nature genetics

Automating resequencing-based detection of insertion-deletion polymorphisms

Tushar R Bhangale[1,2], Matthew Stephens[2,3] & Deborah A Nickerson[1,2]

Structural and insertion-deletion (indel) variants have received considerable recent attention, partly because of their phenotypic consequences. Among these variants, the most common are small indels (~1–30 bp). Identifying and genotyping indels using sequence traces obtained from diploid samples requires extensive manual review, which makes large-scale studies inconvenient. We report a new algorithm, implemented in available software (PolyPhred version 6.0), to help automate detection and genotyping of indels from sequence traces. The algorithm identifies heterozygous individuals, which permits the discovery of low-frequency indels. It finds 80% of all indel polymorphisms with almost no false positives and finds 97% with a false discovery rate of 10%. Additionally, genotyping accuracy exceeds 99%, and it correctly infers indel length in 96% of the cases. Using this approach, we identify indels in the HapMap ENCODE regions, providing the first report of these polymorphisms in this data set.

Recent studies have started to catalog the large number of structural and indel variants present in human populations[1–6]. Of these, the most common are small (~1- to 30-bp) indel polymorphisms[7]. Small indels are important both because of their relative abundance (they are the second most frequent type of polymorphism in humans after nucleotide substitutions) and their functional significance: indels in coding regions can cause severe disruptions in coding sequences[8,9], and indels in promoter regions can alter transcriptional activity[10,11]. Indeed, small indels currently constitute ~24% of all disease-causing mutations reported at the Human Gene Mutation Database[12] (as of August 2006). As the allele frequency spectrum and linkage disequilibrium (LD) characteristics of indels are similar to substitutions[5,7], indels can improve the resolution of genetic maps to uncover a more detailed picture of sequence variation and LD in any region and can have a valuable role in the mapping of complex diseases and traits.

Despite the abundance and potential functional importance of these small indel polymorphisms, no efficient high-throughput technologies currently exist to automatically identify and genotype them in population samples. In principle, as for substitutions[13–15], these tasks can be accomplished by fluorescence-based resequencing. In particular, individuals heterozygous for an indel allele can be reliably identified from the complex pattern of multiple heterozygous peaks (that is, the presence of a peak with a ~50% drop in height compared with a homozygote along with the presence of a second peak of similar height corresponding to the alternate allele) that occur because of mismatches in the two allelic sequences downstream of an indel[7]. This detection of heterozygotes has a central role in comprehensively detecting diallelic polymorphisms because for lower-frequency variants, samples will often not include homozygotes for both alleles. In addition, the pattern of peaks in heterozygotes can be used to identify the inserted or deleted segment relative to a reference sequence. Thus, with recent advances in high-throughput sequencing technology and the rapid increase in resequencing-based polymorphism discovery[16–18], there exists an opportunity for large-scale identification of small indel polymorphisms. However, although indels can be effectively identified and genotyped manually using this pattern[7], for large-scale applications, it is impractical to manually examine every trace. Although existing software tools novoSNP[15], InSNP[19] and Mutation Surveyor (Softgenetics) help to automate this process, these approaches still require extensive manual review of the identified polymorphisms.

In this report, we describe a new algorithm to help automate the identification and genotyping of small (diallelic) indels from sequence trace data. The method detects heterozygous indel patterns using a statistical analysis of the base calls, quality and peak height data obtained from raw sequence traces. In our tests, it is able to identify 80% of indels entirely automatically (without any false positives) and 97% of indels at a false discovery rate (that is, the proportion of false positives among the positive discoveries) of 0.1. Its genotyping accuracy exceeds 99%, and it can correctly infer the indel length in 96% of sites. The algorithm, implemented in a software package (PolyPhred version 6.0) is available from http://droog.mbt.washington.edu/PolyPhred.html. We applied the method to analyze sequence trace data from the ten ENCODE regions, generated as part of the HapMap project, and our method identified 1,244 potential new indel polymorphisms, 1,126 of which (91%) we confirmed to be indels upon manual inspection of the traces. The manual confirmation process for 5 Mb of reference sequence took one person roughly 30 h, demonstrating the potential for large-scale application.

