Biological effect of miR-379 on human prostate cancer cells in different metastatic enviroments

Redmine issue: 4882

NBIS staff: Nima Rafati

06 April, 2021

Redmine issue:	4882
NBIS staff:	Nima Rafati (nima.rafati@nbis.se)
Principal investigator:	Yvonne Ceder(Yvonne.ceder@med.lu.se)
Request by:	James Cassidy(james.cassidy@med.lu.se)
Organisation:	Lund University
Estimated time:	120
Used time:	120

Contents

1 Work log

The aim of this project is to investigate the effect of miRNA-379 on gene expression profile of prostate cancer in different metastatic environments both in in-vitro (regular cell media and osteoblast conditioned medium) as well as in-vivo (mouse liver and bone).

- **2019-09-06:** Meeting with the group to plan the data analysis
- **2019-10-31:** Giving update about QC and trimming of the data as well as selection of samples to test the pipeline
- **2019-12-02:** Giving feedback about high multimapped reads in the tested samples
- **2020-01-22:** Selecting samples with RIN value > 5
- **2020-03-27:** Reporting analysis based on a new pipeline (GSNAP + Disambiguate/XenofilteR)
- **2020-04-03:** Suggesting to check the barcodes used during demultiplexing and asking sequencing platform to provide bcl files generated by sequencing machine
- **2020-08-27:** Confirming errors in used barcodes
- **2020-10-09:** Reporting on the progress and comparison with another prostate cancer cell line from Jividen et al 2018.
- **2020-10-09:** The group confirmed species specific amplification in selected number of samples by qRT-PCR
- **2020-12-18:** Improving the pipeline to generate better mapping and resolving reads clipping issue; This was done by using XenofilteR
- **2021-01-28:** Internal meeting with colleagues at NBIS. Based on the discussions additional analyses were performed.
- **2021-03-03:** Last meeting; we decided to present all the observations in a report.

2 Practical information

2.1 Data responsibilities

Unfortunately, NBIS does not have resources to keep any files associated with the support request; we kindly suggest that you safely store the results delivered by us. In addition, we kindly ask that you remove the files from UPPMAX/UPPNEX. The main storage at UPPNEX is optimized for high-speed and parallel access, which makes it expensive and not the right place for long-term archiving. Please be considerate of your fellow researchers by not taking up this expensive space.

The responsibility for data archiving lies with universities and we recommend asking your local IT for support with long-term data storage. The Data Center at SciLifeLab may also be of help with discussing other options.

Please note that special considerations may apply to human-derived, sensitive personal data. This should be handled according to specific laws and regulations as outlined at the NBIS website.

2.2 Acknowledgements

If you are presenting the results in a paper, at a workshop or at a conference, we kindly remind you to acknowledge us according to the signed NBIS User Agreement:

NBIS staff should be included as co-authors if the support work leads to a publication and when this is merited in accordance to the ethical recommendations for authorship, *i.e.* the ICMJE recommendations. If applicable, please include *Nima Rafati, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University* as co-author. If the above is not applicable, please acknowledge NBIS like so: *Support by NBIS (National Bioinformatics Infrastructure Sweden) is gratefully acknowledged.*

In addition, Uppmax kindly asks you to acknowledge UPPMAX and SNIC. If applicable, please add: *The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project snic2019-30-25 (Storage) snic2019-8-295 (Computation).*

In any and all publications based on data from NGI Sweden, the authors must acknowledge SciLifeLab, NGI and Uppmax, like so: *The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure.*

2.3 Closing procedures

You should soon be contacted by one of our managers, Jessica Lindvall *(jessica.lindvall@nbis.se)* or Henrik Lantz *(henrik.lantz@nbis.se)*, with a request to close down the project in our internal system and for invoicing matters. If we do not hear from you within **30 days** the project will be automatically closed and invoice sent. Again, we would like to remind you about data responsibility and acknowledgements, see the sections on data responsibilities and acknowledgments.

You are naturally more than welcome to come back to us with further data analysis request at any time via the support form. Thank you for using NBIS, we wish you the best of luck with your future research!

3 Methods

In this project we tested different pipelines and here we only present the pipeline showed the most reliable results.

3.1 Genome preparation

We used human genome reference (GRCh38) on Uppmax and downloaded the annotation from gencode (version32). We downloaded mouse (BALB_CJ) genome and annotation from Sanger institute (https://www.sanger.ac.uk/data/mouse-genomes-project/ downloaded 2019Nov). We indexed the genomes by gmap_build from GSNAP(Wu and Nacu 2010).

3.2 QC (00-QC)

We checked quality of the reads by using FastQC (Andrews, n.d.) and merged the results by MultiQC(Ewels et al. 2016). For rRNA contamination we used bbduk from BBMap (version 38.61) (Bushnell 2014).

