Supplementary Materials

Parameter setups

PolyOrigin

For a simulated dataset, the Julia command line used for PolyOrigin (v0.5.6) is given by

```
polyOrigin(genofile, pedfile)
```

where `genofile` specifies input marker data, including genetic map and genotypic data of parents and offspring, and `pedfile` specifies the population mating design. The default settings *epsilon=0.01* and *seqerr = 0.001* are used, specifying the initial value for the internal estimation of dosage error probability and the sequencing error probability in the case of read count data. By default, the input marker map is genetic map (*isphysmap=false*) and it will not be refined (*refinemap=false*), parental phasing assumes only bivalent formations (*chrpairing_phase=22*), and both bivalent and quadrivalent formations are considered for ancestral inference and parental error correction (*chrpairing=44*).

For the real potato dataset with physical map, the Julia command line is given by

```
polyOrigin(genofile,pedfile,
    isphysmap=true, recomrate=1.25,
    refinemap=true, refineorder=false)
```

where the keyword argument *isphysmap* specifies that input map is physical map with marker positions in unit of base pair, and $recomrate$ specified the global constant recombination rate in unit of cM/Mbp. *refinemap=true* indicates the performance of map refinement, and *refineorder=false* indicates the refinement of inter-marker distances but not marker ordering.

TetraOrigin

The Mathematica command line used for TetraOrigin is given by

```
inferTetraOrigin[genofile, epsO, epsF, ploidy, outstem,
    maxStuck -> 5, maxIteration -> 30, maxPhasingRun -> 10,
    bivalentPhasing -> True, bivalentDecoding -> False]
```

where `genofile` specifies the input genotypic data. *epsF* and *epsO* specify the dosage error probability in parents and offspring, respectively. *ploidy*=4 for tetraploids, and *outsem* specifies the string ID of output file. The options maxStuck, maxIteration, and maxPhasingRun for the parental phasing algorithm are re-set to be consistent with PolyOrigin. And the default settings for bivalentPhasing and bivalentDecoding are consistent with PolyOrigin.

For the simulated F1 datasets, we set *epsO* to the true value 0.01. Although the true parental error probability is also 0.01, we set *epsF*=0 because a non-zero setting would result in much longer computational time.

MAPpoly

The parameter setup of MAPpoly was provided by the first author of MAPpoly (M Mollinari, personal communication); see the scripts at `https://github.com/chaozhi/PolyOrigin_Evaluate`. The R command lines used for MAPpoly (v0.2.2.6) are divided into the following steps

```
#step1: read data
dat.dose.csv <- read_geno_csv(file.in  = genofile, ploidy = 4)

#step2: marker filtering
pval.bonf <- 0.05/dat.dose.csv$n.mrk
dat.chi.filt <- filter_segregation(dat.dose.csv,
    chisq.pval.thres =  pval.bonf, inter = FALSE)
dat.seq <- make_seq_mappoly(dat.chi.filt)

#step3: two-point analysis
all.rf.pairwise <- est_pairwise_rf(input.seq = dat.seq)

#step4: parental phasing and marker spacing for a given marker ordering
(1) For the simulated datasets:
map <- est_rf_hmm_sequential(input.seq = dat.seq,
    start.set = 4,
    thres.twopt = 5,
    thres.hmm = 5,
    extend.tail = mark.tail,
    twopt = all.rf.pairwise,
    sub.map.size.diff.limit = 10,
    phase.number.limit = n.ph)
(2) For the potato dataset:
tpt <-  est_pairwise_rf(input.seq = dat.seq0)
dat.seq  <- rf_snp_filter(tpt)
map <- est_rf_hmm_sequential(input.seq = dat.seq,
    start.set = 3,
    thres.twopt = 10,
    thres.hmm = 10,
```

```
    extend.tail = 50,
    twopt = tpt,
    verbose = TRUE,
    tol = 10e-2,
    tol.final = 10e-3,
    phase.number.limit = 40,
    sub.map.size.diff.limit = 5,
    info.tail = TRUE,
    reestimate.single.ph.configuration = TRUE)
map<-filter_map_at_hmm_thres(map, thres.hmm = 0.001)

#step5:re-estimate with a global error epsilon
map.error <- est_full_hmm_with_global_error(input.map = map,
    error = epsilon))

#step6: calculate genotype probability
genoprob <- calc_genoprob_error(input.map = map.error,
    error = epsilon)
```

