bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Non-crossover gene conversions show strong GC bias and unexpected clustering in humans

Amy L. Williams[1,2,3,*,†], Giulio Genovese[3], Thomas Dyer[4], Katherine Truax[4], Goo Jun[5], Nick Patterson[3], Joanne E. Curran[4], Ravi Duggirala[4], John Blangero[4], David Reich[3,6,7], Molly Przeworski[1,2] for the T2D-GENES Consortium

[1] Biological Sciences Department, Columbia University, New York, NY 10027, USA
[2] Department of Systems Biology, Columbia University, New York, NY 10032, USA
[3] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[4] Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78227, USA
[5] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[6] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[7] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
[*] To whom correspondence should be addressed: alw289@cornell.edu
[†] Current affiliation: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

Although the past decade has seen tremendous progress in our understanding of fine-scale recombination, little is known about non-crossover (or "gene conversion") resolutions. We report the first genome-wide study of non-crossover gene conversion events in humans. Using SNP array data from 94 meioses, we identified 107 sites affected by non-crossover events, of which 51/53 were confirmed in sequence data. Our results suggest that a site is involved in a non-crossover event at a rate of 6.7×10^{-6}/bp/generation, consistent with results from sperm-typing studies. Observed non-crossover events show strong allelic bias, with 70% (61–79%) of events transmitting GC alleles ($P=7.9\times10^{-5}$), and have tracts lengths that vary over more than an order of magnitude. Strikingly, in 4 of 15 regions with available resequencing data, multiple (~2–4) distinct non-crossover events cluster within ~20–30 kb. This pattern has not been reported previously in mammals and is inconsistent with canonical models of double strand break repair.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Recombination is a process that deliberately inflicts double strand breaks on the genome during meiosis, leading to their repair as either crossover or non-crossover resolutions. These two outcomes of recombination are accompanied by a short gene conversion tract that fills in the double strand break in one homologous chromosome with the sequence from the other homolog. Whereas crossovers yield chromosomes with multi-megabase long segments from each homolog [1], non-crossover gene conversion tracts have been estimated to span ~50–1,000 bp [2].

Although short, these non-crossover gene conversion tracts affect sequence variation by breaking down linkage disequilibrium (LD) within a localized region, and, in addition to crossovers, are necessary to explain present-day haplotype diversity [3,4]. As an important aspect of recombination biology, characterizing non-crossovers also has potential implications for fertility [5]. While gene conversions also occur at crossover breakpoints, only non-crossover gene conversion events are detectable in pedigrees, and we therefore focus on these, using the shorthand "gene conversion" in what follows.

Despite the importance of gene conversion, much remains to be determined about its biological determinants and its effects. Notably, we know little about the overall frequency of gene conversion in mammals. Previous estimates of the frequency of gene conversion in humans range from ~1–15 times higher than crossover [2-4,6,7], with this value varying widely in both LD [4,6] and sperm-based [2,7] analyses. Likewise, while crossovers show differential frequencies and localization patterns in males and females [8], no such comparison exists for non-crossover gene conversion events.

Also unclear is the impact of gene conversion events on genome evolution. Cross-species analyses have shown that GC content in highly recombining regions increases over evolutionary time, with GC-biased gene conversion (gBGC) being the hypothesized means for this change [9]. Moreover, because gBGC acts analogously to positive selection, its effects on polymorphism and divergence can confound studies of human adaptation [10]. Although one recent sperm-typing study reported two recombination hotspots that exhibit GC-bias in non-crossover resolutions [7],

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

most of the evidence of gBGC in mammals has been based on cross-species divergence data, which cannot reliably estimate the strength of gBGC.

It is also of interest to characterize the localization of gene conversions with respect to crossover hotspots and to examine their locations relative to other recombination events in a single meiosis. While gene conversion events are assumed to occur at the same hotspots for double strand breaks as crossovers [1], this has only been demonstrated for a limited number of locations in sperm [11]. Among the hotspots examined, the ratio of non-crossover to crossover resolutions varies tremendously [2,7,11]. Furthermore, by considering events in a single meiosis, sperm-based analyses have identified *complex crossovers* in which gene conversions occur near but not contiguous with crossover breakpoints [12]. A genome-wide analysis of gene conversion has the potential to reveal further such features of recombination.

Motivated by these considerations, we carried out a study of meiotic gene conversion in pedigrees—to our knowledge, the first genome-wide assay of *de novo* gene conversion in mammals. We sought answers to the following questions: (1) Do gene conversions localize to the same hotspots as crossovers (as defined in [8])? (2) What is the rate at which a site is a part of a gene conversion tract? This is equivalent to the fraction of the genome affected by gene conversion in a given meiosis. (3) Are there differences in the gene conversion rate or localization patterns between males and females? (4) What is the strength of gBGC across the genome? (5) How long are gene conversion tracts, and how variable in length? (6) Are gene conversion tracts distributed independently of each other in a given meiosis or does more than one event sometimes co-occur in a short interval?

We utilized two different sources of data for our analysis. The primary analysis focused on SNP array data from 32 three-generation pedigrees. These SNP array data provide information from 94 meioses, 47 paternal and 47 maternal, and are informative at 12.0 million sites (markers where we can potentially detect a gene conversion in a parent-child transmission). We followed up with a secondary analysis of a subset of the identified gene conversion events using whole genome sequence data.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

We carried out a study of *de novo* meiotic gene conversion in humans by analyzing Illumina SNP array data at two SNP densities (660k and 1M SNP density arrays; see Methods) from 32 three-generation Mexican American pedigrees [13-15]. The goal was to identify *de novo* gene conversion events, manifested as 1 or more adjacent SNP sites that descend from the opposite haplotype relative to flanking markers (Figure 1a). Identifying these events requires phasing of genotypes in the pedigree in order to infer haplotypes and the locations of switches between parental homologs in transmitted haplotypes.

Two features make locating gene conversions challenging. The first is the density of informative sites. Gene conversions have an estimated mean tract length of 300 bp or less [2,7], but on a SNP array with ~1 million variants, genotyped sites occur on average every 3,000 bp. Thus SNP array data will identify only a small subset of gene conversion events. Moreover, to be informative about gene conversion (and recombination in general), a site must be heterozygous in the transmitting parent, so not all assayed positions are informative.

The second challenge arises from erroneous genotype calls. Errors in SNP array data can in principle confound an analysis of gene conversion because certain classes of errors can mimic gene conversion events (e.g., if a child is truly heterozygous but is called homozygous, or if a parent is homozygous but called heterozygous). Our study design minimizes false positive gene conversion calls by using three-generation pedigrees, as depicted in Figure 1b. The approach requires that a putative gene conversion identified in a child in the second generation also be transmitted to a grandchild (red arrows in Figure 1b). Additionally, the approach validates the genotype of the transmitting parent as heterozygous by requiring that the allele from the non-gene-converted haplotype in that parent be transmitted to at least one child (blue arrow in Figure 1b).These requirements guarantee that a false positive gene conversion will only be called if there are at least two genotyping errors at a site. Specifically, for a false positive to occur, either the recipient of the gene conversion and his or her child must be incorrectly typed, or the parent transmitting the putative gene conversion and the child/children receiving the alternate allele must be in error. This approach decreases the number of events that can be detected since not all

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

gene conversions will be transmitted to a grandchild, but it also greatly reduces the false positive rate. Further details on data quality control measures appear in Methods.

Our approach for identifying gene conversion events consisted of first phasing each three-generation pedigree using the program HAPI [16] (Methods). Next, we identified informative sites relative to each parent in the first generation. These are sites where the parent is heterozygous, the inferred phase is unambiguous, and where, if a gene conversion occurred, both alleles would be transmitted to the children (see Methods). We then examined all apparent double crossover events that occur within a span of 20 informative sites or less. That is, we identified haplotype transmissions that contain switches from one parental haplotype to the other and then switch back to the original haplotype. Most of these recombination intervals span 1 to 3 SNPs and are less than 5 kb, and these are putative gene conversion events. A few loci showed complex patterns with multiple, discontinuous recombination events across several SNPs, with tracts spanning 5 kb or more; these are not counted as gene conversions but are described below.

