Investigation of the effect of copy number variants on the expression of genes and the IQ

“Séminaire interdisciplinaire de bio-informatique"

Course: BIF7002

UQAM - Winter 2022

Report on the seminar given by Dr. Jocelyn Bédard on January 25th 2022

Written by Lei Cao

Table of Contents

Introduction 3
Goal of research 4
Method 4
CARTaGENE cohort 4
Data collection 4
G-factor Generating using Principal Component Analysis (PCA) 4
RNA-seq analysis 5
Results 6
Gene expression: RNA-seq analysis 6
Total counts to Total Reads ratio 6
DESeq2 Normalization 6
Combat correction 7
Blood cell types in samples 7
IQ: g-factor 8
g-factor and Total Absolute expression Z-score (TAZ) 8
CNVs and gene expression 8
Comparison of Mean ExpZ-scores for genes within Exon CNVs vs not in CNVs 8
Comparison of individual CaG ExpZ-scores for genes within Exon CNVs vs not in CNVs 9
Relationship between mean ExpZscores of genes and their frequency in CNVs 9
Relationship between Mean CaG exp Zscore and CNV score 9
Relationship between Mean CaG exp Zscore and pLI 10
Conclusion 10
Prospective 10
References 11

Introduction

The term “copy number variation” (CNV) refers to a genetic change on an intermediate scale, defined as segments longer than 1,000 base pairs but less than 5 megabases in length. CNVs comprise both duplicate sequences (duplications) and genetic material losses (deletions) (Figure 1). Along with inversions and translocations, CNVs are categorised as forms of genome structural variation because they change the structure of the genome (Redon et al., 2006).

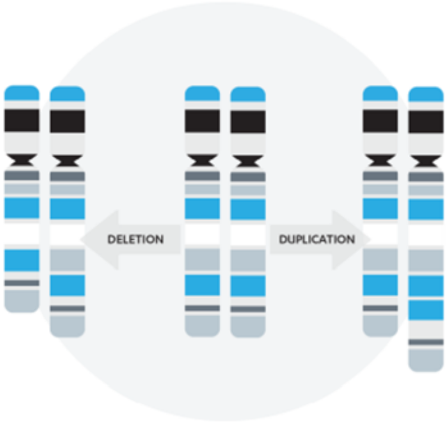


Figure 1

CNVs play a role in a variety of neurodevelopmental and psychiatric diseases. Pathogenic CNVs are detected in 10% to 15% of individuals referred for neurodevelopmental problems using whole genome chromosomal microarrays, which are now routinely used in medical diagnosis. Copy number variants may arise recurrently by nonhomologous recombination in unrelated individuals. Individually, a number of recurrent CNVs have been linked to intellectual impairments, autism spectrum disorders, and schizophrenia (Miller et al., 2010; Kearney et al., 2011).

Studies of CNVs on cognitive traits have been reported. For instance, a study in Iceland’s general population discovered that 26 psychiatric CNVs impair IQ by 15 points or 1 SD on average (Stefansson et al., 2014). Using the UK Biobank’s cognitive tests, 54 loci were shown to be related with worse results, ranging from 0.1 to 0.5 SD (Kendall et al., 2017). Furthermore, CNVs with deletion, based on their gene content and the total probability of loss of function intolerance (pLI), can have a general negative impact on the IQ (Huguet et al., 2018).

Although CNVs have been linked to complex features in humans in certain situations, the processes underlying those links have yet to be fully understood; consequently, a thorough examination of the effects on molecular phenotypes, including gene expression levels, is required. It is generally accepted that CNVs are likely to convey their effect, at least in part, by affecting gene expression. Previous research has shown that structural variants have a significant impact on expression, with big copy number variants (CNVs) alone accounting for 18% of variation in gene expression in cell lines (Stranger et al., 2007).

CNVs can affect gene expression through a variety of methods that go beyond simple gene dosage effects, such as gene regulatory region insertion and deletion, and changes in the physical proximity of genes and regulatory elements (Gamazon and Stranger, 2015).

RNA-sequencing enables precise genome-wide transcription measurement (Wang et al., 2009), allowing for a direct assessment of functional alterations caused by genetic variations.

Goal of research

Based on the previous publication (Huguet et al., 2018), the goal of Dr. Jocelyn Bédard's project was to assess how CNVs affect the expression of genes and to test whether this information could be integrated into Huguet et al's model to help predict the effect on IQ.

