Comprehensive variation discovery in single human genomes

Published in final edited form as:
Nat Genet. 2014 December ; 46(12): 1350–1355. doi:10.1038/ng.3121.

Comprehensive variation discovery in single human genomes

Neil I. Weisenfeld, **Shuangye Yin**, **Ted Sharpe**, **Bayo Lau**, **Ryan Hegarty**, **Laurie Holmes**, **Brian Sogoloff**, **Diana Tabbaa**, **Louise Williams**, **Carsten Russ**, **Chad Nusbaum**, **Eric S. Lander**, **Iain MacCallum**, and **David B. Jaffe**[*]
The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA

Abstract

Complete knowledge of the genetic variation in individual human genomes is a crucial foundation for understanding the etiology of disease. Genetic variation is typically characterized by sequencing individual genomes and comparing reads to a reference. Existing methods do an excellent job of detecting variants in approximately 90% of the human genome, however calling variants in the remaining 10% of the genome (largely low-complexity sequence and segmental duplications) is challenging. To improve variant calling, we developed a new algorithm, DISCOVAR, and examined its performance on improved, low-cost sequence data. Using a newly created reference set of variants from finished sequence of 103 randomly chosen Fosmids, we find that some standard variant call sets miss up to 25% of variants. We show that the combination of new methods and improved data increases sensitivity several-fold, with the greatest impact in challenging regions of the human genome.

Introduction

Accurate determination of an individual's genome is essential for understanding both human hereditary disease and cancer. Advances in genome sequencing have made it possible, at

[*] Corresponding author: jaffe@broadinstitute.org.

AUTHOR CONTRIBUTIONS

D.B.J., I.M and C.R. designed and managed the project. S.Y., N.I.W., T.S., B.L., I.M. and D.B.J. designed algorithms and analyzed data. R.H., L.H., B.S., D.T. and L.W. designed and executed laboratory processes.

D.B.J., C.N. and E.S.L. shaped the results and developed the manuscript, to which all authors contributed. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

URLs

DISCOVAR, http://broad.io/discovar

Segmental Duplication DB,
http://humanparalogy.gs.washington.edu/build37/build37.htm

ACCESSION CODES

NCBI SRA:

SRX297987 (NA12878; primary dataset for this work)
SRX264790, SRX264793 (NA12878; additional PCR-free data)
SRX264763, SRX264794 (NA12892; mother of NA12878)
SRX264762, SRX264792 (NA12891; father of NA12878)
SRX272091, SRX272089 (Fosmid pool data; NA12878; Illumina platform)
SRX272084, SRX272088, SRX272083, SRX272087, SRX276681 (Fosmid pool data; NA12878; PacBio SMRT platform)

GenBank:

PRJNA196715 (BioProject accession for 103 Fosmid reference sequences) relatively low cost, to generate whole genome shotgun (WGS) sequence reads covering nearly the entire human genome. A critical challenge is then to use such WGS reads to fully reconstruct an individual's genome and to identify all of its variation relative to a reference. While there has been substantial progress toward this goal, current methods for calling variants remain imperfect[1–10].

The development of better variant calling methodologies is limited by the difficulty of assessing the accuracy and completeness of a new method. In principle, one could assess variants called by a given method by comparing to a list of 'true' variants, derived from perfect knowledge of the sequence of the DNA source and the reference genome. In practice, it is impossible to gain perfect knowledge of a target genome, so common procedure is to compare to an established reference set of variant calls (e.g. HapMap3[11]). Based on such comparisons, current variant calling methods are estimated to be approximately 99% complete[3].

Because systematic biases against certain genomic regions or variant types cause variants to be missing from both the reference dataset and the set being evaluated, such comparisons may overestimate the completeness of variant call sets. While 'ordinary' variants are readily detected, some variants are particularly challenging to identify – for example, those occurring in low-complexity sequence, segmental duplications and extremely high %GC regions. Importantly, these challenging regions have long been known to contribute to mutations underlying human disease[12,13].

We therefore set out to define a 'truth set' containing *all* variants present in a random sample of the genome, as a foundation for studying the completeness of variant calls. We focused on the well-studied cell line GM12878 (DNA sample identifier NA12878), generated finished sequence for 103 randomly selected Fosmid clones, mapped these back to the human reference sequence, and identified all variants in the ~4 Mb spanned by the Fosmids. When we compared the Illumina 'Platinum' variant call set (based on 100-base reads) to the Fosmid reference set, we found that it omitted ~25% of the variants; these missing entries were highly enriched in challenging variant types.

We then set out to generate improved variant calls by generating better WGS data, without increasing per-base cost, and then analyzing these data with both existing and new methods. Specifically, we obtained WGS data providing approximately 50-fold coverage of NA12878 by using a PCR-free protocol to reduce coverage bias[14] and generated 250-base paired-end reads. These data come at comparable cost to data providing 50-fold coverage using PCR-amplified, 100-base paired-end reads. With these sequence data, we produced variant call sets using the state-of-the-art program, GATK, and a new method, DISCOVAR, which we developed. The DISCOVAR algorithm (**Online Methods;** Supplemental Note) was specifically designed to address challenging variant types; it involves initial alignment of reads to genomic regions followed by careful local assembly.

We show that both GATK and DISCOVAR provide excellent coverage of ordinary variants, but that DISCOVAR provides substantially better coverage of 'challenging' variants.

