Published in final edited form as:

Mol Genet Metab. 2012 March ; 105(3): 382–389. doi:10.1016/j.ymgme.2011.12.014.

Recombination mapping using Boolean logic and high-density SNP genotyping for exome sequence filtering

Thomas C. Markello^{a,b,*}, Ted Han^{a,1}, Hannah Carlson-Donohoe^c, Chidi Ahaghotu^a, Ursula Harper^d, MaryPat Jones^d, Settara Chandrasekharappa^d, Yair Anikster^{a,e}, David R. Adams^{a,b}, NISC Comparative Sequencing Program^f, William A. Gahl^{a,b}, and Cornelius F. Boerkoel^b

^aNational Human Genome Research Institute, NIH, Bethesda, MD 20892-1611, USA

^bNIH Undiagnosed Diseases Program, Office of the Director, NIH, Bethesda, MD 20892-1851, USA

^cUniversity of Minnesota School of Medicine, Minneapolis, MN, 55455, USA

^dGenomics Core, Genome Technology Branch, NHGRI, NIH, Bethesda, MD, USA

^eThe Edmond and Lily Safra Children's Hospital, Sheba Medical Center, and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

^fNational Intramural Sequencing Center, Rockville, MD 20852, USA

Abstract

Whole genome sequence data for small pedigrees has been shown to provide sufficient information to resolve detailed haplotypes in small pedigrees. Using such information, recombinations can be mapped onto chromosomes, compared with the segregation of a disease of interest and used to filter genome sequence variants. We now show that relatively inexpensive SNP array data from small pedigrees can be used in a similar manner to provide a means of identifying regions of interest in exome sequencing projects. We demonstrate that in those situations where one can assume complete penetrance and parental DNA is available, SNP recombination mapping using Boolean logic identifies chromosomal regions identical to those detected by multipoint linkage using microsatellites but with much less computation. We further show that this approach is successful because the probability of a double crossover between informative SNP loci is negligible. Our observations provide a rationale for using SNP arrays and recombination mapping as a rapid and cost-effective means of incorporating chromosome segregation information into exome sequencing projects intended for disease-gene identification.

Keywords

Linkage; Recombination; Mapping; Exome sequencing; Single nucleotide variants

*Corresponding author at: Building 10/10C107, 10 Center Drive, NHGRI, NIH, Bethesda MD 20892-1851, USA. Fax: +1 301 480 3015. markellot@mail.nih.gov (T.C. Markello).

¹Current address: GeneDx, Gaithersburg, MD, 20877-2142, USA.

Conflicts and disclosures

None.

1. Introduction

Identification of genetic loci segregating with disease is often crucial for identification of sequence variants associated with molecularly uncharacterized diseases. Such disease-associated loci have traditionally focused subsequent molecular analyses; however, with the advent of genome and exome sequencing, they are extremely helpful for prioritizing or filtering the thousands of sequence variants detected.

Traditionally, potential disease loci have been identified by linkage analysis using a relatively small number of polymorphic, widely spaced genetic markers. These markers were initially simple biochemical or cytogenetic polymorphisms, then restriction fragment length polymorphisms (RFLPs) or variable number of tandem repeats (VNTRs) and finally microsatellite markers (small tandem repeats or STRs). Recently, widely spaced groups of tightly linked single nucleotide polymorphisms (SNPs) have been employed as genetic markers for linkage studies [1]. These surrogate and actual genotyping techniques produced the first autosomal gene linkage for the Duffy blood group on chromosome 1 [2], later for the CFTR locus on chromosome 7 [3,4] and continue to provide valuable results [5]. These classical linkage strategies generally use widely spaced genotyping markers initially and then follow that with fine mapping. Although computer algorithms such as LIPED [6], MENDEL [7], MLINK [8] and GENEHUNTER [9] have been written to use pedigree and genotyping information to identify genomic regions segregating with the disease phenotype [6], computational limitations and difficulty coping with markers in strong linkage disequilibrium limits use of high density markers in a one-step mapping strategy.

