High-resolution human genome structure by single-molecule analysis

Brian Teague[a], Michael S. Waterman[b], Steven Goldstein[a], Konstantinos Potamousis[a], Shiguo Zhou[a], Susan Reslewic[a], Deepayan Sarkar[c], Anton Valouev[b], Christopher Churas[a], Jeffrey M. Kidd[d], Scott Kohn[a], Rodney Runnheim[a], Casey Lamers[a], Dan Forrest[a], Michael A. Newton[c,e], Evan E. Eichler[d], Marijo Kent-First[f], Urvashi Surti[g], Miron Livny[h], and David C. Schwartz[a,1]

[a]The Laboratory for Molecular and Computational Genomics, Department of Chemistry, Laboratory of Genetics and Biotechnology Center, University of Wisconsin, 425 Henry Mall, Madison, WI 53706-1580; [b]Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089-2910; [c]Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706-1510; [d]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195-5065; [e]Department of Biostatistics and Medical Informatics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706-1510; [f]Department of Animal Science, Department of Biological Sciences, Mississippi State University, 130 Harned Hall, Lee Boulevard, Mississippi State, MS 39762-9698; [g]Department of Pathology, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15213-2536; and [h]Department of Computer Sciences, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706-1685

Edited* by David E. Housman, Massachusetts Institute of Technology, Cambridge, MA, and approved May 6, 2010 (received for review December 17, 2009)

Variation in genome structure is an important source of human genetic polymorphism: It affects a large proportion of the genome and has a variety of phenotypic consequences relevant to health and disease. In spite of this, human genome structure variation is incompletely characterized due to a lack of approaches for discovering a broad range of structural variants in a global, comprehensive fashion. We addressed this gap with Optical Mapping, a high-throughput, high-resolution single-molecule system for studying genome structure. We used Optical Mapping to create genome-wide restriction maps of a complete hydatidiform mole and three lymphoblast-derived cell lines, and we validated the approach by demonstrating a strong concordance with existing methods. We also describe thousands of new variants with sizes ranging from kb to Mb.

structural variation | copy number variation | optical mapping | single-molecule genomics | genome assembly

Recent reports (1–11) have firmly established genome structural variation as an important and pervasive source of genetic polymorphism. Since the initial reports (1, 2) of widespread copy-number variation between the genomes of phenotypically normal individuals, investigators have applied hybridization-based methods (3, 7, 9, 11), computational approaches (5, 6), clone paired-end sequencing (4, 10) and most recently a paired-end sequencing by synthesis approach (8) to the discovery and characterization of structural polymorphism. Others have described phenotypic consequences of these variants, including associations with myocardial infarction, neuroblastoma, autism, and schizophrenia (reviewed recently in ref. 12). Finally, their consistent association with segmental duplications and other classes of repeats (13) provides a mechanistic explanation for their origin (14) and points to a previously unappreciated role in evolution (15) as well as disease.

Unfortunately, despite all efforts, a comprehensive picture of genome structure polymorphism has not yet emerged. Current genome-wide studies of structural variation manifest only modest concordance, possibly due to ascertainment biases arising from the techniques employed. For example, hybridization-based methods (2, 3, 7, 9, 11, 16) are subject to nonspecific hybridization in repeat-rich regions, while clone-based strategies (4, 8, 10) are limited by maximum clone insert sizes and a wide clone size distribution relative to the events they are trying to detect. More recently, several entire human genomes were sequenced using high-throughput methods (17–20), but the difficulty of interrogating repeat-rich regions is compounded by these systems' short read lengths.

In an effort to overcome these challenges, we have applied Optical Mapping to the problem of discerning structural variation in normal human genomes. Optical Mapping (21–35) is a high-throughput system that combines single-molecule measurements with dedicated computational analysis to produce ordered restriction maps from individual molecules of genomic DNA: essentially, a single-molecule realization of traditional restriction fragment length polymorphism mapping (36). Each single-molecule restriction map is a direct measurement of the source genome, free from biases introduced by cloning, amplification, or hybridization. Recent advances in surface chemistry, microfluidics, instrumentation, and algorithms (*SI Text*) have increased our system's throughput so that Optical Mapping is now a viable platform for the analysis of complex eukaryotic genomes, including the human genome. This report presents the analysis of structural variation in four human genomes using Optical Mapping, compares these results to other genome-wide analyses and describes thousands of previously unreported structural variants.

Results

Optical Map Construction. We used Optical Mapping (21–35 and Fig. 1) to generate shotgun single-molecule restriction maps from the genomes of a complete hydatidiform mole (37) (CHM1h-TERT) and three lymphoblast-derived cell lines (GM15510, GM10860, GM18994). High molecular-weight genomic DNA was extracted from the cells with a gentle liquid lysis, then deposited on charged glass surfaces by an array of microfluidic capillary channels (26). The immobilized DNA molecules were digested in situ with the methylation-insensitive restriction endonuclease SwaI, chosen because its moderate average restriction fragment size balances good restriction map resolution with accurate fragment sizing. The digested DNA was stained with the fluorescent dye YOYO-1 and imaged on a laser-illuminated epifluorescence microscopy workstation built from off-the-shelf components. Custom machine-vision software analyzed the micrographs to identify the cleaved DNA fragments, estimate their sizes from their fluorescence, and order collinear fragments, producing ordered restriction maps from single molecules of DNA.

