BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1172–1174
doi:10.1093/bioinformatics/btt744

Genome analysis

Advance Access publication December 26, 2013

ALEA: a toolbox for allele-specific epigenomics analysis

Hamid Younesy[1,2], Torsten Möller[2,3], Alireza Heravi-Moussavi[1], Jeffrey B. Cheng[4], Joseph F. Costello[5], Matthew C. Lorincz[6], Mohammad M. Karimi[1,6,*] and Steven J. M. Jones[1,*]

[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, V5Z 4S6, Canada, [2]Graphics Usability and Visualization Lab, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada, [3]Visualization and Data Analysis Lab, Faculty of Computer Science, University of Vienna, A-1090 Vienna, Austria, [4]Department of Dermatology, University of California San Francisco, San Francisco, California 94143, USA, [5]Brain Tumor Research Center, Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California 94158, USA and [6]Department of Medical Genetics, Life Sciences Institute, The University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada

Associate Editor: Michael Brudno

ABSTRACT

The assessment of expression and epigenomic status using sequencing based methods provides an unprecedented opportunity to identify and correlate allelic differences with epigenomic status. We present ALEA, a computational toolbox for allele-specific epigenomics analysis, which incorporates allelic variation data within existing resources, allowing for the identification of significant associations between epigenetic modifications and specific allelic variants in human and mouse cells. ALEA provides a customizable pipeline of command line tools for allele-specific analysis of next-generation sequencing data (ChIP-seq, RNA-seq, etc.) that takes the raw sequencing data and produces separate allelic tracks ready to be viewed on genome browsers. The pipeline has been validated using human and hybrid mouse ChIP-seq and RNA-seq data.

Availability: The package, test data and usage instructions are available online at http://www.bcgsc.ca/platform/bioinfo/software/alea.

Contact: mkarimi1@interchange.ubc.ca or sjones@bcgsc.ca

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

Received on August 21, 2013; revised on November 25, 2013; accepted on December 9, 2013

1 INTRODUCTION

Each diploid cell consists of two near-identical copies of the genome. Despite the similarity of these paternally and maternally contributed genomes, they differ in certain genomic loci due to heterozygous variants such as single nucleotide polymorphisms (SNPs), indels and structural variations, which can be used to find allele-specific (AS) epigenetic variations. AS events may occur everywhere, including regions devoid of heterozygous variants. However, computational tools can only focus on heterozygous loci where the allelic differences are distinguishable. AlleleSeq (Rozowsky *et al*., 2011) is a computational pipeline for phasing heterozygous variants from trio (father, mother and child) data, constructing diploid personal genome and aligning short sequencing reads to maternal and paternal genomes. iASeq (Wei *et al*., 2012) uses a machine-learning approach to find the allelic imbalance by jointly analyzing multiple ChIP-seq datasets and by improving the AS inference using correlations among them. Allim (Pandey *et al*., 2013) is a more recent AS tool for RNA-seq analysis of F1 individuals that generates a polymorphism-aware diploid reference genome from trio genomic and/or transcriptomic sequences, and corrects the residual mapping bias when measuring the allelic expression imbalance. What limits application of these tools for AS epigenomics studies is that the majority of genomic datasets are generated from single patient samples, which do not include genomic data from the parents of the patients. Also, it is not always practical or desirable to sequence multiple epigenomic marks for each patient as required by iASeq.

To address these limitations, we developed a new pipeline, ALEA, for AS epigenomics analysis. ALEA takes advantage of the available genomic resources for human (The 1000 Genomes Project Consortium) and mouse (The Mouse Genomes Project) to reconstruct diploid *in silico* genomes for human or hybrid mice under study. Then, for each accompanying ChIP-seq or RNA-seq dataset, ALEA generates two wiggle track format files from short reads aligned differentially to each haplotype.

2 METHODS

ALEA consists of the following four modules for which the processes, the input and output files and their formats are shown in Figure 1.

