

RSAT Virtual machine (VM) User Guide

Jacques van Helden

Last update: 2016-06-24

Contents

1	Introduction	1
2	Instantiating an RSAT Virtual Machine on the IFB cloud	1
3	A quick tour of the tools: from gene clusters to motifs	6
3.1	Protocol	6
3.1.1	1. Supported organisms	6
3.1.2	2. Getting genes by name	6
3.1.3	3. Retrieving upstream sequences	6
3.1.4	4. Discovering over-represented k-mers in promoter sequences	16
3.1.5	5. Predicting binding sites in promoter sequences	16
3.1.6	6. Displaying the predicted binding sites	16

1 Introduction

This tutorial explains the steps to instantiate an RSAT Virtual Machine on the cloud of the French Institute of Bioinformatics IFB cloud and perform some basic operations with regulatory sequences and motifs.

2 Instantiating an RSAT Virtual Machine on the IFB cloud

1. Open a connection to the IFB cloud at <http://www.france-bioinformatique.fr/fr/cloud> and click on the link **Se connecter** to access the login window.
2. Cliquez sur le bouton **New Instance**.
3. In the dialog box, select the appliance **RSAT (2016-06)**, fill the **Name** field (for example type ***RSAT-VM**) and click **Run****.
4. The new instance will take a couple of minutes to start. Click periodically on **Show Instances** until your new instance appears with a green light, and the *ssh* and *http* links are active.
5. Click then on the **http** link for the RSAT-VM instance. This brings you to the home page of your own RSAT server.



Figure 1: Login window

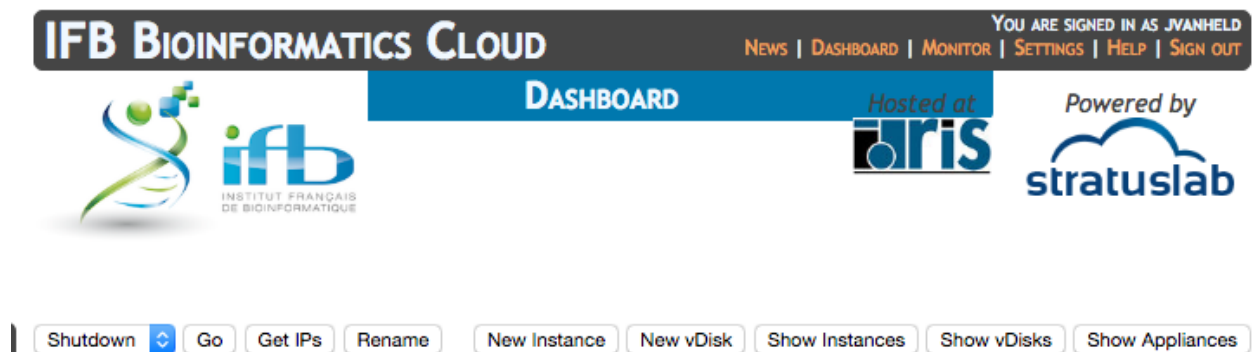


Figure 2: Click on the button **New Instance** to start a new virtual machine.