Author contributions: B.T., M.S.W., S.G., S.R., D.S., A.V., and D.C.S. designed research; B.T., K.P., S.R., C.L., and M.K.-F. performed research; B.T., M.S.W., S.G., K.P., D.S., A.V., C.C., J.M.K., S.K., R.R., D.F., M.A.N., E.E.E., M.K.-F., U.S., and M.L. contributed new reagents/analytic tools; B.T., M.S.W., S.G., K.P., S.Z., S.R., D.S., A.V., C.C., J.M.K., M.A.N., E.E.E., and D.C.S. analyzed data; and B.T., S.G., S.R., and D.C.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: dcschwartz@wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.0914638107/-/DCSupplemental.

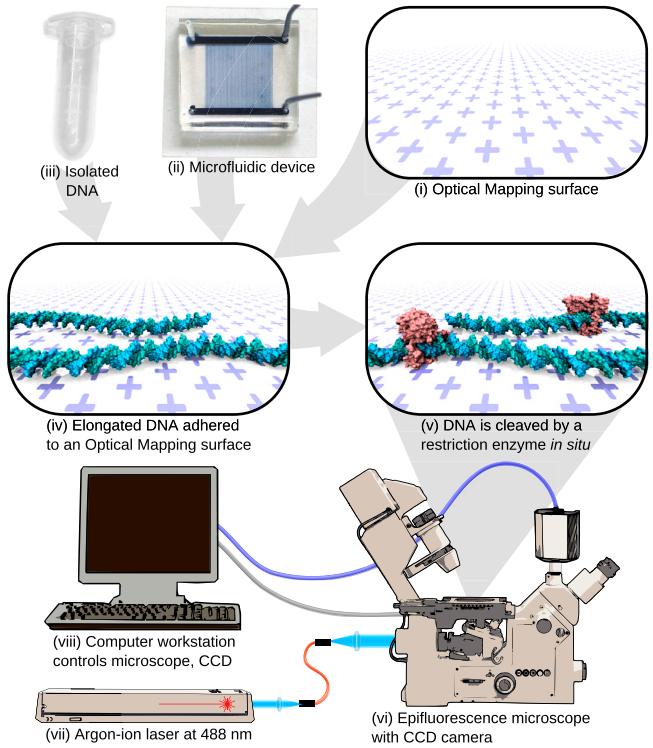

Fig. 1. An overview of the Optical Mapping platform. Bulk microscope cover glass is cleaned with a strong acid, then treated with a silane mixture to make positively charged Optical Mapping surfaces (*i*). A silicon wafer is patterned with standard photolithography techniques, and then replicated into a flexible PDMS microfluidic device (*ii*) using soft lithography. Finally, pure, high molecular-weight DNA (*iii*) is isolated from cultured eukaryotic cells using a gentle detergent-based lysis protocol. The microfluidic device is adhered to the Optical Mapping surface, and the DNA solution is pumped through the microchannels, wherein the DNA is elongated and attached to the Optical Mapping surface via electrostatic interaction (*iv*). The DNA is incubated with a restriction endonuclease (*v*), which cleaves the DNA at its cognate sites. The cleaved DNA is stained and imaged on an epifluorescence microscope (*vi*) illuminated by an argon-ion laser (*vii*) and controlled by a computer workstation (*viii*).

A tight integration between components is responsible for Optical Mapping's high throughput. The microfluidic device confines DNA deposition to a regular geometry, obviating manual microscopy and allowing a single microscope to run for 24 h unattended. Laser illumination and a sensitive CCD camera leverage YOYO-1's high quantum efficiency, reducing per-image exposure time from seconds to tens of milliseconds. Finally, depositing the genomic DNA with capillary flow orients all the molecules in the same direction, facilitating reliable machine vision. These synergies yield a throughput of 50,000–100,000 molecules analyzed every 24 h, allowing data collection for 50-fold coverage of a human genome to be completed on a single microscope in about a month.

