



## OPEN ACCESS

## EDITED BY

Peter S. Linsley,  
Benaroya Research Institute, United States

## REVIEWED BY

Jason Dale Turner,  
University of Birmingham, United Kingdom  
Naresh Doni Jayavelu,  
Benaroya Research Institute, United States

## \*CORRESPONDENCE

Benjamin N. Bimber  
✉ bimber@ohsu.edu

RECEIVED 20 March 2025

ACCEPTED 19 July 2025

PUBLISHED 08 August 2025

## CITATION

Benjamin S, McElfresh G, Kaza M, Boggy GJ, Varco-Merth B, Ojha S, Feltham S, Goodwin W, Nkoy C, Duell D, Selseth A, Bennett T, Barber-Axthelm A, Haese NN, Wu H, Waytashek C, Boyle C, Smedley JV, Labriola CS, Axthelm MK, Reeves RK, Streblow DN, Sacha JB, Okoye AA, Hansen SG, Picker LJ and Bimber BN (2025) An immune-focused supplemental alignment pipeline captures information missed from dominant single-cell RNA-seq analyses, including allele-specific MHC-I regulation. *Front. Immunol.* 16:1596760. doi: 10.3389/fimmu.2025.1596760

## COPYRIGHT

© 2025 Benjamin, McElfresh, Kaza, Boggy, Varco-Merth, Ojha, Feltham, Goodwin, Nkoy, Duell, Selseth, Bennett, Barber-Axthelm, Haese, Wu, Waytashek, Boyle, Smedley, Labriola, Axthelm, Reeves, Streblow, Sacha, Okoye, Hansen, Picker and Bimber. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# An immune-focused supplemental alignment pipeline captures information missed from dominant single-cell RNA-seq analyses, including allele-specific MHC-I regulation

Sebastian Benjamin<sup>1</sup>, GW McElfresh<sup>1</sup>, Maanasa Kaza<sup>1</sup>, Gregory J. Boggy<sup>1</sup>, Benjamin Varco-Merth<sup>1,2</sup>, Sohita Ojha<sup>2</sup>, Shana Feltham<sup>2</sup>, William Goodwin<sup>1,2</sup>, Candice Nkoy<sup>1,2</sup>, Derick Duell<sup>1,2</sup>, Andrea Selseth<sup>2</sup>, Tyler Bennett<sup>2</sup>, Aaron Barber-Axthelm<sup>2</sup>, Nicole N. Haese<sup>1,2</sup>, Helen Wu<sup>1</sup>, Courtney Waytashek<sup>1</sup>, Carla Boyle<sup>1</sup>, Jeremy V. Smedley<sup>1,2</sup>, Caralyn S. Labriola<sup>1,2</sup>, Michael K. Axthelm<sup>1,2</sup>, R. Keith Reeves<sup>3,4</sup>, Daniel N. Streblow<sup>1,2</sup>, Jonah B. Sacha<sup>1,2</sup>, Afam A. Okoye<sup>1,2</sup>, Scott G. Hansen<sup>1,2</sup>, Louis J. Picker<sup>1,2</sup> and Benjamin N. Bimber<sup>1,2\*</sup>

<sup>1</sup>Oregon National Primate Research Center, Oregon Health and Science University, Beaverton, OR, United States, <sup>2</sup>Vaccine and Gene Therapy Institute, Oregon Health and Science University, Beaverton, OR, United States, <sup>3</sup>Division of Innate and Comparative Immunology, Center for Human Systems Immunology, Duke University School of Medicine, Durham, NC, United States, <sup>4</sup>Department of Surgery, Duke University School of Medicine, Durham, NC, United States

**Introduction:** RNA sequencing (RNA-seq) can measure whole transcriptome gene expression from tissues or even individual cells, providing a powerful tool to study the immune response. Analysis of RNA-seq data involves mapping relatively short sequence reads to a reference genome, and quantifying genes based on the position of alignments relative to annotated genes. While this is usually robust, genetic polymorphism or genome/annotation inaccuracies result in genes with systematically missing or inaccurate data. These issues are frequently hidden or ignored, yet are highly relevant to immunologic data, where balancing selection has generated many polygenic gene families not accurately represented in a 'one-size-fits-all' reference genome.

**Methods:** Here we present nimble, a tool to supplement standard RNA-seq pipelines. Nimble uses a previously developed pseudoaligner to process either bulk- or single-cell RNA-seq data using custom gene spaces. Importantly, nimble can apply customizable scoring criteria to each gene set, tailored to the biology of those genes.

**Results:** We demonstrate that nimble recovers data in diverse contexts, ranging from simple cases (e.g., incorrect gene annotation or viral RNA), to complex immune genotyping (e.g., major histocompatibility or killer-immunoglobulin-like receptors). We use this enhanced capability to identify killer-immunoglobulin-

like receptor expression specific to tissue-resident memory T cells and demonstrate allele-specific regulation of MHC alleles after *Mycobacterium tuberculosis* stimulation.

**Discussion:** Combining nimble data with standard pipelines enhances the fidelity and accuracy of experiments, maximizing the value of expensive datasets, and identifying cellular subsets not possible with standard tools alone.

#### KEYWORDS

single-cell RNA-seq (scRNA-seq), T cells, bioinformatics, immunogenetics, major histocompatibility complex (MHC)

## Introduction

RNA-sequencing (RNA-seq) and single-cell RNA-sequencing (scRNA-seq) technologies provide transcriptome-wide quantification in a sample of interest. In the case of scRNA-seq, transcriptomes are captured from individual cells, allowing for high-resolution observations of cellular function and differentiation. These high-dimensional data benefit the analysis of large populations of cells, such as those common in immunologic data. The rapid and accurate production of these data relies on complex software quantification toolchains. The process of transcript quantification is characterized by many technical decision points which, while generally obscured from downstream analysis, have a profound impact on the produced count data, depending on the quantification method of choice and its interaction with the reference genome.

The bioinformatic processing of RNA-seq and scRNA-seq data involves several steps. In most cases, short reads are aligned to a reference genome which is annotated for gene and features. In general, one genome is used to represent the diversity of the entire species. After alignment, an algorithm is run to assign reads to genes/features, producing gene counts. There are many established tools and pipelines for RNA-seq analysis. STAR is a commonly used alignment tool that can align reads by local positional alignment to a reference genome or transcriptome in a splice-aware manner (1). Kallisto performs transcript quantification by pseudoalignment of reads to a reference genome, without undergoing an expensive positional alignment process first (2). Feature calling is sometimes included with the aligner, and is sometimes performed using a separate tool, such as HTseq (3). Especially for scRNA-seq analyses, it is common for vendors to wrap all steps into one pipeline, such as the 10x Genomics CellRanger software. While these tools have differences in their implementation, they each function by aligning all data from a sample to a single reference genome, and they score genes/features using a ‘one-size-fits-all’ logic that treats all genes identically. This approach can work quite well and is probably the desirable approach for most genes.

There are nonetheless situations where standard pipelines are systematically inaccurate or sub-optimal (Figure 1). Complex

regions of the genome, especially gene families with copy number differences and/or segmental duplication, are difficult to accurately assemble when generating reference genomes. If the reference genome is inaccurate or incomplete, this results in feature counting artifacts, such as missing counts for expected genes. If a gene that is transcribed is not represented in the reference genome, the RNA-seq reads from that gene can misalign to the closest available gene, inflating these counts and providing misleading data. Improved genomic assemblies, especially those generated from long read sequencing, will improve this to a point; however, there are gene families with characteristics that remain problematic. Instances where two highly similar genes are encoded in the genome can result in alignment ambiguity or multi-mapped reads, which are often discarded, resulting in lost data. Some gene families have high degrees of variation between individuals, meaning it is extremely difficult to represent genomic diversity of the species with one single reference genome. Examples of these include the major histocompatibility complex (MHC), which is the most variable region of primate genomes, or killer immunoglobulin-like receptors (4–9). In the case of MHC class I, in addition to variable gene content in some species, there is extremely high allelic diversity, with thousands of known alleles (5). While most RNA-seq and scRNA-seq analyses are designed to ignore allelic variation, the identity of the expressed MHC alleles for a subject is critical to antigen recognition, and thus higher resolution genotyping is often needed. Standard RNA-seq pipelines generally treat quantification of all genes identically, which does not permit adaptation of feature calling to match the biology or differing needs of certain gene families.

To address the limitations of standard RNA-seq and scRNA-seq pipelines, we developed nimble, a lightweight tool intended to provide supplemental gene counts to complement standard pipelines (Figure 1). Nimble is designed to be executed against one or more customizable gene spaces, where each gene space contains a focused set of reference sequences to address a specific question. Nimble uses a previously published pseudoalignment engine to align reads against these references (10), followed by customizable logic for feature calling. The combination of these two

