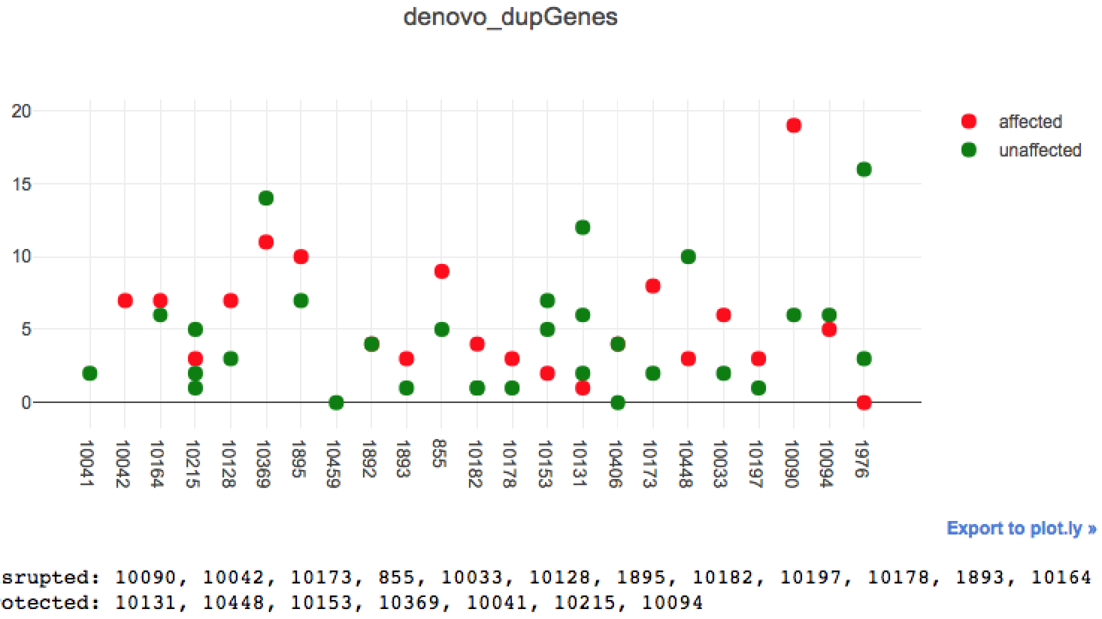
Simplex experiments

January 11, 2018

Main results

Most of the main analysis I did can be summarized in plots like Figure 1:

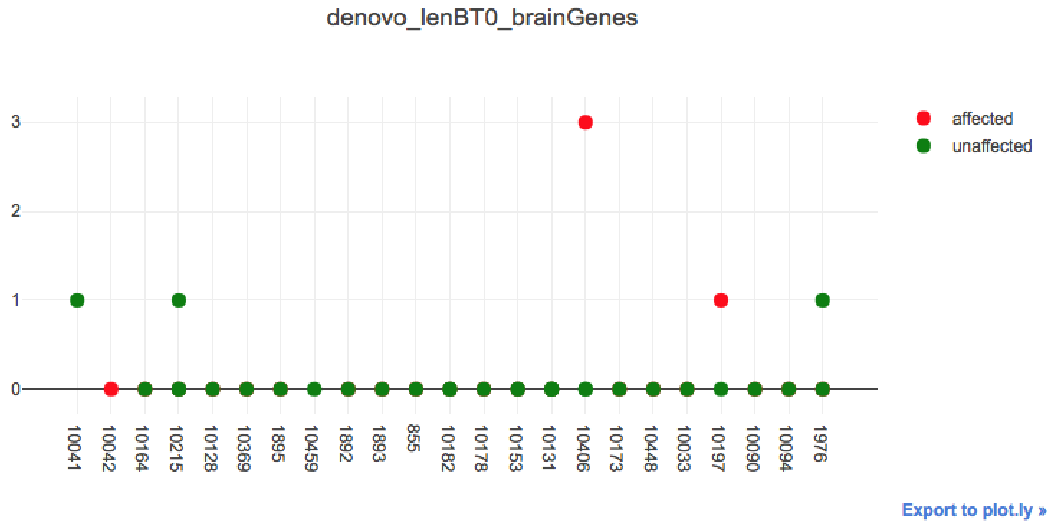


```
Best disrupted: 10090, 10042, 10173, 855, 10033, 10128, 1895, 10182, 10197, 10178, 1893, 10164
Best protected: 10131, 10448, 10153, 10369, 10041, 10215, 10094
```

