HaplotypeRichness.Estimation (HR: HRiD / HRiP)

Concept — HRiD (and HRiP) in a nutshell

Haplotype Richness Drop (HRiD) is a within-population statistic that identifies sudden local declines in haplotype richness as evidence of positive selection. It uses the effective number of haplotypes (nh)—the haplotypic analogue of Kimura–Crow's effective number of alleles—as the diversity measure.

In a selective sweep, the favored haplotype increases in frequency, and the focal region tends to show reduced diversity compared to its immediate flanks. Contrasting the focal window to its neighbors yields high raw HR where such drops occur.

HRiD was first introduced on the X chromosome to leverage male hemizygosity (exact haplotypes) and shown to complement eROHi, iHS, and nSL (*Shihabi et al., 2022*). It proved robust to recombination heterogeneity because it is defined as a local contrast across adjacent windows.

Subsequent work generalized HRiD to autosomes (using phased haplotypes) and used it alongside "classical" scans. The principle remained the same: local drops in nh highlight sweep candidates, while the approach remains complementary to ROH- and EHH-based statistics.

How to read HR, HRiD and HRiP

- **Raw HR before normalization**
    - **HRiD ↔ HR > 1:** center has lower nh than flanks → diversity drop → candidate positive selection.
    - **HRiP ↔ HR < 1:** center has higher nh than flanks → diversity increase → potential balancing selection.
- **After optional Z-score normalization**
    - One-tailed p-values are reported from opposite Z tails:
        - HRiD_Pvalue = pnorm(−Z): small when Z is large positive (HR is high; richness drop).
        - HRiP_Pvalue = pnorm(Z): small when Z is large negative (HR is low; richness increase).

TL;DR

- HRiD detects decreases in haplotype richness (positive selection).

- HRiP flags increases (can indicate balancing selection).

What the function does

HaplotypeRichness.Estimation() computes window-based nh and a local HR statistic from phased VCF data. It optionally applies Z-score normalization and one-tailed p-values for HRiD (decrease) and HRiP (increase). It supports both autosomes and sex chromosomes and includes optional haploid/hemizygosity handling.

Calculation modes (pick one combination)

1. **Bp.based + non-consecutive:** fixed bp windows sliding by slide bp.
2. **Bp.based + consecutiveSNP=TRUE:** for each SNP, analyze a bp-length window centered at that SNP (plus left/right neighbors for the contrast). Ignores slide.
3. **SNP.based + non-consecutive:** fixed SNP-count windows sliding by slide SNPs.
4. **SNP.based + consecutiveSNP=TRUE:** for each SNP, analyze SNP-length windows (center/left/right) anchored at the focal SNP.

In every mode, the function counts unique haplotypes per window (from haplotype strings derived from the VCF), converts them to nh, and then forms the local contrast ratio (the raw HR value). With normalization enabled, it adds Z, HRiD_Pvalue (via −Z), and HRiP_Pvalue (via +Z).

What happens at chromosome edges (edge windows)

HR compares a center window to its neighbors. At the very start or end of a chromosome, symmetric neighbors are not available, so the function uses one-sided contrasts:

- **Sliding-window modes (consecutiveSNP=FALSE):** interior windows use the average of left/right vs center; the first window uses only its right neighbor; the last window uses only its left neighbor.
- **Consecutive-SNP modes (consecutiveSNP=TRUE):** interior SNPs use symmetric left/center/right windows. At the chromosome edges, where symmetric windows cannot fit, the function compares a near window that touches the edge with a far window shifted inward by ~half a window. This preserves an informative contrast at boundaries rather than dropping edge SNPs.

Input & windowing

- **Input (choose one):**
    - vcf: in-memory data.frame with standard VCF columns (CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT) followed by sample genotype columns, or
    - vcf.file.names: character vector of VCF file paths to read and row-bind (ensure sensible chromosome order).
- **Approach:** approach = "Bp.based" or "SNP.based".
- **Window size:** via end and start (length = end − start, or just end if start=0).
- **Sliding:** slide = end/2 by default; ignored when consecutiveSNP=TRUE.

Quality constraints & gap handling

- **minSNP (depends on approach):**
    - **Bp.based:** absolute minimum SNPs per bp window (default 15).
    - **SNP.based:** fraction of the nominal SNP window that must remain after maxGap salvage (default 2/3).
        - Example: end=200, minSNP=0.05 ⇒ at least 10 SNPs required in the kept block.
- **maxGap (SNP.based only):**
    - Maximum allowed bp distance between adjacent SNPs within a SNP window.
    - If exceeded, the window is split at large gaps:
        - in consecutive mode → keep the block containing the anchor SNP,
        - in sliding mode → keep the longest block.
    - HR is computed on the kept block if it passes minSNP.

