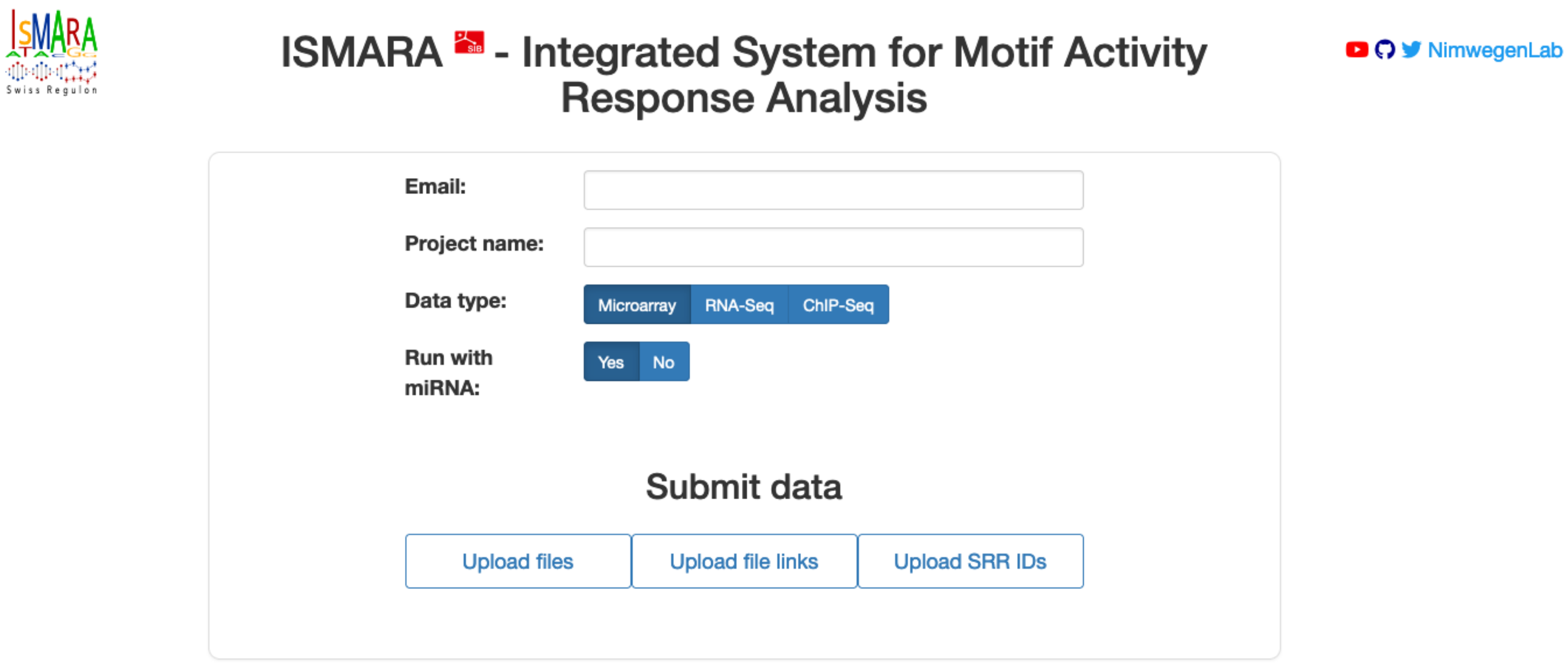
BIOZENTRUM
The Center for
Molecular Life Sciences

Using the ISMARA and CREMA web interfaces.



ISMARA - Integrated System for Motif Activity Response Analysis

Email:

Project name:

Data type:

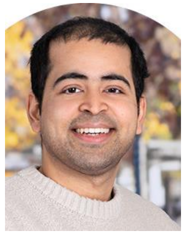
Run with miRNA:

Submit data

Erik van Nimwegen

Mikhail Pachkov

Daan de Groot



Anurag Ranjak

@NimwegenLab

SIB

BIOZENTRUM
The Center for
Molecular Life Sciences

Agenda

- ISMARA genomes and data type support.
- ISMARA upload interface.
- ISMARA uploader.
- CREMA genomes and data type support.
- CREMA upload interface.
- CREMA uploader.
- Averaging replicates, batch effect correction.
- Calculating contrasts between sample groups.

ISMARA: supported species

	Human	Mouse	Rat	Zebrafish	Arabidopsis	Yeast	E. Coli
Promoterome	hg38 + F5	Mm39 +F5	rn6	dr11	TAIR10	S288C R61	RegulonDB 9.3
Genes	20209	22308	22045	25103	31434	4796	4490
Transcripts	68273	49800	28727	44803	52148	6575	4490
Motifs	499	503	503	475		158	
TFs	682	679	650	832	578	158	210
miRNAs	106	99					

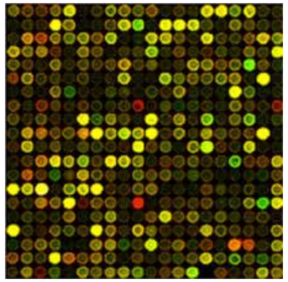
ISMARA: supported data types

Next Generation Sequencing

Mapped reads: .bam and .bed files

Raw reads: .fastq files

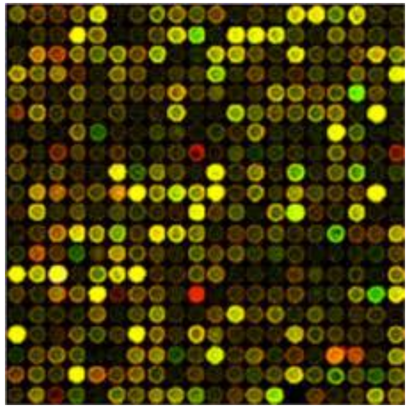
Microarray



Affymetrix .cel files

For human, mouse, rat, yeast, E. coli

ISMARA: microarray processing



- Correction for background and unspecific binding (BioConductor: affy, oligo, gcrma).
- Filtering out non-expressed probes.
- Quantile normalization.
- Log-transformation.

ISMARA: raw read processing (fastq)

RNA-Seq

- Map reads to transcriptome with kallisto algorithm (Bray et al, 2016).
- Count reads per transcript.
- Calculate TPM values for every promoter.
- Log-transform the data.

ChIP-Seq

OBSOLETE

- Map reads to promoter regions with kallisto algorithm (Bray et al, 2016).
- Count reads per promoter region.
- Quantile normalize the counts.
- Log-transform the data.

ISMARA: mapped reads processing (bam/bed)

RNA-Seq

- Count reads per transcript using absolute genomic coordinates.
- Calculate TPM values for every promoter.
- Log-transform the data.

ChIP-Seq

- Count reads per promoter region using absolute genomic coordinates.
- Quantile normalize the counts.
- Log-transform the data.

Please submit raw reads instead of mapped data!

ISMARA file format support

- Supported file formats:

  .cel, .bam, .bed, .fastq (**Proper file extension is important!**)

- File compression support:

  .gz, .tar, .bz2, .zip, .tar.gz

Before submitting mapped reads (bed/bam) make sure that they are mapped to the genome version used by ISMARA!

Shall I compress my files?

Yes! Compressing files significantly reduces the upload time.

- Compress: .cel, .bed, .fastq.
- No compression needed for .bam files.
- No benefits in compressing all files into one archive.

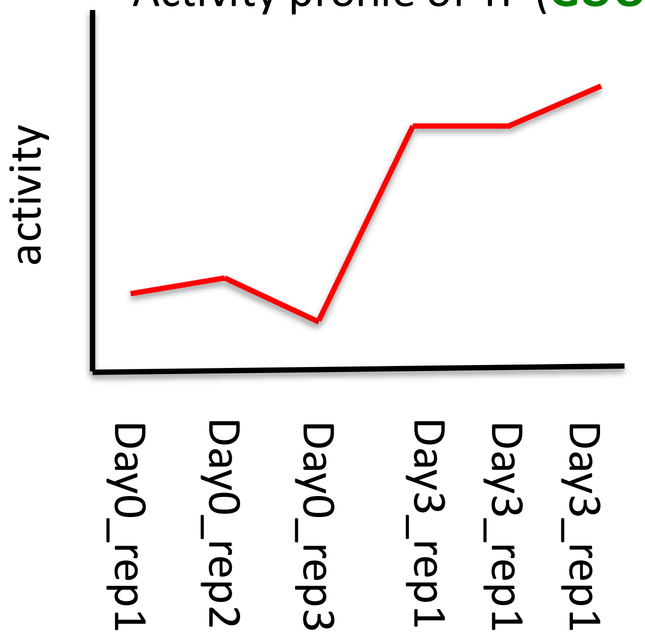
ISMARA supports .zip, .gz, .tar, .tar.gz, .bzip2 formats.

Name your files wisely!

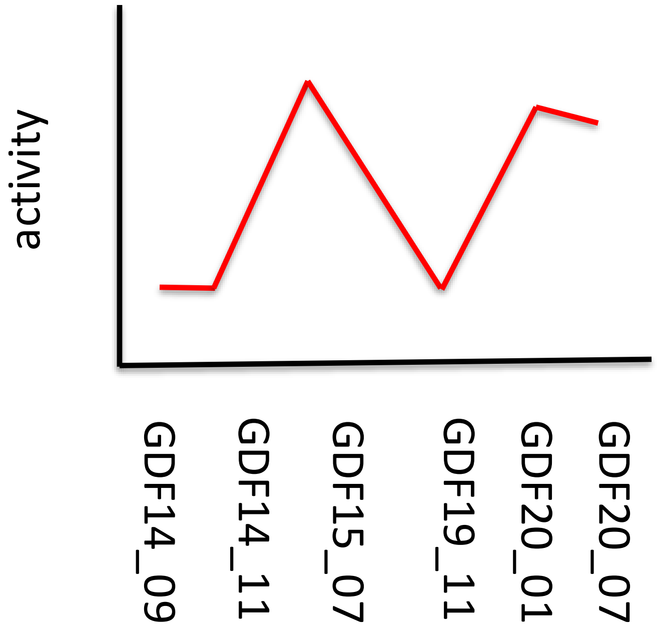
- Sample names should have intuitive meaning.
- Shorter is better (long names can get truncated).
- Samples are shown in alphabetical order.

Sample Order Difference

Activity profile of TF (**GOOD**)



Activity profile of TF (**BAD**)



File naming schemes

GOOD	BAD
Control-rep1.fastq.gz	SRR5134969. fastq.gz
Control-rep2.fastq.gz	SRR5134970. fastq.gz
Treatment1-rep1.fastq.gz	SRR5135011. fastq.gz
Treatment1-rep2.fastq.gz	SRR5135015. fastq.gz
Treatment2-rep1.fastq.gz	SRR5135016. fastq.gz
Treatment2-rep2.fastq.gz	SRR5135017. fastq.gz

Enforcing file order

You can enforce file order with numerical prefixes.
Note leading zeros in the file names.

01_sample1.bed
02_sample2.bed
…
14_sample14.bed
…
22_sample32.bed

with zeros

14_sample14.bed
…
1_sample1.bed
…
22_sample32.bed
…

without zeros

File naming for paired-end FASTQ files

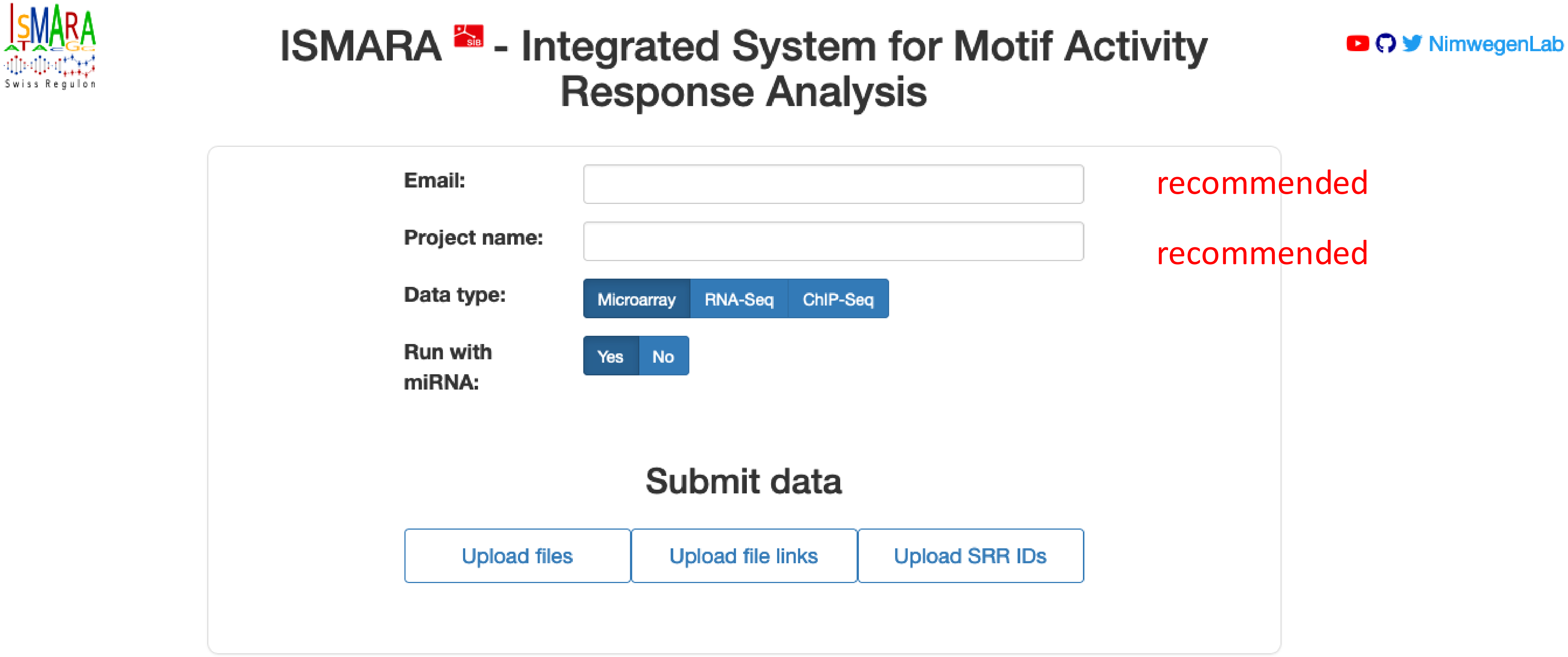
- Paired-end .fastq files require special suffix
- It should be **_R1** for one end and **_R2** for another end.
- The sample name of both files should be the same.