3.3 Trimming

By using trimmomatic(Bolger, Lohse, and Usadel 2014) we trimmed the adapters and filtered out low quality reads. We kept reads that both pairs survived trimming and filtering.

3.4 Alignment (01-BAM)

We aligned trimmed reads on human and mouse genome by STAR(Dobin et al. 2012) and GSNAP(Wu and Nacu 2010). We first evaluated the aligners and the results showed that GSNAP had a better performance Thus, all the results provided here is based on GSNAP alignment. To select species specific reads, we used Disambiguate (Ahdesmäki et al. 2017) and XenofilteR(Kluin et al. 2018) tools. These tools assign reads to corresponding species based on edit distance. We evaluated performance of these tools and XenofilteR could rescue more accurate alignments.

3.5 Post-alignment QC (02-Post-alignment-QC)

We evaluated number of metrics after alignment by using QoRTs(Hartley and Mullikin 2015). We checked the frequency of clipping, drop rate, gene-body coverage, and other metrics.

3.6 Expression analysis (03-Expression)

We extracted fragment counts of all genes by using featurecounts(Liao, Smyth, and Shi 2014). We used reads with mapping quality +20 and pairs that are properly mapped on the same chromosome. For normalization and downstream analysis we used edgeR(Robinson, McCarthy, and Smyth 2009).

All the downstream analysis and visualization are done in R (version 4).

4 Results

Figure 1 shows the study design and 27 samples used to generate RNA-seq data.

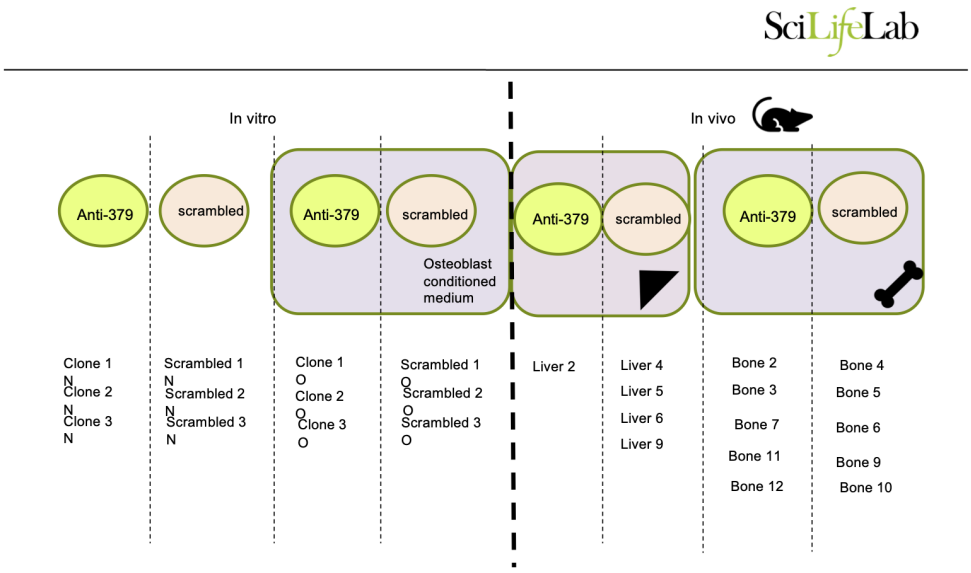


Figure 1: Study design and samples used in this experiment.

Table 1: Summary of sequencing data; Number of trimmed reads.

Sample	Trimmed_paired_reads
Bone10	81102733
Bone11	37971567
Bone12	53948391
Bone2	78067329
Bone3	47754989
Bone4	57020540
Bone5	53046795
Bone6	47358187
Bone7	27381663
Bone9	34191583
Clone1N	55121667
Clone1o	32982410
Clone2N	69161719
Clone2o	11959763
Clone3N	42989660
Clone3o	26328827
Scr1N	44153614
Scr1o	48000823
Scr2N	52635168
Scr2o	49678746
Scr3N	53050567
Scr3o	68393743
liver2	45037303
liver4	87073897
liver5	84114465
liver6	57720097
liver9	84363395

Table 1 shows the number of reads survived the filtering and adapter removal.

4.1 QC

4.1.1 FastQC

The QC results is available in
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/results/00-QC/
By MultiQC we summarized all the fastqc results. The duplication rate is high in the raw reads (Figure 2). Also, GC content seems to be shifted and a bit noisy (Figure 3). These figures together with other statistics are available in multiqc report saved under QC folder.

