Cancer Evolution:
from processes to genomic patterns and back again

International Agency for Research on Cancer
Lyon, France

Nicolas Alcala, PhD
Scientist, Genetic Cancer Susceptibility Group, Section of Genetics

December 1st, 2020

Reminder: cancer evolution shapes observed heterogeneity

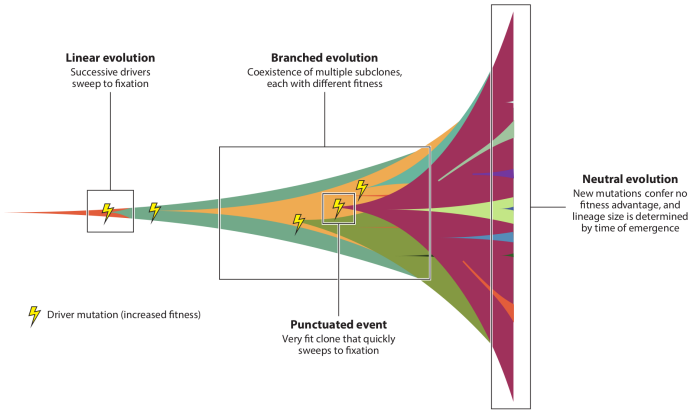
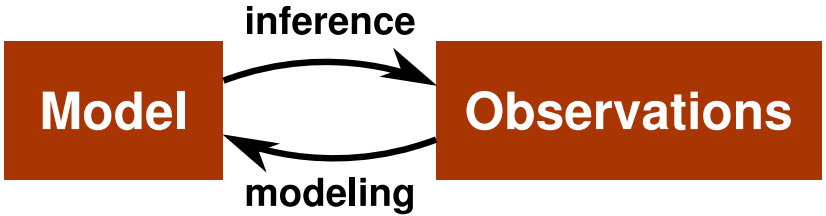


Figure: Schematic view of cancer evolution (Williams *et al.* 2019).

Why do we need models?

Models enable quantitative predictions about genomic patterns that can be tested using statistical inference methods



What are the important parameters?

Evolutionary parameters

- **Cancer cell population size**: gives an idea of tumor progression
- **Cell generation time**: rate of division of self-renewing cells
- **Mutation rate**: how often do alterations appear?
- **Selection coefficients**: do different tumor cells have different growth rates?

Ecological parameters

- **Ecological strategy**: what is the phenotype of the cancer cells? how does it provide an advantage?
- **Carrying capacity/Habitat quality**: what is the extent of vasculature and hypoxic niches? what it the nutrient concentration (e.g., glucose)?
- **Migration rate**: how often do cells migrate/metastasize to other locations?
- **Predation**: how much do immune cells (e.g., tumor-infiltrating lymphocytes) pressure the tumor?

Part I: evolution of the cancer genome

Modeling approaches

Theoretical frameworks are based on **population genetics** (the study of mutations in populations–at micro-evolutionary time-scales) and **molecular phylogenetics** (the study of mutations in species–at macro-evolutionary time-scales)

Modeling approaches

Theoretical frameworks are based on **population genetics** (the study of mutations in populations–at micro-evolutionary time-scales) and **molecular phylogenetics** (the study of mutations in species–at macro-evolutionary time-scales)

Frameworks

- **Mathematical models**: *equations* describing the relationships between processes and observable quantities; allow precise description of the impact of evolutionary processes on genomic patterns
- **Computational models**: *algorithm* describing the effect of processes on observable quantities; allows complex models (e.g., model each cell)

Features

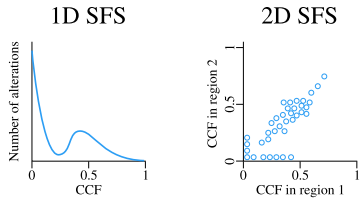
- Spatially implicit or explicit
- Forward-in-time or backward-in-time

What is the main quantity of interest?

The cancer cell fraction (CCF)

Definition: proportion of tumor cells that harbor a given somatic alteration (e.g., mutation, indel, structural variant)

Typical representation: the Site Frequency Spectrum (SFS)



How to get the CCF?

Goal: go from variant allelic fraction (VAF) to cancer cell fraction (CCF)

Method: call single nucleotide variants (SNVs) and copy number variants (CNVs), and correct VAFs for CN state

- Easiest when CN=2 and heterozygous alteration: then $\text{VAF} = \frac{1}{2}\text{CCF} \times \phi$, so

$$\text{CCF} = \frac{2}{\phi}\text{VAF},$$

where $\phi \in [0, 1]$ is the tumor purity. *Example: a heterozygous variant present in $CCF = 100\%$ of tumor cells, given a purity of $\phi = 80\%$, is expected to have a VAF $= \frac{1}{2} \times 1 \times 0.8 = 0.4$.*

How to get the CCF?