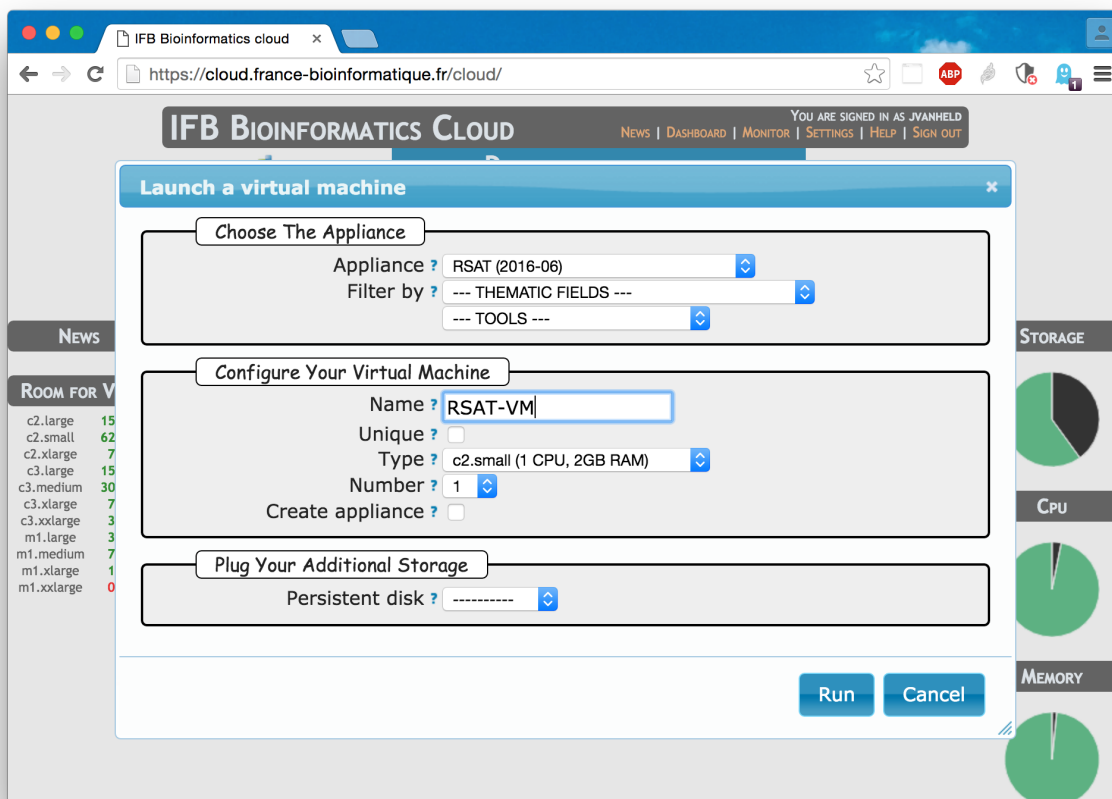


Figure 3: Dialog box for the configuration of a new instance of VM on the IFB cloud.

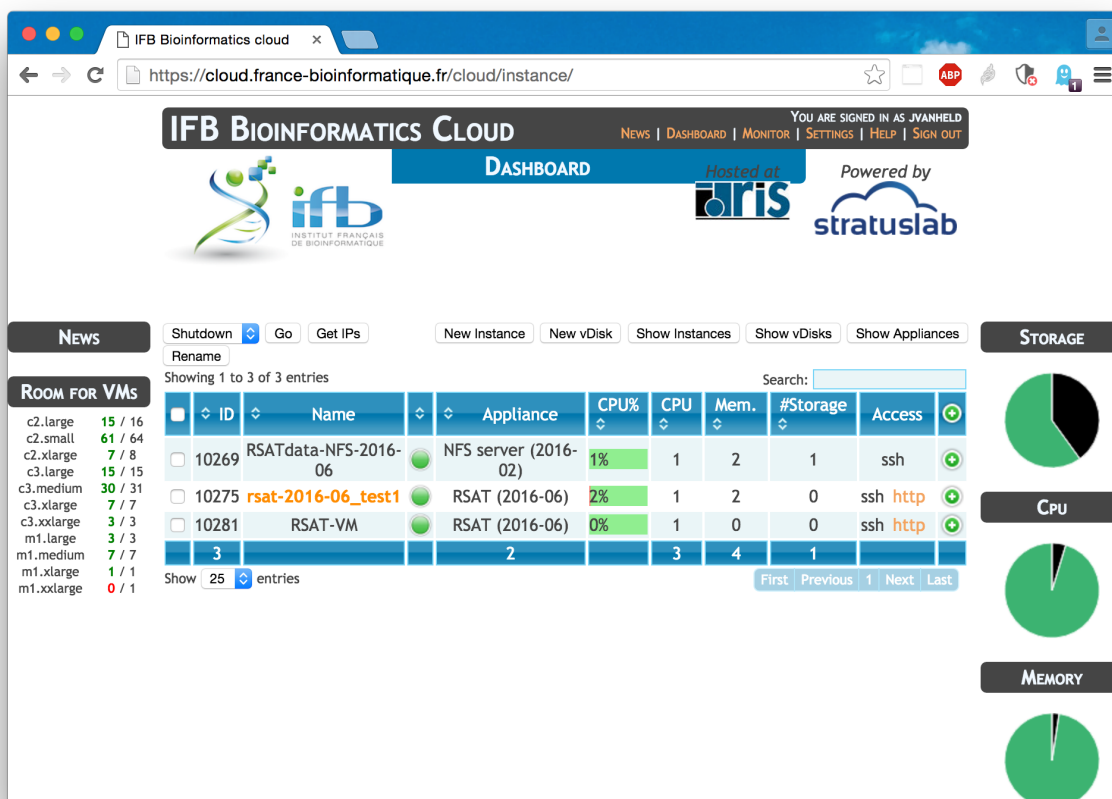


Figure 4: List of running instances on the IFB cloud. The list is user-specific, for this tutorial you only need the RSAT-VM instance.