We ascertained the total number of informative sites in the same way as our gene conversion events. Thus, when calculating the per base pair (bp) rate of gene conversion, the numerator and denominator are identically ascertained (see below and Methods for details).

Identified gene conversions, validation, and localization

Within the 32 three-generation pedigrees, we considered transmissions from a total of 94 first generation meioses (47 paternal, 47 maternal). We identified a total of 107 sites putatively affected by autosomal gene conversion events: 102 with standard ascertainment, and an additional five that are detectable but do not meet all the criteria for inclusion in the rate calculation (Figure 1c; Table S1; Methods). We validated genotype calls for a subset of the putative gene conversions using whole genome sequence data generated by the T2D-GENES Consortium. These data contain genotype calls for 53 of these gene converted sites, of which 51 are concordant with the SNP array calls (Methods, Table S1). Of the two discordant sites, one shows evidence of being an artifact in the sequence data rather than the SNP array data, and for the other, the source of error is unclear (see Methods). Overall, the error rates in these data are low, and in what follows we assume that all 107 detected gene conversion events are real.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gene conversions are thought to localize to the same hotspots as crossovers [1], and studies at specific loci in sperm have supported this hypothesis [11]. To evaluate this question using genome-wide data, we utilized crossover rates that Kong *et al.* estimated based on events identified in an Icelandic pedigree dataset [8]. This genetic map omits telomeres, and thus these rates are only available for a subset of our identified gene conversions. The *de novo* gene conversions show strong enrichment in sites with crossover rate ≥10 cM/Mb (Figure 2a). Indeed, 20 of the 78 events that we can examine (26%) localize to such regions (using only one SNP per gene conversion event), while 4.2% of informative sites have this high of rate. This co-localization is unlikely to occur by chance (P=6.1×10^{-11}, one-sided binomial test), indicating that gene conversions are strongly enriched in crossover hotspots, and providing further validation that the detected gene conversion events are real.

Rate of gene conversion and male and female differences

With a total of 102 ascertained gene converted sites out of 12.0 million informative sites, we can estimate the per bp rate of gene conversion. Assuming the set of informative sites is unbiased with respect to recombination rate, an estimate is given by the number of gene converted sites divided by the number of informative sites. This represents the proportion of the genome affected by gene conversion, or equivalently the probability that a given site will be part of a gene conversion tract per meiosis.

As Figure 2b shows, however, our SNP array data are enriched in regions of high recombination relative to the genome-wide rate, and it is necessary to account for this bias. We therefore estimated the rate of gene conversion in each of five recombination rate intervals based on the HapMap2 recombination map (Figure 2b) by dividing the number of gene conversion sites by the number of informative sites observed in each bin. The overall rate is then the sum of these rates, each weighted by the proportion of the autosomes that occurs in the bin. This procedure yields a sex-averaged rate of R=6.7×10^{-6} per bp per meiosis (and a 95% confidence interval [CI] of 5.2×10^{-6} – 8.4×10^{-6}, calculated by 40,000 bootstrap samples with 10 Mb blocks).

Sperm-typing data have been used to examine the number and tract length of gene conversion events, notably in a study by Jeffreys and May that examined three hotspot loci in detail [2]. That

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

study estimated the number of gene conversion events to be 4–15 times the number of crossovers and the mean tract length to be 55–290 bp. The rate *R* can be calculated as the number of gene conversion tracts in a meiosis multiplied by the tract length, and divided by the genome length. Using the estimates from Jeffreys and May gives $R=2.6\times10^{-6}$ to 5.2×10^{-5}/bp/generation, a range that overlaps our estimates (for a genome-wide crossover rate of 1.2 cM/Mb). Our results are therefore consistent with those from sperm-based analyses, and they are also consistent with several LD-based studies of gene conversion [3,4,6].

Considering the parent of origin of each gene conversion event, we found that the two SNP arrays differ significantly in number of events detected per sex ($P=1.0\times10^{-3}$, χ^2 1 degree of freedom [df] test), with the lower density SNP dataset uncovering fewer male-specific events than expected. This bias may be caused by a lower coverage of the telomeres in the low density SNP array, and makes the analysis of potential differences in gene conversion rate between the sexes difficult. Nevertheless, considering the position of events captured by genotype arrays reveals broad-scale localization differences, with male events more prevalent in the telomeres and female events relatively dispersed throughout the genome (Figure 1c,d). These sex differences in localization are similar to those seen for crossover events [8], as expected from a shared mechanism for broad, megabase-scale control of both types of recombination.

GC-biased gene conversion

GC-biased gene conversion (gBGC) is an important force in the evolution of base composition [9] and has been highlighted as a confounder of the effects of natural selection [10]. To date, sperm-typing analyses have reported hotspots that exhibit allelic bias, but many of these biased transmissions arise from SNP polymorphisms that occur within motifs bound by PRDM9 [12]. Recombinations at these sites typically show under-transmission of the allele that better matches the PRDM9 motif, a phenomenon that can be thought of as a form of meiotic drive. A distinct form of biased gene conversion occurs when AT/GC heteroduplex DNA that arises during the repair of double strand breaks is preferentially repaired towards GC alleles [9]. A recent sperm-typing study reported on two loci that exhibit such biased gene conversion and only impact non-crossover gene conversion events [7]. This sperm-based study is, to our knowledge, the first to demonstrate direct evidence of gBGC in mammals.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Here, we considered the degree of GC-bias genome-wide. We saw no evidence for a difference in GC transmission rate between the two SNP density datasets (P= 0.12, χ^2 1-df test), or between males and females (P=0.69, χ^2 1-df test), and so considered the data jointly. For this calculation, we omitted gene converted sites that occur near crossovers and that are consequently ambiguous as to which strand converted (see below). Of the 100 unambiguous gene conversion sites (which all have an AT allele on one homolog and GC on the other), 70 transmit G or C alleles (70%, 95% CI 61–79%; P=7.9×10^{-5}, two-sided binomial test; Figure 2c). SNP variants at CpG dinucleotides account for 43 of these 100 sites, and these also show GC bias, with 28 CpG sites (65%) transmitting GC alleles, and no evidence of rate difference between transmissions at CpG and non-CpG sites (P=0.48, χ^2 1-df test). By comparison, the sperm-typing study noted above found that 2 of 6 assayed hotspots exhibited detectable levels of gBGC, and these two loci transmitted GC alleles in ~70% of meioses [7].

Gene conversion tract lengths

The data allow us to estimate gene conversion tract lengths, with upper bounds derived from informative SNPs that flank a gene conversion tract and lower bounds given by the distance spanned by SNPs involved in the same tract. Most gene conversion events involve only one SNP, but a total of eleven regions (nine with information from SNP array data only, and two including information from the sequence data) have tracts that include multiple SNPs (as plotted in Figure 3). From these data, we deduce that five of these events have a lower bound on tract length of at least 1 kb while the smallest is at least 94 bp. In turn, one tract is at most 124 bp—only slightly longer than the minimum tract involving more than one SNP (which has length ≥ 94 bp)—and four events have tracts shorter than 1,400 bp. These observations, coupled with the variable length in tracts that occur in the clustered gene conversion events described below (see Figure 4a), suggest that tract lengths are highly variable, and likely span at least an order of magnitude.

We note that, because gene conversions identified using SNP arrays are sparsely sampled, our data may be enriched for gene conversions with longer tracts, since these impact a larger number of sites. This effect would bias an estimate of the mean tract length using the data from this

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

study. It is also possible that some of the longer events result from clustered but separate tracts, as described below.

