Towards an interoperating ecosystem of tools and resources for population genetics in R

Hilmar Lapp[1,2], and Participants in the Population Genetics in R Hackathon

(1) US National Evolutionary Synthesis Center (NESCent), and (2) Center for Genomic and Computational Biology (GCB), Duke University, Durham, NC, USA

Motivation

The broad and inexpensive availability of modern next-generation sequencing and genotyping technologies has led to a wealth of data and analytical methods for population genetics research. There are now dozens of packages available for analyzing and visualizing population genetics data in the popular statistical and mathematical computing platform R. However, this organically grown wealth of methods and packages, combined with the exponential growth of datasets, has also created challenges for researchers to take full advantage of these resources. It can be difficult to know which R packages are best used, and many packages do not interoperate well. A common base class that provides efficient storage of genetic data and promotes interoperability remains lacking even though the need was identified years ago. Algorithm implementations often do not scale well to the kind of large volume datasets that are increasingly common. Creating complex analysis workflows that need to pass data, metadata, and other state information from one package to another can be challenging.

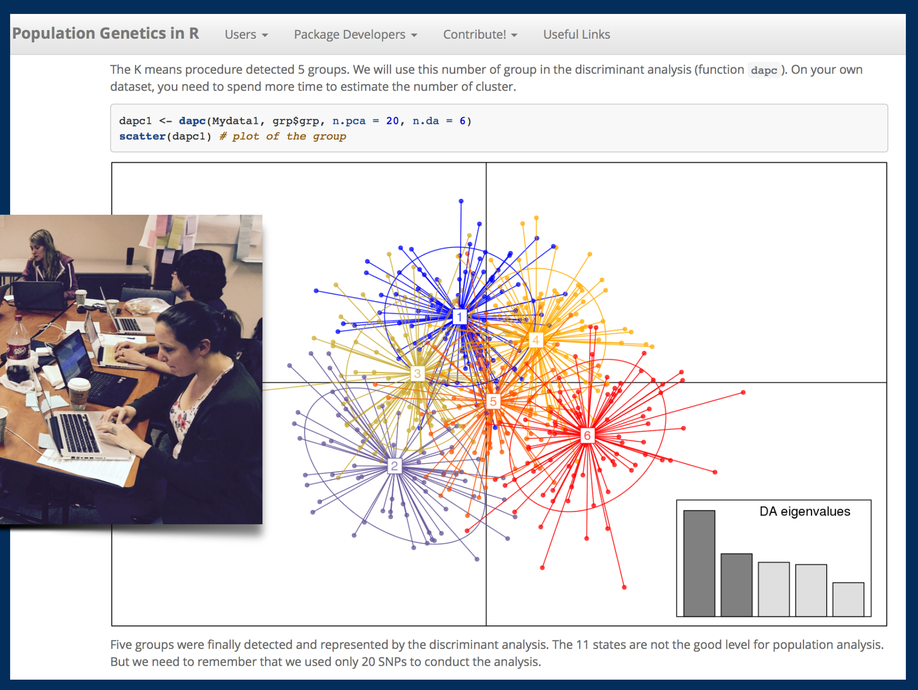
To address these gaps, the Population Genetics in R Hackathon was sponsored by and held at the National Evolutionary Synthesis Center (NESCent) on March 16-20, 2015. The event targeted interoperability, scalability, workflow gaps, and gaps in end-user documentation. Its goal was ultimately to help foster an interoperating ecosystem of tools and resources for both users and researcher-developers. For more details, see https://github.com/NESCent/r-popgen-hackathon

