Supplementary Information for
Integrating mapping, assembly and haplotype-based approaches for calling variants in clinical sequencing applications

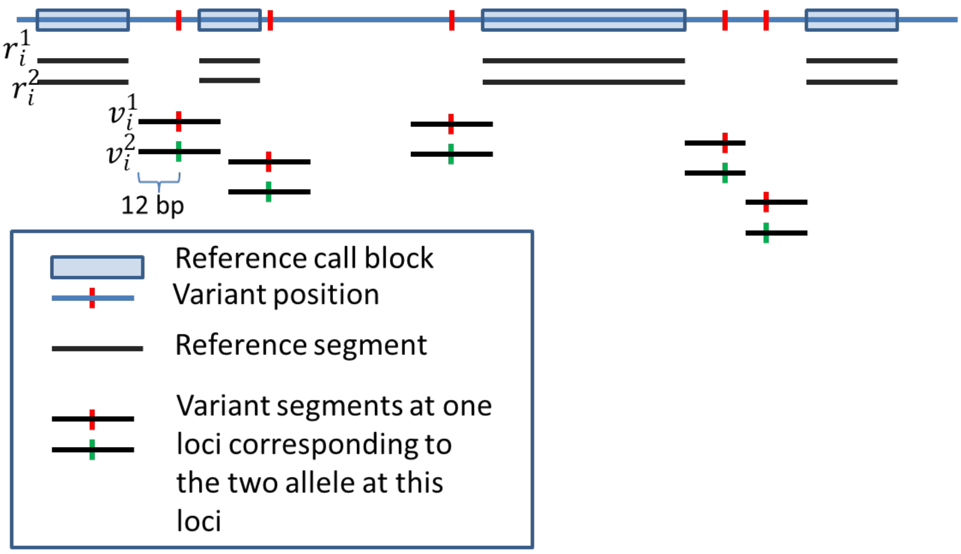
Andy Rimmer[§][†], Hang Phan[§][†], Iain Mathieson[†], Zamin Iqbal[†], Stephen R. F. Twigg[*], WGS500 Consortium[¶], Andrew O. M. Wilkie[*], Gil McVean[†][°], Gerton Lunter[†].

[§]These authors contributed equally to this paper.

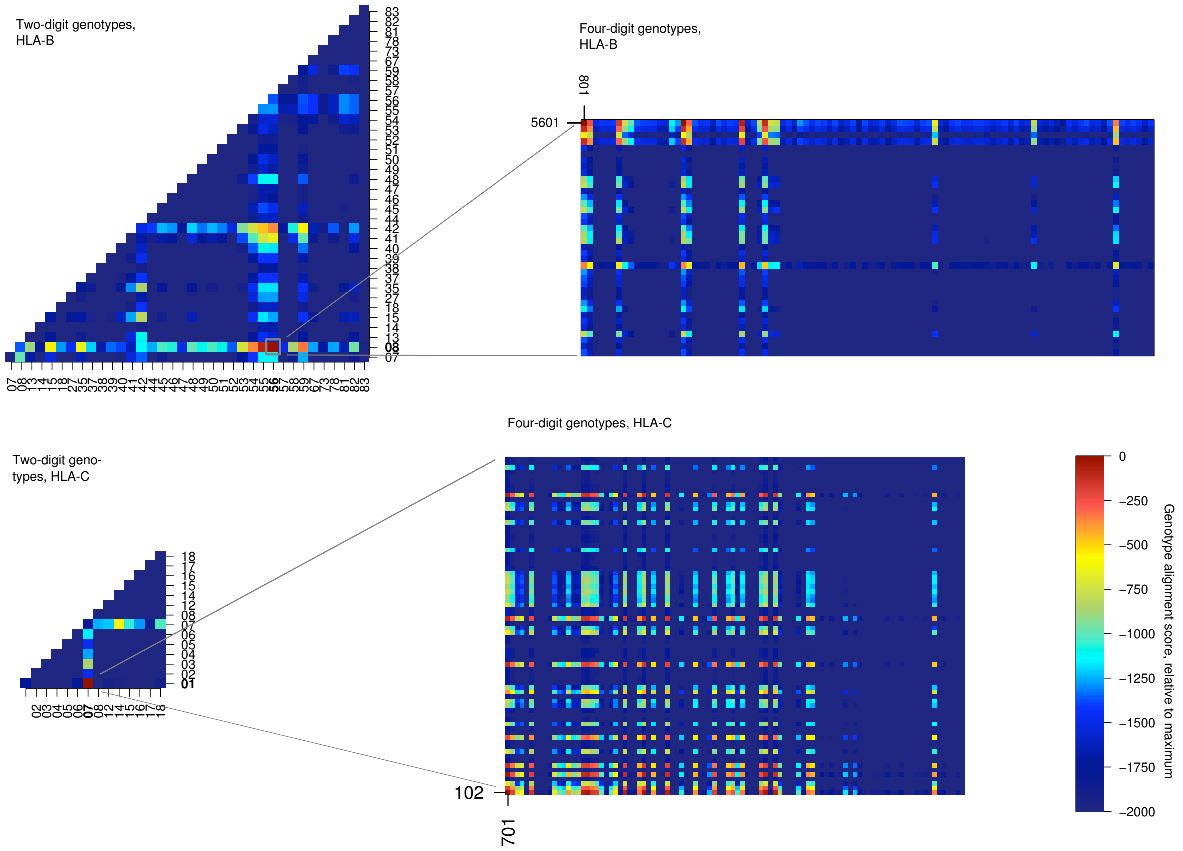
Affiliations:

[†]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX1 3BN, UK
[*]Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK.
[°]Department of Statistics, University of Oxford, South Parks Road, Oxford OX1 3TG, UK.
[¶]A list of members and affiliations is provided in section 6 of this Supplementary Note.

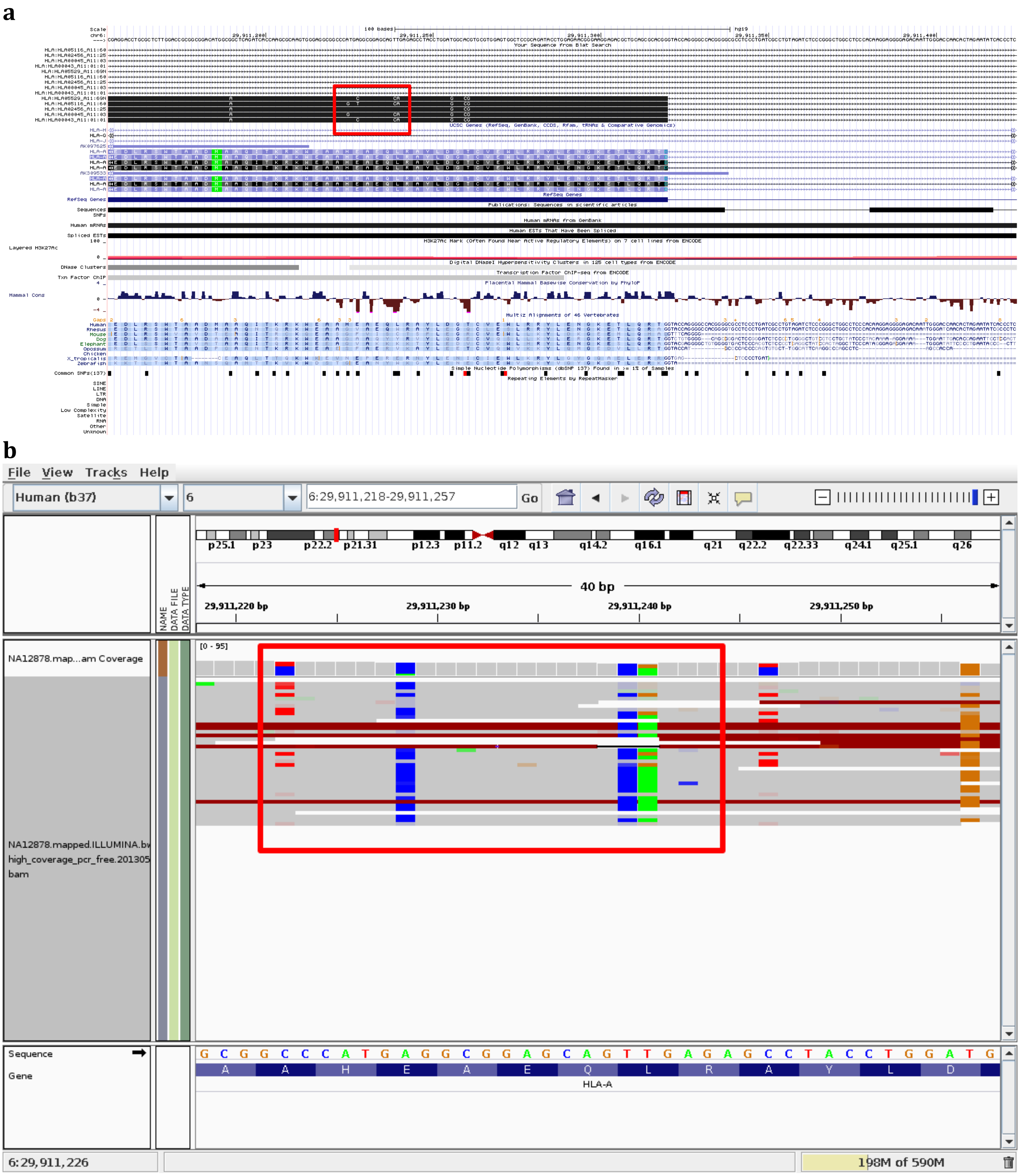
Supplementary Figures



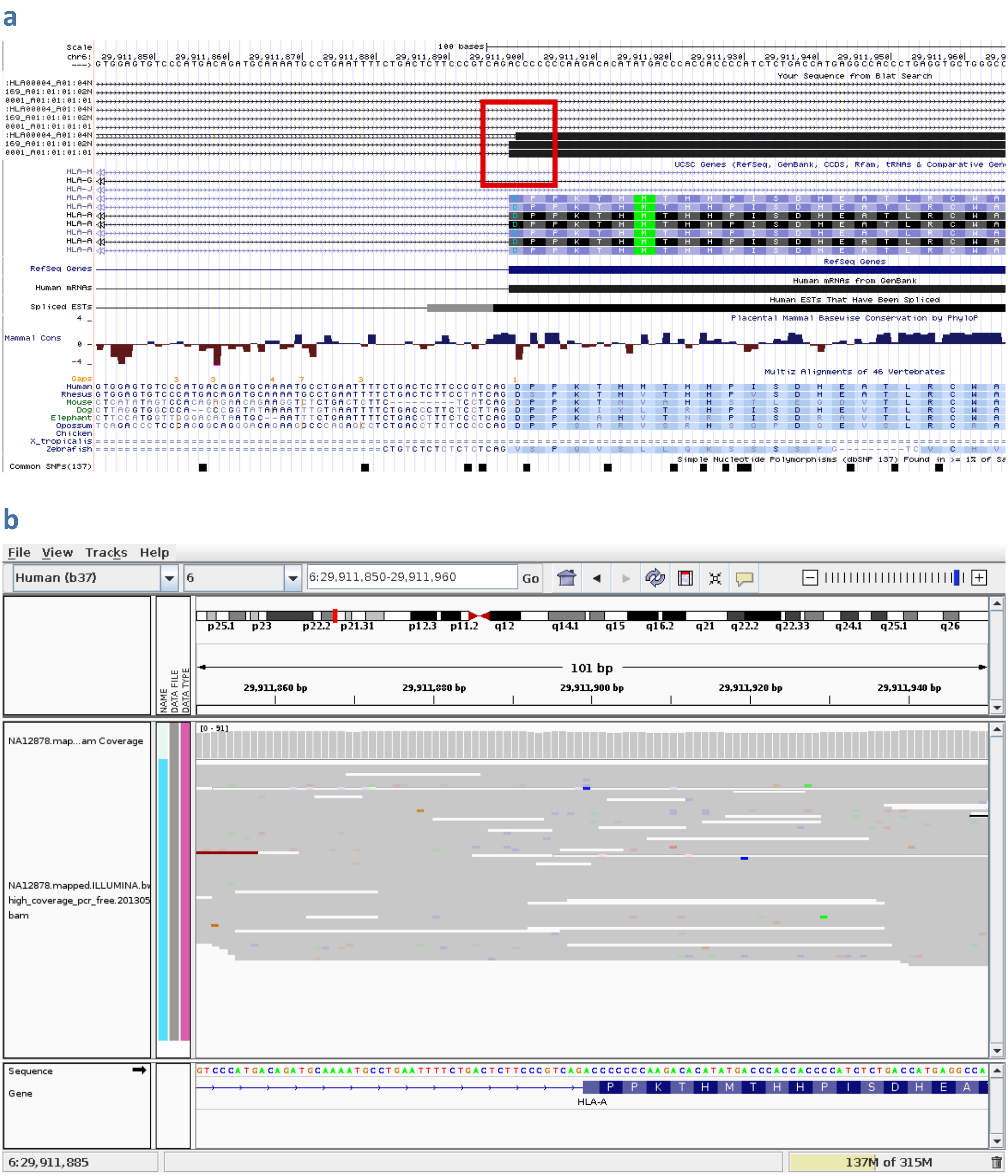
Supplementary Figure 1: Generation of short haplotypes from Platypus variant and reference calls.



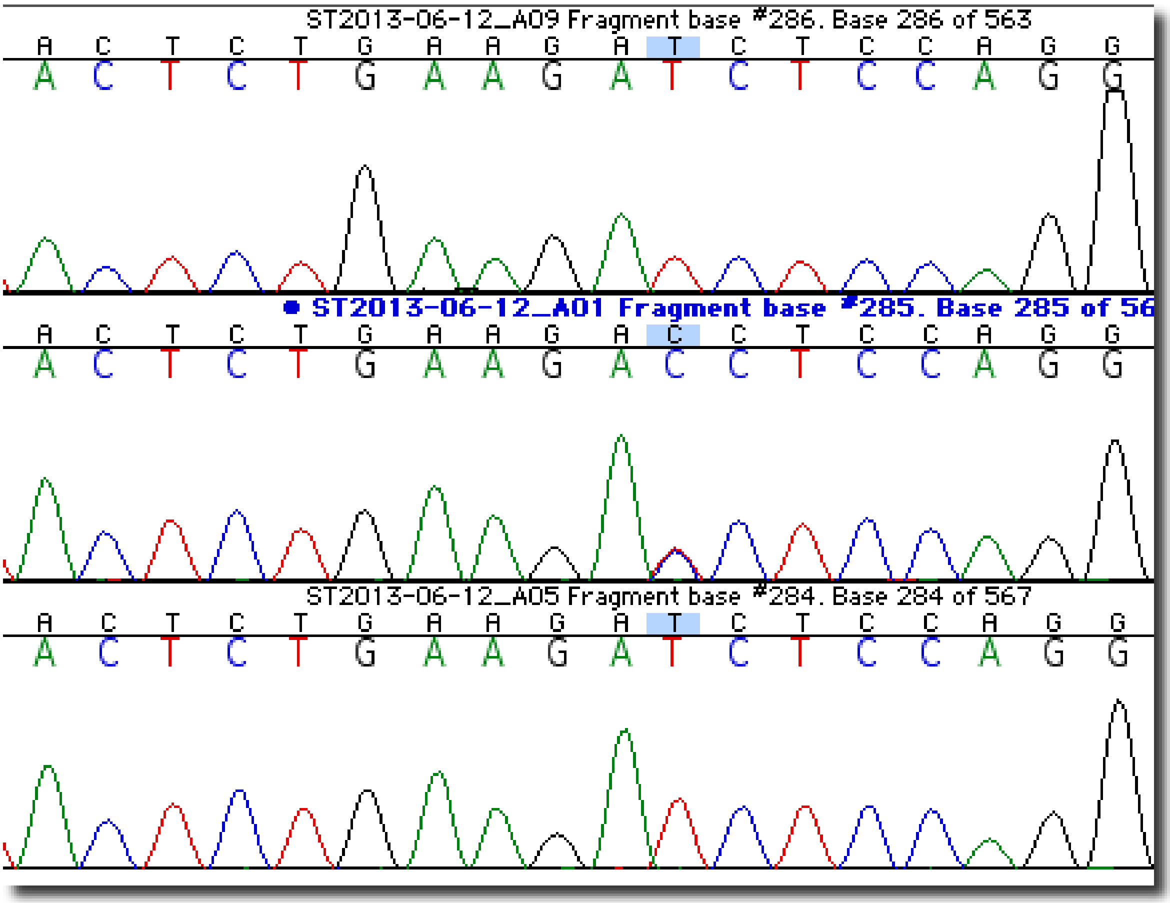
Supplementary Figure 2 Estimated genotypes at the *HLA-B* and *HLA-C* loci for sample NA12878. The two-digit heatmaps recapitulate those of Fig. 3 in the main paper; the four-digit heatmaps show the support for the four-digit genotypes consistent with the two-digit genotype, with the unique best scoring genotype labeled on the axes. In both cases the best scoring genotype is identical to the genotype established using laboratory typing.


a
Scale
chr6:
100 bases
29,911,200
29,911,250
29,911,300
29,911,350
29,911,400
hg19
Your Sequence from Blat Search
HLA:HLA05116_A11:60
HLA:HLA02456_A11:25
HLA:HLA00045_A11:03
HLA:HLA00043_A11:01:01
HLA:HLA05529_A11:69N
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
HLA-H
HLA-G
HLA-J
AK097625
HLA-A
AK309533
RefSeq Genes
Sequences
SNPs
Human mRNAs
Spliced ESTs
Layered H3K27Ac
DNase Clusters
Txn Factor ChIP
Mammal Cons
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Publications: Sequences in scientific articles
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
b
File View Tracks Help
Human (b37)
6
6:29,911,218-29,911,257
Go
p25.1 p23 p22.2 p21.31 p12.3 p11.2 q12 q13 q14.2 q16.1 q21 q22.2 q22.33 q24.1 q25.1 q26
40 bp
29,911,220 bp
29,911,230 bp
29,911,240 bp
29,911,250 bp
NAME
DATA FILE
DATA TYPE
NA12878.map...am Coverage
[0 - 95]
NA12878.mapped.ILLUMINA.bw
high_coverage_pcr_free.201305
bam
Sequence
Gene
G C G G C C C A T G A G G C G G A G C A G T T G A G A G C C T A C C T G G A T G
A A H E A E Q L R A Y L D
HLA-A
6:29,911,226
198M of 590M

Supplementary Figure 3. Ambiguous alleles and supporting reads for allele *HLA-A**11:01. Reads supporting the correct allele A*11:01 exist, but the corresponding SNVs (in red box) were not called due to the existence of reads with aberrant insert sizes (denoted as dark red lines).

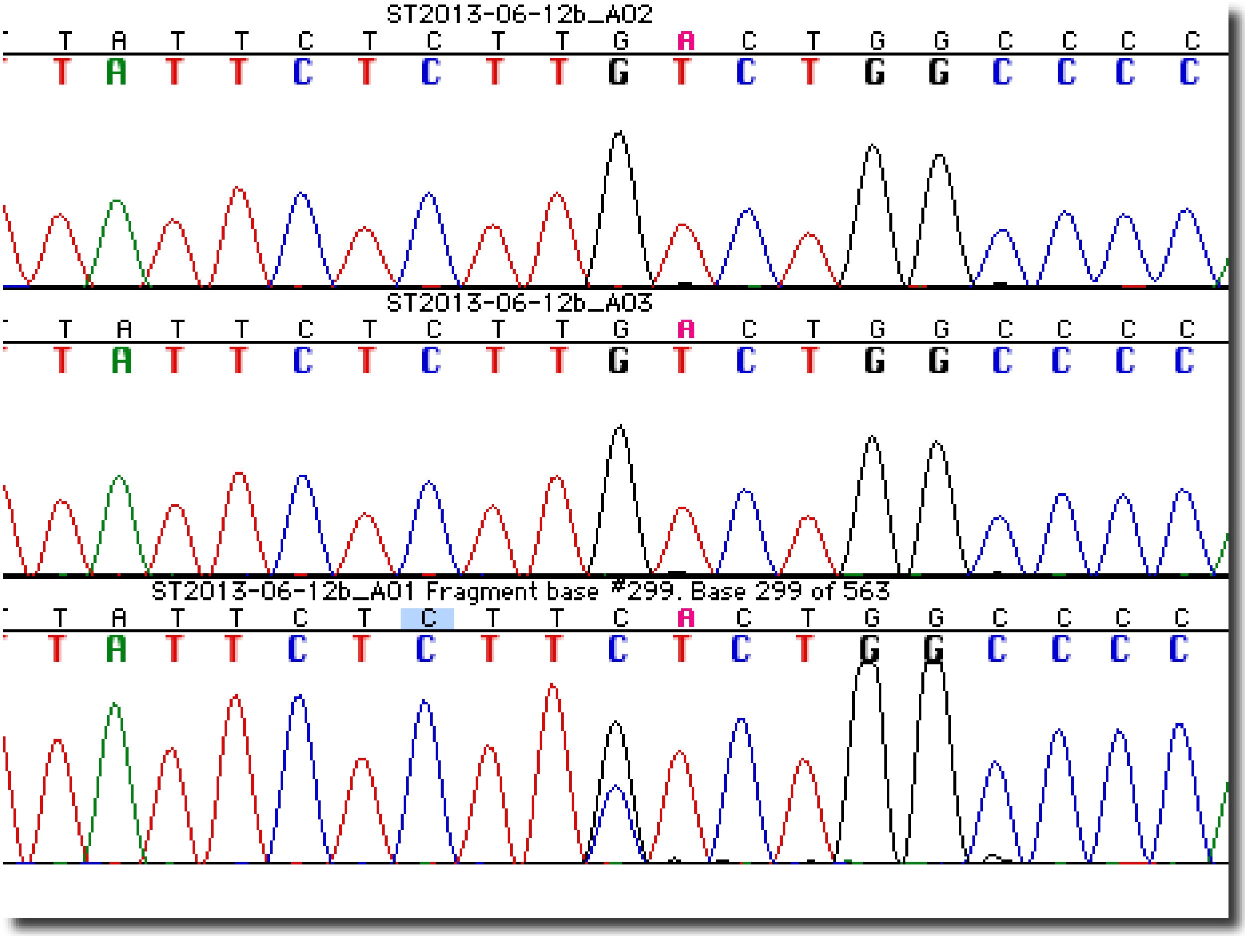


Supplementary Figure 4 Ambiguous alleles and supporting reads for allele *HLA-A**01:01. A length polymorphism in allele A*01:04N (red box) is due to a polymorphic variant 5' of the exon, causing a differential splicing event. While the variant was not called, there are no differences on the exonic portion of the alleles, and consequently the genotyping pipeline did not differentiate between these alleles.



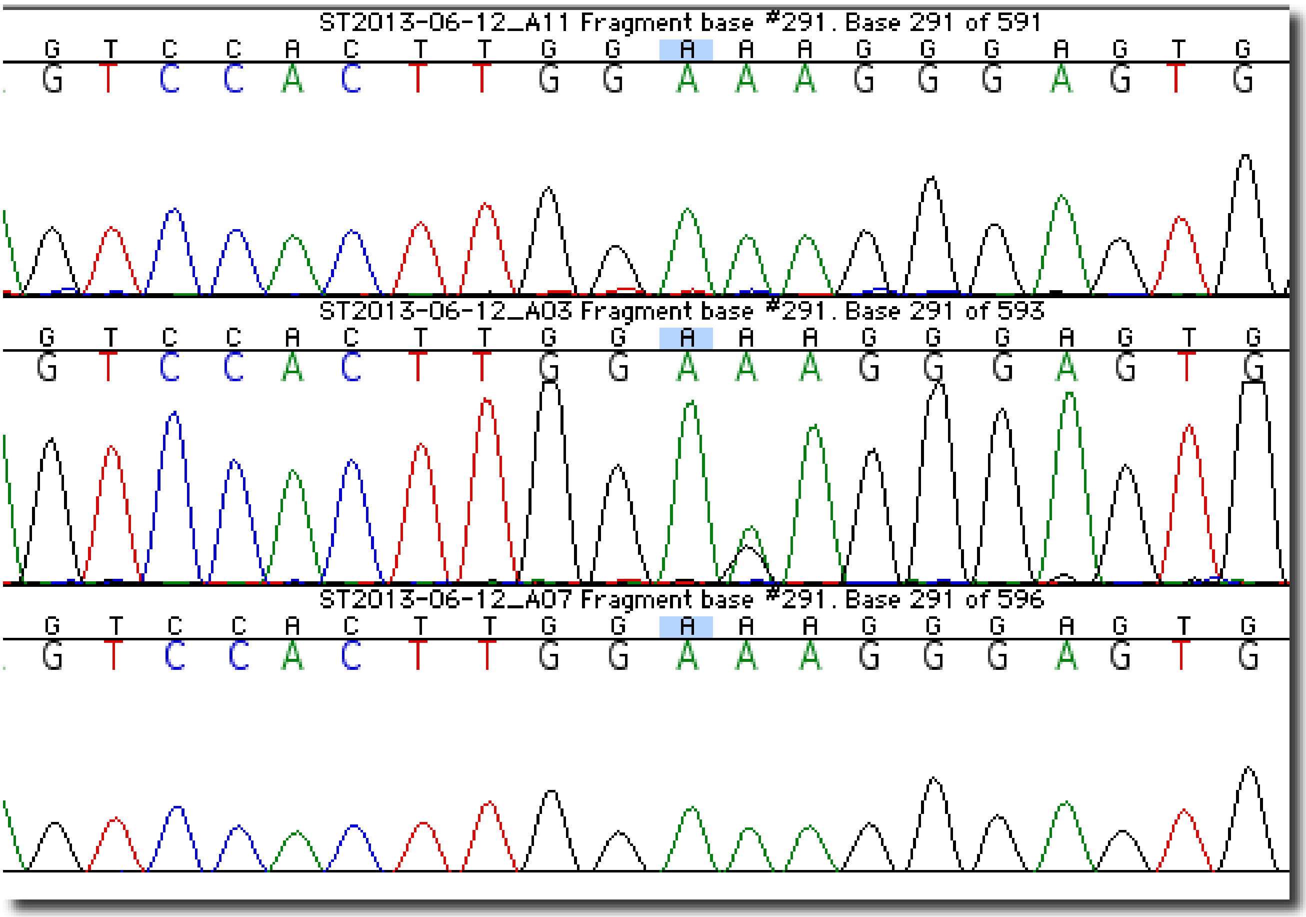
Chr1_6194022_TC_Batch2_A1A5A9.png

Supplementary Figure 5a Validation result for Chr1 position 6194022 (TC)



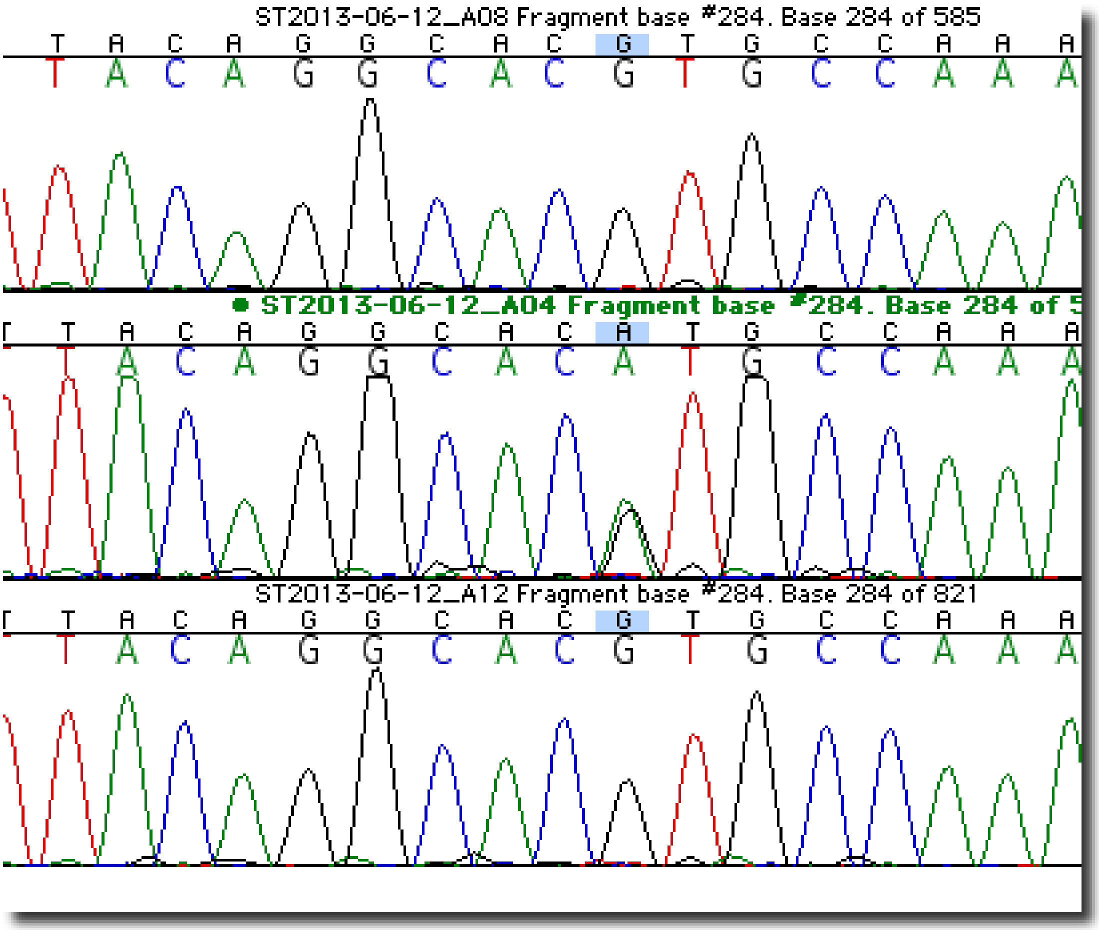
Chr1_185290399_GC_Batch3_A1A2A3.png

Supplementary Figure 5b Validation result for Chr1 position 185290399 (GC)



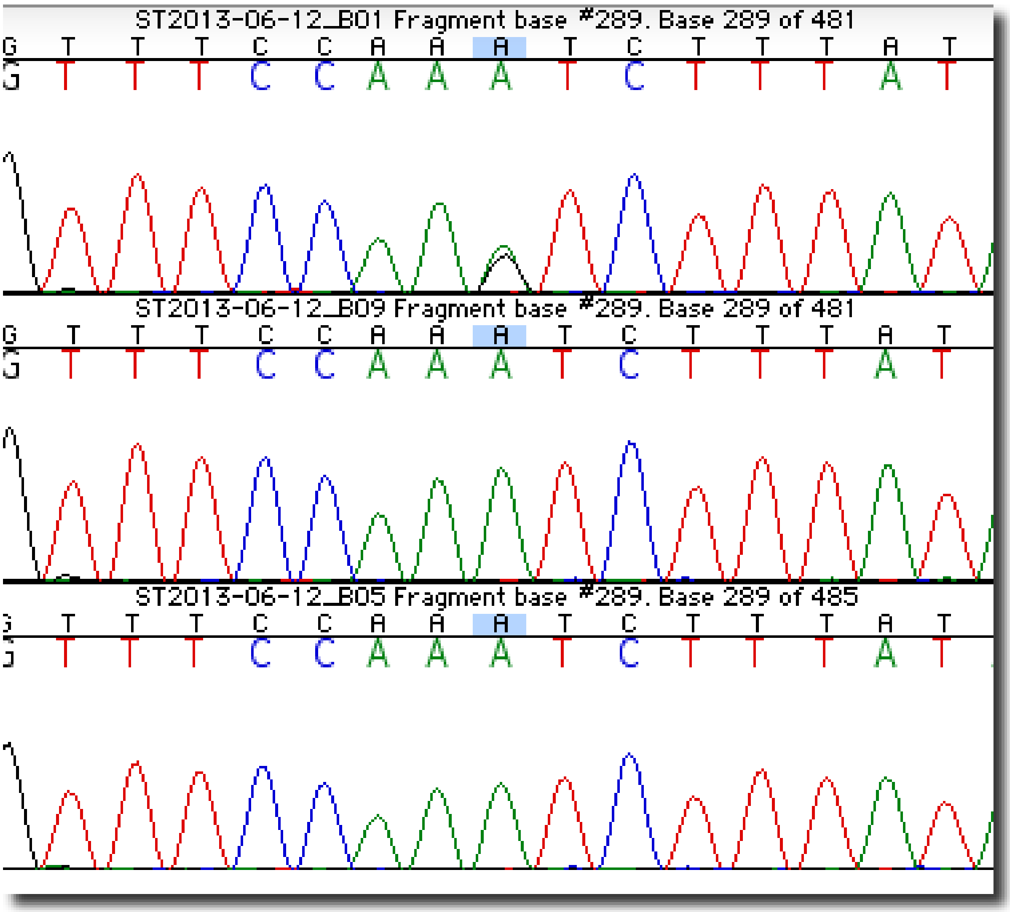
Chr1_212896983_AG_Batch2_A3A7A11.png

Supplementary Figure 5c Validation result for Chr1 position 212896983 (AG)



Chr2_10776159_GA_Batch2_A4A8A12.png

Supplementary Figure 5d Validation result for Chr2 position 10776159 (GA)



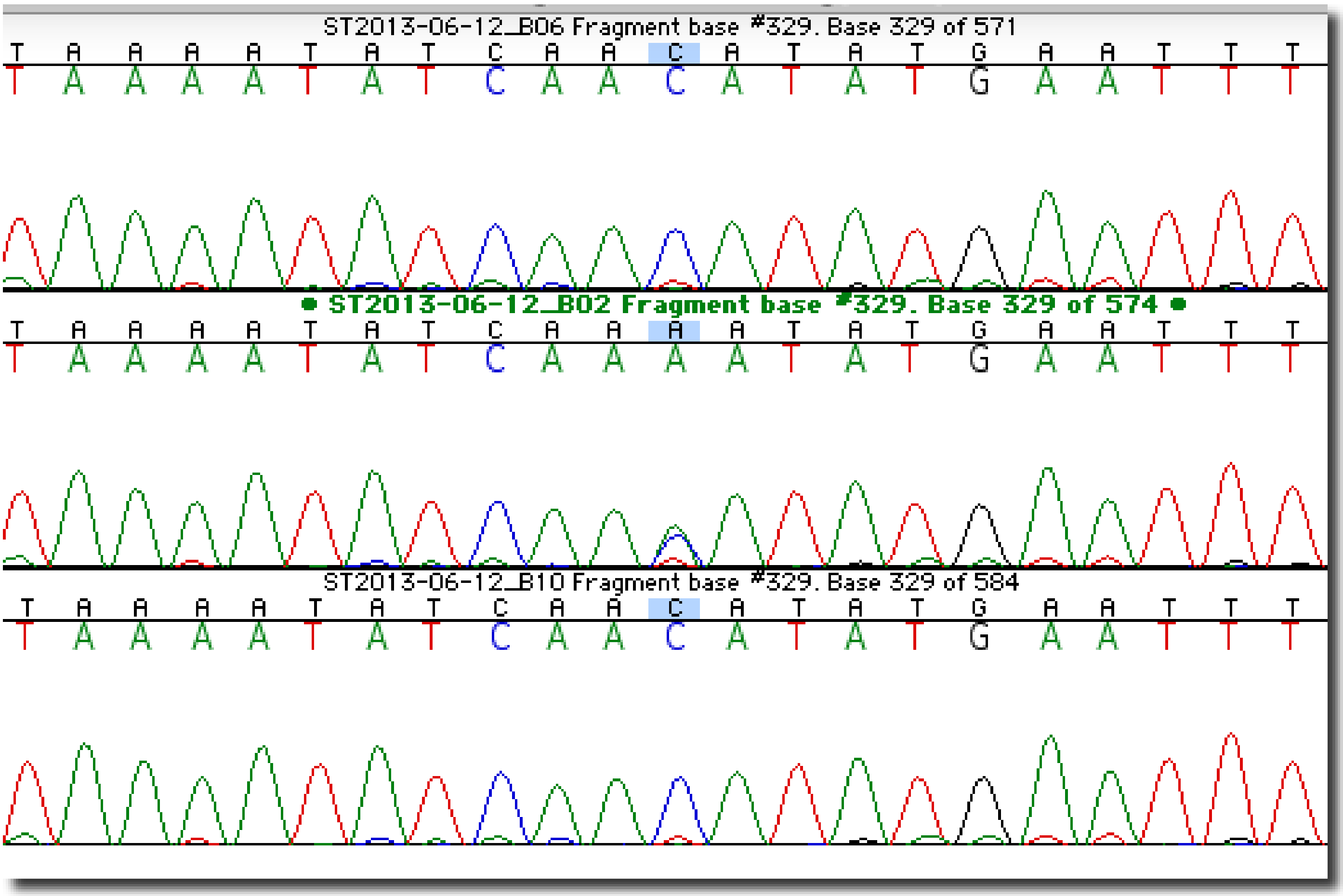
ST2013-06-12_B01 Fragment base #289. Base 289 of 481