Clustered gene conversion tracts in sequence and SNP array data

We used Complete Genomics resequencing data for a subset of samples to more closely examine variants surrounding several of the identified gene conversion events. In order to confidently phase these regions, we required sequence data for both parents and three children (including the gene conversion event recipient); such data were available for two pedigrees. In these pedigrees, there are a total of 15 regions with evidence for a gene conversion event in the SNP array data. Two of these regions are not included in this analysis: for one, the sequence data do not contain genotype call for the putative gene conversion site, while in the other, genotype calls do not match the sequence data. Neither locus shows additional gene conversion sites.

Figure 4a shows the phase for the 13 regions included. In four cases (haplotypes 10–13), multiple discontinuous gene conversion tracts occur within a short interval of less than 30 kb, with discontinuities evident from informative sites located between the gene conversion tracts. The four cases occurred in a single pedigree, three in the mother, and one in the father (haplotype 11). The LD-based genetic map length of the 100 kb around these four regions ranges from 0.034 cM to 0.28 cM. Using these genetic lengths to estimate the probability of gene conversion initiation (Methods), we found that this clustering is highly unexpected, with a probability of observing two independent tracts within the four 100 kb regions ranging from $P=3.7\times10^{-6}$ to 2.4×10^{-4} (considering each region independently).

To check for possible artifacts, we performed Sanger sequencing of the three-generation pedigrees for six regions in three of these four haplotypes, indicated by boxes in Figure 4a. The Sanger sequence data from these regions are concordant with the genotypes from the whole genome sequence data at every site and in all individuals. Moreover, we checked for overlap between these regions and the following resources: (a) recent segmental duplications that have divergence between them of <2% [17]; (b) the 35.4 Mb “decoy sequences” released by the 1000 Genomes Project [18] which contain regions of the genome that are paralogous to sequence from Genbank [19] and the HuRef alternate genome assembly [20]; and (c) regions of the genome

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with excess read mapping in the 1000 Genomes Project [21]. Our quality control procedure already removed individual SNPs that overlap several of these resources (Methods), and this analysis showed no overlap within the regions containing these clustered sites.

The close clustering of gene conversion events occurs in 4 of 15 (27%) cases that we were able to examine, so may be common. As in the case of long tracts, however, our sparse, SNP array-based sampling may be more likely to detect clustered gene conversions (since multiple events may affect a larger proportion of sites), and therefore the rate of clustering may be somewhat lower. Nonetheless, these events are unlikely to be rare.

Indeed, later examination of our array-based data revealed three other clustered gene conversion events as well as six gene conversion events near but disconnected from crossover resolutions (Figure 4b). All events other than two were transmitted in different pedigrees, and those two haplotypes (numbers 18 and 19) are the same events that show clustered gene conversion in sequence data (Figure 4a, haplotypes 11 and 13). These additional observations buttress the evidence for clustered gene conversion and shed light on the distances over which complex crossover may occur. The complex crossover events previously described in humans were seen in assays of relatively short intervals around crossover breakpoints, and suggested that they occurred at a frequency of 0.17% [12]. The results from the current study indicate that additional events may occur farther from the crossover breakpoint, so complex crossover may be more common. Whether the observations at short and longer distances result from the same phenomenon remains to be elucidated.

To our knowledge, this is the first observation of clustered but discontinuous gene conversion tracts in mammalian meiosis, although patterns that resemble those shown in Figure 4a have been reported in meiosis [22,23] and mitosis [24,25] in *S. cerevisiae*. This phenomenon and the distant forms of complex crossover both point to a property of mammalian recombination that is not understood and that is not predicted by canonical models of double strand break repair [1].

Contiguous and clustered recombination events spanning larger distances

In addition to the gene conversion events with tracts that span no more than 5 kb, we identified four longer-range recombination events: two contiguous tracts, and two that showed a clustering

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

pattern (see Figure 5). Each event occurred in a different pedigree, and the contiguous tract that spans ~79 kb was transmitted by a male, while the three others occurred in females. The long contiguous tracts could reflect crossovers in extremely close proximity, as might arise from a crossover-interference independent pathway [26], but the clustered events cannot be explained in this way. For two events, sequence data are available and validate the genotype calls, indicating that the case that spans at least 9 kb in the genotype data is in fact at least 18 kb long, and confirming the case in which clustered events span ~203 kb.

Haplotypes 23 and 26 reside on the p arm of chromosome 8 where a long inversion polymorphism occurs [27]. Single crossovers within inversion heterozygotes can be misinterpreted as double crossover events [28], yet these two recombination events are > 1.7 Mb outside the inversion breakpoints, so should not be affected. One possibility is that the large inversion polymorphism leads to aberrant synapsis between chromosomes during meiosis, leading to complex repair of double strand breaks. In that regard, we note the transmitter of haplotype 23 is heterozygous for tag SNPs for the 8p23 inversion polymorphism [27], and that a sibling inherited a haplotype from the same parent with a crossover at the same position as the end of the tract for haplotype 23. This co-localization may be due effects of the inversion on synapsis; alternatively, this could indicate that the sites are incorrectly positioned, resulting in inaccurate inference of breakpoint locations [28]. The pattern is haplotype 26 is even more complex and difficult to explain by any standard model of recombination.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Non-crossover gene conversion reshuffles haplotypes and shapes LD patterns, at a rate that we estimate to be 6.7×10^{-6}/bp/generation. The heritable and evolutionary effects of gene conversion events occur only at heterozygous sites, so this rate can be meaningfully scaled by human heterozygosity levels. Assuming that $\pi = 10^{-3}$ [29], roughly 19 (95% CI 15–24) variable sites are expected to experience gene conversion in each meiosis (for a euchromatic genome length of 2.9×10^{9} bp). This estimate is on the same order as the number of sites affected by *de novo* mutation in each generation.

In regions that experience gene conversion, our results indicate that there is frequent over-transmission of G or C alleles. Indeed, we observed GC transmission in 70% of events (95% CI 61–79%). More generally, our results provide a direct confirmation of the presence of gBGC, and lend strong support to the hypothesis that it could play a major role in shaping base composition over evolutionary timescales [9].

Considering the distribution of SNPs in gene conversion tracts, we found lengths that vary over more than an order of magnitude, from hundreds to thousands of base pairs. Intriguingly, we also identified several examples of loci where multiple gene conversion tracts cluster within 20–30 kb intervals, as well as instances of complex crossover over extended intervals. As current models do not predict these phenomena, understanding their source will be important for studies of mammalian recombination and may lead to improved population genetic models of haplotypes and LD. A separate study examining *de novo* mutations reported observing regions with gene converted sites across intervals spanning between 2–11 kb [30]. These events may either be long gene conversion tracts or clustered but discontinuous gene conversion events in the same meiosis.

Thus, the results presented here point to a basic feature of human recombination biology that remains to be explained. Going forward, whole genome sequencing of human pedigrees will enable unbiased analyses of *de novo* gene conversion at relatively high resolution. Of particular interest will be systematic examination of tract length distribution and the patterns of clustered gene conversion events revealed by this study.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Samples and sample selection

This study analyzed Mexican American samples from the San Antonio Family Studies (SAFS) pedigrees. SNP array data were generated for these individuals as previously described [13-15]. Our study design required the use of three-generation pedigrees with SNP array data for both parents in the first generation, three or more children in the second generation, one or more grandchildren, and data for both parents for any included grandchildren. Within the entire SAFS dataset of 2,490 individuals, there are 35 three-generation pedigrees consisting of 496 individuals that fit the requirements of this design. As noted below, three of these pedigrees were not included in the analysis, so the overall sample consists of 32 pedigrees and 458 individuals.

Each sample was genotyped using one of the following Illumina arrays: the Human660W, Human1M, Human1M-Duo, or both the HumanHap500 and the HumanExon510S (these latter two arrays together give roughly the same content as the Human1M and Human1M-Duo).

