nature genetics

De novo assembly and genotyping of variants using colored de Bruijn graphs

Zamin Iqbal[1,2,5], Mario Caccamo[3,5], Isaac Turner[1], Paul Flicek[2] & Gil McVean[1,4]

Detecting genetic variants that are highly divergent from a reference sequence remains a major challenge in genome sequencing. We introduce *de novo* assembly algorithms using colored de Bruijn graphs for detecting and genotyping simple and complex genetic variants in an individual or population. We provide an efficient software implementation, Cortex, the first *de novo* assembler capable of assembling multiple eukaryotic genomes simultaneously. Four applications of Cortex are presented. First, we detect and validate both simple and complex structural variations in a high-coverage human genome. Second, we identify more than 3 Mb of sequence absent from the human reference genome, in pooled low-coverage population sequence data from the 1000 Genomes Project. Third, we show how population information from ten chimpanzees enables accurate variant calls without a reference sequence. Last, we estimate classical human leukocyte antigen (HLA) genotypes at *HLA-B*, the most variable gene in the human genome.

Characterization of genetic variants present in an individual, population or ecological sample has been transformed by the development of high-throughput sequencing (HTS) technologies. The standard approach to variant discovery and genotyping from HTS data is to map reads to a reference genome[1–5], thereby identifying positions where the sample contains simple variant sequences. This approach has proven powerful in the study of SNPs[6], short insertion-deletion (indel) polymorphisms[3,5,7,8] and larger structural variation[9–14] in well-characterized genomes, such as the human genome[15–17].

However, the mapping approach has limitations. First, the sample may contain sequence that is absent or divergent from the reference, for example, through horizontal transfer events in microbial genomes[18,19] or at highly diverse loci, such as the classical HLA genes[20]. In such cases, short reads either cannot or are unlikely to map correctly to the reference. Second, reference sequences, particularly of higher eukaryotes, are incomplete, notably in telomeric and pericentromeric regions. Reads from missing regions will often map, sometimes with apparently high certainty, to paralogous regions, potentially leading to false variant calls. Third, samples under study may either have no available reference sequence or it may not be possible to define a single suitable reference, as in ecological sequencing[21]. Fourth, methods for variant calling from mapped reads typically focus on a single variant type. However, in cases in which variants of different types cluster, focus on a single type can lead to errors, for example, through incorrect alignment around indel polymorphisms[6,7]. Fifth, although there are methods for detecting large structural variants, such as using array comparative genomic hybridization (aCGH)[22–25] and mapped reads[11,12,14,26], these cannot determine the exact location, size or allelic sequence of variants. Finally, mapping approaches typically ignore prior information about genetic variation within the species.

Several of these limitations can potentially be solved through *de novo* assembly, which is agnostic with regard to variant type and divergence from any reference. However, although there are established algorithms for *de novo* assembly from HTS shotgun data, which are based on overlap[27–29] or de Bruijn graphs[30–32], current approaches have limitations. Notably, they focus on consensus assembly, treating the sequence as if it is derived from a monomorphic sample (for example, a haploid genome, inbred line or clonal population). Consequently, variation is ignored (processed in the same way as sequencing artifacts) and can lead to assembly errors. Some variation-aware *de novo* assembly algorithms have been developed[31,33–36], but these do not represent a general solution to sequencing experiments in which genetic variation is either the primary concern or unavoidable (for example, in outbred diploid samples, pooled data or ecological samples).

Current assembly methods also typically ignore pre-existing information, such as a reference sequence or known variants. Although variant discovery should not be biased by such information, neither should this information be discarded. For example, in a single outbred diploid sample, it is hard to distinguish paralogous from orthologous variation. However, if variation is also observed in the reference haploid genome, it is likely to be driven by paralogy. Furthermore, current implementations of *de novo* assembly algorithms for HTS data have substantial computational requirements, which make them impractical for large-scale studies on eukaryote genomes.

Here we introduce *de novo* assembly algorithms focused on detecting and characterizing genetic variation in one or more individuals.

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [3]The Genome Analysis Centre, Norwich Research Park, Norwich, UK. [4]Department of Statistics, University of Oxford, Oxford, UK. [5]These authors contributed equally to this work. Correspondence should be addressed to G.M. (mcvean@well.ox.ac.uk).

Received 8 April 2011; accepted 7 November 2011; published online 8 January 2012; doi:10.1038/ng.1028

