SUPPLEMENTARY METHODS

Definition of CS regions from the RG. The RG hg38 (GRCh38) was downloaded from the UCSC Golden Path website. All kmers (with a sliding window of 1 bp, k=30 nt) from GRCh38 were obtained, and the unique kmers (CSs) were retrieved using Bowtie (1). A list containing the start and end positions of regions composed only by CSs was obtained.

Criteria to define inconsistent reads. The reads with: i) a different strand assigned to the PrevCS and PostCS alignment, ii) multiple alignments of the same CS, iii) and no PrevCS alignment were saved to an inconsistencies file and were not used in subsequent analyses

Genotype assignment. The probability of three possible genotypes was computed: homozygous reference (homo-R), heterozygous reference/no-reference (hete-R/NR), and homozygous non-reference (homo-NR). If the assigned genotype was different than homo-R, the allele with the highest frequency other than the reference was obtained and it was considered to be the major allele. Additionally, three probabilities were computed: homozygous major (homo-M), heterozygous major/no major, and homozygous no major (homo-NM). If the genotypes with the highest probability were either homo-NR and homo-NM or hete-R/NR and hete-M/NM or hete-R/NR and homo-NM, there were at least three probable alleles that could be assigned with high probability as the genotype at that particular site. This resulted in an ambiguous genotype calling. Ambiguous SNVs were written to an ambiguous genotype file, and no further analysis was done with these variant sites. For genotype assignment, the reads with the allele N were not taken into account.

Criteria to define a *bona fide de novo* variant. The criteria for defining a possible variant, such as a *bona fide de novo* variant, were: 1) at least 10 reads spanning that site in the child and at least 10 reads in each parent, at least 2 total alignments in the child and at least 2 in each parent; 2) the variant allele should not be contained in more than one high-quality alignment (total alignments) in any parent; 3) the variant allele of the child should be in more than one-fourth of all the reads spanning that site or the variant region could be duplicated in the child genome (15); and 4) the candidate *de novo* SNV must be absent from public SNV databases, such as dbSNP.

Definition of *accessible genome*. All 100-nt windows (with a sliding window of 1 bp) were obtained. For each window, the number of CSs was computed. All consecutive windows with a CS density higher than 0.5 were concatenated, creating a CS-accessible-region. The CS-accessible-regions constituted the callable genome, except for k nucleotides at the start and end of each region. For all simulations and real sequencing data results, only SNVs in the accessible genome were reported.

Simulation experiments. SNVs were introduced into chromosome 12 with a mutation rate of 0.001. The position for every variant site was chosen at random, in addition to the phase and the alternative allele. We used the ART Simulator to generate sequencing reads using the HiSeq Illumina error profile (100 bp paired-end reads) (2), and we applied the COBASI pipeline to call the SNVs. We varied several key parameters, such as sequencing depth, kmer size, minimum coverage for the Signature CSs, absolute value for the RCI, maximum difference in coverage between the Signature CSs, minimum number of whole-VSR alignments, and optimal extension for the partial alignments. To compute Precision-Recall curves, we obtained the number of False-Negative (FN), False-Positive (FP), and True-Positive (TP) calls at different coverage thresholds for each set of parameters. We calculated the Area Under the Curve (AUPR) as a performance score.

In the case of the parent-offspring simulation, SNVs were introduced into chromosome 12 with a mutation rate of 0.001 to create the father diploid chromosome. The position for every variant site was randomly chosen, as well as the phase and the alternative allele. To create the mother diploid chromosome, for every variant site for the father, the phase for the mother was chosen at random. One father chromosome and one mother chromosome were chosen to create the child's pair of chromosomes. *De novo* mutations were introduced in positions not previously mutated in the child with a mutation rate of 3e-7 (39 SNVs). The *de novo* mutation rate was artificially increased to yield a considerable amount of *de novo* SNVs. For all three individuals, we used the ART Simulator to generate sequencing reads using the HiSeq Illumina error profile (100 bp paired-end reads). The coverage for each individual was chosen to resemble our real sequencing experiments, 35x coverage for each parent and 100x coverage for the child. We applied the COBASI pipeline to discover the *de novo* SNVs

www.pnas.org/cgi/doi/10.1073/pnas.1802244115

with the set of parameters that maximizes the APR for each sequencing depth (obtained from one individual simulation). For the child: sequencing depth = 100x, kmer size = 30, minimum coverage for the Signature CSs = 10, absolute value for the RCI = 0.2, maximum difference in coverage between the Signature CSs = 2.0, minimum number of whole-VSR alignments = 3, optimal extension for the partial alignments = 10. For the parents: sequencing depth = 35x, kmer size = 30, minimum coverage for the Signature CSs = 5, absolute value for the RCI = 0.2, maximum difference in coverage between the Signature CSs = 2.0, minimum number of whole-VSR alignments = 2, optimal extension for the partial alignments = 5. We repeated this simulation experiment 20 times and obtained the median values for the FP, FN, and TP calls.

Variant calling using alignment-based pipelines. The best practices guideline (3) was followed to call SNV from 5 (chosen at random) out of the 20 simulations: reads were mapped using BWA, duplicate reads were removed using Picard, local realignment around indels was done, base quality score was recalibrated, genotypes were assigned using GATK HaplotypeCaller and variants were filtered using a hard filter. Finally, *de novo* variants were identified using GATK VariantAnnotator.

TRIO sequencing and COBASI application. DNA from whole blood was extracted using the QIAmp DNA Blood Mini Kit as described by the manufacturer. Three libraries were prepared for the child and one for each parent. The CODIS STRs were determined for each individual. The DNA libraries were sequenced by paired-end Illumina HiSeq 2000 with a read length of 100 bp. The COBASI pipeline was used to discover *de novo* SNVs from the TRIO sequencing data using the same parameters as in *de novo* simulations.

Experimental validation of *de novo* SNVs. PCR primers were designed using the Oligo7 software and manual inspection. PCR was performed using the Accuprime Pfx kit according to the manufacturer's instructions. PCR products were sequenced by Sanger sequencing at Macrogen, Inc. To determine the specific position corresponding to the nucleotide of interest, each Sanger sequence was aligned to the RG using BLAST (4). The genotypes of the sites of interest were determined by manual inspection of the chromatograms.

Probability of one mutation occurring independently at the same site in two unrelated genomes. . There is some disagreement about the human mutation rate (5). However, for the sake of argument, we will assume the worst-case scenario. This means the highest mutation rate, which means 80 new mutations per haploid genome. (6, 7). There are ways to select a set of 80 mutated base pairs in a genome of length nucleotides. Of these, contain a fixed base pair. Therefore, the probability of a fixed base pair being contained in the set of 80 mutations of a genome is (we recover the mutation rate). The probability of any fixed base pair being contained in the set of mutations of two independent genomes is , which is very low. Because of this, any *de novo* SNV is not expected to be found in any population SNV database.

SUPPLEMENTARY TABLES

TABLE S1. Comparison of genomic regions defined as accessible by COBASI and the 1000 Human Genomes Project

DENSITY CUTOFF	BOTH	ONLY COBASI	ONLY 1000HGP	NEITHER	TOTAL COBASI	TOTAL 1000HGP
0	90	10	0	0	100	90
10	88	1	2	8	90	90
20	87	1	3	8	88	90
30	86	1	4	9	87	90
40	83	2	7	9	85	90
50	82	2	8	9	84	90

The callable genome by COBASI was defined in the Methods. In the 1000 Genomes Project, the “accessible genome” was defined based on coverage and mapping quality criteria. Regions with very high or low coverage, as well as many low-quality mapped reads, were defined as unaccessible regions. In the table, several CSs density cutoffs are shown, and the percentage of the genome that is defined as callable by 1) both projects, 2) only COBASI, 3) only 1000HGP, 4) neither project, 5) COBASI, or 6)1000HGP is shown.

TABLE S2. The Area Under the Curve for the Precision-Recall curves (APR) for the COBASI simulation in one individual, part I.

Parameters	**35**x	**50**x	**75**x	**100**x
k=25,RCIV=0.2,total-aln=2	0.932	0.933	0.927	0.915
k=30,RCIV=0.2,total-aln=2	**0.943**	0.952	0.951	0.946
k=25,RCIV=0.2,total-aln=3	0.931	0.939	0.943	0.943
k=30,RCIV=0.2,total-aln=3	0.928	0.951	0.959	**0.961**
k=25,RCIV=0.2,total-aln=4	0.923	0.936	0.942	0.944
k=30,RCIV=0.2,total-aln=4	0.900	0.942	0.956	0.960
k=25,RCIV=0.2,total-aln=6	0.887	0.927	0.938	0.941
k=30,RCIV=0.2,total-aln=6	0.790	0.910	0.947	0.955
k=30,RCIV=0.3,total-aln=2	0.933	0.949	0.951	0.946
k=25,RCIV=0.3,total-aln=2	0.925	0.932	0.926	0.915
k=30,RCIV=0.3,total-aln=3	0.921	0.948	0.959	0.961
k=25,RCIV=0.3,total-aln=3	0.925	0.938	0.942	0.943
k=30,RCIV=0.3,total-aln=4	0.895	0.940	0.956	0.960
k=25,RCIV=0.3,total-aln=4	0.917	0.935	0.942	0.944
k=30,RCIV=0.3,total-aln=6	0.790	0.909	0.947	0.955
k=25,RCIV=0.3,total-aln=6	0.884	0.926	0.938	0.941

One human chromosome (chromosome 12) was mutated (mutation rate = 0.001), and simulated reads were produced for this mutant chromosome. SNVs were called using COBASI by varying three parameters: the kmer size (k), the minimum relative change in coverage to identify a VSR (RCIV), and the minimum number of reads that should contain both SignatureCSs (total-aln). The Area under the curve for the Precision-Recall (APR) curves are shown. To compute the plots, the precision and recall were calculated for different coverage thresholds in a particular simulation. Invariant parameters over these simulations: the extension for alignments of reads containing only the PrevCS (n = 5) for all sequencing depths, the minimum coverage for any Signature CS (rmin = 5 for sequencing depths of 35 and 50´ and rmin = 10 for sequecning depths of 75 and 100´). The set of parameters chosen to perform the parent-offspring simulation are highlighted as bold numbers

TABLE S3. The Area Under the Curve for the Precision-Recall curves (APR) for the COBASI simulation in one individual, part II.

Parameters	35x	100x
ratio=1.5	0.919	0.961
ratio= 2.0	0.943	0.961
ratio=2.5	0.943	0.961
ratio=2.0, n=10	0.941	0.961
ratio=2.0, rmin=5	--	0.960
ratio=2.0, rmin=10	0.943	--

Parameters	35x	100x
ratio=1.5	9411 20	4502 3
ratio=2.0	**6620** **29**	**4451** **5**
ratio=2.5	6561 36	4421 15
ratio=2.0, n=10	6780 46	**4463** **5**
ratio=2.0, rmin=5	--	4591 5
ratio=2.0, rmin=10	6581 32	--

One human chromosome (chromosome 12) was mutated (mutation rate = 0.001) and simulated reads were produced for this mutant chromosome. SNVs were called using COBASI by varying three parameters: the minimum coverage for any Signature CS (rmin), a maximum ratio between the coverage of the Signature CSs (ratio), and the extension for alignments of reads containing only the PreCS (n). The Area under the curve for the Precision-Recall (APR) curves are shown. To compute the plots, the precision and recall were calculated for different coverage thresholds in a particular simulation. Invariant parameters over these simulations: 35x: k =30, RCIV = 0.2, total-aln = 2; 100x: k = 30, RCIV = 0.2, total-aln =3. Default parameters (otherwise mentioned): for all coverage thresholds: n= 5; 35x: rmin = 5 and 100x: rmin = 10. The left table contains the APR score for every simulation, and the right table contains the FN and FP for every simulation. The set of parameters chosen to perform the parent-offspring simulation are highlighted as bolded numbers

TABLE S4. Experimental validation of each predicted *de novo* SNVs.

