

Bioinformatic Approaches in Flow Cytometry

Ryan Brinkman

Department of Medical Genetics, University of British Columbia
Terry Fox Laboratory, British Columbia Cancer Agency

American Thoracic Society

*PG13: FLOW CYTOMETRY IN HUMAN AND MOUSE LUNGS:
FROM EXPERIMENTAL DESIGN TO ANALYSIS*

May 20, 2017

Disclosure

Grants/Research Support: Genocea, Cytobank, De Novo Software, Bristol-Myers Squibb, Providence Healthcare, Celgene, Anixa

Ownership Interest: Cytapex Bioinformatics Inc.

Why Automated Analysis of Flow Cytometry Data?

Automated algorithms for flow cytometry data analysis have reached a level of maturity that enables them to match and in many cases exceed the results produced by human experts.¹

Supervised gating (diagnosis): OpenCyto, flowDensity²

Unsupervised biomarker discovery: FlowReMi, flowType/RchyOptimyx³

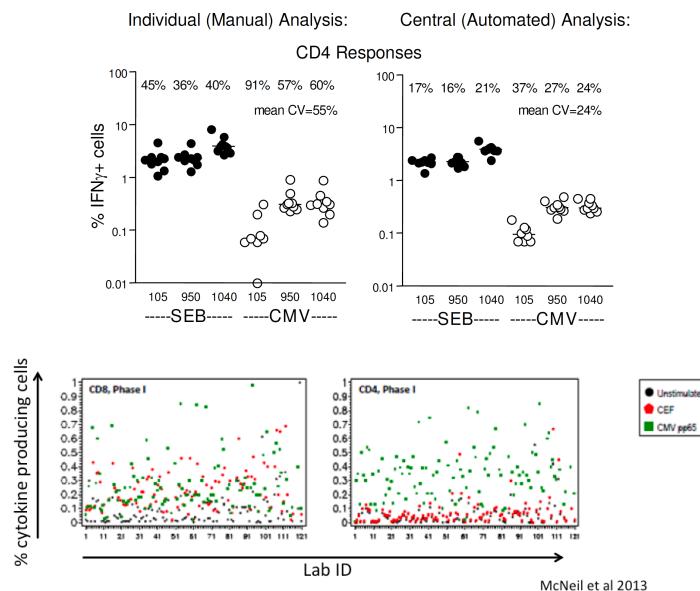
blue text is hyperlinked to primary literature

¹[Aghaeepour et al., Nature Methods \(2013\)](#)

²[Finak et al., Nature Scientific Reports \(2016\)](#)

³[Aghaeepour et al., Cytometry Part A \(2016\)](#)

Most Variation in Cross-center Studies is Due to Gating



Maecker et al. *BMC Immunology*, 2005

McNeil et al. *Cytometry*, 2013

>50 peer-reviewed, free, open source software tools

- ▶ 45 R/BioConductor for data analysis
 - ▶ A **scripted** approach to high throughput data analysis
 - ▶ Non-interactive, self-documented, reproducible
 - ▶ Breaks problem into smaller pieces (packages)
 - ▶ Modules can plug-in & swap-out
 - ▶ Collaborative, cross-platform development environment
- ▶ 9 additional software tools
 - ▶ Java, Python, Matlab, C++
 - ▶ Stand alone (single problem/solution)



Kvistborg *et al.*, *Immunity*. (2016)

Which Automated Analysis Methods to Use ?

The image displays a grid of scientific publications from various journals, each highlighting a different method for automated flow cytometry analysis:

- Cytometry PART A**: Rapid Cell Population Identification in Flow Cytometry Data. (Methodology Article)
- Cytometry Part B (Clinical Cytometry)**: Misty Mountain clustering: application to fast unsupervised flow cytometry gating. (Open Access)
- PNAS**: Automated high-dimensional flow cytometric data analysis. (Research Article)
- BMC Bioinformatics**: Elucidation of Seventeen Human Peripheral Blood B-Cell Subsets and Quantification of the Tetanus Response Using a Density-Based Method for the Automated Identification of Cell Populations in Multidimensional Flow Cytometry Data. (Open Access)
- Cytometry**: Merging Mixture Components for Cell Population Identification in Flow Cytometry. (Methodology Article)
- BMC Bioinformatics**: Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering. (Open Access)
- BMC Bioinformatics**: The curvHDR method for gating flow cytometry samples. (Methodology Article)

Each publication includes a brief description of its methodology or findings, such as 'flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding' and 'The curvHDR method for gating flow cytometry samples'.

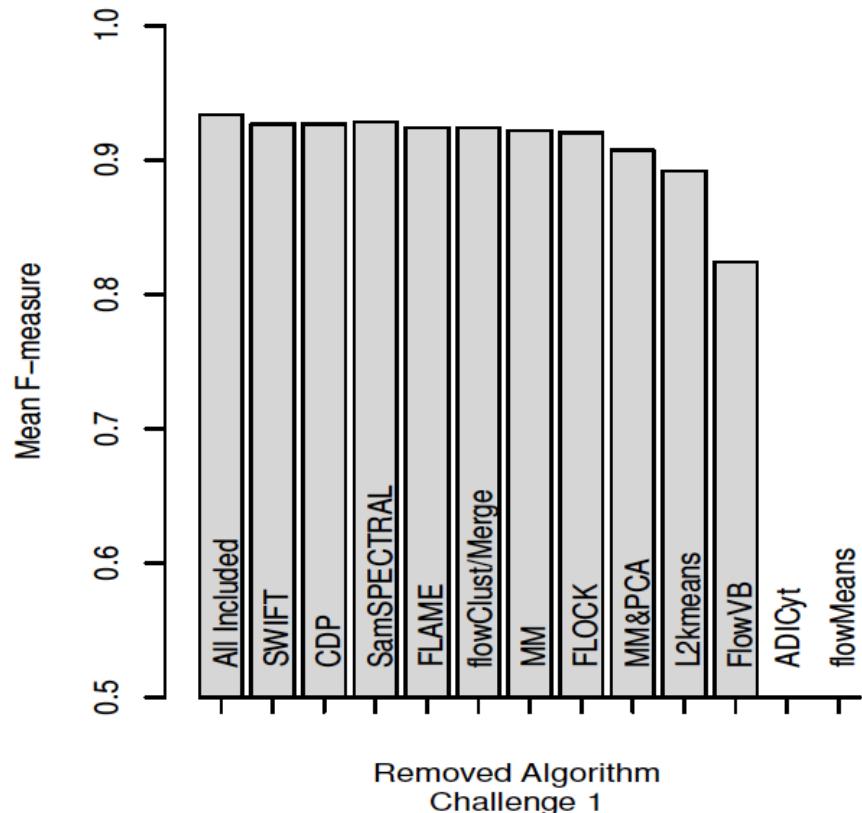
II: FlowCAP (Critical Assessment of Automated Analysis Methods)

