# Characterizing polymorphic inversions in human genomes by single-cell sequencing

Ashley D. Sanders,[1] Mark Hills,[1] David Porubský,[2] Victor Guryev,[2] Ester Falconer,[1] and Peter M. Lansdorp[1,2,3]

[1] Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, V5Z 1L3, Canada; [2] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, NL-9713 AV Groningen, The Netherlands; [3] Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

**Identifying genomic features that differ between individuals and cells can help uncover the functional variants that drive phenotypes and disease susceptibilities. For this, single-cell studies are paramount, as it becomes increasingly clear that the contribution of rare but functional cellular subpopulations is important for disease prognosis, management, and progression. Until now, studying these associations has been challenged by our inability to map structural rearrangements accurately and comprehensively. To overcome this, we coupled single-cell sequencing of DNA template strands (Strand-seq) with custom analysis software to rapidly discover, map, and genotype genomic rearrangements at high resolution. This allowed us to explore the distribution and frequency of inversions in a heterogeneous cell population, identify several polymorphic domains in complex regions of the genome, and locate rare alleles in the reference assembly. We then mapped the entire genomic complement of inversions within two unrelated individuals to characterize their distinct inversion profiles and built a nonredundant global reference of structural rearrangements in the human genome. The work described here provides a powerful new framework to study structural variation and genomic heterogeneity in single-cell samples, whether from individuals for population studies or tissue types for biomarker discovery.**

**[Supplemental material is available for this article.]**

Studies of human genome heterogeneity and plasticity aim to resolve how genomic features underlie phenotypes and disease susceptibilities. Identifying genomic features that differ between individuals and cells can help uncover the functional variants that drive specific biological outcomes. For this, single-cell studies are required to characterize the contribution of rare but functional cellular subpopulations that are important for disease prognosis, management, and progression (Biesecker and Spinner 2013; Macaulay and Voet 2014). Other than sequence variants, structural polymorphisms such as copy number variants (including insertions, deletions, and duplications) and copy-neutral genomic rearrangements (such as translocations and inversions) play major roles in human biology and health (Stankiewicz and Lupski 2010; Alkan et al. 2011). Indeed, polymorphic rearrangements are a common feature of the human genome (Pang et al. 2010) and are implicated in speciation (Feuk et al. 2005; Zody et al. 2008), population diversification (Stefansson et al. 2005; Alves et al. 2014), and many complex diseases, including neurological disorders and cancers (Antonarakis et al. 1995; Bondeson et al. 1995; Koolen et al. 2006; Shaw-Smith et al. 2006; Sharp et al. 2008; Tam et al. 2008; Antonacci et al. 2009; Salm et al. 2012). However, few human inversions have been studied comprehensively to date, and the phenotypic consequences and clinical relevance of most remain undefined (Feuk 2010; Alkan et al. 2011; Alves et al. 2012; Martinez-Fundichely et al. 2014).

Copy-neutral rearrangements, such as inversions, change the orientation of a segment of DNA without altering DNA content, making them difficult to map using currently available tools (Bansal et al. 2007; Wong et al. 2007; Alkan et al. 2011; Alves et al. 2012). Techniques such as karyotyping, fluorescence in situ hybridization (FISH), and optical mapping allow visualization and genotyping of inversions at the single-cell and single-chromosome level. However, the low resolution of these approaches limits their application to mapping large microscopic events that disrupt visible patterns (typically at the megabase-scale), and their low throughput limits the number of cells or individuals that can be studied at a time (Youings et al. 2004; Zody et al. 2008; Antonacci et al. 2009; Feuk 2010; Teague et al. 2010).

High-throughput sequencing (HTS) technologies enable discovery of submicroscopic inversions based on incongruous mapping of paired reads relative to the reference genome (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Pang et al. 2010; Sudmant et al. 2015). While improving throughput and genomic resolution, this approach is prone to false calls, because inversions are often flanked by repetitive DNA that interferes with unambiguous read mapping, and secondary techniques (such as PCR or extensive population-scale sequencing data) are often required to validate and genotype the predicted variant (Feuk et al. 2005; Turner et al. 2006; Antonacci et al. 2009; Pang et al. 2010; Alkan et al. 2011; Mills et al. 2011; Alves et al. 2012; Martinez-Fundichely et al. 2014; Sudmant et al. 2015). Moreover, the requirement of large amounts of DNA for standard HTS techniques prevents the analysis of inversions at the single-cell level, which is essential for exploring cellular heterogeneity, such as in the context of tumor evolution. Consequently, no reported

**Corresponding author: p.m.lansdorp@umcg.nl**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.201160.115.

technique currently enables the discovery and mapping of inversions at high throughput and high resolution, while simultaneously showing the genome-wide structural heterogeneity of single cells.

## Results

### Visualizing genomic rearrangements in single cells by DNA template strand sequencing (Strand-seq)

Strand-seq is a single-cell sequencing technique that identifies parental DNA template strands inherited by daughter cells after mitosis (Falconer et al. 2012). This method takes advantage of the directionality of single-stranded DNA molecules, which can be distinguished as either Crick (C; forward or plus strand) or Watson (W; reverse or minus strand) based on their 5′–3′ orientation (Fig. 1A, i). The thymidine analog 5-bromo-2′-deoxyuridine (BrdU) is incorporated during DNA replication (Fig. 1A, ii), and following mitosis, the BrdU-positive DNA strand is selectively ablated during genomic library construction, ensuring that only the BrdU-negative template strand is sequenced for each chromosome in each single cell. After library construction and HTS, resulting sequence reads are aligned to either the minus or the plus strand of the reference genome using the software package BAIT (Hills et al. 2013), and the DNA template strand inheritance patterns are determined for each chromosome within the cell (Fig. 1A, iii). Strand-seq library construction was automated to generate hundreds of libraries in a single experiment to study cellular heterogeneity at the single-cell level (see Methods).

With respect to the reference assembly, an inversion appears as a localized reorientation in the Watson-Crick state along the DNA strand of a chromosome (Fig. 1B). By sequencing only template strands, inversions are visualized as genomic regions where sequence reads of the inverted DNA segment map to the complementary DNA strand with respect to the surrounding sequence. To survey inversions in a normal human genome, we generated Strand-seq libraries from bone marrow (BM) cells of an adult male donor and filtered libraries based on read depth (>20 reads/Mb) and background (<5%) to ensure template strand states could be accurately assessed (see Methods). In any given Strand-seq library, structural rearrangements were evident as segmental changes in strand orientation along each chromosome (Fig. 1C, arrowheads). Inversions were identified as changes that recurred at the same genomic locations in multiple cells (Fig. 1C, red arrowheads), where a minimum of two cells sharing the inversion was required to distinguish the rearrangement from sporadic sister chromatid exchanges (Fig. 1C, black arrowheads; Supplemental Discussion). Each inversion was further genotyped based on whether one or both chromosomal homologs exhibited the localized strand reorientation (Fig. 1D; Supplemental Fig. S1). With respect to the reference assembly, a genomic locus can be 'homozygous reference' (neither parental homolog carries an inversion, and there is no change in template strand state) (Fig. 1D, left), or can contain a heterozygous rearrangement (a single homolog is inverted, and a WW or CC state switches to a mixed WC state) (Fig. 1D, middle), or a homozygous rearrangement (both homologs are inverted and the template strands completely switch from WW to CC or vice versa) (Fig. 1D, right). It should be noted that when a cell inherits a pair of homologous chromosomes in the WC state (for example, Chromosomes 3, 6, 10, 12, 15, and 21 in Fig. 1C), it is impossible to distinguish between homozygous reference and homozygous inversions (Supplemental Fig. S1), and therefore, WC chromosomes were excluded to accurately calculate allelic frequencies.

In the Strand-seq library shown, we identified 21 putative inversions in this single cell (Fig. 1C, red arrowheads), including a well-documented homozygous inversion on the p-arm of Chromosome 8 (Chr 8) (Hollox et al. 2008; Salm et al. 2012; Alves et al. 2014) and a heterozygous inversion on the q-arm of Chr 7, that are disease-linked (Fig. 1C, boxed regions; Osborne et al. 2001; Tam et al. 2008; Hobart et al. 2010). Strand-seq sequencing reads were BED-formatted and uploaded as custom annotation tracks onto the UCSC Genome Browser (GRCh37/hg19 assembly) (Fig. 1E, i; Kent et al. 2002) to zoom into putative inversions (Fig. 1E, ii). Locating the first read of the inverted region that is in the opposite orientation of the surrounding reads, we mapped the homozygous inversion on Chr 8p23.1 (a 5.21-Mb region at 7,224,122–12,417,334) and the heterozygous inversion on Chr 7q11.22-q11.23 (a 2.7-Mb region at 72,312,825–75,040,174) (Fig. 1E, ii, red bars). These coordinates coincide with previous reports of each inversion (Kidd et al. 2008; Antonacci et al. 2009; Martinez-Fundichely et al. 2014) (also see Fig. 2B), demonstrating how visualizing changes in strand orientation in chromosomes allows us to discover, map, and genotype inversions at high resolution in a single cell.

### Unbiased analysis of inversions using the custom software package, Invert.R

While characterizing genomic rearrangements on the UCSC Genome Browser in individual Strand-seq libraries affords high resolution, it is impractical to independently examine multiple cells for high-throughput studies. To comprehensively characterize inversions in Strand-seq libraries, we developed Invert.R, an R-based (R Core Team 2013) pipeline that systematically interrogates each single-cell library to localize and genotype putative inversions based on read alignment (for details, see Supplemental Methods). Briefly, Invert.R uses a read-based sliding window to calculate the ratio of W and C reads ('W/C ratio') across each chromosome, and plots these values as a histogram to visualize changes in template strand orientation (Fig. 2A). Putative inversions are flagged as genomic regions where the W/C ratio dips below and returns above a background threshold (Fig. 2A, arrows) and genotyped by calculating the overall change in W/C ratio (ΔW/C). A homozygous inversion is classified as a complete change in template state with respect to the surrounding chromosome (from entirely W reads to entirely C reads, or vice versa) and gives a ΔW/C ratio close to 1.0, whereas a heterozygous inversion is classified as a change in template state from entirely W or entirely C reads to a mixture of W and C reads and will give a ΔW/C ratio near 0.5. Invert.R also locates the nearest 5′ and 3′ flanking reads outside the inverted region to predict the 5′ and 3′ breakpoints, assigned as the first base pair position of these reads (Fig. 2A, asterisks).

