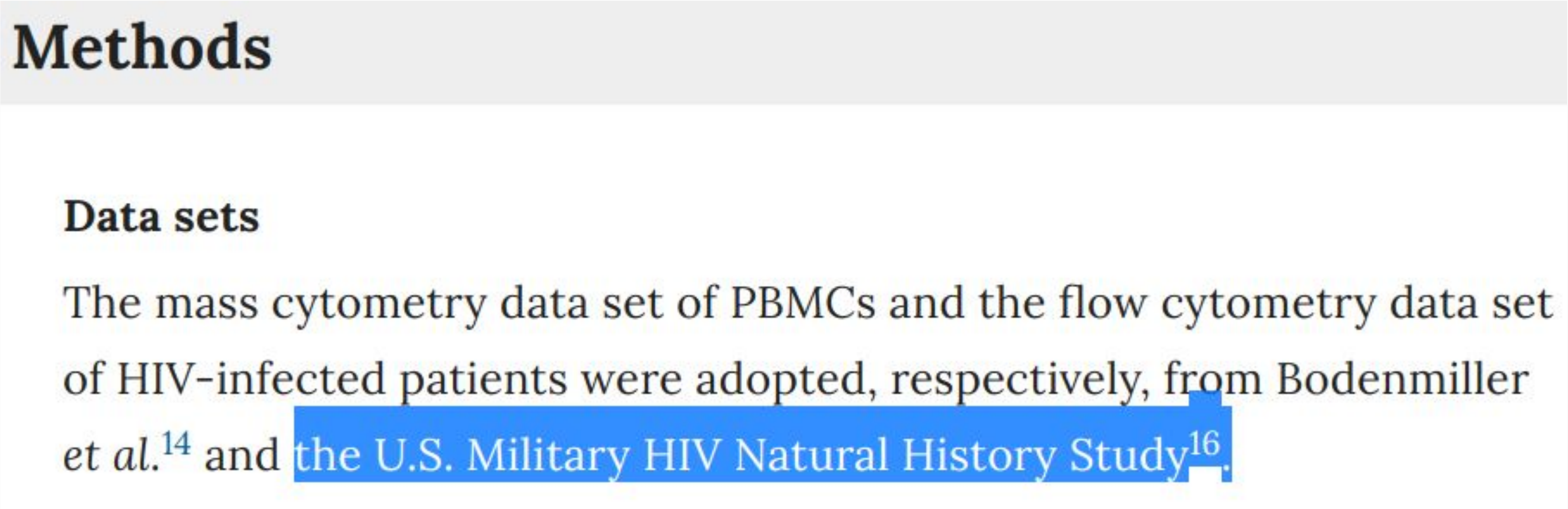
Progress Report

Kuei-Yueh Ko

Sensitive detection of rare disease-associated cell subsets via representation learning

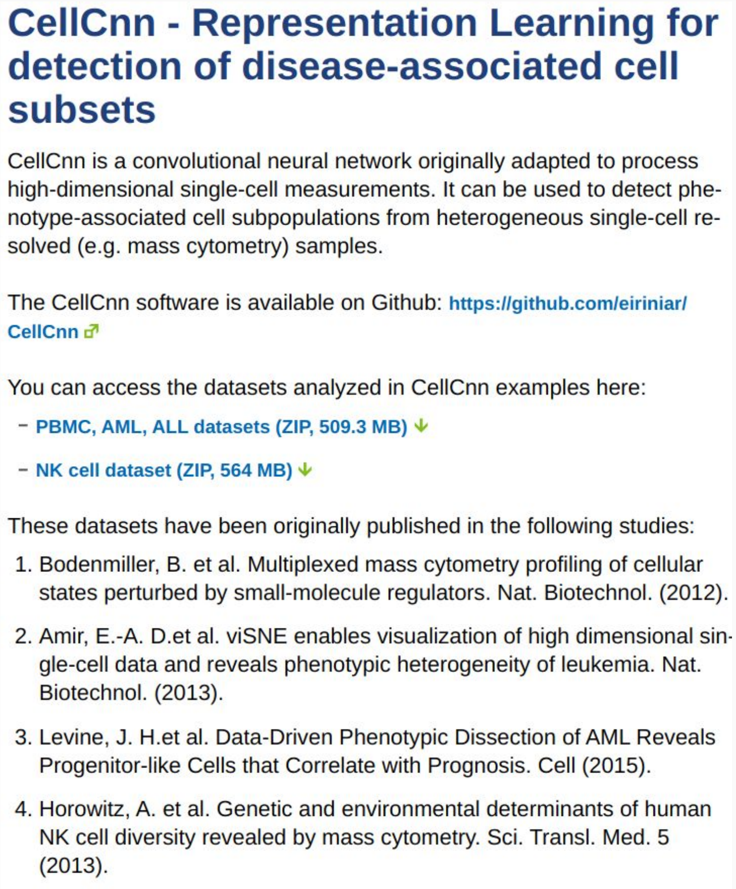


Methods

Data sets

The mass cytometry data set of PBMCs and the flow cytometry data set of HIV-infected patients were adopted, respectively, from Bodenmiller *et al.*[14] and the U.S. Military HIV Natural History Study[16].

Problem --- No HIVdataset



CellCnn - Representation Learning for detection of disease-associated cell subsets

CellCnn is a convolutional neural network originally adapted to process high-dimensional single-cell measurements. It can be used to detect phenotype-associated cell subpopulations from heterogeneous single-cell resolved (e.g. mass cytometry) samples.

The CellCnn software is available on Github: https://github.com/eiriniar/CellCnn

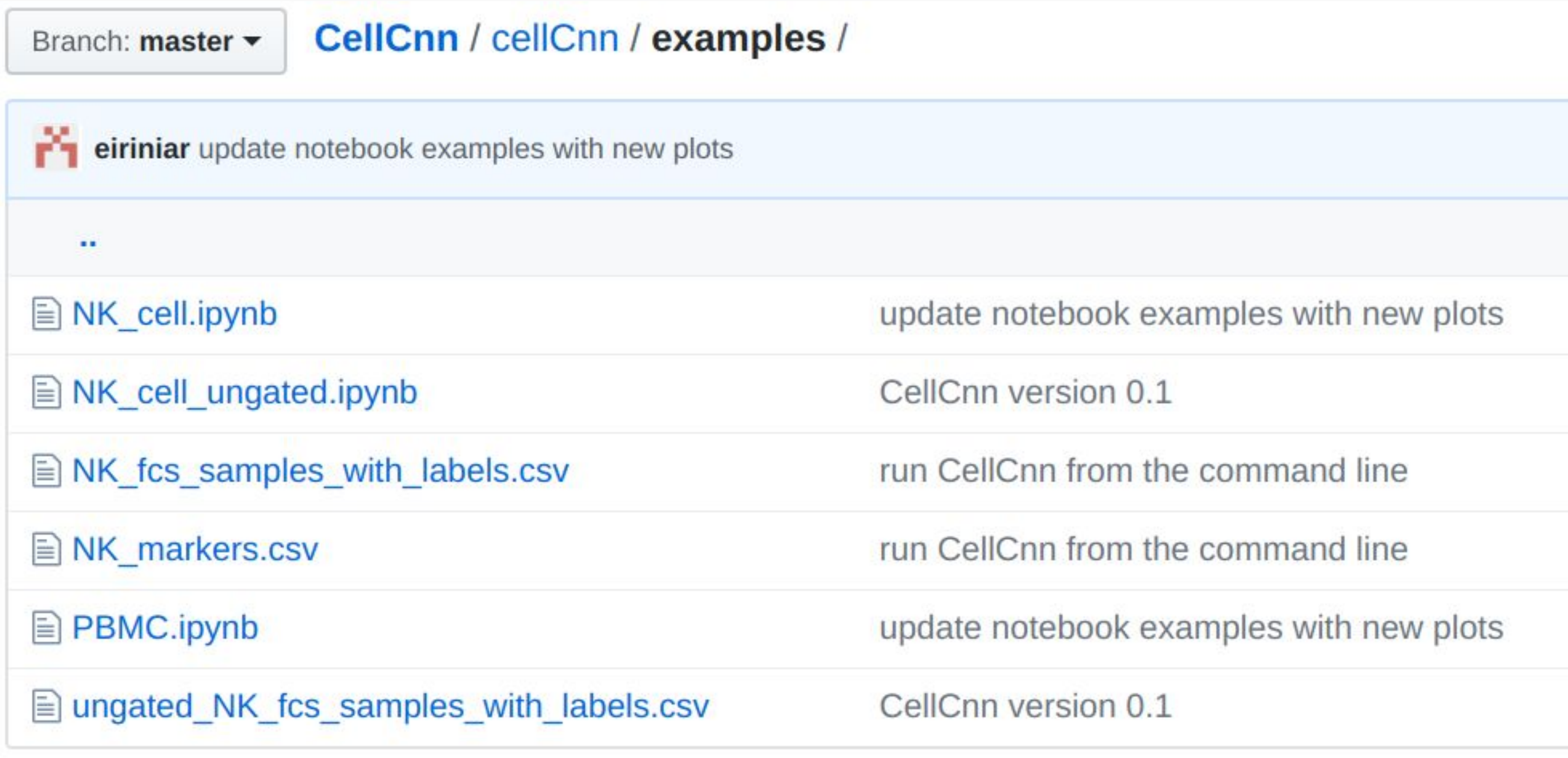
You can access the datasets analyzed in CellCnn examples here:

- PBMC, AML, ALL datasets (ZIP, 509.3 MB)
- NK cell dataset (ZIP, 564 MB)

These datasets have been originally published in the following studies:

1. Bodenmiller, B. et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat. Biotechnol. (2012).
2. Amir, E.-A. D.et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. (2013).
3. Levine, J. H.et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell (2015).
4. Horowitz, A. et al. Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. Sci. Transl. Med. 5 (2013).

Problem --- No CellCnn example on HIV dataset



Branch: **master**

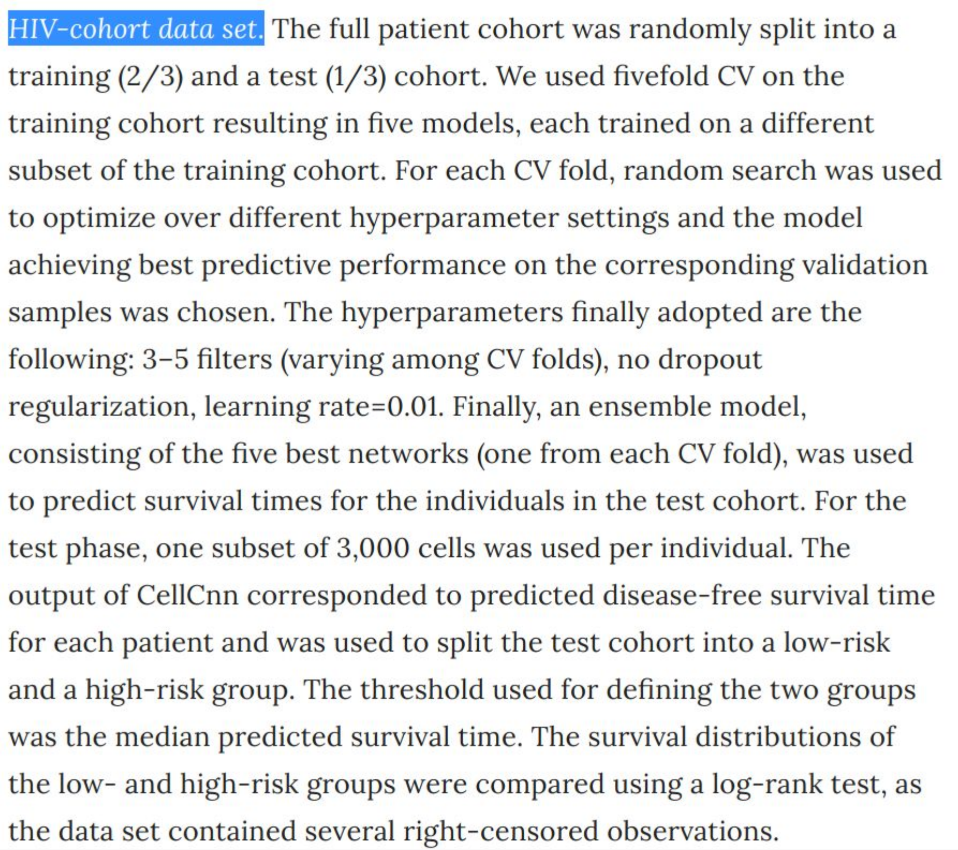
CellCnn / cellCnn / **examples** /

eiriniar update notebook examples with new plots

..

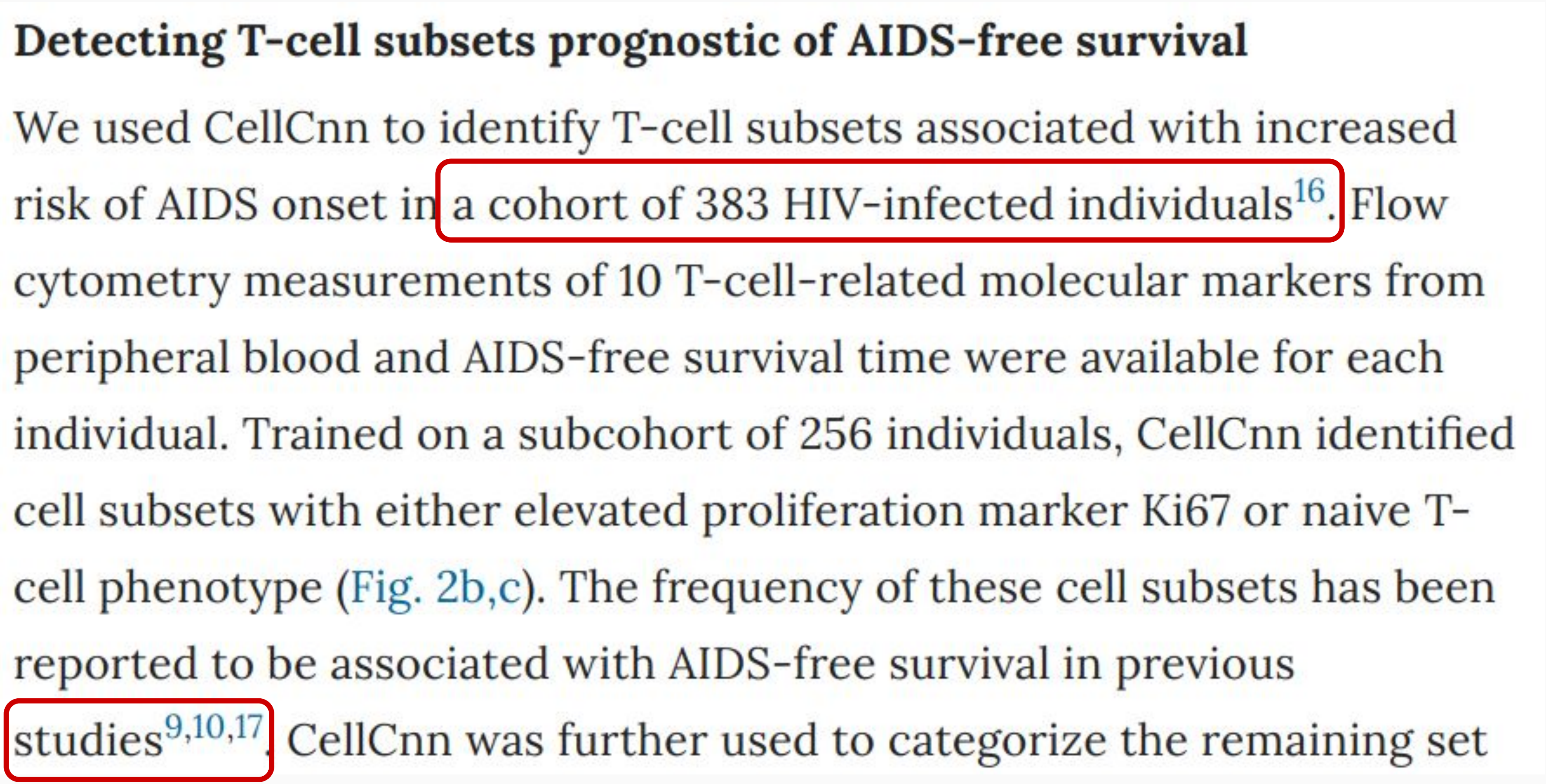
NK_cell.ipynb	update notebook examples with new plots
NK_cell_ungated.ipynb	CellCnn version 0.1
NK_fcs_samples_with_labels.csv	run CellCnn from the command line
NK_markers.csv	run CellCnn from the command line
PBMC.ipynb	update notebook examples with new plots
ungated_NK_fcs_samples_with_labels.csv	CellCnn version 0.1