Goal: go from variant allelic fraction (VAF) to cancer cell fraction (CCF)

Method: call single nucleotide variants (SNVs) and copy number variants (CNVs), and correct VAFs for CN state

- Easiest when CN=2 and heterozygous alteration: then $\text{VAF} = \frac{1}{2}\text{CCF} \times \phi$, so

$$\text{CCF} = \frac{2}{\phi}\text{VAF},$$

where $\phi \in [0, 1]$ is the tumor purity. *Example: a heterozygous variant present in $CCF = 100\%$ of tumor cells, given a purity of $\phi = 80\%$, is expected to have a VAF $= \frac{1}{2} \times 1 \times 0.8 = 0.4$.*

- In region with CN$\neq$ 2: then

$$\text{CCF} = \frac{2(1-\phi) + C_{\text{total}}\phi}{C_{\text{mut}}\phi}\text{VAF},$$

where C_{total} is the total CN in the tumor and C_{mut} is the allele-specific CN.

How to get the CCF?

Goal: go from variant allelic fraction (VAF) to cancer cell fraction (CCF)

Method: call single nucleotide variants (SNVs) and copy number variants (CNVs), and correct VAFs for CN state

- Easiest when CN=2 and heterozygous alteration: then $\text{VAF} = \frac{1}{2}\text{CCF} \times \phi$, so

$$\text{CCF} = \frac{2}{\phi}\text{VAF},$$

where $\phi \in [0, 1]$ is the tumor purity. *Example: a heterozygous variant present in $CCF = 100\%$ of tumor cells, given a purity of $\phi = 80\%$, is expected to have a VAF $= \frac{1}{2} \times 1 \times 0.8 = 0.4$.*

- In region with CN$\neq$ 2: then

$$\text{CCF} = \frac{2(1-\phi) + C_{\text{total}}\phi}{C_{\text{mut}}\phi}\text{VAF},$$

where C_{total} is the total CN in the tumor and C_{mut} is the allele-specific CN.

Note: ideally, inferred simultaneously with evolutionary model

Note 2: although VAFs are informative on tumor evolution, often treated as technical variable!

Approach (i): a simple deterministic forward mathematical model

Williams *et al.* (2016): constant mutation rate + exponential growth (neutral evolution)

Features: mathematical model, spatially implicit, forward-in-time

Approach (i): a simple deterministic forward mathematical model

Williams *et al.* (2016): constant mutation rate + exponential growth (neutral evolution)

Features: mathematical model, spatially implicit, forward-in-time

Model:

- Number of mutations M: $\frac{dM}{dt} = \phi\mu\lambda N(t)$ (i), where μ is the mutation rate and λ the growth rate
- Population size: $N(t) = e^{\lambda\beta t}$ (ii), where β is proportion of surviving cells per division
- Frequency of a mutation (under infinite sites model): $f = \frac{1}{\phi N(t)}$ (iii)

Approach (i): a simple deterministic forward mathematical model

Williams *et al.* (2016): constant mutation rate + exponential growth (neutral evolution)

Features: mathematical model, spatially implicit, forward-in-time

Model:

- Number of mutations M: $\frac{dM}{dt} = \phi\mu\lambda N(t)$ (i), where μ is the mutation rate and λ the growth rate
- Population size: $N(t) = e^{\lambda\beta t}$ (ii), where β is proportion of surviving cells per division
- Frequency of a mutation (under infinite sites model): $f = \frac{1}{\phi N(t)}$ (iii)

Observable values: $M(f)$, the cumulative number of mutations at frequency f, $dM(f)/df$, the number of mutations at frequency f

Solving (i)-(iii) for $M(f)$ gives:

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{\max}}\right), \qquad \frac{dM(f)}{df} = -\frac{\mu}{\beta}\left(\frac{1}{f^2}\right). \tag{1}$$

Approach (i): a simple deterministic forward mathematical model, inference

Eq. 1 is equivalent to a simple linear model

$$M(x) = ax + b,$$

with $x = 1/f$, $a = \frac{\mu}{\beta}$, and $b = -\frac{\mu}{\beta f_{\max}}$.
Thus, we can estimate a scaled mutation rate μ/β, and a general model fit using **linear regression**

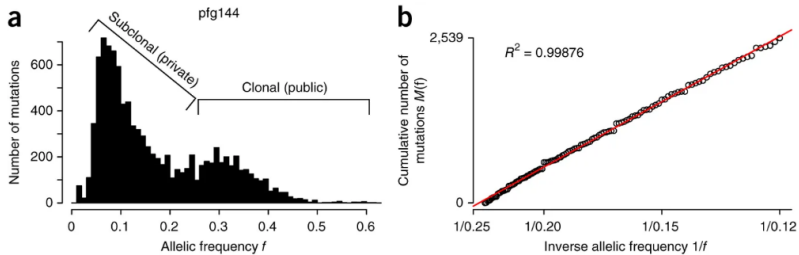


Figure: Gastric cancer whole-genome sequencing data (Williams *et al.* 2016).

