Resolving genomic disorder–associated breakpoints within segmental DNA duplications using massively parallel sequencing

Xander Nuttle[1], Andy Itsara[1], Jay Shendure[1] & Evan E Eichler[1,2]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Published online 29 May 2014; doi:10.1038/nprot.2014.096

The most common recurrent copy-number variants associated with autism, developmental delay and epilepsy are flanked by segmental duplications. Complete genetic characterization of these events is challenging because their breakpoints often occur within high-identity, copy-number polymorphic paralogous sequences that cannot be specifically assayed using hybridization-based methods. Here we provide a protocol for breakpoint resolution with sequence-level precision. Massively parallel sequencing is performed on libraries generated from haplotype-resolved chromosomes, genomic DNA or molecular inversion probe (MIP)-captured breakpoint-informative regions harboring paralog-distinguishing variants. Quantification of sequencing depth over informative sites enables breakpoint localization, typically within several kilobases to tens of kilobases. Depending on the approach used, the sequencing platform, and the accuracy and completeness of the reference genome sequence, this protocol takes from a few days to several months to complete. Once established for a specific genomic disorder, it is possible to process thousands of DNA samples within as little as 3–4 weeks.

INTRODUCTION

Several regions of the human genome are predisposed to recurrent duplication and deletion[1]. Nonallelic homologous recombination (NAHR) between directly oriented segmental duplications, defined as contiguous sequences at least 1 kb in length having at least 90% identity, results in recurrent gain and loss of the intervening sequence[2,3]. Collectively, such events affect hundreds of genes and have been associated with many diseases, including autism, schizophrenia, intellectual disability, epilepsy, macrocephaly, microcephaly, congenital defects and severe obesity, among others[4].

Although a variety of technologies reliably detect recurrent duplications and deletions, determining the breakpoints within duplicated sequences remains a significant challenge. Pulsed-field gel electrophoresis followed by genomic Southern blots was originally used to map breakpoints in patients with Charcot-Marie-Tooth disease type 1A having a duplication at 17p12 (ref. 5). This method required the preparation of high-molecular-weight DNA, an often trial-and-error identification of informative restriction enzymes, and the design of probes adjacent to the breakpoints themselves. When array comparative genomic hybridization (array CGH) methods[6] became the standard for copy-number variant detection, they were often applied for the initial refinement of breakpoints[7,8]; follow-up with long-range PCR, subcloning and capillary sequencing in some cases then enabled the precise delineation of breakpoints[8]. This strategy worked well for breakpoints mapping in unique regions of the genome and would, in principle, prove effective in mapping breakpoints within small segmental duplications (<10 kb). However, it cannot be successfully applied to most recurrent, NAHR-mediated microdeletions and duplications, whose breakpoints map to the largest and most highly identical segmental duplications. In these cases, refinement by array CGH is of limited utility because of probe cross-hybridization. As a result, the researcher can only narrow the breakpoints to within hundreds of kilobases of nearly identical duplicated sequence. Thus, subcloning and sequencing several long-range PCR products across these large duplicated regions would be required for breakpoint resolution—a particularly difficult proposition, given the generation of nonspecific PCR products.

More recent bioinformatics approaches involving the analysis of split-read or read-pair sequence signatures from massively parallel whole-genome sequencing (WGS) data[9,10] are not reliable in these regions. These methods depend on the sequence read or the read-pair itself traversing the junction formed via NAHR. However, short read lengths, short library insert sizes and the paucity of distinguishing variants within breakpoint-containing segmental duplications make detection of a junction-spanning sequence read or read-pair highly unlikely. Finally, breakpoint resolution is often confounded by copy-number polymorphisms, gaps in the reference genome and alternative structural haplotypes affecting breakpoint regions[11,12]. Incomplete knowledge of the sequence, structure and genetic variation at these loci presents a substantial barrier to breakpoint localization regardless of the method used.

Despite its difficulty, accurate breakpoint resolution is crucial for understanding the origins and consequences of recurrent duplications and deletions. Obtaining breakpoint data from multiple independent events may elucidate factors influencing NAHR susceptibility and may help identify potential hotspots. It is becoming apparent, for example, that specific structural configurations within the genome increase susceptibility to some genomic disorders, whereas others are protective with respect to recurrent rearrangements[12,13]. Furthermore, precise breakpoint mapping will reveal the effects of recurrent rearrangements on genes within breakpoint regions. Most such genes have hardly been characterized, and their disruption may contribute to both disease phenotypes and phenotypic variability associated with some genomic disorders[14]. Here we detail a

protocol for resolving breakpoints including a series of experimental approaches (**Fig. 1**), and we provide general guidelines for its successful application to particular cases of interest. Although hybridization-based approaches are still the primary method by which copy-number variants are discovered, we focus mainly on two massively parallel sequencing strategies—analysis of WGS data and targeted capture and sequencing of informative regions using MIPs—as they provide the greatest potential for breakpoint resolution within duplicated sequence. More detailed protocols for the other methodologies outlined here have been previously published[15,16].

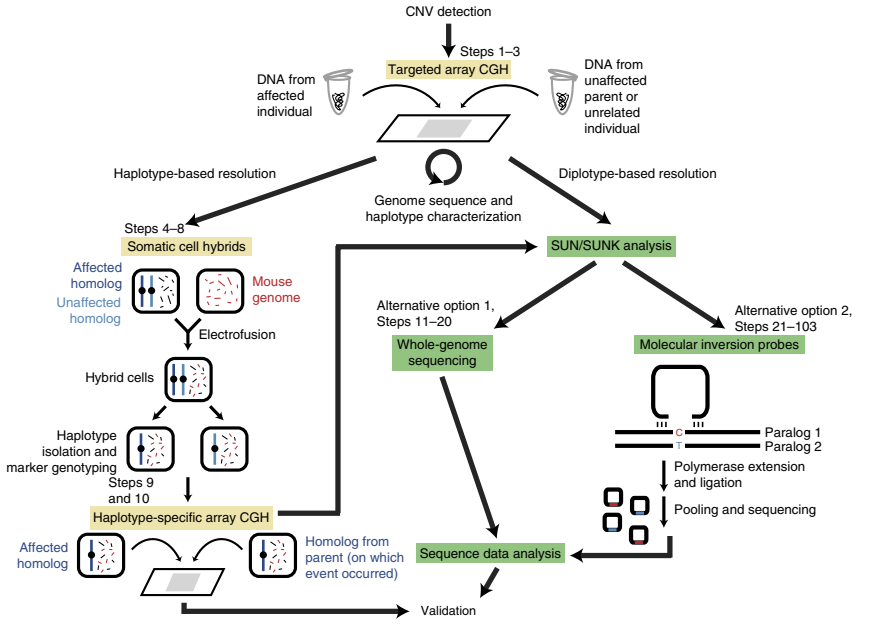

Figure 1 | General workflow for breakpoint resolution. The diagram outlines the typical stages involved in sequence-based breakpoint resolution and indicates some relevant associated experiments and computational analyses. Optional subsections of the procedure described briefly in this protocol are highlighted in yellow, whereas strategies for attaining sequence-level breakpoint resolution covered in more detail are highlighted in green. Targeted array CGH is displayed near the top of the figure because until recently it was the method of choice for breakpoint resolution after CNV detection—the sequencing-based approaches were only developed within the past few years. Today, however, sequencing approaches are sensible starting points for breakpoint resolution. Note that, depending on the particular region of interest, further genomic characterization of the region may be crucial to successfully refining breakpoint locations (see **Box 1** for discussion). Genomic characterization is often an iterative process (circular arrow), and because it facilitates data interpretation for all breakpoint resolution methods it is included in the diagram near the center. CNV, copy-number variant.

Concept and development

We originally leveraged massively parallel WGS to localize breakpoints for three individuals with the 17q21.31 microdeletion syndrome[14]. This recurrent microdeletion usually occurs via NAHR (**Fig. 2a**) between directly oriented segmental duplications that are ~145 kb in length and have >99% sequence identity. To refine breakpoints within the segmental duplications, singly unique nucleotides (SUNs) were identified from a sequence alignment between paralogs (**Fig. 2b**). The idea of using paralogous sequence variants to characterize duplicated regions is not new[17–21]. SUNs, however, represent a specific type of paralogous sequence variant because they uniquely distinguish one paralog from all other sequences in the genome, thereby enabling accurate sequence and copy-number genotyping for specific paralogs genome-wide[22]. Conceptually, any sequencing read carrying a SUN can be unambiguously assigned to a specific paralog even though it maps within segments of nearly identical sequence. Furthermore, quantifying read-depth from sequencing data over SUNs helps refine recurrent deletion and duplication breakpoints (**Fig. 2c**). Specifically, in the case of patients with the 17q21.31 microdeletion syndrome, we used WGS together with SUN read-depth analysis to narrow breakpoints to intervals of <25 kb (**Fig. 3**; ref. 14).

More recently, we developed a conceptually similar method based on targeted capture of SUN-containing regions using MIPs, and we applied it to resolve breakpoints for NAHR-mediated deletions and duplications of *RHD* (encoding Rh blood group, D antigen) to ~6 kb (**Fig. 4**; ref. 23). We have found that these sequencing-based approaches ultimately refine breakpoints within segmental duplications to the highest attainable sequence-level resolution and compare favorably in time and expense to more conventional approaches.

Applications and limitations

This protocol focuses on breakpoint resolution; however, read-depth analysis using SUNs has more broadly enabled genetic characterization of duplicated genes[22–24]. Specifically, this approach has been used to genotype paralog-specific copy number, to discover structural variation, to detect interlocus gene conversion and to assay paralog-specific gene expression. Although these studies were conducted on human DNA, the method should prove similarly useful, in principle, for large duplicated regions in the genome of any organism. Whether or not sequencing data can be used for any of the mentioned purposes, however, depends on the identification of SUNs. Because the level of attainable breakpoint resolution is determined by these markers, careful analysis of their density and spatial distribution should always serve as an initial step to assess whether using a sequencing approach described here makes sense for a particular case. To that end, we provide tables including the number of 30-bp SUN *k*-mers (SUNKs)[22] in different-sized windows (1, 5, 10, 25 and 50 kb) across the GRCh37 reference genome (http://eichlerlab.gs.washington.edu/breakpoint_protocol_supplementary_data).

