TRANSCRIPTION FACTOR BINDING SITE AND REGULATORY SNP META-ANALYSIS BASED ON CHIP-SEQ DATA

Endre Barta
DEOEC, BMBI, Clinical Genomics Center
ABC, Agricultural Genomics and Bioinformatics Group
barta.endre@unideb.hu
barta@abc.hu

REGULOMIX
consortium

What is the regulatory SNP (rSNP)?

- A SNP that is in a regulatory region (transcription factor binding site)
- The transcription factor affinity to the binding site is depending on the allele
- The phenotypic change is the altered gene expression

- How to find TFBSs genome-wide?

```
                    MGF
ovine   -1196 GAAGTCAAACCATTCTGAGAAATAGAAATTTTC_____AATTTACTCTGTACAAGCT
caprine -1194 GAAGTCAAACTATTCTGAGAAATAGAAATTTTC_____AATTTACTCTGTACAATCT
bovine  -1203 GAAGTCAAACCATTCTGAGAAATAGAAATTTTC_____AATTTACTCTGTACAAGCT
rabbit  -1025 GAAACCAA_CTATTCTGAGAAATAGGAAC_____AACTTATTCCACATAGGTT
human   -1275 AAAACCAA_CTATTCTGAGAAACAGAAATTTCATATTACTATTTTACTCTGTGTAATCC
mouse   -1037 GAAGCAAAC___TTCTAAGAAATAGAAAGCCAGGGC_____TTATTGTGGAGAAGCT
                **  **   **  **** ***** ** **           *** *           *
                                                                    MGF
ovine   -1144 GAATAAACATTT_AGGATCATGACCATCACTCTGAACCTTCCAAGACATGAAATACTTC
caprine -1142 GAATAAACATTT__AGGATCATGACCATCACTCTGAAGCTTCCAAGACATGAAATACTTC
bovine  -1151 GAATAAACATTT_AGGGTCATGACCATCACTCTGAACATTCAAGATATGAAATACTTC
rabbit  -978 GAATAACTACTCT_AGCATTCATAATGTCA_____TTC
human   -1216 AAATAAGTATTCT_AGGGTTTAGAATATTGTTCCTAACTTTCTAAGACATGCAATACTTC
mouse   -988 GCACAAGTATTCTCAGGATTCAGAACAATGTTCCCTGCTTTCTAAGACGTGTAATGGTTC
                *  **  *** **  ***    *   *                        *     ***
                              CB1
ovine   -1085 TAAGAAAAAGA_____GGAAT_C_AACACAATGCTGTTATTAATTCTTGTTTGTAT
caprine -1083 TAAGAAAAAGA_____GGAATAC_AACACAATGCTGTTATTAATTCTTGTTTGTAT
bovine  -1092 TAAGAAAAAGA_____GGAATAC_AACACAATGCTGTTATTAATTCTTGTTTGTAT
rabbit  -946 CAAGAAACAAACAGGAACACAGAAATA_ACTACAATGCCAATATTATCTCCTGTTTCTCT
human   -1157 TAAGAAACAGACAGGAATATAGACATA_ACTACAATGCTGATATTATTTCTTGTTTGTAT
mouse   -928 CCAGAAACAGGCAGGAATACATAAAAATTAACTACAATGCTGACATTATTTCTGCTCTGCAT
                 ****** *                 * ******* **** ** * **
                        CB2
ovine   -1036 CCTGGCAGTCCTACATCGATTCCTGTAAATACCACAATCTGATCAGGACTTCATAGATAA
caprine -1033 CCTGGCAGTCCTACATCGATTCCTGTAAATACCACAATCTGATCAGGACTTCACAGATAA
bovine  -1042 CCTGGCAGTCCTACATCAATTCCTGTAATCCACA___TTGGGTGAGGACTTCATAGAAAA
rabbit  -887 TCAGGAAGTCTTACATTGATTCTTGTAAGTCACCAATCTTGCCAAGCCTTCACAGCCAA
human   -1098 TCTGGCAGTCCTACATTAATTCCTGTAAGTTCCTCAATCTTGGTCAGGTCTTCACAGCCAA
mouse   -868 TCTGGCTGCTCTGCAGGGATTCCAGTAAATACC____TCTGATGAGGTTTTCACA____A
                * **  *  *      * ****  ****  **   *    *    * ****  **   *
                       CB3
ovine   -976 AATGAAATCACAGTTAACATTTTTTTGTGGAGAAATGTAAGCGAAAG_AGATATTCTTTC
caprine -973 AATGAAATCACAGTTAACATTTTTTTGTGGAGAAATGTAAGCAAAAGCAGATAGTCTTTC
bovine  -985 TG___AATCACAGTTAACATTTTT__GTGGAGAAATGTAACGAAAACGAGATATTCTTTC
rabbit  -827 GGTAAAGTTGCAGTTAACATTAT_____TC
human   -1038 AAT_AAGTCACAGTTAACATTATTTCTTT_AGAAAAGCAAGTAAAAGAAAATATGCTTTC
mouse   -816 AGCAAAGTCATTGTTAGCATTACCCTGGGGTGGGGGGAGGGACAGGTAAAGCAGGATATC
                   * *  **** **** * *
                                    B6   CB4        YY1
ovine   -917 CTTAATTATCTAGG_AAAATTATTTGGTTAGTGGTATTTTACCAAAATACCCCCCATAT
caprine -913 CTTAATTACCTAGG_AAAATTATTTGGTTAGTGGTATTTTACCAAAATACCCCCCATAT
bovine  -930 CTTAATTATGTAGG_AAAATTATTTGGTTACGAGTATTTTACTAAAATACCCCCATTT
rabbit  -802 CCAAGAAATTCAAGTAAAAGGATTTGTTCAACTAT_TTTTACCAAAATATCCATCATATC
human   -980 CTGAATTATCTAAGCAAAAT__TTTGTTTAGCTATTTTTTGCCAAAATATTCTCAATGTC
mouse   -756 ATT_____TAGGCAAAAT_ACATGTTTGGCTTCATTTTACTAAAATTTTTACAGCATT
                     * * ****  **   **  *          **** *  *****
                                        GR-half rev  B3_rev
ovine   -858 GGTGGCTTTAAGATATATATTTGTAAGTCAGGATAAGCCGTCTTTGAAACAGAACATT
caprine -854 GGTGGCTTTAAGATATATATTTGTAAGTCAGGATAAGCCATCTTTGAAACAGAACATT
bovine  -871 GGTGGCTTTAAGATATATATTTGTAAGTCAG_ATAAGCCGTCTTTGAAACAGAACATT
rabbit  -743 AGTGATTTTATAAAGTGTGATCT__AAATCACAATCTGATGTCTTTGAAACAAAACAATT
human   -922 AGTGGTTTGTGATTTATGCTCT__AAGTCAAAATGA_CTTTCCTTAAAACAGAACAATT
mouse   -704 GGCAGTTTTGCAATCCATGCGCT__AAATCAGAATGATCTGACTCTAAAATAGAACAATG
                *    ***  *      *   *  ** *** ** ** *  * *** *****
                          YY1 rev
ovine   -798 ATTCTGAATTTAGTTA_TTAATTTTGTACAGCCAGAATGGTTCA
caprine -794 ATTCTGAATTTAGTTATTTATTTAATTTTGTACAGCCAGAATGGTTCA
bovine  -812 ATTCTGAATTTAGTTATTTATTTAATTTTGTACATCCAGAATGATTCA
rabbit  -685 ATCTCTGATTCAAGAATTTAAC__TATACATCCTAAATGGTT_A
human   -865 ATCTCCAATTCAATAATTTAATAATTTTTTTTCAAAATGGTTCA
mouse   -646 ATTCCTAACTCAGTTACTTAATATTGGGCATCCAAAATGGGACA
                **      *   **  *              *  * ****  **
```

Known transcription factor binding sites and conserved motifs in κ-casein gene promoter region
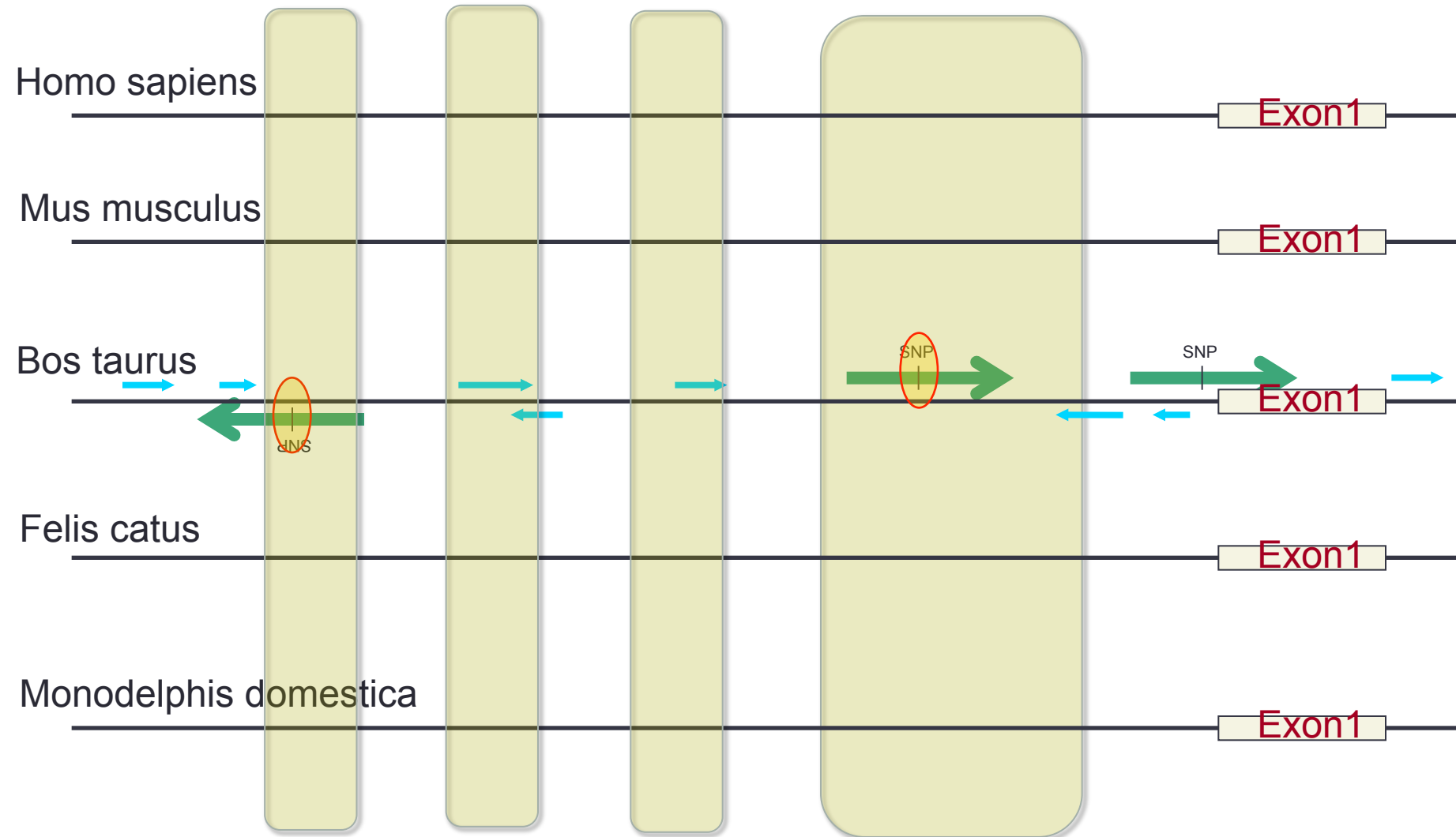
Gerencsér, Á., Barta, E., Boa, S., Kastanis, P., Bősze, Zs. and Whitelaw, C.B.A. (2002) Comparative analysis on the structural features of the 5' flanking region of κ-casein genes from six different species. *Genet. Se. Evol.* **34**:117-128

