APPLICATION NOTE

Genomics & Informatics
Vol. 9(3) 138-141, September 2011

http://dx.doi.org/10.5808/GI.2011.9.3.138
pISSN 1598-866X eISSN 2234-0742

A Program for Efficient Phasing of Three-Generation Trio SNP Genotype Data

Sanghoon Song and Sangsoo Kim*

Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

Abstract

Here, we report a computer program written in Python, which phases SNP genotypes and infers inherited deletions based on the pattern of Mendelian inheritance within a trio pedigree. When tiered trio genotypes that encompass three generations are available, it narrows a recombination event down to a region between two consecutive heterozygous markers. In addition, the phase information that is inferred from the upper trio that is formed by one of the parents and grandparents can be propagated to phase the genotypes of the lower trio that is formed by the parents and an offspring.

Availability: The software is freely available to nonprofit users upon request.

Keywords: SNP, phasing, trio, recombination, CNV, Mendelian inconsistency

Introduction

A single nucleotide polymorphism (SNP) refers to a variation at a single nucleotide sequence position of DNA that is observed commonly within a population (typically at least 1%). Due to its high density in human genomes, it is widely used in disease gene mapping. Humans are diploid organisms, having two copies of each autosomal chromosome. A haplotype is a combination of alleles on a chromosome that are transmitted together from the parent. Phasing is a process that resolves a series of consecutive genotypes into a haplotype of alleles that are transmitted or untransmitted. Because disease-causing mutations segregate with a particular haplotype, an association analysis that is based on haplotype may be more powerful than that based on genotype (Gabriel *et al.*, 2002). Hence, there is a paramount interest in phasing population genotypes. There are many statistical algorithms for phasing genotype data of unrelated populations (Stephens *et al.*, 2003; Browning, 2008). Due to their statistical nature, there are some errors and uncertainty in these methods. On the other hand, phasing that is based on Mendelian inheritance patterns within a trio family is straightforward and deterministic, as long as one of the members is homozygous at the locus of interest. This approach fails if all of the members of a trio have heterozygous genotypes. There are a number of statistical approaches that deal with this problem (Marchini *et al.*, 2006). However, these methods are known to be slow or produce some errors (Iliadis *et al.*, 2010). For a pedigree in which tiered trios encompass three generations, the upper trio is formed by one of the parents and his or her grandparents, while the lower trio is formed by the parents and their offspring. Of loci with a minor allele frequency of 20%, 5.1% is expected to be heterozygous for all of the members of a trio (Marchini *et al.*, 2006), while it drops to 0.8% for a three-generation pedigree. The deterministic method that is mentioned above can be extended to these multi-generation trio cases. In such cases, it is also possible to infer an accurate de novo recombination map by comparing the parental phases that have been inherited from grandparents with those that have been transmitted to an offspring. In this paper, we report an efficient software program that implements all of these processes in one pot. During trio phasing, some markers may deviate from Mendelian inheritance patterns. It is called Mendelian inconsistency and has served as a basis for the detection of copy number variation (CNV) (Conrad *et al.*, 2006; Freeman *et al.*, 2006). Our program also reports a list of putative deletion CNVs. By considering three generations of trios simultaneously, it can identify more individuals who might have a CNV.

Methods

An overview of the program processes is shown in Fig. 1.

Trio phasing of SNP genotypes based on Mendelian inheritance patterns

Phasing amounts to deduce which parental allele has been transmitted. If a parent has a homozygous genotype and thus the same allele, it is immaterial which one has been transmitted. As we know which allele of a child has been transmitted from the homozygous parent,

*Corresponding author: E-mail sskimb@ssu.ac.kr
Tel +82-2-820-0457, Fax +82-2-824-4383
Accepted 2 September 2011

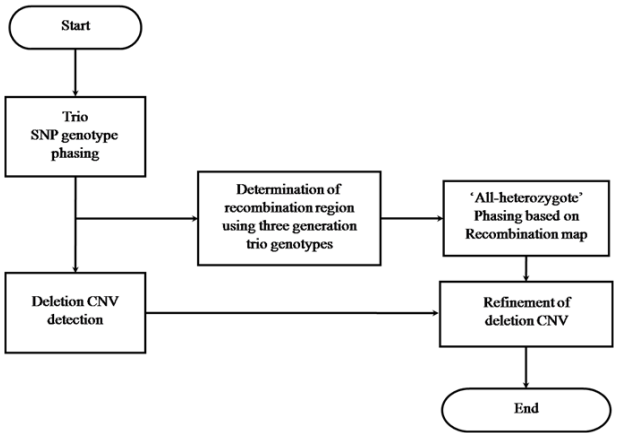

Fig. 1. A flowchart of the program.