[1]Department of Bioengineering, [2]Department of Genome Sciences and [3]Department of Statistics, University of Washington, Seattle, Washington 98195, USA. Correspondence should be addressed to T.B. (tushar@u.washington.edu) or D.A.N. (debnick@u.washington.edu).

Received 5 June; accepted 17 October; published online 19 November 2006; doi:10.1038/ng1925

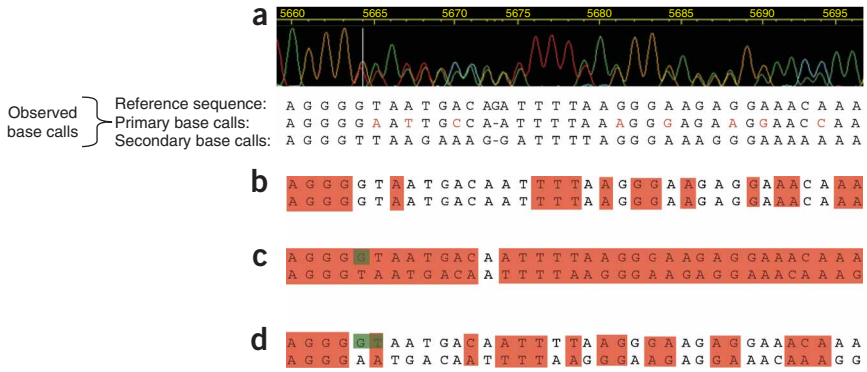

Figure 1 An example of how our algorithm identifies a heterozygous indel trace. (**a**) Heterozygous indel trace along with its observed base calls aligned to the reference sequence. For the purpose of this example, secondary base calls reported by Phred as 'N' (that is, cases in which no secondary base was found) have been replaced by the corresponding primary base calls, suggesting that both the alleles of the individual have the same base at the position. (**b**–**d**) Expected base calls computed from the reference sequence under the assumption that there is (**b**) no indel, (**c**) a deletion of 1 bp at position 5664 or (**d**) a deletion of 2 bp at position 5664. In **c** and **d**, bases highlighted in green in the long allele (top sequences) are deleted in the short allele (bottom sequences), which shifted the downstream bottom sequence to the left. Base calls highlighted in red indicate those expected base calls that can be matched with the observed base calls. The large number of matches in **c** (when compared with **b** and **d**) reflect the fact that this trace comes from an individual with a 1-bp deletion at position 5664.

RESULTS

Our algorithm identifies indels through the characteristic pattern of peaks that occurs in the traces of heterozygous individuals downstream of these polymorphisms (**Fig. 1a**). Distinguishing between this pattern and the pattern of peaks in an ideal non-indel trace, which consists of evenly spaced peaks of similar peak intensities (**Supplementary Fig. 1** online), is straightforward. The main challenge is distinguishing between the pattern created by a heterozygous indel and the noisy peak patterns occurring at the end of each trace due to inconsistent gel migration of very long fragments (**Supplementary Fig. 1**). To do this, our method uses the base calls as well as other information (such as peak heights) obtained from the sequence traces by Phred[20]. In outline, the method proceeds as follows. First, for each possible indel position and indel length, a likelihood ratio is computed, reflecting the extent to which the primary and secondary base calls reported by Phred are more similar to calls expected in a trace of an individual heterozygous for such an indel than to calls expected in traces not containing an indel. This process is illustrated in **Figure 1a–d**. Within each trace, the location and the length of the indel that has the highest likelihood ratio is treated as a possible indel and is subjected to further analysis. In particular, we compute the heights of the peaks both upstream and downstream of the possible indel to see how they compare with the pattern expected in heterozygotes for actual indels (downstream of the indel, we expect to see two peaks at most locations, each of roughly the same height and about half the height of the peaks upstream of the indel). Having computed these trace features, we use them, together with the likelihood ratio, as independent variables in a logistic regression to discriminate between those traces from individuals heterozygous for an indel and those who are not. (The parameters of this logistic regression have been estimated using a training set of traces that have been manually determined to contain, or not contain, an indel). Ultimately, our algorithm provides a score for each location, summarizing the strength of the evidence for an indel at that location. It marks each indel whose score exceeds some user-specified threshold as a potential indel and uses the data on each individual's traces to assign genotypes to that individual, together with an individual-specific score summarizing the confidence that should be placed in each genotype call. The method also provides an estimate of the length of the indel. See Methods for more details.