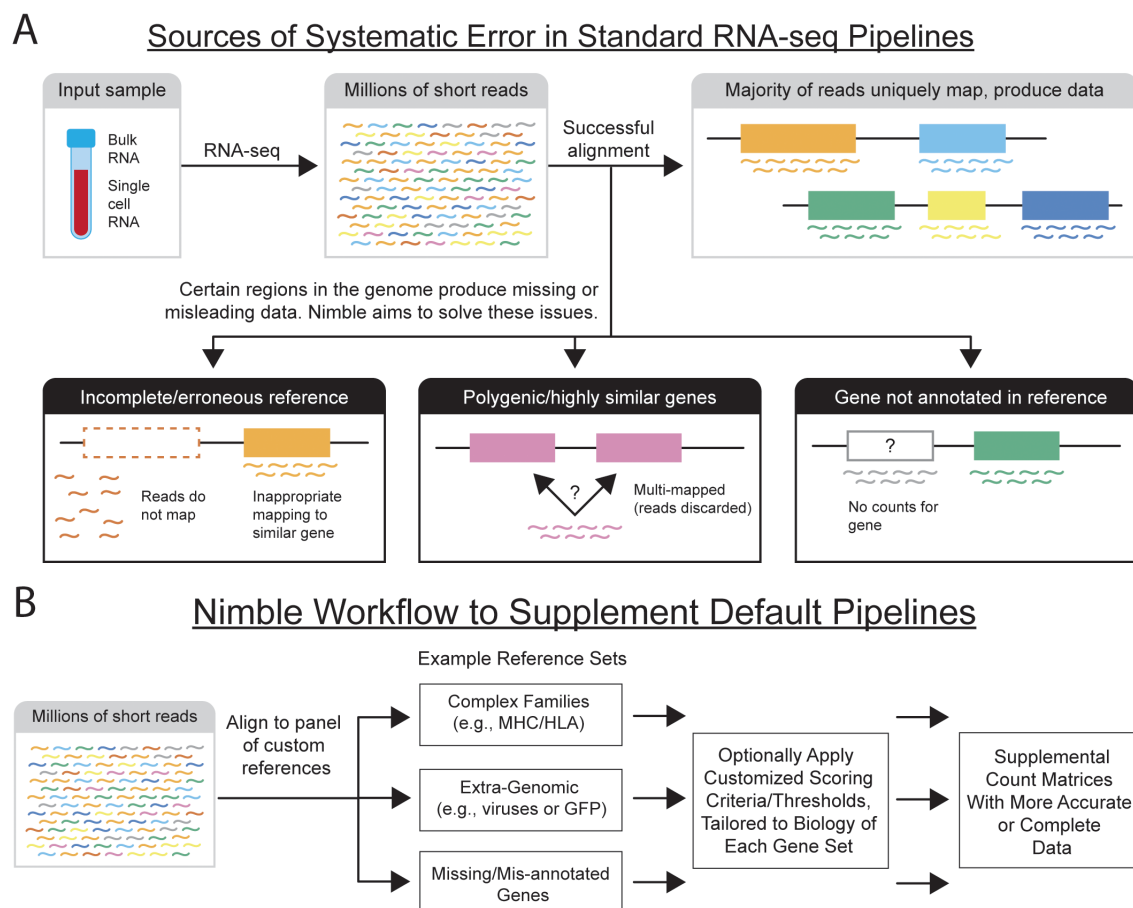


FIGURE 1

Diagram of RNA-seq alignment and potential pitfalls. **(A)** The schematic illustrates the alignment of short read data to a reference genome. In most cases, the short reads uniquely map to that reference, providing unambiguous gene counts. There are nonetheless several examples, with both technical and biological cases, that result in reads being unable to uniquely map to the genome, resulting in either missing or inaccurate gene counts. If the reference assembly is incomplete and lacks a gene, or if that gene is present but not annotated, either missing data or inaccurately aligned counts are produced. Further, when multiple copies of highly similar genes are present, the aligned often cannot uniquely assign the reads to a gene, often resulting in lost data. Finally, certain regions of the genome are highly polymorphic across the population (e.g., MHC) making it virtually impossible to accurately represent the species using one reference genome. **(B)** The schematic illustrates the workflow used by nimble to address deficiencies in standard single-cell and bulk RNA-seq pipelines. Nimble allows the user to create multiple custom reference spaces, where each is generally designed to address a specific need, such as a reference containing the sequences of genes missing or mis-annotated in the reference, or extra-genomic sequences (e.g., a viral genome). This could also involve specialist databases, such as a reference containing all MHC/HLA alleles. Equally important, nimble allows customized feature calling thresholds for each reference. This is critical to support applications like MHC-typing, where higher resolution matches are required than standard feature calling. The result of this pipeline is a set of supplemental count matrices containing per-sample or per-cell counts for the additional genes/features. These data can either be merged with the existing gene counts, or analyzed in parallel, depending on the experimental needs.

capabilities allows nimble to quantify both simple and complex gene families, especially when the biology or characteristics of these genes are problematic for the standard one-size-fits all alignment and feature calling pipelines.

## Results

### Design of nimble and concordance with standard pipelines

Nimble is the combination of a previously developed pseudoaligner and customizable feature-calling algorithm, designed to allow the user to perform targeted quantification of

one or more panels of interest. While nimble is primarily designed to address complex genomic regions, we first constructed a panel of “simple” genes that lack the complex genetics or high intra-species variation that can confound standard RNA-seq pipelines. We processed a single-cell RNA-seq (scRNA-seq) dataset from rhesus macaque peripheral blood mononuclear cells (PBMC) and compared the counts obtained by the CellRanger pipeline using the Mmul10 reference genome (“CellRanger/Mmul10”) against the counts obtained by nimble using this custom gene space (Figure 2). We contrasted un-normalized raw counts in aggregate, prior to downstream normalization or other processing, to provide the most direct comparison of alignment behavior. The results are highly similar both when comparing the total counts per gene (Figure 2), and the per-cell counts (Figure 2). While there is minor variation

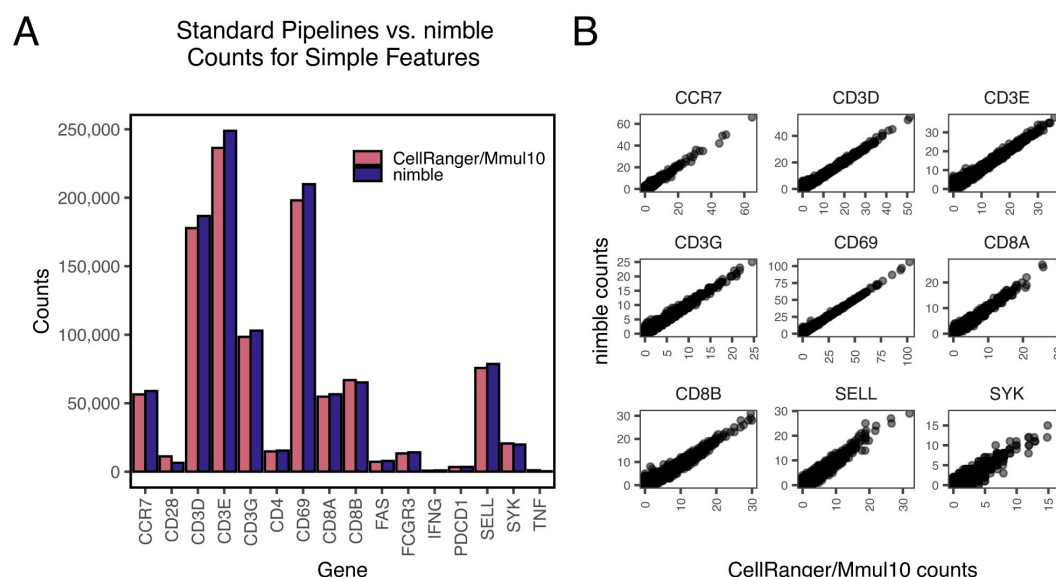


FIGURE 2

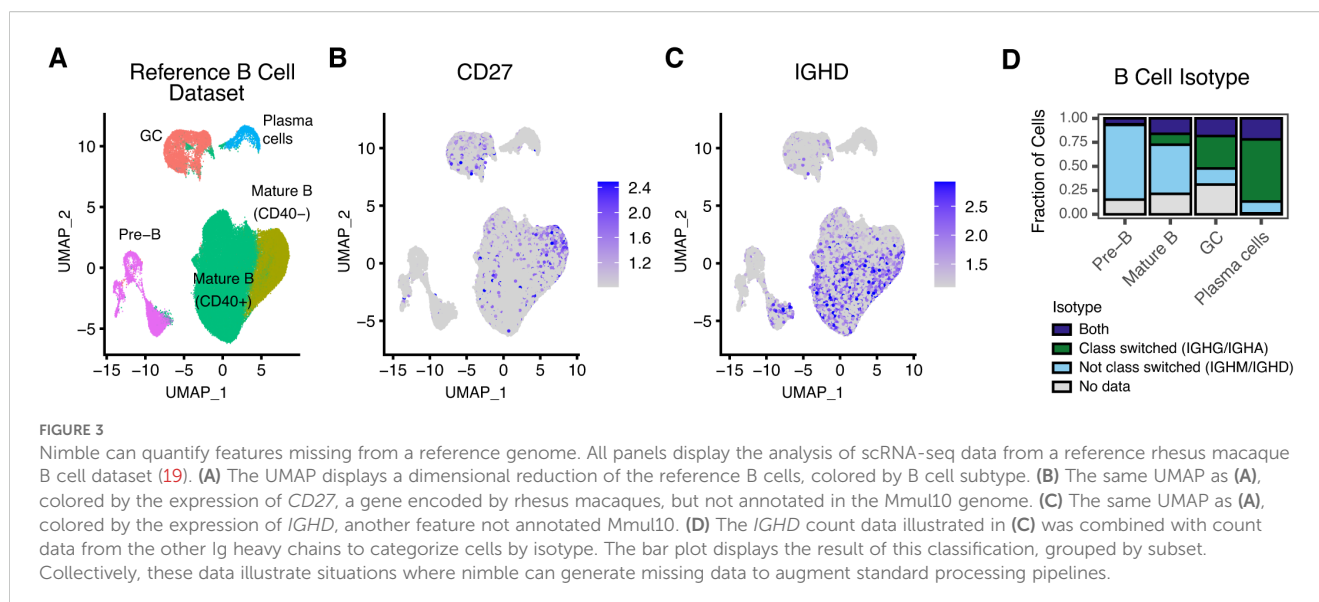
Validation of nimble accuracy relative to established pipelines. All panels display summaries of raw count data generated by processing rhesus macaque scRNA-seq data through either nimble (using a gene space comprised of the mRNA sequences for these genes), or CellRanger (using the Mmul10 reference genome). (A) The bar plot compares the magnitude of raw counts generated by nimble and CellRanger/Mmul10 in aggregate for a set of common immune genes. These genes were selected because they represent “simple” features, present at a single copy in the genome without complex polymorphism expected between subjects. (B) The scatter plot compares the range of counts generated by nimble and CellRanger/Mmul10 for each cell in the dataset. These data demonstrate that both tools produce similar per-cell counts when executed against typical genes.

between the tools, this is likely due to differences in alignment algorithm or scoring thresholds. While nimble is not designed to completely replace standard alignment and feature calling pipelines, to provide a more comprehensive comparison of nimble with standard pipelines we generated a nimble library containing the complete 15,782 genes defined in the MMul\_10 genome, and compared the resulting per-cell counts against the same data processed with CellRanger/MMul\_10. The results were highly concordant, with a Pearson correlation of 0.968 (Supplementary Figure 1). Together, these data indicate that nimble’s alignment pipeline captures similar count data to standard pipelines, establishing nimble’s accuracy when aligning to a straightforward gene space. Nimble’s performance scales with available hardware via thread-level parallelism and will attempt to fully-saturate the provided cores. RAM usage is low, requiring memory only for the reference de Bruijn graph and 50 UMIs of buffered data from the input.bam file. In one example, aligning 491 million paired-end reads to a ~2,200-feature MHC reference completed in 225 minutes on 18 CPUs, sustaining ~36,000 reads/sec. Performance scales in a nearly linear manner with CPU count.

