Computational methods for 4C-seq NGS data analysis

Dimitrios Zisis, BSc, MSc,
PI: Prof. Pawel Krajewski
Institute of Plant Genetics, PAS

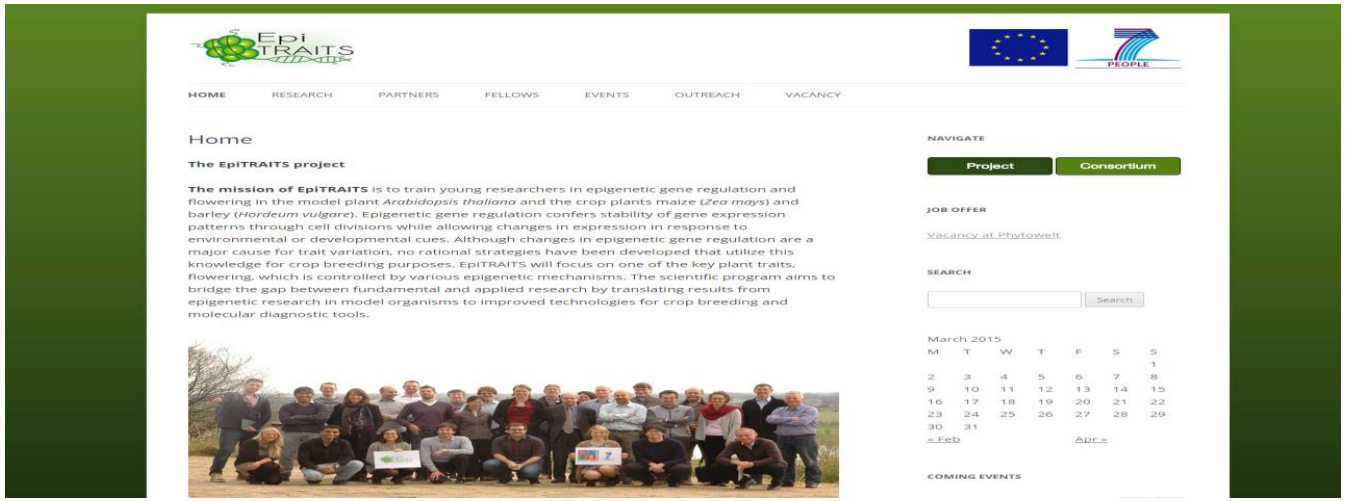
27 June 2016, Poznan

Outline

- EpiTRAITS
- NGS data
- 4C-seq data
- Computations
- Existing Methods
- 4CseqR Analysis
- Conclusions

EpiTRAITS

- **The mission of EpiTRAITS** is to train young researchers in epigenetic gene regulation and flowering in the model plant *Arabidopsis thaliana* and the crop plants maize (*Zea mays*) and barley (*Hordeum vulgare*).
- EpiTRAITS focus on one of the key plant traits, **flowering,** which is controlled by various epigenetic mechanisms.
- The scientific program aims to **bridge the gap between fundamental and applied research** by translating results from epigenetic research in model organisms to improved technologies for crop breeding and molecular diagnostic tools.
- **http://www.epitraits.eu/**



NGS data

- **DNA sequencing** is the process of determining the precise order of nucleotides within a DNA molecule
- Fields:
    i. medical diagnosis,
    ii. biotechnology,
    iii. forensic biology,
    iv. virology and biological systematics.

- The first DNA sequences were obtained in the early 1970s

- High demand for low-cost sequencing has driven the development of **high-throughput sequencing or next-generation sequencing (NGS)**

- NGS are techniques that parallelize the sequencing process, producing thousands or millions of sequences - lower the cost of DNA sequencing.

