Replicating genotype-phenotype associations

Chanock et al., Nature 2007

Fine Mapping versus Replication in Whole-Genome Association Studies

Clarke et al., AJHG 2007

Mari Nelis

Bioinfo J.Club Nov. 2007

- <u>Candidate-gene based studies</u>
  - – questionable genotype-phenotype associations
  - – independent studies have failed to replicate the findings

- <u>Genome-wide association studies</u>
  - – separate true associations from false positives
  - – evaluate initial positive findings

Replicated studies

- Diabetes - peroxisome proliferator-activated receptor-γ (*PPARG*) and transcription factor *TCF7L2*
- Crohn's disease – nucleotide-binding oligomerization domain containing 2 (*NOD2*)
- Age-related macular degeneration – complement factor H (*CFH*)
- Prostate cancer risk – chromosome region 8q24

Instances of non-replication

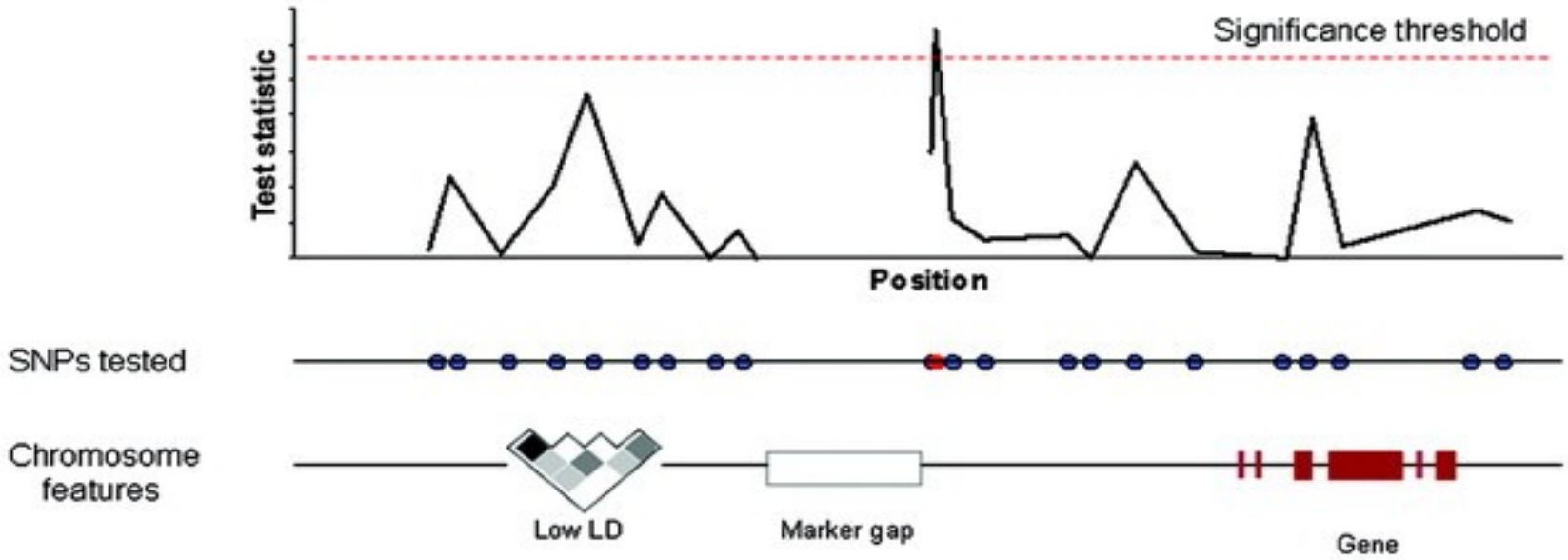
- Small sample size
- Poor study design – lack of comparability between cases and controls
- Follow-up studies analyze different variants

