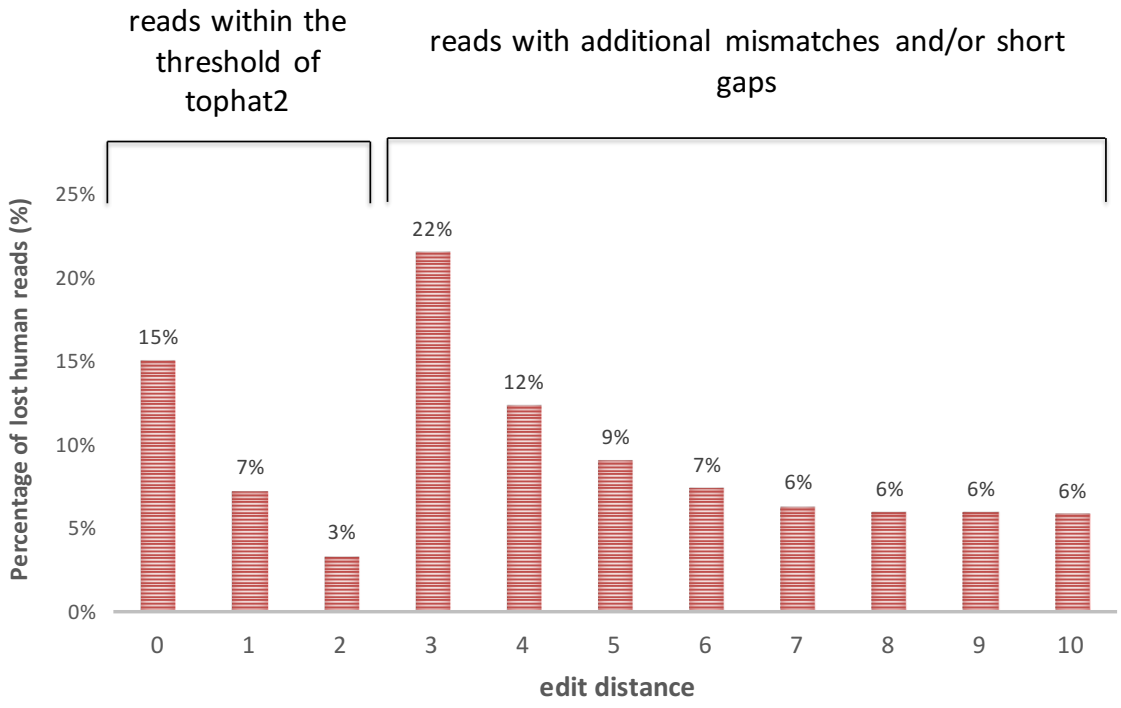

Figure S1. Edit distance of lost human reads. Unmapped reads were remapped to the human references using Megablast. Edit distance was calculated as the minimum number of operations required to transform a read sequence into the corresponding reference subsequence. Reads are grouped by edit distance with the transcriptome or the genome reference. The percentages are the averages across 87 samples.

