BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–6

Approximately independent linkage disequilibrium blocks in human populations - Supplement

Tomaz Berisa [1,*] and Joseph K. Pickrell [1,2]

[1]New York Genome Center, New York, NY, USA
[2]Department of Biological Sciences, Columbia University, New York, NY

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 COVARIANCE MATRIX C

Covariance matrix C in Step 1 is calculated using the shrinkage estimator in Eq. 2.7 from Wen and Stephens (2010):

$$C_{ij} = \begin{cases} \sum_{ij}^{panel}, i = j \\ \exp(-\frac{a_{ij}}{2m}) \sum_{ij}^{panel}, i \neq j \end{cases},$$

where $\sum^{panel}$ is the empirical covariance matrix from the panel, a_{ij} is an estimate of the population-scaled recombination rate between SNPs i and j, and m is the number of individuals sampled from a population.

We calculate matrix C based on the publicly available 1000 Genomes Project Phase 1 dataset (1000 Genomes Project Consortium *et al.*, 2012). We computed this matrix separately in the European, East Asian, and African meta-populations. Genome-wide recombination rates were obtained from Phase 2 HapMap Release 22 (Frazer *et al.*, 2007) and interpolated to all positions in the 1000 Genomes dataset.

2 CORRELATION MATRIX P

The corresponding correlation matrix P in Step 2 is obtained by calculating the square of Pearson product-moment correlation coefficients for each element in the covariance matrix as follows:

$$r_{ij}^2 = \frac{C_{ij}^2}{C_{ii}C_{jj}},$$

where C_{ij} is the covariance between SNPs i and j from matrix C.

3 CONVERSION OF MATRIX P TO VECTOR V

Assuming P is sparse, approximately banded, and approximately block-diagonal (with sporadically overlapping blocks), representing each antidiagonal of P by the sum of it's elements is a straightforward way of approximating the *intensity* of LD between equidistant SNPs around a given ***SNP***. In other words, it can serve as a metric for how close a SNP is to the center of an block of SNPs in LD.

Implementation of the outlined algorithm can further be simplified with the knowledge that P is symmetric.

4 APPLYING LOW-PASS FILTERS OF INCREASING WIDTHS TO V

Applying a low-pass filter to vector V has the effect of filtering out high-frequency fluctuation in the signal, while lower-frequency components remain in tact. In other words, the vector is *"smoothed"*. This is done in order to capture the large-scale variation in LD and discard small-scale changes. Specifically, we utilize a Hann window (Blackman and Tukey, 1958) for the filtering function in order to avoid high frequency components that may remain with a simple rectangular window. The Hann function is a discrete window function given by:

*to whom correspondence should be addressed

$$w[n] = \sin^2\left(\frac{\pi n}{N-1}\right), 0 \leq n \leq N-1,$$

where N is the filter width.

Applying a low-pass filter w to vector V is equivalent to the convolution of those two functions. The convolution of discrete functions f and g is defined as:

$$(f * g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f[m]\, g[n-m].$$

The convolution of two finite sequences is defined by extending the sequences to finitely supported functions on the whole set of integers. Therefore, applying filter w to vector V is equivalent to:

$$V'[n] = (w * V)[n]$$

In order to achieve the required mean segment size, we search for the minimum filter width (in the distance domain, which results in a higher cutoff in the frequency domain) satisfying this constraint. The minimum filter width is used because it results in the least amount of discarded data.

The process of identifying the minimum filter width can be outlined as follows:

1. Set initial filter width to zero
2. Apply filter to vector V
3. Identify and count minima
4. Exit if mean segment size constraint is satisfied, otherwise increase filter width by step value and go to step 2.

Although this approach is valid, it is tedious because it searches every single filter width until it satisfies the search condition. Assuming that increasing the filter width reduces the number of minima monotonously, we can search this space more efficiently by utilizing one-sided binary search with deferred detection. In other words, we start out by doubling the filter width until we encounter a value that satisfies the search condition (but is not minimal), after which we perform a binary search in the space between that value and zero. Also, the deferred detection algorithm has the advantage that if values in the search space are not unique, it returns the smallest index (the starting index) of the region where elements are equal to the search value. Therefore, in the case that multiple filter widths yield the same number of minima, the deferred detection algorithm will return the smallest filter width.