Accuracy of indel detection

To assess accuracy of our algorithm for detecting indels, we applied it to a 'test set' of sequence traces obtained from 16 genes involved in inflammation, lipid metabolism and blood pressure regulation that were resequenced across 24 individuals of African descent and 23 individuals of European descent. To ensure a fair evaluation, there was no overlap between this test set and the training set of sequence data used to train the algorithm. We compared potential indels marked by our method with a local database of 172 'known' indels that had been previously identified by extensive manual inspection of the same traces, and we summarize accuracy with two numbers: the true positive rate (TPR), which is the proportion of indels that the method successfully detects (among all known indels), and the false discovery rate (FDR), which is the proportion of potential indels identified that are not actually indels (among all potential indels identified). As most indels that can be

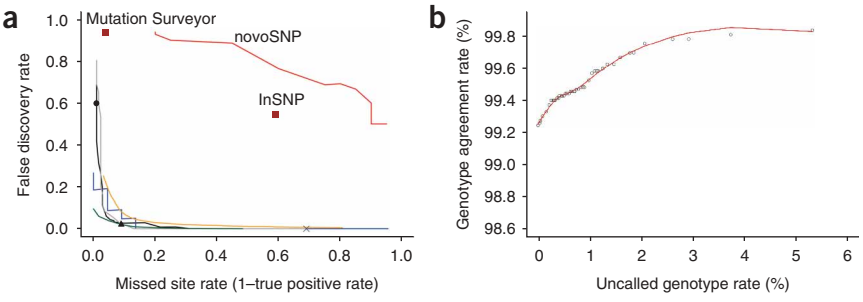

Figure 2 Indel detection and genotyping accuracy. (**a**) Missed site rate versus false discovery rate for different methods and data sets. Detection accuracy for the test genes from SeattleSNPs data (black line) was computed by varying the threshold on the score for accepting site; results for thresholds of 70 (circle), 85 (triangle) and 90 (cross) are highlighted. Gray line represents results obtained after exchanging the training data with the test data. Green line represents results from ENCODE data, and blue line shows results from a smaller data set that was used to evaluate performance of other software tools. Overall, the indel detection accuracy is similar to the SNP detection accuracy of PolyPhred v5 (orange line). Mutation Surveyor and InSNP do not provide any way to obtain results at different levels of TPR. Hence, their results are plotted as two labeled points (brown squares). Results for novoSNP are shown by the red line. (**b**) Genotype agreement rate versus the uncalled (that is, not reported) genotype rate at different threshold values on the genotype score for reporting genotype.

identified using resequencing are only 1 to a few bases in length[7], we considered only indels $\leq$30 bp in length, which constitute $\sim$99% of all indels in this data, to assess the algorithm's performance.

Our method assigns a score to each location, summarizing the strength of the evidence for an indel at that location, and marks those locations whose score exceeds a user-specified threshold as potential indels. As the TPR and FDR vary with the threshold chosen, we compute these quantities for various thresholds (**Fig. 2a**). The figure provides an indication of the trade-off between sensitivity and specificity or, more specifically, between missing indels and falsely identifying locations that are not actually indels. At a TPR of 0.8, the FDR was close to 0; at a TPR of 0.90, the FDR was 0.03; and for a TPR of 0.97, the FDR was 0.1. The method showed similar accuracy when we exchanged the training data (see Methods) with test data (**Fig. 2a**). Performance on a smaller data set (see below), which included a subset of the test data, was similar to performance on the entire test data (**Fig. 2a**). We obtained a slightly higher level of accuracy when we analyzed the much larger ENCODE data sets described below (**Fig. 2a**), although interpretation of these latter numbers is complicated by the fact that the indels were detected using the algorithm that we wish to test, so the TPR will be overestimated. Overall, the particular training data or the size and composition of the test data do not seem to produce a substantial difference in the performance of the method. The level of accuracy is similar to that achieved by current methods for detecting SNPs from sequence trace data[14].

