# Lab 4: Using Chromatin Immunoprecipitation – DNA Sequencing Data to Identify Genes Directly Regulated by Phytochrome Interacting Factors.

In the past three labs, we have analyzed data that has helped us to address the following biological questions.

**How do PIFs regulate hypocotyl elongation?**

**How do PIFs integrate environmental signals?**

We identified sets of genes whose expression was significantly affected by the pifq mutations, determined how expression of the PIF-regulated genes was coordinated in response to various stresses, and identified PIF-regulated genes that may be involved in controlling hypocotyl elongation. To understand the regulatory hierarchy controlling the expression of the PIF-regulated genes, we will identify genes bound by the PIF4 transcription factor in planta. Integrating the list of PIF-regulated genes with a list of genes bound by PIF4 will allow us to address the final biological question.

**What are the direct targets of PIF transcription factors?**

Sites in the genome that are bound by a specific DNA binding protein in vivo can be determined using chromatin immunoprecipitation - DNA sequencing (ChIP-seq) experiments. In brief: 1) the tissue of interest is treated with a crosslinking agent to covalently attach DNA binding proteins to DNA, and nuclei are isolated from the tissue; 2) the chromatin is sheared to small fragments; 3) the DNA binding protein crosslinked with DNA is immunoprecipitated with an antibody against the DNA binding protein; 4) the DNA is

isolated after removal of crosslinks; and 5) the DNA is sequenced. Analysis of the data reveals sites in the genome bound by the DNA binding protein. This approach can be used to identify sites bound by transcription factors, such as PIF4, and modified histones that influence chromatin structure.

We will identify PIF4 binding sites in this lab using the ChIP-seq data described in the Oh et al. (Nature Cell Biol. 14: 802) paper. Be sure to include in your lab report the genotypes of the plants and the antibody that was used for the ChIP experiments. The ChIP-seq data will be analyzed using the computational tool, Model-based Analysis for ChIP Seq (MACS; Genome Biol. 9: R137; http://liulab.dfci.harvard.edu/MACS/). A variety of other computational tools will be used to analyze the ChIP-seq data.

## There are five primary objectives of this lab.

1. Identify PIF4 binding sites using MACS.
2. Analyze the distribution of binding sites relative to the transcription start sites of genes.
3. Identify consensus DNA binding sequences for PIF4.
4. Associate binding sites with genes.
5. Identify genes directly regulated by PIF transcription factors.

# Required Files

**ATTENTION BIS180L** You can download all the files you need for this tutorial as a .zip file in smartsite. `Unit 6 > requiredFiles_smartsite.zip` .

## File Summary

**Output from running `macs callpeak` on `PIF4_ChiPSeq.bam` and `Col-0_ChiPSeq.bam` (control)**

- `PIF4_macs2_peaks.bed` : Peaks found, which are ranges of overlapping reads.

- `PIF4_macs2_summits.bed` : Summits found, which are the single nucleotide tips of these peak

ranges.

- `PIF4_macs2_peaks.xls` : Contains a lot of information including all the peaks, description of all parameters used for peakfinder, and much more.

**Obtained from TAIR**

- `TAIR10_genes.bed` : Coordinates of annotated genes.
- `TAIR10_TSS.bed` : Coordinates of Annotated Gene Transcription Start Sites.
- `Ath.genome.txt` : Chromosome lengths.
- `Athaliana_167.fa` : *Arabidopsis thaliana* genome sequence. This file is too large for Github, but can also be downloaded here https://www.dropbox.com/s/d3ts7qrb136b7ql/Athaliana_167.fa. If you downloaded the requiredfiles.zip file from smartsite, it is included in the file.

