SciLifeLab

Karolinska Institutet | KTH Vetenskap och Konst | Stockholms universitet | Uppsala Universitet

Prioritizing Non coding Variants from Whole genome sequencing project

Sailendra Pradhananga

DNA Club

22/02/2017

Drug induced myelosuppressive toxicity in Lung cancer

SciLifeLab

Lung cancer → Common cancer with high mortality rate

+

=

HIGH Toxicity

LOW Toxicity

- Serious side effect of chemotherapy
- Multitude of genes involved in the traits

- Myelosuppression toxicity
- decrease in blood cells

AIM : Explain variability in myelosuppression response in Lung cancer patients

Previous studies

SciLifeLab

- Whole exome sequencing was performed in 215 lung cancer patients
- Few gene centric SNV variants reported and validated
- Subsequently, scaled up to whole genome sequencing of 96 extreme phenotypes of high and low toxicity

Background

- Whole genome sequencing genotypes large number of variants in either population or individual level
- Convention association suffers from multiple testing correction for complex traits
- Thus prioritizing of variants are done based on specific regions of genome
- In our studies, we are aiming for prioritizing low effect non coding variants for Hi-CAP probe sets

Outline of work

SciLifeLab

- Part1 – Comparing genetic variants from WGS 96 Lung cancer patients sample to SweGen 1000 population sequencing

- Part 2 – Annotating lung cancer non coding variants from ENCODE, FANTOM (Ref) database

- Part 3 – Classify high and low toxicity in lung cancer patients using clustering methods

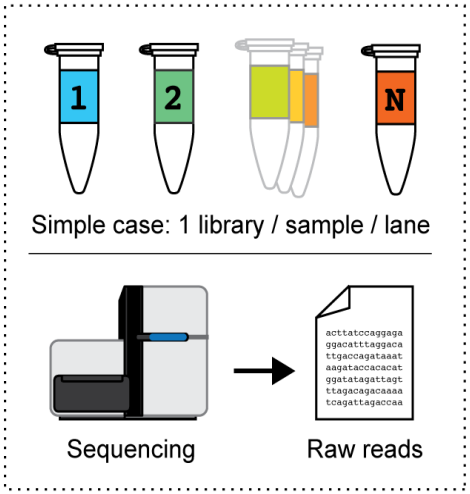
- Part 4 – Enrichment and prioritizing of non coding variants in high and low toxicity phenotype groups

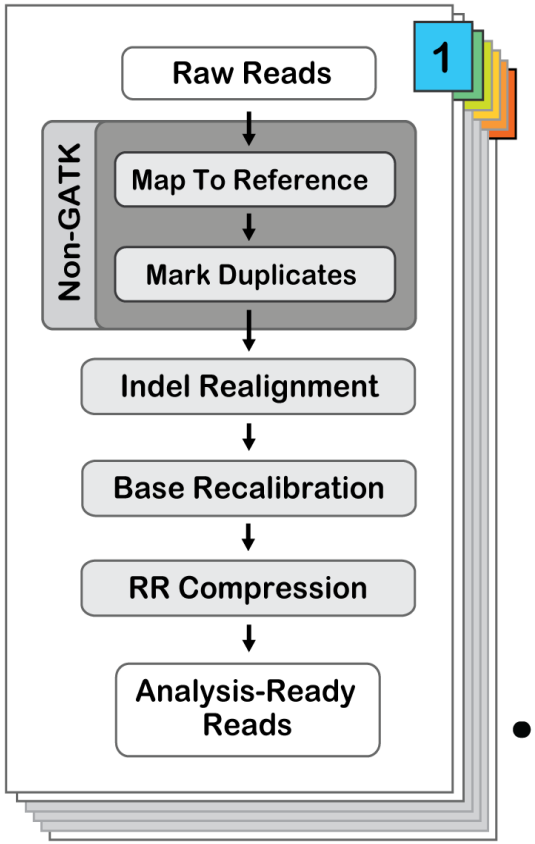
Outline of work

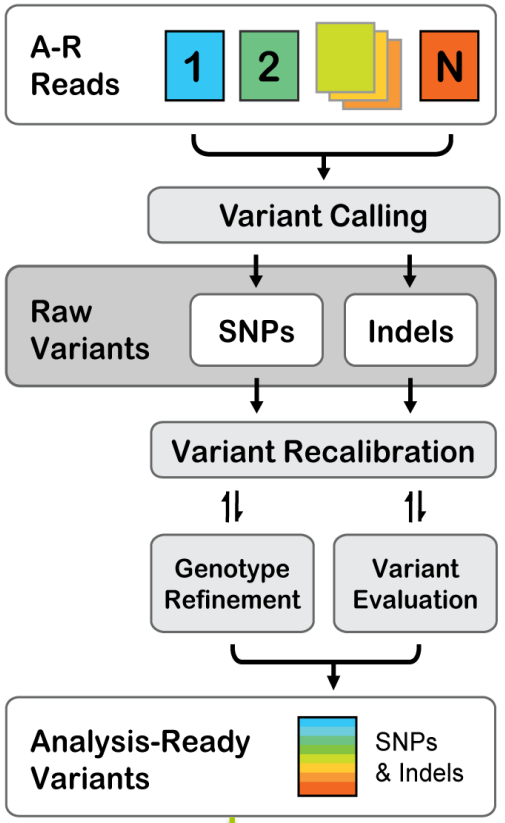
- Part1 – Comparing genetic variants from WGS 96 Lung cancer patients sample to SweGen 1000 population sequencing