Figure 1: de novo duplications in gene regions

You'll see in the X axis each family, and each circle is a child. Some families have more than one child, and there is always only one red circle (affected child) per family. Sometimes there is less than 2 circles per family, in which cases the kid was dropped during QC (or they're just overlayed in the picture). The Y axis counts the number of CNVs in each child, under different conditions (different figures). You'll also see some text under the figure, showing the families with biggest difference (in descending order) between affected and unaffected (red dot is higher, disrupted family), or between unaffected and affected (green dot higher, protected family). For example, in Fig 1 family 10090 has the biggest distance between red and green, and family 10131 has the biggest distance between green and red. The ideal result (i.e. lots of disrupting CNVs) would be lots of families with the red dot higher than green dots, and then when we looked at the (gene) location of those CNVs there would be some sort of consistency among families.

So, Figure 1 shows de novo burden by duplication CNVs, only in regions that code for genes. After going over hundreds of those plots (see things I tried below), these were the more interesting ones:



```
Best disrupted: 10406, 10197
Best protected: 10041, 10215
```

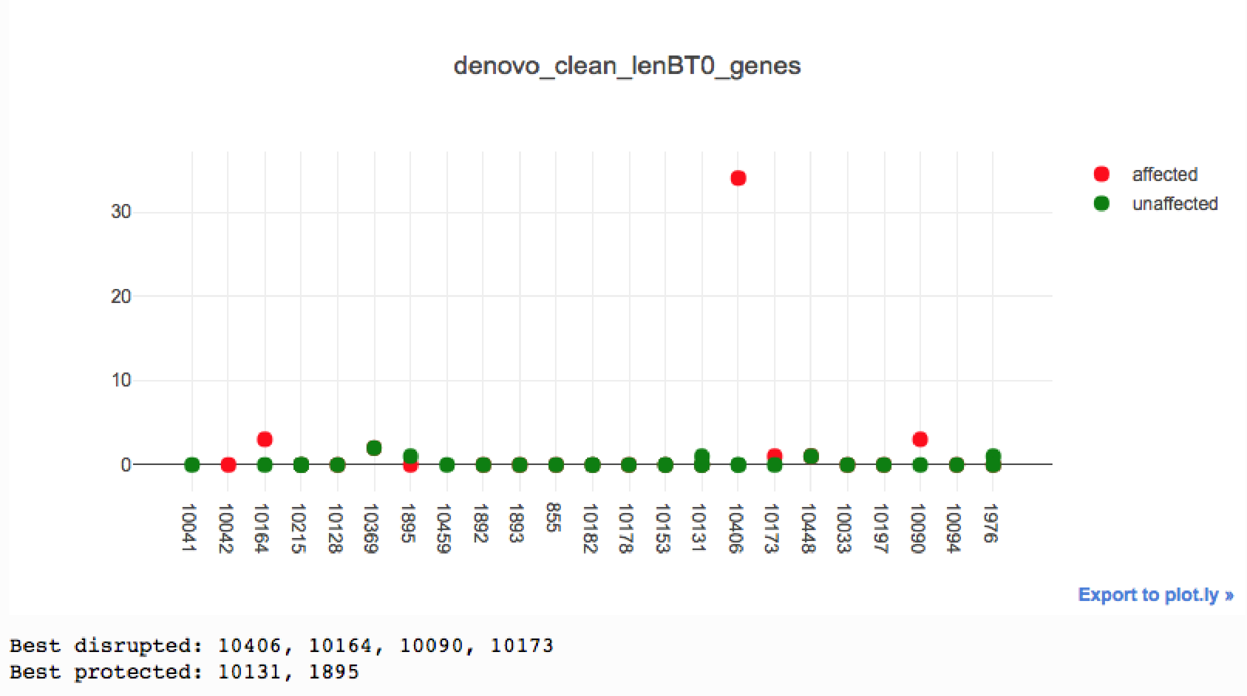
Figure 2: all denovo CNVs expressed in the brain

