A Next-Generation Sequencing Based Analysis of Clonality across 39 Subjects Treated for Lymphoproliferative Disorders Reveals Matching Clones in the Diverse *IGH* Locus

I033

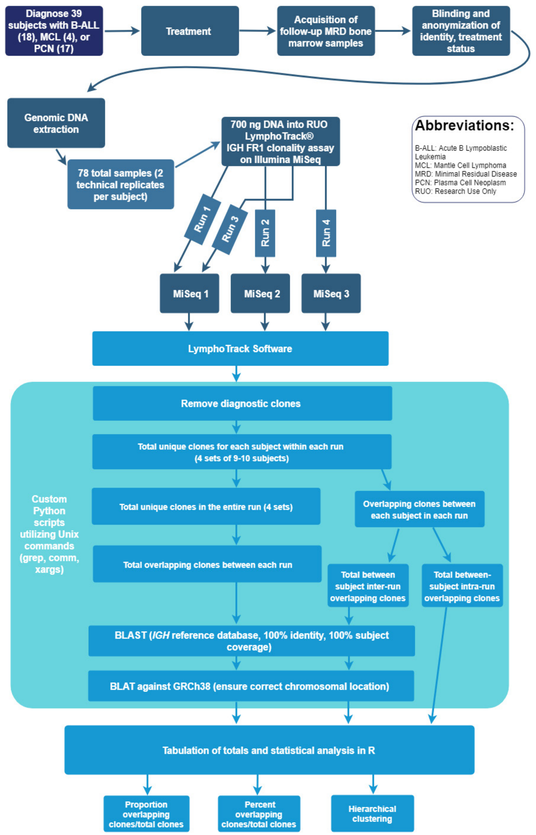
Alyssa M. Zlotnicki[1], Austin Jacobsen[1], Edgar Vigil[1], Kasey Hutt[1], Andrew Carson[1], Ying Huang[1], Khedoudja Nafa[2], Maria Arcila[2,3], and Jeffrey E. Miller[1]

[1]Invivoscribe, Inc., San Diego, United States, [2]Dept of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, [3]Dept of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States

Introduction

The detection of clonality of the rearranged immunoglobulin heavy chain (*IGH*) locus is the basis for early detection of B-cell malignancies. The genomic diversity of this locus has been analyzed in many studies, showing that VDJ rearrangement and the random additions of nucleotides that occur afterwards produce a high diversity across healthy and affected individuals.[1, 2] This is especially true in the complementarity determining region 3 (CDR3) region, which is the most variable region of the *IGH* locus.[4] The resulting possible combinations generated from these rearrangements and additions generates an immune repertoire of > 1 × 10^{12} for Ig receptors[1]. Based on these studies, it is extremely rare to detect the presence of the exact same clone sequence in separate human subject repertoires. Here we present an analysis of the diversity of the *IGH* locus in subjects treated for various lymphoproliferative disorders using a Next-Generation Sequencing based clonality assay.

