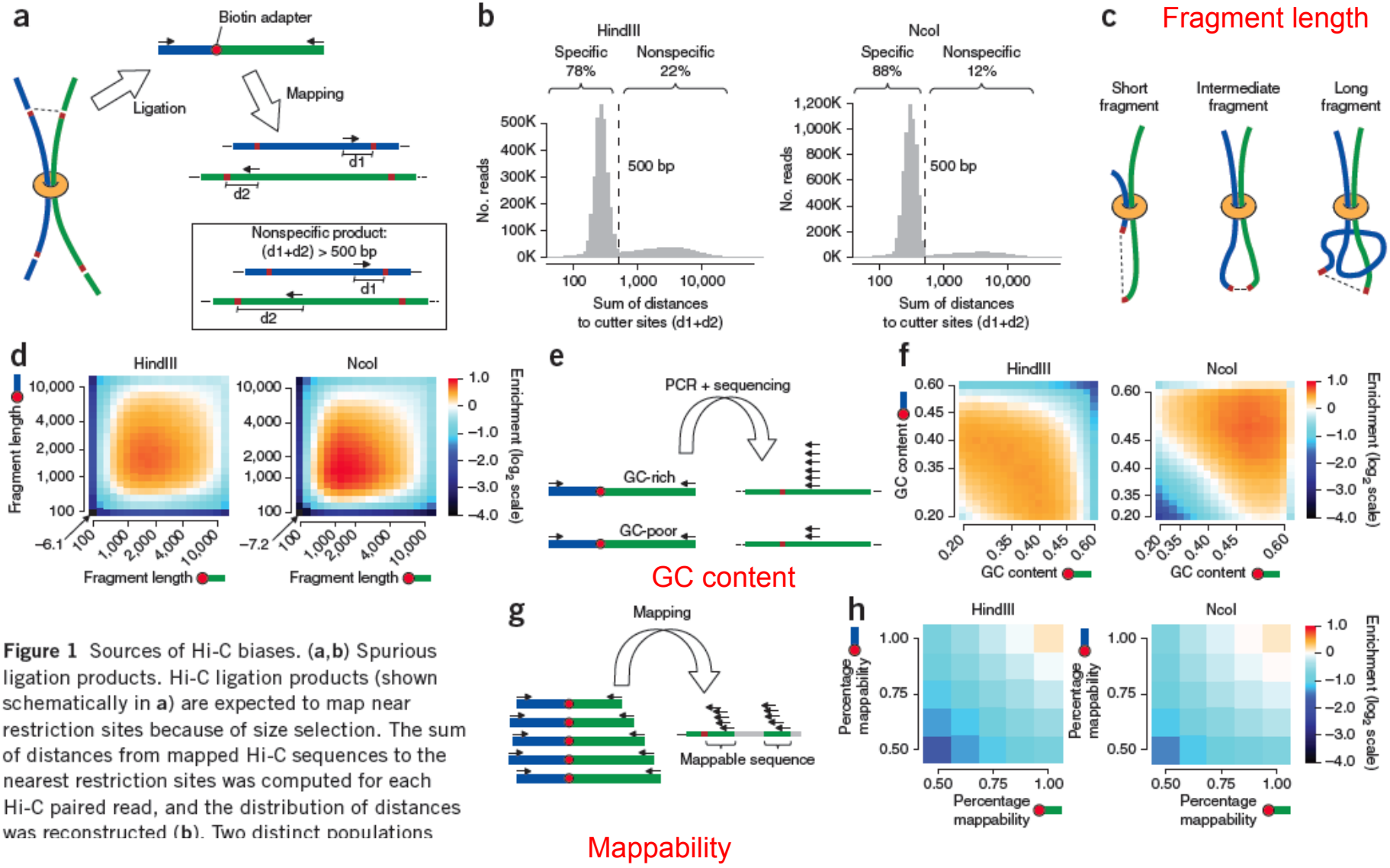


Figure 1 Sources of Hi-C biases. (**a**,**b**) Spurious ligation products. Hi-C ligation products (shown schematically in **a**) are expected to map near restriction sites because of size selection. The sum of distances from mapped Hi-C sequences to the nearest restriction sites was computed for each Hi-C paired read, and the distribution of distances was reconstructed (**b**). Two distinct populations

