eLIFE elifesciences.org

RESEARCH ARTICLE

Figures and figure supplements

Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in *Drosophila*

Evan Witt *et al*

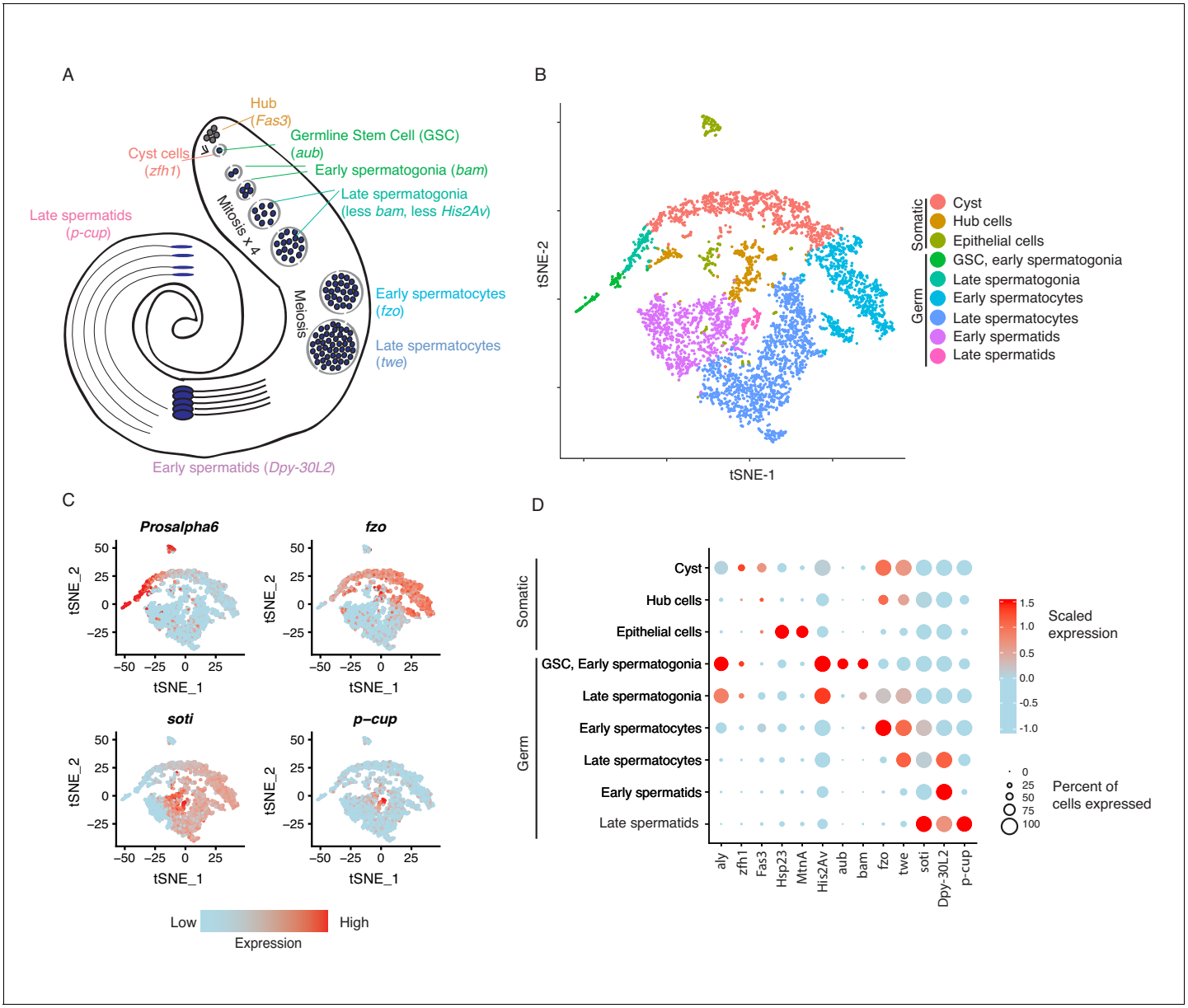

Figure 1. Clustering and cell-type assignment of single cells in Seurat. (**A**) An illustration of the major cell types in the testis, and the marker genes we used to identify them are in brackets. Somatic cells are hub, cyst, and epithelial cells. Spermatogenesis begins with germline stem cells which undergo mitotic divisions to form spermatogonia. These become spermatocytes which undergo meiosis and differentiate into spermatids. (**B**) A t-SNE projection of every cell type identified in the data. (**C**) Examples of marker genes that vary throughout spermatogenesis. *His2Av* is most active in early spermatogenesis, *fzo* and *soti* are active in intermediate and late stages, respectively, and *p-cup* is exclusively enriched in late spermatids. (**D**) Dotplot of scaled expression of marker genes in each inferred cell type. The size of each dot refers to the proportion of cells expressing a gene, and the color of each dot represents the calculated scaled expression value; blue is lowest, red is highest. 0 is the gene's mean scaled expression across all cells and the numbers in the scale are z scores. The cutoffs shown here were chosen to emphasize cell-type-specific enrichment of key marker genes. The genes used to assign each cell type are detailed in the Materials and methods section.

