

Q: Why there would be two peaks in the beginning and the end:

The first and the last blocks are 'ALLIGNED', which means the proportion of 0s, 1s, and 2s are the proportions of 0-type blocks, 1-type blocks, and 2-type blocks.

In the original dataset, among all the 565 **first** blocks, the percentage of 2-type SNP-blocks is only 2%;

Over all 565 individual and over all blocks, the percentage of 2-type SNP-blocks take up ~6%;

After shuffling, the percentage of 2-type SNP-blocks in the first row of blocks would be raised to ~6, that's why we could observe a peak in the beginning.

The same scenario would happen for the very last blocks.

Compare the percentages of 0-blocks, 1-blocks, and 2-blocks, between unshuffled dataset and shuffled datasets, we could see the shuffling normalized the percentages for all blocks.

Unshuffled

	0-blocks,	1-blocks,	<b>2-blocks..</b>
The first blocks:	58.94%	39.12%	1.95%
The last blocks:	57.35%	39.65%	3.01%
Over all blocks:	46.73%	47.03%	6.23%

Shuffle #1 done!

	0-blocks,	1-blocks,	<b>2-blocks..</b>
The first blocks:	46.55%	46.37%	7.08%
The last blocks:	46.90%	46.90%	6.19%
Over all blocks:	46.73%	47.03%	6.23%

Shuffle #2 done!

	0-blocks,	1-blocks,	<b>2-blocks..</b>
The first blocks:	52.39%	40.53%	7.08%
The last blocks:	45.84%	47.96%	6.19%
Over all blocks:	46.73%	47.03%	6.23%

