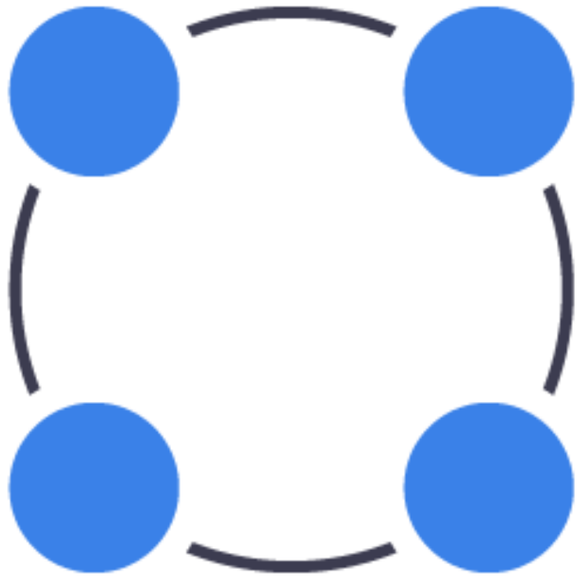


BIOINFORMATICS INSTITUTE

Implementation of GPS-L as Python package

Ivan Dmitrievsky
Tatiana Tatarinova (Prof. at USC)

ADMIXTURE

Software tool for estimation of individual ancestries

Preprocessing — creating zombies

— Modeled individuals

— Belong to K ancestral populations

— Act as basis vectors

Origin assignment procedure

— For every reference origin get mean admixture vector
— Compare given vector to mean vectors
— Choose the closest one

Goals for the summer

— Make a library, so it's easy to incorporate GPS-L into other workflows

— Make an easy to use command-line application

Obvious choice — the oldest possible version of Python.
Luckily for me Pandas supports only Python 2.7 and above.

Wrap around ADMIXTURE and plink

— Everything intermediate is written to disk (implicit global state)

— Python uses exceptions (implicit failures)

Solution

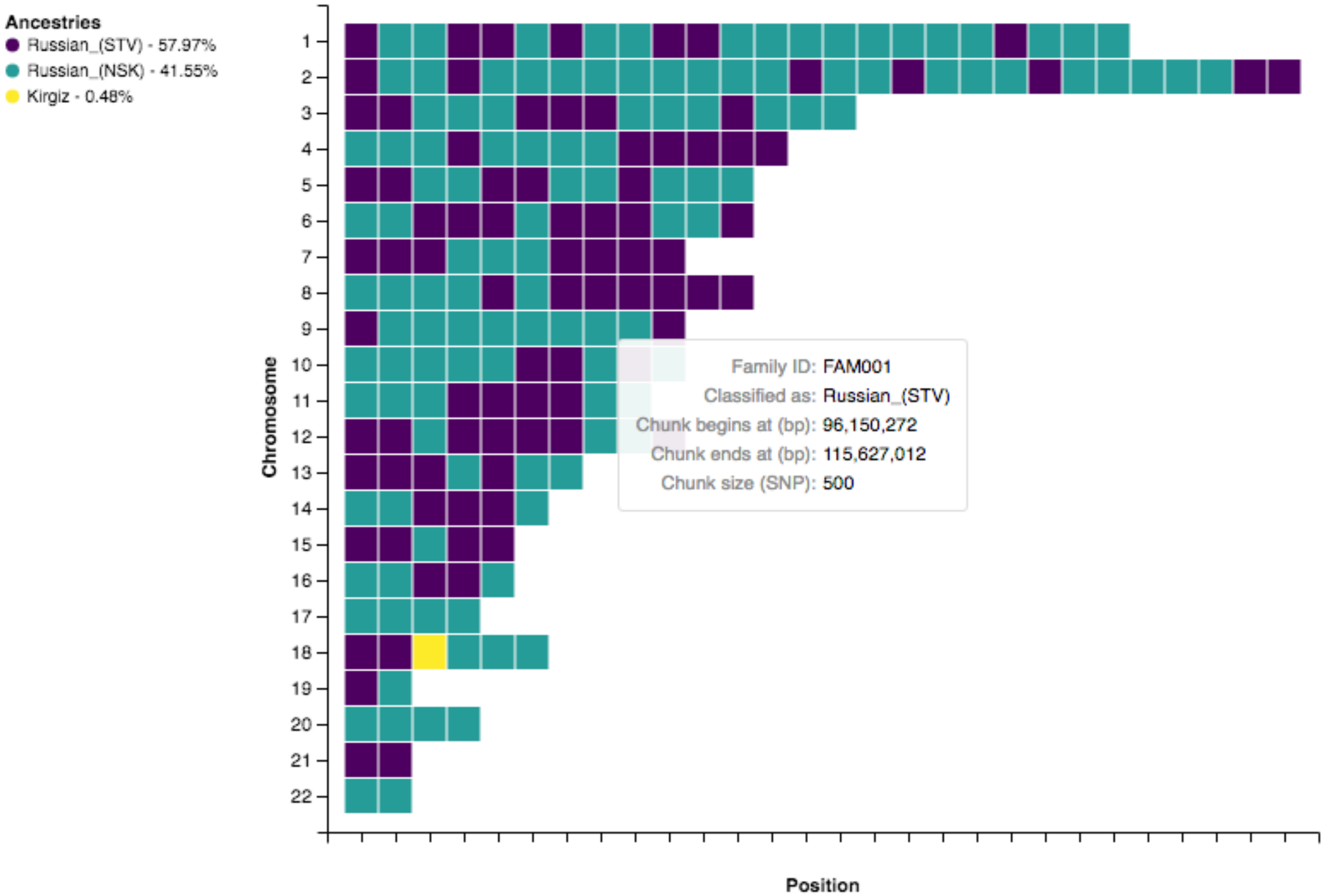
Custom context managers (luckily again Python 2.7 supports them).

