A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree

Supplementary Information

Content

1 Identifying and Assessing Pedigree Consistent Variants

1.1 Informatics pipelines used in this study

We used six different variant calling pipelines and two different sequencing datasets for this study (Table S1). The ILMN PCR-free sequence data was aligned with two different sequence aligners and four different variant callers in addition to assembly-based calls from Cortex (Iqbal et al. 2012). In addition, the SNV calls generated on Complete Genomics sequence data (Drmanac et al. 2010) were included in this study. Within this study, each informatics pipeline was weighted equally for incorporation into our final platinum call set.

Table S1. Sequence data and variant calling pipelines used for this study.

Aligner	Variant Caller	Sequence Data	SNVs	Indels
bwa-mem	GATK3	ILMN PCR-free (2x100bp)	Yes	Yes
bwa-mem	FreeBayes	ILMN PCR-free (2x100bp)	Yes	Yes
bwa-mem	Platypus	ILMN PCR-free (2x100bp)	Yes	Yes
Isaac	Strelka	ILMN PCR-free (2x100bp)	Yes	Yes
NA	Cortex	ILMN PCR-free (2x100bp)	Yes	Yes
CGTools 2.0	CGTools 2.0	CGI (2x50bp)	Yes	No

1.2 Identifying the inheritance vectors for the pedigree

We used the SNV calls from the GATK3 pipeline to generate the inheritance vectors. These variants were filtered to remove sites that showed Mendelian inconsistencies or had missing data. We then used the linkage software package Merlin (Abecasis et al. 2002) to identify the inheritance vectors for the autosomes and Chromosome X in the parents and children of this pedigree. The initial inheritance vectors showed many more crossover events (>8,000) than are expected likely due to either genotyping errors or copy number variations. To correct for these errors, we applied a series of heuristics to merge some of the blocks (e.g. if there are two large inheritance blocks that show the same inheritance of the

parental haplotypes separated by 10 SNVs that show completely different inheritance then we merged the larger blocks together). Additionally, to address gaps expected between the defined inheritance vectors we extended our inheritance regions where possible to maximize the genomic coverage of our inheritance vectors. To minimmize the gaps between inheritance blocks, we utilized the fact that multiple children provided technical replicates, based on the observation that some children inherit the same two chromosomes from their parents. For example, Table S2 illustrates a hypothetical case where a crossover occurs in child 3 somewhere between 3000 and 4000 bases on Chromosome 1. Before this crossover, child 3 inherited the same two haplotypes as child 5 but after this crossover child 3 inherited the same two haplotypes as child 4. We can use the similarity in SNV calls between child 3, 4 and 5 to refine the region of the crossover. For example, if there is a SNV at position 3,500 where both child 3 and 5 have the same genotype as each other but both have different genotypes from child 4 then we can extend the second inheritance vector from 2,001-3,000 to 2,001-3,500.

Table S2. Hypothetical example of inheritance vectors in a region of the genome where the crossover is not well defined. In this example the inheritance vectors define the entire genome except between the 2nd and 3rd rows where the known inheritance blocks are separated by 1000 bp (a crossover occurs somewhere between 3,000 to 4,000 bp). The fact that child 3 changes from having the same two haplotypes as child 5 to the same two haplotypes as child 4 may allow us to close the gap between these inheritance vectors.

Chromosome	Start	Stop	Father	Mother	Child 1	Child 2	Child 3	Child 4	Child 5
Chr1	1	2,000	AB	CD	AC	AD	BC	BD	BC
Chr1	2,001	3,000	AB	CD	AC	AC	BC	BD	BC
Chr1	4,000	5,000	AB	CD	AC	AC	BD	BD	BC
Chr1	5,001	6,000	AB	CD	AD	AD	BD	BD	BC

In addition to identifying and merging the inheritance vectors, we used the founder genotype calls to label the haplotypes consistently. The final inheritance haplotypes are labeled A, B, C & D, where: A & B are the haplotypes that the father (NA12877) inherited from NA12889 (A) and NA12890 (B) and C & D are the haplotypes that the mother (NA12878) inherited from NA12891 (C) and NA12892 (D). For Chromosome X, everything is the same except the father only has haplotype B. Figure S1 shows the number of crossovers in the autosomes for the paternally and maternally derived haplotypes.

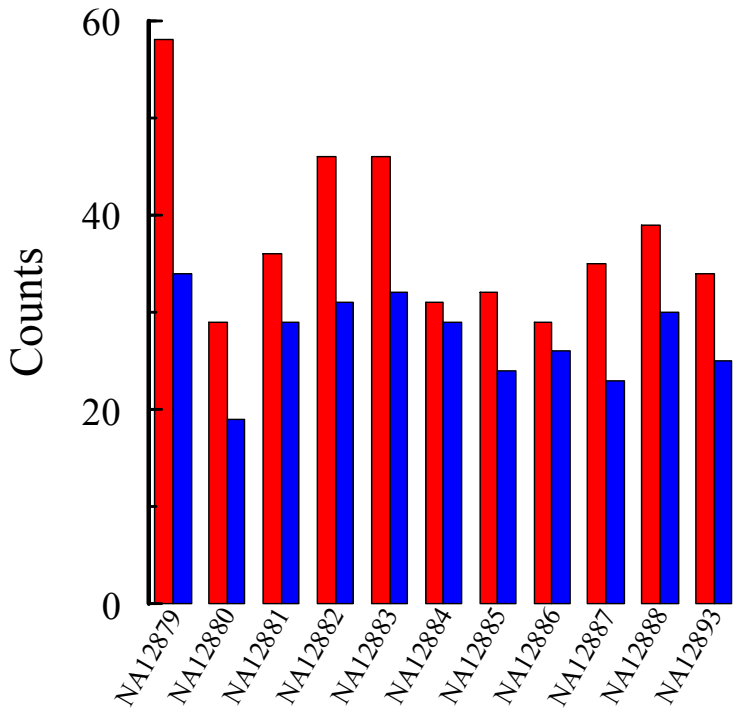

Figure S1. Crossovers identified in each of the eleven children based on our final inheritance vectors in the autosomes. Red bars show the number of crossovers in the maternally inherited haplotypes (inherited from NA12878) and the blue bars show the number crossovers in the paternally inherited haplotypes (inherited from NA12877).

1.3 Identifying platinum variants

Traditional trio analysis provides a good way to confirm the variants specifically within a child because all of the variants in the child must also occur in at least one of the parents (excluding a small number of de novo mutations). In this study however, the aim was to develop high quality truth data that included establishing the correct genotype call in all individuals. This requires that both the alleles and genotypes be correct. A trio analysis is not sufficient to confirm the genotype calls because, for example, of the nine possible (unphased) genotype combinations in the parents, five of the combinations may not produce a Mendel error even if the child is incorrectly genotyped (Table S3). The only time that the child's genotype is unambiguously defined based on the parents is when both parents are homozygous. Excluding the homozygous reference locations this accounts for just ~24% of the SNV positions in a test trio (NA12877-NA12878-NA12882). The problem of possible genotype errors going undetected is even bigger for the parents where there are always at least two possible genotypes that could be called in the one parent without producing a Mendel error (Table S4).

Table S3. Possible genotypes in a child based on the genotypes in the parents.

Father	Mother	Possible GTs in Child
0/0	0/0	0/0
0/0	0/1	0/0 or 0/1
0/0	1/1	0/1
0/1	0/0	0/0 or 0/1
0/1	0/1	0/0 or 0/1 or 1/1
0/1	1/1	0/1 or 1/1
1/1	0/0	0/1
1/1	0/1	0/1 or 1/1
1/1	1/1	1/1

Table S4. Possible genotype combinations in a parent based on the genotypes in the other parent and child.

Father	Possible GTs in Mother	Child
0/0	0/0 or 0/1	0/0
0/0	0/1 or 1/1	0/1
0/0	0/1 or 1/1	1/1
0/1	0/0 or 0/1	0/0
0/1	0/0 or 0/1 or 1/1	0/1
0/1	0/1 or 1/1	1/1
1/1	0/0 or 0/1	0/0
1/1	0/0 or 0/1	0/1
1/1	0/1 or 1/1	1/1

Compared to a trio analysis, when many children are sequenced and the haplotype inheritance identified, the genotypes of the children are completely constrained by the phased parental alleles. At bi-allelic sites this means that there are 14 ($=2^4-2$) possible genotype combinations in the children in addition to the two mono-allelic combinations where every individual is homozygous for the reference allele or every individual is homozygous for the derived allele. Table S5 shows all of the possible genotype combinations based on the phased parental haplotypes within any autosomal chromosome. In Table S5 the father has haplotypes labeled A & B, the mother has the haplotypes labeled C & D, and each of the children inherited a different haplotype combination (AC, AD, BC & BD). Once we have enough children so that all four possible haplotype pairings are represented in the children, sequencing additional siblings provides replicate information that gives additional confidence in the genotype calls.

Table S5. Possible genotype combinations when haplotype inheritance is considered.

