BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 9 2014, pages 1329–1330
doi:10.1093/bioinformatics/btt770

Systems biology

Advance Access publication January 8, 2014

Enhanced flowType/RchyOptimyx: a Bioconductor pipeline for discovery in high-dimensional cytometry data

Kieran O'Neill[1,2,†], Adrin Jalali[1,2,3,†], Nima Aghaeepour[1,2,†], Holger Hoos[4] and Ryan R. Brinkman[1,5,*]

[1]Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC V5Z 1L3, Canada, [2]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, [3]Max-Planck-Institut fur Informatik, Saarland University, 66123, Saarbrucken, Germany, [4]Departments of Computer Science and [5]Departments of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

Associate Editor: Martin Bishop

ABSTRACT

Summary: We present a significantly improved version of the flowType and RchyOptimyx BioConductor-based pipeline that is both 14 times faster and can accommodate multiple levels of biomarker expression for up to 96 markers. With these improvements, the pipeline is positioned to be an integral part of data analysis for high-throughput experiments on high-dimensional single-cell assay platforms, including flow cytometry, mass cytometry and single-cell RT-qPCR.

Availability: FlowType and RchyOptimyx are distributed under the Artistic 2.0 license through Bioconductor.

Contact: rbrinkman@bccrc.ca

Received on October 7, 2013; revised on November 25, 2013; accepted on December 11, 2013

1 INTRODUCTION

Flow cytometry has undergone a 'chromatic explosion' over the past decade and can now measure 17 markers at once for each of hundreds of thousands of individual cells (Chattopadhyay *et al.*, 2008). Since then, mass cytometry has enabled measurement of 30–45 markers/cell (Bendall *et al.*, 2012), whereas single-cell multiplexed RT-qPCR can measure 50–96 messenger RNAs/cell (White *et al.*, 2011). The growth in high-throughput single-cell data continues to outpace development of corresponding bioinformatics techniques (Chattopadhyay *et al.*, 2008). To answer this challenge, we previously developed flowType (Aghaeepour *et al.*, 2012a) and RchyOptimyx (Aghaeepour *et al.*, 2012b). FlowType uses partitioning of cells, either manually or by clustering, into positive or negative for each marker to enumerate all cell types in a sample, e.g. Aghaeepour *et al.* (2013). RchyOptimyx measures the importance of these cell types by correlating their abundance to external outcomes, such as disease state or patient survival, and distills the identified phenotypes to their simplest possible form. These packages have been used to identify several novel cell populations correlated with HIV outcome (Aghaeepour *et al.*, 2012a). More recently, this pipeline has been used to evaluate standardized immunological panels (Villanova *et al.*, 2013), to optimize lymphoma diagnosis (Craig *et al.*, 2013) and to analyze a range of other clinical data (unpublished data).

However, the higher dimensionality of data produced by mass cytometry generates up to $3^{45} \approx 10^{21}$ possible cell types, with an even greater number (up to $3^{96} \approx 10^{45}$) for single-cell qPCR; these magnitudes are beyond the capabilities of flowType and RchyOptimyx. Furthermore, flowType and RchyOptimyx have thus far only treated cells as being either positive or negative for a marker. In practice, many biomarkers can have a range of expression levels such as 'dim' and 'bright'. In this application note, we detail architectural improvements to flowType and RchyOptimyx to overcome these limitations.

2 APPROACH

Our primary challenge was to enable flowType to generate a number of cell types tractable on most common workstations (e.g. those with 4–12 GB of RAM). We hereafter denote the original flowType implementation as flowType-BF (brute force) and the new version as flowType-DP (dynamic programming). Whereas flowType-BF completely enumerates all cell types over all $[1, ..., m]$ markers, we opted in flowType-DP to use a breadth-first strategy of enumerating all cell types defined over a subset of $k \leq m$ markers. We provide a memory use estimation function to assist users in finding a k that fits within the limits of their hardware. To improve computation time in flowType-DP, we implemented a dynamic programming approach, which exploits the fact that cell types can be arranged into a hierarchy, and membership of any given cell type over n markers is equal to the intersection of one of its parent types (over $n-1$ markers) with a single-marker cell type. FlowType-DP first enumerates all cell types involving only one marker by simple partitioning and then iterates over $2, ..., k$ markers, computing all cell types for each level n by set intersections between corresponding cell types in levels $n-1$ and 1.