Sensitive detection of rare disease-associated cell subsets via representation learning



HIV-cohort data set. The full patient cohort was randomly split into a training (2/3) and a test (1/3) cohort. We used fivefold CV on the training cohort resulting in five models, each trained on a different subset of the training cohort. For each CV fold, random search was used to optimize over different hyperparameter settings and the model achieving best predictive performance on the corresponding validation samples was chosen. The hyperparameters finally adopted are the following: 3–5 filters (varying among CV folds), no dropout regularization, learning rate=0.01. Finally, an ensemble model, consisting of the five best networks (one from each CV fold), was used to predict survival times for the individuals in the test cohort. For the test phase, one subset of 3,000 cells was used per individual. The output of CellCnn corresponded to predicted disease-free survival time for each patient and was used to split the test cohort into a low-risk and a high-risk group. The threshold used for defining the two groups was the median predicted survival time. The survival distributions of the low- and high-risk groups were compared using a log-rank test, as the data set contained several right-censored observations.

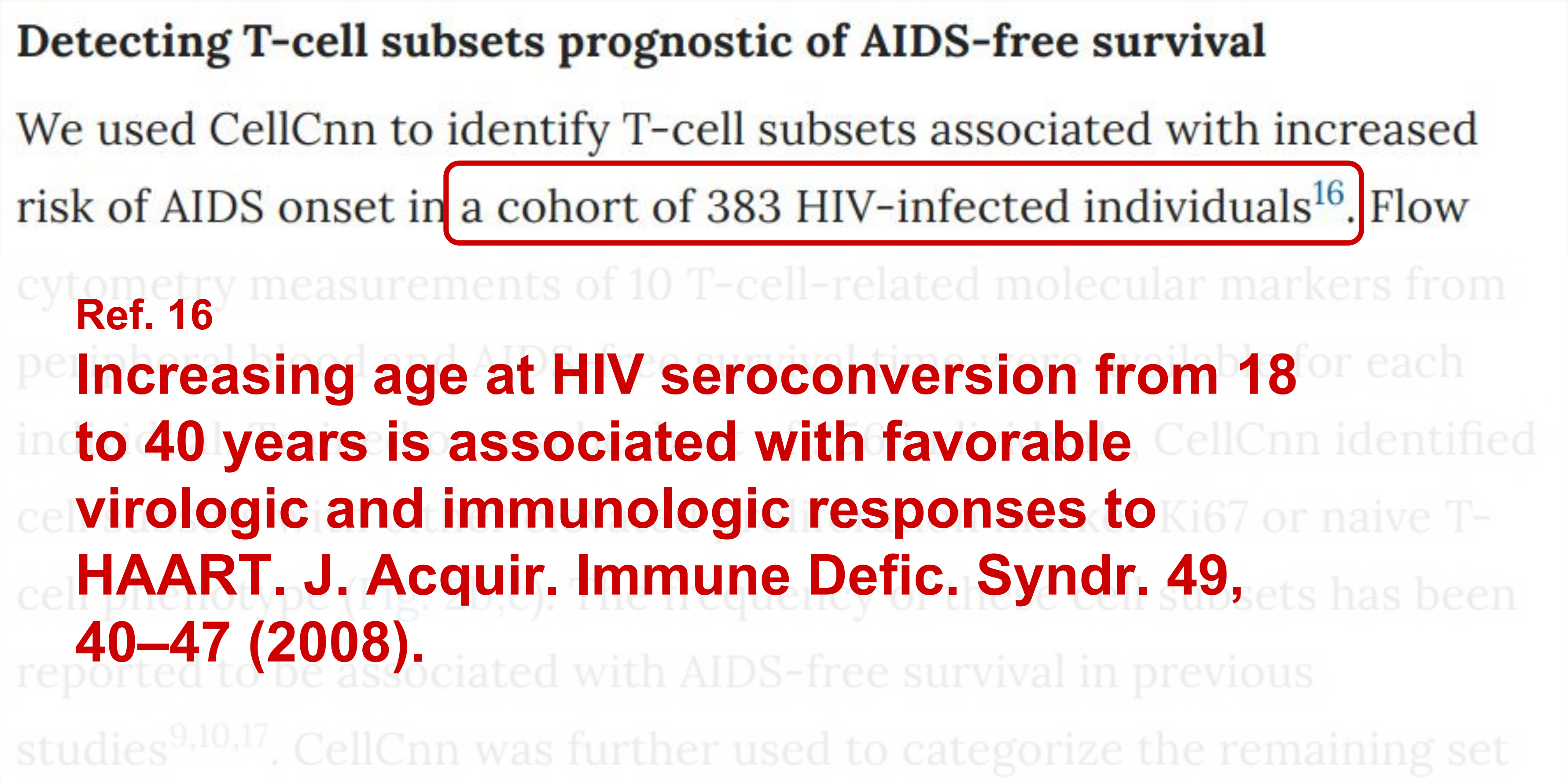
Sensitive detection of rare disease-associated cell subsets via representation learning



Detecting T-cell subsets prognostic of AIDS-free survival

We used CellCnn to identify T-cell subsets associated with increased risk of AIDS onset in a cohort of 383 HIV-infected individuals[16]. Flow cytometry measurements of 10 T-cell-related molecular markers from peripheral blood and AIDS-free survival time were available for each individual. Trained on a subcohort of 256 individuals, CellCnn identified cell subsets with either elevated proliferation marker Ki67 or naive T-cell phenotype (Fig. 2b,c). The frequency of these cell subsets has been reported to be associated with AIDS-free survival in previous studies[9,10,17]. CellCnn was further used to categorize the remaining set

Problem --- I did not find the dataset in the original paper (ref. 16)



