BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 11 2010, pages 1437–1445
doi:10.1093/bioinformatics/btq157

Genetics and population analysis Advance Access publication April 20, 2010

Inferring combined CNV/SNP haplotypes from genotype data

Shu-Yi Su[1,2], Julian E. Asher[3], Marjo-Riita Jarvelin[1,4], Phillipe Froguel[3,5], Alexandra I.F. Blakemore[3], David J. Balding[6] and Lachlan J.M. Coin[1,*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London W2 1PG, UK, [2]Ernest Gallo Clinic and Research Center, Department of Bioinformatics, University of California, San Francisco, CA 94608, USA, [3]Department of Genomics of Common Disease, School of Public Health, Imperial College, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK, [4]Institute of Health, University of Oulu, Oulu, Finland, [5]CNRS 8090-Institute of Biology, Pasteur Institute, France and [6]Institute of Genetics, University College London, London WC 1E 6BT, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Copy number variations (CNVs) are increasingly recognized as an substantial source of individual genetic variation, and hence there is a growing interest in investigating the evolutionary history of CNVs as well as their impact on complex disease susceptibility. CNV/SNP haplotypes are critical for this research, but although many methods have been proposed for inferring integer copy number, few have been designed for inferring CNV haplotypic phase and none of these are applicable at genome-wide scale. Here, we present a method for inferring missing CNV genotypes, predicting CNV allelic configuration and for inferring CNV haplotypic phase from SNP/CNV genotype data. Our method, implemented in the software polyHap v2.0, is based on a hidden Markov model, which models the joint haplotype structure between CNVs and SNPs. Thus, haplotypic phase of CNVs and SNPs are inferred simultaneously. A sampling algorithm is employed to obtain a measure of confidence/credibility of each estimate.

Results: We generated diploid phase-known CNV–SNP genotype datasets by pairing male X chromosome CNV–SNP haplotypes. We show that polyHap provides accurate estimates of missing CNV genotypes, allelic configuration and CNV haplotypic phase on these datasets. We applied our method to a non-simulated dataset—a region on Chromosome 2 encompassing a short deletion. The results confirm that polyHap's accuracy extends to real-life datasets.

Availability: Our method is implemented in version 2.0 of the polyHap software package and can be downloaded from http://www.imperial.ac.uk/medicine/people/l.coin

Contact: l.coin@imperial.ac.uk

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

Received on December 9, 2009; revised on March 15, 2010; accepted on April 7, 2010

1 INTRODUCTION

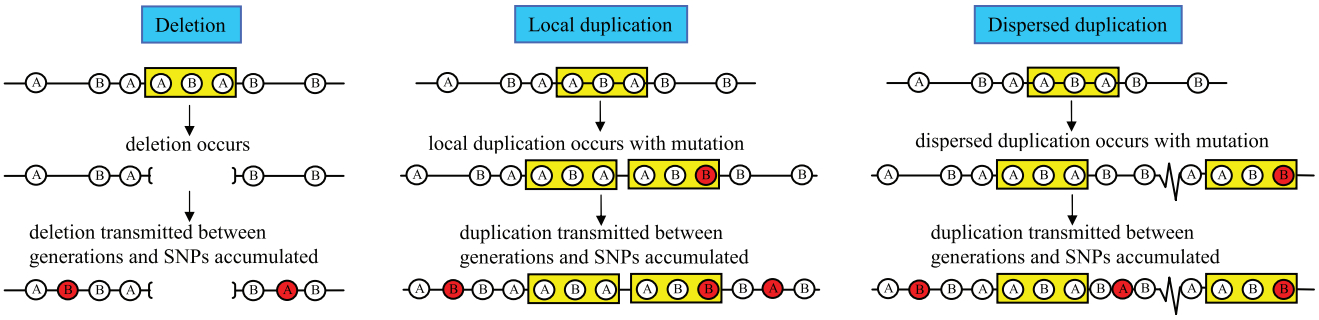
Copy number variations (CNVs) are pervasive in the human genome (Feuk *et al.*, 2006; Redon *et al.*, 2006) and could play a key role in human diversity and disease susceptibility (Conrad *et al.*, 2009; McCarroll and Altshuler, 2007). Despite this, the population genetics of CNVs—and particularly so for duplications—remain relatively poorly understood. Several analytical tools, such as haplotype analysis, which are standard for SNP-based population genetics have yet to be modified to be applicable to complex multi-allelic CNVs.

Several technologies enable high-throughput CNV detection, including array comparative genomic hybridization (aCGH) and SNP genotyping arrays. Many algorithms have been proposed to detect CNV regions and to estimate the integer copy-number (CN) genotypes in each region using these technologies (Colella *et al.*, 2007; Fiegler *et al.*, 2006; Korn *et al.*, 2008; Lai *et al.*, 2005; Olshen *et al.*, 2004; Wang *et al.*, 2007). In particular, using SNP genotyping arrays to simultaneously produce estimates of integer CN and SNP genotype has become popular, particularly as a means to identify both SNPs and CNVs associated with disease. CNV association analyses are conducted either on estimates of integer CN genotype (Barnes *et al.*, 2008; Korn *et al.*, 2008), or using normalized continuous intensity data measurements (Barnes *et al.*, 2008).

As a result of intensive genotyping efforts worldwide for genome-wide association studies, there are many datasets containing inferred CNV regions, CNV genotypes in these regions, as well as SNP genotypes. However, CNV–SNP haplotypes are rarely determined in these datasets, due largely to a lack of algorithmic development in this area. Hence, haplotype-based approaches that have been shown to be more powerful than single-marker analyses (Liu *et al.*, 2007; Mailund *et al.*, 2006; Su *et al.*, 2008a) are not fully exploited in CNV association studies.

