Characterization of somatic events for diffuse large B-cell lymphoma in the Clinical Trial Sequencing Project

Xavier Loinaz, Qing Zhang, Julian Hess, Jialin Ma, David Heiman, Chip Stewart, Gad Getz

Getz Lab, Cancer Genome Computational Analysis Group

ABSTRACT

Under the Clinical Trial Sequencing Project funded by the NIH, we attempt to characterize relevant somatic mutations, indels, copy number alterations, structural variants, and mutational signatures for diffuse large B-cell lymphoma (DLBCL) for a cohort of 128 patients with fresh frozen samples and average sequencing depth of about 20x. We use MuTect1 for calling point mutations, Strelka2 for calling indels; the GATK4 CNV pipeline and ABSOLUTE for calling copy number alterations; apply a consensus filter for the structural variant calls of Manta, SvABA, and dRanger; and use SignatureAnalyzer for finding mutational signatures. A significant finding of our analysis backed by orthogonal methods was the greater frequency of microsatellite indel calls relative to other pipelines run by other research organizations, as well as previously published analyses for DLBCL. We also provide evidence for our somatic copy number alteration calls being more robust.

BACKGROUND

Diffuse large B-cell lymphoma (DLBCL), a type of cancer of B-lymphocytes with specific cellular characteristics, is the most common type of non-Hodgkin lymphoma in the United States as well as the world. More than 18,000 people are diagnosed with DLBCL each year, and DLBCL makes up approximately 30% of all lymphomas.

Within the Clinical Trial Sequencing Project's analysis working group for DLBCL in collaboration with 15 other research groups, we attempt to gain a comprehensive understanding of genetic aberrations that drive the growth and proliferation of DLBCL tumors within patients. The cohort for this analysis also tests two different chemotherapeutic regimens: R-CHOP, which has been the standard treatment for DLBCL for approximately 15 years with a cure rate around 65%, and EPOCH-R, which is a novel regimen. Under our analysis, we may also compare efficacies of either treatment for different subtypes of DLBCL.

At this stage of the project, we are awaiting deeper coverage data for our cohort and have preliminarily run our pipelines for detecting mutations, copy number alterations, and structural variants, which we can validate with other groups and previous studies to prepare our pipelines.

METHODS

This project gave the Getz Lab the opportunity to try its wolF pipeline for the first time, which is a pipeline based off a cluster job management API developed by the lab. The wolF pipeline was used for mutation calling and structural variant calling.

Mutation Calling and Validation
Figure 1 shows a schematic of our mutation calling pipeline. MuTect1 is used for calling point mutations and Strelka2 is used for calling indels. We also feature ContEST, which accounts for cross-contamination that may have occurred in our samples, as well as DeTiN, which accounts for contamination of tumor in our paired normal samples, rescuing certain somatic events that otherwise would have been filtered out. Additionally, postfilters are applied to account for alignment errors, orientation bias, and artefact/germline filtering. GATK's Funcotator tool does gene annotations for mutations.

Copy Number Calling and Validation
For our copy number calling pipeline, we used the standard GATK4 CNV pipeline. In order to account for the relatively lower coverage of our cohort, we decreased the coverage threshold for normal heterozygous sites used for segmentation, as well as modified a parameter to reduce the tendency for segmentation. We then ran ABSOLUTE to call sample tumor purities and GISTIC2 for validation of known CNV drivers for DLBCL compared to prior works (such as *Molecular Subtypes of Diffuse Large B-cell Lymphoma are Associated with Distinct Pathogenic Mechanisms and Outcomes*, Chapuy et al., (2018)).

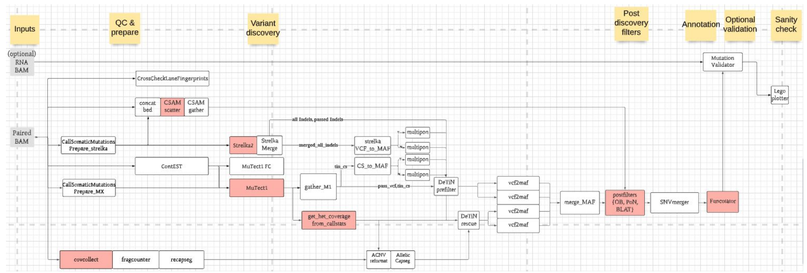

Figure 1. A schematic of wolF's mutation-calling pipeline, showing all tools/workflows along with yellow labels above to show what it accomplishes at each approximate point of the pipeline.

