BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 24 2011, pages 3427–3429
doi:10.1093/bioinformatics/btr578

Genome analysis

Advance Access publication October 17, 2011

Mapping personal functional data to personal genomes

Marcelo Rivas-Astroza[1], Dan Xie[1], Xiaoyi Cao[2] and Sheng Zhong[1,2,*]

[1]Department of Bioengineering and [2]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The sequencing of personal genomes enabled analysis of variation in transcription factor (TF) binding, chromatin structure and gene expression and indicated how they contribute to phenotypic variation. It is hypothesized that using the reference genome for mapping ChIP-seq or RNA-seq reads may introduce errors, especially at polymorphic genomic regions.

Results: We developed a Personal Genome Editor (perEditor) that changes the reference human genome (NCBI36/hg18) into an individual genome, taking into account single nucleotide polymorphisms (SNPs), insertions and deletions, copy number variation, and chromosomal rearrangements. perEditor outputs two alleles (maternal, paternal) of the individual genome that is ready for mapping ChIP-seq and RNA-seq reads, and enabling the analyses of allele specific binding, chromatin structure and gene expression.

Availability: perEditor is available at http://biocomp.bioen.uiuc.edu/perEditor.

Contact: szhong@illinois.edu

Received on July 21, 2011; revised on September 25, 2011; accepted on October 12, 2011

Personal genomics does not stop at obtaining genetic variation. One of the next steps is to analyze the functional consequences of the genetic variation. To enable researchers at large to analyze the functions of individual genomes, large-scale personal genome projects including the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), the Personal Genome Project (Lunshof *et al.*, 2010) and the cancer genome projects (Pleasance *et al.*, 2009, 2010) all provided cells lines from the sequenced individuals. Based on these personal cell lines and their genome sequences, people have started to analyze the variation in transcription factor (TF) binding (Kasowski *et al.*, 2010), chromatin structure (McDaniell *et al.*, 2010) and gene expression (Li *et al.*, 2011) and started to study the association between these molecular-level functional variations and phenotypic variations. In addition, these cell lines and genomic sequences also made it possible to analyze allele-specific epigenetic modifications and gene expression (Turan *et al.*, 2010). These resources and growing research areas require dedicated analysis tools.

One way to analyze individual differences is to map personal data, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) data, onto the human reference genome and then compare the TF binding intensities across individuals. It is hypothesized that using the reference genome for mapping ChIP-seq or RNA-seq reads may introduce errors (McDaniell *et al.*, 2010), especially at polymorphic genomic regions. After all, the polymorphic regions are most likely to exhibit functional variation.

We developed a Personal Genome Editor (perEditor: http://biocomp.bioen.uiuc.edu/perEditor) that changes the reference human genome (autosomal, sex and mitochondrial chromosomes, build NCBI36/hg18) into an individual genome, taking into account single nucleotide polymorphisms (SNPs), insertions and deletions (indels), copy number variation and chromosomal rearrangement. perEditor takes the reference genome in Fasta format and the individual's differences from the reference genome in Variant Call Format (VCF) as inputs. For each difference described in the VCF file, perEditor makes a corresponding change to the reference genome. When the allele information is present in the VCF file, perEditor will keep two genome sequences, representing the two alleles, and make allele-specific changes. After all the data in the VCF file are processed, perEditor outputs the maternal and paternal alleles of the individual genome as Fasta files, ready for mapping ChIP-seq, RNA-seq and other sequence reads.

We quantified the difference in mapping ChIP-seq reads against the reference genome and the individual genome and compared this difference to reported inter-individual differences. Using perEditor and data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010; April 2009 data release, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/2009_04/), we constructed the two alleles of a European individual (GM10847, accession number NA10847) and an African individual (GM18505, accession number NA18505). We re-analyzed ChIP-seq reads of NFκB generated from these individuals (Kasowski *et al.*, 2010) by mapping them to each allele of their individual genome (output of perEditor) as well as to the reference genome (hg18, including autosomal, sex and mitochondrial chromosomes). A total of 55 232 610 raw ChIP-seq reads for the European individual and 64 291 100 raw reads for the African individual were mapped, using the Bowtie program (Langmead *et al.*, 2009) with default parameters and allowing for up to 1 mismatch. When the individual genome was used, more reads became alignable. Taking the maternal allele of GM10847 as an example, a total of 161 150 reads that could not be uniquely aligned to the reference genome could be uniquely aligned; 84.9% of these newly alignable reads overlap with maternal or homozygous SNPs of GM10847. The other 15.1% newly alignable reads were added because a SNP on GM10847 helped to resolve the uniqueness of a read alignment elsewhere. Importantly, 47 825 of the newly alignable reads are located in putative NFκB binding sites (defined as 200 bp windows with 10 or more alignable reads). In contrast, a much smaller number of reads aligned to the reference genome become not

