R Notebook

I'm working on somatic copy number alterations (SCNA) in single cells. Lots of interesting work on this topic is being done by the Kuhn/Hicks lab at USC. The problem I'm focusing on here is clustering of SCNA in single cells.

I've so far found only two software pacakages for SCNA calling and sample clustering, both coming from two labs at Cold Spring Harbor Laboratory (CSHL), with some overlapping personnel.

1. Ginkgo[1]
2. SCclust

Ginkgo comes as an integrated shiny app hosted at CSHL while SCclust is under active development and seems to require extensive configuration and installation.

Each of them seem to proceed from .bam/.bed input files and yield SCNA segmentation profiles and sample dendrograms on a range of metrics (euclidean distance, correlation, etc.)

After trying some of our data in Ginkgo, my PI commented that an unbiased comparison between SCNA profiles for the purposes of building a tree might be deceptive because correlation of some features might be due to similar selective pressures and disease processes rather than shared inheritance between cells.

Some background, it is thought that tumor evolution occurs through clonal evolution. That is, minor changes in the genome of a given cell result in proliferation of that cell and formation of a clone. This is thought to lie behind chemotherapy resistance and relapse. Chemotherapy kills all but a few resistant cells which then grow out as a clone and are refractory to future chemotherapy.

In retinoblastoma as in many cancers, stereotypical SCNA profiles are common. The functional significance of these changes is poorly understood, but it is reasonable to think that certain changes confer a survival advantage. It is therefore reasonable to think that SCNAs might arise in overlapping regions in two clones despite there being no direct relation between the two. If you're trying to infer clones from SCNA data then, it's not enough to look at overall correlation between two cells.

You might be able to distinguish clones on the basis of the breakpoints of SCNAs, as it would be much less likely that two separate clones could develop SCNA in identical chromosomal regions.

I don't understand what clustering method would take that into account. The specifics of clustering is a bit of blind-spot for me. I understand the principles behind different methods (complete, average, ward, etc.) but I'm not clear how to account for this seeming limitation. Doubtless it's a common worry in application of clustering to many datasets.

I'm also still uncertain the best implementation of single cell SCNA analysis to run. Can either method address this issue?

I've found several citations for Ginkgo.

1. One comparison between SCNA called from single cell RNA sequencing data[2].
2. Another method for calling SCNA from single cell RNA sequencing data
3. An application for deriving estimates of chromosomal instability from single cell SCNA[3]
4.

SCclust isn't published yet, thought the PI responsible seems to be deeply involved in single cell bioinformatics.

Information I've found relating to SCclust include CORE[4] called "A Software Tool for Delineating Regions of Recurrent DNA Copy Number Alteration in Cancer"

Caravagna

1. Genomic evolution to cacner relates to NGS and machine learning for inference of explanatory models of epi/genomic events happend in tumor initiation and development.
2. recent work on 'selective advantage' relation among driver mutations in cancer progression and modeling
3. introduce PiCnlc, a pipeline to get pgregressio nmodels from muliomics data.
4. Includes sample stratification, driver selection, fitness-euivalend exlusive alterations, and progression model inference.

References

1. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods* **12**, 1058–1060 (2015).

2. Poirion, O., Zhu, X., Ching, T. & Garmire, L. X. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nature communications* **9**, 4892 (2018).

3. Greene, S. B. *et al.* Chromosomal instability estimation based on next generation sequencing and single cell genome wide copy number variation analysis. *PLoS ONE* **11**, 1–17 (2016).

4. Sun, G. & Krasnitz, A. Chapter 4 CORE : A Software Tool for Delineating Regions of Recurrent. **1878**, (2019).