BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2835–2843
doi:10.1093/bioinformatics/btt503

Genome analysis

Advance Access publication September 3, 2013

A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads

Kaname Kojima*, Naoki Nariai, Takahiro Mimori, Mamoru Takahashi, Yumi Yamaguchi-Kabata, Yukuto Sato and Masao Nagasaki*

Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8573, Japan

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Variant calling from genome-wide sequencing data is essential for the analysis of disease-causing mutations and elucidation of disease mechanisms. However, variant calling in low coverage regions is difficult due to sequence read errors and mapping errors. Hence, variant calling approaches that are robust to low coverage data are demanded.

Results: We propose a new variant calling approach that considers pedigree information and haplotyping based on sequence reads spanning two or more heterozygous positions termed phase informative reads. In our approach, genotyping and haplotyping by the assignment of each read to a haplotype based on phase informative reads are simultaneously performed. Therefore, positions with low evidence for heterozygosity are rescued by phase informative reads, and such rescued positions contribute to haplotyping in a synergistic way. In addition, pedigree information supports more accurate haplotyping as well as genotyping, especially in low coverage regions. Although heterozygous positions are useful for haplotyping, homozygous positions are not informative and weaken the information from heterozygous positions, as majority of positions are homozygous. Thus, we introduce latent variables that determine zygosity at each position to filter out homozygous positions for haplotyping. In performance evaluation with a parent–offspring trio sequencing data, our approach outperforms existing approaches in accuracy on the agreement with single nucleotide polymorphism array genotyping results. Also, performance analysis considering distance between variants showed that the use of phase informative reads is effective for accurate variant calling, and further performance improvement is expected with longer sequencing data.

Contact: nagasaki@megabank.tohoku.ac.jp or kojima@megabank.tohoku.ac.jp

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

Received on March 19, 2013; revised on August 26, 2013; accepted on August 27, 2013

1 INTRODUCTION

Owing to the progress of next-generation sequencing (NGS) technologies, whole-genome sequencing for each individual becomes possible in practical time and with reasonable cost for the identification of disease-associated mutations. Also, individual genome data from case–control study or study with pedigree analysis contribute to the elucidation of unknown disease mechanisms. Because accurate variant calling is required for the analysis of these genomes in a reliable manner, the development of accurate variant callers is demanded.

In most of the variant callers, variants such as single nucleotide polymorphisms (SNPs), insertions and deletions are detected at each position in a reference genome from the information of mapped sequence reads. Because there exist positions with insufficient coverage of reads due to bias in the library preparation and mapping failures at short tandem repeat polymorphic sites or variable number of tandem repeat sites, reliable variant calling is challenging at these sites (Li *et al.*, 2009; Treangen and Salzberg, 2011). To improve the reliability of variant calling, several approaches have been proposed, where additional information such as pedigree information among individuals and linkage disequilibrium among SNPs is considered (Cartwright *et al.*, 2012; Chen *et al.*, 2013; Li *et al.*, 2012).

At present, read length of sequence reads from most of the next-generation sequencers is <250 bp.

However, read length is still extending from the improvement of sequencing protocols, and several sequencers such as Pacific Biosciences' RS sequencer can produce sequence reads with thousands of base pairs although their throughput is not high and their read error rate is relatively high compared with other sequencers such as Illumina HiSeq (Quail *et al.*, 2012). Longer sequence reads are more likely to cover two or more heterozygous genotyping positions, and the use of reads spanning multiple heterozygous positions is effective for accurate phasing (He *et al.*, 2012; Menelaou and Marchini, 2013).

Thus, an approach considering such phase informative reads is promising for reliable variant calling even in low coverage regions.

We propose a new statistical variant calling approach that considers pedigree information and haplotyping based on phase informative reads. In the estimation process, genotyping and haplotyping by the assignment of each read to a haplotype based on phase-informative reads are simultaneously performed in a probabilistic manner. Thus, positions that have low evidence for heterozygosity owing to the lack of mapped sequence reads are rescued by the phase informative reads, and such rescued variant positions also can be used for the evidence on phasing.

*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

In addition, the use of pedigree information is effective not only for phasing but also for estimating variant positions because low coverage regions in each individual are supported by sequence reads covering the same regions in other individuals via pedigree information. From the synergistic effect by the use of phase-informative reads and pedigree information, more accurate variant calling is expected in our proposed approach than other variant callers estimating variants independently on positions and individuals.

In the haplotyping process of our approach, sequence reads are assigned to a more probable haplotype between two haplotypes based on a binomial model, and heterozygous positions are phased by the assigned reads spanning multiple heterozygous positions.

Although homozygous positions are not informative for phasing and cannot be used to distinguish between haplotypes, they are also used for data for the binomial model for phasing. Because the majority of positions are homozygous, information from heterozygous positions is weakened and assignment of reads to haplotypes can be misled. To address this issue, we introduce latent variables that determine zygosity at each position to the model. By filtering out the information from highly probably homozygous positions by using the latent variables, only information from heterozygous positions is selectively used for phasing.

The remainder of this manuscript is organized as follows. We describe a statistical model of our proposed approach and provide how the parameters and genotypes are estimated in Section 2. In Section 3, we compare the performance of our proposed approach and other existing variant callers through a simulation data analysis and real data analysis with human whole-genome sequencing data from Illumina HiSeq 2000 and SNP array genotyping results from Illumina OMNI 2.5 BeadChip. We finally discuss and conclude the performance evaluation results and effective points of our approach in Section 4.

2 METHODS

Our approach uses mapped sequence reads to a reference genome on the SAM/BAM format and pedigree information of individuals as input data, and then variant calling results are returned in variant calling format (VCF) (Danecek *et al.*, 2011) after the estimation. The model of our approach comprises three parts: allele likelihood part, pedigree part and haplotype selection part. Allele likelihood part represents a likelihood function of an allele, given aligned reads at a position under the consideration of sequencing errors and mapping errors. Pedigree part represents the allele transmission from parents to offspring, and haplotype selection part represents selection of haplotype from which sequence reads are generated. In the following sections, we describe the details of these three parts and procedures for parameter estimation and genotype inference in the proposed model.