ST2013-06-12_B09 Fragment base #289. Base 289 of 481

ST2013-06-12_B05 Fragment base #289. Base 289 of 485

Chr2_40874289_AG_Batch2_B1B5B9.png

Supplementary Figure 5e Validation result for Chr2 position 40874289 (AG)



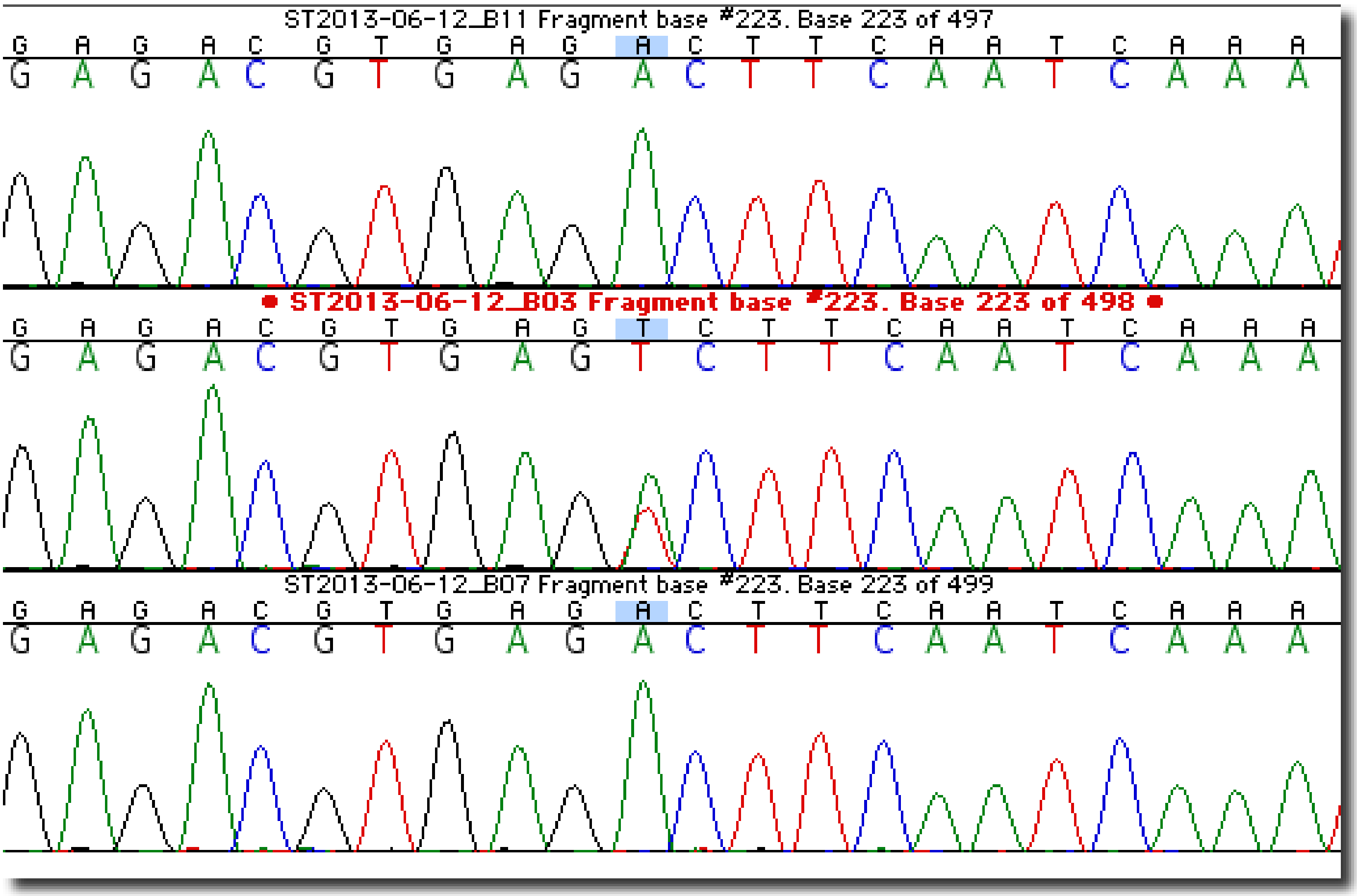
ST2013-06-12_B06 Fragment base #329. Base 329 of 571

● ST2013-06-12_B02 Fragment base #329. Base 329 of 574 ●

ST2013-06-12_B10 Fragment base #329. Base 329 of 584

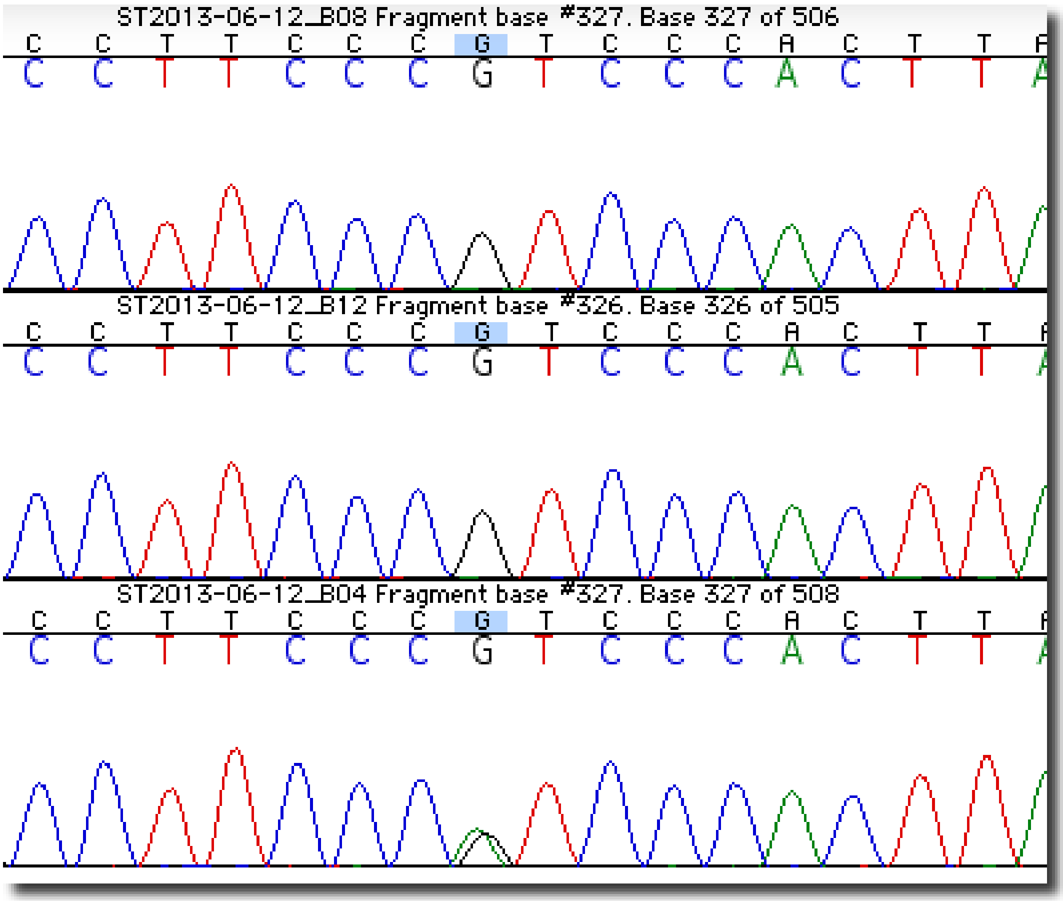
Chr2_41193425_CA_Batch2_B2B6B10.png

Supplementary Figure 5f Validation result for Chr2 position 41193425 (CA)



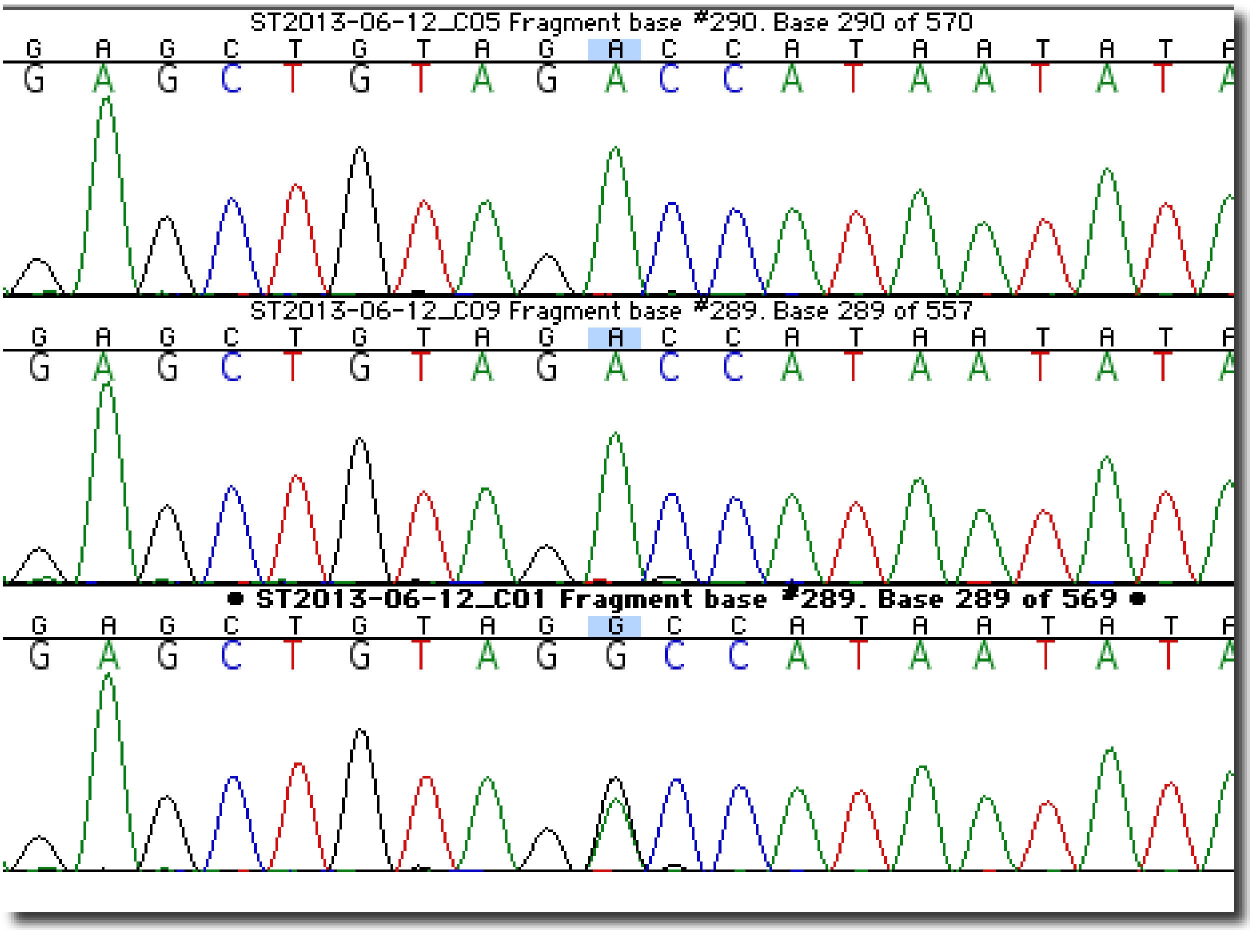
Chr2_64560506_AT_Batch2_B3B7B11.png

Supplementary Figure 5g Validation result for Chr2 position 64560506 (AT)



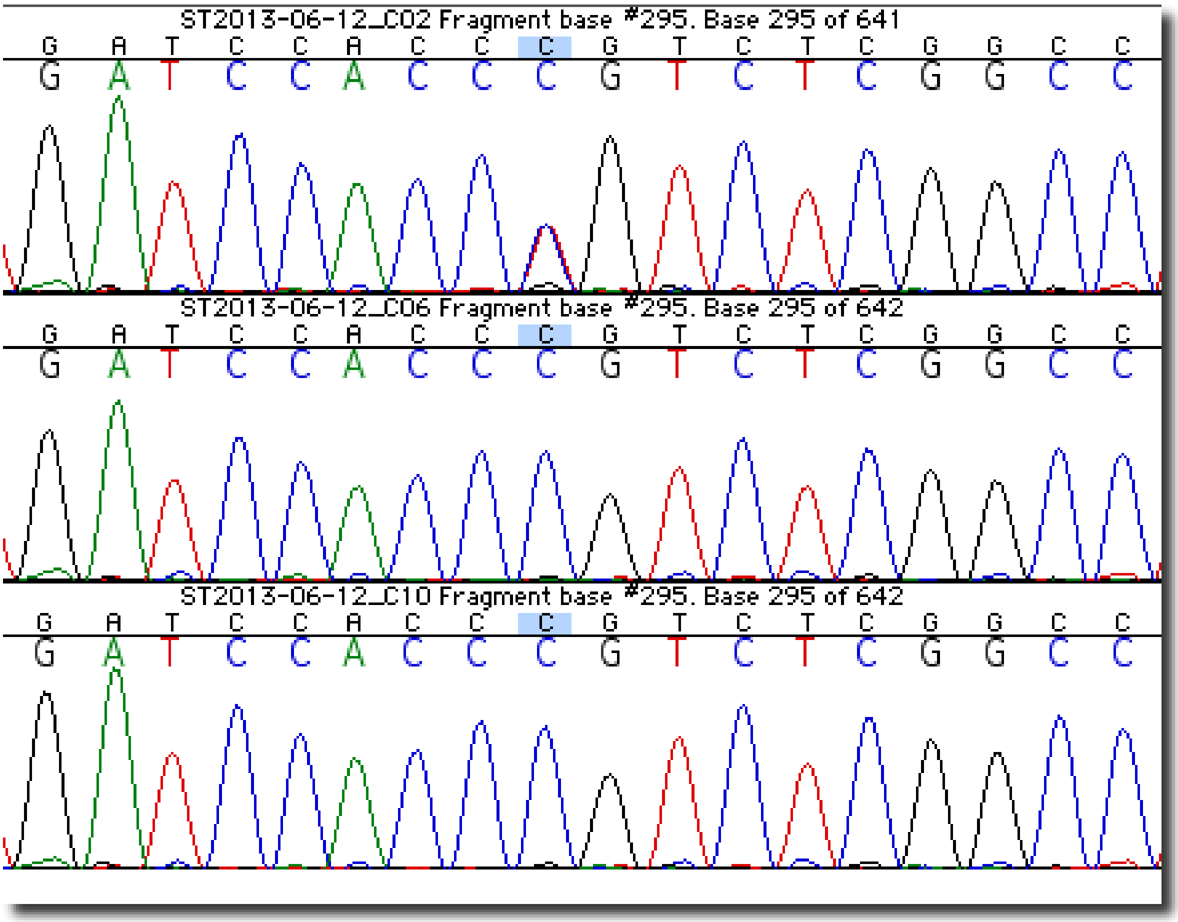
Chr2_85475655_GA_Batch2_B4B8B12.png

Supplementary Figure 5h Validation result for Chr2 position 85475655 (GA)



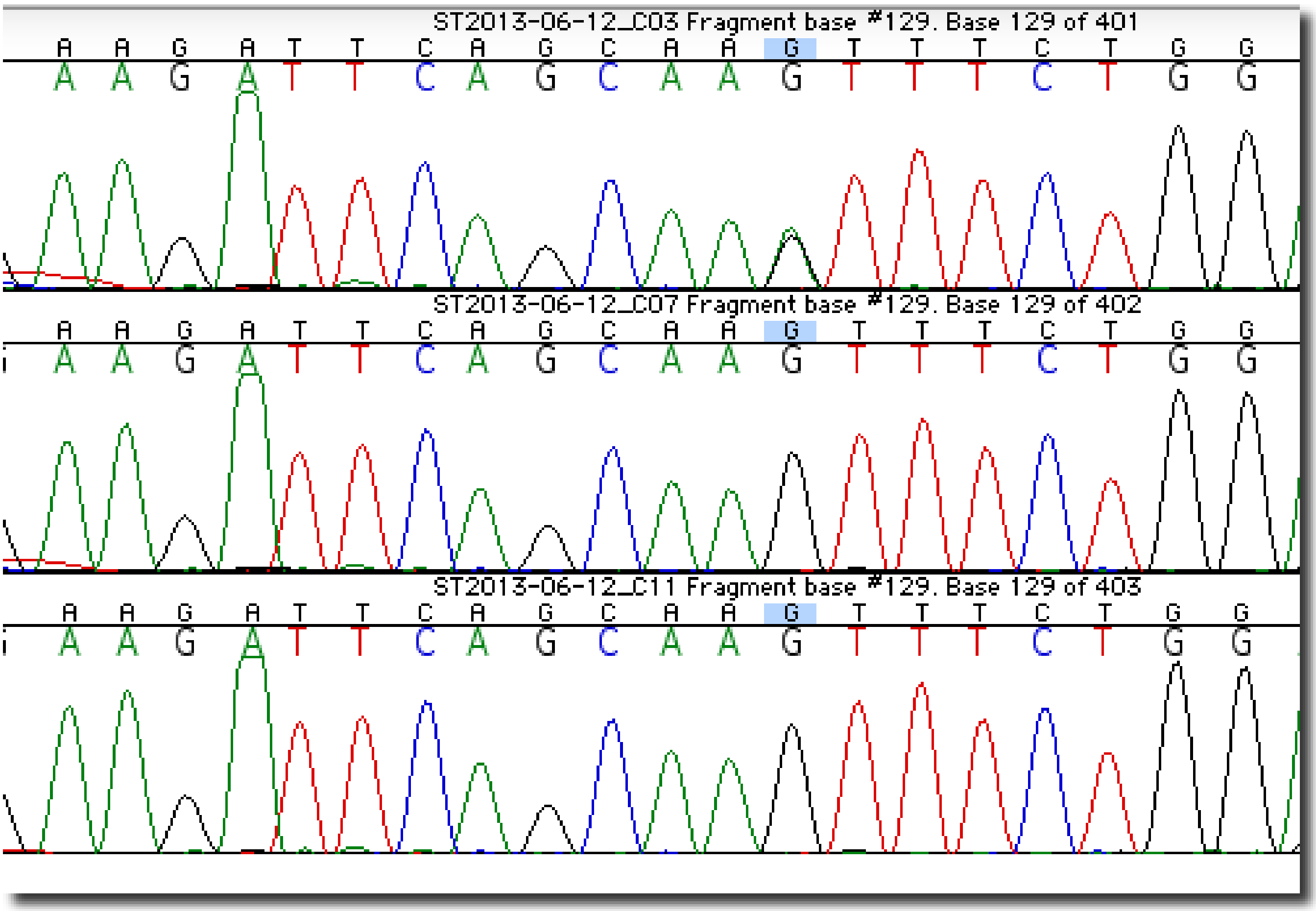
Chr2_139210840_AG_Batch2_C1C5C9.png

Supplementary Figure 5i Validation result for Chr2 position 139210840 (AG)



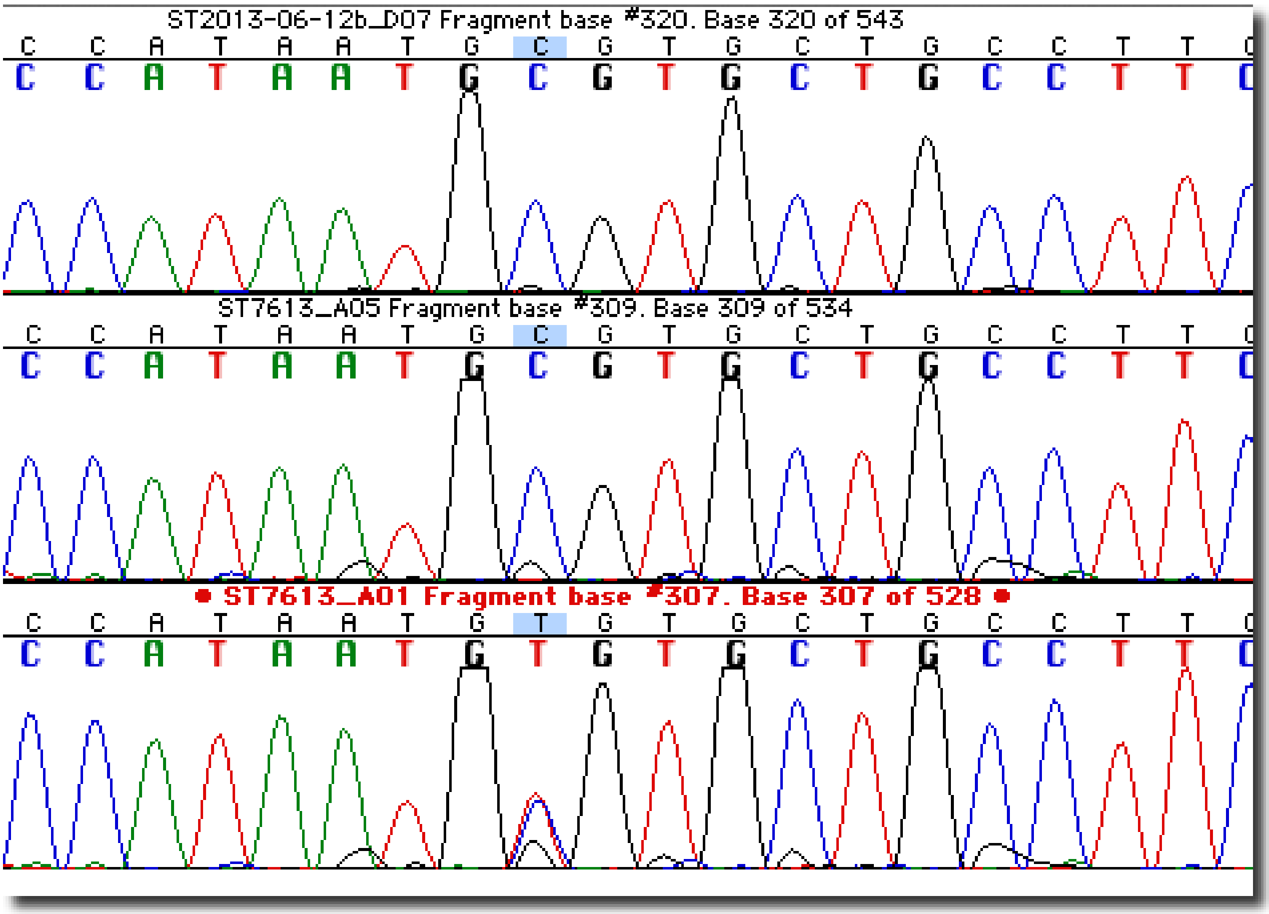
Chr2_144702257_CT_Batch2_C2C6C10.png

Supplementary Figure 5j Validation result for Chr2 position 144702257 (CT)



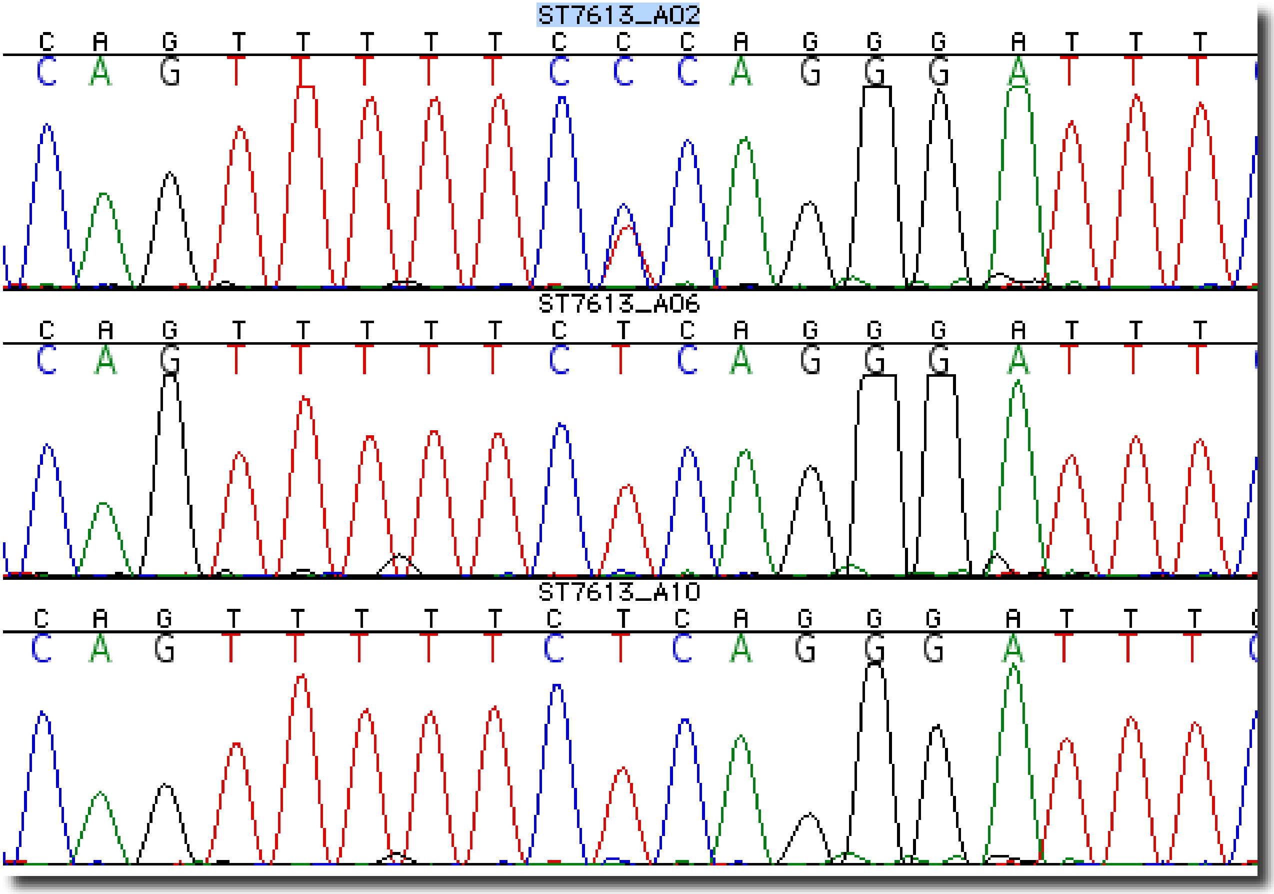
Chr2_148923538_GA_Batch2_C3C7C11.png

Supplementary Figure 5k Validation result for Chr2 position 148923538 (GA)



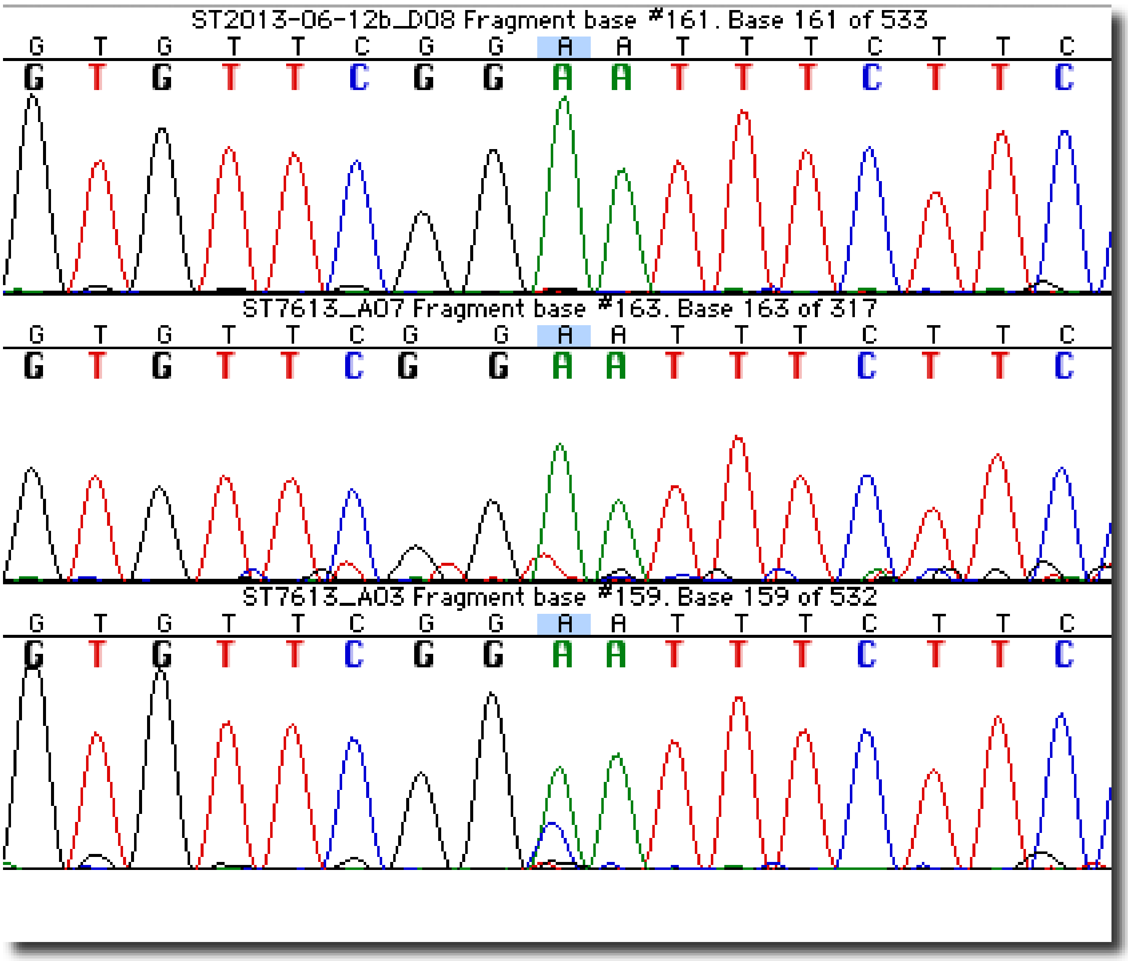
Chr2_159512861_CT_Batch3_D7_Batch1_A1A5.png

Supplementary Figure 51 Validation result for Chr2 position 159512861 (CT)



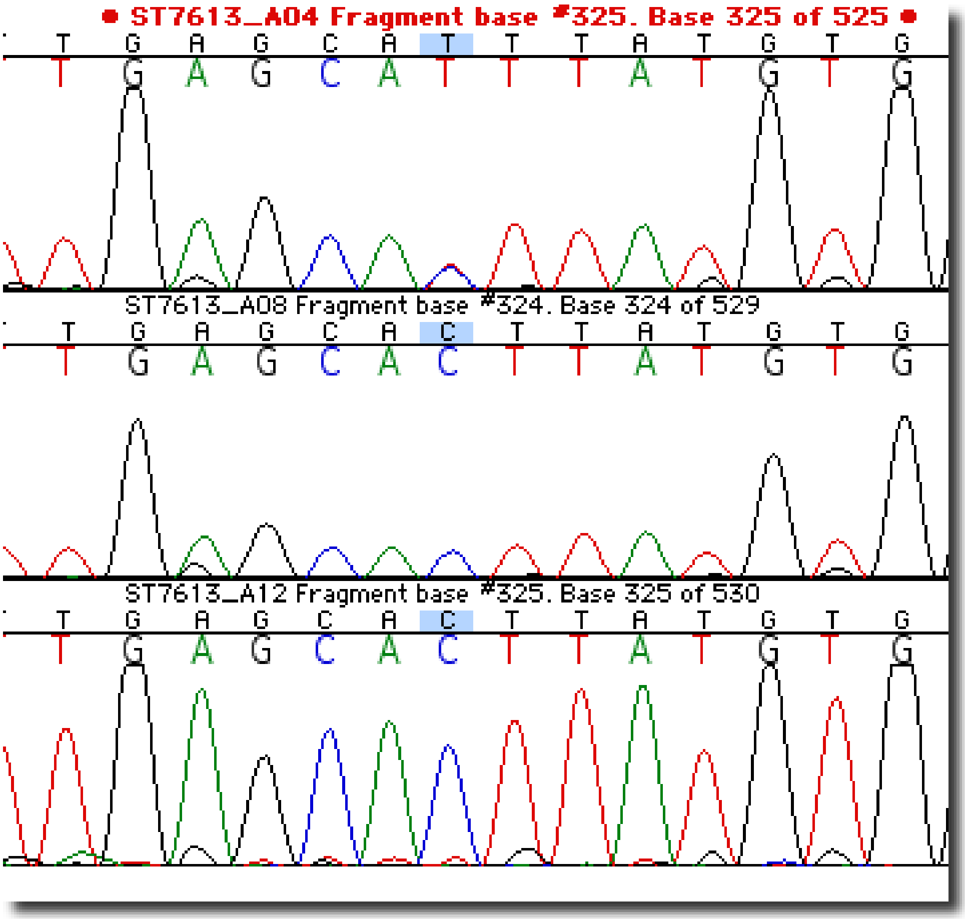
Chr3_1609090_TC_Batch1_A2A6A10.png

Supplementary Figure 5m Validation result for Chr3 position 1609090 (TC)



