Computational analyses of blood cells: Somatic evolution and morphology

José Guilherme Coelho Peres de Almeida
Darwin College

This thesis is submitted for the degree of
Doctor of Philosophy

European Bioinformatics Institute
University of Cambridge
January 1, 2022

Contents

List of Figures

List of Tables

List of acronyms

AML	acute myeloid leukaemia
ASXL1	ASXL transcriptional regulator 1
CH	clonal haematopoiesis
CNN	convolutional neural network
CV	computer vision
DL	deep learning
DNA	deoxyribonucleic acid
DNMT3A	DNA (cytosine-5)-methyltransferase 3A
mCA	mosaic chromosomal alterations
MDS	myelodysplastic syndrome
MIL	multiple instance learning
MILe-ViCe	multiple instance learning of virtual cell types
ML	machine learning
RBC	red blood cell
RDW	red blood cell distribution width
SF3B1	splicing factor 3b subunit 1
SRSF2	serine and arginine rich splicing factor 2
TET2	tet methylcytosine dioxygenase 2
U2AF1	U2 small nuclear RNA auxiliary factor 1
VAF	variant allele frequency
VCT	virtual cell types
WBC	white blood cell
WBS	whole blood slide

1 Discussion

In this dissertation, I have studied the somatic evolution and cytomorphology of the haematopoietic system. To study the somatic evolution of haematopoiesis, I used longitudinal targeted sequencing data from over 300 elderly individuals over a median period of 13 years and 7 single-cell colony derived phylogenies to track CH clones and quantify their driver-specific and lifelong growth patterns. Using simulations, I developed and validated statistical methods that quantify the clonal dynamics using these types of data and applied them, revealing lifelong clonal growth behaviours. Changes to blood cytomorphology were studied using a cohort of over 300 whole blood slide (WBS) from individuals with either splicing factor 3b subunit 1 (*SF3B1*)-mutant myelodysplastic syndrome (MDS), other subtypes of MDS, iron deficiency anaemia or megaloblastic anaemia and case controls using computer vision (CV) methods. I developed methods to detect and characterize blood cells from WBS and showed how the morphometric characterizations of cells obtained using these methods can be used not only to predict clinical conditions, but also to define cytomorphology-disease associations that reveal novel morphological cellular archetypes. Below I discuss the main conclusions and limitations of both works and how they can be further extended and improved to enable their clinical application.

1.1 The longitudinal dynamics and natural history of clonal haematopoiesis

In CH, different drivers are associated with distinct growth rates during old age: while DNA (cytosine-5)-methyltransferase 3A (*DNMT3A*) clones grow on average at 5% per year, others, such as those characterized by U2 small nuclear RNA auxiliary factor 1 (*U2AF1*) mutations and serine and arginine rich splicing factor 2 (*SRSF2*)-P95H, grow 6 and 10 times faster, respectively. While constant during old age, this growth rate is not representative of the lifelong behaviour of clones: factors such as an increasingly competitive oligoclonal landscape lead to clonal deceleration, a behaviour detected in the vast majority of expansions detected in the single-cell colony-derived phylogenies of 7 individuals. While this phenomenon is likely to affect all clones, the driver-specific effect it may exert on clonal dynamics is hard to assert from this study — the estimation of driver-specific dynamics was performed using longitudinal targeted sequencing during old age, which does not permit the study of lifelong alterations to clonal growth. Most mutations, including those in *DNMT3A* and tet methylcytosine dioxygenase 2 (*TET2*) (the two most commonly mutated genes in CH) appear consistently through life. However, mutations in *U2AF1* and *SRSF2*-P95H (two genes associated with splicing) are restricted to old age in CH. Finally, there is a clear association between the gene-specific clonal growth rates and acute myeloid leukaemia (AML) onset risk, and between the site-specific clonal growth rates and positive selection in both MDS and AML.

