The T2T-CHM13 reference genome has more accurate sequences for immunoglobulin genes than GRCh38

Junli Nie[1,2], Julie Tellier[1,2], Ilariya Tarasova[1,2], Stephen L. Nutt[1,2,*] and Gordon K. Smyth[1,3,*]

[*] Correspondence: smyth@wehi.edu.au, nutt@wehi.edu.au

[1] Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.

[2] Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia.

[3] School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, Australia.

Abstract

Background: The production of antibody by members of the B-cell lineage is essential for protective immunity. The clonal selection theory posits that each mature B cell has a unique immunoglobulin receptor, generated through random gene recombination, and when stimulated to differentiate into an antibody-secreting cell has the capacity to produce only a single antibody specificity. Based on this classical, "one cell one antibody" dogma, analysis of single cell RNA sequencing data of antibody-secreting cells should reveal that each cell expresses only a single form of each of the immunoglobulin heavy and light chains. However, when using GRCh38 as the genome reference, many plasma cells appear to express multiple immunoglobulin isotypes.

Results: We show that this false mapping is caused by the inaccurate immunoglobulin sequences provided by GRCh38. The newly published human genome reference T2T-CHM13, due to its more accurate sequences, avoids this false mapping caveat. In addition, further reads mapped to GRCh38 with ambiguity also settle down perfectly.

Conclusions: These studies reveal that the sequences of the immunoglobulin genes within T2T-CHM13 are more accurate than GRCh38. Thus T2T-CHM13 is the reference genome of choice for accurate mapping and identification of the human immunoglobulin genes.

Keywords: Reference genome, T2T-CHM13, GRCh38, Immunoglobin genes, Clonal selection theory

Background

The human genome was first published in 2001[1-3] and has been continually improved over the past 20 years. For the past decade, the standard human reference genome in use by the UCSC Genome Browser, RefSeq, Ensembl and Gencode has been the GRCh38 genome released by the Genome Reference Consortium in 2013[4]. Additional patches have been released regularly, most recently GRCh38.p14 in February 2022 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40). GRCh38 has been accepted as a high quality representation of the 92% of the human genome that is euchromatic and accessible for gene transcription and has been the basis of most human genomic analyses over the past decade[3,4].

Recently, the telomere-to-telomere (T2T) consortium published the T2T-CHM13 assembly, the first essentially complete sequence of a human genome[5]. The T2T-CHM13 assembly provides complete coverage of heterochromatic and repetitive regions of the human genome, and has been found to improve identification of structural variations[6], but the differences between T2T-CHM13 and GRCh38 for protein-coding genes have not been fully explored. In this article, we show that the T2T-CHM13 genome is able to correctly resolve the isotype of antibody-secreting cells whereas the GRCh38 genome does not, demonstrating that the sequences of immunoglobin genes are more accurate in the new T2T-CHM13 assembly.

B cells and immunoglobins form a major arm of the adaptive immune system. Antigens binding to B cell receptors, with suitable co-stimulatory signals, activate B cells to proliferate and differentiate into antibody-secreting cells (ASCs). The antibodies produced by ASCs present a high binding affinity to their target antigens, neutralizing infectious threats, regulating the microbiota, or clearing dead cells from the body. Antibodies are encoded by the immunoglobulin genes (*IGH*, *IGK*, *IGL*) and expressed exclusively by the lymphocytes of the B cell lineage. An antibody molecule consisting of two identical heavy chains and two identical light chains. Both heavy and light chains possess an N-terminal variable region that determines antigen specificity, and a C-terminal constant region. Based on the type of the constant region of the heavy chain, human immunoglobulins are divided into five major isotypes or classes, IgM, IgD, IgG, IgA, and IgE, among which IgG is further divided into IgG1, IgG2, IgG3, and IgG4 and IgA into IgA1 and IgA2 (Fig S1).

An essential step in the formation of a functional humoral immune system is the somatic genomic recombination that produces the variable sequences of each immunoglobulin gene (Fig S1). This process initiates in developing B cells in the bone marrow and occurs in an ordered fashion where the *IGH* locus undergoes sequential Diversity (D) to Joining (J) recombination, followed by Variable (V) to DJ recombination, to produce a single VDJ exon. Upon successful VDJ recombination, the recombination process is inhibited such that the developing B cell only carries a single functional VDJ rearranged *IGH*, a process termed allelic exclusion[7,8,]. Following successful *IGH* rearrangement, developing B cells then undergo a similar (V-J) recombination process in one of the two immunoglobulin light chain loci (*IGK*, *IGL*).

Successful VDJ-recombination results in the circulation of mature naïve B cells expressing two forms of immunoglobulin, IgM and IgD, generated by alternative splicing (Fig S1). Once activated, B cells have the unique capacity to replace the IgM/IgD constant regions with other isotypes (IgG, IgA, or IgE) in a process called immunoglobulin class switch recombination (CSR). CSR results in the VDJ sequences of the *IGH* being recombined to the downstream constant regions of IgG, IgE or IgA by enzymatic deletion of all intervening sequences[9] (Fig S1). The sequence deletion physically ensures that only one isotype is expressed per cell. Key features of the system are that the antibody isotypes have distinct functions[10-15] and that CSR

is differentially induced by infections[16-18], mRNA vaccines[19], autoimmune diseases[20] and in anti-tumour responses[21]. This process thus allows the type of antibody response to be tailored to meet the challenge at hand. At the completion of this stage, B cells undergo differentiation into antibody-secreting cells that transcribe large amounts of a shorter isoform of the *IGH* mRNA resulting in a secreted antibody form of the protein[22].

As outlined above, an important consequence of the VDJ-recombination process, that was captured in clonal selection theory of Burnet[7], and experimentally proven first by Nossal[23], is "one cell, one antibody", i.e., each B cell carries only a single functional and unique immunoglobulin molecule. This theory together with the process of CSR also indicates that, within a single antibody-secreting plasma cell, only one Ig isotype should be present. Advances in single cell RNA sequencing methods now provide the opportunity to validate this classic immunological concept. Here we show that when using the GRCh38 (referred to as hg38) as the genome reference, the results seem to violate the "one cell, one antibody" dogma. However, investigating this phenomenon further, we found that after switching to the newly published T2T-CHM13 (T2T) reference genome, the results agree with the "one cell, one antibody" theory perfectly. Given the essential roles of the immunoglobulin isotypes in human health, and the importance of mapping the functionality B cells in providing insights into the immune responses to infection, vaccination, allergy and in autoimmune conditions, it is necessary to identify them accurately in RNA sequencing data. These study highlights the importance of considering the reference genome in achieving this goal.

Results

Mapping performance is comparable between hg38 and T2T

To compare the performance of hg38 and T2T in isotype identification, two human 10x Genomics Chromium 3' scRNAseq datasets were selected, one generated in-house from enriched bone marrow plasma cells (BMPC) and the other from intestine mucosae published in Martin et al[24]. Two supplementary 10x scRNAseq datasets from King et al[25] (Tonsil samples) and Oetjen et al[26] (Bone marrow samples) were also interrogated to demonstrate the reproducibility of the results. Custom and comparable Cell Ranger references were constructed from the hg38 and T2T genomes, then cell by gene matrices of read counts were obtained from Cell Ranger. Examining the alignment output BAM files, mapping performances of hg38 and T2T at the read level were similar, with approximately the same numbers of effectively mapped sequence read numbers and PCR duplicates (Fig 1A and Fig S2A). After the read alignment and cell demultiplexing, quality control was performed on all single cells, after which, bona fide antibody-secreting cells (ASCs) were selected based on Ig expression and only cells appearing in both hg38 and T2T outputs were retained (Fig S2B). Checking the read numbers and the gene numbers in cells, the mapping performances at cellular level were also comparable between hg38 and T2T (Fig 1B and Fig S3A). In this study, ASCs are the target cells for isotype detection. Expression levels of four key marker genes, JCHAIN, XBP1, IRF4, and PRDM1, were similar between the hg38 and T2T outputs in all datasets (Fig 1C and Fig S3B). Taken together, the general qualities of single plasma cells have few differences between the hg38 and T2T reference genomes.

T2T genome identifies immunoglobulin isotypes more correctly

Given that the general qualities are similar, we next checked the character of isotype gene expression. Visualising the IgG isotype gene (*IGHG1*, *IGHG2*, *IGHG3*, and *IGHG4*) expression of the in house bone marrow plasma cell dataset in UMAP plots (Fig 2A), most of

the cells in T2T output only expressed one isotype gene, while in the hg38 output many cells expressed two isotypes simultaneously. Looking at the scatter plots of the expression levels of any two IgG isotype genes, many *IGHG1*$^+$*IGHG2*$^+$ and *IGHG1*$^+$*IGHG3*$^+$ double positive cells were found in the hg38 output, but few in the T2T output (Fig 2B and Fig S4). Similarly, when examining the IgA isotype gene (*IGHA1* and *IGHA2*) expression of the antibody-secreting cell subset of the intestine mucosa dataset, around half of the cells were *IGHA1*$^+$*IGHA2*$^+$ double positive cells in the hg38 output while virtually all cells were *IGHA1* or *IGHA2* single positive cells in the T2T output (Fig 2C and 3D, also see Fig S5 for another datasets). Furthermore, in the UMAP plot of hg38 output, *IGHA1*$^+$*IGHA2*$^+$ cells clustered together with IGHA1 single positive cells, while *IGHA1* and *IGHA2* single positive cells clearly separated in the T2T output (Fig 2C). Combining the "one cell, one antibody" theory and the fact that fewer double isotype positive cells in the T2T output, we conclude that the T2T genome identifies isotypes more correctly than the hg38 reference.

IgG double positive cells of hg38 output should be single positive

To explore why T2T genome performs better in isotype identification, we examined first the *IGHG1*$^+$*IGHG2*$^+$ cells. As an illustrative example, the cell marked in red in Fig 3A shows high expression for both *IGHG1* and *IGHG2* in the hg38 output but expresses only *IGHG2* in the T2T output (Fig 3A). Looking at the read alignment track plots (Fig 3B), reads were mapped to the 3' end side of the last exon of *IGHG1* (read coverage plot is in blue) and the more 5' region of the last exon of *IGHG2* (red coverage plot is in red) in hg38 output. In T2T output, reads (whose coverage plot is in red) in the more 5' region of the final exon of *IGHG2* in hg38 output remained in the same location of T2T genome, with the same coverage distribution (the coverage plot shapes are similar), but the reads (whose coverage plot is in blue) in the 3' end of *IGHG1* in hg38 output were remapped to the 3' end of *IGHG2* in the T2T output with the same coverage distribution. Therefore, the *IGHG1*$^+$*IGHG2*$^+$ cells were caused by the 3' end regions of *IGHG1* and *IGHG2*. Comparing the consensus sequence of reads around the 3' end region with the sequence of the 3' end regions of *IGHG1* and *IGHG2* from both hg38 and T2T, the numbers of mismatches (red bases) were 2, 1, 0, 7, for *IGHG2*-hg38, *IGHG1*-hg38, *IGHG2*-T2T, and *IGHG1*-T2T, respectively (Fig 3C). Within hg38 output, *IGHG1* had fewer mismatches than *IGHG2*, thus those reads were mapped to *IGHG1* causing the double positive cells, however, in T2T output, it was the *IGHG2* that had 0 mismatch, hence reads were mapped to *IGHG2* like the reads that were mapped the more 5' region (Fig 3C). Taking these together, the *IGHG1*$^+$*IGHG2*$^+$ cells in the hg38 output should be *IGHG2* single positive cells, because the reads in the 3' end region from *IGHG2* single positive cells in the real world are mapped to *IGHG1* incorrectly when using hg38 as genome reference.

Using a similar checking process, it is apparent that *IGHG1*$^+$*IGHG3*$^+$ double positive cells in hg38 output should be *IGHG1* single positive cell (Fig 4 A-C).

T2T reference results in more reads that are mapped to IGHG genes than hg38

In addition to higher accuracy in isotype identification, the T2T output also has more reads that are mapped to IGHG genes. Checking the BAM files, there were more effective reads and PCR duplicates mapped to IGHG genes in T2T output than hg38 output because a considerable number of reads were discarded (Fig 5A and Fig S6A). At single cell level, many cells in the T2T output gained more IGHG gene expression than hg38 (Fig 5B and Fig S6B). Rechecking the *IGHG1*$^+$ (red cell) of T2T output indicated in Fig 4, in addition to reads that were also mapped to *IGHG1* or *IGHG3* in hg38 output (Fig 4B, the right two track plots), more extra reads around the *IGHG1* stop codon were recovered (Fig 5C). Similarly, for the *IGHG2*$^+$ (red

cell) of T2T output indicated in Fig 3, in addition to reads that were also mapped to *IGHG2* or *IGHG1* in hg38 output (Fig 3B, the right two track plots), more extra reads around *IGHG2* stop codon region were recovered (Fig 5D). Tracking those extra recovered reads, most of them belonged to discarded reads in hg38 output (Fig 5A). Those reads were discarded in the hg38 output because they were mapped to regions whose sequences were exactly shared by *IGHG1* and *IGHG2* resulting in being discarded due to the same mapping quality. In contrast, sequences of T2T, at the same location, were distinguishable between *IGHG1* and *IGHG2* so reads could be mapped to *IGHG1* (Fig 5E) or *IGHG2* (Fig 5F) without ambiguity.