We compared the performance of this approach to other available software tools capable of identifying indels, including novoSNP[15], InSNP[19] and Mutation Surveyor. As these tools can process only limited data sets, we evaluated the performances (**Fig. 2a**) on a smaller data set (see Methods). Detection rates for the various approaches on these data were an FDR of 0.94 at a TPR of 0.8 (novoSNP), an FDR of 0.55 at a TPR of 0.41 (InSNP), an FDR of 0.94 at a TPR of 0.96 (Mutation Surveyor) and an FDR of 0.09 at a TPR of 0.95 (PolyPhred 6.0).

Indel length and genotyping accuracy

Determination of indel length from the pattern of peaks in a heterozygote sample is tedious and requires considerable manual analysis. In 95.6% of the identified sites, we found that indel length determined by our method was the same as the one reported by the human expert, suggesting that our approach can substantially reduce the workload in variation analysis.

To assess the method's accuracy in genotyping heterozygotes, we computed the genotype agreement rate (GAR) as the proportion of genotypes for which the call by the algorithm agreed with that by a human expert. Overall, the GAR was 99.3%, but this can be increased by not reporting genotypes that receive low scores. For example, among genotypes with a score of at least 93, the GAR is 99.7%. This increased GAR comes at the expense of 2% missing (that is, not reported) genotypes (**Fig. 2b**).

Analysis of ENCODE regions

We applied our algorithm to sequence traces for the ten ENCODE regions[18] taken from the NCBI trace archive. We were able to align a total of 544,465 traces successfully to the reference sequence and analyzed them to identify 1,244 potential indels (score $\geq$80) of length $\leq$30 bp. Of these, 1,126 seemed to be real based on subsequent manual inspection of the traces (**Table 1**). Distribution of lengths of these indels was similar to that of the indels in the SeattleSNPs data (**Supplementary Fig. 2** online), with 1-bp, $<$5-bp and $<$12-bp indels constituting 46%, 82% and 95% of the indels, respectively. Four of the indels occurred in coding regions, 28 in the 5′ and 3′ UTR regions and 315 in the intronic regions; the remaining indels were located in intergenic regions (**Supplementary Table 1** online). Chromosomal locations and lengths of the indels are provided in **Supplementary Table 2** online. Based on recent interest in the LD characteristics of indels in relation to those of SNPs[5,7,21,22] and whether indels can be effectively assayed by proxy in SNP-based association studies, we compared marker associations versus physical distance for marker pairs in the ENCODE regions containing (i) an indel and a SNP and (ii) two SNPs. As previously observed[5,7] the strength of associations between indels and SNPs and SNP pairs are comparable (**Supplementary Fig. 2**).

Table 1 The number of indels and SNPs discovered in the ENCODE project regions

Region	Chromosome band	Indels	SNPs[a]	Sequencing center
ENr112	2p16.3	142	2,275	Broad
ENr131	2q37.1	169	1,910	Broad
ENr113	4q26	152	2,201	Broad
ENm010	7p15.2	81	1,271	Baylor
ENm013	7q21.13	151	1,807	Broad
ENm014	7q31.33	134	1,966	Broad
ENr321	8q24.11	86	1,758	Baylor
ENr232	9q34.11	55	1,324	Baylor
ENr123	12q12	47	1,792	Baylor
ENr213	18q12.1	109	1,640	Baylor
	Total	1,126	17,944	

[a]SNP counts reported in Table 2 of ref. 18.

DISCUSSION

The development of high-density genetic maps across the human genome provides unparalleled resources for analyzing the association between common sequence polymorphism and common disease[18]. Once applied, these resources are likely to identify the region(s) associated with specific phenotype(s), and subsequent studies will turn to completely cataloguing the variation in the region of interest for specific populations or individuals[23] to identify variants for further molecular analysis. Resequencing has been the gold standard in polymorphism discovery, and with its rapid increase in throughput and reduction in costs, it is expected to remain at the forefront for variant identification. One in every 15 diallelic sites is an indel (**Table 1**). Aside from their value in increasing the resolution of genetic maps, indel polymorphisms can also have an important role as functional variants.