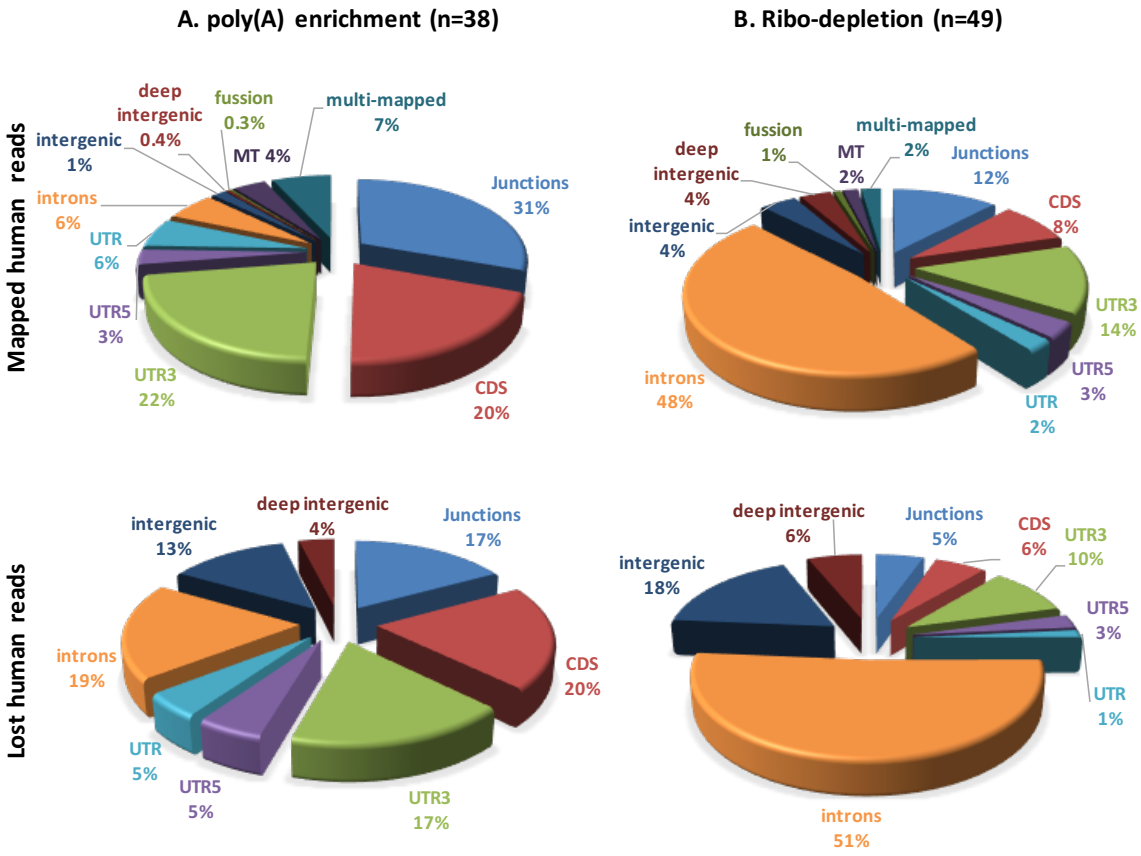

Figure S2. Genomic profile of mapped and lost human reads. (A) RNA-Seq samples were prepared by poly(A) enrichment protocol (n=38). (B) RNA-Seq samples were prepared by ribo-depletion protocol (n=49). Mapped human reads are identified as RNA-Seq reads that mapped to the human reference genome and transcriptome (ENSEMBL hg19 build, ENSEMBL GRCh37 transcripome) via tophat2. Lost human reads are unmapped RNA-Seq reads that aligned to the human reference genome and transcriptome (ENSEMBL hg19 build, ENSEMBL GRCh37 transcripome) via more sensitive Megablast alignment. Single alignment is reported for each read by Megablast. ROP categorizes the reads into genomic categories based on the compatibility of each read from the pair with the features defined by the Ensembl gene annotations. Percentages are calculated as a fraction of reads from a category from the total number of mapped or lost human reads. Junction read is defined as a read spanning exon-exon boundary; CDS, UTR3, UTR5: reads overlapping CDS, UTR3 or UTR5 region; UTR: reads simultaneously overlapping UTR3 and UTR5 regions; intronic: reads overlapping intronic regions; intergenic: reads mapped within the proximity of 1Kb from the gene boundaries; deep intergenic: reads mapped beyond the proximity of 1Kb from the gene boundaries; MT: mitochondrial reads; multi-mapped: reads mapped to multiple locations of the human genome; fusion: reads from the read pair mapped to different chromosomes.

