Assessment of Tumor Purity in CD19$^+$ Sorted Cells from Waldenström Macroglobulinemia Patients and a Comparison of Data Integration Methods Using Neoplastic and Healthy Cell Annotation

O. Theologi[1] , C. Vlachos[1], F. Aktypi[1], M. Sakkou[1], N. Mavrianou-Koutsoukou[1], C. Liacos[1], A. Papadimou[1], F. Theodorakakou[1], M. A. Dimopoulos[1], E. Kastritis[1], T. Bagratuni[1]

1. Department of Clinical Therapeutics, School of Medicine, National and Kapodistrian University of Athens, Greece

Recent advancements in single-cell RNA sequencing (scRNA-Seq) technologies allow the analysis of thousands of cells, providing valuable insights into the cellular heterogeneity of tumors and the biological characteristics of different cell subpopulations. Normal cells infiltrating the tumor tissue perturb the tumor signal in molecular studies. In this study we use computational approaches in order to assess tumor purity on sc-RNA-seq of CD19$^+$ sorted cells from bone marrow aspirates of seven patients with Waldenström macroglobulinemia (WM). We then evaluate various integration methods in terms of limiting technical variability (batch effects) while preserving true biological heterogeneity of malignant and non-malignant cell populations between samples. Integrating data across different patients is crucial to uncovering shared cell states underlying disease pathogenesis.

We analyzed a total of 40,087 single CD19$^+$ cells: 35,131 from WM patients and 4,956 from two healthy donors (HDs), with an average of 4,454.11 (range 1,956–7,731) cells for each sample. To distinguish the malignant from the non-malignant cell populations, we calculated and visualized the IGKC-fraction (IGKC/(IGKC+IGLC2)) of every cell, and calculated the kappa-to-lambda ratio in each cluster for each patient[1]. Malignant clusters display a consistently even lambda and kappa light chain expression among every cell in the cluster, whereas B cells in healthy clusters express lambda and kappa light chain genes in random different proportions.

This was further validated by exploring somatic large-scale chromosomal copy number alterations (CNVs) inferred by inferCNV[2,3]. We examined the inferred CNV regions in each patient looking for chromosomal regions that are associated with WM (chromosomes 6q and 11q or affecting the following genes[4]: *MYD88, CXCR4, BCL2, TP53, PAX5, ATM*, immunoglobulin heavy and light chain. Our findings further support that the vast majority of patient cells are malignant and that there is heterogeneity within the clonal cells of the patients. We observed higher CNVscores[5,6] in malignant clusters and lower medians of the CNVscores in healthy clusters. We performed an unpaired Wilcox test between the inferCNV scores of HD's healthy cells, patient malignant cells and patient healthy cells, under the Bonferroni correction. Statistically significant differences were found between the aforementioned groups.

We have compared seven single-cell data integration tools[7]: BBKNN[8], Scanorama[9], scGen[10], scVI[11], Harmony[12], Seurat.v3[13] (CCA, RPCA). For the evaluation of data integration performance the median LISI score is used to assess batch (sample), individual cell types and

healthy/neoplastic cells mixing. Seurat.v3-CCA, Harmony, scVI tend to remove batch variation whereas scGen prioritizes biological variance conservation by using cell-type information. While Harmony kept each isolated cell label together, it overlapped these populations. We noticed that the above integration methods are prone to overcorrection. Scanorama is less vulnerable to overcorrection because it finds matches between all pairs of datasets. We chose to merge our datasets using Scanorama, to create a scRNA-seq 'panorama' and to proceed with clustering and trajectory analyses using Harmony's embedding[14]. We stress the importance of benchmarking tools when analyzing scRNA-Seq data for WM, to ensure accurate and reproducible results.

Citations:

1. Roider, T., Seufert, J., Uvarovskii, A. et al. Dissecting intratumour heterogeneity of nodal B-cell lymphomas at the transcriptional, genetic and drug-response levels. Nat Cell Biol 22, 896–906 (2020). https://doi.org/10.1038/s41556-020-0532-x

2. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

3. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014 Jun 20;344(6190):1396-401. doi: 10.1126/science.1254257. Epub 2014 Jun 12. PMID: 24925914; PMCID: PMC4123637.