These shotgun single-molecule restriction maps are assembled into genome-wide consensus restriction maps using an iterative process inspired by whole-genome sequence assembly (Fig. 2). Each iteration has two steps, clustering and assembly: The clustering step groups together similar single-molecule maps by aligning them to a reference map (28, 38), and then these clusters are assembled into a new hypothesis map using a Bayesian maximum-likelihood assembler (39). The first iteration uses a reference map derived *in silico* from the National Center for Biotechnology Information (NCBI) build 35 human reference sequence (40), but subsequent iterations of clustering and assembly extend and refine the hypothesis so that the final consensus maps are an accurate representation of the genome being analyzed. Parameters for both the alignment and assembly steps are tuned so

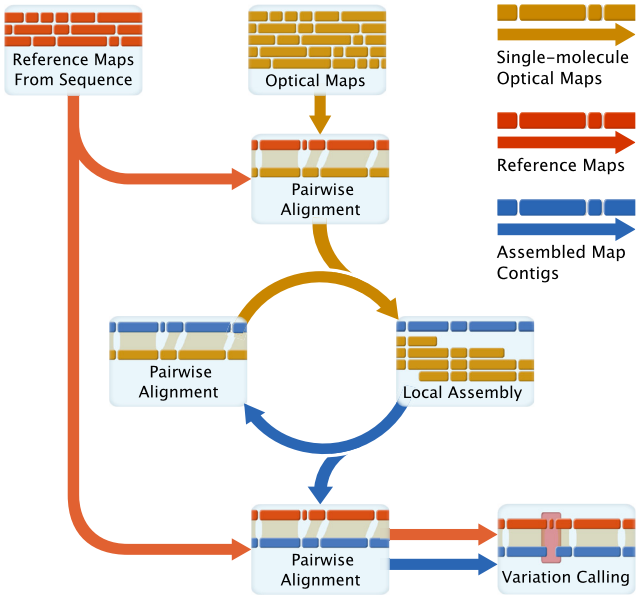

Fig. 2. An overview of the map assembly pipeline. Reference maps are generated *in silico* from the NCBI Build 35 human genome reference sequence (40), and used to seed an iterative process of pairwise alignment (which clusters together similar single-molecule maps) and local assembly (which generates a consensus optical map from a cluster of single-molecule maps). After several iterations of alignment and assembly, the consensus maps are aligned back to the reference map and analyzed for places where the consensus map differs significantly from the reference, indicating potential polymorphisms.

that only high-quality single-molecule maps are present in the final assembly. After eight iterations of assembly, the genome-wide consensus maps thus constructed span as much as 98.6% of the genome and have an average assembly depth of up to 58-fold (Table 1 and Table S1).

The genome-wide consensus maps are highly accurate: In all four genomes, over 95% of the fragments size 10 kb and greater are within 10% of their corresponding reference fragment size. (This regime's fragment sizing error increases with fragments <10 kb; see Fig. S1.) Another indicator of map accuracy comes from gaps in the NCBI build 35 reference sequence that were spanned by the optical maps and then bridged by a subsequent sequence assembly: Because of the iterative nature of the assembly pipeline, optical consensus maps are not confined to finished sequence and frequently span gaps in the reference sequence. Of the 279 gaps in the build 35 sequence, 183 are spanned by at least one assembly, and 156 have reliable size estimates. The optical map's gap size estimates are highly concordant with those that were bridged in the NCBI build 37 reference sequence (Table S2).

Table 1. Optical map collection and assembly statistics

	CHM	GM15510	GM10860	GM18994
Input optical maps	416,284	865,759	1,231,212	1,280,041
Input optical map coverage (fold)	65.91	139.15	214.18	220.82
Assembled optical maps	110,344	237,012	275,198	301,584
Assembled optical map coverage (fold)	18.95	41.85	53.24	57.68
Consensus maps	671	2,915	3,352	7,931
Average consensus map size (kb)	4,094	3,139	3,134	2,574
Sequence scaffold coverage (%)	96.29	97.36	98.62	98.29

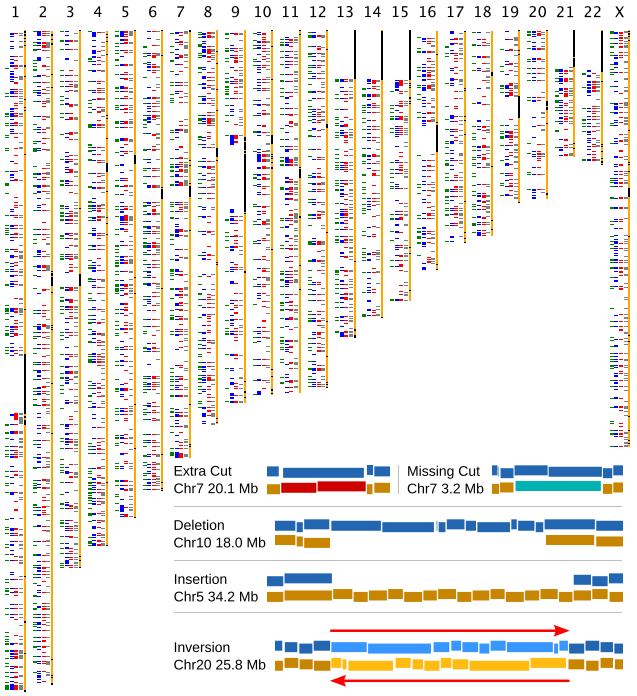

Fig. 3. A representation of the structural variation found in four genomes analyzed by Optical Mapping. Variants from the CHM genome are depicted in green; GM15510 in blue; GM10860 in red; and GM18994 in gray. The inset depicts five example differences from the genome of GM10860: an extra cut, a missing cut, a 250 kb deletion, a 150 kb insertion, and a 150 kb inversion.

