ConTra v3: a tool to identify transcription factor binding sites across species, update 2017

Supplemental Material

Łukasz Kreft[1], Arne Soete[2,4], Paco Hulpiau[2,4], Alexander Botzki[1], Yvan Saeys[2,3] and Pieter De Bleser[2,4*]

[1]VIB Bioinformatics Core, Rijvischestraat 126 3R, 9052 Zwijnaarde-Ghent, Belgium

[2]VIB-UGent Center for Inflammation Research, Technologiepark 927, 9052 Zwijnaarde-Ghent, Belgium

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium

[4]Department of Biomedical Molecular Biology, Ghent University, Technologiepark 927, 9052 Zwijnaarde-Ghent, Belgium

*To whom correspondence should be addressed.

Address for correspondence:

Department of Biomedical Molecular Biology

VIB & Ghent University

Technologiepark 927, B-9052 Ghent, Belgium

Tel: +32.9.3313-693

Fax: +32.9.3313-609

E-mail: pieterdb@irc.vib-ugent.be

Contents

Legends to supplemental data

Supplemental Data S1

The ConTra v3 exploration analysis on the human IL2 promoter region (500 bp upstream) retrieves a table of results shown in (A), by default filtered on a q-value <= 0.25 and information content (IC) of >= 5 bits. The last three columns show the position in the ranking of the three criteria regulatory potential or RP rank, ChIP-seq based hotspot or HS rank and strongest conservation score or CE rank. NF-AT (V$NFAT_Q4_01) and ELF-1 (V$ELF1_Q6) are listed first and second in the table based on their q-value and IC. Both transcription factor binding sites are positioned in the top 10 of each ranking in (A). Octamer (V$OCT_Q6) is on position seven in the list with a very good HS rank of 8 but a bit lower ranked for RP (40) and CE (27). These three transcription factor matrices were selected for visualization also shown in Figure 1.

(B) The UCSC link on the visualization result page of ConTra v3 shows a view of the region of interest with the RefSeq track and ConTra v3 visualization track. Other tracks may be shown depending on the UCSC session of the user. For the view shown in (B) we have added a ConTra v3 exploration track of the same three predicted binding sites to illustrate the three criteria (RP, HS and CE). There are more NF-AT and ELF1 binding sites compared to OCT. Most of them are present both in highly conserved regions (100 vertebrates conservation track) and in hot spot regions (ReMapPublicENCODE track). Only two OCT sites are predicted indicated by the rectangle with orange background in (B). The second site is in a region with modest conservation explaining a lower CE rank for OCT. However both OCT sites are in a hot spot region based on the ReMapPublicENCODE track resulting in a very good HS rank of 8. A user interested in open chromatin regions can choose to show the UCSC track DNAse I hypersensitivity clusters from ENCODE as shown in (B).

A user interested in a specific criterium can remove the default q-value and IC filter and choose to look e.g. only at the top 5 scoring RP hits (C), top 5 HS hits (D) or top 5 scoring CE hits (E) by setting the respective filter on <= 5.

Supplemental Data S2

Riz et al reported an evolutionary conserved Nrf2/NFE2L2 motif in the GABARAPL1 promoter coinciding with an NF-E2 ChIP-seq binding site from the ENCODE project. In the case study in Supplemental Data S2 we show how a user can conduct an exploratory in silico analysis in ConTra v3 and predict how the autophagy-related GABARAPL1 gene is regulated as such.

Supplemental Data S3

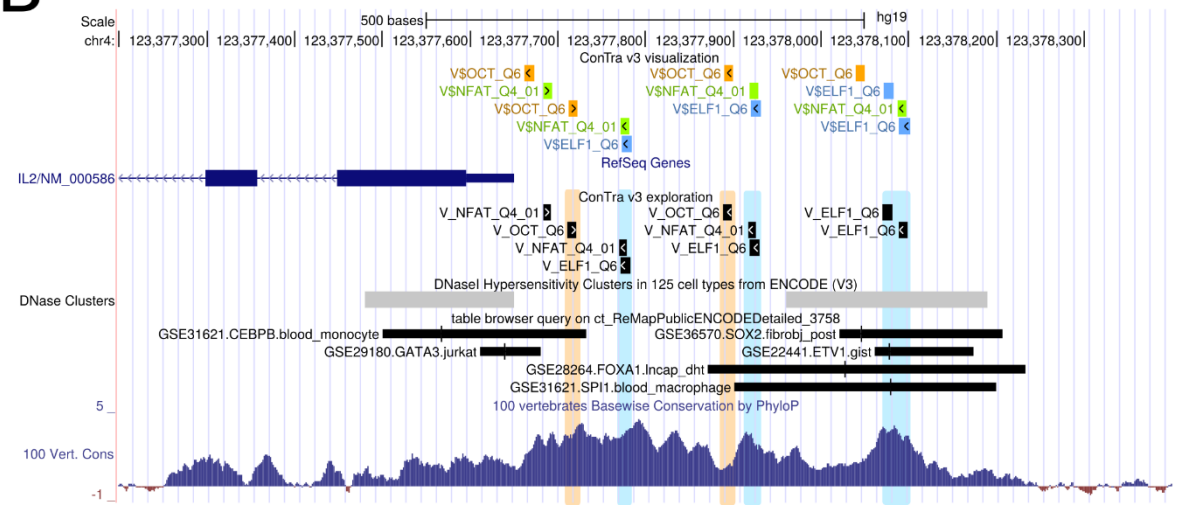
Bianchi et al. reported that intron removal resulted in a marked reduction of UbC promoter activity. Using electrophoretic mobility shift assays the authors demonstrated that Sp1 and Sp3 transcription factors can interact with several sites in the UBC intron. The case study in Supplemental Data S3 of the human UBC intron uses ConTra v3 to identify and visualize these Sp1 binding sites.

Supplemental Data S1

A

Filter / Select	Matrix ID	TF	PWM db	p-value <=	q-value <= 0.25	IC >= 5	Consensus	RP rank <=	HS rank <=	CE rank <=
	V$NFAT_Q4_01	NF-AT	TRANSFAC20113	3.87334e-6	0.0014654	6.23	GWGGAAAMWY	6	1	6
	V$ELF1_Q6	Elf-1	TRANSFAC20113	2.40848e-5	0.0022780	7.236	RDWASAGGAARW	3	9	5
	MA0398.1	SUM1	JASPAR_CORE_2016	3.13096e-4	0.0169217	6.198	AWWATTTWT	5	52	4
	MA0378.1	SFP1	JASPAR_CORE_2016	5.29349e-4	0.0250332	8.663	WYWRDRAAAAWTTTTYYWYKG	2	414	2
	MA0294.1	EDS1	JASPAR_CORE_2016	0.0032594	0.0948557	7.498	CGGAAAAAT	44	5	42
	MA0804.1	TBX19	JASPAR_CORE_2016	0.0025317	0.0948557	12.116	WTTMRCACCTAGGTGYGAAA	24	30	10
	V$OCT_Q6	Octamer	TRANSFAC20113	0.0030462	0.0948557	6.453	TDATTTGCATW	40	8	27
	MA0390.1	STB3	JASPAR_CORE_2016	0.0037323	0.0956778	11.039	GTYYAAAWTTTTTCACTYHKK	41	6	43
	MA0685.1	SP4	JASPAR_CORE_2016	0.0037935	0.0956778	13.827	YWRGCCACGCCCMCTYY	32	14	24

