The SumLINK statistic for linkage analysis: Application to the ICPCG pooled linkage resource

G. Bryce Christensen, Nicola J. Camp and the ICPCG

University of Utah School of Medicine Department of Biomedical Informatics, Division of Genetic Epidemiology
391 Chipeta Way Suite D, Salt Lake City, Utah 84108, USA

Introduction

We propose a novel, genome-wide, linkage-based statistic, "sumLINK," for identification of disease susceptibility loci. Our approach focuses primarily on "linked" pedigrees (those with pedigree-specific LOD ≥ 0.588; equivalent to unadjusted p ≤ 0.05) to identify regions of extreme consistency across powerful pedigrees. The sumLINK statistic is simply the sum of multipoint LOD scores for linked pedigrees at a given point in the genome.
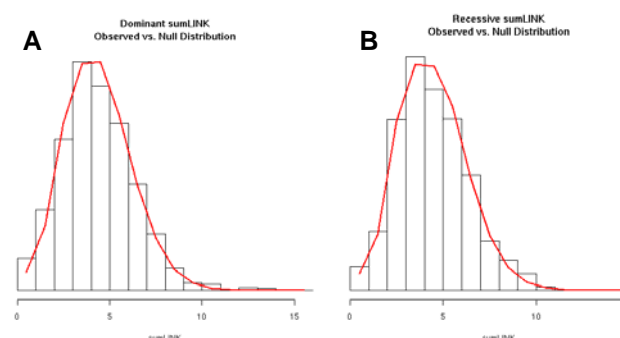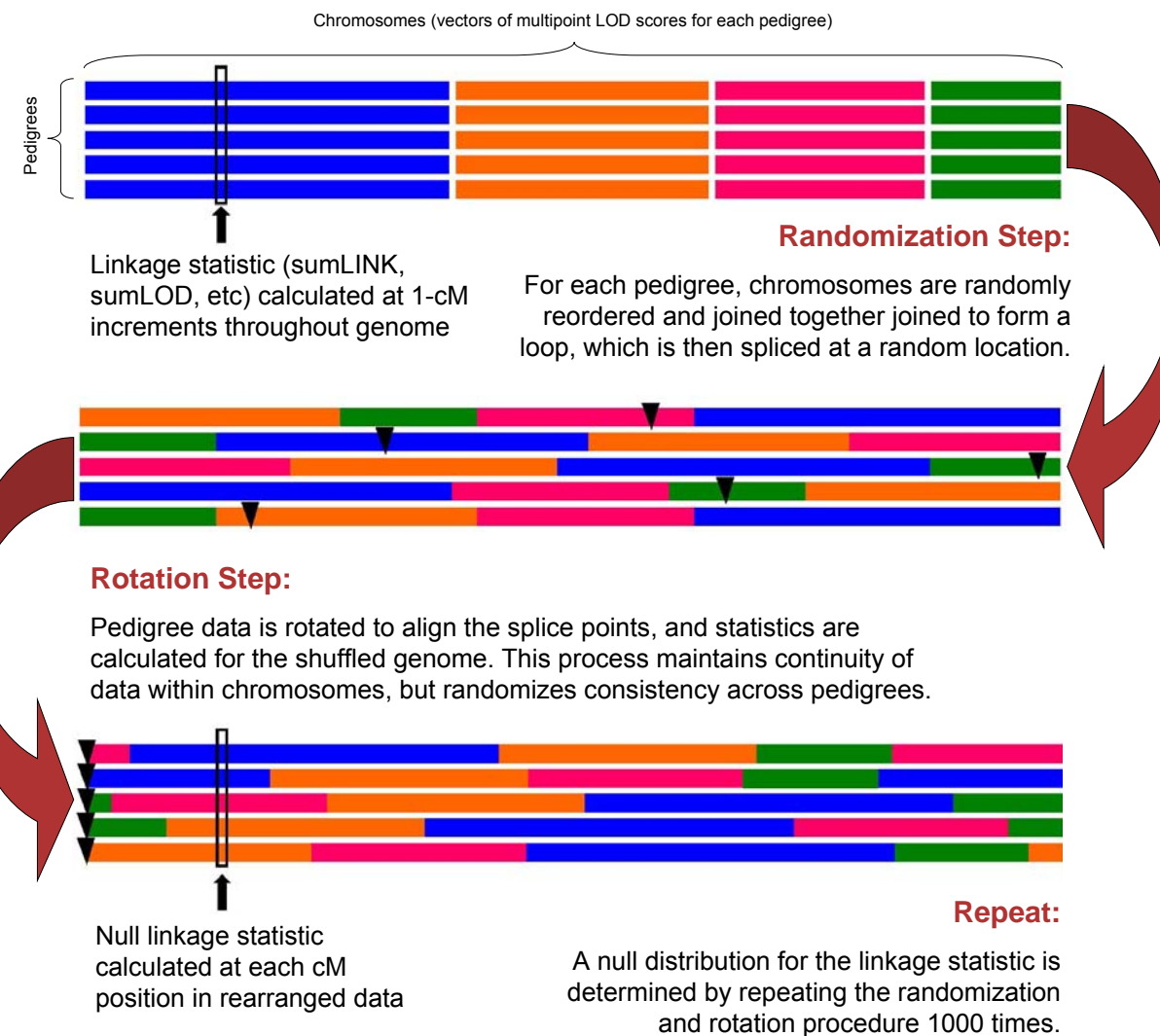
The genetic factors underlying many complex human traits are poorly understood. Linkage findings are often difficult to replicate, and localizing the genes responsible for linkage signals is challenging. We believe that focusing on individually powerful pedigrees may give the greatest opportunity to identify and localize true susceptibility loci and the underlying genes.