*To whom correspondence should be addressed.

Table 1. Functionally relevant alignment differences

		GM10847		GM18505	
		Maternal	Paternal	Maternal	Paternal
New alignments	Total read count	47 825	45 675	15 093	15 589
	Proportion overlapped with SNPs	93.9%	94.0%	92.4%	92.1%
Lost alignments	Total read count	18 944	19 222	6104	5764
	Proportion overlapped with SNPs	40.9%	42.1%	72.3%	69.6%

ChIP-seq data from a European and an African individual were aligned to human reference genome (autosomal, sex and mitochondrial chromosomes, build hg18) and to each allele (maternal, paternal) of their own genomes (GM10847 and GM18505). Differences of aligned reads on putative binding sites are listed in this table. Putative binding sites are loosely defined as 200 bp long windows with 10 or more overlapping ChIP-seq reads. New alignment: a read aligned to an allele on the individual genome that cannot be aligned to hg18. Lost alignment: a read aligned to hg18 that cannot be aligned to the specific allele of the individual genome. For both individuals, there are 2 to 3 times more new alignments than lost alignments. These changes of aligned reads on putative binding sites could influence our understanding of individual variations.

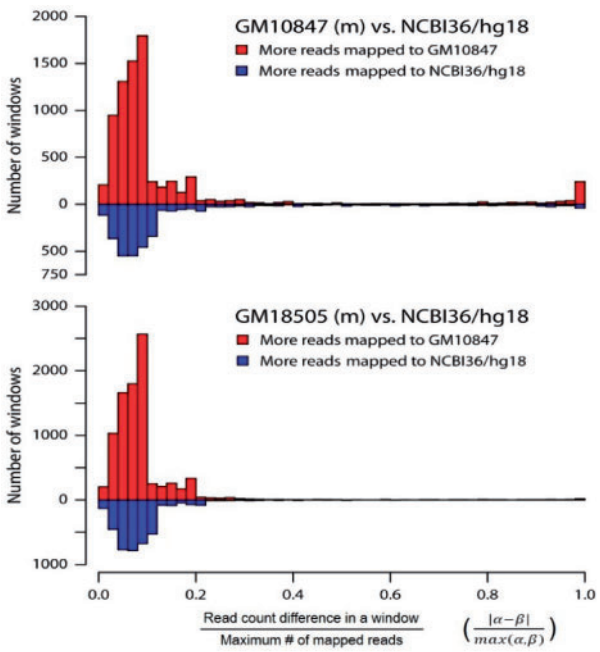

Fig. 1. Relative difference of ChIP-seq reads alignable to individual genome and reference genome. Such relative differences were designed to approximate the differences of estimated binding intensities. For each 200 bp window on the genome, the number of alignable reads to the individual genome (α) was compared to that aligned to the reference genome (β), by taking the relative difference $|\alpha-\beta|/\max(\alpha,\beta)$. The distribution of the windows with 10 or more reads ($\max(\alpha,\beta)\geq 10$) with respect to their relative differences is drawn as a histogram. m, maternal allele. Red bars: $\alpha > \beta$, blue bars: $\alpha < \beta$.

uniquely alignable to the individual genome (Lost alignments, Table 1). Compared with the new alignments, smaller fractions (40–72%) of the lost alignments overlapped with SNPs. These SNP-overlapping reads became not alignable primarily because they had 1 mismatch to the reference genome and had 2 or more mismatches (beyond the threshold) to the particular allele of the personal genome. The rest 28–60% of lost alignments were due to the result that they became not uniquely alignable to the particular allele of the individual's genome (a polymorphism elsewhere produced an identical sequence). These data indicate that mapping to the individual genome may increase the precision of quantifying the binding intensities using ChIP-seq reads, which is essential to explain individual variation (Figs 1 and 2).

