Bioinformatics Laboratory Project

By Michal Alayev
Under the guidance of Dr. Omri Wurtzel

Abstract

The planarian *Schmidtea mediterranea* has extraordinary regenerative ability, and its genome has been extensively researched. The whole planarian body can be regenerated from a single neoblast cell, hence it follows that all its cells contain the same DNA sequence. Evolutionary processes could cause genetic heterogeneity among cells of the same organism, and this angle was never examined before. We used single-cell RNA sequencing data, processed it into a digital expression matrix that counts the number of transcripts for each gene, in each cell.
We collected good-quality mismatches data per position in selected genes, and analyzed it in comparison to bulk RNA sequencing data, in order to verify any trends we may notice. We found positions in genes sequences that had mismatch percentage of 15%-35% the most interesting, suspected to be genetically heterogeneous among cells, because the nucleotide substitution kind and frequency at them cannot be explained easily with known genomic phenomena. We found at least 25 candidate positions that can be looked deeper into, that the mismatches at them can be sorted by the cells they were found in, and reach further conclusion.

Introduction

The planarian *Schmidtea mediterranea* reproduces asexually, by detaching its tail end, and each half regrows the lost parts by regeneration. The neoblasts divide and differentiate, thus resulting in two worms. Since the whole planarian body can be regenerated from a single neoblast, the planarian way of reproduction implies that all of the planarians cells contain the same DNA sequence, generation by generation, starting from the first planarian. Since a lot of cells are divided from the same neoblast, it is possible that throughout all the cell divisions that occurred throughout millions of years, some undifferentiated cells accumulated mutations in their DNA sequence that did not affect functionality. These cells continued to divide, creating a population of cells that contain the same mutations in their DNA, and also creating new planarians that carry the mutation in some of their cells, as a result of their way of reproduction. This way, non-harmful mutations could get fixed, leading to the possibility that different cells of the same planarian might contain different nucleotides at the same position in their DNA.
This assumption leads to our hypothesis that there can be some genetic heterogeneity among cells in the same organism. Our goal was to find out if such genetic heterogeneity exists, by finding positions that demonstrate this heterogeneity among different cells. This can be done

first of all by looking for positions in the transcriptomic reads where at least two different nucleotides, with non-background frequencies, appear in the reads mapped to them.

Methods and Results

First Part - processing Drop-seq sequence data into a digital expression matrix (DGE):

We used single-cell RNA sequencing data of adult Planarians *Schmidtea mediterranea*, from a body section located directly below eyes, enriched for ventral half (Brain 1), which was priorly produced using the SCS method Drop-seq (1). The paired-end fastq data was acquired using the SRA Tool fastq-dump with the option --split-files, on the desired Drop-seq run ID (SRR6829404). A single bam file was generated from the fastq files using the Picard Tool FastqToSam. This bam file was processed using the 'Drop-seq core computational protocol', v1.2, as developed by Jim Nemesh in the McCaroll lab (3) with some modifications, in order to process and align the sequencing data. Briefly, reads were tagged with their associated cell and molecular barcodes, followed by the trimming of 5' primer and 3' polyA sequences. The processed sequences were mapped using bowtie2 to the dd_Smed_v6 transcriptome assembly (asexual Schmidtea mediterranea strain CIW4) (2), with the following 4 sequences added to the transcriptome assembly fasta file for mapping: SMED_11901_V2, dd_Smed_v6_0_0_1, mtRNA_1, and mtRNA_2. The output from alignment was sorted in queryname order, and then merged with the unaligned bam file that had been previously tagged with molecular/cell barcodes, to recover these tags that were "lost" during alignment. Gene/exon annotation tags were added using a GTF file created from the dd_Smed_v6 transcriptome assembly using a python script. The expected number of cells in the run was estimated by extracting the number of reads per cell barcode using BAMTagHistogram module, and then plotting the cumulative distribution of number of reads per cell, using R script. The number of cells at the inflection point (which was 4000 for the data we used), was used as the estimated number of cells that were sequenced. Using the DetectBeadSynthesisErrors module [NUM_BARCODES=2X expected cell number, PRIMER_SEQUENCE=AAGCAGTGGTATCAACGCAGAGTAC], errors in barcode sequence associated with bead synthesis were detected and were either corrected, if possible, or the associated reads were removed. A gene expression matrix was generated for the expected number of cells using the module DigitalExpression [NUM_CORE_BARCODES=expected cell number] (1).