Whole genome sequencing pipeline

SciLifeLab

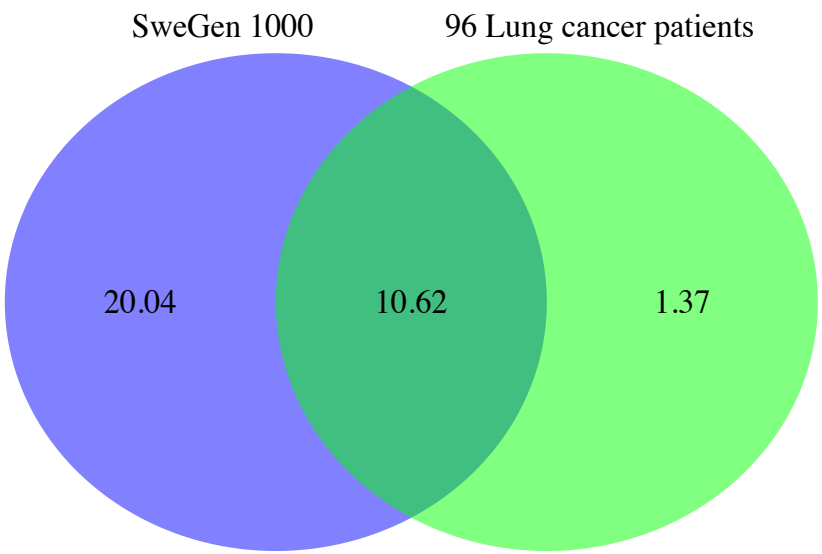






Raw final Variant call in all samples

Comparison of SweGen 1000 with 96 lung cancer sample

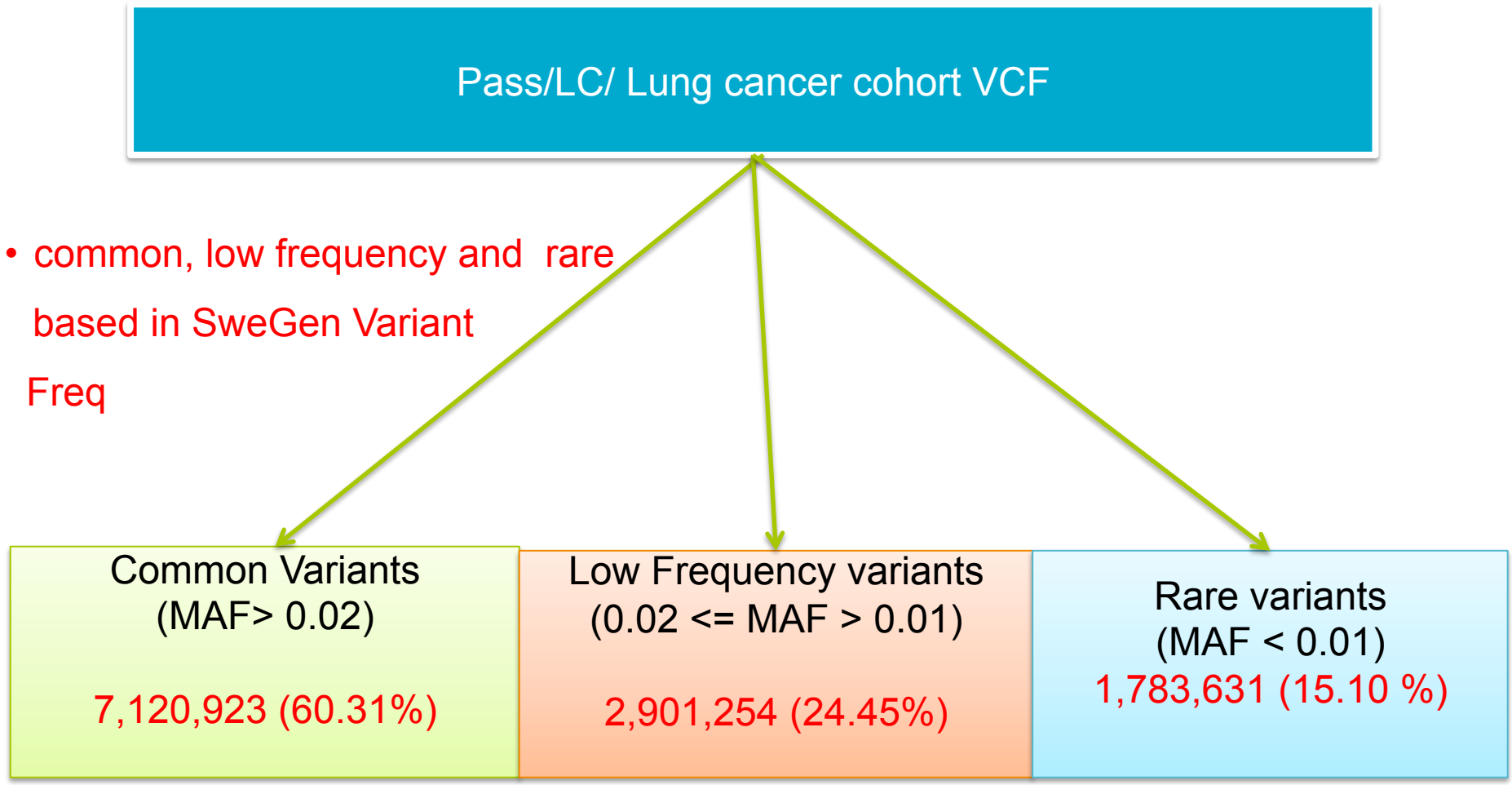


SNV counts bet two datasets in Million

- **70,353 (7.35 %) of novel variants shared by at least 2 individuals**
- **Imputed Allele frequency from SweGen to Lung cancer cohort**