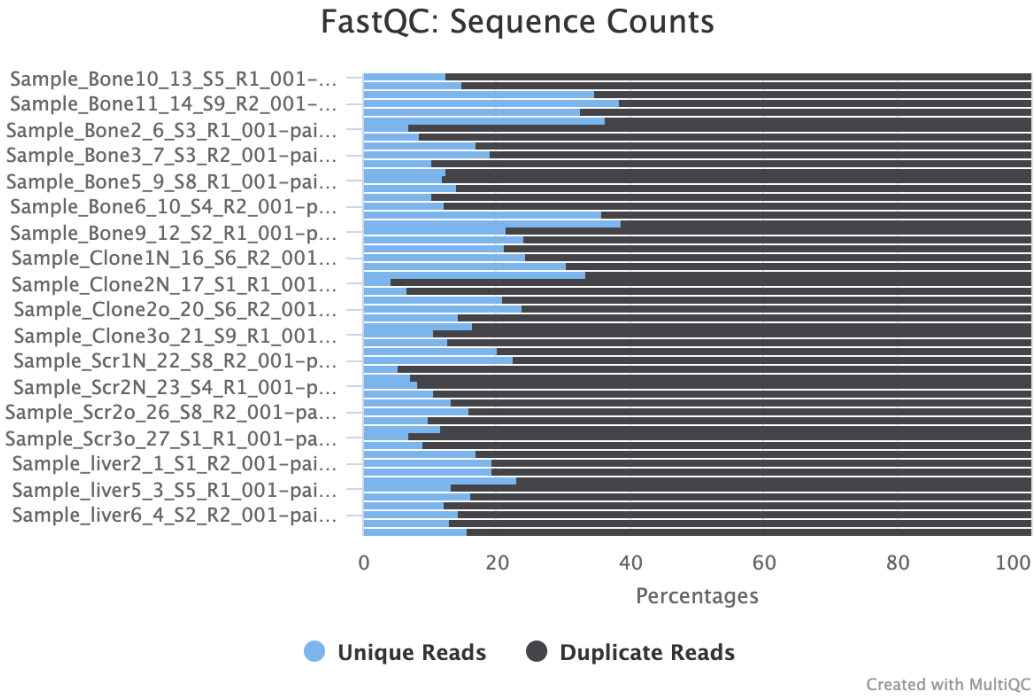


Figure 2: Fraction of duplicate reads.

