nature genetics

Automating sequence-based detection and genotyping of SNPs from diploid samples

Matthew Stephens[1,2], James S Sloan[2], P D Robertson[2], Paul Scheet[1] & Deborah A Nickerson[2]

The detection of sequence variation, for which DNA sequencing has emerged as the most sensitive and automated approach, forms the basis of all genetic analysis. Here we describe and illustrate an algorithm that accurately detects and genotypes SNPs from fluorescence-based sequence data. Because the algorithm focuses particularly on detecting SNPs through the identification of heterozygous individuals, it is especially well suited to the detection of SNPs in diploid samples obtained after DNA amplification. It is substantially more accurate than existing approaches and, notably, provides a useful quantitative measure of its confidence in each potential SNP detected and in each genotype called. Calls assigned the highest confidence are sufficiently reliable to remove the need for manual review in several contexts. For example, for sequence data from 47–90 individuals sequenced on both the forward and reverse strands, the highest-confidence calls from our algorithm detected 93% of all SNPs and 100% of high-frequency SNPs, with no false positive SNPs identified and 99.9% genotyping accuracy. This algorithm is implemented in a software package, PolyPhred version 5.0, which is freely available for academic use.

The detection and genotyping of sequence variations lies at the core of genetic analysis; in addition, all approaches to disease gene mapping ultimately lead to variation discovery across a candidate region or gene to identify genetic variants that affect the trait of interest. Among the numerous methods developed to identify genetic variants (reviewed in ref. 1), DNA sequencing has emerged as the most sensitive and automated[2,3].

Sequence-based approaches to detecting variants can be divided into two groups: those based on detecting sequence differences among 'cloned' DNA samples, which have provided the main source of SNPs currently in dbSNP; and those based on identifying the presence of more than one base in PCR-amplified 'diploid' samples, which have been the preferred approach for large-scale human resequencing projects that aim comprehensively to identify and genotype all variants in targeted genomic regions in a sample of individuals[4]. In both types of approach, the identification of SNPs has been greatly aided by the use of computational and statistical methods.

Methods for detecting SNPs from cloned samples (such as PolyBayes[5] and ssahaSNP[6]) have focused on the use of quality scores that quantify the expected accuracy of each base call[7] to distinguish between genuine differences among sequences and sequencing errors. Methods for detecting SNPs from diploid samples (such as PolyPhred[2] and novoSNP[8]) primarily focus on identifying heterozygous genotypes, which are characterized by the presence of two peaks at a single position in a sequence trace, each roughly half as high as the corresponding peak in a homozygote[9]. Reliably identifying heterozygotes is a key part of SNP detection in diploid samples because, for most SNPs, moderate-sized samples will not include homozygotes for both alleles. Existing algorithms are not, however, sufficiently robust to be used without potentially costly confirmation, either by manual review or by genotyping with an alternative technology.

Here we present a more accurate method to detect and genotype SNPs in sequence traces obtained from diploid DNA samples. The algorithm, implemented in a software package, PolyPhred v5.0, improves on existing approaches in two key ways. First, it takes a more detailed account of systematic variation in peak heights caused by read-specific and sequence-context effects[10], thereby facilitating accurate identification of heterozygotes. Second, it computes a formal statistical measure of the evidence for potential genotypes at each position in each sequencing read. This enables the application of standard statistical methods to combine evidence efficiently across multiple reads for an individual, resulting in exceptional accuracy for 'double-coverage' data (where individuals are sequenced on both the forward and reverse strands). It also provides a quantitative assessment of the confidence in each SNP identified and in each genotype called. This feature is particularly useful for identifying a subset of highly accurate SNP and genotype calls, similar to Phred[11], whose quantitative approach to DNA base-calling revolutionized genome sequencing.

RESULTS

Current sequencing pipelines produce chromatograms (**Fig. 1**) that are used to call bases and to compute a quality score that quantifies the expected accuracy of each base call. The sequences are then assembled (aligned) to a reference sequence and compared. Our algorithm first uses these base calls to identify the 'consensus' (most common) base at each position and then searches for reads that

[1]Department of Statistics and [2]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. Correspondence should be addressed to M.S. (stephens@stat.washington.edu).

Received 16 October 2005; accepted 12 January 2006; published online 19 February 2006; doi:10.1038/ng1746