To validate Invert.R and test the resolution of breakpoint mapping in individual and multiple Strand-seq libraries, we analyzed the previously characterized inversions on Chr 8p23 and Chr 7q11 (Fig. 2B). For the single cell shown in Figure 1C, Invert.R located and genotyped the homozygous Chr 8 inversion (ΔW/C ratio = 0.95) and the heterozygous Chr 7 inversion (ΔW/C ratio = 0.42) (Fig. 2B, i). Note that the ΔW/C ratios are near the expected values of 1.0 and 0.5, respectively, but are smaller due to genomic segments with low read-depth or reads mapping to both template strands, which often flank inversions (Fig. 2B, asterisks). Invert.R-calculated breakpoints closely corresponded to our

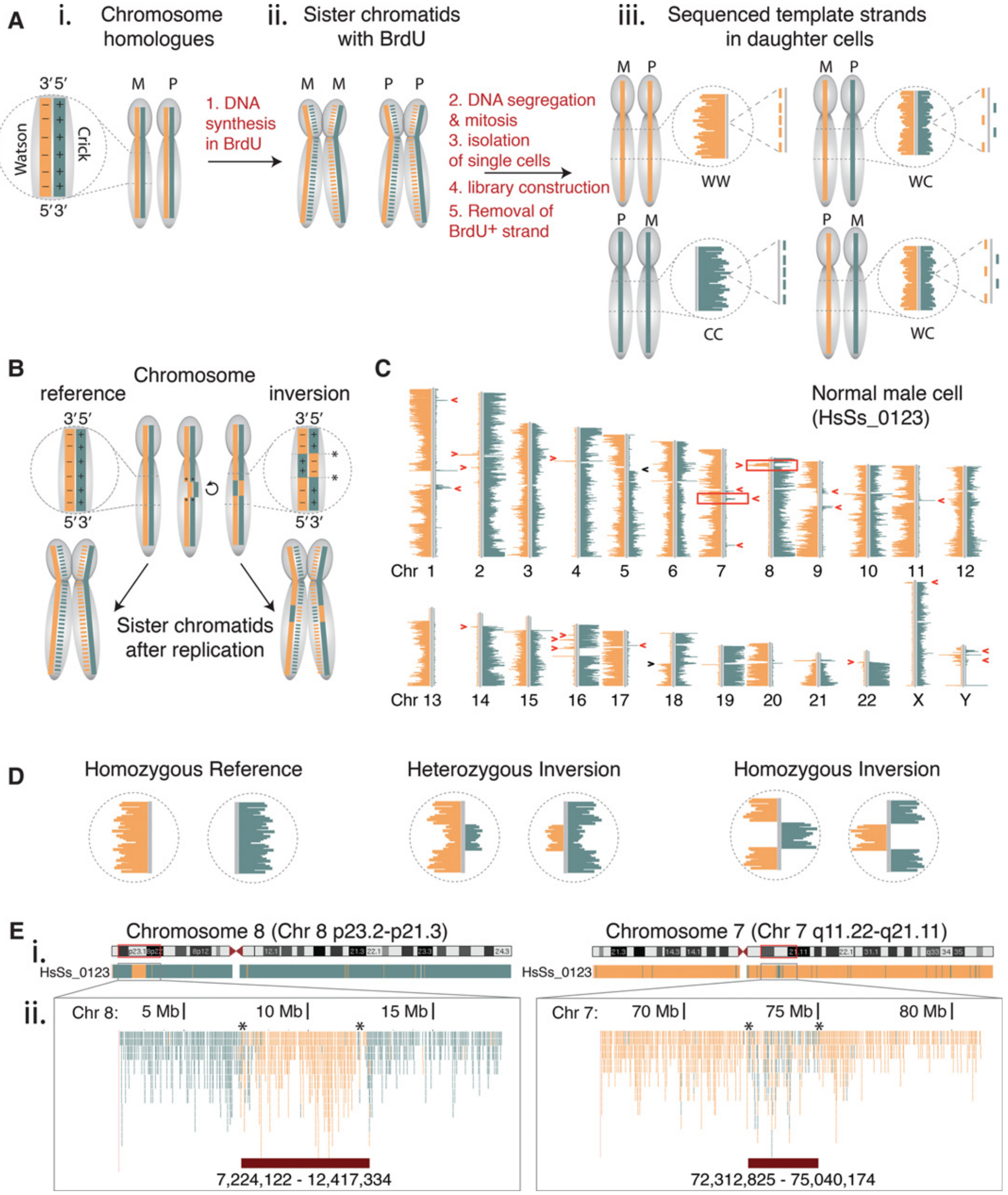

A
i. Chromosome homologues
ii. Sister chromatids with BrdU
iii. Sequenced template strands in daughter cells
3′5′
Watson
Crick
5′3′
M P
1. DNA synthesis in BrdU
M M
P P
2. DNA segregation & mitosis
3. isolation of single cells
4. library construction
5. Removal of BrdU+ strand
WW
WC
P M
CC
WC
B
Chromosome
reference
inversion
3′5′
5′3′
Sister chromatids after replication
C
Normal male cell (HsSs_0123)
Chr 1 2 3 4 5 6 7 8 9 10 11 12
Chr 13 14 15 16 17 18 19 20 21 22 X Y
D
Homozygous Reference
Heterozygous Inversion
Homozygous Inversion
E
i.
Chromosome 8 (Chr 8 p23.2-p21.3)
HsSs_0123
Chromosome 7 (Chr 7 q11.22-q21.11)
HsSs_0123
ii.
Chr 8: 5 Mb 10 Mb 15 Mb
7,224,122 - 12,417,334
Chr 7: 70 Mb 75 Mb 80 Mb
72,312,825 - 75,040,174

**Figure 1.** Visualizing inversions at high resolution in single cells by template strand sequencing (Strand-seq). (*A*) Inheritance patterns of template strands during mitotic division. (*i*) Maternal (M) and paternal (P) chromosome homologs consist of complementary but unique DNA strands, called Watson (W, minus strand, orange solid line) and Crick (C, plus strand, teal solid line). (*ii*) DNA synthesis incorporates BrdU into newly formed strands (dotted lines). These are selectively removed during Strand-seq library construction to generate sequencing reads from the template strands only. (*iii*) When aligned to the reference genome and represented on chromosomal ideograms (using BAIT software), the possible combinations of template strand inheritance for each homolog in daughter cells is shown. For a diploid cell, the template strands of any given chromosome can be inherited as WW, CC, or WC, where each strand represents either the maternal or paternal homolog. (*B*) Inversions are a localized reorientation in the W-C state along each single DNA strand of a chromosome, with respect to the reference genome. Asterisks denote breakpoints. (*C*) A single Strand-seq library from an adult male shows the template strand inheritance patterns of all chromosomes (Chr). Inversions appear as segmental changes in strand orientation along a chromosome (red arrowheads) and are distinguished from sister chromatid exchange events (black arrowheads). Boxed regions on Chr 7 and Chr 8 are previously described inversions. (*D*) Expected Strand-seq results for each possible inversion genotype. (*E*) UCSC Genome Browser view of Strand-seq data of cell shown in *C*, BED-formatted and uploaded as custom tracks for Chr 8p23 (homozygous inversion, *left* panel) and Chr 7q11 (heterozygous inversion, *right* panel). (*i*) Whole chromosome 'packed' view of aligned Crick (teal) and Watson (orange) reads. (*ii*) Zoom of genomic regions containing inversions denoted by boxed regions in *C*. In this 'squished' view, each single aligned C and W read is denoted as an individual teal or orange line, and the chromosomal coordinates of the inversions (*lower* red bars) are manually mapped as the last base pair position of the 5′ and 3′ read flanking the inversion (asterisks).

manual breakpoint predictions for this single cell (for example, compare Chr 8 Invert.R breakpoints of 7,219,983–12,409,738 in Fig. 2B, i, to the manual breakpoints of 7,224,122–12,417,334 in Fig. 1E, ii) as well as previous reports (Kidd et al. 2008; Antonacci et al. 2009; Martinez-Fundichely et al. 2014). To define the limits of detection of Invert.R, we randomly down-sampled a single-cell library and tested the size range of inversions reliably predicted using our pipeline (Supplemental Fig. S2). The range of inversions