DOI: https://doi.org/10.7554/eLife.47138.002

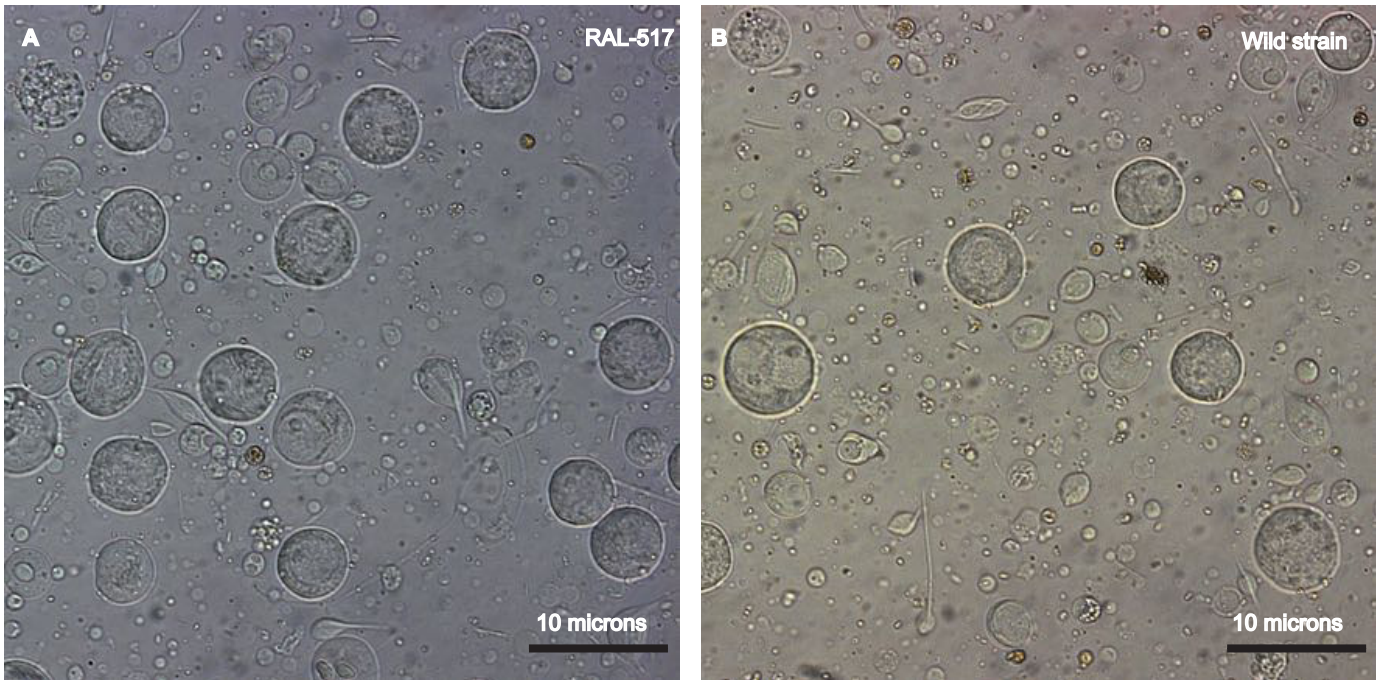

Figure 1—figure supplement 1. Establishing a single cell suspension from *Drosophila* testes. Representative images of single cell suspensions from two *D. melanogaster* strains, RAL 517 (**A**) and our lab's wild strain (**B**). Dissected testes were treated with proteases followed by straining and washes (see Materials and methods). The images contain single cells of various developmental stages, some with tails of various lengths. Cells were imaged using a 40X magnification and scale bars represent 10 μm. Since cells are present in many focal planes, the size of some cells may not correspond to the scale bar shown.

DOI: https://doi.org/10.7554/eLife.47138.003

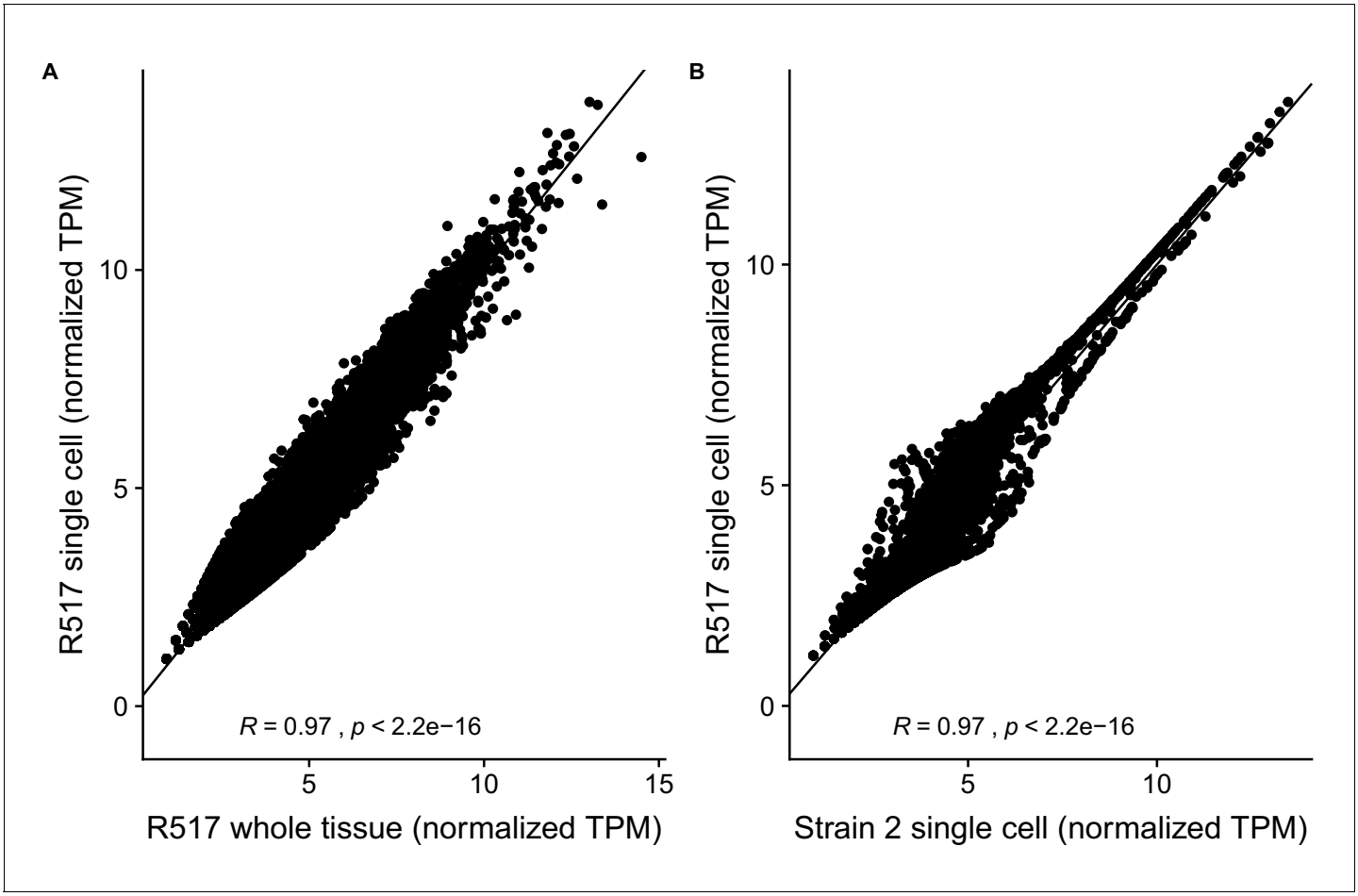

Figure 1—figure supplement 2. Reproducibility of RAL517 single-cell sequencing data. (**A**) Correlation between DESeq2 regularized log-transformed TPM (transcripts per million) of genes in 517 whole-tissue RNA-seq data and our single-cell RNA-seq data. Despite different library strategies and sequencing methods our data correlate extremely well with whole-tissue RNA-seq data indicating that our dataset has captured an accurate sampling of testis-expressed genes and our results are reproducible. (**B**) Correlation between our RAL517 single-cell library and a library of a wild *D. melanogaster* strain from our lab. Despite being from different strains, the libraries show a high correlation in normalized TPM. This result shows, however, that many genes vary between *D. melanogaster* strains, necessitating further work to understand transcriptome evolution on a single-cell level.