MGF (Mammary Gland Factor) binding site:

TTCTRAGAA

How to find rSNPs?

Conserved regions

Homo sapiens
　　　　　　　　　　　　　　　　　　　　　　Exon1

Mus musculus
　　　　　　　　　　　　　　　　　　　　　　Exon1

Bos taurus
　　　　　　　　　　　　　SNP　　　　SNP
　　　　　　　　　　　　　　　　　　　　　　Exon1
SNP

Felis catus
　　　　　　　　　　　　　　　　　　　　　　Exon1

Monodelphis domestica
　　　　　　　　　　　　　　　　　　　　　　Exon1

Implementation (REGULOMIX project)

1. Getting bovine promoter sequences
2. Getting bovine SNPs
3. Mapping known TFBSs and SNPs into the promoter sequences
4. Finding SNPs that are inside a mapped TFBS
5. Determining the conservation of the given TFBS
6. [Provide a list of potential rSNPs with annotation](#)
7. [Genotyping, experimental validations](#)

Result (3072 potential SNPs)

SRACAGGTGKYG	consensus
GGACAGRTGGTG	SNPmotif
GGACAGGTGGTG	bosTau4
GGACAGGTAGTG	hg18:3197897-3197908
GGACAGGTAGTG	panTro2
GGACAGGTAGTG	ponAbe2
GGACAGGTAGTG	rheMac2
GGACAGGAGGTA	mm9
GGACAGGAGGTG	rn4
GGAGAGGTGGTG	cavPor3
GGACAGGTGGTG	equCab2
GGACAGGTGGTG	canFam2
GGACAGAAAGTC	monDom4

RXFP3,
V$E2F_03

Complement C4 Precursor,
V$MYOD_01

|	
TTTSGCGCGMNR	consensus
TAGCGYGCCACG	SNPmotif
TAGCGCGCCACG	bosTau4
TAGCGCGTCACA	hg18:33973132-33973143
TAGCGCGTCACA	panTro2
TAGCGCGTCACA	ponAbe2
TAGCGTGTCACA	rheMac2
TAGCGCGCCACG	mm9
TAGCGCGCCACA	rn4
TAGCGCGCCACA	cavPor3
TAGCGCGCCACG	ochPri2
TATCGCGCCACG	equCab2
TAGCGCGCCACG	felCat3
TAGCGCGCCACG	canFam2
TAGCGTGCTACA	loxAfr2
TAGCGCGCCACG	dasNov2
TAGCGCGCGATA	monDom4

Limitations of the method

- We used only the annotated promoter regions in the search
- The TFBS database was limited
- The TFBS search gives a lot of false positives and negatives

