

# Тестовое задание

Игнат Сонец, ИБГ РАН

[https://github.com/ISonets/ChIP-Seq\\_results\\_v2](https://github.com/ISonets/ChIP-Seq_results_v2)

# Задание

1. Скачать chip-seq для транскрипционного фактора YY1 в виде bed file формата broad или narrow peaks из базы ENCODE.
2. Отобрать Top1000 лучших пиков на основе  $-\log pvalue$ .
3. Для отобранных top1000 пиков выполнить de novo motif discovery используя Homer.
4. Представить визуально полученные результаты для найденных мотивов на основе Homer (можно использовать графические материалы, полученные в результате работы софта Homer). Написать аналитический комментарий с биологической интерпретацией, опираясь на данные литературы и веб сайта Factorbook.
5. Выполнить GO enrichment для генов, полученных посредством ассигнования пик-ген, используя тулу GREAT (<http://great.stanford.edu/public/html/>) и посредством ассигнования пика к ближайшему гену с учетом  $\pm$  цепи (например с помощью bedtools)

Необходимо выслать код и графические материалы по задачам в виде jupyter notebook или в виде презентации.

# 1. Скачивание файлов

Я выбрал YY1 *Homo sapiens* для дальнейшей работы.

Всего в ENCODE выложено 18 экспериментов  
(<https://www.encodeproject.org/genes/7528/>)

Genes / *Homo sapiens*

YY1 (*Homo sapiens*)



Entrez GeneID: [7528](#)

Gene symbol: YY1

Official gene name: YY1 transcription factor

Synonyms: DELTA  
GADEVS  
INO80S  
NF-E1  
UCRBP  
YIN-YANG-1

Gene locations: GRCh38 chr14:100239144-100282788,  
hg19 chr14:100705481-100749125

External resources: [RefSeq:NM\\_003403.5](#)  
[MIM:600013](#)  
[Vega:OTTHUMG00000150479](#)  
[UniProtKB:P25490](#)  
[HGNC:12856](#)  
[ENSEMBL:ENSG00000100811](#)  
[GeneCards:YY1](#)

# 1. Скачивание файлов

Я выбрал 1 из экспериментов (sample GM12892):

## TF ChIP-seq of GM12892

*Homo sapiens* GM12892

**Target:** YY1 ([Factorbook](#))

**Lab:** Richard Myers, HAIB

**Project:** ENCODE



Experiment 







ENCSR000BLT

● released

● 3

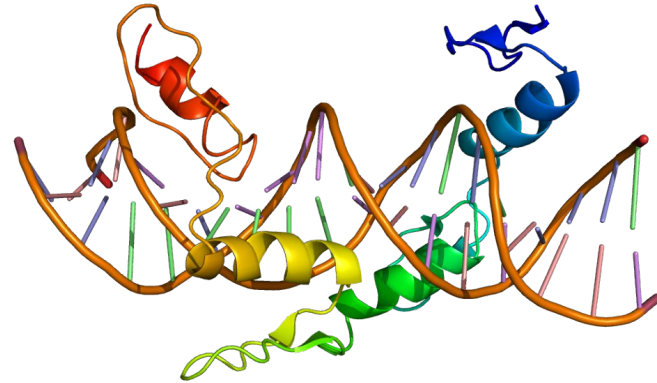
Для дальнейшей работы я взял bed narrowPeak file:

ENCODE4 v1.6.1 GRCh38 (ENCAN010QYV) processed data (22 Files)  ● released  2

Accession	Default	File type	Output type	Isogenic replicate	Mapped read length	Genome assembly	Date added	File size	File status
ENCFF329MHR  	★	bed narrowPeak	IDR thresholded peaks	1, 2		GRCh38	2020-12-28	336 kB	● released
ENCFF993VSK  	★	bigWig	signal p-value	1, 2		GRCh38	2020-12-28	683 MB	● released
ENCFF802MHJ  	★	bigBed narrowPeak	IDR thresholded peaks	1, 2		GRCh38	2020-12-28	622 kB	● released

# Что за YY1?

- повсеместно распространенный ТФ, принадлежит к семейству zinc finger proteins;
- способен к активации и репрессии различных промоторов(имя происходит от инь-ян);
- взаимодействует с гистонтрансферазами и гистондеацетилазами;
- Гетерозиготные делеции YY1, нонсенс- и миссенс-мутации вызывают синдром GADEVS (АД заболевание)



Более подробно здесь: <https://en.wikipedia.org/wiki/YY1>  
и здесь: <https://www.uniprot.org/uniprot/P25490>.

