BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 8 2008, pages 1035–1040
doi:10.1093/bioinformatics/btn074

Genome analysis

Consensus generation and variant detection by Celera Assembler

Gennady Denisov*, Brian Walenz, Aaron L. Halpern, Jason Miller, Nelson Axelrod, Samuel Levy and Granger Sutton

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received on November 3, 2007; revised on January 31, 2008; accepted on February 22, 2008

Advance Access publication March 4, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: We present an algorithm to identify allelic variation given a Whole Genome Shotgun (WGS) assembly of haploid sequences, and to produce a set of haploid consensus sequences rather than a single consensus sequence. Existing WGS assemblers take a column-by-column approach to consensus generation, and produce a single consensus sequence which can be inconsistent with the underlying haploid alleles, and inconsistent with any of the aligned sequence reads. Our new algorithm uses a dynamic windowing approach. It detects alleles by simultaneously processing the portions of aligned reads spanning a region of sequence variation, assigns reads to their respective alleles, phases adjacent variant alleles and generates a consensus sequence corresponding to each confirmed allele. This algorithm was used to produce the first diploid genome sequence of an individual human. It can also be applied to assemblies of multiple diploid individuals and hybrid assemblies of multiple haploid organisms.

Results: Being applied to the individual human genome assembly, the new algorithm detects exactly two confirmed alleles and reports two consensus sequences in 98.98% of the total number 2 033 311 detected regions of sequence variation. In 33 269 out of 460 373 detected regions of size >1 bp, it fixes the constructed errors of a mosaic haploid representation of a diploid locus as produced by the original Celera Assembler consensus algorithm. Using an optimized procedure calibrated against 1 506 344 known SNPs, it detects 438 814 new heterozygous SNPs with false positive rate 12%.

Availability: The open source code is available at: http://wgs-assembler.cvs.sourceforge.net/wgs-assembler/

Contact: gdenisov@jcvi.org

1 INTRODUCTION

Study of single nucleotide polymorphisms (SNPs), insertion/deletion events (indels) and other genetic variations has a number of practical applications, including the identification and treatment of genetically inherited diseases. In the classical model of heredity, different versions of coding sequences, or alleles, may impart a particular phenotype. Mendelian genetic diseases are due to the deleterious effect of an allelic variant (McKusick, 1998), while complex diseases are defined by the combined effect of multiple variant alleles (International HapMap Consortium, 2005). Variants can be used to infer haplotype, or a particular combination of alleles across a chromosome (Clark, 1990; Indap *et al*., 2005; Kim *et al*., 2007a; Lippert *et al*., 2002). Haplotypes have even more power than individual variants in the context of association studies and predicting disease risk (Daly *et al*., 2001; Stephens *et al*., 2001). Reliably determining the variant alleles at a given loci, and linking these variant alleles into haplotype blocks is important for biomedical research.

Whole genome shotgun (WGS) assembly is one of the high-throughput approaches which can be applied to analysis of genetic variations. A number of genome assembly programs have been described in the literature up to date, including TIGR Assembler (Sutton *et al*., 1995), PHRAP (Green, 2005), CAP (Huang and Madan, 1999; Huang *et al*., 2003), the original version of Celera Assembler (Myers *et al*., 2000), GigAssembler (Kent and Haussler, 2001), Euler (Pevzner *et al*., 2001), JAZZ (Aparicio *et al*., 2002), Arachne (Batzoglou *et al*., 2002; Jaffe *et al*., 2003), RePS (Wang *et al*., 2002), Phusion (Mullikin and Ning, 2003) and Atlas (Havlak *et al*., 2004). These programs can, in principle, be used to detect candidate variant positions. For example, Celera Genomics detected candidate SNPs in the human genome through a WGS assembly of 27 million reads representing five different individuals (Istrail *et al*., 2004; Venter *et al*., 2001). However, none of the programs mentioned above is aimed at accurate detection of alleles and variants. A single consensus call is produced for each column in a multiple alignment, as determined by applying the majority rule (in TIGR Assembler, Arachne and the original version of Celera Assembler), selecting the read base of the highest quality (in PHRAP, JAZZ, RePS and Phusion) or using a weighted sums of quality values in reads of opposite orientation (in CAP and PCAP). When reads represented more than one allele and a balanced number of reads from each allele were present, the consensus sequence could alternate between different alleles, thus producing a 'mosaic variant' that matches neither allele (Fig. 1).

