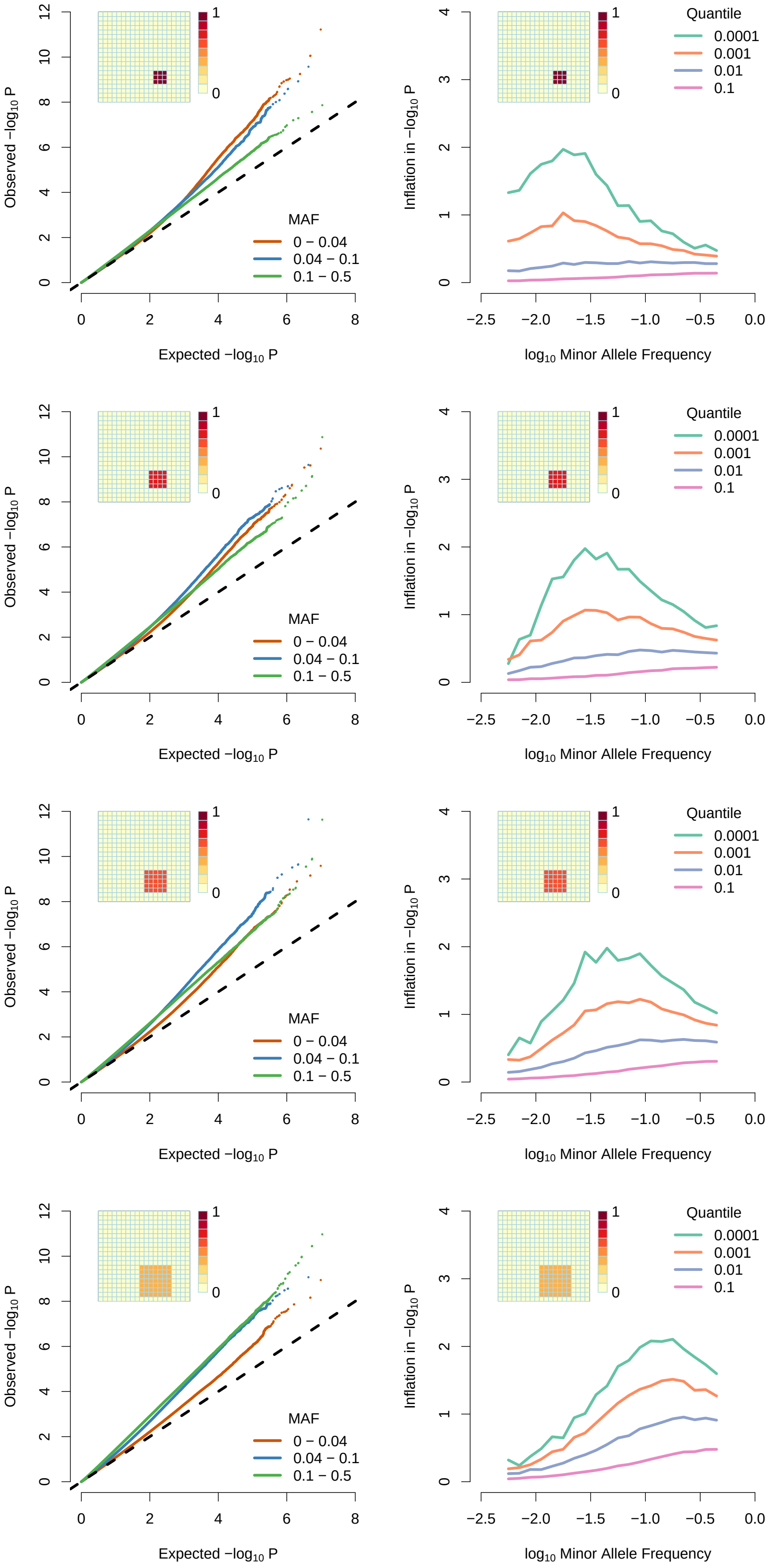
Differential confounding of rare and common variants in spatially structured populations