Model

Given 2 fragment ends a, b, the probability to observe them in a pair-end read

$$P(X_{a,b}) = P_{\text{prior}} \cdot F_{len}(a_{len}, b_{len}) \cdot F_{\text{gc}}(a_{\text{gc}}, b_{\text{gc}}) \cdot M(a) \cdot M(b),$$

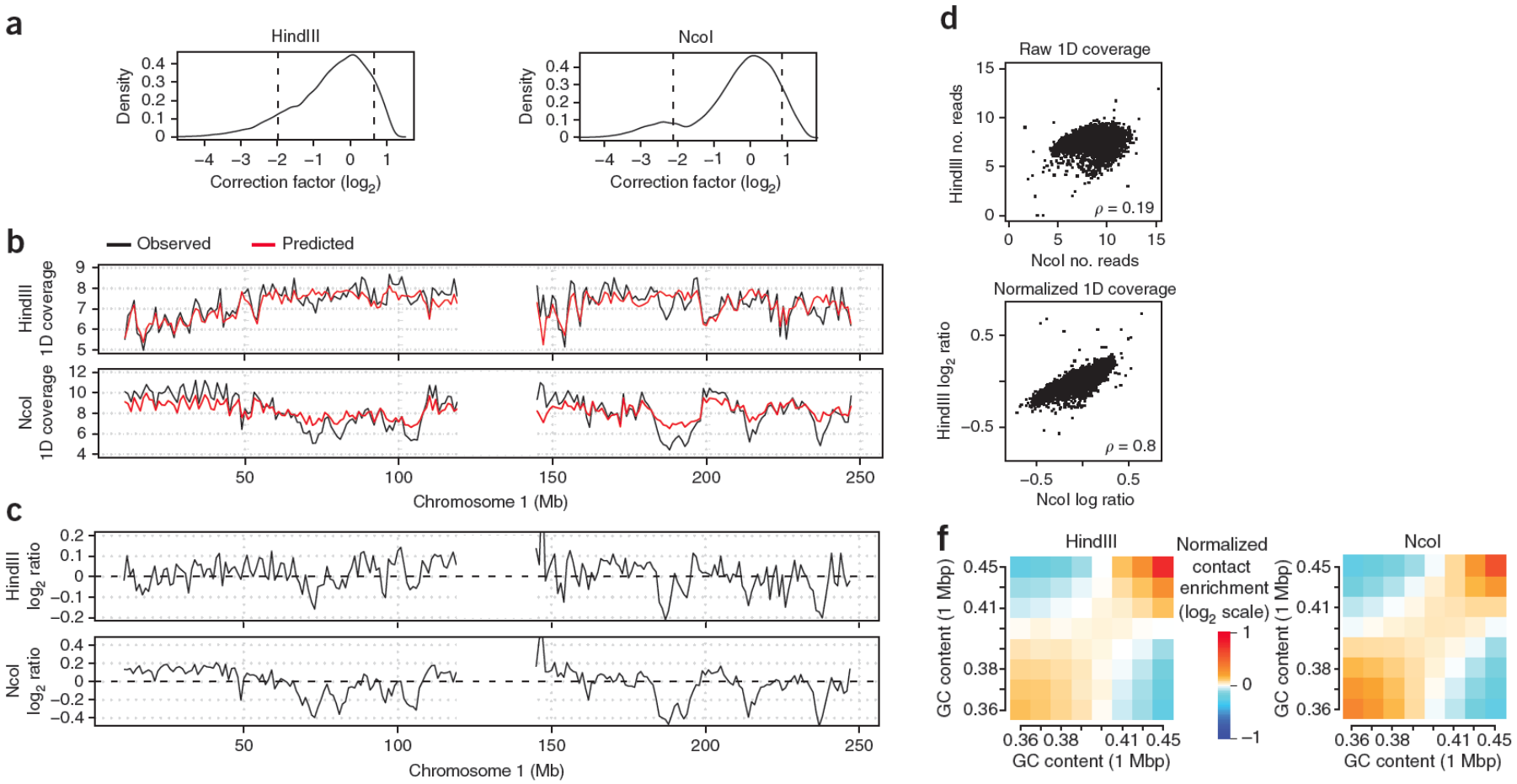
20 by 20 matrices 800 parameters

10^6 pair-end reads -> 10^{12} combinations

$$L(F_{len}, F_{\text{gc}}) = \prod_{\{a,b\} \in I} P(X_{a,b}) \cdot \prod_{\{a,b\} \notin I} (1 - P(X_{a,b}))$$

$$= \prod_{c = (a_{len}, a_{\text{gc}}, b_{len}, b_{\text{gc}})} P(X_{a,b})^{n_c} \cdot [1 - P(X_{a,b})]^{m_c}$$