Haplotypes				Genotypes					
				Father	Mother	Child 1	Child 2	Child 3	Child 4
A	B	C	D	AB	CD	AC	AD	BC	BD
0	0	0	0	0/0	0/0	0/0	0/0	0/0	0/0
0	0	0	1	0/0	0/1	0/0	0/1	0/0	0/1
0	0	1	0	0/0	0/1	0/1	0/0	0/1	0/0
0	1	0	0	0/1	0/0	0/0	0/0	0/1	0/1
1	0	0	0	0/1	0/0	0/1	0/1	0/0	0/0
0	0	1	1	0/0	1/1	0/1	0/1	0/1	0/1
0	1	0	1	0/1	0/1	0/0	0/1	0/1	1/1
1	0	0	1	0/1	0/1	0/1	1/1	0/0	0/1
0	1	1	0	0/1	0/1	0/1	0/0	1/1	0/1
1	0	1	0	0/1	0/1	1/1	0/1	0/1	0/0
1	1	0	0	1/1	0/0	0/1	0/1	0/1	0/1
0	1	1	1	0/1	1/1	0/1	0/1	1/1	1/1
1	1	0	1	1/1	0/1	0/1	1/1	0/1	1/1
1	0	1	1	0/1	1/1	1/1	1/1	0/1	0/1
1	1	1	0	1/1	0/1	1/1	0/1	1/1	0/1
1	1	1	1	1/1	1/1	1/1	1/1	1/1	1/1

In general, many children will need to be included so that at every position in the genome the four haplotype pairings are represented in the children. The fraction of the genome where both parental haplotypes are inherited by at least one child increases as more children are sequenced. When at least two children are sequenced, the probability that each of the parental haplotypes occurs in at least one child is $1 – (1/2)^{n-1}$ where n is the number of children sequenced. The probability that this is true for both parents is the square of the above equation. Using this calculation, we estimated the fraction of the genome where we can detect any genotype errors assuming a diploid genome (Table S6). For this pedigree, the fraction of the phased genome where both parental haplotypes are observed within any child as more children are added to the pedigree agrees with the theoretical prediction.

Table S6. Sensitivity to detect a single genotype error in any member of the family based on the number of offspring included. Here, we assumed no more than a single genotype error per variant position, an assumption that is likely for most positions in the genome. Note, however, that haplotype phasing in large pedigrees also has power to resolve ambiguity in many positions where multiple genotyping errors occur e.g. when eleven children and their parents are sequenced we can identify up to three genotyping errors in over 76% of the genome.

Offspring	Predicted	Observed*
1	0.000	0.000
2	0.250	0.296

3	0.562	0.585
4	0.766	0.781
5	0.879	0.860
6	0.938	0.903
7	0.969	0.941
8	0.984	0.961
9	0.992	0.995
10	0.996	0.999
11	0.998	>0.999

*Represents fraction of phased haplotypes

To identify the pedigree-consistent variants for this study, we generated all of the possible genotype combinations possible (illustrated in Table S5) within each region where we had identified the inheritance vectors. We then compared the genotypes from a single sequencing pipeline for consistency. If the genotypes agreed with exactly one of the predicted genotype combinations then we defined that variant location as accurate for that sequencing pipeline. For example, if the father was homozygous for the reference allele and the mother was heterozygous (represented in the 2nd and 3rd rows in Table S5) then this variant was labeled correct if either of the following conditions were met: (a) every child was homozygous for the reference allele except the children with the AC haplotype and these children were heterozygous; or (b) every child was homozygous for the reference allele except the children that inherited the AD haplotype pairs and these children were all heterozygous. If more than one of the possible genotype combinations agrees exactly with the observed genotypes then the phasing could not be uniquely determined at that position and the variant was excluded. This can only occur if only two of the four possible haplotype pairings are observed in the sequenced children. For example, if all of the children sequenced inherited either: (a) haplotype A from the father and haplotype D from the mother or (b) haplotype B from the father and haplotype C from the mother then we could not phase the locations when all of the members are genotyped as heterozygous. These positions are particularly important to exclude because one of the main failure modes we identified occurs when all the members of the pedigree are called heterozygous, which may be caused by homologous sequence. Additionally, we extended this basic method to also assess loci where three and four alleles are segregating within this pedigree.

Once we identified the variants that are pedigree consistent within each of the six pipelines shown in Table S1, we merged the call sets together to create our final “platinum” variant catalogue. The merged call set contained all of the individual single-pipeline pedigree consistent variants with the following exceptions:

1. If two pipelines identified the same variant as pedigree-consistent but had different genotypes in the parents or children, then we could not resolve the location and exclude these variants.
2. If two pipelines identified the same position as pedigree consistent but with different alleles then we could not resolve that position and exclude these variants from further consideration.

3. If the variant was called homozygous for the alternative allele in the entire pedigree (last row in Table S5), then we ensured that this was not due to systematic biases in the alignment by requiring the variant to be pedigree-consistent in callsets based on at least two aligners (any two of Isaac, bwa or cgi). This same rule was applied when we identified our confident homozygous reference positions.
4. Each variant that did not pass an additional flanking-sequence (*k*-mer) test (cutoff value of 1) was excluded from our final call set. See below for a description of this test.

In addition to the variant positions, we also collated our high confidence invariant positions using the same rules that were applied to positions that were genotyped as homozygous alternative across the pedigree. In this case the position must be called homozygous reference using at least two different call sets based off of different sequence aligners. To further eliminate any possible missed variants in our confident homozygous reference positions, we removed all positions where variant calls were made in any of the samples by any of the sequencing pipelines including variant calls that did not pass the quality filters. In total, this analysis identified 2,737,246,156 bases that we defined as confident homozygous reference across the pedigree and these positions were used to assess false positive rates of variant calling pipelines.

1.4 k-mer test of pedigree-consistent variants

In addition to our checks for pedigree consistency, we also performed a test of the flanking sequence to exclude false positives, incorrect alleles and duplicates from our final call set as far as possible. First we identified the haplotype sequence context (*k*-mer) centered on the variant extending a total of 51bp. In the case of a simple SNV, this would equate to two 51-mers, both containing the 25bp before and after the SNV, and each containing one of the two alleles represented by the SNV. A hypothetical example of this analysis is shown in Figure S2, where there are 24 reads that overlap the SNV, only 15 of these overlap with enough flanking sequence to pass our test. Of these 15 reads, four contain base errors leaving only 11 *k*-mer validated reads (six that confirm the reference allele and five that confirm the alternate allele).

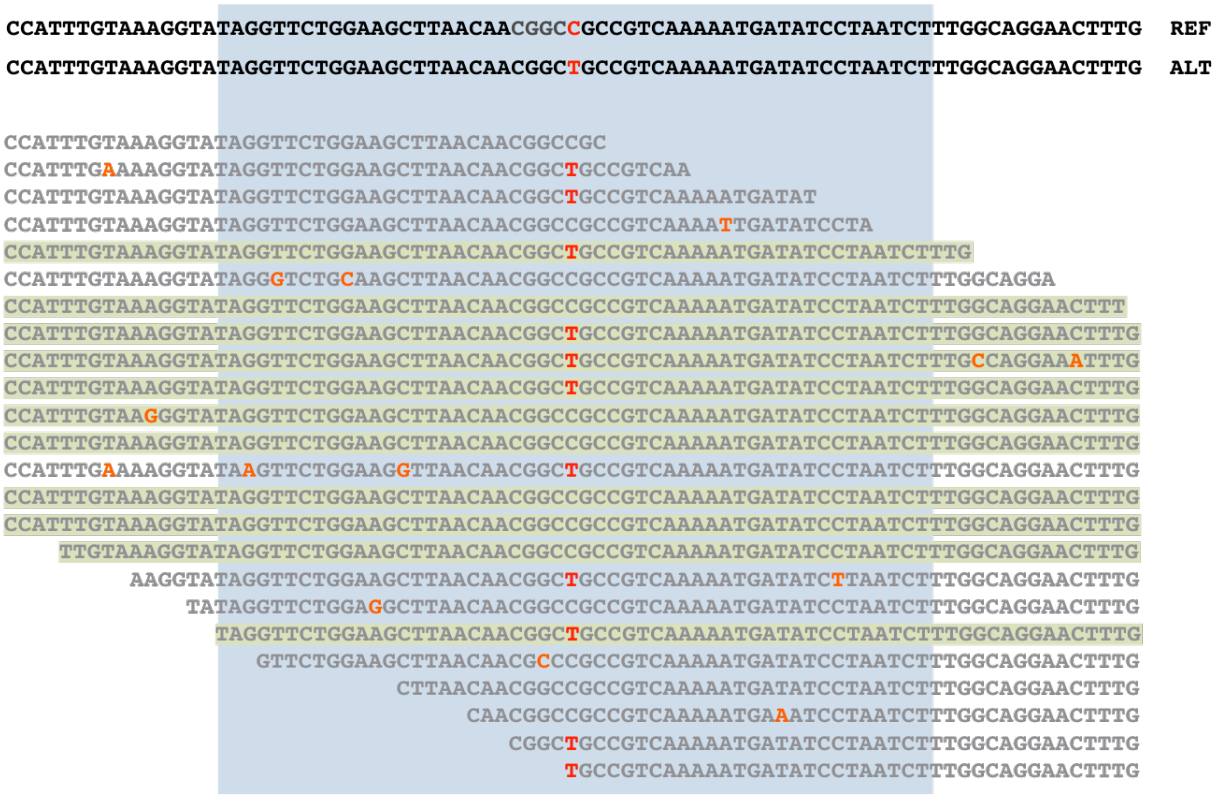

Figure S2. Example of *k*-mer testing to validate the variants and the flanking sequence for a SNV In this example an C->T SNV has been identified and we are testing this SNV +/- 25bp of flanking sequence highlighted by the blue box. All reads that do not contain the full flanking sequence or differ from the predicted sequences do not pass this analysis. Green shading highlights the 11 reads that exactly match the variant and the flanking sequence.