CHR	POS	REF	FATHER	MOTHER	CHILD	STATUS
chr1	24862021	G	G/G	G/G	G/T	OK
chr1	90547932	G	G/G	G/G	GA	OK
chr1	167295816	A	A/A	A/A	A/G	OK
chr1	172427805	G	G/G	G/G	T/G	OK
chr1	207061328	G	G/G	G/G	G/T	NoPCR
chr1	233278131	A	A/A	A/A	A/G	OK
chr2	7834800	G	G/G	G/G	G/C	OK
chr2	24287324	T	T/T	T/T	T/C	OK
chr2	64935802	G	G/G	G/G	G/T	OK
chr2	117515206	A	A/A	A/A	A/G	OK
chr2	159087258	C	C/C	C/C	C/T	BQ
chr2	166134730	G	G/G	G/G	G/A	OK
chr2	174144299	C	C/C	C/C	C/T	OK
chr3	13257366	C	C/C	C/C	C/T	OK
chr3	35344598	T	T/T	T/T	T/A	OK
chr3	84019551	C	C/C	C/C	C/T	OK
chr3	85475191	G	G/G	G/G	G/C	OK
chr3	130405591	G	G/G	G/G	G/T	OK
chr3	154730842	A	A/A	A/A	A/G	OK
chr3	177039650	C	C/C	C/C	C/T	OK
chr3	193814289	G	G/G	G/G	G/T	OK
chr4	12050118	T	T/T	T/T	T/G	OK
chr4	122532439	C	C/C	C/C	C/T	OK
chr4	165308533	C	C/C	C/C	C/T	OK
chr4	183179287	C	C/C	C/C	C/T	OK
chr5	42087606	T	T/T	T/T	T/C	OK
chr6	54488698	A	A/A	A/A	A/T	OK
chr6	110925590	T	T/T	T/T	T/C	OK
chr6	145688494	A	A/A	A/A	A/G	OK
chr6	149023483	C	C/C	C/C	C/T	OK
chr7	8845957	C	C/C	C/C	C/A	OK
chr7	18840247	A	A/A	A/A	A/T	OK
chr7	131254278	G	G/G	G/G	G/T	OK
chr7	148217676	A	A/A	A/A	A/G	OK
chr8	38433070	G	G/G	G/G	G/A	OK
chr8	68845327	T	T/T	T/T	T/A	OK
chr9	74292655	A	A/A	A/A	A/T	OK
chr9	135134043	C	C/C	C/C	C/A	OK
chr10	967661	T	T/T	T/T	T/C	OK
chr10	69932637	A	A/A	A/A	A/C	OK
chr10	124545656	T	T/T	T/T	T/G	NoPCR*
chr11	46199782	A	A/A	A/A	A/C	NoPCR*
chr11	9834859	C	C/C	C/C	C/T	OK
chr11	22218005	G	G/G	G/G	G/T	OK
chr11	57031949	C	C/C	C/C	C/T	OK
chr11	66915741	A	A/A	A/A	A/G	OK
chr11	98890913	G	G/G	G/G	G/A	OK
chr11	120059843	A	A/A	A/A	A/C	OK

chr12	7422099	C	C/C	C/C	C/T	OK
chr13	78641958	C	C/C	C/C	C/T	OK
chr15	81812391	T	T/T	T/T	T/C	OK
chr16	76704617	C	C/C	C/C	C/T	PCRInesp
chr17	61212465	A	A/A	A/A	A/G	OK
chr19	7406505	A	A/A	A/A	A/G	OK
chr20	59356016	A	A/A	A/A	A/C	PrimInes
chrX	87169908	T	T/T	T/T	T/G	OK
chrX	125321179	A	A/A	A/A	A/G	OK

The table contains all the predicted *de novo* SNVs and the results of their experimental validation. Each row shows the chromosome, the geomic position, and the genotype predicted for each individual for each SNV. In the column "experimental status:" OK means that the Sanger sequencing results and the COBASI prediction are consistent for all the individuals. PrimInesp means that no specific primers could be designed because of the presence of a highly repetitive region surrounding the SNV. PCRInesp means that no unique PCR product could be obtained even when specific primers were designed. NoPCR means that no PCR product could be obtained. BQ means that no quality sequence could be obtained even when the sequencing was repeated several times, likely the result of the presence of low-complexity regions (long stretches of poli-dT) found in that specific region.

TABLE S5. Experimental validation for a subset of Mendelian SNVs.

ID	CHR	POS	REF	FATHER	MOTHER	CHILD	STATUS
1	1	108095723	A	G/G	G/G	G/G	OK
2	1	147610227	G	G/G	A/A	G/A	OK
3	2	19463414	G	G/G	G/A	G/A	OK
4	2	161057267	G	G/T	T/T	G/T	OK
5	3	4085479	T	T/T	T/C	T/C	OK
6	3	157221449	G	G/A	G/A	A/A	OK
7	4	107667929	C	C/G	C/C	C/G	OK
8	4	146842576	C	C/C	C/A	C/A	OK
9	5	44277000	C	C/T	C/C	C/T	OK
10	5	80058324	C	C/T	C/T	C/T	OK
11	6	67230929	G	G/G	G/A	G/A	OK
12	6	147785976	C	A/A	C/C	C/A	OK
13	7	77752055	C	G/G	G/G	G/G	OK
14	7	109782464	A	T/T	T/T	T/T	OK
15	8	21514268	T	T/C	T/C	C/C	OK
16	8	27092236	T	T/C	T/T	T/C	OK
17	9	4516070	C	C/T	C/T	C/T	OK
18	9	117229359	C	C/C	C/T	C/T	OK
19	10	8766364	C	C/A	C/A	A/A	OK
20	10	79241132	T	C/C	T/T	T/C	OK
21	11	7937566	C	C/G	C/C	C/G	OK
22	11	6722659	C	C/T	C/C	C/T	OK
23	12	17811890	G	G/T	G/T	G/T	OK
24	12	53864157	A	G/G	G/G	G/G	OK
25	13	74721621	A	G/G	G/G	G/G	OK
26	13	85333647	T	G/G	G/G	G/G	OK
27	14	20550811	G	A/A	G/A	A/A	OK
28	14	55442549	G	A/A	A/A	A/A	OK
29	15	78319135	G	A/A	A/A	A/A	OK
30	15	80916473	G	C/C	C/C	C/C	OK
31	16	5978486	C	C/C	C/G	C/G	OK
32	16	7924503	A	A/A	C/C	A/C	OK
33	17	8758708	A	A/A	A/G	A/G	OK
34	17	72131667	G	A/A	G/A	G/A	OK
35	18	5901790	C	C/G	C/G	G/G	OK
36	18	72951051	T	C/C	T/C	C/C	OK
37	19	19774267	C	C/A	A/A	C/A	OK
38	19	28737356	C	C/T	T/T	T/T	OK
39	20	10778727	T	C/C	T/C	C/C	OK
40	20	64264757	C	C/T	C/C	C/T	OK
41	21	28205983	A	G/G	G/G	G/G	OK
42	21	41715615	G	G/C	G/G	G/C	OK
43	22	23908608	C	G/G	C/G	C/G	OK
44	22	36265520	G	G/A	G/A	A/A	OK
45	X	8928469	T	T/T	C/C	T/C	OK
46	X	22643455	T	C/C	T/T	T/C	OK

The table contains a subset of Mendelian SNVs and the results of their experimental validation. Each row shows the chromosome, the genomic position, and the genotype predicted for each individual for each SNV. In the column, experimental status “OK” means that the Sanger sequencing results are consistent for all individuals.

TABLE S6. Computing time, core number, and RAM required for every stage for the COBASI approach.

	12 N[1] 64Gb RAM	12 N[1] 128Gb RAM	24 N[1] 64Gb RAM	24 N[1] 128Gb RAM
1. ONE TIME PROCESS. CS DATABASE CREATION				
OBTAIN CS DATABASE				
Cut reference genome	00:32	00:32	00:17	00:17
Obtain unique kmers	01:26	01:26	00:58	00:58
Obtain non-overlapping kmers	00:13	00:13	00:10	00:10
TOTAL 1	**02:11**	**02:11**	**01:25**	**01:25**
2a. GENOME-WIDE SNV DISCOVERY				
OBTAIN LANDSCAPE				
Count kmers	02:04	02:36	01:52	01:29
Obtain whole-genome coverage	06:00	03:20	06:00	03:20
Obtain landscape	00:18	00:18	00:12	00:12
SUBTOTAL	**8:22**	**6:14**	**8:04**	**5:01[2]**
GET SIGNATURE REGIONS AND SIGNATURE READS				
Get Variant Signature Regions	00:40	00:40	00:25	00:25
Obtain Signature CSs sequence	00:09	00:09	00:09	00:09
Get Signature Reads	06:00	06:00	06:00	06:00
FILTER READS AND GET SNVs				
Get SNVs	28:40	28:40	23:00	20:00
SUBTOTAL	**35:29**	**35:29**	**29:24**	**26:34**
TOTAL 2a	**43:51**	**41:43**	**37:28**	**31:35[3]**
2b. *DE NOVO*-ORIENTED SNV DISCOVERY (PARENTAL GENOMES)				
GET SIGNATURE REGIONS AND SIGNATURE READS				
Get Variant Signature Regions	00:40	00:40	00:25	00:25
Obtain Signature CSs sequence	01:09	01:09	01:09	01:09
Get Signature Reads	06:00	06:00	06:00	06:00
FILTER READS AND GET SNVs				
Get SNVs	01:51	01:51	01:25	01:10
TOTAL 2b	**9:40**	**9:40**	**8:59**	**8:44**

The first process of the COBASI pipeline is the CS database creation. This process must be done once per reference genome. To discover the *de novo* SNV, all SNVs must be called in the child (Stage 2a), and these positions must be interrogated in the parents (Stage 2b). For every step, the computation time required for different hardware specifications is shown.

[1] N denotes the number of processors.

[2] The whole-genome Variation Landscape can be generated in only 5 hours.
[3] A SNV list from the raw whole-genome sequencing data is generated in less than 36 hours.
[4] If only some regions of interest are chosen for further investigation, the COBASI approach can generate a list of resulting SNVs from the whole-genome sequencing raw data in less than 9 hours