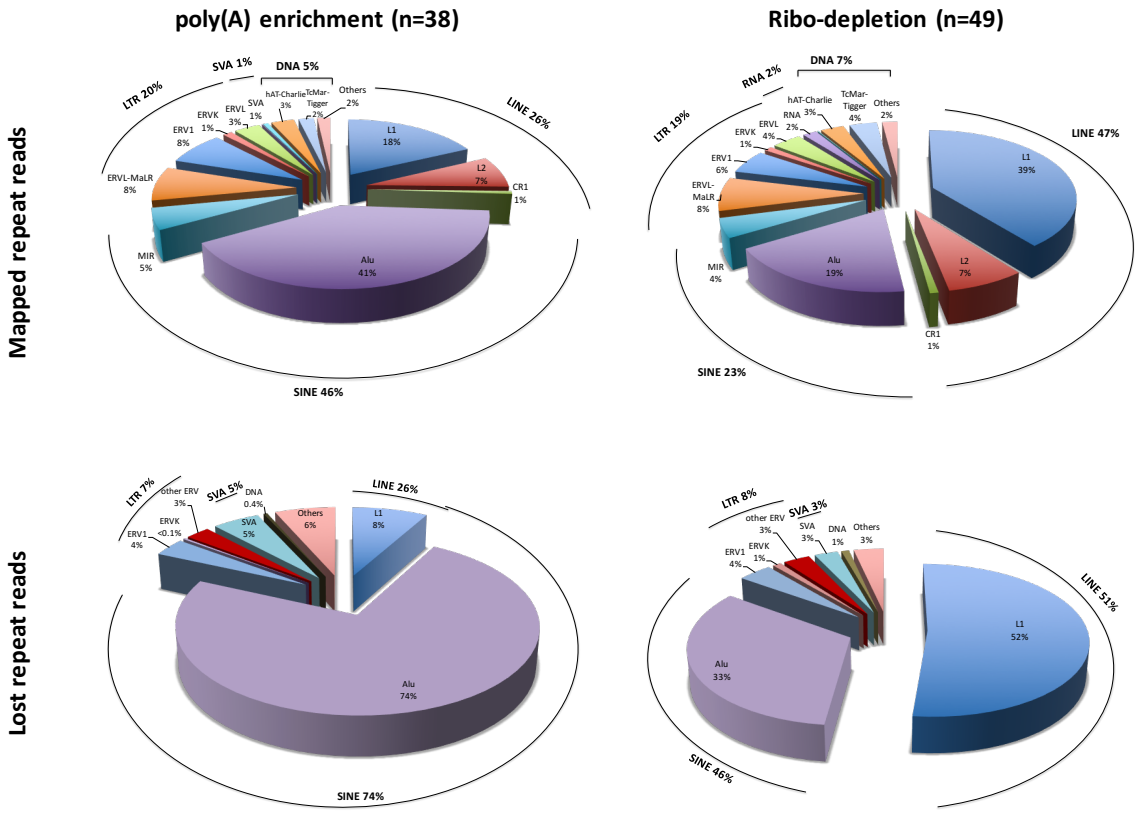

Figure S3. Profile of repeat elements. ROP identifies and categorizes repetitive sequences among the mapped and unmapped reads. RNASeq samples were prepared by poly(A) enrichment protocol (n=38) and ribo-depletion protocol (n=49). Mapped reads were categorized based on the overlap with the repeat instances prepared from RepeatMasker annotation (Repeatmasker v3.3, Repeat Library 20120124). Lost repeat reads are unmapped RNA-Seq reads aligned onto the reference repeat sequences (prepared from Repbase v20.07).

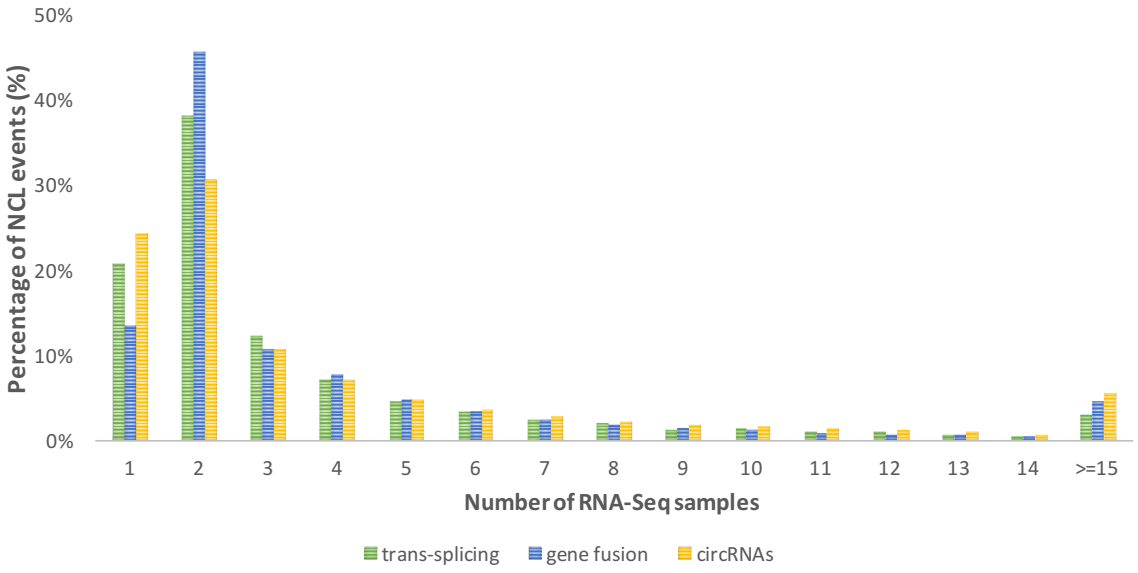

Figure S4. Distribution of non-co-linear (NCL) events across samples. ncSplice custom pipeline was designed to detect NCL events. Trans-splicing events are identified from reads that are spliced distantly on the same chromosome. Gene fusion events are identified from reads spliced across different chromosomes. CircRNAs are identified from reads spliced in a head-to-tail configuration.