Recent advances in the use of high-density genomic data, such as that obtained by whole genome sequencing, have provided a new means of identifying recombination sites. Whole genome data has been used to map recombination intervals and establish detailed haplotypes using small pedigrees. Recombination intervals can be defined precisely and reduce the “search space for candidate genes” [10,11]. Although methods for using SNP data extracted from exome sequence have also been developed [12], exome sequencing provides less contiguous data; large unsequenced introns and intergenic spaces will not provide variants to use for recombination mapping. However, since high-density SNP arrays profile these regions, we hypothesized that using such SNP data would solve this problem at a lower per-individual cost than genome sequencing. Additionally, if SNP arrays could be used for recombination mapping, they would also delineate chromosomal regions of homozygosity [13,14], duplications/deletions, and mosaicism.

The goal of such recombination mapping using SNPs would be to determine all the cases where a series of loci that segregates in a manner consistent with a fully penetrant Mendelian model abruptly changes to a series of loci that is inconsistent with that model. For sufficiently closely spaced loci, these transitions will directly and completely reflect the meiotic recombinations in the parental gametes. In a human genome of length 3200 centiMorgans (cM), a single crossover should occur on average every 30 to 32 Mb if one assumes 1 cM is approximately equivalent to 1 Mb. The random walk probability of a mean free path length between two recombinations, however, makes two recombination events in smaller intervals probable with some real frequency. Such double crossover events have the potential to complicate SNP-based recombination mapping. For any specific pedigree and model of inheritance, the presence of a double crossover occurring between two informative SNP markers will cause a segregation analysis (even homozygosity mapping) to fail to exclude or include genomic regions that could contain the disease-causing mutation. With increasing marker density, however, we hypothesized that there will be a point at which the markers are so close together that the possibility of a double crossover may be ignored with confidence. The question, therefore, is whether the informative SNPs in a high-density SNP

array form a dense-enough array of markers to replicate the situation for whole-genome data, i.e., where double crossovers can be ignored during recombination mapping.

Assuming full penetrance, DNA from both parents, and absence of double recombinants between informative SNPs, we further hypothesized that a Boolean logic strategy applied to high density SNP data would produce the same results as a classical linkage strategy, require minimal computational infrastructure, and integrate readily into a SNP array analysis pipeline (e.g., as a plug-in software to the Illumina Genome Studio). This Boolean logic approach to recombination mapping makes use of logical rule sets that are straightforward to design and implement for small pedigrees. Boolean logic operations can produce the truth tables for all logical genotype states that are, or are not, consistent with full penetrance and Mendelian segregation for a given phenotype at any single genotyped position in the human genome.

To test these hypotheses, we used high density SNP array data and Boolean logic to map recombinations in a manner analogous to that previously shown for whole genome data and show that double crossovers occur infrequently enough that they can be ignored.

2. Results

2.1. Boolean logic analysis of informative SNPs

To determine all the cases where a series of SNP loci segregating in a manner consistent with a fully penetrant Mendelian model abruptly changed to a series of loci inconsistent with that model, we applied a Boolean methodology applicable to all small nuclear families with available parents and all possible inheritance patterns. For example, to identify candidate disease genomic regions for small families with one or two affected children and unaffected parents, we assumed autosomal recessive inheritance and constructed the Boolean expressions shown in Figs. 1a and b. For families with only one affected individual, the only informative loci were those with heterozygous genotypes for both parents (Fig. 1a). In contrast, for families with more than one affected sibling, the Boolean logic also excluded association with disease for SNP loci at which one parent was heterozygous and the other was homozygous (Fig. 1b). Thus, when a second affected sibling is present a larger proportion of genotyped SNPs can be incorporated into the expression for excluding a region.

