ChroMoS Guide (version 1.2)

Background

Genome-wide association studies (GWAS) reveal increasing number of disease-associated SNPs. Since majority of these SNPs are located in intergenic and intronic regions the assessment of their functionality was hindered by the lack of information about regulatory regions. It requires SNP prioritization for initial analysis to be followed by more focused functional analysis.

ChroMoS (Chromatin Modified SNPs) combines genetic and epigenetic data with the goal to facilitate SNP classification and prioritization. To this end the user can provide SNP data in VCF format, dbSNPs or select GWAS SNPs from the local database. The user provides annotations for chromatin state regions obtained from pre-calculated segmentation of epigenomic data for ENCODE 9 cell types. The genome segmentation based on chromatin marks allows predictions of functional elements, such as enhancers and promoters. In fact, six major categories of chromatin states were distinguished: enhancer, insulator, transcribed, repressed and inactive states. Promoter category was further partitioned into 3 states: active, weak and poised based on the expression level of adjacent genes; enhancer class was segregated into strong and weak states. Transcribed regions were separated into strongly and weakly transcribed regions. Also, heterochromatic and repetitive states were isolated based on their H3K9me3 enrichment. Polycomb-repressed regions were defined as well. In total, 15 states were distinguished and this data has been used in ChroMoS. It was shown that disease-associated SNPs were more likely to be situated within strong enhancer regions than neutral dbSNPs. Particularly, it was evident for cell types related to a disease, e.g. lymphoblastoid cell (GM12878) enhancers contained SNPs associated with systemic lupus erythematosus [Ernst et al. (2011), *Nature*].

Based on this data ChroMoS suggests the functional impact of a SNP. In the process, SNPs are assigned to the various chromatin states. The chromatin states were computed applying multivariate hidden Markov model [Ernst et al. (2011), *Nature*]. It uses patterns of chromatin marks to reduce large combinatorial space to an interpretable set of chromatin states. SNPs positioned in enhancer or transcription states can be subjected to differential analysis of transcription factor binding with sTRAP, and SNPs with potential impact on post-transcriptional

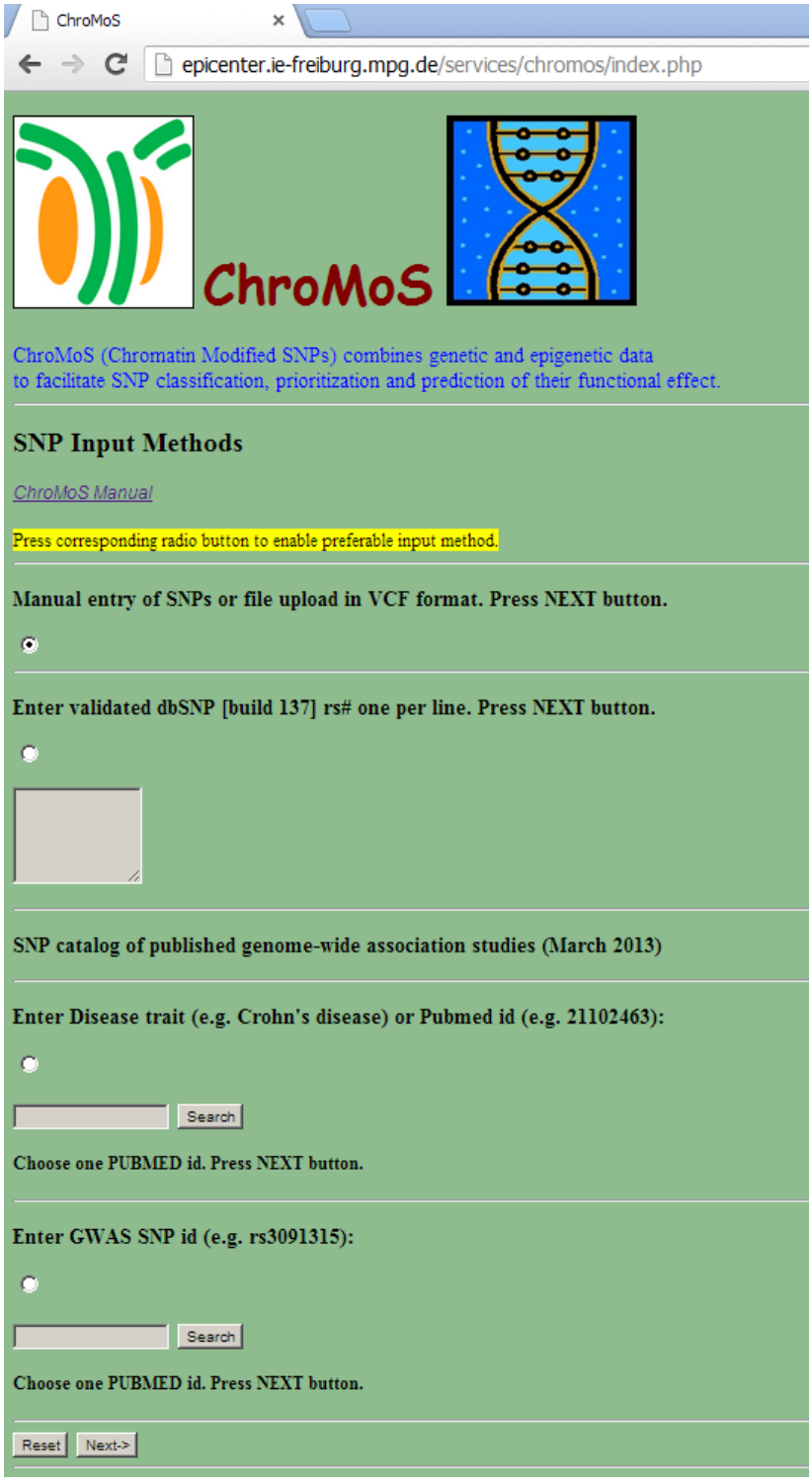
mechanisms are evaluated by MicroSNiPer for a differential binding capacity of annotated miRNA.

