Nucleic Acids Research, 2005, Vol. 33, No. 18 ***e152***
doi:10.1093/nar/gni152

Shotgun haplotyping: a novel method for surveying allelic sequence variation

Sarah J. Lindsay, James K. Bonfield and Matthew E. Hurles*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Received June 24, 2005; Revised and Accepted September 19, 2005

DDBJ/EMBL/GenBank accession numbers: DQ075317, DQ075318, DQ075319, DQ075320

ABSTRACT

Haplotypic sequences contain significantly more information than genotypes of genetic markers and are critical for studying disease association and genome evolution. Current methods for obtaining haplotypic sequences require the physical separation of alleles before sequencing, are time consuming and are not scaleable for large surveys of genetic variation. We have developed a novel method for acquiring haplotypic sequences from long PCR products using simple, high-throughput techniques. This method applies modified shotgun sequencing protocols to sequence both alleles concurrently, with read-pair information allowing the two alleles to be separated during sequence assembly. Although the haplotypic sequences can be assembled manually from the resultant data using pre-existing sequence assembly software, we have devised a novel heuristic algorithm to automate assembly and remove human error. We validated the approach on two long PCR products amplified from the human genome and confirmed the accuracy of our sequences against full-length clones of the same alleles. This method presents a simple high-throughput means to obtain full haplotypic sequences potentially up to 20 kb in length and is suitable for surveying genetic variation even in poorly-characterized genomes as it requires no prior information on sequence variation.

INTRODUCTION

Understanding patterns of genetic variation is essential for studies of the genetic basis of complex diseases, and the reconstruction of evolutionary history. Genetic variation can be assayed in various ways, each of which conveys different levels of information. It has been demonstrated that haplotypes of genetic markers [e.g. single nucleotide polymorphisms (SNPs)] are more informative than genotypes (1) for studying processes such as linkage disequilibrium and recombination, and for examining disease associations. However, haplotypes of genotyped SNPs are inherently biased by the use of a restricted set of individuals for SNP discovery, and will only represent a proportion of the variation present in any individual. Haplotypic sequences are the most informative tools of all.

While much recent attention has focussed on experimental and statistical methods for obtaining haplotypes from genotypic marker data, little progress has been achieved in generating haplotypic sequences. Allelic sequences can be physically separated before amplification and sequencing by constructing somatic cell hybrids (1,2), or using large insert cloning and sequencing (3). Both these methods can be labour-intensive, time consuming and expensive. Alternatively, single molecule dilution (4–6) and allele-specific PCR have been used for the sequencing of single alleles (3) as well as the phasing of known genotypes (7,8). However, due to the difficulty of designing efficient PCR primers for single molecule amplification or allele-specific amplification, these technologies are not applicable to all genomic targets, often require prior knowledge of genotypes, and as a consequence are not scaleable to large surveys of genetic diversity. In principle, single molecule sequencing methods may provide a means to generate large volumes of haplotypic sequences, although the technology is immature at present (9).

In recent years there has been an explosion in methods for statistical inference of haplotypes from genotypic information (10–14), some of which are appropriate for inferring haplotypic sequences from diploid sequences as well as for deducing marker haplotypes. However, all of these statistical methods come with associated error rates, are reliant on the notoriously error-prone process of heterozygote detection in diploid sequence traces (15,16) and are often unreliable when sample sizes are small (17).

We have designed and tested a novel high-throughput method for obtaining haplotypic sequences that requires no prior knowledge of SNP positions or frequency. We demonstrate that it is possible to assemble haplotypic sequences from

*To whom correspondence should be addressed. Tel: +44 (0) 1223 495377; Fax +44 (0) 1223 494919; Email: meh@sanger.ac.uk