Method

CARTaGENE cohort

CARTaGENE (CaG) is a population-based biobank as well as Quebec's largest ongoing prospective health research of men and women. CaG focused on the 40 – 69-year-old age group, which is the most at risk of developing chronic diseases, throughout several Quebec urban areas (Montréal, Québec and Saguenay). The CaG study is one of the few population-based cohorts in the world that stores blood not only for DNA and protein-based science but also for gene expression analyses, allowing multiple systems genomics approaches to identify genetic and environmental factors associated with disease-related quantitative traits (Awadalla et al., 2013; Huguet et al., 2018; Fave et al., 2018).

Data collection

For this study, approximately 3000 participants have been genotyped and a further subset of ~1000 analysed by RNA-Seq. A large set have completed cognitive tests that can likely be used to assess their cognitive function (and predict their IQ) (Figure 2).

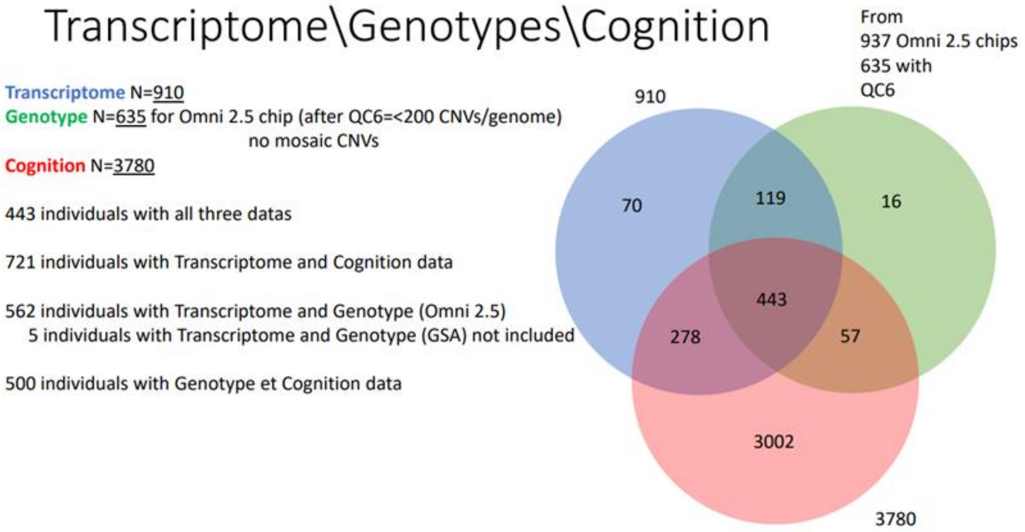

Transcriptome\Genotypes\Cognition
From
937 Omni 2.5 chips
635 with
QC6
Transcriptome N=910
Genotype N=635 for Omni 2.5 chip (after QC6=<200 CNVs/genome)
no mosaic CNVs
Cognition N=3780
443 individuals with all three datas
721 individuals with Transcriptome and Cognition data
562 individuals with Transcriptome and Genotype (Omni 2.5)
5 individuals with Transcriptome and Genotype (GSA) not included
500 individuals with Genotype et Cognition data
910
70
119
16
443
278
57
3002
3780

Figure 2

g-factor Generating using Principal Component Analysis (PCA)

Three cognitive tests were used to generate one variable that could predict IQ: Memory test, Reasoning test and Reaction time test. PC1 (Dim1) was used as g-factor. Value representative of IQ (cognitive function) (Figure 3).

