SOFTWARE TOOL ARTICLE

(ý#50ǒ ĄýĄùĂýĂù úăĈ āõĄĄýĂû õĂø ĄĈă÷ùććýĂû (ýù# øõĈõ ǽĊùĈćýăĂ ˷Ĭ Ĉùúùĉùùćǒ ˰ õĄĄĈăĊùøĬ ˷ õĄĄĈăĊùø ċýĈü ĉùćùĈĊõĈýăĂćǾ

Steven Wingett[1,2], Philip Ewels[3], Mayra Furlan-Magaril[2], Takashi Nagano[2], Stefan Schoenfelder[2], Peter Fraser[2], Simon Andrews[1]

[1]Bioinformatics Group, The Babraham Institute, Cambridge, CB22 3AT, UK
[2]Nuclear Dynamics Programme, The Babraham Institute, Cambridge, CB22 3AT, UK
[3]Epigenetics Programme, The Babraham Institute, Cambridge, CB22 3AT, UK

/ĄùĂ 0ùùĈ 2ùĊýùċ

2ùúùĈùù 3ĉõĈĉćǒ ☑☑

˷ &ùĈüõĉ !č, Northwestern University USA

˰ *ĉõĂ -ǒ 6õąĉùĈýĎõć, Max Planck Institute for Molecular Biomedicine Germany, +ĈĉćùĈ +õý, Max Planck Institute for Molecular Biomedicine Germany

˸ .ý÷ăǻõ .ùćùĉý, Brown University USA

$ýć÷ĉćć ĉüýć õĈĉý÷Ǻù

Comments (0)

!öćĉĈõ÷ĉ

HiCUP is a pipeline for processing sequence data generated by Hi-C and Capture Hi-C (CHi-C) experiments, which are techniques used to investigate three-dimensional genomic organisation. The pipeline maps data to a specified reference genome and removes artefacts that would otherwise hinder subsequent analysis. HiCUP also produces an easy-to-interpret yet detailed quality control (QC) report that assists in refining experimental protocols for future studies. The software is freely available and has already been used for processing Hi-C and CHi-C data in several recently published peer-reviewed studies.

#ăĈĈùćĄăĂøýĂû õĉĉüăĈǒ Steven Wingett (steven.wingett@babraham.ac.uk)

(ăċ ĉă ÷ýĉù ĉüýć õĈĉý÷Ǻùǒ Wingett S, Ewels P, Furlan-Magaril M *et al.* (ý#50ǒ ĄýĄùĂýĂù úăĈ āõĄĄýĂû õĂø ĄĈă÷ùććýĂû (ýù# øõĉõ ǽĊùĈćýăĂ ˷Ĭ Ĉùúùĉùùćǒ ˰ õĄĄĈăĊùøĬ ˷ õĄĄĈăĊùø ċýĉü ĉùćùĈĊõĉýăĂćǾ *F1000Research* 2015, ˻:1310 (doi: 10.12688/f1000research.7334.1)

'ĉõĂĉ ýĂúăĈāõĉýăĂǒ Research supported by the Medical Research Council and the Biotechnology and Biological Sciences Research Council of the UK.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

#ăāĄùĉýĂû ýĂĉùĈùćĉćǒ No competing interests were disclosed.

&ýĈćĉ ĄĉöǺýćüùøǒ 20 Nov 2015, ˻:1310 (doi: 10.12688/f1000research.7334.1)

Introduction

Hi-C is a ligation-based proximity assay utilising the power of massively parallel sequencing to identify three-dimensional genomic interactions[1]. The method (summarised in Figure 1a) involves fixing chromatin to preserve genomic organisation, followed by restriction enzyme digestion of the DNA. Overhanging single-stranded DNA at the ends of restriction fragments are then filled in with the concomitant incorporation of biotin. Fragments in close spatial proximity are ligated together generating a novel "modified restriction site" sequence (see Figure 1b). Following sonication the sheared ligated DNA fragments are enriched by streptavidin pull-down of the biotin residues, and then are ligated between sequencing adapters. The resulting molecule, termed a di-tag, should comprise two different DNA fragments separated by a modified restriction site. Since these two fragments were positioned close to each other during fixation, by analysing the composition of a population of di-tags generated by a Hi-C experiment it is possible to infer genomic three-dimensional organisation.

