American Journal of Human Genetics, Volume *92*

Supplemental Data

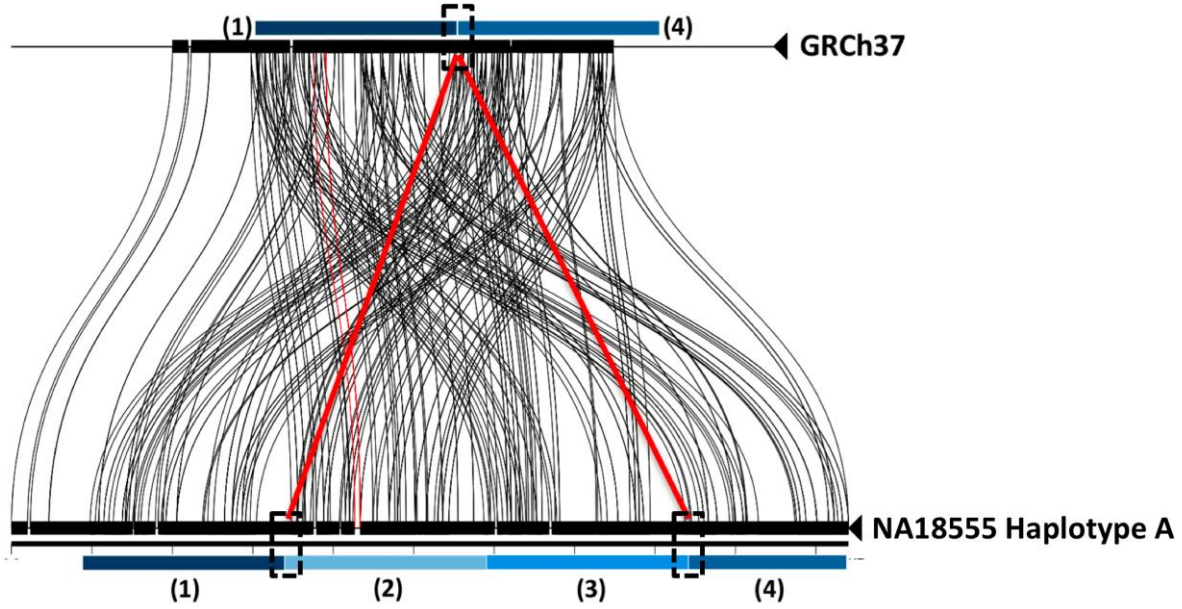
Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation

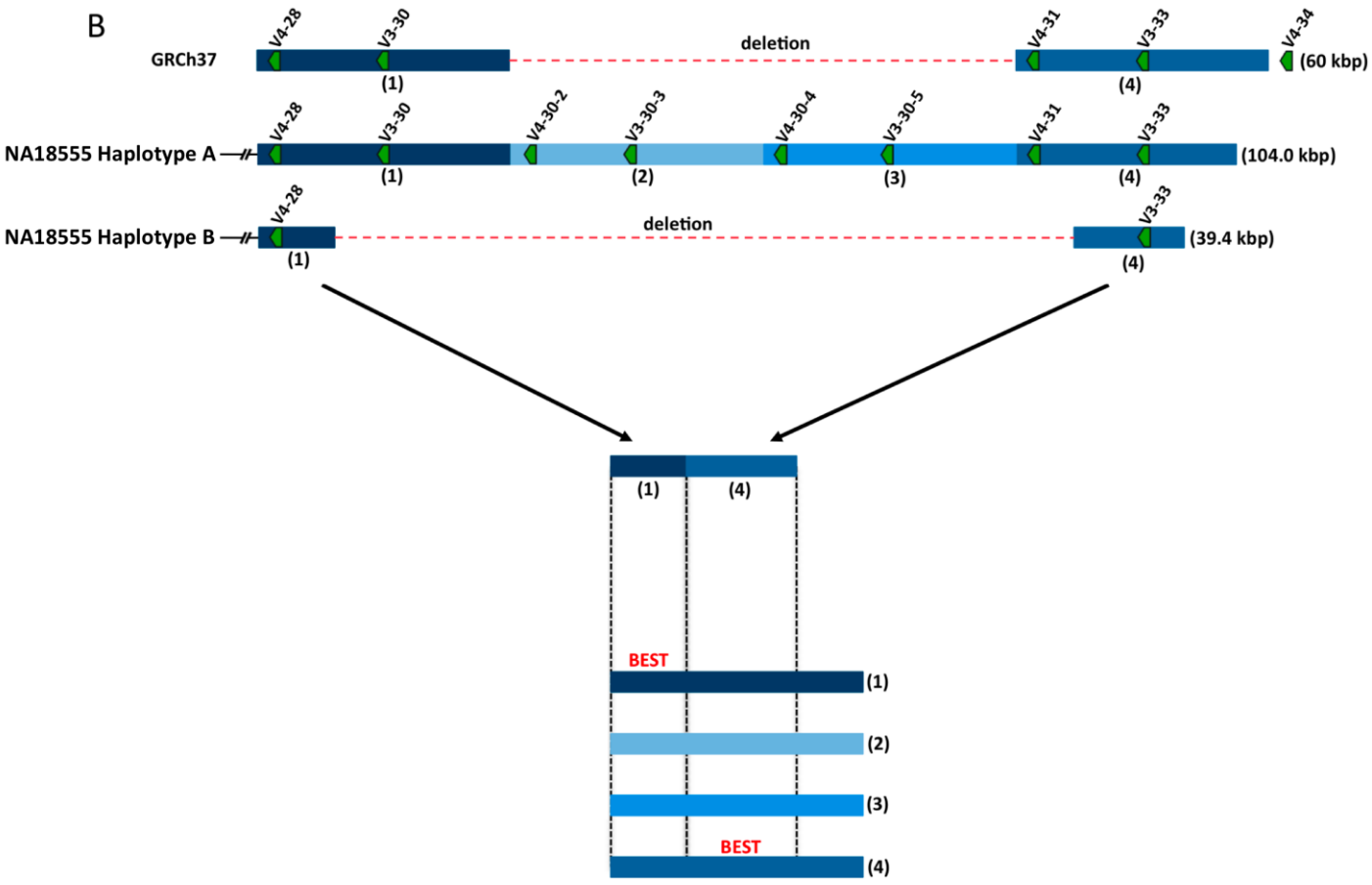
Corey. T. Watson, Karyn M. Steinberg, John Huddleston, Rene L. Warren, Maika Malig, Jacqueline Schein, A.J. Willsey, Jeffrey B. Joy, Jamie K. Scott, Tina Graves, Richard K. Wilson, Robert A. Holt, Evan E. Eichler, and Felix Breden

Supplemental Inventory

Supplemental Figures and Tables

A




B
GRCh37
V4-28
V3-30
deletion
V4-31
V3-33
V4-34
(60 kbp)
(1)
(4)
NA18555 Haplotype A
V4-30-2
V3-30-3
V4-30-4
V3-30-5
(104.0 kbp)
(2)
(3)
NA18555 Haplotype B
(39.4 kbp)
BEST

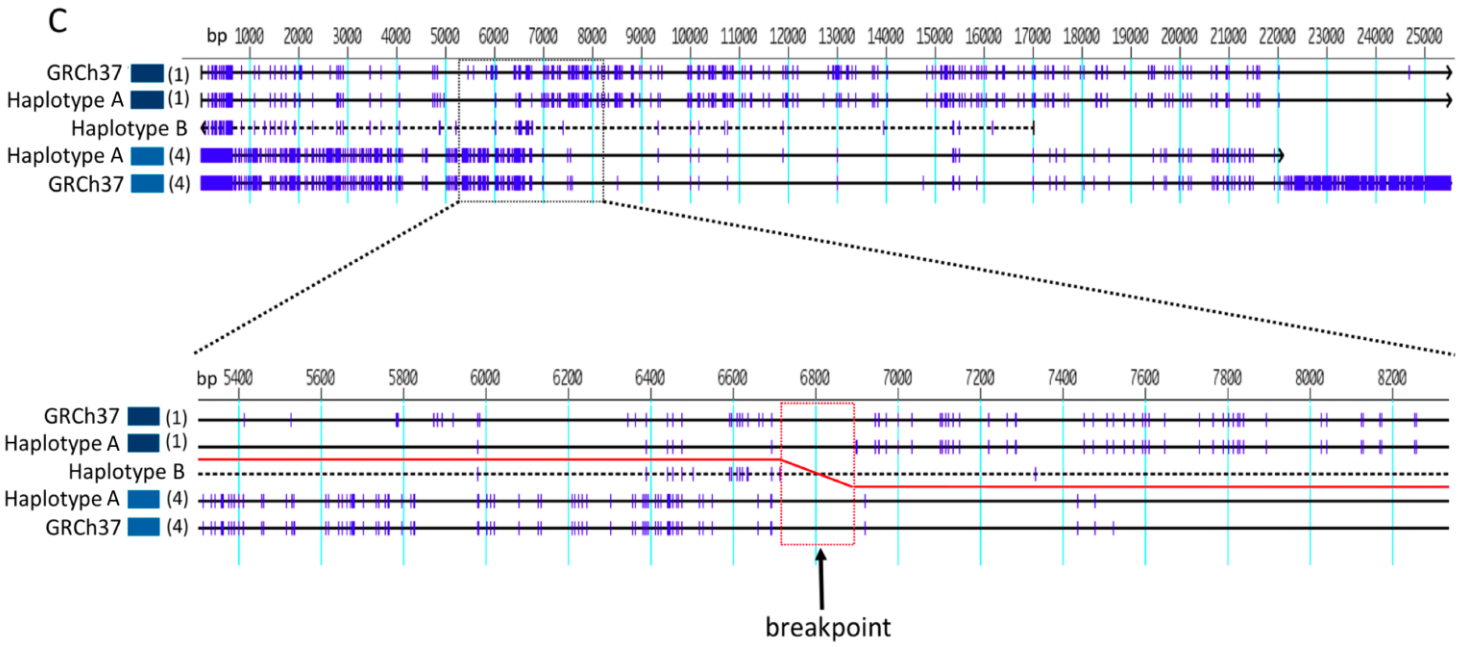


Figure S1. Method used to determine complex events in the CNV hotspot between *IGHV4-28* and *IGHV4-34*. (A) Comparison of NA18555 Haplotype A to a ~100 kbp region of GRCh37 (also see Figure 4, main text). The positions of IGHV genes in GRCh37 are shown in the top panel (red rectangles), with gene name labels indicated on the right. The black bars seen on both haplotypes connected by thin black lines represent regions of shared sequence, either homologous or paralogous sequence blocks, as determined by the program Miropeats (Parsons 1995). The blue shaded bars above GRCh37 and below NA18555 Haplotype A represent a ~25 kbp segment that is shared between the two haplotypes, the structure of which is indicative of a segmental duplication; this duplication block is represented twice in GRCh37 and four times in NA18555 Haplotype A. The four different shades of blue represent distinct versions of this segmental duplication (labelled 1-4). The bold red lines indicate the predicted breakpoints of the complex insertion event in Haplotype A as compared to GRCh37, which are also highlighted by black boxes (see B below). Given the complicated genomic structure and the number of CNV-containing haplotypes identified in this region (full list in Figure 4, main text), the breakpoint analysis method illustrated for the *IGHV3-23* duplication example in Figure 3 of the main text could not be used to identify breakpoints in this region. Thus, we developed the method depicted in (B). (B) Sequence from complete or partial segmental duplications in each of three haplotypes are shown (1-4, NA18555 Haplotype A; 1 & 4, in NA18555 Haplotype B and GRCh37), represented by blue shaded bars as in (A). For this method, the best sequence alignments between all segmental duplications 1-4 are determined; the event boundaries are then predicted from this alignment by searching for the point in the alignment at which the sequence similarities between the aligned sequences change as shown in the bottom panels of (B) and (C). (C) This image shows an expanded and zoomed-in version of a multi-sequence alignment containing sequences of the ~25 kbp segmental duplicates (shown in panels A and B) from GRCh37, and the NA18555 Haplotype A and Haplotype B haplotypes. Alignment bp positions are shown along the top of the diagram. Each sequence is represented by a single horizontal black line, labelled with a haplotype name and the duplication block color and number corresponding to those shown in panels A and B. Blue tick marks on each line indicate single nucleotide differences and gaps between a given sequence and all other sequences in the alignment; thus, tick marks observed at the same position in more than one sequence of the alignment identify positions of shared sequence. The red line tracks the most similar alignments of the middle NA18555 Haplotype B sequence to the other four sequences from GRCh37 and NA18555 Haplotype A. Based on changes in nucleotide similarity on either side of the 249 bp region in which all aligned sequences share 100% sequence identity (within red box, bottom panel), the event breakpoint is presumed to have occurred

within this segment. In the case of NA18555 Haplotype B, it is not clear whether this event arose from a haplotype like that of the human reference genome, which contains segmental duplications 1 and 4, or was mediated by a haplotype like NA18555 Haplotype A containing segmental duplications 1-4. In addition, because NA18555 Haplotype A includes two additional duplication blocks as compared to GRCh37 (panels A and B), it is not clear whether NA18555 Haplotype A occurred from a single event or is the result of more than one round of expansion via segmental duplication.

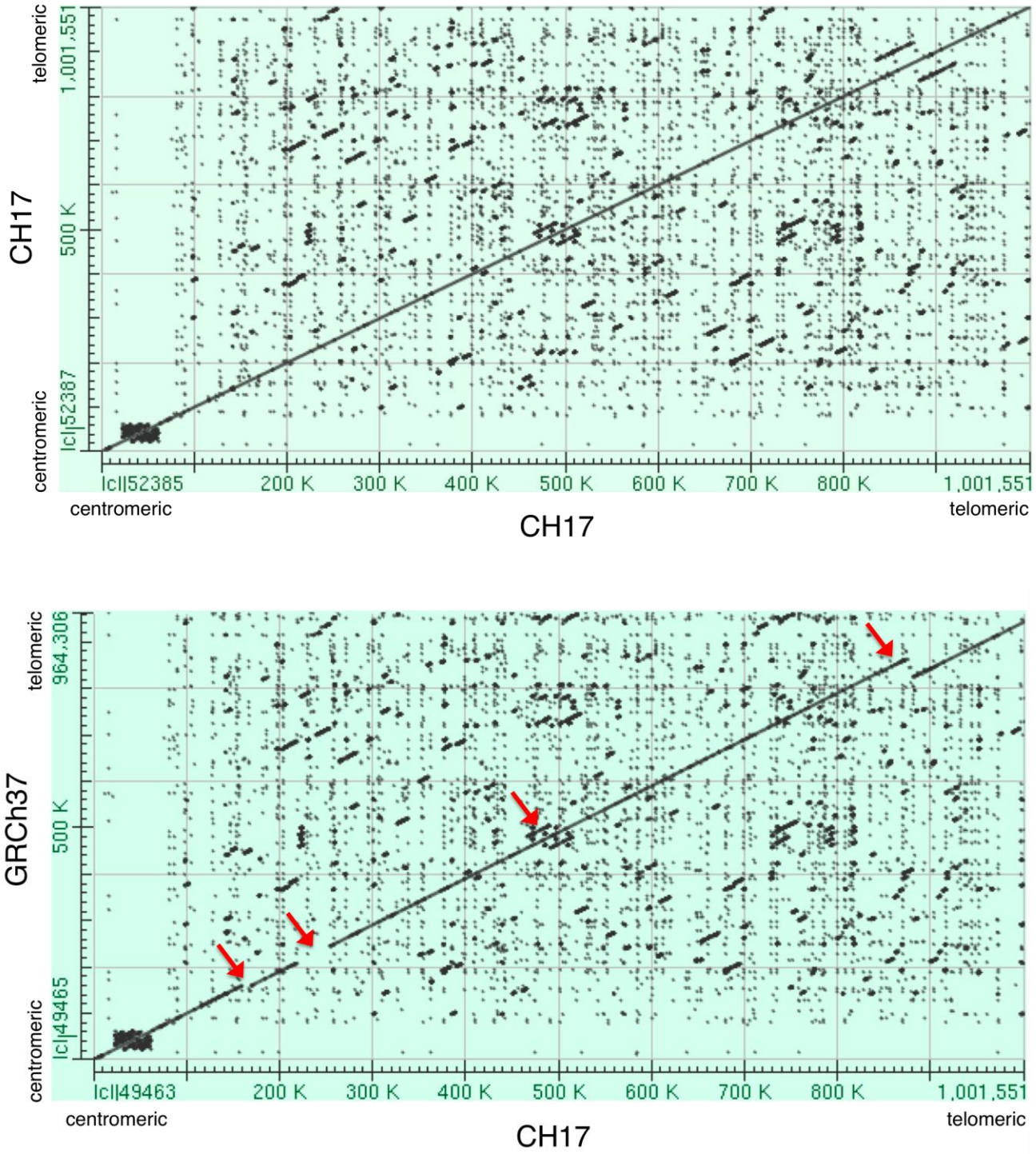


Figure S2. Dotplot analyses of CH17 and GRCh37 haplotypes
Top panel: a self pairwise BLAST of the CH17 haplotype including sequence spanning the IGHJ gene cluster to 49 kbp upstream of *IGHV7-81*. Bottom panel: a pairwise BLAST between the CH17 haplotype and GRCh37 (chr14: 106324366-107268434). Red arrows indicate the positions of CNVs depicted in Figure 1B.

