Supporting Figures

Genotyping allelic and copy number variation in the immunoglobulin heavy chain locus

Shishi Luo[1,2,*], Jane A. Yu[1], Yun S. Song[1,2,3,*]

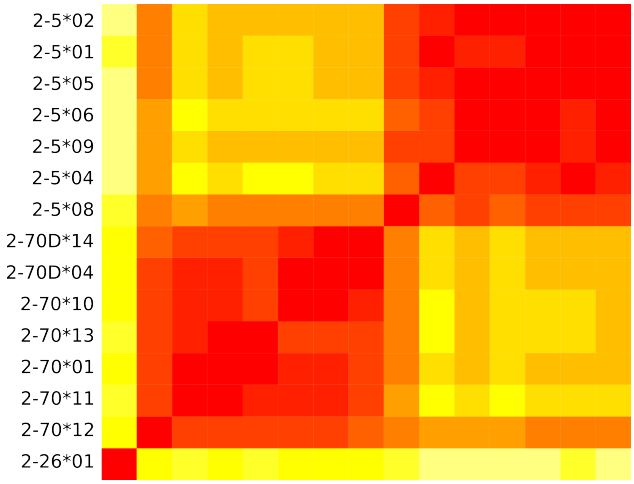

Fig. S1: Hierarchical clustering applied to Hamming distance between all family 2 alleles. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

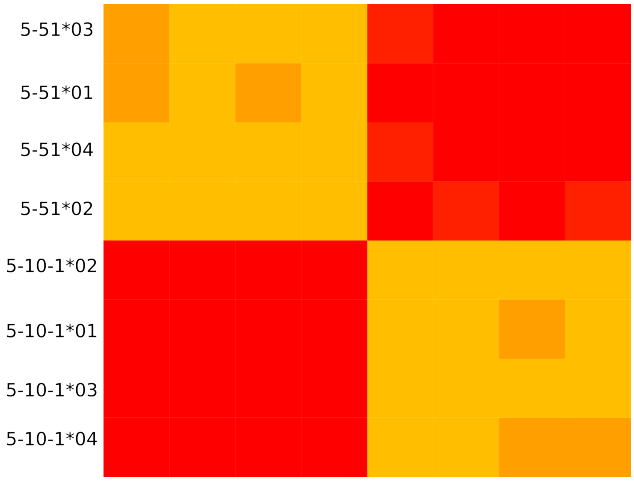

Fig. S2: Hierarchical clustering applied to Hamming distance between all family 5 alleles. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

[1]Computer Science Division, University of California, Berkeley, CA, 94720, USA
[2]Department of Statistics, University of California, Berkeley, CA, 94720, USA
[3]Departments of Mathematics and Biology, University of Pennsylvania, Philadelphia, PA, 19104, USA
*To whom correspondence should be addressed: shishi.luo@berkeley.edu, yss@eecs.berkeley.edu