Apart from improving the sensitivity of association studies, CNV–SNP haplotypes are also invaluable for studying the evolutionary history of CNVs. In particular, many techniques for detecting positive selection rely on accurate phasing (Sabeti *et al.*, 2007). Similarly, CNV–SNP phasing will improve accuracy of estimates of linkage disequilibrium (LD) between SNPs and CNVs (Conrad *et al.*, 2009; de Smith *et al.*, 2008), particularly for multi-allelic CNVs. Identification of the haplotypic background(s) of a given CNV, will also help to distinguish single versus recurrent deletion/amplification events and will also shed light on the age of the CNV (from the size of the extended haplotype containing the CNV).

*To whom correspondence should be addressed.



Deletion	Local duplication	Dispersed duplication
LD between deletion and flanking SNPs (including pre-existing and new SNPs)	LD between duplication states and flanking SNPs (including pre-existing and new SNPs)	-
-	LD within duplication relative to original unduplicated sequence	LD within duplication relative to original unduplicated sequence

Fig. 1. Illustration of the process of forming a deletion and a local/dispersed duplication. The light grey box represents deleted or duplicated region. The deleted or duplicated region is transmitted over generations. The pre-existing SNPs (represented in white circles) and new SNPs (represented in dark grey circles) form the background patterns of haplotypes, from which the correlation between deletion/duplication states and flanking SNPs is captured in our model for non-internal phasing. For internal phasing, our model exploits the correlation between SNPs within duplicated regions.

Methods for inferring haplotypic phase from diploid genotypes are well developed and provide accurate inference of haplotypic phase (Browning and Browning, 2009, 2007; Kimmel and Shamir, 2005; Scheet and Stephens, 2006; Stephens and Scheet, 2005; Su *et al.*, 2008a). For polyploid organisms, two phasing programs, SATlotyper (Neigenfind *et al.*, 2008) and polyHap(v1.0) (Su *et al.*, 2008b), have been proposed. By treating a CN region as a region of variable ploidy, polyHap(v1.0) can also be used for phasing CNV regions providing that the ploidy is fixed for the entire genomic region under investigation (although this can be different for different individuals). Thus, phasing complex CNV regions, in which each individual can have different CNV breakpoints, is a problem that has yet to be fully addressed.

To properly define what is meant by CNV haplotyping, we consider in Figure 1 how a CNV might arise in an ancestral genome, and subsequently be transmitted from one generation to the next. We consider separately the cases of deletion, local and dispersed duplication as illustrated in Figure 1. For both a deletion and local duplication, LD can accumulate between flanking SNPs and the CN state itself. Studies have estimated that the majority of common genotypeable CNVs are well tagged by SNPs (Conrad *et al.*, 2009; de Smith *et al.*, 2008). Exploiting this LD pattern to infer which haplotype contains the CNV is called *non-internal* phasing in our study. For bi-allelic SNPs within duplications, non-internal phasing also provides an estimate of *allelic configuration*—e.g. distinguishing two possible configurations AA/B and AB/A for a genotype AAB. On the other hand, the duplicated regions can themselves contain SNPs, which either arose prior to the duplication event, during the duplication event (through imperfect copying) or subsequently. We refer to exploiting the LD patterns within duplicated regions in order to identify the haplotypes comprising the duplication as *internal* phasing. The difference between internal and non-internal phasing—as applied to diploid genomes—is further illustrated in Figure 2. Non-internal phasing reconstructs haplotypes consisting of CNV–SNP alleles $\{-, A, B, AA, AB, BB, \ldots\}$ from diploid genotypes, but does not phase within duplicated alleles. Internal phasing, on the other hand, reconstructs all of the underlying SNP haplotypes present in the dataset, including those within duplicated regions. Thus, it considers diploid genomes as locally polyploid, where the ploidy is given by the maximum CN. MOCSphaser (Kato *et al.*, 2008b) was the first program developed for inferring (non-internal) CNV–SNP haplotypes using an expectation maximization (EM) algorithm. However, it only accommodates CNs in CNV regions (does not consider variant bases in these regions) and SNP genotypes in non-CNV regions. Another recently proposed non-internal phasing program for CNV haplotype inference, CNVphaser, employed an EM and partition–ligation (PL) algorithms to infer haplotypic phase given identified CNV regions and CNs (Kato *et al.*, 2008a).

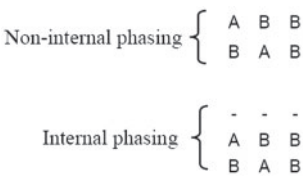


Fig. 2. Illustration of non-internal and internal phasing with a deletion and a single copy amplification. Non-internal phasing considers the genotypes as diploids and treats the duplication and deletion as extra different alleles, whereas, internal phasing considers the genotypes as triploids and introduces an extra chromosome copy. This chromosome will accommodate the extra copy of the duplicated region and will otherwise be set to a deletion state.

In this article, we describe an algorithm for both internal and non-internal CNV haplotype inference from CNV/SNP genotype data, which takes account of the shared haplotype structure between individuals in a population. Our method, polyHap(v2.0) extends the model of polyHap(v1.0) (Su *et al.*, 2008b) to phase complex CNV regions by allowing arbitrary changes of CN within individuals and along the genomic sequence.

To investigate the effectiveness of our approach, we took SNP/CNV genotype data on male X chromosomes and randomly paired these into phase-known diploid and triploid haplotypes. We then investigated how well we could reconstruct the known phase and allelic configuration as well as infer missing CNV genotypes. The results show that our method provides accurate estimates of missing CNV genotypes, allelic configuration and haplotypic phase. We applied polyHap to a region on Chromosome 2 encompassing a short deletion. The results show that polyHap correctly detected a haplotype comprising this deletion.