Although the proposed method is designed to identify new indel polymorphisms and determine their genotypes for the sampled individuals, it can be easily adapted to genotype known indels across a larger set of samples. This approach will also be valuable in diagnostic scanning, as indels are a major form of known disease-causing mutations[12].

We applied our method to the ENCODE regions of the HapMap. Indels were not previously scored in these samples, so these data represent a valuable resource to supplement the existing polymorphism data in these regions and help to provide a more complete picture of sequence variation in these sequences. They also enrich the database with potentially functional sequence variation, as three of the indels identified are predicted to lead to frameshifts in the underlying coding sequences in the ENCODE regions. One of these three indels occurs in a gene involved in hereditary multiple exostoses[24]. For the indels

identified in the SeattleSNPs data, four indels are predicted to cause frameshift changes, one indel disrupts a splice site in the gene *CD36* and another indel alters transcription factor binding, which has been shown to increase the risk of coronary heart disease[25]. Because of the importance of indel polymorphisms, many approaches are emerging to identify and catalog large ($\geq$70 bp) indels[4–6,26]. Our method complements these methods by providing an approach to efficiently identify smaller indels, which form the largest fraction of the indel variation in the genome[7,27,28]. The software tool Polyphred v6 also includes an accurate algorithm to identify and genotype nucleotide substitutions[14] and thus can be used to provide a comprehensive catalog of sequence variation in a highly automated manner in any region that can be amplified and sequenced from the human genome.

Although resequencing can identify indels of lengths up to the length of the PCR product, the proposed method can be computationally slow for detecting indels that are >30 bp in length. We are investigating approaches more suited for indels involving more than 30 bp.

METHODS

SeattleSNPs resequencing data and identification of indels. The resequencing data for this work was produced by SeattleSNPs. Data for the following 18 genes were used as training data to fit the models (see below): *APOLH*, *CY4F2*, *EPHB6*, *F2RL1*, *IL1AP*, *IL1BT*, *IL1R1*, *IL1RN*, *IL21R*, *IL2RB*, *ILK24*, *ILK4R*, *ILKN4*, *KLKN1*, *PLAUR*, *SELEL*, *SFPA2* and *TRPV6*. Data for the following 16 genes were used as the test data to assess the performance of the method: *ABOBG*, *CD36A*, *ESELE*, *F2RL2*, *FAC11*, *FCT10*, *MMPR3*, *MP3K8*, *PLAI1*, *SFPA1*, *SFTPD*, *TIRAP*, *TNFP3*, *TRAF6*, *TRPV5* and *VCAM1*.

The traces have been deposited in NCBI Trace Archive, and all variation data have been deposited into the dbSNP database. All the DNA samples used for variation discovery were obtained from Coriell Cell Repository. The candidate genes were resequenced across two populations: 24 individuals selected from the African American Human Variation Panel (HD50AA; individuals NA17101–NA17116 and NA17133–NA17140) and 23 individuals from Centre d'Etude du Polymorphisme Humain (CEPH) reference panel DNAs (Coriell Cell Repository numbers NA06990, NA07019, NA07348, NA07349, NA10830, NA10831, NA10842, NA10843, NA10842–NA10845, NA10848, NA10850–NA10854, NA10857, NA10858, NA10860, NA10861, NA12547, NA12548 and NA12560). The expected detection rates for these sample sizes are 99% for sites with population mean allele frequency (MAF) of >5% and 87% for sites with population MAF >1% (ref. 29). For each gene, we sequenced the genomic region spanning the longest reference transcript in Entrez Gene, including exons and introns, ~2.5 kb upstream of the gene and ~1.5 kb downstream of the gene. Sequencing and data analysis were performed as described in ref. 16. In brief, overlapping PCR primers were designed to cover the target region with an average amplicon size of ~980 bp and average overlap between amplicons of ~190 bp. The PCR products were sequenced using dye terminator chemistry on ABI 3730 instruments. A total of 115,489 traces (average length, ~650 bp) were generated. The trace data were analyzed using the base-calling software Phred[20,30]. Phred assigns primary and secondary base calls (if the secondary peak is present) to each of the peaks in the traces as well as computing quality values and heights of the primary and the secondary peaks. The method we developed uses these data and the reference sequence to identify indels. The sequence data were mapped onto the reference sequence using Phrap and Cross_match (see URLs section below). The resultant assemblies were visualized using the Consed program[31] in order to correct occasional errors in the alignments and to identify indels. Indels in the data were initially identified and genotyped manually through the identification of heterozygous indel patterns in traces[7].