1.1.1 CH and its role in early detection

It is a well known fact that CH is associated with progression to haematological malignancies and death [1–3]. However, translating this knowledge to the clinic is complicated — while there is a clear association between CH growth dynamics and risk of progression to AML and selection in AML and MDS, most cases of CH will not be associated with transformation to a myeloid malignancy, making it necessary to better understand this process. By showing here how specific drivers in CH are not only associated with relatively well-conserved dynamics, I also demonstrate that mutations associated with an increased risk of progression grow at a much faster rate in healthy individuals. This knowledge can potentially reveal which mutations should be of clinical concern and whether they should be tracked through time. However, several studies, including the work presented here, have highlighted how little is actually known about the prevalence and behaviour of CH, with large clonal expansions having no known driver but similar dynamics to other CH expansions with known drivers [4–6]. Moreover, recent evidence shows the high prevalence of mosaic chromosomal alterations (mCA) and copy number alterations in CH — these can act as confounders, especially considering that some mutations are associated with copy number alterations that directly affect the gene [7, 8].

Closing the gap on CH knowledge. Future studies into CH need to focus much more on the discovery of novel variants, creating a more complete picture of CH before effective and economically-feasible tracking of most CH variants becomes a reality; for this, approaches such as those applied by Genovese *et al.* and Jaiswal *et al.*, where blood whole-exome sequences were obtained for very large cohorts and used to identify blood clones in healthy individuals [1, 2], can be adapted by increasing the sequencing depth to further reveal recurring drivers of CH. Additionally, studies which use single-cell colonies to explore CH, such as the work presented here and that by Mitchell *et al.* can be extremely useful — not only do they allow the practical exploration of the lifelong trajectories of different expansions, they can allow, given sufficient magnitude, for the tentative identification of novel drivers of CH.

However, the cost of deriving a sufficient number of single-cell colonies can act as a considerable obstacle. Poon *et al.* overcame the problem of data availability by using large cohorts where synonymous mutations could be used to track clonal expansions — by doing so, and under the assumption that synonymous mutations can only expand by being passengers in clonal expansions, they were able to identify novel putative drivers. This knowledge can be used to create large panels of potential drivers which can then be further validated. With this knowledge, single-cell colonies could be used to establish the lifelong dynamics of putative drivers. Moreover, through the long-term storage of viable deoxyribonucleic acid (DNA) samples (which constitute relatively small and easily storable aliquots), it could be possible to retrospectively reconstruct the variant allele frequency (VAF) trajectories of clones, similarly to what was done in Chapters 2 and 3, and to better understand the oncogenic trajectories of different mutations and how they evolve as disease onset approached.

Finally, it is worth noting that the collection of clinical data — such as blood counts and behavioural patterns — can help illuminate the mechanisms underlying specific expansions or even stratify individuals regarding the CH risk and risk of progression [3, 9]. Particularly, smoking has been shown to be associated with ASXL transcriptional regulator 1 (*ASXL1*) CH [9], while increased red blood cell distribution width (RDW) is associated with increased risk of AML [3]. Other blood counts have also shown specific associations with CH [2, 3, 9–13], further showing how CH changes haemaotopoiesis. However, the specific mechanisms through which CH these changes are enacted is hard to elucidate through large human sequencing cohorts. For this, studies which combine animal models of CH with expression and epigenetic analyses can be used to reveal how mutations lead to specific changes in haematopoiesis [14].

1.2 The cytomorphology of myelodysplastic syndromes

The use of CV methods to analyse WBS can detect and characterize between thousands and hundreds of thousands of cells as demonstrated in Chapter 4. These vast collections of cells can then be used to predict clinical conditions with morphometric moments and through the use of multiple instance learning (MIL) as shown in Chapter 5. Morphometric moments are individual-specific parametric characterizations of blood cell morphometry. While simple and useful, morphometric moments do not permit a visually intuitive characterization of blood cells. For this reason, I developed multiple instance learning of virtual cell types (MILeViCe), a MIL method that clusters cells into visually coherent and clinically-relevant virtual cell types (VCT) and uses their relative proportions to predict diseases. These VCT reveal novel cellular archetypes of different conditions such as anaemias and MDS and its different subtypes. For instance, red blood cell (RBC) are larger in *SF3B1*-mutant MDS than in other