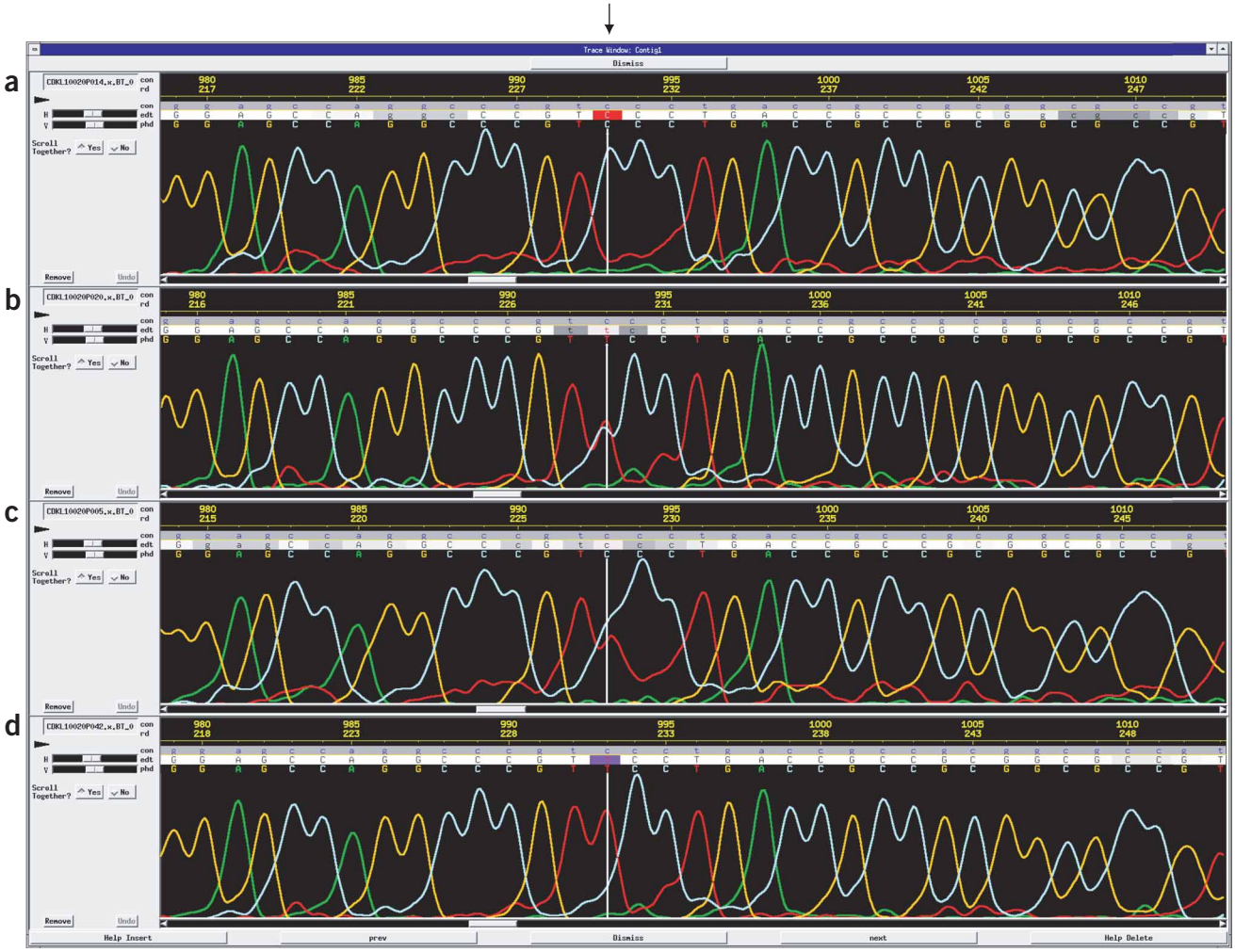

Figure 1 Sequence traces (chromatograms) for four individuals. Shown is a screen shot from the sequence finishing software Consed[17] for four individuals: P014 (**a**), P020 (**b**), P005 (**c**) and P042 (**d**). Consensus position 993 (arrow) is a SNP with consensus base C and minor allele T. The genotypes of P014, P020, P005 and P042 are CC, CT, CT and TT, respectively. The trace for the heterozygote P020 (**b**) shows a clear drop in peak in the C trace and a clean peak in the T trace. By contrast, the trace for the heterozygote P005 (**c**) shows a smaller drop in peak in the C trace and the quality of the trace is lower overall, making the genotype slightly harder to call (see **Fig. 2**).

appear not to be homozygous for the consensus base. A key step in this search is the computation of a series of 'residuals' that measure, at each position in each read, the difference between the observed peak height and the expected peak height for each possible genotype (homozygous for the consensus base, heterozygous or homozygous for a non-consensus base), taking into account sequence-context and read-specific effects (**Fig. 2** and **Supplementary Fig. 1** online). Together with other sequence features including the Phred quality scores, these residuals are used to compute 'log likelihood ratios' (LLRs) that measure the relative strength of the evidence for each possible genotype in each read. For each site in the assembly, our algorithm combines the evidence across all available reads to compute an overall 'score' (0–99) that reflects the strength of the evidence that the site is a SNP. The algorithm then outputs a list of potential SNPs (positions whose score exceeds a user-specified threshold), together with their overall scores and a genotype call for each individual. Each individual genotype call is also assigned a separate score (0–99) that estimates the percentage probability that the call is correct, conditional on the site being a SNP (**Supplementary Fig. 2** online). For example, all three SNPs highlighted in **Figure 2** are assigned an overall score of 99. The individual genotype calls for the six heterozygous positions highlighted are also assigned scores of 99, except for individual P005 at position 993, which received score 96. This lower score reflects the noisier sequence and reduced peak drop in the read for that individual (**Fig. 1c**). See Methods for details.