2.1 Allele likelihood part

Let R_{xi} be the ith read in a SAM/BAM file for individual x. R_{xi} contains its mapping quality score in Phred scale $MAPQ_{xi}$, strings for bases aligned to position k in the reference genome r^k_{xi} and vectors of base quality scores in Phred scale for the aligned bases bq^k_{xi}. Note that r^k_{xi} is just one nucleotide such as 'T' in many cases, but it can be a string with more than one nucleotide for representing insertion, e.g. a string 'TGC' represents an insertion 'GC' right after a base 'T'. The r^k_{xi} can also be a string of length 0 to represent deletion. By using these notations, we give a likelihood function for allele A at position k as:

$$P\left(r^k_{xi}|A, bq^k_{xi}\right) = \sum_{b^k_{xi}=0,1} \sum_{m^k_{xi}=0,1} P\left(r^k_{xi}|A, b^k_{xi}\right)^{I(m^k_{xi}=1)} \times P_{mis}\left(r^k_{xi}\right)^{I(m^k_{xi}=0)} P\left(m^k_{xi}\right) P\left(b^k_{xi}|bq^k_{xi}\right), \quad (1)$$

where m^k_{xi} is a binary variable that takes 1 if the alignment of r^k_{xi} is correct and 0 otherwise. The term b^k_{xi} is a vector of binary variables, and each element indicates the correctness of each base in r^k_{xi}, i.e. if sequencing of a base is correct, the corresponding element takes 1 and 0 otherwise. The term $I(\cdot)$ is an indicator function that returns 1 if condition in its argument is true, and 0 otherwise. As with r^k_{xi}, allele A is represented by a string with nucleotides 'A', 'T', 'G' and 'C'. The term $P(r^k_{xi}|A, b^k_{xi})$ in Equation (1) is the probability of read generation for the correct read alignment, and we represent the probability as:

$$P\left(r^k_{xi}|A, b^k_{xi}\right) = \mathrm{Indel}\left(A, r^k_{xi}\right) \prod_{l=1}^{\min(|A|, |r^k_{xi}|)} P\left(r^k_{xi}[l]|A[l], b^k_{xi}[l]\right),$$

where $A[l]$ is the lth nucleotide of A, $r^k_{xi}[l]$ is the lth nucleotide of r^k_{xi}, $b^k_{xi}[l]$ is the lth value of b^k_{xi} and function Indel represents read skip errors and insertion errors. $|\cdot|$ takes a string or set as its argument and returns length for string or size for set, e.g. $|\mathrm{ATG}| = 3$. We model function Indel by using read skip error rate δ and insertion error rate ι as:

$$\mathrm{Indel}\left(A, r^k_{xi}\right) = \delta^{I(|A|>|r^k_{xi}|)}\left(1-\delta\right)^{I(|A|\le|r^k_{xi}|)} \times \iota^{I(|A|<|r^k_{xi}|)}\left(1-\iota\right)^{I(|A|\ge|r^k_{xi}|)}.$$

In this study, we set δ and ι to 0.001. $P(r^k_{xi}[l]|A[l], b^k_{xi}[l])$ models base substitution error on each base and is given by:

$$P\left(r^k_{xi}[l]|A[l], b^k_{xi}[l]\right) = \begin{cases} 1 & r^k_{xi}[l] = A[l] \;\&\; b^k_{xi}[l] = 1 \\ 1/3 & r^k_{xi}[l] \neq A[l] \;\&\; b^k_{xi}[l] = 0. \\ 0 & \text{otherwise} \end{cases}$$

$P_{mis}(r^k_{xi})$ represents the probability of read generation for misaligned reads. We consider that reads representing indels, i.e. reads with 0 length or >1 nt are generated more probably than reads with 1 nt in the misalignment, and design $P_{mis}(r^k_{xi})$ as:

$$P_{mis}\left(r^k_{xi}\right) = \begin{cases} 1/N_{mis} & |r^k_{xi}| = 1 \\ p_{mis}/N_{mis} & \text{otherwise} \end{cases}, \quad p_{mis} \ge 1.$$

Here, N_{mis} is the normalization factor given by $\sum_{A\in\mathcal{A}_{xk}} 1^{I(|A|=1)} p_{mis}^{I(|A|\neq 1)}$, where $\mathcal{A}_{xk}$ is a set of possible alleles for individual x at position k. $\mathcal{A}_{xk}$ is given by {'A', 'T', 'G', 'C'}∪ null string for deletion. In addition, if there exist reads with nucleotides more than one aligned at position k, the corresponding sequences are added, $\mathcal{A}_{xk}$. In this study, we set p_{mis} to 1.0, i.e. we assume that the read is generated from possible alleles equally probably. $P(b^k_{xi}|bq^k_{xi})$ is factorized as $\prod_l P(b^k_{xi}[l]|bq^k_{xi}[l])$, and each term is given by a binomial distribution with parameter $1 - 10^{-bq^k_{xi}[l]/10}$:

$$P\left(b^k_{xi}[l]|bq^k_{xi}[l]\right) = \left[1 - 10^{-bq^k_{xi}[l]/10}\right]^{I(b^k_{xi}[l]=1)} \times \left[10^{-bq^k_{xi}[l]/10}\right]^{I(b^k_{xi}[l]=0)}$$

$P(m^k_{xi})$ is given by a binomial distribution with parameter $p_{m^k_{xi}}$. We also give a beta distribution with parameters $\alpha_m(1 - 10^{-MAPQ_{xi}/10})$ and $\alpha_m 10^{-MAPQ_{xi}/10}$ as prior distributions of $p_{m^k_{xi}}$. Thus, $p_{m^k_{xi}}$ is updated by considering both probability for alignment reliability of read r^k_{xi} from the model and mapping quality score $MAPQ_{xi}$. We set prior strength α_m to 10.

2.2 Pedigree part

Pedigree part considers statistical relationship among the genotype of individuals in a pedigree. Here, we consider a model of a parent–child

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

trio with child c, mother m and father f. Let G_{xk} be a genotype (A^1_{xk}, A^2_{xk}) at position k for individual $x \in \{c, m, f\}$. We also let t_{mk} and t_{fk} be binary variables that take 1 or 2 to indicate the allele transmission to the offspring, e.g. if t_{mk} is 1, A^1_{mk} is transmitted to the offspring. The joint probability of genotypes for child, mother and father is factorized as:

$$P(G_{ck}, G_{fk}, G_{mk}|t_{fk}, t_{mk}) = P(G_{ck}|G_{mk}, G_{fk}, t_{fk}, t_{mk}) \times P(G_{mk})P(G_{fk})$$