A. Genotype phasing: For human samples, genotypes are typically called from whole genome-sequencing short reads accompanying the epigenomic dataset under AS study. SHAPEIT2 (Delaneau *et al*., 2013) is then used to phase genotypes using the publicly available reference panel of haplotypes, provided by the 1000 Genomes project (McVean *et al*., 2012). We then use the phased haplotypes together with the original unphased variant files in the Variant Call Format (VCF) to create a phased variant file with two haplotypes containing homozygous and phased heterozygous SNPs and indels (Fig. 1a). For mouse datasets, ALEA accepts epigenomic marks from F1 hybrid offspring whose parents are among the 17 inbred strains available from the Mouse Genomes Project (Keane *et al*., 2011).

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

B. Diploid personal genome reconstruction: The output of the phasing module, the phased variant file, together with the reference genome is fed into the second module (Fig. 1b) in which haplotype regions are reconstructed from the individual haplotypes. Two *in silico* genomes are created to be used for the optimal identification of chromosomal or AS effects. For mouse, the two *in silico* genomes represent the parental haplotypes, whereas for human the two genomes will essentially be a mosaic of the parental haplotypes due to the unlikeliness of true long-range phasing. The homozygous SNPs and indels do not need phasing, and they are simply assigned to both parental genomes. Although the homozygous variants are not informative for AS analyses, they may be included to investigate specific epigenomic signatures between individuals.

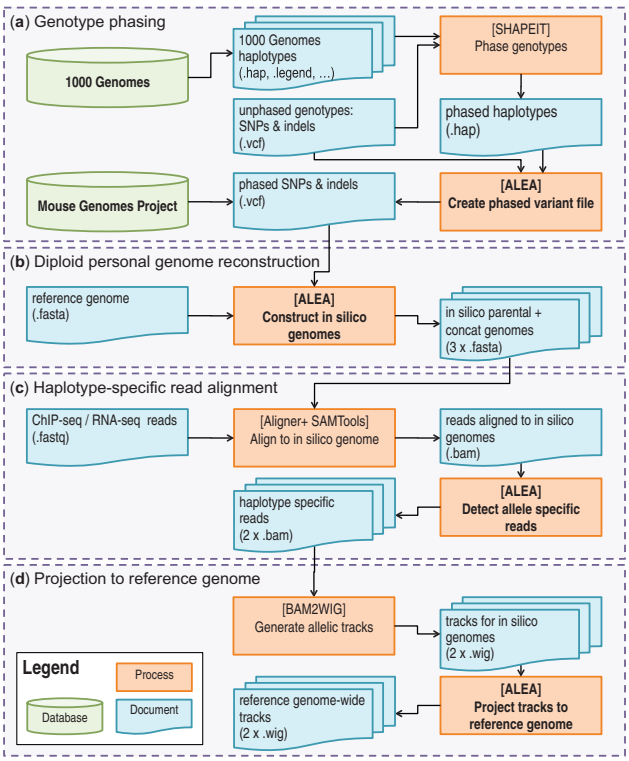

Fig. 1. AS epigenomics analysis pipeline: (**a**) genotype phasing, (**b**) constructing diploid genome, (**c**) finding haplotype-specific short reads and (**d**) projecting the haplotypic read profiles to reference genome

C. Haplotype-specific read alignment: This module aligns short sequencing reads to the *in silico* genomes constructed by the previous module and detects reads that are uniquely aligned only to one of the two genomes (Fig. 1c). Thus, all reads aligned to multiple locations of either haplotypes or aligned uniquely to both haplotypes are filtered out and only AS reads mapping to the regions containing heterozygous SNPs are detected. We provide two methods to detect the AS reads: (i) aligning the reads to a concatenated parental genome and (ii) aligning the reads to each parental genome separately. Details and comparison of the two methods are included in the Section 1 of the Supplementary Material.

D. Projection to reference genome: The fourth module (Fig. 1d) generates two haplotype-specific track files from mapped reads and projects them back to the reference genome. Owing to the existence of indels in the construction of parental *in silico* genomes, the coordinates of the tracks are skewed when compared with the reference genome. However, most visualization tools work based on alignment to reference genomes. Using an index created with the *in silico* genomes, ALEA maps the tracks back to the reference genome.