SUPPLEMENTARY FIGURES

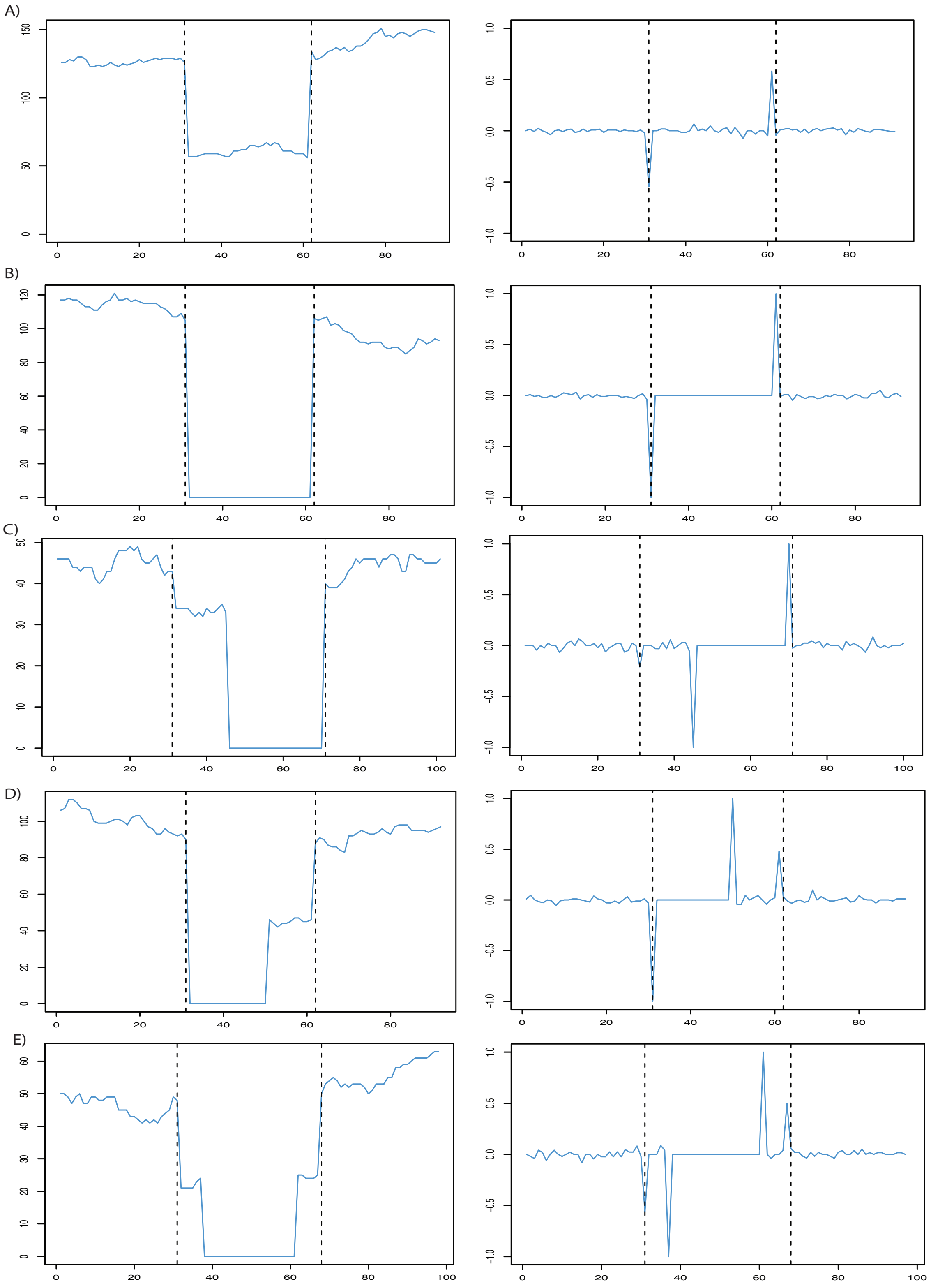
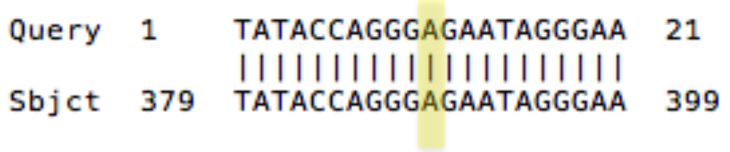
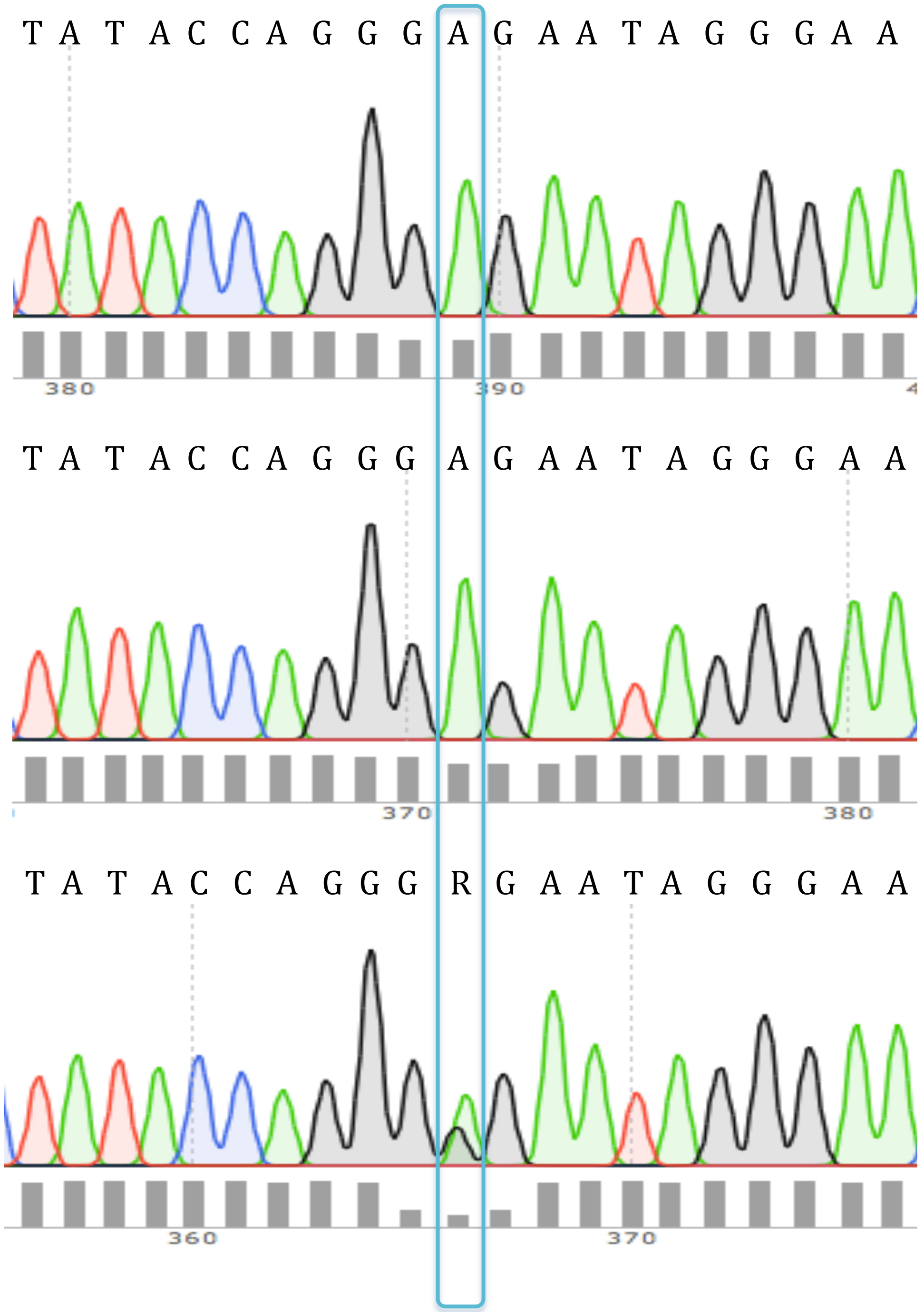


Fig. S1. DIFFERENT TYPES OF VARIANT SIGNATURE REGIONS (VSR). Several variants can be close enough to be concatenated on the same VSR. Depending on their zygosity and chromosomal localization, four different VSR patterns are found. A) A classic VSR formed by only one heterozygous SNV is shown. If there is more than one heterozygous SNV localized on the same chromosome, the SVR is extended. Position 1 on the X axis corresponds to chr7:9,449,337. B) A classic VSR formed by only one homozygous SNV is shown. If there is more than one homozygous SNV localized on the same chromosome, the SVR is extended. Position 1 on the X axis corresponds to chr21:21,616,422. C) A VSR formed when a heterozygous variant is followed by a homozygous SNV is shown. Position 1 on the X axis corresponds to chr17:83,187,030. D) A SNV formed when a homozygous SNV is followed by a heterozygous SNV is shown. Position 1 on the X axis corresponds to chr12:133,163,159. E) A VSR formed when two heterozygous SNVs localized on different chromosomes are found. Position 1 on the X axis corresponds to chr14:104,928,266. Left, the VL for a specific genomic region is shown. Every plot shows the start position of each CS (X axis) and the coverage for each CS (Y axis). Right, the RVL for the same regions is shown. Every plot shows the start position of each CS (X axis) and the RCI values associated with each CS (Y axis). The start positions for the PrevCS and PostCS are shown as dashed vertical lines. The VL and RVL depicted correspond to the child's genome.

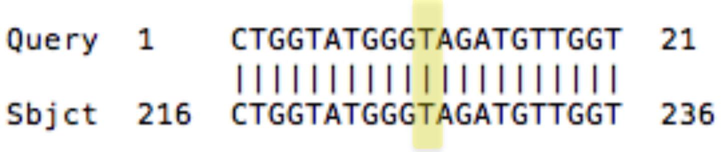
ID	CHR	POSITION	FATHER	MOTHER	CHILD
3	chr1	167295816	A/A	A/A	A/G



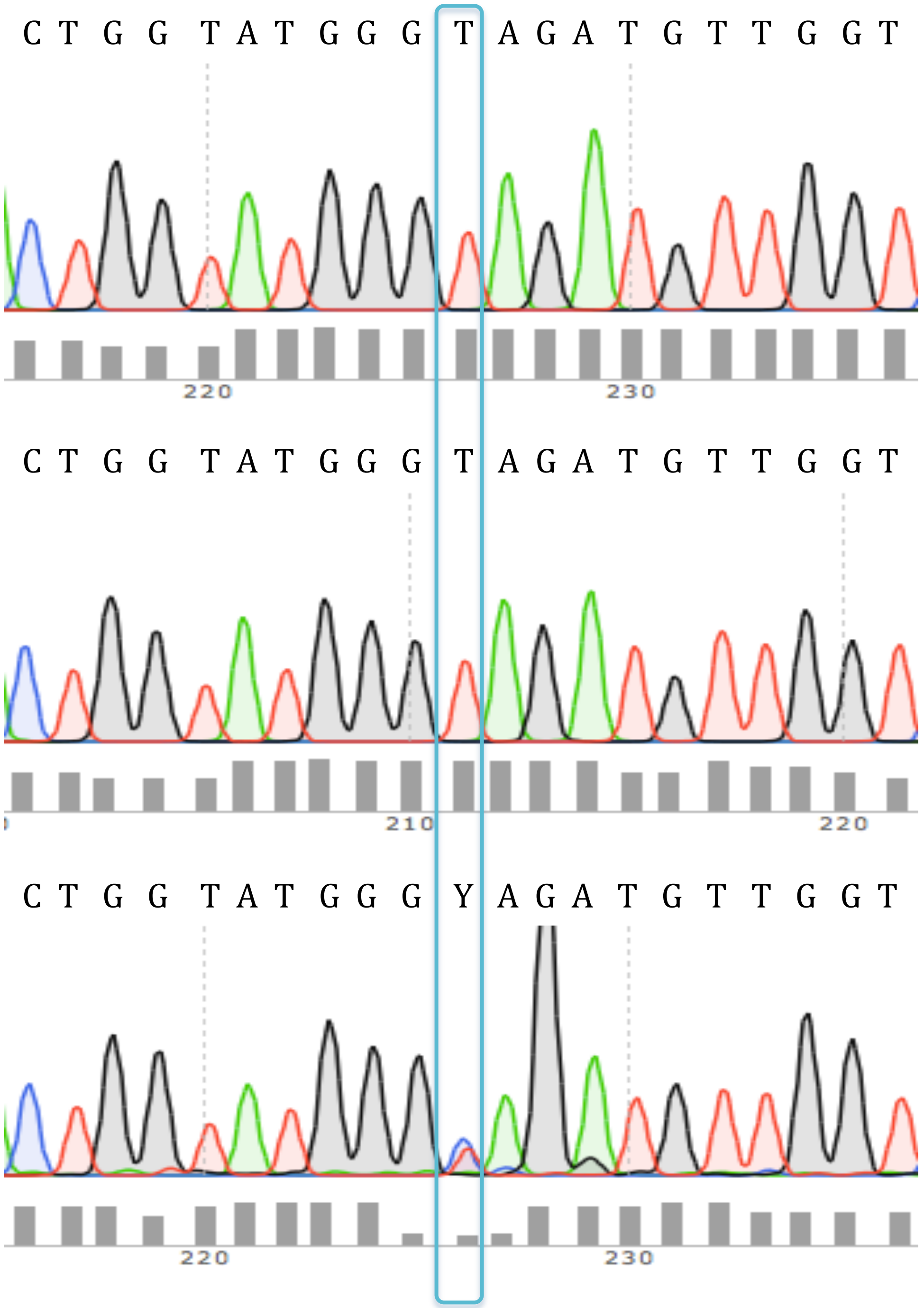
```
Query  1    TATACCAGGGAGAATAGGGAA  21
            |||||||||||||||||||||
Sbjct  379  TATACCAGGGAGAATAGGGAA  399
```