We assessed algorithmic performance by comparing calls from our algorithm with manually confirmed calls obtained from the same sequence data in a local reference database (Methods). We performed comparisons for two sequencing designs (**Supplementary Fig. 3** online): a 'tiled' design, in which amplicons of ~900 bp are sequenced from each end (forward and reverse strands) to provide partially overlapping sequences; and a 'double-coverage' design in which amplicons of ~500 bp are sequenced from each end to provide completely overlapping sequences (in the absence of extensive mononucleotide sequence tracts or insertion-deletion (indel) polymorphisms). In each approach, adjacent amplicons overlap each other by ~100 bp.

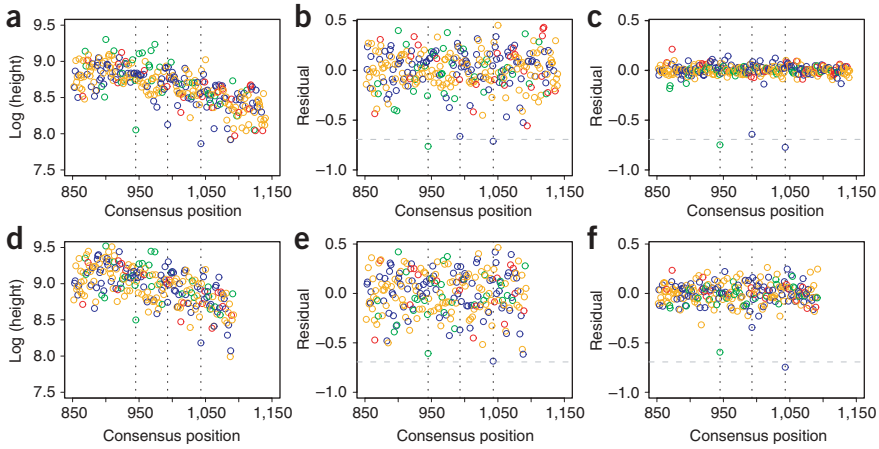

Figure 2 Removal of systematic variation in peak height improves discrimination between heterozygotes and homozygotes. Colors indicate the consensus base at each position (A = green, C = blue, G = orange, T = red). (**a**) Logarithm of the height of the consensus base peak (*y* axis) plotted against position (*x* axis) in a read from individual P020 in **Figure 1**. Heterozygous sites (vertical dotted lines) are characterized by consensus peaks that are smaller than average, but the signal is partially obscured by systematic variation in peak height caused by position in the read and by sequence context. (**b**) Residuals obtained by subtracting from each log (height) value in **a** the average log (height) values of nearby consensus peaks of the same base. Horizontal dashed line indicates the expected value of the residual for heterozygous sites, log (0.5), because peak heights of heterozygotes are expected to be half as big as those of homozygotes). Removing the read-specific effects in this way makes the heterozygotes stand out more clearly. (**c**) Final residuals, obtained by subtracting from each residual in **b** the average value of the corresponding residual in other reads at the same position (using a crude filtering criterion to attempt to include only consensus homozygotes in this average). This corrects for the tendency for the height of the consensus peak at each position, relative to nearby peaks corresponding to the same base, to be consistent across reads, and further enhances the distinction between consensus allele homozygotes and heterozygotes. (**d–f**) Panels **a–c** repeated for a read from individual P005. This read is substantially noisier (the residuals in **f** vary more than in **c**); thus, the heterozygous sites stand out less clearly.

For the tiled design, we used sequence assemblies across 90 individuals for 20 candidate genes, containing a total of 3,092 identified SNPs. For the double-coverage design, we analyzed six genes containing 328 identified SNPs. To augment these data from the double-coverage design, we also examined the performance at 'double-covered' positions in the tiled-design genes, which contained 760 SNPs (see Methods for more details).

Accuracy of SNP detection

We assessed the accuracy of SNP detection by comparing sites marked by our algorithm as potential SNPs, at various score thresholds, with the database of SNPs. We summarize performance at each threshold with two numbers: the proportion of SNPs in the database that were not marked by the method (in other words, the estimated 'false negative' rate, which we term the 'missed SNP rate'), and the proportion of SNPs marked by the method that were not in the database (in other words, one minus the estimated positive predictive value or the 'false discovery rate'). We used the false discovery rate, rather than the false positive rate, because it has a direct interpretation in terms of the reliability of the set of potential SNPs identified by the method.