3 EVALUATION

We used known imprinted genes to validate our AS method for both human and mouse, as these regions are known to show allelic differences in both expression and specific histone marks. We used our AS pipeline on human skin fibroblast RNA-seq and H3K36me3 ChIP-seq datasets generated from the same individual. These data are available for download under the Gene Expression Omnibus accession numbers GSM751277 and GSM817238. Having advance access to unpublished whole genome-sequencing data for this individual, we called and phased the genotype, reconstructed *in silico* genomes and generated separate read profiles for each haplotype for RNA-seq and H3K36me3 data. Figure 2a shows the monoallelic H3K36me3 enrichment of expressed alleles in selected human-imprinted genes. A similar approach was also applied to a recently published dataset for mouse trophoblast cells (Calabrese *et al.*, 2012), including RNA-seq and H3K36me3 ChIP-seq data derived from crosses between CAST/EiJ (Cast) and C57BL/6J (B6) mice. As we expected, H3K36me3 enriches the gene body of active genes and correlates with expression level of selected imprinted genes in a monoallelic manner (Fig. 2b). The details of the quantitative analysis and validation of haplotype-specific wiggle track format files for human and mouse are described in the Section 2 of the Supplementary Material.

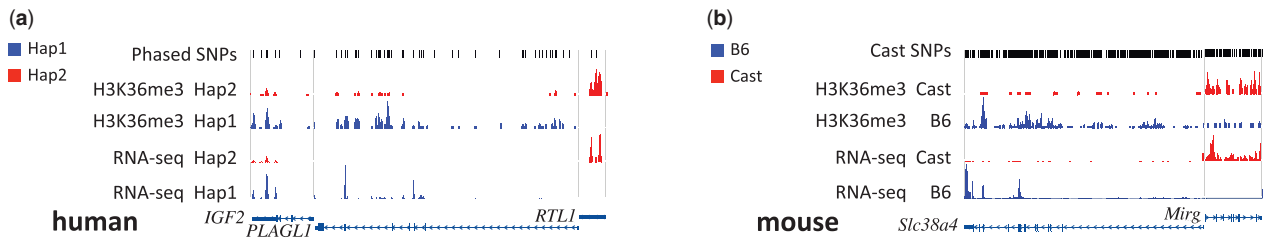

Fig. 2. Viewing haplotype-specific read tracks in the WashU epigenome browser using the same scale for both haplotypes in each gene reveals that AS H3K36me3 enrichment and RNA-seq expression in selected imprinted genes are correlated. (**a**) The tracks show opposite haplotype-specific RNA-seq and H3K36me3 read profiles for IGF2 and PLAGL1 versus RTL1 in human skin fibroblast, consistent with the fact that IGF2 and PLAGL1 have different imprinting status compared with RTL1. (**b**) Slc38a4 (Mirg)-imprinted gene is paternally (maternally) expressed in mouse trophoblast cells. The SNP-poor regions are devoid of reads as expected

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

ACKNOWLEDGEMENT

M.M.K. is a MSFHR post-doctoral fellow. S.J.M.J. is a senior scholar of MSFHR. M.C.L. is a CIHR New Investigator.

Funding: Funding for H.Y. was provided by NSERC PGS-D scholarship. J.B.C. is supported by a Career Development Award from the Dermatology Foundation. This work was supported by Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) funded by Canadian Institutes of Health Research (CIHR) and Genome BC (Grant No. EP2-120591). A full list of funders of infrastructure and research is available at www.bcgsc.ca/about/funding_support.

Conflict of Interest: none declared.

REFERENCES

Calabrese,J.M. *et al.* (2012) Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell*, **151**, 951–963.

Delaneau,O. *et al.* (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Meth.*, **10**, 5–6.

Keane,T.M. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.

McVean,G.A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Pandey,R.V. *et al.* (2013) Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol. Ecol. Resour.*, **13**, 740–745.

Rozowsky,J. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 1–15.

Wei,Y. *et al.* (2012) iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics*, **13**, 681.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016