Rheumatoid arthritis

RNA sequencing and machine learning as molecular scalpels

Myles Lewis and Michael R. Barnes

Standfirst

The application of new technologies including RNA sequencing and machine learning to the analysis of synovial tissue is yielding new insights into the pathology of rheumatoid arthritis, with potential implications for the clinical management of the disease.

Refers to Orange, D. E. et al. Machine learning integration of rheumatoid arthritis synovial histology and RNAseq data identifies three disease subtypes. *Arthritis Rheumatol.* https://doi.org/10.1002/art.40428 (2018) | Stephenson, W. et al. Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nature communications* **9**, 791 (2018)

Main text

Our knowledge of rheumatoid arthritis (RA) pathology has arguably been outstripped by our ability to treat symptoms with biologic drugs, which empirically target key mediators of disease pathology. However, 20-40% of patients remain unresponsive to biologics[1], highlighting our incomplete mechanistic understanding of RA. Several new RNA sequencing (RNA-seq) studies[2-5] from 2018 are offering new insights in RA that complement disease understanding founded on conventional histopathology. In two such studies, Orange et al.[2] developed a machine learning algorithm that correlated RNA-seq and histopathology findings, whereas Stephenson et al.[3] investigated the synovium at the single cell level using single cell RNA-seq (scRNA-seq) facilitated by an intriguing novel 3D-printed microfluidic device. As well as revealing new information regarding the underlying cellular processes of RA, these new methods illustrate the unique clinical insights that 'omics' data can provide.

Machine learning offers an opportunity to integrate detailed knowledge of disease processes with omics data. To set up a machine learning model, Orange et al.[2] assessed the histological features of synovial tissue obtained from patients undergoing joint replacement surgery (123 patients with RA and 6 patients with osteoarthritis (OA)). Robust histological features were selected on the basis of features that were most common and could be consistently identified by different pathologists. Tissue RNA-seq was performed on a subset of 45 synovial samples (39 RA and 6 OA), and consensus clustering suggested the existence of three expression subtypes: low or mixed inflammatory subtypes observed in both RA and OA individuals and a high inflammatory subtype exclusively found in RA individuals. The investigators used the selected histology features to predict the expression

subtype of the remaining 78 individuals using a form of machine learning, known as a support vector machine (SVM), which was trained on the 45 sequenced samples.

The findings from the developed SVM model brought some new insights; for example, the most informative predictor of the high inflammatory subtype was the presence of plasma cells, whereas pain scores correlated with C-reactive protein in patients with the high inflammatory subtype, but not in patients with the low inflammatory subtype, implying the existence of different pain drivers between the high and low inflammatory subtypes. However, it should be noted that SVM training is prone to overfitting (that is, it may use random noise in the data in an over complex and non-reproducible manner) in datasets of small sample sizes and despite the use of cross-validation training, the study lacked external validation. Additionally, the patients had an average disease duration of 14 years and so probably had received past treatment with multiple therapies (indeed, 53% of the patients received a biologic agent before surgery), which might have confounded findings. Orange and colleagues proposed that their SVM algorithm could be used in place of gene expression biomarkers to identify the three inflammatory subtypes and would be cheaper and more widely applicable than using omics platforms. This step seems counter to the prevailing 'omics-revolution' (that is, the drive to use omics platforms in clinical practice), but might offer a middle road to bring omics insights into clinical environments where omics technologies are unavailable.

In an alternative approach for improving the accessibility of omics technologies, Stephenson et al.[3] developed a low-cost, 3D-printed microfluidic instrument that can perform single-cell transcriptome profiling. With this technology, the investigators presented an intriguing molecular view of the synovium by disaggregating synovial tissue from 5 patients with RA, again obtained following joint surgery to sequence 20,387 single cells. On average, 29,651 reads were generated per cell, enabling on average, the detection of 2315 different mRNA species per cell (~10% of the transcriptome). The clustering of individual cellular transcriptomes enabled the detection of 13 distinct clusters that apparently represented distinct autoimmune cell infiltrate and stromal cell populations, including new and distinct fibroblast populations. Although the dataset is probably not the comprehensive 'cell atlas' of the RA synovial tissue that the investigators claim, the methodology could certainly be applied to create such a resource.

Technologically, the method by Stephenson and colleagues is excellent. The manufacturing protocols are provided in an open protocol repository and should enable others to construct microfluidic devices using widely available 3D printer technology at a much lower cost than sourcing commercial alternatives. Although preliminary, this study brings potentially important insight into the structure of synovial tissues by the identification of three transcriptionally distinct fibroblast subpopulations, which could be distinguished in terms of their CD55 and CD90 surface expression: $CD55^+$ fibroblasts (called 'type 1' in this study), which were located in the intimal lining by histology, and $CD90^+$ fibroblasts, which were further subdivided into type 2a and 2b and were located in the synovial sublining.

Although tissue RNA-seq arguably offers only incremental advantage over microarray studies, at increased cost, Stephenson et al. demonstrate the potentially technologically disruptive power of scRNA-seq and the reproducibility between re-aggregated scRNA data and tissue RNA-seq. sc-RNA-seq can unveil the unique picture of the transitions between cell states that characterise the cellular immune response. Furthermore, this technology can differentiate gene expression dynamics that are

masked in bulk, population-averaged measurements.[6] Intriguingly, scRNA-seq is also sensitive to genome variation, enabling the detection of discontinuous transcription, or 'bursting', thus enabling the measurement of both the frequency of such bursts of expression and the magnitude of expression in a manner that is undetectable in tissue.[6]

Even though microfluidic approaches remain prototypic, a technical risk of these approaches, considering the limited stability of RNA, includes RNA degradation and cell lysis, which could compromise cellular transcriptome analysis and jeopardise finite patient biopsy material needed to accurately study RA pathophysiology. As microfluidic instrumentation continues to develop, along with scRNA analysis methodology, we anticipate that the single cell approach could become the method of choice for transcriptomic analysis. However, the currently limited consensus on the technical and analytical processes required for such analysis[7] support a continuing focus on bulk tissue material, perhaps with complementary scRNA-seq studies in the same samples.

