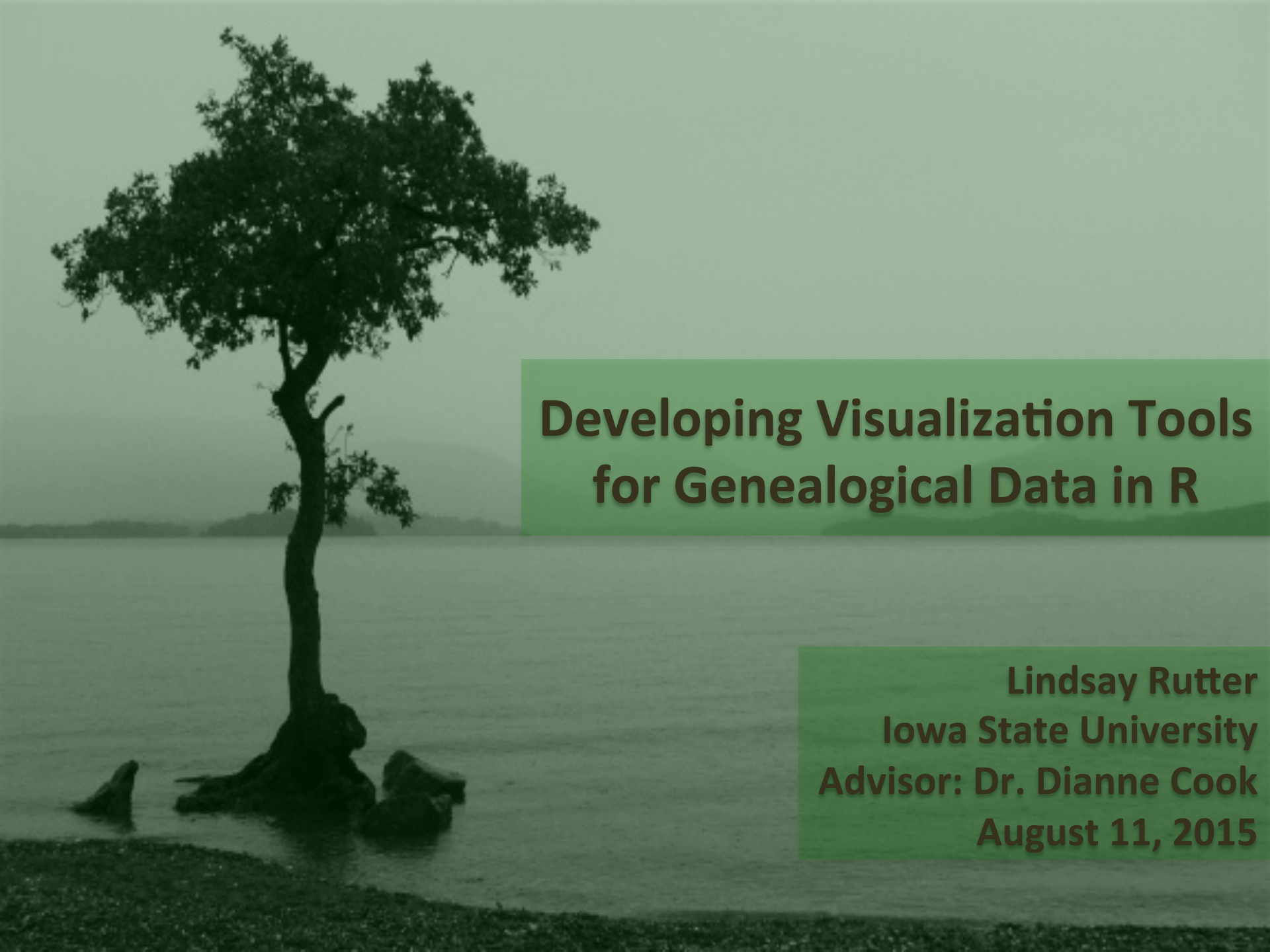


Developing Visualization Tools for Genealogical Data in R

Lindsay Rutter
Iowa State University
Advisor: Dr. Dianne Cook
August 11, 2015

OUTLINE

INTRODUCTION

AVAILABLE TOOLS

GGENEALOGY

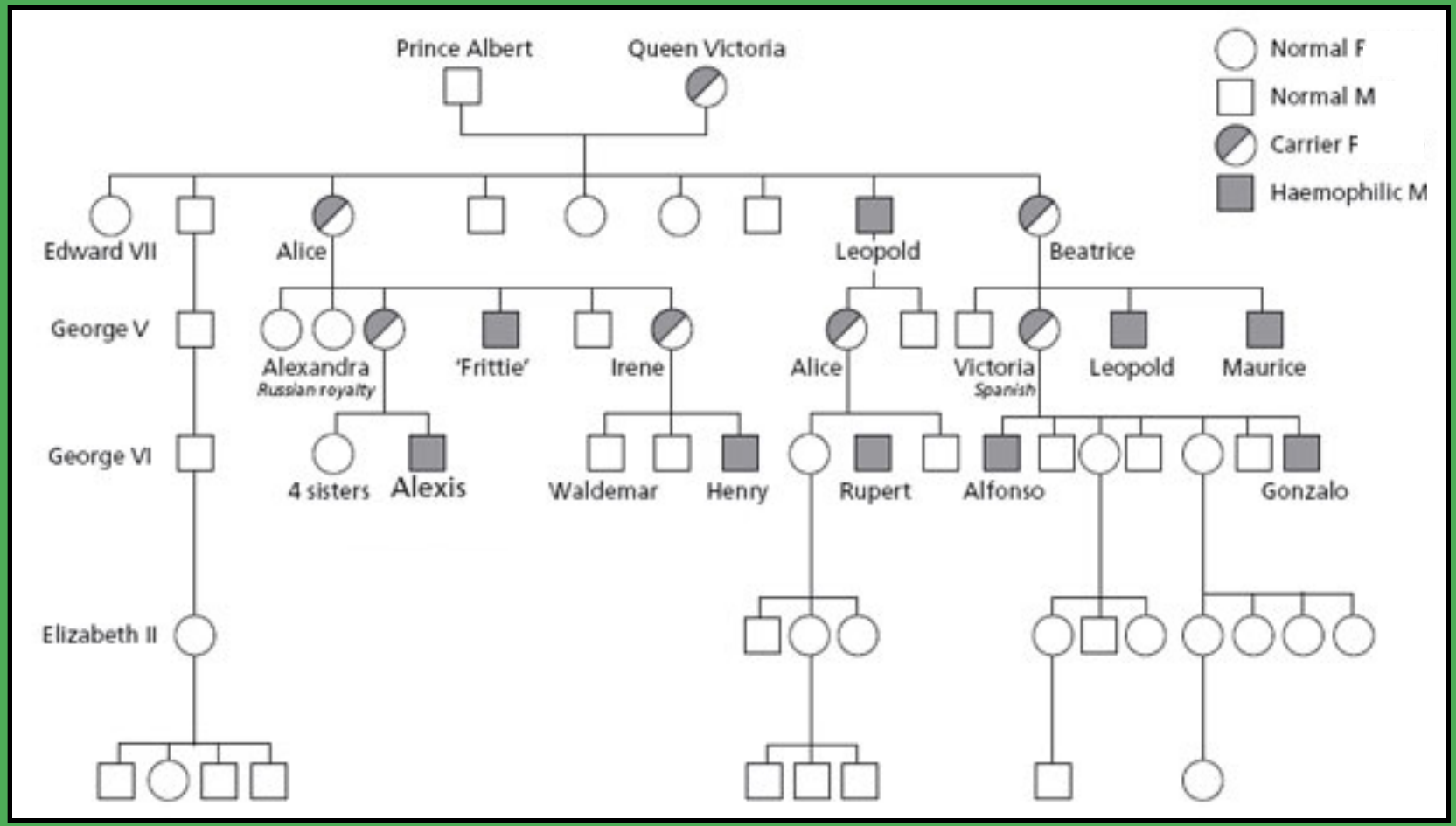
FUTURE OF GGENEALOGY

CONCLUSION

INTRODUCTION: WHAT IS GENEALOGY?

- Study of parent-child relationships
- Provides tools to better understand traits that arise in lineages
  - Desirable (disease resistance)
  - Undesirable (hemophilia)
- Can be represented visually

INTRODUCTION: WHAT IS GENEALOGY?



INTRODUCTION: MOTIVATION

Why do we want to develop and share genealogical visualization tools in R?