RESULTS

Assessing the completeness of variant calling methods

Assessing the accuracy and precision of variant calling methods is hampered by the lack of a true reference variant set and by potential biases in the reference sets used. The traditional approach is to generate sequence reads, align them to a reference sequence, apply the method in question to create a candidate set of variants, and then compare these variants to a reference set. Since reference and candidate sets are typically produced by similar methods, systematic biases could lead to an overestimate of the completeness of these datasets. Indeed, we expect variants from certain regions to be underrepresented – including regions where unique alignment of reads is difficult or impossible such as tandem repeats, segmental duplications, complex rearrangements and sequences missing from the reference genome; regions where sequencing errors occur at high frequency, such as in simple sequence repeats; and regions where sequencing bias occurs, such as sequences at the extremes of GC-content.

To facilitate our investigation of this phenomenon, we divide variants into two categories: *challenging* and *ordinary*. Challenging variants are those (i) in low-complexity sequence (4.3% of the genome, as defined by symmetric DUST[15]); (ii) in segmental duplications (5.5% of the genome, defined[16] based on sequences of length > 1 kb that agree with another sequence at > 90% identity); (iii) in extremely high GC regions (0.1% of the genome, defined[17] by bases occurring in 200-base windows whose middle 100 bases have %GC ≥ 85); and (iv) consisting of long insertions of > 100 bases. Challenging variants are essentially those that occur in specific regions covering ~10% of the genome (apart from the fourth category, which consists of rare events that can occur anywhere). Ordinary variants will refer to all other variants (although some may be difficult to call despite not meeting the above criteria).

Close inspection suggests that challenging variants are significantly underrepresented in at least some current variant call sets. Specifically, we examined Illumina's Platinum variant call set based on 100-base reads from a PCR-free library (which decreases coverage biases arising from PCR amplification), providing ~52x coverage of the genome (measured here as all purity-filtered bases divided by the genome size). This dataset was analyzed by Illumina using an early version of the program GATK[3]. SNPs in this 'Platinum-100' occur at a 2.2-fold lower rate in segmental duplications and a 1.5-fold lower rate in high GC regions compared to the genome as a whole. These deficits seem likely to reflect lower sensitivity for variant calling in these regions, rather than a true biological deficit of polymorphism.

To address the need for an accurate reference, we utilized a Fosmid library from NA12878 and developed finished quality reference sequences for 103 randomly-selected regions comprising 0.1% of the genome. These were generated by sequencing with both Illumina and Pacific Biosciences technologies, and assembling these data using independent methods. To assess how well this worked, we validated portions of the reference sequences using Sanger sequencing. Overall, the data are consistent with an error rate in the Fosmid reference sequences of less than 10^{-5} per base, similar to a finished quality reference. Details are provided in the Online Methods section.

Characterization of variant calls in the Fosmid reference

We found 4486 variants in the 3.8 Mb of Fosmid reference sequence, corresponding to a rate of 1.2 variants per kb (Table 1). Substitutions comprised the vast majority of variants: 77% overall, but 93% outside of low-complexity sequence and 31% in low-complexity sequence. Nearly all of the remaining variants were indels, with a nearly equal number of insertions and deletions (248, 240) as expected. Roughly half of the indels had length 1, with the number falling progressively with size; only 13 had size > 100 bp. There were only two inversion events (25 bp and 4 bp, respectively), which we manually identified.

The variants were not uniformly distributed. They showed striking clustering with 24.9% of the variants in low-complexity sequence that accounts for only 4.3% of the Fosmid reference bases (and the genome). This 5.7-fold enrichment is strongest in bases near A/T homopolymers, which comprise ~10% of the low-complexity sequences in the Fosmids but harbor ~30% of the variants in low-complexity sequence.

Variants tended to occur near other variants. Defining a *cluster* to be a set of at least three variants with adjacent pairs separated by ≤50 bp, we found 115 clusters encompassing 510 variants in total (11.4% of the Fosmid variants). Of clustered variants, 52.0% were in low-complexity sequence (2.1-fold enrichment = 52.0/24.9). The largest two clusters had 49 and 21 variants, respectively, comprised of complex arrangements of indels and substitutions in low-complexity sequence (whose exact representation would change if alignment parameters were perturbed). There are five clusters having 8 or more variants, all in low-complexity sequence. Long low-complexity regions might mutate rapidly, yielding complex relationships between any two given samples. For this reason, clusters of variants in low complexity sequence are unsurprising. The low sequence complexity of these regions would correlate with *in vitro* polymerase replication errors, making these regions very difficult to call correctly.

Among substitutions overall, the ratio of transitions to transversions was 2.1:1, as expected[3,18]. However, the transition/transversion ratio was much lower (54:60 $\approx$ 1:1) for substitution variants within 10 bases of an indel variant. The likely explanation is that sequences containing an indel and substitution in close proximity can often be aligned to a reference sequence in multiple plausible ways, and thus depend on the choice of alignment algorithm (Supplementary Fig. 1).

Evaluation of Platinum-100 variants calls

With the Fosmid reference variant calls in hand, we set out to evaluate the completeness of the Platinum-100 calls (Tables 1,2). We took care to allow for the possibility of complex variants represented in different ways (Supplementary Tables 1a, b).

Overall, the Platinum-100 variant calls missed 25% of the Fosmid variant calls. The rate was much lower (9%) for ordinary variant calls, but much higher (59%) for challenging variant calls. The missing rates were 54% in low-complexity sequence, 78% in segmental duplications, 100% (5/5) in extremely high GC-content and 100% (9 of 9) for long insertion variants. Poor sensitivity in challenging regions combined with higher variation rate in these regions leads to the high false negative rate in the Platinum-100 set. Overall, three-quarters of the missing variants reside in either low-complexity sequence or segmental duplications (Table 1).