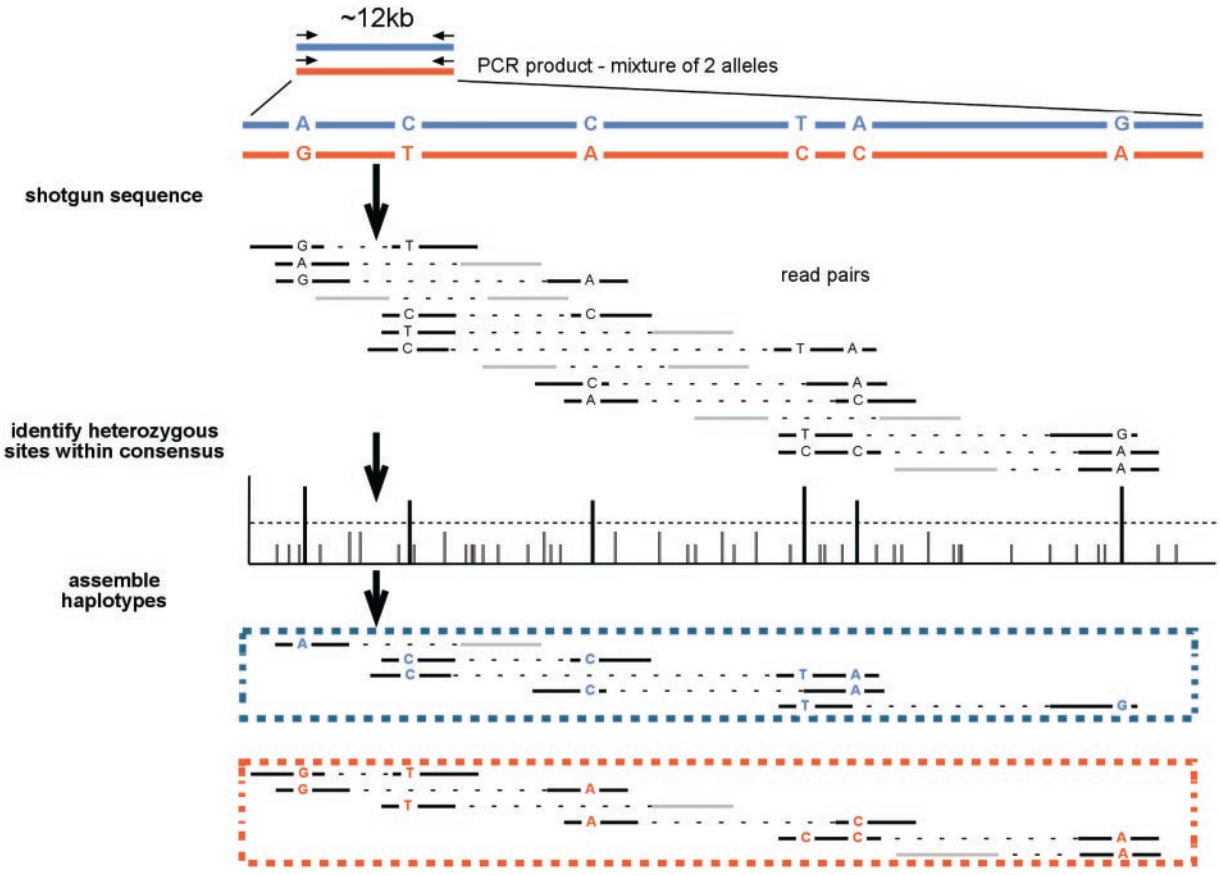

Figure 1. Shotgun haplotyping procedure. A long PCR product of a diploid locus is sheared by sonication and a size range of 2–4.5 kb fragments are cloned. The two alleles are shown in blue and orange. The clones are then shotgun sequenced from both ends to a high coverage across the PCR product and assembled against a consensus. The histogram represents the likelihood for each site that it is heterozygous. Likely heterozygous positions are identified as sites that exceed a threshold value represented by a dashed line. The haplotypes of these heterozygous sites are assembled from read-pair data using a phasing algorithm. Only bases at heterozygous sites are shown and sequence reads containing no heterozygous sites are shown in grey.

a mixture of two alleles using dense read-pair information from a shotgun library of end-sequenced fragments (see Figure 1). Some mathematical consideration has been given to the issue of separating haplotypes from read-pairs (18,19), but these methods have not been applied to real data. We have modified pre-existing assembly software and have successfully generated haplotypic sequences for two 12 kb autosomal regions amplified by long PCR. This novel method presents the opportunity to collect haplotype information from a panel of individuals using a robust and scaleable pipeline, and is particularly applicable to surveying allelic variation in duplicated sequences, such as (pseudo)genes with high sequence similarity.

MATERIALS AND METHODS

DNA subjects

Two anonymous male genomic DNA samples were used as templates for long PCR amplifications: one a gift from Mark Jobling, the other from the ECACC human diversity panel.

Long PCR of test loci

Two ~12 kb long test loci (T1 and T2) on chromosome 17 were amplified and sequenced in this study. All reactions were performed in a 50 µl volume using the 20 kb Expand Plus PCR kit (Roche Applied Science). The reactions were carried out following the manufacturer's protocol using an extension time of 11 min and annealing temperature 57°C. All oligos were synthesized by Sigma Genosys. The two ~12 kb portions of chromosome 17 were amplified using the oligos CMT1AD2-CCACATTACTGCTTCCTCATGTGT and CMT1AINT5-GTTCATGGTTCATGCTGAGGGTTG, CMT1AD1-GGGGGTAGAAAAGGGGTCTCATTTTCC and CMT1AINT3-ATTACAGCTACTGTTGCAGCAGTG. The latter amplicon contains a single exon of the COX10 gene; the former amplicon contains no known coding sequences.

Cloning long PCR products

Long PCR products were purified and cloned using the Expand Cloning kit (Roche Applied Science) and screening for allele-specific clones was carried out by end-sequencing. The reads from both ends were assembled into a Gap4 database (20), and heterozygous sites within these reads were used to identify full-length clones for each allele. The inserts from clones representing each allele were then recovered from the vector by restriction enzyme digestion and gel purification and then a Short Insert Library (SIL) was made for each before they were

shotgun sequenced using the protocol below (without heterozygote detection and phasing).

SIL preparation

SILs were prepared in the vector pUC19 using the methods described by (21), adapted to incorporate a broader size range of inserts (2–4.5 kb).

SIL colony sequencing and assembly

Ampicillin resistant colonies were picked and grown in 384 well format, and then DNA prepared using alkaline lysis. End-sequencing reactions on each clone was then performed to generate read-pair data. Subsequently the sequence data were clipped for vector and quality, and assembled using the Gap4 assembly software (20).

Heterozygote detection within Gap4

The two independent measures of the likelihood of a given position within the sequence assembly of haploid sequences being heterozygous are described in turn.

The first measure of heterozygote likelihood is a modified version of a consensus algorithm that computes consensus probabilities for A, C, G, T and 'gap' at each base position, with each of the five being computed only from sequences containing that base call at that position (such that discrepant bases will not reduce the confidence of that consensus assignment). A traditional consensus algorithm will, for each column of aligned bases in an assembly, compute the most likely consensus base call and assign a probability value of correctness (22). In Gap4, the consensus algorithm takes into account the scaled confidence of each base call (phred score) and whether there are mismatches within the column. From this analysis, the two bases with the highest consensus probability values are considered as the putative heterozygote alleles and the second highest probability indicates the likelihood this being a true heterozygote. This value is then log transformed to improve dynamic range of this measure. This algorithm is capable of detecting both base substitutions and small indel events.

The second measure of heterozygote likelihood is computed by examining the distribution of base calls in a specific column. We expect a heterozygote to be represented by two alleles with an approximate 50:50 ratio represented in the sequence reads across the variant base. In an alignment of 10 bases deep containing a heterozygote we would expect, 5 sequences per allele on an average. Given that a site is heterozygous, the probability of measuring exactly a K versus (N-K) split within N sequences can be computed as a binomial coefficient with an assumed probability of 0.5 for each allele occurring.

Specifically for n trials with k successes where $P(\text{success}) = p$ this is computed as:

$$P(k,n) = \frac{n!}{k!(n-k!)} p^k (1-p)^{n-k}$$

Simplifying this for $p = 0.5$:

$$P(k,n) = \frac{n!}{k!(n-k!)} 0.5^n$$

In early trials we found that this binomial coefficient was overly conservative (true heterozygotes that deviate from a 50:50 ratio were down-weighted), as a result of systematic allelic biases in amplification, cloning and sequencing. The statistical properties of these biases are largely unknown, and therefore difficult to model, but we found that by capping at a maximum of $N = 10$ (and downscaling the true sequence depth accordingly) we obtained good discrimination between true heterozygotes and PCR errors.

These two quality measures (second highest base confidence and binomial coefficient) are then multiplied to give a single confidence value and a threshold applied to produce a set of putative heterozygote sites with associated quality scores. In our early analyses with this algorithm it became apparent that runs of mononucleotides can be extremely variable due to replication slippage during PCR, and that this variability can confuse attempts to phase haplotypes. Thus we down-weighted this type of variant in our algorithm.

There is scope for implementing alternative mechanisms here too without invalidating the subsequent haplotype assembly step. One such method could be to identify 'defined nucleotide positions' (23), although it is not immediately obvious how to apply this two-column based score as a generic single score per column.