Figure 2 shows the de novo CNVs in genes that are expressed in the brain (according to the Allen Brain Atlas). It makes no distinction between deletions and duplications. Something else that made a difference in the results was removing "well-know" CNVs. I used the same dataset the ADHD CNV papers used to discard well-know CNVs, under the idea that we're looking for rare variants. This cleaning process also excluded regions that (according to PennCNV) can be hard to infer signal, such as telomeres, centromeres, and others. Doing such cleaning wiped off the brain results, and significantly reduced the duplication CNV numbers in gene regions (compare Figures 5 and 1).

That affected kid in family 10406 actually had lots of denovo deletions (Figure 4), but it passed all QC metrics, so that's why I kept it. But if we zoom in the <10 region of Figures 3 or 4, we can see other families who also have disruptive (and protective) CNVs. So, the next question is whether the same genes are affected by those CNVs... nope. In other words, for example, even though families 10406, 10164, 10090, and 10173 all have kids with disrupting CNVs (Figure 3), none of them code for the same gene. The 3 latter families have no intersections in gene lists, and although there are some intersections with 10406 (SP3, CREBRF; then NCAPG, LCORL, and ETNK1), I don't think it would be incredibly hard, as there are over 30 to choose from.

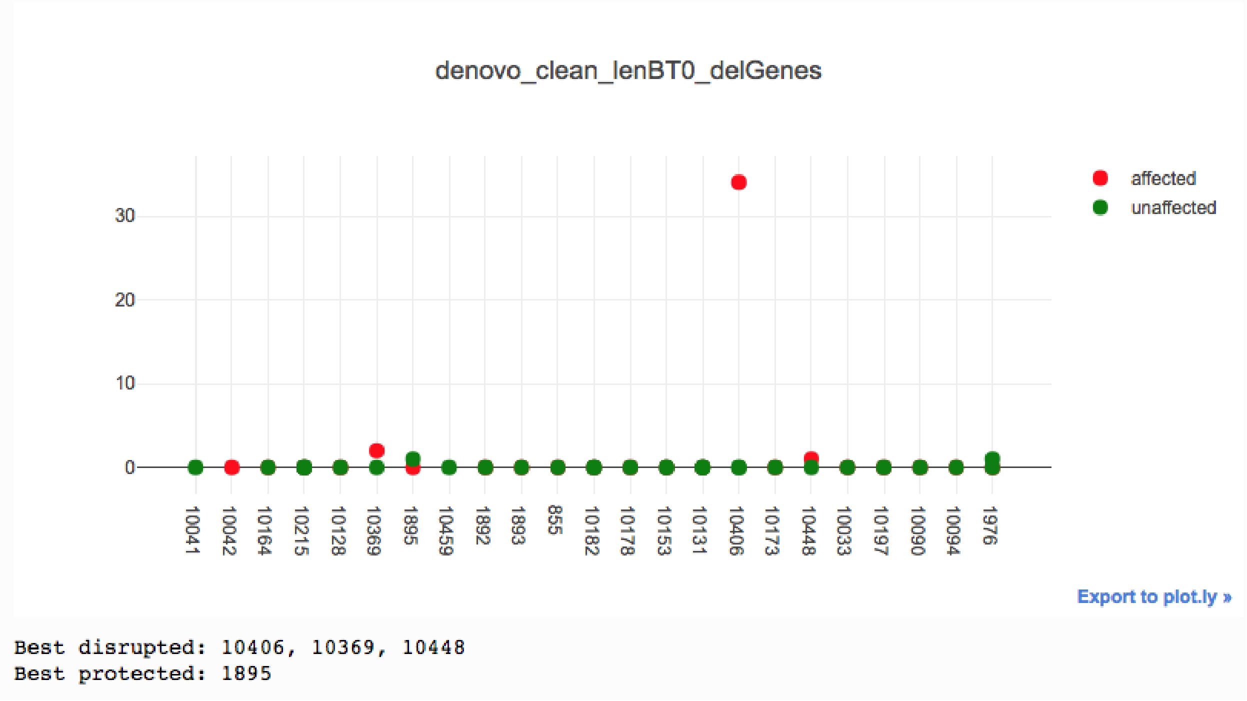
Another analysis I did was gene-based. That means that instead of focusing on how many CNVs were present in each child, I first counted how many kids had gene G with a given CNV. For example, using the clean set of CNVs, we get a table like this (6):