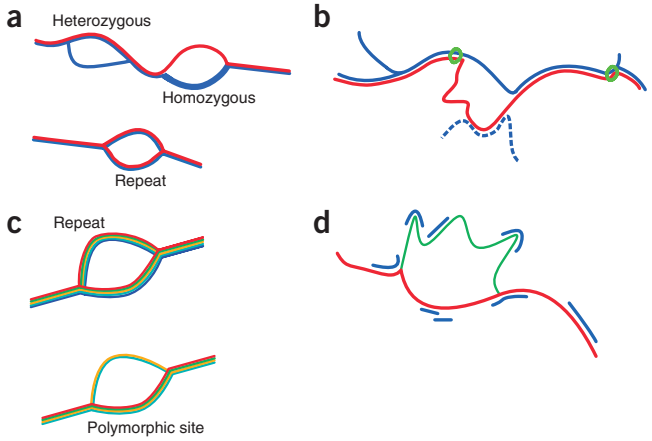

Figure 1 Schematic representation of four methods of variation analysis using colored de Bruijn graphs; line width represents coverage. (**a**) Discovery of variants in a single outbred diploid individual (blue) with a reference sequence (red). True polymorphisms generate bubbles that diverge from the reference, whereas repeat structures lead to bubbles that are also observed in the reference. (**b**) Even when the reference allele (red) does not form a clean bubble, we can identify homozygous variant sites by tracking the divergence of the reference path from that of the sample. On finding a breakpoint, we take the longest contig in the sample (that is, the path as far as the next junction) and ask whether the reference path returns before this point (green circles show the anchoring sequence). The algorithm (path divergence) is not affected by repeat sequences in the reference allele present elsewhere in the genome of the sample (blue dashed line). (**c**) When many samples (each in a different color) are combined, it is possible to distinguish repeat-induced bubbles (in which both sides of the bubble are present in all samples) from true variant sites (in which bubble coverage varies with genotypes and genotypes are in Hardy-Weinberg equilibrium). (**d**) The likelihood of any given genotype can be calculated from the coverage (blue) of each allele (green, red), accounting for contributions from other parts of the genome. In this example, the sample is heterozygous and therefore has coverage of both alleles, although not sufficient to enable full assembly.

These algorithms extend classical de Bruijn graphs[37,38] by coloring the nodes and edges in the graph by the samples in which they are observed. This approach accommodates information from multiple samples, including one or more reference sequences and known variants. We show how the method can detect variation in species without a reference, combine information across multiple individuals to improve accuracy, and genotype known variants. Cortex has already contributed to public datasets as part of the 1000 Genomes Project[17].

RESULTS

Colored de Bruijn graph algorithms

De Bruijn graphs, which represent overlap information in a set of DNA sequences, are widely used in genome assembly and underlie many popular algorithms, including AllPaths-LG[31], SOAPdenovo[32], Abyss[39] and Velvet[30,40]. The graph consists of a set of nodes that represent words of length k (k-mers). Directed edges join k-mers seen consecutively in the input. Variation between genomes generates new nodes and edges. In the simplest case, polymorphisms appear as bubbles within unique contigs. However, more complex structures also arise, for example, where a variant generates a k-mer found in a paralogous location (**Supplementary Fig. 1**).

A colored de Bruijn graph generalizes the original formulation to multiple samples embedded in a union graph, where the identity of each sample is retained by coloring those nodes present in a sample. The samples may reflect HTS data from multiple samples, experiments, reference sequences, known variant sequences or any combination of these. Below, we outline four algorithms for variant discovery and genotyping that make use of the colored de Bruijn graph structure (see **Supplementary Note**).

Bubble calling

The simplest use of colored de Bruijn graphs is to identify variant bubbles in a single diploid individual. This approach may have a high false positive rate because of the difficulty of separating repeat- and variant-induced bubbles. However, inclusion of a haploid reference genome improves the reliability of the algorithm, as most repeat structures will be present in the reference, and any bubble in the reference color must be a repeat (**Fig. 1a**). A reference genome also aids the detection of variants, because only the variant allele contig need be assembled, and is essential for detecting homozygous variant sites. All types of variant can induce bubbles; hence, the process of bubble detection is influenced by variant type only through the graph complexity of the variants and the probability of assembling both sides of the bubble given the sequence coverage, k-mer size and error rate (see **Supplementary Note** for details of a model for predicting power). Haplotypes arising from variants within k bases of each other will naturally be assembled as single compound variants. Our implementation is referred to as the 'bubble-calling' (BC) algorithm (see **Supplementary Note**).

Path divergence

The bubble-calling algorithm relies on the detection of clean bubbles. However, for complex variants (for example, novel sequence insertions, large deletions and inversions), the path of at least one allele is unlikely to generate a clean contig. Nevertheless, in some cases, particularly for deletions, path complexity is restricted to the reference allele. Such cases can be identified by following the (known) reference path through the joint graph and detecting where it diverges from the sample graph (**Fig. 1b**, **Supplementary Fig. 2** and **Supplementary Note**). This 'path-divergence' (PD) algorithm typically identifies only homozygous variants and is biased toward identifying deletions. Nevertheless, the algorithm can substantially increase power to detect some variant types and potentially identifies events of arbitrary size, irrespective of read length.

Multiple sample analysis

The joint analysis of HTS data from multiple samples can substantially improve the power and false discovery rate (FDR) of variant detection. In the simplest case, samples are combined in a single color (for example, in a pooling experiment), and the data can be analyzed as above. However, by maintaining separate colors for each sample, there is additional information about whether a bubble is likely to have been induced by repeats (where many or all samples show coverage of both paths in the bubble; **Fig. 1c**) or errors (where the error-carrying side of the bubble will typically have low coverage). This observation leads to an approach to variant discovery even in species for which there is no suitable reference. We have devised statistical methods that enable probabilistic classification of bubble structures into those arising from errors, repeat structures or variants (see **Supplementary Note**). When a reference genome is available, this approach can still help distinguish true variants from errors and repeat structures that are absent from the reference.