We skip the step of marker grouping and marker ordering by using the simulated genetic map or the real physical map. In step 4 for the simulated datasets with population size $N =$ 200, 100, 50, 30, 20, 15, 10, *extend.tail* = 60, 60, 60, 100, 100, 150, 200, and *phase.number.limit* = 20, 20, 40, 60, 60, 100, 300, repsectively; *extend.tail* = 50 and *phase.number.limit* = 40 for the real potato dataset with $N = 435$. In steps 5 and 6, the dosage error probability *epsilon* is set to the true value for simulating data, and 0.02 for the real potato data, based on the estimation of PolyOrigin.

Supplementary figures

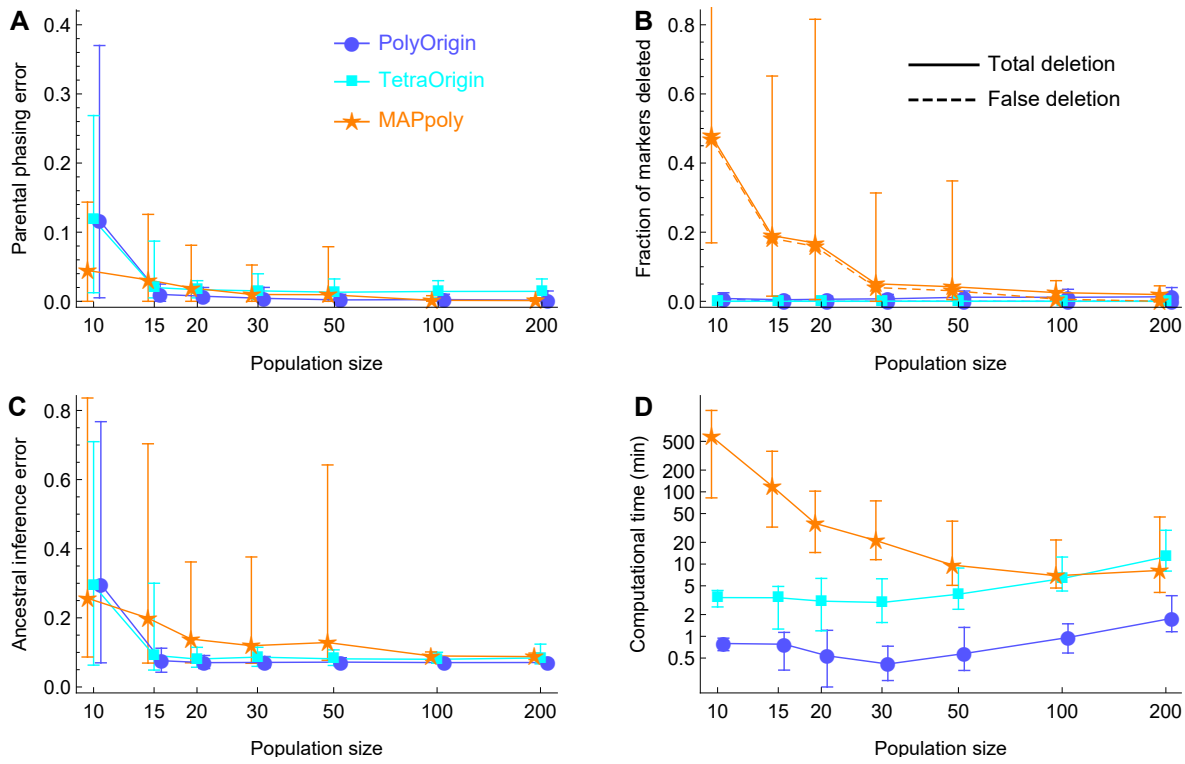

Figure S2: Comparison of PolyOrigin, TetraOrigin, and MAPpoly for the simulated F1 populations without double reduction. The dashed lines in (**B**) denote the fraction of markers that are deleted and have no parental dosage errors.