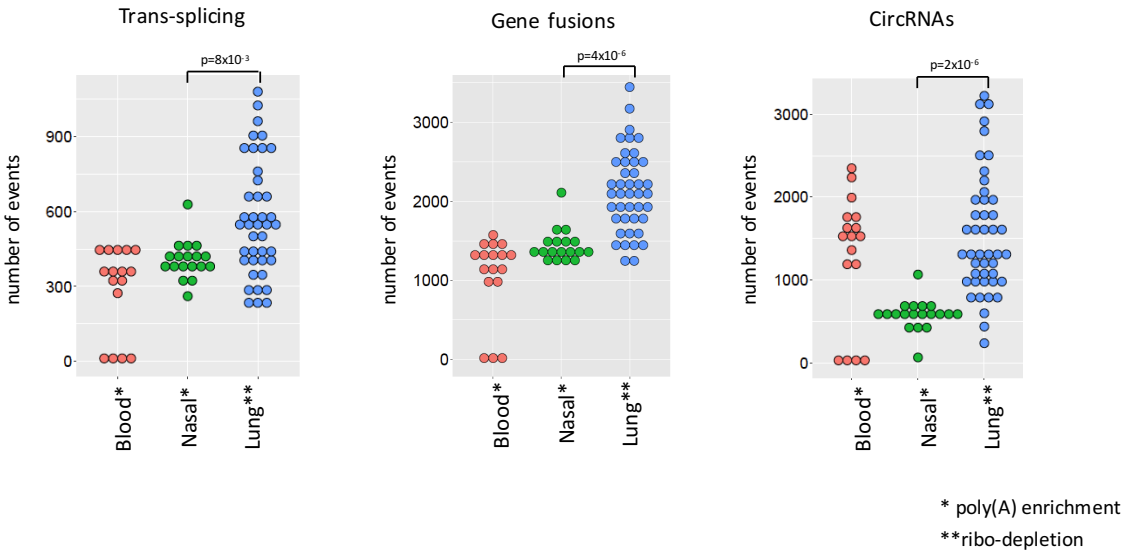

Figure S5. Increased number of NCL events in samples prepared by ribo-depletion protocol. NCL events per sample, detected by ncSplice custom pipeline. Sample were prepared by poly(A) selection (whole blood and nasal epithelium) and ribo-depletion (lung epithelium) protocols. Trans-splicing events are identified from reads spliced distantly on the same chromosome. Gene fusion events are identified from reads spliced across different chromosomes. CircRNAs are identified from reads spliced in a head-to-tail configuration. Samples prepared by ribo-depletion protocol (nasal epithelium versus lung epithelium) show the increase in number of NCL events (p-value < 8 x 10^{-3}). The average increase of 29% of trans-splicing events is detected in the samples prepared by the ribo-depletion protocol, compared to those events in the samples prepared by the poly(A) protocol. The average increase of 32% of gene fusion events is detected in the samples prepared by ribo-depletion protocol, compared to those events of the poly(A) protocol prepared samples. The average increase of 63% of circRNAs events is detected in the samples prepared by ribo-depletion protocol, compared to those prepared by poly(A) protocol.