Having an accurate genome sequence and understanding haplotypic and copy-number variations are crucial foundations upon which all subsequent steps depend (**Box 1**). Sequences corresponding to breakpoint regions are frequently misassembled, even in the finished human reference genome, and they often contain gaps due to the high sequence identity and highly duplicated nature of sequences therein[24–26]. Furthermore, such sequences are enriched for copy-number polymorphisms and structural variation. All these factors complicate SUN identification and warrant

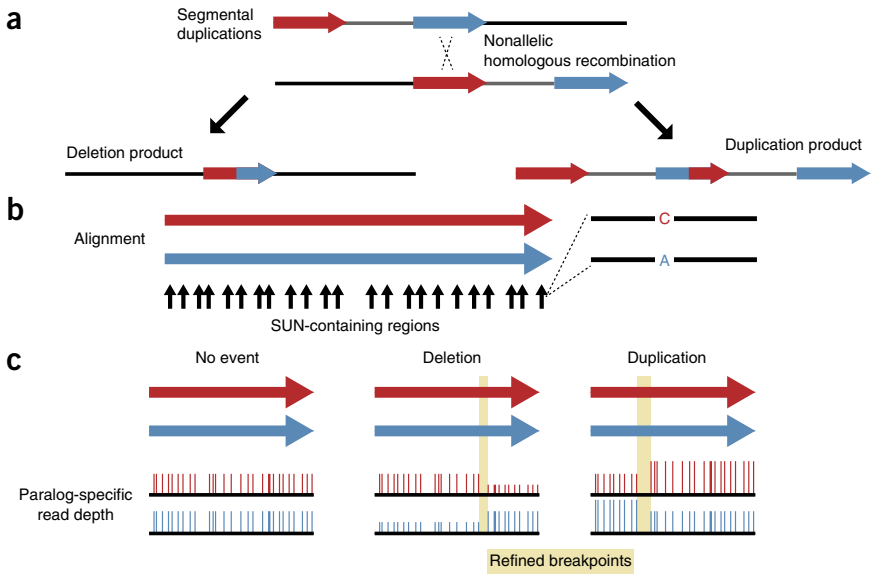

Figure 2 | Sequencing-based breakpoint resolution strategy. (**a**) NAHR between segmental duplications (red and blue arrows) results in the deletion and duplication of the intervening sequence, as well as the proximal end of one of these paralogs and the distal part of the other. (**b**) Alignment of segmental duplication sequences enables the identification of SUNs. (**c**) Quantifying WGS read-depth at each SUN reveals a reciprocal copy-number transition, a signature of NAHR, corresponding to the breakpoint region.

careful consideration before proceeding to study a particular microdeletion or microduplication using this protocol.

In general, the more nearly identical and the higher the total copy of breakpoint-associated segmental duplications, the more difficult breakpoint resolution becomes. The number of SUNs decreases as sequence identity increases, and accurate copy-number prediction grows more challenging as the number of homologous sequences increases. Biological factors can sometimes simplify the procedure. For example, the patients with the 17q21.31 microdeletion syndrome included in our study were all heterozygous for the haplotype on which the microdeletions occurred. As a result, breakpoint resolution amounted to assessing the presence or absence of haplotype-specific SUNs in WGS data[14], a straightforward task compared with inferring deleted regions from a decrease in read-depth over SUNs, which would be necessary if an individual were homozygous for a single haplotype. As a second example, *RHD* deletion and duplication breakpoints were readily refined[23] because the breakpoint-containing paralogs are not present at high total copy number at many locations throughout the genome. Other biological factors, particularly frequent interlocus gene conversion between associated segmental duplications, sometimes complicate accurate breakpoint localization.

As noted above, the precision of breakpoint resolution for recurrent duplications and deletions is inherently limited by the number and spatial distribution of distinguishing variants within associated segmental duplications, and it depends upon where the breaks occur relative to these variants. Breakpoints occurring at different locations within long stretches of identical sequence

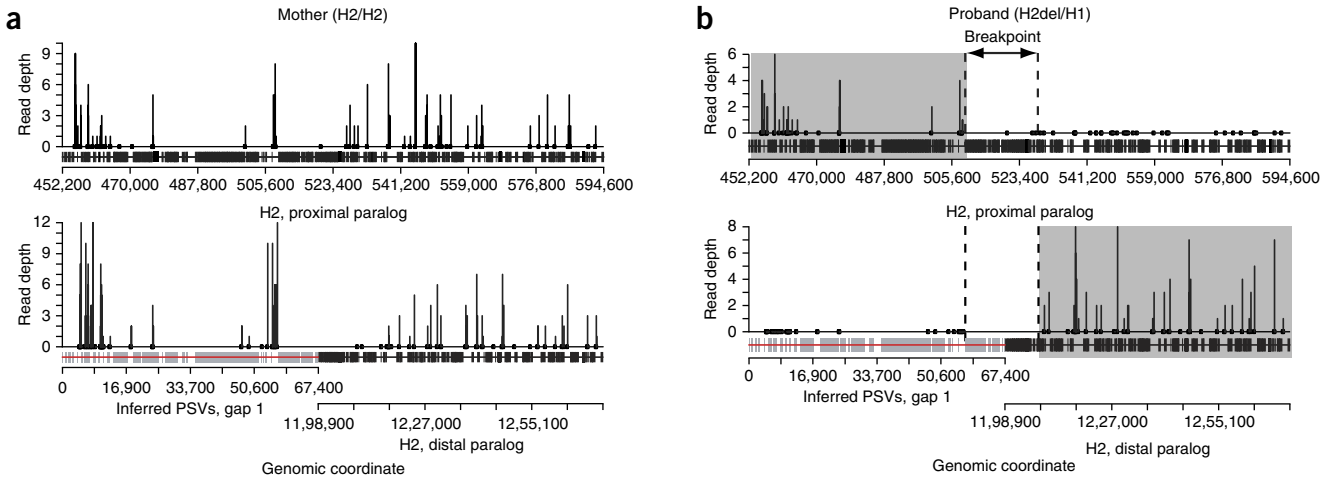

Figure 3 | Resolution of 17q21.31 microdeletion breakpoints using WGS data. (**a**,**b**) Read-depth (vertical lines) at breakpoint-informative paralogous sequence variants (PSVs, dots) is shown over an alignment of the paralogous segmental duplications mediating the microdeletion for the patient's mother (**a**), who lacks the microdeletion, and for the patient (**b**). Structural haplotypes for both chromosomes for each individual are given in parentheses. Variation in read-depth between informative variants occurs even in the absence of any copy-number variation at these loci. However, because the patient is heterozygous for the deletion-bearing haplotype, a read-depth of zero at informative variants is observed over deleted regions, allowing the breakpoint to be unambiguously refined to an ~22-kb window (between dashed lines). Note that a region of sequence (red line) of ~70 kb in size was missing from the reference genome, and it had to be resolved before sequencing data could be accurately interpreted (**Box 1**). Adapted with permission from *The American Journal of Human Genetics*, **90**, Itsara A. *et al.*, Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing, pages 599–613 (2012), with permission from Elsevier (ref. 14).

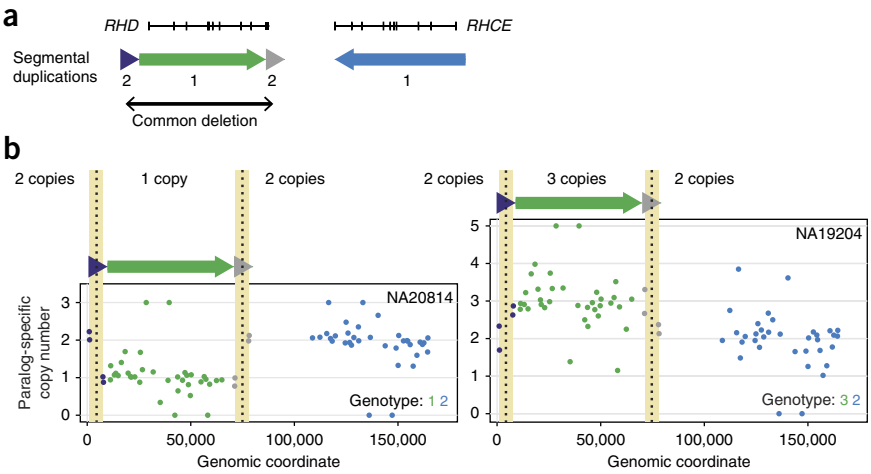

Figure 4 | Resolution of NAHR-associated *RHD* duplication and deletion breakpoints by MIP capture and sequencing. (**a**) NAHR between ~9-kb segmental duplications (dark blue and gray triangles) flanking *RHD* results in its deletion and duplication. (**b**) 39 MIPs were used for genotyping the copy number of *RHD* and flanking segmental duplications in diploid DNA from HapMap individuals NA20814 (having a duplication of *RHD*) and NA19204 (having a duplication of *RHD*). Each point indicates a paralog-specific copy-number estimate, which is calculated at each locus as the product of the paralog-specific relative read-depth and the aggregate estimated copy number. Data from the four MIPs targeting SUN-containing regions in the flanking segmental duplications refine the breakpoints to ~6-kb intervals (yellow highlights). *RHCE*, gene encoding Rh blood group CcEe antigens. Figure modified with permission from Nuttle *et al.*[23], Nature Publishing Group.

between associated segmental duplications will be indistinguishable, as deletion or duplication products of NAHR anywhere in such regions all have identical sequences. Thus, although sequencing-based analysis enables pinpointing breakpoints to one of these identical stretches—providing the highest attainable breakpoint resolution for any recurrent event—in some cases, inferred breakpoint intervals will remain large because there are no additional informative markers to refine the region further.

