Agradecimientos

A mis papás. A mis amigos y profesores del CIMAT, en especial a José Miguel Ponciano y Marcos Capistrán. A Shalev Itzkovitz.

Gracias a CONACyT por los apoyos financieros recibidos para realizar esta tésis.

Abstract

Multicellular tissues are generally maintained and regenerated by populations of stem cells that undergo cell division and differentiation into specific cell types. In order to gain a quantitative understanding of the architecture of such tissues, it is important to characterize them according to the parameters governing the stem and differentiated cell population dynamics, such as division rates, death rates, and population sizes. Here we present a mathematical framework for quantitatively characterizing any such tissue. The framework consists of an analytically solvable model that captures the dynamics of stem and differentiated cell populations and an estimator of cell depth – the number of cell divisions undergone since the zygote – based on microsatellite mutation data. By fitting the model's solution to data on average cell population depth obtained using the estimator and data on cell population size, one can obtain estimates of the biological parameters of interest. We exemplify this framework by applying it to data on mouse oocyte population size and average depth.

I - Introduction

Many organs and tissues in multicellular organisms are maintained by pools of stem cells that continuously divide and differentiate into specific cell types. For example, the epithelial layer of the human colon is organized into crypts, with stem cells residing at the bottom of each crypt (Figure 1a). These epithelial stem cells divide asymmetrically into daughter cells: one stem cell and one differentiated epithelial cell (Yatabe et al. 2001). The daughter epithelial cell gradually migrates to the top of the crypt to replace older and damaged cells, while the daughter stem cell remains at the bottom of the crypt to help maintain the stem cell population. Similarly, a pool of hematopoietic stem cells resides in bone marrow (Figure 1b). This stem cell population is in charge of renewing cells of the blood and immune system (Silverthorn, 2006).

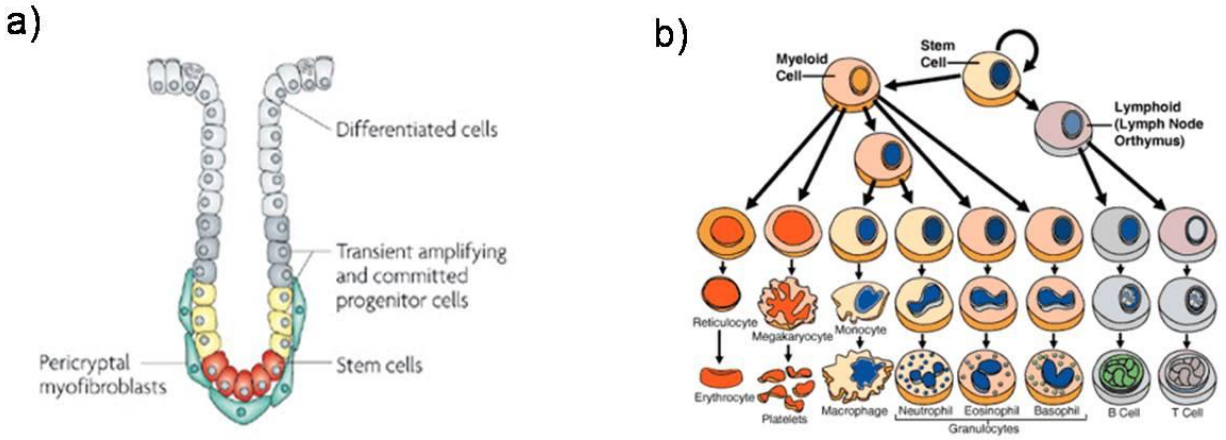


Figure 1: Maintenance of differentiated cell populations through stem cell pools. Differentiated cell populations of different organs, such as the crypts of the human colon (a) and the blood and immune systems (b) are maintained and regenerated through stem cell pools that continuously divide

and undergo differentiation into specific cell types. (Illustrations taken from www.medscape.com and www.diseaseeducation.com)

The existence of this tissue architectural design – where a stem cell pool maintains and regenerates a differentiated cell pool – can result in changes in the size of both the stem cell and differentiated cell populations with time. Importantly, it also leads to changes in the distribution of cell depths – the number of cell divisions undergone since the zygote (Wasserstrom et al., 2008) - of both populations, a fact that is illustrated in Figure 2. For instance, whenever a stem cell of depth *i* (i.e. which has undergone *i* divisions since the zygote) divides asymmetrically, the number of stem cells and differentiated cells of depth *i*+1 increase by one, while the number of stem cells of depth *i* decreases by one. Therefore the size and depth distribution of both cell populations is dictated by the dynamics of the stem cell population.

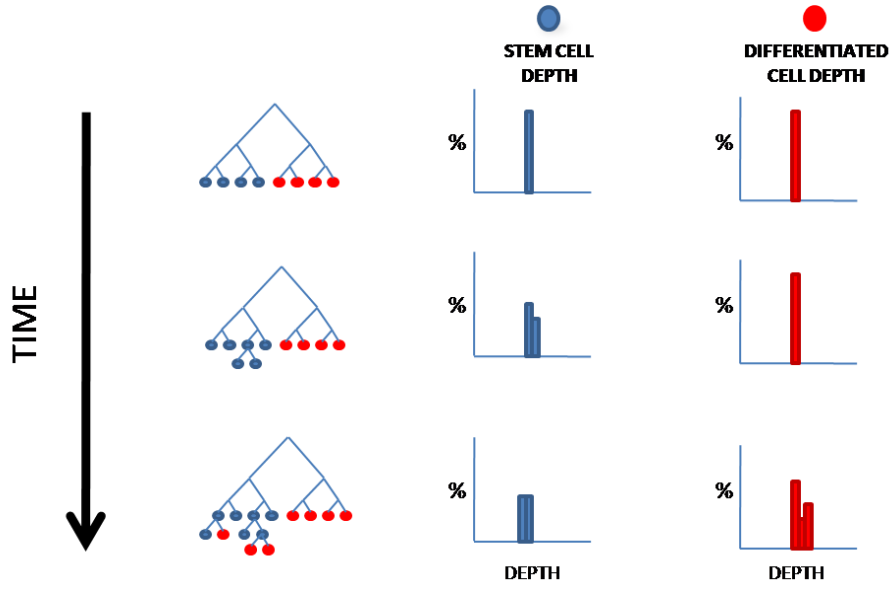


Figure 2: Changes in depth distribution due to stem cell pool division and differentiation. The figure shows a schematic diagram of stem (blue) and differentiated (red) cell populations belonging to a single organism and thus related by a cell lineage tree. As the stem cell population divides, the distribution of cell depth – the fraction of cells that have divided a specific number of times since the root of the tree (zygote) – changes for both cell populations.