Haplotype assembly algorithm

A clustering algorithm was deliberately chosen for phasing the heterozygous sites to facilitate the future use of this method in assembling large insert clones containing two or more copies of a duplicated sequence (J. K. Bonfield unpublished data). Apparently homozygous sites within individual read-pairs are removed to leave only those sites identified as being heterozygous by the above algorithm. Figure 2A shows the read-pairs shown in Figure 1 in this format. A matrix of pairwise similarity between read-pairs is then calculated. The similarity measure takes into account the number of sites at which two read-pairs differ, the number of sites at which they agree and the quality scores of the associated heterozygote calls.

For each heterozygote site j shared by two read-pairs we compute the Pearson correlation coefficient from the two base frequencies in each set of read-pairs as follows:

$$r_j = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5}(x_i - \bar{x})^2 \sum_{i=1}^{5}(y_i - \bar{y})^2}} \quad \text{where} \quad -1 \leqslant r_j \geqslant 1$$

The vectors x and y here are defined as the frequency of base call types (A, C, G, T and gap) at a single SNP site j. Initially x and y will contain just one single base call (with a frequency of one) representing a single read-pair. As clustering progresses and multiple read-pairs are grouped together the absolute magnitude of the x and y vectors will increase.

The combined similarity score between any two read-pairs a and b is then:

$$E_{a,b} = \sum_{j=1}^{n} (r_j s_j t_a t_b)$$

where s_j is the confidence value for this heterozygote site, as described above, and t_a is a measure of the reliability of the

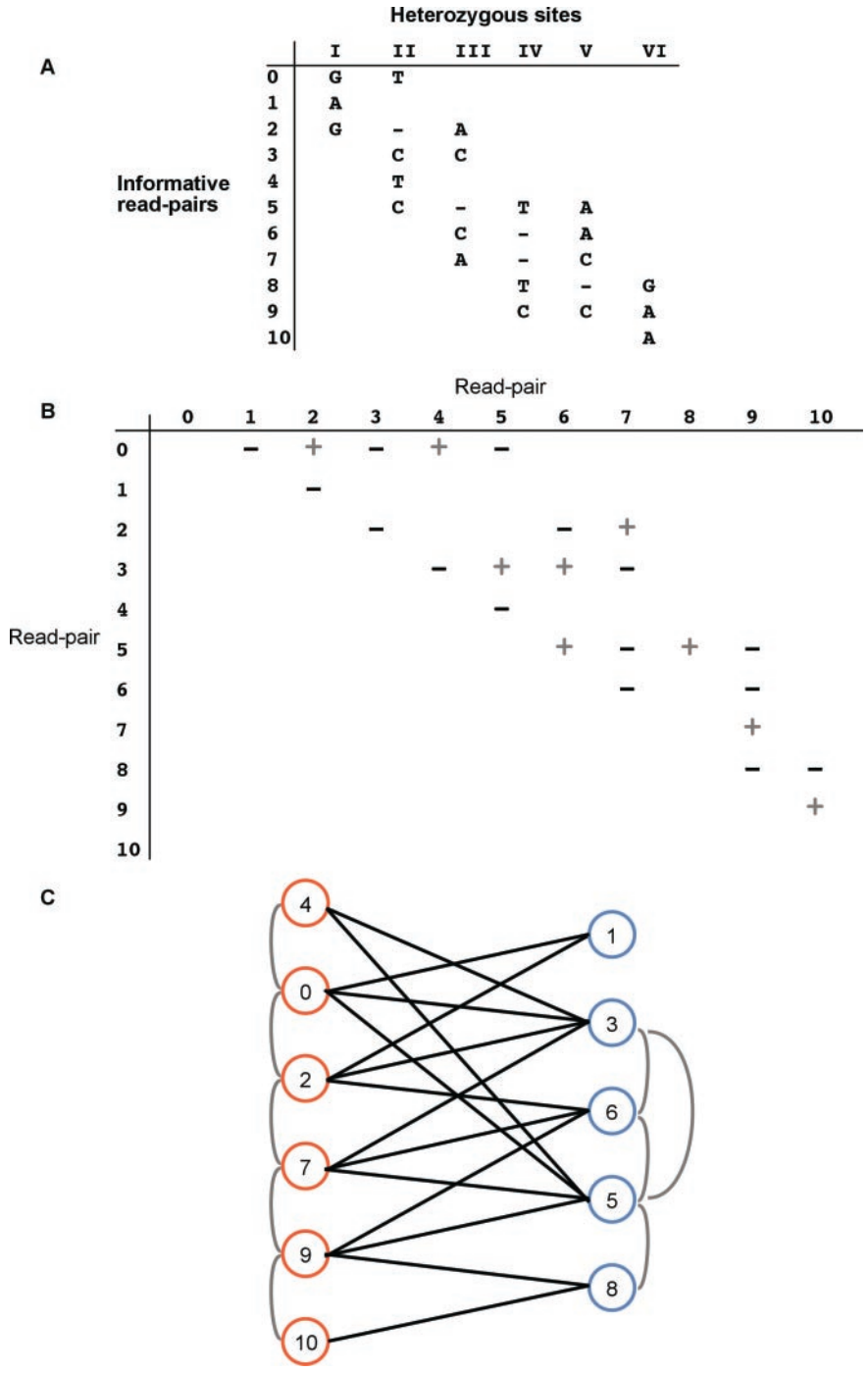

Figure 2. Phasing heterozygous positions from read-pair data. Processing steps applied during the phasing algorithm are shown, using the read-pairs within Figure 1 as a worked example. (**A**) Read-pairs are condensed down to heterozygous positions. (**B**) A matrix of similarity between read-pairs is calculated. Positive and negative similarity scores are indicated by pluses and minuses respectively. In reality these similarity scores differ in magnitude as well as sign. The algorithm used to calculate these similarity scores is detailed in the Materials and Methods. (**C**) Haplotype phasing can be represented by a bipartite graph in which similarity scores within a set are all positive (grey lines), whereas links between the two sets are all negative (black lines). The sets of read-pairs are coloured corresponding to the alleles in Figure 1.

read-pair based on its similarity to the expected size range of clone insert, as previously implemented in Gap4.

Thus read-pairs that cover the same region on the same haplotype will give strongly positive similarity scores, whereas read-pairs that cover the same region on different haplotypes give strongly negative scores. Figure 2B shows a schematic of the matrix of similarity scores for the read-pairs in Figure 2A. This matrix can also be represented as a bipartite graph (Figure 2C) in which all negative links lie between the two sets of read-pairs representing each haplotype. An iterative clustering algorithm is then used to link read-pairs coming from the same haplotype. The two

read-pairs giving the highest positive similarity score in the matrix are merged and the matrix recomputed. This step is then repeated until all pairs within the similarity matrix are below a predefined threshold (defaulting to zero). Using a threshold of zero is most appropriate when we know that only two alleles are present as the lack of a negative score match is typically sufficient to indicate which allele a cluster belongs in. In the case of many chimeric templates or a larger number of homologous sequences we recommend using a higher threshold.

