Next Generation Sequencing – AU Summer course 2025

Marina Ramírez Baños (AU777574)

1. Introduction

This course covers a practical and foundational understanding of current next-generation sequencing (NGS) methodologies. We performed quality control on sequenced data, mapped to reference genomes, conducted VCF variant calling on samples, and bulk RNA sequencing of white clover (*Trifolium repens)* and human testicular tissue data.

T. repens results from the hybridation of *T. pallescens* and *T. occidentale*. We examined data derived from a self-compatible line, S10, and wild accession, Tienshan (Ti) under two experimental conditions – before and after exposure to frost - to explore genomic differences and adaptations. We used two types of reads: High Fidelity (HiFi) PacBio DNA-seq long reads for S10, and Illumina RNA-seq short reads for both S10 an Ti. The reference genomes used included a simplified chromosome 1 (contig1) of *T. occidentale,* a simplified chromosome 1 (contig2) of *T. pallescens,* and a merged from (contig1_2).

Additionally, we performed single-cell RNA-seq of human testicular tissue data to characterize different cell types involved in spermatogenesis.

2. Results and discussion

2.1. Raw data alignment

Quality control of sequencing data

Using Galaxy platform, we conducted data quality control on the PacBio HiFi and Illumina RNA-seq reads (in .fastq format). The analysis was conducted using Falco and MultiQC tools.

The MultiQC report (Figure 1) indicated a high mean sequencing quality – around Phred Score 80 across almost all the read length for PacBio HiFi Reads and around 35 for Illumina RNA-seq reads. Illunima RNA-seq reads show a decline towards the end, around 150 bp, which can be expected due to error accumulation and signal decay during the sequencing cycle. In contrast, PacBio HiFi Reads exhibit in average a stable quality score, but it , it is possible to notice a drop in the quality, compared to the average, in the first 10 and last 20k bp. The variability of the first 10 bp can be explained due to the use of random 10-base tags before sequencing, which can trigger the warning in the “Per Base Sequence Content” in the program. The variability of the last 20k bp can be attributed to the fact that fewer reads are that long.

The GC content (Figure 2) followed a normal distribution, 32% GC for PacBio HiFi reads, and 44% GC for Illumina RNA-seq reads, reflecting the fact that a greater GC is typically found in coding regions of the genome.

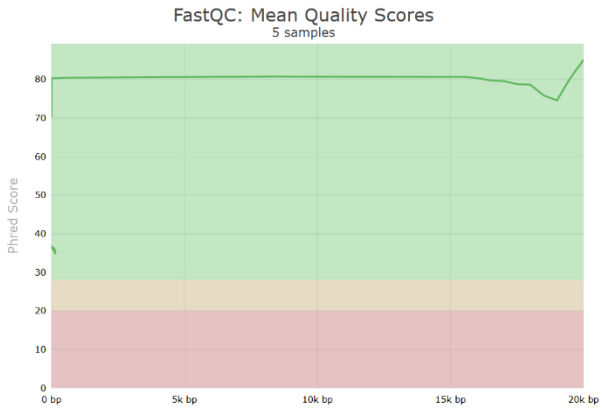

Figure 1. Quality score for RNA-seq and HiFi reads.

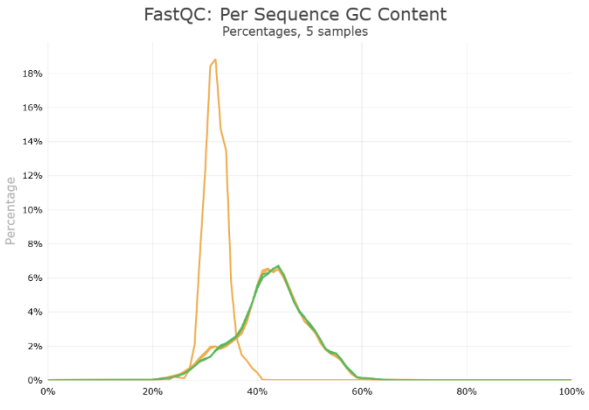

Figure 2. GC content for RNA-seq and HiFi reads.

HiFi data mapping

PacBio HiFi reads were mapped to the white clover reference genomes using minimap2 tool with two different presets: map-hifi (optimized for long accurate reads) and sr (optimized for Illumina-type short reads). The resulting alignments were sorted using samtools and visualized using IGV platform.

Comparison of alignments presets

The quality of read mapping was significantly influenced by the choice of alignment preset configuration (Figure 3). The map-hifi preset shown a optimal performance, a great and uniform coverage along the reference genome, whit few, if any, mismatches. In contrast, applying sr preset to PacBio HiFi read led to suboptimal outcomes, including multiple mismatches, fragmented alignments, reduced coverage, and large gaps.