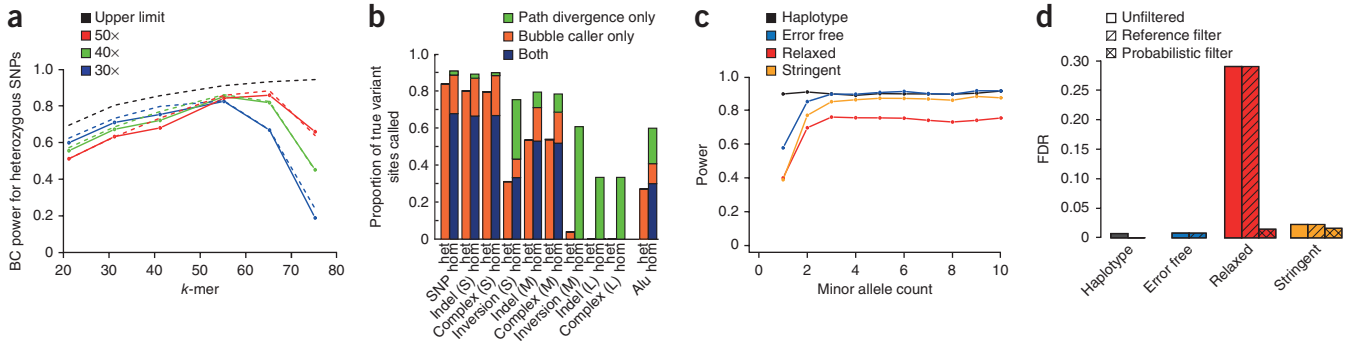

Figure 2 Simulation-based evaluation of Cortex. (**a**) Power of the BC algorithm to detect heterozygous SNPs in a single individual as a function of *k*-mer size. Genome repeat content dictates an upper limit to power (black dashed line), whereas finite sequence coverage reduces power for large *k*-mer size (solid lines, circles) in a predictable manner (dashed lines). Increased coverage reduces power at lower *k*-mer sizes owing to recurrent errors that evade error cleaning. At high *k* (for example, $k = 55$, 50×), power is close to the upper limit. (**b**) Power to detect different variant types in homozygous and heterozygous states using the BC and PD algorithms with 30× coverage (100-bp reads, $k = 55$). For small (S) variants, power is 80–85%. For medium-sized (M) variants, power remains high at homozygous sites but is ≤50% at heterozygous sites. For large (L) events, there is power only for homozygous sites, and this is only achieved with the PD algorithm. (**c**) Power to detect SNP variants using BC in population data (ten individuals, 10× coverage, 100-bp reads, $k = 55$). Fluctuations in coverage reduce power for low-frequency variants. More stringent cleaning increases power because bubbles are less confounded by errors. (**d**) FDR for call sets in **c** before and after classifying bubbles as error-, repeat- or variant-induced. The probabilistic filter reduces FDR under relaxed cleaning from 29% to 1.5% with 1.7% loss in power, and from 2.3% to 1.6% with 1% loss in power, for the stringent cleaning.

Genotyping

Colored de Bruijn graphs can be used to genotype samples at known loci, even when coverage is insufficient to enable variant assembly (**Fig. 1d**). We construct a colored de Bruijn graph of the reference sequence, known allelic variants (which may include those discovered using the above methods) and data from the sample. The likelihood of each possible genotype is calculated, accounting for the graph structure of both the local and genome-wide sequence (see **Supplementary Note**). This approach generalizes to multiple allelic types and, because the algorithm does not require variants to form simple bubble structures, it is possible to genotype complex and compound variants such as those at classical HLA loci.

Graph building and cleaning

We have developed Cortex, a memory-efficient assembler for building and representing colored de Bruijn graphs and for performing variant calling and genotyping from HTS data (see URLs and **Supplementary Note**). The implementation uses an efficient hash table that implicitly encodes the graph; memory use is specified in advance according to a simple formula, and many standard operations have linear or better algorithmic complexity (see **Supplementary Note**). Cortex uses previously undescribed cleaning methods to increase sensitivity (**Supplementary Fig. 3** and **Supplementary Note**). Furthermore, Cortex is the only assembler able to handle multiple eukaryotes simultaneously, for example, 1,000 *Saccharomyces cerevisiae* samples in less than 64 GB of random-access memory (RAM) or 10 humans in less than 256 GB of RAM.

Simulation 1: a single high-coverage diploid genome

We simulated high-coverage (10–50×) sequencing data from a diploid human sample that carries SNPs, indels and structural variants (see **Supplementary Note**). We analyzed data using both the BC and PD methods.

For a variant to be identified successfully, the bubble must both be assembled without gaps and be identifiable in the wider graph. Genome complexity, sequencing depth, read length, *k*-mer size and error rate interact to influence both factors. As *k*-mer size increases, the fraction of SNP sites with unconfounded bubbles ranges from 51% with $k = 21$ to 85% with $k = 75$ in humans (**Supplementary Fig. 4**). Increasing *k*-mer size reduces the risk of error-induced contigs confounding the graph, but it also increases the probability of a *k*-mer containing an error. Furthermore, for a fixed per-base depth and read length, as *k*-mer size increases, the effective depth of each *k*-mer decreases, leading to an increased probability of gaps in the assembly (**Fig. 2a**; see **Supplementary Note**). Consequently, the *k*-mer size that maximized the sensitivity of the BC algorithm in the simulation varied with coverage and read length; being approximately 55 for 30×, 55 for 40× and 65 for 50× with 100-bp reads. The loss in power relative to the theoretical maximum was, however, small. For example, with 50× coverage ($k = 65$ 100-bp reads), we identified 86% of heterozygous SNPs compared to the maximum possibility of 92%. Simulation-based estimates of power closely tracked predictions (**Fig. 2a** and **Supplementary Figs. 5** and **6**).

