Supporting Online Material

I. FAMILY MEMBER AND DNA SAMPLE CHARACTERISTICS 2

Source of genomic DNA 2
Participant characteristics 2

II. SEQUENCE GENERATION 3

Sequencing and assembly 3
Genotype calls and coverage statistics 4

III. POLYMORPHISMS 5

IV FAMILY GENETICS 6

Inheritance state analysis 6
Recombination analysis 8

IV. ERROR IDENTIFICATION AND ANALYSIS 8

Estimation of error rate 10
Missing data inference 13

V. MUTATION ANALYSIS 14

Inference of deletions from inheritance patterns 14
Candidate selection and resequencing 16
Identification of de novo mutations 16
Calculation of mutation rates 19

VI. ANALYSIS OF MUTATIONS AND DISEASE GENES 21

Detrimental mutations 21
SNP frequencies 23
Disease models 24

VII. SUPPLEMENTAL FIGURES 26

Figure S1. Called coverage in all 4 genomes. 26
Figure S2A. Family genome inheritance analysis. 28
Figure S2B. Inheritance information determines uncalled genotypes. 28
Figure S3. Inheritance blocks emerge from the observed variation. 30
Figure S4. PCA plot with the two parents and the HapMap phase 2 populations. 32
Figure S5. Compound heterozygous candidate genes. 33

VIII. SUPPLEMENTAL TABLES 34

Table S1. Insertions and Deletions 34
Table S2. Coverage and SNP distribution. 35
Table S3. Tabulation of potential de novo *mutations (attached file). 37*
Table S4. Reduction in false positive candidate SNPs using inheritance analysis. 39
Table S5. Signal-to-noise enhancement provided by family sequencing 41

IX. SUPPLEMENTAL REFERENCES 46

I. Family member and DNA sample characteristics

Source of genomic DNA

DNA was isolated directly from unpropagated peripheral nucleated white blood cells of individuals, to avoid any genomic point mutations, indels and rearrangements that may accumulate during the establishment of a cell line, and that may obscure gene-phenotype correlation analyses (S*1, S2*).

Participant characteristics

The clinical characteristics of the affected individuals have been reported previously (S*3, S4*), and as kindred #1 in Ng et al. (S*5*) The parents self-report European ancestry. Principal components analysis of ancestry-informative markers confirms tight parental clustering with the CEU HapMap population (Fig. S4). In the entire body of literature for Miller syndrome, there is neither recorded an instance of intergenerational transmission nor of an instance of consanguinity in affecteds. We tested the genomes of this family to confirm the absence of consanguinity. European genomes, subject to a population history including selective sweeps, contain homozygous blocks ranging from 140 kb to 1.9 Mb (S*6*). Tracts longer than this would suggest consanguinity. In our four individuals, there are no homozygous blocks longer than 1.84 Mb; therefore there is no evidence of consanguinity. Miller syndrome has the following aliases listed in OMIM: "postaxial acrofacial dysostosis," "POADS," and "Genee-Wiedemann syndrome." (S*7*) The two children have a pulmonary phenotype, known by these concurrent exome and genome sequencing studies to be primary ciliary dyskinesia, that is similar to the phenotype of cystic fibrosis. The two children and their mother are heterozygotes for the ΔF508 variation in *CFTR* (confirmed by the Complete Genomics data), but this genotype does not account for the pulmonary phenotype of the children. The pulmonary phenotype is recognized as primary ciliary dyskinesia based on the identification of the recessive mutations in *DNAH5*.

IRB approval was obtained from Seattle Children's Hospital and from the Western Institutional Review Board (#2008.0005). All participants provided written consent.

II. Sequence generation

Sequencing and assembly

Complete Genomics, Inc (Mountain View, CA), used their paired end library preparation and sequencing-by-ligation methodology as described recently (S*8*). The average depth of haploid coverage by mapped reads in the sequenced family was 51x in the mother, 88x in the father, 54x in the daughter and 52x in the son. The resulting called coverage is indicated in Fig. S2. Reads were mapped to the NCBI reference genome (NCBI Build 36.1) or recruited by the mapped mate-pair reads for local *de novo* assembly as well as for determining genotyping calls for each reference position for each genome (S*8*). Data for each genome were delivered as lists of sequence variants (SNPs and short indels) relative to the reference genome accompanied with variant confidence scores.

Libraries for sequencing were generated using a four-adaptor protocol (S*8*). Briefly, sequencing substrates were generated by fragmenting genomic DNA followed by recursive cutting with type IIS restriction enzymes and the insertion of directional adaptors. Hundreds of tandem copies of the resulting circular substrates were then replicated with Phi29 polymerase (RCR). The resulting concatamers, referred to as DNA nanoballs (DNBs), were adsorbed to grid-patterned arrays. An unchained probe-anchor ligation sequencing chemistry (cPAL (S*8*)) was then used to independently read up to 10 bases adjacent to each of the eight anchor insertion sites, resulting in 35-base mate-paired reads (70 bases per DNB). Following background removal and image registration, intensities were extracted from the DNB nanoarrays and used to call and score each base. The resulting mate-paired reads were aligned to the reference genome. This process had a yield in mapped sequence bases in the mother, father, daughter and son of the sequenced family of 143.9 Gb, 249.9 Gb, 152.8 Gb and 148.4 Gb respectively. Average discordance rates within all mapped bases in the data ranged from 2.0% to 2.5% over the four genomes; within the highest-scoring 85% of read bases, the discordance rate (which includes true variations) ranged from 0.57% to 0.64%. Within the staggered reads overlapping each genomic position, up to twenty different probes (ten on each strand) assay each of the four bases; thus, base calling errors are largely uncorrelated across reads. The distribution of the mate gap (the genomic distance in bases between the two paired ends of each read) varied by genomic library; the most

probable mate gap within the four sequenced family members was 423 in the mother, 475 in the father, 390 in the daughter and 339 in the son.

At locations selected for likely differences from the reference sequence, mapped reads were assembled into a best-fit diploid sequence with a custom software suite that implements both Bayesian and de Bruijn graph techniques. For each genome and each location, this process yielded diploid reference, variant or no-calls with associated quality scores (S*8*).

We used variant lists defined by CGI's standard variant confidence score thresholds of 20 decibels for homozygous variations and 40 decibels for heterozygous variations – these balance the rates of false positive variant calls (mostly having lower score) and uncalled positions. Insertions as long as 47 bases, deletions as long as 117 bases, and complex insertion-deletion events as long as 93 bases were called relative to the reference genome.

Genotype calls and coverage statistics

By "genotype" we mean: both alleles at a position. There is one genotype position in each of our sequenced genomes corresponding to each position of the reference genome. At some positions, one but not both alleles of a genotype can be called. These positions are considered partially called; for our summary statistics we tabulate these positions as completely uncalled.

Fig. S1 summarizes the coverage statistics for the four sequenced genomes. The reference genome used for our analysis is NCBI Build 36.1, which contains 2,855,343,769 coordinate positions that are not 'N's. For our analyses related to the exome, we use as our operational definition the set of exons included in the UCSC KnownGenes database. This set of exons is larger than the set of exons contained in the CCDS database. The CCDS database has been used as a reference basis for exome sequencing projects (S*5, S9*). The difference in size of the CCDS (for NCBI Build 36.1: 164,217 exons; 27,961,415 bp) and KnownGenes (for NCBI Build 36.1: 235,386 exons; 79,498,653 bp) databases should be considered in any comparison or meta-analysis of the results reported here with results reported from exome sequencing projects. We

employ the CCDS definition of exome once in this report, in Fig. 2, because in that figure we are evaluating the current utility of exome sequencing for inheritance state prediction.

III. Polymorphisms

A very rare, or novel, SNP is a SNP found neither in dbSNP (build 130) (S*10*), the 1000 Genome Project (Pilot 1 release of April 2009) (S*11*), nor identified in the following genomes Venter, Watson, the first Yoruban and Asian genomes (S*12-15*), and the CNV database (S*16*). For operational purposes, we never consider the allele present in the reference genome as a candidate, regardless of reported frequency. Theoretically, very rare reference alleles (perhaps sequencing errors) should be candidates; future bioinformatic pipelines will consider them.

There are a number of copy number variations (CNVs) between the reference genome and our sequenced genomes. Eighty of the largest of these were identified by Comparative Genomic Hybridization with Agilent chips (Agilent Technologies, Santa Clara, CA) containing 1 million probes, evenly spaced throughout genome. Of these, 38 showed heterozygosity in the parents; none were both rare and gene-spanning. Locations spanning CNVs (either identified with the Agilent chip and/or with the HMM) tend to have an excess of SNPs reported both in dbSNP and in our *de novo* analyses. As a result, our false positive rate for SNPs is elevated in CNVs. Deletions in the reference genome with respect to our sequenced genomes result in false negatives (for SNPs and for gene candidates) over the region of the deletions. *Bioconductor* facilitated SNP analyses (S*17*).

In addition to SNPs as described in the main body of the paper, we identified small deletions at 92,945 positions and small insertions at 85,195 positions, ranging from 1 to 117 bp (Table S1).

The intermarker distance in the main body of the paper of 802 bp is based on SNPs that are heterozygous in at least one member of the family. The inter-SNP distance is shorter than this, as many SNPs (defined by reference to a database) are homozygous in all members of this family. The average inter-SNP distance in this family is 617 bp. If MIEs and state consistency errors are excluded, this distance is 772 bp. Across the four

individuals, the inter-SNP distance excluding MIEs and state consistency errors ranges from 1027 bp to 1067 bp.

IV Family genetics

Inheritance state analysis

An “inheritance state” is the pattern, or topology, of Mendelian allele assortment through a pedigree at a given reference position. An inheritance state is only defined in the context of a pedigree; unlike the related concepts of “phase” and “haplotype”, “inheritance state” cannot have meaning when describing a single genome. For the non-pseudoautosomal regions of the X chromosome, there are two states: nonidentical and haploidentical maternal. For the HMM, to prevent overfitting or subjective bias, for each state we set all the emission probabilities of each consistent allele assortment pattern equal to each other. We set the probability of emitting an inconsistent pattern to 0.5%. These probabilities could also have been set to empirically observed frequencies, using the frequencies of patterns as they occur in blocks selected with a heuristic algorithm on the basis of SNP pattern frequency (Fig. S3), or with an iterative HMM parameter-estimation algorithm. Results with empirical emissions were similar to those with uniformly set emissions; therefore we chose the uniformly set emissions to prevent overfitting. In addition to the four inheritance states, we modeled two additional states in the HMM. One of these states was a “Mendelian inheritance error rich” state; for this state the emission probability of an MIE was set to 30%. The MIE-rich states are likely to contain tracts of sequence that are difficult for the Complete Genomics (Mountain View, CA) technology to accurately report. The second of these states was the “compression/CNV” state. Reads that map to very little diverged repeats or to repeat copies not present in the reference assembly can be mismapped. Since 99.85% of the positions within the family are invariant, any differences between two superimposed repeat copies will appear as heterozygous positions for all individuals (often reported with high minor allele frequencies in dbSNP). The compression state is characterized by an excess of these patterns. The emission frequencies for the compression state were set empirically, with the probability of emitting uniform heterozygosity for all four individuals set to 66% (S*18*). The compression state is likely to include tracts of

sequence that are incorrectly assembled in the reference sequence and/or that are likely to include mismapped reads in the Complete Genomics assembly. The use of the term "CNV" to describe this state would be incomplete, as this state is expected to include many sequence elements of the genome that share high similarity to each other. Some of these cannot be anticipated from analysis of the reference genome because the reference genome includes only a single copy.