$P(G_{ck}|G_{mk}, G_{fk}, t_{fk}, t_{mk})$ represents allele transmission, and we model it as:

$$\begin{aligned} P(G_{ck}|G_{mk}, G_{fk}, t_{fk}, t_{mk}) = {} & P(A^1_{ck}|A^1_{mk}; \varepsilon)^{I(t_{mk}=1)} \\ & \times P(A^1_{ck}|A^2_{mk}; \varepsilon)^{I(t_{mk}=2)} \\ & \times P(A^2_{ck}|A^1_{fk}; \varepsilon)^{I(t_{fk}=1)} \\ & \times P(A^2_{ck}|A^2_{fk}; \varepsilon)^{I(t_{fk}=2)} \end{aligned}$$

The conditional probability $P(A^1_{ck}|A^1_{mk}; \varepsilon)$ is given by:

$$P(A^1_{ck}|A^1_{mk}; \varepsilon) = \begin{cases} 1-\varepsilon & A^1_{ck} = A^1_{mk} \\ \varepsilon & \text{otherwise} \end{cases},$$

where ε is *de novo* mutation rate including both germline and somatic mutation rate. We set ε to 2.5×10^{-7} based on the assumption about *de novo* mutation rate in the genome-wide mutation study of parent–offspring trio sequencing data from the 1000 Genomes Project (Cartwright *et al.*, 2012; Conrad *et al.*, 2011). For considering chromosomal recombination, we give a probability on transition of t_{mk} between positions by:

$$P(t_{mk}|t_{mk-1}) = \tau^{I(t_{mk} \neq t_{mk-1})}(1-\tau)^{I(t_{mk}=t_{mk-1})},$$

where τ is chromosomal recombination rate. We set τ to 10^{-8}, as 1 cM is on average $\sim 10^6$ bp in human genome (Collins *et al.*, 1996). For prior probabilities of founders' genotypes G_{mf} and G_{fk}, we assume the following factorization:

$$P(G_{mk}) = P(A^1_{mk})P(A^2_{mk}) \tag{2}$$

$P(A^1_{mk})$ is a prior distribution for allele frequency, and we use a distribution considering reference allele and an empirical distribution from aligned reads given by:

$$P(A^1_{mk}) = \frac{\beta_e P_{ek}(A^1_{mk}, k) + \beta_r I(A^1_{mk} = A^{ref}_k) + \beta_u/|\mathcal{A}_{mk}|}{\beta_r + \beta_d + \beta_u},$$

where P_{ek} is an empirical distribution for allele from aligned reads to position k and A^{ref}_k is a reference allele at position k. β_e, β_r and β_u are non-negative parameters for adjustment. For alleles in neither reference allele nor aligned reads, uniform distribution $1/|\mathcal{A}_{mk}|$ is considered. We deal with β_e as a position-dependent value and set it to the read depth at each position, whereas β_r is set to the coverage of data. Because 1/1000 homologous positions are different between two chromosomes in empirical rough estimate, we set β_u to $0.001 \times \beta_r$.

2.3 Haplotype selection part

Haplotyping part links allele likelihood part and pedigree part by selecting a haplotype from which read is generated. Given a genotype $G_{xk} = (A^1_{xk}, A^2_{xk})$ at position k for individual x, haplotype selection part selects an allele, from which each read is generated, by using binary variable h^k_{xi} in the following manner:

$$\prod_{i \in \mathcal{I}_{xk}} P(r^k_{xi}|A^1_{xk}, bq^k_{xi})^{I(h^k_{xi}=1)} P(r^k_{xi}|A^2_{xk}, bq^k_{xi})^{I(h^k_{xi}=2)} P(h^k_{xi}), \tag{3}$$

where $\mathcal{I}_{xk}$ is a set of indexes of reads for individual x that contain aligned reads at position k. In a simple setting, the allocation of each read to haplotypes is equally probable and independent of positions, i.e. $P(h^k_{xi} = 1)$ and $P(h^k_{xi} = 2)$ are 0.5. Here, instead of $P(h_{xi})$, we consider a conditional probability $P(h^k_{xi}|z_{xk})$ given by:

$$P(h^k_{xi}|z_{xk}) = \begin{cases} p_{h_{xi}} & h^k_{xi} = 1 \,\&\, z_{xk} = 1 \\ 1 - p_{h_{xi}} & h^k_{xi} = 2 \,\&\, z_{xk} = 1 \\ 0.5 & z_{xk} = 0 \end{cases} \tag{4}$$

where $p_{h_{xi}}$ is a rate for the assignment of read R_{xi} to a haplotype and z_{xk} is a binary variable that determines zygosity at position k for individual x and takes value 1 for heterozygote and 0 for homozygote. For paired-end data, $p_{h_{xi}}$ is shared in each read pair.

To represent zygosity with z_{xk}, we introduce a conditional probability $P(z_{xk}|A^1_{xk}, A^2_{xk})$ that is given by:

$$P(z_{xk}|A^1_{xk}, A^2_{xk}) = \begin{cases} 1.0 & \begin{array}{c} A^1_{xk} = A^2_{xk} \,\&\, z_{xk} = 0 \\ \text{or} \\ A^1_{xk} \neq A^2_{xk} \,\&\, z_{xk} = 1 \end{array} \\ 0 & \text{otherwise} \end{cases}$$

The role of z_{xk} is to filter out highly probably homozygous positions from data for read assignment via Equation (4), and hence only highly probably heterozygous positions are used for haplotyping.

$p_{h_{xi}}$ represents one of the two chromosomes from which read R_{xi} comes. Because each read comes from one of the two homologous chromosomes equally probably in an ideal condition, we represent this property with the following formula:

$$P(\sum_{i \in \mathcal{I}_{xk}} p_{h_{xi}}) = \mathcal{N}(\sum_{i \in \mathcal{I}_{xk}} p_{h_{xi}}; |\mathcal{I}_{xk}|/2, \bar{L}|\mathcal{I}_{xk}|/4)$$

where $\mathcal{N}$ represents normal distribution and $\bar{L}$ is the average read length. $\sum_{i \in \mathcal{I}_{xk}} p_{h_{xi}}$ is considered as the number of reads from the same chromosome at position k. Because $\sum_{i \in \mathcal{I}_{xk}} p_{h_{xi}}$ is a continuous value, normal approximation of a binomial distribution with parameter 0.5 is used. Although the mean and variance of the normal distribution approximating a binomial distribution with parameter 0.5 are $|\mathcal{I}_{xk}|/2$ and $|\mathcal{I}_{xk}|/4$, respectively, we set the variance to $\bar{L}|\mathcal{I}_{xk}|/4$ to normalize the effect with the average read length.