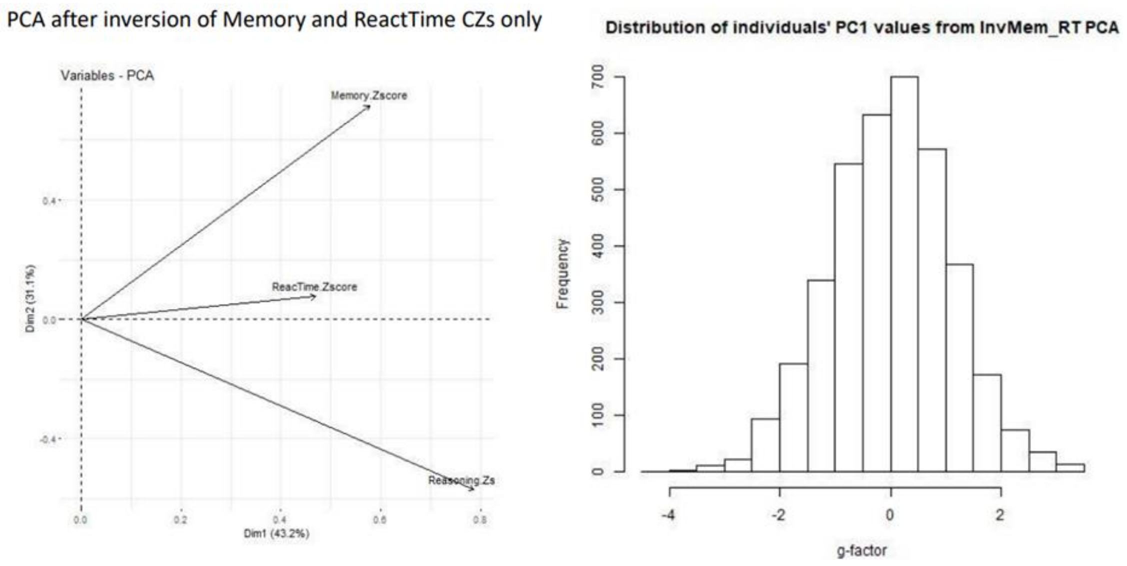
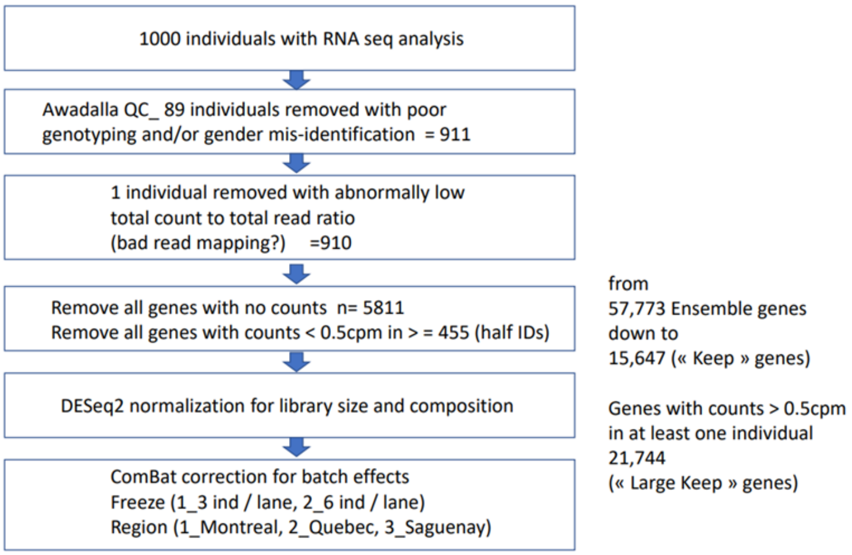


Figure 3

RNA-seq

Flow chart of RNA-seq data selection and processing is shown below (Figure 4).



1000 individuals with RNA seq analysis

Awadalla QC_ 89 individuals removed with poor genotyping and/or gender mis-identification = 911

1 individual removed with abnormally low total count to total read ratio (bad read mapping?) =910

Remove all genes with no counts n= 5811
Remove all genes with counts < 0.5cpm in > = 455 (half IDs)

DESeq2 normalization for library size and composition

ComBat correction for batch effects
Freeze (1_3 ind / lane, 2_6 ind / lane)
Region (1_Montreal, 2_Quebec, 3_Saguenay)

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)

