POSTER # I007 Evaluation of a High-Throughput CDR3 Identification Algorithm for 1136 Clinical Samples Targeting the *IGHV* Leader, *IGH* FR1, and *TRG* Loci

Alyssa M. Zlotnicki[1], Tran Nguyen[1], Ying Huang[1], Julian D'Angelo[1], Javier Velazquez-Muriel[1], Maria Arcila[2] and Jeffery E. Miller[1] | [1]Invivoscribe, Inc., San Diego, United States, [2]Dept of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY

INTRODUCTION

- The identification of the CDR3 region in NGS data is critical for analyses of B/T-cell stereotypy, clonal evolution, and antibody discovery.[1,2,3,4]
- As CDR3 is the most important region conferring binding activity and specificity of those molecules, it is key to have accurate algorithms to find it.
- We present DELTA-CDR3, a new high-throughput, alignment-based, reference-guided CDR3 identification algorithm included in Invivoscribe RUO LymphoTrack® Enterprise Software (DELTA).
- DELTA-CDR3 provides the amino acid sequence of CDR3 for in-frame rearrangements, missing start or end anchors flags when translation is not possible, as well as indicators for unproductive rearrangements.
- Additionally, it improves on the LT-CDR3, the CDR3 identification algorithm included in the RUO LymphoTrack MiSeq Software v2.4.3, by allowing to analyze greater than 200 most prevalent unique clones.

METHODS

- Genomic DNA from 1136 lymphoproliferative disorder clinical samples (908 clonality, 228 follow-up MRD) was run using LymphoTrack® clonality assays, and sequenced on Illumina MiSeq™ for *IGH* FR1, *IGH* Leader, and *TRG* target genes.
- Sequencing results were analyzed using DELTA and RUO LymphoTrack MiSeq Software v2.4.3 (LT).
- The CDR3 regions for the top 200 unique sequences were extracted from both LT and DELTA outputs, and used to analyze the performance of LT-CDR3 and DELTA-CDR3.
- DELTA-CDR3 results were verified by taking a representative clone sequence for all of the *IGH* FR1, *IGH* Leader and *TRG* V families (7, 7 and 8, respectively), and manually comparing the CDR3 sequence identified by DELTA-CDR3 with the junction sequence identified by IMGT/V-QUEST.

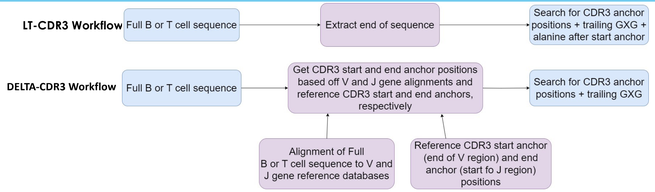

Figure 1: Workflow of LT-CDR3 and DELTA-CDR3 Algorithms

RESULTS

Summary of DELTA-CDR3 Improvements

Table 1: Total Number of CDR3 Sequences Identified in top 200 sequences

Target	LT-CDR3		DELTA-CDR3	
	Count	Percent	Count	Percent
IGH FR1	65542	60.8%	105207	97.6%
IGH Leader	44108	78.1%	54431	96.4%
TRG	19416	34.3%	54007	95.5%
Total	129066	58.4%	213645	98.6%

Table 2: Cases where no CDR3 result was provided by DELTA-CDR3 (3.2% of total sequences across all gene targets)

Target	Missing V or J gene (Impossible to identify CDR3)		Alignment issue	
	Count	Percent	Count	Percent
IGH FR1	2088	81.1%	486	18.9%
IGH Leader	967	47.3%	1076	52.7%
TRG	2507	99.5%	12	0.5%
Total	5562	77.9%	1574	22.1%

Table 3: Runtime Performance

Sequences Analyzed	LT-CDR3	DELTA-CDR3
Top 200	307ms	5ms
Top 10,000	N/A*	64ms

*** LT-CDR3 supports a maximum of 200 sequences for CDR3 analyses**

RESULTS

CDR3 Flag Definitions and Distribution

Flag Definitions:

104CM (**104 Cysteine M**issing or **M**utated): The cysteine preceding the CDR3 region cannot be found at the expected position.

118WM/118FM (**118 Tryptophan/Phenylalanine M**issing or **M**utated): The tryptophan trailing the CDR3 region cannot be found at the expected position.

GXGM (**Glycine-X-Glycine M**issing or **M**utated): The glycine-X-glycine motif immediately following the 118F trailing CDR3 region cannot be found at the expected position.

LGM (**Last Glycine M**issing or **M**utated): The last glycine in the glycine-X-glycine motif immediately following the 118F trailing CDR3 region cannot be found at the expected position.

OOF (**O**ut-**o**f-**F**rame): The CDR3 sequence is not in-frame. This means that either the length of the sequence is divisible by 3, the cysteine at position 104 is not present, and/or the tryptophan/phenylalanine at position 118 is not present.

UNKAA (**Un**known **A**mino **A**cid): The translated version of the CDR3 sequence contains amino acids that are unknown, represented by X in the standard amino acid one-letter code.

