BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 14 2008, pages 1645–1646
doi:10.1093/bioinformatics/btn242

*Genetics and population analysis*

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

Mamoru Kato[1], Yusuke Nakamura[1,2] and Tatsuhiko Tsunoda[1,*]

[1]SNP Research Center, RIKEN, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045 and [2]Human 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

Advance Access publication May 20, 2008

Associate Editor: Martin Bishop

## 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 noise.

**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/SuppInfor.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 phenotypic copy numbers that are not uniquely determined as one number but rather are given an ambiguous value, such as 2 or 3 copies (Komura *et al.*, 2006). In the case of single nucleotide polymorphisms (SNPs), genotypes are experimentally determined, and then, using these genotypes, haplotypes can be computationally inferred (Niu, 2004). Similarly, for trisomic chromosomes, as observed in Down syndrome, there is a method (Clark *et al.*, 2004) which infers three haplotypes from data on three alleles at each SNP site. Since the CNV data are different from such genotypic data, these methods cannot be applied to determine alleles of CNVs as well as haplotypes composed of CNV alleles and SNP alleles, which are necessary for calculating LD between CNV and SNP loci. A recent CNV study (Redon *et al.*, 2006) classified experimental measurements of signal intensities that correlate with copy numbers of individuals and regarded the three clusters as three genotypes; but it is unclear how to treat cases of more than three clusters and how many copies the alleles actually have. For precise analyses of CNV data, it will be necessary to develop techniques that accurately infer CNV alleles and CNV–SNP haplotypes.

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-CNV–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 or-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

*To whom correspondence should be addressed.

a diplotype by a pair of haplotypes separated by '/'. For example, [1, a / 10, b] represents a diplotype composed of haplotype [1, a] and [10, b], in which the first haplotype contains allelic copy number '1' at a CNV locus and SNP allele 'a' at the next SNP locus and the second haplotype contains alleles '10' and 'b' at the same loci, respectively.

Suppose that we have a dataset that lists both phenotypic copy numbers at multiple CNV loci 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. First, for a CNV locus, we list all possible pairs of allelic copy numbers whose total number is the same as the phenotypic copy number at the locus (Fig. 1). See Supplementary Material for the case of ambiguous copy numbers. For a SNP locus, we list a genotype as experimentally determined. Next, from the listed genotypes, we make up all possible diplotype configurations (Fig. 1). After enumerating all diplotype configurations, we iteratively update the frequencies of haplotypes contained in the diplotype configurations. This procedure is essentially the same as in 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.

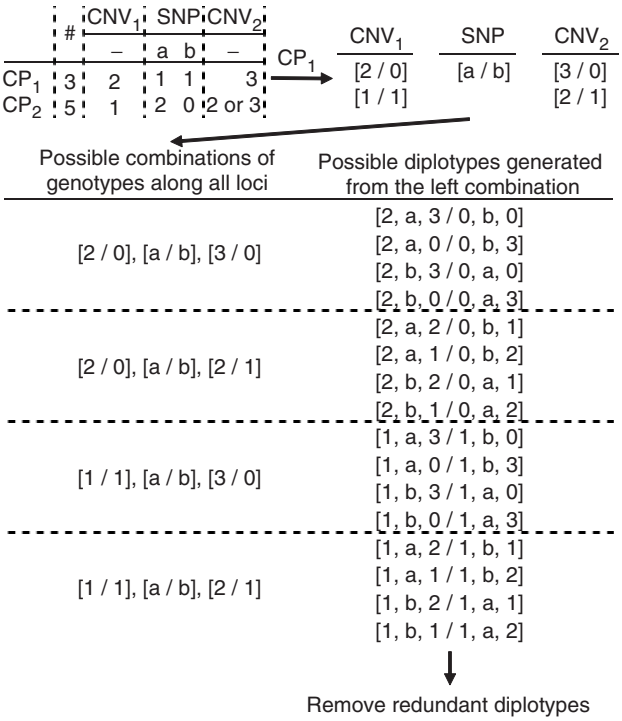


**Fig. 1.** An illustration of the enumeration procedure in our algorithm. The symbol 'CP' in the first table represents the count pattern, which is a unique series of counts along CNV and SNP loci. For example, the count pattern 1 is 2 1 1 3. The symbols '#', and 'a' and 'b' represent the number of individuals with the count pattern, and the SNP alleles, respectively. First, from the count pattern 1, the algorithm lists all possible genotypes consistent with the phenotypic copy number at each CNV locus and also lists the SNP genotype at each SNP locus. Second, the algorithm takes all possible combinations of the listed genotypes along all CNV and SNP loci. Third, it generates all possible haplotype pairs from each genotype combination. Finally, it removes redundant haplotype pairs (diplotypes). This procedure is performed for every count pattern.

## 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 and the frequencies estimated by MOCSphaser. As shown in these files, all true frequencies in the true sets were close to the estimated frequencies, and the estimated frequencies of alleles or haplotypes that were present in the estimated sets but absent in the true sets were all negligibly low.

As example real datasets, we provided 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. Aburatani for information on the Affymetrix GeneChip experiment data.

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

*Conflict of Interest:* none declared.

## 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. 1–8.

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**, 182–189.

Komura,D. *et al.* (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, **16**, 1575–1584.

Niu,T. (2004) 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**, 444–454.

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