Experimental design

Breakpoint mapping generally involves discovery and confirmation of a duplication or deletion, refinement of breakpoint locations and validation. Multiple approaches are available for completing each of these stages, and we consider several refinement strategies here. Although massively parallel sequencing is not the only technology that is applicable to breakpoint refinement for recurrent events, sequencing is ultimately required to obtain sequence-level resolution. Here we overview several breakpoint localization methods, offer suggestions regarding their application in different situations and diagram their relationships to one another in the context of discovery, refinement and validation (**Fig. 1**).

Array CGH (Steps 1–3 and 9–10). A conventional first step in resolving breakpoints is performing array CGH by using a custom-designed, high-density oligonucleotide microarray with probes spanning the region of interest and extending beyond the hypothesized breakpoint locations. This analysis provides useful qualitative information about events of interest, confirming them, defining minimal affected regions and suggesting particular segmental duplications that are likely to harbor breakpoints. In a best-case scenario, initial array CGH will refine breakpoint locations to specific segmental duplications—probe cross-hybridization generally prevents further narrowing breakpoint locations within the duplications. Given the higher resolution of the sequencing-based approaches outlined below, their ability to

Box 1 | Genome sequence and haplotype characterization

To apply sequencing-based components of this protocol, it is crucial to accurately identify SUNs that distinguish breakpoint-associated segmental duplications from one another. How can one tell whether breakpoint-associated sequences are accurate and complete in a reference genome? Unless sequence gaps exist in these regions, evaluating the reference genome with regard to these criteria is challenging. We have found a variety of strategies particularly useful for this purpose: (i) mapping end sequences from clones from large-insert genomic libraries to the reference genome and searching for discordances between mapping locations and the known insert size range[36,37]; (ii) performing fluorescent *in situ* hybridization experiments using probes targeting breakpoint regions and surrounding loci; and (iii) generating and comparing copy-number profiles across breakpoint regions for hundreds of individuals from populations around the globe using massively parallel WGS data[22]. Collectively, these analyses and experiments provide some insight into the variability and complexity of the particular breakpoint regions in question and inform the researcher on whether their further genomic characterization and high-quality sequencing are necessary to develop an alternate reference genome.

In cases where the reference genome is inaccurate or incomplete, we recommend sequencing bacterial artificial chromosomes (BACs) from genomic libraries by using Sanger[24] or Pacific Biosciences technology[38] and assembling a contig for each distinct structural haplotype. BAC libraries from hydatidiform mole source material, containing a single haplotype and lacking allelic variation, are especially useful for high-identity duplications in the human genome[24]. High-quality sequence characterization of a particular set of breakpoint-associated regions with this approach takes from one to several months to complete. Nevertheless, this process is fundamental for accurate SUN identification and the eventual interpretation of sequencing data. Thus, genomic characterization is the starting point for any breakpoint refinement effort whenever the reference genome fails to provide accurate sequence data capturing the full diversity and complexity of breakpoint-associated loci.

assay several pairs of candidate breakpoint-associated segmental duplications in a single experiment and the high cost of custom array design, array CGH is not necessary for breakpoint mapping. However, array CGH provides orthogonal data that are often useful when validating results or when analyzing sequencing data fails to define clear breakpoint intervals.

Haplotype isolation (Steps 4–8). One strategy to achieve breakpoint resolution begins with the generation of human-mouse somatic cell hybrids to physically isolate the chromosome harboring the duplication or deletion and, separately, its parental progenitor[14,15,27]. Isolation of these chromosomes in somatic cell hybrids simplifies breakpoint localization for any method because afterward confounding effects of unaffected homologs need not be considered. This strategy can be leveraged to perform an array CGH experiment comparing the affected and parental chromosomes or to simplify the analysis of sequencing data. For example, we determined that two distinct types of NAHR accounted for the three deletions that we observed in patients with the 17q21.31 microdeletion syndrome. We reached that conclusion on the basis of observing two distinct patterns of haplotype-specific array CGH data[14]. With regard to sequencing data analysis, SUNs can be assessed for the presence or absence rather than relative read-depth when the input library is derived from any chromosome in isolation from its homolog.

Isolation of at least the affected chromosome is, therefore, in theory an optimal early step in any breakpoint mapping effort. However, we recognize that generating somatic cell hybrids is a time-consuming, expensive and highly specialized process, demanding specific expertise and experience, and it is not amenable to large-scale application. Therefore, haplotype isolation–based breakpoint resolution has been applied only for small numbers of cases in which the benefits described above promise considerable improvement in the quality of the results obtained over other options not requiring haplotype isolation. An emerging alternative to creating somatic cell hybrids for achieving haplotype isolation with greater efficiency is subhaploid complexity reduction and sequencing, for instance, with fosmids or *in vitro* dilution[28,29].

Any current haplotype isolation method remains very expensive and time-consuming relative to sequencing, and thus we recommend pursuing such a strategy only if WGS or targeted sequencing after one of the approaches described below fails to define breakpoint intervals. In such cases, haplotype isolation may prove worthwhile, particularly if the individual is homozygous for the structural haplotype on which the event occurred and if the breakpoint-associated segmental duplications have many paralogs throughout the genome—situations that are more challenging than those that we faced in the 17q21.31 study and our *RH* gene family example. Note that a failure of the sequencing-based approaches could indicate inaccuracy of the genomic sequences used for SUN identification, or their irrelevance in that particular case owing to NAHR having occurred on a structural haplotype distinct from those sequences, with accordingly different SUNs (**Box 1**).

As noted above, performing array CGH or haplotype-specific array CGH would probably yield some biological insight when sequencing strategies fail, particularly if array data can be generated for both unresolved and resolved cases involving the same genomic region and compared in a manner akin to our 17q21.31 analysis[14].

WGS of a trio (Steps 11–20). An alternative to haplotype isolation is high-coverage sequencing of diploid genomic DNA from the affected individual and both parents (a trio) and directly assaying sequence read-depth across the breakpoint-informative regions to define the unequal crossover event[14]. Implementation of this approach enables the researcher to achieve sequence-level breakpoint resolution via bioinformatics analyses. Performing WGS provides a comprehensive view of genetic variation across the genome in addition to refined breakpoints, a feature that is particularly valuable in studies exploring genetic bases for variable expressivity of a recurrent event. One limitation of this strategy is its relatively high cost, which currently precludes its application to patient cohorts of even moderate size.

MIP capture and sequencing (Steps 21–103). The majority of reads from sequencing libraries prepared from genomic DNA or even from an isolated affected chromosome are not useful for breakpoint refinement because they do not map to breakpoint regions or do not contain relevant SUNs. The approach described in this paragraph circumvents this problem by first enriching the sequencing library to obtain high coverage precisely over the regions that are most informative for breakpoint delineation[23]. Targeted capture using MIPs is particularly well suited for this purpose, as ~3,000 MIPs can be combined in a single reaction to simultaneously assay many loci[30–32]. MIPs are short oligonucleotides (70–80 bp) used to capture specific genomic targets <200 bp in length. They have been successfully applied to routinely genotype thousands of samples and are generally cost-effective compared with WGS when used to assay moderate to large numbers of samples.

Despite its advantages, MIP capture is not always preferable to WGS for breakpoint resolution. Not all SUNs are targetable using MIPs. Regions of particularly low or high GC content (<30% or >60%) are often refractory to successful capture[30,31], and high-copy repeats cannot be specifically targeted. As currently implemented, analysis of WGS data focuses on absolute read-depth over SUNs, whereas MIP sequence data analysis considers read-depth over SUNs relative to that over captured paralogous targets. This relative analytical framework makes regions of high aggregate copy number difficult to accurately analyze by the MIP method[23]. Furthermore, because such regions often contain very few SUNs, successfully interrogating as many individual SUNs as possible is crucial for breakpoint resolution within these regions. WGS, in principle, yields data for all SUNs, including those that MIPs would not capture. In general, however, we believe that MIP-based breakpoint resolution will prove more broadly useful than its counterpart based on WGS, at least until the cost of the latter approach becomes less prohibitive.

MATERIALS

REAGENTS

- Custom oligonucleotide microarrays (Agilent Technologies)
- Array CGH reagents[16]
- Microsatellite genotyping reagents[33]
- Somatic cell hybrid reagents[15]
- Molecular inversion probes (MIPs; Integrated DNA Technologies)
- T4 DNA ligase buffer with 10 mM ATP (New England BioLabs, cat. no. B0202S)
- T4 polynucleotide kinase (New England BioLabs, cat. no. M0201L)
- Ampligase buffer (Epicentre, cat. no. A1905B)
- 10 mM dNTP mix (Roche NimbleGen, cat. no. 11581295001)
- Hemo Klentaq (New England BioLabs, cat. no. M0332L)
- Ampligase (Epicentre, cat. no. A0110K)
- Nuclease-free water (Ambion, cat. no. AM9906)
- Genomic DNA from individuals to be analyzed **! CAUTION** All human genetic studies must be approved by an institutional review board, and all participating subjects must provide informed consent.
- Exonuclease I (New England BioLabs, cat. no. M0293L)
- Exonuclease III (New England BioLabs, cat. no. M0206L)
- 2× iProof PCR master mix (Bio-Rad, cat. no. 172-5311)
- Primer SLXA_PE_MIPBC_FOR (100 µM; **Supplementary Table 1**, Operon)
- SYBR Green (Life Technologies, cat. no. S-7563, dilute from 10,000× to 100× in DMSO)
- Reverse barcode primers (10 µM; **Supplementary Table 1**; Integrated DNA Technologies)
- Magnetic beads (Agencourt, Beckman Coulter, cat. no. A63881)
- Ethanol 200 proof (Decon Laboratories, cat. no. 2716)
- Deionized water (Milli-Q)
- Elution buffer (Qiagen, cat. no. 19086)
- Sequencing primers (100 µM; **Supplementary Table 1**; Operon)
- HiSeq or MiSeq kit (Illumina, v2 PE 300 cycles)