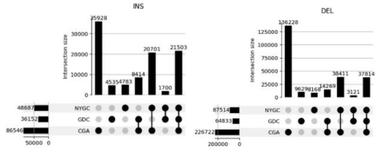

Figure 2. These plots show bars of concordance for our indel calls from wolF's pipeline with those of the New York Genome Center (NYGC) and Genomic Data Commons (GDC) for insertions (left plot) and deletions (right plot). The colored-in dots in the bottom part show the combination of pipelines from which we are measuring the number of concordant calls in its corresponding bar above. The vast majority of these indels tended to be at microsatellite loci.

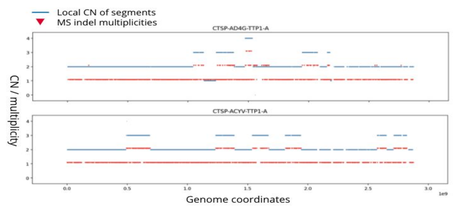

Figure 3. This plot shows for two given tumor samples the local total copy numbers at specific regions in the genome as well as multiplicities of MS indels called as determined by the same Dirichlet processes used to call absolute copy number in ABSOLUTE.

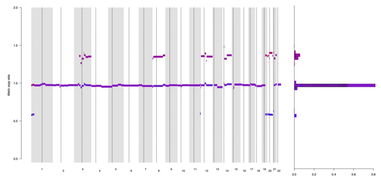

Figure 4. This is an example of the copy number segmentation for a particular tumor-normal pair with our modified GATK4 CNV pipeline. Lack of subclonal events can be a qualitative indication of good CNV calls.

2018 Chapuy et al paper (304 patients):

Gene	# of patients with driver SV event	Proportion of patients with driver SV event
BCL2	62	0.2
IgH-Enh	115	0.38
MYC	24	0.08
PDL1 (CD274)	6	0.02
PDL2 (PDCD1LG2)	9	0.03

wolF (128 frozen pairs):

Gene	# of patients with driver SV event	Proportion of patients with driver SV event
BCL2	24	0.19
IgH-Enh	46	0.36
MYC	16	0.13
PDL1 (CD274)	5	0.04
PDL2 (PDCD1LG2)	3	0.02

Figure 5. Comparison of frequency of structural variants detected within driver genes for patients via wolF's pipeline and that for the Chapuy et al. paper. All frequencies are within 0.05 of each other in terms of absolute difference, indicating fair concordance.

METHODS

Structural Variant Calling and Validation
wolF's structural variant pipeline is comprised of a consensus calling of three different structural variant callers: Manta, SvABA, and dRanger. Events are validated by Breakpointer, and events are further filtered out by a consensus filter in which any events not called by two different structural variant callers (after all events are standardized into dRanger format so as to help identify concordant events) are taken out.

RESULTS

Mutation Calling
A significant finding in our mutation calls was a significantly greater number of microsatellite indels called relative to the pipelines of other groups in our consortium. This was validated through four orthogonal methods: correlation of purity with mutation burden (since greater purity allows for greater indel-calling power but for artefacts this would not be the case), appropriate phasing of indels with one haplotype, high base qualities in regions where indels are called, and indels matching clonal multiplicities of tumor. This last orthogonal validation is shown in Figure 3, and it was determined that greater than 95% of microsatellite indels called are clonal.

Copy Number Calling
For copy number calling, we were able to get segmentation that looked qualitatively acceptable and had less events that appeared to be aberrantly called subclonal based off the other pipelines we tried. There was some degree of oversegmentation, but for the most part this was insignificant for arm-level and larger focal events. Figure 4 shows a segmentation example.

RESULTS

Structural Variant Calling
Through running our structural variant calling pipeline, we were able to were able to replicate similar proportions of structural variants happening within driver genes for DLBCL as determined by *Molecular Subtypes of Diffuse Large B-cell Lymphoma are Associated with Distinct Pathogenic Mechanisms and Outcomes*, Chapuy et al., (2018), as shown in Figure 5. However, one critical driver gene, BCL6, is missing due an annotation issue in the pipeline. This is currently being patched.

CONCLUSION

Overall, although we are still awaiting higher-coverage sequencing data from the Genomic Data Commons, we are close to finalizing results for this preliminary analysis with ~20x sequencing depth on average. This also served to work out any kinks or bugs in wolF's pipeline, as we determined with the structural variant annotator as well as some of Funcotator's annotations with mutations.