Published in final edited form as:

Nat Methods. 2015 August ; 12(8): 780–786. doi:10.1038/nmeth.3454.

Assembly and diploid architecture of an individual human genome via single-molecule technologies

Matthew Pendleton[1,15], **Robert Sebra**[1,15], **Andy Wing Chun Pang**[2,15], **Ajay Ummat**[1,15], **Oscar Franzen**[1], **Tobias Rausch**[3], **Adrian M Stütz**[3], **William Stedman**[2], **Thomas Anantharaman**[2], **Alex Hastie**[2], **Heng Dai**[2], **Markus Hsi-Yang Fritz**[3], **Han Cao**[2], **Ariella Cohain**[1], **Gintaras Deikus**[1], **Russell E Durrett**[4], **Scott C Blanchard**[5], **Roger Altman**[4], **Chen-Shan Chin**[6], **Yan Guo**[6], **Ellen E Paxinos**[6], **Jan O Korbel**[3,7], **Robert B Darnell**[8,9], **W Richard McCombie**[10,11], **Pui-Yan Kwok**[12], **Christopher E Mason**[4,13,14], **Eric E Schadt**[1], and **Ali Bashir**[1]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[2]BioNano Genomics, San Diego, California, USA

[3]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[4]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York, USA

[5]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York, USA

[6]Pacific Biosciences, Menlo Park, California, USA

[7]European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK

[8]Laboratory of Neuro-Oncology, The Rockefeller University, New York, New York, USA

[9]Howard Hughes Medical Institute, New York, New York, USA

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

Correspondence should be addressed to A.B. (ali.bashir@mssm.edu).

[15]These authors contributed equally to this work.

Accession codes. NCBI: whole-genome sequencing data and assemblies for NA127878 have been deposited under BioProject PRJNA253696 with Sequence Read Archive accession numbers SRX627421 and SRX638310.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

E.E.S., A.B., R.S., C.E.M., W.R.M. and R.B.D. conceived the project and provided resources for sequencing and algorithmic analysis. A.B. and E.E.S. provided bioinformatics oversight. J.O.K., M.H.-Y.F., A.M.S. and T.R. performed Illumina SV analysis and PCR validation. R.S., M.P. and E.E.P. prepared long libraries for PacBio sequencing. R.S., Y.G., A.C., S.C.B., R.A. and R.E.D. performed PacBio sequencing and primary analysis of hdf5 data. A.W.C.P., H.D., A.H., T.A., W.S., H.C. and P.-Y.K. generated the BioNano Data, built initial Genome Maps and performed BioNano alignment and SV calling. O.F., A.B. and M.P. performed PacBio SV analysis and validation. A.U., A.B. and C.-S.C. performed error correction and assembly. A.W.C.P. and H.D. built the initial hybrid scaffolding pipeline. A.B. and M.P. refined the hybrid scaffolding pipeline. A.W.C.P., H.D. and A.B. performed scaffold analysis and phasing. A.B., A.W.C.P., M.P. and A.U. generated figures for the main text. A.B., E.E.S., R.S., M.P. and A.W.C.P. primarily wrote the manuscript, though many coauthors provided edits and methods sections.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

[10]The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

[11]The Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

[12]Institute for Human Genetics, University of California–San Francisco, San Francisco, California, USA

[13]Department of Medicine, Division of Hematology/Oncology, Weill Cornell Medical College, New York, New York, USA

[14]The Feil Family Brain and Mind Research Institute, Weill Cornell Medical College, New York, New York, USA

Abstract

We present the first comprehensive analysis of a diploid human genome that combines single-molecule sequencing with single-molecule genome maps. Our hybrid assembly markedly improves upon the contiguity observed from traditional shotgun sequencing approaches, with scaffold N50 values approaching 30 Mb, and we identified complex structural variants (SVs) missed by other high-throughput approaches. Furthermore, by combining Illumina short-read data with long reads, we phased both single-nucleotide variants and SVs, generating haplotypes with over 99% consistency with previous trio-based studies. Our work shows that it is now possible to integrate single-molecule and high-throughput sequence data to generate *de novo* assembled genomes that approach reference quality.

The availability of high-throughput sequencing data has deepened our understanding of human genomes tremendously. Both single-nucleotide variants (SNVs) and small insertions or deletions (indels) can now be reliably genotyped[1,2]. Yet it is not possible to fully characterize all of the variation between any pair of individuals. In fact, though the cost of sequencing has markedly decreased, *de novo* human genome analysis has, to some extent, regressed. Although HuRef and the original Celera whole-genome shotgun assembly have scaffold N50 values (the length such that 50% of all base pairs are contained in scaffolds of the given length or longer) of 19.5 Mb (ref. 3) and 29 Mb (ref. 4), respectively, the best next-generation sequencing (NGS) assemblies have scaffold N50 values of 11.5 Mb (ref. 5), even with the use of high-coverage fosmid jumping libraries. Additionally, NGS technologies have difficulty inferring repetitive structures[6], such as microsatellites, transposable elements, heterochromatin[7] and segmental duplications[8], which is further complicated by gaps and errors in the reference genome.

Existing technologies are constrained by short read lengths and bias. Ensemble-based NGS technologies[9] generate sequence reads of limited length, and even jumping libraries that allow read pairs to span long distances cannot generally resolve structures in highly repetitive regions. Further, NGS technology is prone to systematic amplification and sequence composition biases[10,11]. Amplification-free single-molecule sequencing substantially extends read lengths while also reducing sequencing coverage bias[12]; however, such data require new informatics strategies. Single Molecule Real-Time (SMRT) sequencing using the Pacific Biosciences (PacBio) platform delivers continuous reads from individual molecules that can exceed tens of kilobases in length, albeit with error rates (mainly indels) above 10%. Another recent technology, the NanoChannel Array (Irys System) from BioNano Genomics (BioNano), confines and linearizes DNA molecules up to hundreds of kilobases to megabases in length. Rather than providing direct sequence information, the technology uses nicking enzymes to provide high-resolution sequence motif physical maps, termed ‘genome maps’. *De novo*–assembled genome maps can be used as scaffolds for assembled genomic sequences or compared to a known reference to infer variation or haplotype information. These single-molecule approaches have proven invaluable for the assembly of small genomes[13–17] and in targeted settings for analyzing complex variations in human genomes[18,19].

Here, we present a comprehensive analysis of a diploid human genome based on SMRT sequencing data and single-molecule genome maps from the Irys System. Individually, the assemblies and genome maps markedly improve contiguity and completeness compared with *de novo* assemblies from clone-free, short-read shotgun sequencing data. Moreover, by combining the two platforms, we achieve scaffold N50 values greater than 28 Mb, improving the contiguity of the initial sequence assembly nearly 30-fold and of the initial genome map nearly 8-fold. This represents the most contiguous clone-free human genome assembly to date and is comparable to, or better than, assemblies using mixtures of fosmid or BAC libraries. Furthermore, using reference-based approaches, we are able to better resolve complex forms of structural variation, including tandem repeats (TRs) and multiple colocated events. Additionally, whereas short-read sequencing is restricted to small haplotype blocks, we can generate haplotype blocks several hundreds of kilobases in size, sometimes filling in gaps missed by trio-based analyses.

RESULTS

We sequenced NA12878 genomic DNA across 851 Pre P5-C3 and 162 P5-C3 SMRTcells to generate 24× and 22× coverage with aligned mean read lengths of 2,425 and 4,891 base pairs, respectively. We constructed genome maps using 80× coverage of long molecules (>180 kb) with mean spans of 277.9 kb.

We used an integrated assembly and resequencing strategy (Supplementary Fig. 1). In short, error-corrected PacBio reads were assembled with the Celera Assembler[17] and Falcon (Online Methods) to provide initial sequence contigs. Genome maps were iteratively merged with the assembled sequence contigs to yield final scaffolds. Assembled contigs, genome maps, error-corrected reads and raw PacBio reads were used to detect TRs and SVs in reference analyses. Last, short-read data identified SNVs and indels that were passed, along with PacBio reads, into a two-step phasing pipeline.