On the other hand, numerous post-processing approaches described in the literature can be used to identify variants from an existing assembly or multiple sequence alignment of reads. These include comparative whole genome mapping (Waterston

*To whom correspondence should be addressed.

```
CCACCA---TGTGTGTGTGTGTGTGTGT > Contig 1096708082468

CCACCA----------------GTGTGTGT < Read 1094846374870
||||||||||||||||||||||||||||||
CCACCA----------------GTGTGTGT < Read 1089083146323
||||||                ||||||||
CCACCAGTGTGTGTGTGTGTGTGTGTGT < Read 1089057878307
||||||||||||||||||||||||||||||
CCACCAGTGTGTGTGTGTGTGTGTGTGT < Read 1089053585558
||||||||||||||||||||||||||||||
CCACCAGTGTGTGTGTGTGTGTGTGTGT < Read 1094851158703
||||||                ||||||||
CCACCA----------------GTGTGTGT < Read 1095846055349
```

Fig. 1. An example of multiple alignment of reads from the human genome assembly where a 'mosaic' consensus sequence is produced using the column-by-column approach of the original version of Celera Assembler. The two alleles are 'balanced' in the number of reads. A base (or gap) with maximum sum of quality values across a column is selected to represent a consensus base (the quality values are not shown). The resulting consensus sequence is inconsistent with any of underlying reads.

et al., 2002; Levy *et al*., 2007), restricted representation shotgun (Altshuler *et al*., 2000), Polyphred (Nickerson *et al*., 1997), TRACE_DIFF (Bonfield *et al*., 1998), PolyBayes (Marth *et al*., 1999), AutoSNP (Barker *et al*., 2003), PolyFreq (Wang and Huang, 2005), SEAN (Huntley *et al*., 2006), and others (Jones *et al*., 2004; Kim *et al*., 2007a). Not all of these approaches can be conveniently used with Celera Assembler or most of other assembly programs. Some of the approaches require too detailed trace information which is not available; others focus on detection of SNPs and not applicable to detection of other variants; or may take too long to proceed on large genomes. Most importantly, these post-processing methods do not solve the fundamental problem that WGS assembly ought to produce a set of consensus alleles that are representative of their underlying haploid sequences.

The approach described in this article is an attempt to combine WGS assembly with accurate detection of alleles and variants, and make the first steps toward connecting variants to form increasingly larger haplotype blocks. Our processing goes beyond the standard column-by-column approach taken by the original version of Celera Assembler and most of other programs. It explicitly splits read segments in variant regions into their representative alleles and calls the corresponding number of consensus alleles, rather than a single consensus sequence. It employs a dynamic windowing approach to group closely spaced variants into larger variant regions, where entire portions of reads are processed simultaneously. This guarantees that the consensus allele produced for any given region of the assembly do not contain a mixture of alleles, and the biologically misleading consensus sequence depicted in Figure 1 will not be produced. Finally, it attempts to phase alleles between adjacent variant regions by sorting the alleles so that allele of the same order number in all the regions will represent the same haplotype.