## Quantification of genes missing from the reference genome enhances measurement of B cell class switching

A second straightforward usage of nimble is to quantify genes or features not annotated or misannotated in the reference genome.

While this is less common for the human genome, the genomes of model organisms frequently have less complete or accurate gene models. While gene models can be corrected, generating counts for missing features, at least for most scRNA-seq pipelines, requires repeating the entire alignment. Rhesus macaques encode both *CD27* and immunoglobulin heavy constant delta (*IGHD*), and while the sequence for these genes is present in the MMul\_10 genome, neither are annotated in the NCBI gene build (version 103). Both genes provide useful information about B cell differentiation states (11). To overcome this, we generated a nimble reference containing these genes, along with the remaining Ig heavy chains (*IGHA*, *IGHE*, *IGHM*, *IGHG1*, *IGHG2*, *IGHG3*, and *IGHG4*) to provide a comparison against standard pipelines (Supplementary Table S1). We processed a previously published rhesus macaque reference B cell dataset using this reference space (Figure 3). This dataset contains B cells of multiple differentiation states, including Pre-B cells, mature B cells, germinal center (GC), and plasma cells (Figure 3). Nimble successfully generated missing count data for *CD27*, demonstrating expression primarily in the “innate-like” *CD40<sup>+</sup>* mature B cell cluster, with limited expression among GC cells (Figure 3). *IGHD* is upregulated primarily in pre-B cells and the mature B cell cluster (Figure 3). Finally, we used the nimble immunoglobulin heavy chain expression data to classify B cell class switching status (Figure 3). For each B cell maturation type, we observe a predominant class-switching status: pre-B and mature B cells were predominately not class switched, while germinal center and plasma cells were predominately class-switched, reflecting their antigen-exposed state. Additionally, as cells transition through the



class switch recombination process, we observe many different “mixed” expression states, at a diminished ratio compared to cells that fall into one of the two main class-switching categories. Organizing the cell categories by stage in the B cell maturation and differentiation process, we see the expected transition of B cells from not class-switched to class-switched over time. Taken together, these data demonstrate a case where a new nimble gene space allowed for cell classification beyond what is possible using standard pipelines alone.

## Quantification of extra-genomic features

Many experiments require the quantification of features not encoded by the normal species genome, including the sequences of viral or bacterial pathogens, or exogenous genes (e.g., GFP). A common way to address this situation today is to append the exogenous sequence(s) to the species genome and align data to this new composite “genome”. While this is a viable option much of the time, alterations to the base genome can require re-processing of data, creates issues recombining or merging cohorts (it is more complex to merge counts when not aligned to the identical genomic space). Nimble provides an option to rapidly generate counts for any number of custom features, at any point after the primary alignment is performed, which can either be merged to the primary count matrix or treated separately.

To demonstrate examples of this, we analyzed virally infected cells. First, we performed scRNA-seq on primary normal human dermal fibroblasts experimentally infected with Chikungunya virus (CHIKV; strain SL-15649), as well as uninfected controls (Figure 4A). These were processed on separate lanes, and therefore the CHIKV-exposure status of each cell is known. We began by processing using the standard CellRanger/Mmul10 pipeline. PCA/UMAP analyses revealed two main transcriptional clusters, which largely separate the CHIKV-infected from uninfected cells (Figure 4). We aligned these populations to a

custom nimble gene space containing the CHIKV genome, generating per-cell counts. CHIKV-expression corresponded extremely well with the expected groups, with virtually all CHIKV-exposed cells expressing high levels of CHIKV and no CHIKV detected in the control cells. This indicates that nimble is accurately and specifically detecting CHIKV (Figures 4B).

Because the CHIKV experiment involved experimental viral infection, it was obviously important to quantify CHIKV, and the proper CHIKV reference sequence was known prior to analysis. Therefore, alignment of data to an augmented genome and using standard pipelines would be as effective as nimble. This situation is not always true. Next, we performed scRNA-seq on cells cultured from an adrenal mass detected in an immunosuppressed cynomolgus macaque (12). We hypothesized that a virus was the cause of adrenal mass, but this was not known prior to the experiment. We used nimble to align the reads against a genome containing multiple macaque viruses (Supplementary Table S1). Unsupervised clustering on RNA expression (using the host genome and not viral transcripts), revealed multiple transcriptional clusters (Figure 4). The nimble-generated data identified lymphocryptovirus (LCV) in B cells isolated from this sample (Figure 4E). LCV is a member of the gammaherpesvirus family that naturally infects macaque populations, infects B cells, and can cause lymphomas, especially in immunocompromised subjects (13, 14). Also of note, the per-cell expression of LCV differed between the clusters (Figure 4). This is noteworthy because LCV RNA was not part of the dimensional reduction, and therefore the pattern of clustering is driven by changes in the host transcriptome alone. LCV has complex interactions with the host cell, and this suggests that the host gene expression differences could be the effect of increased viral replication or represent different phases of the viral life cycle. While in this experiment we quantified LCV as a single feature, allowing assignment of cells as LCV-positive or LCV-negative, it would be possible to repeat this nimble analysis using a gene space containing the individual LCV transcripts for a more precise quantification. Together, these examples demonstrate that nimble can generate



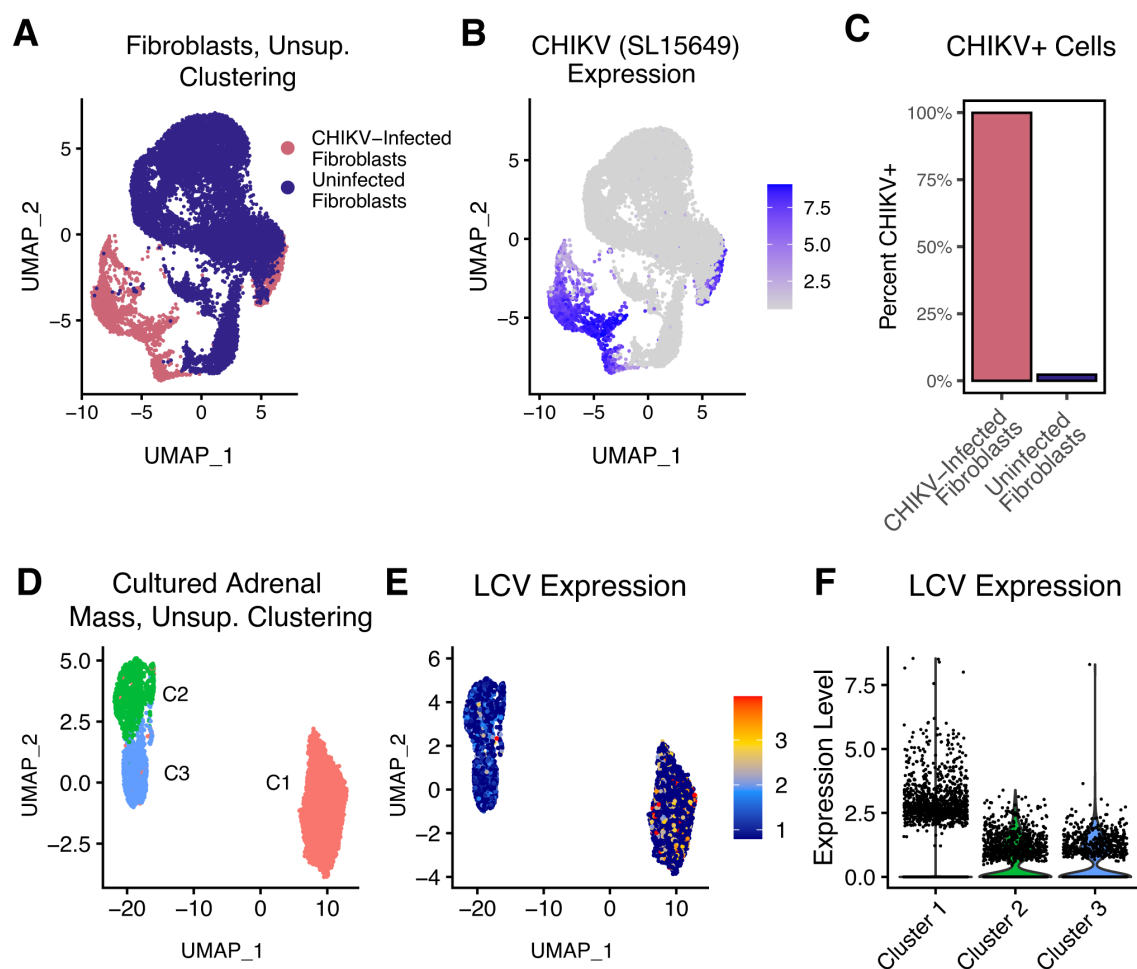


FIGURE 4

Quantification of viral RNA to illustrate detection of extra-genomic features. Panels (A–C) display the results of primary human fibroblasts infected with CHIKV at high MOI, followed by scRNA-seq. Uninfected fibroblasts were included as a control and processed in a physically separate lane. (A) The UMAP displays a dimensional reduction of these cells, colored by infection status. (B) The same UMAP as (A), colored by nimble-generated quantification of CHIKV RNA, demonstrating that CHIKV is specifically detected in CHIKV-infected fibroblasts and absent in uninfected controls. (C) The bar plot quantifies the percentage of CHIKV-positive cells for each fibroblast population from (A). Panels (D–F) display scRNA-seq data generated from B cells obtained from an adrenal mass identified in a cynomolgus macaque. (D) The UMAP displays a dimensional reduction of adrenal mass-derived B cells. (E) The same reduction as (D), colored by nimble-generated LCV expression data. LCV is a ubiquitous opportunistic virus that infects B cells and can induce lymphoma. (F) The violin plot quantifies the same nimble-generated LCV expression data shown in (E), demonstrating upregulation of LCV in B cell cluster 1. Collectively, these data provide two examples where nimble provides a simple solution to quantify transcripts not encoded by the host genome.