In the case where there are nearby variants, the flanking sequence test was modified to be consistent with these additional non-reference positions (e.g. Figure S3). Importantly, where there are clusters of nearby variants, problems with normalization will usually cause both variants to fail the flanking-sequence test. In a hypothetical example illustrated in Figure S3, a deletion in the run of "A"s was represented in two ways: (a) as a deletion of the first base (CA->C ; shown as ALT 1) and (b) as a deletion of the last base (AT->T ; shown as ALT 2). If both of these separate representations of the same variant pass the pedigree check, then this single deletion may appear as two separate, non-overlapping deletions. When we do the *k*-mer check for either of these, we account for the nearby variants and combined this indicates that two bases are deleted (CAAAAAT -> CAAAT ; shown as ALT1+2). Because only one base is truly deleted the *k*-mer for the reference allele is validated (six reads contain this *k*-mer) but the *k*-mer for the alternate allele fails and overall both of these variants will be excluded. Future work will involve normalizing these cases so that a single, correct representation can be identified, resulting in a more complete dataset.

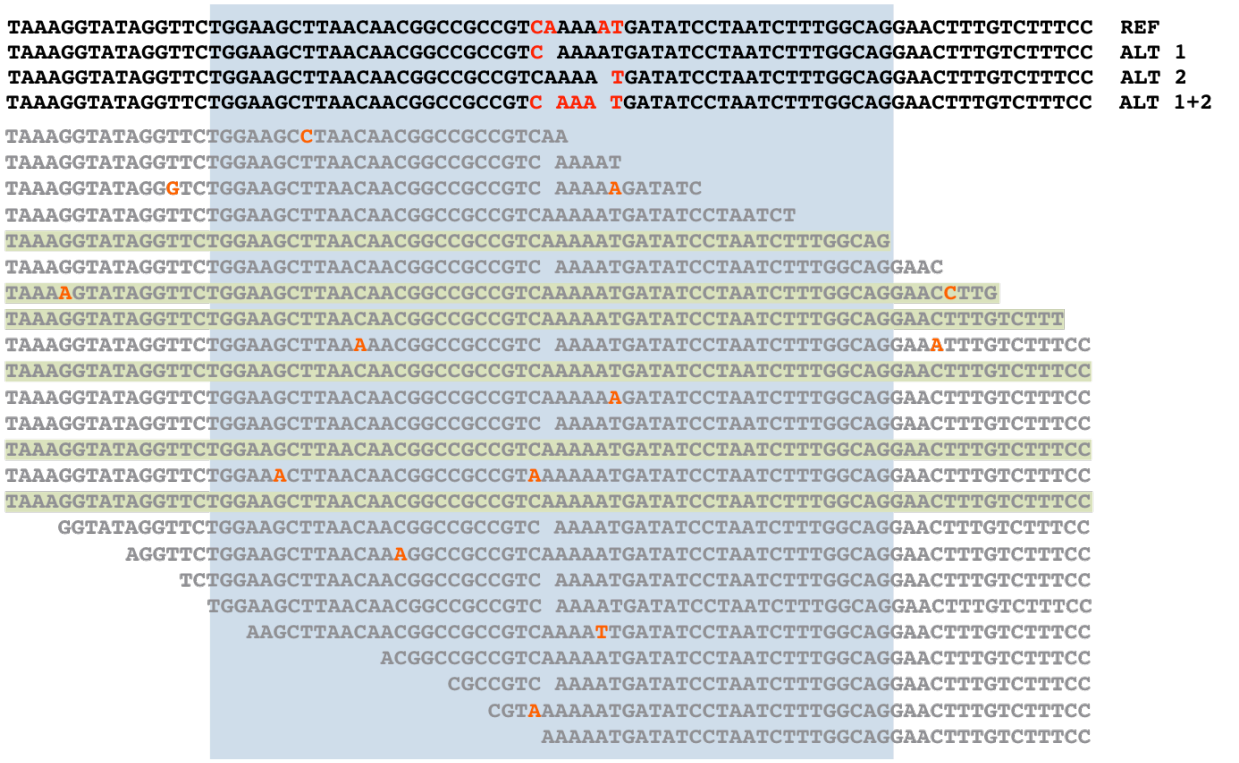

Figure S3. Example of *k*-mer testing to validate the variants and the flanking sequence when there are nearby variants. In this case the *k*-mer test incorporates these variants into the definitions of the *k*-mers. In this example there is one deletion (CA->C) on the left side of the homopolymer (ALT 1). At the same time one of the analysis pipelines has failed to properly left-align the deletion and thus it is called at the right side of the homopolymer (ALT 2 ; AT->T). The resulting non-reference *k*-mer includes both deletions (ALT 1+2). Green shading highlights the six reads that exactly match the reference sequence but no reads match the predicted alternative sequence (ALT 1+2).

For this study, we counted the number of times each of the predicted 51-mers were observed per family member using all reads that aligned within 400bp of the variant based on the bwa-mem alignments. From this we calculated a single whole-pedigree normalized value for each variant. This normalized value represents the average count per predicted haplotype (i.e. total counts divided by the number of sample-haplotypes predicted to have the associated *k*-mer). For example, in the simple SNV case, if the alternate allele occurred in the mother and was passed on to four of the children then the normalized *k*-mer count would be the total alternate-allele 51-mer counts dividing by five (i.e. mother and four children have the alternate allele/haplotype). Likewise, the normalized reference-allele *k*-mer would be the total counts of the reference *k*-mer in the pedigree divided by 21 (i.e. 26 total haplotypes in the pedigree – 5 with the alternate allele). Table S7 shows the number of passing variants based on a minimal number of normalized *k*-mer counts. Because we are only filtering based on the average *k*-mer counts across the pedigree, there may be instances where individual samples do not have any of the predicted 51-mers observed. We also catalogued the number of variants where each *k*-mer genotype was not represented in any of the samples. To test the *k*-mer genotypes we incorporate the ploidy so that to pass our *k*-mer genotype test a heterozygous site needs at least one observation of the reference

51-mer and one occurrence of the alternative 51-mer and a homozygous alternative call needs to have two copies of the alternative 51-mer. Since we only know one of the haplotypes in the founders, we only test for the presence of the inherited *k*-mer in our founder genotype analysis.

Table S7. *K*-mer test of the pedigree-consistent variants using different filtering criteria (*k*-mer cutoff)

		Parents and Children			Founders		
k-mer cutoff	Passing Variants	Failed[1]	Passing[1]	% Failed	Failed[2]	Passing[2]	% Failed
0	5,621,476	2,000,796	73,079,188	2.66	427,869	22,355,514	1.88
1	5,426,236	154,736	70,541,068	0.22	33,132	21,578,102	0.15
2	5,406,318	93,228	70,282,134	0.13	22,746	21,498,700	0.11
3	5,376,937	58,586	69,900,181	0.08	16,596	21,381,649	0.08
4	5,334,472	37,404	69,348,136	0.05	12,294	21,212,562	0.06
5	5,280,210	24,123	68,642,730	0.04	9,203	20,996,673	0.04
6	5,214,385	15,496	67,787,005	0.02	6,858	20,734,821	0.03
7	5,129,161	10,086	66,679,093	0.02	5,083	20,396,064	0.02
8	5,002,685	6,466	65,034,905	0.01	3,756	19,893,326	0.02
9	4,777,927	4,069	62,113,051	0.01	2,702	19,000,344	0.01
10	4,353,650	2,510	56,597,450	0.00	1,886	17,314,366	0.01

[1]Number of sample-variant combinations that pass/fail the *k*-mer GT test in the parents and children
[2]Number of sample-variant combinations that pass/fail the *k*-mer test in the founders

Using a *k*-mer cutoff value of one, there are 154,736 total sample genotype failures (Table S7) within the parents and children at 68,866 variant locations (50,154 + 19,712 from Table S8). It should be noted that this *k*-mer test is very stringent for two reasons: (a) the effective depth is roughly halved because each 100bp read reduces to 49 possible 51-mers and (b) the effective error rate is 51 times higher because any error in the 51-mer will cause it to fail this exact test. For example, a completely random error model with 99.9% base calling accuracy (Q30) would result in an error rate of 5% per 51-mer ($1-0.999^{51}$). This elevated effective error rate will be more pronounced in problematic regions of the genome such as around homopolymers where the raw sequencing base-level error rate is systematically higher. Because this *k*-mer test is so stringent, we do not expect every sample to pass the *k*-mer genotype test (even for true variants) and we do not remove sites that just fail this test in individual samples.

Table S8. Variants that fail a *k*-mer GT test in any sample based on different k-mer cutoffs.

k-mer cutoff	Total Variants	Variants with *k*-mer GT failures		
		Just Pedigree	Just Founders	Both
0	5,621,476	55,423	8,659	209,683
1	5,426,236	50,154	8,659	19,712
2	5,406,318	39,259	8,411	11,866
3	5,376,937	25,203	7,197	7,816
4	5,334,472	15,143	5,785	5,544
5	5,280,210	9,425	4,652	3,981
6	5,214,385	6,188	3,845	2,703
7	5,129,161	4,195	3,162	1,763
8	5,002,685	2,883	2,564	1,104
9	4,777,927	1,915	1,975	662
10	4,353,650	1,178	1,435	407

For this study we chose a minimum, normalized *k*-mer cutoff of 1 such that all variants with at least one 51-mer per predicted haplotype are included in this study. Many of these *k*-mer filtered variants are likely caused by either incompleteness in the variants such that the surrounding reference sequence is not correct, regions of high error rates, or conflicting way different pipelines may represent the same variant such that the combination of the calls are not a correct representation of the region (e.g. Figure S3). For example, just 12.8% of the final platinum variants occur within 25 bp of another platinum variant while 55.9% of the *k*-mer filtered variants occur within 25 bp of another *k*-mer filtered variant. Additionally, we find that compared to our *k*-mer filtered variants, our platinum variants primarily fall within mostly high-confidence (platinum) regions indicating that the *k*-mer filtered variants are more likely to occur near other non-platinum variants or near regions of high error rates (Figure S4). Future work will attempt to resolve these conflicts to create a more complete catalogue.

