Name: Andrew Peters
Student ID: 17232658
Module: Genomics Data Analysis

Assignment: Collaborators working on the Irish Cancer Society BreastPredict project have asked you to analyze data from a ChIP-seq experiment looking at the regulatory role of a specific transcription factor in a human breast cancer cell line. They have asked you to provide a list of genome-wide binding locations, information on any potential regulatory role, and to conduct a de-novo motif dicovery analysis to confirm the binding motif, comparing it to known motif databases.

Commands:

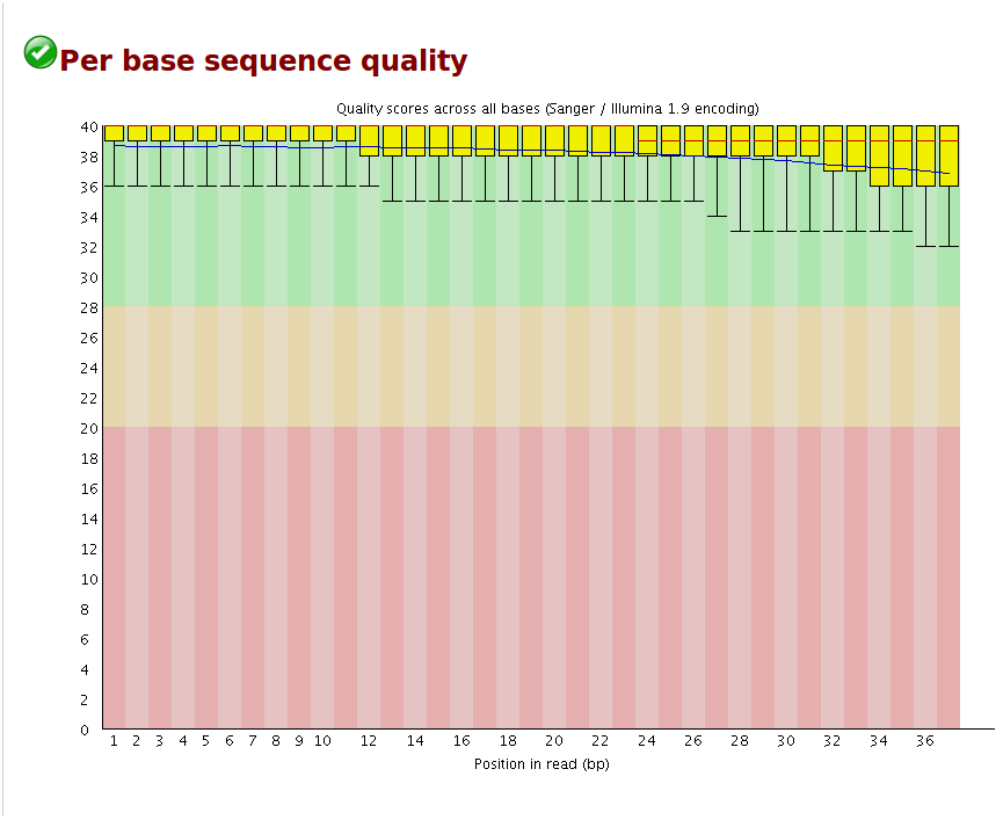
ssh andypetes94@mojo.nuigalway.ie
ssh nextgen2015@syd
qrsh

cd /data4/nextgen2015/users/17232658/
mkdir Assignment 1
cd Assignment 1
cp /data4/nextgen2015/pilib/chip_seq/MA5112_assign1/* ./

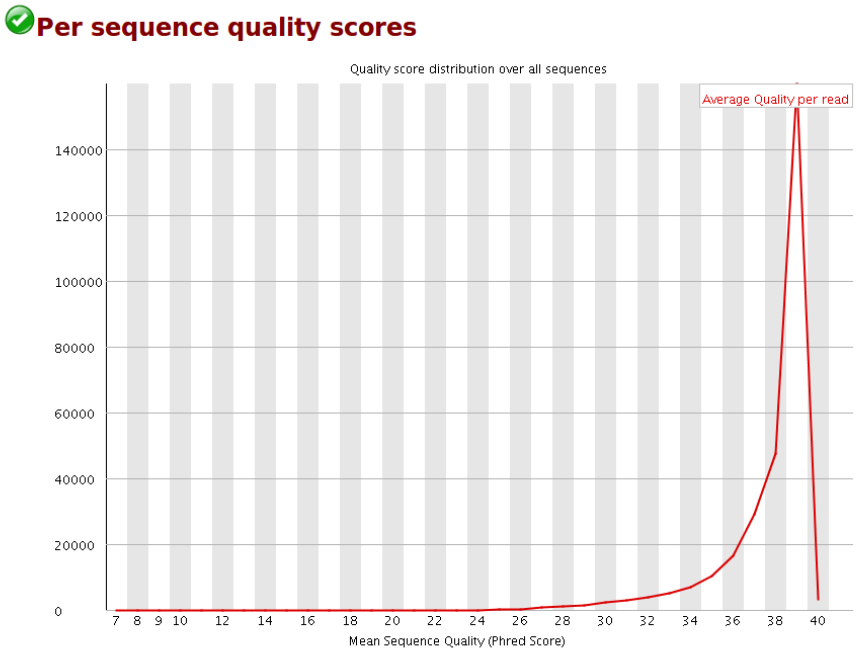
Perform FastQC:
module load fastqc
fastqc input.fastq
fastqc chip_fastqc
cp input_fastqc.html /home/nextgen2015/users/17232658/
cp chip_fastqc.html /home/nextgen2015/users/17232658/

Terminal 2:
scp nextgen2015@syd:/home/nextgen2015/users/17232658/*input_fastqc* ./
scp nextgen2015@syd:/home/nextgen2015/users/17232658/*chip_fastqc* ./

Output - Analysis of Chip_Fastqc QC data:

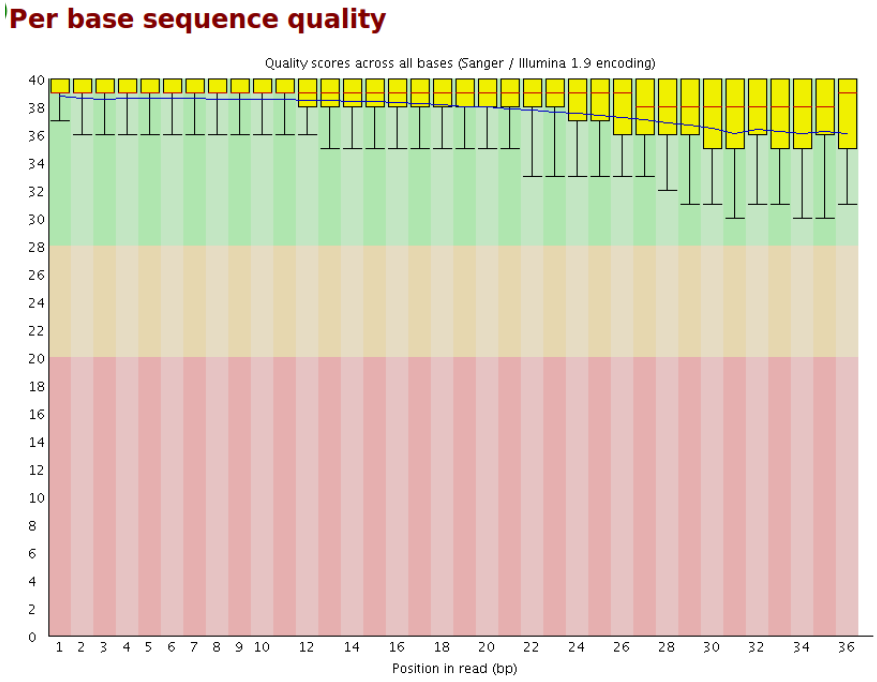


The above graph is an outline of the per base sequence quality. The graph above illustrates how the read quality reduces with increased base length. This is expected as the quality of the read degenerates with increases length. However, based on the above data, every “read” exists within the “green” zone and is therefore of very good quality.

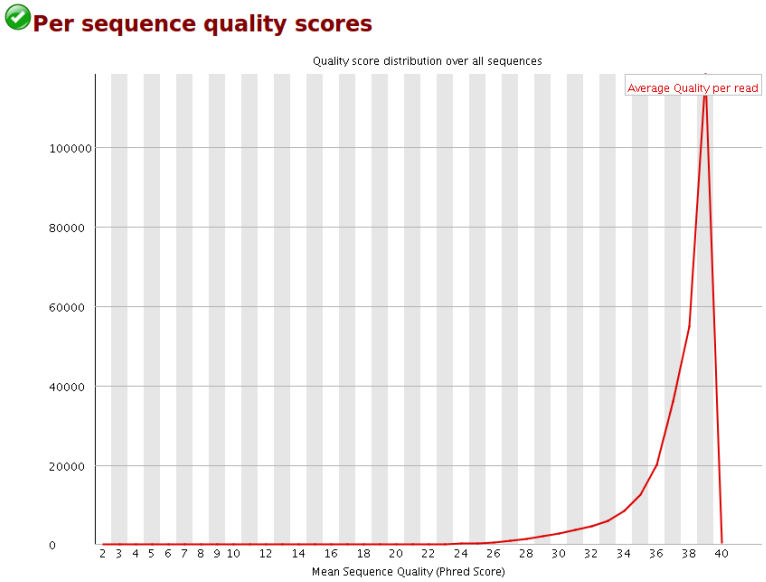


This graph gives a representation of the mean quality “phred scores”. These are a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing. As is observed, the mean quality is around the 40 mark (39). A phred score of 40 is representative of a 1 in 10,000 probability of incorrect base call, or a base 99.99% base call accuracy. It is thererfore intuituve to think that these are extremely good quality base calls.