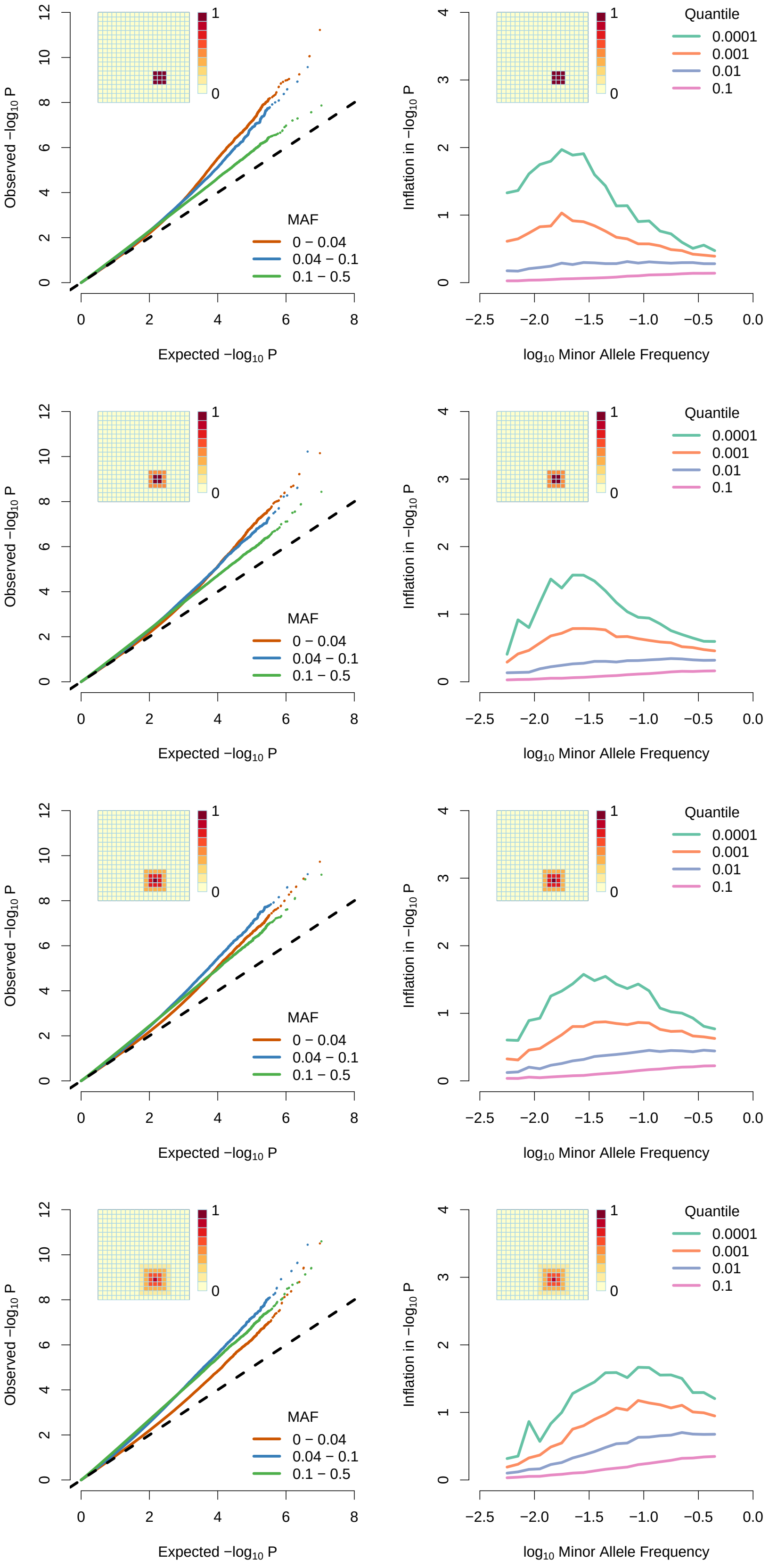
Supplementary information

Iain Mathieson[1*] & Gil McVean[1,2]

