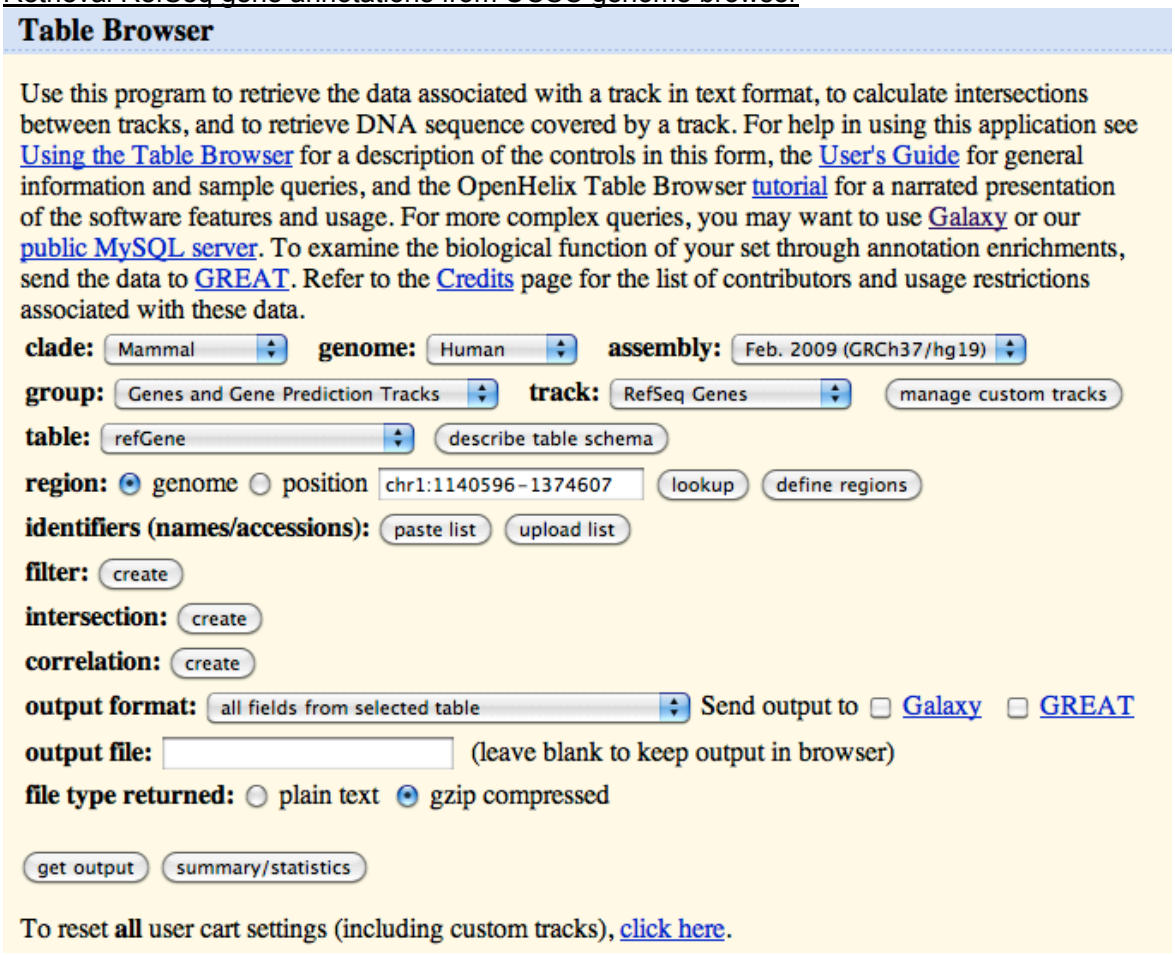
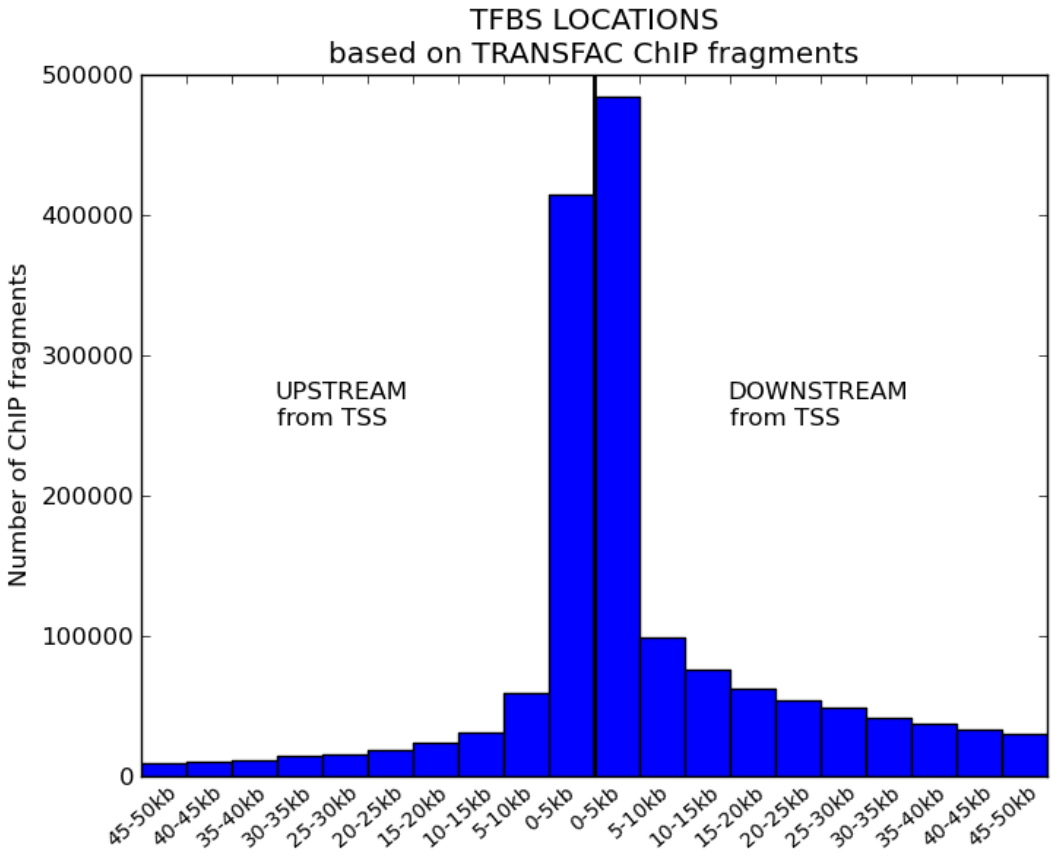
Retrieval RefSeq gene annotations from UCSC genome browser



