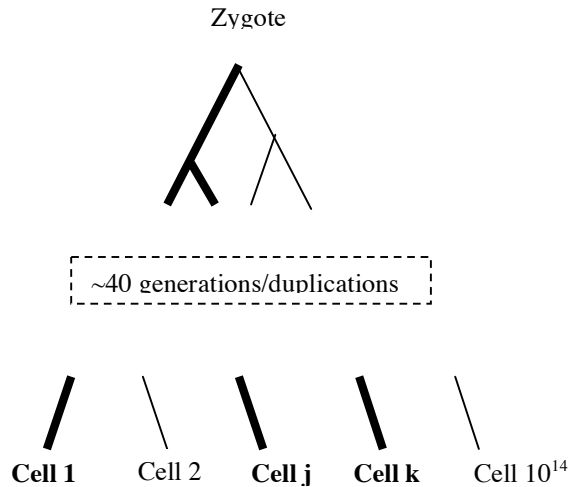


Pilot Study: Somatic Cell Genealogies and Differentiation

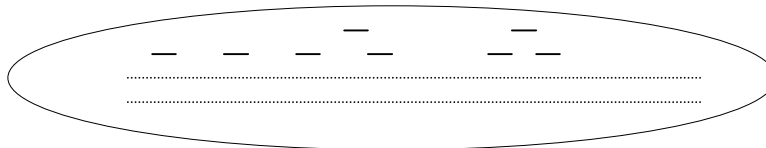
This is a 2-month pilot study in the sense that it is an investigation that will evaluate the feasibility of a certain approach to some big biological questions. It is also quite diverse in the sense that it will imply reading up on somatic mutation rates, phylogeny reconstruction, the biology of the proposed hypotheses and performing simulations evaluating the amount of data needed to distinguish between the hypotheses.

An individual consists of a large number of cells that has been created by a series of duplication that can be traced back to the zygote. The relationship of all cells in an individual can this be described by a traditional phylogeny:



The cell genealogy is enormous with 10^{14} leaves (present cells) and dwarfs the tree of life (with species at its leaves) by a probably a factor of 10^6 . Only a sub-sample of cells (illustrated in **bold**) will be available defining a **sub-genealogy** of the total genealogy. Properties of the leaf-cells can be available. Additionally, genealogies can be available from different species, individuals or developmental stages. Such information creates challenges in combining information in a useful way. A key question is clearly also recoverability of the sub-genealogy from observations in the cells at the leaves.

Traditionally cells of an individual has been viewed as identical (with few exceptions such as cells involved in immune response, egg and sperm cells), but a cell and its duplicated offspring can differ in several heritable properties: Chromosome rearrangements, CG-methylation, microsatellite repeat numbers. (It has for instance been calculated that each cell division is accompanied by about 50 errors in micro-satellites) and Copy Number Variation (CNV). Thus knowing the genomic sequence for all cells would allow the recovery of the cell phylogeny at high degree of certainty (Frumkin et al., 2005). Although for instance 50 events/cell replication could be a high estimate (for this purpose optimistic), the large number of cells descending from a given ancestor could guarantee that the phylogenetically necessary variation always was available. If somatic mutation rate is low, this could be compensated for by more sequencing.



Each cell will have 2 sets of DNA each $3 \cdot 10^9$ bp long – shown as dashed lines. The new techniques will allow determination of random segments on a large scale – shown as little segments. Since the genome is known, the position of the segments can generally be determined. Since comparison is key to this project, there must be enough coverage so that enough common DNA is known for different cells.

There are a series of merging technologies that can yield relevant sequence material on a large scale (Khanna, 2007) that already has the ability to obtain sequences covering 10^8 - 10^9 nucleotides for ~5K€ and price will come down.

Since there are about 10^{14} cells in an adult human, even with much improved sequencing technologies, choosing a strategic smaller set will be important. Knowing the cell tree in conjunction with other data could be a major tool in understanding differentiation. It is clear that this domain of research will experience a major growth in coming years and engaging in both

statistical modelling and empirical research in this field, will be of great value. In this project, we propose research involving both.

Fundamental Questions:

This project doesn't have a theoretical focus, but this kind of data gives rise to many new problems that eventually should be addressed:

- Which is the most efficient way to analyze the data? Clearly the overlap in determined sequences for all cells will be small so the complete tree must be patched together from incomplete data.
- Which conventional tests for trees would be of interest in this setting? Molecular clock? Similar rates for different processes?
- Which branching process has generated the tree?
- For instance *C. elegans* have determinate evolution ie. all individuals have the same cell tree. This is not the case for humans, but the trees can vary to a smaller or larger degree. Quantifying this would be very interesting.
- Traits and positions are associated different cells and leads to new problems: Are different tissues/organs monophyletic? What is the inferred spatial trajectory of cells during ontogeny?