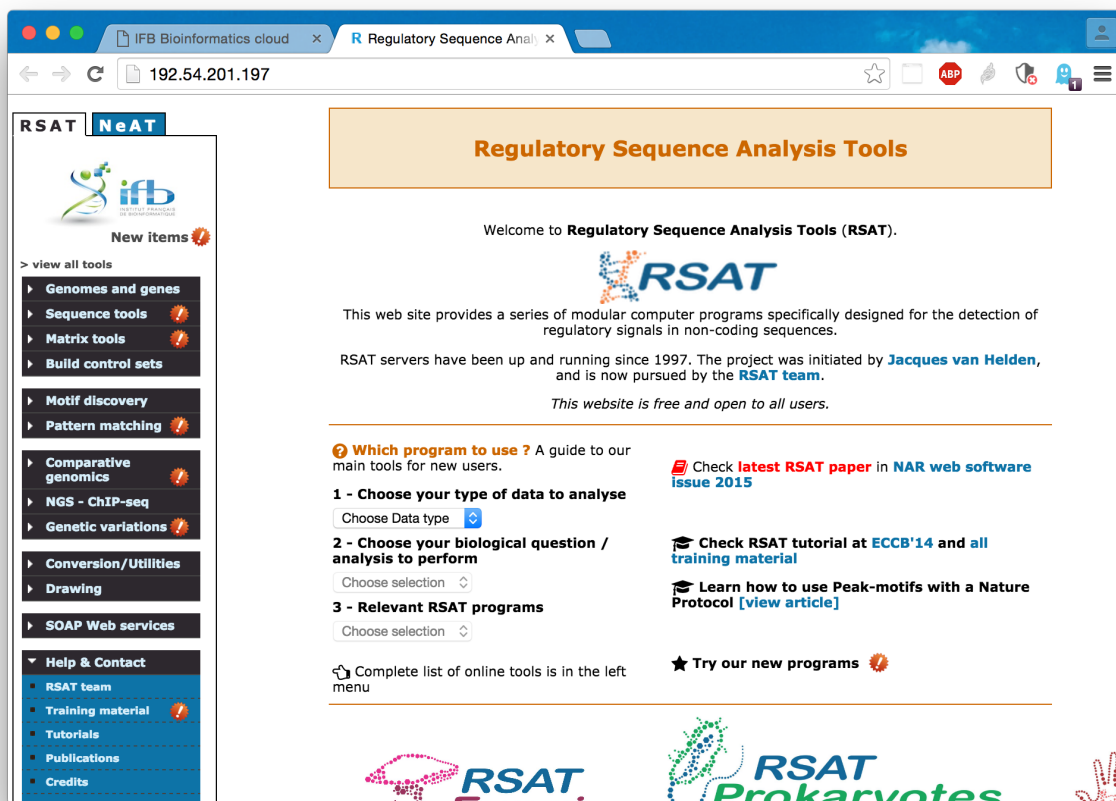


Figure 5: Home page of the RSAT virtual machine.

3 A quick tour of the tools: from gene clusters to motifs

We will run a quick tour of some simple modular tools of the RSAT software suite. We will successively run the following analyses:

1. Get the list of supported organisms.
2. Select a group of yeast genes involved in a common biological process (methionine metabolism and transport), which are supposedly co-regulated by some transcription factors.
3. Retrieve the non-coding sequences located upstream of these genes. These upstream sequences contain the gene promoter and the cis-regulatory elements.
4. Apply an *ab initio* motif discovery approach based on the detection of over-represented k)mers (*oligo-analysis*) in order to detect motifs potentially involved in the transcriptional response of these genes.
5. Scan the promoter sequences to detect the sites (positions) matching the discovered motifs.
6. Generate a feature-map to inspect the position of these sites.

3.1 Protocol

3.1.1 1. Supported organisms

In the left panel, expand the menu **Genomes and genes** and click on the tool **Supported organisms**.

3.1.2 2. Getting genes by name

We will now gather the genes involved in methionine metabolism and transport. In the yeast *Saccharomyces cerevisiae*, these genes are generally named with a prefix MET, followed by one or several numbers.

- a. Under **Genes and genomes**, click **Gene information**.
- b. In the **Organism** menu, select the species *Saccharomyces cerevisiae*.
- c. In the **Query** box, enter 'MET/d+'. This is a regular expression specifying that the gene name should contain the string MET followed by one or several digits ().
- d. Click **GO**.

