Identification of genetic variants using bar-coded multiplexed sequencing

David W Craig[1,3], John V Pearson[1,3], Szabolcs Szelinger[1,3], Aswin Sekar[1], Margot Redman[1], Jason J Corneveaux[1], Traci L Pawlowski[1], Trisha Laub[1], Gary Nunn[2], Dietrich A Stephan[1], Nils Homer[1] & Matthew J Huentelman[1]

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

[1]The Translational Genomics Research Institute, 445 N. 5th St. 5th Floor, Phoenix, Arizona 85004, USA. [2]Illumina, 9885 Town Centre Drive, San Diego, California 92121,

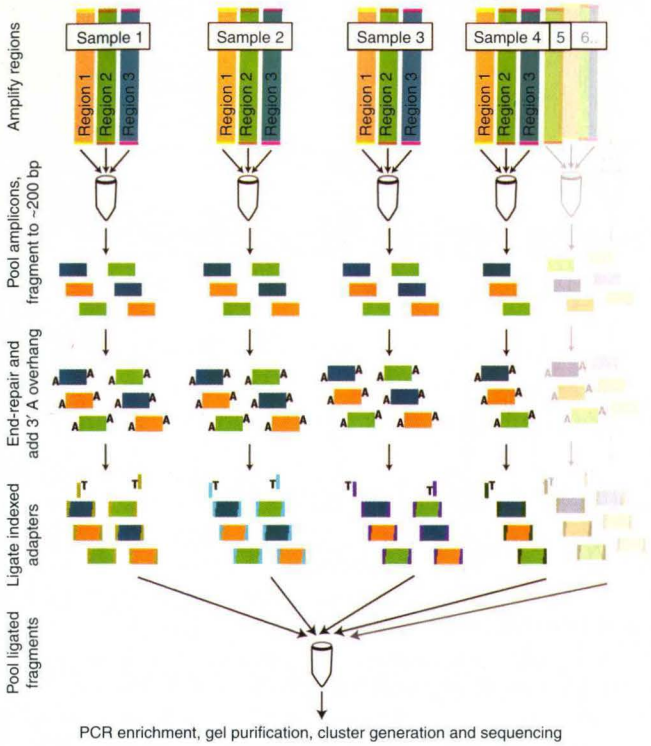

Figure 1 | Schematic describing the preparation of indexed libraries.

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

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 adapters 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

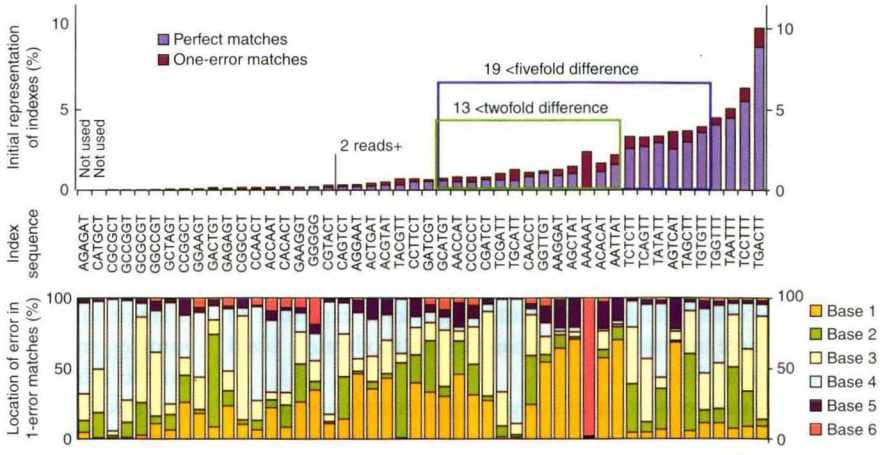

Figure 2 | Comparison of index performance. Index variability in initial sequencing runs (library A) used for evaluating index performance are shown (top). Percentages of reads aligning to the reference sequence are listed by index, without introduction of normalization methods. A total of 30 indexes were