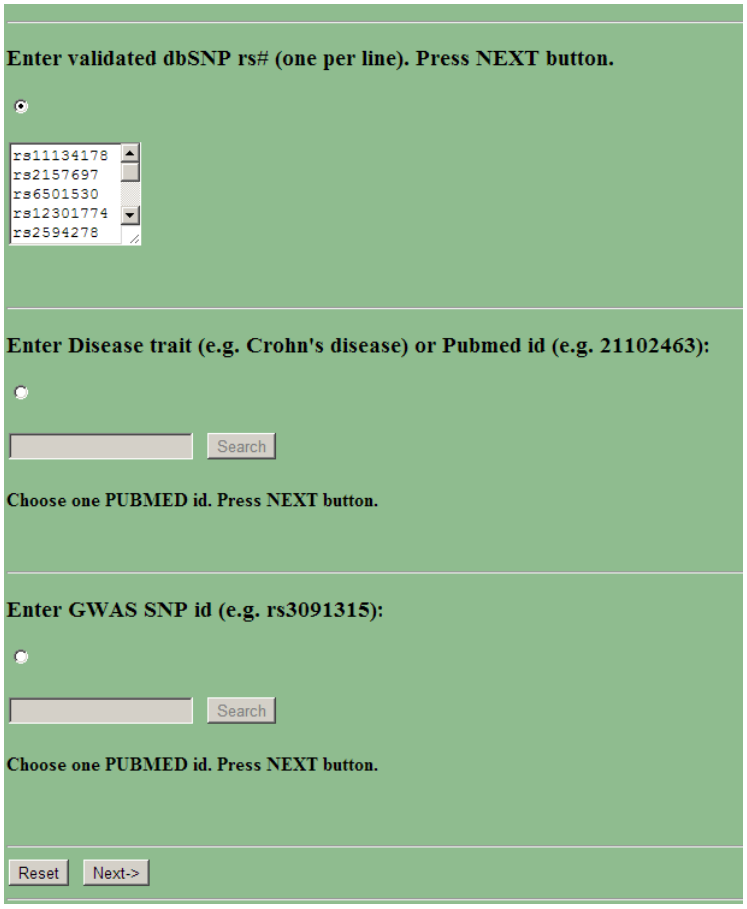
sTRAP, analyzes variations in the DNA sequence and predicts quantitative changes to the binding strength of any transcription factor for which there is a binding model. It suggests possible consequences of sequence variations on regulatory networks. The method was tested against a set of known associations between SNPs and their regulatory effects. Its predictions are robust with respect to different parameters and model assumptions. This tool can serve as important point for routine analysis of disease-associated sequence regions [Manke et al. (2010) Hum Mutat].

MicroSNiPer predicts the impact of a SNP on putative microRNA targets. This application interrogates the 3'-untranslated region and predicts if a SNP within the target site will disrupt/eliminate or enhance/create a microRNA binding site. MicroSNiPer computes these sites and examines the effects of SNPs in real time. It has straightforward graphical representation of the results [Barenboim et al. (2010) Hum Mutat].

ChroMoS Manual

Warning: Firefox web-browser might not display properly a color map of more than 1000 SNPs. Download the map through the web-link.