Assessing DISCOVAR variant calls

To assess the variant calls from DISCOVAR, we generated test data from the human cell line GM12878, which has been used as a testbed for human assembly and variant calling (Supplementary Note, Section 2). We constructed a single PCR-free library from 0.5 μg of genomic DNA and sequenced a single two-lane flowcell on the Illumina HiSeq 2500 instrument to obtain 250-base paired-end reads providing ~60x coverage of the genome (measured here as all read bases divided by the genome size, or ~50x coverage if only purity-filtered bases are included). DISCOVAR is designed to work with data of this type. Similar data have also been generated for NA12878 and her parents as part of the 1000 Genomes Project (Supplementary Note, Section 1), and we also use these data in our analyses.

We set out to compare variant call sets (Supplementary Note, Section 7; Supplementary Data Set 1): (1) the Platinum-100 variant call set described above, based on 100-base paired-end reads from a PCR-free library and the version of GATK[3,7] available at the time it was created; (2) a GATK-250 variant call set that we generated using the 250-base reads described above and the current version of GATK; (3) a DISCOVAR-250 variant call set that we generated by using our DISCOVAR algorithm. The GATK program requires manual selection of various parameters, for which we followed the advice of GATK's authors. The DISCOVAR program runs 'out of the box', without requiring the selection of parameters. Below, we also consider results from the Cortex[5] algorithm.

We then sought to understand the properties of the GATK-250 and DISCOVAR-250 call sets, which benefit from both longer read lengths and algorithmic improvements.

Sensitivity of variant calling

We studied the sensitivity of the variant call sets by comparison with the Fosmid reference variants (Tables 1–3). We began by examining *ordinary* variants. Whereas the Platinum-100 set missed 9% of ordinary variants, GATK-250 missed only 3% and DISCOVAR-250 missed only 1.2%. (Notably, most of the ordinary variants missed by the GATK and DISCOVAR programs are actually identified in the course of the analysis but removed by post-processing steps (Supplementary Note, Section 7). In principle, they could be recovered by using looser settings for their post-processing filters, although this would entail a loss of specificity (Supplementary Table 2)). For *challenging* variants, the differences were striking. Whereas Platinum-100 missed 59% of the challenging variants, GATK-250 missed 33% and DISCOVAR-250 missed 17%. Considering both ordinary and challenging variants (which occur in a ratio of ~2:1), Platinum-100 missed 25%, GATK-250 missed 12% and DISCOVAR-250 missed 6%.

We also tested the completeness of each of the variant call sets relative to the HapMap3 variant collections[11], Supplementary Table 3. The HapMap3 set was obtained by mapping random reads to the genome, rather than by comparison to finished sequence, and it consists primarily of ordinary SNPs; it is a large collection with ~1.5M variants for NA12878. We considered all bi-allelic SNPs reported and genotyped as present in NA12878 in HapMap3 release 27. We then classified each SNP as present or missing in a given variant call set, regardless of whether it was called homozygous or heterozygous. The proportion of missing variants for Platinum-100 was 2.5%, for GATK-250 was 1.8% and for DISCOVAR-250 was 1.2%. These proportions are likely to be slight overestimates: we estimate that roughly 0.6% of HapMap3 are false positives, because they were absent from all three variant call sets as well as from other available sets (including the trio-based call set in the GATK resource bundle 2.3) and, on close inspection, there appeared to be no evidence supporting them. The adjusted proportions (1.9%, 1.2% and 0.8%) are consistent with the HapMap3 set being strongly biased toward variants that are easy to identify by mapping reads to the genome.

Specificity of variant calling

We studied specificity by estimating false positive rates separately for homozygous and heterozygous variants calls (Table 2). We counted a homozygous call as correct if only the given allele is present, and a heterozygous call as correct if both alleles are present. For homozygous calls, we examined variant calls in the regions covered by the 103 finished Fosmids. Any homozygous call that was not present in the Fosmid reference call set was deemed to be a false positive. The measured false positive rates for Platinum-100, GATK-250 and DISCOVAR-250 were 0.85%, 0.74% and 1.94%, with standard error estimates 0.26%, 0.72% and 0.40%, respectively. Most false homozygous calls occur in low complexity sequence (Supplementary Table 4). For heterozygous calls, we examined variants on chromosome X in the male sample NA12891 (father of NA12878). Any heterozygous call on chromosome X (outside the pseudoautosomal regions) must also be a false positive. The measured false positive rates for Platinum-100, GATK-250 and DISCOVAR-250 were 0.83%, 1.82% and 1.44%, with standard error estimates 0.07%, 0.45% and 0.23%, respectively. Weighting by the number of heterozygous and homozygous calls in each call set, the overall false positives rates for Platinum-100, GATK-250 and DISCOVAR-250 are 0.84%, 1.39% and 1.63%, with standard error estimates 0.11%, 0.39% and 0.21%, respectively. The difference between GATK-250 and DISCOVAR-250 is not statistically significant (p=0.71).

As a further check, we examined the ts/tv ratio of ‘novel’ SNPs, defined as those not present in dbSNP v129. The typical value reported for novel SNPs ~2.1[3,18]. When we omit SNPs in the immediate vicinity of an indel (for which, as noted above, the ts/tv ratio is low), the ratio was 1.84 and 1.81 for novel SNPs in GATK-250 and DISCOVAR-250, respectively. The value is 1.62 and 1.63 for novel SNPs present in GATK-250 but not DISCOVAR-250, and novel SNPs present in DISCOVAR-250 but not GATK-250, respectively.

Results for Cortex algorithm

Finally, we also tested the Cortex[5] variant calling algorithm, which has been used as part of the 1000 Genomes Project[1]. Unlike the other programs tested here, Cortex uses a purely *de novo* approach. This approach has the potential advantage of being unbiased by the reference sequence, but is fundamentally challenging to implement because the lack of a reference sequence makes it hard to sort out repeat sequences. We found that Cortex had low sensitivity, with a false-negative rate of 39.3%, and false-positive rates of 3.46% for homozygous variants and 0.33% for heterozygous variants. We note that the developers of Cortex plan to incorporate read-pairing information in a new version, which could substantially improve its performance.