2.4 Parameter estimation and genotype inference

By using allele likelihood part, pedigree part and haplotype selection part, the complete likelihood of our model is given in the following formula:

$$\begin{aligned} & \left[\prod_{x \in \{c,m,f\}} P(p_{h_{x1}}, \ldots, p_{h_{x|\mathcal{I}_x|}})\right] P(t_{m0})P(t_{f0}) \\ & \times \prod_k P(t_{mk}|t_{mk-1})P(t_{fk}|t_{fk-1})P(G_{ck}, G_{mk}, G_{fk}|t_{mk}, t_{fk}) \\ & \times \prod_{x \in \{c,m,f\}} P(z_{xk}|A^1_{xk}, A^2_{xk}) \prod_{i \in \mathcal{I}_{xk}} P(h^k_{xi}|z_{xk}; p_{h_{xi}}) \\ & \times P(p_{m^k_{xi}}|MAQ_{xi}) \prod_{s=1}^{2} P(r^k_{xi}|A^s_{xk}, bq^k_{xi}; p_{m^k_{xi}})^{I(h^k_{xi}=s)}, \end{aligned} \tag{5}$$

where $\mathcal{I}_x$ is a set of indexes of reads for individual x and $P(p_{h_{x1}}, \ldots, p_{h_{x|\mathcal{I}_x|}})$ is the joint probability of $p_{h_{xi}}$ and proportional to $\prod_k P(\sum_{i \in \mathcal{I}_{xk}} p_{h_{xi}})$. In Equation (5), observed variables are r^d_{xd}, bq^k_{xi} and MAQ_{xi}, latent variables are t_{xk}, A^1_{xk}, A^2_{xk}, z_{xk} and h^k_{xi} and parameters are $p_{m^i_{xk}}$ and $p_{h_{xi}}$.

Figure 1 gives a graphical representation of our proposed model, where observed variables are in gray and latent variables and parameters are in white. We use an EM algorithm for parameter estimation, as the model contains many latent variables and the model structure of these variables is complicated. In E-step, the calculation of the marginal probabilities on latent variables such as h^k_{xi} and z_{xk} is required. However, unlike usual hidden Markov models, exact probabilistic inference requires

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

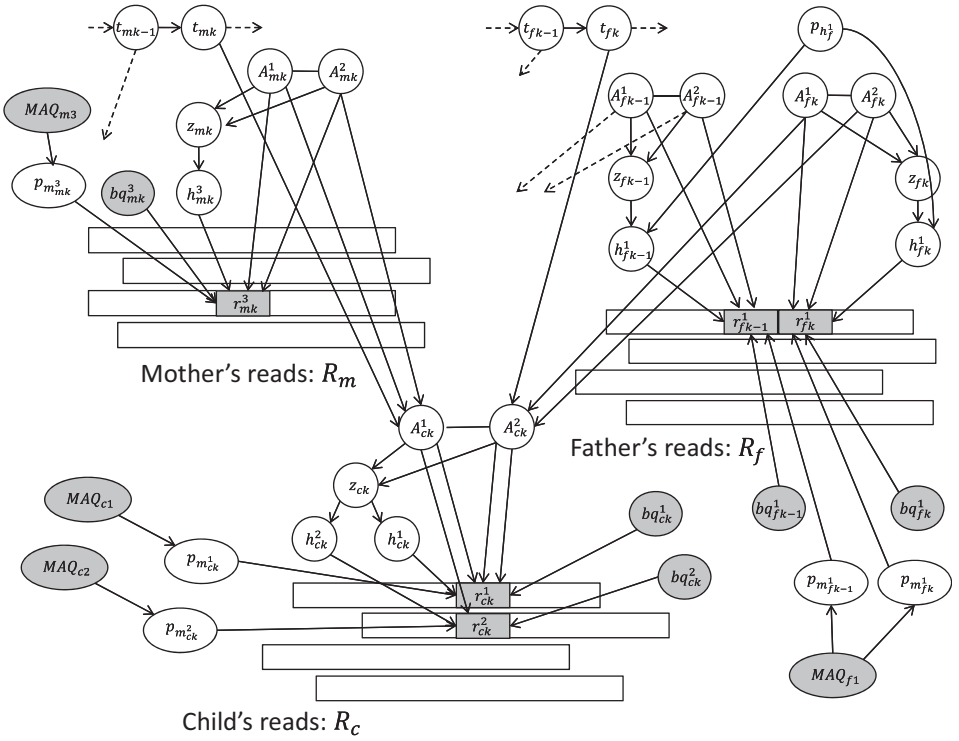

Fig. 1. Agraphical representation of our proposed model, where latent variables and parameters are in white and observed variables are in gray

high computational time when the pedigree size is large and its structure is complicated. To address this type of computational issue, we use loopy belief propagation to calculate the approximated marginal probabilities required in M-step (Murphy *et al.*, 1999; Yedidia *et al.*, 2005). In M-step, $p_{m^k_{xi}}$ and $p_{h_{xi}}$ are updated by using the marginal probabilities calculated in E-step. For details of parameter estimation by the EM algorithm, see Section 1 in the Supplementary Material.

Genotype inference is performed at each position independently. For genotype inference on position k, we search a configuration of latent variables A^1_{ck}, A^2_{ck}, A^1_{mk}, A^2_{mk}, A^1_{fk} and A^2_{fk} that maximizes their marginal probability. We again use loopy belief propagation to calculate the approximated marginal probability of these variables and then apply the max-product algorithm to obtain the configuration maximizing the approximated marginal probability (Weiss and Freeman, 2001).

2.5 Suppression of variant calling at positions next to homopolymer regions and indels

In the base calling on Illumina reads, the position next to a homopolymer region tends to be called as the base comprising the homopolymer because phasing accumulated in synthesis during the Illumina sequencing process affects the base calling result. Thus, if there exists a homopolymer left side of a position in the reference genome, reads from the forward strand tend to have the base comprising the homopolymer at the position, whereas reads from negative strand do not. From this observation, our approach suppresses variant calling at positions satisfying the following three conditions for Illumina reads:

- A position is next to homopolymer whose length is more than two.
- The majority of alternative alleles is the same as the base comprising the homopolymer.
- More than 90% of the aligned reads have the same strand.

In addition, to avoid false positive (FP) variant calling around indels, we use base alignment quality BAQ from SAMtools (Li, 2011) as base quality score.