Example:

control-1_R1.fastq.gz

control-1_R2.fastq.gz

Submitting data



ISMARA - Integrated System for Motif Activity Response Analysis

Email: — recommended

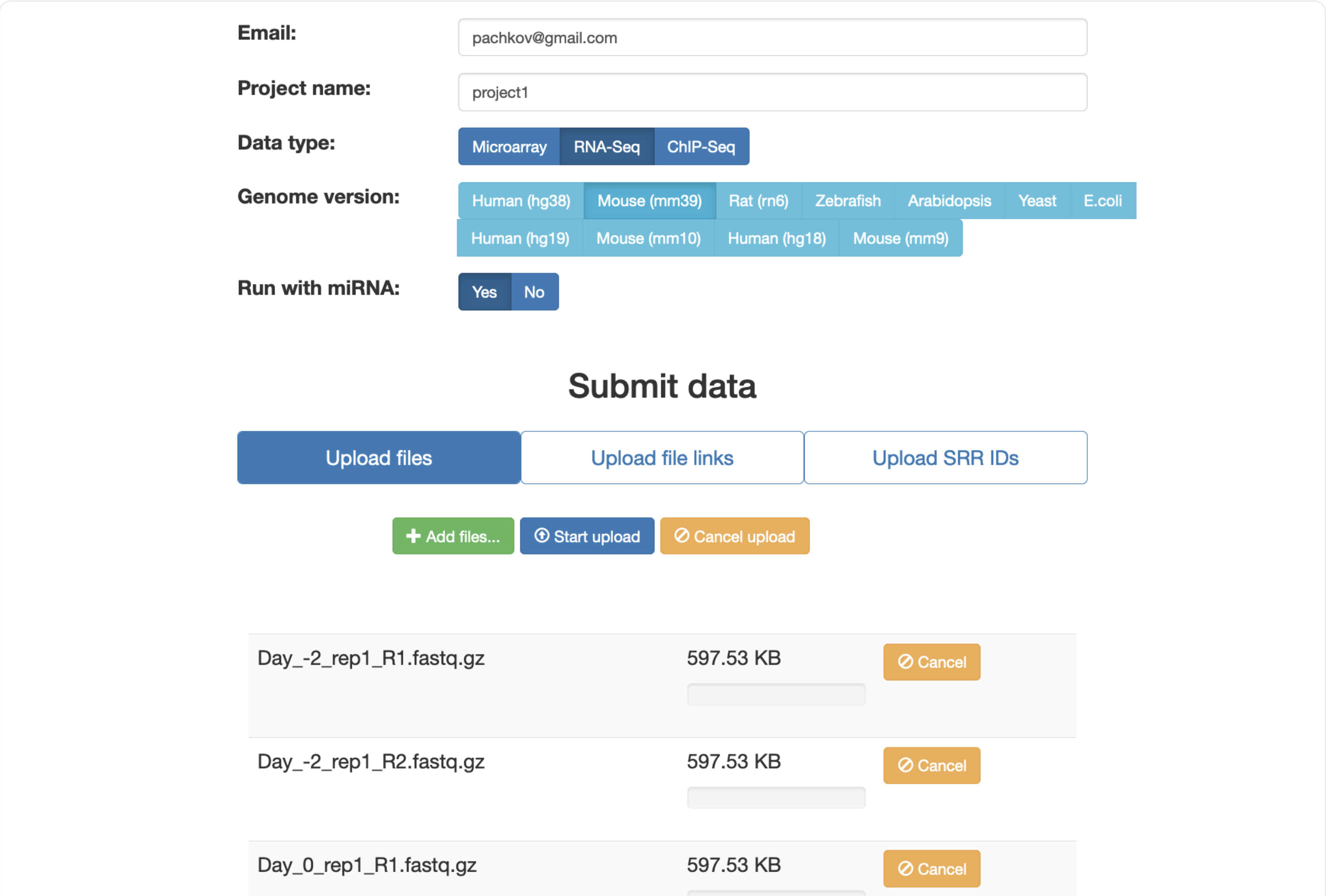
Project name: — recommended

Data type: Microarray | RNA-Seq | ChIP-Seq

Run with miRNA: Yes | No

Submit data

Uploading local files



Email: pachkov@gmail.com

Project name: project1

Data type: Microarray [✓ RNA-Seq] ChIP-Seq

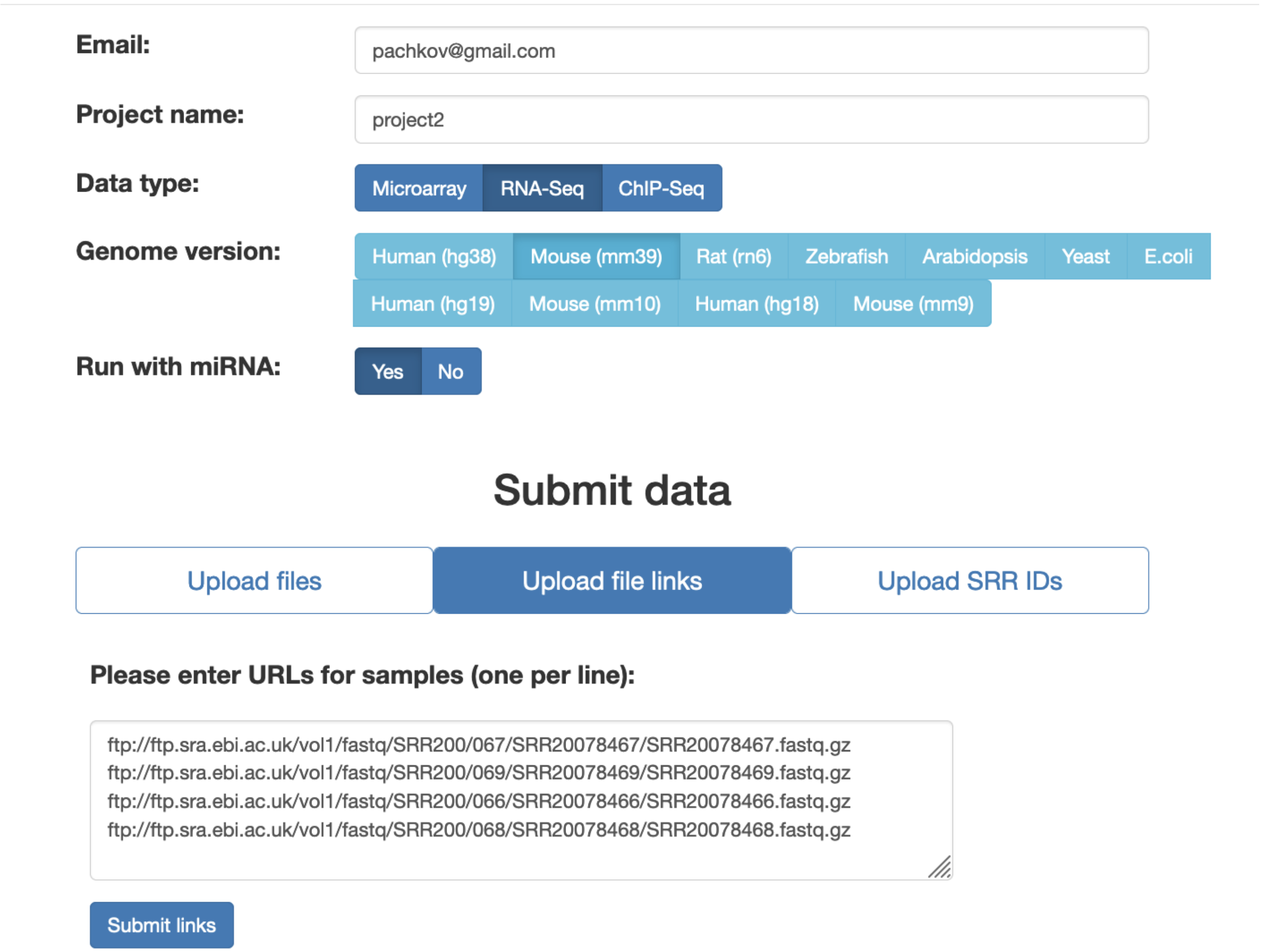
Genome version: Human (hg38) [✓ Mouse (mm39)] Rat (rn6) Zebrafish Arabidopsis Yeast E.coli Human (hg19) Mouse (mm10) Human (hg18) Mouse (mm9)

Run with miRNA: [✓ Yes] No

Submit data

Day_-2_rep1_R1.fastq.gz	597.53 KB	Cancel
Day_-2_rep1_R2.fastq.gz	597.53 KB	Cancel
Day_0_rep1_R1.fastq.gz	597.53 KB	Cancel

Submitting links to data files



Email: pachkov@gmail.com

Project name: project2

Data type: Microarray [✓ RNA-Seq] ChIP-Seq

Genome version: Human (hg38) [✓ Mouse (mm39)] Rat (rn6) Zebrafish Arabidopsis Yeast E.coli Human (hg19) Mouse (mm10) Human (hg18) Mouse (mm9)

Run with miRNA: [✓ Yes] No

Submit data

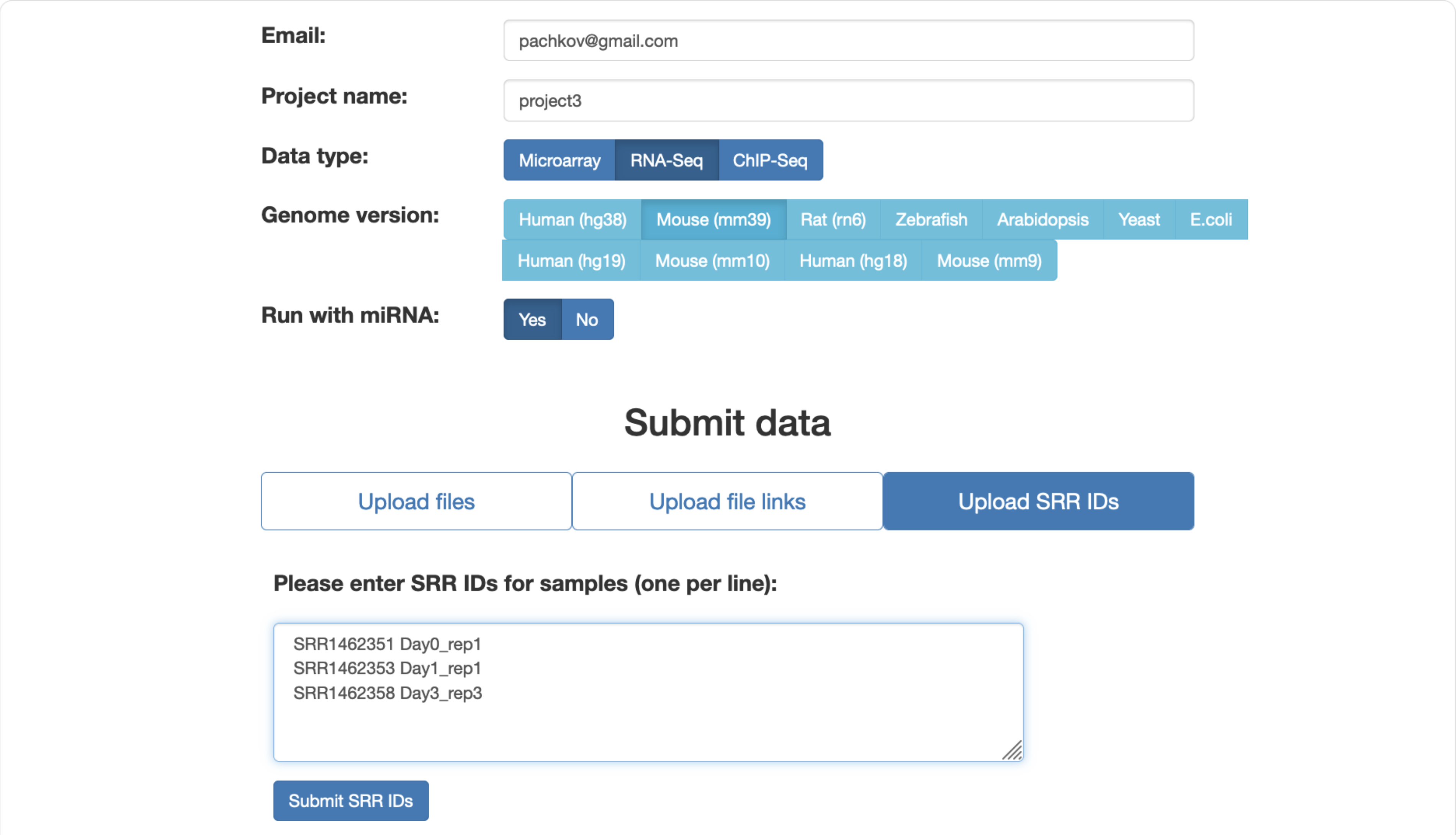
Upload files [✓ Upload file links] Upload SRR IDs