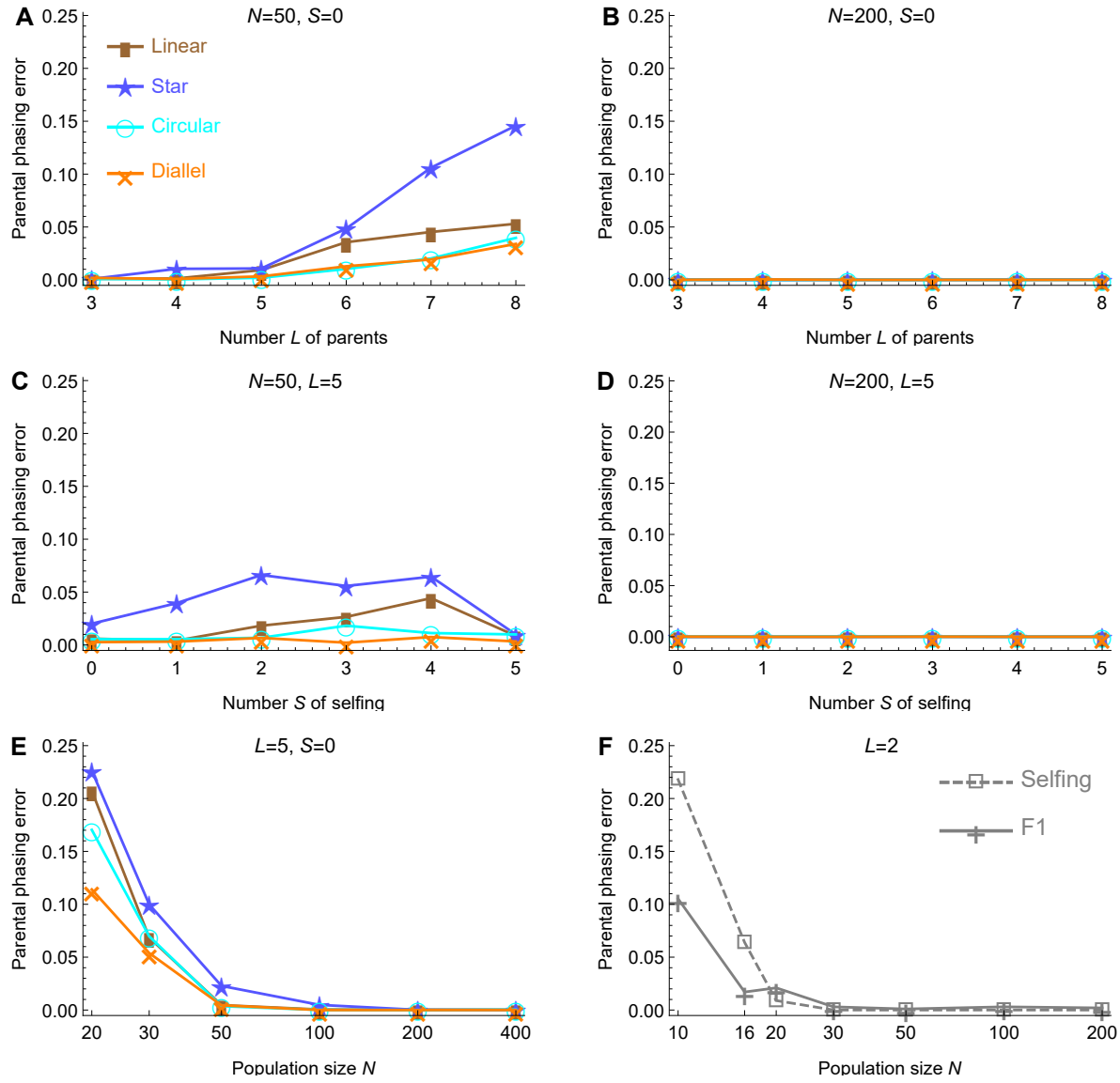

Figure S3: Effect of population design on parental phasing. (**A**&**B**) Effect of the number L of parents for populations with no selfings ($S = 0$) and sizes of $N = 50$ and 200, respectively. (**C**&**D**) Effect of the number S of selfings for populations with $L = 5$ parents and sizes of $N = 50$ and 200, respectively. (**E**) Effect of population size N for $L = 5$ parents. (**F**) Effect of population size N for bi-parental F1 and two independent selfing populations.