- Extremely small p-values in GWA studies should be interpreted carefully

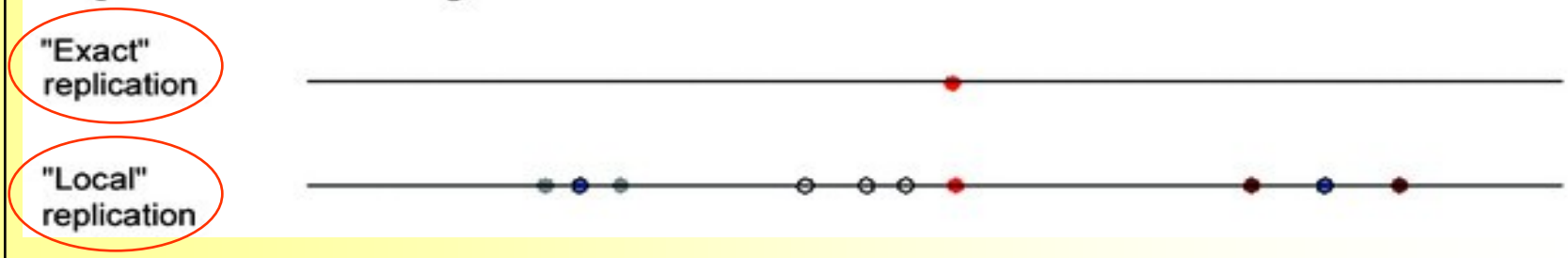
- Genotyping of ancestry informative markers

- Publish the results of initial study, give correct description of sample collection, genotyping, statistical analysis
- Publish negative findings