Like AutoSNP and SEAN, we use redundancy as a simple criterion to discriminate polymorphism from sequencing error at the variant detection step. However, we reinforce this measure of confidence by imposing certain constraints on quality values of bases in the minority allele. We show that these constraints can be calibrated against a set of already known true variants in order to minimize the false variant detection rate, is one of the specific goals of this study. Inference of haplotype blocks is a challenging, NP-hard problem (Lancia *et al*., 2001), which needs to be addressed at both the local and non-local level (Kim *et al*., 2007a). In local haplotyping, variants are connected by reads to form relatively small haplotype blocks. At the non-local, or mate-pair level, haplotype blocks containing mated reads of the same fragment are bridged to form long-range connections. In this study, we consider only the local connection of variants. Long-range haplotype assembly is one of our future goals, but it is currently is not a part of the Celera Assembler processing. To some extent, this problem has been addressed by a greedy haplotype assembly approach (Levy *et al*., 2007), which was employed as a post-processing step based on the variant detection and consensus generation algorithm described here.

2 ALGORITHM

This consensus generation procedure is intended to be executed as the last step in a whole genome sequencing pipeline (Myers *et al*., 2000), and is implemented as such in the Celera Assembler. Our processing applies to a multiple alignment of reads and follows these steps: (i) determines a preliminary, or reference consensus sequence, (ii) identifies the columns of sequence variation using this preliminary consensus, (iii) groups closely spaced columns of sequence variation into regions representing variant alleles, (iv) identifies the alleles in the regions of sequence variation, (v) for each region, determines the weight of each allele, selects the major consensus allele, and outputs the variant alleles as a 'VAR' record and (vi) phases confirmed alleles between adjacent regions of sequence variation, so that their consensus sequences will represent the same haplotype.

Step 1. It employs the column-by-column approach and the majority rule of the original version of Celera Assembler. It sets the preliminary consensus to the base with the maximum sum of quality values in a column. This consensus sequence will be regarded as final outside of regions of variation (which are defined below).

Step 2. The following criteria are used to detect a column of sequence variation in the multiple alignment of reads:

- at least two identically varied bases (or gaps) are present in the column, which are different from the preliminary consensus call and
- the average quality value of the varied bases is at least 21 in the case of SNP type of variation, where both the bases and the preliminary consensus call are not gaps; or the sum of the two highest quality values of the varied bases is at least 60 in the case of indel type of variation, where either the varied base or the preliminary consensus call is a gap.

A justification for using the specified cutoff quality values will be given in RESULTS AND DISCUSSION. Quality values (QVs) of called bases were determined using TraceTuner[1] (Denisov *et al*., 2004). A QV of a gap was defined as the minimal of QVs of its flanking bases.

[1]http://tracetuner.cvs.sourceforge.net/tracetuner/

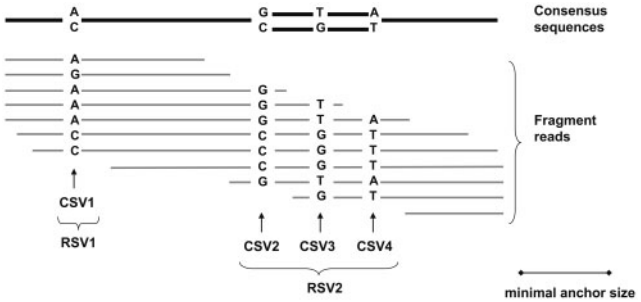

Fig. 2. Schematic illuatration of the main steps of processing. Abbreviations: CSV, column of sequence variation; RSV, region of sequence variation. Bases in columns without sequence variation are not shown. CSV1 forms a separate RSV1 of width 1 bp, because it is spaced from all the other columns of sequence variation by more than MAS bases. Closely spaced columns CSV2, CVS3 and CSV4 are grouped into a single RSV2. Confirmed alleles in RSV1 are represented by read segments 'A' and 'C', and in RSV2 by segments 'C...G...T' and 'G...T...A'. The major consensus sequence for RSV1 is 'A', because this allele has a higher weight than 'C' allele. The major consensus sequence for RSV2 is not 'C...G...T', as would be expected from the highest weight of this allele in RSV2, but 'G...T...A', because this is dictated by phasing of alleles in RSV2 with alleles of the previously processed RSV1.