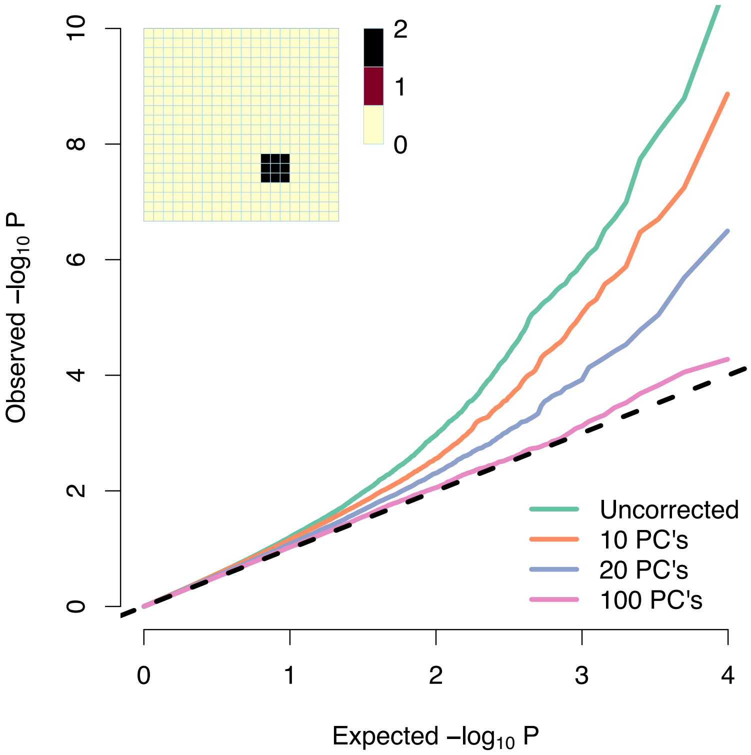
[1] Wellcome Trust Centre for Human Genetics, University of Oxford
[2] Department of Statistics, University of Oxford
[*] Corresponding author; mathii@well.ox.ac.uk



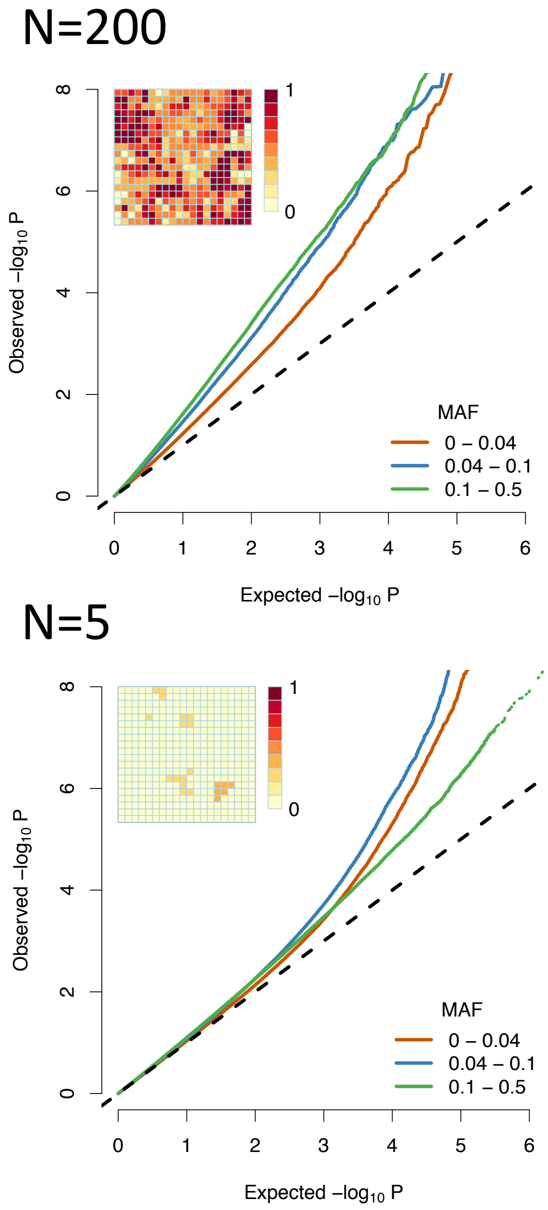
SUPPLEMENTARY FIGURE 1: Effect of the size of the risk area. As the risk area become larger, the peak of maximum inflation moves to higher allele frequencies, and the inflation is spread over a wider range of allele frequencies. The description of these plots is the same as that for Figure 1 in the main text.