These differences stem from inherent differences between the used algorithms: sr aligners are specifically designed for short reads, which have specific error profiles and structural features. As a result, when applied to long reads, they fail to correctly interpret the read structure and generate artificial disjointed alignments and mismatches that impact on the accuracy of the genomic sequence representation. Please, note that the further analysis conducted in this report was using the map-hifi method.

Analysis of coverage patterns and subgenome mapping

The observed coverage patterns reflect the polyploid nature of *T. repens*. When Pac Bio HiFi reads were aligned to the combined reference genome (contig1_2), coverage across most regions remained relatively consistent at around 30–40 reads, with few mismatches.

In contrast, mapping the same reads to the individual contigs separately revealed distinct fluctuations in coverage, with two major peaks around 35 and 70 (Figure 4, 5). The peak at 35 corresponds to regions that are either unique to one of the two subgenomes or where reads align more confidently to one contig. The higher peak at 70 reflects regions of high sequence similarity between the two subgenomes, where reads from both contigs align equally well, effectively doubling the read depth.

Regions exhibiting elevated coverage and balanced allele frequencies close to 50:50 likely reflect the alignment of reads from both subgenomes to the same genomic location, due to high sequence similarity. This overlap can lead to structural inconsistencies in the alignment—such as insertions and deletions—resulting from subtle divergences between the subgenomic sequences.

In contrast, large regions lacking subgenome-specific SNPs may indicate significant sequence divergence, which prevents cross-alignment between subgenomes. Conversely, SNP-rich regions are likely to correspond to conserved loci, where both subgenomes align successfully, often representing homologous or functionally important sequences.

Moreover, the use of a combined reference may artificially increase coverage in highly similar regions, while in more divergent regions it can produce spurious signals of heterozygosity, misrepresenting the true genomic variation.

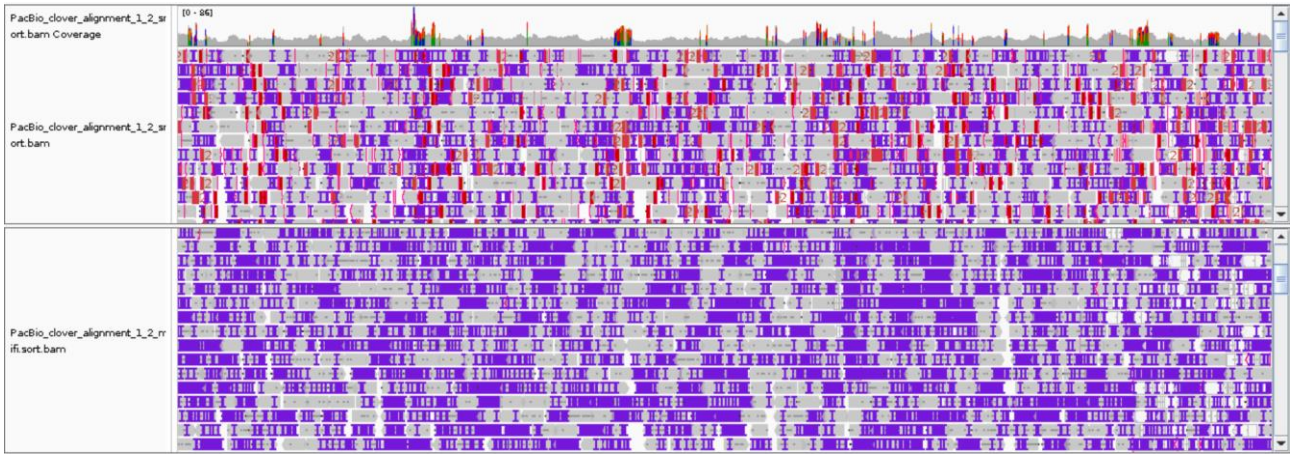

Figure 3. Mapping coverage for HiFi reads, aligned using sr present and mapped to contig1_2 (top) and aligned using map-hifi present mapped to contig1_2 (bottom).

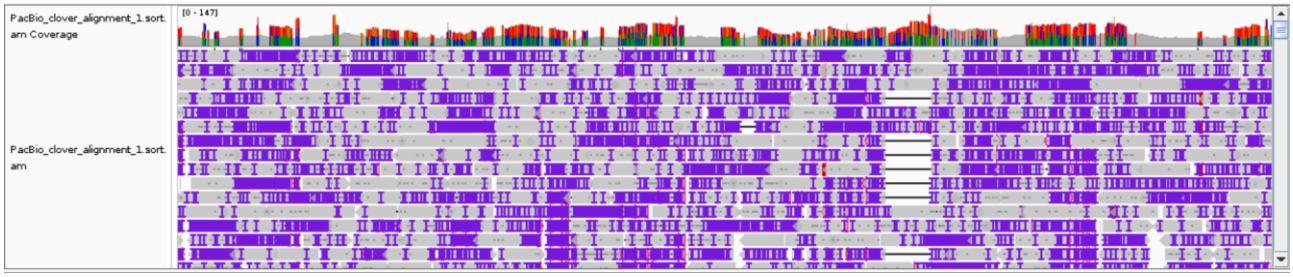

Figure 4. Mapping coverage for HiFi reads aligned using map-hifi present mapped to contig1.