## 2. Подготовка данных

Структура данных (<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>):

### ENCODE narrowPeak: Narrow (or Point-Source) Peaks format

This format is used to provide called peaks of signal enrichment based on pooled, normalized (interpreted) data. It is a BED6+4 format.

1. **chrom** - Name of the chromosome (or contig, scaffold, etc.).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0*, *chromEnd=100*, and span the bases numbered 0-99.
4. **name** - Name given to a region (preferably unique). Use "." if no name is assigned.
5. **score** - Indicates how dark the peak will be displayed in the browser (0-1000). If all scores were "0" when the data were submitted to the DCC, the DCC assigned scores 1-1000 based on signal value. Ideally the average signalValue per base spread is between 100-1000.
6. **strand** - +/- to denote strand or orientation (whenever applicable). Use "." if no orientation is assigned.
7. **signalValue** - Measurement of overall (usually, average) enrichment for the region.
8. **pValue** - Measurement of statistical significance (-log10). Use -1 if no pValue is assigned.
9. **qValue** - Measurement of statistical significance using false discovery rate (-log10). Use -1 if no qValue is assigned.
10. **peak** - Point-source called for this peak; 0-based offset from chromStart. Use -1 if no point-source called.

Here is an example of narrowPeak format:

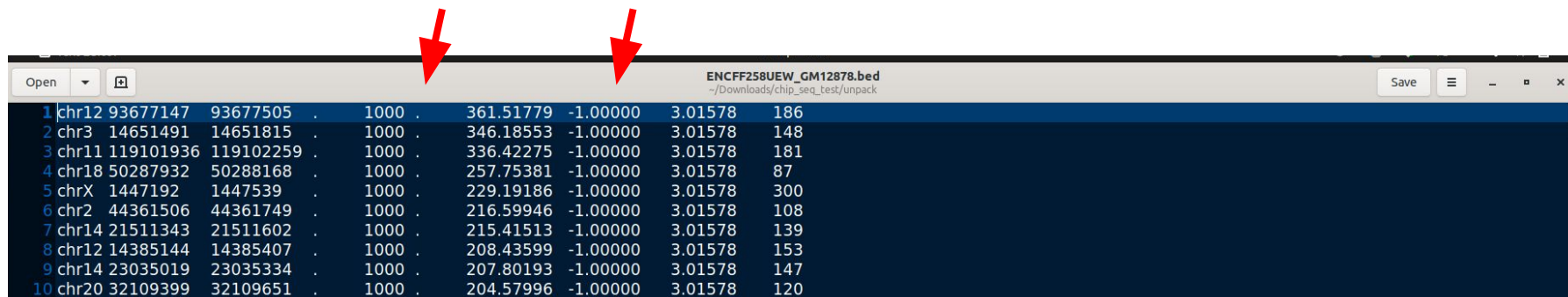
```
track type=narrowPeak visibility=3 db=hg19 name="nPk" description="ENCODE narrowPeak Example"
browser position chr1:9356000-9365000
chr1 9356548 9356648 . 0 . 182 5.0945 -1 50
chr1 9358722 9358822 . 0 . 91 4.6052 -1 40
chr1 9361082 9361182 . 0 . 182 9.2103 -1 75
```

## 2. Подготовка данных

!Есть 2 сложности:

1. pValue не указано (-1), что делает невозможным выбор топ 1000 пиков. Выход: использовать qValue
2. strand не указан(в силу предыдущей обработки данных), что делает невозможным peak assigning с учетом направления цепи. Выход: обойтись без этого параметра.

Среди всех 18 экспериментов отсутствует pValue и strand. Может быть это правила обработки данных перед загрузкой на ENCODE? Или стоило повторить анализ целиком, от сырых .fasta?