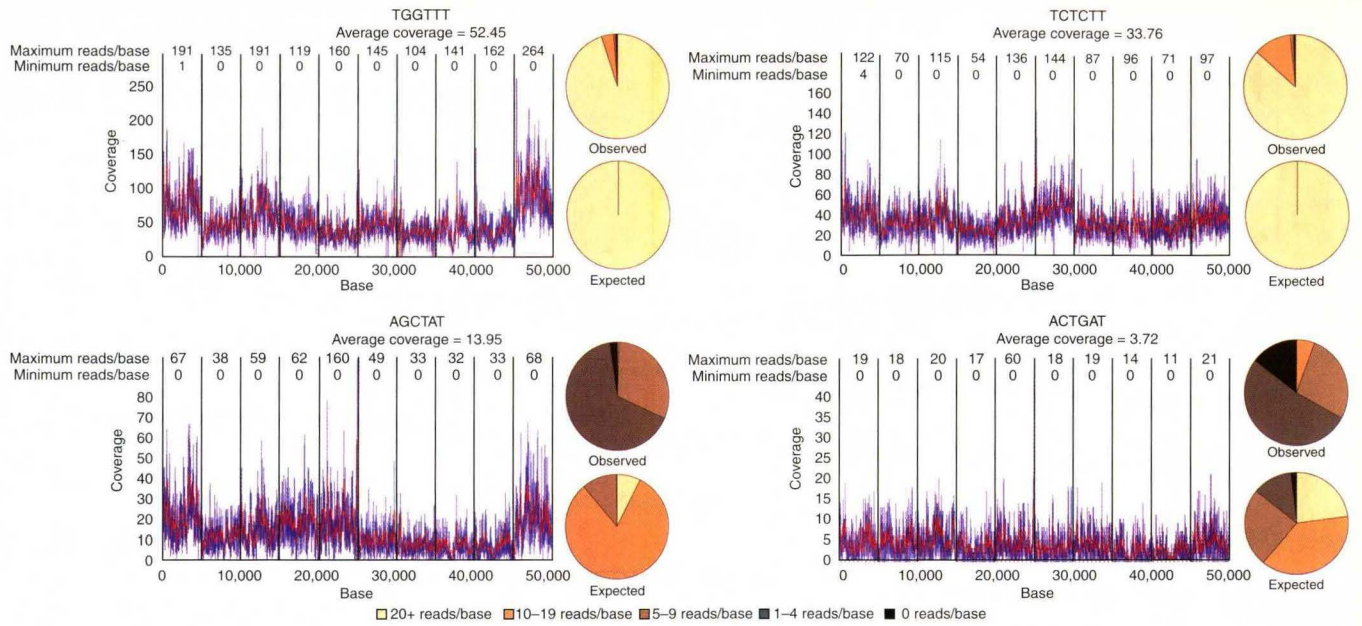

Figure 3 | Relationship between mean and local coverage. Example coverage in 4 individuals sequenced within a single lane of an 8-lane flow-cell for 10 pooled amplicons as part of library A. Amplicons are shown consecutively for each individual. Index sequence and mean coverage for that individual are 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

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.