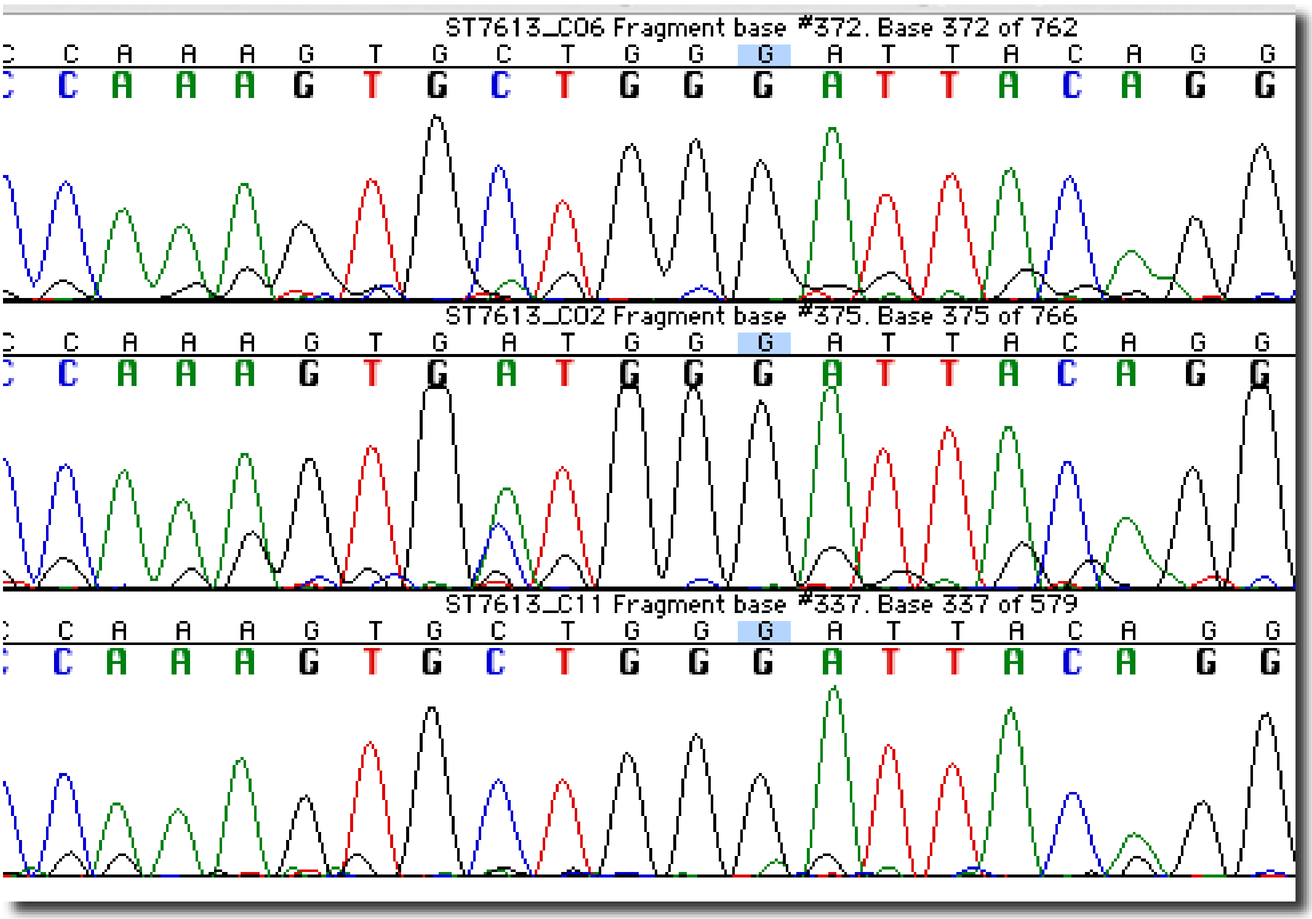
Chr3_152872825_AC_Batch3D8_Batch1A3A7.png

Supplementary Figure 5n Validation result for Chr3 position 152872825 (AC)



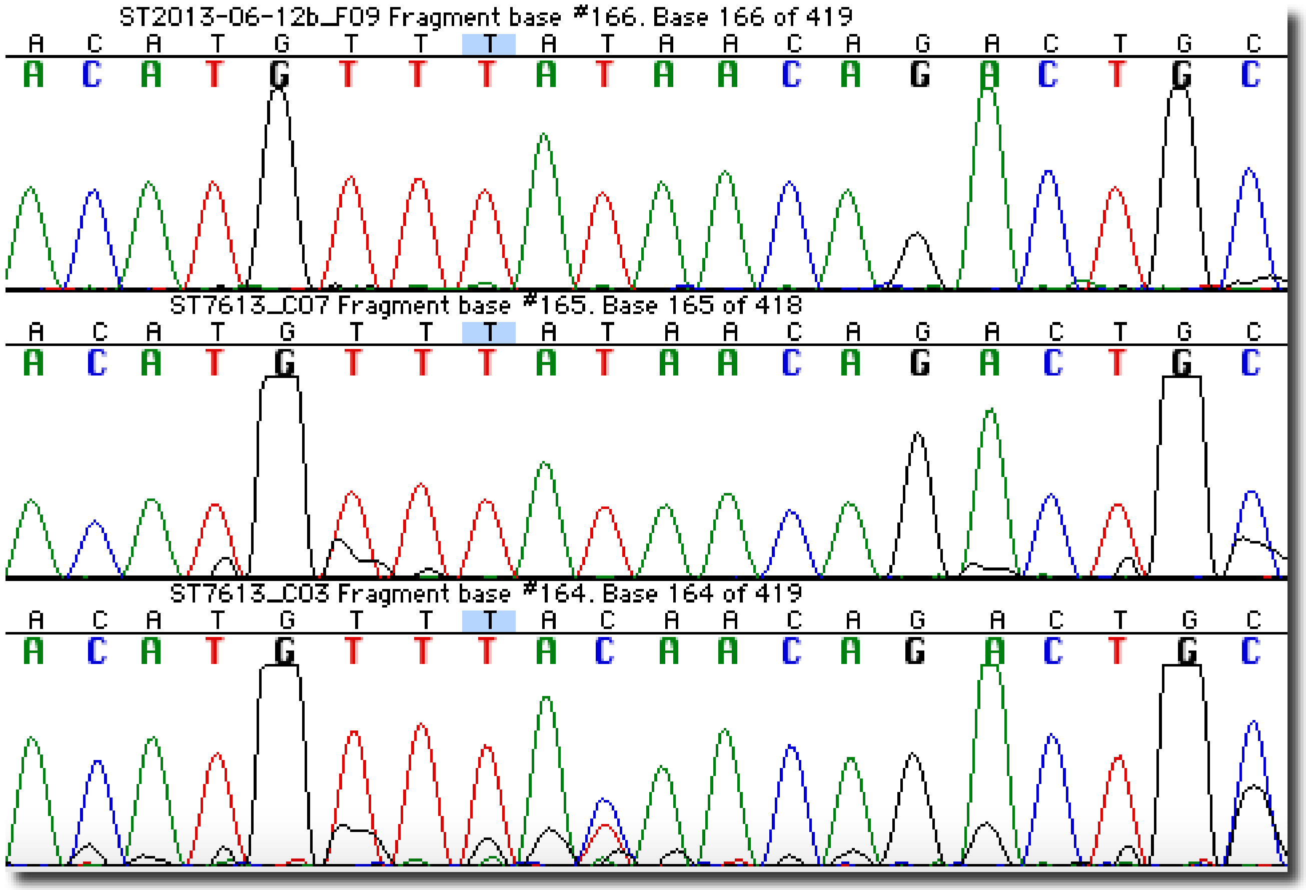
Chr3_183112726_CT_Batch1_A4A8A12.png

Supplementary Figure 5o Validation result for Chr3 position 183112726 (CT)



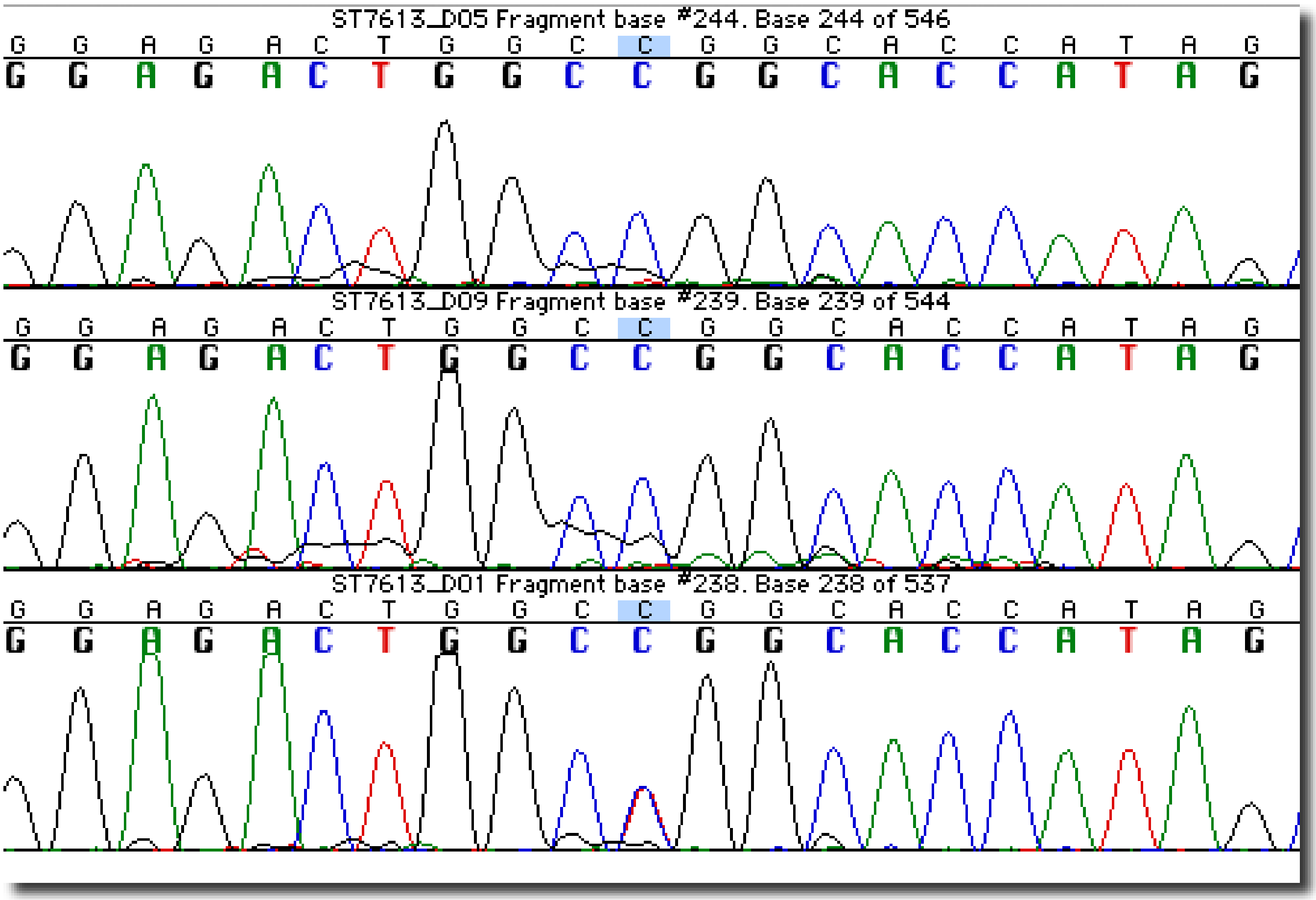
Chr5_5755230_CA_Batch1_C2C6C11.png

Supplementary Figure 5p Validation result for Chr5 position 5755230 (CA)



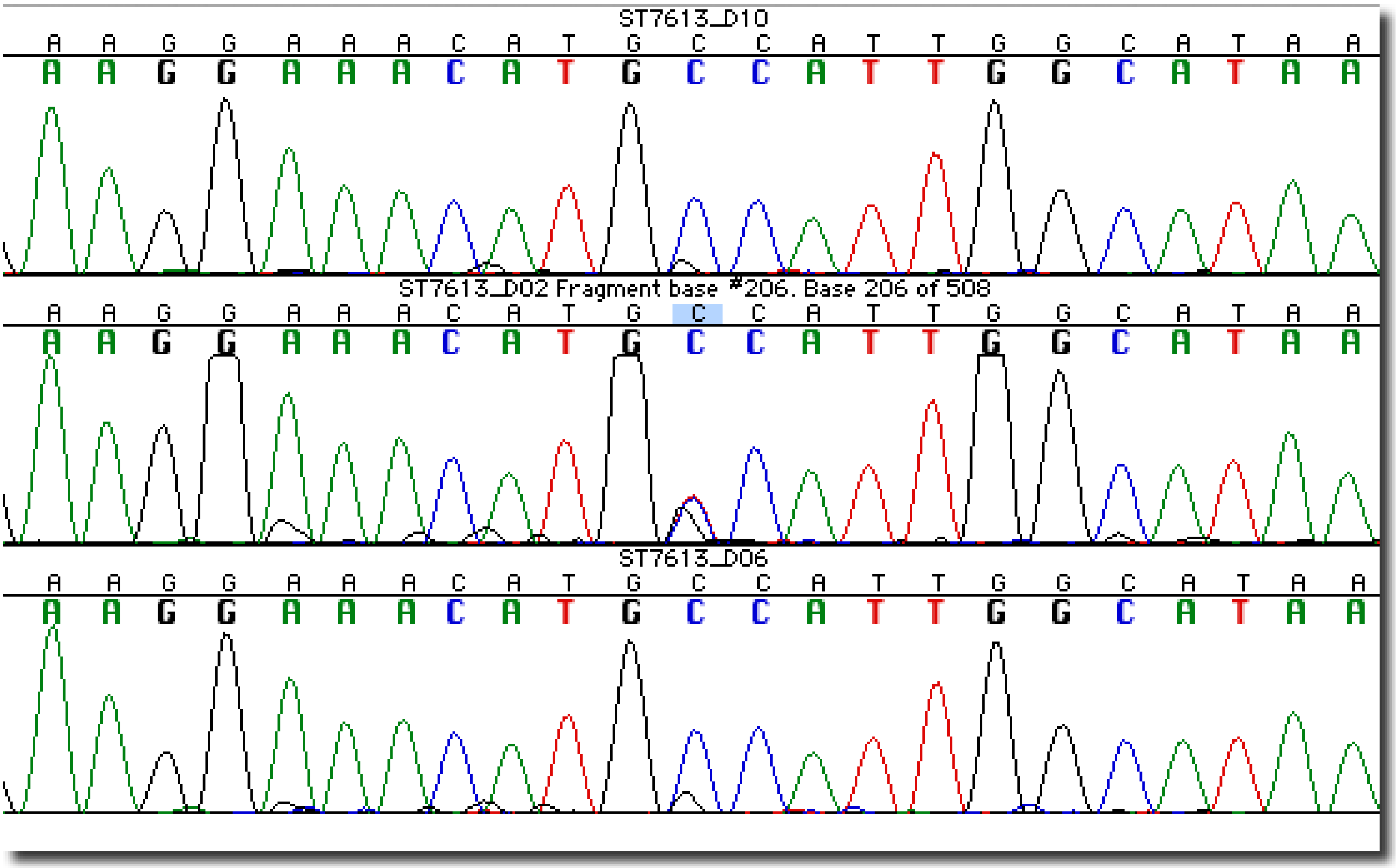
Chr5_108723224_CT_Batch3F9_Batch1C3C7.png

Supplementary Figure 5q Validation result for Chr5 position 108723224 (CT)



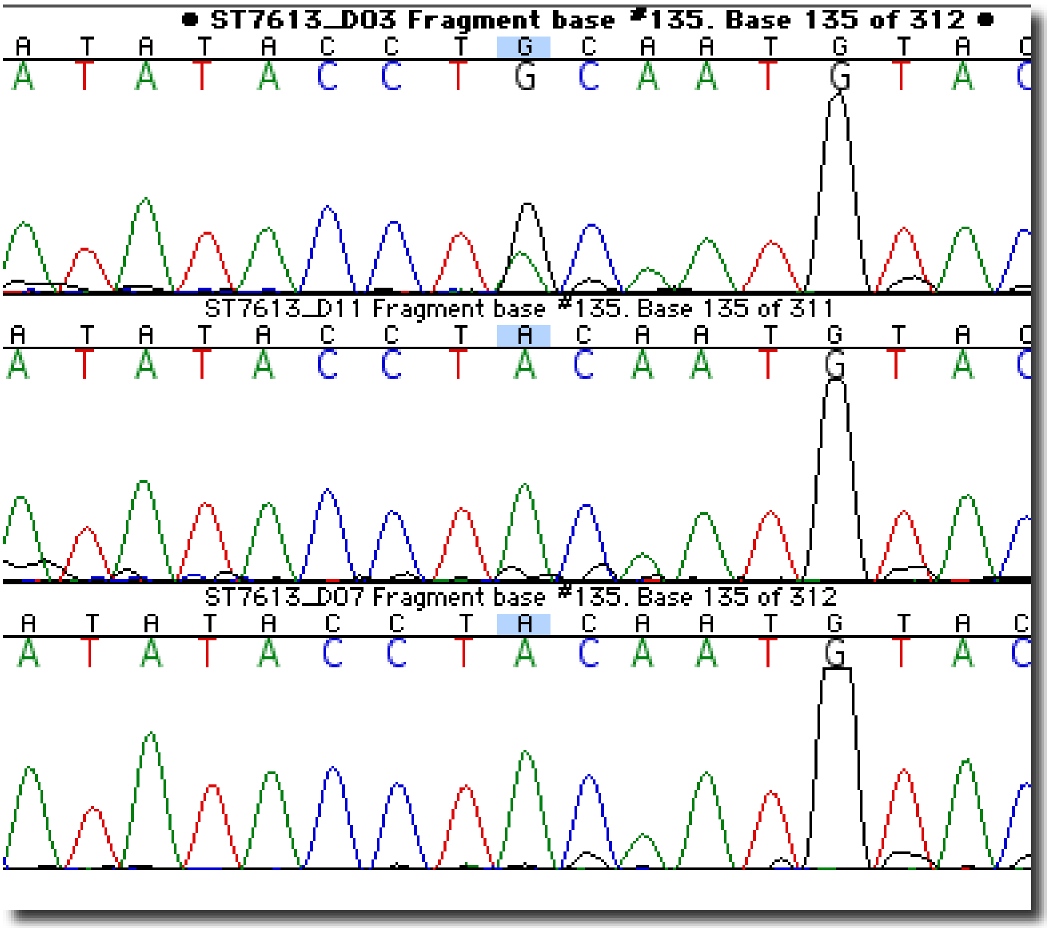
Chr5_122635034_Batch1_CT_D1D5D9.png

Supplementary Figure 5r Validation result for Chr5 position 122635034 (CT)



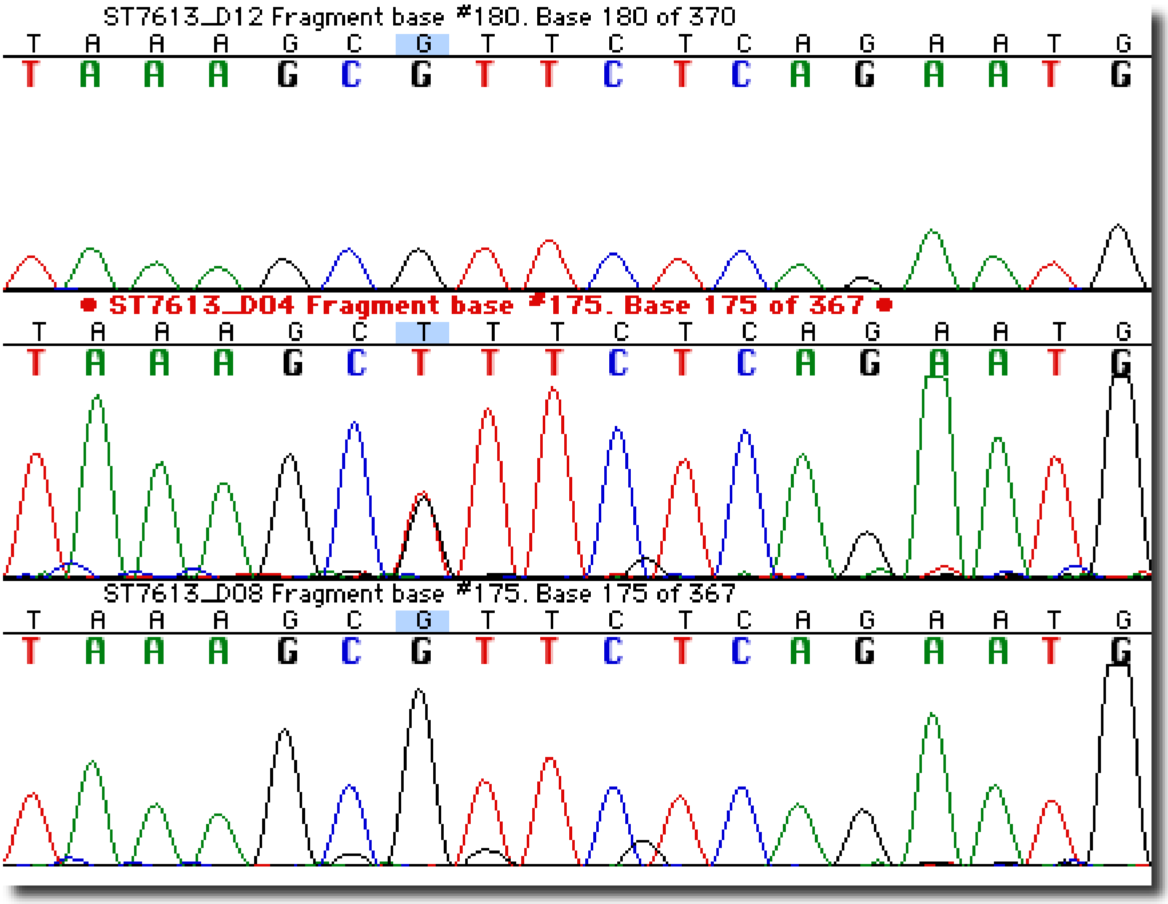
Chr6_51217783_CT_Batch1_D2D6D10.png

Supplementary Figure 5s Validation result for Chr6 position 51217783 (CT)



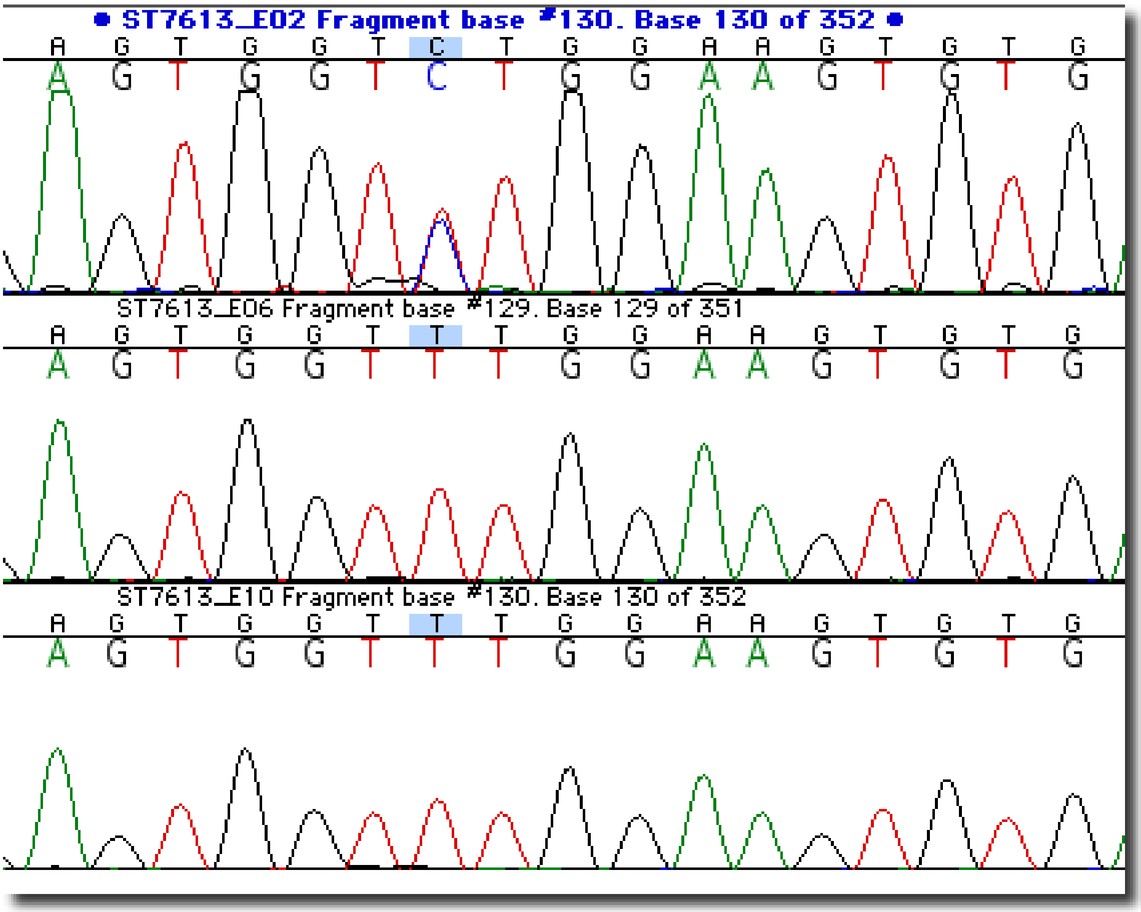
Chr6_53429817_AG_Batch1_D3D7D11.png

Supplementary Figure 5t Validation result for Chr6 position 53429817 (AG)



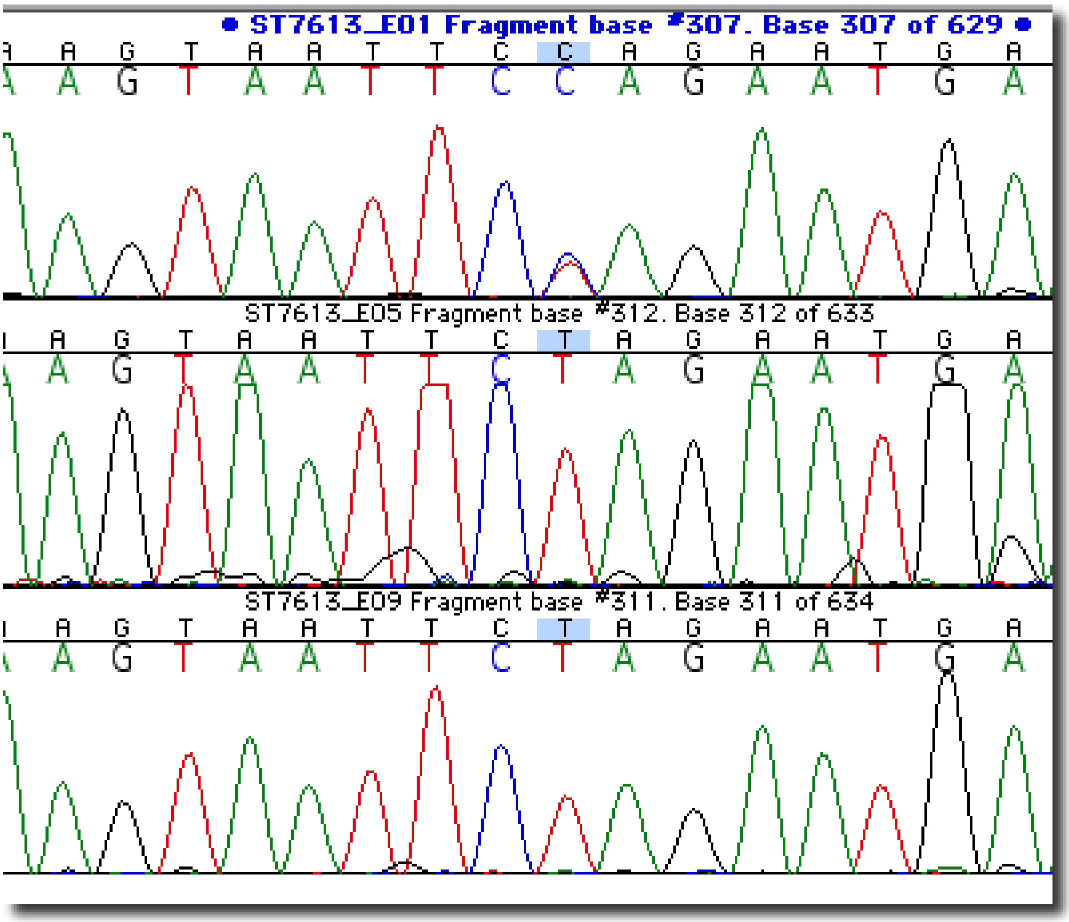
Chr6_87114844_GT_Batch1_D4D8D12.png

Supplementary Figure 5u Validation result for Chr6 position 87114844 (GT)



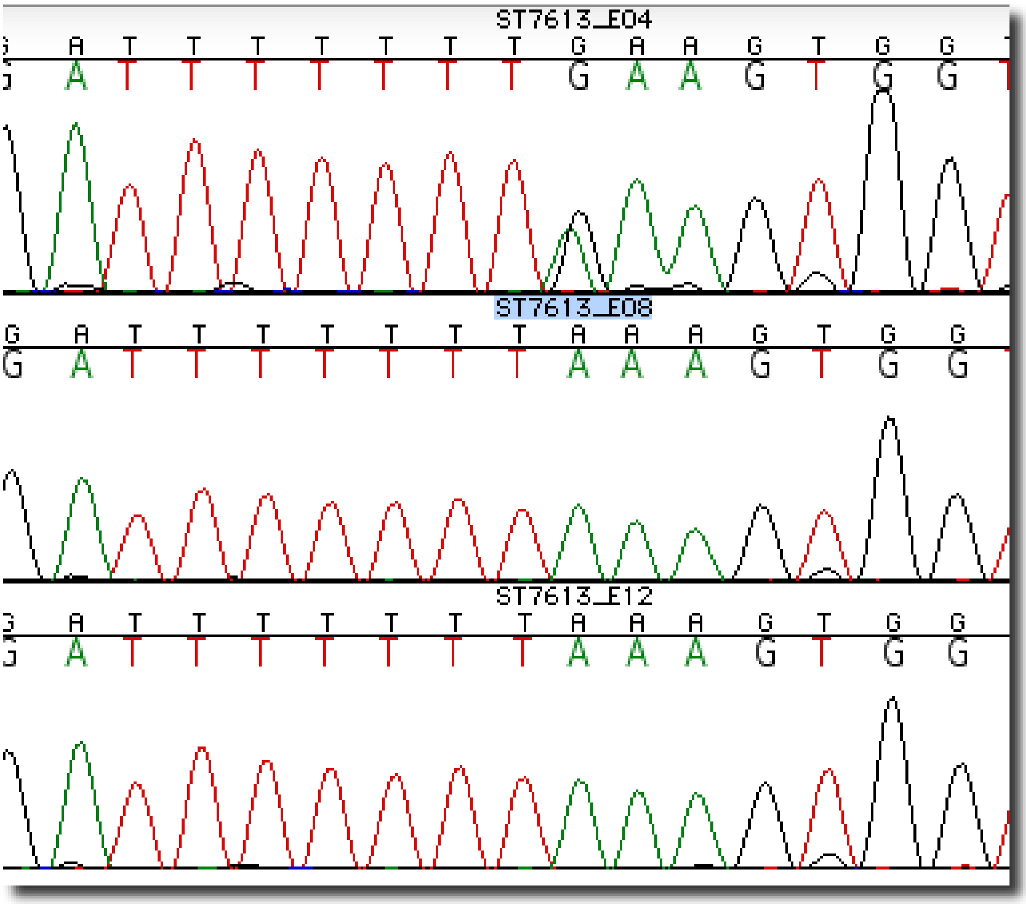
Chr7_90090658_TC_Batch1_E2E6E10.png

Supplementary Figure 5v Validation result for Chr7 position 90090658 (TC)



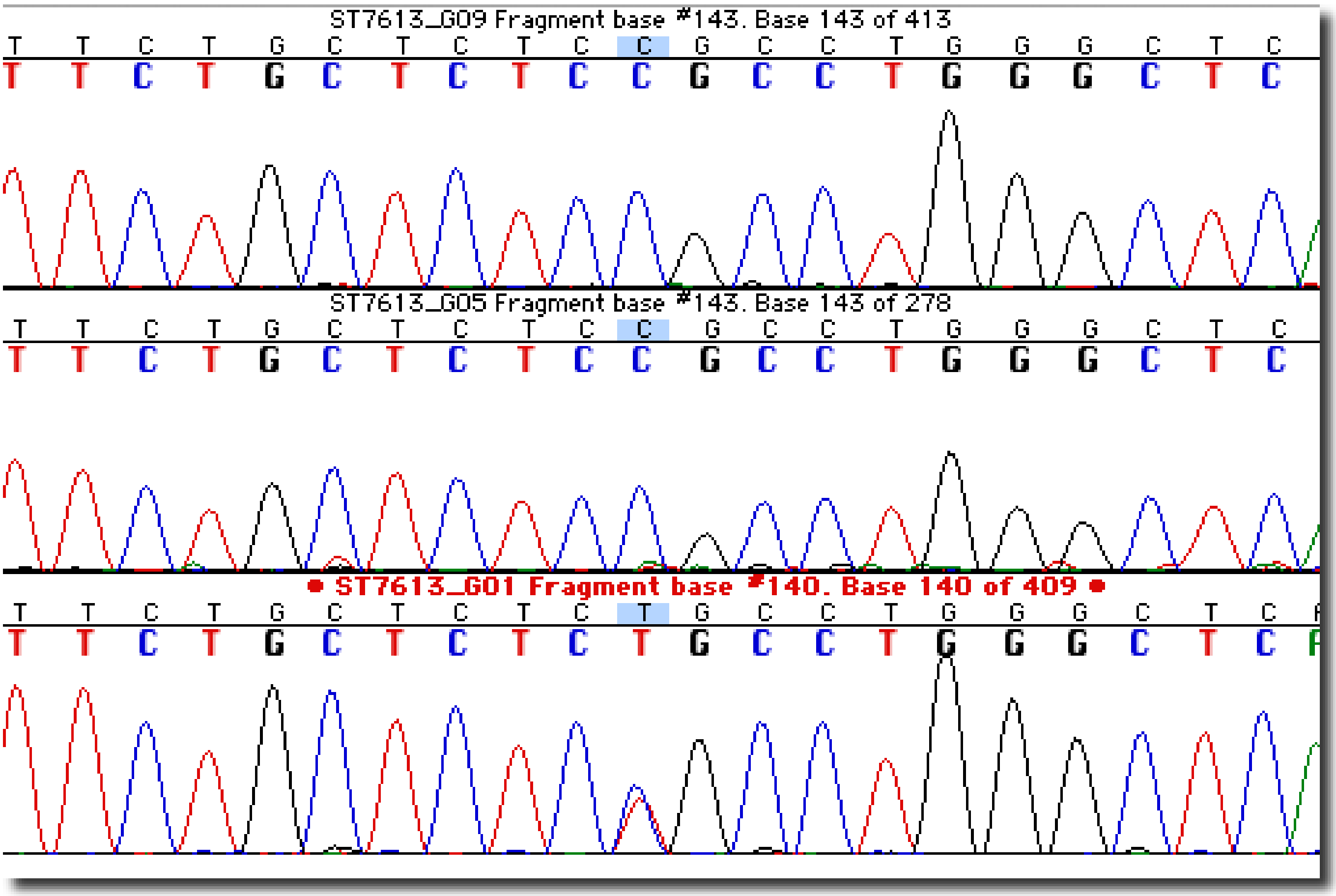
Chr6_104740316_TC_Batch1_E1E5E9.png

Supplementary Figure 5w Validation result for Chr6 position 104740316 (TC)



