STRATEGIES FOR THE DEVELOPMENT OF PANELS OF GENETIC MARKERS FOR HUMAN IDENTIFICATION

Oleg Yasnev, Nadezhda Pilschikova
Advisors: Alexander Pavlov, Anton Bragin

Parseq Lab

Subject

- In forensic science there are genetic tests for human identification
- Main identifier is an allele
- Each allele has a frequency in each population
- Genetic markers:
  - SNP (Single Nucleotide Polymorphisms)
  - STR (Short Tandem Repeat)
- Each genetic marker has 2 or more alleles

Tasks

1. Panel for human identification:

   for a given set of populations create a panel of genetic markers such that it is possible to identify a person with a given confidence level.

2. Panel for a human ancestry determination:

   for a given set of populations create a panel of genetic markers such that it is possible to determine a person's ancestry with a given confidence level.

Task 1: human identification. Subtasks

- Data import
- Data analysis and filtering
- For a given population calculate match probability of a given genetic marker
- For a given population calculate match probability of a set of genetic markers
- For a given set of populations calculate match probability of a set of genetic markers
- Create panel: take unlinked markers with best efficiency
- Test our panel and compare it with existing ones

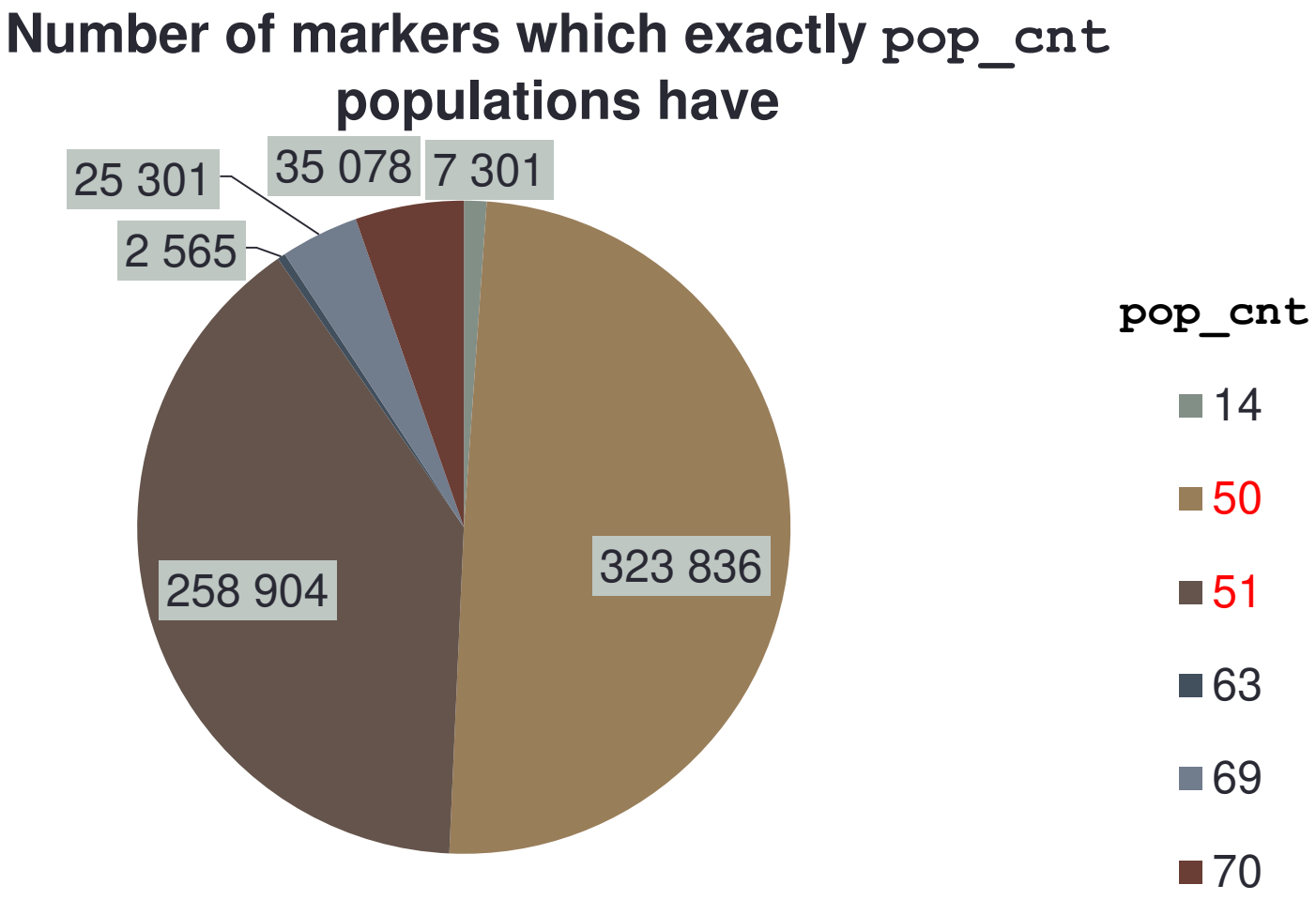
Data import

- ALFRED
- STRBase
- MURKA
- Rutgers Map v.3

Data analysis

Common statistics	
Genetic markers (variants)	28 268 174 (57 370 391)
Populations	473
Cohorts	4 674
Allele frequencies	73 685 223
Discovered variants	2,3% (1 322 256)
Predefined panels	19

Markers distribution by populations



Data filtering

Incorrect data*	
Locuses with 1 allele	58
Allele frequency ∉ [0, 1]	0
Sum allele frequencies ≠ 1 ± 0.01	0,27%

- Filtering scripts for database were written

*Approximation based on half-sized database

Problematic data

- There are 7,4% (5 457 875) of cohorts with different allele frequencies for the same genetic variant in the same population ($\Delta = 0.05$)
- Most of them has $\Delta \leq 0.2$
- But there are even $\Delta = 1$ (!)

