

Identification of genetic variants using bar-coded multiplexed sequencing

David W Craig^{1,3}, John V Pearson^{1,3}, Szabolcs Szalinger^{1,3}, Aswin Sekar¹, Margot Redman¹, Jason J Corneveaux¹, Traci L Pawlowski¹, Trisha Laub¹, Gary Nunn², Dietrich A Stephan¹, Nils Homer¹ & Matthew J Huentelman¹

We developed a generalized framework for multiplexed resequencing of targeted human genome regions on the Illumina Genome Analyzer using degenerate indexed DNA bar codes ligated to fragmented DNA before sequencing. Using this method, we simultaneously sequenced the DNA of multiple HapMap individuals at several Encyclopedia of DNA Elements (ENCODE) regions. We then evaluated the use of Bayes factors for discovering and genotyping polymorphisms. For polymorphisms that were either previously identified within the Single Nucleotide Polymorphism database (dbSNP) or visually evident upon re-inspection of archived ENCODE traces, we observed a false positive rate of 11.3% using strict thresholds for predicting variants and 69.6% for lax thresholds. Conversely, false negative rates were 10.8–90.8%, with false negatives at stricter cut-offs occurring at lower coverage (< 10 aligned reads). These results suggest that > 90% of genetic variants are discoverable using multiplexed sequencing provided sufficient coverage at the polymorphic base.

Genome-wide association, candidate gene and linkage studies have identified thousands of moderately sized genomic regions that are associated with human disease but for which comprehensive resequencing is needed to identify the genetic variant causing the association. In particular, genome-wide association studies have identified hundreds of disease-associated haplotypes, typically spanning 5–100 kb^{1–3}. A logical next step is to identify and resequence all genetic variants within the associated haplotype to identify the functional variants among the many nonfunctional, evolutionarily linked neighboring polymorphisms. Next-generation DNA sequencing technologies are in principle well-suited to this task owing to their capability for high-throughput low-cost sequencing. Although these technologies offer massive sequencing capacity, it is still difficult, time-consuming and/or expensive to resequence large numbers of samples across moderately sized genomic regions (5 kb–1 Mb).

Simultaneous resequencing of a target region in large numbers of individuals is possible by bar-coding or indexing the reads from each individual with a short identifying oligonucleotide^{4–7}. Although indexing has the obvious benefit of multiplexing samples

within a run, DNA indexing offers two key additional advantages: direct measure of base-by-base error rate and reduction of array-to-array or day-to-day variability. Previous pioneering efforts to develop DNA indexing have shown considerable promise, but their adoption is still in its infancy, and considerable challenges remain, including the development of practical and cost-effective approaches for short-read platforms. Beyond these experimental challenges, there are few analytical frameworks that are characterized for discovering and genotyping genetic variants across a targeted interval using multiplexed short-read sequence data from multiple individuals.

Here we report an experimental and analytical approach for simultaneous sequencing of multiple individuals using DNA bar codes, which we call indexes here, on the Illumina Genome Analyzer (GA). We used a six-base index with built-in redundancy for error correction and assessed the performance of the method by resequencing Encyclopedia of DNA Elements (ENCODE) regions of HapMap individuals that have previously been capillary-sequenced. We developed a Bayesian analytical framework that leverages the inherent ability of indexing to measure error and to reflect variability in sequencing coverage.

RESULTS

Experimental design

We amplified multiple 5-kb regions (Supplementary Tables 1 and 2 online) by long-range PCR for 46 individuals genotyped by the ENCODE projects^{1,8} (Fig. 1 and Supplementary Methods online). For each individual, we pooled the amplicons in equimolar amounts, digested them, blunt end-repaired them, added to them an adenosine overhang and ligated these modified amplicons to one of the 46 indexed adapters (Supplementary Tables 3 and 4 online). After ligation, we pooled samples from all individuals into a single sample (referred to as an indexed library), purified the samples, enriched them by PCR amplification and sequenced them on the Illumina GA on a single lane of an 8-lane flow-cell. We prepared two libraries: library A, consisting of ten 5-kb amplicons covering 50 kb and library B, consisting of fourteen 5-kb amplicons covering 70 kb (Supplementary Table 2). Library A contained regions that were previously capillary-sequenced and regions that

¹The Translational Genomics Research Institute, 445 N. 5th St. 5th Floor, Phoenix, Arizona 85004, USA. ²Illumina, 9885 Town Centre Drive, San Diego, California 92121, USA. ³These authors contributed equally to this work. Correspondence should be addressed to D.W.C. (dcraig@tgen.org).

RECEIVED 17 APRIL; ACCEPTED 22 AUGUST; PUBLISHED ONLINE 14 SEPTEMBER 2008; DOI:10.1038/NMETH.1251

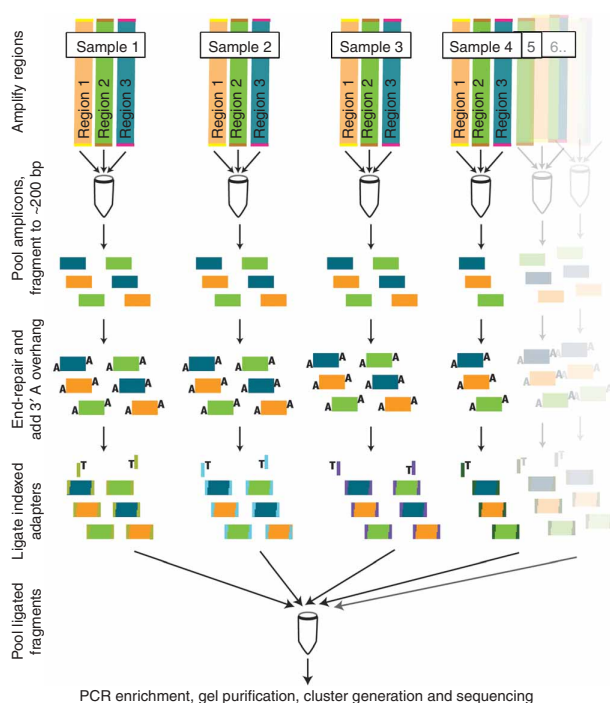


Figure 1 | Schematic describing the preparation of indexed libraries.

were not sequenced as a part of the ENCODE project, whereas library B contained only regions previously sequenced as part of the ENCODE project.

Index design

We used a six-base design, which allowed us to control, tolerate and measure error during base calling of the index. We designed indexes so that one, and in some cases two, sequencing errors could be tolerated without an index being incorrectly identified as being a different valid index. We synthesized only 48 of the 4,096 possible nucleotide combinations (see **Supplementary Table 4** for indexes). Perfect alignment of any index to a randomly generated 6-base sequence should occur at ~0.1% by chance. The sixth base of the index was an obligate thymidine necessary for ligation of the adenosine overhang. The first and fifth bases were identical to detect biases during normalization and calculation of the deconvolution matrix. In practice, we used 46 of the 48 indexes to allow for plate layouts that included positive and negative controls. Although we did not implement this in this study, the use of each of the four nucleotides within an index may provide for higher-accuracy base calling as each base would have to be correctly called at least once within a sequenced read.