If necessary, a final inference step is used to cluster sets of read-pairs that do not overlap shared heterozygous sites, but where both have strong negative similarity scores with a third set. This situation can arise when one haplotype is completely phased, but the other haplotype is split into two, but with all heterozygous sites covered. In this step, the similarity score of two sets of clustered read-pairs that do not overlap a common heterozygous site is modified by adding an additional inferred score as follows.

Let N_a be the set of read-pair clusters with a non-zero similarity score from read-pair cluster a

Let N_b be the set of read-pair clusters with a non-zero similarity score from read-pair cluster b

$N_{a,b} = N_a \cap N_b$ is therefore the set of read-pair clusters 'linked' to both cluster a and cluster b.

We therefore define a 'link score' between read-pair clusters a and b as follows:

$$L_{a,b} = E_{a,b} + \sum_{x \in I_{a,b}} |E_{a,x} + E_{b,x}| - |E_{a,x} - E_{b,x}|$$

Additional clustering (as above) is then performed on this matrix of link scores to finalise the read-pair clusters.

We developed a Graphical User Interface (GUI) to allow inspection and manual curation of the automated haplotyping algorithm (see Supplementary Figure 2). This GUI displays the putative heterozygous sites and the alleles found in the different sets of read-pairs that result from the prior clustering. This display allows the user to exclude apparently false heterozygote sites, and re-run the phasing algorithm. This display also allows the user to inspect the sequence coverage across the region to check for low coverage segments where heterozygous sites might have been missed.

RESULTS

We reasoned that it should be possible to assemble haplotypic sequences from a mixture of two alleles using high density read-pair sequences from randomly sheared fragments of the two alleles. In this way, information on the phase of heterozygous positions present in the two alleles is gained from having two reads from a single template separated by >1 kb. If both reads from a single clone cover a heterozygous site then the phase of these sites is determined (an allele present in the forward read of a single cloned fragment can be assumed to be in phase with the allele present in the reverse read of the same clone). Such a method avoids the problems associated with calling heterozygote positions from diploid sequence data as in this method such positions are revealed by comparisons of multiple haploid sequences, rather than a mixture of signals from both alleles in a single sequence trace.

In practice, the source of the mixture of two alleles comes from the PCR amplification of an autosomal sequence from a diploid organism. In outline (see Figure 1), long PCR is first used to amplify a region of interest. Subsequently, the PCR product is randomly sheared by sonication to produce a set of fragments derived from both amplified alleles. These fragments are cloned and end-sequenced in both directions, resulting in a high coverage of paired sequence reads from both alleles across the length of the original PCR product. These reads are then aligned against one another, variant positions between the two alleles are identified as high quality base differences between the reads from multiple cloned fragments covering the same region, and the haplotypes are phased using read-pair information.

We predicted that several factors may confound this analysis. First, the PCR error rate may be so high that errors might be mistaken for heterozygous positions as they manifest themselves as high quality discrepancies within the alignment of reads. Second, insufficient sequence coverage and/or unequal allele amplification may cause heterozygous sites to be overlooked. Third, a high frequency of chimeric clones resulting from jumping PCR or cloning multiple fragments in a single vector may mislead phasing of SNPs. Finally, an insufficient density of SNPs may make it impossible to bridge the gap between neighbouring SNPs. We addressed each of these concerns in turn and found that after modifying certain parameters of standard shotgun sequencing protocols none of the above factors represented a significant problem (described below).

We estimated the likely PCR error rate by calculating the expected number of PCR errors based on the number of effective cycles of replication in the PCR and the published error rate from the manufacturer. This gives an expected error rate of ~1 PCR error every 6 kb. We also estimated the PCR error rate empirically by comparing the frequency of high quality discrepancies apparent within a shotgun sequence assembly of a clone to that in a shotgun sequence assembly of a PCR product from a haploid sequence. Differences in the frequency of high quality base discrepancies must result from PCR errors as miscalls are present in both assemblies (and are very rare at high quality base calls) and there are no true SNPs in a haploid sequence. This analysis suggested a higher error rate of a PCR error every 2 kb (data not shown). This higher figure was confirmed in subsequent comparisons of cloned haplotypic sequences to the true sequence (see below). Despite this relatively high rate of PCR error, sequence assemblies derived from PCR products of haploid sequence showed that the same PCR error was very rarely observed in more than one read-pair within an assembly, and should therefore be easily distinguishable from real heterozygote sites given sufficient sequence coverage. We found that with an average sequence coverage of 24× or greater, the distribution of allele frequencies of PCR errors (present in <20% of reads at a given site) is distinct from the distribution of minor allele frequencies of true heterozygotes (>20% of reads at a given site), and no true heterozygotes are missed.

The frequency of chimeric clones can only be estimated by observing their presence within a sequence assembly derived from known haplotypes. The identification of chimeric clones within such sequence assemblies (data not shown) indicates that such clones are extremely infrequent (<1%) and do not pose a problem for haplotype assembly.