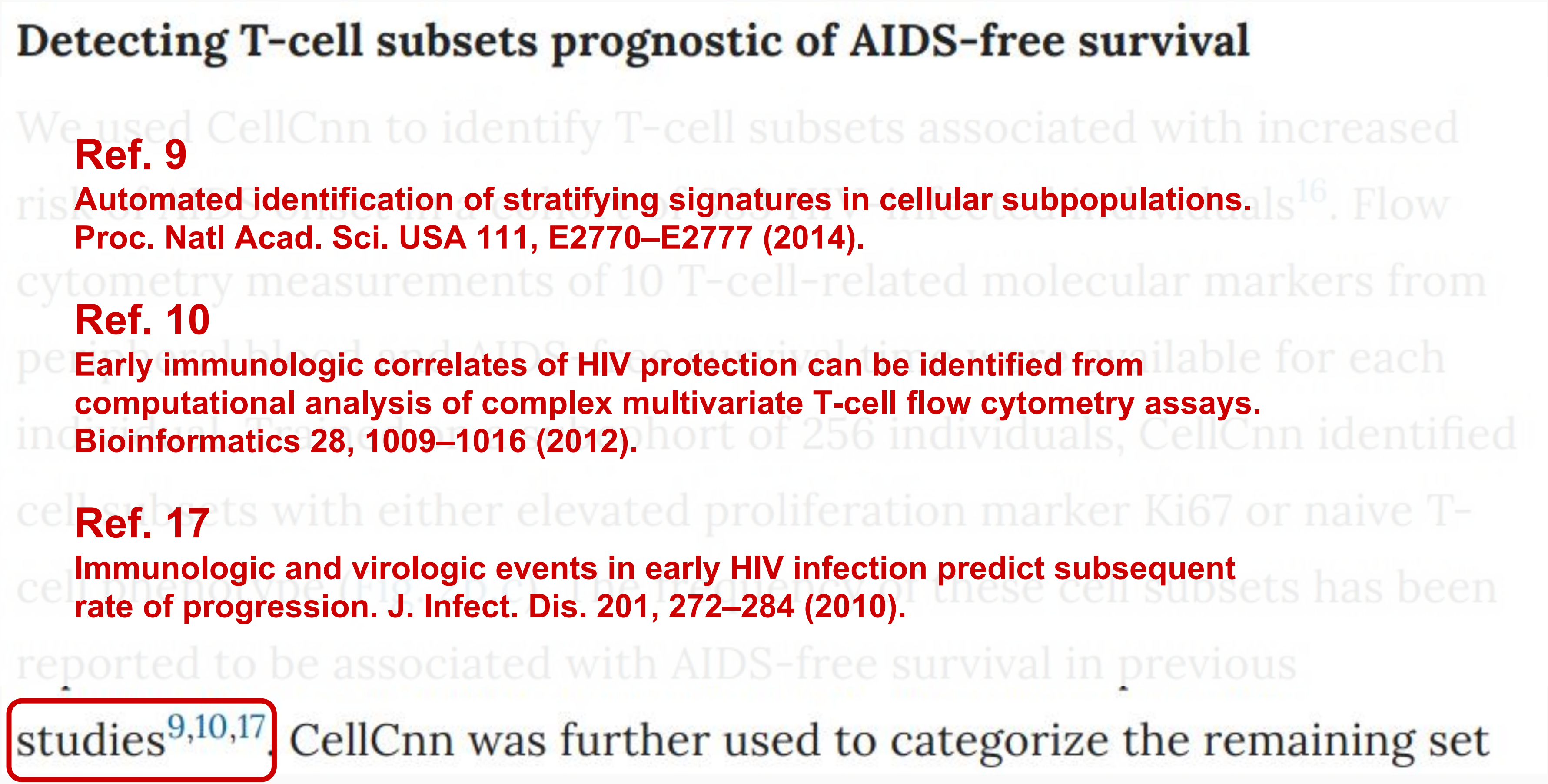
Detecting T-cell subsets prognostic of AIDS-free survival

We used CellCnn to identify T-cell subsets associated with increased risk of AIDS onset in a cohort of 383 HIV-infected individuals[16]. Flow cytometry measurements of 10 T-cell-related molecular markers from pe… for each in… CellCnn identified ce… Ki67 or naive T-ce… sets has been reported to be associated with AIDS-free survival in previous studies[9,10,17]. CellCnn was further used to categorize the remaining set

Ref. 16

Increasing age at HIV seroconversion from 18 to 40 years is associated with favorable virologic and immunologic responses to HAART. J. Acquir. Immune Defic. Syndr. 49, 40–47 (2008).

I tried to search the dataset in other reference cited in this section



Detecting T-cell subsets prognostic of AIDS-free survival

We used CellCnn to identify T-cell subsets associated with increased risk of AIDS onset in a cohort of 383 HIV-infected individuals[16]. Flow cytometry measurements of 10 T-cell-related molecular markers from peripheral blood and AIDS-free survival time were available for each individual. Trained on a cohort of 256 individuals, CellCnn identified cell subsets with either elevated proliferation marker Ki67 or naive T-cell phenotype (Fig. 5b,c). The frequency of these cell subsets has been reported to be associated with AIDS-free survival in previous studies[9,10,17]. CellCnn was further used to categorize the remaining set

Ref. 9

Automated identification of stratifying signatures in cellular subpopulations. Proc. Natl Acad. Sci. USA 111, E2770–E2777 (2014).

Ref. 10

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

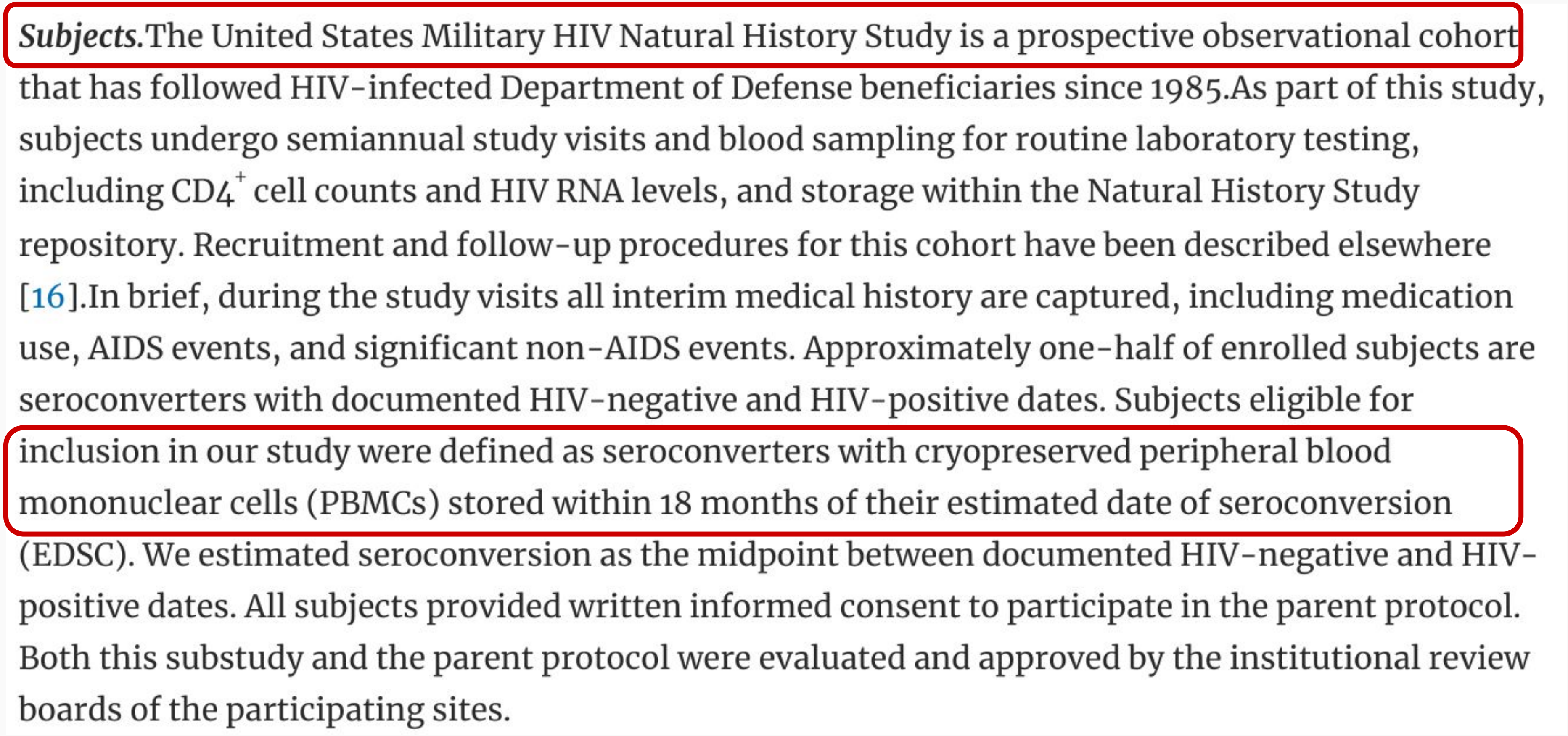
Ref. 17

Immunologic and virologic events in early HIV infection predict subsequent rate of progression. J. Infect. Dis. 201, 272–284 (2010).

