Statistical testing for GBM (Giulio + Guido)

We developed a statistical test to compute p-values for the null hypothesis that clonal mutations are missing from the *margin* samples (labelled as "M"). The test is divided in two steps: a *training*, in which we identify suitable mutations to train a statistical models of expected read counts from whole-exome (WES) data, and a *test,* in which we use the trained model to compute the likelihood a null hypothesis H_0 built from two targeted deep sequencing panels (labelled as "TES1" and "TES2"). The tests are carried out independently on each one of the input samples of a patient, and corrected for multiple testing. In the end, when we reject H_0, we have evidence that some clonal SNVs in the primary tumour are missing from the margin (*true negatives*). This corroborates our phylogenetic analysis, which elucidates that the margin is an ancestor of the primary tumour.

Computing Cancer Cell Fractions (CCFs)

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

Training a statistical model from read-counts

We used CCFs and CNA values to identify putative *clonal* SNVs with:

1. CCF >0.8 in *all* primary samples;
2. the *same CNA status* across all primary samples.

SNVs that do not fulfil these conditions are likely subclonal, and neglected by this analysis. Read counts for clonal SNVs are extracted from Platypus VCF files reporting Number of Variants (NV) and Reads (NR, i.e., coverage).

We corrected read counts for CNA status and tumour purity, before using them to train a statistical for the test. Correction is carried out to estimate how many of the *observed reads* r come from the actual tumour. The correction is a standard procedure which 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 two 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 –

the so-called Variant Allele Frequency (VAF) -- 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, and has been used to model read-counts in several other analysis [CITE.WHO?]; 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 copy number status for the input SNVs, and each WES sample of the primary tumour. By separating read counts by copy number status, we can adjust data for every non-diploid SNV in a more precise and consistent way. If, for a combination of parameters we find less than 10 SNVs, then we cannot train a suitable Beta-Binomial model. When that is the case we removed from downstream analysis this configuration of copy number status and input sample.

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

Testing

We identified testable SNVs from two deep-sequencing targeted panels. Because panels have ~3000x coverage, we require each SNV to have at least *k=10* reads with the variant allele. When that is not the case, the SNV is considered missing from the panel. We particularly care about the ones that are missing in the margin sample (M): if the margin was ancestral to the primary tumour, we expect it to lack some SNVs that are clonal in the primary WES samples.

Giulio 16/4/2018 09:33
Comment [2]: We added this in the last revision of the method.

If a patient has more than one margin sample (e.g., Patient A23), we require the mutation to be absent 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). 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 are corrected as with the training set. Copy number status for mutations in the margin is the same for the training set, by construction. Purity correction instead requires some considerations. Estimation of purity π for margin samples is hard because of the apparent high contamination of normal cells; standard tools like PurBayes and Sequenza have estimated purity values below $\pi = 0.30$ (30%) in most cases (data not shown). To be conservative, however, we have used a fixed, worst-case low purity estimate of $\pi = 0.01$ (1%). This value is much lower than the margin's likely purity, and the power of the test decreases with coverage. Thus, a conservative purity estimate leads us to rescale observed coverage to lower values (i.e., we ``throw away'' coverage from the targeted panels); in practice, this is conservative since 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<k* (with *k=10*) mutated reads at the corrected coverage, given the parameters μ, ρ of the matched Beta-Binomial model

$$H_0 \colon \sum_{w=1}^{k} \mathrm{BetaBin}(v = w | \hat{r}; \mu, \rho).$$

The p-value is the probability of finding less than k reads with the variant allele at the designed locus with coverage $\hat{r}$, given the fact that we expect for a clonal SNV the number of reads to follow a Beta-Binomial model with parameters α, β. Thus, this can be interpreted as the probability of having missed a truly clonal SNV (mutation not detected). Low p-values are evidence for the fact that the missing mutation can be a real *true negative*. The tests are executed at confidence level $\alpha = 0.05$, and corrected for multiple testing with the stringent correction possible (Family-wise Error Rate, via Bonferroni).

Giulio 17/4/2018 15:48
Comment [3]: I think I am reading too much here, maybe this can be removed.