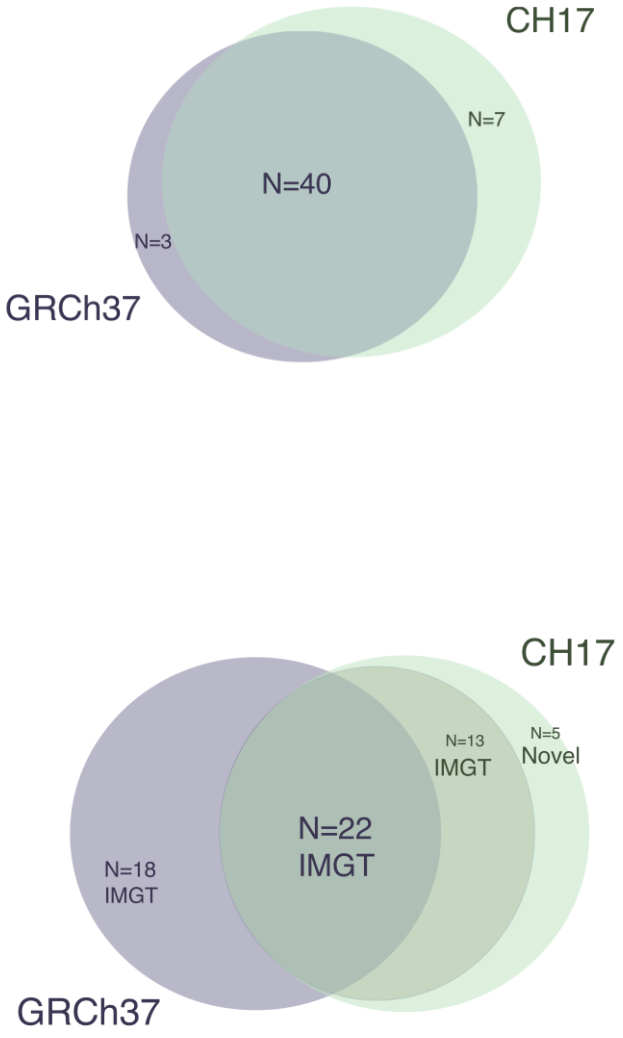


Figure S3. Comparison of IGHV gene content between GRCh37 and CH17 haplotypes. Top panel:Venn diagram depicting IGHV gene overlap between GRCh37 and CH17 haplotypes. As shown there are seven genes that are unique to CH17 and three genes that are unique to GRCh37, which are the result of gene gains and losses associated with CNVs. Bottom panel: Venn diagram depicting the overlap of identified IGHV gene alleles between the GRCh37 and CH17 haplotypes: at the 40 shared genes between the two haplotypes (Figure S3) there are 22 alleles that are found in both GRCh37 and CH17, 18 that are unique to GRCh37, and 18 that are unique to CH17. Five of the 18 alleles unique to CH17 are novel, and were not previously represented in IMGT (www.imgt.org).

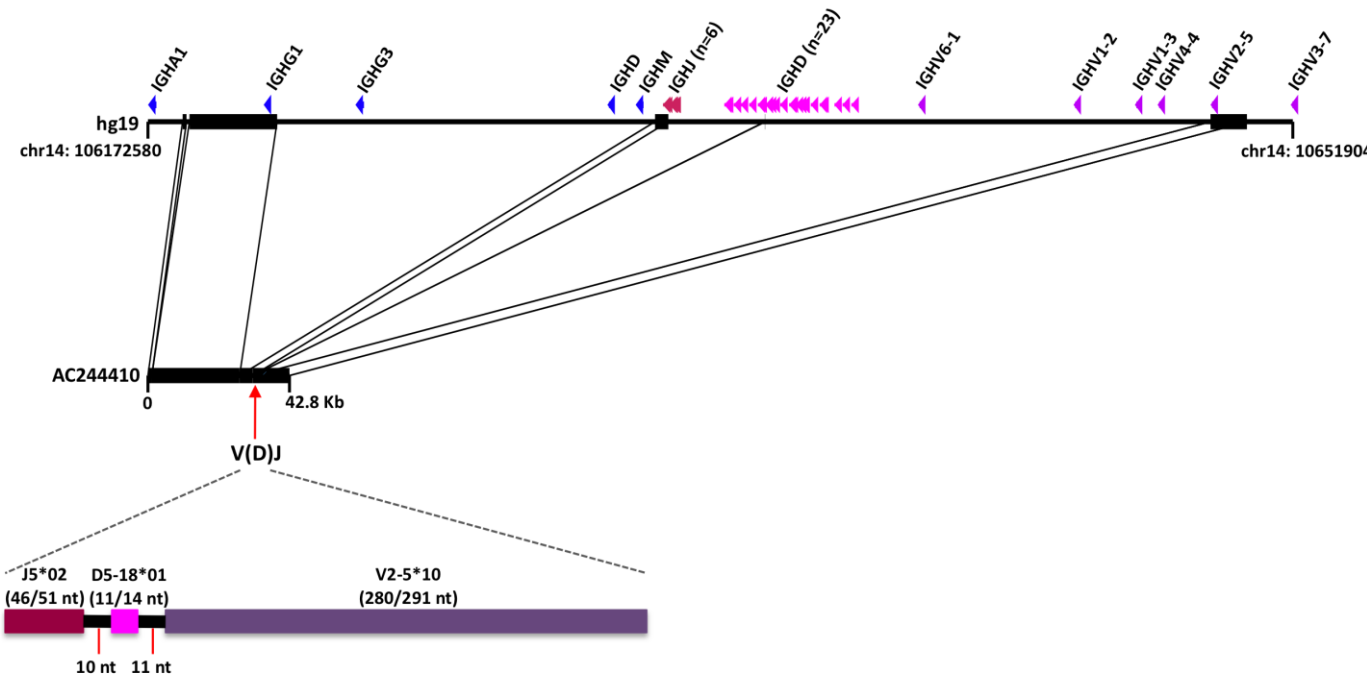


Figure S4. A sequenced V-(D)-J somatic rearrangement
Diagram of the structural comparison of a fosmid clone from the NA18507 library (AC244410) and the GRCh37 IGH reference sequence (chr14:106172580-106519042). The positions and names of IGH genes are indicated by colored triangles and black labels shown above the GRCh37 haplotype: IGHC genes (blue), IGHJ genes (magenta), IGHD genes (pink), and IGHV genes (purple). Regions of shared sequence between the fosmid clone and GRCh37 are indicated by thick black bars connected by black lines between the two haplotypes. Large stretches of sequence found in GRCh37 (depicted by thin horizontal black lines in the GRCh37 haplotype) were deleted in the fosmid; these deleted segments were located between the IGHJ, IGHD, and IGHV genes that were selected during the V-(D)-J somatic rearrangement event. In addition to the loss of large stretches of sequence in each gene region, we observed evidence of class-switch recombination (CSR), junctional diversity (JD), and somatic hypermutation (SHM)—all processes that occur at different stages in developing B cells, either during antibody rearrangement (JD) or antibody/B cell selection post-antigen stimulation (CSR, SHM). A large apparent deletion including *IGHG3*, *IGHD*, and *IGHM* is indicative of CSR. Colored bars at the bottom of the figure show a close-up view of the V-(D)-J rearrangement observed in this fosmid. Black labels indicate the names of the IGHJ gene (magenta bar), IGHD gene (pink bar), and IGHV gene (purple bar) selected in the observed rearrangement, and the most similar known alleles are noted (*IGHJ5*02*, *IGHD5-18*01*, and *IGHV2-5*10*). Numbers in parenthesis indicate the nucleotide similarity between these alleles and the IGHJ, IGHD, and IGHV gene sequences annotated in the fosmid clone; nucleotide differences are indicative of SHM. The number of inserted nucleotides between IGHJ and IGHD, and IGHV and IGHD are also shown and are indicative of JD. Analysis of nine additional clones also confirmed the presence of somatic V-(D)-J rearrangements, CSR, SHM, and JD. Point mutations characteristic of SHM were identified by comparing IGHV, IGHD, and IGHJ gene sequences to known alleles; in each of the ten somatically-rearranged fosmids sequenced, point mutations were observed in the IGHV, IGHD, and IGHJ genes involved in the rearrangement (Table S8). SHM is known to be a highly targeted process that occurs only within the rearranged IGHV, IGHD, and IGHJ genes. In support of this, for clones in which additional IGHV and IGHJ genes were present upstream and downstream of the rearranged IGHV and IGHJ genes, no evidence of SHM was observed (Table S8).

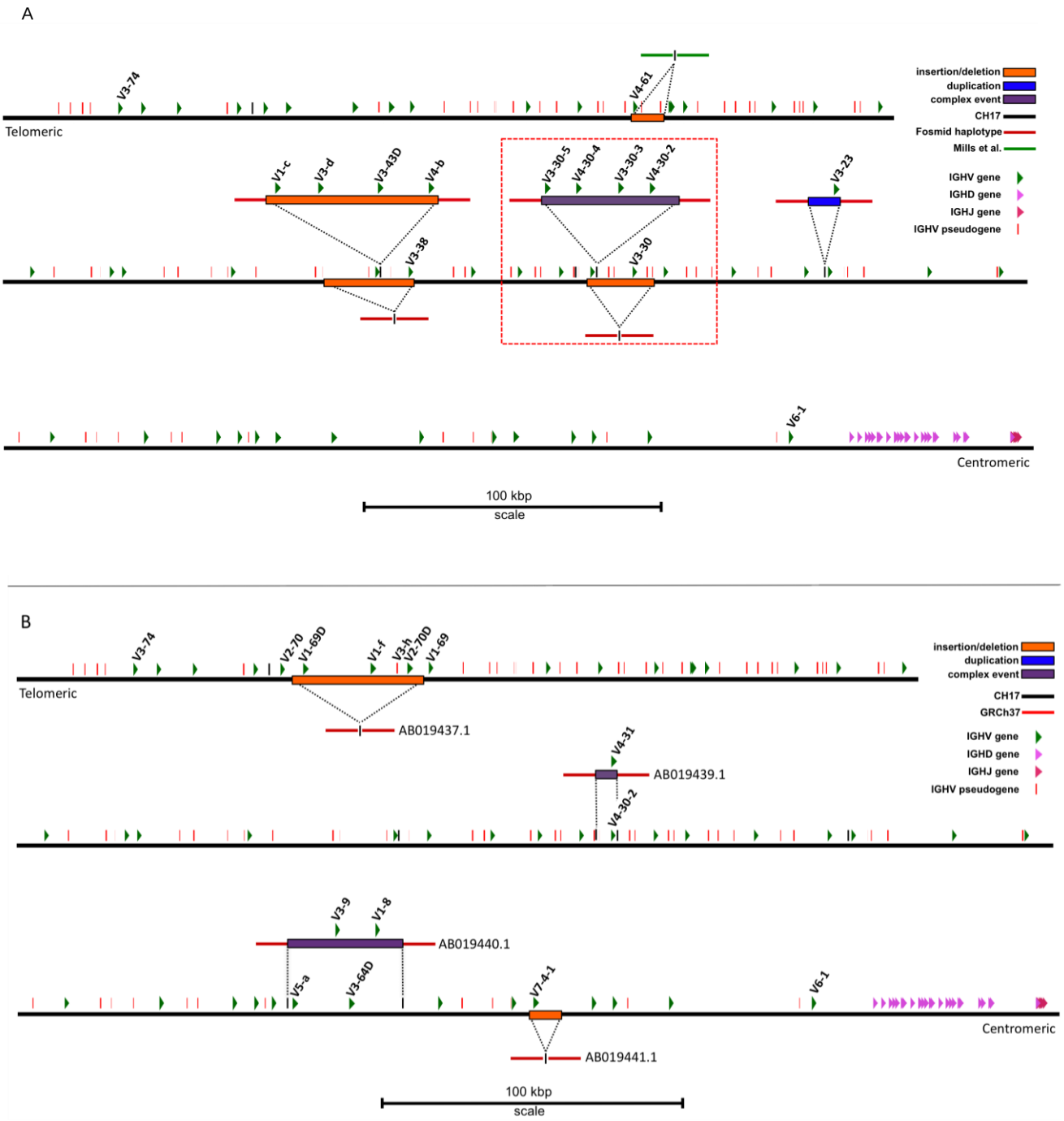


Figure S5. Mapping of CNVs to the CH17 IGH haplotype. (A) The positions of three deletions, two insertions, one duplication, and one complex event characterized from fosmid alternative haplotypes are shown mapped to the CH17 IGH reference (black line) with the same parameters as in Figure 1B. The locus is presented in the same orientation as that depicted by IMGT. Functional and ORF IGHV, IGHD, and IGHJ genes (not to scale) as well as IGHV pseudogenes are shown (to scale); the names of IGHV genes involved in the characterized structural variants are indicated. The large red box indicates a hotspot region of recurrent mutation (see Figure 4 for additional haplotypes associated with this hotspot). The deletion of *IGHV4-61* was identified by Mills et al. [5] (chr14:107,084,861-107,096,738) is also shown. (B) CNVs identified between CH17 and GRCh37 mapped to the CH17 IGH haplotype. Accession numbers for clones from the GRCh37 reference assembly are indicated.

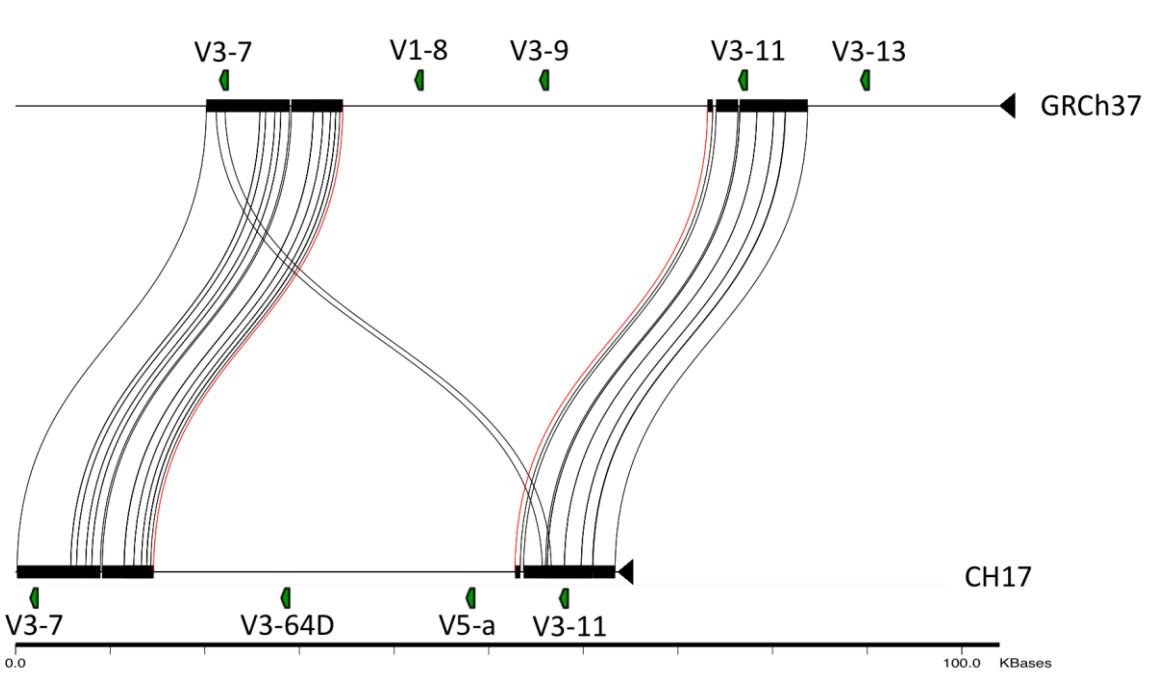
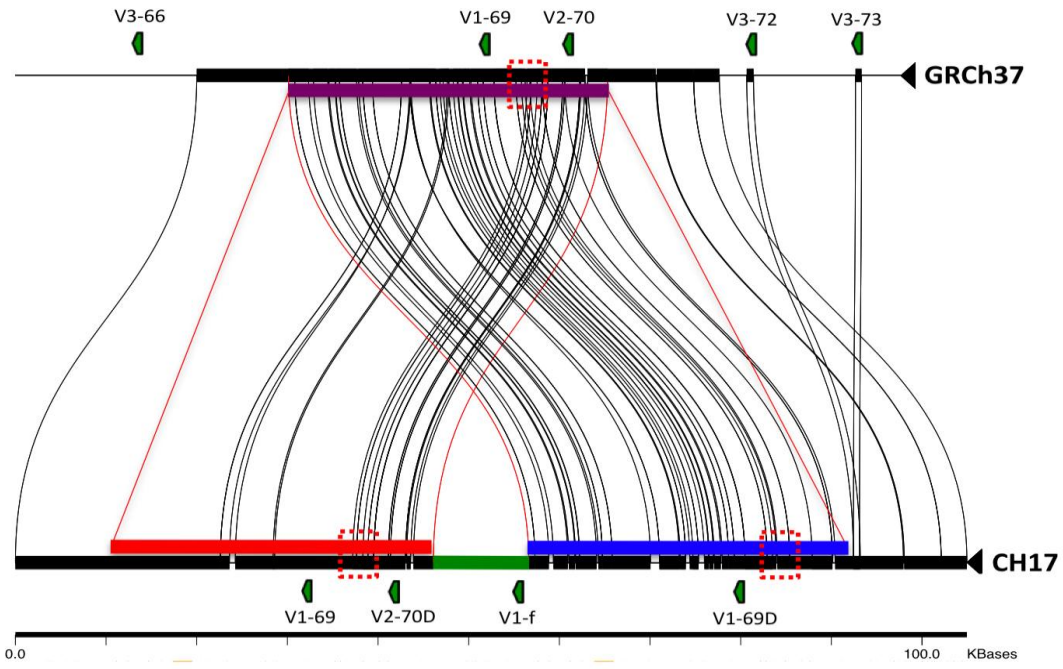
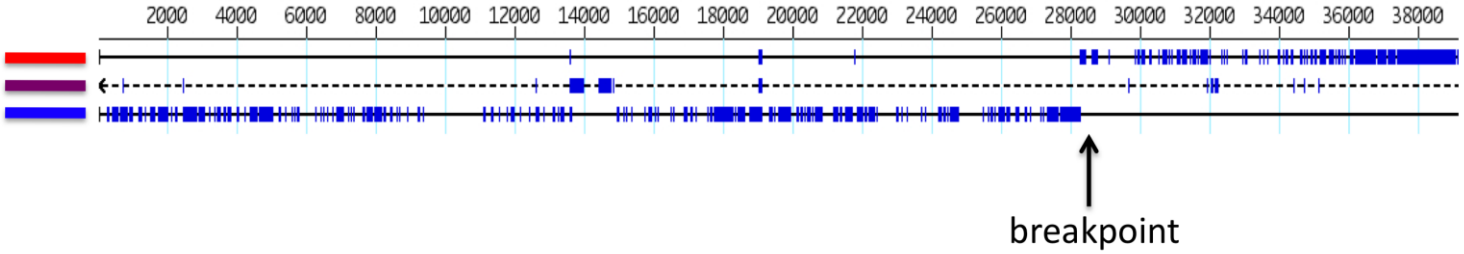


Figure S6. Breakpoint analysis of the *IGHV5-a* and *IGHV3-64D* complex event. A 65 kbp segment from the CH17 haplotype compared to GRCh37 (chr14:106500400-106605932) depicting the complex event that results in the "swap" of ~38 kbp of sequence between the two haplotypes, including either *IGHV1-8* and *IGHV3-9* or *IGHV5-a* and *IGHV3-64D*. Positions of annotated IGHV genes in GRCh37 and CH17 are depicted as green chevrons. The black bars seen on both haplotypes connected by thin black lines represent regions of shared sequence. The breakpoints of the event are indicated by thin red lines. The origins of the haplotype-specific sequences in each of these haplotypes are not known.

