The combination of classical methods from molecular biology with the origin of a sequencing read (Koehler present in multiple copies, leading to ambiguity when determining this mapping step; for example, some sequences in the genome are appreciated that characteristics of the reference genome influence bound by a particular transcription factor) are enriched and then sequences can be assayed for various properties of function. In et al. (2007), DNaseI-seq (Boyle 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.; 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...
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


