Supplementary Information for "Variation in genome-wide mutation rates within and between human families"

Donald F. Conrad[1], Jonathon E. M. Keebler[2,3], Mark DePristo[4], Sarah Lindsay[1], Yujun Zhang[1], Ferran Cassals[2], Youssef Idaghdour[2], Christopher Hartl[4], Carlos Torroja[1], Kiran Garimella[4], Martine Zilversmit[2], Reed A. Cartwright[3,5], Guy Rouleau[6], Mark Daly[4], Eric Stone[3,5], Matthew E. Hurles[1], Philip Awadalla[2] on behalf of the 1000 Genomes Project

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[2] Ste Justine Hospital Research Centre, Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal H3T 1C5, Canada

[3] Bioinformatics Research, North Carolina State University, Raleigh, NC 27695-7566, USA

[4] Program in Medical and Population Genetics, The Broad Institute of Harvard, and MIT, Five Cambridge Center, Cambridge, Massachusetts 02142, USA

[5] Department of Genetics, North Carolina State University, PO Box 7614, Raleigh 27659, USA

[6] Ste Justine Hospital Research Centre, Department of Medicine, Faculty of Medicine, University of Montreal, Montreal H3T 1C5, Canada

April 14, 2011

Display Items

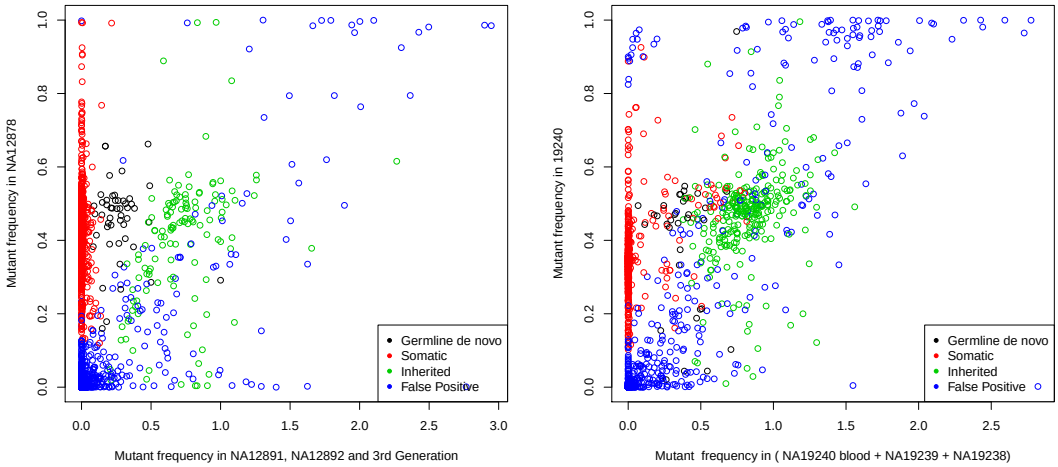

Figure S1: **Visualization of validation results.** Here we display a two-dimensional summary of the CEU (left panel) and YRI (right) read data for each of the validation sites that we were able to classify with confidence. On the x-axis, sites are positioned by the cumulative mutant read frequency in the 3rd generation samples and parents (CEU) or the blood DNA from NA19240 and cell line DNA from the trio parents (YRI). On the y-axis sites are positioned by the mutant read frequency in the cell line of the trio offspring. Mutant read frequencies are weighted summaries of the results from both validation experiments. Sites are colored according to validation status, see inset legend for details.

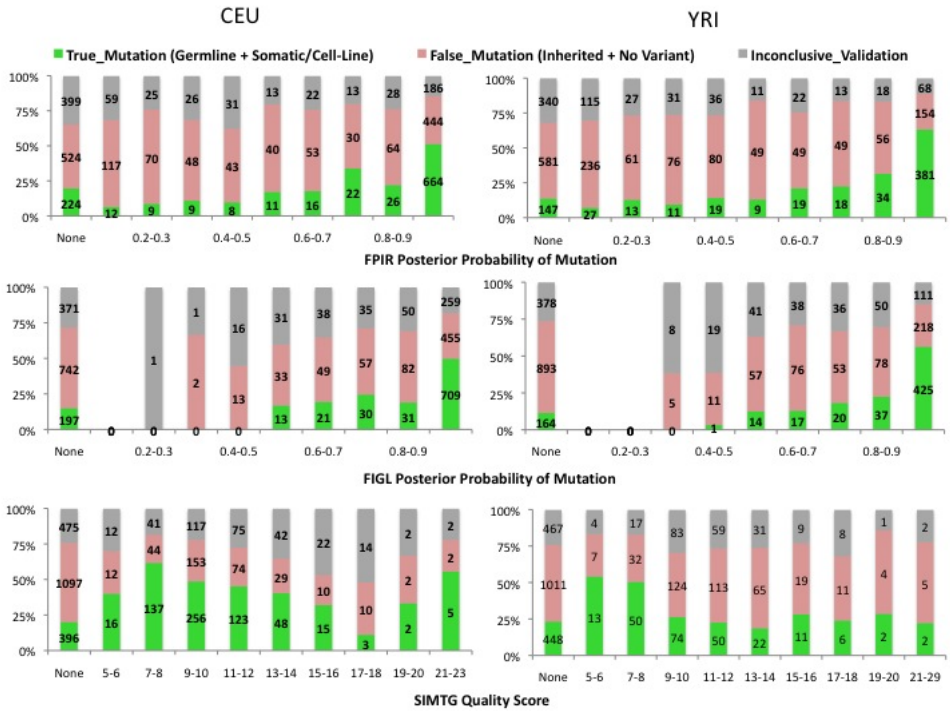

Figure S2: **Validation results stratified by** *de novo* **calling algorithm.** For each method, Family-aware Probabilistic Illumina Read-based method (FPIR), Family-aware Illumina Genotype Likelihood-based method (FIGL), and Sample-Independent Multiple Technology Genotype-based method (SIMTG), the results of the combined validation experiments across all candidate sites are summarized as Inconclusive, True mutation, or False Mutation, and separated by the prediction metric reported by each method. The number of candidate sites falling within each category are shown on the bars. Sites reported in the 'None' columns were not included in the candidate site list contributed by the particular method.

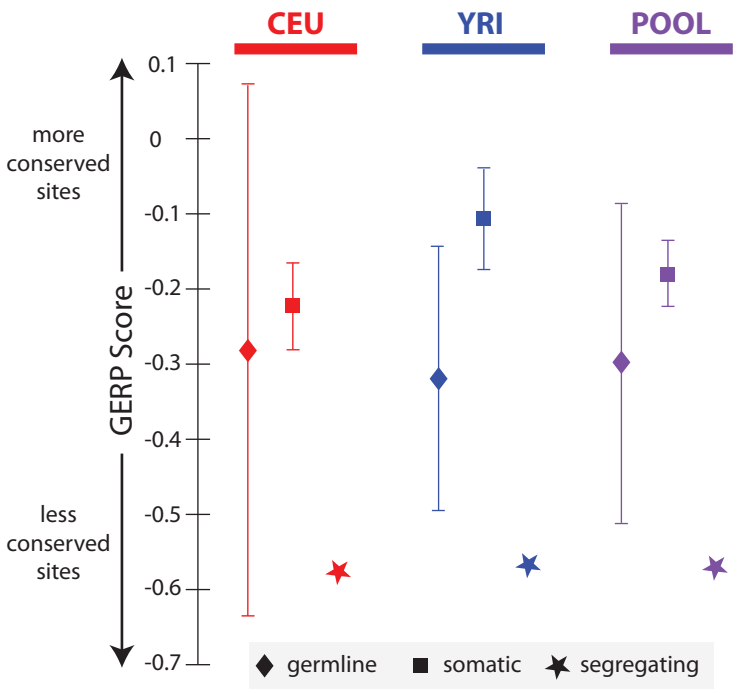

Figure S3: **Analysis of selective constraint on sites of mutation and variation.** For each trio, we calculated GERP scores on sites of germline DNM, non-germline DNM, and sites of segregating variation seen in at least one of the three trio members. The mean plus/minus 1 standard error is plotted for each class. See methods in main text for details.

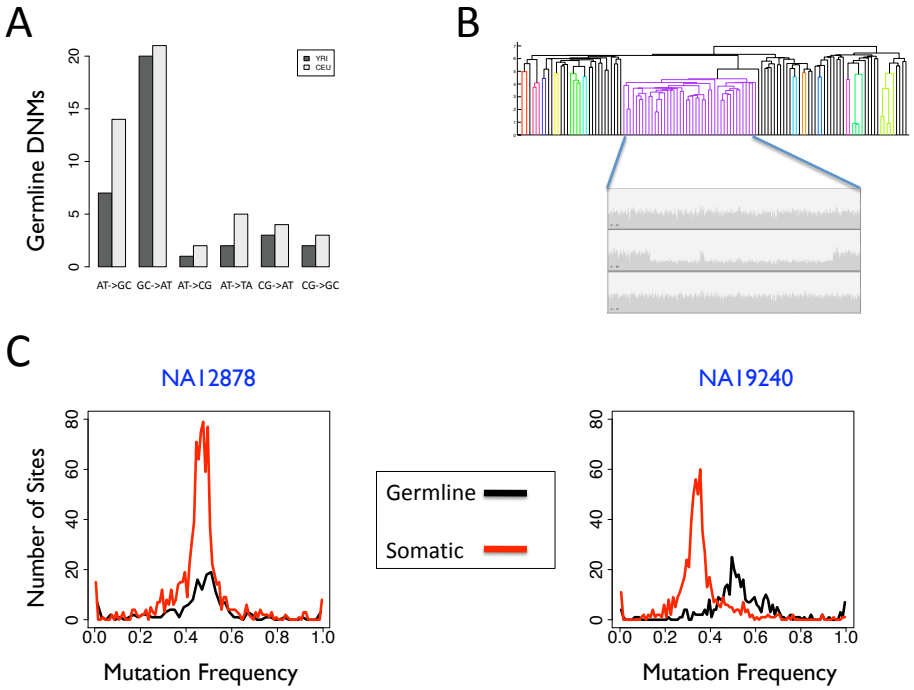

Figure S4: **Comparing germline and somatic** *de novo* **mutations.** (A) The relative proportion of mutations from A:T positions and G:C positions for germline and somatic DNMs merged across trios. Complementary mutations were combined. (B) Clustering of apparent DNMs due to somatic deletions. The top panel shows a dendrogram created by hierarchical clustering of all apparent CEU DNMs on chromosome 19 on the basis of their physical separation. By examining read depth data on the same individuals, we see that physically clustered apparent DNMs are caused by a somatic or cell-line deletion in one of the two trio parents, as illustrated here for an event at 41.8-42.0 Mb in NA12892. (C) Mutant read frequency is highly sensitive to cell line mosaicism. For each trio offspring, the distribution of mutant read frequencies in the Illumina validation data is plotted for germline (DNM and inherited) heterozygous variants (black) and somatic DNMs (red).