A


V3-66
V1-69
V2-70
V3-72
V3-73
GRCh37
CH17
V1-69
V2-70D
V1-f
V1-69D
0.0
100.0
KBases

B


2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
breakpoint

C

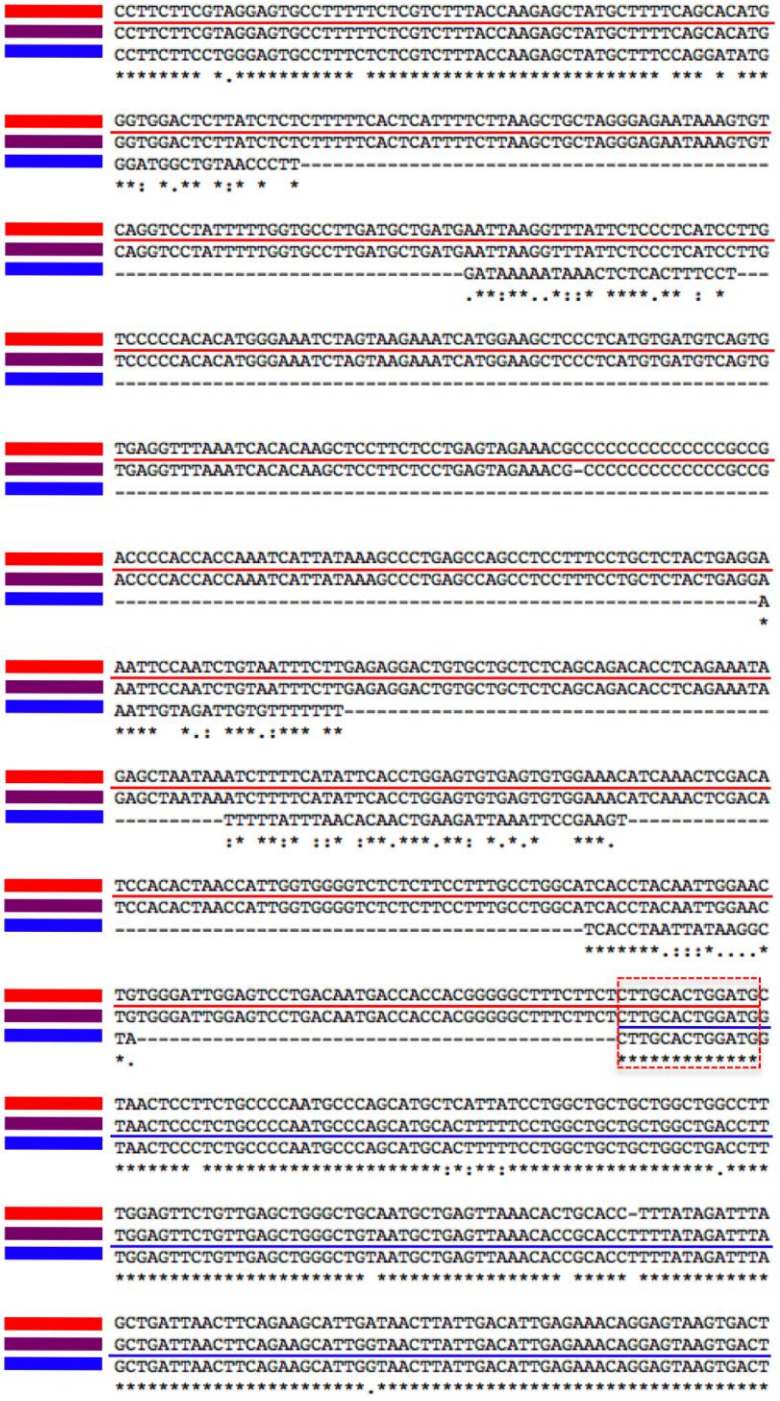
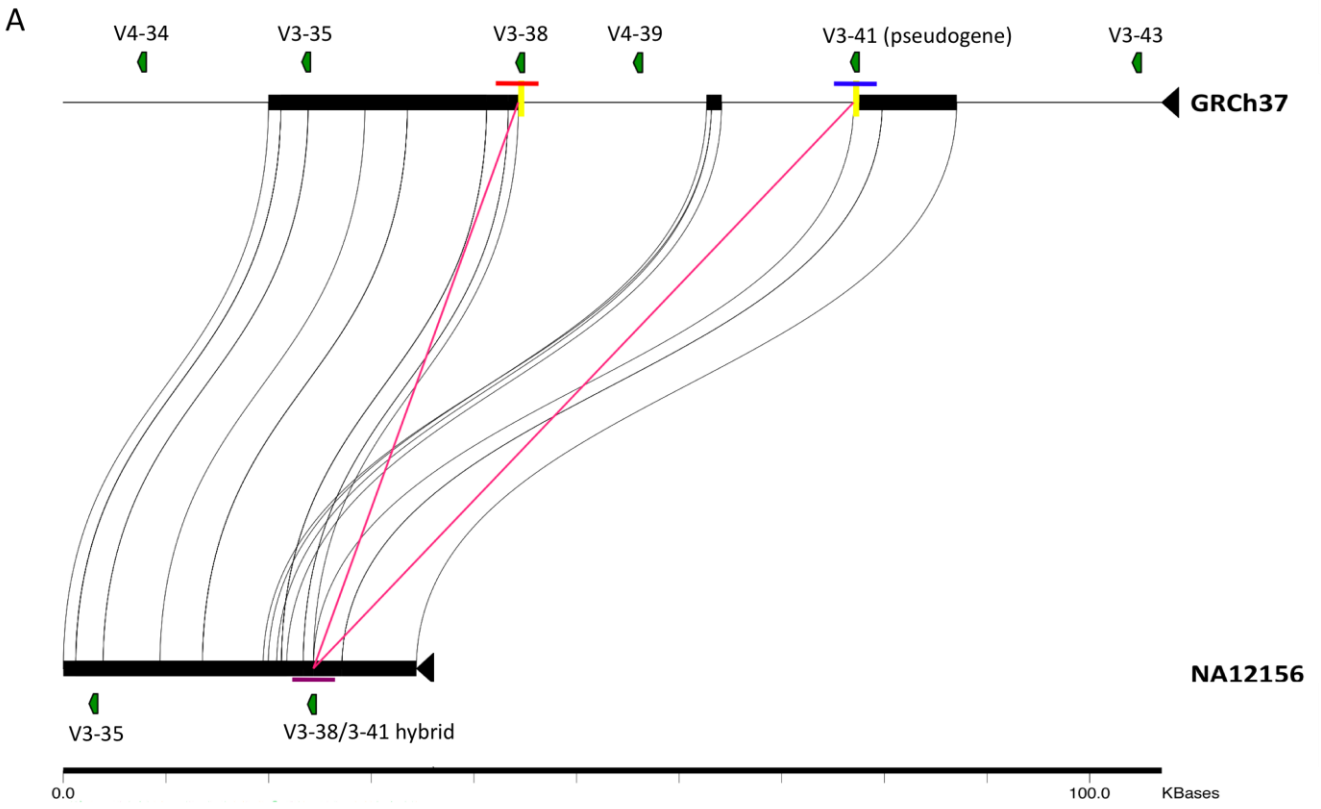


Figure S7. **Breakpoint analysis of *IGHV2-70D, IGHV3-h, IGHV1-f,* and IGHV1-*69D* insertion.** (A) Sequence comparison between ~100 kbp segments from CH17 and GRCh37 reference sequence within the region spanning *IGHV1-69* and *IGHV2-70* (chr14:107121936-107217468). The positions of IGHV genes in GRCh37 are shown in the top panel (red boxes), with gene name labels indicated on the right. The black bars seen on both haplotypes connected by thin black lines represent regions of shared sequence. The purple, red, and blue bars connected by thin red lines represent a ~38 kbp segment that is shared between the two haplotypes, the structure of which is indicative of a segmental duplication, as it is observed twice in the CH17 haplotype (red and blue bars) and once in GRCh37 (purple bar); these segments include copies of *IGHV1-69*, *IGHV2-70*, and either the pseudogene *IGHV3-71* (not depicted) or the pseudogene *IGHV3-h* (not depicted). The segmental duplications in CH17 are split by a region of sequence that is novel to CH17 (green bar), which includes the previously unmapped gene *IGHV1-f*. Functional and ORF genes in CH17 are depicted by green chevrons. The sequence identity between the two segmental duplications in CH17 is 94%, thus, like in the case of the *IGHV3-23* duplication, this segmental duplicate most likely served as substrate for a deletion to occur in the CH17 haplotype, giving rise to the haplotype observed in GRCh37 with a loss of *IGHV1-69D* and *IGHV2-70D*, as well as *IGHV1-f* and *IGHV3-h*. The predicted breakpoints in each haplotype are indicated by red boxes. (B) A three-way alignment of the ~38 kbp segmental duplicates from both haplotypes is shown, based on purple, red, and blue bars shown in (A). Alignment bp positions are shown along the top of the diagram. Each sequence is represented by a single horizontal black line. Blue tick marks on each line indicate nucleotide differences and gaps observed between the aligned sequences. The red line tracks the most similar alignment of the middle GRCh37 duplicate sequence to the other two sequences from CH17. Based on nucleotide similarity, the event breakpoint (black arrow) is presumed to have occurred within an 13 bp region in which all aligned sequences share 100% sequence identity, which is depicted at nucleotide resolution in (C). (C) In this image, the red and blue lines depict the best pairwise sequence alignments between the three sequences, as depicted by red line in (B). The red box indicates 13 bp region where all three sequences align perfectly at predicted breakpoint.


A
V4-34
V3-35
V3-38
V4-39
V3-41 (pseudogene)
V3-43
GRCh37
NA12156
V3-35
V3-38/3-41 hybrid
0.0
100.0
KBases

B

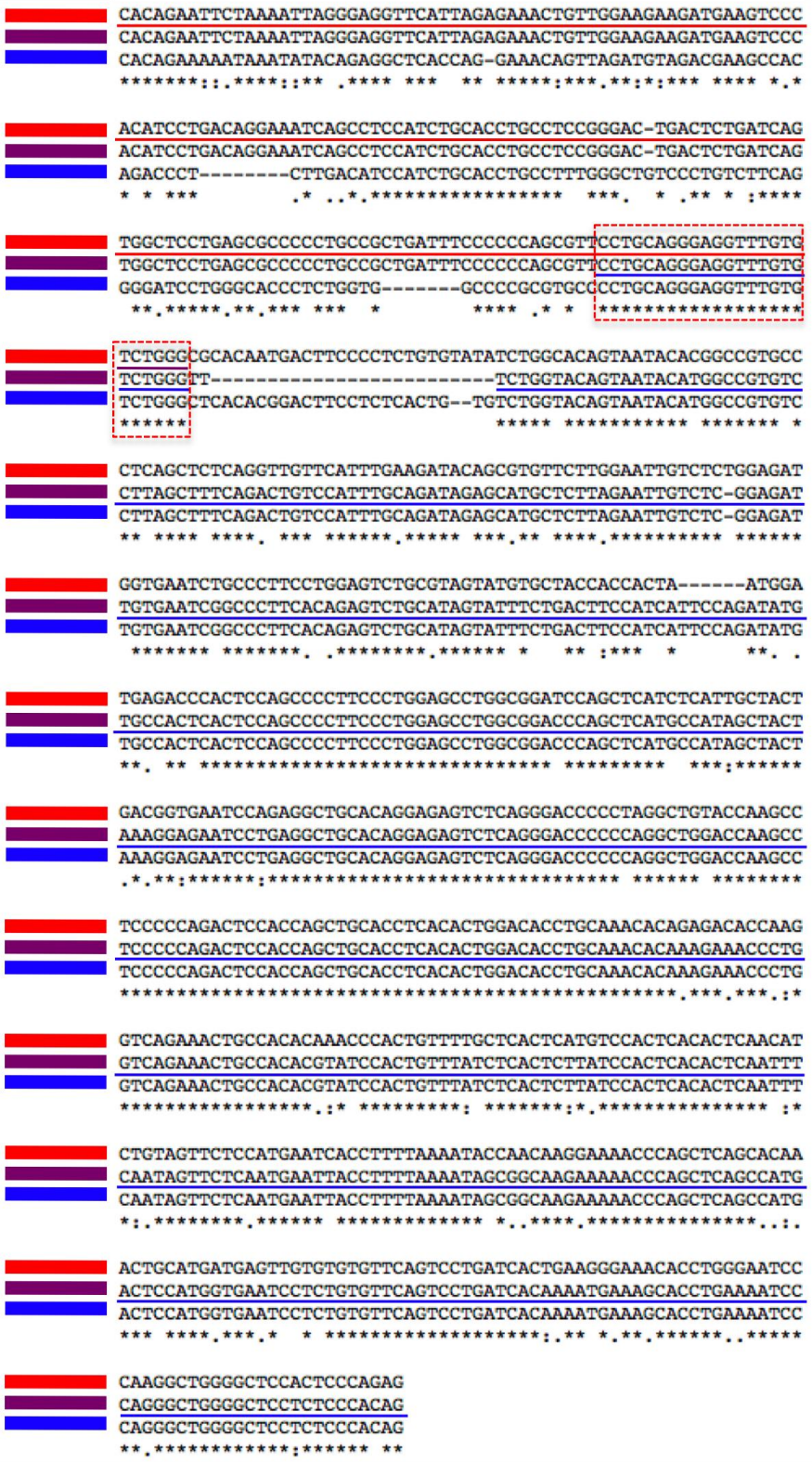
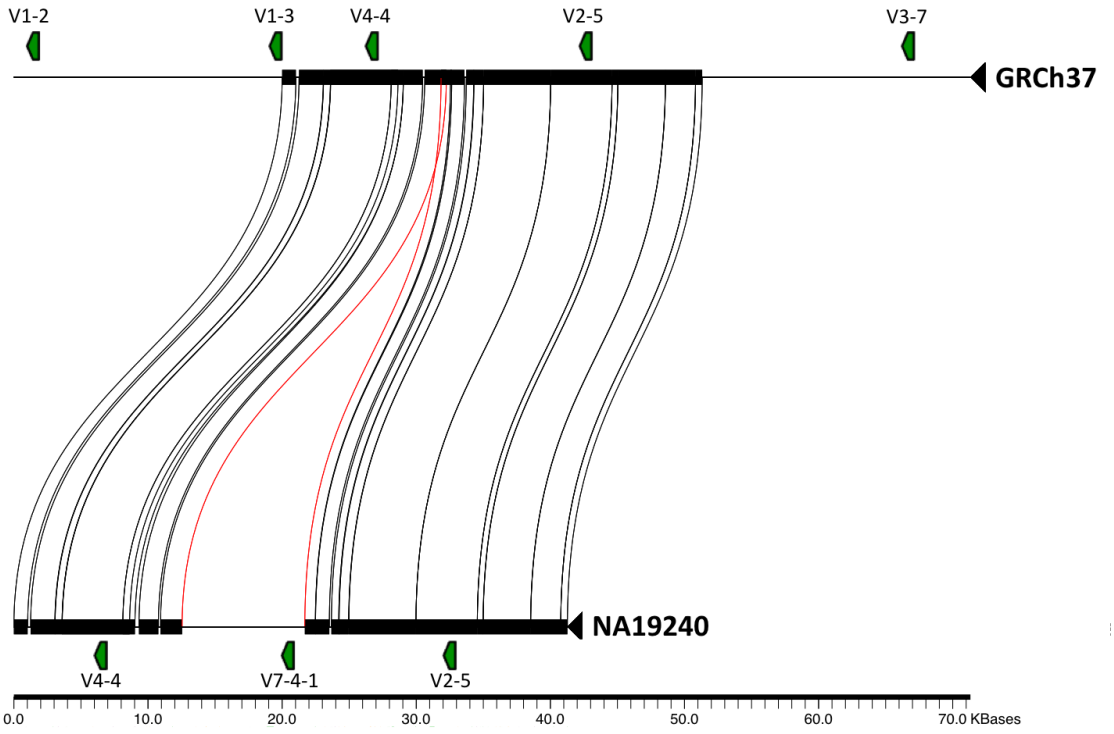