Ref. 9

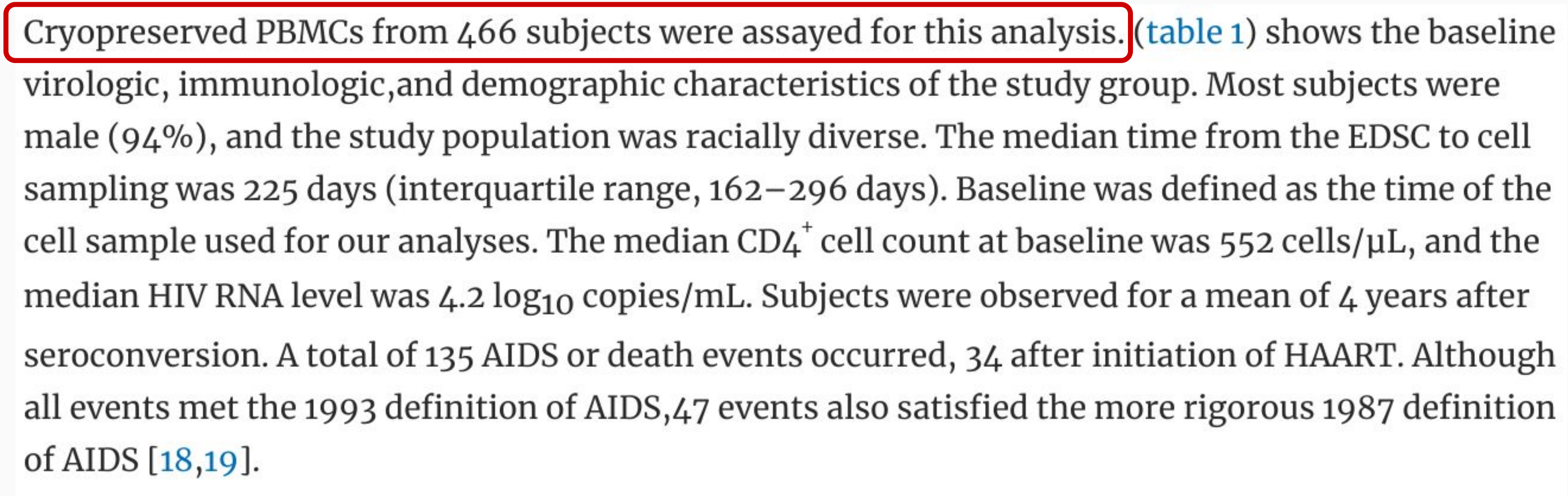
Immunologic and virologic events in early HIV infection predict subsequent rate of progression.

Immunologic and virologic events in early HIV infection predict subsequent rate of progression.



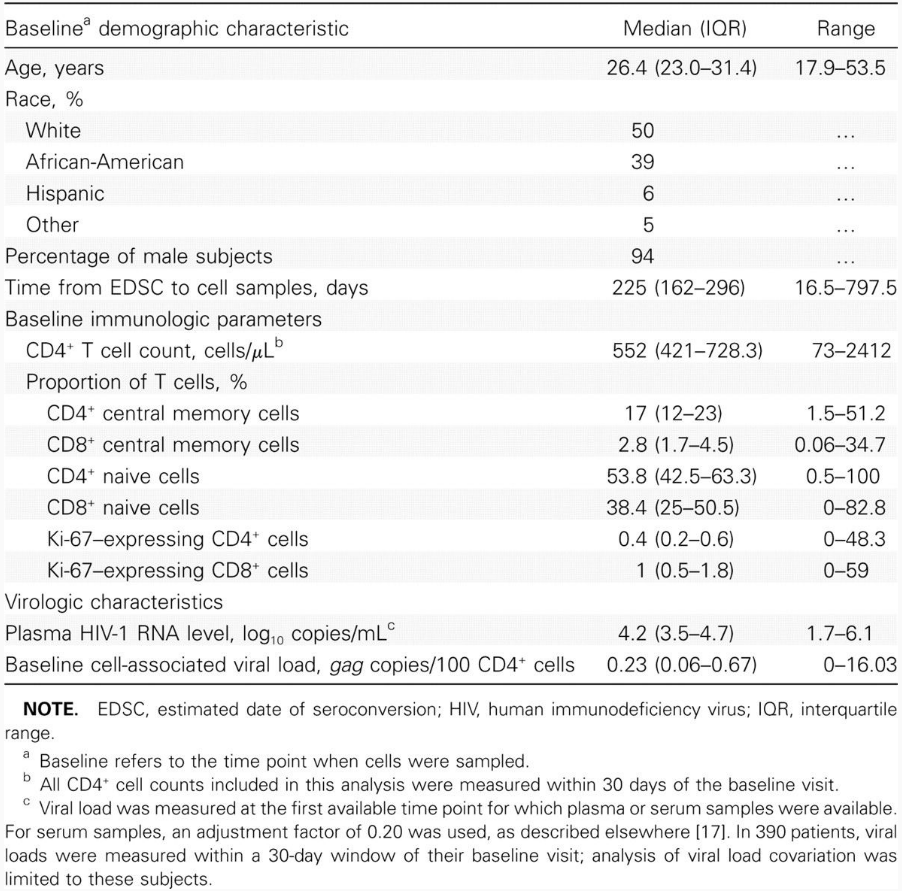
Subjects.The United States Military HIV Natural History Study is a prospective observational cohort that has followed HIV-infected Department of Defense beneficiaries since 1985.As part of this study, subjects undergo semiannual study visits and blood sampling for routine laboratory testing, including $CD4^+$ cell counts and HIV RNA levels, and storage within the Natural History Study repository. Recruitment and follow-up procedures for this cohort have been described elsewhere [16].In brief, during the study visits all interim medical history are captured, including medication use, AIDS events, and significant non-AIDS events. Approximately one-half of enrolled subjects are seroconverters with documented HIV-negative and HIV-positive dates. Subjects eligible for inclusion in our study were defined as seroconverters with cryopreserved peripheral blood mononuclear cells (PBMCs) stored within 18 months of their estimated date of seroconversion (EDSC). We estimated seroconversion as the midpoint between documented HIV-negative and HIV-positive dates. All subjects provided written informed consent to participate in the parent protocol. Both this substudy and the parent protocol were evaluated and approved by the institutional review boards of the participating sites.

Immunologic and virologic events in early HIV infection predict subsequent rate of progression.



Cryopreserved PBMCs from 466 subjects were assayed for this analysis. (table 1) shows the baseline virologic, immunologic,and demographic characteristics of the study group. Most subjects were male (94%), and the study population was racially diverse. The median time from the EDSC to cell sampling was 225 days (interquartile range, 162–296 days). Baseline was defined as the time of the cell sample used for our analyses. The median CD4$^+$ cell count at baseline was 552 cells/µL, and the median HIV RNA level was 4.2 log$_{10}$ copies/mL. Subjects were observed for a mean of 4 years after seroconversion. A total of 135 AIDS or death events occurred, 34 after initiation of HAART. Although all events met the 1993 definition of AIDS,47 events also satisfied the more rigorous 1987 definition of AIDS [18,19].

Immunologic and virologic events in early HIV infection predict subsequent rate of progression.