2 METHODS

Our method employs a hidden Markov model (HMM) to infer an ancestral haplotype for each haplotype at each marker, reflecting the idea that similar haplotypes are likely to have descended from the same ancestral haplotype. Assume we observe the genotypes, $g=(g_1, g_2 \ldots, g_M)$, at M SNPs for each individual. $g_m=\{g_{m1}, \ldots g_{mN}\}$ is an unordered list of the individual's alleles at marker m, where N is the ploidy. For non-internal phasing, we infer haplotypic phase on diploid chromosomes, where N equals to 2. For internal phasing, we consider genotypes as polyploids, where N is set to the maximum CN (ploidy) observed on the individual (Fig. 2). Also, $s_m=\{s_{m1}, \ldots, s_{mN}\}$ and $s'_m=[s_{m1}, \ldots, s_{mN}]$ are the unordered and ordered lists of ancestral haplotypes at marker m, respectively. We write $\pi(s'_m)=[s_{m\pi(1)}, \ldots, s_{m\pi(N)}]$ for a permutation of s'_m, and $\Pi(s_m)$ for the set of all such permutations. Thus, for example, if $s'_m=[1,2]$ there are two permutations, namely $[2,1]$ and $[1,2]$, whereas if $s'_m=[1,1]$ there is only one permutation.

2.1 Emission probability

In our method, each allele is assumed to be descended from one of z ancestral haplotypes, which are the hidden states (haplotype states) in the HMM. The program first learns the ancestral haplotype structure from genotypes jointly for all individuals. Based on this structure, allelic configuration, missing CNV genotypes and CNV haplotypic phase are then inferred. This relationship between the allele and the haplotype hidden state is modelled by the emission probability. In this study, we allow a deletion and a single copy amplification. Thus, the set of possible alleles is {-, A, B, AA, AB, BB} underlying a diploid model when non-internal phasing is considered, while the set of possible alleles is {-, A, B} underlying a polyploid model for internal phasing.

First, we define the emission probability of each genotype given a haplotype state. Let $\theta_{ml_n}(h)$ denote the emission probability of allele h at marker m given the haplotype state l_n in a haploid model, where $h \in$ {-, A, B, AA, AB, BB} for non-internal phasing and $h \in$ {-, A, B} for internal phasing.

We first obtain the emission probability of a list of unordered haplotypes, given an unordered list of haplotype states $\{l_1, \ldots, l_N\}$ by

$$\begin{aligned} p(g_m &= \{h_1, \ldots, h_N\} | s_m = \{l_1, \ldots, l_N\})) \\ &= \textstyle\sum_{\pi \in \Pi(g_m)} \prod_{n=1,\ldots,N} p(g_{mn} = h_{\pi(n)} | s_{mn} = l_n) \\ &= \textstyle\sum_{\pi \in \Pi(g_m)} \prod_{n=1,\ldots,N} \theta_{ml_n}(h_{\pi(n)}), \end{aligned} \quad (1)$$

where $\theta_{ml_n}(h_{\pi(n)}) = p(h_{\pi(n)}|l_n)$.

For non-internal phasing, a given CNV genotype (e.g. AAB) may be consistent with more than one unordered list of haplotype pairs (e.g. AA/B and AB/A). In this case, the observed data is represented as probability distribution p^*_m over unordered haplotype pairs (e.g. such that $p^*_m(AB/A) = p^*_m(AA/B) = 0.5$). We then write

$$p(p^*_m|s_m) = \sum_g p^*_m(g) P(g|s_m) \quad (2)$$

for the emission probability, using Equation (1) to calculate the terms in this sum. We note that a normal copy genotype, e.g. AA is also consistent with two different unordered haplotype pairs, namely A/A as well as AA/-; however, we currently exclude the AA/- haplotype pair from our analysis. Equation (2) can also be used to accommodate uncertain CN genotypes, in which case p^*_m to reflects the probability of each CNV/SNP genotype as calculated by the CNV genotyping algorithm used. If g_m is missing, we set p^*_m to be the uniform distribution over all CNV/SNP genotypes.

2.2 Transition probability

We first briefly describe a basic transition model for internal phasing. We then introduce the extension of this model for non-internal phasing by considering the transitions between the CNs and between the haplotypes. In this extension, a given haplotype hidden state has a fixed CN and there can be multiple haplotype states underlying each CN. In this study, we use eight ancestral haplotype states for internal phasing and nine haplotype states for non-internal phasing of which one haplotype state has the underlying CN = 0 (deletion), four haplotype states have the underlying CN = 1 (normal copy) and four haplotype states have the underlying CN = 2 (a single copy amplification). Note that the CN states are the super states which categorize haplotype states according to their underlying CNs.

2.2.1 A basic haplotype transition model First, we define the transition probability in a HMM from haplotype states k_n to l_n between markers $m-1$ and m by

$$p(s_{mn} = l_n | s_{(m-1)n} = k_n) = \begin{cases} (1-J_m) + J_m \alpha_{ml_n} & l_n = k_n \\ J_m \alpha_{ml_n} & l_n \neq k_n, \end{cases} \quad (3)$$

where J_m is the probability of a jump occurring at marker $m-1$, and α_{ml_n} is the probability that this jump results in the haplotype l_n. For tightly linked markers, J_m is small so that haplotype state changes occur infrequently, but are allowed between any pair of markers. Here, the parameter J_m is independent of the state and α_{ml_n} only depends on the l_n (the 'to') state.

2.2.2 A modified haplotype transition model We further modify this model to allow different models for the transition between CN states, for the transition between haplotype states that have the same CN state, and for the transition between haplotype states that have different CN states. To incorporate the CN state in the transition model, we introduce a hierarchy transition model—the first transition level is the transition between the CN states and the second is between the haplotype states given the CN states (Fig. 3). The idea of using this model is to capture the favoured transition between the CNs.

The transition probability from haplotype state k_n to l_n is then the product of the transition probability between CN states and the transition probability between haplotype states given the CN states

$$\begin{aligned} p(s_{mn} &= l_n | s_{(m-1)n} = k_n) \\ &= p[c(s_{mn}) = c(l_n) | c(s_{(m-1)n}) = c(k_n)] \\ &\times p[ci(s_{mn}) = ci(l_n) | ci(s_{(m-1)n}) = ci(k_n)] \end{aligned} \quad (4)$$

where $c(l_n)$ and $c(k_n)$ are the underlying CN state for haplotype states l_n and k_n, respectively; and $ci(l_n)$ and $ci(k_n)$ are the indices of haplotype states l_n and k_n within the CN states $c(l_n)$ and $c(k_n)$. Both transition probabilities (the two terms of the product in the equation) are calculated based on Equation (3) with different parameters.

