Effect of CNVs on gene expression and the IQ

Jocelyn Bedard

25 January 2022

Introduction to project

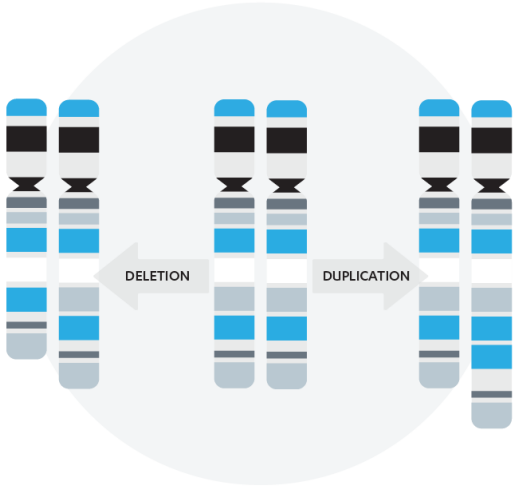
- Based on Huguet, Schramm, *et al.* (2018), we know that DEL CNVs, based on their gene content and the total probability of loss of function intolerance (pLI), can have a general negative impact on the IQ

- It is generally accepted that CNVs are likely to convey their effect, at least in part, by affecting gene expression

- The goal of the project is to assess how CNVs affect the expression of genes and see if we can integrate this information (factor) into Huguet *et al*'s model to help predict their effect on IQ

Huguet G, Schramm C, *et al*. (2018) *JAMA Psy.* 75(5):447

Effect of CNVs on gene expression:



CNVs can affect gene expression in various ways:

- Gene dosage effect: expression of genes within CNV region

  DELs can lead to underexpression (haploinsufficiency), DUPs to overexpression

- Cis effects on genes: if extremeties of CNVs carry regulatory elements

  these can affect the expression of neighbouring genes

- Trans effects on genes: if CNVs contain factors modulating gene expression (e.g. Transcription factors) the level of these can affect the expression of other genes distant from the CNV

CARTaGENE cohort:

General population cohort:

- ~40,000 individuals aged between 40 and 70 for which a lot of historical and physical information, (and in some cases) biological samples have been collected
- A large set have completed congnitive tests that can likely be used to assess their cognitive function (and predict their IQ)
- Aproximately 3000 members have been genotyped and a further subset of ~1000 analysed by RNA-Seq

Awadalla *et al*. (2013) *Int. J. Epidemiology* 42:1285

Transcriptome\Genotypes\Cognition

From
937 Omni 2.5 chips
635 with
QC6

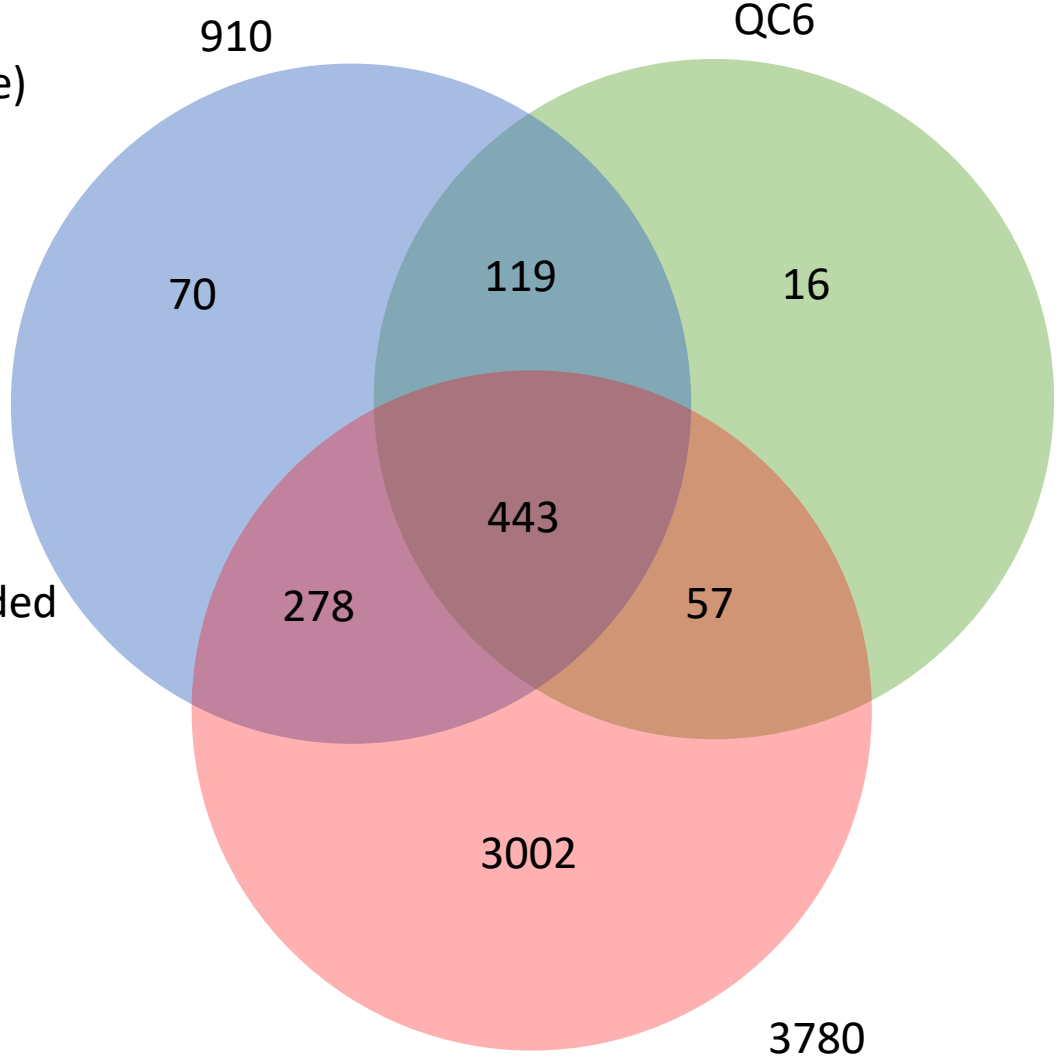
Transcriptome N=910
Genotype N=635 for Omni 2.5 chip (after QC6=<200 CNVs/genome)
no mosaic CNVs
Cognition N=3780

443 individuals with all three datas

721 individuals with Transcriptome and Cognition data

562 individuals with Transcriptome and Genotype (Omni 2.5)
5 individuals with Transcriptome and Genotype (GSA) not included

500 individuals with Genotype et Cognition data



The cognitive tests: Memory

n=3780 (with all 3 CZs)

Standardization:
Z-score= (X-u)/sd

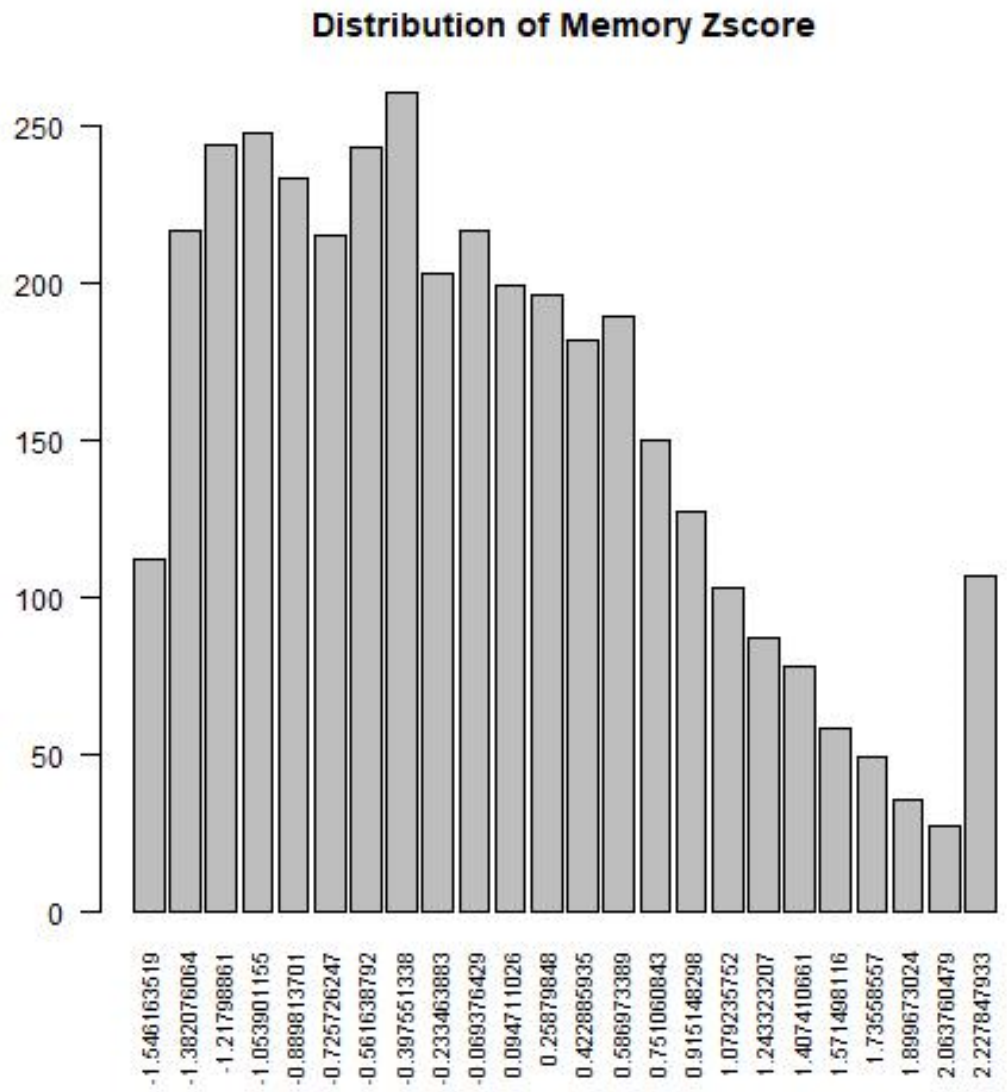
<u>Paired associates learning:</u>

Number of attemps
to correctly identify
location of target (image)

7 targets

Result is
between 7 and 30 incl.
(24 possibilities)

Lower = Better



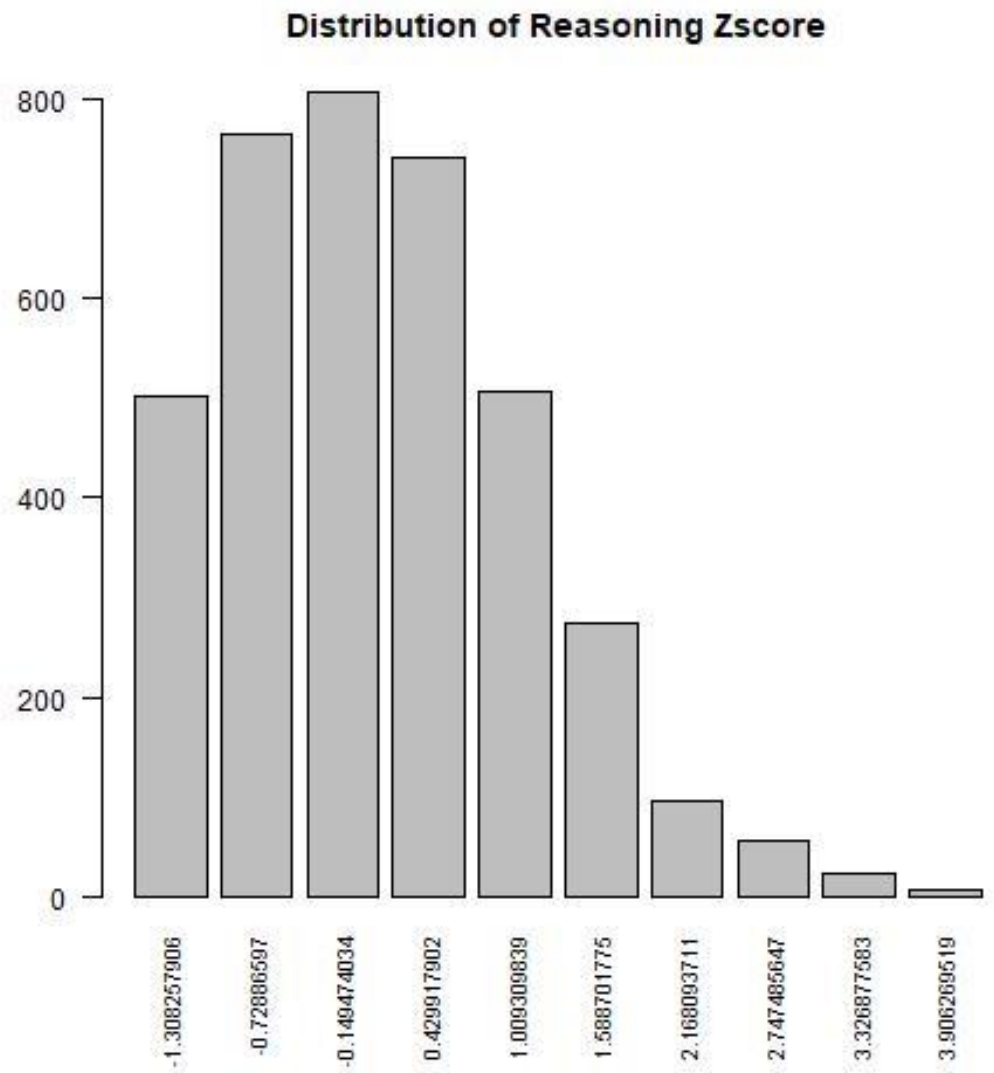
Reasoning

<u>Verbal and Numeric reasoning:</u>

Number of correct answers out of 12 questions in 2minutes

Results theoretically between
0 and 12

Higher=Better



n=3780

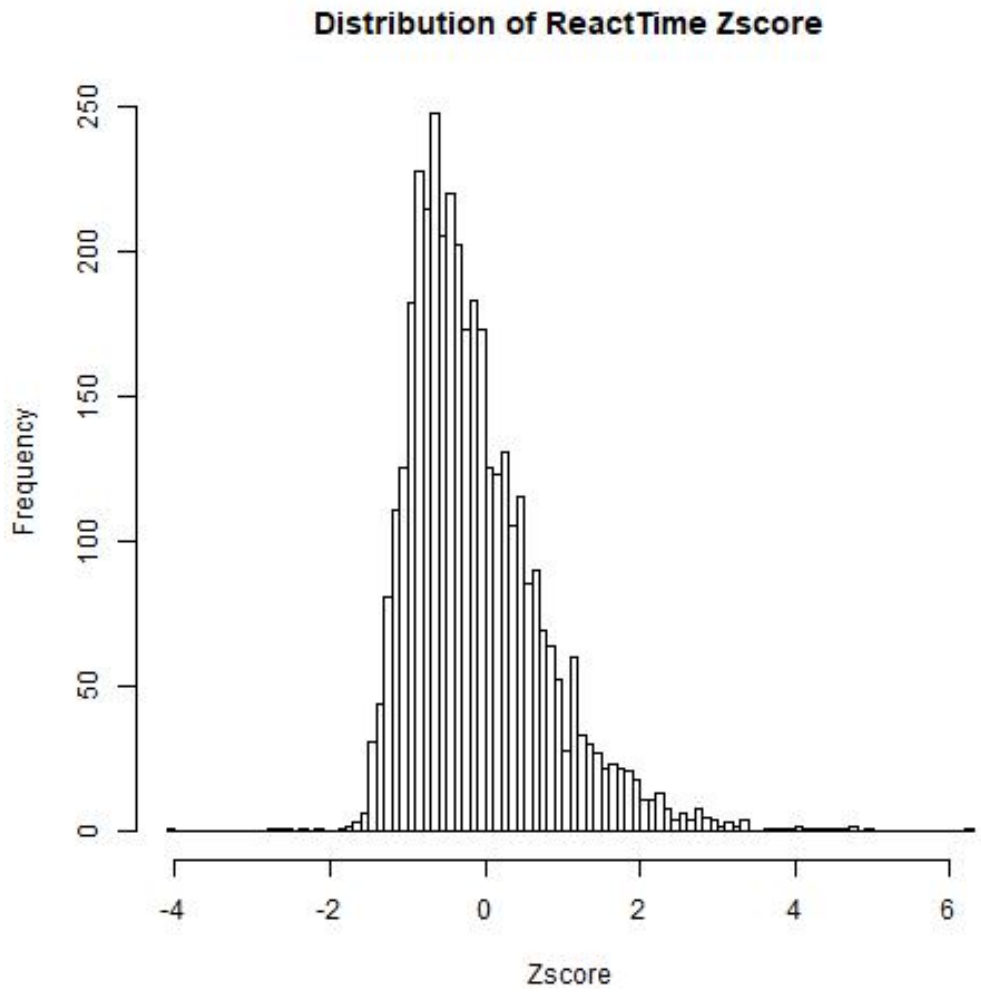
Reaction Time

<u>Reaction time two-choice:</u>

Mean ellapsed time
for response
(over correct answers)