In order to understand the complex interplay between infiltrating immune cells and disease-altered stromal cells in inflamed synovium, future studies will need to build on these pioneering studies in many areas (Figure 1). Although the studies by Orange et al. and Stephenson et al. understandably focused on synovial tissue obtained during joint replacement surgery, such tissue samples are unrepresentative of RA pathology, particularly the clinically important early stages of the disease. Performing synovial biopsies in treatment-naive patients will be critical for distinguishing inflammation in reaction to ongoing joint damage from inflammatory mechanisms driving disease processes. Biopsies performed prior to treatment with steroids, DMARDs and biologics might reveal stronger signals that can differentiate true disease endotypes, just as biopsies for classifying lymphoma are best performed prior to steroid treatment.[8]

In terms of measuring disease activity, obtaining accurate clinical data that incorporates ultrasound and radiographic imaging will enable a more objective estimation of disease burden compared with disease activity measures that rely on subjective patient measures such as pain scores and tender joint count. The integration of data from multiple omics platforms will provide increased depth of knowledge of the disease process in RA. However, with the massive expansion in data, more refined, standardised, machine learning algorithms that are suited to heterogeneous, high-dimensional and statistically noisy datasets will be required. The lack of consensus surrounding the best approaches for analysing data remains a barrier. Furthermore, true validation of biomarkers identified by multi-omics studies will require carefully collected replication cohorts as a standard practice.

Myles Lewis[1] and Michael R. Barnes[2]*

[1]Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London and Barts Health NHS Trust, London, UK

[2]Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London and Barts Health NHS Trust, London, UK

* email: m.r.barnes@qmul.ac.uk

1. Nam, J.L. et al. Efficacy of biological disease-modifying antirheumatic drugs: a systematic literature review informing the 2016 update of the EULAR recommendations for the management of rheumatoid arthritis. *Annals of the Rheumatic Diseases* **76**, 1113-1136 (2017).
2. Orange, D.E. et al. Machine learning integration of rheumatoid arthritis synovial histology and RNAseq data identifies three disease subtypes. *Arthritis Rheumatol* (2018).
3. Stephenson, W. et al. Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nat Commun* **9**, 791 (2018).
4. Mizoguchi, F. et al. Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat Commun* **9**, 789 (2018).
5. Mandelin, A.M., 2nd et al. Transcriptional Profiling of Synovial Macrophages using Minimally Invasive Ultrasound-Guided Synovial Biopsies in Rheumatoid Arthritis. *Arthritis Rheumatol* (2018).
6. Wills, Q.F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* **31**, 748-52 (2013).
7. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5** (2016).
8. Borenstein, S.H., Gerstle, T., Malkin, D., Thorner, P. & Filler, R.M. The effects of prebiopsy corticosteroid treatment on the diagnosis of mediastinal lymphoma. *Journal of Pediatric Surgery* **35**, 973-6 (2000).

Competing interest

The authors declare no competing interests.

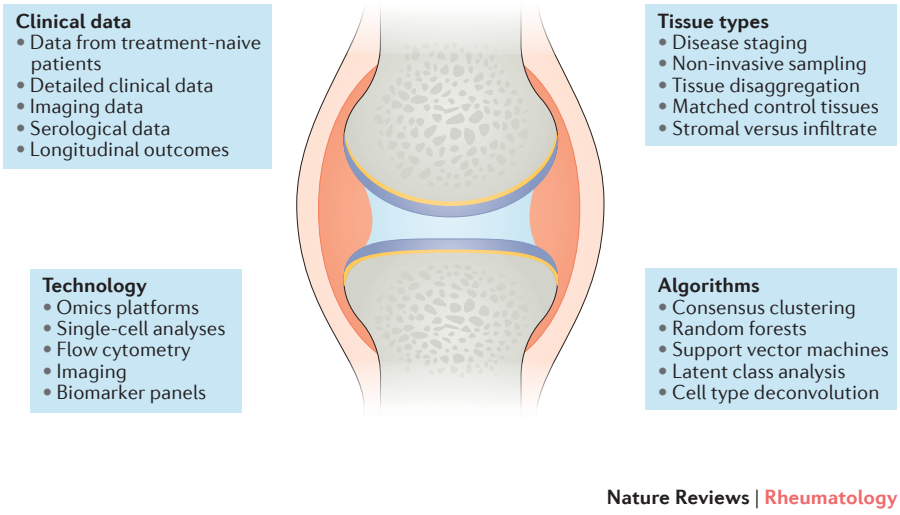

Figure 1. What is needed to gain translational RA disease insight? We highlight key thematic areas (Clinical, Tissue, Technology and Algorithm) that need to be considered in experimental planning to enable biomarker discovery using omic technologies.

Online only

Competing interests

The authors declare no competing interests.

Subject ontology terms

Health sciences / Rheumatology / Rheumatic diseases / Rheumatoid arthritis
[URI /692/4023/1670/498]

Health sciences / Medical research / Translational research
[URI /692/308/575]

Health sciences / Pathogenesis
[URI /692/420]

Health sciences / Health care / Diagnosis / Laboratory techniques and procedures
[URI /692/700/139/1420]