Clustering algorithms and model fit to analyze karyotypic variation from flow cytometry data

Margot Henry, Aleeza Gerstein

Department of Statistics and Microbiology, University of Manitoba, Winnipeg, Manitoba

BACKGROUND

- Karyotypic variation in ploidy (the number of chromosome sets) and aneuploidy (aberrant numbers of chromosomes) is observed in multiple biological contexts, such as cancer cells and fungal microbial populations isolated from ecological, clinical, and industrial environments. In order to understand the dynamics of karyotype subpopulations and their role in adapting populations, we require a computational method to identify different subpopulations and quantify the number of cells within them. Flow cytometry is the gold standard method to measure genome size typically from ~10,000 cells from each population of interest.
- Cells are present in all phases of the cell cycle (Figure 1): G0/G1 prior to DNA replication, S phase during replication, and G2/M when cells have double the DNA but haven't divided. Genome size is determined as the mean of the G0/G1 cells and can be compared to the G0/G1 mean from control populations where the ploidy is known.

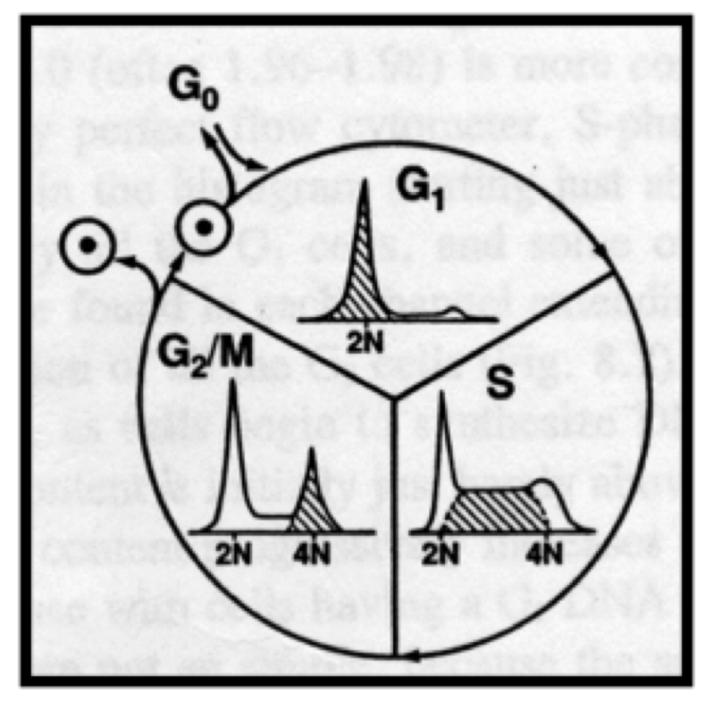
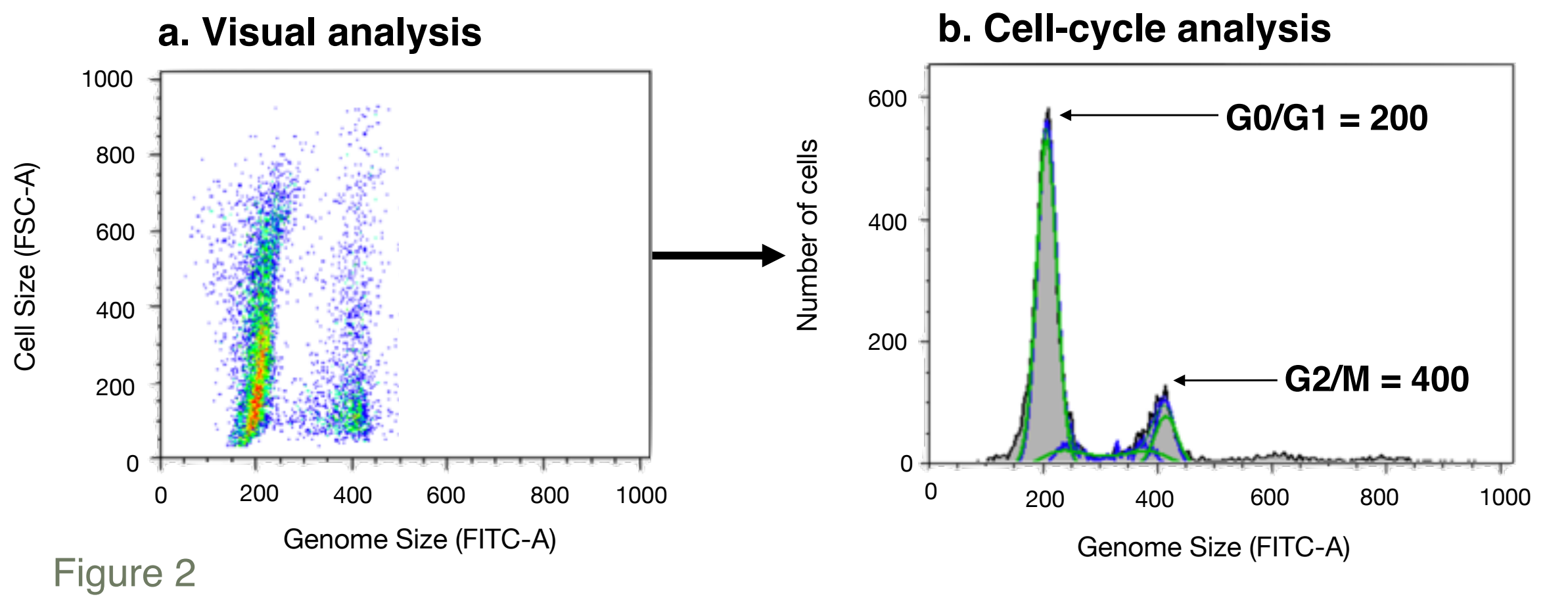