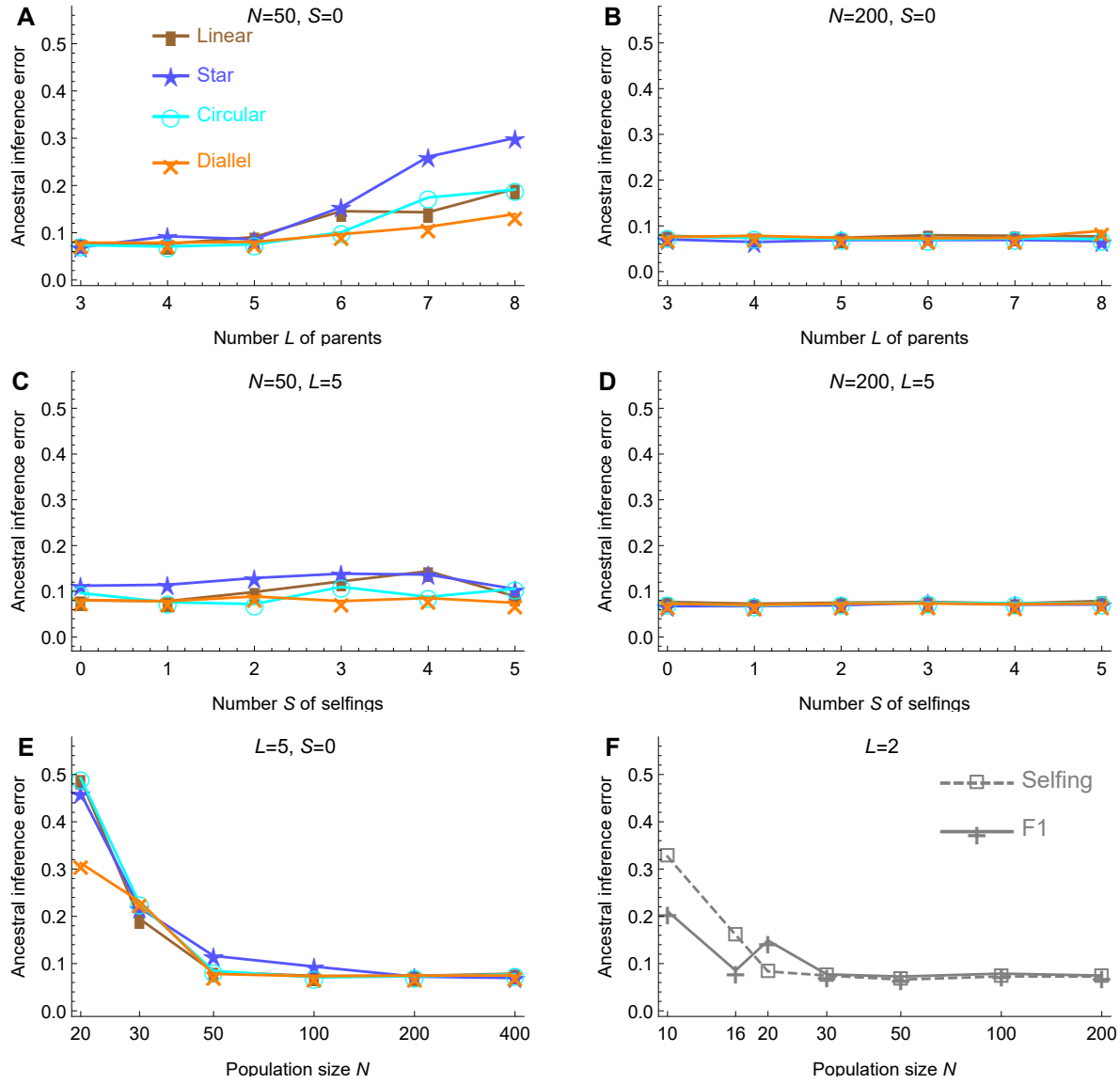

Figure S4: Effect of population design on ancestral inference. (**A**&**B**) Effect of the number L of parents for populations with no selfings ($S = 0$) and sizes of $N = 50$ and 200, respectively. (**C**&**D**) Effect of the number S of selfings for populations with $L = 5$ parents and sizes of $N = 50$ and 200, respectively. (**E**) Effect of population size N for $L = 5$ parents. (**F**) Effect of population size N for bi-parental F1 and two independent selfing populations.