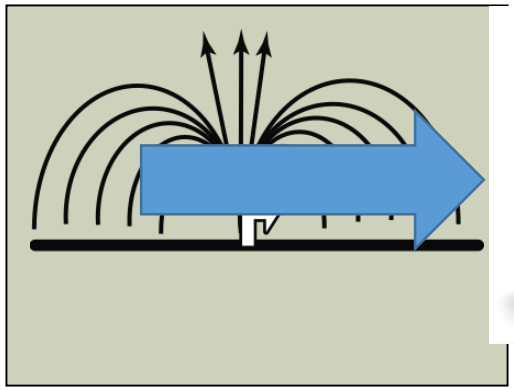
- Next-generation sequencing applications used : Chip-seq, RNA-seq, **4C-seq**

4C-seq data

- 4C-seq technology allows to map the physical interactions' landscape of a given site of interest in a genome-wide manner.

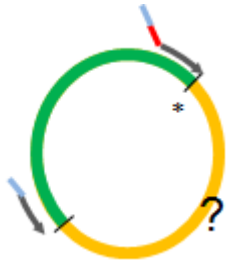
- Goal: Identification of all regions in the genome that contact a genomic site of interest, referred to as **“viewpoint”** or **“bait”**

- Regions that are cross-linked and ligated to the viewpoint are called **“captures”**



4C
One-to-all

Next Generation Sequencing



Computations

- Several data files with size from 4-10 gigabyte (fastq, sam, bam).
- Fastq files represents data sets of NGS reads of length 50, with 11,000,000 to 14,000,000 reads per sample.
- Used tools : R, Bash, Python, Bowtie2, eXpress, samTools
- Used scripts written in R and BASH, running in parallel on a cluster (REEF and INULA) in the Poznan Supercomputing and Networking Center.
- Average time for each script : From 1 minute to 4 hours depending from the parallel processes in each case.
- Used tools and terminal of Ubuntu Linux for filtering and treatment of the data.
- The final data after the analysis in cluster were used for further statistical analysis using script in R and in GenStat 17.

Existing Methods

- Over last few years various methods for the analysis of 4C-seq data have been developed.

- All methods provide basic algorithms for the preprocessing of NGS reads, the creation of in-silico library of restriction fragments and of fragment ends, and alignment (using own or public domain mappers).

- Most important methods:

✓ ***4Cseqpipe*** of van de Werken et al. (2012)

✓ ***FourCseq*** of Klein et al. (2015).

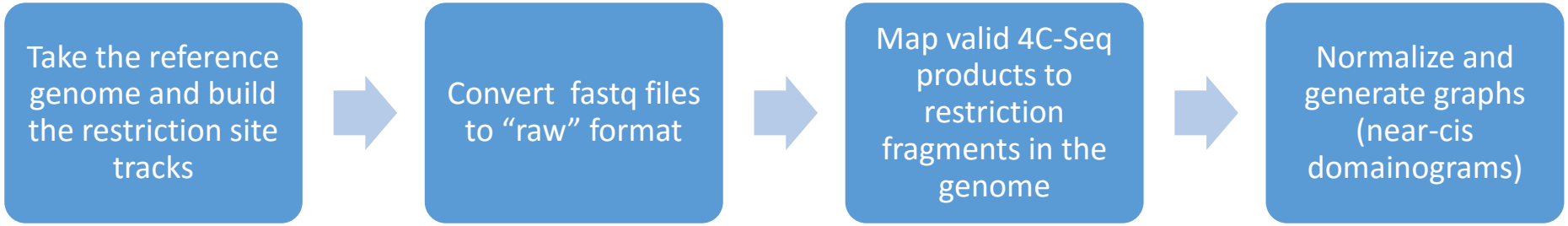
✓ ***fourSig*** of Williams et al. (2014)

✓ **4Cker** of Raviram R et al. (2016)

- We analyzed each step of those methods to identify differences and similarities in the algorithms