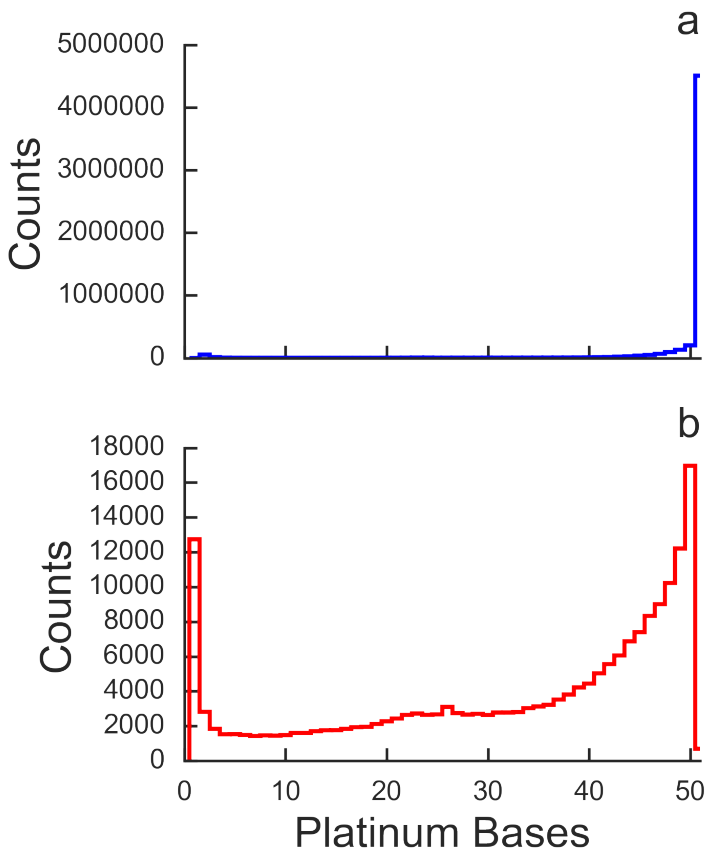

Figure S4. Number of the flanking bases that are part of our final platinum regions for the platinum (a) and *k*-mer filtered (b) variants. Flanking bases represent the region covered by the 51-mers and by definition a *k*-mer filtered variant can have, at most, 50 platinum bases.

1.5 Indel properties

In addition to het:hom ratios we also calculated the size distribution for all the indels and just the ones in coding sequence (Figure S5). Comparing the ins:del ratio with the frequencies derived for these indels in the 1000 Genomes European samples, we observed that the rare indels (between 0% and 5% in the 1000 Genomes European samples) had a ins:del ratio of ~0.5 though the common indels and those not in the 1000 Genomes European samples had and ins:del ration of ~1 (Table S9). Because, the indels that are not in the 1000 Genomes data are predicted to be rare, it is likely that the lower ratio for the "rare" indels that are part of the 1000 Genomes data is due to undercalling rare insertions. Finally, we observed that the indels in this study were more likely to be multi-allelic compared with the SNVs. To examine whether this may be because indels are more likely to occur in regions of high mutation rates such as di-nucleotide repeats we overlapped our indels with STRs (defined here as locations in the reference where motifs of length 2-6bp are repeated at least twice). We found that indels are more twice as likely to occur within STRs and multi-allelic indels are four times as likely to occur within STRs.

Overall, 29.9% of our platinum indels overlap STRs compared with 60.9% of the multi-allelic indels. In contrast, 14.7% of our platinum SNVs overlap STRs compared with 14.8% of the multi-allelic SNVs.

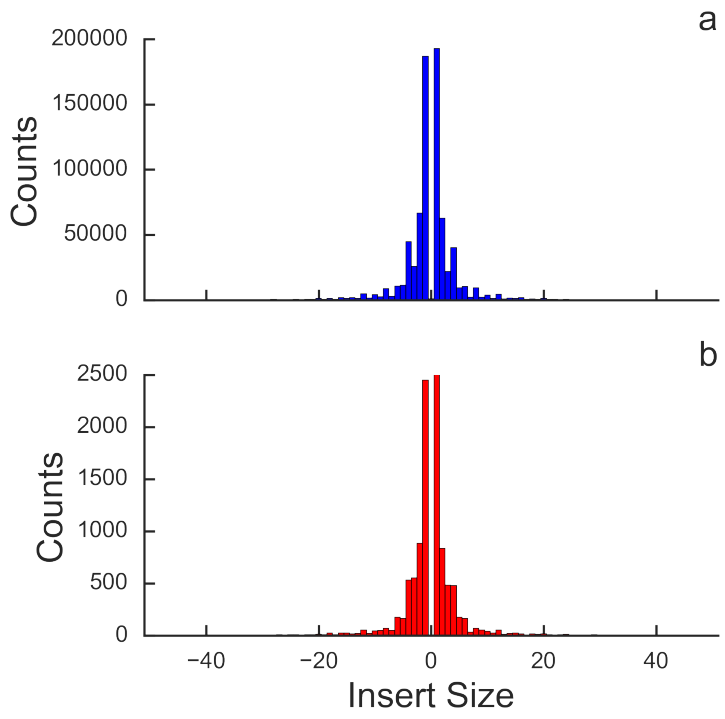

Figure S5. Insert sizes for all of the platinum indels. Panels show the distributions for all of the indels (a) and the indels in coding sequence (b).

Table S9. Ratio of insertions to deletions for the platinum indels versus frequency in the 1000 Genomes European samples.

1kGP Frequency (EUR)	Ins:del	Number*
0%	1.09	249,066
>0% and <5%	0.56	31,735
>5%	0.94	477,121

*Number indicates the number of alleles and not the number of indel positions

1.6 Observed and theoretical het:hom

Summary statistics based on a common reference (hg19) provide a convenient way to compare variant calls from one dataset to other datasets though a better way to assess a dataset is to make a comparison against a theoretical prediction. For example, under the standard neutral model of population genetics, the het:hom ratio will be 2:1. This theoretical prediction is not commonly observed in most samples in part because the reference genome is comprised of sequence taken from samples of different ethnicity violating the panmictic population assumption of the standard neutral model. When

the reference sequence comes from the same ancestry as the sample analyzed, the randomly-mating assumption is a good approximation and we expect that the het:hom ratios will be closer to the theoretical prediction. Because every member of this family is of European ancestry we used each of the four haplotypes that we have defined in this study as a reference genome and recalculated the het:hom ratio in the samples that do not carry that haplotype. For this calculation, we also excluded variants that show evidence of more than one non-reference allele because the theoretical predicted ratio of 2:1 is only valid in the absence of recurrent mutations (i.e. following the infinite sites model). Using the unrelated haplotypes to define the reference alleles the het:hom ratios for all variant types were close to the expected 2:1 ratio (Table S10). While there are differences between the het:hom ratio for each sample, possibly due to population structure, the population-normalized indel and SNV het:hom ratios within a sample are very similar. Qualitatively, this indicates that the indel calls identified here are of similar quality to the SNV calls.

Table S10. Ratio of heterozygous to homozygous variants based on different reference haplotypes

	NA12877			NA12878		
	hg19[1]	NA12891[2]	NA12892[2]	hg19[1]	NA12889[2]	NA12890[2]
SNVs	1.64	2.11	2.09	1.56	1.94	1.94
Indels	1.82	2.09	2.08	1.73	1.93	1.92

[1]Reference allele is defined from hg19
[2]Reference allele is defined from the unrelated founder haplotypes

1.7 Extent of platinum coverage

Assessing the platinum coverage in the autosomes separately from Chromosome X shows that a significantly higher fraction of autosomes has platinum coverage (Table S11). Much of this is likely due to the lower average depth in males, which inevitably leads to a reduction in read depth that compromises variant calling in places. It is likely that the coverage would be significantly better if the variant callers were modified to treat haploid chromosomes differently from diploid chromosomes. See Table S12 for a breakdown of the platinum coverage by gene.

Table S11. Platinum coverage of the autosomes and Chromosome X for different categories of genomic context based on UCSC (hg19).

Category	Autosomes	Chromosome X
hg19	96.96	92.47
Genes[1]	97.51	93.57
Exons[1]	98.13	94.64
ACMG genes	98.78	97.19
ACMG exons	99.49	98.96
LINE	94.66	87.12
SINE	97.16	94.53
LTR	98.09	96.76
DNA	98.87	97.57
Simple repeat	71.77	41.26
Low complexity	89.23	68.65
Satellite	71.59	60.31
Merged Other[3]	72.41	56.73
All Repeats[4]	95.51	89.48
Non Repeats	98.41	97.07

[1]Breakdown of coverage by gene available in supplemental methods

[2]Certain repeats in the genome may be represented in more than one category

[3]"Merged Other" includes a non-redundant merge of categories listed in RepeatMasker as RNA, rRNA, scRNA, snRNA, srpRNA, tRNA, Unknown and Other.

[4]"All_Repeats" is calculated from a single non-redundant merge of all repeat categories.