Please enter URLs for samples (one per line):

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/067/SRR20078467/SRR20078467.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/069/SRR20078469/SRR20078469.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/066/SRR20078466/SRR20078466.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR200/068/SRR20078468/SRR20078468.fastq.gz
```

Submitting SRR IDs

(Sequence Read Archive DB)



Email: pachkov@gmail.com

Project name: project3

Data type: Microarray [✓ RNA-Seq] ChIP-Seq

Genome version: Human (hg38) [✓ Mouse (mm39)] Rat (rn6) Zebrafish Arabidopsis Yeast E.coli Human (hg19) Mouse (mm10) Human (hg18) Mouse (mm9)

Run with miRNA: [✓ Yes] No

Submit data

Upload files | Upload file links | [✓ Upload SRR IDs]

Please enter SRR IDs for samples (one per line):

```
SRR1462351 Day0_rep1
SRR1462353 Day1_rep1
SRR1462358 Day3_rep3
```

Data upload

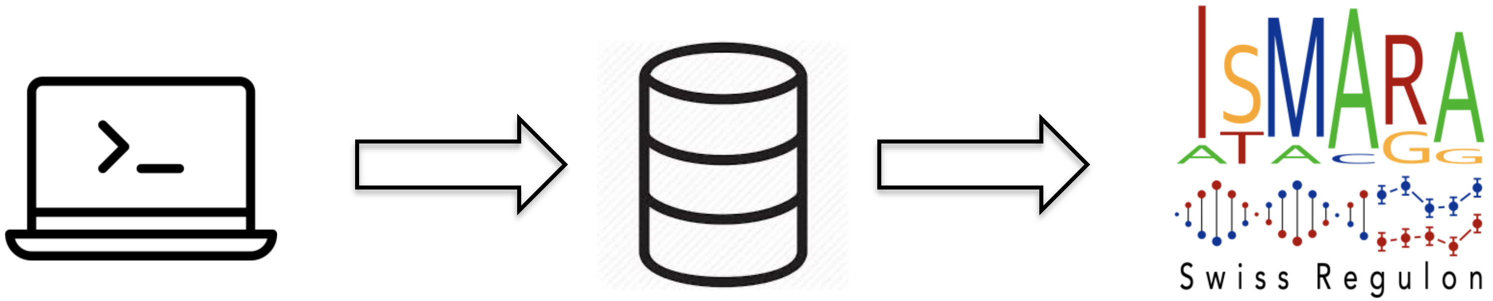
There are currently two possibilities to upload data to the ISMARA webserver:

- **Web interface** ismara.unibas.ch
  - Simple.
  - Requires local access to the data files.

- **ISMARA Uploader** https://github.com/ismara-unibas/upload-client
  - More robust for uploading the large datasets.
  - Requires basic knowledge of the command line.
  - Requires Python environment.

ISMARA uploader

- https://github.com/ismara-unibas/upload-client



User computer

Remote data server
running ISMARA uploader

ISMARA server

- python script which can upload data to the ISMARA server.
- provides all functionality of the web-interface.
- Running environment can be installed with conda package manager.

<u>Standard scenario:</u>

- You connect via terminal to a remote machine wich stores your data.
- Run uploader on the remote machine to upload data to the ISMARA server.

Prepare “file_list”

local files

- “file_list” is a text file which contains paths to files for upload.
- It should be one file path per line.

Example:

```
/path/Sample1.fastq.gz
/path/Sample2.fastq.gz
/path/Sample3.fastq.gz
/path/Sample4.fastq.gz
```

Prepare “file_list”

list of links

Instead of file paths you can use list of links.

Example:

```
https://example.com/data/sample1_R1.fastq.gz
https://example.com/data/sample1_R2.fastq.gz
https://example.com/data/sample2_R1.fastq.gz
https://example.com/data/sample2_R2.fastq.gz
```

Prepare “file_list”

list of SRR IDs

- You can also provide a list of SRR IDs.
- For every SRR you can give a sample name, to be shown in the results.

Example:

```
SRR12345 3hours_rep1
SRR12346 3hours_rep2
SRR12347 3hours_rep3
SRR12348 6hours_rep1
```

ISMARA uploader

Requirements: Python 3, “requests” library

Installation: just download the script

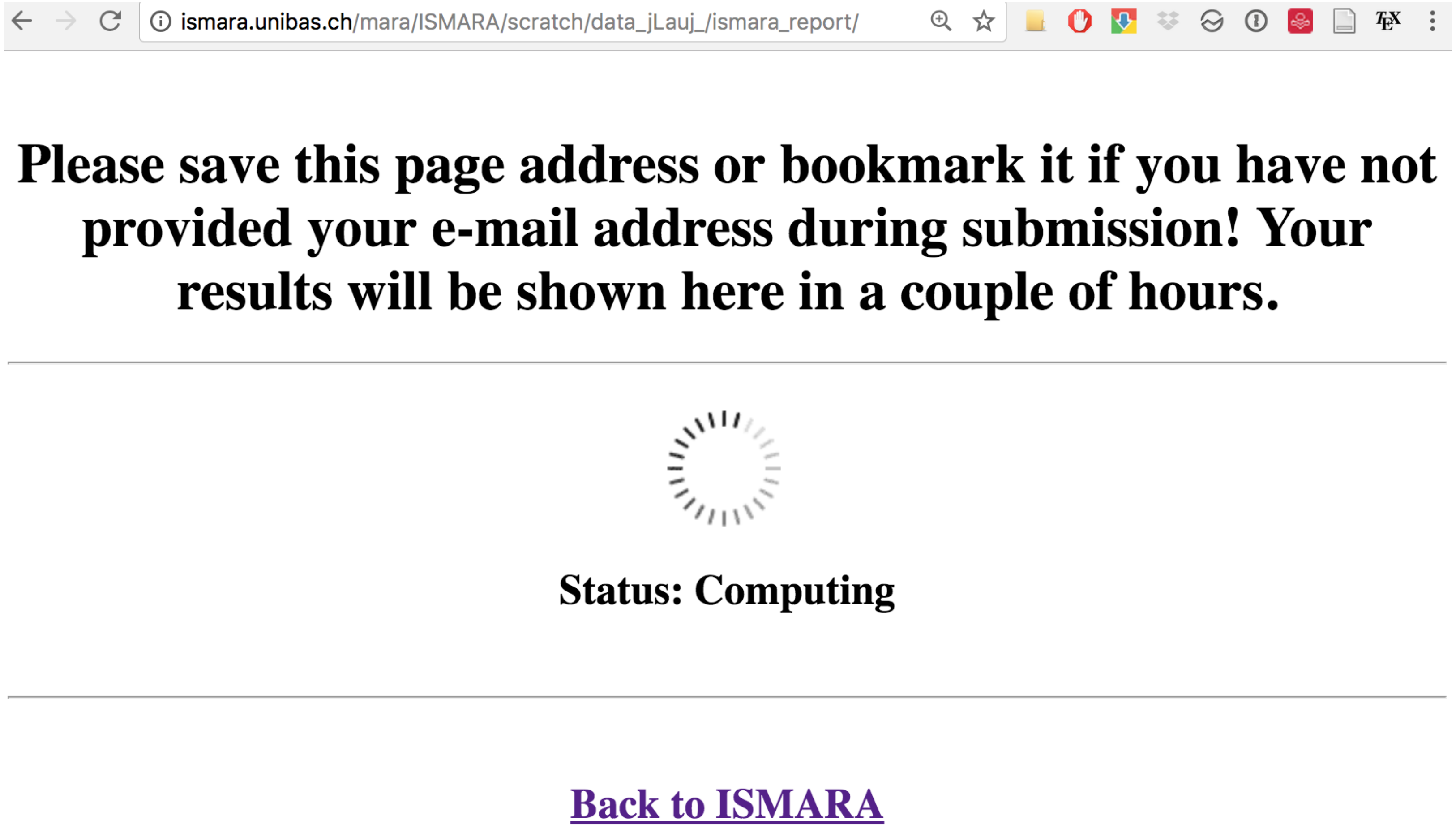
Usage:

```
nohup python ismara_uploader.py -e EMAIL \
         -p PROJECT \
         -t data-type {microarray,rnaseq,chipseq} \
         -o organism id or genome version {human,mouse,hg38,hg19…} \
         --mirna \
         --file-list [file-list] 1> results_link &
```

Output: file “results_link” contains url of the ISMARA results.

Check the GitHub page for documentation!

ISMARA status page



ismara.unibas.ch/mara/ISMARA/scratch/data_jLauj_/ismara_report/

Please save this page address or bookmark it if you have not provided your e-mail address during submission! Your results will be shown here in a couple of hours.

Status: Computing

Back to ISMARA

ISMARA status page

- Shows status of your job (errors if any)
- After ISMARA analysis is finished, results are available through the status page url
- Page automatically reloads, regularly updating its content
- Save this link if you have not provided your email in the submission form

ISMARA running time

ISMARA running time:

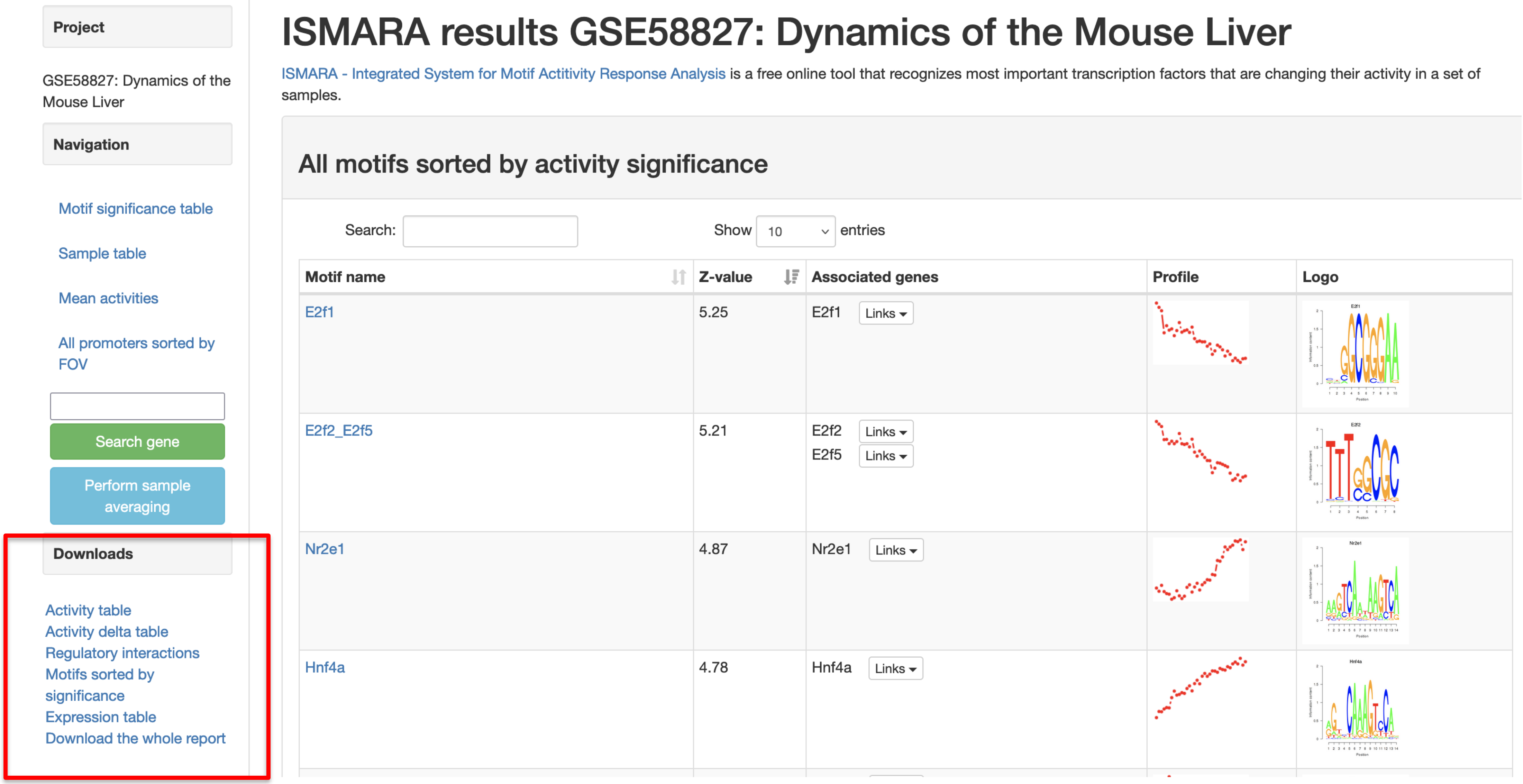
- one to a few hours.
- depends on a dataset size and computational resources availability.

If you do not get your results within 24 hours, this suggests that something is wrong. Please contact us reporting the status page url.

ISMARA storage

- Results are kept on the server for 6 months
- User input data is removed after analysis is complete
- Data available via unique URLs
- Extended security options are available (license required)

ISMARA downloads



Project

GSE58827: Dynamics of the Mouse Liver

Navigation

- Motif significance table
- Sample table
- Mean activities
- All promoters sorted by FOV

Downloads

- Activity table
- Activity delta table
- Regulatory interactions
- Motifs sorted by significance
- Expression table
- Download the whole report

ISMARA results GSE58827: Dynamics of the Mouse Liver

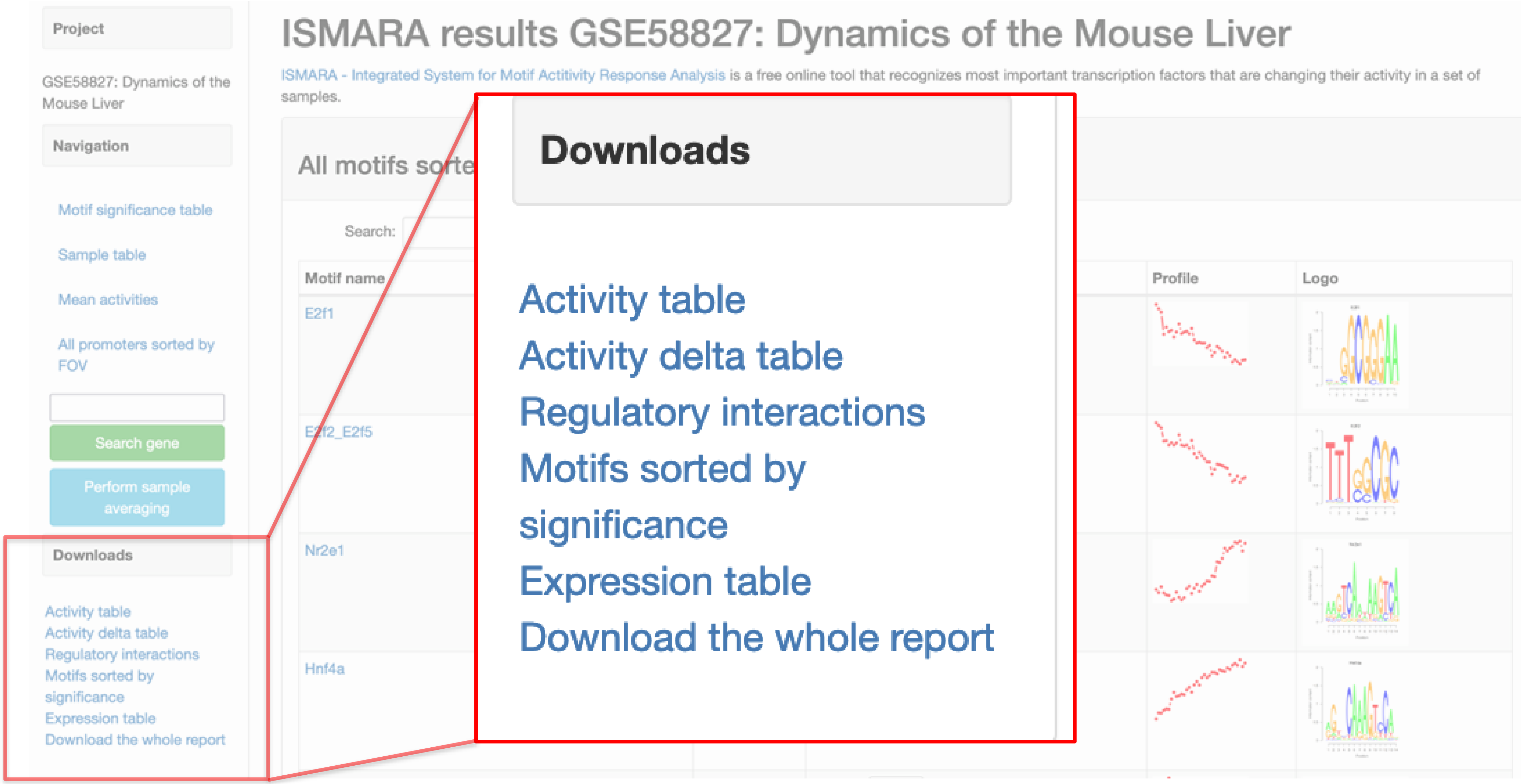
ISMARA - Integrated System for Motif Actitivity Response Analysis is a free online tool that recognizes most important transcription factors that are changing their activity in a set of samples.

All motifs sorted by activity significance

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
E2f1	5.25	E2f1 Links		
E2f2_E2f5	5.21	E2f2 Links E2f5 Links		
Nr2e1	4.87	Nr2e1 Links		
Hnf4a	4.78	Hnf4a Links		

ISMARA downloads



Project

GSE58827: Dynamics of the Mouse Liver

Navigation

Motif significance table

Sample table

Mean activities

All promoters sorted by FOV

Search gene

Perform sample averaging

Downloads

Activity table
Activity delta table
Regulatory interactions
Motifs sorted by significance
Expression table
Download the whole report

ISMARA results GSE58827: Dynamics of the Mouse Liver

ISMARA - Integrated System for Motif Activity Response Analysis is a free online tool that recognizes most important transcription factors that are changing their activity in a set of samples.

All motifs sorte

Search:

Motif name	Profile	Logo
E2f1		
E2f2_E2f5		
Nr2e1		
Hnf4a		

Downloads

Activity table
Activity delta table
Regulatory interactions
Motifs sorted by significance
Expression table
Download the whole report

ISMARA downloads

activity table

Activity table contains activities inferred by ISMARA

- ASCII text
- Tab-separated values

```
#sample   Motif1    Motif2   Motif3
Sample1    0.049     -0.019    -0.035
Sample2    0.046     -0.028    -0.039
Sample3    -0.054    -0.127    -0.009
```

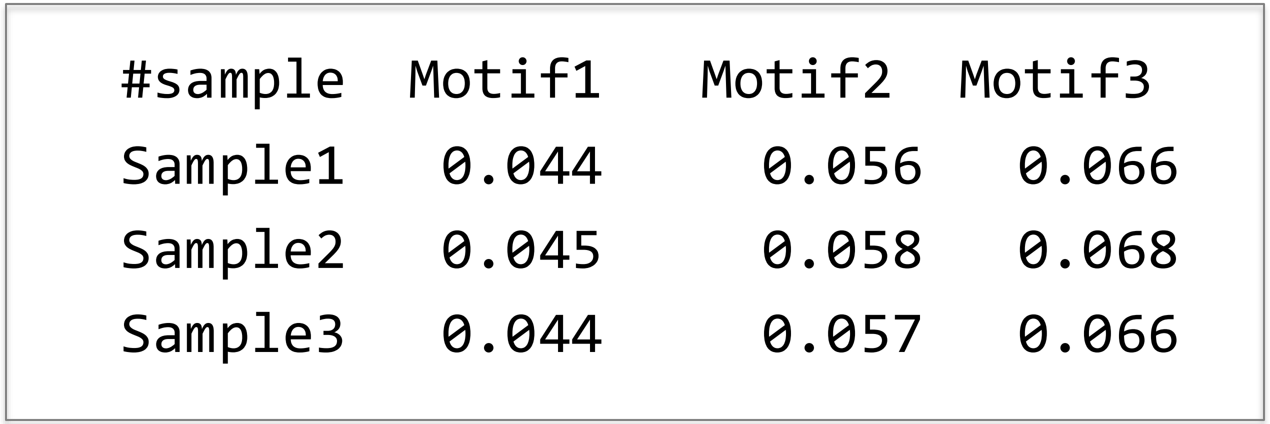
- Activity of motif *m* in sample *s* = predicted expression change in sample *s* resulting from adding one binding site for motif m.

ISMARA downloads

activity deltas table

Activity deltas table

- ASCII text
- Tab-separated values



```
#sample   Motif1    Motif2   Motif3
Sample1    0.044      0.056    0.066
Sample2    0.045      0.058    0.068
Sample3    0.044      0.057    0.066
```

- Delta of motif *m* in sample *s* = error-bar on activity of motif *m* in sample *s*.

ISMARA downloads

regulatory interactions

Regulatory interactions files are available as TAR archive with 1 file for each motif

- Interactions are sorted by log-likelihood score.
- Fields: promoter, log-likelihood score, regulator, promoter annotation.
- Tab-separated values.

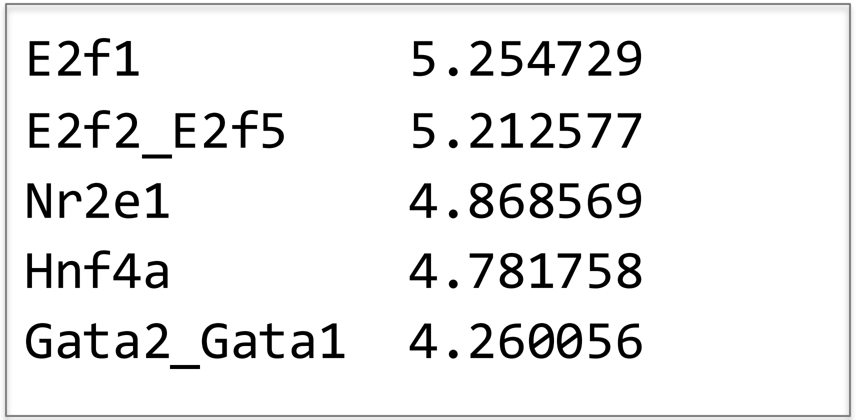
```
Promoter      mm10_v2_chr19_+_39287074_39287104
LL score      95.7766
Motif         Hnf4a
Transcripts   ENSMUST00000003137.8|Cyp2c29|ENSMUSG00000003053.11|cytochrome P450,
family 2, subfamily c, polypeptide 29
```

ISMARA downloads

motif significances

Motif significances table contains list of motifs and their Z-scores

- Motifs are sorted by Z-score.
- Values are tab-separated.



```
E2f1          5.254729
E2f2_E2f5     5.212577
Nr2e1         4.868569
Hnf4a         4.781758
Gata2_Gata1   4.260056
```

- Motif significance = $z_m = \sqrt{\frac{1}{S}\sum_{s}\left(\frac{A'_{ms}}{\delta A'_{ms}}\right)^2}$

ISMARA downloads

expression table

Expression table contains promoter expression values.

- ASCII text.
- Tab-separated values.
- $\log_2$ (transcripts per million transcripts).