Representation of variation in disease genes

We wondered if improvements to assembly and variant calling might enable better detection of disease-causing mutations. While this question cannot be definitively answered using only DNA from healthy individuals, as we have here, we reasoned that increased sensitivity at disease-associated loci could be indicative of power to find disease-causing mutations. To that end we examined two classes of examples. The first is in low-complexity sequence and the second is in segmental duplications.

The first class consists of variable-length trinucleotide repeats associated with disease, and occurring roughly equally often in coding and noncoding sequence. We examined the 14 disease examples shown in ref. [19] that did not involve 100% GC triples (generally recalcitrant to Illumina sequencing), Supplementary Table 5. Briefly, by closely examining the data we determined the true genotype of NA12878 at each locus. Then we checked the Platinum-100, GATK-250 and DISCOVAR-250 call sets to determine if they correctly represented this genotype. Platinum-100 had errors in 11 cases, whereas GATK-250 had errors in 3 cases, and DISCOVAR-250 had no errors.

The second class consists of large genic tandem duplications, illustrated using the 10 kb × 3 copy duplicated region in nebulin (NEB). This gene codes for a very large actin-binding protein, 'a major player in muscle health and disease'[20]. Supplementary Fig. 2 exhibits the DISCOVAR assembly graph for the region containing the duplication proper, roughly chr2:152,435,700–152,465,200 (for assembly including flanking regions for context). The graph is exhibited as a large loop that is traversed in total six times (probably three times each by two chromosomes, although other allelic copy number combinations such as 2+4 cannot be definitively excluded using the data we have).

First we examined variant calls for the region. The number of calls reported for Platinum-100, GATK-250 and DISCOVAR-250 are 0, 13 and 55, respectively, suggesting a marked increase in sensitivity. This increased sensitivity might be explained in the following way. Differences between copies of a duplicated region will be captured by edges in the DISCOVAR assembly graph, and these edges may also contain differences with the reference sequence for the given copy. Thus under such conditions the edge may align to a unique locus on the reference sequence, and further define a variant at that locus. The increased sensitivity at this and other segmental duplication loci has nothing to do with segmental duplications *per se*, but is rather a general consequence of how the assembly graph is built and how variant calls are inferred from it.

Next, direct examination of the DISCOVAR assembly graph suggested that many more variants are readily inferrable from the graph but *not* called as variants by DISCOVAR. For example, any edge that topologically must be part of all loop iterations defines homozygous variants on all three repeat copies. For example edge 6 in Supplementary Fig. 2 has 6 differences with the third repeat copy, indicating 6 homozgyous variants there – which are however not in the variant call set, thus representing a limitation of the DISCOVAR variant calling algorithm.

In other cases there are variants that cannot be localized to a repeat copy. For example edge 4 implies a substitution with each of the three repeat copies, but we do not know which copy. In fact all the assembly tells us (via two edges 4, 5 in a 'bubble') is that both the reference and substituted bases are present in the sample genome. It is plausible that such knowledge (without complete localization) would be enough to infer phenotype, thus medically useful. This particular variant is not called by DISCOVAR, and indeed to do so, adaptation of the VCF format would be needed to allow for cases where a variant is known to occur in one or more of several locations.

DISCUSSION

As decreasing sequencing costs make it possible to undertake large projects to study human disease, there is an increasing need to identify variation among individuals in an accurate and complete manner. It is generally recognized that important classes of variants and parts of the genome are underrepresented in current variant call sets, and that improvements are needed.

Because rigorous assessment of the accuracy and completeness of variant calls requires comparison to finished sequence, we created a complete reference set covering ~3.8 Mb of DNA from the widely studied cell line GM12878. Although the set covers only a fraction of the genome, the regions were selected in a random manner and provide sufficient data to characterize problematic areas of the genome and to assess the performance of variant calling methods.

We found that most challenging variants lie in the ~10% of the genome consisting of low-complexity sequence and segmental duplications. Low-complexity sequence behaves differently than the rest of the genome: it is about six-fold enriched for variants overall, with an additional three-fold enrichment near A/T homopolymers. Variants in low-complexity sequence also tend to be clustered, likely arising from complex evolutionary histories. As a result of the enrichment for variants, the 10% of the genome consisting of low-complexity sequence and segmental duplications harbor ~30% of the variants in the genome. These challenging regions of the human genome have long been known to be associated with mutations underlying human disease[12,13].

Using the reference set, we assessed the performance of the Platinum-100 variant call set. This set was created based on 100-base Illumina reads from a PCR-free library of NA12878, and analyzed, by Illumina, with an early version of the GATK program. We found that it missed only 9% of ordinary (non-challenging) variants but 59% of challenging variants, resulting in 25% of variants being missed overall.

We next turned to improved data and cutting-edge algorithms. We generated a new sequence dataset consisting of 250-base Illumina reads from a PCR-free library of NA12878, at ~60x coverage (~50x PF), using the recently released HiSeq 2500 instrument;

the per-base cost of these longer reads is about the same as for the shorter reads typically generated now from PCR-amplified libraries. Cost per Q30 base is in the range of 10–35% higher. (Supplementary Note, Section 8.)

We analyzed (i) the current version of GATK, a widely-used program that is a workhorse for human genome analysis and (ii) DISCOVAR, a new program reported here that was designed with an eye toward calling challenging variants. In the case of GATK, we consulted with the program's authors to choose multiple parameters that must be set. In the case of DISCOVAR, the program runs 'out of the box' with no parameters to be set.