Community-based evaluation of flow bioinformatics tools

- FlowCAP-I (2010): Matching manual gating (1st gen tools)
- FlowCAP-II (2011): (Too simple) sample classification
- FlowCAP-III (2012): 2nd generation automated gating tools
- FlowCAP-IV (2014): (Hard) Biomarker discovery



There is No Single Best Gating Solution

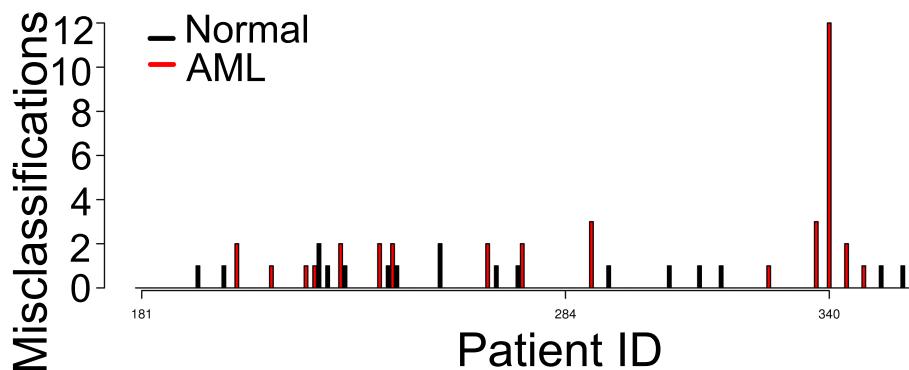


FlowCAP-II: Tools for Clinical Classification

Several algorithms performed perfectly

- 8 tubes of 5 colour assays on 359 subjects;
 - 43 Acute Myeloid Leukaemia vs. 316 healthy donor

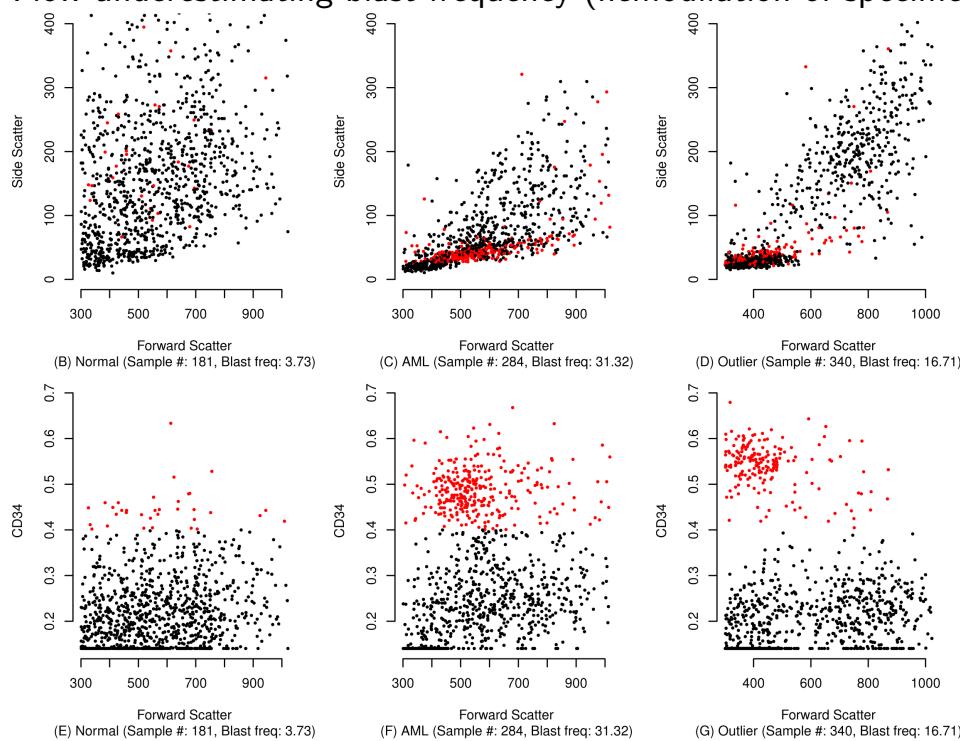
	Sensitivity	Specificity	Accuracy
flowType-FeaLect	1.00	1.00	1.00
flowPeakssvm	1.00	1.00	1.00
SPADE	1.00	1.00	1.00
... n=43



FlowCAP-II: AML Outlier

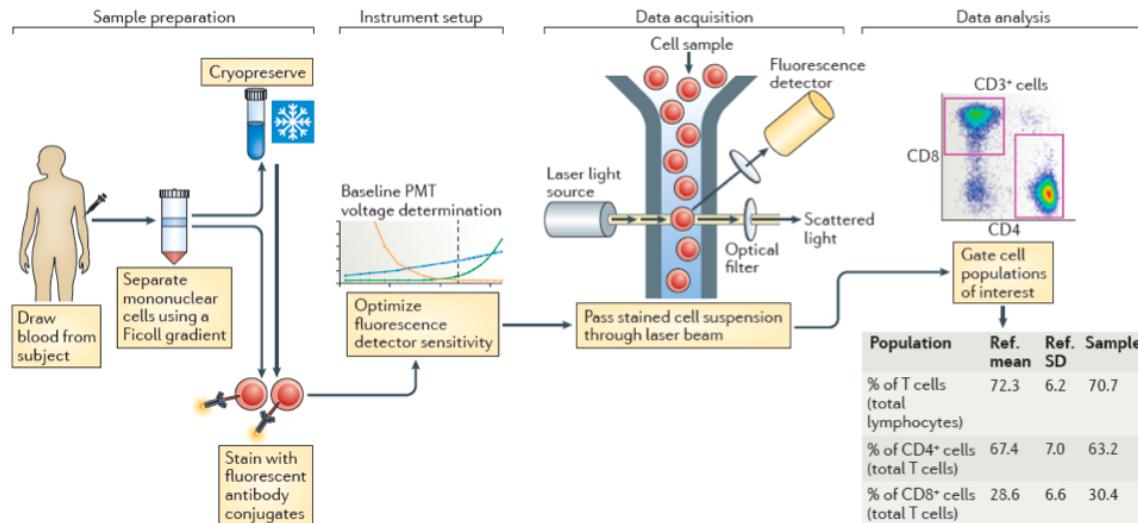
More algorithms is better

High grade myelodysplasia?
 Flow underestimating blast frequency (hemodilution of specimen)?



