Inferring intra-tumoral heterogeneity at single cell resolution

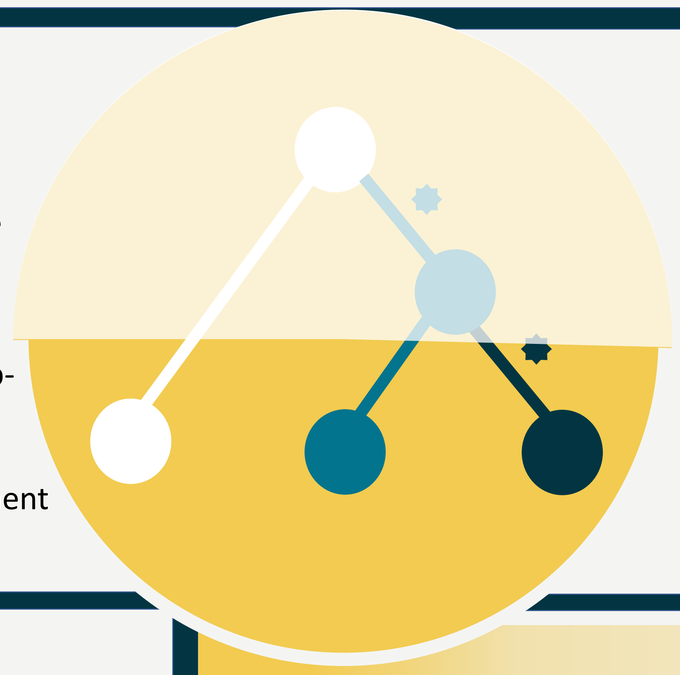
A Bayesian model

Puxue Qiao[1], Davis J. McCarthy[1, 2]

1. Bioinformatics & cellular genomics, St. Vincent's Institute of Medical Research, Melbourne, VIC, Australia
2. Melbourne Integrative Genomics (MIG), the University of Melbourne, VIC, Australia

Motivation

- Characterization of intra-tumoral heterogeneity due to the accumulation of somatic mutations over time is critical to understanding the natural histories of cancer cell populations.
- While useful tools have been developed for inferring sub-clonal structure from scRNA-seq data[1, 2, 3], an integrated model that allows orthogonal sources of information to borrow strength from each other would show improvement in both clustering and clonal profile estimation.