Sample	Experiment I Source	Experiment II Source	Relationship
CEU Family			
NA12891	C	C	Father
NA12892	C	C	Mother
NA12878	C	C	Daughter
NA12877	N	C	Husband
NA12879	C*	C	Granddaughter
NA12880	C*	C	Grandaughter
NA12881	C*	C	Grandaughter
NA12882	C*	C	Grandson
NA12883	C*	C	Grandson
NA12884	C*	C	Grandson
NA12885	C*	C	Granddaughter
NA12886	C*	C	Grandson
NA12887	C*	C	Granddaughter
NA12888	C*	C	Grandson
NA12893	C*	C	Grandson
YRI Family			
NA19238	C,B*	C	Mother
NA19239	C,B*	C	Father
NA19240	C,B	C,B	Daughter

Table S2: **Source of DNA samples used in validation experiments.** B = Whole genome amplified, blood derived DNA, C = cell line DNA, N = not sequenced, B* = PCR products from WGA, blood derived DNA for NA19238 and NA19239 were pooled prior to sequencing in experiment I, C* = PCR products from cell line DNA for all 11 CEU grandchildren were pooled prior to sequencing in experiment I.

Sample	Number Reads (Mill)	% mapped	% properly paired	median read depth	% sites > 0 reads	% sites > 19 reads
Experiment I						
NA12878	31	89.5	82	2945	89	80.5
NA12891	39	89	80	2760	85	73
NA12892	52	88	75	6394	88	80.5
CEU grandchildren	37.5	89	81.5	3981	91.5	87.5
NA19240	34.5	92.3	80	3121	89.4	86.5
NA19238	38.9	85.4	72	1523	74	65.6
NA19239	39.8	90	76.5	2743	89.4	87.3
NA19240 blood	42.5	86	73	2961	89.7	87.5
YRI parent blood	41	83.5	71	3374	90.4	88.9
Experiment II						
NA12878				26	91.9	56.7
NA12891				29	91.7	59.5
NA12892				32	93.9	61.4
CEU grandchildren				21	91.8	51.3
NA19240				14	87.2	42.4
NA19238				16	88.5	45.0
NA19239				17	89.2	46.7
NA19240 blood				17	87.7	46.2
YRI parent blood				NA	NA	NA

Table S3: **Validation Experiment Sequencing Summary Statistics.** Notes for Experiment II: values listed for CEU grandchildren are averages across samples. YRI parental blood DNA was not sequenced in Experiment II.

Supplementary Note: Introduction

This Supplementary Note contains all of the methods used to generate the results in the main text, as well as additional results not described in the main text. This document is organized in a manner that roughly reflects the chronology of the project (Figure 1, main text), with downstream analyses of the validated *de novo* mutations at the end.

Contents

1 Discovery of *de novo* point mutation candidates

Three different algorithms, Sample-Independent Multiple Technology Genotype based method (SIMTG), Family-aware Illumina Genotype Likelihood-based method (FIGL), and Family-aware Probabilistic Illumina Read-based method (FPIR), were developed for DNM discovery. The first approach considers data on each sample independently but jointly considers data from all three sequencing platforms. The subsequent genotypes of family members are compared to identify apparent new alleles present in the offspring and the three genotype confidence values summarized to rank candidate DNMs. The last two are probabilistic approaches that use the Illumina data and jointly analyse data from all three family members simultaneously, although they differ in their underlying statistical methodology. These approaches both make use of a population genetic prior, and model the data for the entire trio jointly. One important difference is in the way each method models sequencing error. The FPIR model assumes a constant sequencing error rate across all reads, which is estimated from the data; on the other hand FIGL uses the base error probabilities attached to each read in the original BAM files and uses these uncertainties when producing likelihooods. The FPIR process filters reads on MapQ and BaseQ prior to fitting the model to data, whereas FIGL uses all reads when producing likelihoods.

1.1 SIMTG

The Sample-Independent Multiple Technology Genotype (SIMTG) protocol was applied independently to read data from each sample of each trio. At each base of the genome, all reads overlapping that base were used to calculate the likelihood of each of the ten possible diploid genotypes:

$$P[XY] = \prod_b P[b \in X, Y = \begin{cases} 1 - 10^{-Q_b/10} & b \in \{X,Y\}, X = Y \\ 1 - \frac{1}{2}10^{-Q_b/10} + \frac{1}{6}10^{-Q_b/10} & b \in \{X,Y\}, X \neq Y \\ \frac{1}{3}10^{-Q_b/10} & b \notin \{X,Y\} \end{cases}$$

(where X, Y are alleles, b a base of the base pileup, and Q_b its quality score) The genotype with maximum likelihood became the called genotype, and was assigned a LOD score (the log-likelihood ratio of the best genotype to the next-best genotype). Those sites identified as heterozygous in the offspring at a confidence of LOD ≥ 5, and homozygous in the parents at a confidence of LOD ≥ 5, comprised the SIMTG DNM call set. For samples sequenced on multiple platforms, two call sets were generated: one using Illumina Genome Analyzer data alone, and one using combined data from Applied Biosystems SOLiD and Roche Life Sciences 454. Sites identified by both call sets as heterozygous (with the same alternate allele) were selected for the SIMTG DNM call set.

1.2 FIGL

The Family-aware Illumina Genotype Likelihood-based method (FIGL) is an approach to calculate the probability of a *de novo* point mutation at a single locus using the joint likelihood of the read-level data for all three trio members and the observed base in the public reference sequence. A likelihood is assigned for all 1000 possible unordered, labeled genotype configurations that the trio may assume. This provides a natural way of accommodating triallelic SNPs. The method is Bayesian, as the calculations incorporate a prior on the probabilities of observing a new mutation, an inherited variant, and the observed sampling configuration of derived alleles among parental chromosomes at a variable site. We will now go into specific details of each step of calculating $P(\text{DNM}|D)$ using FIGL.

Individual genotype likelihoods. The method begins with pre-computed individual genotype likelihoods. In practice these likelihoods were generated by SAMtools 0.1.7 using the Trio Pilot BAM files created from Illumina data and released by the 1000 genomes project (final Trio Pilot release, www.1000genomes.org).

For thoroughness we will briefly outline the SAMtools model that produced these likelihoods, but we refer the reader to the published description for the complete details (Li et al., 2008). The data for a locus consist of n reads, k of which contain the base X and $n-k$ of which contain base Y. A likelihood for each of the 10 possible labeled, unordered genotypes g is calculated as follows:

$$L(D|G=g) = \begin{cases} a_{n,n-k} & g = XX \\ \binom{n}{k}\frac{1}{2}^k\frac{1}{2}^{n-k} & g = XY \\ a_{n,k} & g = YY \\ a_{n,n} & \text{otherwise} \end{cases}$$

The term $a_{n,k}$ is meant to represent the probability of making k base calling errors from n nucleotides. This is a difficult probability to model well, considering that the error process is dependent on the nucleotides at the site to be called, as well as the local sequence context, position of the base in the read, etc. Error probabilities may be correlated within and across reads, and that correlation structure will be locus- and platform-specific. In SAMtools, $a_{n,k}$ is approximated as

$$a_{n,k} = c_{n,k}\prod_{i=0}^{j-1}\varepsilon_{i+1}^{\theta^i}.$$

Here ε_i is the ith smallest base error probability and $c_{n,k}$ is a function of ε_i, details on how to calculate $c_{n,k}$ are provided in (Li et al., 2008). The parameter θ is an unkown that specifies the dependency of errors at a site. By default, SAMtools uses a single, empirically determined, value of $\theta = 0.85$ to model dependency at all sites. A final important point is that SAMtools limits the base quality at any given position to be less than or equal to the mapping quality of the read being analysed.

Trio Model. We receive $L(D|G)$ for each trio member from SAMtools as described above. However the phenomenon we are trying to detect is a property of the trio as a whole so we may increase power and specificity by modeling the data as a whole. We add a subscript to the genotype term to indicate mother, father and child, as G_M, G_F, G_C. Then we write the joint likelihood for the trio as

$$L(G_M, G_F, G_C|D) = L(D|G_M, G_F, G_C)L(G_C|G_M, G_F) \quad (1)$$

$$L(G_M, G_F). \quad (2)$$

$$(3)$$

We now define the likelihood function for each of the terms on right side of the equation.

$L(G_M, G_F)$ is the prior of drawing two genotypes G_M and G_F from the population, while observing the base present in the public reference genome sequence, R. Our prior is loosely derived from the standard neutral coalescent. Empirically we know that two human chromsomes from European populations differ at roughly 1/1000 sites. There are $4N$ generations in a tree of two chromosomes, and $\frac{100}{12}N$ generations in a tree relating 5 chromosomes, so we assume approximately 2/1000 sites will be variable in a sample of 5 human chromosomes. Then the prior for sampling G_M and G_F as homozygous reference is 0.998. For the remaining configurations, conditional on a segregating site, we consider the frequency of the minor allele (will be "1" , 3/5 of the time and "2" 2/5 of the time) and the sampling configuration of the minor allele across individuals (one homozyous derived individual will be less common than two heterozygotes). To make the notation compact consider the unordered parental genotype configuration G_1, G_2. The major allele at a locus is labeled "A" and the minor allele "D". We calculate the prior probability of observing a particular set of alleles at these 5 chromosomes (parents and the reference) as,

$$L(G_1, G_2) = \begin{cases} 0.998 & G1, G2, R = AA, AA, A \\ 0.001 * 3/5 * 4/5 & AA, AD, A \\ 0.002 * 3/5 * 1/5 & AA, AA, D \\ 0.001 * 2/5 * 1/5 & AA, DD, A \\ 0.002 * 2/5 * 2/5 & AD, AD, A \\ 0.001 * 2/5 * 2/5 & AD, AA, D \end{cases}$$

This form that we specify for $L(G_M, G_F)$ is the same one used by the University of Michigan's trio-aware caller to generate SNP genotype calls for Pilot 2 (Trio Pilot) of the 1000 genomes project, and similar in principle to the one used by the FPIR pipeline described below.

$L(G_C|G_M, G_F)$ is the "transmission" likelihood; the likelihood that the child's genotype is G_C given the parent's genotypes are G_F and G_M. In our analysis this function is defined for all possible trio genotype configurations. For configurations compatible with Mendelian inheritance

the likelihoods are just the standard transmission probabilities (0.25, 0.5, or 1). For sites with Mendelian incompatabilities (i.e. sites that look like DNMs) we assign a probability of μ (one mutation) or μ^2 (two mutations) as necessary. In our final analysis we defined $\mu = 2.5 \times 10^{-7}$, as we multiplied our prior expectation of the mutation rate (2.5×10^{-8}) by a factor of 10 to account for the increase in apparent DNMs due to non-germline mutation.