Approach (i): a simple deterministic forward mathematical model, inference

Advantages:

- Mathematically tractable
- Simple, general predictions

Criticisms:

- Other processes can produce similar predictions (Tarabichi *et al.* 2018)
- Deterministic nature does not capture variability of evolutionary processes and does not allow classical hypothesis testing
- Is neutral evolution the null in cancer evolution?

Approach (ii): a more complicated probabilistic forward mathematical model

Caravagna *et al.* (2020): mixture of neutrally evolving (constant mutation rate + exponential growth) and selected clones (exponential growth at rate higher than the neutral growth)

Features: mathematical model, spatially implicit, forward-in-time

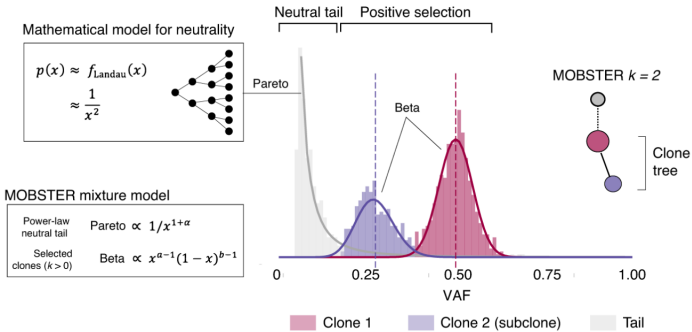


Figure: MOBSTER model (Caravagna *et al.* 2020). The neutral tail is based on the solution to the stochastic Luria–Delbruck model of bacterial growth (Kessler and Levine 2013; Luria and Delbrück 1943).

Approach (ii): a more complicated probabilistic forward mathematical model, inference

Fit a $k+1$ component mixture model, where each alterations comes either from the neutral tail or from one of k selected clones

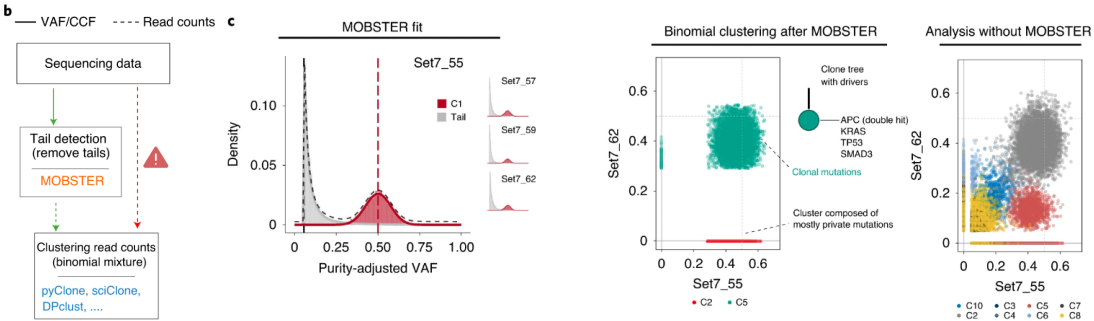


Figure: MOBSTER inference scheme (Caravagna *et al.* 2020).

Approach (ii): a more complicated probabilistic forward mathematical model, inference

Advantages:

- Encompasses many scenarios
- Allows inference of population parameters

Criticisms:

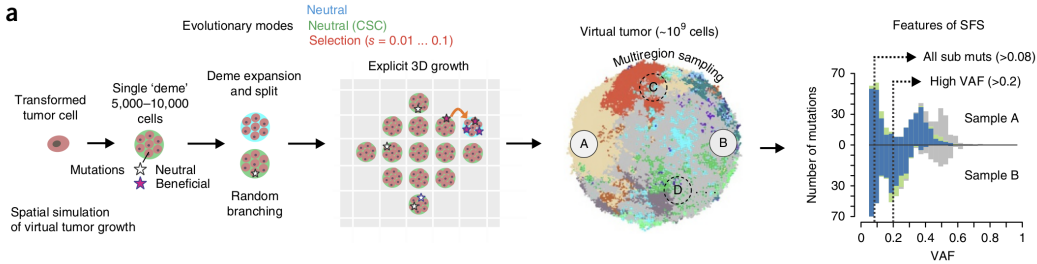
- Spatially implicit ⇒ does not account for spatial heterogeneity
- Only considers exponential growth ⇒ mostly suited for initial progression phase, cannot detect slowing down or tumor regression
- Only considers competition between clones ⇒ not commensalism or cooperation that could help maintain stable proportions of different clones through time

Approach (iii): a forward computational model

Sun *et al.* (2017): constant mutation rate + spatially explicit 3-D growth + selection

Other features: forward-in-time

Model:



Approach (iii): a forward computational model

The model allows to make **quantitative predictions** in different scenarii, and distinguish **qualitative behaviors**:

- neutral or effectively neutral models lead to weak spatial correlation, little differentiation between regions, and no private "bumps" in the SFS