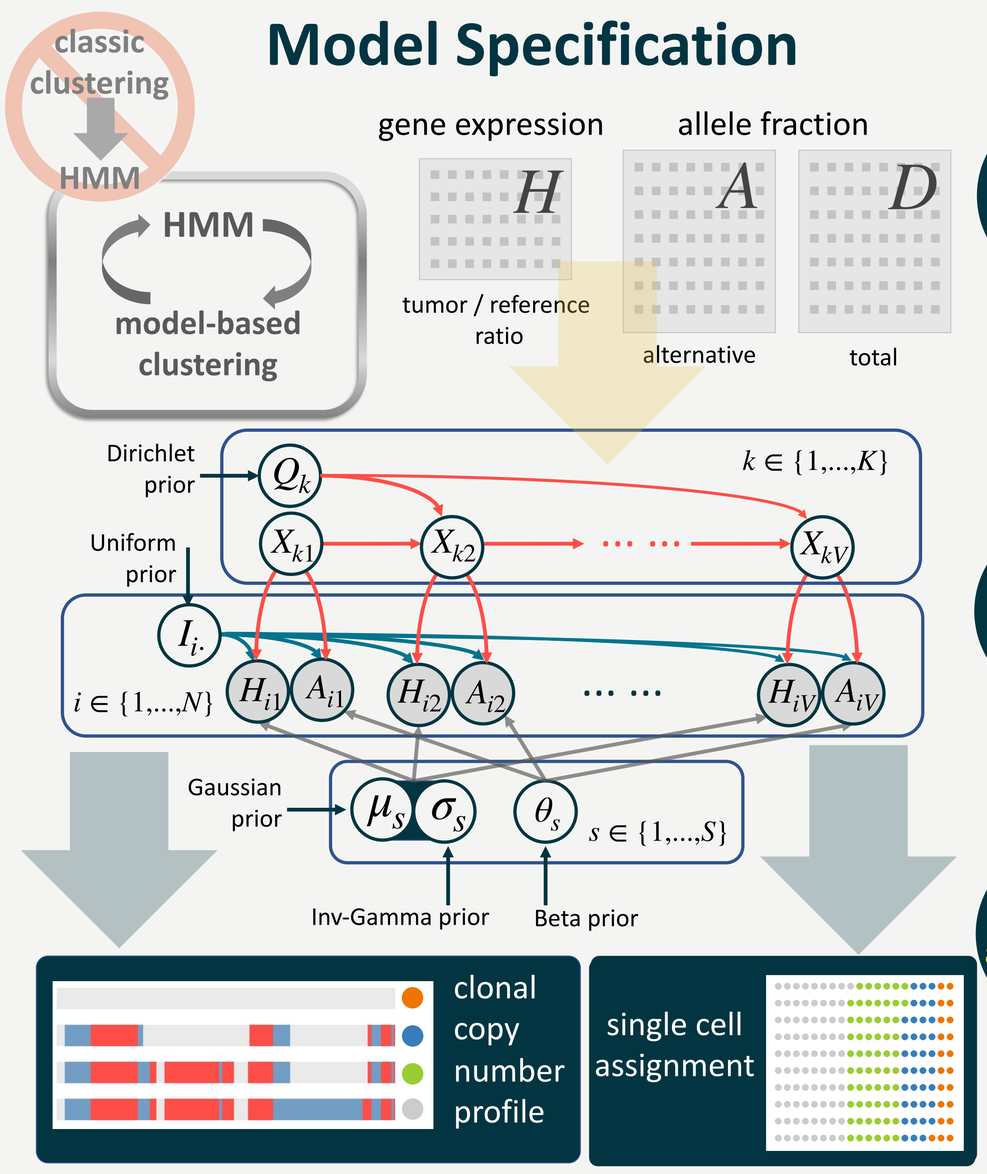


Goal

Build a flexible model for scRNA-seq data that:

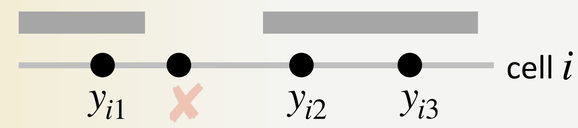
- Assigns the single-cells transcriptome profiles to their clone-of-origin
- Integrates multiple sources of information including expressed variant alleles and gene expression level
- Infers clonal profile w.r.t single nucleotide alterations and copy number alterations

Model Specification



Data

- Cells, $(i = 1, \ldots, N)$ are independent;
- All cells have reads for allele fraction in all (germline) variant positions $(v = 1, \ldots, V)$;
- Every (germline) variant position is covered by a gene in the expression matrix;
- Expression ratio on log scale follows a Gaussian distribution;
- Allele fraction is modelled by a Binomial distribution;



$$Y_{iv} = (H_{iv}, A_{iv}, D_{iv})$$

$$H_{iv} \sim N(\mu_{iv}, \sigma_{iv})$$

$$A_{iv} \sim Binom(\theta_{iv}, D_{iv})$$

Sub-clones

- Copy number profiles are defined at sub-clonal $(k = 1,\ldots,K)$ level;
- Each sub-clonal copy number profile is a (latent) Markov Chain;
- Transition probability is invariant across sub-clones.

$$X_{k\cdot} = (X_{k1}, \ldots, X_{kV}),\ X_{kv} \in \{1,\ldots,S\}$$

$$P(X_{kv} \mid X_{k1}, \ldots, X_{kv-1}) = P(X_v \mid X_{v-1}) \triangleq Q(x_v, x_{v-1})$$

Cell assignment

- Each cell is assigned to one and only one sub-clone;
- All cells in the same sub-clone have the same copy number profile.

$$I_{i\cdot} = (I_{i1}, \ldots, I_{iK}),\ \sum_k I_{ik} = 1$$

$$I_{ik} = I(\text{cell } i \text{ belongs to subclone } k)$$

$$\tilde{\mu}_{iv} = \prod_k \prod_s \mu_s^{I_{ik} I(X_{kv}=s)},\ \text{same for } \tilde{\sigma}_{iv} \text{ and } \tilde{\theta}_{iv}$$

(preliminary) Empirical Results --- in comparison with infercnv [1]