Although a vast amount of information has been obtained regarding the structure and organization of stem and differentiated cell pools in diverse multicellular tissues (Polesskaya et al., 2003; Tulina et al., 2001; Zhang et al., 2003) it is important to gain a better quantitative understanding of the architecture and dynamics of such cell populations. Here we present a mathematical framework that allows the inference of biological parameters governing the dynamics of stem and differentiated cell pools, such as division and death rates, cell population sizes at birth, and average cell depth at birth. The framework consists of i) an analytically solvable model that captures the dynamics of the size and depth distribution of the cell pools and ii) an estimator of cell depth based on microsatellite (MS) mutation data. The model is analytically solvable using generating function techniques, and yields expressions for the size and average depth of the stem and differentiated cell populations as a function of time. These expressions can then be fit to experimental data on average cell depth - obtained using the MS mutation based estimator - and cell population size to obtain estimates of the biological parameters of interest. We illustrate the use of this framework by fitting the model to recent data on the number and average depth of mouse oocytes at different ages, obtaining rough estimates for the size, division rate, and death rates of a newly discovered oocyte stem cell population (Johnson et al., 2004; Johnson et al., 2005; Eggan et al., 2006; Zou et al., 2009).

The thesis is organized as follows. In section II we present the mathematical model and solve for the dynamics of both the size and the average depth of the stem and differentiated cell pools. In section III, we describe basic concepts of microsatellite mutations and, using simulations, compare two different estimators of cell depth based on microsatellite mutation data. The estimator with the best performance – a maximum-likelihood neighbor-joining estimator – has higher efficiency and smaller bias, and is robust to anomalous mutation events. In section IV we exemplify the value of the model and estimator by fitting the model's parameters to data on the number of oocytes and average oocyte depth as a function of age, the last of which was recently obtained using the maximum-likelihood neighbor-joining estimator. This allows us to obtain estimates of the division and death rates of the oocyte stem cell pool.

II – Mathematical Model

The model consists of a regenerating stem cell pool and differentiated cell pool of initial sizes Rt and Qt (Figure 3). Stem cells divide into two daughter cells at a rate α. Each stem cell division event yields two stem cells with probability p, two differentiated cells with probability p, and one stem cell and one differentiated cell (asymmetric division) with probability 1-2p. The fact that the two symmetric

division events occur with equal probability is consistent with the notion that stem cell pool must maintain a constant size in short timescales (Potten et al., 1990). Additionally stem cells die at a rate D_R.

A newly generated differentiated cell enters a transit amplifying stage, where it rapidly divides n times before entering a quiescent stage in which it no longer undergoes cell divisions. Thus every differentiation event yields $f = 2^n$ differentiated cells. Finally, differentiated cells in the quiescent stage disappear at a rate D_Q. This rate includes cell death as well as other cell depletion mechanisms, such as ovulation and atresia in the case of oocytes.

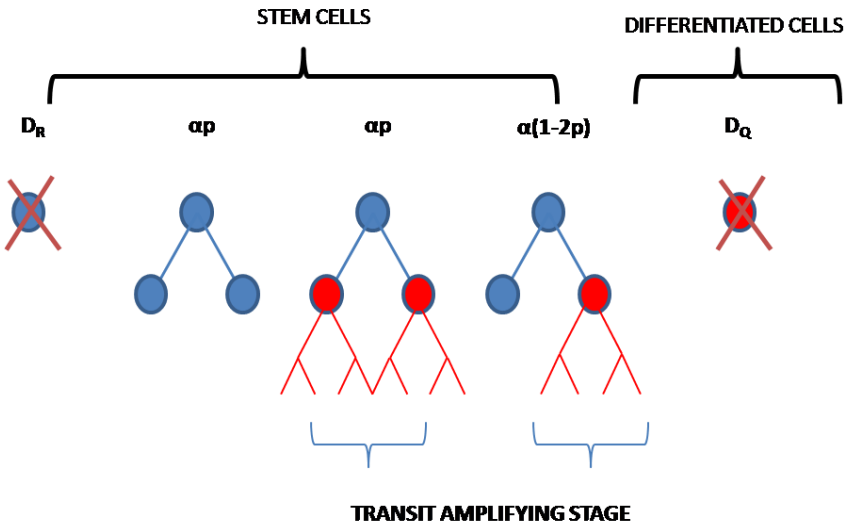


Figure 3: Stochastic model of stem and differentiated cell pool dynamics. The different events that occur in the stem and differentiated cell populations, which affect the size and depth distribution of both pools, include stem cell division (with rate parameter α) and cell death (with rate parameters D_R and D_Q). A stem cell division event can be symmetric (with probability 2p) or asymmetric (with probability 1-2p). Additionally, newly differentiated cells may enter a transit amplifying stage and undergo several more rounds of cell division.

We denote by R_i and Q_i the number of stem cells and quiescent differentiated cells respectively of depth i. For simplicity, all cells have an initial depth of i_0 divisions at birth. We will analyze a model with no transit amplifying stage, i.e. $n = 0$. The dynamics of the stem cell pool can be described by:

$$(1)\qquad \begin{aligned} \frac{dR_i}{dt} &= -(\alpha + D_R)R_i && i = i_0, \\ \frac{dR_i}{dt} &= 2p\alpha R_{i-1} + (1-2p)\alpha R_{i-1} - (\alpha + D_R)R_i = \alpha R_{i-1} - (\alpha + D_R)R_i && i > i_0. \end{aligned}$$

Equation (1) includes terms for population increase due to symmetric division of stem cells of depth $i-1$ into two stem cells, asymmetric divisions of cells of depth $i-1$, as well as terms corresponding to population decrease due to division and death of cells of depth i. Summing equation (1) over all values of i we obtain the dynamics of the total number of stem cells $R = \sum_{i=i_0}^{\infty} R_i$:

$$(2)\qquad \frac{dR}{dt} = -D_R R \Rightarrow R(t) = R_T e^{-D_R t}.$$

Similarly the dynamics of the (quiescent) differentiated pool can be described by:

$$(3)\qquad \begin{aligned} \frac{dQ_i}{dt} &= -D_Q Q_i \quad i = i_0, \\ \frac{dQ_i}{dt} &= 2p\alpha R_{i-1} + (1-2p)\alpha R_{i-1} - D_Q Q_i = \alpha R_{i-1} - D_Q Q_i \quad i > i_0. \end{aligned}$$

Equation (3) includes terms for population increase due to symmetric division of stem cells of depth $i-1$ into two differentiated cells, asymmetric divisions of stem cells of depth $i-1$, as well as terms corresponding to population decrease due to differentiated cell death. Summing equation (3) over all values of i we obtain the dynamics of the total number of differentiated cells $Q = \sum_{i=i_0}^{\infty} Q_i$:

$$(4)\qquad \frac{dQ}{dt} = \alpha R - D_Q Q \Rightarrow \frac{d(e^{D_Q t}Q)}{dt} = \alpha \operatorname{Re}^{D_Q t} = \alpha R_T e^{-(D_R - D_Q)t},$$

which leads to the following expression for the total number of differentiated cells as a function of age:

$$(5)\ Q(t) = \left(Q_T + \frac{\alpha R_T}{D_R - D_Q}\right)e^{-D_Q t} - \frac{\alpha R_T}{D_R - D_Q}e^{-D_R t}.$$

Equation (5) contains four free parameters. However, for some cell populations such as mouse oocytes, the size of the quiescent pool at birth (Q_T) has been measured. In that case, Equation (5) contains three free parameters – death rates of stem cells (D_R) and oocytes (D_Q), and an effective flux (αR_T).