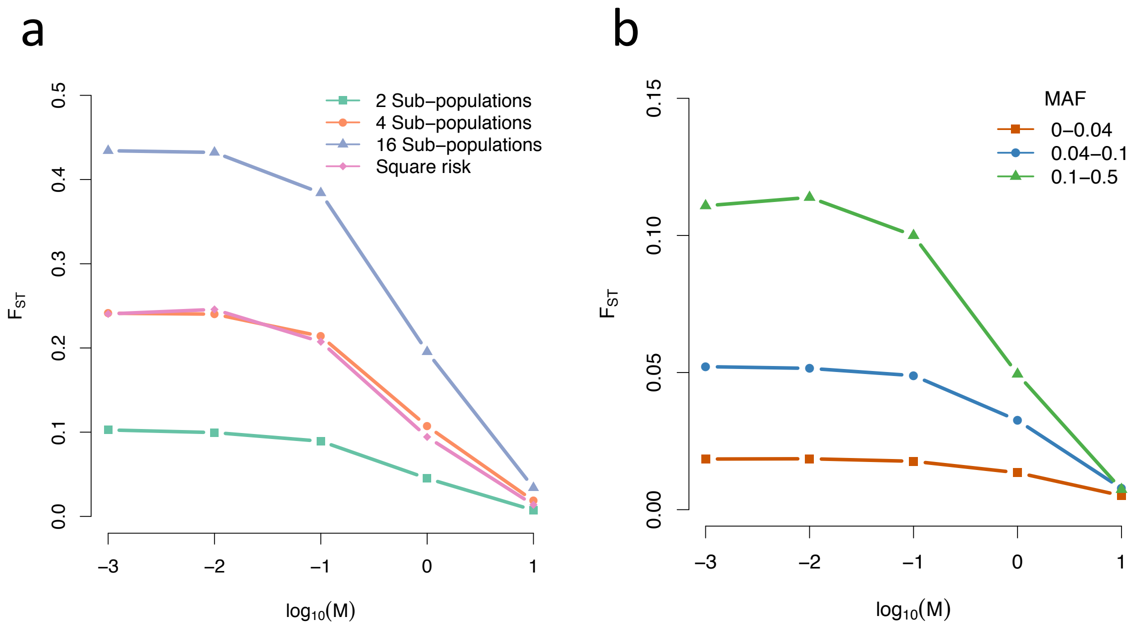
SUPPLEMENTARY FIGURE 2: Effect of the smoothness of the risk area. As the risk area become smoother, the peak of maximum inflation moves to higher allele frequencies, and the inflation is spread over a wider range of allele frequencies. This looks similar to the effect shown in Supplementary Figure 1.



SUPPLEMENTARY FIGURE 3: Ccorrecting using more principal components. Different lines show corrections with different numbers of principal components. This plot is the same design and uses the same parameters as Figure 3b in the main text, but is averaged over only 10 experiments, rather than 100.