T2T output has fewer reads that are mapped to IGHA genes but more distinguishable *IGHA1* and *IGHA2* subclasses

In contrast to IGHG genes, there were fewer effective reads that were mapped to IGHA genes and fewer total IGHA gene expression in single cells in the T2T output compared to the hg38 output (Fig 6A and 6B and Fig S7A and S7B). The decreased effective reads were largely represented in fewer *IGHA1* expression in single cells (Fig 6C and Fig S7C), but for *IGHA2*, the expression level was higher in the T2T output than the hg38 output (Fig 6D and Fig S7C). For those double positive cells in the hg38 output, they should be *IGHA2* single positive (Fig 6E). Taking the red cell in Fig 5E as an example, in the hg38 output, reads mapped to *IGHA1* (the track plot in blue of Fig 6F and the blue region of Fig 6G) were discarded in the T2T output and reads mapped to *IGHA2* were still mapped to *IGHA2* in T2T at the same location (the track plot in red of Fig 6F and the red region of Fig 6G), so the double positive cells became the single positive cells (Fig 6F). Like *IGHG1* and *IGHG2* (Fig 5), there were also extra reads for *IGHA2* in the T2T output retrieved from discarded reads in the hg38 output (the track plot in orange of Fig 6F and the orange region of Fig 6G). Aligning the sequences of the last exons of *IGHA1* and *IGHA2* from both hg38 and T2T and summarising the characteristics, the last exons of *IGHA1* or *IGHA2* could be divided into 3 regions (Fig 6G). For the "blue" region of hg38, the sequences were distinguishable incorrectly between *IGHA1* and *IGHA2*, hence in the hg38 output, IGHA gene reads in this region were all mapped to *IGHA1*, regardless those reads belong to $IGHA1^+$ cells or $IGHA2^+$ cells; while for the "blue" region of T2T, the sequences were exactly the same between *IGHA1* and *IGHA2*, so all reads in this region were discarded due to multi-mapping (Fig 6F, the left), causing reduced effective reads for IGHA genes (Fig 6A, 6B and 6C). For the "orange" region of hg38, the sequences were shared between *IGHA1* and *IGHA2*, so reads in this region were discarded; while for T2T, sequences were distinguishable, thus more reads were recovered in this region (Fig 6F, the middle), evaluating the *IGHA2* expression level largely (Fig 6D). For the "red" region, it was distinguishable in both hg38 and T2T and reads in this region were mapped to the corresponding genes correctly in both hg38 and T2T (Fig 6F, the right). Taking these together, in the hg38 output, all *IGHA2* single positive cells in the real world become $IGHA1^+IGHA2^+$ cells; reads in the blue region were mapped to *IGHA1* making them become $IGHA1^+$ in addition to their true *IGHA2*+ identity, also making the higher *IGHA1* expression. while for the T2T output, the reads in the blue region were all discarded resulting in reduced *IGHA1* expression, but removal of artificial double positive cells. Furthermore, T2T has one more distinguishable region, the orange region, allowing the higher *IGHA2* expression to be determined.

Discussion

"One cell one antibody" is a cornerstone of the clonal selection theory that has been central to our understanding of the immune system for over 60 years[27]. The fact that the T2T output fits this theory, while the widely used hg38 output does not, prompted us to explore the underlying

cause of this uncertainty. We conclude that the T2T output aligns best with the "one cell one antibody" dogma and that sequences of the *IGH* constant regions provided in the T2T reference genome are more accurate than hg38. Thus, T2T is the genome of choice for accurate mapping of the immunoglobulin loci.

T2T is the first published complete human genome employing the long read sequencing technology[5]. T2T is sequenced from a homozygous complete hydatidiform mole (CHM) cell line, representing a complete haplotype[5]. In comparison, the current hg38 genome is derived from multiple individuals[4]. The sequences of the subisotypes of IgG and IgA are highly similar with each other, especially for the most 3' exons, thus the accuracy of the genome references in these regions is of great importance in identifying the correct subisotype. The combination of the single individual origin and the long-read sequencing technology that can span longer fragments and cover more distinguishable points between orthologous genes results in more correct sequences for isotype-related gene regions.

Considering the importance of immunoglobulin isotype in immune regulation and diseases, it is necessary to identify the isotypes correctly. Based on this study, T2T genome reference is a better option than hg38. Nevertheless, when analysing scRNAseq data, using hg38 is still feasible for isotype identification. In those IgG double positive cells, large proportion of isotype reads are mapped correctly and those reads mapped to the false isotype represented a smaller percentage. We also found that the normalised expression level of the "error isotype" is also lower than the "correct isotype". Therefore, the error isotype can be ignored manually for IgG double positive cells. For IgA double positive cells, even though the error *IGHA1* expression is always higher than the correct *IGHA2* expression, the double positive cells should always be *IGHA2* single positive in the real world and T2T output. For this study, supported by 4 independent datasets and more than 10 individuals, *IGHG1* and *IGHG2* double positive cells should be *IGHG2* single positive; *IGHG1* and *IGHG3* double positive cells should be *IGHG1* single positive; and *IGHA1* and *IGHA2* should be *IGHA2* single positive. Though manual correction is possible, using T2T is highly recommended for two reasons. Firstly, T2T it provides the correct results directly and the secondly more discarded reads of hg38 output can be retrieved and thus higher gene expression for *IGHG1*, *IGHG2*, and *IGHA2* is registered.

For bulk RNAseq, T2T reference is also the optimum choice if isotype genes need to be quantified precisely, as manual correction of the hg38 output based on the cellular origin is not possible. The manual correction at the single cell level is based on the precondition that there are usually more correct isotype reads in single cells, thus those "error reads" in those double positive cells can be removed manually, but for bulk RNAseq, the cellular origin of reads cannot be traced since all cells are mixed and processed together, therefore error reads cannot be found and removed. The sole approach to identify error reads for bulk RNAseq data is to localise the coordinates of reads. In this study, the exact regions that cause the double positive cells have been located. In hg38 output, reads mapped to the 3' end of the last exon of *IGHG1* should belong to IGHG2; reads mapped to the 3' end of the last the last exon of *IGHG3* should belong to IGHG1. Reads mapped to the last exon of *IGHA1* or *IGHA2* should also be adjusted based on the conclusion of Fig 6G. Performing such correction would be highly tedious, compared to the simplicity of using the T2T genome reference.

Conclusions

In summary, we conclude that for analysis of immunoglobulin isotype-determining genes, the sequences of T2T genome reference are more accurate than GRCh38 (hg38). It is noteworthy, that this study was solely focused on immunoglobulin constant region family. Whether other

highly related gene families also present similar mapping problems warrants further investigation.

Methods

Public data pre-processing

10x-format FASTQ files were generated for each of the four datasets. The bone marrow plasma cell (BMPC) data was generated in-house using Cell Ranger 5.0.0. Metadata is available from GEO accession GSM7181915 and FASTQ files from SRA accession SRX20001597.

Metadata for the intestine mucosae dataset is available from GEO accession GSM3972026 and FASTQ files were downloaded from SRA accession SRX6584269.

Metadata for the tonsil dataset is available from GEO accession GSM5051497 and a BAM file containing sequence reads was downloaded from SRA accession SRX9986858. This BAM file was converted to FASTQ files using the Cell Ranger 7.0.0 utility bamtofastq.

Metadata from the bone marrow dataset is available from GEO accession GSM3396184 and a BAM file was downloaded from SRA accession SRX4720060. The BAM file was converted to FASTQ files using the Cell Ranger 7.0.0 utility bamtofastq.

Cellranger references

To make a reference genome for cellranger, a FASTA file of genomic sequences and a GTF file of gene annotation is required. For GRCh38, the FASTA file was GRCh38.primary_assembly.genome.fa and the backbone GTF file was gencode.v35.annotation.gtf, both downloaded from https://www.gencodegenes.org. For T2T, the FASTA file was chm13v2.0_maskedY.fa and the backbone annotation was UCSC GENCODEv35 CAT/Liftoff v2, downloaded from https://github.com/marbl/CHM13.

Only protein coding genes are kept to make the results more reproducible between different annotations[28]. In addition, to make the results comparable between GRCh38 and T2T, only genes that appear in both the GRCh38 and T2T GTF annotations are retained. Genes were matched by symbol and rtracklayer[29] was used to read the GTF and help with subsetting.

Finally, the cellranger references are made by mkref of cellranger 7.0.0, with the FASTA files and filtered GTF files as inputs.

Cellranger mapping and quantification

Cellranger count (7.0.0) is used to get the gene x barcode matrix. All FASTQ files are mapped to GRCh38 and T2T separately. For each dataset, there are two outputs, the GRCh38 (hg38) one and the T2T one.

To generate Figure 1A, Figure 5A, and Figure 6A, bam files generated by cellranger count are used to check the general mapping performance. Information of xf tag and GN tag is extracted by samtools[30], in which, GN represents which gene a read is assigned to and xf represents the kind of a read (effective read, PCR duplicate, or discarded).

Single cell quality control and Seurat pipeline

The matrix outputs from cellranger count were further processed following the general Seurat[31] pipeline, in which quality control, normalisation, and dimension reduction is performed. For quality control, cells that have >10% mitochondrial gene expression were removed; ribosomal genes were removed; genes that are expressed in fewer than 0.5% of total cells were removed;

and cells with too few reads (usually 1500 as threshold) or expressed genes (usually 450 as threshold) were removed. Finally, only cells that have >10% immunoglobulin gene expression and appear in both hg38 and T2T output are defined as antibody-secreting cells for downstream analysis and for generating Figures 1BC, 2, 3A, 4A, 5B and 6BCDE.

Alignment track plot visualisation

To generate Figure 3B, Figure 4B, Figure 5 C and B, and Figure 6F, bam files of particular genes and cells are generated by samtools based on the read names, CR tag (representing cell barcodes), GN tag, and xf tag (only effective reads and PCR duplicate reads are included). Alignment track plots are produced with those specific bam files with R Gviz package[32].

Sequencing alignment

A Plasmid Editor (ApE) software[33] (https://jorgensen.biology.utah.edu/wayned/ape) was used to generate Fig 3C, Fig 4C, Fig 5E and Fig 5F.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets analysed during the current study are available from the NCBI Sequence Read Archive experiments https://www.ncbi.nlm.nih.gov/sra/SRX20001597, https://www.ncbi.nlm.nih.gov/sra/SRX6584269, https://www.ncbi.nlm.nih.gov/sra/SRX9986858 and https://www.ncbi.nlm.nih.gov/sra/SRX4720060.

Competing interests
The authors declare they have no competing interests.

Funding
This work was supported by the National Health and Medical Research Council Australia (1155342, 1160830 to S.L.N; 1058892 to G.K.S; IRIISS to WEHI), Cancer Council Victoria (Grant-in-Aid to J.T. and S.L.N.), the Chan Zuckerberg Initiative (EOSS4 grant number 2021-237445 to G.K.S.), the University of Melbourne (Melbourne Research Scholar to J.N.) and by Victorian State Government Operational Infrastructure Support.

Authors' contributions
JN conceived the project, performed bioinformatics analyses and wrote the original draft of the manuscript. JT performed experiments, provided supervision and contributed to writing the manuscript. IT performed bioinformatics analyses. SLN and GKS supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
Matthew Robertson from 10X genomics provided technical support with regard to custom reference making and BAM file dissection.

References

1 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001). doi: 10.1126/science.1058040

2 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001). doi: 10.1038/35057062

3 Collins, F. S., Lander, E. S., Rogers, J., Waterston, R. H. & Conso, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004). doi: 10.1038/nature03001

4 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849-864 (2017). doi: 10.1101/gr.213611.116

5 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022). doi: 10.1126/science.abj6987

6 Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, abl3533 (2022). doi: 10.1126/science.abl3533

7 Burnet, F. M. A modification of Jerne's theory of antibody production using the concept of clonal selection (Reprinted from Australian Journal of Science, vol 20, 1957). *Nat Immunol* **8**, 1024-1026 (2007).

8 Vettermann, C. & Schlissel, M. S. Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunol Rev* **237**, 22-42 (2010). doi: 10.1111/j.1600-065X.2010.00935.x

9 Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000). doi: 10.1016/S0092-8674(00)00078-7

10 Jefferis, R. & Kumararatne, D. S. Selective Igg Subclass Deficiency - Quantification and Clinical Relevance. *Clin Exp Immunol* **81**, 357-367 (1990). doi: 10.1111/j.1365-2249.1990.tb05339.x

11 Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol* **5**, 520 (2014). doi: 10.3389/fimmu.2014.00520

12 Collins, A. M. & Jackson, K. J. L. A temporal model of human IgE and IgG antibody function. *Front Immunol* **4**, 235 (2013). doi: 10.3389/fimmu.2013.00235

13 Aalberse, R. C., Stapel, S. O., Schuurman, J. & Rispens, T. Immunoglobulin G4: an odd antibody. *Clin Exp Allergy* **39**, 469-477 (2009). doi: 10.1111/j.1365-2222.2009.03207.x

14 Steffen, U. *et al.* IgA subclasses have different effector functions associated with distinct glycosylation profiles. *Nat Commun* **11,** 120 (2020). doi: 10.1038/s41467-019-13992-8

15 Lai, K. N. *et al.* IgA nephropathy. *Nat Rev Dis Primers* **2**, 16001 (2016). doi: 10.1038/nrdp.2016.1

16 Hammarstrom, L. & Smith, C. I. E. IgG Subclasses in Bacterial-Infections. *Monogr Allergy* **19**, 122-133 (1986).

17 Skvaril, F. IgG Subclasses in Viral-Infections. *Monogr Allergy* **19**, 134-143 (1986).

18 Ferrante, A., Beard, L. J. & Feldman, R. G. Igg Subclass Distribution of Antibodies to Bacterial and Viral-Antigens. *Pediatr Infect Dis J* **9**, S16-S24 (1990).