MDS subtypes, whereas poikilocytic RBC are more prevalent in other MDS subtypes. Additionally, *SF3B1*-mutant MDS has a larger relative prevalence of hypolobulated neutrophils when compared with other *SF3B1* subtypes, whereas hypersegmented neutrophils, prevalent in both megaloblastic anaemia and iron deficiency anaemia, are larger in the megaloblastic anaemia. Finally, I show how the predictive performance of most tasks generalizes to an external validation cohort. However, when considering the prediction of whether a WBS belongs to an individual with either MDS or anaemia, MILe-ViCe underperforms. I posit that this is likely to be due to artefacts potentially introduced by the preparation and digitalisation of WBS and show how a smaller subset of features, uncorrelated with dataset of origin, can alleviate the drop in the external validation performance.

1.2.1 Computational cytomorphology — a potential path to the clinic

Different clinically-relevant conditions can be predicted from WBS using only computational methods and, through the use of MILe-ViCe, new cellular archetypes may be discovered in situations where little expert annotation is required. However, clinical application requires these methods to be applicable to a much larger number of clinical conditions, implying the necessity for much larger cohorts of digitalised and diagnosed WBS. Additionally, generalization to other cohorts, digitalized with a different scanner and prepared under different standards, is still unsatisfactory — the most logical path towards a possible solution for this lack of generalization is the use of multi-centre training cohorts. Additionally, regarding the characterization of blood cells, improvements can be made — large databases of annotated cells can enable the extraction of morphologically-meaningful features with no need for feature design. This can be achieved by training classification convolutional neural network (CNN) models and using the feature vectors produced by these models as morphological descriptors of different cells, or by training auto-encoder models to characterize the morphology of white blood cell (WBC) and RBC.

The relevance of multi-centre cohorts. Machine learning (ML) and deep learning (DL) systems enable the high-throughput and reproducible analysis of WBS in a clinical context. However, particular hurdles are still in the way — the absence of very large, multi-centre cohorts impedes the realistic assessment of predictive power. Moreover, most studies are retrospective (using available and previously stored data), which is not representative of point-of-care where predictions are expected on a regular basis for new slides [15]. Large multi-centre cohorts also enable a more realistic learning of the true disease signal, which should be independent of cohort-specific noise. An additional factor to consider is that diagnostic conventions change — how these prediction algorithms change is also worthy of consideration: if an augmented intelligence setting is considered, where the predictive algorithm is merely an assistant to the expert which is expected to confirm the diagnoses, a reinforcement learning strategy should be considered, where the algorithm is iteratively refined by a team of experts. Federated learning, a decentralized ML paradigm that permits guarantees data-privacy and better data governance [16], could be combined with large multi-centre databases and reinforcement learning to create a powerful semi-automated diagnostic system capable of constant updates. However, if diagnostic criteria change, new training cohorts are possibly necessary — in the worst case scenario, a disease subtype may be stratified further, requiring expert annotation or confirmation (it may also be possible that a classification algorithm will already, implicitly, cluster disease subtypes within its feature space). This last circumstance would require the retraining of these models.

Digitaliser agnostic models? Finally, it is worth considering whether generalization to other digitalisers should be an objective of these models. Companies working on the automated analysis of WBS focus, for now, exclusively on the development of protocols which work exclusively with their digitalizers. These methods perform differential counting of WBC and generic cellular detection in WBS with some abnormality detection capabilities by leveraging large expert-annotated collections of WBC [17, 18]. Combined with MILe-ViCe, which enables the definition of novel and clinically-relevant cellular archetypes, the robust cellular detection and characterization capability of these systems can be not only expanded but used for cell type discovery. It is also worth noting that the work presented in Chapters 4 and 5 shows how RBC morphology, generally not considered by these systems, can be useful to establish novel cytomorphology-disease.