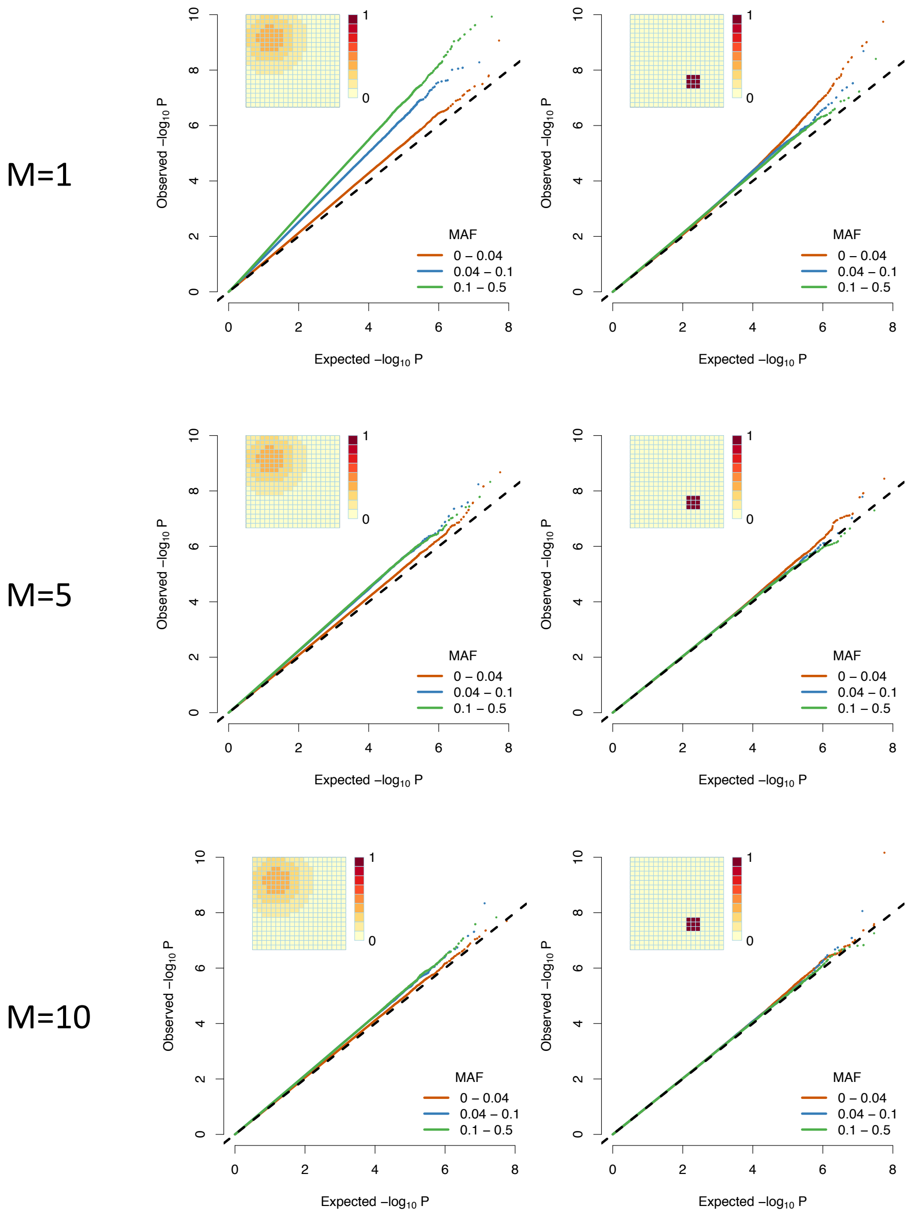
Supplementary Figure 4: Effect of background genetic risk from rare variants. Each of these plots shows the qq plot for a model where the phenotype is influenced by N rare variants at independent loci, not tested for association. The effect of each rare variant is drawn from a normal distribution with mean 0 and standard deviation ranging from 0.2 in the N=200 case to 1 in the N=1 case. The grids in the top left corners of each plot show examples of the resulting spatial pattern of (absolute) phenotypic mean, which was resimulated for each genealogy. Other parameters are the same as in Figure 1 in the main text.