Reference management

— Save datasets by id for easy referencing
— Generate and keep zombies (time-consuming task)
— Merge datasets ahead of time (reducing queries time)

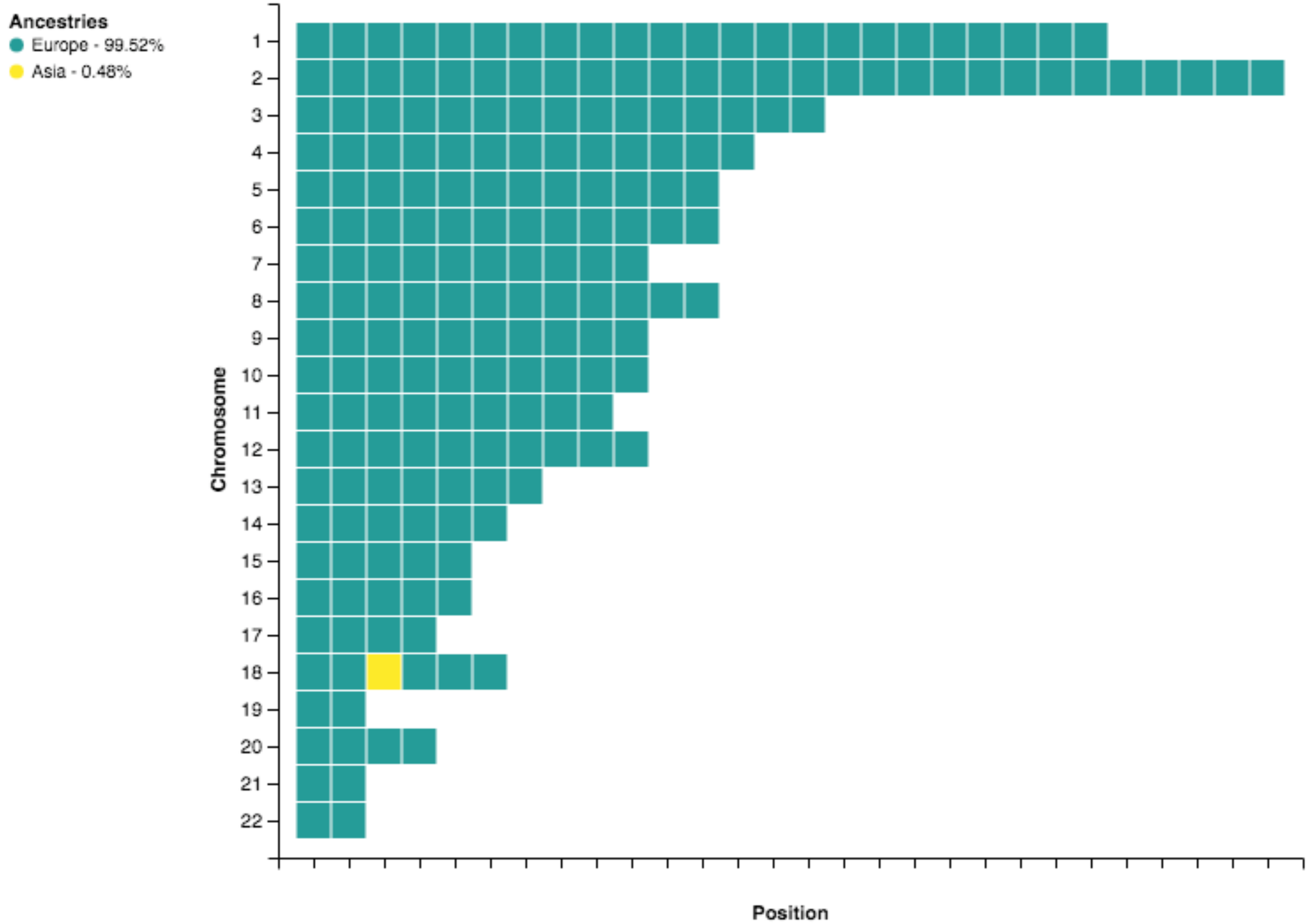
Conversion from other formats

Generic and 23 And Me exported CSV



Export as PNG View Source

Default



Export as PNG View Source

Eurasia

Pluggable pipeline

— Users can implement their own partition strategies
— Users can implement their own classification methods

When framework is in place, it's a lot easier to experiment with different steps of the pipeline.