- TFBSs might also occur in non-conserved regions

Solution:
    ChIP-seq shows genome-wide where are the TFBSs

ChIP-seq (Chromatin immunoprecipitation)
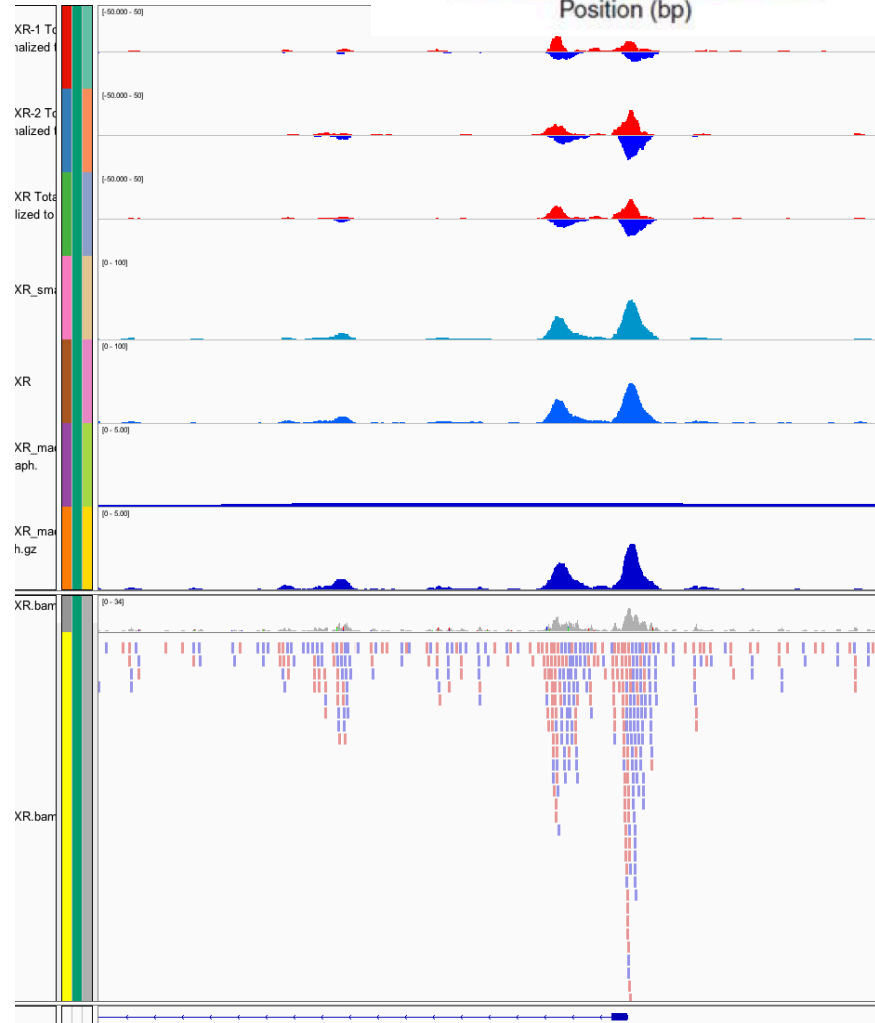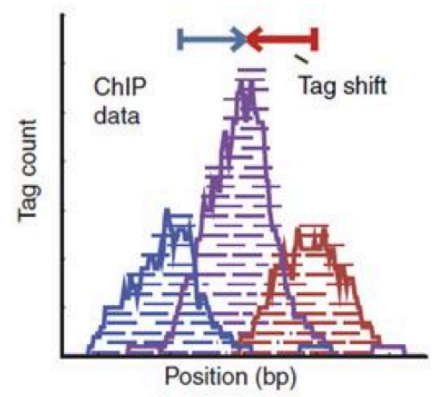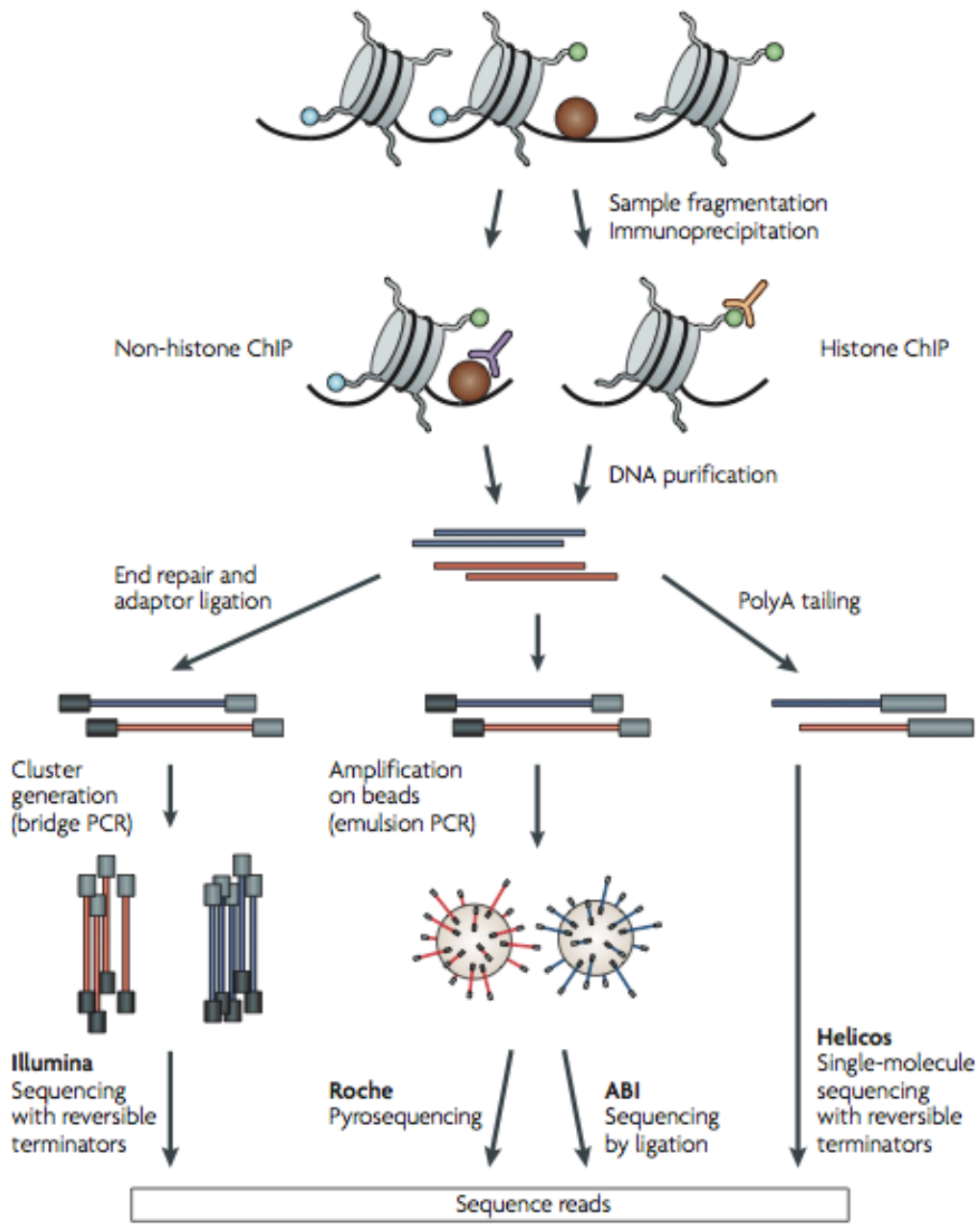
Different TFBSs are active in different tissues
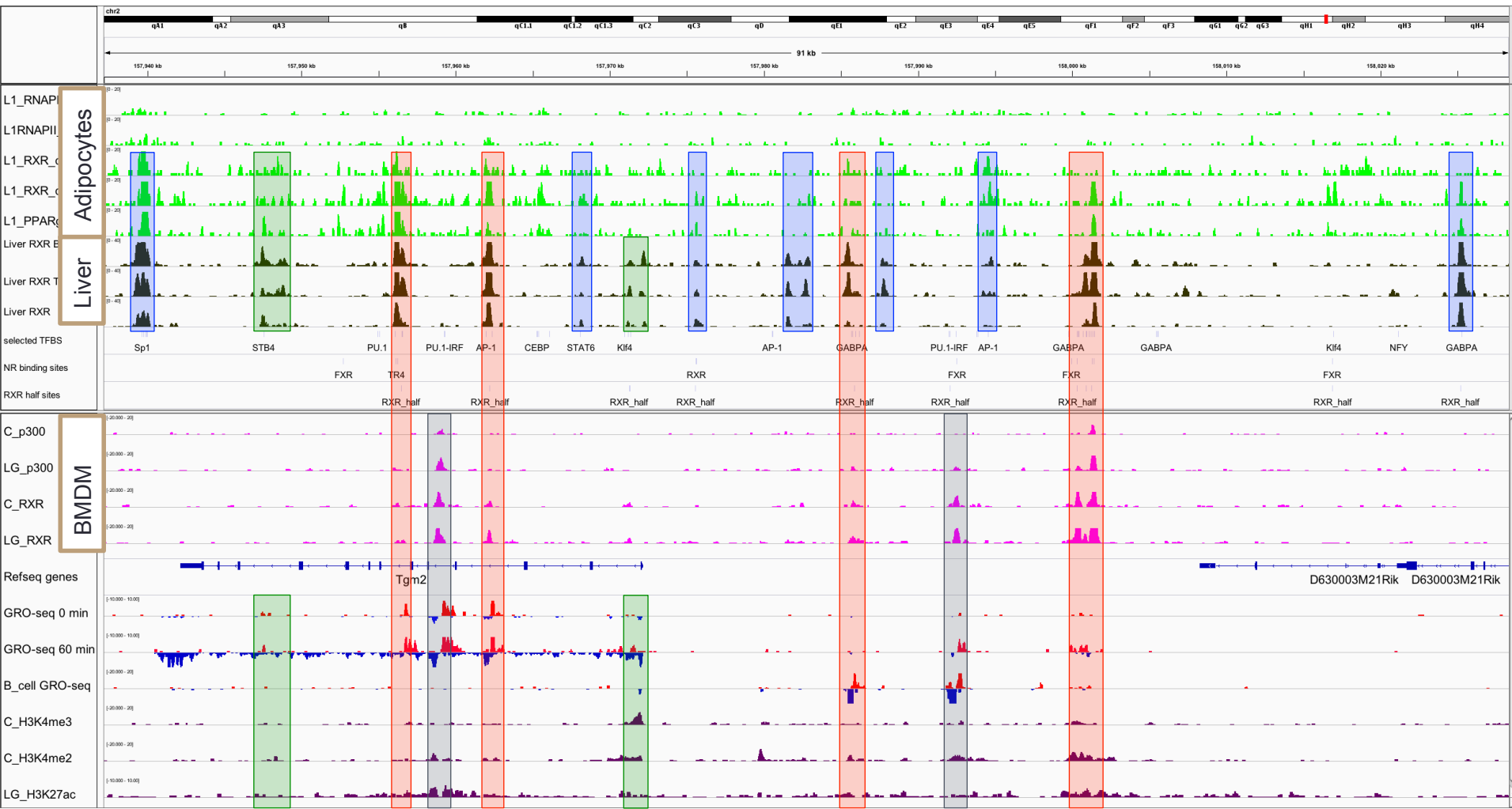
Goals

- Creating a ChIP-Seq peak database based on public human and mouse ChIP-seq data
- Creating a *de novo* motif database
- Investigating SNPs found in motifs
- Exploring the processes affecting on TFBS evolution

- Comparing human and mouse results
- Finding possible phenotype for these rSNP

```
[barta@ngsdeb TFBSlists]$ cat hs_TFBS.lst
hs_MonocyteDerived_macrophage_normal_PU1_n        ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/litesra/SRX/SRX093/SRX093189        factor
hs_MultipleMyeloma_MM1S_cancer_Med1_DMSO          ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/litesra/SRX/SRX129/SRX129117        factor
hs_MultipleMyeloma_PlasmaCell_cancer_Cdk9_DMSO    ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX203/SRX203391        factor
hs_myoblast_myoblast_normal_DUX4_DUX4lentivirus   ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/litesra/SRX/SRX107/SRX107284        factor
hs_NeuralCrest_hNCC_normal_NR2F1_fromH9hESC       ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX059/SRX059367        factor
[barta@ngsdeb TFBSlists]$
```

NCBI-SRA database → **BWA (Burrows-Wheeler Alignment Tool)**
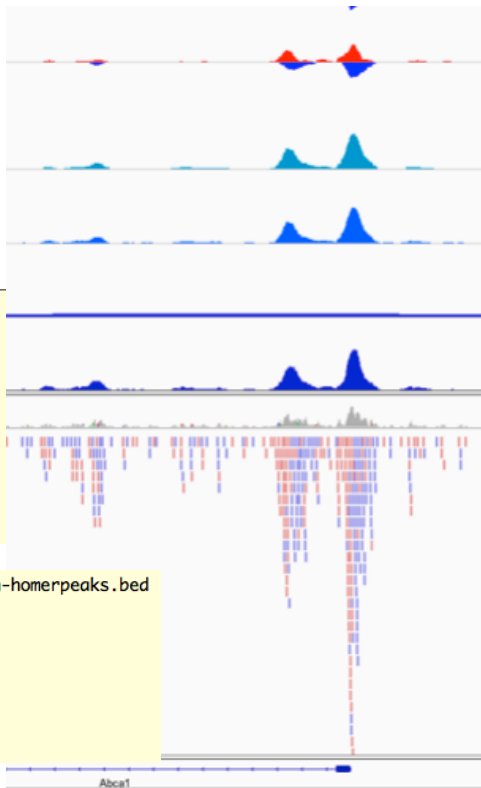
Short sequence reads
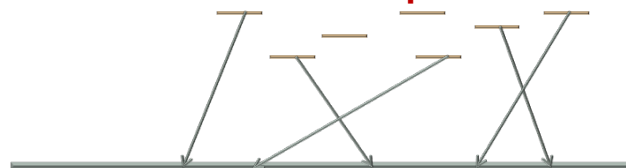
Reference genome

↓

HOMER findPeaks or MACS

peak

Reference genome

Short sequence reads

```
@HWI-H148:101:C0PPFACXX:2:1101:2058:1997 1:N:0:
GTCATGGAATATGGCGAGAAAACTGAAAATCACGGAAAATGAGAAATACACACTT
+
CCCFFFFFHHHGHJJJJJJJJJJJJJJJJJIGIIIIIJGIJJFHGGHIJFHIJJJJJJJ
@HWI-H148:101:C0PPFACXX:2:1101:2546:1986 1:N:0:
GTNAGATTTCCTGTCCAGTATATGGGTTATGGAGTTCTGGAGTTCTACAGAATTC
+
;=#4AD;D?DFDFIFF<EA?IHIC@DAFG<CGCC?CGBBF?F9:?B??DGBGFFE
```

```
[barta@ngsdeb bed]$ head -20 hs_MonocyteDerived_macrophage_normal_PU1_n-homerpeaks.bed
chr1    121485114    121485301    chr1-3    4495.5    +
chr1    121484665    121484852    chr1-2    1548.7    +
chr7    61969267     61969454     chr7-1    796.2     +
chr1    145030110    145030297    chr1-1    532.0     +
chr19   27738400     27738587     chr19-21  505.2     +
chr19   27732632     27732819     chr19-11  500.8     +
chr10   42385094     42385281     chr10-76  477.1     +
```
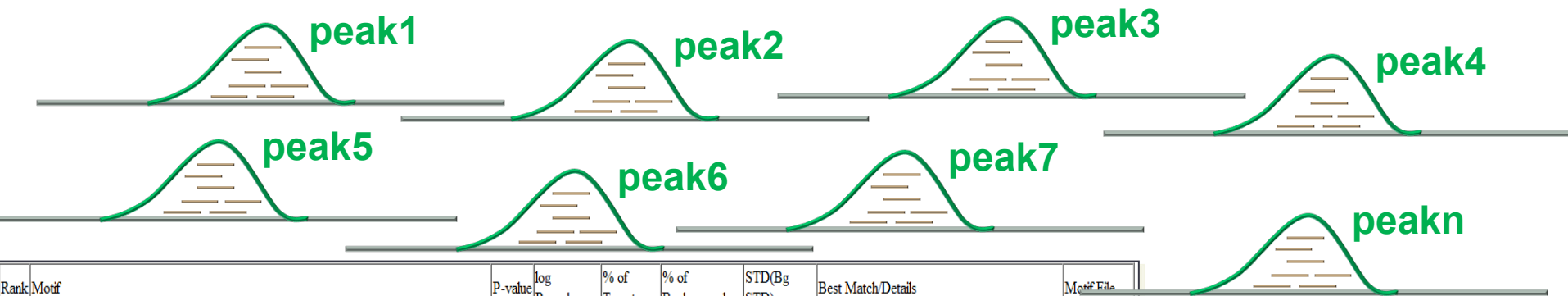
Abca1