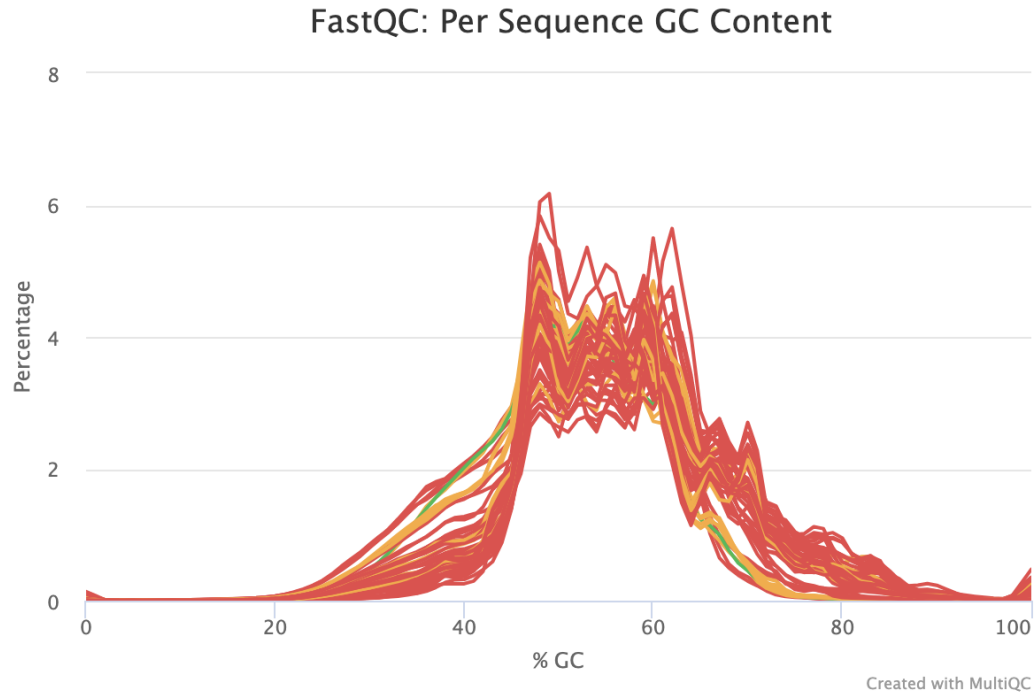


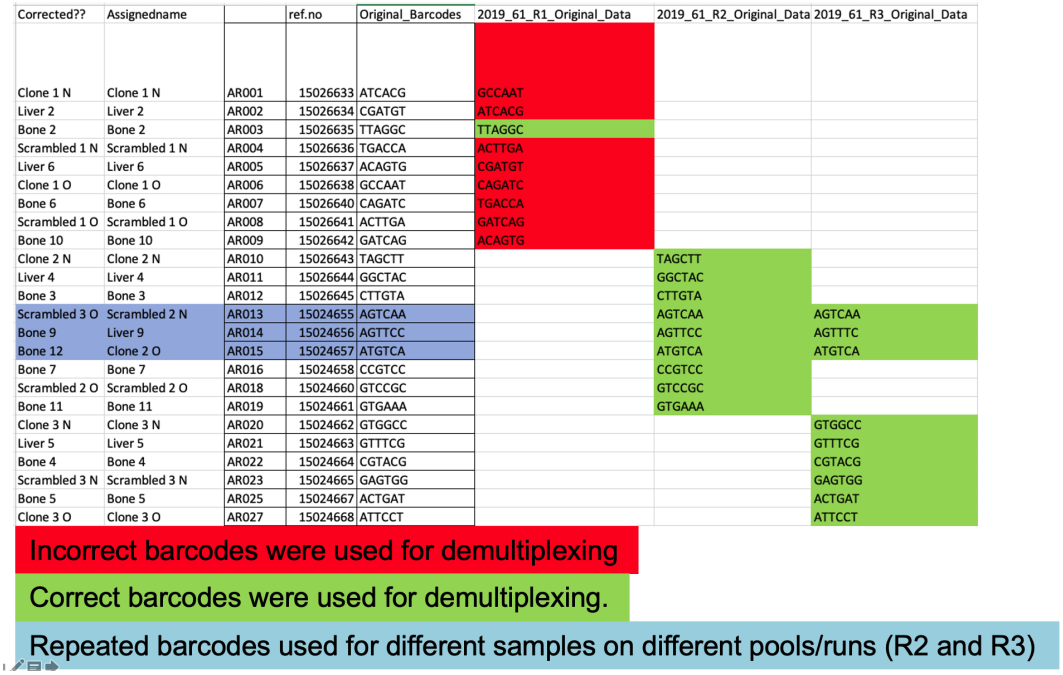
Figure 3: GC content across all samples.

4.1.2 rRNA contamination

We checked the data for presence of rRNA in the data by bbduk. All the rRNA sequences of both genomes were extracted from the annotation files. The contamination level was very low (0-0.07%). The results are available in
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/results/00-QC/

4.1.3 Barcodes

During the analysis we noticed an inconsistency between barcodes used for demultiplexing/sequencing procedure and barcodes used in the lab. Figure 4 indicates that almost all the barcodes in pool1 are shuffled. In following results 17 samples used from which 5 are from pool1 (Clone1N, liver2, Clone1o, Scr1o, Bone10). Only Bone2 seems to have the correct barcode.



Corrected??	Assignedname		ref.no	Original_Barcodes	2019_61_R1_Original_Data	2019_61_R2_Original_Data	2019_61_R3_Original_Data
Clone 1 N	Clone 1 N	AR001	15026633	ATCACG	GCCAAT		
Liver 2	Liver 2	AR002	15026634	CGATGT	ATCACG		
Bone 2	Bone 2	AR003	15026635	TTAGGC	TTAGGC		
Scrambled 1 N	Scrambled 1 N	AR004	15026636	TGACCA	ACTTGA		
Liver 6	Liver 6	AR005	15026637	ACAGTG	CGATGT		
Clone 1 O	Clone 1 O	AR006	15026638	GCCAAT	CAGATC		
Bone 6	Bone 6	AR007	15026640	CAGATC	TGACCA		
Scrambled 1 O	Scrambled 1 O	AR008	15026641	ACTTGA	GATCAG		
Bone 10	Bone 10	AR009	15026642	GATCAG	ACAGTG		
Clone 2 N	Clone 2 N	AR010	15026643	TAGCTT		TAGCTT	
Liver 4	Liver 4	AR011	15026644	GGCTAC		GGCTAC	
Bone 3	Bone 3	AR012	15026645	CTTGTA		CTTGTA	
Scrambled 3 O	Scrambled 2 N	AR013	15024655	AGTCAA		AGTCAA	AGTCAA
Bone 9	Liver 9	AR014	15024656	AGTTCC		AGTTCC	AGTTTC
Bone 12	Clone 2 O	AR015	15024657	ATGTCA		ATGTCA	ATGTCA
Bone 7	Bone 7	AR016	15024658	CCGTCC		CCGTCC	
Scrambled 2 O	Scrambled 2 O	AR018	15024660	GTCCGC		GTCCGC	
Bone 11	Bone 11	AR019	15024661	GTGAAA		GTGAAA	
Clone 3 N	Clone 3 N	AR020	15024662	GTGGCC			GTGGCC
Liver 5	Liver 5	AR021	15024663	GTTTCG			GTTTCG
Bone 4	Bone 4	AR022	15024664	CGTACG			CGTACG
Scrambled 3 N	Scrambled 3 N	AR023	15024665	GAGTGG			GAGTGG
Bone 5	Bone 5	AR025	15024667	ACTGAT			ACTGAT
Clone 3 O	Clone 3 O	AR027	15024668	ATTCCT			ATTCCT

Figure 4: List of barcodes and samples in groupes in three pools.