This entire process is done by the following scripts: 'Drop-seq_alignment_before_R.sh', 'plot_cumdist.R', 'Drop-seq_alignment_after_R.sh', 'create_gtf.py' (SM 1).

We examined the correctness of the outputted DGE considering the assumption that cells highly expressing markers of a specific cell type, should not express high-expressed markers of other cell types. Due to this assumption, we performed the following test: cells that co-express the epidermal markers dd_Smed_v6_332_0_1 and dd_Smed_v6_69_0_1, should not express the muscle marker dd_Smed_v6_323_0_1. The results of this test on our created DGE showed

that out of 1341 cells (33.53% of all cells represented in the DGE) that co-express the epidermal markers, 1327 cells (98.96%) do not express the muscle marker, as desired, which indicates the consistency of our DGE and the correctness of the process we performed to create it.

Second Part - collecting mismatches data per position in selected genes

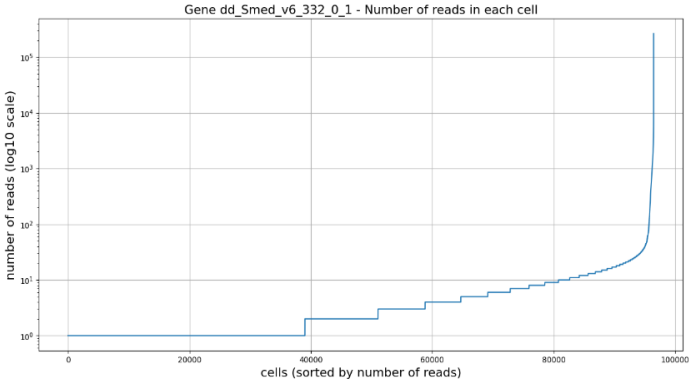
We manually selected numerous genes from the planarian transcriptome that met the following requirements: (i) the gene is associated to a specific tissue with high power (according to table S2 in digiworm (1)), (ii) the DGE contains many reads mapped to the gene or many cells in the DGE contain such reads, and (iii) the gene's contig identifier prefix is below 1000, if possible. For example, we chose the gene dd_Smed_v6_69_0_1 that was associated to the epidermal tissue with power of 0.982, was expressed in 39.65% of the cells in the DGE (total of 1586 cells), there were 32,391 reads in the DGE that were mapped to the gene, and the number 69 which is the prefix of the contig identifier is below 1000. The 38 selected genes are listed in the file genes_to_use.xlsx (SM 2).

The alignment bam file that was produced and cleaned using the DetectBeadSynthesisErrors module in the first part (one step before creating the DGE itself), named bowtie_gene_exon_tagged_clean.bam, was processed to collect data for those selected genes. Ultimately, per each gene we created a file holding the information of good quality mismatches per position in the gene sequence, based on all the reads in the bam file that were mapped to the gene. The following steps were taken to accomplish this per each gene:

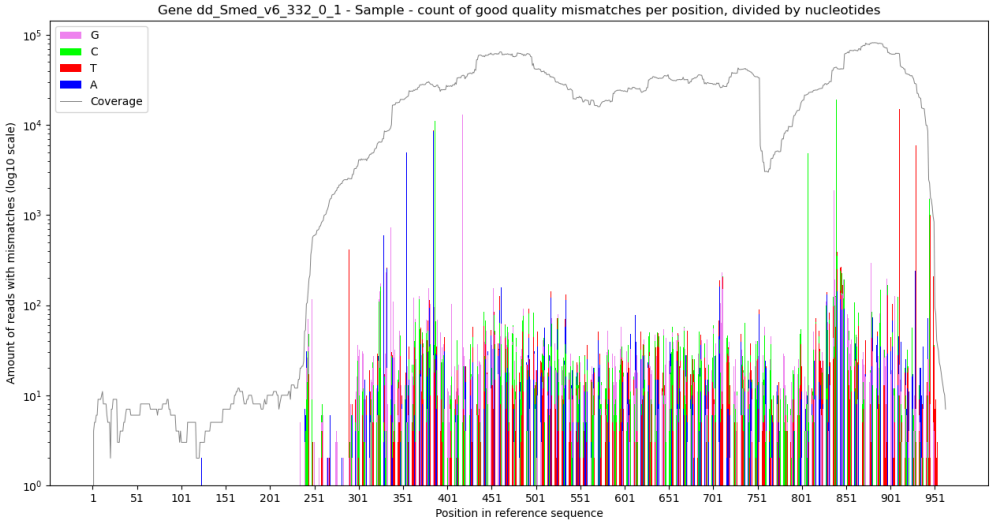
Firstly, using Samtools, we created from the bam file a smaller bam file (referred to as 'gene bam' henceforth) containing only reads that were mapped to the specified gene, removing reads that contained indels, and converting read bases identical to the aligned reference base to the = sign.

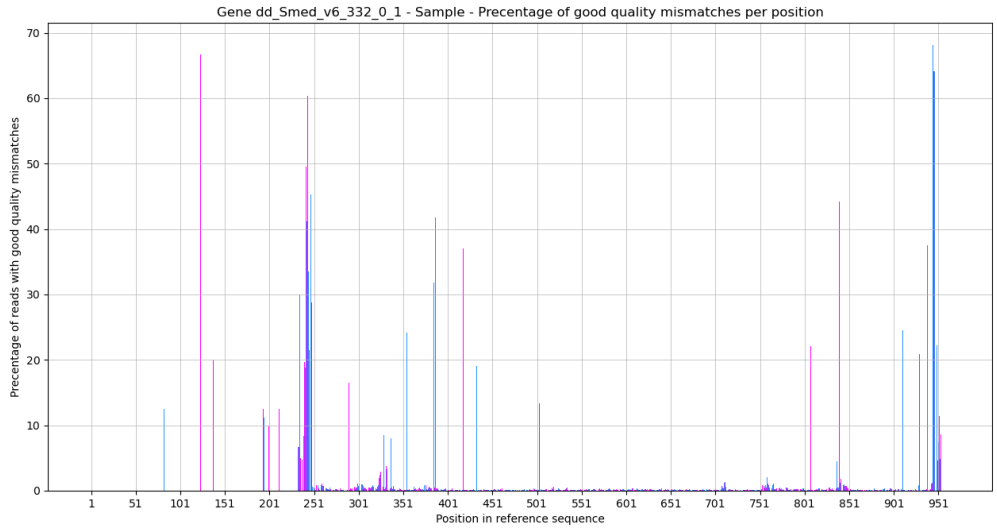
Secondly, using python scripts, the gene bam was converted to csv file, keeping only the relevant data for further processing: read name, leftmost mapping position, read sequence, ASCII of sequencing qualities per base, cell barcodes (XC tag), and number of mismatches in the alignment of the read (NM tag). The column names in the csv file are 'name', 'pos', 'seq', 'qual', 'tag', and 'NM', respectively.

Thirdly, since there were cells that contained numerous reads compared to cells that contained thousands of reads, and it could cause some data bias, we decided to sample reads from cells. In order to set the threshold, we plotted the number of reads per cell for reads with no insertions, mapped to the gene dd_Smed_v6_332_0_1, and discovered that most of the cells (85.66%) contained up to 10 reads, while the rest contained hundreds and thousands of reads. This specific gene



