Human Genome Structural Variation

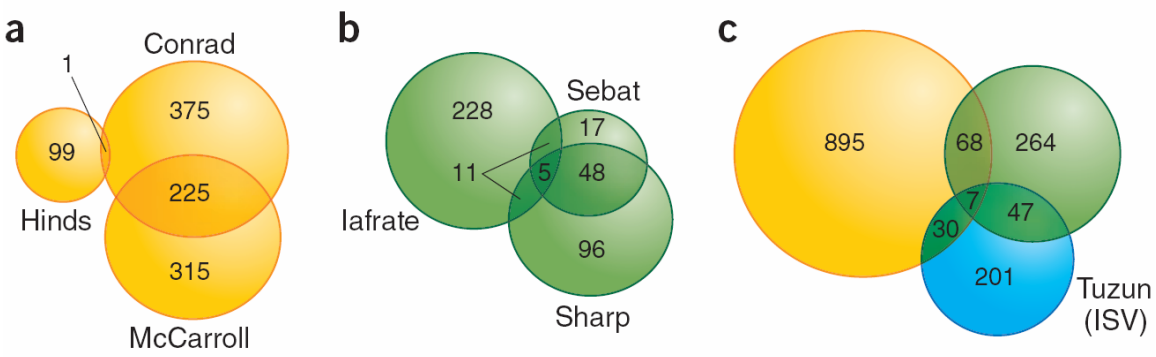
January 23, 2006

Evan Eichler, David Altshuler, Debbie Nickerson,
and members of the Medical Sequencing and Annotating the Human working groups

Rationale. The last three years have brought considerable progress in understanding the nature and patterns of single nucleotide polymorphism within the human species. Of the estimated 10-15 million common SNPs, a large fraction have already been discovered and 3.8 million SNPs converted to genotyping assays, providing the community with a framework to investigate associations between common SNPs and human disease (Consortium 2005; Hinds et al. 2005b). By contrast, our understanding of structural variation within the human genome lags far behind. Several recent publications (Fredman et al. 2004; Iafrate et al. 2004; Sebat et al. 2004; Conrad et al. 2005; Gonzalez et al. 2005; Hinds et al. 2005a; McCarroll et al. 2005; Sharp et al. 2005; Stefansson et al. 2005; Tuzun et al. 2005) have described large scale (>50 kb) and intermediate-size structural variation (>500 bp) in the human genome, revealing that:

1) structural variation is common—two "normal" individuals differ by several hundred insertions, deletions and inversions (>1 kb) ; ~1500 sites in the genome have been identified [or estimated] to have structural variants (Fig. 1).
2) structural variation alters gene structure—over 1,000 genes currently map to or near regions known to harbor structural variation-- and
3) structural variation can be associated with disease and disease susceptibility in the human population (Buckland 2003; Stankiewicz et al. 2003; Gonzalez et al. 2005).
4) structural variation is largely uncharacterized: a comparison of the seven major genome-wide studies of structural variation (Iafrate et al. 2004; Sebat et al. 2004; Conrad et al. 2005; Hinds et al. 2005a; McCarroll et al. 2005; Sharp et al. 2005; Tuzun et al. 2005) show that the vast majority of the identified sites (80%, 2,172/2,721) do not overlap (Figure 1) (Eichler 2006), indicating that most structural variation remains undiscovered. However, it should be noted that current technology is largely driving this description and limits comparison.

Figure 1: Comparison of studies of structural human genome variation based on a) deletion polymorphisms (median size=7 kb), b) Copy-number variants (>100 kb) and c) an intersection of the sum of each with intermediate size variation (~15 kb).



