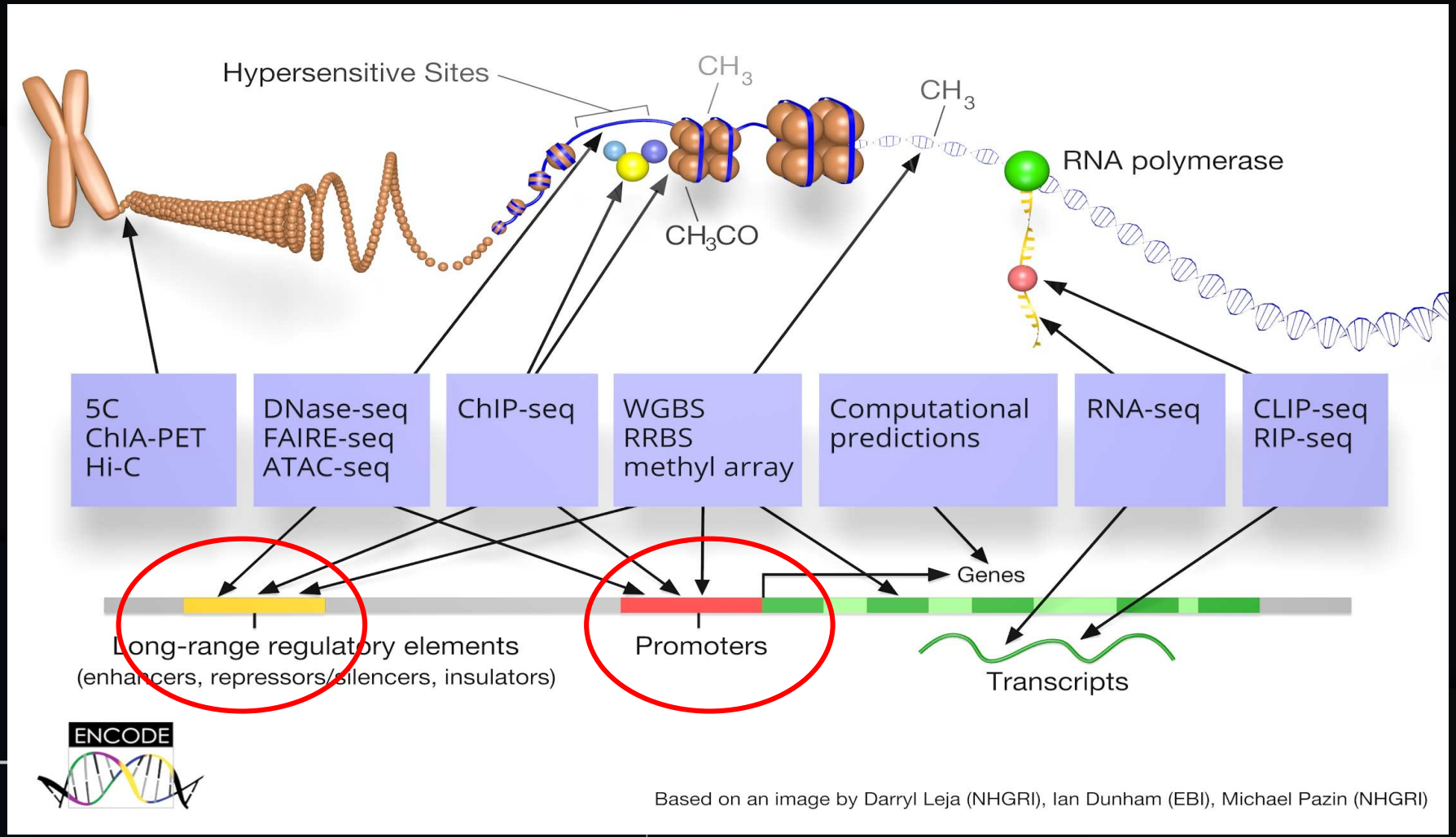
Analisi di dati di ENCODE genomici ed epigenomici

Linguaggi di Programmazione:

- -SQL
- -R
- -Python


Hypersensitive Sites
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
ENCODE
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Tipi di Dato

```
| Tables in snpImpactResource
+---------------------------+
| CellLines
| FunctionalElement
| GeneSymbols
| GeneSymbols_Symbols
| INDELs
| INDELs_FunctionalElement
| INDELs_PFM
| MotifDatabases
| PFM
| PFM_GeneSymbols
| PFM_MotifDatabases
| SNPs
| SNPs_FunctionalElement
| SNPs_PFM
| Symbols
+---------------------------+
```

Tutte le tabelle del DataBase
In rosso sono le tabelle utilizzate

Server: MariaDB

Container: Singularity

Dati quantitativi iniziali

- SNPs = 14810175
- Elementi Funzionali = 759
- CellLines = 238
- PFM = 5424 (SNPs_PFM = 5352)

3 esempi di records

SNPs

SNPID	rsid	chrom	pos	ref	alt	maf1000genomes	mafTOPMed
1	rs575272151	1	11008	C	G	NULL	NULL

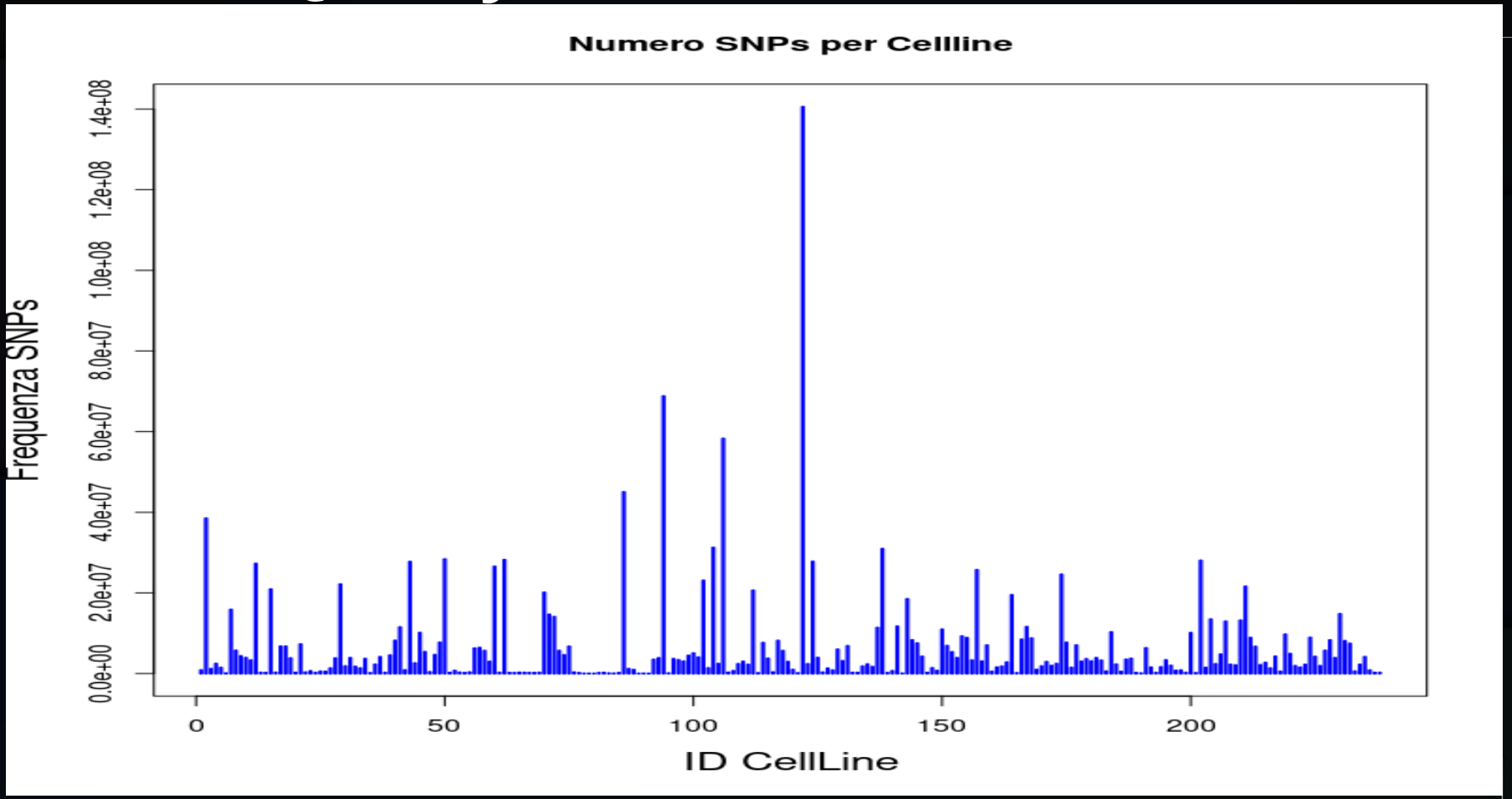
SNPs_FunctionalElement

SNPID	ElementID	CELLLINEID	countExperiments	fileType
1	33	122	1	broad

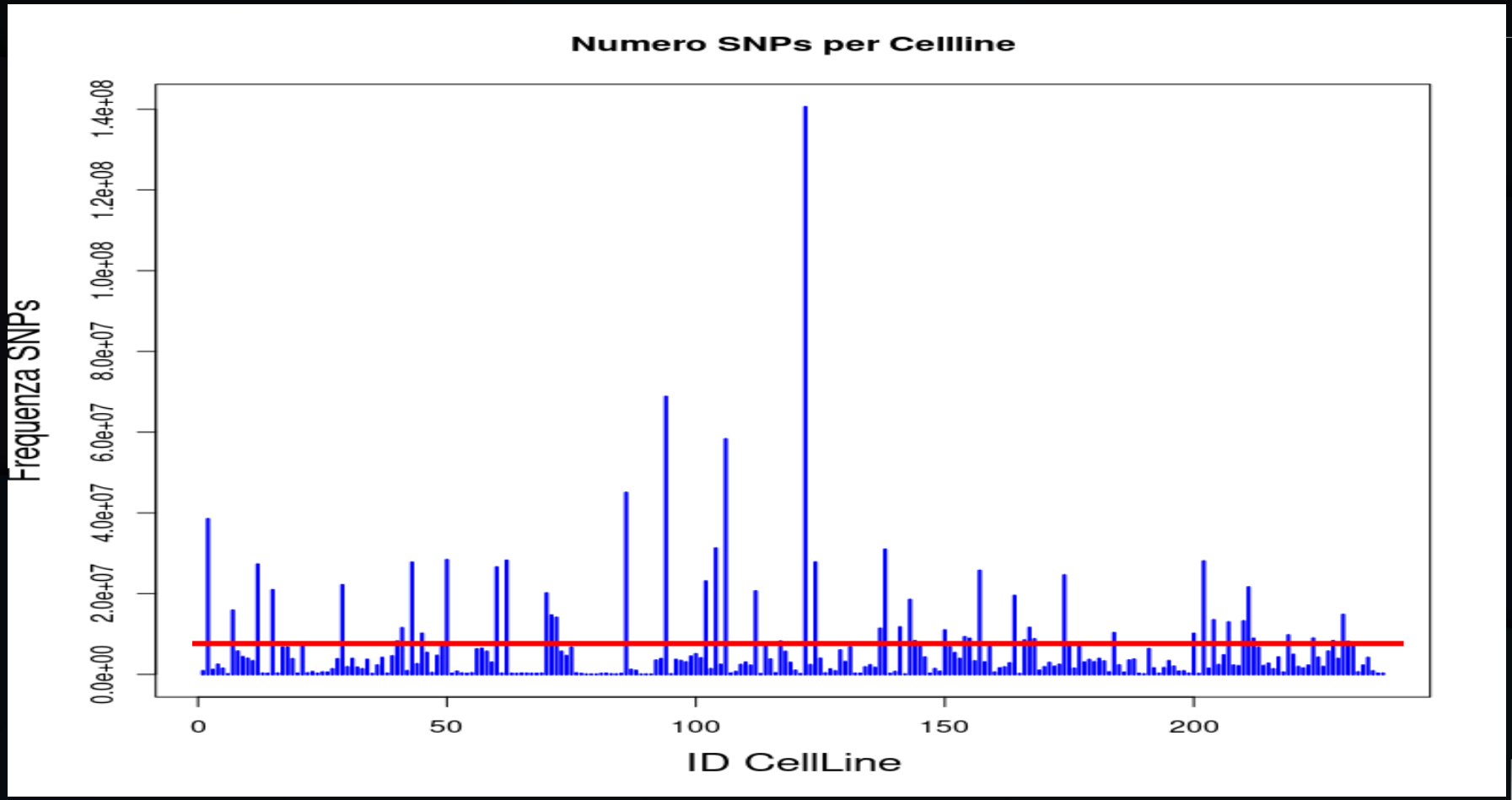
PFM

SNPID	PFMID	start	strand	type	scoreRef	scoreALT
1	34384	11007		match	6.22488	6.80984