specific quantification for extra-genomic features, such as viruses. It also provides an example where iterative alignment against custom gene spaces could be advantageous over modifying the base genome and repeating the entire analysis.

## Resolution of complex, multigenic families such as NKG2 and KIRs

The examples shown thus far would be possible using standard scRNA-seq analysis pipelines, although there are situations when it might be more convenient or flexible to generate these data using nimble. There are nonetheless many gene families, particularly

those with gene duplication or variable copy number, where aligning data and calling features in one-size-fits-all logic creates artifacts. When aligning RNA-seq data to a reference, an important technical decision point, which is often obscured from the end-user, is whether to discard alignments that are mapped to multiple features. These ambiguous “multi-mapped” reads are often discarded in standard pipelines, which can result in the systematic loss of biologically important data. The NKG2 genes are a family of cell surface receptors expressed on NK cells and a subset of T cells (15, 16). This includes the inhibitory receptor NKG2A and the activating receptors NKG2C, and NKG2E. In concert with CD94, these receptors can recognize MHC-E ligands (16, 17). NKG2D is a separate activating receptor that binds

numerous MHC-I ligands (17, 18). These receptors are an important part of NK cell signaling and can modulate T cell signaling and thus quantifying these is important for the study of these cells. As is common among many expanded families of genes, there is a high degree of sequence similarity between NKG2C and E, which has the potential to result in multi-mapping of reads and loss of data. To examine the role of multi-mapping in standard RNA-seq pipelines, we obtained a reference scRNA-seq dataset consisting of T and NK cells from eight tissues and 47 rhesus macaques (19). We processed these data using the standard CellRanger/Mmul10 pipeline, along with nimble using a custom gene space containing NKG2 and KIR genes (Supplementary Table S1). Further, because we understood the biology of these genes differed from the genome as a whole, we executed nimble in a mode to allow and report multi-feature hits for these gene families. For NKG2A and NKG2D, which are both single-copy and have relatively unique sequences, we observe high concordance between nimble and the standard CellRanger/Mmul10 pipelines; however, there is a significant difference for NKG2C and NKG2E (Figure 5). Because NKG2C and NKG2E have high sequence similarity, many short Illumina reads will match both genes and no aligner will be capable of uniquely differentiating them; however, because NKG2C/E are functionally related activating receptors, there can be value in capturing and quantifying those ambiguous hits, especially if this results in a significant increase in data. This is exactly what we observe: by tolerating and reporting multi-mapped data, nimble recovers 45% more counts than the standard pipelines (Figure 4).

In addition to the NKG2 family, the killer immunoglobulin-like receptors (KIRs) are a well-characterized polygenic gene family also involved in NK and T cell signaling (7, 20, 21). KIRs are polygenic, with 23 genes present in rhesus macaques, which can be divided into activating and inhibitory KIRs (20, 21). This raises similar alignment/ambiguity issues as NKG2C/E; however, KIRs present a second layer of difficulty for standard pipelines. The gene number and gene content are variable between KIR haplotypes, which is a challenge when trying to use a 'one-size-fits-all' genome to represent the entire species. Because the reference genome represents one possible haplotype, individuals will encode different configurations of KIRs than represented in the reference. Further, polygenic regions with segmental duplications are notoriously difficult to sequence and are often poorly represented in reference genomes. The Mmul10 genome contains only 11 KIR genes and thus cannot provide accurate data for this gene family. We processed the same T and NK data using a gene space containing all published rhesus macaque KIR sequences (Supplementary Table S1). This reference gene space contained allele-level sequences; however, we executed nimble in a mode to aggregate results to the KIR gene level (Figure 5). As expected, KIRs are primarily expressed in NK cells, although expression was detected in gamma/delta ( $\gamma\delta$ ) and to a lesser degree alpha/beta ( $\alpha\beta$ ) T cells. KIR2DL4 and KIR3DS1 were the most detected KIRs. Collectively, these data illustrate the ability to nimble to recover data missed from standard pipelines by employing a more complete reference gene space and executing feature calling using logic more appropriate to the biology of the target genes.

## Characterization of NKG2 and KIR expression in NK and T cells

The enhanced NKG2 and KIR data obtained by nimble allow more detailed characterization of the expression patterns of these functionally important receptors. Because the data in Figure 5 are derived from a comprehensive single-cell atlas, they provide an ideal dataset in which to characterize expression patterns (19). Expression for NKG2 and KIR genes was variable by tissue, although this likely reflects the composition of the T and NK cells at these sites. For example, the lowest NKG2 and KIR expression was detected in lymph nodes, which are sites dominated by naive and central memory cells (Figure 5D). To resolve expression in more detail, we subset to effector-differentiated cells and performed dimensionality reduction (Figure 5). The subsets defined in Figure 5 are explained in greater detail in the original publication (19). As expected, the NKG2 and KIR expression is more common in NK cells relative to T cells. Most NK cells are positive for NKG2A, NKG2C, or both (Figure 5). CD16+ NK cells, which are more cytotoxic, have a higher fraction of NKG2C+ cells relative to the CD16- NK subset. T cells were more likely to be NKG2D+ without NKG2A/C/E; however, approximately 10% of effector memory T cells ( $T_{EM}$ ) express NKG2C/E. T follicular helper ( $T_{FH}$ ) cells had the highest fraction of NKG2D+ cells. Gamma/delta T cell subsets showed intermediate NKG2 expression relative to NK cells and  $\alpha\beta$  T cells, with MAIT-like T  $\alpha\beta$  cells and TRGV9+  $\gamma\delta$  T cells being the population with the largest fraction of NKG2A+ cells. KIR expression was also primarily detected in NK cells (Figure 5). Cells are more likely to express inhibitory KIRs alone, and a higher fraction of CD16- NK cells expressed KIRs relative to the CD16+ NK cells, the opposite of NKG2 expression. Among the T cell subsets,  $\gamma\delta$  T cells had intermediate levels of KIR positivity, and KIR expression was nearly absent from  $T_{FH}$  cells, and KIR expression is relatively rare in most  $\alpha\beta$  T cells. A notable exception was tissue resident memory T cells ( $T_{RM}$ ), which are  $\alpha\beta$  T cells differentiated for rapid response to antigenic stimulation (22–24). Nearly a third of these cells expressed KIRs, which is a significant difference from other  $\alpha\beta$  T cell subsets and even higher than CD16+ NK Cells. In total, these data illustrate additional information that can be gleaned from existing scRNA-seq datasets, by combining standard RNA-seq pipelines with targeted alignment and feature calling that are adapted to the biology of complex gene families.

## Quantifying major histocompatibility class I and II allelic expression

The major histocompatibility complex (MHC) is among the most polymorphic in the genome and presents multiple challenges for traditional RNA-seq analyses. Because of the high importance of MHC/HLA genotyping and the unique challenges, an entire field has emerged dedicated to MHC genotyping (5). The MHC is divided into class I and II loci. In humans, there are three MHC class I loci (termed human leukocyte complex or HLA): HLA-A, -B,