Figure 1

- The Dean-Jett-Fox algorithm is a well-known algorithm that fits a cell-cycle algorithm to a cell population.
- Figure 2 is a visual analysis of cell size (FSC-A) and genome size (FITC-A). From the visual analysis you can see two distinct cell clusters representing the G0/G1 cells (centered around FITC-A = 200) and G2/M cells (centered around FITC-A = 400). A small number of cells are in S phase (in between the two clusters). b) Cell-cycle analysis. The green fitted line represents the Dean-Jett Fox algorithm.


a. Visual analysis

b. Cell-cycle analysis

Figure 2

OBJECTIVE

To develop a new, open-source, method to quantify karyotypic variation in populations from flow cytometry data in an unbiased fashion.

Table 1: Overview of *PloidyCluster* functions

Function	Aim	Purpose	Statistical technique or algorithm
CellCycle	1	Fit a cell-cycle algorithm to a cell population	Dean-Jett-Fox cell-cycle algorithm
FittedError	1	Test how well a cell population fits the Dean-Jett-Fox model	Method of least squares error
FindClusters	2	Find the clusters within the population	Hierarchal clustering algorithm implemented from *hclust* in the fastcluster R package
FindSubPops	2	Find the subpopulations and the number of cells in each one	Use *CellCycle* and *FittedError* to pair the clusters found by *FindClusters*

METHODS

Aim 1: Build a new R package, *PloidyCluster,* to implement the Dean-Jett-Fox model. Develop a statistical method to measure model fit to flag atypical populations that contain subpopulations.

Dean-Jett-Fox algorithm:

- This model assumes that cells in the G0/G1 and G2/M clusters are Gaussian distributed and that the G2/M cluster mean is ~1.75 the size of the G0/G1 mean. For the S phase cells, the model fits a second order polynomial.
- The algorithm is made of three parts: the normal distribution of the G0/G1 peak (1), the normal distribution of the G2/M peak (2), and the second order polynomial of the S phase(3). The parameters A,B and C being decided by methods of least squares error.

(1) G0/G1: $F_1(x) = \frac{N_1}{\sqrt{2\pi}\sigma_1} exp\left[-\frac{(x-x_1)^2}{2\sigma_1{}^2}\right]$

(2) G2/M: $F_2(x) = \frac{N_2}{\sqrt{2\pi}\sigma_2} exp\left[-\frac{(x-x_2)^2}{2\sigma_2{}^2}\right]$

(3) S phase: $F_s(x) = \sum_{j=x_1}^{x_2} f(x_j) \frac{1}{\sqrt{2\pi}\sigma_1\frac{x_j}{x_1}} exp\left[-\frac{(x-x_j)^2}{2(\sigma_1\frac{x_j}{x_1})^2}\right]$

For asynchronous populations:

$$f(x_j) = A + Bx_j + Cx_j^2$$

For synchronous populations:

$$f(x_j) = A + Bx_j + Cx_j^2 + \frac{N_s}{\sqrt{2\pi}\sigma_s} exp\left[-\frac{(x_j - x_1 s)^2}{2\sigma_s{}^2}\right]$$

Aim 1 workflow:

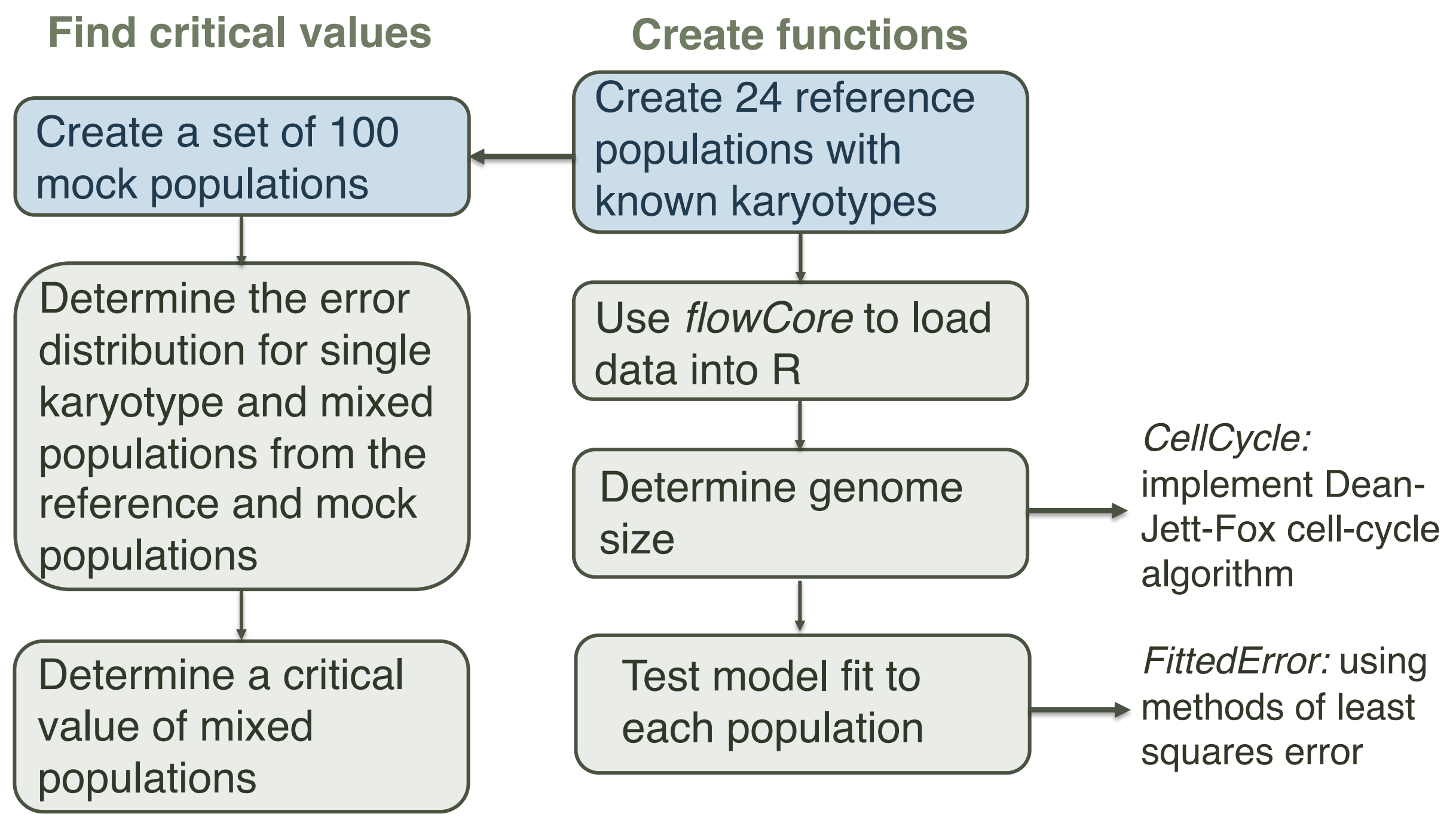
- The workflow consist of two parts, finding critical values and creating functions. Below is a diagram of the workflow, the empirical lab work is indicated in blue, and the computational work as green.

Find critical values:

- This workflow is used to create cutoff points and define critical values to implement into our functions.