I still need to do some stats to see the chances of 4 kids having CNVs in that gene, but they were not all affected kids, so it's not as exciting. It becomes a bit more interesting if we think of combinations of genes. For example, say you're in trouble if you have two (or more) disrupted of the genes in this list. Figure 7 dives into the first 7 rows in figure 6:



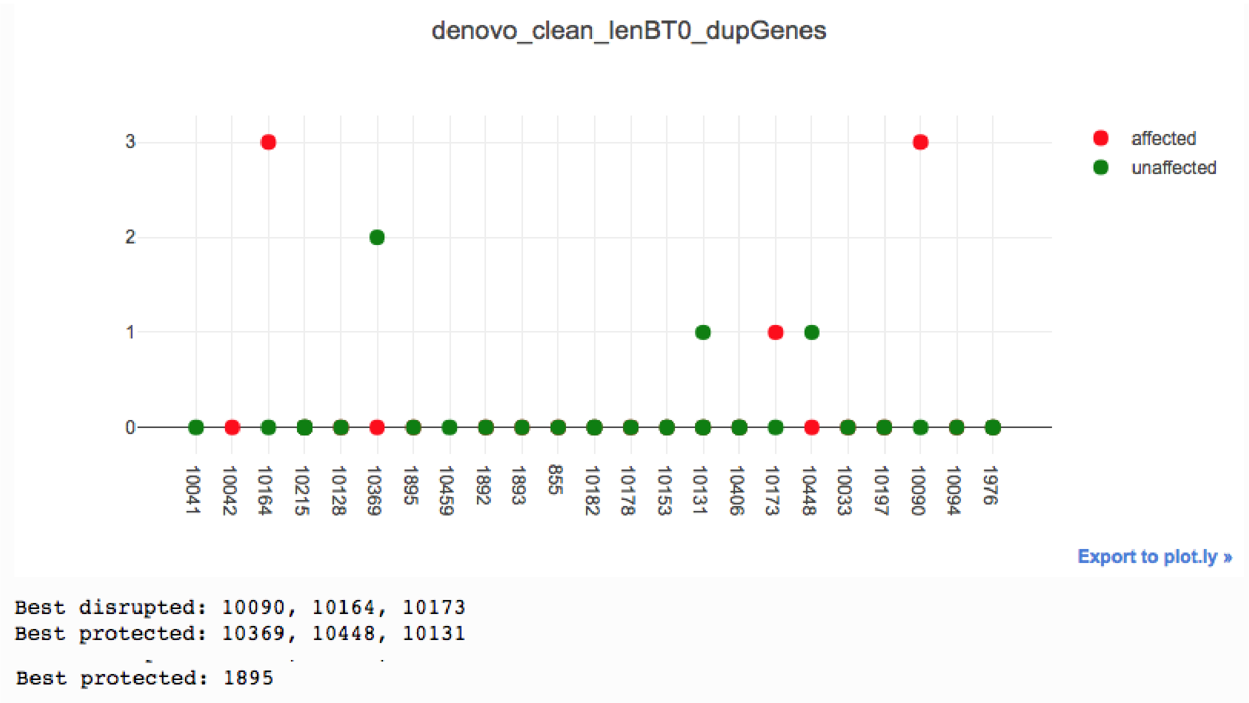
```
Best disrupted: 10406, 10164, 10090, 10173
Best protected: 10131, 1895
```