For example, membership of the cell type $CD45^{++}CD117^{+}CD34^{-}$ is computed as follows:

$$\begin{aligned}&\{CD45^{++}CD117^{+}CD34^{-}\}\\ &= \{CD45^{++}CD117^{+}\} \cap \{CD34^{-}\}\\ &= \{CD45^{++}\} \cap \{CD117^{+}\} \cap \{CD34^{-}\}\end{aligned}$$

To allow partitioning into levels other than positive and negative, we used a string representation for cell types. The string has

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

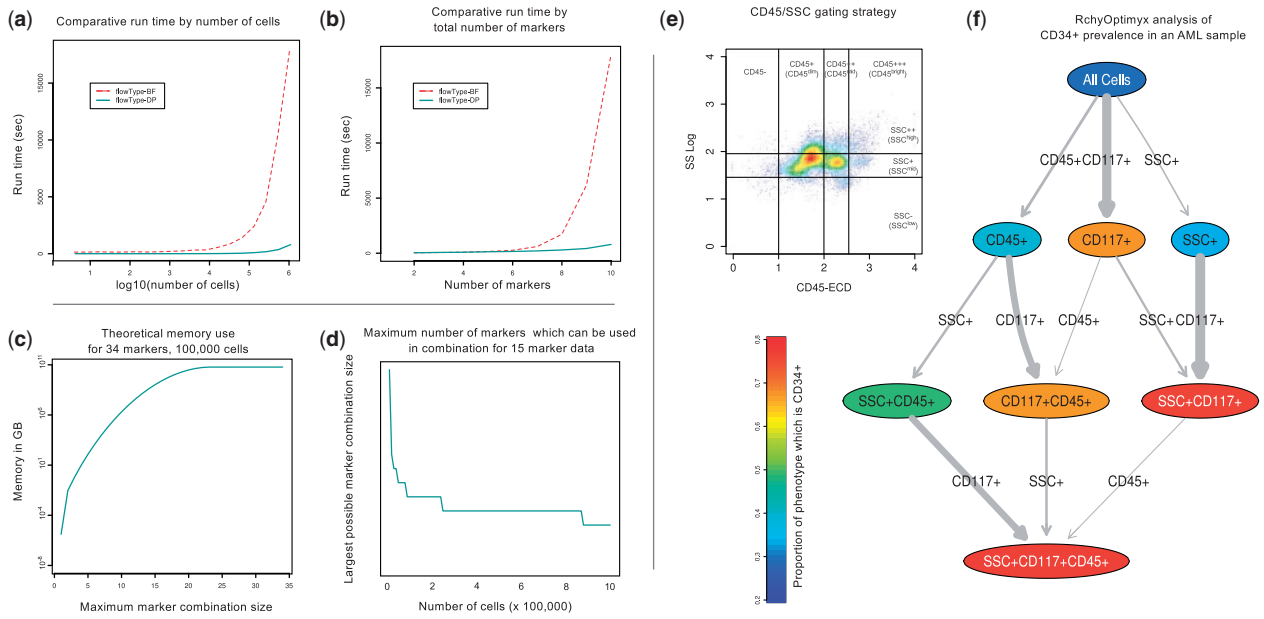

Fig. 1. (**a** and **b**) Run time comparison of flowType-DP to flowType-BF in terms of number of cells (a) and number of markers (b). (**c** and **d**) Possible thresholds for marker combinations using flowType-DP for typical mass cytometry data (c) and polychromatic flow cytometry data (d). (**e** and **f**) Three/four partition flowType-generated RchyOptimyx-visualized cell type hierarchy on a bone marrow sample from a patient with AML. Cell population identification strategy used for SSC and CD45, with the CD34-enriched subset highlighted (e). RchyOptimyx analysis showing CD34 enrichment (f)