was chosen since it was expressed broadly in the cells represented in the DGE (73.25% of the cells) with a very large total amount of reads (146,511), so it provided some reasonable perspective. According to that, we sampled 10 reads from cells that contained more reads than that, and kept all the reads from cells that contained up to 10 reads.
Afterwards, we created a bam file containing only the reads we kept after the sampling, and calculated the coverage they provided for each position in the gene sequence, using Samtools.
Lastly, we collected all mismatches with sequencing quality score of at least 20, which considered as good quality, per each position in the gene sequence (this data was saved as a dictionary, elaboration will follow below). Using the coverage data, we calculated per position the good-quality mismatches amount and percentage, the amount of each nucleotide that appeared as mismatch, and its percentage out of all mismatches at the position and out of all reads covering the position. Note that for the nucleotide that matches the reference in each position, the values of amount and percentage would be 0, since we display only information of good-quality mismatches (matches and low-quality mismatches would complete the percentage in each position to 100%).
For each gene, the final file that summarizes the mismatches information per position is called 'X_sample_good_quality_mismatch_info.csv', where X is the number of the gene, e.g., for the gene dd_Smed_v6_332_0_1, the file's name would be '332_sample_good_quality_mismatch_info.csv'.
We also saved a dictionary for the good-quality mismatches, where keys are the positions in the gene sequence, and values are the mismatch bases, their sequencing quality, and the tag of the cell in which the read containing the mismatch was found. The dictionary was saved as pickle file named 'X_sample_pos_good_quality_mm_dict.pkl', where X is the number of the gene, as explained above. This dictionary can be loaded with pickle and used for further exploration, specifically for sorting the mismatches by the cells they were found in, and examining variance among cells.
In addition, for each gene we have created two bar plots: one shows the amount of good-quality mismatches per position, divided by nucleotides, and the other shows the percentage of good-quality mismatches per position. The files are named 'X_sample_barplot_mm_by_nucleotides.pkl' and 'X_sample_barplot_mm_precentage.pkl', respectively, where X is the gene number. These plots were saved as pickle files, so they can be loaded with the script 'show_plot.py' as interactive figures, because the bars in the plots are tightly packed, and it can be difficult to observe when saved as a regular image.
For example, these are the plots saved as an image for the gene dd_Smed_v6_332_0_1:



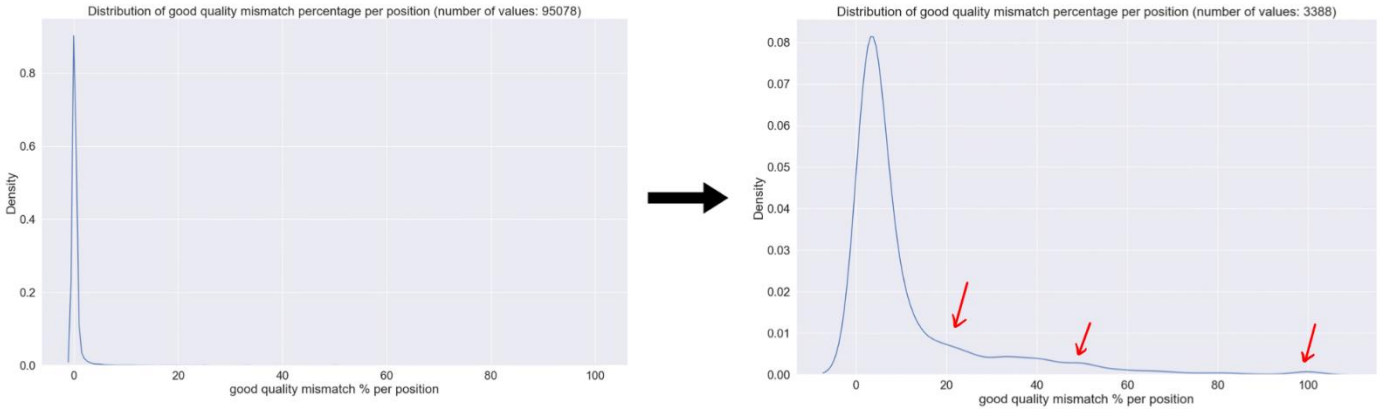


The entire described above process is done using the shell script 'loop_over_genes.sh', that gets as input the cleaned alignment bam and a txt file containing the names of genes to use, and for each gene runs the shell script 'process_BAM_for_gene.sh' which performs the described process for a specified gene (SM 1).