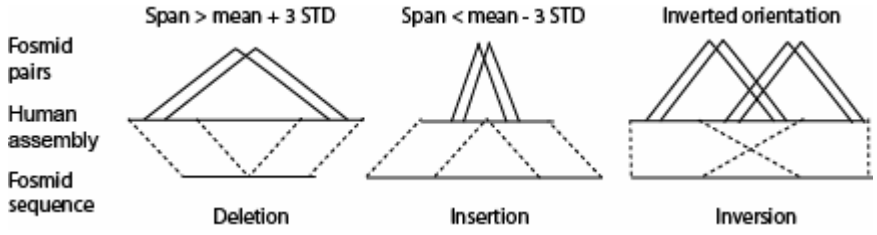
Significance: We propose that the Human Genome Structural Variation Project, be based on a recently developed fosmid paired-end sequencing strategy. We further propose that this effort utilize the DNA samples used by the HapMap Project, resulting in an integrated database that relates common SNP and structural variation.

The data and clone resources generated by the Project will be important, making it possible to provide sequence-level characterization, frequency information and patterns of LD with SNPs for most common forms of human structural variation (90% of the common variation >6 kb). It should be emphasized that this approach will not immediately detect events < 6 kb in length. However, the clone framework that emerges may be used to explore smaller events which are expected to be much more numerous (Bhangale et al. 2005). For example, putative deletion polymorphisms as small as 500 bp in size have been discovered by clustered SNP genotyping error (McCarroll et al., 2006, Conrad et al., 2006) among HapMap samples. These may be subsequently recovered in a specific clone corresponding to that individual to resolve the structure of the event. As structural variants are candidates for disease and disease association, and as whole genome association studies are ongoing, it becomes critical to understand which structural variants can be found by an LD approach, and which cannot. Moreover, this project will provide resources to complement SNP-based studies of human variation (i.e. allow the recovery of specific haplotypes within clones for further characterization), a baseline for proposed studies of somatic structural variation (i.e. Cancer Genome Project), a sequence-defined gold standard for future genotyping and studies of structural variation, and as an ancillary benefit, a much deeper catalogue of SNPs and small structural variants. The impact of structural variation on complex and single Mendelian disease is unknown. Structural variant events (1 kb -10 Mb in size) currently account for 5% of all Mendelian disease within the Human Genome Mutation Database. This is likely an underestimate due to technological limitations in their discovery.

Paired-end Sequence Approach: We propose to extend a recently published strategy to characterize structural variation by the identification of clusters of discordant fosmid sequence pairs (Tuzun et al. 2005). The strategy maps paired end-sequences from a fosmid genome library (representing a single individual) to the human genome reference sequence assembly. This creates a clone tiling-path of the second human genome, identifying discordant regions where multiple fosmids show discrepancy by length and/or orientation: these regions become putative sites of insertion, deletion and inversion. As such, it define the "edges" of each structural variant event. The pilot study cited above was based on the single fosmid library for which a large set of paired-end (~1.1 million fosmid end- sequences pairs) was generated at the Whitehead Center for Genome Research. The analysis and subsequent sequencing demonstrated proof-in-principle of this approach to detect intermediate-sized structural variants ranging in size from *5 kb to 350 kb* (Median=15.2 kb) for a single individual (See attachment #1 pdf, Tuzun2005).

The power of this approach is that it simultaneously links discovery and generation of clones that can be used for sequence resolution. It is unbiased with regard to frequency. We note that the small insert size of fosmids (~40kb) limits their utility in mapping larger structural variations embedded within segmental duplications where long-range continuity is required. As other efforts are screening the HapMap samples with methods such as arrayCGH, ROMA and SNP arrays, many such large-scale variants may be discovered – but without clone resources they will not be understood in detail. ***Thus, the construction of corresponding large-insert BAC clone libraries in a subset of the HapMap samples, followed by sequencing of clones spanning larger-scale variants, should also be considered a long-term priority of the project.***

Fig. 2. Paired-end sequence Approach: Fosmid End-sequence pairs which are discordant by length (> |3 STD|) or by orientation when mapped against the genome assembly flag potential insertions, deletions or inversions. Sequence analysis of 150 fosmid clones, to date, has confirmed ~85% of the structural variation that was detected using a threshold of two more independent fosmids with length discordancy (>|3 STD| beyond the mean) . Fingerprint analysis confirms 30% if a threshold of two standard deviations beyond the mean is used with a bias toward confirmed deletions.