In this modification, we allow that the parameters J_m depend on the k_n ('from') state, denoted as J_{mk_n} and α_m is related to both the k_n and l_n ('from' and 'to') states, denoted as $\alpha_{mk_nl_n}$. To capture linkage disequilibrium between

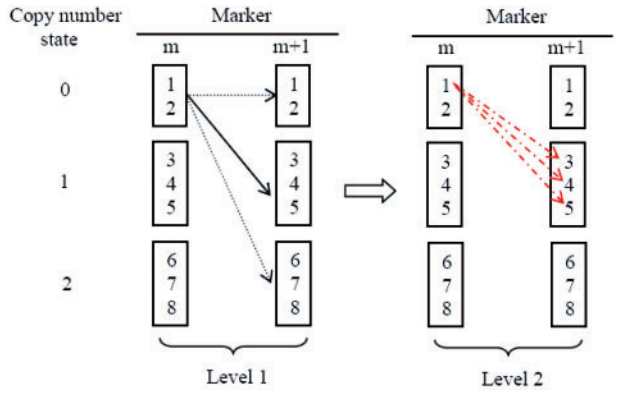


Fig. 3. Illustration of two levels of the transitions based on the haploid model. Each box represents the CN state and the numbers in the box are the assigned haplotype states. The first level of the transition (which is between the CN states) can be considered as the transition between the boxes. The dashed and solid lines in the left panel give an example of the possible transitions from CN state 0 (the solid line represents the most likely transition). The second level of transition (which is between the haplotype states) can be considered as the the transition between the numbers given the boxes, where the dashed lines give an example of the transitions from the haplotype State 1 to the haplotype States 3, 4 and 5 given that the transition between CN states is from 0 to 1. Note that the number of haplotype states in each CN state can be specified by users 1.

duplication states and flanking SNPs, we use Equation (3) with parameters J_{mk_n} and $\alpha_{mk_nl_n}$ to compute the transition probability between CN states and between haplotype states given the transition occurring in different CN states. We use the basic transition model (the parameter J_m is independent of the state and α_{ml_n} only depends on the l_n state) to calculate the transition probability between the haplotype states given the same CN state.

2.2.3 Polyploid transition model We use the modified haplotype transition model for non-internal phasing and basic haplotype transition model for internal phasing. Based on these transition models, the transition probability between unordered lists of haplotype states $k=k_1,\ldots,k_N$ and $l=l_1,\ldots,l_N$ at marker m is given by

$$p(s_m = \{l_1,\ldots,l_N\} | s_{(m-1)} = \{k_1,\ldots,k_N\}) = \sum_{\pi\in\Pi(s_m)} \left(\prod_{n=1,\ldots,N} p(s_{mn} = l_{\pi(n)} | s_{(m-1)n} = k_n)\right), \quad (5)$$

2.3 The prior and computation

We use Dirichlet priors on all of our parameters. We let $\theta_{ml} \sim \text{Dirichlet}(u_\theta \mathbf{m}_\theta)$, where $\mathbf{m}_\theta$ is the uniform vector with each element equal to $1/H$ (H is the length of allele space), and $\alpha_{m\cdot} \sim \text{Dirichlet}(u_\alpha \mathbf{m}_\alpha)$ where $\mathbf{m}_\alpha$ is the uniform vector with each element equal to $1/z$ (z the number of ancestral haplotypes). We let $J_m \sim \text{Beta}(u_J(1-e^{-d_m r}), u_J e^{-d_m r})$ where d_m is the physical distance between consecutive markers and $r=10^{-8}$ per based pair in the population, reflecting the background recombination rate. We use $u_\theta = u_\alpha = 1$ and $u_J = 10^5$ for initialization of the EM algorithm and $u_\theta = u_\alpha = u_J = 0.1$ for the maximization step.

Although our HMM has many parameters, approximate posterior mode estimates are readily obtained using the Baum–Welch algorithm, which is a form of the EM algorithm. The parameters in the model are updated at each step of the EM algorithm given the observed genotype data. The training process might converge to a local maximum of the likelihood function, which is a typical problem for the EM algorithm. To deal with this problem, we combine the results from 10 repetitions of the EM algorithm with different start values. In our model, the first-order Markov chain is employed to model ancestral haplotypes across the sequence. Thus, the number of EM iterations does not depend on the number of markers. A default number of iterations is 25 for each repetition of the training process, which can be specified in our parameter file.

After obtaining the estimates of parameters at each repetition, a specified number of haplotypes are sampled from the posterior distribution conditional on the genotype data of a given individual (Su *et al.*, 2008b). Here, we obtain 100 samples for each repetition. The most likely haplotype is then inferred from all the sampled haplotypes across the 10 repetitions of the EM algorithm. The certainty rate of this estimate is the fraction of times it is sampled. Because we consider only a small number (e.g. 10) of local modes of the posterior distribution for the HMM parameters, the certainty value is not the probability of the imputed genotype under the model, which would require integration over the posterior distribution, but it may serve as a reasonable approximation to this probability.

3 SIMULATION STUDY

In this section, we present the details of the simulation study to evaluate the performance of our method for inferring allele configurations, CNV–SNP haplotypes and missing CNV genotypes. We simulated phase-known datasets based on data obtained from French and Finnish population cohorts, respectively, with different technologies for obtaining the CN status and using different genotyping chips. The French dataset contains fewer samples but denser CNV–SNP genotypes, while the Finnish dataset contains more samples but less dense genotypes.

3.1 The French samples

We obtained data for X chromosomes from 48 males of northern French origin who were genotyped both on the Illumina 1M platform and 244K aCGH platform. The 244K aCGH chips, custom-designed for focussed investigation for putative CNV regions, provide information on the locations of CNV regions as well as CNs in these regions across the entire genome (de Smith *et al.*, 2007). In aCGH, test and reference DNA samples, which are labelled differentially with fluorescent tags, are competitively hybridized into genomic arrays. The fluorescence ratio of test and reference hybridization signal is then determined at different positions along the genome, which provides information on the difference in CNs between test and reference samples.