Max:2000 ms
If under 50 ms, scored as anticipation

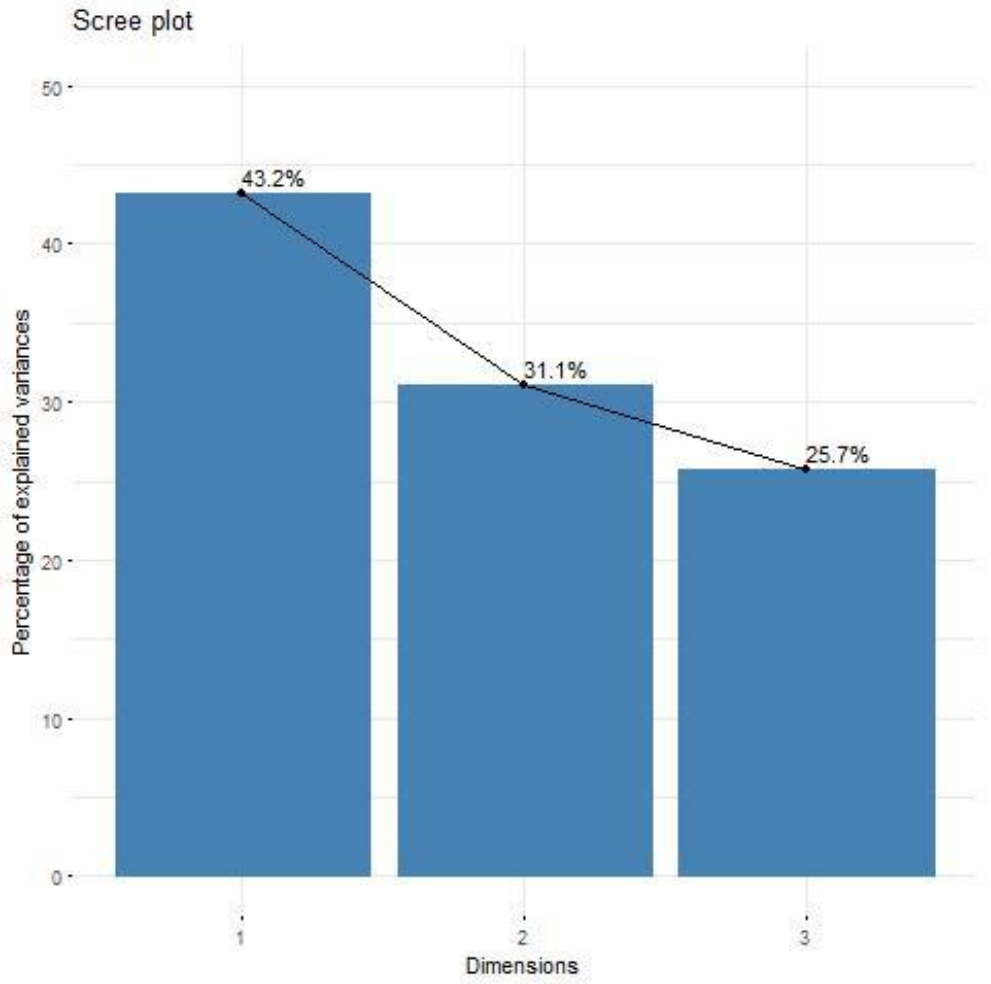
Lower = Better



n=3780

Principal Component Analysis

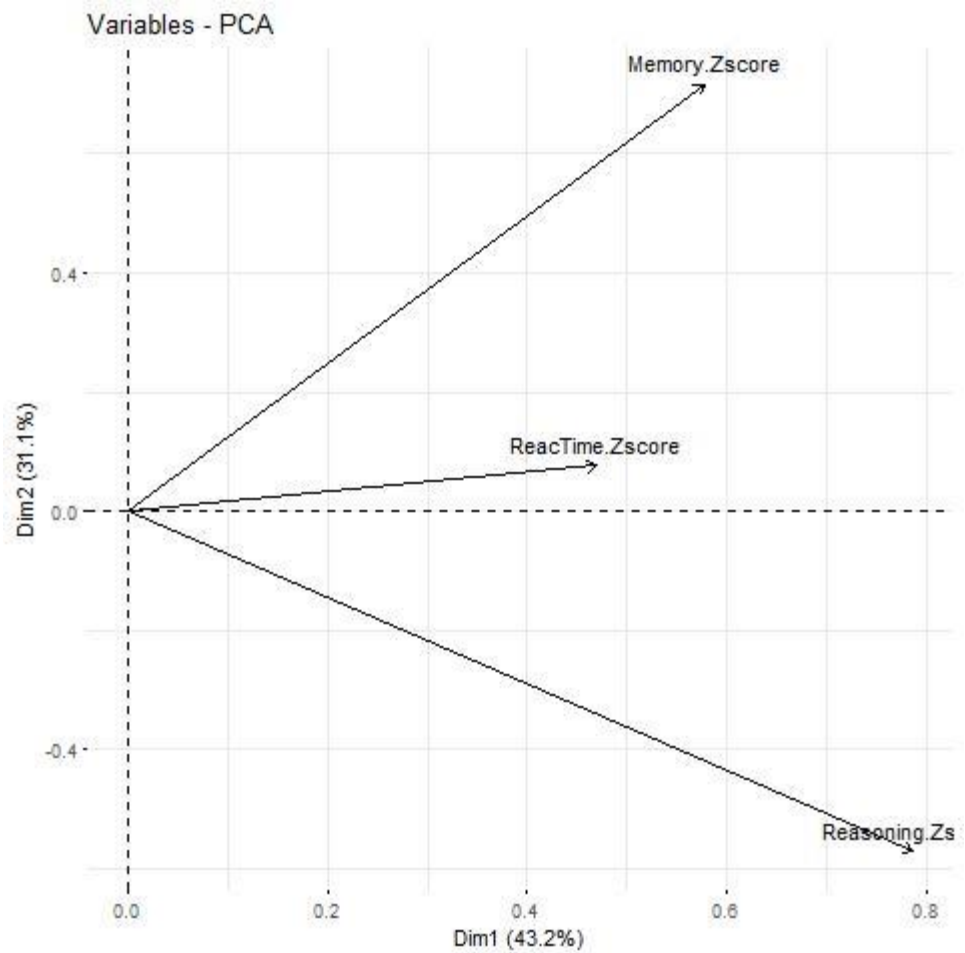
Proportion of variance represented by each PC

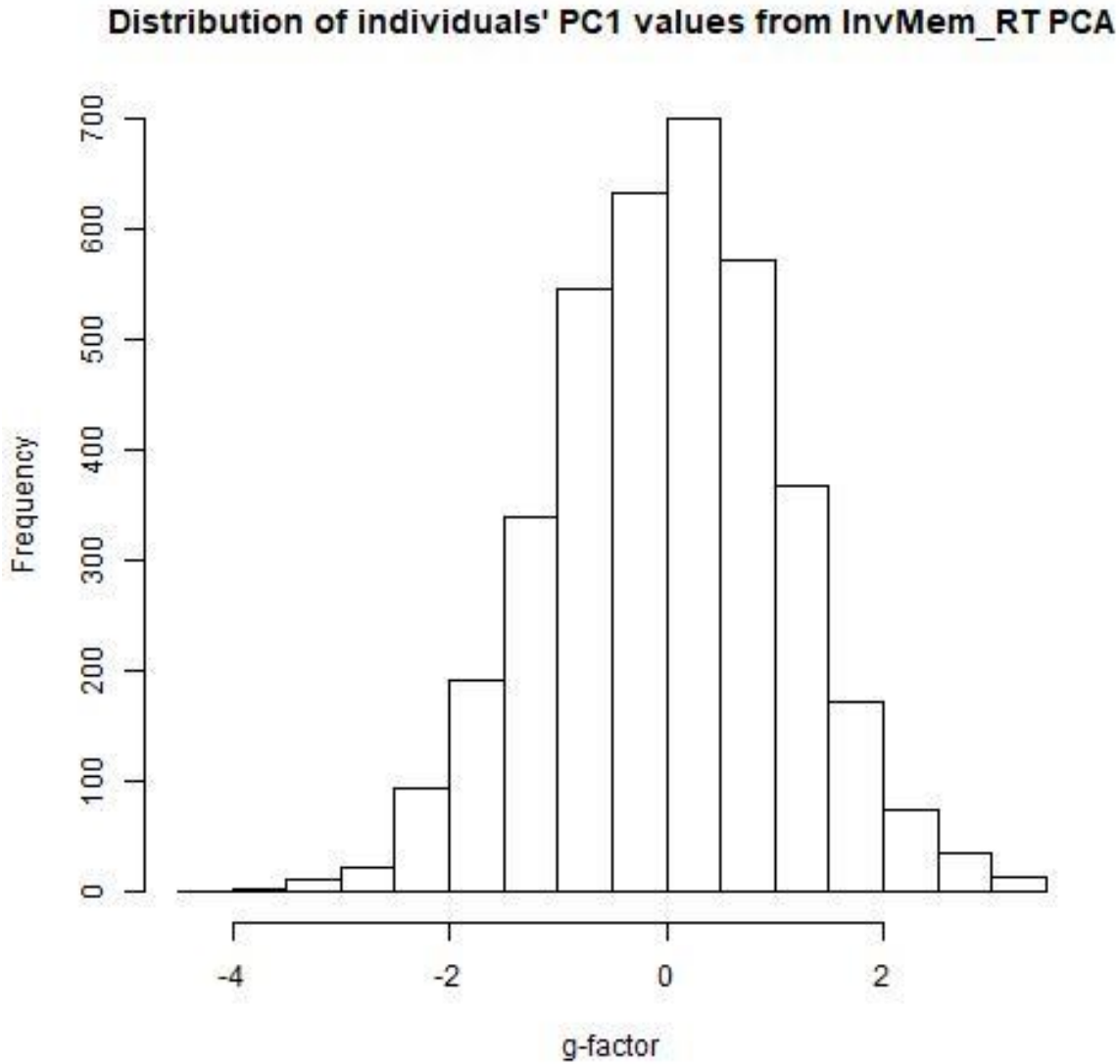


Note: PCA carried out with inverted Memory and Reaction time Z-score (zscore * -1

i.e. all Zscore results become positively correlated with IQ

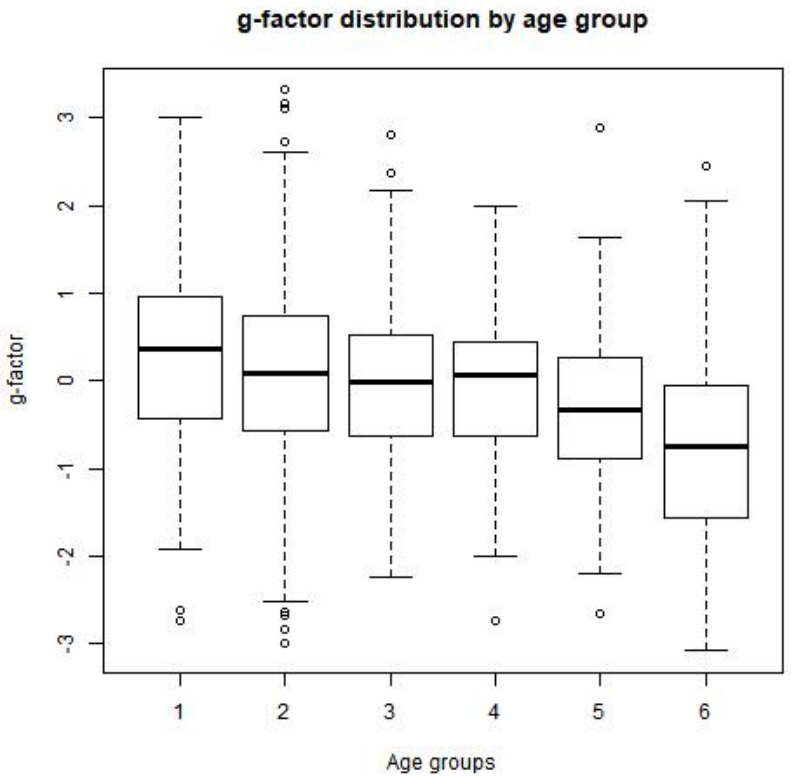
PCA after inversion of Memory and ReactTime CZs only

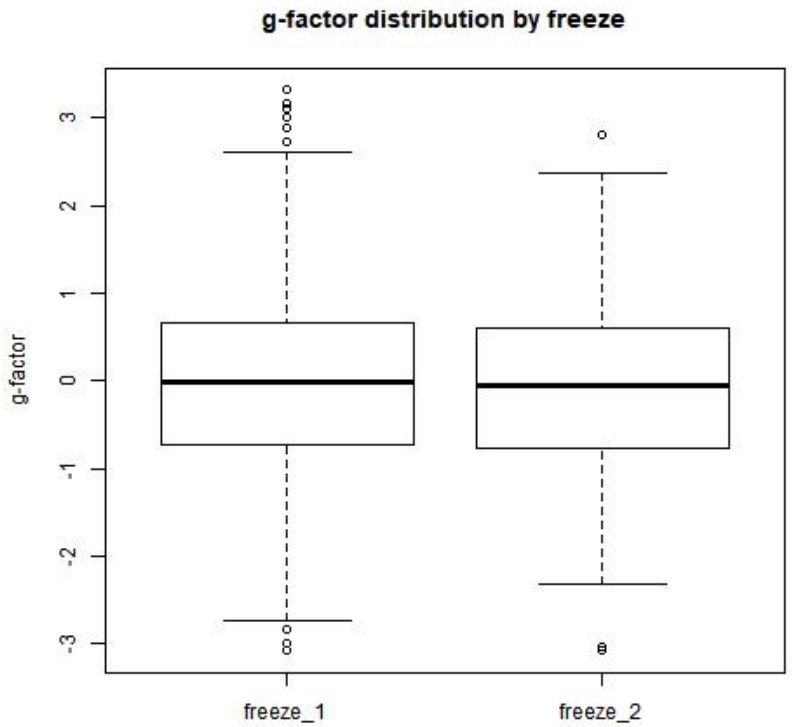


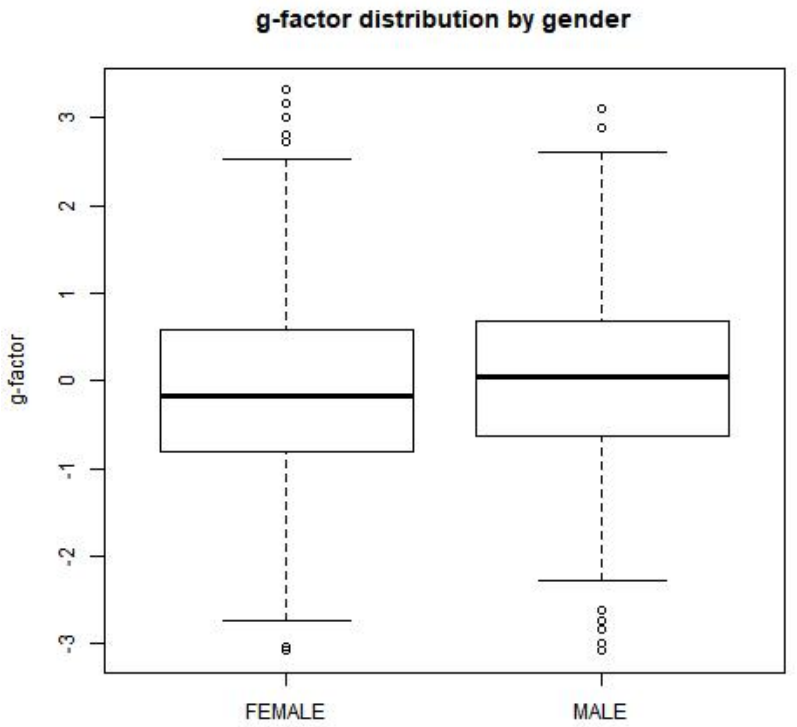


PC1 (Dim1) was used as G-factor. Value representative of IQ (cognitive function).