EQUIPMENT

- 1.7-ml Eppendorf tube incubators (VWR)
- Hybridization oven (Shel Lab)
- Microarray scanner (Agilent Technologies)
- Agilent feature extraction software
- Microsatellite genotyping equipment[33]
- Somatic cell hybrid equipment[15]
- Microsoft Excel software
- Microcentrifuge (Eppendorf)
- 1.7-ml PCR tube minicentrifuge (Fisher Scientific)
- 96-well cold blocks (Eppendorf, cat. no. Z606634)
- Lightcycler (Bio-Rad)
- Optical qPCR tubes, eight-strip (Bio-Rad, cat. no. TLS-0851)
- Optical qPCR caps, eight-strip (Bio-Rad, cat. no. TCS-0803)
- 200-µl eight-strip PCR tube minicentrifuge (Fisher Scientific)
- Magnet tube rack (Life Technologies, MagnaRack, cat. no. CS15000)
- eGel system (Life Technologies, cat. no. G6512ST)
- Qubit DNA quantification system (Life Technologies, cat. no. Q32871)
- Illumina MiSeq, HiSeq 2000 or HiSeq 2500 sequencer
- Hardware (64-bit computer running Linux with at least 5 GB RAM—a high-memory, multicore computer is best)
- Software (**Box 2**)

REAGENT SETUP

MIPs Order MIPs with the following specifications: a scale of 25 nmol, with standard desalt purification, in deep-well plates, shipped wet-frozen, with a full yield per well, at a concentration of 100 µM in IDTE buffer (1× TE buffer) at pH 8. Upon their arrival, store the MIPs at 4 °C (for up to ~1 year) or −80 °C (for long-term storage).

Reverse barcode primers Order reverse barcode primers (**Supplementary Table 1**) with the following specifications: a scale of 25 nmol, with standard desalt purification, in deep-well plates, shipped wet-frozen, with a full yield per well, at a concentration of 100 µM in buffer/IDTE (1× TE buffer) at pH 8. Prepare each working reverse barcode primer plate by adding 20 µl of each reverse barcode primer and 180 µl of elution buffer to each well of a new 96-well plate and mixing thoroughly by pipetting up and down.

EQUIPMENT SETUP

WGS Depending on the desired turnaround time, either the Illumina HiSeq 2000 or the HiSeq 2500 can be used for sequencing (see http://www.illumina.com/systems/sequencing.ilmn and http://res.illumina.com/documents/systems/hiseq/datasheet_hiseq_systems.pdf for comparisons of various sequencing platforms). The final sequence coverage for each individual should be >15× (follow standard protocols with 101-bp paired-end reads and 300–500-bp insert libraries). This level of coverage ensures that several breakpoint-informative sequence reads will be obtained. Sequencing runs should include forward and reverse reads of at least 100 bp each, as well as an 8-bp index read. Sequences of the index reads will be the reverse complements of sample barcode sequences incorporated into the sequencing library during library preparation.

MIP sequencing Depending on the desired coverage per MIP target and the desired turnaround time, any of the Illumina MiSeq, HiSeq 2000 or HiSeq 2500 instruments can be used for sequencing (see http://www.illumina.com/systems/sequencing.ilmn and http://res.illumina.com/documents/systems/hiseq/datasheet_hiseq_systems.pdf for comparisons of various sequencing platforms). For a pool of ~2,000 MIPs, up to 192 samples can be pooled and sequenced to good coverage on a single lane of HiSeq 2000 with 101-bp paired-end reads and an 8-bp index read. With the reverse barcode primers provided here (**Supplementary Table 1**), up to 384 samples can be pooled and analyzed in a single run. Sequencing runs should include forward and reverse reads of at least 100 bp each, as well as an 8-bp index read. Sequences of the index reads will be the

Box 2 | Software setup

Custom analysis programs.

Download all custom analysis programs from GitHub (http://github.com/xnuttle/breakpoint_resolution_wgs_mips) and save them in a single directory, for example, '/software/brkpt/'. Define the environmental variable BRKPT_SOFTWARE by using the command below or including it as a line in your bash profile, replacing the example path shown here with the actual path to the directory where the custom analysis programs have been saved:

```
$ export BRKPT_SOFTWARE=/software/brkpt
```

mrFAST

Download mrFAST[35] version 2.6.0.0 from SourceForge (http://mrfast.sourceforge.net/). Refer to the mrFAST user manual (http://mrfast.sourceforge.net/manual.html) for detailed instructions on setup and use.

R

Download R version 2.15 from the R Project website (http://www.r-project.org). Install the R package 'ggplot2' by running the following command in the R console:

```
> install.packages("ggplot2")
```

reverse complements of sample barcode sequences incorporated into the sequencing library during library preparation. Sequencing primers are reported in **Supplementary Table 1**.

Hardware Many programs benefit from parallelization in a parallel computing environment, such as a high-performance Linux-based cluster integrated with network-available storage. Specifically, we use a Linux-based high-performance cluster with 110 nodes with an aggregate 1048 CPU cores. We have 491 terabytes (TB) of usable network-available storage, a mix of EMC storage area network (SAN)-based storage (22 TB), a CORAID storage server (48 TB), three large Sun Microsystems storage servers (131 TB) and three Dell SAS servers (290 TB). To facilitate the rapid analysis of data across systems, all storage can be made available to all cluster nodes, application servers and desktop systems. The cluster queuing system is Sun Grid engine 6.1.

PROCEDURE

Targeted array CGH ● TIMING 2–3 d (plus the time it takes to receive microarrays)

▲ CRITICAL Many steps in the PROCEDURE specify using a centrifuge or a microcentrifuge to 'spin down' a sample (contained in plates or tubes), but they do not provide information regarding the centrifugation speed, duration and temperature. For these steps, the exact centrifugation speed, duration and temperature do not matter, so long as the centrifugation effectively ensures that all liquids in plates or tubes collect at the bottom of the wells or tubes.

▲ CRITICAL The implementation of this subsection of the PROCEDURE (Steps 1–3) is optional (see Experimental design for a relevant discussion on its suggested implementation).

1| Design a custom oligonucleotide microarray covering the region of interest by using the Agilent eArray design suite (http://www.genomics.agilent.com/en/product.jsp?cid=AG-PT-122&tabId=AG-PR-1047&_requestid=1207000). Probes should be 60 bp in length, and they should be designed at a high density over the target region (approximately one probe every 900 bp). To ensure adequate coverage of the region of interest, probe design should cover a larger region including at least 25 kb of flanking sequence on each side of the breakpoint-associated segmental duplications. The final probe set should include a substantial fraction of probes targeting unique regions outside of the region of interest (e.g., a genomic backbone), as such regions facilitate calibration of diploid state (see Step 3 below). Order the oligonucleotide arrays.

■ PAUSE POINT Ordered custom oligonucleotide arrays take several weeks to months to arrive.

2| Perform array CGH experiments using a sample from the individual whose DNA has the duplication or deletion under investigation and a sample from an individual whose DNA does not to approximate regions where breakpoints occur. Provided that DNA from the affected individual and both parents is available in sufficient quantity for these experiments (250 ng per individual per hybridization) and desired follow-up experiments, we recommend performing three separate array CGH experiments (comparing the affected individual to the mother, the affected individual to the father and the mother to the father). A protocol detailing the array CGH procedure has been previously published[16].

3| Analyze the array CGH data with the Agilent feature extraction software (http://www.genomics.agilent.com/en/Microarray-Scanner-Processing-Hardware/Feature-Extraction-Software/?cid=AG-PT-144&tabId=AG-PR-1050) and custom analysis programs to generate plots of $\log_2$ fluorescence intensity ratios across the spatial extent of the targeted region. For quality control, we recommend following the manufacturer's instructions, ensuring that the derivative log ratio spread is <0.23 per sample. To highlight probes signaling a deletion or a duplication, use different colors to visualize points with $\log_2$ ratios of >1.5 s.d. from the mean $\log_2$ ratio in the experiment. If probes in the region of interest constitute a substantial fraction of the total probe set (>5%), exclude these probes when calculating the mean and s.d. of $\log_2$ ratios. Note that the signals from deletions and duplications affecting duplicated sequences will not be as strong as those observed for such events affecting unique sequences because the relative loss or gain of DNA is smaller for the sequences originally present at higher copy numbers.

Isolation of the affected chromosome and its parental progenitor(s) ● TIMING 2–3 months

▲ CRITICAL The implementation of this subsection of the PROCEDURE (Steps 4–8) is optional (see Experimental design for a relevant discussion on the merits of implementing it).

4| Genotype DNA from the affected individual, as well as from both parents, using microsatellite markers (Marshfield map[34]) along the chromosome of interest, including multiple markers in proximity to the deletion or duplication locus (within ~5 Mb). The Marshfield map is a collection of short-tandem repeat markers developed in the 1990s to map human genetic traits, integrate physical mapping data and assess patterns of recombination[34]. Obtaining genotypes at these markers will enable the experimenter to infer the haplotypes of the affected individual and of both parents. A protocol detailing microsatellite genotyping has been previously published[33].

5| Generate human-mouse somatic cell hybrids, using standard materials and protocols, by performing electrofusion of human Epstein-Barr virus–transformed lymphoblast cells with mouse E2 cells and expanding transformants for 18 d.

The specific protocol that we use is described by Highsmith *et al.*[15]. Please note that the materials and equipment necessary to implement this step are also reported in the study by Highsmith *et al.*[15].

