Identification of Enriched Variants using Genotype Recovery Algorithm

Cancer progression and evolution is characterised by the non-linear and repetitious cycles of clonal expansion with the accumulation of effective genetic variants, and ongoing clonal selection without disrupting the adaptive genetic landscape of tumour microenvironment [1]. Throughout the process of clonal expansion, cellular heterogeneity of tumour is maintained with the presence of mature tumour cells, both cancer and non-cancer stem cells and non-cancer cells, contributing to the stochastic transcriptional regulation of genetically identical cells allowing them to respond to such diverse microenvironment to progress further at the expense of normal cells [2].This admixture of various cell types within tumour tissue ecosystem also determines the overall purity (cellularity) of the tumour samples. So, to calculate the true tumour sample coverage (a total estimate of both major and minor read counts from the tumour cells only), the number of reads from “normal” tissue within tumour tissue has to be deducted based on the cellularity of the given tumour sample. Utilising these aspects of clonal expansion of tumour, we used binomial model to derive any deviation in the distribution of minor alleles in samples, potentially affecting both genotypes and phenotypes, or in this case, emergence of resistance to MEK-inhibitors.

In this study, a cohort of 59 unrelated germline samples sequenced on the name exome capture method were used to estimate the background frequency for minor alleles observed under the same sequencing protocol. This background frequency was used as a reference to quantify the allele frequency changes in three THP-clones treated with MEK-inhibitors (parental clones, clones treated with AS703026 and Seluminitib). Significant changes in read counts for minor alleles and the direction of allelic change was used to quantify clonal expansion in all three THP-clones.

In order to calculate the new estimate for alternate read counts in treatment samples, we used only good quality bases (that are less likely to be sequencing errors). To obtain these pile up statistics for good quality bases (minimum base quality score >20) overlapping each genomic position (SNPs) were obtained from individual bam files (for all controls and treatment samples) using in-house R script utilizing pileup program in R package, 'Rsamtool' [3]. We then calculated the normal allele frequency using major and minor read counts in control samples for each SNP for both reference base calls (0/0) and non-reference base calls (0/1, 1/1). The normal allele frequency in the case of missing genotypes/coverage was set to be an error estimate of 0.001 (probability to observe non reference base due to inherent sequencing errors). Using the cellularity (which in our case was taken to be 100% (pure tumour culture), and the normal allele frequency from control samples using both reference and non-reference base calls, we then calculated the expected major and minor read counts in our samples contributed by the normal cells. After subtracting the major counts in normal (normal in tumour) from the major counts in samples (THP treatments) and minor counts in normal (normal in tumour) from the minor counts in samples (THP treatments), we then calculated the new estimates for true major count and minor counts in true tumour in samples. The estimates for true tumour coverage was also calculated by subtracting the normal coverage in samples from the sum of major read counts and minor read counts in the samples.

One-sided binomial test was performed to calculate the P-value for any deviation of minor read counts in true cancer samples than expected. The binomial model uses an assumption that the observed outcomes are dichotomous and independent to each other. In a binomial distribution described as **X ~ B(n, p)**, where **X** is a binomially distributed random variable (alternate read counts in true tumour in samples), **n** is the total number of experiments (total number of reads in true tumour in samples) and **p** is the probability of successful outcomes (allele frequency in controls) for each experiment (SNP), respectively. The expected value of **X** is calculated to be **E[X] = np**, and the variance to be **Var[X] = np (1 – p).** The test statistics (**z**-score) calculated to be **z = (X-np)/sqrt(np(1-p))**, where **z** is the number of standard deviations an observation (observed minor read counts) is away from the mean, **np**. The calculated **z**-score (>6) was chosen and converged to P-value (p-value <10-9) to identify the significantly enriched variants with respect to both reference base calls and non-reference base calls independently. The putatively resistant variants in three THP clones were then clustered based on the calculated P-values using fuzzy clustering method based on Euclidian distance and c-means objective function (weighted square error function) where each data point is assigned a membership value to be grouped into each nth cluster to a certain degree [4]. The *m* value (value of fuzzy *c*-means parameter) of 1.8 was chosen to identify a total of nine clusters (*c*=9) appropriately for clustering.

**Results**

The fuzzy clustering of sample1 (parental clone), sample2 (AS703026) and sample3 (Seluminitib) has identified the enriched SNPs (potentially contributing to MEK resistance in the cancer samples). Using reference calls, a total of 621 (cluster 1), 88 (cluster6) and 22 SNPs (cluster3) were found to be significantly enriched in Seluminitib, AS703026 and in both clones, respectively. Whereas, using non-reference calls, only seven (cluster1), three (cluster4) and one SNP (cluster 2) were found to be enriched in Seluminitib, AS703026 and in both clones, respectively, as compared to the controls.

**References**

1. Gatenby RA, Gillies RJ: **A microenvironmental model of carcinogenesis**. *Nat Rev Cancer* 2008, **8**(1):56-61.

2. Mannello F, Ligi D: **Resolving breast cancer heterogeneity by searching reliable protein cancer biomarkers in the breast fluid secretome**. *BMC Cancer* 2013, **13**:344.

3. Morgan M, Pagès H, Obenchain V, Hayden N: **Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import**. In*.* R package version 1.23.0, <http://bioconductor.org/packages/release/bioc/html/Rsamtools.html>.

4. Futschik M: **Mfuzz: Soft clustering of time series gene expression data**. In*.*: R package version 2.30.0. <http://www.sysbiolab.eu/software/R/Mfuzz/index.html;> 2015.