Baseline[a] demographic characteristic	Median (IQR)	Range
Age, years	26.4 (23.0–31.4)	17.9–53.5
Race, %		
White	50	...
African-American	39	...
Hispanic	6	...
Other	5	...
Percentage of male subjects	94	...
Time from EDSC to cell samples, days	225 (162–296)	16.5–797.5
Baseline immunologic parameters		
$CD4^+$ T cell count, cells/μL[b]	552 (421–728.3)	73–2412
Proportion of T cells, %		
$CD4^+$ central memory cells	17 (12–23)	1.5–51.2
$CD8^+$ central memory cells	2.8 (1.7–4.5)	0.06–34.7
$CD4^+$ naive cells	53.8 (42.5–63.3)	0.5–100
$CD8^+$ naive cells	38.4 (25–50.5)	0–82.8
Ki-67–expressing $CD4^+$ cells	0.4 (0.2–0.6)	0–48.3
Ki-67–expressing $CD8^+$ cells	1 (0.5–1.8)	0–59
Virologic characteristics		
Plasma HIV-1 RNA level, $\log_{10}$ copies/mL[c]	4.2 (3.5–4.7)	1.7–6.1
Baseline cell-associated viral load, *gag* copies/100 $CD4^+$ cells	0.23 (0.06–0.67)	0–16.03

NOTE. EDSC, estimated date of seroconversion; HIV, human immunodeficiency virus; IQR, interquartile range.

[a] Baseline refers to the time point when cells were sampled.

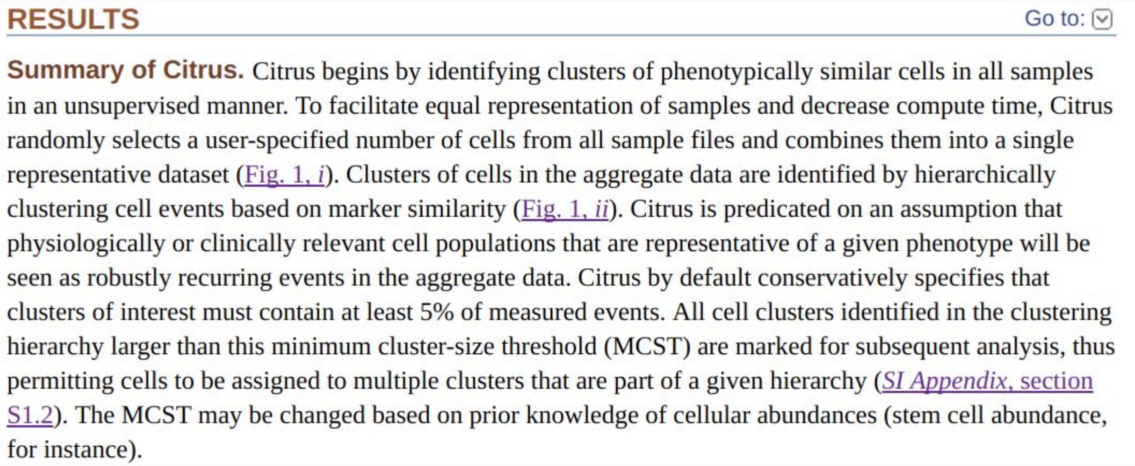
[b] All $CD4^+$ cell counts included in this analysis were measured within 30 days of the baseline visit.

[c] Viral load was measured at the first available time point for which plasma or serum samples were available. For serum samples, an adjustment factor of 0.20 was used, as described elsewhere [17]. In 390 patients, viral loads were measured within a 30-day window of their baseline visit; analysis of viral load covariation was limited to these subjects.

Ref. 10
Automated identification of stratifying signatures in cellular subpopulations

Automated identification of stratifying signatures in cellular subpopulations

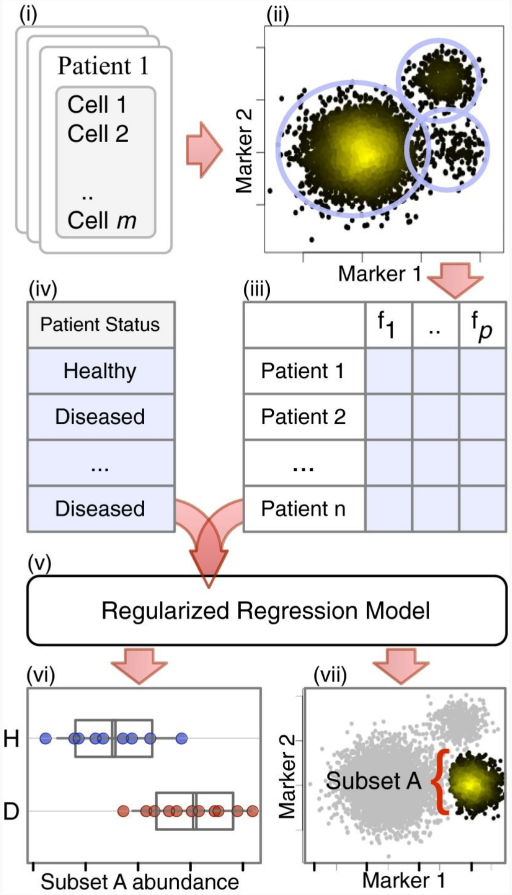
Citrus begins by identifying clusters of phenotypically similar cells in all samples in an unsupervised manner.



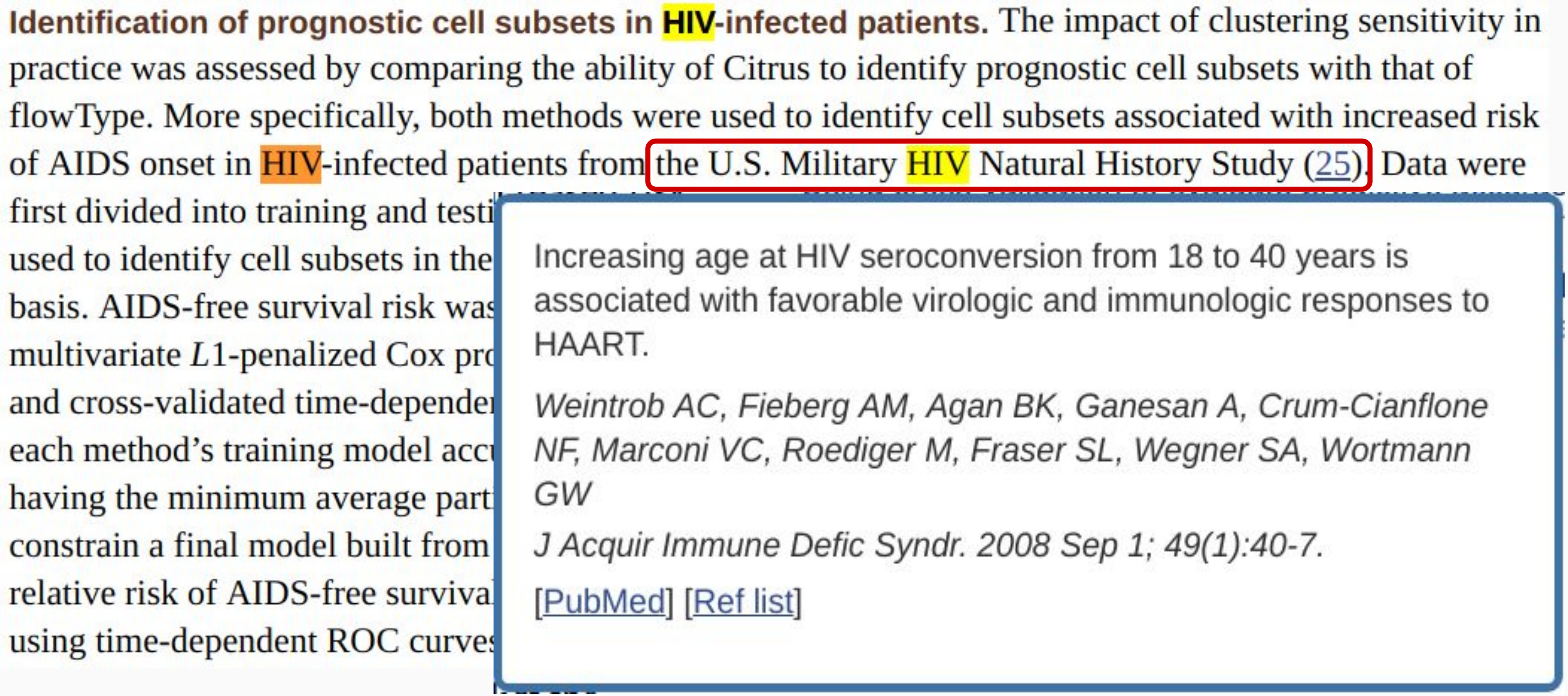
RESULTS

Go to: ⌄

Summary of Citrus. Citrus begins by identifying clusters of phenotypically similar cells in all samples in an unsupervised manner. To facilitate equal representation of samples and decrease compute time, Citrus randomly selects a user-specified number of cells from all sample files and combines them into a single representative dataset (Fig. 1, *i*). Clusters of cells in the aggregate data are identified by hierarchically clustering cell events based on marker similarity (Fig. 1, *ii*). Citrus is predicated on an assumption that physiologically or clinically relevant cell populations that are representative of a given phenotype will be seen as robustly recurring events in the aggregate data. Citrus by default conservatively specifies that clusters of interest must contain at least 5% of measured events. All cell clusters identified in the clustering hierarchy larger than this minimum cluster-size threshold (MCST) are marked for subsequent analysis, thus permitting cells to be assigned to multiple clusters that are part of a given hierarchy (*SI Appendix*, section S1.2). The MCST may be changed based on prior knowledge of cellular abundances (stem cell abundance, for instance).