**Files needed for step one** *files are given above. No need to download these unless you really want to try with the raw data*

From Dropbox: PIF4_ChiPSeq.bam, PIF4_ChiPSeq.bam, and Athaliana_167

---

# Module 1: Identifying DNA sequences Bound by PIF4 in planta

PIFs are basic helix-loop-helix transcription factors that bind with specific DNA sequence motifs to activate or repress the transcription of genes. In this module, we will analyze the ChIP-seq data from Oh et al. paper using MACS to identify statistically-significantly bound regions by PIF4. The numbers of reads from this dataset are given below. Col-0 ChIP seq is the control.

| | # total reads | # uniquely mapping reads |
|---|---|---|
| PIF4_ChIPSeq. Bam | 39,463,565 | 13,208,418 |
| Col-0_ChIPSeq.bam | 14,556,313 | 3,417,136 |

Once PIF4 binding sites have been identified, we will use two tests to assess the validity of these binding sites. First, transcription factors will often but not always bind with the 5' flanking region of genes to regulate transcription. We will use software from the BEDTool suite (http://bedtools.readthedocs.org/en/latest/ ; Bioinformatics 26: 841) to determine the position of PIF4 binding sites relative to the transcription start site of genes.

Second, PIF4 is a DNA sequence-specific binding protein. If the ChIP-seq experiments are valid, the expectation is that a specific DNA motif will be enriched in the PIF4 bound regions. We will use a de novo motif finder, MEME-ChIP (http://meme.nbcr.net/meme/cgi-bin/meme-chip.cgi) to identify a consensus binding sequence for PIF 4.

# Exercise #1: Identify PIF4 binding sites (designated "peaks" by MACS).

**Attention BIS180L - You will not be running the first command because of file input size and computational time.**

1. Use the `callpeak` function to identify PIF4 binding sites. You must designate your treatment and your control.

```
macs callpeak -t PIF4_ChiPSeq.bam -c Col-0_ChiPSeq.bam -n PIF4 -g 1.118e8 -q 0.05
```

2. Look up the usage at from at MACS. Make notes on what parameters were run on the above command to obtain the output files that were given to you.

3. From this analysis, you will obtain five files, three of which will be used for the

subsequent analyses: `PIF4_macs2_peaks.xls`, `PIF4_macs2_peaks.bed` and `PIF4_macs2_summits.bed` and `PIF4_macs2_peaks.xls files`. Here is a guide to BED formated files.

4. The `PIF4_macs2_peaks.xls` file contains a description of all of the parameters used for the analysis and summarizes information generated by the analysis. Look at the output file section in MACS2 documentation for more information on what is contained in this file.

## Questions (Answers to these questions should be included in the Discussion section of your lab report).

**Question #1-1**: Why were plants of the indicated genotype used for the ChIP experiment? Why was the antibody selected for use in this experiment? The plants used for the ChIP experiments were grown differently than were the plants used for the RNA-seq experiments. Is this difference likely to influence the interpretation of results? Refer to Oh *et al*, 2012.

**Question #1-2**: What ChIP-seq reads were compared to obtain the PIF4 binding sites?

**Question #1-3**: How does the number of binding regions that you calculated compare with the number reported in the paper? Is the average width of the PIF4 peaks of a reasonable size? Refer to Oh *et al*, 2012. Find the average width of you PIF4 peaks in R using the `PIF4_macs2_peaks.bed` file. Explain your answer.

**Question #1-4**: Discuss how the p-value, the q-value, and the fold-enrichment for the binding sites relate to each other. Examine information in `PIF4_macs2_peaks.xls` file.

# Exercise #2: Analyze the distribution of PIF4 binding sites relative to the transcription start site of genes

Now we are going to use the BEDTOOLS, a powerful suite of tools to deal with genomic data. We want to know what genomic regions are closest to the peaks found earlier. In order to accomplish this we will use the `bedtools closest` function. First look at the usage manual page by typing `bedtools closest` and read about the general usage of the tool.