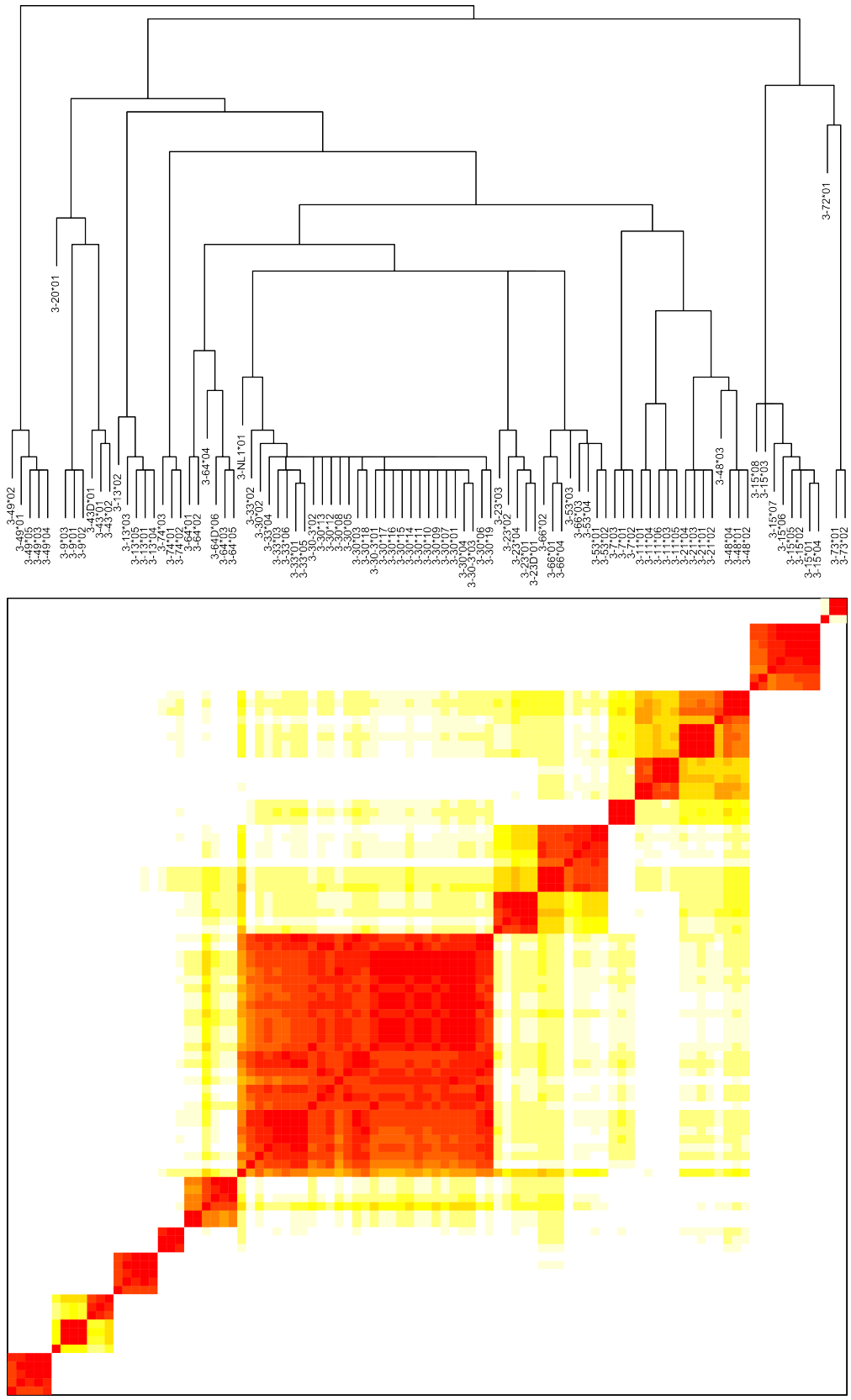

Fig. S3: Hierarchical clustering applied to Hamming distance between all family 3 alleles. Allele labels are in cladogram above matrix. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