To assess the power of Cortex to detect a range of variants of different types and sizes, we applied the BC and PD algorithms at a single point (30×, 100-bp reads, $k = 55$; see **Supplementary Note**). For isolated SNPs, short indels (1–100 bp) and small complex combinations of SNPs and indels (1–100 bp), we had 80% power to detect heterozygous sites and 90% power to detect homozygous variant sites (**Fig. 2b**). For moderately-sized (100–1,000 bp) indels and complex variants, power was 50% and 75–80% for heterozygous and homozygous sites, respectively. For large variants (1–50 kb), we had power to detect only homozygous variant sites (~35%), entirely through PD. These sensitivity estimates were attained with an FDR of 2%.

Simulation 2: population-based variant calling

We simulated sequence data from ten diploid individuals based on human chromosome 22 (100-bp reads, 10× per individual; see **Supplementary Note**). We analyzed data at the level of individual haplotypes, error-free reads and error-containing reads ($k = 55$) under two cleaning thresholds (relaxed and stringent), and compared two filtering approaches. First, we removed bubbles present in the reference. Second, we used a probabilistic model to classify bubbles as arising through errors, repeats or true variants (see **Supplementary Note**).

Table 1 Comparison of 1000 Genomes and Cortex calls to fosmid data

Variant type	1000 Genomes[a]	Cortex		Bubble caller		Path divergence	
		All	High confidence[b]	All	High confidence	All	High confidence
SNP (Hom.)	1,085 (0)	1,071 (4.0)	605 (0.5)	1,057 (3.9)	591 (0.5)	340 (8.5)	144 (1.4)
SNP (Het.)	2,350 (28)	1,155 (32)	1,029 (32)	1,155 (32)	1,029 (32)	0 (–)	0 (–)
Indel (Hom.)	64 (0)	96 (6.3)	20 (0.0)	79 (6.3)	16 (0.0)	37 (5.4)	5 (0.0)
Indel (Het.)	127 (29)	67 (40)	43 (30)	67 (40)	43 (30)	0 (–)	0 (–)
Complex (Hom.)	–	258 (1.9)	202 (1.5)	112 (2.7)	77 (1.3)	174 (1.7)	139 (2.2)
Complex (Het.)	–	161 (26)	137 (25)	161 (26)	137 (25)	0 (–)	0 (–)

Het., heterozygous; Hom., homozygous.

[a]Values reported are the number of each variant per genotype combination called, and those in parentheses are the percentage of cases in which only the reference allele was observed in the fosmid sequence data. [b]High-confidence call set requires $\log_{10}$ (Bayes factor) for the reported genotype to be at least 4.

At $k = 55$, 10% of SNPs fail to make clean bubble structures (**Supplementary Fig. 4**); however, coverage was sufficient such that only for rare variants ($N < 3$) was there a substantial loss of power (**Fig. 2c**). With realistic levels of sequence error, power dropped by an additional 10% under the relaxed cleaning threshold but was recovered under the more stringent cleaning threshold because confounding error contigs were removed. FDR with relaxed cleaning was 29%, but probabilistic classification reduced this to 1.5% with only a 1.7% loss in power (**Fig. 2d**). The more stringent cleaning approach had an FDR of 2.3% before classification and 1.6% after, with a 1.0% loss in power. In contrast, removal of bubbles present in the reference had only a marginal effect on FDR (29.0% and 2.3% for the two cleaning thresholds, respectively), as most false calls are read-error driven.

Case 1: variant calling in a high-coverage human genome

We analyzed high-coverage data (26×, 100-bp reads, $k = 55$; see **Supplementary Note**) from a single individual of European ancestry (NA12878 from the Utah residents of Northern and Western European ancestry (CEU)) for whom independent validation data were available through 3 Mb of fosmid sequence (median length of 40 kb), selected to contain structural variation[41]. This sample has been analyzed using mapping-based strategies in the 1000 Genomes Project[17] (63× of mostly 36-bp paired-end reads), thus enabling comparison of Cortex with alternative strategies. The fosmid data enabled us to estimate an upper bound for FDR for variants of different types from the fraction of sites called as homozygous variant, where the fosmid sequence contains only the reference allele. A detailed discussion of the validation results can be found in the **Supplementary Note**.

After cleaning, the de Bruijn graph for NA12878 had 2,777,352,792 nodes (unique k-mers) compared to 2,691,115,653 in the reference sequence (cleaning reduced the initial number of nodes by 23%). The BC algorithm identified 2,686,963 bubbles, of which 5.6% were removed because both sides of the bubble were also present in the reference. The PD algorithm identified 528,651 deviations from the reference, of which 39.8% were not identified by BC. The union of the BC and PD call sets included 2,245,279 SNPs, 361,531 short indels (insertion-to-deletion ratio in the 5–30-bp range of 1:1.3 for BC and 1:1.7 for PD compared with 1:3.7 for the 1000 Genomes calls) and 1,100 larger or more complex variants.