B



C

Matrix ID	TF	PWM db	p-value	q-value	IC	Consensus	RP rank	HS rank	CE rank
			<=	<=	>=		<= 5	<=	<=
F$SFP1_01	SFP1	TRANSFAC20113	8.33776e-6	0.0015772	3.128	NNNNNAAAAAATTTTNNNNNN	1	62	1
V$ELF1_Q6	Elf-1	TRANSFAC20113	2.40848e-5	0.0022780	7.236	RDWASAGGAARW	3	9	5
F$SUM1_02	SUM1	TRANSFAC20113	1.10558e-4	0.0069711	2.815	AAAATTTTT	4	36	3
MA0398.1	SUM1	JASPAR_CORE_2016	3.13096e-4	0.0169217	6.198	AWWATTTWT	5	52	4
MA0378.1	SFP1	JASPAR_CORE_2016	5.29349e-4	0.0250332	8.663	WYWRDRAAAAWTTTTYYWYKG	2	414	2

D

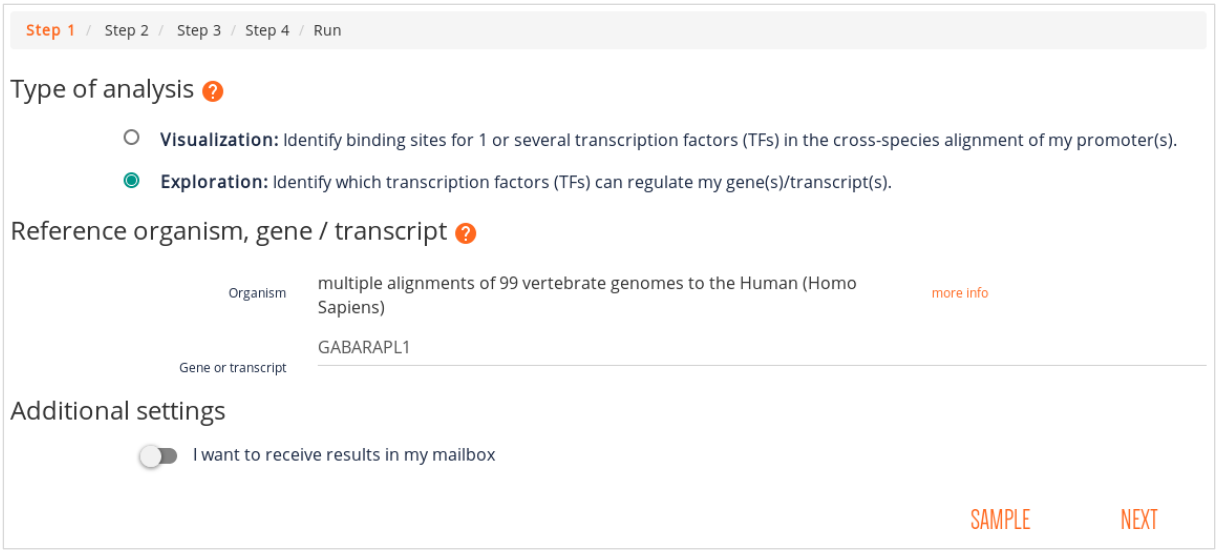
Matrix ID	TF	PWM db	p-value	q-value	IC	Consensus	RP rank	HS rank	CE rank
			<=	<=	>=		<=	<= 5	<=
V$NFAT_Q4_01	NF-AT	TRANSFAC20113	3.87334e-6	0.0014654	6.23	GWGGAAAMWY	6	1	6
M6365_1.02	NFATC4	Homo_sapiens	1.87355e-5	0.0022780	1.877	NNNTTTTCCN	8	2	7
M6363_1.02	NFATC2	Homo_sapiens	3.22389e-5	0.0024393	2.189	NNTTTTCCA	7	3	8
M6364_1.02	NFATC3	Homo_sapiens	0.0024786	0.0948557	2.502	ANTTTTCCA	43	4	41
MA0294.1	EDS1	JASPAR_CORE_2016	0.0032594	0.0948557	7.498	CGGAAAAAT	44	5	42

E

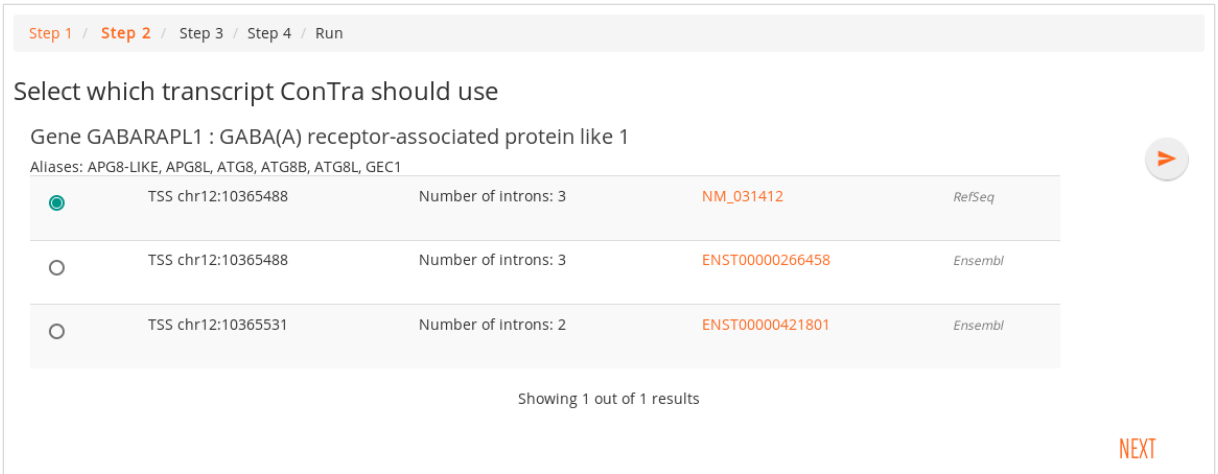
Matrix ID	TF	PWM db	p-value	q-value	IC	Consensus	RP rank	HS rank	CE rank
			<=	<=	>=		<=	<=	<= 5
F$SFP1_01	SFP1	TRANSFAC20113	8.33776e-6	0.0015772	3.128	NNNNNAAAAAATTTTNNNNNN	1	62	1
V$ELF1_Q6	Elf-1	TRANSFAC20113	2.40848e-5	0.0022780	7.236	RDWASAGGAARW	3	9	5
F$SUM1_02	SUM1	TRANSFAC20113	1.10558e-4	0.0069711	2.815	AAAATTTTT	4	36	3
MA0398.1	SUM1	JASPAR_CORE_2016	3.13096e-4	0.0169217	6.198	AWWATTTWT	5	52	4
MA0378.1	SFP1	JASPAR_CORE_2016	5.29349e-4	0.0250332	8.663	WYWRDRAAAAWTTTTYYWYKG	2	414	2