DOI: https://doi.org/10.7554/eLife.47138.004

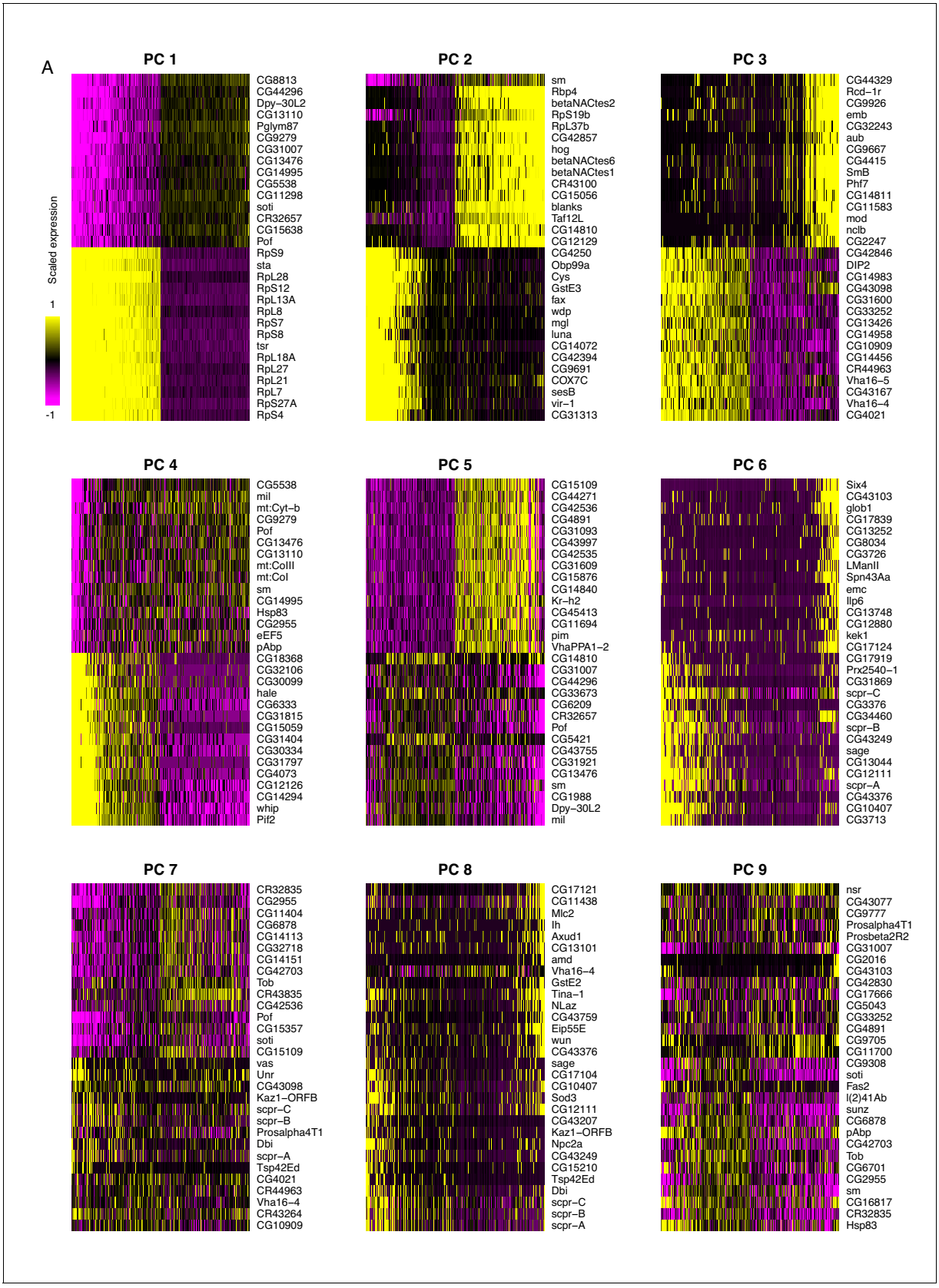

Figure 1—figure supplement 3. Principal component analysis of testis-expressed genes. Horizontally, each line is 500 randomly selected cells, and vertically, the expression of the 15 genes with the highest positive and negative scores for the principal component. For PC 1, one interpretation could

Figure 1—figure supplement 3 continued on next page

Figure 1—figure supplement 3 continued

be that *soti*, a marker of late spermatocytes/early spermatids, is negatively correlated with the expression of many ribosomal protein genes. This is consistent with our finding that ribosomal protein genes peak in early spermatogenesis. It is worth noting that the higher numbered PCs become more and more diffuse, as they each explain a smaller proportion of variance than the PCs 1 and 2.