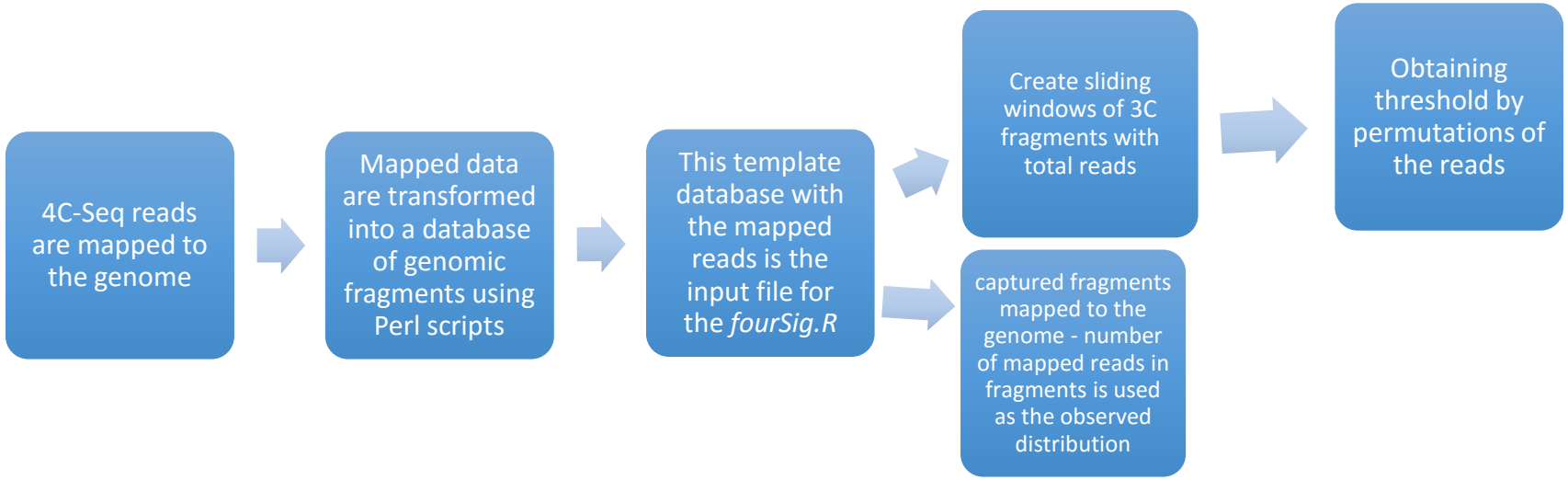
Existing Methods

- *4Cseqpipe* of van de Werken et al. (2012) provides a full schema for analyzing 4C-seq experiments,
- It considers weighted mapping of the reads according to the uniqness of the fragments
- Different normalization methods for remote and for near contacts.
- *4Cseqpipe* is a set of for several programs



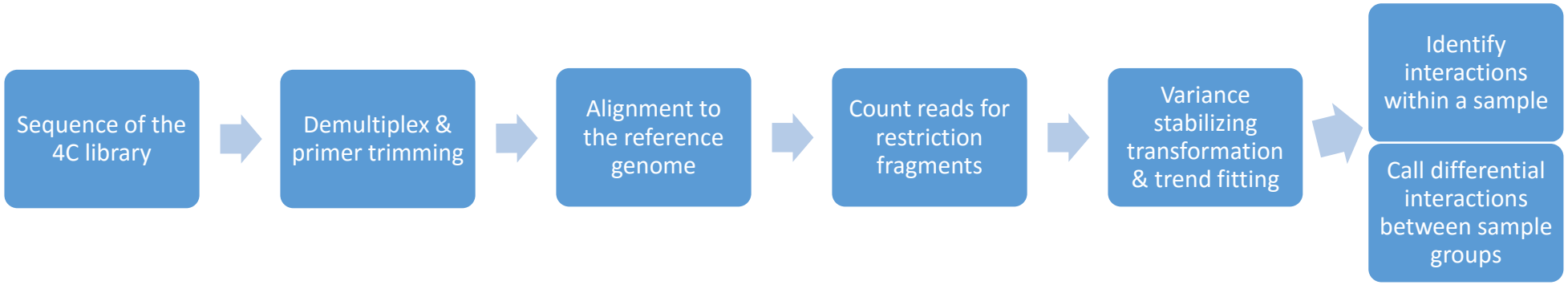
Existing Methods

- *fourSig* of Williams et al. 2014 is a software suite, written in Perl and *R*
- Goal : Detect and identify statistically significant interactions from 4C-Seq data
- A template database is produced with the mapped reads from the 4C-Seq data and this is used as the input file for the *fourSig.R* program
- A method for determining significant interactions and interactions based on the likelihood that they are reproducible.