Figure 4

Correction model:

```
counts_lm= lm(tCaG_ComBat_CaG_Blood_RNASeq_noNA_IDs[,i]~

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

Library normalisation and covariate (both biological and technical) adjustments were performed using linear regression modelling to account for differences between samples, experimental batch effects, and unwanted RNA-seq-specific technical variations. After correction, count values was converted to Z-scores.

Results

<u>Gene expression: RNA-seq analysis</u>

Total counts to Total Reads ratio

Figure 5 shows the comparison 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. Low count/read outlier sample #11113726 was removed from matrix, resulting a new CaG_matrix (n=910 individuals)

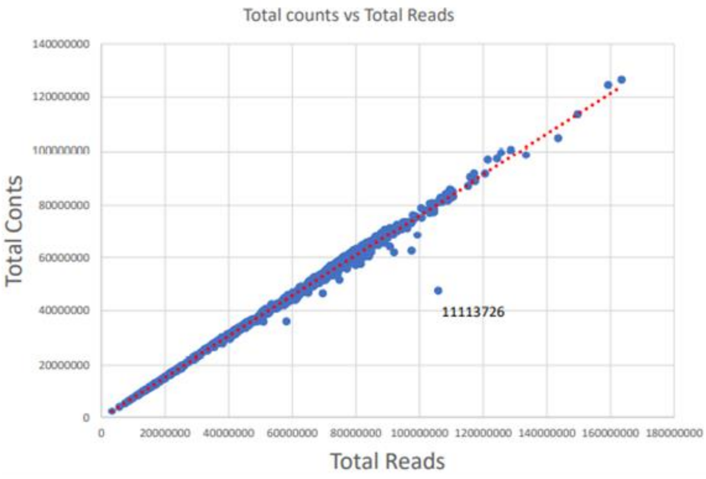


Figure 5

DESeq2 Normalization

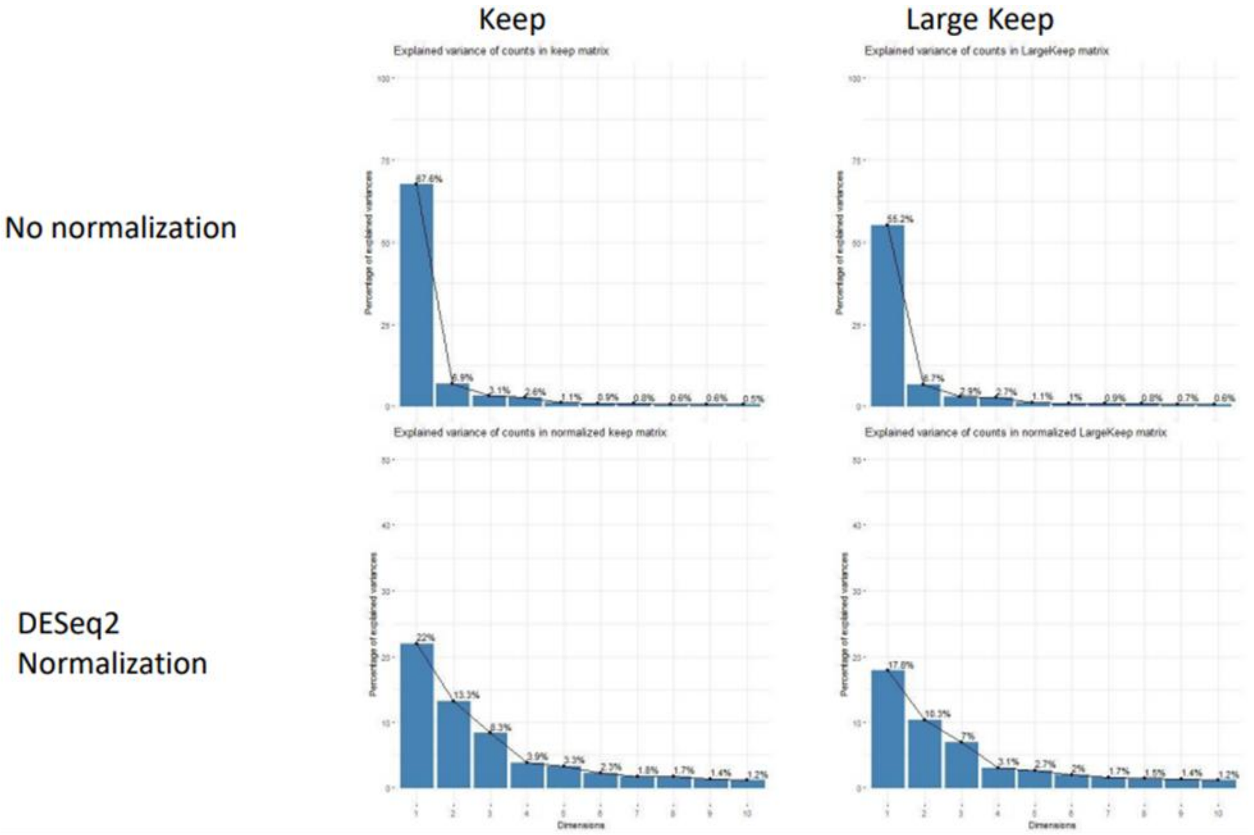


Figure 6

Figure 6 indicates that a large part of variation accounted for by PC1 is associated to library size.

Combat correction

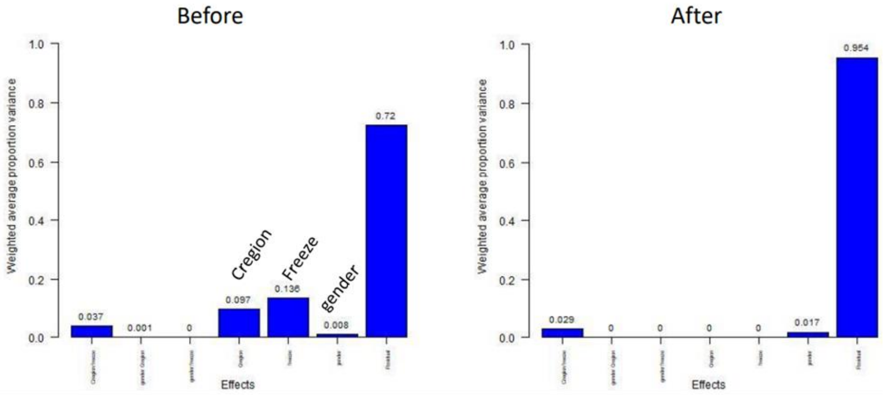


Figure 7

Factors Freeze and region had large effect on the Keep expression matrix before ComBat correction and these effects were significantly reduced after. ComBat library is used for batch effects (Figure 7).

Blood cell types in samples

Two methods were used for assessing blood cell types in this study. CIBERSORT is an analytical tool to provide an estimation of the abundances of member cell types in a mixed cell population, using gene expression data. Quantification was carried out with Cibersort with expression counts of 524/547 genes in LM22 signature matrix. Hematometry was also carried out in most of the individuals in CaG cohort.