INTRODUCTION: R

OPEN-SOURCE

Academic statisticians
Package system

FLEXIBLE

C/C++
Shiny

CROSS-PLATFORM

Collaboration

GRAPHICS

Data visualization
Interactive

OUTLINE

INTRODUCTION

AVAILABLE TOOLS

GGENEALOGY

FUTURE OF GGENEALOGY

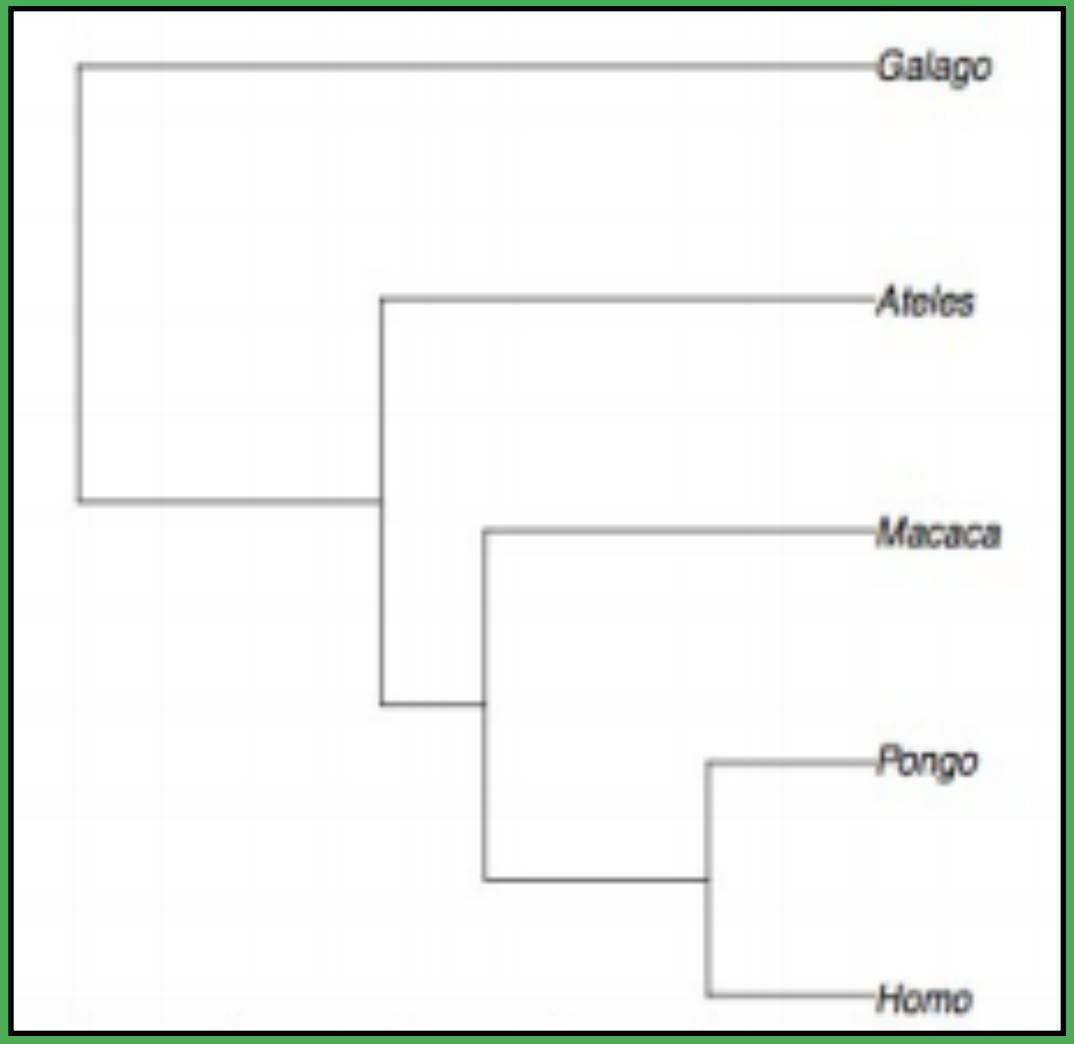
CONCLUSION

CURRENT TOOLS: APE

TITLE:
APE: Analyses of phylogenetics and evolution in R language

AUTHOR:
Emmanuel Paradis

LATEST UPDATE:
May 29, 2015

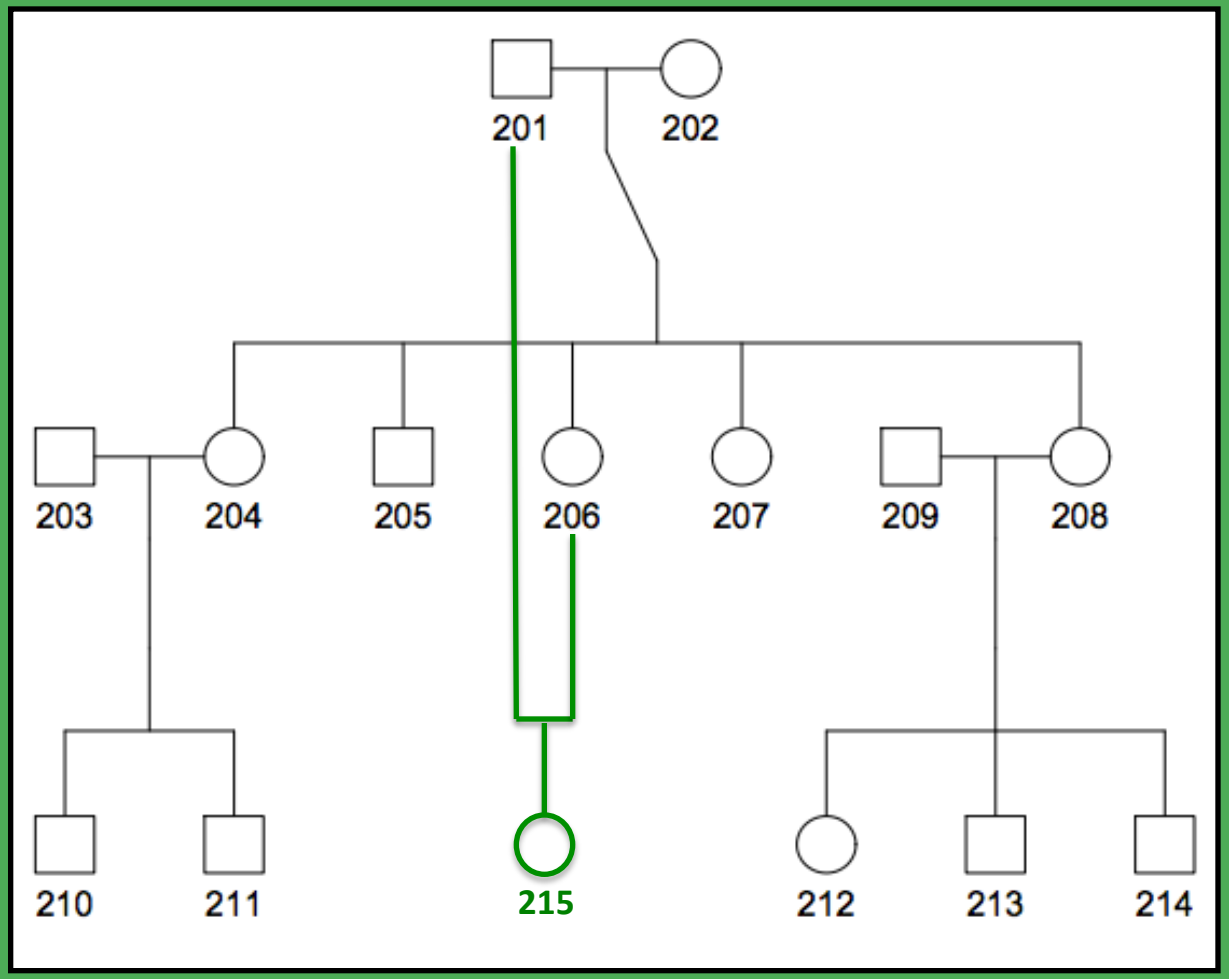


CURRENT TOOLS: KINSHIP2

TITLE:
Kinship2: Pedigree functions

AUTHOR:
Terry Therneau
Jason Sinnwell

LATEST UPDATE:
August 3, 2015