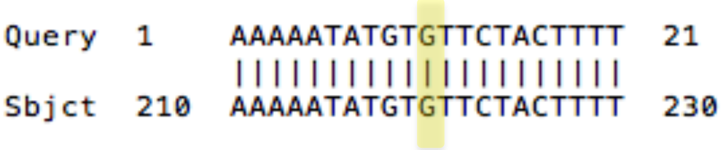
ID	CHR	POSITION	FATHER	MOTHER	CHILD
8	chr2	24287324	T/T	T/T	T/C



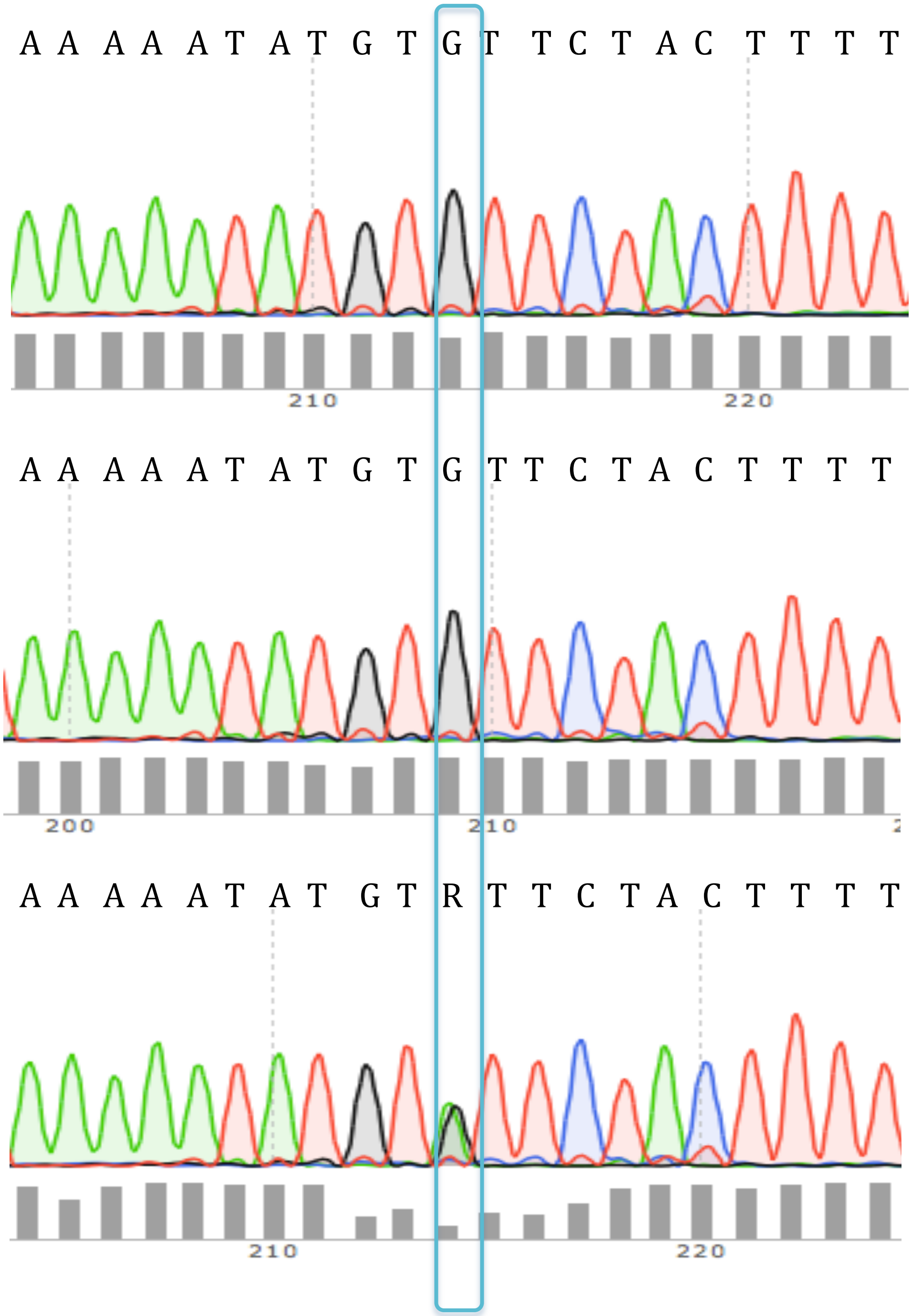
```
Query  1    CTGGTATGGGTAGATGTTGGT  21
            |||||||||||||||||||||
Sbjct  216  CTGGTATGGGTAGATGTTGGT  236
```



ID	CHR	POSITION	FATHER	MOTHER	CHILD
12	chr2	166134730	G/G	G/G	G/A

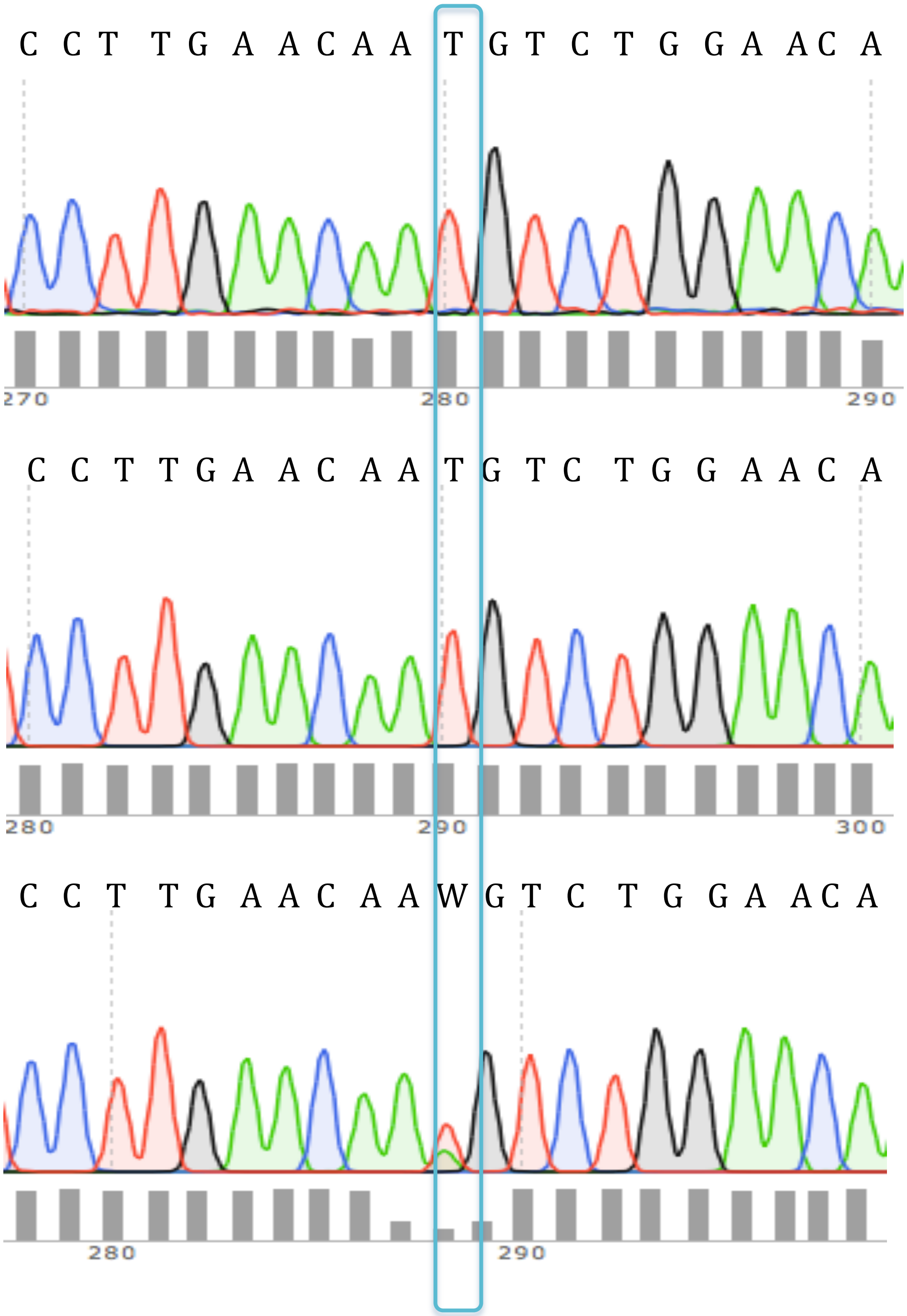


```
Query  1    AAAAATATGTGTTCTACTTTT  21
            |||||||||||||||||||||
Sbjct  210  AAAAATATGTGTTCTACTTTT  230
```

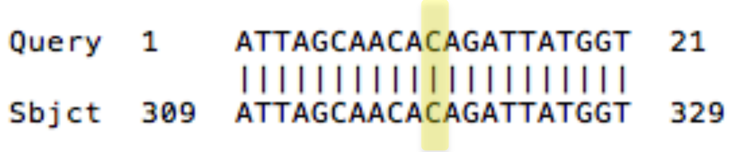


ID	CHR	POSITION	FATHER	MOTHER	CHILD
15	chr3	35344598	T/T	T/T	T/A

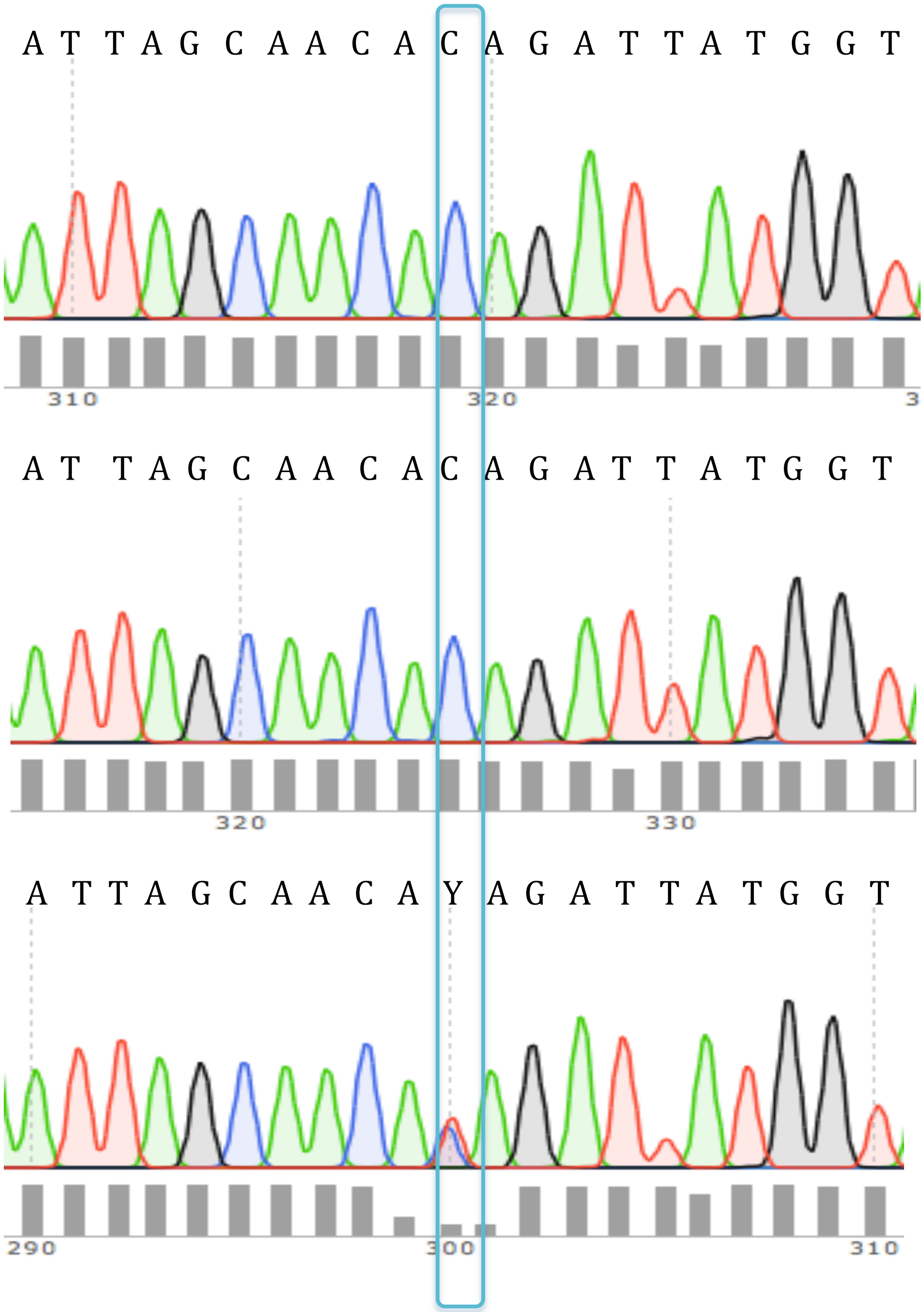
```
Query  1    CCTTGAACAATGTCTGGAACA  21
            |||||||||||||||||||||
Sbjct  274  CCTTGAACAATGTCTGGAACA  294
```