Supplemental Data S2

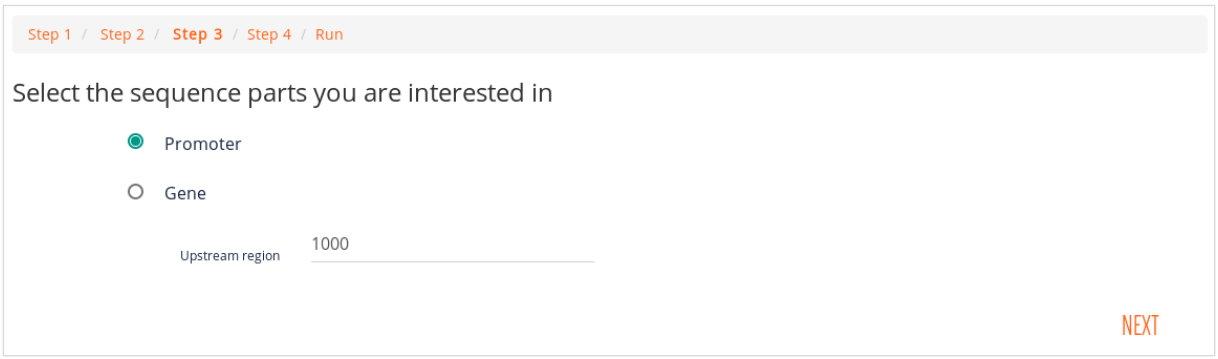
Step 1: Choose Exploration as Type of analysis and type the GABARAPL1 gene as input.



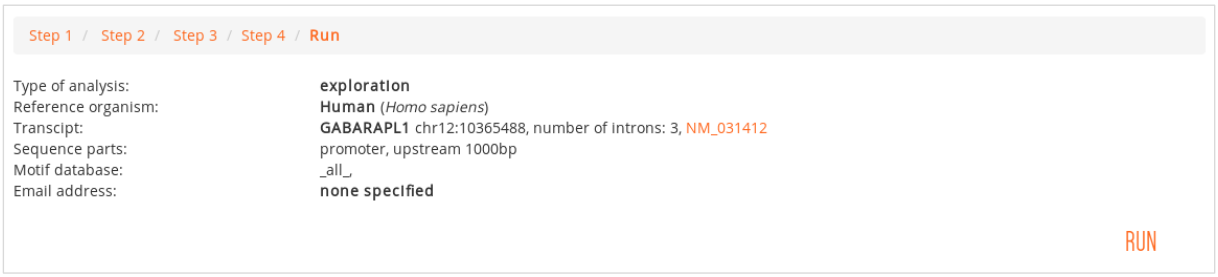
Step 2: There is one RefSeq transcript NM_031412 (default selection) and two Ensembl transcripts for the GABARAPL1 gene. Click Next.

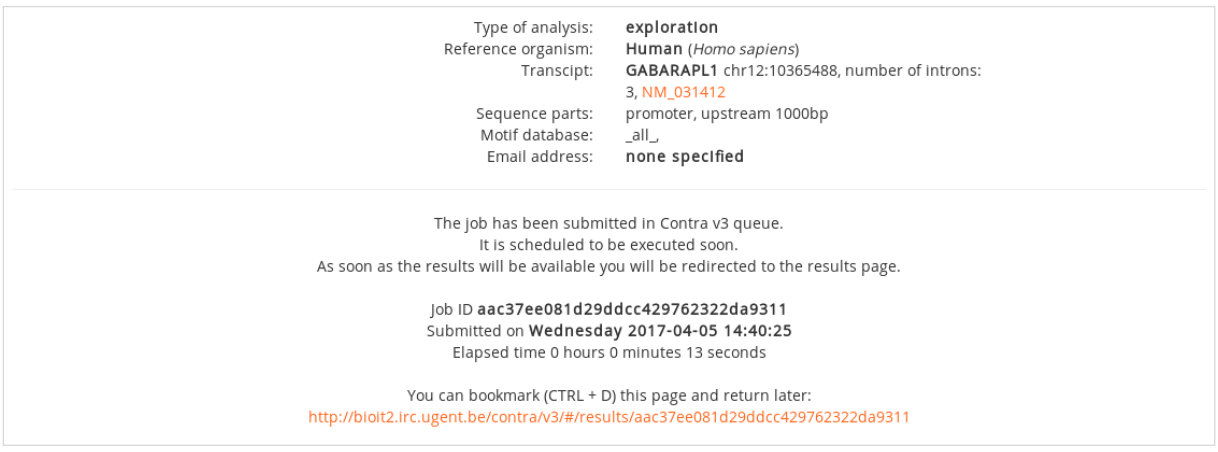


Step 3: Select Promoter and enter 1000 upstream region. Click Next.

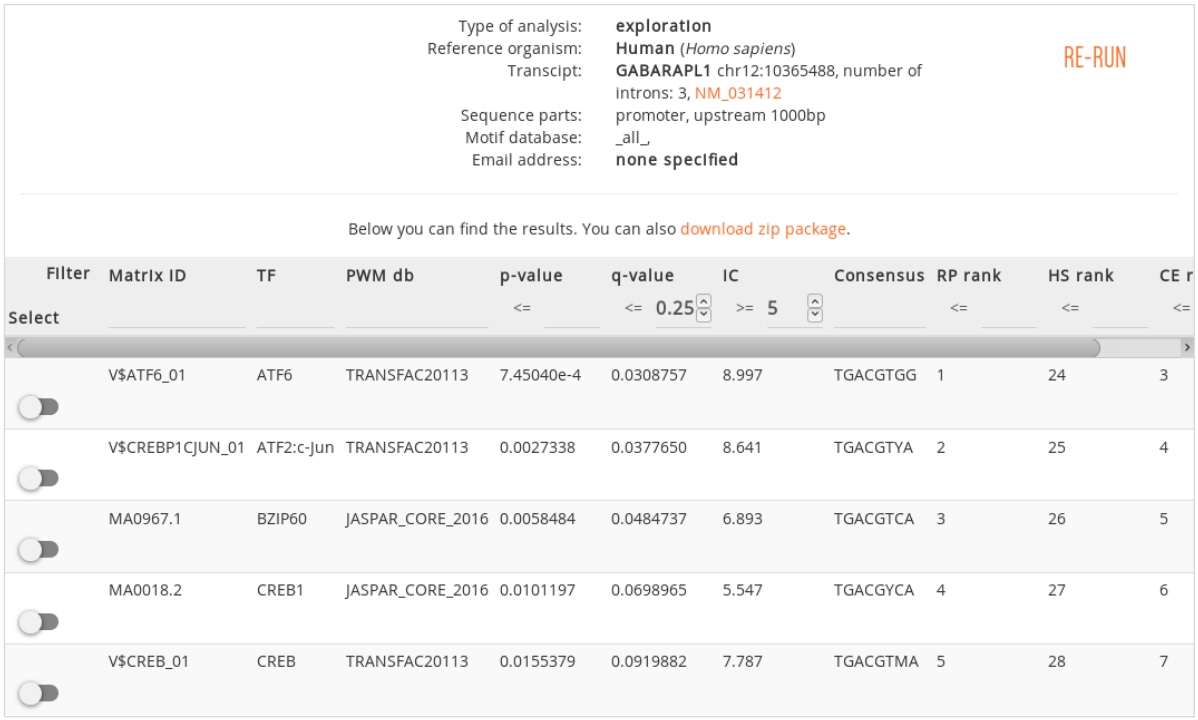


Step 4 to set the stringency and choose the transcription factor binding sites is **skipped in exploration mode**. Before starting the analysis, a user can review the input parameters by clicking on the relevant step in the breadcrumb. Clicking the Run button will submit the job in the server queue and look for all motifs in the database in the region of interest.





