Statistical testing for GBM (Giulio)

We developed a statistical test to compute p-values for the null hypothesis that clonal mutations are missing from the margin sample (M). The test is divided in two steps: a *training*, in which we identify some mutations to train a statistical models of expected read counts from WES, and a *test,* in which we use the trained model to compute the likelihood for the null hypothesis, given the targeted deep sequencing panels.

The test is carried out independently on each one of the input samples.

Training a read-counts model

We first computed CCF values and CNA from WES of the primary tumour samples, as discussed in **GEORGE'S-PART**.

From these data, we isolated all SNVs that are putative *clonal,* as estimated by requiring their CCF > 0.8 in *all* primary samples. Then, we selected SNVs with the *same CNA status* across all primary samples, and discarded the others. We stored their read counts from Platypus VCF files reporting Number of Variants (NV) and Reads (NR, i.e., the coverage), as well as the CNA status for the mutation.

We corrected read counts for CNA status and tumour purity, before using them to train a *Beta-Binomial* model of read counts. Correction is carried out to estimate how many of the *observed reads r* come from the actual tumour. The correction uses tumour purity π (here estimated from WES data) and tumour copy number status c as follows

$$\hat{r} = \frac{c * \pi}{c * \pi + 2 * (1 - \pi)} * r$$

Factor 2 in the correction is the diploid copy number status of normal cells. After correction, $\hat{r}$ is then the number of reads, out of r, that are estimated coming from tumour cells.

Giulio 31/12/2017 12:52

Comment [1]: Shall we say that this correction assumes that 1 allele is mutated out of c copies (i.e., SNVs happen before amplifications?)? If I well remember this should be more conservative no?

With the corrected coverage value, we can fit a Beta-Binomial with number of trials $\hat{r}$, and number of success NV (i.e., the number of mutated reads out of $\hat{r}$). This statistical model describes the probability of observing v mutated reads at coverage $\hat{r}$ as a function of the Beta distribution $\mathrm{B}(\alpha, \beta)$

$$\mathrm{BetaBin}(v|\hat{r}; \alpha, \beta) = \frac{\mathrm{B}(v + \alpha, \hat{r} - v + \beta)\binom{\hat{r}}{v}}{\mathrm{B}(\alpha, \beta)}$$

This compound model extends binomial sampling with over-dispersion effects; technically, it is a Binomial distribution whose parameter follows a Beta with hyper-parameters α and β.

A Beta-Binomial model was trained for each combination of copy number status for the input SNVs, and each WES sample of the primary tumour. Groups of SNVs with less than 10 mutations are removed from downstream analysis.

To learn the model parameters we used the Maximum Likelihood fitting procedure vglm for *vector generalized linear models* that is implemented in the R package VGAM.

Testing

We first identified testable SNVs from the two deep-sequencing targeted panels. These must have 0 mutated reads (NV = 0) in the margin sample (M), as annotated in the Platypus VCF files computed for the panels, and must be flagged clonal in the WES samples, as discussed in the training. If a patient has more than one margin sample (e.g., Patient A23 has margin samples M1 and M2), we required the mutated reads counts to be 0 across *all* margin samples. SNVs selected in this way appear indeed clonal in the primary tumour, but are missing in the margin biopsy from the targeted panels.

Some patients have no testable mutations (e.g. Patient 42); in this case their testing is aborted. The SNVs that we detected from both targeted panels are pooled together, and their number of reads (NR) from the VCF files is stored; if an SNV is detected from both panels, we sum the NR values from both panels.

Read counts from the margins need to be corrected as done for the training set. Copy number status for mutations in the margin is the same for the training set, but purity correction requires some considerations. Estimation of purity π for margin samples is hard because of the apparent high contamination of normal cells; so we used a fixed very low purity value $\pi = 0.01$ (1%). This value is much lower than the margin's purity that we could have estimated by external tools such as PurBayes and Sequenza (data not shown). Smaller π's value rescale observed coverage to lower values (i.e., we ``throw away'' coverage from the targeted panels); in practice, this is a conservative choice because it renders harder to pass the statistical test (i.e., harder to reject the null).

The set of SNVs to test is divided according to the copy number status, and matched against the trained models. Each group is tested independent against all models trained from the different primary regions. The null hypothesis H_0 for testing a group of SNVs is the probability of detecting NV=0 mutated reads at the corrected coverage, given the parameters of the matched Beta-Binomial model

$$H_0\text{: BetaBin}(v = 0|\hat{r}; \alpha, \beta).$$

The p-value can be interpreted as the probability of a "negative SNV call" (NV=0, mutation not detected) to be a *true negative*, given the expected read counts if it was a clonal mutation. The tests are executed at confidence level 0.05, and corrected for multiple testing with the stringent correction possible (Familywise Error Rate, via Bonferroni).

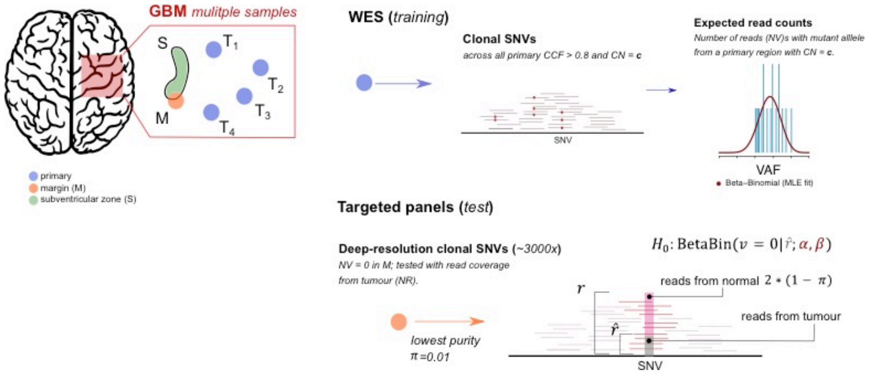
GBM mulitple samples
S
T$_1$
T$_2$
T$_3$
T$_4$
M
● primary
● margin (M)
● subventricular zone (S)
WES (training)
Clonal SNVs
across all primary CCF > 0.8 and CN = c
SNV
Expected read counts
Number of reads (NV)s with mutant alllele from a primary region with CN = c.
VAF
● Beta–Binomial (MLE fit)
Targeted panels (test)
Deep-resolution clonal SNVs (~3000x)
NV = 0 in M; tested with read coverage from tumour (NR).
H_0: BetaBin($v = 0|\hat{r}; \alpha, \beta$)
reads from normal $2 * (1 - \pi)$
r
$\hat{r}$
reads from tumour
lowest purity
π =0.01
SNV

Figure 1. Multiple Hypothesis Testing

With multiple GBM samples we detect clonal SNVs via WES of primary tumour regions, and train a Beta-Binomial model of the expected VAF of clonal mutations, for each region and copy number status. Then, from deep-sequencing of two targeted panels we detect clonal SNVs that have no variant reads in the margin. That probability under the training model is our null hypothesis, which we correct for multiple hypotheses (Bonferroni).