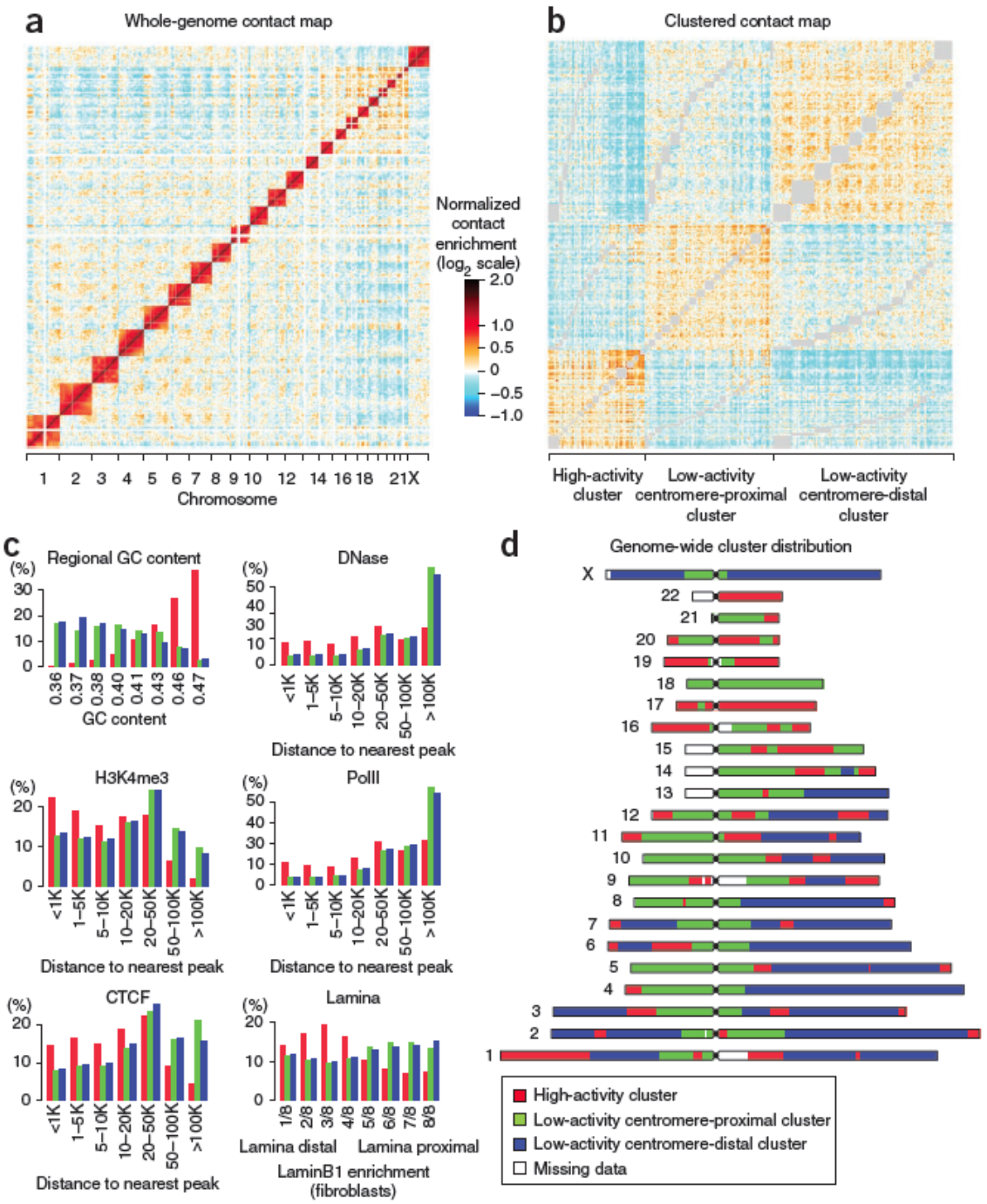
After correction



Clustering of the corrected matrix

Remarks:

- assume particular sources of bias
- very computationally expensive: 800 parameters, all possible pairs of fragments