DOI: https://doi.org/10.7554/eLife.47138.005

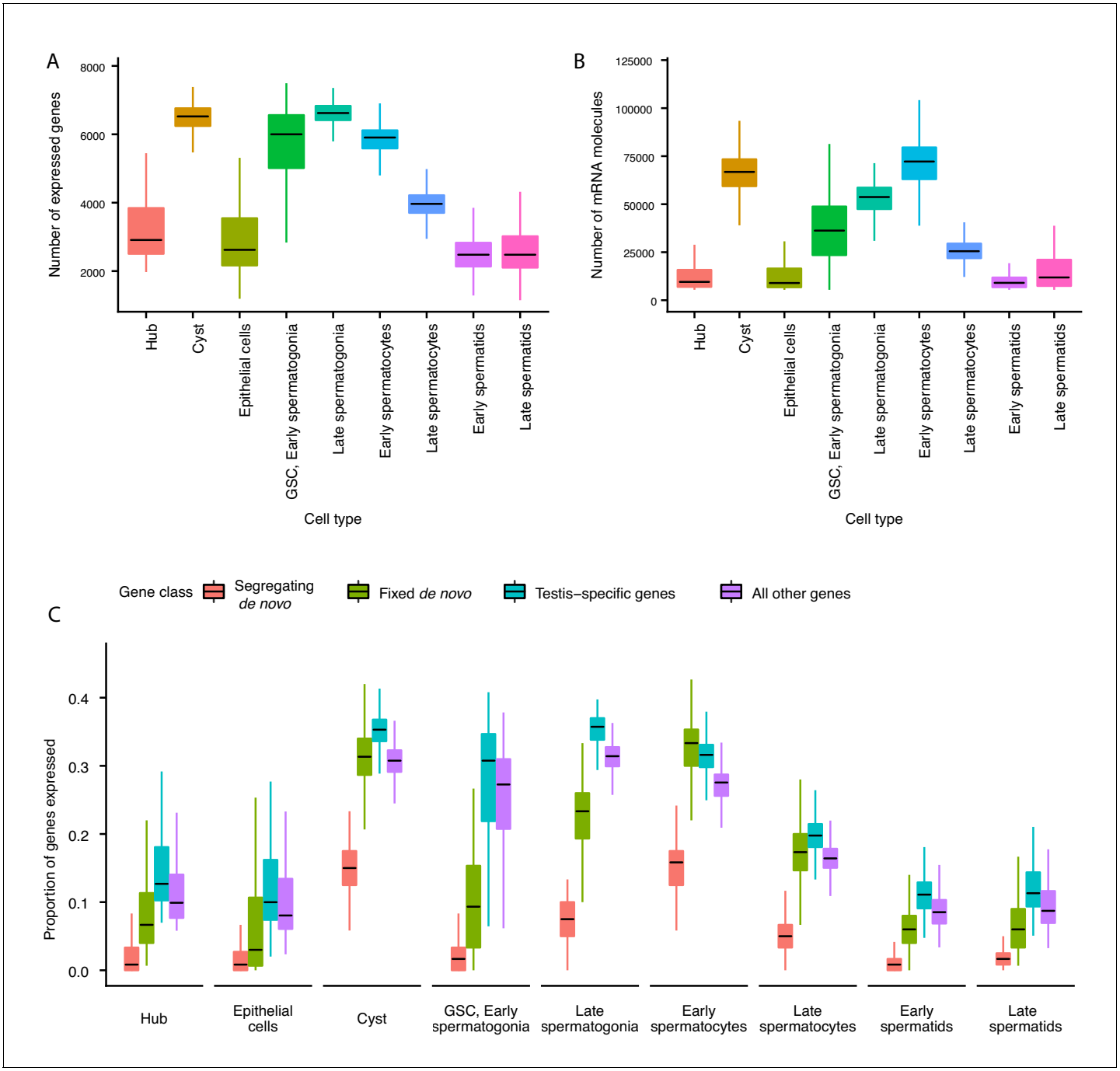

Figure 2. Gene expression and RNA content through spermatogenesis. (**A**) Boxplots of the number of genes expressed in each cell, binned by assigned cell type. Late spermatogonia and early spermatocytes express the most genes, and spermatids the least. (**B**) The number of Unique Molecular Indices (UMIs) detected for each cell, a proxy of RNA content. By this metric RNA content peaks in early spermatocytes, and is reduced thereafter by post-meiotic transcriptional suppression. (**C**) The proportion of segregating de novo, fixed de novo, testis-specific, and all genes expressed in every cell. For each cell, we counted the number of each class of gene with non-zero expression and divided it by the total number of genes of that type, grouping by cell type. For every cell type except spermatocytes, segregating de novo genes are the least commonly expressed, fixed de novo genes are more commonly expressed and all genes are most commonly expressed. In every cell type except early spermatocytes, a smaller proportion of fixed de novo genes are expressed than testis-specific genes, but early and late spermatocytes express similar proportions of fixed de novo genes and testis-specific genes. It is important to note that this measure looks at the number of genes of each type detected in a cell, not the expression level of each, and does not distinguish between high and low expression.