Proposal: In principle, the discovery and analysis of human structural variation involves three straightforward steps 1) identification 2) sequencing to resolve the structure of each variant, and 3) genotyping in larger samples to establish frequency and LD characteristics. While all three are the long-term objectives of the project, the initial effort should focus on developing a plan to sensitively identify variants across the genome, with subsequent steps to sequence structural variants to understand the nature of the base changes, as well as sufficient genotyping to understand the frequency and association patterns of structural variants. One of the most important aspects of this proposal is the construction of a high quality clone framework for additional human genomes which has additional uses beyond the discovery of structural variation (see below).

To put this proposal into practice — in particular the development of a clone resource of this magnitude — raises many scientific and logistical issues. We outline a plan based on eight primary considerations.

1. **DNA Samples:** HapMap samples are unequivocally an ideal source material because (a) they are well characterized, (b) have been consented for genome-wide variation discovery with full data release, and (c) already have been characterized by genotyping of 3.6M SNPs, making it possible to correlate structural and single-nucleotide variation. As an example, these samples have already been characterized for large-scale structural variation by BAC-based arrayCGH (Locke et al.,in preparation) and other efforts are underway to characterize these samples for copy number variants using other platforms. Combined with a bottom-up sequence resource, these samples provide the ideal set for cross-referencing different platforms. Two important considerations are the number and diversity of samples. While a final decision requires a broader and more in-depth consideration, there are good reasons to bias towards African samples to maximize discovery and characterization in the most diverse samples. Studies of SNP variation show that Africans show at least 10% more diversity than Non-African populations, and preliminary studies of structural variation in the HapMap samples shows a similar trend. One proposal would include 48 samples (24 YRI, 12 CEU, 6 JPT and 6 HCB). If we assume 100% sensitivity and a neutral model of evolution, 12 samples from each population will capture 90% of variants with an allele frequency of 10% but will have 95% power to identify variants (>6 kb in length) with a minimum allele frequency of 5% *or greater* (Eberle and Kruglyak 2000). These numbers are likely overly-optimistic, however, since sensitivity will be less than 100%, and there is reason to expect purifying selection (and thus a skew towards lower allele frequencies) among structural variants. As NHGRI has already authorized Agencourt to construct 9 fosmid libraries consisting of 5 YRI, 2 CEU, 1 CHB and 1 JPT, we propose that 19 additional YRI, 10 CEU, 5 CHB and 5 JPT be prepared (a total of 39 additional libraries). Samples should be selected based on maximal SNP diversity within respective populations, excluding samples where cell line artifacts had been documented and based on representation within the ENCODE sequencing project. Samples should all be female to provide comparable coverage of the X chromosomes and selected from offspring of parent-offspring trios so that heritability of the structural variant can be assessed. We

recognize that larger structural variation within complex regions of the genome (Y chromosome and segmental duplications) will not be adequately captured by a fosmid paired-end approach. Such variation (eg Spinal muscular atrophy region, Schmutz, 2004) is typically embedded in high identity sequence and requires long range continuity in large insert vectors to traverse (Rozen et al. 2003; Skaletsky et al. 2003). We propose the development of a complementary set of ***14 BAC libraries*** from HapMap males corresponding to the 14 major branches of the Y chromosome genealogy(Consortium 2002). Library construction of such a BAC resource is a more long-term endeavour, but once developed could be end-sequenced and mapped against the human assembly as described for the BACs. Due to the larger insert size, this would entail far fewer clones per individual (~180,000 end-sequence pairs). These data would complement the fosmid analysis and serve two long-term goals: characterizing larger structural variants (>100 kb) and allow comprehensive characterization of human Y chromosome diversity.