Third part - analyzing the mismatches data:

All the files that summarize the mismatches information that were created per gene, were merged to a single file named 'all_genes_mm_per_pos_info.csv', containing the following data per gene and per position in the gene sequence: coverage at the position, reference base, good-quality mismatch percentage, amount of each nucleotide that appeared as mismatch, and its percentage out of all reads covering the position.

We created a distribution of good-quality mismatch percentage per position, using all the data from the genes we selected. We found, unsurprisingly, that most of the values were 0%-2%. Since the density at this area of the distribution was very high, we removed values below 2%, and got a more examinable distribution, as follows:



Mismatch percentage of up to 2% was determined as background- sequencing errors. Beyond that, we found some interesting areas in the new distribution, as marked above: around 100%, 50%, and 20%, and each one was examined closely. We gathered all positions in the genes that the mismatch percentage in them fell into one of these ranges: 93%-100%, 40%-60%, 15%-35%, while noting the mismatch percentage of each nucleotide that appeared at the position and was different from the reference base. This data was compared manually to bulk RNA sequencing data ('ddv6_unc22_RNAi_2.BAM' and 'ddv6_unc22_RNAi_3.BAM'), displayed in IGV, in order to verify any trends we may notice. The bulk RNA data was not filtered as the single-cell RNA data to include only mismatches with good quality, since this data is inherently of good quality. This information is fully shown in the files named 'mm_percentage_range_X-Y.xlsx' where X-Y specify the mismatch percentage range (SM 3).
From each percentage range we chose some representative positions, while discarding positions with too low coverage, and trying not to choose too many positions from the same gene. For each range of mismatches considered, we observed the following:

- 19 positions with mismatch percentage of 93%-100% were found, and only 5 of them were compared to bulk data, since the others had only 1-2 reads covering the position in the single-cell RNA data. Two positions had very low coverage in the bulk data, so they could not be addressed properly. In two other positions (22 in dd_Smed_v6_432_0_1 and 19 in dd_Smed_v6_432_0_1) there was substitution from A to G with frequency above 92%, which can be due to an error in the reference sequence. In the fifth position we compared, 25 in dd_Smed_v6_432_0_1, the substitution frequency differed between the bulk data and the single-cell data, with A to C substitution frequency of 72% in the bulk data, compared to 100% in the single-cell data.

- We found 175 positions with mismatch percentage of 40%-60%, and 31 of them were selected for examining. At 18 positions, the reference base appeared in 48%-65% of the reads mapped to the position, and another nucleotide appeared with frequency completing to almost 100%, since in some cases other nucleotides were appearing in up to 1% of the reads, as

background. For example, at position 10,797 in dd_Smed_v6_636_0_1, the reference base T appeared in 50% of the reads in the bulk data, while C appeared in 49% of the reads in the bulk data and in 40% of the reads in the single-cell data. It is likely that at these positions there are two alleles in the diploid genome of the organism, while only one allele is shown in the reference sequence. There were two positions where in the single-cell data there was G to T substitution in about half of the reads, while in the bulk data this substitution happened in most of the reads. For example, at position 290 in dd_Smed_v6_61_0_1 the reference base was G, but T appeared in 96% of the reads in the bulk data, while appearing in only 50% of the reads in the single-cell data. Two other positions (948 in dd_Smed_v6_332_0_1 and 670 in dd_Smed_v6_478_0_1) were located at the polyA tail in the gene's reference sequence, and had substitution of A to T in about half of the reads in the single-cell data. The frequency of such substitution in the bulk data was significantly low, possibly due to a much lower number of reads mapped to those positions.