CACAGAATTCTAAAATTAGGGAGGTTCATTAGAGAAACTGTTGGAAGAAGATGAAGTCCC
CACAGAATTCTAAAATTAGGGAGGTTCATTAGAGAAACTGTTGGAAGAAGATGAAGTCCC
CACAGAAAAATAAATATACAGAGGCTCACCAG-GAAACAGTTAGATGTAGACGAAGCCAC
ACATCCTGACAGGAAATCAGCCTCCATCTGCACCTGCCTCCGGGAC-TGACTCTGATCAG
ACATCCTGACAGGAAATCAGCCTCCATCTGCACCTGCCTCCGGGAC-TGACTCTGATCAG
AGACCCT--------CTTGACATCCATCTGCACCTGCCTTTGGGCTGTCCCTGTCTTCAG
TGGCTCCTGAGCGCCCCCTGCCGCTGATTTCCCCCCAGCGTTCCTGCAGGGAGGTTTGTG
TGGCTCCTGAGCGCCCCCTGCCGCTGATTTCCCCCCAGCGTTCCTGCAGGGAGGTTTGTG
GGGATCCTGGGCACCCTCTGGTG-------GCCCCGCGTGCCCTGCAGGGAGGTTTGTG
TCTGGGCGCACAATGACTTCCCCTCTGTGTATATCTGGCACAGTAATACACGGCCGTGCC
TCTGGGTT-----------------------TCTGGTACAGTAATACATGGCCGTGTC
TCTGGGCTCACACGGACTTCCTCTCACTG--TGTCTGGTACAGTAATACATGGCCGTGTC
CTCAGCTCTCAGGTTGTTCATTTGAAGATACAGCGTGTTCTTGGAATTGTCTCTGGAGAT
CTTAGCTTTCAGACTGTCCATTTGCAGATAGAGCATGCTCTTAGAATTGTCTC-GGAGAT
CTTAGCTTTCAGACTGTCCATTTGCAGATAGAGCATGCTCTTAGAATTGTCTC-GGAGAT
GGTGAATCTGCCCTTCCTGGAGTCTGCGTAGTATGTGCTACCACCACTA------ATGGA
TGTGAATCGGCCCTTCACAGAGTCTGCATAGTATTTCTGACTTCCATCATTCCAGATATG
TGTGAATCGGCCCTTCACAGAGTCTGCATAGTATTTCTGACTTCCATCATTCCAGATATG
TGAGACCCACTCCAGCCCCTTCCCTGGAGCCTGGCGGATCCAGCTCATCTCATTGCTACT
TGCCACTCACTCCAGCCCCTTCCCTGGAGCCTGGCGGACCCAGCTCATGCCATAGCTACT
TGCCACTCACTCCAGCCCCTTCCCTGGAGCCTGGCGGACCCAGCTCATGCCATAGCTACT
GACGGTGAATCCAGAGGCTGCACAGGAGAGTCTCAGGGACCCCCTAGGCTGTACCAAGCC
AAAGGAGAATCCTGAGGCTGCACAGGAGAGTCTCAGGGACCCCCCAGGCTGGACCAAGCC
AAAGGAGAATCCTGAGGCTGCACAGGAGAGTCTCAGGGACCCCCCAGGCTGGACCAAGCC
TCCCCCAGACTCCACCAGCTGCACCTCACACTGGACACCTGCAAACACAGAGACACCAAG
TCCCCCAGACTCCACCAGCTGCACCTCACACTGGACACCTGCAAACACAAAGAAACCCTG
TCCCCCAGACTCCACCAGCTGCACCTCACACTGGACACCTGCAAACACAAAGAAACCCTG
GTCAGAAACTGCCACACAAACCCACTGTTTTGCTCACTCATGTCCACTCACACTCAACAT
GTCAGAAACTGCCACACGTATCCACTGTTTATCTCACTCTTATCCACTCACACTCAATTT
GTCAGAAACTGCCACACGTATCCACTGTTTATCTCACTCTTATCCACTCACACTCAATTT
CTGTAGTTCTCCATGAATCACCTTTTAAAATACCAACAAGGAAAACCCAGCTCAGCACAA
CAATAGTTCTCAATGAATTACCTTTTAAAATAGCGGCAAGAAAAACCCAGCTCAGCCATG
CAATAGTTCTCAATGAATTACCTTTTAAAATAGCGGCAAGAAAAACCCAGCTCAGCCATG
ACTGCATGATGAGTTGTGTGTGTTCAGTCCTGATCACTGAAGGGAAACACCTGGGAATCC
ACTCCATGGTGAATCCTCTGTGTTCAGTCCTGATCACAAAATGAAAGCACCTGAAAATCC
ACTCCATGGTGAATCCTCTGTGTTCAGTCCTGATCACAAAATGAAAGCACCTGAAAATCC
CAAGGCTGGGGCTCCACTCCCAGAG
CAGGGCTGGGGCTCCTCTCCCACAG
CAGGGCTGGGGCTCCTCTCCCACAG

Figure S8. **Breakpoint analysis of *IGHV3-38* and *IGHV4-39* deletion.** (A) Sequence comparison between fosmid clone AC244497 from individual NA12156 to GRCh37 in the region spanning *IGHV3-34* to *IGHV3-43* depicting a deletion of *IGHV4-39* and *IGHV3-38*, the breakpoints of which are different from those predicted previously (Chimge et al., 2005). The positions of IGHV genes in GRCh37 and the NA12156 haplotype are depicted as green chevrons with gene name labels indicated above. The black bars seen on both haplotypes connected by thin black lines represent regions of shared sequence, with the thin red connecting lines indicating the positions of the predicted breakpoints from the comparison of the two haplotypes. The shorter horizontal red, blue, and purple lines represent the regions of the predicted breakpoints in the fosmid (purple) and reference genome (red and blue). (B) A three-way sequence alignment based on sequences from each of these breakpoints, indicated by three colored lines (red, blue, and purple) corresponding to those shown in (A). In this image the red and blue lines depict the best pairwise sequence alignments between the three sequences. The red box indicates the 24 bp region where all three sequences align perfectly at the predicted breakpoint. The gene *IGHV3-38* and pseudogene *IGHV3-41* are included in the regions of extended homology (*i.e.*, within the shown aligned sequences).

A


V1-2
V1-3
V4-4
V2-5
V3-7
GRCh37
NA19240
V4-4
V7-4-1
V2-5
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0 KBases

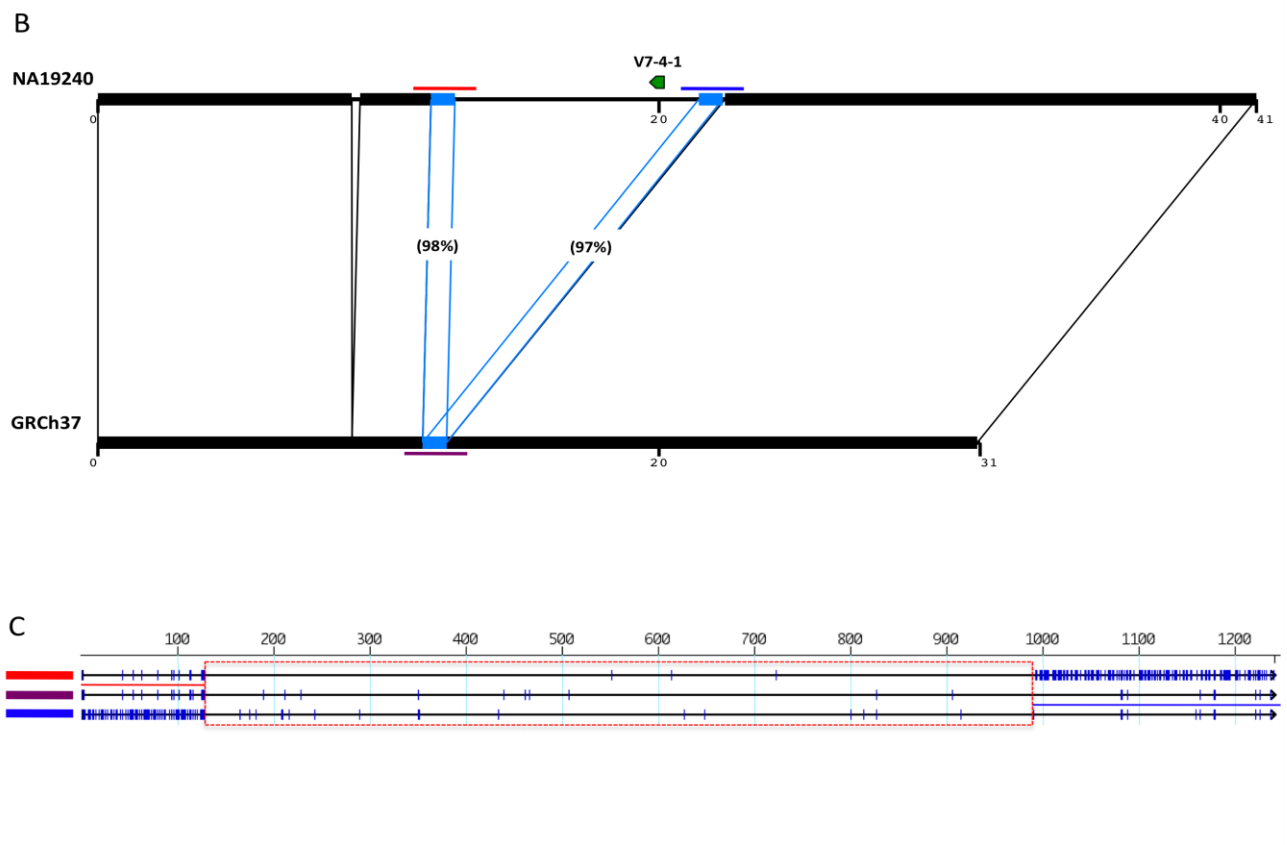
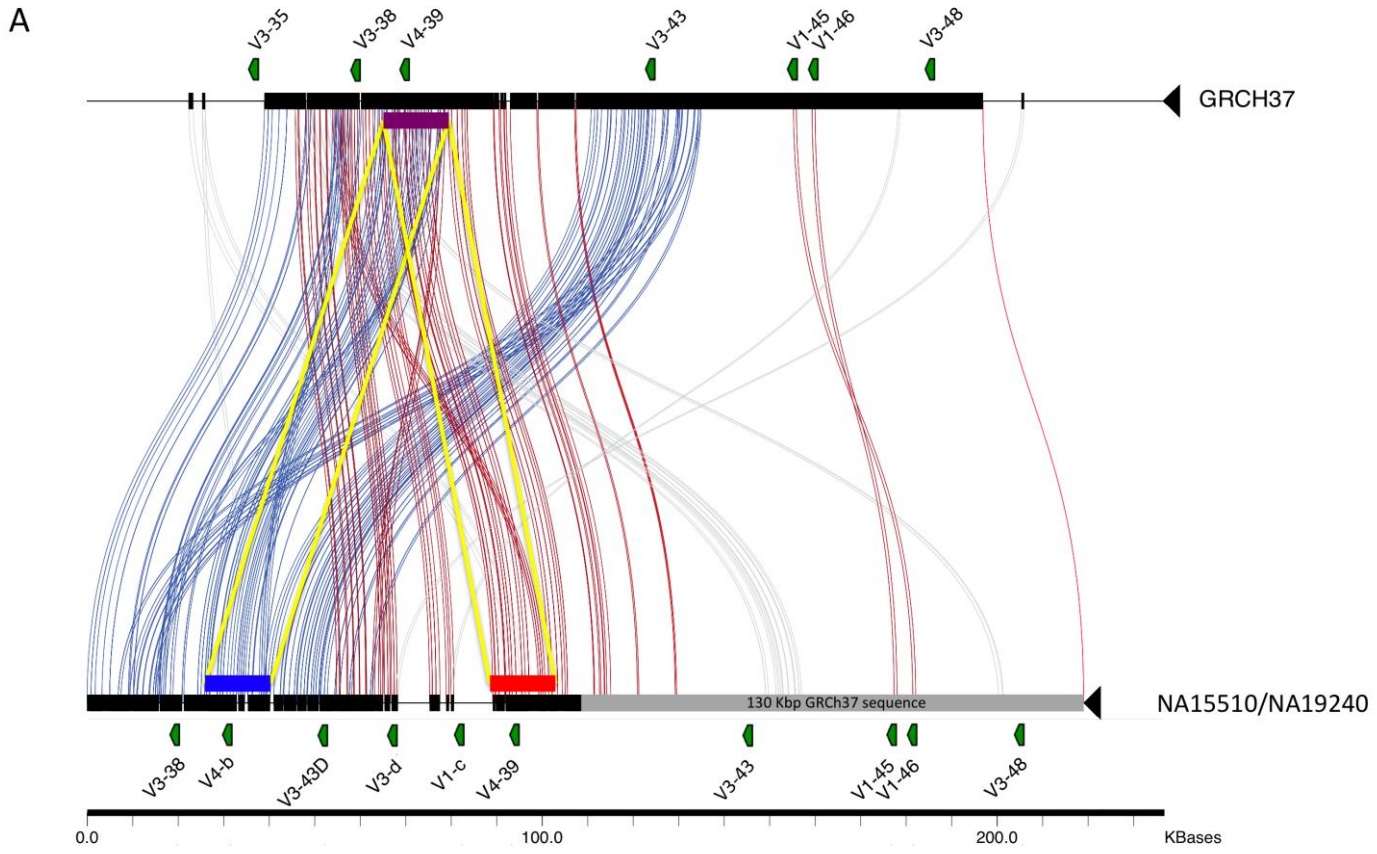


Figure S9. **Breakpoint analysis of the *IGHV7-4-1* insertion.** (A) Sequence comparison between fosmid clone AC241513 (individual NA19240) and GRCh37 in the region spanning *IGHV1-2* to *IGHV4-4*. IGHV genes in GRCh37 and the NA19240 haplotype are depicted as green chevrons, with gene names indicated above each gene. The black bars seen on both haplotypes connected by thin black lines represent regions of shared sequence, with the thin red connecting lines indicating the positions of the predicted breakpoints from the comparison of the two haplotypes. (B) Image showing the same alignment as in (A). Again, the black bars on both haplotypes connected by thin black lines represent regions of shared sequence; however, also shown is a ~860 bp repeat segment identified by BLAST, which occurs at the predicted breakpoint. This segment is repeated twice in the fosmid haplotype, but only once in the reference assembly. The sequence identities between each segment found in the fosmid haplotype and that in the reference genome are indicated in parentheses. The horizontal red, blue, and purple lines spanning the regions of the predicted breakpoints in the fosmid (red and blue) and reference genome (purple) are shown and represent the positions of the sequences used in the alignment shown in C. (C) A three-way alignment of sequences spanning the breakpoints in the fosmid and reference assembly, indicated by red, blue, and purple lines, corresponding to those shown in B. In the image, each sequence is represented by a single horizontal black line, and blue tick marks on each line indicate nucleotide differences and gaps observed between the aligned sequences. The red and blue lines between the sequences in the image depict the best pairwise sequence alignments between the three sequences. The red box indicates the ~860 bp region (shown in B) where all three sequences align with high similarity. Although these data indicate that the event breakpoint is predicted to occur within this 860 bp region, sequence similarity data do not allow for the exact breakpoints to be identified. Similar results were observed for the *IGHV7-4-1* insertions identified from CH17 and NA12878.