SCDR3 (**S**hort **CDR3**): The length of the CDR3 region falls below the minimum experimentally observed (based off experimentally derived (not *in silico*) sequences recorded in IMGT GENE-DB[5,6]) length of CDR3s.

LCDR3 (**L**ong **CDR3**): The length of the CDR3 region falls above the maximum experimentally observed (based off experimentally derived (not *in silico*) sequences recorded in IMGT GENE-DB[5,6]) length of CDR3s.

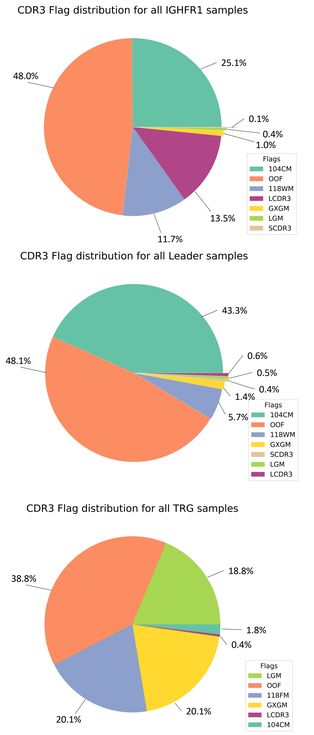

Figure 2: CDR3 Flag distribution for *IGH* FR1, *IGH* Leader, and *TRG* respectively.

COMPARISON OF PERFORMANCE IN TERTIARY ANALYSIS

Table 4: Comparison of top 6* and unique* major IGH CLL CDR3 subsets[3] between DELTA-CDR3 and LT-CDR3 output for 80 IGH FR1 MRD follow-up samples sorted by DELTA-CDR3 frequency

**The top 6 subsets plus 1 additional subset uniquely found by DELTA-CDR3 were chosen from the 19 major CLL subsets[3] due their relevancy*

Number of clone clusters identified for each subset using DELTA-CDR3 (total clusters=12709) and LT-CDR3 (total clusters=3629)

Subset Number[3]	DELTA-CDR3: Ranked 1st by p-value		LT-CDR3: Ranked 1st by p-value		DELTA-CDR3: Ranked 2nd by p-value		LT-CDR3 Ranked 2nd by p-value		CDR3 Pattern[3]
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	
#5	1667	13.1%	494	13.6%	606	4.8%	105	2.9%	ARxxxxxx[AVLI]xxxYYYYxMDx
#28A	489	3.8%	35	1.0%	763	6.0%	92	2.5%	ARxxxGxxYYYYYGMDx
#202	405	3.2%	18	0.5%	80	0.6%	2	0.06%	A[KRH]xxxGxx[AVLI]xxxDx
#1	261	2.1%	15	0.4%	112	0.9%	0	0.0%	ARx[NQ]W[AVLI]xxxxFDx
#31	232	1.8%	148	4.1%	251	2.0%	186	5.1%	ARxxxxxxxxxxxxYYYxMDx
#7C	224	1.8%	88	2.4%	65	0.5%	19	0.5%	AxxxxxxDFW[ST]GYxxxxYYYxxDx
#6**	11	0.09%	0	0.0%	21	0.2%	0	0.0%	ARGGxYDY[AVLI]WGSYRxx[DE][AVLI]FDx
# clusters where no subset p < 0.05	8482	66.7%	2700	74.4%	10023	78.9%	2878	79.3%	N/A

***CLL subset #6 is only found in results generated by DELTA-CDR3*

Background: A tertiary analysis that utilizes the CDR3 region was performed on the results of both LT-CDR3 and DELTA-CDR3 to see the impact of the new algorithm on downstream analysis. A motif analysis was used to determine if shared clone motifs were similar to motif patterns observed in clinical samples. This comparison in the differences between DELTA-CDR3 and LT-CDR3 was performed against a set of 19 stereotyped *IGH* CDR3 subset motifs derived from 7424 clinical chronic lymphocytic leukemia (CLL) samples.[3] The CLL subsets motifs are associated with certain clinical attributes (e.g. age of patient, aggressiveness of disease). Motif matches between the clones and the subsets would indicate that 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). 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.

Tertiary Analysis: Clusters of CDR3 regions were generated using 3 criteria: 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 (mean length 19), using MEME's sites2meme[8] with pseudocounts=0.0001 and background amino acid frequencies derived from analyses of vertebrate polypeptides.[7] Matches between cluster PSSM and CLL subsets were determined with MEME's tomtom,[8] and similarly we compared each cluster PSSM with the PSSM of the random sequences. A match was considered insignificant if its p-value was greater than the p-value of a random sequence or 0.05 (whichever lower), and excluded from cluster motif classification.

Results: DELTA-CDR3's improved CDR3 detection (in terms of raw CDR3 identification) is able to classify more sequences into subsets, generate more specialized clusters (vs the more generalized LT-CDR3), and identify a unique subset (subset #6) that is associated with an enriched **frequency of NOTCH1 mutations**,[4] in comparison to LT-CDR3. This demonstrates that the DELTA-CDR3 algorithm enables more precise downstream (tertiary) classification as compared to LT-CDR3.

