Cancer CRISPR Targets

Michael Bradshaw

Introduction

The purpose of this project was to investigate ways of identifying cancer mutations that could serve as CRISPR targets for a potential therapeutic. The ideal target would be the oldest and most ubiquitously present mutation. Here I show how regions of the genome with abnormal amounts of mutation could be identified and subsequently how the mutations within these regions could be ranked in a way I hope can proxy age of the mutation. The methods used for this project are publicly available on Github and is ready for use on any variant call format file.

Methods & Materials

Code availability:

All code used in this project and instructions on how to replicate the results and run the tool can be found on Github at https://github.com/MSBradshaw/CancerCRISPR.

Data Source:

Fastq files from lung cancer cell line NCIH2170_LUNG were downloaded from https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR8652088. This data was generated as part of the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012).

Alignment & Calling:

Fastq files were aligned to the HG37 reference genome using bwa mem (Li 2013). Variants were called with HaplotypeCaller from GATK.

Results

25kb Bins:

Due to their size, full-size high resolution version of all figures can be found on this project's Github https://github.com/MSBradshaw/CancerCRISPR

To determine if there were regions of the genome with increased variability, hopefully due to an amplicon caused by the cancerous mutations, I split each chromosome up continuously from start to end into 25kb bins. Within these bins I then counted the number of calls in the VCF with an approximate read depth (DP) >= 10 (Figure 1, first column of plots). To aid in understanding how each bin’s count compared to bins in the same chromosome, a z-score was also calculated (Figure 1, second column of plots).

In about half the chromosomes (2,3,5,6,10,11,12,14,17,20,21,X) a bin exists with a z-score >= 5. Considering in normally distributed data, >99% of the data falls within 3 standard deviations of the mean (a z-score -3 to 3), these z-scores we see over 5 are very notable. This means there are certainly regions within this cancer cell line that are more mutated than others.
It is important to note that some of these high z-score spikes fall very near to centromeres and telomeres of the chromosome, known low quality regions of the genome which may explain the increase in variation. Good examples of this can be seen in chromosomes 20 and X where there is a large z-scores spike in the bin immediately to the right of where there is a complete absence of sites (assumedly the location of the centromere) in the first column plots.

[Figure 1]

Figure 1. Column 1: number of calls in the VCF with a DP >= 10. The x-axis is the position within a chromosome, y-axis is the number of calls with DP >= 10 at a given bin position. All bins in this figure are 25,000 bases in length, there is no overlap between bins. Please note that the scale of the y-axis can differ widely from one plot to another. Column 2: z-score of the number of calls with DP >= 10 in a bin. The x-axis is the position within a chromosome, y-axis is the z-score of the number of calls with DP >= 10 at a given bin position. When calculating z-score the sample is the number of sites with DP >10 in a bin compared to the mean and standard deviation of all bins in the same chromosome.

Sliding Window:

Bin sizes and locations used in figure 1 were arbitrary and rather high granularity. To address this, I did a similar analysis but using a sliding window across each chromosome. The size of the window was set to 1,000 bases (a far more feasible CRISPR target than 25kb) and with each iteration or “slide” the window was moved by 100 bases. Results of this approach can be seen in Figure 2. Using a sliding window with a step size smaller than the window size means there will be overlap between nearby window measurements, resulting in very steep parabolic peaks in the plots. These overlaps and semi-duplicate measurements will later allow us to pinpoint the window location with the greatest number of mutations.

As observed in Figure 1, there are still peaks very near the centromeres of the chromosomes (Figure 2, Chromosome 16 for example), but with the increased granularity of Figure 2 we can see new peaks that are not obvious close to the centromeres these include:

- the tallest peaks in chromosomes 3,5 & 6,
- the 4th peak with z-score over 50 in chromosome 7,
- second tallest in chromosome 9,
- 4th peak with z-score > 50 in chromosome 11,
- first peak chromosome 12

- second tallest peak chromosome 19

[Figure 2]

Figure 2. Column 1: number of calls in the VCF with a DP >= 10. The x-axis is the position within a chromosome, y-axis is the number of calls with DP >= 10 at a given window position. Column 2: y axis is the z-score of the number of calls with DP >= 10 in a window, x-axis the chromosomal location. Windows with no calls with DP >= were excluded from this plot and z-score calculations. *It should be noted that although figures in the two columns look identical, they are not, this is not an error.* This is due to the fact that the average mean for the chromosome is ~2 and the average standard deviation is ~2 in all chromosomes, resulting in z-scores and actually values that are extremely similar and visually indistinct in a plot.