Figure 3: CLEAN denovo CNVs in gene coding regions



```
Best disrupted: 10406, 10369, 10448
Best protected: 1895
```

Figure 4: CLEAN deletion CNVs in gene coding regions



```
Best disrupted: 10090, 10164, 10173
Best protected: 10369, 10448, 10131

Best protected: 1895
```

Figure 5: CLEAN de novo duplications in gene regions

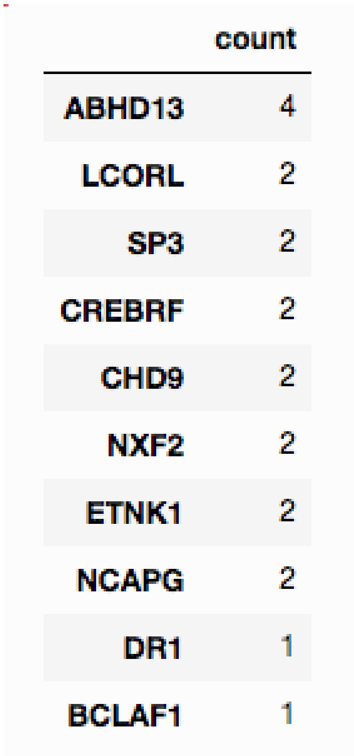
In this case, 400122 has 2 affected genes, 400123 has 3, 400178 has 6. What are the chances of that happening by chance? Need to check... but it would also be cool to check a metric like that against other phenotypes, like symptom counts.

As the results above were somewhat intriguing, but not necessarily all pointing to the same direction, the last step I took was to try to find correlations between CNV burden and different phenotypic variables. Brain variables weren't as good as I'd like, mostly because there are only 25 kids (out of the possible 51 in the simplex study) with imaging, and that's even before taking QC into consideration. Wendy confirmed that.

Still, in Figure 8 I show an exploratory search of correlations between different gene burden pipelines (Y) and phenotypes (X). By pipelines I mean different combinations of the things I tried (see below). I'm not too worried about multiple comparison corrections yet, as most of the pipelines in the Y axis can be removed. But I do need to calculate appropriate p-values for each cell in the heatmaps as the N for each phenotype varies. All in all, there seems to be some interesting phenotypes correlated with CNV burden. I think we're in a good place that, if we add a few more phenotypes and burden pipelines, some cooler stuff will come up.

Tried (but no better results)

- Different exome analysis CNV tools: XHMM, ExomeCopy, cn.mops, Conifer
- PennCNV (SNP array CNV)