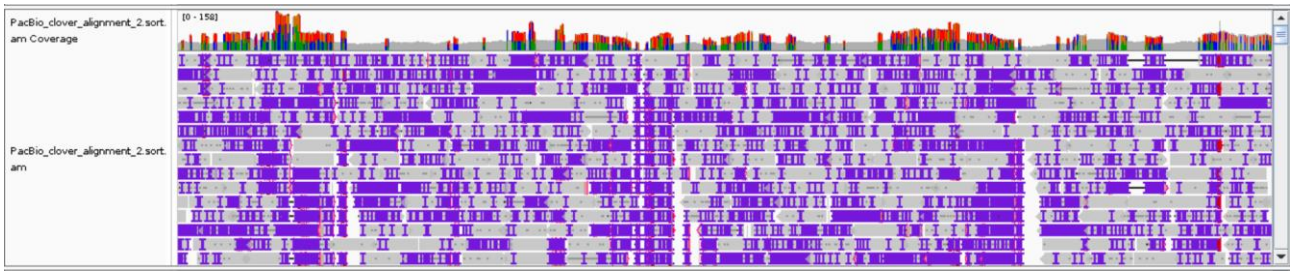

Figure 5. Mapping coverage for HiFi reads aligned using map-hifi present mapped to contig2.

Alignment statistics and mapping quality

When comparing the alignment statistics for reads mapped to merged versus individual contigs is possible to detect some major differences. While the majority of reads successfully align to reference genome, around the 15% of the reads are aligned to multiple locations. These phenomena can originate due the existence of repetitive sequences or highly similar regions between subgenomes, making the reads to not be assigned to a unique location.

RNA-seq mapping

The RNA-seq dataset included twelve samples—six from the Ti variant and six from the S10 variant—with half of each group subjected to frost treatment. Alignment was conducted using the STAR aligner against the combined contig1_2 reference genome. Unique mapping rates were above 83% for Ti and 86% for S10 reads.

2.2. Variant calling analysis

Variant calling analysis on aligned PacBio HiFi and Illumina RNA-seq reads was performed using contig1_2 as reference and bcftools, which generates VCF to store SNPs. The aim of the analysis was identifying subgenome SNPs (mapping HiFi reads to contigs) and genomic variations between S10 and Ti genotype.

A significant concentration of variants was observed near the beginning of contig2, which was unexpected given that the reference genome was constructed from real reads. Therefore, true SNPs should theoretically be absent in this region. To determine whether these variants represented false positives (errors that were generated, for instance, during the merge of both contigs) or genuine differences, we compared alignments against the combined contig1_2 reference with those against the individual contigs (Figure 6). If reads expected to map exclusively to contig1 align only there—and similarly for contig2—this suggests the presence of false positive calls.

We filtered for false positives (or SNPs with low quality) adjusting relevant parameters, as the quality scores, mapping quality, and depth sequencing (Table 1). Additionally, we plotted the SNPs quality, mapping quality and depth for the PacBio HiFi reads plotted against the merged and individual contigs as well as with S10 and Ti alignments to reference genomes. Before filtering, the number of SNPs detected for PacBio HiFi VCF files was: HiFi contig 1 had 22918 SNPs; HiFi contig 2 had 22588 SNPs; and HiFi contig1_2 had 1433 SNPs. After applying the defined filters: HiFi contig_2 has 21033; HiFi contig_1 has 21756; and HiFi contig1_2 has 1013.

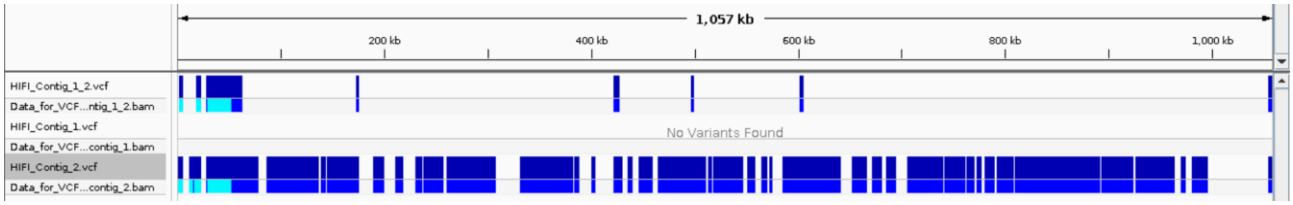

Figure 6. Variant called from HiFi reads mapped to contig1_2 and contig2.