$L(D|G_M, G_F, G_C)$ is the joint data likelihood given a particular trio genotype configuration. It is calculated simply by multiplying the three single genotype likelihoods together:

$$L(D|G_M, G_F, G_C) = L(D|G_M)L(D|G_F)L(D|G_C).$$

Posterior Probabilities. The ultimate output of this DNM caller is a posterior probability that a site contains a DNM, calculated using the following approach. Let M, D, and C be 10-element vectors containing the likelihoods of all 10 possible genotypes for the mother, father, and child, respectively. Then a rescaled version of the joint trio likelihood surface is obtained with the following steps:

1. $P = M \otimes D$
2. $F = P \otimes C$
3. $T = F \odot R$
4. $X = T \odot Y$

Where $\otimes$ is the Kronecker product operation, $\odot$ is the Schur product operation, R is the matrix of transmission likelihoods corresponding to each trio configuration, and Y is the matrix of $L(G_M, G_F)$ corresponding to the prior on each trio configuration. The maximum likelihood trio configuration compatible with DNM, x_{i-max}, x_{j-max}, is identified, and the posterior probability is calculated as:

$$P(\text{is DNM}|D) = \frac{x_{i-max}, x_{j-max}}{\sum_{i,j} x_{i,j}}$$

1.3 FPIR

The Family-aware Probabilistic Illumina Read-based (FPIR) algorithm is a probabilistic method to identify candidate *de novo* mutations starting from read-level data. The approach considers each genomic site separately and uses the aligned reads for each individual within a trio to simultaneously infer all three genotypes.

Model. A probabilistic model was constructed to account for the uncertainty and error in the process of de novo mutation discovery (Cartwright et al. in prep). The method makes use of the relatedness between individuals and produces posterior probabilities of *pedigrees* at each site rather than posterior probabilities on individual *genotypes* to facilitate correctly determining the family members genotypes. While the data for a single site is considered jointly among the family members, within a single individual the data for each site is treated independently. This simplifies calculations without sacrificing much accuracy in terms of modeling mutation patterns. Polymorphic sites are treated as independent from one another since closely linked double heterozygotes will be rare when the per-site diversity level is low. In fact, the vast majority of sites are expected to be non-polymorphic, in which case linkage between sites has little to no effect.

Conceptually the data in our model belongs to one of two categories: 1) the *observed data*, R, consists of aligned sequence reads from each individual and 2) the *hidden data*, H, from which the observed data is derived, is comprised of the actual parental and offspring genotypes, the pattern of inheritance, mutation events, how the chromosomes are sampled by the sequencing reads, and sequencing error.

The probability of the data at a site is $P_S(R, H|\Theta)$, where Θ contains the parameters of the model: the population per-site diversity parameter, θ, the per-site, per-generation mutation rate, μ, and ε, the per-site sequencing error rate. For this application the parameters are assumed to have fixed per-locus values of $\theta = 0.001$, $\mu = 2.0 \times 10^{-7}$, and $\varepsilon = 0.01$ across the entire genome. Although the germ-line mutation rate has been estimated to be on the order of 2.0×10^{-8}, we increased this by an order of magnitude to account for the possibility of cell-line and somatic mutation events in addition to those occuring within the germ-line. For a single family including two parents and one offspring (hereafter referred to as a trio), these three events are indistinguishable. The generative model works as follows:

1. Starting from the root of the pedigree, the parental alleles m_a, m_b, f_a, and f_b are sampled from a population at equilibrium allowing up to three segregating alleles. The distribution of these alleles is calculated in a coalescent framework utilizing θ and allowing for at most two mutations on the coalescent genealogy. These four alleles are the founders of the pedigree and form the genotypes of the two parents.

2. From this sample of alleles, one is transferred from each parent to the offspring, with the possibility of germ-line mutation at rate , to form the child genotype $o_a o_b$. The allele on chromosome a in the offspring is arbitrarily labeled as the allele inherited from the mother.

3. Because a trio offers little to no power to distinguish somatic and germ-line mutations, we assume that the somatic mutation rate is 0.

4. The genotypes are sampled by sequencing with an error rate of ε per base, producing the observed data $R = \{R_M, R_F, R_O\}$. There are N_{RM} reads in total sampling the mother

at this site, N_{RF} reads from the father, and N_{RO} offspring reads. Each read samples a chromosome at random.

Probability of the observed data. The probability of the full data at a single site given the parameters is reduced into several kernel functions:

$$P(R, H|\Theta) = P(m, f|\theta) \times P(o|m, f, \mu) \times P(R_M|m, \varepsilon) \times P(R_F|f, \varepsilon) \times P(R_o|o, \varepsilon)$$

The joint probability of the parent genotypes times the probability of the child genotype given the parents times the probabilities of the nucleotide reads at the site given the individual genotypes.

The hidden data, H, is not directly observed so we calculate the marginal probability of the observed data, R, given the parameters:

$$P(R|\Theta) = \sum_{H} P(R, H|\Theta)$$

This can be evaluated using the tree-peeling algorithm.

Mutation and Error Kernel. The Jukes-Cantor (1969) substitution model is used for the mutation and error kernels. For the mutation kernel the possible inheritance patterns are summed over within this kernel. This results in the following

$$P(o|m, f, \mu) = \left\{ \begin{array}{ll} \frac{1}{4} + \frac{3}{4}e^{-\mu} & \text{if } o_a = m_a = m_b \\ \frac{1}{4} + \frac{1}{4}e^{-\mu} & \text{if } o_a = m_a \neq m_b \text{ or } o_a = m_b \neq m_a \\ \frac{1}{4} + \frac{3}{4}e^{-\mu} & \text{if } o_a \neq m_a \text{ and } o_a \neq m_b \end{array} \right\} \times \left\{ \begin{array}{ll} \frac{1}{4} + \frac{3}{4}e^{-\mu} & \text{if } o_b = f_a = f_b \\ \frac{1}{4} + \frac{1}{4}e^{-\mu} & \text{if } o_b = f_a \neq f_b \text{ or } o_a = f_b \neq f_a \\ \frac{1}{4} + \frac{3}{4}e^{-\mu} & \text{if } o_b \neq f_a \text{ and } o_a \neq f_b \end{array} \right\}$$

The error kernel is constructed similarly to the mutation kernel but must consider the entire set of reads for that individual at that site. For example:

$$P(R_{Oi}|o, \varepsilon) = \left\{ \begin{array}{ll} \frac{1}{4} + \frac{3}{4}e^{-\varepsilon} & \text{if } R_{Oi} = o_a = o_b \\ \frac{1}{4} + \frac{1}{4}e^{-\varepsilon} & \text{if } R_{Oi} = o_a \neq o_b \text{ or } R_{Oi} = o_b \neq o_a \\ \frac{1}{4} + \frac{3}{4}e^{-\varepsilon} & \text{if } R_{Oi} \neq o_a \text{ and } R_{Oi} \neq o_b \end{array} \right\}$$

and

$$P(R_O|o, \varepsilon) = \prod_{i=1}^{N_{RO}} P(R_{Oi}|o, \varepsilon)$$

Population Kernel. In the population kernel the probability of a sample of four alleles constituting the two genotypes of the parents is defined using coalescent theory under the finite sites model of mutation. The nucleotides m_a, m_b, f_a, and f_b constitute a sample of four alleles from a finite, randomly-mating population, with effective size of N_e. Allowing up to three different allele states amongst this sample of four, the following allele spectra are possible:

1. 4-0-0: all alleles are the same
2. 3-1-0: two alleles, with the minor allele occuring once
3. 2-2-0: two alleles, with the minor allele occuring twice
4. 2-1-1: three allele states

When calculating the probability of each parental genotype sample, the possible specific nucleotide used (assuming all are equal in frequency) and the possible order in which they occur must be considered. For simplicity we will describe $P(m, f|\theta)$ based on these spectra and include corrections for nucleotides and orderings. In order to calculate the probability of allele spectra, we use coalescent theory. Mutations are allowed to occur continuously along the genealogical tree connecting these four sampled chromosomes, back to their most recent common ancestor. In general the distribution of genotype patterns is given by the integration of the probability of mutations occurring over the length of the genealogical tree and summing the results across all groupings of mutations that can produce the possible allele spectrum.

$$P_{mf}(4-0-0|\theta) = \frac{1}{4}(\frac{6}{6+11\theta} + \frac{121\theta^2 \times 0.163}{2(3+11\theta)(6+11\theta)})$$

$$P_{mf}(3-1-0|\theta) = \frac{1}{12\times 4}(\frac{48.7\theta}{36+132\theta} + \frac{121\theta^2 \times 0.271}{2(3+11\theta)(6+11\theta)})$$

$$P_{mf}(2-2-0|\theta) = \frac{1}{12\times 3}(\frac{17.3\theta}{36+132\theta} + \frac{121\theta^2 \times 0.225}{2(3+11\theta)(6+11\theta)})$$

$$P_{mf}(2-1-1|\theta) = \frac{1}{24\times 6}(\frac{121\theta^2 \times 0.341}{2(3+11\theta)(6+11\theta)})$$

Detecting *de novo* mutations. To find candidate sites that have de novo mutations, the above model is used to estimate the probability that a site contains a mutation in the child. This can be easily estimated from the probability that the site does not contain a mutation:

$$P(\text{at least 1 mutation}|R,\Theta) = 1 - \frac{\sum_H I(H=\emptyset)P(R,H|\Theta)}{\sum_H P(R,H|\Theta)}$$

The indicator function, $I(H = \emptyset)$, is 1 if a history contains no mutations and 0 otherwise. Due to the relatively small number of histories that contain no mutations this can be easily calculated, especially when coupled with the tree-peeling algorithm.

The probability is used to rank sites by order of those most likely to have a mutation event. Sites with probability above 10% were considered candidates.

Data pre-processing. Stringent quality filters were applied to the pileup files of each individual. Filters were customized to the genome being analyzed. For any given site, reads were removed with mapping quality below a set threshold (MapQ $<$ 63-67 depending on sample), or base quality below a set threshold (BaseQ $<$ 21 for all samples).

Taken together these values provide a measure of the degree of confidence one can have when determining a genotype based upon a set of reads. The data structure uses discrete read-level base observations where one observation carries as much weight as any other. To approximate this uniformity in the real data, minimum quality filters were applied using levels set by a detailed investigation of the MapQ and BaseQ value distribution in each individual pileup. In addition to the quality filters, due to the ability of many reads from repetitive regions to align incorrectly in tandem, sites with a read depth greater than a maximum threshold defined by specific pileup investigation were ignored in the analysis (read depth $>$ 54-75 depending on the sample).

1.4 Filtering and merging of calls to create validation lists

After DNM discovery, we attempted to remove artifactual calls by applying a common set of filters to the FIGL and FPIR callsets. These filters fell into three broad categories: (i) proximity to other known variants, (ii) overlap with primary genome sequence known to be problematic for mapping and assembly, and (iii) other properties of the read-level data. In order to assess the impact of our assumptions about what filters were appropriate, we decided to leave the SIMTG set of candidates sites unfiltered. We used the following filters:

- **Simple Repeats:** Union of bases spanned by track of the same name, from UCSC human genome assembly hg18. This covered 53240703bp.
- **segmental duplications:** Union of bases spanned by track of the same name, from UCSC human genome assembly hg18. This covered 142256390bp.
- **dbSNP:** Union of bases spanned by single nucleotide variants in dbSNP build 129, excluding single nucleotide variants discovered only in NA12878 and NA19240 by Kidd et al. (2008). This covered 14944456bp.
- **GSVC 42M probe CNV map:** Union of bases spanned by CNV calls described in (Conrad et al., 2010). This covered 157719857bp.
- **Read Depth:** Sites where at least one trio member has no mapped Illumina reads. This covered 191658532bp in CEU and 190614209 bp in YRI.
- **Broad multiple realignment regions:** Sites where heterozygous genotype calls in either NA12878 or NA19240 are transformed to homozygous genotype calls after multiple sequence

realignment, as implemented by the Broad Institute GATK. This covered 1278665bp in CEU and 1835556bp in YRI.