Index performance

We generated 3–10 million short-read (32- or 42-base) sequences for each lane of an

8-lane flow cell, though early sequencing runs exhibited greater variability in the number of sequenced reads. After filtering using Illumina analysis pipeline defaults, approximately 45–50% of the reads remained. We observed a large spread in the number of counts per index (**Fig. 2**). Although we did not identify a systematic reason for the initial spread in index performance, weaknesses in index design were obvious in some cases. For example, 'AAAAAT' was frequently read as 'AAAAAAT', perhaps because of an oligonucleotide synthesis bias. A few indexes that were not well-represented were complementary to other sections of the adaptor sequence, possibly hindering adaptor formation. Resequencing the same library gave nearly the identical distribution of reads regardless of run performance, indicating that the distribution is likely not due to a post-PCR enrichment step. Furthermore, recreating libraries and sequencing DNA from different individuals in additional sequencing runs did not substantially alter performance for indexes that were substantially underrepresented or overrepresented. Of the 46 initial indexes, 19 indexes varied by less than a factor of 5 between the most and least common index, and 13 indexes varied by less than a factor of 2. Although some of the initial index variability was consistent between sequencing runs, retrospective analysis of the products by gel electrophoresis after ligation of adaptors suggested that a portion of the index variance may have been due to subtle differences in DNase digestion of pooled amplicons, whereby the number of available ligation targets was higher for samples that were digested with higher efficiency. In runs after sequencing of these initial libraries (data not shown), we observed that quantification and normalization of the amount of ligated adaptor before pooling, using gel electrophoresis of the PCR-enriched products or quantitative PCR, reduced index variability such that the index with highest number of reads aligning to the reference sequence was observed fivefold more frequently than the index with the fewest number of aligned reads. By comparison, the same ratio was 11 fold without quantification of the ligated primers before pooling. Although future studies may improve index variability still further, it may be effectively managed without substantially affecting workflow, by

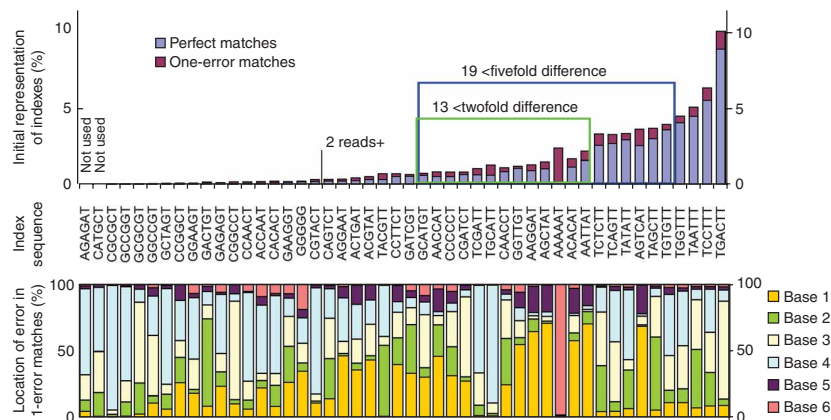


Figure 2 | Comparison of index performance. Index variability in initial sequencing runs (library A) used for evaluating index performance are shown (top). Percentages of reads aligning to the reference sequence are listed by index, without introduction of normalization methods. A total of 30 indexes were present in >0.05% of all aligned reads. There were 19 indexes with less than fivefold difference in index frequencies, which we used in subsequent studies. The bottom graph shows the location of errors by base for each index.

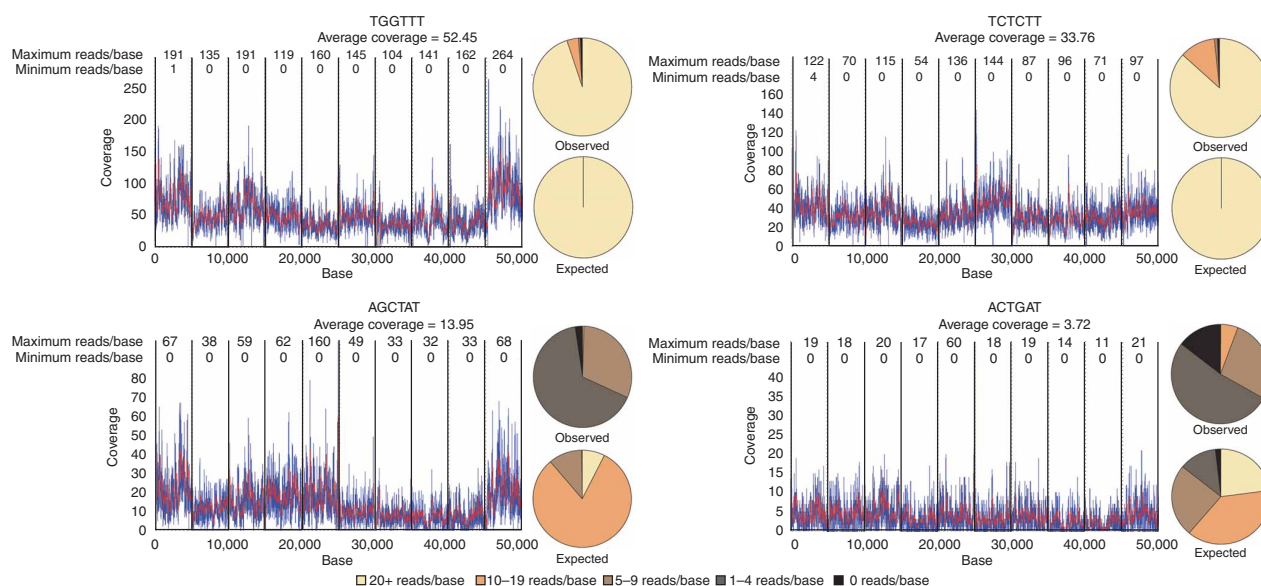


Figure 3 | Relationship between mean and local coverage. Example coverage in 4 individuals sequenced within a single lane of an 8-lane flow-cell for 10 pooled amplicons as part of library A. Amplicons are shown consecutively for each individual. Index sequence and mean coverage for that individual are shown above each graph. The maximum and minimum coverage is shown for each amplicon at the top. Pie charts show the observed distribution of bases across all amplicons and the expected distribution determined from a Poisson distribution of the mean coverage, binned as indicated.

requiring higher average coverage within a study, by sequencing in two lanes with different indexes or by sequestering samples with deficient coverage for later runs.

Index-level coverage

As shown for a subset of library A, coverage across individual 5-kb amplicons was even and generally free of large gaps (Fig. 3). We observed base-to-base variability in the coverage, as expected from alignment of short reads. We observed some deviation from the expected Poisson distribution both between amplicons and within an amplicon. Clearly, amplicon-to-amplicon variability contributes to some extent to the departure from the expected Poisson distribution. For a given index, we observed an approximately 1.5- to 2.0-fold difference between the amplicons with the most and fewest number of reads. Inspecting gel images for selected amplicons confirmed that these observed differences within regions were largely due to uneven pooling of amplicons. The observed amplicon-to-amplicon variability is likely due to the fact that we used median concentrations across the plate when pooling amplicons for an individual, rather than separately pipetting the reaction for each amplicon based on its concentration.

Comparing a given amplicon across indexes (that is, across individuals), there was clearly some base-level correlation in coverage based on the positions of spikes and valleys within the coverage plots (Fig. 3). Within a single amplicon, there was also departure from a Poisson distribution, evident from the fact that the same bases had little or no coverage across individuals. Indeed, there is consistency between individuals with regard to bases that were under or over-represented. The rank correlation coefficient between indexes at a given base averaged 0.408, suggesting that local sequence (or base order) accounts for slightly less than half of the base-to-base variability in coverage.