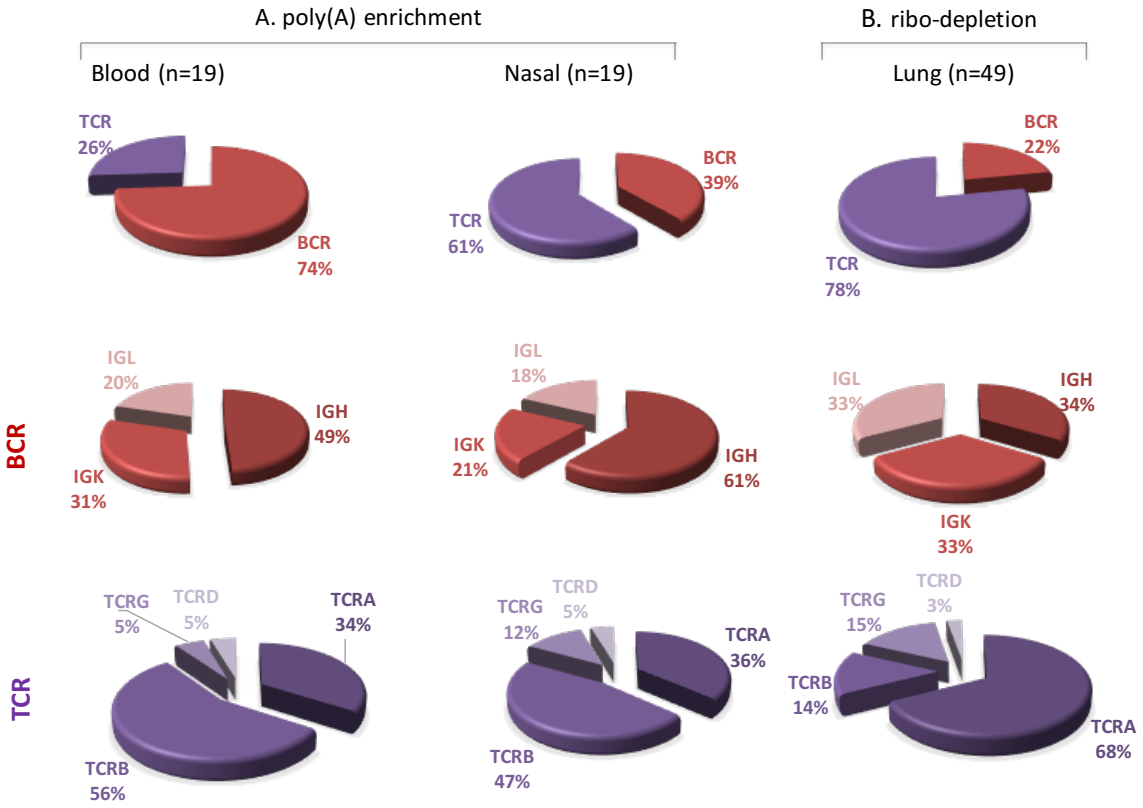

Figure S6. . Percentage of immune reads mapped to B-cell receptor (BCR) and T-cell receptor (TCR) loci. (A) RNA-Seq samples were prepared by poly(A) enrichment protocol (whole blood and nasal epithelium). (B) RNA-Seq samples were prepared by ribo-depletion protocol (lung epithelium). Immune reads that are entirely mapped to BCR and TCR genes are identified by tophat2. Immune reads with extensive somatic hyper mutations (SHM) and reads arising from V(D)J recombination are identified by IgBlast. Blood samples show larger fraction of reads mapped to BCR locus, while nasal and lung epithelium samples show larger fraction of reads mapped to TCR locus. BCR are composed of heavy (IGH) and light chains. Among the reads mapped to BCR locus, the number of reads mapped to immunoglobulin heavy locus (IGH), immunoglobulin kappa locus (IGK), and immunoglobulin lambda locus (IGL) is determined. Among the reads mapped to TCR locus, the number of reads mapped to T cell receptor alpha locus (TCRA), T cell receptor beta locus (TCRB), T cell receptor gamma locus (TCRG), and T cell receptor delta locus (TCRD) is determined.

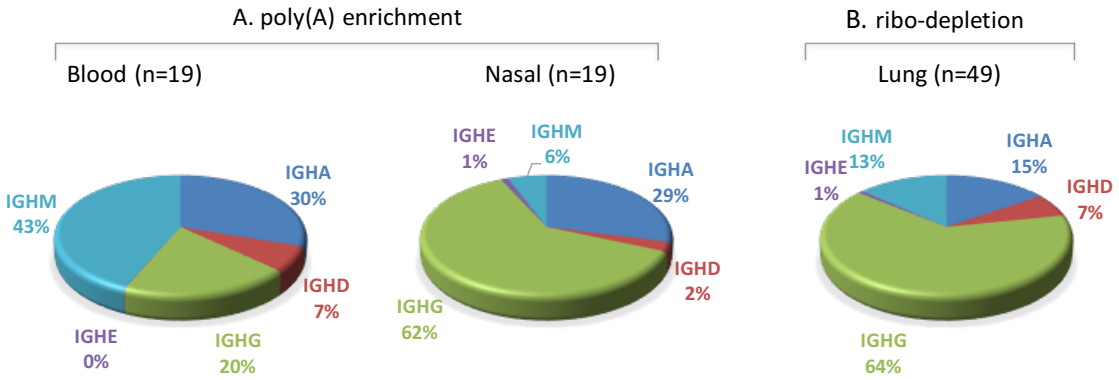

Figure S7. Percentage of immune reads mapped to genes encoding the constant region of immunoglobulin heavy locus (IGH). (A) RNA-Seq samples were prepared by poly(A) enrichment protocol (whole blood and nasal epithelium). (B) RNA-Seq samples were prepared by ribo-depletion protocol (lung epithelium). Immune reads that are entirely mapped to IGHA (Immunoglobulin Heavy Constant Alpha), IGHD (Immunoglobulin Heavy Constant Delta), IGHG (Immunoglobulin Heavy Constant Gamma), IGHE (Immunoglobulin Heavy Constant Epsilon), and IGHM (Immunoglobulin Heavy Constant Mu) are identified by tophat2.