After a few seconds, the result form should appear, a table with the genes whose name matched the query, followed by a table of links to the result files, and another table **Next Step** of possible tools for the next step of the analysis.

3.1.3 3. Retrieving upstream sequences

- a. In the **Next step** table of the gene-info result, click on the button **retrieve sequence**. The **Retrieve sequence** form is displayed, where the organism and gene query box have automatically been filled with the results of your gene-info query.
- b. Leave all other parameters unchanged and click **GO**. After a few seconds, the result page is displayed.
- c. Optionally, in the table **Result files**, click on the link to the sequence file (fasta), to inspect the result.
- d. Come back to the retrieve-seq result page. In the **Next step** table which appears at the bottom of this result page, click on the button **oligo-analysis**.

RSAT - Supported organisms

RSAT instance: RSATVM-IFB-2016-06

Organisms supported: 1535

Group specificity: None

ID	nb	source	last_update
Abiotrophia_defectiva_atcc_49176_GCA_000160075.2	1	ensemblgenomes	2016-06-23.123035
Absidia_idahoensis_var_thermophila.Lramosa_hybrid_454_Illumina.30	2	ensemblgenomes	2016-06-23.043103
Acaryochloris_marina_mbic11017_GCA_000018105.1	3	ensemblgenomes	2016-06-23.123037
Acetobacter_aceti_1023_GCA_000691125.1	4	ensemblgenomes	2016-06-23.043754
Acetobacteraceae_bacterium_at_5844_GCA_000245075.1	5	ensemblgenomes	2016-06-23.043810
Acetobacterium_woodii_DSM_1030_uid88073	6	NCBI	2016-06-23.043817
Acetohalobium_arabaticum_DSM_5501_uid51423	7	NCBI	2016-06-23.123041
Acetonebma_longum_dsm_6540_GCA_000219125.2	8	ensemblgenomes	2016-06-23.123043
Acholeplasma_laidlawii_PG_8A_uid58901	9	NCBI	2016-06-23.123044
Achromobacter_xylosoxidans_A8_uid59899	10	NCBI	2016-06-23.123046
Acidaminococcus_fermentans_dsm_20731_GCA_000025305.1	11	ensemblgenomes	2016-06-23.123047
Acidimicrobiae_bacterium_YM16_304_uid193703	12	NCBI	2016-06-23.123049
Acidimicrobium_ferrooxidans_DSM_10331_uid59215	13	NCBI	2016-06-23.123050
Acidiphilium_cryptum_JF_5_uid58447	14	NCBI	2016-06-23.123052
Acidithiobacillus_ferroxidans_atcc_23270_GCA_000021485.1	15	ensemblgenomes	2016-06-23.123054
Acidithrix_ferroxidans_GCA_000949295.1	16	ensemblgenomes	2016-06-23.123056

Figure 6: List of supported organisms on the RSAT Virtual Machine of the IFB cloud. The current version (June 2016) supports 1535 species, whose genomes were downloaded from various sources (EnsemblGenomes, NCBI).

IFB Bioinformatics cloud x R Regulatory Sequence Analysis x

192.54.201.197

RSAT NeAT

RSAT - gene-info

Returns the information about genes (CDS, mRNA, ...) specified either by their identifier, name, or by any supported synonym. Searches can also be done by specifying a sub-string of the gene descriptions. Regular expressions are supported.

Organism not
seeing your favorite organism in the list ? Contact us to have it installed

[Gene queries](#)

Upload gene list from file
 No file chosen

Feature type ☒ gene ☐ mRNA ☐ CDS

☐ Full string matching

☐ Match queries against description

Output ☒ display

[MANUAL MAIL](#)

Genomes and genes

- supported organisms
- gene information
- Infer operons
- get orthologs
- random gene selection

Sequence tools

- Matrix tools
- Build control sets

Motif discovery

- Pattern matching

Comparative genomics

- NGS - ChIP-seq
- Genetic variations

Figure 7: Query form of the **gene-info** tool.