Haploid / hemizygosity support (optional)

Set HaploidExistance = TRUE if some chromosomes are haploid (e.g., X outside PAR in XY systems). The function will:

1. Ask for the list of haploid chromosome IDs.

2. Derive haplotypes and flag hemizygous individuals when identical alleles exceed HemizygosityProportion (default 0.95).
3. Compute HR on haploids separately before merging with autosomes.

Include PAR as a separate "chromosome" or exclude it to avoid mixing diploid and hemizygous regions. This mirrors HRiD's original motivation on the X chromosome, where male hemizygosity provides exact haplotypes.

Normalization & p-values

If WithNormalisation = TRUE, raw HR values are standardized to Z-scores either globally (all chromosomes) or per chromosome (NormalisationByChr = TRUE). One-tailed p-values are then returned for both directions:

- HRiD_Pvalue = pnorm(−Z) → targets decreases in richness (positive selection).
- HRiP_Pvalue = pnorm(Z) → targets increases in richness (candidate balancing selection).

Log10-transformed p-values (*_LogPvalue) are included for Manhattan-style plotting.

Output columns (most common)

- **Sliding (non-consecutive) modes:** Window, CHROM, start.Index, start.position, end.Index, end.position, na (number of unique haplotypes), nh (effective number of haplotypes), HR_value, plus optional Z_value, HRiD_Pvalue, HRiD_LogPvalue, HRiP_Pvalue, HRiP_LogPvalue.
- **Consecutive-SNP mode:** same as above (no Window column), with SNPname and position for the focal SNP (anchor).

Arguments

Core

- approach — "Bp.based" or "SNP.based".
- consecutiveSNP — FALSE for sliding windows; TRUE for per-SNP evaluation.
- start, end, slide — window length and step. Defaults: start=0, end=1e6 (bp) or 30 (SNPs), slide=end/2.
- minSNP — Bp.based: absolute count (default 15). SNP.based: fraction retained after salvage (default 2/3).

- maxGap — (SNP.based only) split threshold in bp; default 1e6.

Haploids (optional)

- HaploidExistance — TRUE if haploid chromosomes are present.
- HemizygosityProportion — fraction of identical alleles to flag hemizygosity (default 0.95).

Normalization

- WithNormalisation — TRUE to add Z and p-values.
- NormalisationByChr — FALSE to normalize across all chromosomes.

Note: A full argument reference, including defaults and mode-specific behavior, is documented in the inline comments of HaplotypeRichness.Estimation().

Installation / Usage

This repository is function-first (no package scaffolding).

From the repo root:

source("R/HaplotypeRichness.Estimation.R")

R version & dependencies

- Tested on R ≥ 4.0.
- Uses only base R (utils::txtProgressBar, read.table, etc.).

Quick start (with bundled example)

The repo includes:

- example/example.R — a runnable script demonstrating HRiD/HRiP on a toy dataset.
- example/example.RData — contains a small toy_vcf data.frame (phased VCF-like) and precomputed result objects for quick inspection.

Troubleshooting

- **"Unknown genotype separator"** — Input must be phased or consistently encoded. The function accepts both | (phased) and / (unphased), but mixing them will fail. If both appear, harmonize first (e.g., convert all to |).

Note: using phased input (|) is recommended. Unphased (/) input may artificially inflate haplotype counts, since allele order is not canonicalized.

- **Entire sample column is NA** — If all genotypes are NA for a sample, delimiter detection can fail. Remove or impute.
- **CHROM not numeric** — Convert labels (e.g., X→23, Y→24) and strip chr prefixes.
- **Rows not sorted** — SNP rows must be sorted by POS within each chromosome.
- **Window boundary mismatch warnings** — Benign if windows contain no SNPs. Otherwise check for duplicate POS or inconsistent CHROM labels.
- **NA p-values / sd=0** — Too few windows or zero variance in HR_value. Increase window size or relax minSNP.
- **Many windows dropped** — maxGap may be too strict; increase (e.g., 3e5–1e6) or switch to bp-based windows.
- **Edge windows NA** — Very large windows on short chromosomes may not fit. Reduce end or use consecutiveSNP=TRUE.
- **Interactive prompt** — Triggered by HaploidExistance=TRUE. For batch runs, set FALSE.
- **Slow runtime / high memory** — For large VCFs, run per chromosome, increase slide, or save intermediates with saveRDS().
- **SNPname is NA (consecutiveSNP=TRUE)** — Happens if VCF ID is missing (.). Use position or populate IDs.

Citation

Background and original X chromosome application:

Shihabi, M. *et al.* (2022).

Identification of Selection Signals on the X-Chromosome in East Adriatic Sheep: A New Complementary Approach. *Front. Genet*. 13:887582. doi: 10.3389/fgene.2022.887582

This repository generalizes HRiD beyond its original X chromosome application, providing a practical implementation for autosomes in phased datasets to detect signatures of positive selection, while additionally reporting HRiP as the opposite tail that may indicate balancing selection.