DOI: https://doi.org/10.7554/eLife.47138.006

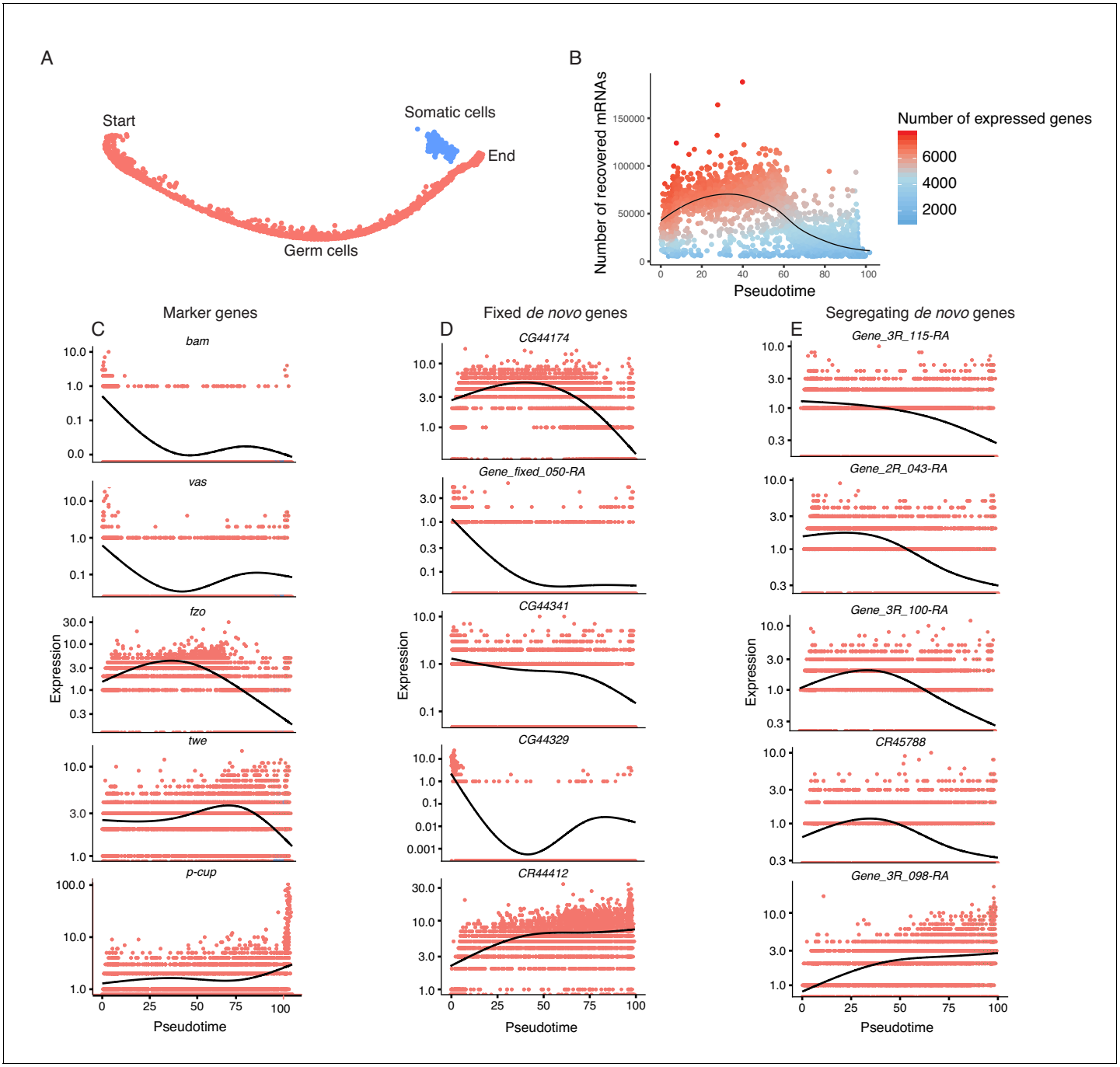

Figure 3. Pseudotime approximates the developmental trajectory of spermatogenesis. (**A**) We aligned every cell from our testis sample along an unsupervised developmental trajectory. From the expression of marker genes, we found somatic cells (blue) which were forced onto the developmental trajectory. For further analysis we disregard this branch (See Materials and methods, ***Figure 3—figure supplement 1***). Spermatogenesis begins at the far-left end of the trajectory. (**B**) The relative RNA content per cell peaks in mid-spermatogenesis, and declines during spermatid maturation, as approximated by the number of UMIs detected per cell. The number of genes expressed declines as well. The black line is a Loess-smoothed regression of the data, which should be thought of as a general trend among stochastic data and not a mathematical model. (**C**) Loess-smoothed expression of marker genes along the red germ cell lineage assigned in panel A. Along this lineage, the relative expression of marker genes is consistent with their temporal dynamics inferred from previous work. (**D**) Fixed de novo genes show a variety of expression patterns, including biphasic, early-biased, and late-biased. (**E**) Segregating de novo genes are often biased towards early/mid spermatogenesis.

DOI: https://doi.org/10.7554/eLife.47138.007

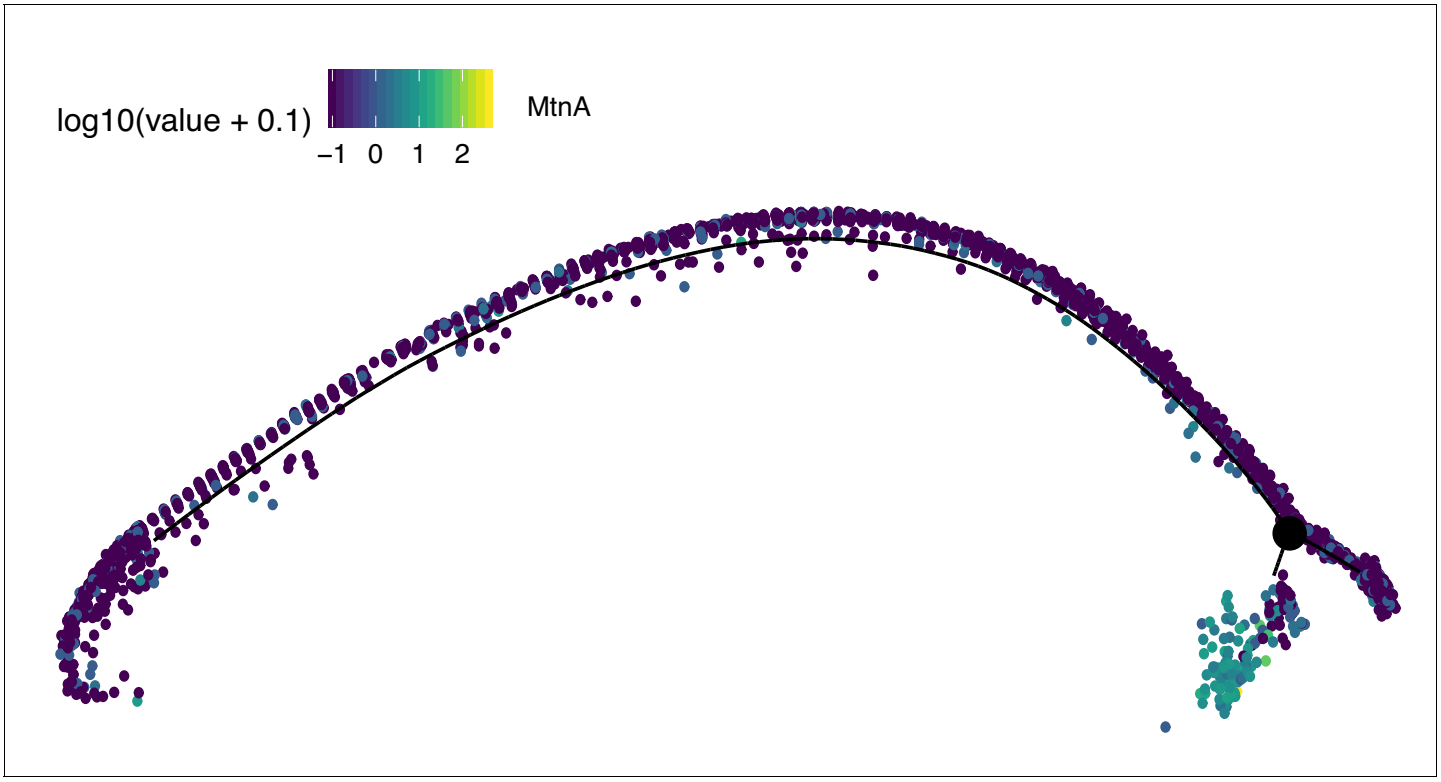

Figure 3—figure supplement 1. Assignment of somatic branch of pseudotime trajectory. This is the same pseudotime developmental trajectory from ***Figure 3A***, but each cell has been colored according to its expression of *MtnA*, a marker of somatic cells. This led us to conclude that state three in ***Figure 2A*** is mostly somatic cells and is not part of the germ lineage since it is enriched in *MtnA*.