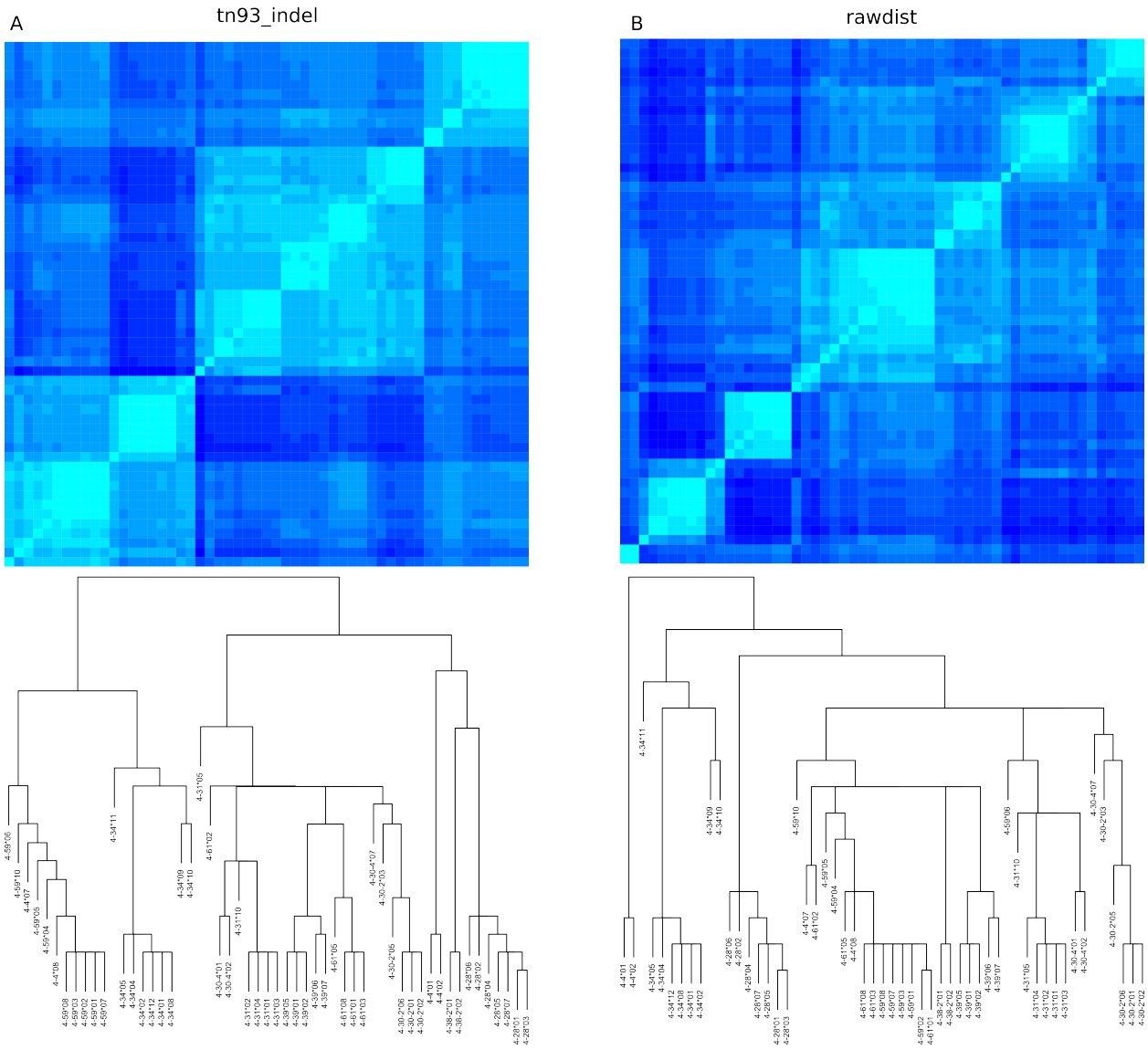

Fig. S4: Hierarchical clustering applied to all family 4 alleles. (A) Simple average of 'TN93' evolutionary distance and indel distance. (B) Hamming distance. Allele labels are in cladogram below matrix. Because TN93 and indel distances cannot be interpreted in terms of nucleotide similarity, the distances in each matrix have been normalized by the maximum value in the matrix for comparison. Heatmap color scale is cyan=0 and blue=1. The clustering that uses 'TN93' and 'indel' distances is cleaner and was used to define the operational segments for family 4 in Table 1.