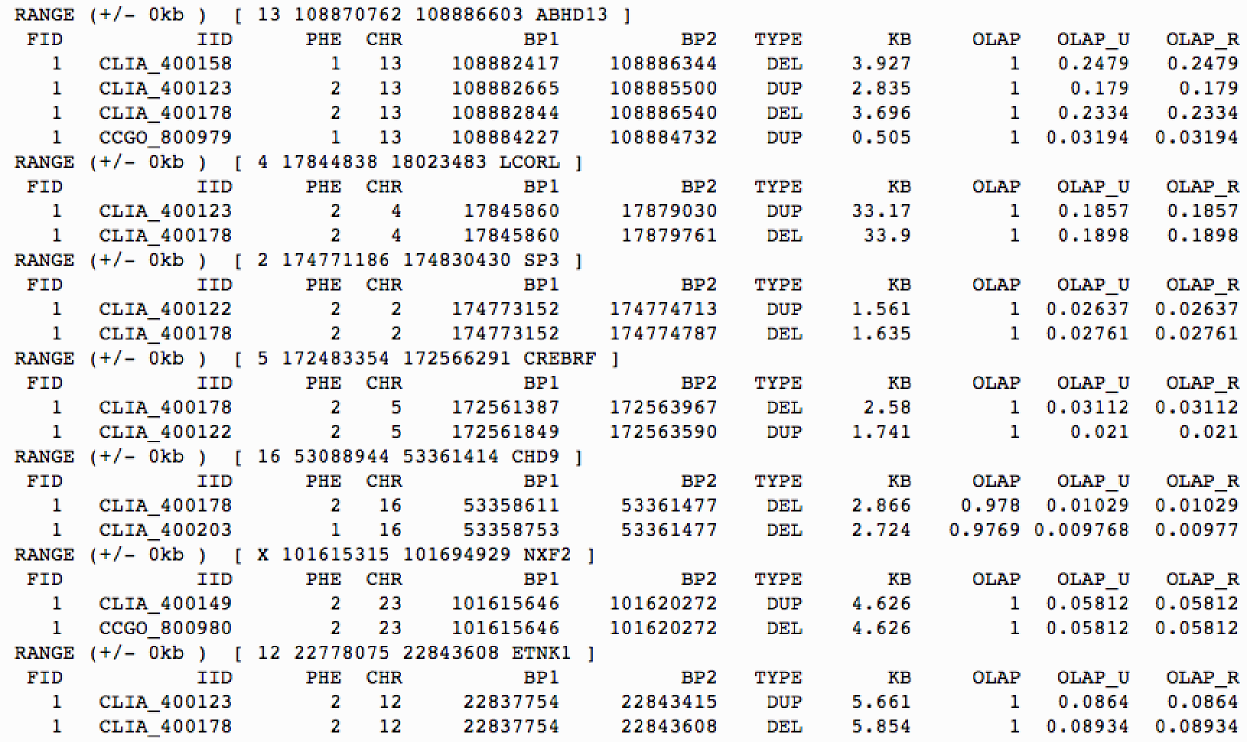
	count
ABHD13	4
LCORL	2
SP3	2
CREBRF	2
CHD9	2
NXF2	2
ETNK1	2
NCAPG	2
DR1	1
BCLAF1	1

Figure 6: Number of kids with CNVs in that Gene

- Filtering CNVs based on length, CNV type (dup/del), origin (denovo, inherited, all), number of markers, gene-coding regions, brain-expressed regions, literature-based search, removing well-known variants