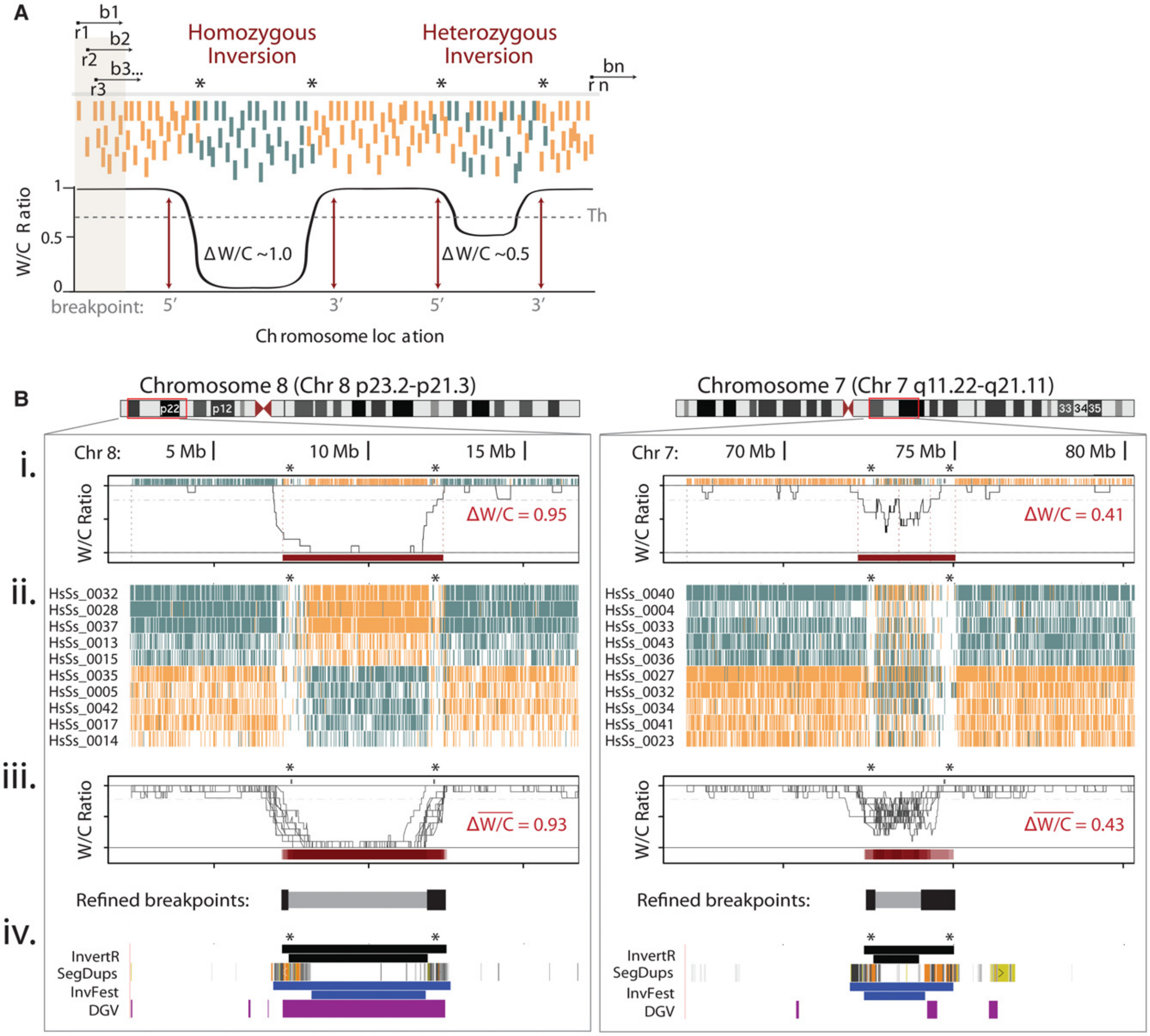


**Figure 2.** High-throughput characterization of inversions in multiple single cells using Invert.R. (*A*) Illustration of Invert.R, which steps along each read (r) to survey a user-defined bin (b) and calculates the proportion of Watson (W, orange) and Crick (C, teal) reads (W/C ratio). The bin moves sequentially ($b_1$ to $b_n$) along every read ($r_1$ to $r_n$), and the W/C ratio calculation is repeated and plotted as a histogram (black line). Putative inversions are localized to the genomic region (breakpoints) where the W/C ratio passes a threshold (Th) and genotyped based on the magnitude of change (Δ W/C) at the inverted segment. (*B*) Zoom of localized Invert.R histograms and corresponding Strand-seq libraries of a homozygous (Chr 8p23; *left* panel) and heterozygous (Chr 7q11; *right* panel) inversion, viewed in UCSC Genome Browser (red box). (*i*) Invert.R output of a single library (shown in Fig. 1E) with predicted breakpoints (dotted lines) and corresponding ΔW/Cs of each inversion (red bar). Asterisks denote regions with low read depth that often flank inversions. (*ii*) UCSC Genome Browser view of 10 additional libraries from the same donor (*iii*) with overlaid Invert.R histograms. Sequence gaps (gray bars *above* histograms) and a heat map of the overlapping inversion predictions (red bars *below* histograms) are included. The minimal inverted region (inverted segment predicted in 80% of cells, gray bar *below* histogram) and flanking breakpoint ranges (inverted segment predicted in 20% of cells, black bars *below* histogram) calculated from all 10 cells. This placed the breakpoint ranges to 7,183,914–7,404,466 (5′) and 11,880,370–12,489,771 (3′) for the Chr 8 inversion, and 72,380,014–72,659,960 (5′) and 73,989,814–75,007,165 (3′) for Chr 7. (*iv*) Simultaneous view of inversions mapped by Invert.R (black), in relation to segmental duplications (SegDups), and previously reported inversions in the Database of Genomic Variants (DGV, purple) and the Human Polymorphic Inversion Database (InvFest, blue). For Invert.R and InvFest, the minimal inverted region is represented as the *lower* bar in the track, with the maximal inverted regions (outermost breakpoint ranges) represented as the *upper* bar.

detectable in a given cell was inversely correlated to the sequencing depth of the library. However, even at very low genomic coverage (0.05×), inversions larger than 25 kb were called by Invert.R.

To refine inversion breakpoints, we analyzed 10 additional cells from the same individual (Fig. 2B, ii). Invert.R mapped the Chr 8p23 and Chr 7q11 inversions to the same genomic location in each cell (Supplemental Fig. S3) and showed a high degree of overlap between the W/C ratio distributions in overlaid histograms (Fig. 2B, iii). This concordance demonstrates that changes in strand orientation represent bona fide structural variants that are accurately found using our approach. Inversion breakpoints were narrowed by finding the consensus between inversion predictions (see Supplemental Methods), with the proportion of overlapping calls graphically depicted as heat maps under the overlaid histograms (Fig. 2B, iii, red bars). The minimum inverted region was defined as the overlap present in at least 80% of the cells (Fig. 2B, iii, gray bars) and the maximum inverted region (which defines the outer limits of the inversion) as the overlap present in at least 20% of the cells (Fig. 2B, iii, black bars). This localized the maximum Chr 8 inversion to 7,183,914–12,489,771, and the maximum Chr 7 inversion to 72,380,014–75,007,165 (Fig. 2B, iii). The precise breakpoints are predicted to reside between the minimum and maximum inverted regions, which for these two inversions had a resolution of 220.5 kb–1.02 Mb and overlapped with large blocks of segmental duplications (Fig. 2B, iv). This resolution coincided very closely with previous reports, including those listed in the Database of Genomic Variants (DGV) (MacDonald et al. 2014), and the Human Polymorphic Inversion Database (InvFest) (Fig. 2B, iv; Martinez-Fundichely et al. 2014), highlighting how inversion breakpoints can be accurately mapped using Invert.R.

### Characterizing polymorphic inversions in a multidonor population of single cells

With a robust method to accurately localize and genotype inversions, we set out to explore polymorphic inversions across the human genome. To investigate the extent of cellular heterogeneity within a sample population, we generated 47 Strand-seq libraries from a pool of 353 separate cord blood (CB) donors. We selected this sample to survey the spectrum of common inversions in multiple individuals simultaneously and rapidly characterize the distribution and allelic frequency of polymorphism in the normal human genome. Individual cells from the pooled CB sample were sorted as single cells and daughter cells arising after a single cell division in 5 µM BrdU were isolated and prepared for library construction (see Methods). Libraries were filtered for a minimum read depth of 20 reads/Mb, showed an average of 204 reads/Mb, with genomic coverage ranging from 0.01–0.11× per library (Supplemental Table S1). Assuming equal donor cell contributions, the majority of cells (98%) in our population likely represent a unique human genome, and collectively, they represent a normal mixed donor population. We analyzed each library independently with Invert.R and overlaid the resulting histograms of the W/C ratios for each chromosome (Supplemental Fig. S4). Genomic regions where a segmental change in strand orientation was mapped to the same location in at least two cells were flagged as regions of interest (ROIs) that contain putative inversions (Supplemental Fig. S4, red bars). Some ROIs flagged by Invert.R were in complex genomic regions that contained reference assembly gaps or several distinct template strand states (Supplemental Fig. S5). These were manually refined and characterized (by Invert.R) based on the number of W and C reads at each locus, using Fisher's exact test to determine the best-fit genotype (see Methods).

To distinguish between true polymorphisms and other genomic features, we examined the frequency that each ROI was called as heterozygous (i.e., WC, containing a ratio of W and C reads that best fit a heterozygous genotype) (Supplemental Fig. S6). We identified 46 ROIs that were called WC in ≥80% of cells (herein denoted as AWC, Always WC) (Supplemental Table S2). In some cases, AWC regions coincided with large blocks of segmental duplications with complex architecture (for example, the inversion breakpoint on Chr 8p23 in Fig. 2B, asterisks; Supplemental Fig. S7); in other cases, they did not (see ROIno.10.7 in Supplemental Fig. S5). The AWC regions ranged from 4105 bp to 1.5 Mb in size, and together comprised 15.3 Mb (0.5%) of the genome. Although 67.4% (31) overlapped with inversion entries in the DGV, AWCs are unlikely to be inversions because we would expect a higher homozygous frequency in our population if they represented a common inversion. Indeed, none of the AWCs were in Hardy-Weinberg equilibrium (HWE) (Supplemental Table S2; Wang and Shete 2012). Instead, we hypothesize that these are underrepresented repetitive sequences in the human reference assembly, which physically occur at several genomic locations but are represented in the reference assembly at a single locus. For instance, the pseudoautosomal regions (PARs) are present on both sex chromosomes but only represented on Chr X (as per convention, the regions on Chr Y were masked [represented as 'N'] in the assembly these data were aligned to) (The 1000 Genomes Project Consortium 2012). Consequently, reads originating from both Chr X and Chr Y PARs align only to Chr X, often appearing as WC (see tip of Chr Xp of Supplemental Fig. S4). Moreover, the overall average densities at AWCs was 5.7-fold greater than the average read depth at other loci (1700 versus 300 reads/Mb, with the highest density found at centromeric ROIno.19.2, averaging over 32,000 reads/Mb). This supports the hypothesis that these sequences are present in multiple copies but are collapsed into one locus in the reference assembly. To test our hypothesis, we aligned reads mapping to AWCs to short tandem repeat (STR) sequences recently described and appended to the human reference assembly (Chaisson et al. 2015) and found that 41 (89%) AWCs contained reads mapping to at least one STR. For instance, the AWC at ROIno.10.7 (Supplemental Fig. S5) mapped to 14 STRs on 13 different chromosomes, explaining why this region always appears as WC in Strand-seq libraries.