DOI: https://doi.org/10.7554/eLife.47138.008

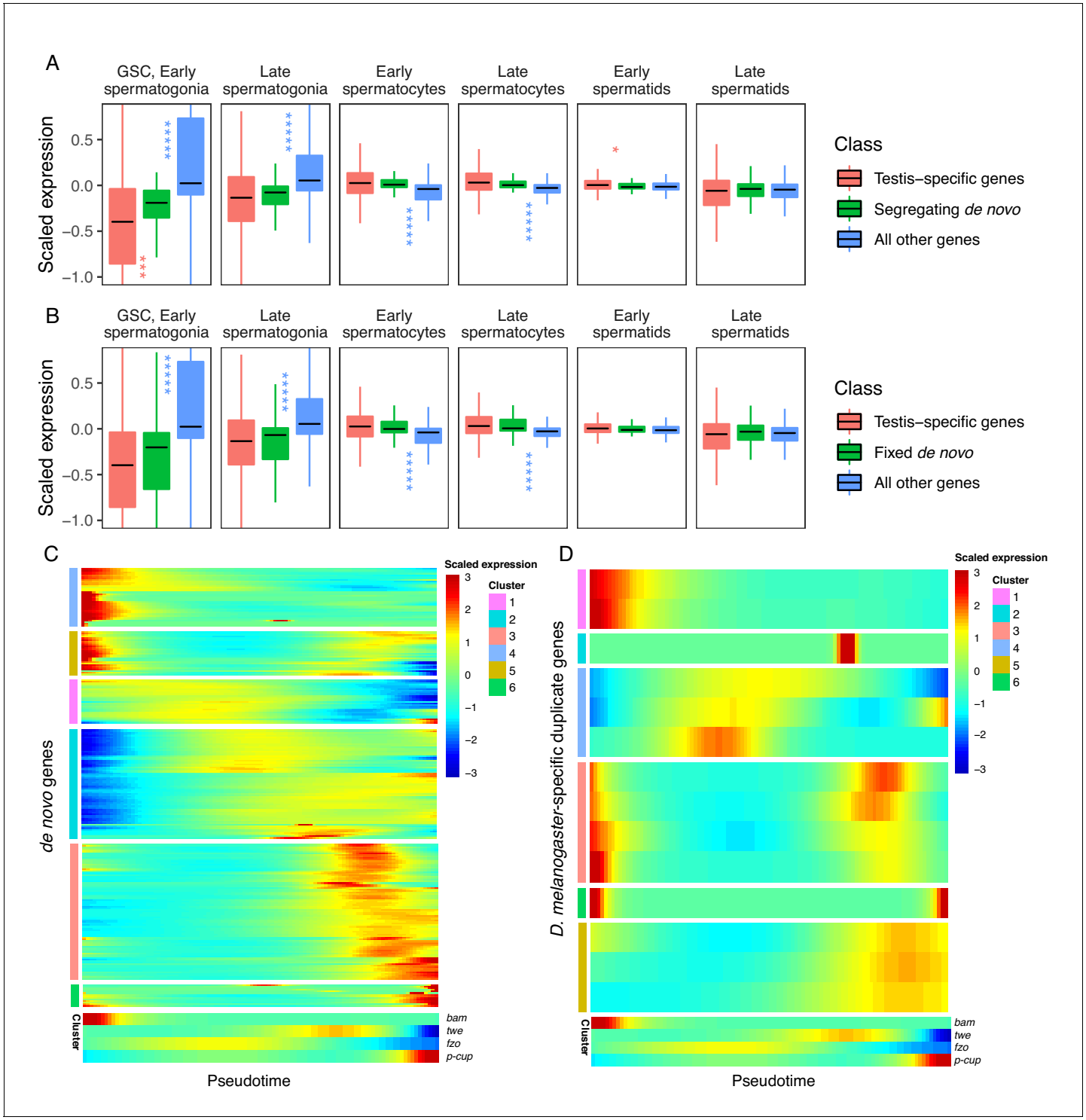

Figure 4. Expression bias of young genes. Spermatogenesis starts at GSC, early spermatogonia and proceeds rightward. (**A**) The scaled expression distribution of segregating de novo genes in each cell type, compared with the distribution of every other gene and testis-specific genes. For every gene, 0 is its mean scaled expression in a cell type, and the Y axis corresponds to Z scores of deviations higher or lower than that mean value. Asterixis represent Hochberg-corrected p values. The color of the asterixis indicates which gene set is being compared to de novo genes, and their placement above or below the boxplots indicates that gene set's relationship (higher or lower) to de novo genes. By this measure, de novo genes are biased downwards in early spermatogenesis and upwards in early spermatids. (**B**) The scaled expression patterns of fixed de novo genes are typical of testis-specific genes. (**C**) The scaled expression of detected fixed de novo genes across pseudotime (left to right), clustered by monocle's plot_pseudotime_heatmap function. While most de novo genes are biased towards intermediate cell-types, a small portion of de novo genes are most

Figure 4 continued on next page

Figure 4 continued

expressed during early and late spermatogenesis. (**D**) The scaled expression of *melanogaster*-specific duplicate genes over pseudotime. Despite being a similar evolutionary age to fixed de novo genes, young duplicate genes are more likely to be biased towards early and late spermatogenesis.