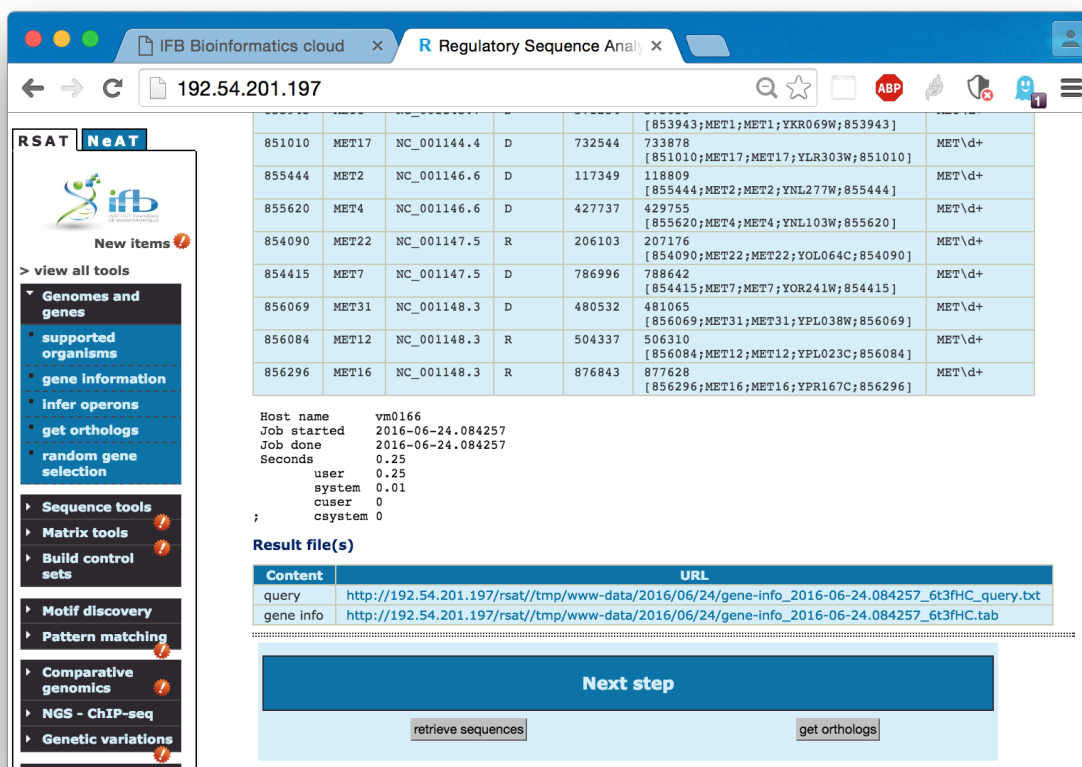


Figure 8: Result page of the **gene-info** tool.

IFB Bioinformatics cloud x R Regulatory Sequence Analysis x

192.54.201.197

RSAT NeAT

RSAT - oligo-analysis

Analysis of oligomer occurrences in nucleotidic of peptidic sequences

Reference: van Helden, J., André, B. and Collado-Vides, J. (1998). . J Mol Biol 281, 827-42.

Warning !! For **vertebrate** genomes, analyses of complete promoters from **co-expressed gene groups** return **many false positive** (i.e. if you submit a random set of genes, you always get plenty of highly 'significant' motifs). This is likely to come from the heterogeneity of human sequences (mixtures of GC-rich and GC-poor promoters). However, analyses of **ChIP-seq peaks** return **very good** results. See the program *peak-motifs*.

Sequence transferred from previous query

Mask

Sequence type

☒ **purge sequences (highly recommended)**

Oligomer counting mode

Oligomer lengths ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☒ 6 ☒ 7 ☒ 8

Note: motifs can be larger than oligo sizes (oligos are used as seed for building matrices)

☒ **prevent overlapping matches**

Count on ☒ **return reverse complements together in the output**

Background model

☒ **Genome subset**

☒ **Organism** not seeing your favorite organism in the list ? Contact us to have it installed

☐ **Taxon**

Figure 9: **oligo-analysis** query form.