5 FINE-TUNING SEGMENT BOUNDARIES WITH LOCAL SEARCH

Identifying minima in the filtered (i.e., *"smoothed"*) vector corresponds to identifying large-scale fluctuation in LD. Since the filtered vector represents large scale variation, it is impervious to shorter-range fluctuation. This makes the minima identified in the previous step a good starting point for identifying their final values. The final value for a segment boundary is found as follows:

1. Initialize segment boundaries b_m to minima identified in Step 4.
2. Define search space SSP_m for boundary b_m as $[\frac{b_{m-1}+b_m}{2}, \frac{b_m+b_{m+1}}{2})$.
3. For each boundary, find ***SNP*** $l \in SSP_m$ which minimizes its *outer sum*: $\sum_{i<l}\sum_{j>l} e_{ij}, P = (e_{i,j})$

Fig. 1 illustrates the local search procedure on the simplified example from the manuscript. The full line and transparent orange rectangle represent an initial breakpoint b_m (identified in Step 4 of the main algorithm) and its corresponding *outer sum*. The dashed lines represent *outer sums* for a subset of candidate ***SNPs*** in SSP_m. The goal of this search procedure is to find the ***SNP*** with a minimal *outer sum*. In the provided illustration, the local search procedure would identify the ***SNP*** corresponding to the green dashed line as a final breakpoint.

This process can be computationally optimized by pre-calculating the total outer sum (defined by all initial breakpoints) once, after which the local search procedure simply updates this total sum with the difference between two adjacent *outer sums* (add/subtract one row and subtract/add one column, depending on the direction of the search) as the search progresses.

6 COMPARISON OF BREAKPOINTS FOR CROHN'S DISEASE GWAS AND HEIGHT GWAS

To compare the LD-aware breakpoints to uniform breakpoints we ran fgwas (Pickrell, 2014) on GWAS of Crohn's disease (Jostins *et al.*, 2012) and height (Wood *et al.*, 2014). We used the set of LD-aware breakpoints calculated using European populations. For both sets of

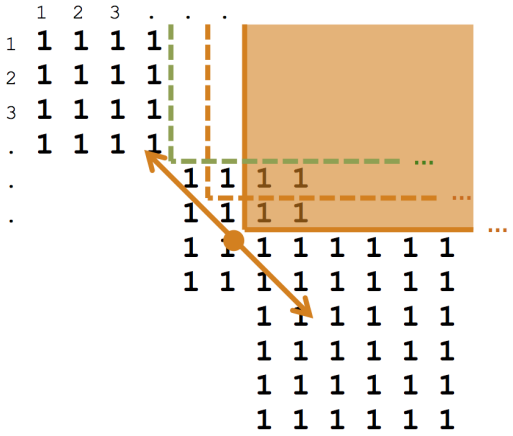

Fig. 1. Illustration of local search.

#	SNP_1	SNP_2	r^2
1	rs80262450	rs12955302	0.203
2	rs2361543	rs2242576	0.914
3	rs2838517	rs3804031	0.255
4	rs2266963	rs5998509	0.804

Table 1. Significant SNPs from Crohn's disease GWAS (Jostins *et al.*, 2012) placed in different blocks (as defined by uniform breakpoints) with $r^2 > 0.1$

#	SNP_1	SNP_2	r^2
1	rs3118903	rs1327646	0.246
2	rs928758	rs7280828	0.352
3	rs62396185	rs806794	0.274
4	rs11774218	rs16905189	0.351

Table 2. Significant SNPs from height GWAS (Wood *et al.*, 2014) placed in different blocks (as defined by uniform breakpoints) with $r^2 > 0.1$

breakpoints, we extracted all regions with a posterior probability of association greater than 0.9, and within each region, we extracted the individual variant with the maximum posterior probability of association. In Tables 1-2, we list the pairs of these variants that fall into separate regions according to the uniform breakpoints but which which have $r^2 > 0.1$. Using the LD-aware breakpoints we find no such pairs of SNPs.

7 LD BLOCK STATISTICS

The number of independent blocks for each ancestry is a consequence of the mean segment size (expressed in number of SNPs and provided as input to the algorithm). For the published blocks, we used 10^4 SNPs for the mean segment size and the resulting number of independent blocks (per chromosome and in total) are shown in Fig. 2. Box plots for block sizes in each of the populations are shown in Fig. 3. The largest blocks correspond to regions around each chromosome's centromere, while 95.9%, 94.6%, and 94.8% of the remaining blocks are within 2 standard deviations of the mean (not including centromere blocks) for African, Asian, and European populations, respectively.

8 RECOMBINATION ACTIVITY AT BREAKPOINTS

Fig. 4 shows mean recombination activity (calculated from genetic maps from the HapMap 2 Project for 1000 Genomes Project variants, available at https://github.com/joepickrell/1000-genomes-genetic-maps) at the presented breakpoints (and uniform breakpoints) across populations. LDetect breakpoints are well above: 1. the genome wide average of 1cM/Mb; and 2. mean recombination rate at uniformly spaced breakpoints. Therefore, we can conclude that the presented breakpoints fall in areas of high recombination.