6| Genotype the somatic cell hybrid colonies (50–100) for the same microsatellite markers as above, according to published protocols[15,33]. Generating somatic cell hybrids yields several colonies, most of which do not contain the affected chromosome or a parental homolog in isolation. Genotyping microsatellites from several colonies enables the experimenter to identify these colonies of interest, and it provides insight into the integrity of the affected and parental homologous chromosomes that they contain.

7| Genotype a denser panel of microsatellite markers, including several in close proximity (~1 Mb) to the deletion or duplication event, using DNA from a single colony that has the affected chromosome intact and in isolation and from four single colonies each having one parental homolog intact and in isolation. Follow the same published protocols as above[15,33]. The genotyping results will provide insight into the specific parental chromosome or chromosomes involved in the NAHR event, recombination patterns and the timing of the NAHR event in meiosis (i.e., meiosis I or II).

8| Propagate a single colony harboring the intact, isolated affected chromosome and a single colony harboring the intact, isolated parental homolog involved in NAHR (determined in the previous step). If two parental homologs were involved in the deletion or duplication event (i.e., NAHR was interchromosomal rather than interchromatidal), propagate at least three single colonies, one harboring each intact, isolated parental homolog involved in NAHR and one harboring the intact, isolated affected chromosome. Follow the published protocol[15] to propagate relevant colonies, and isolate DNA from each colony that can be used for haplotype-specific array CGH or sequencing experiments.

Haplotype-specific array CGH ● TIMING 2–3 d

▲ **CRITICAL** The implementation of this subsection of the PROCEDURE (Steps 9–10) is optional, but Steps 9 and 10 can only be performed after the completion of Steps 4–8 (see Experimental design for details and discussion).

9| Implement Steps 1–3 described above by using DNA from one expanded somatic cell hybrid colony harboring the affected chromosome as test and another harboring its progenitor as reference to obtain $\log_2$ fluorescence intensity ratios. Provided that NAHR was interchromatidal in origin (the most typical case), this experiment allows for the direct comparison of the affected chromosome with an effectively isogenic background over the region of interest. This comparison is possible because the rearranged chromosome differs from the parental donor chromosome primarily as a result of the duplication or deletion event.

10| Assuming a model of NAHR, identify all directly orientated segmental duplications by using the 'segmental duplications' track on the University of California Santa Cruz (UCSC) Genome Browser (http://www.genome.ucsc.edu/) or the whole-genome assembly comparison pipeline[3], if sequences of interest are not accurate or complete in the reference genome (**Box 1**). This pipeline defines all segmental duplications over 1 kb in length and having >90% sequence identity in a genome. For all duplication pairs, analyze $\log_2$ ratio patterns over the spatial extent of the targeted region[14]. Comparing the observed $\log_2$ ratios with expected $\log_2$ ratios under different hypothesized NAHR scenarios enables the experimenter to define candidate breakpoint-harboring duplication pairs and eliminate others from consideration.

Massively parallel WGS and SUNK analysis ● TIMING variable; 1–3 weeks, depending on the sequencing platform

▲ **CRITICAL** Please note that the implementation of this subsection of the PROCEDURE (Steps 11–20) is an alternative to the MIP-based approach (Steps 21–103, see Experimental design for relevant discussion).

11| Sequence libraries prepared from genomic DNA or haplotype-resolved chromosomes from the affected individual and both parents on an Illumina HiSeq according to the manufacturer's instructions and specifications detailed above (see WGS under Equipment Setup). Please note that sequencing will take anywhere from ~1 d to multiple weeks to complete, depending on the platform used. Once the sequencing run is complete, data should be stored and backed up before beginning the analysis. Please note as well that the first few steps of the analysis (Steps 12–15 below) do not require the sequencing data and can be completed while you are waiting for the sequencing run to finish.

12| Obtain paralogous breakpoint-associated sequences and align them using an alignment program such as ClustalW2 (http://www.clustal.org/clustal2/). Make two fasta files from the alignment output, one containing the first aligned sequence and the other containing the second aligned sequence. These fasta files should include '-' characters within the nucleotide sequences at positions corresponding to alignment gaps. Both unaligned and aligned breakpoint-associated sequences must start with the first base in the alignment and end with the last base in the alignment. Name the unaligned sequences 'prox.fasta' and 'dist.fasta' (corresponding to the proximal and distal breakpoint-associated segmental duplications

mediating the rearrangement) and the aligned sequences 'prox_aligned.fasta' and 'dist_aligned.fasta'. Make sure that the names of the sequences correspond to the file names (e.g., the file 'prox.fasta' should have '>prox' as its first line). Save all sequences in the same directory and name this directory 'brkpt_WGS', henceforth referred to as the project directory.

13| Determine the reference sequence coordinates corresponding to the contig sequences—specifically, the reference coordinates of the first and last bases in the 'prox.fasta' and 'dist.fasta' files. Create a tab-delimited text file in the project directory detailing this information, with the chromosome name (i.e., 'chr1') in the first column, the base-1 start coordinate in the second column and the base-1 end coordinate in the third column. Name this file 'seqs.refcoords'. These regions must be listed in order of their reference genomic coordinates.

14| Identify breakpoint-informative SUNKs (36 bp) and SUNs by running the script 'wgs_analysis_pt1.sh' from the project directory on a high-memory machine. This program will generate the text files 'brkpt.suns', 'brkpt.sunks' and 'brkpt.sunsunks'. It requires high memory (for an example test run, the 'top' command showed that VIRT was 52.0*g* and RES was 27*g*) to run and takes several hours to finish; we thus recommend running it overnight. Please note that the following Step 15 can be completed before this step finishes and before the sequencing run is complete.

```
$ bash $BRKPT_SOFTWARE/wgs_analysis_pt1.sh
```

? TROUBLESHOOTING

15| Create a tab-delimited file listing the names of all samples pooled in the sequencing run in the first column and their corresponding barcode reverse complement sequences in the second column. Name this file 'brkpt.barcodekey' and save it in the project directory.

16| After the sequencing run has completed, follow the manufacturer's instructions regarding bcl conversion to convert raw sequencing base call data to qseq text files. Make a new directory in the project directory called 'raw_qseq_files' and store the qseq text files in this new directory. Do not compress these files: running the script in the next step will do that.

17| Change into the 'raw_qseq_files' directory and run the script 'wgs_analysis_pt2.sh' to generate gzipped fastq files that will be searched for breakpoint-informative SUNKs. These files will be generated in the 'raw_qseq_files' directory and moved to a directory within the parent directory called 'fastqs'.

```
$ cd raw_qseq_files
$ bash $BRKPT_SOFTWARE/wgs_analysis_pt2.sh
```

18| Quantify read-depth over breakpoint-informative SUNs by running the script 'wgs_analysis_pt3.sh' from the 'raw_qseq_files' directory. Directories for each sample in the 'brkpt.barcodekey' file will be created within the 'fastqs' directory. In each of these sample directories, the final output is written to the file 'brkpt.suns.depth', with the last column of this file showing the observed read depth over each breakpoint-informative SUN.

```
$ bash $BRKPT_SOFTWARE/wgs_analysis_pt3.sh
```

19| Analyze and visualize the data in R. To view data for a single individual, copy that individual's 'brkpt.suns.depth' file and the '$BRKPT_SOFTWARE/pdf_brkpt_WGS.r' file to a directory that R can access; open R and set the R working directory to that directory. Next, run the following commands in the R console to generate a file with a name like 'sample_brkpt.pdf', written to that directory. Replace 'sample' in the first command below with the name of the individual:

```
> indiv<-"sample"
> source("pdf_brkpt_WGS.r")
```

20| Manually inspect the pdf file with a name like 'sample_brkpt.pdf' showing read-depth data over breakpoint-informative SUNs for the individual of interest. Breakpoint signatures should be apparent as a decrease or increase of paralog-specific read-depth over the extent of the deletion or duplication (e.g., **Fig. 2c**; see ANTICIPATED RESULTS for further discussion).

MIP design ● TIMING 1 d, plus 1–2 weeks to receive oligonucleotides

▲ CRITICAL As mentioned above, Steps 21–103 below should be performed as an alternative to the WGS approach (Steps 11–20, see Experimental design for relevant discussion).

21| Obtain paralogous breakpoint-associated sequences and align them using an alignment program such as ClustalW2 (http://www.clustal.org/clustal2/). Make two fasta files from the alignment output, one containing the first aligned sequence and the other containing the second aligned sequence. These fasta files should include '-' characters within the nucleotide sequences at positions corresponding to alignment gaps. Both unaligned and aligned breakpoint-associated sequences must start with the first base in the alignment and end with the last base in the alignment. Name the unaligned sequences 'prox.fasta' and 'dist.fasta' (corresponding to the proximal and distal breakpoint-associated segmental duplications mediating the rearrangement) and the aligned sequences 'prox_aligned.fasta' and 'dist_aligned.fasta'. Make sure that the names of the sequences correspond to the file names (e.g., the file 'prox.fasta' should have '>prox' as its first line). Save all sequences in the same directory and name this directory 'brkpt_MIPs', henceforth referred to as the project directory.

22| Generate initial MIP designs. The script 'mip_design_pt1.sh' calls several programs to design an initial set of MIPs targeting breakpoint-informative SUNs, detailed in the output file 'brkpt.mipdesign'. It typically takes 30 min–2 h to run. Run this script from the project directory:

```
$ bash $BRKPT_SOFTWARE/mip_design_pt1.sh
```

23| Import data from the tab-delimited text file 'brkpt.mipdesign' into Microsoft Excel, so that each column in the file is imported into a separate column in the spreadsheet and the data begin in position A1. The first line of the file 'prox.mipdesign' details the meaning of the data in each of the columns except for the last column, which contains the oligo sequences for all MIPs initially designed.