- We found 385 positions with mismatch percentage of 15%-35%, and 60 of them were selected for examining. At 29 positions there was a substitution to one other base with frequency of 15%-35%, that was similar in both bulk and single-cell data. At 9 of these positions there was G to A substitution, e.g., at position 206 in dd_Smed_v6_146_0_2 the base A appeared in 16% of the reads instead of the reference base G, in both the bulk data and in the single-cell data. At 4 positions out the 29, there was A to G substitution, for example, at position 95 in dd_Smed_v6_3_0_1 the base G appeared instead of the reference base A in 19% of the reads in the bulk data, and in 21% in the single-cell data. Probably there was RNA editing of A->I, and I was interpreted as G by the sequencer. At 6 positions out the 29, there was C to T substitution, for example, position 323 in dd_Smed_v6_478_0_1 contained 32% T in the bulk data instead of the reference base C, while 26% in the single-cell data. Probably there was RNA editing of C->U, and U was interpreted as T by the sequencer.
In addition, there were 5 positions with mismatch frequency of around 30% in the single-cell data, while it was 40% in the bulk data. These positions are likely to be not-isogenic spots in the diploid genome of the organism, containing two different alleles, at the edge of the possible distribution. For example, position 620 in dd_Smed_v6_241_0_1 that contained 53% of the reference base G in the bulk data, and 46% A in the bulk data, while only 34% in the single-cell data. There were also two positions (941 and 943 in dd_Smed_v6_1440_0_1) located at the polyA tail in the gene's reference sequence, that had A to T substitution frequency of around 20% in the single-cell data, but 0% in the bulk data, maybe due to a much lower number of reads mapped to those positions in the bulk data, that failed to exhibit this trend.
Moreover, we found one interesting position (167 in dd_Smed_v6_146_0_2) where 3 different nucleotides appeared in the reads mapped to it- the reference base G with 71%, T with 24%, and A with 6% in the bulk data (in the single-cell T appeared in 11% of the reads, and A appeared in 4%).
Another special and interesting case was substitution from A to C in one position, and from C to A at a near position, with inverse relation of the frequency, since each read that contained C to A substitution, also contained A to C substitution downstream. For example, in the gene

dd_Smed_v6_332_0_1 at position 385 there was C to A substitution with frequency of 32% in the bulk data, and at position 387 there was A to C substitution with frequency of 32%.
This also happened at positions 178 and 184 in dd_Smed_v6_68_0_1.
Lastly, there were 12 positions where the mismatch frequency in the single-cell data did not match the frequency in the bulk data, for example, position 976 in dd_Smed_v6_37_0_1 which had A to G substitution frequency of 17% in the single-cell data, but only 1% in the bulk data.

After examining the above ranges, we decided to also gather all positions with 75%-82% mismatch percentage, to see if there are positions where some base appeared more frequently than the reference base- less frequent than in the case it can be explained as an error in the reference sequence itself (around 100%), but more frequent than in the case it can be explained as a second allele (around 50%). Such positions were scarce and we found only 11, while noticing that all of them located at the start or end of the gene sequence, with low coverage. Two positions, 5 and 7, in the gene dd_Smed_v6_629_0_1 exhibited what we were looking for, with about 80% substitution frequency in the single-cell data, and around 65% in the bulk data (from T to G in position 5, and T to A in position 7).
These full results are summarized in the file named 'results_summary.xlsx' (SM 3), with a possible explanation of the observation in each examined position.

It should be mentioned that in all the mismatch percentage ranges we examined, among the positions we selected to compare with the bulk data, there were some positions that could not be compared since there were no reads mapped to these positions in the bulk data. These positions are also listed in the file 'results_summary.xlsx' (SM 3).

Discussion

We examined single-cell RNA sequencing reads that were mapped to 38 genes chosen carefully, we looked at the mismatches at each position, and deeply inspected selected positions that the good-quality mismatch percentage at them fell in one of the ranges we decided to examine in the conducted distribution of mismatch percentage per position.
Only a few positions had mismatch percentage of 93%-100%, it is likely that there was an error in the reference sequence at these positions, or there were very little reads, so even a few random mismatches could cause the mismatch percentage to be very high.
Among the positions that had mismatch percentage of 40%-60%, most of the cases reconcile with the fact that there is a second allele in the planarian diploid genome, that do not appear in the recorded reference sequence.
The mismatch percentage ranged between 15%-35% is the range we believe can actually hold within it positions that are suspected to be genetically heterogeneous among cells, because the mismatches frequency at them cannot be explained easily with known genomic phenomena, like 100% or 50% mismatch frequency. We found some positions that the mismatches at them