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	TGACCTT	1e-771	-1.777e+03	76.95%	3.42%	17.9bp (38.8bp)	Esrrb(NR)/mES-Esrrb-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2	CCTTCCCCCA	1e-15	-3.568e+01	3.96%	0.66%	24.3bp (29.2bp)	MA0079.2_SP1/Jaspar More Information | Similar Motifs Found	motif file (matrix)
3	AAGGACGCATCCTT	1e-13	-3.035e+01	0.70%	0.00%	21.9bp (0.0bp)	MA0136.1_ELF5/Jaspar More Information | Similar Motifs Found	motif file (matrix)
4	GACATTCCAG	1e-12	-2.889e+01	4.31%	0.99%	29.8bp (30.7bp)	TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)

HOMER de novo motif finding

re-mapping de novo matrices to peaks (HOMER annotatePeaks

Reference genome

(a)

raw sequence reads

reference genome

peak region

de novo TFBSs

G G G C G C c G C C A T T T
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

>AAVATGGCGGCGCCC_rev   4-
AAVATGGCGGCGCCC,BestGuess:YY1(Zf)/
Promoter/Homer(0.866)   14.4488
-71.552202   0   T:27.0(2.99%),B:
36.7(0.09%),P:1e-31

	A	C	G	T
1	A	C	G	T
2	0.001	0.083	0.915	0.001
3	0.168	0.084	0.664	0.084
4	0.236	0.001	0.762	0.001
5	0.001	0.831	0.084	0.084
6	0.083	0.001	0.915	0.001
7	0.001	0.913	0.085	0.001
8	0.002	0.844	0.070	0.084
9	0.001	0.001	0.997	0.001
10	0.001	0.997	0.001	0.001
11	0.001	0.997	0.001	0.001
12	0.997	0.001	0.001	0.001
13	0.001	0.001	0.001	0.997
14	0.001	0.394	0.354	0.252
15	0.001	0.002	0.084	0.913
	0.001	0.001	0.083	0.915

occurence of the base in a motif position

SNP from dbSNP (**G/A**)

motif matrix

bin1   bin10

(b)

SNPs from dbSNP

Comparing with other species' bases in the same position

Comparing with GWAS SNPs

From multiz whole genome alignments

hg19	G
panTro2	G
gorGor1	G
ponAbe2	A
rheMac2	A
papHam1	C

SNP table containing metadata e.g. orthologous bases, motif matrix base frequencies

Orthologous SNP groups according to conservation

evolutionary observations

Pipeline of ChIP-Seq analysis

Reads / tags
(fastq format)

BWA — Burrows-Wheeler Alignment Tool

Hypergeometric Optimization of Motif EnRichment

Alignment files (bam)

Homer
makeTagdirectory
makeUCSCfile
findMotifsGenome
annotatePeaks

MACS2 — Model Based Analysis for ChIP-Seqdata

findPeaks(bed)
bedgraphs
motif logos
meta-
histograms
/ metagenes

bam   bed   bed

Differential Binding Analysis of ChIP-Seqpeak data

bash, bedtools awk, mergeBed

intersectBed, awk, sed...

bed   rpkm-like table   bed

DiffBind
edgeR, limma

intersectBed, awk, sed...

Chipster
edgeR, limma

+ GRO-Seqtags

+ eTr-s (GRO-Seq)

Functional annotation of binding and modification loci

PeakAnnotator

Clusters, Heatmaps, Plots,
Profiles, Venn diagrams
Peak lists, bed files

Overlaps,
Peak lists,
bed files

Clusters, Heatmaps,
Profiles, Plots,
Peak lists, bed files

Barta E: Command line analysis of ChIP-seq results. EMBNET JOURNAL 17:(1) pp. 13-17. (2011)

Implementation

- Collecting Transcription Factor ChIP-seq experiments from NCBI SRA database *(526 human, 363 mouse)*
- Downloading and processing raw data:
  - Mapping reads to hg19 and mm10 reference genomes
  - Defining peaks
  - Searching and filtering de novo motifs
  - Mapping back de novo motifs to peaks -> TFBS
    (hs: 10 million redundant, 6.6 million non-redundant 1.86 million region, mm: 11, 4.8, 2.2)
  - Searching SNPs in TFBSs -> rSNP *(hs: 805 , mm 671 thousand)* and in background
  - Searching orthologs in rSNPs positions
  - Analysing datas

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	TGTTTACATA	1e-12661	-2.915e+04	70.91%	15.19%	40.5bp (65.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2	CTTGGCAG	1e-578	-1.332e+03	27.14%	16.52%	54.0bp (65.5bp)	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)

```
>TGAGTCAYSS    1-
TGAGTCAYSS,BestGuess:Jun-AP1(bZIP)/K562-
cJun-ChIP-Seq/Homer(0.970)    6.52458
-232.587472    0    T:169.0(21.20%),B:
1158.4(2.47%),P:1e-101
0.001    0.001    0.001    0.997
0.001    0.001    0.941    0.057
0.997    0.001    0.001    0.001
0.001    0.300    0.609    0.090
0.001    0.001    0.001    0.997
0.137    0.861    0.001    0.001
0.997    0.001    0.001    0.001
0.079    0.370    0.144    0.406
0.166    0.363    0.261    0.209
0.186    0.413    0.277    0.124
```
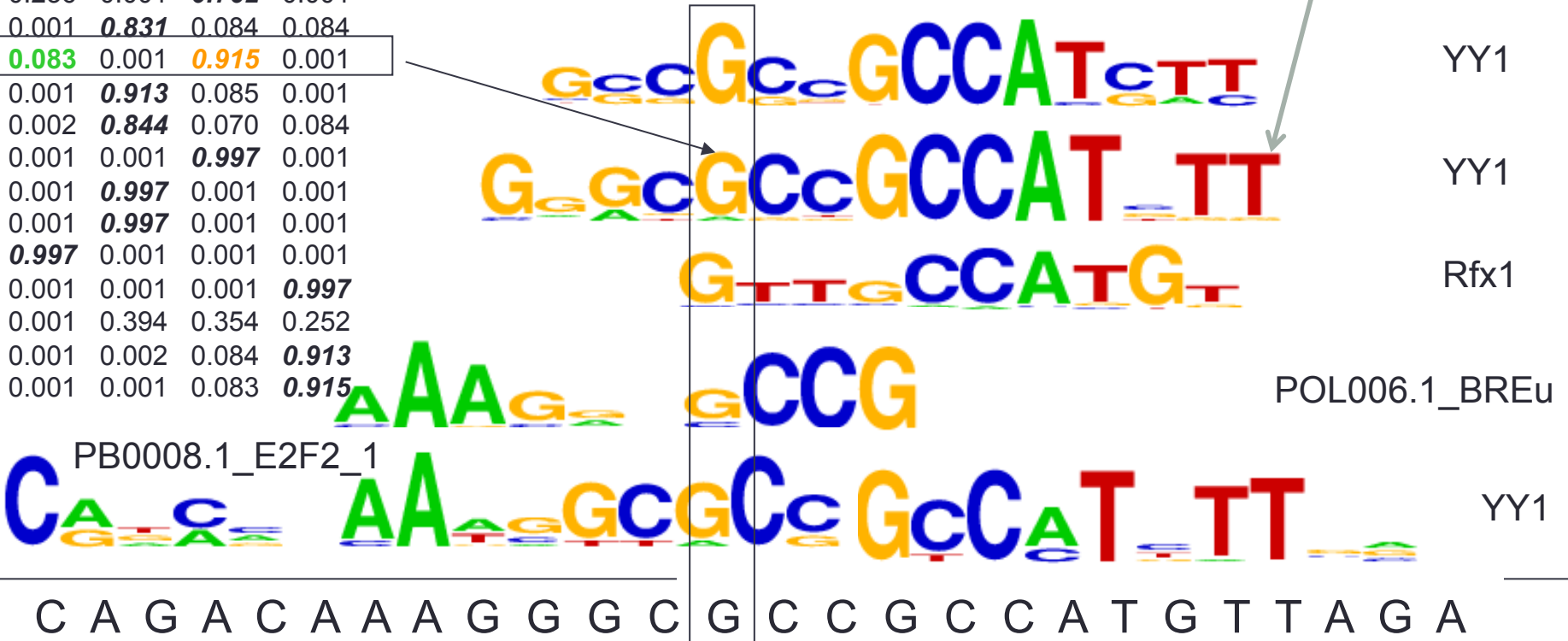
```
probafileok_1.00081_52_5        chr7    122443591       122443606       -       AAGCCATAAATCACA 15      35      /molbio/projects/rSNP/julcsi/
probafileok_1.00081/selectedmotifs_1.00081/rev_opp_motifs/hs_AcuteLymphoblasticLeukaemia_SEM_cancer_HOXA9_n§1_rev.motif
        DATGATTTATGGBTY HOXA9/HSC-Hoxa9-ChIP-Seq(GSE33509)/Homer(0.913)-rev      17.79   0.86    1e-98   9.017108        /molbio/projects/
rSNP/hs_ChIP-seq-anal/analysis/hs_AcuteLymphoblasticLeukaemia_SEM_cancer_HOXA9_n/homer/
hs_AcuteLymphoblasticLeukaemia_SEM_cancer_HOXA9_n_homerpeaks.txt        122443557       122443645       0.679733333333333       0.0428
        0.997   0.001   probafileok_1.00081_52122443595         probafileok_1.00081_52C         T       CCCCCCCCCCCCC-TACCCC-CC-CC-CCCCTT---
        5       0.507   0.191   C       A       0.507   0.270   2       0.237   1       0.246258426966292       0.186933333333333       0.249
```