This table contains 22,285 unique HGNC names. Not all protein-coding genes are present: 509 names of genes that code for protein are not included in the RefSeq table.

TFBS location analysis:



Conclusion:

1. Upstream interval = [TSS+10kb, TSS[- [Transcript interval of nearest upstream gene] - (for all transcripts that overlap with [TSS+10kb, TSS[: union of their [exons in CDS])
2. Do not work with 5'UTR and 1st intron addition to upstream interval, but work with an interval for the downstream region too. Downstream interval = [TSS, TSS-25kb] - [Transcript interval of nearest downstream gene] - (for all transcripts that overlap with [TSS, TSS-25kb]: union of their [exons in CDS])

Algorithm:
Prerequisites: UCSC RefSeq table -> SQL-table (e.g. SQLITE3) with index on chromosome and position interval
Main loop: for each transcript T in the table:

- Create the interval Ireg = [TSS+10kb, TSS-25kb]
- Get all transcripts that overlap with Ireg and determine Sexons = (all [exons in CDS] intervals)
- Get transcript intervals for nearest upstream and downstream gene = Stranscripts
- Regulatory region for T = Rreg,T = Ireg - (Sexons U Stranscripts)

Special transcripts: Some RefSeq IDs map to two or more chromosomal positions (sometimes not even in each others proximity but on a different chromosome): for these transcripts determination of a regulatory region is impossible. To avoid losing to many transcripts:

1. Make sure the set of all transcripts associated with a ID don't form an interval when combined. If so, the transcripts can still be used.
2. Some of the transcripts map to special chromosomes (chromosomes with an

underscore in their name): these can be omitted before trying step 1.

Prerequisites: UCSC RefSeq table -> SQL-table (e.g. SQLITE3) with index on chromosome, strand and position interval. But also on refSeqID.