Table 1 shows the number of informative SNPs in the Boolean segregation analysis for each of 34 UDP families with participating parents and one or more participating children. The total number of informative loci is equal to the number of SNPs where both parents are heterozygous when there is only one affected sibling but includes other parental genotypes when there are two or more affected siblings. For all chip types and all families, there were between 40,000 and 46,000 informative SNPs with the exception of UDP1647, a family in which the parents had significant consanguinity. Even for that family, however, there were more than 33,000 informative loci and the average inter-locus spacing was <82 kb. As the number of affected siblings increased, the number of informative loci ranged from 171,000 to 459,000. This SNP density provided an average inter-locus distance of <20 kb; with the exception of a small number of regions not interrogated by the SNP chip, few intervals approached 100 kb in size.

2.2. Simulation defines a 100 kb minimum double recombination distance in 5 generations

Having shown that recombination mapping can define candidate genomic regions of inclusion or exclusion for disease genes, we addressed the risk of a double recombination causing recombination mapping to fail to include genomic regions containing the disease-causing mutation. For our test, we focused on chromosome 6, which has an estimated

recombination rate of 2.426 crossovers per meiosis for females and of 1.376 crossovers per meiosis for males [11,15]. Assuming that these rates were uniform along the chromosome and that there was no interference between crossovers (e.g., Haldane mapping function) [16], we performed a Monte Carlo simulation for inheritance of a mutant *NT5E* allele from a common female ancestor over five generations to produce a shared region of identity by descent in five affected siblings (Fig. 2A). Using recombination lengths derived from the HapMap high resolution recombination map [17] and the known position of *NT5E*, we repeated this simulation 7 million times and found that no simulations gave a homozygous interval of less than 100 kb (Fig. 2B), a value much smaller than nearly all inter-locus distances between informative SNPs as defined above and also much smaller than the 22.4-Mb critical region of homozygosity identified in the family of UDP797 [18]. Additionally, if one were to take interference between crossovers into account (e.g., Kosambi mapping function) [16], it would only make the simulated homozygous interval larger and thus a double recombination less likely between informative SNPs.

2.3. The smallest identified double recombination is 215 kb

To substantiate further the low risk of a double recombination between informative SNPs, we asked if a double recombination had ever been reported across an interval as small as 100 kb, and what the interval size of a double crossover might be at a recombination hotspot [19,20] or a recombination jungle [21]. We reviewed the literature using the NCBI Entrez search engine for PubMed and identified only one list of the shortest human double crossovers in Supplemental Table 3 by Fledel-Alon et al. [22]. Of the 27 reported double crossovers in this table, the smallest interval was 710 kb. Examination of our own data identified a double crossover on chromosome 12 with the minimal distance between the two recombinations of 215 kb (Fig. 3). This occurred within a known recombination hot spot that has a recombination fraction nearly 27 times the average (27 cM/Mb). Reassuringly, this double crossover region was readily detected by 23 sequential informative SNPs that occurred between the two crossovers.

2.4. Recombination mapping agrees with multipoint analysis

Having shown that SNP-based recombination mapping was able to identify genomic regions potentially associated with disease and that the risk of a double recombination between informative SNPs was practically negligible, we asked whether the candidate disease loci identified by recombination mapping corresponded with those that would be identified by traditional linkage. Using the family of UDP2753 to test this, we found that the regions of multipoint linkage all coincided with the regions identified by recombination mapping (Fig. 4A).