A
V3-35
V3-38
V4-39
V3-43
V1-45
V1-46
V3-48
GRCH37
130 Kbp GRCh37 sequence
NA15510/NA19240
V3-38
V4-b
V3-43D
V3-d
V1-c
V4-39
V3-43
V1-45
V1-46
V3-48
0.0
100.0
200.0
KBases

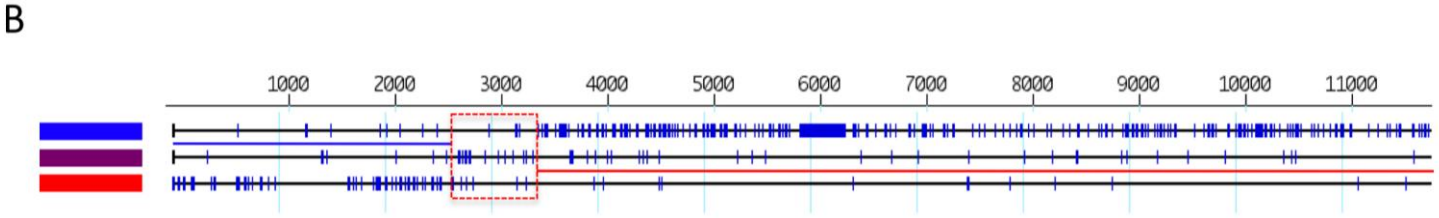
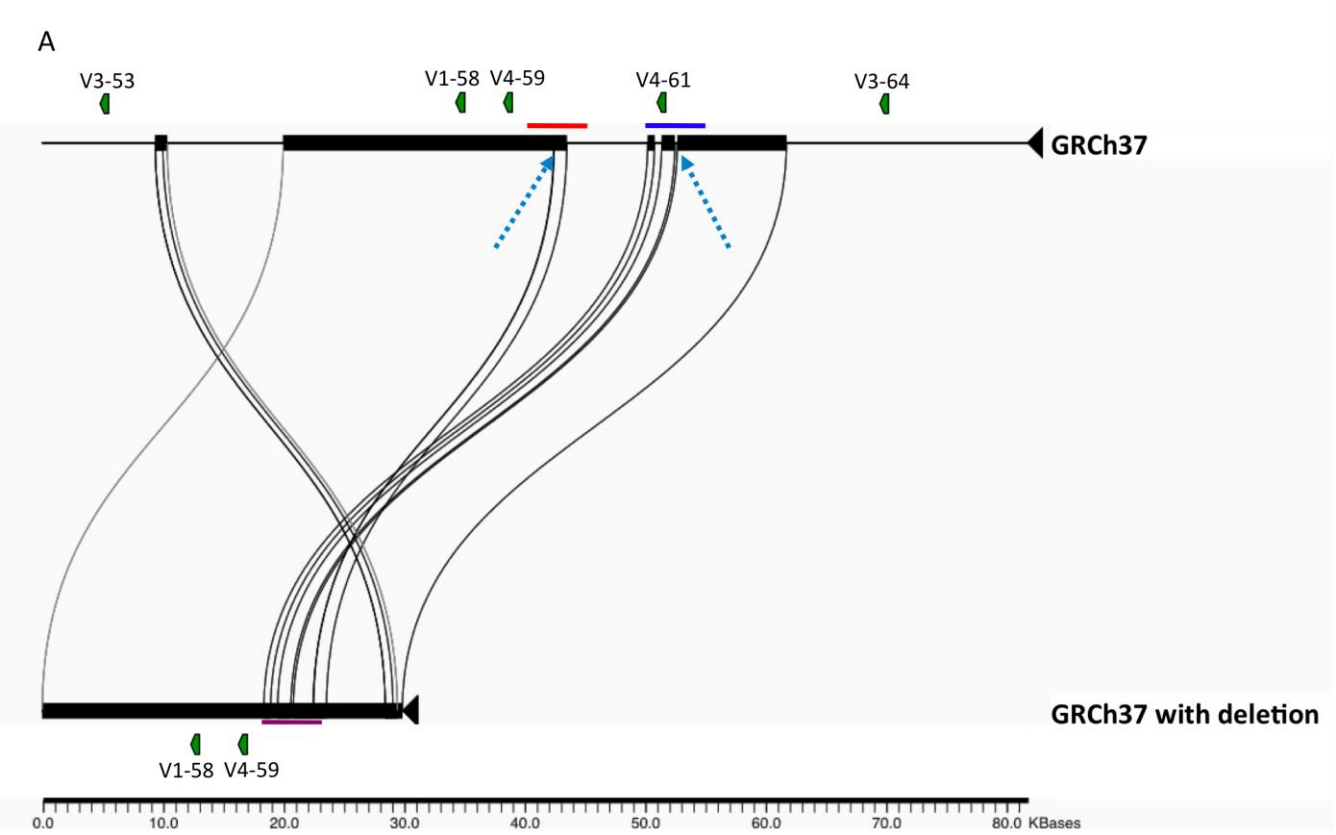

B
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000

Figure S10. **Breakpoint analysis of the *IGHV4-b, IGHV3-43D, IGHV3-d* and *IGHV1-c* insertion.** (A) Image of sequence comparison between GRCh37 in the region spanning *IGHV3-35* to *IGHV3-48* to a haplotype characterized from fosmid clones of NA15510/NA19240, depicting a complex insertion of the genes *IGHV4-b*, *IGHV3-43D*, *IGHV3-d*, and *IGHV1-c*. For the construction of this image, approximately 130 Kbp of sequence from GRCh37 (gray bar) was added to the NA15510/NA19240 haplotype, so that the duplication structure of the region could be better visualized, in particular, the duplication blocks containing *IGHV3-43* and *IGHV3-43D*. The positions of IGHV genes in GRCh37 and the haplotype constructed from fosmid sequences are shown as green chevrons. The black bars shown on both haplotypes connected by thin blue, red, and black lines represent regions of shared sequence and reflect the complex duplication structure in the region. As shown, the NA15510/NA19240 haplotype insertion contains duplicated sequence of *IGHV4-39*, *IGHV3-38*, *IGHV3-43*, and flanking sequences and provides evidence that *IGHV4-b* and *IGHV3-d* are close paralogs of *IGHV4-39* and *IGHV3-38*, and *IGHV3-43D* and flanking sequences are duplicated from the *IGHV3-43* gene region. Red, blue, and purple blocks, connected by yellow lines, depict 11.7 kbp segmental duplications (identified by BLAST) found twice in the fosmid haplotype and once in the reference. (B) A three-way alignment of these segmental duplications indicates they are likely to have mediated this event, although the exact breakpoints cannot be identified. In the image (B), each sequence is represented by a single horizontal black line, and blue tick marks on each line indicate nucleotide differences and gaps observed between the aligned sequences. The red and blue lines between the sequences depict the best pairwise sequence alignments between the three sequences. The red box indicates a ~800 bp region within which the reference duplicate shows low sequence similarity with both duplicates of the fosmid haplotype, inhibiting the identification of event breakpoints within this sequence; however, the patterns of sequence similarity between the reference and fosmid segmental duplicates on either side of the red box suggest that the event crossover was likely to have occurred within this segment. Nucleotide differences seen between the three sequences in this region may reflect the accumulation of mutations following the formation of the variant, implying that this event could be older than the other insertion/deletions examined in this study. However, further sequencing will be required to address this hypothesis.


A
V3-53
V1-58
V4-59
V4-61
V3-64
GRCh37
GRCh37 with deletion
V1-58
V4-59
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0 KBases

B

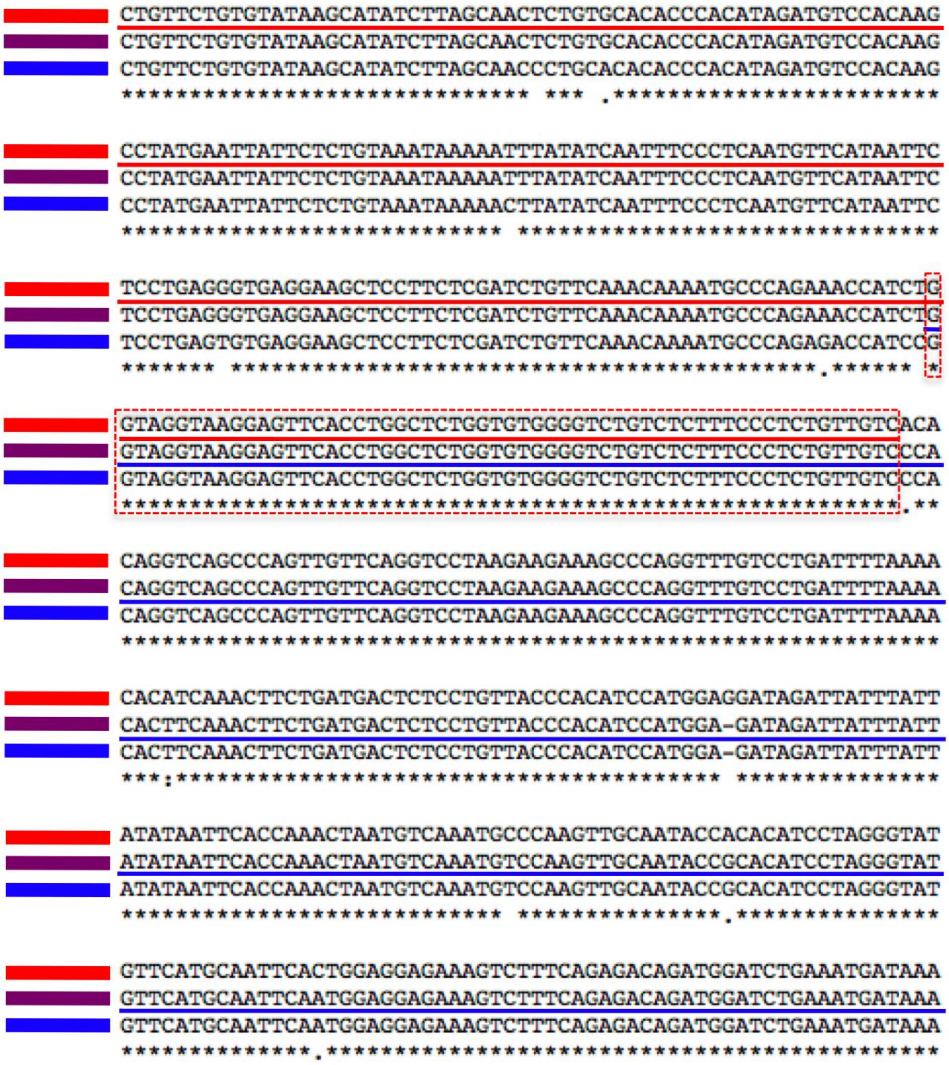


Figure S11. **Breakpoint analysis of the *IGHV4-61* deletion.** (A) Image depicting the *IGHV4-61* deletion identified by Mills *et al.* (2011)(chr14:107084861-107096738). The sequence comparison shows GRCh37 in the region spanning *IGHV3-53* to *IGHV3-64* to a "dummy" haplotype using GRCh37 sequence with the *IGHV4-61* sequence masked out, based on previously identified breakpoints (Mills et al., 2011). The positions of IGHV genes in GRCh37 are shown in the top panel (red boxes), with gene name labels indicated on the right. The black bars shown on both haplotypes connected by thin blue and black lines represent regions of shared sequence. Blue dotted arrows indicate the breakpoints of the event in the modified GRCh37 haplotype. Red, blue, and purple horizontal lines represent sequences spanning breakpoints (blue dotted arrows) in each haplotype aligned in B. (B) A three-way alignment spanning the breakpoints of this event. The red and blue lines depict the best pairwise sequence alignments between the three sequences, and the red box indicates the 58 bp region where all three sequences align perfectly at the predicted breakpoint. In the image (B), each sequence is represented by a single horizontal black line, and blue tick marks on each line indicate nucleotide differences and gaps observed between the aligned sequences. Two 5.6 kbp duplication segments (0.95% identity) containing the genes *IGHV4-59* and *IGHV4-61* and the pseudogenes *IGHV3-60* and *IGHV3-62* were identified at the breakpoints using BLAST (not depicted here).

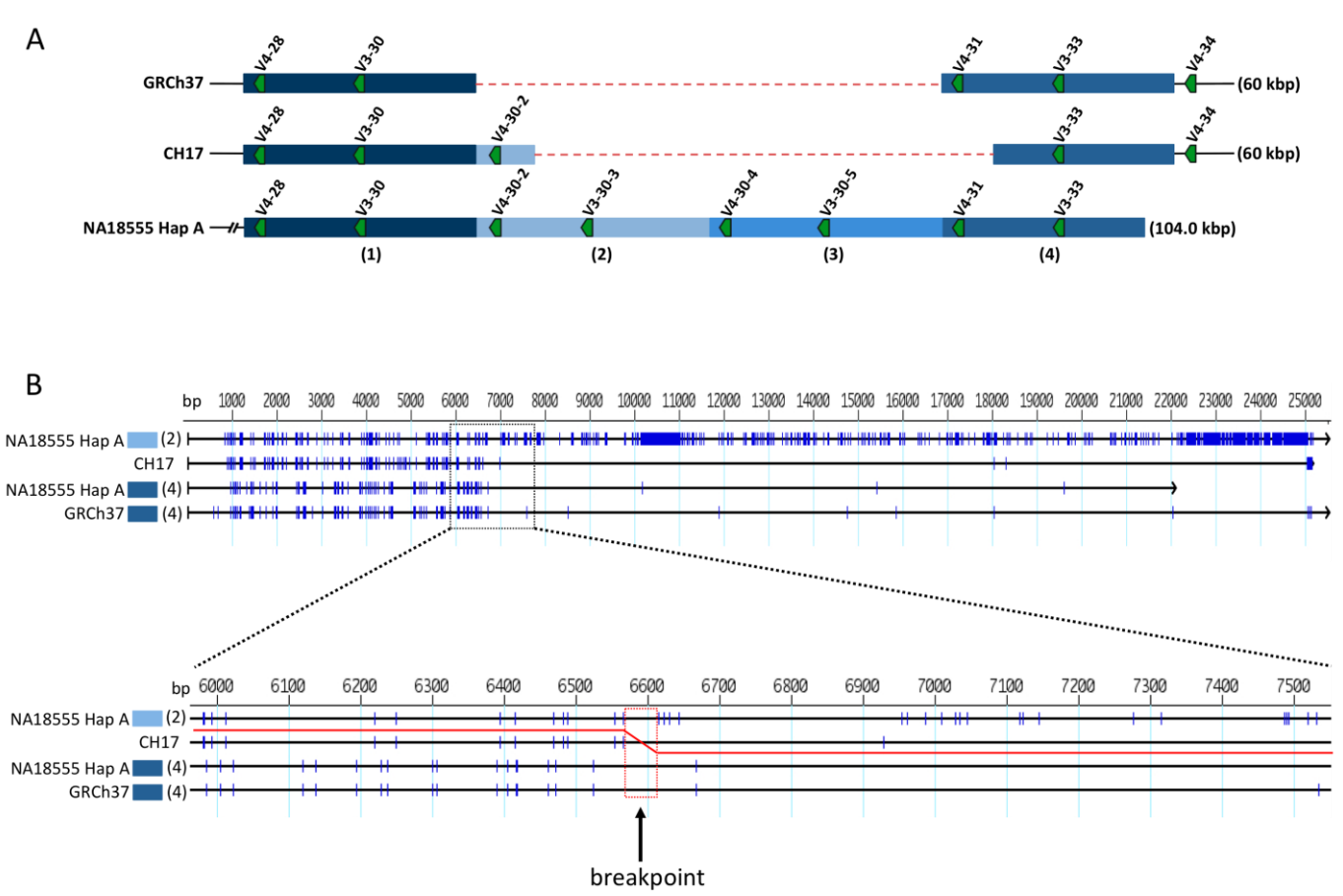


Figure S12. Breakpoint analysis of the CH17 Haplotype complex event.
(A) Depicts GRCh37, CH17, and NA18555 Haplotype A (also see Figure 4, main text). The haplotypes shown were determined using the method described in the Materials and Methods (main text) and Figure S1. Red dotted lines represent deletions compared to NA18555 Haplotype A. (B) This image shows an expanded and zoomed-in version of a multi-sequence alignment containing sequences of the ~25 kbp segmental duplicates (shown in panel A) from GRCh37, CH17, and NA18555 Haplotype A. Alignment bp positions are shown along the top of the diagram. Each sequence is represented by a single horizontal black line, labelled with a haplotype name and the duplication block color and number corresponding to those shown in panel A. Blue tick marks on each line indicate single nucleotide differences and gaps between a given sequence and all other sequences in the alignment. The red line tracks the most similar alignments of the middle CH17 sequence to the other three sequences from GRCh37 and NA18555 Haplotype A. Based on changes in nucleotide similarity on either side of the red box (bottom panel), the event breakpoint is presumed to have occurred within a 47 bp region in which all aligned sequences share 100% sequence identity. Importantly, if this event had simply been compared to GRCh37, it is likely that this variant would have been mischaracterized.