The DISCOVAR-250 and GATK-250 variant calls had much higher sensitivity than Platinum-100 set. For ordinary variants, the false negative rates were fairly similar (1.2% for DISCOVAR and 3% for GATK vs 9% for Platinum-100). As expected, the sensitivity for challenging variants was much lower. Notably, DISCOVAR had a much lower false negative rate for challenging variants (17% for DISCOVAR vs. 33% for GATK). Across all types of variants, DISCOVAR provides the lowest false negative rate (6% for DISCOVAR vs. 12% for GATK and 25% for Platinum-100).

The greater sensitivity of DISCOVAR is associated with a modest decrease in specificity, adding about 0.8% false positives, as compared to Platinum-100. This suggests that an ideal program would provide a transparent and effective method for trading off sensitivity for specificity, by binning calls and providing an accurate false positive prediction for each bin. This appears to be a hard problem, and has not been accomplished by any of the tested programs.

DISCOVAR was designed to work well for the data type that we generated (~60x total coverage by 250-base paired reads from a PCR-free library, from fragments of size ~450 bp; see Supplementary Fig. 3). Because of the way it closes read pairs, DISCOVAR would not work with 100-base reads from fragments of the same size, and reducing coverage by a factor of two or more would also be outside the specification of the current version. Furthermore, the error correction relies, to some degree, on the PCR-free nature of the data. The use of high quality-score discrepancies to differentiate reads from different loci precludes the inclusion of data that may contain errors generated during PCR-based library preparation.

DISCOVAR provides a more complete inventory of an individual's genetic variants than had been previously possible. As such, it adds to the tools that can be used to probe the genetic basis of disease. It may be particularly useful in cases where targeted or exome sequencing fails to find causal mutations. As data costs drop and the ability to interpret variants improves, we anticipate even wider applicability.

DISCOVAR is freely available for academic use. Although not described here, the program can also be used to call variants on non-human genomes and to assemble small genomes *de novo*.

METHODS

Developing a reference variant call set

To study the completeness of variant calling methods, we sought to generate a complete reference set of variant calls based on finished genomic sequence. We utilized a Fosmid library previously developed from the GM12878 lymphoblastoid cell line, plated an aliquot, randomly selected 106 clones, divided the clones into two pools, and sequenced the pools using Illumina and Pacific Biosciences technologies (see Supplementary Note, Section 3). We then assembled the data to obtain finished quality sequences for 103 of the Fosmids, spanning 3.8 Mb of genomic DNA.

The Fosmids comprise ~0.1% of the human genome, which is simpler to assemble than an entire genome. Nonetheless, we took care to assemble the sequence data with two independent methods. The first method employed the DISCOVAR algorithm (described below). It yielded complete sequence of 103 Fosmids, which we refer to here as the Fosmid reference sequences. (We omitted three Fosmids: one could not be assembled because it consisted of highly repetitive heterochromatin sequence, one could not be assembled due to insufficient coverage, and one consisted of sequence from Epstein-Barr virus, which is used to create lymphoblastoid cell lines.) We aligned the Fosmid reference sequences to the human genome reference sequence (version hg19), excluding three Fosmids whose placements were ambiguous (Supplementary Note, Section 7e). We then parsed the alignment into variant calls, defining a Fosmid reference call set based on 100 Fosmids that provides a reasonably unbiased representation of the genome.

We reconciled these variant calls with those obtained from a second independent approach. The second method used the HGAP program[22] to assemble the Pacific Biosciences data, followed by iCorn[23] to correct the assemblies using the Illumina data. We obtained complete assemblies for 57 Fosmids (with the remainder having insufficient coverage in the Pacific Biosciences data, owing to variation in molarity within the pools). For 32 of the Fosmids, there was perfect agreement with the Fosmid reference sequences obtained from the first method. For the remaining 25 Fosmids, we manually analyzed every difference (Supplementary Tables 7, 8). We found 9 errors in the Fosmid reference sequences, 29 errors in the HGAP+iCorn assemblies, and 2 cases where both were wrong. From these data, we infer an error rate of ~1/200,000 bases in our Fosmid reference sequences. We corrected the detected errors, which had little effect on our analysis (Supplementary Note, Section 3).

We also experimentally assessed the accuracy of the reference sequences for 26 of the Fosmids (Supplementary Note, Section 4). We identified 162 variants that were not present in the Platinum-100 variant set, reasoning that such variants could represent errors in the Fosmid reference sequences. These were grouped into 134 loci. We selected 95 cases that best suited primer design, amplified the sequences from the Fosmid DNA and Sanger sequenced the amplicon. In 93 cases, we obtained a clear PCR product and reads that aligned to the targeted Fosmid. The Fosmid reference sequence was confirmed in 87 of these 93 cases. In 5 cases, we could not adequately resolve the discrepancy. In the remaining case, the Sanger reads revealed a 20-base insertion (5 copies of GAAA) that had been missed in the Fosmid reference sequence.

Overall, the data are consistent with an error rate in the Fosmid reference sequences of less than 10^{-5} per base, consistent with 'finished' quality[24]. Our analyses do not account for mutations in the Fosmids relative to the source DNA, but past analyses for BACs suggest a comparable or lower error rate[25], and the rate for the Fosmid pool protocol used here could be lower because the Fosmids were not passaged or subjected to single-cell cloning.

The Fosmid reference sequences are available in GenBank (Supplementary Table 9; Accession Codes).

DISCOVAR algorithm: sequence assembly

Given the limitations of current variant calling approaches, we set out to design a new algorithm, optimized for an improved data type, that might achieve improved results, especially for challenging variants. This new algorithm, DISCOVAR, works by (i) assigning sequence reads to genomic regions (taking all read pairs for which at least one read had been mapped to the region, and without any filtering) (ii) assembling the genome on a region-by-region basis, and (iii) calling variants within each region. Assembly of regions (rather than the whole genome at once) is easier, however may also result in artifacts because of mapping errors, and cannot assemble long sequences that are novel to the sample.