3 PERFORMANCE EVALUATION

3.1 Simulation data analysis

We evaluated our proposed approach, which hereafter we call PedigreeCaller, by using synthetically generated parent–offspring trio sequence data. We first generated genome sequences of chromosome 21 for two Utah residents with Northern and Western European ancestry (CEU) individuals NA12286 and NA12287 according to the variant calling and phasing results in a VCF file released on November 23, 2010, by the 1000 Genomes Project. We then synthetically generated genome sequences of their child by randomly choosing one of the genome sequences for each individuals and recombining them with recombination rate of 1.0×10^{-8}. The number of variants for these three individuals is 149 947.

From these genome sequences, we generated paired-end sequence reads and put 0.1% bases substitution errors. As base quality scores for the reads, we uniformly gave Q30 to each base, which corresponds to 0.1% error. We generated three types of datasets with the following conditions:

- Read length is 100 bp, and insert size is normally distributed with mean 400 bp and standard deviation 50 bp.
- Read length is 500 bp, and insert size is normally distributed with mean 2000 bp and standard deviation 250 bp.
- Read length is 1000 bp, and insert size is normally distributed with mean 4000 bp and standard deviation 500 bp.

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

The BAM files for these datasets were obtained by mapping the sequence reads to UCSC hg19 reference genome for chromosome 21 with BWA–MEM (Li, 2013). To evaluate the performance of our proposed approach with various read coverage, we downsampled the BAM files with Picard DownsampleSam (http://picard.sourceforge.net/) and obtained BAM files with read coverage of 5, 10, 20 and 40× for each individual. These downsampled BAM files were realigned with Genome Analysis Toolkit(GATK) Indel Realigner (DePristo *et al.*, 2011).

For the comparison with existing approaches, we used the following four methods: GATK Unified Genotyper (DePristo *et al.*, 2011; McKenna *et al.*, 2010), BCFtools with SAMtools mpileup (Li *et al.*, 2009), TrioCaller (Chen *et al.*, 2013)and PolyMutt (Li *et al.*, 2012). In Unified Genotyper and SAMtools, variant calling is performed for each individual independently, i.e. no pedigree information is considered. TrioCaller takes variant calling results from other callers as input data, and re-estimates variants by considering pedigree information and linkage disequilibrium. Note that linkage disequilibrium information is limited in this experiment, as only three individuals are considered. For the input of TrioCaller, we used results from SAMtools as is in the instruction of TrioCaller users' example. PolyMutt takes genotype likelihood information from other variant callers as its input data, and estimates variants by considering pedigree information. As is instructed in PolyMutt web page, we obtained likelihood information from SAMtools in Genotype Likelihood Format (GLF) and used it as the input data of PolyMutt. We used Unified Genotyper and PolyMutt with default options. For SAMtools/BCFtools, we used the commands described in SAMtools Web site without variant filtering with depth (http://samtools.sourceforge.net/mpileup.shtml). For TrioCaller, we set '–round' option to 100 and used default setting for other options.

Table 1 summarizes the performance of variant detection on our proposed approach and these four methods: the numbers of true positives (TPs), FPs, accuracy, recall, precision and F-

Table 1. Comparison on the variant detection performance of PedigreeCaller, PolyMutt, TrioCaller, SAMtools and GATK-Unified Genotyper for simulation datasets with read coverage of 5 and 10× and read length of 100, 500 and 1000 bp

Read coverage	Read length (bp)	Method	Number of TPs	Number of FPs	Recall	Precision	F-measure
5×	100	PedigreeCaller	137 206	654	0.9150	0.9953	0.9535
		PolyMutt	136 383	781	0.9095	0.9943	0.9500
		TrioCaller	135 162	1076	0.9014	0.9921	0.9446
		GATK	104 231	297	0.6951	0.9972	0.8192
		SAMtools	118 661	2202	0.7914	0.9818	0.8763
	500	PedigreeCaller	138 040	504	0.9206	0.9964	0.9570
		PolyMutt	137 506	781	0.9170	0.9944	0.9541
		TrioCaller	136 303	1282	0.9090	0.9907	0.9481
		GATK	104 406	312	0.6963	0.9970	0.8199
		SAMtools	120 514	2292	0.8037	0.9813	0.8837
	1000	PedigreeCaller	137 638	357	0.9179	0.9974	0.9560
		PolyMutt	137 602	810	0.9177	0.9941	0.9544
		TrioCaller	135 863	1524	0.9061	0.9889	0.9457
		GATK	104 406	316	0.6963	0.9970	0.8199
		SAMtools	120 751	2314	0.8053	0.9812	0.8846
10×	100	PedigreeCaller	146 738	291	0.9786	0.9980	0.9882
		PolyMutt	146 122	235	0.9745	0.9984	0.9863
		TrioCaller	145 718	441	0.9718	0.9970	0.9842
		GATK	137 984	401	0.9202	0.9971	0.9571
		SAMtools	141 308	406	0.9424	0.9971	0.9690
	500	PedigreeCaller	147 207	180	0.9817	0.9988	0.9902
		PolyMutt	146 590	207	0.9776	0.9986	0.9880
		TrioCaller	146 089	586	0.9743	0.9960	0.9850
		GATK	138 087	387	0.9209	0.9972	0.9575
		SAMtools	142 326	422	0.9492	0.9970	0.9725
	1000	PedigreeCaller	147 349	133	0.9827	0.9991	0.9908
		PolyMutt	146 734	198	0.9786	0.9987	0.9885
		TrioCaller	146 511	251	0.9771	0.9983	0.9876
		GATK	138 580	383	0.9242	0.9972	0.9593
		SAMtools	142 854	373	0.9527	0.9974	0.9745

Note: The best result on each condition is underlined.

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

measure for the sequencing data with read coverage of 5 and 10×. Results for 20 and 40× are summarized in Section 2 of the Supplementary Material. In variant detection, if a genotype estimated at a position and true genotype at the position are not homozygous for the reference allele, the estimated genotype is counted as TP. On the other hand, if true genotype is homozygous for the reference allele, the estimated genotype is counted as FP. Note that TPs could contain estimated genotypes that are different from their corresponding true genotypes, e.g. if an estimated genotype and true genotype at a position with reference allele A are (A, B) and (B, B), respectively, the estimated genotype is counted as TP. Recall and precision are given by $\frac{\#\text{ of TPs}}{\#\text{ of trues}}$ and $\frac{\#\text{ of TPs}}{\#\text{ of TPs}+\#\text{ of FPs}}$, respectively. F-measure is given by the harmonic mean of recall and precision as $2\times\frac{\text{recall}\times\text{precision}}{\text{recall}+\text{precision}}$, and achieves overall performance by capturing the trade-off of recall and precision. These measures are valued between 0 and 1, and larger value is better.