19 Irrgang, P. *et al.* Class switch toward noninflammatory, spike-specific IgG4 antibodies after repeated SARS-CoV-2 mRNA vaccination. *Sci Immunol* **8**, eade2798 (2023). doi: 10.1126/sciimmunol.ade2798

20 Zhang, H. Z. *et al.* Serum IgG Subclasses in Autoimmune Diseases. *Medicine* **94**, e387 (2015). doi: 10.1097/MD.0000000000000387
21 Sharonov, G. V., Serebrovskaya, E. O., Yuzhakova, D. V., Britanova, O. V. & Chudakov, D. M. B cells, plasma cells and antibody repertoires in the tumour microenvironment. *Nat Rev Immunol* **20**, 294-307 (2020). doi: 10.1038/s41577-019-0257-x
22 Word, C. J., Mushinski, J. F. & Tucker, P. W. The Murine Immunoglobulin Alpha-Gene Expresses Multiple Transcripts from a Unique Membrane Exon. *Embo J* **2**, 887-898 (1983). doi: 10.1002/j.1460-2075.1983.tb01518.x
23 Nossal, G. J. V. & Lederberg, J. Antibody Production by Single Cells. *Nature* **181**, 1419-1420 (1958). doi: 10.1038/1811419a0
24 Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493-1508 (2019). doi: 10.1016/j.cell.2019.08.008
25 King, H. W. *et al.* Integrated single-cell transcriptomics and epigenomics reveals strong germinal center-associated etiology of autoimmune risk loci. *Science Immunology* **6**, eabh3768 (2021). doi: 10.1126/sciimmunol.abh3768
26 Oetjen, K. A. *et al.* Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *Jci Insight* **3**, e124928 (2018). doi: 10.1172/jci.insight.124928
27 Hodgkin, P. D., Heath, W. R. & Baxter, A. G. The clonal selection theory: 50 years since the revolution. *Nat Immunol* **8**, 1019-1023 (2007). doi: 10.1038/ni1007-1019
28 Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310-315 (2022). doi: 10.1038/s41586-022-04558-8
29 Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842 (2009). doi: 10.1093/bioinformatics/btp328
30 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). doi: 10.1093/bioinformatics/btp352
31 Hao, Y. H. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 (2021). doi: 10.1016/j.cell.2021.04.048
32 Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol* **1418**, 335-351 (2016). doi: 10.1007/978-1-4939-3578-9_16
33 Davis, M. W. & Jorgensen, E. M. ApE, A Plasmid Editor: A Freely Available DNA Manipulation and Visualization Program. *Front Bioinform* **2**, 818619 (2022). doi: 10.3389/fbinf.2022.818619

Figure Legends

Figure 1 Mapping performance is comparable between hg38 and T2T

A) Classification of reads for the bone marrow plasma cell (top) and intestinal mucosa (bottom) datasets for each reference genome. The barplots show the total read count for each sample and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons. Only effective reads were used for downstream analyses.

B) Reads per cell (left) and genes detected per cell (right) for the bone marrow plasma cell (top) and intestinal mucosa (bottom) datasets. Here and subsequently, only bona fide ASCs are shown. The hg38 and T2T cell-wise counts are compared by scatter plots and by side-by-side boxplots. Boxplots show median, quartiles, upper and lower hinges.

C) Cell-wise expression of plasma cell marker genes for the bone marrow plasma cell (left) and intestinal mucosa (right) datasets. Expression is represented as the log1p of reads per ten thousand (expression value = $\log1p(y/L*10^4)$ where y is the read count for the gene and L is total read count for the cell.). Hg38 and T2T expression values are compared by scatter plots and side-by-side boxplots.

Figure 2 The T2T genome identifies immunoglobulin isotypes more accurately

A) UMAP dimension reduction plots of individual cells (ASCs) for the BMPC dataset. In each plot, cells are coloured by expression level of the indicated gene for a particular IgG subclass. Plots are shown for hg38 (top) and T2T (bottom). Here expression values are represented the same as Fig 1C.

B) Paired scatter plots correlating the expression levels of pairs of IgG subclass genes for the BMPC ASCs. Plots are shown for hg38 (top) and T2T (bottom). Here expression is represented as reads per hundred (RPH=y/L*100 where y is the read count for the gene and L is total read count for the cell.)

C) UMAP dimension reduction plots of individual cells (ASCs) for the intestinal mucosa dataset. Cells are coloured by expression level of the indicated gene for a particular IgA subclass. Plots are shown for hg38 (top) and T2T (bottom). Here expression values are represented the same as Fig 1C.

D) Scatter plots correlating the expression levels of *IGHA1* and *IGHA2* for intestinal mucosa ASCs. Plots are shown for hg38 (top) and T2T (bottom). Here expression is represented as reads per hundred (RPH=y/L*100 where y is the read count for the gene and L is total read count for the cell.)

Figure 3 Separating *IGHG1*$^+$ and *IGHG2*$^+$ plasma cells

A) Scatter plots correlating *IGHG1* and *IGHG2* expression for BMPC ASCs. Left plot for hg38 and right plot for T2T. These plots are repeated from Figure 2B except that the cell with highest *IGHG2* expression is highlighted.

B) Coverage plots for reads aligned using hg38 from the highlighted cell in (A). Top left plot shows reads mapped to *IGHG1* using hg38 and bottom left shows reads mapped to *IGHG2* using hg38. Only the 3'-most exon of the genes are shown. The plots on the right show coverage plots for the same reads but using T2T. The top right plot shows that all the reads mapped to *IGHG1* using hg38 map instead to *IGHG2* using T2T. The bottom right plot shows that all reads mapped to *IGHG2* using hg38 map also map to *IGHG2* using T2T.

C) Detailed alignment results for one illustrative sequence read incorrectly assigned to *IGHG1* by hg38 in (B). The full sequence of the read is shown together with the best matches to the *IGHG2*-hg38, *IGHG1*-hg38, *IGHG2*-T2T and *IGHG1*-T2T genome regions. Mismatched bases are marked in red. The read maps perfectly to *IGHG2*-T2T and, using T2T, the

assignment to *IGHG2* is unambiguous. Using hg38, however, the result is reversed with fewer mismatches for *IGHG1*-hg38 than for *IGHG2*-hg38. The coverage plot merges the right two plots from (B) on the same scale, i.e., shows *IGHG2*-T2T coverage for all reads mapped by hg38 to either *IGHG1* or *IGHG2*. The aligned position of the illustrative read is shown by a red bar.

Figure 4 Separating *IGHG1*$^+$ and *IGHG3*$^+$ plasma cells

A) Scatter plots correlating *IGHG1* and *IGHG3* expression for bone marrow ASCs. Left plot for hg38 and right plot for T2T. These plots are repeated from Figure 2B except that the cell with highest *IGHG1* expression is highlighted.

B) The plots on the left show coverage plots for reads aligned using hg38 from the highlighted cell in (A). Top left plot shows reads mapped to *IGHG3* using hg38 and bottom left shows reads mapped to *IGHG1* using hg38. Only the 3'-most exon of the genes are shown. The plots on the right show coverage plots for the same reads but using T2T. The top right plot shows that all the reads mapped to *IGHG3* using hg38 map instead to *IGHG1* using T2T. The bottom right plot shows that all reads mapped to *IGHG1* using hg38 map also map to *IGHG1* using T2T.

C) Detailed alignment results for one illustrative sequence read incorrectly assigned to *IGHG3* by hg38 in (B). The full sequence of the read is shown together with the best matches to the *IGHG1*-hg38, *IGHG3*-hg38, *IGHG1*-T2T and *IGHG3*-T2T genome regions. Mismatched bases are marked in red. The read maps perfectly to *IGHG1*-T2T and, using T2T, the assignment to *IGHG1* is unambiguous. Using hg38, however, the result is reversed with fewer mismatches for *IGHG3*-hg38 than for *IGHG1*-hg38. The coverage plot merges the right two plots from (B) on the same scale, i.e., shows *IGHG1*-T2T coverage for all reads mapped by hg38 to either *IGHG1* or *IGHG3*. The aligned position of the illustrative read is shown by a red bar.

Figure 5 T2T reference results in more reads that are mapped to IGHG genes than hg38

A) Classification of reads that are assigned to IGHG genes for the BMPC dataset for each reference genome. The barplots show the total read count of IGHG genes and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons.

B) Percentage of IGHG gene reads per cell for the BMPC dataset. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

C) Top, the coverage plot from Fig 4C with the same scale as the bottom coverage plot; bottom, the coverage plot of reads that map to *IGHG1* perfectly using T2T but with ambiguity when using hg38. The small red bar denotes the location of the stop codon.

D) Top, the coverage plot from Fig 3C with the same scale as the bottom track plot; bottom, the coverage plot of reads that map to *IGHG2* perfectly using T2T but with ambiguity when using hg38. The small red bar denotes the location of the stop codon.

E) Detailed alignment results for one illustrative sequence read with assignment ambiguity between *IGHG1* and *IGHG2* when using hg38 in (C). The full sequence of the read is shown together with the best matches to the *IGHG1*-hg38, *IGHG2*-hg38, *IGHG1*-T2T and *IGHG2*-T2T genome regions. Mismatched bases are marked in red. The read maps perfectly to *IGHG1*-T2T and, using T2T, the assignment to *IGHG1* is unambiguous. Using hg38, however, the result is ambiguous with the same number of mismatches between *IGHG1*-hg38 and *IGHG2*-hg38. The coverage plot is from the bottom plot of (C).

F) Detailed alignment results for one illustrative sequence read with assignment ambiguity between *IGHG1* and *IGHG2* when using hg38 in (D). The full sequence of the read is shown together with the best matches to the *IGHG2*-hg38, *IGHG1*-hg38, *IGHG2*-T2T and *IGHG1*-

T2T genome regions. Mismatched bases are marked in red. The read maps perfectly to *IGHG2*-T2T and, using T2T, the assignment to *IGHG2* is unambiguous. Using hg38, however, the result is ambiguous with the same number of mismatches between *IGHG2*-hg38 and *IGHG1*-hg38. The coverage plot is from the bottom plot of (D).

Figure 6 T2T output has fewer reads that are mapped to IGHA genes but better distinguishes *IGHA1* from *IGHA2*

A) Classification of reads that are assigned to IGHA genes for the intestinal mucosa (intestine) dataset for each reference genome. The barplots show the total read count of IGHA genes and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons.

B) Percentage of IGHA gene reads per cell for the intestine dataset. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

C) Percentage of *IGHA1* reads per cell for the intestine dataset. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

D) Percentage of *IGHA2* reads per cell for the intestine dataset. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

E) Scatter plots correlating expression levels of *IGHA1* and *IGHA2* using hg38 (left) and T2T (right). The cell with highest *IGHA2* expression is highlighted in red. These plots are repeated from Figure 2D except that the cell with highest *IGHA2* expression is highlighted.

F) The plots on the top show coverage plots for reads aligned using hg38 from the highlighted cell in (E). Top left plot shows reads mapped to *IGHA1* using hg38 (these reads are discarded when using T2T because of mapping ambiguity) and top right shows reads mapped to *IGHA2* using hg38 (these reads also map *IGHA2* using T2T). Only the 3'-most exon of the genes are shown. The plots on the bottom show coverage plots for reads aligned using T2T. The bottom middle plot shows reads that map *IGHA2* perfectly, and using T2T, the assignment to *IGHA2* is unambiguous (these reads are discarded when using hg38 due to mapping ambiguity). The bottom right plot shows that all reads mapped to *IGHA2* using hg38 map also map to *IGHA2* using T2T.

G) A schematic plot of the 3'-most exon of *IGHA1* or *IGHA2*. The top text explains the result of sequence alignment between *IGHA1* and *IGHA2* for the T2T or hg38 genomes. The bottom text explains the read mapping outcome when reads mapped to the indicated regions with specific colours. The colours used here have the same meaning as the coverage plots of (F), i.e., the reads of coverage plots from (F) with one colour locate on the region in the same colour.