2.5. Recombination mapping filters exome sequence variants

Having shown concordance of Boolean SNP-based recombination mapping and microsatellite multipoint linkage analysis for the family of UDP2753 (Figs. 4A and B), we asked if the recombination map could effectively filter variants identified by exome sequencing. Assuming complete penetrance and recessive inheritance, we built a Boolean logic strategy for the six family members of UDP2753. Exome sequencing of the six family members produced 112,936 sequence variants. These variants were filtered to 1) exclude common polymorphisms, 2) require segregation by recombination mapping, 3) require autosomal recessive segregation and 4) exclude known polymorphic markers and pseudogenes that appear in all UDP exome data. Following imposition of these filters, the remaining variants were 51025, 4638, 198 and 2, respectively. The last identified a pair of compound heterozygous mutations in *KCTD7* (Ensemble ENST00000275532:c.[682G>T];[812T>A], p.[D229Y];[L271H]) (Fig. 5), a gene that resides on chromosome 7 in a region

detected by both recombination mapping and multipoint linkage analysis and that has been associated with progressive myoclonic epilepsy type 3 (OMIM # 611726) [23].

3. Discussion

We have shown that SNPs from high-density SNP arrays can be used for recombination mapping in a manner analogous to that previously shown for whole genome sequence data [10,11]. The SNP data can be quickly and easily analyzed using Boolean rule sets, which are implementable in commercially available software. The identification of recombination segments consistent with a given segregation model is not confounded by double recombination events. We further showed that the candidate disease loci identified by recombination mapping correspond to those identified by traditional linkage analysis methods, although the identified regions were not able to obtain a consensus-threshold LOD score due to small pedigree size. Lastly, we showed that the candidate disease inclusion or exclusion regions identified by this method enable filtering of variants identified by exome sequencing, and we provide an example in which such filtering identified the disease-causing variant.

The ability to perform Boolean logic recombination mapping on small family pedigrees arose because of three recent and distinct advances in human genetics. The first was the construction of high-resolution recombination maps along with recombination lengths corresponding to physical genetic map lengths. These data are available at the HapMap web site (<http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en>). While the high-resolution HapMap data provide no direct differential recombination data with regard to gender, it can be interpolated using gender specific recombination maps [15]. These maps indicate approximately 100 recombination events for females and 60 events for males per generation. The validity of this difference in gender specific recombination was confirmed in one family using whole genome sequencing [11]. These data define the highest rates of recombination for the genome and give adequate limits to estimate the smallest regions of double crossovers and regions of identity by descent that must be covered by the density of genetic markers; for practical disease discovery experiments, this estimate is roughly 100–200 kb.

We recognize that the precept that no recombination occurs between very closely spaced genetic markers must be violated at the actual site of each recombination; this takes place on average 60 or 100 times in each male or female haploid genome, respectively [11]. With intervals of 68 kb, there will be 40,000 regions in which these 60 or 100 events could happen, but the simple conjoint possibility of two crossovers occurring in a single interval remains vanishingly small, and can be practically discounted, even with a correction for multiple regions and two or more meioses in a pedigree.

The second recent advance in human genetics facilitating SNP-based recombination mapping was the publication of several whole human genomes. There were consistently 2 million variations between Caucasian and Asian genomes and the reference genome, and up to 4.5 million variants for the divergent Yorban African genome [10,24]. The 1000 genomes project (www.1000genomes.org) lists more than 9.8 million single nucleotide variations with minor allele frequencies greater than 1%. Thus, there appear to be adequate numbers of informative genotypes in every diploid human genome to determine the recombination patterns for every meiotic event in small family pedigrees when parents and two or more siblings are genotyped either with a high density SNP array or with whole genome sequencing.

The third advance facilitating SNP-based recombination mapping is the cost/performance improvement in the ability to determine SNP genotypes, first by array oligonucleotide hybridization and now by direct sequencing with second and third generation technologies.

Genotyping chips quietly passed the density necessary to detect enough informative SNPs per small family pedigree at the transition from about 500K to 1 million SNPs per chip. The cost to analyze a family was less than \$1000 in 2010 using 700K chips, which are as informative for locating recombinations and excluding missed double crossovers as higher density chips. Similar prices could be within reach for whole genome sequencing of a quartet in the next few years. Along with improvements in density and cost of variant detection for SNP chips, the capacity of desktop computation has increased to allow Boolean operations to be performed over entire sets of variants for a small pedigree within seconds.