Structural Variation Discernment. To identify sites of structural variation, we compared the consensus restriction maps to a restriction map generated *in silico* from the NCBI build 35 human genome reference sequence (40) (Fig. 3 and Table S3). Individual molecule maps are subject to a number of sources of random error, including missing restriction sites resulting from incomplete digestion, extra cuts from random breakage and nonspecific enzyme activity, sizing errors from random variation in dye incorporation, and the absence of small fragments that have desorbed from the Optical Mapping surface. However, the assembly of many single-molecule maps at each locus of a consensus map allows us to assign each difference a confidence score based on the probability that they arose solely due to Optical Mapping error; this allows us to reliably discern variation even in situations with high stochastic error, such as size variation in small fragments. Low-confidence ($p > 0.05$) variants were removed, and the remaining differences were filtered conservatively to remove several classes that, based on past experience, are less likely to represent bona fide sequence variants. A final manual curation step resolved regions that were not amenable to automated analysis. Details regarding these filtering and curation steps can be found in *SI Text*.

We summarize our findings in Table 2. The variants include simple events such as extra restriction sites, missing restriction sites, and insertions and deletions with size differences ranging from megabases down to 3 kb. They also include more complex events such as inversions and large discordant regions. In total, we discerned 4,205 unique variants in the four genomes we analyzed. We note that the smaller total from CHM is likely due to the reduced number of single-molecule restriction maps collected from the limited sample available, resulting in a loss of statistical power. We hypothesize, however, that this lower power is somewhat compensated for by the fact that a complete hydatidiform mole is effectively monoploid (37), eliminating the effect of diploidy on an assembler that wasn't designed to accommodate mixed haplotypes.

We also intersected the variants from each genome with variants of the same type from the other three genomes (Table 2). We note that over a third of the variants we report were observed in multiple genomes, giving us confidence that these results are due to polymorphism and not the spurious result of cell culture artifacts or other random processes. [The infrequency of culture-induced artifacts is also supported by analyses of the HapMap parent-progeny trios (9).] We suggest that the 322 variants common to all four genomes might be due to assembly errors in the NCBI build 35 reference sequence, or they might represent polymorphisms for which the reference sequence reports a minor frequency allele.

Comparison to Other Platforms. To validate the variants discerned by Optical Mapping, we carefully compared them to results reported by other investigators who used different technologies to analyze some of the same samples (Table 3). The reference platform's results were filtered to remove variants not amenable to detection by Optical Mapping (e.g., inversions that were contained entirely within a single SwaI restriction fragment), and the remaining variants were compared to the consensus map. Table 3 gives an overview of these comparisons, along with the intersections of other technologies' results; notes on each variant's comparison to the optical consensus map are included in Table S4, and a detailed example comparing several fosmid end-sequencing (FES) and paired-end mapping (PEM) variants to the corresponding optical map is presented in Fig. S2.

Because FES and PEM technologies have the ability to estimate insertion and deletion sizes independent of probe placement or density, we also compared the sizes of variants discerned with these technologies to the corresponding Optical Mapping variants. To increase the likelihood that the findings from each dataset represent the same sequence-level event, we only included Optical Mapping results that matched one-to-one with an FES- or PEM-derived observation. We were left with 84 pairs of observations for FES and 82 for PEM, several of which were discarded after manual curation (e.g., to remove several that overlapped gaps or were parts of large-scale discordances). A linear model fit to the remaining pairs has an R^2 of 0.95 and a slope of 0.97 for FES, and an R^2 of 0.94 and a slope of 0.98 for PEM, indicating strong agreement between these two methods and Optical Mapping (Fig. S3 and Fig. S4).

Table 2. Summary of structural variants discerned by Optical Mapping

	Variant type					Variant intersections				
	EC	MC	Ins	Del	Other	Unique	Int.1	Int.2	Int.3	Total
CHM	465	446	165	183	96	471	283	273	322	1355
GM15510	556	348	447	297	105	616	387	417	322	1753
GM10860	584	352	631	350	86	777	447	411	322	2003
GM18994	535	409	523	384	90	735	443	411	322	1941
Total	2140	1555	1766	1214	377	2599	780	504	322	

EC, extra cut; MC, missing cut; Ins, insertion; Del, deletion; Int.1, variant intersects with a variant from one other map; Int.2, intersects with 2 other maps; Int.3, intersects with all three other maps.