Post processing of Variant files

SciLifeLab

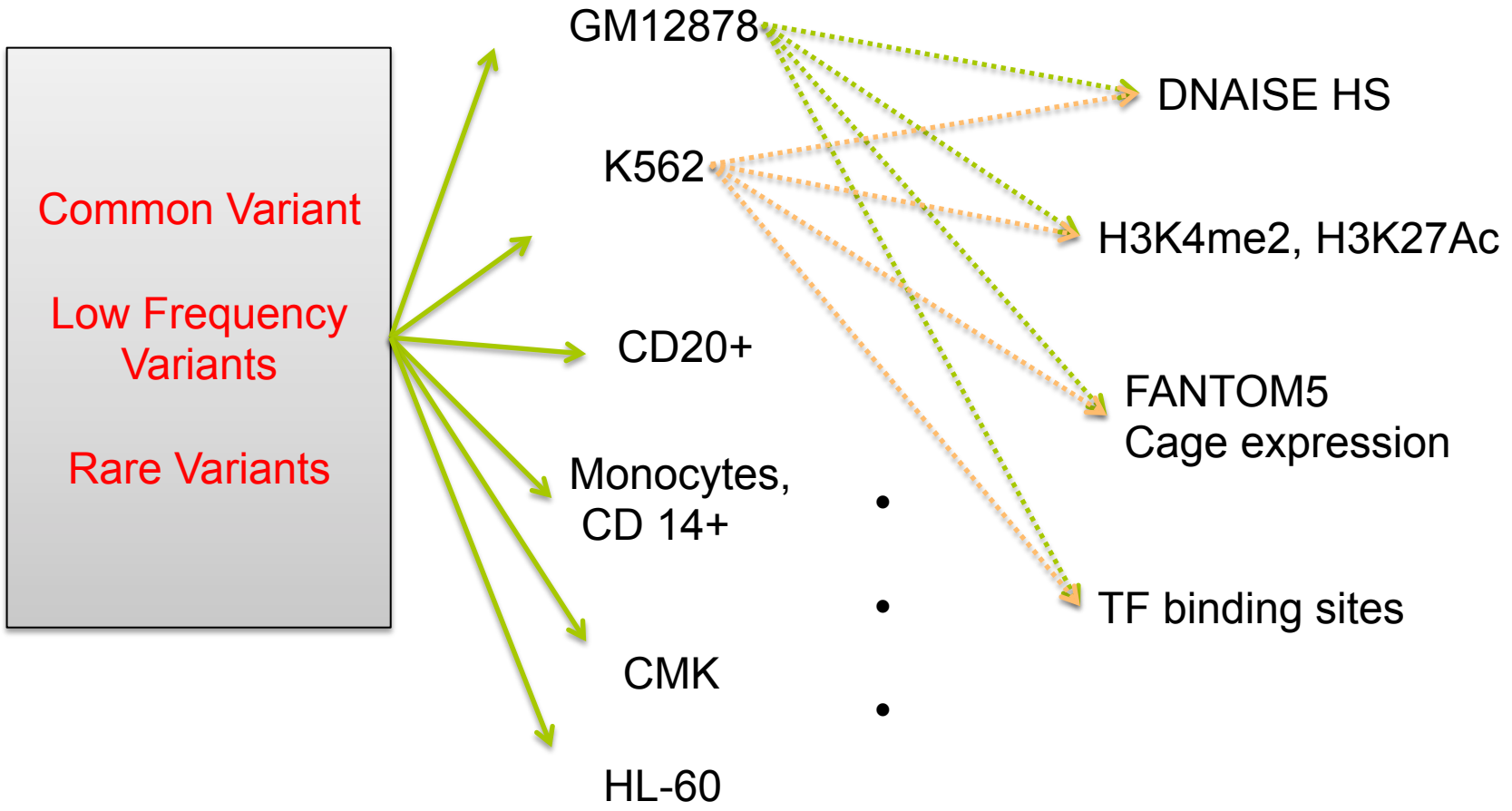


Outline of work

- Part 2 – Annotating lung cancer non coding variants from ENCODE, FANTOM (Ref) database