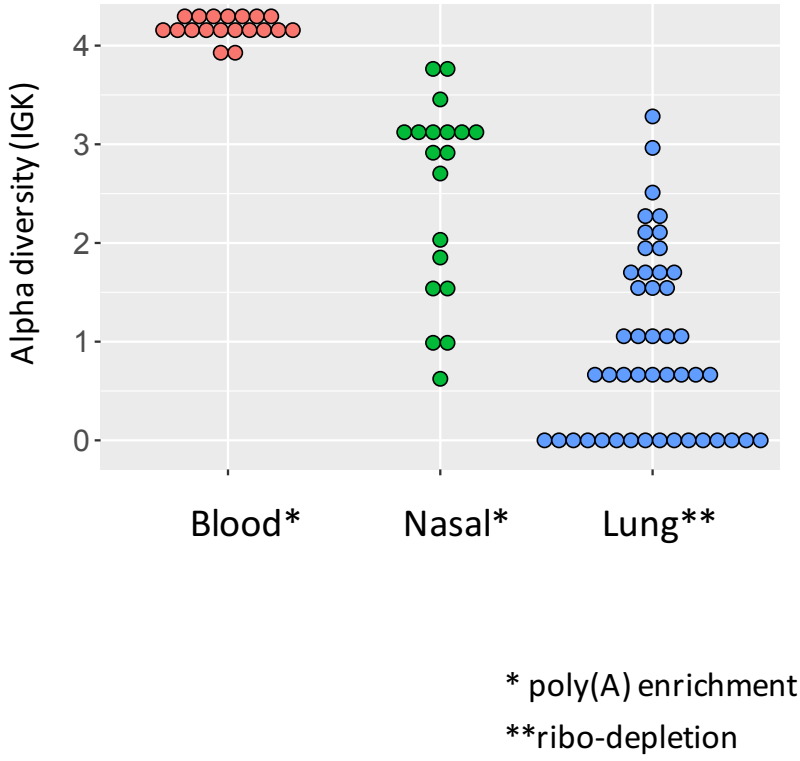

Figure S8. Combinatorial diversity of immunoglobulin kappa locus (IGK) locus across tissues. Samples were prepared by poly(A) selection (whole blood and nasal epithelium) and ribo-depletion (lung epithelium) protocols. The combinatorial diversity of IGK locus is determined based on the recombinations of the VJ gene segments. Alpha diversity is measured using the Shannon entropy incorporating the total number of VJ combinations and their relative proportions. Mean alpha diversity for blood samples was 4.2, for nasal samples, was 2.5, and for lung, was 1.0.

.