- We select the largest cohort, if there are several
- It is not always optimal

Calculate match probability (mp)

One marker and one population

- Suppose marker has two alleles A and B
- Possible genotypes are: AA, BB, AB, BA
- $mp = (P(A) \cdot P(A))^2 + (P(B) \cdot P(B))^2 + (2 \cdot P(A) \cdot P(B))^2$

- Common case (k alleles)
- Go over all possible genotypes (allele pairs)
- If it is a homozygote (AA), $p_i = (P(A) \cdot P(A))^2$
- If it is a heterozygote (AB), $p_i = (2 \cdot P(A) \cdot P(B))^2$ (as AB and BA are the same)
- $mp = \sum p_i$

Calculate match probability (mp)

One marker and set of populations

- p_i – probability of i-th population in region
- mp_i – match probability of marker and i-th population
- If there is no data for a population consider it to be the worst ($mp_i = 1$)
- $mp = \sum p_i \cdot mp_i$

- For now we do not know p_i
- We suppose all populations to be equally probable

Calculate match probability (mp)

Set of markers and set of populations

- Calculate match probability for each marker independently
- Get a list of markers with match probability
- Suppose markers are inherited independently
- $mp = \prod mp_i$

- Next step is to select best markers and create a panel of minimum size

Compare model results

- Match probability of SNPforID (panel of 52 SNP markers) is 3.0×10^{-21}
- The best marker has $mp = 0.375$
- The best panel of 52 markers has

$$mp = 0.375^{52} = 7.1 \times 10^{-23}$$

- The model seems to be correct!

Create a panel

If all markers were unlinked

- Greedy algorithm: we take markers with best mp until we reach the confidence level

But markers can be linked

- Problem: from the list of markers select the minimum number of unlinked markers to reach the confidence level
- We know
  - match probability of each marker
  - distance between markers in centimorgans (from Rutgers Map)

What to do with markers linkage

- We do not know genotypes frequencies in populations
- Should suppose the worst case

- Possible approaches:
  1. Not consider linkage. Suitable only for testing.
  2. Distance threshold. Can be rough.
  3. Markers efficiency depends on physical distance. Details are not obvious.

Dynamic algorithm

- Suppose n – number of markers, mp_i – match probability of i-th marker, d – distance threshold
- $MP[i]$ – minimum match probability with using first i markers
- If we do not use i-th marker, $MP[i] = MP[i-1]$
- If we use i-th marker, $MP[i] = mp_i \cdot MP[prev(i)]$,
  $prev(i)$ – is the largest index j such that $d_i - d_j < d$
- $MP[0] = 1$ (no markers – worst case)
- $MP[i] = \min\{MP[i-1], mp_i \cdot MP[prev(i)]\}$

Dynamic algorithm

- Complexity: $O(n)$

- There is no limitations for panel size
- We can get the maximum best panel and then remove the least efficient markers – not necessary optimal

- For multiple chromosomes we can linearly order chromosomes, use the single coordinate system, and use gaps between chromosomes larger than distance threshold

- Not obvious how to consider the third linkage approach

Graph algorithm

- Build directed graph
- Vertices: *source*, *sink* and *n* vertices corresponding to markers
- From *source* draw edge to each vertex. Weight is match probability
- From each vertex draw edge to *sink* of weight 0
- Consider linkage approaches:
  1. From current vertex draw edge to each lower ones
  2. Draw edge only if physical distance is larger than threshold
  3. Edge weight depends on physical distance between markers

Graph algorithm

- Now implemented with Bellman-Ford algorithm
- Complexity: $O(|V| \cdot |E|) = O(n^3)$ (worst case)
- Core of the algorithm remains. Only edges weights change
- For multiple chromosomes we can concatenate $source$ and $sink$

Test our panel

- Took panel “19 markers”*
- Took the same 40 populations and get list of markers for them
- All of 19 markers were in the list
- Maximum $mp_i = 0.42$, so $mp \leq 0.42^{19} = 6.9 \times 10^{-8}$
- In the article $mp < 1 \times 10^{-6}$
- Our results correlate with articles ones

*K.K. Kidd et al. Developing a SNP panel for forensic identification of individuals (2006)

Implementation

- PostgreSQL
- Scripts for DB creation and data import (SQL and Python)
  full DB and test DB (only 21 chr)
- Scripts for data analysis and filtering (SQL)
- Functions (PL/pgSQL):
  - **`get_panel`** – get list of markers with match probabilities for a given set of populations
  - **`get_panel_match_prob`** – get match probability of a given panel

Benchmarks

Before optimization	
100 markers	130 sec
After	
10000 markers	40 sec

- Optimization: ~300 times faster

- On a full database:

Populations	Time, min	Markers
1	10	650 000
40	75	660 000

Current progress

- Full database with genetic variants, populations, cohorts, panels, chromosome coordinates
- Scripts for DB creation, data import, analysis and filtering
- Can get a list of markers for a set population or estimate a panel in reasonable time
- First version of algorithm to consider markers linkage
- Our model and algorithm are proved theoretically
- Results correlate with existing ones