1. Sometimes files are not formatted exactly how you want them, especially when changing between different operating systems. You can get inserts symbols that you can't necessarily see unless you look for them. For instance when working with MACS2, sometimes it inserts these pesky `^M`s which cause line breaks in the wrong places. Want to see them? Use the text editor `Vim`

   ```
   vi -b TAIR10_TSS.bed
   ```

   To get rid of them type the command below. Where `<Ctrl-V>` is literally holding down `ctrl` and pressing `V`. Same with `<Ctrl-M>`.

   ```
   :%s/<Ctrl-V><Ctrl-M>//g
   ```

   To save and quit type

   ```
   :wq
   ```

   Check the `PIF4_macs2_summits.bed` file too. Are they there? Without checking for these, the output can be wrong. How do you know when it will mess up your results? You have to just be hyper aware of formatting inputs and outputs of the program you are using. In future work: always refer to manuals and check files, run them with and without formatting differences and see if this changes your results.

2. Now you can run the closestBed function using the `PIF4_macs2_summits.bed` and the `TAIR10_TSS.bed file`, which will give you an idea of where the summits are in relation to the transcription start sites. Consult the usage manual and make notes on what the

below command is doing.

bedtools closest -D “b” -a PIF4_macs2_summits.bed -b TAIR10_TSS.bed > closestSummitsOut.bed

3. This creates `closestSummitsOut.bed`.

```
vi -b closestSummitsOut.bed
```

To get rid of them type command below. Where `<Ctrl-V>` is literally holding down `ctrl` and pressing `V`. Same with `<Ctrl-M>`.

```
:%s/<Ctrl-V><Ctrl-M>//g
```

To save and quit type

```
:wq
```

4. Now let’s do some visualization. Make a histogram of the `closestSummitsOut.bed` file generated in step 1. We want to get an idea of how far from the start codon of a gene is PIF4 binding.

a. Upload `closestSummitsOut.bed` into R. Use `read.table()` and set header to false.

b. Make sure you understand the columns. In this case V12 is the distance from the transcription start site.

c. Make histogram that illustrates where in relation to the transcription start site of genes does are the PIF4 binding summits (center of PIF peaks) are found.

# Questions

**Question #1-5**: Where are most PIF4 binding sites relative to the Transcription Start Site (TSS)? Explain whether the distribution of binding site is what you would expect of a transcription factor.

**Question #1-6**: In Module 8, you will identify genes whose transcription is thought to be directly regulated by PIF4. Based on the distribution obtained here, specify the maximum distance that a gene directly regulated by PIF is likely to be from the PIF4 binding site.

# Exercise #3: Identify consensus sequence for regions bound by PIF4

1. Use the `fastaFromBed` tool from bedtools to extract the DNA sequences of your PIF4 binding sites. This tool will extract the sequence for the genomic intervals specified in the peak.bed files (and also the random peak file they will generate with the following tool) and return a fasta format file. (using your `PIF4_macs2_peaks.bed` file and the genome sequence file `Athaliana_167.fa)`. This analysis will produce a .fasta file containing the nucleotide sequences of all of the regions bound by PIF4 in this experiment.

```
fastaFromBed -fi Athaliana_167.fa -bed PIF4_macs2_peaks.bed -fo peakSeq.fa
```

2. Submit the PIF4 bound DNA sequences to the MEME-ChIP website at: http://meme.nbcr.net/meme/cgi-bin/meme-chip.cgi or . Make sure to enter your email address and a description of the sequence set in the allocated space before submitting your job. You will receive an email message with a hyperlink to your result page. (It may take a while for the results from the analysis to be available.)