Since we identified more than 3 million informative SNPs, all crossover sites could be determined with precision. Our median resolution of 2.6 kb, with a few sites localized within a 30-bp window (Fig. 1), is a substantial improvement over a recently published median resolution of 93 kb (S*19*). The methodologies and data used by the HapMap project to predict recombination rates are distinct from our methodology and data. Therefore the two independent sets of results confirm each other and help to establish that an HMM analysis of inheritance states of complete genome sequence is a sensitive and precise method for determining the boundaries of inheritance state blocks. Our results reaffirm that the majority (~59%) of recombinations occur in and around hotspots of chromosomal recombination (S*20*). In the main text, when we refer to a recombination taking place in a hotspot, the bioinformatic interpretation is that for 92 of the 155 regions (median length 2.6 kb) in which we calculate a recombination to have taken place, they contain a HapMap hotspot or the closest HapMap position is a hotspot.

"Reverse pedigree analysis" has recently been reported as an alternative approach to inheritance-state block identification (S*21*). *SNPtrio* encodes genotyping data in a data structure similar to an inheritance pattern, and infers recombination positions with an heuristic algorithm. Coop et al. report a similar heuristic algorithm (S*19*). HMM algorithm implementations, such as the one we report here, are likely to be slower than *SNPtrio*. However, the difference in speeds of the two algorithms is unlikely to create a bottleneck in any computational pipeline for the identification of recombination locations.

Visual inspection of Figure 2 demonstrates that pedigrees of less than a nuclear family of four (trios or other sets of two or three individuals) produce a signal too noisy for precise prediction of recombination locations and inheritance states. Trios have no informative sites at all. Also, without complete genomic information, less robust inference is possible. Exome information is very fragmented; note that in Figure 2, the exome

information is scaled up for visibility – the intensity of the signal is much less than for complete genome sequence.

Recombination analysis

In principle, recombinations can take place in different meioses at positions that are within a few basepairs or a few kilobasepairs of each other. If these happen in the two meioses of the same parent, leaving too few informative SNPs in between to create a signal that the HMM (or any other algorithm) would recognize as a distinct state block, they are likely to not be observed, leading to an underestimate of the recombination events. They can be observed if they occur in different parents. In only one case did a recombination in the mother and the father occur that would have been too close to resolve had they been in the same parent. The number of recombinations that we missed is likely small, particularly because crossover interference is expected to inhibit such recombinations.

In HapMap data, a hotspot is defined as a region with ≥10 cM/Mb (S*22*). Fig. 1 shows the distribution of maximum recombination values in 1000 Monte Carlo replicates of a same-length set of 155 windows at random locations in the genome. 5.423% of the 155,000 windows have a value >= 10 cM/Mb. A p-value for finding 92 hotspots in our observed data in 155 windows is thus $0.0542392^{92} = 3.5 \times 10^{-117}$. The sex differences in recombination rate are known to be most prominent around the centromeres (S*23*), and, indeed in this dataset, the crossover nearest the centromere was maternal in 20 of 22 autosomes (Fig. 1).

IV. Error identification and analysis

High sequencing error rates, relative to the frequencies of true genetic variants, represent a significant challenge in mining DNA sequence data. We have used our approach to family genome inheritance analysis to identify approximately 70% of the errors in the set of assembled genomes for this family. Mendelian inheritance errors (MIEs) occur when the pattern of alleles observed in a child is inconsistent with assortment of the parental alleles. We observed 59,243 such MIEs, which may be the result of false base calls, mismapped reads, or more rarely, inherited deletions that are

heterozygous in one or both parents. Much more rarely, MIEs may be observed as a result of *de novo* mutations.

Based on the HMM results and the co-occurrence of MIEs consistent with inheritance of deletions, we identified 365 deletions in this family, ranging from 37 bp to >700 kb, with a total of >7 Mb. These inherited deletions explain only 1,761 of the MIEs, however, and all the deletion data together explain about 5,700 of the 59,243 MIEs. Most inherited short indels do not result in an observed MIE (e.g., one hemizygous and one homozygous parent will only result in an MIE if the parents have different alleles).

Adjusting for MIEs explained by inheritance of deletions, and considering that the *de novo* mutation rate is about three orders of magnitude lower than the sequencing error rate (S*24*), the count of MIEs can be reliably used to estimate the rate of genotyping error, as ~1.0×10^{-5} per genotype position per individual. The average error rate, however, is a somewhat misleading concept. At positions that are homozygous in all four individuals (99.85% of all positions), the error rate is only 8.0×10^{-6} (such errors lead to false heterozygosity calls); at variable positions this rate is 1.7×10^{-4} (real variants miscalled as reference or as an incorrect variant). We determined these error rates with several independent methods described below, including exome sequencing and the resequencing of ~60,000 positions, each method yielding consistent estimates.

Overall, ~70-75% of all DNA sequence errors in a four-person pedigree can be detected through family genome inheritance analysis.

Once the inheritance state has been established for a region of the genome, positions with allele patterns inconsistent with the inheritance state of a region can be inferred to be errors (Fig. S2B) (S*25*). These “state consistency errors” almost always derive from sequencing errors or from assembly discordances. Overall about three quarters of all errors in a four-person pedigree can be detected through family genome inheritance analysis. These errors may be corrected through targeted resequencing, or labeled and sequestered (i.e., placing them in a category with distinct confidence statistics) in downstream analyses. By excluding these detected errors, the accuracy of the Complete Genomics sequence in the context of this pedigree rises from 99.999% to 99.9997%. Of the 323,255 novel SNPs we identified, 17,959 were state consistency errors and 40,110

were MIEs. Of the resulting reportable 265,186 SNPs, we estimate a false discovery rate of 14%, determined as described below.

Estimation of error rate

We applied Complete Genomics technology as a resequencing approach for analyzing each of four genomes. In resequencing, a mapping algorithm maps raw reads to the positions of a reference genome. We define a no-call as a report of a genotype of "NN" at a reference position. We define an error, only at positions that are not no-calls, as the report a genotype at a reference position that is not the genotype that would be reported at this position if we knew the true sequences of both haplotypes of an individual's genome.

There are other types of errors that could be considered for a *de novo* sequencing project, or even in a resequencing project. For example, descriptions of inversions and the sequences and locations of novel insertions might be predicted, and these predictions would have associated error statistics. For our main analysis we neither make such predictions nor provide error statistics.

MIEs and state consistency errors are observations that may occur at all positions in the genome (except completely uncovered positions). MIEs and state consistency errors are detected through internal inconsistencies in a data set (i.e., between the genomes of the family members). For this reason, because no external reference is needed to identify them, counts of MIEs and state consistency errors provide a gold-standard estimate for the overall genome error rate. Error estimates based on comparison to external reference sequences or to data obtained via resequencing (or genotyping) by alternative technologies may suffer from biases or errors in the reference sequences or technologies. Most errors arising from sequencing technology imperfections will result in MIEs or state consistency errors. MIEs and state consistency errors may also be observed at positions with a *de novo* mutation, an unrecognized indel in at least one member of the family, or with unusual inheritance events such as uniparental disomy.

The method of error rate detection described in the main text, that of comparing blocks between the two children that are identical by descent (IBD), provides an error estimate based on observations of the frequency that the Complete Genomics technology

produces a genotype call that is inconsistent with a replicate assay of the same technology on the same true genotype. Error rates can also be estimated by comparisons with other technologies.

Across a conservative portion of the exome (as defined by the Consensus CDS project), we could directly compare error rates with an independent data set, as reported in Ng *et al*. (S*5*) For this comparison there were 1377 discordant genotypes out of 49,998,778 positions. Of the 1377 genotypes, 969 can be attributed, based on parental status in the Complete Genomics data, to errors in the exome data; 408 can be attributed to errors in the Complete Genomics data. These data result in an estimated overall error rate across the conservative exome of 8.16×10^{-6} per base pair. This estimate is similar to the estimate derived from MIE counts; the difference can be attributed to the restriction of the analysis to the exome rather than to the entire genome, which might bias the data towards easier-to-sequence regions of the genome.

By repeating error-rate computations over distinct subsets of our data, we could quantify the variation in error rate across different subsets of the genome. The error rate is higher in repetitive sequence, in copy number variations (CNVs), and near telomeres. The error rate in the exome was 8.1×10^{-6}, less than in other regions, and nearly identical to the exome error rate derived from a comparison with the Ng *et al.* dataset.

Given that 99.85% of genomic positions are homozygous and identical in all four individuals, even though the error rate is lower than at positions with some variability, 97% of errors occur at invariant positions. A false call of an allele at an invariant position, if in a child, will be observed as a MIE; if in a parent, it will be observed as a state consistency error ~50% of the time (if the inheritance state permits the observation, as described below). We observed 52,009 MIEs with this “invariant position” error pattern: 22,126 with the unexpected observation in the daughter’s genotype, and 29,883 with the unexpected observation in the son’s genotype. When there is an error in a parental genotype, a state consistency error will be observed when a falsely called base does not appear in either child’s genotype in regions where both alleles of that parent are expected to have been inherited, one to each sibling. Thus, maternal errors in nonidentical and haploidentical paternal blocks will be observed as state consistency

errors, as will paternal errors in nonidentical and haploidentical maternal blocks. We observed 11,300 and 10,690 such state consistency errors, respectively.

Of the approximately 30,000 errors we expect per genome (a total of 120,000 expected errors in all four genomes), we observe 59,243 MIEs and 25,725 state consistency errors, or about 70% of all errors. Therefore, for many types of analyses, our effective error rate (for purposes of computing false positives) is 30% of our genotyping error rate. For example, we do not report novel SNPs at positions that are MIEs or state consistency errors, so rather than a per base false discovery rate of 1.1×10^{-5}, our rate is 3.3×10^{-6}, resulting in about 36,000 falsely reported SNPs across all four genomes analyzed.

We also estimated error from a resequencing analysis of 25 regions of length 200 kb that were chosen at random from the genome. Within these regions, there were 3,188,347 sites with Complete Genomics calls in all 4 individuals. 2,577,718 of these sites could be called with *Maq* in the resequencing data. Among those sites, there were 430 discordances, for a discordance rate of 430 / (2,577,718 x 4) = 4.17×10^{-5} per site per genome. This approximates the expected error rate in the Illumina (San Diego, CA) resequencing process. We can conclude that the Complete Genomics error rate is much less than this rate, which is consistent with a rate of 1.0×10^{-5} to 1.1×10^{-5}.

We can detect variation when as few as one of the four individuals is called at a position, so our predicted false negative rate for SNP discovery is dominated by the percent of genome not covered by the Complete Genomics assembly in any of the four genomes, and is therefore approximately 5% (3% for the exome) (Table S2). In the remaining 95% of the genome, our false negative rate for detection of recessive positions is approximately 1.7×10^{-4}, which is our estimated genotyping error rate at heterozygous sites. We expect 1.1×10^{-5} x 2,855,343,769 $\approx$ 30,200 errors per genome that might create false positive observations of novel SNPs, or a total of ~121,000 falsely reported novel SNPs for our set of four genomes. We observe 323,255 novel SNPs, including these false positive errors. Therefore, without identification of MIEs (40,110 of them at novel SNP locations), the false discovery rate would be 29%. With identification of MIEs, but not state consistency errors (17,959 of them at novel SNP locations), only possible with information from the four-person pedigree, the false discovery rate for SNPs would be

19%. Because we can identify both MIEs and state consistency errors, our actual false discovery rate for novel SNPs is 14%. This rate of 14% comes from ~18,000 undetected errors per parental genome. This false discovery rate of 14% is relatively high because relatively few true SNPs are novel, so even with a low sequencing error rate and with detection of approximately 70% of these errors, a number of sequencing errors are still reported as novel SNPs. The false discovery rate for SNPs in a single genome sequenced with the same technology without the context of a pedigree might be ~16% (~18,000 false SNPs per 110,000 reported SNPS). Our "raw" false discovery rate of 29% is nearly twice that because we are reporting the combined rate from two founders (the parents). These calculations assume that sites that are homozygous and different from the reference sequence are not reported as novel SNPs. Conservative analysis, as we perform for sites uniformly homozygous in our pedigree of four, suggests that the reference sequence may be in error at these positions.