Methods



Results: Proportion of Overlapping Clones to Total Clones Between Subjects

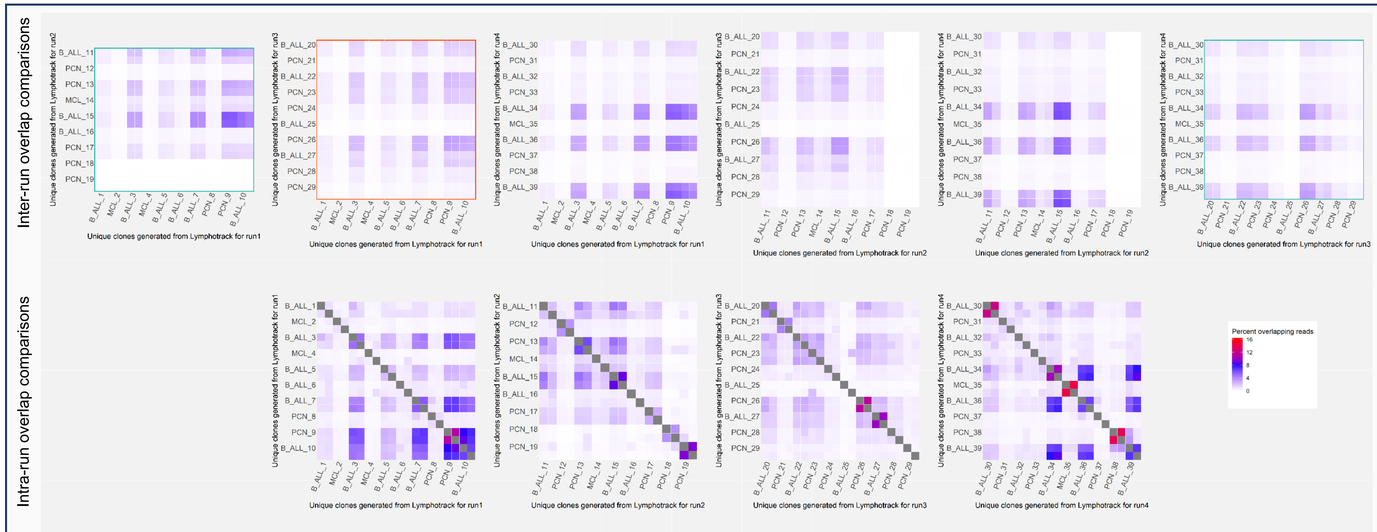


Figure 1: Percent of unique clones shared between the repertoires of any two given subjects within a sequencing run or between two different sequencing runs. The first row contains comparisons between subjects between different sequencing runs, and the second row contains comparisons between subjects within the same sequencing run. The latter set of comparisons is presented in order to demonstrate the level of expected overlap within a run due to barcode-hopping/crossover, and to provide as a basis for comparison to the former. The orange box indicates an inter-run comparison between runs sequenced on the same MiSeq®. The turquoise boxes indicate an inter-run comparison between runs sequenced on the same day.

Results: Proportions of Overlapping Clones Based on Run Average and Disease Types

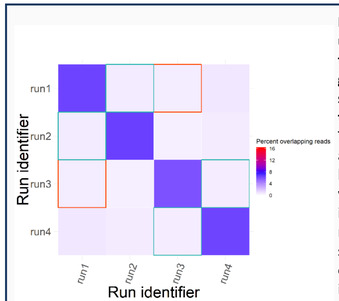


Figure 2a: Average percent of unique clones shared between the repertoires of any two given subjects within a sequencing run or between two different sequencing runs. This plot represents the average number of overlaps between any two subjects within the two given run identifiers, e.g. the average number of overlaps between subjects in run1 and run2. The orange box and turquoise boxes indicate inter-run comparisons between runs sequenced on the same MiSeq® and same day, respectively.

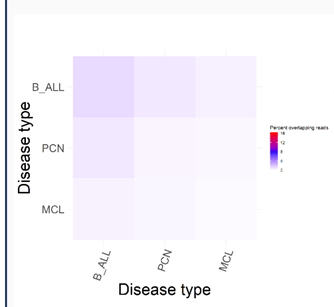


Figure 2b: Percent of unique clones shared between the repertoires of subjects with the same disease between any two different sequencing runs. This plot represents the average number of overlaps between any two subjects in different sequencing runs that were treated for a particular disease type or types: e.g. the average number of clones shared between subjects with B-ALL and subjects with PCN between any two given subjects between any given run.