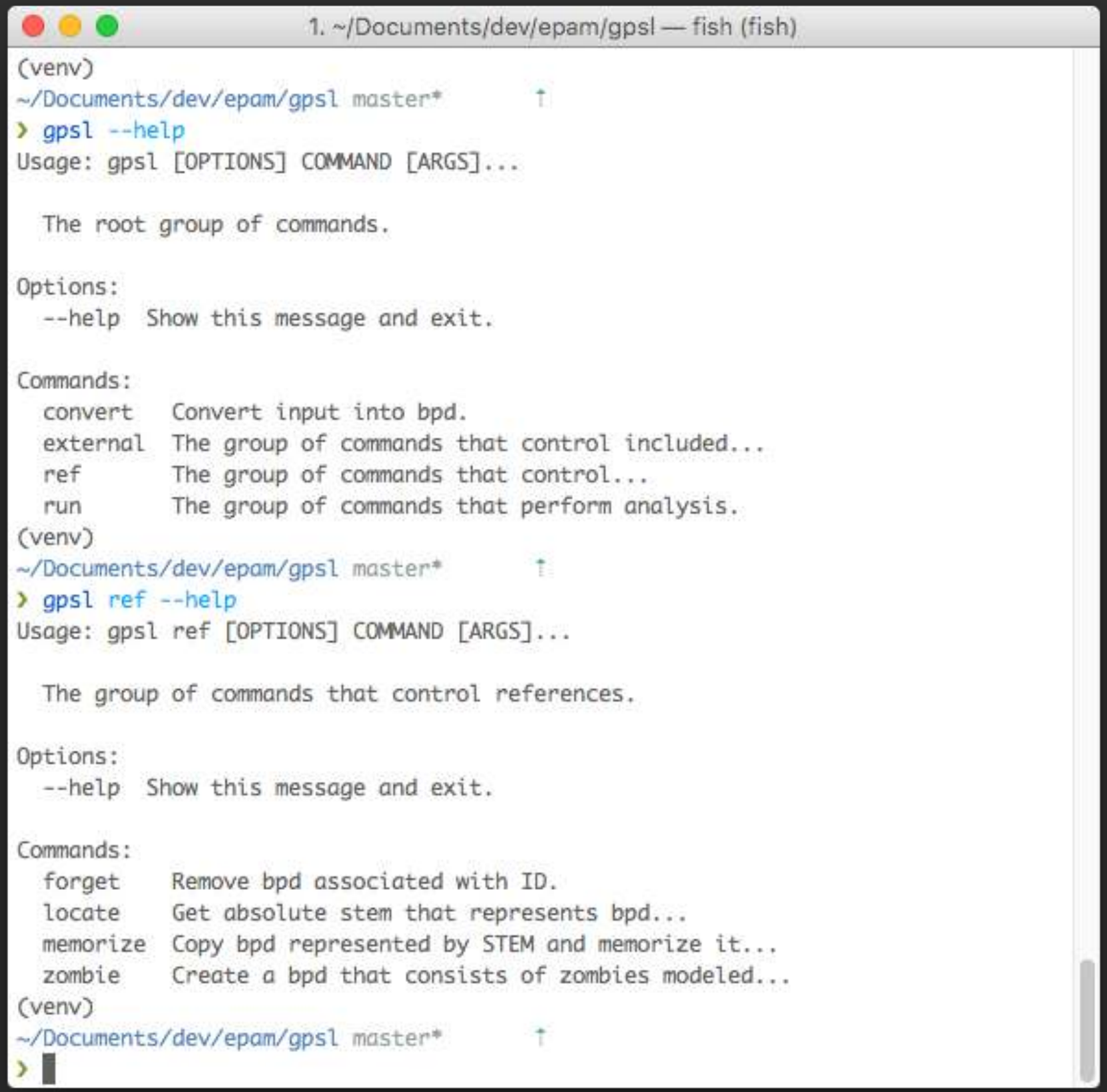
Command line interface

- Support relative paths
- Check for various FS prerequisites
- Have argument types/possible values noted in help
- Report progress to the user
- Print helpful error messages

Command line interface

- Support relative paths ✓
- Check for various FS prerequisites ✓
- Have argument types/possible values noted in help ✓
- Report progress to the user ?
- Print helpful error messages ✓

```
INFO: Started preparing complete bpd.
INFO: Merging data with merged AOT reference and zombies.
INFO: Left-hand bpd has 145032 SNPs before merging.
INFO: Right-hand bpd has 299038 SNPs before merging.
INFO: Finding and extracting common SNPs.
INFO: Trying to merge bpds.
INFO: Extracting mismatching SNPs.
INFO: Trying to merge bpds one more time.
INFO: Resulting bpd has 60705 SNPs.
INFO: Finished merging bpds.
INFO: Got bpd complete.
INFO: Started figuring out a partition.
INFO: Sorted SNPs according to bp units on chromosomes.
INFO: Started extracting chunks according to partition.
Extracting chunks according to partition  [###-------------------------------]    8%  0d 00:01:24
```


1. ~/Documents/dev/epam/gpsl — fish (fish)
(venv)
~/Documents/dev/epam/gpsl master*
› gpsl --help
Usage: gpsl [OPTIONS] COMMAND [ARGS]...
The root group of commands.
Options:
--help Show this message and exit.
Commands:
convert Convert input into bpd.
external The group of commands that control included...
ref The group of commands that control...
run The group of commands that perform analysis.
(venv)
~/Documents/dev/epam/gpsl master*
› gpsl ref --help
Usage: gpsl ref [OPTIONS] COMMAND [ARGS]...
The group of commands that control references.
Options:
--help Show this message and exit.
Commands:
forget Remove bpd associated with ID.
locate Get absolute stem that represents bpd...