The screenshot shows a text editor window titled "ENCF258UEW\_GM12878.bed" with a file path of "~/Downloads/chip\_seq\_test/unpack". The editor displays a table of genomic data with 8 columns. Two red arrows point to the 6th and 7th columns, which contain the values -1.00000 and 3.01578 respectively for the first row.

Line	chr	start	end	score	strand	value	count
1	chr12	93677147	93677505	1000	.	-1.00000	3.01578 186
2	chr3	14651491	14651815	1000	.	-1.00000	3.01578 148
3	chr11	119101936	119102259	1000	.	-1.00000	3.01578 181
4	chr18	50287932	50288168	1000	.	-1.00000	3.01578 87
5	chrX	1447192	1447539	1000	.	-1.00000	3.01578 300
6	chr2	44361506	44361749	1000	.	-1.00000	3.01578 108
7	chr14	21511343	21511602	1000	.	-1.00000	3.01578 139
8	chr12	14385144	14385407	1000	.	-1.00000	3.01578 153
9	chr14	23035019	23035334	1000	.	-1.00000	3.01578 147
10	chr20	32109399	32109651	1000	.	-1.00000	3.01578 120

## 2. Подготовка данных

В Jupyter Notebook, используя pandas, я отобрал топ 1000 пиков на основе qValue.

!Чем больше значение qValue, тем более значим этот пик (в .bed qValue в  $-\log(10)$  форме). Также отбирал пики на основе signalValue и score (при условии равенства qValue).

In [4]: bed\_Np\_orig

Out[4]:

	chrom	chromStart	chromEnd	name	score	strand	signalValue	pValue	qValue	peak
0	chr6	33789021	33789449	.	1000	.	483.48822	-1.0	4.62705	217
1	chr20	10434942	10435377	.	1000	.	482.86921	-1.0	4.62705	233
2	chr19	49817916	49818380	.	1000	.	474.64664	-1.0	4.62705	220
3	chr1	110034384	110034814	.	1000	.	473.88893	-1.0	4.62705	163
4	chr16	87392080	87392393	.	1000	.	473.05367	-1.0	4.62705	151
...	...	...	...	...	...	...	...	...	...	...
20791	chr10	114096482	114096798	.	610	.	4.94941	-1.0	0.15952	158
20792	chr1	23168119	23168435	.	562	.	4.94706	-1.0	0.15941	158
20793	chr19	19643547	19643863	.	775	.	4.93649	-1.0	0.15790	158
20794	chr20	33663756	33664072	.	584	.	4.92921	-1.0	0.15810	158
20795	chr1	226407239	226407555	.	687	.	4.92817	-1.0	0.15803	158

20796 rows × 10 columns

In [29]: bed\_Np\_qVal\_sorted\_top1000

Out[29]:

	chrom	chromStart	chromEnd	name	score	strand	signalValue	pValue	qValue	peak
0	chr6	33789021	33789449	.	1000	.	483.48822	-1.0	4.62705	217
1	chr20	10434942	10435377	.	1000	.	482.86921	-1.0	4.62705	233
2	chr19	49817916	49818380	.	1000	.	474.64664	-1.0	4.62705	220
3	chr1	110034384	110034814	.	1000	.	473.88893	-1.0	4.62705	163
4	chr16	87392080	87392393	.	1000	.	473.05367	-1.0	4.62705	151
...	...	...	...	...	...	...	...	...	...	...
995	chr19	37779469	37779720	.	1000	.	241.56200	-1.0	4.62705	129
996	chr5	160119233	160119510	.	1000	.	241.56193	-1.0	4.62705	144
997	chr12	113185303	113185565	.	1000	.	241.14477	-1.0	4.62705	122
998	chr20	45881132	45881385	.	1000	.	241.03950	-1.0	4.62705	120
999	chr6	31664869	31665115	.	1000	.	240.95224	-1.0	4.62705	144

1000 rows × 10 columns



### 3. *De novo* motif discovery (HOMER)

<http://homer.ucsd.edu/homer/>

Workflow:

1. HOMER был установлен в conda;
2. `configureHomer.pl -install hg38;`
3. `findMotifsGenome.pl top1000_peaks.bed hg38 de_novo_motifs_top/`