4.2 Trimming

By using Trimmomatic, we kept reads that both pairs survived the trimming and reads with +36 bases length. Also, we removed adapter sequences from the reads. Trimmed reads are located in
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/data/Trimmed-reads/

Also FastQC of the trimmed reads are located in
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/results/00-QC/FastQC_Trimmed/

4.3 Alignment and Post-alignment QC

We first aligned the reads on BALB_CJ and human genome separately. Then, we used XenofilteR with default values. This tool assigns the reads to corresponding genome based on edit distance. Reads with smaller number of mismatches will be assigned. By using this method we could rescue reads with more

reliable alignments and improve clipping rate of the reads. In original mappings, aligners clipped edges of the reads in order to find a better match on the genome. This resulted into mapping of shorter sequences on multiple location which in turn increases multi-maping rate in the alignment and we observed this issue in this dataset. Thus, it is important that reads be assigned to correct target and by using XenofilteR we could significantly improve this metric (Figure 5).

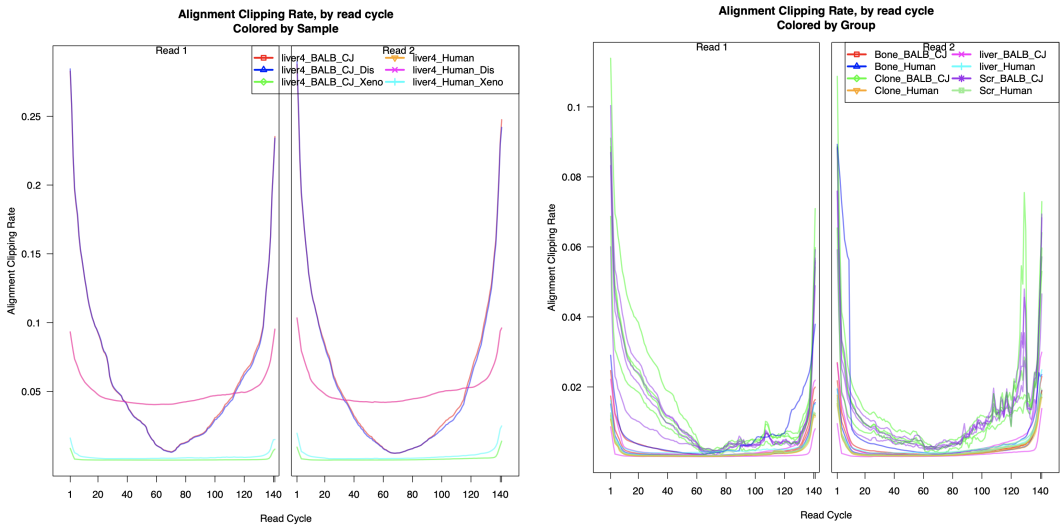


Figure 5: Clipping rate of reads left: V shape line shows a lot of clipping at the edges while U shape is more expected with lower rate at the edges. By using XenofilteR reads, with more reliable alignment to corresponding genomes, were rescued with significant improvement in clipping rate.

All the bam files are in
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/results/01-BAM/

4.4 Expression analysis

We extracted expression values of all the genes in annotation files by using featurecounts. After extracting the expression values we normalized the data and generated TMM values (trimmed mean of M-values) by edgeR to check overall expression pattern among all the samples. In addition to this dataset, we analyzed three PC3 cell line samples from Jividen et al. 2018(Jividen et al. 2018) (SRR7943936, SRR7943937, SRR7943938; all the generated bam files and expression values of these samples are saved together with the dataset in this project.).

Figure 6 shows clustering of the samples. Test samples tend to cluster together while all other samples are scattered both in the alignment on human and BALB_CJ genome. Mouse tissues tend to have a better clustering in the alignment on BALB_CJ genome which is expected except liver2 (liver2 is one of the samples from pool1 with inconsistent barcode).