We now obtain the solution for the dynamics of the distribution of cell depths for the stem cell and differentiated cell pools. For simplicity we shift the depth variable to i-i_0. We then obtain an equation for the rate of change of the fraction of stem cells of depth i, $P_i(t) = \frac{R_i(t)}{R(t)}$:

$$(6)\quad \begin{aligned} \frac{dP_i}{dt} &= \frac{1}{R}\frac{dR_i}{dt} - \frac{R_i}{R^2}\frac{dR}{dt} = \alpha P_{i-1} - \alpha P_i \quad i > 0, \\ \frac{dP_i}{dt} &= -\alpha P \quad i = 0, \end{aligned}$$

where we substituted using equations (1) and (2). We can interpret $P_i(t)$ as the discrete probability distribution of stem cell ages. Using (6) and the corresponding probability generating function $F(x,t) = \sum_{i=0}^{\infty} P_i(t)x^i$ we obtain the following differential equation:

$$(7)\quad \frac{\partial F}{\partial t} = \sum_{i=0}^{\infty}\frac{dP_i}{dt}x^i = -\alpha P_0 + \alpha\sum_{i=1}^{\infty}P_{i-1}x^i - \alpha\sum_{i=1}^{\infty}P_i x^i = \alpha(x-1)F(x,t),$$

which has the following solution:

$$(8)\quad F(x,t) = F(x,0)e^{\alpha(x-1)t} = e^{\alpha(x-1)t},$$

where we assumed that at t=0, all renewing cells belong to the single age group $i = 0$, i.e. $F(x,0) = P_0(0) = 1$. We can now solve for the temporal dynamics of the depth distribution of stem cells by using the relation $P_i(t) = \frac{1}{i!}\frac{\partial^i F}{\partial x^i}\Big|_{x=0}$:

$$(9)\quad P_i(t) = \frac{1}{i!}\alpha^i t^i e^{-\alpha t}.$$

Equation (9) suggests that the depth distribution of stem cells drifts to larger values with age, as seen in Figure 4:

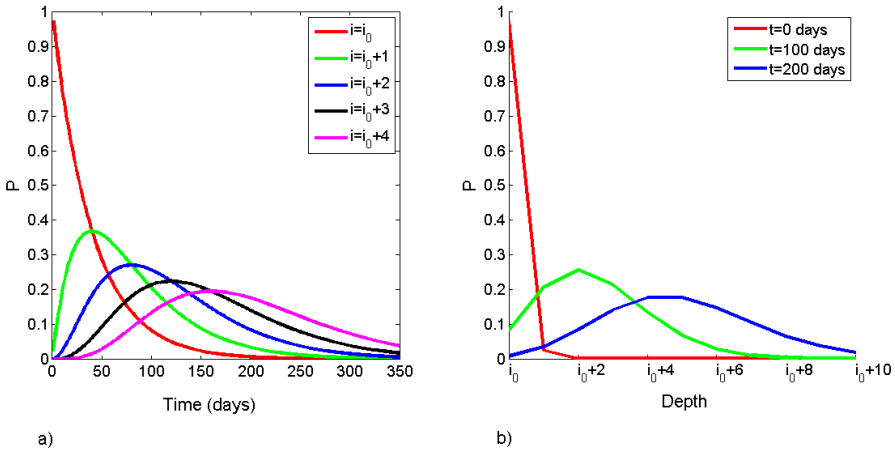


Figure 4: Dynamics of depth distribution of stem cells. a) shows the change in time in the fraction of stem cells with depth *i*, b) shows the depth distribution at different times. using α=1/40 days.

Using the expression for the generating function (8), we can now solve for the average depth as a function of time of the stem cell population:

$$(10)\quad \frac{d\langle i\rangle_R}{dt} = \frac{\partial}{\partial t}\left(\frac{\partial F}{\partial x}\right)\Big|_{x=1} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial t}\right)\Big|_{x=1} = \alpha F\big|_{x=1} + \alpha(x-1)\frac{\partial F}{\partial x}\Big|_{x=1} = \alpha$$
$$\Rightarrow \langle i\rangle_R(t) = i_0 + \alpha t.$$

Thus the average depth of the stem cell population increases linearly with age.

Thus far we have solved for the population size of both the stem cell and the differentiated cell pools (equations 2 and 5) and for the average depth of the stem cell population (equation 10). We now wish to obtain the dynamics of the average depth of the differentiated cell population. Using equations (3) and (9), and the relation $P_i(t)R(t) = R_i(t)$ we first solve the temporal dynamics of the number of quiescent differentiated cells at different depths:

$$(11)\quad \frac{d\left(e^{D_Q t}Q_i\right)}{dt} = \frac{\alpha R_T}{(i-1)!}\alpha^{(i-1)}t^{(i-1)}e^{-(\alpha+D_R-D_Q)t},$$

which has the following solution:

$$(12)\quad Q_i = \begin{cases} Q_T e^{-D_Q t} & i = 0, \\ \dfrac{\alpha R_T}{(i-1)!(D_R+\alpha-D_Q)^i}\alpha^{i-1}e^{-D_Q t}\Gamma(i,(D_R+\alpha-D_Q)t) & i > 0, \end{cases}$$

where $\Gamma(i,x) = \int_0^x t^{(i-1)}e^{-t}dt$ is the incomplete Gamma function. We can finally obtain the dynamics for the average differentiated cell depth:

$$(13)\quad \langle i \rangle_Q = \frac{\sum_{i=0}^{\infty} iQ_i(t)}{\sum_{i=0}^{\infty} Q_i(t)} = i_0 + \frac{\eta(e^{\eta t}-1)+\alpha\left((\eta t-1)e^{\eta t}+1\right)}{\eta(e^{\eta t}-1)+Q_T\eta^2/\alpha R_T},$$

where $\eta = D_Q - D_R$.

Equation (13) has four free parameters: η, α, αR_T, and i_0. In principle, two of these can be obtained by fitting equation (5) to data on the number of occytes as a function of time (see Section IV), leaving equation (13) with only two free parameters. These can then be obtained by fitting the model to data on average oocyte depth as a function of age.