Step 3. Regions of sequence variation are the regions in between at least MAS non-variant columns, where parameter MAS (=11 by default) stands for minimal anchor size (Fig. 2). By grouping the columns of variation into regions, we consider these variable columns as part of a particular allele.

Step 4. It involves splitting of read segments between appropriate number of alleles. In this context, a 'read segment' refers to the portion of a fragment read spanning the region of sequence variation.

Allele is defined as a group of read segments all having the same sequences of bases (except gaps). If an allele includes more than one read, it is termed 'confirmed'; otherwise the allele is unconfirmed.

Step 5. Each allele is assigned a weight defined as the sum of average quality values for the spanning segments of its reads. Alleles are reverse sorted with respect to their weights, and the first (highest weighted and, normally, confirmed) allele is used for the major consensus sequence of a region. Sequences of alternate confirmed alleles are stored in an XML-formatted variation, or 'VAR record', together with other characteristics of the region of sequence variation, such as its size and location in the corresponding contig, the total number of aligned reads, the weight of each confirmed allele, the number and ids of the reads in each confirmed allele, and other relevant metadata.

Step 6. All the regions of sequence variation are re-processed in increasing order of their locations, starting from the second leftmost region of each contig. For each region, an attempt is made to phase its confirmed alleles with confirmed alleles of the previously processed region. Phasing of alleles is performed by first determining a 1 : 1 mapping of all confirmed alleles of previously processed region on confirmed alleles of the currently processed region, and then sorting alleles of the current region according to this mapping. A prototype allele is considered to be mapped on image allele if the two alleles share most of their reads. If the attempt to phase the confirmed alleles of the two regions is successful, then the major consensus sequence of the currently processed region will be reset to the sequence of its first sorted allele, which may no longer be the allele with the maximum weight, and the VAR record of this region will be updated accordingly.

3 RESULTS AND DISCUSSION

Figure 2 schematically illustrates the main steps of our processing, which include detecting columns of sequence variation in the multiple alignment of reads, grouping closely spaced columns into regions, splitting of read segment between alleles in the regions and, finally, connecting adjacent regions of variation to produce small haplotype blocks. Two regions of sequence variation are shown. The first region is of size 1 and corresponds to a SNP. The second region represents a variant of size >1. It is formed by grouping three columns of sequence variation separated by two narrow (less than MAS columns each) regions without sequence variation. In either region, two confirmed alleles are detected. Alleles of the first regions are sorted in the reverse order of their weights: 'A', 'C' and 'G'. Alleles of the second region are sorted in the order 'G...T...A', 'C...G...T', ..., 'G...-...-', which is dictated by their phasing with alleles of the previously processed first region. In either region, major consensus sequence is set to the sequence of the first sorted allele.

The processing described above facilitated identification and comparison of alternate alleles within individual human genome (Levy *et al*., 2007). It is important to mention that the allele processing approach presented here does not make any preliminary assumptions about the number of alleles present in the input data. This number is determined directly from the consensus process. In particular, in the case of a genome of an individual human, exactly two confirmed alleles have been detected in as many as 98.9% of regions of sequence variation (Table 1), thus confirming the validity of the approach we use. (Table 1). The total number of such regions is 2 051 331, a huge reduction compared to 67 682 594 columns containing at least one variant base or gap in a multiple alignment of ~32 million fragments, and is far more consistent with previous estimates of variant frequency in the human genome. In the genome assembly of yet another diploid organism, the palm oil tree (*Elaeis guineensis*, estimated genome size 1.8 Gb, read coverage 4.3X), our algorithm detected exactly two confirmed alleles in 96.2% of total 858 258 regions of variation.