Results: Subset Motifs Found in the CDR3 Regions of Clones Shared Between All Runs

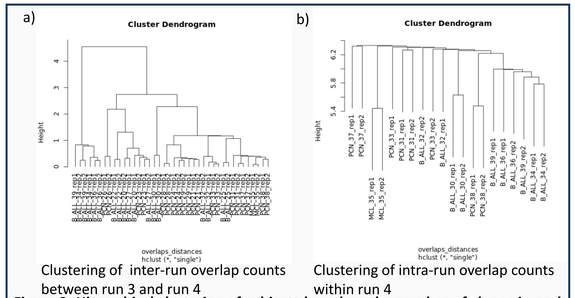
Table 1: Top 6 subsets found in 267 clusters of clones shared by all runs based on the major *IGH* CDR3 subsets found in the repertoires of CLL subjects by Agathangelidis et. al. (2012)

Subset Number	Number of clone clusters						CDR3 pattern
	Ranked 1st by p-value		Ranked 2nd by p-value		Ranked 3rd by p-value		
	Count	Percent	Count	Percent	Count	Percent	
#31	44	16.5%	0	0%	7	2.62%	ARxxxxxxxxxxxxYYYxMDx
#64B	0	0%	42	15.7%	0	0%	A[KRH][DE]xx[AVLI]VVxxx[AVLI]xYYYYGMDx
#28A	7	2.62%	2	0.749%	33	12.4%	ARxxxGxxYYYYYGMDx
#5	0	0%	7	2.62%	11	4.12%	ARxxxxxx[AVLI]xxxYYYYxMDx
#2	12	4.49%	0	0%	0	0%	[AVLI]x[DE]xxxM[DE]x
#12	0	0%	9	3.37%	0	0%	ARDxxYYDSSGYY[ST]xxxDx
# clusters where no subset p < 0.05	194	72.7%	197	73.8%	204	76.4%	N/A

Background: There were 1279 unique clone sequences shared between all four runs. Given the high variability of the *IGH* CDR3 region, it would be rare to observe patterns in this region due to random chance, and so the CDR3 regions of the shared clones were extracted for comparison. To determine if shared clone motifs were similar to motif patterns observed in clinical samples, the comparison was performed against a set of 19 *IGH* CDR3 subset motifs derived from 7424 clinical chronic lymphocytic leukemia (CLL) samples; the largest study of its kind at its time of publication.[3] Motif matches between the clones and the subsets would indicate: 1) a strong consensus group can be formed from the clones, showing that some sort of genetic event (e.g. clonal evolution) has caused sequence convergence to develop post-treatment. 2) these motifs are possibly clinically-relevant, given they share motifs that are associated with certain clinical attributes (e.g. age of patient, aggressiveness of disease), and gives support to the idea that a similar clustering analysis performed with treatment information for the diseases analyzed here (B-ALL, MCL, PCN), may reveal motifs similarly associated with clinical outcomes, given that clones in *IGH* locus cause all aforementioned diseases, although this would not be the case if convergence is related to the treatment itself.

Methods: Clusters of CDR3 clone regions were generated using three of the criteria followed in other CDR3 clustering practices: 50% amino acid identity, 70% amino acid similarity based on physio-chemical properties, and same CDR3 amino acid sequence length.[3] Position-specific scoring matrices (PSSMs) were generated for each cluster, each CLL subset, and 31 random sequences with mean length 19 using MEME's sites2meme with pseudocounts=0.0001 and background amino acid frequencies derived from analyses of vertebrate polypeptides.[5,6,7] Each cluster PSSM was compared to the set of CLL subset and random sequences PSSMs using MEME's tomtom to determine motif matches. Any motif match that had a p-value higher than that of a random sequence or 0.05, whichever was lower, was considered an insignificant match and was not used for cluster motif classification.

Results: Hierarchical Clustering



Clustering of inter-run overlap counts between run 3 and run 4

Clustering of intra-run overlap counts within run 4

Figure 3: Hierarchical clustering of subjects based on the number of clones in each subject that overlap with at least one other run.
a) clustering between subjects in two different MiSeq® runs based on unique clone count between subjects. Other inter-run clustering graphs showed similar patterns. b) clustering between subjects within the same MiSeq® run based on unique clone count between subjects. Other intra-run clustering graphs showed similar patterns. Hierarchical clustering was performed using R's hclust function, using the single agglomeration method. Counts were scaled using z-score scaling before clustering.