CONCLUSIONS

- DELTA-CDR3 interrogates the CDR3 region more effectively than LT-CDR3 by identifying 166% more CDR3s, reducing runtime by 98.2%, and providing flags that add genomic context surrounding rearrangements.
- This genomic context can enable researchers to investigate patterns in mutation in unproductive sequences, and matches CDR3 information provided by the IMGT/V-QUEST reference tool.[5,6]
- The improved DELTA-CDR3 enables greater utility of specific tertiary repertoire analyses, allowing relevant mutated and prognostic CLL subsets to be identified. Further interrogation of these CDR3 sequences could aid in the development of future diagnostic or prognostic indicators for CLL or other lymphoproliferative disorders.[3,4] Identifying more CDR3s provides more data to characterize a sample's immune repertoire and enables tracking CDR3 clonal evolution and stereotypy.
- Above we present an example for CLL where the identification of more CDR3 sequences by DELTA-CDR3 enhances the utility of the tertiary repertoire analysis.

REFERENCES

1. Petrova-Drus K, Syed M, Yu W, Hutt K, Zlotnicki AM, Huang Y, Kamalska-Cyganik M, Maciag L, Wang M, Ma YG, Ho C, Moung C, Yao J, Nafa K, Baik J, Vanderbilt CM, Benhamida JK, Liu Y, Zhu M, Durham B, Ewalt MD, Salazar P, Rijo I, Baldi T, Mato A, Roeker LE, Roshal M, Dogan A, Arcila ME. Clonal Characterization and Somatic Hypermutation Assessment by Next-Generation Sequencing in Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma: A Detailed Description of the Technical Performance, Clinical Utility, and Platform Comparison. *J Mol Diagn*. 2023 Jun;25(6):352-366. doi: 10.1016/j.jmoldx.2023.02.005. Epub 2023 Mar 23. PMID: 36963483; PMCID: PMC10243287.
2. Rustad EH, Misund K, Bernard E, Coward E, Yellapantula VD, Hultcrantz M, Ho C, Kazandjian D, Korde N, Mailankody S, Keats JJ, Akhlaghi T, Viny AD, Mayman DJ, Carroll K, Patel M, Famulare CA, Op Bruinink DH, Hutt K, Jacobsen A, Huang Y, Miller JE, Maura F, Papaemmanuil E, Waage A, Arcila ME, Landgren O. Stability and uniqueness of clonal immunoglobulin CDR3 sequences for MRD tracking in multiple myeloma. *Am J Hematol*. 2019 Dec;94(12):1364-1373. doi: 10.1002/ajh.25641. Epub 2019 Oct 21. PMID: 31571261; PMCID: PMC7449571.
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. Sutton LA, Young E, Baliakas P, Hadzidimitriou A, Moysiadis T, Plevova K, Rosi D, Kminkova J, Stalika E, Pedersen LB, Malcikova J, Agathangelidis A, Davis Z, Mansouri L, Scarfò L, Boudjoghra M, Navarro A, Muggen AF, Yan XJ, Nguyen-Khac F, Larrayoz M, Panagiotidis P, Chiorazzi N, Niemann CU, Belessi C, Campo E, Strefford JC, Langerak AW, Oscier D, Gaidano G, Pospisilova S, Davi F, Ghia P, Stamatopoulos K, Rosenquist R; ERIC, the European Research Initiative on CLL. Different spectra of recurrent gene mutations in subsets of chronic lymphocytic leukemia harboring stereotyped B-cell receptors. *Haematologica*. 2016 Aug;101(8):959-67. doi: 10.3324/haematol.2016.141812. Epub 2016 May 19. PMID: 27198719; PMCID: PMC4967575
5. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D413-22. doi: 10.1093/nar/gku1056. Epub 2014 Nov 5. PMID: 25378316; PMCID: PMC4383898.
6. Manso T, Folch G, Giudicelli V, Jabado-Michaloud J, Kushwaha A, Nguefack Ngoune V, Georga M, Papadaki A, Debbagh C, Pégorier P, Bertignac M, Hadi-Saljoqi S, Chentli I, Cherouali K, Aouinti S, El Hamwi A, Albani A, Elazami Elhassani M, Viart B, Goret A, Tran A, Sanou G, Rollin M, Duroux P, Kossida S. IMGT® databases, related tools and web resources through three main axes of research and development. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D1262-D1272. doi: 10.1093/nar/gkab1136. PMID: 34875068; PMCID: PMC8728119.
7. Beals, M., Gross, L., & Harrell, S. (1999). *Amino Acid Frequency*. University of Tennessee, Knoxville. http://www.nimbios.org/~gross/bioed/webmodules/aminoacid.htm
8. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Res*. 2015 Jul 1;43(W1):W39-49. doi: 10.1093/nar/gkv416. Epub 2015 May 7. PMID: 25953851; PMCID: PMC4489269.

invivoscribe®