The g-factor in the individuals analysed by RNA-Seq (n=562)

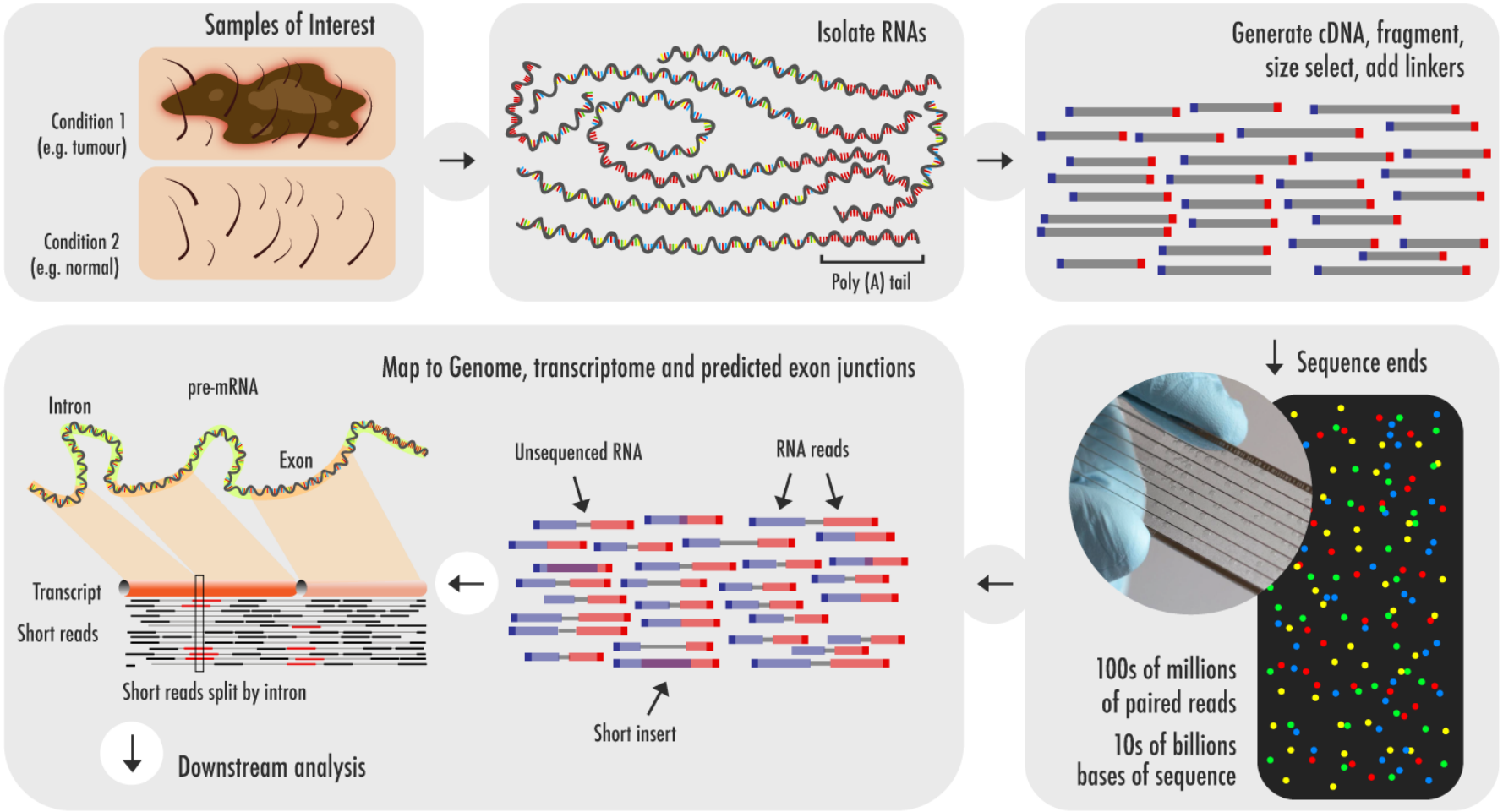






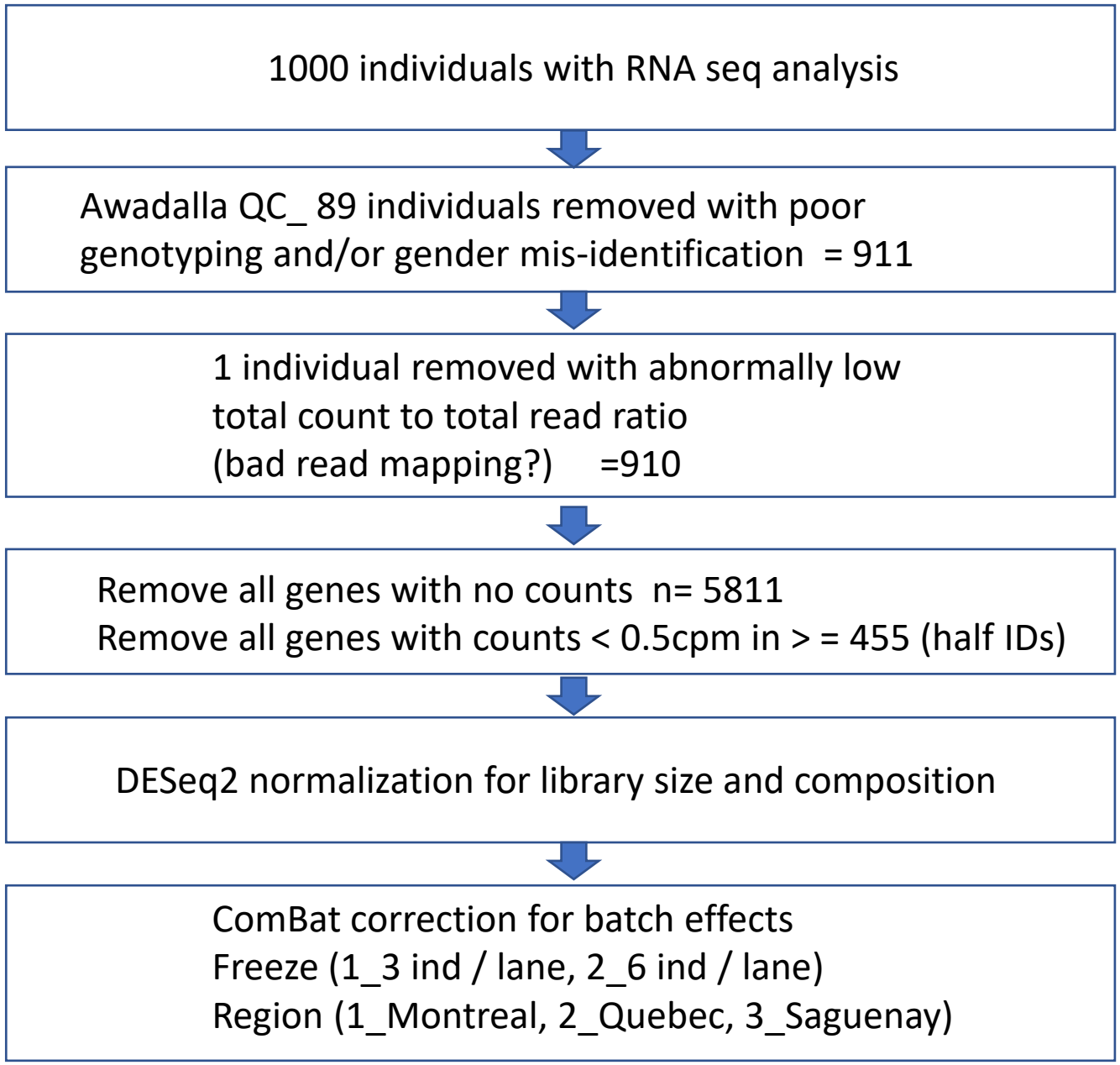
Groups of 5 years
from 40 to 70

RNA-Seq analysis:



- After alignment, number of reads aligned per gene are counted and count is representative of gene expression level (as well as gene length)
- In our case we obtained a count matrix from CaG containing counts for 911 individuals

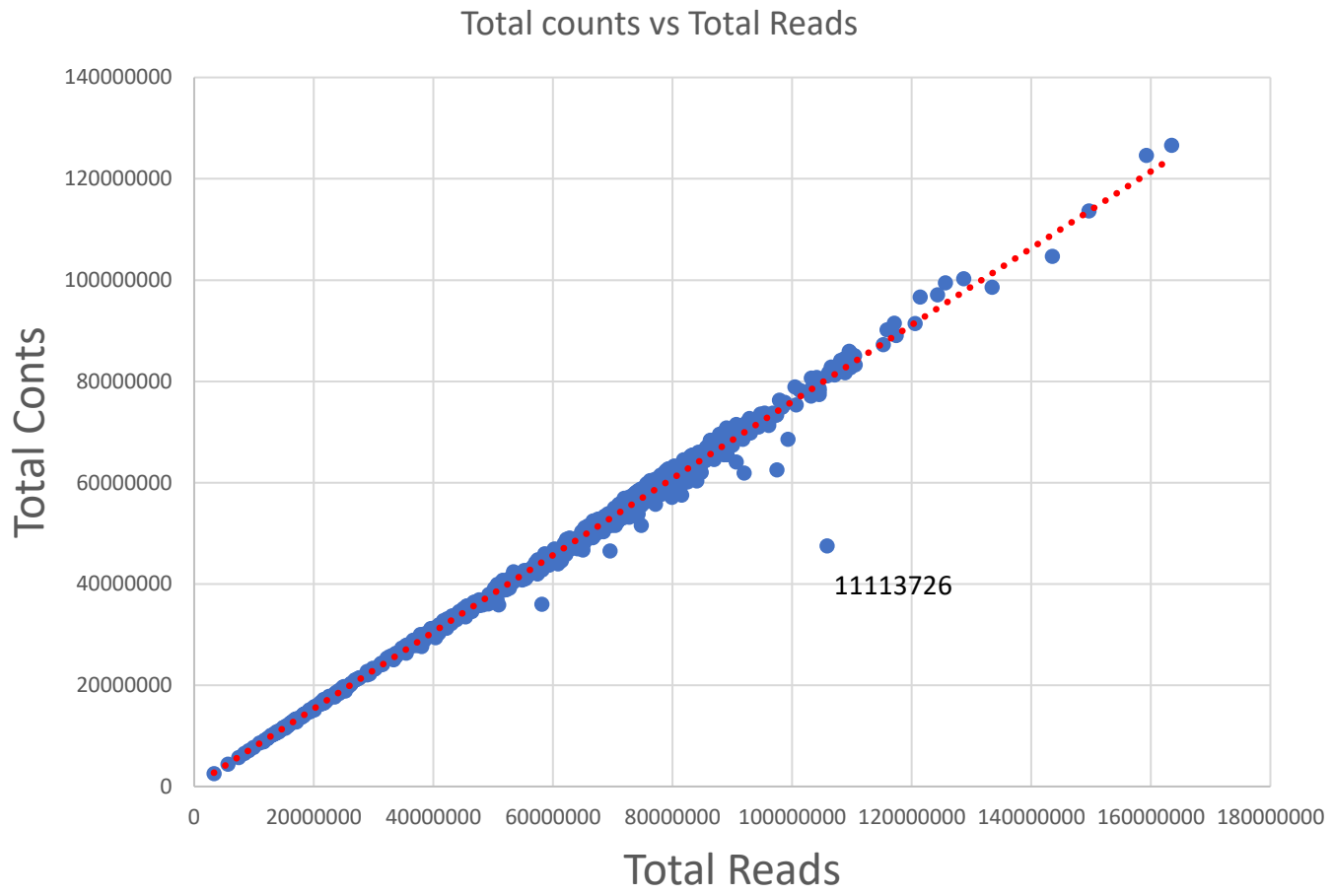
Flow chart of RNAseq data selection and processing:



from
57,773 Ensemble genes
down to
15,647 (« Keep » genes)

Genes with counts > 0.5cpm
in at least one individual
21,744
(« Large Keep » genes)

Compairison of total counts (aligned reads) to total reads making up all libraries (obtained for each sample (individual))



Generally ~75% of reads were aligned successfully to a unique gene

11113726
Count / read ≈ 45%
outlier

Remove low count/read outlier sample #11113726 from matrix (new CaG_matrix n=910 individuals)

Before correcting for possible technical/biological variables affecting counts / gene expression

Remove genes with very low expression (filtering):

Two approaches:

Keep genes with >0.5 cpm in > ½ individuals (455) = 15,647 genes (CaG keep Matrix)

Keep genes with >0.5 cpm in >= 1 individual = 21,744 genes (CaG LargeKeep Matrix)

Normalize matrix for library size and composition:

- Composition means affect of overexpression / underexpression of one gene on apparent expression level of other genes

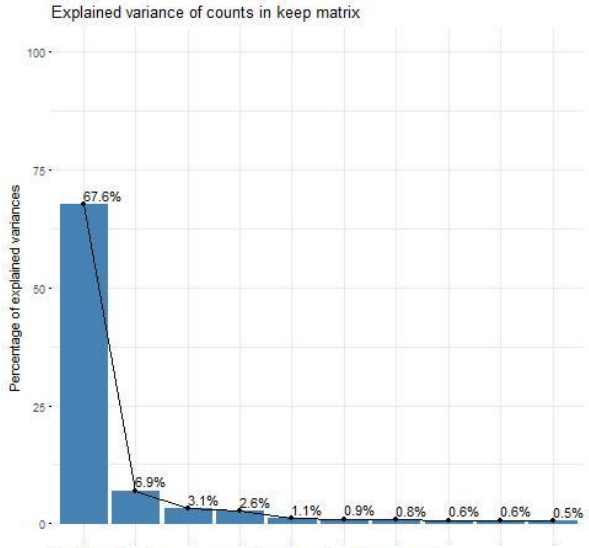
Used DESeq2 software which utilizes scaling factors method

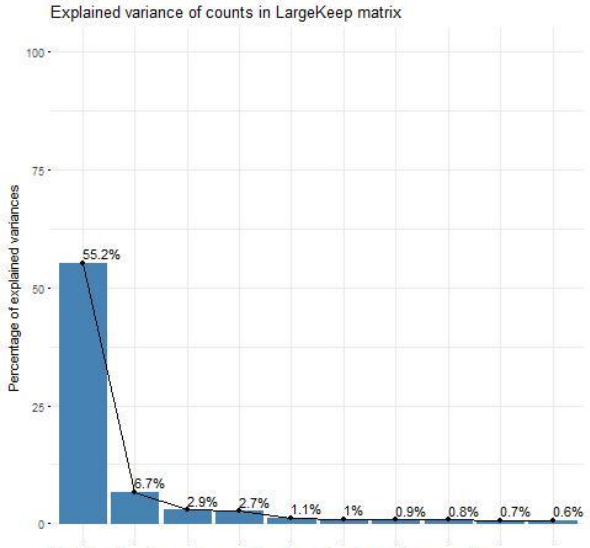
PCA analysis on gene counts to assess affect of normalization of CaG keep Matrix with DESeq2

Keep

Large Keep

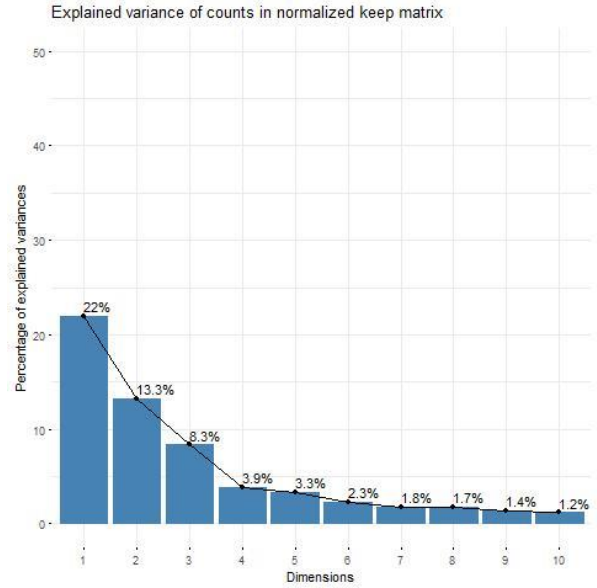
No normalization

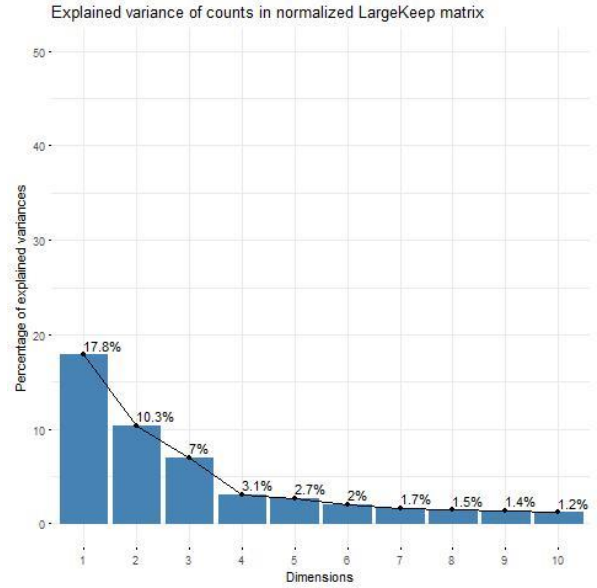




Most likely
a large part of variation
accounted for by PC1
is associated
to library size