Error reduction and alignment strategy

Depending on alignment rules, aligning a short read to a reference sequence reduces the sequencing error rate at the cost of limiting discovery. We aligned 35-base-pair sequences, allowing for only a single error. We were thus essentially limited to identifying single-base substitutions in an aligned read, while limiting error to 1/35 or 2.8%, as explained below. We also required that two stretches of 11 or more consecutive bases match the reference sequence or that the read have at least one stretch of 15 consecutive matches to the reference sequence. In both cases, our aligner required that the final 2 bases match the reference sequence to insure that we did not overalign an error at the final base. We chose the rules for alignment largely to control error, and a randomly generated sequence would falsely align in less than 0.1% of alignments in a 100-kb region. Given our tolerance for 1 error in alignment, we expected a maximum per-base error rate of 2.3% (1 error in 35 bases = ~2.8%).

One would expect that we would have greater difficulty detecting closely neighboring single-nucleotide polymorphisms (SNPs) because we mostly limited our aligner to one nonconsecutive mismatch. However, the short-reads stochastically overlapped, and neighboring genetic variants were observed by alignment of multiple sequences not spanning both variants.

Polymorphism discovery

Polymorphism discovery is a primary goal for resequencing an association interval for a genome-wide association study, particularly under the common variant hypothesis. Indeed, in some cases one may only wish to know which bases are polymorphic for custom genotyping on a separate platform.

We first provide an intuitive explanation of our analysis approach for polymorphism discovery (see Methods for

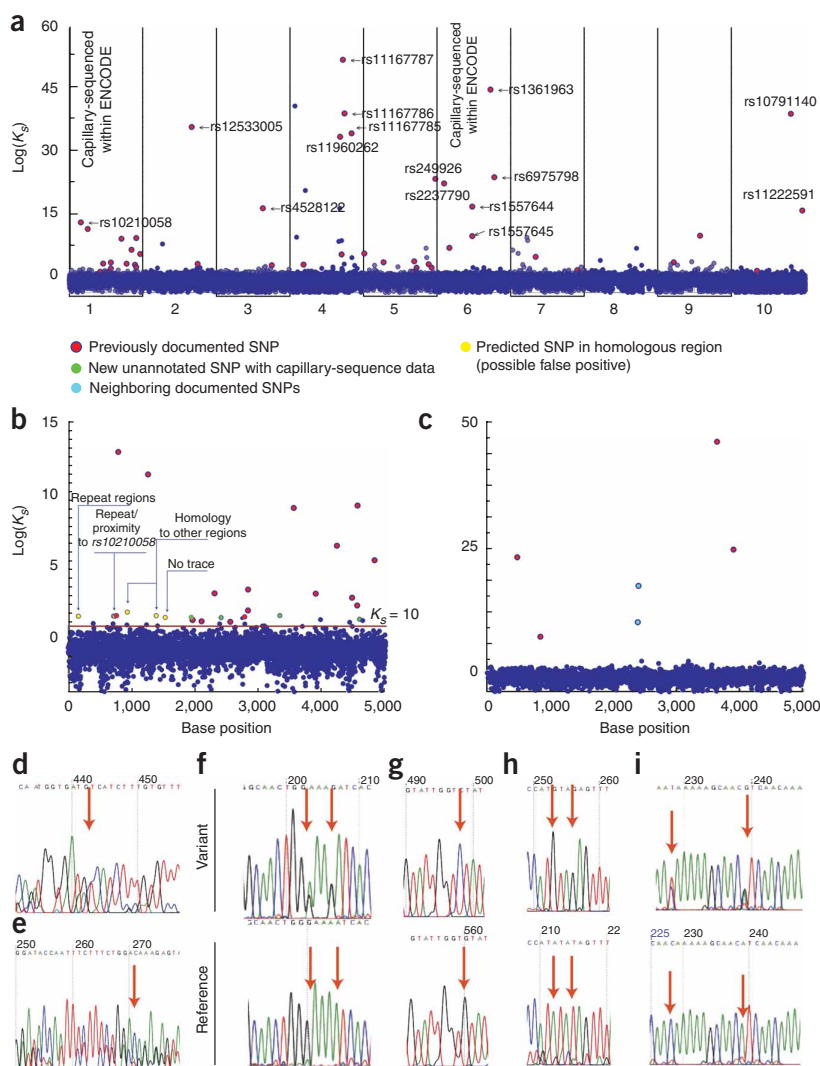


Figure 4 | Discovery of variant bases by simultaneous analysis of all individuals. **(a)** The Bayes factor for polymorphism discovery (K_s) is plotted for each of the 10 sequenced 5-kb amplicons from library A. Exact positions matching known polymorphisms are colored as red spheres, and the dbSNP identifier is provided for most SNPs with high K_s values. **(b,c)** Magnified views of amplicon 1 **(b)** and amplicon 6 **(c)** to compare variants predicted by indexed-multiplexed sequencing to previous deep capillary sequencing results for the same individuals as part of the ENCODE project. **(d,e)** Example traces of predicted SNPs in homologous regions with ambiguous trace data. **(f–i)** Examples of sequence traces validating the discovery of new SNPs not previously annotated in ENCODE capillary sequencing traces. The top row shows traces from HapMap individuals with the rare variant genotype and the bottom row shows reference traces. Similar analysis was conducted on library B (shown in **Supplementary Fig. 1**). Red arrows mark predicted SNPs in capillary-sequence data.

We analyzed sequenced regions base by base for all individuals by calculating a polymorphism discovery Bayes factor, where K_s is the Bayes factor for polymorphism discovery for the s^{th} base as derived in equation 2 (see Methods). An example plot of K_s across each base (of 50 kb) is shown in **Figure 4** for library A; we conducted a similar analysis for library B (**Supplementary Fig. 1** online).

We next evaluated false positive and false negative rates to assess our experimental and analytical framework for variant discovery (**Table 1** for library B and **Fig. 5** for both library A and B). False positives are particularly difficult to quantify as not all polymorphic sites are known, even in previously resequenced regions. In our analysis,

to be defined as a false positive, a variant must not exactly match the location of variants within the Single Nucleotide Polymorphism database (dbSNP) and must not have trace sequencing data indicating a previously missed variant. In some cases, trace sequence data were not available or were unreliable. Consequently, the false positive rate is expected to be an upper estimate as the exact position must be validated as polymorphic by an existing database. We determined false negative rates by calculating whether a base known to be polymorphic in our library of HapMap individuals reached previously specified K_s thresholds of 3, 10, 100 or 1,000. This calculation of false negative rates does have some bias, as it does not take into account coverage of the polymorphic base.

As expected, setting a higher threshold for K_s gives fewer false positives. For library B, as K_s increased from 10 to 1,000, the false positive rate decreased from 69.6 to 11.3% (**Table 1**). Likewise, with fixed coverage, we observed the false negative rate increasing from 10.8 to 90.8% as K_s increased from 10 to 1,000. A more detailed discussion of false-negative and false-positive rates is available in the **Supplementary Methods**. We found that false negatives occurred when the cumulative coverage of individuals with the

detailed equations). We used Bayes factors to compare the probability that the distribution of mismatched bases arises from sequencing error to the probability that the distribution of mismatches arises from diploid polymorphism. For example, if 20% of reads for a given base were nonconcordant with the reference sequence across all individuals, and the nonconcordant bases were due to the presence of a SNP, one would expect each individual to be homozygous (0% or 100% concordance with reference) or heterozygous (concordance split 50:50). In contrast, if the 20% nonconcordant bases were due to sequencing error, then the number of nonconcordant bases for each individual would follow a binomial distribution around 20% (for example, person 1, ~20.5%; person 2, ~19.3%; person 3, ~20.7%; and so forth). As described below, the error estimates required to calculate the probability of a genetic variant being a true variant are readily obtainable when samples from individuals are indexed and multiplex-sequenced. Additionally, indexed and multiplexed sequencing removes run-to-run biases, which would confound these estimates if all aspects of experimental design were not properly randomized.