4. Conclusion

In summary, if one can assume complete penetrance and the absence of double recombinants between informative SNPs and has DNA from both parents, then high-density SNP arrays, or the intrinsic density of informative variants obtained during whole genomic sequencing, allow recombination mapping to identify genomic regions included or excluded from association with disease for families with two or more offspring. Application of these recombination maps as a filter for candidate variants from exome sequencing reduces the number by 24 to 92%, even for very small pedigrees.

5. Materials and methods

5.1. Patients and DNA

Individuals enrolled in the study gave informed consent or assent approved by the Institutional Review Board of the National Institutes of Health (Clinical Protocol 76-HG-0238). DNA was prepared by drawing 10 mL of peripheral blood from each individual in 34 families of two parents and two or more siblings, using the Puregene kit (Qiagen, Inc) according to the manufacturer's protocol.

5.2. SNP analysis

Genomic DNA was submitted to the NHGRI core lab and run on one of three generations of SNP oligo arrays: the 1MDuo, the OmniQuad1M or the OmniExpress12 (Illumina Corp, San Diego CA). All members of each family were run on the same type of chip. Each family was analyzed individually. Only SNPs that were genotyped successfully with two alleles in all family members were used for the Boolean analyses. Call rates were >98% before quality control processing and were typically >99.7%.

5.3. Exome sequencing (ES)

DNA from the UDP2753 family (two parents and four children) was submitted to the NIH Intramural Sequencing Center (NISC) for ES. Agilent SureSelect Kits (Agilent Inc. Santa Clara, CA USA) were used for library generation without modification. Sequencing was performed on an Illumina GAIIX. The resulting sequence was assembled using the NISC pipeline. Variants identified as candidates for causing disease were verified with capillary Sanger sequencing in the UDP laboratory and also at a CLIA certified commercial laboratory (GeneDX, Gaithersburg, MD).

5.4. Sanger sequencing

Genomic DNA was amplified by PCR using the primer pairs UDP2753-1-F: CAGCCTTTTGAGCACCCTG and UDP2753-1-R: TTA CTCCAGAGCCACCAAG and UDP2753-2-F: AGCGGATTGTGGAGATCG and UDP2753-2-R: CACACACCCCGATGCACT. The PCR products were purified using the Qiagen PCR purification kit (Qiagen, Valencia, CA) and sequenced by Macrogen (Rockville, MD) using

dye-terminator chemistry (Applied Biosystems, Foster City, CA). Using sequence NM_153033.4 as a reference, we numbered the KCTD7 cDNA sequence such that the adenine of the ATG translation initiation codon is +1.

5.5. Recombination mapping by Boolean logic

Boolean logic trees were constructed in the filter function of Illumina Genome Studio (Version 2010.1 Illumina Inc, San Diego CA, USA) and could be run within seconds on a standard desktop computer (Dell Precision 690 with Windows XP64 or Hewlett Packard xw6600 workstation with Windows XP64). The results of the Boolean analysis yielded all SNPs segregating or not segregating with the phenotype. The ends of blocks of non-segregating SNPs visualized in the genome viewer of Genome Studio define the genomic positions for an exclusion bed file to be used as a filter of sequence variants identified by ES.

5.6. Traditional linkage analysis

For UDP2753, we genotyped all available family members using markers from the Applied Biosystems Prismv2.5 Linkage Mapping Set according to the manufacturer's instructions and Genotyper version 2.5 (Applied Biosystems). We set the disease allele frequency to 0.0001 and the penetrance parameters to zero for carriers and non-carriers and 1.0 for affected individuals. After checking for pedigree/genotyping errors [25], a standard Pedfile was constructed and traditional multipoint linkage analysis was performed on the Biowolf cluster at the NIH intramural computer center using Allegro version 1.1b [26]. We used marker allele frequencies according to the number of observed alleles and all grandparental genotypes were available for both parents. Multipoint linkage was performed across the entire genome. Fine mapping markers were included for chromosome 7. The ABI allele on chromosome 16 (D16s3140) had two alternative primers run to resolve ambiguities in the ABI primer set at this locus.