## Homer de novo Motif Results (de\_novo\_motifs\_top/)

Known Motif Enrichment Results

Gene Ontology Enrichment Results

If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into [STAMP](#)

More information on motif finding results: [HOMER](#) | [Description of Results](#) | [Tips](#)

Total target sequences = 1000

Total background sequences = 43231

\* possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bq STD)	Best Match/Details	Motif File
1		1e-876	-2.018e+03	71.20%	2.57%	32.5bp (68.0bp)	YY1/MA0095.2/Jaspar(0.976) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
2		1e-61	-1.413e+02	37.60%	15.86%	55.9bp (75.0bp)	ETV6/MA0645.1/Jaspar(0.940) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
3		1e-44	-1.023e+02	11.50%	2.26%	50.2bp (69.0bp)	TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer(0.960) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
4		1e-33	-7.805e+01	1.40%	0.00%	46.3bp (9.3bp)	NKX2-5/MA0063.2/Jaspar(0.809) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
5		1e-29	-6.724e+01	25.90%	12.61%	55.2bp (82.3bp)	BMYY(HTH)/Hela-BMYB-ChIP-Seq(GSE27030)/Homer(0.751) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
6		1e-24	-5.552e+01	38.20%	23.64%	56.5bp (78.9bp)	YY2/MA0748.2/Jaspar(0.635) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
7		1e-22	-5.282e+01	8.70%	2.44%	48.4bp (73.2bp)	THAP1/MA0597.1/Jaspar(0.827) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
8		1e-22	-5.131e+01	26.90%	14.84%	56.3bp (81.8bp)	CEBPB/MA0466.2/Jaspar(0.760) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
9		1e-19	-4.564e+01	4.90%	0.92%	50.7bp (79.1bp)	GFY(?) /Promoter/Homer(0.912) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
10		1e-14	-3.286e+01	1.10%	0.03%	40.0bp (84.0bp)	PB0203.1_Zfp691_2/Jaspar(0.617) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
11		1e-14	-3.249e+01	3.60%	0.71%	49.4bp (68.3bp)	SPIC/MA0687.1/Jaspar(0.632) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
12		1e-14	-3.225e+01	3.40%	0.64%	50.6bp (81.0bp)	PB0004.1_Atf1_1/Jaspar(0.881) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
13		1e-13	-3.215e+01	12.90%	6.24%	61.4bp (76.8bp)	NRF1(NRF)/MCF7-NRF1-ChIP-Seq(Unpublished)/Homer(0.861) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
14		1e-13	-3.104e+01	1.00%	0.02%	61.8bp (108.7bp)	OSR2/MA1646.1/Jaspar(0.654) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
15		1e-12	-2.866e+01	4.00%	0.99%	61.5bp (73.9bp)	GFY(?) /Promoter/Homer(0.889) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
16 *		1e-11	-2.729e+01	0.70%	0.01%	46.8bp (32.7bp)	ZNF652/HepG2-ZNF652-Flag-ChIP-Seq(Encode)/Homer(0.678) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
17 *		1e-11	-2.630e+01	3.00%	0.62%	55.9bp (66.4bp)	ZBTB7C/MA0695.1/Jaspar(0.771) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
18 *		1e-9	-2.144e+01	0.70%	0.02%	53.9bp (62.7bp)	Bapx1(Homeobox)/VertebralCol-Bapx1-ChIP-Seq(GSE36672)/Homer(0.571) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
19 *		1e-9	-2.131e+01	13.90%	8.12%	62.1bp (83.8bp)	NFIA/MA0670.1/Jaspar(0.761) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
20 *		1e-8	-1.925e+01	2.60%	0.64%	63.5bp (83.9bp)	SP1/MA0079.4/Jaspar(0.852) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
21 *		1e-7	-1.764e+01	0.90%	0.06%	48.8bp (59.7bp)	TEAD1(TEAD)/HepG2-TEAD1-ChIP-Seq(Encode)/Homer(0.618) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
22 *		1e-7	-1.699e+01	1.90%	0.40%	60.0bp (64.4bp)	RHOXF1/MA0719.1/Jaspar(0.605) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
23 *		1e-6	-1.413e+01	4.00%	1.67%	53.7bp (80.8bp)	MeT2a(MADS)/HL1-MeT2a.biotin-ChIP-Seq(GSE21529)/Homer(0.604) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
24 *		1e-4	-1.144e+01	0.50%	0.03%	51.0bp (40.2bp)	SMAD5/MA1557.1/Jaspar(0.759) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
25 *		1e-4	-1.105e+01	0.30%	0.01%	65.4bp (26.2bp)	E2F1(E2F)/Hela-E2F1-ChIP-Seq(GSE22478)/Homer(0.666) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>