We also identified 24 regions that had a high (≥80%) homozygous frequency across all cells (Supplemental Fig. S6). We predict that these regions are either minor alleles or misoriented segments of the human reference assembly, as we previously identified and confirmed in the mouse (Falconer et al. 2012). These regions ranged in size from 18.9 kb to 1.7 Mb and collectively comprised 8.4 Mb (0.27%) of the human genome, and 12 (50%) overlapped with DGV-identified inversions (Supplemental Table S3). They included two ROIs on Chr 10q11 (ROIno.10.9 and ROIno.10.10) (Supplemental Fig. S5), which match fragments recently resolved as assembly misorientations using targeted, long-read sequencing of BAC clones (Chaisson et al. 2015; EE Eichler, pers. comm.). Interestingly, while ROIno.10.10 appeared misoriented (100% of individuals were homozygous), some individuals in the population showed a polymorphism at ROIno.10.9, suggesting that it is an inversion where the minor allele is represented in the reference assembly (Supplemental Fig. S5). We also found most individuals (82%) exhibited a homozygous inversion at ROIno.16.22, with a small proportion (18%) harboring a heterozygous inversion at the locus (Supplemental Table S3). Since this ROI is flanked by two misoriented fragments (ROIno.16.21 and ROIno.16.23), we predict that the entire contig (GL000125.1) is misoriented, and ROIno.16.22 is a rare inversion that falls within this misorientation. Finally, the largest misoriented fragment (ROIno.1.13) fell on Chr 1q21, overlapped 19 inversions in the DGV, and encompassed 27 unique genes. This fragment was misoriented in 93% of the population we sampled, and the breakpoints disrupted several *NBPF* paralogs of a tumor-suppressor gene family associated with neuroblastoma (Vandepoele et al. 2005; Dumas et al. 2012; Andries et al. 2015). This shows that errors in genome assemblies can appear as structural variants using conventional techniques but are more accurately annotated using our Strand-seq approach.

The remaining 111 ROIs were heterogeneous between cells and represent polymorphic inversions in the sampled population (Supplemental Table S4). In total, these polymorphisms comprised 34.9 Mb (1.13%) of the genome, 40 of which (36%) did not overlap with inversions listed in the DGV (Fig. 3A, i; Supplemental Table S4). They ranged in size from 16.5 kb to 3.9 Mb, with a median of 175 kb (Fig. 3A, i). Ninety-five percent of the inversions identified were below 1 Mb in size (Fig. 3A, i, gray box), which marks the limit of detection for traditional cytogenetic techniques (Youings et al. 2004; Feuk 2010), and 87% were above the 50-kb detection range of nontargeted HTS approaches (Sudmant et al. 2015). When we compared the genotypes of all polymorphic loci between each cell (Supplemental Fig. S8), we observed extensive heterogeneity in the structural composition of each genome, with cells clustering based on similar inversion profiles. This suggests that the relatedness between individual cells in a heterogeneous sample (for example, in defined human populations or tumor samples) can be visualized by the set of inversions mapped in single Strand-seq libraries.

The polymorphisms showed a distinct genomic distribution, with over half (51.4%) of the inversions present on just five autosomes (Chromosomes 7, 9, 15, 16, and 17), whereas six autosomes (Chromosomes 3, 8, 10, 12, 19, and 21) together contained only 9% of the inversions (Fig. 3A, ii; Supplemental Table S4). We did not observe any inversions on Chr 13 or Chr 18 in our population. The allelic frequencies of the autosomal inversions ranged from 0.05 to 0.89, with 87% (86) in HWE (Supplemental Fig. S9; Supplemental Table S4). The ROIs not in HWE typically had a high proportion of heterozygous cells. Whereas six (46%) were adjacent to centromeres or telomeres (e.g., ROIno.4.1 and ROIno.17.7), others encompassed genes with disease associations, including ROIno.17.4, that contains the kinase *MAP2K* (Kim and Choi 2010), and ROIno.16.17, that contains a p53 target gene *TP53TG3* (Fig. 3C, i, asterisk; Ng et al. 1999).

For chromosomes harboring multiple inversions, we observed clusters forming large blocks of highly polymorphic domains. For instance, a ∼20-Mb domain surrounding the Chr 7 centromere (p12.1-q11.13) harbored seven distinct polymorphic inversions (Fig. 3B, i). To visualize the inheritance patterns of those present on Chr 7, we performed a cluster analysis based on genotype (Fig. 3B, ii) and noticed no obvious correlation with genomic distance (Fig. 3B, arrowheads and asterisks). We also identified 13 distinct polymorphisms in a ∼20-Mb region on the p-arm of Chr 16 (Fig. 3C, i). The inversions here clustered into distinct blocks based on frequency, where one block contained very rare inversions (such as ROIno.16.6 and ROIno.16.18) and another block contained highly prevalent inversions with frequencies above 0.8 (including ROIno.16.9 and ROIno.16.11) (Fig. 3C, ii, arrowheads; Supplemental Table S4). We identified 24 inversions with frequencies >50%, indicating that these alleles in the human reference assembly do not represent the common variant found in our sampled population (Supplemental Fig. S9, dotted line). We mapped several at the pericentromeric region of Chr 9, where recombination events have proven difficult to visualize due to the high number of segmental duplications in this area (for details, see Chr 9 in the interactive PDF, Supplemental Data File S1). Taken together, these analyses demonstrate that rearrangements cluster in polymorphic domains within the human genome, and that allelic states can be used to study inversion haplotypes and discern relationships between inversions and single cells in a heterogeneous sample.

## A genome-wide map of an individual's inversions reveals their distinct inversion profile

We have shown how analyzing multiple Strand-seq libraries with Invert.R can be applied to investigate the distribution and frequency of inversions in a population. To map the entire set of inversions present within an individual genome and define their inversion profile, we analyzed 140 Strand-seq libraries from the BM of a single adult male. We merged data from all WW and CC chromosomes to create a large composite file that preserved directionality while increasing read depths (37.7- to 66.6-fold) for each chromosome (Supplemental Fig. S10). Using Invert.R, we identified 132 ROIs, which we further refined by removing the AWC regions identified in the pooled donor population (see Methods).

Upon genotyping the refined ROIs, we identified 86 inversions, totaling 34.4 Mb (1.11%) of the male genome (Fig. 4A, blue circles; Supplemental Table S5). The inversions ranged in size from 1750 bp to 4.0 Mb, with a median of 197 kb (Supplemental Fig. S11). Notably, 38 (44%) did not overlap with inversions listed in the DGV, while 48 (56%) overlapped with inversions identified in the pooled donor population (Supplemental Fig. S11; Supplemental Table S5). More than one-third of the inversions (31) mapped to the polymorphic domains found on Chromosomes 7, 9, and 16, whereas no inversions were found on Chromosomes 13, 18, or 19 (Fig. 4A, lower table).

To explore differences between inversion profiles of two individuals, we generated another 106 Strand-seq libraries from the CB of a newborn female (see Methods) and repeated the analysis described above for the adult male. Here, we located 60 inversions, which comprised 23.3 Mb (0.77%) of the female genome (Fig. 4A, pink triangles; Supplemental Table S6). The sizes ranged from 740 bp to 2.15 Mb, with a median of 245 kb (Supplemental Fig. S11). Of these, 24 (40%) did not overlap with inversions listed in the DGV, 38 (63%) overlapped with inversions found in the male inversion profile, and 48 (80%) overlapped with inversions from the pooled donor population analysis (Supplemental Fig. S11; Supplemental Table S6). The polymorphic domain on Chr 16 contained 15% of the inversions characterized in this inversion profile, and again no inversions were found for Chr 13 or Chr 18, along with Chr 8 (Fig. 4A, lower table).

To compare the two inversion profiles in greater detail and visualize the inversions in the context of different genomic features, we created Circos diagrams (Krzywinski et al. 2009) for each chromosome (Fig. 4B; Supplemental Data File S1). This allowed us to simultaneously visualize the presence and genotype of inversions between the individuals and pooled donor population and the presence of misoriented fragments and AWC regions, and to compare these to DGV-listed inversions and gene densities at specific loci. On Chr 15q13, we found a large (∼2 Mb) heterozygous inversion in the female inversion profile (fCB.15.4) and pooled donor population (ROIno.15.9) that matches a known inversion in a gene-rich region (Supplemental Data File S1). A small (∼94 kb) heterozygous inversion on Chr 17q12 present in the male inversion profile (mBM.17.4) and the population (ROIno.17.10) was found near a 3′ breakpoint of a reported inversion, but does not actually overlap with it (Supplemental Data File S1). Neither inversion profiles harbored a known Chr 17q21 inversion that we identified in the population (ROI.no.17.16) (Supplemental Data File S1). We