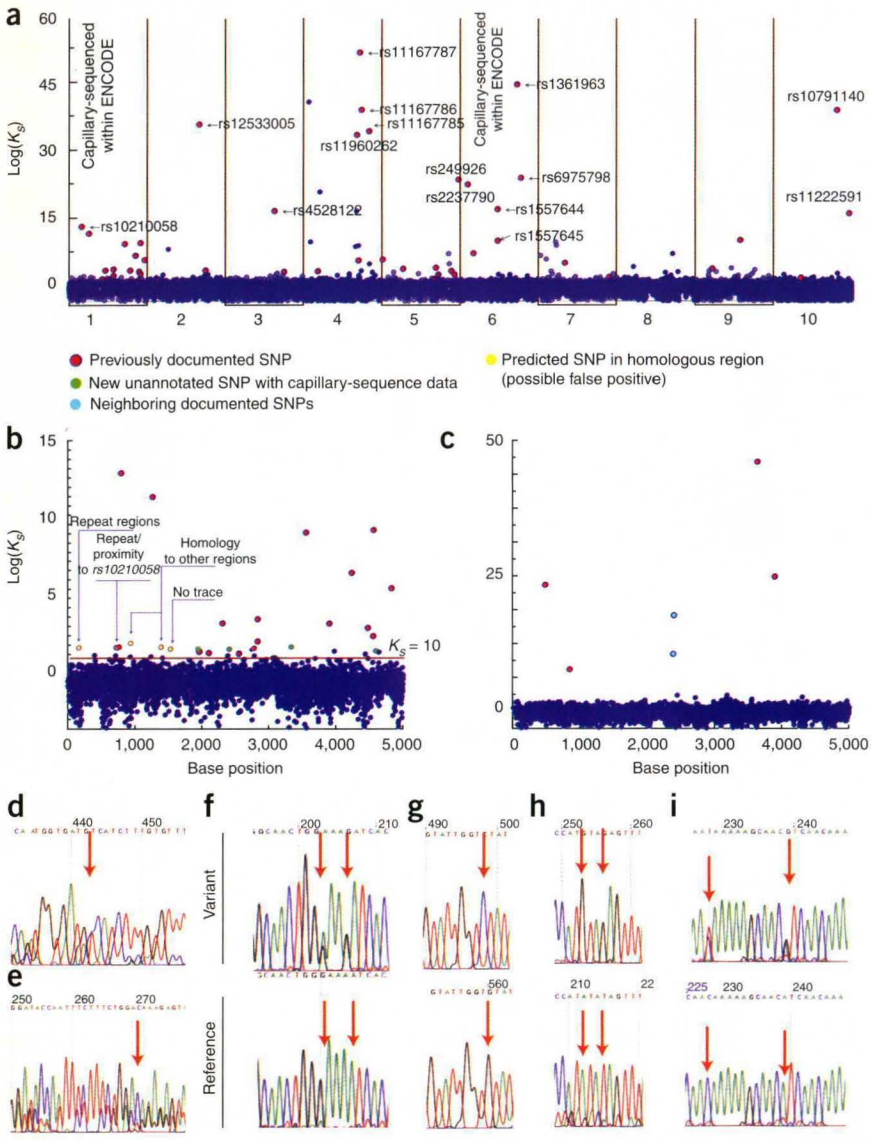

Figure 4 | Discovery of variant bases by simultaneous analysis of all individuals. (**a**) The Bayes factor for polymorphism discovery (K_s) is plotted for each of the 10 sequenced 5-kb amplicons from library A. Exact positions matching known polymorphisms are colored as red spheres, and the dbSNP identifier 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.

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 con-

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

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

[a]Predicted 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). [b]Validated in dbSNP or NCBI trace archive. [c]Not identified in dbSNP or NCBI trace archive. [d]Rates 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 ith 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,

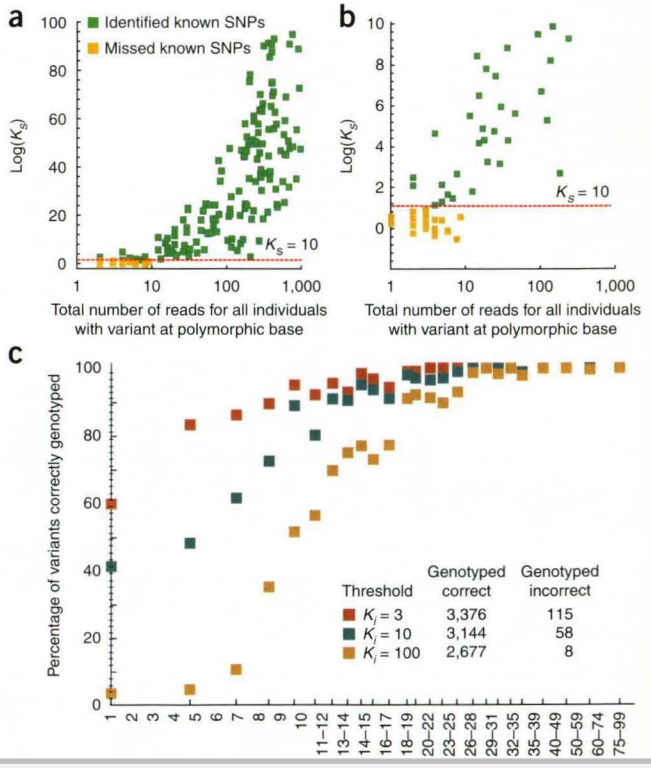

Figure 5 | Relationship between base-level coverage and Bayes factor for polymorphism discovery and variant genotyping. (**a**) Total coverage across those individuals with a nonreference genotype at a known polymorphism. (**b**) Magnification of the graph in **a**. (**c**) The percent of the time the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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.