one integer character for every marker, denoting the partition, or zero if the marker is not used. Values $1, ..., n$ denote partitions 1 to n. For example, if the set of markers were {CD3, CD45, CD13, CD117, CD34}, the cell type $CD45^{++}CD117^{+}CD34^{-}$ would be represented by `03021`. RchyOptimyx uses a dynamic programing algorithm for efficiently constructing k-shortest paths (Eppstein, 1998). We modified RchyOptimyx' graph construction component to be able to handle more than one partition per marker.

3 RESULTS AND DISCUSSION

We evaluated flowType-DP against flowType-BF on a 10-marker dataset available from Flow Repository (ID FR-FCM-ZZZK) (Aghaeepour *et al*., 2012a). FlowType-DP showed a substantial speedup over flowType-BF, which increases exponentially with the number of cells and markers. For example, at 10^6 cells and 10 markers, flowType-DP is 14 times faster (see Fig. 1a and b). Comparison on larger datasets was not possible due to the limitations of flowType-BF.

We also computed the limits for k on a hypothetical machine with 12 GB of RAM for samples representative of mass cytometry (Fig. 1c) and polychromatic flow cytometry (Fig. 1d), both of which would be intractable for flowType-BF. FlowType and RchyOptimyx are now able, within the memory of a common workstation (12 GB), to analyze 34-marker data.

Finally, to demonstrate the importance of several partitions per marker, we applied flowType and RchyOptimyx to an acute myeloid leukemia (AML) sample from Flow Repository (ID FR-FCM-ZZYA) (Fig. 1e and f). CD34 is a stem cell marker typically expressed on AML blast cells. These blasts are also known to have dimly positive CD45 expression and low side scatter (SSC) (Vial and Lacombe, 2001). By partitioning CD45 and SSC into four and three partitions and naively running flowType and RchyOptimyx to search for CD34-enriched cell types, we were able to find that the $SSC^{low}CD45^{dim}$ cell type had a high proportion of $CD34^{+}$ cells, as expected. This would not have been possible with only two partitions for each of CD45 and SSC.

Funding: ISAC scholar program, CIHR/MSFHR scholarship for strategic training in bioinformatics, UBC 4YF scholarship, NIH/NIBIB (EB008400), Canadian Cancer Society (700374), the Terry Fox Research Institute and the Terry Fox Foundation.

Conflict of Interest: none declared.

REFERENCES

Aghaeepour,N. *et al.* (2012a) Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. *Bioinformatics*, **28**, 1009–1016.

Aghaeepour,N. *et al.* (2012b) RchyOptimyx: cellular hierarchy optimization for flow cytometry. *Cytometry A*, **81**, 1022–1030.

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Bendall,S.C. *et al.* (2012) A deep profiler's guide to cytometry. *Trends Immunol.*, **33**, 323–332.

Chattopadhyay,P.K. *et al.* (2008) A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology*, **125**, 441.

Craig,F. *et al.* (2013) Computational analysis optimizes the flow cytometric evaluation for lymphoma. *Cytometry B Clin. Cytom.*, [Epub ahead of print, doi: 10.1002/cytob.21115].

Eppstein,D. (1998) Finding the k shortest paths. *SIAM J. Comput.*, **28**, 652–673.

Vial,J.P. and Lacombe,F. (2001) Immunophenotyping of acute leukemia: utility of CD45 for blast cell identification. *Methods Cell Biol.*, **64**, 343–358.

Villanova,F. *et al.* (2013) Integration of lyoplate based flow cytometry and computational analysis for standardized immunological biomarker discovery. *PLoS One*, **8**.

White,A.K. *et al.* (2011) High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl Acad. Sci. USA*, **108**.