Results: The exploration results page shows a table with the top hits filtered by q-value <= 0.25 and information content (IC) >= 5 which gives the users an initial idea about putative regulators. For a longer list users can choose to change or remove the filters and/or search for specific transcription factors (TF) in the (preferably unfiltered) list.



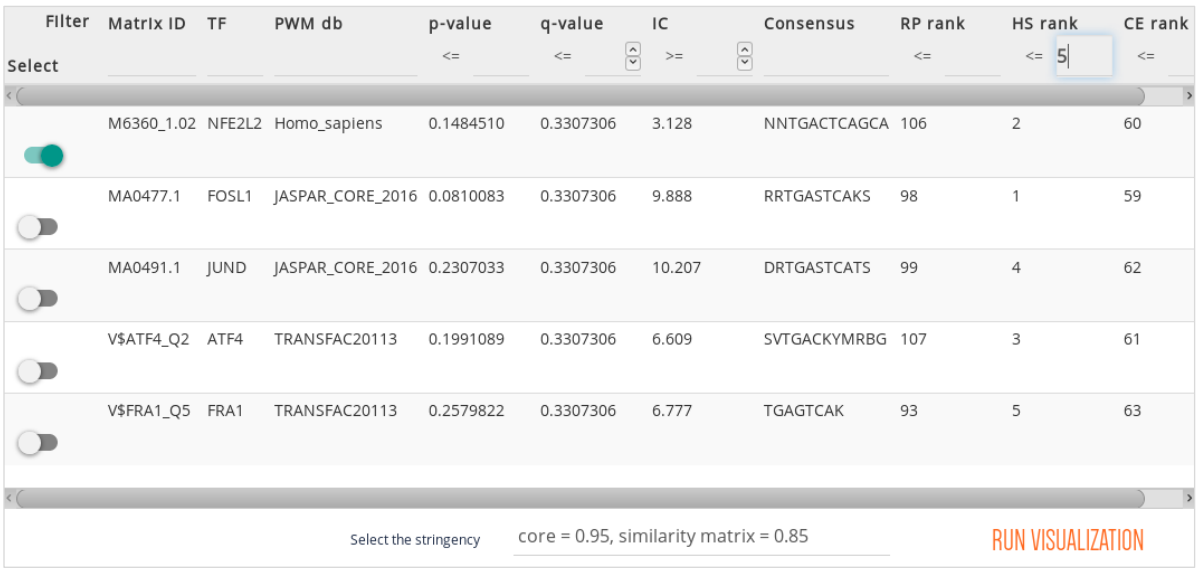
Type of analysis: **exploration**
Reference organism: **Human** (*Homo sapiens*)
Transcipt: **GABARAPL1** chr12:10365488, number of introns: 3, NM_031412
Sequence parts: promoter, upstream 1000bp
Motif database: _all_,
Email address: **none specified**

RE-RUN

Below you can find the results. You can also download zip package.

Filter / Select	Matrix ID	TF	PWM db	p-value	q-value	IC	Consensus	RP rank	HS rank	CE r
				<=	<= 0.25	>= 5		<=	<=	<=
	V$ATF6_01	ATF6	TRANSFAC20113	7.45040e-4	0.0308757	8.997	TGACGTGG	1	24	3
	V$CREBP1CJUN_01	ATF2:c-Jun	TRANSFAC20113	0.0027338	0.0377650	8.641	TGACGTYA	2	25	4
	MA0967.1	BZIP60	JASPAR_CORE_2016	0.0058484	0.0484737	6.893	TGACGTCA	3	26	5
	MA0018.2	CREB1	JASPAR_CORE_2016	0.0101197	0.0698965	5.547	TGACGYCA	4	27	6
	V$CREB_01	CREB	TRANSFAC20113	0.0155379	0.0919882	7.787	TGACGTMA	5	28	7

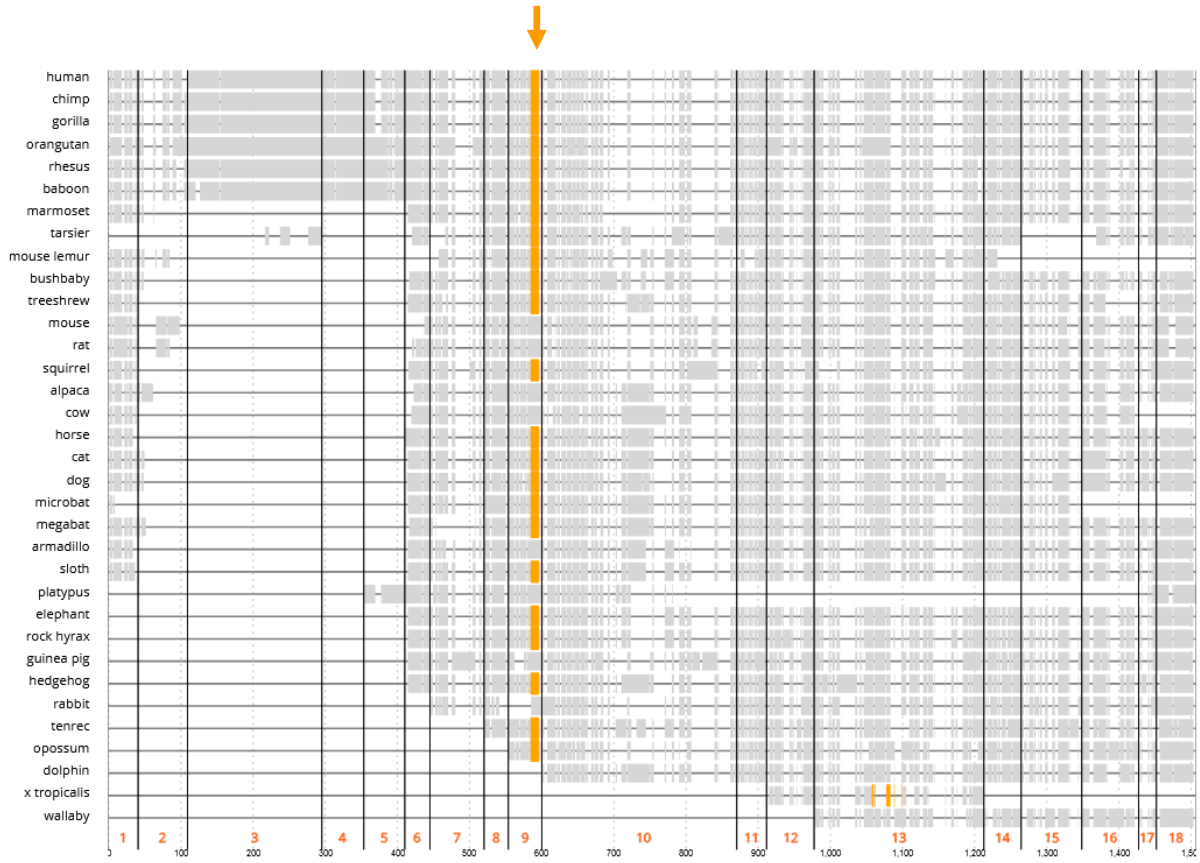
We advise users also to look at the individual ranked product analysis. This can be done by filtering on either regulatory potential (RP rank), ChIP-seq based hotspot (HS rank) or strongest conservation i.e. conserved element score (CE rank). Below the top 5 “ChIP-seq based” hotspot results (HS rank). Specific binding sites can be switched on to select them and run visualization.