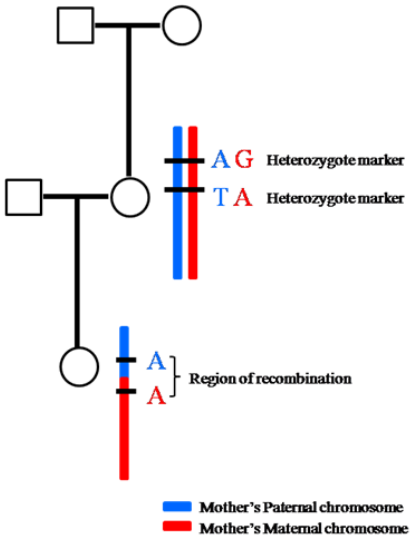

Fig. 2. Determination of the region where recombination must have occurred. Two heterozygous markers are considered. Using the upper trio, the genotypes of the mother can be resolved into those inherited from the grandfather and those from the grandmother, respectively, shown in blue and red. Using the lower trio, the chromosome that has been transmitted from the mother to the offspring can be inferred. A comparison of this chromosome with those two resolved maternal chromosomes can pinpoint the region of recombination, which is marked by two neighboring heterozygous markers.

we can deduce that the other allele of the child must have been transmitted from the opposite parent. Once we know which allele of a child is paternal and which is maternal, we can also deduce the untransmitted allele of a parent. Thus, the genotypes of both parents can be phased. If both parents have the same heterozygous genotype, it is possible to phase it as long as the child has a homozygous genotype in a similar way as above. If the genotype of the child as well as both parents is heterozygous, this simple rule can not phase the genotypes. We call this an 'all heterozygote' case, and our software flags it as unphased.

CNV deletion detection

Some SNPs may violate Mendelian inheritance patterns and can not be phased as above. For example, a parental homozygote allele may not be found in the child. If this is reproduced in multiple individuals and thus is due to neither experimental error nor de novo mutation, it can be due to a parental deletion mutation that has been transmitte d to the child. This concept has been instrumental in discovering the deletion polymorphisms of CNVs (Conrad *et al.*, 2006; Freeman *et al.*, 2006). By treating such a deletion as one of the alleles, our software can phase it as well.

Determination of recombination loci based on three- generation trio genotype data

The trio phasing that we have implemented above can not tell the phases of the parental chromosomes before meiosis but can for those after the recombination. If we can phase the genotype of a parent independently, we can compare the two phases before and after meiosis and deduce the approximate locus where *de novo* recombination has occurred. The parental phase before meiosis can be inferred if the genotypes of both grandparents are available through the trio phasing that we have implemented above (Fig. 2). For an SNP that is heterozygous for the parent, it can be determined whether the allele that passed on to the offspring is from the parent's maternal or paternal chromosome. The location of the recombination can hence be localized to the region that is spanned by the two closest flanking heterozygous markers in the parent (Kong *et al.*, 2010). We call this the 'recombination region'.

Phasing 'all heterozygote' cases using three-generation trio genotype

We have mentioned a so-called 'all-heterozygote' case above, in which all three members of a trio family have heterozygous genotypes. In such a case, phasing is still possible if the genotypes of both grandparents of either parent are available and at least one of them is homozygous. For example, let us assume that the genotypes of the mother's paternal and maternal genotypes are available (Fig. 3). In this case, the lower trio members all have heterozygous 'TC' genotypes ('all-heterozygote'), while the grandfather has a homozygous geno-

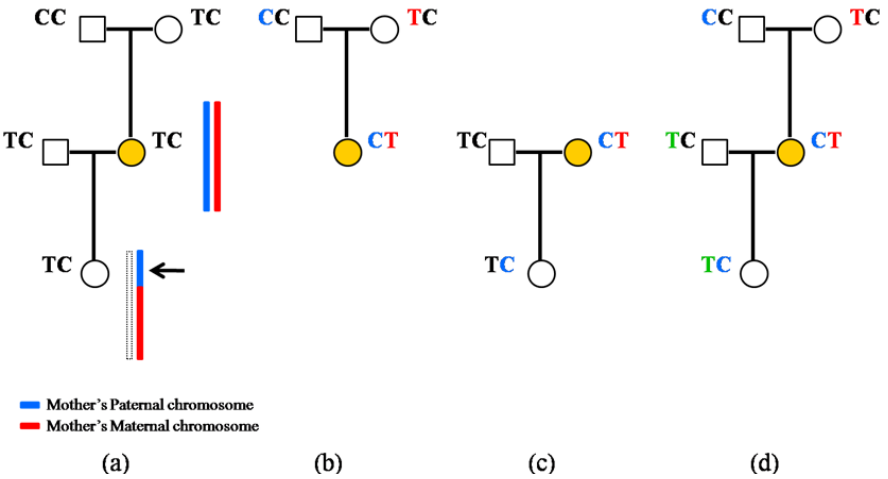

Fig. 3. Phasing 'all heterozygous' genotypes based on three-generation trio genotypes. (a) The pedigree and unphased genotypes of the family. The blue and red bars next to the mother represent the chromosomes inherited from her father and mother, respectively. The recombination pattern of these two chromosomes is depicted next to the child, whose paternal status is not known, and the chromosome is represented by a white bar. The approximate location of the marker is shown by an arrow. (b) The phased genotypes of the upper trio. The blue and red alleles represent those transmitted from the grandfather and the grandmother, respectively, to the mother, while the black alleles are untransmitted. (c) According to the recombination map in (a), the maternal allele of the child must have originated from the grandfather, which is 'C'. (d) The final phasing status of the pedigree. Referring to the child's phase, the father's genotype can be phased. The allele transmitted from the father is shown in green.