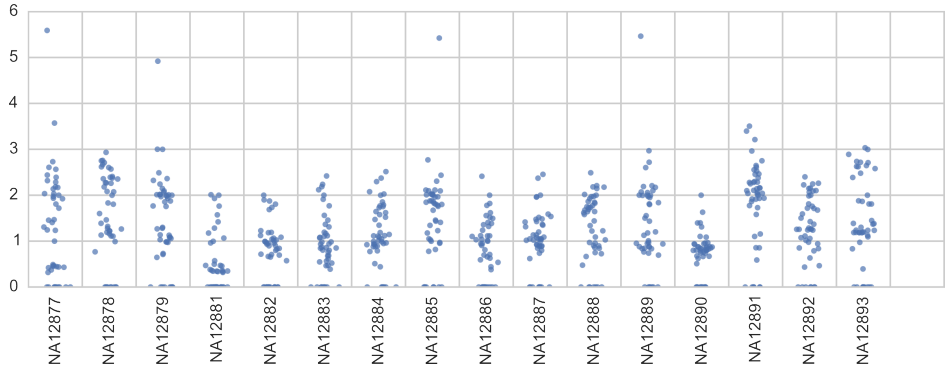

Fig. S5: Dotplots of coverage calls for each individual in the Platinum Genomes dataset. The data points are the same as in Fig. 4 but grouped by individuals. Y axis is normalized read coverage depth.

```
# Aligned_sequences: 2
# 1: 7-4-1*04
# 2: 7-4-1*04_5
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 296
# Identity:     291/296 (98.3\%)
# Similarity:   291/296 (98.3\%)
# Gaps:           0/296 ( 0.0\%)
# Score: 1435.0
#
#
#=====================================

7-4-1*04           1 caggtgcagctggtgcaatctgggtctgagttgaagaagcctggggcctc     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
7-4-1*04_5         1 CAGGTGCAGCTGGTGCAATCTGGGTCTGAGTTGAAGAAGCCTGGGGCCTC     50

7-4-1*04          51 agtgaaggtttcctgcaaggcttctggatacaccttcactagctatgcta    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
7-4-1*04_5        51 AGTGAAGGTTTCCTGCAAGGCTTCTGGATACACCTTCACTAGCTATGCTA    100

7-4-1*04         101 tgaattgggtgcgacaggcccctggacaagggcttgagtggatgggatgg    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
7-4-1*04_5       101 TGAATTGGGTGCGACAGGCCCCTGGACAAGGGCTTGAGTGGATGGGATGG    150

7-4-1*04         151 atcaacaccaacactgggaacccaacgtatgcccagggcttcacaggacg    200
                     ||||||||||||||||||||||.||||||||||||||||||||||||||||
7-4-1*04_5       151 ATCAACACCAACACTGGGAACCTAACGTATGCCCAGGGCTTCACAGGACG    200

7-4-1*04         201 gtttgtcttctccttggacacctctgtcagcatggcatatctgcagatca    250
                     |||||||||||||.||||||||.||||||||||||||||.||.||||||||
7-4-1*04_5       201 GTTTGTCTTCTCCATGGACACCTCCGTCAGCATGGCATATCTTCATATCA    250

7-4-1*04         251 gcagcctaaaggctgaggacactgccgtgtattactgtgcgagaga    296
                     ||||||||||||||||||||||||||||||||||||||||||||||||
7-4-1*04_5       251 GCAGCCTAAAGGCTGAGGACACTGCCGTGTATTACTGTGCGAGAGA    296
```

Fig. S6: Pairwise alignment of the putative 7-4-1 allele, 7-4-1*04_5, with its closest matching IMGT allele, 7-4-1*04. The allele 7-4-1*04_5 was found in individuals NA12877, NA12878, NA12879, NA12883, NA12884, NA12886, NA12888, NA12891, and NA12893. Pairwise alignment was performed using the online EMBOSS Water tool.