We briefly mention that the dynamics of the stem cell and differentiated cell populations could alternatively be analyzed as a multivariate stochastic birth and death process (Allen, 2003). This would involve formulating birth and death rates for each population's depth group (i.e. the group of cells of a particular depth). For example, the birth and death rate of the group of regenerating cells of depth *i* would be given by

$$(14)\quad \begin{aligned} \lambda_{N_i} &= \alpha N_{i-1}, \\ \mu_{N_i} &= -(\alpha + D_R)N_i, \end{aligned}$$

where N_i is the size of the cell group of depth *i*. One could then solve the coupled Kolmogorov forward equations

$$\frac{dp_{N_i}}{dt} = \lambda_{(N-1)_i} p_{(N-1)_i} + \mu_{(N+1)_i} p_{(N+1)_i} - \left(\lambda_{N_i} + \mu_{N_i}\right) p_{N_i} \tag{15}$$

to obtain expressions for the probability distribution of the size of each depth group as a function of time, as well as its expectation value and variance.

Importantly we note that, due to the linearity of the deterministic model developed above, the solutions obtained (i.e. the fraction of cells belonging to each age group as a function of time) correspond to expectation values for the size of each depth group in the stochastic model formulation. To verify this we simulated the stochastic dynamics of stem cell and differentiated cell pools using the Matlab programming framework. Figure (5) shows the results obtained in a single instantiation of the simulations, illustrating that the analytical solution fits the average dynamics of population size and mean depth.

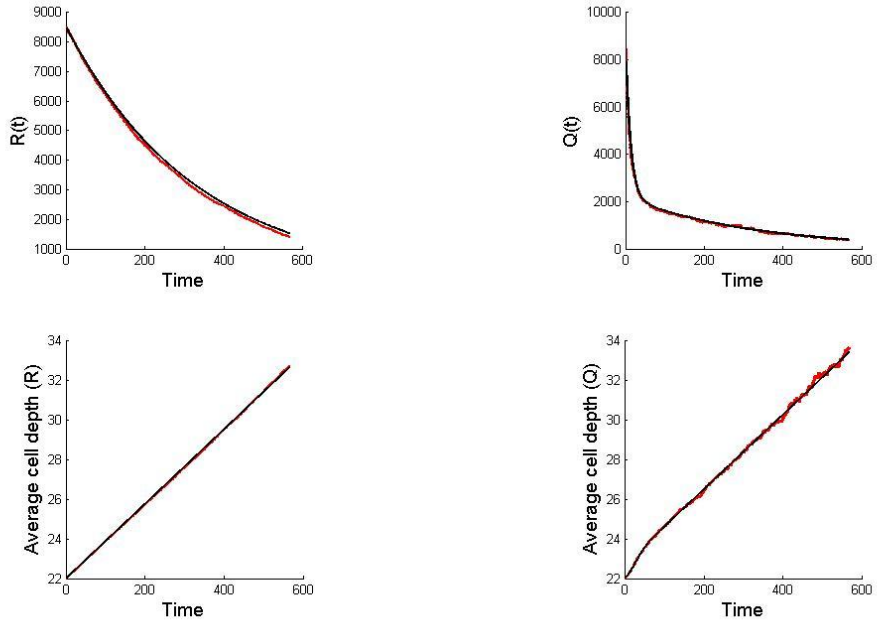


Figure 5: Stochastic simulation of stem cell and differentiated cell population dynamics. A stochastic simulation of the model developed above was performed using the Matlab programming environment. The red line shows the simulation results for population sizes and average depths of both the stem and differentiated cell pools. The black line shows the analytical solution to the deterministic model presented here. The following parameter values were used for the simulation: $\alpha = 1/53$ (days^{-1}), $D_R = 1/330$ (days^{-1}), $D_Q = 1/13$ (days^{-1}), $R_T = Q_T = 8500$ (cells), $i_0 = 22$ (cell divisions).

To summarize, in this section we developed a mathematical model for the size and depth dynamics of a stem and differentiated cell populations. We analytically solved for the size of both pools as a function of age (Equations 2 and 5). Additionally, using generating function techniques, we solved for the average depth of both populations as a function of time (Equations 10 and 13). Finally we showed, using stochastic simulations, that the analytical solutions obtained correspond to the expectation value for the size of each age group in the stochastic birth and death process formulation of the model. In Section IV we will use equations (5) and (13) to estimate the values of the free parameters for the case a mouse oocyte population by fitting the model to data on oocyte number and average depth as a function of age.

III - Estimating Depth from Microsatellite Mutation Data

The model developed above yields an analytical expression for the distribution of cell depth as a function of time. In order to use this model to obtain estimates of cell population parameter values for multicellular tissues, we need to fit the expressions to experimental data on average cell depth as a function of time. Recent developments in detecting microsatellite length differences of single cells taken from the same organism (Frumkin et al., 2005, Wasserstrom et al., 2008) allow estimating average cell depth of a cell population as a function of time. In this section, we first explain basic notions of microsatellite mutations, including the Single Step Mutation (SSM) model and its associated transition probability. We then describe a typical cell sampling experiment and compare two different estimators of cell depth according to their efficiency and bias. The estimators, in contrast to previously developed model-free estimators (Wasserstrom et al., 2008), are based on the SSM model of microsatellite mutation. We show through simulations that the maximum-likelihood neighbor-joining estimator has highest efficiency and lowest bias, and is more robust to mutation events that deviate from the SSM.

Microsatellite Mutations: Basic Concepts

Microsatellites are repetitive DNA elements dispersed throughout an organism's genome. They are generally defined as consisting of a basic repeat unit of six base pairs or less, and they vary in repeat number from a few dozen to hundreds of repeats. One example is $(AT)_n$, where AT is the basic repeat unit and n is the number repeats. A single eukaryotic genome can contain hundreds of thousands of

distinct microsatellite loci of different length and basic repeat unit dispersed throughout its chromosomes (Frumkin et al., 2005).

Importantly, during somatic cell division, microsatellites undergo stochastic changes in their repeat number. The mechanism underlying this change in repeat number is generally considered to be complementary strand mispairing due to strand slippage during DNA replication by the enzyme DNA polymerase (Figure 6) (Ellegren, 2000). The rates at which such changes in microsatellite repeat number occur vary among different organisms but are generally in the order of 10^{-5} (probability of mutation per locus per cell division event) (Frumkin et al., 2005). In cells with mutations in the Mismatch Repair System (MMR – the genetic system in charge of repairing slipped strand mispairing events), this rate is highly increased to the order of 10^{-2}.