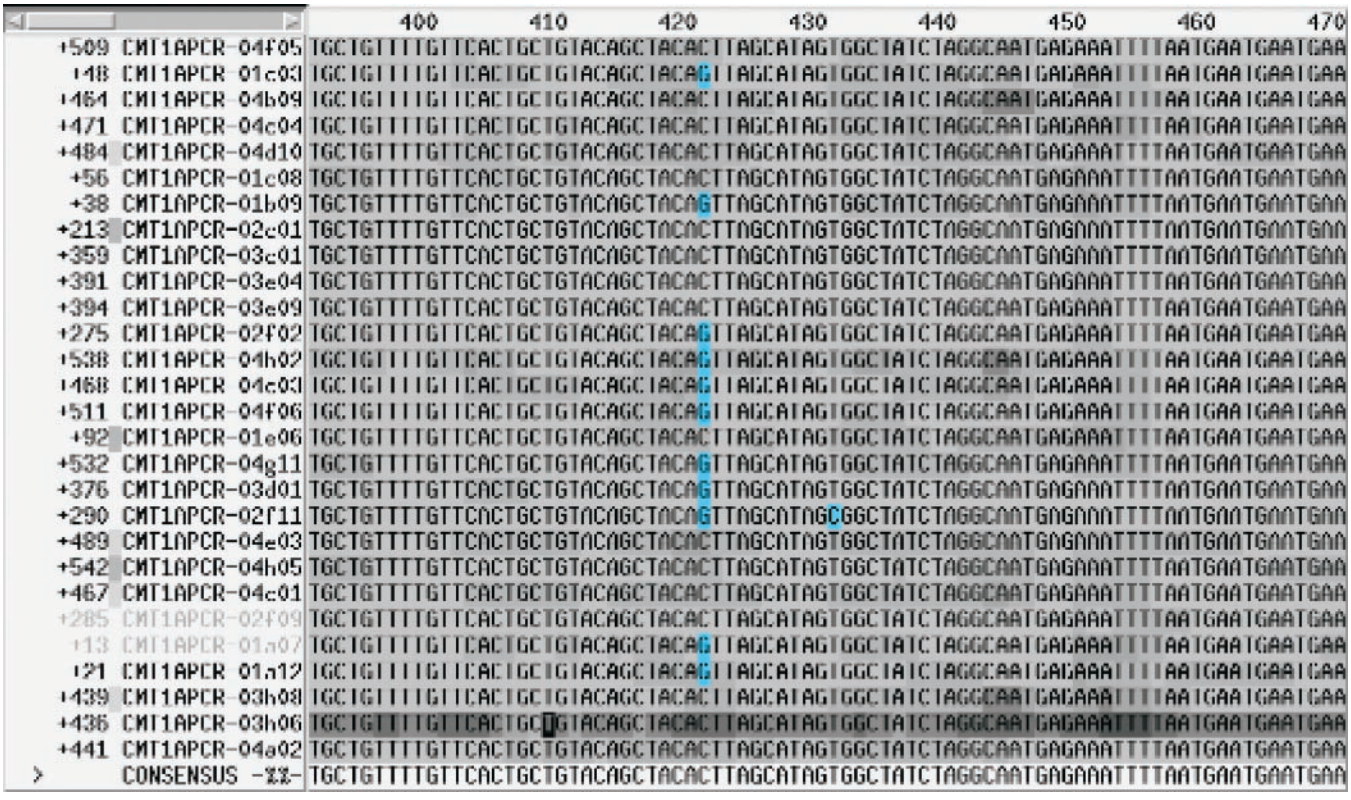

Figure 3. Identifying heterozygous positions within a sequence assembly. A screenshot of the Gap4 sequence assembly software in which bases flagged in blue represent high quality discrepancies from the consensus. The phred (quality) scores of each base call are shown in greyscale, from light grey (good) to dark grey (poor). A singleton base discrepancy with a high phred score is indicative of a PCR error. A discrepant base call found in approximately half of all sequence reads across a given position is indicative of a heterozygous site.

The average density of heterozygous positions has been estimated to be ~0.75 per kb within humans (24). Therefore, it is entirely possible that gaps of several kb exist between neighbouring heterozygous sites. We therefore modified existing shotgun sequencing protocols to generate read-pairs from a broader size distribution of sheared fragments (2–4.5 kb) and found in our validation experiments (see below) that this size range was sufficient to phase across 12 kb at both test loci.

Heterozygote detection and haplotype assembly

Once all the read-pairs derived from a single PCR product are loaded into the GAP4 sequence assembly software, and discrepant sites are flagged, it becomes trivial to identify true heterozygous sites by eye (see Figure 3). It is then possible to phase these alleles manually by seeking read-pairs in which two or more heterozygous sites are found and assigning the observed allele at each site to the same phase. Although it is possible to infer the phase of the unobserved alleles having observed the phase present in a given read-pair, we developed a set of rules in which each allele had to be bound into its phase assignment by at least two observations of that phase in two independent read-pairs. This process of manual haplotyping is laborious and prone to human error. Consequently, we decided to automate heterozygote detection and subsequent haplotype assembly.

Likely locations of heterozygous base substitutions and small indels were identified computationally by combining two independent measures of the likelihood of a heterozygote existing at a given base position within an assembly of aligned shotgun sequences. The first measure takes into account the quality of the base calls and is based on a modified consensus algorithm, while the second measure considers how closely the frequency of discrepant base calls at a given base position corresponds to the 50:50 ratio expected for a true heterozygote. These two quality measures are combined to give a single value and a threshold applied to produce a set of potential heterozygote positions with associated quality scores.

It is important to note that the process of heterozygote detection need not be exhaustive as the purpose is to identify sufficient heterozygote positions to split the read-pairs into two sets representing the two haplotypes and not to find all such positions. Once the sequences of the two haplotypes have been assembled, they can be aligned against one another to identify comprehensively all heterozygous positions.

Having identified computationally a set of heterozygous positions within the sequence assembly, we developed an algorithm to automatically phase these heterozygous positions and identify the two sets of read-pairs that derive from each haplotype. It has been previously shown that one mathematical method for reconstructing haplotypes from dense sequencing information is to estimate the bipartite graph that separates the read-pairs originating from the two haplotypes (see Materials and Methods). However, it has also been shown that estimating this bipartite graph becomes NP-hard (exact inference of the bipartite graph is not computationally feasible) when analysing data comprised of pairs of non-overlapping sequence reads (18). Therefore, to split the read-pairs into two sets that

could subsequently be used to assemble the sequences of the two alleles, we chose to develop a heuristic haplotyping algorithm based on the iterative clustering of read-pairs that share the same alleles at heterozygous sites.

The clustering algorithm splits the read-pairs into two major sets that represent the two alleles. A minority of read-pairs falls into neither major set. These represent read-pairs containing no heterozygous positions. Both major sets of read-pairs (corresponding to the two haplotypes) are then assembled independently and the resultant consensus sequences of each allele exported for downstream analyses.

Validating shotgun haplotyping

To validate the protocol described above we compared haplotypic sequences for two ~12 kb long PCR products (T1 and T2) obtained by shotgun haplotyping with sequences of full-length clones of each allele. Both regions are well-annotated and contain no unusual sequence features, including GC content and distribution of simple repeats.

For each of the test loci we generated 384 paired read sequences. After removing failed reads (i.e. low quality reads and vector reads) we retained 598 reads for T2, but a higher frequency of failed reads for T1 required an additional 96 read-pairs to be sequenced to raise the coverage above 500 reads. This gave an average coverage across the test loci of 34× and 24× for T1 and T2 respectively, although coverage varied along both amplified loci, with lowest coverage at the ends.