Analysis of Input_Fastqc QC data:



Just like the “chip” data, the input per base sequence quality is very good also. This is again evident as all reads exist within the “green” region.



Again,the per sequence quality scores are very positive, with an average quality phred score of in and around 40. This is representative of a of a 1 in 10,000 probability of incorrect base call, and is therefore indicative good read quality.

Overall, it is evident from the independent fastqc analysis that the read's are of good quality and should be used for further research.

Commands - Perform MultiQC:

```
nano do_qc.sh
#!/bin/bash
for f in *.fastq;
do
fastqc $f;
done
multiqc .;
chmod a+x do_qc.sh

./do_qc.sh
```

After this
multiqc_data
multiqc_report.html
are generated.

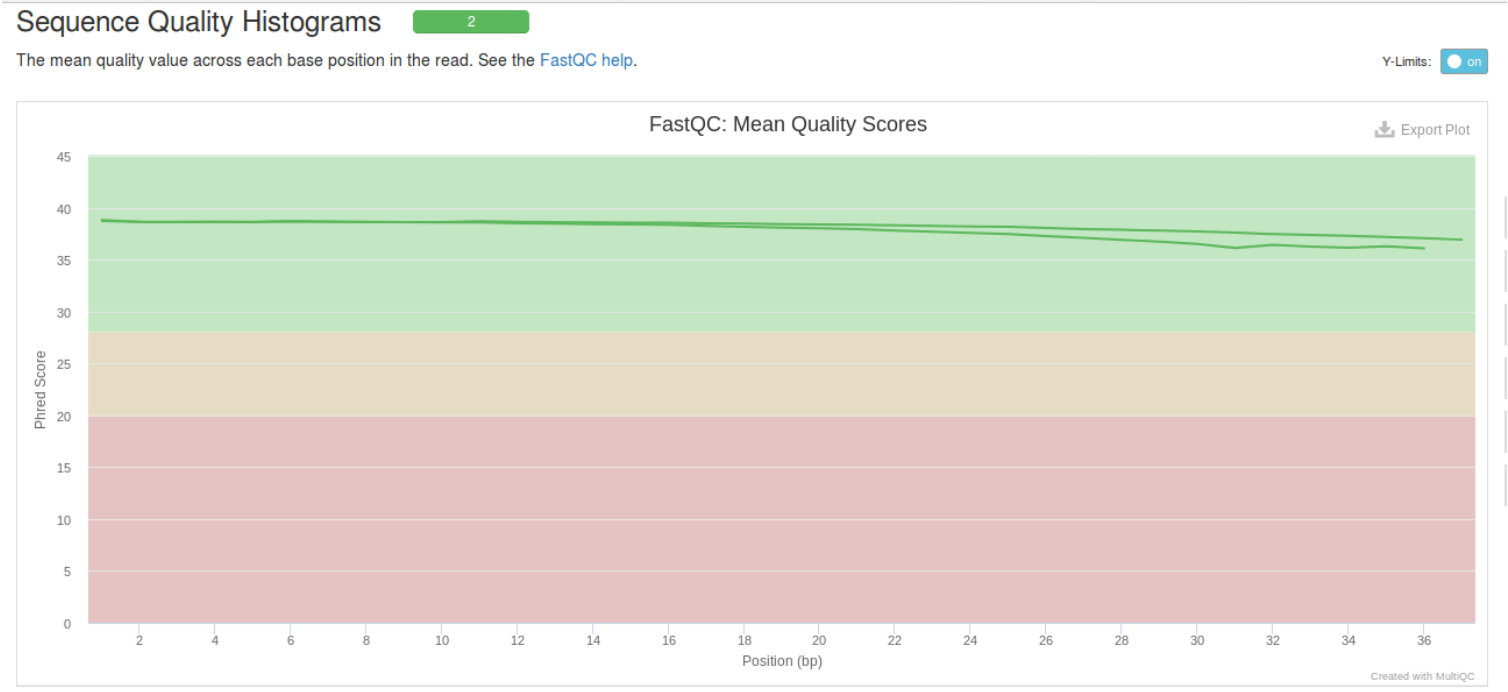
Copy "multiqc_report.html" to local machine:

```
cp multiqc_report.html /home/nextgen2015/users/17232658/
```

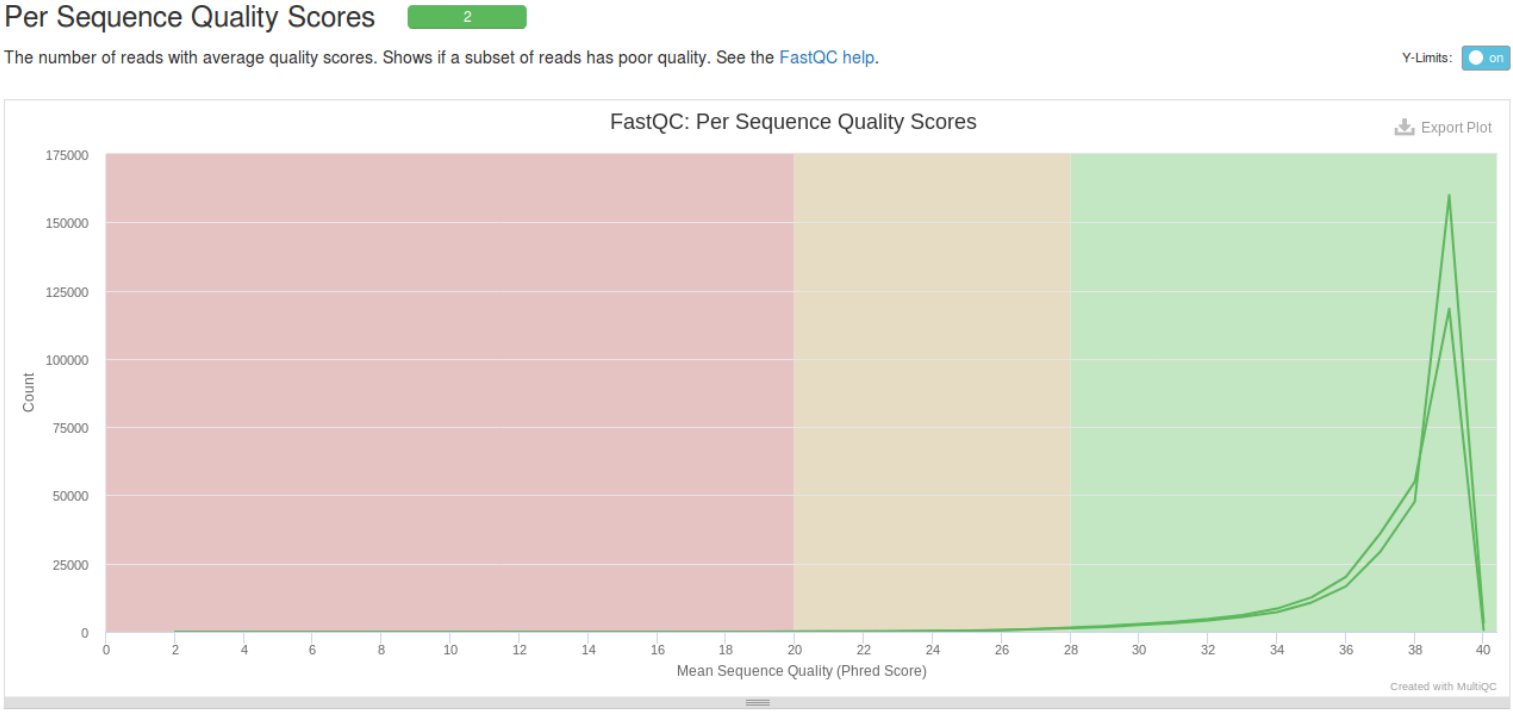
Terminal 2:

```
scp nextgen2015@syd:/home/nextgen2015/users/17232658/* ./
```

Output - Analysis of QC data:



This “sequence quality histogram” exhibits the mean quality value across each base position in the read. The above line is the “chip” data, and therefore has a slightly higher mean quality score than the “input” data. However, both datasets have very good quality as evidenced by high Phred scores existing within the “green” zone.



Here, the higher “peak” is from the “chip” dataset and therefore represents a higher concentration of very good reads (around Phred score of 39). As is evidenced by the above graph, it is clear that both datasets have very good per sequence quality scores as they exist in the “green” zone.