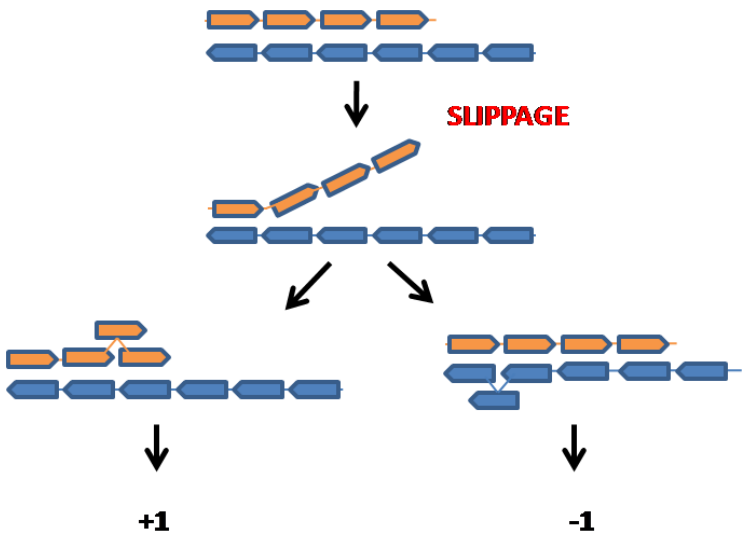


Figure 6: Single-step model of microsatellite mutation. During DNA replication prior to cell division, the leading edge replicating strand (orange) stochastically slips and loses contact with the complementary bases of the template strand (blue). Due to the repetitive nature of microsatellite sequences, base mispairing might occur, leading to either an increase or decrease of repeat number in one of the daughter strands.

The fact that repeat number changes are associated to cell divisions allows one to use microsatellites as molecular clocks to estimate the number of cell division events that a cell has undergone since the existence of the zygote (Figure 7). Towards this end, several models of

microsatellite mutations have been developed, one of which is the symmetric Single Step Mutation (SSM) model (Ohta and Kimura, 1973). Under the symmetric SSM model, a microsatellite increases or decreases its repeat number by one with probability mu per cell division event. Thus the SSM model is equivalent to a random walk.

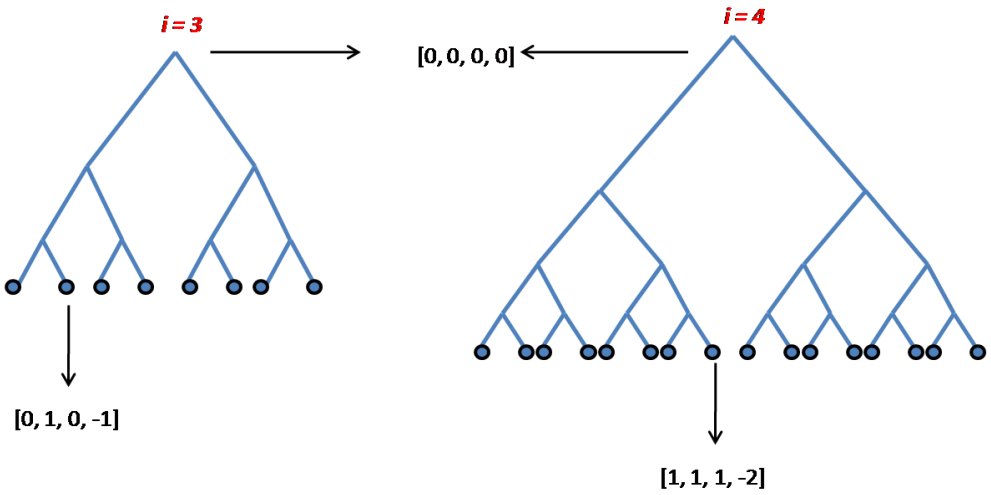


Figure 7: Use of microsatellites as molecular clocks. The figure shows two schematic cell populations related by lineage trees of different depth. Cells are sampled from each population, and the microsatellite repeat numbers of four loci are measured, where loci of the root of each tree is considered to be of length zero. The deeper cells tend to show more MS mutation events, allowing the use of such loci as molecular clocks to estimate cell depth (Wasserstrom et al., 2008).

Comparison of Two Different Depth Estimators

In a typical cell sampling experiment several (*s*) single cells are sampled from a single organism and modern molecular biology techniques (see Frumkin et al., 2005) are used to measure the length of *l* microsatellite loci in each cell. The resulting data set is a microsatellite length profile for each cell; *s* vectors of length *l*. This data set is amenable to statistical analysis to estimate the number of divisions undergone by each cell. Here we compare two different estimators of cell depth that are based on the SSM model. Both estimators are used for a typical cell sampling experiment and are based on lineage tree reconstruction via the Neighbor-Joining algorithm. However, each method uses a different distance estimate, where distance refers to the number of cell divisions separating any pair of sampled cells. The first distance estimate is the mean squared deviation in allele sizes; the second is a maximum likelihood estimate based on the symmetric SSM model of microsatellite mutations. We show that the maximum

likelihood Neighbor-Joining method provides the most efficient and less biased depth estimation, which is also the most robust to deviations from the symmetric step-wise model.

The mean squared difference estimate for the distance between cells *i* and *j* is given by:

$$(16) \qquad d(i,j) = \frac{1}{l}\sum_{k=1}^{l}\left(x_k^i - x_k^j\right)^2$$

where x_k^i and x_k^j are the number of repeats in the k-th microsatellite allele of the *i*-th and *j*-th sampled cells, respectively, and l is the total number of alleles. Since the SSM model of microsatellite mutation is equivalent to a *random walk* (Felsenstein, 2003), $d(i,j)$ is on average proportional to μt, where *t* is the number of cell divisions separating the two cells, and μ is the mutation rate per cell per division event.

An alternative estimator for the number of divisions separating two cells is the maximum likelihood estimator. The transition probability for the one-step microsatellite mutation model yields an expression for the maximum likelihood estimate of the number of divisions separating two cells (Felsenstein, 2003; Watkins 2007). Given a mutation rate per cell per division event μ, the transition probability is the likelihood that after *t* cell divisions the repeat number of a microsatellite locus will increase by *j* repeat units. Assuming symmetric mutations (equal probability of increasing and decreasing the repeat number) and following (Felsenstein, 2003) we write the transition probability as

$$(17) \qquad P(j;t) = e^{-\mu t}\sum_{k=0}^{\infty}\left(\frac{\mu t}{2}\right)^{j+2k}\frac{1}{(k+j)!\,k!}$$

Let $d_{i,j} \equiv \left\{d_1^{i,j}, d_2^{i,j}, ..., d_l^{i,j}\right\} = \left\{\left|x_1^i - x_1^j\right|, \left|x_2^i - x_2^j\right|, ..., \left|x_l^i - x_l^j\right|\right\}$ be the vector of absolute repeat number differences between cell i and cell j. Then, assuming equal mutation rates for all microsatellite loci, the log-likelihood of observing that particular data set is given by the following expression:

$$(18) \qquad \log\left(L(D_{i,j} \mid \mu t)\right) = \sum_{\beta=1}^{l}\log\left(e^{-\mu t}\sum_{k=0}^{\infty}\left(\frac{\mu t}{2}\right)^{d_\beta^{i,j}+2k}\frac{1}{\left(k+d_\beta^{i,j}\right)!\,k!}\right),$$

where the above expression is obtained by taking the natural logarithm of the product of the likelihood functions for each MS locus. In practice, this expression can be numerically evaluated for reasonable

values of *k* and for a range of values of *t*. Then one can find the value of *t* in this range that maximizes the log-likelihood for a given microsatellite repeat number data.