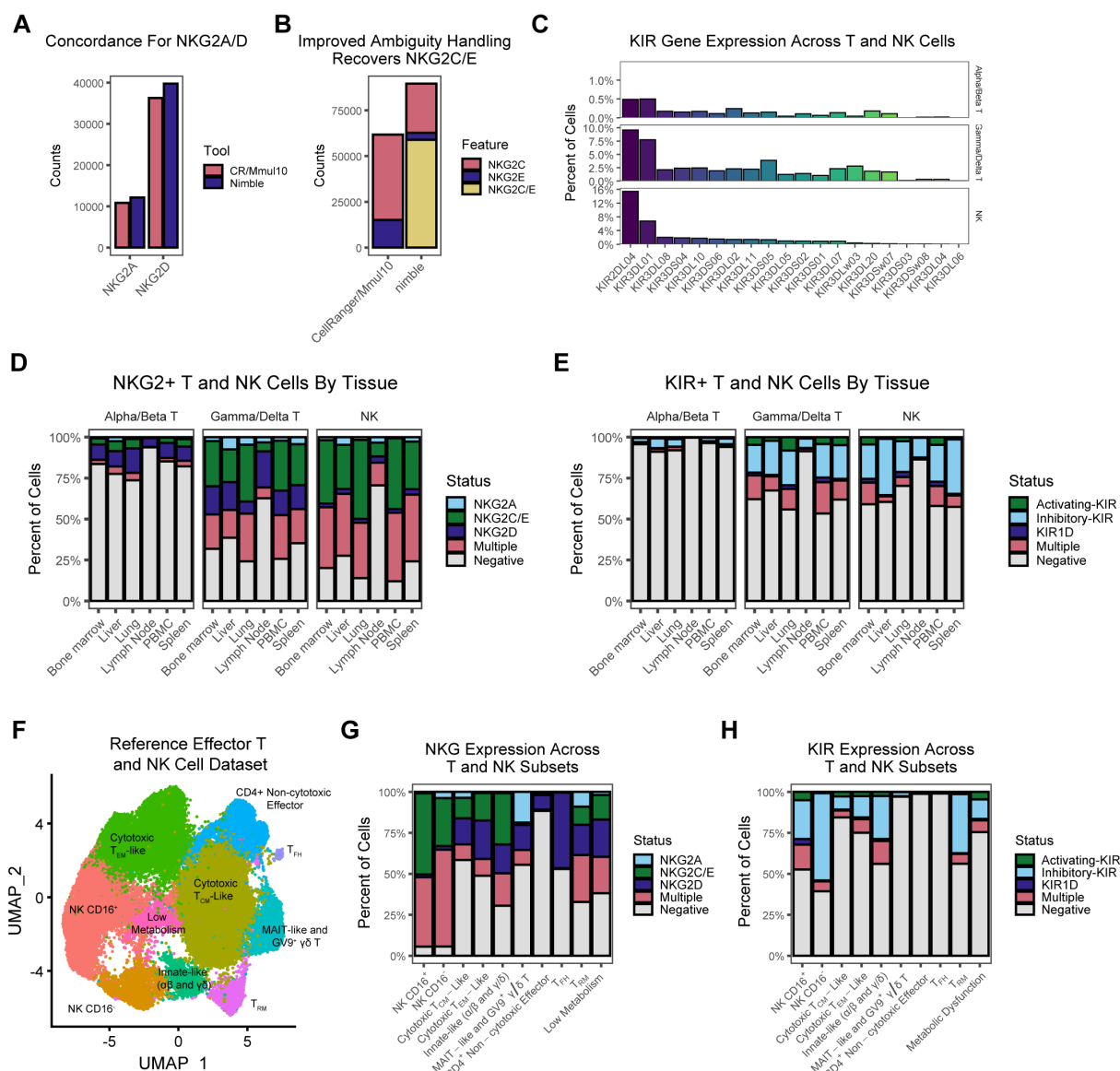


FIGURE 5

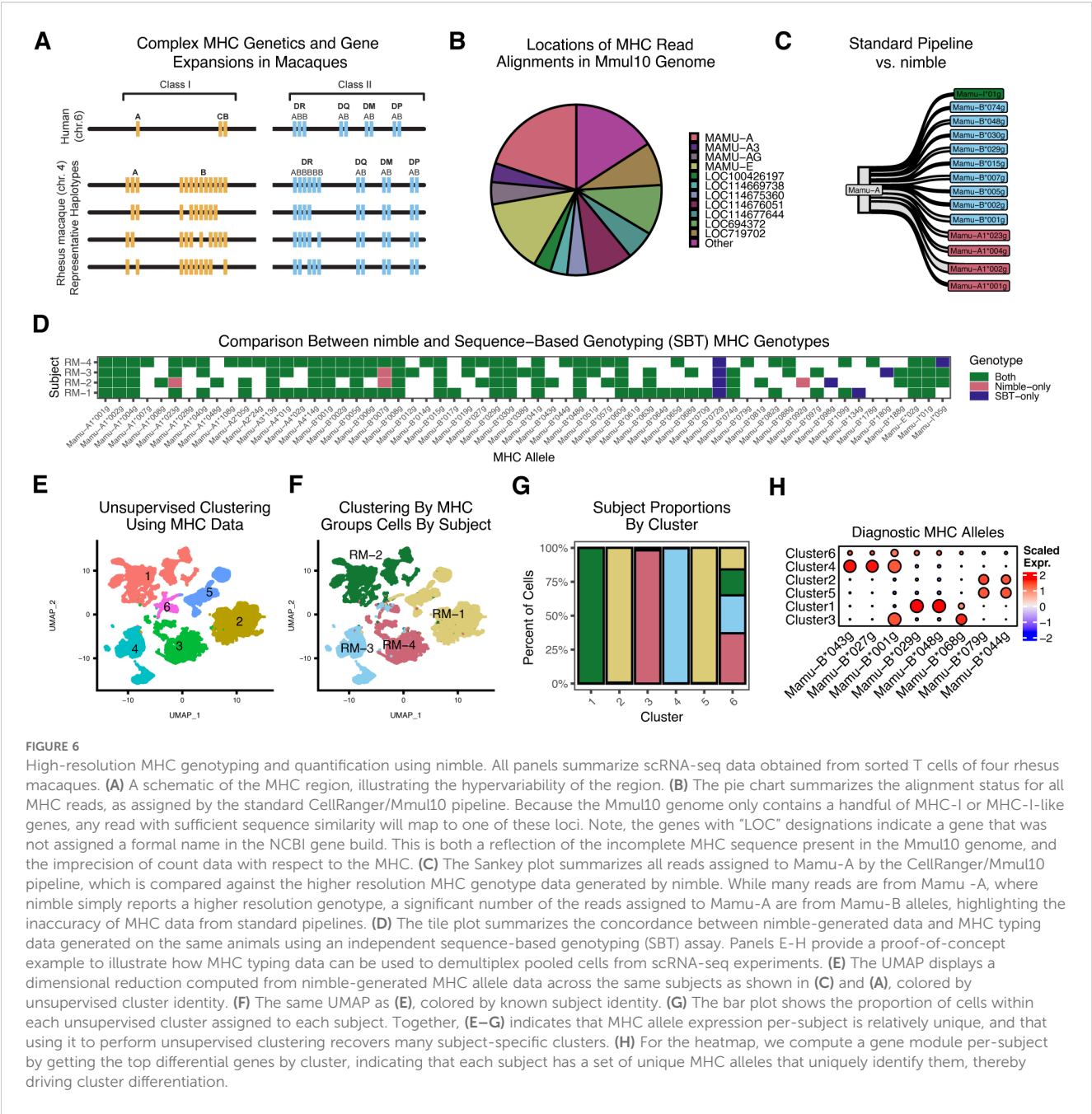
Summarization of NKG and KIR expression across T and NK populations. (A) The bar plot displays the magnitude of aggregated counts for *NKG2A* and *NKG2D*, demonstrating that nimble and CellRanger/Mmul10 exhibit similar behavior for these two features. (B) The bar plot displays the magnitude of aggregated counts for *NKG2C*, *NKG2E*, as well as reads that mapped to both *NKG2C* and *NKG2E*, which are two functionally similar genes with high sequence similarity. Critically, the difference in ambiguity resolution strategies between CellRanger/Mmul10 and nimble leads to a significant disparity in the number of counts generated for these features. Because nimble can be configured to retain and report ambiguous results, we can recover more counts for the activating *NKG2C/E* case than standard pipelines. (C) The bar plots display the percentage of cells across the RIRA T and NK cell types that are positive for various KIR genes. (D) The bar plots display the percentage of RIRA T and NK cells that express nimble-generated NKG data across various RIRA tissue types. (E) A similar set of bar plots as (D), displaying the percentage of RIRA T and NK cells that express nimble-generated KIR data across various RIRA tissue types. (F) The UMAP displays a dimensional reduction of a population of effector T and NK cells colored by RIRA subtype. (G) The bar plot represents the percent of cells that express NKG across the clusters we defined in (F). (H) A similar bar plot to (G), representing the percent of cells that express KIR across the clusters we defined in (F).

and -C; however, in macaques there are a variable number of MHC-A and -B loci, ranging from 20–24 distinct alleles per macaque (Figure 6) (25). The latter creates similar challenges as KIRs and presents a major problem for reference genome design. Further, most MHC loci have extremely high allelic diversity, with thousands of unique alleles per species (5). These allele-specific polymorphisms alter the peptide binding potential of MHC alleles, meaning that genotyping at high resolution is required. This is quite

distinct from the approach used for RNA-seq quantification of most genes, where pipelines are generally designed to ignore subject-specific polymorphism within a gene. Together, these present twin problems: 1) the single reference genome cannot adequately represent species-level diversity, and 2) genotyping at a much higher resolution is required than for typical genes.

To demonstrate the ability of nimble to generate high-resolution and accurate MHC typing from scRNA-seq data, we





processed scRNA-seq data from four rhesus macaques against a reference space with 2,379 MHC-I and MHC-II alleles (Supplementary Table S1). Unlike prior figures, nimble was run in a mode to report only perfect sequence matches, which is essential for the MHC, where nucleotide differences as little as a single base pair change alter peptide binding potential, necessitating high-resolution allele-level genotyping. While the database contained all known rhesus macaque MHC alleles, the resulting data were summarized by lineage (e.g. two-digit typing) (5, 26). We began by isolating the raw sequence reads nimble detected as aligned to MHC and compared the nimble result against the CellRanger/Mmul10 alignments. As noted above, the Mmul10 genome is incomplete and inaccurate across the MHC region, due

to the high complexity and polymorphism of that region. It nonetheless contains a handful of genes annotated as MHC or MHC-like. The majority of reads nimble identified as MHC were aligned to these loci (Figure 6). While MHC loci are highly polymorphic, polymorphism is clustered in the peptide-binding region with most regions of the gene being much more conserved. Because the Mmul10 reference genome only contains a handful of MHC loci, and these loci do not represent the diversity of the species, all reads from MHC will pile up against the best available target. This is shown in greater detail when comparing the reads assigned to the gene annotated as MHC-A in Mmul10 against the higher resolution genotyping generated by nimble (Figure 6). This shows that while many of the reads assigned to MHC-A by standard

pipelines are MHC-A alleles, a significant number of MHC-B alleles are also aligned to this gene. This highlights the fact that standard pipelines generate incorrect data from complex loci, in large part because they are simply not designed to differentiate this level of complexity. To evaluate the accuracy of nimble MHC data, we genotyped the subjects using an alternative sequence-based genotyping (SBT) assay, commonly employed for MHC genotyping of rhesus macaques (27). This assay involves PCR amplification of a small amplicon spanning the most variation portion of MHC-I alleles, using primers conserved across most MHC-I alleles. The genotypes obtained from each assay were overwhelmingly concordant, supporting nimble's accuracy. The small number of discrepancies can be explained by methodological differences. The most notable was the lack of Mamu-B\*072 detection in the nimble/scRNA-seq data. When we inspected the raw nimble results, Mamu-B\*072 alignments were detected; however, they were always ambiguous with different MHC alleles and thus discarded by nimble's featurizing calling logic. Because scRNA-seq data randomly samples regions of mRNA molecules, as opposed to the targeted sequencing of SBT, and because many MHC alleles are highly similar, it may not always be possible to resolve every allele. Depending on the goal of the analysis, nimble's alignment and scoring parameters could be adjusted. These data nonetheless demonstrate a resolution of MHC genotyping far superior to standard pipelines.