## 3. De novo motif discovery (HOMER)

Всего найдено 25 мотивов, из них 10 HOMER считает возможно ложноположительными (может ли быть дело в стандартных параметрах запуска?). Motifs logo и последовательности доступны на Github.

### 3. *De novo* motif discovery – сравним с Factorbook



Среди de novo мотивов найден YY1/MA0095.2 => совпадение с опубликованными данными.

### 3. Known motifs discovery (HOMER)

HOMER также предоставил известные ему мотивы (79 штук):



#### Homer Known Motif Enrichment Results (de\_novo\_motifs\_top)

[Homer de novo Motif Results](#)

[Gene Ontology Enrichment Results](#)

[Known Motif Enrichment Results \(txt file\)](#)

Total Target Sequences = 1000, Total Background Sequences = 43206

Rank	Motif	Name	P-value	log P-value	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	SVG
1		YY1(Zf)/Promoter/Homer	1e-748	-1.724e+03	0.0000	627.0	62.70%	985.1	2.28%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
2		Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	1e-56	-1.294e+02	0.0000	327.0	32.70%	5696.0	13.16%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
3		Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	1e-54	-1.259e+02	0.0000	320.0	32.00%	5576.3	12.88%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
4		Flh1(ETS)/CD8-FL1-ChIP-Seq(GSE20898)/Homer	1e-52	-1.215e+02	0.0000	377.0	37.70%	7455.3	17.22%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
5		ELF1(ETS)/Jurkat-ELF1-ChIP-Seq(SRA014231)/Homer	1e-50	-1.167e+02	0.0000	292.0	29.20%	4992.6	11.53%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
6		ETS(ETS)/Promoter/Homer	1e-47	-1.095e+02	0.0000	215.0	21.50%	3067.0	7.09%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
7		ETV4(ETS)/HepG2-ETV4-ChIP-Seq(ENCODE)/Homer	1e-46	-1.059e+02	0.0000	378.0	37.80%	7983.3	18.44%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
8		ETV1(ETS)/GIST48-ETV1-ChIP-Seq(GSE22441)/Homer	1e-36	-8.388e+01	0.0000	341.0	34.10%	7525.3	17.38%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
9		GABPA(ETS)/Jurkat-GABPa-ChIP-Seq(GSE17954)/Homer	1e-36	-8.373e+01	0.0000	276.0	27.60%	5437.4	12.56%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
10		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer	1e-32	-7.417e+01	0.0000	273.0	27.30%	5658.4	13.07%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
11		TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer	1e-30	-6.916e+01	0.0000	67.0	6.70%	479.6	1.11%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>
12		Etv2(ETS)/ES-ER71-ChIP-Seq(GSE59402)/Homer	1e-28	-6.456e+01	0.0000	217.0	21.70%	4247.8	9.81%	<a href="#">motif file (matrix)</a>	<a href="#">svg</a>

## 4. GO enrichment (GREAT)

<http://great.stanford.edu/public/html/index.php>

Была произведена доп. обработка файла топ-1000 пиков:

1. `cut -f 1-6 top_1000_peaks.bed > top_1000_GREAT.bed`  
(.bed и .narrowPeak имеют 6 одинаковых колонок, но остальные различны; чтобы избежать ошибок, сохранил 1-6 колонки отдельно)
2. `cat top_1000_GREAT.bed | sed 1d > top_1000_GREAT_noheader.bed`  
(GREAT выдает ошибку при наличии заголовка)

