Information items	Should this be included? Why?	How to report the information?
Protocol information (e.g. in situ Hi-C, restriction enzyme used, crosslinking method, etc)		*Full reference protocol.*
Metadata (*who performed the experiments, when, protocol specific information such as biotin labeling, amplification cycle numbers, etc*)		
Cell sample information (*cell pellet characterization, tissue processing procedures, cell line passage numbers, etc*.)		
Sequencing information (instrument, sequencing depth and length, optic duplicates, etc)		
Sequencing QC information (read quality metric, etc) *[[[Sequenced Read Pairs, Normal Paired, Chimeric Paired, Chimeric Ambiguous, Unalignable, Ligation Motif Present, Alignable (Normal+Chimeric Paired), Unique Read Pairs, PCR Duplicates, Optical Duplicates, Library Complexity Estimate, Intra-fragment Read Pairs, Below MAPQ Threshold, Hi-C Contacts, Ligation Motif Present, 3' Bias (Long Range), Pair Type % (L-I-O-R), Inter-chromosomal, Intra-chromosomal, Short Range (<20Kb), Long Range (>20Kb)]]]*		
Initial data processing (alignment algorithms, *with*		

parameters used and reference genome)		
Sample description (Species, cell type, treatment condition, etc)		
Percentage of cis reads		
Reproducibility metric		
Resolution metric		
Raw chromatin contact matrix		
Normalized contact matrix		
Chromatin domain calls	This seems outside scope of metadata. It is data analysis	
Chromatin loop calls	This seems outside scope of metadata. It is data analysis	

Hi-C Library Statistics and Quality Control

Excerpted from:
Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell, 2014.

A Hi-C experiment can fail in a number of ways. Failures are sometimes obvious when viewing Hi-C heatmaps, but their underlying cause can be difficult to diagnose. Here, we describe quality metrics that we calculate on all of our Hi-C libraries. Together, these quality metrics can detect a variety of failure mechanisms. Note that, prior to performing a high-resolution Hi-C experiment, we often sequenced 200K – 2M reads from a "test aliquot" before deciding whether the library quality was sufficiently high to justify deep sequencing. As an example, Table S2 shows library statistics for our *in situ* GM12878 libraries used in our primary and replicate experiments. Note that the "replicate experiment" discussed in the main text is in fact the aggregate of the eight biological replicates shown in Table S2. We discuss these statistics in detail below.

II.d.1. Standard sequencing and alignment statistics: At the top of the table we record a series of statistics that characterize the sequencing and alignment. In a high-quality sequencing run, few reads are unmapped. If more than 10% of reads fail to align, it typically indicates either a problem with the sequencing run, or sample contamination. The frequency of chimeras is an indicator of the frequency of long-range ligation junctions in the data, although the specific value seen depends on numerous experimental parameters, such as aligner and read length. A sudden anomaly in this value as compared to experiments with similar parameters can suggest a failure of the ligation step.

II.d.2. Duplicate frequency: High duplication rate indicates low molecular complexity (Lander and Waterman, 1988). A Hi-C library was considered a good candidate for deeper sequencing if our complexity estimates suggested it consisted of hundreds of millions or billions of unique contacts. In general, in constructing high resolution Hi-C maps, we found that it was more efficient to sequence many replicate libraries at lower depth (e.g. one lane on an Illumina HiSeq, or less than 20% of the total library complexity) rather than sequencing many lanes of a single library, since the latter strategy typically leads to extensive duplication. In addition, the duplicate removal step requires vastly more time and memory if a single library is sequenced very deeply.

For instance, if we estimated that an otherwise high-quality library contains 1.5 billion contacts, we usually planned to sequence at most two lanes from that library on an Illumina HiSeq (~300M reads total).

The exact molecular complexity was calculated using the Picard tools formulation of the Lander-Waterman equation (Lander and Waterman, 1988), which entails solving the following equation for the molecular complexity *m* (measured in molecules):

$R/m = 1 – e^{N/m}$

Here *R* is the number of distinct reads observed, *N* is the total number of reads sequenced. Note that "optical duplicates" created by the Illumina sequencing process rather than PCR were not included in either *R* or *N*. In our experience, the resulting value is typically an underestimate of the true library complexity.

II.d.3. Fraction of "Hi-C contacts": After duplication removal, we filter out read pairs where both ends align to the same fragment. If this step filters out over 20% of read pairs, it indicates that the library failed in the restriction, fill-in or ligation steps of the protocol and thus is not a good candidate for deeper sequencing.

We also filter out read pairs where the mapping quality of either read falls short of the desired threshold. (In the example of Table S2, the threshold is MAPQ > 0.) The rest of the quality metrics are calculated using the list of read pairs that remain once all filtering was completed. We refer to these read pairs as "contacts."

II.d.4. Ligations: This statistic measures how often a ligation junction is found inside a read. (A ligation junction is the sequence created when the ends of two filled-in restriction fragments ligate to one another. For MboI, the ligation junction sequence is GATCGATC. For HindIII, the sequence is AAGCTAGCTT.) A paucity of ligation junctions in a Hi-C library suggests that the ligation failed. Note that there can be many causes of such a failure, ranging from a bad batch of DNA ligase to rupture of the cell nuclei. This statistic is also dependent on sequence read length and insert size. We typically sequenced a 300-500bp insert using 101bp PE reads, and observed that ligation rates tended to fall into the 30-40% range. Of course, with shorter reads and longer insert sizes, this value tends to be smaller. With longer reads and shorter inserts, it is larger.