Table 1. Results of a test run of the program with a genotype dataset of 194 Korean trios

Item	Chr 1	Chr 2	Chr3	Chr4
'All-heterozygous' loci in at least one trio family	5.79%	5.80%	5.82%	5.88%
The loci whose phase was not resolved in at least one trio, even after three generation were used (29 families)	0.90%	0.93%	0.91%	0.98%
Average number of recombination events per chromosome	7.97	10.14	8.55	7.69
Average number of inferred deletion CNVs per individual	8.38	8.44	7.89	8.83

type, which enables phasing of the upper trio (Fig. 3b). At this stage of the process, we still can not tell which allele of the mother has been transmitted to the offspring. By phasing the neighboring SNPs in both the upper and lower trios, we can construct a recombination map as described above and tell whether the maternal part of this locus originated from the grandfather or grandmother (Fig. 3a). The locus must be outside of the so-called 'recombination region' mentioned above, within which the grandparental origin of the allele is uncertain. If we can be sure that the locus inherited the allele from the maternal grandfather, as shown in Fig. 3a, the genotype of the offspring can be phased (Fig. 3c), and subsequently, that of the father can be phased as well (Fig. 3d).

Results and Discussion

The algorithm has been implemented as a computer program, written in Python, executable in both the Linux and Windows operating systems. The software reads in files that are formatted as used by PLINK that is, PED, MAP formats and outputs three files, one each for 'recombination region', 'all-heterozygote' information, and the final phases. It is freely available upon request.

We applied this software to a genotype dataset of 194 trio families collected in Korea using Illumina 370K-Duo SNP chips (Park *et al.*, 2009). About 6% of the SNPs were so-called 'all-heterozygous', whose genotypes of all three trio members were heterozygous (Table 1). This is similar to the frequency that is expected for trios 5.1% of loci with a minor allele frequency of 20% (Marchini *et al.*, 2006). Among the fami-

lies in the dataset, three-generation trio genotype data are available for 29 families. After phasing, based on single-tier trio genotypes, followed by recombination mapping, based on two-tier trio genotypes, the 'all-heterozygous' genotypes were phased as described. Phasing by this method failed for about 1% of the SNPs, as none in the pedigree had a homozygous genotype (Table 1). As a byproduct of our algorithm, the *de novo* recombination events per chromosome were counted, ranging from 7 to 10 (Table 1). In addition, putative CNVs were inferred, based on Mendelian inconsistency: 2237 of them had a frequency of 1% or more. Park *et al*. also analyzed an expanded dataset that included basically the same trio families in this analysis, as well as 199 additional families of Korean-Vietnamese origin (Park *et al*., 2009). They had filtered the candidate CNVs by examining their cluster images, which we did not have access to, reporting putative 1029 CNVs, which was about half of the number of our result. Among 1029 CNVs that were reported by Park *et al*., 1014 were found in our results, achieving a recovery rate of 98.5%. The small portion of CNVs that were observed by Park *et al*. and not by us may be due to the difference in sample sizes. The genotype phasing information obtained here covers more than 99% of the SNP markers in the dataset. The phasing result from our software can be fed into other statistical phasing programs, such as PHASE, in order to resolve the phases of the rest of the markers. The highly accurate phased haplotypes that have been obtained here may serve as seeds for and thus facilitate the phasing of genotypes of unrelated individuals (Howie *et al*., 2009).

Acknowledgements

The authors are grateful to Dr. Jong Young Lee and colleagues in the National Institute of Health, Korea Center for Disease Control, for kindly providing the trio genotype dataset used in this work. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, and Technology (NRF-2010-0021811).

References

Browning, S.R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* 124, 439-450.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genetics.* 38, 75-81.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W., and Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949-961.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5, e1000529.

Iliadis, A., Watkinson, J., Anastassiou, D., and Wang, X. (2010). A haplotype inference algorithm for trios based on deterministic sampling. *BMC Genet.* 11, 78.

Kong, A., Thorleifsson, G., Gudbjartsson, F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, B., Jonasdottir, A., Gylfason, A., Kristinsson, K., Gudjonsson, S., Frigge, L., Helgason, A., Thorsteinsdottir, U., and Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467, 1099-1103.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., Donnelly, P., and International HapMap Consortium. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78, 437-450.

Park, M., Kim, D.J., Kim, K.J., Hong, C.B., Kim, Y.J., Cheong, H.S., Shin, H.D., Lee, E.J., Kim, H.N., Chung, H.W., Kim, E.K., Lee, J.Y., and Kim, H.L. (2009). Gene associations of common deletion polymorphisms in families with Avellino corneal dystrophy. *Biochem. Biophys. Res. Comm.* 387, 688-693.

Stephens, M., and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73, 1162-1169.