In the two ~12 kb loci we observed 26 and 35 heterozygous base substitutions for T1 and T2 respectively, and the ratio of the sequence reads assigned to individual haplotypes was 1.08:1 and 1.11:1 This suggests that allele-specific amplification efficiencies were not dramatically different, and are unlikely to pose a problem. The median frequency of the less frequently observed allele at a heterozygous site was 43%.

In order to test the accuracy of our phased haplotypes we cloned the long PCR products from the same PCR as was used for the shotgun haplotyping into a cosmid vector. We then screened these clones by sequencing across heterozygous positions at either end of the allele to identify a full-length clone for each allele. A single clone for each allele was then shotgun sequenced.

We then compared cloned allelic sequences to the haplotypes obtained from shotgun haplotyping. All of the heterozygous positions we identified during shotgun haplotyping also differed between the sequences of the two cloned alleles (see Table 1). We demonstrated that at both loci we had correctly phased the heterozygous positions that we had identified (see Table 1). At both test loci our automatic haplotyping algorithm successfully assembled the same haplotypes that we had identified manually, thus showing that our heuristic approach is sufficiently powerful to be of general utility.

The clone sequences contained additional putatively heterozygous sites over and above the heterozygous sites identified during shotgun haplotyping. a priori, these could represent false heterozygotes resulting from PCR errors present in a single clone but absent from the genomic template, or true heterozygotes overlooked during shotgun haplotyping. There was no evidence that these putatively heterozygous sites fell into regions of low sequence coverage, as might be expected of missed heterozygotes. We sequenced the original PCR product across all 11 putatively heterozygous sites apparent in the clone sequences of T1. All 11 sites were definitively homozygous (see Figure 4). Therefore, all of these putatively heterozygous sites appear to be PCR errors within single clones. This gives an estimate of PCR error rate (11 sites in 24 kb of cloned sequence) that accords closely with our independent estimate of one PCR error for every 2 kb.

DISCUSSION

We successfully applied shotgun haplotyping to the sequencing of 12 kb allelic sequences at two autosomal locations. We corroborated these haplotypic sequences by checking them against sequences from clones derived from the same source PCR product. The heuristic haplotyping algorithm we developed was able to successfully assemble haplotype sequences for both loci.

A priori we identified four possible problems with this methodology, but none of these proved to be a major issue. An average sequence coverage of 24× across a 12 kb sequence was sufficient to clearly identify all heterozygous sites, as allele-specific biases in sequence coverage were minimal. We developed criteria that distinguished well between true heterozygous sites and PCR errors. Furthermore, chimeric read-pairs were too low in frequency to confound attempts to phase these heterozygous sites, although rates of jumping PCR might vary between genomic loci. Finally, the distribution of heterozygous sites within both test loci was sufficient to phase both haplotypes contiguously.

In principle, the maximum distance between neighbouring heterozygous sites that we should be able to phase across is equal to the size of the largest clone inserts in our shotgun haplotyping library. In order to ensure contiguous haplotyping across 12 kb, we expanded the size range of these inserts up to 4.5 kb, beyond that typically used in shotgun sequencing. We observed a higher density of heterozygous positions in both of our test loci than the genome average of one every 1.2 kb. This means that the observed distances between heterozygous positions are shorter than the genome average (Supplementary Figure 1). However, simulations indicate that for the genome average nucleotide diversity of 0.000751 (24) we should expect distances between neighbouring heterozygous sites to exceed 4.5 kb only 3.4% of the time. On the rare occasion that a distance between neighbouring heterozygous sites is beyond the limit of shotgun haplotyping, then a variety of methods could be used to link the dislocated haplotypes. For example, double allele-specific PCR (25) between the nearest sites in the two haplotype segments should reveal the remaining phase information. Alternatively, two sequence reads across the same sites on a single clone would also allow the two segments to be phased.

Here, shotgun haplotyping has been applied to long PCR amplifications of diploid loci, but in principle it is applicable to any diploid sequence template. Long PCR is typically able to amplify up to 20 kb in complex genomes. No prior information is required on the sequence within the amplified product. As such it is an ideal method for investigating sequence variation at loci about which little is known, e.g. within gaps in genome sequences, or in organisms in which no genomic sequence information is available.

Table 1. Sequence coverage and haplotypes identified at two test loci

(a) Haplotype information for region T1

No. of successful reads	Average coverage[a]	No. of variants	No. of PCR errors	No. of reads assigned to each haplotype	No. of chimeras	Allele	408	638	1293	2241	2869	2946	3248	3441	4059	4197	4499	5970–5907
558	34×	27	11	242/262	6	A1	G	G	A	––––	A	C	T	T	T	A	C	G
						A2	C	A	G	ATTT	T	A	C	C	C	T	T	A
Frequency of less frequently observed allele							40	48	47	34	50	48	49	45	46	50	45	44

Allele	6325–6328	7188–7191	7839–7842	7846–7849	7929–7932	8476–8479	8733–8736	8849–8852	9884–9887	9991–9994	10033–10036	10237–10240	10278–10281	10285–10288	10737–10740
A1	A	A	G	T	A	G	C	C	A	T	C	T	C	G	G
A2	G	C	A	C	T	A	T	T	G	C	T	C	T	A	A
	43	47	50	46	48	48	42	41	46	42	36	38	27	37	20

(b) Haplotype information for region T2

No. of successful reads	Average coverage[b]	No. of variants	No. of PCR errors	No. of reads assigned to eachhaplotype	No. of chimeras	Allele	656	1182	1244	3776	3832	4942	5104	5126	5141	5165	5573	5681	6678	6864	7309	7863	7873	8005
598	24×	38	14	220/199	3	A1	C	T	TTAT	C	T	T	T	T	T	C	G	A	A	T	G	C	G	G
						A2	T	C	––––	T	C	C	C	C	C	T	A	G	G	C	A	T	T	A
Frequency of less frequently observed allele							39	43	47	47	42	49	48	48	48	47	37	30	48	44	44	38	38	42

Allele	8122	8387	8396	8517	8548	9016	9180	9208	9360	9510	10401	10972	10982	11224	11239	12554	12585	12862	12960–12956	12964–12960
A1	T	T	G	-	A	A	G	T	G	C	T	C	A	C	TGAG	T	T	T	T	C
A2	A	C	A	T	G	G	A	A	C	T	C	A	G	T	––––	C	C	C	C	G
	43	40	40	40	44	42	50	50	36	40	47	40	40	42	45	38	38	41	42	42

[a]Average read length 735 bp.
[b]Average read length 496 bp.