Commands:
#Obtain reference sequence for Chromosome 21:
wget --timestamping
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/chr21.fa.gz

gunzip chr21.fa.gz

module load bowtie

#Build a model using the human genome chromosome 21 sequence.

bowtie2-build chr21.fa hg19

#Generate alignment of model just built and data.

bowtie2 -x hg19 -U chip.fastq -S chip.sam

bowtie2 -x hg19 -U input.fastq -S input.sam

#Convert SAM reads to bam.

samtools view -Sb chip.sam > chip.bam

samtools view -Sb input.sam > input.bam

#Remove duplicates.
samtools rmdup chip.bam chip.rmdup.bam
samtools rmdup input.bam input.rmdup.bam

#Sort the files - this automatically adds the .bam so we can leave it off the output files.
samtools sort chip.rmdup.bam chip.rmdup.sorted
samtools sort input.rmdup.bam input.rmdup.sorted

#Index the files - creates .bai files.
samtools index chip.rmdup.sorted.bam
samtools index inputrmdup.sorted.bam

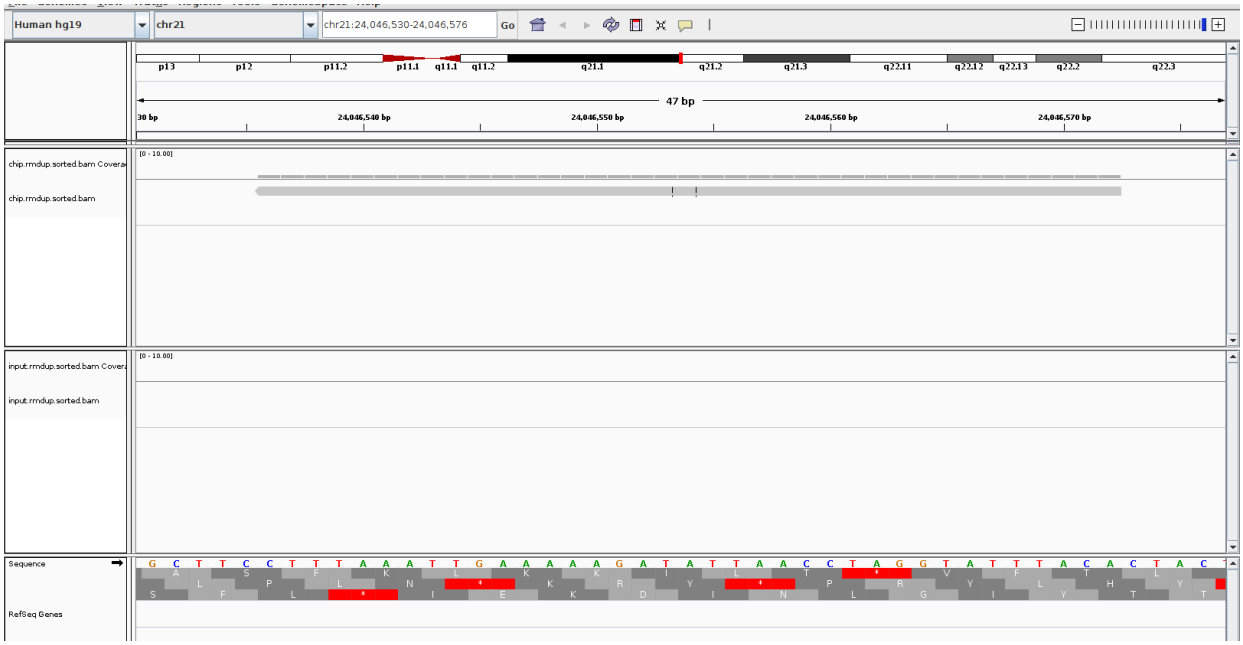
#Get the mapping stats.
samtools flagstat chip.rmdup.sorted.bam > chip_mappingstats.txt
samtools flagstat input.rmdup.sorted.bam > input_mappingstats.txt

cp *chip.rmdup* /home/nextgen2015/users/17232658/
cp *input.rmdup* /home/nextgen2015/users/17232658/

New Terminal:
scp nextgen2015@syd:/home/nextgen2015/users/17232658/*chip.rmdup* ./
scp nextgen2015@syd:/home/nextgen2015/users/17232658/*input.rmdup* ./
cd /home/user9/Desktop/

mkdir Chip_Seq
mv *chip.rmdup* /home/user9/Desktop/Chip_Seq/
mv *input.rmdup* /home/user9/Desktop/Chip_Seq/
igv

IGV Analysis:



I zoomed in on the postition chr21:24,046,530 – 24,046,576, and identified a potential motif that was in the chip data set and not the input dataset. The nucleotide sequence and associated amino acid profile is present immediately below at the bottom. Perhaps in this case there was loss of the initial (input) genetic material corresponding to this region before sequencing.

I zoomed in on the postition chr21:24,046,676 – 24,046,722, and identified a sequence in the input file that wasn't present in the chip dataset. This would indicate that perhaps this region isn't a TF-binding motif as it is not persent in the chip_seq dataset after antibody precipitation.

Peak Calling (MACS2):
Chip-seq has been used in recent times to characterise the *cistome* and in recent projects such as the *BrestPredict* to determine transcription factor binding sites and motif discovery. However, Chip-seq presents some limitations. These include that the Chip tags only represent the ends of the Chip fragments and not the precise protein binding location; Chip-seq data exhibit regional biases. *MACS* offers four important utilities for predicting protein-DNA interaction sites from ChIP-Seq data to address these limitations. First, MACS improves the spatial resolution of the predicted sites by empirically modeling the distanced and shifting tags by d/2. Second, MACS uses a dynamic λ parameter to capture local biases in the genome and improves the robustness and specificity of the prediction. Third, MACS can be applied to ChIP-Seq experiments without controls, and to those with controls with improved performance. Finally, MACS is easy to use and provides detailed information for each peak, such as genome coordinates, p-value, FDR, fold_enrichment, and summit (peak center).

Commands:
macs2
macs2 callpeak -t chip.bam -c input.bam -f BAM -g hs -n macs_out --call-summits -B

-t is the treatment file
-c is the control
-f is the file format
-g represents the genome size (hs = *homo sapiens*)
-n is the output file prefix
--call-summits is a secondary processing of peaks to deconvolve adjacent binding events
-B store the fragment pileup, control lambda, -log10pvalue and -log10qvalue scores in bedGraph files for upload to genome browser.

cp *macs_out* /home/nextgen2015/users/17232658/

New Terminal:
scp nextgen2015@syd:/home/nextgen2015/users/17232658/* ./
mv *macs_out* /home/user9/Desktop/Chip_Seq/

Peak Annotations (GREAT):

GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. These cis-regulatory regions are identified via experimental methods, such as our Chip-seq data.

Commands:
awk '{print "chr21" " "$2" "$3}' macs_out_peaks.xls > chip.bed
awk '{print "chr21\t" $2"\t" $3}' chip.bed > chip_tab.bed
bedtools getfasta -fi chr21.fa -bed chip_tab.bed -fo human_output

cp human_output /home/nextgen2015/users/17232658/
cp chip_tab.bed /home/nextgen2015/users/17232658/

New Terminal:

scp nextgen2015@syd:/home/nextgen2015/users/17232658/*chip_tab* ./

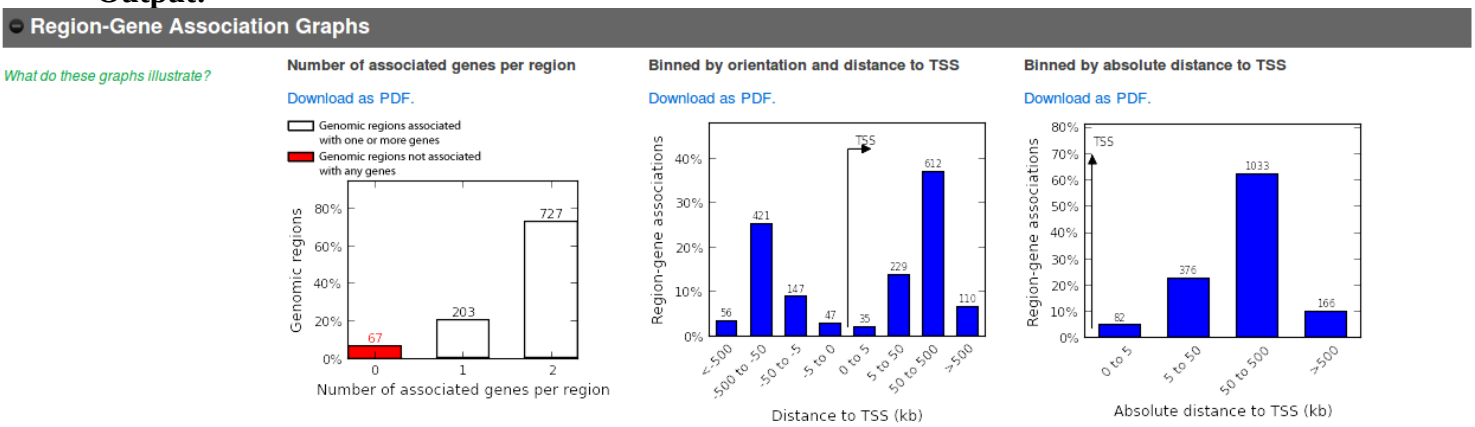
scp nextgen2015@syd:/home/nextgen2015/users/17232658/*human_output* ./

cp *human_output* /home/user9/Desktop/Chip_Seq/

cp *chip_tab* /home/user9/Desktop/Chip_Seq/

Web Browser:
http://bejerano.stanford.edu/great/public/cgi-bin/greatWeb.php
Selected “Species Assembly” as “Human: GRCh37” & attached “chip_tab.bed” to “test regions”

Output:



For all the above graphs, displayed is statistics about the association of input genomic regions to the TSS of all the genes putatively regulated by the genomic regions. In all cases the y-axis displays percentages. Based on the graphs, 67 genomic regions don’t associate with any gene, 203 regions associate with 1 gene and 727 regions associate with 2 genes. The next 2 graphs represent the approximate distance in kilobases of the genomic regions to the tanscriptioanl start sites.