Most of the samples—19 out of the 32 analyzed pedigrees containing 269 individuals—have SNP data derived from arrays with roughly equivalent content and ~1 million genotyped sites. We analyzed all these samples across the SNPs shared among these arrays, with data quality control applied collectively to all samples and sites (see below). After quality control filtering, 896,375 autosomal SNPs remained for the analysis of gene conversion.

Data for the other 13 out of 32 analyzed pedigrees comprise 189 individuals and were analyzed on a lower density SNP arrays. The majority of the samples in these pedigrees (105 individuals) have SNP array data from ~660,000 genotyped sites. The other samples (84 individuals) have higher density genotype data available, but because other pedigree members have only lower density data, we omit these additional sites from analysis. After quality filtering, this lower SNP density dataset contained 513,283 autosomal sites.

Quality control procedures applied to full dataset

Initially, sites with non-Mendelian errors, as detected within the entire SAFS pedigree, were set to missing. We next ensured that the locations of the SNPs were correct by aligning SNP probe

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequences to the human genome reference (GRCh37) using BWA v0.7.5a-r405 [31]. Manifest files for each SNP array list the probe sequences contained on the array and we confirmed that these probe sequences are identical across all arrays for the SNPs shared in common among them. We retained only sites that (a) align to the reference genome with no mismatches at exactly one genomic position and that (b) do not align to any other location with either zero or one mismatches.

We updated the physical positions of the SNPs in accordance with the locations reported by our alignment procedure and utilized SNP rs ids contained in dbSNP at those locations. We omitted sites for which multiple probes aligned to the same location. Some sites had either more than two variants or had non-simple alleles (i.e., not A/C/G/T) reported by dbSNP, and we removed these sites. We also filtered three sites that had differing alleles reported in the raw genotype data as compared to those reported for the corresponding sites in the manifest files. We filtered a small number of sites for which the manifest file listed SNP alleles that differed from those in dbSNP at the aligned location.

Some SNPs are listed in dbSNP as having multiple locations or as "suspected," and we removed these sites from our dataset. We also removed sites that occur outside the "accessible genome" as reported by the 1,000 Genomes Project [29] (roughly 6% of the genome is outside this), and sites that occur in regions that are segmentally duplicated with a Jukes-Cantor K-value of <2% (this value closely approximates divergence between the paralogs) [17]. Finally, we removed sites that occur within a total of 17 Mb of the genome that receive excess read alignment in 1,000 Genome Project data [21].

We next conducted more standard quality control measures by performing analyses on two distinct datasets: (1) including all individuals that were genotyped at ~1 million SNPs (1,932 samples) and (2) including all 2,490 samples. On the densely typed dataset, we first removed any site with ≥1% missing data and those for which a test for differences between male and female allele frequencies showed $|Z|\geq 3$. We then removed 29 samples with ≥2% missing data. Next we examined the principal components analysis (PCA) plots [32] generated using (a) the genotype data and (b) indicators of missing data at a site. These plots generally show an absence of outlier

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

samples, and the genotype-based PCA plot appears consistent with the admixed history of the Mexican Americans (results not shown).

For the datasets that include samples typed at lower density, we first removed sites with ≥1% missing data and sites with male-female allele frequency differences with $|Z|\geq 3$. This filtering step yields SNPs of high quality that are shared across all SNP arrays, including the lower density Human660W array. Next we removed 30 samples with ≥2% missing data. Lastly, we examined PCA plots generated using (a) genotype and (b) missing data at each site, and these plots are again generally as expected with an absence of outlier samples (results not shown).

Phasing and identifying relevant recombination events in three-generation pedigrees

We performed minimum-recombinant phasing on the three-generation pedigrees using the software HAPI [16], but with minor modifications because this program phases nuclear families independently. Specifically, our approach phased nuclear families starting at the first generation family. After this completed, we phased the families from later generations while utilizing the haplotype assignments from the first generation. Our approach assigned the phase at the first heterozygous marker to be consistent across generations in the individuals shared between the two nuclear families. (Shared individuals are members of the second generation who are a child in one family and a parent in another.) This approach helps produce consistent phasing across generations and does not introduce extra recombinations since the phase assignment at the first marker on a chromosome is arbitrary.

After phasing, our method for detecting gene conversions also handled sites with inconsistent phase between the families (though in practice nearly all sites have consistent phase assignments between families). This method excluded sites that have inconsistent phase and that occur within a background of flanking markers with consistent phase; we examined these sites individually and confirmed that they do not represent gene conversion events, but are likely driven by genotyping errors. When 10 or more informative SNPs in succession are inconsistent across families, we assumed that a crossover event went undetected in one of the generations, and inverted the phase for the relevant individuals in order to identify putative gene conversion events.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We analyzed the inferred haplotype transmissions to identify sites that exhibit recombination from one haplotype to the other and then back again. The detection approach identified any recombination events that switch and revert back to the original haplotype within ≤ 20 informative SNPs.

Pedigree-specific quality control and determination of informative sites

Genotypes are only informative for which haplotype a parent transmits—and therefore recombination—at sites where the parent is heterozygous. We employed a pedigree-specific quality control measure by only considering sites in which all individuals in the full three-generation pedigree have genotype calls and no missing data; other sites are omitted. This requirement helps address possible structural or other complex variants that are specific to a particular pedigree and that may adversely affect genotype calling (as evidenced by a lack of a genotype call for some individual in that pedigree at the given site).

Because gene conversions occur relatively infrequently, it is unlikely that the same position will experience gene conversion in multiple generations. We therefore excluded sites that exhibit gene conversion in any grandchild (i.e., locations with potential gene conversion events transmitted from the second generation). We applied this filter regardless of the gene conversion status in earlier generations in order to obtain unbiased ascertainment of events and informative sites. We also excluded sites that exhibit potential gene conversion events from a given parent and where that parent only transmits one haplotype. In this case, the genotype from the transmitting parent is likely to be in error and to be homozygous; given this consideration, we considered the site as invalid for both parents.

In principle, all children in the second generation are useful for studying meiosis in their parents, but to reduce false positives, we only analyzed a subset of the these children. Specifically, we only analyzed a child if data for his/her spouse and one or more of their children (grandchildren in the larger pedigree) were available.

We counted a site as informative (or not) relative to a given parent and a given child if sufficient data for relatives were available and if it satisfied five requirements. First, we required the parent to be heterozygous at the site. Second, as shown in Figure 1b, we required the allele that the

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

given parent transmitted to the child also be transmitted to at least one grandchild. Third, in any series of otherwise informative sites, we counted all but the first and last sites as informative since we detect gene conversion events as haplotype switches relative to some previous informative site. Fourth, except at sites that are putatively gene converted, we required that a second child to have received the same haplotype as the child that is potentially informative. This requirement helps to ensure the validity of the heterozygous genotype call of the parent. As an example, consider a pedigree with four children, three of whom received a haplotype 'A' at some site and the fourth of whom received haplotype 'B'. If the fourth child were to receive a gene conversion at some subsequent position, it would receive haplotype 'A', and thus all four children would receive the same haplotype. This scenario violates the requirement that the non-gene converted allele be transmitted to at least one second-generation child. Thus, in this example, the fourth child is not informative at this example site (where it is the sole recipient of haplotype 'B'). Note however that this site could be informative in the other children if they meet the other requirements listed here.

Finally, we required that the site be phased unambiguously across two generations, and that if a gene conversion had occurred, the phase at the site would remain unambiguous in the first generation. Sites in which all individuals in a nuclear family are heterozygous have ambiguous phase. Thus, if a given child is homozygous at a marker but all other individuals in the family are heterozygous, the child is not informative at that site since a gene conversion event would lead the child to be heterozygous. We note that it is possible to identify putative gene conversions when a child receives a haplotype that has recombined from otherwise ambiguous phase to be homozygous at this type of marker. Indeed, we identified five such putatively gene converted sites, but did not include them when calculating the rate of gene conversion since the denominator does not include ambiguously phased sites and is therefore ascertained differently.