The reference assembly is reported as a haplotype. Historical error rates with respect to genome sequencing have generally been based on haplotypic error. An individual's genome contains two haplotypes. Unless otherwise specified, we report error rates as per individual per position (i.e., per genotype). Since an error at either haplotype at a given position will result in a genotype error, if we reported a haplotype error rate, it would be about half our reported genotype error rate.

Our estimates of error rates are more accurate than similar estimates made for genotyping arrays. The most informative positions for estimating error rate are those that are invariant or have low heterozygosity. Such positions are relatively rare on genotyping arrays because array data is centered on common SNPs, and so are enriched for positions at which some individuals in a pedigree are heterozygous. The high heterozygosity in genotyping studies also degrades the fraction of detectable MIEs and state consistency errors in such studies.

Missing data inference

For our identification and prioritization of disease candidate alleles, we infer missing data using the called alleles of other individuals together with inheritance state. The *CES1* candidate variants were uncalled in three of four individuals, and were called as a

homozygous probably detrimental very rare allele in one of the children. Because *CES1* is in an identical block (Fig. 1), missing data inference could confidently infer that the other child was also homozygous for the probably detrimental very rare allele, and that the parents were most likely heterozygous, making *CES1* a strong candidate. One of the *DNAH5* candidate variants was uncalled in one of the parents. Because *DNAH5* is in an identical block (Fig. 1), the presence of a probably detrimental very rare allele in this parent could be confidently inferred, making *DNAH5* a strong candidate. If the search were to have been restricted to only fully called positions in all 4 individuals, these two candidates would have been missed. However, in the particular dataset reported here, at least one of these candidates (*DNAH5*) would have been found with high confidence even without allele inference, as in this case there was adequate information present in the called genotypes of the two children alone.

V. Mutation analysis

Inference of deletions from inheritance patterns

The presence of a hemizygous deletion, unreported by the sequence technology, in one or both of the parents at a site with at least two alleles can result in an observed MIE. For example, the observed inheritance pattern [aa, ab, aa, bb] (the genotype order in the pattern is [mother/parent1, father/parent2, daughter/child1, son/child2]) is an observed MIE. It can arise from a sequence error, but it could also represent a hemizygous deletion in parent 1 that was inherited by child 2, the real underlying inheritance pattern being [a-, ab, aa, b-]. This particular explanation is only possible in a nonidentical state (the children inherit opposite alleles from the parents). Likewise, in a maternal haploidentical state the observed inheritance pattern [aa, bb, aa, aa] could represent the inheritance of a hemizygous deletion in the mother by both children [aa, -b, a-, a-]. Each mode of deletion inheritance results in a specific inheritance pattern. These patterns are observed as MIEs. In these patterns, the donor and recipient of the deletion appear homozygous.

We scanned the genomes for inherited hemizygous deletions by looking for regions containing two or more MIEs that may be explained by the same deletion inheritance pattern and are either immediately adjacent or are separated by non-MIEs consistent

with the (transparent) inheritance of a deletion (the deletion donor and recipient are homozygous). The frequency of long runs of variable sites consistently homozygous in one parent and one child that are supported by just one MIE suggesting hemizygosity in those individuals is very high. The occurrence of two neighboring MIEs consistent with the same deletion inheritance was rare enough to include even those not further supported by flanking non-MIEs showing homozygosity in the individuals that are predicted to have the hemizygous deletions. With a minimum requirement of two neighboring MIEs consistent with the same gap inheritance or one MIE embedded in at least 40 consecutive variable sites homozygous in the deletion carrying individuals, we predicted 366 inherited hemizygous deletions. These are supported by 736 apparent MIEs and cover 8.25 Mbp and 5479 variable sites.

At two locations the predicted gaps helped to refine the location of a crossover site. For example on chromosome 19, the HMM had predicted a transition from the maternal haplo-identical to the non-identical state between position 56,822,642 and 56,842,413, but the MIE-rich region 56,824,238 to 56,840,731 is consistent with a gap inherited from the mother to both children (not possible in the non-identical state), suggesting the crossover to have taken place in the subregion 56,840,731 to 56,842,413 instead.

The accuracy of the hemizygous gap predictions was partially confirmed by our resequencing efforts. Of the 270 MIEs that were confirmed by genotyping according to at least one filter, 226 (84%) are located in the 8.25 Mbp of predicted hemizygous gaps above (see Supplemental table 3). Up to 32 more MIEs could be located in shorter hemizygous deletions (resulting in only a single apparent MIE and spanning fewer than 40 variable sites) that were not included in our set of 736 gap predictions.

Deletions spanning only one or a few informative markers cannot be confidently identified by this analysis. However, such deletions are often identified by the direct results of the assembly of the short read sequences, and do not require computational methods for identification. The accuracy of the hemizygous deletion predictions was partially confirmed by our resequencing efforts. Of the 270 sites with MIE inheritance patterns that were confirmed by at least one genotyping method, 226 (84%) are located in the 8.25 Mbp of predicted hemizygous deletions above. Up to 32 more may be located in shorter gaps that resulted in only a single observed MIE.

Candidate selection and resequencing

We initially identified 49,720 *de novo* mutation candidates among the 2.3 billion bases that were successfully genotyped in each parent-offspring trio (2,333,121,607 bases in the daughter and 2,336,234,940 bases in the son). We then excluded positions for which unique probes could not be successfully designed and positions in error-prone and compression states, resulting in 33,937 potential mutations among 1,825,738,754 bases in the daughter and 1,830,066,433 bases in the son (total of 3,655,805,187 bases). Next we designed a custom Agilent SureSelect array with ~1 million features to capture the regions surrounding the 33,937 candidates (as described in Ng et al.) (S*9*). In addition to these regions, we selected probes for 25 regions of length 200 kb that were chosen at random from the genome for the purpose of obtaining an additional empirical estimate of sequencing error. Genomic DNA samples were sonicated, ligated to adapters suitable for subsequent sequencing on the Illumina GA2 (Illumina Inc., San Diego, CA), size-selected, amplified, and hybridized to the array.[7] Four 76-base and one 36-base lanes of DNA captured and released from the array were sequenced on the Illumina GA2, and aligned to the NCBI 36.1 reference genome using the *ELAND (extended)* pipeline (Illumina, Inc.). An average of 931 Mb of filtered data were successfully aligned by *ELAND* for each genome, based on the summary statistics for the sequencing runs.

Identification of de novo mutations

To define the list of candidate *de novo* mutations we used three base-calling algorithms with the Illumina sequencing data: 1) the default settings of *Maq* (Illumina; http://maq.sourceforge.net/maq-man.shtml) (S*26*), 2) the *perl* script maq.pl (co-distributed with *Maq*; maq.pl applies additional filters), and 3) a binomial method described immediately below.

To perform the binomial method, we identified all the genomic positions at each genome covered by at least eight reads. For each position, we tabulated the number of reads supporting an A, C, G or T call relative to the top strand of the reference sequence, and sorted them by decreasingly observed frequency.

At homozygous positions, a specific nucleotide is expected to be observed significantly more frequently than the other three: these are expected at low, “noise” levels. At

heterozygous positions, two specific nucleotides are expected at equivalent levels, and significantly more frequently than the remaining two, which should be observed at equivalently low levels. We used a simple binomial test (with a probability cutoff of 0.01) to ascertain statistical equivalence (or difference) between the calls observed for different nucleotides. For example, a position for which the calls were A=13, C=11, G=1, T=1 can be confidently called heterozygous A/C, since A=13 and C=11 can be observed at high probability by a process producing A and C at random, but C=11 and G=1 are an improbable outcome if the process is expected to produce C and G equally. A position with A=5, C=0, G=0, T=12 cannot be called with confidence under this model.

The binomial, *Maq*, and *maq.pl* methods confirmed 40, 35 and 53 candidate mutations in the daughter, and 35, 52, and 59 in the son. Only 28 candidates (11 in the daughter, 17 in the son) were confirmed by all three filters. We consider a mutation confirmed if the genotype calls in the resequencing data match the original genotype calls in the whole-genome data for each member of the parent-offspring trio used to initially identify the candidate mutation. This excludes all sites with a no call or discordant genotype for any member of the trio.

None of the 28 mutations confirmed by all three filters had been previously reported, whereas for each individual filter one third to one half of the confirmed positions had been reported. In addition, each of the 28 mutations confirmed by all three filters was originally identified in only one of the two children. In contrast, many of the candidates confirmed by only one of the three filters were originally identified in both children (Table S3). These sites are likely to be prone to sequencing errors across a variety of short-read technologies.

Some of the candidate *de novo* mutations that are not included in our final list are likely to be true. However, many have properties that intuitively seem unlikely to be attributes of a *de novo* mutation, such as being embedded in a region rich in MIEs or an assembly compression, being present in dbSNP, or (for putative germline SNPs) confirmation in one but not the other sibling. Nearly all of the MIEs that showed no novel alleles but a pattern of observed allele assortment that was inconsistent with Mendel's rules (Fig. S2A) could be explained by a pattern of inheritance of one or two heterozygous deletions.

To measure the quality of the combined *Maq*+maq.pl+binomial filter, we analyzed the 25 regions that we randomly selected for resequencing. The discordance rate between the original Complete Genomics data and the resequencing data was 4.2×10^{-5} per genotype for the default *Maq* settings vs. 2.3×10^{-5} for the combined Maq+maq.pl+binomial filter. Given that the error rate in the Complete Genomics data reported here is $\sim 1.1 \times 10^{-5}$, the error rate for the default *Maq* algorithm was approximately 3.1×10^{-5} vs. 1.2×10^{-5} for the three filters combined. The incorrect confirmation of a mutation could only result from an erroneous genotype call in all three filters that matched the original erroneous call in the whole-genome data. Therefore, we estimate that the false positive rate for confirming a mutation is approximately 1.2×10^{-5} per candidate mutation.

Although the false positive rate of our method was quite low, the false negative rate was substantial. Because both the binomial and the maq.pl algorithms have a tendency to report a true heterozygote as a "no call," our false negative rate estimate needed to represent sites with the exact pattern of a *de novo* mutation, with two homozygous reference calls and one heterozygote call. To match this pattern, we examined sites in the 25 random regions where one to two individuals were heterozygous and two to three individuals were homozygous reference in the Complete Genomics data. For each of these sites, we constructed one or two "trios", with two homozygous reference individuals representing the parents and one heterozygous individual representing the child with the candidate mutation. For each trio, we fail to detect a mutation if any of the three filters fails to confirm the genotype call for any member of the trio, and so our estimated false negative rate is the fraction of constructed "trios" that could not be confirmed by all three filters. The vast majority of such failures were the result of a no call in one or more of the filters. Sites that were erroneously called in the original data would cause us to overestimate the false negative rate, so to minimize this effect we restricted our analysis to sites that did not contain an inheritance error where the heterozygous allele was previously reported. In this process, we did not confirm 1832 mutations out of a total of 2768 constructed "trios", for a false negative rate estimate of 0.662 (95% C.I. approximately 0.644 – 0.680).