Table 3. Summary of OM results compared to other platforms

Query platform	Reference platform: Fosmid end sequencing	Paired-end mapping	Affymetrix SNP 6.0	Tiling array CGH	Optical Mapping
Fosmid end sequencing		92/196 (47%)	262/564 (46%)	262/564 (46%)	58/141 (41%)
Paired-end mapping	62/109 (57%)		146/163 (90%)	461/641 (72%)	114/473 (24%)
Affymetrix SNP 6.0	562/9527 (6%)	173/753 (23%)		17628/217344 (8%)	93/314 (30%)
Tiling array CGH	686/9527 (7%)	631/826 (76%)	17628/217344 (8%)		127/1599 (8%)
Optical Mapping	**108/206 (52%)**	**96/231 (42%)**	**33/54 (61%)**	**127/247 (51%)**	

A comparison of structural variant detection overlap between several technological platforms when applied to the same samples. Each cell shows the number of variants from the reference platform's results that were detected by the query platform. The reference platform's variants are first filtered to remove those that the query technology is not expected to be able to detect; for a full description of the filters used, consult *SI Text*. Fosmid end sequencing data from refs. 4 and 10; paired-end mapping data from ref. 8; Affymetrix CNV data from ref. 9; tiling array CGH data from ref. 11.

Optical Mapping Complements Other Platforms. As we were performing the comparison detailed above, we noted a number of common cases where Optical Mapping complements the results of another platform. A particularly striking example involves large gains in sequence discerned by hybridization-based platforms: such results can indicate additional copies of a sequence, but give no insight into the genome structure that engenders the gain in sequence. Optical Mapping's ability to resolve structural details can bring clarity to this situation, as exemplified by Fig. 4: The Affymetrix 6.0 SNP oligonucleotide microarray indicated a 290 kb gain in sequence on GM10860 chromosome 16, and the optical map identifies this event as an inverted tandem duplication.

End-sequencing strategies, on the other hand, are limited in their ability to resolve sequence insertions larger than their insert- or fragment-size, while Optical Mapping is subject to no such constraints. For example, FES analysis reported by Kidd et al. (10) demonstrated that clusters of fosmids with only one aligned end can indicate the presence of an insertion that was too large to be captured by the fosmid library. Of the eleven clusters identified by Kidd et al., eight have clear support in the optical map and a ninth comes from a region of large discordance between the optical map and the reference genome, making the presence of extra sequence likely (Table S4). (Several have since been spanned by sequence and closely agree with the optical map-derived estimate.) A detailed example, including micrographs of some of the DNA molecules that support this conclusion, is presented in Fig. 5. We also find evidence that fosmids with one aligned end that occur outside of clusters might indicate smaller insertions: An interval-intersection permutation test (see *SI Text* for details) reveals a significant intersection with optical map-discerned insertions ($p < 0.0001$).

Optical Mapping Reveals Variants Inaccessible to Other Platforms. We wanted to determine if Optical Mapping's unique properties quantitatively affect the variants it is able to discern. We focused on repeat-rich regions, because repeats are closely associated with structural variants (13, 14) but can hamper discernment efforts. We examined the performance of Optical Mapping and the two most current technologies, paired-end mapping (8) and tiling array comparative genome hybridization (CGH) (11), by classifying each variant as detected by Optical Mapping, detected by the alternate technology, or detected by both. We then compared the proportions of these classes from the entire genome with those subsets that intersect the 6 most common classes of repeat from the University of California Santa Cruz (UCSC) Genome Browser's RepeatMasker database (41) (Fig. 6*A*). While the proportion of Optical Mapping-discerned results compared to PEM is about the same in repeat-rich regions as in the entire genome, the repeat-intersecting proportion significantly increases when Optical Mapping is compared to the hybridization-based technology (χ^2 test, $p < 10^{-7}$). We interpret this as evidence that Optical Mapping has a similar power to discern variants in repeat-rich regions as PEM, but a greater capacity in this regard than tiling array CGH.

We also compared the distributions of insertion and deletion sizes between Optical Mapping, PEM, and tiling array CGH. (Fig. 6*B*). Optical Mapping is the only platform that does not evidence a strong bias toward the detection of deletions, perhaps due to its lack of reliance on a reference sequence either for probe selection or to anchor end-sequences.

Discussion

Pervasive natural variation in genome structure plays an increasingly acknowledged role in human health and evolution. The full