Figure 3a,b shows how these two performance measures vary with threshold, both for our algorithm and for the most recent previous version of PolyPhred (v4.29). In all data sets, at a given missed SNP rate our algorithm substantially reduced the corresponding false discovery rate as compared with PolyPhred v4.29. For example, for a missed SNP rate of 0.03, the estimated false discovery rates for our algorithm versus PolyPhred v4.29 were 0.20 versus 0.87 (for the tiled design), 0.05 versus 0.63 (for the double-coverage design), and 0.02 versus 0.65 (for double-covered positions from the tiled design). In addition, for genes with double coverage, at a score threshold of 99 only 7% of SNPs were missed and there were essentially no false discoveries. (Seven positions received a score of 99 but were not in the database of SNPs; however, subsequent re-examination of the

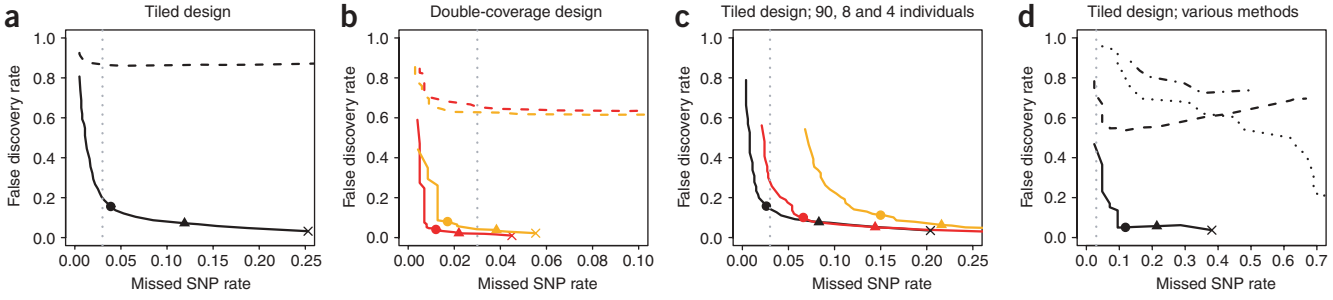

Figure 3 Missed SNP rate versus false discovery rate for different data sets. Shown are results from our algorithm with varying threshold values (1–99); results for thresholds of 70 (circle), 95 (triangle) and 99 (cross) are highlighted. Note that the scale of the *x* axis varies across panels; for reference, the vertical gray dotted line in each indicates a missed SNP rate of 3%. (**a**,**b**) Comparison of our algorithm (unbroken line) and PolyPhred v4.29 (dashed line) for the tiled-design genes (**a**), and for the double-coverage genes (orange) and the tiled-design genes restricted to double-covered positions (red; **b**). (**c**) Variation in accuracy of our method with assembly size. Colors correspond to assemblies of 90 (black), 8 (red) and 4 (orange) individuals. (**d**) Comparison of accuracy of our method (unbroken line) with PolyPhred v4.29 (dashed line), novoSNP (dotted line) and Mutation Surveyor (dotted-dashed line), on two tiled-design genes with assemblies of eight individuals.

Table 1 Frequency spectrum of 'missed', 'found' and 'all' SNPs in the 20 tiled-design genes

	Minor allele frequency				
	0.01	0.02	0.03	0.04	≥0.05
Missed SNPs	0.84	0.11	0.02	0.02	0
Found SNPs	0.49	0.09	0.05	0.03	0.34
All SNPs	0.5	0.09	0.05	0.03	0.33

Missed SNPs are real SNPs assigned a score of <70 by our method; found SNPs are real SNPs assigned a score of ≥70.

chromatograms showed that six of these positions were real SNPs and the seventh was within a polymorphic microsatellite.) Notably, the SNPs missed by our algorithm have a low minor allele frequency (MAF; average MAF of missed SNPs at a threshold of 70 ≈ 0.01; **Table 1**). The algorithm misses so few SNPs, however, that the frequency spectrum of marked SNPs differs little from that of all SNPs (**Table 1**).

Effects of errors in the database

Two types of potential error in the database could influence the estimated error rates: first, the database could contain some positions that are not SNPs; second, some SNPs could be missing from the database, such as SNPs missed by previous versions of PolyPhred, SNPs in particularly difficult regions (for example, within microsatellites) because these are deliberately omitted from the database, and SNPs that were (mistakenly) not confirmed by the analyst.

To assess how these errors influence estimated error rates, we visually examined positions in the chromatograms where calls from our algorithm disagreed with the database. First, for the 20 genes sequenced by tiled design, we examined the 120 SNPs in the database that were 'missed' by our algorithm (defined here as SNPs assigned a score of <70) to see whether some of these might not be SNPs. On re-inspection of the chromatograms, only three appeared not to be real SNPs, suggesting that the database contains few errors of the first type described above. (These three errors represent ~0.1% of all SNPs in the database.)