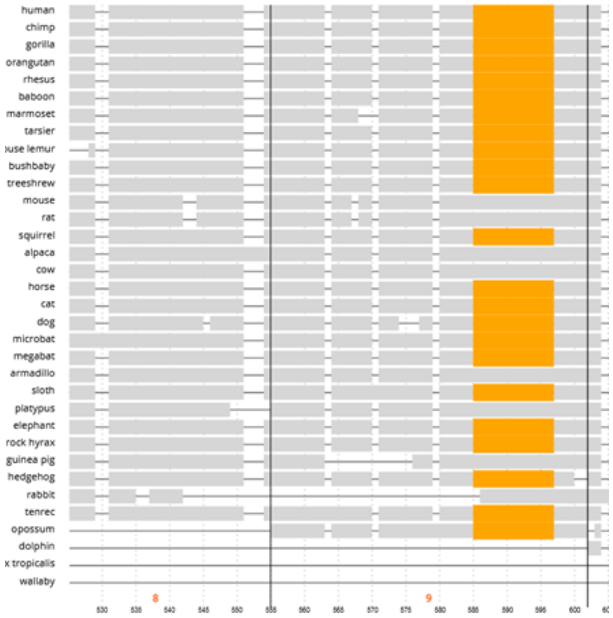
Filter / Select	Matrix ID	TF	PWM db	p-value	q-value	IC	Consensus	RP rank	HS rank	CE rank
				<=	<=	>=		<=	<= 5	<=
	M6360_1.02	NFE2L2	Homo_sapiens	0.1484510	0.3307306	3.128	NNTGACTCAGCA	106	2	60
	MA0477.1	FOSL1	JASPAR_CORE_2016	0.0810083	0.3307306	9.888	RRTGASTCAKS	98	1	59
	MA0491.1	JUND	JASPAR_CORE_2016	0.2307033	0.3307306	10.207	DRTGASTCATS	99	4	62
	V$ATF4_Q2	ATF4	TRANSFAC20113	0.1991089	0.3307306	6.609	SVTGACKYMRBG	107	3	61
	V$FRA1_Q5	FRA1	TRANSFAC20113	0.2579822	0.3307306	6.777	TGAGTCAK	93	5	63

Select the stringency: core = 0.95, similarity matrix = 0.85

RUN VISUALIZATION



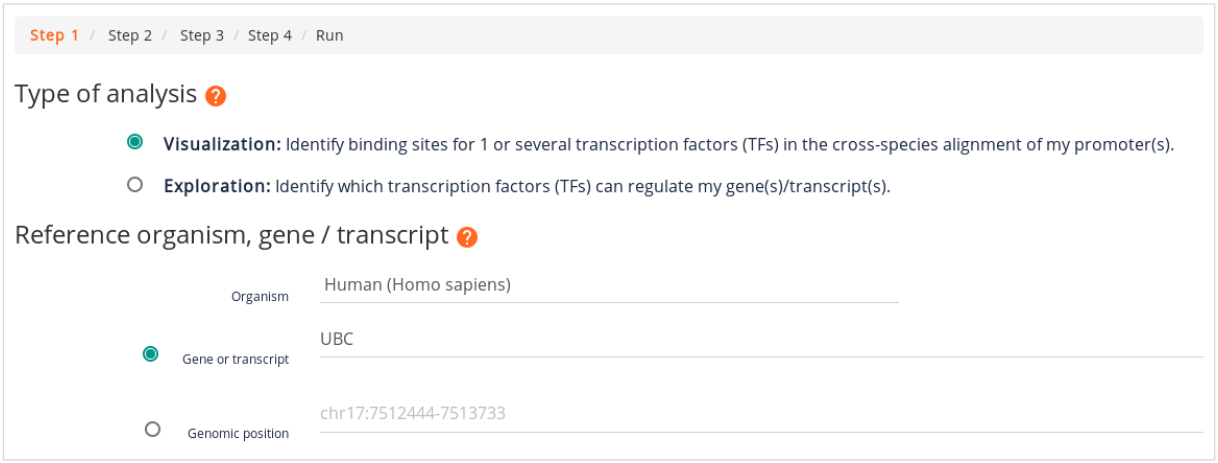
Selection of NFE2L2/NRF2 reveals the highly conserved site reported by Riz et al in the graphical visualization overview (figure above) and can be zoomed in (figure below left). Below the graphical overview users can see the conserved NFE2L2/NRF2 binding site at nucleotide level (below right).