Annotation of WGS variants

SciLifeLab



Genetic variant from 96 Lung cancer sample

Relevant cell/ tissue type

Non coding Functional markers in ENCODE and FANTOM5

Approach

SciLifeLab

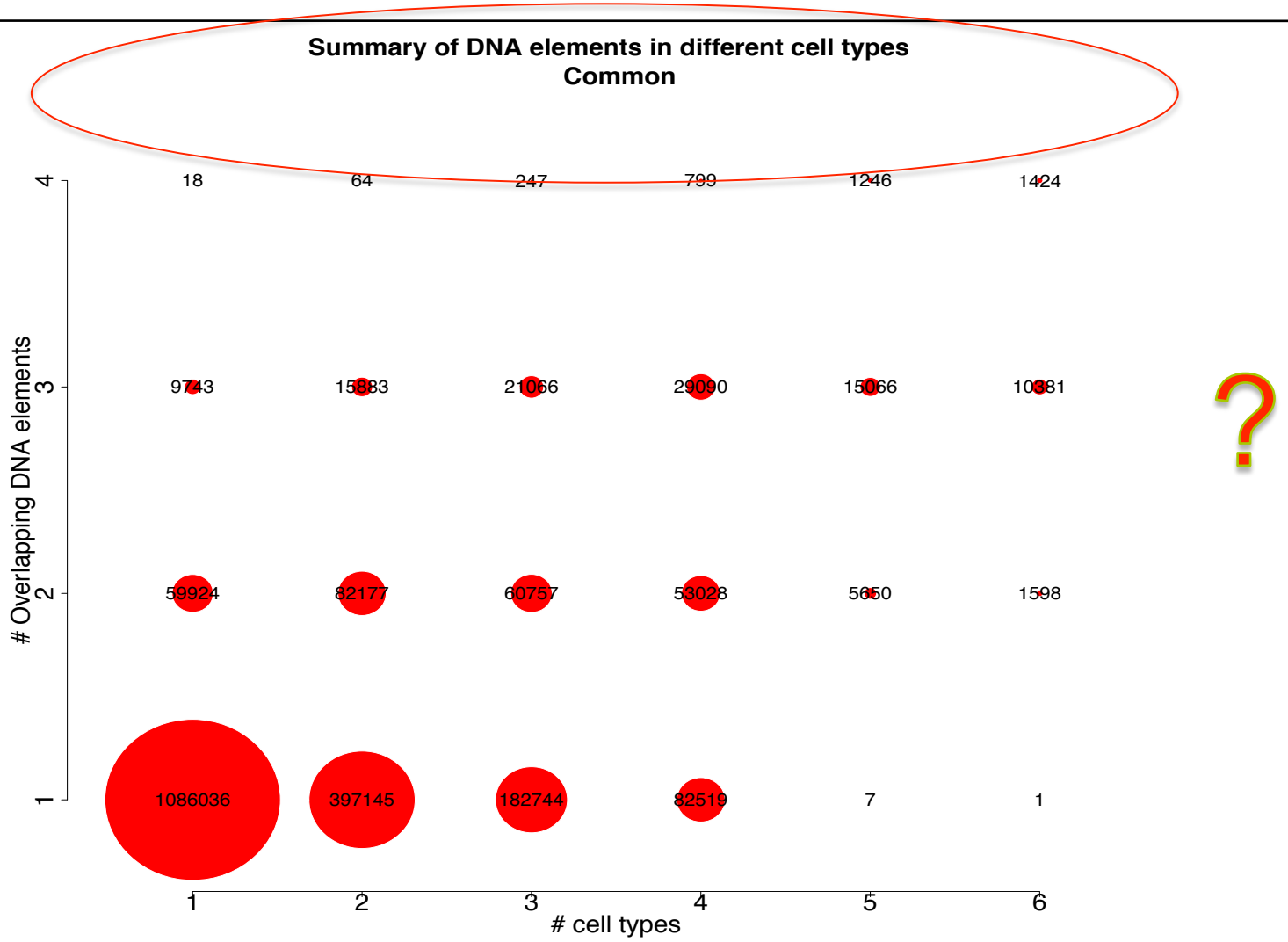
- For each variants we annotated with different functional noncoding markers of relevant cell types.

- In the initial study we annotated variant with individual cell type.

- Subsequently, aggregated approach to find variants active in different cell-types with different functional elements.