Second, for two tiled-design genes and all six double-covered genes, we examined the 78 positions not in the database that our method marked as possible SNPs (those assigned a score of ≥70). From visual inspection of the chromatograms, we classified each position as 'probably a SNP', 'real variation, but microsatellite or indel rather than SNP', or 'probably not a variant'. Of the 78 positions, 22 (28%) were classified as probably a SNP, 22 were probably not a real variant and the remainder were associated with indels or microsatellite variation. The 22 (apparently) real SNPs missing from the database represent ~3% of all SNPs originally detected in these genes.

In summary, errors in the database seem to have negligible effect on the estimated missed SNP rates but may substantially inflate the estimated false discovery rates (by a factor of 3–4 at a threshold of 70, if positions that are marked as potential SNPs owing to variation in microsatellites or indels are not considered false discoveries). For this reason, **Figure 3** probably understates the actual performance of the algorithm.

Performance on smaller samples

We took random (nested) subsets of four and eight individuals from the 90 individuals sequenced across nine of the tiled-design genes (the first nine genes), and applied our method to these smaller assemblies. In these new alignments, we manually identified the positions of those SNPs in the database that were polymorphic in the reduced samples and estimated error rates as before. We found that SNP detection was only slightly less accurate in eight individuals than in 90 individuals, but there was a notable reduction in accuracy when data from only four individuals was used (**Fig. 3c**). This reduction in accuracy may reflect the fact that our approach uses consensus homozygotes as a standard against which to compare other reads, which is expected to work well provided that there are at least several such homozygotes (which will be the case for most SNPs in samples from eight individuals, but not in samples from four individuals).

Comparison with other methods

We compared performance of our algorithm with two other heterozygote detection algorithms: novoSNP v2.0.3 (ref. 8) and Mutation Surveyor v2.61 (Softgenetics). Owing to limitations in the size of the assemblies that we could analyze by these methods, we obtained results for only two genes in assemblies of eight individuals. Our algorithm substantially outperformed the others on these data (**Fig. 3d**). For example, for a missed SNP rate of 0.15 the false discovery rates were 0.05 (our algorithm), 0.55 (PolyPhred v4.29), 0.70 (novoSNP) and 0.85 (Mutation Surveyor).

Dependence of performance on sequence quality

The performance of our algorithm, and indeed any algorithm, for SNP detection will depend on the quality of the sequence data being analyzed. To examine this dependence, we computed the false-positive and missed SNP rates at a score threshold of 70 as a function of the average Phred quality[11] of the available reads at each position for the 20 tiled-design genes (**Fig. 4**). Caution is necessary in interpreting the estimated error rates for data of very low quality, because typically it will be hard even for human experts to discern the 'truth' from low-quality data; thus, real SNPs in such regions are likely to be missing from the database. In addition, **Figure 4** was produced without distinguishing between positions for which data on both the forward and the reverse strands were available (which in the tiled design tend to occur at the ends of reads, where data quality is lower), and those for which data on only one strand were available. Nonetheless, **Figure 4** suggests that an average quality of roughly 30 is sufficient for the algorithm to achieve near-optimal performance and that performance degrades relatively sharply for lower quality data.

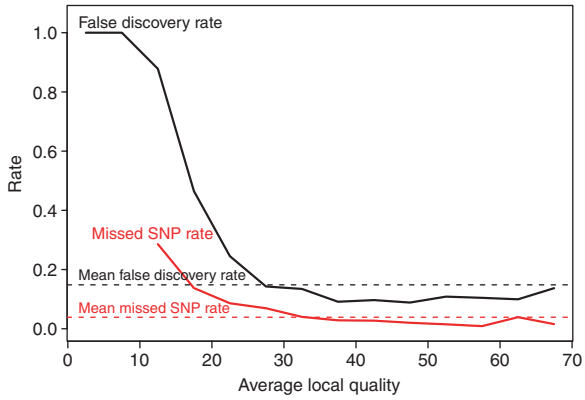

Figure 4 Dependence of performance on sequence quality. Unbroken lines show how the false positive rate and missed SNP rate for our algorithm, at a score threshold of 70, vary as a function of the average (across reads) of the Phred quality score (see Methods). Horizontal dotted lines show the corresponding mean values of these rates for our data, across all positions, irrespective of quality.

Table 2 Proportion of genotype calls from the algorithm that agreed with reference calls

	All genotype calls			High-confidence genotype calls		
Data set	All SNPs	High frequency	Low frequency	All SNPs	High frequency	Low frequency
20 tiled-design genes	99.7	99.2	>99.9	99.9	99.7	>99.9
Six double-covered genes	99.8	99.6	>99.9	99.9	99.8	>99.9
Perlegen-typed SNPs	99.2	99	99.9	99.8	99.7	>99.9

High frequency indicates SNPs with a MAF of ≥0.05; low frequency indicates other SNPs. High-confidence genotype calls indicates genotype calls assigned a score of 99 by the algorithm (see text for details).

