Outline

Claim 1: Genomic variation in CLL is reflected on gene expression:
IGHV status and Trisomy12 are main drivers of gene expression variability. In line with this, expression of the 150 most variable genes is regulated by low-, intermediate and high programmed methylation groups. Other known variants show distinct expression profiles as TP53, SF3B1, Del11q22.3, Notch1 and Del13q14. In total 7 of 13 variants showed distinct expression signatures. Expression of Marker genes ZAP70 and CD38 is up-regulated in U-CLL.

Questions to solve:
1.How do we define distinct? All variants result in significantly differentially expressed genes using Deseq2. They vary in the number of significant genes (see figure **??**) and how good they cluster on the expression of their differentially expressed genes. What is the best way to formalize this? Can we use the associations (t-test) of the variants and principal components of their significant genes (see figure 1) 2.Should we include an extra figure for the methylation groups? I already included the t-sne in figure 2, but I think it is a remarkable finding. Expression of the most variable genes are clearly associated to them and we can explain it by the expression of some multi-regulatory transcription factors (that have been proposed by Oakes et al.)

Claim 2: IGHV status and Trisomy12 interact (in an epistatic way) determining gene expression:
Sample with hypermutated IGHV and trisomy12 unravel gene cluster of different ways of epistatic interaction, so called mixed epistasis. We find sets of genes, that are affected by buffering, suppression, masking and inversion. These cluster represent..? Drug sensitivity data (and prognosis?) support this evidence of epistatic interaction.

Questions to solve:
4. Which of the other data sets can we use to support our model of epistasis? To what degree do we find epistasis in Mek/Erk inhibitors? Do we find evidence of epistasis in prognosis/metabolic data?

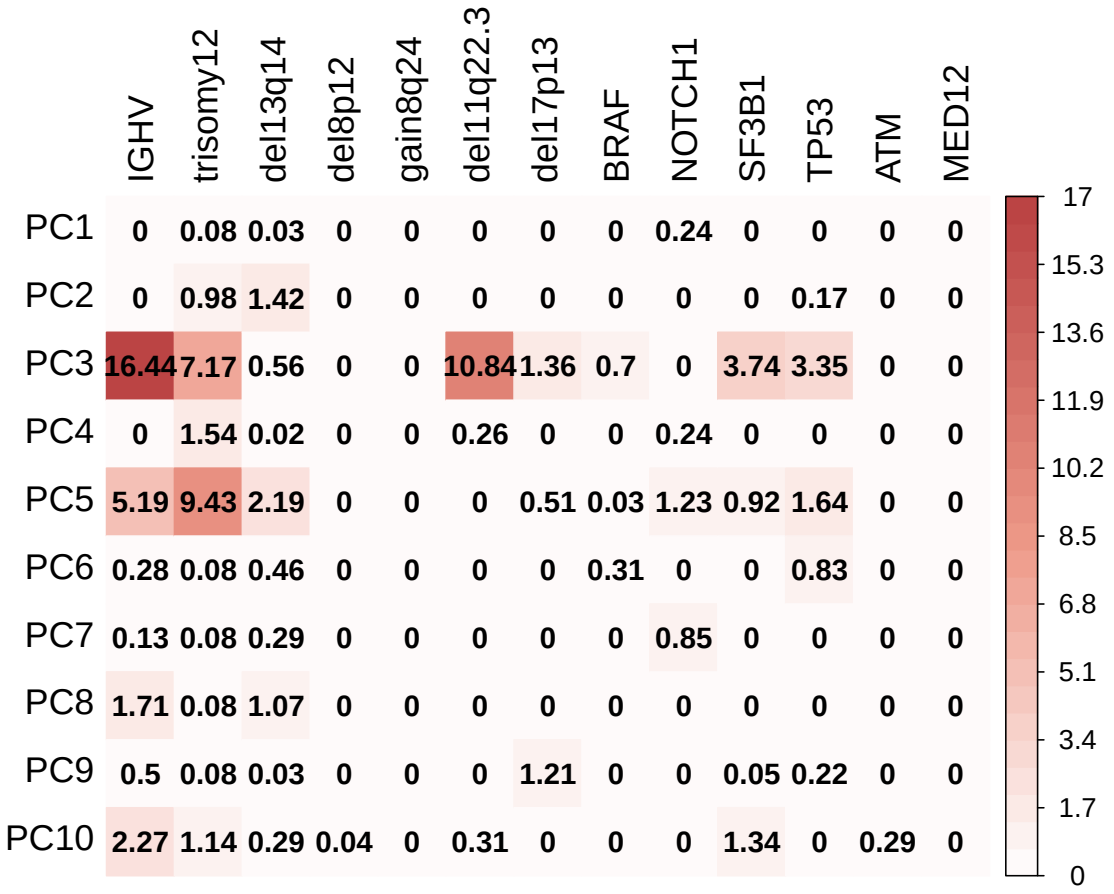

Figure 1: **Associations of variants and principal components in differentially expressed genes:** -$\log_{10}$(p.adjust) of t-test for associations between variants and the principal components within gene expression of significant ($p_{adj} < 0.01$, $\log_2$fold change > 2) results from differential expression analysis. In total 7 of 13 variants show significant associations to at least one of the first principal components describing the variance within gene expression based on results from differential expression analysis. *fixme: This analysis and figure has some merit, but it's too abstract / too complex for a main figure. I think it'd better to just state the number of DE genes (at a fixed FDR, say 10%) for each mutation. Also, I have an idea for a PCA plot which we can discuss on the phone. Basically: choose top 2000 (or so) genes (or the union of the above?), for each mutation compute the difference of means between groups of samples with and without, which results in 13 vectors of length 2000, then display these 13 points in a 2D PCA.*