II.d.5. Proximity to 5' and 3' restriction fragment ends: For long-range contacts (defined as intrachromosomal and over 20Kb apart, or interchromosomal), we looked at both read ends to see if the end is closer to the 5' or 3' end of the restriction fragment and to which strand the read maps. When Hi-C libraries are generated using a six-cutter restriction enzyme and, after the shearing step, are size selected for 300-500bp molecules, we find that the large fragment size to insert size ratio causes most contacts (>85%) to come from the 3' ends of fragments. A much lower value indicates that the restriction enzyme had not cut effectively. Note that when the fragment size to insert size ratio declines (i.e., when a four-cutter restriction enzyme is used) we find that the 5' to 3' bias is markedly attenuated.

II.d.6. Percentage of contacts at various distances: We broke down contacts into intrachromosomal and interchromosomal contacts. We then further subdivided the intrachromosomal contacts to short range (<20 Kb) and long range (>20 Kb) contacts.

A crucial metric is the percentage of long-range intrachromosomal contacts. In successful Hi-C libraries, we found that at least 15% of unique reads were long-range intrachromosomal contacts. Lower values usually indicated that the experiment had failed. If more than 40% of unique reads are long-range intrachromosomal contacts, a library was considered a good candidate for sequencing. If the fraction was above half, a library was considered an excellent candidate for sequencing. In general, this value was one of the statistics we found most important to scrutinize in performing cost-effective high-depth Hi-C.

A library with many interchromosomal contacts and a paucity of contacts at shorter distances (i.e., absence of both a strong diagonal and robust distance decay effects in the intrachromosomal contact matrices) suggests that the library comprises mostly random ligation products, likely due to the rupture of a large fraction of nuclei.

II.d.7. Percentage of contacts by read pair type: We broke down intrachromosomal contacts by type: in a "left" pair, both ends map to the reverse strand. In a "right" pair, both reads map to the forward strand. In an "inner" pair, the ends map to different strands and point (5' to 3') towards each other. In an "outer" pair, reads land on opposite strands but point away from one another. If the chimeras observed are due to proximity ligation, this statistic should be random, i.e., each pair type should account for roughly 25% of contacts. Thus, the distance at which the percentage of each pair type converges to 25% is a good indication of the minimum distance at which it is meaningful to examine Hi-C contact patterns. For six-cutter restriction enzymes, such as HindIII and NcoI, this distance is approximately 30 Kb. For four-cutter restriction enzymes (MboI, DpnII), this distance is approximately 3 Kb (Figure S1D). Note that the existence of read pairs in the "right" and "left" configuration is rarely seen outside of Hi-C experiments. DNA-Seq reads, for instance, are by design all "inner" pairs; "jumping libraries" tend to produce outer pairs.

Table S2 (Excerpts)

	Primary (HIC*_br1)	Bio Rep 1 (HIC*_br2)	Bio Rep 2 (HIC*_br3)	Bio Rep 3 (HIC*_br4)	Bio Rep4 (HIC*_br5)
Sequenced Reads	3.6B	314M	389M	178M	669M
Normal Paired	2.7B (75%)	244M (78%)	305M (78%)	124M (70%)	477M (71%)
Chimeric Paired	563M (16%)	48M (15%)	59M (15%)	41M (23%)	124M (18%)
Chimeric Ambiguous	153M (4%)	11M (4%)	12M (3%)	7M (4%)	26M (4%)
Unmapped	187M (5%)	11M (4%)	13M (3%)	6M (3%)	43M (6%)
Alignable Reads	3.2B (90%)	291M (93%)	364M (93%)	165M (93%)	600M (90%)
Duplicates	300M (8%)	42M (13%)	15M (4%)	3M (2%)	18M (3%)
Unique Reads	2.9B (81% / 100%)	250M (80% / 100%)	348M (89% / 100%)	163M (91% / 100%)	582M (87% / 100%)
Intra-fragment	43M (1% / 1%)	2M (1% / 1%)	20M (5% / 6%)	6M (3% / 4%)	26M (4% / 5%)
Low Mapping Quality	268M (7% / 9%)	22M (7% / 9%)	31M (8% / 9%)	15M (8% / 9%)	55M (8% / 9%)
HiC Contacts	2.6B (73% / 89%)	226M (72% / 91%)	297M (76% / 85%)	142M (79% / 87%)	501M (75% / 86%)
Inter chromosomal	644M (18% / 22%)	51M (16% / 20%)	70M (18% / 20%)	31M (18% / 19%)	105M (16% / 18%)
Intra chromosomal	2B (55% / 68%)	176M (56% / 70%)	227M (58% / 65%)	110M (62% / 68%)	395M (59% / 68%)
Intra Short Range (<20 Kb)	602M (17% / 20%)	47M (15% / 19%)	82M (21% / 23%)	38M (21% / 24%)	147M (22% / 25%)
Intra Long Range (≥20 Kb)	1.4B (39% / 47%)	129M (41% / 51%)	145M (37% / 42%)	72M (40% / 44%)	248M (37% / 43%)
Ligations	950M (27% / 32%)	80M (26% / 32%)	97M (25% / 28%)	76M (42% / 47%)	222M (33% / 38%)
3’ bias (long range)	68% - 32%	69% - 31%	69% - 31%	75% - 25%	72% - 28%
Read Pair type (L-I-O-R)	25-25-25-25	25-25-25-25	25-25-25-25	25-25-25-25	25-25-25-25