Pedigrees included in the analysis

Three out of the 35 available three-generation pedigrees were excluded from our analysis. One pedigree is an outlier for gene conversion rate: in it, we detected nine putative gene conversions out of ~208,000 informative sites—suggesting a rate roughly an order of magnitude higher than suggested by other pedigrees. All nine of these gene conversion events are homozygous in the

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

recipient, and that recipient has a missing data rate that is more than double any other gene conversion recipient. The other two excluded pedigrees failed phasing because of a bug in the software and were therefore excluded.

Quality filtering of double recombination events in close proximity

Our method identified all double recombination events (defined as switches from one haplotype to the other and then back again) that span 20 informative sites or fewer. We examined the haplotype transmissions at each such reported event by hand to ensure that segregation to all children matches expectations. A few sites exhibited gene conversion events in the same interval in two or more children. Because gene conversion is relatively rare, it is unlikely that these are true gene conversion events. Additionally, some sites were consistent with gene conversion events transmitted to the same child from both parents; these are again unlikely to be real and are more likely caused when a child is homozygous for one allele but called homozygous for the opposite allele. We therefore considered these cases false positives.

Although we omitted sites in which grandchildren exhibit putative gene conversion events that occur at a single site, the software did not filter putative gene conversions that span multiple sites. We examined all events by hand, and excluded three reported gene conversion events in which the grandchildren either exhibit putative gene conversions longer than one SNP (therefore undetected) or show aberrant genotype calls.

The main text describes four long-range recombination events. For all these events, the recombined alleles at every site were transmitted to the third generation with no apparent recombinations or gene conversion events in the third generation. We excluded two other events with unexpected transmissions to the grandchildren. Specifically, one 4-SNP contiguous tract shows transmission to the third generation for three of the four recombined SNPs, but one SNP in middle of the tract was not transmitted and shows an apparent gene conversion in the third generation. The other 18-SNP long contiguous tract shows a putative gene conversion transmitted from the opposite parent across this same interval.

Validating gene conversion events

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We tested for overrepresentation of either heterozygous or homozygous genotype calls in the recipient of the putative gene conversions. Overrepresentation would suggest bias and possibly artifactual detection of gene conversions, but we saw no evidence of bias (P=0.92, two-sided binomial test). This analysis excludes the five sites identified using non-standard ascertainment and which are homozygous by detection.

Of the 458 individuals that we analyzed using SNP array data, 98 were whole genome sequenced by the T2D-GENES Consortium and we were therefore able to check concordance of genotype calls. We attempted validation on all sites for which data were available for the transmitting parent or a recipient (either the child or a grandchild) of the putative gene conversion (Table S1). Within these 98 samples, genotype calls were available for 53 of the putative gene converted sites (of the 107 total); 42 of these sites include data for both the transmitting parent and a gene conversion recipient. One additional site had data available for relevant samples, but the sequence data do not contain calls for that position. We compared genotypes for every available parent, child, partner of the gene conversion recipient, and children of the recipient (grandchildren in the larger pedigree). The genotype calls for all inspected individuals are concordant between the two sources of data for 51 of the 53 sites. One of the inconsistent sites shows a discordant genotype call between the datasets for the recipient of the gene conversion, but a concordant call for his child (the grandchild in the pedigree). This inconsistency suggests that the genotype data may in fact be correct. The other discrepancy occurs at a site where sequence data were unavailable for the recipient of the gene conversion. Here, the genotype call for the transmitting parent is discordant between the two sources of data, and the error source is ambiguous; we retained this site in the analyses.

Crossover and recombination rates

Crossover rates are those reported by deCODE [8] based on crossovers detected in large Icelandic pedigrees. The original map is reported for human genome build 36 and was lifted over to build 37 coordinates. This map is estimated to have resolution to roughly 10 kb, and we therefore computed recombination rates in cM/Mb using the genetic distances from the map across 10 kb windows and divided by this (10 kb) window size. Because this map omits relatively large telomeric segments, we did not have rates for many sites from the SNP arrays

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and from the identified gene conversion events. We used linear interpolation to obtain rates at sites within the range of the map but not directly reported. The proportion of sites in the "autosomal genome" in Figure 2a derives from all sites within the reported positions in the autosomal genetic map.

The HapMap2 LD-based recombination rates are from the genetic map generated by the HapMap Consortium [33] using LDhat [34] that was subsequently lifted over to human genome reference GRCh37. We used analogous methods for calculating recombination rates from this map as for the crossover map mentioned above, including a window size of 10 kb and linear interpolation. A few sites on the higher density SNP data (12 of 896,387) fall outside the interval of positions reported in the map.

Inclusion criteria for gene conversion and GC-bias rate calculations, crossover hotspots, and tract lengths

Five gene conversion events were identified with non-standard ascertainment and are inappropriate for inclusion in estimating the rate of gene conversion. However, these sites are not expected to show bias with respect to allelic composition and we therefore included them when calculating the strength of GC-bias.

Somewhat more complex cases are gene conversion sites that occur near crossover events (Figure 4b, haplotypes 17–22). In most, a single site appears to have been involved in the gene conversion event, and is followed by a single site that reverts to the first haplotype, and then followed by a crossover. Depending on whether one considers the "background haplotype" to be the one upstream of the gene conversion and crossover, or downstream, the site that was in the gene conversion tract differs. Thus which site was gene-converted is ambiguous. To simplify the examination of GC-bias, we excluded these sites from consideration. However, to estimate the rate of gene conversion genome-wide, rather than exclude these sites—which would bias our rate calculation downwards—we instead included both possibilities in the rate calculation, and gave each of them a weight of 0.5, while other sites have a weight of 1. There are two effects of this weighting. First, if the recombination rate bin differs across these sites, they each contribute the weight of half a site to the rate calculation for those bins. Most sites fall into the same rate bin

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and therefore have the same effect as counting a single site. The second effect of weighting these sites is that, in one case, we cannot tell whether 2 SNPs were gene-converted or only 1 SNP was. In this case, we counted the event as 1.5 gene-converted sites. Finally, we observed one instance of two putatively gene converted sites separated from a crossover by three informative sites. The three informative sites span 19.6 kb—longer than our threshold for gene conversion events. In this case, we considered the two sites (which form a tract of length at least 264 bp) as definitive gene conversions with weight 1.

For estimating the number of sites with crossover rate ≥10 cM/Mb, we included only 1 SNP per tract and weighted ambiguous cases by 0.5 as above. Additionally, two ambiguous sites have crossover rates that straddle this threshold, with one site slightly less, the other slightly more. To be conservative in estimating a P-value, we considered these sites as falling below the threshold.

To examine tract lengths, we omitted all but one ambiguous event. For the one included ambiguous event, the two possibilities have tract lengths ≥1,615 bp and ≥365 bp (upper bounds are more than 25 kb for both). We included the shorter of these lengths (365 bp) since this lower bound holds for both possibilities.

Examination of regions containing clustered gene conversions

We calculated the probability of two gene conversion events occurring within the four intervals in which we observed clustered gene conversion by rescaling the genetic distances of these regions as reported in the LD-based map. (Note that this map includes some of the historical effects of gene conversion [35].) We earlier estimated the per bp rate of gene conversion R, and $R=N\times l/G$ where N is the number of gene conversion events that occur in a meiosis, l is the average tract length of these events, and G is the total genome length. The genome-wide average rate of *initiation* of gene conversion at a bp is simply $N/G = R/l$. For an interval with genetic map length d cM, we estimated the rate of initiating a gene conversion as $r=d/c\times R/l$, where c=1.2 cM/Mb is the average genome-wide rate of crossover. The probability of two independent gene conversion tracts (conservatively assuming lack of interference among events) is then $P=r^2$. This calculation assumes the HapMap2 map accurately represents the relative rate of both crossover and gene conversion events in an interval; a test for difference between the observed locations of

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

gene conversion sites and expected locations based on this map are generally consistent with this assumption (P=0.15, χ^2 4-df test).