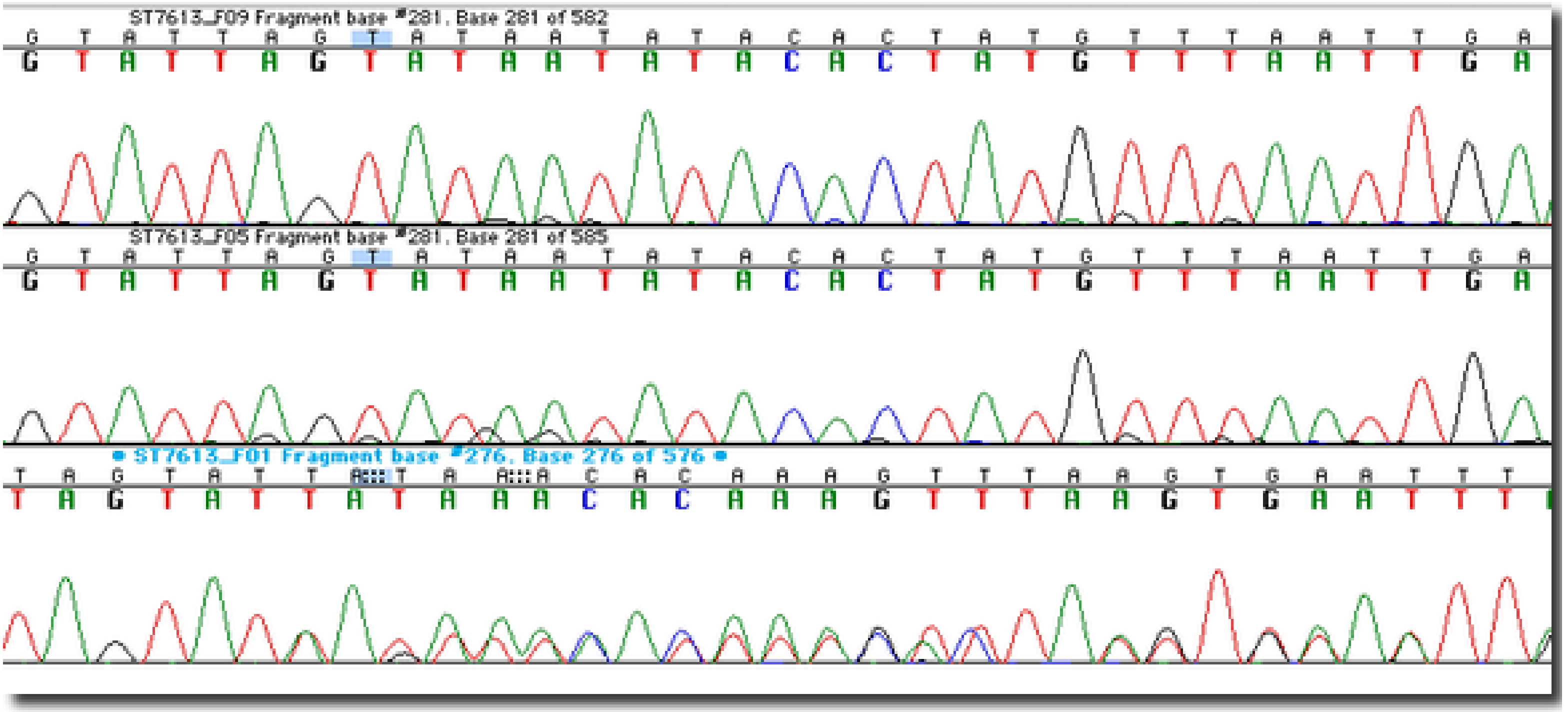
Chr8_123851128_AG_Batch1_E4E8E12.png

Supplementary Figure 5z Validation result for Chr8 position 123851128 (AG)



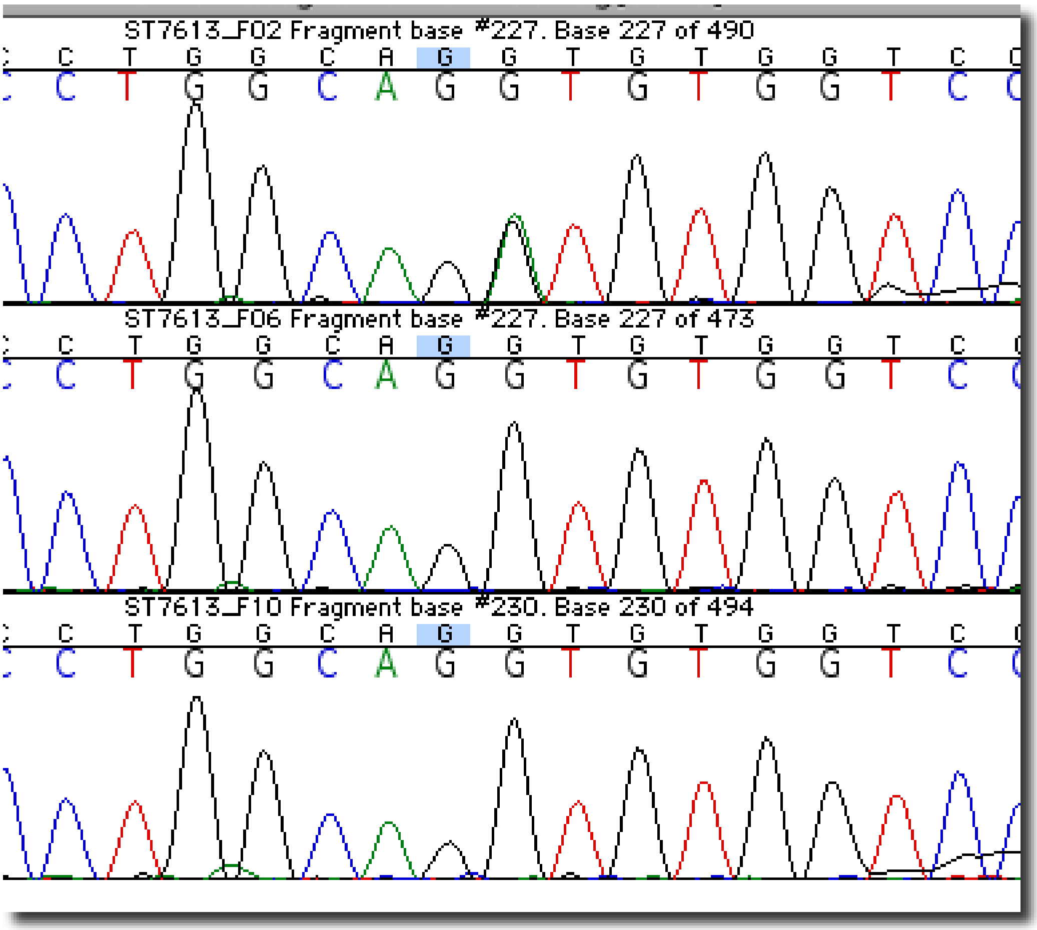
Chr9_12891072_CT_Batch1_G1G5G9.png

Supplementary Figure 5y Validation result for Chr9 position 12891072 (CT)



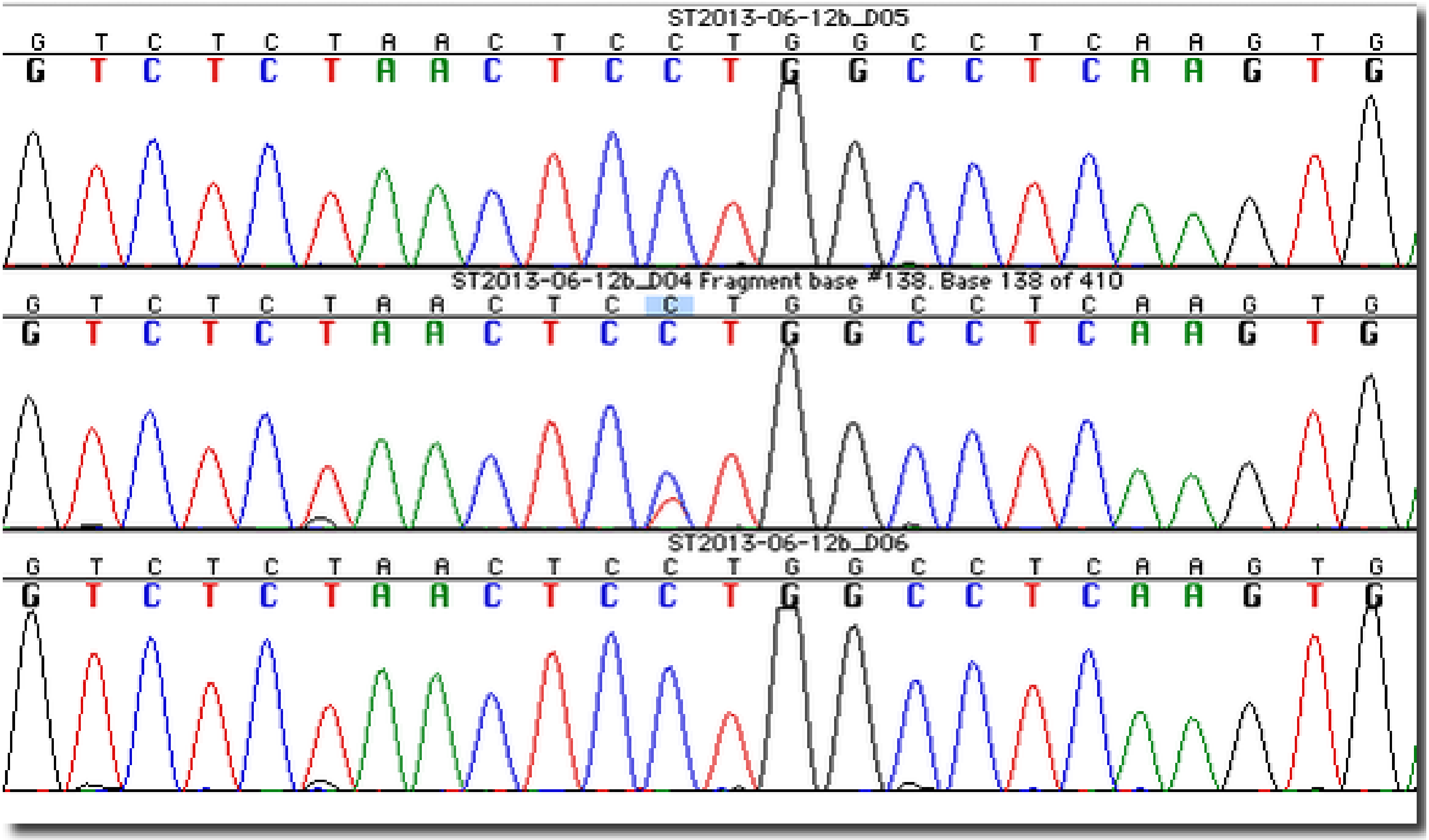
Chr9_29696950_TAGTADel_Batch1_F1F5F9.png

Supplementary Figure 5z Validation result for Chr9 position 29696950 (TAGTA Del)



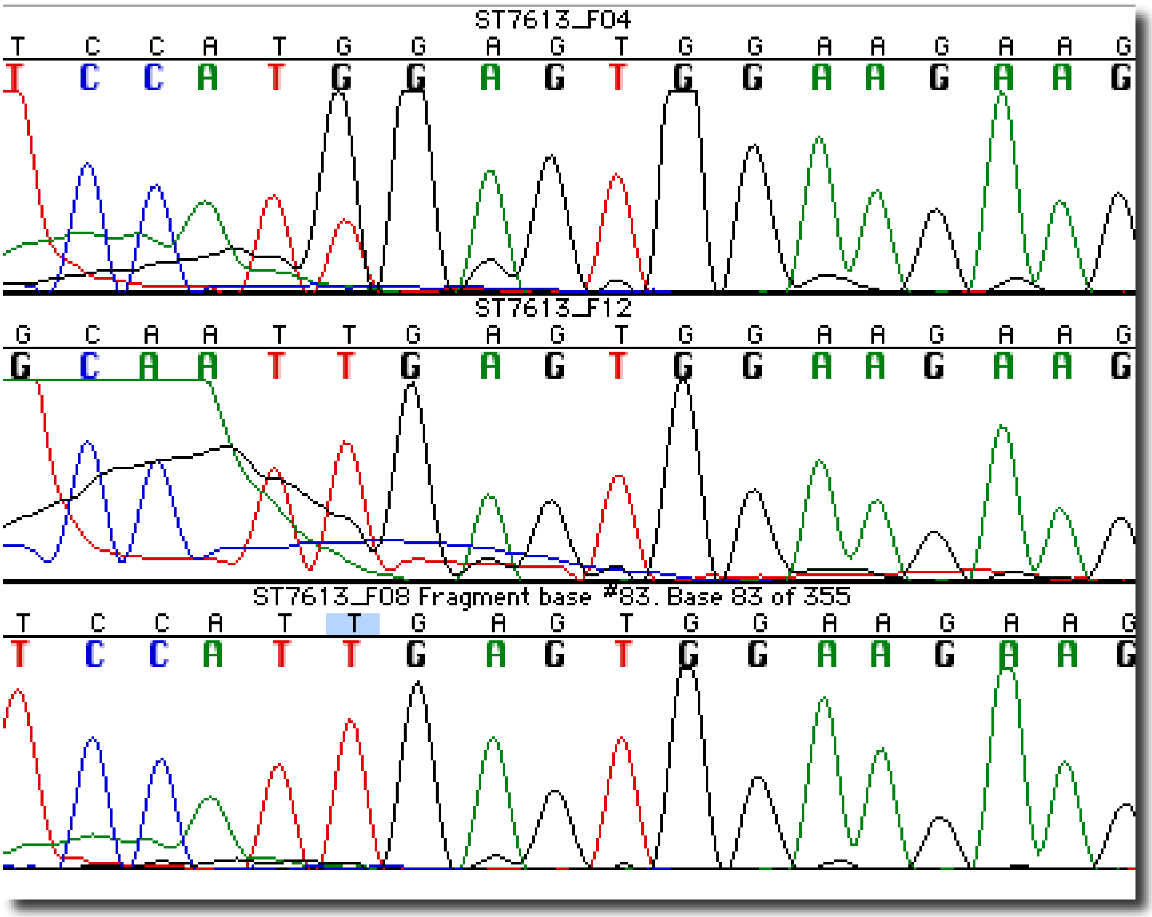
Chr9_35399475_GA_Batch1_F2F6F10.png

Supplementary Figure 5aa Validation result for Chr9 position 35399475 (GA)



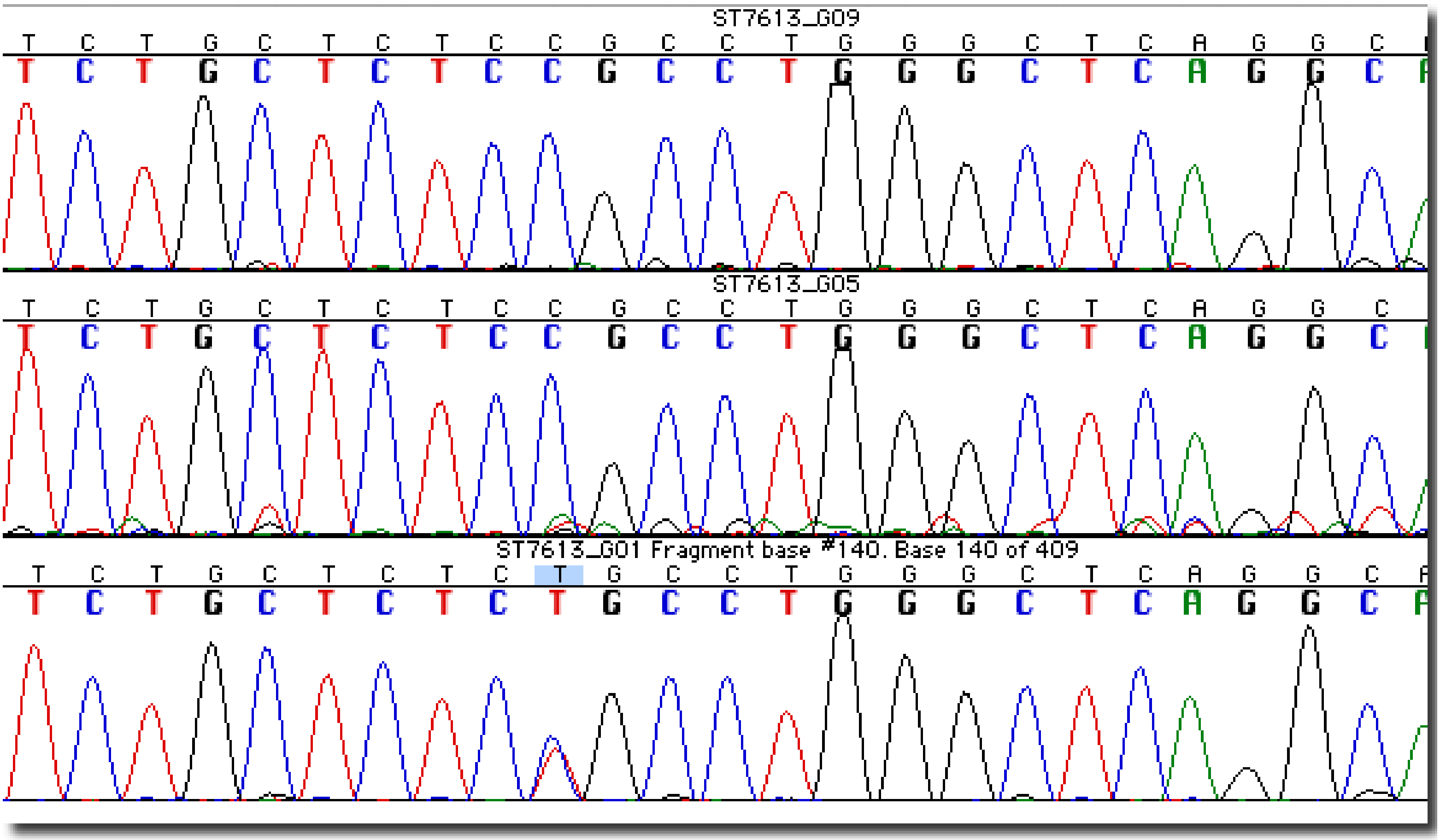
Chr9_99876399_GA_Batch3_D4D5D6_A7A8A9_reverse.png

Supplementary Figure 5ab Validation result for Chr9 position 99876399 (GA)



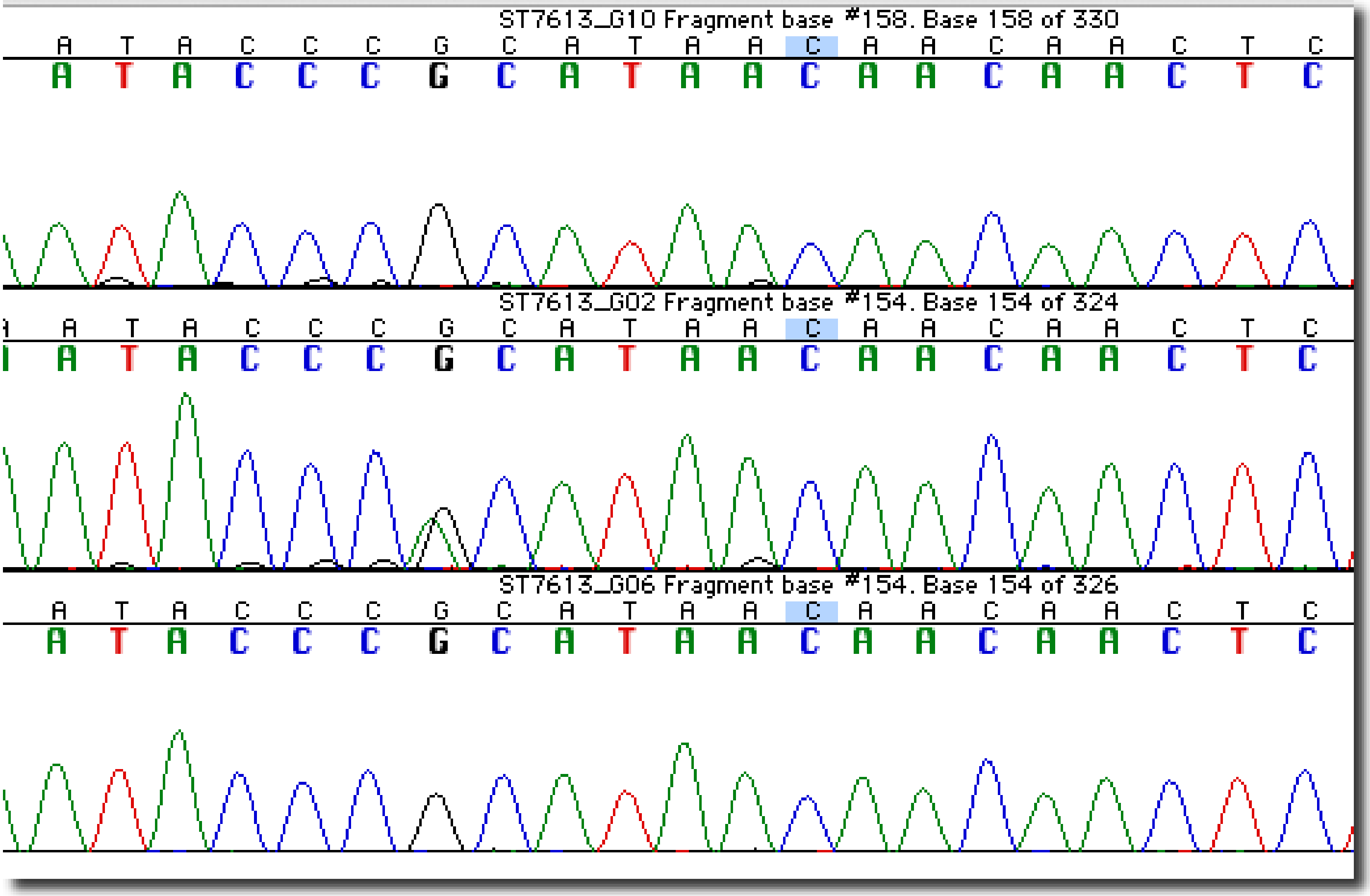
Chr9_110594946_TG_Batch1_F4F8F12.png

Supplementary Figure 5ac Validation result for Chr9 position 110594946 (TG)



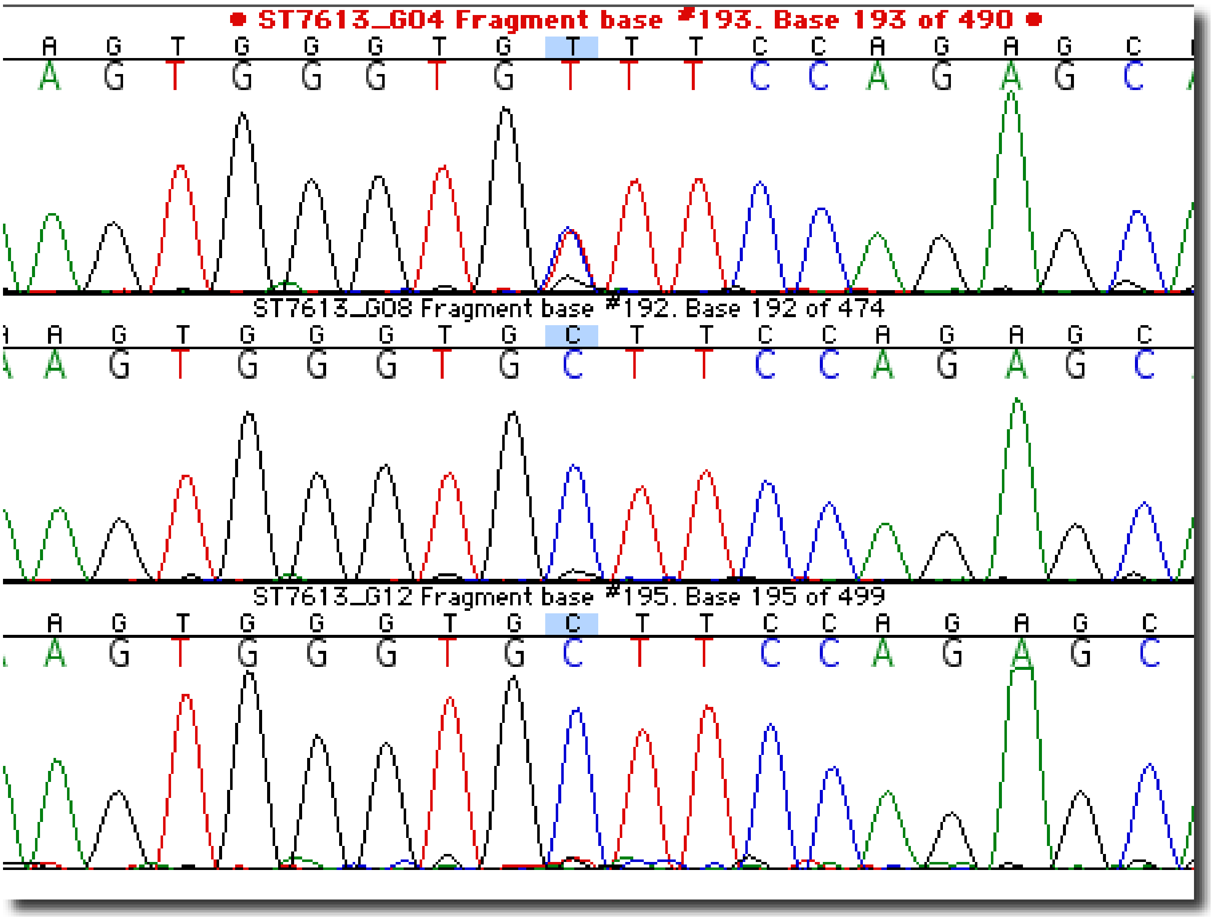
Chr9_112891072_CT_Batch1_G1G5G9.png

Supplementary Figure 5ad Validation result for Chr9 position 112891072 (CT)



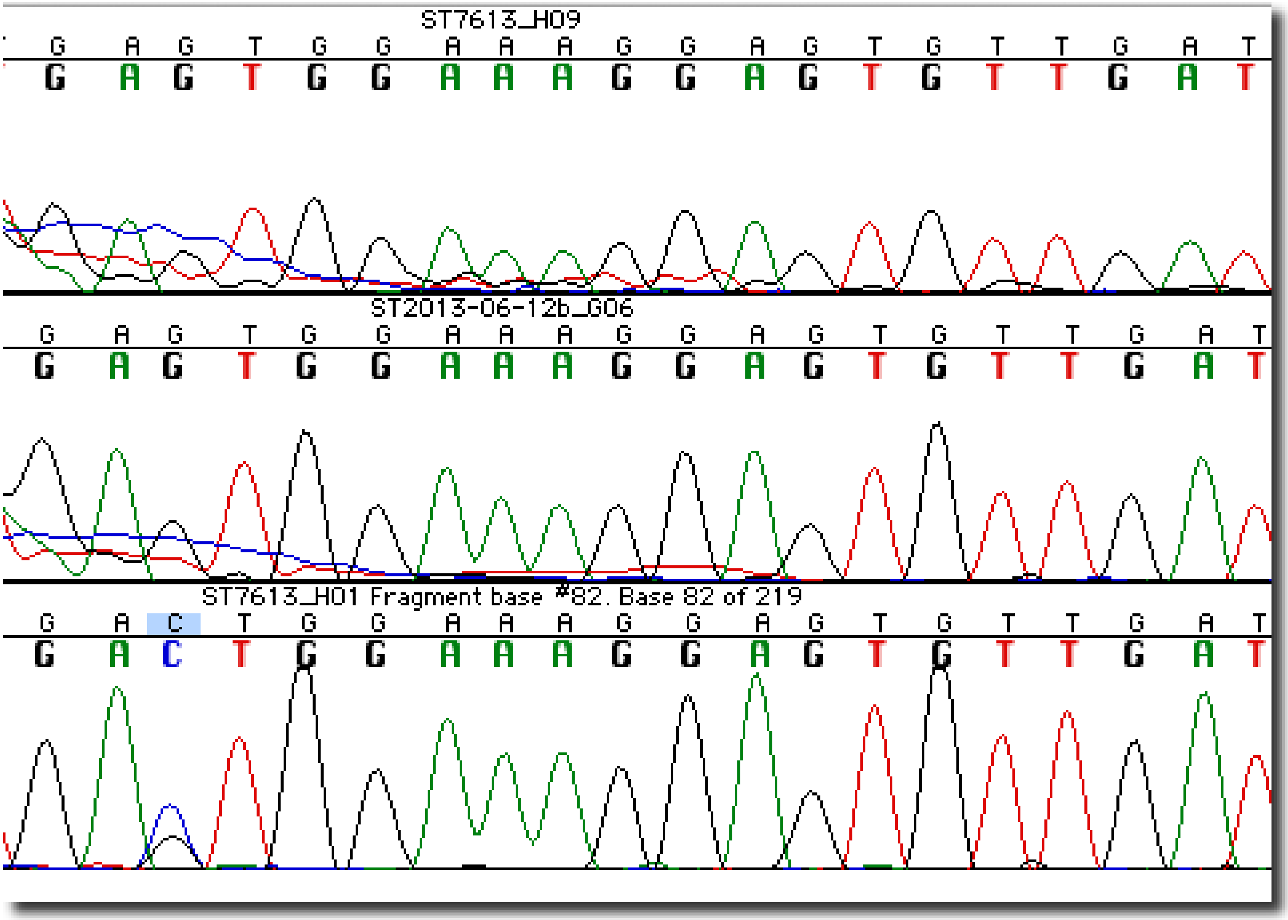
Chr10_2491373_GA_Batch1_G2G6G10.png

Supplementary Figure 5ae Validation result for Chr10 position 2491373 (GA)



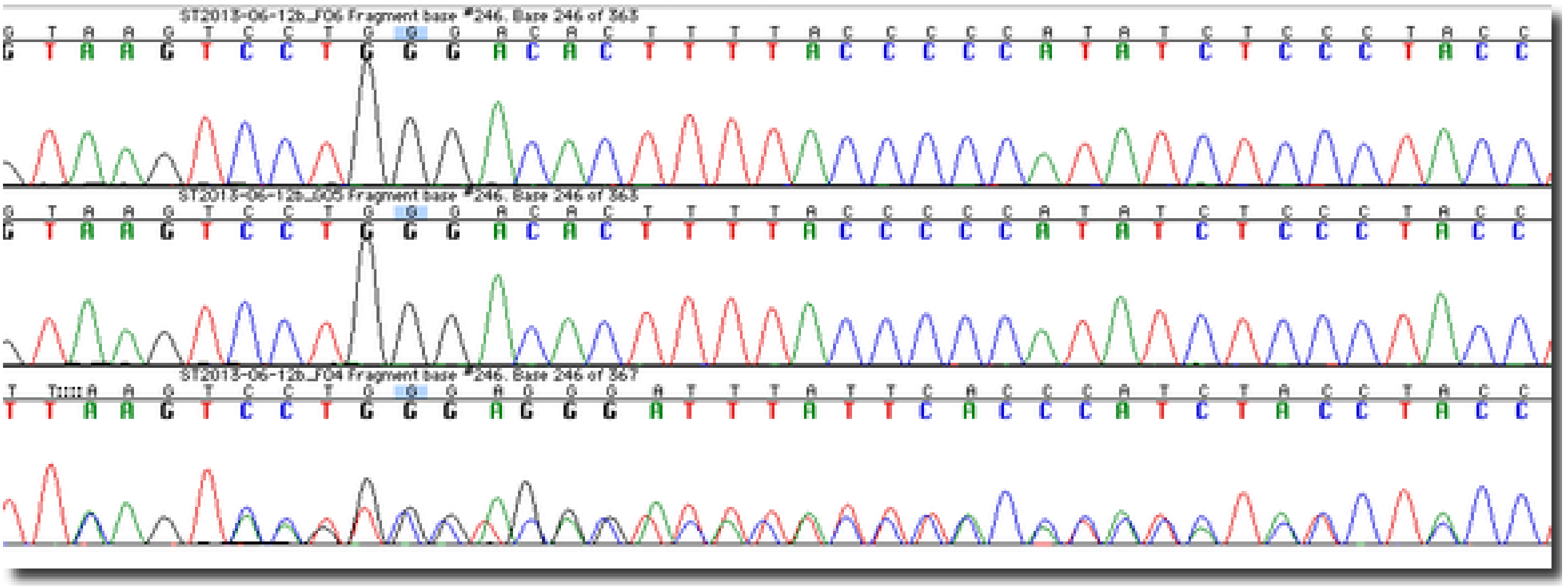
Chr11_16842521_CT_Batch1_G4G8G12.png

Supplementary Figure 5af Validation result for Chr11 position 16842521 (CT)



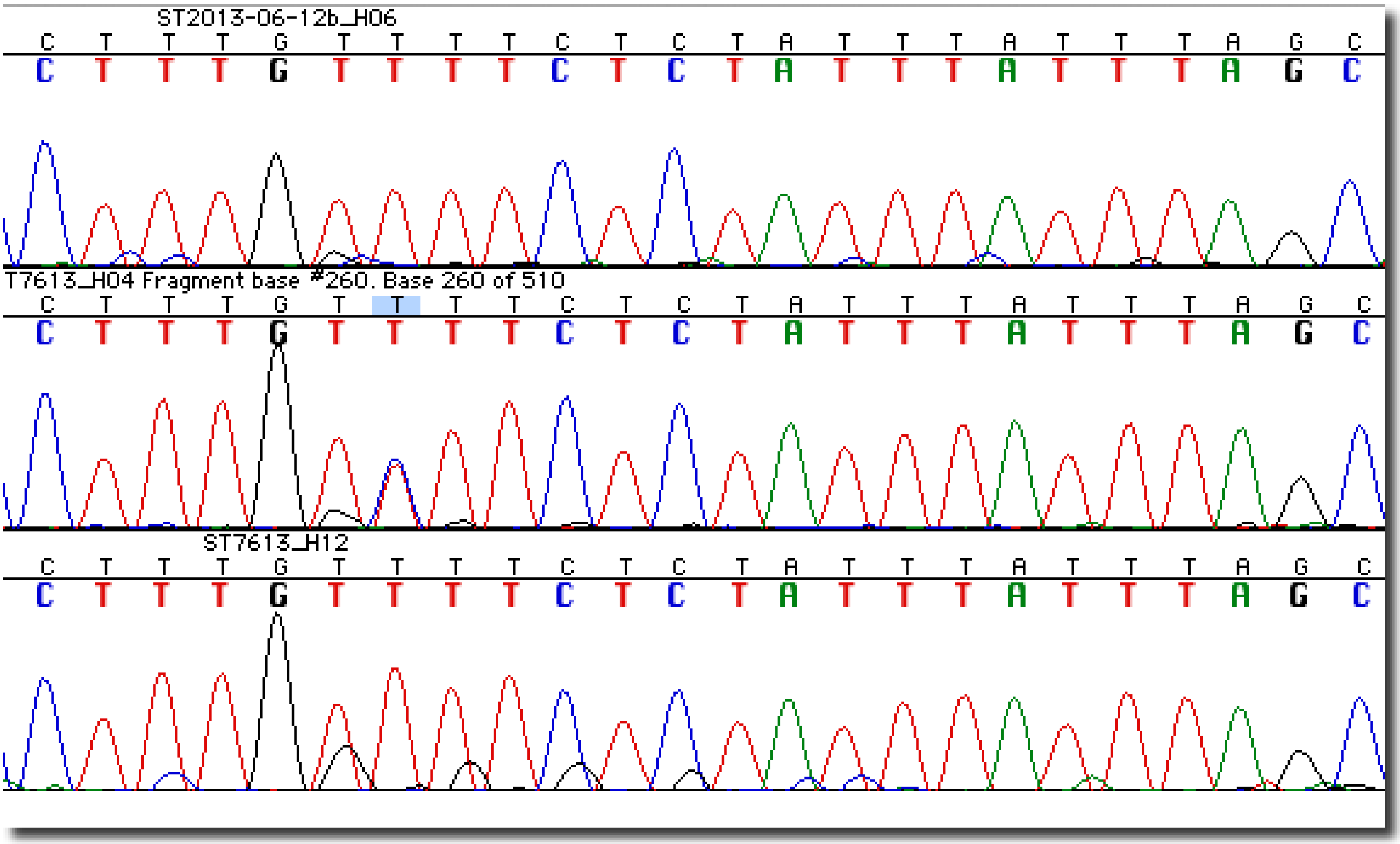
Chr11_57039493_CG_Batch3G6_Batch1H1H9.png