24| Sort the Excel spreadsheet by column S and delete all rows having a value in column S that includes 'snp'. This action will ensure that all remaining MIP designs have hybridization arms targeting sequences that are identical between both breakpoint-associated sequences.

25| Sort the Excel spreadsheet by column G, which contains the total number of copies of the extension hybridization arm sequence found in the human genome. Delete all rows having values in this column greater than a threshold cutoff. The exact value of this cutoff will differ on a case-by-case basis, but a good general rule to follow is that the cutoff should be approximately two times the number of copies of breakpoint-associated duplicated sequences in the haploid genome of interest. For example, if the breakpoint-associated sequences have no paralogous sequences elsewhere in the genome, this cutoff should be set to around 4. Increasing this cutoff will increase the number of MIPs designed and potentially the spatial resolution of refined breakpoints, but it may result in more off-target MIP capture events. Columns L–M contain the alignment coordinates of regions targeted by the remaining MIPs, so sorting by column L and examining the values in columns L–M in comparison with the length of the aligned sequence 'prox_aligned.fasta' provides some sense of the spatial resolution afforded by the remaining MIPs.

26| Sort the Excel spreadsheet by column K, which contains the total number of copies of the ligation hybridization arm sequence found in the human genome, and then delete all rows having values in the column greater than the same threshold cutoff imposed in Step 25.

27| Sort the Excel spreadsheet by column L to order the remaining MIPs by the alignment coordinates of the regions that they target. Copy all sequences in column N, paste them into a new text file named 'target.seqs' created in the project directory by using a command-line text editor such as vim (http://www.vim.org/) and save the text file. Next, run the following command from the project directory to calculate GC content for all remaining MIP target regions. This information will be taken into account by a later program that selects a set of MIPs that have good potential for successfully capturing targets harboring breakpoint-informative SUNs:

```
$ $BRKPT_SOFTWARE/calculate_target_GC target.seqs > target.gc
```

28| Import data from the text file 'target.gc' into column X of the Excel worksheet such that the data begin in position X1. Cut the data from column X and paste them to the same rows in column S. Copy all data in the spreadsheet, paste them into a new text file named 'brkpt.filtered.mipdesign' created in the project directory by using a command-line text editor such as vim (http://www.vim.org/) and save the text file. Next, run the script 'mip_design_pt2.sh' from the project directory to generate a file containing the final MIP oligos to order:

```
$ bash $BRKPT_SOFTWARE/mip_design_pt2.sh
```

29| Order MIPs (see Reagent Setup for order specifications). The final set of MIP oligonucleotides to order is specified in the last column of the file 'brkpt.filtered.mippicks' in the project directory.
■ **PAUSE POINT** Ordered MIPs typically take 1–2 weeks to arrive.

MIP pooling and 5′ phosphorylation ● TIMING 2–3 h
30| On the same day that the ordered MIPs are received in plates, allow these plates to thaw in a refrigerator at 4 °C .

31| Remove thawed MIP plates from the refrigerator (maintained at 4 °C), and spin them down in a centrifuge.

32| For each MIP plate, pool 5 µl of each MIP into a 1.7-ml Eppendorf tube. This pooling can be easily accomplished with an eight-channel 0.5–10-µl mechanical pipette to first pool MIPs from each row of a plate into an eight-tube strip of 200-µl PCR tubes and then pooling the contents of each PCR tube into an Eppendorf tube.

33| Pool together the MIP pools for each plate by combining P µl of each plate pool into a new 1.7-ml Eppendorf tube, where P = 0.1 times the number of MIPs in each plate pool, calculated separately for each plate pool.

34| Add B µl of 10× T4 DNA ligase buffer with 10 mM ATP and K µl of T4 polynucleotide kinase to the same 1.7-ml tube containing the final MIP pool from Step 33. Here B = (5/26) × (volume of MIPs in that pool) and K = (7/26) × (volume of MIPs in that pool).

35| If the total volume in the 1.7-ml tube after Step 34 is <50 µl, add enough nuclease-free water to bring the final volume to 50 µl, then vortex the tube, spin it down in a microcentrifuge and then transfer the tube contents to a 200-µl PCR tube. Otherwise, use enough nuclease-free water to bring the final volume up to V µl, where V is the smallest multiple of 50 above the current total volume; vortex the mixture, centrifuge it as detailed above and then split up the final volume into multiple 200-µl PCR tubes, each containing 50 µl of solution.

36| Phosphorylate the MIPs by incubating the PCR tubes from Step 35 in a thermocycler. Set the thermocycler to run at 37 °C for 45 min, followed by 65 °C for 20 min and then 4 °C indefinitely, with a heated lid at 105 °C. Do not place the tubes in the thermocycler until the block temperature reaches 37 °C.

MIP capture ● TIMING ~1 d
37| After the reactions from Step 36 have completed (the 4 °C stage is reached in the thermocycler), combine all reaction products from Step 36 into a single 1.7-ml Eppendorf tube to form a stock of phosphorylated MIPs.
■ **PAUSE POINT** Phosphorylated MIPs can be stored at 4 °C for up to ~1 year.

38| Calculate the concentration C of each phosphorylated MIP in the stock, where $C = 10/V$ µM, and where V = the final volume of the stock determined in Step 35.

39| Calculate the volume of MIP stock that should be added per MIP capture reaction. 1 ng of genomic DNA contains ~330 haploid genome copies. 200 ng of genomic DNA will be used per reaction; therefore, ~66,000 haploid genome copies will be present per reaction. Each MIP should be present at a ratio of 800 copies per haploid genome copy, so ~52,800,000 copies of each MIP are needed per reaction, or 8.8×10^{-5} pmol. Thus, you will need to add M µl of MIP stock per reaction, where $M = 8.8 \times 10^{-5}/C$, and C = the concentration in µM calculated in Step 38. Because the value of M is typically very low, it is likely to be necessary to dilute the phosphorylated MIP stock with elution buffer to form a working stock solution and recalculate M on the basis of the concentration of the working stock. M should ideally be between 0.1 and 1 µl.

40| Dilute genomic DNA samples to be analyzed to the same concentration S, which should be between 10 and 25 ng/µl.

41| Label an Eppendorf plate with information about the experiment and add U µl of samples from Step 40 to it, where $U = 200/S$, and S is the concentration of each sample from Step 40. Make sure to record the contents of each well for future reference.

42| Remove two 96-well cold blocks from a freezer maintained at –20 °C.

43| Make a working stock of 0.25 mM dNTPs by combining 1 µl of 10 mM dNTP mix with 39 µl of nuclease-free water.

44| Set up a master mix for the MIP capture reactions in a 2-ml Eppendorf tube, keeping the tube and its contents on ice when not adding components. In addition to the DNA to be analyzed, each reaction mixture contains 2.5 µl of Ampligase buffer, *M* µl of working MIP stock (as calculated in Step 39), 0.032 µl of 0.25 mM dNTPs, 0.32 µl of Hemo Klentaq, 0.01 µl of Ampligase and *W* µl of nuclease-free water, where $W = 25 - (U + 2.5 + M + 0.032 + 0.25 + 0.32 + 0.01)$, and *U* is the volume of the sample added per well determined in Step 41. Make enough master mix to prepare more reaction mixtures than the capture reactions to be performed, so as to allow for pipetting error. Preparation of this excess volume is recommended particularly if the experimenter is planning, as we recommend (Step 46), to transfer the master mix first to eight-well 200-µl PCR tubes so that the eight-channel 5–100-µl pipette can be used to add the master mix to the plate with samples.

45| Place the labeled plate with samples into a 96-well cold block.

46| Vortex the master mix, spin it down in a microcentrifuge and add *MM* µl of the master mix to each well of the plate with samples, where $MM = 25–U$, and *U* is the volume of sample added per well determined in Step 41. We recommend first transferring the master mix to eight-well 200-µl PCR tubes (placed in the remaining 96-well cold block) so that the eight-channel 5- to 100-µl pipette can be used to add the master mix to the plate. Mix the master mix with the samples by pipetting up and down a few times with the tips in the wells.

47| Seal the plate with a PCR plate seal.
▲ **CRITICAL STEP** The plate must be sealed very well, particularly along the edges, or some samples will probably evaporate during the capture reaction.

48| Spin down the plate in a centrifuge.

49| Incubate the sealed plate in a thermocycler for ~23 h. Set the thermocycler to run at 95 °C for 10 min, followed by 60 °C indefinitely, with a heated lid at 105 °C. Do not place the tubes in the thermocycler until the block temperature reaches 95 °C.

Exonuclease treatment ● TIMING ~1 h

50| When the capture reactions from Step 49 have nearly completed (~22.5 h after they were placed in the thermocycler), remove the Ampligase buffer from the freezer (–20 °C) and allow it to thaw completely.

51| After the capture reactions from Step 49 have been completed (~23 h after they were placed in the thermocycler), remove two 96-well cold blocks from the freezer (–20 °C) and allow them to thaw for ~5 min.

52| Remove the plate from the thermocycler, place it into a 96-well cold block and spin it down in a centrifuge.

53| Set up a master mix for the exonuclease reactions in a 1.7-ml Eppendorf tube, keeping the tube and its contents on ice when not adding components. In addition to the capture reaction products, each reaction contains 0.5 µl of exonuclease I, 0.5 µl of exonuclease III, 0.2 µl of Ampligase buffer and 0.8 µl of nuclease-free water. Make enough master mix to allow for pipetting error, particularly if you are planning to transfer the master mix first to eight-well 200-µl PCR strip tubes, so that the eight-channel 0.5–10-µl pipette can be used to add the master mix to the plate with samples (recommended).

54| Vortex the master mix, spin it down in the microcentrifuge and add 2 µl of the master mix to each well of the plate with capture reactions. We recommend first transferring the master mix to eight-well 200-µl PCR strip tubes (placed in the remaining 96-well cold block) so that the eight-channel 0.5–10-µl pipette can be used to add the master mix to the plate. Mix the master mix with the capture reactions by pipetting up and down a few times with the tips in the wells.