Result I: Community resource of population genetics analysis vignettes

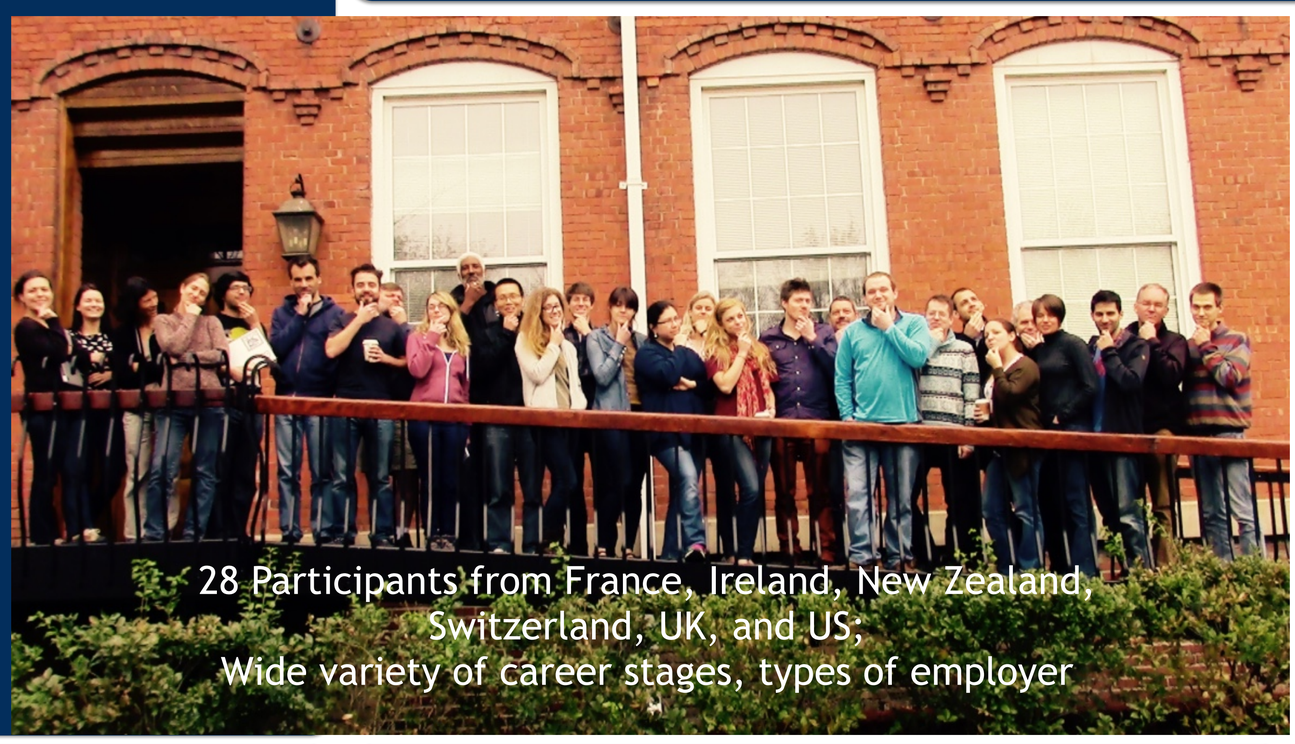
Motivation: Identifying which R packages can be recruited to answer a particular biological question in population genetics is difficult.

Results:

- Website generated from Rmarkdown + Knitr
- Source and website hosted on Github
- Pull requests auto-tested by Circle CI, master branch changes auto-rebuild the website
- Eight vignettes (5 biological) currently



Five groups were finally detected and represented by the discriminant analysis. The 11 states are not the good level for population analysis. But we need to remember that we used only 20 SNPs to conduct the analysis.

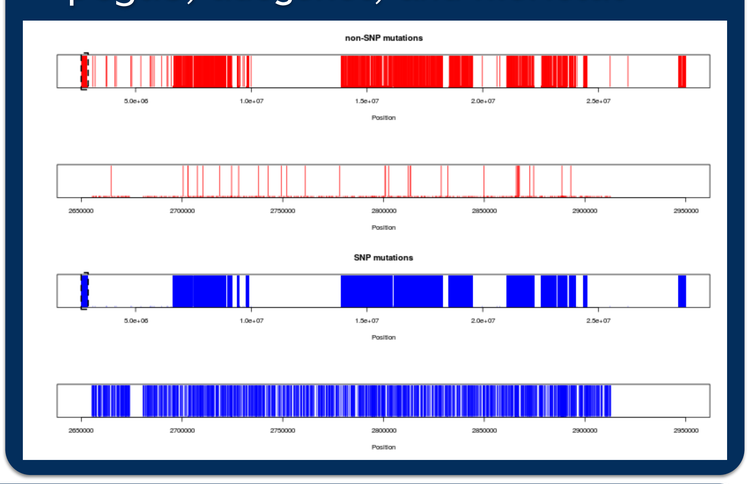


Result II: Streamlined, scalable, interoperable handling of VCF data

Motivation: Reading Variant Calling Format (VCF) does not scale well to large-volume data, and interoperability between packages is poor.