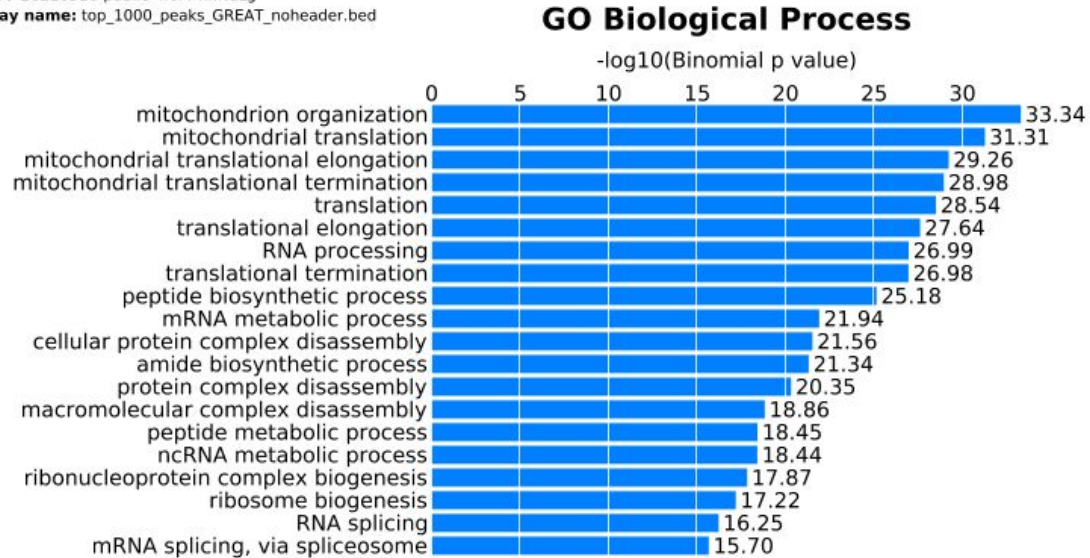
Параметры запуска:

- геном Human: GRCh38 (UCSC hg38, Dec. 2013)
- Background regions: whole genome
- Associating genomic regions with genes: Proximal: 5 kb upstream, 1kb downstream; Distal: up to 1000 kb (стандарт)

## 4. GO enrichment (GREAT)

Job ID: 20210916-public-4.0.4-hhNuLg

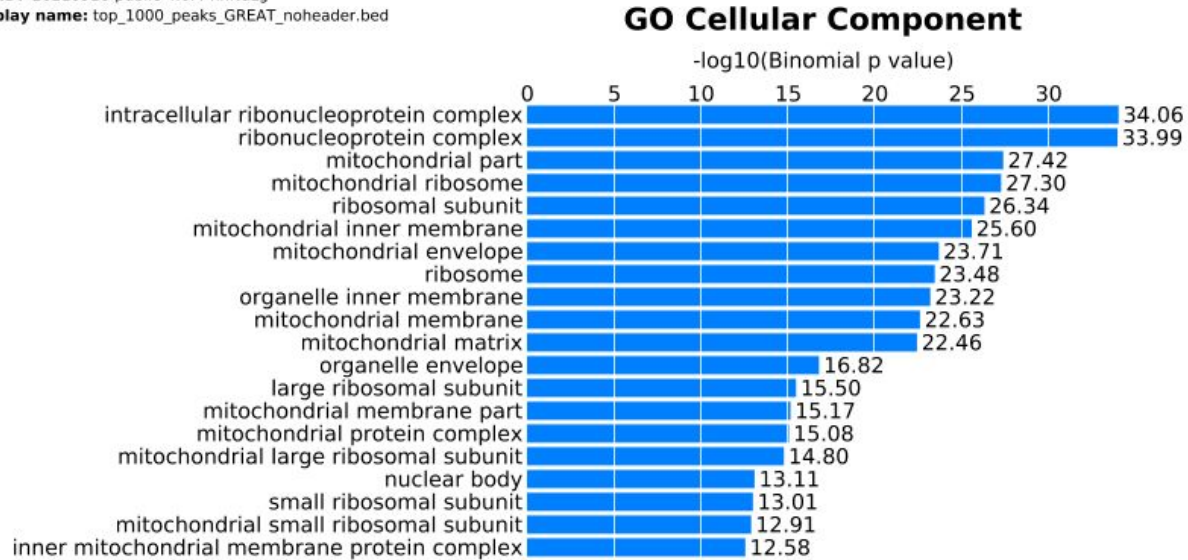
Display name: top\_1000\_peaks\_GREAT\_noheader.bed