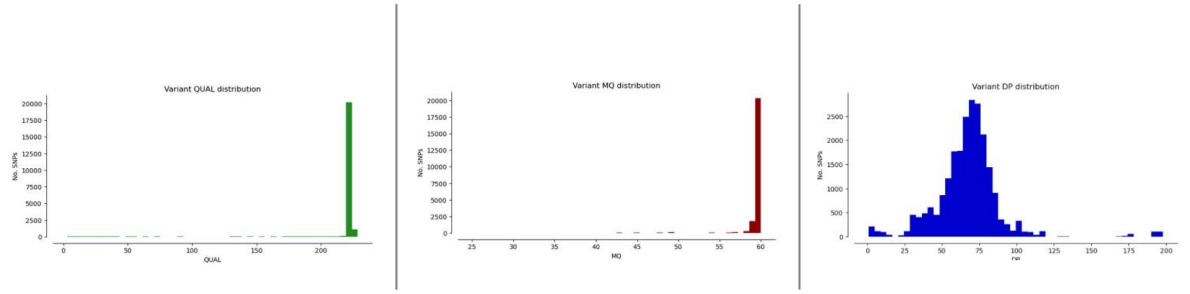

Figure 7. Variant quality distribution (left); mapping quality distribution (middle); and variant depth distribution (right) for SNPs from the aligned HiFi reads against contig1 (unfiltered).

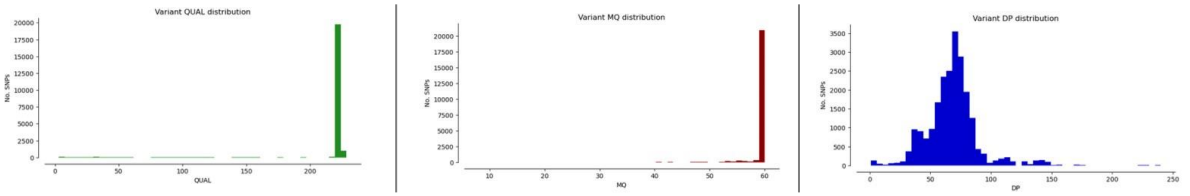

Figure 8. Variant quality distribution (left); mapping quality distribution (middle); and variant depth distribution (right) for SNPs from the aligned HiFi reads against contig2 (unfiltered).

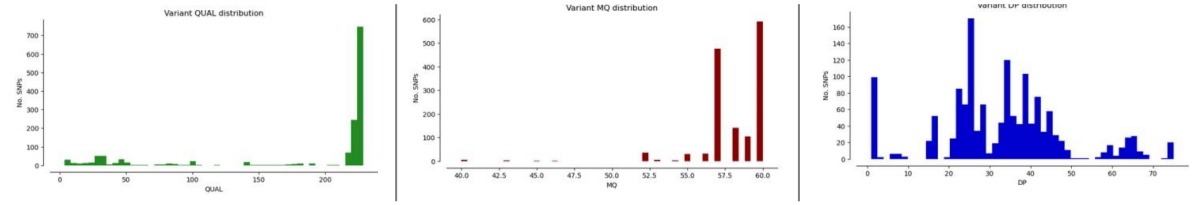

Figure 9. Variant quality distribution (left); mapping quality distribution (middle); and variant depth distribution (right) for SNPs from the aligned HiFi reads against contig1_2 (unfiltered).

