High-resolution human genome structure by single-molecule analysis


*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; *Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089-2910; *Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706-1510; *Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195-5065; *Department of Biostatistics and Medical Informatics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706-1510; *Department of Animal Science, Department of Biological Sciences, Mississippi State University, 130 Harned Hall, Lee Boulevard, Mississippi State, MS 39762-9688; *Department of Pathology, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15213-2536; and *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.

Results

Optical Map Construction. We used Optical Mapping (21–35) 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 SswI, chosen because its moderate average restriction fragment length is compounds 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.


The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

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.
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 microscopv 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 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.

Table 1. Optical map collection and assembly statistics

<table>
<thead>
<tr>
<th></th>
<th>CHM</th>
<th>GM15510</th>
<th>GM10860</th>
<th>GM18994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input optical maps</td>
<td>416,284</td>
<td>865,759</td>
<td>1,231,212</td>
<td>1,280,041</td>
</tr>
<tr>
<td>Input optical map coverage (fold)</td>
<td>65.91</td>
<td>139.15</td>
<td>214.18</td>
<td>220.82</td>
</tr>
<tr>
<td>Assembled maps</td>
<td>110,344</td>
<td>237,012</td>
<td>275,198</td>
<td>301,584</td>
</tr>
<tr>
<td>Assembled optical map coverage (fold)</td>
<td>18.95</td>
<td>41.85</td>
<td>53.24</td>
<td>57.68</td>
</tr>
<tr>
<td>Consensus maps</td>
<td>671</td>
<td>2,915</td>
<td>3,352</td>
<td>7,931</td>
</tr>
<tr>
<td>Average consensus map size (kb)</td>
<td>4,094</td>
<td>3,139</td>
<td>3,134</td>
<td>2,574</td>
</tr>
<tr>
<td>Sequence scaffold coverage (%)</td>
<td>96.29</td>
<td>97.36</td>
<td>98.62</td>
<td>98.29</td>
</tr>
</tbody>
</table>
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 map generated from the NCBI build 35 reference sequence, or they might represent polymorphisms for which the reference sequence reports a minor frequency allele.

Table 2. Summary of structural variants discerned by Optical Mapping

<table>
<thead>
<tr>
<th>Variant type</th>
<th>CHM</th>
<th>GM15510</th>
<th>GM10860</th>
<th>GM18994</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>465</td>
<td>556</td>
<td>584</td>
<td>535</td>
<td>2140</td>
</tr>
<tr>
<td>MC</td>
<td>446</td>
<td>348</td>
<td>352</td>
<td>409</td>
<td>1555</td>
</tr>
<tr>
<td>Ins</td>
<td>165</td>
<td>447</td>
<td>631</td>
<td>523</td>
<td>1766</td>
</tr>
<tr>
<td>Del</td>
<td>183</td>
<td>297</td>
<td>350</td>
<td>384</td>
<td>1214</td>
</tr>
<tr>
<td>Other</td>
<td>96</td>
<td>105</td>
<td>86</td>
<td>90</td>
<td>377</td>
</tr>
<tr>
<td>Unique</td>
<td>471</td>
<td>616</td>
<td>777</td>
<td>735</td>
<td>2599</td>
</tr>
<tr>
<td>Int.1</td>
<td>283</td>
<td>387</td>
<td>447</td>
<td>443</td>
<td>780</td>
</tr>
<tr>
<td>Int.2</td>
<td>273</td>
<td>417</td>
<td>411</td>
<td>411</td>
<td>504</td>
</tr>
<tr>
<td>Int.3</td>
<td>322</td>
<td>322</td>
<td>322</td>
<td>322</td>
<td>322</td>
</tr>
<tr>
<td>Total</td>
<td>1355</td>
<td>1753</td>
<td>2003</td>
<td>1941</td>
<td>6299</td>
</tr>
</tbody>
</table>

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 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).
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 may indicate smaller insertions: An interval-intersection permutation test (see SI Text) 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. 6B). 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 ($\chi^2$ 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. 6B). 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

---

Table 3. Summary of OM results compared to other platforms

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

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.

---

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).
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.

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 somatic rearrangements that are 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 Cyto genetics 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 supplemented with 2 mM L-glutamine and 15% FBS (Invitrogen). Genomic DNA was extracted using a liquid lysis followed by treatment with protease 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 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 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 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 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 surface.

We began with a hypothesis consensus map generated using a maximum-likelihood Bayesian assembler to generate 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.).

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.