Because our algorithm does not impose any upper limit on the number of processed haploid alleles, it is applicable to multiploid assemblies as well, particularly to genome assembly of multiple diploid individuals. Furthermore, it could be used by metagenomics projects to generate assemblies of complex environmental samples comprising multiple prokaryotic strains (Tringe and Rubin, 2005; Venter *et al*., 2004; Yooseph *et al*., 2007), assuming the read coverage is sufficient for detecting confirmed alleles. Indeed, such strains are viewed as analogous to eukaryotic haplotypes (Chen and Pachter, 2005).

Further, by grouping individual columns of variation into regions, our processing allowed detection of variants of size >1.

Table 1. Statistics of regions of sequence variation detected in the human genome assembly (genome size ~3.1 Gb) read coverage 7.5X and MAS = 11. Confirmed alleles normally consist of reads without base calling errors, since it is unlikely that exactly the same random sequencing error will occur in two separate experiments at exactly the same position. Regions with more than two confirmed alleles in a diploid genome assembly represent either misassembly (e.g. collapsed repeats) or groups of reads with systematic base calling error, rather than the genetic variation. On the other hand, regions of variation with only one or no confirmed alleles are normally spanned by low-quality reads (with average quality value < 20) possessing base calling errors

Statistic	Value
Regions with < 2 confirmed alleles	3431
Regions with 2 confirmed alleles	2 030 492
Regions with > 2 confirmed alleles	17 408
Bases in regions with < 2 confirmed alleles	66 458
Bases in regions with 2 confirmed alleles	3 431 851
Bases in regions with > 2 confirmed alleles	135 910
Maximal size of haplotype block, bp	27 223
Average size of haplotype block, bp	816

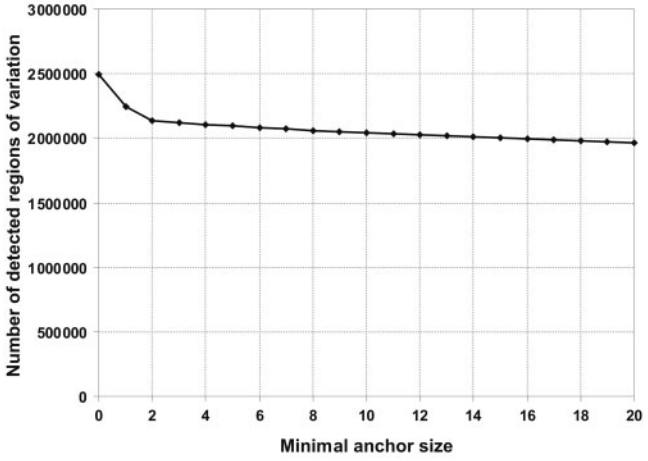

Fig. 3. Number of regions of sequence variation detected in the human genome assembly as a function of the MAS. With MAS = 0, no columns with variation will be grouped into regions and with MAS = 1, only immediately adjacent columns with variation will be grouped.

For example, Levy *et al*. (2007) detected 53 823 block substitutions of size 2–206; 263 923 heterozygous indels of size ranging between 1 and 321 bp and mean size 2.4 bp; and over 28 000 other complex variants. All alleles in such a variant are consistent with at least one of underlying reads. In our processing, this is controlled by parameter MAS, which models a correlation distance between two mutation events, such as SNPs: two mutation events will be regarded as independent if they occur in loci separated by a region with no sequence variation, and the size of this region is equal or greater than MAS. Experimental studies have shown that, indeed, variants in close physical proximity are often strongly correlated. Furthermore, this correlation structure, or linkage disequilibrium, is complex and varies from one region of the genome to another (Hinds *et al*., 2005). In this study, we used the default setting MAS = 11. With MAS = 0, no columns of sequence variation will be grouped into regions, so the column-by-column consensus calling approach will be used, although alleles will still be detected. As follows from the data presented in Figure 3, if we vary MAS parameter between 0 and 11, the total number of detected regions of variation will drop by 18.5%, but a subsequent increase of this parameter from 11 to 20 would only result in an additional drop of about 2%. Thus, the total number of detected regions of variation does not change dramatically when MAS parameter exceeds the value of 11. Using larger MAS may result in accumulation of base calling errors, so that every read segment will become unique, and no confirmed alleles will be detected in a region of variation. The total number of 460 373 regions of sequence variation of size > 1 was detected in the multiple alignment of reads using the default MAS = 11. By comparing the consensus sequences of the regions with consensus sequence computed using column-by-column approach and the majority rule, we found that our default processing fixes 33 269 occurrences of mosaic haploid representation of a diploid locus similar to that shown in Figure 1.