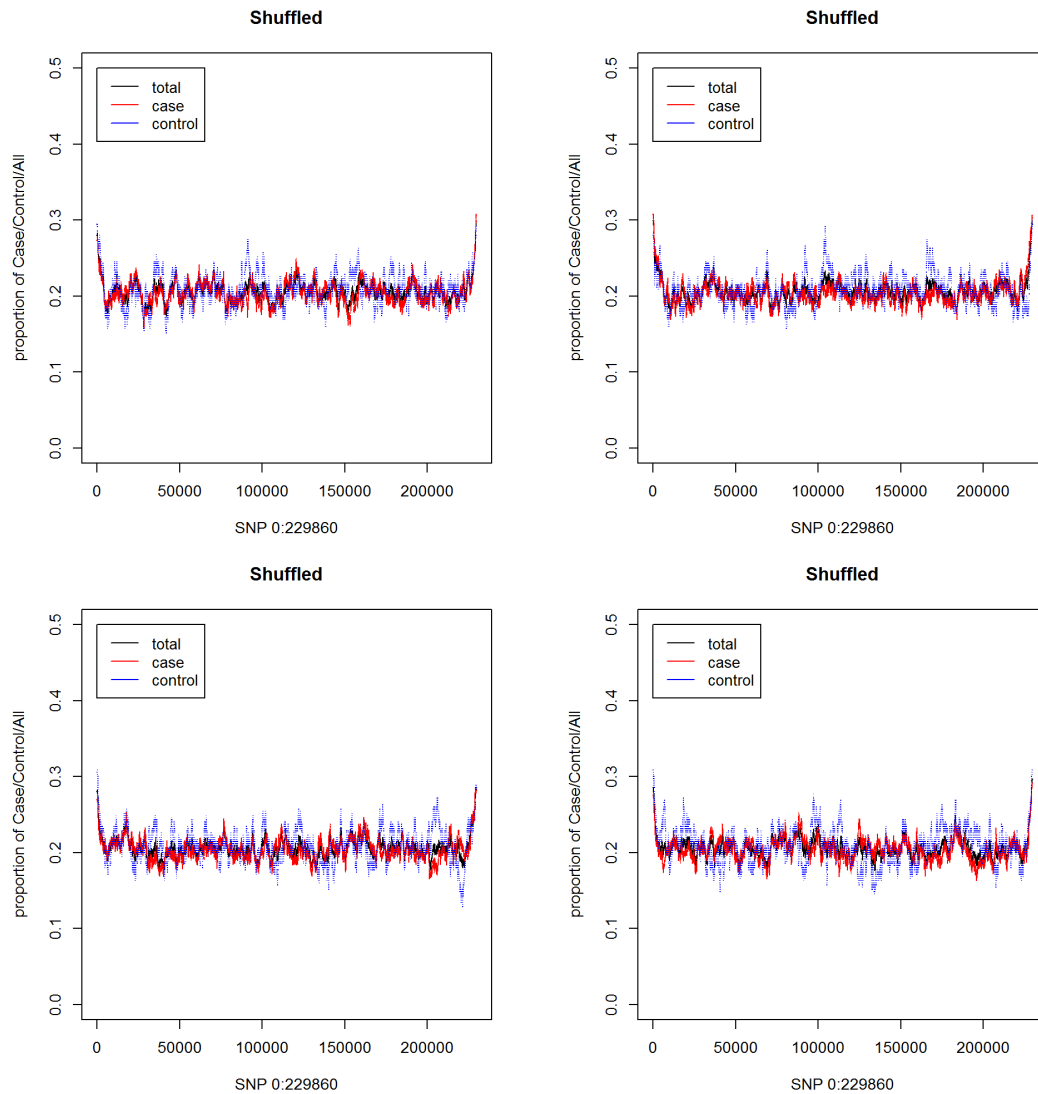


Fig 1 ancestry proportion of normally shuffled dataset

As shown in Fig 1, this is not what we expected.

Solution:

How about we link the last block with the first block for each individual, then made a 'circle' of SNP-blocks.

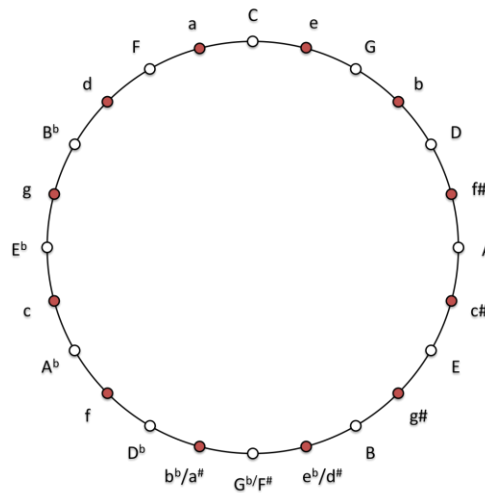


Fig 2 a circle linked blocks for one individual

After we build a circle linked SNP-blocks for a person, we shuffle it with permutation method (3x).

Align all circled SNP-blocks like this: this is a visualization of 6 individuals aligned together.



Fig 3 Align circled SNP-blocks

Then, pick a random SNP position instead of purposely pick a Block start; Open the circle, get a new array of 229860 SNPs.

Align all 565 arrays of SNPs to get a huge matrix. Still, we will get a 229860\*565 matrix.

Calculate the ancestry proportion for the unshuffled and circled shuffled datasets:

The ancestry proportion plot of first 10,000 SNPs are shown in Fig 3.