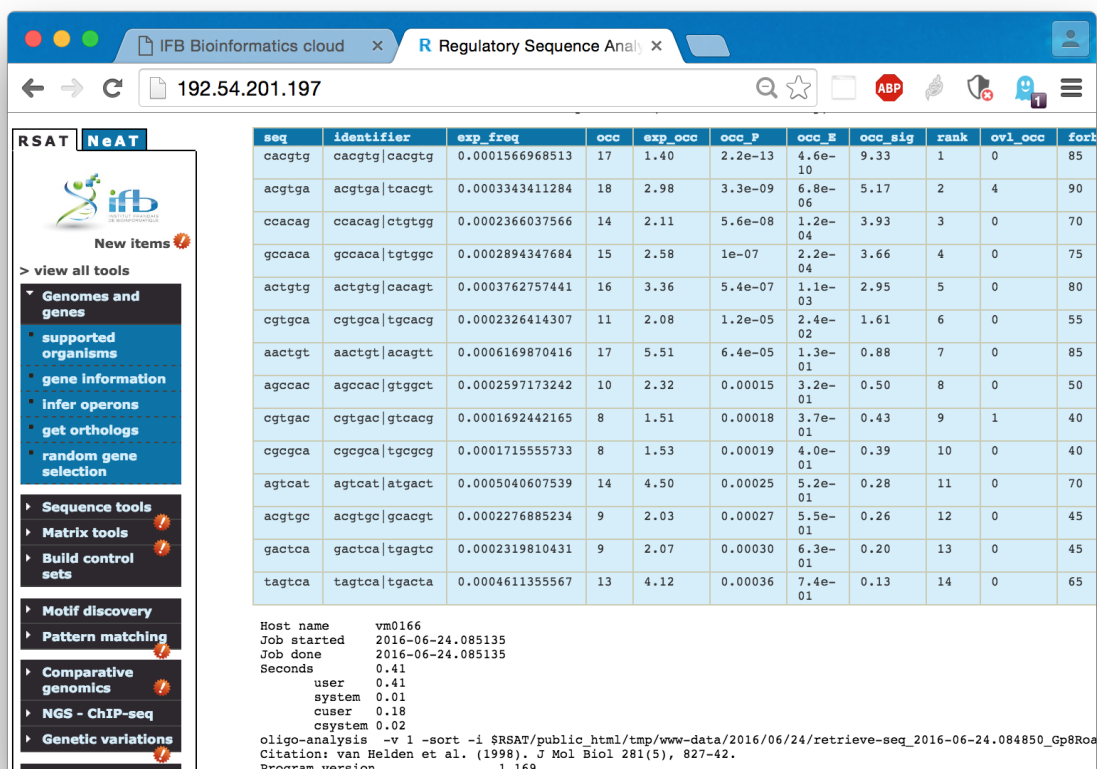


Figure 10: Primary result of **oligo-analysis**: list of over-represented k-mers.

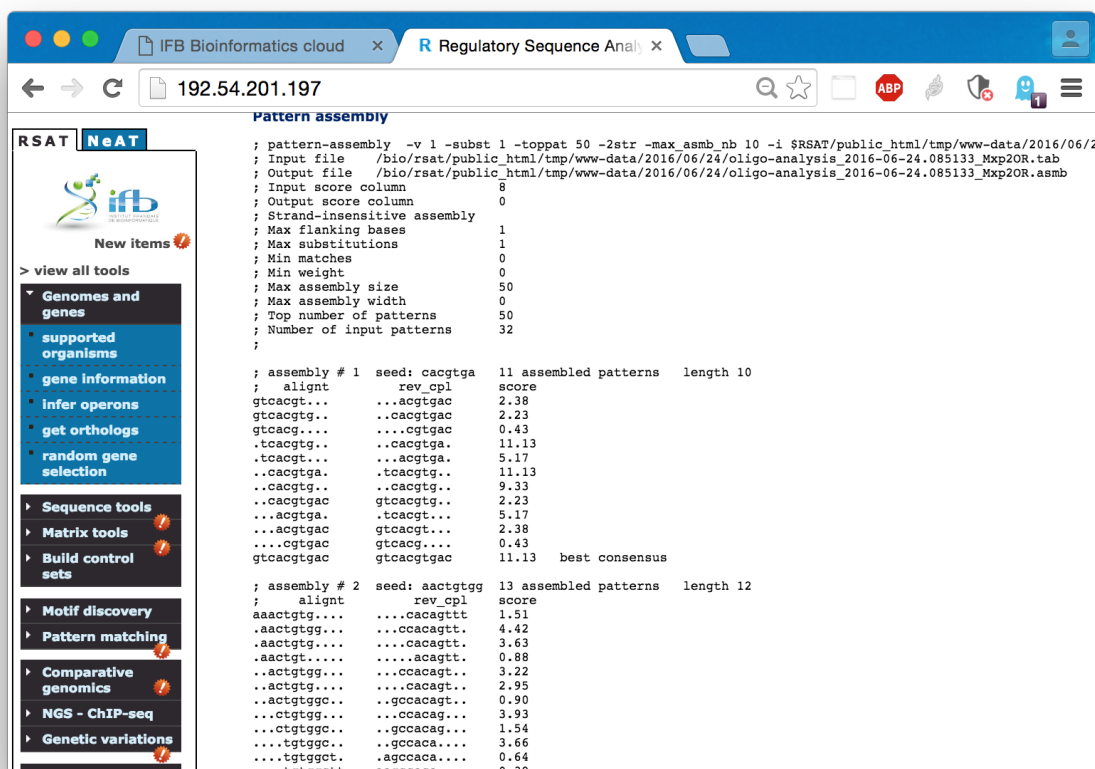


Figure 11: Assembly of the over-represented k-mers detected by **oligo-analysis**.