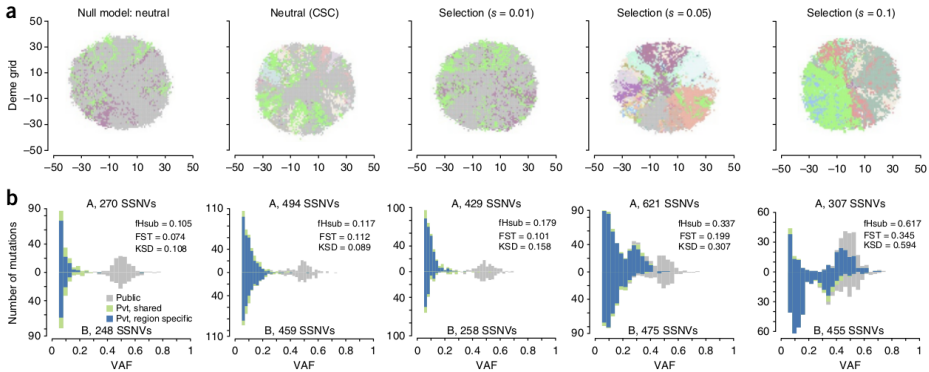


Figure: Predictions under several models (Sun *et al.* 2017).

Approach (iii): a forward computational model, inference

Compute summary statistics, compare their values with that of many simulations ($\sim 10^6$); Retain the parameters of simulations with stats closest to the observations (Approximate Bayesian Computation)

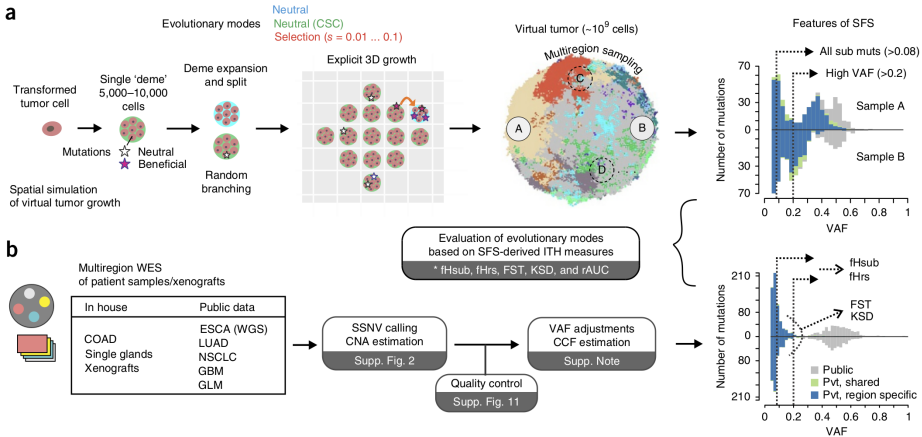


Figure: Model schematic and inference method (Sun *et al.* 2017).

Approach (iii): a forward computational model, inference

In practice, summary statistics enable to separate the 2 qualitative behaviors

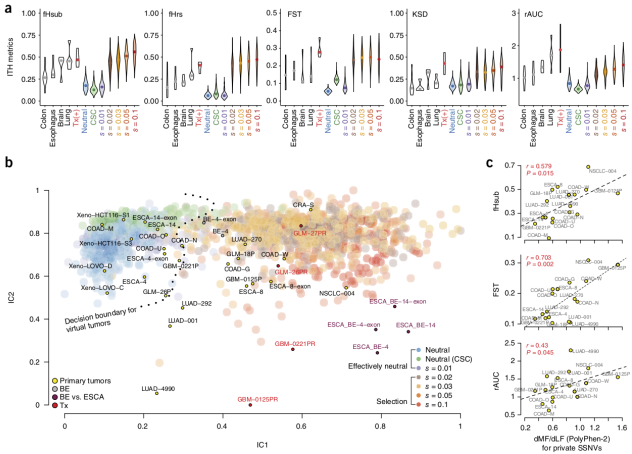


Figure: Application to tumor genomic data (Sun *et al.* 2017).

Approach (iii): a forward computational model, inference

Also extended to study tumor/metastases co-evolution and infer type of seeding (mono- or polyclonal)

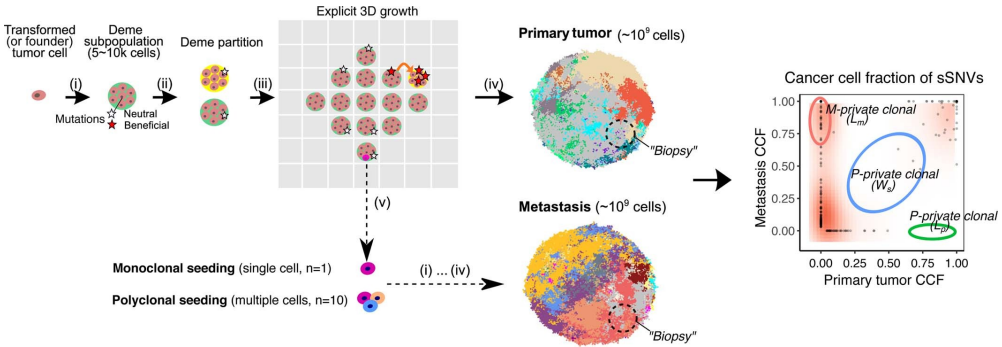


Figure: Model schematic (Hu *et al.* 2020).