Figure 1

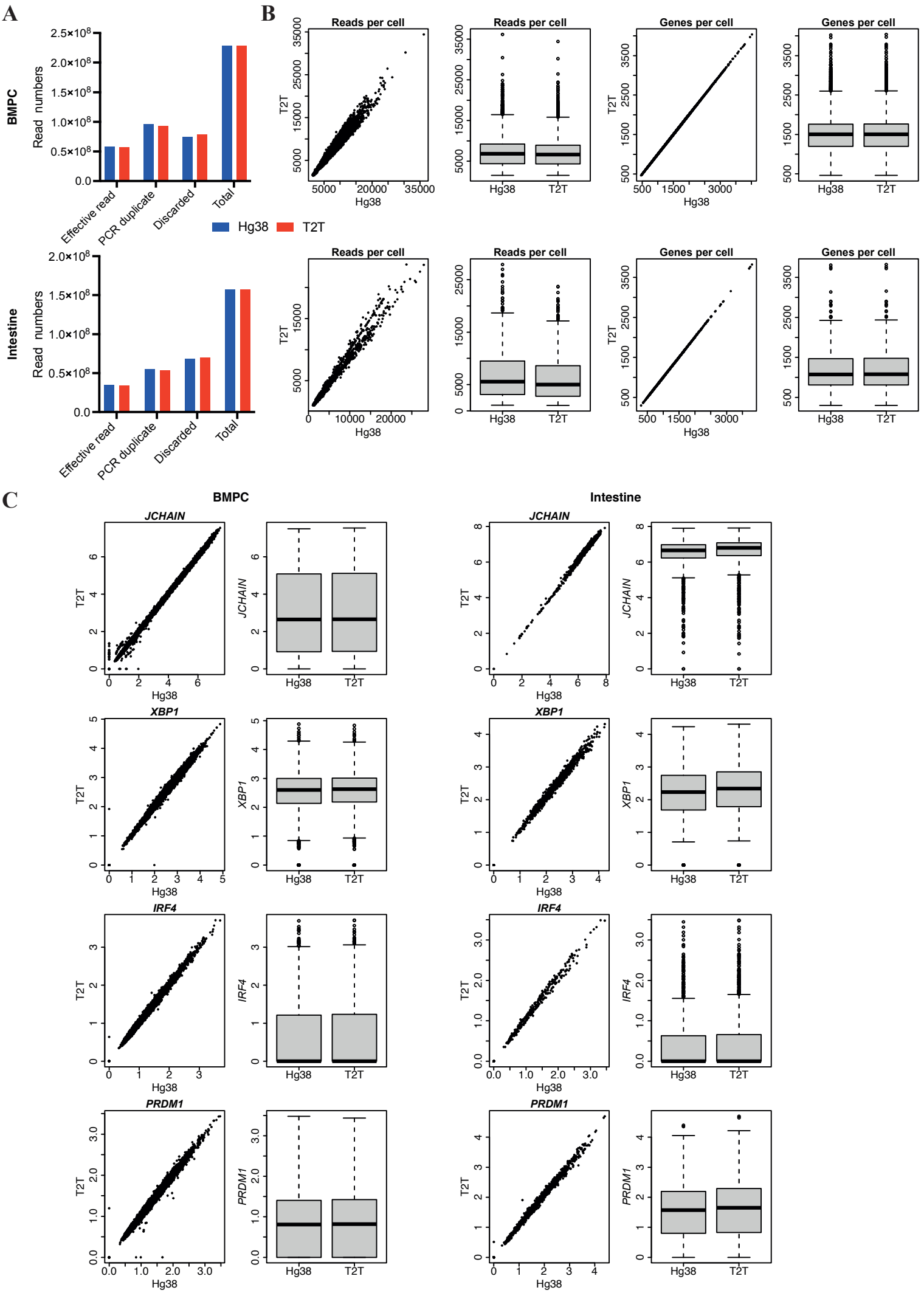

A
B
C
BMPC
Intestine
Read numbers
Effective read
PCR duplicate
Discarded
Total
Hg38
T2T
Reads per cell
Genes per cell
JCHAIN
XBP1
IRF4
PRDM1

Figure 2

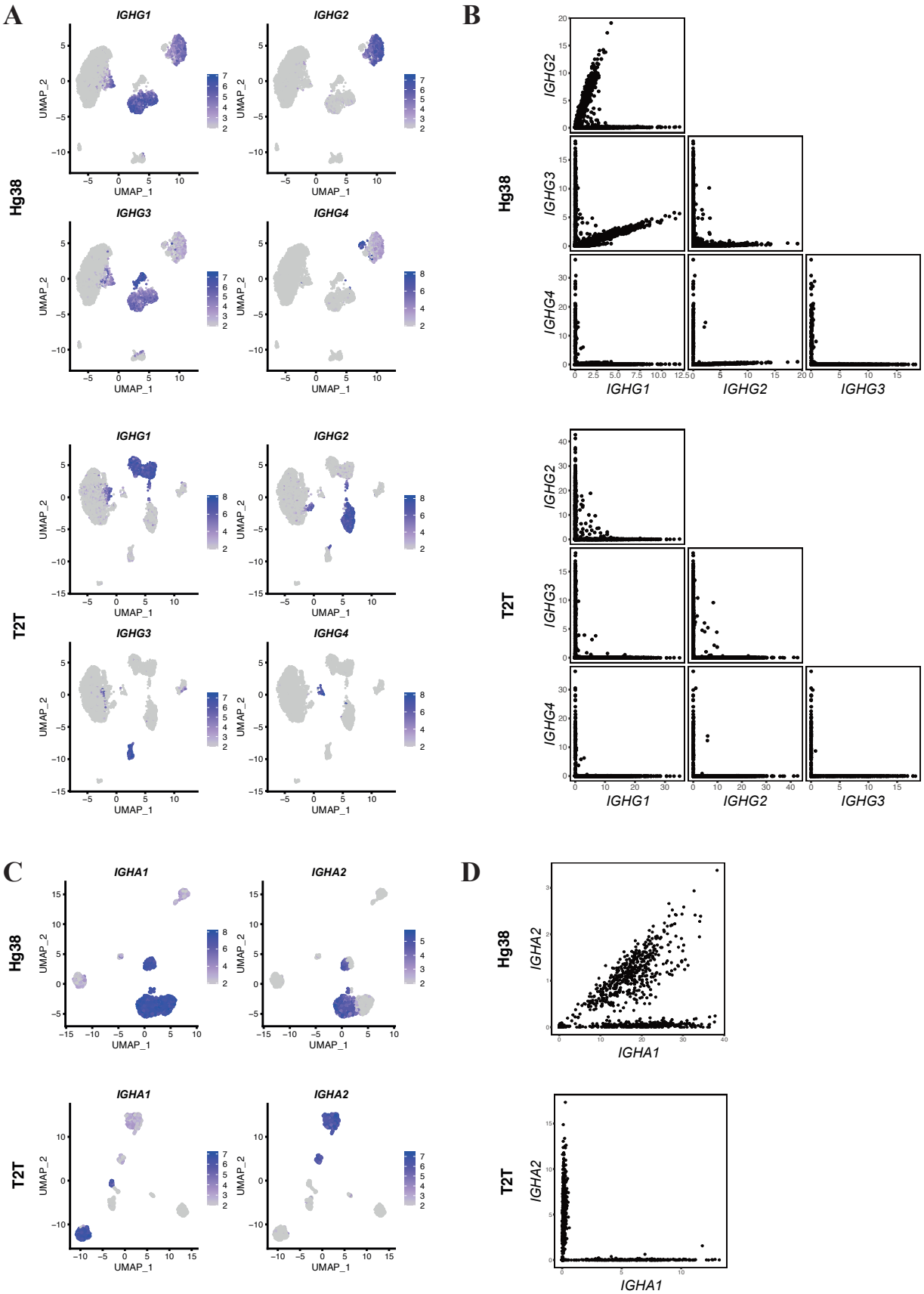
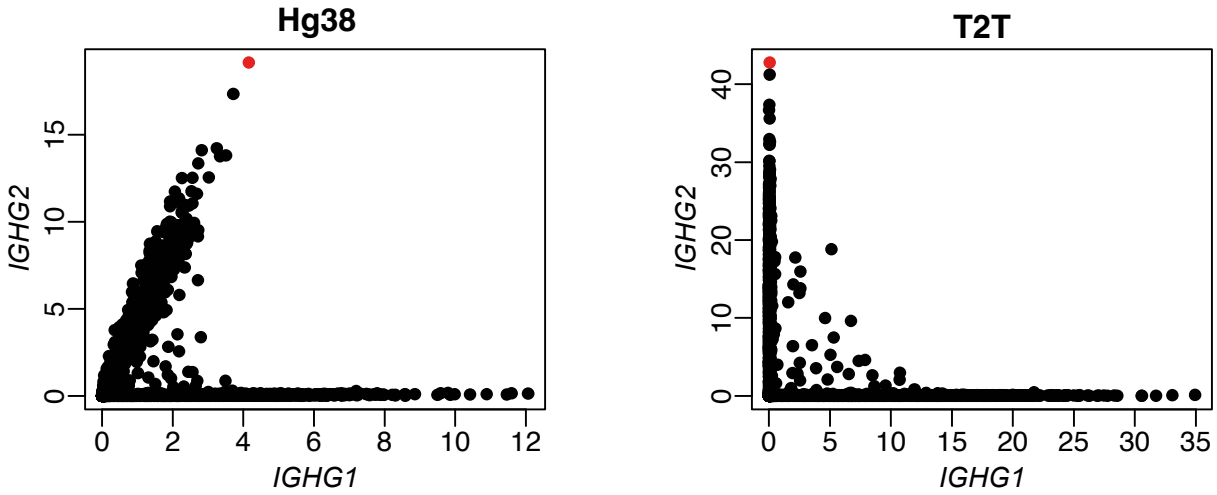
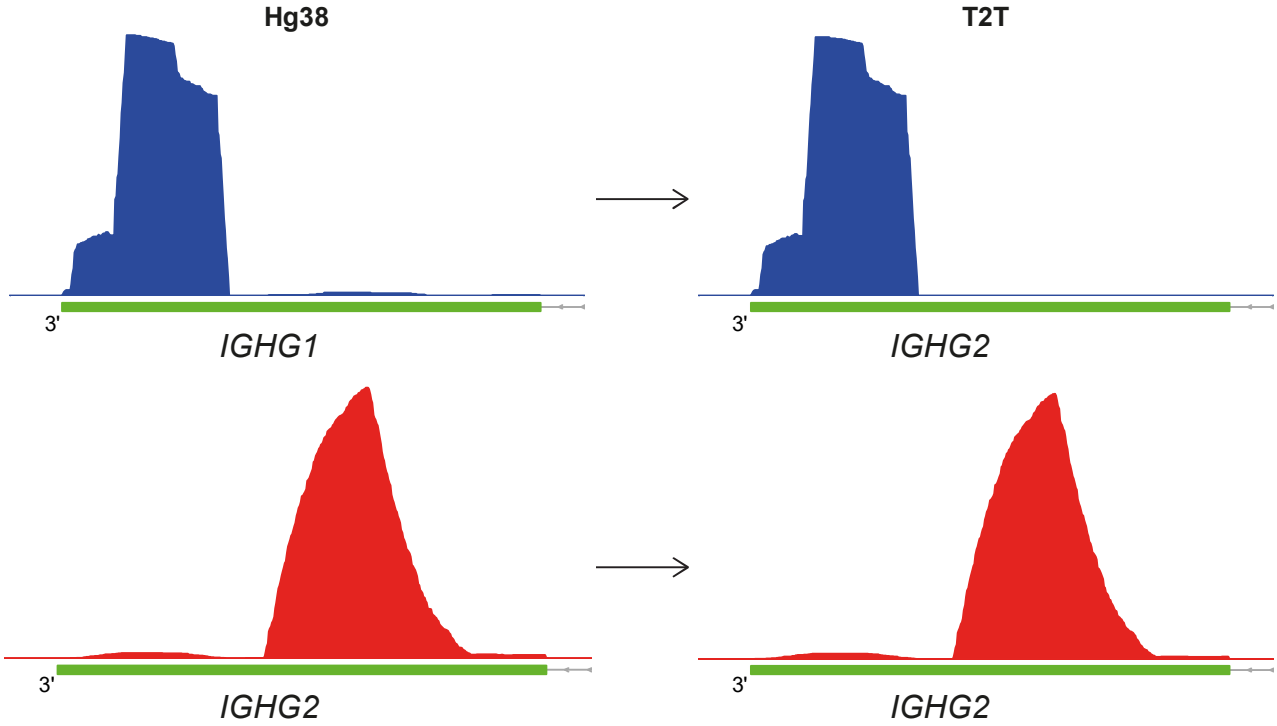