Figure 12: Position-specific matrices and logo representations of the motifs discovered. by oligo-analysis

RSAT NeAT

IFB Bioinformatics cloud x R Regulatory Sequence Analysis x

192.54.201.197

RSAT - dna-pattern

Search a pattern (string description) within a DNA sequence

Query pattern(s)

```
oligo-analysis -v 1 -sort -i $RSAT/public_html/tmp/www-data/2016/06/24/retrieve-seq_2016-06-24.084850_Gp8Roa.fasta.purged -format fasta -lth occ_sig 0 -uth rank 50 -return occ,proba,rank -2str -noov -quick_if_possible -seqtype dna -bg upstream-noorf -org Saccharomyces_cerevisiae -pseudo 0.01 -l 6 -o
```

Sequence transferred from previous query

Mask non-dna

Search strands both strands ☒ **prevent overlapping matches**

Return ☒ match positions **Origin** end **flanking** 4

☒ sequence limits

☐ non ACGT characters

☐ match counts **min count** 0

☐ match count table ☐ totals ☐ match scores ☐ match rank ☐ sort

☐ matching statistics

Substitutions 0

Output ☒ display

GO Reset DEMO [MANUAL TUTORIAL MAIL](#)

Genomes and genes

- supported organisms
- gene information
- infer operons
- get orthologs
- random gene selection

Sequence tools

Matrix tools

Build control sets

Motif discovery

Pattern matching

Comparative genomics

NGS - ChIP-seq

Genetic variations

Figure 13: Web form of the **dna-pattern** tool, which allows to scan sequences with string-based motifs (k-mers, consensus, regular expressions, ...).

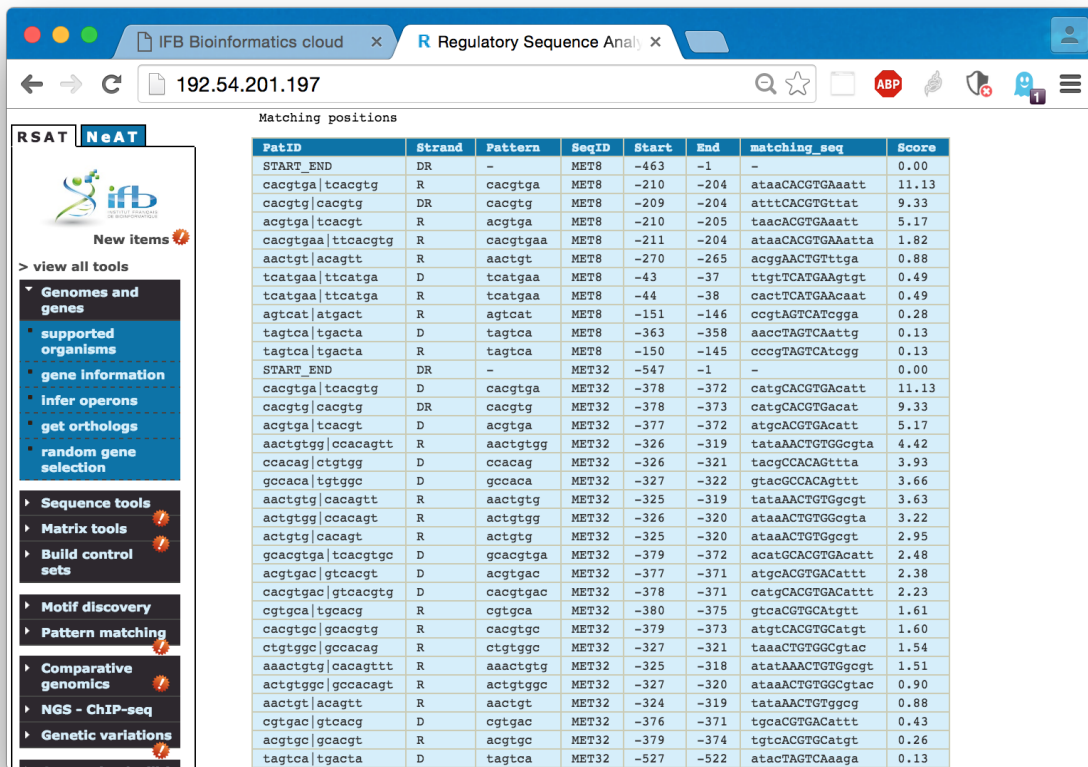


Figure 14: Matching positions for the over-represented k-mers in the yeast MET genes

- 3.1.4 4. Discovering over-represented k-mers in promoter sequences
- 3.1.5 5. Predicting binding sites in promoter sequences
- 3.1.6 6. Displaying the predicted binding sites

RSAT - feature map

Generates a graphical map of features localized on one or several sequences.