We used a smaller data set comprising four genes with 26 indels to compare the performance of our method with other software tools, as the other tools can process only limited amount of data. These genes spanned a total of 40 kb reference sequence and were resequenced using 75 PCR amplicons. Two of these genes (*F2RL2* and *SERPINE1*) were from the SeattleSNPs data, and two genes (*ACTB* and *ALAD*) were resequenced as a part of the Environmental Genome Project across eight individuals (Coriell Cell Repository numbers NA15385A, NA15063, NA15506, NA15341, NA15242, NA15352, NA15078 and NA15365A).

Indel detection algorithm. The algorithm consists of the following steps:

1. For each trace, use the given reference sequence to compute the probability of the observed base calls, conditional on the observed quality scores, under the assumption that there is no indel anywhere in the trace. This probability is computed as follows: let $X = (X_1, X_2, \ldots, X_n)$ denote the observed base calls, where X_j is the ordered pair of primary and secondary base calls at the j^{th} position in the trace, and n is the length of the trace in bp. Let $Q = (Q_1, Q_2, \ldots, Q_n)$ denote the corresponding observed quality scores, and let $Y^{\text{non-indel}} = (Y_1, Y_2, \ldots, Y_n)$ denote the corresponding 'expected' base calls determined from the reference sequence (**Fig. 1**). We assume that

$$\Pr(X|Y^{\text{non-indel}}, Q) = \prod_{j=1}^{n} \Pr(X_j|Y_j^{\text{non-indel}}, Q_j),$$

where $\Pr(X_j|Y_j^{\text{non-indel}}, Q_j)$ is determined from tables we generated using observed and expected base calls in a large number of traces determined by manual inspection not to contain an indel (**Supplementary Methods** online).

2. For each trace, compute the probability of X, conditional on Q, under the assumption that the trace contains a heterozygous indel of length k beginning at site i, for $-30 \leq k \leq 30$, $k \neq 0$ and $|k| < i < n - |k|$ (here, a negative value of k represents an insertion relative to the reference sequence). This probability is again based on using the reference sequence to determine the 'expected' base calls, $Y^{i,k} = (Y_1^{i,k}, Y_2^{i,k}, \ldots, Y_n^{i,k})$, if there is an indel of length k at location i (**Fig. 1c,d**), and then assuming that

$$\Pr(X|Y^{i,k}, Q) = \prod_{j=1}^{n} \Pr(X_j|Y_j^{i,k}, Q_j).$$

This expression assumes that the vectors of the observed base calls (X) and the expected base call ($Y^{i,k}$) are properly aligned. In practice, however, errors in the alignment of traces downstream of the indel can disrupt the alignment of the observed and expected base calls. To overcome this, we align X with $Y^{i,k}$ using a dynamic programming algorithm (**Supplementary Methods** and **Supplementary Fig. 3** online) similar to the global pairwise sequence alignment algorithm described in ref. 32. For $j < i$, the probabilities $\Pr(X_j \mid Y_j^{i,k}, Q_j)$ are assumed to be the same as for traces containing no indel, whereas for $j \geq i$, the probabilities are determined from tables created using observed and expected base calls downstream of a heterozygous indel in traces manually determined to contain a heterozygous indel (**Supplementary Methods**). Let $\hat{i}$ and $\hat{k}$ denote the values of i and k that maximize $\Pr(X|Y^{i,k}, Q)$.