CHROMOSOME	AFR	ASN	EUR
1	202	114	134
2	221	123	145
3	186	105	123
4	187	105	123
5	171	94	111
6	166	97	113
7	152	85	100
8	148	80	95
9	113	64	75
10	129	73	86
11	129	72	85
12	125	70	83
13	94	54	63
14	86	49	57
15	78	44	51
16	82	46	55
17	72	40	48
18	75	43	49
19	58	34	40
20	59	33	39
21	37	21	25
22	35	21	25
Total	*2605*	*1467*	*1725*

Fig. 2. Number of independent blocks per chromosome and in total for African (AFR), Asian (ASN), and European (EUR) populations.

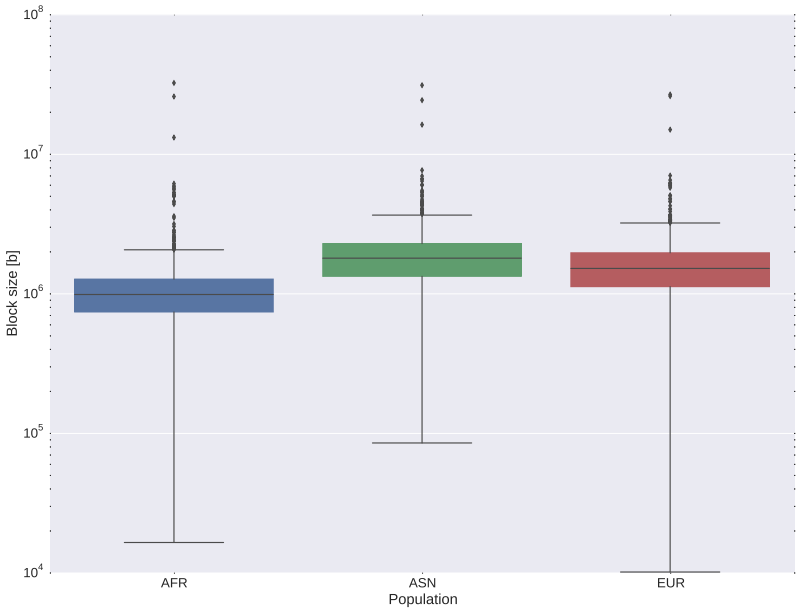

Fig. 3. Box plot of block sizes for African (AFR), Asian (ASN), and European (EUR) populations. Ends of the whiskers represent the datums still within 1.5 interquartile range of the lower and upper quartiles.

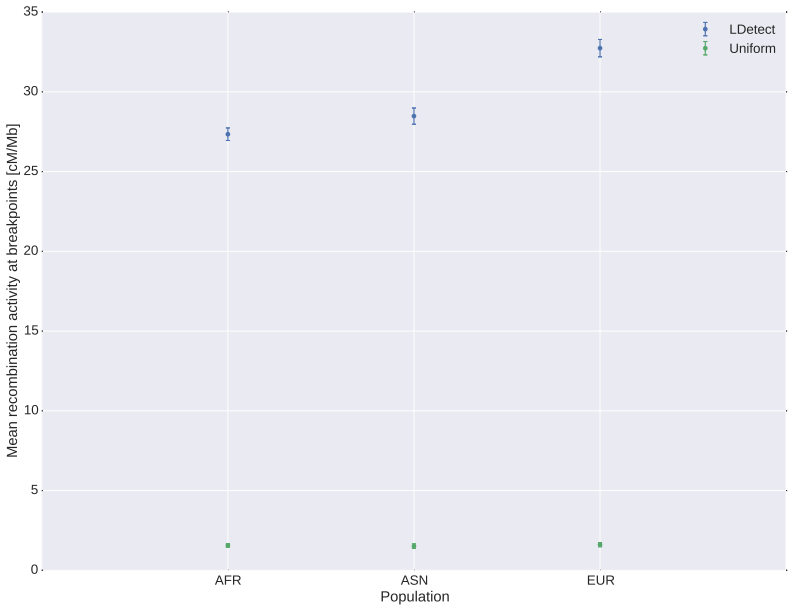

Fig. 4. Mean recombination activity at breakpoints for ldetect and uniform breakpoints (across populations). Error bars are equal to the standard error of the mean.

REFERENCES

1000 Genomes Project Consortium *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering - part i. *Bell System Technical Journal*, **37**(1), 185–282.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., *et al.* (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, **449**(7164), 851–861.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., *et al.* (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**(7422), 119–124.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, **94**(4), 559–573.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, **4**(3), 1158.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., *et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, **46**(11), 1173–1186.