Main loop: for each refSeqID in the table:

1. Retrieve all transcripts associated with ID, most of the times there will only be one transcript. If there are more find out if these regions can be merged (first with all locations, if that is not possible only with normal chromosome locations). If not skip this ID.
2. Create the interval Ireg = [TSS+10kb, TSS-25kb]. Keep strandedness in mind.
3. Get all transcripts that overlap with Ireg and determine Sexons = (all [exons in CDS] intervals). Strandedness is not important.
4. Get transcript intervals for nearest upstream and downstream gene = Stranscripts. Strandedness is again not important. BUT extend downstream interval by regulatory region. For search of upstream and downstream gene the strandedness is important, but because we need both this requirement can be dropped.
BUT: these regions must be infinitely extended to one side depending on there streamyness.
5. Regulatory region for T = Rreg,T = Ireg - (Sexons U Stranscripts)

Post processing:
Group regions by HGNC name and try to reduce the number of intervals to a minimum by merging overlapping intervals.

Build in progress monitoring!

Procedure via BED-tools:
For defining the regions and their limit RefSeq IDs (only protein-coding transcripts with prefix NM_) are included, overlapping regions are merged. This is a correct approach for scoring regions, but for combining RefSeq transcript regions to a score for a gene the RefSeq ID cannot be used. Because this results in a problem: for each regions that contains more than one RefSeq ID, this region gets mapped to more than once. The solution is simple: use HGNC gene names to regions. This gives the following approach:

1. **Score for combined transcripts (combined so that overlap is removed) = max(cluster-buster homotypic CRM score)**
2. **Score for HGNC = max(associated combined transcripts)**

1. Preprocessing of RefSeq transcripts:
    1. Some RefSeq IDs map to two different regions (sometimes not even in each others proximity but on a different chromosome).
        - Possibly others are pseudogenes?
        - Most have a provisional status like OR4F3 or CTAG1B. Locations are determined by BLATing the whole genome for possible locations?
    2. Some RefSeq IDs map to special chromosomes (chromosomes with an underscore in their name), these are omitted to avoid removal because of duplication (cfr. first step).
    3. When creating the LUT for 5'UTR and first intron regions, some irregularities are encountered. These genes are omitted. 2 table inconsistencies exist:
        - For a RefSeq annotation that has no coding sequence (for non-coding transcripts the cdsStart = cdsEnd position) its ID should start with NR_, this is not always the case (e.g. NM_001195202).
        - A RefSeq annotation that has no coding sequence should have an exon count of one, this is not always the case (e.g. NR_036634)

        - CAVE: Definition of 5'UTR in this context = DNA sequence between TSS and start of CDS. This definition does not correspond with the strict biological meaning of the 5'UTR, which is defined on the final mRNA strand and contains only exons. In my definition the 5'UTR can contain introns. For this reason the first intron can completely fall into the 5'UTR region.
    4. RefSeq IDs for transcripts (prefix NR_, leaving only NM_) are omitted to avoid the problem that more than one HGNC name gets mapped to the same transcript region (cfr later). Some NR_ regions, like some miRNAs, range over many protein-coding transcripts that otherwise wouldn't overlap.
2. RefSeq transcript IDs (MM_/NR_ prefixes) are merged and the merged transcripts are considered as a union in the whole process of defining regulatory regions. In most of the cases these transcripts are associated with the same HGNC gene name.
3. 10kb upstream regions are determined for all these transcript units.
4. These regions are limited by previous genes (ignoring strandedness in limitation process).
    - CAVE: theoretical loss of gene IDs by limiting process? If two genes are located exactly after each other. Can be checked by looking at genes for which 5'UTR and introns are present and that don't get into the final pool of regions.
5. The 5'UTR regions is appended to the upstream region and the 1st intron regions are appended to the pool of regions.
6. For each regions the cluster-buster score is calculated and are ranked based on this score. Orderstatistics combines multiple region rankings.
7. Finally regions are grouped based on common HGNC name and these groups are ranked based on the best rank of one their members.

BUT: Still there are a couple of HGNC names that map to the combined transcript unit and thus are always ranked next to each other (they have the same score). When both are present in a gene set query and are ranked very high this results in a bias: the resulting recovery curve will be higher because the recovery curve will increment not by one but by two when encountering these genes.