>AAVATGGCGGCGCCC_rev   4-AAVATGGCGGCGCCC,BestGuess:YY1(Zf)/Promoter/Homer(0.866)   14.4488 -71.552202   0   T:27.0(2.99%),B:36.7(0.09%),P:1e-31
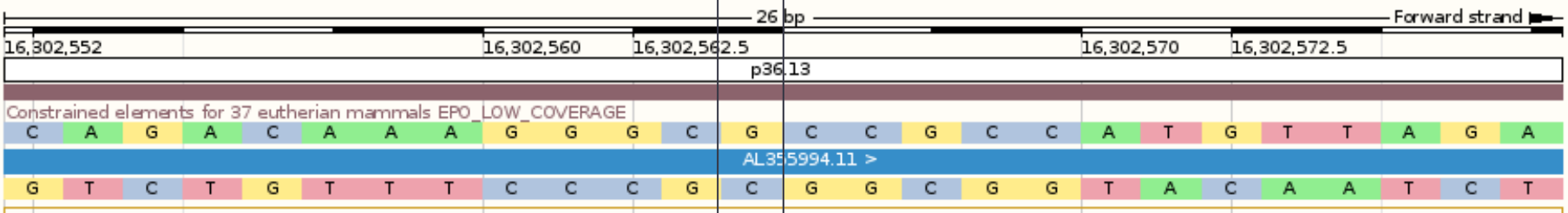
0.001	0.083	*0.915*	0.001
0.168	0.084	*0.664*	0.084
0.236	0.001	*0.762*	0.001
0.001	*0.831*	0.084	0.084
0.083	0.001	*0.915*	0.001
0.001	*0.913*	0.085	0.001
0.002	*0.844*	0.070	0.084
0.001	0.001	*0.997*	0.001
0.001	*0.997*	0.001	0.001
0.001	*0.997*	0.001	0.001
0.997	0.001	0.001	0.001
0.001	0.001	0.001	*0.997*
0.001	0.394	0.354	0.252
0.001	0.002	0.084	*0.913*
0.001	0.001	0.083	*0.915*

hs_lymphoblastoid_GM12878_normal_Mxi1_n

rs149795722 **G**/**A**

YY1

YY1

Rfx1

POL006.1_BREu

PB0008.1_E2F2_1

YY1

C A G A C A A A G G G C G C C G C C A T G T T A G A

chr1

26 bp   Forward strand ▶

16,302,552   16,302,560   16,302,562.5   16,302,570   16,302,572.5

p36.13

Constrained elements for 37 eutherian mammals EPO_LOW_COVERAGE

C A G A C A A A G G G C G C C G C C A T G T T A G A

AL355994.11 >

G T C T G T T T C C C G C G G C G G T A C A A T C T
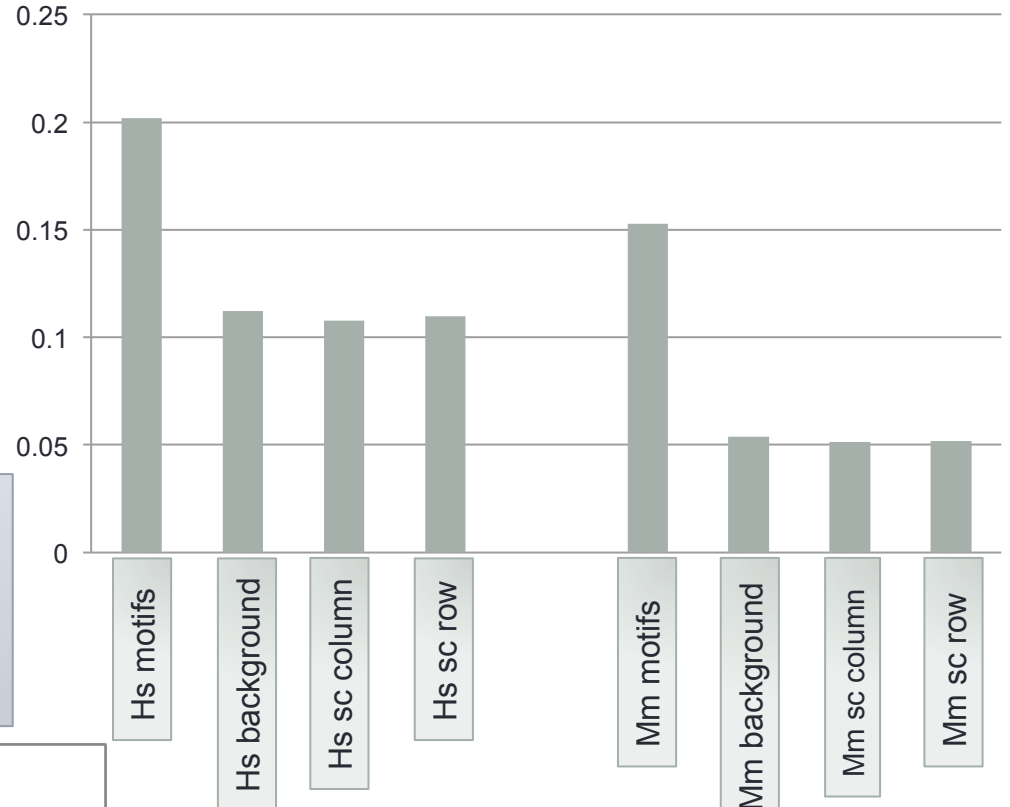
Basic statistics of the pipeline

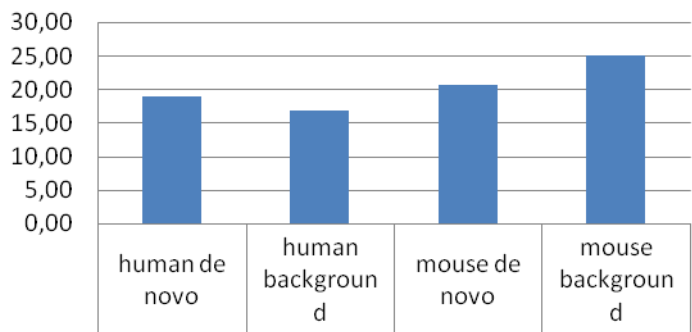|	SRA sample number	Our sample number	sample name	Redundant TFBS number (million)	Redundant TFBS lenght (megabasepair)	merged region number (million)	merged region lenght (megabasepair)	region's occupancy on the genome (human & mouse%)	Redundant SNP number (million)	Unique SNP number (million)	SNP density (SNP number per 1000 bp)	phastcons average
human	8323	480	*human_motif*	10,3	127,1	2,89	42,4	1,3	2,31	0,8	19	0.15
			human_background	18,7	208,8	1,86	41,5	1,3	3,5	0,7	16,9	0.054
mouse	7224	304	*mouse_motif*	11,1	140,5	2,23	32,6	1,2	2,53	0,67	21,5	0.2
			mouse_background	10,2	112,6	1,58	30,7	1,1	2,76	0,77	25,7	0.11

Hypothesis 1: There are less SNPs in TFBSs

Hypothesis 2: TFBSs are highly conserved

Background = 100 basepairs long genomial regions mapping one by one – (exons, repetitive sequences, ChIP-seq peaks)
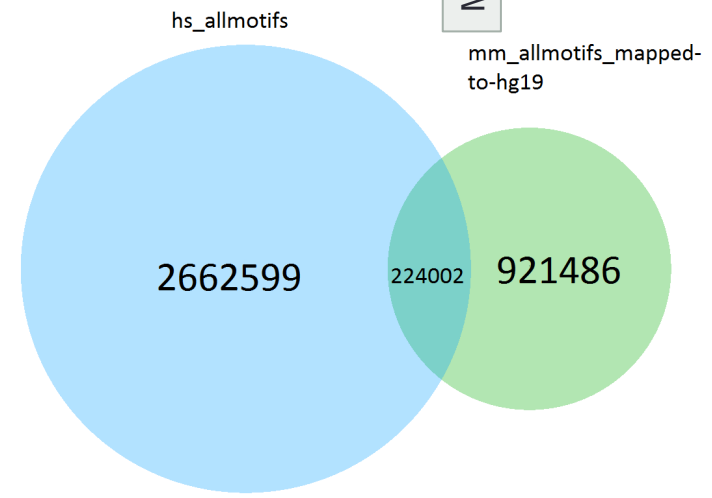
PHASTCONS average values