Ielected the “View all region-gene associations” from the drop-down menu of “Global Controls”.

All genomic region-gene association tables (997 regions, 157 genes)

Job ID: 20180206-public-3.0.0-evl1Po

Display name: chip_tab.bed

What do these tables show?

Genomic region -> gene association table Download table as text.

Region	Gene (distance to TSS)
unnamed	NONE
unnamed	NONE
unnamed	NONE
unnamed	NONE
unnamed	TPTE (+414,522)
unnamed	TPTE (+66,349)
unnamed	TPTE (-35,911)
unnamed	TPTE (-153,653)
unnamed	POTED (-258,105)
unnamed	POTED (-216,411)
unnamed	POTED (-89,954)
unnamed	POTED (-84,886)
unnamed	POTED (+12,443), LIPI (+584,329)
unnamed	POTED (+17,855), LIPI (+578,917)
unnamed	POTED (+73,838), LIPI (+522,934)
unnamed	POTED (+74,921), LIPI (+521,851)
unnamed	POTED (+78,984), LIPI (+517,788)
unnamed	POTED (+94,994), LIPI (+501,778)
unnamed	POTED (+247,060), LIPI (+349,712)
unnamed	POTED (+247,714), LIPI (+349,058)

Gene -> genomic region association table Download table as text.

Gene	Region (distance to TSS)
ABCG1	unnamed (-132,878), unnamed (-112,672), unnamed (-98,805), unnamed (-50,787), unnamed (-25,562), unnamed (+43,878), unnamed (+71,906)
ADAMTS1	unnamed (-35,024), unnamed (+33,061), unnamed (+257,211)
ADAMTS5	unnamed (-74,468), unnamed (+86,080)
ADARB1	unnamed (-84,393), unnamed (-77,799), unnamed (-46,021), unnamed (-19,880), unnamed (-4,361), unnamed (-1,037), unnamed (+28,228), unnamed (+31,618), unnamed (+67,251), unnamed (+73,808), unnamed (+76,465), unnamed (+117,156), unnamed (+122,417), unnamed (+161,555), unnamed (+185,708)
AGPAT3	unnamed (-110,369), unnamed (-108,047), unnamed (-44,819), unnamed (+4,732), unnamed (+30,202), unnamed (+52,515), unnamed (+54,881)
AIRE	unnamed (+9,098)
APP	unnamed (-219,491), unnamed (-219,491), unnamed (-218,386), unnamed (-183,426), unnamed (+3,591), unnamed (+22,332), unnamed (+30,334), unnamed (+31,897), unnamed (+44,653), unnamed (+54,084), unnamed (+63,783), unnamed (+81,470), unnamed (+147,717), unnamed (+199,893), unnamed (+207,219), unnamed (+235,527), unnamed (+292,148)
ATP5J	unnamed (-432,037), unnamed (-413,296), unnamed (-405,294), unnamed (-403,731), unnamed (-390,975), unnamed (-381,544), unnamed (-371,845), unnamed

The tables generated link each input genomic region to the genes it regulates according to the association rule, adopted by the *GREAT* algorithm. The first table shows how each genomic region relates to their accompanying gene(s) and the second table conveys the association of gene's with their genomic regions.

Summaries for TPTE Gene

Entrez Gene Summary for TPTE Gene

This gene encodes a PTEN-related tyrosine phosphatase which may play a role in the signal transduction pathways of the endocrine or spermatogenic function of the testis. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Mar 2014]

GeneCards Summary for TPTE Gene

TPTE (Transmembrane Phosphatase With Tensin Homology) is a Protein Coding gene. Among its related pathways are Metabolism and Glycerophospholipid biosynthesis. GO annotations related to this gene include *ion channel activity* and *protein tyrosine phosphatase activity*. An important paralog of this gene is TPTE2.

UniProtKB/Swiss-Prot for TPTE Gene TPTE_HUMAN,P56180

Could be involved in signal transduction.

Gene Wiki entry for TPTE Gene

After further inspection into the *TPTE* gene, I found out that it is involved in endocrine and spermatogenic pathways in the testis, and therefore didn't further consider this gene for further discovery in this particular *BreastPredict* project breast cancer cell line.

Looked at GO Molecular Function:

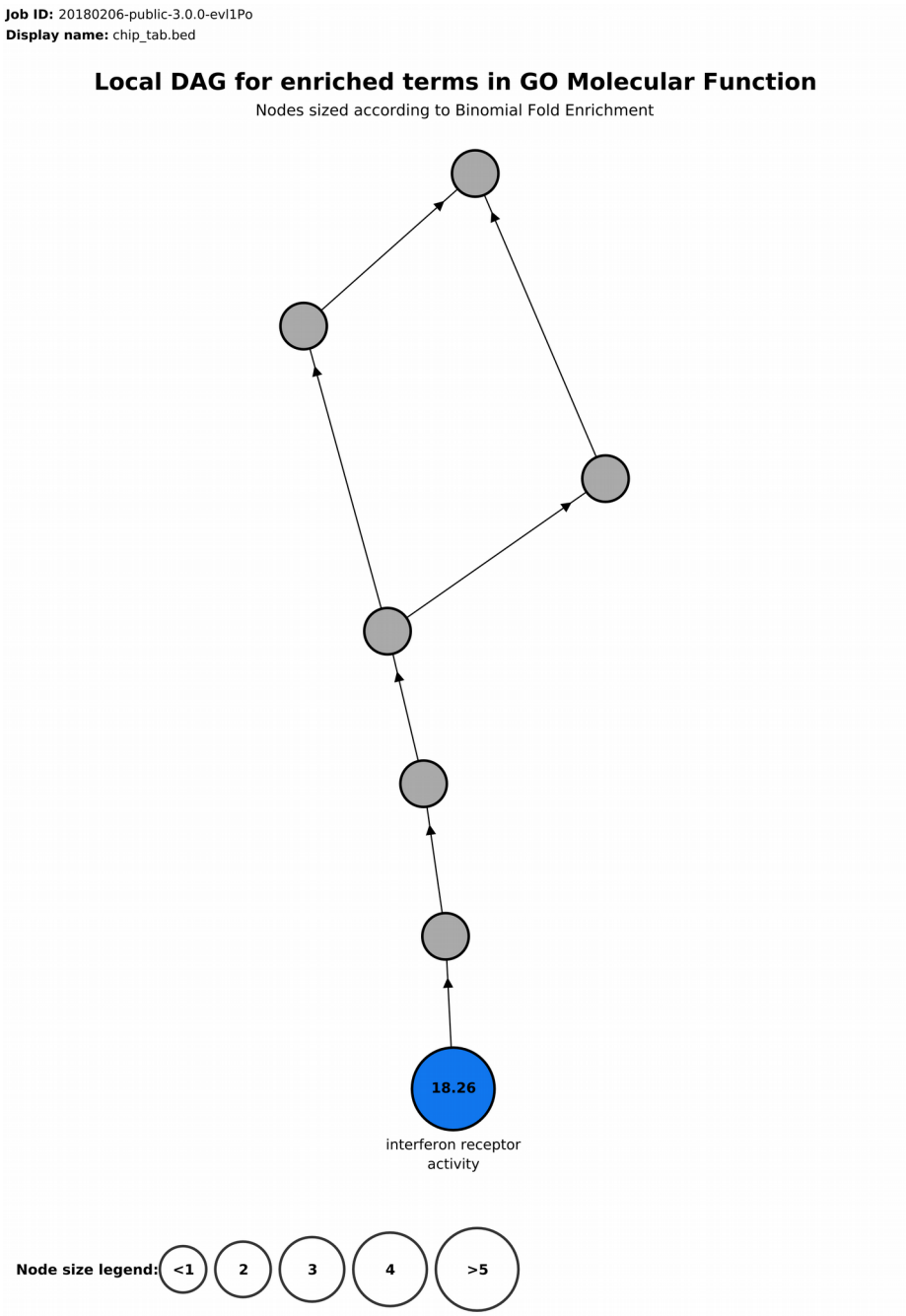
GO Molecular Function (1 term)

Table controls: Export | Shown top rows in this table: 20 Set | Term annotation count: Min: 1 Max: Inf Set | Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
interferon receptor activity	154	**1.0118e-5**	**2.4231e-4**	18.2637	5	0.50%	1	**2.3543e-2**	68.9465	3	5	1.91%

The test set of 997 genomic regions picked 157 (1%) of all 18,041 genes.
GO Molecular Function has 3,688 terms covering 15,090 (84%) of all 18,041 genes, and 189,388 term - gene associations.
3,688 ontology terms (100%) were tested using an annotation count range of [1, Inf].