Feature list (Format: dna-pattern)

```
; dna-pattern -v 1 -pl $RSAT/public_html/tmp/www-data/2016/06/24/dna-
pattern_2016-06-24.085247_L5UujD.pat -i $RSAT/public_html/tmp/www-
data/2016/06/24/retrieve-seq_2016-06-24.084850_Gp8Roa.fasta.fasta
-format fasta -return sites -origin 0 -N 4 -return limits -noov -2str -subst 0
; Citation: van Helden et al. (2000), Yeast 16(2), 177-187.
; Input file $RSAT/public_html/tmp/www-data/2016/06/24/retrieve-//
```

File Choose File No file chosen

Title

☒ **Legend** ☒ **Scalebar** step auto

☒ **Sequence names** **Orientation** horizontal

Display limits From auto To auto origin 0

Map dimensions Length 500 thickness 25 spacing 2

Color palette color Color File Choose File No file chosen

Background color (R,G,B) 220,220,220

Feature handle none

Feature thickness max auto min auto ☒ **Proportional to score**

☒ **Dynamic map**

Figure 15: Web form of the **feature-map** tool.

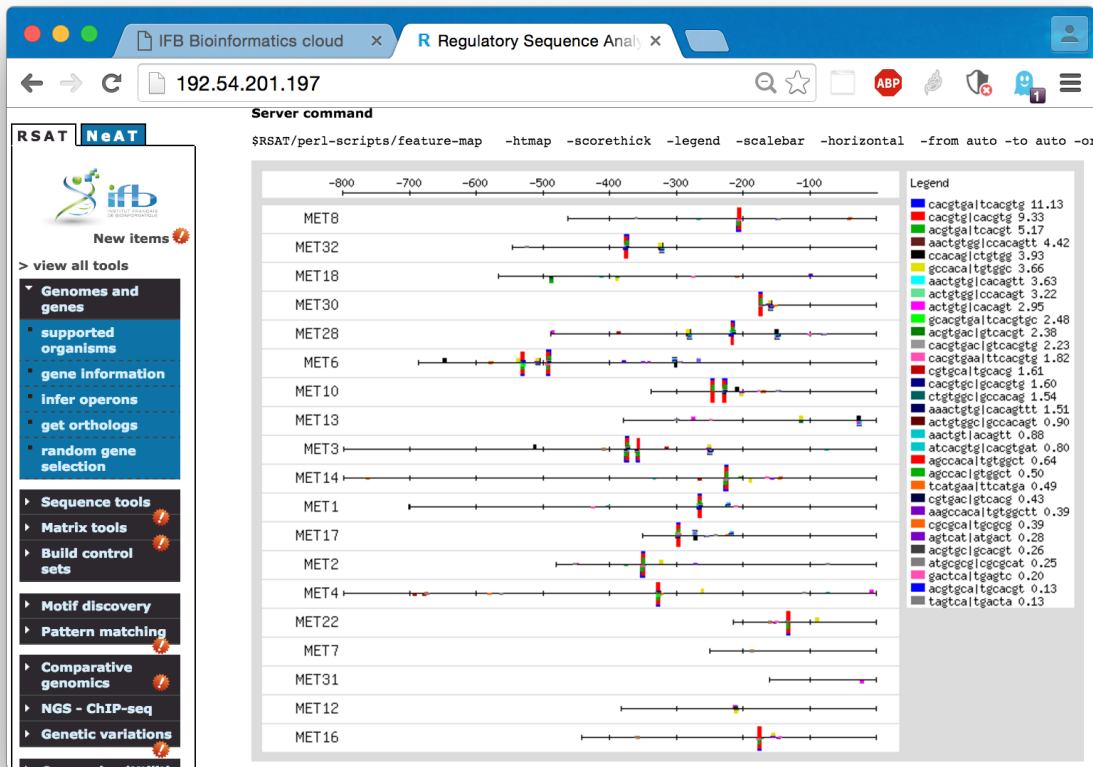


Figure 16: Feature map of the over-represented k-mers discovered by oligo-analysis and matched with dna-pattern in the previous steps. The putative transcription factor binding sites, which are revealed by clumps of mutually overlapping k-mers, corresponding to different fragments of the motifs.