Prima Query fatta al DataBase

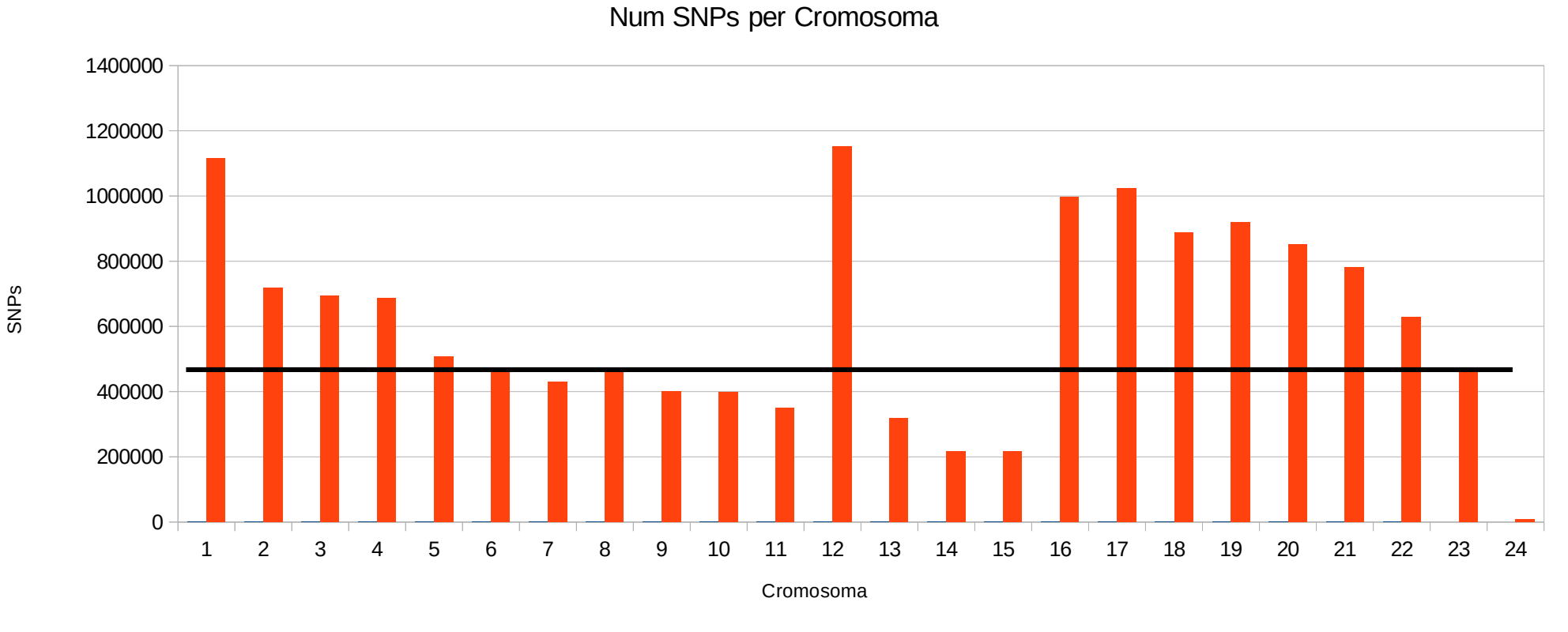


Creare threshold di riferimento per i ricercatori



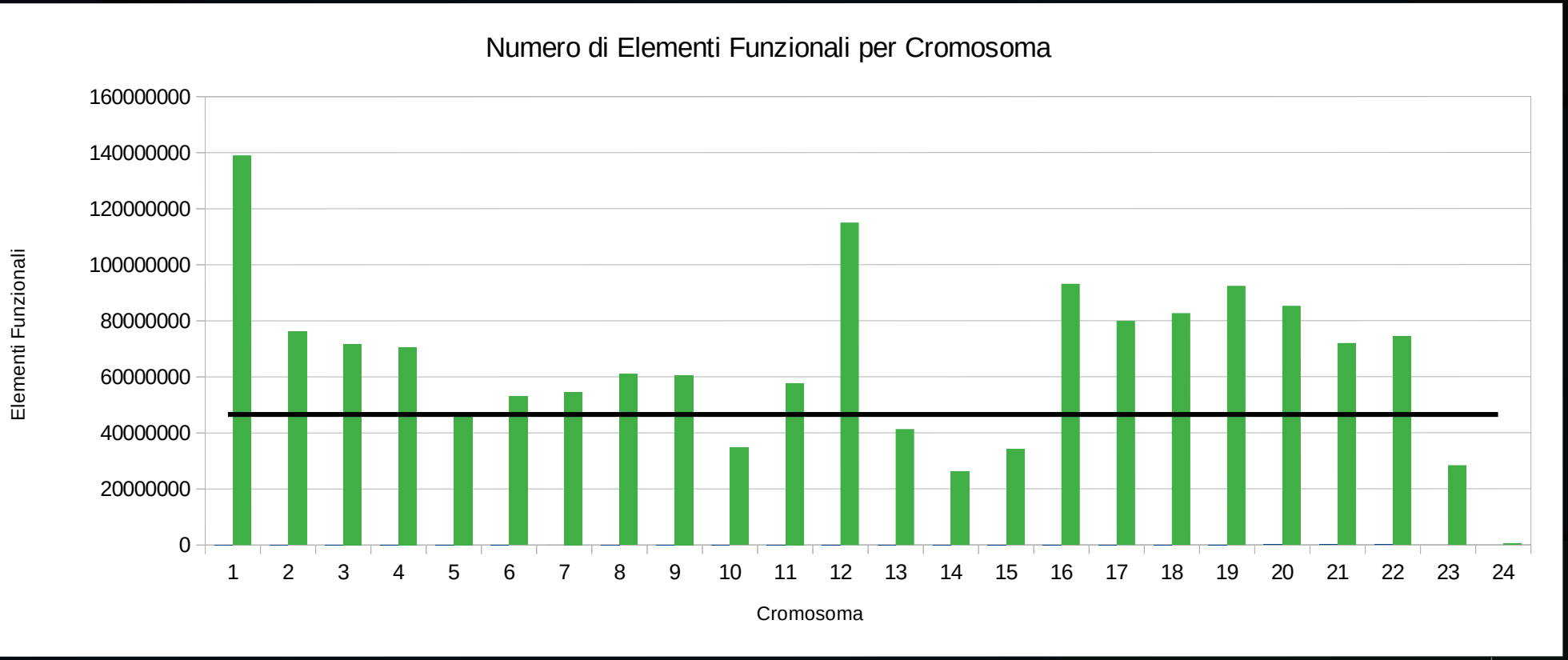
SNPs per Chr

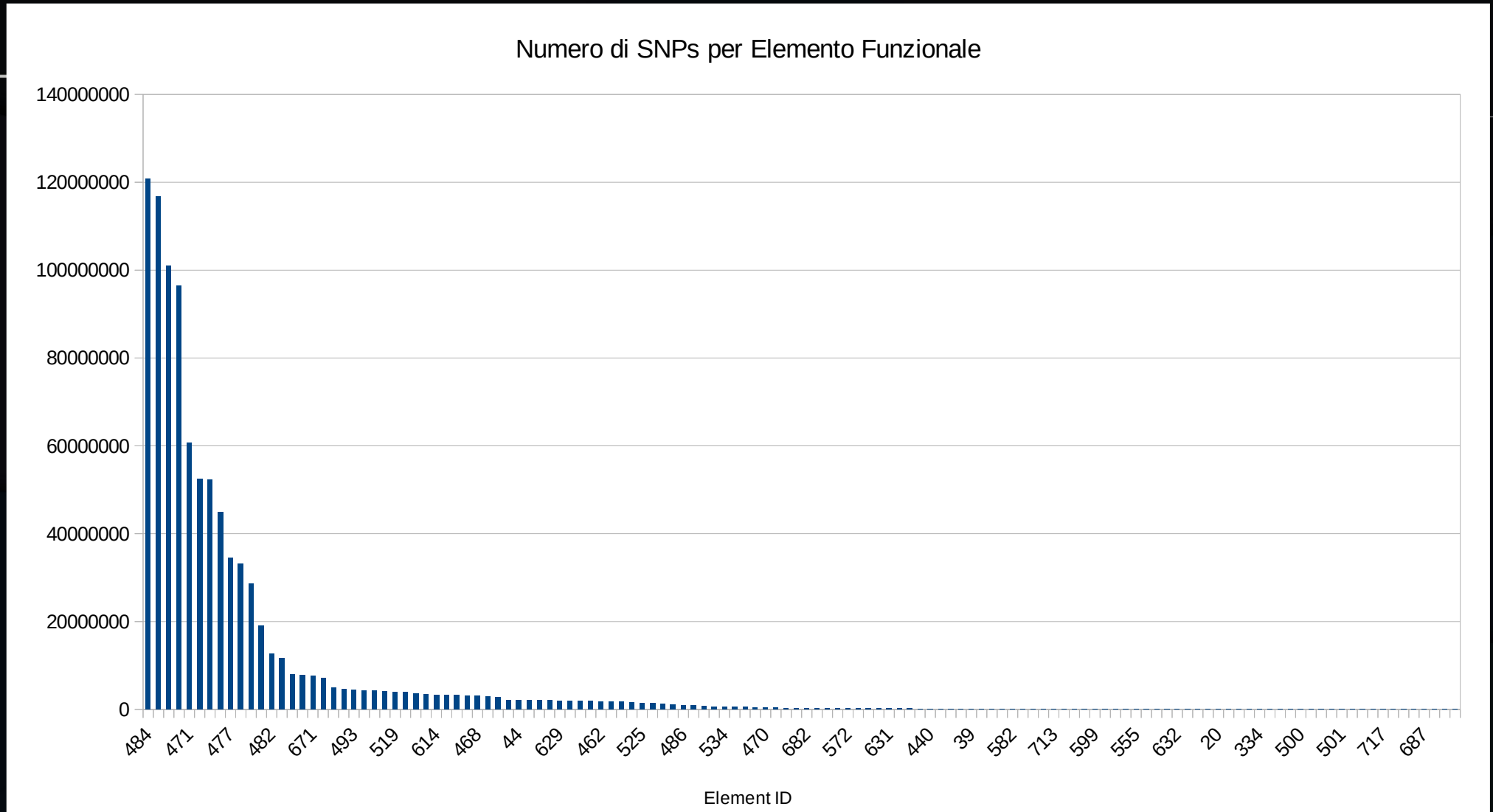
In questo caso non discrimino fra SNPs diversi, ne' per CellLine diverse



** Chr 23=X, 24=Y*

Elementi Funzionali per Chr




Numero di SNPs per Elemento Funzionale
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
484
471
477
482
671
493
519
614
468
44
629
462
525
486
534
470
682
572
631
440
39
582
713
599
555
632
20
334
500
501
717
687
Element ID