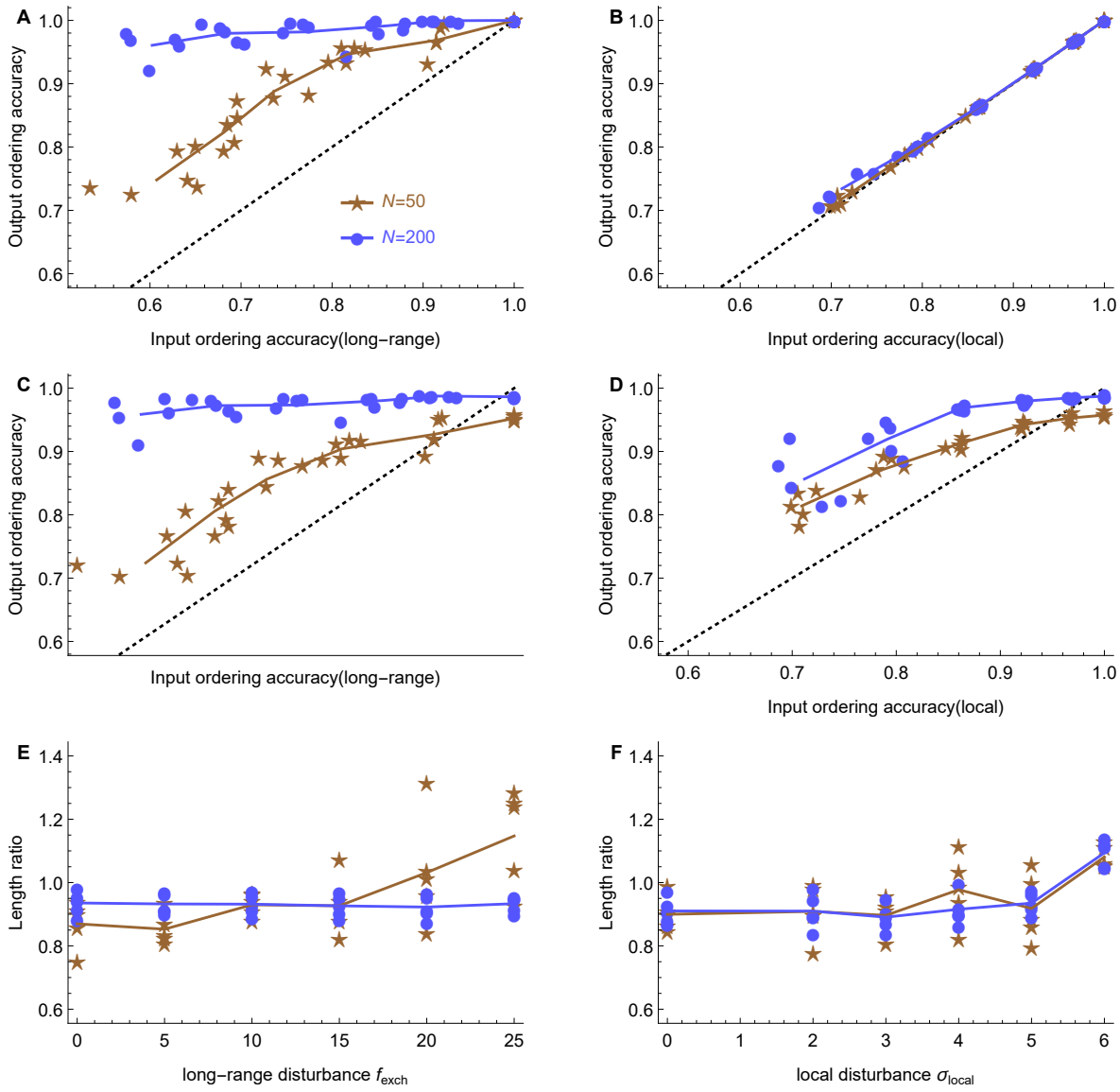

Figure S5: Refinement of the input genetic maps in the presence of long-range or local disturbances in the diallel populations with no selfings ($S = 0$) and $L = 5$ parents. The dotted lines denoting $y = x$. (**A**&**B**) Improvement of marker ordering after the first step of parental phasing and before the second step of map refinement. (**C**&**D**) Improvement of marker ordering after the second step of map refinement. (**E**&**F**) Ratio of final estimated genetic length to the true value.