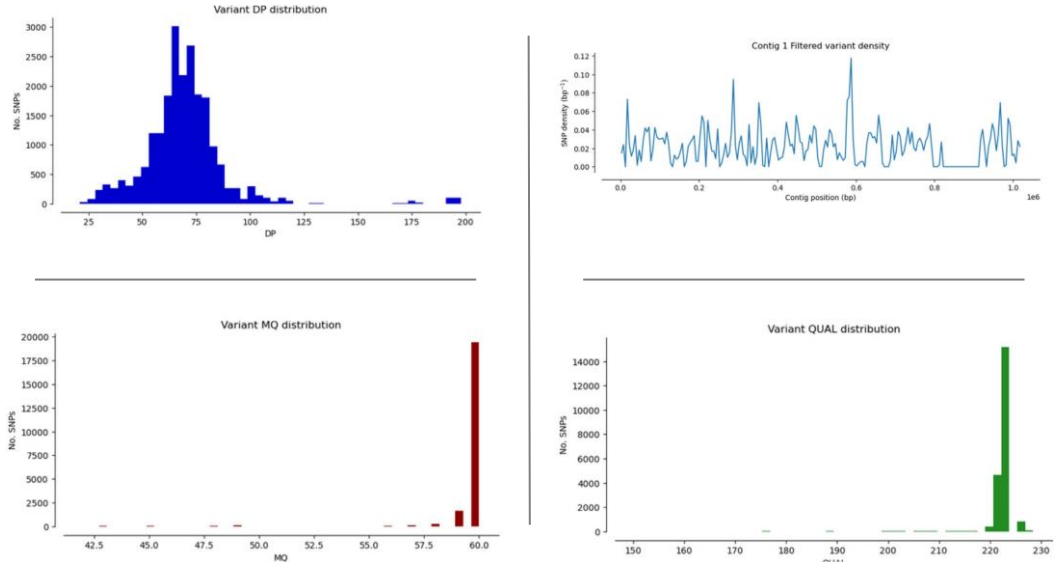

Figure 10. Filtered variant depth distribution (top left); mapping quality distribution (bottom left); variant density (top right); variant quality distribution (bottom right) for SNPs from the aligned HiFi reads against contig1.

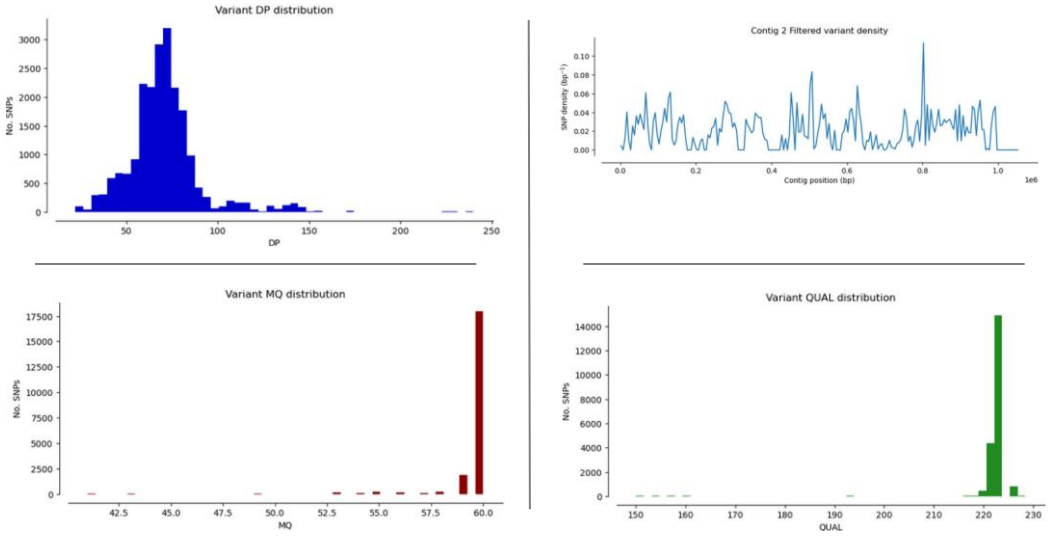

Figure 11. Filtered variant depth distribution (top left); mapping quality distribution (bottom left); variant density (top right); variant quality distribution (bottom right) for SNPs from the aligned HiFi reads against contig2.

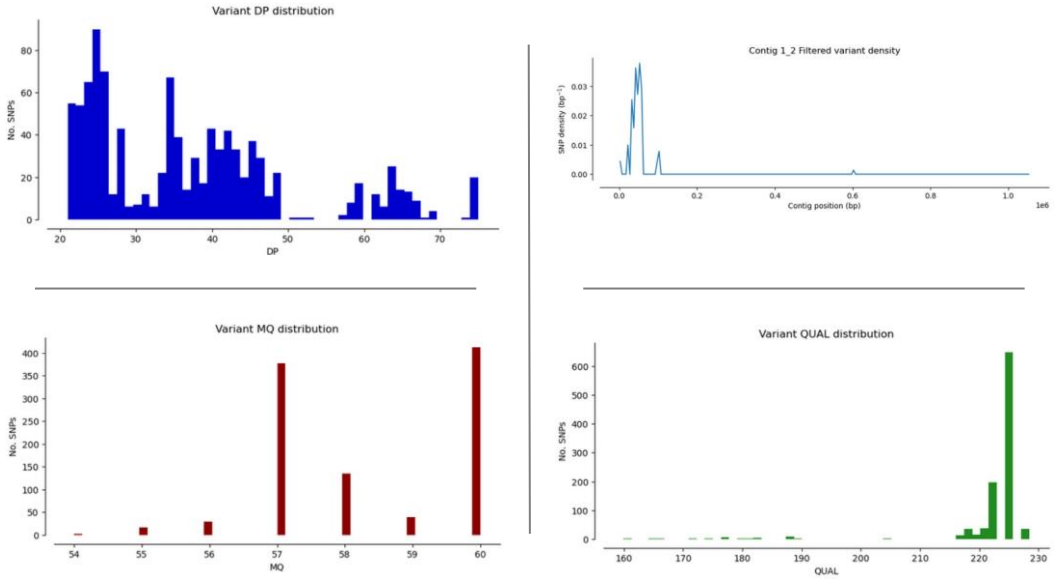

Figure 12. Filtered variant depth distribution (top left); mapping quality distribution (bottom left); variant density (top right); variant quality distribution (bottom right) for SNPs from the aligned HiFi reads against contig1_2.