Create functions:

- Create *CellCycle* and *FittedError*


Find critical values

Create a set of 100 mock populations

Determine the error distribution for single karyotype and mixed populations from the reference and mock populations

Determine a critical value of mixed populations

Create functions

Create 24 reference populations with known karyotypes

Use *flowCore* to load data into R

Determine genome size

Test model fit to each population

CellCycle: implement Dean-Jett-Fox cell-cycle algorithm

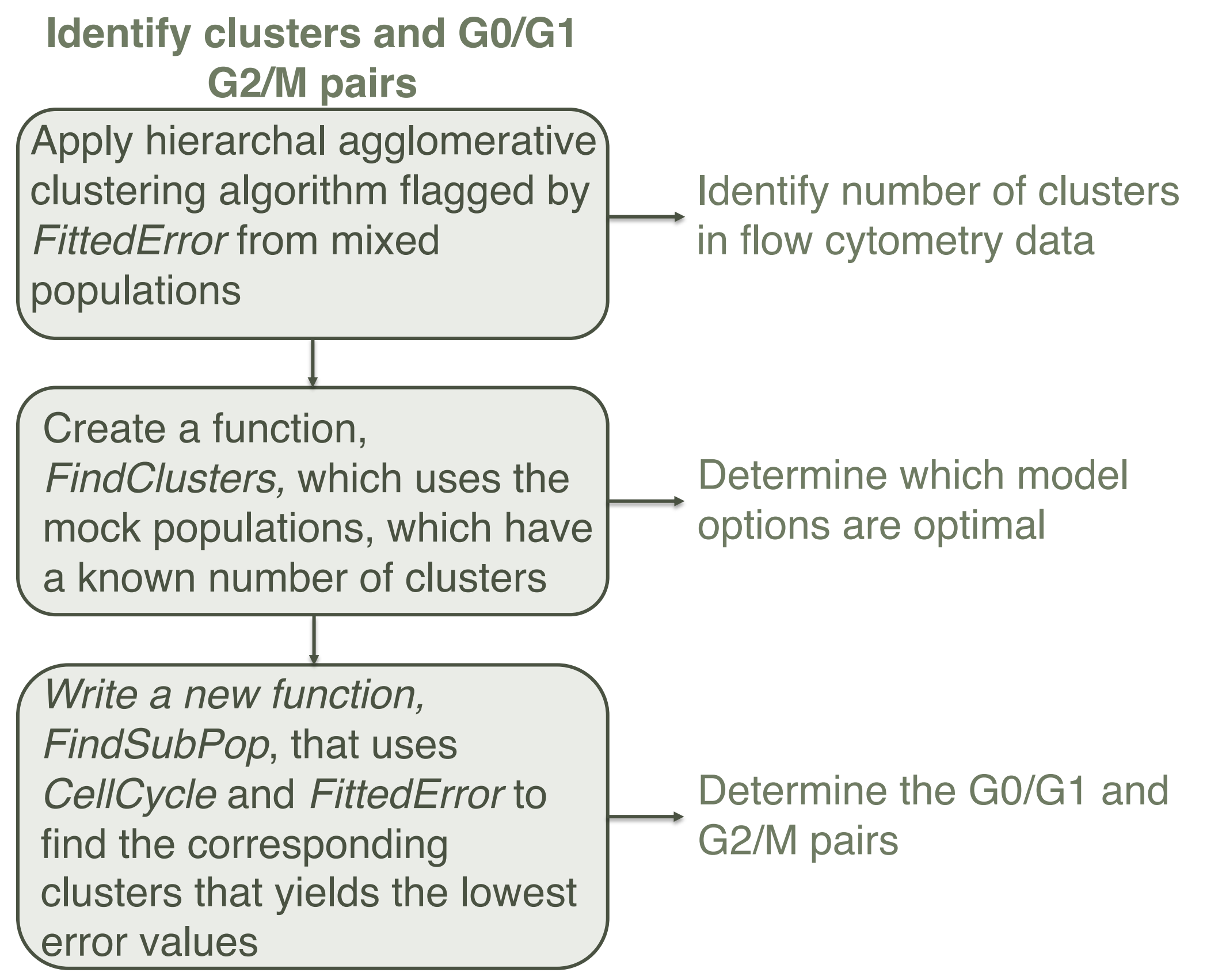
FittedError: using methods of least squares error

METHODS

Aim 2: Use unsupervised machine learning to identify and quantify karyotypic sub-populations.