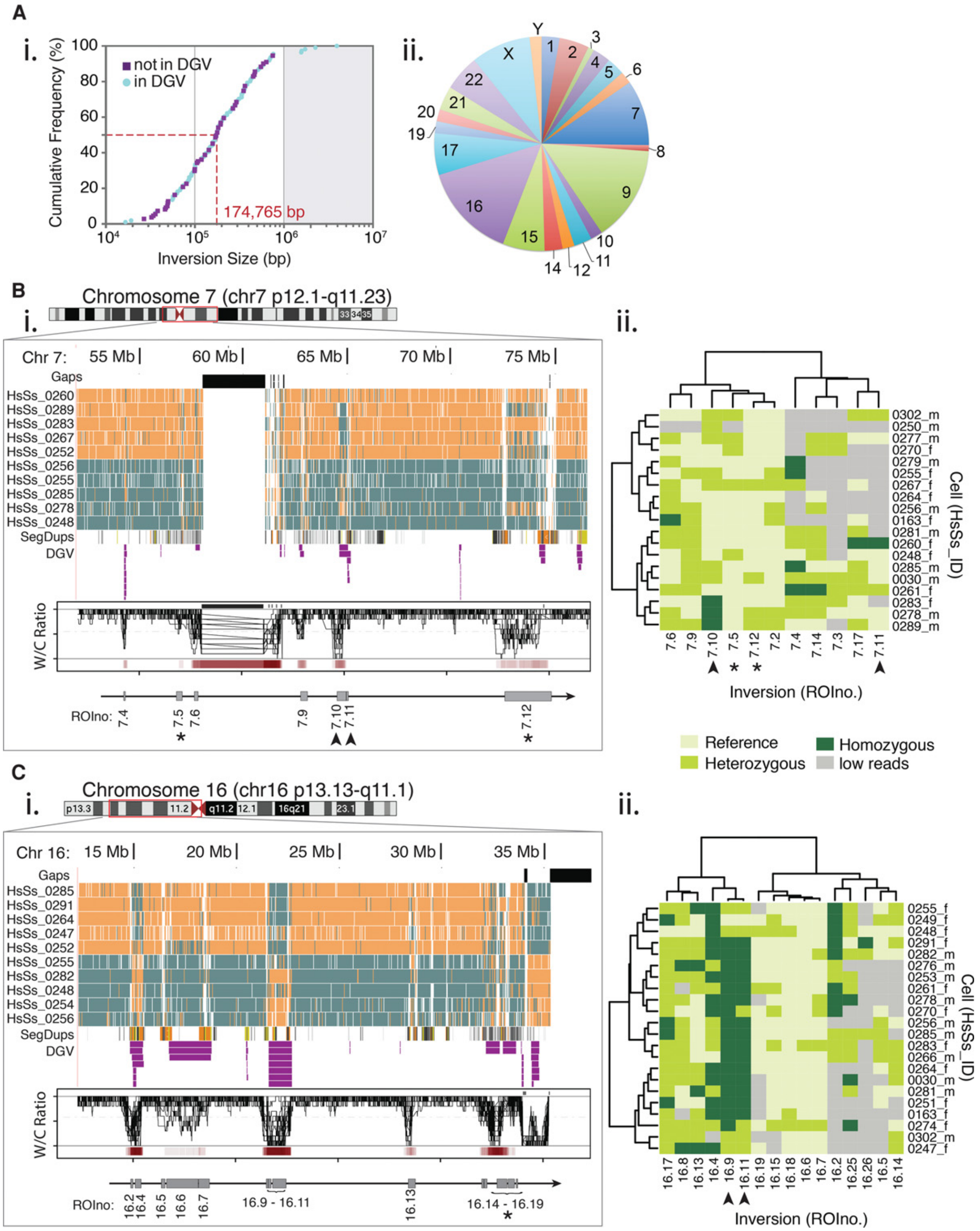
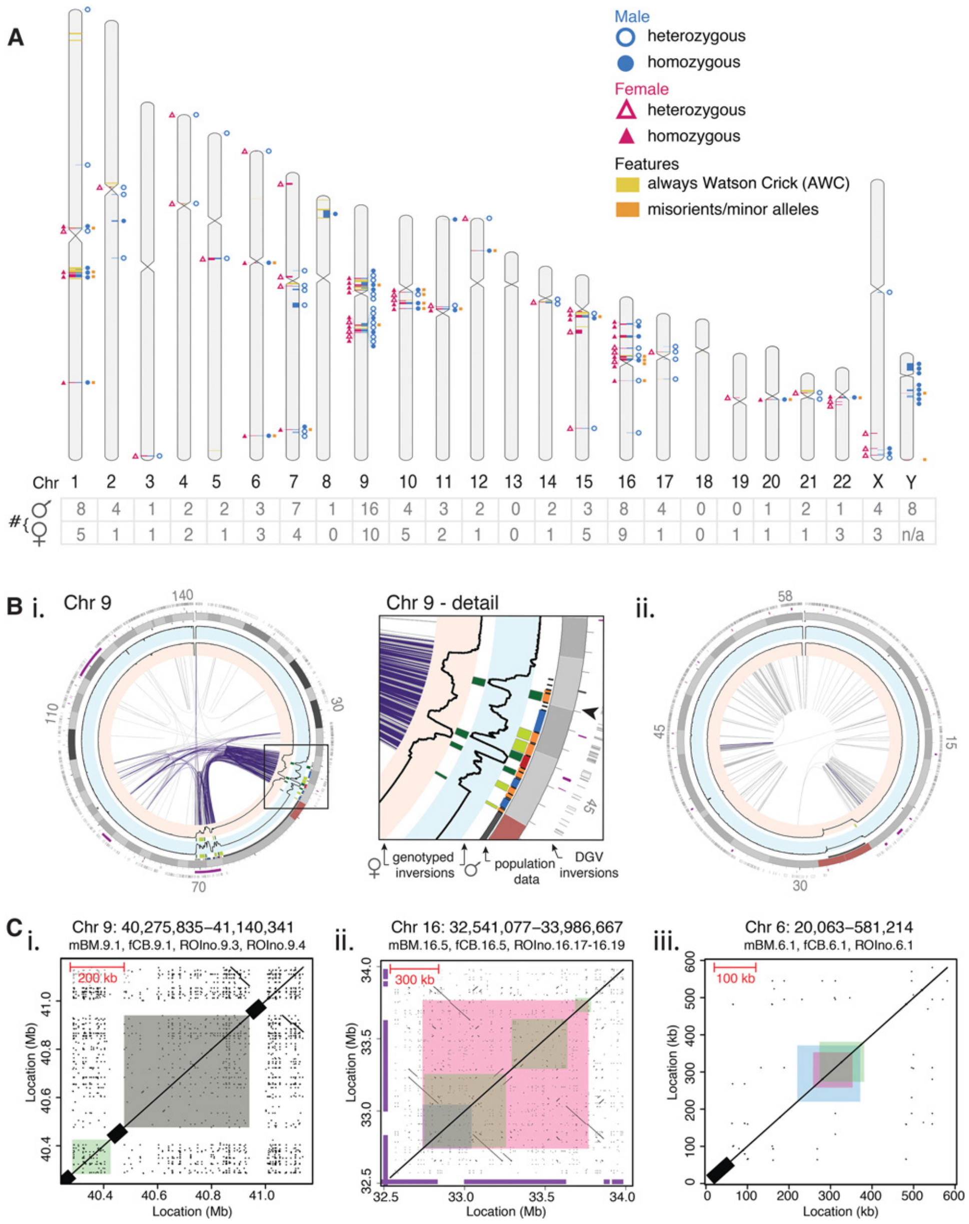


**Figure 3.** Polymorphic inversions mapped in multiple single cells of a pooled donor population. (*A*) Size and genomic distributions of 111 polymorphic inversions identified in the pooled donor population. (*i*) The cumulative frequency of inversion sizes in base pairs (bp), divided into new inversions (purple squares) and those overlapping with the Database of Genomic Variants entries (blue circles). The median inversion size (dashed red line) is well below the 1-Mb detection limit of traditional cytogenetic techniques (gray shading). (*ii*) Distribution of the total number of inversions present on each chromosome. (*B*, *C*) Polymorphic domains (red box) mapped to Chr 7 (*B*) and Chr 16 (*C*). Asterisks and arrowheads denote specific inversions highlighted in the main text. (*i*) Detail of the domains shown in the UCSC Genome browser 'packed' view for 10 representative Strand-seq libraries, along with tracks for sequence gaps (black), segmental duplications (SegDups), and inversions identified in the DGV (purple). Corresponding overlaid Invert.R histograms of W/C ratios and inversion frequency heat maps (red bars) are shown in the *lower* panel. The polymorphic inversions (gray boxes) and corresponding ROIno identifiers are shown *below*. (*ii*) Clustered heat maps of the genotyped inversions (*x*-axis) identified in each cell (*y*-axis). Inversions are depicted as pale green (homozygous reference), medium green (heterozygous), or dark green (homozygous). In some cases, too few reads were present in the region to genotype the ROI (gray).



| Chr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # ♂ | 8 | 4 | 1 | 2 | 2 | 3 | 7 | 1 | 16 | 4 | 3 | 2 | 0 | 2 | 3 | 8 | 4 | 0 | 0 | 1 | 2 | 1 | 4 | 8 |
| # ♀ | 5 | 1 | 1 | 2 | 1 | 3 | 4 | 0 | 10 | 5 | 2 | 1 | 0 | 1 | 5 | 9 | 1 | 0 | 1 | 1 | 1 | 3 | 3 | n/a |

**Figure 4.** Genome-wide comparison of inversion profiles of an adult male and newborn female. (*A*) The inversion profile characterized for a male (blue circles, *right*-hand side) and female (pink triangles, *left*-hand side). Each inversion (plotted using Idiographica v2.2 [Kin and Ono 2007]) was genotyped as either heterozygous (empty symbols) or homozygous (filled symbols). The number of inversions per chromosome is listed *below* (table). The location of Always Watson Crick regions (yellow) and misorients or minor alleles (orange) are also depicted. (*B*) Invert.R histograms (black lines) for the adult male (blue background) and newborn female (pink background) were overlaid on Circos plots, with all inversions plotted (heterozygous in light green, homozygous in dark green). See Supplemental Data File S1 for other chromosomes. Palindromic intra-chromosomal segmental duplications (purple lines) correlate with the inversion load of each chromosome. (*i*) Chr 9 contains several inversions clustered within palindromic segmental duplications (purple links). Structural differences of inversions were seen between the two donors within this complex region of the genome (Chr 9, detail). The arrowhead marks the area depicted in *C*, *i*. (*ii*) Nonpalindromic segmental duplications (gray links) are common on Chr 19, which contains a single inversion. (*C*) Dot plots illustrate the genomic architecture of inversions, which can be flanked by (*i*) reference assembly gaps (black bars on diagonal axis), (*ii*) palindromic segmental duplications, or (*iii*) nonrepetitive sequence. Sequence coordinates that were self-aligned are listed *above* each plot, with the inversions found in the male (mBM; blue), female (fCB; pink), and pooled donor population (ROIno; green) highlighted. Inversions listed in the Database of Genomic Variants are shown (purple bars on *x*- and *y*-axes). See Supplemental Data File S2 for all other inversions.

also identified multiple polymorphic inversions at the centromere of Chr 9 that distinguished the two inversion profiles (Fig. 4B, i), including four heterozygous inversions located on the p-arm (Chr 9p13-p11) in the male inversion profile that did not overlap with any inversions listed in the DGV (Fig. 4B, i, Chr 9 detail).