In all scenarios in Table 1, PedigreeCaller outperforms other methods in recall and F-measure. For the datasets with read coverage of 10× and read length of 100 bp, PolyMutt outperforms other methods including PedigreeCaller in precision. Also, for the datasets with read coverage of 5× and read length of 100 and 500 bp, GATK outperforms other methods including PedigreeCaller in precision. However, recalls of GATK in these conditions are low and hence its F-measures are worse than those of other methods. In the datasets, except for the conditions with 5× and 100 and 500 bp, precisions of PedigreeCaller are higher than those of other methods.

To see the effect of read length to PedigreeCaller, we focus on the difference on the performance between PedigreeCaller and PolyMutt. For dataset with read coverage of 10× and read length of 100 bp, the number of TPs on PedigreeCaller is 616 more than that on PolyMutt, but the number of FPs on PedigreeCaller is 46 more than that on PolyMutt. On the other hand, for dataset with read coverage of 10× and read length of 500 bp, the number of TPs on PedigreeCaller is 617 more than that on PolyMutt and the number of FPs on PedigreeCaller is 27 less than that on PolyMutt. In addition, in the dataset with read length of 1000 bp, the number of TPs on PedigreeCaller is 615 more than that on PolyMutt and the number of FPs on PedigreeCaller is 65 less than that on PolyMutt.

Table 2 summarizes the performance of genotype concordance on our proposed approach and four existing methods for the sequencing data with read coverages of 5 and 10×. Results for 20 and 40× are summarized in Section 2 of the Supplementary Material. In genotype concordance, if a genotype estimated at a position and true genotype at the position are the same and not homozygous for the reference allele, the estimated genotype is counted as TP. On the other hand, if the genotype of the estimated variant is different from the true genotype, the estimated variant is counted as FP. Note that the recall and precision for genotype concordance are respectively the same as sensitivity and PPV used for evaluation of variant callers in You *et al.* (2012). Also, note that the number of trues is not equal to the sum of the number of TPs and the number of false negatives (FNs) because some trues can be in FPs.

Similar to variant detection, all the conditions in Table 2, PedigreeCaller outperforms other methods in recall and F-measure.

3.2 Real data analysis

To evaluate performance on real sequencing data, we use 100-bp paired-end sequencing data of HapMap CEU parent–offspring trio comprising NA12878 (child), NA12891 (father) and NA12892 (mother) (Conrad *et al.*, 2011). The data were sequenced with Illumina HiSeq 2000 with read coverage of 45× for each individual, and stored as BAM files after mapping to UCSC hg19 with Burrows–Wheeler Aligner (Li and Durbin, 2009). The average insert size is ∼300 bp.

We downsampled the sequencing datasets to 5, 10, 20 and 40× for each individual with Picard DownsampleSam. These downsampled BAM files were realigned with GATK Indel Realigner, and base quality scores are recalibrated with GATK Base Quality Score Recalibration. We evaluate the performance by assessing the concordance of estimated variants from these variant callers for datasets with various read coverages with SNP array genotyping results from Illumina OMNI 2.5 BeadChip. The number of SNP sites considered is 2 310 349, and the number of variants for these three individuals is 2 122 147.

We also assessed the concordance with the 1000 Genomes Project and commonly estimated variants by the five callers for the dataset of 40×. Results for those cases are given in the Supplementary Material.

Table 3 summarizes the performance of variant detection on PedigreeCaller and four existing methods for NA12878, NA12 891 and NA12 892 from the real datasets with read coverages of 5, 10, 20 and 40×.

PedigreeCaller outperforms other four variant callers in F-measure for all the read coverages. Also, PedigreeCaller achieves the best performance on the number of TPs and recall rate except for the dataset of 40×. Although Unified Genotyper produces better recall rate than that of PedigreeCaller in the 40× data, the result of Unified Genotyper contains four times as many FPs than PedigreeCaller, and hence its F-measure is worse than that of PedigreeCaller. Although recall of SAMtools is low for all the datasets, the results of SAMtools contain the least FPs for all the datasets. In precision, SAMtools outperforms other variant callers including PedigreeCaller except for the datasets with read coverages of 5, 10, 20×, and gives the same performance as PedigreeCaller for the dataset with read coverage of 40×. The performance gaps between PedigreeCaller and other variant callers are larger in lower coverage data, e.g. the maximum gap on F-measure in the 40× data is 0.0025, whereas that in the 5× is >0.1. Because read data is insufficient for accurate variant calling in low coverage data such as the 5× data, the effects from pedigree information and haplotyping are high, and the larger performance gaps can be obtained in lower coverage data. This is also observed in the results between TrioCaller and SAMtools. The results of TrioCaller contain more TPs and less FPs than those of SAMtools in the 5 and 10× data, whereas SAMtools achieves better performance in the numbers of TPs and FPs than TrioCaller in the 20× data.

Table 4 summarizes the performance of genotype concordance on PedigreeCaller and four existing methods for NA12878, NA12891 and NA12892: the numbers of TPs, FPs, true negatives (TNs), FNs, accuracy, recall, precision and F-measure for the sequencing data with read coverages of 5, 10, 20 and 40×. Accuracy is given by $\frac{\#\text{ of TPs}+\#\text{ of TNs}}{\#\text{ of TPs}+\#\text{ of FPs}+\#\text{ of TNs}+\#\text{ of FNs}}$.