Not any column containing two or more variable bases will be identified as a variant. To minimize the number of false-positive and false-negative variants, we have optimized the parameters of the variant detection procedure as follows.

In the case of the SNP type of variation, where both the variable base and the reference consensus call is not a gap, we require that the average quality value of the bases representing a minority allele exceeds a certain cutoff value. This cutoff value was calibrated against a subset of ~1.5 million detected SNPs which matched those in the dbSNP database. This was considered as a set of bona fide SNP variation. Specifically, Figure 4 shows two histograms of SNPs detected in the human genome assembly as a function of QV_{ave}, the average quality value of bases of the minority allele. The first histogram corresponds to SNPs matching dbSNP, as determined from those SNP positions in our assembly that match the NCBI version 36 assembly, and therefore interpreted as true SNPs. The second one corresponds to all the remaining, newly detected SNPs and clearly shows the presence of two maxima. Position of the second maximum coincides with position of the (only) maximum on the first histogram. This maximum in both distributions which is shifted to higher QV_{ave} is therefore interpreted as corresponding to the true SNPs. The first maximum, observed only in the second histogram at lower QV_{ave}, is interpreted to correspond to false (spurious) SNPs. To better separate the newly detected true SNPs from false SNPs, we have set the cutoff value of QV_{ave} to the position of the minimum located between the two maxima in the second histogram, $QV_{ave} = 21$. Furthermore, we fitted the area underneath the second maximum to a Gaussian shape, which thus represented our guess regarding the newly detected true SNPs, and we interpreted the remaining area under this histogram as corresponding to false SNPs. This allowed us to partition all the detected SNPs into four groups: 1 778 993 true positives (TP), represented by all the true SNPs with $QV_{ave} \geq 21$; 431 878

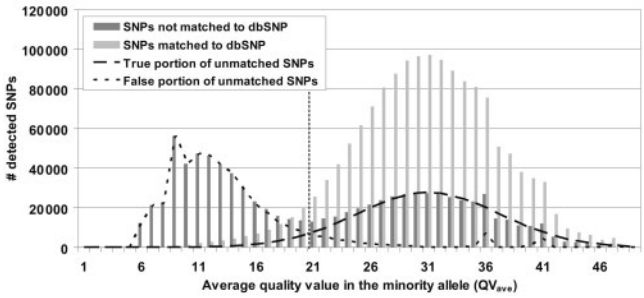

Fig. 4. Histograms used to calibrate parameters of the variant detection procedure: number of heterozygous SNPs detected from the human assembly as a function of average quality value in the minority allele. The dark gray bars represent newly detected SNPs, and the light gray bars represent already known SNPs. The dashed Gaussian curve represents our guess regarding the true portion of the new SNPs, and the dotted curve represents the remaining, false portion of the new SNPs. The area under the portion of the dashed curve corresponding to $QV \geq 21$ represents 438 814 newly detected heterozygous SNPs.

true negatives (TN), represented by all the false SNPs with $QV_{ave} < 21$; 61 842 false positives (FP), representing all the false SNPs with $QV_{ave} \geq 21$ and 115 813 false negatives (FN) representing all the true SNPs with $QV_{ave} < 21$. The false positive rate and false negative rate were estimated as follows: FP rate $= FP/(FP + TN) = 12\%$ and FN rate $= FN/(FN + TP) = 6\%$.