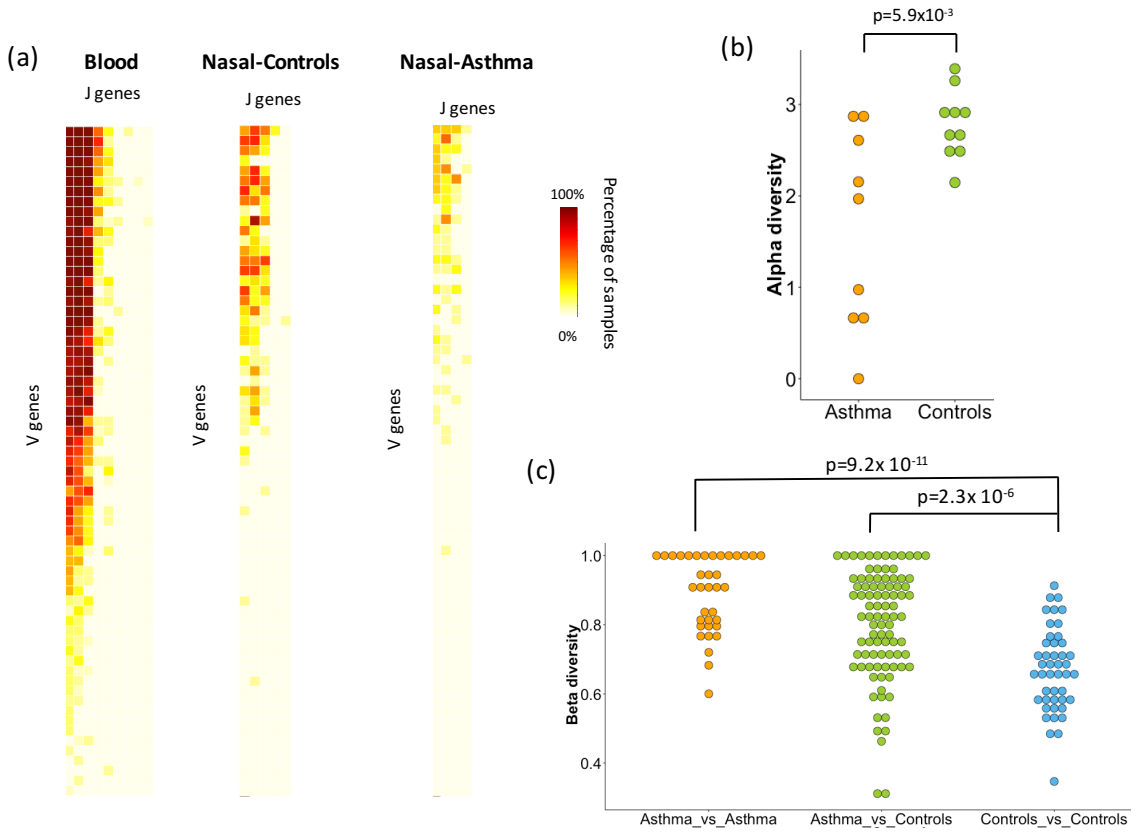

Figure S9. Combinatorial diversity of immunoglobulin lambda locus (IGL) locus differentiates disease status. (a) Heat map depicting the percentage of RNA-Seq samples supporting particular VJ combination for whole blood, nasal epithelium of healthy controls and asthmatic individuals. Each row corresponds to a V gene and each column corresponds to a J gene. (b) Alpha diversity is measured using the Shannon entropy incorporating the total number of VJ combinations and their relative proportions. Nasal epithelium of asthmatic individuals exhibits decreased combinatorial diversity of IGK locus compared to that of healthy controls (p-value=5.9×10^{-3}) (c) Compositional similarities between the samples in terms of gain or loss of VJ combinations of IGK locus are measured using the Sørensen–Dice index across pairs of samples from the same group (Asthma, Controls) and pairs of sample from different groups (Asthma versus Controls). Lower level of similarity is observed between nasal samples of the asthmatic individuals compared to the unaffected controls (p-value<9.2×10^{-11}). Nasal samples of the unaffected controls are more similar to each other than to the asthmatic individuals (p-value<2.3×10^{-6}).

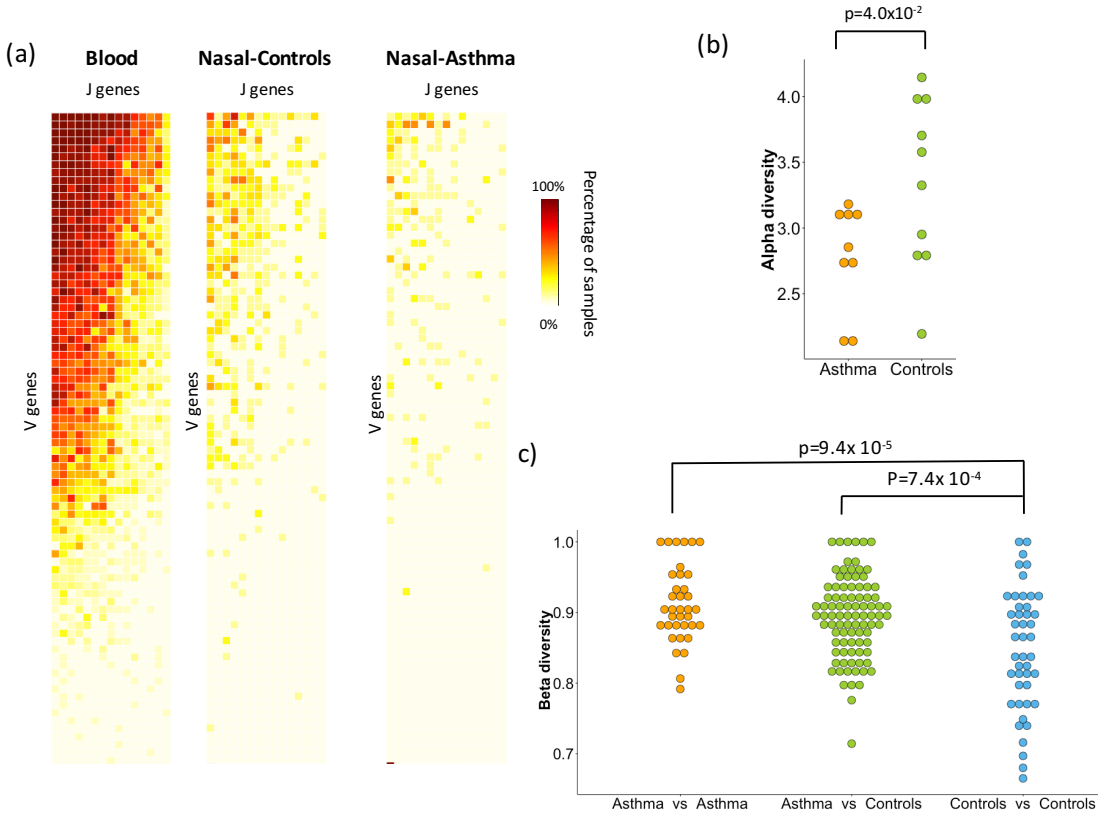

Figure S10. Combinatorial diversity of T cell receptor beta (TCRB) locus differentiates disease status. (a) Heat map depicting the percentage of RNA-Seq samples supporting of particular VJ combination for whole blood, nasal epithelium of healthy controls and of asthmatic individuals. Each row corresponds to a V gene and each column corresponds to a J gene. (b) Alpha diversity is measured using the Shannon entropy incorporating the total number of VJ combinations and their relative proportions. The nasal epithelium of asthmatic individuals exhibits decreased combinatorial diversity of IGK locus compared to that of healthy controls (p-value = 4.0×10^{-2}) (c) Compositional similarities between the samples in terms of gain or loss of VJ combinations of IGK locus are measured using the Sørensen–Dice index across pairs of sample from the same group (Asthma, Controls) and pairs of sample from different groups (Asthma versus Controls). Lower level of similarity is observed between nasal samples of asthmatic individuals compared to unaffected controls (p-value < 9.4×10^{-5}). Nasal samples of unaffected controls are more similar to each other than to the asthmatic individuals (p-value < 7.4×10^{-4}).