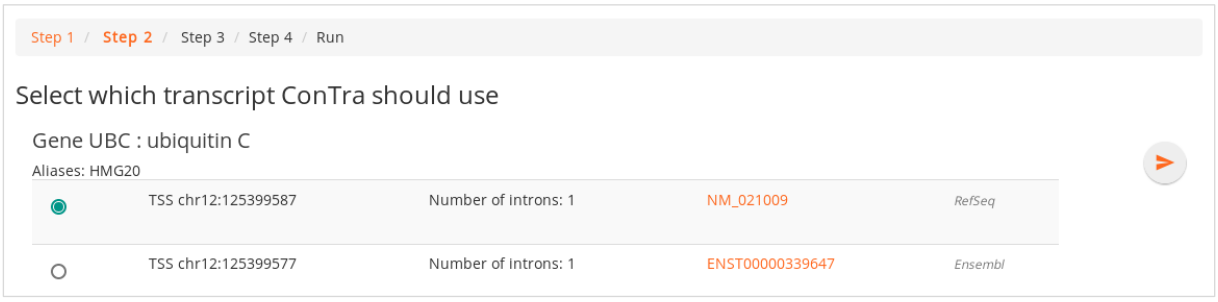

```
human          --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
chimp          --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
gorilla        --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
orangutan      --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
rhesus         --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
baboon         --GTCCTTCAT-CTGACT-CCTCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
marmoset       --GTCCTTCAC-CTGA---CATCTCTT-CAGATTCCTGAGTCACGCTCTGTT----
tarsier        --GCCCTTCAC-CTTACA-CATCTCTT-CAAATTCCTGAGTCACGCTCTGTT----
mouse lemur    --GCCCTTCAC-CTGACA-CGTCTCTC-CGAACTCCTGAGTCACGCTGTGCC----
bushbaby       --GCCCTTCAC-CTTACA-CATATTTC-AAAATTCCTGAGTCACGCTCCGCT----
treeshrew      --GCCCTTCAC-CTTAGG-CATCTCTT-CAAACTCCTGAGTCACGCTCTGTT----
mouse          --GTTCTTCAC-CCT-TA-CACTTCTT-CAAATCCATGAGTCATGCTCTGTT----
rat            --GTTCTTCAC-CCT-TA-CACTTCTT-CAAATCCATGAGTCACGCTGTGTT----
squirrel       --GCCCTTTGC-TTTGCA-CACCCTTT-CAAATTCCTGAGTCACGCTCTGTT----
alpaca         ACGCCCTTCAG-CTTATA-CGTCTCAT-CAAATCCCTGAGTCATGCTCTGTT----
cow            --GCCCTTCAC-CGGCTA-CATCTCTT-CAGTTCCCTGAGTCATGCTCTGTT----
horse          --GCCCTTCAC-CTTATA-CATCTCTT-CAAATTCCTGAGTCATGCTCTGTT----
cat            --GCCCTTCAC-CTTATA-CATCTCCT-CAAATTCCTGAGTCATGCTCTGTT----
dog            --GCCCTTCAG-CTTCTA-CAT---CT-CAAATTCCTGAGTCATGCTCTGTT----
microbat       --GCCCTTCAC-CTAATA-CGTCTCTT-CGAATTCCTGAGTCACGATTTGTT----
megabat        --ACCCTTCAC-CTAATA-TGCCTTTT-GGAATTCCTGAGTCACGTTTTGTT----
armadillo      --ACCCTTCAT-CTTACG-CATCTCCT-CAGATGCCTGAGTCACTCTCTGTT----
sloth          --GCCCTTCAT-CGCAGGACATCTTCT-CAGATTCCTGAGTCACGCTGTGTT----
platypus       ---CCCTCAGT-CTTCCA-CGGAGCAG-AAGGTCCTTGAGTCACCCCGTGTT----
elephant       --GCCCTTCAT-ATTACA-GGTCTCTT-CAAATTCCTGAGTCATGCTCAGTT----
rock hyrax     --GCCCTTCAC-CTTACA-TATCTCTT-CAAACTCCTGAGTCATGCTCTGTT----
guinea pig     --GCCCTTCAC-------------CTT-CAAACGCCTGAGTCACGCTCCGTT----
hedgehog       --GCCCTGCAC-TTTAAA-CATCTCTT-CAAATTCCTGAGTCAAGCTG--TT----
rabbit         ---------------------------------CCCGAGTCACGCCCTGCTGACC
tenrec         --GCCCTTCACGCTTACA-CAGCTCTT-CAAATTCCTGAGTCATGCTCTGTT----
opossum        ---TCTTTCCC-CGTTCT-CTCCCCTTCCAAAATCCTGAGTCACGCTGTT-T----
dolphin        ------------------------------------------------TT----
x tropicalis   --------------------------------------------------------
wallaby        --------------------------------------------------------
```

Supplemental Data S3

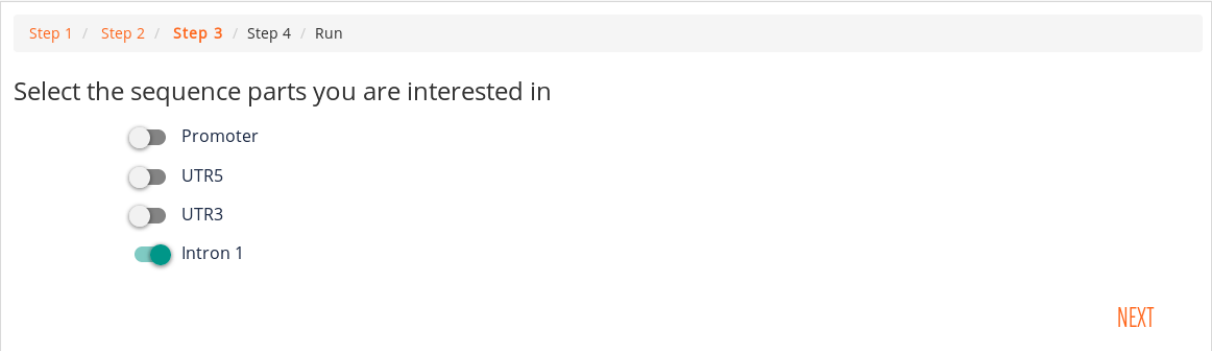
Step 1: Choose Visualization as Type of analysis and enter the human UBC as reference organism and gene respectively.



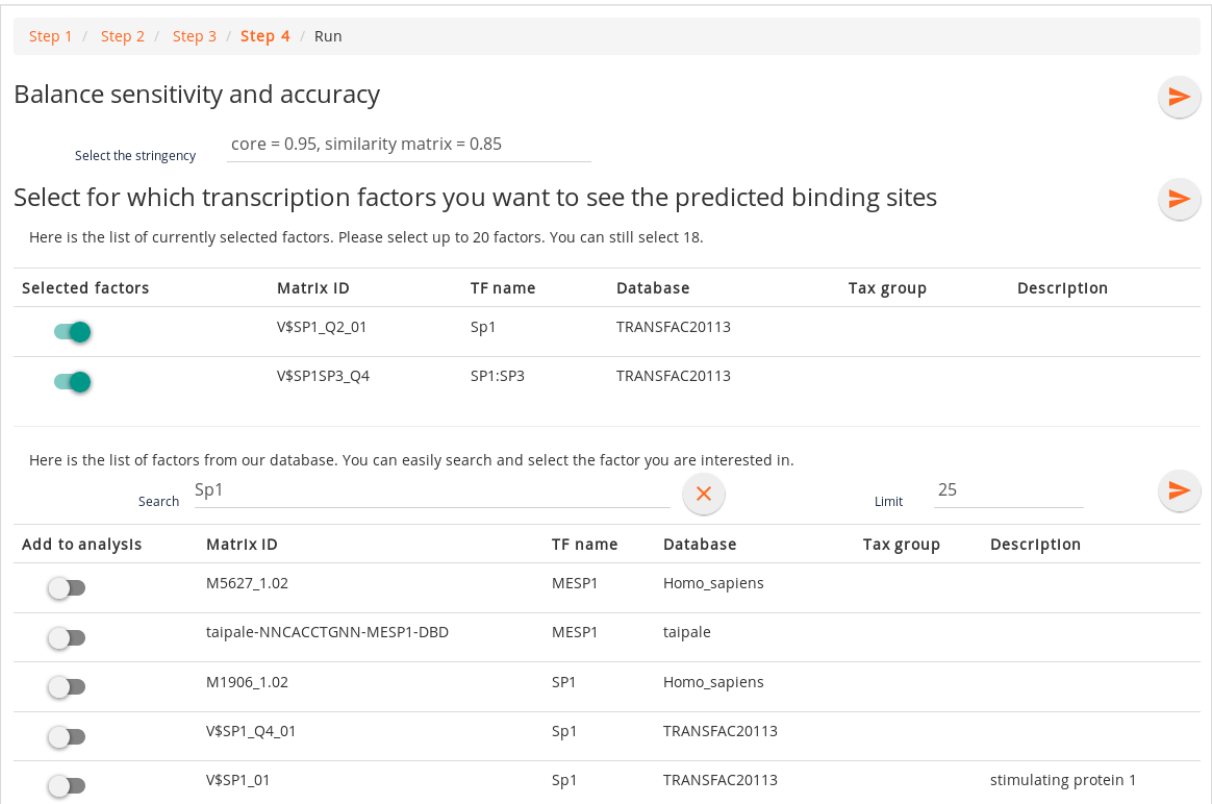
Step 2: There is one reference transcript for the UBC gene. The default choice is the NCBI RefSeq NM_021009. A user could also choose to use the Ensembl transcipt ENST00000339647. Click Next. Tip: when a long list of transcripts is retrieved, clicking the orange arrow on the right of the transcript table will automatically jump to the Next button on the bottom of the page.