The Cortex and 1000 Genomes call sets have different properties arising from differences in experimental design and analysis approach. Only 80% of the genome is accessible to the 1000 Genomes SNP calls[17], but power in these regions is high. In contrast, at $k = 55$, more than 85% of the genome is accessible to Cortex, but power is reduced (by approximately 40% at heterozygous and <5% at homozygous sites) owing to fluctuations in coverage. Thus, whereas the call-set sizes for homozygous SNPs and short indels in the fosmid footprint were similar, Cortex called only half the number of heterozygous sites (**Table 1**). Across the genome, Cortex detected variation at 87% of sites called as homozygous alternative by the 1000 Genomes Project and 67% of sites called as heterozygous (both our model prediction and comparison to HapMap 2 sites gave equivalent figures; **Supplementary Note** and **Supplementary Table 1**).

SNP variants identified by Cortex have diagnostic properties, such as transition-transversion ratio (Cortex = 2.02 (BC) and 2.1 (PD), 1000 Genomes = 2.07) and dbSNP rate (Cortex = 92.7% (BC) and 95.6% (PD), 1000 Genomes = 92.1%; dbSNP 129) that are comparable to the 1000 Genomes calls[17].

The overall FDR for Cortex SNP calls was 4%, reduced to 1.5% by applying a homopolymer filter and selecting high-confidence calls (25% reduction in call set; **Table 1** and **Supplementary Note**). None of the 1000 Genomes homozygous SNP calls was invalidated (**Table 1**). Of the 43 invalidated homozygous SNP calls from Cortex, 35 were called as heterozygous sites by the 1000 Genomes Project; hence, the FDR for Cortex was probably <1%. Short indels (1–100 bp) had similar FDRs (0% for high-confidence set), but whereas the 1000 Genomes calls were restricted to variants less than 30 bp in length, both the BC and PD approaches identified indels of longer than 100 bp (**Fig. 3a**).

Across the genome, Cortex identified 138,262 complex variants, consisting of phased SNPs (74%), closely sited SNPs and indels (25%) and complexes of insertions, deletions and local rearrangements (1%). FDR for complex variants was low (2.7% for BC and 1.7% for PD; **Supplementary Tables 2** and **3**). Although mapping-based approaches can call closely sited variants of different types, these are often filtered out. However, our results indicate that Cortex can identify complex variants with an FDR comparable to that for simple variants. **Figure 3b** shows examples of complex variants that were validated in the fosmid data.

Case 2: detection of novel sequence from population graphs

We constructed three pooled population graphs for 164 humans (CEU, Yoruba from Ibadan (YRI) and Han Chinese from Beijing (CHB) + Japanese from Tokyo (JPT)) sequenced at low coverage (2–4×) in the 1000 Genomes Project[17] (see **Supplementary Note**). By including the reference sequence as a fourth color, we identified 21,281 novel contigs of ≥100 bp (<90% homology to any reference sequence) totaling 3.2 Mb. The novel unique sequence load carried by a typical individual was 1.4 Mb for CEU, 1.5 Mb for YRI and 1.5 Mb for CHB + JPT populations, respectively. Of this, 93% was estimated to be allelic, and copy-number estimates for other sequences ranged up to 6.3 (**Fig. 4a**). On average, 45 kb per individual is homologous to a known gene, and we saw strong overrepresentation for matches to variants at classical HLA and killer cell immunoglobulin-like receptor (KIR) loci, which are both known to be highly variable in sequence,

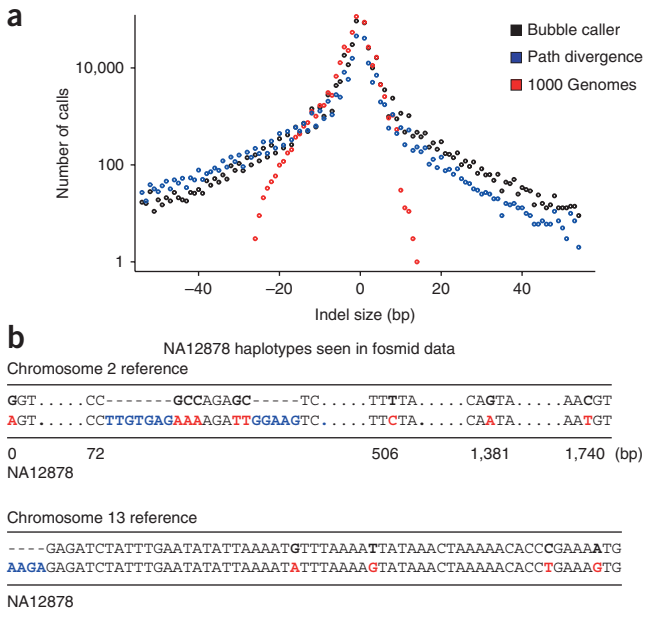

Figure 3 Structural and complex variants identified in a single high-coverage genome. (**a**) Size distribution of short indels discovered in NA12878 from 26× coverage of 100-bp reads analyzed using the BC (black) and PD (blue) algorithms. Also shown is the number of indels of different sizes called by mapping-based approaches from 63× coverage on the same sample within the 1000 Genomes Project[17] (red). Although the 1000 Genomes Project calls more small variants, only the two Cortex algorithms can detect longer variants, which are typically too short to be detected by larger-structural-variation discovery methods. The PD caller shows bias toward calling deletions for larger variant sizes. (**b**) Two examples of complex variants identified in NA12878 using Cortex and validated in the independent fosmid data. Above, the PD algorithm assembles a haplotype of more than 1.7 kb containing a number of phased SNPs (red) and a complex indel event (inserted material is blue). Below, the BC algorithm assembles two haplotypes containing four phased SNPs and an insertion.

structure and copy number. Some sequences showed strong differentiation between populations. For example, we found three contigs in the YRI population homologous to olfactory receptor genes, which were absent from other populations (**Fig. 4a**).