2.3. Bulk RNA sequencing

To analyse bulk RNA-seq data and identify differentially expressed genes (DEGs), we employed edgeR, organizing the dataset with genes as rows and samples as columns. The data preprocessing included normalization of gene counts, removal of rows with zero counts, and a logarithmic transformation using log2(counts + 1). Principal component analysis (PCA) and heatmaps provided valuable insights to detect differences between genotypes and treatment conditions (Figure 13). PC1 captured the majority of variation between the S10 and Ti samples, while PC2 highlighted variation attributable to the experimental treatments. Additionally, to detect changes in gene expression due to treatment, excluding genotype effects, we conducted a generalized quasi-likelihood test. The resulting DEGs—from Ti, S10, and the comparison excluding genotypic effects—were combined and visualized through a Venn diagram (Figure 14).

Some genes only showed significant differential expression when both genotypes were analysed together, likely due to the increased sample size. However, in species-specific analyses, differential expression was not statistically significant.

Further examination revealed multiple duplicated genes when comparing Ti and S10 genotypes. This observation aligns with the allotetraploid nature of white clover, which possesses two homologous copies of each gene.

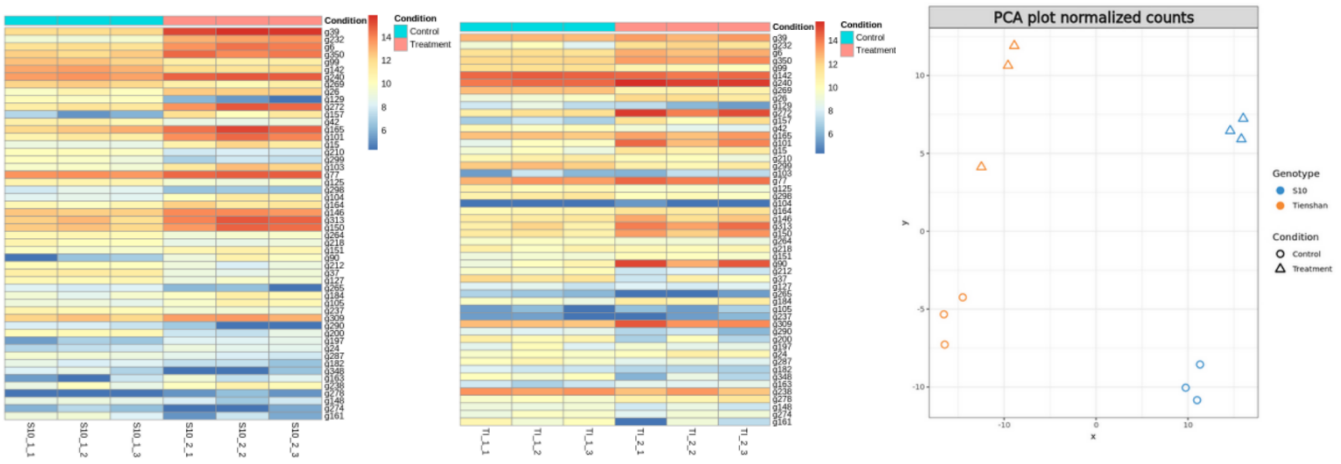

Figure13. Heatmaps for differential expression analysis, comparing control against treatment condition (left for S10; middle for Ti); PCA plot for raw counts (right).

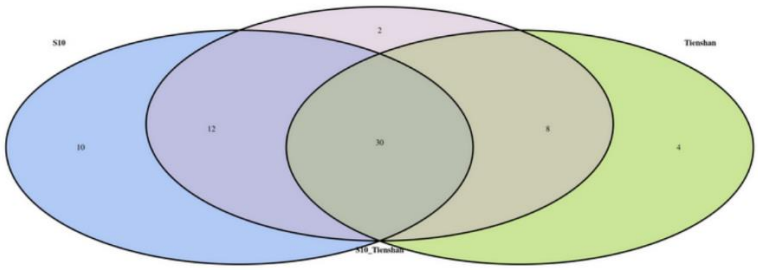

Figure 14. Venn Diagram showing DEGs from Ti, S10, and excluding genotypic effects.

Own question

The experimental question addresses how gene expression changes induced by treatment compare between two white clover lines, S10 and Ti, focusing on genes commonly differentially expressed in both. To answer the question, we illustrated using a scatter plot this comparison (Figure 15), by plotting the logFC values for each shared gene, with the x-axis representing S10 and the y-axis representing Ti. The diagonal line represents equal expression changes between the lines. Genes positioned near this line exhibit similar regulation in both lines, while those distant from the line indicate differential expression patterns, revealing potential line-specific responses to treatment.

