Reproducible research, algorithms, and data

Benjamin Haibe-Kains
Princess Margaret Cancer Centre
University Health Network
University of Toronto
Ontario Institute of Cancer Research

2 open postdoc positions:
Re radiomics and single-cell RNA-seq

June 25, 2016
Replicability, reproducibility and reusability

▷ Implement and document your functions
→ Replicability

▷ Adapt your functions to similar datasets
→ Reproducibility

▷ Extend your functions to datasets generated in different settings (samples, platforms, normalization, ...)
→ Reusability
Building upon previous work

If you can do it with your own functions, you can do it with published algorithms

→ **genefu** R package reproducing published molecular subtyping classifiers and gene “signatures” with common interface
  
  + my own models

This holds true for datasets

→ **MetaGxData** data pack and ovarian ($n=3,752$) cancers
Hard to fully replicate results!

▷ Devil is in the details

▷ Try to reproduce the figures of the main paper
  ○ Exact same results ~10%
  ○ Approximately the same ~50%
  ○ The remaining 40%, well… I guess we are not smart enough to understand the methods section...

▷ Start communicating with the authors early on, most are willing to help

▷ Tons of unit testing and documentation

▷ Make your code and documentation publicly available to get the community to scrutinize your work
Same algorithm, different implementations, different results
Meta-analysis and comparative studies

With functions and data in hand, hard to resist the temptation to further challenge your model:

▷ Is my model robust?
▷ Is my model’s performance reproducible in multiple independent datasets?
▷ How does my model compare to competitors?

→ **survcomp** R package to compare the prognostic value of published and new gene signatures
Conclusion

Prototyping, implementing, documenting, testing, sharing, fixing, testing, extending, sharing, …

This cycle is vital in my lab where code is scrutinized and tested by multiple members before public release.

This helped me improve my Science and truly value the benefits of data and code sharing.
Acknowledgements

BHK lab

Princess Margaret Cancer Centre

▷ Deena Gendoo
▷ Gregory Chen
▷ Natchar Ratanasirigulchai

Collaborators

▷ Markus Schroeder
▷ Levi Waldron
▷ Aleix Prat
▷ Joel Parker
Thank you for your attention!

Questions?