- **Pindel short indel calls:** Sites within 100bp (+ or -) of a small indel call in NA12878 or NA19240, made by the program "pindel", Ye et al. (2009). This covered 11044933bp in CEU and 17065637bp in YRI.

Taking the union of all filters 468051448bp were filtered in CEU and 473283432bp in YRI.

Our validation philosophy was to identify as many DNMs as possible, thus we used a permissive threshold for calling, generated a long list of variants for each trio, and attempted to validate all of these experimentally. In each trio we included all post-filtered sites assigned a posterior probability greater than 10% by either FPIR or FIGL, as well as the top 500 unfiltered SIMTG calls not present in the union of the FPIR and FIGL sets. This led to 2750 candidate calls in the YRI trio and 3236 in the CEU trio.

2 Validation Experiments

We attempted to validate all 2750 candidate DNMs in the YRI trio and all 3236 in the CEU trio using two parallel approaches based on next-generation sequencing.

Samples. We sequenced genomic DNA from all 6 lymphoblastoid cell lines (LCLs) that were used to generate the Trio Pilot data (but note, perhaps different lots of cells). In order to separate germline from somatic (or cell-line) mutations, we screened additional DNA samples with the same validation assays (Table S2). The CEU trio is part of a larger, 15 member CEPH/UTAH pedigree (number 1463, which includes the partner of NA12878 and 11 of her children). We included DNA from the 11 grandchildren to confirm germline status by inheritance. For the YRI trio (Y117), the Coriell Institute provided genomic DNA for individuals NA19238, NA19239, NA19240 extracted from the same primary blood samples that were used to generate their LCLs. Upon receipt at the Sanger, 10ng of blood DNA from each sample was whole-genome amplified using Genomiphy-HY DNA amplification kit from GE Health, following the manufacturer's protocol. After amplification samples were ethanol precipitated and sent for PCR.

2.1 Validation Experiment I

Design. The first experiment comprised nested PCR amplification of putative DNMs followed by read-pair sequencing of pooled PCR products on the Illumina platform.

Nested PCR primers. Whole genome SNP genotype calls were obtained from the 1000 genomes project for both trios and this information was used in primer design. The internal amplicon sizes range from 95bp-200bp, with about 20% greater than 100bp. We were able to design informative assays for 91% sites using this approach.

The sequencing experiment was to designed to use 54bp pair-end reads. As most internal amplicons were 100bp, the amplicon was designed to be centered on the site of interest, thus providing double coverage. In some cases (e.g. where the internal amplicon was > 108bp) the internal amplicon was designed so that the site of interest was within 20bp of one end of the PCR product (and not in the primer).

For YRI, external primer pairs could not be designed for 26 sites, and internal pairs couldn't be designed for 228 sites. For CEU, external primers could not be designed for 26 sites, and internal pairs couldn't be designed for 250 sites. Inspection of the local sequence context indicated that most of these problem regions were highly repeat-rich. In total we designed 22,520 primers, and conducted about 111,000 PCRs.

Sample allocation. Illumina GAII was used to generate nine lanes of sequence, distributed as follows:

CEU - One lane each of pooled PCR product from cell line DNAs NA12878, NA12891, NA12892. One lane in which the pooled PCR products from all 11 grandchildren are themselves pooled together. Four lanes were sequenced in total.

YRI - One lane each of pooled PCR product from cell line DNAs NA19238, NA19239, NA19240. One lane of pooled PCR product from blood DNA of NA12878. One lane in which the PCR product from blood DNA of NA19238, NA19239 are themselves pooled together. Five lanes were sequenced in total.

Mapping and Postprocessing. Reads were mapped against the entire reference genome sequence (NCBI36) using BWA (Li and Durbin, 2009). Some post-processing was done using GATK (v1.0.2873,McKenna et al. (2010)). Mapping and coverage statistics for Validation Experiment I are provided in Table S3. The following pipeline was implemented for mapping and post-processing of reads. Program names are given for each step, followed by arguments.

- BWA aln, BWA sampe
- samtools import, samtools sort, samtools flagstat, samtools index
- GATK -T CountCovariates DBSNP dbsnp_130_b36.rod cov ReadGroupCovariate cov QualityScoreCovariate cov CycleCovariate cov DinucCovariat
- GATK -T TableRecalibration
- GATK -T RealignerTargetCreator
- GATK -T IndelRealigner

2.2 Validation Experiment II

Design In the second experiment we resequenced all candidate *de novo* mutations using Agilent's SureSelect Target Enrichment System and ABI SOLID3 Plus sequencing. Agilent's eArray was used to design a SureSelect library of 120-bp oligos (baits). Genomic coordinates of putative DNMs were uploaded into eArray and initially regions defined by RepeatMasker and Tandem Repeat Masker Finder were omitted from the design process and no overlap of the baits with the repeats was allowed. However only 2432 baits could be designed in this case. Stringency of the masks was then gradually relaxed which resulted in baits designed for 5921 (99%) of loci.

Capture experiments were performed independently on each DNA sample. A single Agilent SureSelect capture design was used targeting all sites on the CEU and YRI validation lists, meaning that putative DNMs in the CEU trio were sequenced in the YRI samples and *vice versa*. SOLiD fragment library construction and sequencing on octet slides was done following the manufacturers instructions. The SOLiD3 Plus reads from each sample were aligned independently using ABIs Bioscope v1.2 software with the default settings to NCBI build 36 of the human reference genome. The resulting alignments were stored in 19 separate BAM files, which were sorted and indexed using samtools v0.1.7a. The GATK library was used to normalize base qualities within the BAM files by testing for and removing effects of di-nucleotide and cycle co-variation (plots not shown).

Sample allocation CEU- DNA from NA12878, NA12891, NA12892, cell lines from all 11 grandchildren (NA12879, NA12880, NA12881, NA12882, NA12883, NA12884, NA12885, NA12886, NA12887, NA12888, NA12893), and NA12877 were sequenced.

YRI- DNA from NA19238, NA19239, NA12940, and blood DNA from NA19240 were sequenced.

The 3,236 putative DNM sites in the CEU trio had a mean coverage of 50.2x per sample across all samples; 2,422 sites had at least 5x coverage for each of NA12891, NA12892, NA12878, and at least one grandchild. The 2,750 YRI candidate DNM sites had a mean coverage of 38.2x per sample; 1,650 sites had at least 5x coverage for each of NA19238, NA19239, NA19240, and the blood sample from NA19240. Additional summaries of coverage and depth are in Table S3.

2.3 Validation Analyses

The goal of the validation experiments was to classify each putative *de novo* into one of 4 categories: germline *de novo*, non-germline *de novo*, variant inherited from the parents, or a false positive call (ie. no evidence of variation in any sample). We thought of this goal as a model selection problem and created a modeling framework that would allow us to evaluate the joint likelihood of the data from both validation experiments (denoted D_1 and D_2) under each model.

For the CEU trio, we model separately the data from each parent's cell line, the cell line of NA12878, and the pooled data from the 3rd generation. Recall that in validation Experiment 1, these 3rd generation DNAs are pooled into a single lane of sequencing, while in Experiment 2

each sample was sequenced independently. For the YRI trio, the modeling is analogous, except the blood DNAs of NA19238 and NA19239 were pooled prior to sequencing in Experiment 1 and were not sequenced in Experiment 2. We will present the modeling approach using the CEU family as an example.

Let m_m^1, m_r^1 be the number of reads from the mother with the mutant allele and reference allele, respectively, in Experiment 1 (indicated by superscript), and $m_T^1 = m_m^1 + m_r^1$. Use the same convention for g, c, d, the data for the 3rd generation, the trio child, and the father, and use a superscript '2', as in m_m^2, to indicate data coming from the second experiment. We assume that in the case of a heterozygous locus, the number of mutant reads from each lane should be binomially distributed with an appropriate parameter defined by the experimental design, f_o^1 in the case of the pooled offspring in Experiment 1, f_s^1 for all other samples in Experiment 1. We allow for different parameter values in Experiment 2, f_o^2 and f_s^2, to accommodate effects such as differences in reference bias during mapping.

In the case of a sample that is homozygous reference at the mutant position, the number of reads with the mutant allele should be Poisson distributed with some rate equal to the error rate. This error rate is predefined and considered fixed for Experiment 2, denoted e^2, but for Experiment 1, which often produced $>$ 1000X coverage at the mutant site, we attempt to estimate locus-specifice error rates from the data. The four models (germline DNM, non-germline DNM, inherited, false positive) differ from one another in the configurations of binomial and Poisson likelihoods, but they also differ in the way the Poisson rate is estimated from the data, as described below, giving $e_I^1, e_{II}^1, e_{III}^1, e_{IV}^1$. Putting this all together, the likelihood for the first model is written:

Model I True Germline *de novo*.

$$L(M_I|D_1) = \text{Bin}(g_m^1; g_T^1, f_o^1) * \text{Pois}(m_m^1, m_T^1 * e_I^1) \tag{4}$$
$$* \text{Pois}(d_m^1, d_T^1 * e_I^1) * \text{Bin}(c_m^1; c_T^1, f_s^1) \tag{5}$$
$$\tag{6}$$

and

$$L(M_I|D_2) = \text{Bin}(g_m^2; g_T^2, f_o^2) * \text{Pois}(m_m^2, m_T^2 * e^2) \tag{7}$$
$$* \text{Pois}(d_m^2, d_T^2 * e^2) * \text{Bin}(c_m^2; c_T^2, f_s^2) \tag{8}$$
$$\tag{9}$$

and then

$$L(M_I|D_1, D_2) = L(M_I|D_1)L(M_I|D_2)$$

The likelihoods for the other models follow naturally.

Parameter values and parameter estimation. There are two sets of parameters that are needed to specify each model, (1) the expected mutant read frequencies for lanes containing germline variants, and (2) the sequencing error rate.

The expected mutant read frequency for sites of germline variants in the pooled grandchildren, f_o, differs from what is expected for germline variants present in a single unpooled sample, f_s. For Experiment 1 we define these as $f_o^1 = 0.25$ and $f_s^1 = 0.5$. For Experiment 2 we use $f_o^1 = 0.175$ and $f_s^1 = 0.35$, which accounts for the reference bias observed while doing alignments in color space.