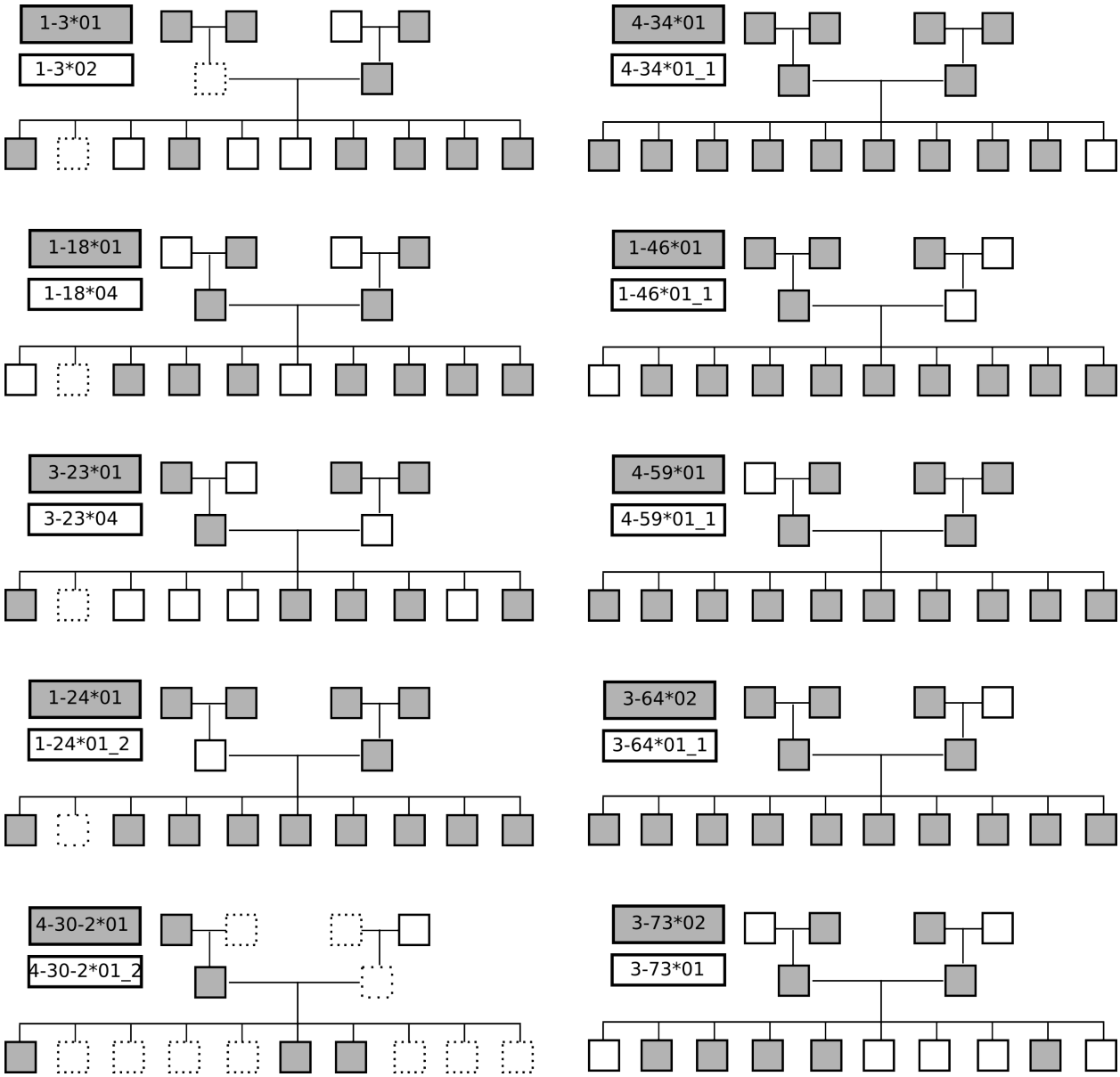

Fig. S7: Allele calls arranged according to family pedigree. Only segments for which there were two alleles in the family are shown (colored grey and white). Individuals who did not carry the segment are denoted by boxes with dashed outlines.