We performed Sanger sequencing on individuals from the three-generation pedigrees in which clustered gene conversions occurred. Assayed samples included both parents, all children (including the gene conversion recipient), the partner of the gene conversion recipient, and all grandchildren of that couple. Overall, sequencing included 11 or 12 samples for each of the three regions examined. We manually examined chromatograms to determine genotype calls. For most variant positions, the sequence quality was sufficient to easily call genotypes, though for a minority of sites, we did not call all samples. Still, sufficient data were available at sites intended for validation to verify either the gene conversion recipient or his/her grandchild and thereby confirm the status of the gene-converted allele. The available Sanger-based calls were concordant with the re-sequencing data for all sites and samples.

The main text describes an additional analysis that checked the regions for potential mismapping from paralogous sequences elsewhere in the genome.

Sanger Sequencing

We ran Primer3 (http://bioinfo.ut.ee/primer3/) using the initial presets on the human reference sequence from targeted regions to obtain primer sequences. For the suggested primer designs, we performed a BLAST against the human reference to ensure that each primer is unique, and ordered primers from Eurofins Operon. We tested each primer using the temperature suggested during primer design on DNA at a concentration of 10ng/uL and checked on a 2% agarose gel. For any primer with poor performance, we conducted a temperature gradient, and, if needed, a salt gradient until we found a PCR mix that performed well. Next we performed PCR on the samples of interest, running a small quantity on a 2% agarose gel. We then cleaned the PCR sample using Affymetrix ExoSAP-IT and ran sequencing reactions twice for each sample using Life Technologies BigDye Terminator v3.1 Cycle Sequencing Kit. Finally, we purified each sample using Life Technologies BigDye XTerminato Purification Kit and placed these onto the 3730xl DNA Analyzer for sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank Scott Keeney and Maria Jasin for helpful discussions and Melanie Carless for bioinformatics support. We thank Swapan Mallick for sharing a version of the deCODE crossover map in GRCh37 coordinates. A.L.W. was supported by NIH Ruth L. Kirschstein National Research Service Award number F32 HG005944. This work was supported by NIH GM83098 to M.P. and was done while M.P. was a Howard Hughes Medical Institute Early Career Scientist. D.R. is a Howard Hughes Medical Institute Investigator. T2D-GENES project data generation was supported by NIH grants U01 DK085501, U01 DK085524, U01 DK085526, U01 DK085545, and U01 DK085584.

Competing Interests

The authors declare that no competing interests exist.

References

1. Baudat F, Imai Y, de Massy B (2013) Meiotic recombination in mammals: localization and regulation. Nat Rev Genet 14: 794-806.
2. Jeffreys AJ, May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet 36: 151-156.
3. Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, et al. (2001) Lower-Than-Expected Linkage Disequilibrium between Tightly Linked Markers in Humans Suggests a Role for Gene Conversion. Am J Hum Genet 69: 582-589.
4. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, et al. (2001) Gene Conversion and Different Population Histories May Explain the Contrast between Polymorphism and Linkage Disequilibrium Levels. Am J Hum Genet 69: 831-843.
5. Hassold T, Hall H, Hunt P (2007) The origin of human aneuploidy: where we have been, where we are going. Human Molecular Genetics 16: R203-R208.
6. Gay J, Myers S, McVean G (2007) Estimating Meiotic Gene Conversion Rates From Population Genetic Data. Genetics 177: 881-894.
7. Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, May CA (2014) Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. PLoS Genet 10: e1004106.
8. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467: 1099-1103.
9. Duret L, Galtier N (2009) Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics 10: 285-311.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

10. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23: 273-277.
11. Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, et al. (2011) Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proceedings of the National Academy of Sciences 108: 12378-12383.
12. Webb AJ, Berg IL, Jeffreys A (2008) Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proceedings of the National Academy of Sciences 105: 10471-10476.
13. Mitchell BD, Kammerer CM, Blangero J, Mahaney MC, Rainwater DL, et al. (1996) Genetic and Environmental Contributions to Cardiovascular Risk Factors in Mexican Americans: The San Antonio Family Heart Study. Circulation 94: 2159-2170.
14. Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, et al. (1999) Linkage of Type 2 Diabetes Mellitus and of Age at Onset to a Genetic Location on Chromosome 10q in Mexican Americans. The American Journal of Human Genetics 64: 1127-1140.
15. Hunt KJ, Lehman DM, Arya R, Fowler S, Leach RJ, et al. (2005) Genome-Wide Linkage Analyses of Type 2 Diabetes in Mexican Americans: The San Antonio Family Diabetes/Gallbladder Study. Diabetes 54: 2655-2662.
16. Williams A, Housman D, Rinard M, Gifford D (2010) Rapid haplotype inference for nuclear families. Genome Biology 11: R108.
17. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent Segmental Duplications in the Human Genome. Science 297: 1003-1007.
18. 1000 Genomes Project Human Decoy Sequences (37d5). ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/.
19. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. (2014) GenBank. Nucleic Acids Res 42: D32-D37.
20. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The Diploid Genome Sequence of an Individual Human. PLoS Biol 5: e254.
21. Genovese G, Handsaker Robert E, Li H, Kenny Eimear E, McCarroll Steven A (2013) Mapping the Human Reference Genome s Missing Sequence by Three-Way Admixture in Latino Genomes. Am J Hum Genet 93: 411-421.
22. Globus ST (2013) From Start to Finish: Fine Scale Mapping of Meiotic Double Strand Breaks and Gene Conversion Tracts Reveals New Insights Into Homologous Recombination: Cornell University.
23. Martini E, Borde V, Legendre M, Audic S, Regnault B, et al. (2011) Genome-Wide Analysis of Heteroduplex DNA in Mismatch Repair–Deficient Yeast Cells Reveals Novel Properties of Meiotic Recombination Pathways. PLoS Genet 7: e1002305.
24. St. Charles J, Petes TD (2013) High-Resolution Mapping of Spontaneous Mitotic Recombination Hotspots on the 1.1 Mb Arm of Yeast Chromosome IV. PLoS Genet 9: e1003434.
25. Yin Y, Petes TD (2013) Genome-Wide High-Resolution Mapping of UV-Induced Mitotic Recombination Events in Saccharomyces cerevisiae. PLoS Genet 9: e1003894.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, et al. (2009) Broad-Scale Recombination Patterns Underlying Proper Disjunction in Humans. PLoS Genet 5: e1000658.
27. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, et al. (2009) Characterization of six human disease-associated inversion polymorphisms. Human Molecular Genetics 18: 2555-2566.
28. Broman KW, Matsumoto N, Giglio S, Martin CL, Roseberry JA, et al. (2003) Common long human inversion polymorphism on chromosome 8p. In: Goldstein DR, editor. Science and Statistics: A Festschrift for Terry Speed. IMS Lecture Notes-Monograph Series. pp. 237-245.
29. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
30. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44: 1277-1281.
31. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
32. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2: e190.
33. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
34. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science 304: 581-584.
35. Hellenthal G, Stephens M (2006) Insights into recombination from population genetic variation. Current Opinion in Genetics & Development 16: 565-572.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Titles and Legends