2) **Fosmid Library Construction and End-Sequencing.** In addition to the 9 libraries currently being end-sequenced, we propose that 39 arrayed fosmid libraries mentioned above should be constructed at a depth of 12-fold physical coverage per individual (~ 1 million fosmid clones per individual). End-sequencing of 960,000 clones @ (1.92 million reads) per individual will provide >99% genome coverage of each haplotype (6 fold coverage) (Lander Waterman, 1988). The effective sequence coverage that can be mapped to a best location in the genome is less (>98%) than this theoretical estimate due to the fact that ~30% of clones will not map uniquely due to genome repeat structure, insufficient sequence quality or lack of paired-end sequence data (singletons). An analysis of one fosmid library (Tuzun et al. 2005) indicates that the majority of clones that do not map to a best location are low quality or singletons. Increased read length, sequence quality and mate-pair correspondence will increase mapping power. The most important parameter for fosmid library construction is careful attention to library insert size distribution. Since the number of structural variants will increase logarithmically as a function of reduced size, narrowing the insert size distribution will significantly increase the yield of structural variant sites. Agencourt has recently demonstrated that standard deviations as low as 1.8 kb (as opposed to the standard 2.5-3.5 kb) may be obtained with double pulsed-field gel electrophoresis preparations. This will allow smaller variants (>6 kb) to be detected. Additional recovery will arise from combining libraries because minor allele variants within the human genome reference will become transparent.

   A substantial ancillary benefit of this proposal is that SNPs and small structural variants can be mined from this data, creating a SNP database that is nearly complete to minor allele frequencies of 5% and below.

3) **Clone Distribution.** There will be at least two types of demands for clones from this resource: (a) sequencing of clones discordant by length or by orientation to define inversions, deletions and insertions, and (b) follow-up of positive “hits” in whole genome or candidate gene association studies, allowing rapid and complete characterization of all SNP and structural variants on the associated haplotype(s).

Given that the latter use will result in a steady demand for clone libraries, means of distributing these libraries and individual clones should be developed. Existing BAC distribution centers or long-term repositories such as Coriell are possible options. Quality control of this resource is key. Ideally, given the investment in sequencing and value of the clones, each individual library (~1 million clones) corresponding to 2,500 plates, would be stored for distribution (requiring, on average, 1 large 28 cubic ft -80 C freezer per individual). A staged approach may reduce the number of clones that need to be stored. For example, structural variant clones could be readily identified and a minimum tiled framework could be identified per haplotype (combining HapMap SNP and end-sequence data) to rearray a subset of clones.

4) **Structural Variation Detection.** All end sequence pairs will be mapped to the genome. An initial algorithm has been developed but there is certainly need for further improvement. The number of clones corresponding to sites of structural variation will be substantial (over a hundred per individual, resulting in several thousand locations at which at least one individual is variant). Algorithmic improvements are required to detect structural variants duplicated sequence, to capture insertion events > 40 kb, and to detect smaller variants (<6 kb). Conservatively, the current threshold is 8 kb, but with improvements in library preparation mentioned above, and with many-fold coverage of different individuals, discovery of variants >6 kb should become routine (particularly for those seen in more than one individual). We recommend that these 48 samples should be established as a reference set for cross-platform validation. As the same samples are being intensively studied by both array-based platforms (Nimblegen, BAC-based array CGH) and sequence based platforms (i.e. SNP genotyping errors detection), these "gold standard" clones would provide further value to the community. Specifically, sequence-validated structural variants would provide a better understanding of false positives and false negatives and to provide a standard for new technology development.