Aim 2 workflow:

- The workflow consist of identifying the number of clusters in a cell population and determining the G0/G1 and G2/M pairs


Identify clusters and G0/G1 G2/M pairs

Apply hierarchal agglomerative clustering algorithm flagged by *FittedError* from mixed populations

Identify number of clusters in flow cytometry data

Create a function, *FindClusters*, which uses the mock populations, which have a known number of clusters

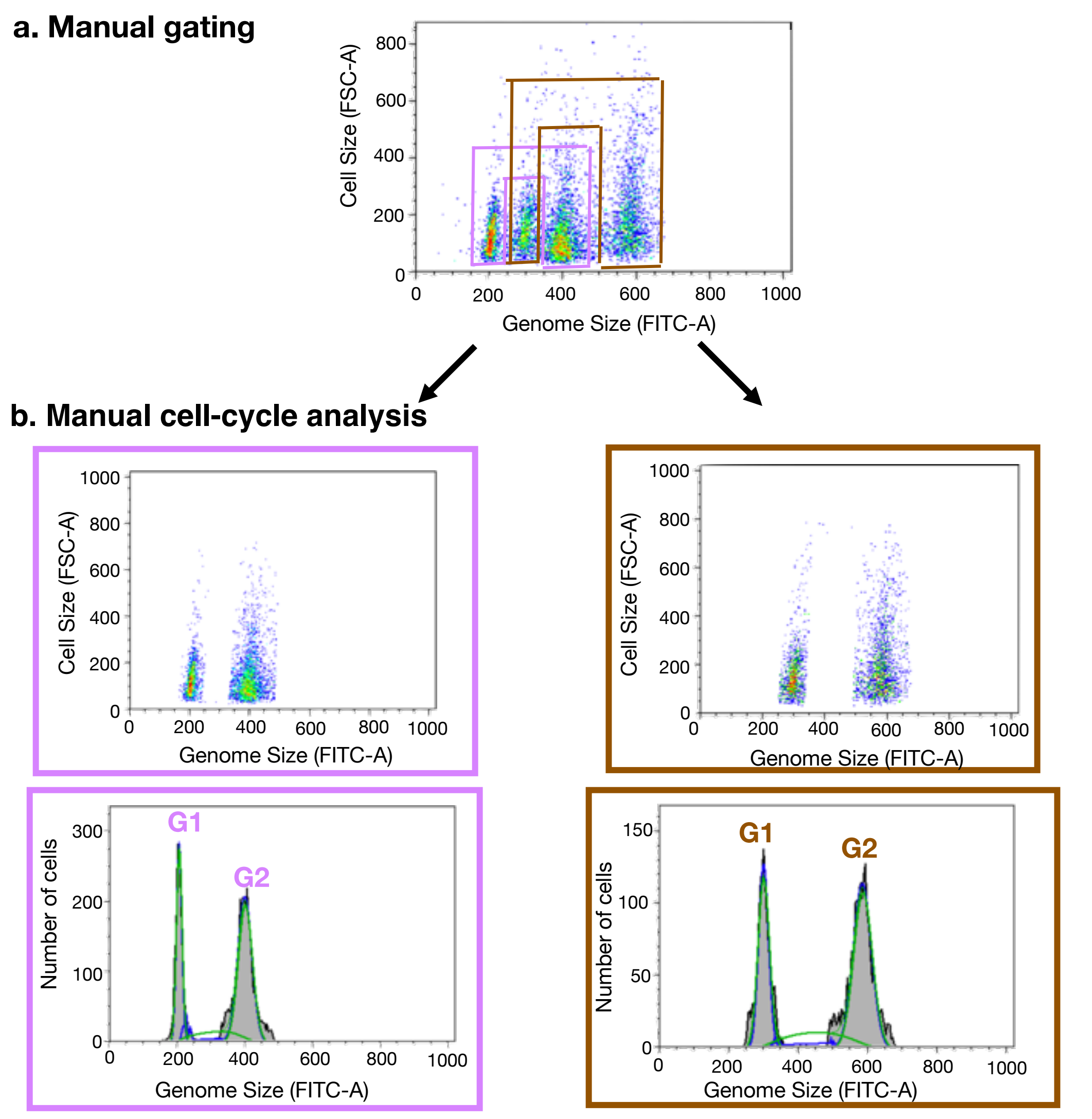
Determine which model options are optimal

Write a new function, FindSubPop, that uses *CellCycle* and *FittedError* to find the corresponding clusters that yields the lowest error values

Determine the G0/G1 and G2/M pairs

The existing workflow for a mixed-populations in Flow-Jo

- In Figure 3 a) we conduct manual gating (pink and brown gates) to match up G1 & G2 clusters based on visual analysis. In b) we repeat visual and cell-cycle analysis from cells in each manual gate separately.