FlowCAP-III: Reducing Variability in Translational Immunology

Sample prep (BD lyoplates), Instrumentation, Acquisition, Automated Analysis

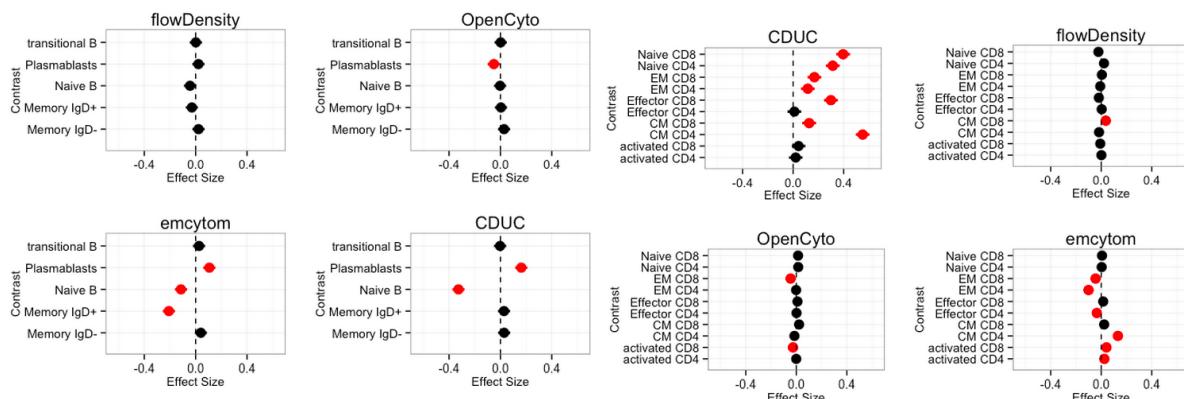


Maecker et al. *Nature Reviews Immunology*, 2012

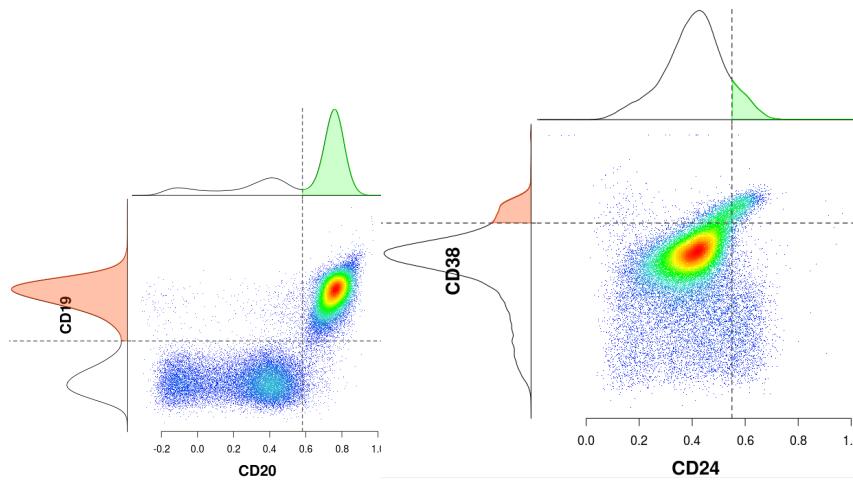
FlowCAP-III: Supervised Analysis for “Diagnostics”

2 automated tools can match human gating

- 9 clinical sites, 4 replicates of cryopreserved cells per site.
- Centralized gating of data based on a consensus best approach.
- Automated algorithms vs. centralized gating.



flowDensity



FlowCAP-III conclusion: Stop manually gating
Supervised gating can match manual analysis when it is data-driven

- ▶ Automated gating is unbiased relative to manual gating
- ▶ Variability is as low or lower than manual gating
- ▶ Even when biased, the bias is associated with populations that have low cell counts and CV is lower than manual gating
- ▶ Not following SOPs can result in large variability

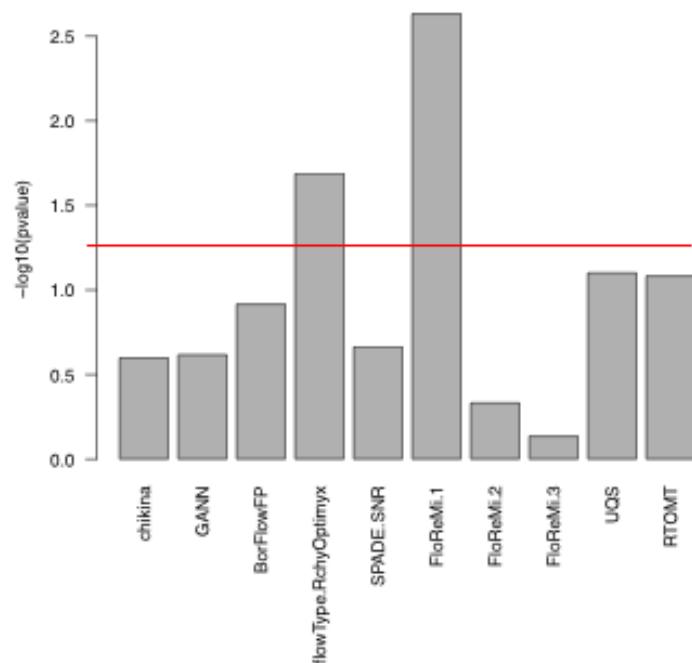
*Finak *et al.*, *Nature Scientific Reports* (2016)

FlowCAP-IV: Unsupervised clustering and classification Biomarker discovery

- ▶ 388 patients (split training and test)
- ▶ 14 parameter data
- ▶ Predict survival time (onset of HIV) & identify biomarkers
- ▶ Thorough manual analysis (NIH/VRC) had failed to identify any biomarker

*Aghaeepour *et al.*, *Cytometry A* (2016)

FlowCAP-IV: Unsupervised Clustering and Classification
Two similar methods (included flowDensity/flowType) had significant results on test data



*p-value not adjusted for testing multiple cell populations, but not algorithms

III: flowType/RchyOptimyx for discovery

Immunologic and Virologic Events in Early HIV Infection Predict Subsequent Rate of Progression

Anuradha Ganesan,^{1,a} Pratip K. Chattopadhyay,^{2,a} Tess M. Brodie,² Jing Qin,³ Wenjuan Gu,⁴ John R. Mascola,² Nelson L. Michael,⁵ Dean A. Follmann,³ and Mario Roederer,² for the Infectious Disease Clinical Research Program HIV Working Group^b

¹National Naval Medical Center, Infectious Disease Clinical Research Program, Uniformed Services University, ²Vaccine Research Center, National Institute of Allergy and Infectious Diseases, and ³Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, ⁴Biostatistics Research Branch, Scientific Application International Corporation-Frederick, Frederick, and ⁵United States Military HIV Research Program, Walter Reed Army Institute of Research, Rockville, Maryland

United States Military HIV Natural History Study

- ▶ PBMCs of 466 HIV⁺ personnel and beneficiaries from Army, Navy, Marines, and Air Force.
- ▶ 13 surface markers and Ki-67 (cell proliferation).
- ▶ Clinical Data: Survival times including 135 events^a

^aAn event is defined as progression to AIDS or initiation of HAART.

Manual Gating Results

- ▶ Frequency of **long-lived Memory Cells (CD127⁺)** has a positive correlation.
- ▶ Frequency of **cells with high proliferation (Ki-67⁺)** has a negative correlation.
- ▶ Can we find what they have found? Can we find more?