Equations (16) and (18) allow one to estimate the number of cell divisions between pairs of cells in the sample set. In order to use this information to obtain for each cell an estimate of cell depth, the cell lineage tree relating the sampled cells is reconstructed via the Neighbor-Joining algorithm (Saitou and Nei, 1987). The depth of each cell can then be read off the tree as the sum of branch lengths between each cell and the root of the tree.

The Neighbor-Joining algorithm we applied has the following stages (Felsenstein, 2003):

1. For each cell in the sampled set compute $u_i = \sum_{j \neq i}^{n} \frac{D_{i,j}}{n-2}$.

   a. For the mean squared difference estimate, let $D_{i,j} = \frac{1}{l}\sum_{k=1}^{l}\left(x_k^i - x_k^j\right)^2$ where x_k^i and x_k^j are the repeat number of the *k*-th microsatellite locus of the *i*-th and *j*-th cells in the sample.

   b. For the maximum likelihood estimate, $D_{i,j} = t$, where t is the number of divisions obtained by maximizing the log-likelihood function (18).

   2. Choose cells i and j for which $D_{i,j} - u_i - u_j$ is smallest.

   3. Join cells i and j and compute the distance from each cell to the new node:

$$v_i = \frac{1}{2}D_{i,j} + \frac{1}{2}\left(u_i - u_j\right)$$

$$v_j = \frac{1}{2}D_{i,j} + \frac{1}{2}\left(u_j - u_i\right).$$

   4. Compute the distance between the new node, (i, j) and all other cells or nodes:

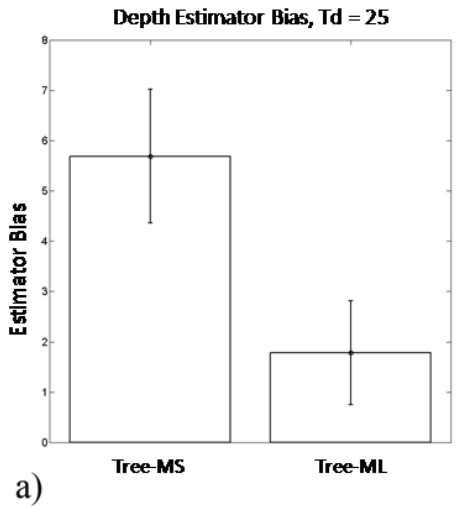
$$D_{(i,j),k} = \frac{\left(D_{i,k} + D_{j,k} - D_{i,j}\right)}{2}$$

   5. Repeat until only one node remains.

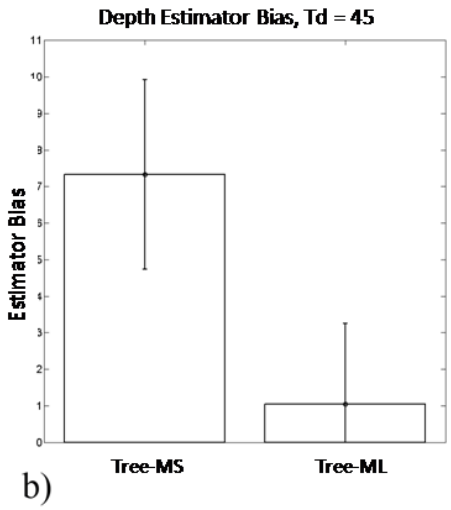
   6. Reroot the tree using the median of all sampled cells.

The performance of both estimators was evaluated by simulating random microsatellite mutations over cells of fixed depth (either 25 divisions or 45 divisions) and computing the efficiency and bias of the different depth estimates. An estimator is more efficient the smaller the variability in its values. Mutations were simulated according to the SSM model, where at every cell division a microsatellite locus *i* was assigned a small probability μ_i of increasing or decreasing its repeat number by one. Mutation rates were calibrated according to the mutation rate estimates obtained from mouse ex-vivo trees (Frumkin et al., 2005). Additionally, signal dropout, where the length of a microsatellite allele in a particular cell cannot be measured due to amplification failure, was incorporated into the simulations using the dropout rates of experimental data. Finally, we evaluated the estimators' robustness to scenarios where microsatellite mutations deviate from the single-step model, and to stochastic errors in signal analysis that result in significantly high repeat numbers. To measure robustness to non-single-step events, we simulated microsatellite mutations according to a two-step model (Di Rienzo, 1994), where the size of the step is geometrically distributed with parameter β. To test robustness to signal errors, which may occur due to noisy capillary signal or poor allele separation, for each sampled cell we randomly chose a set of loci at different rates and added relatively large repeat numbers for these loci.

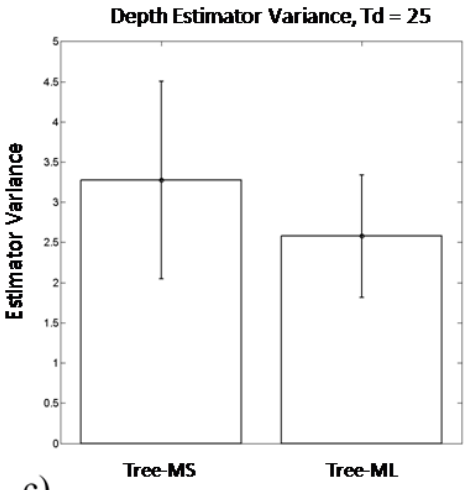
We find that the maximum likelihood depth estimation is more efficient and less biased than the mean square estimate (Figure 8 a-d). In addition the method is more robust to deviations from a single-step model and to low rates of signal errors (Figure 8 e-f). We therefore chose this method for estimating cell depth from microsatellite mutation data.