5) **Sequencing of Structural Variant Clones.** Structural variant clones should be characterized (i.e. fingerprinting to eliminate rearranged clones) and sequenced completely to confirm the nature of the structural variation. The goal should be to sequence these inserts to *high quality standard.* This is of paramount importance since available sequence data has shown that fosmids fall into two general types a) those representing clear-cut insertion/deletions and b) those embedded in complex regions of the genome including inversions. Sequencing of clones in complex regions of the genome (tandem duplicate gene families such as CYP2D6, GSTT2, etc) will require additional effort and resources. In addition to structural variation, analysis will likely uncover regions of misassembly or collapse within the human genome reference and will, therefore, enhance the baseline human sequence as well. Whether all variant clones need be sequenced, or a subset, cannot be determined until the diversity of variant types is characterized.

6) **Deeper Sequencing of a Set of Regions as a 'Gold Standard'.** A valuable component of the HapMap Project was selection and deeper sequencing / genotyping of a set of 10 ENCODE regions as a "gold standard" to define patterns of variation

over long, contiguous regions, and against which the quality and completeness of the genome-wide HapMap could be evaluated. While valuable, this project was not able to completely sample all SNP variation, nor made any attempt to characterize structural variants, as it was based on PCR-based sequencing of diploid 500 bp fragments.

We thus propose that the current project include selection of the same or a similar set of representative regions, and that these regions be completely sequenced from haploid clones by identification and sequencing of a tiling path of fosmids (and/or BACs) from each haplotype from a collection of the HapMap DNA samples. (The size and scale of the HapMap ENCODE project -- 5 Mb of regions sequenced in 48 individuals -- would seem a minimum projected scale for such a project.) The completion of this arm of the Project would extend understanding of human DNA sequence variation to a level of unbiased detail that has yet to be attempted. Such a set of sequenced regions would be a gold-standard reference that would allow typing technologies to be compared for their ability to detect various kinds of structural variants.

7) **Genotyping**. A major goal should be to genotype discovered variants in the full set of HapMap samples, providing an integrated map of SNP and structural variation. Currently, no single technology could adequately genotype all forms of structural variation. Thus, in addition to exploiting current technologies, a complementary effort should be initiated to explore novel technologies and improvements in existing technology. We emphasize that this technology development and deployment as validated should proceed concurrently with identification and discovery genome-wide. As above, these 48 samples and the corresponding sequenced variants will serve as a reference for analysis and evaluation of genotyping methods, with cross-platform validation the key. Among the many different assays that have been promulgated to detect copy number differences are array CGH, Nimblegen custom fine tiling path, Illumina BeadARRAY, mendelian inconsistencies in SNP data (deletions), with quantitative information on copy number provided by methods such as RT-PCR, DASH, QMPSF, MAPH, and MLPA, and known breakpoints detected using PCR/RFLP or array CGH. (Virtually no genotyping technology has been developed to rapidly genotyping inversions.) As structural variants are sequence resolved, it should be straightforward to directly detect breakpoints using existing SNP genotyping methods. PCR-based capture and sequencing may represent a gold standard.

As genotyping methods for structural variants are validated and proven cost-effective, a major objective should be to genotype these variants in the full set of HapMap samples. This will allow correlation of single nucleotide and structural variation, provide an assessment of the allele frequency spectrum and help understand the frequency of recurrence. While genotyping costs are a significant issue, identifying that subset of common structural variants that show LD to nearby SNPs can reduce costs in subsequent association studies (since the information may be indirectly obtained through associated SNPs). The immediate benefit to the human genetics

community will be twofold: 1) distinction of structural variants in linkage disequilibrium with flanking SNPs from those that may have arisen due to recurrent mutational events and 2) the development of a robust set of genotyping assays for subsequent disease association studies.

8) **Data coordination and analysis.** We recommend that data coordination and analysis teams be established, mirroring successful models used by the International HapMap Project and other large genome projects. Third parties not linked to a particular platform, Center or approach should be engaged to establish and monitor QA/QC procedures, and disseminate the results. A long-term goal should be to establish quality standards for each assay and to incorporate these data along with genotype frequency information into the Data Coordination Center and dbSNP.