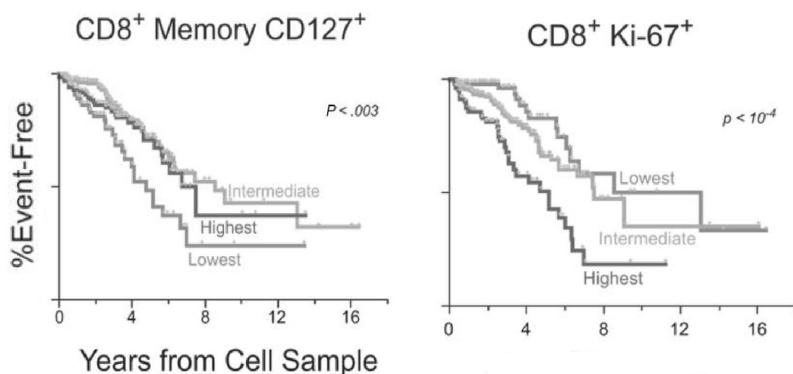
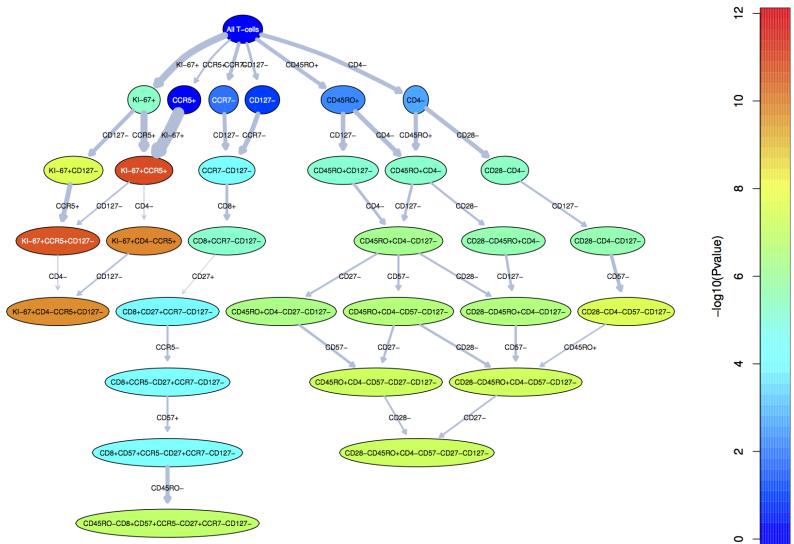


Figure from Ganesan et. al., JID, 2011.

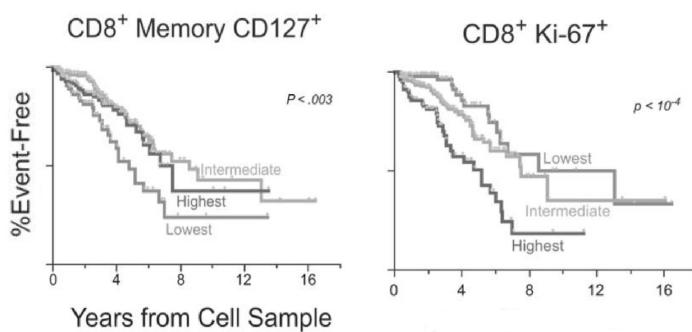
Summarizing 466 patient/ 16 parameter dataset in 1 figure

Annotate a large number of cell populations ID'd by other methods (e.g., manual gating, SPADE, flowType) in terms of importance.

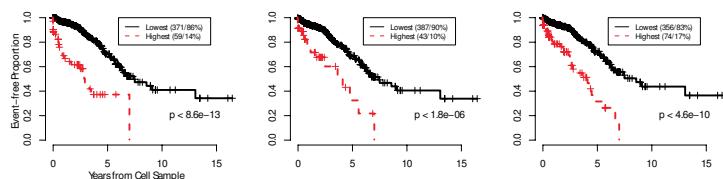


Colour is p-value; arrow thickness is it's change
O'Neill *et al.*, *Bioinformatics*, 2014

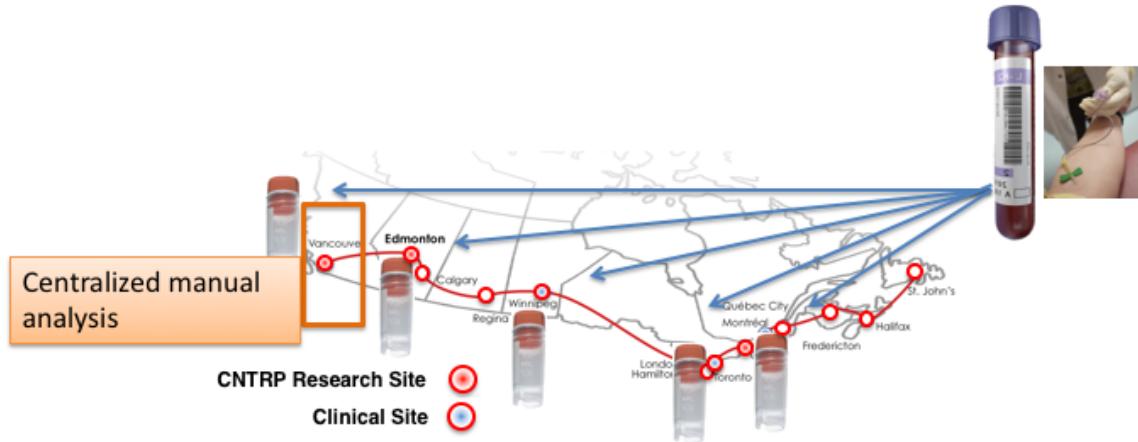
Manual analysis:



Computational analysis:

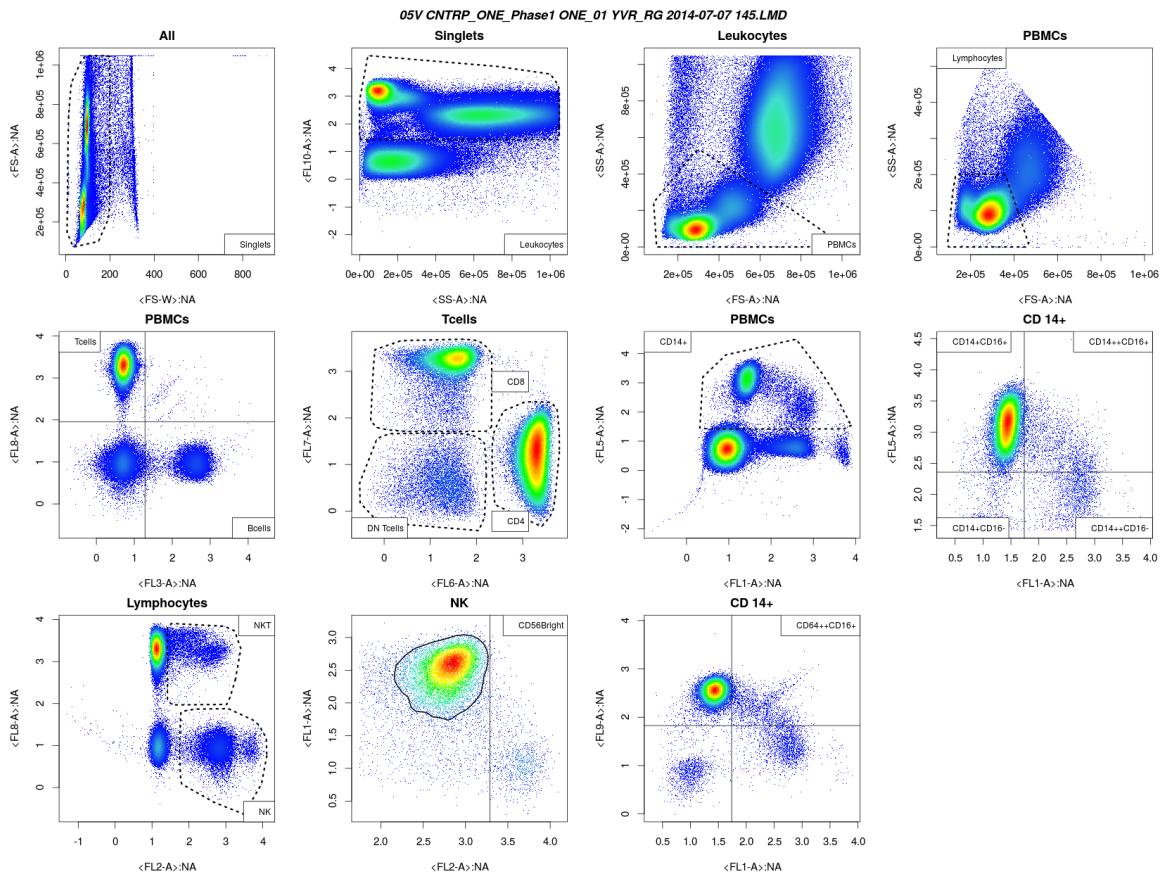


IV: flowDensity for “Diagnosis” Using Beckman Coulter Duraclone panels (ONE Study*)

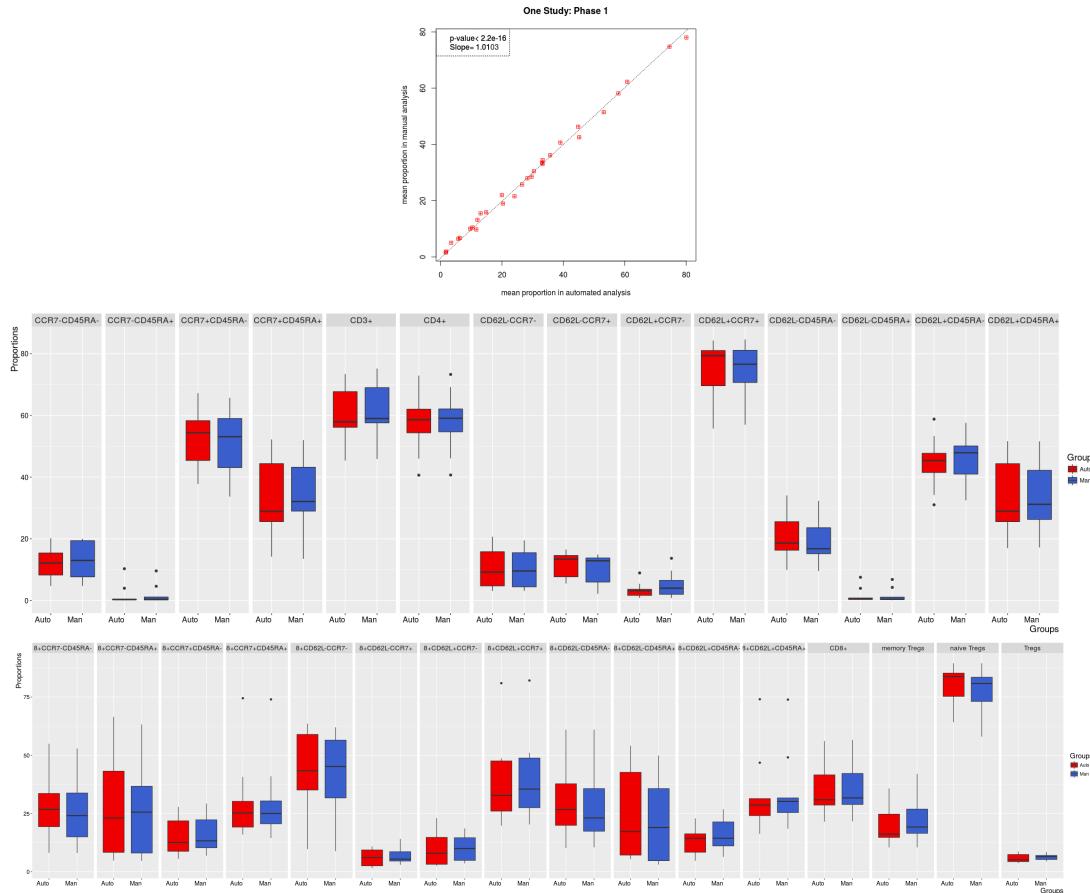


*www.onestudy.org
Streitz et al., *Transplant Res.* 2013.

Automated gating - Panel 1



Automated vs. manual comparison (healthy donors): Panel-4



V: Putting it all together in real life International Mouse Phenotyping Consortium

Input:

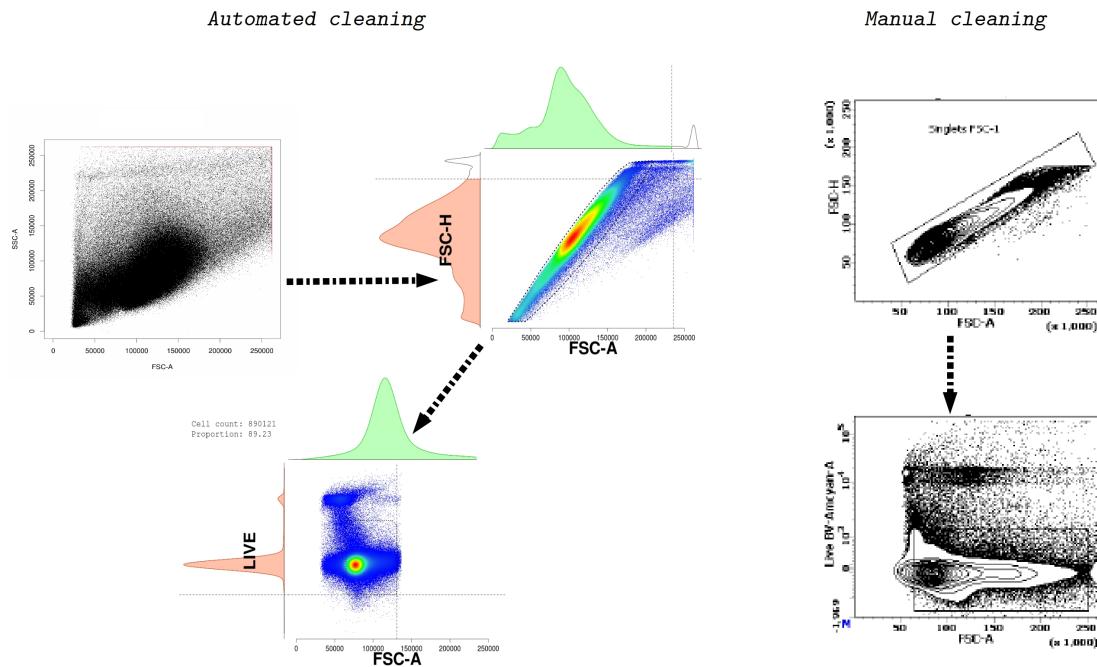
- ▶ 20,000 mouse knockouts (KO) from 9 facilities
- ▶ 77,000 8-12 colour FCS files using 2 panels