Assembly

Assembly performance on NA12878 varies across the multiple technologies and data sets generated in this study (Fig. 1 and Table 1). The initial genome maps have a substantially higher scaffold N50 (4.6 Mb versus 0.9 Mb, approximately fivefold higher) than the more comprehensive SMRT sequencing assembly, albeit without single-base resolution. The longer genome maps anchor sequence contigs across difficult repeat regions (4,007 contigs merged via genome maps), as expected; but notably, the hybrid approach improves the genome mapping assembly nearly as dramatically, with 848 instances of long-read contigs bridging genome maps. This suggests an independent contig fragmentation mechanism between sequence-based and genome map assemblies. In addition to long repeat regions and intervals with low nick-site density, the genome map assembly may break around ‘fragile sites’ (where two nick sites are proximally located on opposite strands), leading to biased DNA double-strand fragmentation[20,21]. We observed a significant enrichment in the density of fragile sites within 20 kb of genome map ends compared to all expected fragile sites in the human genome ($P < 5.0 \times 10^{-261}$ assuming a Poisson site distribution; Supplementary Fig. 2). The complementarity of break mechanisms between contigs (repeats) and genome maps (fragile sites) supports a stronger merged assembly.

To reduce misassemblies, we compared SMRT contig and genome map assemblies to identify inconsistent regions. Such inconsistencies could be the result of assembly errors or alternative haplotypes; 31 ‘junctions’ (alignments with at least three unaligned contig and genome map labels upstream or downstream of the aligned region) were identified, and corresponding contigs and genome maps were removed. Hybrid scaffolding was run on the remaining data, which resulted in 377 hybrid scaffolds with a scaffold N50 of 13.6 Mb (Table 1). A second round of scaffolding to progressively incorporate the more aggressive sequence assembly, generated by Falcon, resulted in 202 hybrid scaffolds with a scaffold N50 of 31.3 Mb. Though the number of joins was not substantial between iterations, the practical impact on contiguity was substantial. For example, in chromosome 18, the V1 scaffold contains three and four scaffolds in the p and q arms, respectively, whereas the V2 scaffold yields single scaffolds in both arms.

The hybrid scaffold is smaller (2.76 Gb) than the initial genome map (2.92 Gb), with 82% (2.5 Gb) of the sequence contigs anchored within scaffolds. Including sequence contigs that could not be anchored (owing to insufficient mapping quality or representation of alternate haplotypes) leads to a revised scaffold N50 of 28.4 Mb and a genome totaling 3.16 Gb (Table 1 and Supplementary Table 1).

Contiguity and accuracy of scaffolds relative to hg19—We compared our assembly against the published Allpaths-LG NA12878 assembly, which used short-read sequencing of insert and fosmid libraries[5]. A high-level comparison of the two assemblies, using metrics from refs. 22 and 23, is shown in Supplementary Table 1. Normalizing to hg19, our assembly has a higher contig N50 (886 kb versus 19 kb), scaffold N50 (26 Mb versus 10 Mb) and ‘scaffold accuracy’ (98.7% versus 94.9%), which represents the odds of being correctly connected at a distance of 100 kb. Additionally, fewer hg19 reference bases are missing (14.9% versus 7.6%), and more new assembly sequence was potentially added (58 Mb versus 9 Mb).

However, our sequence identity compared to hg19 is lower (99.7% versus 99.8%); though some of this deviation may be due to detection of true variants or alternative alleles, much of it represents miscalled small indels that result from the higher, indel-based error rate of SMRT sequencing. Such errors can be resolved by mapping short-read data[1] to contigs and using variant calls to correct contigs, leading to sequence identity consistent with the Allpaths-LG assembly (Supplementary Table 2). Using heterozygous SNVs, over 2 Gb of sequence was resolvable into haplotype blocks, with a haplotype N50 of 145 kb (Supplementary Table 3). Last, we measured the structural fidelity of both scaffolds by performing nick-site mapping (or *in silico* nick-site mapping) relative to each other and hg19 (Supplementary Fig. 3). Fewer chimeras were observed in the V2 scaffolds; moreover, when discrepancies existed between V2 and Allpaths scaffold mapping, the V2 scaffold was 15 times more likely to be consistent with hg19 (Supplementary Results).

Reference-based analyses

Phasing—Phasing was performed using a combination of short-and long-read approaches, enabling long haplotype blocks with low switch error rates and resolving unphased variants from trio-based approaches. SNVs and indels previously identified by deep Illumina sequencing of the NA12878 trio[24] represented 2,367,085 heterozygous events (1,925,040 phased by trio analysis), far more than those detected by PacBio sequencing alone (Supplementary Results). The consistency of SMRT sequencing–based phasing with trio results markedly improved when SNV filtering was performed with a modified reference (Supplementary Table 1 and Online Methods). The switch error rate was estimated by measuring concordance of the predicted haplotype blocks with the SNVs labeled by trio-based phasing. After filtering, the estimated switch error rate was reduced to 0.1, which was lower than the estimated switch error rate of 0.9 in HuRef[25]. There is a trade-off between reducing the switch error rate and eliminating SNVs from analysis (using multiple parameters for both long- and short-read data sets; Supplementary Fig. 4). To increase accuracy, 369,785 SNVs were eliminated from the analysis; however, a similar number of additional variants not amenable to trio-based phasing were resolved (314,630) via the long reads. For all heterozygous TR and SV events, we used local phasing in an attempt to assign events to either the maternal or paternal haplotype. Both alleles were assigned to distinct haplotypes for 9,196 TRs and 3,562 SVs. The assignment approach also serves to assess the accuracy of heterozygous calls. True events should have distinct haplotypes assigned to the two read clusters; here, 97% of predicted heterozygous SVs form consistent haplotypes (Supplementary Fig. 5 and Supplementary Results).

Structural variation—SV calls from PacBio data were generated using the *de novo* sequence assembly and read-mapping approaches (Online Methods). Some SVs evaluated by locally comparing our *de novo* sequence assembly to syntenic intervals of hg19 (using tools developed in ref.26) were shared with those previously detected in the CHM1 haploid cell line[26] for both insertions (39%) and deletions (12%). Although each genome clearly had many unique variants, they were largely comparable in the magnitude of calls (Supplementary Result and Supplementary Table 4).

SVs were further evaluated by aligning individual raw and error-corrected reads to hg19 (Supplementary Tables 5 and 6 and Supplementary Fig. 6). Short-read calls from tandem duplications, deletions and inversions were evaluated on both short and long insert data using Delly[27]. These were compared to the PacBio call set and further evaluated by manually inspecting break-point-spanning reads (Supplementary Result and Online

Methods). Of the callable short-read predictions, 95% agreed in approximate variant type with the PacBio data (Table 2). Substantially more SVs were predicted in the PacBio data set (Supplementary Table 5), even when considering assembly- mapping or read-mapping approaches separately. For insertions, as expected, the most frequent mobile element insertions corresponded to *Alu* elements, L1s and SVAs (SINE-VNTR-Alu) (Supplementary Table 4 and Supplementary Fig. 7), elements known to be active in the human genome and expanded in the human lineage[28,29]. To estimate the false discovery rate of the mapping-based calls without short-read support, we interrogated events using long-range PCR. For PCR validation, SV predictions were divided into bins on the basis of predicted event size (200–500 bp, 500–1,000 bp, 1,000–1,500 bp, 1,500–5,000 bp). Of the 59 successful PCR reactions, 58 positively supported the predicted event (Supplementary Results and Supplementary Table 7). Additionally, whole-genome data from Tru-seq, an orthogonal long-read platform, were largely consistent with PacBio calls (Supplementary Result).