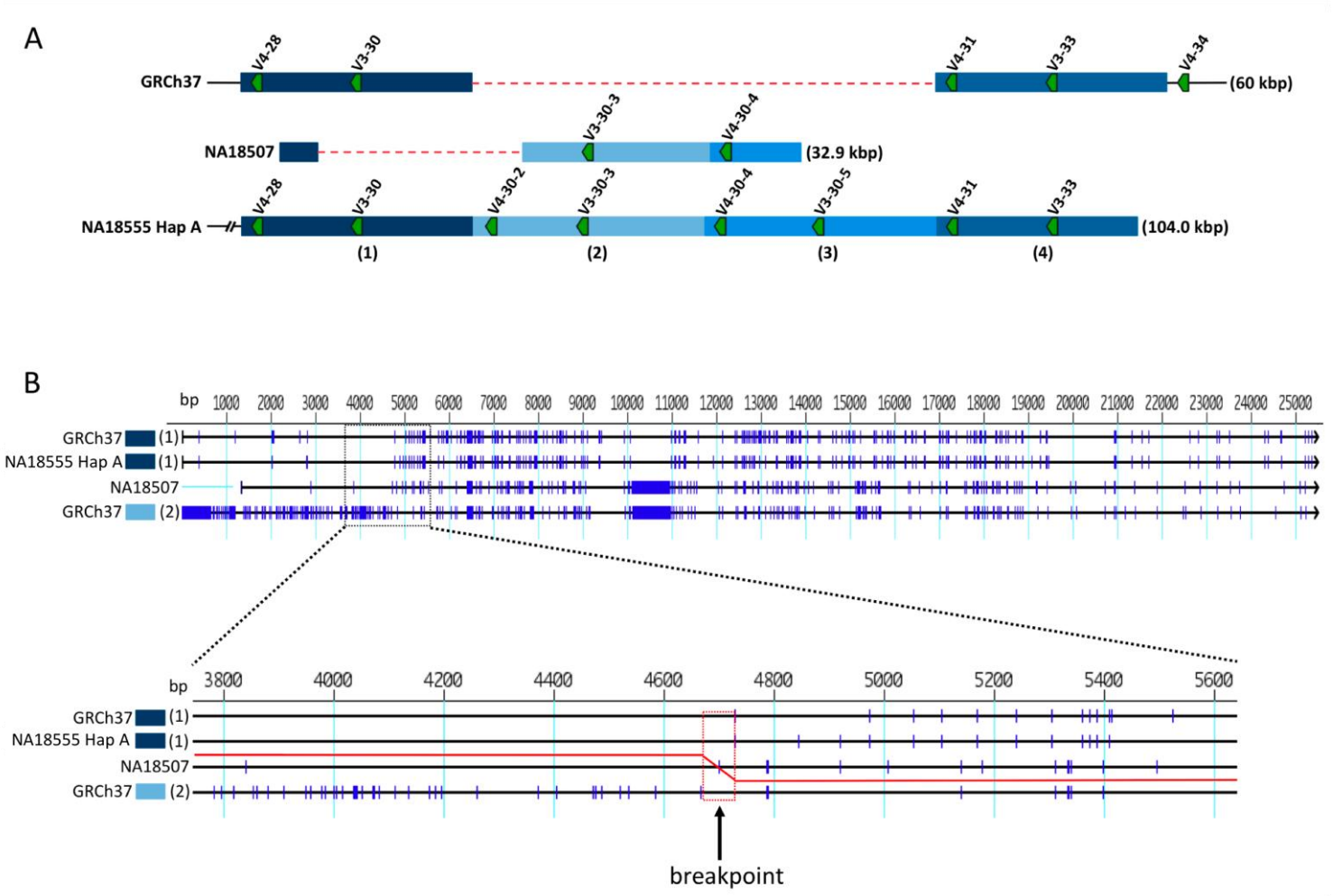


Figure S13. Breakpoint analysis of the NA18507 Haplotype complex event.
Depicts GRCh37, NA18507, and NA18555 Haplotype A (also see Figure 4, main text). The haplotypes shown were determined using the method outlined in the Materials and Methods (main text) and Figure S1. (B) This image shows an expanded and zoomed-in version of a multi-sequence alignment containing sequences of the ~25 kbp segmental duplicates (shown in panel A) from GRCh37, NA18507, and NA18555 Haplotype A. Alignment bp positions are shown along the top of the diagram. Each sequence is represented by a single horizontal black line, labelled with haplotype name and the duplication block color and number corresponding to those shown in panel A. Blue tick marks on each line indicate single nucleotide differences and gaps between a given sequence and all other sequences in the alignment. The red line tracks the most similar alignments of the middle NA18507 sequence to the other three sequences from GRCh37 and NA18555 Haplotype A. Based on changes in nucleotide similarity on either side of the red box (bottom panel), the event breakpoint is presumed to have occurred within a 61 bp region in which all aligned sequences share 98% sequence identity.

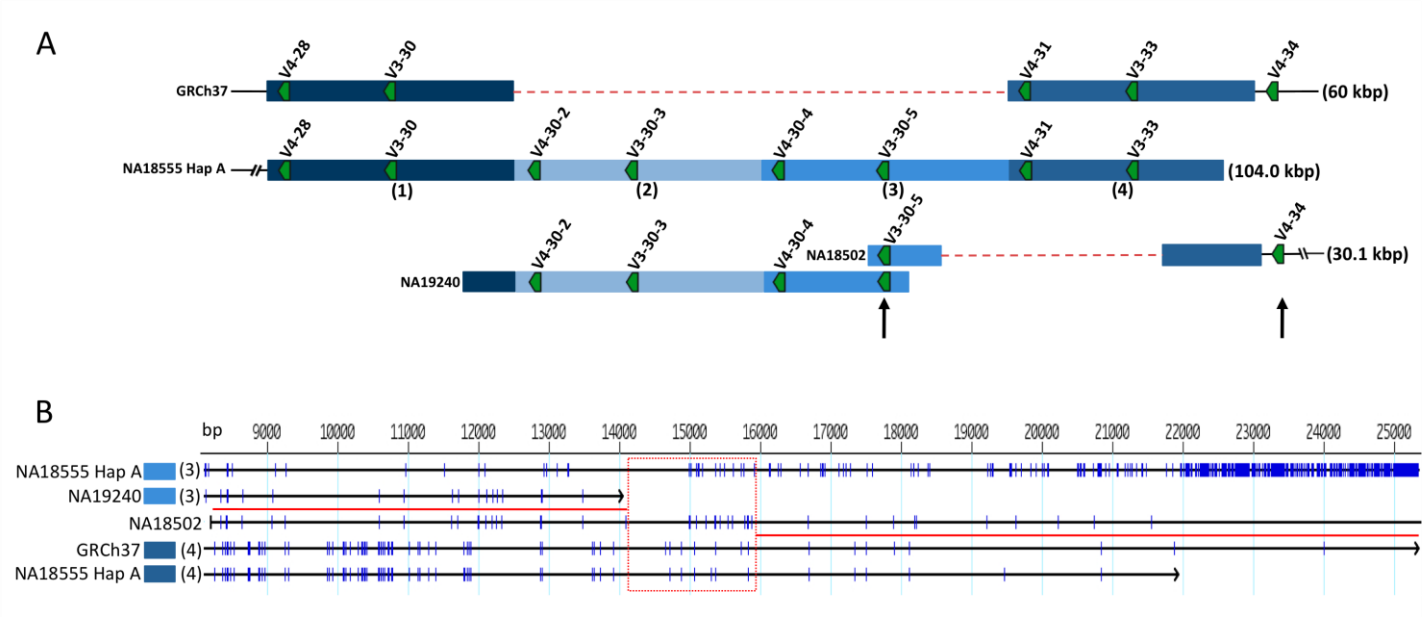


Figure S14. Breakpoint analysis of NA18502 Haplotype complex event.
(A) Depicts a GRCh37, NA18555 Haplotype A, NA18502, and NA19420 clone AC234301(also see Figure 4, main text). The haplotypes shown were determined using the method described in the Materials and Methods (main text) and Figure S1 above. Black arrows indicate gene synteny between the NA18502 haplotype and other haplotypes in the comparison. (B) This image shows a multi-sequence alignment containing sequences from a portion of the ~25 kbp segmental duplicates from GRCh37, NA18555 Haplotype A, and NA19420 clone AC234301 that overlap the NA18502 haplotype (see panel A). Alignment bp positions are shown along the top of the diagram. Each sequence is represented by a single horizontal black line, labelled with a haplotype name and the duplication block color and number corresponding to those shown in panel A. Blue tick marks on each line indicate single nucleotide differences and gaps between a given sequence and all other sequences in the alignment. The red line tracks the most similar alignments of the middle NA18502 haplotype sequence to the other four sequences in the alignment. The NA19420 clone AC234301, although only a partial haplotype, was used to inform the placement of NA18502 based on their sequence similarity in segmental duplication block 3. Both haplotypes are from individuals of the Yoruban population, and thus shared SNPs that differentiate these haplotypes from duplication block 3 in NA18555 Haplotype A are likely an example of population-specific nucleotide variation. Such variation inhibits the delineation of breakpoints for this event; nonetheless, the breakpoint is predicted to have occurred within the red box, based on changes in nucleotide similarity on either side of this box. Additional haplotype resequencing in this region from African populations would likely improve resolution with respect to tracking the crossover events associated with this variant. Again, as noted in Figure S12 for CH17, it is likely that the NA18502 would have been mischaracterized if only analyzed in the context of GRCh37.

Clone Name	Accession	Sample ID	Library	Pop	Chr	Start	End	Variants
ABC7_000043068300_J5	AC244483	NA18517	ABC7	YRI	chr14	106298007	107175503	somatic rearrangement
ABC7_000042374900_I23	AC244478	NA18517	ABC7	YRI	chr14	106611603	106842070	mismapped end sequences
ABC8_000041787500_L13	AC244410	NA18507	ABC8	YRI	chr14	106183100	106505148	somatic rearrangement
ABC8_000043241700_M22	AC244412	NA18507	ABC8	YRI	chr14	106196182	106507379	somatic rearrangement (not annotated)
ABC8_000040989700_P12	AC244459	NA18507	ABC8	YRI	chr14	106199798	106683306	somatic rearrangement
168825_ABC8_4_1_000000787522_D4	AC244496	NA18507	ABC8	YRI	chr14	106304430	107025286	somatic rearrangement
ABC8_000000718240_A7	AC244411	NA18507	ABC8	YRI	chr14	106780964	106789660	V4-28 to V4-34 region complex indel
ABC8_000005713749_P16	AC244460	NA18507	ABC8	YRI	chr14	107091316	107111772	no variant
170215_ABC9_3_2_000043829600_C23	AC244396	NA18956	ABC9	JPT	chr14	106174436	106327048	somatic rearrangement (not annotated)
170215_ABC9_3_2_000041260500_H2	AC244491	NA18956	ABC9	JPT	chr14	106175047	106327110	somatic rearrangement (not annotated)
170215_ABC9_3_2_000041282600_M4	AC244399	NA18956	ABC9	JPT	chr14	106178228	106809945	somatic rearrangement (not annotated)
170215_ABC9_3_2_000043834200_J7	AC244400	NA18956	ABC9	JPT	chr14	106178814	106364907	somatic rearrangement (not annotated)
170215_ABC9_3_2_000041279400_F8	AC244487	NA18956	ABC9	JPT	chr14	106193084	106380204	somatic rearrangement (not annotated)
170215_ABC9_3_2_000046007900_D17	AC244397	NA18956	ABC9	JPT	chr14	106193099	106380194	somatic rearrangement (not annotated)
172343_ABC9_3_5_000046294400_K6	AC244398	NA18956	ABC9	JPT	chr14	106203239	106837343	somatic rearrangement (not annotated)
172343_ABC9_3_5_000043993300_H10	AC206018	NA18956	ABC9	JPT	chr14	106689750	106718606	V3-23 duplication
170215_ABC9_3_2_000043849600_N9	AC244473	NA18956	ABC9	JPT	chr14	106726513	106758523	V3-23 duplication
170215_ABC9_3_2_000041239600_E19	AC244395	NA18956	ABC9	JPT	chr14	106917878	106948696	no variant
174222_ABC10_2_1_000044551900_K9	AC244480	NA19240	ABC10	YRI	chr14	106037759	106192107	somatic rearrangement (not annotated)
173650_ABC10_2_1_000044117600_K12	AC244470	NA19240	ABC10	YRI	chr14	106060871	106499156	somatic rearrangement
173650_ABC10_2_1_000044141400_O24	AC244495	NA19240	ABC10	YRI	chr14	106183931	106341954	somatic rearrangement (not annotated)
174552_ABC10_2_1_000044789200_G15	AC244482	NA19240	ABC10	YRI	chr14	106327549	106532103	somatic rearrangement
173650_ABC10_2_1_000044509600_J4	AC241513	NA19240	ABC10	YRI	chr14	106471831	106503095	V7-4-1 insertion
174552_ABC10_2_1_000044748900_N6	AC244463	NA19240	ABC10	YRI	chr14	106543383	106573224	no variant

174552_ABC10_2_1_000001587670_O14	AC244430	NA19240	ABC10	YRI	chr14	106543403	106573227	no variant
174552_ABC10_2_1_000044687300_P14	AC234301	NA19240	ABC10	YRI	chr14	106792347	106801572	V4-28 to V4-34 region complex indel
173650_ABC10_2_1_000044084700_I10	KC162926	NA19240	ABC10	YRI	chr14	106846461	106921833	V4-b, V3-43D, V3-d, V1-c insertion
173650_ABC10_2_1_000044145400_L1	AC233755	NA19240	ABC10	YRI	chr14	106881362	106937986	V4-b, V3-43D, V3-d, V1-c insertion
174552_ABC10_2_1_000001585770_E3	AC234135	NA19240	ABC10	YRI	chr14	107159176	107165232	V1-69D, V1-f, V3-h, V2-70D insertion
1201894_ABC11_2_1_000049604900_H9	AC244405	NA18555	ABC11	CHB	chr14	106148239	106887042	somatic rearrangement (not annotated)
1200277_ABC11_2_1_000048179600_D4	AC244467	NA18555	ABC11	CHB	chr14	106163452	107103691	somatic rearrangement
1200277_ABC11_2_1_000048290000_H9	AC244500	NA18555	ABC11	CHB	chr14	106164280	106900701	somatic rearrangement
1201894_ABC11_2_1_000048039500_G23	AC244481	NA18555	ABC11	CHB	chr14	106326549	106850130	somatic rearrangement
1201894_ABC11_2_1_000048019600_E23	AC244464	NA18555	ABC11	CHB	chr14	106756521	106820610	V4-28 to V4-34 region complex event
1200277_ABC11_2_1_000047355000_A14	KC162924	NA18555	ABC11	CHB	chr14	106770972	106812962	V4-28 to V4-34 region complex event
1201894_ABC11_2_1_000049598600_E10	KC162925	NA18555	ABC11	CHB	chr14	106785862	106825714	V4-28 to V4-34 region complex event
1200277_ABC11_000047150400_I4	AC231260	NA18555	ABC11	CHB	chr14	106787111	106801207	V4-28 to V4-34 region complex event
1200277_ABC11_2_1_000047354200_D2	AC244456	NA18555	ABC11	CHB	chr14	106788077	106804909	V4-28 to V4-34 region complex event
1200277_ABC11_2_1_000048009100_A5	AC244477	NA18555	ABC11	CHB	chr14	106978743	107012102	no variant
174779_ABC12_000048925600_G6	AC245090	NA12878	ABC12	YRI	chr14	106474006	106503223	V7-4-1 insertion
174779_ABC12_000046655100_E11	AC244490	NA12878	ABC12	YRI	chr14	107246863	107279721	no variant
1209408_ABC14_4_1_000050109000_A2	AC244484	NA12156	ABC14	CEU	chr14	106067185	106203209	somatic rearrangement (not annotated)
1207854_ABC14_4_1_000000945814_H10	AC244494	NA12156	ABC14	CEU	chr14	106067187	106203209	somatic rearrangement (not annotated)
1207854_ABC14_4_1_000001038114_M15	AC244486	NA12156	ABC14	CEU	chr14	106067188	106203215	somatic rearrangement (not annotated)
1207854_ABC14_4_1_000000947614_C21	AC244488	NA12156	ABC14	CEU	chr14	106196455	106897980	somatic rearrangement
1207854_ABC14_4_1_000000954514_B22	AC244492	NA12156	ABC14	CEU	chr14	106717227	106747270	V3-23 duplication
1210216_ABC14_4_1_000050416300_K24	AC244476	NA12156	ABC14	CEU	chr14	106821728	106837355	no variant
1207854_ABC14_4_1_000001051114_P8	AC244497	NA12156	ABC14	CEU	chr14	106842036	106909061	V4-39 and V3-38 deletion
1207856_ABC14_6_1_000000986905_I21	AC244468	NA12156	ABC14	CEU	chr14	106990476	107035740	no variant
H_HN-0366M09	AC245243	NA18502	ABC24	YRI	chr14	107083960	107133221	V4-28 to V4-34 region complex event
G248P82590E10	AC244393	NA15510	G248	NA	chr14	106080189	106215603	somatic rearrangement (not annotated)

G248P89855C11	AC244450	NA15510	G248	NA	chr14	106086128	106223924	somatic rearrangement (not annotated)
G248P801347C7	AC244449	NA15510	G248	NA	chr14	106219577	106398180	somatic rearrangement (not annotated)
G248P86528D1	AC241995	NA15510	G248	NA	chr14	106864253	106891653	V4-b, V3-43D, V3-d, V1-c insertion
G248P802396B6	AC234225	NA15510	G248	NA	chr14	106881362	106937986	V4-b, V3-43D, V3-d, V1-c insertion
CH17-185P21	AC246787	CH17	CH17	NA	chr14	106252705	106457005	no variant
CH17-262H11	AC244226	CH17	CH17	NA	chr14	106341680	106786536	V7-4-1 insertion
CH17-224D4	AC247036	CH17	CH17	NA	chr14	106457021	106667214	V7-4-1 insertion
CH17-76I19	AC245085	CH17	CH17	NA	chr14	106371554	106580630	V7-4-1 insertion, V1-8, V3-9/V5-a, V3-64 complex event
CH17-60O17	AC245166	CH17	CH17	NA	chr14	106665750	106874148	V4-28 to V4-34 region complex event
CH17-268I9	AC244452	CH17	CH17	NA	chr14	106851717	107090287	V4-28 to V4-34 region complex event
CH17-212P11	AC245369	CH17	CH17	NA	chr14	107041060	107186635	V1-69D, V1-f, V3-h, V2-70D insertion
CH17-314I7	AC245023	CH17	CH17	NA	chr14	107104086	107268418	V1-69D, V1-f, V3-h, V2-70D insertion
CH17-226J7	AC245094	CH17	CH17	NA	chr14	107135811	107268404	V1-69D, V1-f, V3-h, V2-70D insertion

Table S1. Names, accessions, libraries, coordinates and variants contained within each clone presented in main text.