Biological Test Case and Data Analysis:

Clearly this approach has potential in the analysis in any tissue – healthy as well as pathological (for instance cancers). In this pilot project we have chosen to zoom in on a smaller feasible project involving a specific project involving neurological diseases: The human nervous system consists almost exclusively of post-mitotic cells. An estimated 10^{12} cells make 10^{14} synaptic connections. Given that these cells cannot be regenerated or replaced, 'errors' or variations at the level of the genome are likely to be less well tolerated than in tissues with high cell turnover such as blood or epithelia. Age-dependent neurodegenerative disease, where cell death is often anatomically restricted (eg; cortex, motor system, cerebellum etc) may arise due to genomic variations which arise through somatic evolution in cell groups which are ontogenetically related. Analysis of somatic cell genealogy in the brain may therefore give important clues to the anatomical distribution of neurodegeneration and provide a window through which to observe a stochastic contribution to neurodegenerative disease, information which cannot be derived from analysis of DNA from whole blood. An unexpectedly high degree of heritable genetic variation in both coding (eg; copy number polymorphism) and non-coding DNA is emerging from whole genome studies in populations. It is plausible that the underlying biological factors driving this variation through meiotic divisions also operate during somatic cell division. There are recently identified examples where copy number variation can give rise to Parkinson's Disease.

One possible experimental test of this hypothesis would be to take a population of mice all derived from one embryonic stem cell (ie a series of clones) and to sample specific brain regions (cortex, cerebellum, hippocampus, spinal cord, etc) either by dissecting out groups of cells (likely to be >1000) or individual cells (by laser capture micro-dissection) and analyse these for genetic variation. The current limitations to this idealised experimental approach are at many levels: obtaining cloned mice, LCD is time consuming, DNA yield is too low, analysis is still quite expensive. However as a prelude to an experimental approach theoretical modelling should address the following: i. with different estimates of mutation rates (point mutations, large scale duplications etc.....precise rates for any of these are unknown currently) what is the probability of a population of 10^{12} cells all derive d from one progenitor (an assumed pan-neuronal progenitor in the developing embryo) if sampled at a low rate yielding an output that could allow construction of a genealogical tree? ii. What would be the effect of modelling in a series of cloned animals?

2 month pilot project:

Week 1: Read references below. The books should only be used for reference.

Week 2: Plan the remaining period. Following will have to be addressed: number of cells sampled (n – probably less than 10), number of clones sampled from each cell (k : 10^6 - 10^8), length of each clone (L : 15-500), shape of tree to simulate under, rates of the involved somatic mutations.

Week 3-4: Write Phylogeny-Genome simulator. A diploid 1 chromosome genome can assumed for simplicity. This is related according to the chosen phylogeny, on each branch the mutational process is applied generating leaf-cell genomes. The leaf cell genomes are cover with clones whose sequences can be known with a certain error.

Week 5-6: Simulations are performed to check the program and get a feeling for key quantities: How many clones of a given length is needed to obtain a certain level of overlap between different genomes? How many similar cells must be sequences to bring error rate of observed differences down to an acceptable level?

Week 7-8: Application of simulation to specific hypothesis and final report write-up. Report writing is encouraged to be done continuously during the period.

References

- Bender A, Krishnan KJ, Morris CM, Taylor GA, Reeve AK, Perry RH, Jaros E, Hersheson JS, Betts J, Klopstock T, Taylor RW, Turnbull DM. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nat Genet.* 2006 May;38(5):515-7
- Felsenstein (2002) "Inferring Phylogenies" Sinauer
- Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. "Genomic Variability within an Organism Exposes Its Cell Lineage Tree." *PLoS Comput Biol.* 2005 Oct;1(5):e50. Epub 2005 Oct 28
- Khanna VK (2006) "Existing and emerging technologies for DNA finger printing, sequencing, bio- and analytical chips: A multidisciplinary approach unifying molecular biology, chemical and electrical engineering" *Biotech. Adv.* 25:85-98.

- Kim JY, Tavaré S, Shibata D. "Human hair genealogies and stem cell latency." *BMC Biol.* 2006 Feb 3;4:2.
- Kim JY, Tavaré S, Shibata D. "Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions." *Proc Natl Acad Sci U S A.* 2005 Dec 6;102(49):17739-44. Epub 2005 Nov 28.
- Ro, S. and B.Rannala (2001) "Methylation patterns and mathematical models reveal dynamics cell turnover in the human colon" *Proc.Natl.Acad.Sci.*98.10519-21.
- Semple C and M Steel (2003) "Phylogenetics" OUP
- Shibata D, Tavaré S. "Counting divisions in a human somatic cell tree: how, what and why?" *Cell Cycle.* 2006 Mar;5(6):610-4. Epub 2006 Mar 15. Review.
- Watase K, Venken KJ, Sun Y, Orr HT, Zoghbi HY. Regional differences of somatic CAG repeat instability do not account for selective neuronal vulnerability in a knock-in mouse model of SCA1. *Hum Mol Genet.* 2003 Nov 1;12(21):2789-95