DOI: https://doi.org/10.7554/eLife.47138.009

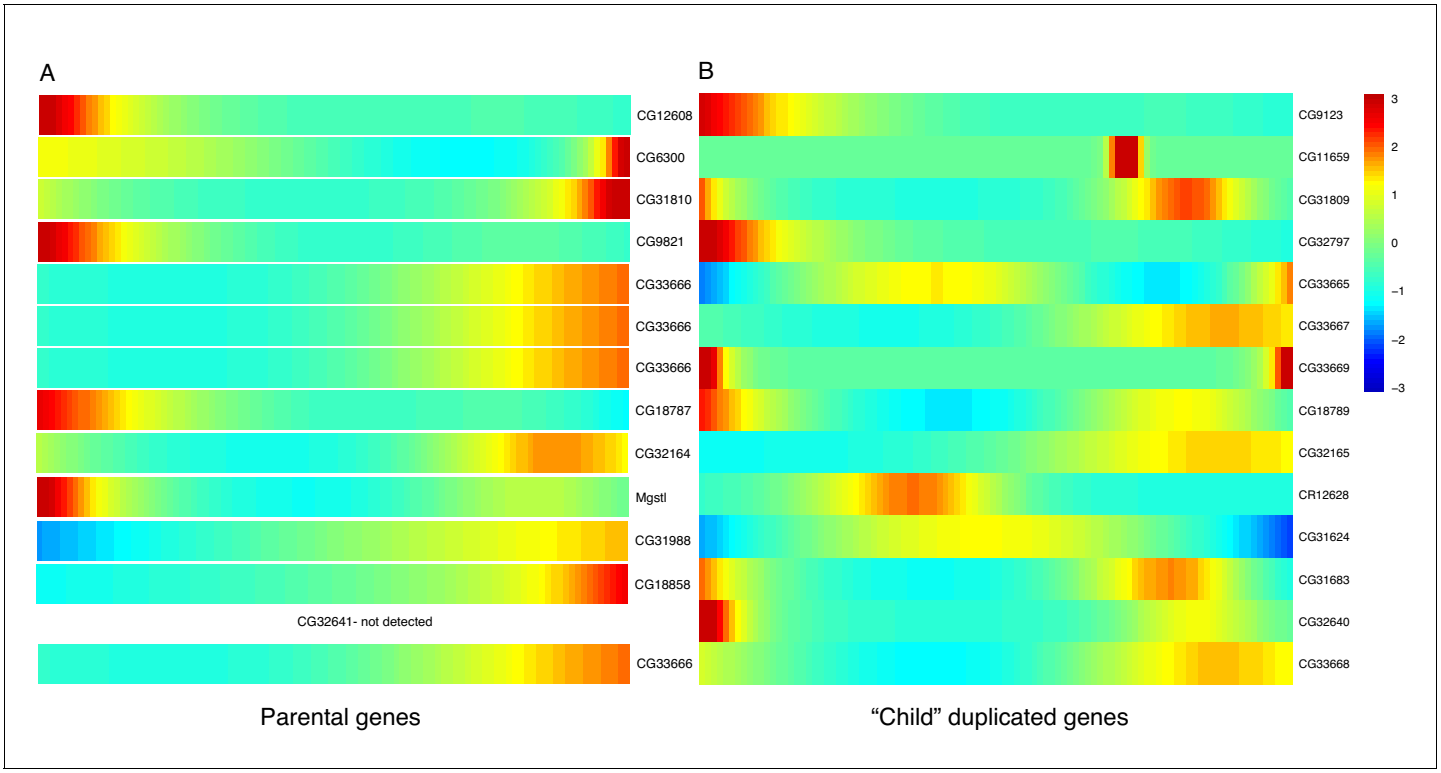

Figure 4—figure supplement 1. Expression heatmaps of parental and derived duplicated genes. A row is a gene, shown as it progresses along pseudotime from left to right. (**A**) The scaled expression of a set of parental copies of duplicated genes, plotted in pseudotime. (**B**) The scaled expression of the derived copies of duplicated genes.