2 Analysis of Pedigree Inconsistent Variants

2.1 Likely Mosaic in NA12889

There were a high number of sites that failed the founder *k*-mer analysis in NA12889 on the distal end of Chromosome 11 (see Figure 2a). As an additional review of this region we utilized the fact that because all of the platinum variants are phased, we can separate all of the heterozygous SNVs in any of the founders into the transmitted (to NA12877 or NA12878 as appropriate) or non-transmitted allele. Doing this for the SNVs in NA12889, we identified that these *k*-mers that failed to validate in the founders occur in a region where the read counts for the transmitted alleles are substantially lower than the non-transmitted alleles (Figure S6a). Specifically, for the 13,000 heterozygous SNVs in NA12889 on the distal end of Chromosome 11, there are, on average of ~11 reads supporting the transmitted allele while there are an average of ~42 reads supporting the non-transmitted allele. By comparison, Figure S6b shows the same analysis for another founder (NA12890) showing the expected 1:1 ratio of the transmitted and non-transmitted alleles.

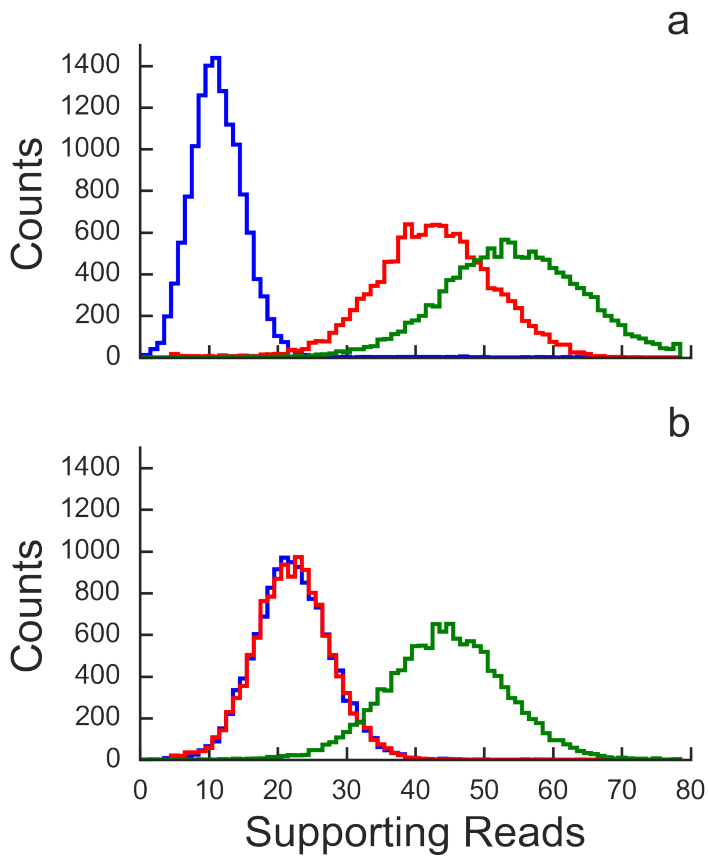

Figure S6. Read depth for the different haplotypes showing the depth distribution for the transmitted haplotype (blue) and the non-transmitted haplotype (red) and the total depth (green). The likely mosaic region in NA12889 (a) and the same region in NA12890 (b).

2.2 High quality pedigree-inconsistent SNVs

We examined the 334,652 positions where at least two pipelines made identical calls in the parents and eleven children but the genotypes were inconsistent with the transmission of the haplotypes ("high-quality" failures). We classified the failure modes of these SNVs into four categories: 1) positions where all individuals are heterozygous for the SNV; 2) positions where the genotypes are consistent with the occurrence of a hemizygous deletion segregating in the pedigree; 3) positions where all individuals are homozygous for the reference allele except for a heterozygous call in one individual; 4) the remaining positions not covered by the first three categories and where the genotypes are not consistent with the inheritance vectors. We analyzed each category to evaluate the possible underlying reasons for failure, by examining properties of the failing SNVs including read depth, read counts of each allele, spacing in

the genome, and extended this analysis outside of this pedigree by identifying deviations from Hardy-Weinberg equilibrium in a population of samples of European ancestry (unpublished data).

Category 1: SNVs in this category represent positions that are heterozygous in all thirteen individuals. This observation is expected where SNVs lie in regions of paralogous sequence, i.e. arising from duplications. We found that the average sequence depth across all these sites was elevated almost three-fold compared to a genome-wide average for the platinum variants (138.3x versus 50.6x, respectively), indicating the predominance of duplications or higher order copy number variants (CNVs) underlying these SNVs (Figure S7). We studied these variants in a cohort of ~2,000 unrelated samples of European descent (unpublished data). In the European cohort, a representative subset (84.0% of all 75,362 with a minor allele frequency >4%) significantly deviated from Hardy-Weinberg Equilibrium (HWE) and 98.9% of the SNVs that deviated from HWE had an excess of heterozygous calls consistent with a duplication in the population. Most of these SNVs (173,629 or 90.9%) also occur within 1kb of another SNV exhibiting the same failure mode, consistent with multiple adjacent SNVs lying within the same duplication. Together these observations indicate that most of these SNVs failed the pedigree inheritance test primarily because they overlap CNVs, and not because of errors in the data or the algorithms used in this analysis.

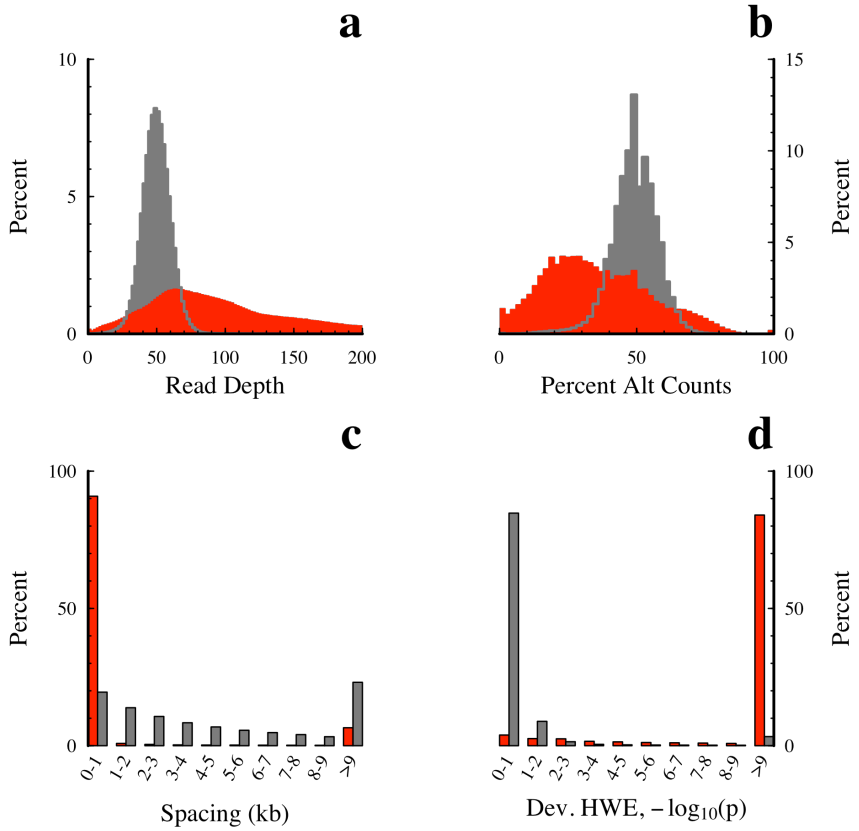

Figure S7. Category 1 SNVs (red) and the pedigree-consistent SNVs (grey). (a) Mean depth in the category 1 failures compared with heterozygous sites in platinum samples. (b) Fraction of reads showing the non-reference allele. (c) Distribution of spacing between category 1 failures compared to a random sampling of the same number of platinum variants. (d) Deviation from HWE (significance based an analysis of ~2,000 samples of European ancestry for the SNVs with MAF>4%).

Category 2: SNV genotypes in this category failed the pedigree validation test but would be consistent with transmission if variants are co-located within a hemizygous deletion that also segregates in the family. Based on this hypothesis, we predicted the haploid or diploid state of all sites in this category and examined the mean sequence depth of each group compared to the Chromosome X read depth in males or females of this pedigree. Our results matched expectations: read depth for the predicted haploid group was close to male Chromosome X read depth (25.3x and 25.7x, respectively) while the mean depth for predicted-diploid sites (50.1x) was close to Chromosome X read depth in the females (50.1x and 48.3 respectively) (Figure S8). The occurrence of duplications underlying most of the SNVs was also consistent with two other observations. First, 83.9% of these SNV sites were clustered within 1kb of one another. Second, many of the SNVs were significantly out of HWE in the European cohort (~72.8% of the 2,918 SNVs with MAF above 2% ; $p < 10^{-9}$) with a marked underrepresentation of heterozygous genotypes.