DISCOVAR differs from most assembly approaches in the following critical way: at every stage it attempts to preserve bases in reads that appear to be present in the sample DNA, including bases distinguishing homologous chromosomes and repeat copies. For early approaches[26,27] based on Sanger sequencing this would not have made sense because coverage levels were too low. Subsequent work[28] axiomatized removal of biological differences during error correction as deliberate 'data corruption', and in fact the removal of differences in error correction or graph simplification continued in other assemblers[9,29]. In our prior work[30], differences were not eliminated in the assembly graph *per se*, however in practice they were often missing from the final assembly, in part because of assembly linearization during the process of scaffolding. In our current approach, differences are preserved from beginning to end.

The assembly process has two main components: error correction and graph simplification (for details, see Supplementary Note, Section 5).

(i) Error correction—DISCOVAR starts by correcting errors in the reads, and as part of this process, closing read pairs (Supplementary Fig. 4). The goal is to remove most laboratory errors in the sequence data, while preserving biological differences, including polymorphisms. Because the laboratory process introduces indel errors at a very low rate, the error correction algorithm targets only substitution errors. (Systematic indel errors, for example in long homopolymers, would be more likely to be uncorrected, and may result in ambiguity 'bubbles' in the assembly graph.) To correct a read, DISCOVAR tries to find the 'true friends' of that read – *i.e.* reads from the same locus on the same chromosome (as described below). In particular, where possible because of a proximate polymorphism, DISCOVAR will exclude reads arising from the homologous chromosome from the friend set. This behavior differs from other error correction algorithms that find friends[31,32].

DISCOVAR uses each read as the 'founder' of a stack of friends, which consists of a gap-free multiple alignment among the founder and its friends (exhibited as a matrix of bases, with one row for each read and one column for each position). To do this, for each founder, DISCOVAR seeds alignments on perfect matches to define the initial friend reads associated to the founder (some of which are not true friends). These friends are arranged in a stack, with each friend (or its reverse complement) positioned at a given offset relative to the founder.

Posed in this fashion, the central problem of error correction is to remove false friends from the stack. This is easy in cases where the friend and the founder both have high quality sequence at a column in the matrix corresponding to a base position where true and false loci differ. We handle this case by scanning the stack to find columns in which the founder and a friend both have quality ≥ 30, but different base calls. Such friends are declared false and removed from the stack.

For each read pair, we form friend stacks for its two reads. The stack for the right-hand read is reverse complemented, so that both stacks are on the same DNA strand. These stacks are truncated on both ends (Supplementary Fig. 4c), so that the stacks do not extend beyond the original DNA fragment. Next we derive the consensus sequences for both stacks. These consensus sequences typically extend beyond the read ends and usually across the gap (if any) between the reads.

We note that the algorithm design relies on the existence of an overlap between the consensus sequences, and thus the original fragments should be shorter than four times the read length. This is consistent with the laboratory design.

DISCOVAR next finds overlaps between the consensus sequences, allowing for the possibility of more than one overlap. For each overlap, we then merge the left and right stacks, yielding a joint stack, and find its consensus. This 'joint consensus' is a single sequence extending from end to end on the original DNA fragment defining the pair, thus 'closing' the pair. We require that the joint consensus for a given overlap is unambiguously determined. For example, for a pair where the reads have a gap between them on their originating DNA fragment, and there is a SNP in the gap, the pair may not be closed. Typically there is exactly one overlap between the left and right consensus sequences, yielding a single unambiguous joint consensus. However it is possible for there to be multiple overlaps, each yielding an unambiguous joint consensus. In such cases (~1 per 1000 pairs in human), the pair is not closed, but instead we report as its 'closure' a modified pair, whose left member is obtained by taking the longest sequence that the consensus sequences share on the left, and whose right member is similarly obtained.

(ii) Graph creation—The pair closures are now formed into a unipath graph[33]. For this we set the minimum overlap K to ~0.18 of the median closure length. This graph would have rare gaps arising from low coverage. Most such gaps can be eliminated by including the hg19 reference sequence (for the region) as an additional input to the graph construction, and therefore we do this.

(iii) Graph improvement—DISCOVAR next simplifies and improves the graph (Supplementary Note, Section 5). There are two goals: to remove branches that are not present in the sample (likely artifacts of the laboratory process), and to 'pull apart' the graph. We illustrate the latter using the operation of *graph reconstruction* (Supplementary Fig. 5). Briefly, each pair closure may be represented as a path through the graph, or 'closure path'. Graph reconstruction takes the closure paths, each expressed as a sequence of assembly edges, and glues the closure paths together along some (but not all) proper overlaps between them, yielding a new assembly graph. If we have two paths "abc" and "cde", then those two paths overlap along c, and if glued together would yield "abcde". Gluing many paths together yields a graph. Supplementary Fig. 5 illustrates the method using a simple but common scenario where two genomic regions come together in an assembly along a repeat r. Depending on the length of r, and the particularities of the data, and which proper overlaps are excluded, the reconstructed assembly may be either the same assembly, or a 'pulled apart' assembly that represents the genome itself (Supplementary Fig. 5a–d). Exclusion of overlaps depends on a detailed analysis of the closure path counts at a particular locale (Supplementary Fig. 5e). Graph reconstruction is akin to the approaches in EULER[28] and RockBand[29], however differs from them in how it thresholds simplification.