DESeq2
Normalization





Other variables to take into acount:

Region:

Montréal	Québec	Saguenay
527	246	138

Freeze:

Freeze1	Freeze2
690	221

Gender:

FEMALE	MALE
456	455

Age:

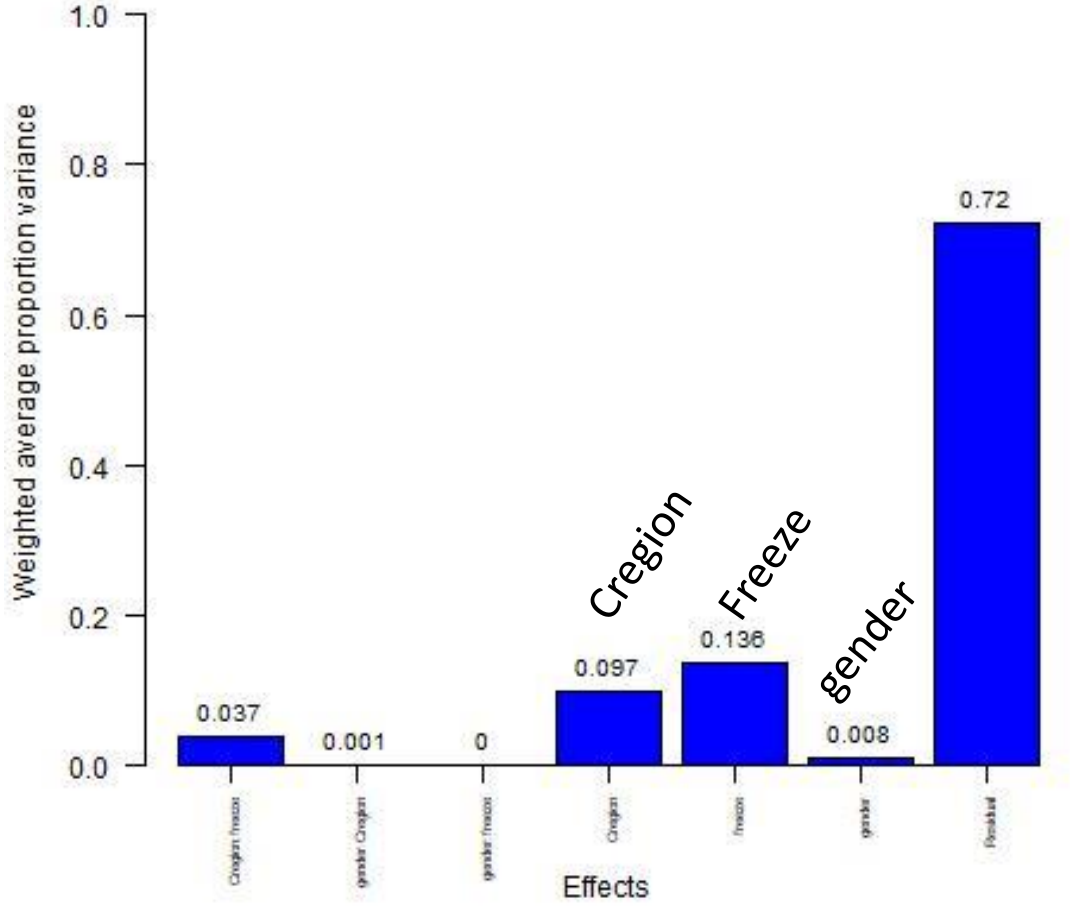
1	2	3	4	5	6
40-44	45-49	50-54	55-59	60-64	65-70
196	212	155	103	102	143

Hematology

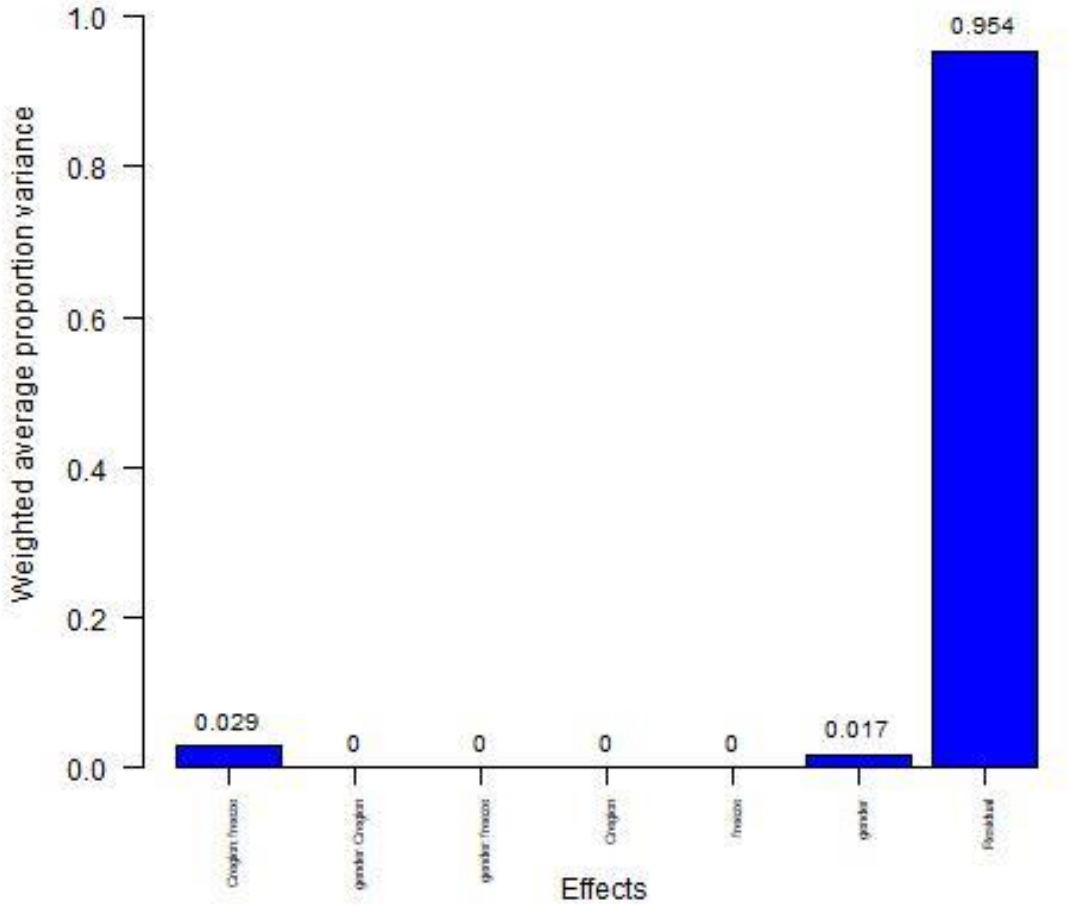
Favé *et al*. (2018) *Nat Commun* 9(1):827

Evaluation of factors (Effects) influencing the variance in the Keep expression matrix before and after ComBat correction for Freeze and region of origin

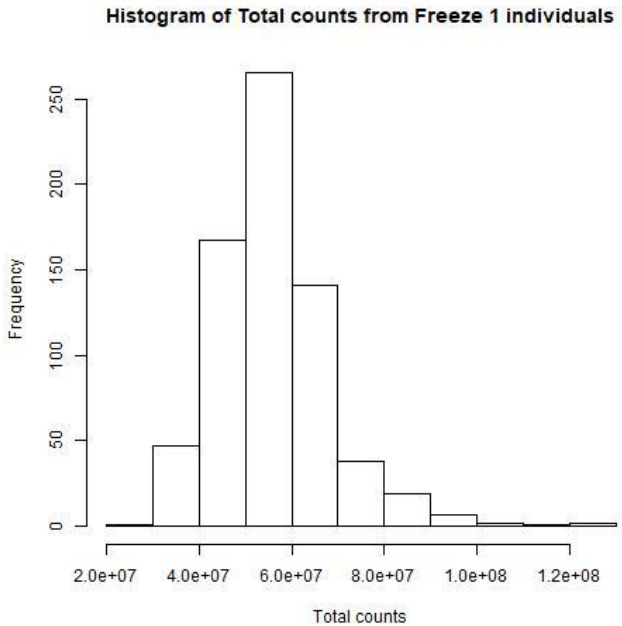
Before

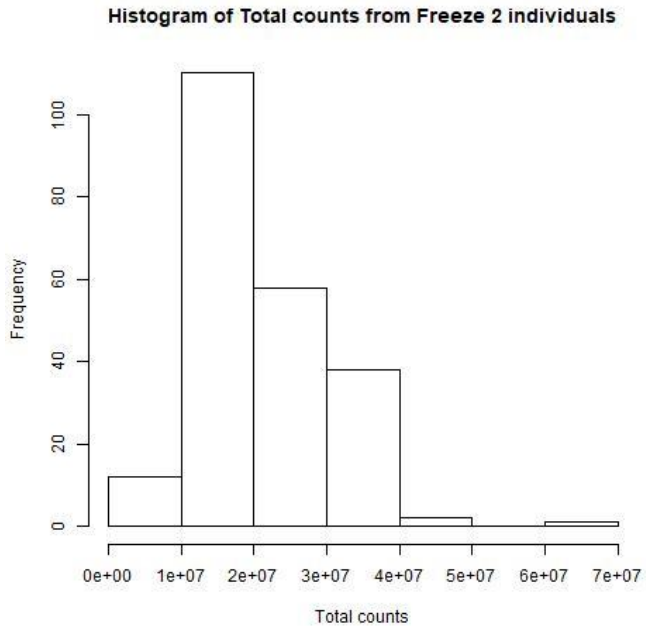


After

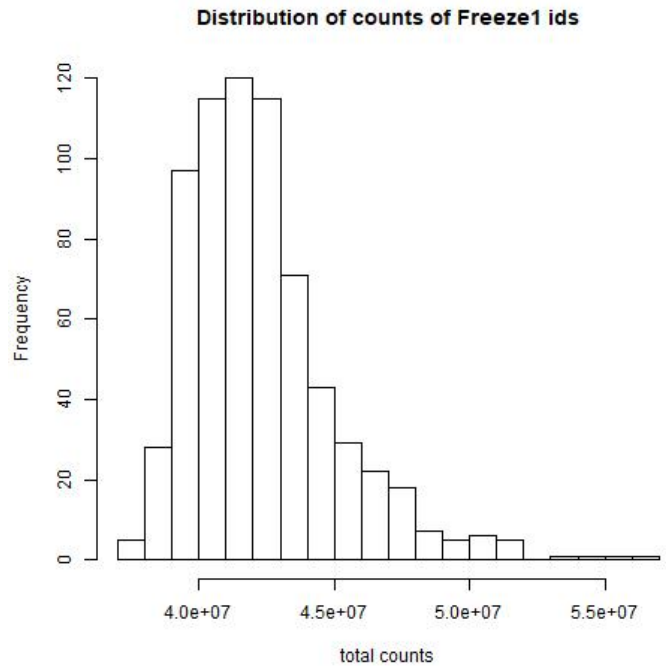


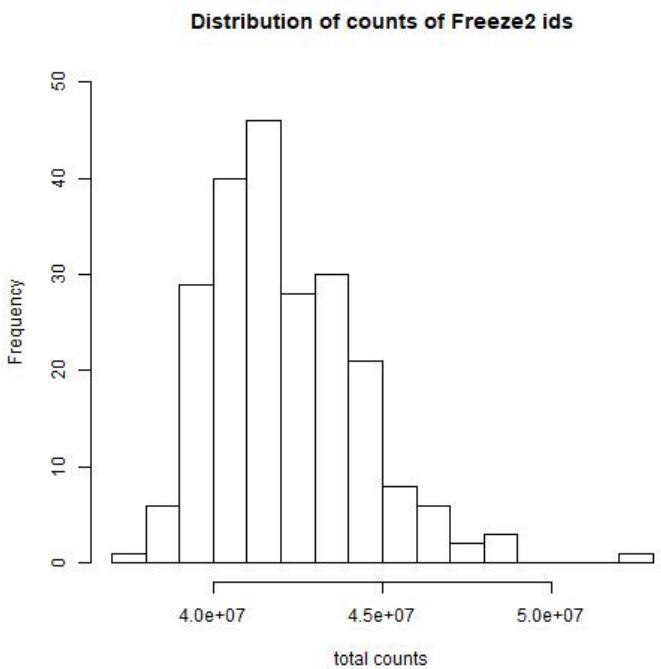
Raw Counts



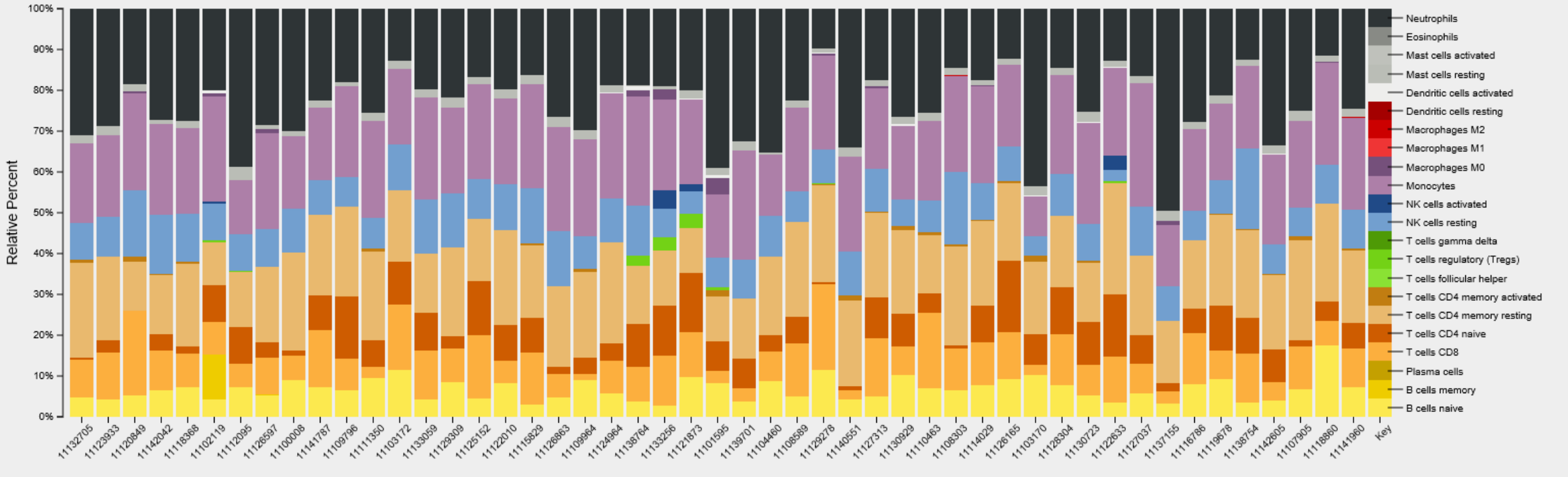


Normalized And ComBat corrected counts





Relative amount of different blood cell types in samples (individuals)



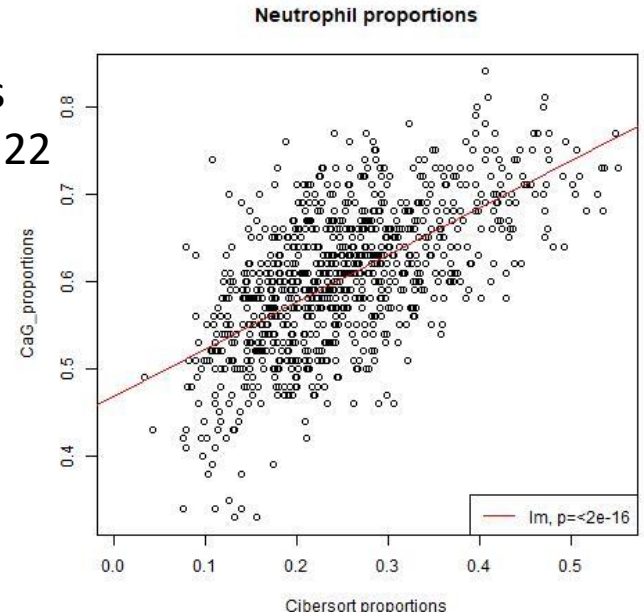
Representative image of 99 /910 individuals analysed by RNA-Seq