## 4. GO enrichment (GREAT)

Job ID: 20210916-public-4.0.4-hhNuLg

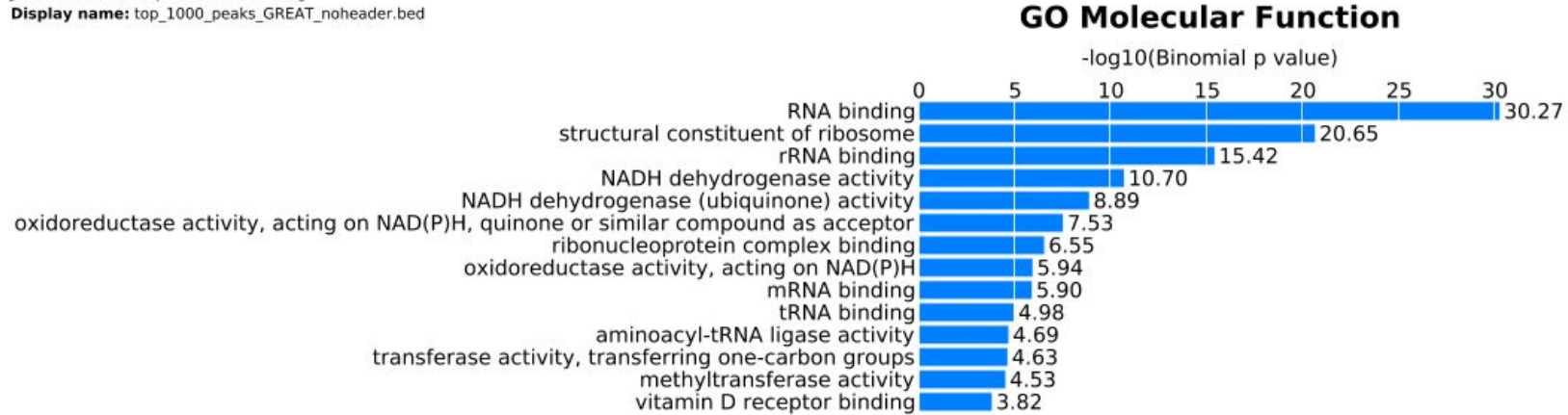
Display name: top\_1000\_peaks\_GREAT\_noheader.bed





## 4. GO enrichment (GREAT)

**Job ID:** 20210916-public-4.0.4-hhNuLg  
**Display name:** top\_1000\_peaks\_GREAT\_noheader.bed





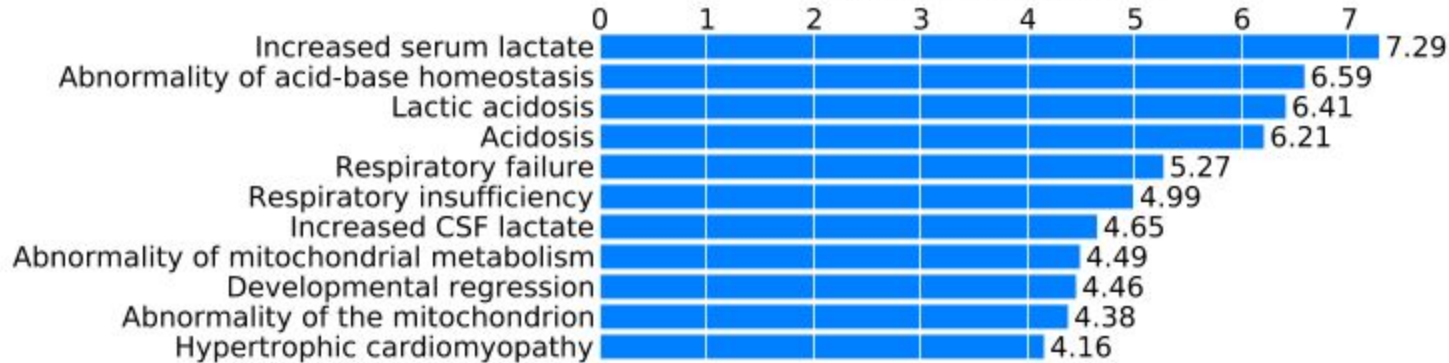
## 4. GO enrichment (GREAT)