Some automated systems for WBS analysis incorporate the automatic preparation of WBS from blood samples [18]. This eliminates most of the variability associated with slide preparation and digitalization. In realistic terms, a relatively simple solution — training models which are optimized for detecting and characterizing cells in specific systems of slide preparation and digitalization and testing them as diagnosis and prognosis tools prospectively — will provide the strongest evidence for their generalization as a clinically applicable and transferrable solution. Models for disease prediction from WBS — such as those developed and presented in Chapter 5, whose underperformance in an external validation cohort is likely to be associated with WBS preparation and digitalisation — become much more applicable in these circumstances. Nonetheless, it should be noted that specific sources of noise, associated with the specific composition of different populations, may still be a problem for these models — it has been shown that incidence and 5-year survival of blood cancers, or the incidence of specific haemoglobinopathies (such as thalassemia, a condition characterized by decreased or defective haemoglobin production), differs by ethnicity and race [19, 20]. The aetiology of such differences is unknown, with factors such as diet, socio-economic conditions and biology being possible causes, and the changes they elicit on the cytomorphology of blood are far from being studied. Approaches such as the one presented here, where morphometric trends and cell types are established in association to specific conditions, could help reveal the morphological signatures of different ethnicities and races, possibly creating better informed diagnostic practices by defining novel patient strata which are associated with clinical outcome in specific diseases.

Bibliography

1. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. en. *N. Engl. J. Med.* **371,** 2477–2487 (Dec. 2014).
2. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. en. *N. Engl. J. Med.* **371,** 2488–2498 (Dec. 2014).
3. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559,** 400–404 (July 2018).
4. Poon, G. Y. P., Watson, C. J., Fisher, D. S. & Blundell, J. R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. en. *Nat. Genet.* **53,** 1597–1605 (Nov. 2021).
5. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *bioRxiv* (2021).
6. Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. en. *bioRxiv,* 2021.08.12.455048 (Aug. 2021).
7. Gao, T. *et al.* Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. en. *Nat. Commun.* **12,** 338 (Jan. 2021).
8. Saiki, R. *et al.* Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis. en. *Nat. Med.* **27,** 1239–1249 (July 2021).
9. Dawoud, A. A. Z., Tapper, W. J. & Cross, N. C. P. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. en. *Leukemia* **34,** 2660–2672 (Oct. 2020).
10. Bick, A. G., Weinstock, J. S., Nandakumar, S. K., Fulco, C. P., *et al.* Inherited causes of clonal hematopoiesis of indeterminate potential in TOPMed whole genomes. *BioRxiv* (2019).
11. Cordua, S. *et al.* Prevalence and phenotypes of JAK2 V617F and calreticulin mutations in a Danish general population. en. *Blood* **134,** 469–479 (Aug. 2019).
12. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. en. *Blood* **130,** 742–752 (Aug. 2017).
13. Zehir, A. *et al.* Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell* **21,** 374–382.e4 (2017).
14. Asada, S. & Kitamura, T. Clonal hematopoiesis and associated diseases: A review of recent findings. en. *Cancer Sci.* **112,** 3962–3971 (Oct. 2021).
15. Eckardt, J.-N., Bornhäuser, M., Wendt, K. & Middeke, J. M. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. en. *Blood Adv* **4,** 6077–6085 (Dec. 2020).
16. Rieke, N. *et al.* The future of digital health with federated learning. en. *NPJ Digit Med* **3,** 119 (Sept. 2020).

17. *CellaVision DM1200 Digital Cell Morphology System* en. https://www.siemens-healthineers.com/hematology/systems/cellavision-dm1200-digital-cell-morphology-system. Accessed: 2021-11-30.

18. *ADVIA 120 Hematology* en. https://www.siemens-healthineers.com/hematology/systems/advia-120-hematology-system. Accessed: 2021-11-30.

19. Kirtane, K. & Lee, S. J. Racial and ethnic disparities in hematologic malignancies. en. *Blood* **130,** 1699–1705 (Oct. 2017).

20. Lorey, F. W., Arnopp, J. & Cunningham, G. C. Distribution of hemoglobinopathy variants by ethnicity in a multiethnic state. en. *Genet. Epidemiol.* **13,** 501–512 (1996).