DISCOVAR algorithm: Variant calling

Once the graph assembly has been simplified, DISCOVAR creates a list of variants by comparing the assembly to the reference sequence (Supplementary Note, Section 6; Supplementary Fig. 6). The graph assembly can encode variants either as alternative paths in the graph or as unambiguous edges containing base differences with the reference.

In order to identify variants from the assembly graph (Supplementary Fig. 6a), its edges are aligned to the reference (Supplementary Fig. 6b). In many cases this would be enough to successfully call variants, but for some loci it is not possible due to erroneous or missing alignments. The causes of this are varied, and include: alignment artifacts, low complexity or repetitive sequence, insertions or deletions near the ends of edges, or larger insertions in an edge.

To deal with these difficult loci, we next build an acyclic graph whose edges are anchored to the reference. The anchor points, edge order, and edge adjacencies are determined from the edge alignments. Initially only high quality alignments are used to build the graph, eliminating problems caused by erroneous alignments. This excludes some valid edges, which either aligned poorly or not at all. These can be safely recovered by adding back unused paths through the original assembly graph that are consistent with both the anchored acyclic graph and the reference (Supplementary Fig. 6c).

Given this anchored acyclic graph, variants can then be identified by comparing the edges to the portions of the reference to which they are anchored (Supplementary Fig. 6d). Probabilities are then computed by realigning the original reads to the edges of this graph and scoring these alignments using the read quality scores.

DISCOVAR can be run on a human genome in about two days, using a small cluster (see Supplementary Note, Section 7a).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Broad Institute Genomics Platform for generating data for this project and Jim Bochicchio for project management. We thank the 1000 Genomes Project for early access to 250-base reads from the CEPH trio. We thank Illumina for use of sequence data from the 17-member CEPH pedigree, and Mike Eberle for detailed information on these data. We thank Zam Iqbal for detailed advice regarding running Cortex on 250-base reads, and Mauricio Carneiro and Eric Banks for detailed advice regarding running GATK on the same data. We thank Heng Li for early access to bwa-mem and for his extensive examination of our manuscript. We thank Sheila Fisher for help preparing the data cost estimate. We thank Mark Daly, Mark DePristo, Stacey Gabriel, Zam Iqbal, Daniel MacArthur, Dan Neafsey and Mike Zody for helpful comments. This project has been funded in part with Federal funds from the National Human Genome Research Institute, National Institutes of Health, Department of Health and Human Services, under grants R01HG003474 and U54HG003067, and in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C.

References

1. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
2. Albers CA, et al. Dindel: accurate indel calls from short-read data. Genome Res. 2011; 21:961–73. [PubMed: 20980555]
3. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–8. [PubMed: 21478889]
4. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012; 22:1154–62. [PubMed: 22522390]
5. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012; 44:226–32. [PubMed: 22231483]
6. Li B, et al. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. 2012; 8:e1002944. [PubMed: 23055937]
7. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303. [PubMed: 20644199]
8. O'Fallon BD, Wooderchak-Donahue W, Crockett DK. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinformatics. 2013; 29:1361–6. [PubMed: 23620357]
9. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22:549–56. [PubMed: 22156294]
10. Wang Y, Lu J, Yu J, Gibbs RA, Yu F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 2013; 23:833–42. [PubMed: 23296920]
11. International HapMap Consortium et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–8. [PubMed: 20811451]
12. Verkerk AJ, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991; 65:905–14. [PubMed: 1710175]
13. Lupski JR, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell. 1991; 66:219–32. [PubMed: 1677316]
14. Kozarewa I, et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009; 6:291–5. [PubMed: 19287394]
15. Morgulis A, Gertz EM, Schaffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006; 13:1028–40. [PubMed: 16796549]

16. She X, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 2004; 431:927–30. [PubMed: 15496912]
17. Ross MG, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013; 14:R51. [PubMed: 23718773]
18. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011; 27:1157–8. [PubMed: 21320865]
19. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet. 2005; 6:743–55. [PubMed: 16205714]
20. Labeit S, Ottenheijm CA, Granzier H. Nebulin, a major player in muscle health and disease. FASEB J. 2011; 25:822–9. [PubMed: 21115852]
21. Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979; 7:1–26.
22. Chin CS, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10:563–9. [PubMed: 23644548]
23. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010; 26:1704–7. [PubMed: 20562415]
24. Chain PS, et al. Genomics. Genome project standards in a new era of sequencing. Science. 2009; 326:236–7. [PubMed: 19815760]
25. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–45. [PubMed: 15496913]
26. Myers EW, et al. A whole-genome assembly of Drosophila. Science. 2000; 287:2196–204. [PubMed: 10731133]
27. Venter JC, et al. The sequence of the human genome. Science. 2001; 291:1304–51. [PubMed: 11181995]
28. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001; 98:9748–53. [PubMed: 11504945]
29. Zerbino DR, McEwen GK, Margulies EH, Birney E. Pebble and Rock Band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. PLoS One. 2009; 4:e8407. [PubMed: 20027311]
30. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011; 108:1513–8. [PubMed: 21187386]
31. Batzoglou S, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002; 12:177–89. [PubMed: 11779843]
32. Kao WC, Chan AH, Song YS. ECHO: a reference-free short-read error correction algorithm. Genome Res. 2011; 21:1181–92. [PubMed: 21482625]
33. Butler J, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008; 18:810–20. [PubMed: 18340039]