Primary Study



Replication Study



Disease prevalence – 0.05

Genotype relative risk – 1.3

Freq of high-risk allele – 0.25

First stage

500 000 markers -> 3000 cases, 3000 controls

Replication study

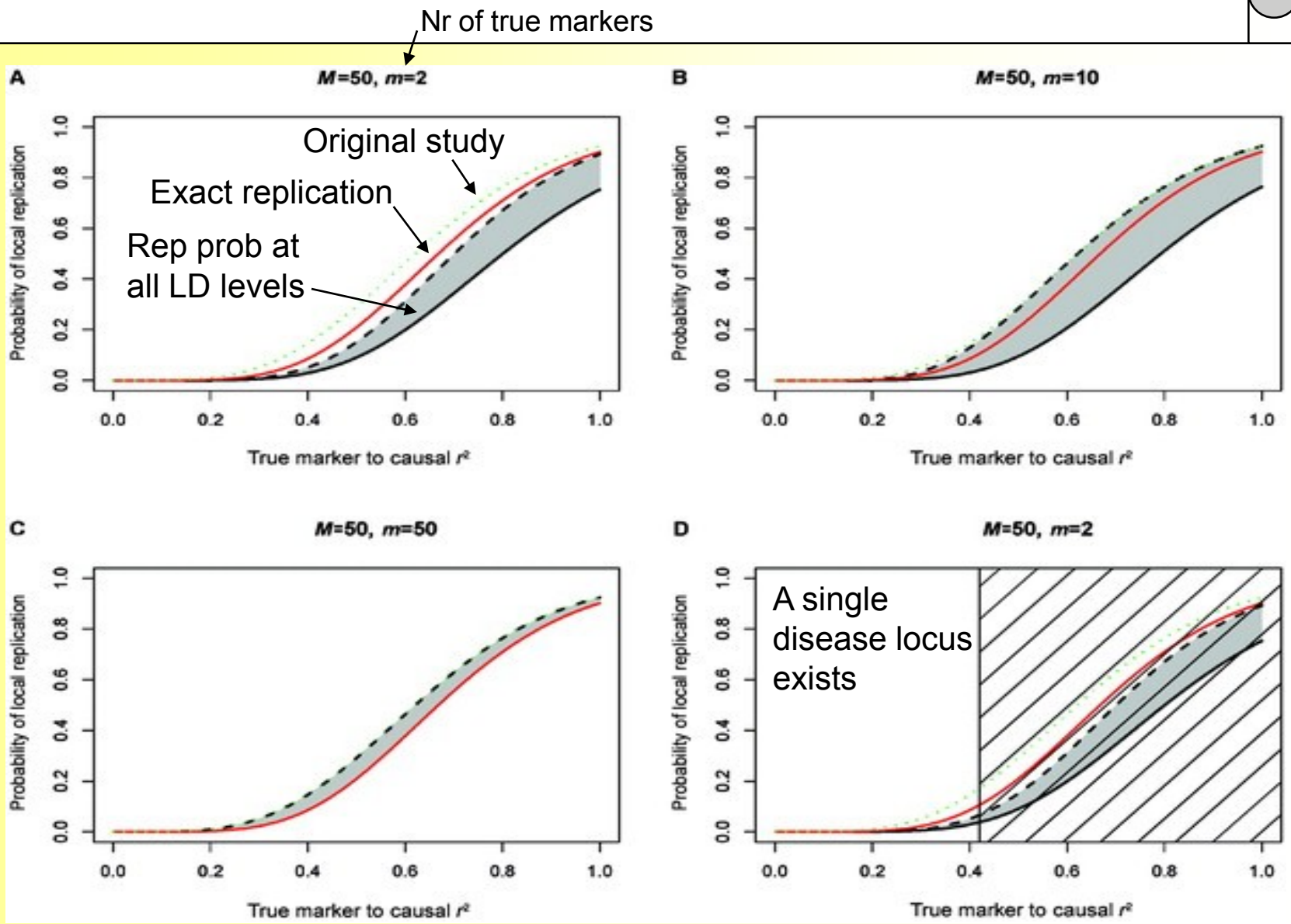
K=10 regions

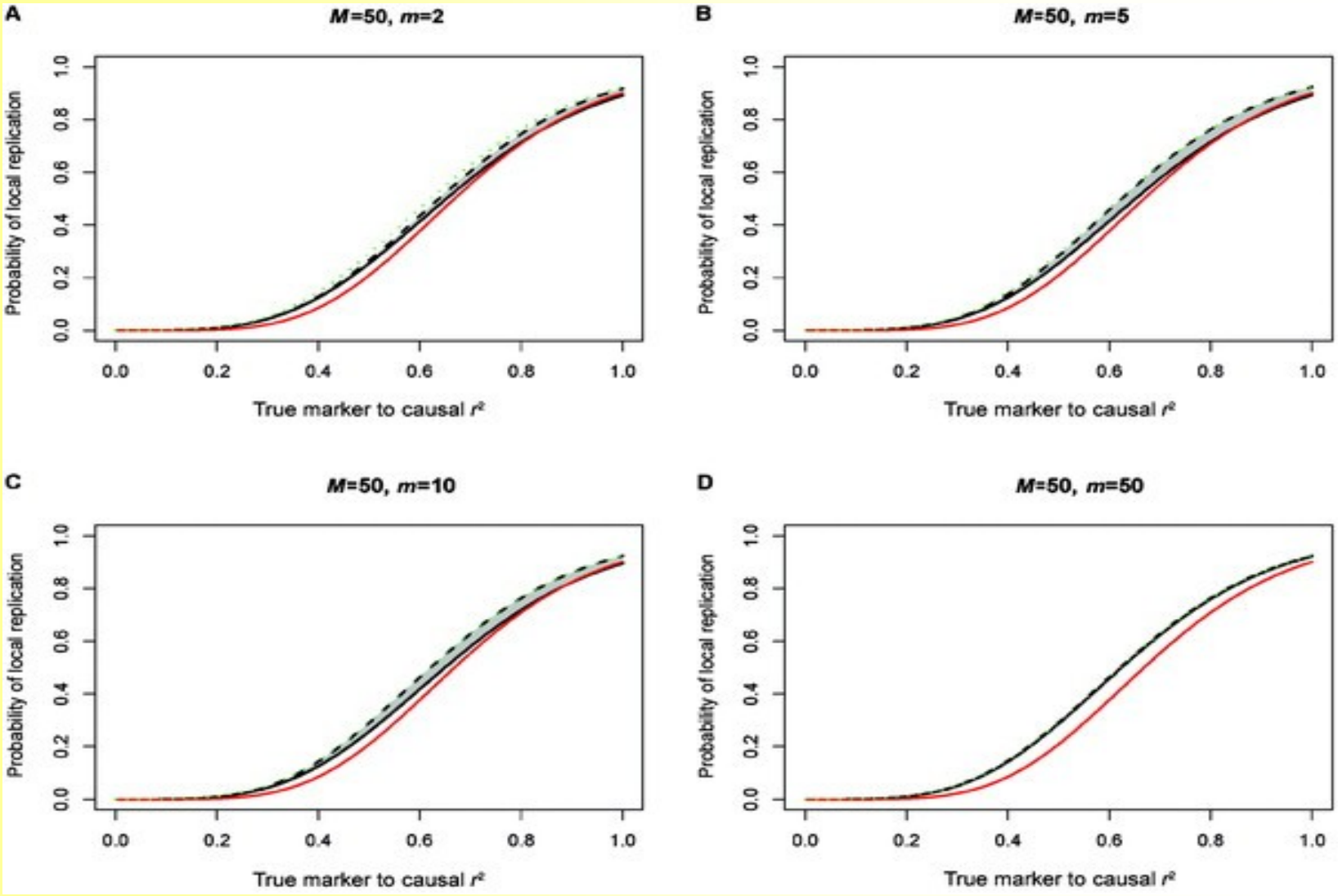
1500 cases, 1500 controls

Type I error rate – 0.05

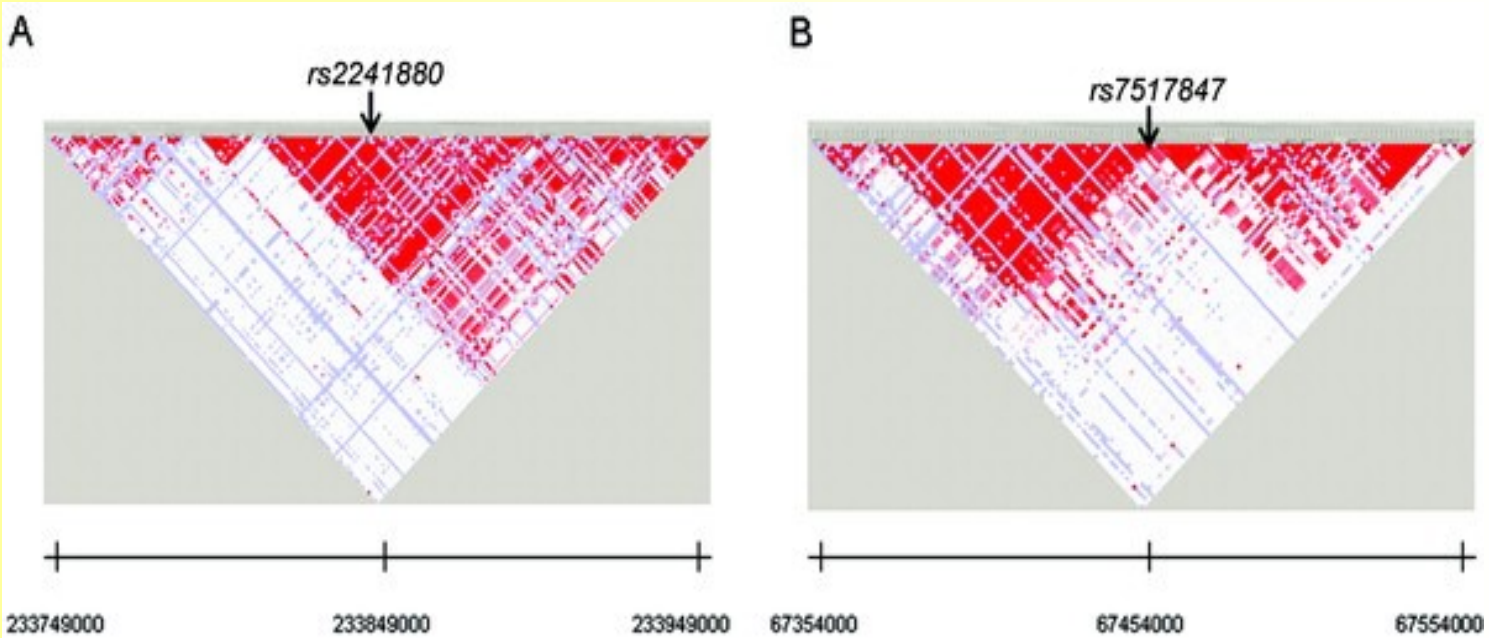
Bonferroni corrected rate $\alpha' = 0.05/[10 \times (50-m+1)]$ (true markers are dependent)

$\alpha' = 0.05/(10 \times 50)$ (true markers are independent)


Nr of true markers
A
M=50, m=2
Original study
Exact replication
Rep prob at
all LD levels
Probability of local replication
True marker to causal r²
B
M=50, m=10
C
M=50, m=50
D
M=50, m=2
A single
disease locus
exists