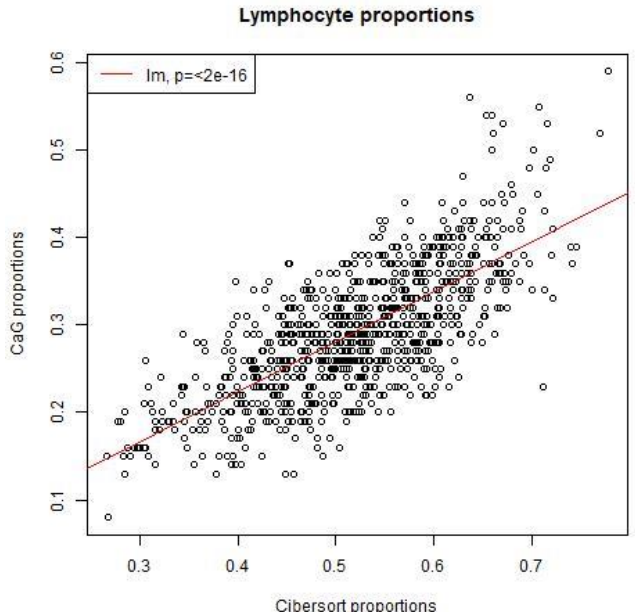
Quantification carried out with Cibersort with expression counts of 524/547 genes in LM22 signature matrix

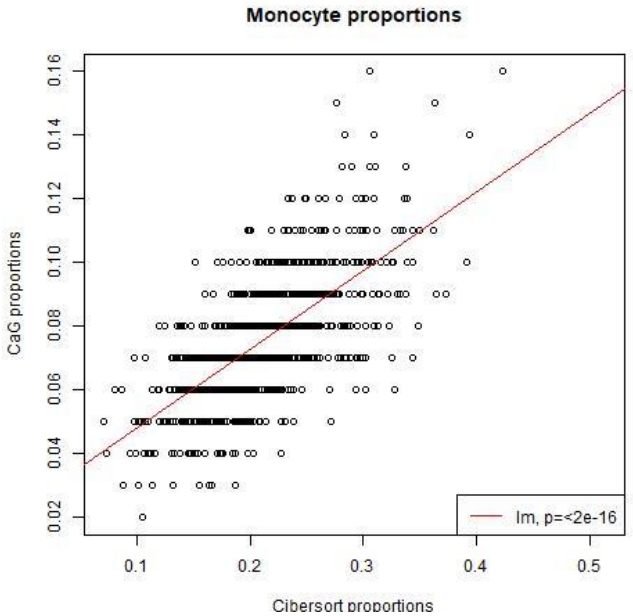
Used this data initially to do last modifications on CaG count matrix using a linear regression model Including age and gender before converting counts to Zscores.

Comparison of blood cell type proportions from Cibersort and CaG

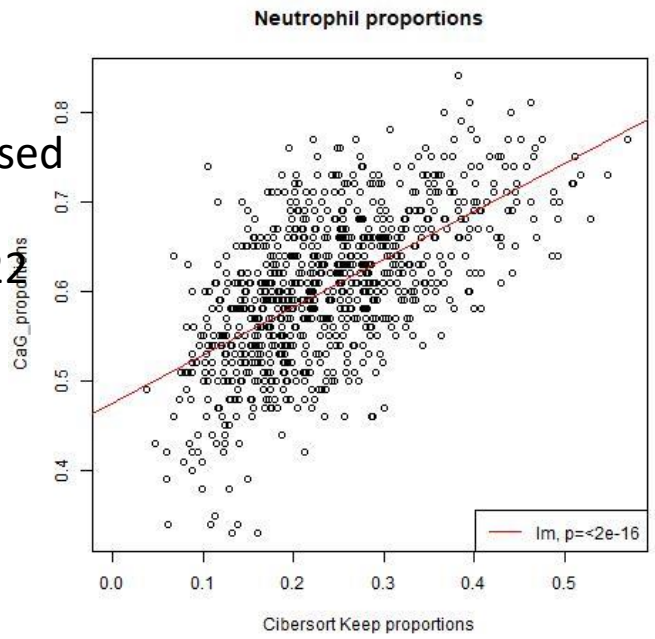
~525 genes of LM22

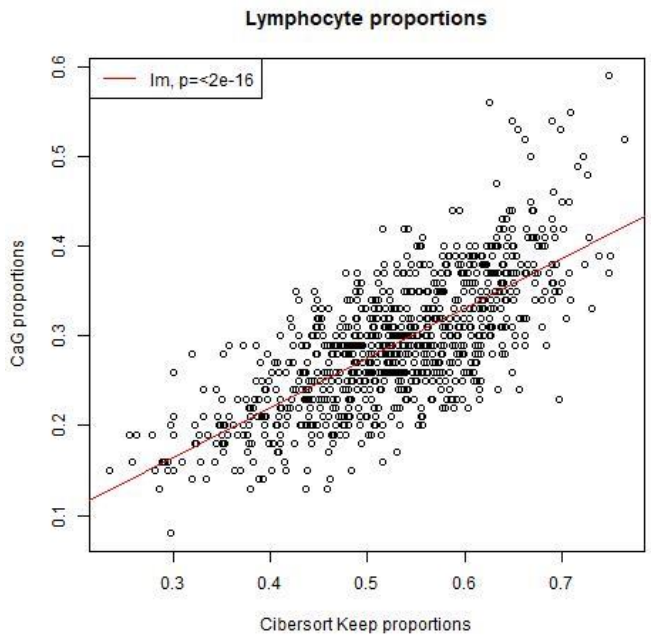


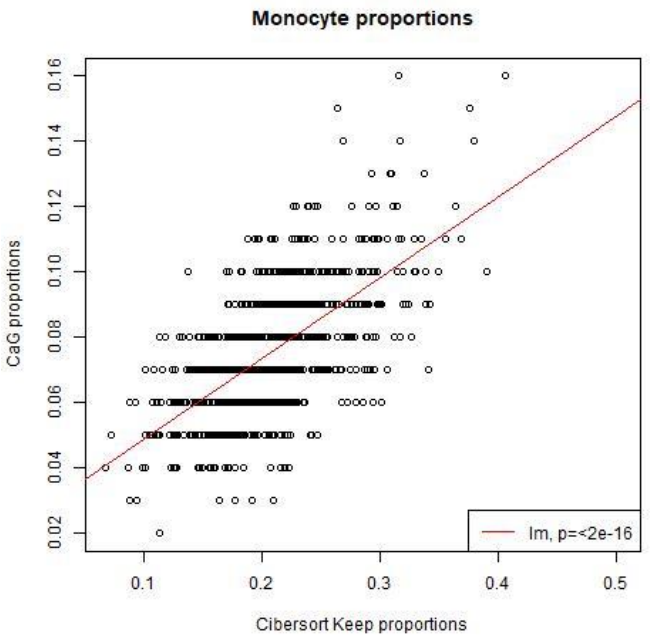




~425 Well expressed genes of LM22

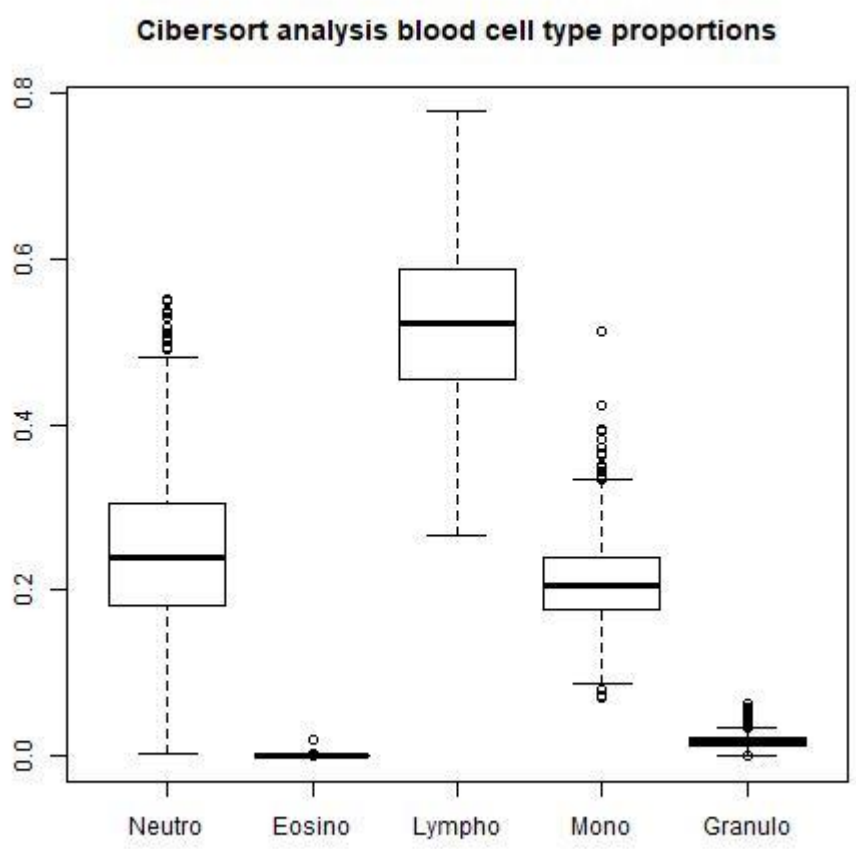




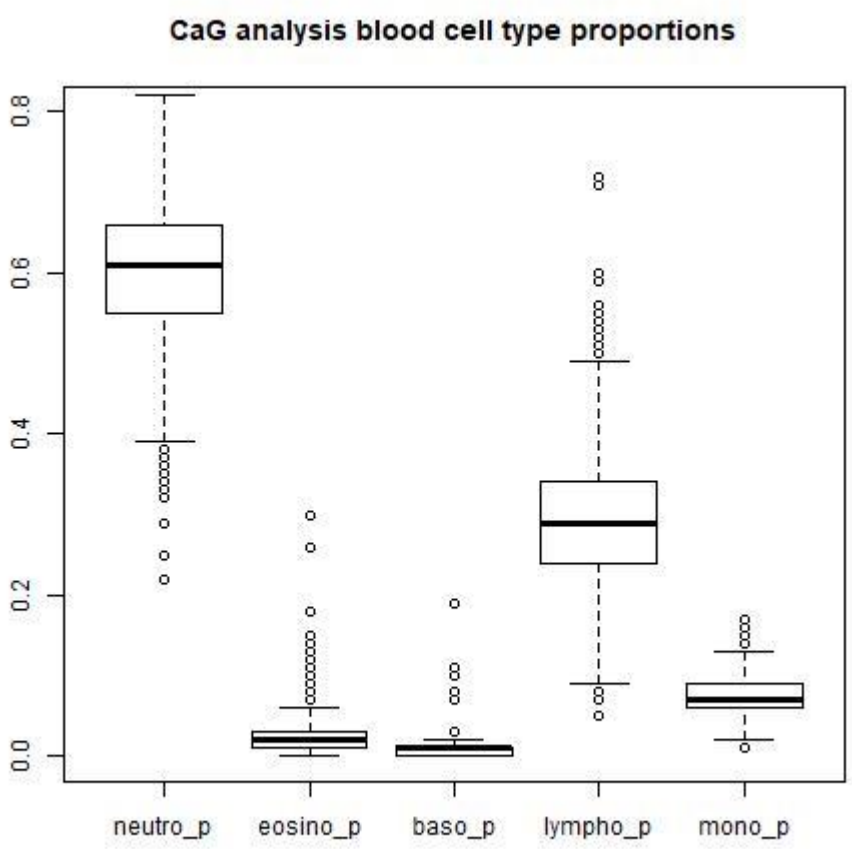


Comparison of blood cell type proportions from Cibersort and CaG

Cibersort prediction data



CaG experimental data



For cibersort counts matrix correction
I used 7 most important cell types
(1 B cell,
3 T cell,
1 NK,
Monos,
and Neutros)

For CaG analysis, 14 individuals with RNASeq
data have no blood data (i.e. 896 individuals with both data vs 910 with cibersort)

(Because of this I lose 24 CNVs fom 2494 CNV data file)

For cibersort groups:
Lymphos = all B cells, T cells and NK cells
Monos = Monocytes, Macrophages and Dendric cells
Granulos = Mast Cells

Linear regression models used to correct for blood composition, Age and Gender

Correction model:

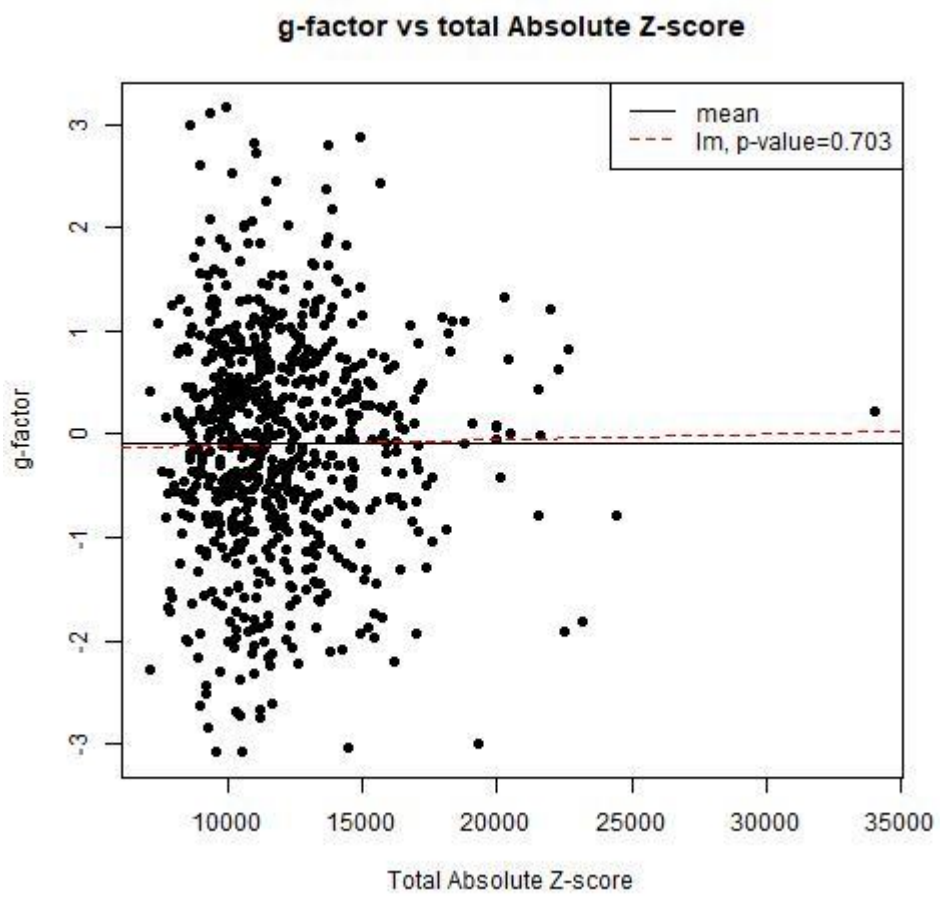
counts_lm= lm(tCaG_ComBat_CaG_Blood_RNASeq_noNA_IDs[,i]~

Neutros+Eiosinos+Basos+Lymphos+Monos+Age+gender)

After correction, convert count values to Z-scores

Global effect analysis:

Look for relationship between G-Factor and Total Absolute expression Z-score

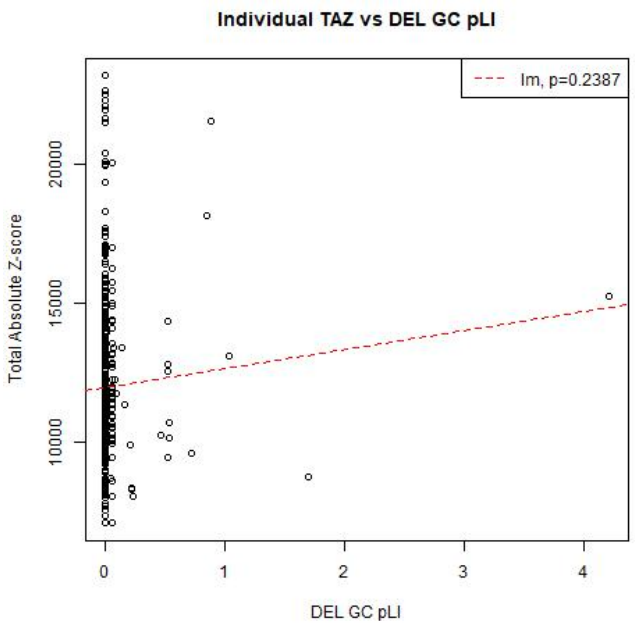


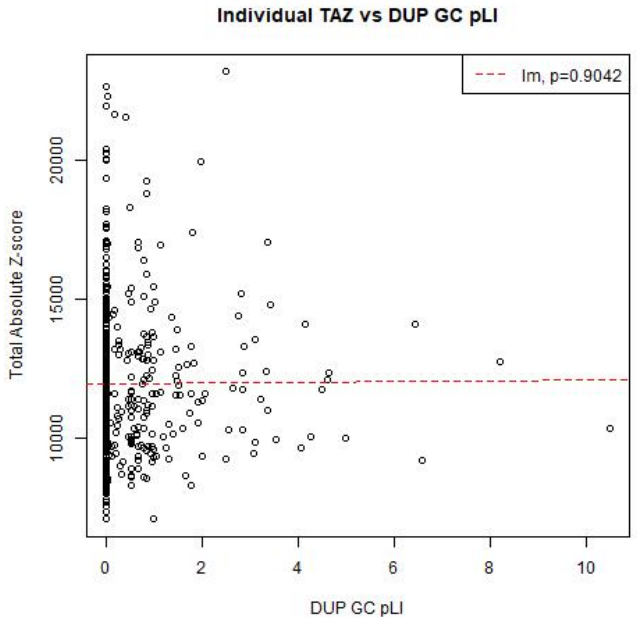
Counts for each gene converted to Z-scores

Z-score of 0 means average expression level,
Z-score above 0 (+) = overexpression,
Z-score below 0 (-) =underexpression

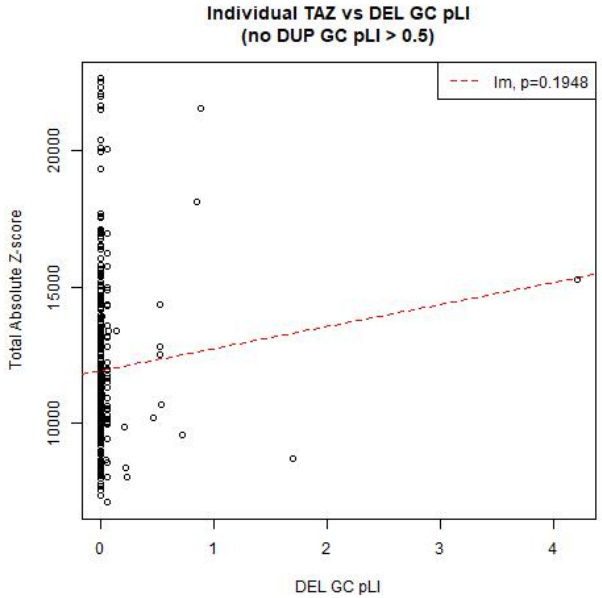
Total absolute Z-score represents overall level of deregulation

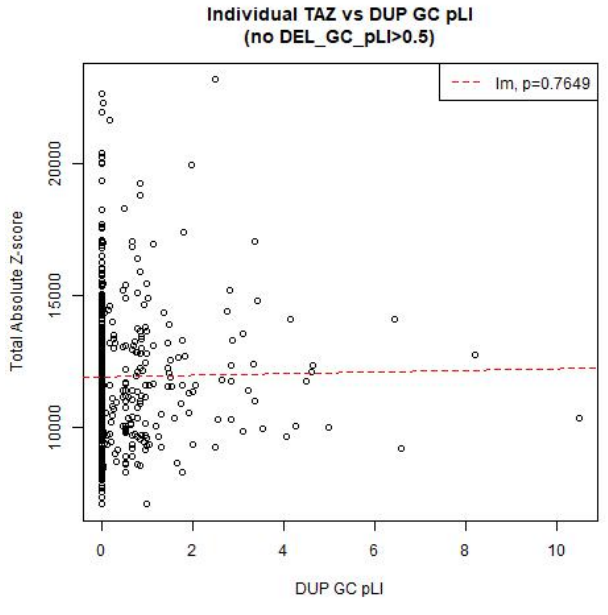
Look at relationship between TAZ and total GC pLI of individuals





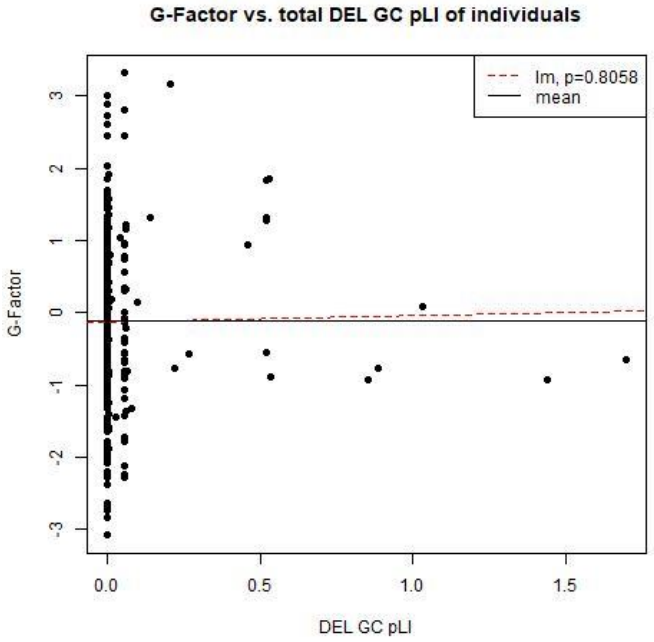
GC = gene complete
i.e. count only pLI of genes completely within CNV

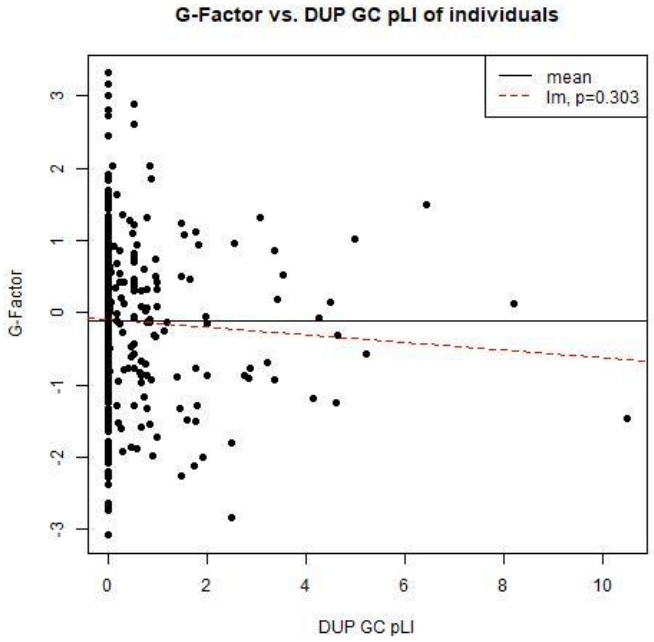


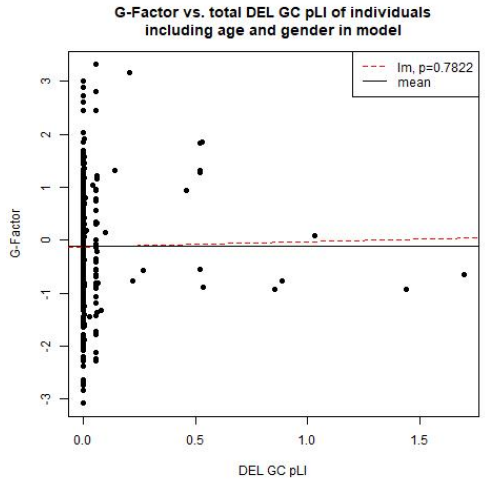


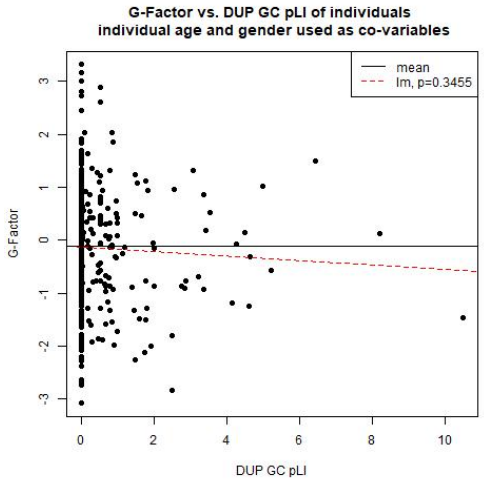
Also no significant correlation with non-GC pLI
or after removal of high pLI outliers

Look for relationship between G-Factor and pLI of individuals



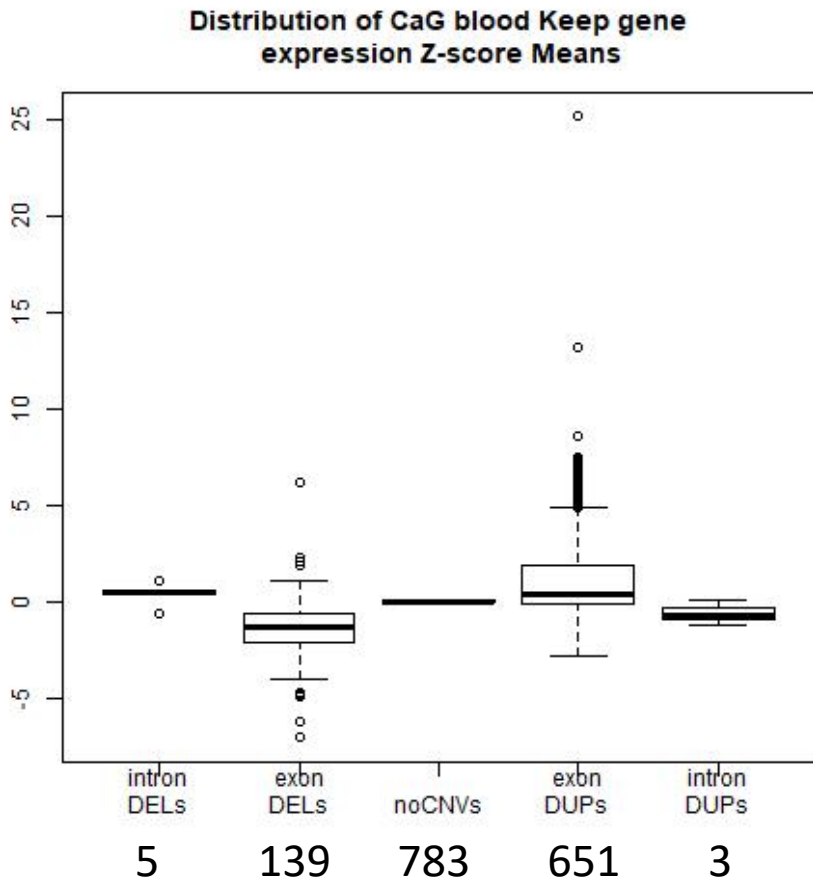


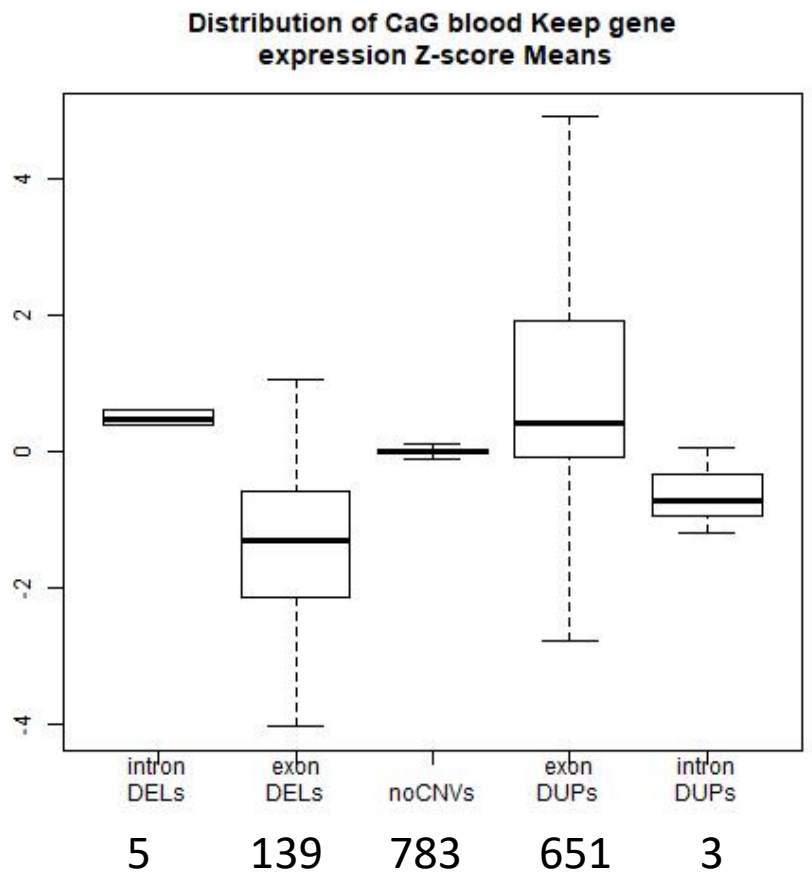




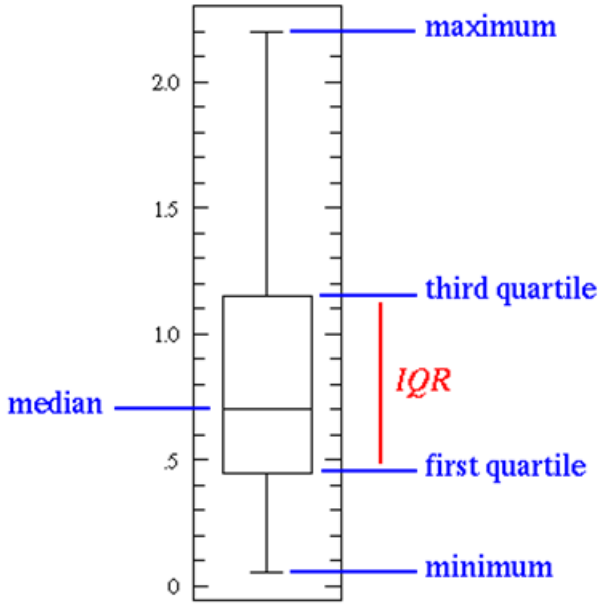
Comparison of Mean ExpZ-scores for genes within CNVs vs not in CNVs

2470 good CNVs total
in 550 individuals
(avg of 4.5 CNV / individual)