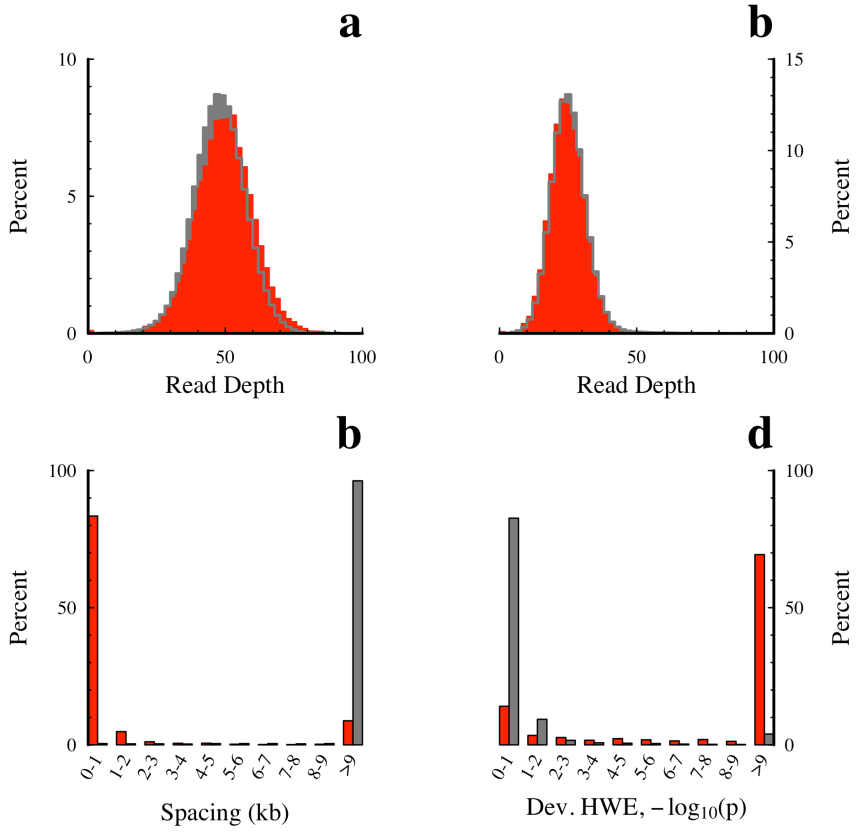

Figure S8. SNVs that fail the pedigree-consistency check because they potentially overlap hemizygous deletions (red) or category 2 failures compared with values from the pedigree-consistent SNVs (grey). (a) Mean depth in the category 2 failures where the sample is predicted to be diploid compared with platinum SNV depths for females on Chromosome X. (b) Mean depth in the category 2 failures where the sample is predicted to be haploid compared with platinum SNV depths for males on Chromosome X. (c) Distribution of spacing between category 2 failures compared to a random sampling of the same number of platinum variants. (d) Deviation from HWE (significance based an analysis of ~2,000 samples of European ancestry for the SNVs with MAF>2%).

Category 3: Another failure mode is characterised by positions where all of the samples are homozygous for the reference allele except for a single sample that is genotyped as heterozygous. These singletons may be either false positive calls or true mutations occurring either in the individual or during culture of the cell line. Compared to the previous two categories the Category 3 variants had normal depth and no clustering though there was an excess of low derived allele frequencies in this category (Figure S9 and S10). We found that 48.8% (24,299) of these singletons were called identically in two independent datasets generated by different sequencing chemistries and analysis pipelines (Complete Genomics and Illumina), which lends support to the hypothesis that many of these anomalies are likely to be true mutations. The depth is roughly the same in the singletons that were validated (51.6x) as the ones that were not validated (48.3x). In the technical replicates the derived allele frequencies were significantly

lower for the unvalidated SNVs (17.6% vs 40.6%). Even though the unvalidated singletons had a lower derived allele frequency, the high depth at these sites means that each singleton SNV was observed in almost 7 reads. These sites include 91.8% (874/952) of the putative cell line mutations previously identified in NA12878 (Conrad et al. 2011).

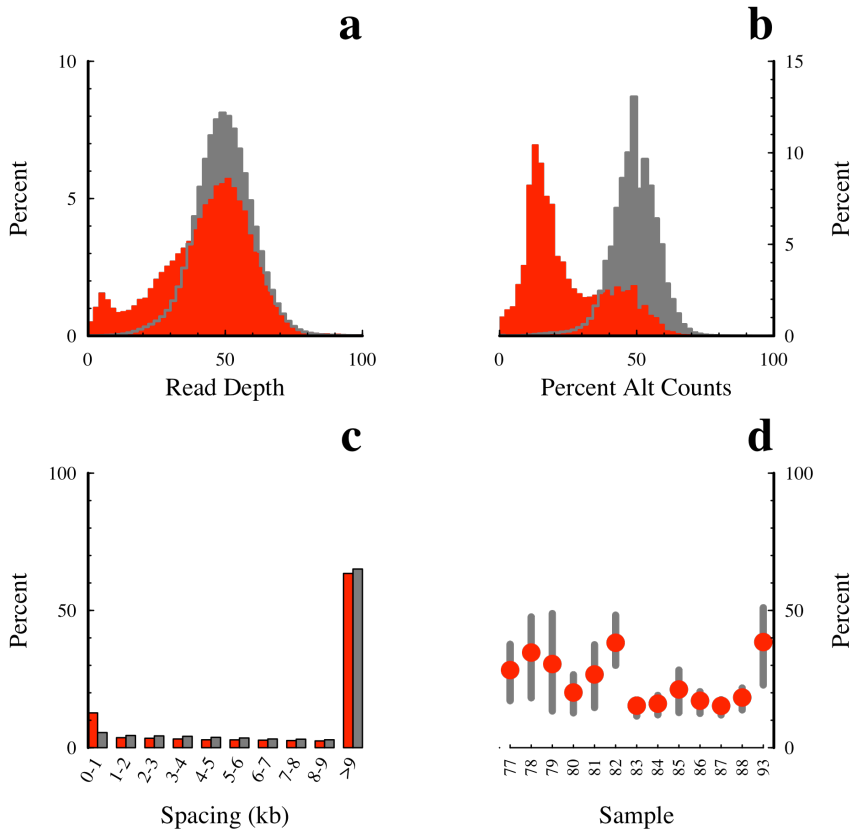

Figure S9. SNVs observed in only a single sample but called by two different software pipelines. (a) Mean depth in the category 3 failures compared with platinum SNVs. (b) Fraction of reads showing the non-reference allele. (c) Distribution of spacing between category 3 failures compared to a random sampling of the same number of platinum variants. (d) Mean value of the fraction of reads showing the non-reference allele represented in (b) according to sample.

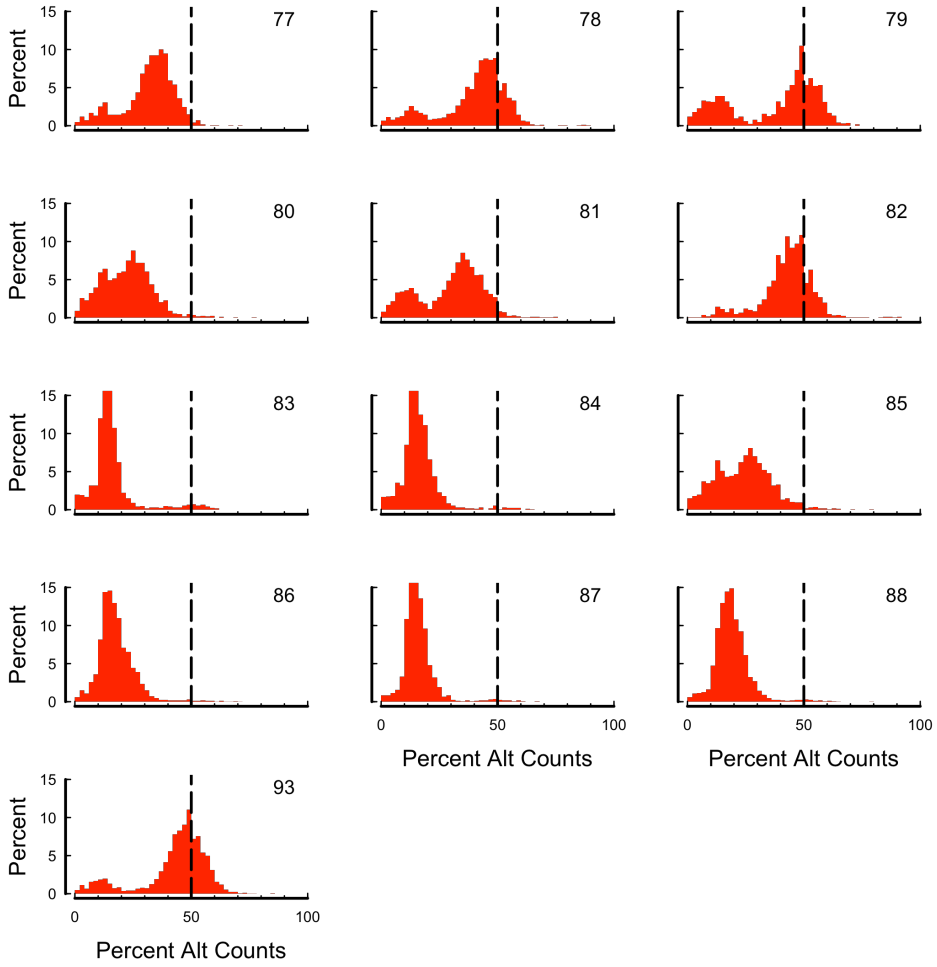

Figure S10. Percentage of reads containing alternate alleles by sample for the category 3 SNVs. As shown in Figure S9, the mean depth at these sites is approximately the same as at the pedigree-consistent locations. Multiple peaks may be representative of different cell line passages and different mean frequencies may represent the differences in the histories of the cell lines between individual samples.

Category 4: The remaining SNVs in this part of the analysis are probably accounted for by a variety of reasons including: genotyping errors, undetected somatic deletions, and undetected germline CNVs. Similar to the failures where every sample is heterozygous, duplications can cause Mendel failures due to incorrect genotypes. When we combine this group of variants with the ones that were heterozygous in every sample (as a combined group possible due to duplications), we identified (~1900) clusters of failed SNVs representing almost 70% of all of the high-quality failured SNVs.

Figure S11 shows the region of the cell line deletion in NA12878 highlighted in the main manuscript from sequence data from four different sources. The black line shows the sequence data used in this study while the other lines show the depth from other sequencing experiments on the same sample. The variable depth in this region indicates that the cell line is heterogeneous, different sample lots have

different amounts of the cell line deletions.