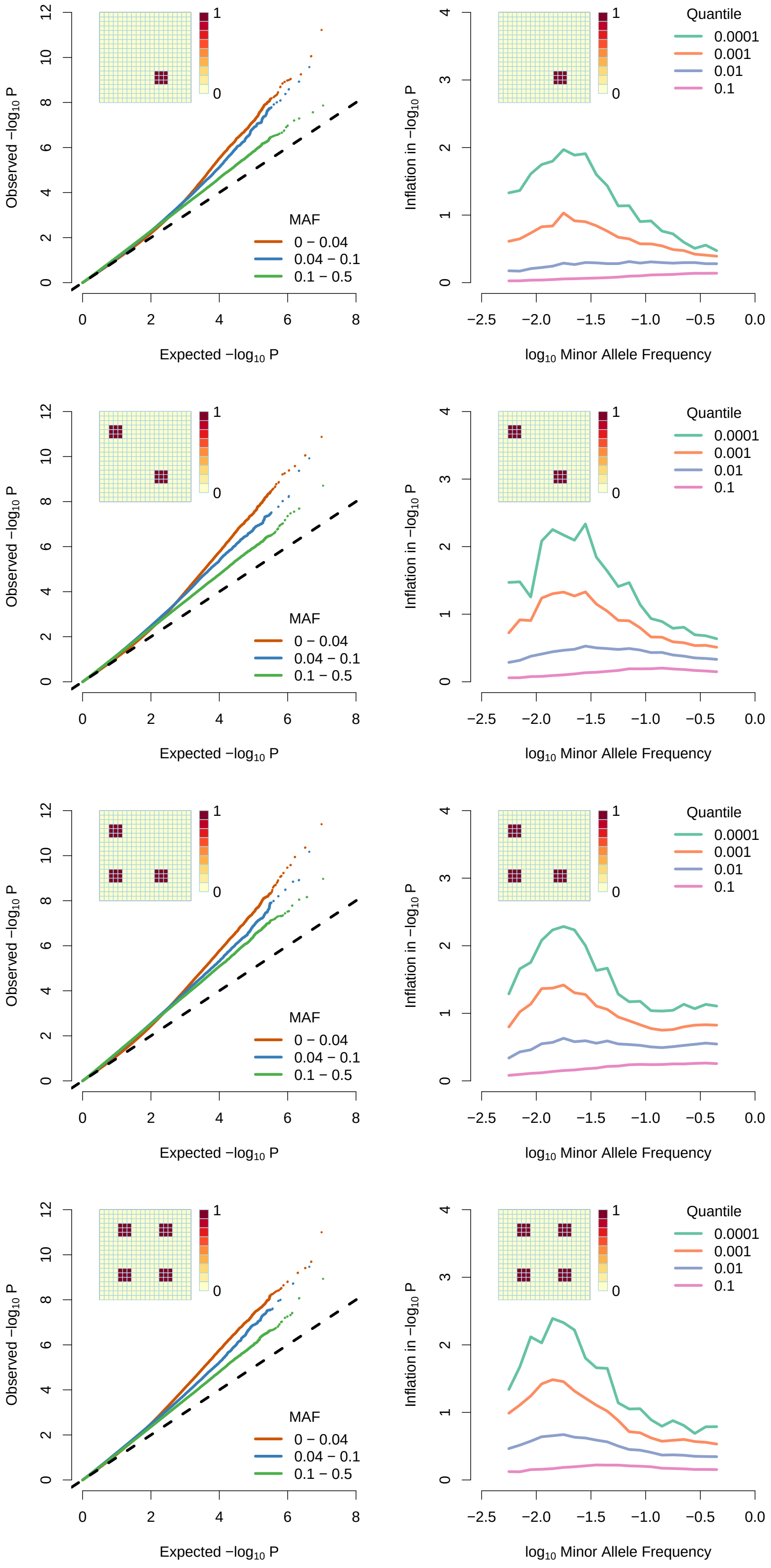
SUPPLEMENTARY FIGURE 5: F_{ST}, calculated (a) for different subdivisions of the grid at different migration rates and (b) for the case of 2 sub-populations (green squares in a) but separating rare, low frequency and common variants.

For (a), we split the grid into 2, 4, and 16 equally sized blocks, and also consider two sub-populations defined by being either inside our outside the small, sharp risk function used in the results (pink diamonds). F_{ST} increases with the number of sub-populations. If we look at the case of two sub-populations (green squares), we see that F_{ST} in our model with $M = 0.01$ is around 0.1 which would be high for human populations.

In (b) we consider only two sub-populations but compute F_{ST} separately for rare, low frequency and common variants. F_{ST} varies dramatically with the frequency of the variants used to calculate it and the overall value is much closer to the value for common variants. This is one reason why it is not an appropriate measure of structure when considering rare variants. Further, F_{ST} measures average differentiation, but as long as there any variants that are highly differentiated, the effects we have described could still persist.



SUPPLEMENTARY FIGURE 6: Effect of higher migration rates. Although overall stratification is much reduced as the migration rate increases, the qualitative effect persists. Other than the migration rate, all parameters are the same as in Figure 1 in the main text..



SUPPLEMENTARY FIGURE 7: Effect of multiple risk areas. As the number of areas of the same size increases, the peak of maximum inflation does not move, and the inflation becomes more pronounced. Other than the risk distribution, all parameters are the same as in Figure 1 in the main text.