CNV regions on non-pseudo-autosomal regions of the male X chromosome were identified from 244K aCGH chip data using the ADM2 algorithm developed by Agilent Technologies (Santa Clara, CA, USA), which recursively searches for CNV intervals based on log R ratios (LRRs) of fluorescent signals from probes between test and reference DNA sample (de Smith *et al.*, 2007). A single sample from the Coriell Cell Repository (NA15510) was used as reference. The boundary and size of the CNV intervals are defined on the basis of the positions of the first and last array probes identified as lying within the CNV. The integer CN of the CN region was set to 0 if the average LRR of probes within the region was less than −0.5 (i.e. deletion) and was set to 2 otherwise (i.e. amplification on male haploid background). Haploid SNP genotypes in non-CNV regions were obtained from BeadStudio, using the Illumina 1M chip. Within amplified regions, two-copy SNP genotypes were estimated from a Gaussian mixture model using the B-allele frequency from BeadStudio. For this dataset, we analysed a 2.7 Mb non-pseudo-autosomal region of the X-chromosome (151 881 226–154 588 828 bp based on NCBI build 36) This region has 1904 aCGH probes (equally 1 probe for every 1.4 kb) and 1058 Illumina SNP probes.

3.2 The Finnish samples

We also assessed our method using a larger dataset from the Northern Finland Birth Cohort (NFBC), from which we obtained non-pseudo-autosomal X-chromosome genotype data on 695 Finnish males assayed on Illumina Hap370 chips. aCGH data were unavailable for this cohort; however, Illumina's BeadStudio software generates the log ratio of observed to expected fluorescent signal intensity (LRR), as well as a normalized measure of relative signal intensity between the two SNP alleles the B allele frequency (BAF), which can be used to detect CNV regions and infer CN genotypes (Colella *et al.*, 2007; Wang *et al.*, 2007). Haploid SNP genotypes were obtained from BeadStudio, while two copy SNP genotypes within amplifications were obtained on the basis of BAF. For this dataset, we analysed a 20.9 Mb region on the X chromosome (19 502 220–40 491 848 bp based on NCBI build 36), which contains 2149 markers.

3.3 Simulation of phase-known genotypes

We randomly combined SNP/CNV genotypes on male X chromosomes into pairs to create diploid genomes with up to four copies for non-internal phasing (Fig. 2). We created 24 'non-internal' phase-known diploid genomes in the French dataset and 347 genomes in the Finnish dataset. These samples were inappropriate for internal phasing as the 'internal' haplotypes comprising the amplifications on each X-chromosome copy are not known. Thus, to evaluate internal phasing, we masked X-chromosome amplifications, and randomly grouped these X-chromosomes into 15/231 French/Finish triploid genomes, so that we obtain internal + external phase-known SNP/CNV genotype data with up to three copies.

3.4 Switch error rate

The switch error rate for each individual is defined as $\psi/(n-1)$, where n denotes the number of heterozygous sites for that individual and ψ the minimal number of switches needed to recover the true haplotypes. We assumed that at most one switch could occur between consecutive heterozygous sites.

For each individual, we determined if there was a switch by comparing the inferred haplotypes to the true haplotypes. If a discrepancy is identified at a heterozygous marker m, a switch error is counted and a switch is introduced in the inferred haplotypes to ensure that it matches the true haplotypes up to marker m. To identify a discrepancy, it is only necessary to compare haplotype sets as far back as to distinguish N distinct preceding haplotypes (N is the ploidy), which in diploids requires looking back to the previous heterozygous marker only.

4 RESULTS

4.1 Missing data imputation

We first examined the accuracy of our method for missing data imputation with both French and Finnish data. In each dataset, 5% and 10% of genotypes with one to four copies of alleles were set as missing at random, respectively. We report the proportion of missing genotypes for each CN that were estimated incorrectly (imputation error rate). Table 1 shows the imputation error rate in the French and Finnish datasets, respectively. Overall, our method provides accurate estimates of missing genotypes. For both missing rates (5% and 10%), our method gives an imputation error rate <0.09.

Table 1. Error rate for estimation of missing genotype

	CN of genotype			
Missing rate	1	2	3	4
French dataset				
5%	0.020	0.034	0.060	0.0
10%	0.009	0.030	0.090	0.0
Finnish dataset				
5%	0.062	0.075	0.053	0.028
10%	0.050	0.081	0.050	0.027

Table 2. The distribution of CN and error rate of estimation of allele configuration at heterozygous sites

	CN of genotype			
	1	2	3	4
French dataset				
CNs	1155	18 318	6317	754
Heterozygous genotypes	0	2932	1075	94
Error rate of allelic configuration	NA	NA	0.119	0.0
Finnish dataset				
CNs	5609	664 847	60 010	15 237
Heterozygous genotypes	0	210 239	24 106	1572
Error rate allelic configuration	NA	NA	0.016	0.188

4.2 Allelic configuration inference

We assess the performance of our method for inferring allelic configuration on a pair of haplotypes (such as AA/B versus A/AB). Table 2 presents the distribution of CNs observed on all markers and the error rate of estimated allele configurations. In the French data, there are 6317 and 754 3-CN and 4-CN genotypes of which 1075 and 94 are heterozygous, respectively. The allelic configuration is ambiguous for all of these heterozygous 3-CN and 4-CN genotypes (excluding genotypes AAAB and ABBB for which AA/AB and AB/BB, respectively, are the only possible configurations). The allelic configuration error rate amongst these ambiguous 3-CN and 4-CN genotypes is 0.119 and 0.0, respectively. In the Finnish data, the corresponding error rates are 0.016 and 0.188, based on 24 106 and 1572 heterozygous 3-CN and 4-CN genotypes.