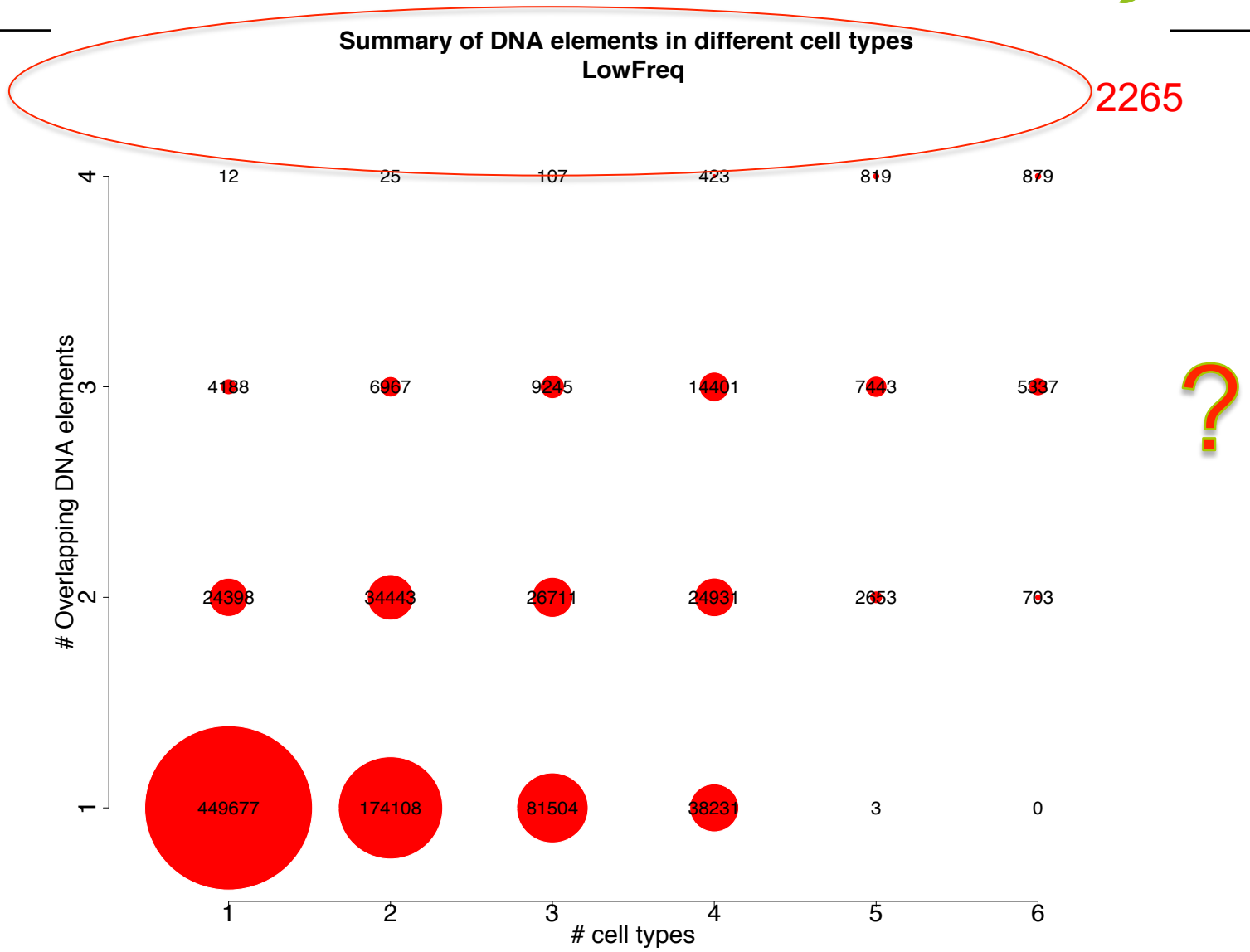
SciLifeLab

3798 variants



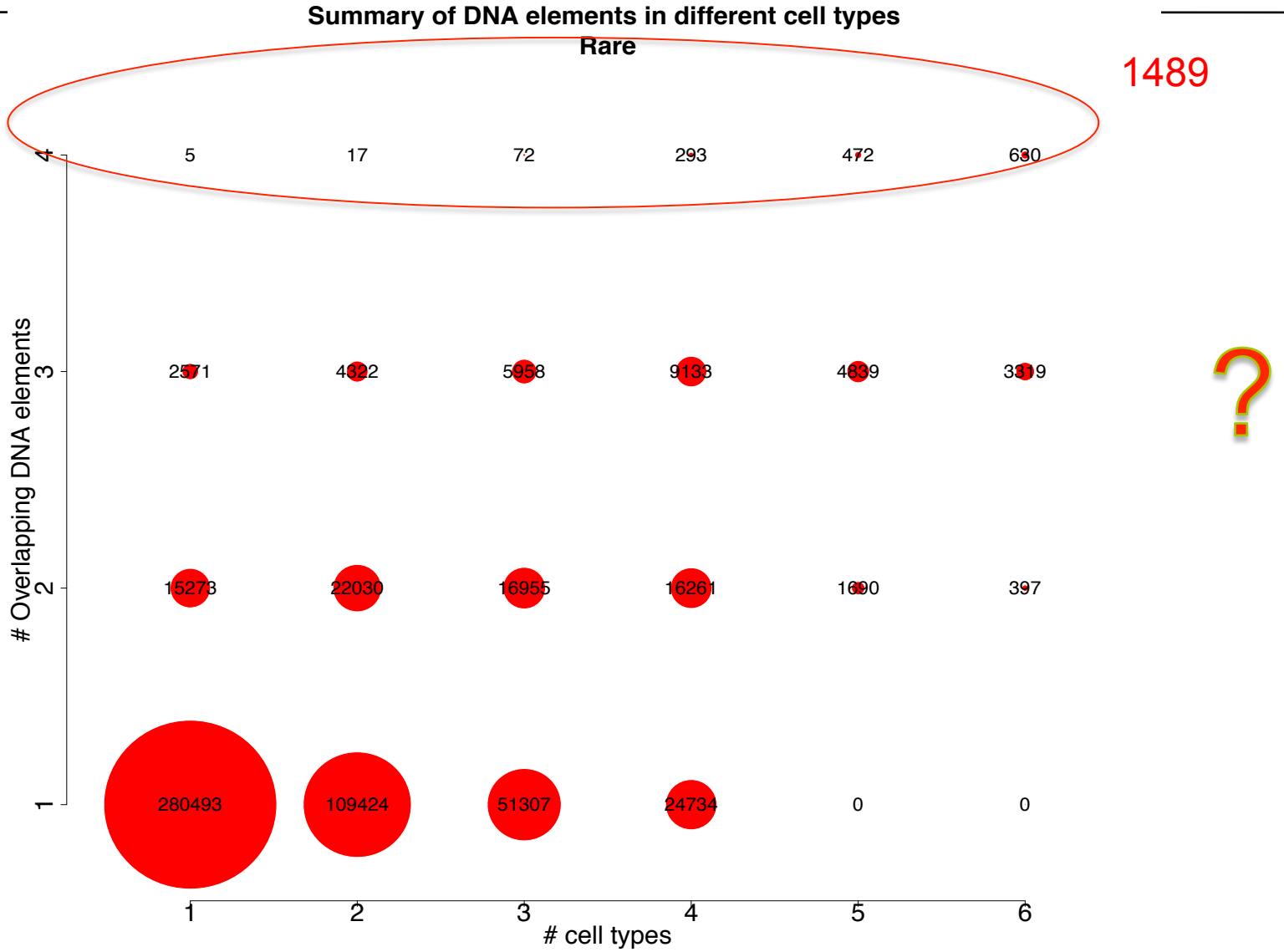
2,116,613 Common variants annotated with non coding functional marks

SciLifeLab



907,208 low frequncy annotated with histone marks

SciLifeLab



1489

?

570,195 rare variants annotated with histone marks

Thus we have…

SciLifeLab

- To devise and prioritize variants based on low and high toxicity of lung cancer patients

- We used the blood cell count values platelets(TPK), leucocytes(LPK), neutrophils(NPK) after drug administration

- Cluster each variants in high and low toxicity for each TPK, LPK and NPK phenotype and combined phenotype

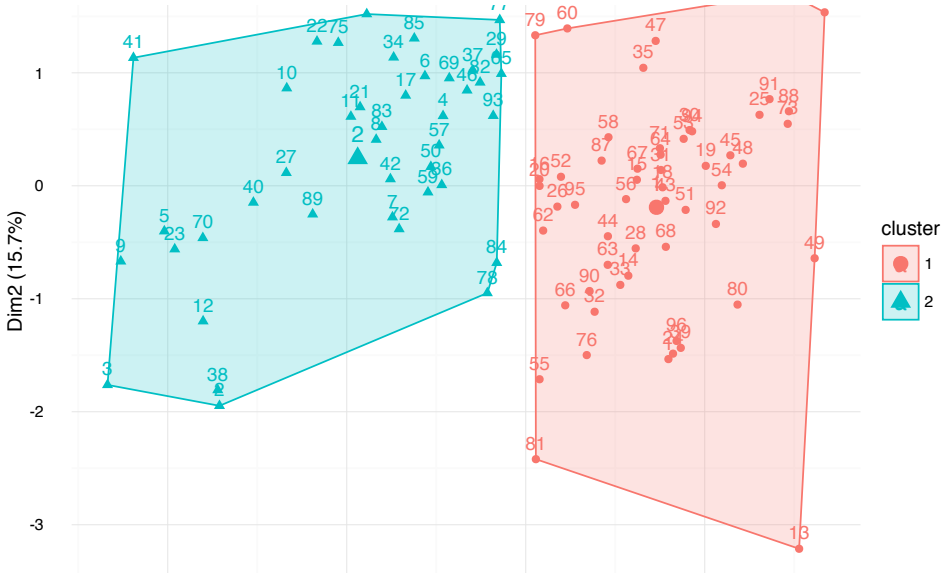
Outline of work

- Part 3 – Classify high and low toxicity in lung cancer patients using clustering methods