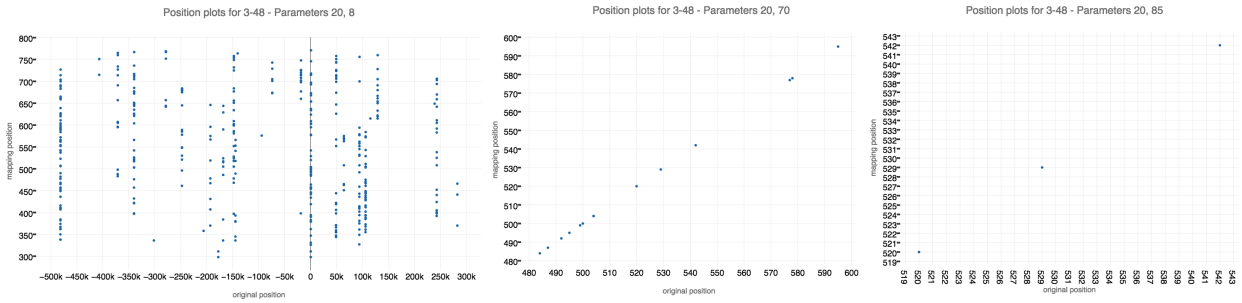

Fig. S8: Mapped position versus original position of the start of each 250 bp read whose alignment exceeds the score threshold for segment 3-48. Axis values are centered at position chr14:1,062,766,005. (A) With default Bowtie2 local alignment threshold of $20 + 8.0\ln(L)$, where L is the read length, reads originally from pseudogenes or similar functional segments are incorrectly mapped to 3-48, as seen by multiple vertical strips of dots. (B) With the threshold increased to $20 + 70\ln(L)$, a single diagonal row of dots indicates that only reads from 3-48 are mapped to segment 3-48. (C) When the threshold is increased to $20 + 85\ln(L)$ however, this is too restrictive and too few reads are mapped. Assessing analogous plots for the rest of the segments led to a threshold of $20 + 70\ln(L)$ being chosen. The README of the package provides more detail on how to modify the threshold. (Coordinates are for chromosome 14 on GRCh37).

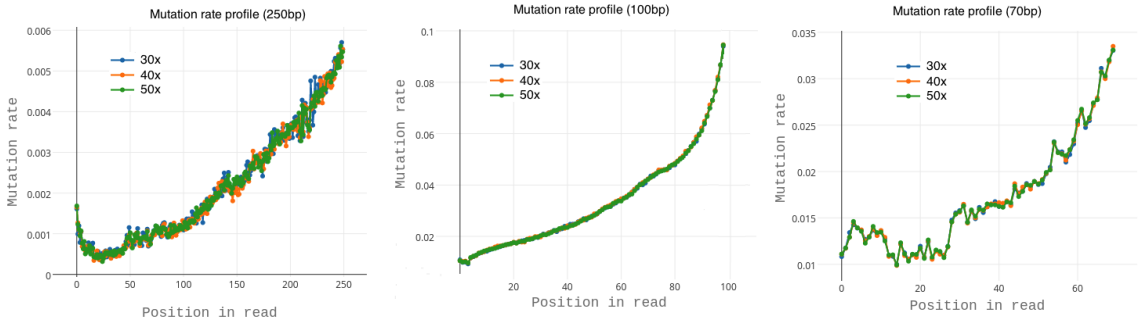

Fig. S9: Error profiles of simulated reads under default ART parameters. Plots are shown for 30x, 40x, and 50x coverages and are only displayed for GRCh37 (plots for GRCh38 are similar). Note the high error rates for 100 bp and 70 bp reads. This difference is attributed to the fact that ART automatically selects one of several built-in read quality profiles according to the read length provided. Mutation rates are computed by first calculating, for each position in the read, the number of mismatches between the position of the simulated nucleotide and the original nucleotide. The number of mismatches was then divided by the total number of reads.

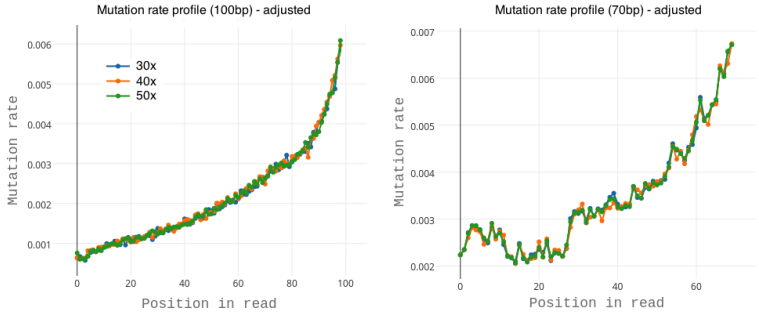

Fig. S10: Error profiles of simulated reads after parameter adjustment. To make the profiles for 70 bp and 100 bp comparable to that of 250 bp, the parameter for quality score shifting (-qs and -qs2) was used: 12.896 for 100bp and 7.99 for 70bp.