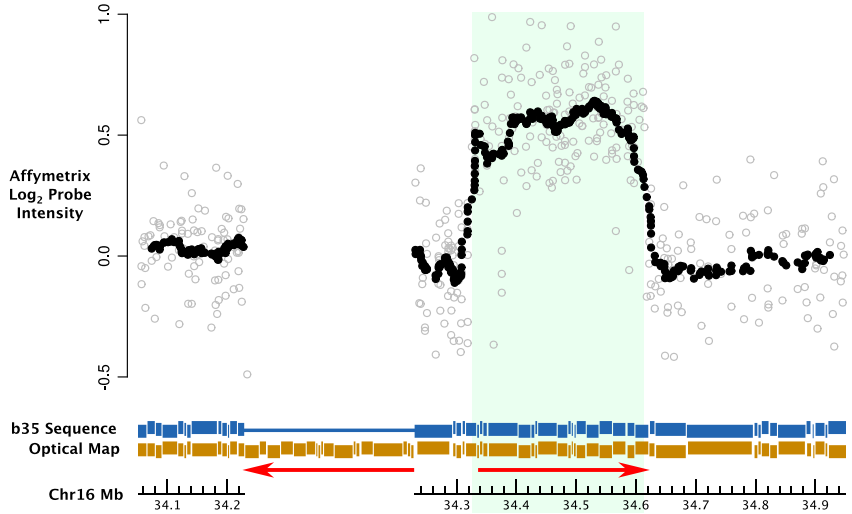

Fig. 4. The optical map complements hybridization-based approaches. The optical map reveals that the gain in sequence detected by the Affymetrix SNP 6.0 platform (shaded region) is due to an inverted tandem duplication at this locus (red arrows).

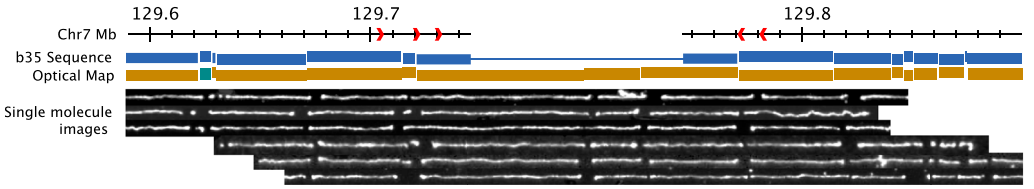

Fig. 5. A large insertion from GM15510 chromosome 7. Optical Mapping indicates a 90 kb insertion, confirming the large insertion that was indicated by a cluster of singleton fosmids reported by Kidd et al. (10) (red arrows). Included below the map is a montage of several of the single-molecule images that give evidence to support this insertion.

extent of this role, however, is obscured by the absence of an accurate, comprehensive, and unbiased method for analyzing a genome's structure. Current techniques are biased by the physical and biological principles on which they are based, limiting both the types of variants they can ascertain and the regions of the genome that are open to them.

To address these limitations, we have applied Optical Mapping to the discovery of structural variation in normal human genomes. Once limited to clones and prokaryotes, the technology has advanced to become an inexpensive, high-throughput platform for analyzing genome structure of complex eukaryotes including humans. Its scale of discernment ranges from kilobases to megabases, and it is not subject to ascertainment biases imposed by amplification, cloning, or hybridization.

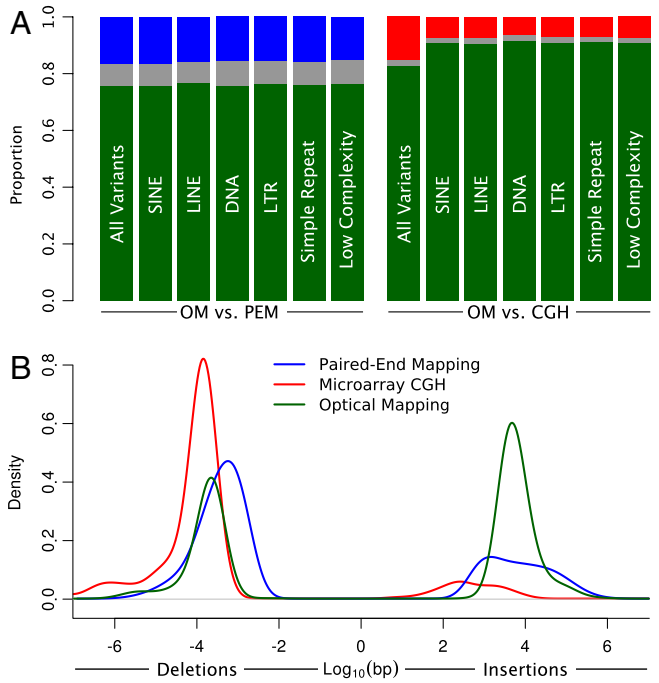

Fig. 6. (*A*) Optical Mapping has greater ability to discern variation in repeat-rich regions than hybridization-based technologies. The first bar in each section is a genome-wide representation of variants discerned only by Optical Mapping (green), only by an alternate technology (blue for PEM, red for CGH), or by both technologies (gray). (For example, in the first bar of the PEM comparison, 76% of the variants were found only by Optical Mapping, 17% were found only by PEM, and 7% were found by both technologies.) Subsequent bars represent the same proportions, but include only variants that intersect with various classes of repeat. The proportions are substantially the same when comparing Optical Mapping to PEM, but Optical Mapping detects a greater proportion of variants intersecting repeats when compared to hybridization-based technologies (χ^2 test, $p < 10^{-7}$). (*B*) Optical Mapping-discerned variants are more evenly distributed between insertions (median size, 4.5 kb) and deletions (median size, 4.3 kb). We compared the sizes of indels discovered with Optical Mapping to platforms based on end-sequencing and hybridization. Indel size density was estimated for each dataset using a Gaussian kernel with a bandwidth of 0.3. Negative sizes represent deletions, while positive sizes are insertions. The Optical Mapping indels are more evenly distributed between insertions and deletions, perhaps due to the platform's unique ability to detect large novel insertions.