Boxplot interpretation

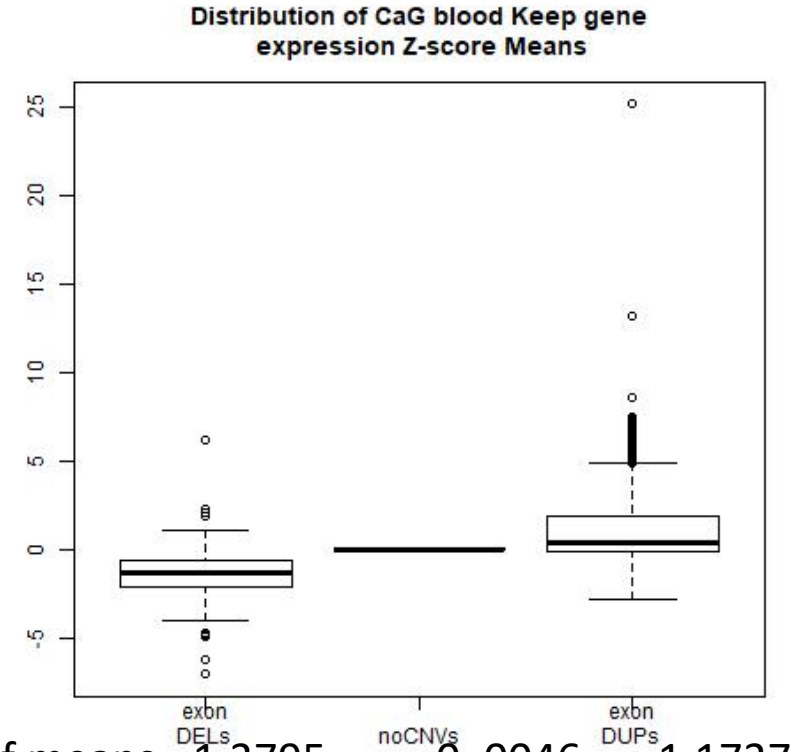


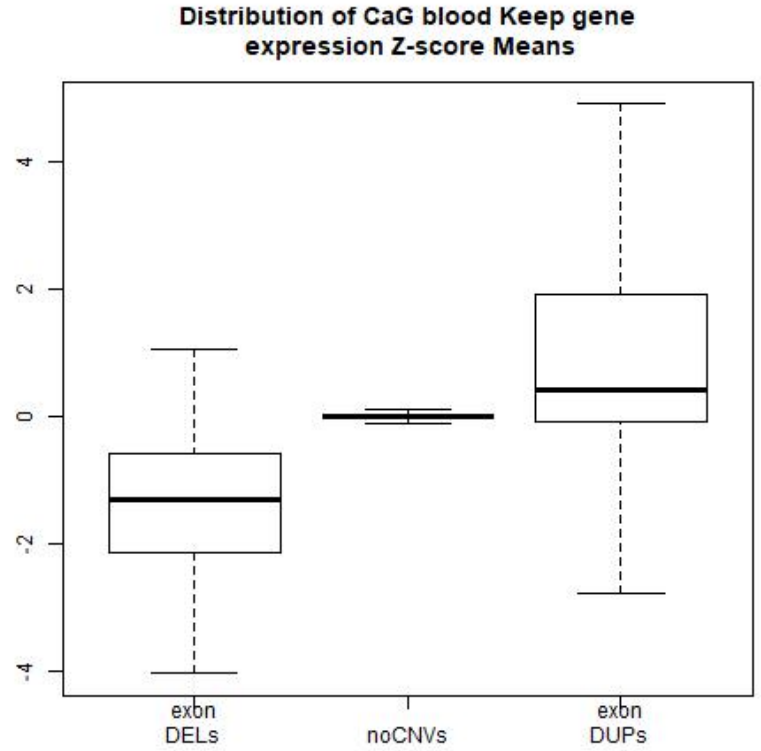
783 genes

Note: some individuals have combinations of genes in different CNV class
exon_CNV= exonic, 5'UTR, ncRNA_exonic, intron_CNV= intronic, ncRNA_intronic

Comparison of Mean ExpZ-scores for genes within Exon CNVs vs not in CNVs

783 genes





Means of means	-1.3795	0 .0046	1.1737
No of genes	139	783	651

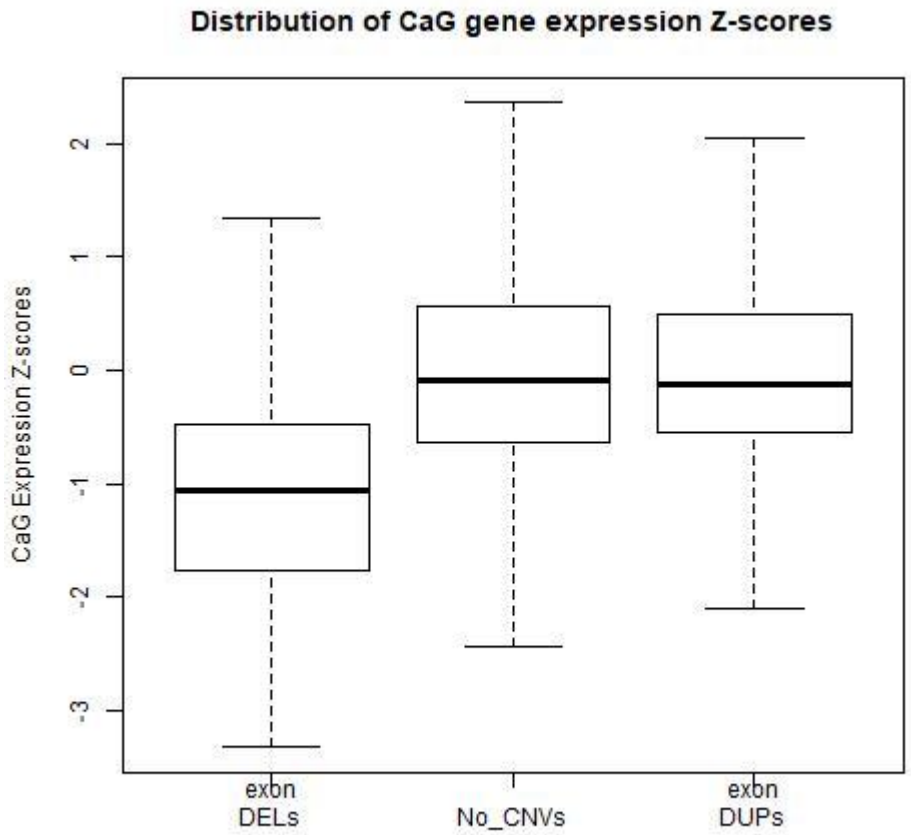
Medians of means	-1.1848	0 .0078	0.4840
No of genes	139	783	651

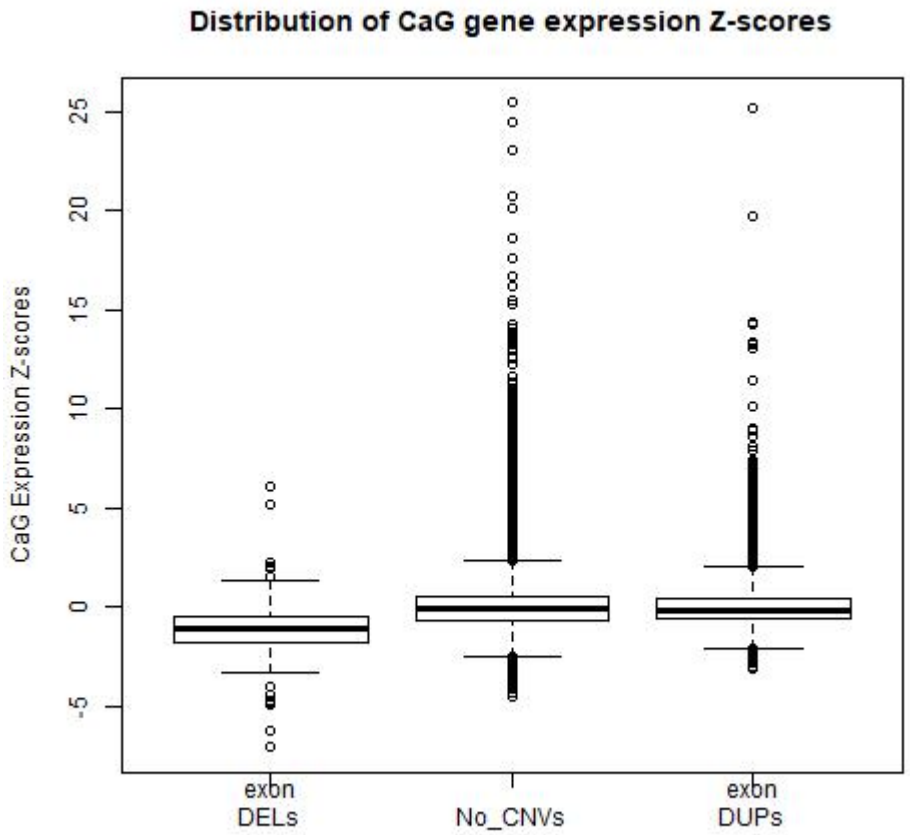
Analysis of CaG expression Zscore means:

Kruskal-Wallis (non-parametric test for difference between means) p-value = 3.256469e-76

Pairwise Wilcoxon tests: DEL ~ NoCNVs p= 1.553495e-48 DUP~NoCNVs p= 5.171651e-31 DEL~DUP p= 1.188058e-48

Comparison of individual CaG ExpZ-scores for genes within Exon CNVs vs not in CNVs





Mean Z-score	-1.0866	0.00433	0.41280

Analysis of CaG expression Zscore means:

Kruskal-Wallis (non-parametric test for difference between means) p-value <2.2e-16

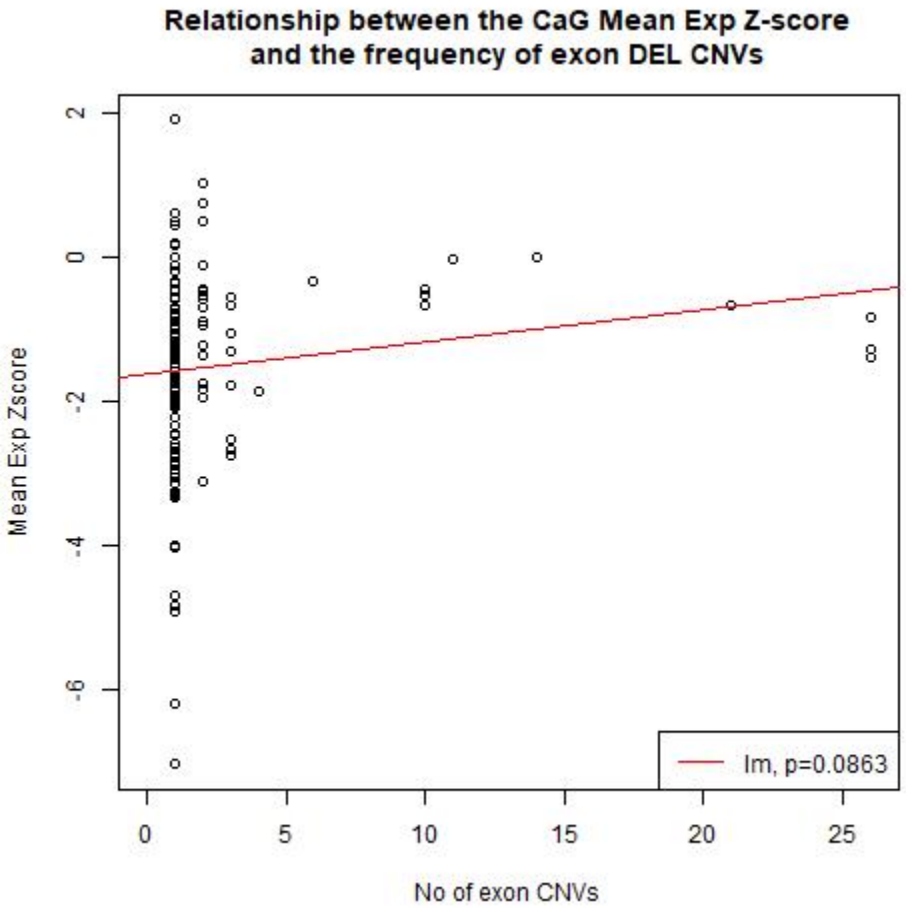
Pairwise Wilcoxon tests: DEL ~ NoCNVs p <2.2e-16 DUP~NoCNVs p 0.01 DEL~DUP p <2.2e-16

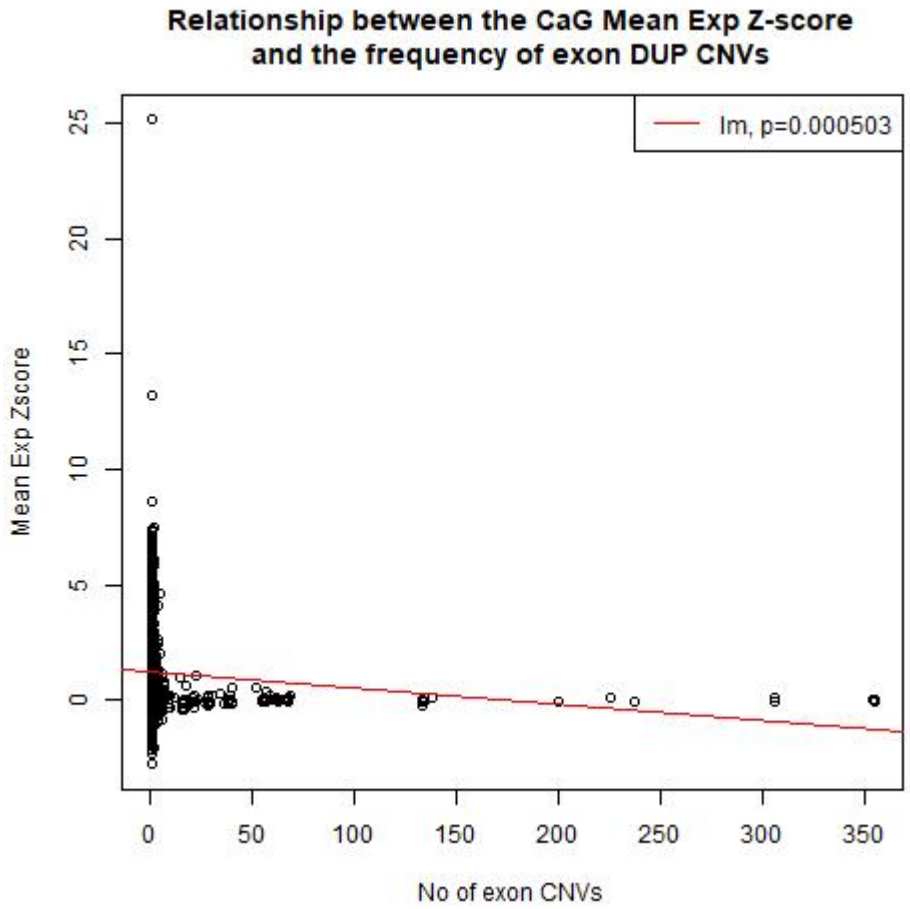
Relationship between mean ExpZscores of genes and their frequency in CNVs