Approach (iii): a forward computational model, inference

Advantages:

- Allows arbitrarily complex scenarios that fit the biology

Criticisms:

- Computationally intensive (large parameter space to explore)
- Identifiability is unknown (worked in this case but might not for other applications)

Approach (iv): a clonal mixture model

Miller *et al.* (2014): clone identification + phylogenetic reconstruction
Other features: no assumption about selection, growth, or mutation rate, backward-in-time
Mixture model: Given N individuals sequenced at M variant positions, the $N \times M$ theoretical CCF matrix **D** is:

$$\mathbf{D} = \mathbf{A}.\mathbf{T}, \tag{2}$$

with **A** a $N \times K$ mixture matrix that gives the proportion of each of K clones in the N samples, and **T** a $K \times M$ genotypes matrix that gives the variants of each clone. The *observed* CCF matrix is $\tilde{\mathbf{D}} \sim Binom(\mathbf{X}, \mathbf{D})$, where **X** is the sequencing coverage matrix.
Tree model: assumes that **A** is tree-like, i.e., it satisfies the sum rule

$$d_{\text{cluster}Y} = d_{\text{clone}Y} + \sum_{X_i \in \text{direct subclones of} Y} d_{\text{cluster}X_i},$$

where d_X is the CCF of X

Approach (iv): a clonal mixture model, inference

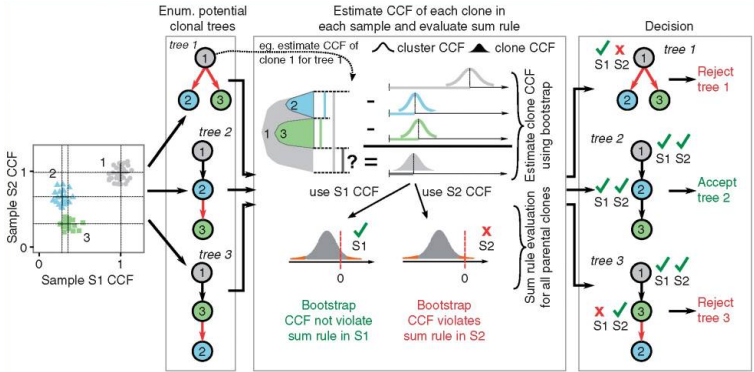


Figure: Example mixture inference (Dang *et al.* 2017)

Approach (iv): a clonal mixture model, inference

Results: Finds peculiar evolutionary patterns (convergent evolution)

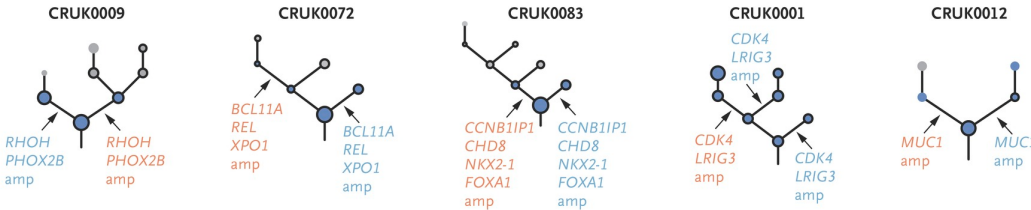


Figure: Phylogenetic tree of non–small–cell lung cancer (Jamal-Hanjani *et al.* 2017).

Approach (iv): a clonal mixture model, inference

Results: Attempts at estimating population size N_e (with phylogenetic software BEAST)

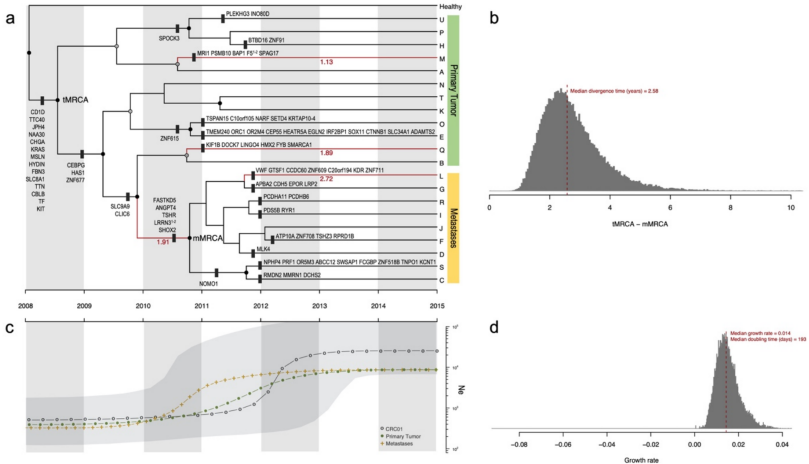


Figure: Phylogenetic tree of non–small–cell lung cancer (Alves *et al.* 2019).