Methods

The sumLINK method is ideally suited for large data resources. For this example we utilized 190 pedigrees from the International Consortium for Prostate Cancer Genetics (ICPCG) with clinically aggressive prostate cancer. A conventional linkage study of this resource was published previously. (Schaid et al., 2006)

Dominant and recessive max LOD scores were computed for each pedigree at 1-cM increments throughout the 22 autosomes (3502 cM total length). The sumLINK was then calculated for both inheritance models by adding all of the LOD scores that exceeded 0.588 at each cM position. Results were compared to the better known "sumLOD" statistic, which is the sum of all positive pedigree LOD scores at each point. The significance of the sumLINK and sumLOD metrics were assessed empirically by a unique shuffling method that simulates the expected consistency of linked pedigrees under null conditions. Peaks with a magnitude that occurs with a frequency less than 0.05 per genome were considered genome-wide significant, and peak heights occurring less than once per genome were considered suggestive evidence for linkage.

Derivation of Empirical Null Distribution

Chromosomes (vectors of multipoint LOD scores for each pedigree)

Linkage statistic (sumLINK, sumLOD, etc) calculated at 1-cM increments throughout genome

Randomization Step:

For each pedigree, chromosomes are randomly reordered and joined together joined to form a loop, which is then spliced at a random location.

Rotation Step:

Pedigree data is rotated to align the splice points, and statistics are calculated for the shuffled genome. This process maintains continuity of data within chromosomes, but randomizes consistency across pedigrees.

Null linkage statistic calculated at each cM position in rearranged data

Repeat:

A null distribution for the linkage statistic is determined by repeating the randomization and rotation procedure 1000 times.

Results

- **SumLINK**
  - Significant linkage evidence identified on chromosome 20 in dominant model
  - Suggestive linkage evidence on chromosome 11 in recessive model
  - Suggestive evidence on chr 2 in both models
- **SumLOD**
  - Replication of significant result on chr 20
  - Significant evidence for recessive linkage to chr 11

Table: SumLINK and SumLOD peaks, and the expected frequency of peaks of similar or greater magnitude based on the empirical null distribution.

	Chromosome	Position (cM)	Observed Statistic	Frequency per genome
Dominant SumLINK	20*	59	13.85	0.005
	2	69	10.84	0.625
Recessive SumLINK	11*	90	10.94	0.610
	2	68	10.62	0.885
Dominant SumLOD	20*	59	30.31	0.028
	11*	89	27.65	0.649
Recessive SumLOD	11*	89	27.98	0.007

* Replicates ICPCG finding (Schaid et al., 2006)   █ Significant Result

Discussion

- SumLINK analysis replicated ICPCG findings on chromosomes 20 and 11, but not 6.
- New suggestive locus found on chromosome 2
- Shuffling method performed well, and may be appropriate for testing significance of other statistics with unknown distributions
- Method is ideal for pooled data resources, as it requires no sharing of raw data.
- SumLINK loci have good potential for gene localization, as several linked pedigrees exist beneath each peak identified.

Distribution of SumLINK

A: Dominant sumLINK Observed vs. Null Distribution
B: Recessive sumLINK Observed vs. Null Distribution

The distribution of the sumLINK is difficult to determine theoretically, but can be derived empirically. The histograms in figures A and B show the distribution of the observed sumLINK statistic for the dominant and recessive models, respectively. The red lines show the empirical distribution of those statistics as determined by 1000 repetitions of the shuffling procedure described above.

The University of Utah
Biomedical Informatics