```
#promoter   Sampe1          Sample2         Sample3
prom1       4.21900323481   3.87669279321   4.02108886991
prom2       1.51146874145   0.73990012059   0.95424591736
prom3       4.97351148778   4.50373729065   4.86135208071
```

ISMARA downloads

full report

The report archive contains:

- all html report pages for off-line browsing
- and all downloadable files

Features missing in report archive:

- gene search function
- promoters sorted by FOV page
- averaging functionality

CREMA: supported species

	Human	Mouse	Rat	Zebrafish
Genome	h19	mm10	rn6	dr11
Motifs	499	503	503	475
TFs	682	679	650	832

CREMA: supported data types

Next Generation Sequencing

Required data:

- Raw reads in FASTQ format.
- Sample description file in TSV format.

Supported data types:

- ATAC-Seq and DNase-Seq DNA accessibility data.
- ChIP-Seq histone modification data (H3K4me1, H3K4me3, *etc.*).

Sample annotation

samples.tsv

For proper processing of the data we need description of the files in your dataset.

Description provided in a .TSV file of the following form:

```
 sample    type   fq1     fq2
Cond1      fg     /a/a.fastq.gz
Cond1      fg     /a/b.fastq.gz
Cond1      bg     /a/c.fastq.gz
Cond2      fg     /a/d_1.fastq.gz    /a/d_2.fastq.gz
Cond2      bg     /a/e_1.fastq.gz    /a/e_2.fastq.gz
```

It contains condition name, files associated to a condition and type of the sample (fg/bg).

Sample annotation

samples.tsv

It is allowed

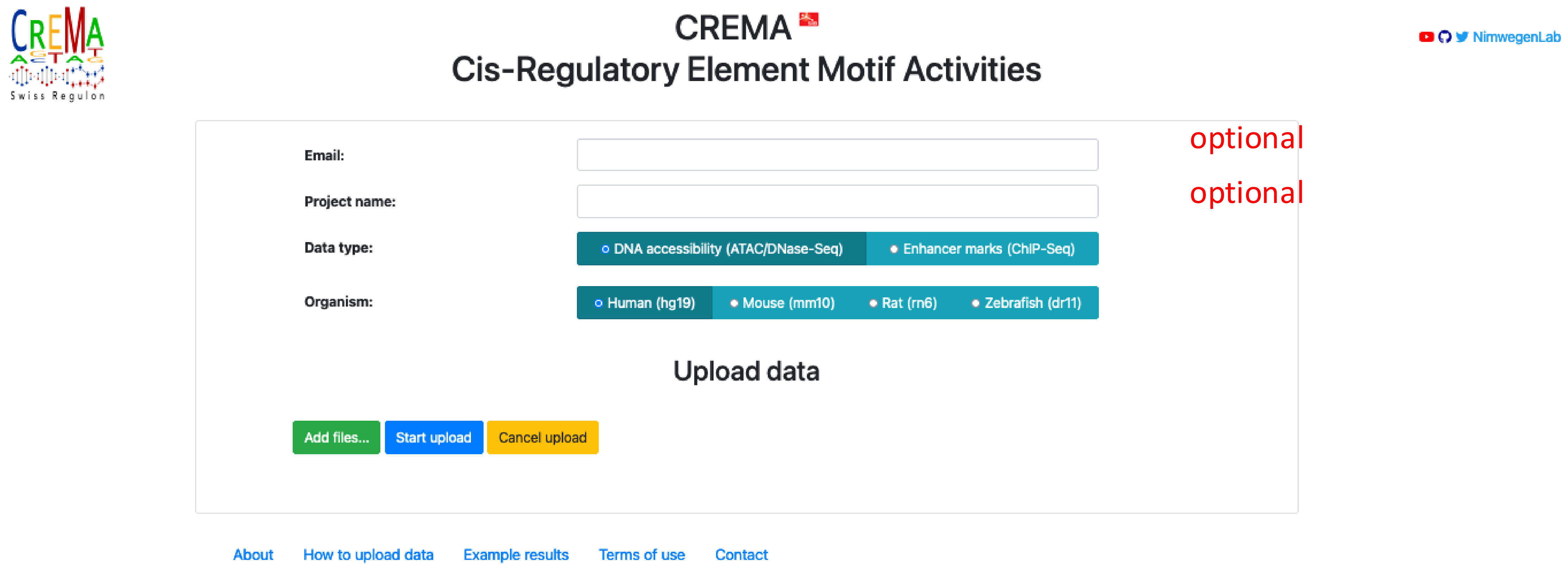
- multiple files per sample
- mix single-end and paired-end data

```
sample    type   fq1      fq2
Cond1     fg     /a/a.fastq.gz
Cond1     fg     /a/b.fastq.gz
Cond1     bg     /a/c.fastq.gz
Cond2     fg     /a/d_1.fastq.gz    /a/d_2.fastq.gz
Cond2     bg     /a/e_1.fastq.gz    /a/e_2.fastq.gz
```

Naming rules

- Sample names should be comprehensive.
- Sample names should not be long.
- Order of sample names in the plots is defined by order of sample names in `samples.tsv` file.
- FASTQ filenames have no effect on sample names shown in the report.
- There are no requirements for FASTQ filenames of paired-end reads.

CREMA web interface



CREMA

Cis-Regulatory Element Motif Activities

NimwegenLab

Email: optional

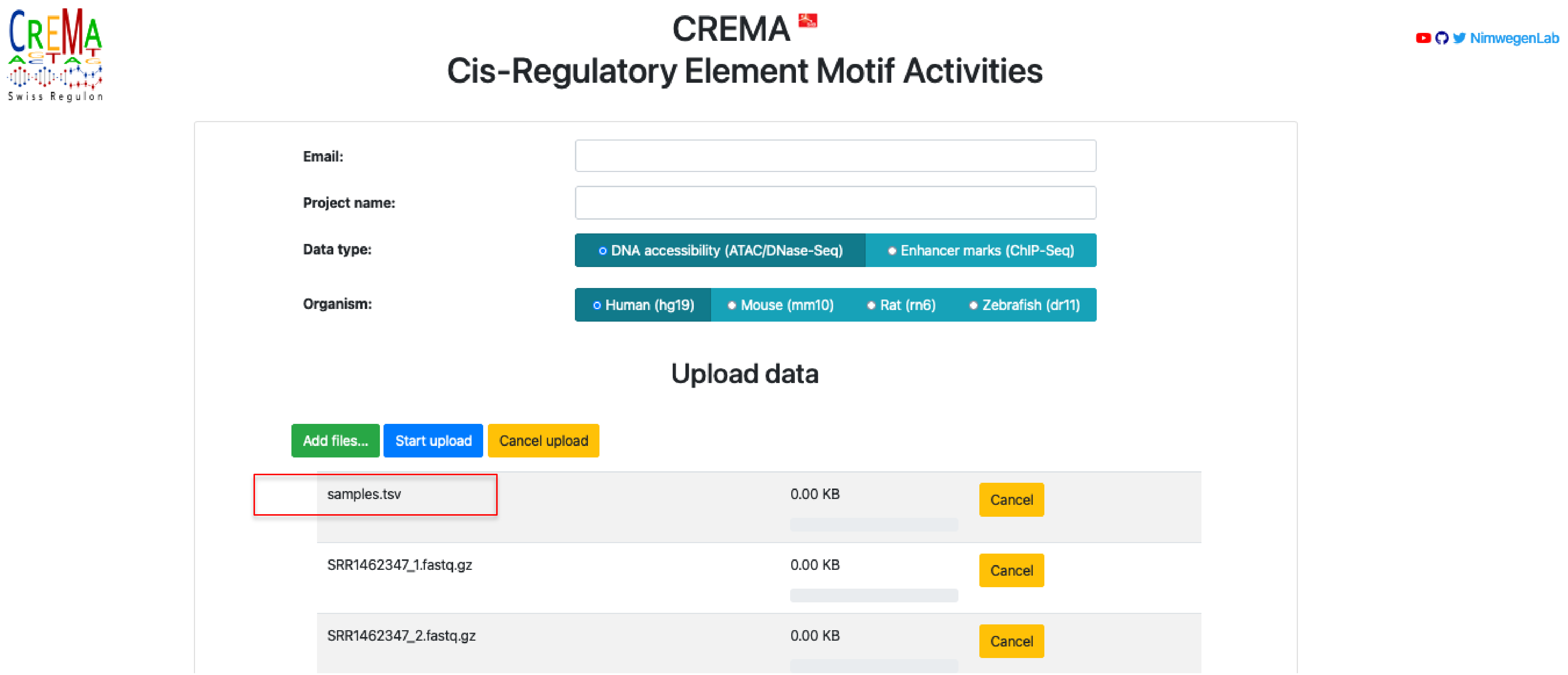
Project name: optional

Data type: ○ DNA accessibility (ATAC/DNase-Seq) ● Enhancer marks (ChIP-Seq)

Organism: ○ Human (hg19) ● Mouse (mm10) ● Rat (rn6) ● Zebrafish (dr11)

Upload data

CREMA web interface



CREMA
Cis-Regulatory Element Motif Activities

Email:

Project name:

Data type: ○ DNA accessibility (ATAC/DNase-Seq) ● Enhancer marks (ChIP-Seq)

Organism: ○ Human (hg19) ● Mouse (mm10) ● Rat (rn6) ● Zebrafish (dr11)

Upload data

samples.tsv 0.00 KB

SRR1462347_1.fastq.gz 0.00 KB

SRR1462347_2.fastq.gz 0.00 KB

Uploading a links or SRR IDs

You can add URLs or SRR IDs to the samples.tsv file. The corresponding FASTQ files will be downloaded automatically and added to the dataset.

There could be multiple URL/SRR per condition.

```
sample      type  fq1    fq2
condition1    fg  /data/file1.fastq.gz
condition1    bg  /data/file2.fastq.gz
condition2    fg SRR12345
condition2    fg SRR12346
condition2    bg  SRR12347
condition3    fg  https://example.com/1_1.fastq.gz https://example.com/1_2.fastq.gz
condition3    bg  https://example.com/2_1.fastq.gz https://example.com/2_2.fastq.gz
```

Dataset upload

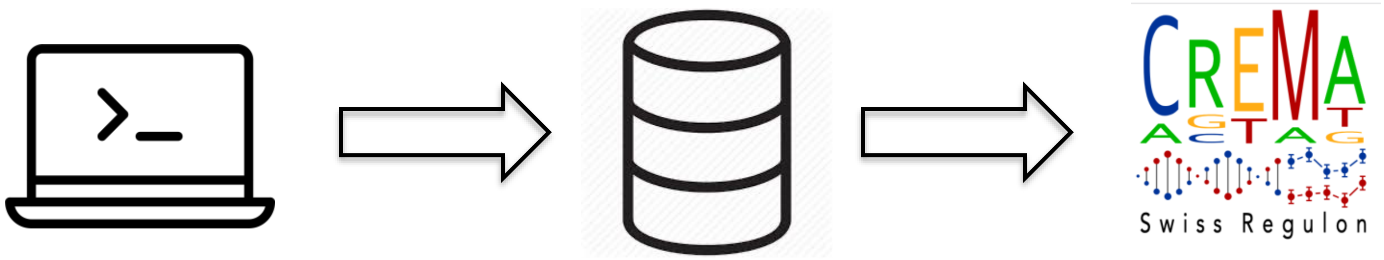
There are currently two possibilities to upload data to the CREMA webserever:

- **Web interface** crema.unibas.ch
  - Simple.
  - Requires local access to the data files.

- **CREMA Uploader** github.com/ismara-unibas/crema_uploader
  - More robust for uploading the large datasets.
  - Require basic knowledge of the command line.
  - Requires Python environment.

CREMA uploader

- https://github.com/ismara-unibas/crema_uploader



User computer

Remote data server
running CREMA uploader

CREMA server

- python script which can upload data to the CREMA server.
- provides all functionality of the web-interface.
- Running environment can be installed with conda package manager.

<u>Standard scenario:</u>

- You connect via terminal to a remote machine wich stores your data.
- Run uploader on the remote machine to upload data to the CREMA server.

CREMA uploader

Requirements: Python 3, “requests” library

Installation: just download the script

Usage:

```
nohup python crema_uploader.py -e EMAIL \
         -p PROJECT \
         --data-type {chip-seq, atac-seq} \
         -o genome version {hg19,mm10, rn6, dr11} \
         --file-list TSV_FILE 1> results_link &
```

Output: file “results_link” contains url of the CREMA results.

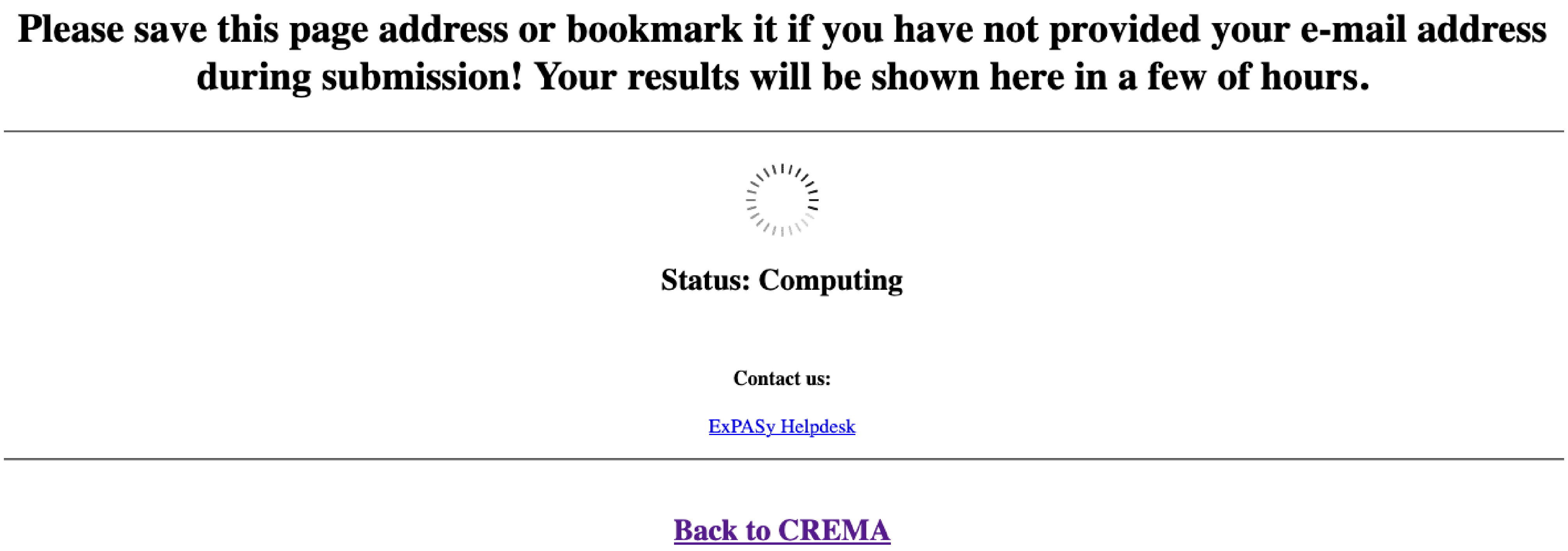
Check the GitHub page for documentation!

CREMA uploader

Like the web interface, CREMA uploader supports TSV files containing local paths, URLs and SRR IDs.