4.3 Inference of haplotypic phase of CN state relative to flanking SNPs (non-internal phasing)

We assessed the performance of our method for haplotypic phase inference using the switch error rate. In this case, the CNV/SNP alleles consist of $\{-, A, B, AA, AB, BB\}$ and we do not distinguish the order of alleles within an amplification. Hence, when calculating switch error rate, homozygous 3-CN genotypes (A/AA or B/BB) are considered as heterozygous sites as the CNV/SNP alleles are different for each haplotype. Homozygous genotypes 4-CN genotypes (AA/AA or BB/BB) are still considered as homozygous

Table 3. Switch error rate for non-internal phasing

	CN on the second site (N_2)			
CN on the first site (N_1)	1	2	3	4
French dataset				
1	0.0009 (1062)	0.26 (15)	0.571 (7)	NA (0)
2	0.2 (15)	0.036 (2866)	0.413 (29)	0.0 (1)
3	0.333 (6)	0.322 (31)	0.0008 (6150)	0.0 (2)
4	NA (0)	NA (0)	0 (3)	0.057 (87)
Finnish dataset				
1	0.067 (1373)	0.360 (3022)	0.396 (232)	0.142 (7)
2	0.383 (2810)	0.071 (204 467)	0.264 (2551)	0.386 (101)
3	0.282 (436)	0.158 (2163)	0.001 (56 688)	0.188 (303)
4	0.333 (9)	0.235 (289)	0.357 (112)	0.076 (286)

sites. In calculating the switch error, we excluded the sites where the allelic configurations were incorrectly inferred.

For the French data, the overall switch error rate is 0.015. We then classified transitions by the 'from' and 'to' genotype CN (denoted by $N_1 \rightarrow N_2$) to get error rates in Table 3. The number of observed $N_1 \rightarrow N_2$ transitions is shown in brackets. Overall, the switch error rates are <0.34, apart from two cases where the error rates are 0.57 and 0.41 at heterozygous sites with CNs $1 \rightarrow 3$ and $2 \rightarrow 3$, respectively. The reduced accuracy in the French data at such sites is due to the fact that the number of observations is small, and moreover may not all occur at the same CN breakpoints. Accuracy for these CN transitions is improved by increasing the population size as can be seen in the corresponding results for the Finnish data.

Figure 4 shows the error rate for each transition in the Finnish data distributed according to certainty score. In general, the estimate with the higher certainty rate provides the more reliable inference. However, we observed a low proportion of estimates that have a high certainty rate (>0.9) in some cases, such as CNs $4 \rightarrow 3$. Supplementary Figure S1 shows that the level of LD between SNPs and CNVs, as measured by r^2, is inversely correlated with switch error.

To compare the results with those from CNVphaser and MOCSphaser, we chose three and eight sites in two different CNV regions from the French data. The maximum number of CNV sites used in the original CNVphaser article (Kato *et al.*, 2008a) is eight. MOCSphaser could not be run on the eight site data because it ran out of memory on a 32 GB machine. We also attempted to run CNVphaser using the same number of sites as presented for polyHap in Table 3, but we found that the scale of our simulated dataset was not computationally feasible for CNVphaser.

CNVphaser and MOCSphaser both return a posterior probability distribution over possible haplotypes given the observed genotypes. We selected the haplotype with the highest probability as the inferred haplotype. We show the number of individuals whose genotypes are not correctly phased at any heterozygous sites in Table 4. The CNV genotypes at three sites are all correctly phased by both polyHap and CNVphaser/MOCSphaser. For the genotypes at eight sites, the results from polyHap show that only one individual has a single switch error over all the sites, while most of the inferred haplotypes from CNVphaser are incorrect. The allele configurations are incorrectly inferred in most heterozygous sites by CNVphaser.

Previous studies have also used fastPhase and Beagle for CNV phasing (Conrad *et al.*, 2009). This approach is limited to phasing bi-allelic CNVs relative to flanking SNPs not in CNV regions, which is achieved by recoding bi-allelic CNV genotypes as SNP genotypes. To compare our method to this approach, we removed multi-allelic CNVs from the Finnish dataset, and also masked SNPs within CNV regions, and finally encoded CNV genotypes as SNP genotypes. We then ran each of fastPhase/Beagle and polyHap on this dataset (Supplementary Table 1). Comparing with Table 3, we see that switch error rates for polyHap have markedly increased in most cases due to loss of information from masking SNPs. Comparing algorithms on the masked dataset (Supplementary Table 1), we see that polyHap and fastPhase had comparable switch error rates between SNPs and CNVs with different CN transitions, while Beagle had higher error rates on these CN transitions except for CN $3 \rightarrow 2$.

Finally, to test polyHap's accuracy on a non-simulated dataset we successfully phased a region on chromosome 2 containing a known short (<3 kb) deletion at 229.467 mb, using data generated by a 244K Agilent array CGH chip (de Smith *et al.*, 2007). A consistent haplotype including this deletion was detected (Supplementary Fig. S2). This deletion has been previously verified by polymerase chain reaction (PCR) across the breakpoints followed by sequencing (de Smith *et al.*, 2007).

4.4 Inference of haplotypic phase of SNPs within CN states (internal phasing)

Internal phasing can be considered as a tool for further investigating haplotypic phase of duplicated alleles locally. Thus, we report the switch error rate at sites that have the same CN. Note that here we only consider up to a single copy amplification at the genotype level. Table 5 gives the switch error rate between a pair of consecutive heterozygous sites, which have the same CN. The count of each pair of CN is shown in parentheses. For both French and Finnish datasets, the error rates are ≤0.08 for locally inferring haplotypic phase of duplicated alleles.

5 DISCUSSION

We have presented a method for inferring haplotypic phase for CNV/SNP genotype data among unrelated individuals. Our method allows CNV regions and ploidy to vary along the sequence and between the individuals. Our program accommodates both CNV and SNP genotype data and infers missing genotypes and haplotypic phase for both types of data. Our method allows uncertainty in the CN assignment by representing the CNV genotype as a probability distribution over multiple CNV genotypes.