Clustering results

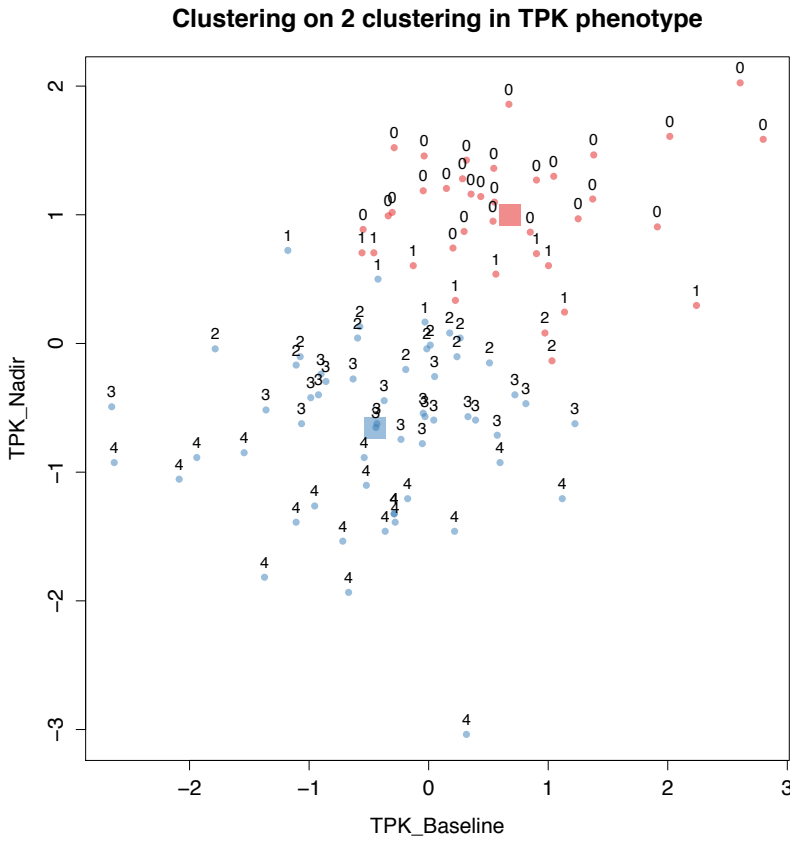
SciLifeLab

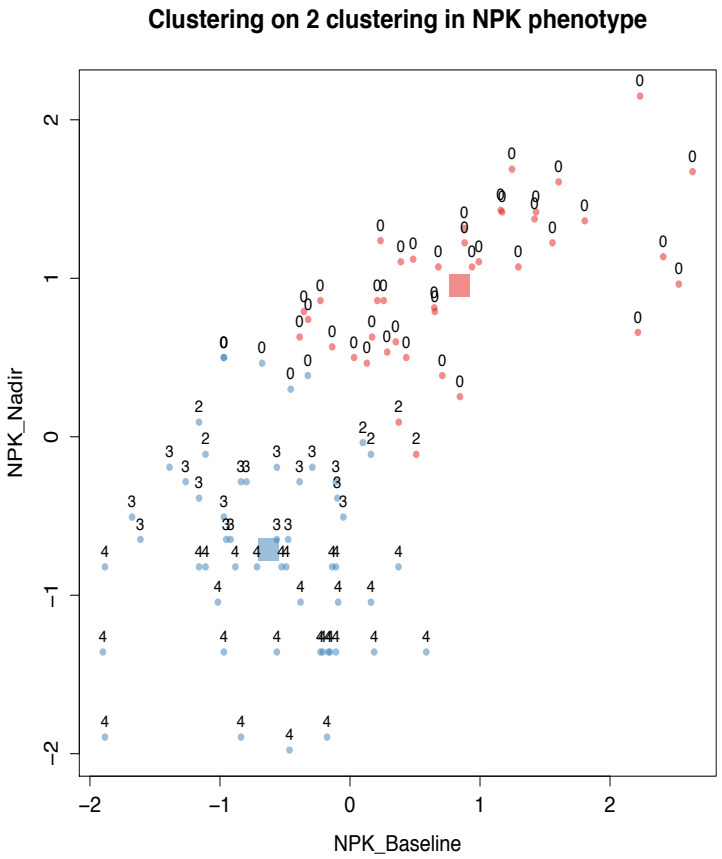
- Using unsupervised Kmeans clustering at K=2
- Blood count values of all phenotype as an input



HT	LT
54	42

For the individual phenotypes





LT (red)	HT (Blue)
38	58

HT(Blue)	LT(Red)
55	41

Outline of work

- Part 4 – Enrichment and prioritizing of non coding variants in high and low toxicity phenotype groups

Imputing toxicity score in each phenotypic groups

Hypothesis:

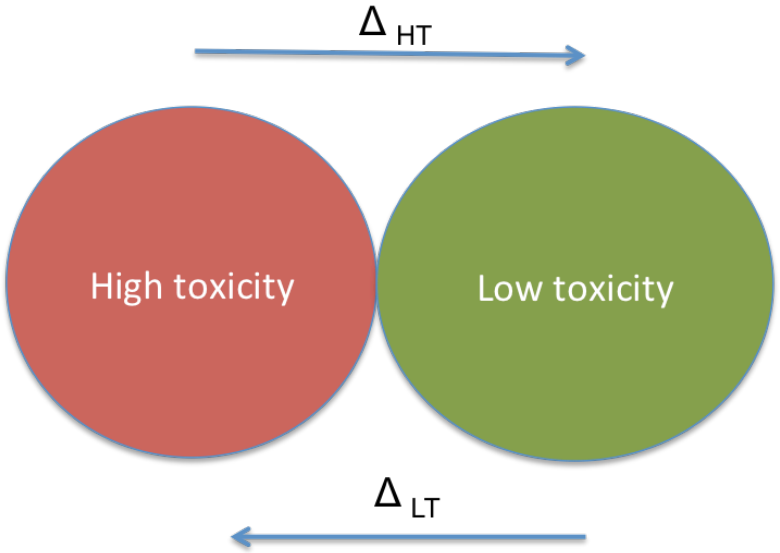
In each toxicity group enrichment of variants is attributed in delta value which is ratio of minor allele frequency in each group.

- For each variants in all group

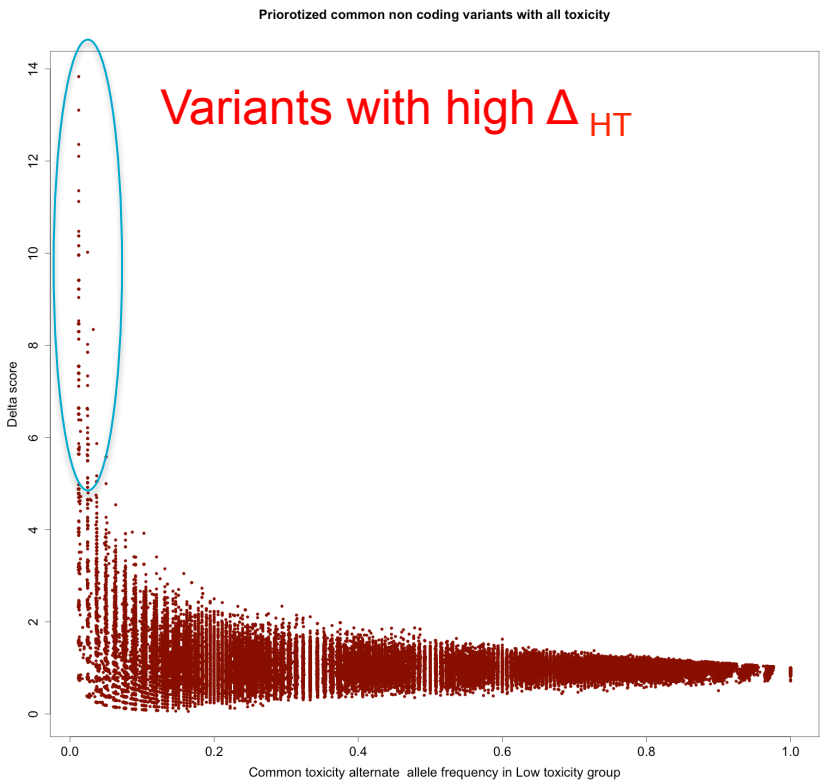
Delta score defined as:

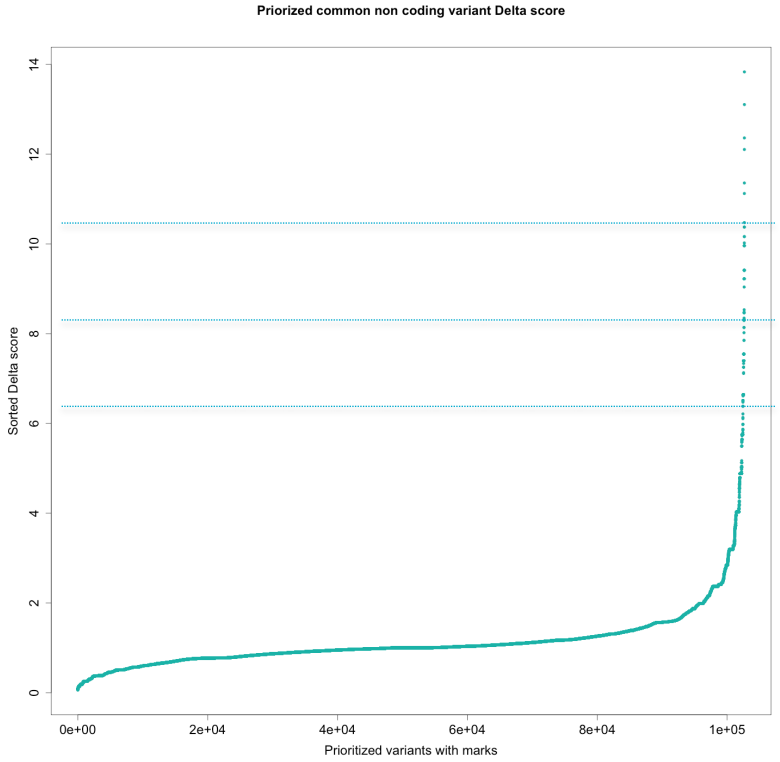
Δ_{HT} = (MAF_HT/MAF_LT)

Δ_{LT} = (MAF_LT/MAF_HT)



Toy example
Common ann_level 3 / phenotype all





?