```
sample     type   fq1    fq2
condition1    fg  /data/file1.fastq.gz
condition1    bg  /data/file2.fastq.gz
condition2    fg SRR12345
condition2    fg SRR12346
condition2    bg  SRR12347
condition3    fg  https://example.com/1_1.fastq.gz https://example.com/1_2.fastq.gz
condition3    bg  https://example.com/2_1.fastq.gz https://example.com/2_2.fastq.gz
```

CREMA status page



Please save this page address or bookmark it if you have not provided your e-mail address during submission! Your results will be shown here in a few of hours.

Status: Computing

Contact us:

ExPASy Helpdesk

Back to CREMA

CREMA status page

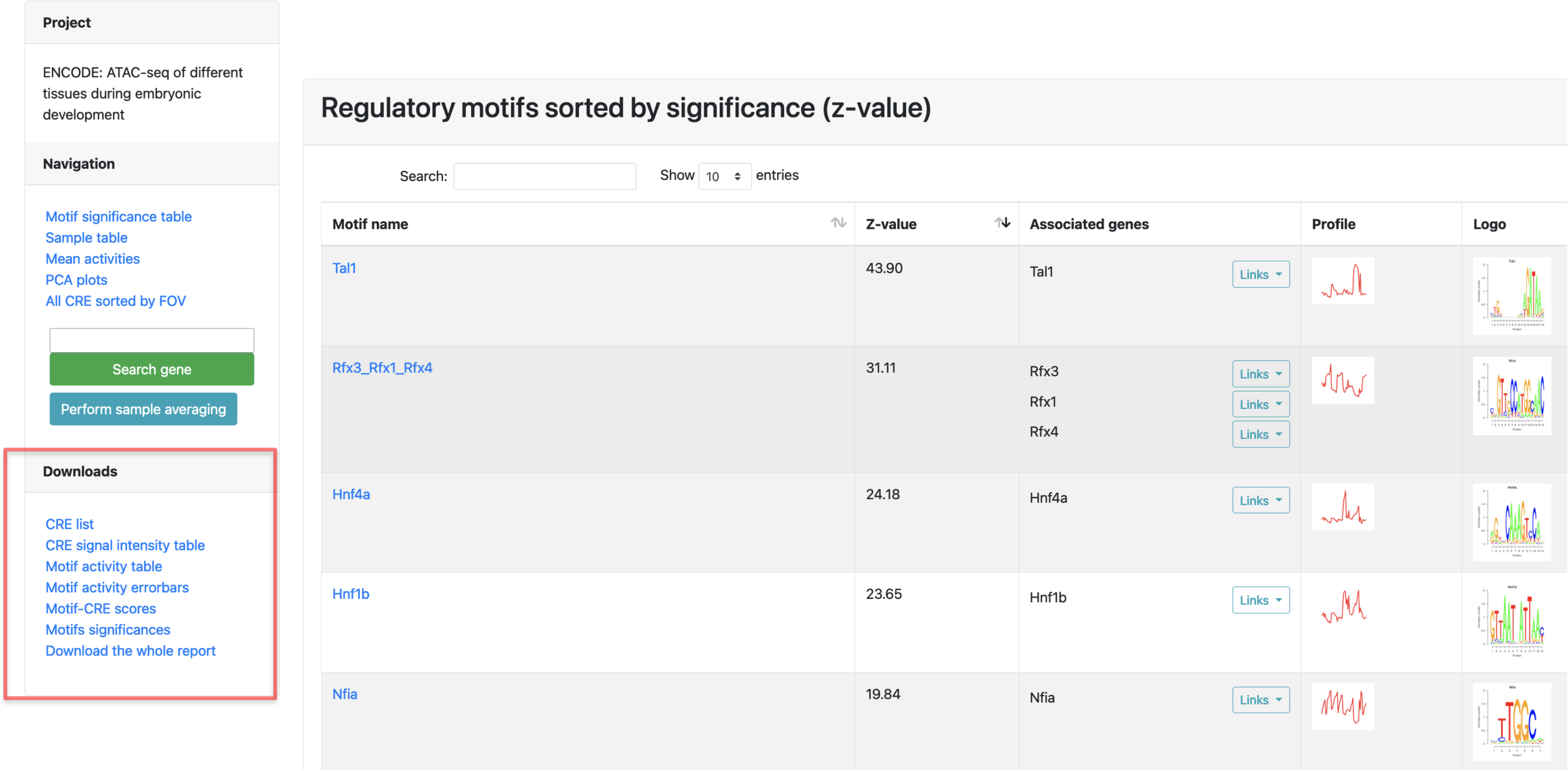
- Shows status of your job (errors if any)
- After CREMA analysis is finished, results are available through the status page url
- Page automatically reloads, regularly updating its content
- Save this link if you have not provided your email in the submission form

CREMA running time

ISMARA running time ranges from a few hours to a few days depending on the size of a dataset and availability of computational resources.

If you do not get your results within 48 hours that might indicate that something is wrong. Please contact us reporting the status page url.

CREMA downloads



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Downloads

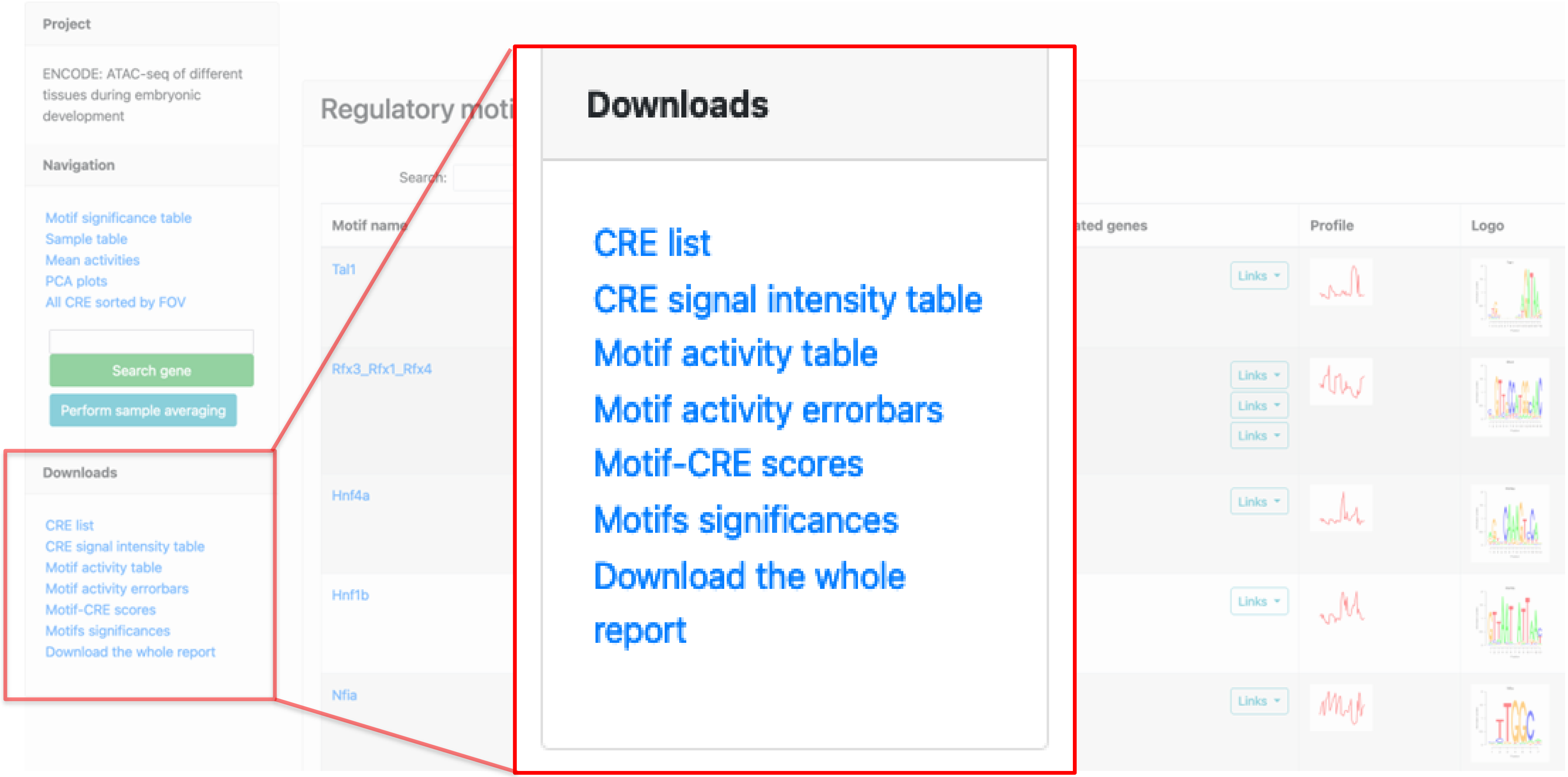
- CRE list
- CRE signal intensity table
- Motif activity table
- Motif activity errorbars
- Motif-CRE scores
- Motifs significances
- Download the whole report

Regulatory motifs sorted by significance (z-value)

Search: Show 10 entries

Motif name	Z-value	Associated genes		Profile	Logo
Tal1	43.90	Tal1	Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Rfx1 Rfx4	Links Links Links		
Hnf4a	24.18	Hnf4a	Links		
Hnf1b	23.65	Hnf1b	Links		
Nfia	19.84	Nfia	Links		

CREMA downloads



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Downloads

- CRE list
- CRE signal intensity table
- Motif activity table
- Motif activity errorbars
- Motif-CRE scores
- Motifs significances
- Download the whole report

Downloads

- CRE list
- CRE signal intensity table
- Motif activity table
- Motif activity errorbars
- Motif-CRE scores
- Motifs significances
- Download the whole report

CREMA downloads

CRE list

- ASCII text
- Tab-separated values
- Columns:

chromosome:	chr10
CRE start:	103367406
CRE end:	103367811
CRE length:	405
CRE ID:	mm10_chr10_103367406_103367811
Transcript with closest TSS:	ENSMUST00000218844

Transcript information:

ENSMUST00000218844|Slc6a15|ENSMUSG00000019894|solute carrier family, member 15|175|0.963813707301511 5

distance to CRE ↗ ↖ association probability

CREMA downloads

CRE signal intensity table

CRE signal intensity table contains *log*(normalized read counts)

- ASCII text
- Tab-separated values

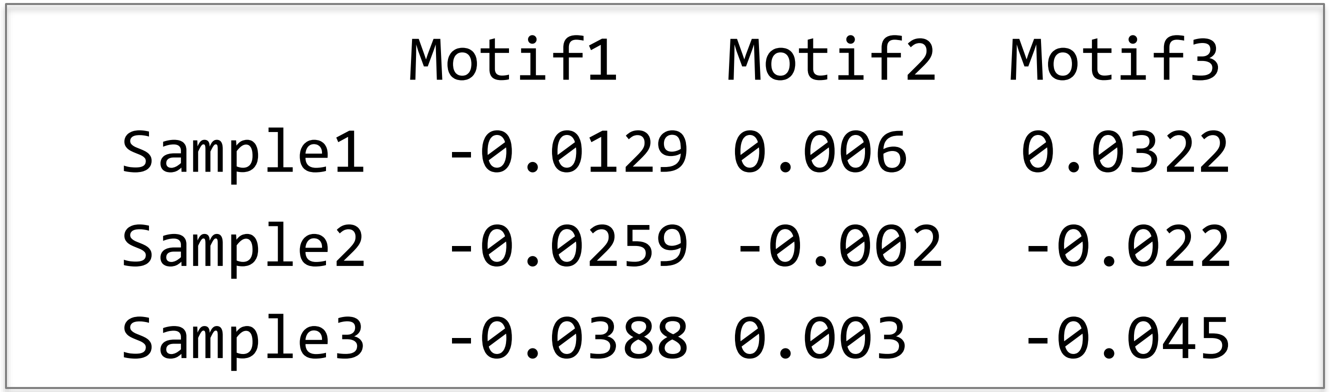
	Sampe1	Sample2	Sample3
CRE1	2.515	3.027	3.229
CRE2	2.092	2.936	2.312
CRE3	1.661	2.096	2.783

CREMA downloads

activity table

Activity table contains activities inferred by CREMA

- ASCII text
- Tab-separated values