Tandem repeats—TRs represent an important source of variation that are associated with a broad range of diseases[30] but are not easily addressed by NGS technologies. The combination of single-molecule and long-read approaches not only identified large TRs outside the range of short-read approaches but also suggested that many TRs are substantially larger than indicated in hg19 (Fig. 2). Some of these differences may be explained by allelic variation at a given locus (Fig. 2b), but there is clearly a systematic underrepresentation of repeat signals in the reference (Fig. 2a), with certain regions showing increased variability (Supplementary Results and Supplementary Fig. 8). Small events were compared to short-read predictions by RepeatSeq[31] and showed >90% concordance (Supplementary Fig. 9). As events become larger, they seem to be more consistent with the reference. However, this is most likely due to the exponential read-length distribution observed in SMRT sequencing. As fewer reads reliably span larger TR intervals, high-quality alignments are more likely to be observed when consistent with the reference. Thus, very large TRs (Fig. 2c) can only be directly examined using genome mapping data. An example of this is a TR within the *LPA* gene on chromosome 6q26 (the ~5.6-kb ‘kringle-IV’ type 2 (KIV2)-like domain; Fig. 2c); long molecules spanning over 100 kb are needed to reliably span the TR. Correctly identifying its multiplicity is particularly relevant as *LPA* size has been associated with plasma lipoprotein level and risk of cerebrovascular and cardiovascular diseases in the human population[32].

Characterization of variation

Large variation via assembly and scaffolding—The scaffolding results largely validate the layout predicted by hg19 (Fig. 1); however, large structural variation events are observed in the hybrid scaffold (Supplementary Table 8), and a number of large SVs were directly identified by genome mapping (Supplementary Table 9). To validate these events, we compared the genome maps, and the raw molecules used to construct them, to hg19. For example, a 206.6-kb insertion seems to be a large TR expansion (Fig. 3a). A number of raw molecules spanning the event support the BioNano assembly, whereas no spanning molecules confirm the smaller reference allele. The observed difference could be due to variability in the population or artificial compression from traditional assembly approaches. In another example, a 577.3-kb inversion spans previously unresolved regions in hg19 (Fig.

3b), suggesting potential misassembly during BAC tiling layout. This is supported by higher concordance with the updated GRCh38 (hg38) assembly. Other large events (Supplementary Fig. 10) show our assembly to be more consistent with hg38, suggesting that the hg38 assembly has fixed errors in hg19 or is more representative of dominant haplotypes. Yet, despite these improvements, gaps still persist in hg38. Our sequence assembly resolves 28 previously defined ‘interstitial gap’ intervals[26], yielding 34 kb of assembled sequence that spans 621 kb in hg38 (Supplementary Table 10). The resulting gap sequence is enriched for simple repeats (Supplementary Table 10), consistent with previous long-read gap closure results in hg19 (ref.26).

Complexity of variant sequences—As mentioned earlier, spanning long reads enable direct observation of breakpoints and inserted sequences. This allows one to distinguish mobile element insertions that only contain the repeat element from those that contain other inserted sequences (Supplementary Fig. 7a). Although many of these events may be duplications not derived from mobile element insertion, some of these intervals are the result of SVA[33,34] or L1 (ref.35) DNA transduction: for example, a 5′ truncated SVA element mediating the 3′ transduction of a proximal *Alu* sequence (Supplementary Fig. 11). SMRT sequencing long reads can also be used to distinguish subtle SV insertions within TR intervals[36] (Supplementary Table 11). A common feature of these internal SVs appears to be distinct repeat substructures within the putative inserted SV. In one example, the canonical reference repeat of AGG is interrupted by three distinct (but related) repeat subpatterns (Fig. 2d).

Spanning long reads also elucidate complex rearrangements typically missed by conventional NGS in which multiple events are located together. The assembly-based approach identified 4.2% of events as complex, and whereas Delly short-read predictions were largely confirmed, substantially greater complexity was observed in the variants (Table 2), with 3.4% showing added complexity. Inversions seem to be particularly enriched for complexity (55%; Supplementary Table 6), a feature we are exploring further in the context of a large population cohort (T.R., M.H.-Y.F., A.M.S., A.B. and J.O.K., unpublished data). We find predicted inversions located together with insertions (Fig. 4a), deletions (Fig. 4a,b) and duplications (Fig. 4c). A number of inversions also showed overlapping boundaries (for example, inverted repeat structure at the inversion boundaries), making it challenging to resolve precise breakpoints (Supplementary Table 6). Another arrangement frequently observed in the data is the insertion or deletion of proximally located sequences, which we refer to as proximal duplicated or deleted substrings. These appear in both forward (Fig. 4d) and inverted (Fig. 4e) orientation, and highly similar substrings can excise and insert multiple times within the same genomic interval (Fig. 4e). These events are particularly challenging to detect using short reads and are often mischaracterized as tandem duplications or inversions (Table 2).

DISCUSSION

Our analysis of the NA12878 genome shows that combining complementary technologies yields results that are superior to those from any single technology. Long contigs from SMRT sequencing facilitate unambiguous mapping to genome maps; the 800-kb N50 (far longer than those observed in standard short-read approaches) and absence of fragile sites for our sequence contigs also make it very likely for a contig to bridge multiple genome maps. This leads to scaffolds that are far more contiguous than sequence contigs or genome maps alone. Analogously, although long reads elucidate SVs far better than short reads and provide breakpoint-level precision, some events (Fig. 3) contain repeat lengths that only genome maps can accurately resolve. Last, the high accuracy and depth of low-cost short-read data provide reliable SNV and indel calls that increase overall accuracy and improve phasing precision of long reads. Together, these technologies allowed us to resolve long-standing assembly discrepancies.

Additional improvements are needed to extend the impact of our assembly approach. The high cost and run time of long- read sequencing (Supplementary Results and Supplementary Table 12) are the most obvious concerns, but yields continue to rise, and recent algorithmic developments for overlapping long-read data (the most time-consuming step in assembly) have reduced run times substantially[37,38]. Additionally, our current assembly approach is not fully integrated; sequence contigs and genome maps are separately assembled before scaffolding. Integrated methods such as AGORA have been restricted to genomes with single complete genome maps or to simple bacterial genomes but could potentially lead to better anchoring of sequence contigs within scaffolds, better N50 values[39] and better haplotype resolution (by extending existing string graph algorithms, which have the potential to directly reconstruct *de novo* haplotypes from sequencing data)[40]. Such approaches could obviate mapping-based SV detection, especially in the context of large SVs. Simultaneous SV and haplotype resolution, along with integrating statistical phasing strategies, could yield phasing results on par with the >500-kb block length recently reported using statistically aided, long-read haplotyping in NA12878 (ref.41).

Even in the absence of divergent haplotypes, regions such as centromeres and large segmental duplications remain difficult to resolve and can lead to misassemblies. In some cases (Supplementary Fig. 9a), we cannot determine with absolute certainty whether a rearrangement or inversion has occurred owing to the high similarity of regions flanking the breakpoints, though previous studies have shown this region to be unstable[42]. Molecular maps have the potential to span regions of high similarity at great depth, as individual molecules can exceed 1 Mb in size. However, their nonrandom breakage can lead to systematic failures in detection. This limitation can be mitigated by creating multiple genome maps that use distinct recognition sequences (using high-quality sequence contigs to bridge across maps). Resolving repetitive regions is more than simply an issue of ‘completeness’; these regions have been shown to mediate large-scale rearrangements in the genome[43,44].