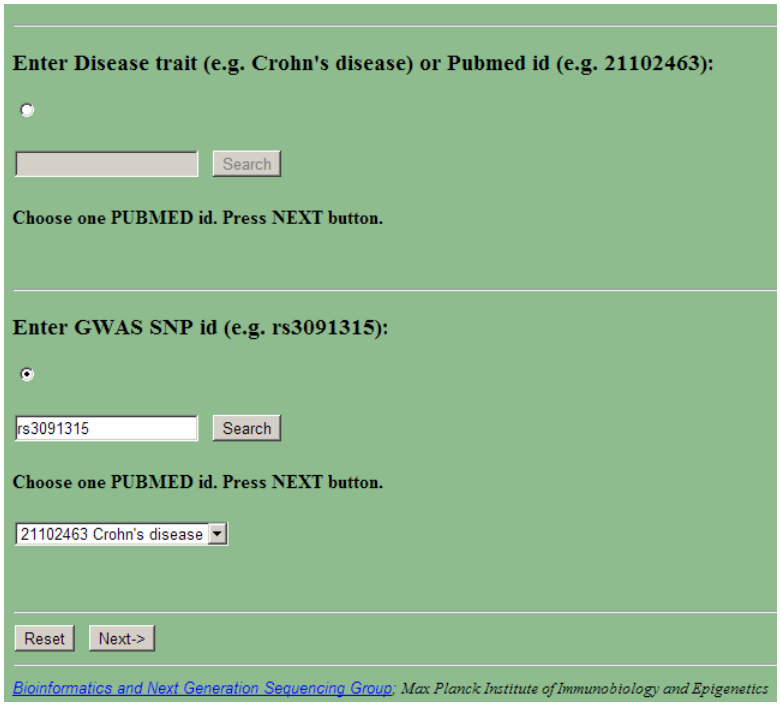




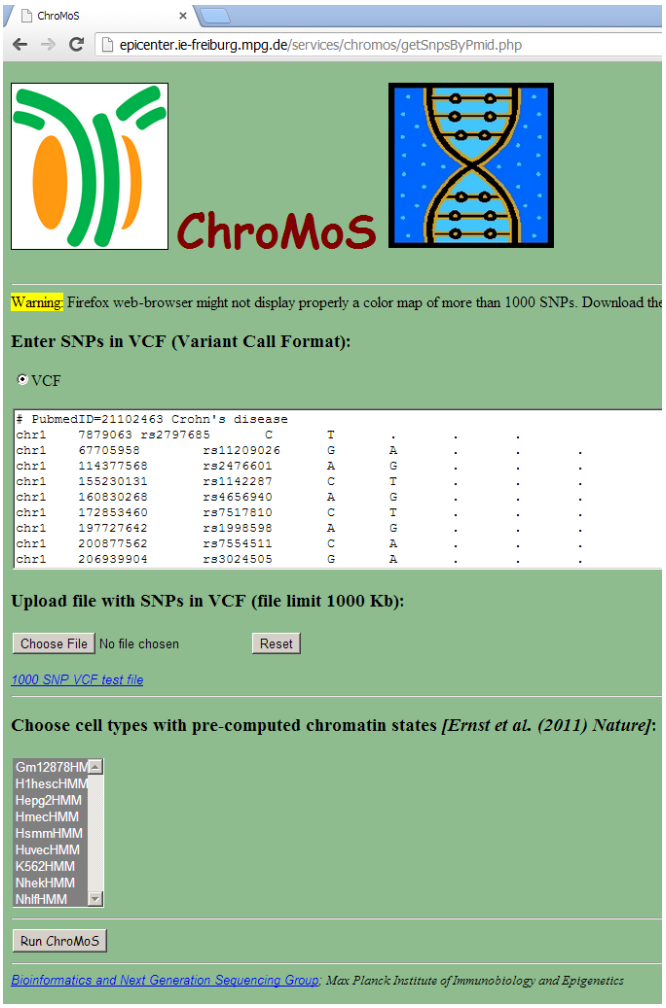
(2) Entry of validated dbSNP rs# (~45 mln dbSNPs). One rs# per line. It can be any dbSNP not necessarily from GWAS catalog. After pressing **Next** button these SNPs appear in the second page SNP area in VCF.

(3) Entry by disease trait (e.g. Crohn's disease) or Pubmed id (e.g. 21102463). Click second from the top radio button. Enter **Crohn's disease**. Click **Search** button. It retrieves a list of all currently published Crohn's disease GWAS studies including unique Pubmed IDs.

Choose a certain Pubmed id e.g. **21102463**. Press **Next** button. ChroMoS retrieves all 71 SNPs belonging to GWA study with PMID 21102463 and displays them on the next page.



(4) A user can also retrieve PMID by entering SNP id (e.g. rs3091315) and after pressing **Search** button choose proper PMID. Press **Next** button. Chromos retrieves all 71 SNPs belonging to GWA study with PMID: 21102463 and displays them on the next page.



On the second page 71 SNPs from GWA study with PMID 21102463 are displayed in VCF. On this stage user can add her own data by entering her data in the same format. If a user wants to upload only her own SNP file in VCF she can use **Choose File button**. In this case all data in VCF text area are erased. Pressing **Reset button** will recover original data. We provide a test file of 1,000 SNPs in VCF. It can be pasted to VCF area or uploaded as a VCF file directly from the local computer.