Esempio dei filtri per i dati

- File Type: broad / narrow or match/change/change
- Contare distinti elementi oppure contare la somma delle frequenze dei singoli elementi
- Non mettere i valori NULL
- Discriminare per la lunghezza delle PFM Probability Frequency Matrix

Possibile filtro dati per i PFM

| type file | scoreRef | scoreALT | name Element | length | maxScore | SNPs ID | chrom | pos

Uso di 3 tabelle: SNPs_PFM, PFM, SNPs

Per ogni PFM e SNP ho uno score di riferimento, uno score massimo teorico ed uno score Alternativo nella cellula mutata.

Voglio filtrare gli score piu’interessanti con:

| ((ScoreALT – ScoreRef) / maxScore)*100 | > 10 %

Filtro gli score che differiscono del + o - 10% rispetto a quello di riferimento.

| SNPID | scoreRef | scoreALT | maxScore | PFMID |
| 1 | 6.22488 | 6.80984 | 9.62551 | 34384 |

Possibili informazioni utili ricavabili DataBase

Marcare le regioni dei cromosomi ad alta frequenza di mutazioni di SNPs.

Capire se un particolare SNPs (ex C → G) incide di piu’ su alcuni geni piuttosto che su altri

Marcare le posizioni dei cromosomi a cui “si attaccano” i differenti Elementi Funzionali