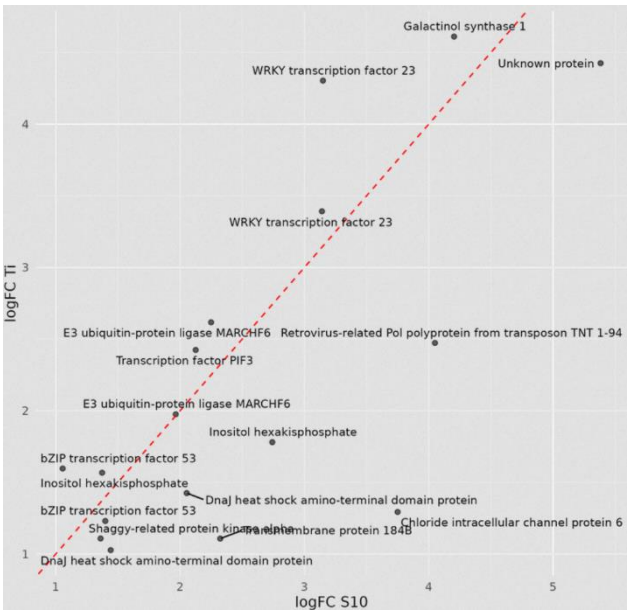

Figure 15. Scatter plot representing the differential expression of common DEGs from Ti and S10.

The plot shows that some of the common DEGs with larger logFC for both genotypes are regulated in oppo-

site direction, such as the genes encoding for Retrovirus-related Pol polyprotein from transposon TNT 1-94, Chloride intracellular channel protein 6, and WRKY transcription factor 23.

2.4. scRNA-seq analysis

Testis data preprocessing

Single-cell RNA-seq testis data was process using scanpy framework, following a quality control and normalization procedure. The first filtering steps focused on excluding poor-quality cells that could interfere in the analysis: potential multiplets (cells with >40000 transcripts), and cells with very high (>8000) or low (<800) gene counts – possible stressed. To reduce technical noise, we also removed genes expressed in less than 10 cells. To improve data quality, cells with over 20% mitochondrial transcripts or more than 10% MALAT1 expression were excluded. High levels of these markers typically indicate stressed, dying, or disrupted cells, which can introduce technical noise. Filtering based on these criteria helps ensure that downstream analyses reflect true biological signals.

Potential doublets were identified using Scrublet, which estimated an 8% doublet rate; cells with scores above 0.1 were excluded to prevent artifacts from mixed transcriptomes. Following quality control, expression data were normalized with TPM (transcripts per million) to account for sequencing depth, then log-transformed to reduce variance. Genes with high variability (variance > 0.05; 14965 genes) were selected, and the top 15,000 were scaled to zero mean and unit variance for downstream dimensionality reduction and clustering.

Dimensionality reduction and cluster identification and assignment

PCA was first conducted to identify the primary sources of variation in the dataset, with the top 15 principal components capturing the majority of variance. Despite this, some overlap between cell populations remained. To enhance separation and better visualize cellular heterogeneity, UMAP projections were generated based on these components (Figure 16). Clustering was guided by the expression of known marker genes, allowing the assignment of specific cell types. Additionally, applying the Leiden algorithm with a resolution of 0.5 further refined the clustering by grouping transcriptionally similar cells, providing a more distinct and interpretable representation of the various cell types involved in spermatogenesis.

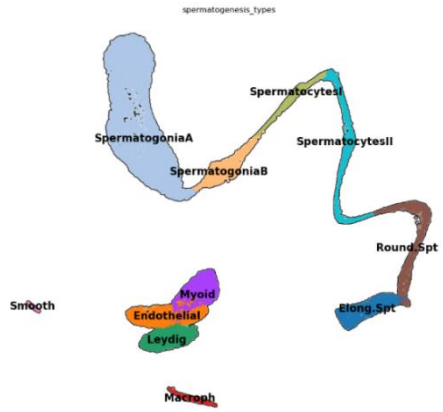

Figure 16. UMAP visualization coloured by identified cell-type clusters.

Gene Enrichment Analisis

Macrophages were analysed to identify uniquely enriched genes and pathways. The top 50 cluster-specific marker genes were submitted to Enrichr for functional analysis. The results revealed enrichment in transcriptional and translational regulation, as well as MHC-related immune functions, aligning with their role as immune cells.

Comparison between healthy and infertile individuals

Pseudotime analysis in healthy individuals showed that round and elongated spermatids display higher variability, suggesting they undergo more advanced differentiation than spermatogonia A, B, and spermatocytes I. Using the ingest tool, annotations and clustering from healthy samples were transferred to cryptozoospermic individuals, generating a new UMAP where clustering patterns were largely similar between groups. However, pseudotime distributions differed notably, with healthy men showing greater variability across most cell clusters (Figure 17). This suggests that spermatogenic differentiation is diminished in men with cryptozoospermia compared to healthy controls, despite comparable cell type composition.