On smaller scales, we have shown that a major benefit of continuous long reads is the ability to directly observe structural variants. With the exception of deletions, most approaches for whole-genome sequencing structural variant analysis depend on either breakpoint analysis[45,46] or local realignment and reassembly[47,48], thus often inferring large events from indirect evidence. A much richer landscape of structural variation is observed when using direct evidence (Fig. 4). Reliance on breakpoint deconstructions often leads to incomplete or incorrect assignment of events, as we also observed in the context of inversions and mobile element insertions. Long-read approaches could be particularly useful and cost effective in validating mobile element insertions in repeat-dense areas using targeting strategies such as transposon-seq[49].

Given the large degree of variability between any two genomes, we are approaching a paradigm where short-read, reference-based approaches are no longer the sole gold standard for variant analysis, both for exome and genome sequencing[50]. This study provides a framework for integrating multiple platforms: high-quality short reads for SNVs and indels, long reads for structural variation, and long-read assembly and genome maps for large-scale genome rearrangements. By using a collection of technologies, we can finally begin to circumvent biases induced by overreliance on a single reference genome. As long-read technologies mature, fully *de novo* approaches will increasingly become a standard practice, and inference of variation will be replaced by a more direct, comprehensive characterization of genome variation that will in turn accelerate our understanding of the complex phenotypes such variations induce.

METHODS

Methods and any associated references are available in the online version of the paper.

ONLINE METHODS

BioNano data generation and analysis

High-molecular-weight DNA extraction, DNA labeling and data collection— NA12878 cells were washed with PBS, and the final cell pellet was resuspended in cell suspension buffer (CHEF Genomic DNA Plug Kit). Cells were embedded in a thin LMP agarose layer and lysed, protease treated and washed. Purified DNA embedded in a thin agarose layer was labeled following the IrysPrep Reagent Kit protocol (BioNano Genomics). Briefly, DNA was digested with Nt.BspQI nicking endonuclease (New England BioLabs) for 2 h at 37 °C. Nicked DNA was then incubated for 1 h at 50 °C with fluorescently labeled dUTP and Taq Polymerase (New England BioLabs). Taq ligase (New England BioLabs) was used in the presence of dNTPs for ligation of nicks. Recovered DNA was counterstained with YOYO-1 (Life Technologies).

Labeled and counterstained DNA samples were loaded into IrysChips (BioNano Genomics) and run on the Irys (BioNano Genomics) imaging instrument. Data were collected for each sample until >50-fold coverage of long molecules (>180 kb) was achieved. The IrysView (BioNano Genomics) software package was used to detect individual linearized DNA molecules using the YOYO-1 counterstain and to determine the localization of labeled nick sites along each DNA molecule. Sets of single-molecule maps for each sample were then used to build a full genome assembly.

De novo assembly of genome maps—*De novo* assembly of single molecules was accomplished using a custom BioNano assembler software program based on an Overlap-Layout-Consensus paradigm[51–53]. First, we started with pairwise comparison of all molecules longer than 180 kb and nine labels to find all overlaps with $P < 1 \times 10^{-10}$, then we constructed a draft consensus map on the basis of these overlaps. The draft map was further refined by mapping single molecules to it and recalculating the label positions. Next the consensus maps were extended by aligning overhanging molecules to the consensus maps and calculating a consensus in the extended regions. Finally, the consensus maps were compared and merged where patterns matched with $P < 10^{-15}$. The process of extension and merge was repeated five times before a final refinement was applied to 'finish' all genome maps. The result of this assembly is a genome map set entirely independent of any known reference or external data.

PacBio data generation and analysis

NA12878 genomic DNA library preparation was prepared using high-molecular-weight DNA (20–50 kb) extracted from the Coriell control sample, and sequencing was performed using a modified method that was primarily based on the manufacturer's instructions and reflects the XL-C2 and P5-C3 sequencing enzyme and chemistries. Detailed description of sequencing and summary of results can be found in the Supplementary Result.

Error correction and assembly—Error correction of all reads was performed using Falcon, following the general principles proposed in ref.15 (Supplementary Note 1). In short, all long reads greater than 3 kb were first aligned to one another using Blasr[54]. These reads were then grouped together by selecting the top alignments (using a coverage cutoff of 40). A consensus was formed for each read; the resulting read was trimmed at the ends to eliminate potential chimeras and low-quality sequence (here we require at least 5× coverage of a given base). Error-corrected reads were passed to the Celera Assembler to form contigs (Supplementary Note 1). The resulting raw contigs were passed back to the Quiver pipeline (SMRTAnalysis v.2.2.0) on the subset of raw reads corresponding to the newer chemistry to provide the final, high-quality sequences. These sequences were then merged with genome mapping scaffolds to produce our initial V1 assemblies. For scaffolding purposes, an alternative long-read assembly was generated using a modified form of the Falcon pipeline, which yielded a more aggressive assembly with a 2.1 Mb N50 (Supplementary Fig. 12 and Supplementary Table 1). Final assemblies were cleaned to remove contaminants (Supplementary Note 1).

Short-read mapping and variant calling and correction—Short reads from ref.1 were mapped using BWA-MEM[55] (version 0.7.12-r1039) with default parameters. Variants were called using Freebayes[56] (version 0.9.18-3-gb72a21b) with default parameters but were filtered for variants with Q50 or higher.

PacBio structural variation—Events were called using the methodology from ref.26, PBHoney[57] and a custom pipeline. For comparison to the haploid CHM1 assembled breakpoint data set more directly, only events spanned by the *de novo* assembly were considered (though some alternative haplotypes persisted in the assembly). A brief description of the custom pipeline follows (see Supplementary Note 2 for detailed overview). Reads were first aligned to the reference using Blasr via an iterative process. First, the full read was mapped, maintaining the top ten highest-scoring alignments relative to the reference. Next, unmapped portions of each read were extracted from the input read set and remapped to the reference to identify potential highly divergent rearrangements that were missed in the initial mapping step. The top alignment for each query was then passed into a three-state HMM to identify potential insertions and deletions contained within a single long-read alignment. The HMM is needed because of a lack of affine gap alignment[58] in the initial Blasr results (1.3.1) and to reduce false positive calls in high-degeneracy raw PacBio reads, both of which lead to sporadic alignment of query bases within a true deletion region (or reference bases within a true insertion region). Note that using an updated version of Blasr with affine gap alignment (Supplementary Note 2) resulted in improved specificity at improved, or similar sensitivities, for most read types and event categories (Supplementary Table 5).

For more complex or larger rearrangements, a secondary step was performed in which a directed alignment graph is created. Alignments (nodes) were ordered relative to their position on the query; a directed edge was drawn between alignments *a* and *b* if the end of alignment *a* preceded the end of alignment *b*. A source, *s*, node and a sink, *k*, node are created for each query, and two edges, (*s*, *u*) and (*u*, *k*), are added to each alignment node (Supplementary Fig. 13). A simple dynamic programming algorithm determines the highest-scoring path from source to sink, where overlapping alignments are rescaled to eliminate double counting of overlapping intervals. This highest-scoring path is returned if it indicates a nonreference alignment path for the query. Although this step was not rate-limiting, it is shown in ref.59 that sparse dynamic programming approaches can yield O(*n* log *n*) runtimes, and these approaches have been applied in the context of detecting rearrangements[60,61]. The resulting set of individual read calls is then clustered by event type across the entire data set to yield the final set of predicted SVs. Both error-corrected and raw PacBio reads were processed via the same protocol (Supplementary Note 2).

Detection of structural variants with CLRs differs from paired reads in that the detection of SVs often implies complete resolution of the spanned event. However, given the potential for chimeras, and specifically the known issue of inverted tandem repeat–like chimeras due to missed adaptor sequence, singleton events are not sufficient to accurately resolve a sequence[13]. Therefore, to provide and validate predictions, we performed consensus calling across all putative events using partial order alignments of all ‘event’-containing reads[62]. For insertions, these consensus sequences were then scanned through Dfam HMMs to identify putative repeats using “nhmmscan” with default parameters[63].