Supplementary Figure 5ag Validation result for Chr11 position 57039493 (CG)



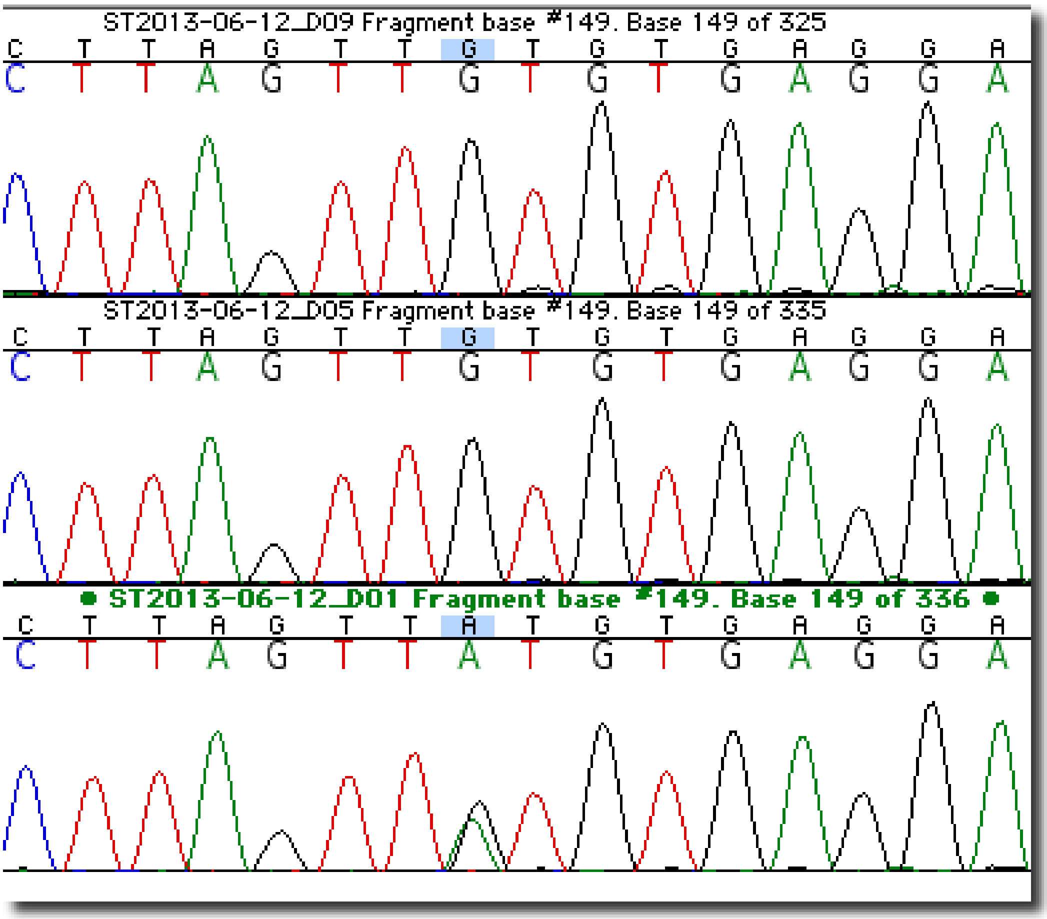
Chr13_31156078_del_Batch3_F4F6G5.png

Supplementary Figure 5ah Validation result for Chr13 position 31156078 (Del)



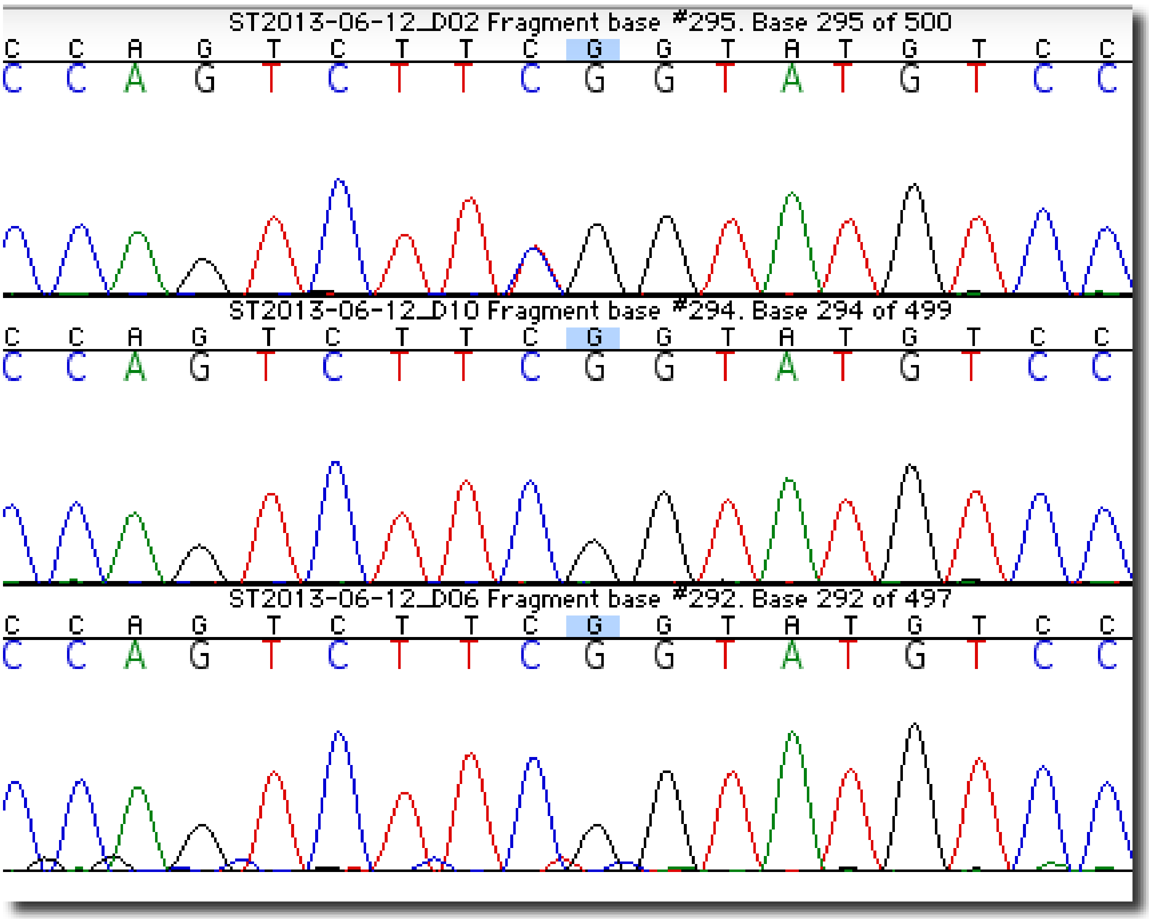
Chr13_47704010_TC_Batch3H6_Batch1H4H12.png

Supplementary Figure 5ai Validation result for Chr13 position 47704010 (TC)



Chr13_65989893_GA_Batch2_D1D5D9.png

Supplementary Figure 5aj Validation result for Chr13 position 65989893 (GA)



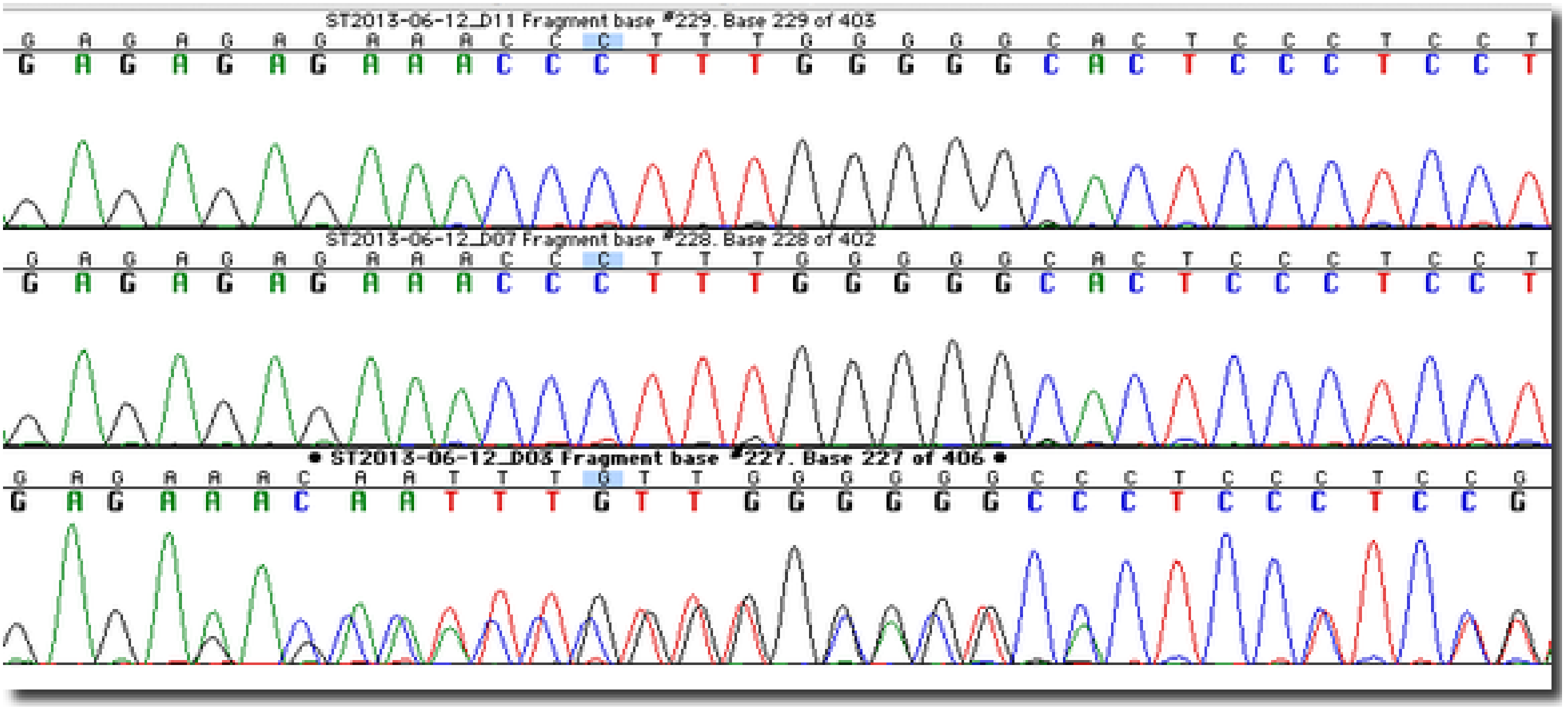
ST2013-06-12_D02 Fragment base #295. Base 295 of 500

ST2013-06-12_D10 Fragment base #294. Base 294 of 499

ST2013-06-12_D06 Fragment base #292. Base 292 of 497

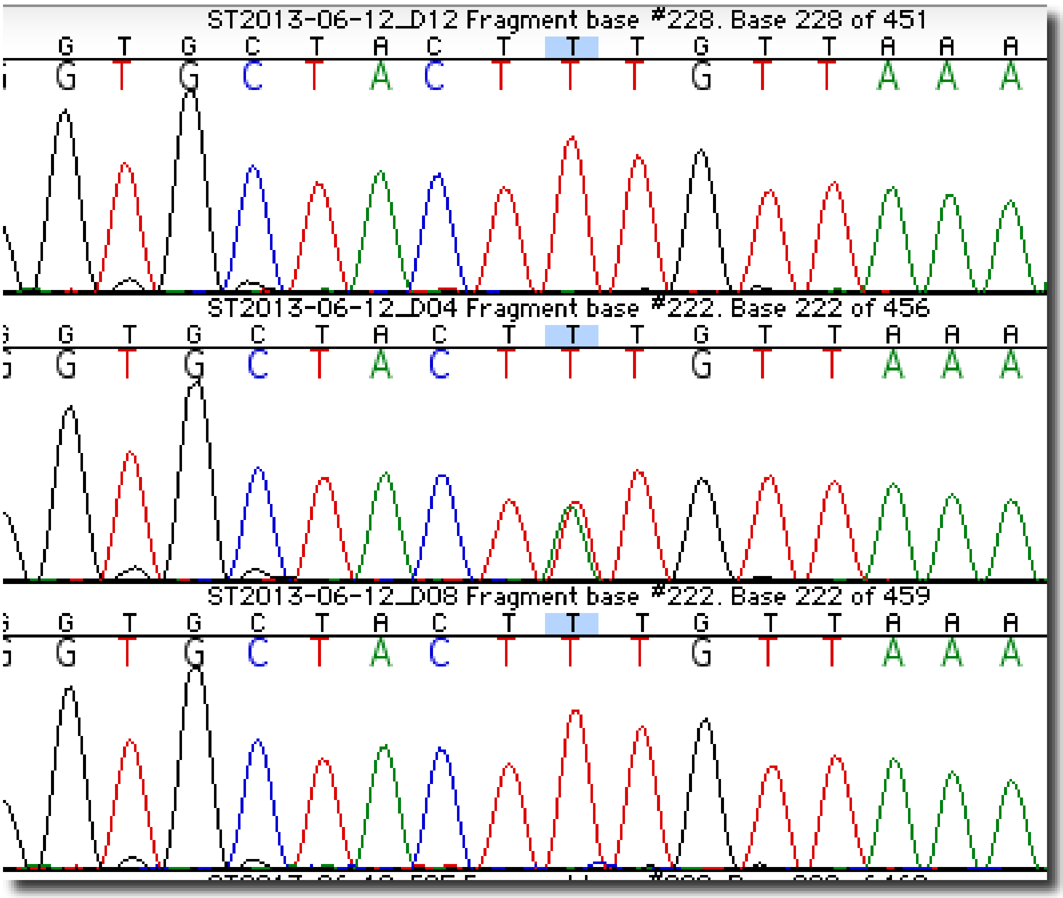
Chr13_82710605_CT_Batch2_D2D6D10.png

Supplementary Figure 5ak Validation result for Chr13 position 82710605 (CT)



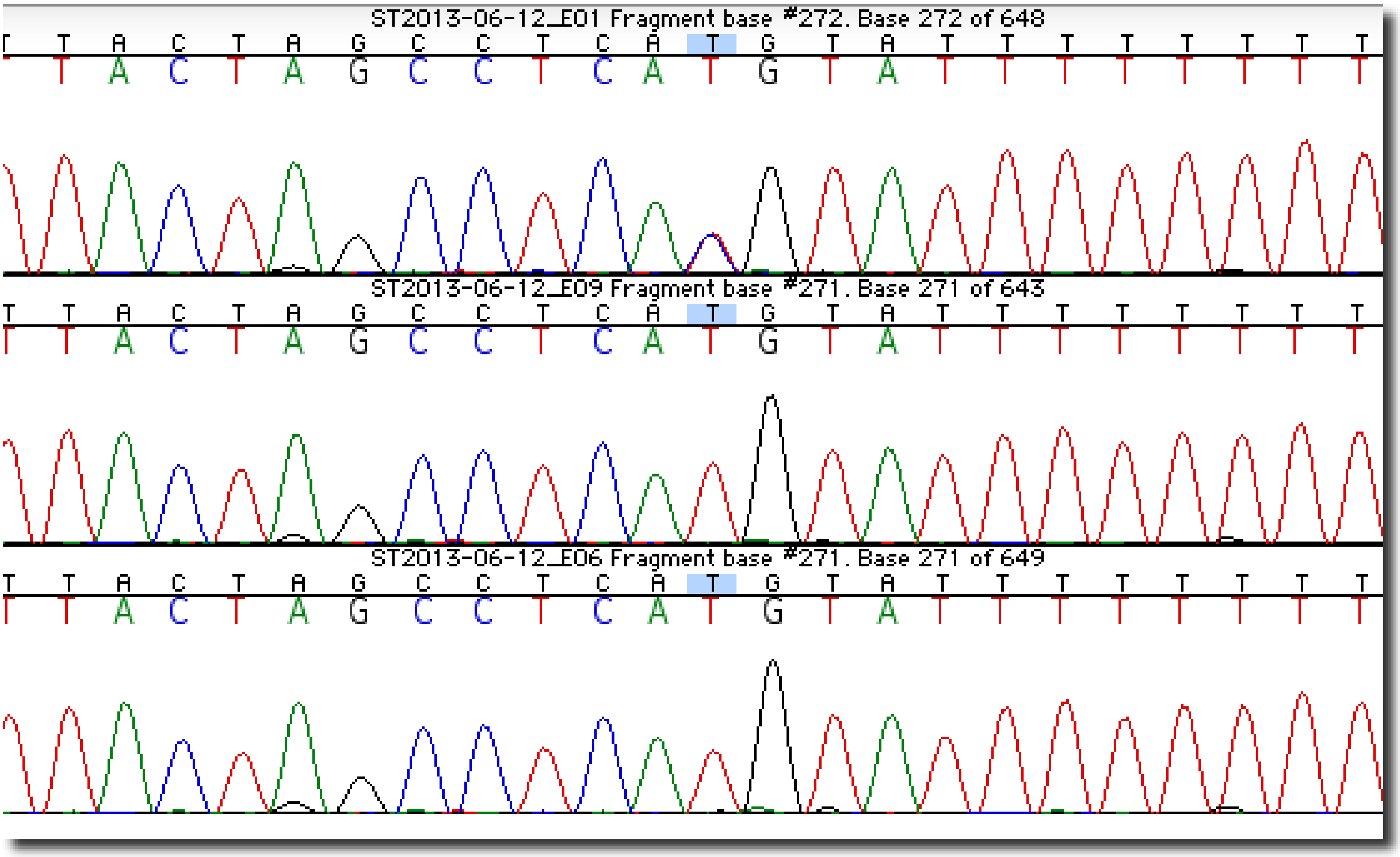
Chr14_31721479_AGAGDel_Batch2_D3D7D11.png

Supplementary Figure 5al Validation result for Chr14 position 31721479 (AGAG Del)



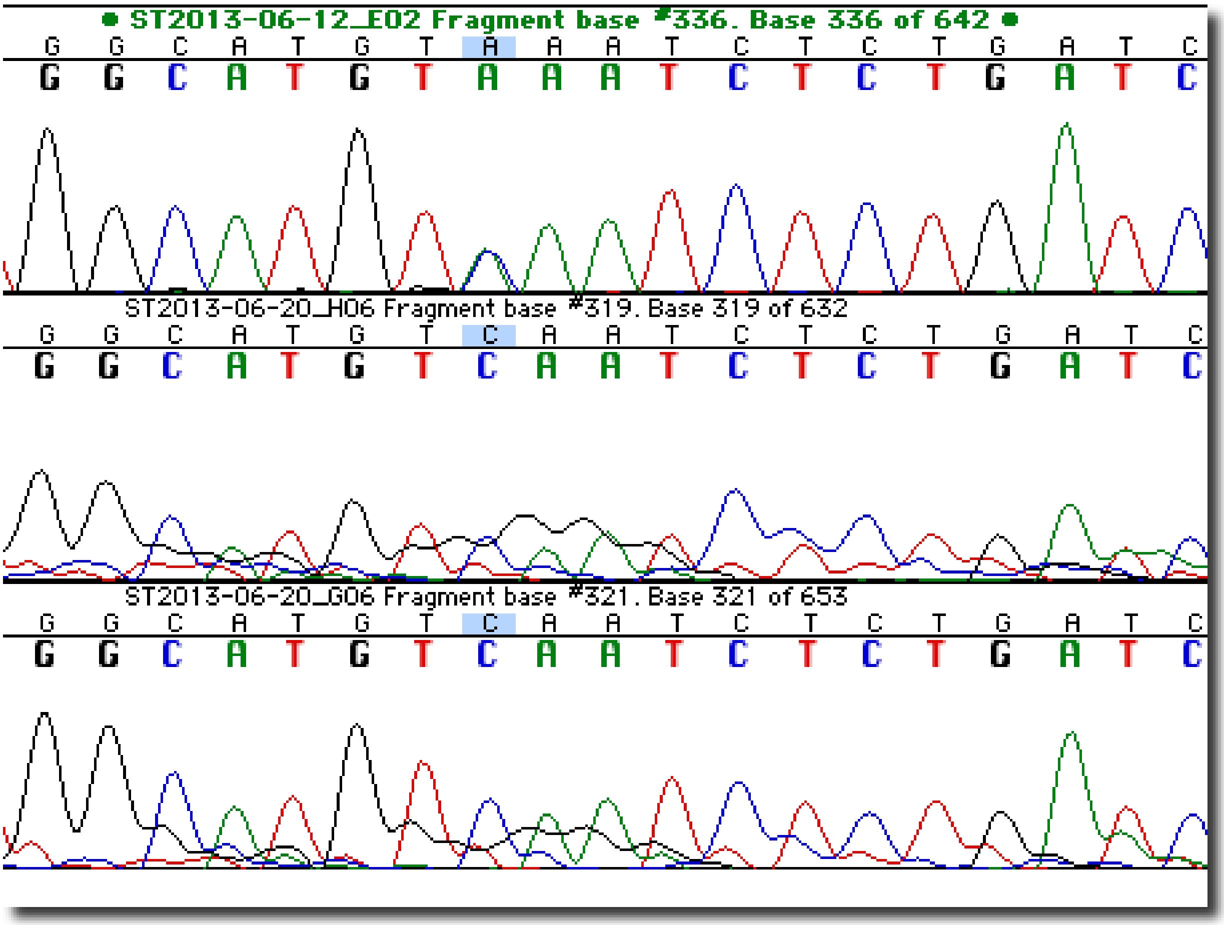
Chr14_45560955_TA_Batch2_D4D8D12.png

Supplementary Figure 5am Validation result for Chr14 position 45560955 (TA)



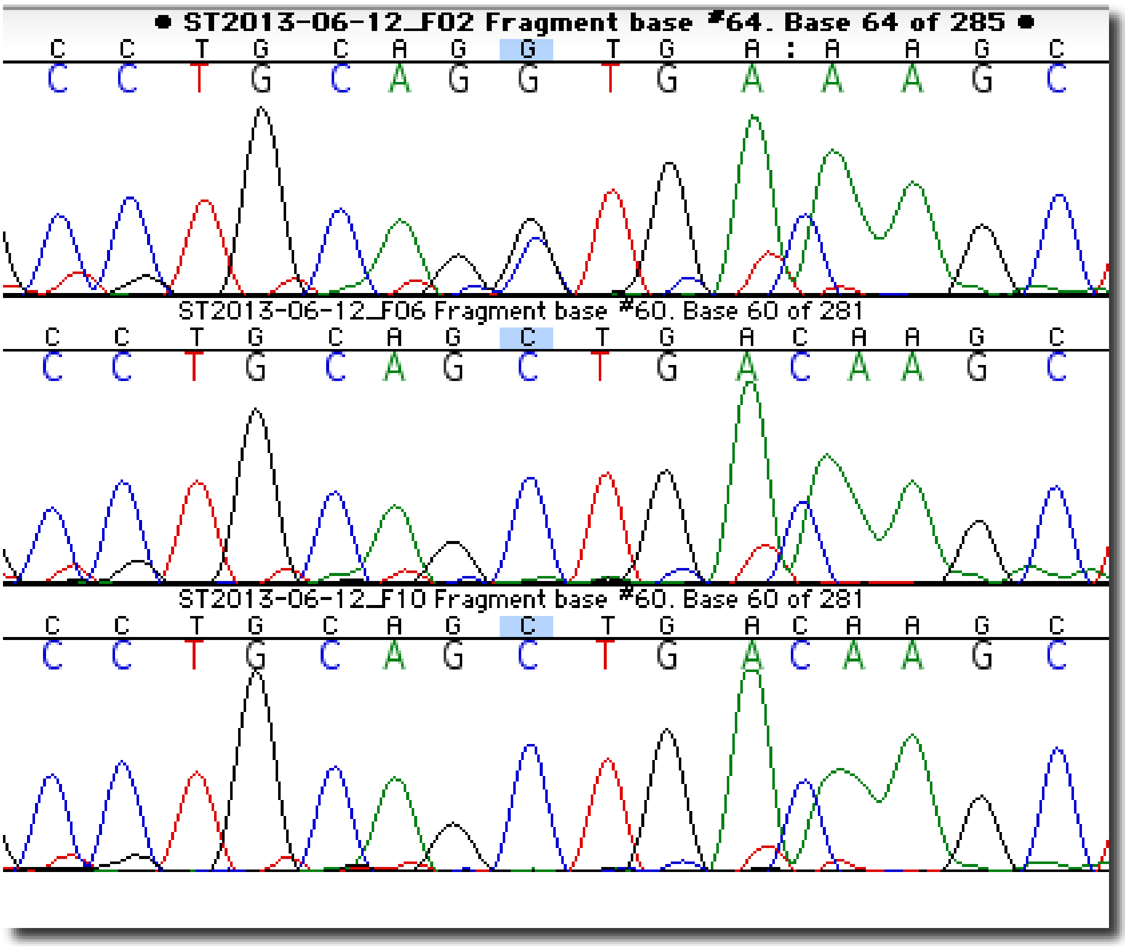
Chr14_106220319_TC_Batch2_E1E6E9.png

Supplementary Figure 5an Validation result for Chr14 position 106220319 (TC)



Chr15_27719721_CA_Batch4_G6H6_Batch2_E2.png

Supplementary Figure 5ao Validation result for Chr15 position 27719721 (CA)



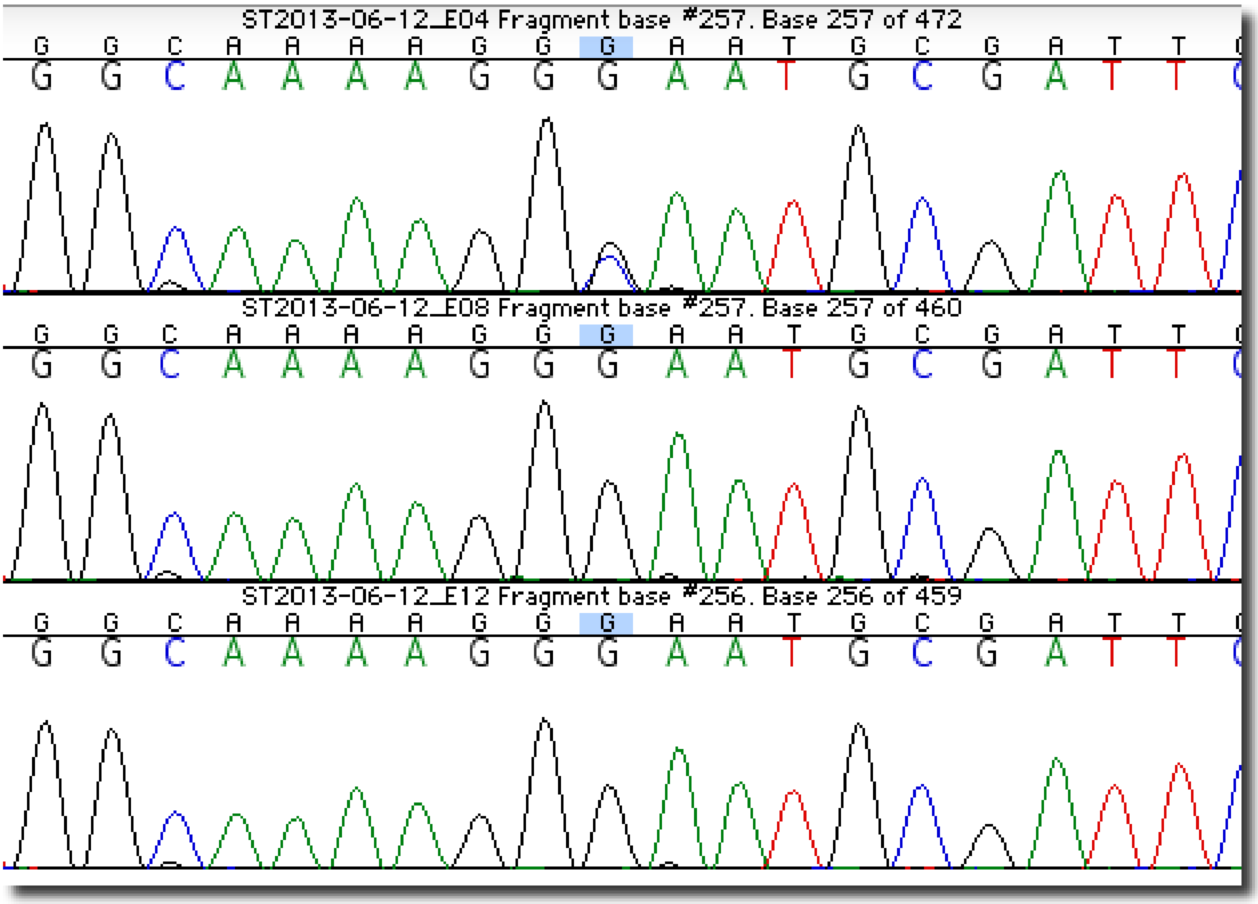
● ST2013-06-12_F02 Fragment base #64. Base 64 of 285 ●

ST2013-06-12_F06 Fragment base #60. Base 60 of 281

ST2013-06-12_F10 Fragment base #60. Base 60 of 281

Chr16_8462342_CG_Batch2_F2F6F10.png

Supplementary Figure 5ap Validation result for Chr16 position 8462342 (CG)



ST2013-06-12_E04 Fragment base #257. Base 257 of 472

G G C A A A A G G G A A T G C G A T T

G G C A A A A G G G A A T G C G A T T

ST2013-06-12_E08 Fragment base #257. Base 257 of 460

G G C A A A A G G G A A T G C G A T T

G G C A A A A G G G A A T G C G A T T

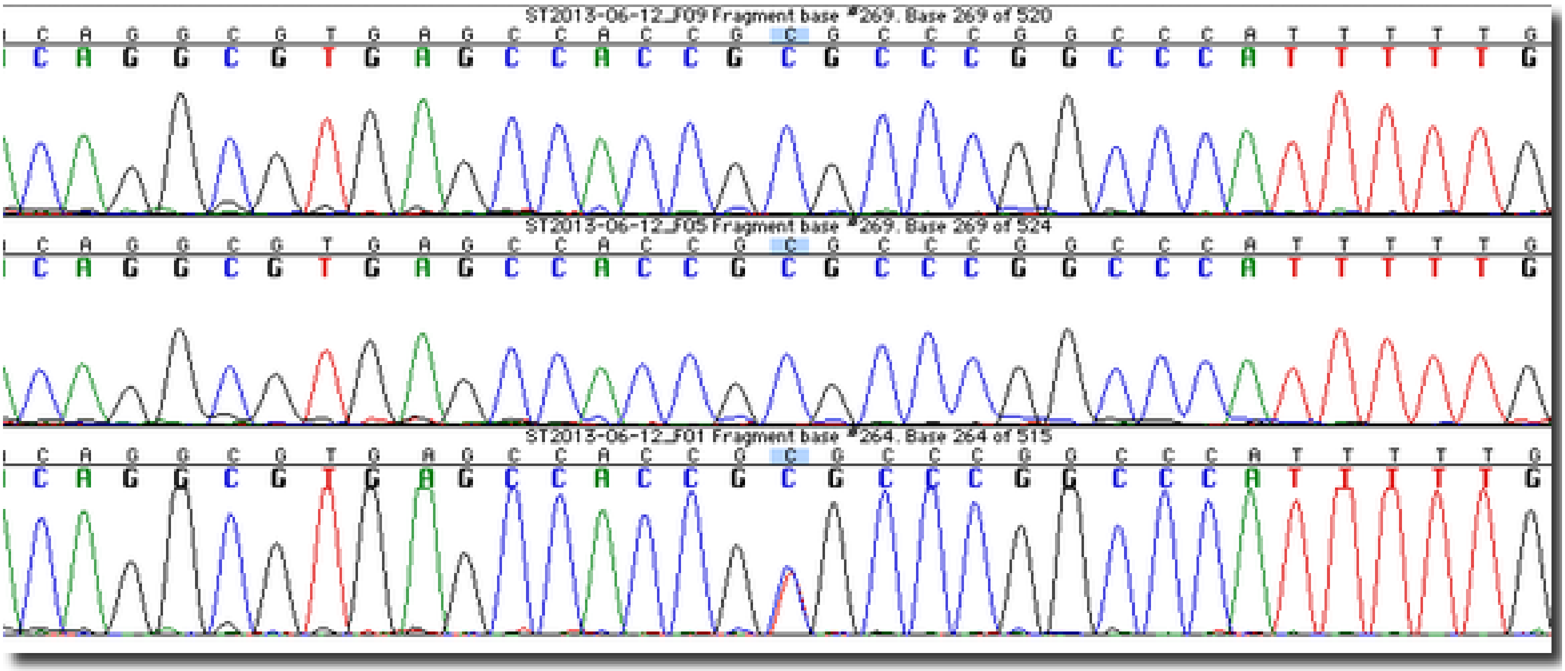
ST2013-06-12_E12 Fragment base #256. Base 256 of 459

G G C A A A A G G G A A T G C G A T T

G G C A A A A G G G A A T G C G A T T

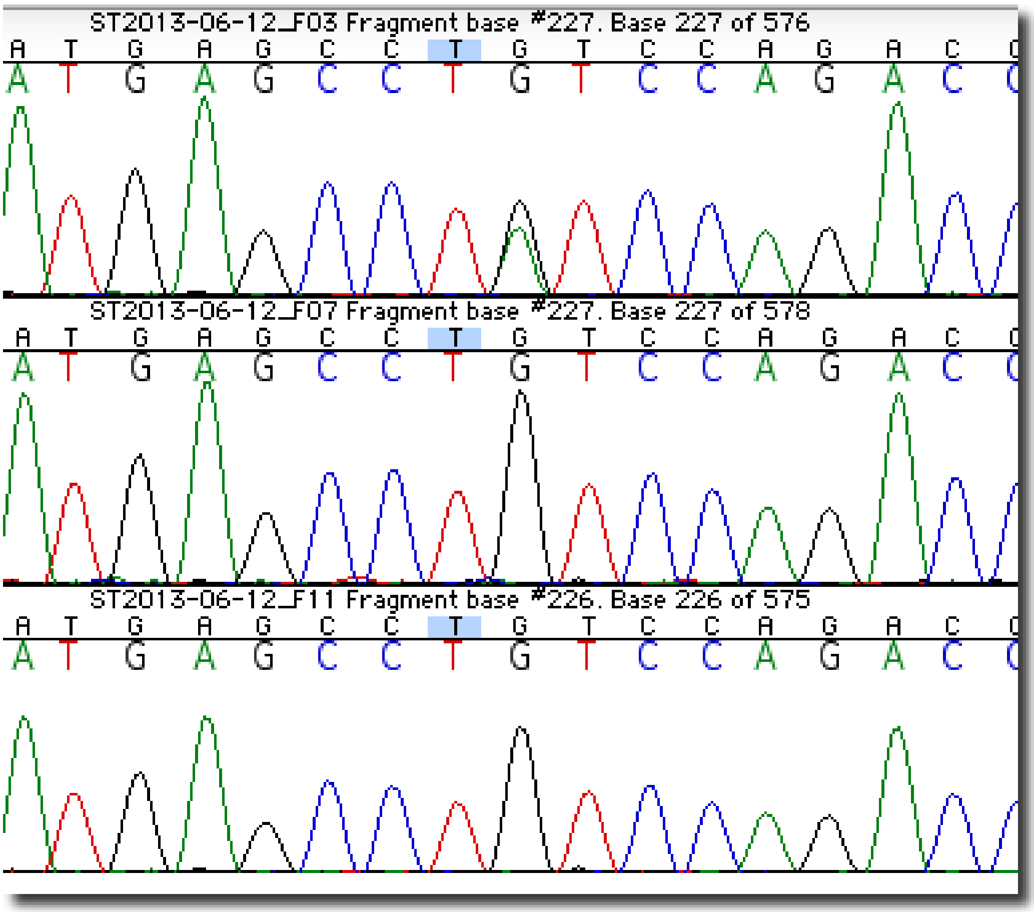
Chr16_53768897_GC_Batch2_E4E8E12.png

Supplementary Figure 5aq Validation result for Chr16 position 53768897 (GC)