A
B
C
D
Hg38
T2T
IGHG1
IGHG2
IGHG3
IGHG4
IGHA1
IGHA2
UMAP_1
UMAP_2

Figure 3

A


Hg38
T2T
IGHG2
IGHG1
0
5
10
15
0
2
4
6
8
10
12
0
10
20
30
40
0
5
10
15
20
25
30
35

B


Hg38
T2T
3'
IGHG1
3'
IGHG2
3'
IGHG2
3'
IGHG2

C

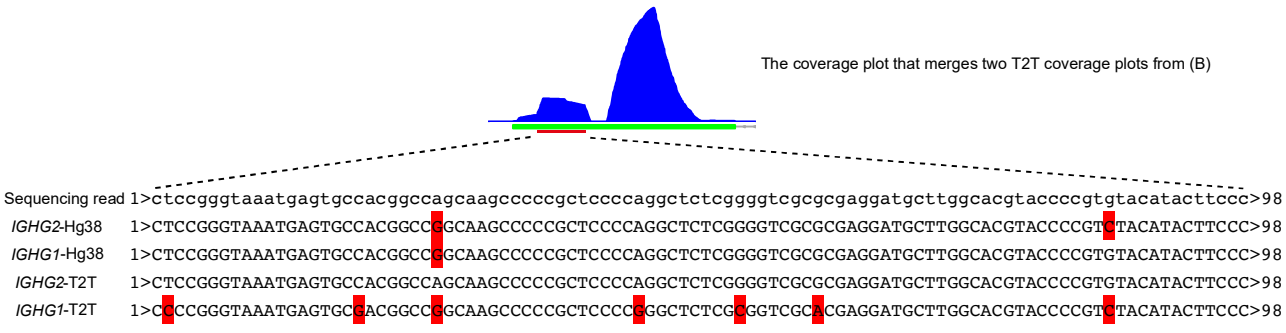
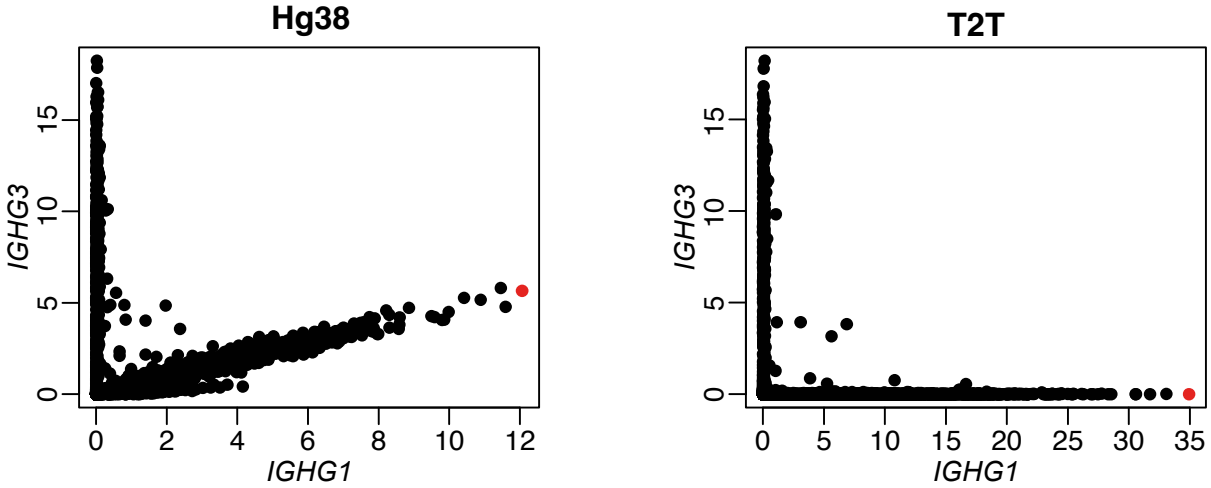
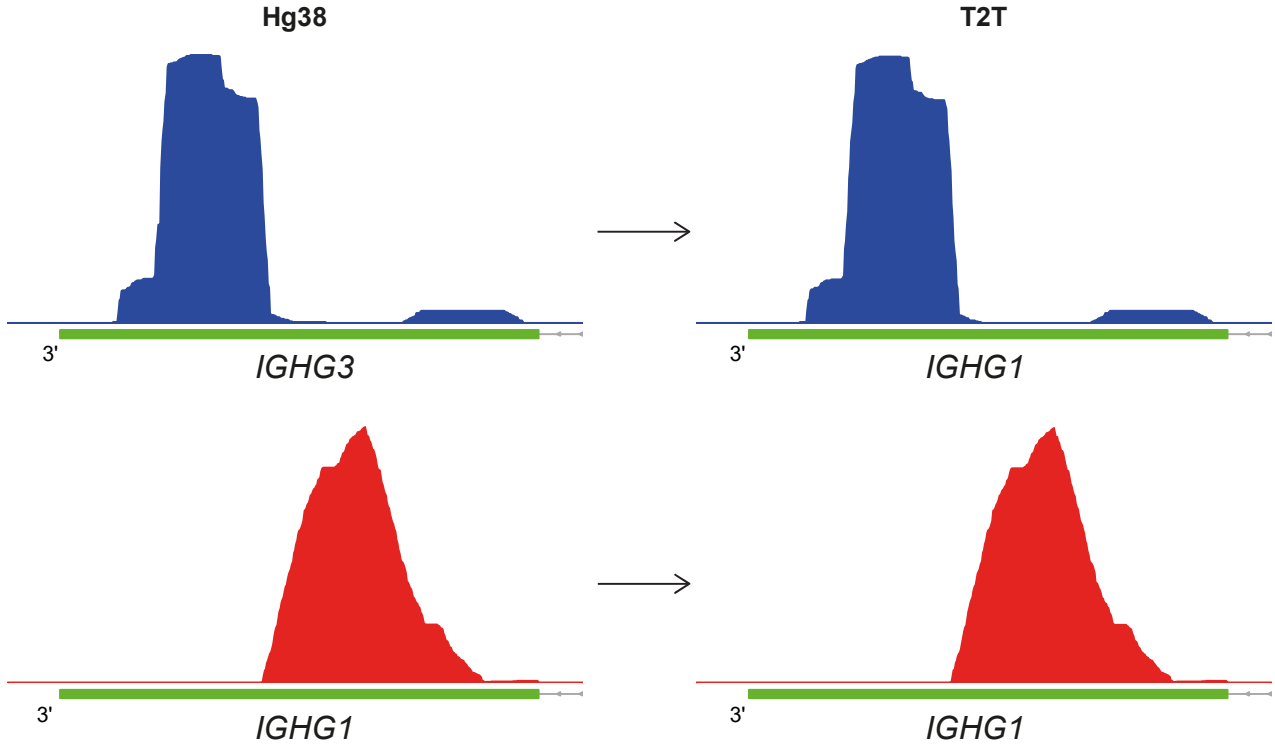

The coverage plot that merges two T2T coverage plots from (B)
Sequencing read 1>ctccgggtaaatgagtgccacggccagcaagcccccgctccccaggctctcggggtcgcgcgaggatgcttggcacgtaccccgtgtacatacttccc>98
IGHG2-Hg38 1>CTCCGGGTAAATGAGTGCCACGGCCGGCAAGCCCCCGCTCCCCAGGCTCTCGGGGTCGCGCGAGGATGCTTGGCACGTACCCCGTGTACATACTTCCC>98
IGHG1-Hg38 1>CTCCGGGTAAATGAGTGCCACGGCCGGCAAGCCCCCGCTCCCCAGGCTCTCGGGGTCGCGCGAGGATGCTTGGCACGTACCCCGTGTACATACTTCCC>98
IGHG2-T2T 1>CTCCGGGTAAATGAGTGCCACGGCCAGCAAGCCCCCGCTCCCCAGGCTCTCGGGGTCGCGCGAGGATGCTTGGCACGTACCCCGTGTACATACTTCCC>98
IGHG1-T2T 1>CCCCGGGTAAATGAGTGCGACGGCCGGCAAGCCCCCGCTCCCCGGGCTCTCGCGGTCGCACGAGGATGCTTGGCACGTACCCCGTGTACATACTTCCC>98

Figure 4

A


Hg38
T2T
IGHG3
IGHG1
0
5
10
15
0
2
4
6
8
10
12
0
5
10
15
20
25
30
35

B


Hg38
T2T
3'
IGHG3
IGHG1
3'
IGHG1
3'
IGHG1
3'
IGHG1

C

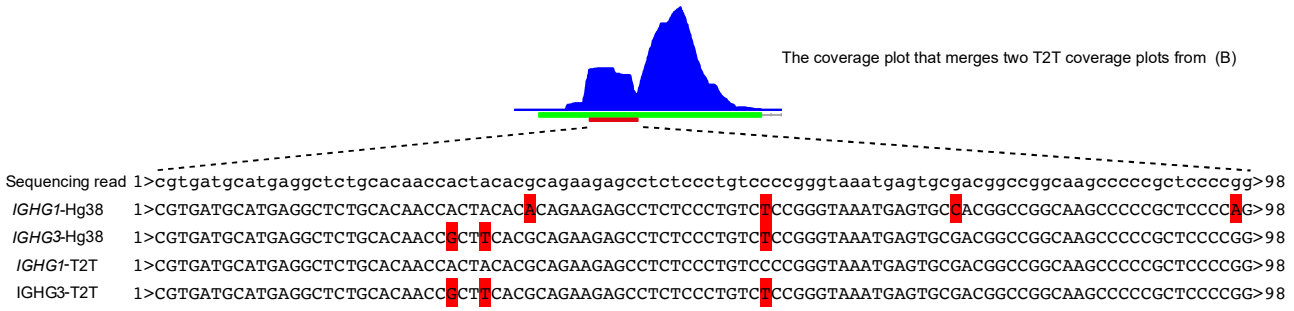

The coverage plot that merges two T2T coverage plots from (B)
Sequencing read 1>cgtgatgcatgaggctctgcacaaccactacacgcagaagagcctctccctgtccccgggtaaatgagtgcgacggccggcaagcccccgctccccgg>98
IGHG1-Hg38 1>CGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCACGGCCGGCAAGCCCCCGCTCCCCAG>98
IGHG3-Hg38 1>CGTGATGCATGAGGCTCTGCACAACCGCTTCACGCAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCGACGGCCGGCAAGCCCCCGCTCCCCGG>98
IGHG1-T2T 1>CGTGATGCATGAGGCTCTGCACAACCACTACACGCAGAAGAGCCTCTCCCTGTCCCCGGGTAAATGAGTGCGACGGCCGGCAAGCCCCCGCTCCCCGG>98
IGHG3-T2T 1>CGTGATGCATGAGGCTCTGCACAACCGCTTCACGCAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCGACGGCCGGCAAGCCCCCGCTCCCCGG>98