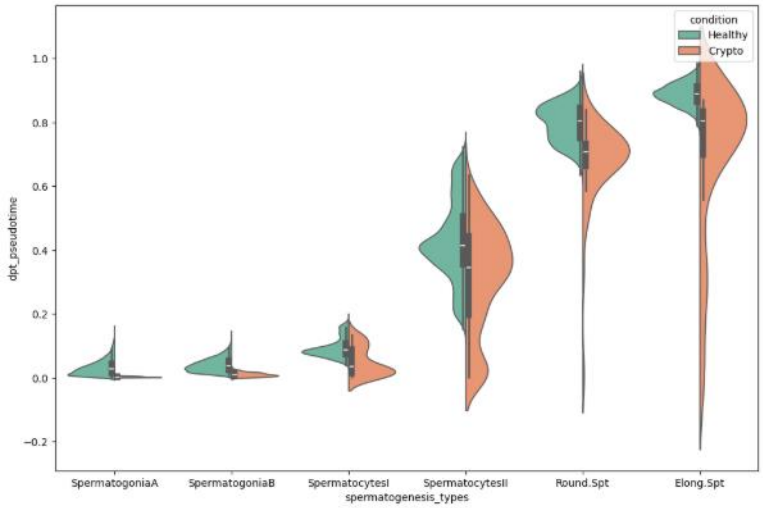

Figure 17. Pseudotime distribution of healthy group versus cryptozoospermia group.

Cross-dataset differential expression

We conducted a differential expression analysis comparing healthy and infertile individuals using pseudobulk samples, grouping the data by cell type, condition, and sample. Pseudobulk samples were transformed into annotated dataset objects, with genes lacking expression in all cells being excluded from the analysis. Results were visualized using a volcano plot ($p < 0.001$; absolute logFC >2). The analysis revealed that most genes were downregulated in the infertility group, suggesting that gene inactivation may play a significant role in male fertility.

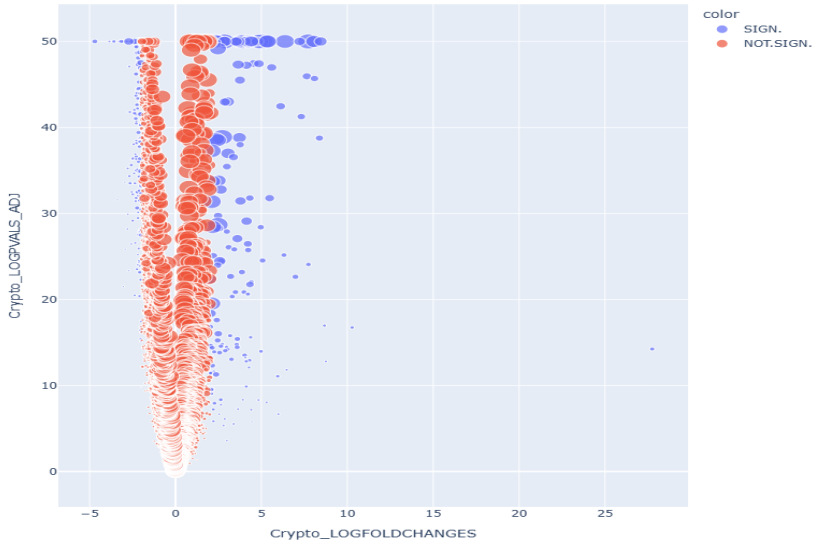

Figure 18. LogFC of healthy group versus infertile group ($p < 0.001$; absolute logFC >2).

Table 1. Cutoff values and relevant parameters.

Differentially expressed genes		
Cutoffs selected	FDR	0,05
	logFC	1
Number of genes	**up**	**down**
S10 frost	28	24
Tienshan frost	22	20
S10 versus Tienshan	27	25

Variants					
Filtering criteria selected:		**DP**	**QUAL**	**MQ**	
	HiFi	>20	>150	>40	
	Ti	>20	>20	>15	
	S10	>20	>20	>15	
Number of calls		**no call**	**Hom Ref**	**Het**	**Hom Alt**
Contig 1 HiFi subgenome (A)			0	27768	1427
Contig 2 HiFi subgenome (B)			0	27546	1412
Contigs1&2 HiFi			0	326	656
S10			0	12	0
Tienshan (C)			0	612	702

Number of calls	A het C het	A het C hom alt	A hom alt C het	A hom alt C hom alt
Overlap A and C	58	149	8	7

Number of calls	B het C het	B het C hom alt	B hom alt C het	B hom alt C hom alt
Overlap B and C	92	162	9	6