Output:

- ▶ Supervised automated gating of 72 cell populations of interest
 - ▶ Lower CVs than manual analysis
- ▶ Unsupervised analysis to phenodeviants
 - ▶ ID all (>59,000) cell populations
 - ▶ ID all biomarkers that could have been ID'd by manual analysis
 - ▶ ID biomarkers that would have been missed by manual analysis
 - ▶ No false positives

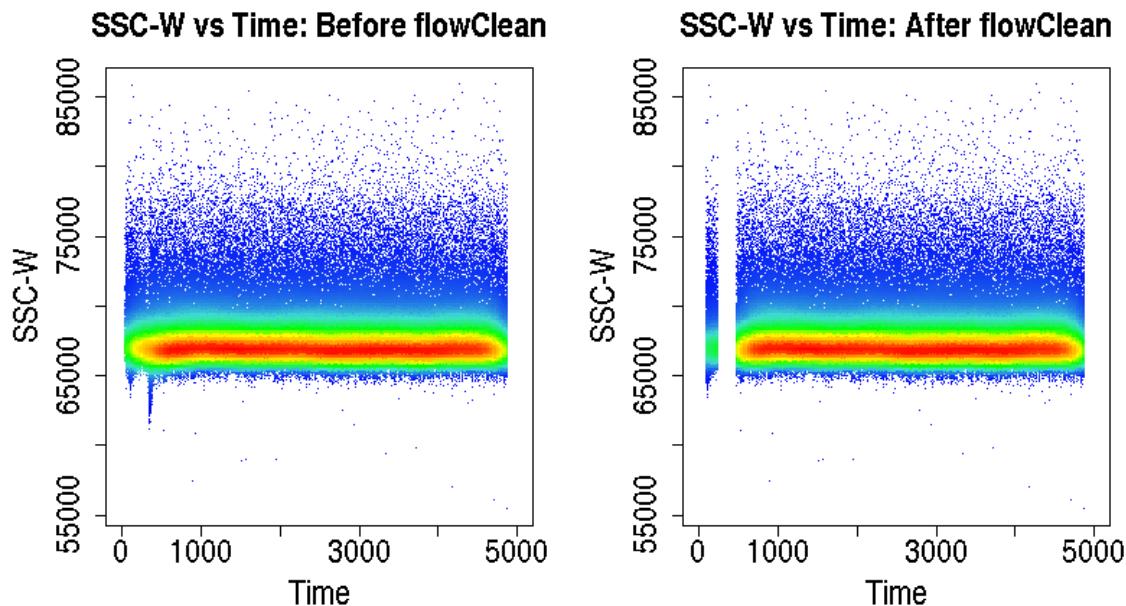
Step 1. Preprocessing

Remove dead cells and doublets



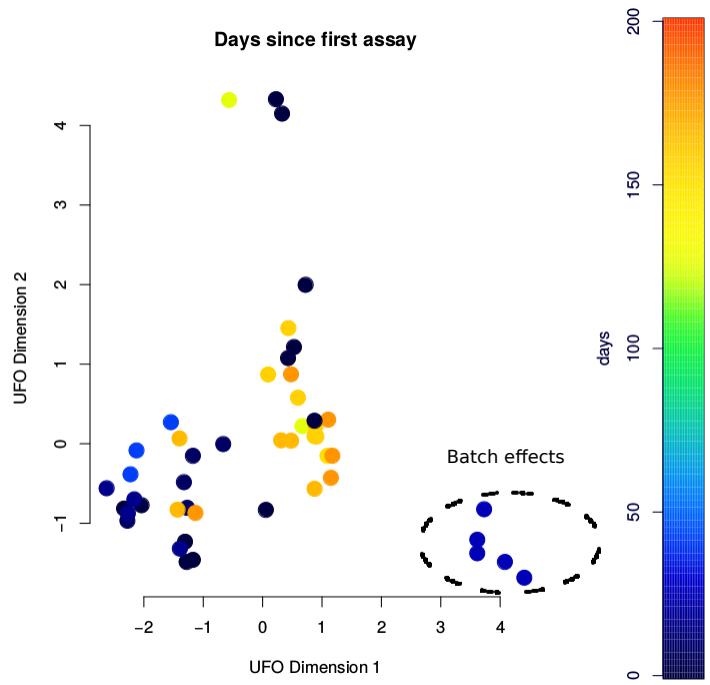
Step 2. QC

flowClean, flowAI, flowCut remove spurious events (e.g., flow cell blockage)



