Focal Individual Analysis Workflow

This document describes the complete workflow for creating and analyzing focal individual databases for kinship analysis using likelihood ratios (LRs). The workflow starts after creating a database of unrelated individuals and focuses on analyzing how well related individuals can be identified among unrelated individuals.

Prerequisites

Before starting this workflow, you must have:

- **Unrelated database**: `output/sim_processed_genotypes_unrelated_database.csv`
- **Parent-child data**: `output/sim_processed_genotypes_{POPULATION}_parent_child_combined.csv`
- **Allele frequencies**: `data/df_allelefreq_combined.csv`
- **Core loci definitions**: `data/core_CODIS_loci.csv`

Workflow Steps

Step 1: Extract Focal Individual

Script: `extract_focal_individual.sh`

Purpose: Extract a single individual from the parent-child database to serve as the "focal individual"

Usage:

```bash
./extract_focal_individual.sh <population> [focal_individual_number]
# Examples:
./extract_focal_individual.sh AfAm           # Extract first individual
./extract_focal_individual.sh AfAm 3         # Extract third individual
```

Inputs:

- `output/sim_processed_genotypes_{POPULATION}_parent_child_combined.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_individual_{POPULATION}.csv`

- `output/focal_database/{POPULATION}/focal_individual_{POPULATION}_{N}.csv` (for additional focals)

Output Format:

```
csv

population,focal_id,locus,focal_allele1,focal_allele2
AfAm,1234,D8S1179,12,15
AfAm,1234,D21S11,28,30
...
```

Step 2: Create Focal-Unrelated Pairs

Script: `create_focal_pairs.sh`

Purpose: Create all possible pairs between the focal individual and every individual in the unrelated database

Usage:

```
bash

./create_focal_pairs.sh <population>
```

Inputs:

- `output/sim_processed_genotypes_unrelated_database.csv`
- `output/focal_database/{POPULATION}/focal_individual_{POPULATION}.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_unrelated_pairs_{POPULATION}.csv`

Output Format:

```
csv

population,relationship_type,pair_id,locus,focal_allele1,focal_allele2,ind2_allele1,in
AfAm,unrelated_focal,1,D8S1179,12,15,14,16
AfAm,unrelated_focal,1,D21S11,28,30,29,31
...
```

Step 3: Simulate Related Individuals

Script: `simulate_relationships_4focal.R`

Purpose: Create simulated individuals with known relationships to the focal individual

Usage:

```bash
Rscript simulate_relationships_4focal.R <population> <num_pairs_per_relationship>
# Example:
Rscript simulate_relationships_4focal.R AfAm 50
```

Inputs:

- `output/focal_database/{POPULATION}/focal_individual_{POPULATION}.csv`
- `data/df_allelefreq_combined.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_full_siblings_focal_pairs_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_half_siblings_focal_pairs_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_cousins_focal_pairs_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_second_cousins_focal_pairs_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_all_relationships_{POPULATION}.csv` (combined)

Relationship Types Simulated:

- `full_siblings_focal`: k0=1/4, k1=1/2, k2=1/4
- `half_siblings_focal`: k0=1/2, k1=1/2, k2=0
- `cousins_focal`: k0=7/8, k1=1/8, k2=0
- `second_cousins_focal`: k0=15/16, k1=1/16, k2=0

Step 4: Calculate Likelihood Ratios

Script: `calculate_lrs_4focal.R`

Purpose: Calculate likelihood ratios for all focal-related pairs using parent-child and full-sibling hypotheses

Usage:

```bash
Rscript calculate_lrs_4focal.R <population>
```

Inputs:

- `output/focal_database/{POPULATION}/focal_all_relationships_{POPULATION}.csv`
- `data/df_allelefreq_combined.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_all_relationships_with_LR_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_combined_LR_{POPULATION}.csv`

LR Columns Added:

- `LR_parent_child_{POPULATION}` (for each population: AfAm, Cauc, Hispanic, Asian)
- `LR_full_siblings_{POPULATION}` (for each population)

Combined LR Columns:

- `{loci_set}_LR_parent_child_{POPULATION}` (for each loci set and population)
- `{loci_set}_LR_full_siblings_{POPULATION}` (for each loci set and population)

Step 5: Calculate LRs for Focal-Unrelated Pairs

Script: `focal_unrelated_lr_script.R`

Purpose: Calculate likelihood ratios for focal-unrelated pairs (these should have low LRs)

Usage:

```bash
Rscript focal_unrelated_lr_script.R <population>
```

Inputs:

- `output/focal_database/{POPULATION}/focal_unrelated_pairs_{POPULATION}.csv`
- `data/df_allelefreq_combined.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_unrelated_with_LR_{POPULATION}.csv`
- `output/focal_database/{POPULATION}/focal_unrelated_combined_LR_{POPULATION}.csv`

Step 6: Ranking Analysis

Script: `analyze_focal_ranking.R`

Purpose: Analyze how well related individuals rank against the unrelated database

Usage:

```bash
Rscript analyze_focal_ranking.R <population>
```

Inputs:

- `output/focal_database/{POPULATION}/focal_combined_LR_{POPULATION}.csv`
- `output/unrelated_database/{POPULATION}/unrelated_LR_{POPULATION}.csv`

Outputs:

- `output/focal_database/{POPULATION}/focal_individual_ranking_summary_{POPULATION}.csv`
- Console output with detailed ranking analysis

Output Directory Structure

```
output/
├── focal_database/
│   ├── {POPULATION}/                                    # e.g., AfAm, Cauc,
Hispanic, Asian
│   │   ├── focal_individual_{POPULATION}.csv            # Extracted focal
individual
│   │   ├── focal_unrelated_pairs_{POPULATION}.csv       # Focal vs unrelated pairs
│   │   ├── focal_unrelated_with_LR_{POPULATION}.csv     # Focal-unrelated with LRs
│   │   ├── focal_unrelated_combined_LR_{POPULATION}.csv # Combined LRs for focal-
unrelated
│   │   ├── focal_all_relationships_{POPULATION}.csv     # All simulated
relationships
│   │   ├── focal_all_relationships_with_LR_{POPULATION}.csv # With individual LRs
│   │   ├── focal_combined_LR_{POPULATION}.csv           # Combined LRs for related
pairs
│   │   ├── focal_individual_ranking_summary_{POPULATION}.csv # Ranking analysis
results
│   │   ├── focal_validation_overall_{POPULATION}.csv    # Validation statistics
│   │   ├── focal_validation_by_family_{POPULATION}.csv  # Family-specific
validation
│   │   ├── timing_log_LR_{POPULATION}.csv              # Performance timing
│   │   └── family_{FOCAL_ID}/                          # Individual family
directories
│   │       ├── full_siblings_focal_pair_1.csv
│   │       ├── full_siblings_focal_pair_2.csv
│   │       ├── full_siblings_focal_all_pairs.csv
│   │       └── ...
│   └── combined_analysis/
│       └── multi_focal_summary.txt                     # Summary for multiple
focals
└── unrelated_database/
    └── {POPULATION}/
        └── unrelated_LR_{POPULATION}.csv               # Unrelated database with
LRs
```

Multi-Focal Analysis

Script: `create_focal_database_multi.sh`

For analyzing multiple focal individuals simultaneously:

Usage:

bash

```
./create_focal_database_multi.sh [num_pairs_per_relationship] [num_focal_individuals]
# Examples:
./create_focal_database_multi.sh 50 3 AfAm Cauc    # 3 focals for AfAm and Cauc
./create_focal_database_multi.sh 50 2              # 2 focals for all populations
```