Inversions were broken up into several distinct categories for custom analysis: (i) spanned inversions with a single contiguous alignment, (ii) spanned inversions with multiple subalignments and (iii) inversions in which only a single breakpoint is observed (Supplementary Fig. 14). Additionally, all inversion calls were passed through an additional step that used custom dot-plotting script followed by manual analysis from spanning reads to reduce false positives.

Genome map structural variation—The structural variation algorithm begins by aligning genome maps to the reference (hg19). The alignment algorithm uses a Smith-Waterman style dynamic programming algorithm where the units of comparison are distance intervals between detected label sites (as opposed to base pairs). Intervals are compared using maximum likelihood and a noise model designed for BioNano data. Both orientations are aligned separately for each map, and the best scoring alignment for each genome map, over the entire reference, is recorded. The algorithm scores each interval: positive scores are given when the interval in the reference and genome maps agree according to the model, and negative scores are given when they do not agree. If there are outliers in the alignment, or if one side of the genome map does not align (i.e., all intervals have a negative score), the map is split at the outlier positions. Here, we set the outlier cutoff to be 10^{-6}, which represents the probability of the interval being similar by random chance. The split subintervals are then realigned to the entire reference (with each piece again permitted to align in either orientation). In the case of a large insertion or deletion (>3 kb), the split portions of the map on either side of the SV will align next to one other. In the case of inversions, the split portions will again align next to each other, but in opposite orientations.

Tandem repeats—The tandem repeat detection pipeline uses PacBio long reads, alignment of reads to hg19 (via Blasr) and a reference TR table (available from University of California–Santa Cruz) as input. Only the top scoring alignment was used for each read, and only reads which had at least 100 bp of sequence anchoring upstream and downstream of the TR were considered for analysis. The method is based on the work in ref.36. A summary proceeds as follows. First, a three-stage dynamic programming algorithm step more robustly identifies the TR region in the long read and the putative boundaries. The TR region is then passed into the pairHMM-based method to provide a more robust and probabilistic estimate of TR multiplicity using the appropriate error profile. The processed pairHMM output is used for clustering. The objective of the clustering routine is to call the allele based on the estimated number of TRs for all the reads that span a particular TR event. We keep track of several key features on the clustering, the binomial probability of the clustering split, the minimum number of reads in each TR allele, the total number of reads given as input (to distinguish potential copy number abnormalities that could lead to spurious calls) and the c-separation, which specifies the separation between the means of two clusters. We return this information along with the cluster means and s.d. for each cluster. The sequences identified within each cluster of a TR event are used as an input for the partial order alignment (POA) consensus generation step. SVs internal to TRs are obtained by traversing the raw pairHMM output to find intervals in the query that are of low quality relative to the consensus TR element. TRs were filtered to exclude those that are segmental duplications as well as those that are contained within another repeat (if two TRs overlap but are not fully contained, we excluded the region). In a situation where two alternative TRs were present and both were completely contained within one another, we selected the TR with larger period.

Phasing—SNVs and indels identified by high-depth Illumina sequencing of the NA12878 trio[24] were used as a starting point for phasing (Broad Institute, GATK 2.5; Haplotype Caller version 2.7). Long reads were phased relative to hg19 using HapCut[64] version 0.7, which implements a graph-based optimization heuristic and has been previously applied to phase HuRef. PacBio reads have a well-known reference bias in which SNVs are more likely to be called as the reference owing to alignment artifacts created by the high insertion and deletion error rates of raw reads[65]. To mitigate this process, we first created a ‘variant-free' version of hg19 in which all known variant bases in the reference were converted to 'Ns'. Reads were then mapped to this and assessed for consistency with short-read trio predicted variants in NA12878. Short reads were also mapped to assembled PacBio contigs, and heterozygous positions were phased using the same approach (after correction of high-quality homozygous variants). Variants were included in phasing if they were covered by at least ten PacBio reads with at least 25% of reads supporting both alleles and 20 Illumina reads with at least 25% of reads supporting both alleles.

To phase tandem repeats and structural variants, we extracted reads spanning the event of interest. Reads were separated into two sets on the basis of which allele they supported (in the case of ambiguous alignment, the read was discarded from analysis). SNVs immediately upstream or downstream of a putative event were used to assess phase by performing variant free alignment, as described above. For each allele, a consensus SNV was called at each position, and a label (maternal or paternal) was placed on the allele on the basis of aforementioned trio calls. In many cases, insufficient SNVs were available in the flanking region, and these regions were listed as ambiguous and eliminated from haplotype consistency analysis.

We attempted to phase each of the predicted tandem repeats and structural variants by placing them in the context of the previous trio-based phasing. Each tandem repeat and structural variant was evaluated to produce 'high-confidence' heterozygous calls. The set of reads corresponding to each allele of a high-quality call was retrieved and used to locally determine the haplotype. In short, each read set for an allele was evaluated at all known SNV or indel locations proximal to the SV, where at least two reads covered the event using a POA alignment. As before, the reference SNV or indel position was eliminated from the reference sequence to eliminate reference bias. A consensus haplotype was then established for each allele; if the consensus haplotypes were consistent with trio-based phasing, we then assigned each allele to its corresponding haplotype.

Hybrid scaffolding

The scaffolding pipeline takes two input files, a sequence contig map file and a genome map file. Here, the sequence contig map file was generated by running an '*in silico* digest' on the PacBio contigs. There are two steps in the scaffolding process. In the first step, which used BioNano's alignment tool RefAligner[51,52], the sequence maps were compared to BioNano genome maps to find their best matches. Only those sequence maps with more than seven labels were used for comparison. Those sequence-genome map pairs with for which three consecutive sequence labels did not agree with genome maps were flagged and were not used in the next step. These pairs can potentially be chimerical assemblies, haplotype discrepancies or mismatches. During the second step, filtered sequence maps and genome maps were merged with RefAligner using a P value of 1×10^{-10} to create hybrid scaffold maps. The merge process was performed in a recursive pairwise manner. The pairs between sequence maps and genome maps were ranked on the basis of their similarity and were merged in order. The process was repeated until all pairs were merged, and the results became the first version of our hybrid scaffold maps (V1). We then re-ran the hybrid scaffolding pipeline to further merge the V1 scaffolds with additional sequence contigs and generated our V2 hybrid scaffold maps. Finally, to anchor the original sequences and generate FASTA sequences, we realigned the sequence maps with the V2 hybrid scaffolds using custom scripts (Supplementary Note 3), and any V2 scaffolds formed solely from Falcon overlaps (5) that did not have Celera-assembled contigs mapping support were eliminated (3). Alignment of sequence contigs and genome maps to the V2 scaffold is provided in Supplementary Tables 13 and 14. The resulting V2 scaffolds were aligned to hg19 in both nick and sequence space (Supplementary Fig. 15).

Delly short-read SV calls

The Illumina Platinum Genomes (http://www.illumina.com/platinumgenomes/) were used to discover SVs in NA12878 using short-read sequencing data. The SV prediction software Delly was run on two independent multi-sample data sets, both of which included NA12878. The first data set was the 17-member CEPH pedigree sequenced to 50× depth using a standard paired-end, short-insert library. The second data set was a family trio sequenced to >30× depth using a long-insert mate-pair library. We used the multi-sample Delly version to subsequently filter SVs on the basis of the genomic site itself and the available genotype information. For the long insert trio, we required a minimum SV size of 1 kb, at least three supporting paired ends for the SV site and a median paired-end Phred-scaled mapping quality >20. In addition, for copy-number variable events we required that at least one sample in the pedigree trio was a noncarrier with increased or decreased read depth for deletions or duplications, respectively. This filter was added to exclude false positive SV predictions due to repeat-induced mismappings, reference assembly errors or incomplete reference sequences. For balanced inversions, we could not apply a read-depth filter, and thus inspected SV predictions with low paired-end support manually. For the short-insert SV predictions, we used the same filtering approach except that we lowered the minimal required SV size to 250 bp owing to the decreased mean and s.d. of the fragment size distribution compared to the long-insert library.