a. Manual gating

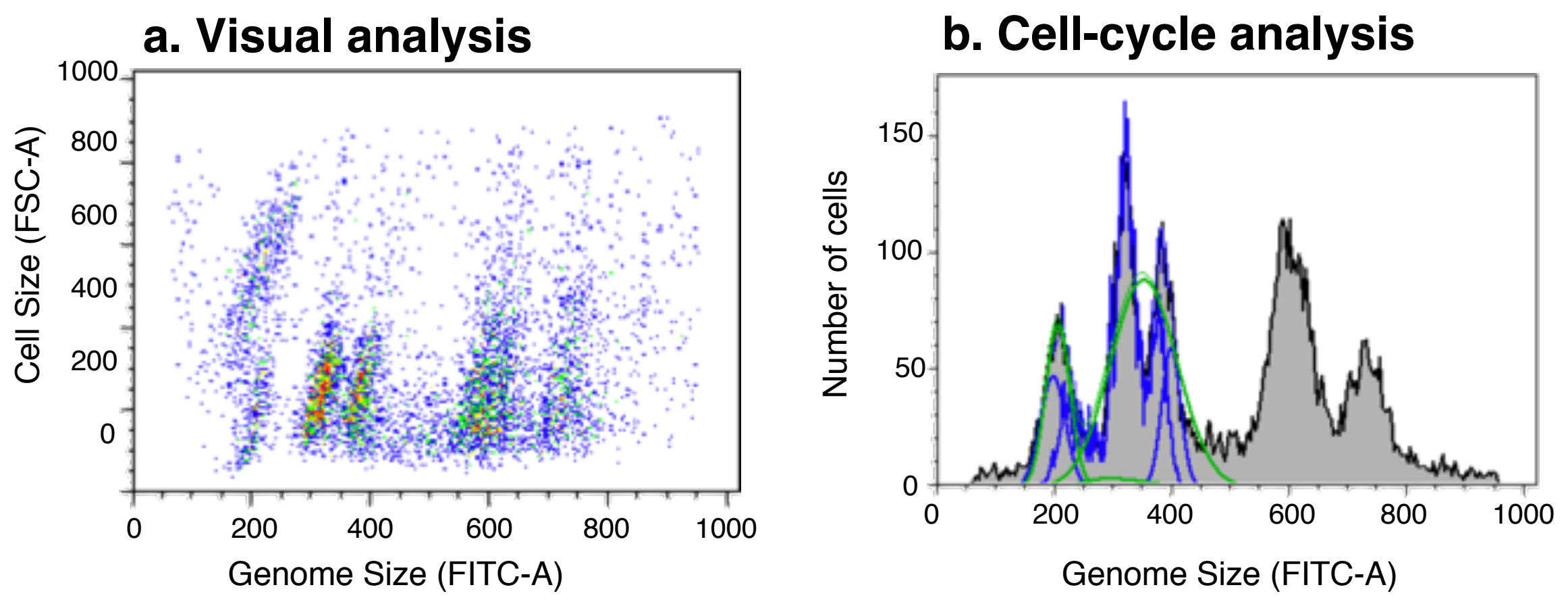
b. Manual cell-cycle analysis

Figure 3

METHODS

Proposed workflow for mixed-populations

- We will automate this manual process by using hierarchal clustering to identify the number of clusters instead of manual gating and use methods of least square error (*FitterError*) and the Dean-Jett-Fox algorithm (*CellCycle*) to identify the G0/G1 and G2/M pairs.
- This will eliminate human bias and will account for populations with more then two subpopulations. (Figure 4)


a. Visual analysis

b. Cell-cycle analysis

Figure 4

The previous workflow can not be applied when there are more then two subpopulations or the visual analysis is not easy to manually gate.

Aim 3: Implement the developed functions into *PloidyCluster* and release as a Bioconductor package.

- Write help documentation and a comprehensive user guide ("vignette") and upload the package to the R Bioconductor package repository.
- For each population, the user will be provided with:
  - An indication of model fit
  - The number of subpopulations
  - The number of cells within each subpopulation
  - Calculated genome size (G0/G1 mean).

SIGNIFICANCE

- Provide an unbiased measure of model fit
- Able to quantify populations that contain multiple subpopulations
- Can be applied in karyotype variation in cancer cells

ACKNOLEDGEMENTS

- I acknowledge the support of the Visual and Automated Disease Analytics (VADA) Program for their funding.
- Special thank you to Dr. Aleeza Gerstein and the all the members of the MicroStats lab.

Gerstein MICRO STATS Lab

The VADA Program
Visual and Automated Disease Analytics
Graduate Training Program