55| Seal the plate with a PCR plate seal.

56| Spin down the plate in a centrifuge.

57| Incubate the sealed plate in a thermocycler. Set the thermocycler to run at 37 °C for 45 min, followed by 95 °C for 2 min and then 4 °C indefinitely, with a heated lid at 105 °C. Do not place the tubes in the thermocycler until the block temperature reaches 37 °C.
■ **PAUSE POINT** Exonuclease-treated capture reactions can be stored at 4 °C for several days. Nevertheless, we recommend proceeding to quantitative PCR within a day of finishing the exonuclease treatment.

Quantitative PCR ● TIMING ~1 h

58| After the reactions from Step 57 have been completed (the 4 °C stage has been reached in the thermocycler), remove the plate from the thermocycler and spin it down in a centrifuge.

59| Remove a 96-well cold block from the freezer (−20 °C).

60| Remove the reverse barcode primer plate (see Reagent Setup) from the refrigerator (4 °C) and spin it down in a centrifuge.

61| Set up a master mix for eight (or as many samples are being analyzed plus one, if fewer than seven samples are being analyzed) quantitative PCRs in a 1.7-ml Eppendorf tube, keeping the tube and its contents on ice when not adding components. In addition to exonuclease reaction products and reverse primers to be added separately, each reaction contains 12.5 µl of 2× iProof PCR master mix, 0.125 µl of forward primer solution (SLXA_PE_MIPBC_FOR, 100 µM; **Supplementary Table 1**), 0.125 µl of SYBR Green 100× and 6 µl of nuclease-free water. Prepare enough master mix to allow for pipetting error.

62| Put one strip of eight-well optical PCR tubes in the cold block, vortex the master mix, spin it down in a microcentrifuge and add 18.75 µl of master mix to each optical PCR tube.

63| Add 5 µl of exonuclease reaction products from the first column of the exonuclease reaction plate, except for the last row, to the eight-well optical PCR tubes with the eight-channel 0.5–10-µl pipette. Mix the exonuclease reaction products with the master mix by pipetting up and down a few times with the tips in the wells. Add 5 µl of nuclease-free water to the optical PCR tube without added exonuclease reaction product to serve as a negative control, and then mix by pipetting up and down a few times with the tip in the well.

64| Add 1.25 µl of reverse barcode primers (**Supplementary Table 1**) from the first column of the reverse barcode primer plate to the eight-well optical PCR tubes with the eight-channel 0.5–10-µl pipette. Mix by pipetting up and down a few times with the tips in the wells.
▲ **CRITICAL STEP** Take extreme care not to contaminate the different reverse barcode primer solutions with each other—always use new pipette tips when you are working with this plate.

65| Seal the eight-well optical PCR tubes using an eight-cap strip of optical qPCR caps and spin them down in a microcentrifuge.

66| Incubate the sealed plate in a light thermocycler. Set the light thermocycler to run at 98 °C for 30 s, followed by 30 cycles at 98 °C for 10 s, 60 °C for 30 s, 72 °C for 20 s, a plate read and 72 °C for 10 s, followed by 72 °C for 2 min and then 4 °C indefinitely, with a heated lid at 105 °C. Do not place the tubes in the thermocycler until the block temperature reaches 98 °C.

67| Determine approximately how many cycles (*Y*) the quantitative PCRs took until the fluorescence curves reached their plateaus by manually inspecting the fluorescence curves. Also ensure that this value was tightly distributed between individual curves (the highest value should be within two cycles of the lowest value, excluding the value corresponding to the negative control curve, which should plateau far later than all other curves or not plateau at all). Typical values for *Y* are in the 17–24-cycle range, depending on the MIP pool used and the initial amount of DNA added to each MIP capture reaction.
? TROUBLESHOOTING

PCR ● TIMING ~1 h

68| Remove two 96-well cold blocks from the freezer (−20 °C).

69| Remove the reverse barcode primer plate from the refrigerator (4 °C) and spin it down in a centrifuge.

70| Set up a master mix for the PCRs in a 15-ml Falcon tube, keeping the tube and its contents on ice when not adding components. In addition to exonuclease reaction products and reverse primers to be added separately, each reaction contains 12.5 µl of 2× iProof PCR master mix, 0.125 µl of forward primer (SLXA_PE_MIPBC_FOR, 100 µM; **Supplementary Table 1**) and 6.125 µl of nuclease-free water. Make enough master mix to allow for pipetting error, particularly if you are planning to transfer the master mix first to eight-well 200 µl-PCR tubes so that the eight-channel 5–100-µl pipette can be used to add the master mix to the plate with samples (recommended).

71| Label a 96-well PCR plate with information about the experiment, and place it into a cold block.

72| Vortex the master mix, spin it down in a microcentrifuge and add 18.75 µl of the master mix to each well of the plate with capture reactions. It is recommended to first transfer the master mix to eight-well 200-µl PCR tubes (placed in the remaining 96-well cold block) so that the eight-channel 5–100-µl pipette can be used to add the master mix to the plate.

73| Add 5 µl of exonuclease reaction products from each column of the exonuclease reaction plate to the PCR plate with the eight-channel 0.5–10-µl pipette. Mix the exonuclease reaction products with the master mix by pipetting up and down a few times with the tips in the wells.

74| Add 1.25 µl of reverse barcode primers (**Supplementary Table 1**) from each column of the reverse barcode primer plate to the PCR plate with the eight-channel 0.5–10-µl pipette. Mix by pipetting up and down a few times with the tips in the wells.
▲ **CRITICAL STEP** Take extreme care not to contaminate the different reverse barcode primer solutions with each other; always use new pipette tips when you are working with this plate.

75| Seal the PCR plate with a PCR plate seal.

76| Spin down the PCR plate in a centrifuge.

77| Incubate the sealed PCR plate in a thermocycler. Set the thermocycler to run at 98 °C for 30 s, followed by *Y* cycles (where *Y* is the value determined in Step 67) at 98 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s, followed by 72 °C for 2 min and then 4 °C indefinitely, with a heated lid at 105 °C. Do not place the tubes in the thermocycler until the block temperature reaches 98 °C.
■ **PAUSE POINT** PCR products can be stored at 4 °C for multiple days, although we recommend proceeding directly to cleanup (see below).

Cleanup ● TIMING 30 min
78| About 30 min before the reactions from Step 77 have completed (the 4 °C stage has been reached in the thermocycler), remove the magnetic beads from the refrigerator (4 °C) and incubate them at room temperature (~21 °C).

79| After the reactions from Step 77 have been completed (the 4 °C stage has been reached in the thermocycler), remove the PCR plate from the thermocycler and spin it down in a centrifuge.

80| Pool 5 µl of each PCR into a 1.7-ml Eppendorf tube. Pooling can be easily achieved with an eight-channel 0.5–10-µl mechanical pipette to first pool PCR products from each row into an eight-tube strip of 200-µl PCR tubes and then pooling the contents of each PCR tube into an Eppendorf tube.

81| Add 1.8 µl of magnetic beads per microliter of pooled PCR product to the PCR pool. Mix well by pipetting up and down several times.

82| Incubate the PCR product pool with added beads at room temperature for 10 min.

83| Place the tube containing the PCR product pool and magnetic beads into a magnet tube rack and wait for 5 min.

84| Prepare 20 ml of fresh 70% (vol/vol) ethanol from 200-proof ethanol and deionized water in a 50-ml Falcon tube.

85| Taking care not to disturb the beads, while keeping the Eppendorf tube in the magnet tube rack, pipette the clear solution and discard into an empty 100-ml beaker.

86| Add fresh 70% (vol/vol) ethanol to the tube in the magnet tube rack such that the beads are fully covered (1 ml usually works well) and incubate the tube for 30 s.

87| Taking care not to disturb the beads, while keeping the Eppendorf tube in the magnet tube rack, pipette the 70% (vol/vol) ethanol and discard it into an empty 100-ml beaker.

88| Add fresh 70% (vol/vol) ethanol to the tube in the magnet tube rack such that the beads are fully covered (1 ml usually works well) and incubate the tube for 30 s.

89| Taking care not to disturb the beads, while keeping the Eppendorf tube in the magnet tube rack, pipette the 70% (vol/vol) ethanol and discard it into an empty 100-ml beaker.

90| Allow the beads to air-dry for 5 min.

91| Remove the tube with beads from the magnet tube rack and add to it 100 µl of elution buffer, making sure to bring the beads into solution.

92| Mix the tube contents well by pipetting up and down at least ten times.

93| Place the tube with the beads in elution buffer back into the magnet tube rack and wait for 1 min.

94| Taking care not to disturb the beads, while keeping the Eppendorf tube in the magnet tube rack, pipette out the elution buffer (which now contains cleaned-up DNA from the PCRs) and transfer it to a new 1.7-ml Eppendorf tube. Label the tube with details regarding its contents. This tube contains the final library for sequencing (to be performed in Step 96).
■ **PAUSE POINT** Cleaned, pooled PCR products can be stored at 4 °C for months or at −20 °C indefinitely.

Size confirmation by gel electrophoresis ● TIMING 15 min–2 h

95| Run 4 µl of the final library for sequencing on a gel by using the eGel system or by manually setting up an electrophoresis experiment with an agarose gel (1–2% (wt/vol)). Ensure that only a single major band of ~274 bp, corresponding to the desired MIP PCR products, is present.
? TROUBLESHOOTING

Massively parallel sequencing ● TIMING variable; 1 d–3 weeks, depending on the sequencing platform

96| Set up a sequencing run on an Illumina MiSeq or HiSeq according to the manufacturer's instructions and specifications detailed above (see MIP sequencing under Equipment Setup), by using the final library prepared in Step 94 and the sequencing primers. We recommend measuring concentration with the Qubit dsDNA HS assay (included with the Qubit DNA quantification system). Please note that sequencing will take from ~1 d to multiple weeks to complete, depending on the platform used. Once the sequencing run is complete, store and back up data before beginning the analysis. The first step of the analysis (Step 97) does not require the sequencing data and can be completed while waiting for the sequencing run to finish.