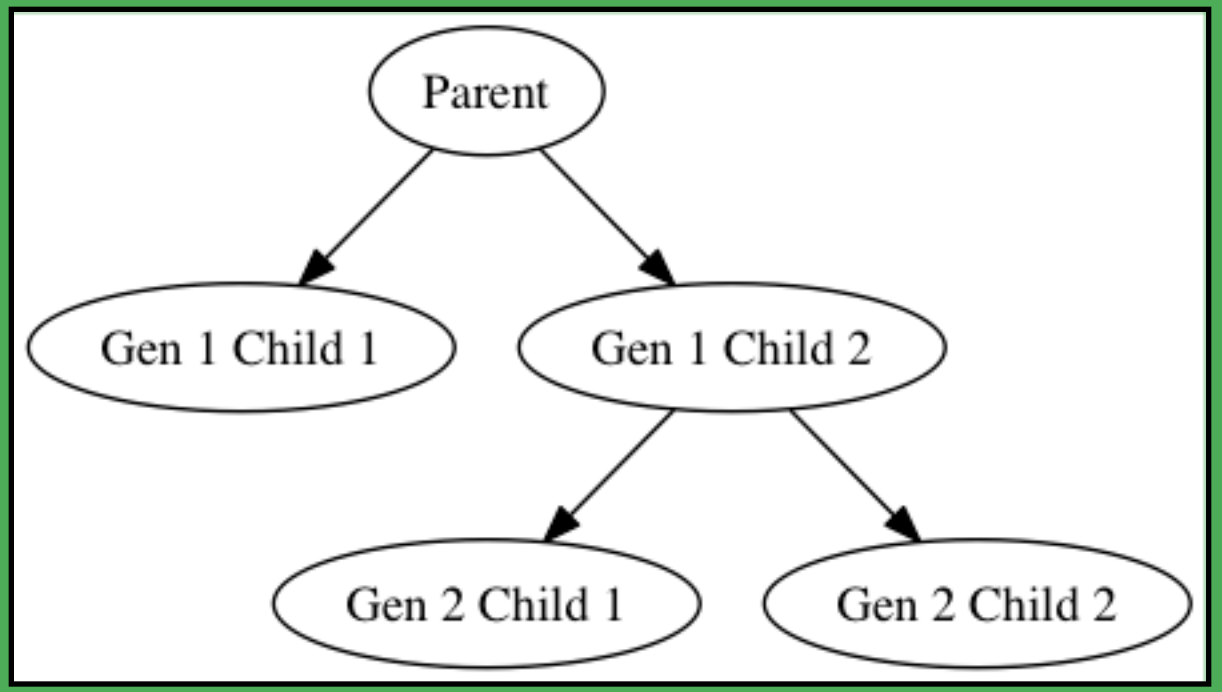


CURRENT TOOLS: OTHERS

TITLE:
Pedigree: Pedigree functions

AUTHOR:
Albart Coster

LATEST UPDATE:
November 3, 2013



Graph Drawing Software (GraphViz, Cytoscape)

OUTLINE

INTRODUCTION

AVAILABLE TOOLS

GGENEALOGY

FUTURE OF GGENEALOGY

CONCLUSION

GGENEALOGY: EXAMPLE DATASET

- Soybean variety data collected from
  - Field trials
  - Genetic studies
  - USDA bulletins
- Data frame of 412 rows (parent-child relations)
- Each variety (230)
  - Developmental years
  - Copy number variants (CNV)
  - Single nucleotide polymorphisms (SNP)
  - Protein content and yield