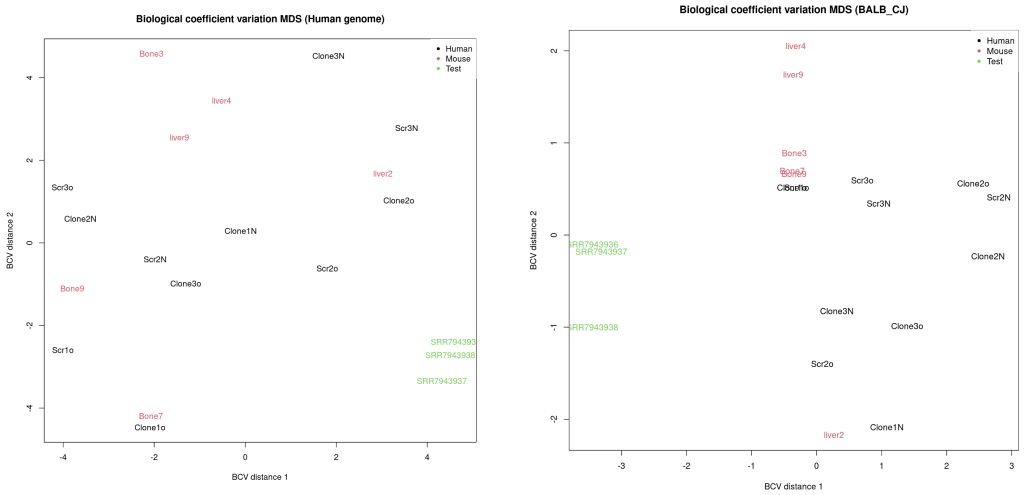


Figure 6: Multidimensional scaling (MDS) plot showing biological variation among samples (Left: human right: BALBCJ) .

We also checked the expression distribution of genes in all the samples. Figure 7 and 8 shows expression values distribution in human and BALB_CJ, respectively. Two genes were used to validate species specific expression that are highlighted in these figures; GUSB (blue) and PGK1 (red). qRT-PCR confirms species specific expression in selected number of samples while in some of the samples here we see inconsistent pattern and it is due to the barcodes issue highlighted before.

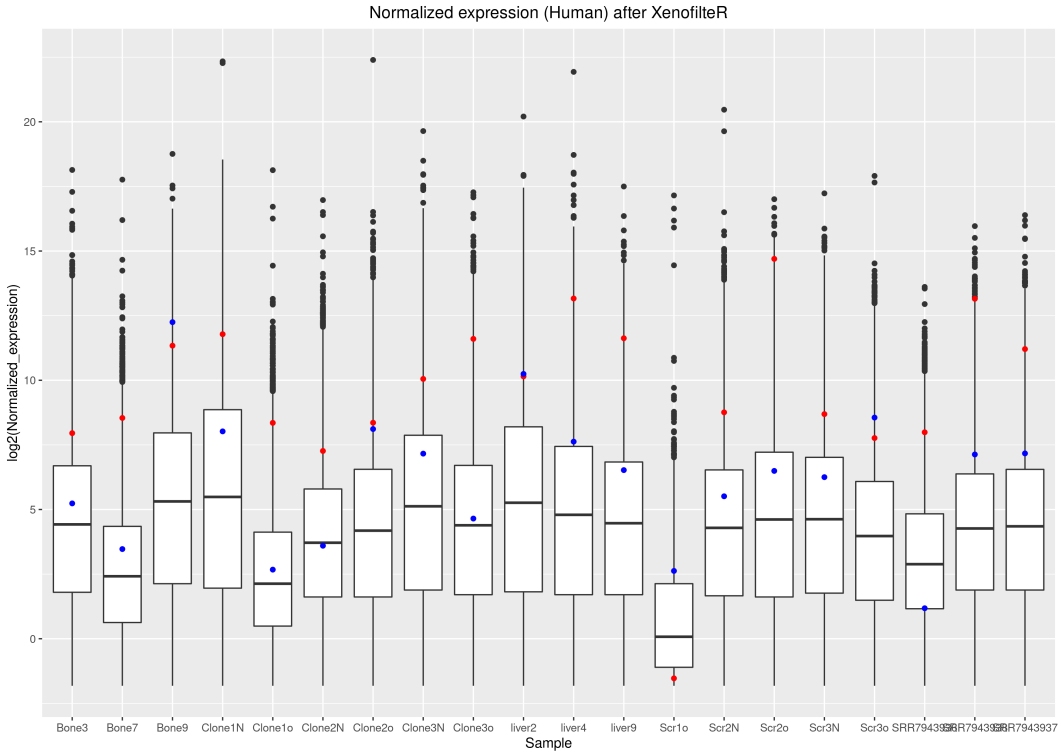


Figure 7: Expression distribution of genes in human genome across all samples. GUSB (blue), PGK1(red).