This optimized SNP detection procedure was not used by Levy *et al*. (2007). Instead, a set of filters was applied to the initially detected 'raw' set of new non-synonymous coding SNPs, which resulted in a 3.2 fold reduction of its size. After that, a manual trace inspection of the filtered variants was performed, which allowed estimation of their final FP rate at ~34%. Thus, our optimized variant detection procedure allows ~2.8 fold reduction of the FP rate compared to the use of filtering approach.

In the case of indel type of variation, where either varied base or the preliminary consensus call was a gap, the cutoff value of the sum of the two highest quality values of varied base was determined by visually inspecting traces in the vicinity of the candidate indel locations. We found that, in many cases, false positive indels were detected in the regions of homopolymer runs (i.e. repeats of a single base in a read). Using the QV cutoff value of 60, however, reduces detection of such false positive indels to negligibly small number. Out of total number 1 795 049 of variants of size one detected at MAS = 11, there were 170 027 indels and 1 624 122 SNPs. The cutoff QV values specified above will generally work only with assembly of reads produced by Sanger sequencing. However, our preliminary analysis indicates that the approach to calibrating QV cutoffs against a set of known true SNPs presented above can be generalized and extended to hybrid assemblies of Sanger and 454 reads (Goldberg *et al*., 2006). Furthermore, indel detection cutoff QVs could be calibrated in a similar way, and FP and FN rates could be further reduced by applying simultaneously more than one constraint on QVs.

In addition to grouping closely spaced columns of sequence variation into regions, the second approach to connecting variants discussed in this article is phasing of alleles in adjacent regions of variation to form even larger haplotype blocks. The first approach is only applicable to variants located in close physical proximity, less than MAS bases apart. While the second approach can link regions spaced at larger distances comparable with read length and results in forming blocks of size up to ~27 kb (Table 1), it is more prone to potential errors than the first approach.

As an illustrative example, consider six hypothetical overlapping reads numbered $1, \ldots, 6$ and spanning four adjacent regions of variation $1, \ldots, 4$. Assume that the read segments are partitioned between two confirmed alleles as follows:

Region 1: Allele1 $= \{1, 2, 3\}$ Allele2 $= \{4, 5, 6\}$,
Region 2: Allele1 $= \{1, 2, 4\}$ Allele2 $= \{3, 5, 6\}$,
Region 3: Allele1 $= \{1, 4, 5\}$ Allele2 $= \{2, 3, 6\}$,
Region 4: Allele1 $= \{4, 5, 6\}$ Allele2 $= \{1, 2, 3\}$.

According to our phasing algorithm, the alleles of the same order number will be considered 'phased' between any couple of adjacent regions, because they share most of reads. Thus, all the regions of variation will be considered to belong to the same haplotype block. Nevertheless, because of the slow 'drift' of reads between alleles in adjacent regions, phases get completely inversed between regions 1 and 4. Alleles with the same order number in these regions represent opposite haplotypes, so the major consensus sequence of the entire haplotype block will be a mixture of two haploid sequences.

Direct detection of phase inversion events in the human genome assembly is complicated, because variant regions 1 and 4 would, on average, be connected by less than one read. Special care needs to be taken to reduce the risk of phase inversion in haplotype assembly (Kim *et al*., 2007b). As a simple measure of such risk, we consider the number of reads which are found simultaneously in both confirmed alleles. This number is 644 572 with our default processing, and it will increase dramatically, by about four-fold up to 2 748 025, if the quality value constraints described in Step 2 of our algorithm are relaxed, so that any redundancy in a column of aligned bases is regarded as polymorphism. This suggests that, by rejecting spurious SNPs and indels, our tuned procedure may help avoid phase inversion and improve the accuracy of haplotype assembly.