```
         Motif1   Motif2  Motif3
Sample1  -0.0129  0.006    0.0322
Sample2  -0.0259  -0.002  -0.022
Sample3  -0.0388  0.003   -0.045
```

- Activity of motif *m* in sample *s* = predicted expression change in sample *s* resulting from adding one binding site for motif

CREMA downloads activity errorbars table

Activity errorbars table contains error bars inferred inferred by CREMA

- ASCII text
- Tab-separated values

```
         Motif1    Motif2   Motif3
Sample1   0.003     0.007   0.015
Sample2   0.004     0.008   0.016
Sample3   0.004     0.008   0.016
```

- Error-bar on activity of motif *m* in sample *s*.

CREMA downloads

Motif-CRE scores

Motif-CRE score files are available as TAR archive with 1 file for each motif

- Interactions are sorted by log-likelihood score
- Fields: promoter, log-likelihood score, regulator, promoter annotation
- Tab-separated values

```
CRE            mm10_chr16_87268014_87268522
LL score       6.12251
Motif          Hsf2
Transcript     ENSMUST00000054442|N6amt1|ENSMUSG00000044442|N-6 adenine-specific
DNA methyltransferase 1 (putative)|85917|0.08330647427020685
```

CREMA downloads

motif significances

Motif significances contains list of motifs, their significances and Z-values across all conditions.

- Motifs are sorted by Z-score
- Values are tab-separated

	significances	Sample1	Sample2	Sample3
Tal1	43.896	-25.672	-27.113	-24.202
Rfx3_Rfx1_Rfx4	31.110	10.006	-4.054	-10.816
Hnf4a	24.182	-11.727	-7.589	-1.126

- Motif significance = $z_m = \sqrt{\frac{1}{S}\sum_s \left(\frac{A'_{ms}}{\delta A'_{ms}}\right)^2}$

CREMA downloads

report

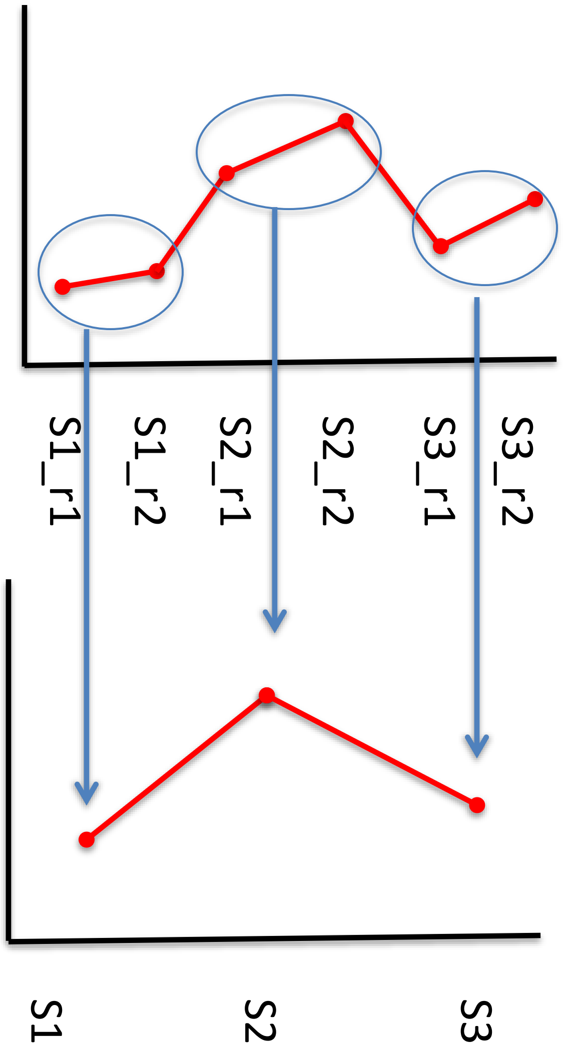
The report archive

- Contains compressed CREMA report directory for off-line browsing
- Contains html pages which are available on-line
- Includes activity, activity errorbars, regulatory interactions files, CRE signal table

Features missing in local report copy

- Gene search function
- Promoters sorted by FOV page
- Averaging functionality

Averaging activities

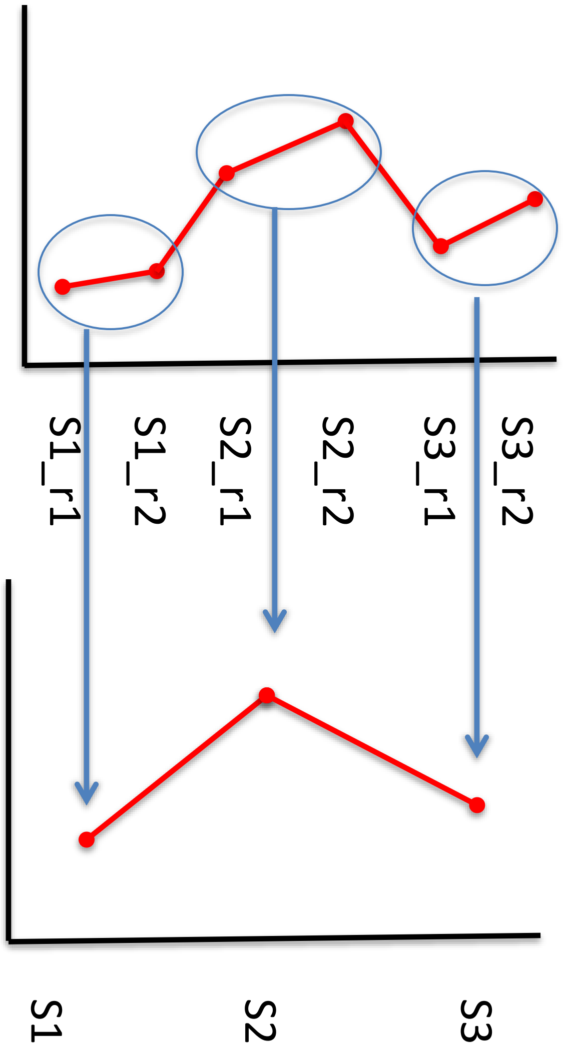


- Divides samples into different groups.
- Calculate average activity and corresponding errorbar per group.
- Calculate significances of motifs across groups.
- Identifies regulators with little variation within a group but significant variation across the groups.

Examples:

- replicate averaging.
- tissue-type averaging.
- age averaging.

Replicate averaging



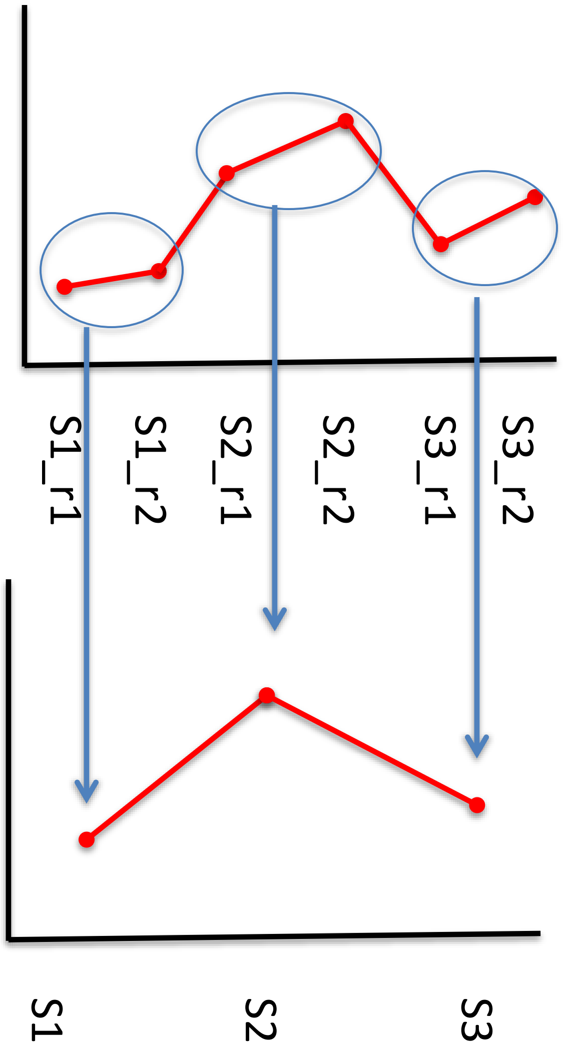
We assume that the activities across group *g* are normally distributed around some unknown mean $\bar{A}_{mg}$ with unknown variance σ^2_{mg}

$$P(A_{ms}|\bar{A}_{mg},\sigma_{mg}) = \frac{1}{\sqrt{2\pi}\,\sigma_{mg}}\exp\left[-\frac{1}{2}\frac{\left(A_{ms}-\bar{A}_{mg}\right)^2}{\sigma^2_{mg}}\right]$$

Then the probability of the data given $\bar{A}_{mg}$ and σ^2_{mg}
Is the following:

$$P(D|\,\bar{A}_{mg},\sigma_{mg}) = \prod_{s\in g}\frac{1}{\sqrt{2\pi(\sigma^2_{mg}+\sigma^2_{ms})}}\exp\left[-\frac{(A^*_{ms}-\bar{A}_{mg})^2}{2(\sigma^2_{mg}+\sigma^2_{ms})}\right]$$

Replicate averaging



$$P(D|\bar{A}_{mg},\sigma_{mg}) = \prod_{s\in g} \frac{1}{\sqrt{2\pi(\sigma_{mg}^2+\sigma_{ms}^2)}} \exp\left[-\frac{(A^*_{ms}-\bar{A}_{mg})^2}{2(\sigma_{mg}^2+\sigma_{ms}^2)}\right]$$