GGENEALOGY: PLOT SHORTEST PATH

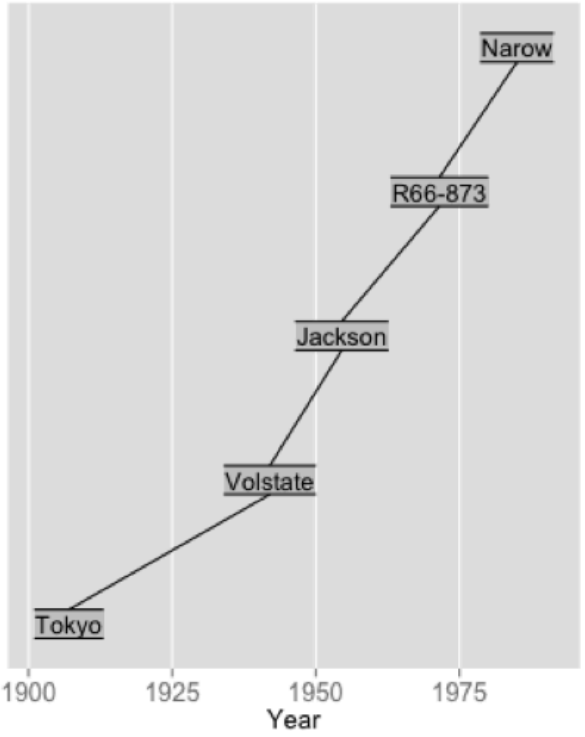


Fig. 1: The shortest path between varieties Tokyo and Narow is strictly composed of a unidirectional sequence of parent-child relationships.

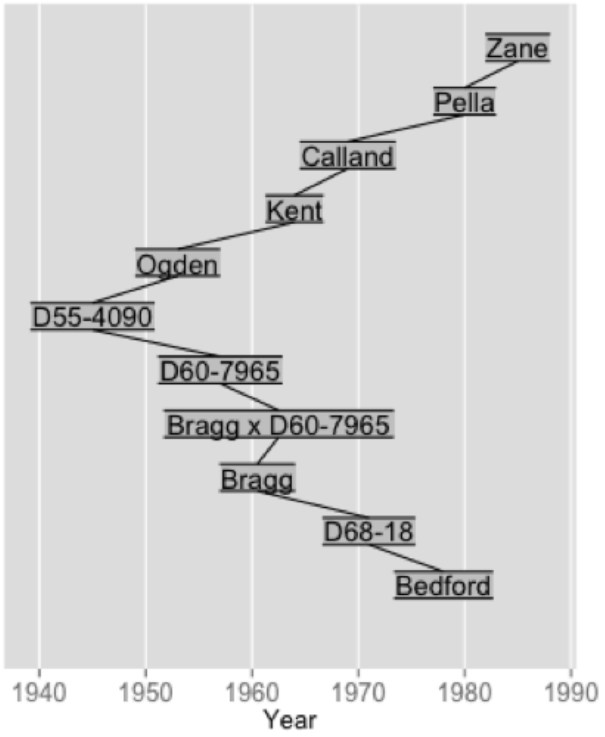


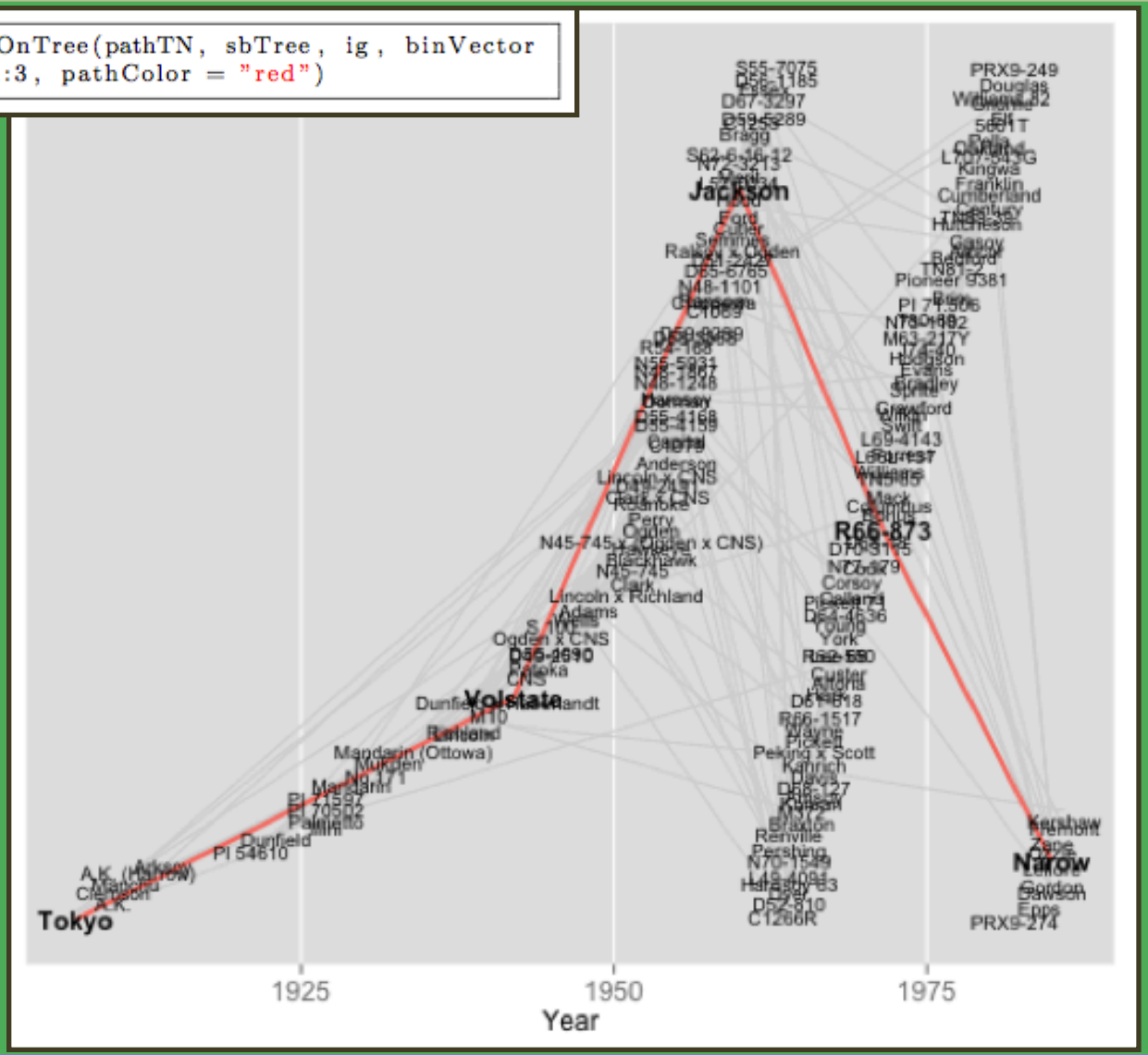
Fig. 2: The shortest path between varieties Zane and Bedford is not strictly composed of unidirectional parent-child relationships, but have a cousin-like relationship.

```
pathTN <- getPath("Tokyo","Narow", ig,
    sbTree)
plotPath(pathTN)
```

```
pathZB <- getPath("Zane","Bedford", ig,
    sbTree)
plotPath(pathZB)
```