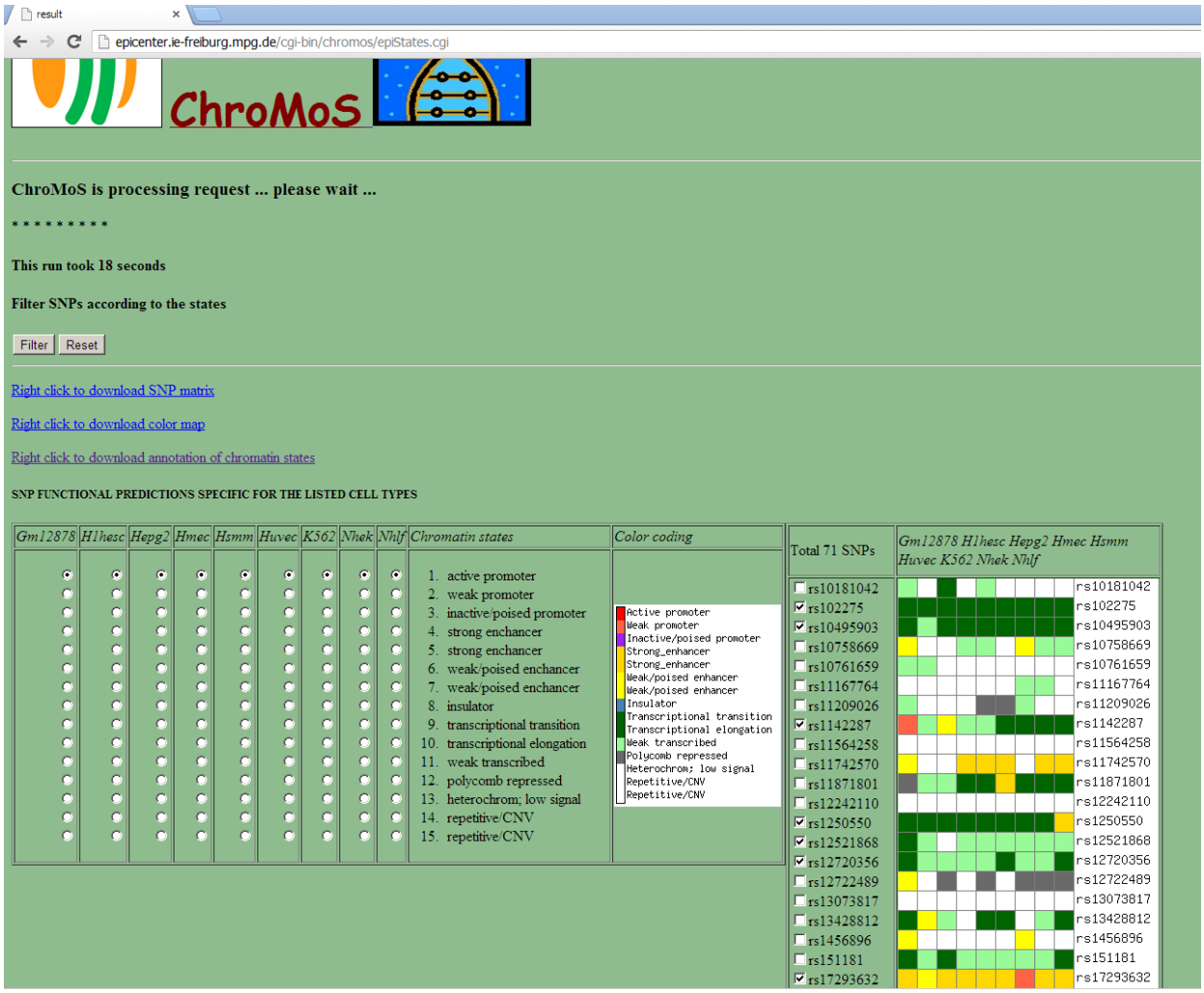
Important: one SNP record has to be in one continuous line. If this is not a case, text field should be stretched by grabbing lower right corner of the VCF text area.

A user can select one or more available cell types with pre-computed chromatin states in bed-format [Ernst et al. (2011), *Nature*] by Ctrl-Click and press **Run Chromos button.** This invokes Perl CGI script which utilizes bedtools [Quinlan and Hall (2010), *Bioinformatics*] intersecting SNP coordinates with coordinates of chromatin states and, subsequently, matrix2png [Pavlidis and Noble (2003), *Bioinformatics*], which provides color map of 15 states for each cell type.

On *ChroMoS* result page a user can also download digital matrix based on which color map is created and use in other tools. Table includes color map with SNP id aligned to color code of chromatin states. Column names display a number of SNPs and chosen cell types. Warning: Firefox web-browser has some limitation on displaying large PNG files (above ~ 1,000 SNPs) and alignment for large files is not exact, too. Opera web-browser has also graphical limitations.