Figure 8a shows one example of the comparison of blood cell type proportions from Cibersort and CaG. There was a significant correlation in lymphocyte measured with abovementioned two methods. Albeit this linear relationship, proportions of different cell types predicted from both datasets were significantly different (Figure 8b).

Blood cell type proportion from CaG data was further used for the linear regression for its accuracy.

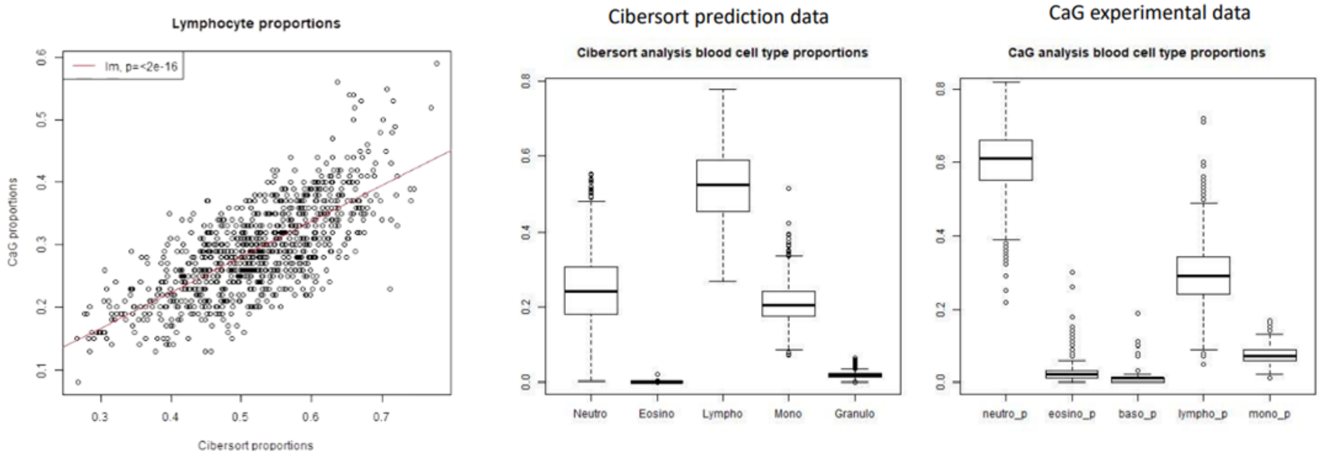


Figure 8a

Figure 8b

IQ: g-factor

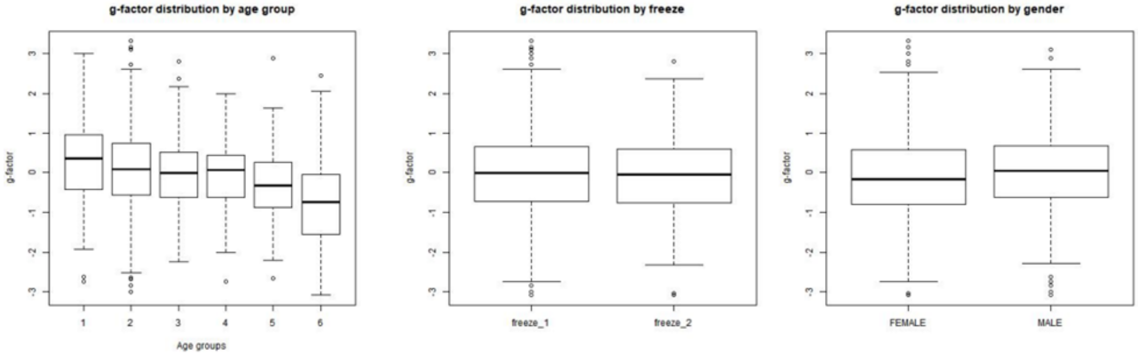


Figure 9

The g-factor variations cross-different variables (age, freeze and gender) are shown in Figure 9. With age between individuals from 40 to 70 year-old, cognitive function seems to decrease in older age group. There is no significant difference in term of freeze and gender.

g-factor and Total Absolute expression Z-score (TAZ)

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.

No significant relationship between g-factor and absolute expression Z-score by linear regression analysis (p=0.703) (Figure 10).

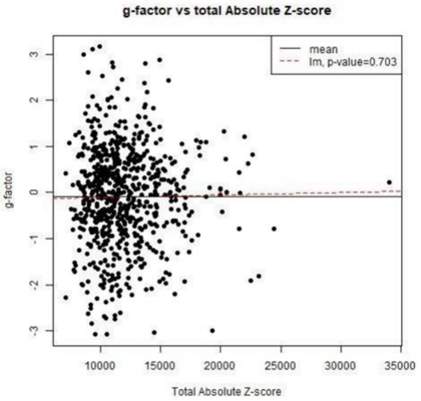
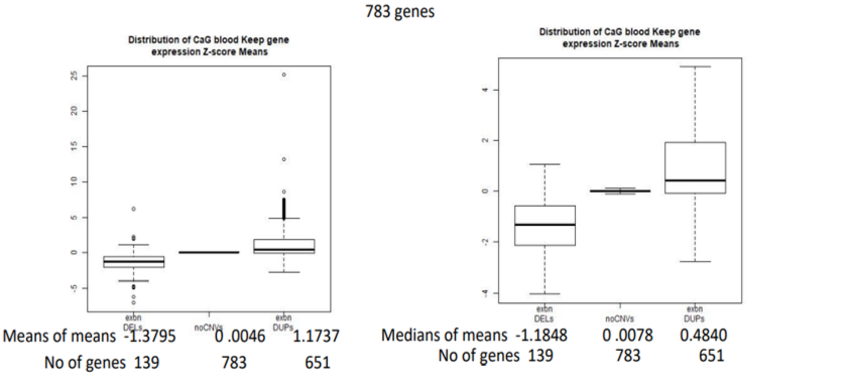


Figure 10

CNVs and gene expression

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

For each gene, the mean of expression value was calculated. Figure 11 shows significant down regulation of gene expression in DEL CNVs in exon comparing to gene expression without CNVs, and significant up-regulation of gene expression in DUP CNVs with or without taking into account of outliers.