Figure 5

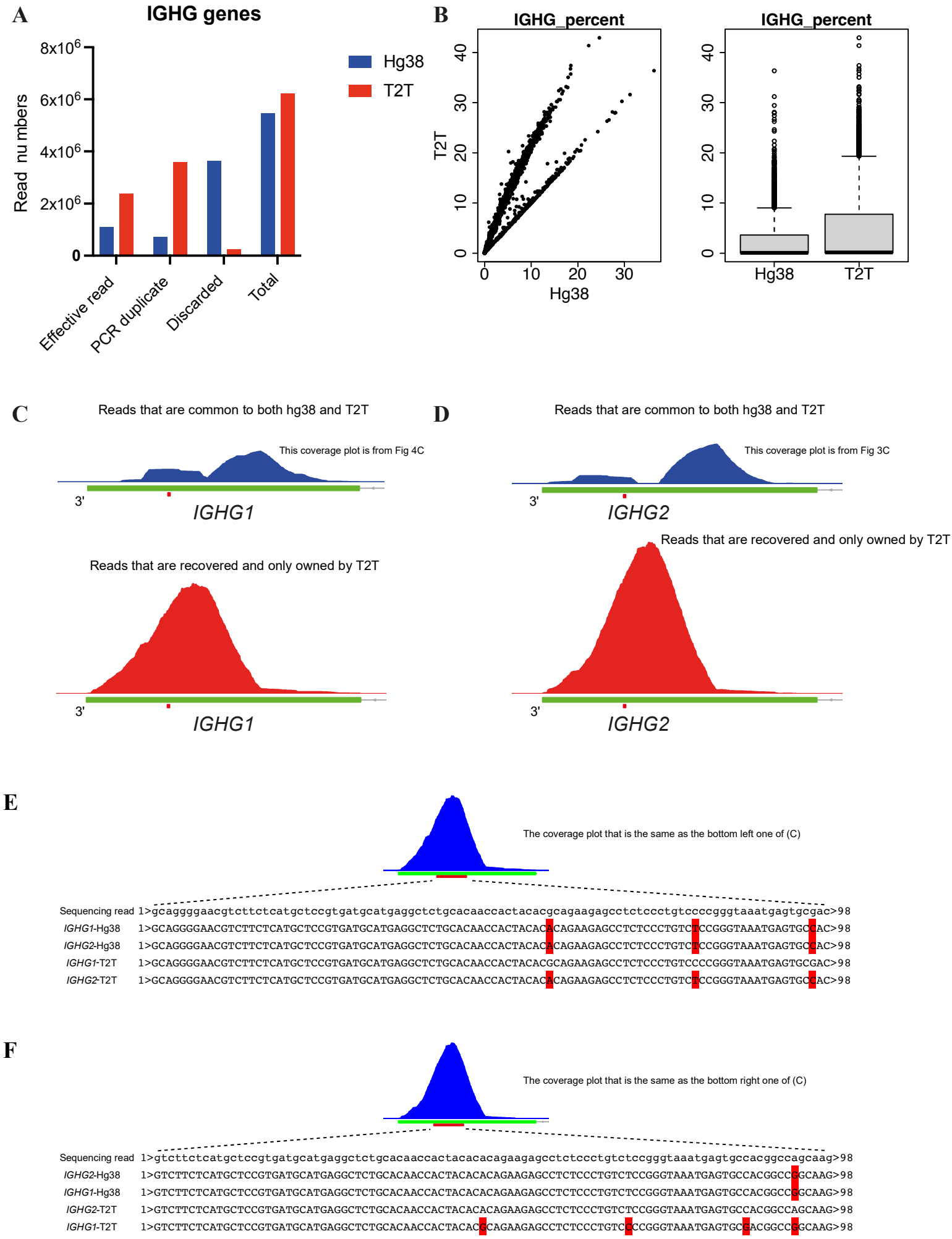
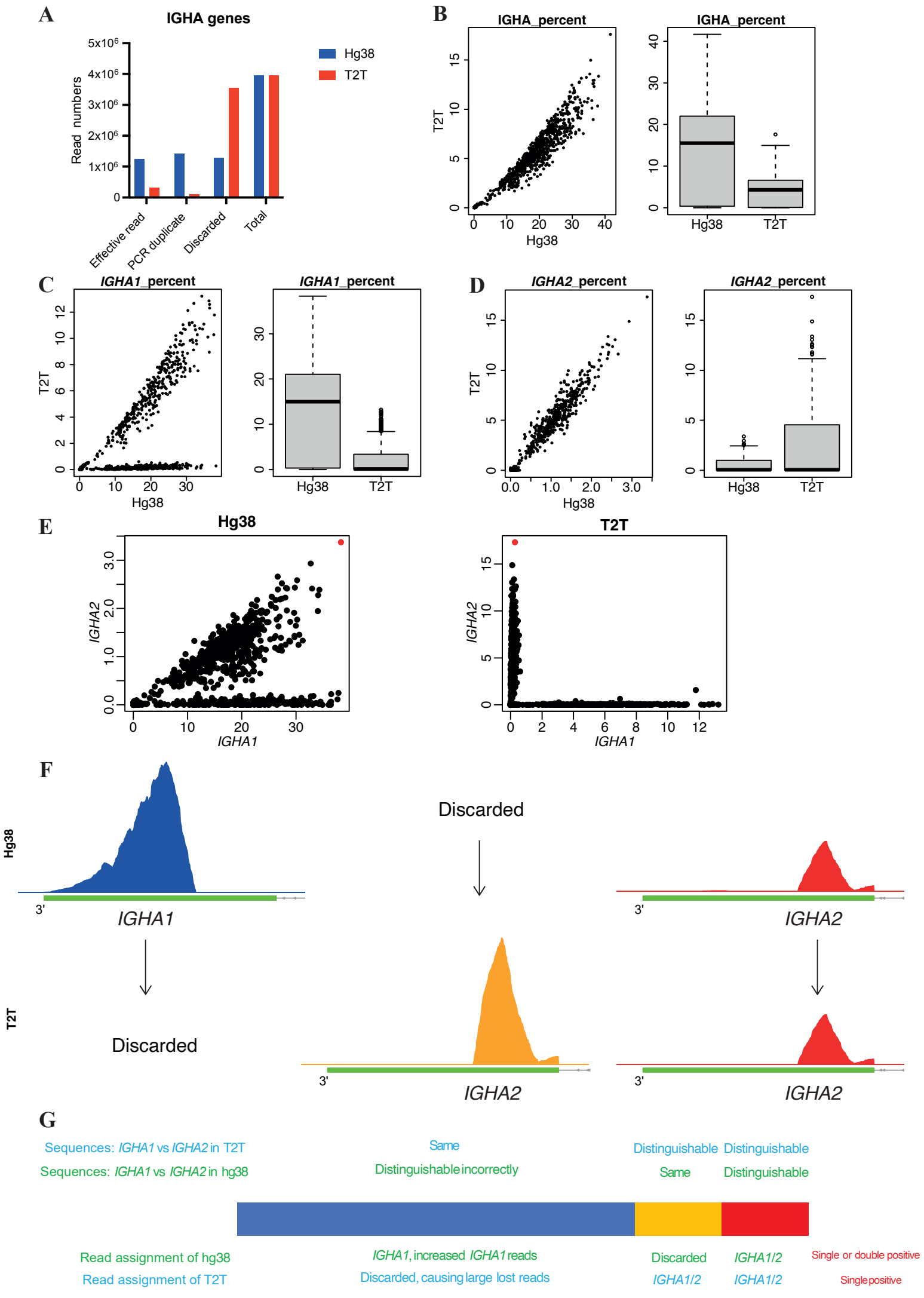

A
IGHG genes
Read numbers
8x10^6
6x10^6
4x10^6
2x10^6
0
Hg38
T2T
Effective read
PCR duplicate
Discarded
Total
B
IGHG_percent
T2T
Hg38
IGHG_percent
Hg38
T2T
C
Reads that are common to both hg38 and T2T
This coverage plot is from Fig 4C
3'
IGHG1
Reads that are recovered and only owned by T2T
3'
IGHG1
D
Reads that are common to both hg38 and T2T
This coverage plot is from Fig 3C
3'
IGHG2
Reads that are recovered and only owned by T2T
3'
IGHG2
E
The coverage plot that is the same as the bottom left one of (C)
Sequencing read 1>gcaggggaacgtcttctcatgctccgtgatgcatgaggctctgcacaaccactacacgcagaagagcctctccctgtccccgggtaaatgagtgcgac>98
IGHG1-Hg38 1>GCAGGGGAACGTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCAC>98
IGHG2-Hg38 1>GCAGGGGAACGTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCAC>98
IGHG1-T2T 1>GCAGGGGAACGTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACGCAGAAGAGCCTCTCCCTGTCCCCGGGTAAATGAGTGCGAC>98
IGHG2-T2T 1>GCAGGGGAACGTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCAC>98
F
The coverage plot that is the same as the bottom right one of (C)
Sequencing read 1>gtcttctcatgctccgtgatgcatgaggctctgcacaaccactacacacagaagagcctctccctgtctccgggtaaatgagtgccacggccagcaag>98
IGHG2-Hg38 1>GTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCACGGCCGGCAAG>98
IGHG1-Hg38 1>GTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCACGGCCGGCAAG>98
IGHG2-T2T 1>GTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACACAGAAGAGCCTCTCCCTGTCTCCGGGTAAATGAGTGCCACGGCCAGCAAG>98
IGHG1-T2T 1>GTCTTCTCATGCTCCGTGATGCATGAGGCTCTGCACAACCACTACACGCAGAAGAGCCTCTCCCTGTCGCCGGGTAAATGAGTGCGACGGCCGGCAAG>98

Figure 6


A
IGHA genes
Read numbers
5x10^6
4x10^6
3x10^6
2x10^6
1x10^6
0
Hg38
T2T
Effective read
PCR duplicate
Discarded
Total
B
IGHA_percent
T2T
Hg38
IGHA_percent
Hg38
T2T
C
IGHA1_percent
T2T
Hg38
IGHA1_percent
Hg38
T2T
D
IGHA2_percent
T2T
Hg38
IGHA2_percent
Hg38
T2T
E
Hg38
IGHA2
IGHA1
T2T
IGHA2
IGHA1
F
Hg38
Discarded
3'
IGHA1
3'
IGHA2
T2T
Discarded
3'
IGHA2
3'
IGHA2
G
Sequences: IGHA1 vs IGHA2 in T2T
Sequences: IGHA1 vs IGHA2 in hg38
Same
Distinguishable incorrectly
Distinguishable Distinguishable
Same Distinguishable
Read assignment of hg38
Read assignment of T2T
IGHA1, increased IGHA1 reads
Discarded, causing large lost reads
Discarded
IGHA1/2
IGHA1/2
IGHA1/2
Single or double positive
Single positive

Supplementary Figure Legends

Figure S1 Somatic recombination of the *IGH* locus in human B cells
Schematic plots showing the structure of the human immunoglobulin heavy chain gene (*IGH*) and the process of VDJ recombination and class switch recombination. The first row shows the structure of *IGH* gene locus in germ line state, where red rectangles represent *IGHV* genes, green rectangles represent *IGHD* genes, and blue rectangles represent *IGHC* genes. The second row shows the structure of *IGH* gene locus after VDJ recombination during B cell development in which a single VDJ exon is synthesised using one IGHV gene, one IGHD gene and one IGHJ gene while the IGHC genes remain unchanged. At this stage, the VDJ exon will splice to *IGHM* or *IGHD* by alternative splicing to generate IgM or IgD. Once B cells become activated, class switch recombination is performed, after which the DNA sequence between VDJ exon and any *IGHC* gene except *IGHM* and *IGHD* (*IGHA1* is shown as an example) will be removed as shown in the bottom row. After class switch recombination, the VDJ exon will splice to the most 5' remaining IGHC, in this case *IGHA1* to generate IgA1. "//" in red or green means that some *IGHV* genes or *IGHD* genes are not shown, while "//" in blue means that a long DNA segment is omitted due to space limitations.

Figure S2 Mapping performance and cell clustering is comparable between hg38 and T2T
A) Classification of reads for the tonsil (left) and bone marrow (right) datasets for each reference genome. The barplots show the total read count for each sample and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons. Only effective reads were used for downstream analyses.
B) t-SNE plots of all cells for the intestinal mucosa (top), tonsil (middle) and bone marrow (bottom) datasets for each reference genome. Antibody-secreting cells (highlighted in blue) cluster separately from other cells and were selected for downstream isotype analyses.

Figure S3 Cellular metrics for quality control and key gene expression are comparable between hg38 and T2T
A) Reads per cell (left) and genes detected per cell (right) for the tonsil (top) and bone marrow (bottom) datasets. Here and subsequently, only bona fide ASCs are shown. The hg38 and T2T cell-wise counts are compared by scatter plots and by side-by-side boxplots. Boxplots show median, quartiles, upper and lower hinges.
B) Cell-wise expression of plasma cell marker genes for the tonsil (left) and bone marrow (right) datasets. Hg38 and T2T expression values are compared by scatter plots and side-by-side boxplots. Here expression values are represented the same as Fig 1C.

Figure S4 The T2T genome identifies IgG subclasses more accurately
Paired scatter plots correlating the expression levels of pairs of IgG subclass genes for intestinal mucosa (top), tonsil (middle) and bone marrow (bottom) ASCs. Plots are shown for hg38 (left) and T2T (right). Expression is represented as reads per hundred (RPH).

Figure S5 The T2T genome identifies IgA subclasses more accurately
Scatter plots correlating *IGHA1* and *IGHA2* expression for the BMPC (top), tonsil (middle) and bone marrow (bottom) datasets. Left plots are for hg38 and right plots for T2T.

Figure S6 T2T has more reads effectively mapped to IGHG genes than hg38
A) Classification of reads that are assigned to IGHG genes for the intestine (top), tonsil (middle) and bone marrow (bottom) datasets for each reference genome. The barplots show the total

read count of IGHG genes and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons.
B) Percentage of IGHG gene reads per cell for the intestine (top), tonsil (middle) and bone marrow (bottom) datasets. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

Figure S7 T2T has fewer reads mapped to *IGHA* but better distinguishes *IGHA1* from *IGHA2*
A) Classification of reads that are assigned to IGHA genes for for the BMPC (left), tonsil (middle) and bone marrow (right) datasets for each reference genome. The barplots show the total read count of IGHA genes and also the number of effective reads, the number of reads identified as PCR duplicates and the number of reads discarded for other reasons.
B) Percentage of IGHA genes reads per cell for the BMPC (left), tonsil (middle) and bone marrow (right) datasets. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.
C) Percentage of *IGHA1* (top) and *IGHA2* (bottom) reads per cell for the BMPC (left), tonsil (middle) and bone marrow (right) datasets. Hg38 and T2T results are compared by scatter plots and side-by-side boxplots.