5.7. Simulation of double recombinations on chromosome 6 at the *NT5E* locus

Based on published recombination frequencies [11,15], chromosome 6 has an estimated recombination rate of 2.426 crossovers per female meiosis and of 1.376 crossovers per male meiosis. We assumed that these recombination rates applied to the *NT5E* locus. Adopting the most conservative or stringent strategy to identify the smallest possible double recombination interval, we chose a mapping function with no interference between crossovers (e.g., Haldane mapping function) [16]. Using the HapMap high resolution recombination map [17], the known position of *NT5E*, and the preceding parameters, we performed a Monte Carlo simulation for inheritance of a mutant *NT5E* allele from a common female ancestor over five generations to produce a shared region of identity by descent in five affected siblings. We repeated the simulation 7 million times.

Acknowledgments

The authors appreciate the advice and assistance of the Genomics Core of the NHGRI. This work was supported by the NIH Undiagnosed Diseases Program (Office of Rare Disease Research, Office of the Director, NIH; NHGRI; and NIH Clinical Center) and by the Intramural Research Program of the NHGRI.

References

1. Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet.* 2005; 77:754–767. [PubMed: 16252236]
2. Donahue RP, Bias WB, Renwick JH, McKusick VA. Probable assignment of the Duffy blood group locus to chromosome 1 in man. *Proc Natl Acad Sci USA.* 1968; 61:949–955. [PubMed: 5246559]

3. Collins FS. Identifying human disease genes by positional cloning. *Harvey Lect.* 1990; 86:149–164. [PubMed: 2152136]
4. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science.* 1989; 245:1066–1073. [PubMed: 2475911]
5. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* 2008; 322:881–888. [PubMed: 18988837]
6. Ott J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet.* 1974; 26:588–597. [PubMed: 4422075]
7. Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet Epidemiol.* 1988; 5:471–472. [PubMed: 3061869]
8. Cottingham RW Jr, Idury RM, Schaffer AA. Faster sequential genetic linkage computations. *Am J Hum Genet.* 1993; 53:252–263. [PubMed: 8317490]
9. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet.* 1996; 58:1347–1363. [PubMed: 8651312]
10. Roach JC, Glusman G, Hubley R, Montsaroff SZ, Holloway AK, Mauldin DE, Srivastava D, Garg V, Pollard KS, Galas DJ, Hood L, Smit AF. Chromosomal haplotypes by genetic phasing of human families. *Am J Hum Genet.* 2011; 89:382–397. [PubMed: 21855840]
11. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science.* 2010; 328:636–639. [PubMed: 20220176]
12. Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.* 2011; 12:R85. [PubMed: 21917141]
13. Jiang H, Orr A, Guernsey DL, Robitaille J, Asselin G, Samuels ME, Dube MP. Application of homozygosity haplotype analysis to genetic mapping with high-density SNP genotype data. *PLoS One.* 2009; 4:e5280. [PubMed: 19399176]
14. Li X, Self SG, Galipeau PC, Paulson TG, Reid BJ. Direct inference of SNP heterozygosity rates and resolution of LOH detection. *PLoS Comput Biol.* 2007; 3:e244. [PubMed: 18052545]
15. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. *Nat Genet.* 2002; 31:241–247. [PubMed: 12053178]
16. Tan YD, Fornage M. Mapping functions. *Genetica.* 2008; 133:235–246. [PubMed: 17934869]
17. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Wayne MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima

- Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–861. [PubMed: 17943122]
18. St Hilaire C, Ziegler SG, Markello TC, Brusco A, Groden C, Gill F, Carlson-Donohoe H, Lederman RJ, Chen MY, Yang D, Siegenthaler MP, Arduino C, Mancini C, Freudenthal B, Stanescu HC, Zdebek AA, Chaganti RK, Nussbaum RL, Kleta R, Gahl WA, Boehm M. NT5E mutations and arterial calcifications. *N Eng J Med*. 2011; 364:432–442.
 19. Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG. Genetic analysis of variation in human meiotic recombination. *PLoS Genetics*. 2005:e1000648.
 20. Meyers S, Bottolo L, Freeman C, McVean G, Donnell P. A fine-scale map of recombination rates and hotspots across the human genome. *Science*. 2005; 310:321–324. [PubMed: 16224025]
 21. Cheung VG, Burdick JT, Hirschmann D, Morley M. Polymorphic variation in human meiotic recombination. *Am J Hum Genet*. 2007; 80:526–530. [PubMed: 17273974]
 22. Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, Coop G, Przeworski M. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet*. 2009; 5:e1000658. [PubMed: 19763175]
 23. Azizieh R, Orduz D, Van Bogaert P, Bouschet T, Rodriguez W, Schiffmann SN, Pirson I, Abramowicz MJ. Progressive myoclonic epilepsy-associated gene KCTD7 is a regulator of potassium conductance in neurons. *Mol Neurobiol*. 2011; 44:111–121. [PubMed: 21710140]
 24. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–59. [PubMed: 18987734]
 25. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*. 1998; 63:259–266. [PubMed: 9634505]

26. Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsson A. Allegro version 2. Nat Genet. 2005; 37:1015–1016. [PubMed: 16195711]

\$watermark-text

\$watermark-text

\$watermark-text

Fig. 1.
Boolean filters for families with two children. a: recombination mapping inclusion filter logic tree for one affected and one unaffected sibling with both parents available for genotyping. b: recombination mapping exclusion filter logic tree for two affected siblings and both parents available for genotyping.

Fig. 2. Monte Carlo simulation of Identity By Descent (IBD) intervals on chromosome 6 for five affected siblings of 3rd cousin parents. A: pedigree depicting segregation of a disease locus on chromosome 6 that gives rise to arterial calcification in five affected siblings of third cousin parents via identity by descent. The chromosome with the mutant locus is depicted in yellow (arrow) and non-mutant chromosomes introduced in subsequent generations are depicted in purple. B: a plot of the frequency versus simulated interval length of conjoint of IBD on chromosome 6 for the five siblings depicted in panel A. Simulations of recombination events occurring through the three generations were performed 7 million times. At each meiotic (gametogenetic) event a random number of recombinations were generated for chromosome 6 that averaged 2.426 crossovers per meiosis for females and 1.376 crossovers per meiosis for males. Inset: Close-up of the region near zero (box). Fewer

than one incidence of an identity by descent interval under 200 kb will occur among 20 ensembles of 64,000 simulations.

\$watermark-text

\$watermark-text

\$watermark-text

Fig. 3.

Candidate region for the smallest double crossover interval detected in humans. Genotypes are shown for SNP positions rs268346, rs7961855, rs7962930 and rs12320939 from top to bottom. The absolute size of the double recombination between rs268346 and rs12320939 is 367,200 bp and the absolute size from rs7961855 to rs7962930 is 215,631 bp with mean size estimate of 295,429 bp for this double recombinant. All four loci were confirmed by Sanger sequencing; the haplotype phases in the parents were derived by genotyping the grandparents.

Fig. 4.