Table 1

Some categories of challenging variants

	Fosmid reference variants										
category	number	% of each category missing from call set					% of total missing variants in each category				
		Platinum-100	GATK-250	CORT EX-250	DISCO VAR-250	union	Platinum-100	GATK-250	CORT EX-250	DISCO VAR-250	union
low complexity	1115	53.6	32.1	63.9	16.2	11.6	53.5	64.5	40.4	66.5	64.8
near A/T homopolymers	360	68.3	45.0	74.4	17.5	13.9	22.0	29.2	15.2	23.2	25.1
segmental duplication	351	78.1	37.3	90.3	17.7	16.5	24.5	20.4	18.0	22.8	29.1
long insertion	9	100.0	88.9	100.0	77.8	77.8	0.8	1.4	0.5	2.6	3.5
extremely high GC	5	100.0	60.0	80.0	60.0	60.0	0.4	0.5	0.2	1.1	1.5
challenging	1402	58.8	33.0	69.1	16.8	12.9	73.6	83.2	55.0	86.4	91.0
ordinary	3084	9.5	3.0	25.7	1.2	0.6	26.4	16.8	45.0	13.6	09.0

Several categories of challenging variants are described and compared to the Fosmid reference set. The “number” shows the raw count of events in regions defined by the Fosmid reference set, while subsequent columns show the false negative rate in each category of variant (“% of each category missing from call set”), as well as the breakdown by category of all false negatives (“% of total missed variants in each category”). Variants are categorized by type and/or region of the genome as follows: Low complexity: bases identified as having low complexity by symmetric DUST with default parameters[15]. Near A/T homopolymers: bases lying within 5 bases of a run of 15 or more identical A or T bases. Of 360 such events, all but 8 were labeled low complexity by DUST. Segmental duplication: a segmental duplication as defined in the Segmental Duplication DB[16] (see URL section). Long insertion: an insertion of > 100 bases Extremely high GC: defined[17] by bases occurring in 200-base windows whose middle 100 bases have %GC 85. Challenging: union of low complexity, segmental duplication, long insertion and extremely high GC categories. Ordinary: the complement of challenging. This table has been adjusted to reflect the manual corrections of Supplementary Tables 1a, b.

Table 2

Estimated sensitivity and specificity of variant call sets

call set	read length	%FN	#heterozygous/#homozygous	%FP		
				heterozygous variants	homozygous variants	all variants
Platinum	100	25.0 ± 2.5	1.49	0.83 ± 0.07	0.85 ± 0.26	0.84 ± 0.11
GATK	250	12.3 ± 1.8	1.54	1.82 ± 0.45	0.74 ± 0.72	1.39 ± 0.39
Cortex	250	39.3 ± 2.6	1.39	0.33 ± 0.18	3.46 ± 0.61	1.64 ± 0.28
DISCOVAR	250	6.0 ± 1.2	1.57	1.44 ± 0.23	1.94 ± 0.40	1.63 ± 0.21

For each of four variant call sets, we estimated the percent of false negatives (%FN) and false positives (%FP). False negative rates were estimated using 100 randomly selected Fosmids, as described in the text. False positive rates for homozygous variants were estimated by computing the fraction of homozygous calls that were not in the Fosmid reference set. In both cases standard errors were obtained by bootstrapping, using 1000 bootstrap samples from the set of 100 Fosmids[21]. False positive rates for heterozygous variants were estimated by dividing the number of heterozygous events observed in the 100 Mb region of X from 10–110 Mb on NA12878's father by the number observed on the same region for NA12878, then dividing by 1.8 = (heterozygous calls per Mb of genome)/(heterozygous calls per Mb of X), from the Platinum-100 call set for NA12878, thus correcting for the difference between X and the genome. Standard errors were obtained using 1000 bootstrap samples from the set of 100 regions of size 1 Mb obtained by segmenting the 100 Mb region. Intermediate calculations for false negatives and positives are shown in Supplementary Table 6. #heterozygous/#homozygous: genome-wide ratio of number of heterozygous calls divided by number of homozygous calls. %FP for all variants: the average of %FP for heterozygous and homozygous variants, weighted by #heterozygous/#homozygous. The FN and FP (homozygous) values in the this table are corrected values, after taking account of manual corrections from Supplementary Table 1b. Data for the Platinum-100 call set had 48x PF Q30 coverage, while data for the 250-base analyses had 40x raw Q30 coverage, and 39x PF Q30 coverage.

Table 3

Classification of Fosmid variants

Called by				Total	Substitutions	Insertions by size in bp				Deletions by size in bp			
D	G1	G2	C			1	2–10	11–100	>100	1	2–10	11–100	>100
•	•	•	•	2542	2151	117	67	6	0	109	87	5	0
				200	109	9	21	16	7	17	14	4	3
•				253	145	25	22	14	1	32	10	4	0
	•			9	6	0	0	0	0	1	2	0	0
		•		39	25	1	6	2	0	2	1	2	0
			•	2	0	0	0	0	0	2	0	0	0
•	•			36	20	5	1	0	0	7	3	0	0
•		•		489	373	21	42	12	1	7	17	16	0
•			•	14	6	0	1	0	0	4	0	2	1
	•	•		10	5	0	1	0	0	3	1	0	0
	•		•	3	1	0	0	0	0	2	0	0	0
		•	•	3	0	0	0	0	0	2	1	0	0
•	•	•		725	542	52	37	1	0	41	43	9	0
•	•		•	38	26	6	0	0	0	5	1	0	0
•		•	•	118	65	12	10	9	0	4	10	8	0
	•	•	•	5	2	0	1	0	0	2	0	0	0
Total				4486	3476	248	209	60	9	240	190	50	4

We classify all the variants that are in the Fosmid truth call set, and which are obtained from the Fosmid reference sequences (thus representing single haplotypes). For each of four call sets D = DISCOVAR-250, G1 = Platinum-100, G2 = GATK-250 and C = Cortex, and each of 2^4 = 16 possible call combinations, variants are classified as substitutions, insertions and deletions, and by size. This table has been modified to reflect the manual corrections in Supplementary Table 1a.