Calculation of mutation rates

The 28 *de novo* candidates that we report for purposes of rate estimation are 1) not in a region rich in MIEs, 2) not embedded in a likely assembly compression, 3) not known SNPs, and each one is 4) present in only one of the two children. In addition, we estimate that the false positive rate of our confirmation step is approximately 1.2×10^{-5}. Therefore this small set was likely to have zero false positives. As a final check, we evaluated these SNPs with Sequenom (San Diego, CA) MassArray genotyping and confirmed the calls for all 28 *de novo* candidates (Table S3).

To incorporate the uncertainty in the false negative rate into the confidence interval of the mutation rate, we first note that the mutation rate estimate is derived from the outcome of two random variables. Let the first random variable equal a Poisson with unknown parameter λ, to represent the number of identified *de novo* mutations. Let the second random variable equal a binomial with parameters n=2768 and unknown p, to represent the number of confirmations in the estimate of the false negative rate. The true mutation rate, μ, is equal to:

$$\mu = \frac{\lambda}{p} \bullet 2 \bullet 3{,}665{,}805{,}187 \propto \frac{\lambda}{p}$$

The log likelihood function for λ *and* p is*:*

$$\ln L(\lambda, p) = \ln\left[\frac{e^{-\lambda}\lambda^{28}}{28!}\binom{2768}{936}p^{936}(1-p)^{2768-936}\right]$$

The maximum likelihood estimate for λ/p *is* equal to 28/(0.338)=82.8. We estimate the 95% confidence interval for λ/p from a 2-parameter likelihood ratio test using a χ^2 approximation. Then the upper and lower confidence limits for λ/p are the respective minimum and maximum values of λ/p for which the following condition holds:

$$-2 * \left[\ln L(\lambda, p) - \ln L(28, 0.338)\right] < \chi^2_{\alpha=0.05, df=2}$$

A solution to this equation by numerical methods yields a 95% confidence interval for λ/p of 50.0 to 127.8. Then the 95% confidence interval for μ is 6.8×10^{-9} to 1.7×10^{-8}.

The false negative and false positive rates were estimated based on the resequencing data from the randomly selected regions. However, these were tiled at a different density than the candidate de novo mutations, which adds uncertainty to the estimate of the

intergenerational mutation rate. Specifically, the randomly selected regions were tiled at one probe per ~24 bases whereas the candidate mutations were targeted with 4 probes (2 identical probes in each orientation) corresponding to a single location.

Although our mutation rate estimate is representative of approximately two-thirds of the genome, we excluded regions of the genome that may be more mutable; inclusion of these regions might result in a higher reported mutation rate. This two-thirds fraction of the genome is not a random sample; instead it necessarily represents the proportion of the genome that could be reliably sequenced with two different technologies. If regions of the genome that are more difficult to sequence are also more mutable, the mutation rate in the remaining one third of the genome will be higher than 1.1x10-8 per site. In addition, we were only able to identify single-nucleotide mutational events. Although the comparable phylogenetic estimates also include only single-nucleotide substitutions, recent evidence suggests that a fraction of these substitutions may be the result of multi-nucleotide mutational events (S*27*). A summary of the mutation rate calculation is:

28 mutations / (1.83 billion diploid base pairs x 2 individuals)
= 7.6×10^{-9} per diploid base pair before adjusting for false negatives

$7.6 \times 10^{-9} / 2 = 3.8 \times 10^{-9}$ per haploid base pair before adjusting for false negatives

$3.8 \times 10^{-9} / (1-0.662) = 1.1 \times 10^{-8}$ per haploid base pair

Because CpG sites mutate at a rate 10 to 12 times higher than other sites (S*28*), they provide an indicator of the mutability of a genomic region. In the 1.83 billion bases we surveyed for mutations, the proportion of CpG sites is 1.8%. In the remaining 1 billion bases of the reference sequence, this proportion is 2.3%. From this factor alone, we estimate that the mutation rate is at least 4% higher in the third of the genome that we could not survey. We expect to see a modest increase in intergenerational mutation rate estimates over time, resulting from the incorporation of more mutable regions of the genome as sequencing technology improves.

None of the confirmed *de novo* mutations are on the Y chromosome and, as expected, all mutations are at positions homozygous reference in both parents, so cannot be assigned to a parental origin based on allele assortment pattern. We cannot estimate the ratio of maternal:paternal mutations. We could not confidently estimate a *de novo* indel mutation rate. It is unlikely that many, if any, of the observed 28 mutations are due to

mutations in the first few non-germline somatic cell divisions of an individual. If that were the case, our estimated germline mutation rate would be lower, and become inconsistent with phylogenetic estimates.

VI. Analysis of mutations and disease genes

Detrimental mutations

We used the KnownGenes database from UCSC supplemented with a list of 718 miRNAs as the implementation of our definition of a "gene" for purposes of considering gene candidates for inheritance models. We enumerate all missense and nonsense mutations, together with mutations in miRNAs, UTRs, non-translated transcripts, splice sites and nearby sequences, and highly conserved regions. This approach will miss many detrimental variants, such as at enhancers that are not conserved across species; this approach will falsely count many variants, such as missense variants that do not alter function.

We use a PhastCons28 score ≥500 for our operational definition of highly conserved sequence (S*29, S30*). For operational implementation, a non-coding transcript is any transcript in the UCSC known genes database that does not code for a protein. This operational implementation includes, for example, many miRNA transcripts.

Our set of "potentially detrimental" changes has two types: 1) specific positions with specific changes, which are most probably detrimental, and 2) ranges of more vaguely functional positions, in which changes might be detrimental (S*31-33*). The specific positions include: non-initiation (altering the ATG initiation codon), nonsense, missense and splice. The ranges include: near splice, noncoding, UTR. These positions may or may not be conserved across species; those that are conserved are somewhat more likely to convey a detrimental phenotype or reduction in fitness as a result of a substitution.

We compared the list of all possible detrimental changes to the list of previously observed SNPs. This list includes dbSNP130 and the 1000Genomes data set. In dbSNP there are 13,742,160 changes from reference (13,458,407 when looking only at chromosomes 1 through 22, X, and Y), 1000Genomes mentions 21,883,431 changes.

The two sets overlap by 7,889,443 changes from reference. A SNP not listed in a database likely has «0.1% allele frequency in the population.

The same position can be listed with one or more specific detrimental mutations, and at the same time it can be included in a range of positions where changes might be potentially detrimental. If a specifically listed position is not in a range, we count only its specific changes from reference as most probably detrimental. For positions in ranges but not listed specifically, we consider all three possible changes from reference. For positions both listed specifically and in a range, we do the combination: the specifically listed mutation is counted as most probably detrimental, and all other possible mutations are counted as possibly detrimental.

We tabulated counts over chromosomes 1 through 22, X, and Y. Across all chromosomes, there were 33,087,176 positions with at least one probably detrimental possible substitution (core splice: 859,727; nonsense: 3,689,392; missense: 28,444,221; non-initiation: 93,836). At these 33,087,176 positions, there are 81,455,871 possible substitutions of which 181,958 are reported in SNP databases. Across all chromosomes, there were 105,144,446 positions with possibly detrimental substitutions of which 5,744,554 are included in the tabulation of probably detrimental (UTR: 36,230,881; noncoding transcript: 18,814,972; in vicinity of a splice site: 28,746,967). At these 105,144,446 positions, there are 236,325,827 possible substitutions, excluding probably detrimental substitutions. Of these, 762,378 are reported in SNP databases.

Previous authors have speculated that inability to distinguish detrimental variation from neutral variation will limit the utility of screening methodologies that attribute function to SNPs (S*31, S34*). However, we demonstrate that for at least two disorders, current approaches to assigning SNP function (i.e., identifying missense and nonsense SNPs in coding sequences) were adequate for identifying candidates that fit our recessive models. Inheritance analysis will become more powerful with the improvement of bioinformatic characterization of the effects of variation and mutation throughout the genome on function, dysfunction, and fitness.

For Figure 3 in the main text, averages for compound heterozygote gene candidates do not include cases where both children are missing, as a substantial alteration in the

definition of “compound heterozygote” occurs in these instances: the definition becomes that of any gene in which one parent has a dominant variant. The concept of “candidate allele” or “candidate gene” is Boolean, in that a gene either is a candidate or is not. Probabilistic, or Bayesian, classification of genes as candidates would be more robust, but has not been presented here in order to allow for concise exposition. Because the choice of genes are classified as candidates or not, the precise numbers in Figure 3 are not robust. For example, if only positions fully called in all individuals were considered, there would be fewer candidates for every plotted scenario in the Figure. However, although the values for each data point might not be robust, the power of family context to substantially decrease candidates is seen with all choices for the definition of candidate gene.

SNP frequencies

Thirty or more cases of Miller syndrome have been reported in scientific publications (S*35-44*). It is difficult to accurately predict incidence because of the potential for acquisition bias and for diagnostic uncertainty. The observed incidence of a disease in a population is a result of the genetic model, penetrance as influenced by environmental and stochastic effects, and the fraction of cases detected by medical surveillance. For a simple recessive model with 100% penetrance, and with uniform population mixing, the incidence of the disease is the square of the causative allele frequency. If we assume between 500,000 and 5,000,000,000 births were surveyed to observe 30 cases, then the disease incidence is $6x10^{-9}$ to $6x10^{-6}$. The square root of the incidence, and predicted causative allele frequency under a simple recessive model, is $7.7x10^{-5}$ to $2.4x10^{-3}$.

There are 29 SNPs in dbSNP with a frequency estimate less than $1x10^{-3}$ (as of August 2009). However, there are likely to be millions, if not billions, of very rare SNPs in the world human population. Therefore it is unlikely that a SNP solely responsible for Miller syndrome, assuming it exists, is in dbSNP or any other database. As genome sequence information accumulates, estimated bounds on SNP frequencies will improve. Also, as many more genomes are sequenced, and very rare and rare SNPs are submitted to databases, the upper bound estimate for population frequency of any SNP not in any database will drop. This increasing SNP frequency information will further empower methodologies for matching genetic variants to disease models based on using the

expected frequency of a disease-causing variant, although the complexity of the bioinformatics analysis will increase, as it will no longer be sufficient to equate absence from SNP databases with a frequency < 0.1%. As of mid 2009, any SNP or variation previously seen is exceptionally unlikely to be etiologic for a rare recessive disorder.

For compound heterozygote analysis, a SNP seen in both parents is unlikely to be rare enough to be consistent with known disease incidence; these were also considered to be exceptionally unlikely candidates for etiology.

Disease models

For purposes of constraining disease candidates, compression and error-prone blocks were considered to be part of the inheritance state block that encompassed them.

A strong conclusion from published evidence is that the inheritance mode of Miller syndrome is recessive (S*38, S45, S46*). Because there is neither evidence of consanguinity in this family nor in any reported case of Miller syndrome siblings, a compound heterozygote recessive model might be a better fit to the data than a model of recessive inheritance of a single defective allele at a unique position (S*38*). It has also been postulated that all cases of Miller syndrome might be due to *de novo* mutations acting under a dominant inheritance model (S*45, S46*). Some of the variability in Miller syndrome phenotypes may be due to dietary availability of pyrimidines; dietary availability (both *in utero* and perinatally) may be affected by either maternal or fetal genetic backgrounds (S*47*). *DHODH* mutations in yeast and *Drosophila* show variable phenotypes in different background genotypes, including differences in the severity of the wing defect in *Drosophila (S48)*. Therefore we expect that *DHODH* variations in humans might also demonstrate variable expressivity (S*49, S50*).