can be explained by RNA editing, e.g. A->G substitutions that can be explained by A->I editing, where I was interpreted as G by the sequencer, or C->T substitutions that can be explained by C->U editing, where U was interpreted as T. However, many other positions that we found had other base substitutions, e.g. T->C, C->A, A->C, A->T, T->A, G->A, G->T, that need further research to find out what they mean and what are they caused by. We also noticed some positions that are located at the polyA tail in the reference sequence that had A->T substitution, and it is also a possible phenomenon to focus on and examine in order to give a proper explanation.
Some positions we examined (in all the ranges), had different mismatches percentage between the single-cell data and bulk data. This inconsistency could stem from the facts that (i) we sampled reads from the single-cell data using cell tags, in attempt to get a non-biased perspective, since there were cells with a few single reads versus cells with thousands of reads, but such sampling could not be done in the bulk data since there are no cell tags, (ii) the single-cell data was filtered so we took into account in our calculations only good-quality mismatches, but we did not filtered the bulk data since we assumed that it is inherently of good quality, and (iii) at some positions where was significant difference in coverage between the two data sources. These matters could affect the mismatch percentage per position that was calculated in both data sources and the comparison we performed for each selected position.
Another analysis limitation we encountered was lack of data in the bulk RNA for some positions we wanted to inspect, so we could not verify the trends that we saw in the single-cell data at these positions.

The ultimate goal of this research is to find out if there is some genetic heterogeneity among different cells in the same organism. We found at least 25 candidate positions in some genes, that had nucleotide substitution kind and frequency that could be explained with the known genomic phenomena. We can look deeper into those suspicious positions, sort the mismatches at them by the cells they were found in, and reach further conclusions. This can be done easily using the dictionary we created X_sample_pos_good_quality_mm_dict.pkl for each gene, and the python script 'cell_mm_variance.py' that creates a detailed report of the number of reads in each cell that contained a mismatch in the specific position, divided by the nucleotide that appeared there, and creates a heatmap.
Another further steps that could be taken are (i) looking at single-cell data of more other body parts, isolating from them cells of specific type (e.g. only neoblasts), and check for genotypic heterogeneity among these cells, instead of searching for it among all cell types, as we did. And (ii) obtaining more bulk RNA data, in order to compare those suspicious positions (mostly in the start and end of genes sequences) that were found in the single-cell data but had no reads mapped to them in the bulk data we used.

In conclusion, we provided a detailed procedure for processing Drop-seq sequence

data into detailed single-cell RNA sequencing data divided by cells, in order to obtain information of the transcriptomic reads sequences as they appear in each cell, with the variance among them. We make this data available in files specified in the supplementary materials, and hope that it will help understand further interesting discoveries about genomic heterogeneity among cells of the same planarian organism.

References

1. Fincher, Christopher T et al. “Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*.” *Science (New York, N.Y.)* vol. 360,6391 (2018): eaaq1736. doi:10.1126/science.aaq1736
2. Rozanski, A., Moon, H., Brandl, H., Martín-Durán, J. M., Grohme, M., Hüttner, K., Bartscherer, K., Henry, I., & Rink, J. C.
   PlanMine 3.0—improvements to a mineable resource of flatworm biology and biodiversity
   Nucleic Acids Research, gky1070. doi:10.1093/nar/gky1070 (2018)
3. Nemesh, James. "Drop-seq core computational protocol." *McCarroll Laboratory http://mccarrolllab. com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1. 2Jan2016. pdf* (2016).

Supplementary Materials

The supplementary materials attached to this paper in a zip file, and contain:
SM 1 - all scripts we used for our analysis.
SM 2 - table of genes we selected for our analysis, and txt file containing the genes' names.
SM 3 - files containing the full results and summary of the results.
SM 4 - table listing all created files, which script generated them, their location, and content.