Approach (iv): a clonal mixture model, inference

Advantages

- (relatively) Hypothesis-free
- Infers the history of the tumor–the sequence of alteration events

Criticisms

- Oversimplification of the data (the actual tree has thousands of branches)
- Weakly informative about processes (selection, growth)

Part II: evolution of the cancer transcriptome

Modeling approaches

Theoretical frameworks are based on **quantitative genetics** (the study of the evolution of phenotypes in populations) and **phylogenetics of traits** (the study of the evolution of phenotypes in species)

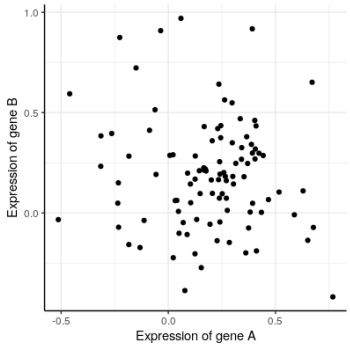
Most traits have **complex genetic architectures**, that are often unknown

What is the main quantity of interest?

Molecular traits

Definition: quantitative molecular variables (e.g., gene expression) that have an important role in cancer initiation and progression

Typical representation: a scatterplot, where each dimension represents a trait of interest



How to get the right molecular traits?

Goal: identify molecular the variables that are important for cancer evolution

Methods:

- (Supervised) use biological knowledge of a cancer type to choose relevant pathways
- (Unsupervised) leverage the between-patient molecular variation within a cohort to identify major sources of variation

An equilibrium deterministic mathematical model

Hausser and Alon (2020): directional selection + constraints (trade-off)

Features: mathematical model, unsupervised, equilibrium model (selection had enough time to act)

An equilibrium deterministic mathematical model

Hausser and Alon (2020): directional selection + constraints (trade-off)

Features: mathematical model, unsupervised, equilibrium model (selection had enough time to act)

Model:

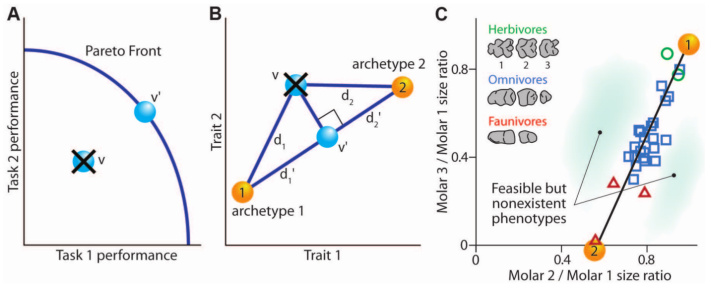
- cancer cells have to perform **multiple tasks** (e.g., cell division, energy production, interaction with immune system, invasion and tissue remodeling)
- cells **cannot excel at everything** (e.g., the optimal phenotype for cell division will not be optimal in terms of interaction with immune system) $\Rightarrow$ **trade-offs**

An equilibrium deterministic mathematical model

Hausser and Alon (2020): directional selection + constraints (trade-off)

Features: mathematical model, unsupervised, equilibrium model (selection had enough time to act)

- **Pareto front**: set of phenotypes that cannot be improved at all tasks at once (increasing some performances requires reducing performance at other tasks)
- **Archetype**: optimal phenotype for a task



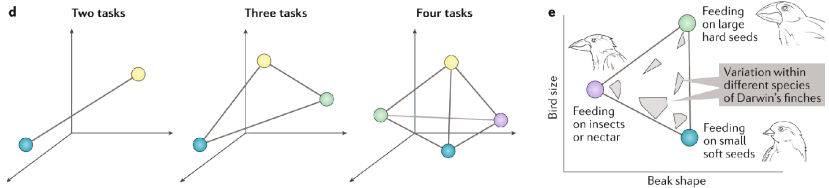
An equilibrium deterministic mathematical model

Hausser and Alon (2020): directional selection + constraints (trade-off)

Features: mathematical model, unsupervised, equilibrium model (selection had enough time to act)

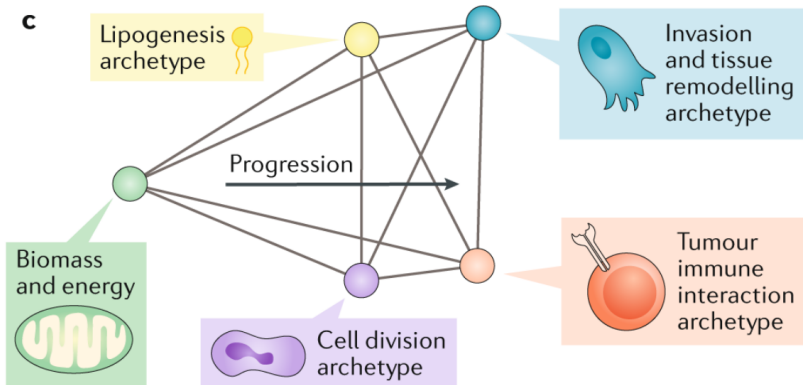
Results:

- the Pareto front has a **simple geometric shape** that only depends on the number of tasks (2 tasks: line, 3 tasks: triangle, four tasks: tetrahedron, etc)
- position within the front depends on the strength of selection for each task (strong selection for a tasks: specialist, equal selection for each task: generalist)



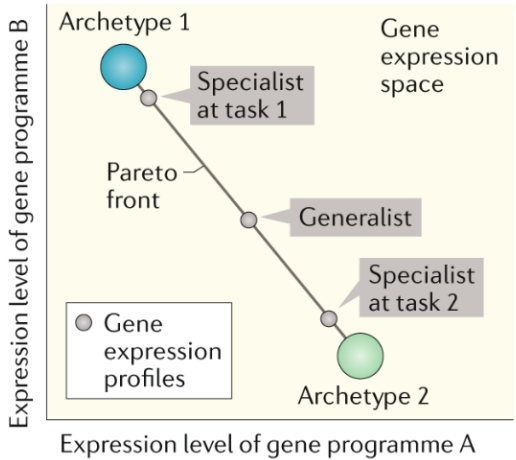
An equilibrium deterministic mathematical model

Hausser *et al.* (2019): most cancer types experience trade-offs between the same 5 tasks



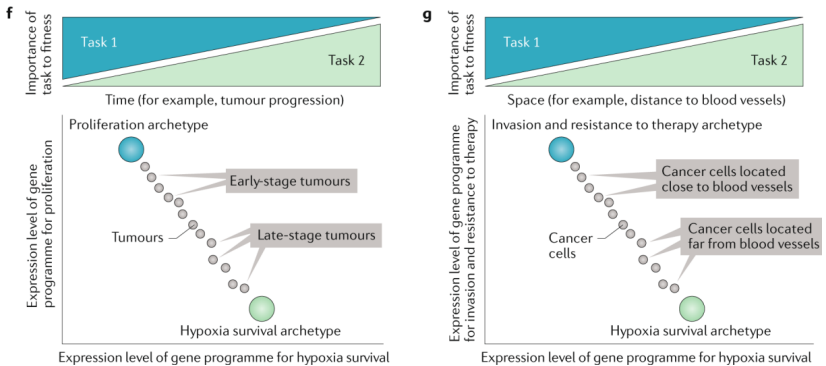
An equilibrium deterministic mathematical model

Hausser *et al.* (2019): we expect cancer transcriptomes to be a proxy for the molecular phenotypes facing trade-offs between tasks



An equilibrium deterministic mathematical model

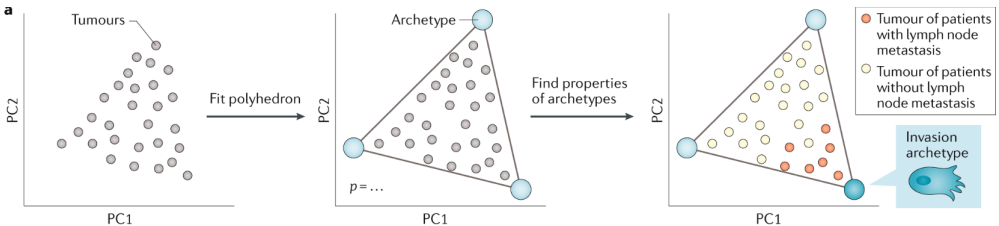
Hausser *et al.* (2019): The optimal position in the Pareto front can vary through time and space



An equilibrium deterministic mathematical model

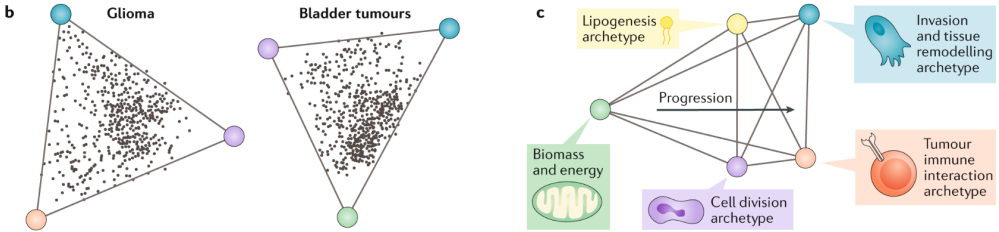
Hausser *et al.* (2019): Method

- Create a low-dimensional representation of the data (PCA)
- Fit a simplex and identifying vertices (archetypes)
- Interpret archetypes through gene-set enrichment analyses