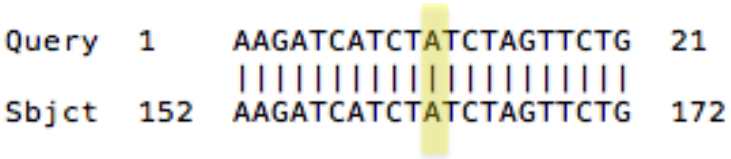
ID	CHR	POSITION	FATHER	MOTHER	CHILD
20	chr3	177039650	C/C	C/C	C/T



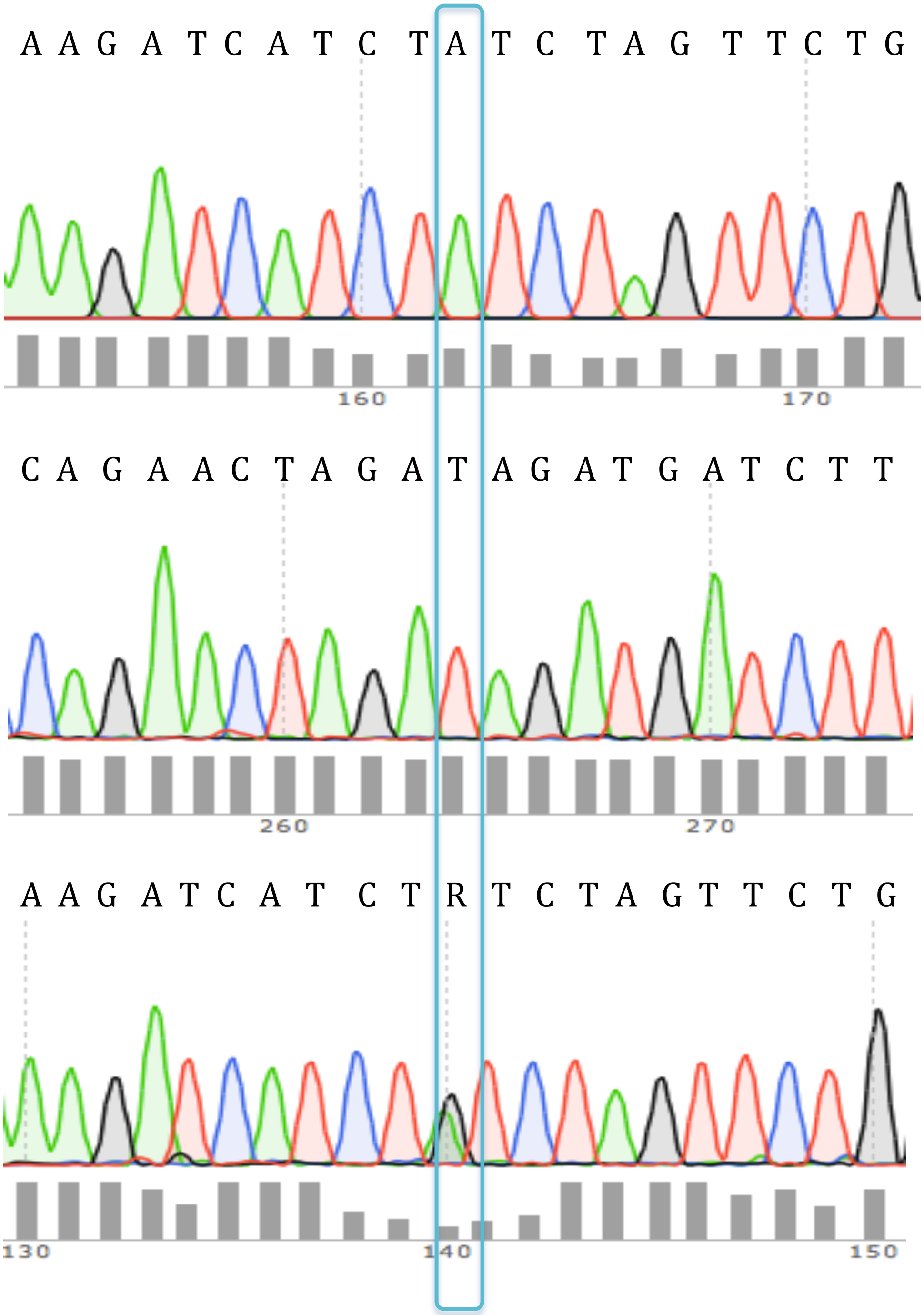
```
Query  1    ATTAGCAACACAGATTATGGT  21
            |||||||||||||||||||||
Sbjct  309  ATTAGCAACACAGATTATGGT  329
```



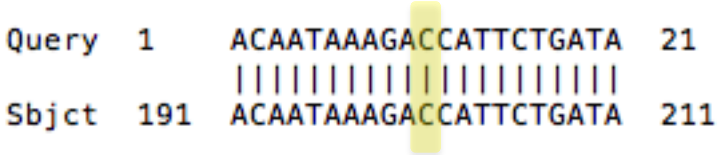
ID	CHR	POSITION	FATHER	MOTHER	CHILD
29	chr6	145688494	A/A	A/A*	A/G



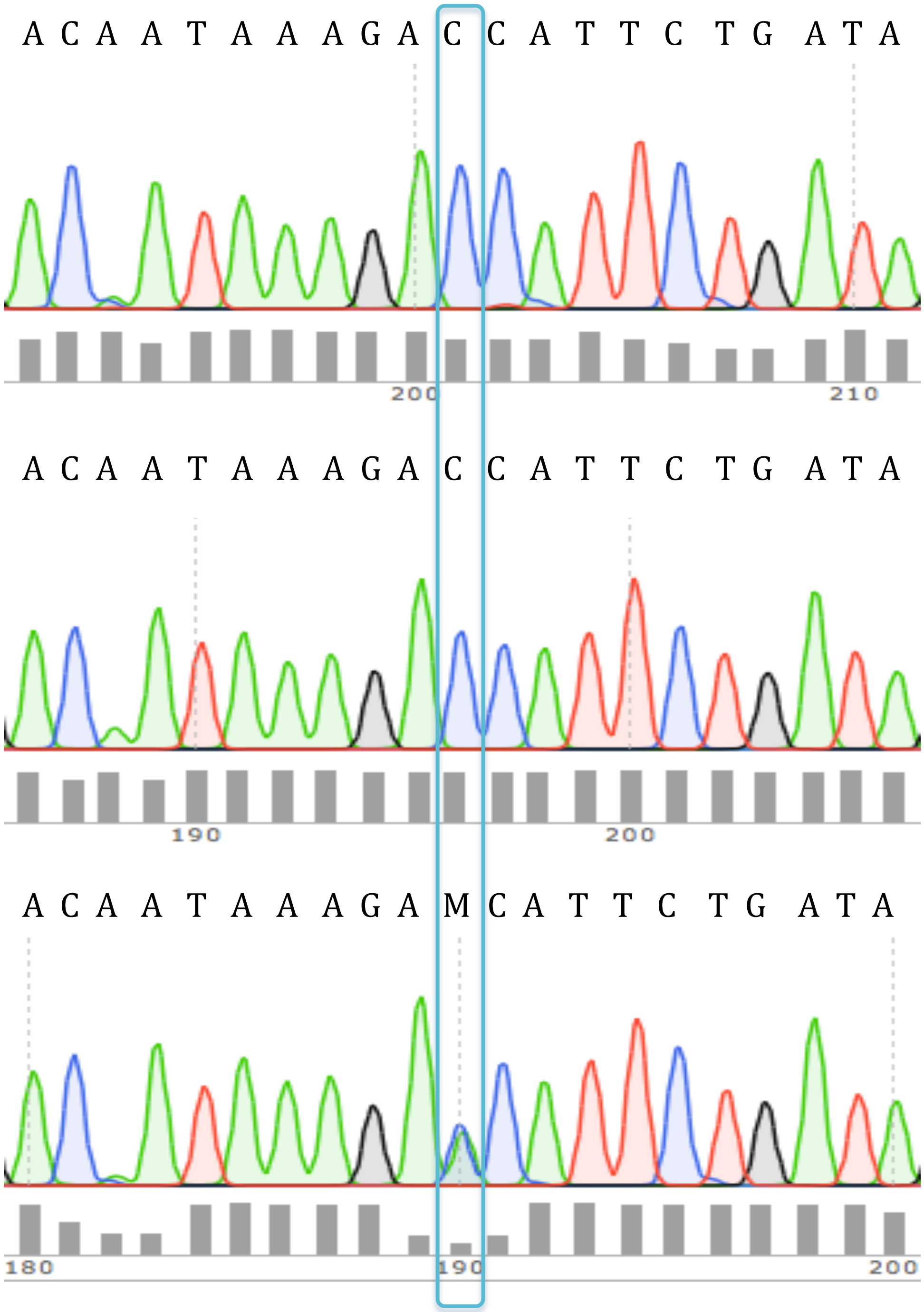
```
Query  1    AAGATCATCTATCTAGTTCTG  21
            |||||||||||||||||||||
Sbjct  152  AAGATCATCTATCTAGTTCTG  172
```



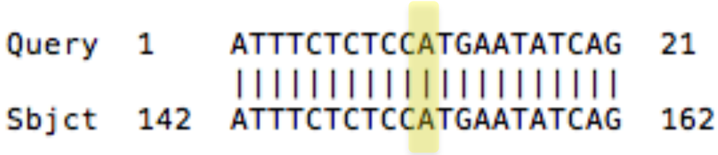
ID	CHR	POSITION	FATHER	MOTHER	CHILD
31	chr7	8845957	C/C	C/C	C/A