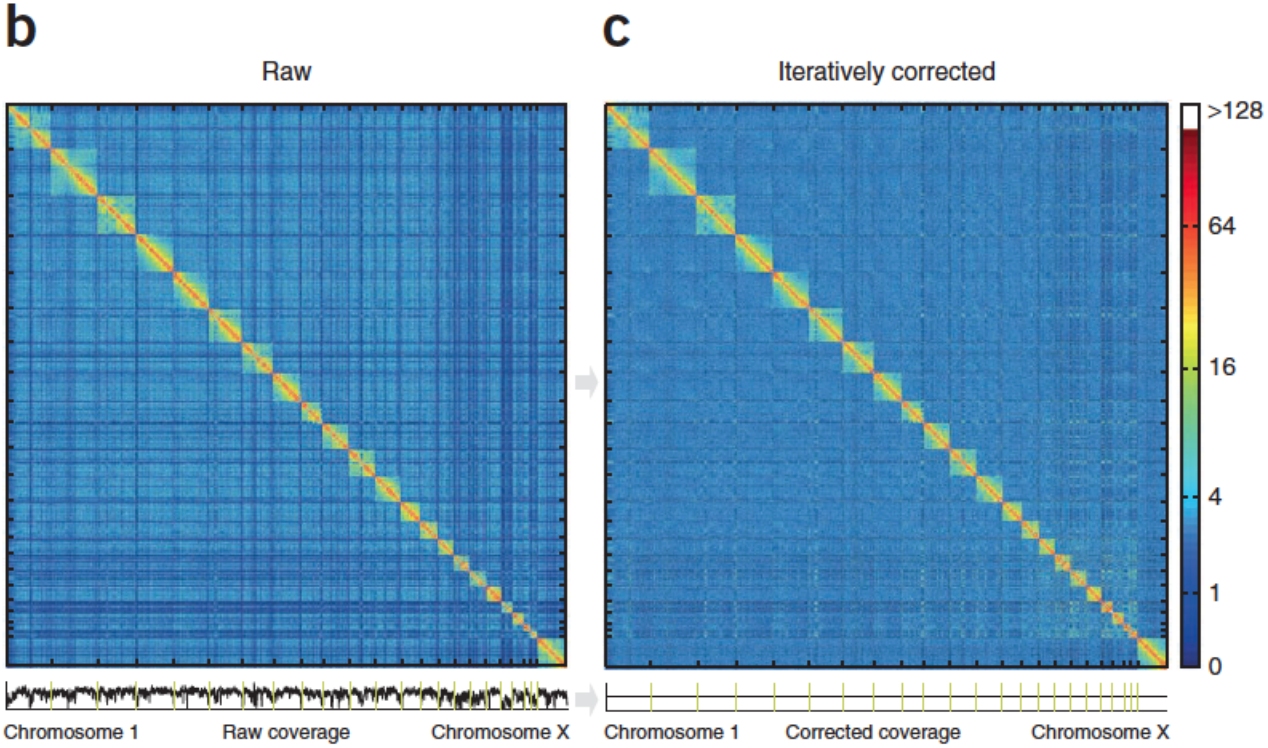


A “better” approach: ICE

Iterative correction of Hi-C data reveals hallmarks of chromosome organization

Maxim Imakaev[1,5], Geoffrey Fudenberg[2,5], Rachel Patton McCord[3], Natalia Naumova[3], Anton Goloborodko[1], Bryan R Lajoie[3], Job Dekker[3] & Leonid A Mirny[1,2,4]

NATURE METHODS | VOL.9 NO.10 | OCTOBER 2012 | 999



Assumption:
All loci should have equal visibility