Importantly, we observed almost perfect overlap between several inversions predicted in our data sets, and those validated using orthogonal techniques. For instance, the breakpoints of the inversions we mapped at Chr 16p12.3 showed >86% overlap with an inversion characterized in six independent studies using combinations of BAC clone sequencing, mate-pair sequencing, optical mapping, PCR, and FISH technologies (Supplemental Fig. S12; Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008, 2010; Pang et al. 2010; Teague et al. 2010). Of note, in some genomes from our population study, we observed multiple inversions with distinct genotypes at this locus (see Fig. 3C), suggesting that the region is more complex than previously reported and multiple recombination events have generated nonrecurrent but overlapping breakpoints here (Gu et al. 2008). These complex events can only be seen at the population level, and more in-depth studies are required to better resolve the region. Overall, we identified 21 inversions that each showed a >80% concordance to known inversions predicted using alternative technologies (Supplemental Table S7). These data help highlight the accuracy of our technique to predict inversion breakpoints.

Inversion breakpoints are often characterized by segmental duplications, which are thought to play a role in genomic rearrangements (Emanuel and Shaikh 2001; Samonte and Eichler 2002; Bailey et al. 2004; Sharp et al. 2008). To investigate whether the inversions we identified are also flanked by these repeats, we added both palindromic (inverted orientation to each other) and nonpalindromic (in direct orientation) intra-chromosomal segmental duplications to the Circos plots (Fig. 4B; Supplemental Data File S1). We found a positive correlation between the percent of bases inverted and segmental duplications per chromosome ($r^2 = 0.70$, $P < 0.001$), which was strongest for palindromic ($r^2 = 0.78$) versus nonpalindromic ($r^2 = 0.66$) segmental duplications (Supplemental Fig. S13). This correlation is highlighted at the polymorphic domains identified on Chr 7 and Chr 16 and the Chr 9 centromere, where inversion breakpoints mapped to clusters of palindromic segmental duplications (Fig. 4B, i; Supplemental Data File S1). Conversely, Chr 19, which only had a single rare inversion, was enriched for nonpalindromic intra-chromosomal segmental duplications (Fig. 4B, ii).

We further analyzed the sequences and surrounding genomic regions of the 257 inversions (137 unique) identified in all our data sets and created dot plots for each self-alignment (see Methods; Supplemental Data File S2). This revealed that 110 (43%) inversions were bordered by reference assembly gaps (Fig. 4C, i), which prevented analysis of the sequences directly flanking these variants. Of the remaining 147 inversions, 71 (48.3%) were flanked by palindromic (Fig. 4C, ii), and 18 (12.2%) were flanked by nonpalindromic segmental duplications (Supplemental Data File S2), suggesting that they were formed by nonallelic homologous recombination (Zody et al. 2008; Kidd et al. 2010; Stankiewicz and Lupski 2010; Dittwald et al. 2013). These were distinct from 58 (39.5%) inversions that were not flanked by any segmental duplications (Fig. 4C, iii). These inversions may have arisen by an alternative mechanism, or the flanking sequences may have diverged since the recombination event.

Finally, we tested levels of linkage disequilibrium to see if they were disrupted at the breakpoints of inverted loci (Stefansson et al. 2005; Bansal et al. 2007; Caceres and Gonzalez 2015). For every population in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012; Sudmant et al. 2015), we calculated the level of linkage disequilibrium (LD) between neighboring single nucleotide variants (SNVs) 5′ and 3′ to each breakpoint and summarized this as the mean for all inversions in our data set (Supplemental Fig. S14). We observed disrupted linkage patterns in many populations, where clear dips in LD values were evident directly abutted to the breakpoint locations. This effect was most profound in Asians and Europeans, where the standard deviation of LD values at the inversion breakpoints were 1.5- and 1.7-fold greater than randomly selected genomic loci. This suggests that these SNVs are not closely linked because they are physically separated by the inverted segment. Additionally, for the South Asian population, the mean LD within the inversion was significantly greater ($P < 1.5^{-16}$) than outside the inversion. These results suggest that the polymorphic inversions identified in our analysis differentially impact linkage patterns and/or recombination rates of nearby SNVs in human populations.

## Discussion

Here, we describe an unbiased and rapid methodology for high-resolution, high-throughput mapping and genotyping of genomic rearrangements. Using Strand-seq, we can now reliably visualize structural polymorphisms at the single-cell level with the genomic resolution afforded by HTS technologies. By maintaining genomic directionality, we circumvent the need for high-coverage sequence data to impute variants based on incongruous mapping signatures (Tuzun et al. 2005; Korbel et al. 2007; Alkan et al. 2011; Ritz et al. 2014); instead, we have developed a robust tool for discovering and characterizing large (kilobase-scale) rearrangements that visibly and statistically alter template strand states of chromosomes. Combined with Invert.R, we can generate a complete inversion profile at a fraction of the cost and time of conventional techniques. In addition, the ability to multiplex hundreds of single cells allows us to rapidly screen a population in a single experiment. This work represents a major advancement to studying copy-neutral rearrangements, which can be applied to uncover rare cells in heterogeneous samples, study levels of genetic mosaicism within individuals, and explore genomic variants between demographics.

By investigating the structural heterogeneity within a population of pooled donor cells, we mapped and genotyped 111 polymorphisms in 47 single cells simultaneously. Our criteria for inversion calls required that we observed a nonreference genotype at a given locus in a minimum of two cells, limiting our analysis to common inversions with a minor allele frequency >0.021. To identify more rare variants, a greater number of donors will be required (see Supplemental Discussion). The low sequence coverage of our Strand-seq libraries limits the size of detectable variants, and the smallest inversion discovered in our single-cell population study was ∼17 kb. While already orders-of-magnitude improved from traditional single-cell approaches, we expect the resolution can be pushed further with future improvements in library preparation protocols. Overall, our novel strategy of pooling multiple individuals in a single experiment represents a more high-throughput, unbiased, and comprehensive study of inversions than previous targeted approaches (Turner et al. 2006; Zody et al. 2008; Antonacci et al. 2009). These improvements allowed us to identify genomic regions that contain clusters of polymorphic inversions, including two 20-Mb domains on Chr 7 (p12.2-q21.11) and Chr 16 (p13.2-q11.2). These domains correspond to genomic locations predicted to recombine (Bailey et al. 2004) and may represent hotspots for structural variation in the human genome. Conversely, we found that Chr 13 and Chr 18 had no observable inversions, suggesting that genomic rearrangements may be suppressed on specific chromosomes.

By examining allelic frequencies of the rearrangements found in our population, we identified 24 rare alleles and likely sequence misorients in the reference assembly. We have previously shown that even highly sequenced reference assemblies such as the mouse contain misoriented regions that can be identified by Strand-seq and corrected in future builds (Falconer et al. 2012; Hills et al. 2013). Accurately annotating these regions in the human reference has immediate applications for genetic association studies. For instance, the 1.7-Mb misoriented fragment identified on Chr 1q21 encompasses several neuroblastoma-associated genes, which may have implications for which *NBPF* paralogs are associated with the disease and can be used as appropriate biomarkers (Vandepoele et al. 2005; Dumas et al. 2012; Andries et al. 2015). We also identified 46 AWC regions that likely point to repetitive sequences present in the genome that have not yet been placed in the reference assembly. Several AWC regions and misorients coincide with reported inversions in the DGV, suggesting that these are false-positive records. The recent advancement of long-read single-molecule sequencing (Chaisson et al. 2015) can help further refine the sequence and orientation at these complex genomic regions.

In generating multiple Strand-seq libraries from a single donor and combining these data to rapidly generate an inversion profile, we describe a new framework for characterizing structural rearrangements genome-wide. This strategy offers far greater efficiency (in terms of time, cost, resolution, and sensitivity) compared to alternative HTS approaches that require deep genomic coverage (up to 135×) to discover inversions (Tuzun et al. 2005; Bansal et al. 2007; Korbel et al. 2007; Kidd et al. 2008; Sharp et al. 2008; Ahn et al. 2009; Pang et al. 2010). Indeed, we characterized variants smaller than 1 kb and up to 4.5 Mb, (matching or surpassing the sensitivity of other studies [Tuzun et al. 2005; Bansal et al. 2007; Kidd et al. 2010; Martinez-Fundichely et al. 2014; Sudmant et al. 2015]). We also discovered several new inversions not previously described, many of which were within blocks of segmental duplications where variant mapping has previously been challenging (Feuk 2010; Alkan et al. 2011; Alves et al. 2012; Martinez-Fundichely et al. 2014). By accessing these complex and repetitive regions of the genome, we created a new catalog of human inversions that influence linkage patterns of local SNVs. Future studies will better test for population-specific effects on allele inheritance patterns, and we expect that the disruption of LD values will be most pronounced for larger inversions enriched in specific subpopulations. Overall, our approach paves the way for new studies of the molecular pathways driving recombination, whether specific sets of inversions are co-inherited in defined populations, and whether they act cooperatively to inform phenotypes.