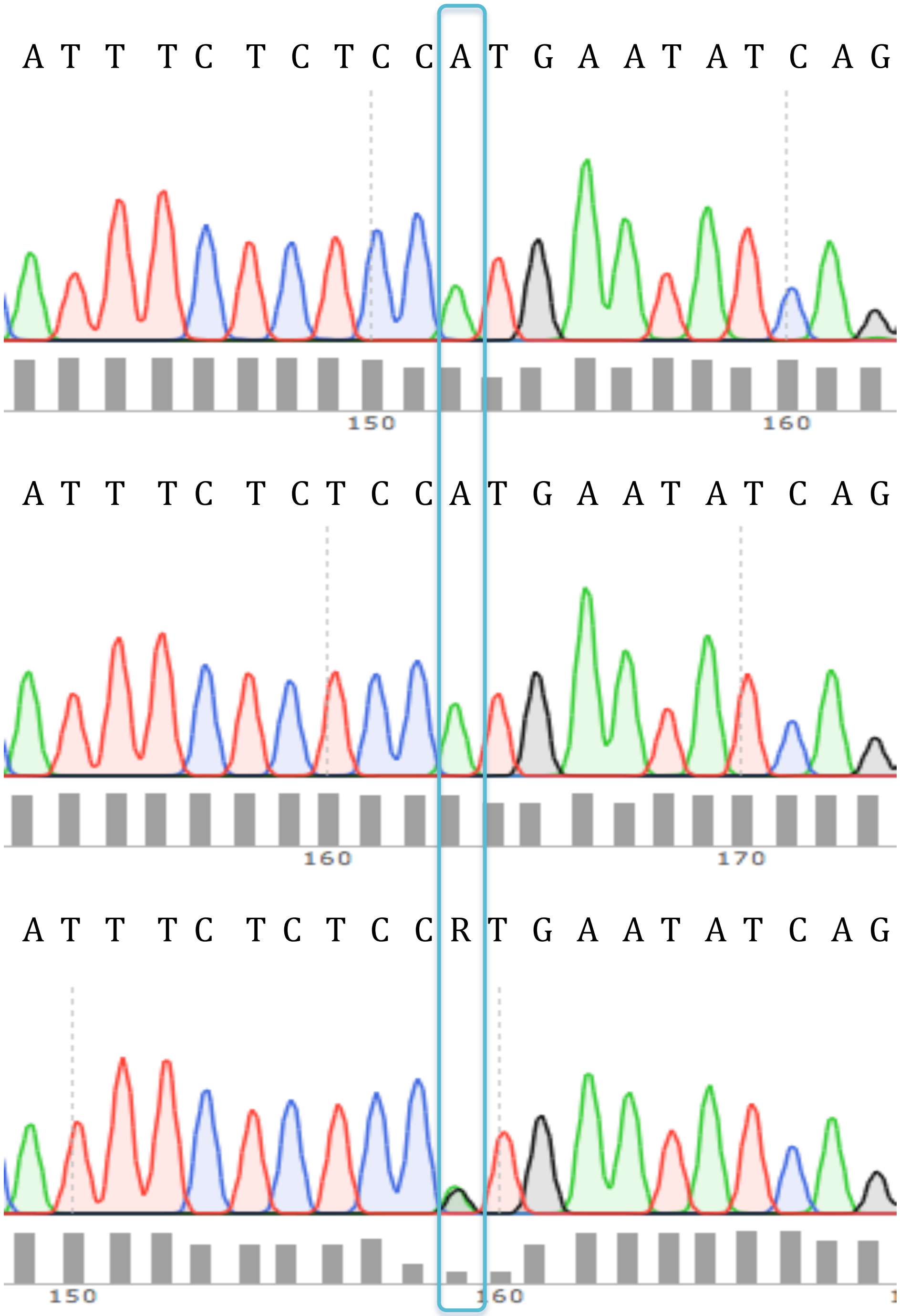
```
Query  1    ACAATAAAGACCATTCTGATA  21
            |||||||||||||||||||||
Sbjct  191  ACAATAAAGACCATTCTGATA  211
```



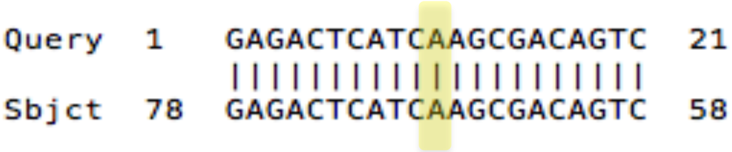
ID	CHR	POSITION	FATHER	MOTHER	CHILD
34	chr7	148217676	A/A	A/A	A/G



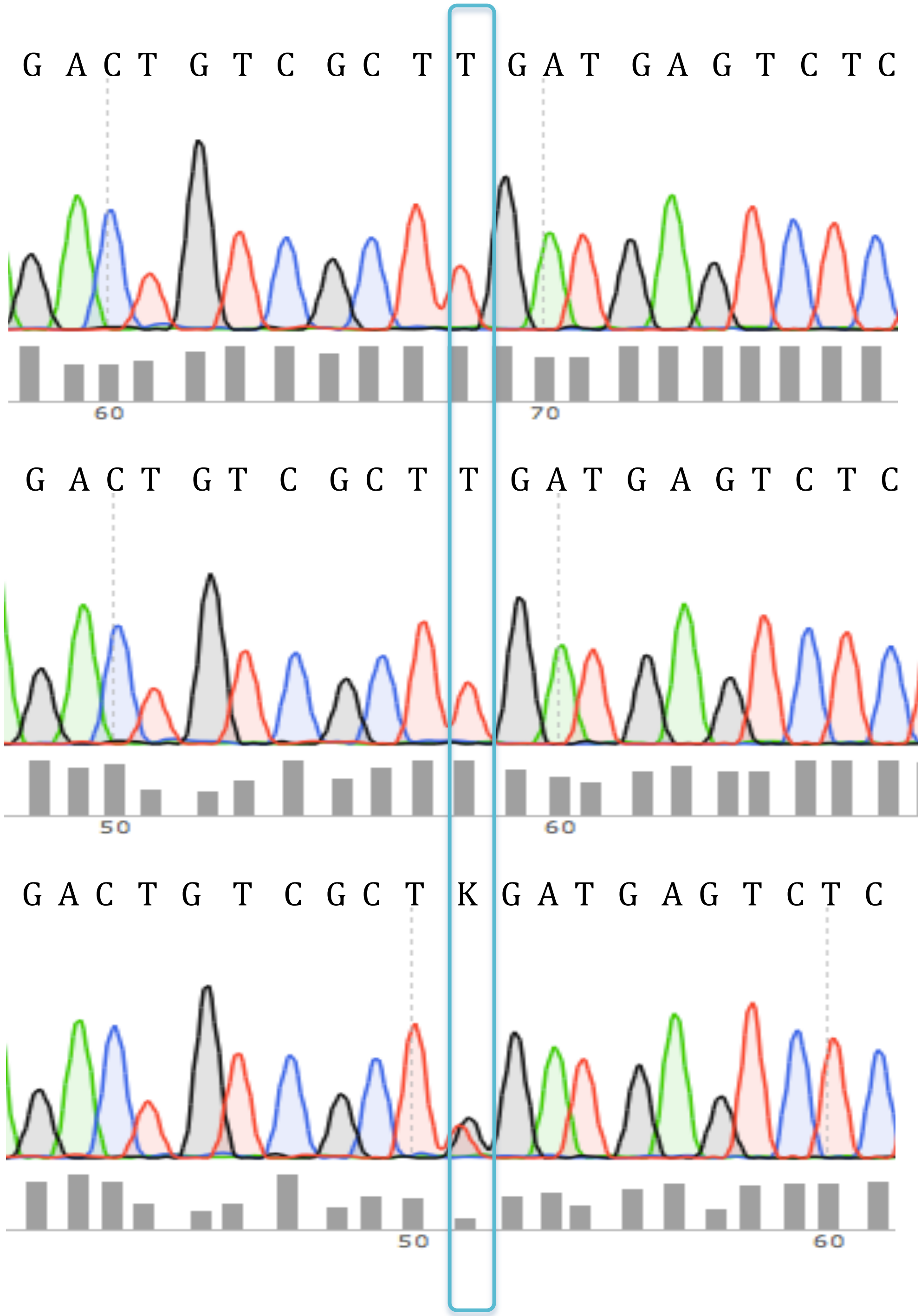
```
Query  1    ATTTCTCTCCATGAATATCAG  21
            |||||||||||||||||||||
Sbjct  142  ATTTCTCTCCATGAATATCAG  162
```



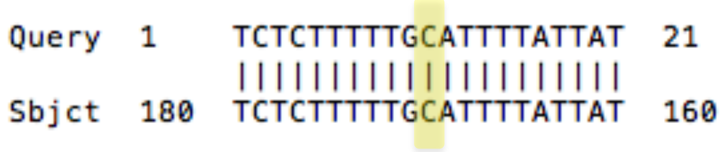
ID	CHR	POSITION	FATHER	MOTHER	CHILD
40	chr10	69932637	A/A*	A/A*	A/C*



```
Query  1   GAGACTCATCAAGCGACAGTC  21
           |||||||||||||||||||||
Sbjct  78  GAGACTCATCAAGCGACAGTC  58
```



ID	CHR	POSITION	FATHER	MOTHER	CHILD
45	CHR11	63582494	C/C*	C/C	C/T*



```
Query  1    TCTCTTTTTGCATTTTATTAT  21
            |||||||||||||||||||||
Sbjct  180  TCTCTTTTTGCATTTTATTAT  160
```

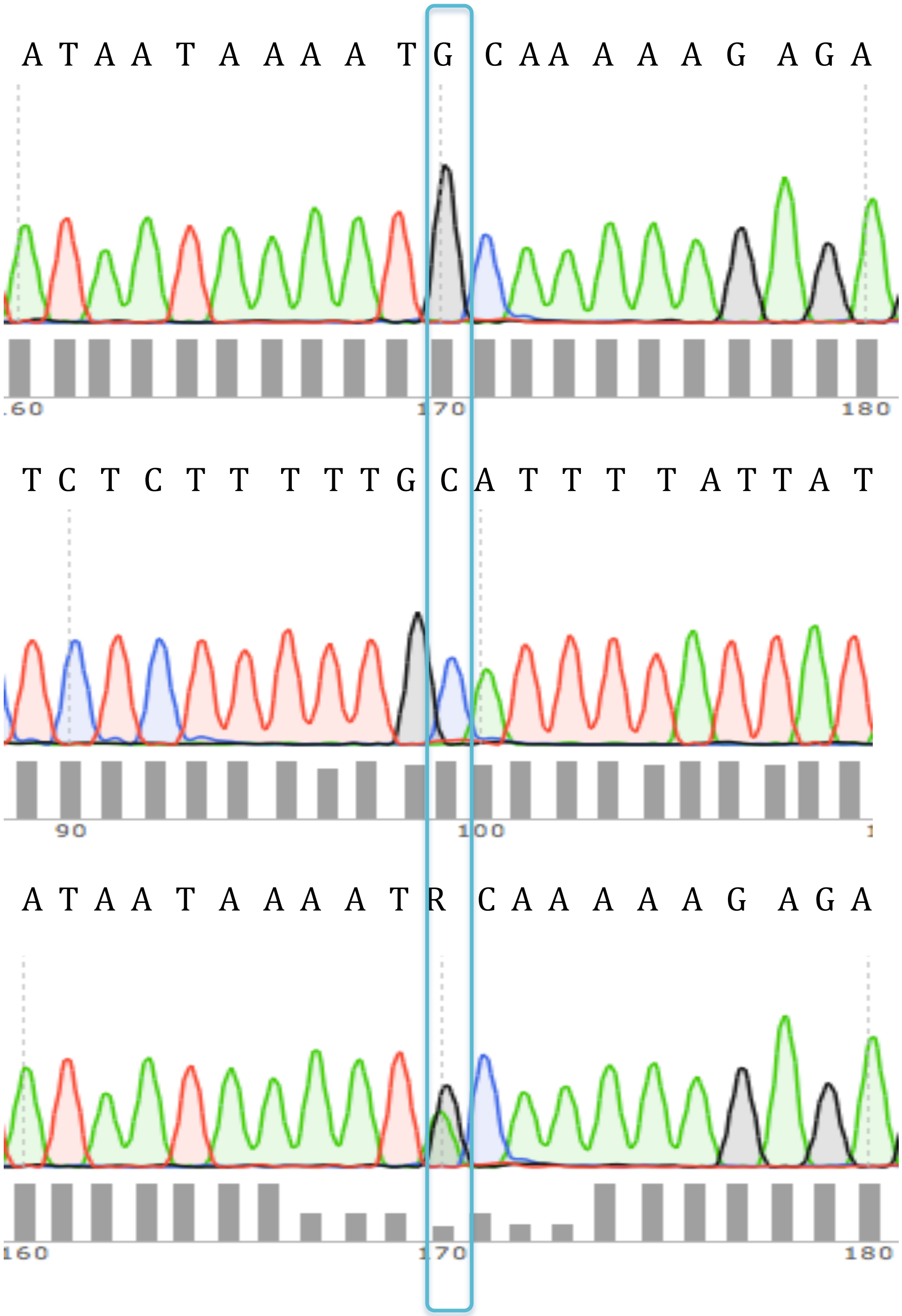
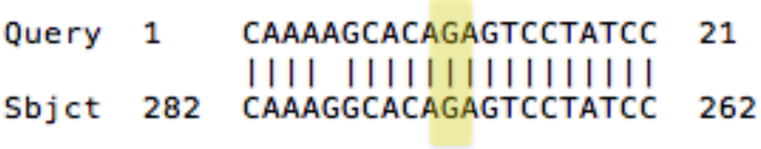
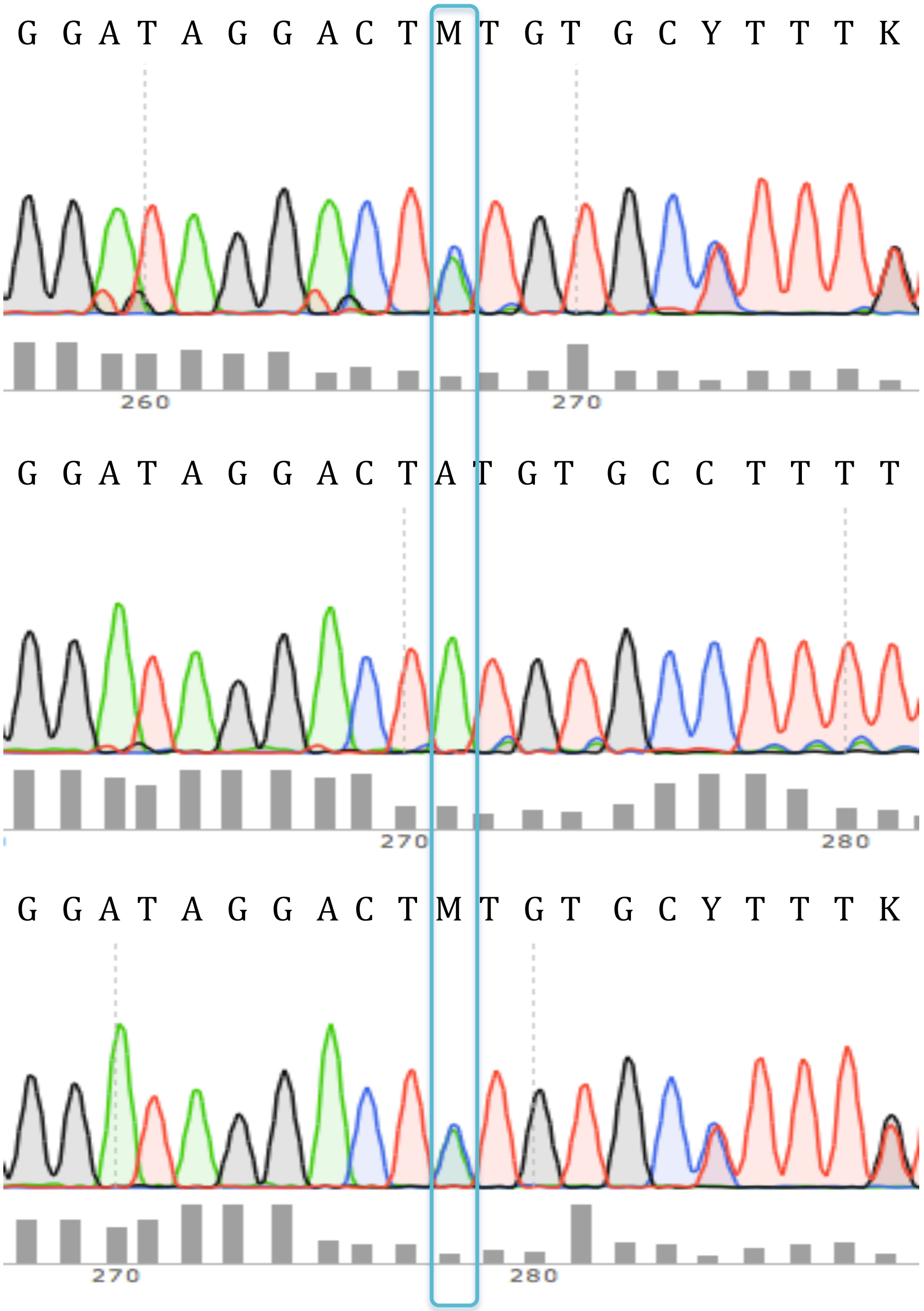