Data analysis ● TIMING ~3 h

97| Create a tab-delimited file listing the names of all samples pooled in the sequencing run in the first column and their corresponding reverse barcode primer reverse complement sequences in the second column. Name this file 'brkpt.barcodekey' and save it in the project directory.

98| After the sequencing run has completed, follow the manufacturer's instructions regarding bcl conversion to convert raw sequencing base call data to qseq text files. Make a new directory in the project directory called 'raw_qseq_files' and store the qseq text files in this new directory. Do not compress these files: running the script in the next step will do that.

99| Change into the new directory and run the script 'mip_analysis_pt1.sh' to generate gzipped fastq files that will be used for mapping. These files will be generated in the 'raw_qseq_files' directory and moved to a directory within the parent directory called 'mrfast_mapping_input'.

```
$ cd raw_qseq_files
$ bash $BRKPT_SOFTWARE/mip_analysis_pt1.sh
```

100| Map reads with mrFAST[35] to a custom genome, consisting of chromosomes 'prox' and 'dist' containing the sequences in 'prox.fasta' and 'dist.fasta', respectively. Refer to the mrFAST user manual for detailed mapping instructions. Use the parameters '--pe --max 160 --min 144 -e 4 --discordant-vh --seqcomp --outcomp --maxoea 500'. Gzipped input fastq files can be found in the 'mrfast_mapping_input' directory in the project directory, and mapping output files should be written to the 'mrfast_mapping_output' directory in the project directory.

101| Parse the mapping output gzipped sam files to generate a file containing paralog-specific read counts for each individual-MIP combination ('brkpt.mipcounts') by running the script 'mip_analysis_pt2.sh' from the 'mrfast_mapping_output' directory:

```
$ bash $BRKPT_SOFTWARE/mip_analysis_pt2.sh
```

102| Analyze and visualize the data in R. Copy the 'brkpt.mipcounts' file, the 'brkpt.barcodekey' file, the '$BRKPT_SOFTWARE/pdf_brkpt_MIP.r' file and the '$BRKPT_SOFTWARE/mipplot_brkpt.r' file to a directory that R can access; open R and set the R working directory to that directory. Next, run the following commands in the R console to generate the file 'brkpt.pdf', written to that directory:

```
> base_name<-"brkpt"
> source("pdf_brkpt_MIP.r")
```

103| Manually inspect each page of the pdf file 'brkpt.pdf' corresponding to data for each individual. The breakpoint location can be narrowed to the interval between two MIPs indicating a reciprocal paralog-specific copy-number transition (see **Fig. 4** and ANTICIPATED RESULTS).

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
14	Error message: 'Memory allocation failed!'	Insufficient computing memory	Rerun the script on a high-memory machine. Alternatively, use the less memory-intensive, successfully implemented strategy for identifying SUNKs detailed in our study of patients having the 17q21.31 microdeletion syndrome[14]
67	Fluorescence curves for non-water reactions do not all plateau around the same value	DNA concentrations of samples used were not all within a narrow range	Measure the concentrations of samples used. If they are not all within a narrow range (within a few ng/μl), dilute the samples so that they come to be all within a few ng/μl of each other and perform the MIP capture and exonuclease treatment again with these samples
95	Smaller band present in addition to band of ~274 bp in size	Substantial number of smaller amplicons in PCR products	Perform the cleanup steps again using a smaller ratio than 1.8 (as low as 0.6) for the volume of magnetic beads added per microliter of pooled PCR product. Perform size confirmation by gel electrophoresis to ensure that having followed the modified cleanup protocol effectively removed the smaller band from the final library for sequencing

● TIMING

Steps 1–3, targeted array CGH (optional): 2–3 d, plus time to receive microarrays
Steps 4–8, isolation of the affected chromosome and its parental progenitor(s) (optional): 2–3 months
Steps 9 and 10, haplotype-specific array CGH (optional): 2–3 d
Steps 11–20, massively parallel WGS and SUNK analysis (alternative to Steps 21–103): variable; 1–3 weeks, depending on the sequencing platform
Steps 21–29, MIP design (note that Steps 21–103 are alternative to Steps 11–20): 1 d, plus 1–2 weeks to receive MIP oligonucleotides
Steps 30–36, MIP pooling and 5′ phosphorylation: 2–3 h
Steps 37–49, MIP capture: ~1 d
Steps 50–57, exonuclease treatment: ~1 h
Steps 58–67, quantitative PCR: ~1 h
Steps 68–77, PCR: ~1 h

Steps 78–94, cleanup: 30 min
Step 95, size confirmation by gel electrophoresis: 15 min–2 h
Step 96, massively parallel sequencing: variable; ~1 d–3 weeks, depending on the sequencing platform
Steps 97–103, data analysis: ~3 h

ANTICIPATED RESULTS

The end product of WGS or MIP-based analysis as detailed in the present protocol should be a copy-number profile corresponding to paralog-specific read-depth over informative SUNs that distinguish the two segmental duplications that were involved in an unequal crossover event (NAHR). WGS analysis will yield plots resembling those reported in **Figure 3**. The plot for an individual where an unequal crossover event has occurred should reveal a transition in paralog-specific copy number over the spatial extent of aligned breakpoint-associated sequences (e.g., **Fig. 3b**). The region lacking breakpoint-informative SUNs where the transition in copy number occurs defines the boundary or the breakpoint interval. Perfect sequence identity between breakpoint-associated segmental duplications over this region prohibits the breakpoint interval from being refined any further. Notably, paralog-specific read-depths of zero will only be observed in a few specific cases: when individual chromosomal haplotypes have been isolated, or when the proband has a deletion and is heterozygous for the structural haplotype on which the deletion occurred.

MIP analysis yields similar results to those obtained from WGS analysis, including a plot showing paralog-specific read-count frequencies on the *y* axis and the alignment coordinate on the *x* axis. As in the WGS analysis, the plots will show two patterns of paralog-specific relative read-depth over the spatial extent of aligned breakpoint-associated sequences, separated by a region lacking breakpoint-informative SUNs that defines the breakpoint interval. This plot will differ slightly from the example in **Figure 4** in that it will show relative rather than absolute paralog-specific copy-number estimates, and it will use alignment coordinates rather than genomic coordinates.

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

ACKNOWLEDGMENTS We thank J. Huddleston for assistance in testing analysis software and preparing it for public access; P. Sudmant for analyzing the number of SUNKs in windows across the reference genome; and T. Brown for assistance with manuscript preparation. X.N. is supported by a National Science Foundation Graduate Research Fellowship under grant no. DGE-1256082. This work was supported by US National Institutes of Health grants HG004120 and HG002385 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS X.N., A.I., J.S. and E.E.E. developed the protocol. X.N. and E.E.E. wrote the paper, with input and approval from all coauthors.

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

1. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
2. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
3. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
4. Stankiewicz, P. & Lupski, J.R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
5. Lupski, J.R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
6. Pinkel, D. *et al.* GH (CGH). US Patent 5,976,790 (1999).
7. Wang, N.J. *et al.* High-resolution molecular characterization of 15q11-q13 rearrangements by array CGH (array CGH) with detection of gene dosage. *Am. J. Hum. Genet.* **75**, 267–281 (2004).
8. Sahoo, T. *et al.* Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat. Genet.* **40**, 719–721 (2008).
9. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
10. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
11. Sharp, A.J. *et al.* Segmental duplications and copy number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
12. Zody, M.C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
13. Girirajan, S. & Eichler, E.E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum. Mol. Genet.* **19**, R176–R187 (2010).
14. Itsara, A. *et al.* Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *Am. J. Hum. Genet.* **90**, 599–613 (2012).
15. Highsmith, W.E., Meyer, K.J., Marley, V.M. & Jenkins, R.B. Conversion technology for the separation of maternal and paternal copies of any autosomal chromosome in somatic cell hybrids. *Curr. Prot. Hum. Genet.* **3**, 3.6 (2007).
16. van den Ijssel, P. & Ylstra, B. Oligonucleotide array CGH. in *Methods in Molecular Biology* vol. 396: Comparative Genomics, Volume 2 (ed. Bergman N.H.) 207-221 (Humana Press, 2007).
17. Horvath, J.E., Schwartz, S. & Eichler, E.E. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Res.* **10**, 839–852 (2000).
18. Saxena, R. *et al.* Four *DAZ* genes in two clusters found in the *AZFc* region of the human Y chromosome. *Genomics* **67**, 256–267 (2000).
19. Tilford, C.A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
20. Estivill, X. *et al.* Chromosomal regions containing high-density and ambiguously mapped putative single-nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* **11**, 1987–1995 (2002).
21. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet.* **36**, 861–866 (2004).
22. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
23. Nuttle, X. *et al.* Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat. Methods* **10**, 903–909 (2013).
24. Dennis, M.Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
25. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40**, 96–101 (2008).

26. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–414 (2013).
27. Carlson, C. *et al.* Molecular definition of 22q11 deletions in 151 velo-cardio-facial syndrome patients. *Am. J. Hum. Genet.* **61**, 620–629 (1997).
28. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
29. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
30. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
31. Turner, E.H. *et al.* Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).
32. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
33. Renshaw, M.A., Giresi, M. & Adams, J.O. Microsatellite fragment analysis using the ABI PRISM 377 DNA sequencer. in *Methods in Molecular Biology*, vol. 1006: Microsatellites (ed. Kantartzi, S.K.) 181–196 (Humana Press, 2013).
34. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
35. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
36. Tuzun, E., Bailey, J.A. & Eichler, E.E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14**, 493–506 (2004).
37. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).
38. Chin, C.S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).