There are practical implications for these results. First, the novel sequences, particularly those matching genes or strongly differentiated between populations, are candidates for functionally relevant polymorphism. Second, the combined population graphs provide a summary of human genome diversity against which it is possible to map sequence data from future studies. We have released the novel sequence contigs, population estimates of the per-genome copy number, the combined population graph and tools for aligning reads to the graph (see **Supplementary Note**).

Case 3: using population information to classify bubbles

We applied Cortex to data from ten Western Chimpanzees (50-bp reads, average coverage 6×; see **Supplementary Note**). Bubbles were identified after using the relaxed cleaning threshold and classified probabilistically. Power was estimated by comparison to previous SNP genotype data on the same samples[42].

Across the genome, we identified 3.5 million variants, of which 2.7 million were single-nucleotide variants. The probabilistic filter classified 153,921 of these as repeats, of which 69% were bubbles in the reference. For bubbles classified as SNPs, we estimated FDR from the fraction of sites that were also bubbles in the reference, here, 3.5% (compared to 6.5% before classification). This estimate is substantially greater than that predicted from simulations. However, manual inspection revealed that many of these sites were segregating in the sample and, therefore, are either polymorphic segmental duplications (not currently considered in the classification process) or allelic variants misassembled as paralogous in the reference. Power compared to the SNP genotype data was 55% before classification and 54% after. Thus, probabilistic classification results in a data set with low FDR at a small cost to power and can be applied to any species, regardless of reference availability. The relationship between allele count and detection rate closely follows the theoretical predictions (**Fig. 4b**).

Case 4: genotyping simple and complex variants

We applied the genotyping algorithm to both simple variants, here HapMap2 SNPs[43], and complex variants, specifically *HLA-B* genotypes, using the sequencing data from NA12878 described above and high-coverage sequence data from an individual of African origin (NA19240). Both individuals have classical HLA alleles typed in a previous project[44]. At HapMap2 SNP sites, we report discordance of 1.1% (BC) at high-confidence sites (**Supplementary Tables 4–7**; see **Supplementary Note**). Discrepancies were driven by sites called as homozygous variant by the BC algorithm and heterozygous by HapMap2—a result of stochastic loss of coverage of *k*-mers spanning the reference allele.

Classical HLA allele genotyping, of importance in many areas of medical genetics, is laborious and expensive. Although DNA sequencing represents the gold standard for quality, most genotyping is performed through a mixture of PCR amplification and oligonucleotide

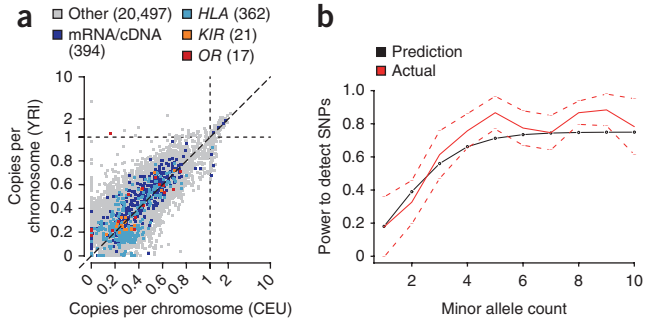

Figure 4 Population analysis with Cortex. (**a**) Estimates of mean copy number per genome in CEU and YRI populations for novel sequence contigs identified from analysis of pooled population graphs for 164 humans sequenced to low depth (2× to 4×) by the 1000 Genomes Project. Contigs are all at least 100 bp long and have <90% homology to the reference genome (as determined by BLAST). Allelic variation lies in the interval (0,1), whereas copy number–variable sequence can be present up to 6.3 times. Variants are annotated by whether they show significant homology to known transcripts, including *HLA*, *KIR* and olfactory receptor (*OR*) genes as specific categories, which are clearly enriched. We note the presence of several OR matches that show approximately 20% frequency in the YRI population but seem to be absent from the CEU set. (**b**) Power to detect SNP variants previously analyzed through SNP genotyping in HTS data from ten chimpanzees (6× coverage, 50-bp reads analyzed at $k = 31$). Empirical estimates (red, with normal approximation to binomial confidence intervals shown dashed) closely track predictions (black) from the theoretical model.