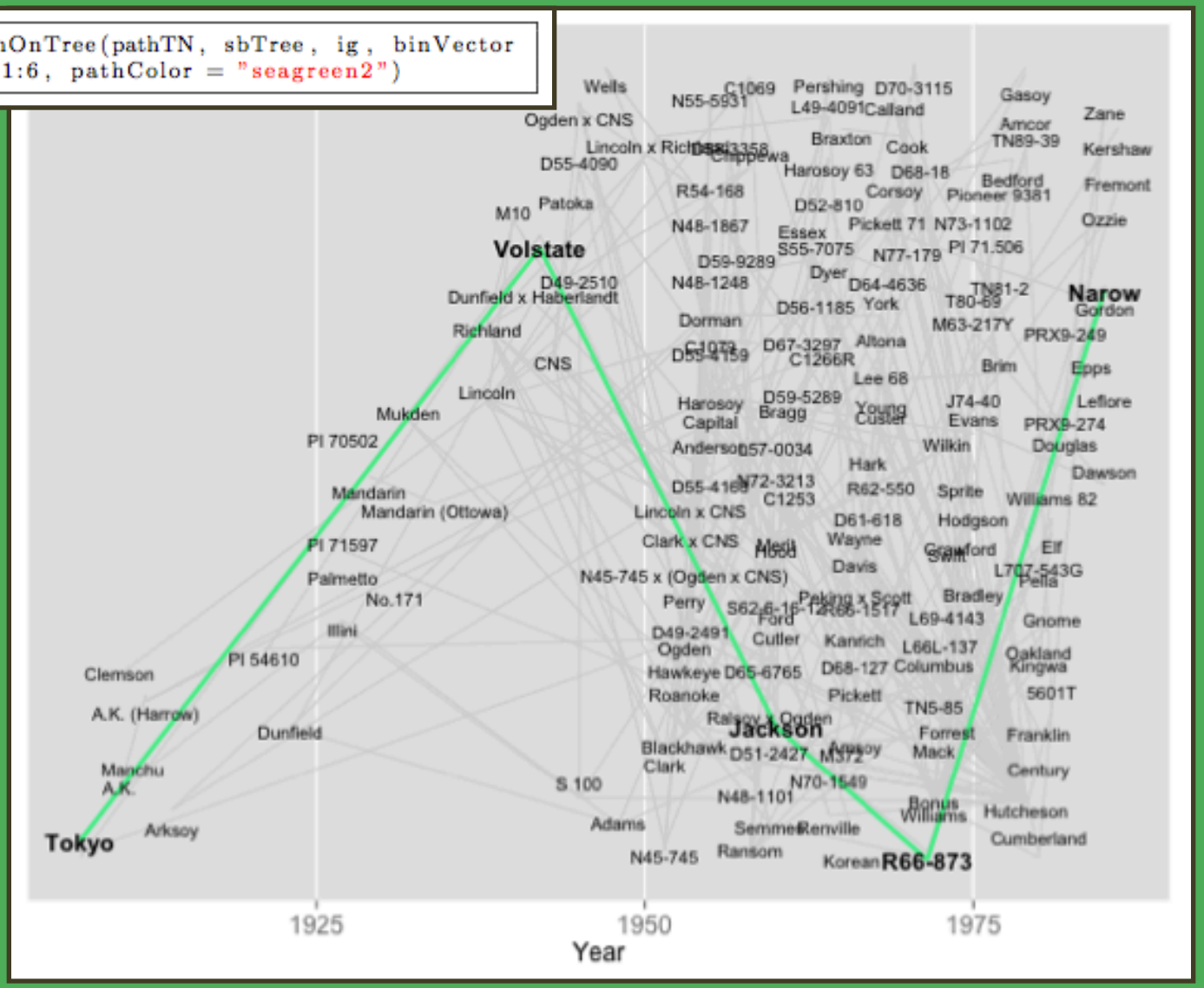
GGENEALOGY: PLOT PATH ON TREE

```
plotPathOnTree(pathTN, sbTree, ig, binVector
    = 1:3, pathColor = "red")
```