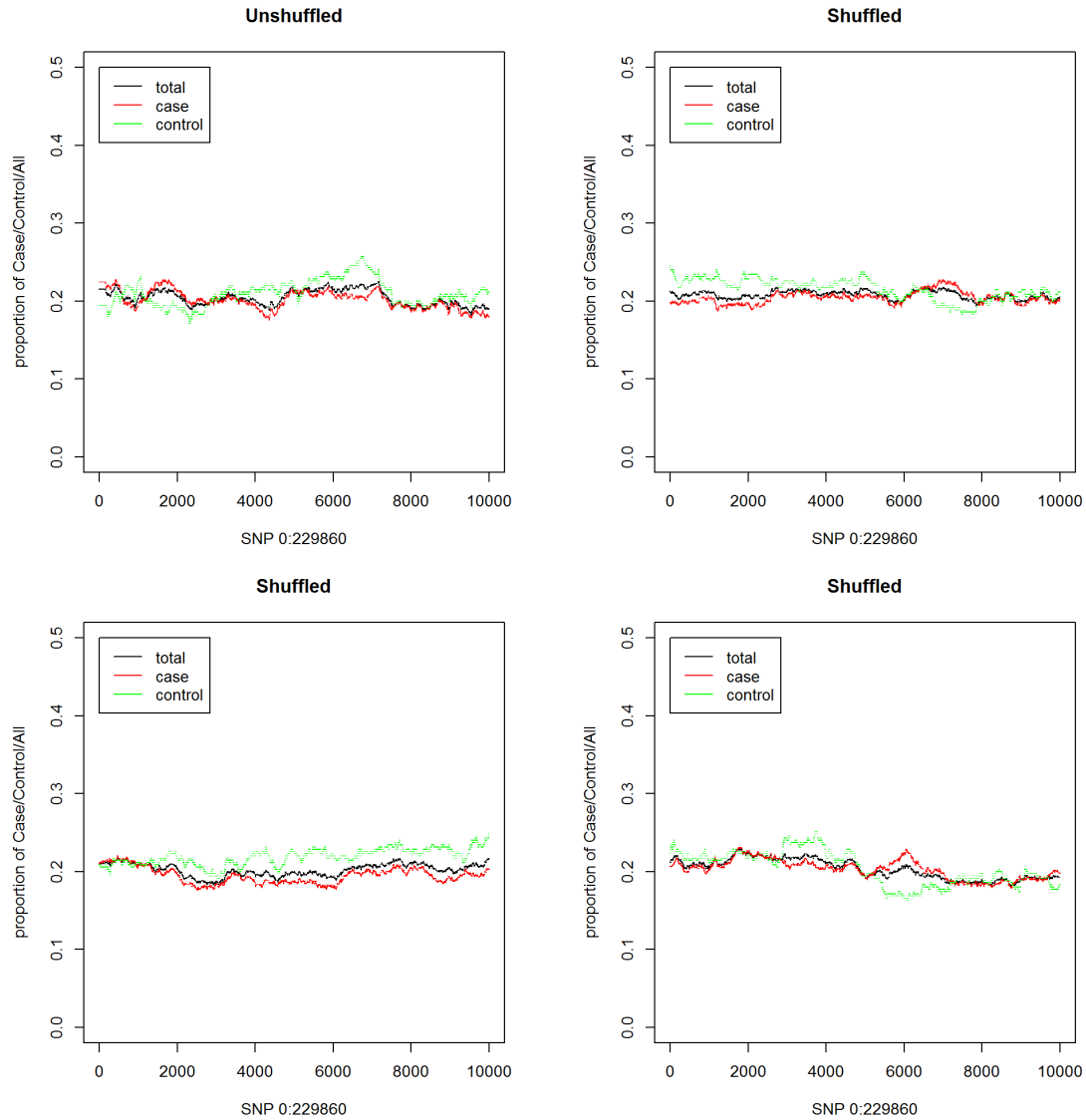


Fig 4 1-10,000 SNPs ancestry proportion plot.

The ancestry proportion plot of all 229,860 SNPs are shown in Fig 5.