Table 1 | Polymorphism discovery and individual variant calling

Polymorphism discovery by K_s threshold ^a				
Threshold (K_s)	Polymorphisms predicted	True positives ^b	False positives ^c	False negatives ^d
3	932	112	88.0%	9.2%
10	352	107	69.6%	10.8%
100	131	99	24.4%	32.3%
1,000	106	94	11.3%	90.8%

Individual variant calling or genotyping (AA, AB, BB) by K_i threshold		
Threshold (K_i)	Genotyped correctly	Genotyped incorrectly
3	3,376	115
10	3,144	58
100	2,677	8
1,000	2,397	7

^aPredicted polymorphic bases at a given threshold for K_s were evaluated by comparison to known polymorphisms within dbSNP and to ENCODE capillary sequencing traces. False negatives rates reflect that greater base coverage is required to exceed larger K_s thresholds and that many polymorphisms become insufficiently covered for polymorphism discovery at these levels (see Fig. 5 for relation between coverage and K_s).

^bValidated in dbSNP or NCBI trace archive. ^cNot identified in dbSNP or NCBI trace archive. ^dRates of polymorphism discovery were evaluated irrespective of coverage. False negatives rates were calculated across Libraries A and B. False positive rates were calculated using only library B since not all regions of library A were previously resequenced within the ENCODE project.

rarer variant was less than 10 reads (Fig. 5). Highlighting the dependence of false negatives on coverage, all polymorphisms that were covered by 20 or more reads (summed across individuals known to differ from the reference) had a $K_s > 1,000$. Overall, we observed that 90% of variants were detectable, though designing experiments for an average of greater than 20 reads will be essential for controlling false negatives.

While analyzing bases with a $K_s > 100$ for false positives using US National Center for Biotechnology Information (NCBI)-archived ENCODE traces, we discovered new SNPs that were evident in visual reinspection of capillary traces but that had not been annotated in dbSNP (Fig. 4f h). These examples demonstrate that index-based resequencing can identify new variants even in heavily sequenced and heavily annotated regions. Notably, within library B, two variants with a $K_s > 100$ were not SNPs but actually insertions (with dbSNP, rs11279266 is a 1-bp insertion and rs10555419 is a 6-bp insertion). Thus, it is possible to identify genetic variants explicitly not allowed within the alignment scheme.

Genotyping individuals at known polymorphisms

As false negatives are clearly tied to coverage, we explored the influence of coverage further by analyzing sequenced regions in an individual-by-individual analysis. Derived in equation 3 (see below), K_i is the analogous Bayes factor for the i th individual having the rarer allele at a known polymorphic base. Conceptually, it can be thought of as a specific individual's contribution to K_s . We calculated the percentage of variants correctly identified in an

individual given a certain number of reads (Fig. 5c). For example, when the coverage for a base was ~ 20 reads (averaging from 16 to 24), we detected >80 90% of the bases at $K_i > 10$, with a false-positive rate of 1.6%. In comparison to polymorphism discovery, the low false positive rates of genotyping at a known polymorphic base are due to the fact that we were no longer assessing thousands of bases for a rare event but rather assessing samples from a few dozen individuals for a more frequent event.

DISCUSSION

Our experience suggests that achieving adequate coverage is one of the most important factors in the design of a multiplexed targeted resequencing experiment. Depending on assumptions made in the experiment, the desired coverage (and as a consequence, the cost) can vary substantially. Key considerations include whether the objective is

(i) discovering genetic variants for genotyping by a separate method such as custom SNP genotyping, (ii) conducting polymorphism discovery and variant calling within one sequencing experiment, and/or (iii) exhaustively resequencing for all common and rare variants.

Exhaustive polymorphism discovery is the next major phase for genome-wide association studies. Indexing of short-reads was surprisingly robust at polymorphism identification. For example,

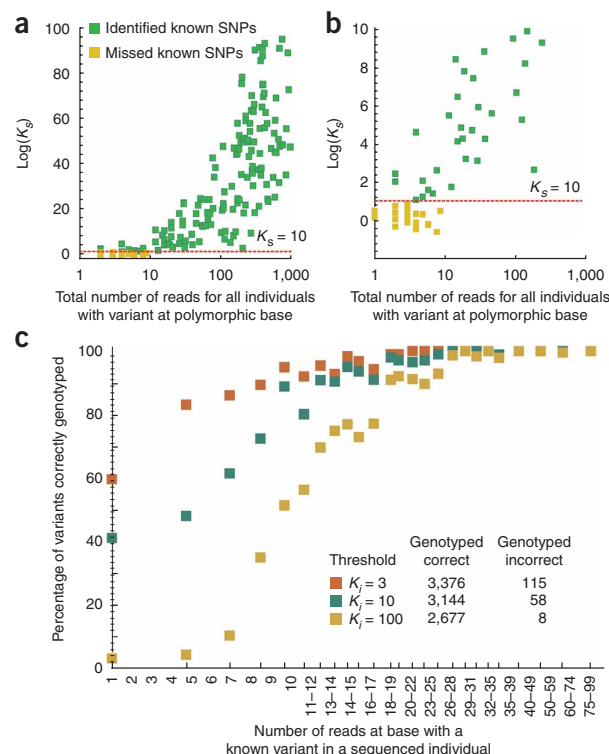


Figure 5 | Relationship between base-level coverage and Bayes factor for polymorphism discovery and variant genotyping. (a) Total coverage across those individuals with a nonreference genotype at a known polymorphism. (b) Magnification of the graph in a. (c) The percent of the time the correct genotype was determined versus the coverage of the variant within the individual. Plots contain cumulative statistics using variant discovery and genotyping within both libraries A and B.

even when we used a highly restrictive error alignment scheme, we identified a new coding SNP 3 bases from an annotated SNP. Additionally, it is encouraging to see the discovery of insertions using an alignment scheme only allowing substitutions. Finally, we show that an automated analysis strategy for short-read data can uncover new variants, even in regions that had been previously sequenced for variant discovery using samples from these individuals in the HapMap database.

Based on the false positive and false negative rates, a critical factor will be how to balance coverage and cost. Using $K_s > 1,000$, we observed a low false positive rate and a high false negative rate. As we required that the exact base must be polymorphic in an existing database, the actual false-positive rate may be lower. False negatives are due to the additional coverage (> 10 reads) required for overcoming higher K_s thresholds. Considering the substantial base-to-base variability (Fig. 3), one would not simply want to design for an average coverage of 10 reads. Rather, by designing for ~ 50 reads or more, one may minimize both the false negative and false-positive rates, given coverage variability of short-read sequencing.

Although whole-genome sequencing may be the primary motivator for improvements in sequencing technology, it is clear that next-generation technologies are immediately useful for focused, hypothesis-driven sequencing of linkage peaks, groupings of candidate genes or sequencing the entire known coding sequence of the human genome. In this report, we developed per-individual indexing of pooled PCR amplicons to carry out targeted sequencing. However, it is straightforward to envision using other sample-preparation methods such as genome partitioning^{9,12}. One could replace the pooled amplicons from our experimental strategy with total genomic DNA and complete the partitioning by a hybridization approach after pooling ligated amplicons. Indeed, variant discovery through the resequencing of all candidate regions implicated in a disease across genomes of dozens, and possibly hundreds, of individuals could be considerably accelerated by merging multiplex capture, indexing and next-generation sequencing approaches into a single protocol.