Chr16_84096015_CT_Batch2_F1F5F9.png

Supplementary Figure 5ar Validation result for Chr16 position 84096015 (CT)



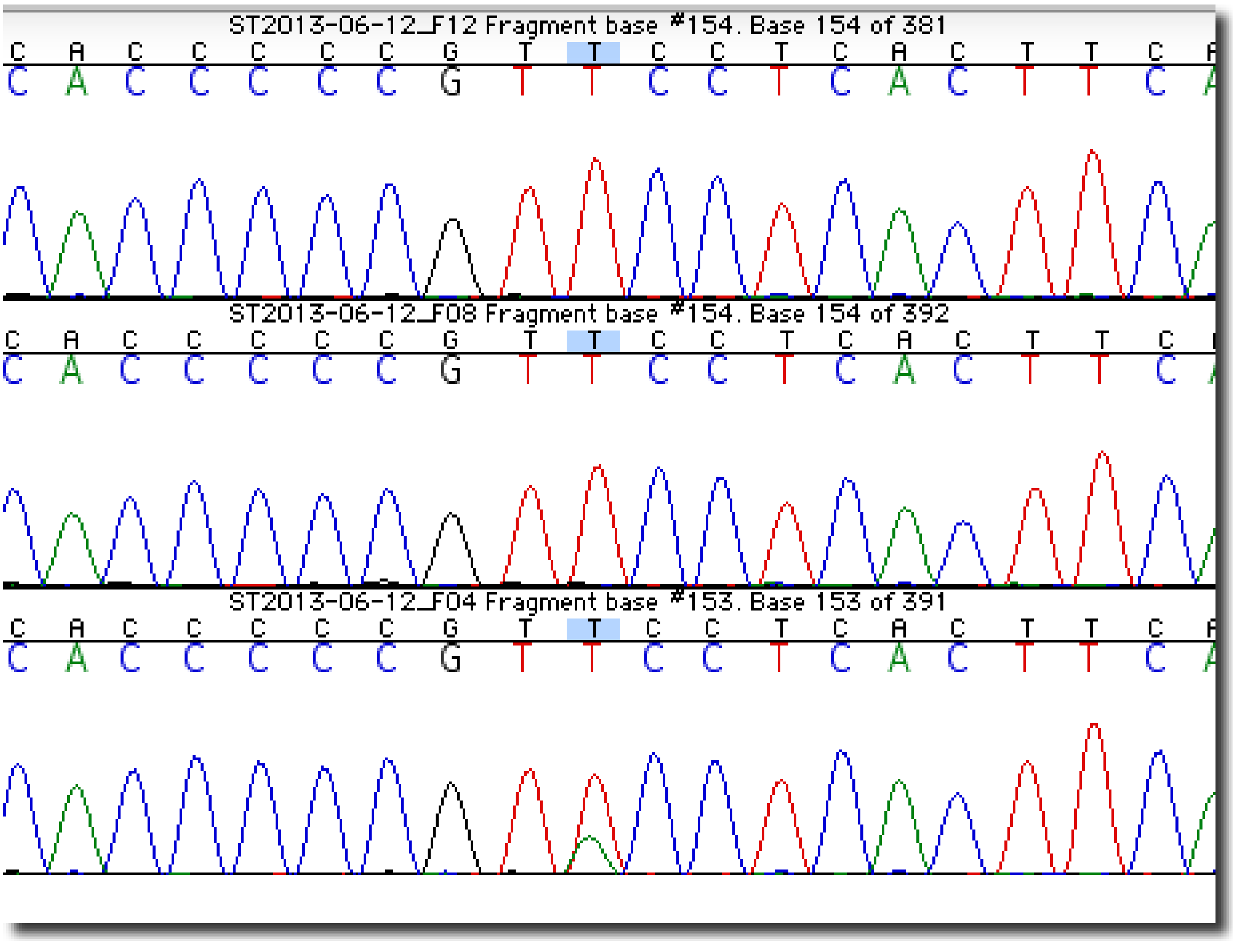
ST2013-06-12_F03 Fragment base #227. Base 227 of 576

ST2013-06-12_F07 Fragment base #227. Base 227 of 578

ST2013-06-12_F11 Fragment base #226. Base 226 of 575

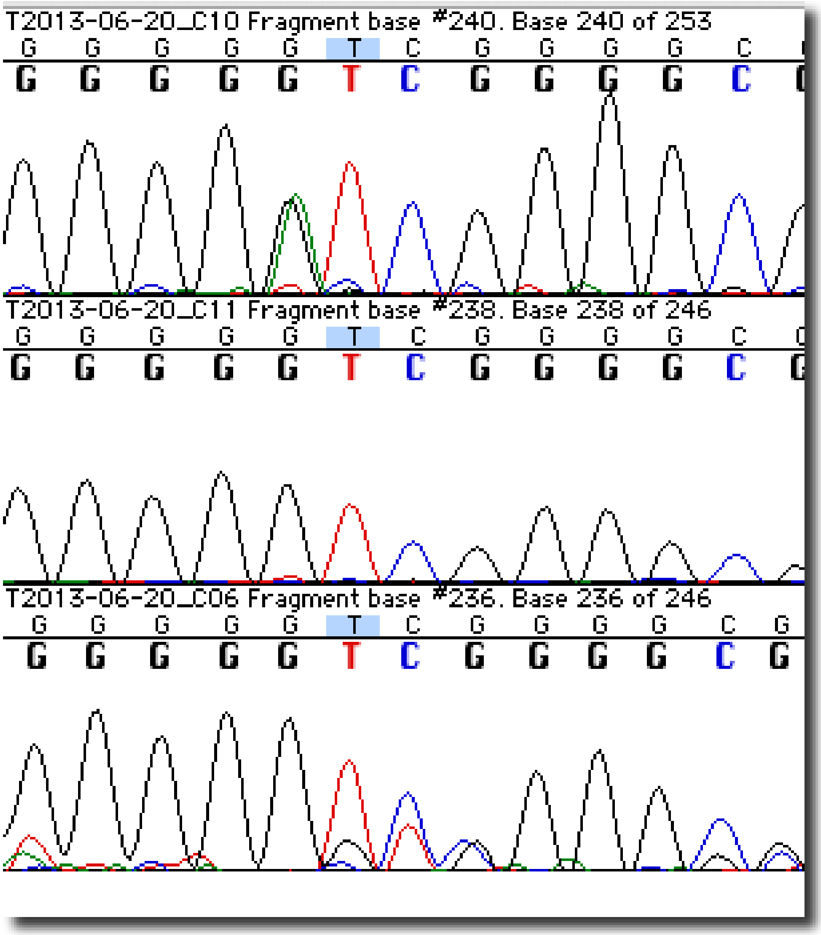
Chr16_85011201_GA_Batch2_F3F7F11.png

Supplementary Figure 5as Validation result for Chr16 position 85011201 (GA)



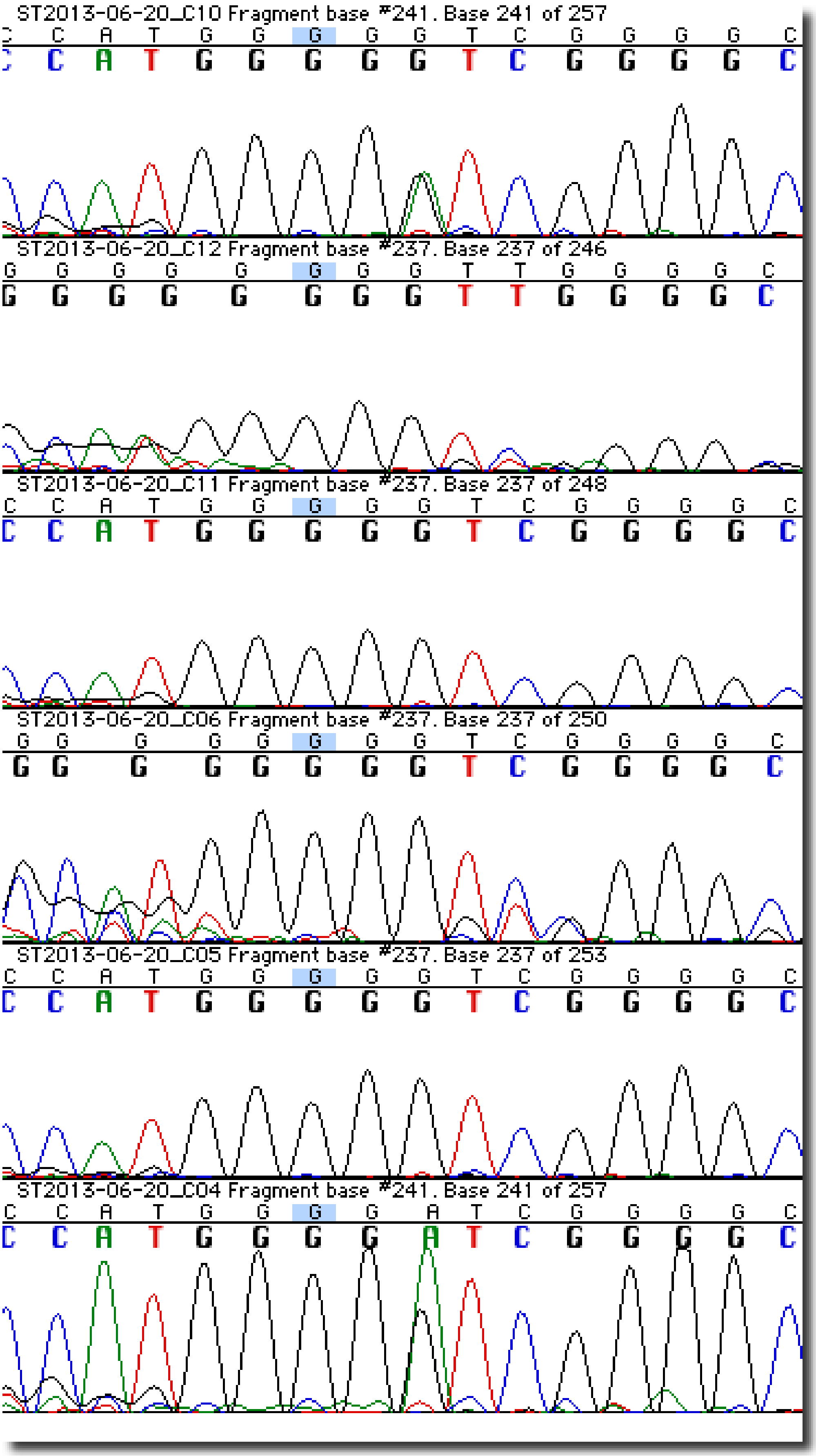
Chr17_15854155_TA_Batch2_F4F8F12.png

Supplementary Figure 5at Validation result for Chr17 position 15854155 (TA)



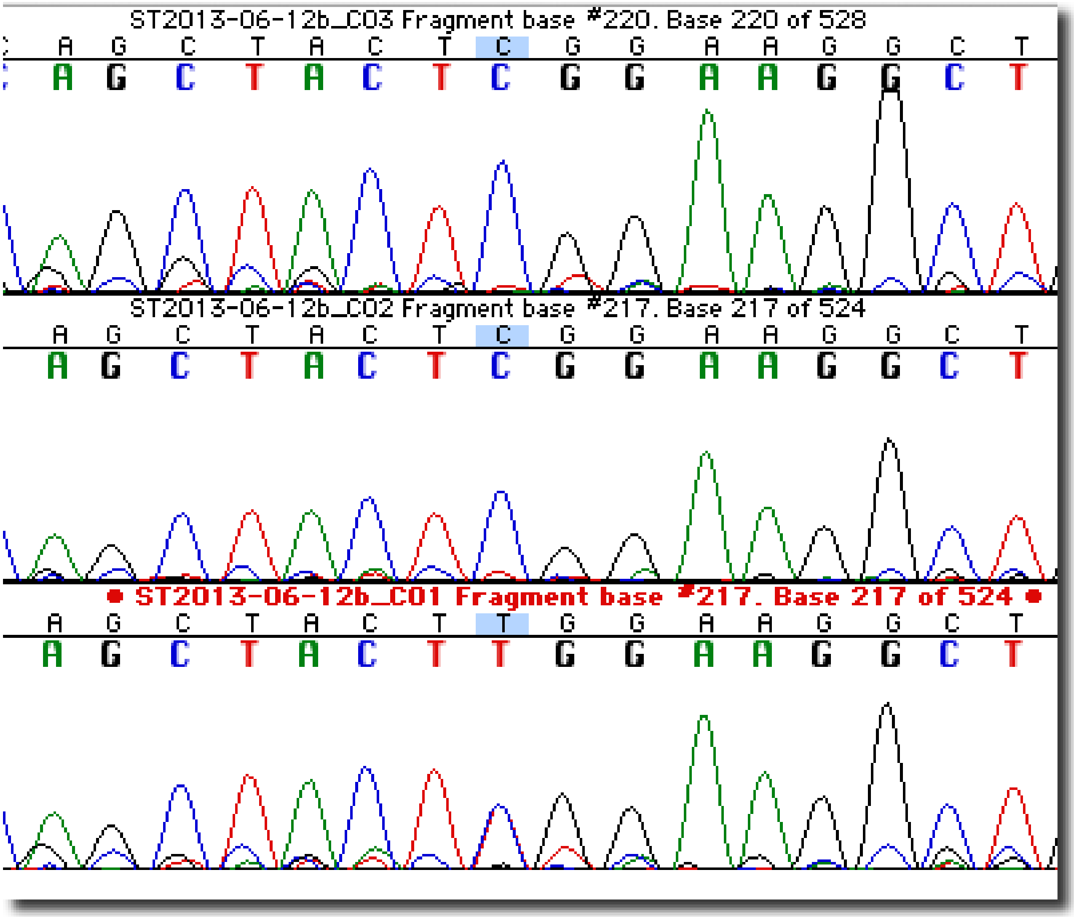
Chr17_17562750_CT_Batch4_C6C10C11_reverse.png

Supplementary Figure 5au Validation result for Chr17 position 17562750 (CT)



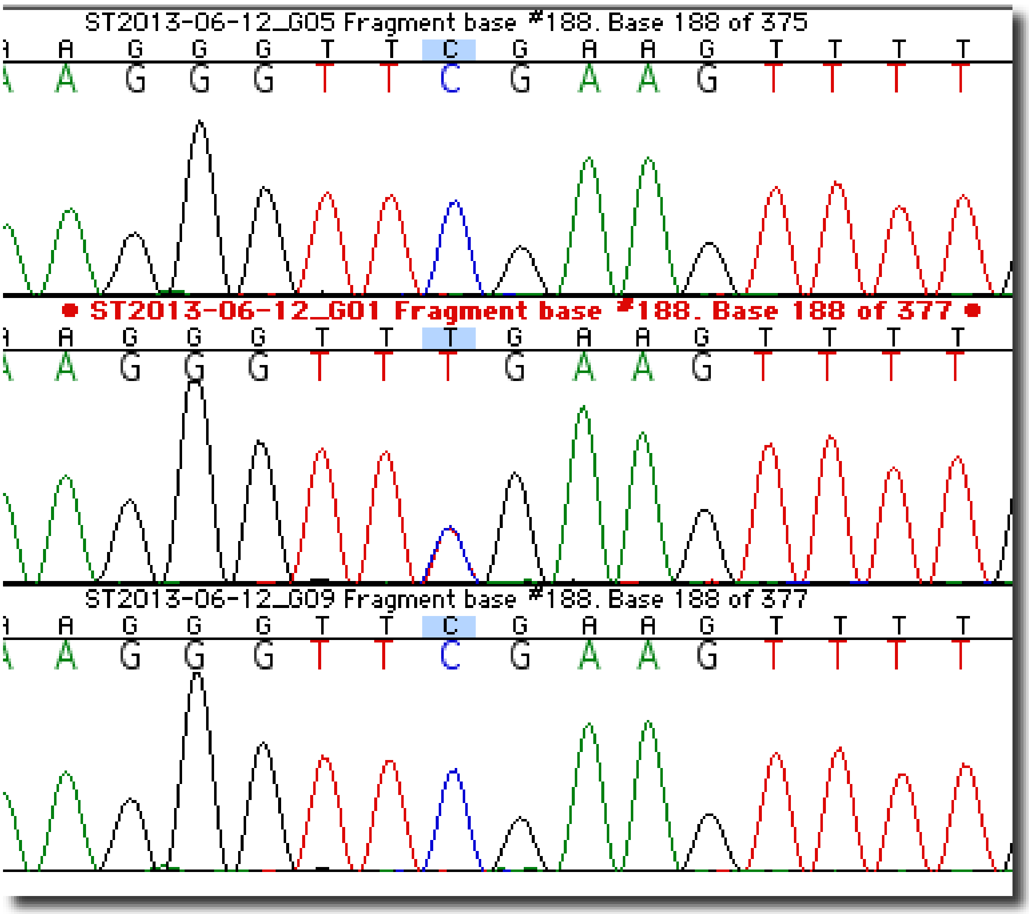
Chr17_17585029_CT_Batch4_C4C5C6C10C11C12_reverse.png

Supplementary Figure 5av Validation result for Chr17 position 17585029 (CT)



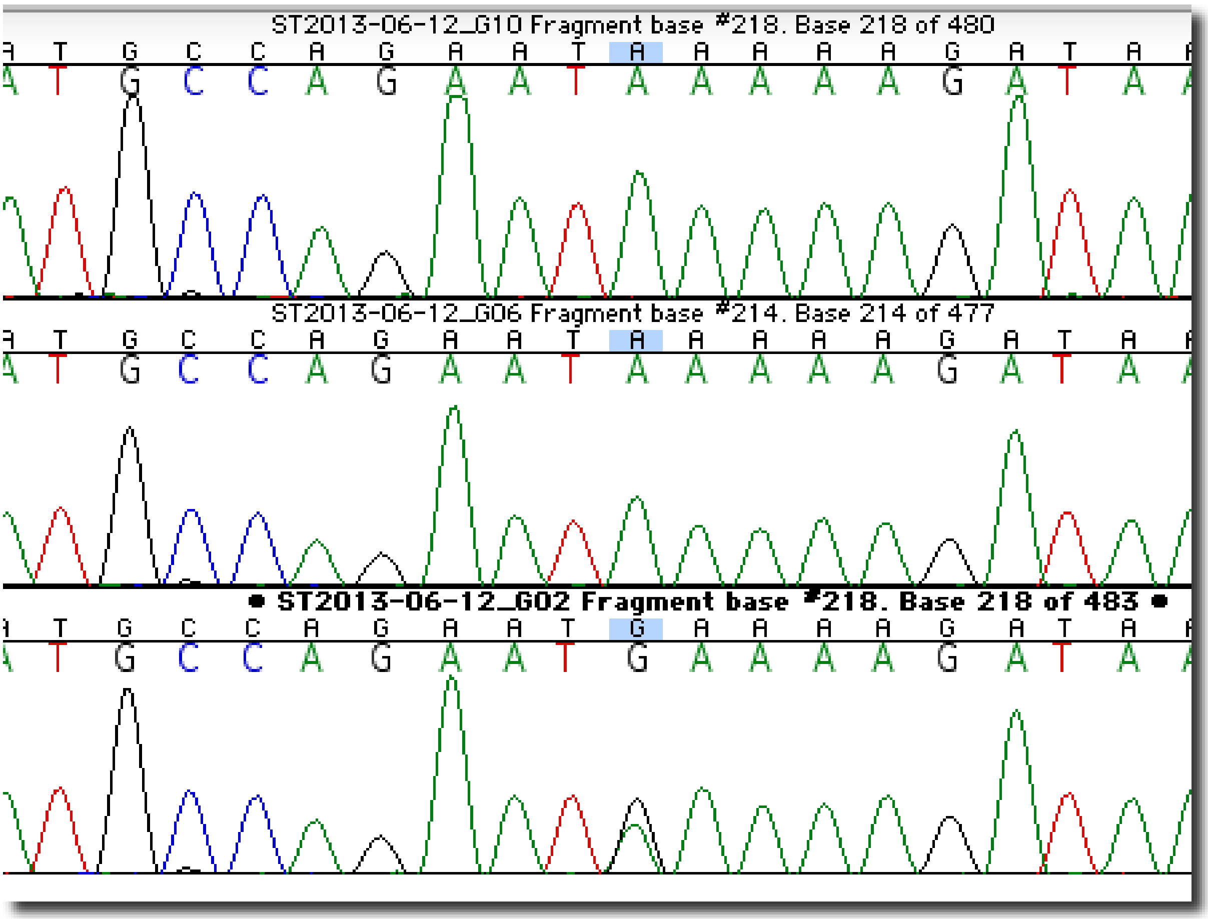
Chr17_49830172_CT_Batch3_C1C2C3.png

Supplementary Figure 5aw Validation result for Chr17 position 49830172 (CT)


ST2013-06-12_G05 Fragment base #188. Base 188 of 375
A A G G G T T C G A A G T T T T
● ST2013-06-12_G01 Fragment base #188. Base 188 of 377 ●
A A G G G T T T G A A G T T T T
ST2013-06-12_G09 Fragment base #188. Base 188 of 377
A A G G G T T C G A A G T T T T

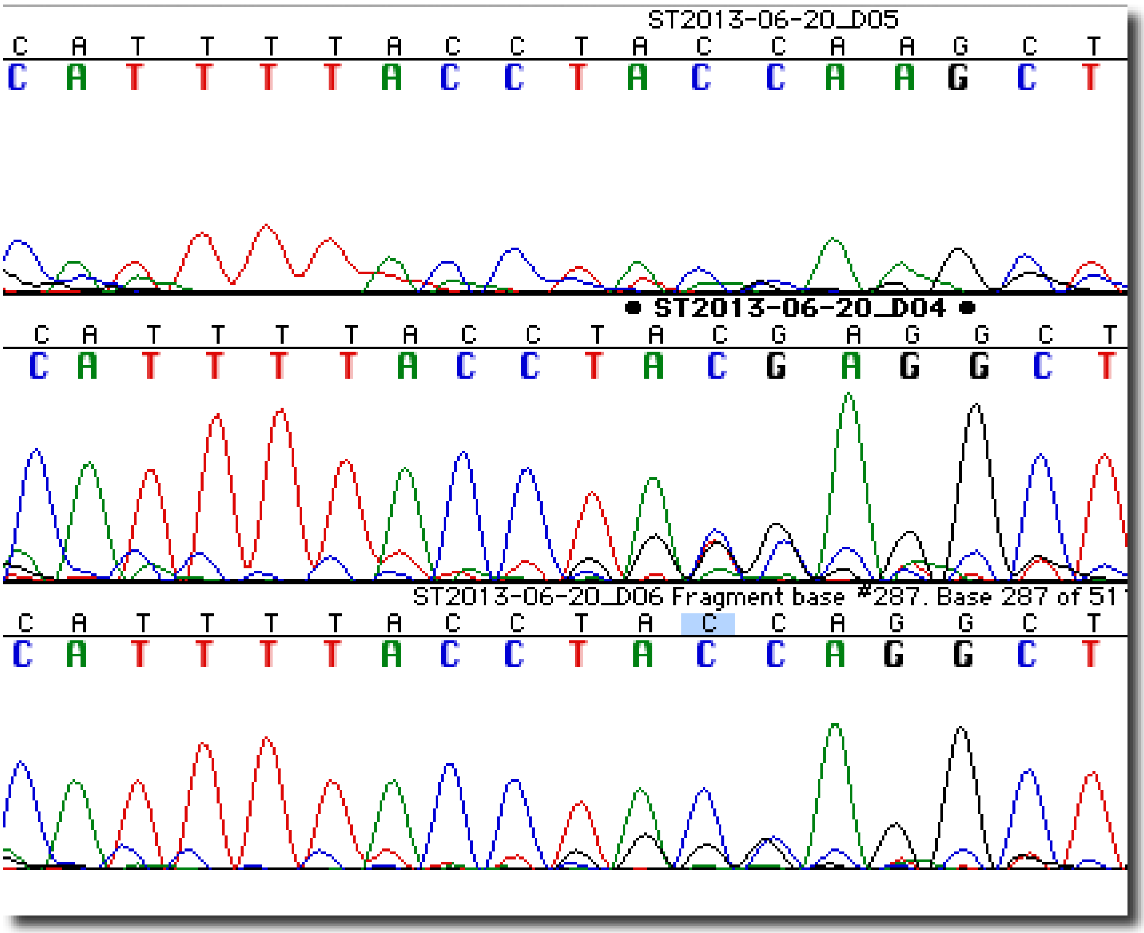
Chr18_3905873_CT_Batch2_G1G5G9.png

Supplementary Figure 5ax Validation result for Chr18 position 3905873 (CT)



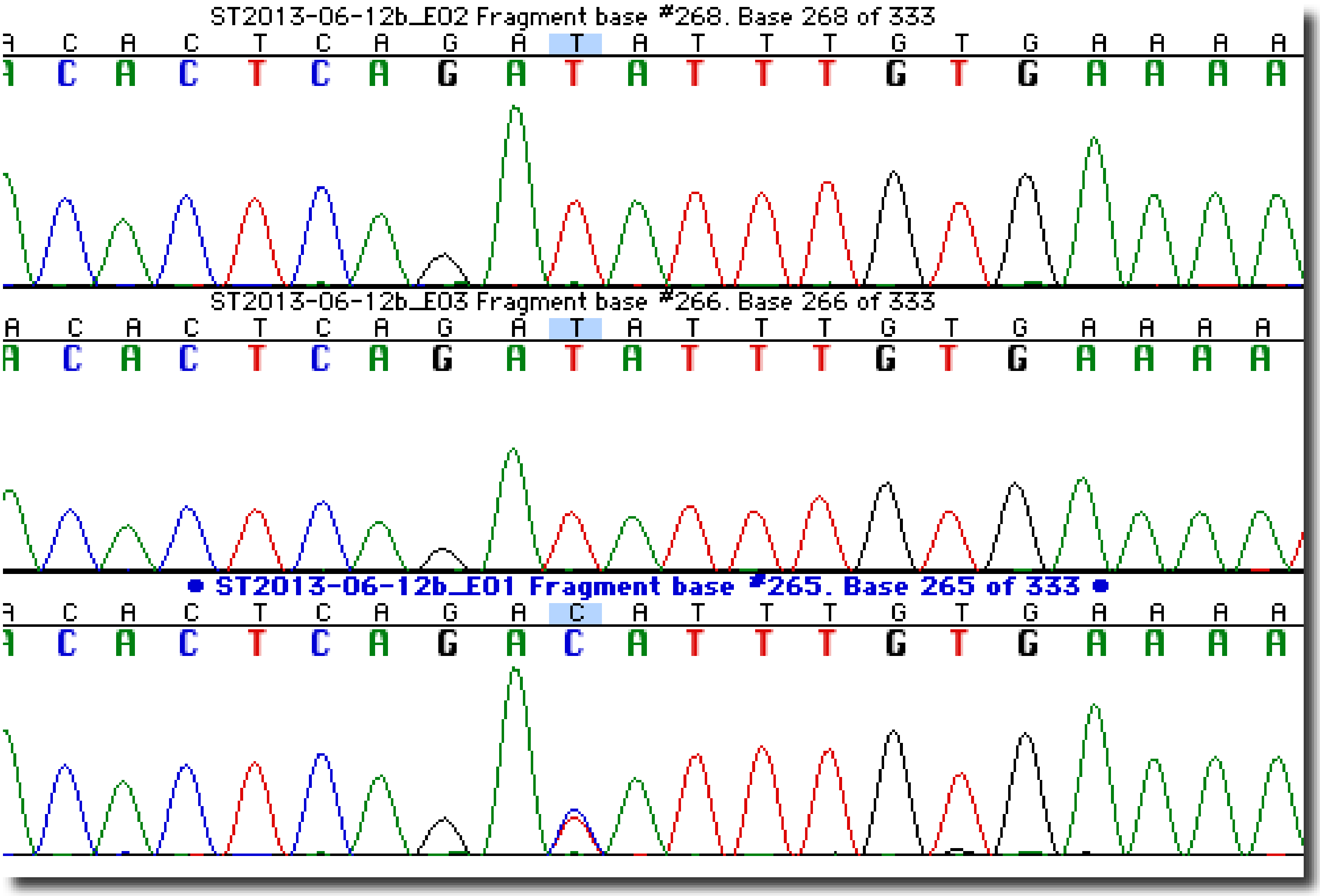
Chr18_53547194_AG_Batch2_G2G6G10.png

Supplementary Figure 5ay Validation result for Chr18 position 53547194 (AG)



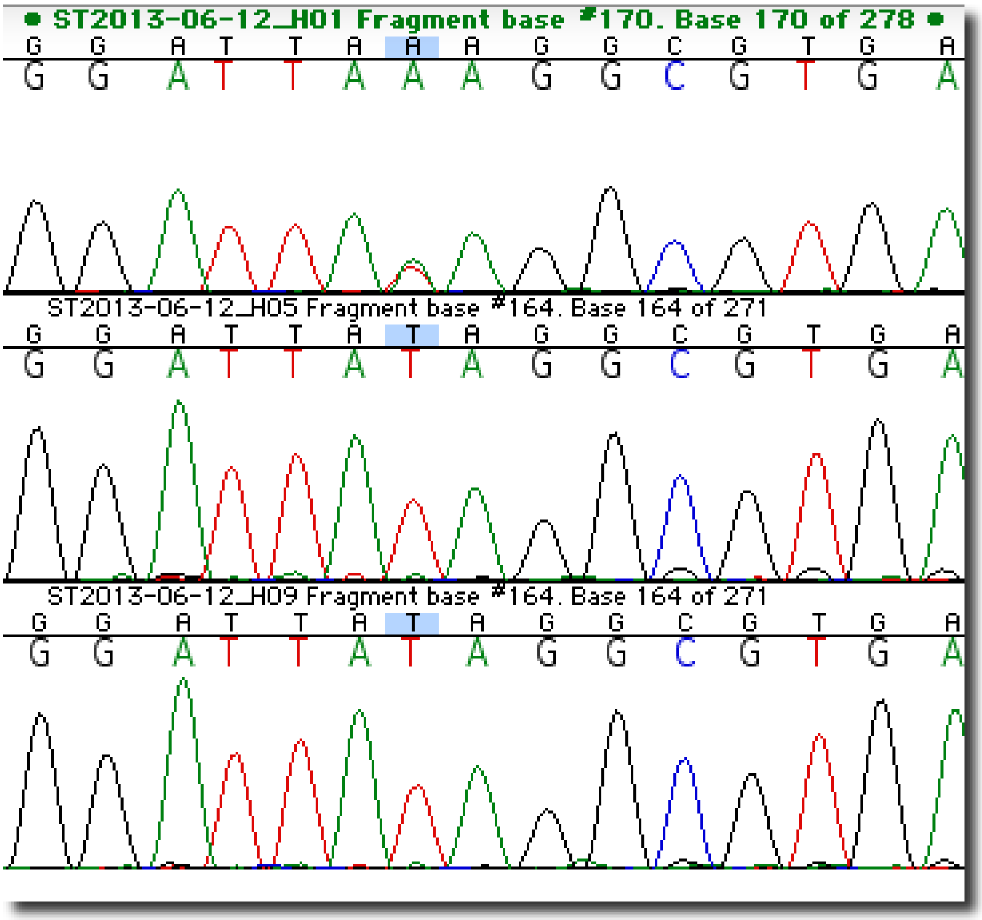
Chr19_17562750_GA_Batch4_D4D5D6D10D11D12_reverse.png

Supplementary Figure 5az Validation result for Chr19 position 17562750 (GA)



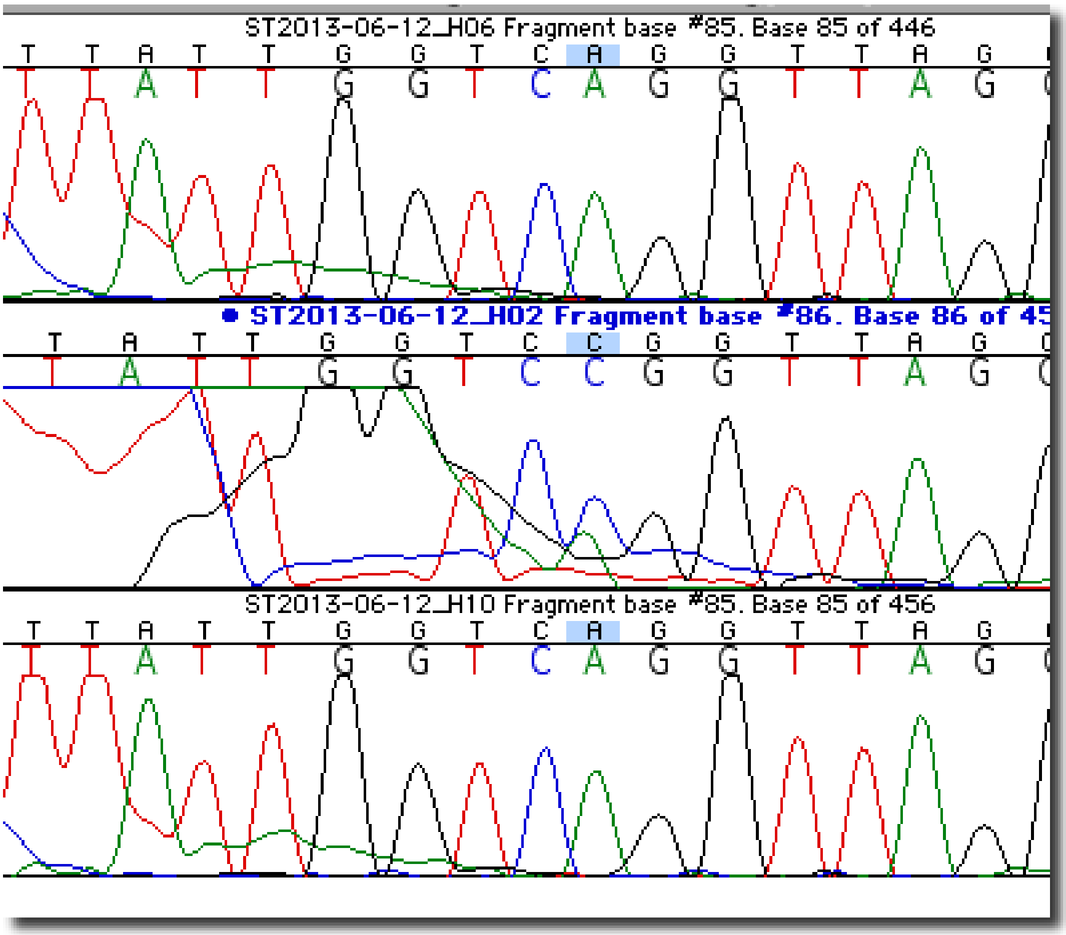
Chr21_29597244_TC_Batch3_E1E2E3.png

Supplementary Figure 5ba Validation result for Chr21 position 29597244 (TC)