Δ HT as high as 14 observed in Common variant levels

However still thresholding of Delta is not certain

Furthermore,

SciLifeLab

- Similar pattern observed in all other phenotypes

- Enrichment of variants were observed in two groups

Enrichment status

SciLifeLab

We already had variants that have been annotated with different functional markers in different cell types. So basic premise is that with increase of annotation level in variants, we are adding to functionality to variants with removal of random variants. In order to test the hypothesis we define the following terms a and b in which is defined as :

$$a = \frac{\#\ of\ non-annotated\ variants\ with\ \Delta > threshold}{Total\ number\ of\ variants}$$

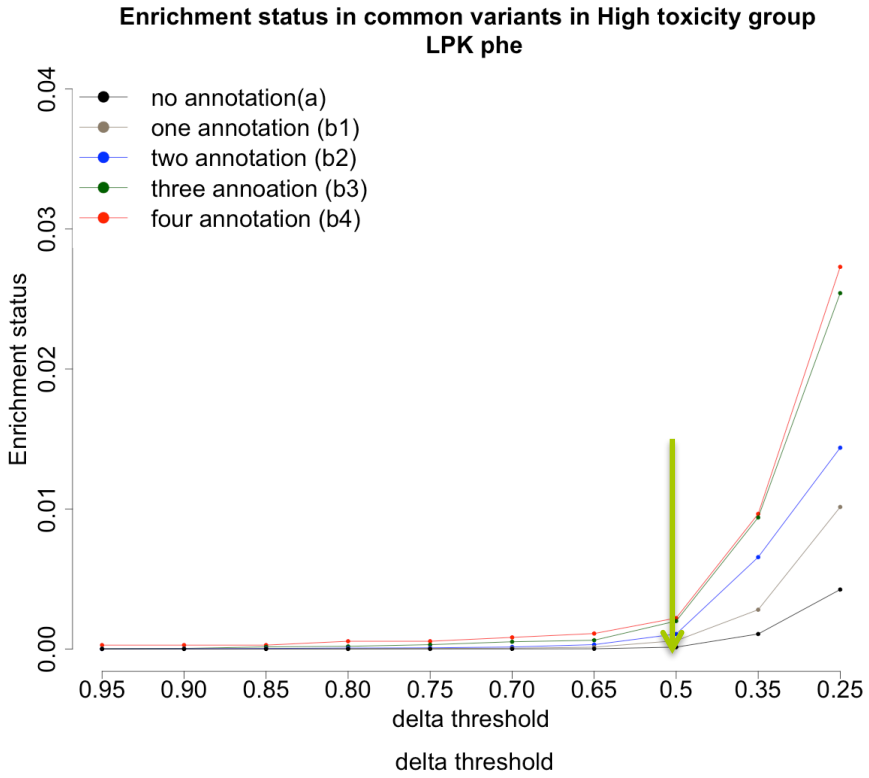
$$b = \frac{\#\ of\ annotated\ variants\ with\ x\ levels\ \ \&\&\ \ \Delta > threshold}{Total\ number\ of\ variants\ with\ x\ levels}$$

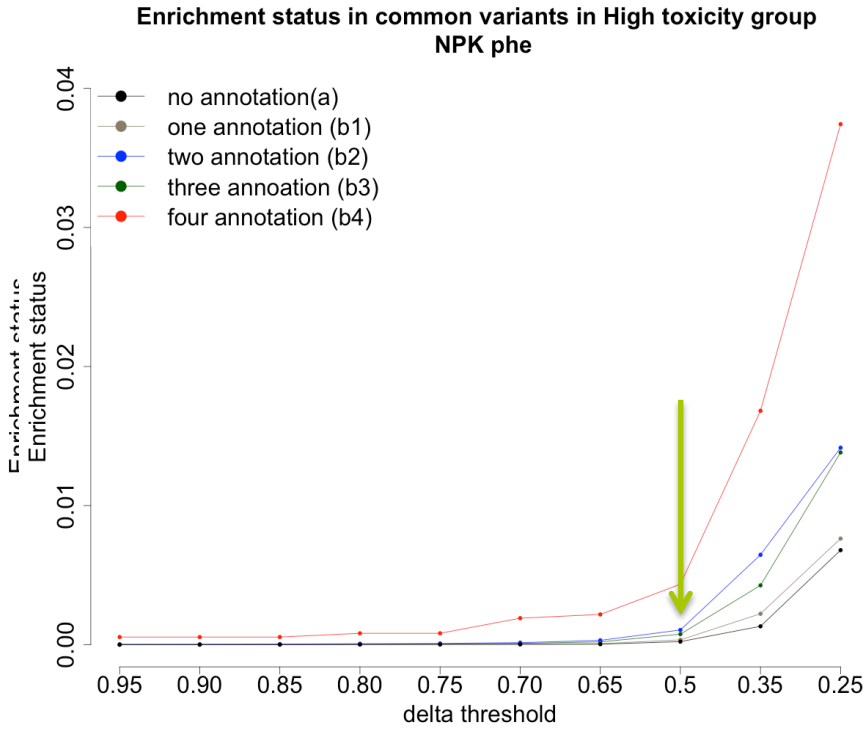
$$where\ x = 1, 2, 3, 4$$

$$where\ threshold = 0.95, 0.90, 0.85, 0.80, 0, 75, 0.70, 0.65, 0.5, 0.35, 0.25$$

Enrichment status b/a (enrichment status) in different phenotype

Common variants with enrichment status





Trading off between number of variant and annotation level

Delta Threshold of 0.5 and Annotation level = 3

Finally,

SciLifeLab

- Using threshold of 0.5 and Annotation level of 3, we prioritized 350 variants with functionality marks for high toxicity from 2million common variants

- Similarly, using the same principal we have prioritized 293 enriched variants with non coding function for all low toxicity groups

Future works

SciLifeLab

- Prioritized variant list of 3000 from 15 M variants
- Design probe-set for Hi-CAP studies with prioritized variants for selected cell lines MolM1, CMK
- Study interaction dynamics in variants with HiCAP analysis

Acknowledgements

- Prof. Joakim Lundeberg
- Asst Prof. Pelin Sahlen
- Assoc Prof. Henrik Green
- Niclas Bjorn
- Complex disease group