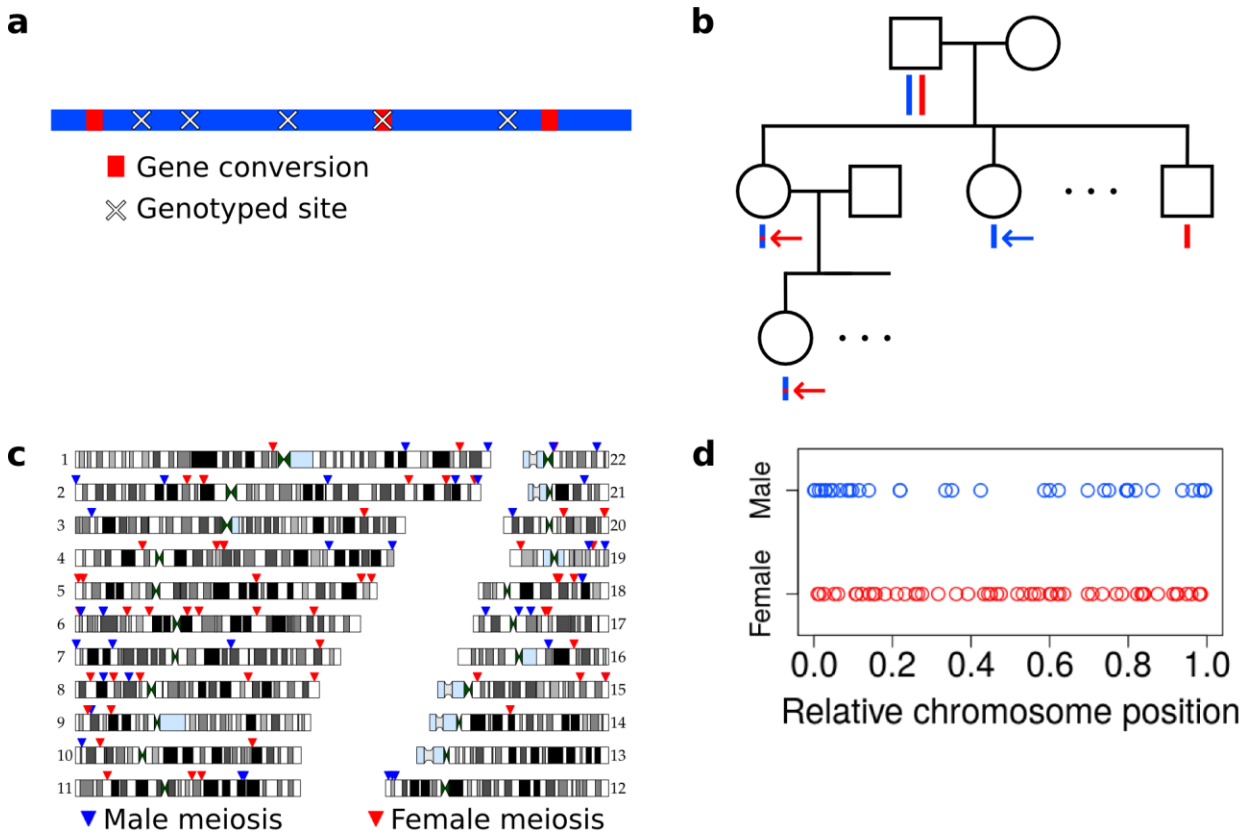

Figure 1. Gene conversion detection. a, Pictorial representation of a haplotype transmission including gene conversion events. A parent has two copies of each chromosome but transmits only one copy to his or her children. That copy is composed of DNA segments from the parent's two homologs; i.e., it is formed by recombination between these two haplotypes. Here, the two haplotypes in the parent are colored in blue and red, and switches in color represent sites of recombination. The figure only depicts short gene conversion events and no crossovers. Overlaid on this haplotype are × symbols representing sites assayed by the SNP array. In this example, only one gene conversion has a SNP array site within it and only that gene conversion can be identified. **b**, To avoid calling false positive gene conversion events driven by genotyping error, we required putative gene conversion events first to be detected in a second generation child (top red arrow) and also transmitted to a third generation grandchild (bottom red arrow). We also required that the allele from the non-gene converted haplotype in the parent (first generation) be transmitted to at least one child in the second generation (blue arrow). This study design ensures

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

that false positive gene conversions will only occur if there are two or more genotyping errors at a site. All 32 pedigrees included in this study have genotype data for both parents, at least three children, one or more grandchild, and both parents of included grandchildren. **c**, Genomic locations of the gene conversion sites that we detected are indicated by arrowheads, with red arrowheads representing gene conversion events from female meioses, and blue from male meioses. Many of the male gene conversion events localize to the telomeres. **d**, Relative chromosomal positions of events, stratified by the sex of the transmitting parent.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

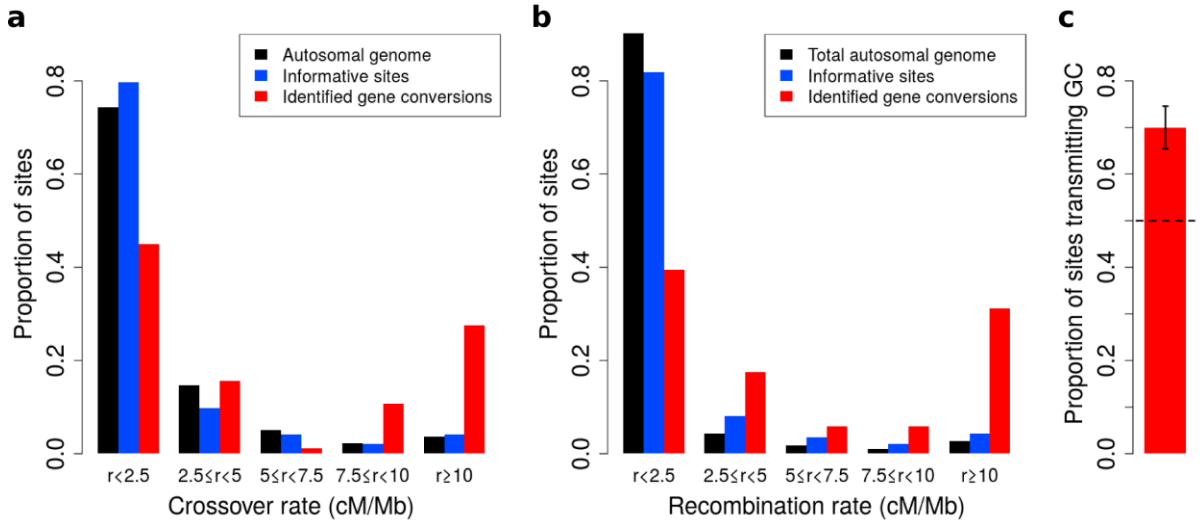

Figure 2. Localization of gene conversions in hotspots and rate of GC vs. AT allele transmissions. a, Histogram of proportions of sites that fall into five ranges of crossover rates [8] in the autosomal genome, all informative sites, and the identified gene conversion events (see Methods). Because this map excludes telomeric regions, some sites are excluded. **b**, Same as in **a**, but rates are from the HapMap2 LD-based recombination map [33]. This map does not exclude the telomeres and provides rate information for all gene conversion sites and nearly all sites from the SNP arrays (see Methods). **c**, Rate of GC allele transmissions: 70 out of 100 gene conversions transmit GC alleles. Thus, GC alleles are transmitted in 70% of gene conversion events (95% CI 61–79%; $P=7.9\times10^{-5}$, two-sided binomial test). Plot shows standard error bars.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

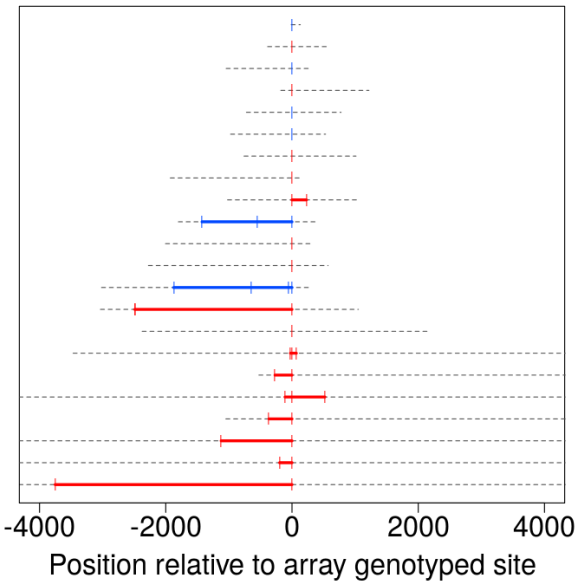

Figure 3. Tract lengths for identified gene conversions. Tract lengths derived from a total of 22 gene conversions that either have 2 or more SNPs in a tract or have maximum length of ≤5 kb. Each line corresponds to a gene conversion tract; lower bounds on length appear in color, with red corresponding to tract lengths informed by SNP array data and blue corresponding to tract lengths from sequence data. Gray dashed lines represent the region of uncertainty surrounding the tract length, with end points being the upper bound on tract length. Tracts are sorted by upper bound on tract length.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