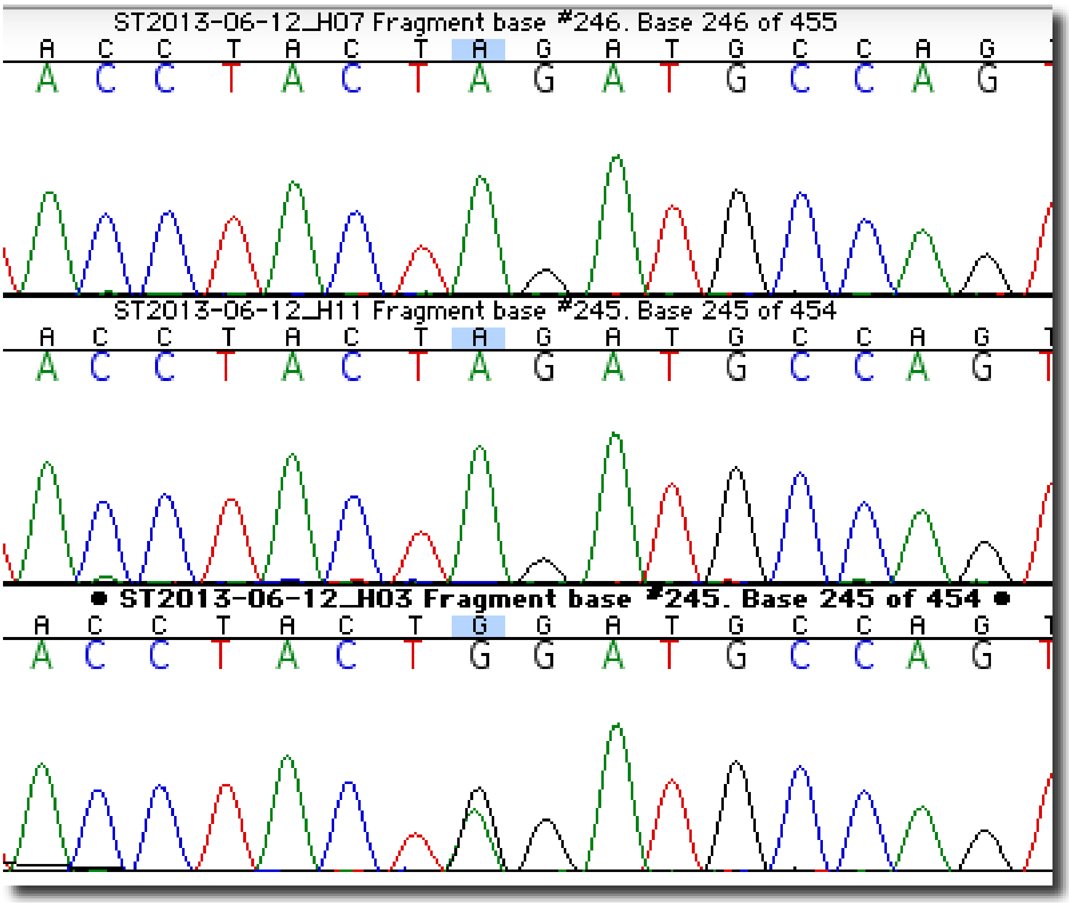
Chr22_42260599_TA_Batch2_H1H5H9.png

Supplementary Figure 5bb Validation result for Chr22 position 42260599 (TA)


ST2013-06-12_H06 Fragment base #85. Base 85 of 446
T T A T T G G T C A G G T T A G
• ST2013-06-12_H02 Fragment base #86. Base 86 of 45
T A T T G G T C C G G T T A G C
ST2013-06-12_H10 Fragment base #85. Base 85 of 456
T T A T T G G T C A G G T T A G

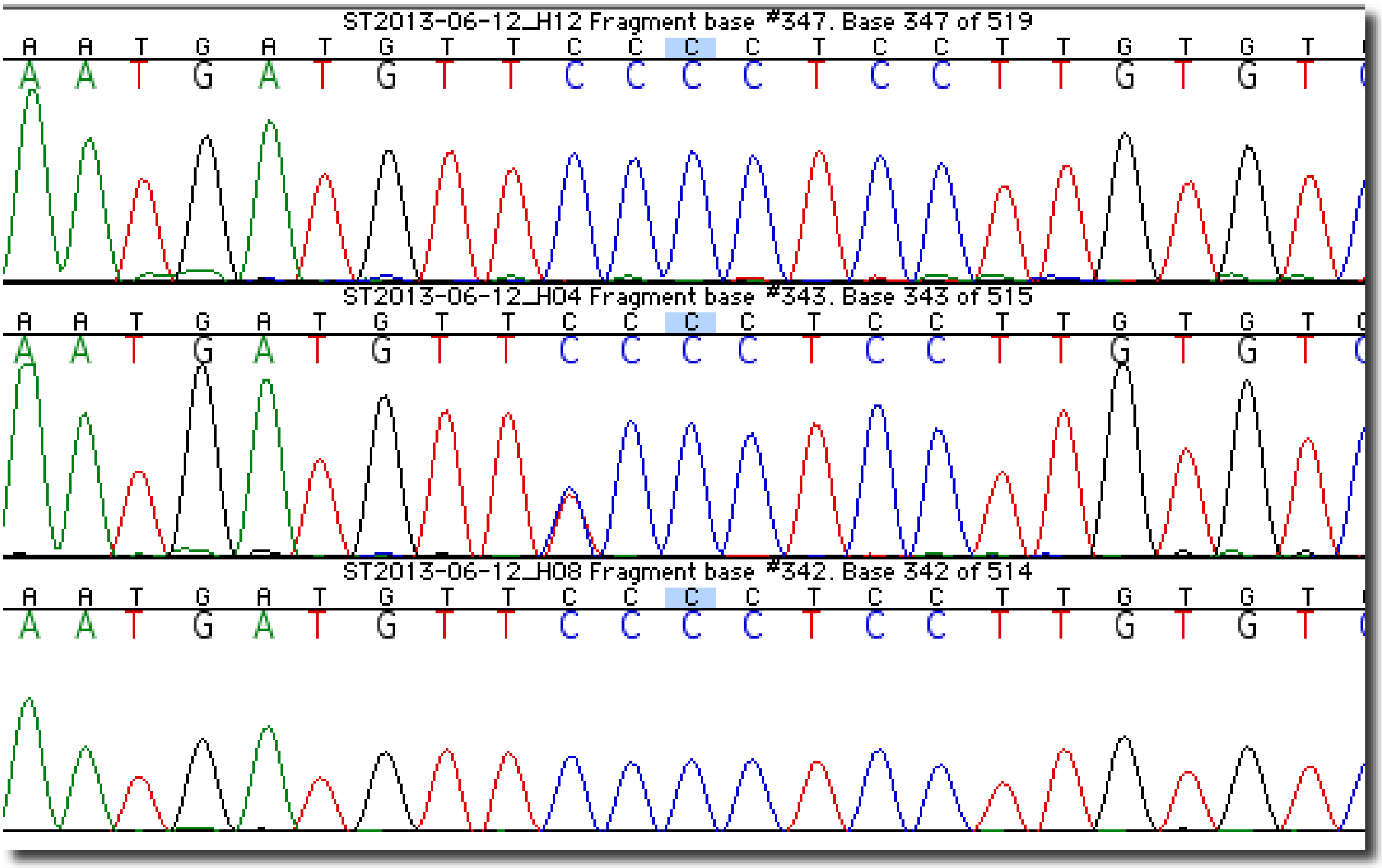
Chr22_48380824_AC_Batch2_H2H6H10.png

Supplementary Figure 5bc Validation result for Chr22 position 48380824 (AC)



ChrX_12965284_AG_Batch2_H3H7H11.png

Supplementary Figure 5bd Validation result for ChrX position 12965284 (AG)



ChrX_75334345_CT_Batch2_H4H8H12.png

Supplementary Figure 5be Validation result for ChrX position 75334345 (CT)

Supplementary Tables

chrom	position	ref	alt	status	comment	potential reason for probe failure
1	6505934	A	G	FP	no evidence	-
1	27995521	T	C	likely FP	repetitive, complex	-
1	144620091	C	T	likely TP	strong evidence, low mapq reads	nonspecific hybridization
1	144871755	A	T	TP	strong evidence	hybridization failure due to nearby SNP
1	144871782	A	G	TP	strong evidence	hybridization failure due to nearby SNP
1	148178291	A	G	likely TP	strong evidence, low mapq reads	nonspecific hybridization
1	245850513	C	G	TP	strong evidence	unclear
10	39141706	T	C	TP	strong evidence; low complexity	nonspecific hybridization
11	71712875	T	C	FP	no evidence	-
11	99690428	T	G	TP	strong evidence	unclear
12	11420866	C	T	TP	strong evidence, low mapq reads	nonspecific hybridization
12	53431298	T	A	TP	strong evidence	unclear
12	113543517	A	C	likely FP	weak evidence	-
12	123200334	T	C	TP	strong evidence	unclear
14	81609703	A	C	FP	no evidence	-
15	89400023	A	G	TP	strong evidence, low mapq reads	nonspecific hybridization
16	1291318	G	C	TP	strong evidence	hybridization failure due to nearby SNP
16	3119304	A	G	TP	strong evidence	hybridization failure: nearby 1bp ins
16	87925439	T	C	TP	strong evidence	unclear
17	21319007	G	A	TP	strong evidence	hybridization failure; variant is a MNP
17	45820022	A	C	FP	weak evidence	-
17	55183813	A	G	TP	strong evidence, low complexity	nonspecific hybridization
17	79514378	C	T	TP	strong evidence, low complexity	nonspecific hybridization
18	14537970	C	T	TP	strong evidence	unclear
19	9011412	C	T	TP	strong evidence	unclear
19	36336437	T	C	FP	weak evidence	-
19	43382368	T	G	TP	strong evidence	hybridization failure due to nearby SNP
19	44223138	G	T	TP	strong evidence	unclear
19	48305586	G	A	likely FP	repetitive, complex	-
2	95537568	C	T	TP	strong evidence	hybridization failure due to nearby SNP
2	130899940	T	C	TP	strong evidence	unclear
2	230914604	C	A	FP	no evidence	-
22	29661524	A	C	FP	weak evidence	-
4	1389156	T	C	TP	strong evidence, low mapq reads	nonspecific hybridization
4	88537088	A	G	likely TP	strong evidence, low mapq reads	nonspec. hybridization; nearby deletion
5	140573461	G	A	TP	strong evidence	nearby SNP
6	31733466	T	C	TP	strong evidence	unclear
6	83892687	C	A	FP	no evidence	-
7	61887144	C	T	TP	strong evidence	hybridization failure due to nearby SNP
7	100647594	A	C	TP	strong evidence, low mapq reads	nonspecific hybridization; nearby SNP
8	143957129	G	T	TP	strong evidence	nonspecific hybridization
X	140994200	T	C	FP	weak evidence	-

Supplementary Table 1. Axiom NA12878 homozygous reference sites where variants were called by Platypus

Chrom	pos	ref	alt	read support	coverage	freq	notes
1	103741404	a	g	1	42	**0.024**	
2	16388891	t	a	2	58	**0.034**	
2	82223415	t	c	1	54	**0.019**	
2	188011122	a	g	5	65	0.077	
3	27870886	a	g	4	49	0.082	
3	79759331	g	a	0	46	**0.000**	
3	97612327	g	a	1	57	**0.018**	
5	89633475	t	g	4	67	0.060	
5	92276047	a	g	1	56	**0.018**	
5	124141465	a	t	0	58	**0.000**	
5	163540981	c	t	1	44	**0.023**	
6	14030291	g	t	1	64	**0.016**	
6	143177923	a	c	4	74	0.054	
6	154170812	a	c	2	43	**0.047**	
7	92322120	a	g	3	53	0.057	
7	158804334	c	a	12	52	0.231	1
9	72601126	c	a	3	58	0.052	
9	113373306	a	c	0	67	**0.000**	
12	117872773	c	t	0	72	**0.000**	2
14	21623000	t	c	3	35	0.086	
15	79168281	g	a	9	86	0.105	
16	34745406	c	a	5	81	0.062	
16	83795994	t	a	0	61	**0.000**	3
18	51634409	a	g,t	0	54	**0.000**	4
20	31207586	g	a	4	49	0.082	
X	70800319	a	t	4	56	0.071	
X	139556130	t	c	1	49	**0.020**	
X	148644064	g	t	5	60	0.083	

Supplementary Table 2. Characteristics of missed cell line artefact calls from Conrad *et al.*
1, Many reads classified as "bad reads" and filtered out. 2, 35 bp insertion called at 117872774; multiple alleles apparently segregating, but difficult to interpret. 3, AT→A homozygous deletion called at 83795993, shifting A^8 homopolymer across T causing local T→A change; likely miscall in Conrad *et al.*. 4, CAA→C deletion called at 51634408 shifting GA repeat across AA causing local A→G change; likely miscall in Conrad *et al.*.

Sample	chrom	pos	hom ref	het	hom alt	call
1	2	144702257	1	0	0	reference
1	4	3633192	0.298	0.697	0.005	(uncertain)
1	16	84096015	1	0	0	reference
2	1	23258855	1	0	0	reference
2	21	19452145	0.044	0.956	0	**variant**
2	22	21067722	0.918	0.082	0	(uncertain)
2	X	114350757	0	n/a	1	**variant**
3	15	58072392	0	1	0	**variant**
3	19	7448875	1	0	0	reference
4	5	33849611	0.978	0.022	0	reference
4	5	66193799	1	0	0	reference
4	7	62704425	1	0	0	reference
4	11	6852763	1	0	0	reference
4	17	11846721	1	0	0	reference
5	3	177421362	1	0	0	reference
5	11	82271220	1	0	0	reference
5	15	22692408	0.733	0.263	0.004	(uncertain)
6	5	105524432	1	0	0	reference
6	17	77672293	1	0	0	reference
7	16	70172890	0.118	0.734	0.148	(uncertain)
7	5	162887	1	0	0	reference
7	1	97434858	1	0	0	reference
7	3	68320105	1	0	0	reference
7	20	15872969	1	0	0	reference
8	5	45092855	0.999	0.001	0	reference
8	12	121042254	1	0	0	reference
9	1	17208049	0.92	0.08	0	(uncertain)
9	15	70490476	1	0	0	reference
9	15	86026190	1	0	0	reference
9	18	14868627	0.698	0.302	0	(uncertain)
9	20	47382781	1	0	0	reference
10	11	108971824	1	0	0	reference
11	4	95021335	1	0	0	reference
11	6	29873313	0.644	0.355	0	(uncertain)
11	6	141323853	1	0	0	reference
11	9	108508983	0	1	0	**variant**
11	17	3757189	0.191	0.804	0.005	(uncertain)
12	6	17579959	0.994	0.006	0	reference
12	8	117398796	1	0	0	reference
12	11	130984345	0.997	0.003	0	reference
12	17	11192920	0.012	0.963	0.025	**variant**
13	10	72626957	1	0	0	reference
13	14	86146440	0	1	0	variant
13	14	86146453	0	1	0	variant
14	2	80004141	0	1	0	**variant**
14	6	31289078	0.113	0.850	0.037	**(uncertain)**
14	6	34752422	1	0	0	reference
14	19	43581323	0.008	0.992	0	**variant**
15	4	59343903	1	0	0	reference
15	10	471495	1	0	0	reference
15	12	57270338	1	0	0	reference
15	13	79441515	1	0	0	reference

Supplementary Table 3. Imputed genotypes posteriors at polymorphic predicted DNM loci.

	HLA-A		*HLA-B*		*HLA-C*	
True allele (genotype)	**A*11:01 /**	**A*01:01**	**B*56:01 /**	**B*08:01**	**C*07:01 /**	**C*01:02**
Estimated alleles	A*11:01:01	A*01:01:01:01	B*56:01:01	B*08:01:01	C*07:01:01:01	C*01:02:01
	A*11:03	A*01:01:01:02N			C*07:01:01:02	C*01:02:11
	A*11:25	A*01:04N				
	A*11:60					

Supplementary Table 4. Estimated genotypes of MHC Class-I genes in NA12878 using Platypus calls.

Platypus	SNP			Indel			Complex variant		
	0/0	0/1	1/1	0/0	0/1	1/1	0/0	0/1	1/1
Absent in fosmid	1732	2021	8	193	241	12	13	28	2
Present in fosmid	399	1729	1891	516	333	250	30	25	17

Samtools	SNP			Indel			Complex variant		
	0/0	0/1	1/1	0/0	0/1	1/1	0/0	0/1	1/1
Absent in fosmid	1824	2115	36	162	238	45	0	0	0
Present in fosmid	306	1768	1962	508	206	382	58	2	2

GATK	SNP			Indel			Complex variant		
	0/0	0/1	1/1	0/0	0/1	1/1	0/0	0/1	1/1
Absent in fosmid	1665	1844	11	249	332	25	0	0	0
Present in fosmid	659	1596	1772	452	314	289	57	5	1

Supplementary Table 5. Coincidence counts of called genotypes and fosmid presence/absence status. For each table the union of calls made from fosmid sequence, and the calls made by the relevant caller (Platypus, Samtools or GATK) from short read data within the extent of the fosmid sequences were considered.

	Sample1	Sample2	Sample3	Sample4
95% WGS	33	34	32	36
99% WGS	48	51	44	61
95% Exome	33	34	32	52
99% Exome	47	49	43	N/A

Supplementary Table 6. Required per-sample coverage thresholds to obtain 95% or 99% sensitivity for *de novo* variant detection in whole-genome data. Thresholds are specified separately for the whole genome (WGS) and the exome.

Supplementary Note

Table of contents

Detailed description of the calling algorithm

1.1 Read processing

The algorithm takes as input one or more BAM files containing reads mapped to a reference genome, and an indexed FASTA file of the reference sequence. The BAM data are processed in chunks, sorted by the mapping position in the reference sequence of the reads. To reduce disk load the algorithm first loads all reads mapping to a fairly large region (default 100 kb; option bufferSize) into memory. Mates of paired-end reads are also loaded, including those that map outside the region, for use by the assembler later. The read-group tags are used to match reads to samples, so that multiple BAM files may contribute to data from one sample, or a single BAM file may contain data from multiple samples.

While this approach works well for modest numbers of samples, it does not scale to large sample sizes (~1000 and up). The reason is that in order to fit all reads into memory, the region size needs to shrink as the sample size increases, increasing the number of region load operations. Each of these operations involves access to a BAM file, which itself requires loading the index file to identify the physical index into the file to read. For large numbers of samples, loading the index, rather than loading the reads, becomes the bottleneck. To avoid this, Platypus can optionally (option cacheFileOffsets) generate the regions to process, and load the required file offsets into memory. In this way each index file only needs to be read once.

After the reads are read into memory, an optional de-duplication step (option --filterDuplicates) removes reads that likely represent PCR copies of the same molecule, as base-calling or indel errors in these may otherwise be mistaken for real variants. In addition, reads having fewer than 20 high quality bases (option minGoodQualBases), defined as those with a base quality score of at least 20 (option minBaseQual), are also discarded.

By default, reads that have low mapping quality scores (below 20; option minMapQual), and paired-end reads that are marked by the read mapper as improperly paired, are also discarded. When Platypus uses assembly to suggest additional candidate variants, these reads can optionally be used (option assembleBadReads).

1.2 Construction of candidate variants

Three sources contribute to the list of candidate variants that are considered by the algorithm:

(i) Single-nucleotide variants (SNVs), multi-nucleotide variants (MNVs), short indels and small replacements as reported in the CIGAR string and read sequences of the BAM files

(ii) Alleles constructed using local assembly

(iii) Any auxiliary variants provided from a (bgzip-compressed, tabix-indexed) VCF file.

Each of these sources are optional and controlled by the user (options: getVariantsFromBAMs; assemble; source); e.g. Platypus can be made to genotype only a list of specific alleles fed from a source VCF, while for a normal variant calling run, only variants from read alignments and local assembly are used.

1.2.1 Variants from read alignments

Read alignments as reported by the CIGAR string in the BAM input files are used directly to produce SNP and indel candidate variants. When two or more sequence mismatches occur next to each other, they are merged to form a multi-nucleotide polymorphism (MNP) candidate.

When multiple mismatches appear within a specified distance (option minFlank), they are merged to form a sequence replacement polymorphism.

Candidate variants that occur closer than 10 bp (configurable with option minFlank) to either of the read's edges are discarded.

1.2.2 Variants from assembly

Assembly is done in short (by default 1.5 kb) regions at a time. As input, the assembler uses the reference sequence in the region, the reads mapping directly to that region, and their mates (whether mapping to the region or not).

The assembler starts by building a standard (colored) de Bruijn graph. Since the orientation of all reads (including that of unmapped mates, and of mates mapping outside the region) is known, a directed graph is built for the forward strand only. The algorithm uses a fixed *k*-mer size specified by the user, with a default of 15 (option: assemblerKmerSize).

Next, any "bubbles" in the de Bruijn graph that both begin and end at the reference are identified using an exhaustive loop-avoiding depth-first search. This algorithm produces a candidate allele as soon as the reference sequence is found again, and then backtracks in order to find further alleles. In detail, the algorithm works as follows. First, the algorithm traverses the reference sequence and checks whether the current node has an outgoing edge to a non-reference node; if not, a subsequent reference node is examined until a qualifying node is found, or until no more reference nodes exist. When a qualifying node is identified, the search is started by traversing the graph in the edge order, in a "depth first" fashion. During this traversal, every node traversed is marked as "used", and nodes marked as such are avoided. The search continues until either a reference node is found, or no further outgoing edges to un-used non-reference nodes are found. In the first case, the full path from initial to final reference node is recorded. In both cases, the "used" mark is cleared, and the algorithm backtracks to consider alternative graph traversals until all possible paths are recorded.

This algorithm will always return non-self-intersecting paths, and is not hampered by either repetitive elements causing complex loops in the larger graph, nor by the existence of multiple paths sharing subsets of the graph.

From the resulting paths, reference sequence is trimmed away as much as is possible, and the resulting sequence is used as a candidate variant. Variants constructed in this way include SNPs, MNVs, indels, larger variants including deletions up to the window size, and clusters of these.

1.2.3 Candidate variants from VCF files

When a VCF file is provided as input, Platypus collects all variants listed in it (independent of genotype or filter status) and adds these to the list of candidate variants.

1.3 Priors on candidates

Platypus assigns priors at the candidate variant generation stage. The following priors are used:

- SNPs (1 bp changes): 0.33×10^{-3}. This number was obtained by assuming that SNPs occur at a density of 10^{-3} per base, and that each of the three alternative nucleotides are equally probable. The density is appropriate for single samples, and a slight under-estimate when calling from multiple samples (e.g., for 1000 samples the density is closer to 10^{-2}), providing a slightly conservative prior.

- Indels (pure insertion or deletion of sequence): we use a model that estimates the local indel rate from local sequence similarity as described in ref. 5, and Supplementary Information section 8. This assigns for example a prior of 4×10^{-4} to 1-bp indels in a homopolymer context of length 4, and a prior of 0.6×10^{-2} to 1-bp indels in a homopolymer context of length 10. Deletions in complex contexts (homopolymer run length of 3 or less) are assigned a prior of $5\times10^{-5}\times L_I(n)$, corresponding to one deletion in complex sequence for every 20,000 bp (see [5]), and an indel length distribution $L_I(n)$, for where $L_I(n)=0.25\times(3/4)^n$. Similarly, insertions are assigned a prior of $5\times10^{-6}\times\ L_I(n)$, to reflect the fact that insertions in complex sequence are about 10x less frequent than deletions (see [5]). The indel length distribution $L_I(n)$ penalizes long indels too strongly; to ensure that long indels are called, we capped the prior from below at a value of 10^{-10}.

- Replacements (>1 bp without changing the sequence length): $\max(10^{-10}, 5\times10^{-5}\times L_R(n))$, where the length distribution $L_R(n)= 0.9\times (0.1)^n$.

- Complex variants (>1bp length-changing events): 5×10^{-6}, independent of length.

1.4 Windowing

Platypus calls alleles and thereby variants in small genomic windows. This section explains how these windows are constructed.

First, any length-changing variant is left-aligned, and the set of candidates is filtered for low support. Candidates are kept if they are supported by at least 2 (option: minReads) reads with bases of quality 20 (option: minBaseQual) or above. Candidates from the assembly stage are kept if the accumulated base quality exceeds 40 (minReads*minBaseQual) for each of the variant's nucleotides. Next, the candidates are sorted by position, and grouped by position. Groups are subsequently merged if they contain overlapping variants.

Platypus then constructs windows by combining groups containing variants that are within 15 bp of each other. If, as a result, the window has become longer than the read length, or the total number of candidate variants within the window exceeds 8, the window is split into smaller windows. If the window is not skipped, but contains more than 8 variants, Platypus prioritizes candidates from assembly, and further prioritizes candidates by read support. In this way the maximum number of candidate variants considered per window is limited to 8 (option maxVariants). For certain designs, including those with large numbers of samples, and those targeting highly diverse regions, it will be beneficial to increase this threshold.

1.5 Haplotype generation

The most straightforward way to build haplotypes out of a set of candidate variants is to generate all combinations of candidate variants, and apply each set of variants to the reference sequence. For *n* variants, this results in 2^n haplotypes (disregarding excluded combinations due to overlap). For small *n* this is acceptable, but the procedure results in excessive runtime in a minority of windows with many candidate variants. To address this, we implemented an alternative stepwise prioritization algorithm that is applied for larger *n*, and that keeps the number of haplotypes that need to be considered under control. With default settings, which allows a maximum of 8 variants in a single window, this algorithm is not used, and all possible combinations are considered.

If the total number of haplotypes that must be considered exceeds a specified limit (the default is 256, or 2^8), we first pick the variant that is most strongly supported by the data, and build a haplotype from this variant. This haplotype goes into a set *H*. Next, Platypus goes through all remaining variants, and iteratively adds all combinations of existing haplotypes with that variant to *H*. Every time a new haplotype *h* is added, the set of all haplotypes currently in *H* is ranked by the maximum of likelihoods of the heterozygote (*r*, *h*) genotype across all samples, where *r* denotes the reference. At each step the top *K* haplotypes are kept, where *K*=256 by default; this number is configurable (option: maxHaplotypes). The intuition behind this approach is that in most cases, reads either support the reference or an alternative haplotype, and the likelihood of (*r*,*h*) is a reasonable proxy of the true marginalized likelihood. In cases where the true genotype is heterozygous non-reference (h_1,h_2), each candidate haplotype that approximates h_1 is penalized equally by forcing h_2 to be explained by *r*, so that the true h_1 has a good chance of making it in the list.

In the event of duplicate haplotypes being produced, i.e. the same exact sequence is produced by more than one combination of variants, the combination with the highest overall prior probability is used.

1.6 Calculation of haplotype likelihoods

This section, together with the following sections, explains how the genotype likelihoods are calculated.

The first step in the calculation of the genotype likelihood, is the calculation of the haplotype likelihood $p(\ r \mid h\ )$, where *r* and *h* denote read and haplotype respectively. These are calculated by aligning a read to the haplotype sequence. The underlying model is a hidden Markov model (HMM), and the likelihood of a read given a haplotype can be calculated using the Forward algorithm. Because this likelihood calculation needs to be performed many times, we used the Viterbi algorithm instead, which admits a more efficient implementation, and which calculates the probability of the most likely path through the HMM, i.e. the probability of the most likely alignment given the data, rather than the likelihood which would marginalize over all possible alignments. In practice, and with the proper choice of gap priors, the Viterbi algorithm is a good practical approximation of the full Forward algorithm.

The alignment algorithm includes models for base mismatches and indel errors. Mismatches are scored by adding up the Phred quality scores of mismatching bases. The likelihood of indel errors is modeled using position-dependent gap-opening Phred scores. These are pre-calculated based on the propensity of indel errors to occur given the (reference) sequence context. This model, the same as is used by Dindel[2] is a simplified version of the model for the indel prior, and considers homopolymers only, rather than homopolymers and more general tandem repeats. Note that the alignment algorithm does not model SNPs and indel mutations, since the haplotype *h* represents the hypothesized true sequence.

1.7 Estimation of haplotype frequencies

After P(*r* | *h*) is calculated for all combinations of reads and haplotypes, an Expectation-Maximization (EM) algorithm is run to estimate the frequency of each haplotype $h_1,\ldots,h_a$, under a diploid genotype model:

$$L(\ R\ |\{h_i,f_i\}_{i=1\ldots a}) = \prod_{\substack{samples\\ s}} \sum_{\substack{haplotypes\\ i,j}} f_i f_j \prod_{\substack{reads\\ r\in R_s}} \left(\tfrac{1}{2}p(\ r \mid h_i\ ) + \tfrac{1}{2}p(\ r \mid h_j)\right)$$

Here, f_i denotes the frequency of haplotype h_i in the population; *a* is the number of alleles considered, *R* and R_s denote the set of all reads, and reads from sample *s* respectively, and the sum over haplotypes extends over all ordered pairs (*i*,*j*), i.e. genotypes. In the formula above we implicitly integrate out the latent variable that determines the two haplotypes of the sample *s*. This formula holds for both heterozygous and homozygous genotypes.

1.8 Calling variants and genotypes

The posterior support for any variant is computed by comparing the likelihood of the data given all haplotypes, and the likelihood given only those haplotypes that do not include a particular variant, i.e., the likelihood in a nested model where the frequencies of haplotypes that do not include the variant are fixed to 0. For the latter model, the frequencies not fixed to 0 are scaled up to account for the estimated frequency of the excluded haplotypes:

$$P(v|R) = \frac{P(v)L(R\ |\{h_i, f_i\}_{i=1\ldots a})}{P(v)L(R|\{h_i, f_i\}_{i=1\ldots a}) + (1-P(v))L(R|\{h_i, \frac{f_i}{1-F_v}\}_{i\in I_v})}$$

where *P(v)* is the prior probability of observing variant *v*, I_v is the set of haplotype indices *i* for which h_i does not contain *v*, and $F_v = \sum_{i\in I_v} f_i$. The likelihood of reads given haplotypes and their frequencies is computed as

$$L(R|\{h_i, f_i\}_{i=1\ldots a}) = \prod_{samples} \sum_{\substack{haplotypes \\ i,j}} f_i f_j \prod\nolimits_{r\in R} (\frac{1}{2}p(r|h_i) + \frac{1}{2}p(r|h_j))$$

Variants are called when their posterior support exceeds a threshold (by default Phred score 5), using these frequencies as a prior.

Genotype likelihoods for a particular variant are calculated by marginalizing over the genotypes at other variant sites within the window being considered. The best likelihood is reported as a genotype call, and the posterior for this call is calculated in the usual way (as prior times likelihood of the call, divided by the sum of prior times likelihood over all genotypes considered), and reported as a "genotype quality Phred score" (the Phred-scaled probability of the call being wrong, or 10 times negative 10-log of one minus the posterior) in the per-sample GQ field. Using the maximum-likelihood estimates of haplotype frequencies estimated from the data itself as priors when calling haplotypes and variants works well, but tends to bias genotype calls particularly for small pedigrees and single samples. To address this, we replace the estimated frequencies by a flat prior when calling genotypes if the number of samples is below 25. This also affects the reported genotype quality, but does not affect the reported genotype likelihoods.

1.9 Reference Calling

As well as calling variants, Platypus is able to generate explicit reference calls, when there is strong support for the reference allele. This is done by examining the sequence data that maps to loci between the windows identified by Platypus to call variants; because Platypus did not construct windows, there were no candidate variants in these regions, and the read coverage is indicative of the confidence in the absence of variants. In addition, Platypus examines windows where there are candidate variants but Platypus does not make a variant call. The reference calls are output in blocks, as standard VCF, using a format first described at https://sites.google.com/site/gvcftools/home/about-gvcf. The start position, end position, and size of the reference-call block are included in the VCF INFO column. To quantify the confidence in the reference calls, we compute a Phred-scaled quality score, which is output in the QUAL column. There are 2 components which may go into this QUAL value:

1. A beta-binomial P-value, using the same parameters as used for the allele bias test, which quantifies the probability of a real variant not being seen due to low coverage or allele-biased coverage. This is used as an upper limit on the quality.
2. If there are variant candidates seen in the window, then we compute the highest posterior *p* out of all the candidates but with a flat prior across all alleles (including the reference), and assign a probability of 1-*p* to the reference allele.

1.10 Filtering

Finally, a number of filters are applied on a per-variant basis:

1.10.1 Allele bias (alleleBias)

The allele bias filter identifies variants that show support in too few reads compared to the expectation under heterozygous segregation in a diploid organism. Specifically, it rejects variants if (*i*) the fraction of reads supporting the variant allele is less than the minimum of 0.5 and a user-specified threshold frequency (default 20%; configurable via the –minVarFreq option); and (*ii*) the *p* value under a binomial model with a Beta prior is less than 0.001. We used a Beta prior to account for small biases which may exist even for good calls, for instance due to mapping or PCR biases. The parameters for the Beta distribution are fixed at α=20 and β=20.

1.10.2 Strand bias (strandBias)

The strand bias filter identifies variants whose support is skewed in terms of reads mapping to the forward and reverse strands, relative to the distribution seen in all reads. We use the distribution seen overall, rather than say a binomial distribution centered around a fraction of 0.5, because certain experimental designs can give rise to *bona fide* strand biases. Examples include exon capture, and mapping biases due to the existence of an anchoring point to one end of the sequence, but not the other.

Specifically, the reads supporting the variant are tested against a Beta-binomial distribution with parameters α and β such that that smallest of these parameters is 20, and the parameters are such that the mean of the distribution equals the ratio observed in all reads. Variants are accepted if the *p* value exceeds 0.001.

1.10.3 Bad reads (badReads)

This filter triggers when across reads supporting a variant, the median of the minimum base quality close to the focal site (default 7 bp either side, configurable using --badReadsWindow) is too low (default 15 or less, configurable using --badReadsThreshold); this identifies systematic local sequencing issues causing an excess of read errors, which are not always accurately reflected in the read quality scores. It also triggers when more than a fraction of reads are filtered out for the candidate generation stage; the default for this is 0.7 (configurable using –filteredReadsFrac).

1.10.4 Mapping quality (MQ)

We compute the root-mean-square mapping-quality of all reads covering the variant site, and filter the variant if this value is less than 40 (configurable using –rmsmqThreshold). This statistic is computed using all reads mapping to the region, before any filtering is applied. We find that the default filter threshold works well for Stampy and BWA-mapped reads; for other mappers this threshold may need to be reduced.

1.10.5 Quality over depth (QD)

In order to avoid calling variants using data from many low quality reads, we compute a value (the QD score) that reflects the total evidence in favor of the variant per read supporting the variant. As a proxy for the total evidence, we use the Phred-scaled variant posterior as reported in the QUAL field of the VCF file. To avoid a downward bias in low-coverage samples, we add the Phred-scaled prior of the variant to this.. The resulting score is divided by the number of reads that support the variant. Variants with a QD value of less than 10 (configurable using –qdThreshold) are flagged as suspicious.