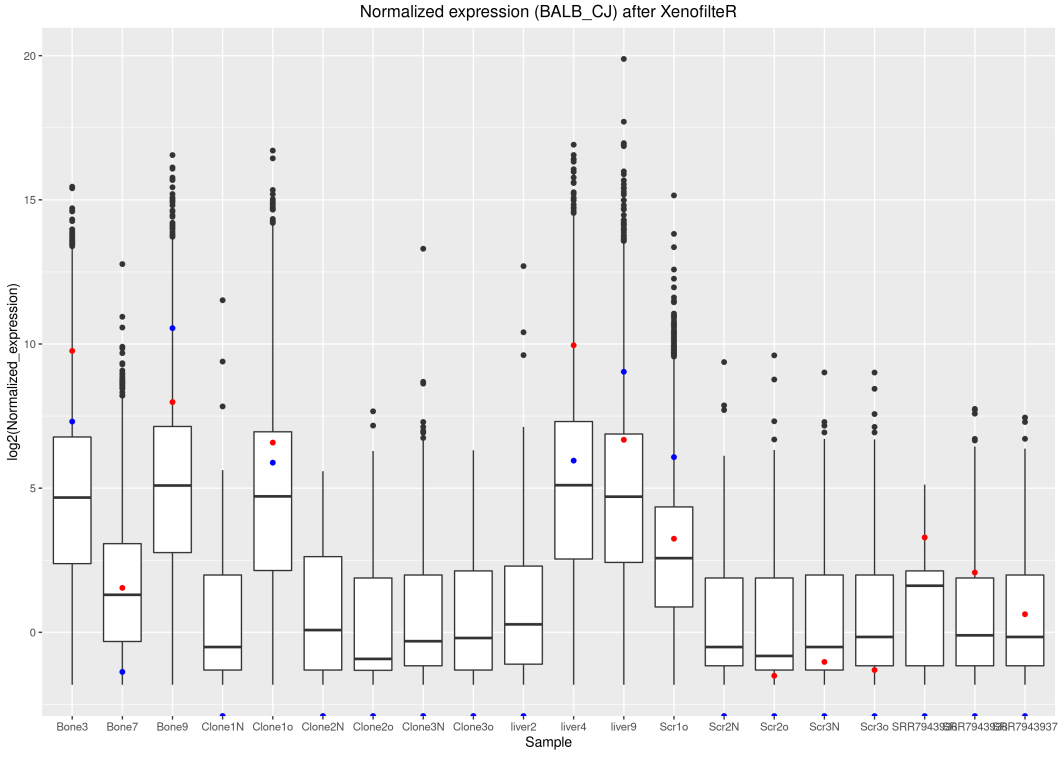


Figure 8: Expression distribution of genes in BALBCJ genome across all samples. GUSB (blue), PGK1(red).

Moreover, by looking into fraction of reads that were used for quantification of the genes or expression analysis is quite small (Figure 9). Mouse tissues show high fraction of reads assigned to mouse (BALB_CJ) features (genes) compared to human. Unexpectedly, this fraction is quite small in both genomes for human cell lines samples.

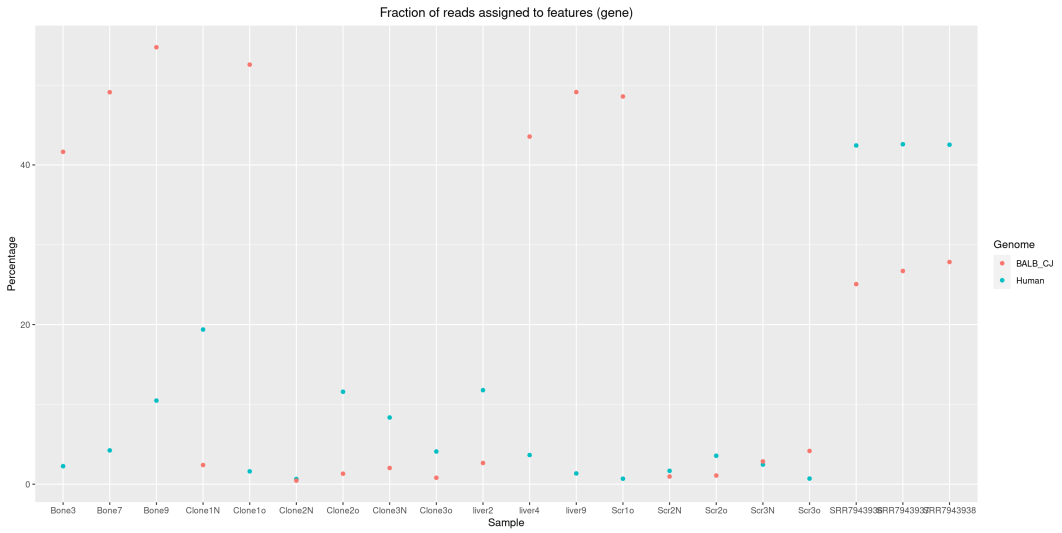


Figure 9: Fraction of reads assigned to features (genes) in both genomes.