Although our current understanding of how inversions impact human health is limited, we predict combinations of specific inversions can be used together to better understand ancestry and disease susceptibilities. For instance, the large inversion on Chr 8p23 exhibits a clinal distribution correlating with geographic distance from Ethiopia (Salm et al. 2012) and confers a reduced risk of autoimmune diseases (Hollox et al. 2008; Salm et al. 2012; Alves et al. 2014). The gene-rich polymorphisms found on Chr 7q11 and 15q13 correspond to inversions associated with complex neurological disorders, including mental impairments (Osborne et al. 2001; Tam et al. 2008; Hobart et al. 2010), seizures (Koolen et al. 2006; Sharp et al. 2008), or schizophrenia (International Schizophrenia Consortium 2008; Stankiewicz and Lupski 2010). A Chr 17q21 inversion that is common in Europeans (with a minor allele frequency of 0.2) predisposes children of heterozygous carriers to a microdeletion syndrome associated with developmental delays (Stefansson et al. 2005; Koolen et al. 2006; Zody et al. 2008; Donnelly et al. 2010). In our pooled donor population, we found nine heterozygous individuals for this inversion (and a similar allelic frequency of 0.23). Previously, the size and complexity of this genomic locus made this inversion difficult to localize (Cardone et al. 2008; Antonacci et al. 2009), but our ability to rapidly genotype multiple individuals simultaneously demonstrates how this may be useful as an early clinical diagnostic test.

Our single-cell approach to map structural rearrangements genome-wide offers a new opportunity to assay how different inversions operate in different populations. To date, studies have focused on single inversions in isolation and have not been able to consider (due to technical constraints) a more global analysis of inversions, such as whether groups or pairs act antagonistically or synergistically to impact individuals and health. Now that we can define inversion profiles and perform rapid population-based studies to compare multiple genomes simultaneously, we can finally investigate the evolutionary importance and phenotypic consequences of sets of inversions in specific populations.

## Methods

### Cell culture and Strand-seq library construction

Experiments were performed with fresh or previously frozen primary human hematopoietic cells, derived from (1) a 27-yr-old male cadaveric cryopreserved bone marrow sample (NTR00165) acquired from the Northwest Tissue Centre (Seattle, WA), (2) a female cord blood sample (C9053) acquired fresh (i.e., never frozen) from the Stem Cell Assay Laboratory (Vancouver, Canada), or (3) a cyropreserved pooled cord blood sample (CB7), where umbilical cord blood was harvested from 353 newborn donors, pooled together, and cryopreserved for banking by Stem Cell Assay Laboratory. All human cells were collected according to procedures approved by the University of British Columbia Research Ethics Board.

Human hematopoietic cells derived from bone marrow (adult male donor; library identification numbers HsSs_0001 - HsSs_0140) or cord blood (newborn female donor; HsSs_0141 - HsSs_0246, or 353 pooled donors; HsSs_0247 - HsSs_0315) were lineage-depleted (Human Progenitor Cell Enrichment kit, Stem Cell Technologies), and CD34$^+$ cells were FACS-sorted and plated in serum-free medium (Stemspan, Stem Cell Technologies), supplemented with human-recombinant (rh) growth factors SCF (100 ng/mL), Flt-3L (100 ng/mL), TPO (50 ng/mL), ±EPO (3 U/mL), and GM-CSF (20 ng/mL), as described (Mayani et al. 1993; Notta et al. 2011), all from Stem Cell Technologies. 5-bromo-2′-deoxyuridine (BrdU) was added to culture medium at a final concentration of 5 µM for one cell division (between 3 and 5 d), and daughter cells were isolated either by manual micromanipulation or by FACS-sorting nuclei based on quenching of Hoechst fluorescence by BrdU (Latt et al. 1977). To isolate nuclei, cells were resuspended in staining buffer (100 mM Tris-HCl [pH 7.4], 154 mM NaCl, 1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.2% BSA, and 10 µg/mL Hoechst 33258) and lysed using Nonidet-P40 (0.6% v/v; US Biological).

Isolated cells or nuclei were transferred into lysis buffer (5 µL; Nuclei EZ, Sigma), and Strand-seq library construction was performed on micrococcal nuclease-digested genomic DNA using the modified paired-end protocol (Illumina), described in Falconer et al. (2012), with minor modifications. Library preparation was scaled for a 96-sample format using an Agilent Bravo Automated Liquid Handling Platform; reaction volumes were reduced and all enzymatic step reactions were purified using solid-phase reversible immobilization paramagnetic beads Agencourt AMPure XP beads (1.8× vol. for all pre-adapter ligation reactions, and 0.8× vol. for all post-adapter ligation reactions; Agencourt AMPure, Beckman-Coulter), followed by EB buffer elution (6–10 µL; Qiagen).

### Illumina sequencing and sequence alignment

Completed libraries were pooled for two rounds of size selection using a 2% and then 1% agarose gel (E-Gel Ex, Invitrogen) to excise the 200- to 400-bp range. Library size distribution was confirmed using an Agilent High Sensitivity chip (Agilent), and the final concentration was determined using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen). Libraries were submitted for sequencing to the Michael Smith Genome Sciences Centre (Vancouver, Canada), where paired-end 100-nt reads were generated on the HiSeq 2000 (SBSxx) platform. For read metrics of each individual library, see Supplemental Table S1. The .fastq files were aligned to the human reference assembly (hg19/GRCh37, released Feb 2009) and analyzed using the open source software 'Bioinformatic Analysis of Inherited Templates' (BAIT) (Hills et al. 2013).

### Inversion localization

To manually visualize strand inheritance states of chromosomes, Strand-seq libraries were converted into BED files (using a modification of BEDTools [Quinlan and Hall 2010] bamToBed, implemented through BAIT [Hills et al. 2013]) and uploaded onto the UCSC Genome Browser (http://genome.ucsc.edu/) as custom annotation tracks on the hg19/GRCh37 assembly (Feb. 2009) (Kent et al. 2002). Library reads had duplicates removed and were filtered with a minimum mapping quality score of >10, and putative inversions were manually validated using the Genome Browser's navigation tool. Inversion breakpoints were manually located to the first base pair position of the first read present in a putative inversion.

To bioinformatically assess inversions in Strand-seq libraries, the custom software package, Invert.R, was developed (see Supplemental Methods). For single-cell libraries, duplicate reads were removed, mapping quality was set to >10, baseline threshold was set to 0.8, and a bin size of 25 reads was used, and the minLibs variable was set to 2 (see Supplemental Methods for details). To generate a directional composite file from multiple single cells of the same donor, the reverse complement of every read in the BED files of WW chromosomes was combined with all the reads from CC BED files for the corresponding chromosome (Supplemental Fig. S10). W/C ratios of the composite files were calculated using Invert.R. Here, a more stringent mapping quality ($q = 20$) and a larger bin ($b = 250$ reads) were applied to account for the higher read depth of each file, and minLibs was set to 1. Invert.R outputs an ROI file for further analysis.

### Inversion analysis

For the pooled donor population data set, Invert.R-identified ROIs were confirmed by visualizing Strand-seq libraries on the UCSC Genome Browser, and if required, they were refined by redefining start and stop coordinates based on read depths (e.g., Chr 10 and Chr 16) or gaps in the reference assembly (e.g., Chr 9). In male cells, ROIs falling within PAR1 or PAR2 were removed. The refined ROIs were genotyped using Invert.R (minReads = 10, and bg = 0.02) and allelic frequencies calculated (see Supplemental Methods). Autosomal ROIs were tested for Hardy-Weinberg equilibrium using the HWExact test (HardyWeinberg package [v1.5.4] [https://CRAN.R-project.org/package=HardyWeinberg]) and found to be in HWE when $P > 0.05$. ROIs were classified by counting the frequency of cells with a heterozygous or homozygous state. If a minimum of 10 cells showed a heterozygous frequency ≥80%, the region was defined as AWC, whereas if they had a homozygous frequency ≥80%, it was defined as a potential misorient or minor allele. If there were fewer than 10 cells at an ROI with ≥80% homozygous or heterozygous frequency, it was not classified. AWC sequences were aligned to short tandem repeat sequences that consisted of 13,305 unique regions, collectively encompassing 4,433,533 nt of sequence. The size of the STRs ranged between 24,489 and 31 nt in length, with a median of 333 nt. Polymorphisms were identified as ROIs where at least two cells showed different allelic states. To generate clustered heat maps of the polymorphisms (using the heatmap.2 function of gplots [v2.14.2] [https://CRAN.R-project.org/package=gplots]), ROIs were subdivided based on chromosome, and a distance matrix of genotyped cells was calculated by the Manhattan method (dist function) and hierarchically clustered by Ward's method (hclust function). To generate cell-by-cell heat maps of all ROIs, cells were clustered by the daisy pairwise dissimilarity method in cluster (v1.15.3) (https://CRAN.R-project.org/package=cluster).

For single-donor inversion profiles, ROIs identified by Invert.R were refined by removing regions overlapping the AWC regions and sequence gaps in the reference assembly, using BEDTools (Quinlan and Hall 2010) genomeCoverageBed function. For the male inversion profile, ROIs on Chr Y were manually refined, and any falling within the PARs were removed. To genotype the refined ROIs in Invert.R, background (bg) was set to 0.1 for both single donors, and minReads was set to 100 for the male and 50 for the female, to account for the different read densities in the composite files (the final average reads/Mb was 3311 for the male versus 1444 for the female). To generate Circos (v0.76) (Krzywinski et al. 2009) plots for each chromosome, data were formatted to include the Invert.R histograms of the W/C ratios (male in blue, female in pink) and genotyped inversion profiles for each individual (heterozygous inversions in light green and homozygous inversions in dark green), the classified ROIs identified in the pooled donor population (inner ring; AWCs in blue, misorients or minor alleles in red, and polymorphic inversions in orange), all inversions listed in the DGV (outer fuchsia bars), and Refseq genes listed in the UCSC Genome Browser (outermost gray bars). Intra-chromosomal segmental duplications were added as links, subdivided as palindromic (dark purple) or nonpalindromic (gray).