Accuracy of genotyping

We compared genotype calls from our algorithm with three reference sets of SNP genotypes: (1) 253,735 genotype calls from sequence data for the tiled-design genes; (2) 17,451 genotype calls from sequence data for the double-coverage genes; and (3) 68,384 genotypes that were called identically by two independent platforms, namely, from sequence data using earlier versions of PolyPhred followed by manual review, and from a genotyping platform developed by Perlegen Sciences[12]. This last set of genotypes comes from 47 individuals typed at ~1,900 SNPs in 120 autosomal genes sequenced by Seattle-SNPs. The ~1,900 SNPs are a subset of the SNPs detected in the project that were subsequently genotyped by Perlegen. They are enriched for high-frequency SNPs because Perlegen preferentially selected high-frequency SNPs.

Overall, calls from our algorithm agreed with 99.2–99.8% of the reference calls. For high-confidence genotype calls (those assigned a score of 99 by the algorithm) the rate of agreement rose to 99.8–99.9% (**Table 2**). The increased accuracy of the high-confidence calls comes at the cost of not making genotype calls for 4–6% of genotypes (see **Supplementary Fig. 4** online, which also shows the trade-off between genotyping accuracy and missing genotype calls as the threshold for accepting calls is varied). Agreement rates were higher at low-frequency SNPs than at high-frequency SNPs. One reason why low-frequency SNPs are easier to genotype is that, at a very low frequency SNP, simply calling every genotype a major allele homozygote achieves a low genotyping error rate. The added difficulty of accurately genotyping high-frequency SNPs partly explains the lower overall agreement rates with the Perlegen calls (reference set (3)), because this data set is enriched for high-frequency SNPs.

The high rate of agreement between calls from our algorithm and the reference calls suggests that our algorithm is highly accurate. Even assuming that all differences between the algorithm and reference calls reflect errors in the algorithm calls, the agreement rates for high-confidence calls suggest genotyping error rates of 0.1–0.3%, representing a several-fold reduction in error rates as compared with PolyPhred v4.29. For example, both high- and low-confidence calls in PolyPhred v4.29 used without manual review produced genotyping error rates of >1.2% among high-frequency SNPs (**Supplementary Fig. 4** online).

DISCUSSION

With its decreasing costs and rapidly expanding scale, DNA sequencing seems likely to continue to have a key role in genetic analysis. The identification of genomic regions associated with biologically or medically important phenotypes will inevitably be followed by sequencing of these regions in appropriate individuals. In addition, genome-wide resequencing of a moderate number of individuals in multiple populations seems to be a natural follow-up to the current International HapMap Project[13], as a further step towards comprehensively cataloguing human genetic variation.

The algorithm that we describe represents a considerable advance in methods for SNP detection and genotyping and can greatly facilitate, and in some cases automate, such efforts. For genes with double coverage we found that our algorithm, used with a high score threshold (99), can reliably identify and genotype SNPs without manual intervention. In some contexts, the missed SNP rate for this strategy (~7% in the genes that we analyzed) may be acceptable, particularly because most missed SNPs had a MAF of ~1%. In other contexts such as mutation detection, it may not be acceptable; and in clinical settings, manual review of key calls seems likely to continue to be standard practice. Nonetheless, the ability to assign confidence scores that distinguish between high and lower quality calls marks a very important step forward in the automation of sequence-based SNP detection and genotyping.

Our data characterize the performance of the algorithm under various conditions, providing a useful guide to issues of design and choice of threshold. For example, the very high accuracy for double-coverage data suggests that, in many contexts, the reduction or elimination of manual review could more than compensate for the increased cost of producing double-coverage data. Regarding the choice of threshold, we observe that most positions assigned score of <70 are either not real SNPs or have sequence data that are not sufficiently informative even for human experts to make a confident call. Thus, although choice of suitable threshold will inevitably be project-specific, a threshold of 70 may be appropriate for projects that aim to identify a large proportion of all SNPs with only a moderate amount of manual review.

Despite the gains in performance, some limitations remain. The algorithm's performance on small assemblies (four or fewer individuals) could be improved, although even here it considerably outperforms other available methods. Other minor improvements to the algorithm should also be possible, including replacing some of the simple *ad hoc* filters that are currently used with more sophisticated approaches. Finally, the algorithm takes no account of the locations of microsatellite loci and makes no attempt to identify or genotype indels, which represent roughly 6% of all biallelic variation[14]. To rectify this, we have implemented in the software an optional filter that allows the user to 'demote' scores assigned to positions that appear to fall within poly tracts or microsatellites, and we are currently working to extend the algorithm to provide an automatic method to identify and genotype indels. Taken together, these measures should further improve SNP detection accuracy, because a substantial proportion of the few remaining 'false discoveries' are due to indels or microsatellites.