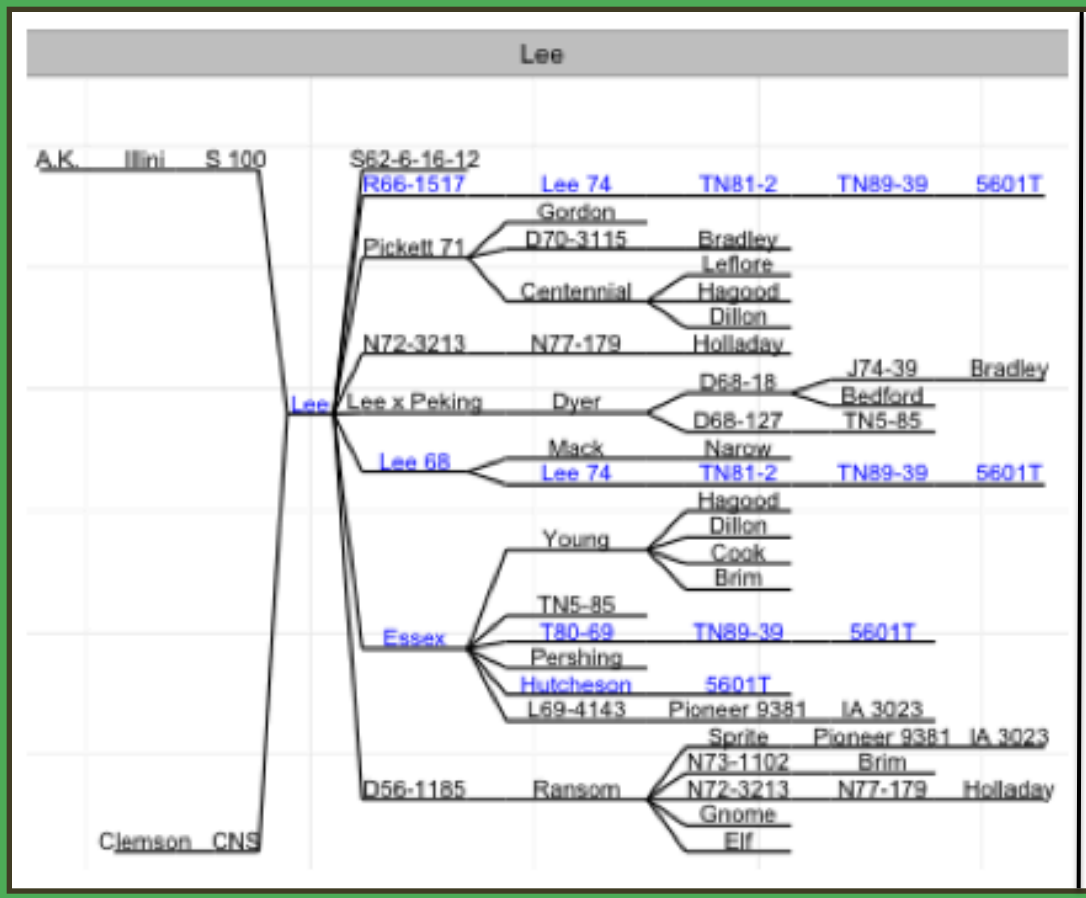


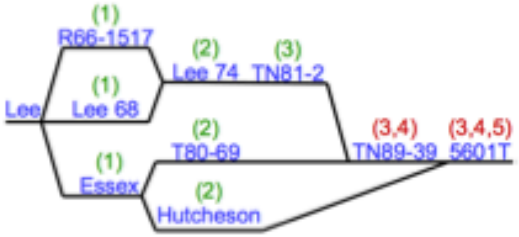
GGENEALOGY: PLOT PATH ON TREE

```
plotPathOnTree(pathTN, sbTree, ig, binVector
    = 1:6, pathColor = "seagreen2")
```