Fig. S2. CHROMATOGRAMS FOR EACH *DE NOVO* SNV. Only 10 SNVs designated as *de novo* by COBASI were chosen at random. For every SNV, a table with the chromosome and genomic coordinates along with the genotype designated for every trio individual is shown. Also, an alignment between the HRG and the father Sanger sequence is shown. This alignment is centered on the SNV position, which is highlighted in a yellow rectangle. The Sanger sequence chromatograms for that genomic region are shown for the father, mother, and child. The SNV position is highlighted in a blue rectangle, and the chromatogram quality metrics are illustrated as bars at the bottom of each chromatogram.

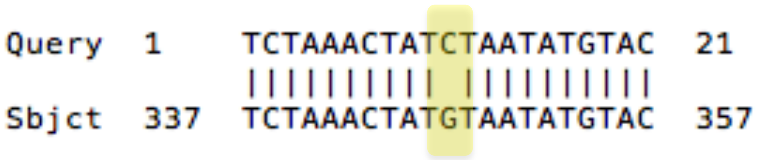
ID	CHR	POSITION	FATHER	MOTHER	CHILD
4	chr2	161057267	G/T*	T/T*	G/T*



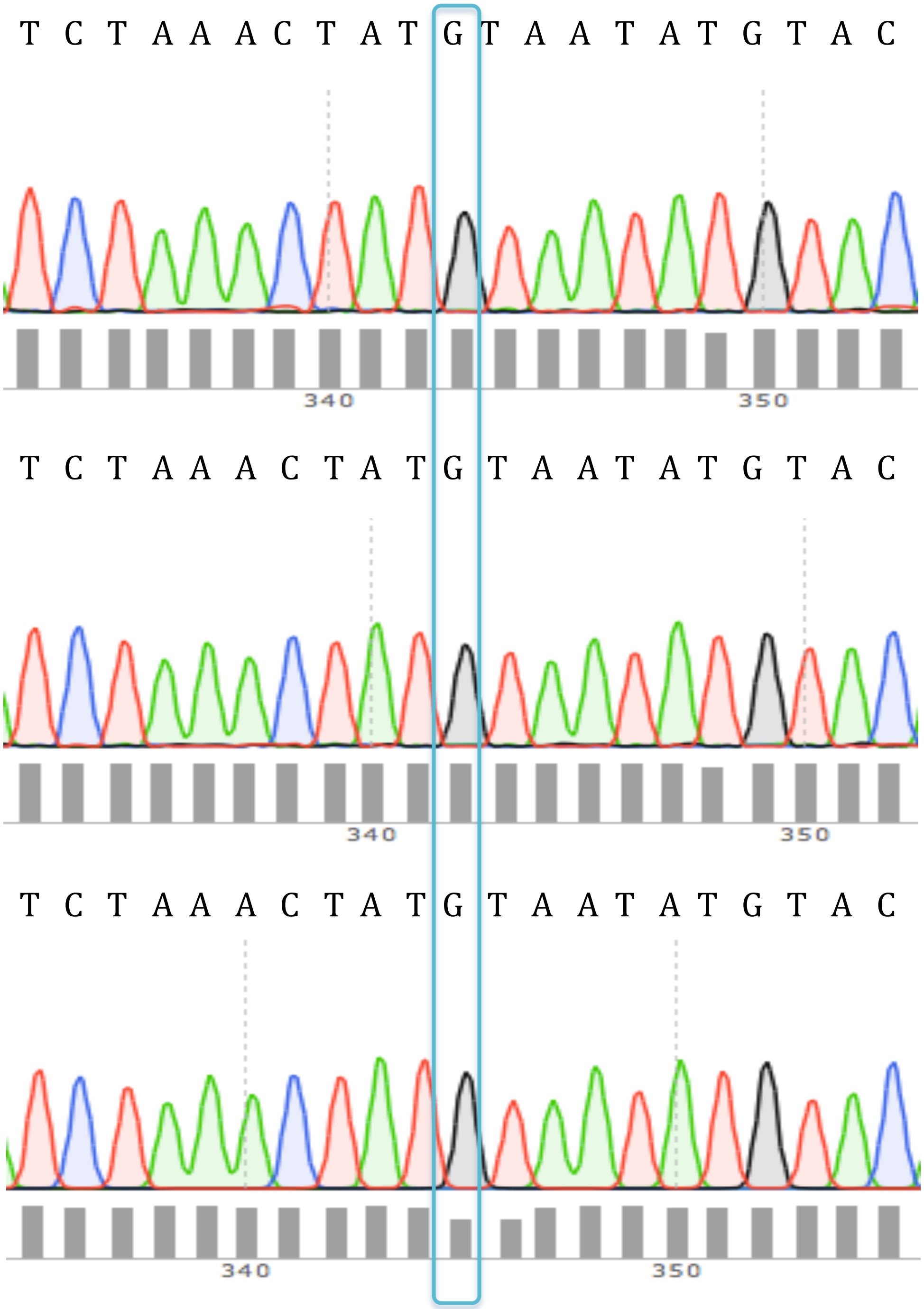
```
Query  1    CAAAAGCACAGAGTCCTATCC  21
            |||| ||||||||||||||||
Sbjct  282  CAAAGGCACAGAGTCCTATCC  262
```



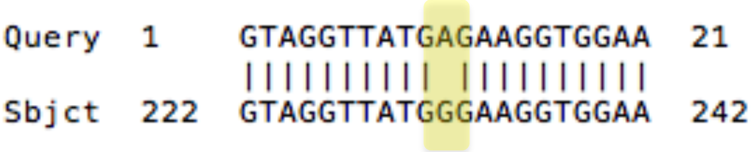
ID	CHR	POSITION	FATHER	MOTHER	CHILD
13	chr7	77752055	G/G	G/G	G/G



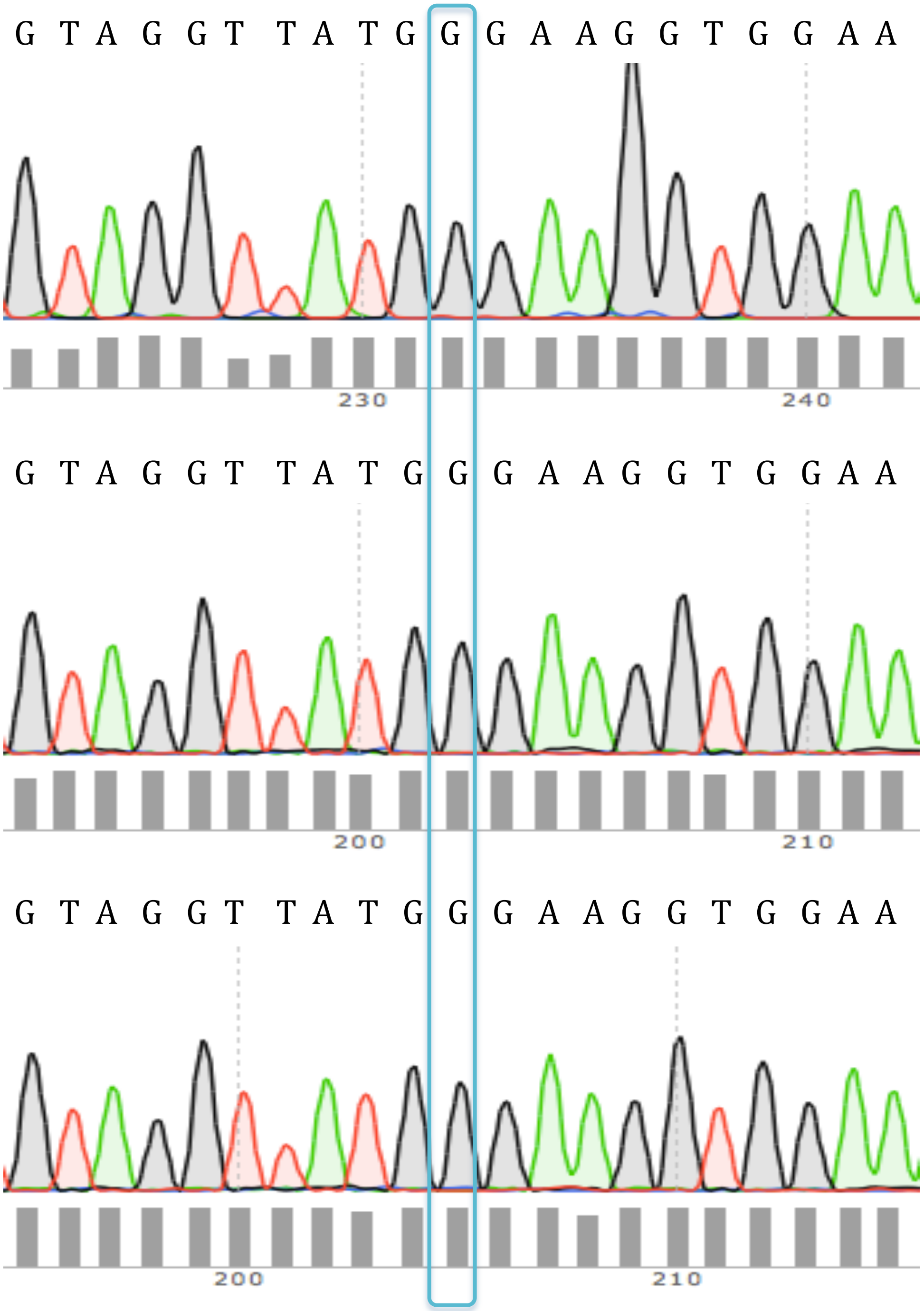
```
Query  1    TCTAAACTATCTAATATGTAC  21
            |||||||||| ||||||||||
Sbjct  337  TCTAAACTATGTAATATGTAC  357
```