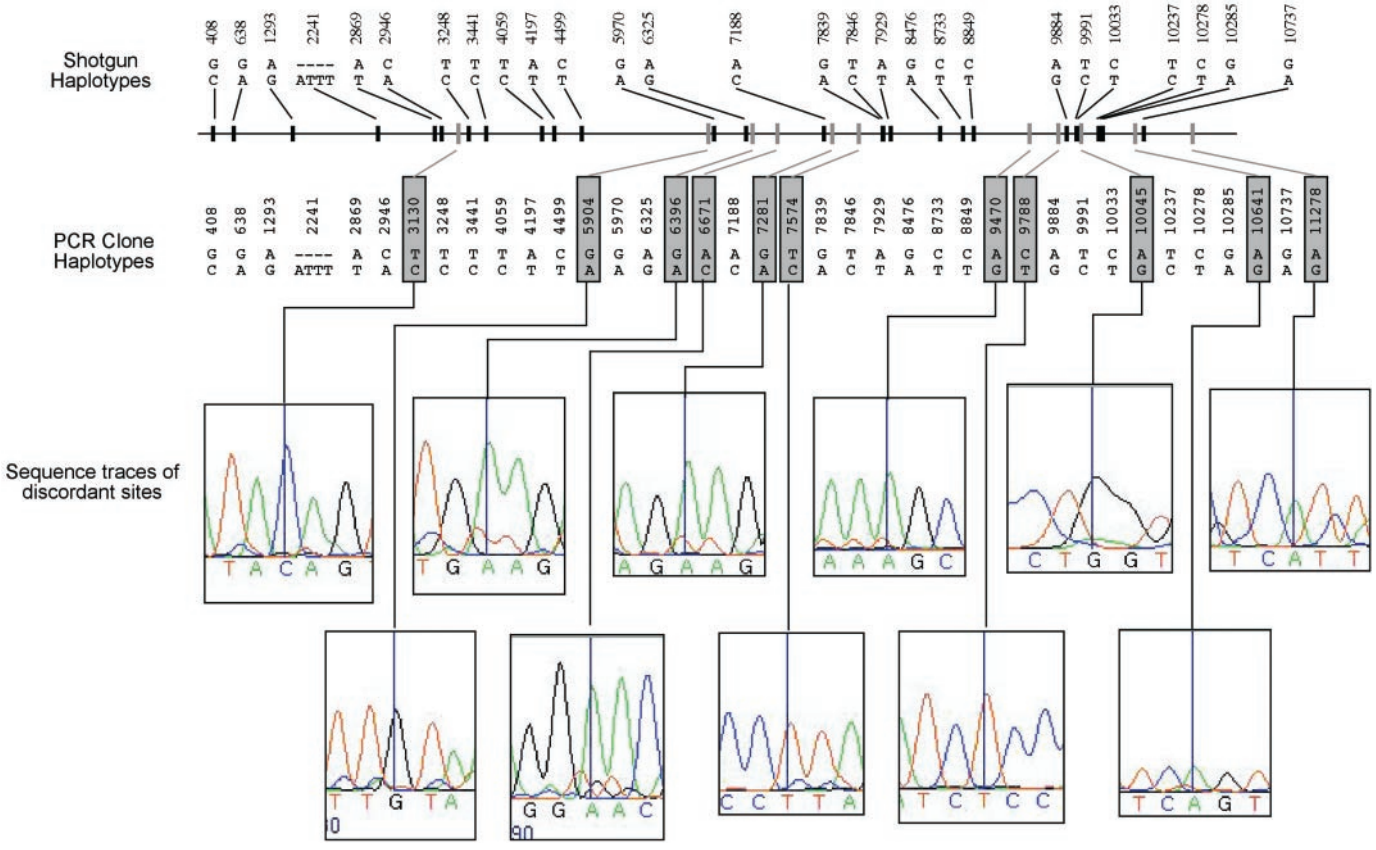

Figure 4. Comparison of shotgun haplotypes and PCR clone haplotypes. Heterozygous sites identified during the shotgun haplotyping of T2 are displayed above the apparent heterozygous sites identified between two sequenced PCR clones of the same alleles. The positions of the apparent variant sites within the amplicon are displayed on the line between the shotgun and PCR clone haplotypes. Allelic states on the same row indicate the haplotypes that were identified by our phasing algorithm. The positions of heterozygous sites within the sequence assembly are given above each site. The 11 sites (shown in grey) identified in the clone sequences but not the shotgun haplotyping were subsequent sequenced across on the raw PCR product from the same individual. All 11 discrepant sites appear to be homozygous in the PCR product, indicating that their presence in the PCR clone sequences reflects clone-specific PCR errors. Supplementary Figure 1: Distribution of the distances between neighbouring heterozygous sites observed in our validation experiments compared to a simulation based on the genome average nucleotide diversity. A histogram showing the distribution of distances between the heterozygous sites we identified in our two test loci is compared to a line indicating the distribution expected given a nucleotide diversity of 0.000751, the genome-wide average (24). As expected given the higher density of heterozygous sites identified in our test loci, the observed distribution is skewed towards shorter distances between variant sites. Supplementary Figure 2: Screenshot of the Graphical User Interface used to inspect the automatic phasing algorithm. The front panel displays (from left to right) the sequence coverage across the assembly, the location of apparent heterozygous sites, the number of instances of the different alleles and the read-pair clusters in which each allele is found. Individual sites can be eliminated from the analysis and the phasing algorithm re-run. The panel behind shows the assembled read-pairs sorted by read-pair cluster.

Shotgun haplotyping identifies all sequence variation at a locus, including both base substitutions and small indels. It is therefore applicable to projects where it is desirable to characterize all variation exhaustively, especially studies that seek to contrast levels of variation between loci. In common with all other methods that seek to discover genetic variation in a population by analysing one individual at a time, a large number of individuals must be analysed to capture the majority of the variant sites in the population.

Our novel method presents the opportunity to collect haplotype information from a panel of individuals using a robust and scaleable pipeline. One of the major advantages is that once a robust long PCR amplification has been optimized, it is quick and easy to obtain comparative data from a panel of individuals. Given the sequence coverage required, the consumables costs are appreciable. However, this is cancelled out by lower labour costs compared to lower throughput protocols such as long PCR cloning. Furthermore, the cost of sequencing is falling, whereas that of labour is rising. As we have shown, sequences of cloned PCR products are susceptible to PCR errors. Data of comparable accuracy to those obtained by shotgun haplotyping could only be generated from a cloning-based project by completely sequencing four cloned long PCR products, two for each allele. Moreover, the pooling of overlapping PCR products from the same individual prior to shotgun haplotyping should allow the generation of haplotypic sequences longer than those amplified in a single PCR.