Next, we asked how strongly the genomes used for mapping would affect our understanding of individual variation. The difference in reads alignable to an individual genome and to the reference genome was calculated for each 200 bp window covering the whole genome. We focused on the windows with 10 or more ChIP-seq reads for further analysis, because these regions are more likely to be binding sites (Table 2). Taking the maternal allele of GM10847 as an example, a total of 1356 windows (putative binding sites) showed a difference of 5 or more alignable reads. In terms of relative changes between individual and reference genomes (absolute difference of alignable reads divided by the maximum alienable reads), a total of 3794 windows showed 10% or larger changes, and 852 windows showed very strong (50% or larger) changes (Fig. 1). These data indicate the precision of inferred binding intensity can be increased by 10% or more on thousands of binding sites. This improvement is on the same scale as reported individual variation of NFκB binding (comparing ChIP-seq data of two individuals by using human reference genome hg18 for mapping) (Kasowski *et al.*, 2010).

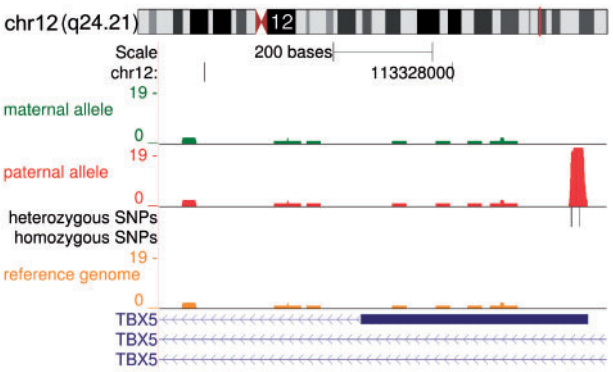

Fig. 2. An example of the genome differences and the allele differences. The NFκB ChIP-seq reads were mapped to the two alleles of the GM10847 genome (green and red tracks), as well as to the human reference genome (hg18, orange track). The height of colored bars represents the number of overlapping ChIP-seq reads on that genomic location. A strong peak was observed (19 overlapping ChIP-seq reads) in the second promoter of the TBX5 gene on the paternal allele of GM10847, overlapping with two of her heterozygous SNPs. This peak does not show up in her maternal allele or in the reference genome.

Table 2. Differences in inferred binding intensities α

	GM10847		GM18505	
Difference	Maternal	Paternal	Maternal	Paternal
$\lvert\alpha-\beta\rvert \geq 5$	1356	1376	324	284
$\lvert\alpha-\beta\rvert \geq 10$	869	875	89	84
$\frac{\lvert\alpha-\beta\rvert}{\max(\alpha,\beta)} \geq 10\%$	3794	3827	3500	3369
$\frac{\lvert\alpha-\beta\rvert}{\max(\alpha,\beta)} \geq 50\%$	852	865	85	79

The number of putative binding sites with defined differences are listed. α: number of reads aligned to this site in individual genome. β: number of reads aligned to this site in reference genome. Putative binding sites are loosely defined as 200 bp long windows with 10 or more overlapping ChIP-seq reads.

Funding: NIH DP2-OD007417; NSF DBI 08-45823; NSF DBI 09-60583.

Conflict of Interest: none declared.

REFERENCES

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,M. *et al.* (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, **333**, 53–58.

Lunshof,J.E. *et al.* (2010) Personal genomes in progress: from the human genome project to the personal genome project. *Dialog. Clin. Neurosci.*, **12**, 47–60.

McDaniell,R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

Pleasance,E.D. *et al.* (2009) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Pleasance,E.D. *et al.* (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184–190.

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

Turan,N. *et al.* (2010) Inter- and intra-individual variation in allele-specific dna methylation and gene expression in children conceived using assisted reproductive technology. *Plos Genet.*, **6**, e1001033.