sum of each row (column)=1

Model

Matrix T: “true” relative contact probabilities $\sum_j T_{ij} = 1$ equal visibility

For each loci i, there is a bias factor B_i (summarize all sorts of sources)
the expected contact frequency between i and j is given by $B_i B_j T_{ij}$

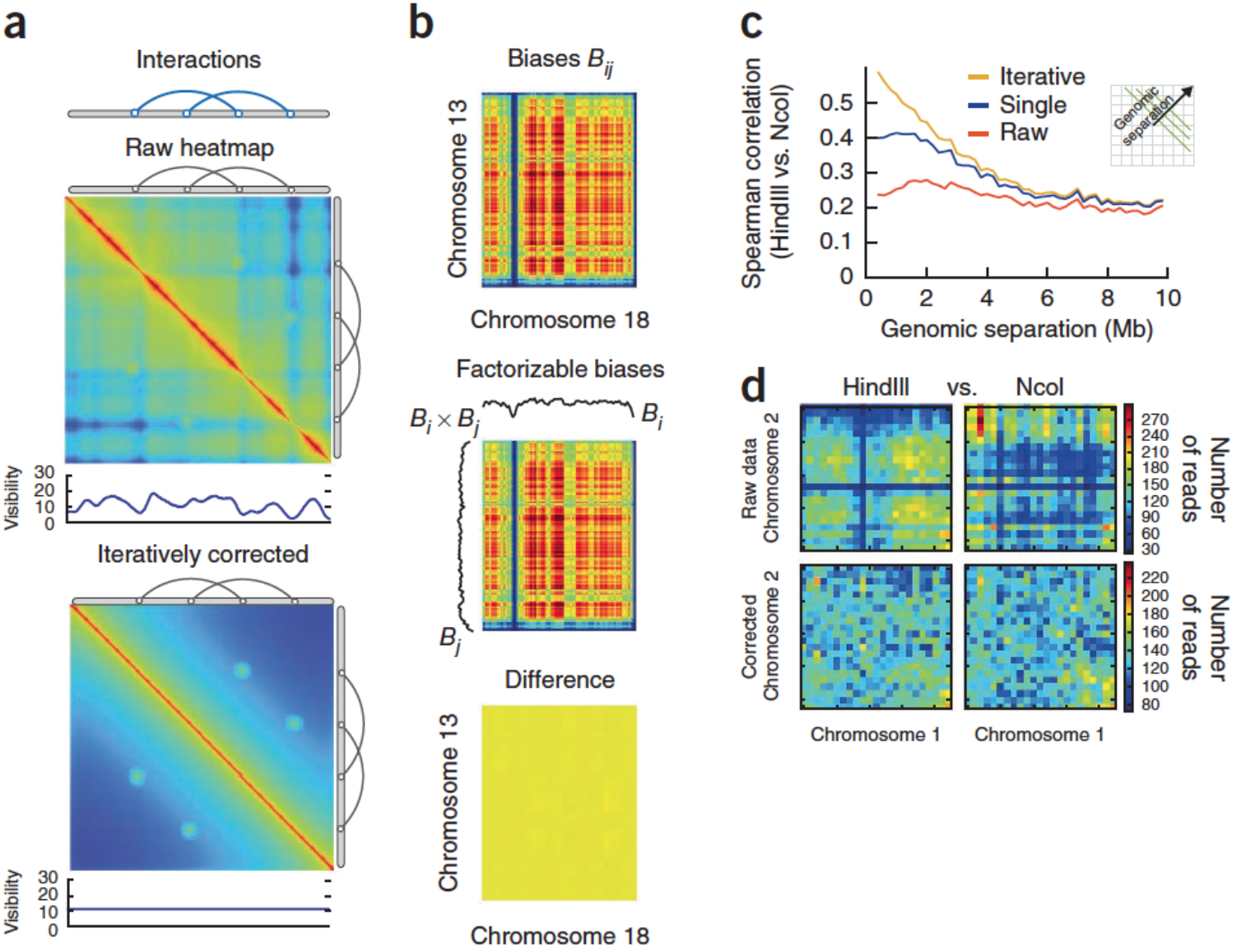
Given the observed contact frequency $O_{ij,}$ they inferred the values of T and B by maximum likelihood.