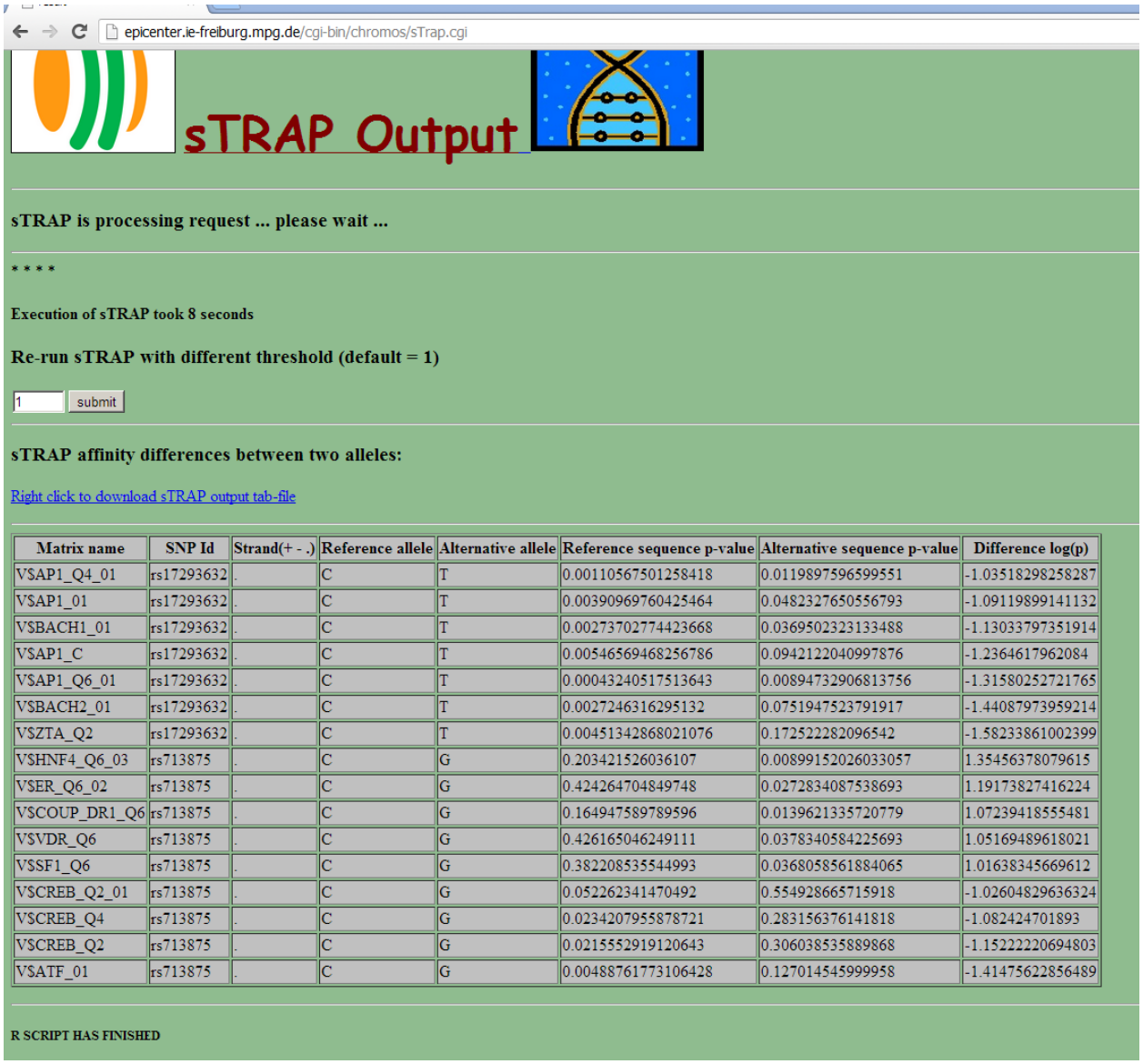
Next, a user should decide which way she prefers to filter results. One option is to use radio buttons in order to create certain pattern of states, e.g. "active promoter" in all 9 cell types. It is helpful for large SNP sets with only several cell types, or else, this type of selection likely produces empty set. Currently, the limitation for upload is 10,000 SNPs. If the SNP set consists of only several hundred SNPs, we suggest visually examining color map and manually checking out SNPs of interests (e.g. SNPs in the enhancer state in all 9 cell types).

If user starts manually checking out SNPs, pattern filtering is disabled. In order to return to pattern filtering and clear checkboxes user has to press **Reset button**. In this example 11 SNPs were checked out, and then **Filter button** was pressed.



On the next page filtered SNPs with color code are displayed. Then, in order to test if SNPs affect transcription factor binding a user can send SNPs to sTRAP [Manke et al., (2010) Hum Mutat.] selecting SNPs and pressing **Submit button.** Since sTRAP is computationally intensive, there is a limit of 60 SNPs to submit to sTRAP. Initial threshold is equal to one which displays