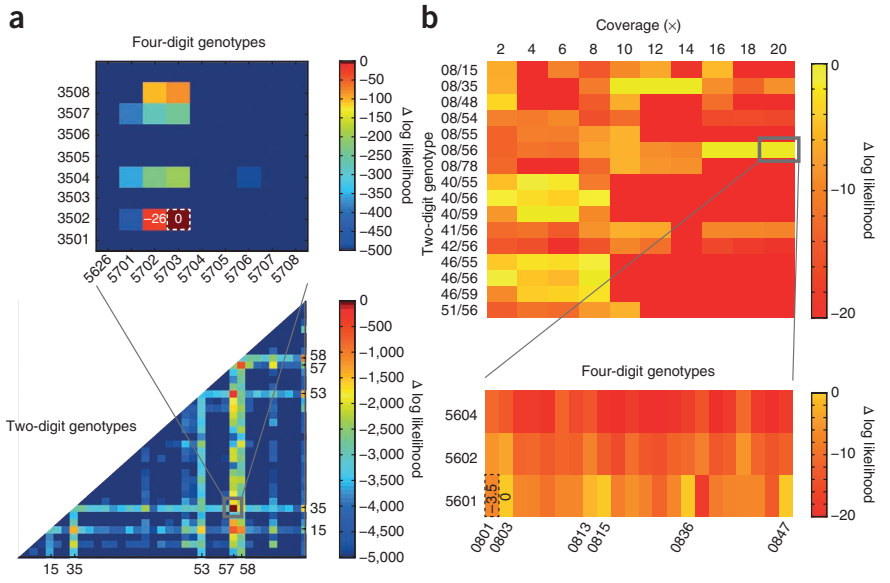

Figure 5 *HLA-B* genotyping from HTS data using Cortex. (**a**) Heat plot showing the likelihood surface for *HLA-B* genotypes for NA19240 (the child of the Yoruban trio from the 1000 Genomes Project[17]). Below, the likelihood surface at two-digit resolution (represented by the most likely genotype among all compatible alleles); above, an expanded view of the most likely two-digit genotype (B*57/B*35). The maximum-likelihood estimate (MLE) genotype is B*57:03:01/B*35:01:01, which agrees with the four-digit resolution data generated previously using standard experimental methods[44] (MLE shown by dashed line; difference in log likelihood from MLE shown for selected genotypes). Other alleles identified as possible at the two-digit level (HLA-B*15, HLA-B*53 and HLA-B*58) are closely related to those present in the sample. B*53 is known to be a product of gene conversion from B*35 (ref. 52), and B*58 is a split antigen from B*15 with sister serotype B*57. (**b**) Heat plot showing the likelihood surface at two-digit resolution for selected *HLA-B* genotypes for NA12878 (above) as a function of sequence depth. The expanded view (below) shows four-digit resolution at 20×. At low coverage, there is no consistent, most-likely genotype; from 10× to 14×, the most likely is B*08/B*35, which switches to B*56:01/B*08:xx (where xx = 3, 13, 15, 36 and 47; all have log likelihood within 0.03 units) at 16×. Laboratory-based typing gives B*56:01/B*08:01, which is 3.2 units in log likelihood less than the MLE.

hybridization. HTS genome-wide data have the potential to provide classical HLA sequence information, but sequence diversity, structural variation and extensive paralogy in the region currently restrict mapping-based approaches. To evaluate the performance of Cortex for genotyping *HLA-B*, we constructed a graph containing the reference genome, all 1,429 known *HLA-B* alleles and data from each high-coverage sample as separate colors (see **Supplementary Note**). We calculated the likelihood of all 1,021,735 possible genotypes. For NA19240, the most likely genotype (B*57:03:01/B*35:01:01) agreed at four-digit resolution with previous data obtained using classical typing methods[44] and was very strongly supported (likelihood ratio of ~10^{23}; **Fig. 5a**). For NA12878, the most likely genotype contained B*56:01 and could not distinguish between B*08:03, B*08:15, B*08:36, B*08:47 and B*08:13 for the second allele (**Fig. 5b**). The laboratory-based genotype was B*56:01/B*08:01, which differs from the maximum-likelihood estimate by 3.2 units of log likelihood and agrees at the two-digit level with the graph-based estimate. We note that laboratory typing was performed by primer amplification and oligonucleotide hybridization, which can often lead to minor ambiguities at the four-digit level. By subsampling NA12878 data to generate graphs using coverage between 2× and 20×, we found that 16× coverage was required to attain two-digit agreement with laboratory-based typing (**Fig. 5b**).

DISCUSSION

We have introduced a new approach to combining *de novo* assembly with variant detection and genotyping from HTS data. We use colored de Bruijn graphs to represent information from multiple sources and a mixture of graph-analytical and statistical approaches to detect variants of different types and, subsequently, genotype. Our method, to our knowledge, is the first *de novo* assembly-based variant caller, although previous work has made steps toward reference-free variant calling[45,46]. Technically, the key advance is the development of a highly efficient de Bruijn graph implementation. This efficiency enables data from multiple samples, as well as reference sequences and known variants, to be included in a single graph structure that preserves sample identity through the use of colors. For single high-coverage genomes, the algorithms provide power to detect and genotype simple and complex variants. However, the main strength of the approach lies in the simultaneous analysis of multiple genomes, which enables powerful and accurate approaches to variant detection without the need for a reference genome. This makes possible HTS analysis of genetic variation in any species. It could also provide an approach for detecting changes between highly related genomes, as in tumor-normal pairs in cancer genomics[47] or bacteria in transmission chains[48].

We have also developed a simple mathematical model to describe assembly of de Bruijn graphs from HTS data, which has two practical benefits. First, the model has predictive power both in simulated and empirical data, so it can guide experimental design. Second, the model can be used to calculate the likelihood of any particular genome sequence given HTS data and an estimate of error rates; one application of this is genotyping complex variants, such as the classical HLA loci.

Last, the Cortex algorithms have several limitations. Most notably, we did not use read-pair information to improve local assembly, which can be of substantial value around repeat sequences. However, there are established algorithms for using read-pair information to disambiguate de Bruijn graphs[30,31,39,40,49,50]. There is also the potential for error correction[51], which can compensate for the loss of coverage caused by errors. There are, however, more fundamental challenges in using de Bruijn graphs, including the greater need for error correction as the *k*-mer size increases, the lack of any natural way of encoding read-pair information and the potential for graph explosion as more individuals are included in the graph. Nevertheless, multicolored graphs provide one solution to the obvious inadequacy of representing the genetic composition of a species by a single haploid reference.