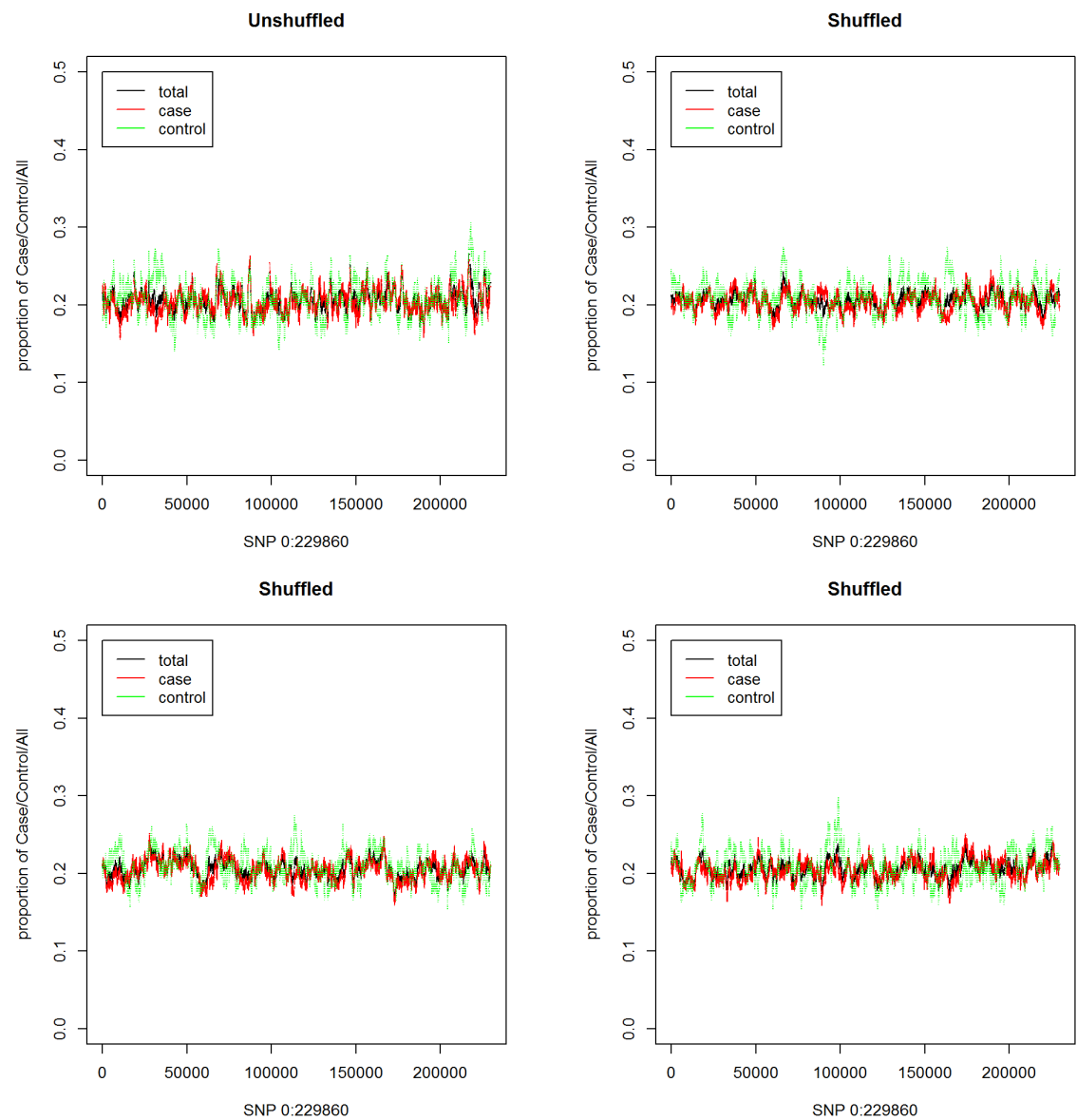


Fig 5 1-229860 SNPs ancestry proportion plot.

After circle-shuffling.

Pass the matrix to mixscore for ADM analysis, get the result.