3. Compute the following five features:

(i) the log-likelihood ratio (*LLR*):

$$LLR = \log(\Pr(X|Y^{\hat{i},\hat{k}}, Q) / \Pr(X|Y^{\text{non-indel}}, Q))$$

(ii) the length L of the trace downstream of $\hat{i}$: $L = n - \hat{i}$

(iii) a measure of the goodness of fit of the observed data to the indel model (GOF), being the log-likelihood for the data downstream of the indel divided by L:

$$\text{GOF} = \log\left(\prod_{j \geq \hat{i}} \Pr(X_j|Y_j^{\hat{i},\hat{k}}, Q)\right) \Big/ L$$

(iv) a feature that summarizes the ratio of secondary to primary peak heights at the potential heterozygous peaks in the trace downstream of the indel (htratio):

$$\text{htratio} = \text{median}(h2_j/h1_j : \hat{i} \leq j \leq n, b2_j \neq \text{'N'}),$$

where $h1_j$ and $h2_j$ denote the primary and the secondary peak heights, respectively, reported by Phred; $b2_j$ is the secondary base call at the j^{th} peak in the trace and $b2_j \neq$ 'N' is the condition that a secondary base is reported at the j^{th} peak by Phred (for real indels, htratio tends to have a value close to 1);

(v) a feature that summarizes the relative drop in the primary peak heights downstream of the indel position compared with the upstream trace (drop):

$$\text{drop} = \sum_{b\in\{A,C,G,T\}} d_b w_b,$$

where

$$d_b = \sum_{j<\hat{i};b1_j=b} \log(h1_j) \Big/ \sum_{j<\hat{i}} 1[b1_j = b] - \sum_{j\geq\hat{i};b1_j=b;b2_j\neq\text{'}N\text{'}} \log(h1_j) \Big/ \sum_{j\geq\hat{i}} 1[b1_j = b,\ b2_j \neq \text{'}N\text{'}]$$

and

$$w_b = \sum_{j\geq\hat{i}} 1[b1_j = b, b2_j \neq \text{'}N\text{'}] \Big/ \sum_{j\geq\hat{i}} 1[b2_j \neq \text{'}N\text{'}],$$

where $1[.]$ represents the indicator function, and $b1_j$ denotes the primary base call at the j^{th} peak.

4. Eliminate traces with $L \leq 5$ from further analysis. For the remaining traces, use the five features, along with suitable transformations and interaction terms (**Supplementary Table 3** online) as independent variables in a logistic regression model to compute a score (LLR_{trace}) for each trace, quantifying the strength of the evidence for an indel in that trace.

5. Combine information across multiple traces to identify likely locations of indel polymorphisms. First, indel positions ($\hat{i}$) from the traces with $LLR_{\text{trace}} > -0.5$ are mapped on to the reference sequence to create a list of likely locations of indel polymorphisms in the gene. This results in clusters of points where every cluster corresponds to either a true indel locus or a false positive locus. We therefore apply a clustering algorithm to these points to identify putative indel loci (see **Supplementary Methods** for details). For every putative locus, we compute a log-likelihood ratio score (LLR_{locus}) using a logistic regression model (**Supplementary Table 4** online) that uses the following two features as independent variables: (i) the highest value of the LLR_{trace} at the cluster and (ii) the proportion of traces with $LLR_{\text{trace}} > -0.5$ among the traces that align at that locus. A score with range 0–99, computed using LLR_{locus} as

$$\lfloor 100 \exp(LLR_{\text{locus}})/(1+\exp(LLR_{\text{locus}})) \rfloor,$$

is then assigned to the locus.

6. At every potential indel locus, determine the genotypes (homozygous or heterozygous) of each individual using a method similar to the expectation maximization (EM) algorithm–based approach described in ref. 14 as follows:

(i) Initialize the current estimate $\hat{f}$ of the minor allele frequency to 0.01.

(ii) For every individual, use $\hat{f}$ to compute the probability that the genotype is heterozygous, using

$$\Pr(\text{het})/\Pr(\text{hom}) = \hat{f}/(1-\hat{f})\exp(LLR_{\text{trace}})$$

.

(iii) Compute the new value of $\hat{f}$ as the average of the Pr(het) values for the individuals. If $\hat{f} < 0.01$, set $\hat{f} = 0.01$.

(iv) Return to (ii).

At the end of the second iteration, if Pr(het) > Pr(hom), the individual's genotype is classified as heterozygous; otherwise, it is classified as homozygous. A score with a range of 50–99 is then assigned to the genotype call by multiplying the corresponding probability by 100 and rounding the result to integer value.