5 Concluding remarks

The pipeline implemented in this project has improved the assignment of the reads to two different genomes (BALB_CJ and human). Issues hindering the downstream analysis in this project are:

- Shifted and a bit noisy distribution of GC-content.
- High rate of duplicated reads.
- Swapped barcodes which has affected samples in pool1. This also resulted in inconsistency between RNA-seq data and qRT-PCR results.
- High rate of multimapped reads which is result of high rate of duplicated reads.
- Highlighted issues has reflected in dispersed clustering of the samples.
- Quite varilable fraction of reads assigned to features in both genomes. It was mostly evident in human cell-line samples.

The downstream analysis and interpretation of the results are subjected to bias because of unexpected features observed in the data. Thus, by this report we have summarized the analysis pipeline and QC metrics used to evaluate the data. All the scripts, this report, and results are available on Uppmax:
/crex/proj/snic2019-30-25/private/UserDirectories/SMS_4882_19_Prostate_Bulk_RNA_Seq/

Also you can find scripts and this report and results (except bam files) on github:
https://github.com/NBISweden/SMS_4882_19_Prostate_Bulk_RNA_Seq

Scripts are under *code* directory:
GSNA_BALB_CJ/generate-commands.sh for alignment of reads on BALB_CJ genome.
GSNA_Human/generate-commands.sh for alignment of reads on human genome

6 Reproducibility

List of tools and packages used in this project:
- FastQC 0.11.9
- MultiQC 1.9
- Trimmomatic 0.36
- STAR 2.7.2b
- GSNAP gmap-gsnap/2017-09-11
- samtools 1.10
- QoRTs 1.3.6
- StringTie 2.1.4
- featureCounts 2.0.0
- Disambiguate 1.0
- XenofilteR 0.0.99

Matrix products: default BLAS: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
locale: [1] sv_SE.UTF-8/sv_SE.UTF-8/sv_SE.UTF-8/C/sv_SE.UTF-8/sv_SE.UTF-8 attached base packages:
[1] stats graphics grDevices utils datasets methods base
other attached packages: [1] dplyr_1.0.3 kableExtra_1.3.1 captioner_2.2.3 knitr_1.31 [5] reshape2_1.4.4 edgeR_3.32.1 limma_3.46.0 ggplot2_3.3.3
loaded via a namespace (and not attached):
[1] Rcpp_1.0.6 plyr_1.8.6 pillar_1.4.7 compiler_4.0.3 [5] tools_4.0.3 digest_0.6.27 viridisLite_0.3.0 evaluate_0.14 [9] lifecycle_0.2.0 tibble_3.0.6 gtable_0.3.0 lattice_0.20-41
[13] pkgconfig_2.0.3 rlang_0.4.10 rstudioapi_0.13 yaml_2.2.1
[17] xfun_0.20 xml2_1.3.2 httr_1.4.2 withr_2.4.1
[21] stringr_1.4.0 generics_0.1.0 vctrs_0.3.6 webshot_0.5.2
[25] locfit_1.5-9.4 grid_4.0.3 tidyselect_1.1.0 glue_1.4.2
[29] R6_2.5.0 rmarkdown_2.7 bookdown_0.21 purrr_0.3.4
[33] magrittr_2.0.1 scales_1.1.1 ellipsis_0.3.1 htmltools_0.5.1.1 [37] rvest_0.3.6 colorspace_2.0-0 stringi_1.5.3 munsell_0.5.0
[41] crayon_1.3.4

Reference

Ahdesmäki, Miika J., Simon R. Gray, Justin H. Johnson, and Zhongwu Lai. 2017. "Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples." *F1000Research* 5 (January): 2741. https://doi.org/10.12688/f1000research.10082.2.

Andrews, S. n.d. "FastQC A Quality Control tool for High Throughput Sequence Data." citeulike-article-id:11583827%20http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bushnell, Brian. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner," March.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Hartley, Stephen W, and James C Mullikin. 2015. "QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments." *BMC Bioinformatics* 16 (1): 224. https://doi.org/10.1186/

s12859-015-0670-5.

Jividen, Kasey, Katarzyna Z Kedzierska, Chun-Song Yang, Karol Szlachta, Aakrosh Ratan, and Bryce M Paschal. 2018. "Genomic analysis of DNA repair genes and androgen signaling in prostate cancer." *BMC Cancer* 18 (1): 960. https://doi.org/10.1186/s12885-018-4848-x.

Kluin, Roelof J C, Kristel Kemper, Thomas Kuilman, Julian R de Ruiter, Vivek Iyer, Josep V Forment, Paulien Cornelissen-Steijger, et al. 2018. "XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data." *BMC Bioinformatics* 19 (1): 366. https://doi.org/10.1186/s12859-018-2353-5.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2014. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Wu, Thomas D, and Serban Nacu. 2010. "Fast and SNP-tolerant detection of complex variants and splicing in short reads." *Bioinformatics* 26 (7): 873–81. https://doi.org/10.1093/bioinformatics/btq057.