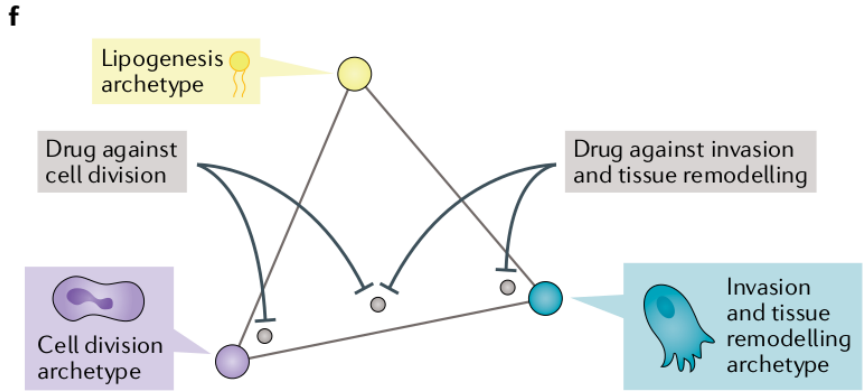
An equilibrium deterministic mathematical model

Hausser *et al.* (2019): Results



An equilibrium deterministic mathematical model

Hausser *et al.* (2019): Clinical implications



Take-home message

Interpret the geometry of the data

1. **Look at the molecular variation** (preferably after applying some linear dimensionality reduction)
2. Think about the **constraints in the data**: which genes are never overexpressed simultaneously? what does it say about the trade-offs between biological functions?
3. **Interpret the vertices and edges** of the distribution: do they have clear molecular phenotypes corresponding to pure tumoral strategies?

RARE CANCERS GENOMICS

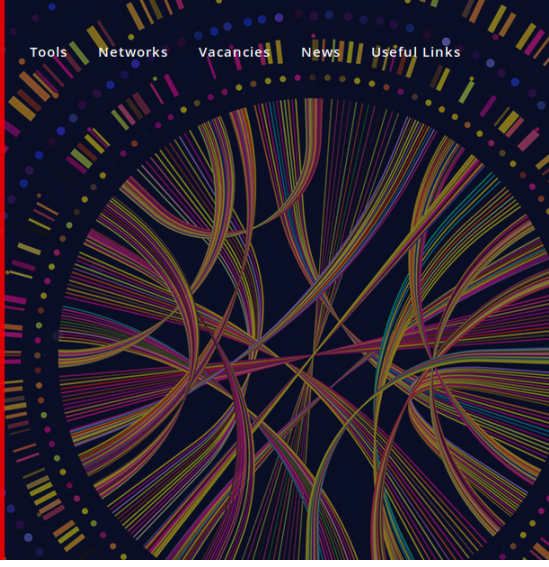
Home Study Design Projects People Tools Networks Vacancies News Useful Links

Contact Us

*

Rare Cancers Genomics

Multidisciplinary and multi-omics molecular characterisation of rare cancers



References

Alves, J. M., S. Prado-Lopez, J. M. Cameselle-Teijeiro, and D. Posada, 2019 Rapid evolution and biogeographic spread in a colorectal cancer. bioRxiv : 623850.

Caravagna, G., T. Heide, M. J. Williams, L. Zapata, D. Nichol, *et al.*, 2020 Subclonal reconstruction of tumors by using machine learning and population genetics. Nature Genetics **52**: 898–907.

Dang, H., B. White, S. Foltz, C. Miller, J. Luo, *et al.*, 2017 Clonevol: clonal ordering and visualization in cancer sequencing. Annals of oncology **28**: 3076–3082.

Hausser, J., and U. Alon, 2020 Tumour heterogeneity and the evolutionary trade-offs of cancer. Nature Reviews Cancer : 1–11.

Hausser, J., P. Szekely, N. Bar, A. Zimmer, H. Sheftel, *et al.*, 2019 Tumor diversity and the trade-off between universal cancer tasks. Nature communications **10**: 1–13.

Hu, Z., Z. Li, Z. Ma, and C. Curtis, 2020 Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. Nature Genetics : 1–8.

Jamal-Hanjani, M., G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. Watkins, *et al.*, 2017 Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine **376**: 2109–2121.

Kessler, D. A., and H. Levine, 2013 Large population solution of the stochastic luria–delbrück evolution model. Proceedings of the National Academy of Sciences **110**: 11682–11687.

Luria, S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance. Genetics **28**: 491.

Miller, C. A., B. S. White, N. D. Dees, M. Griffith, J. S. Welch, *et al.*, 2014 Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS computational biology **10**: e1003665.

Sun, R., Z. Hu, A. Sottoriva, T. A. Graham, A. Harpak, *et al.*, 2017 Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature genetics **49**: 1015.

Tarabichi, M., I. Martincorena, M. Gerstung, A. M. Leroi, F. Markowetz, *et al.*, 2018 Neutral tumor evolution? Nature genetics **50**: 1630.

Williams, M. J., A. Sottoriva, and T. A. Graham, 2019 Measuring clonal evolution in cancer with genomics. Annual review of genomics and human genetics **20**.

Williams, M. J., B. Werner, C. P. Barnes, T. n. Graham, and A. Sottoriva, 2016 Identification of neutral tumor evolution across cancer types. Nature genetics **48**: 238.