DOI: https://doi.org/10.7554/eLife.47138.010

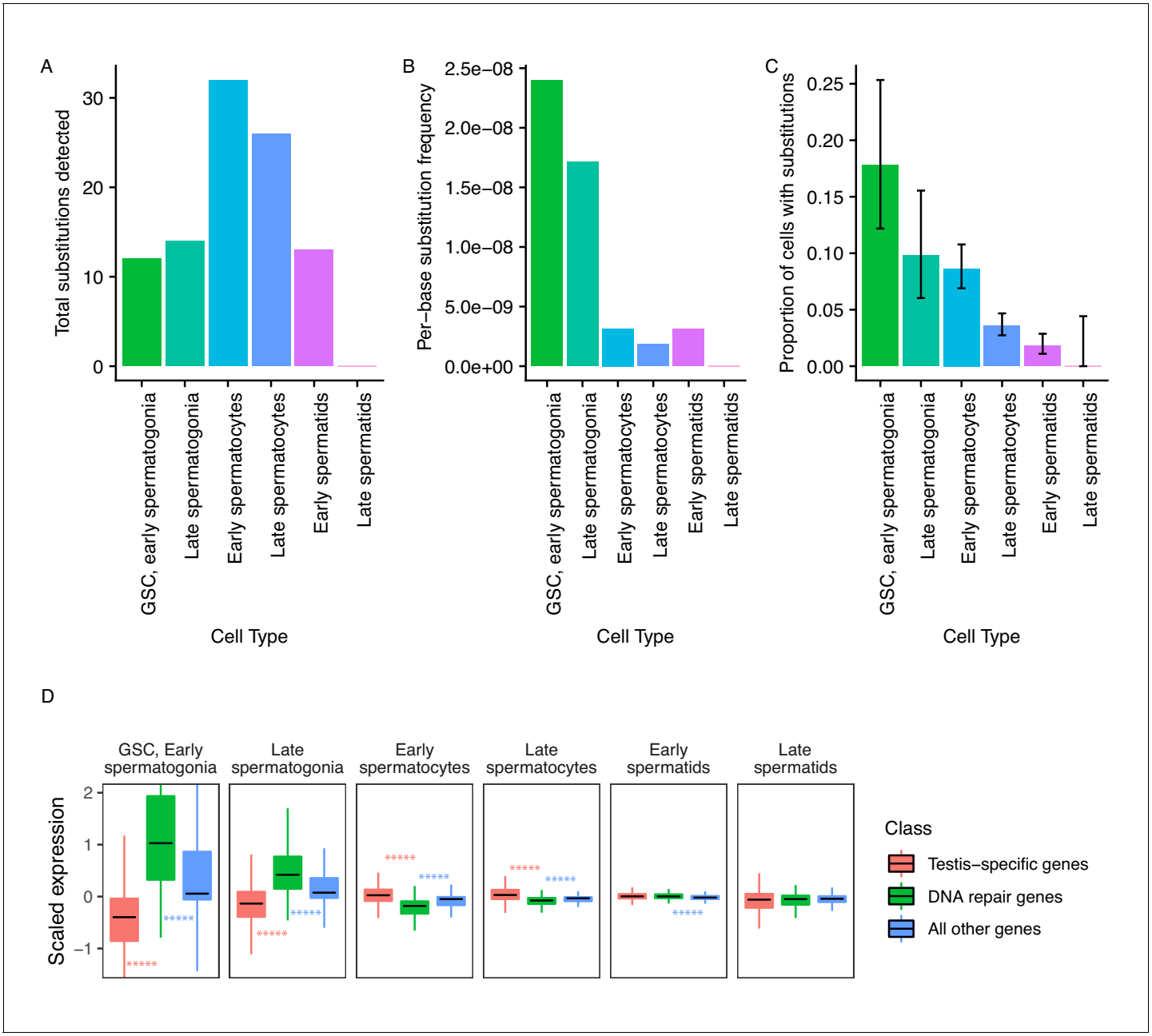

Figure 5. Abundance of putative de novo germline mutations. (**A**) For every cell type, the total number of high-quality polymorphisms identified. Out of 2590 candidate variants, we excluded all substitutions that could be found in any somatic cell, leaving 73 variants. We then counted clustered polymorphisms as single mutational events and removed variants that could have resulted from RNA editing. See Materials and methods for details. (**B**) Dividing the number of polymorphisms in a cell type by the number of cells of that type, and the number of bases covered with at least 10 reads in that cell type (***Supplementary file 5***) yields an approximate relative substitution frequency for each cell type. By this metric, substitutions are most prevalent in early spermatogenesis, and decrease in relative abundance during spermatid development. This could be due to the apoptosis of mutated cells, or the systematic repair of DNA lesions during spermatogenesis. (**C**) The proportion of cells of each type with at least one identified germline lesion. Error bars are the 95 percent confidence intervals for each proportion. A Chi-square test for trend in proportions gives a p value of 2.20E-16, indicating strong evidence of a linear downward trend. (**D**) DNA repair genes are generally biased towards early spermatogenesis, statistically enriched compared to the distribution of all other genes. (Wilcoxon adjusted p value < 0.05).

DOI: https://doi.org/10.7554/eLife.47138.013

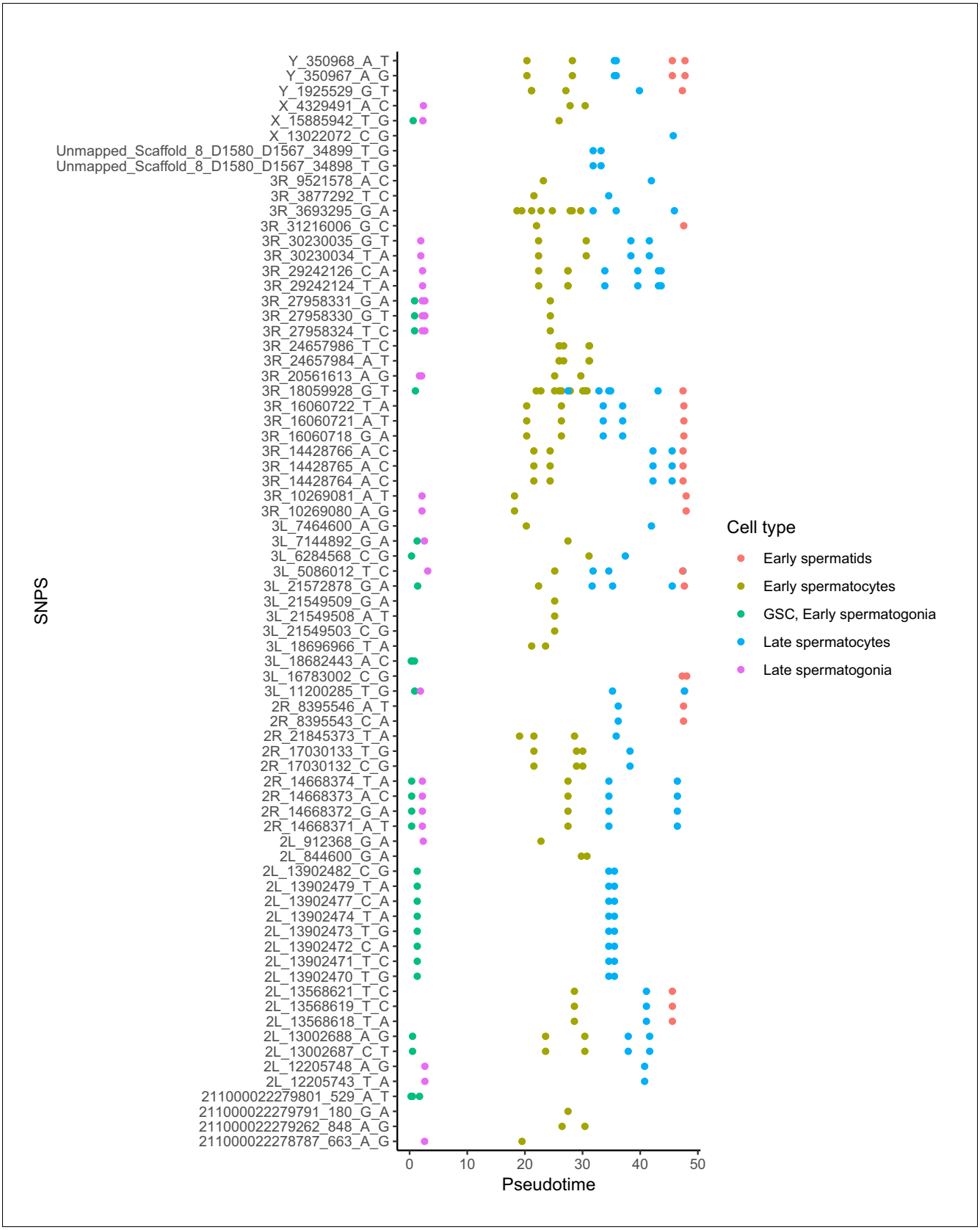

Figure 5—figure supplement 1. Alignment of germline mutations along pseudotime. For the 73 germline SNPS, we plotted every cell containing the SNP according to its pseudotime value inferred by monocle. Some SNPs are found in multiple cell types. Other SNPs are actually clustered close to

Figure 5—figure supplement 1 continued on next page

Figure 5—figure supplement 1 continued

other SNPS found in the same cells, such as the SNPS from 2L_13902470 to 2L_13902482. For the purposes of calculating mutational abundance in ***Figure 5***, we considered clusters of SNPS within 10 bp of each other to be a single mutational event, to prevent clusters of SNPs from biasing our inferred mutational abundance. Raw numbers of SNPs per cell type and corrected mutational events for each cell type are available in ***Supplementary file 5***.

DOI: https://doi.org/10.7554/eLife.47138.014