For the error rate parameters we use a fixed value for Experiment 2, $e^2 = 0.005/3$. For Experiment 1 we use a model-dependent estimation process. This algorithm works as follows:

- Set the default error rate, $e_d^1 = 0.003$.
- Model I, germline *de novo*. Set e_I^1 to the minimum mutant read frequency from the mother and the father. If both parents are missing data, $e_I^1 = e_d^1$. If $e_I^1 > 0.1$ or $e_I^1 < e_d^1$, set $e_I^1 = e_d^1$. The rationale here is that we don't believe we can accurately measure an error rate below the default rate with the data from one sample. If the observed mutant frequency is too high then the site is probably polymorphic (or some other artifact is present).
- Model II, non-germline *de novo*. Set e_{II}^1 to the mutant read frequency in the combined set of data from the mother, the father, and the grandchildren. If the total number of reads in these samples is < 100, or if $e_{II}^1 > 0.1$ or if $e_{II}^1 = 0$, set $e_{II}^1 = e_d^1$.
- Model IIIa, variant inherited from the mother. Similar to Model I. Set e_{III}^1 to the minimum mutant read frequency from the father. If the father missing data set to default. If $e_{III}^1 > 0.1$ or $e_{III}^1 < e_d^1$, set $e_{III}^1 = e_d^1$.
- Model IIIb, variant inherited from the father. Analogous to Model IIIb.
- Model IV, False positive. Under this model, all samples should be homozygous for the reference allele and all mutant reads are artifacts of base calling or alignment. Estimate e_{IV}^1 as the proportion of mutant reads in the combined set of all reads from all lanes.

Model selection. Classification of sites is done by maximum likelihood. We require that 2 times the difference in log likelihood between the best-fitting model and the next best fitting model to be greater than 15 in order to classify the site; otherwise we consider the data uninformative and assign the site a "no call".

Additional filters. Prior to generating base counts for each site, we used the following filters on the read level data from Experiment 1:

- remove all reads less than 50 bp long, after removing soft clipped sequence
- remove reads with base quality less than 15 at the mutant position
- remove reads with 4 or more mismatches to the reference

We also implemented a simple filter on the validation assignments themselves, based on the observed read depth across all samples and experiments:

CEU: if the total coverage of NA12878, NA12891, or NA12892 is less than 100 reads in Experiment 1 AND less than 10 reads in Experiment 2, we do not call this site. If the total coverage across all grandchildren is less than 100 reads in Experiment 1 AND less than 100 reads in Experiment 2, we do not call this site.

YRI: if the total coverage of NA19238, NA19239, or the blood DNA from NA19240 is less than 100 reads in Experiment 1 AND less than 10 reads in Experiment 2, we do not call this site.

Finally, there were a small number of loci with data properties that were not well captured with our model-based validation process, but identified as problematic and removed by hand.

Validation Results. The goal of the validation experiments was to classify each putative *de novo* into one of 4 categories: germline *de novo*, non-germline *de novo*, variant inherited from the parents, or a false positive call (ie. no evidence of variation in any sample). Sites that could not be unambiguously classified were allocated to a fifth category, "no call". The counts for each category in the CEU trio were: germline *de novo*, 49; somatic *de novo*, 952; inherited, 129; false positive, 1304; no call, 802. The counts for each category in the YRI trio were: germline *de novo*, 35; somatic *de novo*, 634; inherited, 335; false positive, 1065; no call, 681. These counts have also been stratified by analysis algorithm and plotted in Figure S2. A visualization of the mutant read frequencies from the combined validation experiments is presented in Figure S1.

3 Estimation of mutation rate

The sex-averaged mutation rate for each trio was estimated by correcting for the false negative rate in DNM discovery by the combination of the three DNM discovery algorithms, correcting for the false negative rate in DNM validation (the proportion of putative DNMs that were classified as being inconclusive after validation) and dividing by the number of bases that passed the genome filters and thus had been scrutinized for DNMs. Uncertainty is estimated from the Poisson confidence intervals on the number of DNMs observed. The sex-specific rates were estimated by scaling the sex-averaged rates by the proportion of haplotyped DNMs that were ascribed to paternal and maternal gemlines, with uncertainty estimated from the Poisson confidence intervals on the numbers of haplotyped DNMs ascribed to either parental germline. We go into specific details of all calculations in the following sections.

3.1 Power analysis

We estimated the experiment-wide false negative rate during the discovery phase in two ways; first, by simulation, and second, empirically, by comparing sensitivity of the different discovery algorithms to the set of validated germline DNMs. Due to the fact that properties of the read data vary systematically among samples, and that some groups used sample-specific information (i.e. the multiple platform data available for the trio offspring but not the parents), we did not consider the read data to be exchangable among individuals. Our simulation strategy involved using the HapMap 3 SNP data for these samples to identify all $\sim 30,000$ chromosome 1 heterozygous sites in the trio child and measuring each discovery algorithm's sensitivity at these sites using the 1000 genomes trio pilot data. As true DNMs are extremely rare, our reasoning was that the major source of false negatives in the real analysis would be due to miscalling a child with a DNM as a homozygote reference, and not miscalling a parent as heterozygote at a DNM position. In the simulation, sites at which a trio configuration with a heterozygous child was assigned a posterior probability > 0.10 (FIGL and FPIR) or a quality score equivalent to the threshold used in discovery (SIMTG) were treated as *de novo* calls. The virtue of having all three groups estimate false negative rates on a common dataset is that it models covariance in calling between algorithms. Based on these results we estimate the experiment-wide (across all algorithms) false negative rate of 7% in the YRI trio and 4% in the CEU trio.

We also assessed our false negative rate empirically, by *post-hoc* analysis of the validation results. Briefly, by comparison to the union of validated germline DNMs called by all three methods, we can obtain caller-specific false negative rates. Based on these results, we can estimate false negative rates considering complete dependence (3% YRI, 4.44% CEU) or independence (0.3% YRI, 0.12% CEU) among calling algorithms.

Finally, we examined the impact of whole genome amplification as a potential source of false negatives in the YRI validation. If allele dropout occurs at a significant rate during WGA, it is possible that the mutant allele will sometimes be lost specifically in the blood of NA19240, leading us to artifactually classify the site as a cell-line mutation. To empirically gauge the importance of this behavior, we studied 335 sites which could be unambiguously classified as inherited variants on the basis of cell line DNA alone. In the Illumina validation data only 1 site showed strong support of allelic dropout in the blood of NA19240, chr19:14131791. At this locus there are 4079 and 4442 reads with the reference and mutant alleles from the cell line of NA19240, while there are 3846 and 0 reads from the WGA'd blood of NA19240. Looking at these same 335 sites in the SOLiD data, there was no detectable evidence of allelic dropout (but note, not all sites are informative). Together these numbers suggest a lower limit of $1/335 = 0.3\%$ for the rate of misclassifying germline DNMs as non-germline DNMs as a result of allele dropout during WGA of NA19240 blood.

Conservatively assuming complete dependence in the sensitivity of callers, both the empircal and simulation approaches to estimating power suggest that we have missed 5% of the true DNMs in the portion of the CEU genome analyzed by all three groups, and 3.2% in the analogous section

of the YRI genome.

3.2 Rate Estimate

The overall mutation rate provided in the main text is based on only for the portion of the genome analyzed by all 3 centers (i.e. not filtered by the FPIR/FIGL-specific filters). The total number of bases interrogated was 2,555 Mb in CEU and 2,549 Mb in YRI. Based on our validation results the number of DNMs in the "unfiltered" CEU genome is estimated to be

45 observed validated DNMs $*$[2802 sites of attempted validation /2197 called sites]$*1/(1-0.04)$

and for YRI

35 observed validated DNMs $*$[2332 sites of attempted validation /1782 called sites]$*1/(1-0.07)$

dividing by the total number of base pairs interrogated gives the following point estimates of the rate:

$$\text{CEU} = 1.17 \times 10^{-8}$$

$$\text{YRI} = 0.97 \times 10^{-8}$$

These values are derived using the simulation-based estimation of experiment-wide power, which we believe to be the most accurate way of estimating the false negative rate (FNR). Given that there is some uncertainty in the actual FNR of the project, we have analyzed the relationship between the actual FNR and the inferred germline mutation rate. If the actual FNR has been underestimated by a factor of 10 (that is, the actual rate is in the range of $40\% - 60\%$), which seems very unlikely, our estimates of the mutation rate will be only 2-3 times lower than the true rate.

3.3 Performance of New Sequencing Technologies.

To assess the sensitivity and specificity of newer technologies, we examined call sets for NA12878 and NA19240 generated from newer technology at sites identified as germline, somatic, or false-positive from validation. We found that with as little as $0.0\% - 7.8\%$ loss in sensitivity, newer technologies are $71.5\% - 93.5\%$ more specific at these sites. Specifically, we applied the SIMTG protocol to whole-genome Illumina HiSeq sequencing of NA12878 (McKenna et al., 2010). Of the 1,304 sites identified as false-positive heterozygote calls from the 1000 Genomes Trio Pilot DNM validation, 1,104 were identified as homozygous reference when using the HiSeq data, significantly

reducing the false-positive rate on these sites (by 93.5%). In addition, we considered Complete Genomics (Drmanac et al., 2010) calls for both offspring (NA12878 and NA19240), and found that the false-positive reduction rates were also quite high: 80.5% and 71.5% respectively. This suggests that future DNM analyses using new data may be significantly more specific due to improvement in sequencing technologies.

During the writing of this report we examined call sets for NA12878 and NA19240 made with two of the newest sequencing platforms (CG and HiSeq references). We found that the newer technologies reduce the number of false positive calls in our validation list by 71.5%-93.5% with as little as 0%-7.8% loss in sensitivity for true DNMs. Fourty-nine validated germline mutations were identified in CEU in the present study, 47/49 were id'd by CG and 49/49 by HiSeq, while 886/952 somatic DNMs were called by CG and 929/952 by HiSeq, and only 255/1304 false positives were called by CG and 85/1304 by HiSeq. In the YRI trio, 35/38 validated germline DNMs were called by CG, 501/634 somatic DNMs, and 301/1055 false positives.

4 Properties of *de novo* mutation

4.1 Parent-of-origin

We attempted to identify the parental origin for each of the germline *de novos*, using three different approaches. Full results are available in Table S1, which is a stand-alone file.

Haplotyping by direct observation of phase in Trio Pilot data. For both CEU and YRI, we identified all haplotype informative sites within 5kb of the *de novo*. These are sites in which the child is heterozygous, and either, (i) only one of the parents is heterozygous, or (ii) each parent is homozyogous for a different allele. We then mined the 1000 genomes data from NA12878 and NA19240 to identify individual reads (454) or pairs of reads (SOLiD and Illumina GA) that span both the haplotype informative site and the *de novo* site, thus allowing direct observation of the haplotype phase from a single molecule.

Segregation analysis of CEU ***de novos*****.** The CEU trio is part of a larger, three-generation CEPH pedigree. NA12878 bore 11 offspring, from all of which cell lines are available. We generated SNP genotype calls from the cell line DNAs all 11 offspring using the Affymetrix 6.0 oligo array. We merged these genotype calls with the genotypes produced by the International HapMap project on NA12878, NA12891, and NA12892, and jointly phased all 14 samples using a pedigree-aware algorithm implemented in BEAGLE 3.0.3 (Browning and Browning, 2009).