A dominant model would require either very low penetrance or a germline mutation since both parents are unaffected. Very low penetrance seems unlikely, given the known instances of two affected siblings with Miller syndrome. In the context of exact knowledge of inheritance states, a dominant variant could not be in a nonidentical block, as in order to share the *de novo* mutation, the children must be at least haploidentical. The mutation is constrained to be in a region spanning 80.2% of the genome. The resulting pattern of alleles in the pedigree following a *de novo* mutation would show up

as an MIE, with absence of the mutation in the parents, and heterozygous presence in both children. We observed 748 such SNPs, which we tested by resequencing. Three were confirmed, but none had predicted functional significance, and we ruled out the dominant model for Miller syndrome.

Relaxing the constraint that an etiologic variant be very rare increases the number of candidates (Tables S4 & S5). For example, considering all SNPs regardless of frequency, we identified 36 candidates for compound heterozygous genes, one of which was *DNAH3*, a paralog of *DNAH5* (Fig. S5). Under simple recessive models, relaxing the population frequency threshold to 10% would permit seven candidates. The candidates resulting from relaxing the frequency thresholds are more likely to have no functional effect or be an artifact of analysis than the candidates meeting strict criteria, as these more permissive candidates have weaker, less confident, predictions for functional effect (because barring balancing selection, detrimental variation is eliminated from the population, and so the higher the frequency of an allele, the less likely it is to be detrimental). Also, candidates such as *DNAH5* and *CES1* that have candidacy based partially on inferred data are also slightly less likely candidates than they would have been had they been based on fully called data, as there remains some uncertainty in the process of inference.

The predicted amino acid changes for the three genes that fit the compound heterozygote mode are: *DHODH* (chr16:70608443 G>R, chr16:70612611 G>A), *DNAH5* (chr5:13845155 R>Q, chr5:13917742 R>stop), and *KIAA0556* (chr16:27691998 E>K, chr16:27696565 R>H). The predicted amino acid changes for the missense SNPs in the *CES1* gene are chr16:54424450 I>R and chr16:54424458 V>P.

One of the non-coding recessive candidates disrupts a putative acceptor splice site just 5' of a previously unannotated upstream exon in SP9, the mouse ortholog of which is implicated in embryonic skeletal malformation.

VII. Supplemental Figures

Figure S1. Called coverage in all 4 genomes.

Sequence reads were mapped to the NCBI reference genome. Because each individual's genome is diploid, single nucleotide polymorphisms (SNPs) and insertion-deletion polymorphisms (indels) were tabulated for both alleles of each individual genome for each position in the reference genome. The observed genotype sequence is therefore the pair of base-calls for both alleles at each chromosomal coordinate position. For some positions, only one allele, or neither, could be confidently called; these positions are denoted as "not covered." In the four individuals (mother, father, daughter, son), the percent of fully called sequence was, respectively: 85%, 91%, 91%, and 92%. In this pedigree 96% of all of the reference positions were genotyped in at least one family member; 81% were genotyped in all four (Table S2). The reference genome was divided into seven non-overlapping classes: Exome, Unique, CNVs, and four classes of repetitive elements. The UCSC KnownGenes collection operationally defined the exome. *RepeatMasker 3.2.8* output defined the repetitive element classes. The intersection of the UCSC segmental duplications and DGV structural variation collections defined the CNV regions. All remaining sequence was designated "Unique." The repetitive elements classed were: Interspersed ("Int" - complex sequence repeats ranging from 100bp to over 10kb), Simple (short stretches of low complexity sequence or tandem repeats), Young (<10% diverged from consensus) and Old (≥10% diverged from the consensus). The lower segment of each column represents the fraction of the sequence class fully called. The middle segment (lighter color) represents the fraction of the sequence class partially called (i.e, one of two alleles called). The upper segment (lightest color) is the remaining fraction: fully uncalled. The horizontal bar at the bottom of the main graph depicts the proportion of the genome attributed to each sequence class. M, mother; F, father; D, daughter; S, son.

Figure S1

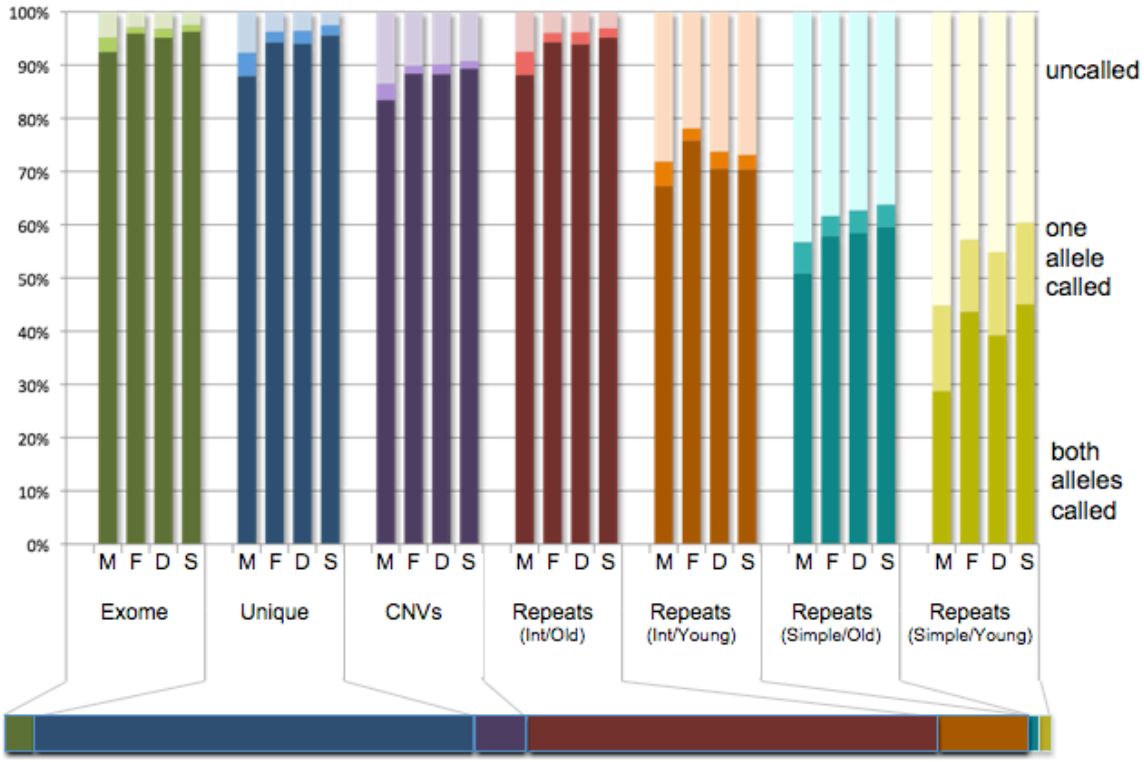

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
uncalled
one allele called
both alleles called
M F D S
Exome
Unique
CNVs
Repeats (Int/Old)
Repeats (Int/Young)
Repeats (Simple/Old)
Repeats (Simple/Young)

Figure S2A. Family genome inheritance analysis.

(a) There are four possible states of allele inheritance, depending on whether the children inherited the same alleles from both parents (red), share only a maternal allele (green) or a paternal allele (yellow), or share none (blue). Inheritance states are observed in large contiguous blocks: transitions between blocks correspond to recombinations (orange arrows). (b) Ten possible family genotype patterns for biallelic positions are consistent with one or two inheritance states. For each pattern (white boxes), the parental genotypes are shown on top (father to the left, mother to the right). The most frequent allele in the parents is denoted by “a”; in case of equal frequency, “a” denotes the most frequent allele in the children. Thus, “aa+ab” means the father is homozygous for the most frequent allele, while the mother is heterozygous. A “/” symbol is used to indicate that the order is not important. (c) One biallelic family genotype pattern, and the single monoallelic pattern, are consistent with all inheritance states, and therefore uninformative (brown). (d) Five genotype patterns are Mendelian Inheritance Errors (MIE), with a novel allele in a child. (e) Six genotype patterns would require that both alleles observed in a child derive from the same parent.

Figure S2B. Inheritance information determines uncalled genotypes.

Graphical conventions are as in Fig. S2A. (a) An uncalled allele with only one possible outcome that is consistent with Mendel's rules. (b) Two examples of uncalled alleles that can be determined based on prevalent state. The uncalled allele on the left (ab/ab, aa/an) resolves as "a" in "identical" context, and as "b" in either haploidentical context. Its presence in "nonidentical" context would represent a state consistency error. In the example on the right (ab+aa, aa/an) the SNP is consistent with all four states, and in turn the prevalent state determines the uncalled allele, as shown.

Figure S2A

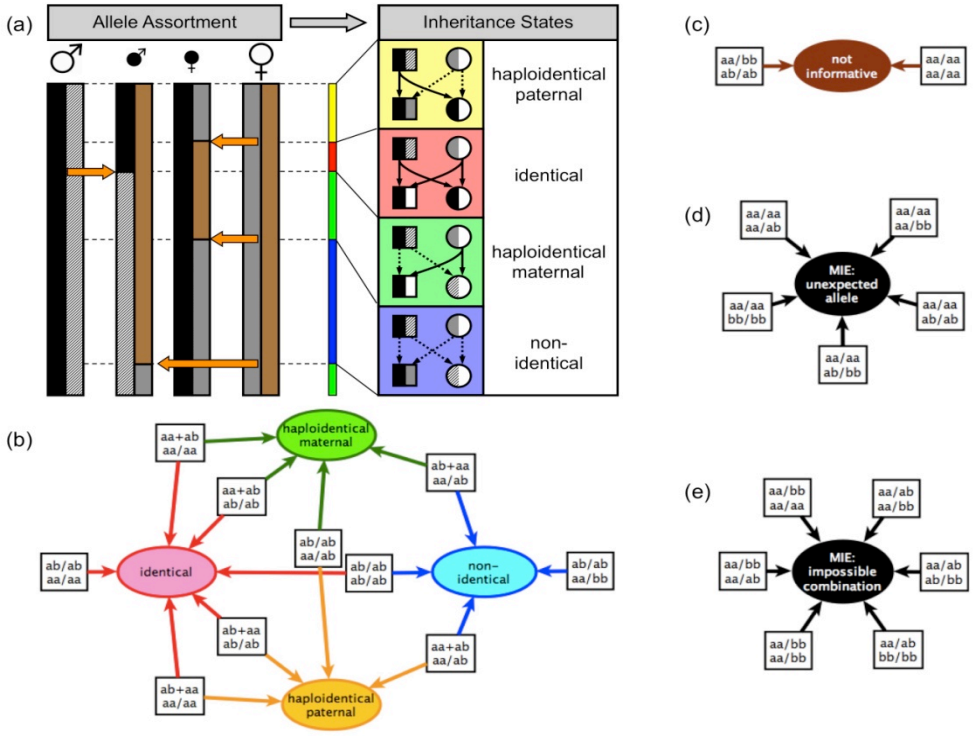


Figure S2B

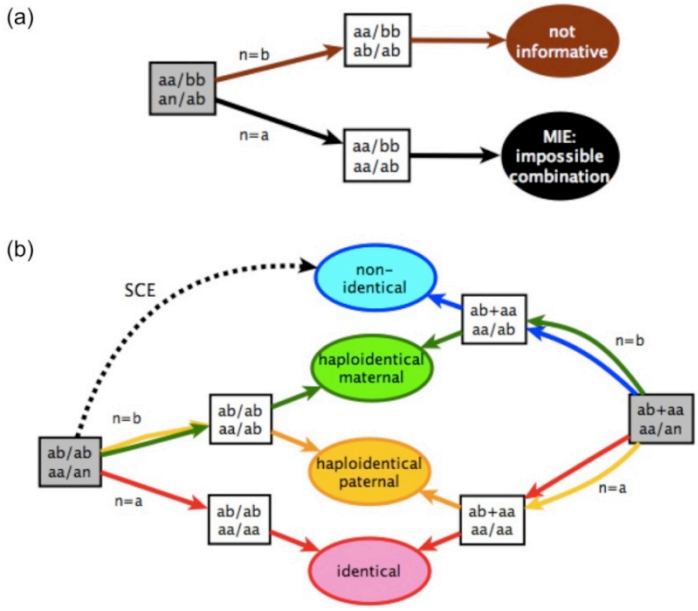


Figure S3. Inheritance blocks emerge from the observed variation.