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

Table 2. Comparisonon genotype concordance of PedigreeCaller, PolyMutt, TrioCaller, SAMtools, and GATK Unified Genotyper for simulation datasets with read coverages of 5 and 10× and read length 100, 500, and 1000 bp

Read coverage	Read length (bp)	Method	Number of TPs	Number of FPs	Recall	Precision	F-measure
5×	100	PedigreeCaller	132 465	5395	0.8834	0.9609	0.9205
		PolyMutt	131 606	5558	0.8777	0.9595	0.9168
		TrioCaller	128 073	8165	0.8541	0.9401	0.8950
		GATK	98 791	5737	0.6588	0.9451	0.7764
		SAMtools	112 068	8795	0.7474	0.9272	0.8277
	500	PedigreeCaller	133 741	4803	0.8919	0.9653	0.9272
		PolyMutt	132 892	5395	0.8863	0.9610	0.9221
		TrioCaller	129 307	8278	0.8624	0.9398	0.8994
		GATK	99 054	5664	0.6606	0.9459	0.7779
		SAMtools	114 023	8783	0.7604	0.9285	0.8361
	1000	PedigreeCaller	133 456	4539	0.8900	0.9671	0.9270
		PolyMutt	133 068	5344	0.8874	0.9614	0.9229
		TrioCaller	129 140	8247	0.8612	0.9400	0.8989
		GATK	99 135	5587	0.6611	0.9466	0.7785
		SAMtools	114 225	8840	0.7618	0.9282	0.8368
10×	100	PedigreeCaller	146 103	926	0.9744	0.9937	0.9839
		PolyMutt	145 479	878	0.9702	0.9940	0.9820
		TrioCaller	144 059	2100	0.9607	0.9856	0.9730
		GATK	137 114	1271	0.9144	0.9908	0.9511
		SAMtools	140 439	1275	0.9366	0.9910	0.9630
	500	PedigreeCaller	146 746	641	0.9787	0.9957	0.9871
		PolyMutt	146 038	759	0.9739	0.9948	0.9843
		TrioCaller	144 523	2152	0.9638	0.9853	0.9745
		GATK	137 283	1191	0.9155	0.9914	0.9520
		SAMtools	141 510	1238	0.9437	0.9913	0.9669
	1000	PedigreeCaller	146 971	511	0.9802	0.9965	0.9883
		PolyMutt	146 215	717	0.9751	0.9951	0.9850
		TrioCaller	144 984	1778	0.9669	0.9879	0.9773
		GATK	137 852	1111	0.9193	0.9920	0.9543
		SAMtools	142 086	1141	0.9476	0.9920	0.9693

Note: The best result on each condition is underlined.

Similar to the case of variant detection, PedigreeCaller outperforms these three variant callers in accuracy and F-measure on genotype concordance for all the coverages. The performance gaps between PedigreeCaller and other variant callers are larger in lower coverage data, e.g. the maximum gaps on F-measure and accuracy for the datasets of 40× are 0.002 and 0.001, respectively, whereas gaps on F-measure and accuracy for the dataset of 5× are >0.1 and 0.05, respectively. Because read data is insufficient for accurate variant calling in low coverage data such as the 5× data, the effects from pedigree information and haplotyping are high, and the larger performance gaps can be obtained in lower coverage data. This is also observed in the results between TrioCaller and SAMtools. The results of TrioCaller contain more TPs and less FPs than those of SAMtools in the 5 and 10× data, whereas SAMtools achieves better performance in the numbers of TPs and FPs than TrioCaller in the 20× data.

Figure 2 shows relationship between the physical distance of variants and the performance gap of PedigreeCaller and other existing variant callers in the 5× data. The x-axis indicates the distance between positions in base pairs. The y-axis indicates the F-measure difference given by the F-measure of PedigreeCaller subtracted by that of other variant caller for the SNP positions from which there exists a position with heterozygous genotype call within a distance indicated by the x-axis. Here, we restrict the position with heterozygous genotype call to the position where at least two heterozygous genotypes are estimated by PedigreeCaller among the trio members. As shown in Figure 2, PedigreeCaller shows the strong performance for the SNP positions distant from other variants within 100 bp. Because the read length in our data is 100 bp, this result implies that phase-informative reads effectively work for haplotyping and consequently improve the performance of variant calling. Also, we can observe a slow decline of the performance gap for the SNPs >100 bp away from other variants. PedigreeCaller uses paired-end reads for phase-informative reads as well, and less paired-end reads are available as phase-informative reads for more distant positions. Therefore, this slow decline of the performance gap is due to the decrease of the haplotyping effect by paired-end reads.

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

Table 3. Comparison on variant detection performance of PedigreeCaller, PolyMutt, TrioCaller, SAMtools, and GATK based on SNP array genotyping results for real datasets

Read coverage	Method	Number of TPs	Number of FPs	Recall	Precision	F-measure
5×	PedigreeCaller	1867 118	8247	0.8798	0.9956	0.9341
	PolyMutt	1818 625	11 437	0.8570	0.9938	0.9203
	TrioCaller	1788 512	21 692	0.8428	0.9880	0.9096
	GATK	1524 138	3021	0.7182	0.9980	0.8353
	SAMtools	1516 241	1467	0.7145	0.9990	0.8331
10×	PedigreeCaller	2014 055	4259	0.9491	0.9979	0.9729
	PolyMutt	1987 388	4365	0.9365	0.9978	0.9662
	TrioCaller	1975 516	7627	0.9309	0.9962	0.9624
	GATK	1900 812	6024	0.8957	0.9968	0.9436
	SAMtools	1875 413	2288	0.8837	0.9988	0.9377
20×	PedigreeCaller	2043 875	3373	0.9631	0.9984	0.9804
	PolyMutt	2036 495	3899	0.9596	0.9981	0.9785
	TrioCaller	2031 028	4593	0.9571	0.9977	0.9770
	GATK	2039 673	9859	0.9611	0.9952	0.9779
	SAMtools	2014 731	2879	0.9494	0.9986	0.9734
40×	PedigreeCaller	2045 752	3035	0.9640	0.9985	0.9810
	PolyMutt	2045 021	3390	0.9637	0.9983	0.9807
	TrioCaller	2038 818	4205	0.9607	0.9979	0.9790
	GATK	2052 441	14 153	0.9672	0.9932	0.9800
	SAMtools	2035 897	3008	0.9594	0.9985	0.9785

Note: The best result on each condition is underlined.

Table 4. Comparison on genotype concordance of PedigreeCaller, PolyMutt, TrioCaller, SAMtools and GATK based on SNP array genotyping results for real datasets