For each validated germline DNM, we defined the paternally and maternally inherited haplotypes in NA12878, using a 200 SNP window centered on the *de novo* position. We then classified each grandchild as having inherited the paternal or maternal haplotype from NA12878, by selecting the NA12878 haplotype with the most similarity to a haplotype in that grandchild. We then counted the number of reads of the mutant allele observed in paternal and maternal

haplotype carriers. Sites with at least 0.5% mutant read frequency one of the two haplotype backgrounds were assigned a parent of origin, which was simply determined by selecting the parent whose haplotype showed the greatest mutant read frequency.

Molecular haplotyping of YRI *de novos* We pursued a second experimental haplotyping strategy similar in spirit to our use of the Trio Pilot data, but more directed to the task. We use the same set of haplotype-informative of sites, this time designing long-range PCR assays that included both the *de novo* and the nearest informative site. Successfully amplified bands were inserted into plasmids, the plasmids were used to transform bacteria, and we attempted to sequence 6 independent plasmids from both T7 and SP6 for each transformation.

4.2 Functional Impact

It might be expected that somatic mutations would be under less selective constraint than germline variation. We used Genomic Evolutionary Rate Profiling (GERP) to quantify the extent of selective constraint at the site of each DNM (Cooper et al., 2005). GERP scores measure conservation as the difference between expected and observed rates of nucleotide substitution at a given human base. GERP scores are position-specific and estimated from aligned orthologous sequences, in this case from genomic alignments of 16 amniote species in Ensembl 58 (Compara.16_amniota_vertebrates_Pecan). GERP scores were extracted for genotypes found in the 1000 genomes trio pilot, as well as sites from the combined candidate de novo mutation lists from the CEU and YRI families (Table S1). The positions of these sites were converted from the hg18 coordinate system to hg19 using the LiftOver web tool available at the UCSC Genome Browser website (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

Using GERP to infer whether DNMs were occurring at selectively constrained sites, we found that somatic DNMs and germline DNMs arose at similiarly constrained sites in the CEU and YRI families (t-test p-values 0.8 and 0.4 respectively). Across both families we only observed a single coding germline DNM, a synonymous variant. The somatic DNMs appear at more conserved sites relative to inherited variants within each family (t-test, CEU p= 0.003 and YRI p= 0.02, Figure S3). We observed that 16/17 somatic DNMs in proteincoding sequences were missense mutations (Table 1, main text). This proportion of missense to synonymous somatic DNMs is significantly higher than similar ratios for segregating sites in these populations (1000 Genomes Project Consortium (2010), Fisher's exact test, $p < 0.0004$), which might reflect a relative lack of selective constraint, or adaptive selection in cell culture for somatic DNMs. We did not observe evidence that somatic DNMs were enriched as a result of somatic hypermutation, nor did we observe individual somatic DNMs that represent plausible candidates for mutations conferring a selective advantage for growth in cell culture.

4.3 Mutation context and content

As a first pass at dissecting the mutation process operating in different cell lineages, we examined the ratio of transition to transversion mutants in different subsets of the data. We examined the ratio in somatic vs. germline DNMs (0.9 vs. 2.5 in CEU and 1.1 vs. 3.4 YRI), DNMs of maternal vs. paternal germline origin (2.6 vs. 3.0), and germline DNMs in the YRI and CEU. We noted that germline DNMs have a significantly higher transition:transversion ratio than somatic DNMs (Fisher's exact test , $p < 3 \times 10^{-5}$), but that these ratios are not significantly different ($p > 0.05$) between families for either germline or somatic DNMs, and thus likely reflect fundamental differences between germline and somatic mutational mechanisms (Figure S4a). Both germline and somatic DNMs exhibited a significant mutational bias towards A/T composition (binomial test p= 0.01 and 0.006 respectively), although this mutation bias was significantly stronger at germline DNMs than at somatic DNMs (Fisher's exact test, p= 0.04) and was similar to previously reported germline mutation bias in humans (Lynch, 2010). We observed a higher fraction of CpG mutations in germline DNMs than somatic DNMs, but, with small numbers involved, this comparison was not statistically significant.

Analysis of context-specific mutation. In order to obtain a more detailed description of the validated mutations, we created a 4-letter labelling system that describes (A) the ancestral base at the mutated site, (B) the first base 5' to the mutation, (C) the first base 3' to the mutation, and (D) the mutant base. As a point of comparison, we created a relative rate matrix describing common polymorphism in the human intergenic region (IGR), using all SNPs reported in CEU by 1000 genomes project. Positions where the chimpanzee reference genome differs from the human reference genome were excluded, to increase the probability that the human reference allele is the ancestral allele. The matrix is calculated in the following way. For the ith premutation triplet, x_i, we count the number of occurrences of that triplet in the IGR of the reference genome. For each such triplet there are 3 possible non-reference alleles that can be observed. We count the number of times a particular triplet/variant combination occurs in the 1000 genomes project data, n_{ij}, and then compute entry z_{ij} of the matrix as $\frac{n_{ij}}{x_i}$.

The matrix derived from CEU chromosome 1 (IGR1) has an extremely high correlation (> 0.99) with the matrix derived from CEU chromosome 2 (IGR2), and also a high correlation with the matrix derived from YRI chromosome 1. The slope is always very close to 1, indicating that the table of mutation frequencies is only slightly variable regardless of which chromosome or population is examined.

To test the hypothesis that validated somatic mutations from NA12878 and NA19240 originate from the same mutation processes as common germline variants, we used the following method. Our test statistic is the correlation coefficient between the 192 elements in the mutation matrix calculated from the germline IGR polymorphisms and the matrix calculated from the n somatic DNMs in either NA12878 or NA19240. We generated a null distribution for this statistic by simulating n mutations from IGR2, calculating a new rate matrix, and then computing the sample

correlation coefficient with IGR1. We repeat this process 1000 times, and calculate the p-value as the proportion of simulated correlation coefficient smaller than the observed. By this analysis, the correlation of the germline polymorphism IGR with the somatic variants in NA12878 (0.90) and NA19240 (0.88) are extremely significant (p=0)

4.4 Properties of cell cultures

Comparison of the proportion of sequence reads supporting the mutant allele for somatic DNMs and germline variants revealed that the degree of clonality of the cell-lines derived from the two trio offspring is quite different; while the CEU cell-line seems fully clonal the YRI cell-line is only 70% clonal (Figure S4b). An additional source of apparent DNMs that result from somatic or cellline mutational processes events are clusters of variants present in the offspring but not either parent as a result of a deletion in one of the parental cell-lines (Figure S4c).

Identification of somatic deletions. It has been known for some time that clustered DNMs can be used to reliably identify both germline ((Conrad et al., 2006), (McCarroll et al., 2006) and somatic deletions ((Redon et al., 2006)). To quantify the number of somatic deletions in each trio, genotypes were assessed on each individual independently (e.g. with no knowledge of familial structure). We then looked for loci at which the offspring may have inherited an allele that was subsequently deleted in the parental cell lines. This was done by 1) Calculating the intervals of homozygosity in each parent, 2) identifying all sites at which Mendelian inheritance had been violated, and 3) Identifying regions of parental homozygosity with an elevated rate of Mendelian violations. A region of homozygosity in either parent in which at least four Mendelian violations were present, of which at least one was a *de novo* event, was considered a candidate somatic deletion. Of 15 candidate deletions in YRI, and 26 candidate deletions in CEU, two parental somatic deletions could be visually confirmed based on sequence coverage, both in the CEU trio (chr19:41,814,852-41,881,619 and chr15:81,093,996-81,225,674, both in the father). In each of the identified regions, more than 67% of de novo events were at dbSNP sites, suggesting that these are indeed somatic deletions.

The problem of cell line deletions is more pressing for the CEU trio than the YRI trio. We are only worried about deletions in the parental cell lines, and we have blood DNA from the YRI parents, therefore we should be able to exclude these artifacts in YRI. There is no way to definitively exclude small somatic deletion in the CEU parents at the moment (large events should be detected by our CGH experiments).

However, we believe we have cleaned the majority of CNV-based artifacts from the pilot 2 data by doing the following:

- filtering de novo calls falling within the high-resolution CNV map made by 42 million probe CGH array; this map includes CNVs calls from all trio members ((Conrad et al., 2010)).
- filtering de novos that fall in dbSNP sites

- manually censoring a small number of the remaining sites that are validated as germline and apparently clustered:

  in CEU those are- 1:245658580, chr1:245658581, chr12:41132850, chr12:41132851.

  and in YRI: chr1:37316638, chr1:37316639, chr10:87090349, chr10:87090351, chr17:4510332, chr17:4510336, chr6:5358109, chr6:5358111, chr8:67177220, chr8:67177222.

4.5 Transcription-coupled repair in cell lines

We assessed the impact of transcription-coupled repair (Bohr et al., 1985; Pleasance et al., 2010) in the two cell lines using the following procedure. We obtained array-based expression data for about 17000 genes previously generated on the unrelated Phase I HapMap samples and calculated the median expression level for each gene, separately within the YRI and CEU populations. These two sets of medians were highly correlated. We next annotated the number of somatic mutations falling within each gene. We fit a linear model to the number somatic mutations falling within each gene, using population-specific expression level and gene size as covariates. Both covariates were significant in each model (CEU: expression $p < 0.002$, gene size $p < 1 \times 10^{-15}$; YRI: expression $p < 0.002$, gene size $p < 1 \times 10^{-15}$). There was no significant strand bias.

References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–73, 2010.

V. A. Bohr, C. A. Smith, D. S. Okumoto, and P. C. Hanawalt. DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell*, 40:359–69, 1985.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84: 210–23, 2009.

D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, 38:75–81, 2006.

D. F. Conrad et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464:704–12, 2010.

G. M. Cooper et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, 15:901–13, 2005.

R. Drmanac et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327:78–81, 2010.

J. M. Kidd et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:56–64, 2008.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–60, 2009.

H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18:1851–8, 2008.

M. Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A*, 107:961–8, 2010.

S. A. McCarroll et al. Common deletion polymorphisms in the human genome. *Nat Genet*, 38: 86–92, 2006.

A. McKenna et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20:1297–303, 2010.

E. D. Pleasance et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463:184–90, 2010.

R. Redon et al. Global variation in copy number in the human genome. *Nature*, 444:444–54, 2006.

K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25:2865–71, 2009.