It is necessary to first calculate CNV genotypes prior to running our program. In particular, polyHap does not accommodate a continuous measurement in place of the integer CN genotypes. polyHap can include—in principle—an arbitrary maximum number

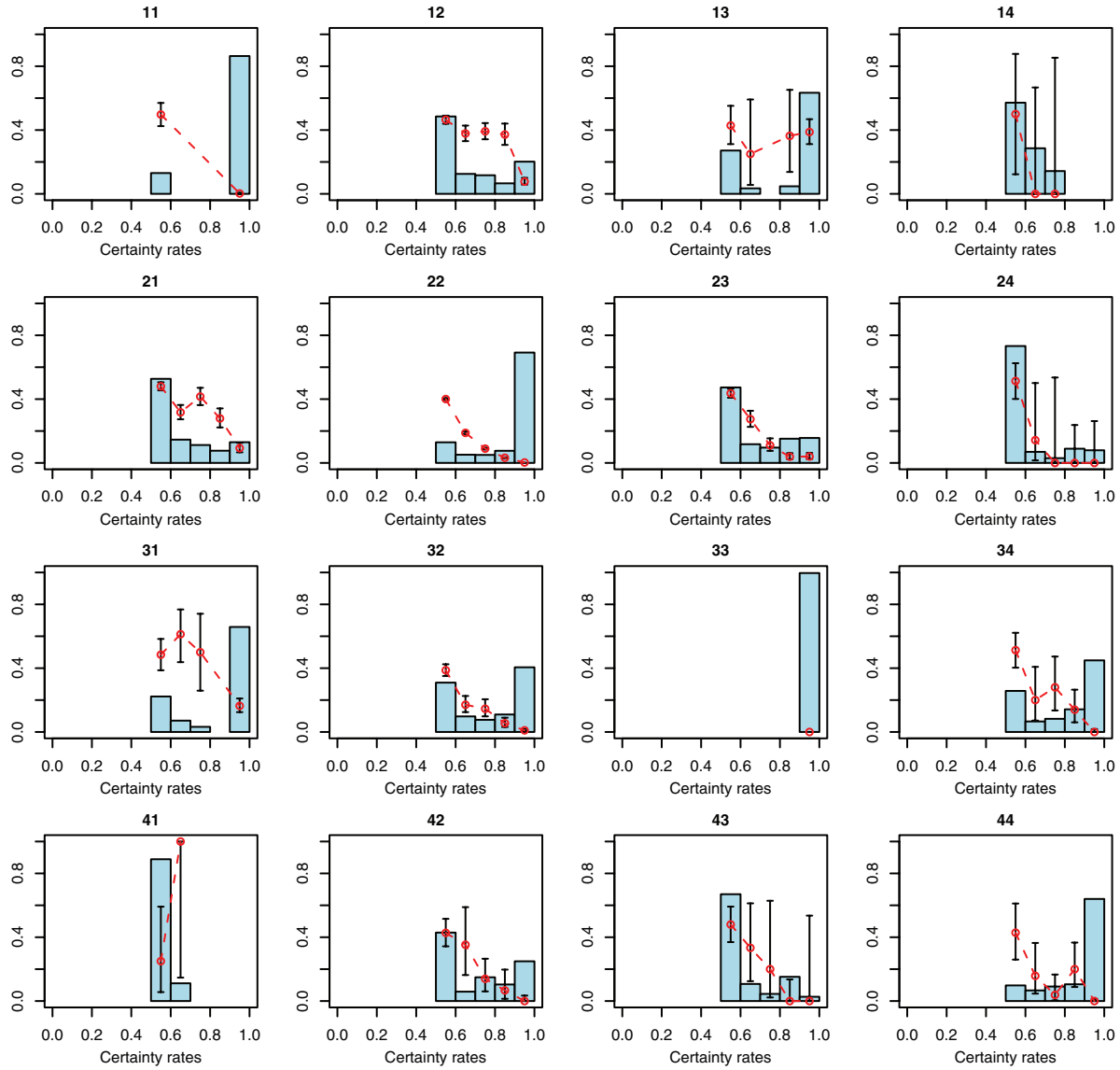


Fig. 4. Histogram of certainty scores and switch error rate in each bin from the Finnish dataset. The circles indicate average switch error rates within each histogram bin. The error bar of each switch error rate is based on a 95% equal-tailed Bayesian interval given the prior Beta(1/2, 1/2). The number on the top of each cell graph represents CNs at a pair of heterozygous sites (first digit is the CN at the first site and second is the CN at the second site).

of copies. However, as the computational complexity scales roughly as # copies^2 for non-internal phasing, and as $e^{\#\text{ copies}}$ for internal phasing, meaning that internal phasing is feasible for up to 6 copies, and non-internal for up to 20 copies. Similarly, polyHap cannot model complicated structural rearrangements, including inversions and translocations.

polyHap requires a pre-defined number of ancestral haplotypes. In this study, we use eight ancestral haplotype states for internal phasing (two CN = 0 and six CN = 1 states) and nine haplotype states for non-internal phasing (one CN = 0, four CN = 1 and four CN = 2 states). We have also tried different numbers of ancestral haplotypes and found that the results are comparable. Here, we would suggest using higher number of ancestral haplotypes when dealing with rare variants. The choice of ancestral haplotype number usually does not depend on the sample size but rather on the number of haplotypes present in the population. Thus, if a very diverse, heterogeneous population or a mixture of several populations were being analysed, then it would be advisable to include more states.