Still to try

- Dive deeper into gene-based analysis (e.g. permutation analysis for combination of genes)
- Contrast with multiplex samples (waiting on Tri + Sijung new pipeline for final results)
- Check if worth doing anything with imaging for the 25 kids.
- Add more phenotypes for comparison, and more pipelines in the Y axis
- Play with tool-specific tuning parameters (e.g. HMM priors, variant threshold sensitivities)



```
RANGE (+/- 0kb )  [ 13 108870762 108886603 ABHD13 ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400158        1   13      108882417     108886344     DEL     3.927         1    0.2479    0.2479
   1    CLIA_400123        2   13      108882665     108885500     DUP     2.835         1     0.179     0.179
   1    CLIA_400178        2   13      108882844     108886540     DEL     3.696         1    0.2334    0.2334
   1    CCGO_800979        1   13      108884227     108884732     DUP     0.505         1   0.03194   0.03194
RANGE (+/- 0kb )  [ 4 17844838 18023483 LCORL ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400123        2    4       17845860      17879030     DUP     33.17         1    0.1857    0.1857
   1    CLIA_400178        2    4       17845860      17879761     DEL      33.9         1    0.1898    0.1898
RANGE (+/- 0kb )  [ 2 174771186 174830430 SP3 ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400122        2    2      174773152     174774713     DUP     1.561         1   0.02637   0.02637
   1    CLIA_400178        2    2      174773152     174774787     DEL     1.635         1   0.02761   0.02761
RANGE (+/- 0kb )  [ 5 172483354 172566291 CREBRF ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400178        2    5      172561387     172563967     DEL      2.58         1   0.03112   0.03112
   1    CLIA_400122        2    5      172561849     172563590     DUP     1.741         1     0.021     0.021
RANGE (+/- 0kb )  [ 16 53088944 53361414 CHD9 ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400178        2   16       53358611      53361477     DEL     2.866     0.978   0.01029   0.01029
   1    CLIA_400203        1   16       53358753      53361477     DEL     2.724    0.9769  0.009768   0.00977
RANGE (+/- 0kb )  [ X 101615315 101694929 NXF2 ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400149        2   23      101615646     101620272     DUP     4.626         1   0.05812   0.05812
   1    CCGO_800980        2   23      101615646     101620272     DEL     4.626         1   0.05812   0.05812
RANGE (+/- 0kb )  [ 12 22778075 22843608 ETNK1 ]
 FID           IID       PHE  CHR            BP1           BP2    TYPE        KB      OLAP    OLAP_U    OLAP_R
   1    CLIA_400123        2   12       22837754      22843415     DUP     5.661         1    0.0864    0.0864
   1    CLIA_400178        2   12       22837754      22843608     DEL     5.854         1   0.08934   0.08934
```

Figure 7: Samples with CNVs in each gene. PHE=2 is ADHD

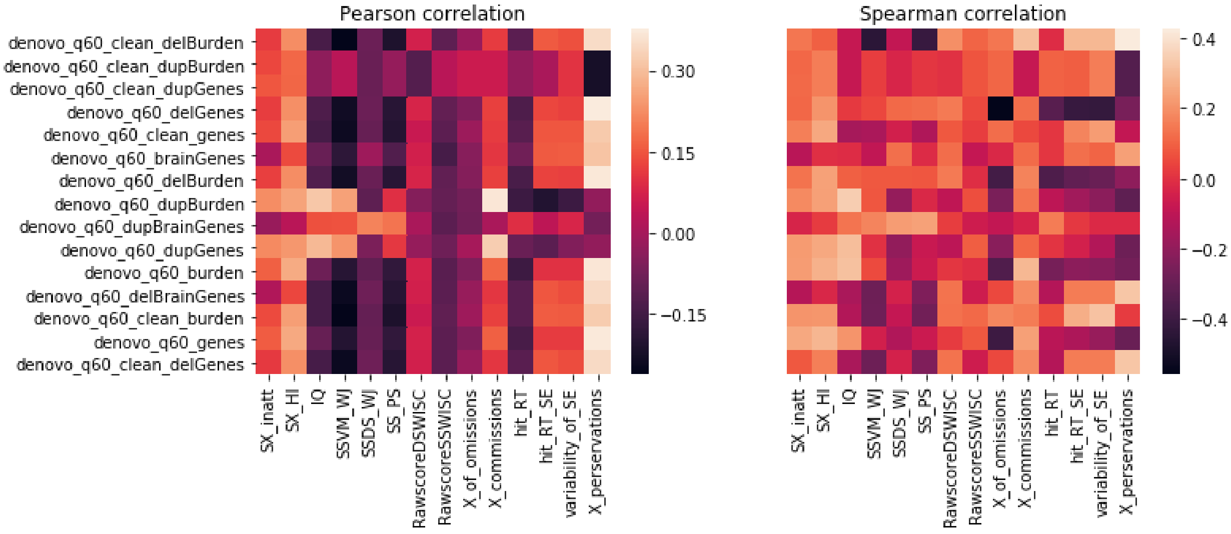


Figure 8: CNV to phenotype correlation