## High-resolution MHC typing from scRNA-seq can assign scRNA-seq transcriptomes to subject

Due to cost, it is common to pool samples in single-cell RNA-seq experiments. Multiple methods exist to demultiplex samples, including cell hashing reagents and genotype-based approaches (28–31). Due to the high subject-to-subject diversity of MHC, these MHC genotypes should provide a mechanism to assign cells to subjects as well. To test this, we performed dimensionality reduction and unsupervised clustering on the MHC genotypes generated by nimble, using the same four rhesus macaques, which identified six clusters (Figure 6). This method was effective at separating cells by subject (Figure 6F). Pairwise differential gene expression analyses between these clusters identified a handful of MHC alleles with cluster-specific expression patterns, which correspond to alleles uniquely expressed by one or two subjects (Figure 6). These data illustrate one practical usage for high-resolution MHC genotyping from scRNA-seq data.

## Differential regulation of individual MHC alleles following *Mtb* lysate exposure

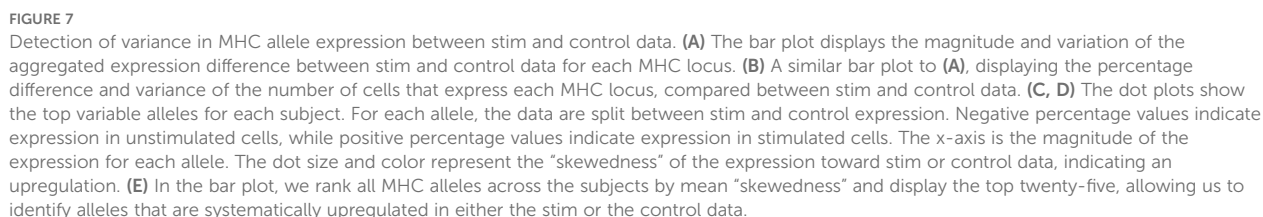
MHC expression can be altered in response to pathogens, although most data measure global regulation of all alleles from a given MHC locus, since allele resolution of expression has been difficult to measure (32, 33). To examine this, we generated a dataset

with bronchial alveolar lavage cells from a cohort of 12 rhesus macaques vaccinated against *Mycobacterium tuberculosis* (*Mtb*). The recovered cells were overwhelmingly macrophages. We divided these cells and exposed one fraction to *Mtb* lysate for 6 hours, with the remaining cells serving as a control, followed by scRNA-seq. We used nimble to quantify MHC-I and MHC-II expression levels. First, we quantified MHC at the locus level to identify any systematic changes across a whole MHC locus (Figures 7A, B). Data were quantified as either the mean expression level or the percent of cells expressing each locus. Across all subjects, we see changes in mean expression of <5% in *Mtb* stim or control unstimulated cells, with generally high variance, especially for MHC-B, suggesting expression may vary by allele. The percentage of cells expressing MHC loci varies by <10% in *Mtb* exposed or control unstimulated cells, except for Mamu-I. There was a small overall increase in the percentage of cells expressing any MHC-A and MHC-B alleles.

When the data are summarized at the allele-level, a more complex picture emerges (Figures 7C, D). First, the per-cell expression of each MHC allele varies heavily both at rest and post-stimulation. Certain MHC alleles are expressed by nearly all cells (e.g., Mamu-A1\*004), while some are only expressed by a small fraction of cells (e.g., Mamu-A1\*008 or Mamu-DRB1\*10). Second, individual alleles behave differently after *Mtb* exposure, with some alleles increasing significantly relative to controls, some unchanged, and some even decreasing (e.g., Mamu-DRB5\*03). When summarized across all rhesus macaques, Mamu-B\*053 showed the highest increase after exposure, while Mamu-DRB1\*10 showed the highest decrease (Figure 7E). These data, while proof-of-concept, demonstrate that there is high variability in expression at the level of MHC allele, and that allele-specific regulation occurs. This level of information is largely undetected by standard analysis pipelines. These changes could have implications for antigen selection in vaccines and could contribute to the protective effects of certain MHC alleles.

## Discussion

Dominant RNA-seq and scRNA-seq pipelines are designed to produce reliable quantification across the genome as a whole. Reads are typically aligned to a single reference genome that is intended to represent the genomic diversity of the entire species. While these can be very effective, there are gene families and genomic regions with characteristics that are problematic for standard pipelines, especially polygenic regions with differences in gene content between subjects. These include regions where it is extremely difficult for one reference genome to faithfully represent the diversity of the species (e.g. MHC/HLA or KIR), and situations where the species simply encodes multiple copies of highly similar genes (e.g. NKG2C/E). There are also more mundane situations where standard pipelines can fail or be sub-optimal, including detection of exogenous RNA (e.g., a virus) or simple errors in the gene model. Here, we presented nimble, a novel and flexible tool to address these gaps, especially where the underlying biology doesn't



We demonstrate the value of nimble through multiple examples. To validate accuracy, we demonstrate concordance between nimble and the standard CellRanger pipeline, using a set of typical single-copy, relatively conserved genes. Nimble can be

used to address technical issues in the genome or gene model, as shown for *CD27* and *IGHD*, or for the quantification of viral expression. Nimble is especially powerful for complex, polygenic gene families, because it can perform feature calling using settings more appropriate for each gene family. We show that nimble can recover otherwise discarded expression data for *NKG2* genes and used nimble to characterize *NKG2* and *KIR* expression in T and NK cells. These analyses identified previously unreported enrichment of inhibitory KIR expression in tissue resident memory T cells. Finally, nimble can resolve high-resolution MHC expression data, which revealed significant expression differences at the allele-level, and allele-specific changes after *Mtb* exposure. Collectively, these demonstrate a range of situations where nimble can augment standard RNA-seq pipelines to recover potentially valuable data.

There are many efforts to improve the quality of reference genomes, including the new generation of so-called telomere-to-telomere (T2T) genomes (34, 35). While these may help with some of the issues outlined here, they will not address them all. While newer builds may provide contiguous sequence across the entire chromosome, they still represent the entire genomic diversity of the species with one reference. This is not adequate for highly polymorphic regions, especially when gene content differs between subjects. Newer genomic representations, such as pangenome graphs, seek to represent multiple haplotypes in a single reference and could address some of this (36). While the latter is unquestionably a more biologically appropriate way to represent intra-species variation, this is a significant change, and new generations of software will be required to take advantage of this new genomic representation. Finally, even if a perfect representation of genomic diversity existed, some gene families have characteristics that require customized logic in the scoring of gene and allele-level counts.

Nimble, as the name suggests, is designed to be lightweight and flexible. We presented a set of examples where it is useful, but others may exist. One potential use-case is quantification of isoforms, in which case the reference might contain a handful of isoforms for the gene of interest. This supplemental alignment and count data makes it feasible to identify previously missed patterns of expression across diverse species and cell types.

## Methods

### Nimble aligner

The data presented here were processed using a novel toolchain developed for the purpose of aligning RNA-seq data to arbitrary reference spaces. Nimble provides various facilities for curating these custom reference libraries, aligning sequence data, and reporting properties of the alignment data for the purpose of quality control. The tool takes RNA-seq data in a variety of formats and a set of custom reference libraries as input and produces one count matrix per library. To create a custom gene space, the user can provide a set of Entrez identifiers, a CSV, or a FASTA file. The nimble library file produced allows the user to

customize values for aligner filter behavior, such as minimum read length for a passing alignment, the maximum allowable mismatches, or sequence trimming strictness, among several others. The tool and detailed documentation about its usage and configuration options is available on GitHub (<https://github.com/BimberLab/nimble>).

Nimble incorporates a previously developed, multithreaded pseudoalignment algorithm to align RNA-seq data to these custom gene spaces (10). First, if the input is a bam file, it is sorted using samtools (37). Then, the custom gene space reference sequences are used to construct a Debruijn graph, which allows for rapid and accurate quantification of counts per input gene without needing to perform expensive positional alignment. All input RNA-seq data is trimmed with a reimplementation of Trimmomatic's MAXINFO algorithm (38) and aligned to the input gene spaces, producing either counts-per-gene matrices or counts-per-molecule matrices, depending on the format of the input data. Ambiguous alignments to two or more features may be kept or discarded, depending on the user settings.

The nimble alignment pipeline provides several additional layers of filtration for the alignment count matrix, depending on the format of the input data. All alignments are subject to alignment length, mismatch, and trimming filters. In the case of paired-end input reads, there are optional filters for asserting read-pair alignment orientations relative to the reference space, and several options for producing a single set of calls from differing alignments between two sequences in the same read-pair. Finally, nimble can transform the counts-per-molecule matrix produced from 10x scRNA-seq input data into a counts-per-cell matrix by intersecting on the molecule and cell barcodes to conform to the expected data format for downstream packages like Seurat.