METHODS

Amplification and pre-ligation sample preparation. Two amplicon libraries (library A and B; see **Supplementary Table 2** for specific regions) were constructed from individually amplified 5-kb regions using long-range PCR. Regions comprising a library were chosen from the ENCODE project and provide a sampling of different genomic region types. Only a portion of the ENCODE regions have been previously sequenced as part of the SNP discovery portion of the ENCODE project. Library A was a composite of previously sequenced regions and regions that had not been sequenced in ENCODE. Library B was entirely composed of regions that had been previously sequenced. Flanking primers for each amplicon (**Supplementary Table 5** online) were manually selected. Although we did not screen for the existence of known polymorphisms within the primer sequences, this would be advisable in future experiments.

Base calling and alignment. Illumina GA images were analyzed using a modified Illumina GA (0.2.2.6) processing pipeline. Descriptions of the modifications and all scripts are available in **Supplementary Methods**. A default matrix deconvolution file was

used for base-calling by 'Bustard', based on a control phi-X library provided by Illumina. Following base-calling, a script was used to access all sequences, regardless of quality score. We deviated from the default cut-offs provided by Illumina's 'GERALD' process because we found that sequence quality was better controlled by matching to an index (46/4,096 or 0.1% by chance) or matching with 1 or fewer errors to the reference sequence. Bases were aligned by progressively truncating the sequence at the 5' end until a unique alignment was obtained with a probability of stochastic alignment of less than 1%. This approach is distinct from recently described short-read alignment schemes¹³ but has some of the same features.

Aligned sequences were summarized into a binomial of counts agreeing (a_i) or disagreeing (b_i) with the reference sequence for the i^{th} individual of n total individuals for each base (s). The error rate (θ_s) is the percentage of reads disagreeing with the reference sequence across all samples. Model 1 (M_1) assumes that the error rate for an individual equals the error rate for all individuals, or $\theta_i = \theta_s$. Therefore, we can estimate θ_s as:

$$\hat{\theta}_s = \sum_{i=1}^n \left(\frac{a_i}{a_i + b_i} \right). \quad (1)$$

Model 2 (M_2) assumed that for some individual i , we have $\theta_i \neq \theta_s$. To calculate this likelihood, we used a hyperprior offset (σ_s) for three possible genotypes. The hyperprior can be thought of as conditioning on the zygosity of the individual and thus reflecting the uncertainty of the genotype of the individual at a given base. In this analysis, we focused on the detection of biallelic SNPs, but triallelic or other types of SNPs could be considered under more complex models. The Bayes factor for a base position is

$$K_s \frac{p(\theta_s|M_1)}{p(\theta_s|M_2)} = \frac{\prod_{i=1..n} \left(\frac{a_i + b_i}{a_i} \right) \times \hat{\theta}_s^{b_i} \times (1 - \hat{\theta}_s)^{a_i}}{\int_0^1 \prod_{i=1..n} \left(\frac{a_i + b_i}{a_i} \right) \times \theta_s^{b_i} \times (1 - \theta_s)^{a_i} \times p(\theta_s|M_2, \sigma_s) \times p(\sigma_s) \times d\theta}. \quad (2)$$

In equation 2, K_s is the Bayes factors across all individuals and is calculated for each SNP. The value for K_s is effectively identifying polymorphisms at a base, and determination of the individuals that show the rare variant is accomplished by K_i :

$$K_i \frac{p(\theta_s|M_1)}{p(\theta_s|M_2)} = \frac{\left(\frac{a_i + b_i}{a_i} \right) \times \hat{\theta}_s^{b_i} \times (1 - \hat{\theta}_s)^{a_i}}{\int_0^1 \left(\frac{a_i + b_i}{a_i} \right) \times \theta_s^{b_i} \times (1 - \theta_s)^{a_i} \times p(\theta_s|M_2, \sigma_s) \times p(\sigma_s) \times d\theta}. \quad (3)$$

The analysis for rare variants in the above equation uses the prior probability of $p(\sigma_s)$ derived from the distribution across all individuals and all bases initially assuming all other individuals are homozygous for the reference allele, which is a reasonable assumption given that new variants are assumed to be rare. By successive iterations, $p(\sigma_s)$ is then recalculated by calling variants AA, AB and BB (where A and B refer to the polymorphism alleles) under a given threshold (for example, $K_i = \{3, 10, 100\}$) and recalculating $p(\sigma_s)$ based on these variant calls. Plots of Bayes factors versus error rates are available in **Supplementary Fig. 2** online to show distribution of Bayes factors and their correlation with error rate.

If a SNP is known, it is also reasonable to use other prior information at a base, such as allele frequency in the population. However, prior information on the location and allele frequency of known SNPs was not used in this study to better evaluate the effectiveness of the underlying framework.

Additional methods. Descriptions of region amplification, amplicon pooling, fragmentation, preparation of the 48 adapters, end repair, ligation, PCR enrichment, cluster generation and sequencing on the Illumina Genome Analyzer are available in **Supplementary Methods**.

Accession code. NCBI Short Read Archive: SRA001811.

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

ACKNOWLEDGMENTS

We acknowledge funding from the state of Arizona, US National Heart Lung and Blood Institute (U01 HL086528), the Stardust foundation, Science Foundation Arizona, and National Institute for Neurological Disorders and Strokes (R01 NS059873).

AUTHOR CONTRIBUTIONS

D.W.C., J.V.P., M.J.H., G.N. and D.A.S. contributed to initial experimental design. S.S., A.S., M.R., J.J.C., T.L. and T.L.P. contributed to development and execution of exact experimental protocols. J.V.P., D.W.C. and N.H. contributed to the development of bioinformatics and analysis pipelines.