For each chromosome, the upper graph depicts the number of informative SNPs supporting each of the four possible inheritance states: "identical" (red), "haploidentical maternal" (green), "haploidentical paternal" (yellow) and "nonidentical" (blue). SNPs consistent with two inheritance states contribute 0.5 weight to each. SNP counts are binned in non-overlapping 1 Mb windows; within each window, the four inheritance states are sorted by decreasing level of support. This directly leads to the visual identification of large blocks of consistent inheritance, bounded by recombination events. The lower graph (gray) depicts the density of state consistency errors: SNPs inconsistent with the inheritance state in which they are embedded. This graph is topped at 100 state consistency errors/Mb. State consistency errors are rare compared to informative SNPs. For improved visualization, the scale of the lower graph is 16x that of the upper graph.

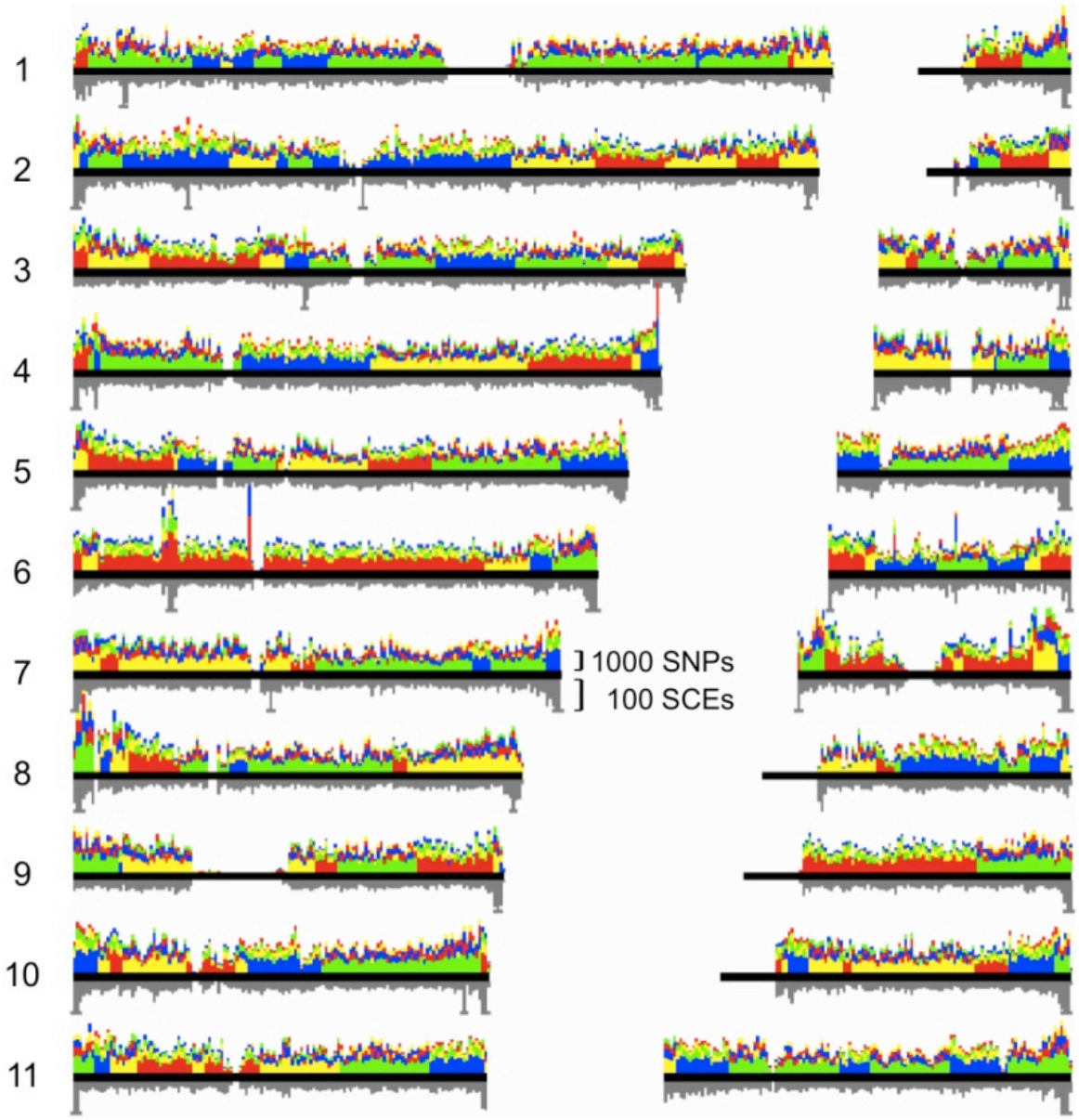

1
2
3
4
5
6
7
8
9
10
11
1000 SNPs
100 SCEs

Figure S4. PCA plot with the two parents and the HapMap phase 2 populations.

The parents cluster with the Utah HapMap sample. To build the dataset, the phased HapMap phase 2 dataset was merged with our complete genome sequence data. The PCA was generated from pairwise genetic distances for each individual in the dataset. All 2.19 million SNPs were included that met the following criteria: 1) unambiguous genotypes in both parents (no nocalls), 2) present in phased HapMap phase 2, and 3) the forward and reverse complements of the polymorphic alleles do not match (to avoid SNPs that might have been typed on the minus strand in HapMap). The tight clustering of the parents with the Utah HapMap sample is a demonstration of the accuracy of our data, in that we can merge it with another highly accurate dataset without introducing artifactual population structure to the individuals in the two datasets.

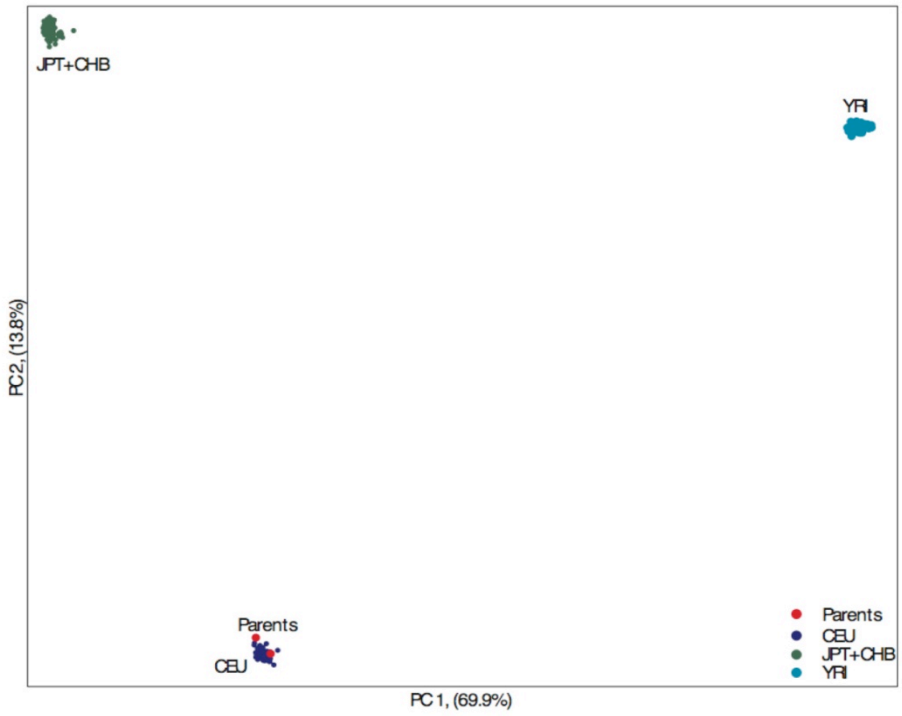
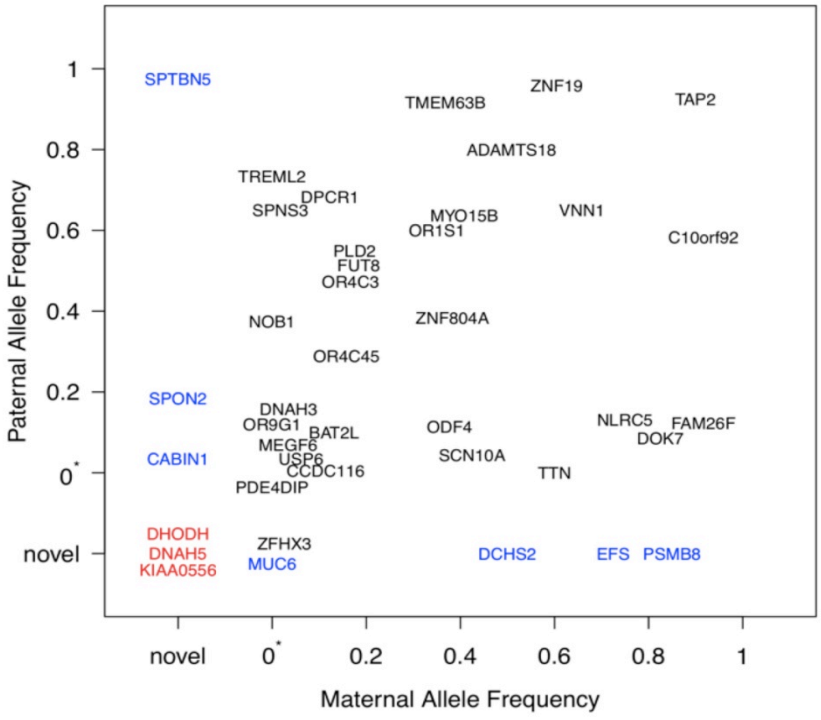


Figure S5. Compound heterozygous candidate genes.

These genes may be considered as candidates for Miller syndrome, or any recessive phenotype shared by both children. Rare diseases such as Miller syndrome would be more probably encoded by a gene near the origin than a more common recessive disease. Any allele with a frequency above 20% is unlikely to be detrimental, as it would have been purged from the population by selection unless the negative selection was balanced by a positive effect of that allele (*51*). 0*, reported in a SNP database, but without a reported frequency.



VIII. Supplemental Tables

Table S1. Insertions and Deletions

	Max Effective Length	Total count in genome	Number in "unique sequence"	Number in repetitive sequence	Number in CNVs	Number in Exome
Insertions	47	638,371	275,923	324,889	25,402	12,157
Deletions	117	618,494	237,838	347,684	22,215	10,757
Deletion-Insertion	93	247,211	91,461	136,022	14,908	4,820

Table S2. Coverage and SNP distribution.

Rare reference allele loci are positions at which all four individuals in the family analyzed in this report are homozygous for an allele different from the reference allele. Chromosome Y is considered to be fully called in females for purposes of the statistics in this table. There are 2,855,343,769 unamibiguous positions in the reference genome; for purposes of this table if a genome is fully called at all of these positions, that genome is reported as 100% called.