Depth Estimator Bias, Td = 25
Estimator Bias
Tree-MS
Tree-ML

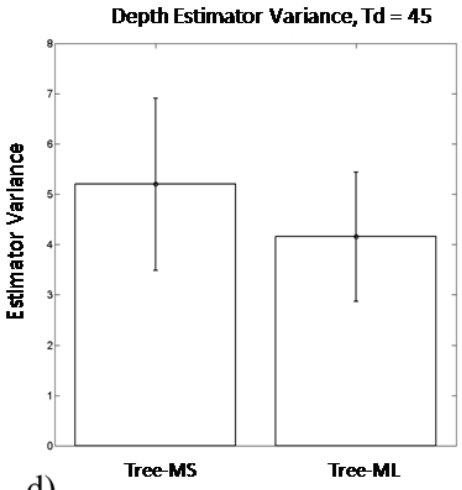
a)


Depth Estimator Bias, Td = 45
Estimator Bias
Tree-MS
Tree-ML

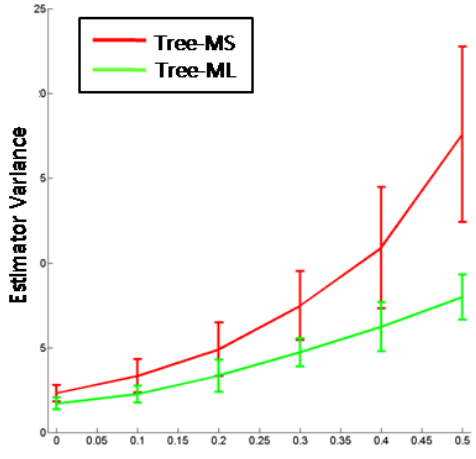
b)


Depth Estimator Variance, Td = 25
Estimator Variance
Tree-MS
Tree-ML

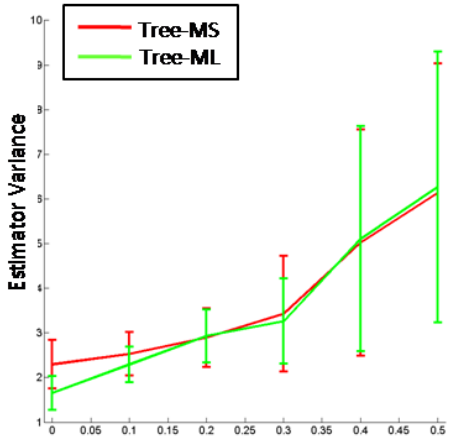
c)


Depth Estimator Variance, Td = 45
Estimator Variance
Tree-MS
Tree-ML

d)


Tree-MS
Tree-ML
Estimator Variance
Rate of deviation from single step mutations

e)


Tree-MS
Tree-ML
Estimator Variance
Rate of signal error

f)

Figure 8: Comparison of depth estimator methods. The estimator bias (a,b) efficiency (c,d) and robustness to deviations from the mutation model (e) and to signal errors (f) is shown. Results are for cells sampled from a binary tree of depth Td=25 (a,c,e,f) and Td=45 (b,d).

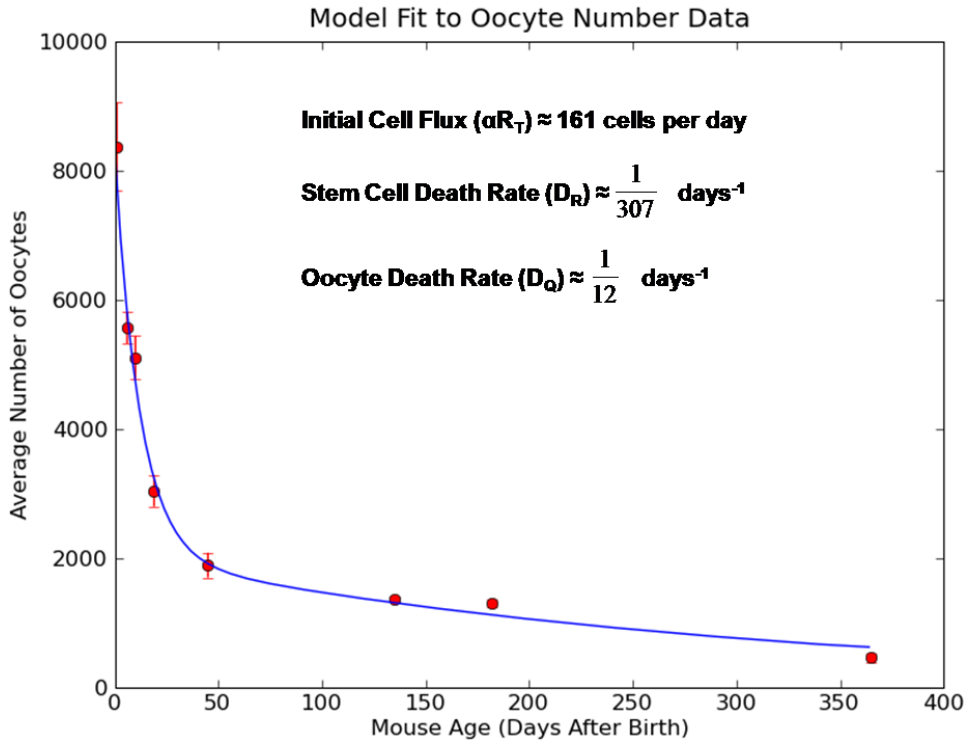
IV - Model Parameter fit to Experimental Data

In this section we illustrate the value of the model and depth estimator by obtaining estimates of five biological parameters governing the dynamics of a mouse oocyte cell pool. Specifically, we fit equation (5) to data on oocyte number as a function of age (Bristol-Gould et al., 2006) yielding estimates for initial stem cell flux (αR_T), stem cell death rate (D_R), and differentiated cell death rate (D_Q). These three parameter estimates, along with a recently obtained data set on average oocyte depth as a function of age, are then used along with equation (13) to obtain values for the initial differentiated pool depth (i_0) and the stem cell division rate (α). This last estimate allows decoupling the two factors from the stem cell flux (αR_T), yielding the size of the stem cell pool at birth (R_T).

We performed the nonlinear least squares parameter fits using the Levenberg-Marquardt algorithm implemented in the optimization package of the Python Scipy module (Marquardt, 1963). The Levenbeg-Marquardt algorithm interpolates between the Gauss-Newton Method and the method of Gradient Descent through the use of a damping factor λ in the linear equation for the step δ. The algorithm was initialized by choosing biologically reasonable values for the different parameters. It would be interesting to perform a global minimization algorithm such as DIRECT search (Kolda, 2003) to explore whether other parameter values that minimize the error function exist. Confidence intervals were obtained for the parameter estimates by performing the fitting algorithm using the upper and lower bounds of the depth and oocyte number data. It would also be interesting to obtain confidence intervals through a different method, such as parametric bootstrapping for the cell depth data using the simulations of the one-step model of MS mutation.

Figure (9) shows the parameter estimates obtained using the oocyte number and average depth data sets. At age t = 0 days, 161 stem cells divide per day (C.I. 142-180). Stem cells live on average 307 days ($1/D_R$) (C.I. 290-325), while oocytes live on average 12 days ($1/D_Q$) (C.I. 10-14). Finally,

differentiated oocytes have an average cell depth of 22 cell divisions at birth (C.I. 21-23), and stem cells divide approximately once every 70 days (1/α) (C.I. 63-80).