Table 4. Comparison between polyHap and CNVphaser/MOCSphaser

	Number of individuals having switch error		
Number of sites	polyHap	CNVphaser	MOCSphaser
3	0	0	0
8	1	24	NA

Table 5. Switch error rate for internal phasing with same CN

	CN on a pair sites		
	1→ 1	2 → 2	3 → 3
French dataset	0 (34)	0.005 (1034)	0.070 (3514)
Finnish dataset	0.002 (351)	0.056 (864)	0.080 (2 29 244)

The results from the simulation study demonstrate that our program provides accurate estimates of missing genotypes, allele configuration and haplotypic phase for both CNV and SNP data. Our method gives an imputation error rate <0.09 for imputing missing genotypes with one to four copies of alleles. Also, our method provides accurate estimates of allele configurations on a pair of haplotypes, with an error rate <0.19. Furthermore, polyHap successfully identified a haplotype comprising a short deletion on chromosome 2. Our method gives encouraging results for inferring CNV haplotypic phase over different CNs at heterozygous sites. Although there are several situations where the switch error rate is >0.3, this might result from rare haplotypes in the dataset, and the accuracy here would be improved by using a larger population sample. Also, reliable phase inferences can be distinguished using the uncertainty estimates. In general, a higher certainty rate indicates higher accuracy of the estimate.

polyHap outperforms two existing methods for phasing CNV-SNP haplotypes—CNVphaser and MOCSphaser—in terms of accuracy and capacity of dealing with large-scale datasets. Comparing our method with fastPhase/Beagle for phasing bi-allelic CNV, polyHap is comparable to fastPhase and gives more accurate estimates than Beagle in most cases of CN transitions. One advantage of our new method over fastPhase/Beagle for phasing CNV–SNP haplotypes is that polyHap is designed for inferring CNV–SNP haplotypes and is able to accommodate some properties of CNV that differ from SNPs and to deal with multi-allelic CNV.

Our program provides two different levels of CNV phasing—non-internal and internal. With internal phasing, the individual is considered as polyploid, and thus the phasing process is similar to that described for polyHap (Su *et al.*, 2008b). Internal phasing enables inference of the duplicated and original haplotype, but does not say which chromosome copy contains the amplification. Non-internal phasing, on the other hand provides information about which chromosome copy contains the CNV, but not the internal structure of duplications. By providing both options, our program enables the researcher to choose a suitable level of phasing for the specific purposes of the study.

Our method is faster than CNVphaser, and is feasible for genome-wide analyses using a computing cluster. The computing time for the French dataset with nine ancestral haplotype states and 10 repetitions of the EM-training algorithm (containing 1106 markers on each of 24 individuals) was ∼0.8 h on a 8 GB computer, while the Finnish dataset (containing 2149 markers on each of 347 individuals) took 1.5 h on a 16 GB computer. The computing time increases linearly with the number of markers and individuals.

Modelling the haplotypic background of CNVs will provide a better understanding of the evolutionary processes affecting CNVs. Moreover, it will help us to better model CNV–phenotype associations—to make CNV–disease associations more robust by simultaneously identifying the underlying haplotype harbouring the CNV and to disentangle associations between CNVs and phenotype from associations with flanking SNPs.

ACKNOWLEDGEMENTS

We thank Rob Sladek for providing Illumina data and Adam de Smith for providing aCGH data. The DNA extractions, sample quality controls, biobank up-keeping and aliquotting for the NFBC was performed in the national Public Health Institute, Biomedicum Helsinki, Finland. Genotyping of the NFBC samples was supported by the National Institute of Mental Health.

Funding: Research Council UK fellowship (to L.J.M.C.); Genome Canada and Genome Quebec funded genotyping on the French samples; the NFBC1966 received financial support from the Academy of Finland (project grants 104781, 120315, 132797, and Center of Excellence in Complex Disease Genetics); University Hospital Oulu, Biocenter, University of Oulu, Finland; the European Community's Fifth/Seventh Framework Programme (EURO-BLCS, QLG1-CT-2000-01643, FP7/2007-2013); NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01); ENGAGE project (HEALTH-F4-2007-201413); the Medical Research Council (centre grant G0600705); the Wellcome Trust (project grant GR069224), UK; the National Institute of Health Research (NIHR) Biomedical Research Centre Programme at Imperial College; the DNA extractions, sample quality controls, biobank up-keeping and aliquotting were performed in the National Public Health Institute, Biomedicum Helsinki, Finland, and supported financially by the Academy of Finland and Biocentrum Helsinki.

Conflict of Interest: none declared.

REFERENCES

Barnes,C. *et al*. (2008) A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.*, **40**, 1245–1252.

Browning,B.L. and Browning,S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Amer. J. Hum. Genet.*, **84**, 210–223.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Amer. J. Hum. Genet.*, **81**, 1084–1097.

Colella,S. *et al*. (2007) QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, **35**, 2013–2025.

Conrad,D.F. *et al*. (2009) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

de Smith,A.J. *et al*. (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.*, **16**, 2783–2794.

de Smith,A.J. *et al*. (2008) Small deletion variants have stable breakpoints commonly associated with Alu elements. *PLoS ONE*, **3**, e3104.

Feuk,L. *et al*. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Fiegler,H. *et al*. (2006) Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.*, **16**, 1566–1574.

Kato,M. *et al*. (2008a) An algorithm for inferring complex haplotypes in a region of copy-number variation. *Am. J. Hum. Genet.*, **83**, 157–169.

Kato,M. *et al*. (2008b) MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data. *Bioinformatics*, **24**, 1645–1646.

Kimmel,G. and Shamir,R. (2005) A block-free hidden Markov model for genotypes and its application to disease association. *J. Comput. Biol.*, **12**, 1243–1260.

Korn,J.M. *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Liu,J. *et al*. (2007) Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. *PLoS Genet.*, **3**, 0421–0430.

Mailund,T. *et al*. (2006) Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, **7**, 454.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**(Suppl. 7), S37–S42.

Neigenfind,J. *et al*. (2008) Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics*, **9**, 356.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Redon, R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sabeti,P.C. *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Stephens,M. and Scheet,P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, **76**, 449–462.

Su,S.-Y. *et al*. (2008a) Disease association tests by inferring ancestral haplotypes using a hidden Markov model. *Bioinformatics*, **24**, 972–978.

Su,S.-Y. *et al*. (2008b) Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics*, **9**, 513.

Wang,K. *et al*. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.