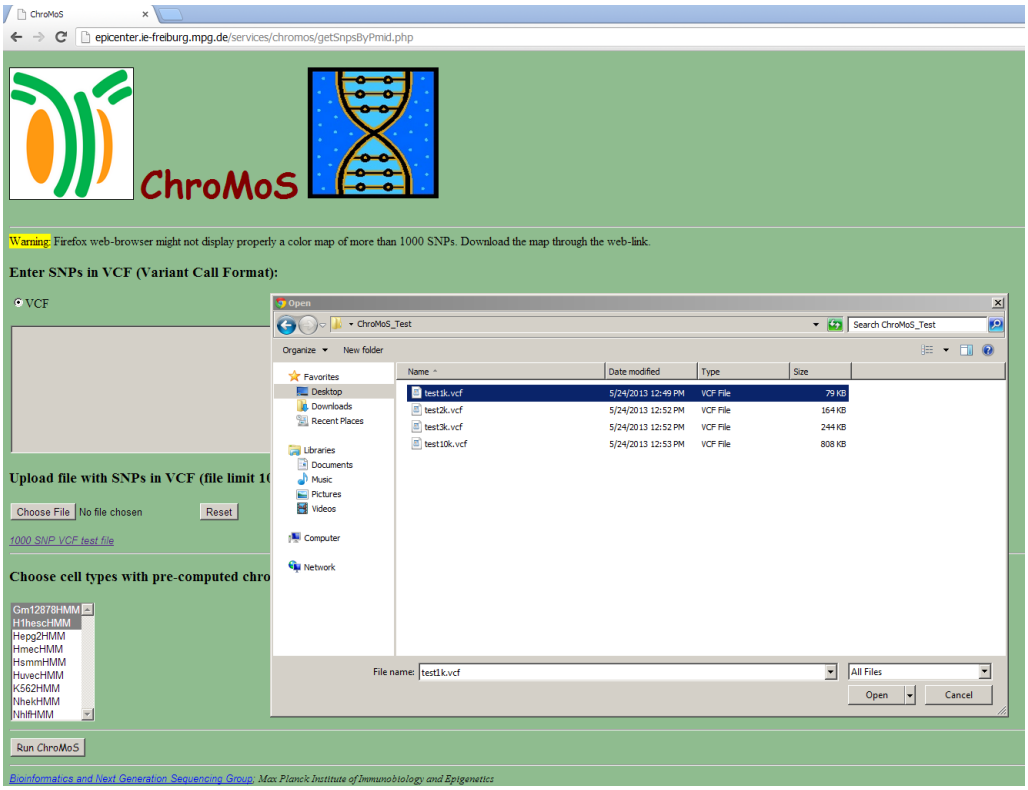
only significant candidate SNPs for impact on transcription factor binding sites. However, if there is an empty result table a user can decrease threshold (e.g. 0.6) and re-run sTRAP.