The 1000 Genomes Consortium (Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Steering Committee: David Altshuler (Co-Chair)[2-4], Richard M. Durbin (Co-Chair)[1], Gonçalo R. Abecasis[5], David R. Bentley[6], Aravinda Chakravarti[7], Andrew G. Clark[8], Francis S. Collins[9], Francisco M. De La Vega[10], Peter Donnelly[11], Michael Egholm[12], Paul Flicek[13], Stacey B. Gabriel[2], Richard A. Gibbs[14], Bartha M. Knoppers[15], Eric S. Lander[2], Hans Lehrach[16], Elaine R. Mardis[17], Gil A. McVean[11,18], Deborah A. Nickerson[19], Leena Peltonen*, Alan J. Schafer[20], Stephen T. Sherry[21], Jun Wang[22,23], Richard K. Wilson[17]

Production Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[14], David Deiros[14], Mike Metzker[14], Donna Muzny[14], Jeff Reid[14], David Wheeler[14] **BGI-Shenzhen** Jun Wang (Principal Investigator)[22,23], Jingxiang Li[22], Min Jian[22], Guoqing Li[22], Ruiqiang Li[22,23], Huiqing Liang[22], Geng Tian[22], Bo Wang[22], Jian Wang[22], Wei Wang[22], Huanming Yang[22], Xiuqing Zhang[22], Huisong Zheng[22] **Broad Institute of MIT and Harvard** Eric S. Lander (Principal Investigator)[2], David Altshuler[2-4], Lauren Ambrogio[2], Toby Bloom[2], Kristian Cibulskis[2], Tim J. Fennell[2], Stacey B. Gabriel (Co-Chair)[2], David B. Jaffe[2], Erica Shefler[2], Carrie L. Sougnez[2] **Illumina** David R. Bentley (Principal Investigator)[6], Niall Gormley[6], Sean Humphray[6], Zoya Kingsbury[6], Paula Kokko-Gonzales[6], Jennifer Stone[6] **Life Technologies** Kevin J. McKernan (Principal Investigator)[24], Gina L. Costa[24], Jeffry K. Ichikawa[24], Clarence C. Lee[24] **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[16], Hans Lehrach (Principal Investigator)[16], Tatiana A. Borodina[16], Andreas Dahl[25], Alexey N. Davydov[16], Peter Marquardt[16], Florian Mertes[16], Wilfiried Nietfeld[16], Philip Rosenstiel[26], Stefan Schreiber[26], Aleksey V. Soldatov[16], Bernd Timmermann[16], Marius Tolzmann[16] **Roche Applied Science** Michael Egholm (Principal Investigator)[12], Jason Affourtit[27], Dana Ashworth[27], Said Attiya[27], Melissa Bachorski[27], Eli Buglione[27], Adam Burke[27], Amanda Caprio[27], Christopher Celone[27], Shauna Clark[27], David Conners[27], Brian Desany[27], Lisa Gu[27], Lorri Guccione[27], Kalvin Kao[27], Andrew Kebbel[27], Jennifer Knowlton[27], Matthew Labrecque[27], Louise McDade[27], Craig Mealmaker[27], Melissa Minderman[27], Anne Nawrocki[27], Faheem Niazi[27], Kristen Pareja[27], Ravi Ramenani[27], David Riches[27], Wanmin Song[27], Cynthia Turcotte[27], Shally Wang[27] **Washington University in St. Louis** Elaine R. Mardis (Co-Chair) (Co-Principal Investigator) [17], Richard K. Wilson (Co-Principal Investigator)[17], David Dooling[17], Lucinda Fulton[17], Robert Fulton[17], George Weinstock[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], John Burton[1], David M. Carter[1], Carol Churcher[1], Alison Coffey[1], Anthony Cox[1], Aarno Palotie[1,28], Michael Quail[1], Tom Skelly[1], James Stalker[1], Harold P. Swerdlow[1], Daniel Turner[1]

Analysis Group: Agilent Technologies Anniek De Witte[29], Shane Giles[29] **Baylor College of Medicine** Richard A. Gibbs (Principal Investigator)[14], David Wheeler[14], Matthew Bainbridge[14], Danny Challis[14], Aniko Sabo[14], Fuli Yu[14], Jin Yu[14] **BGI-Shenzhen** Jun Wang (Principal Investigator)[22,23], Xiaodong Fang[22], Xiaosen Guo[22], Ruiqiang Li[22,23], Yingrui Li[22], Ruibang Luo[22], Shuaishuai Tai[22], Honglong Wu[22], Hancheng Zheng[22], Xiaole Zheng[22], Yan Zhou[22], Guoqing Li[22], Jian Wang[22], Huanming Yang[22] **Boston College** Gabor T. Marth (Principal Investigator)[30], Erik P. Garrison[30], Weichun Huang[31], Amit Indap[30], Deniz Kural[30], Wan-Ping Lee[30], Wen Fung Leong[30], Aaron R. Quinlan[32], Chip Stewart[30], Michael P. Stromberg[33], Alistair N. Ward[30], Jiantao Wu[30] **Brigham and Women's Hospital** Charles Lee (Principal Investigator)[34], Ryan E. Mills[34], Xinghua Shi[34] **Broad Institute of MIT and Harvard** Mark J. Daly (Principal Investigator)[2], Mark A. DePristo (Project Leader)[2], David Altshuler[2-4], Aaron D. Ball[2], Eric Banks[2], Toby Bloom[2], Brian L. Browning[35], Kristian Cibulskis[2], Tim J. Fennell[2], Kiran V. Garimella[2], Sharon R. Grossman[2,36], Robert E. Handsaker[2], Matt Hanna[2], Chris Hartl[2], David B. Jaffe[2], Andrew M. Kernytsky[2], Joshua M. Korn[2], Heng Li[2], Jared R. Maguire[2], Steven A. McCarroll[2,4], Aaron McKenna[2], James C. Nemesh[2], Anthony A.

Philippakis[2], Ryan E. Poplin[2], Alkes Price[37], Manuel A. Rivas[2], Pardis C. Sabeti[2,36], Stephen F. Schaffner[2], Erica Shefler[2], Ilya A. Shlyakhter[2,36] **Cardiff University, The Human Gene Mutation Database** David N. Cooper (Principal Investigator)[38], Edward V. Ball[38], Matthew Mort[38], Andrew D. Phillips[38], Peter D. Stenson[38] **Cold Spring Harbor Laboratory** Jonathan Sebat (Principal Investigator)[39], Vladimir Makarov[40], Kenny Ye[41], Seungtai C. Yoon[40] **Cornell and Stanford Universities** Carlos D. Bustamante (Co-Principal Investigator)[43], Andrew G. Clark (Co-Principal Investigator)[8], Adam Boyko[43], Jeremiah Degenhardt[8], Simon Gravel[43], Ryan N. Gutenkunst[44], Mark Kaganovich[43], Alon Keinan[8], Phil Lacroute[43], Xin Ma[8], Andy Reynolds[8] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Paul Flicek (Co-Chair, DCC) (Principal Investigator)[13], Fiona Cunningham[13], Javier Herrero[13], Stephen Keenen[13], Eugene Kulesha[13], Rasko Leinonen[13], William M. McLaren[13], Rajesh Radhakrishnan[13], Richard E. Smith[13], Vadim Zalunin[13], Xiangqun Zheng-Bradley[13] **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator)[45], Adrian M. Stütz[45] **Illumina** Sean Humphray (Project Leader)[6], Markus Bauer[6], R. Keira Cheetham[6], Tony Cox[6], Michael Eberle[6], Terena James[6], Scott Kahn[6], Lisa Murray[6] **Johns Hopkins University** Aravinda Chakravarti[7] **Leiden University Medical Center** Kai Ye[46] **Life Technologies** Francisco M. De La Vega (Principal Investigator)[10], Yutao Fu[24], Fiona C.L. Hyland[10], Jonathan M. Manning[24], Stephen F. McLaughlin[24], Heather E. Peckham[24], Onur Sakarya[10], Yongming A. Sun[10], Eric F. Tsung[24] **Louisiana State University** Mark A. Batzer (Principal Investigator)[47], Miriam K. Konkel[47], Jerilyn A. Walker[47] **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[16], Marcus W. Albrecht[16], Vyacheslav S. Amstislavskiy[16], Ralf Herwig[16], Dimitri V. Parkhomchuk[16] **US National Institutes of Health** Stephen T. Sherry (Co-Chair, DCC) (Principal Investigator)[21], Richa Agarwala[21], Hoda M. Khouri[21], Aleksandr O. Morgulis[21], Justin E. Paschall[21], Lon D. Phan[21], Kirill E. Rotmistrovsky[21], Robert D. Sanders[21], Martin F. Shumway[21], Chunlin Xiao[21] **Oxford University** Gil A. McVean (Co-Chair) (Co-Chair, Population Genetics) (Principal Investigator)[11,18], Adam Auton[11], Zamin Iqbal[11], Gerton Lunter[11], Jonathan L. Marchini[11,18], Loukas Moutsianas[18], Simon Myers[11,18], Afidalina Tumian[18] **Roche Applied Science** Brian Desany (Project Leader)[27], James Knight[27], Roger Winer[27] **The Translational Genomics Research Institute** David W. Craig (Principal Investigator)[48], Steve M. Beckstrom-Sternberg[48], Alexis Christoforides[48], Ahmet A. Kurdoglu[48], John V. Pearson[48], Shripad A. Sinari[48], Waibhav D. Tembe[48] **University of California, Santa Cruz** David Haussler (Principal Investigator)[49], Angie S. Hinrichs[49], Sol J. Katzman[49], Andrew Kern[49], Robert M. Kuhn[49] **University of Chicago** Molly Przeworski (Co-Chair, Population Genetics) (Principal Investigator)[50], Ryan D. Hernandez[51], Bryan Howie[52], Joanna L. Kelley[52], S. Cord Melton[52] **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)[5], Yun Li (Project Leader)[5], Paul Anderson[5], Tom Blackwell[5], Wei Chen[5], William O. Cookson[53], Jun Ding[5], Hyun Min Kang[5], Mark Lathrop[54], Liming Liang[55], Miriam F. Moffatt[53], Paul Scheet[56], Carlo Sidore[5], Matthew Snyder[5], Xiaowei Zhan[5], Sebastian Zöllner[5] **University of Montreal** Philip Awadalla (Principal Investigator)[57], Reed A. Cartwright[79], Ferran Casals[58], Youssef Idaghdour[58], Jonathan Keebler[58], Eric A. Stone[58], Martine Zilversmit[58] **University of Utah** Lynn Jorde (Principal Investigator)[59], Jinchuan Xing[59] **University of Washington** Evan E. Eichler (Principal Investigator)[60], Gozde Aksay[19], Can Alkan[60], Iman Hajirasouliha[61], Fereydoun Hormozdiari[61], Jeffrey M. Kidd[19,43], S. Cenk Sahinalp[61], Peter H. Sudmant[19] **Washington University in St. Louis** Elaine R. Mardis (Co-Principal Investigator)[17], Ken Chen[17], Asif Chinwalla[17], Li Ding[17], Daniel C. Koboldt[17], Mike D. McLellan[17], David Dooling[17], George Weinstock[17], John W. Wallis[17], Michael C. Wendl[17], Qunyuan Zhang[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], Cornelis A. Albers[62], Qasim Ayub[1], Senduran Balasubramaniam[1], Jeffrey C. Barrett[1], David M. Carter[1], Yuan Chen[1], Donald F. Conrad[1], Petr Danecek[1], Emmanouil T. Dermitzakis[63], Min Hu[1], Ni Huang[1], Matt E. Hurles[1], Hanjun Jin[64], Luke Jostins[1], Thomas M. Keane[1], Si Quang Le[1], Sarah Lindsay[1], Quan Long[1], Daniel G. MacArthur[1], Stephen B. Montgomery[63], Leopold Parts[1], James Stalker[1], Chris Tyler-Smith[1], Klaudia Walter[1], Yali Xue[1], Yujun Zhang[1] **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)[65,66], Michael Snyder (Co-Principal Investigator)[43], Alexej Abyzov[65], Suganthi Balasubramanian[67], Robert Bjornson[66], Jiang Du[66], Fabian Grubert[43], Lukas Habegger[65], Rajini