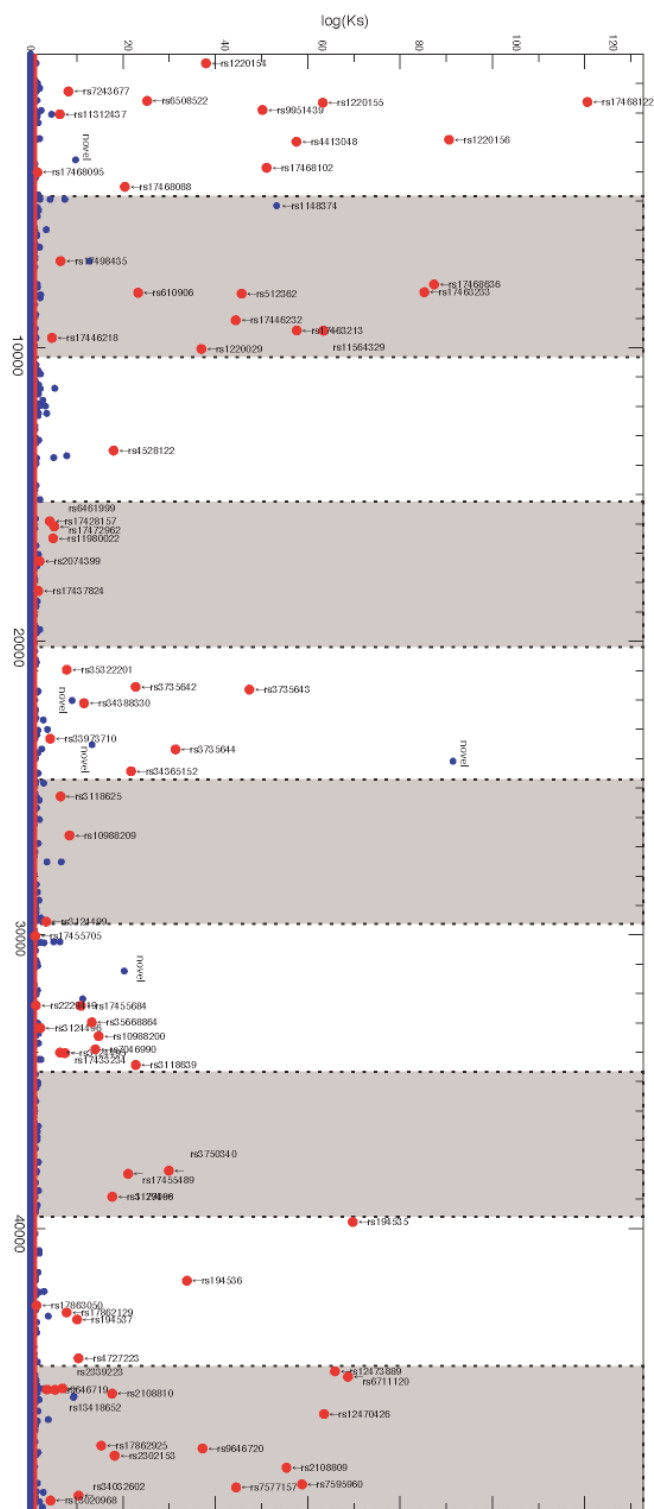
COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

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

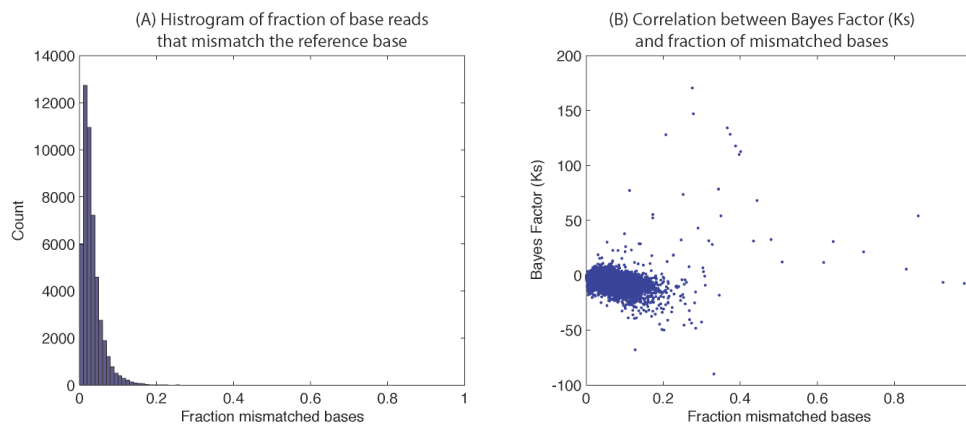
1. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
3. Zondervan, K.T. & Cardon, L.R. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2**, 2492–2501 (2007).
4. Meyer, M., Stenzel, U., Myles, S., Prüfer, K. & Hofreiter, M. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.* **35**, e97 (2007).
5. Parameswaran, P. *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* **35**, e130 (2007).
6. Milosavljevic, A. *et al.* Pooled genomic indexing of rhesus macaque. *Genome Res.* **15**, 292–301 (2005).
7. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237 (2008).
8. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
9. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
10. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
11. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
12. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
13. Jeck, W.R. *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942–2944 (2007).

Supplementary Figure 1.



Supplemental Figure 1. Library B polymorphism discovery. Similar to Figure 4, variant discovery plot for Library B. As with Library A, inspection of traces from NCBI Trace Archive confirmed polymorphic bases that were previously not annotated.

Supplementary Figure 2



Supplementary Figure 2. Bayes factors vs. non-reference bases alignments. (A) A histogram of the fraction of mismatches at each base indicating that a substantial percentage of bases have 15% of the aligned bases disagreeing with the reference base (B.) Scatterplot of K_s vs. fraction of mismatched. The lack of a positive correlation indicates that most bases favor the model of a binomial distribution even as the fraction of mismatched bases increases. Taken together, these data suggest that most bases sample error, as expected, from a binomial distribution across individuals. Those bases that do not are possible polymorphisms.

Supplementary Table 1

Region #	Amplicon Primer Sequence	Region (Build 36.1)
Region 1	Gagatgggtctcctgagtc	chr2:234189304+234194258
	Aaccaactgcatgctttc	
Region 2	Ccagccaaactgacgtacc	chr7:113839924+113844896
	Aacagcccaattcaacttc	
Region 3	gaccaaggaaacaaccaacc	chr20:33747673-33752665
	caggaaagcccacatacacc	
Region 4	ctaactcagcggcttgtcc	chr2:220085623+220090673
	tcacgtccttttgagacc	
Region 5	gggagctcacgatatcaagg	chr11:130743490+130748475
	caggaagcagcagctctagg	
Region 6	gctgggtgttgatatttc	chr1:149678673-149683603
	cagtaggcaaggacacatgc	
Region 7	taacagtgggctgaaagc	chr5:142043190-142048159
	cttgggagctccaggtacg	
Region 8	gtgaaaccagggacaatgg	chr5:141971045-141976067
	ctcctcacctgcagaacc	
Region 9	gccatgggagttaacagagg	chr7:126531683-126536662
	ttgctaccattgccattcc	
Region 10	tggccagtttgtttcagg	chr11:116250173-116255189
	tattgggacgaaggattgc	
Region 11	agtagggtagcttgatgc	chr13:112347206+112352113
	ccgcatgactttgtttgg	
Region 12	gaccttgtatctgccttc	chr16:26053412+26058370
	gaggggctcctaaagttgc	
Region 13	atggagtttgcctttgttc	chr11:63970822+63975969
	agttgtccctgtggctatgg	
Region 14	tcagtgtgctgtactcatgc	chr16:41060-46001
	aggacacctggggattacg	

Supplementary Table 1. Primers. Primers for amplification of targeted regions within Library A.

Supplementary Table 2.

Encode Regions	Multiplex Pool	Chr	Physical Position Region (36.1)	Region Features
ENr231	A,B	1	149678673-149683603	Moderate non-exonic conservation, High Gene density
ENr331	A	2	220085623-220090673	High-exonic conservation, High Gene density
ENr131	A,B	2	234189304-234194258	Low-exonic conservation, High Gene density
ENr212	A	5	141971045-141976067	Moderate non-exonic conservation, Moderate Gene density
ENr212	A	5	142043190-142048159	Moderate non-exonic conservation, Moderate Gene density
ENm014	A	7	126531683-126536662	7q31.33
ENm017	B	7	127008375-127012881	7q32.1
ENm017	B	7	127013024-127018100	7q32.1
ENm017	B	7	127016306-127021248	7q32.1
ENm010	B	7	27246985-27251931	HOXA Cluster
ENm012	B	7	89709475-89714560	FOXP2
ENr335	B	9	130810649-130815588	High-exonic conservation, High Gene density
ENr335	B	9	130908740-130913660	High-exonic conservation, High Gene density
ENm003	A	11	116250173-116255189	Apo_cluster
ENr312	A	11	130743490-130748475	High-exonic conservation, Low Gene density
ENr123	B	12	38507675-38512629	Low-exonic conservation, High Gene density
ENr123	B	12	38783227-38788147	Low-exonic conservation, High Gene density
ENr123	B	12	38903255-38908283	Low-exonic conservation, High Gene density
ENr132	A	13	112347206-112352113	Low-exonic conservation, High Gene density
ENr213	B	18	23783347-23788366	Moderate-exonic conservation, Low gene density
ENr213	B	18	23808628-23813466	Moderate-exonic conservation, Low gene density
ENr333	A	20	33747673-33752665	High-exonic conservation, High Gene density

Supplementary Table 2. Selected Regions. Two different indexed pools were sequenced across a series of runs (referred to as multiplex Library A and B). The ENCODE region name, and key ENCODE attributes are listed.