Discussion and Conclusions

The proportions of overlapping clones (where an overlapping clone was defined as a exact match between two clones across the entire length of their sequences) between subject samples across runs are lower than what is seen within runs, but is distinctly present across runs. This demonstrates that the overlaps between runs do not mimic the pattern of overlap seen from what you would expect with sample contamination, which is seen in the intra-run samples due to barcode-hopping. The number of overlaps seen between runs sequenced on the same MiSeq® was not substantially higher or lower than other inter-run comparisons, which would indicate that these overlaps are not due to contamination. Clustering revealed overlap counts were similar based on run or subject obtained from, further indicating the substantive presence of these overlaps. These overlaps were not expected to be present at the level observed in each subject's repertoire given the high diversity of the *IGH* locus and the number of possible unique Ig receptor sequences. This could be explained by convergence of these *IGH* FR1 sequences in response to treatment. Clustering of the CDR3 regions of *IGH* clones shared by all sequencing runs revealed that a strong consensus subset could be generated via clustering based on amino acid identity, amino acid similarity, and sequence length. When these motifs were compared to 19 major CLL subsets[3], the proportion of sequences that were classified and unclassified were similar to those found by Agathangelidis et. al. (2012) (27.3%/72.7% as compared to 30%/70%), which indicates that the diversity of clones is not higher or lower than expected by this metric. The most significant CLL subset found in the most shared clones in our dataset (#31) has the characteristics of being found in younger subjects and subjects with aggressive disease. An analysis using subject samples with known treatment status may inform future interpretations of results from post-treatment subject samples, particularly considering clonal evolution and tracking future clones. If shared clones were found to be strongly associated with the subject's treatment type, it would provide evidence to suggest that their convergence is treatment driven. This would inform the analysis of the efficacy of newly developed drugs, as this would indicate that certain clones may be generated from the drug in addition to the disease. It is also possible that these clones are a sign of early progression of the disease that remain artifacts post-treatment. Analysis of clones present throughout the course of treatment would reveal if this is the case.

References

1. Ho C, Arcila ME. Minimal residual disease detection of myeloma using sequencing of immunoglobulin heavy chain gene VDJ regions. Semin Hematol. 2018 Jan;55(1):13-18. doi: 10.1053/j.seminhematol.2018.02.007. Epub 2018 Feb 23. Review. PubMed PMID: 29759147.
2. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012 Jul;13(5):363-73. doi: 10.1038/gene.2012.12. Epub 2012 May 3. Review. PubMed PMID: 22551722.
3. Agathangelidis A, Darzentas N, Hadzidimitriou A, Brochet X, Murray F, Yan XJ, Davis Z, van Gastel-Mol EJ, Tresoldi C, Chu CC, Cahill N, Giudicelli V, Tichy B, Pedersen LB, Foroni L, Bonello L, Janus A, Smedby K, Anagnostopoulos A, Merle-Beral H, Laoutaris N, Juliusson G, di Celle PF, Pospisilova S, Jurlander J, Geisler C, Tsaftaris A, Lefranc MP, Langerak AW, Oscier DG, Chiorazzi N, Belessi C, Davi F, Rosenquist R, Ghia P, Stamatopoulos K. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. Blood. 2012 May 10;119(19):4467-75. doi: 10.1182/blood-2011-11-393694. Epub 2012 Mar 13. PubMed PMID: 22415752; PubMed Central PMCID: PMC3392073.
4. VanDyk L, Meek K. Assembly of IgH CDR3: mechanism, regulation, and influence on antibody diversity. Int Rev Immunol. 1992;8(2-3):123-33. Review. PubMed PMID: 1318346.
5. Dyer, K. F. 1971. The quiet revolution: A new synthesis of biological knowledge. Journal of Biological Education 5:15-24
6. King, J. L. and T. H. Jukes. 1969. Non-Darwinian evolution. Science 164:788-798
7. Beals, M., Gross, L., & Harrell, S. (1999). Amindo Acid Frequency. Retrieved September 13, 2019, from http://www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm.

AMP, November 05-09, 2019, Baltimore, Maryland, USA