Number of SNPs / total lenght of regions *1000

	human de novo	human backgroun d	mouse de novo	mouse backgroun d
Number of SNPs / total lenght of regions *1000	19,00	16,82	20,63	24,95

hs_allmotifs

mm_allmotifs_mapped-to-hg19

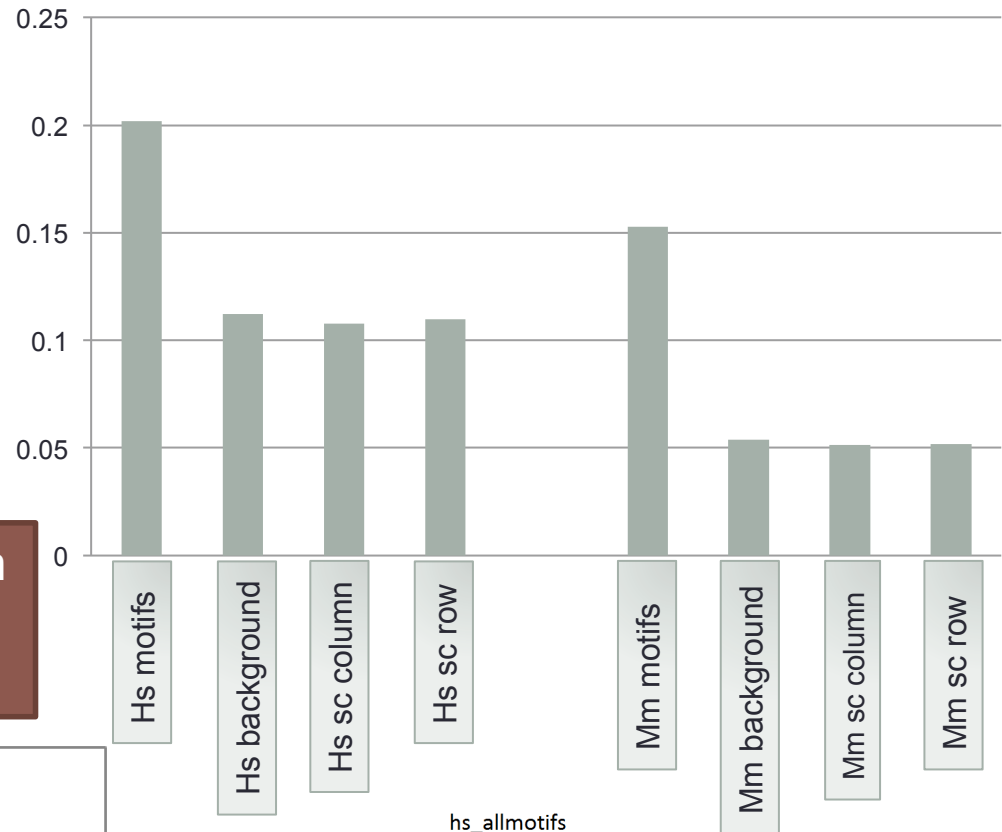

2662599   224002   921486

PHASTCONS average values

Hypothesis 1: There are less SNPs in TFBSs

Hypothesis 2: TFBSs are highly conserved

?
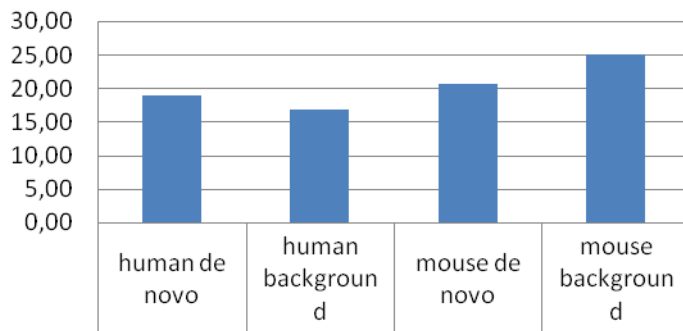
Phastcons values are higher but human and mouse TFBSs overlap barely (<10%)

Number of SNPs / total lenght of regions *1000

	human de novo	human background d	mouse de novo	mouse background d
Number of SNPs / total lenght of regions *1000	19,00	16,82	20,63	24,95

hs_allmotifs

mm_allmotifs_mapped-to-hg19

2662599   224002   921486

SNP: A/G   SNP: A/C

type1	type2	type6	type7	type8	type9	type11	
A	A	A	A	A	A	A	HG
A					G	T	
A					G	T	PRIMATES
A					A	A	
A					A	A	
A					A	T	
A					G	T	
A	G	A	C	A	C	T	OTHER MAMMALS
A	G	G	C	A	T	A	
A	G	A	C	G	T	A	
A	G	A	A	G	G	T	
A	G	G	A	G	G	T	

Type 1:
- The major allele is totally conserved in all of the species
- There is a strong evolutional pressure in the given position
- The fact that an other allele also can be found in human population may be important from functional aspects

SNP: A/G   SNP: A/C

type1   type2   type6   type7   type8   type9   type11

HG

PRIMATES

OTHER MAMMALS

Type 2:
- The minor allele is conserved in other species
- Probably the minor allele is the determinant in a given position of the TFBS matrix
- The fact that the null allele is fixed in human may denote that the TFBS is not active

SNP: A/G   SNP: A/C

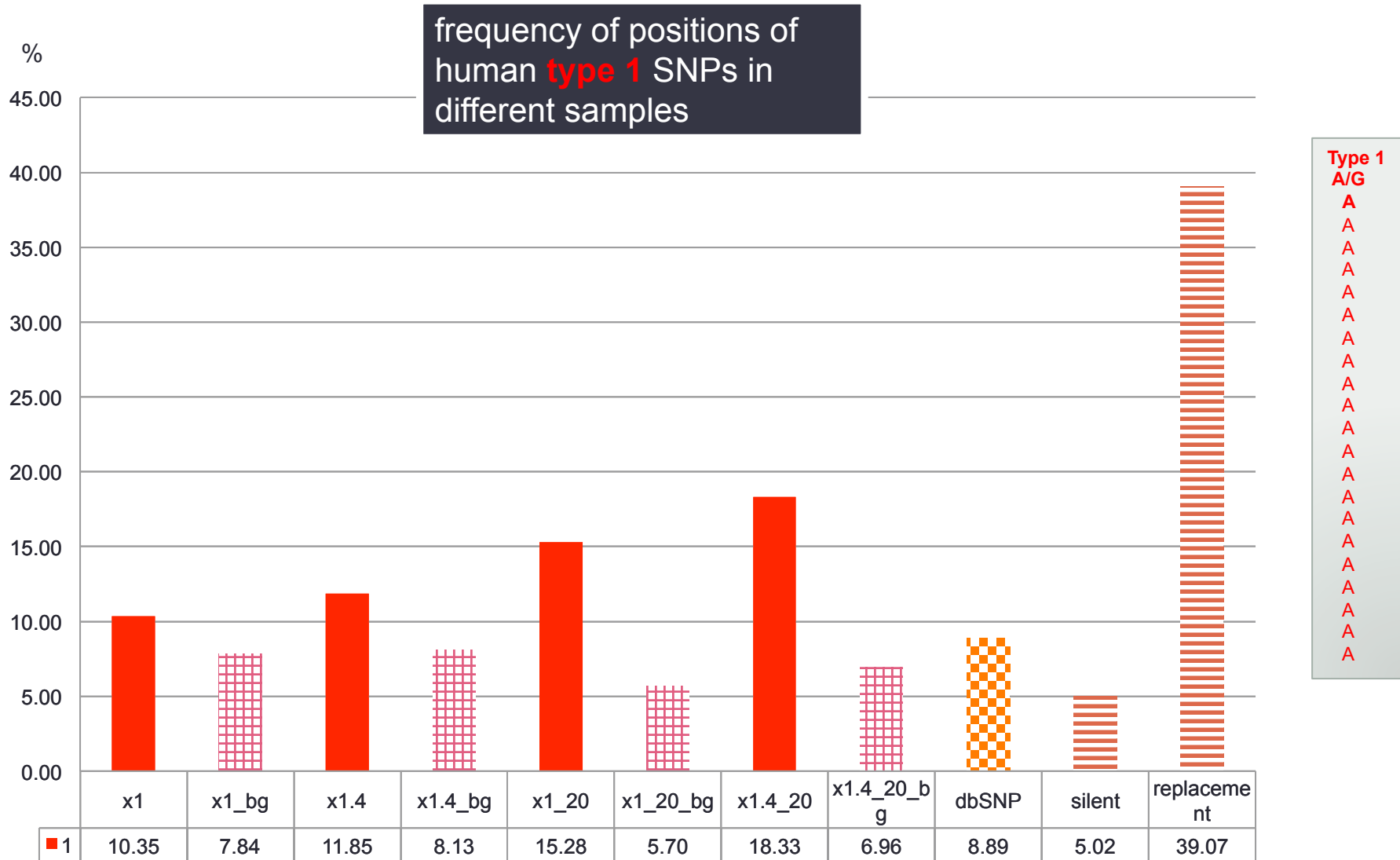| type1	type2	type6	type7	type8	type9	type11	
A	A	A	A	A	A	A	HG
A	G	A	C	G	G	T	
A	G	G	C	T			
A	G	A	C	A			
A	G	G	A	G			
A	G	G	A	G			
A	G	A	A	G			
A	G	A	C	A			
A	G	G	C	A			
A	G	A	C	G			
A	G	A	A	G	G	T	
A	G	G	A	G	G	T	OTHE

Type 6 (transition), 7 (transversion):
- The major and minor alleles are also conserved in other species
- It may mean that:
  - It is only important in a given position to be purin or pirimidin
  - Both of the two alleles may have important role, that's why they are both fixed

Hypothesis 3: In TFBSs there are more SNPs in conserved positions

frequency of positions of human **type 1** SNPs in different samples

	x1	x1_bg	x1.4	x1.4_bg	x1_20	x1_20_bg	x1.4_20	x1.4_20_bg	dbSNP	silent	replacement
1	10.35	7.84	11.85	8.13	15.28	5.70	18.33	6.96	8.89	5.02	39.07

Type 1 A/G
A A A A A A A A A A A A A A A A A A A A A

x1, x1.4: stringency of motif searching
1: at least 4 non-primate orthologs, 20: at least 20

Hypothesis 4: The bases important in transcription factor binding are more conserved evolutionarily (type 1)

human de novo motifs, **major** allele frequencies in ChIP-seq peaks and background

	0_0.1	0.1_0.2	0.2_0.3	0.3_0.4	0.4_0.5	0.5_0.6	0.6_0.7	0.7_0.8	0.8_0.9	0.9_1
mot_ma1	8.17	8.96	9.18	9.87	10.82	12.32	11.48	14.43	15.79	17.91
mot_ma2	1.05	0.61	0.44	0.39	0.42	0.43	0.52	0.34	0.31	0.23
bg_ma1	7.88	8.11	8.22	7.68	8.04	7.84	8.21	8.40	8.44	8.64
bg_ma2	0.52	0.53	0.53	0.56	0.50	0.58	0.59	0.60	0.58	0.54

Type 1 A/G	Type2 A/G
A	A
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G
A	G

Hypothesis 5: Selection can be against the null (G) allele (type 1) = e.g. there is A in every species, the G is less than 0.1 in TFBS matrix, however G occures in the population in humans