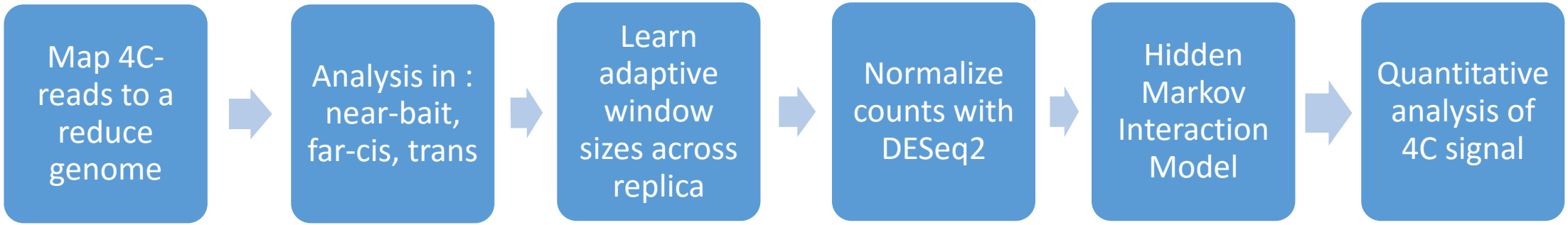
Existing Methods

- *FourCSeq* of Klein et al. (2015) is a computational method developed in R that provides a pipeline for the analysis of multiplexed 4C sequencing data.
- Goal: To detect specific interactions between DNA elements and identify differential interactions between conditions.
- The statistical analysis in R starts with individual „bam" files for each sample as inputs and ends with the contact profiles.
- Provides also a function in R for the statistical differential analysis of the contact profiles (based on the DESeq2 algorithm designed earlier for RNA-Seq data analysis).



Existing Methods

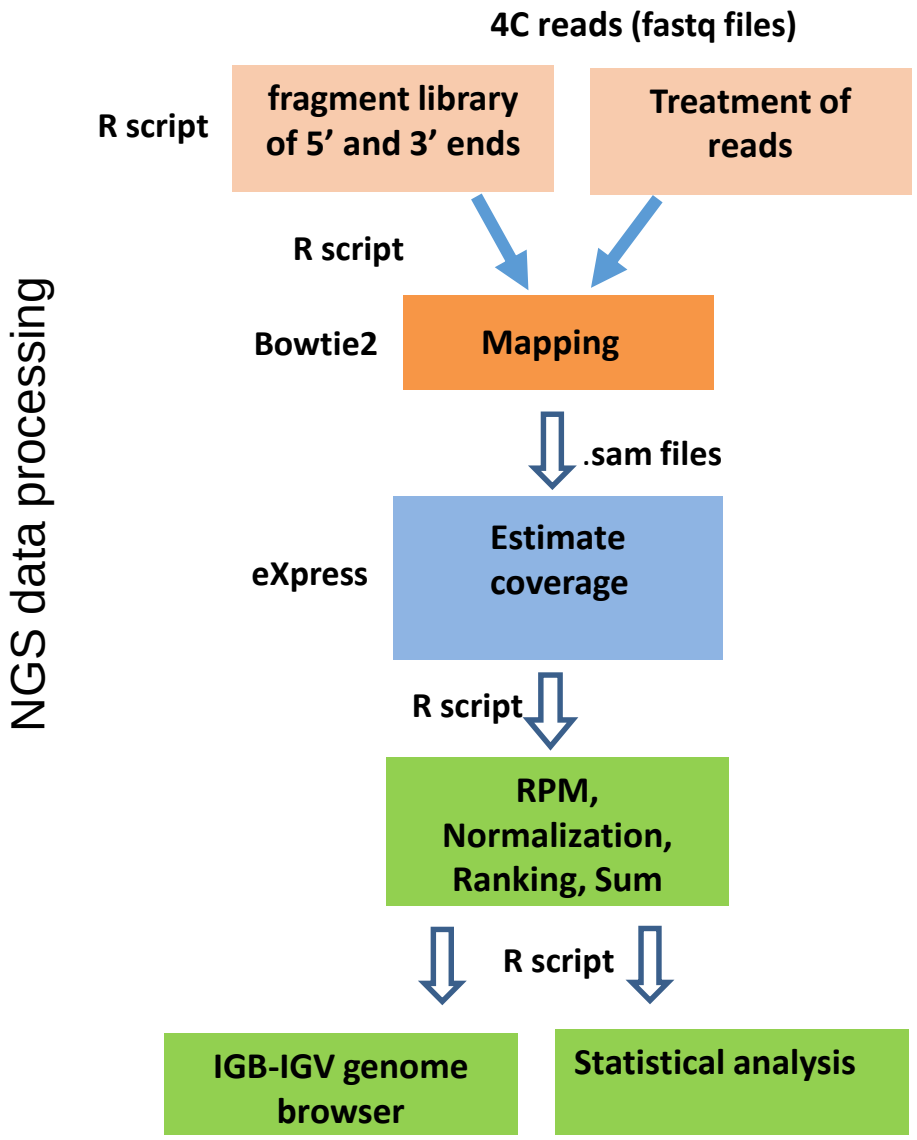
- **4Cker** of Raviram R et al. (2016) is a computational method developed in R that provides a pipeline for identification of regions across the genome that interact with 4C bait.
- Goal: To detect specific interactions between DNA elements and identify differential interactions between conditions.
- 4C-ker is using the Hidden Markov Model to partition the genome into windows that interact with the bait.