Selected the "Viualize this table" drop-down menu and selected "visualise shown terms in hierarchy".



This visualization generates a directed acyclic graph (DAG) based on the enriched terms from a single ontology-specific table from a GREAT job. Enriched terms are shown in blue. Nodes have been sized according to Binomial Fold Enrichment. Here is also an option of holding "ctrl" and clicking a node to toggle whether the name/statistic of the term is shown. Therefore the above map shows that *interferon receptor activity* is an enriched term.

After further analysis into this term, I found that interferon-α/β receptor (IFNAR) is a virtually ubiquitous membrane receptor which binds endogenous type 1 interferon (IFN) cytokines. This maybe of no surprise in a breast cancer cell line as innate immune signaling would expect to be accelerated in patients fighting cancer and therefore possess high *interferon receptor activity*.

I then scrolled down to “Disease Ontology”

Disease Ontology (6 terms) Global controls

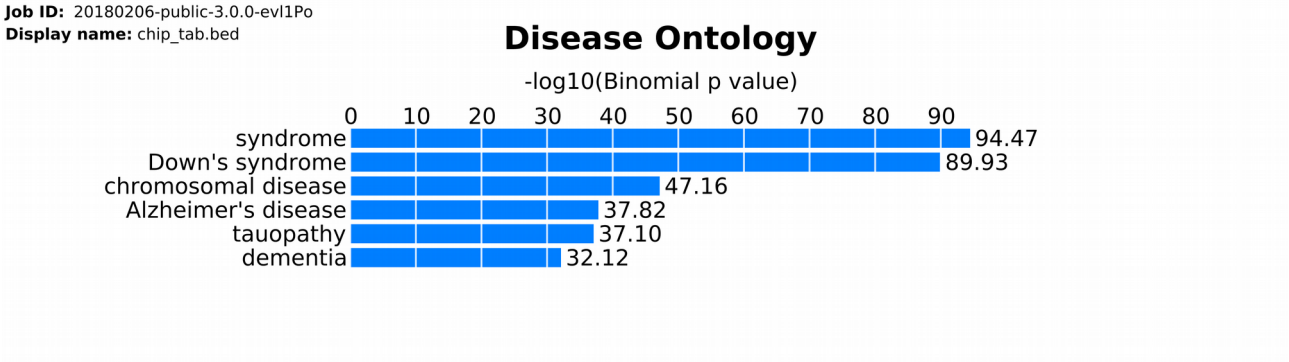
Table controls: Export | Shown top rows in this table: 20 Set | Term annotation count: Min: 1 Max: Inf Set | Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
syndrome	3	3.3589e-95	2.5024e-92	2.4632	494	49.55%	1	4.6907e-7	2.7244	45	1,898	28.66%
Down's syndrome	4	1.1638e-90	6.5029e-88	24.7610	89	8.93%	2	6.3554e-6	26.8125	7	30	4.46%
chromosomal disease	16	6.8729e-48	9.6006e-46	7.5301	89	8.93%	3	1.7557e-2	8.2079	7	98	4.46%
Alzheimer's disease	24	1.5084e-38	1.4047e-36	3.6103	140	14.04%	4	2.1073e-2	3.8304	13	390	8.28%
tauopathy	25	8.0036e-38	7.1552e-36	3.5536	140	14.04%	5	2.0765e-2	3.7534	13	398	8.28%
dementia	30	7.5550e-33	5.6285e-31	3.1795	140	14.04%	7	4.5539e-2	3.3569	13	445	8.28%

The test set of 997 genomic regions picked 157 (1%) of all 18,041 genes.
Disease Ontology has 2,235 terms covering 7,886 (44%) of all 18,041 genes, and 232,324 term - gene associations.
2,235 ontology terms (100%) were tested using an annotation count range of [1, Inf].

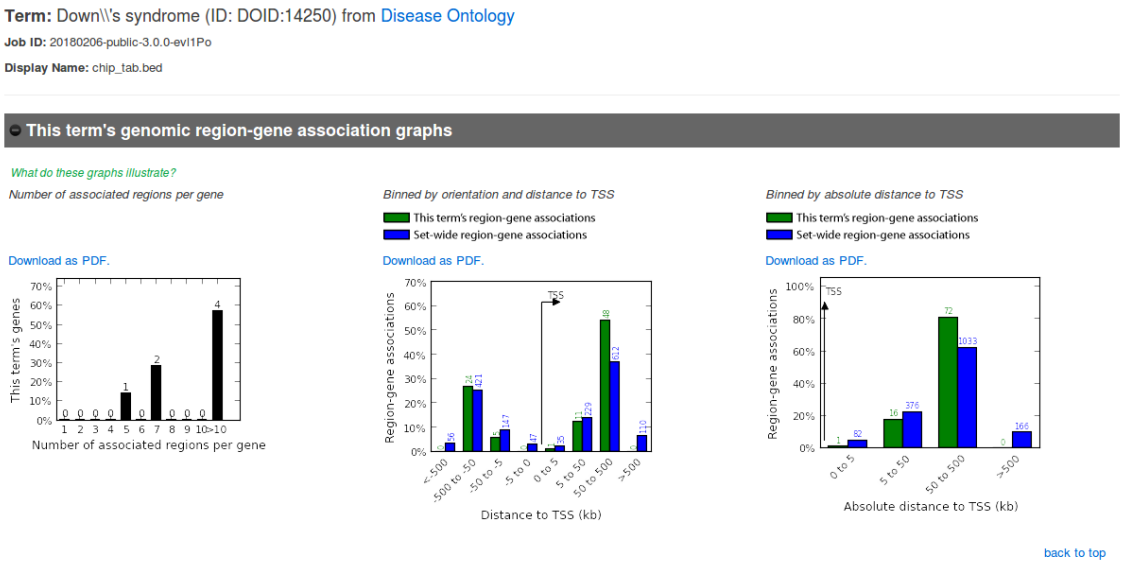
From here I selected the “Visualize the table” drop down and selected the option “Bar chart of current sorted value.”

Job ID: 20180206-public-3.0.0-evl1Po
Display name: chip_tab.bed



This bar chart is based on a single ontology-specific table from a *GREAT* job.

Therefore, none of these disease ontolgies are directly linked to breast cancer.
After “clciking” on the “Down’s syndrome” link, the following results were displayed:



These graphs display similar data to that found in the region-gene associated output earlier, except the later two graphs compare the ontology gene-region associations with the set-wide gene-region associations as indicated by the seperate colours (green and blue respectively).

Meme:
MEME searches for statistically significant motifs from the input sequence set. In our case, the input sequence set is the Chip-seq data. In this way, MEME can discover the binding sites for the shared transcription factor in the set of promoters or the common protein-protein binding domains in the set of proteins. Therefore, our objective is to identify the motif and subsequently the transcription factor present in the breast cancer cell line in accordance with the *BreastPredict* project.

Commands:
module load python
module load meme
meme human_output -dna -mod zoops -minw 6 -maxw 26 -nmotifs 5 -maxsize 1000000 -o meme_chi_out
cp *meme_chi* /home/nextgen2015/users/17232658/

New Terminal:

scp nextgen2015@syd:/home/nextgen2015/users/17232658/*chip_tab* ./

Looked at the “meme_html” output file.

MEME

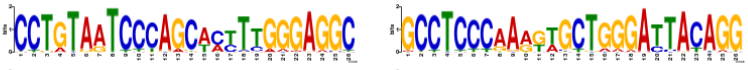
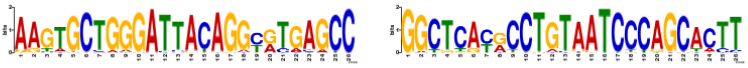
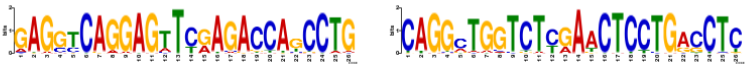
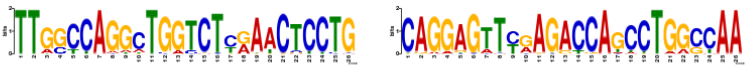
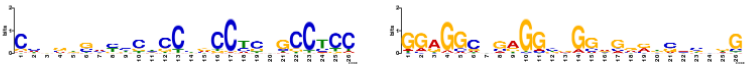
For further information on how to interpret these results or to get a copy of the MEME software please access http://meme.nbcr.net.