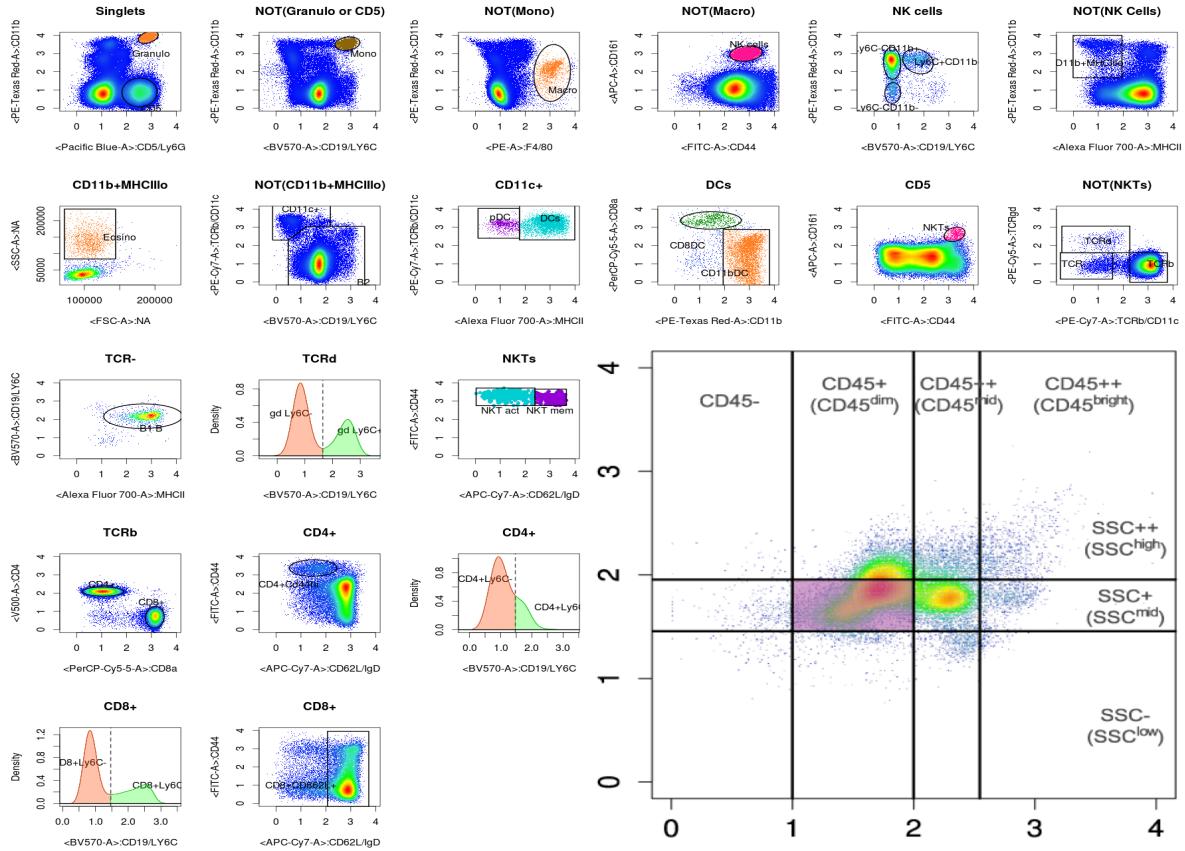
Step 3. ID Batch Effects

UFO IDs for batch effects via dimension reduction

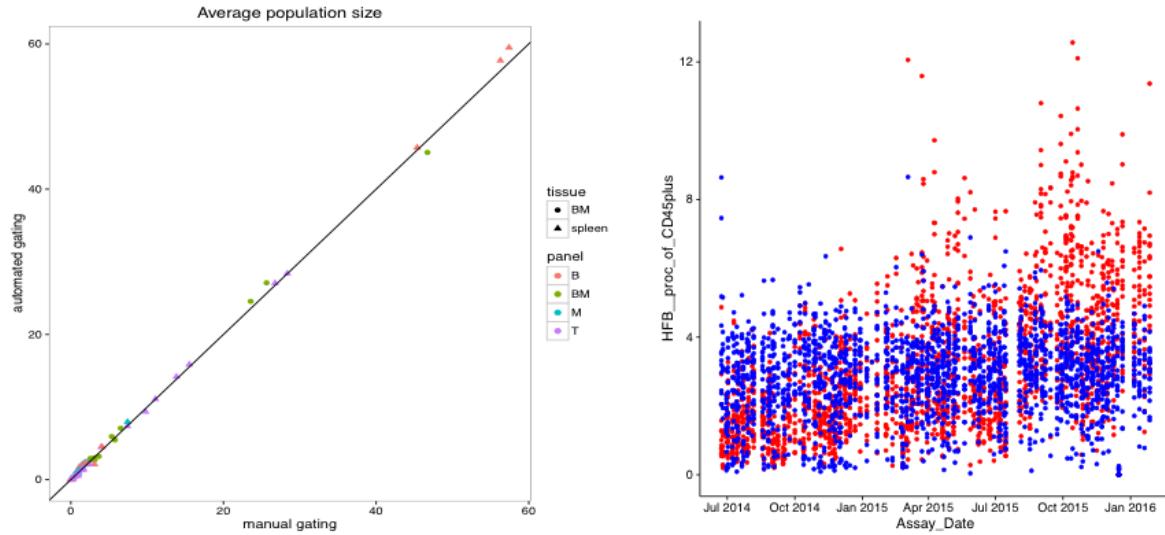


*Courtesy of Nima Aghaeepour.

Step 4. flowDensity

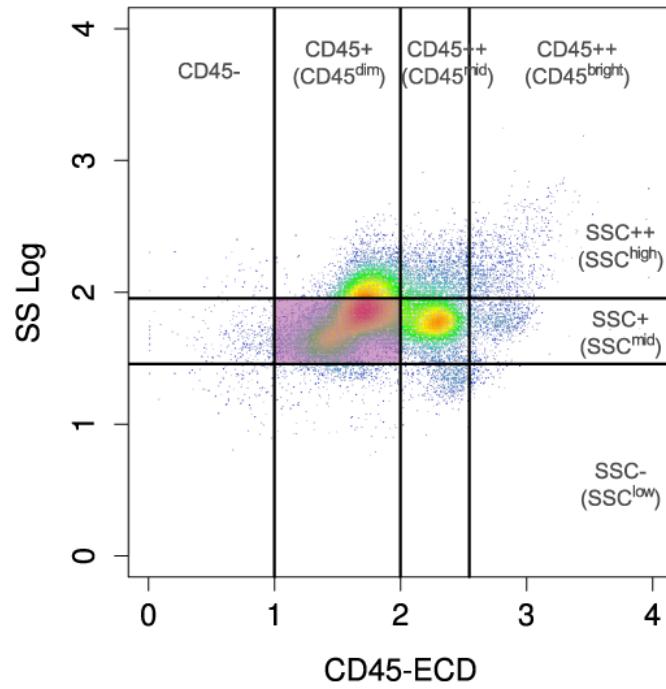


Automated Analysis is Robust, Reproducible and Reduces Variability



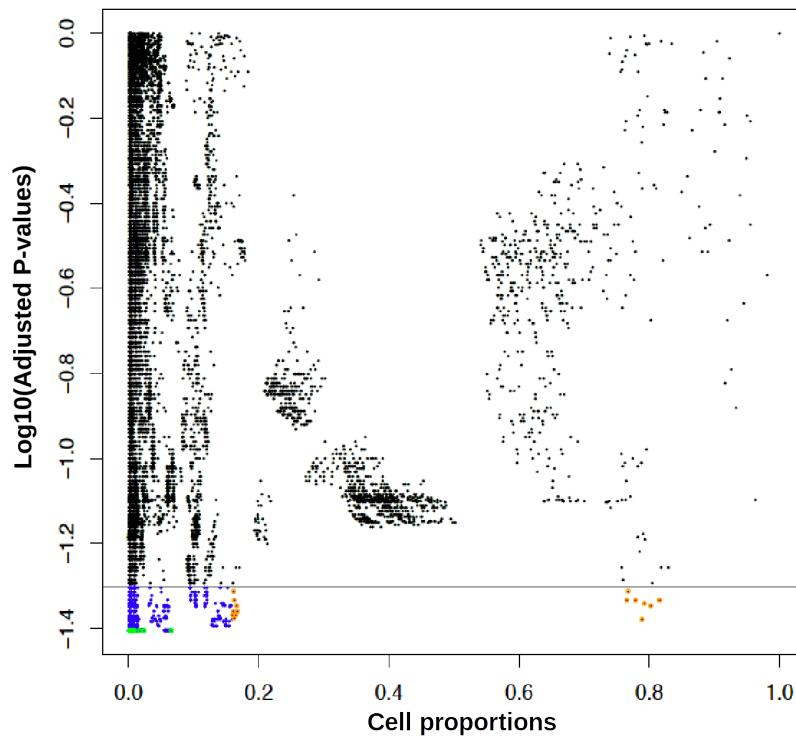
Step 5. FlowType flowType Immunophenotype Extraction Concept

