Review of Methods for Single Cell SCNA Detection and Clustering

# The issue

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.

# The Software

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

1. [Ginkgo](http://qb.cshl.edu/ginkgo/?q=/EGgIcLtaAcUEbqb1VaIa) (Garvin et al., 2015)
2. [SCclust](https://github.com/KrasnitzLab/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.

The third comes as a bioconductor package and integrates karyotype heterogeneity.

1. [Aneufinder](https://bioconductor.org/packages/release/bioc/html/AneuFinder.html) (Bakker et al., 2016)

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.)

# The Problem

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.

# So Which To Use?

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

## Ginkgo

Ginkgo seems to be an implementation of the method laid out in Baslan et al. (2015)

I’ve found several citations for Ginkgo and/or Baslan et al. (2015).

1. One comparison between SCNA called from single cell RNA sequencing data (Poirion et al., 2018).
2. Another method for calling SCNA from single cell RNA sequencing data (Patel et al., 2014).
3. An application for deriving estimates of chromosomal instability from single cell SCNA (Greene et al., 2016)
4. A method for improved haplotype phasing by relying on whole genome amplification data (Satas and Raphael, 2018)
5. An application of single cell SCNA for parafin-embedded samples (Martelotto et al., 2017)
6. A method of whole genome amplification proposed to reduce error and improve precision of single cell SCNA (Chen et al., 2017)

## Aneufinder

I’ve found some for Aneufinder:

1. A method for improved haplotype phasing by relying on whole genome amplification data (Satas and Raphael, 2018)

## SCclust

SCclust isn’t published yet, though the PI responsible seems to be deeply involved in single cell SCNA work. Information I’ve found relating to SCclust includes:

1. CORE called “A Software Tool for Delineating Regions of Recurrent DNA Copy Number Alteration in Cancer” (Sun and Krasnitz, 2019)

# References

Bakker, B., Taudt, A., Belderbos, M.E., Porubsky, D., Spierings, D.C., Jong, T.V. de, Halsema, N., Kazemier, H.G., Hoekstra-Wakker, K., Bradley, A., et al. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biology *17*, 1–15.

Baslan, T., Kendall, J., Ward, B., Cox, H., Leotta, A., Rodgers, L., Riggs, M., D’Italia, S., Sun, G., Yong, M., et al. (2015). Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Research *125*, 714–724.

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., and Xie, X.S. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science *356*, 189–194.

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., and Schatz, M.C. (2015). Interactive analysis and assessment of single-cell copy-number variations. Nature Methods *12*, 1058–1060.

Greene, S.B., Dago, A.E., Leitz, L.J., Wang, Y., Lee, J., Werner, S.L., Gendreau, S., Patel, P., Jia, S., Zhang, L., et al. (2016). Chromosomal instability estimation based on next generation sequencing and single cell genome wide copy number variation analysis. PLoS ONE *11*, 1–17.

Martelotto, L.G., Baslan, T., Kendall, J., Geyer, F.C., Burke, K.A., Spraggon, L., Piscuoglio, S., Chadalavada, K., Nanjangud, G., Ng, C.K., et al. (2017). Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. Nature Medicine *23*, 376–385.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

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

Satas, G., and Raphael, B.J. (2018). Haplotype phasing in single-cell DNA-sequencing data. Bioinformatics *34*, i211–i217.

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