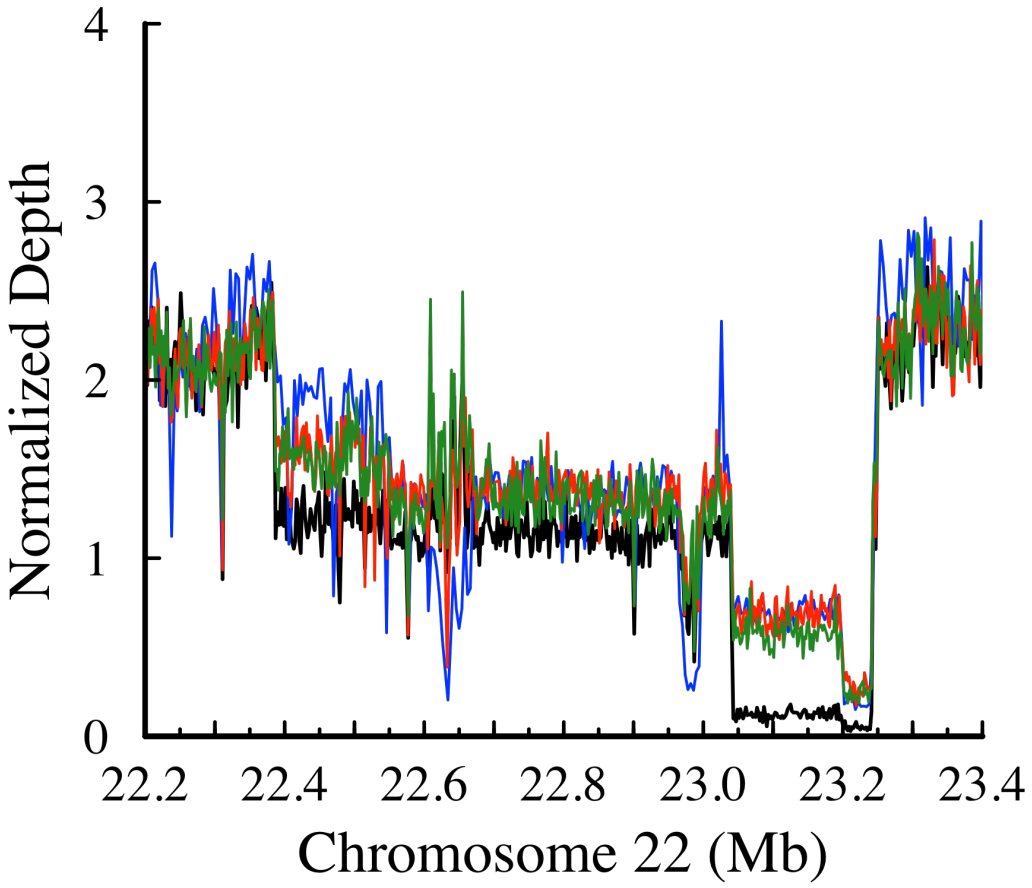

Figure S11. Different sequencing experiments show different percent representation of the somatic deletions observed on Chromosome 22 in sample NA12878. The black line shows the depth based on the sequence data used in this study compared with sequence data from the 1000 Genomes Pilot project (blue), the NIST Genome-in-a-Bottle sample (red) and another run generated within Illumina based on different DNA obtained from Coriell. Overall, the data used in this study shows a significantly higher representation of the somatic deletion(s).

As an additional assessment of the different failure modes and we compared our calls with two separate CNV call sets: 1) a population level database of common duplications and deletions (Sudmant et al. 2015) and 2) CNV calls on this pedigree using Canvas (Roller et al. 2016). For the population-level CNVs

we limited our analysis the CNVs that were predicted to occur in at least 5% of the samples; for the Canvas CNVs we used all calls that were made in at least two samples. When we overlapped these two CNVs call sets with our high quality failures and our platinum variants, we found that the high quality failures (Categories 1-4) are much more likely to overlap the CNVs (Table S13). Both the Category 1 and Category 4 variants are much more likely to overlap the duplications compared to the platinum variants. Likewise, the Category 2 variants are much more likely than the platinum variants to overlap deletions though we also see that the Category 2 variants have a higher rate of overlap with the population-level (Sudmant et al. 2015). Supplementary Table S14 provides the overlap of Category 1-4 SNVs and platinum variants for each of the CNVs analyzed here.

Table S13. Percentage of SNVs that overlap different types of CNVs

			Population CNVs		
CNV Type	CAT1	CAT2	CAT3	CAT4	platinum
Duplication	34.76%	20.51%	5.64%	27.96%	2.30%
Deletion	0.05%	37.76%	0.42%	0.19%	0.35%
Del. & Dup.	0.02%	1.27%	0.04%	0.22%	0.02%
			Canvas CNVs		
Duplication	13.04%	1.50%	0.48%	5.20%	0.24%
Deletion	0.48%	58.87%	1.00%	1.84%	0.19%

2.3 SNVs co-segregating with CNVs

In many regions where we observed likely CNVs there were some SNVs that were pedigree-consistent and others that were not pedigree-consistent (marked as blue or red dots respectively in the examples shown in Figure 3). This is expected: to illustrate we have highlighted possible genotype calls that may occur in an area where the SNVs co-segregate with a duplication (Table S15). In this hypothetical example the father has two copies of a duplicated region on haplotype B. One copy contains the reference allele and the other copy contains the alternative allele. Because the variant callers are not aware of the correct ploidy in this region some of the genotypes will be incorrect in this family because they are calculated according to a diploid model. As an example, the father is genotyped as diploid heterozygous (0/1) each time even though he is triploid at this site. In the first row of Table S15 (status "PASS"), the resulting diploid genotype calls are consistent with a diploid model of inheritance. Conversely, in the next three rows (status "Mendel Error") the resulting diploid genotypes are not consistent with a diploid model of inheritance.

Table S15. Example of how a CNV may produce both pedigree-consistent and pedigree-inconsistent GTs.

Haplotypes				Father	Mother	Child 1	Child 2	Child 3	Child 4	Status
A	B	C	D	AB	CD	AC	AD	BC	BD	
0	01	0	0	0|01 (01)	0|0	0|0	0|0	01|0 (01)	01|0 (01)	PASS
1	01	0	0	1|01 (01)	0|0	1|0	1|0	01|0 (01)	01|0 (01)	Mendel Error

0	01	1	0	0|01 (01)	1|0	0|1	0|0	01|1 (01)	01|0 (01)	Mendel Error
0	01	0	1	0|01 (01)	0|1	0|0	0|1	01|0(01)	01|1 (01)	Mendel Error

2.4 Double crossovers and gene conversion

Excluding the category 1-3 SNVs described above, there were 6,127 category 4 SNV failures that have genotypes consistent with double crossovers or gene conversions. To exclude putative gene conversion that could be caused by other factors such as overlapping CNVs, we filtered out all SNVs where any of the 13 samples had depth less than 25x, depth greater than 75x or where there are fewer than 10 supporting reads for every predicted allele. These cutoff values were chosen because for the platinum sites: (a) 99% of the SNVs have read depth greater than 25x; (b) 99% of the SNVs have read depth less than 75x; and (c) 99% of the alleles (both reference and alternative) at heterozygous positions had at least 10 supporting reads. After removing these sites there were 589 SNVs that could be caused by gene conversion (Table S16). By merging nearby SNVs (<10kb) where all of the gene conversion evidence was consistent (same child and consistent parental origin), we identified 322 total regions of which 103 were consistent with a gene conversion on the paternal haplotype, 110 were consistent with gene conversion from the maternal haplotype and 109 were consistent with gene conversion on either the maternal or paternal haplotypes. A previous study estimates that a non-crossover event (gene conversion) occurs at a rate of ~5.7×10^{-6}/bp/generation (Williams et al. 2015). That estimate combined with the ~4.25M heterozygous SNVs in the parents predicts that 266 of the SNVs will exhibit a gene conversion in the 11 children, or about half the number of gene conversion events compared to the present analysis. Many of the SNVs in Table S16 show evidence of CNVs due to skewed allele counts between the paternally and maternally derived chromosomes. Follow up targeted sequencing of the original 6,127 SNVs in the pedigree and accurately identifying CNVs would provide a more accurate estimate of gene conversion in this pedigree.

3 Assessing Variant Calling Performance

3.1 Performance measured against the platinum regions from this study

We measured recall and precision for NA12878 against our platinum truth data using four different sequencing pipelines (Platypus, FreeBayes, GATK3 and Strelka ; see Table S1) using both single sample mode (Table S17) and joint calling using the parents (Table S18). Table S19 shows the relative improvements gained by using joint calling. For this analysis we used the same sequence data on NA12878 that was used for the rest of this study though we sampled the data to approximate depths of 30x, 40x and 50x (full data) to understand the affects of sequencing to different depths. It should be noted that different sequencing experiments will likely yield lower values. Additionally, this assessment requires that both the genotypes and alleles agree exactly and in some difficult regions the same variants may be represented differently (e.g. Figure S3). More sophisticated comparison and/or normalization algorithms are needed to deal with the problems that occur when the same variant(s) is represented differently.

Table S17. Recall and precision calculated using the platinum data as a reference for different sequencing pipelines on sequence data for NA12878 sampled down to 30x, 40x and 50x (all of the data).