A recent variation of the protocol involves enriching Hi-C libraries for di-tags in which one or both reads align to pre-selected regions of a genome[2–4]

HiCUP uses Bowtie[11] or Bowtie 2[12] to map Hi-C di-tags, allow proportion of these internal fragments arose from the genomic background. Alternatively, the presence of these internal fragments may be explained by the aberrant incorporation of biotin, possibly a result of DNA restriction digestion at non-canonical sites.

Validating insert size. HiCUP places aligned di-tags on an *in silico* digested genome to calculate the theoretical length of the Hi-C insert and removes those not falling within the range set by the size-selection step of the protocol (Figure 2g). Explanations for such discrepancies include a read being incorrectly mapped or a putative di-tag containing multiple internal fragments or dangling ends. It is also possible that sequence variation between the sample DNA and reference genome leads to the loss or creation of restriction sites in the sample material. While such events are not common, the hallmark of restriction enzyme site generation is sometimes observed in Hi-C datasets, manifesting as an aggregation of reads orientated towards the novel restriction site.

Hi-C [illegible] visualisation.

Archive accession SAMEA2421733) were CHi-C libraries of foetal liver cells from mouse (strain C57BL/6) embryos at day 14.5 of development.

Assuming equal numbers of each barcode, all 16 barcode-barcode permutations should occur with equal frequency. Crucially, the barcodes are incorporated into the di-tags before PCR amplification, providing a test for the origin of the duplicates: those representing independent Hi-C events are most likely to possess differing barcodes (Figure 4a), whereas di-tags with a single common origin amplified by PCR should have identical barcodes (Figure 4b).

The two independent libraries were sequenced and the resulting FASTQ reads were classified by barcode i.e. the first four base pairs of the polynucleotide. The reads were mapped and filtered with HiCUP, but duplicates were retained. The barcode sequences were then quantified revealing that 71.3% corresponded to a pre-defined sequence in Sample 1 and 71.5% in Sample 2. Each valid barcode was then quantified. For both samples the barcode CCTT was most prevalent (Sample 1: 9,889,602; Sample 2: 16,368,793), and for both samples the barcode CGCT was the least common (Sample 1: 2,991,820; Sample 2: 5,002,346).

Despite the deviation from the ideal and expected result in which all barcode combination frequencies were equal, it was still possible to address the source of duplicate di-tags. To achieve this, the barcode combination of each duplicate was recorded. Duplicates arising from PCR amplification should have had identical barcode combinations, in contrast to *bone fide* initial Hi-C interaction events. Table 1 shows the result of this quantification, demonstrating that the overwhelming majority of duplicate di-tags are delimited by identical barcode combinations, almost certainly a result of PCR amplification.

To assess the impact of retaining duplicate di-tags we detected significant interactions in HiCUP-processed CHi-C datasets using the CHiCAGO pipeline (five datasets were processed: Samples 1 and 2 are described previously and Samples 3, 4 and 5 were generated from mouse embryonic stem cells and have the Gene Expression Omnibus accession GSM1888519). The analysis showed that removing duplicate di-tags substantially reduced the number of called significant interactions (see Table 2). This was not surprising owing to the vast number of theoretical interactions, meaning that fragment-fragment contacts repeated only a small number of times were likely to be statistically significant.

Consequently, when considering that PCR results in certain di-tags being amplified disproportionately[14], and that only an AMc[C9numbern AM)-14(c[C

over-calling of significant interactions becomes more problematic as the proportion of duplicate di-tags increases.

[redacted]

[redacted]

No competing interests were disclosed.

Research supported by the Medical Research Council and the Biotechnology and Biological Sciences Research Council of the UK.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

. ă ÷ă ā ĄùĈýĂû ýĂĈùĆùćĈć ċùĆù øýć÷İăćùøŏ