Automated identification of stratifying signatures in cellular subpopulations



Identification of prognostic cell subsets in HIV-infected patients. The impact of clustering sensitivity in practice was assessed by comparing the ability of Citrus to identify prognostic cell subsets with that of flowType. More specifically, both methods were used to identify cell subsets associated with increased risk of AIDS onset in HIV-infected patients from the U.S. Military HIV Natural History Study (25). Data were
first divided into training and testi
used to identify cell subsets in the
basis. AIDS-free survival risk was
multivariate *L*1-penalized Cox pro
and cross-validated time-depender
each method's training model accu
having the minimum average part
constrain a final model built from
relative risk of AIDS-free survival
using time-dependent ROC curves

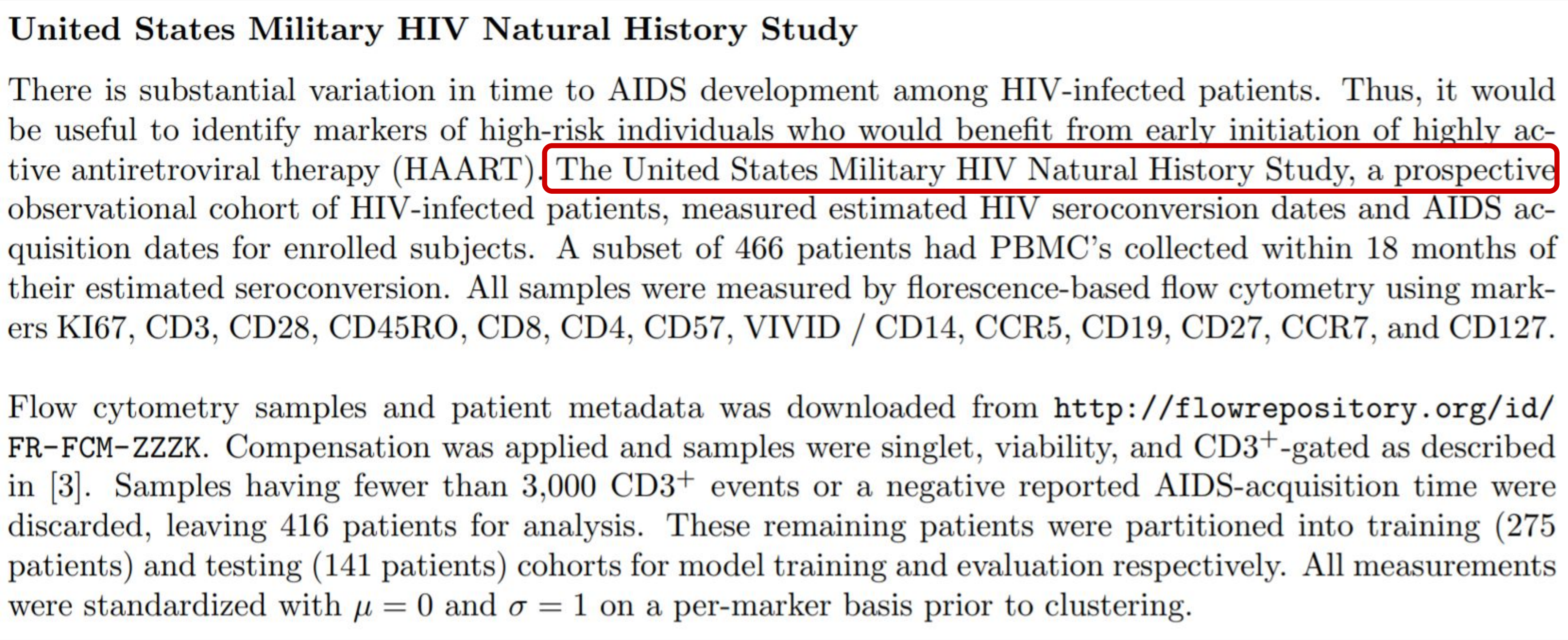
Increasing age at HIV seroconversion from 18 to 40 years is associated with favorable virologic and immunologic responses to HAART.

Weintrob AC, Fieberg AM, Agan BK, Ganesan A, Crum-Cianflone NF, Marconi VC, Roediger M, Fraser SL, Wegner SA, Wortmann GW

J Acquir Immune Defic Syndr. 2008 Sep 1; 49(1):40-7.

[PubMed] [Ref list]