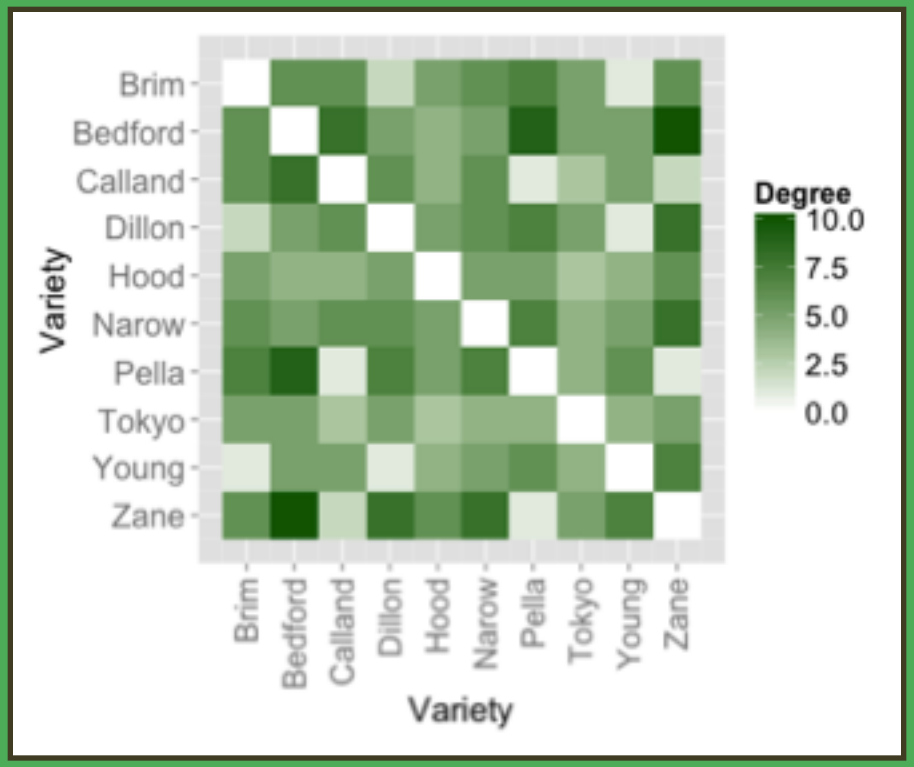
GGENEALOGY: PLOT GENERATIONS






```
plotAncDes("Lee", sbTree, mAnc = 6, mDes =
    6, vCol = "blue")
```

GGENEALOGY: PLOT DISTANCE MATRIX




```
varieties <- c("Brim", "Bedford", "Calland",
    "Dillon", "Hood", "Narow", "Pella", "
    Tokyo", "Young", "Zane")
plotDegMatrix(varieties,ig,sbTree,"Variety",
    "Variety", "Degree") + ggplot2::scale_
    fill_continuous(low="white", high="
    darkgreen")
```

OUTLINE

INTRODUCTION

AVAILABLE TOOLS

GGENEALOGY

FUTURE OF GGENEALOGY

CONCLUSION

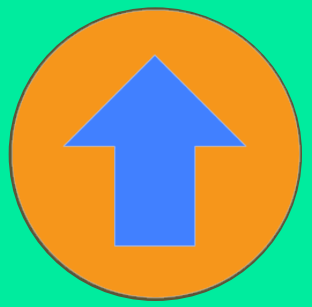
FUTURE: GGENEALOGY

EXPANSION

- LARGER DATASETS
- INTERACTIVE GRAPHICS

TESTING

- BARLEY DATASET
- MATHEMATICS GENEALOGY PROJECT



SUBMISSION

- DEVTOOLS (PASS CRAN STANDARDS)
- ROXYGEN2 (DOCUMENTATION, HELP FUNCTIONS)

OUTLINE

INTRODUCTION

AVAILABLE TOOLS

GGENEALOGY

FUTURE OF GGENEALOGY

CONCLUSION

CONCLUSION

R PACKAGE CONSTRUCTION

IMPROVED READABILITY

COLLABORATION VERSION CONTROL

VERSION UPDATE USER FEEDBACK

GGENEALOGY

GGBIO

REFERENCES

1) APE: analyses of phylogenetics and evolution in R language. E. Paradis and J. Claude and K. Strimmer. Bioinformatics. 20: 289-290 (2004).
2) Pedigree: Pedigree functions. Albart Coster (2013).
3) Kinship2: Pedigree functions. Terry Therneau and Jason Sinnwell (2015).
4) Graphviz: Graph visualization software (http://www.graphviz.org/).
5) Pedigrees of soybean cultivars released in the United States and Canada." Theodore Hyivitz, C.A. Newell, S.G. Carmer. College of Agriculture, University of Illinois at Urbana-Champaign (1977).
6) Mathematics Genealogy Project. Department of Mathematics, North Dakota State University (http://genealogy.math.ndsu.nodak.edu/).
7) Shiny: A web application framework for R (http://shiny.rstudio.com/).
8) Devtools: Tools to make developing R packages easier. Hadley Wickham and Winston Chang (2015).
9) Roxygen2: In-source documentation for R. Hadley Wickham, Peter Danenberg, Manuel Eugster (2014).
10) Ggbio: an R package for extending the grammar of graphics for genomic data. Tengfei Yin, Dianne Cook, and Michael Lawrence. Genome Biology. 13:8 (2012).

THANK YOU

DR. DIANNE COOK

SUSAN VANDERPLAS

DR. MICHELLE GRAHAM

DR. WILLIAM BEAVIS

MATH GENEALOGY PROJECT

YOU!

CONTACT

LINDSAY RUTTER
LRUTTER@IASTATE.EDU

DR. DIANNE COOK
DICOOK@MONASH.EDU