3. Run the `shuffleBed` tool on the `PIF4_macs2_peaks.bed` file to generate nucleotide sequences of a set of randomly selected genomic regions that have a similar size distribution as your PIF4 peaks.

```
shuffleBed -i PIF4_macs2_peaks.bed -g Ath.genome.txt > randomPeaks.bed
```

Now convert `randomPeaks.bed` to fasta as you did above.

```
fastaFromBed -fi Athaliana_167.fa -bed randomPeaks.bed -fo randomPeakSeq.fa
```

4. Submit the random peak sequences obtained with fasta from bed to MEME-ChIP.

## Questions

**Question #1-7**: What is the relationship between the most significant DNA motif discovered and the known binding sites of PIF transcription factors? How does the analysis performed with randomly selected DNA sequences affect your conclusion about the motif discovery experiment?

**Question #1-8**: How many peaks contain the most significant DNA motif discovered? Speculate about the biological significance of PIF4 binding sites that do not contain this motif.

# Module 2: Identification of Genes Directly Regulated by PIF Transcription Factors

In the previous exercises, we identified DNA sequences that were bound by PIF4 and the relative distribution of these regions to the transcription start sites of genes. We will use this information to associate DNA binding sites to specific genes using a function that assigns DNA binding regions to the closest gene.

We will then compare the list of genes bound by PIF4 with the lists of genes that are statistically-significantly differentially expressed, both upregulated and downregulated, in pifq

mutants relative to wild type. Tests will be conducted to determine whether a statistically significant number of PIF-regulated genes are bound by PIF4.

# Exercise #4: Identify genes that are bound by PIF4 and regulated by PIFs.

1. Use the `bedtools closest` function with your `PIF4_macs2_peaks.bed` file and the `TAIR10_genes.bed` reference file. This will assign peaks to the closest *A. thaliana* annotated gene.

```
bedtools closest -D "b" -a PIF4_macs2_peaks.bed -b TAIR10_genes.bed > closestPeaksOut.
```

2. You will obtain a file that lists for each peak 1) the coordinates and the gene ID of the closest gene and 2) the distance (in bp) that separates the peak from the gene. (A distance of 0 means the peak and the overlap by at least 1 bp. The distance given can be from the 5' or 3' end of a gene.)

3. In Question #1-6, you specified the maximum distance that a gene directly regulated by PIF is likely to be from a PIF4 binding site. Use this distance and sort using the `subset()` function in R to generate a list of genes that are specifically associated with PIF4 binding sites. *don't forget to check for* `^M` *s when loading the file*

4. Designate this file as the "PIF4-bound" **gene list**. Remove duplicates from the "AGI" column with the `duplicated()` and `unique()` functions in R.

---

## How to use on `duplicated()` in this context**

1. Check out this small summary of the funtions: http://www.cookbook-r.com/Manipulating_data/Finding_and_removing_duplicate_records/

2. Ask what if there are duplicates in the gene column (V9) and see the dimentsions of the

file to get an idea of how large it is.

```
duplicated(dataframe$V9)
dim(dataframe)
```

3. Now we want to specify to remove the duplicates, this gets a little tricky, but make sure you understand what is going on.

```
uniqueDataframe <- subset(dataframe,!duplicated(dataframe[,9]))
```

4. Now double check this auctually did something by checking the dimensions of new dataframe.
5. Does this make sense? Why couldn't we just use the `unique()` function?
6. Identify genes that are direct targets of PIF transcription factors.

---

1. Create separate lists of genes that are upregulated and downregulated in the pifq mutant using the fold-change and FDR that you consider to be most relevant. *Use table generated in previous lab*.
2. Using the web-based tool Venny http://bioinfogp.cnb.csic.es/tools/venny/, identify genes that are in both the "PIF4 bound" and "Upregulated" gene lists. Repeat for the "Downregulated" gene list.

# Questions

**Question #2-1**: Why is it necessary to know the sites in genome that are bound by a transcription factor to identify direct target genes?

**Question #2-3**: Discuss the significance of the over- or under- representation of PIF-regulated genes that are bound by PIF4. Based on the results, speculate on whether PIF4

more likely to activate or repress genes in wild-type plants.

**Question #2-4**: What do the results suggest about the role of the direct target genes of PIFs in controlling hypocotyls elongation?