The top 50 peaks were then extracted from the sliding window results. In order to do this and receive truly distinct peaks the overlapping windows had to be controlled for. This was done by sorting all windows by their z-score in descending order then iteratively moving down the list while removing all windows down stream that overlapped with the current window.

Ranking mutations within a window:

Once isolated, the calls within the windows of interest were extracted from the VCF. Calls were then ranked and sorted by what I will call the “ratio of the sum of absolute differences of allelic depths(AD)” (Equation 1) which I will refer to as just “ratio” from now on. The goal behind this metric was to prioritize mutations that were more different than expected. My assumption is that a greater difference translates to the mutation being older (because it is more abundant).

$$ratio = \frac{\sum_{i=0}^{size\ of\ x} \left| x_{i - max(x)} \right|}{sum(x)}$$

Equation 1. Let x be a list of the allelic depths as reported in the VCF, of length 2 or more. The *max(x)* function denotes the maximum value in the list x. The *sum(x)* function denotes the sum of the values found in *x*.

The results of this can be found on the Github in the file entirited ‘results.txt’.

I will now highlight the top 5 calls from the highest window in the results not adjacent the centromere. This window is at chr6:32632300-32633300. This region falls within the gene

HLA-DQB1 (major histocompatibility complex). According to gnomAD this gene is fairly mutation tolerant, in the pLoF section it has pLI = 0.01 (closer to 1 mean more intolerant to mutation) and observed vs expected (oe) of 0.45 (closer to 0.0 is worse). Despite this, gnomAD has no information or record of SNPs in the same location as the top 5 in my results (Table 1).

id	chr	start	end	ref	alt	AD	ratio	Known SNP
0	6	32633185	32633186	A	C,G	0,29,5	1.558824	No
1	6	32633211	32633212	G	A,C	0,4,23	1.555556	No
2	6	32633284	32633285	C	T	0,31	1	No
3	6	32632473	32632474	T	G	0,5	1	No
4	6	32632482	32632483	A	G	0,3	1	No

Table 1. Calls with the highest ratio in the highest ranked window not adjacent to a centromere. Known SNP column denotes if gnomAD had a record of this SNP already.

Conclusion & Discussion

Here I have shown how the frequency of mutations and a static related to allele frequency could be used to identify mutations as potential CRISPR targets. I have also created a simple and easy to use tool for doing so.

Inorder for the results produced by this tool to be truly useful, additional filtering of the results is still required. It is known anything near or within highly variable regions needs to be removed and not considered, but it will also be necessary to filter out mutations known to be germline (or non-tumor specific) mutations. Inforder to do this a second VCF from a non-tumor sample from the same patient would be required. Given that I was working specifically with a cancer cell line this non-tumor information was not available. Filtering to considering only polymorphic sites with 3 or more observed alleles could also be an effective way to super saturate the results with windows with genuine cancer mutations.

The ratio used for ranking the calls within a window is certainly not perfect. One problem with it is that it will always give a ratio >1 to any call with more than 1 alternate allele. In hindsight this metric also does not account for the genotype of the call, it worked from the assumption that each call was supposed to be homozygous. The metric could be improved by focusing on differences from expected allele frequencies given the genotype. For instance if the call is labeled heterozygous (0/1) it may be better to rank allele balances based on their difference from 0.5 or from 0.0 and 1.0 if the call is homozygous.

While an effective ratio representing the allele balance could be used as a proxy for the age of a mutation (the older and thus more widely present it is, the better a target for CRISPR) the ability to build some hierarchical tree describing the relatedness of mutations so that the oldest mutation, or root of the tree, could be discovered. This tree could be constructed by referring back to the aligned reads and analysing the co-occurrence of these mutations in addition to their overall abundance. One challenge with this possible approach to the problem is that it would be limited by the length of the sequencing reads.