Haraksingh[65], Justin Jee[65], Ekta Khurana[67], Hugo Y.K. Lam[43], Jing Leng[65], Xinmeng Jasmine Mu[65], Alexander E. Urban[43,68], Zhengdong Zhang[67]

Structural Variation Group: **BGI-Shenzhen** Yingrui Li[22], Ruibang Luo[22] **Boston College** Gabor T. Marth (Principal Investigator)[30], Erik P. Garrison[30], Deniz Kural[30], Aaron R. Quinlan[32], Chip Stewart[30], Michael P. Stromberg[33], Alistair N. Ward[30], Jiantao Wu[30] **Brigham and Women's Hospital** Charles Lee (Co-Chair) (Principal Investigator)[34], Ryan E. Mills[34], Xinghua Shi[34] **Broad Institute of MIT and Harvard** Steven A. McCarroll (Project Leader)[2,4], Eric Banks[2], Mark A. DePristo[2], Robert E. Handsaker[2], Chris Hartl[2], Joshua M. Korn[2], Heng Li[2], James C. Nemesh[2] **Cold Spring Harbor Laboratory** Jonathan Sebat (Principal Investigator)[39], Vladimir Makarov[40], Kenny Ye[41], Seungtai C. Yoon[40] **Cornell and Stanford Universities** Jeremiah Degenhardt[8], Mark Kaganovich[43] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Richard E. Smith[13], Xiangqun Zheng-Bradley[13] **European Molecular Biology Laboratory** Jan O. Korbel[45] **Illumina** Sean Humphray (Project Leader)[6], R. Keira Cheetham[6], Michael Eberle[6], Scott Kahn[6], Lisa Murray[6] **Leiden University Medical Center** Kai Ye[46] **Life Technologies** Francisco M. De La Vega (Principal Invesigator)[10], Yutao Fu[24], Heather E. Peckham[24], Yongming A. Sun[10] **Louisiana State University** Mark A. Batzer (Principal Investigator)[47], Miriam K. Konkel[47], Jerilyn A. Walker[47] **US National Institutes of Health** Chunlin Xiao[21] **Oxford University** Zamin Iqbal[11] **Roche Applied Science** Brian Desany[27] **University of Michigan** Tom Blackwell (Project Leader)[5], Matthew Snyder[5] **University of Utah** Jinchuan Xing[59] **University of Washington** Evan E. Eichler (Co-Chair) (Principal Investigator)[60], Gozde Aksay[19], Can Alkan[60], Iman Hajirasouliha[61], Fereydoun Hormozdiari[61], Jeffrey M. Kidd[19,43] **Washington University in St. Louis** Ken Chen[17], Asif Chinwalla[17], Li Ding[17], Mike D. McLellan[17], John W. Wallis[17] **Wellcome Trust Sanger Institute** Matt E. Hurles[1] (Co-Chair) (Principal Investigator), Donald F. Conrad[1], Klaudia Walter[1], Yujun Zhang[1] **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)[65,66], Michael Snyder (Co-Principal Investigator)[43], Alexej Abyzov[65], Jiang Du[66], Fabian Grubert[43], Rajini Haraksingh[65], Justin Jee[65], Ekta Khurana[67], Hugo Y.K. Lam[43], Jing Leng[65], Xinmeng Jasmine Mu[65], Alexander E. Urban[43,68], Zhengdong Zhang[67]

Exon Pilot Group: **Baylor College of Medicine** Richard A. Gibbs (Co-Chair) (Principal Investigator)[14], Matthew Bainbridge[14], Danny Challis[14], Cristian Coafra[14], Huyen Dinh[14], Christie Kovar[14], Sandy Lee[14], Donna Muzny[14], Lynne Nazareth[14], Jeff Reid[14], Aniko Sabo[14], Fuli Yu[14], Jin Yu[14] **Boston College** Gabor T. Marth (Co-Chair) (Principal Investigator)[30], Erik P. Garrison[30], Amit Indap[30], Wen Fung Leong[30], Aaron R. Quinlan[32], Chip Stewart[30], Alistair N. Ward[30], Jiantao Wu[30] **Broad Institute of MIT and Harvard** Kristian Cibulskis[2], Tim J. Fennell[2], Stacey B. Gabriel[2], Kiran V. Garimella[2], Chris Hartl[2], Erica Shefler[2], Carrie L. Sougnez[2], Jane Wilkinson[2] **Cornell and Stanford Universities** Andrew G. Clark (Co-Principal Investigator)[8], Simon Gravel[43], Fabian Grubert[43] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Paul Flicek (Principal Investigator)[13], Richard E. Smith[13], Xiangqun Zheng-Bradley[13] **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[21], Hoda M. Khouri[21], Justin E. Paschall[21], Martin F. Shumway[21], Chunlin Xiao[21] **Oxford University** Gil A. McVean[11,18] **University of California, Santa Cruz** Sol J. Katzman[49] **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)[5], Tom Blackwell[5] **Washington University in St. Louis** Elaine R. Mardis (Principal Investigator)[17], David Dooling[17], Lucinda Fulton[17], Robert Fulton[17], Daniel C. Koboldt[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], Senduran Balasubramaniam[1], Allison Coffey[1], Thomas M. Keane[1], Daniel G. MacArthur[1], Aarno Palotie[1,28], Carol Scott[1], James Stalker[1], Chris Tyler-Smith[1] **Yale University** Mark B. Gerstein (Principal Investigator)[65,66], Suganthi Balasubramanian[67]

Samples and ELSI Group: Aravinda Chakravarti (Co-Chair)[7], Bartha M. Knoppers (Co-Chair)[15], Leena Peltonen (Co-Chair)*, Gonçalo R. Abecasis[5], Carlos D. Bustamante[43], Neda Gharani[69], Richard A. Gibbs[14], Lynn Jorde[59], Jane S. Kaye[70], Alastair Kent[71], Taosha Li[22], Amy L. McGuire[72], Gil A. McVean[11,18], Pilar N. Ossorio[73], Charles N. Rotimi[74], Yeyang Su[22], Lorraine H. Toji[69], Chris Tyler-Smith[1]

Scientific Management: Lisa D. Brooks[75], Adam L. Felsenfeld[75], Jean E. McEwen[75], Assya Abdallah[76], Christopher R. Juenger[77], Nicholas C. Clemm[75], Francis S. Collins[9], Audrey Duncanson[20], Eric D. Green[78], Mark S. Guyer[75], Jane L. Peterson[75], Alan J. Schafer[20]

Writing Group: Gonçalo R. Abecasis[5], David Altshuler[2-4], Adam Auton[11], Lisa D. Brooks[75], Richard M. Durbin[1], Richard A. Gibbs[14], Matt E. Hurles[1], Gil A. McVean[11,18]

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.
2 The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
3 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
4 Dept of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115 , USA.
5 Center for Statistical Genetics and Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
6 Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.
7 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
8 Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA.
9 US National Institutes of Health, 1 Center Drive, Bethesda, Maryland 20892, USA.
10 Life Technologies, Foster City, California 94404, USA.
11 Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.
12 Pall Corporation, 25 Harbor Park Drive, Port Washington, New York 11050 USA.
13 European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK.
14 Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.
15 Centre of Genomics and Policy, McGill University, Montréal, Québec H3A 1A4, Canada.
16 Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany.
17 The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA.
18 Dept of Statistics, University of Oxford, Oxford OX1 3TG, UK.
19 Dept of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.
20 Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
21 US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA.
22 BGI-Shenzhen, Shenzhen 518083, China.
23 Dept of Biology, University of Copenhagen, Denmark.
24 Life Technologies, Beverly, Massachussetts 01915, USA.
25 Deep Sequencing Group, Biotechnology Center TU Dresden,Tatzberg 47/49, 01307, Dresden, Germany.
26 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany.
27 Roche Applied Science, 20 Commercial Street, Branford, Connecticut 06405, USA.
28 Department of Medical Genetics, Institute of Molecular Medicine (FIMM) of the University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
29 Agilent Technologies Inc., Santa Clara, California 95051, USA.
30 Dept of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA.
31 US National Institutes of Health, National Institute of Environmental Health Sciences, 111 T W Alexander Drive, Research Triangle Park, North Carolina 27709, USA.
32 Dept of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA.
33 Illumina, San Diego, California 92121, USA.
34 Dept of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.
35 Dept of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA.
36 Center for Systems Biology, Dept Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachussetts 02138, USA.
37 Dept of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
38 Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.

39 Depts of Psychiatry and Cellular and Molecular Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, California 92093, USA.
40 Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA.
41 Dept of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
42 Dept of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA.
43 Dept of Genetics, Stanford University, Stanford, California 94305, USA.
44 Dept of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA.
45 European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany.
46 Molecular Epidemiology Section, Medical Statistics and Bioinformatics, Leiden University Medical Center, 2333 ZA, The Netherlands.
47 Dept of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA.
48 The Translational Genomics Research Institute, 445 N Fifth Street, Phoenix, Arizona 85004, USA.
49 Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA.
50 Dept of Human Genetics and Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois 60637, USA.
51 Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA.
52 Dept of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
53 National Heart and Lung Institute, Imperial College London, London SW7 2, UK.
54 Centre Nationale de Génotypage, Evry, France.
55 Depts of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
56 Dept of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.
57 Dept of Pediatrics, Faculty of Medicine, University of Montréal, Ste. Justine Hospital Research Centre, Montréal, Québec H3T 1C5, Canada.
58 Dept of Medicine, Centre Hospitalier de l'Université de Montréal Research Center, Université de Montréal, Montréal, Québec H2L 2W5, Canada.
59 Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.
60 Dept of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington 98195, USA.
61 Dept of Computer Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.
62 Dept of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, UK.
63 Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.
64 Center for Genome Science, Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122-701, Korea.
65 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
66 Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.
67 Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
68 Dept of Psychiatry and Behavioral Studies, Stanford University, Stanford, California 94305, USA.
69 Coriell Institute, 403 Haddon Avenue, Camden, New Jersey 08103, USA.
70 Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK.
71 Genetic Alliance, 436 Essex Road, London, N1 3QP, UK.

72 Center for Medical Ethics and Health Policy, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.
73 Dept of Medical History and Bioethics, University of Wisconsin--Madison, Madison, Wisconsin 53706, USA.
74 US National Institutes of Health, Center for Research on Genomics and Global Health, 12 South Drive, Bethesda, Maryland 20892, USA.
75 US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA.
76 The George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA.
77 US Food and Drug Administration, 11400 Rockville Pike, Rockville, Maryland 20857, USA.
78 US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.
79 Dept of Ecology and Evolutionary Biology, Rice University, Houston, TX 77251, USA

* Deceased