Hypothesis 6: In some TFBSs the null allele (type 2) is more frequent= e.g. there is G in every species, the G > 0.9 in TFBS matrix, but A is more frequent in human

human de novo motifs, **minor** allele frequencies in ChIP-seq peaks and background

	Type 1 A/G	Type2 A/G
	A	A
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G
	A	G

Type 1

Type 2

	0_0.1	0.1_0.2	0.2_0.3	0.3_0.4	0.4_0.5	0.5_0.6	0.6_0.7	0.7_0.8	0.8_0.9	0.9_1
mot_mi1	16.35	10.85	8.98	9.53	9.15	9.97	8.87	8.40	7.96	6.36
mot_mi2	0.26	0.38	0.50	0.51	0.59	0.68	0.85	1.35	1.83	2.95
bg_mi1	8.65	7.64	7.65	7.83	7.71	8.33	8.47	8.95	9.32	8.61
bg_mi2	0.55	0.58	0.59	0.53	0.56	0.44	0.42	0.48	0.42	0.52

Hypothesis: If both purin or both pirimidin bases are equally important in the TFBS (A >0.45 & G>0.45), then both alleles can be conserved evolutionarily
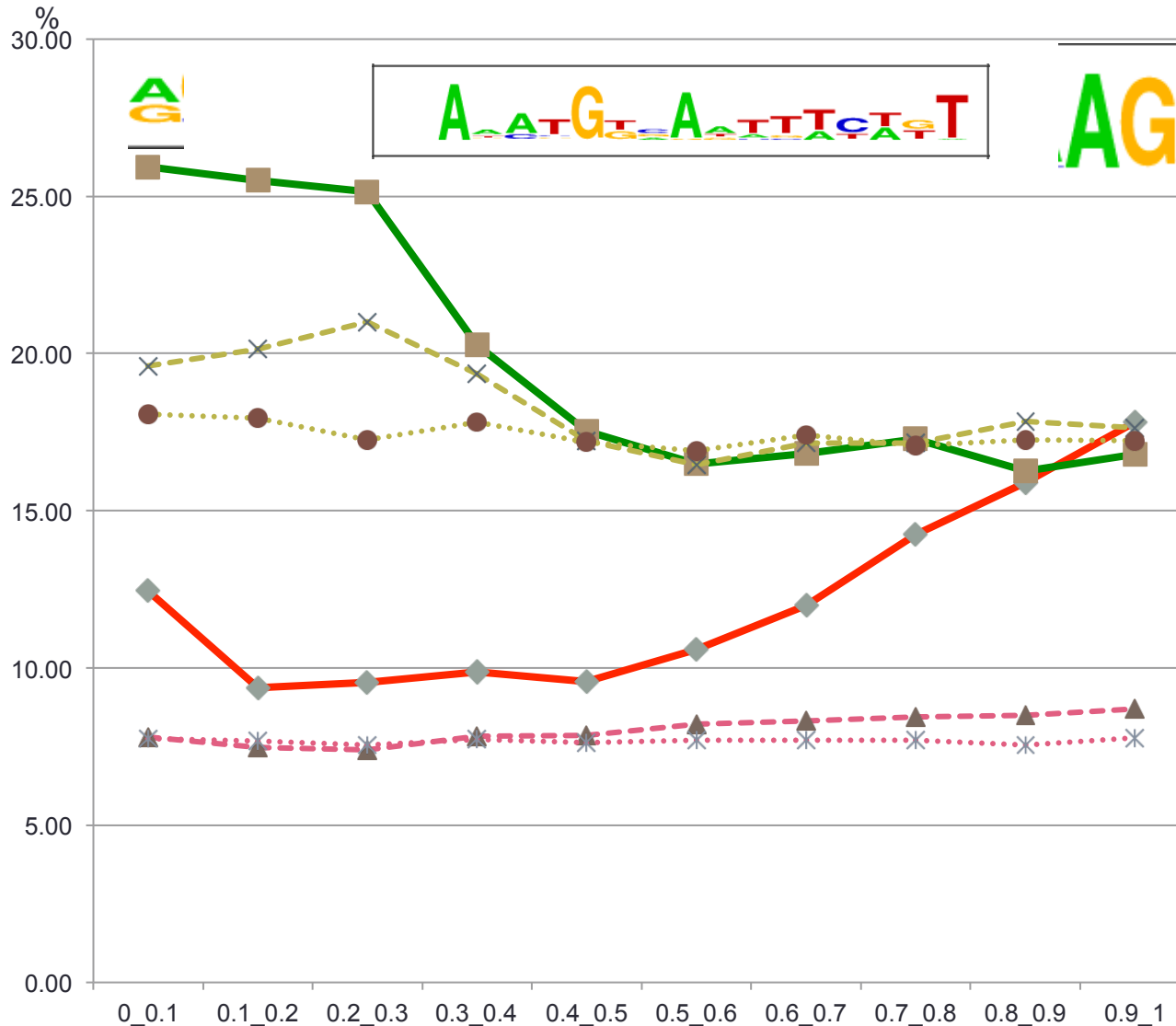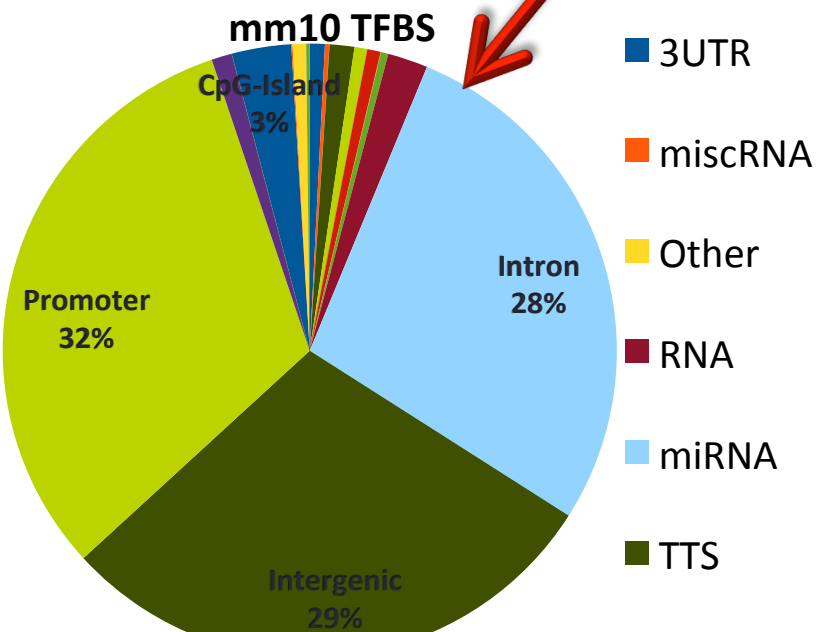
human de novo and scrambled motifs, **abs(0.5-major)+abs(0.5-minor)** allele frequencies in ChIP-seq peaks and background

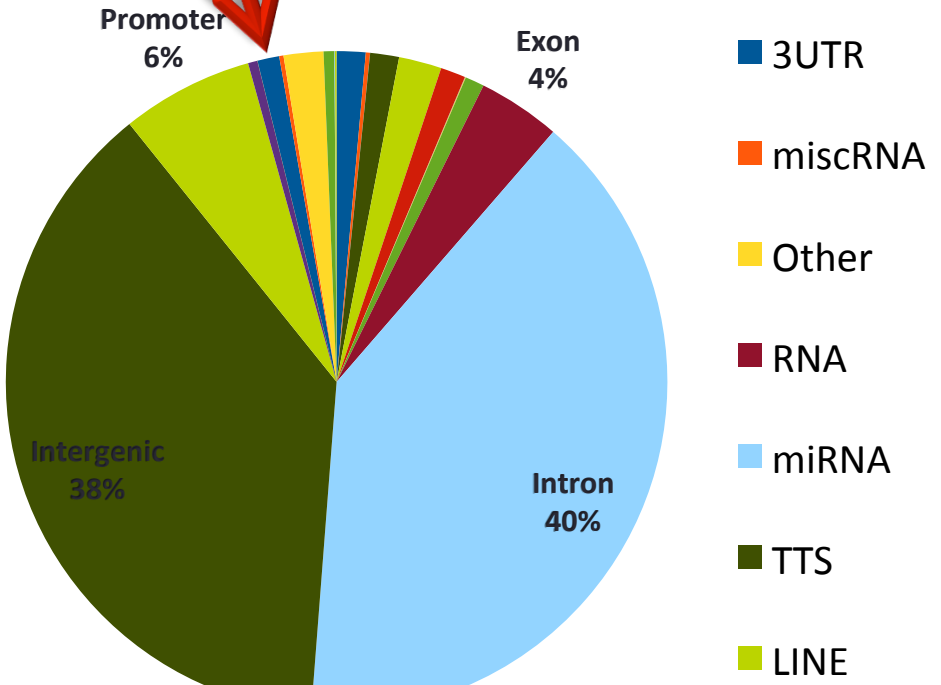Motif control (scrambled): we mixed the allele frequencies in the selected de novo motif matrices