For all deletion regions in which the Illumina prediction did not symmetrically overlap with a PacBio predicted event by at least 80%, the events were manually evaluated by visual examination of dot plots for all reads either spanning the event or with alignments disrupted within 5 kb of the event. In cases where a single event existed within this 5-kb interval, we considered the event ‘validated’ even if the predicted Illumina boundaries were inconsistent with the predicted boundaries from PacBio. All duplications and inversions were manually validated given the heightened complexity observed in these event types.

PCR validation

A custom primer design pipeline (A.M.S., T.R. and J.O.K., unpublished data) using the Primer3 algorithm[66], and BLAST[67] was used to design specific PCR primers for the different SV types. The search for specific primers was repeated iteratively until a maximum product size of 6,000 bp, after which the locus was excluded from validation. For both deletions and insertions, a pair of primers was placed outside flanking the predicted SV. PCR yields bands at the size expected on the basis of the reference genome: that is, smaller or larger than what is expected for deletions and insertions, respectively. The band pattern allows distinguishing between 0/0, 0/1 and 1/1 genotypes.

PCR primers were obtained from Sigma. PCR reactions were performed using 10 ng of genomic NA12878 DNA (Coriell) in 20-μl volumes using the Sequalprep Long PCR reagents (Life Technologies) in a 96-well plate using the DNA Engine Tetrade 2 thermocycler (Bio-Rad). PCR conditions were: (i) 94 °C for 3 min; (ii) ten cycles at 94 °C for 10 s, 62 °C for 30 s and 68 °C for 8 min and 25 cycles of 94 °C for 10 s, 60 °C for 30s and 68 °C for 10 min; and (iii) a final cycle of 72 °C for 5 min. PCR products were analyzed on a 0.8% agarose gel stained with Sybr Safe Dye (Life Technologies) and a 100-bp ladder and 1-kb ladder (NEB). If necessary, gel bands were cut with a scalpel, gel extracted with the Nucleospin Gel and PCR Cleanup kit (Macherey-Nagel) and sent for capillary sequencing (GATC Biotech AG).

Software availability

The corresponding software built and used (as described above) can be found at https://bitbucket.org/znfinger/na12878_architecture and is also provided as Supplementary Software. For external software, we have provided links to the build used when available. Otherwise, we have included this information within the project README. Parameter files for the error-correction and assembly specs are provided in the “param_file” subdirectory of the repository.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by institutional support from the Icahn Institute for Genomics and Multiscale Biology, R01 HG005946, U01 HL107388, R01 DK098242-01, R01 MH106531, US National Institutes of Health (NIH) U41HG007497, the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, the STARR Consortium, the WorldQuant Foundation, the Pershing Square Foundation, the Genomics & Epigenomics Core Facilities and SMRT Sequencing Center at Weill Cornell Medical College, and through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. DNA samples were provided by the Coriell Institute for Medical Research and the US National Institute of Standards and Technology (NIST). We would also like to thank T. Zichner for assistance with the design of validations and M. Chaisson for assistance with running Blasr, the assembly-based SV pipeline, and in performing the CHM1 comparison.

References

1. Zook JM, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014; 32:246–251. [PubMed: 24531798]
2. Lam HYK, et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 2012; 30:78–82. [PubMed: 22178993]
3. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
4. Istrail S, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci USA. 2004; 101:1916–1921. [PubMed: 14769938]
5. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA. 2011; 108:1513–1518. [PubMed: 21187386]
6. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

7. Human Genome Sequencing Consortium International. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. [PubMed: 15496913]

8. Pang AWC, Macdonald JR, Yuen RKC, Hayes VM, Scherer SW. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. G3 (Bethesda). 2014; 4:63–65. [PubMed: 24192839]

9. Schadt EE, Turner S, Kasarskis A. A window into third generation sequencing. Hum Mol Genet. 2010; 19:R227–R240. [PubMed: 20858600]

10. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

11. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. [PubMed: 21293372]

12. Ross MG, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013; 14:R51. [PubMed: 23718773]

13. Rasko DA, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011; 365:709–717. [PubMed: 21793740]

14. Bashir A, et al. A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol. 2012; 30:701–707. [PubMed: 22750883]

15. Chin CS, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10:563–569. [PubMed: 23644548]

16. Ribeiro FJ, et al. Finished bacterial genomes from shotgun sequence data. Genome Res. 2012; 22:2270–2277. [PubMed: 22829535]

17. Koren S, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012; 30:693–700. [PubMed: 22750884]

18. Huddleston J, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014; 24:688–696. [PubMed: 24418700]

19. Patel A, Schwab R, Liu YT, Bafna V. Amplification and thrifty single-molecule sequencing of recurrent somatic structural variations. Genome Res. 2014; 24:318–328. [PubMed: 24307551]

20. Hastie AR, et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. PLoS ONE. 2013; 8:e55864. [PubMed: 23405223]

21. Lam ET, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012; 30:771–776. [PubMed: 22797562]

22. Salzberg SL, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012; 22:557–567. [PubMed: 22147368]

23. Maccallum I, et al. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol. 2009; 10:R103. [PubMed: 19796385]

24. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. [PubMed: 21811232]

25. Bansal V, Halpern AL, Axelrod N, Bafna V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. Genome Res. 2008; 18:1336–1346. [PubMed: 18676820]

26. Chaisson MJP, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015; 517:608–611. [PubMed: 25383537]

27. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28:i333–i339. [PubMed: 22962449]

28. Carter AB, et al. Genome-wide analysis of the human Alu Yb-lineage. Hum Genomics. 2004; 1:167–178. [PubMed: 15588477]

29. Myers JS, et al. A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet. 2002; 71:312–326. [PubMed: 12070800]

30. Mason CE, et al. Location analysis for the estrogen receptor-alpha reveals binding to diverse ERE sequences and widespread binding within repetitive DNA elements. Nucleic Acids Res. 2010; 38:2355–2368. [PubMed: 20047966]

31. Highnam G, et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res. 2013; 41:e32. [PubMed: 23090981]

32. Kamstrup PR. Lipoprotein(a) and ischemic heart disease–a causal association? A review. Atherosclerosis. 2010; 211:15–23. [PubMed: 20106478]

33. Damert A, et al. 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. 2009; 19:1992–2008. [PubMed: 19652014]

34. Xing J, et al. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci USA. 2006; 103:17608–17613. [PubMed: 17101974]

35. Ejima Y, Yang L. *Trans* mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Hum Mol Genet. 2003; 12:1321–1328. [PubMed: 12761047]

36. Ummat A, Bashir A. Resolving complex tandem repeats with long reads. Bioinformatics. 2014; 30:3491–3498. [PubMed: 25028725]

37. Myers, G. Algorithms in Bioinformatics. Brown, D.; Morgenstern, B., editors. Springer; 2014. p. 52-67.

38. Berlin K, et al. Assembling large genomes with single-molecule sequencing and locality sensitive hashing. bioRxiv. 201410.1101/008003

39. Lin HC, et al. AGORA: Assembly Guided by Optical Restriction Alignment. BMC Bioinformatics. 2012; 13:189. [PubMed: 22856673]

40. Myers EW. The fragment assembly string graph. Bioinformatics. 2005; 21(suppl 2):ii79–ii85. [PubMed: 16204131]

41. Kuleshov V, et al. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol. 2014; 32:261–266. [PubMed: 24561555]

42. Antonacci F, et al. Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. Nat Genet. 2014; 46:1293–1302. [PubMed: 25326701]

43. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008; 1:4. [PubMed: 19014668]

44. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. Annu Rev Genomics Hum Genet. 2006; 7:407–442. [PubMed: 16780417]

45. Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. PLoS Comput Biol. 2008; 4:e1000051. [PubMed: 18404202]

46. Tuzun E, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005; 37:727–732. [PubMed: 15895083]

47. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]

48. Li S, et al. SOAPindel: Efficient identification of indels from short paired reads. Genome Res. 2013; 23:195–200. [PubMed: 22972939]

49. Iskow RC, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010; 141:1253–1261. [PubMed: 20603005]

50. Fuentes Fajardo KV, et al. Detecting false-positive signals in exome sequencing. Hum Mutat. 2012; 33:609–613. [PubMed: 22294350]

51. Nguyen JV. Genomic Mapping: A Statistical and Algorithmic Analysis of the Optical Mapping System. PhD thesis, Univ Southern California. 2010

52. Anantharaman, T.; Mishra, B. Algorithms Bioinformatics WABI. Gascuel, O.; Moret, BME., editors. Springer; 2001. p. 27-40.