If you use MEME in your research, please cite the following paper:
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

DISCOVERED MOTIFS | BLOCK DIAGRAMS OF MOTIFS | PROGRAM INFORMATION | EXPLANATION

DISCOVERED MOTIFS

Motif Overview

Motif 1	• 9.5e-319 • 52 sites	
Motif 2	• 1.3e-253 • 43 sites	
Motif 3	• 2.2e-238 • 43 sites	
Motif 4	• 1.7e-215 • 38 sites	
Motif 5	• 3.4e-189 • 252 sites	

Further Analysis

Submit all motifs to MAST FIMO GOMO BLOCKS Mouse-over buttons for more information.

Above are "logo's" of the top identified motifs, with the left sequence representative of the normal sequence and the right representative of the complemetary sequence.

Motif 1:

MOTIF 1

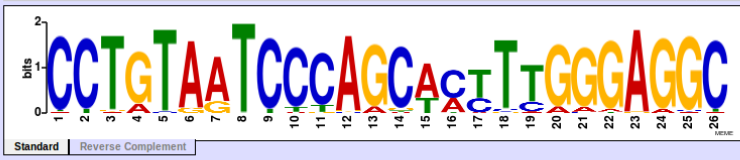
Next To

Summary

E-value	9.5e-319
Width	26
Sites	52

show more

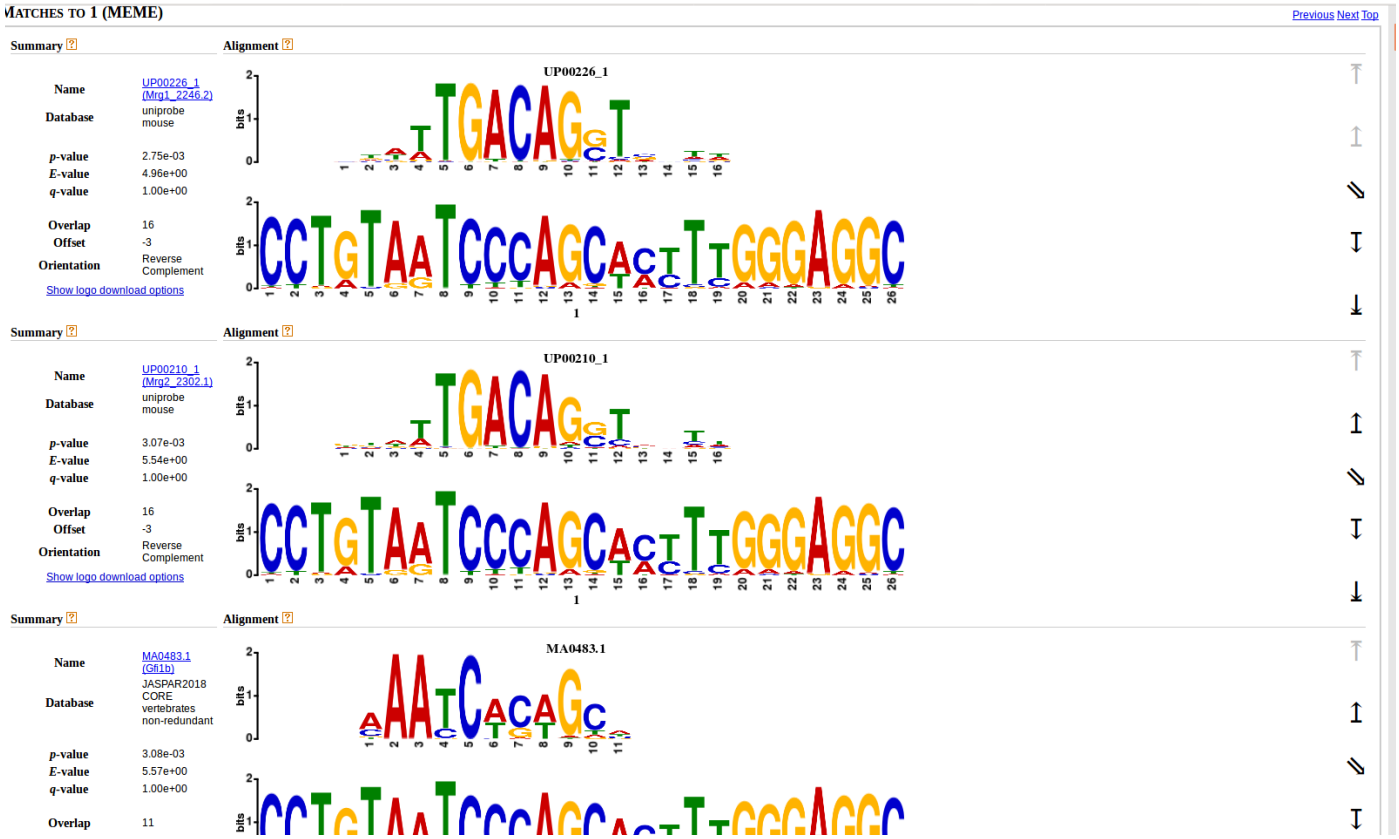
Sequence Logo



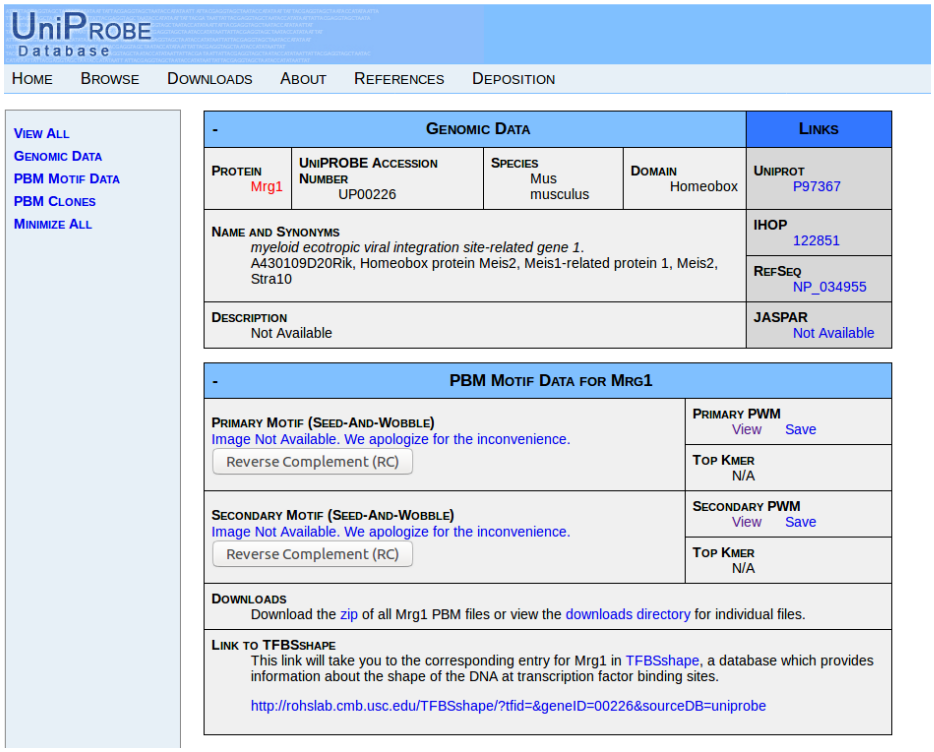
Standard | Reverse Complement

Download LOGO Orientation: standard SSC: off Format: web (png) Width: 26 cm Height: 7.5 cm Download

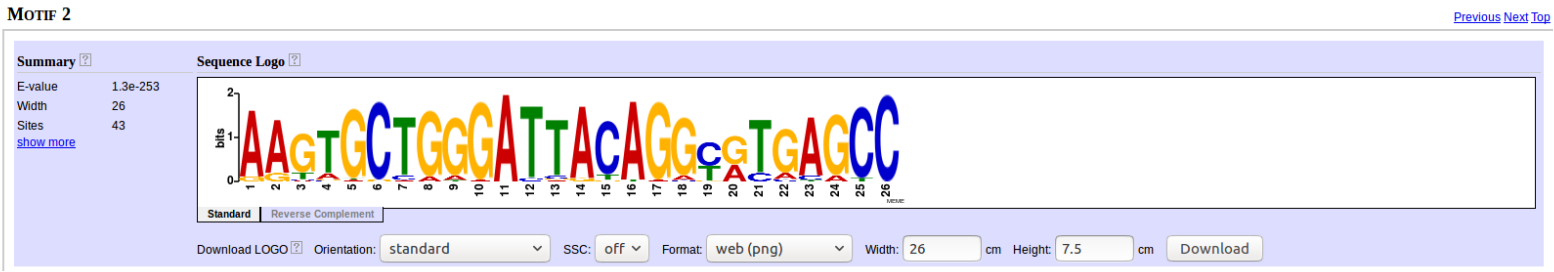
TOMTOM: I used the "TOMTOM" tool at http://meme-suite.org/. *Tomtom* compares one or more motifs against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match. I attached the "html" file and ran the search using the default settings, i.e. searching a vertebrate databsae. I obtained the folowing results:



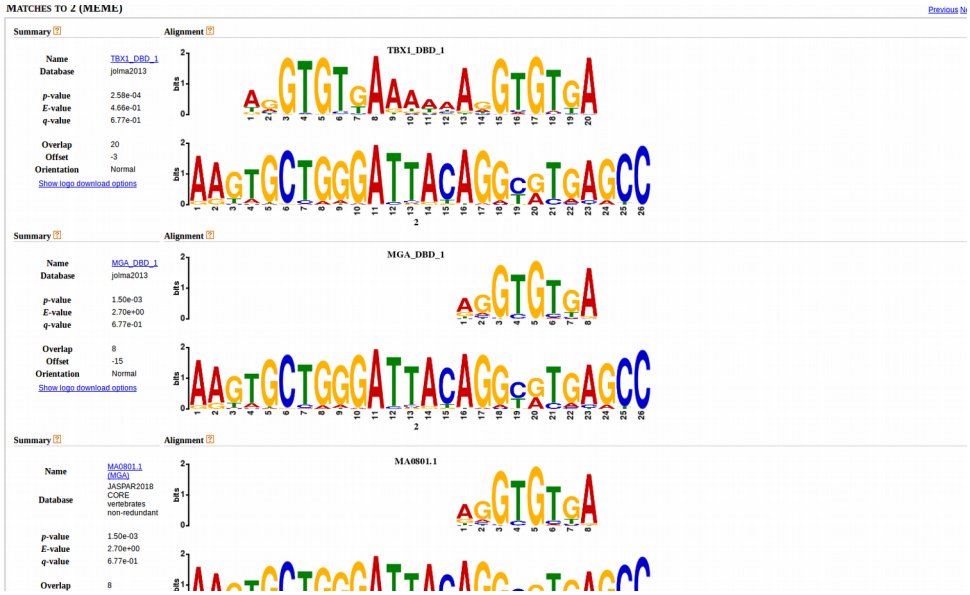
Above is an example of motifs located various databases, with the known motif at the bottom, compared to the databse located motif on the top. On the left, there is some information regarding the motifs, such as p-values, that are considered signifcant at a given significance threshold (p-value); an e-value which is an estimate of the number of (equally or more interesting) motifs one would expect to find by chance if the letters in the input sequences were shuffled; orientation as to whether the motif interacts with the normal sequence or complemtary sequence; the database by which the motif was located. After further inspection of the "top hit" the "uniprobe mouse", I located the following data about the "top hit".



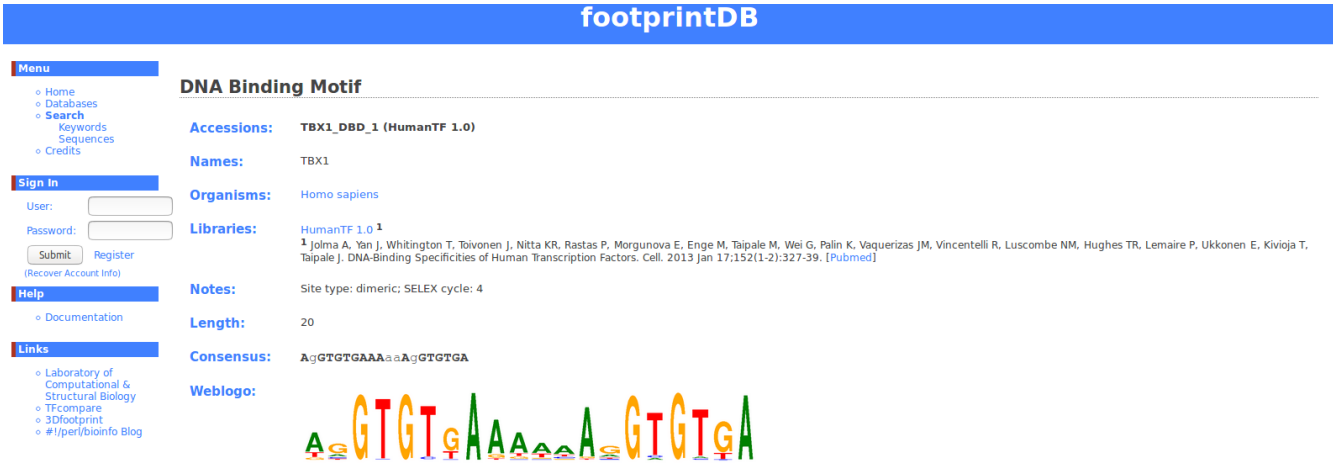
Motif 2:



The following is a set of the “top hits” observed on “TomTom:”



I inspected the “top hit” further at the “jolma2013” database:

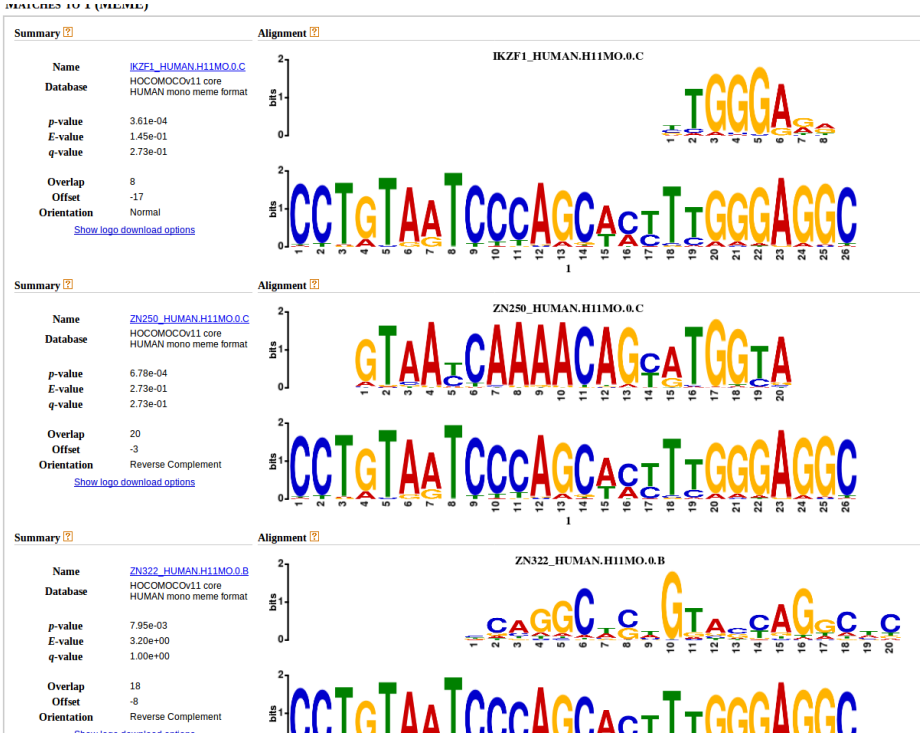


As is observable, this motif is found in *Homo sapians*. Therefore, this may have more relevance in the Irish Cancer Society BreastPredict project as the Chip-seq assay was performed on a specific transcription factor in a human breast cancer cell line.

In order to conduct de-novo motif dicovery analysis to confirm the binding motif in association with the Irish Cancer Society BreastPredict project , I used the “TOMTOM” tool at http://meme-suite.org/ and this time compared with “Human (*homo sapiens)* DNA” and used the default human database.

Below is the results when comparing the “top hit” using the human databases for the first 2 motifs.

Motif 1:



After selecting the “top hit” on the “HOCOMOCOv11 core HUMAN database”, which is the *IKZF1 HUMAN.H11MO.0.C* I obtained the following information:

HOCOMOCO Home Human TFs Mouse TFs Tools Downloads Help Search:

Model info

Transcription factor	IKZF1 (GeneCards)
Model	IKZF1_HUMAN.H11MO.0.C
Model type	Mononucleotide PWM
LOGO	
LOGO (reverse complement)	
Data source	Integrative
Model release	HOCOMOCOv9
Model length	8
Quality	C
Motif rank	0
Consensus	bTGGGARd
Best auROC (human)	
Best auROC (mouse)	
Peak sets in benchmark (human)	
Peak sets in benchmark (mouse)	
Aligned words	61
TF family	Factors with multiple dispersed zinc fingers{2.3.4}
TF subfamily	Ikaros{2.3.4.4}
HGNC	13176
	10320

This transcription factor has a **consensus** sequence of *bTGGGARd* which clearly mimics the later end of the motif. To reinforce the protein's and motif's complementarity, the *MEME*software estimated an accompanying p-value 2.61e-04 of and e-value of 1.45e-01 respectively. This "e-value" is the expected that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. Therefore, the lower the "e-value", the less likely it is the *MEME* database will generate a random transcription factor binding to the motif. Therefore, this motif alignment and acommpanying p-value and e-value points towards the initial transcription factor used in the human breast cancer cell line by the *BreastPredict* project.

However, upon further inspection of the gene's summary, it appears to be expressed in fetal and adult hemo-lymphopoeitic tissue, and therefore may not be the transcription factor used as part of the *BreastPredict* project.