Step 3: Select Intron 1 and click Next

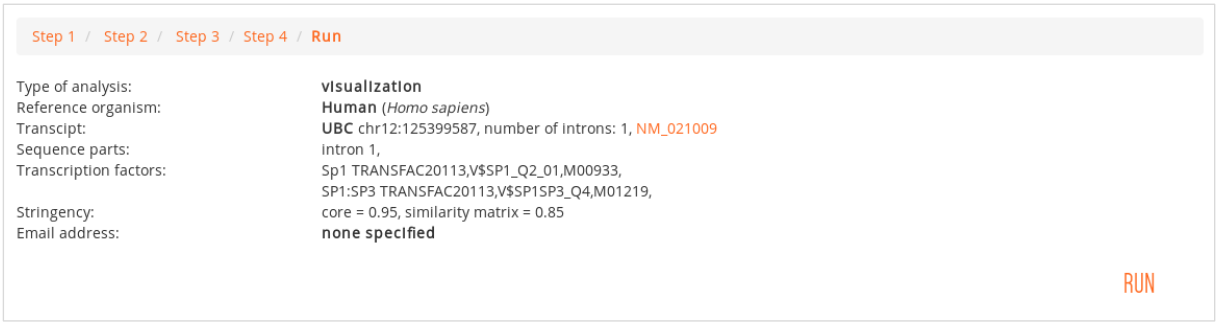


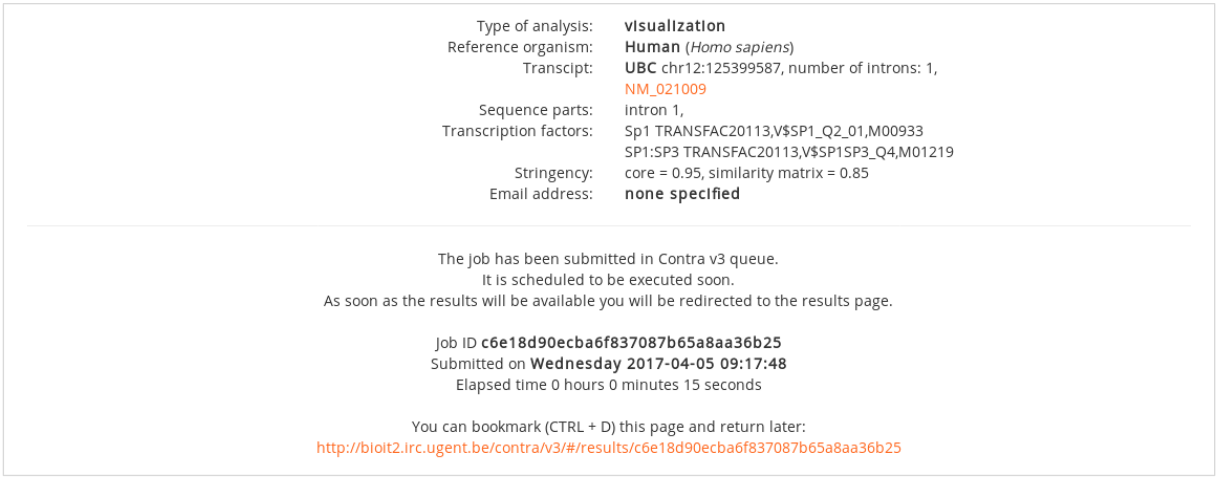
Step 4: The default stringency of 0.95 core similarity and 0.85 matrix similarity is a good start for most analyses. A more experienced user could choose another stringency based on e.g. the quality of the matrix. Searching for the transcription factor Sp1 retrieves several matrices. We have selected the Transfac Sp1 matrix V$SP1_Q2_01, indicating a high quality matrix of 2 (Q2) and the Transfac Sp1/Sp3 matrix V$SP1SP3_Q4.



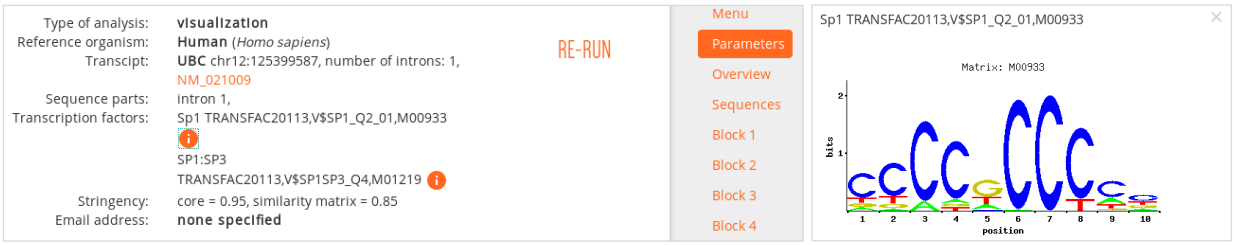
Run: Before starting the analysis a user can review the input parameters. Input can still be modified by clicking on the relevant step in the breadcrumb with the previously chosen input filled in.

Clicking the Run button will submit the job in the queue on the server.