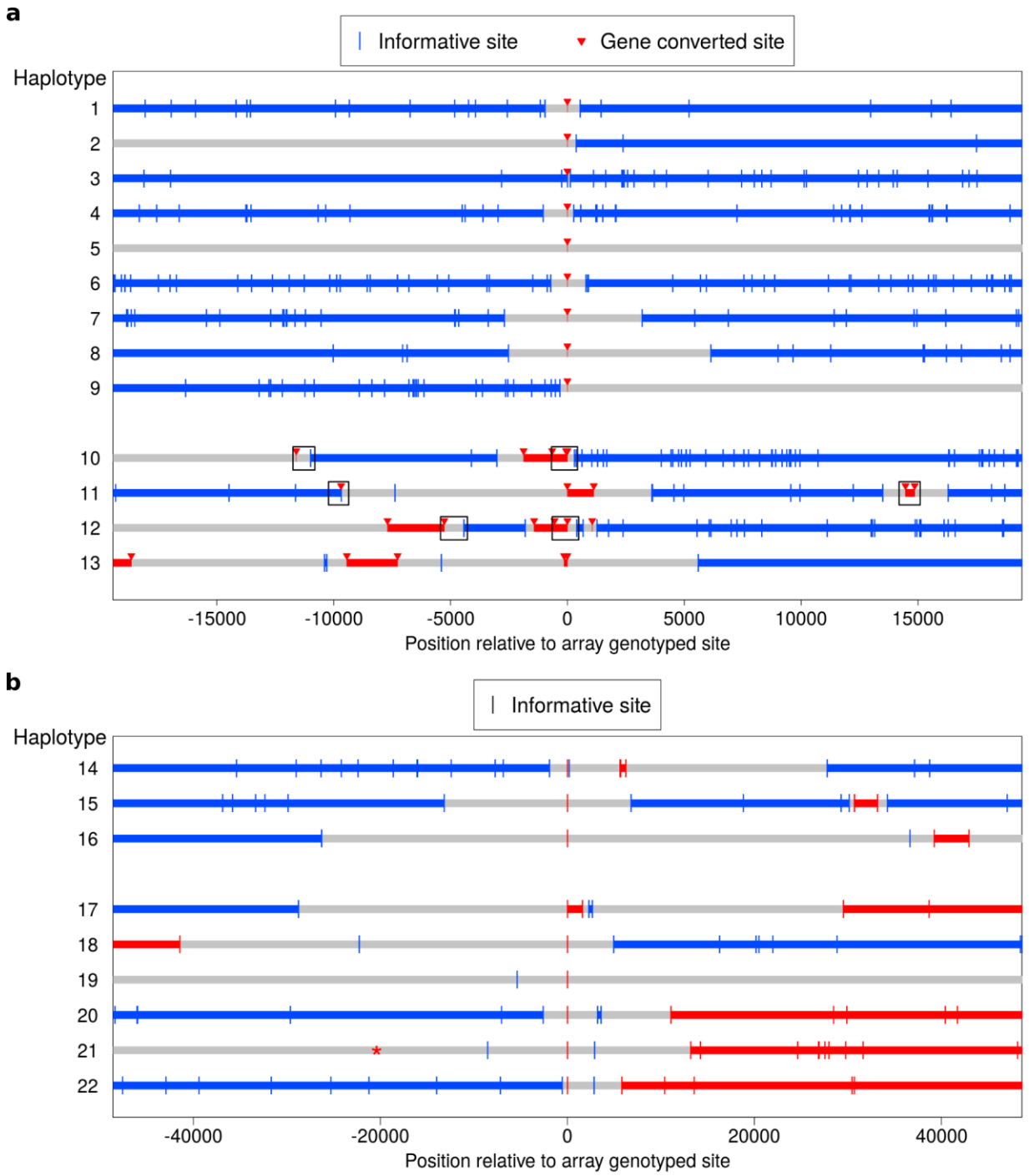

Figure 4. Clustered gene conversion events evident in re-sequence data. a, Recombination patterns derived from whole genome sequence data for the region surrounding 13 gene conversion events originally identified in the SNP array data. Each horizontal line represents a haplotype transmission from a single meiosis, and position 0 on the x-axis corresponds to gene

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

conversion sites identified in the SNP array data. Blue lines depict haplotype segments that derive from the parental homolog transmitted in the wider surrounding region, with blue vertical bars depicting informative sites. Red lines depict segments from the opposite homolog and are putative gene conversion events, with red arrows indicating informative sites. Grey lines are regions that have ambiguous haplotypic origin. For haplotypes 1–9, only a single site exhibits gene conversion. For haplotypes 10–13, several gene conversions appear in a short interval near each other but separated by informative SNPs from the background haplotype. Boxes indicate regions for which we preformed Sanger sequencing (see text). **b**, Clustered recombination events identified in the SNP array data; note the different scale on the x-axis compared with panel **a**. Here, haplotypes 14–16 are clustered gene conversion events while haplotypes 17–22 occur near but not contiguous with crossover events (note the switch in haplotype color between the left and right side of the plot). It is uncertain whether the sites descending from the blue or the red haplotype represent gene conversion events (Methods); thus the plot uses the same symbol for both types of informative sites. Haplotype 19 also appears to have resulted from a crossover, but with informative sites more distant than the range of the plot. Haplotype 21 contains an informative marker that is ambiguous in the third generation and therefore was not detected initially, but it is plotted here with a * symbol. The ambiguous phase in the third generation is consistent with neighboring sites and not indicative of an incorrect genotype call.

bioRxiv preprint doi: https://doi.org/10.1101/009175; this version posted September 18, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

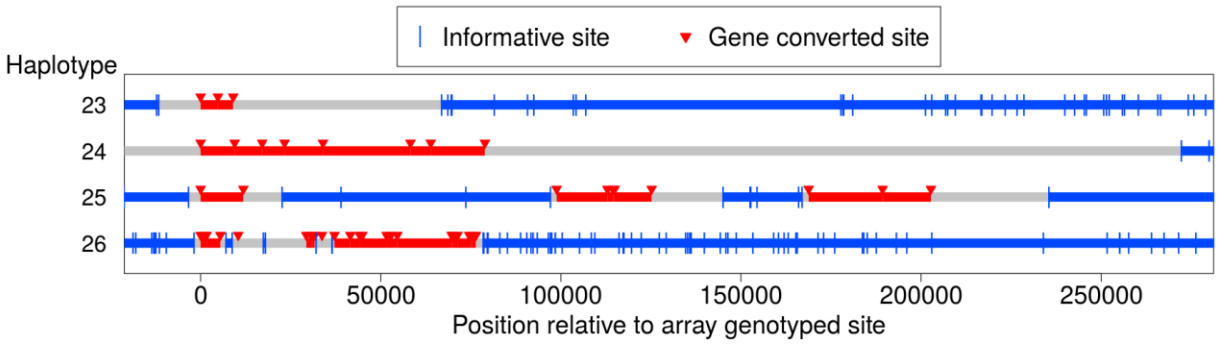

Figure 5. Long-range recombination events observed in sequence data. Shown are two contiguous recombination tracts with length ≥ 9 kb and ≥ 79 kb as well as two sets of clustered long-range recombination events that span ~200 kb and ~76 kb.