URLs. Cortex, http://cortexassembler.sourceforge.net/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

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

ACKNOWLEDGMENTS

We would like to thank the members of the 1000 Genomes Project Consortium for discussion, suggestions and sequencing data. We thank B. Ahiska, A. Auton, E. Birney, R. Durbin, G. Lunter, J. Woolf and D. Zerbino for discussion, two anonymous reviewers for their comments and members of the PanMap Project and the Genomics Core at the Wellcome Trust Centre for Human Genetics for access to sequence data. Z.I. is funded by a grant from the Wellcome Trust (WT086084/Z/08/Z to G.M.). The sequencing of NA12878 was performed by the Wellcome Trust Sequencing Core at Oxford, under a grant from the Wellcome Trust (090532/Z/09/Z).

AUTHOR CONTRIBUTIONS

Z.I. and G.M. designed the study, developed the mathematical models and wrote the manuscript. M.C. and Z.I. developed the variant discovery algorithms, designed the multicolor graph data structures and implemented software. Z.I. performed simulations and analyses for cases 1, 3 and 4. I.T. and Z.I. performed analyses for case 2. P.F. contributed to early plans for Cortex.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
3. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
4. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
5. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
6. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
7. Albers, C.A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).
8. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* **6**, 473–474 (2009).
9. Hajirasouliha, I. *et al.* Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26**, 1277–1283 (2010).
10. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
11. Korbel, J.O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).
12. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
13. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
14. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
15. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
16. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
17. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
18. Ge, F., Wang, L.S. & Kim, J. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.* **3**, e316 (2005).
19. Beiko, R.G., Harlow, T.J. & Ragan, M.A. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 14332–14337 (2005).
20. Holcomb, C.L. *et al.* A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue Antigens* **77**, 206–217 (2011).
21. Fonseca, V.G. *et al.* Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nat. Commun.* **1**, 98 (2010).
22. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
23. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
24. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
25. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
26. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
27. Myers, E.W. Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* **2**, 275–290 (1995).
28. Myers, E.W. The fragment assembly string graph. *Bioinformatics* **21** (suppl. 2), ii79–ii85 (2005).
29. Simpson, J.T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).
30. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
31. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
32. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
33. Jones, T. *et al.* The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci. USA* **101**, 7329–7334 (2004).
34. Vinson, J.P. *et al.* Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res.* **15**, 1127–1135 (2005).
35. Kim, J.H., Waterman, M.S. & Li, L.M. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.* **17**, 1101–1110 (2007).
36. Donmez, N. & Brudno, M. Hapsembler: an assembler for highly polymorphic genomes. in *Research in Computational Molecular Biology, Lecture Notes in Computer Science* Vol. 6577 (eds. Bafna, V. & Sahinalp, S.), 38–52 (Springer, Berlin, Heidelberg, 2011).
37. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
38. Idury, R.M. & Waterman, M.S. A new algorithm for DNA sequence assembly. *J. Comput. Biol.* **2**, 291–306 (1995).
39. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
40. Zerbino, D.R., McEwen, G.K., Margulies, E.H. & Birney, E. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read *de novo* assembler. *PLoS ONE* **4**, e8407 (2009).
41. Kidd, J.M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
42. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**, 876–879 (2010).
43. The International HapMap Consortium. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
44. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
45. Ratan, A., Yu, Z., Hayes, V.M., Schuster, S.C. & Miller, W. Calling SNPs without a reference sequence. *BMC Bioinformatics* **11**, 130 (2010).
46. Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.-F. & Lacroix, V. Identifying SNPs without a reference genome by comparing raw reads. in *String Processing and Information Retrieval—17th International Symposium* (eds. Chavez, E. & Lonardi, S.) 147–158 (Los Cabos, Mexico, 2010).
47. Ding, L., Wendl, M.C., Koboldt, D.C. & Mardis, E.R. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum. Mol. Genet.* **19**, R188–R196 (2010).
48. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
49. Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
50. Chaisson, M.J., Brinza, D. & Pevzner, P.A. *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.* **19**, 336–346 (2009).
51. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
52. Allsopp, C.E. *et al.* Sequence analysis of HLA-Bw53, a common West African allele, suggests an origin by gene conversion of HLA-B35. *Hum. Immunol.* **30**, 105–109 (1991).

ONLINE METHODS

Full details of the methods used in this paper can be found in the **Supplementary Note**. These include definitions of terminology and the following sections: (i) mathematical model for power to detect variants, (ii) variant calling algorithms (bubble caller and path-divergence caller), (iii) probabilistic classification of graph structures as repeat, error or variant induced, (iv) genotyping of simple and complex variants in a de Bruijn graph, (v) description of Cortex software implementation, (vi) error-cleaning algorithms for high-coverage samples and low-coverage populations, (vii) simulation 1: high-coverage single-genome, (viii) simulation 2: low-coverage population, (ix) case 1: analysis of high-coverage human sample NA12878, (x) case 2: pooled assembly of 164 samples from the 1000 Genomes Pilot, (xi) case 3: using probabilistic classification of bubbles on ten chimpanzees, (xii) case 4: genotyping of simple and complex variants, (xiii) making appropriate choices of parameters for experimental design and (xiv) release of source code and 1000 Genomes population graph.