For ten markers: $3^{10} \approx 60,000$ possible cell populations
CD45/SSC gating strategy



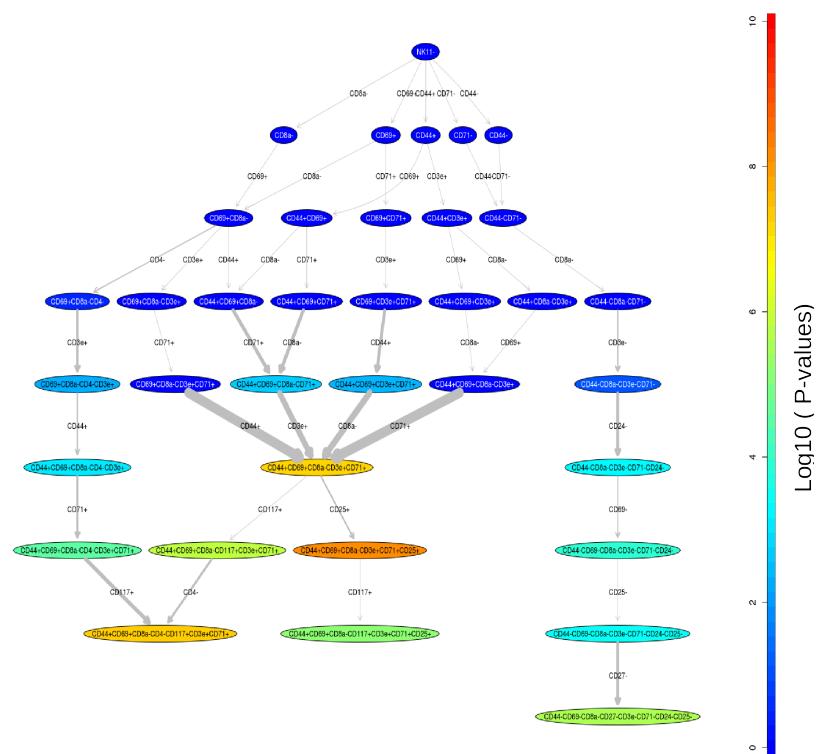
Step 5. FlowType

ID significant Immunophenotypes (from 60,000 cell populations / KO)



Step 6. RchyOptimyx

Most significant phenotypes



flowType/RchyOptimyx and flowDensity

Automated analysis for discovery and diagnosis in Big Flow Cytometry Data

- ▶ flowDensity: Pipeline for Diagnosis
 - ▶ Finds what you want to find, how you want find it
 - ▶ Based on density estimation techniques
 - ▶ Seconds per FCS file
 - ▶ Identical to the manual practice of 2D gating
 - ▶ Guaranteed* (lower CV, same range), or your money back
- ▶ flowType/RchyOptimyx: Pipeline for Discovery
 - ▶ You split FCS files into groups
 - ▶ Pipeline finds best cell populations that correlate with that split
 - ▶ One graph summary of very large datasets
 - ▶ Can be used as input to large multi-group studies

*When your gating strategy is supported by your data/metadata/standardization

Acknowledgements

R/BioConductor.org flow cytometry infrastructure

Genentech Robert Gentleman, all BioConductor contributors

Data Standards

ISAC DSTF Josef Spidlen (BCCA), Wayne Moore (Stanford), Adam Triester (ex-FlowJo), Chris Bray (VSH), Michael Goldberg (BD), David Novo (DeNovo), David Parks (Stanford) & all other past and current members

FlowCAP

Coordinating Committee Nima Aghaeepour (BCCA), Greg Finak (FHCRC), Raphael Gottardo (FHCRC), Tim Mosmann (U Rochester), Richard H. Scheuermann (JCVI)

Data providers and participants flowcap.flowsite.org

flowType/RchyOptimyx

BCCA Nima Aghaeepour, Adrin Jalali, Kieran O'Neill, Habil Zare,

Lympoma Fiona Craig (UPenn)

IMPC Adam Liang, Lucie Abeler-Dorner, Adrian Hayday (Sanger Centre), Hervé Luche (CIPHE), Justin Meskas (BCCA)

flowDensity

BCCA Jafar Taghiyar, Mehrnoush Malekeshmaeli, Radina Droumeva

HIP-C Holden Maecker (Stanford), Bob Nussenblatt, Phil McCoy (NIH) and collaborating HIP-C Consortium centers

CNTRP Megan Levings, Sabine Ivison (CFRI)/

\$

Funding NIH (NIBIB, NIAID), HIP-C, TFRI & TFF, CCS, MSFHR, NSERC, GenomeBC, Genome Canada, CIHR

Additional references

FCM data wrangling using R/BioConductor

Tool	Link	Use
flowCore ¹	PMC2684747	FCM data pipeline
flowClean	PMID: 26990501	Time-based data QA
plateCore	PMC2777006	Multiwell plate analysis
flowQB	Early view Cyto A	Instrument standardization
fda/GuassNorm	PMC3648208	Normalization
flowUtils	PMC4874733	Gating standard support
FlowRepositoryR	BioConductor	API for flowRepository
FCS standard	PMC2892967	Used by every FCM instrument
MIFlowCyt	PMC2773297	Data annotation

¹15,371 unique IP downloads, top 5% in [BioConductor](#)

FCM Automated Analysis Algorithms

Automated cell population identification

Tool	Link	About
flowClust	PMC2684747	First robust clustering tool
flowBin	PMC4426837	Multi-tube clustering
flowMeans ¹	PMC3137288	FlowCAP-I best performer
flowMerge	PMC2798116	Model-based gating
SamSPECTRAL	PMC2923634	Spectral clustering
flowDensity ^{2,3}	PMC4325545	Supervised gating

¹Best combination speed/performance [FlowCAP-I](#)

²Co-best performing approach in [FlowCAP-III](#)

³Used by the 2 approaches with significant performance in [FlowCAP-IV](#)

FCM Automated Analysis Algorithms

Biomarker identification and other post-gating tools

Tool	Link	About
flowType ¹	PMC3315712	Full-D biomarker identification
Feialect ²	PMC3549810	Features for diagnosis
RchyOptimyx ³	PMC3998128	Full-D biomarker visualization
flowCL	PMC4393520	Semantic cell population labelling

¹Used by the 2 approaches with significant performance in [FlowCAP-IV](#)

²Perfect sensitivity, specificity and accuracy in [FlowCAP-II](#)

³Used by 1/2 of approaches with significant performance in [FlowCAP-IV](#)