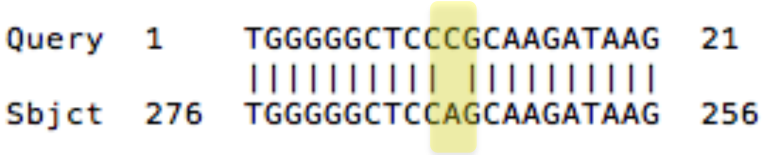
ID	CHR	POSITION	FATHER	MOTHER	CHILD
24	chr12	53864157	G/G	G/G	G/G



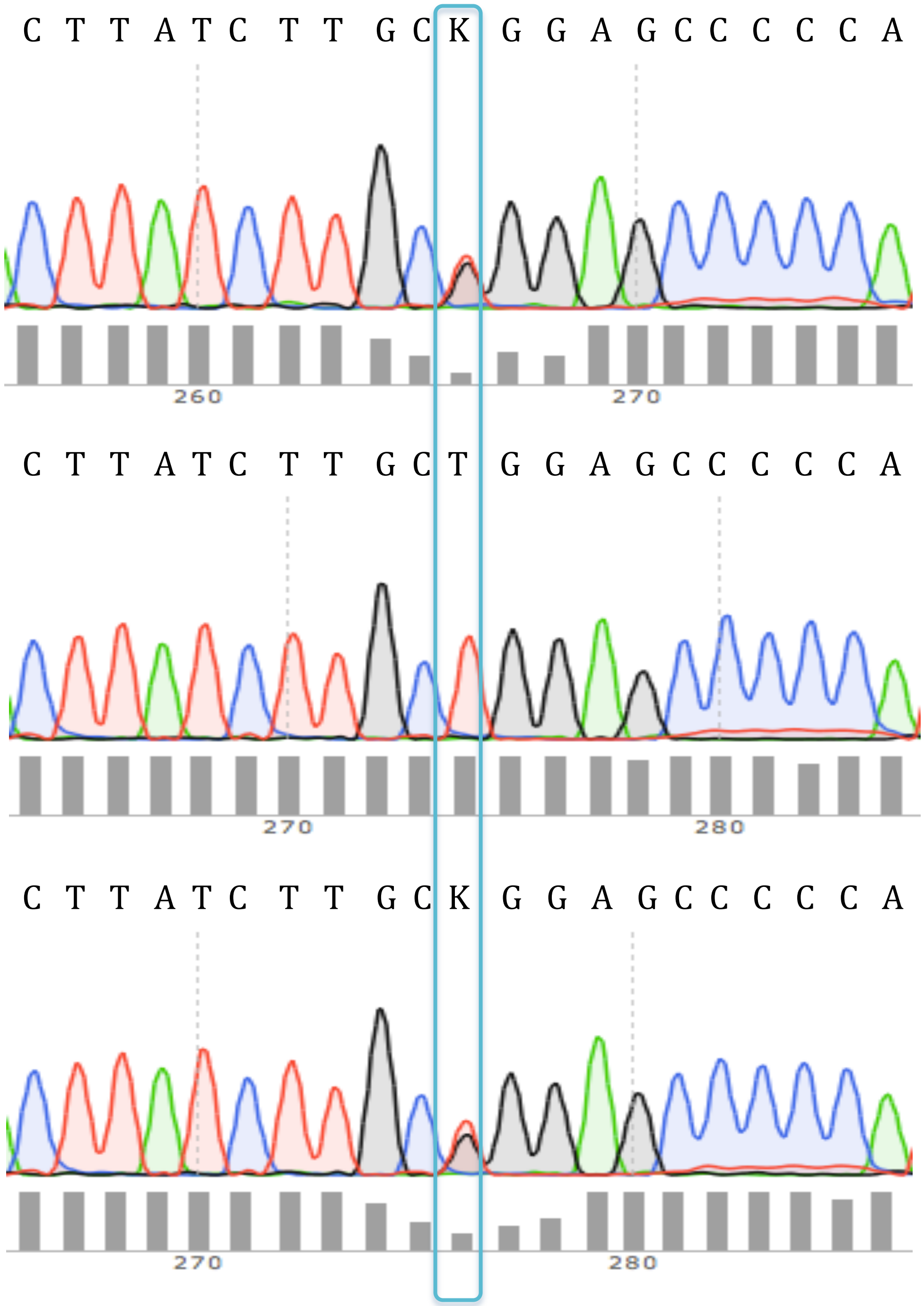
```
Query  1    GTAGGTTATGAGAAGGTGGAA  21
            |||||||||| ||||||||||
Sbjct  222  GTAGGTTATGGGAAGGTGGAA  242
```



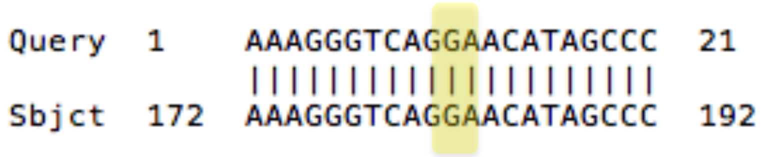
ID	CHR	POSITION	FATHER	MOTHER	CHILD
37	chr19	19774267	C/A*	A/A*	C/A*



```
Query  1    TGGGGGCTCCCGCAAGATAAG  21
            ||||||||||| ||||||||||
Sbjct  276  TGGGGGCTCCAGCAAGATAAG  256
```



ID	CHR	POSITION	FATHER	MOTHER	CHILD
42	chr21	41715615	G/C	G/G	G/C



```
Query  1    AAAGGGTCAGGAACATAGCCC  21
            |||||||||||||||||||||
Sbjct  172  AAAGGGTCAGGAACATAGCCC  192
```

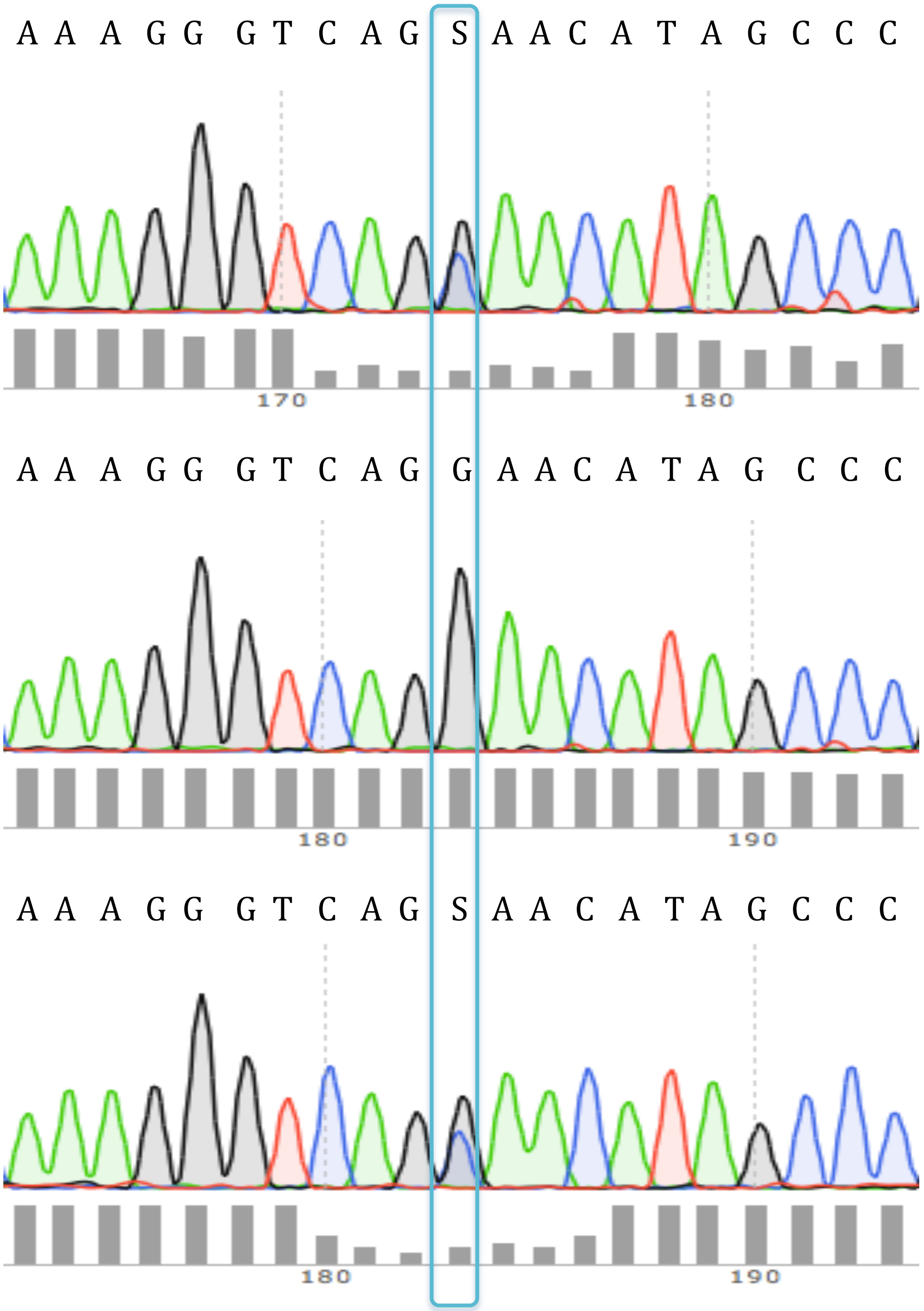


Fig. S3. CHROMATOGRAMS FOR EACH EXPERIMENTALLY VALIDATED MENDELIAN SNV. Only 5 SNVs designated as mendelian by COBASI were chosen at random. For every SNV, a table with the chromosome and genomic coordinates, along with the genotype designated for every trio individual is shown. Also, an alignment between the HRG and the father Sanger sequence is shown. This alignment is centered on the SNV position, which is highlighted in a yellow rectangle. The Sanger sequence chromatograms for that genomic region are shown for the father, mother, and child, and the SNV position is highlighted in a blue rectangle. The chromatogram quality metrics are illustrated as bars at the bottom of each chromatogram.

SUPPORTING REFERENCES

1. Langmead B, Trapnell C, Pop M Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 485(7397):237-241.
2. Huang W, Li L, Myers JR, Marth GT (2012) ART: A next-generation sequencing read simulator. *Bioinformatics* 28(4):593-594.
3. Van der Auwera GA, Carneiro MO, Hartl C, *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics.11(1110):11.10.1-11.10.33.
4. Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ. (1990) Basic local alignment search tool. *J Mol Biol.* 215:403-410.
5. Li B, et al. (2012) A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet* 8(10):e1002944.
6. Michaelson JJ, et al. (2012). Whole genome sequencing in autism identifies hotspots for de novo germline mutation. Cell 151(7):1431-1442.
7. Sanders SJ, et al. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237-241.