Kinship LENGTH_CM Distribution Analysis Report

Overview

This report documents the exploratory data analysis (EDA) of kinship LENGTH_CM (centiMorgans) distributions across 7 kinship categories (kinship_1 through kinship_6, and kinship_UN).

Generated: November 25, 2025
Data Source: `/mg6-18t/prj/2506_LHC_SPO_Kinship/model_improvement/cbj_process/Analysis/Proto/ibis_test/gene`

Objectives

1. Create individual LENGTH_CM distribution line plots for each kinship category
2. Show both count and percentage (%) on individual plots
3. Create a combined overlay plot for comparing all kinship distributions
4. Filter UN data to remove duplicate experimental variants (X-1_vs_X-2 patterns)

Methodology

Data Processing

Input Files:

- kinship_1.tsv through kinship_6.tsv (kinship 1-6)
- kinship_UN.tsv (unrelated pairs)

Filtering (UN file only):

- Removed records matching pattern `X-1_vs_X-2` where X ∈ [1-6]
- Rationale: `-1` and `-2` suffixes represent same sample with different experimental runs
- Records removed: 135 out of 6,805,861
- Final UN records: 6,805,726

Distribution Calculation

- **X-axis**: LENGTH_CM range (1 to MAX value, binned by integer cM)
- **Y-axis (Individual Plots)**:
  - Primary: Count (absolute number of segments)
  - Secondary: Percentage (% of total records in that kinship)
- **Y-axis (Combined Plot)**:
  - Percentage only (%) for cross-kinship comparability

Visualization

- Individual plots: Dual Y-axis (count + percentage) with distinct colors
- Combined plot: 7 overlaid lines with unique colors for each kinship category
- All plots: 150 DPI resolution, high contrast markers and lines

---

Key Findings

Summary Statistics

Kinship	Total Records	LENGTH_CM Range	Max Count	Max %	Peak Location
1	190,374	0.00-53.74 cM	26,974	14.17%	1 cM
2	226,391	0.00-130.68 cM	20,105	8.88%	1 cM
3	345,553	0.00-53.35 cM	25,614	7.41%	1 cM
4	686,469	0.00-47.70 cM	34,076	4.96%	1 cM
5	561,483	0.00-43.52 cM	22,806	4.06%	1 cM
6	94,083	0.00-35.83 cM	3,380	3.59%	1 cM
UN	6,805,726	0.00-249.15 cM	232,317	3.41%	1 cM

Observations

1. **Total IBD Segments**: ~9.0 million segments across all kinship categories
2. **Close vs. Distant Relatives**:
   - **Kinship 1-2** (closest): Shorter maximum LENGTH_CM (~50-130 cM), more concentrated distributions
   - **Kinship 5-6** (distant): Shorter maximum LENGTH_CM (~35-43 cM), lower peak percentages
   - **Unrelated (UN)**: Much longer LENGTH_CM range (up to 249 cM), but lower concentration
3. **Distribution Pattern**: All kinship categories show peak concentration at 1 cM bin
4. **Length Differentiation**:
   - Close relatives (1-3): 14.17%, 8.88%, 7.41% at peak
   - Distant relatives (4-6): 4.96%, 4.06%, 3.59% at peak
   - Unrelated: 3.41% at peak

Output Files

All visualization outputs saved to: `reports/kinship_length_analysis/`

Individual Plots (7 files)

- `kinship_1_length_distribution.png` - Kinship 1 (parents/siblings)
- `kinship_2_length_distribution.png` - Kinship 2 (grandparents/aunts-uncles)
- `kinship_3_length_distribution.png` - Kinship 3 (first cousins)
- `kinship_4_length_distribution.png` - Kinship 4 (first cousins once removed)
- `kinship_5_length_distribution.png` - Kinship 5 (second cousins)
- `kinship_6_length_distribution.png` - Kinship 6 (more distant)
- `kinship_UN_length_distribution.png` - Unrelated (filtered)

Combined Plot (1 file)

- `kinship_combined_length_distribution.png` - All 7 kinship categories overlaid

Technical Details

Script Information

- **Location**: `scripts/kinship_length_analysis.py`
- **Language**: Python 3.11
- **Dependencies**: pandas, matplotlib, numpy
- **Execution Time**: ~30 seconds

Data Verification

- All files successfully loaded from `data/raw/`
- No missing values in LENGTH_CM or kinship columns
- UN filtering validation: 135 records removed (0.002% of data)

Next Steps

This analysis provides baseline understanding of kinship LENGTH_CM distributions:

1. **Visualization Task**: ✅ Complete
2. **Model Training Task**: Pending receipt of processed IBD data from 채병주
3. **Feature Engineering**: Will use segment length distributions identified in this analysis
4. **Model Comparison**: Results will be compared against 2024 study under same conditions

---

Notes

- X-axis bins start from 1 cM (values < 1 cM rounded to 1)
- UN file filtering removed only 135 duplicate samples (~0.002%), minimal impact
- Combined plot uses percentage (%) to account for different record counts per kinship
- All plots include grid overlay for easier value reading

Report generated by `scripts/kinship_length_analysis.py`
Documentation: `docs/kinship_length_analysis/`