Traditional microsatellite Multipoint Linkage and SNP Based Boolean recombination mapping results for family UDP2735. A: multipoint linkage analysis for 422 ABI markers across the entire human genome of family UDP2753. The chromosome lengths are aligned sequentially along the abscissa. The base pair distances for each chromosome are derived from build 36/hg18 of the human reference genome. Only LOD scores greater than -2 are shown for the multipoint linkage results (black lines). The blue shaded regions depict inclusion regions based on Boolean recombination mapping. B: multipoint linkage analysis for chromosome 7 for family UDP2753 with 21 markers: D7s531, D7s517, D7s513, D7s507, D7s5493, D7s516, D7s484, D7s510, D7s519, D7s502, D7s669, D7s630, D7s657, D7s2504, D7s486, D7s530, D7s640, D7s495, D7s2511, D7s798, D7s550. All LOD scores > -2 are shown, the maximum LOD was 0.815. Regions consistent with Boolean recombination mapping are shaded in yellow. The red arrow gives the approximate location of the *KCTD7* gene, which contains the disease-causing mutations. The base pair distances for chromosome 7 are derived from build 36/hg18 of the human reference genome.

Fig. 5. Segregation of *KCTD7* mutations through the family of UDP2753. The Sanger sequence traces show that the affected individuals were compound heterozygous for the c.682G>T mutation inherited from their father and the c.812T>A mutation inherited from their mother. These *KCTD7* missense mutations respectively encode the amino acid changes p.D229Y and p.L271H.

Table 1

Informative SNPs for 34 UDP families.

UDP#	Affected siblings	Un-affected sibs	1 M duo	Omni quad	Omni express	Total informative SNPs ^a	Included SNPs ^b	Excluded SNPs ^c
369	1	1		✓	✓	40,840	29,728	11,112
658	1	1		✓	✓	40,935	28,782	12,153
798	1	1		✓	✓	43,332	29,610	13,722
984	1	1		✓	✓	49,339	39,759	9,580
1019	1	1		✓	✓	39,983	30,718	9,265
1137	1	1		✓	✓	41,151	30,081	11,070
2700	1	1		✓	✓	40,658	30,936	9,722
2473	1	1		✓	✓	46,227	32,822	13,405
2128	1	1		✓	✓	41,425	31,837	9,588
2518	1	1		✓	✓	41,320	31,583	9,737
3755	1	1		✓	✓	58,496	38,995	19,501
1757	1	1		✓	✓	40,690	32,566	8,124
1418	1	2		✓	✓	49,854	24,726	25,128
2009	1	2		✓	✓	52,020	27,818	24,202
2296	1	2		✓	✓	49,136	30,453	18,683
2717	1	2		✓	✓	49,031	25,450	23,581
1943	1	2		✓	✓	41,569	23,093	18,476
1033	1	2		✓	✓	41,429	24,262	17,167
2717	1	2		✓	✓	40,723	21,078	19,645
563	1	3	✓		✓	57,236	23,964	33,262
1647	1	3			✓	33,033	8872	24,161
1846	1	3		✓		47,351	18,494	28,857
337	2	0		✓		232,324	14,086	218,238
577	2	0	✓			336,998	38,440	298,558
606	2	0		✓		240,237	13,995	226,242
809	2	0			✓	201,820	8638	193,182
929	2	0	✓			272,574	14,078	258,496
2610	2	0		✓		210,177	7899	202,268

\$watermark-text

\$watermark-text

\$watermark-text

UDP#	Affected siblings	Un-affected sibs	1 M duo	Omni quad	Omni express	Total informative SNPs ^a	Included SNPs ^b	Excluded SNPs ^c
2652	2	0	✓			247,013	10,601	236,412
760	2	1	✓			225,076	9603	215,473
2146	2	2		✓		171,721	4759	166,962
2248	2	2	✓			266,797	25,252	241,545
2753	2	2	✓			226,537	5895	220,642
797	5	0	✓			459,277	1791	457,486

^aTotal informative SNP loci have parental genotypes that produce information for recombination mapping either for locus inclusion or exclusion.

^bIncluded SNP loci are parental genotypes that produce information for recombination mapping, and that ultimately designate regions that segregate with the phenotype of interest.

^cExcluded SNP loci are parental genotypes that produce information for recombination mapping, and that ultimately designate regions that do not segregate with the phenotype of interest.