The Optical Mapping results presented here confirm the prevalence of natural structural variation in the human genome. We present evidence for over 4,000 unique structural variants from four normal human genomes, with sizes ranging from several thousand to several million base pairs. We present the substantial overlap in the four sets of variants as evidence that the variations we detect are not random experimental error, but instead represent actual sequence-level differences between the analyzed genomes and the NCBI build 35 reference sequence. We support this assertion with discrete observations representing single molecules of DNA from the genomes under study. And we propose that the substantial number of unique variants discerned in just four individuals suggests many additional variants remain undiscovered in the human population as a whole.

We also show that these results confirm and complement the results of other technologies. We show a close concordance with both fosmid end sequencing (4, 10) and paired-end mapping (8), though the Optical Mapping results are not limited to the small insertions available to these mapping methods. The Optical Mapping results also bring structural insight to insertions and deletions discovered by hybridization-based methods, and are not limited to regions of the genome amenable to unique probe design. These advantages lead to a more balanced distribution of insertions and deletions, an indication of Optical Mapping's low systematic ascertainment bias and its ability to reveal structural variants inaccessible to other platforms. We also note Optical Mapping's ability to handle balanced events such as inversions and rearrangements, areas of genome structural variation that other high-throughput methods are just beginning to explore.

The Optical Mapping platform's freedom from dependence on sequence for de novo variant discovery comes at the price of lower resolution than sequence-based approaches: The endpoints of any individual event can only be resolved to the nearest restriction site. We are addressing this shortcoming by developing alternative enzymological methods that increase marker density and add sequence information to mapped molecules (42). We are anticipating these experimental advances by developing algorithms that take advantage of the additional information content to, for example, confidently separate multiple genotypes at a single genomic locus. These advances, combined with nanoconfinement techniques to dramatically increase analyte density (43, 44), promise the elucidation of complex sequence-level events such as the somatic rearrangements that are a hallmark of cancer genomes.

Materials and Methods

Sample Preparation. The complete hydatidiform mole cells used in this study were graciously provided by Urvashi Surti, director of the Pittsburgh Cytogenetics Laboratory, whose laboratory has a long-standing interest in hydatidiform moles. The cultured primary cells from the case CHM1 were immortalized using human telomerase reverse transcriptase (hTERT) to generate the CHM1hTERT cell line. The lymphoblast-derived cell lines were ordered from the Coriell Cell Repository (Coriell Institute) and cultured using standard eukaryotic cell culture techniques in RPMI medium 1640 supplemented with 2 mM L-glutamine and 15% FBS (Invitrogen). Genomic DNA was extracted using a liquid lysis followed by treatment with proteinase K (Bioline USA); details are available in *SI Text*.

Optical Mapping. Full experimental details regarding Optical Mapping protocols are available in *SI Text*. Briefly, Optical Mapping surfaces were prepared by acid-cleaning microscope cover glass, then treating it with a silane solution to impart a positive charge. A device comprising an array of microfluidic channels was fabricated using soft lithography and adhered to an Optical Mapping surface. A dilute DNA solution was pumped through the microchannels under parabolic flow conditions (26), causing the DNA to adhere to and stretch along the Optical Mapping surface via electrostatic interactions and flow-mediated forces. Thus presented and immobilized, the DNA was digested with the restriction endonuclease SwaI (New England Biolabs), then stained with YOYO-1 fluorescent dye (Invitrogen) and imaged on a Zeiss 135M inverted microscope (Carl Zeiss MicroImaging) at 63× magnification. The micrographs were analyzed by our automated machine-vision processing pipeline whose ultimate output was a set of ordered restriction maps of single DNA molecules (26, 31).

Genome-Wide Consensus Map Assembly. Full details of our assembly algorithm are available in *SI Text*. Briefly, genome-wide consensus map assembly is an iterative process wherein similar single-molecule maps are clustered by pairwise alignment to a hypothesis genome consensus map; these clusters are then assembled using a maximum-likelihood Bayesian assembler to generate a new hypothesis map. We began with a hypothesis consensus map generated *in silico* from the Build 35 human genome reference sequence (40), but the iterative nature of the assembler ensures that subsequent hypotheses are more and more representative of the genome under analysis. Empirically, eight iterations appear to be sufficient to generate an accurate, comprehensive consensus map of a mammalian genome.