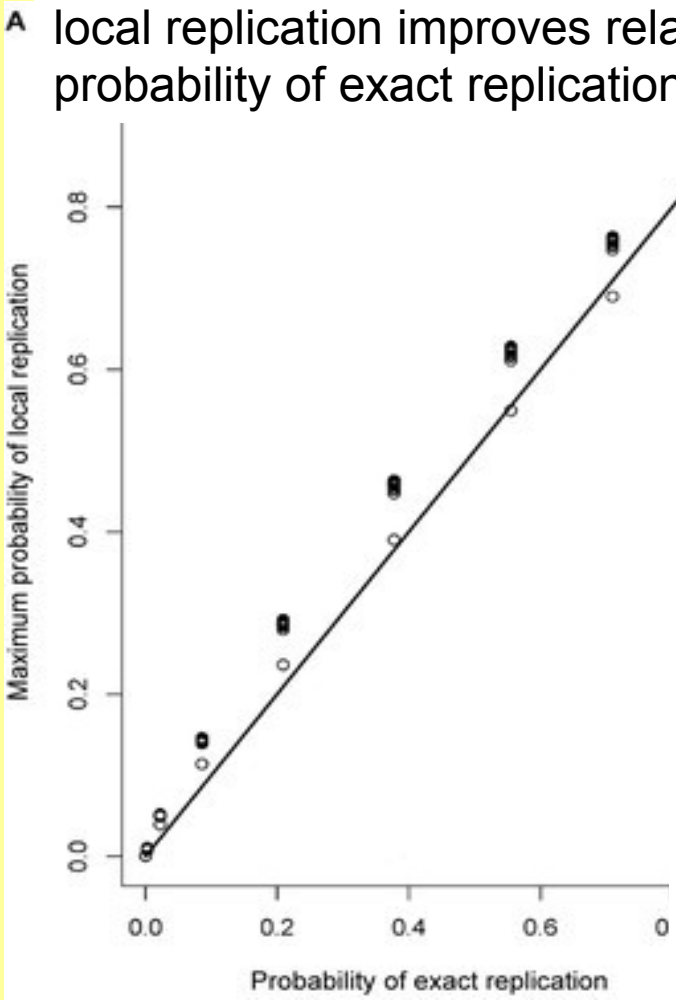


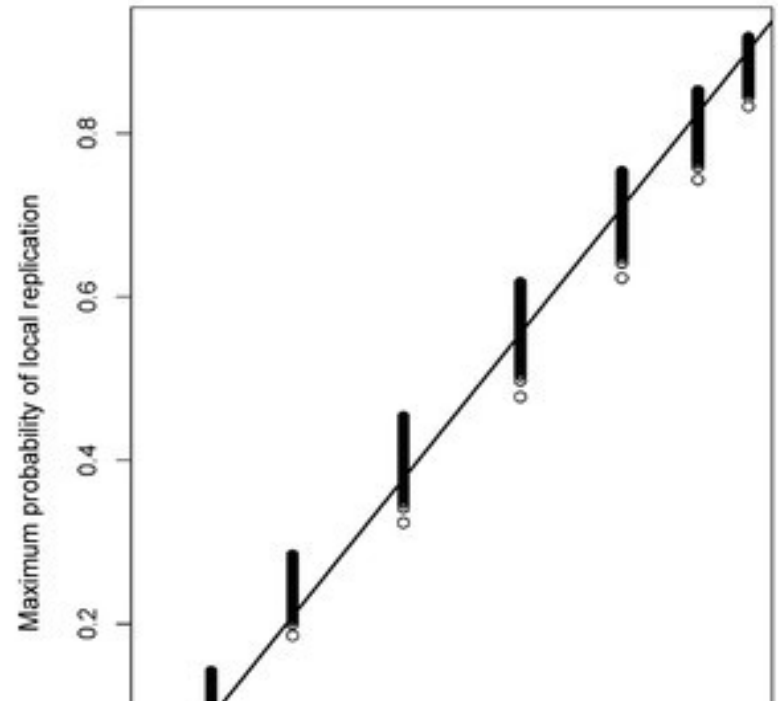
One of the selected markers is in perfect LD with causal variant