In this study, we made only an initial step in haplotype block assembly, which did not take into account the mate-pair information. Our analysis indicated that 52.3% of the detected variants could be phased, or connected by reads into haplotype blocks. However, the average length of a haplotype block was only 816 bp and the largest block was ~27 kb long (Table 1). As discussed in (Levy *et al*., 2007), using mate-pair information and an aggressive greedy approach would allow assembly of much larger haplotypes, with half of genome covered by blocks spanning >200 kb of sequence each. However, the accuracy of this assembly has not been explored. We hope that the algorithm and software presented in this article will serve as a convenient framework for further analysis of this challenging problem.

ACKNOWLEDGEMENTS

This work was supported by the J. Craig Venter Institute and by the National Institute of General Medical Sciences (R01 GM077117-01 to G.S.). The authors would like to thank Pauline Ng and Wengwah Lee for the help in processing dbSNP and palm oil tree data and valuable comments.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Aparicio,S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**, 1301–1310.

Barker,G. *et al*. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.

Batzoglou,S. *et al*. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res*., **12**, 177–189.

Bonfield,J.K. *et al*. (1998) Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Res*., **26**, 3404–3409.

Chen,K. and Pachter,L. (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol*., **1**, 106–112.

Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol*., **7**, 111–122.

Daly,M.J. *et al*. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet*., **29**, 229–232.

Denisov,G.A. *et al*. (2004) A system and method for improving the accuracy of DNA sequencing and error probability estimation through application of a mathematical model to the analysis of electropherograms. *US Patent* 6681186.

Goldberg,S.M. *et al*. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl Acad. Sci. USA*, **103**, 11240–1124.

Green,P. (2005) PHRAP documentation. http://www.phrap.org.

Havlak,P. *et al*. (2004) The Atlas genome assembly system. *Genome Res*., **14**, 721–732.

Hinds,D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.

Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res*., **9**, 868–877.

Huang,X. *et al*. (2003) PCAP: a whole-genome assembly program. *Genome Res*., **13**, 2164–2170.

Huntley,D. *et al*. (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. *Bioinformatics*, **22**, 495–496.

Indap,A.R. *et al*. (2005) Analysis of concordance of different haplotype block partitioning algorithms. *BMC Bioinformatics*, **6**, 303.

International HapMap Consortium,T.I.H. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Istrail,S. *et al*. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA*, **101**, 1916–1921.

Jaffe,D.B. *et al*. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*., **13**, 91–96.

Jones,T. *et al*. (2004) The diploid genome sequence of Candida albicans. *Proc. Natl Acad. Sci. USA*, **101**, 7329–7334.

Kent,W.J. and Haussler,D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res*., **11**, 1541–1548.

Kim,J.H. *et al*. (2007a) Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi. *Genome Res*., **17**, 1101–1110.

Kim,J.H. *et al*. (2007b) Accuracy assessment of diploid consensus sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **4**, 88–97.

Lancia,G. *et al*. (2001) SNPs problems, complexity, and algorithms. *Lect. Notes Comput. Sci*., **2161**, 182–193.

Levy,S. *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biol*., **5**, 2113–2144.

Lippert,R. *et al*. (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinform*., **3**, 23–31.

Marth,G.T. *et al*. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Gen*., **23**, 452–456.

McKusick,V.A. (1998) *Mendelian Inheritance in Man*. Johns Hopkins University Press, Baltimore.

Mullikin,J.C. and Ning,Z. (2003) The phusion assembler. *Genome Res*., **13**, 81–90.

Myers,E.W. *et al*. (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

Nickerson,D.A. *et al*. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*., **25**, 2745–2751.

Pevzner,P.A. *et al*. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Stephens,J.C. *et al*. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.

Sutton,G. *et al*. (1995) TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol*., **1**, 9–19.

Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet*., **6**, 805–814.

Venter,J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Venter,J.C. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Wang,J. and Huang,X. (2005) A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage. *BMC Bioinformatics*, **6**, 220.

Wang,J. *et al*. (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res*., **12**, 824–831.

Waterston,R.H. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Yooseph,S. *et al*. (2007) The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol*., **5**, e16.