Figure S1

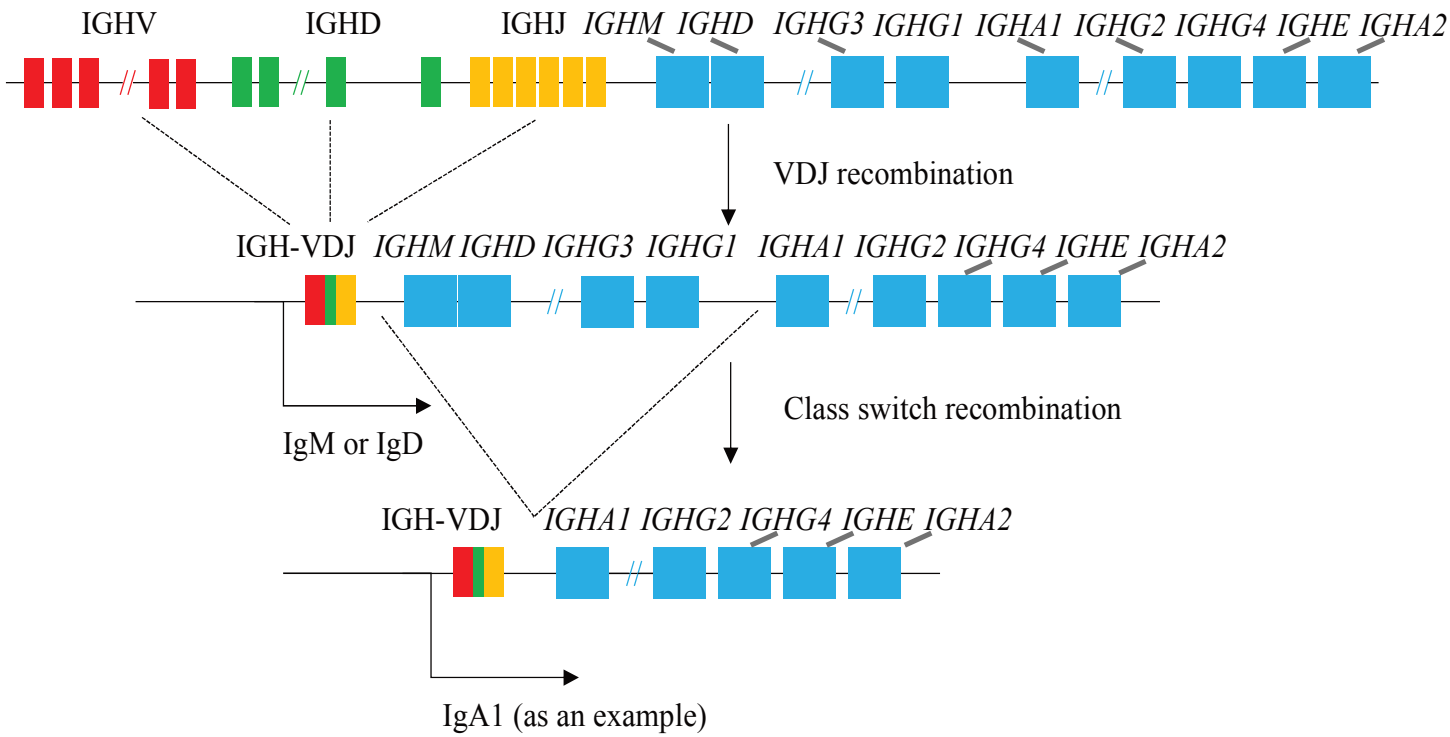

IGHV
IGHD
IGHJ
IGHM
IGHD
IGHG3
IGHG1
IGHA1
IGHG2
IGHG4
IGHE
IGHA2
VDJ recombination
IGH-VDJ
IGHM
IGHD
IGHG3
IGHG1
IGHA1
IGHG2
IGHG4
IGHE
IGHA2
IgM or IgD
Class switch recombination
IGH-VDJ
IGHA1
IGHG2
IGHG4
IGHE
IGHA2
IgA1 (as an example)

Figure S2

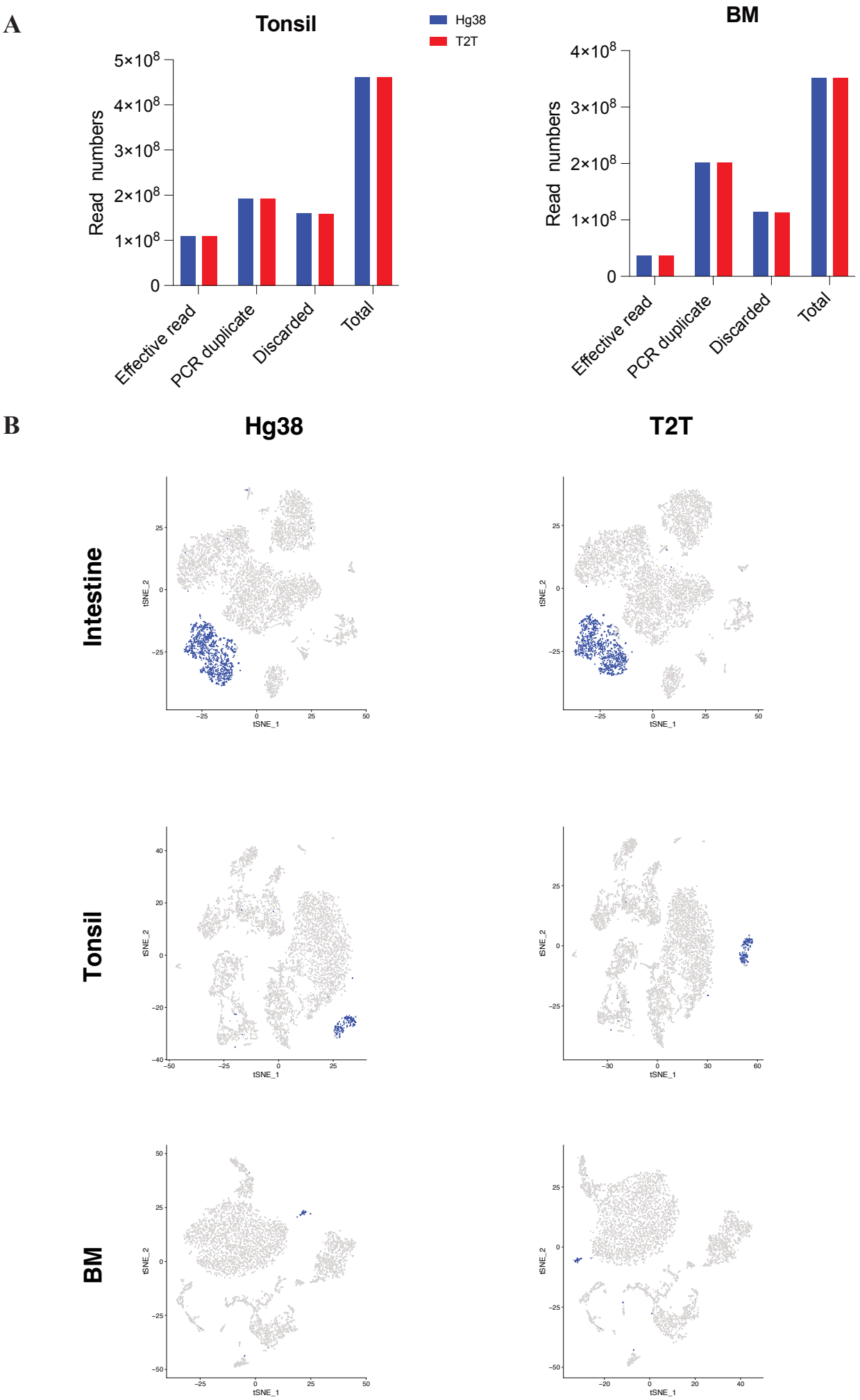

A
Tonsil
BM
Hg38
T2T
Read numbers
5×10^8
4×10^8
3×10^8
2×10^8
1×10^8
0
Effective read
PCR duplicate
Discarded
Total
B
Hg38
T2T
Intestine
Tonsil
BM
tSNE_1
tSNE_2

Figure S3

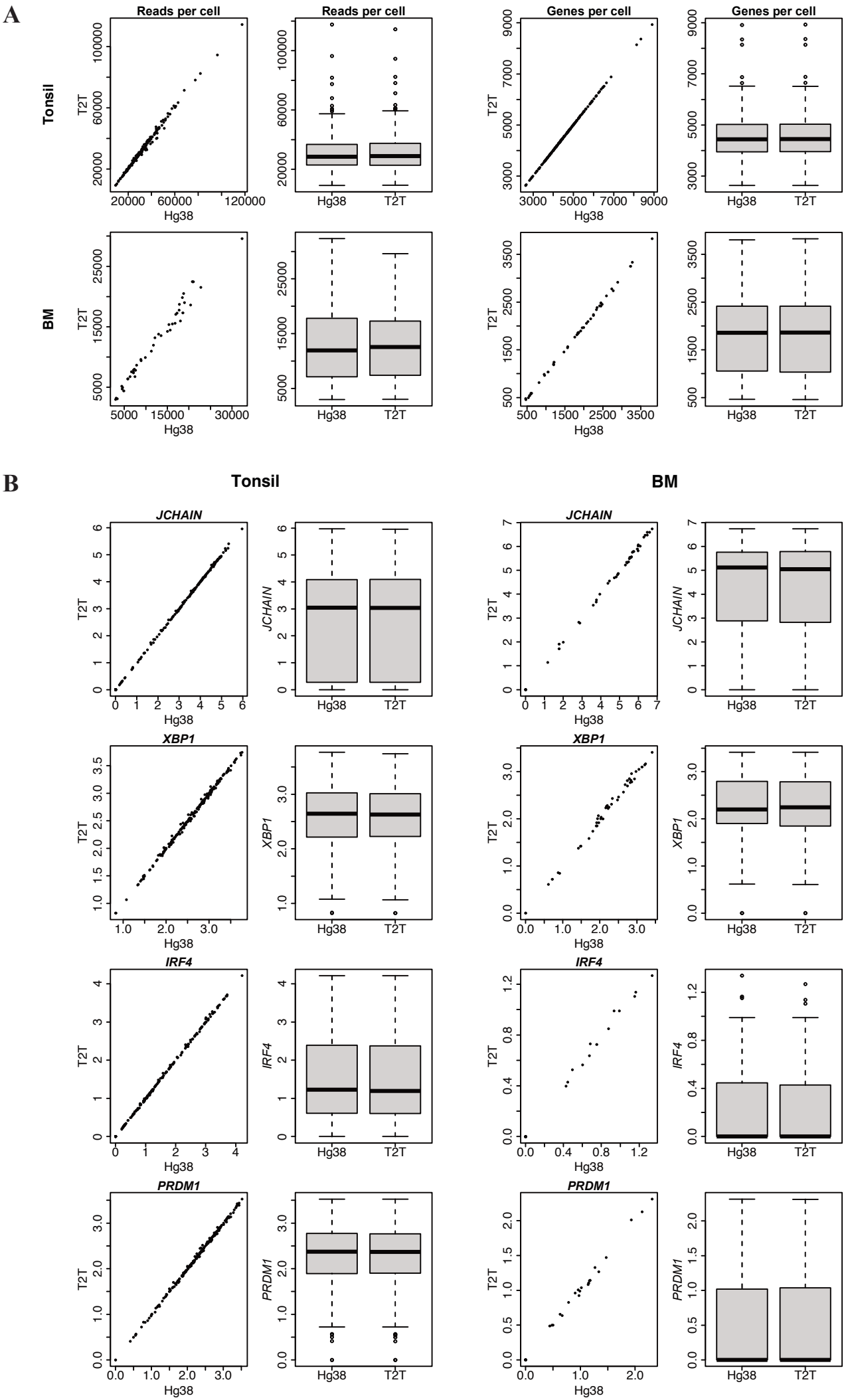

A
Reads per cell
Reads per cell
Genes per cell
Genes per cell
Tonsil
BM
T2T
Hg38
B
Tonsil
BM
JCHAIN
XBP1
IRF4
PRDM1