DEL CNVs

DUP CNVs

Mean ExpZcores of 135 genes

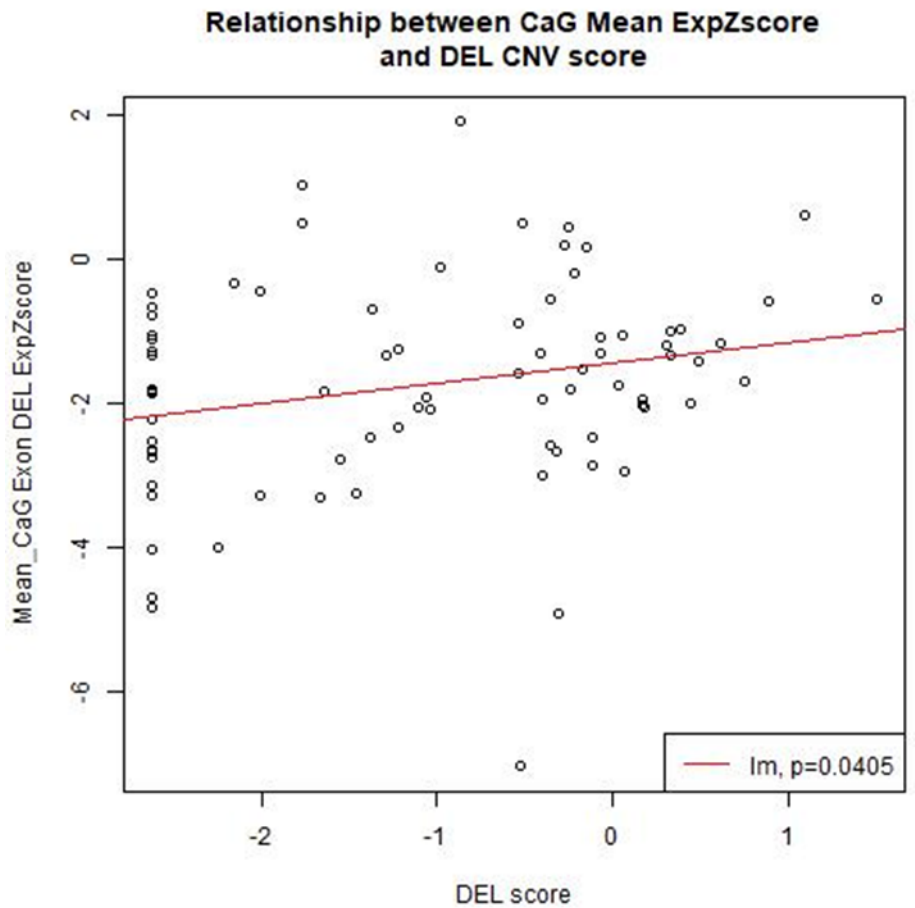


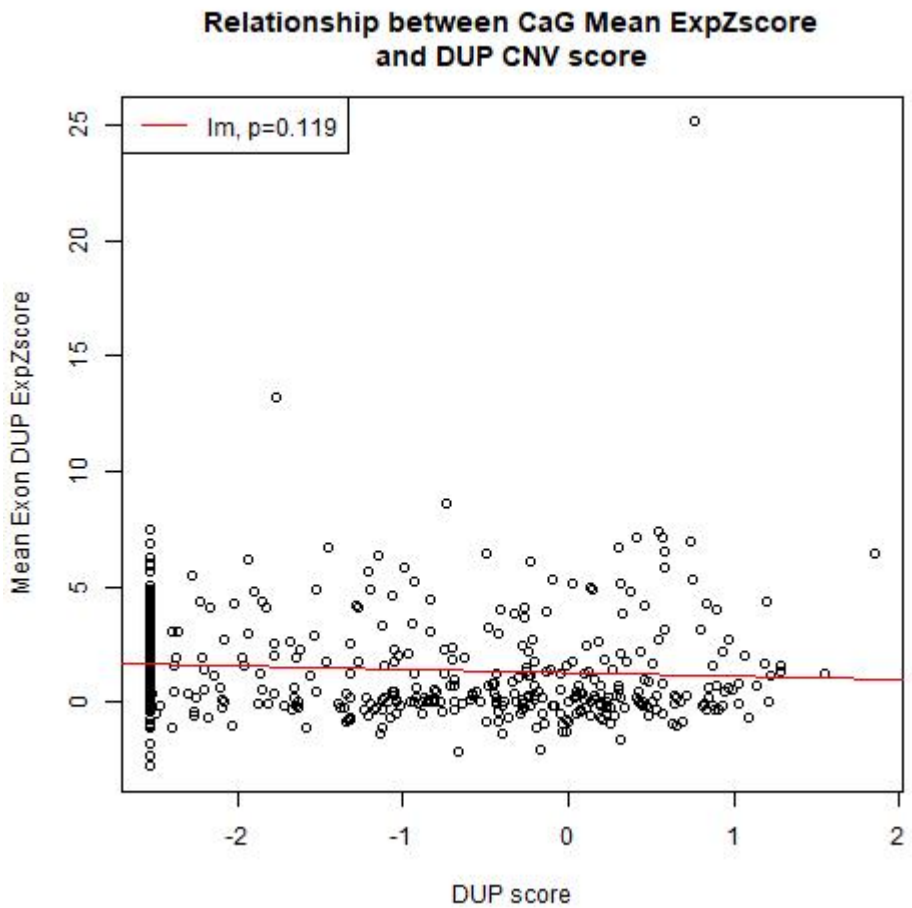


Mean ExpZscores of 651 genes

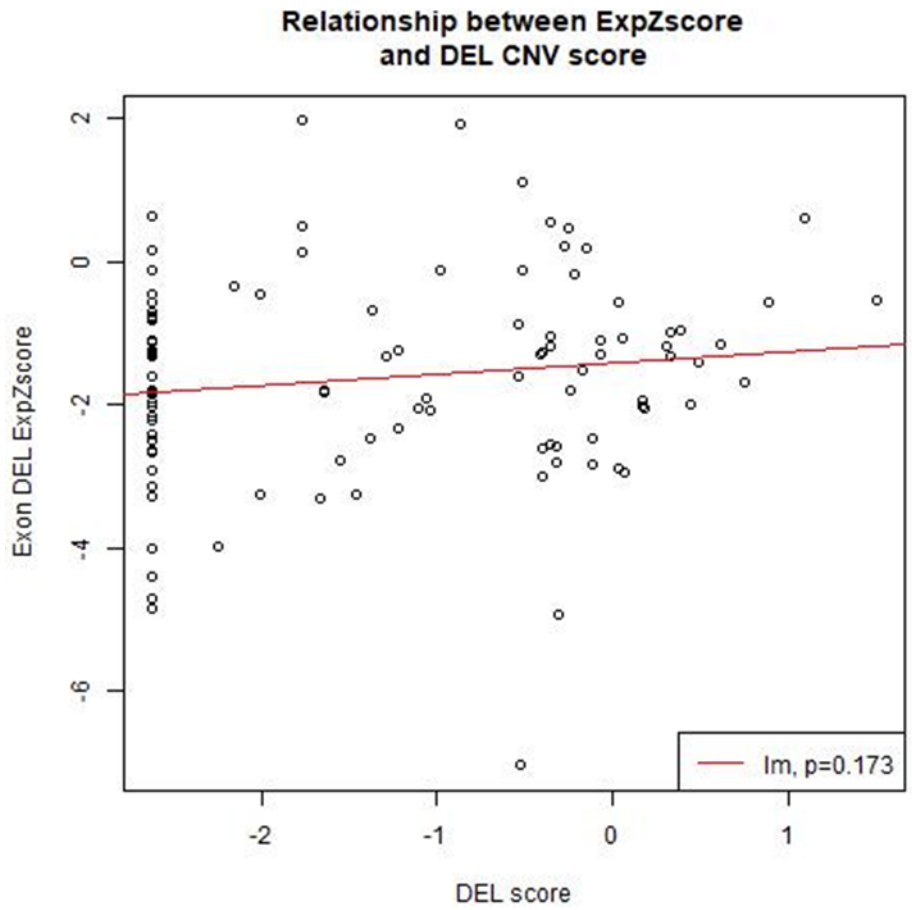
IDs with CNVs, with DEL CNVs, with DUP CNVs
frequency of CNVs

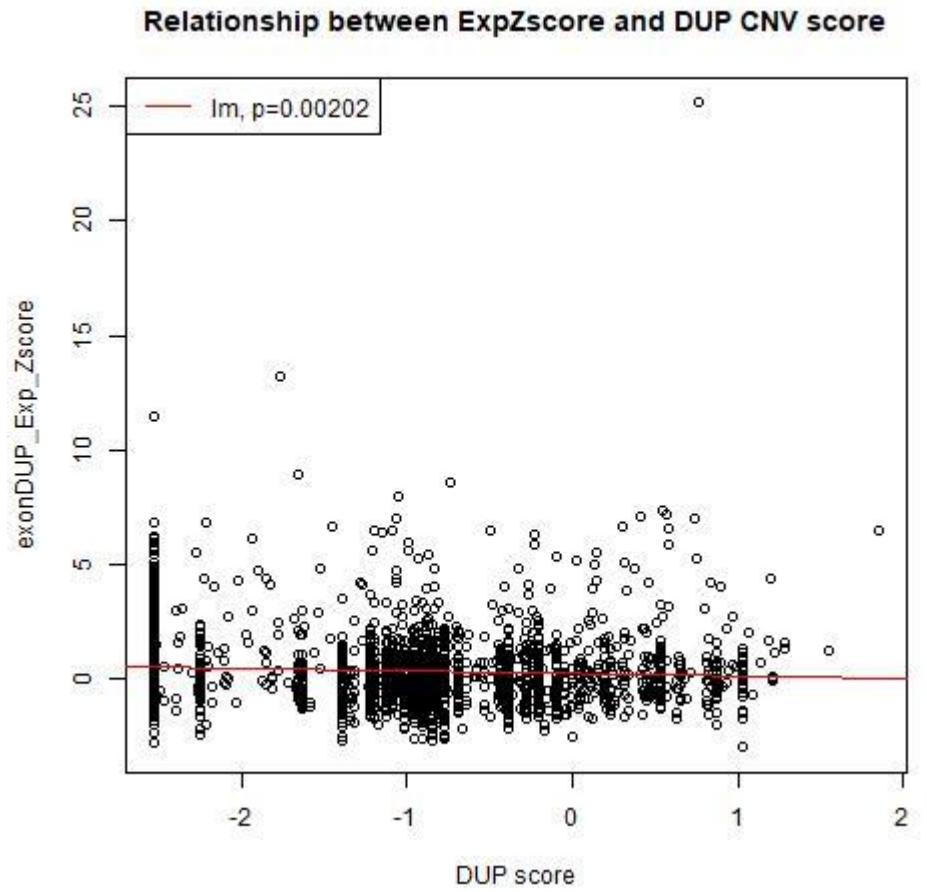
Relationship between Mean CaG exp Zscore vs CNV score



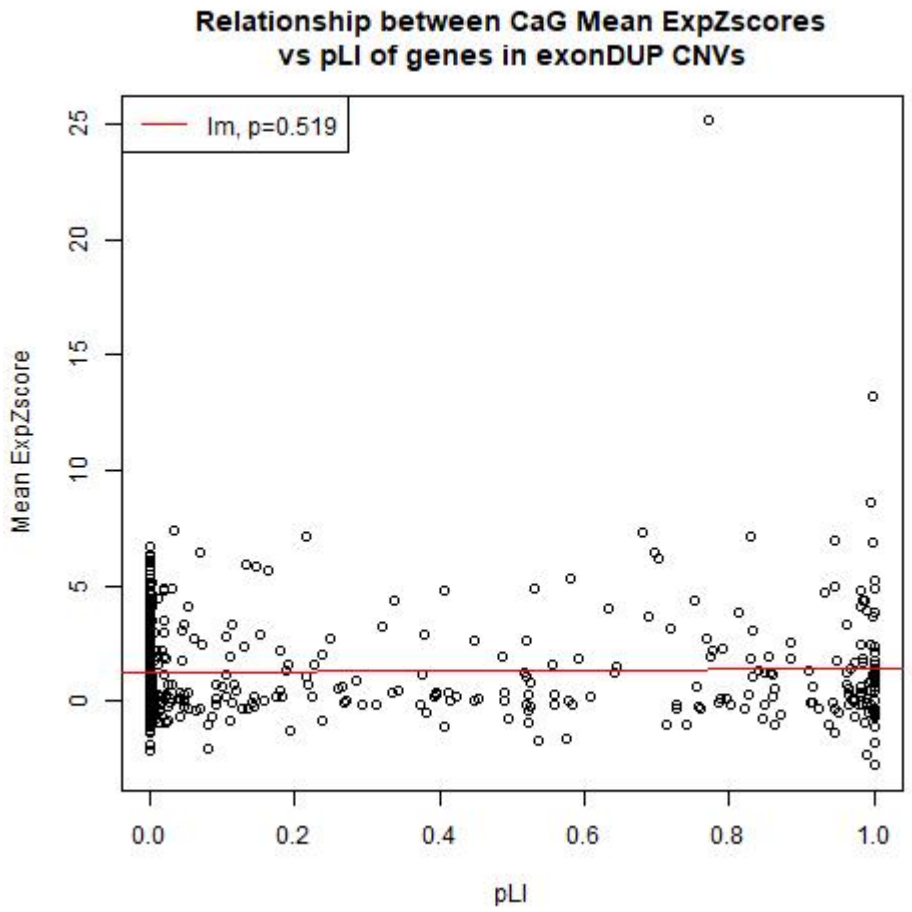


Relationship between individual CaG CNV gene ExpZscores and CNV score



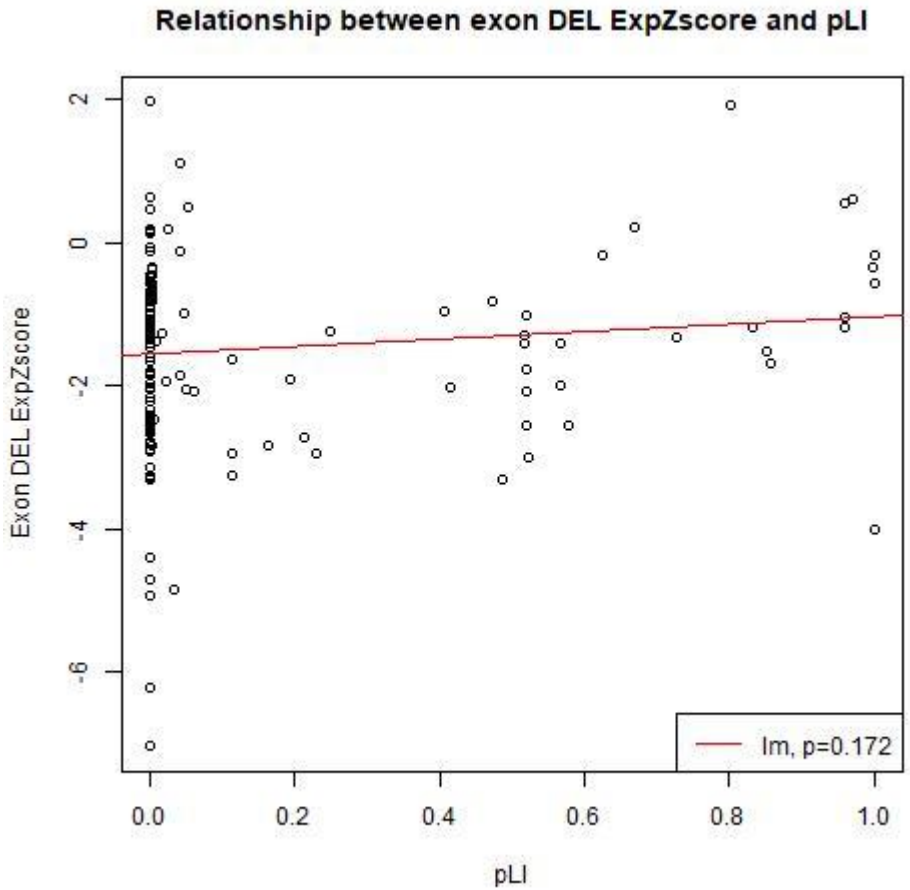


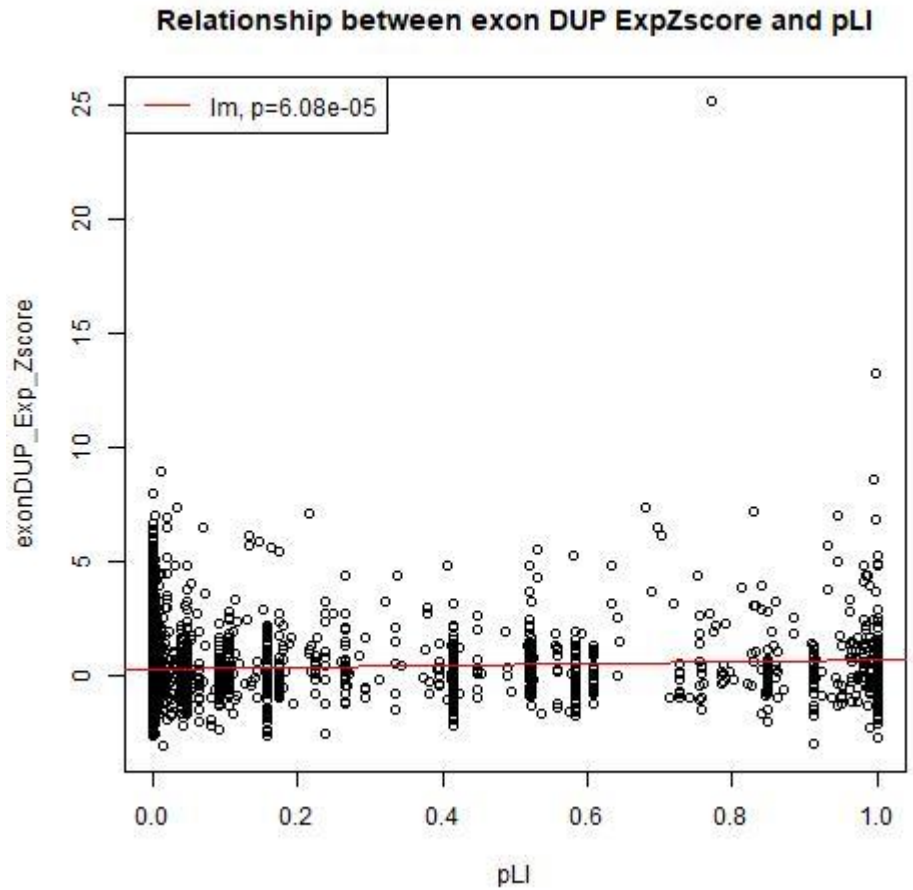
Relationship between Mean CaG exp Zscore vs pLI



Note: exonDUP Gene ENSG00000163945 with Zscore -0.693 has two pLI values (<2.44e-12) (i.e. correponds to two different ENST IDs

Relationship between individual CaG CNV gene ExpZscores and pLi





Conclusions:

- There appears to be no clear corrlation between TAZ and G-factor
- The frequency and importance (CNV score / pLI) of DEL CNVs appear to be correlated with their effect on the expression of genes contained within them
- The expression of genes contained in important DEL and DUP CNVs appear to be up-regulated and down-regulated, respectively
- This suggests that a form of compensation occurs to prevent significant impacts of the CNV gene-expression
- **More data is required to reach strong conlusions**

Aknowledgements:

Sébastien Jacquemont

Guillaume Huguet

Catherine Schramm

Boris Chaumette

Jean-Louis Martineau

Maude Auger

Thank You