Automated identification of stratifying signatures in cellular subpopulations

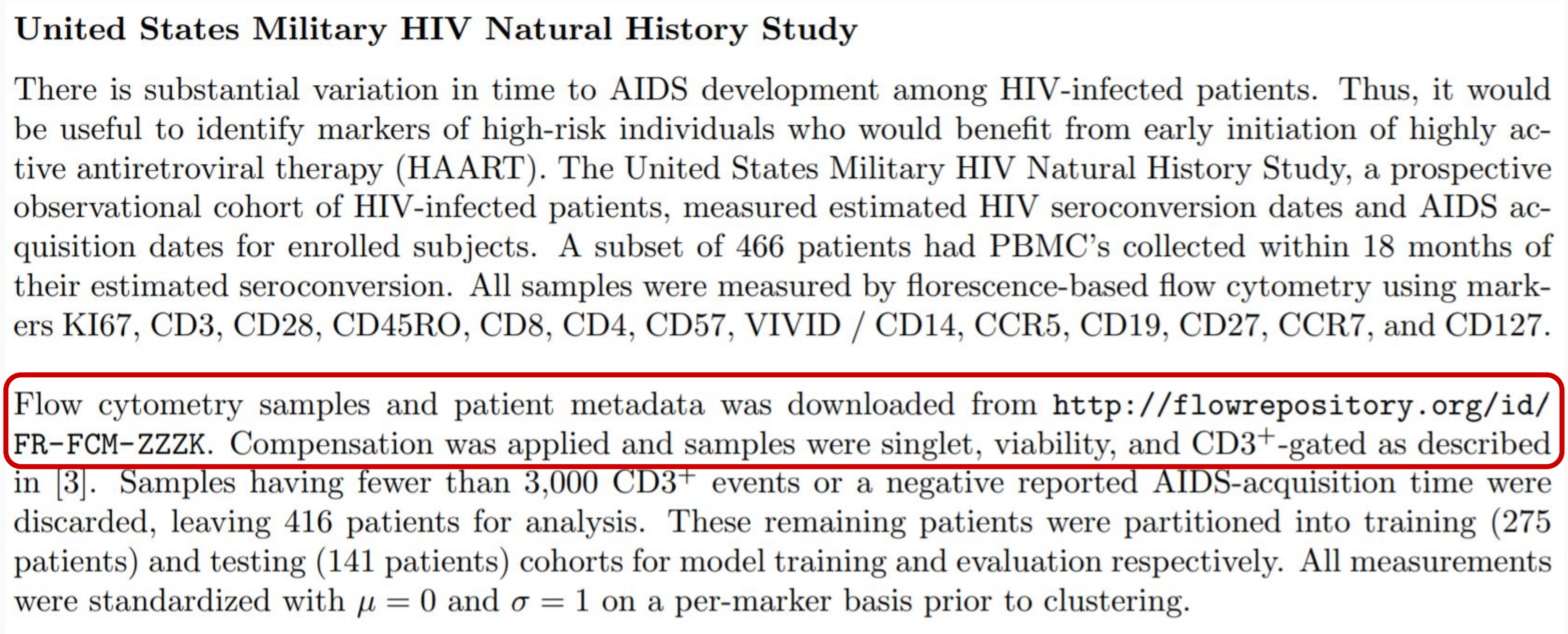


United States Military HIV Natural History Study

There is substantial variation in time to AIDS development among HIV-infected patients. Thus, it would be useful to identify markers of high-risk individuals who would benefit from early initiation of highly active antiretroviral therapy (HAART). The United States Military HIV Natural History Study, a prospective observational cohort of HIV-infected patients, measured estimated HIV seroconversion dates and AIDS acquisition dates for enrolled subjects. A subset of 466 patients had PBMC's collected within 18 months of their estimated seroconversion. All samples were measured by florescence-based flow cytometry using markers KI67, CD3, CD28, CD45RO, CD8, CD4, CD57, VIVID / CD14, CCR5, CD19, CD27, CCR7, and CD127.

Flow cytometry samples and patient metadata was downloaded from `http://flowrepository.org/id/FR-FCM-ZZZK`. Compensation was applied and samples were singlet, viability, and $CD3^+$-gated as described in [3]. Samples having fewer than 3,000 $CD3^+$ events or a negative reported AIDS-acquisition time were discarded, leaving 416 patients for analysis. These remaining patients were partitioned into training (275 patients) and testing (141 patients) cohorts for model training and evaluation respectively. All measurements were standardized with $\mu = 0$ and $\sigma = 1$ on a per-marker basis prior to clustering.

Automated identification of stratifying signatures in cellular subpopulations

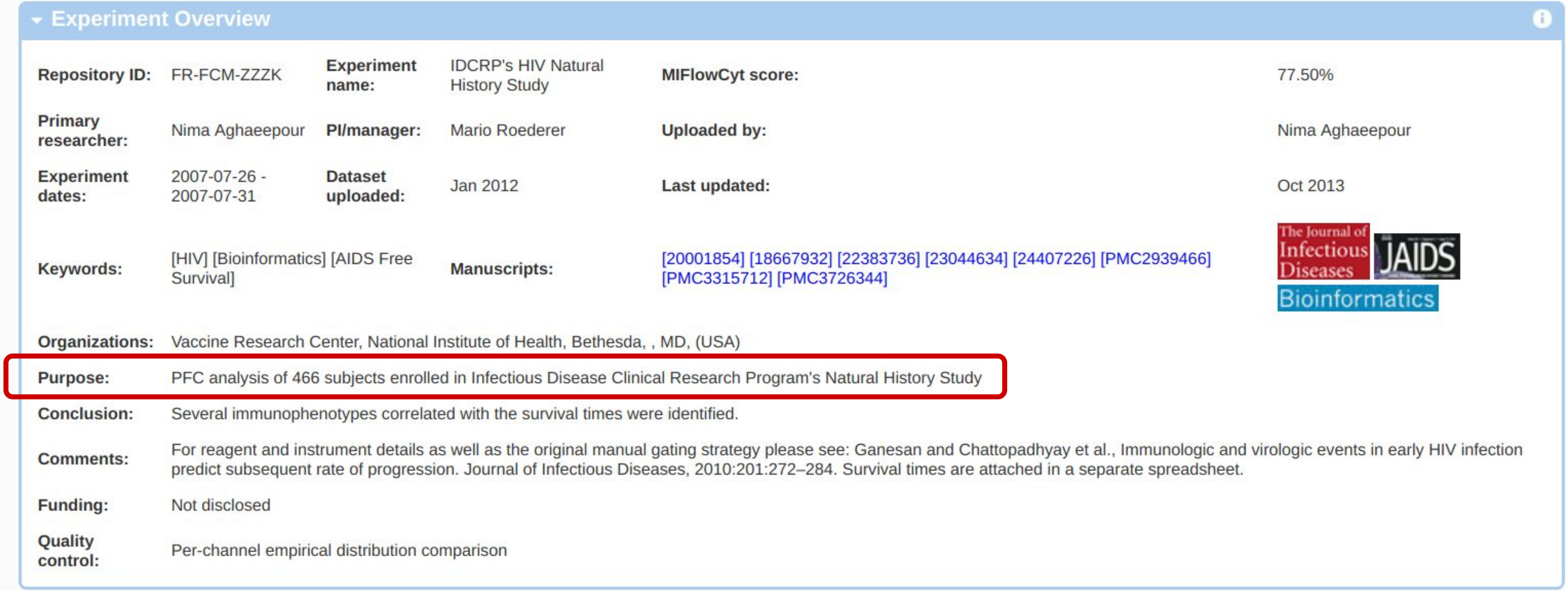


United States Military HIV Natural History Study

There is substantial variation in time to AIDS development among HIV-infected patients. Thus, it would be useful to identify markers of high-risk individuals who would benefit from early initiation of highly active antiretroviral therapy (HAART). The United States Military HIV Natural History Study, a prospective observational cohort of HIV-infected patients, measured estimated HIV seroconversion dates and AIDS acquisition dates for enrolled subjects. A subset of 466 patients had PBMC's collected within 18 months of their estimated seroconversion. All samples were measured by florescence-based flow cytometry using markers KI67, CD3, CD28, CD45RO, CD8, CD4, CD57, VIVID / CD14, CCR5, CD19, CD27, CCR7, and CD127.

Flow cytometry samples and patient metadata was downloaded from `http://flowrepository.org/id/FR-FCM-ZZZK`. Compensation was applied and samples were singlet, viability, and CD3$^+$-gated as described in [3]. Samples having fewer than 3,000 CD3$^+$ events or a negative reported AIDS-acquisition time were discarded, leaving 416 patients for analysis. These remaining patients were partitioned into training (275 patients) and testing (141 patients) cohorts for model training and evaluation respectively. All measurements were standardized with $\mu = 0$ and $\sigma = 1$ on a per-marker basis prior to clustering.

Automated identification of stratifying signatures in cellular subpopulations



Experiment Overview

Repository ID:	FR-FCM-ZZZK	**Experiment name:**	IDCRP's HIV Natural History Study	**MIFlowCyt score:**	77.50%
Primary researcher:	Nima Aghaeepour	**PI/manager:**	Mario Roederer	**Uploaded by:**	Nima Aghaeepour
Experiment dates:	2007-07-26 - 2007-07-31	**Dataset uploaded:**	Jan 2012	**Last updated:**	Oct 2013
Keywords:	[HIV] [Bioinformatics] [AIDS Free Survival]	**Manuscripts:**	[20001854] [18667932] [22383736] [23044634] [24407226] [PMC2939466] [PMC3315712] [PMC3726344]		The Journal of Infectious Diseases JAIDS Bioinformatics

Organizations: Vaccine Research Center, National Institute of Health, Bethesda, , MD, (USA)

Purpose: PFC analysis of 466 subjects enrolled in Infectious Disease Clinical Research Program's Natural History Study

Conclusion: Several immunophenotypes correlated with the survival times were identified.

Comments: For reagent and instrument details as well as the original manual gating strategy please see: Ganesan and Chattopadhyay et al., Immunologic and virologic events in early HIV infection predict subsequent rate of progression. Journal of Infectious Diseases, 2010:201:272–284. Survival times are attached in a separate spreadsheet.

Funding: Not disclosed

Quality control: Per-channel empirical distribution comparison

PFC analysis of 466 subjects enrolled in Infectious Disease Clinical Research Program's Natural History Study