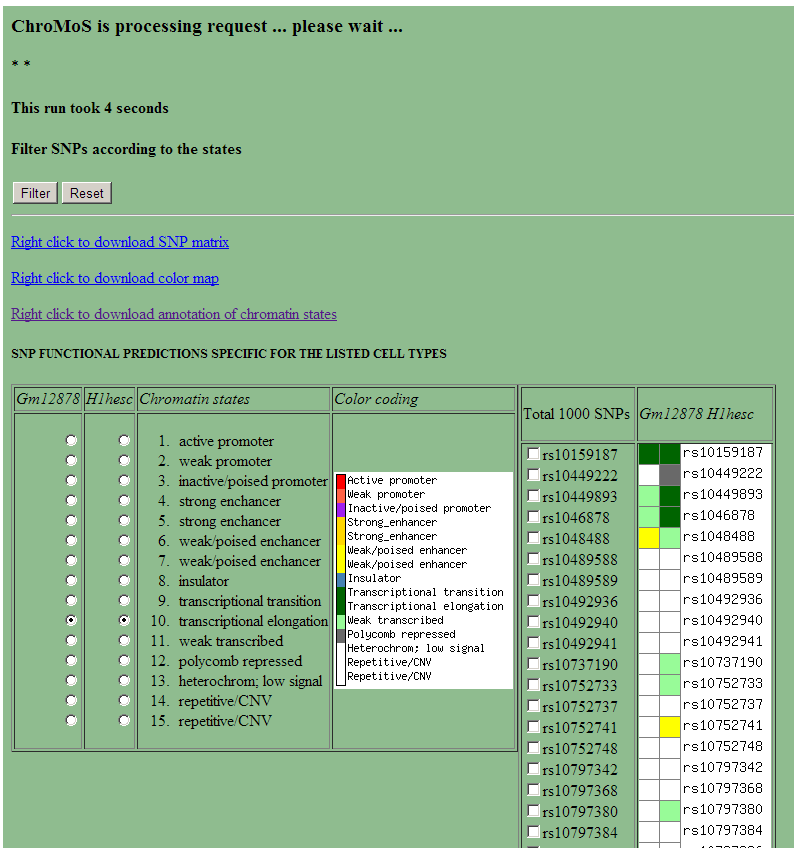


Matrix name	SNP Id	Strand(+ - .)	Reference allele	Alternative allele	Reference sequence p-value	Alternative sequence p-value	Difference log(p)
V$AP1_Q4_01	rs17293632	.	C	T	0.00110567501258418	0.0119897596599551	-1.03518298258287
V$AP1_01	rs17293632	.	C	T	0.00390969760425464	0.0482327650556793	-1.09119899141132
V$BACH1_01	rs17293632	.	C	T	0.00273702774423668	0.0369502323133488	-1.13033797351914
V$AP1_C	rs17293632	.	C	T	0.00546569468256786	0.0942122040997876	-1.2364617962084
V$AP1_Q6_01	rs17293632	.	C	T	0.00043240517513643	0.00894732906813756	-1.31580252721765
V$BACH2_01	rs17293632	.	C	T	0.0027246316295132	0.0751947523791917	-1.44087973959214
V$ZTA_Q2	rs17293632	.	C	T	0.00451342868021076	0.172522282096542	-1.58233861002399
V$HNF4_Q6_03	rs713875	.	C	G	0.203421526036107	0.00899152026033057	1.35456378079615
V$ER_Q6_02	rs713875	.	C	G	0.424264704849748	0.0272834087538693	1.19173827416224
V$COUP_DR1_Q6	rs713875	.	C	G	0.164947589789596	0.0139621335720779	1.07239418555481
V$VDR_Q6	rs713875	.	C	G	0.426165046249111	0.0378340584225693	1.05169489618021
V$SF1_Q6	rs713875	.	C	G	0.382208535544993	0.0368058561884065	1.01638345669612
V$CREB_Q2_01	rs713875	.	C	G	0.052262341470492	0.554928665715918	-1.02604829636324
V$CREB_Q4	rs713875	.	C	G	0.0234207955878721	0.283156376141818	-1.082424701893
V$CREB_Q2	rs713875	.	C	G	0.0215552919120643	0.306038535889868	-1.15222220694803
V$ATF_01	rs713875	.	C	G	0.00488761773106428	0.127014545999958	-1.41475622856489

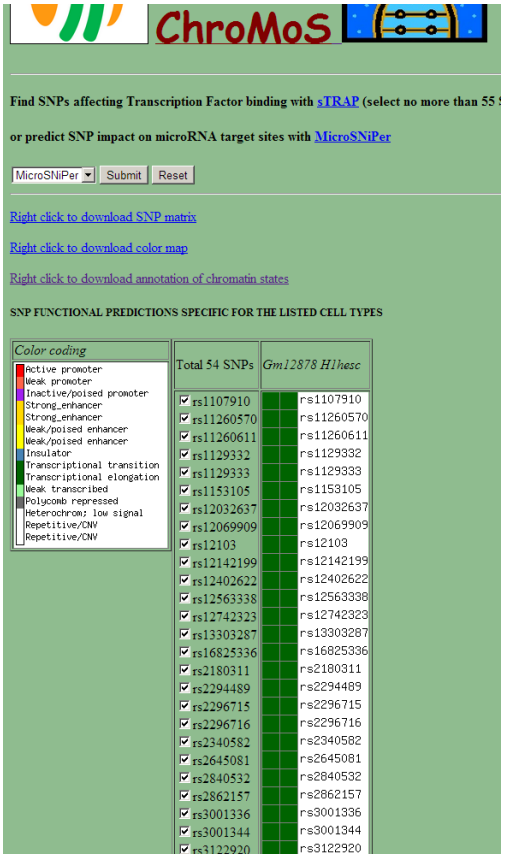
The sTRAP result page will display transfac matrix names grouped by SNPs. The transcription factors with reduced affinity receive a negative ratio of p-values and those with increased binding get a positive ratio. On the sTRAP result page user can re-run sTRAP with a different threshold. On each step a user can download data in tab-format.

To demonstrate integration with MicroSNiPer [Barenboim *et al.* (2010) *Hum Mutat]*, we download 1,000 SNPs sample file with **Choose File button**. We select two cell types GM12878 and H1hesc and press **Run ChroMoS button**.