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

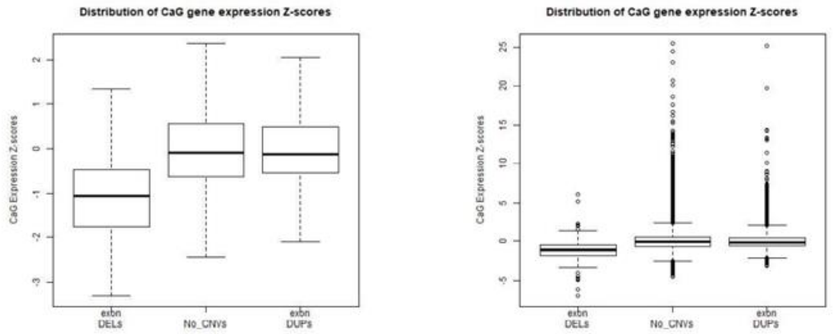
Medians of means -1.1848 0 .0078 0.4840
No of genes 139 783 651

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

Figure 11

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

Similar results were found in each individual gene.



Mean Z-score -1.0866 0.00433 0.41280

Analysis of CaG expression Zscore means:

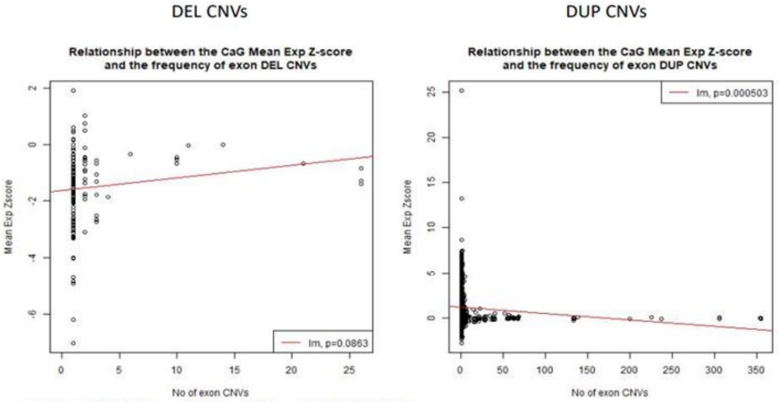
Kruskal-Wallis (non-parametric test for difference between means) p-value <2.2e-16

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

Figure 12

Relationship between mean ExpZscores of genes and their frequency in CNVs

In DEL CNVs, higher frequency of CNVs was marginally associated with higher gene expression; in contrast, in DUP CNVs, higher frequency of CNVs was significantly associated with lower gene expression (Figure 13).



Mean ExpZcores of 135 genes

Mean ExpZscores of 651 genes

IDs with CNVs, with DEL CNVs, with DUP CNVs

frequency of CNVs

Figure 13

Relationship between Mean CaG exp Zscore and CNV score

The importance of CNVs is presented by CNVs score which are derived from pLI and gene content. Figure 14 shows that more important DEL CNVs had a higher gene expression. However, no significant relation was found in DUP CNVs and gene expression.

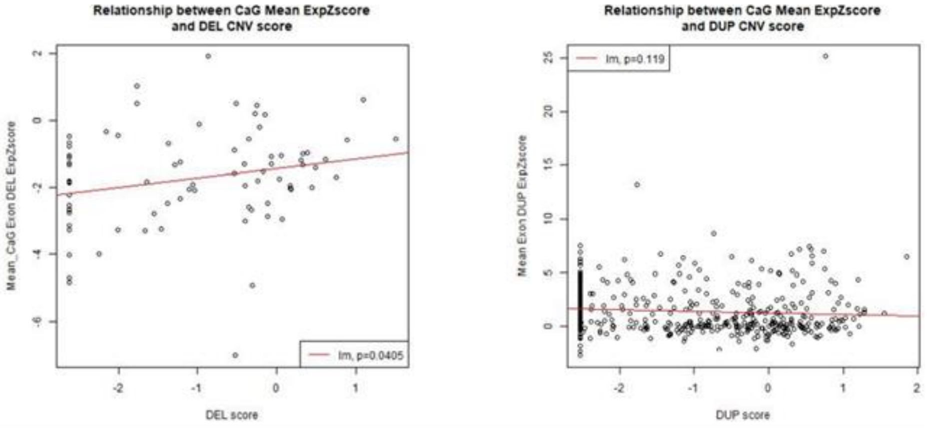


Figure 14

Relationship between Mean CaG exp Zscore and pLI

The higher pLI of individual was associated with a higher gene expression in DEL CNVs. However, no significant relation was found in DUP CNVs and pLI (Figure 15).