53. Valouev A, Schwartz DC, Zhou S, Waterman MS. An algorithm for assembly of ordered restriction maps from single DNA molecules. Proc Natl Acad Sci USA. 2006; 103:15770–15775. [PubMed: 17043225]

54. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): theory and application. BMC Bioinformatics. 2012; 13:238. [PubMed: 22988817]

55. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

56. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. 2012. Preprint at http://arxiv.org/abs/1207.3907

57. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014; 15:180. [PubMed: 24915764]

58. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982; 162:705–708. [PubMed: 7166760]

59. Eppstein D, Galil Z, Giancarlo R, Italiano GF. Sparse dynamic programming I: linear cost functions. J ACM. 1992; 39:519–545.

60. Brudno M, et al. Glocal alignment: finding rearrangements during alignment. Bioinformatics. 2003; 19:i54–i62. [PubMed: 12855437]

61. Dubchak I, Poliakov A, Kislyuk A, Brudno M. Multiple whole-genome alignments without a reference organism. Genome Res. 2009; 19:682–689. [PubMed: 19176791]

62. Lee C. Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics. 2003; 19:999–1008. [PubMed: 12761063]

63. Wheeler TJ, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013; 41:D70–D82. [PubMed: 23203985]

64. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. 2008; 24:i153–i159. [PubMed: 18689818]

65. Carneiro MO, et al. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012; 13:375. [PubMed: 22863213]

66. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007; 23:1289–1291. [PubMed: 17379693]

67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

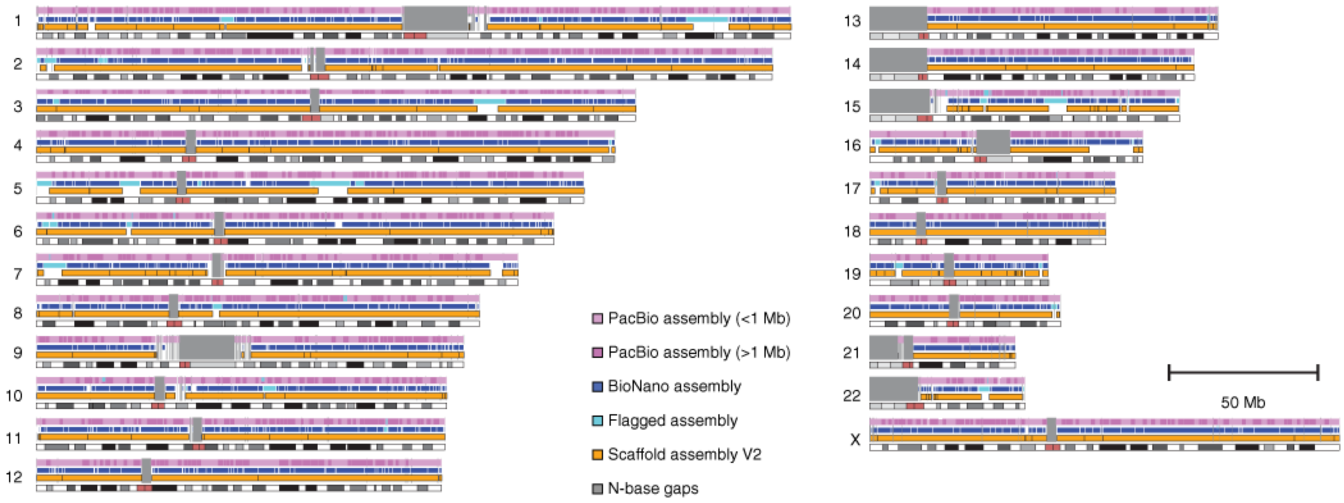

Figure 1.

De novo assembly and scaffold layout. PacBio sequence contigs. Genome maps and scaffold V2 are shown in order from the top of each chromosome, with the hg19 reference at the bottom. Possible chimeras identified by comparison of sequence contigs and genome maps (but not those that persist in the V2 scaffold) are indicated in cyan (flagged assembly). Ideogram and Giemsa banding for hg19 is plotted at the bottom of each chromosome in grayscale, with centromeres highlighted in light red. 'N' gaps in hg19 are shaded with gray in the background of all assemblies and scaffolds.

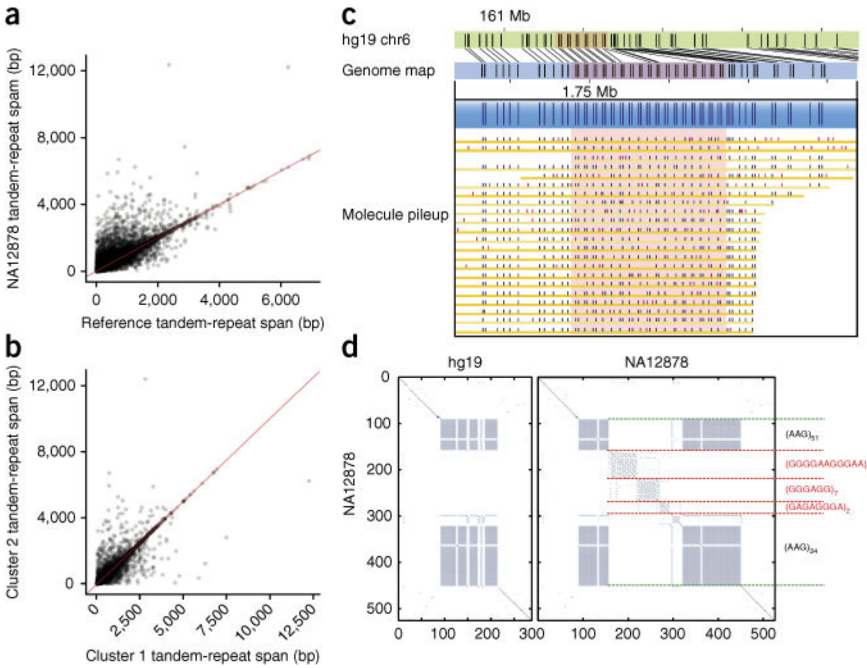

Figure 2.

Tandem-repeat detection from single molecules predicts a large divergence from reference. (**a**) Tandem-repeat span comparisons between predicted NA12878 alleles and hg19. (**b**) Length comparisons of each predicted heterozygous tandem-repeat locus in NA12878. (**c**) Copy-number difference at the *LPA* kringle domain (light red) between NA12878 (blue) and hg19 reference (green; chr6, chromosome 6). Spanning molecules (yellow) confirm that an expansion has occurred. In the molecule pileup view, dark blue represents mapped molecule labels, and red represents unmapped labels. Each tick on the scale represents a distance of 50 kb. (**d**) Left, a dot plot showing an expansion within a tandem repeat versus hg19. Right, a self-self dot plot of NA12878 indicates that the insertion contains repeated sequences that diverge from the original AAG repeat.