Previous GeneCards Identifiers for IKZF1 Gene

GC07P050314

Search aliases for IKZF1 gene in PubMed and other databases

Summaries for IKZF1 Gene

Entrez Gene Summary for IKZF1 Gene

This gene encodes a transcription factor that belongs to the family of zinc-finger DNA-binding proteins associated with chromatin remodeling. The expression of this protein is restricted to the fetal and adult hemo-lymphopoietic system, and it functions as a regulator of lymphocyte differentiation. Several alternatively spliced transcript variants encoding different isoforms have been described for this gene. Most isoforms share a common C-terminal domain, which contains two zinc finger motifs that are required for hetero- or homo-dimerization, and for interactions with other proteins. The isoforms, however, differ in the number of N-terminal zinc finger motifs that bind DNA and in nuclear localization signal presence, resulting in members with and without DNA-binding properties. Only a few isoforms contain the requisite three or more N-terminal zinc motifs that confer high affinity binding to a specific core DNA sequence element in the promoters of target genes. The non-DNA-binding isoforms are largely found in the cytoplasm, and are thought to function as dominant-negative factors. Overexpression of some dominant-negative isoforms have been associated with B-cell malignancies, such as acute lymphoblastic leukemia (ALL). [provided by RefSeq, May 2014]

CIViC summary for IKZF1 Gene

GeneCards Summary for IKZF1 Gene

IKZF1 (IKAROS Family Zinc Finger 1) is a Protein Coding gene. Diseases associated with IKZF1 include Immunodeficiency, Common Variable, 13 and Ikzf1-Related Common Variable Immune Deficiency. Among its related pathways are Innate Lymphoid Cell Differentiation Pathways and Development of pulmonary dendritic cells and macrophage subsets. GO annotations related to this gene include *transcription factor activity, sequence-specific DNA binding* and *protein heterodimerization activity*. An important paralog of this gene is IKZF3.

UniProtKB/Swiss-Prot for IKZF1 Gene IKZF1_HUMAN,Q13422

Transcription regulator of hematopoietic cell differentiation (PubMed:17934067). Binds gamma-satellite DNA (PubMed:17135265, PubMed:19141594). Plays a role in the development of lymphocytes, B- and T-cells. Binds and activates the enhancer (delta-A element) of the CD3-delta gene. Repressor of the TDT (fikzfterminal deoxynucleotidyltransferase) gene during thymocyte differentiation. Regulates transcription through association with both HDAC-dependent and HDAC-independent complexes. Targets the 2 chromatin-remodeling complexes, NuRD and BAF (SWI/SNF), in a single complex (PYR complex), to the beta-globin locus in adult erythrocytes. Increases normal apoptosis in adult erythroid cells. Confers early temporal competence to retinal progenitor cells (RPCs) (By similarity). Function is isoform-specific and is modulated by dominant-negative inactive isoforms (PubMed:17135265, PubMed:17934067).

The gene accompanying the second motif is ZN250_HUMAN.H11MO.0.C had a big overlap of 20 and an accompanying p-value of 6.78e-04 and e-value of 2.73e-01:

HOCOMOCO Home Human TFs Mouse TFs Tools Downloads Help Search:

Model info

Transcription factor	ZNF250 (GeneCards)
Model	ZN250_HUMAN.H11MO.0.C
Model type	Mononucleotide PWM
LOGO	
LOGO (reverse complement)	
Data source	ChIP-Seq
Model release	HOCOMOCOv11
Model length	20
Quality	C
Motif rank	0
Consensus	TACCATRCTGTTTTGRTTAC
Best auROC (human)	0.9043242591801471
Best auROC (mouse)	
Peak sets in benchmark (human)	4
Peak sets in benchmark (mouse)	
Aligned words	360
TF family	More than 3 adjacent zinc finger factors{2.3.3}
TF subfamily	unclassified{2.3.3.0}
HGNC	13044
EntrezGene	58500 (SSTAR)
UniProt ID	ZN250_HUMAN
UniProt AC	P15622 (TFClass)
Comment	
Downloads	pcm pwm

This is a zinc finger transcription factor (250) from which the data source obtained is “Chip-seq”. So perhaps this potentially is the Transcription factor immunprecipitated by the “Chip-seq” experiment, which is consistent with the miniscule value and accompanying large nucleotide overlap with this binding motif. Based on the proteins’s summary, it doesnt specify a particular role or tissue this operates in and therefore perhaps requires further experimentation to validate it’s specificity and activity.

Summaries for ZNF250 Gene

GeneCards Summary for ZNF250 Gene

ZNF250 (Zinc Finger Protein 250) is a Protein Coding gene. Among its related pathways are Gene Expression. GO annotations related to this gene include *nucleic acid binding* and *transcription factor activity, sequence-specific DNA binding*. An important paralog of this gene is ZNF879.

UniProtKB/Swiss-Prot for ZNF250 Gene ZN250_HUMAN,P15622

May be involved in transcriptional regulation.

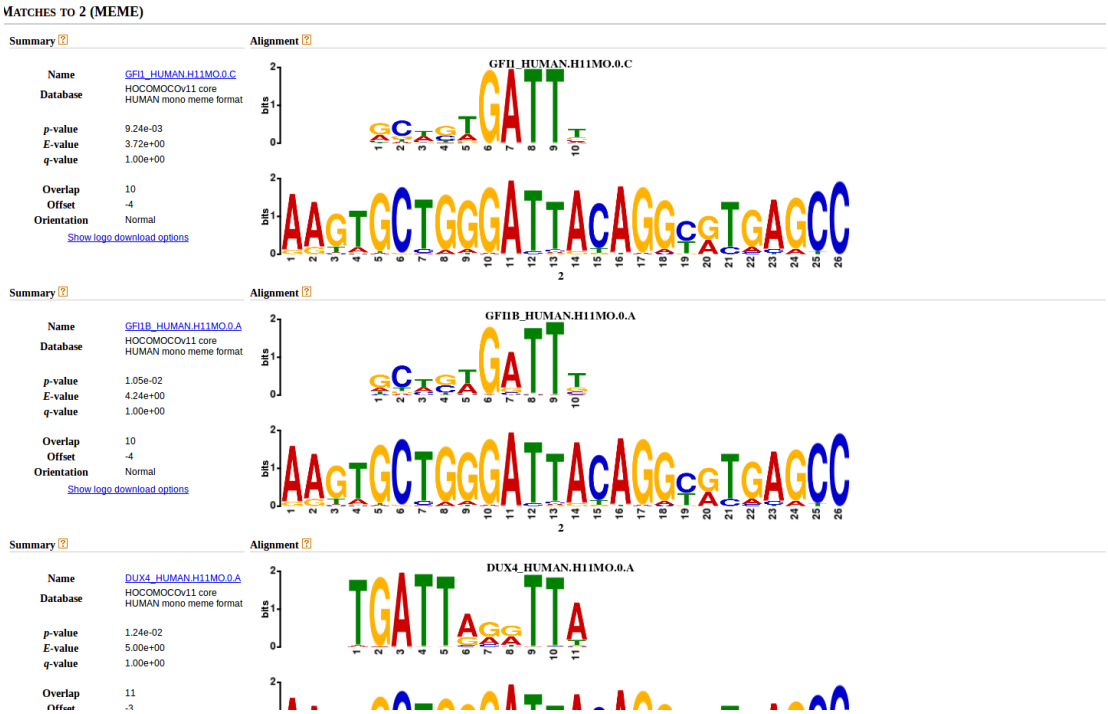
Additional gene information for ZNF250 Gene NEW!

HGNC Entrez Gene Ensembl UniProtKB

Search for ZNF250 at DataMed
Search for ZNF250 at HumanCyc

No data available for Entrez Gene Summary , CIViC summary , Tocris Summary , Gene Wiki entry , PharmGKB "VIP" Summary , fRNAdb sequence ontologies and piRNA Summary for ZNF250 Gene

Motif 2:



After following up on the “top hit”, I found the following information on the “HOCOMOCOv11 core HUMAN database” regarding motif 2.

Model info

Transcription factor	GFI1 (GeneCards)
Model	GFI1_HUMAN.H11MO.0.C
Model type	Mononucleotide PWM
LOGO	
LOGO (reverse complement)	
Data source	Integrative
Model release	HOCOMOCOv9
Model length	10
Quality	C
Motif rank	0
Consensus	RShSWGATTb
Best auROC (human)	
Best auROC (mouse)	
Peak sets in benchmark (human)	
Peak sets in benchmark (mouse)	
Aligned words	152
TF family	More than 3 adjacent zinc finger factors{2.3.3}
TF subfamily	GFI1 factors{2.3.3.21}
HGNC	4237
EntrezGene	2672 (SSTAR)
UniProt ID	GFI1_HUMAN
UniProt AC	Q99684 (TFClass)
Comment	
Downloads	pcm pwm

Although this could potentially the specific transcription factor immunoprecipitated in the *BreastPredict* project, it has a lower p-value than the potential proteins that interact at motif 1. Therefore, I feel the transcripton factors *IKZF1* and *ZNF250* are more likely viable candidates that were immunoprecipitated in the *chip-seq* assay condcuted by the *BreastPredict* project. However, the *IKZF1* transcription factor doesn't seem to function in breast tissue and the *ZNF250* transcription factor has no indication of operating in this tissue. Therefore, perhaps further validation of function is required in both of these transcription factors to elucidate their function and identify any links with breast cancer cell lines.