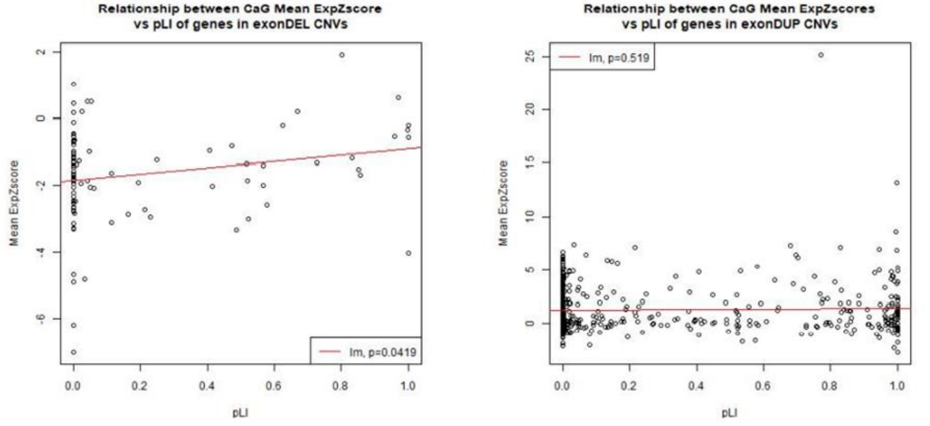


Figure 15

Conclusion

This study has investigated the how CNVs (deletion and duplication) affected gene expression measured by RNA-sequencing, and the association between gene expression and IQ in CaG cohort. There appears to be no clear corrlation between gene expression and IQ. 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 CNVs gene-expression. Further analyses with more data is required to reach strong conclusions.

Prospective

Despite the importance of CNVs in brain illnesses and the tissue-specific nature of transcriptional control, the difficulty in obtaining brain samples has hampered efforts to understand the functional consequences of CNVs in the brain. Hence, it may be important in future studies to perform detailed analyses of gene expression profiling in peripheral blood lymphocytes, which could be further used to predict the gene expression profile in brain regions.

References

Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, Hamet P, Laberge C; CARTaGENE Project. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. 2013 Oct;42(5):1285-99. doi: 10.1093/ije/dys160. Epub 2012 Oct 15.

Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genomics. 2015 Sep;14(5):352-7. doi: 10.1093/bfgp/elv017. Epub 2015 Apr 27.

Favé MJ, Lamaze FC, Soave D, Hodgkinson A, Gauvin H, Bruat V, Grenier JC, Gbeha E, Skead K, Smargiassi A, Johnson M, Idaghdour Y, Awadalla P. Gene-by-environment interactions in urban populations modulate risk phenotypes. Nat Commun. 2018 Mar 6;9(1):827. doi: 10.1038/s41467-018-03202-2.

Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019 May;51(5):772-776. doi: 10.1038/s41588-019-0383-1. Epub 2019 Apr 8.

Huguet G, Schramm C, Douard E, Jiang L, Labbe A, Tihy F, Mathonnet G, Nizard S, Lemyre E, Mathieu A, Poline JB, Loth E, Toro R, Schumann G, Conrod P, Pausova Z, Greenwood C, Paus T, Bourgeron T, Jacquemont S; IMAGEN Consortium. Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. JAMA Psychiatry. 2018 May 1;75(5):447-457. doi: 10.1001/jamapsychiatry.2018.0039.

Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST; Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 2011 Jul;13(7):680-5. doi: 10.1097/GIM.0b013e3182217a3a.

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010 May 14;86(5):749-64. doi: 10.1016/j.ajhg.2010.04.006.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature. 2006 Nov 23;444(7118):444-54. doi: 10.1038/nature05329.

Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, Bjornsdottir G, Walters GB, Jonsdottir GA, Doyle OM, Tost H, Grimm O, Kristjansdottir S, Snorrason H, Davidsdottir SR, Gudmundsson LJ, Jonsson GF, Stefansdottir B, Helgadottir I, Haraldsson M, Jonsdottir B, Thygesen JH, Schwarz AJ, Didriksen M, Stensbøl TB, Brammer M, Kapur S, Halldorsson JG, Hreidarsson S, Saemundsen E, Sigurdsson E, Stefansson K. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014 Jan 16;505(7483):361-6. doi: 10.1038/nature12818. Epub 2013 Dec 18.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007 Feb 9;315(5813):848-53. doi: 10.1126/science.1136678.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484.