Analysis of ENCODE resequencing data. Traces for the ten ENCODE regions were obtained from the NCBI Trace Archive. Owing to the large memory requirements, in order to make this data amenable to analysis, each of the ten regions was further divided into five subregions of 100 kb each. For the five ENCODE regions resequenced at the Broad Institute, the traces were assigned to the subregions based on the chromosomal locations of these traces available at the trace archive. For the traces resequenced for the remaining five regions, the chromosomal locations were not available in the archive. Therefore, traces were assigned to the subregions based on how well they aligned to the reference sequence of the subregion. We used the base-call sequence and the quality values provided at the trace archive for each of the traces and Cross_match to perform these alignments. Traces that were successfully assigned to the subregions were analyzed using Phred to determine the base calls and quality values. For each of the 50 subregions, an assembly of traces was constructed by mapping the trace sequences on to the corresponding reference sequence using Cross_match and Consed. The resulting assemblies were then analyzed using the indel detection algorithm to identify indel loci. Trace data for the indels detected by the method were then manually examined to confirm the indels at the loci.

URLs. MutationSurveyor is available from Softgenetics at http://www.softgenetics.com. Our algorithm, implemented in a software package (PolyPhred version 6.0) is available from http://droog.mbt.washington.edu/PolyPhred.html. For SeattleSNPs, see http://pga.gs.washington.edu/. For the NCBI Trace Archive, see http://www.ncbi.nlm.nih.gov/Traces. For the Coriell Cell Repository, see http://locus.umdnj.edu/ccr. For Phrap and Cross_match, see http://www.phrap.org. For the Environmental Genome Project, see http://egp.gs.washington.edu.

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

ACKNOWLEDGMENTS

The authors thank the past and present members of the SeattleSNPs team for their efforts in variation discovery and the PolyPhred development team, including J. Sloan and P. Robertson. This work was supported by grants from the US National Institute of Health (HL66682 to D.A.N. and HG/LM02585 to M.S.).

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. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
2. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
3. Albertini, A.M., Hofer, M., Calos, M.P. & Miller, J.H. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell* **29**, 319–328 (1982).
4. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
5. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
6. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
7. 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).
8. Othman, M. *et al.* Identification and functional characterization of a novel 27-bp deletion in the macroglycopeptide-coding region of the GPIBA gene resulting in platelet-type von Willebrand disease. *Blood* **105**, 4330–4336 (2005).
9. deSanctis, L. *et al.* Familial PAX8 small deletion (c.989_992delACCC) associated with extreme phenotype variability. *J. Clin. Endocrinol. Metab.* **89**, 5669–5674 (2004).
10. Karban, A.S. *et al.* Functional annotation of a novel NFKB1 promoter polymorphism that increases risk for ulcerative colitis. *Hum. Mol. Genet.* **13**, 35–45 (2004).
11. Lin, S.C. *et al.* Correlation between functional genotypes in the matrix metalloproteinases-1 promoter and risk of oral squamous cell carcinomas. *J. Oral Pathol. Med.* **33**, 323–326 (2004).
12. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
13. 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).
14. Stephens, M., Sloan, J.S., Robertson, P.D., Scheet, P. & Nickerson, D.A. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38**, 375–381 (2006).
15. Weckx, S. *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res.* **15**, 436–442 (2005).

16. 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).
17. Livingston, R.J. *et al.* Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14**, 1821–1831 (2004).
18. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
19. Manaster, C. *et al.* InSNP: a tool for automated detection and visualization of SNPs and InDels. *Hum. Mutat.* **26**, 11–19 (2005).
20. 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).
21. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
22. Newman, T.L. *et al.* High-throughput genotyping of intermediate-size structural variation. *Hum. Mol. Genet.* **15**, 1159–1167 (2006).
23. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
24. Ahn, J. *et al.* Cloning of the putative tumour suppressor gene for hereditary multiple exostoses (EXT1). *Nat. Genet.* **11**, 137–143 (1995).
25. Rockman, M.V. *et al.* Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.* **14**, 1531–1539 (2004).
26. Eichler, E.E. Widening the spectrum of human genetic variation. *Nat. Genet.* **38**, 9–11 (2006).
27. Weber, J.L. *et al.* Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**, 854–862 (2002).
28. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
29. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
30. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
31. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
32. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).