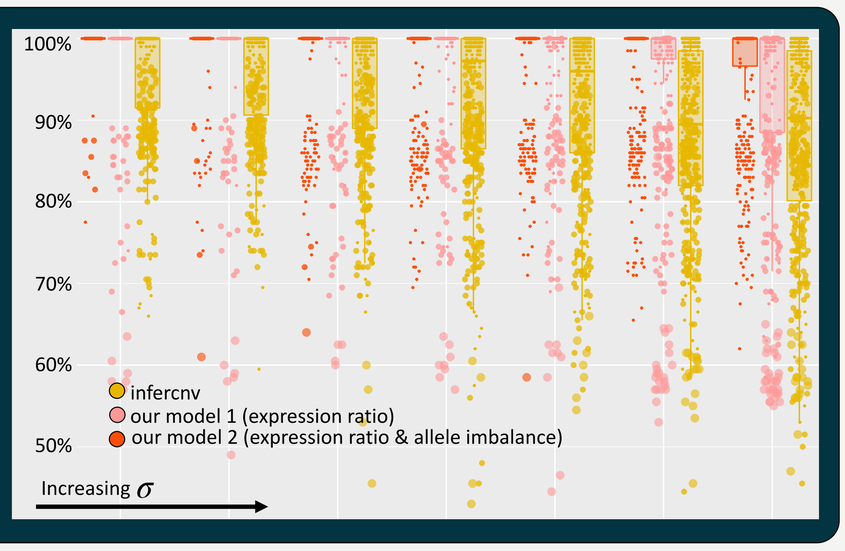
Simulation Setup

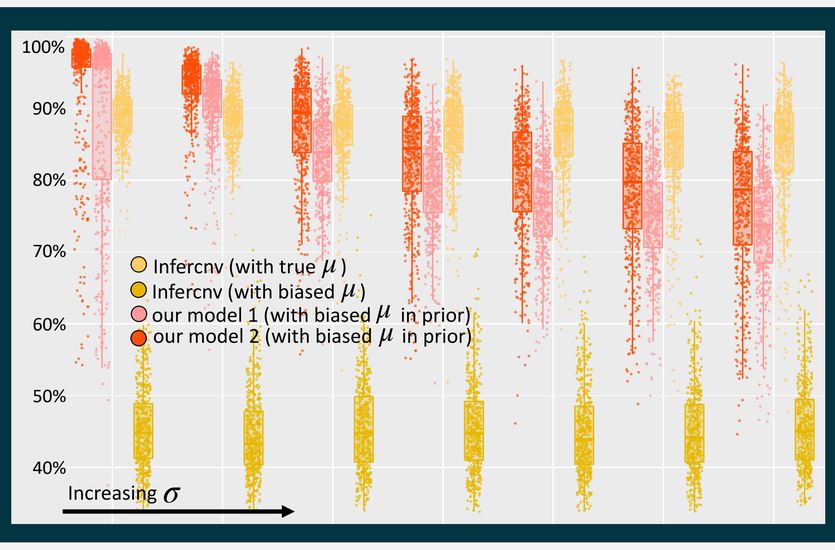
- $200\,(N)$ simulated cells
- $500\,(V)$ genes each cell
- $3(S)$ copy number states $(\mu = -1, 0, 0.6)$
- $4\,(K)$ clusters/sub-clones
- Variance $\sigma = 0.1, \ldots, 0.7$ $\Rightarrow$different difficulty levels
- 500 simulations for each value of σ

Improved Clustering

(higher percentage of correctly clustered cells)

- Incorporating spatial dependency while clustering improves accuracy
- Integrating multiple sources of information improves robustness





More robust copy number state estimation

(higher percentage of correctly identified CN states)

- In our nested HMM, state-specific means of expression ratio are treated as latent variables instead of required parameters, thus much less sensitive to biased input information.
- If the true values are provided, infercnv performs the best, but a mild bias can lead to all state lost.

Conclusion

- Our model runs clustering while incorporating spatial dependency via nested HMMs. We show that it outperforms popular existing tool that also does clustering and adopts HMM, but separately.
- We show that our model is flexible in the sense that it can integrate various sources of information to improve both clustering and CN state estimation performance. So will be interesting if both types of data are available.
- Next Steps:
  - Enable unknown number of clusters and/or CN states
  - Investigate the distribution of expression ratio
  - Test on real data

Reference

[1]. Tickle T, Tirosh I, et al. (2019). *inferCNV of the Trinity CTAT Project.*. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[2]. J. Fan, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell rna-seq data. Genome research, 28(8):1217–1227, 2018.
[3]. D. J. McCarthy, R. Rostom, Y. Huang, et al. Cardelino: computational integration of somatic clonal substruc- ture and single-cell transcriptomes. Nature Methods, 17(4):414–421, 2020.