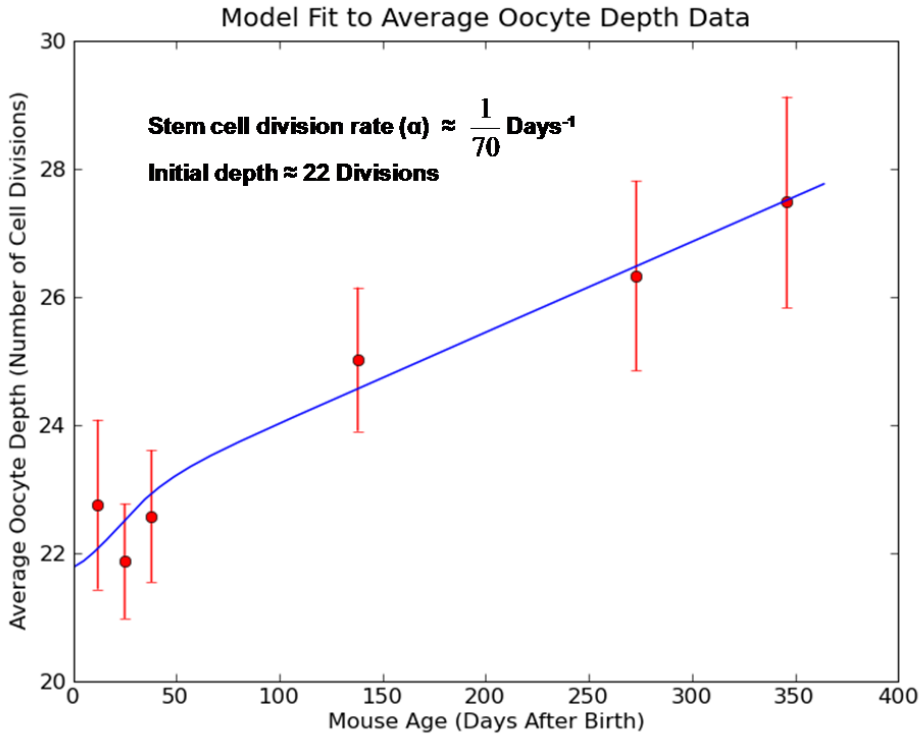


Figure 9: Model fit to oocyte number and average cell depth as a function of mouse age. Data on mouse oocyte number as a function of age was obtained from Bristol-Gould et al. 2006. We fit the three free parameters of Equation (5) (Section II) to this data, obtaining estimates for stem and oocyte death rates as well as the initial cell flux. Data on average oocyte depth as a function of age was recently obtained using microsatellite mutation data and the depth estimation method outlined in section III. Error bars correspond to the standard error of the mean (SEM) obtained by grouping two mice from the same age-group. We fit equation (13) (Section II) to this data to obtain stem cell division rate and initial oocyte depth.

V – Conclusion

In this thesis we developed a mathematical model to describe the dynamics of the size and depth distribution of general stem cell and differentiated cell populations. The model, consisting of two sets of coupled differential equations, is analytically solvable using generating function techniques. Using stochastic simulations, we showed that the solution accurately describes the dynamics of the expectation values of population size and average depth. We then analyzed two estimators of cell depth according to their efficiency and bias. We found that the maximum-likelihood Neighbor-Joining (ML-NJ) estimator has higher efficiency and smaller bias.

We illustrated the value of the model and the depth estimator by performing a nonlinear best fit to data on oocyte average depth and population size as a function of mouse age. The fit yields estimates for the parameters governing the dynamics of an oocyte stem cell population. Although the existence of the oocyte stem cell pool was recently confirmed (Zou et al., 2009), it would be interesting to experimentally verify the range of values for the parameter estimates.

Finally we emphasize that the framework developed here to estimate cell population parameters is applicable to any multicellular tissue that is organized into a stem cell pool that maintains and replenishes a differentiated cell pool. Therefore it is in principle feasible to utilize this framework to characterize the dynamics governing diverse multicellular organs.

References

Allen, L. *An Introduction to Stochastic Processes with Biology Applications.* Prentice Hall (2003).

Bristol-Gould, S. K. et al. Fate of the initial follicle pool: empirical and mathematical evidence supporting its sufficiency for adult fertility. *Dev Biol* **298**, 149-54 (2006).

Di Rienzo, A.D., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., and Freimer, N. B. Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA*, **91**, 3166-70 (1994).

Eggan, K., Jurga, S., Gosden, R., Min, I. M. & Wagers, A. J. Ovulated oocytes in adult mice derive from non-circulating germ cells. *Nature* **441**, 1109-14 (2006).

Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* **16**, 400-402, (2000)

Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates (2003).

Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol* **1**, e50 (2005).

Johnson, J., Canning, J., Kaneko, T., Pru, J. K. & Tilly, J. L. Germline stem cells and follicular renewal in the postnatal mammalian ovary. *Nature* **428**, 145-50 (2004).

Johnson, J. et al. Oocyte generation in adult mammalian ovaries by putative germ cells in bone marrow and peripheral blood. *Cell* **122**, 303-15 (2005).

Kolda, T.G., Lewis, R.M., Torczon, V. Optimization by Direct search: new perspectives on some classical and modern methods. *SIAM Review*, **45** 385-482 (2003).

Marquardt, D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics* **11**: 431–441 (1963).

Ohta, T., and Kimura, M. The model of mutation appropriate to calculate the number of electrophoretically detectable alleles in a genetic population. *Genet Res* **22,** 201-204 (1973).

, A., P. Seale, Rudnecki, M. A. Wnt signaling induces the myogenic specification of resident $CD45^+$ adult stem cells during muscle regeneration. *Cell* **113**: 841–852 (2003)

Potten, C. S. & Loeffler, M. Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt. *Development* **110**, 1001-20 (1990).

Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25 (1987).

Silverthorn, D. U. *Human Physiology: An Integrated Approach*. Benjamin Cummings (2006).

Tulina, N., Matunis, E. Control of stem cell self-renewal in *Drosophila* spermatogenesis by JAK-STAT signaling. *Science* **294**: 2546–2549 (2001).

Wasserstrom, A. et al. Estimating cell depth from somatic mutations. *PLoS Comput Biol* **4**, e1000058 (2008).

Watkins, J. C. Microsatellite evolution: Markov transition functions for a suite of models. *Theor Popul Biol* **71,** 147-59 (2007).

Yatabe, Y., Tavare, S., Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc Natl Acad Sci U S A* **98**, 10839-844 (2001).

Zhang et al. Identification of the haematopoietic stem cell niche and control of the niche size. *Nature*, **425,** 836-841 (2003).

Zou, K. et al. Production of offspring from a germline stem cell line derived from neonatal ovaries. *Nat Cell Biol* **11**, 631-6 (2009).