METHODS

Reference databases. The local reference database of manually confirmed SNPs and genotypes, used to assess algorithmic performance, was compiled by applying an earlier (and algorithmically different) version of PolyPhred[2] to

the same sequence traces, followed by manually reviewing the chromatograms at positions flagged by PolyPhred as possible SNPs (see ref. 4 for details). The 20 tiled-design genes were randomly selected from those sequenced as part of the Environmental Genome Project[15]. The genes analyzed were *ACTB*, *ALAD*, *BNIP2*, *ERCC3*, *GAD2*, *MDM4*, *MRE11A*, *MSH3*, *RPA4*, *TNFRSF1B*, *BIRC3*, *CTNND1*, *CYP2J2*, *FGF13*, *GCLC*, *HMOX2*, *HPRT1*, *MMP2*, *RAD21* and *TP53*. The six double-covered genes were *ODC1* and *ABP1*, sequenced across 90 individuals; and *FASLG*, *TNFSF13B*, *CD28* and *CD72*, sequenced across 47 individuals.

Criteria for a 'double-covered' position. In the tiled-design genes, we counted a position as double-covered if at least 50 of the 90 individuals had reasonable quality data on both strands. At most sites our definition of 'reasonable quality' was a Phred quality score, Q, of ≥ 25. Because heterozygote sites tend to have low a Q value[2], however, at known heterozygote positions i in a read we required that the average of Q at sites $i-2$ and $i+2$ was ≥ 25.

Quality-specific false-positive and missed SNP rates. To produce **Figure 4**, for each potential SNP assigned a score of ≥ 70 and for each real SNP, we computed the average, across reads, of the Phred quality score, Q, at that position (with the same adjustment of Q for heterozygous sites used above). We then grouped the positions into bins of average quality 0–5, 5–10, 10–15,…, 65–70. Within each bin we computed the false discovery rate as the number of positions in that bin assigned a score of ≥ 70 that were real SNPs, divided by the total number of positions in that bin assigned a score of ≥ 70. Similarly, we computed the missed SNP rate for each bin as the number of real SNPs in that bin assigned a score of <70 divided by the total number of real SNPs in that bin.

Algorithmic details. The algorithm has the following steps.

1. From an alignment of sequence data on multiple individuals, we identify a 'consensus base' (the most common base) at each site. (In our comparisons we used existing alignments that had been produced during the original resequencing project but, to mimic the use of fresh alignments, we removed all manually inserted tags identifying low-quality data.)

2. For each site in each read, we compute a measure of the strength of the evidence for each of three possible genotypes: homozygous with two copies of the consensus base (hom); heterozygous with one copy of the consensus base (het); or homozygous with no copies of the consensus base (min for 'minor allele homozygote'). We do this by using multiple logistic regression, with various relevant sequence features as explanatory variables, to compute two LLRs for each read:

$$\mathrm{LLR_{het}} = \log \frac{\Pr(\text{sequence features} \mid \text{genotype is het})}{\Pr(\text{sequence features} \mid \text{genotype is hom})}$$

and

$$\mathrm{LLR_{min}} = \log \frac{\Pr(\text{sequence features} \mid \text{genotype is min})}{\Pr(\text{sequence features} \mid \text{genotype is hom})}.$$

The sequence features used in the logistic regression include (functions of) (1) residuals that measure the deviation of peak height observed from that expected under each genotype class, taking into account read-specific effects and sequence-context effects (see **Fig. 2** for example); (2) the read-specific variance of such residuals, to account for differing peak-height variability among the sequence reads; and (3) Phred quality scores, which take account of the overall quality of the trace data, including features such as peak shape and spacing. The coefficients in the logistic regression were estimated using sequence traces and a list of known SNPs and genotypes for a training set of six genes from the Environmental Genome Project. (None of these six genes was included in the genes subsequently used to test the algorithms.)

3. If sequence data are available from both forward and reverse strands for an individual at a particular site, we sum $\mathrm{LLR_{het}}$ for the two strands and $\mathrm{LLR_{min}}$ for the two strands to obtain, for each individual $\overline{\mathrm{LLR}}_{\mathrm{het}}$ and $\overline{\mathrm{LLR}}_{\mathrm{min}}$, respectively, which measure the overall evidence in each individual's trace data for het and min genotypes relative to the hom genotype.

4. At each position we identify the largest values of $\overline{\mathrm{LLR}}_{\mathrm{het}}$ and $\overline{\mathrm{LLR}}_{\mathrm{min}}$ across all individuals, denoting these maxima $\widetilde{\mathrm{LLR}}_{\mathrm{het}}$ and $\widetilde{\mathrm{LLR}}_{\mathrm{min}}$, respectively. We assign the position a score,

$$S = 100(1 - 1/(1+(f_{\mathrm{het}}/f_{\mathrm{hom}})\exp(\widetilde{\mathrm{LLR}}_{\mathrm{het}})+(f_{\mathrm{min}}/f_{\mathrm{hom}})\exp(\widetilde{\mathrm{LLR}}_{\mathrm{min}}))),$$

