BASIC: BCR assembly from single cells

Stefan Canzar[1,†], Karlynn E. Neu[2,†], Patrick C. Wilson[2], and Aly A. Khan[1,*]

[1] Toyota Technological Institute at Chicago, Chicago IL 60637, USA
[2] Committee on Immunology, The Knapp Center of Lupus and Immunology Research, The University of Chicago, Chicago IL 60637, USA

Background B cells form an important component of the adaptive immune system. They possess the remarkable capacity to recognize antigens through the B-cell receptor (BCR, Figure 1A), which is generated through a series of somatic rearrangements and mutations [1][2]. Recent advances in single cell RNA-sequencing (scRNA-seq) offer a high-throughput means of profiling all transcripts expressed in a single B cell. However, the assembly of full-length BCR sequences from scRNA-seq is a non-trivial problem that neither current reference-based assembly methods nor *de novo* assembly methods address. Thus, the lack of efficient methods for assembling BCR sequences is a major roadblock in studying B-cell biology at a single cell level.

Results Here, we present a novel semi-*de novo* assembly method to determine the full-length sequence of the BCR in single B cells from scRNA-seq data, called BASIC (BCR assembly from single cells). Briefly, BASIC performs semi-*de novo* assembly in two stages (Figure 1B). First, BASIC uses known variable and constant regions in both chains to identify anchor sequences. Second, BASIC performs *de novo* assembly to stitch together the anchor sequences. To demonstrate the utility and accuracy of our method, we subjected single B cells from a human donor to scRNA-seq, assembled the full-length heavy and the light chains, and experimentally confirmed these results by using single cell primer based nested PCRs and Sanger sequencing. Importantly, errors in Sanger sequencing, where specific nucleotides are unresolved and reported as N, were resolved by BASIC and match known germline sequences. Furthermore, we compared our method with a state-of-the-art *de novo* transcript assembly program and report better accuracy for BCR assembly with BASIC (see Figure 1C for an example). In sum, BASIC correctly assembles full-length BCR sequences and demonstrated better performance when compared to a state-of-the-art *de novo* transcript assembly method.

Conclusion BASIC enables investigators to assemble BCR sequences from scRNA-seq data and study B-cell repertoire. We experimentally validated sequences assembled by BASIC, and show it to be robust to potential noise associated with different PCR pre-amplification cycles. The algorithm underlying

[†] Equal contribution
[*] Corresponding author: aakhan@ttic.edu

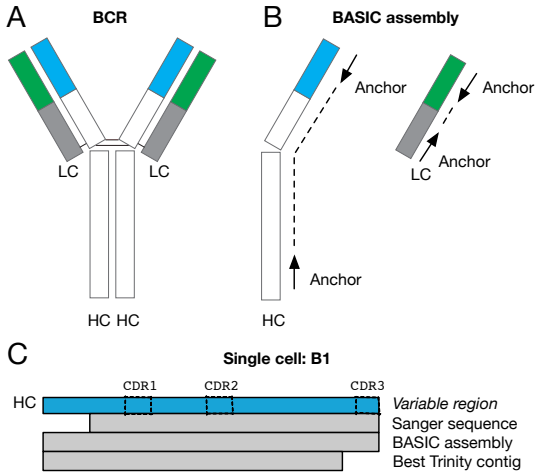

Fig. 1. A) The BCR is a large 'Y' shaped protein complex composed of two identical heavy chains (HC) and two identical light chains (LC). The variable regions are colored blue and green in the paired chains. The complementarity determining regions (CDRs) are those parts of the variable regions that participate in the binding of antigens. B) Anchors are stitched together to assemble the HC and LC. C) Illustration of the HC variable region sequence for single cell B1 along with: the Sanger sequence, the BASIC assembled sequence, and the best contig reported by Trinity. Note the absence of CDR3 from the Trinity contig, and the BASIC assembly extending past the 5' PCR primer site used in the Sanger sequence.

BASIC also serves as a principled approach to assemble other diverse genes associated with immunological repertoire using scRNA-seq, such as HLA and TCR genes. BASIC is available at: http://ttic.uchicago.edu/∼aakhan/BASIC

References

1. Davies, D.R., Padlan, E.A., Segal, D.: Three-dimensional structure of immunoglobulins. Annual review of biochemistry 44(1), 639–667 (1975)
2. Edelman, G.M.: Dissociation of γ-globulin. Journal of the American Chemical Society 81(12), 3155–3156 (1959)