9) **Public Dissemination.** All corresponding fosmid end-sequences and assembled fosmid insert sequences must be deposited in the trace repository and Genbank, respectively, according to the Bermuda and Ft. Lauderdale standards. Quality standards for genotyping and sequencing data are critical. Similarly, public dissemination will benefit from deposition of data on genome browsers (UCSC, ENSEMBL) (i.e. humanparalogy.gs.washington.edu/ structuralvariation and http://tcag.edu). Broad public dissemination of data will facilitate inclusion of these variants in association studies and in detailed functional analyses of candidate regions.

Capacity and Production: We outline the necessary sequence capacity and summarize other logistical issues with respect to this project:

- End-sequence pairs: 960,000 pairs X 2 ends/pair X 39 individuals X 1.1 (pass rate of 90%)= ~85 million reads; an additional 5 million reads (180,000 pairs X 2ends/pair X 14 individuals) would be generated from BACs for a total of ~90 million reads of sequence. Fosmid end-sequencing costs are estimated to be 30-50% more expensive than standard plasmid WGS. This would entail 15-18 months capacity of one of the current large sequencing centers (A. Felsenfeld, personal communication)
- Fosmids: We estimate that ~5000 fosmids will need to be completely sequenced to follow up on structural variants that are detected (item 5 above). However, this estimate is highly dependent on the allele frequency distribution of variants which is largely unknown and may increase with a tighter library insert size distribution. Fosmid sequencing requires high quality sequence (>8 X sequence redundancy). We project 500 reads per fosmid sequencing project (average Q20 length =650). An addition of 2.5 million reads are required, although this may increase as additional demand for the clone framework emerges such as complete sequencing of specific haplotypes. For the structural variation analysis, it is important to distinguish standard fosmid sequencing from more problematic clones which will be enriched in this set (repeats, duplicated gene families, etc). Our preliminary analysis of 150 fosmid sequencing projects suggests that 40% of the clones will belong to the latter category and require additional effort to finish.

For item 6 above (sequencing a set of fosmids from defined regions as a gold standard), we estimate that a total of about 2M Q20 kb (approximately 3M reads) will be required, assuming ~8X coverage, 5Mb of regions, and 48 individuals. Additional effort will be entailed in picking tiling paths and finishing the proportion of clones that cannot be sequenced to high quality with 8X coverage.

- Storage: Each individual library (~1 million clones) corresponds to 2,500 plates and will require on average, 1 large 28 cubic ft -80 C freezer per individual. A coordination and distribution center should be established including the costs of ~48 freezers for this purpose. The strength of this proposal is contingent upon rapid and efficient access to the underlying clones.

References

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14:59-69

Buckland PR (2003) Polymorphically duplicated genes: their relevance to phenotypic variation in humans. Ann Med 35:308-15

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2005) A high-resolution survey of deletion polymorphisms in the human genome. Nat Genet

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-320

The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339-48

Eberle MA, Kruglyak L (2000) An analysis of strategies for discovery of single-nucleotide polymorphisms. Genet Epidemiol 19 Suppl 1:S29-35

Eichler EE (2006) Widening the spectrum of human genetic variation. Nat Genet 38:9-11

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861-6

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell R J, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. Science

Hinds DA, Kloek AP, Frazer KA (2005a) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005b) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-9

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2005) Common deletion polymorphisms in the human genome. Nat Genet

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of massive palindromes in human and ape Y chromosomes. Nature

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525-8

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78-88

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier LW, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A (2003) The male-specific region of the human Y chromosome: A mosaic of discrete sequence classes. Nature

Stankiewicz P, Inoue K, Bi W, Walz K, Park SS, Kurotaki N, Shaw CJ, Fonseca P, Yan J, Lee JA, Khajavi M, Lupski JR (2003) Genomic disorders: genome architecture results in susceptibility to DNA rearrangements causing common human traits. Cold Spring Harb Symp Quant Biol 68:445-54

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, et al. (2005) A common inversion under selection in Europeans. Nat Genet 37:129-37

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727-32