where $f_{\mathrm{het}}/f_{\mathrm{hom}} = 0.00051$ and $f_{\mathrm{min}}/f_{\mathrm{hom}} = 0.00019$ are, respectively, the relative frequencies of het versus hom and min versus hom genotypes in the training set of six genes. We round scores down to the next integer, giving a range 0–99, and mark positions whose score exceeds (or equals) a user-specified threshold as potential SNPs. (The score S uses Bayes Theorem to compute the posterior probability that a (pseudo-) individual with LLRs $\widetilde{\mathrm{LLR}}_{\mathrm{het}}$ and $\widetilde{\mathrm{LLR}}_{\mathrm{min}}$ has a het or min genotype, using the relative frequency of each genotype in the training set as a prior. It gives high weight to positions where at least one individual is a clear het or a clear min, and even higher weight to sites where both het and min individuals appear to be present. We also experimented with other approaches that gave higher scores to positions where there appeared to be multiple heterozygote individuals, but we found that these were more susceptible to identifying false positive SNPs, presumably owing to systematic effects that we had not fully taken into account.)

5. At each potential SNP, we obtain an estimate, $\hat{f}$, of the putative MAF, f, by applying one iteration of the EM algorithm[16] for maximizing the likelihood under an assumption of Hardy-Weinberg equilibrium and independence of trace data across individuals:

$$\Pr(\mathrm{Data}|f) \propto \prod_i [(1-f)^2 + 2f(1-f)\exp(\overline{\mathrm{LLR}}^i_{\mathrm{het}}) + f^2\exp(\overline{\mathrm{LLR}}^i_{\mathrm{min}})]$$

where the product is over all individuals i, and $\overline{\mathrm{LLR}}^i_{\mathrm{het}}$ and $\overline{\mathrm{LLR}}^i_{\mathrm{min}}$ are the LLRs for individual i. The EM algorithm is initialized at $f = 0.05$, and if $\hat{f} < 0.05$ we set $\hat{f} = 0.05$. (The aim here is to get only a very rough idea of the frequency of the minor allele to improve genotyping accuracy, particularly for high-frequency SNPs. The limit of 0.05 limits the informativeness of the prior on genotypes used in the next step.)

6. At each potential SNP for each individual, we compute a probability of each genotype using

$$\Pr(\mathrm{hom}) \propto (1-\hat{f})^2$$

$$\Pr(\mathrm{het}) \propto 2\hat{f}(1-\hat{f})\exp(\overline{\mathrm{LLR}}_{\mathrm{het}})$$

$$\Pr(\mathrm{min}) \propto \hat{f}^2\exp(\overline{\mathrm{LLR}}_{\mathrm{min}})$$

As above, these expressions come from application of Bayes Theorem, using Hardy-Weinberg equilibrium and $\hat{f}$ to specify a 'prior' for each genotype. The constant of proportionality is determined by assuming these probabilities must sum to 1. Each individual has its genotype classified as hom, het or min, according to which has the highest probability, and this probability (as a rounded-down percentage) is assigned as a score to that genotype call. The specific base, or pair of bases, involved in each genotype is then determined as described in the Supplementary Information.

The algorithm is described in more detail in the Supplementary Methods.

Availability of traces. Traces used in this study have been deposited in the National Center for Biotechnology Information (NCBI) Trace Archive.

URLs. PolyPhred v5, http://droog.mbt.washington.edu/PolyPhred.html; Environmental Genome Project, http://egp.gs.washington.edu; NCBI Trace Archive, http://www.ncbi.nlm.nih.gov/Traces/; SeattleSNPs, http://pga.gs.washington.edu; SoftGenetics, http://www.softgenetics.com/me/index.htm. Traces and manual genotype calls are available at http://chum.gs.washington.edu/poly_data.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors thank past and present members of the Nickerson lab for compiling the databases that were used to develop, train and test our algorithm. This work was supported by US National Institutes of Health (NIH) grants (1RO1HG/LM-02585 to M.S., and ES-15478 and HL-66682 to D.A.N.). P.S. was supported by an NIH training grant (T32 HG00035-06).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Carlson, C.S., Newman, T.L. & Nickerson, D.A. SNPing in the human genome. *Curr. Opin. Chem. Biol.* **5**, 78–85 (2001).
2. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
3. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
4. Carlson, C.S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* **33**, 518–521 (2003).
5. Marth, G.T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456 (1999).
6. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
7. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
8. Weckx, S. *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res.* **15**, 436–442 (2005).
9. Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W. & Nickerson, D.A. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23**, 138–144 (1994).
10. Parker, L.T. *et al.* AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns. *Biotechniques* **21**, 694–699 (1996).
11. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using Phred. I. accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
12. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
13. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
14. Bhangale, T.R., Rieder, M.J., Livingston, R.J. & Nickerson, D.A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**, 59–69 (2005).
15. Olden, K. & Wilson, S. Environmental health and genomics: visions and implications. *Nat. Rev. Genet.* **1**, 149–153 (2000).
16. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **34**, 1–38 (1977).
17. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).