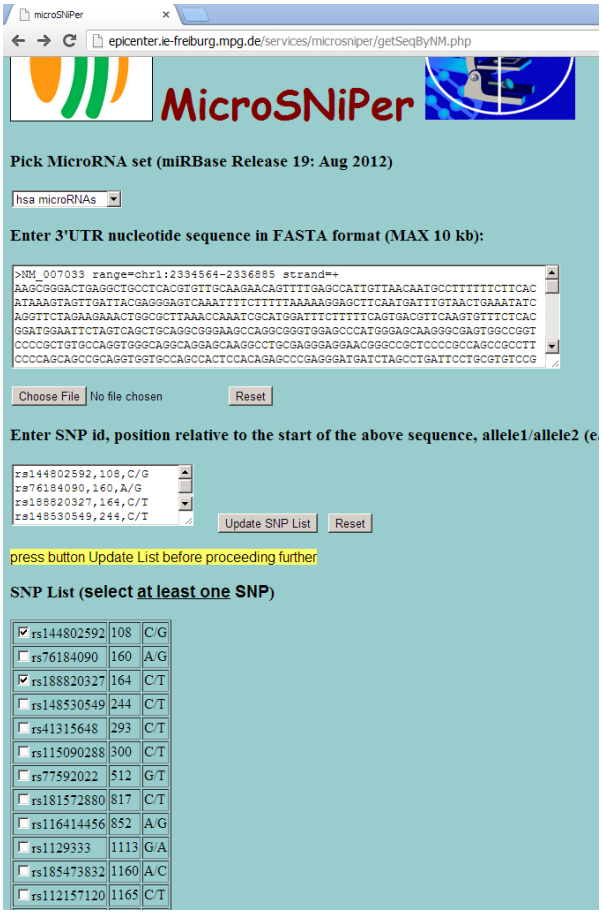


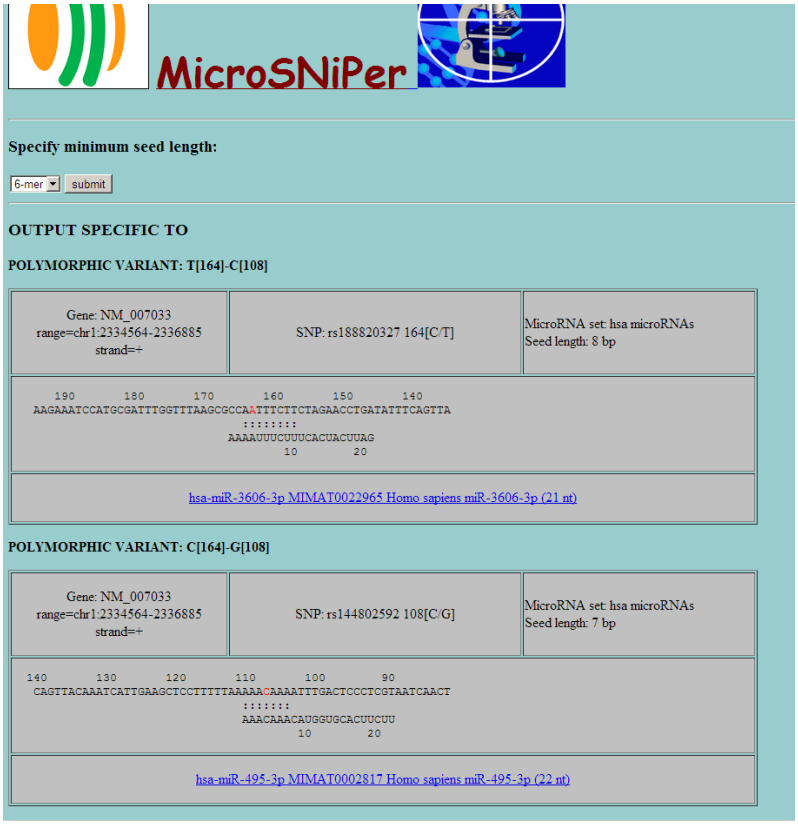


On ChroMoS result page we choose out of 1,000 SNPs all SNPs which are in *transcriptional elongation* state by pressing **radio button pattern filtering**. Pressing **Filter button** will bring another page.



On the filter result page there are 54 SNPs which are in *transcriptional elongation* state in both cell types. There is a possibility that some of them are in 3'UTR and can have an impact on microRNA target sites. In order to send these SNPs to integrated tool Microsniper a user has to choose MicroSNiPer from a menu on the top of the page. All SNPs will be automatically checked out. By pressing **Submit button** user send them to ***ChroMoS to MicroSNiPer*** page. On this stage a user can also add her SNPs in suggested format. Then, user tests if some of these SNPs are in 3'UTRs of RefSeq genes by pressing **Find SNPs in 3'UTRs button.**





Program filters SNPs for presenting in 3'UTRs and creates a table with radio buttons. User has to choose a single SNP from the table, and subsequently a transcript NM_id from the dropdown list. User also can choose validated dbSNPs (default) or a set of HapMap SNPs on the top of the page. Pressing **Next button** inputs this data to a routine MicroSNiPer workflow. A SNP selected with radio button is added to the list of validated dbSNPs (or HapMap SNPs) positioned within chosen 3'UTR. On MicroSNiPer page a user can also add her own SNPs. Then, user presses **Update SNP List button**, check out SNPs of interest (limit 6 SNPs) and presses **Run Microsniper button.** User can also go directly to MicroSNiPer main page on the http://epicenter.ie-freiburg.mpg.de/services/microsniper/.

REFERENCES

Barenboim, M. et al. (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets, Hum Mutat, 31, 1223-1232.

Ernst, J. et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types, Nature, 473, 43-49.

Franke, A. et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci, Nat Genet, 42, 1118-1125.

Manke, T. et al. (2010) Quantifying the effect of sequence variation on regulatory interactions, Hum Mutat, 31, 477-483.

Pavlidis, P. and Noble, W.S. (2003) Matrix2png: a utility for visualizing matrix data, Bioinformatics, 19, 295-296.

Pruitt, K.D. et al. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res, 35, D61-65.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 26, 841-842.