We numerically find the value of $\sigma^2{}_{mg}$ which maximizes the expression above. Assuming an uniform prior over mean activity $\bar{A}_{mg}$ we find that $P(A_{mg}/D)$ is a gaussian with mean

$$\bar{A}^*_{mg} = \frac{\sum_{s\in g} \frac{A^*_{ms}}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}{\sum_{s\in g} \frac{1}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}$$

and error

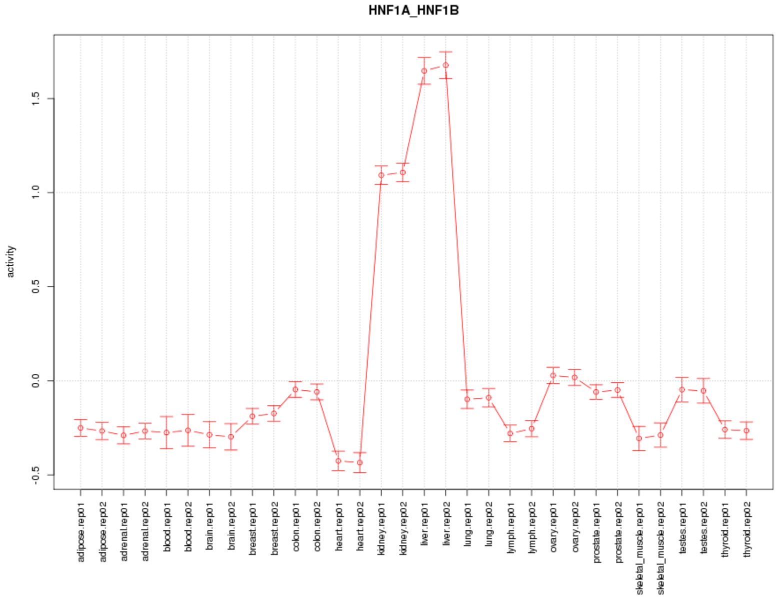
$$\bar{\sigma}^*_{mg} = \sqrt{\frac{1}{\sum_{s\in g} \frac{1}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}}$$

Replicate averaging

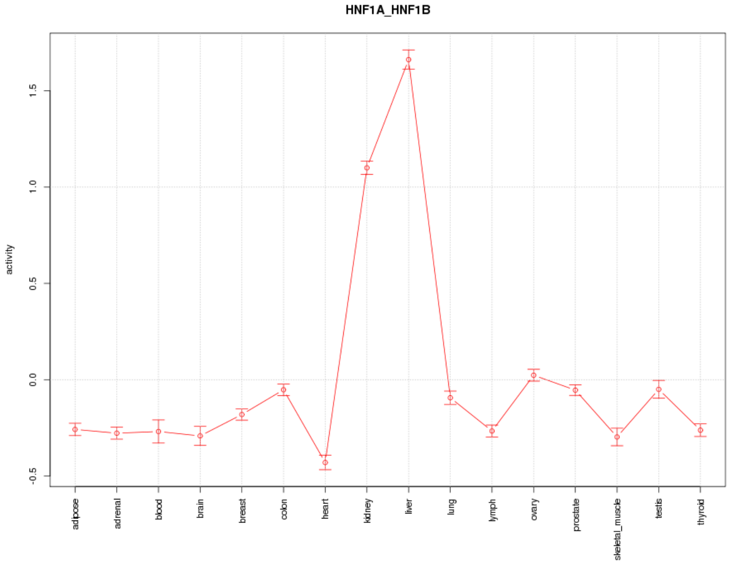
Illumina Body Map 2 averaged over replicates (GSE30611)

HNF1A_HNF1B activity profiles for

Original results



Averaged results



Averaging live example

Gene expression profiling of epithelial and mesenchymal subpopulations within immortalized human mammary epithelial cells (GSE28681, Scheel et al. Cell 2011)

Microarray experiment

Samples:

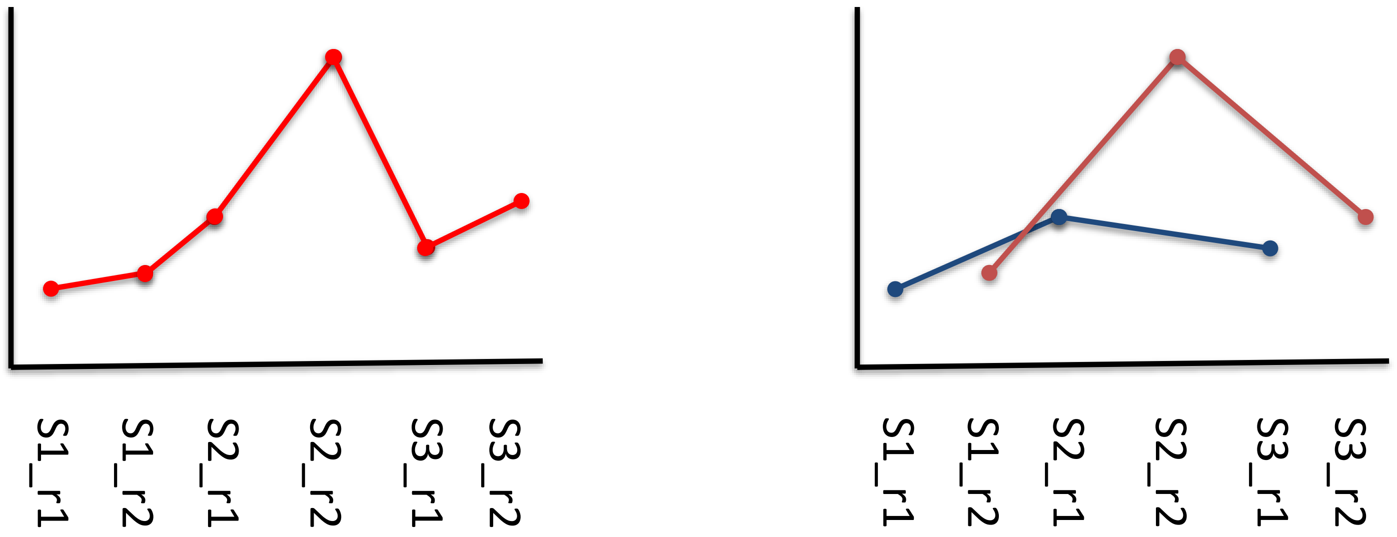
- epithelial cells (HMLE); 2 replicates
- 3 subpopulations of mesenchymal cells (HMLE); 2 replicates

Let’s see what is the difference between epithelial cells and mesenchymal cells subsets.

Effects of replicate averaging

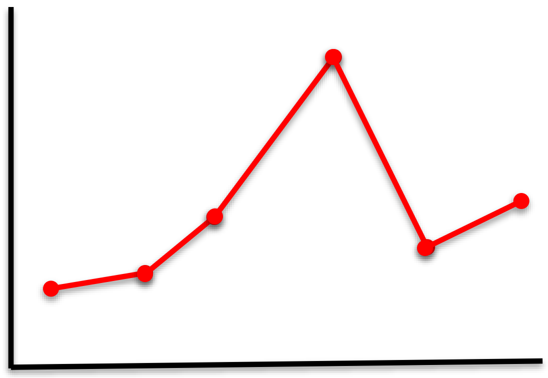
- Recalculated: Activities, error bars, z-values are recalculated and corresponding tables and plots
- Remain unchanged: Target list, regulatory network, activity/expression correlation plot, StringDB image, gene enrichment tables

Batch effect correction



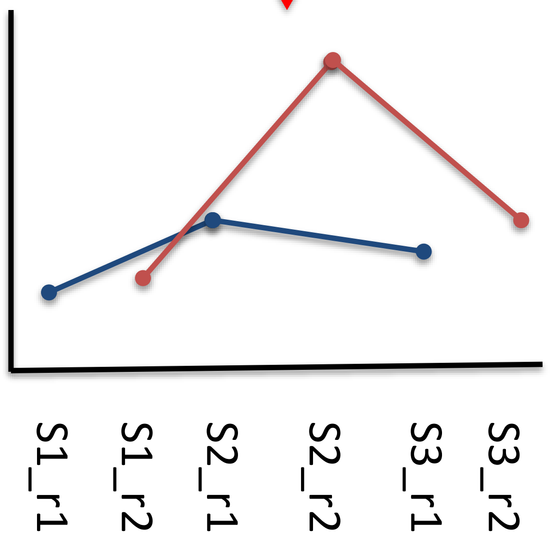
There may be systematic differences between each batch of measurements that are not of interest. To remove such batch effects, ISMARA will standardize the activities of each batch by normalizing the average and variance of the activities across all samples in a batch.

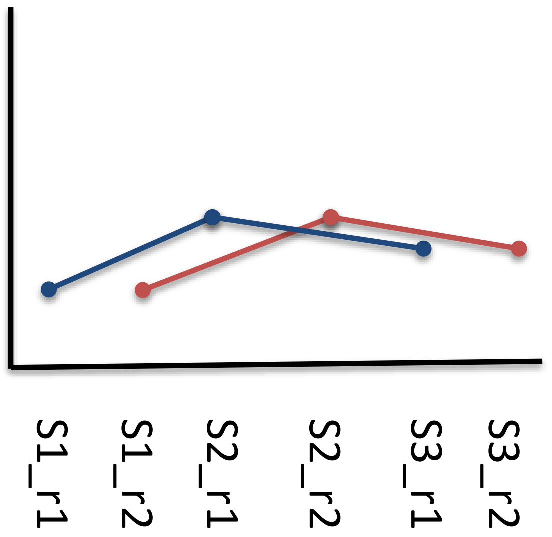
Batch effect correction



Standardization procedure

$$A^{*}_{mb} = \frac{A_{mb} - \bar{A}_{mb}}{s_{mb}} \qquad dA^{*}_{mb} = \frac{dA_{mb}}{s_{mb}}$$





Averaging with batch effect correction live example

Gene expression profiling of epithelial and mesenchymal subpopulations within immortalized human mammary epithelial cells (GSE28681, Scheel et al. Cell 2011)

Microarray experiment

Samples:

- epithelial cells (HMLE); 2 replicates
- 3 subpopulations of mesenchymal cells (HMLE); 2 replicates

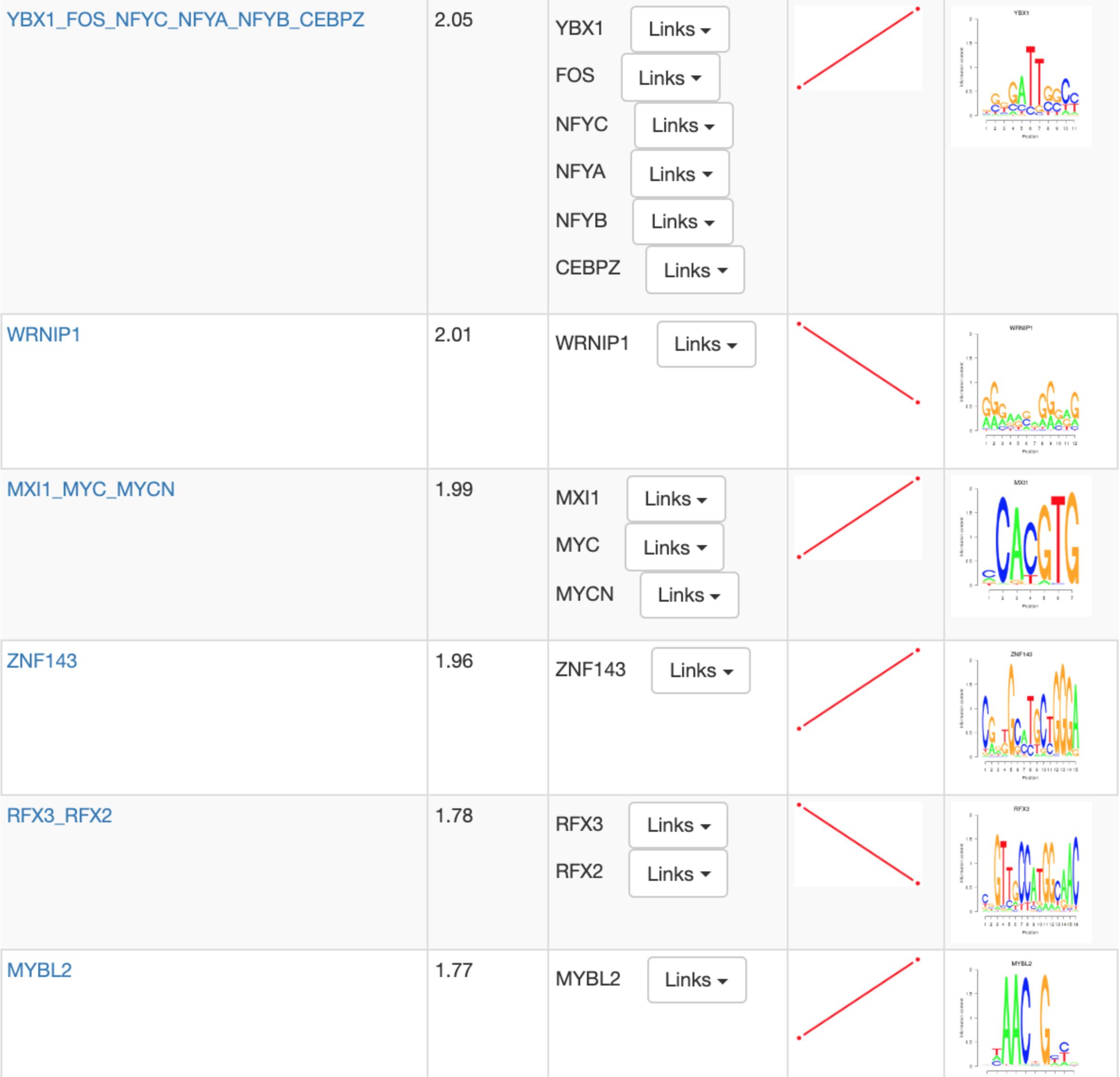
Every first replicate is a first batch. Every second replicate is a second batch.

Motifs dis-regulated in tumor cells

Dataset: GNF atlas of 79 tissues and cell lines + NCI atlas of 60 reference cancer cell lines

- Samples were divided into two groups: cancer samples and non-cancer samples.
- Average activities, error bars and Z-values were calculated for these groups.
- Top motifs are strongly associated with cancers.

Top motifs in cancers *vs* non-cancers dataset



YBX1_FOS_NFYC_NFYA_NFYB_CEBPZ	2.05	YBX1 FOS NFYC NFYA NFYB CEBPZ	Links ▾ Links ▾ Links ▾ Links ▾ Links ▾ Links ▾		
WRNIP1	2.01	WRNIP1	Links ▾		
MXI1_MYC_MYCN	1.99	MXI1 MYC MYCN	Links ▾ Links ▾ Links ▾		
ZNF143	1.96	ZNF143	Links ▾		
RFX3_RFX2	1.78	RFX3 RFX2	Links ▾ Links ▾		
MYBL2	1.77	MYBL2	Links ▾		

Motifs with high activity in cancer cells

Motifs with low activity in cancer cells