Figure S4

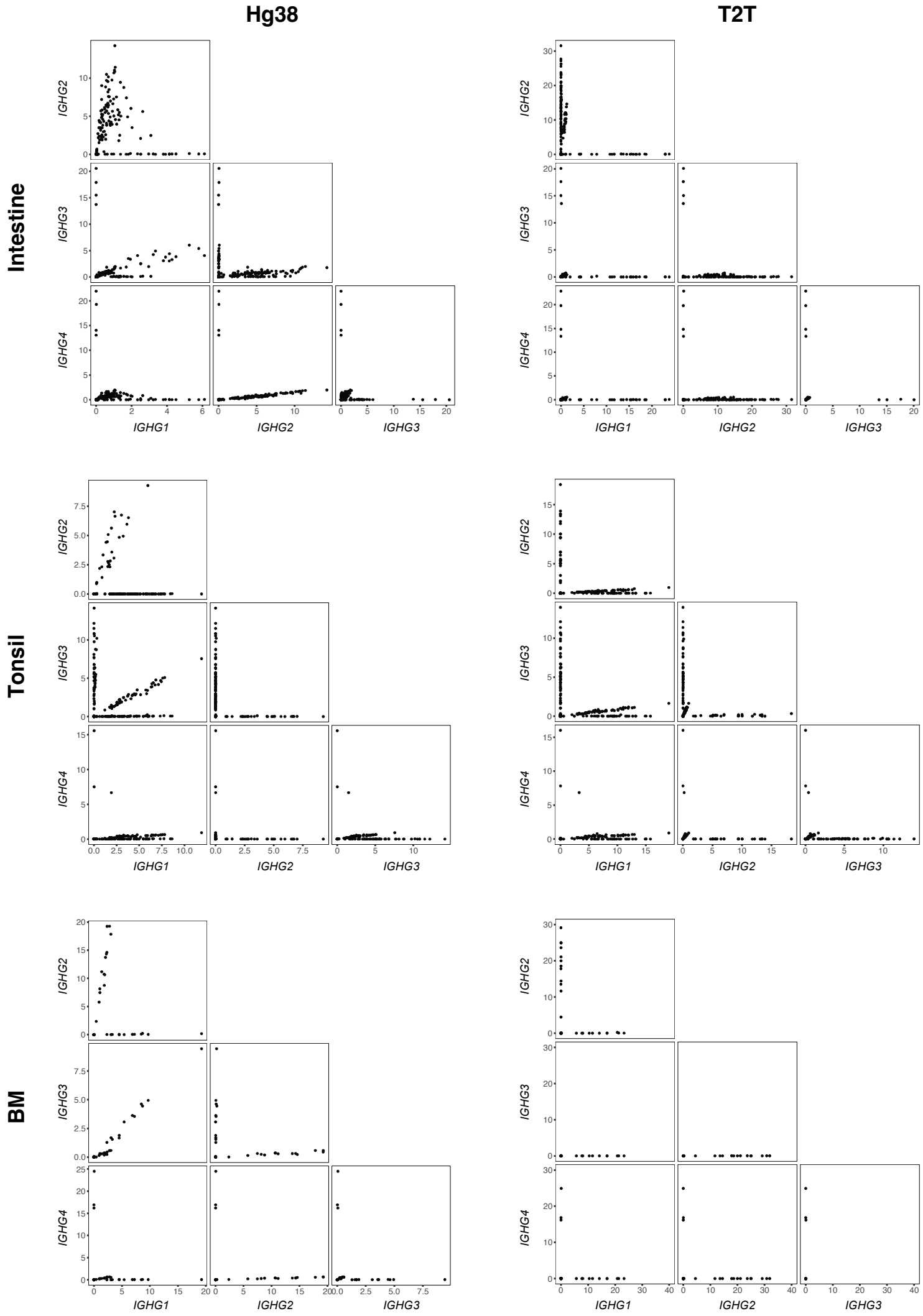


Figure S5

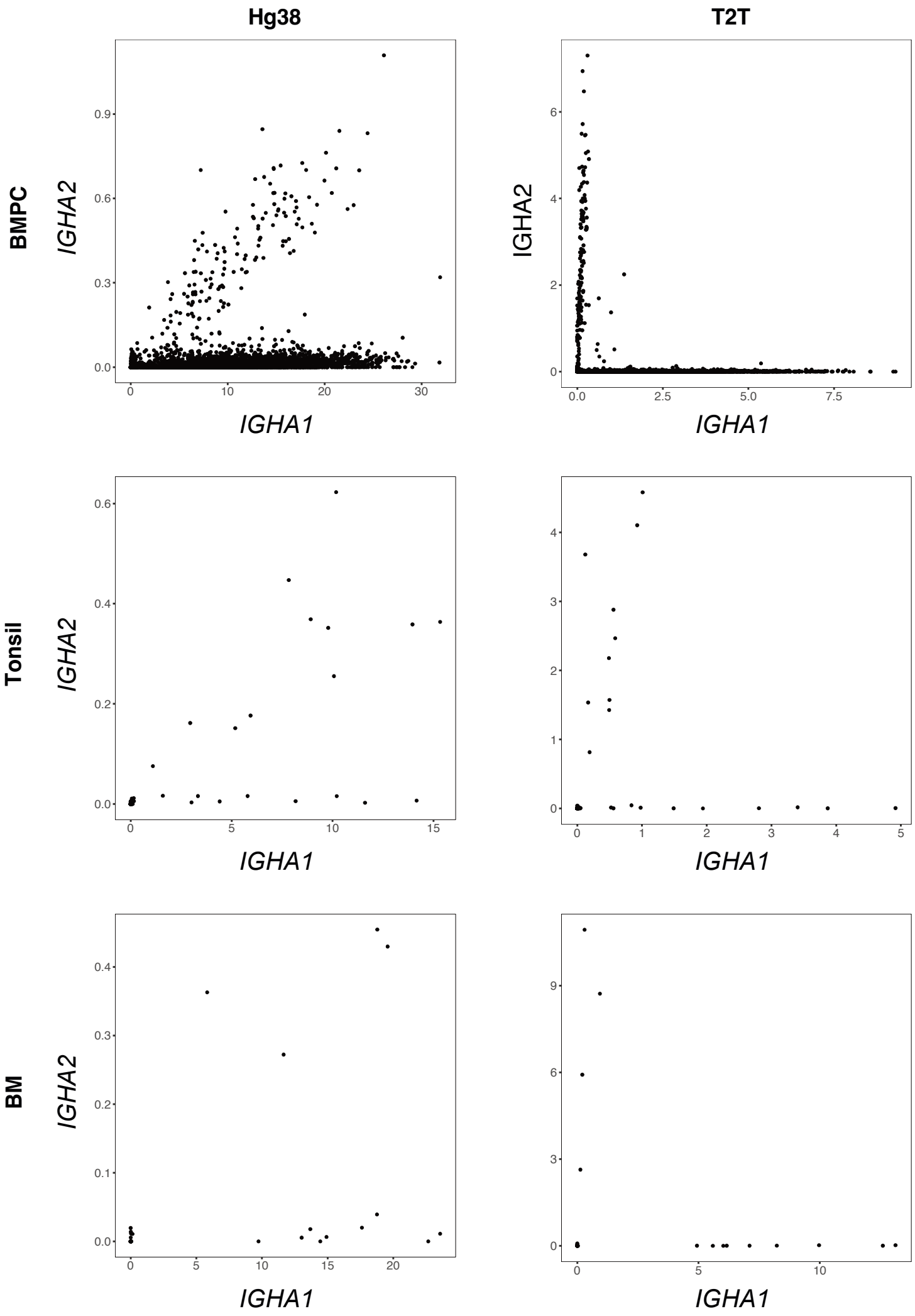


Figure S6

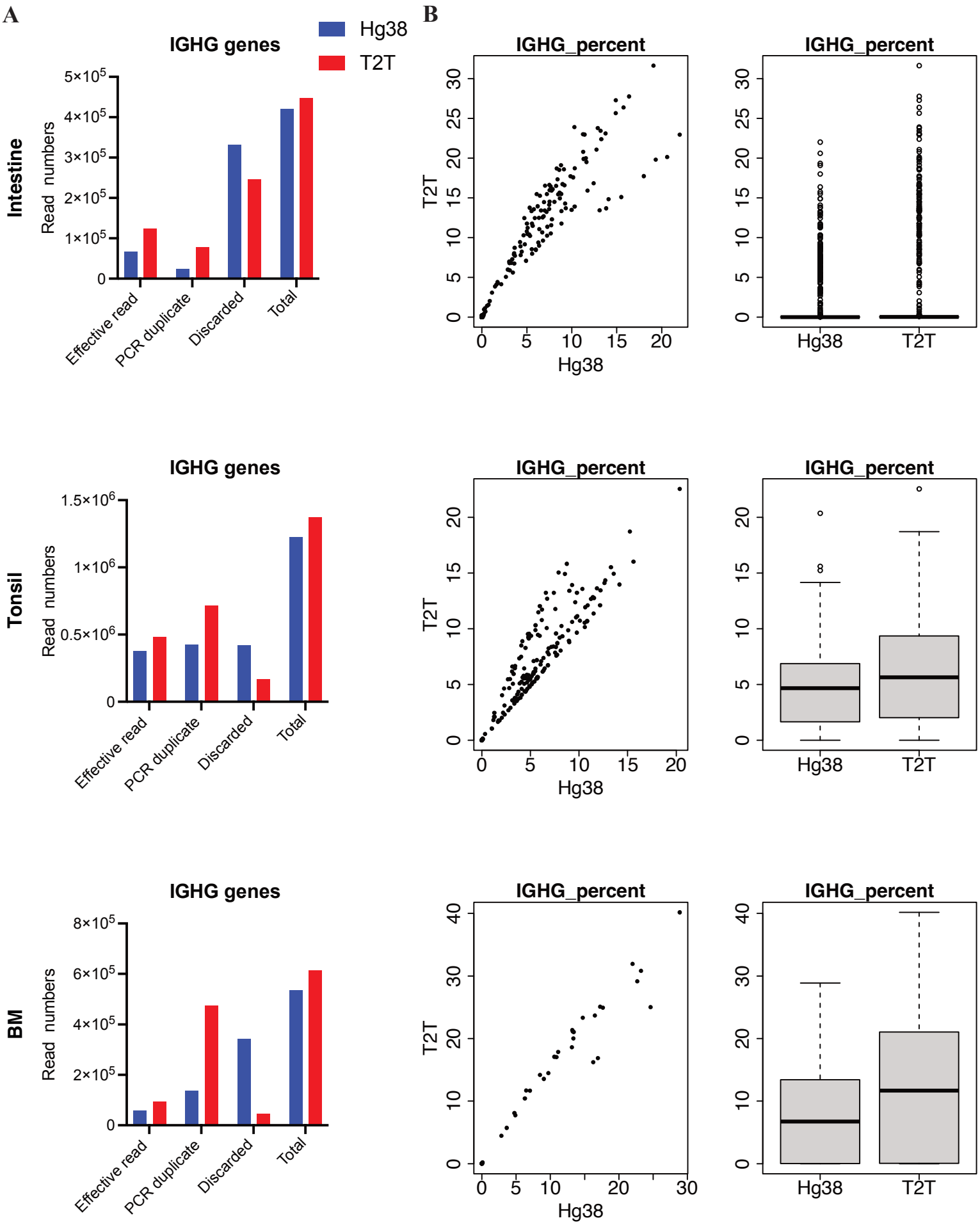
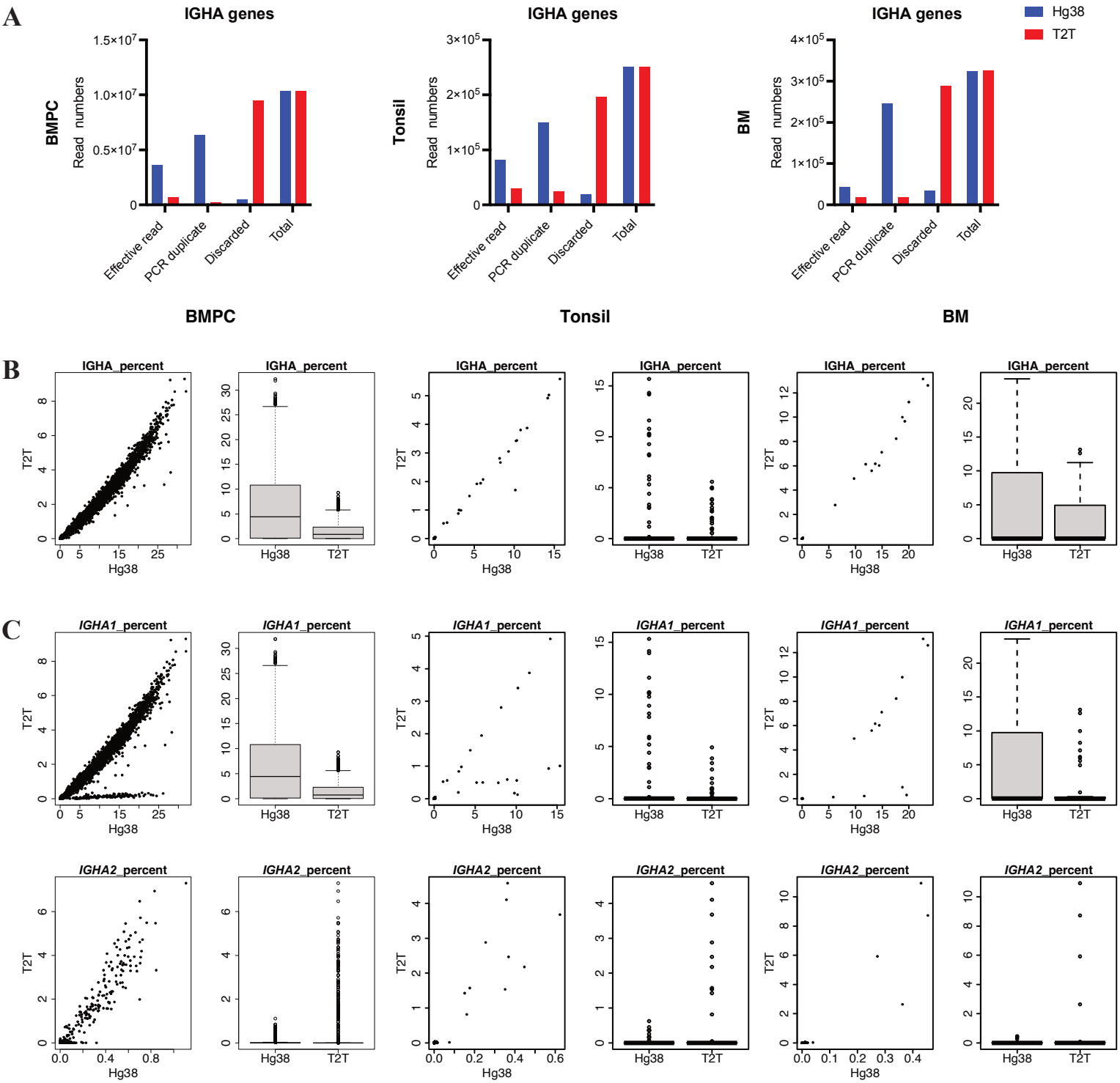

A
B
Hg38
T2T
IGHG genes
Intestine
Tonsil
BM
Read numbers
Effective read
PCR duplicate
Discarded
Total
IGHG_percent
T2T
Hg38

Figure S7


A
IGHA genes
BMPC
Tonsil
BM
Read numbers
Effective read
PCR duplicate
Discarded
Total
Hg38
T2T
B
IGHA_percent
C
IGHA1_percent
IGHA2_percent