Genbank Accession	Clone name	chr	GRCh37 start	GRCh37 end	Join alignment length	Join % id	Join details
AC246787.2	CH17-185P21	14	106252715	106456986	115128	100	
AC244226.3	CH17-262H11	14	106341680	106786536	91177	100	
AC247036.2	CH17-224D4	14	106457039	106667195	1479	100	alignment length < 2000 (join supported by additional overlapping clone AC245085.2, not shown)
AC245166.2	CH17-60O17	14	106665750	106874148	22460	100	
AC244452.3	CH17-268I9	14	106851717	107090287	49239	100	>0 gap of 500bp or more, not simple sequence (result of segmental duplication that cannot be fully resolved)
AC245369.3	CH17-212P11	14	107041060	107186635	129883	99.95	>0 gap of 50bp or more, not simple sequence (result of segmental duplication that cannot be fully resolved)
AC245023.2	CH17-314I7	14	107104086	107268418			

Table S2. Minimal tiling path of CH17 contig with join details.

Variant	Forward (5'-3')	Reverse (5'-3')	Probe (5'-3')
V7-4-1 Non-insertion	GCTTGGGGAATCAACGACCG	TGTAGCCTACTTGATTGCAGTG	NA
V7-4-1 Insertion	GCTTGGGGAATCAACGACCG	GGGATCCTGCAAATGCCTCTG	NA
V1-8/V3-9 (hg19)	AGTCTGAAATTGGGAGTCCA	CGCATTAACCCAGTCAGATG	NA
V5-a/V3-64 (CH17)	TCTCCACACCTTGGAGCTTG	GTGGGAGCCACAGTCAGTTC	NA
V4-b/V3-43/V3-d/V1-c Insertion	GATTGTGCCAGGACTCACAG	CCCTTCATCTTGGACCTCAG	NA
V4-b/V3-43/V3-d/V1-c Non-insertion	TGAAAATCTGAGCGATGTGC	TGTTTCTAGGATGCAGACATGG	NA
V1-f Insertion	GGGACATTTCTGTGAGTCCAG	CCAGGAACTGATGGGGACTT	NA
V1-69_1 Duplication	GCTGAGGCAGGAGATTTAATCCT	CAGCTCACTGCAAACTCTGCTT	CAAGACCTACAGCCTCAGA

Table S3. Primer and probe list for breakpoint PCR and TaqMan copy number assays

Sample ID	Population	Included in LD analysis
NA18530	CHB	*
NA18534	CHB	*
NA18536	CHB	*
NA18543	CHB	*
NA18544	CHB	*
NA18546	CHB	*
NA18548	CHB	*
NA18549	CHB	*
NA18557	CHB	*
NA18559	CHB	*
NA18595	CHB	*
NA18596	CHB	*
NA18597	CHB	*
NA18599	CHB	*
NA18602	CHB	*
NA18606	CHB	*
NA18610	CHB	*
NA18613	CHB	*
NA18614	CHB	*
NA18615	CHB	*
NA18616	CHB	*

NA18617	CHB	*
NA18618	CHB	*
NA18619	CHB	*
NA18625	CHB	
NA18626	CHB	*
NA18627	CHB	*
NA18628	CHB	*
NA18630	CHB	*
NA18631	CHB	*
NA18634	CHB	*
NA18638	CHB	*
NA18639	CHB	*
NA18640	CHB	*
NA18641	CHB	*
NA18642	CHB	*
NA18643	CHB	*
NA18645	CHB	*
NA18647	CHB	*
NA18740	CHB	*
NA18745	CHB	*
NA18747	CHB	*
NA18748	CHB	*
NA18749	CHB	*
NA18757	CHB	*

HG00174	FIN	*
HG00177	FIN	*
HG00178	FIN	*
HG00180	FIN	*
HG00181	FIN	
HG00182	FIN	*
HG00183	FIN	*
HG00186	FIN	*
HG00188	FIN	*
HG00190	FIN	*
HG00271	FIN	*
HG00272	FIN	*
HG00277	FIN	*
HG00284	FIN	*
HG00285	FIN	*
HG00310	FIN	*
HG00311	FIN	*
HG00312	FIN	*
HG00313	FIN	*
HG00320	FIN	*
HG00323	FIN	*
HG00324	FIN	*
HG00325	FIN	*
HG00326	FIN	*

HG00327	FIN	*
HG00329	FIN	*
HG00331	FIN	*
HG00334	FIN	*
HG00335	FIN	*
HG00336	FIN	*
HG00338	FIN	*
HG00341	FIN	*
HG00342	FIN	*
HG00343	FIN	*
HG00350	FIN	*
HG00351	FIN	*
HG00353	FIN	*
HG00355	FIN	*
HG00356	FIN	*
HG00360	FIN	*
HG00361	FIN	*
HG00368	FIN	
HG00369	FIN	*
HG00372	FIN	*
HG00373	FIN	*
HG00380	FIN	
HG00382	FIN	*
HG00383	FIN	*

HG00097	GBR	*
HG00100	GBR	*
HG00103	GBR	*
HG00104	GBR	*
HG00106	GBR	*
HG00107	GBR	
HG00108	GBR	*
HG00110	GBR	*
HG00111	GBR	*
HG00113	GBR	*
HG00115	GBR	
HG00116	GBR	*
HG00117	GBR	*
HG00119	GBR	*
HG00120	GBR	*
HG00124	GBR	*
HG00128	GBR	*
HG00129	GBR	*
HG00130	GBR	*
HG00136	GBR	*
HG00139	GBR	*
HG00141	GBR	*
HG00142	GBR	*
HG00145	GBR	

HG00148	GBR	*
HG00149	GBR	*
HG00150	GBR	*
HG00151	GBR	*
HG00156	GBR	*
HG00231	GBR	*
HG00232	GBR	*
HG00234	GBR	*
HG00235	GBR	*
HG00236	GBR	*
HG00237	GBR	*
HG00238	GBR	*
HG00240	GBR	*
HG00243	GBR	*
HG00244	GBR	*
HG00247	GBR	*
HG00251	GBR	*
HG00253	GBR	*
HG00255	GBR	*
HG00256	GBR	*
HG00258	GBR	*
HG00262	GBR	*
HG00263	GBR	*
HG00264	GBR	*

HG01500	IBS
HG01501	IBS
HG01503	IBS
HG01506	IBS
HG01507	IBS
HG01509	IBS
HG01510	IBS
HG01513	IBS
HG01515	IBS
HG01516	IBS
HG01525	IBS
HG01527	IBS
HG01530	IBS
HG01531	IBS
HG01602	IBS
HG01612	IBS
HG01613	IBS
HG01615	IBS
HG01619	IBS
HG01624	IBS
HG01625	IBS
HG01626	IBS
HG01631	IBS
HG01672	IBS

HG01678	IBS
HG01679	IBS
HG01684	IBS
HG01686	IBS
HG01694	IBS
HG01695	IBS
HG01697	IBS
HG01700	IBS
HG01704	IBS
HG01707	IBS
HG01708	IBS
HG01746	IBS
HG01747	IBS
HG01761	IBS
HG01762	IBS
HG01770	IBS
HG01771	IBS
HG01773	IBS
HG02223	IBS
HG02224	IBS
HG02232	IBS
HG02233	IBS
HG02235	IBS
HG02238	IBS

NA18939	JPT	*
NA18941	JPT	*
NA18946	JPT	*
NA18954	JPT	*
NA18955	JPT	
NA18957	JPT	*
NA18962	JPT	*
NA18963	JPT	*
NA18977	JPT	*
NA18979	JPT	
NA18993	JPT	
NA19001	JPT	
NA19002	JPT	*
NA19009	JPT	*
NA19010	JPT	*
NA19054	JPT	*
NA19055	JPT	*
NA19056	JPT	*
NA19057	JPT	*
NA19058	JPT	*
NA19059	JPT	*
NA19060	JPT	*
NA19062	JPT	*
NA19063	JPT	*

NA19064	JPT	*
NA19065	JPT	*
NA19066	JPT	*
NA19067	JPT	*
NA19068	JPT	*
NA19070	JPT	*
NA19072	JPT	*
NA19074	JPT	*
NA19075	JPT	*
NA19076	JPT	*
NA19077	JPT	*
NA19078	JPT	*
NA19079	JPT	*
NA19080	JPT	*
NA19081	JPT	*
NA19082	JPT	*
NA19083	JPT	*
NA19084	JPT	*
NA19085	JPT	*
NA19086	JPT	
NA19087	JPT	*
NA19088	JPT	*
NA19027	LWK	
NA19028	LWK	*

NA19035	LWK	*
NA19046	LWK	*
NA19307	LWK	*
NA19309	LWK	*
NA19310	LWK	*
NA19313	LWK	*
NA19315	LWK	*
NA19317	LWK	*
NA19321	LWK	*
NA19334	LWK	*
NA19347	LWK	*
NA19350	LWK	*
NA19352	LWK	*
NA19360	LWK	*
NA19372	LWK	*
NA19376	LWK	*
NA19377	LWK	*
NA19379	LWK	*
NA19380	LWK	*
NA19382	LWK	*
NA19383	LWK	*
NA19384	LWK	*
NA19390	LWK	*
NA19393	LWK	*

NA19396	LWK	*
NA19397	LWK	*
NA19399	LWK	*
NA19403	LWK	*
NA19404	LWK	*
NA19431	LWK	*
NA19434	LWK	*
NA19435	LWK	*
NA19436	LWK	*
NA19438	LWK	*
NA19440	LWK	*
NA19444	LWK	*
NA19448	LWK	*
NA19451	LWK	*
NA19455	LWK	*
NA19456	LWK	*
NA19457	LWK	*
NA19463	LWK	*
NA19468	LWK	*
NA19470	LWK	*
NA19471	LWK	*
NA19472	LWK	*
NA21295	MKK	
NA21297	MKK	

NA21306	MKK
NA21333	MKK
NA21336	MKK
NA21339	MKK
NA21352	MKK
NA21353	MKK
NA21355	MKK
NA21368	MKK
NA21371	MKK
NA21417	MKK
NA21418	MKK
NA21420	MKK
NA21434	MKK
NA21435	MKK
NA21436	MKK
NA21448	MKK
NA21451	MKK
NA21473	MKK
NA21491	MKK
NA21509	MKK
NA21510	MKK
NA21520	MKK
NA21529	MKK
NA21577	MKK

NA21578	MKK	
NA21582	MKK	
NA21587	MKK	
NA21596	MKK	
NA21611	MKK	
NA21616	MKK	
NA21631	MKK	
NA21632	MKK	
NA21649	MKK	
NA21685	MKK	
NA21689	MKK	
NA21719	MKK	
NA21722	MKK	
NA21738	MKK	
NA21742	MKK	
NA21743	MKK	
NA21744	MKK	
NA21774	MKK	
NA21776	MKK	
NA21784	MKK	
NA20505	TSI	*
NA20506	TSI	*
NA20508	TSI	*
NA20509	TSI	*

NA20515	TSI	*
NA20517	TSI	*
NA20518	TSI	*
NA20519	TSI	*
NA20520	TSI	*
NA20522	TSI	*
NA20524	TSI	*
NA20527	TSI	*
NA20529	TSI	*
NA20531	TSI	*
NA20535	TSI	*
NA20538	TSI	*
NA20541	TSI	*
NA20543	TSI	*
NA20581	TSI	*
NA20582	TSI	*
NA20752	TSI	*
NA20753	TSI	*
NA20754	TSI	*
NA20755	TSI	*
NA20758	TSI	*
NA20759	TSI	*
NA20761	TSI	*
NA20765	TSI	*

NA20768	TSI	*
NA20769	TSI	*
NA20770	TSI	*
NA20771	TSI	*
NA20773	TSI	*
NA20783	TSI	*
NA20786	TSI	*
NA20787	TSI	*
NA20795	TSI	*
NA20799	TSI	*
NA20803	TSI	*
NA20804	TSI	*
NA20805	TSI	*
NA20807	TSI	*
NA20808	TSI	*
NA20809	TSI	*
NA20811	TSI	*
NA20813	TSI	*
NA20815	TSI	*
NA20819	TSI	*
NA18486	YRI	*
NA18487	YRI	*
NA18498	YRI	*
NA18499	YRI	*

NA18510	YRI	*
NA18511	YRI	*
NA18520	YRI	*
NA18867	YRI	*
NA18868	YRI	*
NA18873	YRI	*
NA18907	YRI	*
NA18909	YRI	*
NA18910	YRI	*
NA18916	YRI	*
NA18917	YRI	*
NA18923	YRI	*
NA18924	YRI	*
NA18934	YRI	*
NA19095	YRI	*
NA19107	YRI	*
NA19108	YRI	*
NA19113	YRI	*
NA19114	YRI	*
NA19117	YRI	*
NA19118	YRI	*
NA19121	YRI	*
NA19122	YRI	
NA19146	YRI	*

NA19147	YRI	*
NA19149	YRI	*
NA19175	YRI	*
NA19181	YRI	
NA19182	YRI	
NA19184	YRI	
NA19185	YRI	*
NA19189	YRI	*
NA19190	YRI	*
NA19197	YRI	*
NA19198	YRI	*
NA19213	YRI	*
NA19225	YRI	*
NA19226	YRI	
NA19235	YRI	*
NA19236	YRI	*
NA19247	YRI	*
NA19248	YRI	*
NA19256	YRI	*
NA19257	YRI	*

Table S4. Samples used in this study and their respective populations

hg19	CH17	Variant	Novel Allele notes 1	Novel Allele notes 2
*V7-81*01 orf*	region not covered			
*V3-74*01*	*V3-74*01*			
*V3-73*02*	*V3-73*02*			
*V3-72*01*	*V3-72*01*			
*V2-70*13*	*V2-70*01*	allelic/SYN		
not present	*V1-69D*01*	insertion		
not present	*V1-f*01*	insertion		
not present	*V3-h*01 (pseudo)*	insertion		
not present	*V2-70D*04*	insertion		
*V1-69*06*	*V1-69*06*			
*V3-66*03*	*V3-66*03*			
*V3-64*02*	*V3-64*02*			
*V4-61*08*	*V4-61*01*	allelic/NON-SYN		
*V4-59*01*	*V4-59*01*			
*V1-58*02*	*V1-58*01*	allelic/NON-SYN		
*V3-53*01*	*V3-53*02*	allelic/NON-SYN		
*V5-51*01*	*V5-51*01*			
*V3-49*03*	*V3-49*04*	allelic/NON-SYN		
*V3-48*02*	*V3-48*03*	allelic/NON-SYN		
*V1-46*01*	*V1-46*01*			
*V1-45*02*	*V1-45*02*			
*V3-43*01*	*V3-43*01*			
*V4-39*01*	*V4-39*01*			
*V3-38*02 orf*	*V3-38*02 orf*			
*V3-35*01 orf*	*V3-35*01 orf*			

*V4-34*01*	*V4-34*01*			
*V3-33*01*	*V3-33*01*			
*V4-31*02*	*V4-30-2*01*	complex event		
*V3-30*03*	*V3-30*18*	allelic/NON-SYN		
*V4-28*03*	*V4-28*07*	allelic/SYN	Similar to *01 (295 nt/296 nt)	Variant in 1KG; rs56407212 (C/T)
*V2-26*01*	*V2-26*01*			
*V1-24*01*	*V1-24*01*			
*V3-23*01*	*V3-23*04*	allelic/NON-SYN		
*V3-21*01*	*V3-21*01*			
*V3-20*01*	*V3-20*02*	allelic/NON-SYN	Similar to *01 (293 nt/294 nt)	Variant in 1KG; rs112170273 (NON-SYN)
*V1-18*01*	*V1-18*04*	allelic/SYN	Similar to *01 (293 nt/294 nt)	Variant in 1KG; rs72695948 (SYN)
*V3-16*02 orf*	*V3-16*02 orf*			
*V3-15*01*	*V3-15*01*			
*V3-13*01*	*V3-13*05*	allelic/NON-SYN		
*V3-11*01*	*V3-11*06*	allelic/NON-SYN		
*V3-9*01*	*V5-a*03*	complex event		
*V1-8*01*	*V3-64D*06*	complex event		
*V3-7*01*	*V3-7*03*	allelic/SYN		
*V2-5*01*	*V2-5*10*	allelic/NON-SYN		
not present	*V7-4-1*01*	insertion		
*V4-4*07*	*V4-4*02*	allelic/NON-SYN		
*V1-3*02*	*V1-3*01*	allelic/NON-SYN		
*V1-2*02*	*V1-2*04*	allelic/NON-SYN		
*V6-1*01*	*V6-1*01*			
*D1-1*01*	*D1-1*01*			
*D2-2*02*	*D2-2*02*			

*D3-3*01*	*D3-3*01*			
*D4-4*01*	*D4-4*01*			
*D5-5*01*	*D5-5*01*			
*D6-6*01*	*D6-6*01*			
*D1-7*01*	*D1-7*01*			
*D2-8*01*	*D2-8*01*			
*D3-9*01*	*D3-9*01*			
*D3-10*01*	*D3-10*01*			
*D4-11*01 orf*	*D4-11*01 orf*			
*D5-12*01*	*D5-12*01*			
*D6-13*01*	*D6-13*01*			
*D1-14*01 orf*	*D1-14*01 orf*			
*D2-15*01*	*D2-15*01*			
*D3-16*01*	*D3-16*01*			
*D4-17*01*	*D4-17*01*			
*D5-18*01*	*D5-18*01*			
*D6-19*01*	*D6-19*01*			
*D1-20*01*	*D1-20*01*			
*D2-21*02*	*D2-21*02*			
*D3-22*01*	*D3-22*01*			
*D4-23*01 orf*	*D4-23*01 orf*			
*D5-24*01 orf*	*D5-24*01 orf*			
*D6-25*01*	*D6-25*01*			
*D1-26*01*	*D1-26*01*			
*D7-27*01*	*D7-27*01*			
*J1*01*	*J1*01*			

*J2*01*	*J2*01*			
*J3*02*	*J3*02*			
*J4*02*	*J4*02*			
*J5*02*	*J5*02*			
*J6*03*	*J6*03*			

Table S5. Genes, variants and alleles of CH17 compared to GRCh37
Alleles in red represent differences between GRCh37 and CH17 haplotypes; those alleles highlighted in yellow were not previously described.

