**Bioinformatics analysis and integration of all data set into a model**

For each next generation sequencing analysis, we are doing the following: quality check and cleaning of the reads with FastQC (1) and cutadapt (2) programs followed by alignment of the reads to TAIR 10 genome (<https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release> )

using TopHat2 (3) and Bowtie2 (4) for RNA-seq and edgeR (5) and/or DEseq2 (6) packages in R program for ChIP-seq.

 Peak calling of ChIP-seq data are done using MACS2 (7) program and association of peaks to genes are done with BEDTools program (8). Peak pattern are analyzed using deeptools 2.0 program (9). Statistically significant changes in peak pattern between the genotypes are analyzed using the edgeR package algorithm.

For visualization of the data we are using IGV (10).

For mRNA-seq:

To obtain a broad view on the differences between the transcripts profiles of leaves and calli, a Principal Component Analysis (PCA) is conducted

For example when we compered callus and Leaves mRNA-seq in three replicates preforming the PCA samples profiles could be clearly distinguished



To do Gene Ontology analysis we are using the BiNGO application on Cytoscape platform (11)

**Constructing a model for epigenetics marks regulating transcriptional states**.

Following the bioinformatics analysis we are generating tables with list of genes according to the questions that we are asking. Then we merged the data and extract the information we need.

For example we would like to know what genes harbor the H3K27me3 marks and are not expressed in WT callus, but in the *emf2* mutant callus acquire H3K4me3 and are expressed. For this set of genes there might be a competition between the PRC2 and the TrxG complexes on the binding site or that the marks itself can prevent the TrxG from setting the H3K4me3 mark.

We are making the following lists:

1. All genes marked by H3K27me3 in WT callus

2. All silenced genes in WT callus

3. All the expressed genes in the *emf2* mutant callus

4. All the genes marked with H3K4me3 in the *emf2* mutant callus

Then we combine table 1 and 2 and extract the genes appeared in both: having the mark and are silenced

Next we combine table 3 and 4 and extract the genes appeared in both: having the mark and are expressed

Next we combine the two lists to extract the genes that in WT are marked by H3K27me3 and silenced and in the *emf2* are marked by H3K4me3 and are expressed

As soon as we will have the genome wide binding sits for EMF2 in WT callus and in the TrxG triple mutant callus as well as the binding sites for the TrxGs in WT and in the *emf2* mutant we can conclude on the competition and the interaction between the two complexes and between the complexes and the opposing marks.

References:

1. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>, 2010.

2. Marcel, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.

3. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biology 2013. **14**(4)

4. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, . **9**(4): p. 357-U54

5. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

6. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12).

7. Zhang, Y., et al., *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology, 2008. **9**(9)

8. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

9. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic Acids Research, 2016. **44**(W1): p. W160-W165.

10. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Briefings in Bioinformatics, 2013. **14**(2): p. 178-192.

11. Maere S, Heymans K, Kuiper M. *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. 21(16):p.3448-9