BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2144–2146
doi:10.1093/bioinformatics/btr354

Genome analysis

Advance Access publication June 19, 2011

False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions

Joseph K. Pickrell[1,*,†], Daniel J. Gaffney[1,2,*,†], Yoav Gilad[1,*] and Jonathan K. Pritchard[1,2,*]

[1]Department of Human Genetics and [2]Howard Hughes Medical Institute, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Sequencing-based assays such as ChIP-seq, DNase-seq and MNase-seq have become important tools for genome annotation. In these assays, short sequence reads enriched for loci of interest are mapped to a reference genome to determine their origin. Here, we consider whether false positive peak calls can be caused by particular type of error in the reference genome: multicopy sequences which have been incorrectly assembled and collapsed into a single copy.

Results: Using sequencing data from the 1000 Genomes Project, we systematically scanned the human genome for regions of high sequencing depth. These regions are highly enriched for erroneously inferred transcription factor binding sites, positions of nucleosomes and regions of open chromatin. We suggest a simple masking procedure to remove these regions and reduce false positive calls.

Availability: Files for masking out these regions are available at eqtl.uchicago.edu

Contact: pickrell@uchicago.edu; dgaffney@uchicago.edu; gilad@uchicago.edu; pritch@uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2011; revised on June 6, 2011; accepted on June 8, 2011

The combination of classical methods from molecular biology with high-throughput sequencing, as used in ChIP-seq (Johnson *et al.*, 2007), DNaseI-seq (Boyle *et al.*, 2008) and MNase-seq (Schones *et al.*, 2008), has dramatically increased the scale at which genomic sequences can be assayed for various properties of function. In each of these experiments, sequences of interest (for example, sites bound by a particular transcription factor) are enriched and then sequenced. These resulting sequences are then mapped back to a reference genome to determine the position of their origin. It is well appreciated that characteristics of the reference genome influence this mapping step; for example, some sequences in the genome are present in multiple copies, leading to ambiguity when determining the origin of a sequencing read (Koehler *et al.*, 2011).

A related issue that has received less attention in this context is the existence of sequences that appear to be present in a single copy in the available reference genome, but which, in reality, are present in multiple copies in all or some individuals. Such sequences could potentially cause artifactual peaks in sequencing-based assays (Hesselberth *et al.*, 2009; Zhang *et al.*, 2008), and can be identified as regions of high sequencing depth in genomic DNA (Bailey *et al.*, 2002; Vega *et al.*, 2009). To screen for potentially problematic regions, we used data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010). Specifically, we downloaded the Illumina sequencing reads derived from low-coverage sequencing of 57 Nigerian individuals, mapped the reads to the human genome and then calculated the coverage at each base in the genome using only uniquely mapped reads. For full details on the data used, see the Supplementary Material.

An example of a problematic genomic region is presented in Figure 1A. In the genome sequencing data, there are multiple clear peaks of reads, suggesting the presence of collapsed repeats in the reference genome [the reference genome used throughout is hg18; ∼10% of these regions are no longer problematic in hg19 (Supplementary Fig. S1)]. We find that 0.1% of the genome has read depth at least twice the median, and 0.01% of the genome has read depth at least 15 times the median (Fig. 1B). We identified contiguous regions where the read depth exceeded thresholds corresponding to the top 0.1, 0.01, and 0.001% of the per-base read depths, merging regions which fall within 50 bases of each other. At the 0.1% threshold, there are 34 359 such high depth regions (HDRs), with a mean size of 188 bases.

We then asked whether HDRs were indeed causing false positive peaks of coverage in high-throughput molecular biology assays. To do this, we used data from DNase-seq (Pique-Regi *et al.*, 2011), MNase-seq (Schones *et al.*, 2008) and ChIP-seq from various transcription factors (ENCODE Project Consortium, 2007) and histone marks (Wang *et al.*, 2008) (Supplementary Material). In the example in Figure 1A, the regions found to be copy number variable also show up as sensitive to DNaseI and show extremely high read depth in the MNase assay. Overall, of the top 0.1% of 200 base pair windows in the genome with the greatest read depth in the DNase-seq experiment, 3% overlap HDRs, and of the top 0.1% of 200 base pair windows in the MNase-seq experiment, 26% overlap HDRs. Additionally, across many ChIP-seq experiments on both transcription factors and histone modifications, we see