memorize Copy bpd represented by STEM and memorize it...
zombie Create a bpd that consists of zombies modeled...
(venv)
~/Documents/dev/epam/gpsl master*
›

Memoization

— Users rerun their analysis with slightly tweaked arguments

— Application can try to reuse intermediate results from the previous run

— It does so only when it is safe

It helps to have pure-ish functions (explicit arguments) for implementation.

Welcome to GPSL's documentation!

Welcome to GPSL's documentation. This documentation is divided into different parts. I recommend that you get started with Installation and then head over to the Quickstart. Besides the quickstart, there is also a more detailed Tutorial that shows how to work with different options, formats, and references. If you'd rather dive into the internals of GPSL, check out the API Reference.

Some parts of this documentation are missing. These parts are listed in The State of This Documentation. If you can suggest anything else on improving this documentation — get in touch.

Todo

Provide contact information.

GPSL depends on two external tools: the ADMIXTURE software tool for maximum likelihood estimation of individual ancestries and the plink whole genome association analysis toolset. These tools are not documented here. If you want to dive into their documentation, check out the following links:

- ADMIXTURE Documentation
- plink Documentation

User's Guide

This part of the documentation, which is mostly prose, begins with some background information about GPSL, then focuses on step-by-step instructions for working with GPSL.

Conclusion

— Write libraries, it's easier to use them
— Provide at least basic documentation of internals in this case
— GPSL exists, works, but needs a little bit more polishing before releasing

Thank you