Job ID: 20210916-public-4.0.4-hhNuLg

Display name: top\_1000\_peaks\_GREAT\_noheader.bed

### Human Phenotype

$-\log_{10}(\text{Binomial p value})$



## 4. GO enrichment – bedtools

<https://bedtools.readthedocs.io/en/latest/index.html>

Т.к. в .bed не указан strand, назначение пика к гену м.б. затруднительным в плане интерпретации(нет смысла указывать флаг -s(учитывать цепь в процессе)).

Workflow:

1. bedtools установлен в conda;
2. С GENCODE был загружен.gff3 файлы для сборки hg38 (<https://www.gencodegenes.org/human/> );
3. bedtools sort top\_1000\_GREAT\_noheader.bed/bedtools sort hg38\_full.gff3;
4. bedtools closest -a top\_1000\_GREAT\_sorted.bed -b hg38\_full\_sorted.gff3 > top\_1000\_closest.txt  
(находим ближайшие гены);
5. grep "ID=exon" top\_1000\_closest.txt > top1000\_exons\_only.txt (извлекаем экзоны);
6. cut -f 15 top1000\_exons\_only.txt | cut -d';' -f6 | sort | uniq | cut -d=' ' -f2 > top1000\_GO\_ready.txt  
(достаем имена генов)

## 4. GO enrichment – bedtools

Для GO enrichment был использован веб-сервис <http://geneontology.org/> и <http://pantherdb.org/> .

Параметры:

- референс – Homo sapiens (all genes in database);
- датасет – GO biological process complete;
- тест – Fisher exact;
- коррекция – FDR.

Всего обнаружено обогащение по 327 биологическим процессам.

## 4. GO enrichment – HOMER

Для верификации результатов я решил попробовать сделать GO enrichment с помощью аннотации пиков HOMER.

Workflow:

1. `annotatePeaks.pl top_1000_peaks.bed hg38> HOMER_annotated_peaks.txt`
2. `cut -f 16 HOMER_annotated_peaks.txt | sort | uniq | wc -l` (862 аннотированных гена)
3. `cut -f 16 HOMER_annotated_peaks.txt | sort | uniq > HOMER_gene_names.txt`(получить гены списком)
4. Воспользоваться <http://geneontology.org/> и <http://pantherdb.org/> для GO enrichment.  
Параметры аналогичны использованным для GO enrichment с bedtools.

Всего обнаружено обогащение по 326 процессам. Множество совпадений с результатами анализа генов, обнаруженных bedtools, а также с результатами работы GREAT.

# GO enrichment – интересные находки

- + регулирует трансляцию в митохондриях;
- + регулирует транскрипцию мРНК и ее сплайсинг;
- способствует сборке большой субъединицы рибосомы;
- участвует в встраивании белков в мембрану митохондрий;
- + регулирует инициацию транскрипции РНК-полимеразой 2;
- участвует в катаболизме мРНК;
- участвует в дифференцировке нейронов, клеточной миграции и морфогенезе;
- участвует в регуляции врожденного иммунного ответа;
- участвует в регуляции В-клеточного иммунитета.

# Итого

1. Выделено топ-1000 пиков для образца GM12892 *Homo sapiens* из исходных файлов, несмотря на выбор другой метрики.
2. Выполнен de novo motif search с HOMER.
3. Выполнен GO enrichment с GREAT и bedtools.

Код и результаты доступны на [Github](#).

# Тестовое задание

Игнат Сонец, ИБГ РАН

[https://github.com/ISonets/ChIP-Seq\\_results\\_v2](https://github.com/ISonets/ChIP-Seq_results_v2)