To compare inversion predictions between different data sets, we used the findOverlap tool of GenomicRanges (v2.14) (Lawrence et al. 2013), with minimum overlap set to 1 kb. To interrogate segmental duplications, the 'Segmental Dups' track was downloaded from the 'Repeats' group of the UCSC Genome Table Browser (Karolchik et al. 2014). The track was filtered for intra-chromosomal entries (i.e., the duplicated region fell on the same chromosome) and then subdivided into nonpalindromic (duplicated region in same orientation) and palindromic (duplicated region in inverted orientation) segmental duplications. The Table Browser was also used to extract the 'Refseq Genes' track from the 'Genes and Gene Predictions' group. Inversions reported in the DGV on the hg19/GRCh37 genome assembly were downloaded from the DGV database (MacDonald et al. 2014). Inversions listed in the InvFest (Martinez-Fundichely et al. 2014) database were lifted from hg18/GRCh36 to hg19/GRCh37 using the UCSC liftOver tool. To assess surrounding segmental duplications not present in the UCSC track, the entire inversion plus 200 kb of sequence upstream and downstream was self-aligned in a pairwise fashion using LASTZ (step = 20, seed match = 12, exact = 20, identity = 90 using the gapped, no chain, and no transition options), and the output used to generate dotplots in R (R Core Team 2013). ROIs and additional tracks were plotted as overlays onto these dotplots.

### Analysis of linkage disequilibrium

To analyze the level of LD at predicted inversions, we downloaded phased VCF files from the 1000 Genomes Project, phase 3 population data (Sudmant et al. 2015). The level of LD was calculated using VCFtools (Danecek et al. 2011) for pairs of SNVs with a MAF > 0.1 and within a 500-kb region upstream of or downstream from each inversion breakpoint, independently for all populations. To summarize LD across all inversions, the mean LD at each position was calculated and plotted for 400 SNVs spanning the inversion breakpoint (i.e., 200 5′ and 200 3′ of the breakpoint), with the SNVs found within the inversion always plotted on the right-hand side. To test LD at random genomic coordinates, 100 breakpoints were simulated from any position in the genome, and LD was calculated for the five continental populations as described.

## Data access

The Invert.R software is publicly available through SourceForge (https://sourceforge.net/projects/strandseq-invertr/), and the execution file is available in the Supplemental Information. The Strand-seq library sequence data from this study have been submitted to the NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA273996; BioSample accession numbers are SAMN03350247–SAMN03350539, inclusive.

## Acknowledgments

We thank Geraldine Aubert for discussions and critical review of the manuscript, along with Evan E. Eichler for personal communications and sharing his data regarding Chr 10 assembly misorientations. We also thank David Knapp, Connie Eaves, and the Stem Cell Assay Laboratory for providing and processing the cord blood samples, and the Terry Fox Laboratory Flow Cytometry Core Facility for cell sorting support. A.D.S. was supported by a Four Year Doctoral Fellowship from the University of British Columbia, and a Vanier Graduate Scholarship from the Natural Science and Engineering Research Council. We gratefully acknowledge support from the Canadian Institutes of Health Research (RMF-92093), the US National Institutes of Health (R01GM094146), the Canadian Cancer Society, and the Terry Fox Foundation (grants 018006 and 105265) as well as an Advanced Grant from the European Research Council (ERC Grant 294740 to P.M.L.).

*Author contributions:* A.D.S. performed all experiments, wrote bioinformatic programs, interpreted results, and wrote the manuscript. M.H. wrote bioinformatic programs, helped with data analysis. and helped write the manuscript. D.P. helped with data analysis. V.G. helped with results interpretation and commented on the manuscript. E.F. automated library construction, interpreted results, and helped write the manuscript. P.M.L. conceived the study and helped write the manuscript.

## References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19:** 1622–1629.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376.

Alves JM, Lopes AM, Chikhi L, Amorim A. 2012. On the structural plasticity of the human genome: chromosomal inversions revisited. *Curr Genomics* **13:** 623–632.

Alves JM, Chikhi L, Amorim A, Lopes AM. 2014. The 8p23 inversion polymorphism determines local recombination heterogeneity across human populations. *Genome Biol Evol* **6:** 921–930.

Andries V, Vandepoele K, Staes K, Berx G, Bogaert P, Van Isterdael G, Ginneberge D, Parthoens E, Vandenbussche J, Gevaert K, et al. 2015. NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer* **15:** 391.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* **18:** 2555–2566.

Antonarakis SE, Rossiter JP, Young M, Horst J, de Moerloose P, Sommer SS, Ketterling RP, Kazazian HH Jr, Negrier C, Vinciguerra C, et al. 1995. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood* **86:** 2206–2212.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol* **5:** R23.

Bansal V, Bashir A, Bafna V. 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res* **17:** 219–230.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14:** 307–320.

Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, Carlberg BM, Pettersson U. 1995. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet* **4:** 615–621.

Caceres A, Gonzalez JR. 2015. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res* **43:** e53.

Cardone MF, Jiang Z, D'Addabbo P, Archidiacono N, Rocchi M, Eichler EE, Ventura M. 2008. Hominoid chromosomal rearrangements on 17q map to complex regions of segmental duplication. *Genome Biol* **9:** R28.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158.

Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodriguez Rojas LX, Elton LE, Scott DA, Schaaf CP, et al. 2013. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* **23:** 1395–1409.

Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Mehdi SQ, Kajuna SL, Barta C, Kungulilo S, Karoma NJ, Lu RB, et al. 2010. The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am J Hum Genet* **86:** 161–171.

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, et al. 2012. DUF1220-domain copy number implicated in human brain-size pathology and evolution. *Am J Hum Genet* **91:** 444–454.

Emanuel BS, Shaikh TH. 2001. Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet* **2:** 791–800.

Falconer E, Hills M, Naumann U, Poon SS, Chavez EA, Sanders AD, Zhao Y, Hirst M, Lansdorp PM. 2012. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* **9:** 1107–1112.

Feuk L. 2010. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* **2:** 11.

Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* **1:** e56.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *PathoGenetics* **1:** 4.

Hills M, O'Neill K, Falconer E, Brinkman R, Lansdorp PM. 2013. BAIT: organizing genomes and mapping rearrangements in single cells. *Genome Med* **5:** 82.

Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, Rios CM, Kimberley KW, Gregg RG, Bray-Ward P. 2010. Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. *Am J Med Genet C Semin Med Genet* **154C:** 220–228.

Hollox EJ, Barber JC, Brookes AJ, Armour JA. 2008. Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *Genome Res* **18:** 1686–1697.

International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455:** 237–241.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The

UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42:** D764–D770.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143:** 837–847.

Kim EK, Choi EJ. 2010. Pathological roles of MAPK signaling pathways in human diseases. *Biochim Biophys Acta* **1802:** 396–405.

Kin T, Ono Y. 2007. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* **23:** 2945–2946.

Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, et al. 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* **38:** 999–1001.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645.

Latt SA, George YS, Gray JW. 1977. Flow cytometric analysis of bromodeoxyuridine-substituted cells stained with 33258 Hoechst. *J Histochem Cytochem* **25:** 927–934.

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118.

Macaulay IC, Voet T. 2014. Single cell genomics: advances and future perspectives. *PLoS Genet* **10:** e1004126.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42:** D986–D992.

Martinez-Fundichely A, Casillas S, Egea R, Ramia M, Barbadilla A, Pantano L, Puig M, Caceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res* **42:** D1027–D1032.

Mayani H, Dragowska W, Lansdorp PM. 1993. Cytokine-induced selective expansion and maturation of erythroid versus myeloid progenitors from purified cord blood precursor cells. *Blood* **81:** 3252–3258.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65.

Ng CC, Koyama K, Okamura S, Kondoh H, Takei Y, Nakamura Y. 1999. Isolation and characterization of a novel TP53-inducible gene, *TP53TG3*. *Genes Chromosomes Cancer* **26:** 329–335.

Notta F, Doulatov S, Laurenti E, Poeppl A, Jurisica I, Dick JE. 2011. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333:** 218–221.

Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, et al. 2001. A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* **29:** 321–325.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11:** R52.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

R Core Team. 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, Raphael BJ. 2014. Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics* **30:** 3458–3466.

Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, et al. 2012. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Res* **22:** 1144–1153.

Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet* **3:** 65–72.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* **40:** 322–328.

Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, et al. 2006. Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* **38:** 1032–1037.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61:** 437–455.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37:** 129–137.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81.

Tam E, Young EJ, Morris CA, Marshall CR, Loo W, Scherer SW, Mervis CB, Osborne LR. 2008. The common inversion of the Williams–Beuren syndrome region at 7q11.23 does not cause clinical symptoms. *Am J Med Genet A* **146A:** 1797–1806.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, et al. 2010. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci* **107:** 10848–10853.

Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* **3:** 439–445.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727–732.

Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F. 2005. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* **22:** 2265–2274.

Wang J, Shete S. 2012. Testing departure from Hardy–Weinberg proportions. *Methods Mol Biol* **850:** 77–102.

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, et al. 2007. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* **80:** 91–104.

Youings S, Ellis K, Ennis S, Barber J, Jacobs P. 2004. A study of reciprocal translocations and inversions detected by light microscopy with special reference to origin, segregation, and recurrent abnormalities. *Am J Med Genet A* **126A:** 46–60.

Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, et al. 2008. Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat Genet* **40:** 1076–1083.

*Received October 22, 2015; accepted in revised form June 13, 2016.*