Coverage	Percent of Genome Fully Called (out of 2.85 Gbp)	Percent of "unique sequence" Genome fully called (out of 1.2 Gbp)	Percent of Repetitive Genome fully called (out of 1.43 Gbp)	Percent of CNV Genome fully called (out of 142 Mbp)	Percent of Exome fully called (out of 78 Mbp)
mother	85%	88%	82%	83%	92%
father	91%	94%	89%	88%	96%
daughter	91%	94%	88%	88%	95%
son	92%	95%	89%	89%	96%
Intersection of daughter & son	88%	92%	85%	85%	94%
all four (i.e., pedigree coverage)	81%	86%	78%	77%	90%
any of the four	96%	98%	94%	95%	98%
Statistics for Characteristics of Positions					
	Total count in genome	Number (fraction) in "unique sequence"	Number (fraction) in repetitive sequence	Number (fraction) in CNVs	Number (fraction) in exome
SNPs	3,665,772	1,542,175 (42.1%)	1,879,975 (51.3%)	168,610 (4.6%)	75,012 (2.0%)
Rare Reference Allele Loci	805,738	333,056 (41.3%)	421,485 (52.3%)	35,839 (4.4%)	15,358 (1.9%)
MIEs	59,243	21,723 (36.7%)	31,185 (52.6%)	4,788 (8.1%)	1,547 (2.6%)
state consistency errors	25,725	10,284 (40.0%)	12,704 (49.4%)	2,078 (8.1%)	659 (2.6%)
SNP positions with at least one no call	1,183,152	311,092 (34.7%)	516,354 (57.6%)	55,782 (6.2%)	13,616 (1.5%)
SNP positions with a no call in mother	793,065	209,074 (36.6%)	322,711 (56.5%)	31,999 (5.6%)	7,477 (1.3%)
SNP positions with a no call in father	497,766	116,649 (33.1%)	206,814 (58.7%)	23,974 (6.8%)	5,061 (1.4%)
SNP positions with a no call in daughter	430,223	95,393 (31.8%)	175,406 (58.5%)	23,574 (7.9%)	5,336 (1.8%)
SNP positions with a no call in son	434,473	90,283 (29.7%)	187,443 (61.6%)	21,655 (7.1%)	4,730 (1.6%)

Table S3. Tabulation of potential de novo *mutations (attached file).*

The columns are:

Chrom: chromosome

Site: coordinate

State: the inheritance state of the position as defined in the main text

MIEtype: unexp(ected allele) or wrong (combination)

Ref: Base call in reference sequence (hg18)

Genotype: 'ab' genotype ('a' = reference); m=mother, f=father, d=daughter, s=son)

maq: PASS if all four genotypes were confirmed by maq

maq.pl: PASS if all four genotypes were confirmed by maq.pl SNPfilter

Beyond maq's calls, this script qualifies SNPs based on the quality of the context: "Rule out SNPs that are covered by few reads (specified by -d), by too many reads (specified by -D), near (specified by -w) to a potential indel, falling in a possible repetitive region (characterized by -Q), or having low-quality neighboring bases (specified by -n)." Default values of parameters: -d = 3, -D = 256, -w = 3, -Q = 40, -n = 20; maq.pl SNPfilter doesn't disqualify any of the "unexpected allele" MIEs. It does reject several "wrong combination" MIEs.

binom: PASS if all four genotypes were confirmed by the binomial method

Massarray: PASS if all four genotypes were confirmed by Sequenom MassArray analysis; 'ab'-genotype if a different genotype was observed. (1) Genotyping failed in the paternal genome and could in principal be 'ab', but all three filters on array results suggest it is indeed 'aa'. (2) Genotyping failed in the paternal genome, but this is inconsequential since in the haploidentical paternal state the 'b' allele in the son only should have come from the mother. (3) This genotype is a state consistency error. All other new genotypes resulting from Sequenom analysis fit the local inheritance patterns. (4) Genotyping failed in the mother, while in a haploidentical maternal region the 'b'-allele in both kids would have been inherited from the mother. Considering that the potential *de novo* mutation represents a known allele and that the capture array coverage was very weak for the maternal genome, we predict this to be a false call rather than a *de novo* mutation.

Assessment: An unexpected allele is labeled a "DE_NOVO" mutation if all three capture array analysis methods or one such method and the MassArray results confirmed all four genotypes. DE_NOVO(dtr) and DE_NOVO(son) for new alleles present in the daughter or son. FALSE_CALL: the MIE was resolved by resequencing. HEMIZ_GAP: the apparent MIE resulted from the inheritance of a hemizygous deletion. Uncertainty is indicated by a question mark. Calls are considered certainly false when the unexpected allele is a known allele or when MassArray genotyping revealed a Mendelian inheritance pattern. Hemizygous gaps are considered certain explanations for a wrong combination MIE when the site is within a predicted hemizygous gap of a type that can explain the observed genotype.

Change: transition or transversion, with specific bases indicated

Context: trinucleotide context of the mutation site. Lowercase if repeat.

KnownSNP: indicates if the unexpected allele in the child(ren) has been reported before in dbSNP, the 1000 Genome Project, the Venter, Watson, the first Yoruban or Asian genomes, or the CNV database. The dbSNP identifier is given when the SNP appeared in dbSNP build 130.

maq-quals: quality scores assigned by maq (for mother, father, daughter, son)

Coverage: resequencing coverage (for mother, father, daughter, son)

Mapability: based on the wgEncodeDukeUniqueness35bp track

Deletion?: Indicates the type of inheritance of a hemizygous deletion could, in principal, explain the observed “wrong combination MIE” given the known inheritance state. F = father, M = mother, S = son, D = daughter. For example, M->D&S is a deletion inherited by both children (possible in either the identical or haploidentical maternal state).

Predicted gap extent: Maximum extent of hemizygous gaps (inherited in the same fashion as indicated in the previous column) as predicted by the gap prediction analysis described above.

Table S4. Reduction in false positive candidate SNPs using inheritance analysis.

Genomic inheritance analysis excludes many false positive candidates and enhances the signal to noise of family based complete genome sequencing. This table shows benefits in together removing both correct and falsely called sequence variants from candidacy. Correctly called variants are removed if they occur in a region with an inheritance state inconsistent with the inheritance mode (recessive, compound heterozygous, or dominant). Falsely called variants are removed regardless of inheritance state. Inheritance analysis permits: determination of regions of the chromosome that were inherited from each parent in the affected siblings, inference of missing data, and identification of positions that are errors because their inheritance pattern is inconsistent with the inheritance state of the region or are MIEs. These factors permit a reduction in the number of false positive candidates for disease models. For example, if this information were ignored, there would be nine variation candidates (in seven genes) for very rare recessive missense mutations. Using this information, there are only two candidates (linked in the same gene, *CES1*). False positives inherent to true genotypes, such as a very rare SNP coincidentally with the correct inheritance mode and that is not detrimental or is detrimental but affecting another phenotype, cannot be eliminated with inheritance analysis: better functional prediction or the additional of individuals to resolve the inheritance mode would be necessary to remove these false positives. The recessive categorization tabulates simple recessive patterns; these do not include compound heterozygote patterns. Dominant inheritance patterns tabulated in this table match sets of genotypes for the pedigree in which one parent is homozygous for a common allele, and one parent and both children are heterozygous for a candidate detrimental allele. Such a dominant model, to be realized in this family as causing a phenotype shared by the two children but not by the parents, would require the phenotype to not be penetrant in the heterozygous parent. Such low penetrance is unlikely for the observable phenotypes in this family; therefore the counts for the dominant model represent a control analysis for estimating the false positive rate of dominant candidates for a disease if a family of four, with a phenotype consistent with dominant inheritance, were to be analyzed. IS; exact knowledge of inheritance state.

	SNP frequency	all (SNPs)	conserved (SNPs)	UTR (SNPs)	Noncoding (SNPs)	near splice (SNPs)	missense (SNPs)	splice (SNPs)	nonsense (SNPs)	non-initiation (SNPs)	compound heterozygote (genes)
With IS: candidates for compound heterozygote patterns	all	256759	2940	2719	1471	2287	967	17	10	5	66
	very rare	25651	421	305	213	203	145	0	1	1	3
	known	231108	2519	2414	1258	2084	822	17	9	4	NA
With IS: dominant candidates	all	589108	6781	5830	3196	5023	1990	38	23	10	NA
	very rare	62019	1104	692	434	509	322	7	6	2	NA
	known	527089	5677	5138	2762	4514	1668	31	17	8	NA
With IS: recessive candidates	all	60489	543	526	297	520	149	0	1	1	NA
	very rare	528	2	1	5	1	2	0	0	0	NA
	known	59961	541	525	292	519	147	0	1	1	NA
No IS: dominant or compound heterozygote pattern candidates	all	673348	7743	6422	3651	5587	2188	40	26	10	66
	very rare	73196	1333	764	542	582	369	8	7	2	3
	known	600152	6410	5658	3109	5005	1819	32	19	8	NA
No IS: recessive candidates	all	119264	1134	907	579	936	277	3	1	2	NA
	very rare	1634	53	8	25	9	9	0	0	0	NA
	known	117630	1081	899	554	927	268	3	1	2	NA

Table S5. Signal-to-noise enhancement provided by family sequencing.

Sequencing and analysis of family members excludes many false positives and enhances the signal to noise of complete genome sequencing. A large increase in the number of excluded false positives occurs at the transition from three sequenced to four sequenced individuals. This increase is made possible not just from inference of missing data but also because inheritance analysis permits determination of inheritance states, and consequently exclusion of large portions of the genome from consideration. For most functional categories, the number of false positive candidates drops several fold when three individuals are sequenced rather than one, and by about an additional order of magnitude as a fourth individual is added to the set sequenced. (M, mother; F, father; D, daughter; S, son). “Very rare” is operationally defined as a SNP not in a known database (e.g., dbSNP or the 1000 Genome collection) and is equivalent to a frequency < 0.1%; “common” is operationally defined as a SNP in a known database. For the top row in each panel, when all four individuals are sequenced (i.e., none are excluded), all candidate SNPs are constrained to be in "identical" blocks. Inheritance state information is greatly impoverished in combinations of two or three individuals with only two children included, and is completely absent in other combinations of individuals, so no constraint from state blocks is applied to limit candidates when less than all four individuals are considered. An initiation mutation alters the ATG of a mRNA. Panel A: rare SNP candidates for a recessive inheritance model; Panel B: common SNP candidates for a recessive inheritance model; Panel C: rare SNP candidates that might form half of a compound heterozygote variation pair, and the resulting number of candidate genes; Panel D: common SNP candidates that might form half of a compound heterozygote variation pair, and the resulting number of candidate genes.