How does our method compare to recent impressive developments in single molecule haplotyping (26)? Linking emulsion PCR (LE-PCR) (27), M1-PCR (5) and PCR colonies ('polonies') (28) are all single molecule haplotyping methods that are capable of generating empirical (as opposed to inferred) haplotypes over similar physical distances to shotgun haplotyping. However, all of these methods require that the variable sites to be haplotyped are known before performing

the experiments, and do not exhaustively characterize the sequence variation within the genomic segment being investigated. A more relevant comparison might be with recent developments in highly-parallel single molecule sequencing (29,30). The short (6–120 bp) reads generated with current protocols do not lend themselves to haplotyping applications, however, the prospect of using paired reads is encouraging (29), although the ability to generate a broad range of physical distances between paired reads will be critical for haplotyping applications.

How does our method compare to the statistical inference of haplotypic sequences from genotypic sequence data? The statistical approach relies on the ability to identify heterozygous sites within diploid sequence traces, and error rates can vary wildly depending on the size of the locus, the number of individuals genotyped and the haplotype diversity (8,31). There is no gold standard method for the automatic identification of heterozygous bases in diploid sequence traces: all methods come with associated error rates. Moreover, the accuracy of statistical haplotype inference declines as fewer sequences of the same locus are analysed. Shotgun haplotyping occupies a niche in comparative sequencing methods by generating accurate haplotypic sequences across all population sizes. Furthermore, methods of statistical haplotype inference often include explicit models of sequence evolution and consequently will be compromised when studying variation at a locus at which unusual evolutionary processes (e.g. gene conversion) operate. We are applying shotgun haplotyping to studies of sequence variation in segmental duplications, where gene conversion is known to occur (S. J. Lindsay and M. E. Hurles, manuscript in preparation).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank David Willey and Bob Plumb for technical assistance, and Mark Jobling for the gift of genomic DNA. We would also like to thank Stephan Beck and Mark Jobling for critical comments on an earlier version of the manuscript. This work was funded by the Wellcome Trust. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Douglas,J.A., Boehnke,M., Gillanders,E., Trent,J.M. and Gruber,S.B. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genet*., **28**, 361–364.
2. Yan,H., Papadopoulos,N., Marra,G., Perrera,C., Jiricny,J., Boland,C.R., Lynch,H.T., Chadwick,R.B., de la Chapelle,A., Berg,K. *et al*. (2000) Conversion of diploidy to haploidy. *Nature*, **403**, 723–724.
3. Martinez-Arias,R., Bertranpetit,J. and Comas,D. (2002) Determination of haploid DNA sequences in humans: application to the glucocerebrosidase pseudogene. *DNA Seq*., **13**, 9–13.
4. Burgtorf,C., Kepper,P., Hoehe,M., Schmitt,C., Reinhardt,R., Lehrach,H. and Sauer,S. (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res*., **13**, 2717–2724.
5. Ding,C. and Cantor,C.R. (2003) Direct molecular haplotyping of long-range genomic DNA with M1-PCR. *Proc. Natl Acad. Sci. USA*, **100**, 7449–7453.
6. Ruano,G., Kidd,K.K. and Stephens,J.C. (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc. Natl Acad. Sci. USA*, **87**, 6296–6300.
7. Michalatos-Beloin,S., Tishkoff,S.A., Bentley,K.L., Kidd,K.K. and Ruano,G. (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res*., **24**, 4841–4843.
8. Yu,C.E., Devlin,B., Galloway,N., Loomis,E. and Schellenberg,G.D. (2004) ADLAPH: A molecular haplotyping method based on allele-discriminating long-range PCR. *Genomics*, **84**, 600–612.
9. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nature Rev. Genet*., **5**, 335–344.
10. Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol*., **7**, 111–122.
11. Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol*., **12**, 921–927.
12. Stephens,M., Smith,N.J. and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet*., **68**, 978–989.
13. Wang,L. and Xu,Y. (2003) Haplotype Inference by maximum parsimony. *Bioinformatics*, **19**, 1773–1780.
14. Halperin,E. and Eskin,E. (2004) Haplotype reconstruction from genotype data using Imperfect Phylogeny. *Bioinformatics*, **20**, 1842–1849.
15. Nickerson,D.A., Tobe,V.O. and Taylor,S.L. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*., **25**, 2745–2751.
16. Weckx,S., Del-Favero,J., Rademakers,R., Claes,L., Cruts,M., De Jonghe,P., Van Broeckhoven,C. and De Rijk,P. (2005) novoSNP, a novel computational tool for sequence variation discovery. *Genome Res*., **15**, 436–442.
17. Adkins,R.M. (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genet*., **5**, 22.
18. Bonizzoni,P., Della Vedova,G., Dondi,R. and Li,J. (2003) The Haplotyping Problem: An Overview of Computational Models and Solutions. *J. Comput. Sci. Technol*., **18**, 675–689.
19. Lippert,R., Schwartz,R., Lancia,G. and Istrail,S. (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief Bioinform*., **3**, 23–31.
20. Staden,R., Beal,K.F. and Bonfield,J.K. (2000) The Staden package, 1998. *Methods Mol. Biol*., **132**, 115–130.
21. McMurray,A.A., Sulston,J.E. and Quail,M.A. (1998) Short-insertlibraries as a method of problem solving in genome sequencing. *Genome Res*., **8**, 562–566.
22. Bonfield,J.K. and Staden,R. (1995) The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res*., **23**, 1406–1410.
23. Tammi,M.T., Arner,E., Britton,T. and Andersson,B. (2002) Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. *Bioinformatics*, **18**, 379–388.
24. Sachidanandam,R., Weissman,D., Schmidt,S.C., Kakol,J.M., Stein,L.D., Marth,G., Sherry,S., Mullikin,J.C., Mortimore,B.J., Willey,D.L. *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
25. Jeffreys,A.J. and Neumann,R. (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genet*., **31**, 267–271.
26. Kwok,P.Y. and Xiao,M. (2004) Single-molecule analysis for molecular haplotyping. *Hum. Mutat*., **23**, 442–446.
27. Wetmur,J.G., Kumar,M., Zhang,L., Palomeque,C., Wallenstein,S. and Chen,J. (2005) Molecular haplotyping by linking emulsion PCR: analysis

of paraoxonase 1 haplotypes and phenotypes. *Nucleic Acids Res*., **33**, 2615–2619.
28. Mitra,R.D., Butty,V.L., Shendure,J., Williams,B.R., Housman,D.E. and Church,G.M. (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl Acad. Sci. USA*, **100**, 5926–5931.
29. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
30. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
31. Niu,T. (2004) Algorithms for inferring haplotypes. *Genet. Epidemiol*., **27**, 334–347.