HAPLO	V GENE*allele	Heptamer (only fam1)	bp	Octamer	bp	TATA	bp	ATG	splice	7mer	bp	9mer	notes
NCBI	V3-74*01			ATGCAAAT	18	AAGAAAA	90	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-74*01			ATGCAAAT	18	AAGAAAA	90	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-73*02			ATGCAAAT	19	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-73*02			ATGCAAAT	19	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-72*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGCG	23	ACACAAACC	
CH17	V3-72*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGCG	23	ACACAAACC	
NCBI	V2-70*13			ATGCAAAT	26	TTCAAAA	41	yes	CTGT/AGGG	CACAGAG	23	ACAAGAACC	
CH17	V2-70*01			ATGCAAAT	26	TTCAAAA	41	yes	CTGT/AGGG	CACAGAG	23	ACAAGAACC	
CH17	V1-69D*01	CTCATGC	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
CH17	V1-f*01	CTCATGA	2	ATGCAAAT	19	TAAATAC	80	yes	AGGT/AGGC	CACAGTG	23	TCAGAAACC	first description
CH17	V2-70D*04			ATGCAAAT	26	TTCAAAA	41	yes	CTGT/AGGG	CACAGAG	23	ACAAGAACC	
NCBI	V1-69*06	CTCATGC	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
CH17	V1-69*06	CTCATGC	2	ATGCAAAT	26	TAAATAT	81	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
NCBI	V3-66*03			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-66*03			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-64*02			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	GCAGAAACC	
CH17	V3-64*02			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	GCAGAAACC	
NCBI	V4-61*08			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	
CH17	V4-61*01			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	
NCBI	V4-59*01			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
CH17	V4-59*01			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
NCBI	V1-58*02	CTCATGA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGC/AGGT	CACAGTG	23	TCAGAAACG	
CH17	V1-58*01	CTCATGA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGC/AGGT	CACAGTG	23	TCAGAAACG	
NCBI	V3-53*01			ATCCAAAT	18	ATGAAAA	98	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	

CH17	V3-53*02			ATGCAAAT	18	ATGAAAA	98	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V5-51*01			ATGCAAAT	18	ACTTAAA	79	yes	AGGT/AGGA	CACAGTG	23	CTAAAACCC	
CH17	V5-51*01			ATGCAAAT	18	ACTTAAA	79	yes	AGGT/AGGA	CACAGTG	23	CTAAAACCC	
NCBI	V3-49*03			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-49*04			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-48*02			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-48*03			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V1-46*01	CTCATGA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
CH17	V1-46*01	CTCATGA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
NCBI	V1-45*02	CTCATCA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGAT	CACAGTG	23	TCAGAAACC	
CH17	V1-45*02	CTCATCA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGT/AGAT	CACAGTG	23	TCAGAAACC	
NCBI	V3-43*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACAAAAACC	
CH17	V3-43*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACAAAAACC	
NCBI	V4-39*01			ATGCAAAT	39	TTAAATT	58	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
CH17	V4-39*01			ATGCAAAT	39	TTAAATT	58	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
NCBI	V3-38*02 orf			____	__	____	___	yes	AGGT/AGGT	TACACAG	23	ACACAAACC	
CH17	V3-38*02 orf			____	__	____	___	yes	AGGT/AGGT	TACACAG	23	ACACAAACC	
NCBI	V3-35*01 orf			ATGCAAAT	18	ATAAAAA	95	yes	AGGT/AGGT	CACTGAG	23	ACACAAACC	
CH17	V3-35*01 orf			ATGCAAAT	18	ATAAAAA	95	yes	AGGT/AGGT	CACTGTG	23	ACACAAACC	
NCBI	V4-34*01			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
CH17	V4-34*01			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACAAAAACC	
NCBI	V3-33*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-33*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V4-30-2*01			ATGCAAAT	38	TTAAATT	59	yes	ATGT/AGGG	CACAATG	23	ACACAAACC	first description
NCBI	V3-30*03			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	

CH17	V3-30*18			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V4-28*03			ATGCAAAT	38	TTAAATT	59	yes	ATGT/AGTG	CACAGTG	23	ACACAAACC	
CH17	V4-28*07 novel			ATGCAAAT	38	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	
NCBI	V2-26*01			ATGCAAAT	26	TTCAAAA	41	yes	CTGT/AGGG	CACAGAG	23	ACAAGAACC	
CH17	V2-26*01			ATGCAAAT	26	TTCAAAA	41	yes	CTGT/AGGG	CACAGAG	23	ACAAGAACC	
NCBI	V1-24*01	CTCATGA	2	ATGCAAAT	19	TAAATAC	80	yes	AGGC/AGGC	CACAGTG	23	TCAGAAACC	
CH17	V1-24*01	CTCATGA	2	ATGCAAAT	19	TAAATAC	80	yes	AGGC/AGGC	CACAGTG	23	TCAGAAACC	
NCBI	V3-23*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-23*04			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-21*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-21*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-20*01			ATGCAGGT	17	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACG	
CH17	V3-20*02 novel			ATGCAGGT	17	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACG	
NCBI	V1-18*01	TTCATGA	2	ATGCAAAT	12	TATAGAT	76	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
CH17	V1-18*04 novel	TTCATGA	2	ATGCAAAT	12	TATAGAT	76	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
NCBI	V3-16*02 orf			ATGCAAAT	18	ATGAAAA	94	yes	AGGT/AGGT	TCCTGTG	23	ACACAAACC	
CH17	V3-16*02 orf			ATGCAAAT	18	ATGAAAA	95	yes	AGGT/AGGT	TCCTGTG	23	ACACAAACC	
NCBI	V3-15*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-15*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-13*01			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-13*05 novel			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V3-11*01			ATGCAAAT	18	ATAAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-11*06 novel			ATGCAAAT	18	ATAAAAA	101	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V5-a*03			ATGCAAAT	18	ACTTAAA	79	yes	AGGT/AGGA	CACAGTG	23	CTAAAACCC	first description

CH17	V3-64D*06 novel			ATGCAAAT	18	ATGAAAA	101	yes	AGGT/AGAT	CACAGTG	23	ACACAAACC	first description
NCBI	V3-7*01			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
CH17	V3-7*03			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	
NCBI	V2-5*01			ATGCAAAT	26	TTGAAAA	42	yes	ATGT/AGGG	CACAAAG	23	ACAAAAACC	
CH17	V2-5*10			ATGCAAAT	26	TTGAAAA	42	yes	ATGT/AGGG	CACAAAG	23	ACAAAAACC	
CH17	V7-4-1*01	CTCATGA	2	ATGCAAAT	8	TAAATAT	80	yes	AGGT/AGGT	CACAGTG	23	TCAAAAACC	first description
NCBI	V4-4*07			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	
CH17	V4-4*02			ATGCAAAT	39	TTAAATT	59	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	
NCBI	V1-3*02	CTCATGA	2	ATGCAAAT	8	TGACTAT	77	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
CH17	V1-3*01	CTCATGA	2	ATGCAAAT	8	TGACTAT	77	yes	AGGT/AGGT	CACAGTG	23	TCAGAAACC	
NCBI	V1-2*02	CTCATGA	2	ATGCAAAT	19	TAAATAC	82	yes	AGGT/AGGA	CACAGTG	23	TCAGAAACC	
CH17	V1-2*04	CTCATGA	2	ATGCAAAT	19	TAAATAC	82	yes	AGGT/AGGA	CACAGTG	23	TCAGAAACC	
NCBI	V6-1*01												
CH17	V6-1*01												
G248	V1-c	CCCATGA	2	ATGCAAAT	19	TAAATAT	81	yes	AGGA/AGGT	CACAGTG	23	TCAGAAAGC	first description
G248, ABC10	V3-d			ATGCAAAT	18	ATGAAAA	101	yes	AGAT/AGGT	CACAGTG	23	ACACAAACC	first description
ABC10	V4-b			ATGCAAAT	39	TTAAATT	58	yes	ATGT/AGGG	CACAGTG	23	ACACAAACC	first description…structure and seqs match V4-39
ABC11	V3-30-3			ATGCAAAT	18	ATGAAAA	100	yes	AGGT/AGGT	CACAGTG	23	ACACAAACC	first description…structure and seqs match V3-30
ABC11	V4-30-2			ATGCAAAT	38	TTAAATT	59	yes	ATGT/AGGG	CACAATG	23	ACACAAACC	first description…same as CH17
ABC11	V4-30-4			ATGCAAAT	38	TTAAATT	59	yes	ATGT/AGGG	CACAATG	23	ACACAAACC	first description

Table S6. RSS motifs identified in CH17 and fosmid haplotypes

Fosmid (Accession)	Overlap CNV?	Gene	Allele	Most similar allele	Variant (NS/SYN)	In 1KG/dbsnp 135?	SNP/s	Notes
ABC8-718240A7	yes	*IGHV3-30-3*	*03, novel	V3-30-3*01; 293/294 nts				Complex event; This gene resides at locus IGHV3-30-3; however, we found allele V3-30*04 here, although this was different by only one bp.
ABC8-718240A7	yes	*IGHV4-30-4*	*07, novel	V4-30-2*05, *01; 288/297	3 NS; 5 SYN	gene not in reference		Complex event: This gene resides at locus IGHV4-30-4; however, allele found here may be hybrid between alleles from multiple loci.
ABC8-5713749P16		*IGHV4-59*	*01					
ABC8-5713749P16		*IGHV4-61*	*02					
ABC9-41239600E19		*IGHV3-43*	*01					
ABC9-43993300H10	yes	*IGHV3-21*	*01					
ABC9-43993300H10	yes	*IGHV3-23*	*01					Duplication
ABC9-43849600N9	yes	*IGHV3-23*	*01					Duplication
ABC9-43849600N9	yes	*IGHV1-24*	*01					
ABC9-43849600N9	yes	*IGHV2-26*	*01					
ABC10-44509600J4	yes	*IGHV4-4*	*08, novel	V4-59*08; 288/291 nt	1 SYN	yes; at IGHV4-59 locus, see notes	1KG_14_107083337	This gene resides at IGHV4-4 locus, but shares sequence with V4-4 and V4-59 alleles. 1KG_14_107083337 shared with V4-59*05. The other two SNPs differentiating this allele from IGHV4-59*08 are also found in IGHV4-59 and IGHV4-4 alleles.
ABC10-44509600J4	yes	*IGHV7-4-1*	*02					Insertion
ABC10-44509600J4	yes	*IGHV2-5*	*10					
ABC10-44687300P14	yes, partial	*IGHV4-30-2*	*01					Complex event
ABC10-44687300P14	yes, partial	*IGHV3-30-3*	*01					Complex event
ABC10-44687300P14	yes, partial	*IGHV4-30-4*	*07, novel	*01; 297/299 nt	1 NS; 1 SYN	gene not in reference		Complex event

ABC10-44687300P14	yes, partial	*IGHV3-30-5*	*02					Complex event; Alleles of IGHV3-30-5 and IGHV3-30 are indistinguishable, but this gene occupies position of IGHV3-30-5 locus.
ABC10-1585770E3	yes, partial	*IGHV1-69*	*14, novel					Insertion
ABC10-1585770E3	yes, partial	*IGHV2-70D*	*14, novel		1 SYN	yes	rs112642180 (A/C)	Insertion
ABC10-1585770E3	yes, partial	*IGHV3-h*	*02					Insertion
ABC10-1585770E3	yes, partial	*IGHV1-f*	*01					Insertion
ABC10-44748900N6	yes	*IGHV3-9*	*03, novel	*01; 297/298 nt	1 NS	yes	rs8020204 (A/G)	
ABC10-44084700I10	yes	*IGHV3-38*	*03, novel	*02; 287/288 nt	1 SYN	yes	rs58308167 (C/A)	
ABC10-44084700I10	yes	*IGHV4-b*	*02					Insertion
ABC10-44145400L1	yes	*IGHV4-b*	*02					Insertion
ABC10-44145400L1	yes	*IGHV3-43D*	*03, novel	*01; 294/298 nt	2 NS; 2 SYN	yes	rs2467912 (T/C); rs2467911 (G/A); rs4977158 (T/G); rs2467910 (T/C)	Insertion
ABC10-44145400L1	yes	*IGHV3-d*	*01					Insertion
ABC11-47355000A14	yes	*IGHV4-28*	*01					
ABC11-47355000A14	yes	*IGHV3-30*	*18					Complex event
ABC11-47355000A14	yes	*IGHV4-30-2*	*06, novel	*01; 298/299	1 NS	gene not in reference		Complex event
ABC11-47150400I4	yes	*IGHV3-30*	*18					Complex event
ABC11-47150400I4	yes	*IGHV4-30-2*	*06, novel	*01; 298/299	1 NS	gene not in reference		Complex event
ABC11-47150400I4	yes	*IGHV3-30-3*	*01					Complex event
ABC11-47354200D2	yes	*IGHV3-30-3*	*01					Complex event
ABC11-47354200D2	yes	*IGHV4-30-4*	*01					Complex event

ABC11-47354200D2	yes	*IGHV3-30-5*	*18					Complex event; Alleles of IGHV3-30-5 and IGHV3-30 are indistinguishable, but this gene occupies position of IGHV3-30-5 locus.
ABC11-49598600E10	yes	*IGHV3-30-5*	*18					Complex event; Alleles of IGHV3-30-5 and IGHV3-30 are indistinguishable, but this gene occupies position of IGHV3-30-5 locus.
ABC11-49598600E10	yes	*IGHV4-31*	*03					
ABC11-49598600E10	yes	*IGHV3-33*	*01					
ABC11-48019600E23	yes	*IGHV2-26*	*01					
ABC11-48019600E23	yes	*IGHV4-28*	*01					
ABC11-48019600E23	yes	*IGHV3-33*	*01					
ABC11-48009100A5		*IGHV3-48*	*06					
ABC12-48925600G6	yes	*IGHV4-4*	*02					
ABC12-48925600G6	yes	*IGHV7-4-1*	*01					
ABC12-48925600G6	yes	*IGHV2-5*	*10					
ABC12-46655100E11		*no functional V genes*						
ABC14-954514B22	yes	*IGHV3-23*	*01					
ABC14-954514B22	yes	*IGHV3-23*	*01					
ABC14-954514B22	yes	*IGHV1-24*	*01					
ABC14-986905I21		*IGHV3-48*	*02					
ABC14-986905I21		*IGHV3-49*	*05					
ABC14-986905I21		*IGHV5-51*	*01					
ABC14-1051114P8		*no functional V genes*						
WI2-1707G1	yes	*IGHV3-d*	*01					Insertion
WI2-1707G1	yes	*IGHV1-c*	*01					Insertion
WI2-1707G1	yes	*IGHV4-39*	*07					

WI2-3737C11	yes	*IGHV3-d*	*01					Insertion
WI2-3737C11	yes	*IGHV1-c*	*01					Insertion
WI2-3737C11	yes	*IGHV4-39*	*07					
ABC24-366M9	yes	*IGHV3-30*	*02					
ABC24-366M9	yes	*IGHV4-34*	*01					

Table S7. IGHV gene and allelic variants identified from annotated fosmid clones

CLONE	IGHC	IGHJ	IGHD	IGHV	V gene SMs	2nd V gene	2nd V gene SMs	Notes
ABC7_000043068300_J5	M	4*02	2-15*01	1-69*12	10	NA	NA	
ABC8_000041787500_L13	G1	5*02	5-18*01	2-5*10	11	NA	NA	
ABC8_000040989700_P12	G1	4*02	3-10*01	3-20*01	8	NA	NA	
ABC8_4_1_000000787522_D4	M	4*02	3-22*01	3-49*03	4	NA	NA	
ABC10_2_1_000044117600_K12	G1	5*02	2-21*02	2-5*10	18	NA	NA	
ABC10_2_1_000044789200_G15	NI	5*02	2-21*02	2-5*10	18	3-7*01	0	
ABC11_2_1_000048179600_D4	A1	6*03	2-21*01	4-59*01/08	28	4-61*01	0	
ABC11_2_1_000048290000_H9	A1	5*02	2-21*01	4-39*01	24	NA	NA	Fosmid sequence contains 109 bp insertion within IGHV4-39 gene, which may be artifactual. This was removed for IMGT V-Quest analysis.
ABC11_2_1_000048039500_G23	NI	4*02	5-24*01	3-33*01	27	4-34*01/ 3-35*01	0/0	
ABC14_4_1_000000947614_C21	G1	2*01	4-23*01	4-39*01	8	NA	NA	

Table S8. V-(D)-J somatic rearrangement events identified in fosmid clones
SM, somatic mutation; NI, C gene not included in fosmid sequence; NA, not applicable