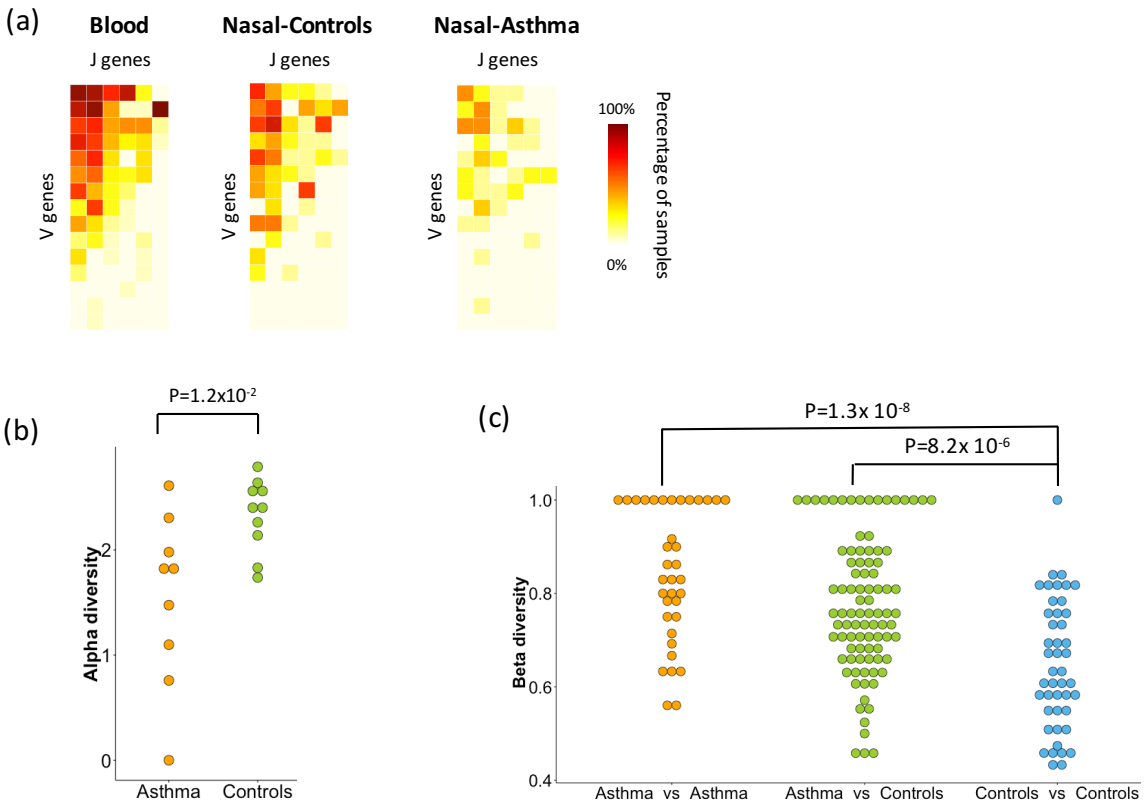

Figure S11. Combinatorial diversity of T cell receptor gamma (TCRG) locus differentiates disease status. (a) Heat map depicting the percentage of RNA-Seq samples supporting of a particular VJ combination for whole blood, nasal epithelium of healthy controls and asthmatic individuals. Each row corresponds to a V gene and each column corresponds to a J gene. (b) Alpha diversity is measured using the Shannon entropy incorporating the total number of VJ combinations and their relative proportions. Nasal epithelium of asthmatic individuals exhibits decreased combinatorial diversity of IGK locus compared to that of healthy controls (p-value = 1.2×10^{-2}, ANOVA). (c) Compositional similarities between the samples in terms of gain or loss of VJ combinations of IGK locus are measured using the Sørensen–Dice index across pairs of sample from the same group (Asthma, Controls) and pairs of sample from different groups (Asthma versus Controls). Lower level of similarity is observed between nasal samples of asthmatic individuals compared to unaffected controls (p-value < 1.3×10^{-8},). Nasal samples of unaffected controls are more similar to each other than to the asthmatic individuals (p-value < 8.2×10^{-6}).