### Animal subjects

All study macaques were housed at the Oregon National Primate Research Center (ONPRC) in animal biosafety level 2 rooms with autonomously controlled temperature, humidity, and lighting. Macaques were fed commercially prepared primate chow twice daily and received supplemental fresh fruit or vegetables daily. Fresh, potable water was provided via automatic water systems. During all protocol time points, body weight and complete blood counts were collected and animals underwent anesthesia support and monitoring. The ONPRC Institutional Animal Care and Use Committee approved macaque care and all experimental protocols and procedures. The ONPRC is a Category I facility. The American Association for Accreditation of Laboratory Animal Care fully accredits the Laboratory Animal Care and Use Program at the ONPRC. It has an approved assurance (no. A3304-01) for the care and use of animals on file with the National Institutes of Health Office for Protection from Research Risks. The Institutional Animal Care and Use Committee adheres to national guidelines established in the Animal Welfare Act (7 U.S. Code, sections 2131–2159) and the Guide for the Care and Use of Laboratory Animals, Eighth Edition, as mandated by the U.S. Public Health Service Policy.



## Tissue collection and processing

Cell isolation from PBMC and solid tissues were acquired and processed to single-cell suspensions using previously published methods, summarized below (5, 19). Liver, spleen, and mesenteric lymph node biopsies were collected by a minimally invasive laparoscopic procedure (39). For lung samples, animals were humanely euthanized, and caudal lung lobe samples were collected during necropsy. Bone marrow cells were harvested from the humerus or iliac crest by flushing with R10 media. Peripheral blood mononuclear cells (PBMCs) were isolated from freshly collected ACD-A treated blood utilizing Ficoll-Paque density centrifugation (20). Lymph nodes (LN) and spleen samples were homogenized as previously described (40) while liver and lung samples were enzymatically digested with DNase and collagenase (41, 42). Prior to processing, cells were filtered using 70  $\mu$ m strainers. Cells were quantified using a Countess II (Thermo Fisher), aliquoted, diluted as required for single-cell RNA sequencing (typically 500–1,500 cells/ $\mu$ L), and kept on ice prior to processing. Mononuclear cells were isolated from bronchoalveolar lavage (BAL) as previously described (43).

## Cell hashing

Cell hashing was used for most scRNA-seq samples, with the MULTI-Seq lipid labeling system (31), using commercially available lipid modified oligos (Sigma Aldridge LMO001). Cells were labeled with barcoded lipids as follows: 15  $\mu$ L MultiSeq solution 1 (LMO stock, diluted in PBS to 400nM) was added, along with 45  $\mu$ L of the barcode solution (10  $\mu$ M barcode oligo, diluted in PBS to 400nM), giving a final working concentration of 200nM for LMO and 200nM for the barcode oligo. Next, pipet mix and incubate for 5 min at 4°C. Add 10  $\mu$ L of the MultiSeq co-anchor solution (50  $\mu$ M Co-A stock, diluted in PBS to 2  $\mu$ M), then pipet mix and incubate for 5 min at 4°C. Wash twice with 1 mL cold PBS, spinning cells at 700 g for 5 min at 4°C, and then resuspend in 200  $\mu$ L R10 (which will quench LMOs). Samples were pooled, followed by GEM generation on the 10x instrument.

## Single-cell RNA sequencing

The isolated single cell suspensions were then processed for single-cell RNA sequencing using the 10x Genomics Chromium platform, using 5' v2 or HT chemistry, following the manufacturer's protocols, including feature barcoding library preparation. To improve capture of MULTI-Seq fragments, we added the following primer, 5'-CCTTGGCACCCGAGAATTCC-3', at 2.5  $\mu$ M to the 10x cDNA synthesis step. Generation of VDJ enriched libraries followed manufacturer's instructions with the exception that macaque-specific TCR constant region primers were used in place of human-specific TCR enrichment primers for macaque cells (19). Primer pairs were used to amplify the alpha,

beta, delta, and gamma TCR chains. The concentration of the alpha constant region primer was increased relative to the beta primer to improve amplification. PCR conditions for both reactions were as follows: lid temp 105°C, 98°C 0:45, 12 cycles of: 98°C 0:20, 60°C 0:30, 72°C 1:00, followed by 72°C 1:00 and 4°C hold. Sequence libraries were sequenced using Illumina chemistry, on either Novaseq or HiSeq instruments (Illumina).

## Single-cell RNA-seq processing

Raw sequence reads were processed using 10X Genomics Cell Ranger software (version 8.0.1). The resulting sequence data were aligned to the MMul\_10 genome (assembly ID: GCF\_003339765.1) with NCBI gene build 103. Cell demultiplexing used a combination of algorithms, including GMM-demux, demuxEM and BFF, implemented using the cellhashR package (30, 44, 45). Droplets identified as doublets (i.e. the collision of distinct sample barcodes) were removed from downstream analyses. We additionally performed doublet detection using DoubletFinder, and removed doublets from downstream analysis (46). Data were otherwise processed as previously described (19). Analyses utilized the Seurat R package, version 4.2 (47). Multiple scRNA-seq datasets are used, including many previously published datasets. A complete listing of the SRA accession numbers for datasets generated for this manuscript are available in [Supplementary Table S2](#). Adrenal mass B cell samples were obtained from a previously described case (12). The Rhesus Macaque Immune Atlas (RIRA) dataset was used for multiple analyses, with the input gene expression data from NIH GEO database, accession GSE277821, and NIH BioProject PRJNA1163395 (19).

## Major histocompatibility complex analysis

Genotyping for Major Histocompatibility Complex class I (MHC-I) allele was performed using a PCR amplicon-based method, as previously described<sup>78,79</sup>. For nimble-generated MHC data, normalization was performed per cell by dividing the raw reads for each MHC allele by the sum of reads from each MHC locus (i.e. total MHC-A, total MHC-B, etc.).

## Chikungunya virus infection

Primary normal human dermal fibroblasts (NHDFs) were experimentally infected with Chikungunya virus strain SL-15649, obtained from Dr. Mark Heise (University of North Carolina at Chapel Hill). NHDFs were plated into 6-well plates, cultured in DMEM containing 10% FBS and 1X PSG, and incubated overnight at 37°C with 5% CO<sub>2</sub>. Cells were infected in triplicate wells with CHIKV SL-15649 at a multiplicity of infection equal to 1. At 24 hours post infection the cells were trypsinized and washed twice with DMEM-10 and once with PBS.



## Mycobacterium lysate exposure assay

Mononuclear cells isolated from bronchoalveolar lavage (BAL) fluid were incubated at 37°C under a humidified 5% CO<sub>2</sub> atmosphere. These cells were rested for 4 hours, and then either exposed to Mtb lysate (BEI NR-14822 at 6uL/Test) for 4 hours or cultured without Mtb lysate as a control. After incubation, cells were processed using the 10x Genomic Chromium system, as described above.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/[Supplementary Material](#).

## Ethics statement

Ethical approval was not required for the studies on humans in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used. The animal study was approved by ONPRC Institutional Animal Care and Use Committee. The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

SB: Formal analysis, Methodology, Software, Writing – original draft. GM: Formal analysis, Methodology, Software, Writing – review & editing. MK: Formal analysis, Methodology, Software, Writing – review & editing. GB: Formal analysis, Methodology, Software, Writing – review & editing. BV-M: Investigation, Resources, Writing – review & editing. SO: Investigation, Resources, Writing – review & editing. SF: Investigation, Resources, Writing – review & editing. WG: Investigation, Resources, Writing – review & editing. CN: Investigation, Resources, Writing – review & editing. DD: Investigation, Resources, Writing – review & editing. AS: Investigation, Resources, Writing – review & editing. TB: Investigation, Resources, Writing – review & editing. AB-A: Investigation, Resources, Writing – review & editing. NH: Investigation, Resources, Writing – review & editing. HW: Investigation, Resources, Writing – review & editing. CW: Investigation, Resources, Writing – review & editing. CB: Investigation, Resources, Writing – review & editing. JVS: Investigation, Resources, Writing – review & editing. CL: Investigation, Resources, Writing – review & editing. MA: Investigation, Resources, Writing – review & editing. RR: Funding acquisition, Writing – review & editing. DS: Funding acquisition, Writing – review & editing. JBS: Funding acquisition, Writing – review & editing. AO: Funding acquisition, Writing – review & editing. SH: Funding acquisition, Investigation, Resources, Writing – review & editing. LP:

Funding acquisition, Writing – review & editing. BB: Conceptualization, Methodology, Software, Supervision, Writing – original draft.

## Funding

The author(s) declare financial support was received for the research and/or publication of this article. This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID) grants and contracts 75N93019C00070 (to LP), P01AI177688-01 (to LP) R01AI161010 (to RR), R01AI129703 (to JBS), U19 AI142759-01 (to DS), as well as Bill and Melinda Gates Foundation grant INV-002377 (to LP), INV-055706 (subaward to BB), and the Oregon National Primate Research Center Core grant from the National Institutes of Health, Office of the Director (P51OD011092). The research reported in this publication used computational infrastructure supported by the Office of Research Infrastructure Programs, Office of the Director, of the National Institutes of Health under Award Number S10OD034224.

## Acknowledgments

We also thank Dr. Katinka Vigh-Conrad for assistance with figures.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2025.1596760/full#supplementary-material>

### SUPPLEMENTARY FIGURE 1

Comparison of nimble to CellRanger for genome-wide alignment. While nimble is not designed to completely replace standard alignment and feature

calling pipelines, to provide a more comprehensive comparison of nimble with standard pipelines we generated a nimble library containing the complete 15,782 genes defined in the MMuL10 genome, and compared the resulting per-cell counts against the same data processed with CellRanger/MMuL10. The scatter plot presents the counts for each gene obtained using nimble relative to the CellRanger pipeline. Results were highly concordant, with a Pearson correlation of 0.968 (Supplementary Figure 1). Together, these data indicate that nimble's alignment pipeline captures similar count data to standard pipelines, establishing nimble's accuracy when aligning to a straightforward gene space.