Results:

- apex: new R package extending ape for multiple genes. On Github and CRAN.
- pegas now has fast scanning and reading of VCF files
- Genetics data object (`genind`) now scalable and interoperable between `pegas`, adegenet, and hierfstat

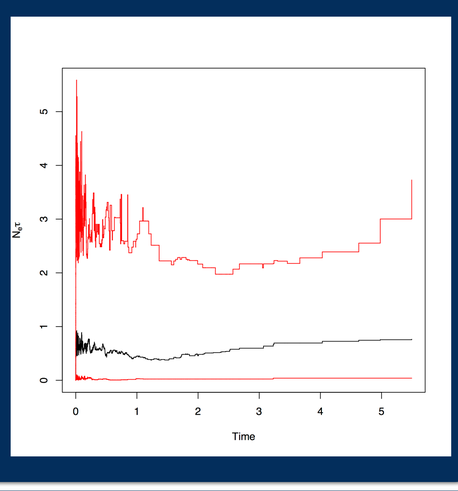


Result V: R package for calculating effective population size in multiple ways

Motivation: Some algorithms for calculating N_e are in different packages, others are not available in R.

Results:

- New multiNe package
- Hosted on Github



Result IV: Identifying and visualizing outliers in multivariate summary statistics space

Motivation: Multivariate statistics are a common product in population genetic analysis. R tools to summarize and visualize these and detected outliers interactively are lacking.

Results:

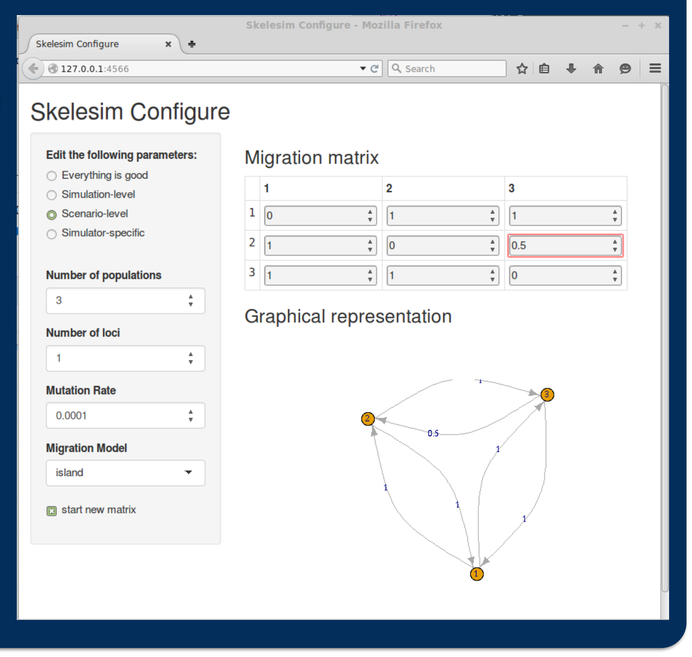
- New MINOTAUR package
- Hosted on Github
- Uses a `shiny`-powered web user-interface

Result III: Use-cases and power-testing for simulation

Motivation: There are simulator type, parameters, summary metrics, etc to choose when implementing simulations for population genetic questions. There is little guidance on how to decide.

Results:

- New skeleSim package
- Hosted on Github



Acknowledgments. The event was supported by the US National Evolutionary Synthesis Center (NESCent, http://nescent.org, NSF EF-0905606). H. Lapp is supported by Duke University's Center for Genomic and Computational Biology (GCB). We are indebted to the enthusiasm and energy of the hackathon participants that made the event a success.