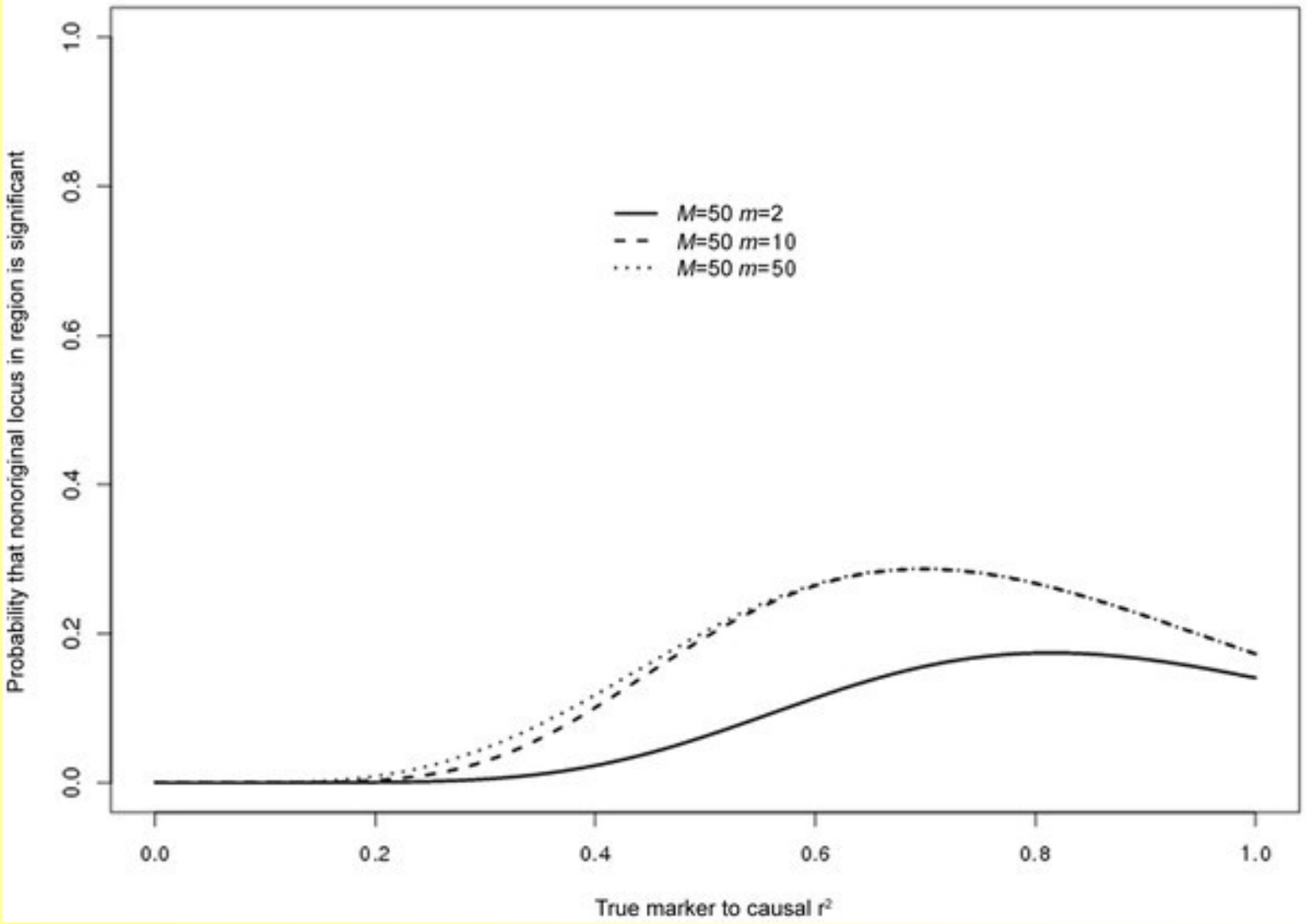
Crohn disease, 100 kb regions

high intermarker LD –
the probability of exact replication decreases, the maximum probability of local replication improves relative to the probability of exact replication





low intermarker LD - the exact strategy is generally optimal when the initial finding is strongly associated with disease and that the local strategy is increasingly dependent on additional marker selection as the strength of the initial finding decreases



Maximum probability that a locus exceeds the significance threshold in a replication study but is different from the locus initially identified

Conclusions

- The effectiveness of the **local strategy** increases with the number and strength of true markers among the additional markers included in the replicate study

- When the original marker is strongly associated with disease (either because there is a large effect or because it is highly correlated with the causal variant) then an **exact strategy** is the best approach