Table S5 (panel A).

family members excluded	All (SNPs)	Very rare (SNPs)	Conserved (SNPs)	UTR (SNPs)	noncoding: transcripts that do not code for proteins (SNPs)	near splice: sites in the intron that might affect splicing (SNPs)	missense (SNPs)	Splice: key donor and acceptor bases (SNPs)	nonsense (SNPs)	initiation (SNPs)
none	60489	528	2	1	5	1	2	0	0	0
missing: S	208420	1668	33	17	16	12	9	0	0	0
missing: D	230919	1918	57	14	25	8	7	0	0	0
missing: F	235734	788	15	7	6	7	4	0	0	0
missing: M	139812	617	15	2	6	5	4	0	0	0
missing: S,D	789832	27186	568	203	274	191	139	2	4	0
missing: S,F	437888	2199	40	22	21	16	9	0	0	0
missing: D,F	456545	2505	71	23	28	11	8	0	0	0
missing: S,M	356730	2237	46	24	23	13	11	0	0	0
missing: D,M	376177	2451	63	19	32	12	9	0	0	0
missing: M,F	816218	1729	51	10	26	10	9	0	0	0
missing: S,D,M	1959524	119813	1925	1309	825	1029	571	6	10	3
missing: S,D,F	1861024	111381	1940	1284	900	960	601	13	17	1
missing: S, M, F	1159074	5292	94	39	39	30	20	0	0	0
missing: D, M, F	1181721	5630	125	47	49	37	17	1	0	0
MEANS:										
missing one	203721	1248	30	10	13	8	6	0	0	0
missing two	538898	6385	140	50	67	42	31	0	1	0
missing three	1540336	60529	1021	670	453	514	302	5	7	1

Table S5 (panel B).

family members excluded	Known (SNPs)	Conserved (SNPs)	UTR (SNPs)	noncoding: transcripts that do not code for proteins (SNPs)	near splice: sites in the intron that might affect splicing (SNPs)	missense (SNPs)	Splice: key donor and acceptor bases (SNPs)	nonsense (SNPs)	Initiation (SNPs)
none	59961	541	525	292	519	147	0	1	1
S	206398	2053	2007	1036	1735	543	10	4	2
D	228606	2283	2014	1133	1882	574	8	4	2
F	234805	2383	1997	1035	1865	575	13	2	3
M	139063	1441	1208	748	1164	366	5	1	5
S,D	756961	8050	7489	4245	6701	2290	34	28	5
S,F	435198	4564	4064	2075	3555	1146	24	10	3
D,F	453521	4743	4060	2233	3683	1165	23	6	4
S,M	354011	3692	3489	1883	3020	975	19	4	6
D,M	373210	3918	3508	2009	3274	1052	17	5	5
M,F	814285	9256	8392	4512	7365	2755	42	12	11
S,D,M	1824652	20044	18460	9950	16219	5773	79	55	19
S,D,F	1735899	19671	18321	9849	15649	5702	100	53	16
S, M, F	1153163	12714	11730	6135	10116	3604	63	20	12
D, M, F	1173527	12881	11852	6331	10450	3694	62	18	12
MEANS:									
missing one	202218	2040	1807	988	1662	515	9	3	3
missing two	531198	5704	5167	2826	4600	1564	27	11	6
missing three	1471810	16328	15091	8066	13109	4693	76	37	15

Table S5 (panel C).

family members excluded	All	Very rare (SNPs)	Conserved (SNPs)	UTR (SNPs)	noncoding: transcripts that do not code for proteins (SNPs)	near splice: sites in the intron that might affect splicing (SNPs)	missense (SNPs)	Splice: key donor and acceptor bases (SNPs)	nonsense (SNPs)	initiation (SNPs)	compound heterozygote (genes)
none	256759	25651	421	305	213	203	145	0	1	1	3
S	1081419	113971	1969	1207	843	947	565	15	9	2	31
D	1096928	115109	2013	1258	783	999	515	11	7	2	27
F	488794	43961	718	547	289	380	236	6	2	2	23
M	488810	43966	718	547	289	380	236	6	2	2	12
S,D	1963559	214224	3587	2432	1534	1865	700	18	11	4	37
S,F	1118583	131671	2290	1451	962	1130	708	21	16	4	58
D,F	1166766	145595	2661	1620	983	1282	700	14	16	2	50
S,M	1157890	133633	2329	1481	967	1156	714	21	16	4	49
D,M	1189541	145515	2667	1630	946	1285	693	14	15	2	41
M,F	488796	43963	718	547	289	380	236	6	2	2	28
S,D,M	4154347	271527	4631	3149	1915	2480	1005	43	26	6	NA
S,D,F	3959401	262235	4526	3033	1912	2408	1008	36	26	6	NA
S, M, F	1069645	123053	2070	1384	885	1067	666	20	15	4	72
D, M, F	1118848	136668	2438	1570	864	1224	664	12	15	2	64
MEANS:											
missing one	788988	79252	1355	890	551	677	388	10	5	2	23
missing two	1180856	135767	2375	1527	947	1183	625	16	13	3	44
missing three	2575560	198371	3416	2284	1394	1795	836	28	21	5	68

Table S5 (panel D).

family members excluded	Known (SNPs)	Conserved (SNPs)	UTR (SNPs)	noncoding: transcripts that do not code for proteins (SNPs)	near splice: sites in the intron that might affect splicing (SNPs)	missense (SNPs)	Splice: key donor and acceptor bases (SNPs)	nonsense (SNPs)	Initiation: A mutation that alters the ATG of a mRNA (SNPs)	compound heterozygote (genes)
none	231108	2519	2414	1258	2084	822	17	9	4	66
S	961493	10534	9482	5071	8176	3019	43	21	12	256
D	974559	10722	9537	5115	8255	2935	46	31	12	259
F	443438	4982	4480	2326	3920	1435	24	14	7	217
M	443448	4983	4480	2326	3920	1435	24	14	7	169
S,D	1727735	19375	17777	9431	15107	4138	81	36	17	374
S,F	982879	10791	9676	5138	8377	3045	45	20	13	478
D,F	1014438	11190	9878	5382	8586	2991	49	32	12	461
S,M	1018278	11049	9947	5304	8640	3073	48	20	13	451
D,M	1036735	11329	10030	5429	8742	2997	53	31	12	472
M,F	443438	4982	4480	2326	3920	1435	24	14	7	312
S,D,M	3855842	43409	40121	21466	34544	6052	225	50	40	NA
S,D,F	3672784	41970	38971	20953	33382	5977	216	48	38	NA
S, M, F	942563	10311	9305	4859	8049	2931	41	18	12	655
D, M, F	975451	10662	9534	5053	8277	2901	48	29	12	660
MEANS:										
missing one	705735	7805	6995	3710	6068	2206	34	20	10	215
missing two	1037251	11453	10298	5502	8895	2947	50	26	12	425
missing three	2361660	26588	24483	13083	21063	4465	133	36	26	658

IX. Supplemental References

S1. B. Borrell, *Nature* **463**, 858 (Feb 18, 2010).
S2. P. D. Lewis, J. S. Harvey, E. M. Waters, D. O. Skibinski, J. M. Parry, *Mutagenesis* **16**, 503 (Nov, 2001).
S3. R. M. Fineman, *J Pediatr* **98**, 87 (Jan, 1981).
S4. M. Miller, R. Fineman, D. W. Smith, *J Pediatr* **95**, 970 (Dec, 1979).
S5. S. B. Ng *et al.*, *Nat Genet* **42**, 30 (Jan, 2010).
S6. K. Tang, K. R. Thornton, M. Stoneking, *PLoS Biol* **5**, e171 (Jul, 2007).
S7. J. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, *Nucleic Acids Res* **37**, D793 (Jan, 2009).
S8. R. Drmanac *et al.*, *Science*, 1181498 (November 5, 2009).
S9. S. B. Ng *et al.*, *Nature* **461**, 272 (Sep 10, 2009).
S10. S. T. Sherry, M. Ward, K. Sirotkin, *Genome Res* **9**, 677 (Aug, 1999).
S11. R. Durbin, D. Altshuler, L. D. Brooks, A. Felsenfeld, J. McEwen. (www.1000genomes.org, 2009).
S12. J. Wang *et al.*, *Nature* **456**, 60 (Nov 6, 2008).
S13. S. Levy *et al.*, *PLoS Biol* **5**, e254 (Sep 4, 2007).
S14. D. R. Bentley *et al.*, *Nature* **456**, 53 (Nov 6, 2008).
S15. D. A. Wheeler *et al.*, *Nature* **452**, 872 (Apr 17, 2008).
S16. A. J. Iafrate *et al.*, *Nat Genet* **36**, 949 (Sep, 2004).
S17. R. C. Gentleman *et al.*, *Genome Biol* **5**, R80 (2004).
S18. M. Wirtenberger, K. Hemminki, B. Burwinkel, *Am J Hum Genet* **78**, 520 (Mar, 2006).
S19. G. Coop, X. Wen, C. Ober, J. K. Pritchard, M. Przeworski, *Science* **319**, 1395 (Mar 7, 2008).
S20. N. Arnheim, P. Calabrese, I. Tiemann-Boege, *Annu Rev Genet* **41**, 369 (2007).
S21. J. C. Ting, E. D. Roberson, D. G. Currier, J. Pevsner, *BMC Med Genet* **10**, 93 (2009).
S22. I. Tiemann-Boege, P. Calabrese, D. M. Cochran, R. Sokol, N. Arnheim, *PLoS Genet* **2**, e70 (May, 2006).
S23. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am J Hum Genet* **63**, 861 (Sep, 1998).
S24. M. W. Nachman, S. L. Crowell, *Genetics* **156**, 297 (Sep, 2000).
S25. L. Wang, Z. Wang, W. Yang, *BMC Bioinformatics* **10**, 216 (2009).
S26. H. Li, J. Ruan, R. Durbin, *Genome Res* **18**, 1851 (Nov, 2008).
S27. J. C. Walser, L. Ponger, A. V. Furano, *Genome Res* **18**, 1403 (Sep, 2008).
S28. Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (Sep 1, 2005).
S29. A. Siepel *et al.*, *Genome Res* **15**, 1034 (Aug, 2005).
S30. G. Glusman, *Proceedings of VII IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2009)*, (May, 2009).
S31. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, *Am J Hum Genet* **80**, 727 (Apr, 2007).

S32. A. R. Boyko *et al.*, *PLoS Genet* **4**, e1000083 (May, 2008).
S33. K. E. Lohmueller *et al.*, *Nature* **451**, 994 (Feb 21, 2008).
S34. J. N. Hirschhorn, D. Altshuler, *J Clin Endocrinol Metab* **87**, 4438 (Oct, 2002).
S35. G. W. Stevenson, S. C. Hall, B. S. Bauer, F. A. Vicari, F. L. Seleny, *Can J Anaesth* **38**, 1046 (Nov, 1991).
S36. D. Donnai, H. E. Hughes, R. M. Winter, *J Med Genet* **24**, 422 (Jul, 1987).
S37. O. Cacchione, A. Guadagni, B. Persichetti, G. Logoluso, D. Barbuti, *Radiol Med* **84**, 650 (Nov, 1992).
S38. L. Neumann, J. Pelz, J. Kunze, *Am J Med Genet* **64**, 556 (Sep 6, 1996).
S39. D. Barbuti, C. Orazi, A. Reale, C. Paradisi, *Eur J Pediatr* **148**, 445 (Feb, 1989).
S40. K. H. Chrzanowska, J. P. Fryns, M. Krajewska-Walasek, L. Wisniewski, H. Van den Berghe, *Clin Genet* **35**, 157 (Feb, 1989).
S41. A. L. Ogilvy-Stuart, A. C. Parsons, *J Med Genet* **28**, 695 (Oct, 1991).
S42. M. Richards, *Anaesthesia* **42**, 871 (Aug, 1987).
S43. R. Rochels, *Klin Padiatr* **193**, 238 (May, 1981).
S44. J. Vigneron, M. Stricker, P. Vert, J. M. Rousselot, M. Levy, *J Med Genet* **28**, 636 (Sep, 1991).
S45. M. Robinow, G. F. Johnson, J. Apesos, *Am J Med Genet* **25**, 293 (Oct, 1986).
S46. P. Meinecke, H. R. Wiedemann, *Am J Med Genet* **27**, 953 (Aug, 1987).
S47. C. T. Van Buren, F. Rudolph, *Nutrition* **13**, 470 (May, 1997).
S48. J. M. Rawls, J. W. Fristrom, *Nature* **255**, 738 (Jun 26, 1975).
S49. X. Ma, A. M. Tarone, W. Li, *PLoS One* **3**, e1922 (2008).
S50. S. L. Ooi *et al.*, *Trends Genet* **22**, 56 (Jan, 2006).
S51. S. R. Sunyaev, W. C. Lathe, 3rd, V. E. Ramensky, P. Bork, *Trends Genet* **16**, 335 (Aug, 2000).