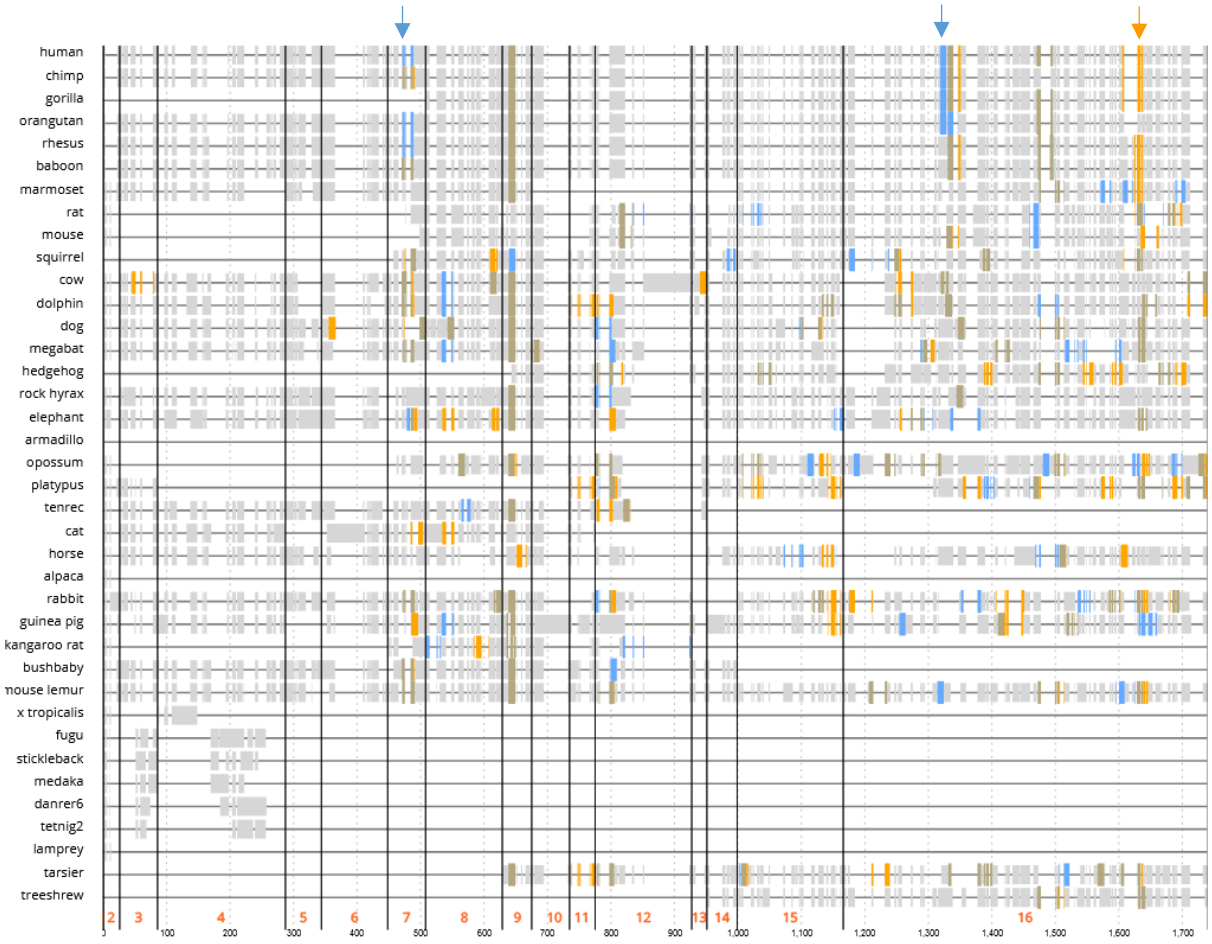




Results: The results page consists of several sections indicated by a navigation menu on the right. The first sections shows the input parameters and a re-run option shown on the left below. For each TF an info button is available which can show the sequence logo (shown on the right below) representing the position weight matrix (PWM) for the detected binding site.

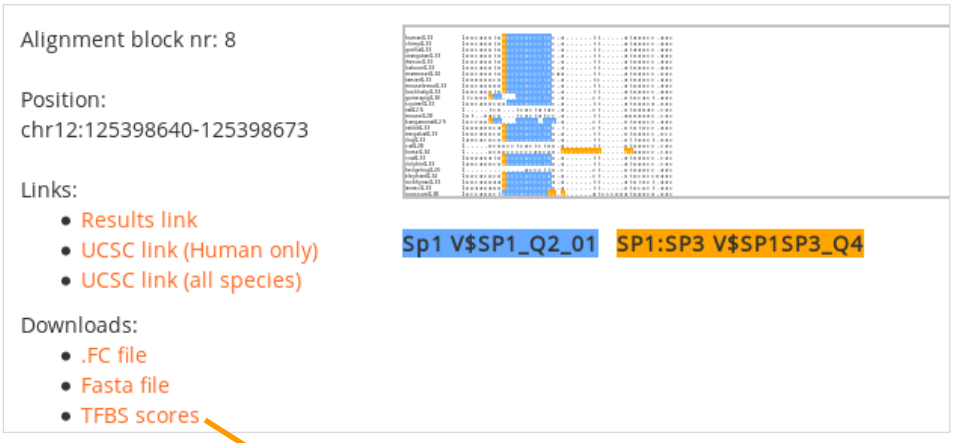


Next in the overview section both a graphical overview and sequence alignment overview is shown. Several conserved transcription factor binding sites (TFBS) have been identified in this example and most of them can easily be spotted in the graphical overview in the figure below.



In this example we selected two similar PWMs (Sp1 and Sp1:Sp3). When multiple sites are predicted on the same sequence ConTra mixes colors. However because we have analyzed a large region in this example overlapping conserved sites might not be that obvious in the graphical overview but can be easily seen in the detailed multiz alignment blocks in the third results section. Alignment block nr. 8 shown below illustrates a highly conserved Sp1/Sp3 site.

A fasta and feature color (.fc) file for each alignment block can be downloaded to generate high quality figures with Jalview (explained on the help page of ConTra v3). The TFBS scores can also be downloaded as a tab separated summary table. For each species the detected TFBS (matrix), position in the alignment block, strand, scores and sequence hit are shown.



Species	Matrix	Position	Strand	Core_Score	Matrix_Score	Seq
Human	V$SP1_Q2_01	10	(+)	0.966	0.945	ccCCACCctg
Chimp	V$SP1_Q2_01	10	(+)	0.966	0.945	ccCCACCctg
Gorilla	V$SP1_Q2_01	10	(+)	0.966	0.945	ccCCACCctg
orangutan	V$SP1_Q2_01	10	(+)	0.966	0.942	ccCCACCcta
Rhesus	V$SP1_Q2_01	10	(+)	0.966	0.966	ccCCACCcca
Baboon	V$SP1_Q2_01	10	(+)	0.966	0.942	ccCCACCcta
Marmoset	V$SP1_Q2_01	10	(+)	1.000	0.993	ccCCGCCccg
tarsier	V$SP1_Q2_01	2	(-)	1.000	0.905	gggGGCGGcc
tarsier	V$SP1_Q2_01	10	(+)	1.000	0.969	ccCCGCCctg
Mouse	V$SP1_Q2_01	10	(+)	0.966	0.945	ccCCACCctg
Bushbaby	V$SP1_Q2_01	10	(+)	0.966	0.966	ccCCACCcca
Guinea	V$SP1_Q2_01	7	(+)	0.966	0.952	ccCCACCctc
squirrel	V$SP1_Q2_01	10	(+)	0.966	0.930	tcCCACCccg
Kangaroo	V$SP1_Q2_01	7	(+)	1.000	0.938	cgCCGCCccg
Rabbit	V$SP1_Q2_01	10	(+)	1.000	0.969	ccCCGCCctg
Megabat	V$SP1_Q2_01	10	(+)	0.966	0.945	ccCCACCctg
Dog	V$SP1_Q2_01	10	(+)	0.966	0.969	ccCCACCccg
Cow	V$SP1_Q2_01	10	(+)	0.966	0.942	ccCCACCcta
Dolphin	V$SP1_Q2_01	10	(+)	0.966	0.952	ccCCACCctc
Elephant	V$SP1_Q2_01	10	(+)	0.966	0.966	ccCCACCcca
rock	V$SP1_Q2_01	10	(+)	0.966	0.966	ccCCACCcca
Tenrec	V$SP1_Q2_01	10	(+)	0.966	0.966	ccCCACCcca
Opossum	V$SP1_Q2_01	9	(+)	0.966	0.935	gcCCACCccc
Human	V$SP1SP3_Q4	9	(+)	0.971	0.903	gccCCACCctg
Chimp	V$SP1SP3_Q4	9	(+)	0.971	0.903	gccCCACCctg
Gorilla	V$SP1SP3_Q4	9	(+)	0.971	0.903	gccCCACCctg

In the display options of the graphical overview visualization of species and transcription factor binding sites can be switched off and on again.