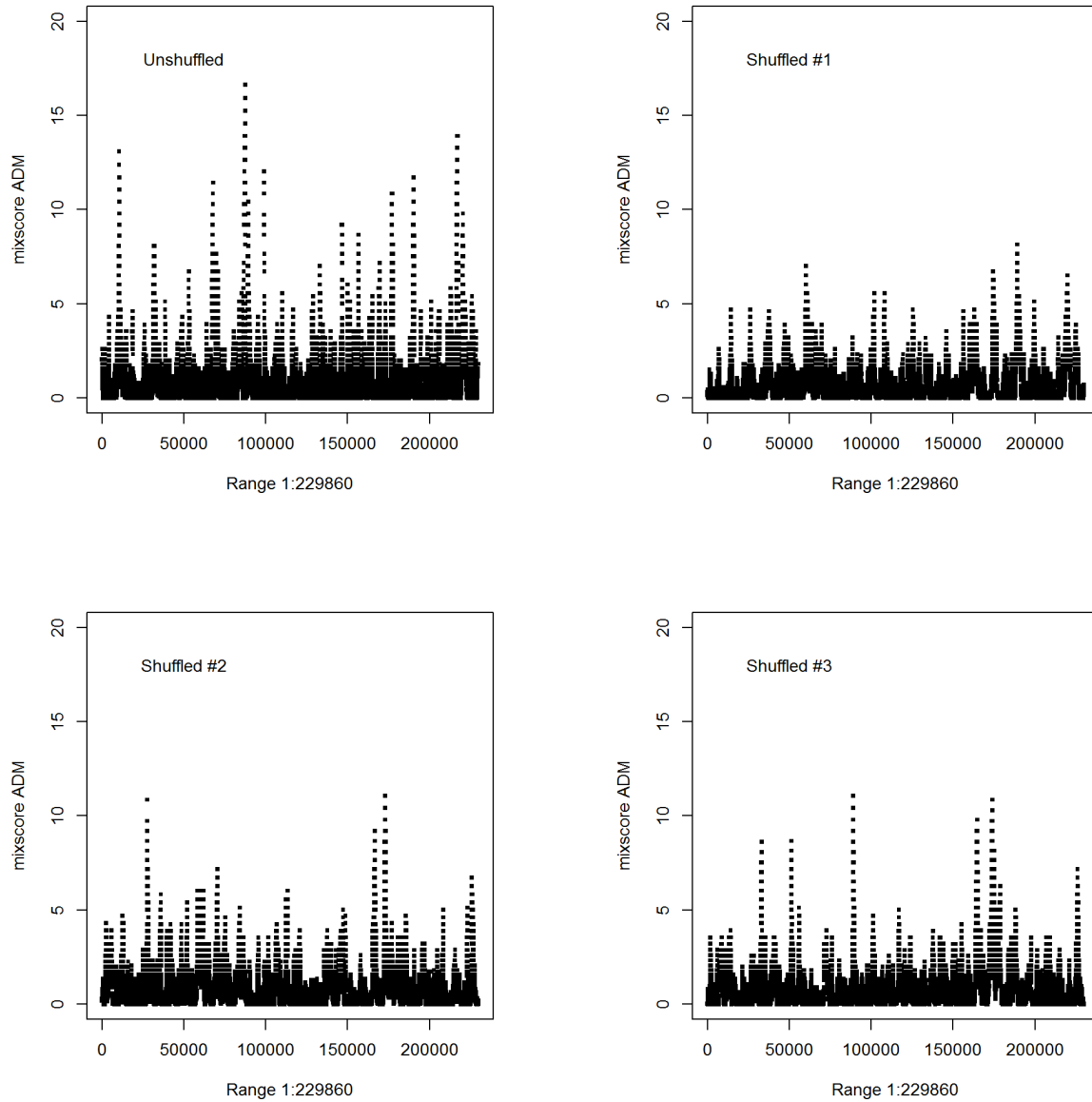


Fig 6 mixscore ADM results of circle-shuffled datasets.

From the results, we could see the mixscore results from circle-shuffled datasets have smaller variance.