4. Nguyen-Khac F, Lambert J, Chapiro E, Grelier A, Mould S, Barin C, Daudignon A, Gachard N, Struski S, Henry C, Penther D, Mossafa H, Andrieux J, Eclache V, Bilhou-Nabera C, Luquet I, Terre C, Baranger L, Mugneret F, Chiesa J, Mozziconacci MJ, Callet-Bauchu E, Veronese L, Blons H, Owen R, Lejeune J, Chevret S, Merle-Beral H, Leblondon V; Groupe Français d'Etude de la Leucémie Lymphoïde Chronique et Maladie de Waldenström (GFCLL/MW); Groupe Ouest-Est d'étude des Leucémie Aiguës et Autres Maladies du Sang (GOELAMS); Groupe d'Etude des Lymphomes de l'Adulte (GELA). Chromosomal aberrations and their prognostic value in a series of 174 untreated patients with Waldenström's macroglobulinemia. Haematologica. 2013 Apr;98(4):649-54. doi: 10.3324/haematol.2012.070458. Epub 2012 Oct 12. PMID: 23065509; PMCID: PMC3659998.

5. Sun, H., Fang, T., Wang, T. et al. Single-cell profiles reveal tumor cell heterogeneity and immunosuppressive microenvironment in Waldenström macroglobulinemia. J Transl Med 20, 576 (2022). https://doi.org/10.1186/s12967-022-03798-6

6. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, Liu L, Huang D, Jiang J, Cui GS, Yang Y, Wang W, Guo D, Dai M, Guo J, Zhang T, Liao Q, Liu Y, Zhao YL, Han DL, Zhao Y, Yang YG, Wu W. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019 Sep;29(9):725-738. doi: 10.1038/s41422-019-0195-y. Epub 2019 Jul 4. Erratum in: Cell Res. 2019 Aug 13;: PMID: 31273297; PMCID: PMC6796938.

7. Luecken, M.D., Büttner, M., Chaichoompu, K. et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods 19, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

8. Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, Jong-Eun Park, BBKNN: fast batch alignment of single cell transcriptomes, Bioinformatics, Volume 36, Issue 3, February 2020, Pages 964–965, https://doi.org/10.1093/bioinformatics/btz625

9. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol 37, 685–691 (2019). https://doi.org/10.1038/s41587-019-0113-3

10. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. Nat. Methods 16, 715–721 (2019).

11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053–1058 (2018).

12. Korsunsky I, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P, Raychaudhuri S (2019). "Fast, sensitive, and accurate integration of single cell data with Harmony." bioRxiv. https://doi.org/10.1101/461954.
13. Stuart, T. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21 (2019).
14. Bagratuni T, Aktypi F, Theologi O, Sakkou M, Verrou KM, Mavrianou-Koutsoukou N, Patseas D, Liacos C, Skourti S, Papadimou A, Taouxi K, Theodorakakou F, Kollias G, Sfikakis P, Terpos E, Dimopoulos MA, Kastritis E. Single-cell analysis of MYD88L265P and MYD88WT Waldenström macroglobulinemia patients. Hemasphere. 2024 Feb 8;8(2):e27. doi: 10.1002/hem3.27. PMID: 38435423; PMCID: PMC10878187.

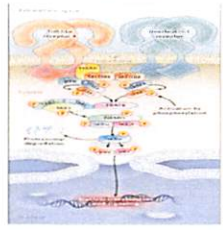
Clinical Records

Patients

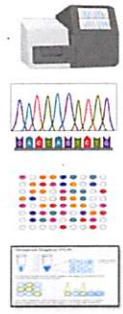
[N=385]

- ✓ BM [N=356]
- ✓ PB [N=22]
- ✓ FFPE [N=7]

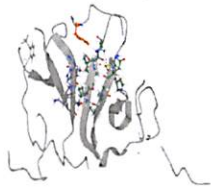
MYD88 mutation
[N=385]



- Sanger sequencing [N=233]
- NGS [N=246]
- AS-PCR [N=322]
- dd-PCR [N=214]



CXCR4 mutation
[N=362]



- Sanger sequencing [N=273]
- NGS [N=246]
- AS-PCR [N=318]