Supplementary Table 3

Phos	5'-3' index	Sequence	5'-3' index
P-	TTT TTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAA AAT
P-	GGT TGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAA CCT
P-	CCT TCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAA GGT
P-	AAT TAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAA TTT
P-	GTG TGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAC ACT
P-	TGG TTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAC CAT
P-	ACG TAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAC GTT
P-	CAG TCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAC TGT
P-	CTC TCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAG AGT
P-	AGC TAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAG CTT
P-	TCC TTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAG GAT
P-	GAC TGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAG TCT
P-	ATA TAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAT ATT
P-	CGA TCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAT CGT
P-	GCA TGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAT GCT
P-	TAA TTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAT TAT
P-	GTT GGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCA ACT
P-	TGT GTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACA CAT
P-	ACT GAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCA GTT
P-	CAT GCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCA TGT
P-	TTG GTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACC AAT
P-	GGG GGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCC CCT
P-	CCG GCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCC GGT
P-	AAG GAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCC TTT
P-	ATC GAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCG ATT
P-	CGC GCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCG CGT
P-	GCC GGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCG GCT
P-	TAC GTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACG TAT
P-	CTA GCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCT AGT
P-	AGA GAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCT CTT
P-	TCA GTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACT GAT
P-	GAA GGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCT TCT
P-	CTT CCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGA AGT

Phos	5'-3' index	Sequence	5'-3' index
P-	AGT CAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGA CTT
P-	TCT CTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGA GAT
P-	GAT CGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGA TCT
P-	ATG CAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGC ATT
P-	CGG CCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGC CGT
P-	GCG CGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGC GCT
P-	TAG CTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGC TAT
P-	TTC CTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGG AAT
P-	GGC CGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGG CCT
P-	CCC CCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGG GGT
P-	AAC CAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGG TTT
P-	GTA CGA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGT ACT
P-	TGA CTA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGT CAT
P-	ACA CAA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGT GTT
P-	CAA CCA	GATCGGAAGAGCTCGTATGCCGCTCTTCTGCTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGT TGT

Supplementary Table 3. Full oligonucleotide sequences for indexed adapters.

Supplementary Table 4.

DNA Indexes Appended to Adapter Sequence			
AAAAAT	CAACCT	GAAGGT	TAATTT
AACCAT	CACACT	GACTGT	TACGTT
AAGGAT	CAGTCT	GAGAGT	TAGCTT
AATTAT	CATGCT	GATCGT	TATATT
ACACAT	CCAACCT	GCATGT	TCAGTT
ACCAAT	CCCCCT	GCCGGT	TCCTTT
ACGTAT	CCGGCT	GCGCGT	TCGATT
ACTGAT	CCTTCT	GCTAGT	TCTCTT
AGAGAT	CGATCT	GGAAGT	TGACTT
AGCTAT	CGCGCT	GGCCGT	TGCATT
AGGAAT	CGGCCT	GGGGGT	TGGTTT
AGTCAT	CGTACT	GGTTGT	TGTGTT

Supplementary Table 4. DNA Indexes Appended to Each Adapter. A total of 48 different 6-mer index sequences were chosen from the 4096 different possible 6-mers. Each index is designed so that the 1st and 5th base are identical and represent an XOR-based checksum of bases 2-4 so that the index sequence remains identifiable even with an uncalled or low-quality base.

Supplementary Table 5.

Library A dbSNP identifier	log(Ks)	Coverage
rs12533005	51.1	355
rs11167787	74.0	713
rs11960262	47.7	455
rs11167786	55.7	467
rs10210058	18.7	452
rs34012	8.1	417
rs1042597	4.6	297
rs13027376	2.9	297
rs13400017	7.9	397
rs11167785	48.9	528
rs1361963	63.8	859
rs6975798	34.1	443
rs152524	5.4	372
rs1557644	24.0	774
rs10791140	55.6	1645
rs907676	5.1	179
rs11222591	22.7	761
rs250108	5.1	390
rs1872858	14.1	168
rs2237790	31.9	686
rs4114768	16.4	642
rs11563720	10.0	541
rs249926	33.5	1514
rs17092980	0.5	601
rs4291502	4.0	533
rs4528122	23.5	58
rs12705964	4.5	950
rs7727653	4.4	398
rs17099096	3.2	362
rs17864670	5.0	405
rs9919599	2.4	40
rs11955552	4.3	825
rs681524	6.9	611
rs17864673	13.1	1009
rs17868309	4.6	815
rs10221563	9.3	374
rs11975039	1.2	832
rs17099102	1.7	709
rs17099100	3.2	598
rs7119590	2.1	1707
rs10241421	-1.7	0
rs17099249	-1.5	17
rs12292614	-1.3	0
rs1042590	-1.7	191
rs17867764	-1.3	0
rs3088078	-0.7	0
rs11222590	-2.5	5
rs17863762	0.1	0
rs1126806	-1.5	0

Library B dbSNP	Log(Ks)	Coverage
rs13289043	0	0
rs17445742	0	0
rs17863844	0	0
rs17437810	0.5724	1
rs3814492	0.3001	1
rs17468650	0.138	2
rs17468601	-0.0202	3
rs17446218	4.6375	4
rs2229419	1.1383	4
rs9945664	-0.3719	4
rs17437824	1.6533	6
rs4291502	-0.1137	6
rs17445496	-0.1526	6
rs17468095	1.4706	7
rs10988209	8.4219	15
rs4727223	10.3836	16
rs3118625	6.4789	16
rs17428157	4.1719	16
rs11980022	4.8702	18
rs7046990	14.0197	20
rs17862129	7.7949	20
rs4528122	17.9849	26
rs3124495	7.4348	27
rs10988200	14.7299	29
rs3118639	22.777	30
rs13020968	4.3019	40
rs4363925	5.6175	50
rs17468088	20.4266	110
rs194537	10.06	116
rs13418652	5.2823	137
rs3750340	29.9307	144
rs7243677	8.1997	152
rs17446232	44.4083	170
rs7577157	44.4728	172
rs4800835	24.5798	174
rs2302153	18.1967	212
rs17468102	51.1039	219
rs7595960	58.7743	227
rs194536	33.8634	239
rs9951439	50.2121	243
rs17463213	57.6308	260
rs1220155	63.2521	271
rs512362	45.6942	272
rs194535	69.7629	275
rs1220029	36.9822	278
rs12470426	63.5206	288
rs11564329	63.3966	291
rs1530380	62.8385	291
rs12473889	65.888	298
rs9646720	37.2413	343
rs2108809	55.3984	358
rs4768189	44.4388	358
rs17463233	85.2413	367
rs1220154	37.9816	376
rs7962260	64.8659	385
rs17468636	87.3576	392
rs4413048	57.5525	453
rs17468122	120.5659	469
rs610906	23.2828	470
rs1148374	53.2736	571
rs6508522	25.2233	583
rs6711120	68.73	608
rs1220156	90.5985	678

Supplementary Table 5. Polymorphism coverage and K_s. List of SNPs containing known variants, their coverage, and their K_s in sequenced regions for Library A (left) and Library B (right).

