MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data

Mamoru Kato¹, Yusuke Nakamura¹,² and Tatsuhiko Tsunoda¹,∗

¹SNP Research Center, RIKEN, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045 and ²Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Received on February 26, 2008; revised on April 26, 2008; accepted on May 19, 2008

ABSTRACT

Summary: Detailed analyses of the population-genetic nature of copy number variations (CNVs) and the linkage disequilibrium between CNV and single nucleotide polymorphism (SNP) loci from high-throughput experimental data require a computational tool to accurately infer alleles of CNVs and haplotypes composed of both CNV alleles and SNP alleles. Here we developed a new tool to infer population frequencies of such alleles and haplotypes from observed copy numbers and SNP genotypes, using the expectation-maximization algorithm. This tool can also handle copy numbers ambiguously determined, such as 2 or 3 copies, due to experimental noises.

Availability: http://emu.src.riken.jp/MOCSphaser/MOCSphaser.zip

Contact: tsunoda@src.riken.jp

Supplementary information: Additional materials can be found at http://emu.src.riken.jp/MOCSphaser/Supplinfo.doc

1 INTRODUCTION

Recent high-throughput experimental technologies have produced a vast amount of data on copy number variations (CNVs), which are variations of the number of DNA segments that are 1 k bases or larger, in human individuals, and their universality has been increasingly recognized (Redon et al., 2006). Since DNA segments of this size often include entire genes and their regulatory regions, CNVs are likely to have a major influence on phenotypic traits such as disease susceptibility (Feuk et al., 2006).

To perform population-genetic analyses such as analyses of allele frequencies and linkage disequilibrium (LD), alleles or haplotypes have to be determined. However, current high-throughput technologies cannot determine genotypes (pairs of alleles) of CNVs but instead measure only phenotypic copy numbers, which are the total numbers of allelic copies over two homologous chromosomes (Conrad and Hurles, 2007). Moreover, because of experimental noise, such technologies often produce ambiguous copy numbers, which are neither uniquely determined as one number because of experimental noise or limitations, nor uniquely determined as multiple numbers by concatenating these numbers. We denote such an ambiguous number by concatenating these numbers by ‘or’. For example, when copy numbers greater than 6 are impossible experimentally indistinguishable. We denote this number using ‘>6’. An ot-type ambiguous copy number indicates that several candidate numbers are suggested. We denote such an ambiguous number by concatenating these numbers by ‘or’. For example, when a copy number is either 2 or 3, we denote this equivocal state by ‘2or3’. A greater-type ambiguous copy number indicates that copy numbers over a certain value are experimentally indistinguishable. We denote this number using ‘>’.

In this study, we developed a new computational tool that infers population frequencies of allelic copy numbers as well as those of CNV–SNP haplotypes from a mixture of the data of both phenotypic copy numbers at CNV loci and genotypes at SNP loci. This tool can also handle the phenotypic copy numbers that are ambiguously determined. We tested this tool using simulated datasets and showed a good accuracy of the inference. We here introduce a tool called MOCSphaser (mixture-of-CNVS–SNP phaser), which is a command-line tool written in the Perl language.

2 ALGORITHM

Let us call an allelic copy number the number of allelic copies at a CNV locus on a chromosome. We denote an allelic copy number by its number. Let us call a phenotypic copy number the total number of allelic copies over two homologous chromosomes. Let us call an ambiguous (phenotypic) copy number a phenotypic copy number that is not uniquely determined as one number because of experimental noise or limitations. Ambiguous copy numbers are classified into ‘or-type’ and ‘greater-type’. An ot-type ambiguous copy number indicates that several candidate numbers are suggested. We denote such an ambiguous number by concatenating these numbers by ‘or’. For example, when a copy number is either 2 or 3, we denote this equivocal state by ‘2or3’. A greater-type ambiguous copy number indicates that copy numbers over a certain value are experimentally indistinguishable. We denote this number using ‘>’. For example, when copy numbers greater than 6 are impossible to discern, we denote this equivocal state by ‘>6’. We denote SNP alleles by the letters ‘a’ and ‘b’. We denote a haplotype with multiple loci by a series of alleles separated by ‘,’ per each locus, and denote

© 2008 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/14/1645/182014
by guest
on 27 July 2018

[15:24 4/7/03 Bioinformatics-btn242.tex]
We examined the performance of our algorithm using simulation. We packaged the MOCSphaser program together with example files, all true frequencies in the true sets were close to the estimated frequencies. From pre-defined, true haplotype frequencies under Hardy–Weinberg equilibrium. The simulated datasets were generated by random sampling from each CNV locus and SNP genotypes at multiple SNP loci for unrelated individuals (Fig. 1). From this dataset, we used the expectation–maximization (EM) algorithm to estimate haplotype frequencies under the assumption of Hardy–Weinberg equilibrium. For the first, we used the expectation–maximization (EM) algorithm, which was derived from the EM algorithm of SNP haplotype frequency estimation (Excoffier and Slatkin, 1995); for details of this procedure, see Supplementary Material. We examined the performance of our algorithm using simulation tests; see Supplementary Material for the results.

3 EXAMPLE

We packaged the MOCSphaser program together with example datasets, which consisted of four simulated datasets and eight real datasets. The simulated datasets were generated by random sampling from pre-defined, true haplotype frequencies under Hardy–Weinberg equilibrium. We also packaged files containing the true frequencies of allele or haplotype frequencies, and the estimated frequencies of alleles or haplotypes from Nigerian (YRI) in the HapMap populations (The International HapMap Consortium, 2007). As example real datasets, we provide experimental CNV data (Hosono et al., 2008), which were measured by quantitative PCR, on CYP2D6 and MRGPRX1 genes for individuals of European descent from Utah, USA (CEU) and for individuals of the Yoruba from Nigeria (YRI) in the HapMap populations (The International HapMap Consortium, 2007). We also provided real mixture data of these CNVs and neighboring SNPs that we arbitrarily selected as samples.

ACKNOWLEDGEMENTS

M.K. developed the algorithms and wrote the paper; T.T. checked the algorithms and reviewed the paper; Y.N. reviewed the paper. We thank T. Kawaguchi for implementing the phasing algorithm into MOCSphaser and T. Morizono for coding the simulation algorithm. We acknowledge N. Hosono and M. Kubo for information on the quantitative PCR data and S. Ishikawa and H. Abaratani for information on the Affymetrix GeneChip experiment data.

Funding: This work was partly supported by JSPS.KAKENHI (20790269).

REFERENCES

Clark,A.G. et al. (2004) Trisomic phase inference. In Istrail,S. et al. (eds) Computational Methods for SNPs and Haplotype Inference, Lecture Notes in Bioinformatics 2983, Springer-Verlag, Heidelberg, pp. 349–359.

Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. Nat. Genet., 39, S30–S36. Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol., 12, 921–927. Feuk,L. et al. (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum. Mol. Genet., 15, R57–R66.

Hosono,N. et al. (2008) Multiplex PCR-based real-time invader assay (mPCR-RETINA): a novel SNP-based method for detecting allelic asymmetries within copy number variation regions. Hum. Mutat., 29, 182–199.

Komura,D. et al. (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res., 16, 1757–1754.

Lee,T. (2006) Algorithms for inferring haplotypes. Genet. Epidemiol., 27, 334–347.

Redon,R. et al. (2006) Global variation in copy number in the human genome. Nature, 444, 444–444.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature, 449, 851–861.

Conflict of Interest: none declared.