$$O_{ij} = B_i B_j T_{ij}$$

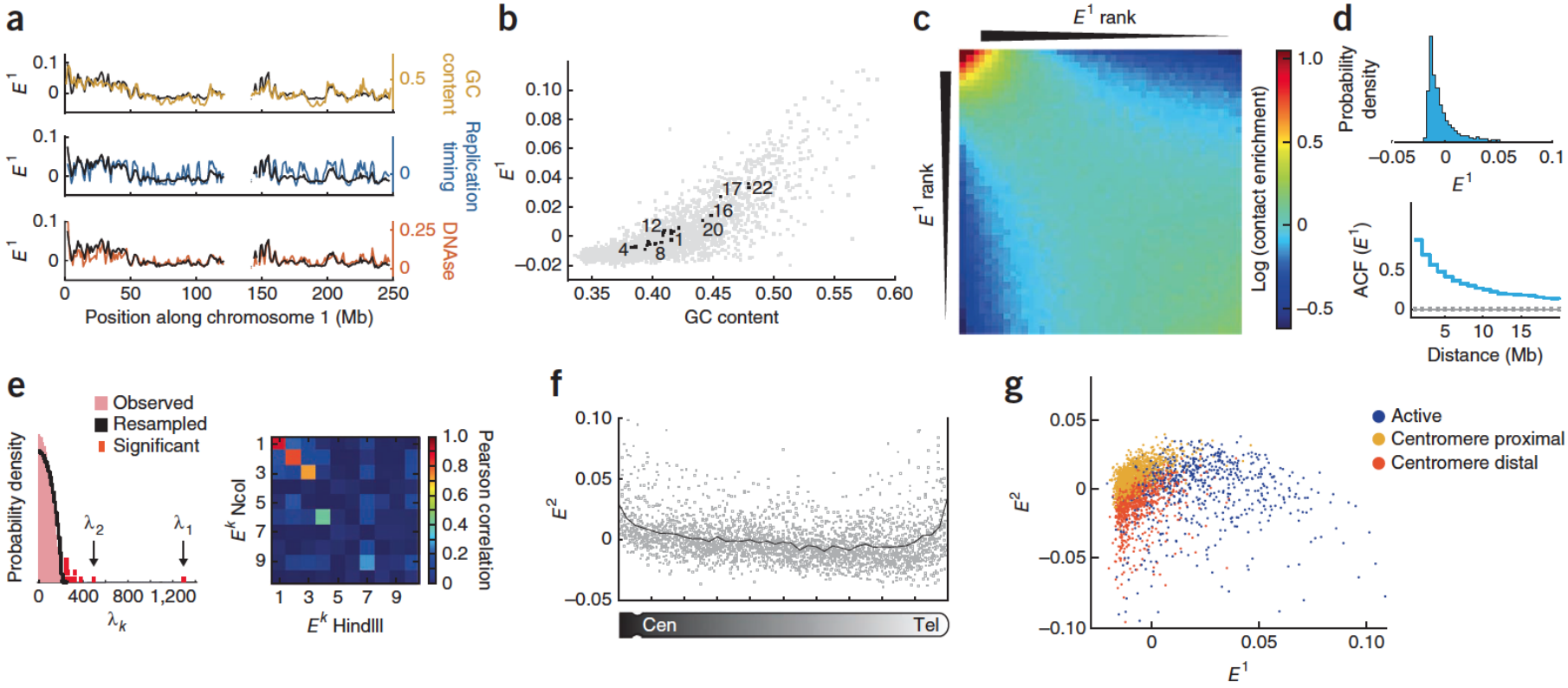
$$\sum_{j=1}^{N} T_{ij} = 1, \forall i$$