mm_allmotifs_TFBS

hs_allmotifs_TFBS
mapped-to-mm10

2006746

224048

1007169
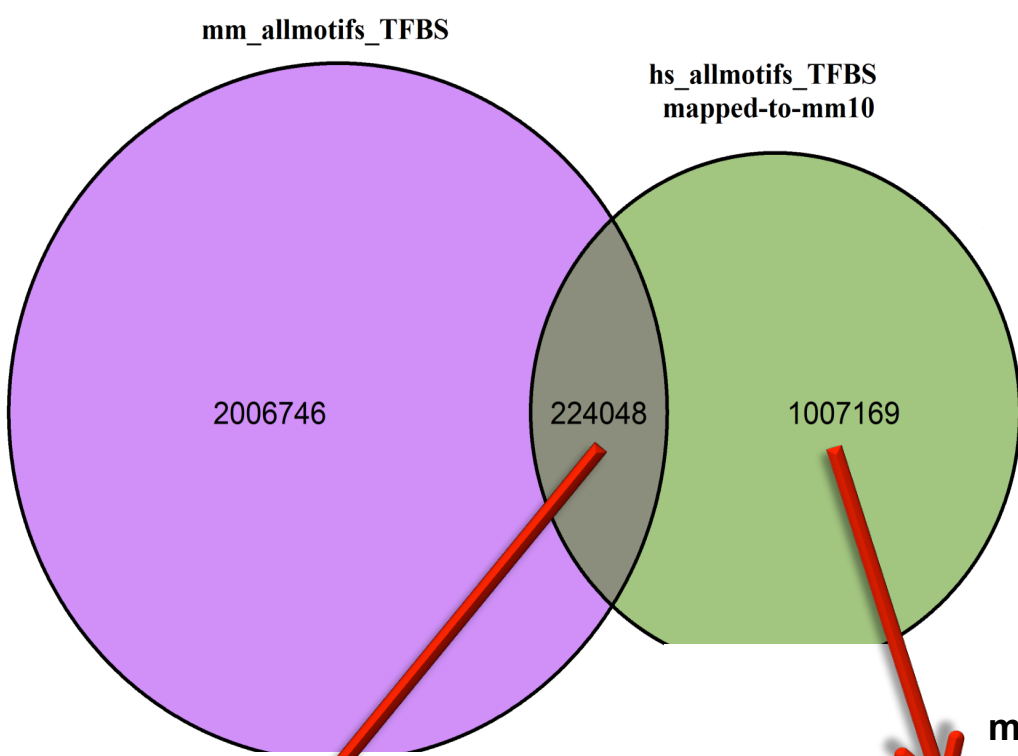
50% of the human TFBS regions can be mapped to the mouse genome using liftOver

hg19 mapped to mm10 overlap with mm10 TFBS

CpG-Island 3%

Promoter 32%

Intron 28%

Intergenic 29%

- 3UTR
- miscRNA
- Other
- RNA
- miRNA
- TTS

hs_allmotif_mapped-to-mm10_no_overlap_with_mm10_TFBS

Promoter 6%

Exon 4%

Intergenic 38%

Intron 40%

- 3UTR
- miscRNA
- Other
- RNA
- miRNA
- TTS
- LINE

SNP statistics

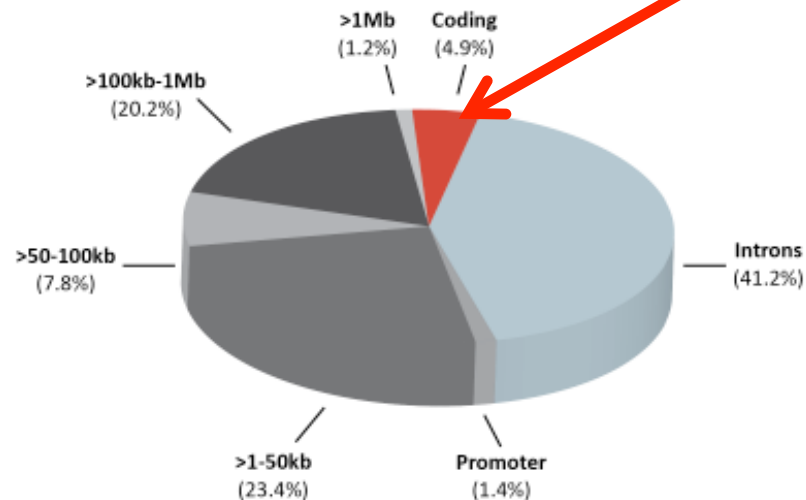|	human		mouse	
	de novo	background	de novo	background
number of regions	2.9 million	1.9 million	2.2 million	1.6 miilion
total length of regions	42.4 million bps	41.5 million bps	32.6 million bps	30.7 miilion bps
total length of peaks	793.4 million bps	64.5 million bps	632.1 miilion bps	55.5 million bps
number of SNPs	806 thousand	698 thousand	672 thousand	766 thousand
number of SNPs / total length of regions *1000	19.00	16.82	20.63	24.95
number of regions / total length of peaks *1000	3.64	28.88	3.53	28.40
number of SNPs found in non-coding gwas	129	101		

A

920 Genome Wide Association Studies

- Parasitic/bacterial Disease 1%
- Neurological/behavioral 13%
- Quantitative traits 7%
- Radiographic parameters 1%
- Miscellaneous 7%
- Serum Metabolites 11%
- Lipid 3%
- Viral Disease 2%
- Kidney, Lung, Liver 3%
- Aging 9%
- Hematologic parameters 3%
- Drug metabolism 4%
- Autoimmune disease 12%
- Diabetes 4%
- Cardiovascular 9%
- Cancer 12%
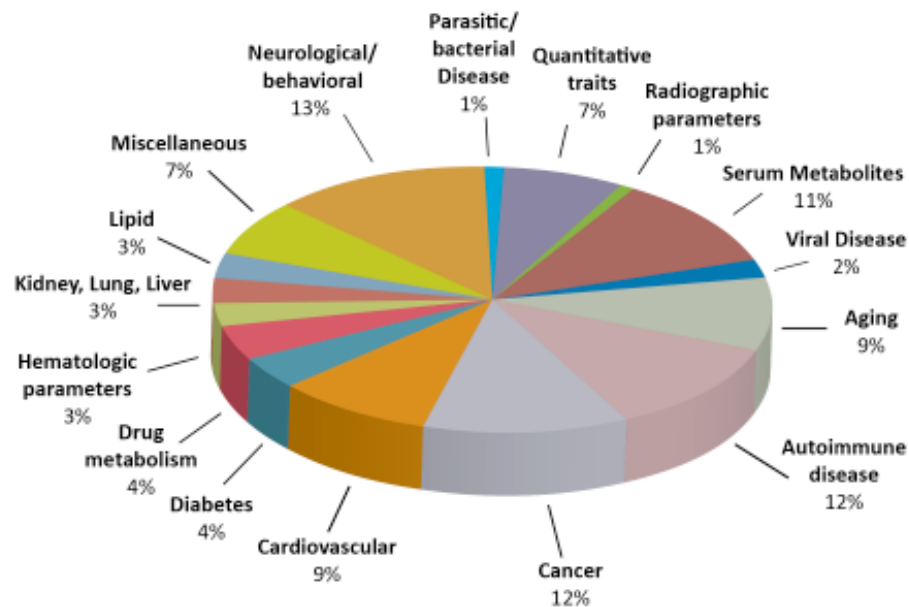
B

Distribution of GWAS SNPs vs. RefSeq

- >1Mb (1.2%)
- Coding (4.9%)
- >100kb-1Mb (20.2%)
- >50-100kb (7.8%)
- Introns (41.2%)
- >1-50kb (23.4%)
- Promoter (1.4%)

Matthew T. Maurano et al.: **Systematic Localization of Common Disease-Associated Variation in Regulatory DNA**
Science. 2012 September 7; 337(6099): 1190–1195.

rs285480

(A;A) (A;G) (G;G)

http://www.snpedia.com/index.php/Rs285480

Scale (phenotype)	SNP	Location	Gene	Unadjusted *P* value	EMP1	BONF
ZKPQ-Sensation Seeking	rs285480	Chromosome 1q23.3	*RXRG*	3.83e-7	5.094e-6	NS
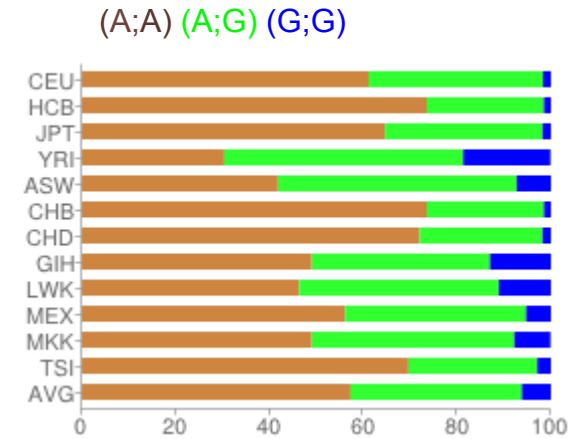
ZKPQ, Zuckerman–Kuhlman Personality Questionnaire
Unadjusted *P*, unadjusted (nominal) genome-wide association study *P* value
BONF, EMP1 *P* value after Bonferroni correction for 702 866 single nucleotide polymorphisms (SNPs);
EMP1, empirical significance levels, adaptive permutation-calculated *P* value (max permutation =100 000 000 000)
NS, not significant

Ney Alliey-Rodriguez et al.: **Genome-wide association study of personality traits in bipolar patients**
Psychiatr Genet. 2011 August; 21(4): 190–194. doi: 10.1097/YPG.0b013e3283457a31

name	In_DHS	LOG(P)	sample_size	DNA_region	gene
rs285480	Y	6.39794	1007	intronic	RXRG

primates
GGGGAAGGGG
other mammals
GGGGA-AC-GGGAGG---A-A-----

hs_EmbryonicKidney_HEK293TREx_normal_ZNF263_n§1.motif

>**G**GGGAGSACYS   1-GGGAGSACYS,BestGuess:Znf263(Zf)/K562-Znf263-ChIP-Seq/Homer(0.964)
8.19068 -452.343749   0   T: 276.0(32.32%),B:1345.7(2.99%),P: 1e-196

	A	C	G	T
G	**0.172**	0.163	***0.634***	0.031
G	0.110	0.035	***0.782***	0.073
G	0.005	0.006	***0.988***	0.001
A	***0.960***	0.023	0.016	0.001
G	0.001	0.001	***0.997***	0.001
C/G	0.001	***0.418***	0.379	0.203
A	***0.828***	0.034	0.137	0.001
C	0.003	***0.862***	0.126	0.009
T/C	0.183	***0.267***	0.145	***0.405***
G/C	0.143	***0.340***	***0.372***	0.144

Summary

- We have compiled and analyzed a human and a mouse comprehensive TFBS database.
- We have shown special evolutionary characteristics at the TFBSs

- The database is suitable for further studies
- A web interface is under development

Thank you for your attention!

- REGULOMIX consortium
- **Júlia Koller**
- **Erik Czipa**
- Dávid Jónás
- Gergely Nagy

- Tibor Nagy
- Levente Szabadkai