Supplementary Methods

Amplification of targeted regions. 46 HapMap individuals were whole genome amplified by RepliG (Qiagen) and the concentration determined by replicate measures using the Quant-iT PicoGreen dsDNA kit (Invitrogen). For each HapMap individual, the ENCODE regions were amplified by long-range PCR in a 25ul reaction volume using 75ng template DNA, 1X PfuUltra Buffer, 2mM dNTPs (total), 400nM each of the forward and reverse primers, and 0.5ul PfuUltra II HS DNA polymerase (Stratagene) per reaction. The thermal cycler conditions were: a denaturation step at 95°C for 2 minutes, 30 cycles consisting of 95°C for 20 seconds, an annealing temperature specific to each primer for 20 seconds, and 68°C for 3 minutes, and a final extension of 68°C for 5 minutes. To generate a sufficient amount of amplicons, the initial PCR products were put through a second PCR reaction. This second-round reaction used 2ul of the initial PCR product, 2mM dNTPs (total), 400nM each of the forward and reverse primers, 1.5ul PfuUltra II HS DNA polymerase (Stratagene), and 1x PfuUltra buffer in a 100ul reaction. The thermal cycler conditions were the same for both the first and second round PCR. Products were purified on QiaQuick 96 well columns (Qiagen) and quantified by taking optical density and Picogreen measurements. A 4-fold range of concentrations was observed after the second round of PCR so the products were titrated such that equimolar amplicons from each individual were pooled and the combined total DNA was ~5µg. Pooled amplicons were digested to 200-300 bp fragments using DNase I enzyme (NEB). Fragmented pools were then blunt end repaired using T4 DNA Polymerase, DNA polymerase I Klenow fragment, and T4 polynucleotide kinase enzyme (NEB). Subsequently, dATP incorporation was performed to the blunt ended amplicons with Klenow Fragment 3'-5' exo minus enzyme (NEB). DNA was purified after each step using Illumina-recommended Qiagen 96 well purification columns.

Indexing adapter preparation. A total of 48 different adapters were produced to index sequenced fragments (See supplementary tables 2-6). The adapter sequence began with the oligonucleotide sequences provided by, and shared with the permission of, Illumina, (© 2006 Illumina, Inc.; 5' P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG and 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT) and were followed by the index in the forward and reverse directions respectively. Lyophilized, indexed adapters were dissolved in 10mM Tris-HCl pH 7.8 (Sigma) to a 100uM stock concentration. A 10X annealing buffer was made containing 100mM Tris-HCl pH 7.8 and 0.5M NaCl (Gibco). Indexed adapter pairs were combined to a 50uM final concentration with 10X annealing buffer for a final concentration of 1X. Subsequently, a step down annealing reaction was performed, where the indexed adapter mix was incubated at 95°C for 2 minutes followed by a series of cooling steps of 1°C per minute from 95°C to 25°C.

Adapter ligation, PCR enrichment and sequencing of indexed amplicons. A unique indexed-adapter sequence was ligated to each HapMap individuals' adenylated amplicon pool. Ligation was performed at 20°C for 2 hours followed by 16°C for 16 hours using T4 DNA ligase enzyme (NEB). Ligated amplicons were then pooled for all individuals to be sequenced on the same flow cell lane. The pooled ligated amplicons (referred to as a library) were loaded onto a 4% agarose gel and 150-200 bp fragments were gel-purified using Qiagen gel purification columns. Libraries were PCR-enriched using Phusion DNA polymerase PCR mix and adapter compatible primers 1.1 and 2.1 (Illumina). Each PCR product was loaded onto a 4% agarose gel and fragments of 150-200bp were gel-purified using Qiagen gel purification columns. Each library was quantified using Nanodrop ND-1000 and diluted to 10nM working concentration using EB buffer. Each library was loaded on a lane of the flowcell for cluster generation and libraries were sequenced on an Illumina GA. Initial cluster counts typically ranged from 1M to 2M clusters per flow cell lane. In later runs, cluster counts of approximately 8-10M clusters per

flow cell lane were typically observed during indexed sequencing. Also in later runs (not shown), we found that the same indexing strategy of ligating barcoded adapters was effective in paired-end sequencing on the Genome Analyzer II upgraded system. In these runs, 5-base indexes were utilized. With 5-base paired-end indexing runs, indexes from both the forward and reverse reads are sequenced for a total of 10 indexed bases.

Calculation of false-positives. False negative rates were determined by calculating if a base known to be polymorphic in our library of HapMap individuals reached previously specified K_s thresholds. Within Library A, 152 SNPs were listed in dbSNP. Of these SNPs, 50 SNPs have at least one individual varying in genotype from the other sequenced individuals. We observed 41 of the 50 SNPs had a base-level $K_s > 10$, with 40 of 41 exceeding $K_B > 100$. Of the 9 false-negative SNP positions, 1 SNP was triallelic and not entirely compatible with our biallelic analysis model, 1 SNP was found only in 2 individuals and the other 7 SNPs were found in only 1 individual. Similar results were obtained for Library B (table 1). The false negatives appear to result from lack of coverage (figure 5a) and not simply because only 1 individual had the variant: of the 7 single-event false negatives, 3 individuals had less than 2 reads at the variant position. False positives are more difficult to quantify since not all polymorphic sites are known, even in previously resequenced regions. In our analysis, to be defined as a false positive, the base must reach a pre-specified K_s , must not exactly match the location of variants within dbSNP, and must not have trace sequencing data indicating a missed variant. Library B, which is entirely composed of regions with existing ENCODE sequencing data, was used to specifically estimate false-positives (table 1). Immediately evident by visual inspection (Figure 4) is an overall low false positive rate for bases with large K_s values; most high-ranking bases ($K_s > 100$) are also listed in dbSNP (see Figure 4a). Particularly within the two ENCODE sequenced regions, 25 of the top 25 ranked bases were at the exact position of the SNP. In previously non-sequenced regions, we could identify many candidate variants. For instance, in region 2, there were 20 bases with $K_s > 3$. For the 8 highest ranking bases, 4 were unambiguously confirmed as novel SNPs. Of the remaining 4 bases, 1 did not have sufficient trace data, 1 was neighboring an identified SNP and neighboring a repeat region, 1 was in a repeat region, and 2 were in regions highly homologous to other regions in the genome. For the remaining 12 high-ranking bases, we were not able to confirm the existence of a SNP. For these bases, trace data frequently exhibited multiple reads, was poor in quality in both read directions, or was within a location with high sequence homology (see examples of difficult-to-assess traces in Figure 4d and 4e). In some cases, high quality sequence data was not available for the exact individuals driving the ranking. However, it is likely that these 12 unconfirmed candidate variants are false-positives and we treat them as such for our analysis. Indeed, the most prominent source of false positives with our approach appears to arise from homology to other regions or polymorphic sites elsewhere in the genome. In practice, one would expect lower false positive rates for less complex regions.