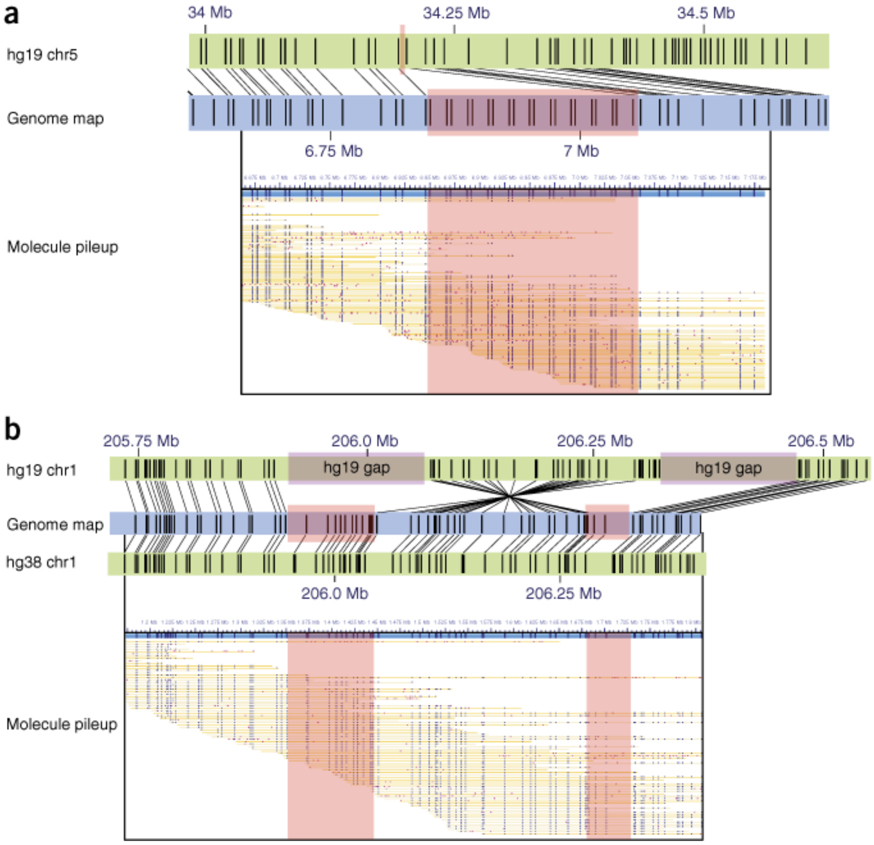

Figure 3.

De novo maps identify large structural variants. (**a**,**b**) Alignment of genome maps (blue) to *in silico* maps of hg19 (green) for a 206-kb insertion at 5p13.2 (**a**) and a 577-kb inversion at 1q32.1 (**b**). Below each event, all of the individual long molecules spanning the region of interest are shown to confirm homozygosity of the predicted event. The insertion locus in **a** and the boundaries of the predicted inversion in **b** are highlighted in light red. The predicted inversion (and resolution of gapped sequences) is consistent with the updated hg38 assembly.

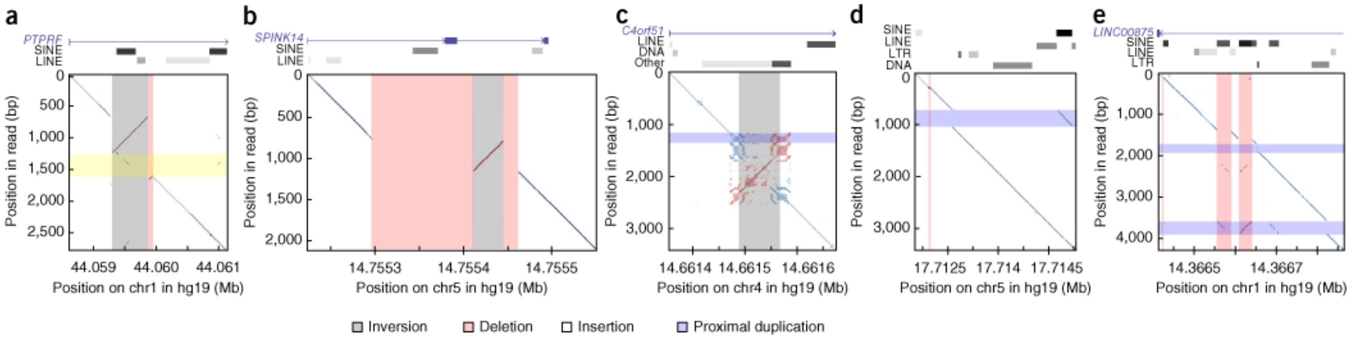

Figure 4.

CLRs highlight multiple colocated SVs and complex SV structures. Dot plots of a single error-corrected read (*y* axis) versus the corresponding reference regions (*x*-axis) for complex events in NA12878. Above each dot plot are gene annotations, known repeats (including short interspersed elements (SINE), long interspersed elements (LINEs), long terminal repeats (LTRs)) and other biologically relevant features. (**a**) Chromosome 1 (Chr1): 44058631–44061135, inversion with a trailing insertion and deletion (supported by 17/31 spanning raw reads). (**b**) Chr5:147552243–147555736, inversion with preceding and trailing deletion (20/34). The larger deletion eliminates an exon in *SPINK14*. (**c**) Chr4:146613545–146616773, inversion with potential duplication (6/11). (**d**) Chr5:17711870–17715038, proximally duplicated substring (10/26). (**e**) Chr1:143664130–143668633, a complex region with multiple events (9/34), including deletion of neighboring AluSG and AluU elements, expansion of a small tandem repeat and insertion of an AluY element at a nearby location.

Table 1

Assembly and scaffold summary statistics

	Scaffold NG50/N50	Contig NG50/N50	No. of contigs (scaffolds)	Mean contig (scaffold) length	Maximum contig (scaffold) length	Assembly size
Sequence assembly (Celera)	NA/NA	908 kb/906 kb	22,433	135 kb	6.5 Mb	3.04 Gb
Genome maps (BioNano)	4.5 Mb/4.6 Mb[a]	NA/NA	1,039	2.8 Mb	26.6 Mb	2.92 Gb
V1 scaffolds	12.2 Mb/13.6 Mb[a]	NA/NA	377	7.3 Mb	50.2 Mb	2.74 Gb
V2 scaffolds	28.8 Mb/31.1 Mb[a]	1.1Mb/1.4Mb[b]	202	13.5 Mb	81.4 Mb	2.76 Gb[a]

[a] N50/NG50 of the nick maps for the scaffold is generated in this step. NG50 is the length such that 50% of all base pairs of the known or estimated genome size are contained in scaffolds (or contigs) of the given length or longer.

[b] Corresponds to sequences derived from splitting scaffolds on Ns. NA, not available.

Table 2

Summary of SVs validated from Delly predictions

Delly call type	Illumina calls			PacBio calls								
	Illumina library with call			Deletions		Insertions				Unsupported		
	Long	Short	Both	Pdel	Simple	Pdup	Tandem duplication	Simple	Inversions	Reference only	Cov	Complex
Tandem duplication	22	19	7	0	1	2	29	0	0	5	7	4
Inversion	21	31	13	0	3	18	1	0	23	1	2	17
Deletion	22	691	24	9	701	1	0	2	0	5	11	8

Two separate paired-end Illumina libraries of different sizes were constructed and sequenced from NA12878 and were subjected to the Delly structural variation finder. Events that were confirmed by both the long and short libraries are enumerated under ‘Both’. PacBio Calls include the number of Delly predictions that were confirmed using manual examination of dot plots. The unsupported column includes any Delly events for which we found either no evidence for the indicated SV despite reasonable coverage (reference only) or for which we had insufficient coverage to reliably confirm the corresponding Delly call. Cov, coverage; PDel, proximal deleted substring; PDup, proximal duplicated substring.