	30x				40x				50x			
	Indels		SNVs		Indels		SNVs		Indels		SNVs	
	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.
Platypus	70.29	97.2	95.35	99.9	70.42	97.25	95.45	99.91	70.48	97.28	95.49	99.91
FreeBayes	85.19	92.37	99.17	99.67	86.26	92.8	99.27	99.73	86.88	93.04	99.31	99.77
GATK3	92.28	96.2	96.28	99.95	93.5	96.58	98.13	99.94	94.72	96.78	98.64	99.94
Strelka	93.67	95.84	97.23	99.93	94.36	96.32	97.37	99.96	94.41	96.54	97.4	99.96

Table S18. Recall and precision calculated using the platinum data as a reference for different sequencing pipelines using joint calling (NA12878, NA12891 & NA12892).

	30x				40x				50x			
	Indels		SNVs		Indels		SNVs		Indels		SNVs	
	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.
Platypus	70.07	97.07	95.36	99.85	70.33	97.02	95.39	99.86	70.5	97.02	95.4	99.86
FreeBayes	85.71	92.09	99.33	99.63	86.51	92.51	99.38	99.7	86.96	92.75	99.41	99.74
GATK3	93.08	96.42	97.92	99.92	94.14	96.68	98.63	99.92	95.15	96.81	98.89	99.92
Strelka	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table S19. Relative improvements to the recall and precision measured against the platinum data as the reference when joint calling is employed (i.e. Table S18 minus Table S17).

	30x				40x				50x			
	Indels		SNVs		Indels		SNVs		Indels		SNVs	
	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.
Platypus	-0.22	-0.13	0.01	-0.05	-0.09	-0.23	-0.06	-0.05	0.02	-0.26	-0.09	-0.05
FreeBayes	0.52	-0.28	0.16	-0.04	0.25	-0.29	0.11	-0.03	0.08	-0.29	0.1	-0.03
GATK3	0.8	0.22	1.64	-0.03	0.64	0.1	0.5	-0.02	0.43	0.03	0.25	-0.02
Strelka	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

3.2 Performance measured against the NIST confident regions

We also measured recall and precision against the NIST truth data for the same variant calls used for Table S20. In general, the recall is 0.6-2.5% higher for SNVs measured against the NIST data but the precision is up to 1.6% lower (Tables S20 and S21). Of particular note is that there is a significant increase in both recall (1-2.2%) and precision (1.3-2.5%) of indels when using the bwa aligments whereas both recall and precision are lower for the Strelka calls based on Isaac alignments. This highlights the possibility that biases may be introduced into comparisons when the same tools are not used to build up the starting truth data (most of the data for NIST was aligned using bwa). Likewise, the Platinum calls of this study may need to be supplemented to incorporate any pedigree-consistent variant calls that are identified with new and improved informatics pipelines.

Table S20. Recall and precision calculated using the NIST data as the reference.

	30x				40x				50x			
	Indels		SNVs		Indels		SNVs		Indels		SNVs	
	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.
Platypus	71.84	98.49	97.9	99.56	71.79	98.52	97.98	99.6	71.78	98.57	98.01	99.65
FreeBayes	87.37	94.78	99.84	98.06	88.31	95.23	99.9	98.29	88.87	95.52	99.91	98.48
GATK3	93.75	98.71	98.78	99.89	94.64	98.86	99.71	99.75	95.72	98.89	99.8	99.69
Strelka	93.39	95.04	98.68	99.87	93.9	95.57	98.78	99.91	93.94	95.86	98.8	99.92

Table S21. Changes in recall and precision estimations using the NIST data as a reference compared with the results obtained using the platinum data as the reference. Positive values indicate that the estimates are higher when using the NIST data to assess the variant calls.

	30x				40x				50x			
	Indels		SNVs		Indels		SNVs		Indels		SNVs	
	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.	recall	prec.
Platypus	1.55	1.29	2.55	-0.34	1.37	1.27	2.53	-0.31	1.3	1.29	2.52	-0.26
FreeBayes	2.18	2.41	0.67	-1.61	2.05	2.43	0.63	-1.44	1.99	2.48	0.6	-1.29
GATK3	1.47	2.51	2.5	-0.06	1.14	2.28	1.58	-0.19	1	2.11	1.16	-0.25
Strelka	-0.28	-0.8	1.45	-0.06	-0.46	-0.75	1.41	-0.05	-0.47	-0.68	1.4	-0.04

4 Comparison Against Other Studies

4.1 Variants in NIST and 1kGP that are not included in this database

Both the NIST dataset and the 1000 Genomes data include variant calls on NA12878 that were not in our final catalogue of truth variants and it is important to understand the reasons for this. For example, the NIST calls contain 62,946 SNVs that were not included in our platinum catalogue. Upon inspection, all except 287 of the ~63k SNVs were identified in our Platinum data analysis but were subsequently excluded by our inheritance test or by additional filters. The majority (72.6%) passed our pedigree check

in at least one pipeline but failed subsequent filters (see Table S22). For example, 22,728 (36.1%) failed our *k*-mer test, 15,051 (23.9%) were homozygous alternate alleles in all individuals but this result was limited to a single aligner, and 7,864 (12.5%) had conflicting calls in our study. The remaining 16,972 were not pedigree consistent, and of these 18.3% (3,105) overlap with our high quality failures. While the absence of these variants in our database does not negate the possibility that they are true positives, the exact genotypes are ambiguous in our study and we would therefore exclude these variants from benchmarking tests. Of the 60,057 indel calls in the NIST dataset but absent from the Platinum catalogue, 98.3% overlapped an initial call in our study but were later removed by one or more of our quality filters. Similar to the SNVs, the indels specific to the NIST data may be real, indicating the need for further analysis and refinement of the Platinum catalogue in the future.

There were also 224,651 SNVs in the 1kGP calls for NA12878 that were not included in our final call set (Table S22). Examining the 1KG SNVs in more detail, 46% (103,885) were pedigree-consistent but filtered out in our study, and 42% (95,339) were observed in this study but were not pedigree-consistent. For these variants, while we cannot exclude the possibility that these are true positives, the exact genotypes are ambiguous in our study and we would therefore exclude these variants from benchmarking tests. For the remaining 25,427 that were not detected in this study, we suspect most or all them to be false positives on the basis that they were not observed in any of the members of this pedigree. Of the 159,163 indel calls in the 1kGP dataset for NA12878 but absent from the Platinum catalogue, 94.0% overlap an initial call in our study but were later removed by one or more of our quality filters. Similar to the SNVs, the indels specific to the 1kGP data may be real, indicating a need for further analysis and refinement of the Platinum catalogue in the future.

Table S22. Counts of SNVs in the NIST data and 1kGP but not in platinum calls according to why they were excluded from this call set.

Category	Number in NIST	Number in 1kGP
Pedigree consistent but failed additional rules[1]		
Single Hom Alts	15,051	31,021
Kmer fails	22,728	55,057
Overlap conflicts	7,864	17,490
Possible platinum	44	317
Sub Total	45,687	103,885
High quality failures (category 1-4)[2]		
All heterozygous	588	5,793
Possible deletions	356	2,170
Singletons	1,706	1,301
Hamming one	347	1,867
Hamming >one	108	2,105
Sub Total	3,105	13,236
Low quality failures (not included in any analysis)[3]		
Full data	3,109	44,700
Missing data	10,758	37,403

Sub Total	13,867	82,103
	Likely false positives[4]	
Not in pedigree	287	25,427
Total not in PG	62,946	224,651

[1]SNVs that were called consistently with the transmission in the parents and all children but failed the additional rules for inclusion as platinum

[2]SNVs that were not consistent with the transmission but were called by at least two software pipelines and the GTs were consistent where multiple callers made a call

[3]SNVs that either had missing data, or were not called by more than one pipeline or were not consistent between callers

[4]SNVs not seen in either the parents on eleven children of this pedigree

4.2 Coverage of Difficult Regions

Overall, the Platinum catalogue contains over 800k SNVs and 223k indels that are not included within the NIST call set. One of the primary benefits of using a pedigree compared to alternative methods such as replication analysis is that we can identify variants in genomic regions that may be more difficult or where a single variant calling pipeline performs well and delivers platinum variants, though the other pipelines fail. As an assessment of how these more "difficult" regions are included in this study we separated SNVs and indels into two categories: 1) those called as platinum by all of the alignment based methods (defined here as "consistent"); and 2) those identified as platinum by only a subset of the alignment based methods (defined here as "difficult"). For these two sets of variant descriptors we compared our sub-call sets (within NA12878) against the NIST call set. For the consistent variants the NIST call set includes just 69.4% (indels) and 82.6% (SNVs) of our consistent sites highlighting that even for these "easy" sites the pedigree method is more sensitive (Table S23). For the platinum variants in the "difficult" category, only 45.4% of the indels and 61.9% of the SNVs are also in the NIST call set highlighting that the pedigree method allows us to identify many more SNVs and indels in "difficult" regions. These "difficult" platinum variants are useful for improving the software tools because they highlight areas where certain methods perform worse than others and these regions can be targeted for improvement.

Table S23. Overlap and concordance between NIST and Platinum Genomes by variant descriptor.

Variant Type	Consistent Variants			Difficult Variants		
	Number	Recall	Concordance	Number	Recall	Concordance
SNVs	2,612,894	82.65	100.00	912,018	61.93	99.99
Indels	265,688	69.39	>99.99	262,647	45.36	99.76

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet **30**: 97-101.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV et al. 2011. Variation in genome-wide mutation rates within and between human families. Nat Genet **43**: 712-714.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science **327**: 78-81.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet **44**: 226-232.

Roller E, Ivakhno S, Lee S, Royce T, Tanner S. 2016. Canvas: versatile and scalable detection of copy number variants. Bioinformatics doi:10.1093/bioinformatics/btw163.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. Science **349**: aab3761.

Williams AL, Geneovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R et al. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. bioRxiv doi:10.1101/009175.