Read coverage	Method	Number of TPs	Number of FPs	Number of TNs	Number of FNs	Recall	Precision	F-measure	Accuracy
5×	PedigreeCaller	1 795 123	80 242	4 800 654	255 029	0.846	0.957	0.898	0.952
	PolyMutt	1 747 830	82 232	4 797 464	303 522	0.824	0.955	0.884	0.944
	TrioCaller	1 683 205	126 999	4 787 209	333 635	0.793	0.930	0.856	0.934
	GATK	1 440 989	86 170	4 805 880	598 009	0.679	0.944	0.790	0.901
	SAMtools	1 428 057	89 651	4 807 434	605 906	0.673	0.941	0.785	0.900
10×	PedigreeCaller	2 001 903	16 411	4 804 642	108 092	0.943	0.992	0.967	0.982
	PolyMutt	1 976 330	15 423	4 804 536	134 759	0.931	0.992	0.961	0.978
	TrioCaller	1 953 026	30 117	4 801 274	146 631	0.920	0.985	0.951	0.974
	GATK	1 884 099	22 737	4 802 877	221 335	0.888	0.988	0.935	0.965
	SAMtools	1 859 807	17 894	4 806 613	246 734	0.876	0.990	0.930	0.962
20×	PedigreeCaller	2 040 441	6807	4 805 528	78 272	0.961	0.997	0.979	0.988
	PolyMutt	2 033 754	6640	4 805 002	85 652	0.958	0.997	0.977	0.987
	TrioCaller	2 022 855	12 766	4 804 308	91 119	0.953	0.994	0.973	0.985
	GATK	2 035 619	13 913	4 799 042	82 474	0.959	0.993	0.976	0.986
	SAMtools	2 012 134	5476	4 806 022	107 416	0.948	0.997	0.972	0.984
40×	PedigreeCaller	2 043 075	5712	4 805 866	76 395	0.963	0.997	0.980	0.988
	PolyMutt	2 042 700	5711	4 805 511	77 126	0.963	0.997	0.980	0.988
	TrioCaller	2 031 073	11 950	4 804 696	83 329	0.957	0.994	0.975	0.986
	GATK	2 049 088	17 506	4 794 748	69 706	0.966	0.992	0.978	0.987
	SAMtools	2 033 955	4950	4 805 893	86 250	0.958	0.998	0.978	0.987

Note: The best result on each condition is underlined.

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

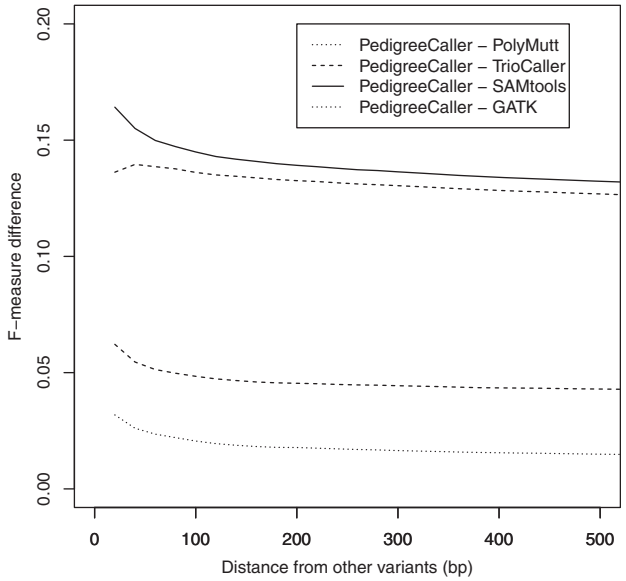

Fig. 2. A plot representing performance gaps between PedigreeCaller and other existing variant callers with respect to distance between target SNP positions and other closest variants in the 5× data. The x-axis indicates distance in base pairs between target SNP positions and other variants. The y-axis indicates the performance gap between PedigreeCaller and other variant callers in F-measure

4 DISCUSSION AND CONCLUSION

We proposed a new statistical variant calling approach that considers pedigree information and haplotyping based on phase-informative reads. In a naïve way, modeling of haplotyping based on phase-informative reads may fail to estimate accurate results owing to undesirable influence from homozygous positions in the estimation of the assignment of each read to a haplotype. To address this issue, we introduced latent variables that determine zygosity, and avoided the undesirable influence from homozygous positions by using the estimated zygosity in the latent variables.

Through variant calling in the practical NGS data and the comparison of the estimated variants with the SNP array genotyping results, we showed that our approach outperforms existing variant callers, including variant callers that consider pedigree information. Also, the analysis of the performance on the positions located close to other heterozygous variants showed that our approach is more powerful in such positions than other variant callers, especially for SNPs distant from other variants within 100 bp. Because read length in the data used for the evaluation is 100 bp, the result of the analysis implies that haplotyping based on phase-informative reads with these variants effectively works for the improvement of the performance in our approach. We also observed the performance improvement by the effect of haplotyping based on paired-end reads through the slow decline of the performance gap between our approach and other variant callers on the SNP positions distant >100 bp from other variant positions.

Owing to the rapid progress of NGS technologies, read length is extending in the current sequencing platforms. In addition, systematically new technologies such as the Nanopore technology (http://www.nanoporetech.com/) and the Moleculo technology (http://www.moleculo.com/) are now under development, and these technologies are considered to enable the production of more accurate and longer sequencing data. Therefore, further accurate variant calling is expected in our approach by using longer reads available in the near future.

ACKNOWLEDGEMENTS

The authors thank the anonymous referees for their constructive suggestions and comments, which improved the quality of this paper. Illumina HiSeq 2000 sequencing data and OMNI 2.5 SNP array genotyping results for the CEU parent-offspring trio were kindly provided by Illumina, Inc. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

Funding: This work was supported (in part) by MEXT Tohoku Medical Megabank Project.

Conflict of Interest: none declared.

REFERENCES

Cartwright,R.A. *et al.* (2012) A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. *Stat. Appl. Genet. Mol. Biol.*, **11**.

Chen,W. *et al.* (2013) Genotype calling and haplotyping in parent-offspring trios. *Genome Res.*, **23**, 142–151.

Collins,A. *et al.* (1996) A metric map of humans: 23,500 loci in 850 bands. *Proc. Natl Acad. Sci. USA*, **93**, 14771–14775.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

He,D. *et al.* (2012) Hap-seq: an optimal algorithm for haplotype phasing with imputation using sequencing data. *Lect. Notes Comput. Sci.*, **7262**, 64–78.

Li,H. (2011) Improving SNP discovery by base alignment quality. *Bioinformatics*, **27**, 1157–1158.

Li,B. *et al.* (2012) A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, **8**, e1002944.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Li,H. and Durbin,R. (2009) Fast and accurate short-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al.* (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Menelaou,A. and Marchini,J. (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, **29**, 84–91.

Murphy,K.P. *et al.* (1999) Loopy belief propagation for approximate inference: an empirical study. In: *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden, pp. 467–475.

Quail,M.A. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.

Weiss,Y. and Freeman,W.T. (2001) On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Trans. Inf. Theory*, **47**, 736–744.

Yedidia,J.S. *et al.* (2005) Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory*, **51**, 2282–2312.

You,N. *et al.* (2012) SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, **28**, 643–650.

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