1. HGNC website -> 19364
2. UCSC Protein coding genes -> 18975

cat RefSeq-hg19-refGene.tbl | egrep 'NM_[0-9]+' | cut -f13 | sort -u | wc -l

BUT: Only 17684 HGNC gene names remain, but according to analysis there should be 18057 names that remain. Who are these genes and why do they get lost?

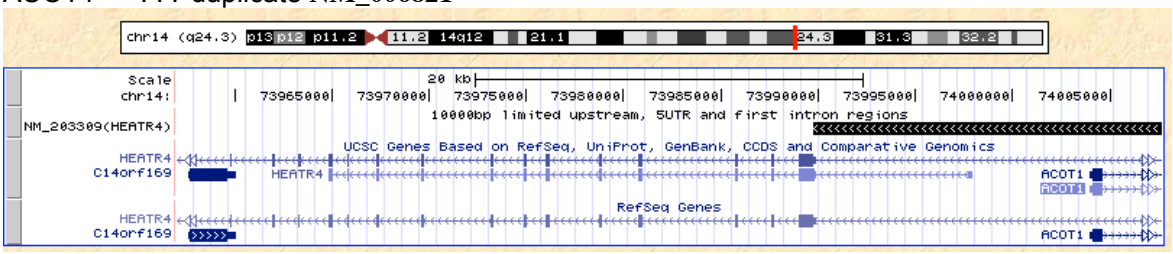
gene-analysis.sh	**18057**
cat refseq-tx-5utr-intron1.lut | sed '1d' | egrep -w 'NM_[0-9]+' | cut -f6 | sort -u | wc -l	**18699**
cat hg19-upstream10000-limit-5utr-intron1.geneid | wc -l	**17684**

cat refseq-tx-5utr-intron1.lut | sed '1d' | egrep -w 'NM_[0-9]+' | cut -f6 | sort -u | wc -l
18699

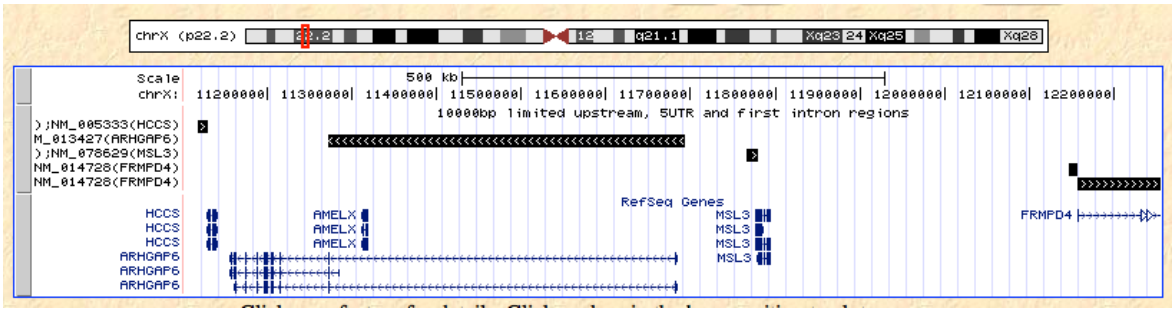
cat refseq-tx-5utr-intron1.lut | egrep -w 'NM_[0-9]+' | awk 'NR > 1 { print $1 "\t" $2 "\t" $3 "\t" $5 "(" $6 ")\t0\t" $4 }' | mergeBed -s -nms | awk '{ print $1 "\t" $2 "\t" $3 "\t" $4 "\t0\t" $5 }' | cut -f4 | tr ';' '\n' | cut -d'(' -f2 | sed -e 's/)//' | sort -u | wc -l
18698 (SGIP1 gets lost -> why: NR > 1 bij awk hoeft niet meer want grep heeft de header al weggehaald)

create-regions.sh
17684 (cfr. hgnc-names.setdiff file:

duplicate mappings are back: No.
ACOT1 -> ??? duplicate **NM_006821**



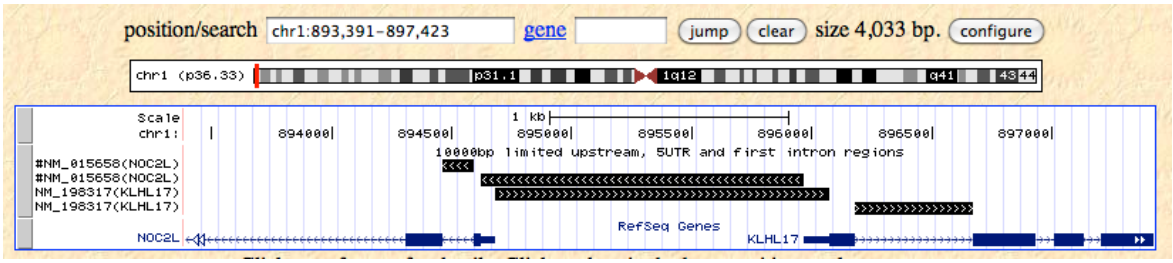
And also AMELX (NM_182680) -> created by ARHGAP6 (NM_013427)? Yes substract BED operation removes its ...

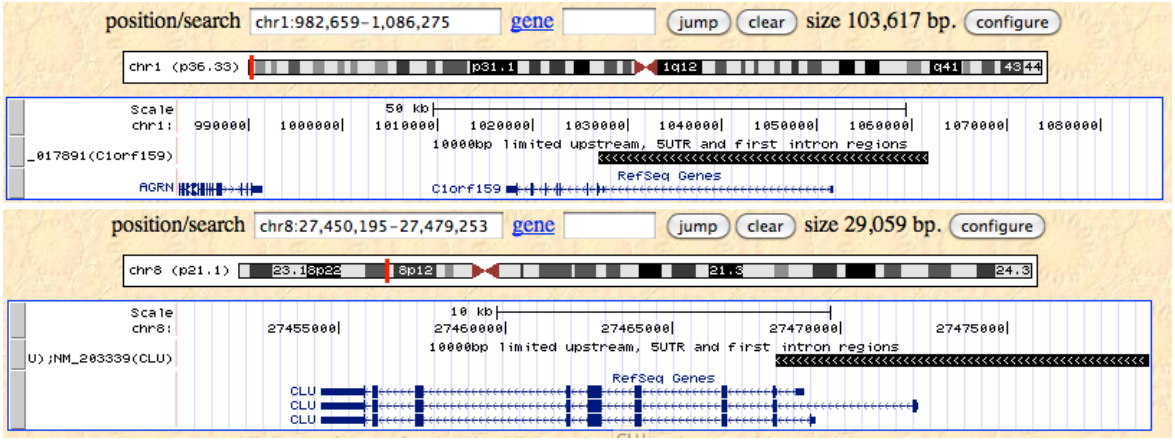


Solutions:
Do not BED subtract transcripts but just introns ...

<u>Cases:</u>

1. Normal case: e.g. KLHL17 and NOC2L
    - 10kb upstream limited by previous genes (ignoring strandedness in limitation process)
    - Extension with 5'UTR region and first intron
2. Sometimes the first intron completely overlaps with the 5'UTR region (contradiction when regarded in strict biological sense): e.g. C1orf159
3. Multiple RefSeq transcripts can overlap and most of the time these regions map to a unique HGNC gene. These regions are merged and considered as one unit. For the determination of the 5'UTR and 1st intron of this unit the 5'UTR regions are merged, the same for the 1st introns. In the case of CLU the 5'UTR regions are merged and the merger of the 1st introns happens to overlap with the merged 5'UTR region and so is removed.




position/search
chr1:982,659-1,086,275
gene
jump
clear
size 103,617 bp.
configure
chr1 (p36.33)
p31.1
1q12
q41
43
44
Scale
chr1:
990000
1000000
1010000
1020000
1030000
1040000
1050000
1060000
1070000
1080000
50 kb
10000bp limited upstream, 5UTR and first intron regions
_017891(C1orf159)
RefSeq Genes
AGRN
C1orf159
position/search
chr8:27,450,195-27,479,253
gene
jump
clear
size 29,059 bp.
configure
chr8 (p21.1)
23.1
8p22
8p12
21.3
24.3
Scale
chr8:
27455000
27460000
27465000
27470000
27475000
10 kb
10000bp limited upstream, 5UTR and first intron regions
U);NM_203339(CLU)
RefSeq Genes
CLU
CLU
CLU