Solve equations by iteration.

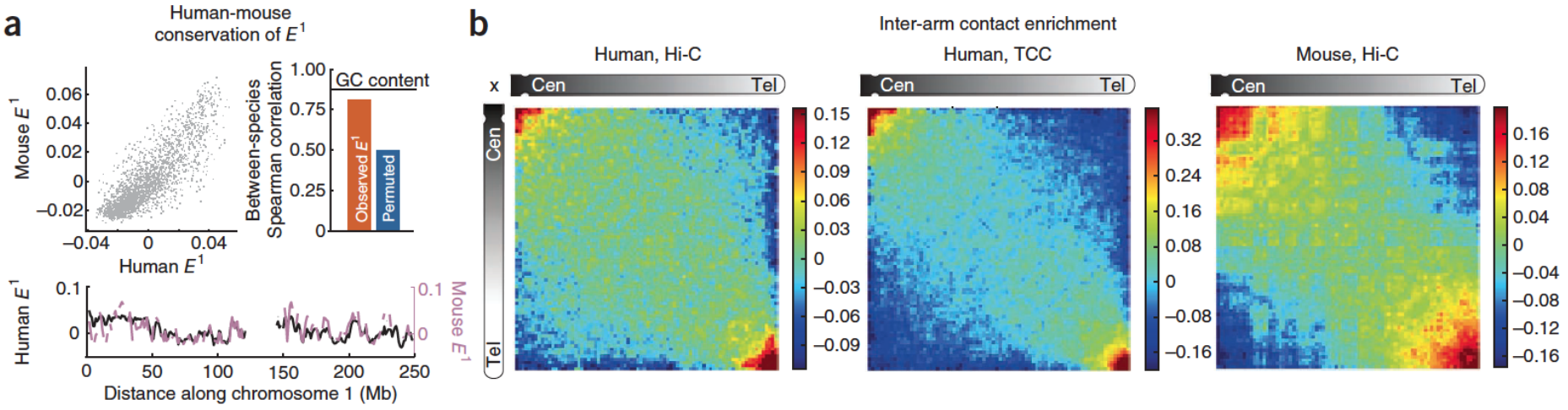
Validation


a
Interactions
Raw heatmap
Visibility
30
20
10
0
Iteratively corrected
Visibility
30
20
10
0
b
Biases B_{ij}
Chromosome 13
Chromosome 18
Factorizable biases
$B_i \times B_j$
B_i
B_j
Difference
Chromosome 13
Chromosome 18
c
Spearman correlation (HindIII vs. NcoI)
0.5
0.4
0.3
0.2
0.1
0
Iterative
Single
Raw
Genomic separation
0
2
4
6
8
10
Genomic separation (Mb)
d
HindIII
vs.
NcoI
Raw data Chromosome 2
Corrected Chromosome 2
270
240
210
180
150
120
90
60
30
220
200
180
160
140
120
100
80
Number of reads
Number of reads
Chromosome 1
Chromosome 1

Biology revealed from eigenvectors of T



Cross species analysis



Pro and Limitation of ICE

- Unbiased comparison of Hi-C data between data sets, cell types and organisms
- Only for genome-wide matrix of contacts, NOT suitable for techniques like ChIA-PET compared to Yaffe and Tanay
- Operates on binned data, there's a resolution limit compared to Yaffe and Tanay