Structural Variation Calling. After 8 rounds of iterative assembly, the consensus maps were aligned back to the Build 35 reference sequence to identify places where the two maps differ significantly. We discarded differences that were not statistically significant ($p > 0.05$) based on an appropriate statistical test of the underlying single-molecule map fragments. We also applied a set of empirically derived filters to account for other sources of error in the Optical Mapping process. For additional detail, see *SI Text*. A final manual curation step served to elucidate hard-to-automate variants such as large inversions.

ACKNOWLEDGMENTS. We thank Alex Lim for his preliminary data, and Jonathan Pritchard and Steven McCarroll for stimulating conversations and early access to data. This work was supported by National Institutes of Health Grants R01 HG000225 and R33 CA111933 (D.C.S.), National Research Service Award T32 GM008349 (S.R.), National Research Service Award T32 GM07215 (B.T.), and National Library of Medicine Training Grant 5T15 LM007359 (B.T.).

1. Lucito R, et al. (2003) Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res* 13:2291–2305.
2. Sebat J, et al. (2004) Large-scale copy number polymorphism in the human genome. *Science* 305:525–528.
3. Sharp AJ, et al. (2005) Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 77:78–88.
4. Tuzun E, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.
5. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38:75–81.
6. McCarroll SA, et al. (2006) Common deletion polymorphisms in the human genome. *Nat Genet* 38:86–92.
7. Redon R, et al. (2006) Global variation in copy number in the human genome. *Nature* 444:444–454.
8. Korbel JO, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426.
9. McCarroll SA, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40:1166–1174.
10. Kidd JM, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64.
11. Conrad DF, et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712.
12. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genom Hum G* 10:451–481.
13. Bailey JA, Eichler EE (2006) Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat Rev Genet* 7:552–564.
14. van Ommen GJB (2005) Frequency of new copy number variation in humans. *Nat Genet* 37:333–334.
15. Perry GH, et al. (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA* 103:8006–8011.
16. Shen F, et al. (2008) Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC Genet* 9:27.
17. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5:e254.
18. Wheeler DA, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–876.
19. Wang J, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456:60–65.
20. Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 27:847–852.
21. Schwartz DC, et al. (1993) Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* 262:110–114.
22. Jing J, et al. (1998) Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proc Natl Acad Sci USA* 95:8046–8051.
23. Lai Z, et al. (1999) A shotgun optical map of the entire Plasmodium falciparum genome. *Nat Genet* 23:309–313.
24. Lin J, et al. (1999) Whole-genome shotgun optical mapping of Deinococcus radiodurans. *Science* 285:1558–1562.
25. Zhou S, et al. (2003) Whole-genome shotgun optical mapping of Rhodobacter sphaeroides strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome Res* 13:2142–2151.
26. Dimalanta ET, et al. (2004) A microfluidic system for large DNA molecule arrays. *Anal Chem* 76:5293–5301.
27. Valouev A, Zhang Y, Schwartz DC, Waterman MS (2006) Refinement of optical map assemblies. *Bioinformatics* 22:1217–1224.
28. Valouev A, et al. (2006) Alignment of optical maps. *J Comput Biol* 13:442–462.
29. Valouev A, Schwartz DC, Zhou S, Waterman MS (2006) An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci USA* 103(43): 15770–15775.
30. Zody MC, et al. (2006) DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature* 440:1045–1049.
31. Zhou S, Herschleb J, Schwartz DC (2007) *New High Throughput Technologies for DNA Sequencing and Genomics*, ed KR Mitchelson (Elsevier, Amsterdam), pp 266–301.
32. Zhou S, et al. (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics* 8:278.
33. Ananiev GE, et al. (2008) Optical mapping discerns genome wide DNA methylation profiles. *BMC Mol Biol* 9:68.
34. Church DM, et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* 7:e1000112.
35. Zhou S, et al. (2009) A single molecule scaffold for the maize genome. *PLoS Genet* 5:e1000711.
36. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331.
37. Jacobs PA, Wilson CM, Sprenkle JA, Rosenshein NB, Migeon BR (1980) Mechanism of origin of complete hydatidiform moles. *Nature* 286:714–716.
38. Sarkar D (2006) On the analysis of optical mapping data. PhD thesis (University of Wisconsin, Madison, WI).
39. Anantharaman T, Mishra B, Schwartz DC (1999) Genomics via optical mapping. III: Contiging genomic DNA. *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, Menlo Park, CA), pp 18–27.
40. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.
41. Rhead B, et al. (2010) The UCSC genome browser database: Update 2010. *Nucleic Acids Res* 38:D613–D619.
42. Ramanathan A, et al. (2004) An integrative approach for the optical sequencing of single DNA molecules. *Anal Biochem* 330:227–241.
43. Jo K, et al. (2007) A single-molecule barcoding system using nanoslits for DNA analysis. *Proc Natl Acad Sci USA* 104:2673–2678.
44. Jo K, Schramm TM, Schwartz DC (2009) A single-molecule barcoding system using nanoslits for DNA analysis : Nanocoding. *Methods Mol Biol* 544:29–42.