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

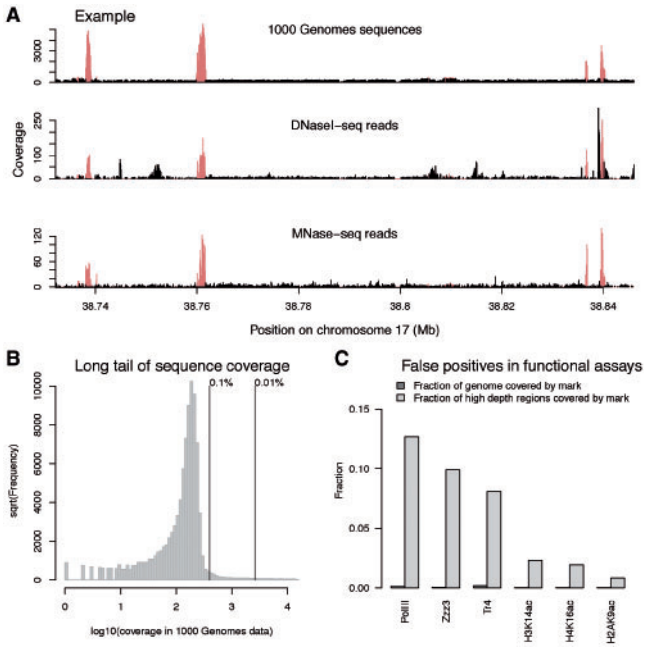

Fig. 1. Sequences absent from the reference genome cause spurious peaks of sequencing reads. (**A**) An example of such a region. In each panel, we plot the density of uniquely mapped sequencing reads from three sources: the Illumina data from low coverage sequencing of Yoruba individuals from the 1000 Genomes Project (summed across all individuals), a study of DNaseI hypersensitivity (Pique-Regi *et al.*, 2011) and a study of MNase sensitivity (Schones *et al.*, 2008). In the first of these, copy number is expected to be approximately constant. In red are regions that we call as high depth regions at a threshold of 0.1%. (**B**) A long tail of very high read depth for sequences present once in the human reference. Using the coverage from the 1000 Genomes Project data, we plot the histogram of the coverage at each base (using 500 Mb of sequence). Marked are the positions corresponding to the top 0.1 and 0.01% of the distribution. (**C**) Collapsed repeats cause false peaks of sequencing reads in functional assays. For each experiment, we plot the fraction of the genome covered occupied by the mark, as well as the fraction of the HDRs covered by the mark. For the ChIP-seq on transcription factors, we used the binding sites called by the ENCODE Project (ENCODE Project Consortium, 2007) using PeakSeq (Rozowsky *et al.*, 2009). For the ChIP-seq on histone modifications (Wang *et al.*, 2008), we split the genome into windows of 200 bases and called the most extreme 0.1% of windows as bound. Shown are selected experiments; for all experiments see Supplementary Figures S2 and S3.

enrichment of signal in HDRs (Fig. 1C, Supplementary Figs S2 and S3). The magnitude of enrichment of ChIP-seq peaks in HDRs depends on the choice of peak-calling algorithm; peaks called using PeakSeq (Rozowsky *et al.*, 2009) are dramatically enriched in HDRs (Fig. 1C), while peaks called using MACS (Zhang *et al.*, 2008) are not (Supplementary Fig. S4). This is likely attributable to the different choices in how to use the control lanes in the two algorithms (Supplementary Material).

We next sought to confirm that HDRs are due to collapsed repeats in the genome (as opposed to, for example, biases due to GC content or other sequence properties during library construction or Illumina sequencing). First, we examined the impact of GC content of a region on sequencing coverage. As expected, there is a relationship between GC content and coverage, but this effect is too small to account for the dramatic peaks of coverage we see in the data (Supplementary Fig. S5). Next, we examined copy number data generated on separate individuals using the orthogonal technology of array CGH (Conrad *et al.*, 2010). The intensities of array probes falling in HDRs is dramatically higher than that of control probes, often approaching the limit of the dynamic range of the array (Supplementary Fig. S6). Finally, we examined the overlap of HDRs with annotated repeats. Of the most extreme outliers in the coverage distribution (at the 0.01% point in the distribution of coverage), 92% of the regions overlap annotated repeats. Of these repeats, 81% are satellite DNA and the remainder largely consist of L1 retrotransposons and Alu elements. We conclude that the majority of HDRs are indeed collapsed repeats, with the caveat that some fraction may be copy number variable across individuals (Vega *et al.*, 2009).

We suggest a simple masking procedure to remove false positive calls due to collapsed repeats. This can be done in two ways: first, we have generated BED files with the coordinates of regions we suggest masking out (available at eqtl.uchicago.edu). Files are available at five different cutoffs. Alternatively, we have made available a FASTA file with the sequences present in these regions. If this FASTA file is included in the reference genome when mapping, sequencing reads from these regions will no longer map uniquely to the genome and can be filtered out.

In summary, we have identified a set of genomic regions in humans which are likely to generate spurious peaks in any assay involving high-throughput sequencing, and have provided a resource for screening out these regions. Similar approaches will be feasible in other organisms as resequencing data from multiple individuals becomes available. Screening out these regions will be particularly useful in studies, like DNase-seq and MNase-seq, where there is no natural control experiment [apart from copy number quantification via whole genome sequencing (Kharchenko *et al.*, 2011), which remains impractical for species with large genomes], and will aid interpretation in ChIP-seq experiments where 'control' lanes contain biologically relevant signal (Rozowsky *et al.*, 2009; Vega *et al.*, 2009).

ACKNOWLEDGEMENTS

We thank the 1000 Genomes Project and the ENCODE Project for making their data public and easily accessible.

Funding: Howard Hughes Medical Institute (to Jon.K.P.), National Institutes of Health (MH084703-01 to Jon.K.P., GM084996 to Y.G.).

Conflict of Interest: none declared.

REFERENCES

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Bailey,J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, **447**, 799–816.

Hesselberth,J.R. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Johnson,D.S. *et al*. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kharchenko,P.V. *et al*. (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**, 480–485.

Koehler,R. *et al*. (2011) The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics*, **27**, 272–274.

Pique-Regi,R. *et al*. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.

Rozowsky,J. *et al*. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Schones,D.E. *et al*. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Vega,V.B. *et al*. (2009) Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS One*, **4**, e5241.

Wang,Z. *et al*. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.