4CseqR Analysis

- Application of above methods for *Arabidopsis* data revealed some limitations of each method.
- Need for a full data analysis procedure including comparison of different samples.
- For this reason we developed a 4C-seq data processing schema — 4CseqR.
- Characteristics of the pipeline :
  - ✓ Based on tools commonly used for NGS data analysis,
  - ✓ The main algorithm is developed anew in R
- Usage :
  - ✓ For data for different viewpoints, genotypes or conditions.
  - ✓ Experiments including replications that allow to test the significance of differences between treatments.

4CseqR Analysis



After the design of fragment libraries next step is to use own scripts in R and existing tools to:

1. Find the reads from the original files which are legitimate 4C reads (**primer + restr.site)**
2. Mapping to library of fragment ends (50 bp) – **all mapping positions, up to 2 mismatches, no mismatches** in the restriction sequence
3. Processed by **eXpress** to estimate fragment end coverage
4. Normalize coverage by computing ranks within classes of **blind-non blind, short-long fragment ends, short-long fragments**
5. Ranks within each category are expressed **from 0 to 1**
6. Sum of coverage of 5' and 3' end. **Total coverage from 0 to 2**
7. Use this results to **visualize** in a genome browser (IGV)
8. **Perform Statistical analysis** by own script in R for the :
    1. Analysis of Variance (ANOVA)
    2. Pairwise Comparison
    3. Adjustment of p-values
    4. Graph and plot generation
    5. Additional data tables for further exploration

Conclusions

After the analysis and comparisons we can conclude that:

• Before normalization, our method (*4CseqR*) produces results similar to the results of *4Cseqpipe*, whereas after normalization the computed coverage is different- Normalized results closer to the *FourCseq*.

• Different normalizations may have a large effect on comparisons between samples (differential analysis).

•Parallel processing in INULA Cluster - Speeding up computations and data analysis.

•Stable environment and bigger storage space in Cluster.

•Synchronize – Share working scripts and data with other people.

Acknowledgments

UvA:

- Maike Stam
- Iris Hövel
- Rurika Oka

IGR PAN:

Paweł Krajewski

Hanna Ćwiek-Kupczynska

Aneta Sawikowska