1.10.6 Posterior variant quality (Q20)

Although variants are called, and included in the output VCF, if they have a phred-scaled posterior exceeding 5, all variants with posteriors below 20 are flagged as suspicious.

1.10.7 Sequence context (SQ)

Polymerase slippage in low-complexity regions are a known cause of spurious indel calls, and in the alignment model we account for a higher incidence of these errors. However, in certain instances, particularly when two different but substantial low-complexity regions abut, compensating errors in both low-complexity regions can result in sequence that does not support an indel, but does support an artefactual SNP. Although these occurrences are rare, when they occur these false calls are difficult to spot, because the bases can be of high quality. To avoid calling these variants, we compute a sequence complexity statistic that measures the contribution of the two most frequent nucleotides among the 21 around a site; if this measure exceeds 95%, SNP calls are flagged as suspicious.

2 Calling variation from whole-genome high-coverage data

2.1 Data and calls

We obtained BWA-aligned data of the CEU trio NA12878/NA12891/NA12892 from the 1000 Genomes website, ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120117_ceu_-trio_b37_decoy/CEUTrio.HiSeq.WGS.b37_decoy.NA12878.clean.dedup.recal.20120117.bam, and similarly for the NA1891 and NA12892 individuals.

GATK calls made from these data at the Broad institute were downloaded from the Broad FTP site, ftp.broadinstitute.org/gsapubftpanonymous/bundle/2.3/b37/CEUTrio.HiSeq.WGS.-b37.bestPractices.phased.b37.vcf.gz. These calls were made using the latest UnifiedGenotyper with default arguments run on the NA12878 data set (see www.broadinstitute.org/gatk/guide/article.php?id=1213). We used the default filters as provided in the VCF file.

GATK Haplotype Caller calls were made with GATK 2.5, using best practices as described on the GATK website.

Samtools calls were made using Samtools version 0.1.18 (revision 982:295) with the default options. For filtering we used the recommended protocol available at http://sourceforge.net/apps/-mediawiki/samtools/-?title=SAM_protocol.

For Platypus calls we used version 1.0 with default options, and used the default filters.

Calls from Platypus, GATK and Samtools are available at http://www.well.ox.ac.uk/platypus-paper-data

2.2 Post-processing multi-nucleotide variants

Callers differ in their handling of multi-nucleotide variants (MNVs), with some reporting such variants as separate SNVs each in their own VCF record, while others report these as a single multi-nucleotide alternative allele in a single VCF record. In order to achieve a fair comparison of sensitivity for SNVs, we post-processed each VCF file so that length-invariant replacements (such as e.g. ACG→TCC) were expanded into their "constituent" SNVs (A→T and T→C in the example).

2.3 Variants from fosmid data

To obtain a set of validated variant calls, we used existing fosmid data collected as part of the 1000 Genomes project (ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120627_-NA12878_fosmid_data). This set of mapped and aligned fosmid sequences were in part selected to be enriched for structural variation[1], and contain many segments with a high density of variants that are difficult to interpret. We therefore filtered these sequences by the following criteria:

1. Fosmid sequence was required to map in at most 2 pieces; and if mapping in two pieces, these were required to be at most 60kb apart.

2. Fosmids were required not to overlap with any other; we identified these cases by hand and removed fosmids with identifiers 262360140, 290775907, 257471685, 209360443, 226246976, 281428032 and 224858439.

3. Fosmids were required to have been sequenced using Sanger sequencing, rather than next-generation sequencing technologies, to avoid possible coincidence of sequencing artefacts.

4. The density of mutations with respect to the reference was required not to exceed 3 indels or 6 SNVs in any 1 kb window.

This procedure resulted in 224 fosmid sequences covering 4.8 Mb.

We next called variants from the aligned fosmid sequences directly from the CIGAR string in the BAM file, under the assumption that the reported alignments are correct. This is reasonable since the sequences are long, virtually error-free, and contain variants at low to medium density. In order also to call complex variation, we merged calls that were directly adjacent; for instance, an insertion next to a deletion would be converted into a replacement call, and two adjacent SNV calls would be converted into a multi-nucleotide variant. This resulted in 4004 SNV calls, 1020 indel calls in length ranging from 1 to 701 bp, 5 multi-nucleotide variants and 52 complex variants.

Using this set of fosmid variants, we obtained for each call set a 2x3 coincidence matrix with rows marked as “Present/absent in fosmid”, and columns marked with “0/0”, “0/1”, “1/1” according to the called genotype. At normal biallelic sites, the reference is encoded as 0, and the alternative as 1. When an algorithm makes a multi-allelic call, all allelic variants are considered in turn, and this variant is considered the alternative encoded as “1”, while both the reference as well as any other variant is encoded as “0”. The numbers in the coincidence matrix were obtained by considering the union of variants present in the call set of interest and those observed in the fosmid data. These data are presented in Supplementary Table 5.

For any call set, “fosmid sensitivity” was computed as the fraction of variants observed in fosmids that achieved a 0/1 or 1/1 genotype. Following Ref. 1, we computed “fosmid FDR” as the fraction of homozygous alternative calls that were not observed in the fosmid data.

The data used for this analysis can be downloaded at http://www.well.ox.ac.uk/platypus-paper-data.

2.4 Comparison to AXIOM SNP chip data

The AXIOM SNP chip data were obtained from the 1000 Genomes project (ftp://ftp.ncbi.nih.-gov/1000genomes/ftp/phase1/analysis_results/supporting/axiom_genotypes/ALL.wex.axiom.20120206.snps_and_indels.genotypes.vcf.gz)

These SNP chip data include the genotypes of several individuals from the 1000 Genomes project, but we only use the variants where the genotype of NA12878 is different from “0/.” “./0” and “./.”. Denote by n_{rr} , n_{ra} and n_{aa} the number of positions where the chip genotype is “0/0” “0/1” and “1/1” respectively; let n'_{rr}, n'_{ra} and n'_{aa} be the number of such AXIOM positions where the variant caller makes a call with identical genotype as in the chip data (of “0/0” “0/1” and “1/1” respectively). Then the genotype concordance is calculated as

$$gtcon = \frac{n'_{rr} + n'_{ra} + n'_{aa}}{n_{rr} + n_{ra} + n_{aa}}$$

Let $n_v = n_{ra} + n_{aa}$, and let n'_v be the number of corresponding positions in the variant call set where the genotype is called as either “0/1” or “1/1”, then AXIOM sensitivity is calculated as $\frac{n'_v}{n_v}$.

Finally, let n_m be the number of positions where the chip genotype is "0/0" but the variant caller's genotype is not, then the monomorphic call rate is calculated as $\frac{n_m}{n_{rr}}$.

2.5 Platypus calls at homozygous reference AXIOM sites

Of the 258,914 sites that were called as homozygous reference for NA12878 by the AXIOM chip, 69 were called by Platypus as either heterozygous (64) or homozygous alternative (5). Of these 69 sites, 27 were present in a public catalogue of "high quality" variants for NA12878 collated by the GeT-RM project (http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/). Each of these calls was identified by two orthogonal technologies, and we consider these true positives.

We manually checked each of the remaining 42 calls, and classified them into "false positive", "likely false positive", "likely true positive", and "true positive" based on the raw data present in the BAM files (see Supplementary Table 1). Most calls (30/42) were classified "true positive" (27) or "likely true positive" (3); a plausible reason for hybridization failures or nonspecific hybridization for the relevant AXIOM probe could be identified for most of these calls (20/30). The remaining 12 calls were classified as "false positive" (9/12) or "likely false positive" (3/12). We therefore estimate the rate of false positive calls at AXIOM homozygous reference sites at 12/258,914 = 0.005%.

2.6 Comparison to Conrad cell line artefact calls

Before Bayesian filtering we call 924 of the 952 cell line artefacts that Conrad et al. discovered in the NA12878 cell lines [9]. We visually inspected the data at the 28 missing call sites to identify the possible cause of the missed call (see Supplementary Table 2):

- Lack of coverage was not the reason for missing these calls; coverage across the 28 sites ranged from 35 to 86 with a mean of 57.5.
- In 24/28 cases the variant was supported by very few (<10%) reads: in 13/28 cases fewer than 5% of reads supported the variant, in another 11/28 between 5% and 10% of reads supported the variant. In our experience this level of allelic unbalance is rarely seen at heterozygous sites, and we believe these cases are likely due to non-clonality of the cell line sample;
- In 2/28 cases clear support for an indel was present, and the indel was called by Platypus; misaligned reads did show support for the Conrad calls, so that these sites are likely mis-calls in the Conrad data set.
- One site showed a complex pattern which looked like the segregation of >2 alleles; Platypus called a 35 bp insertion close to the site, but it was difficult to interpret whether this could explain the Conrad call.
- One site showed substantial (12 out of 52 reads) support for the reported Conrad SNV, but Platypus did not call it because too many reads in the region were classified as 'bad reads'.

2.7 Calling indels

Samtools reports indel calls with trailing tandem repeats, to reflect ambiguities in indel call positions in the absence of a left-alignment convention. In order to simplify comparisons between Samtools and the fosmid calls, we removed these repetitive trailers. All callers report left-aligned indels, so left alignment was not performed.

2.8 Large deletions, MNPs, complex replacements, TE insertions.

We used the whole-genome data for NA12878 (described earlier) to assess Platypus's ability to identify large (>50bp) deletions and insertions. Platypus identified 28,089 bi-allelic large variants, of which 481 are insertions and 27,608 are deletions. Of the insertions, 403 are pure insertions and 78 complex replacements, and of the deletions 21,778 are pure insertions and 5,830 are complex. The large imbalance in favour of deletions is expected, largely due to the difficulty in assembling large contigs which are not present in the reference sequence. Reads mapping to newly inserted sequence will either not map anywhere, or will map to a different part of the reference genome. In cases were both reads of a pair do not map anywhere, we currently do not recover these reads for assembly/variant-calling.

The peak in the deletion length spectrum at ~300bp is due to polymorphic Alu insertions that are represented in the reference but which are not present in either or both alleles in NA12878. Platypus calls fewer Alu insertions (8), which is due to the difficulty in reconstructing the complete sequence. In order to suggest a variant candidate, the current assembly algorithm needs to assemble a complete contig which goes from the reference to new sequence and back to the reference. For Alu events, which are ~300bp in length, this requires pulling in all the reads which map in the Alu sequence, and whose mates map in the flanking sequence. Depending on the fragment size distribution, it may not always be possible to access the whole insertion with read-pair information.

Specificity for large variant calls of all kinds is expected to be high, as the assembly algorithm needs a large amount of evidence in order to suggest a candidate. For large insertions, particularly, it must be able to assemble a contig of ≥ 50 bp with good support along the whole contig.

3 Calling SNPs and indels from exome-capture data

3.1 Data and calls

This analysis was performed on BWA-aligned whole-exome capture sequence data of the CEU trio NA12878/NA12891/NA12892 from the 1000 Genomes website, ftp.1000genomes.ebi.ac.-uk/vol1/ftp/technical/working/20120117_ceu_trio_b37_decoy/CEUTrio.HiSeq.WEx.b37_decoy.NA12878.clean.dedup.recal.20120117.bam, and similarly for the NA1891 and NA12892 individuals. GATK calls, with both the Unified Genotyper and the Haplotype Caller, were made from these data using GATK version 2.5, and using best practices as described on the GATK website. Samtools calls were made using Samtools version 0.1.18 (r982:295) with default options and filters as described on the website. For Platypus calls we used version 1.0 with default options, and used the default filters.

All call sets were intersected with the whole-exome capture target region, and with a high-quality set of coding regions derived from the Vertebrate Genome Annotation database genes (vega.sanger.ac.uk), before further analysis. Multi-nucleotide variants and complex variants were processed in the same way as for the whole-genome sequencing data, and comparisons to the AXIOM chip were also performed in the same way.

Calls from Platypus, GATK and Samtools are available at http://www.well.ox.ac.uk/platypus-paper-data

3.2 Filtering Platypus Calls

For the whole-exome data, we used a slightly different set of filters for the Platypus callset, in order to take into account biases in the capture process, and the generally different coverage profile of exome data compared to whole-genome data. Specifically, we used a lower threshold (18% rather than 20%) for the minimum variant allele frequency, as the exome data shows a somewhat stronger bias against variant-alleles than is normally seen in whole-genome data, possibly due to capture bias. Other filters were the same as in the whole-genome sequence analysis.

3.3 Transition/transversion ratio

We observe a much higher transition/transversion ratio amongst SNPs in the exome data compared to the whole-genome data (~3,1 as opposed to ~2.21 for the whole-genome data). This is expected due to a combination of transversions being removed by purifying selection, and the generally higher GC content of exonic regions of the genome leading to a high proportion of CpG sites, which are known to have a high rate of transition mutations.

4 Calling *de novo* variation in parent-offspring trios

We analyzed 15 trios that were sequenced as part of a clinical re-sequencing project recently carried out at the Wellcome Trust Centre for Human Genetics, UK. Consent was obtained to study these DNA samples for the purpose of identifying a disease-causing mutation; we undertook the analyses presented here as part of that effort.

We called variants in each trio jointly, and subsequently applied the Bayesian DNM filter as described in Section 4.1.

4.1 Bayesian detection of *de novo* variation

In order to obtain a reliable set of candidate *de novo* variants, it is necessary to post-process the call set. This is because downward fluctuation in coverage in either parent can occasionally lead to parental heterozygote loss (heterozygous parental variants receiving homozygous reference genotype calls), and therefore patterns that appear to support a *de novo* variant in the child.

To avoid this, we designed a Bayesian filter to identify loci where there exists substantial evidence for a variant segregating in the child, as well as substantial evidence for the same variant *not* segregating in either parent. Specifically, this filter weighs the evidence for the data under a model for Mendelian segregation, and a model for *de novo* variation. We use a prior of 2×10^{-8} for *de novo* mutations to encode our prior belief that any offspring holds ~70 *de novo* mutations[2].

Define M to be the set of genotypes $gt = (gt^c, gt^f, gt^m)$ of the individuals in the trio, that are consistent with a Mendelian segregation pattern; the superscript c, f, and m refer to the child, father and mother respectively, and M is defined taking account of sex chromosomes and the child's sex. Similarly, define N to be the set of genotypes that are consistent with a *de novo* mutation. If we denote by d the data supporting the called genotypes, then the likelihood of d given a Mendelian segregation pattern is

$$L(d|M) = \sum_{gt\in M} L(d|gt) = \sum_{gt\in M} L(d^c|gt^c)L\left(d^f\middle|gt^f\right)L(d^m|gt^m)$$

where d^c refers to the data supporting the child genotype, and similarly for the other trio members; the per-sample likelihoods on the right-hand side are reported by Platypus in the GL field. The likelihood $L(d|N)$ of the data given a *de novo* mutation is defined analogously. The posteriors for a Mendelian segregation pattern, and a pattern supporting a *de novo* mutation, are

$$P(M|d) = \frac{L(d|M)p(M)}{L(d|M)p(M) + L(d|N)p(N) + L(d|R)p(R)}$$

$$P(N|d) = \frac{L(d|N)p(N)}{L(d|M)p(M) + L(d|N)p(N) + L(d|R)p(R)}$$

where $P(N)=2\times10^{-8}$ is the prior probability of the occurrence of a *de novo* mutation, $P(M)= 1\times10^{-3}$ is the prior probability of the occurrence of a variant following a Mendelian segregation pattern, and and $P(R)=1-P(N)-P(M)$ is the probability of a site being homozygous reference throughout the trio. We accept a variant as a *de novo* mutation if the posterior probability is larger than 0.5, and reject it otherwise.

Only variants that pass the standard filters of Platypus are considered as potential *de novo* variants. In addition, we require that the genotype quality, as reported in the GQ field, of each genotype call is at least 30, and that the ratio of the number of reads supporting the variant (reported in the VCF with the NV tag) to reads supporting the reference (reported as NR) is at most 0.03 in the parents. We also require at least 8 reads supporting the variant to be observed in the child, in order to remove a category of false positives driven by otherwise unrecognized segmental duplications or presence of repetitive sequence not included in the reference sequence leading to mapping issues.

4.2 Estimating the 95% credible interval for the FDR from validation data

Suppose that the true false discovery rate among DNMs is *p*. The probability of observing k false positives among n validation experiments is

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Assuming a uniform prior on *p* – a conservative choice – the CDF of the posterior after observing *k* can be written as

$$P(p < q|n,k) = \frac{\int_{p=0}^{q} \binom{n}{k} p^k (1-p)^{n-k} \mathrm{d}p}{\int_{p=0}^{1} \binom{n}{k} p^k (1-p)^{n-k} \mathrm{d}p} = (n+1) \binom{n}{k} B_q(k+1, n+1-k)$$

where B_q is the incomplete Beta function. For *n*=68, *k*=5, the 95% credible interval of this distribution is [0.033,0.163].

4.3 Explaining a high incidence of DNM calls at known polymorphic sites

Across the 15 trios we analyzed, we observe substantial numbers of apparent *de novo* mutations (DNMs) that have previously been identified as polymorphic in the 1000 Genomes project, ranging from 1 to 5 per sample (52 across all samples). We explored three possible explanations of this observation.

4.3.1 Algorithmic errors

The simplest explanation is that DNM predictions at polymorphic sites may be explained by inaccuracies in the algorithm, through missing variation in the parents. We have manually reviewed the underlying data, and although we find a small number of instances where the data are hard to interpret and where errors in genotyping may have occurred, in the majority of cases we find that the call is supported by clean reads, and the parents show strong support for a homozygous reference call at the site.

4.3.2 Sample contamination

A second explanation is that the samples, particularly the child's, may have been contaminated by a small fraction of human DNA from another source. Platypus is able to call variants with as few as 3 reads supporting them. Consider a site where the contaminant supports a heterozygous variant that is absent from both parents. Given an overall coverage of 25X, and a contamination fraction z, the probability that this site receives coverage by 3 reads from the appropriate allele is Poisson(25 z/2,3). Consider now a single chromosome of the contaminant. Under the assumptions of neutral coalescent theory, the polymorphisms present in the population have a frequency distribution proportional to 1/f (with some lower frequency cutoff, corresponding approximately to the inverse of the sample size, in order to make the distribution normalizable). In a contaminant's haploid chromosome, the probability of occurrence of a variant at a site where a polymorphism is segregating with frequency *f* is *f*, so that the density of polymorphisms in the chromosome is

$$C\int_0^1 f\frac{1}{f}df = C,$$

where C is a constant. Since the probability of observing a variant in one chromosome but not another, at a site where a polymorphism is segregating at frequency f, is $f(1\text{-}f)$, the total density of heterozygous sites in a pair of chromosomes is $\pi = C\int_0^1 (2f - 2f^2)\frac{1}{f}df = C$, we find that the constant C equals the heterozygosity, 0.001 per nucleotide. Now consider a single chromosome of the contaminant, and the four parental chromosomes. The density of sites at which all parental chromosomes are in the reference state, but the contaminant shows the alternative allele, is

$$C\int_0^1 (1-f)^4 f\frac{1}{f}df = \frac{1}{5}\pi$$

Since the contaminant contains two chromosomes, we double this number, ignoring the contribution of homozygous alternative alleles in the contaminant. This corresponds to approximately $3\times10^9\times\frac{2}{5}\pi = 1.2\times10^6$ sites. To account for an average of 2 polymorphic sites called as a DNM by contamination, we need that

$$1.2\times10^6\times\text{Poisson}(25z/2,3) = 3$$

or approximately a contamination fraction of z=0.20 %.

While according to this analysis, a small contamination fraction can explain the observed number of DNMs at polymorphic sites, we do not *a priori* favor this explanation. The reason is that the observed number of DNM calls at polymorphic sites is consistent with a constant rate across all 15 samples, but the expected number of spurious DNM calls is a strongly varying function of z. If contamination were the explanation of these calls, a very low but non-zero and highly reproducible level of contamination has to be hypothesized, which seems unlikely.

4.3.3 Homoplasies through mutation rate variation

A third explanation is that variations in mutation rates may increase the number of homoplasies. Mutational hotspots for both SNPs[3] and indels[4] are well documented, and mutations rates are known to vary at longer length scales as well[6].

A particularly striking example of variable mutation rates in the human genome is the hyper-mutation of methylated CpGs. As a result, CpGs are known to have mutation rates of about 20-fold above non-CpG sites[5], resulting in CpG sites occurring in the human genome at about 0.8% frequency – an under-representation by about five-fold with respect to their expected frequency of 4% – but accounting for 25% of mutations[7].

We find that the 1000 Genomes phase 1 variant collection contains 38,248,779 SNPs of which 6,716,544 occur at CpG sites, while CpG sites in the genome cover 56,490,714 bp (counting two base-pairs for each CpG site; we did not separately analyze CpG islands as they represent only 2% of all CpG sites[8]), so that a fraction 6,716,544/56,490,714 = 11.89% of CpG bases are known to be polymorphic. For the non-CpG sites, the corresponding fraction is 31,532,235/2,808,294,509 = 1.12%. In the 1000 Genomes collection CpG mutations account for 17.56% of all SNPs. Taking this fraction as representative for DNMs, we expect a fraction 0.1756*0.1189 + 0.8244*0.0112 = 0.030 or 3% of DNMs to occur on a polymorphic background, using the 1000 Genomes Phase 1 release 3 set as a reference for known polymorphic sites.

Because a relatively high fraction of CpG sites are already known to be polymorphic, homoplastic DNMs are expected to be enriched with CpG mutations. Using the numbers above, we expect 0.1756*0.1189 / 0.030 = 70% of homoplastic DNMs to be CpG-related mutations, compared to 17.56% among known SNPs.

4.4 Local haplotype structure around predicted homoplastic DNMs

The haplotype structure around homoplastic DNMs provides indirect validation for the majority of the 52 calls reported in the paper. For each call we imputed the genotype at the DNM locus from neighboring genotype calls at polymorphic loci in a 0.5 Mb window using phased 1000 Genomes data as a scaffold. We expect true homoplastic DNMs to occur on a different haplotype background than the known polymorphic variant, so that the genotype at the predicted DNM locus would be imputed as homozygous reference. If instead these calls were due to genotyping errors in either parent, the child would be expected to carry the haplotype on which the variant segregates, and a heterozygous (or homozygous) genotype would be imputed.

In detail, for each DNM call, we obtained all variant calls in the child within a 500 kb window centered around the locus. The resulting VCF file was converted into IMPUTE2's .gen format using the vcf2impute_gen from the IMPUTE2 website (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), and the call at the DNM locus was removed. We then ran IMPUTE2[10] on this file, using the pre-formatted Phase 1 release 3 reference obtained from the website http://mathgen.stats.ox.ac.uk/impute/data_download/-1000G_phase1_integrated.html), using standard options, and imputed genotype posteriors were extracted. Results are presented in Supplementary Table 3.

We find that of the 52 calls, 34 were unequivocally imputed as homozygous reference (posterior > 0.95), consistent with homoplasy. Genotypes at 9 sites were confidently imputed as heterozygous, consistent with the existence of a mutation that is heterozygous in one of the parents, but that has been called as homozygous reference, leading to a DNM miscall in the child. At the remaining 9 sites, the genotype inference did not result in a confident genotype call and neither possibility could be confidently excluded.

4.5 Estimating coverage requirements for calling DNMs

Platypus achieves its specificity for DNM in trio designs by using a Bayesian filter after standard variant calling. In the paper we show that we achieve good sensitivity and specificity on actual data sets. These data sets are characterized by relatively high average coverage. In this section we use this data to estimate the minimum average coverage required to achieve comparable specificity, and at least 95% or 99% sensitivity.

This analysis is based on the premise that the Bayesian filter achieves a given false positive rate mostly independently of the coverage; the task is to identify the minimum coverage to attain reasonable sensitivity under this filter. In order to call DNMs, we apply a Bayesian filter that requires positive evidence for the presence of a mutation in the child, as well as positive evidence for its absence in both parents. Independently of this, to achieve high specificity we additionally require a minimum number of reads (default 8) to support the variant in the child.

The Bayesian filter uses as prior probability for a DNM of $p_{\mathrm{DNM}}=2\times10^{-8}$, and as prior on a SNP a probability of $p_{\mathrm{SNP}}=1\times10^{-3}$. Under the model specified above, the posterior probability of observing N and M parental reads supporting the reference under the hypothesis of a DNM is p_{DNM}, while the posterior of observing the same under the hypothesis of a SNP is $p_{\mathrm{SNP}}(½\times1\times2^{-N} + ½\times1\times2^{-M} + ½\times½\times2^{-\mathrm{N}-\mathrm{M}}) \approx ½\ p_{\mathrm{SNP}}\ 2^{-\min(N,M)}$ to good approximation. The Bayes factor is

$$(p_{\mathrm{DNM}}/p_{\mathrm{SNP}})\ 2^{1+\min(N,M)}$$

which exceeds 1 if min(N,M) > -1 + $\log_2 p_{\mathrm{SNP}}/p_{\mathrm{DNM}}$ = 14.6, so at least 15 reads are required in both parents. In addition, at least 8 reads supporting the variant (on one of the haplotypes) are required in the child.

To estimate the probability that these conditions are fulfilled in a data set of average coverage K, we use the empirical coverage distributions of the data sets we analyzed and probabilistically re-scaled these to mimic a coverage-K data set. In detail, we computed the empirical coverage histogram $H_N(i)$, giving the fraction of genomic sites with observed coverage *i* in the data set of average coverage *N*. We model the re-scaled coverage at any site *x* that has observed coverage *i* in the original data set as a Poisson distribution with rate $\boldsymbol{\lambda}=r\ i$, where *r* is the scaling factor.

To model the parental total distributions, we use *r=K/N*, whereas to model the single-haplotype coverage in the child we use *r=K/2N*. We compute the probability that a draw *j* from one distribution exceeds the threshold (15 or 8) by computing 1 minus the cumulative distribution function. We then computed the probability that all three samples exceed their respective coverage thresholds of 15, 15 and 8 respectively, by multiplying these probabilities. Finally, we averaged the results over all observed coverages *i* weighted by the empirical histogram $H_N(i)$; we do this last, rather than averaging over the observed coverage histogram first and then multiplying out, to explicitly account for the fact that fluctuations in coverage will often be systematic, driven e.g. by GC biases, and therefore correlate across the three samples. This results in the expression

$$\sum_{i=0}^{\infty} H_N(i) \left(e^{-K/N} \sum_{j=15}^{\infty} \frac{(iK/N)^j}{j!} \right)^2 \left(e^{-K/2N} \sum_{j=8}^{\infty} \frac{(i\,K/2N)^j}{j!} \right)$$

We then varied the target average coverage *K* in order to identify the minimum target average coverage required to achieve 95% (or 99%) sensitivity.

In order to be robust against different types of GC biases and other technical biases, we carried out this procedure on a range of data sets. We also carried out the procedure independently on the whole (accessible) genome, and the exome, to account for the higher average GC content in exomes. Results are shown in Supplementary Table 6. We find that an average coverage of 32-36X is sufficient to achieve 95% sensitivity for DNMs across the whole genome. An average coverage of 32-34X also achieves 95% sensitivity within the exome for 3 out of the 4 samples we analyzed; the fourth sample requires 52X for 95% sensitivity, and no target coverage was found that would achieve 99% sensitivity due to >1% of sites in the exome receiving no read coverage in this sample, presumably because of high GC biases.

From these results we conclude that to robustly achieve 95% sensitivity for DNMs in a trio-based design using Platypus and the Bayesian filter as described in this paper, it is advisable to aim for an average coverage of 35x, with the caveat that in order to achieve this sensitivity particularly in exomes it will be important to keep GC biases under control in the library preparation stage.

The data used for this analysis can be downloaded at http://www.well.ox.ac.uk/platypus-paper-data

5 HLA genotyping from Platypus calls

The HLA region is under long-term balancing selection, and as a result has accumulated a large number of haplotype and extensive diversity, making calling challenging. By design, Platypus calls local variants, regions with confident reference support, and provides limited linkage information within calling windows of the order of one read length. To turn these local calls into HLA genotypes, we developed an algorithm that assesses the evidence provided by Platypus against all known HLA genotypes.

Specifically, from the list of variant calls $V = \{V_i\}$ and reference calls $R = \{R_i\}$ of Platypus, a set H of haplotype pairs (v_i^1, v_i^2) and $(r_i^1 \equiv r_i^2)$ was generated. The v_i^1 and v_i^2 segments were obtained by first clustering variants (and intervening reference sequence) by the reported calling window. These clusters were then extended by 12 bp up- and downstream from the variant or variants, and trimmed to ensure no overlap with any other haplotype pair. Where necessary, these haplotype pairs were further trimmed to ensure that they did not extend beyond the exonic region of the relevant HLA gene, using coordinates taken from the UCSC database (ftp://hgdownload.ucsc.edu/goldenPath/hg19/database) (see Supplementary Figure 1).

For each HLA gene targeted for genotyping (*HLA A*, *B* and *C*), we downloaded the known alleles from the IMGT/HLA database (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/; see also hla.alleles.org). Let $A = \{a_1, a_2, \ldots\ a_N\}$ be the set of known alleles available for a gene. We query the short segment set H against A using BLAST, generating BLAST scores which show the similarity of the short segments and the HLA alleles. Here, we interpret the BLAST scores as proportional to the log probability of the variant (or reference) segment being v_i given haplotype a_j, apart from a constant additive term; in short:

$$\log p(v_i|a_j) \propto \begin{cases} \max\ (S(v_i, a_j) & \text{if } L(alignment) = |v_i| \text{ and Percent identity} = 100\% \\ & 0 \text{ otherwise} \end{cases}$$

Here, we take the maximum across BLAST scores since BLAST occasionally identifies multiple partial matches against a single allele a_j. The interpretation of BLAST scores as proportional to the log probability is justified since the score is the sum of base mismatch penalties and gap open/extend penalties for the optimal alignment, which have an interpretation as probabilities when the alignment model is represented as a hidden Markov model. Given the probabilities defined above, the (scaled) log likelihood of the genotype $g = (a_i, a_j)$ is calculated as:

$$\log LK(g = (a_i, a_j)|D) \propto \sum \log p\left(V_k \middle| g = (a_i, a_j)\right) + \sum \log p\left(R_k \middle| g = (a_i, a_j)\right),$$

where

$$\log p\left(V_k \middle| g = (a_i, a_j)\right) = \max\left(\log p(v_k^1|a_i) + \log p(v_k^2|a_j), \log p(v_k^1|a_j) + \log p(v_k^2|a_i)\right)$$

since the phasing of each local pair of haplotypes is unknown, and

$$\log p\left(R_k \middle| g = (a_i, a_j)\right) = \log p(r_k^1|a_i) + \log p(r_k^2|a_j) \quad (\text{note that } r_k^1 \equiv r_k^2)$$

The pairs of alleles that achieve the highest (scaled) maximum likelihood value are chosen as the candidate genotypes for the *HLA* gene in question.

5.1 Ambiguous *HLA-A* genotype calls

The pipeline described above successfully genotyped the *HLA-B* and *HLA-C* loci up to the 4-digit resolution at which the sample NA12878 was previously typed in the laboratory. However, *HLA-A* was only genotyped uniquely and correctly up to 2 digit resolution; at 4 digit resolution ambiguities remained (see Supplementary Table 4). We here discuss the reasons for the remaining ambiguity.

5.1.1 Allele *HLA-A**11:01

For the allele that was laboratory-typed as A*11:01, Platypus identified 5 possible alleles (A*11:01:01, A*11:03, A*11:25, A*11:60, A*11:69N) with equal support.

The three alleles A*11:03, A*11:25 and A*11:60 differ from each other and the likely true allele A*11:01:01 at positions 29,911,225, 29,911,228, and 29,911,239-40 on chromosome 6 (build NCBI37) (Supplementary Figure 3a). Reads covering that location do support the correct SNPs (Supplementary Figure 3b), and Platypus makes the correct calls, but these calls were filtered out by the badReads filter triggering due to a fraction of reads with high implied fragment sizes.

5.1.2 Allele *HLA-A**01:01

For this allele Platypus identified 4 possible alleles (A*01:01:01:01; A*01:01:01:02N; A*01:01:38 and A*01:04N) with equal support.

Allele A*01:01:01:02N differs from the reference allele A*01:01:01:01 by a 4bp deletion 4 bases downstream of exon 2, causing no splicing changes, resulting in an exonic sequence identical to that of the reference allele. Since the current HLA genotyping pipeline only considers the exonic allele sequence, these alleles were not distinguished.

Allele A*01:04N differs from the reference allele by a 1 bp deletion at the beginning of exome 3, causing a frameshift and a length polymorphism of the allele (Supplementary Figure 4a). Read data did not support the 1 bp deletion and supported the reference allele (Supplementary Figure 4b), but the only impact of the deletion is a length polymorphism at the exon boundary and the HLA genotyping pipeline currently does not distinguish such alleles.

6 500 Whole-Genome Sequences (WGS500) Consortium: names and affiliations of authors

Steering Committee: Peter Donnelly (Chair)[1], John Bell[2], David Bentley[3], Gil McVean[1], Peter Ratcliffe[1], Jenny Taylor[1,4], Andrew Wilkie[4,5]

Operations Committee: Peter Donnelly (Chair)[1], John Broxholme[1], David Buck[1], Jean-Baptiste Cazier[1], Richard Cornall[1], Lorna Gregory[1], Julian Knight[1], Gerton Lunter[1], Gil McVean[1], Jenny Taylor[1,4], Ian Tomlinson[1,4], Andrew Wilkie[4,5]

Sequencing & Experimental Follow up: David Buck (Lead)[1], Christopher Allan[1], Moustafa Attar[1], Angie Green[1], Lorna Gregory[1], Sean Humphray[3], Zoya Kingsbury[3], Sarah Lamble[1], Lorne Lonie[1], Alistair Pagnamenta[1], Paolo Piazza[1], Guadelupe Polanco[1], Amy Trebes[1]

Data Analysis: Gil McVean[1] (Lead), Peter Donnelly[1], Jean-Baptiste Cazier[1], John Broxholme[1], Richard Copley[1], Simon Fiddy[1], Russell Grocock[3], Edouard Hatton[1], Chris Holmes[1], Linda Hughes[1], Peter Humburg[1], Alexander Kanapin[1], Stefano Lise[1], Gerton Lunter[1], Hilary Martin[1], Lisa Murray[3], Davis McCarthy[1], Andy Rimmer[1], Natasha Sahgal[1], Ben Wright[1], Chris Yau[6]

[1]The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

[2]Office of the Regius Professor of Medicine, Richard Doll Building, Roosevelt Drive, Oxford, OX3 7LF, UK

[3]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex, CB10 1XL, UK

[4]NIHR Oxford Biomedical Research Centre, Oxford, UK

[5]Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK

[6]Imperial College London, South Kensington Campus, London, SW7 2AZ, UK

Supplementary References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* Jan 7 2000;100(1):57-70.
2. Kong A, Frigge ML, Masson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature.* Aug 23 2012;488(7412):471-475.
3. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics.* Feb 2012;44(2):226-232.
4. Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. *PLoS biology.* Feb 3 2009;7(2):e1000027.
5. Montgomery SB, Goode DL, Kvikstad E, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome research.* May 2013;23(5):749-761.
6. Lunter GA, Miklos I, Song YS, Hein J. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *Journal of computational biology : a journal of computational molecular cell biology.* 2003;10(6):869-889.
7. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics.* Feb 1 2012;28(3):311-317.
8. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research.* Mar 2012;22(3):568-576.
9. Conrad DF, Keebler JE, DePristo MA, et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics.* Jul 2011;43(7):712-714.
10. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *Plos Genet.* Jun 2009;5(6).