## References

- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
- Sullivan DK, Min KHJ, Hjorleifsson KE, Luebbert L, Holley G, Moses L, et al. kallisto, bustools, and kb-python for quantifying bulk, single-cell, and single-nucleus RNA-seq. *Nat Protoc*. (2025) 20(3):587–607. doi: 10.1038/s41596-024-01057-0
- Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*. (2022) 38:2943–5. doi: 10.1093/bioinformatics/btac166
- Parham P. MHC class I molecules and KIRs in human history, health and survival. *Nat Rev Immunol*. (2005) 5:201–14. doi: 10.1038/nri1570
- Robinson J, Barker DJ, Marsh SGE. 25 years of the IPD-IMGT/HLA database. *HLA*. (2024) 103:e15549. doi: 10.1111/tan.15549
- Nomura T, Matano T. Association of MHC-I genotypes with disease progression in HIV/SIV infections. *Front Microbiol*. (2012) 3:234. doi: 10.3389/fmicb.2012.00234
- Parham P, Guethlein LA. Genetics of natural killer cells in human health, disease, and survival. *Annu Rev Immunol*. (2018) 36:519–48. doi: 10.1146/annurev-immunol-042617-053149
- Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SG. IMGT/HLA Database—a sequence database for the human major histocompatibility complex. *Nucleic Acids Res*. (2001) 29:210–3. doi: 10.1093/nar/29.1.210
- Walker BD, Korber BT. Immune control of HIV: the obstacles of HLA and viral diversity. *Nat Immunol*. (2001) 2:473–5. doi: 10.1038/88656
- Genomics, x. Rust Pseudoaligner. Available online at: <https://github.com/10XGenomics/rust-pseudoaligner> (accessed April 01, 2025).
- Wu YC, Kipling D, Dunn-Walters DK. The relationship between CD27 negative and positive B cell populations in human peripheral blood. *Front Immunol*. (2011) 2:81. doi: 10.3389/fimmu.2011.00081
- Wu HL, Weber WC, Shriver-Munsch C, Swanson T, Northrup M, Price H, et al. Viral opportunistic infections in Mauritian cynomolgus macaques undergoing allogeneic stem cell transplantation mirror human transplant infectious disease complications. *Xenotransplantation*. (2020) 27:e12578. doi: 10.1111/xen.12578
- Marr-Belvin AK, Carville AK, Fahey MA, Boisvert K, Klumpp SA, Ohashi M, et al. Rhesus lymphocryptovirus type 1-associated B-cell nasal lymphoma in SIV-infected rhesus macaques. *Vet Pathol*. (2008) 45:914–21. doi: 10.1354/vp.45-6-914
- Wu HL, Weber WC, Waytashek CM, Boyle CD, Reed JS, Bateman KB, et al. A model of lymphocryptovirus-associated post-transplant lymphoproliferative disorder in immunosuppressed Mauritian cynomolgus macaques. *PLoS Pathog*. (2024) 20:e1012644. doi: 10.1371/journal.ppat.1012644
- Carville A, Evans TI, Reeves RK. Characterization of circulating natural killer cells in neotropical primates. *PLoS One*. (2013) 8:e78793. doi: 10.1371/journal.pone.0078793
- Walter L, Petersen B. Diversification of both KIR and NKG2 natural killer cell receptor genes in macaques - implications for highly complex MHC-dependent regulation of natural killer cells. *Immunology*. (2017) 150:139–45. doi: 10.1111/imm.12666
- Wroblewski EE, Parham P, Guethlein LA. Two to tango: co-evolution of hominid natural killer cell receptors and MHC. *Front Immunol*. (2019) 10:177. doi: 10.3389/fimmu.2019.00177
- Siemaszko J, Marzec-Przyślak A, Bogunia-Kubik K. NKG2D natural killer cell receptor-A short description and potential clinical applications. *Cells*. (2021) 10(6):1420. doi: 10.3390/cells10061420
- Rhesus Immune Reference Atlas (RIRA): A multi-tissue single-cell landscape of immune cells(2025). Available online at: <https://github.com/BimberLab/RIRA> (accessed April 01, 2025).
- Bimber BN, Evans DT. The killer-cell immunoglobulin-like receptors of macaques. *Immunol Rev*. (2015) 267:246–58. doi: 10.1111/imr.12329
- Robinson J, Guethlein LA, Maccari G, Blokhuis J, Bimber BN, de Groot NG, et al. Nomenclature for the KIR of non-human species. *Immunogenetics*. (2018) 70(9):571–83. doi: 10.1007/s00251-018-1064-4
- Kumar BV, Ma W, Miron M, Granot T, Guyer RS, Carpenter DJ, et al. Human tissue-resident memory T cells are defined by core transcriptional and functional signatures in lymphoid and mucosal sites. *Cell Rep*. (2017) 20:2921–34. doi: 10.1016/j.celrep.2017.08.078
- Bromley SK, Akbaba H, Mani V, Mora-Buch R, Chasse AY, Sama A, et al. CD49a regulates cutaneous resident memory CD8(+) T cell persistence and response. *Cell Rep*. (2020) 32:108085. doi: 10.1016/j.celrep.2020.108085
- Walzer T, Marcais A, Saltel F, Bella C, Jurdic P, Marvel J. Cutting edge: immediate RANTES secretion by resting memory CD8 T cells following antigenic stimulation. *J Immunol*. (2003) 170:1615–9. doi: 10.4049/jimmunol.170.4.1615
- Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE. Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res*. (2004) 14:1501–15. doi: 10.1101/gr.2134504
- Robinson J, Malik A, Parham P, Bodmer JG, Marsh SG. IMGT/HLA database—a sequence database for the human major histocompatibility complex. *Tissue Antigens*. (2000) 55:280–7. doi: 10.1034/j.1399-0039.2000.550314.x
- Wiseman RW, Karl JA, Bimber BN, O'Leary CE, Lank SM, Tuscher JJ, et al. Major histocompatibility complex genotyping with massively parallel pyrosequencing. *Nat Med*. (2009) 15:1322–6. doi: 10.1038/nm.2038
- Boggy G, Bimber BN. cellhashR: An R package designed to demultiplex cell hashing data(2021). Available online at: <https://github.com/bimberlab/cellhashr> (accessed April 1, 2025).
- Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*. (2018) 19:224. doi: 10.1186/s13059-018-1603-1
- Xin H, Lian Q, Jiang Y, Luo J, Wang X, Erb C, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. *Genome Biol*. (2020) 21:188. doi: 10.1186/s13059-020-02084-2
- McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods*. (2019) 16:619–26. doi: 10.1038/s41592-019-0433-8
- Greene JM, Wiseman RW, Lank SM, Bimber BN, Karl JA, Burwitz BJ, et al. Differential MHC class I expression in distinct leukocyte subsets. *BMC Immunol*. (2011) 12:39. doi: 10.1186/1471-2172-12-39
- Ting JP, Trowsdale J. Genetic control of MHC class II expression. *Cell*. (2002) 109 Suppl:S21–33. doi: 10.1016/s0092-8674(02)00696-7
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. (2022) 376:44–53. doi: 10.1126/science.abj6987
- Zhang S, Xu N, Fu L, Yang X, Ma K, Li Y, et al. Integrated analysis of the complete sequence of a macaque genome. *Nature*. (2025) 640(8059):714–21. doi: 10.1038/s41586-025-08596-w
- Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. (2022) 604:437–46. doi: 10.1038/s41586-022-04601-8
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170
- Zevin AS, Moats C, May D, Wangari S, Miller C, Ahrens J, et al. Laparoscopic technique for serial collection of liver and mesenteric lymph nodes in macaques. *J Vis Exp*. (2017) (123):55617. doi: 10.3791/55617

40. Pitcher CJ, Hagen SI, Walker JM, Lum R, Mitchell BL, Maino VC, et al. Development and homeostasis of T cell memory in rhesus macaque. *J Immunol.* (2002) 168:29–43. doi: 10.4049/jimmunol.168.1.29
41. Kauffman KD, Sallin MA, Sakai S, Kamenyeva O, Kabat J, Weiner D, et al. Defective positioning in granulomas but not lung-homing limits CD4 T-cell interactions with Mycobacterium tuberculosis-infected macrophages in rhesus macaques. *Mucosal Immunol.* (2018) 11:462–73. doi: 10.1038/mi.2017.60
42. Burwitz BJ, Wettengel JM, Muck-Hausl MA, Ringelhan M, Ko C, Festag MM, et al. Hepatocytic expression of human sodium-taurocholate cotransporting polypeptide enables hepatitis B virus infection of macaques. *Nat Commun.* (2017) 8:2146. doi: 10.1038/s41467-017-01953-y
43. Hansen SG, Zak DE, Xu G, Ford JC, Marshall EE, Malouli D, et al. Prevention of tuberculosis in rhesus macaques by a cytomegalovirus-based vaccine. *Nat Med.* (2018) 24:130–43. doi: 10.1038/nm.4473
44. Gaublomme JT, Li B, McCabe C, Knecht A, Yang Y, Drokhlyansky E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun.* (2019) 10:2907. doi: 10.1038/s41467-019-10756-2
45. Boggy GJ, McElfresh G, Mahyari E, Ventura AB, Hansen SG, Picker LJ, et al. BFF and cellhashR: analysis tools for accurate demultiplexing of cell hashing data. *Bioinformatics.* (2022) 38(10):2791–801. doi: 10.1093/bioinformatics/btac213
46. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* (2019) 8:329–337.e324. doi: 10.1016/j.cels.2019.03.003
47. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* (2018) 36:411–20. doi: 10.1038/nbt.4096