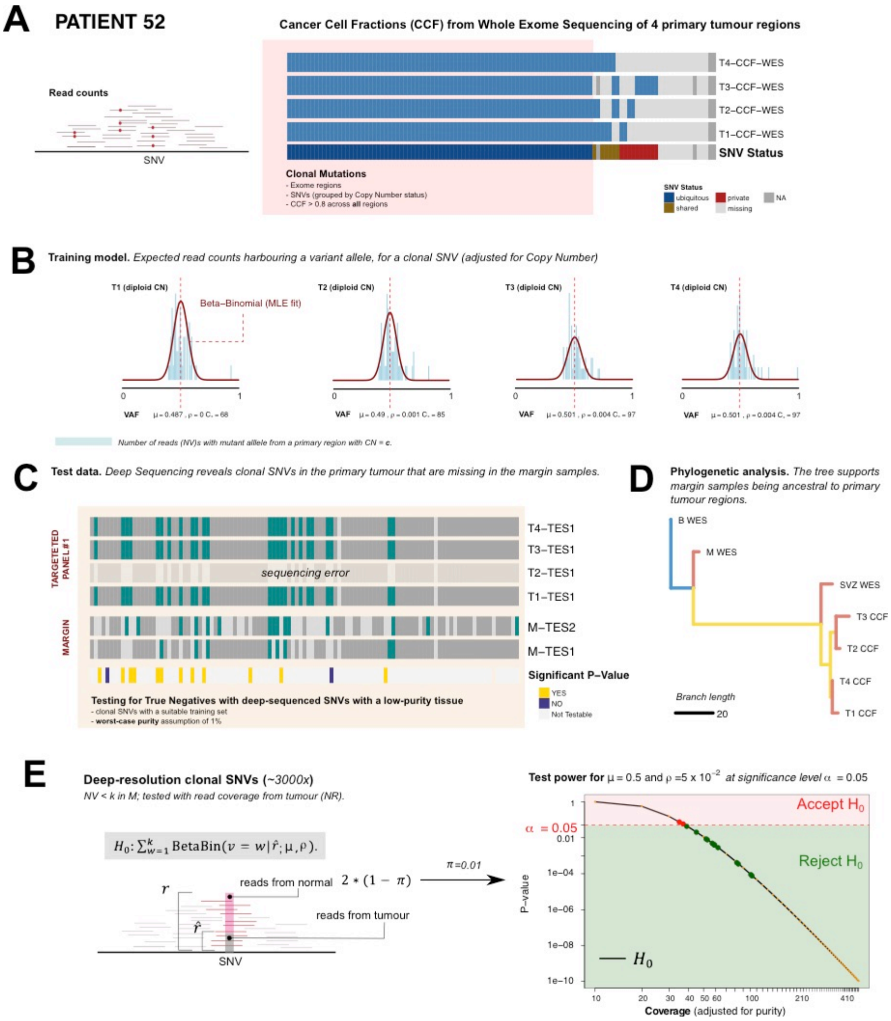

Figure 1. Methodology.

A. *We detect clonal SNVs in multiple GBM samples via WES of primary tumour regions. Clonal SNVs have CCF >0.8 and the same CNA status across all samples.* ***B.*** *Using read counts from clonal SNVs, we train a Beta-Binomial model of their expected Variant Allele Frequency (VAF) of clonal mutations, for each primary sample and copy number status. This model captures, for each clonal SNV, how many reads we shall expect harbouring the variant allele.* **C.** *We use deep-sequencing data from two targeted panels to identify which clonal SNVs are missing (or at least less have than k reads) in every margin sample. The presence of these SNVs suggests that margin samples might be ancestral to primary ones.* **D.** *Standard phylogenetic reconstruction confirms that margin samples are ancestral to the primary regions.* **E.** *A statistical test for SNVs identified in panel C is used to test a null hypothesis that their read counts are generated from the distribution of read counts for clonal mutations in the primary regions. Rejecting the null means that we have evidence for the fact that these SNVs are not clonal in the margin, providing further evidence that the margin is ancestral to the primary regions. The power of the test increases with higher coverage; we use a conservative setting for the test, as we estimate the actual number of reads coming from tumour cells in a worst-case purity scenario of* $\pi = 0.01$ (1%)*. The test is also corrected for Multiple Hypothesis Testing via Bonferroni.*