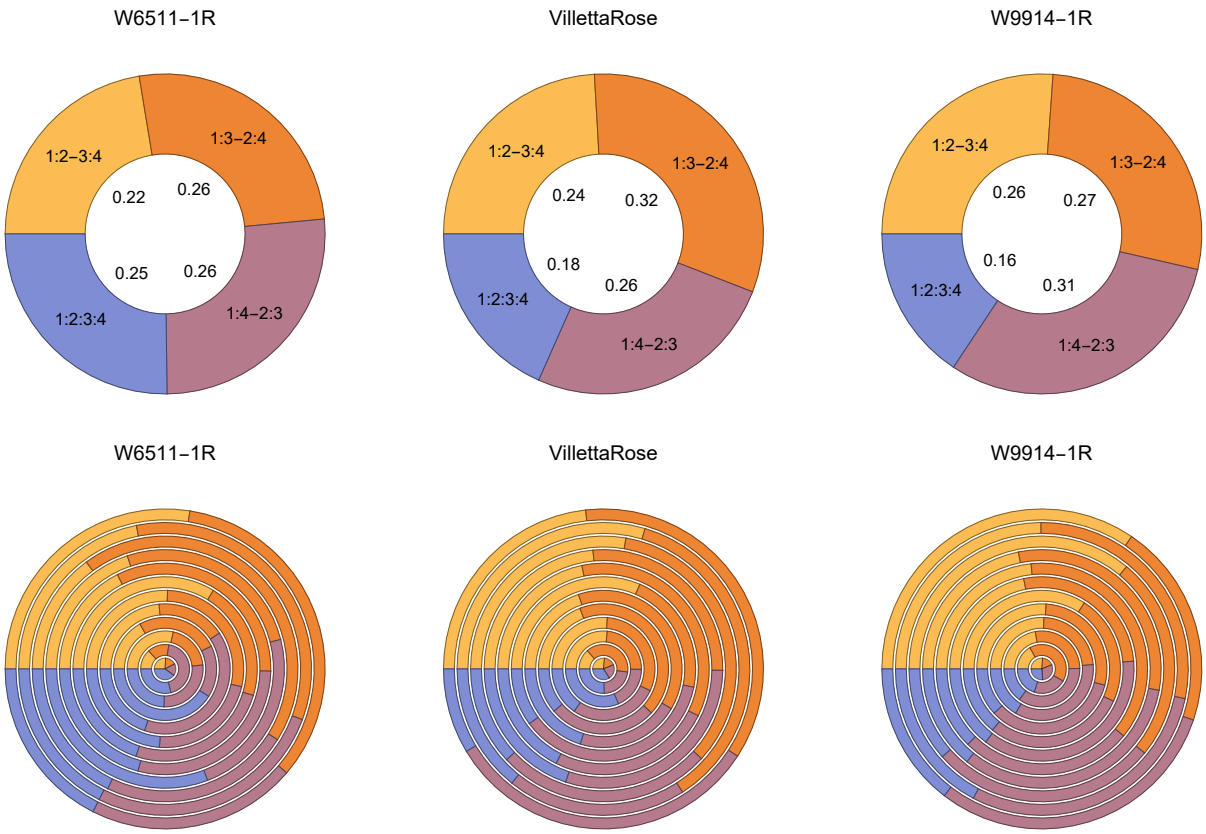

Figure S6: The proportion of valent configurations for the 12 chromosomes of potato in the 3x3 half-diallel with parents VillettaRose, W6511-1R, and W9914-1R. The proportion was calculated based on the maximum possible configurations for each offspring and each chromosome. The configuration 1:2:3:4 refers to a quadrivalent, while the other three refer to bivalent pairs (the colon separates paired homologs). Each bottom panel denotes the proportions among the 12 chromosomes starting from the inner, and the upper panels denote the averages over chromosomes

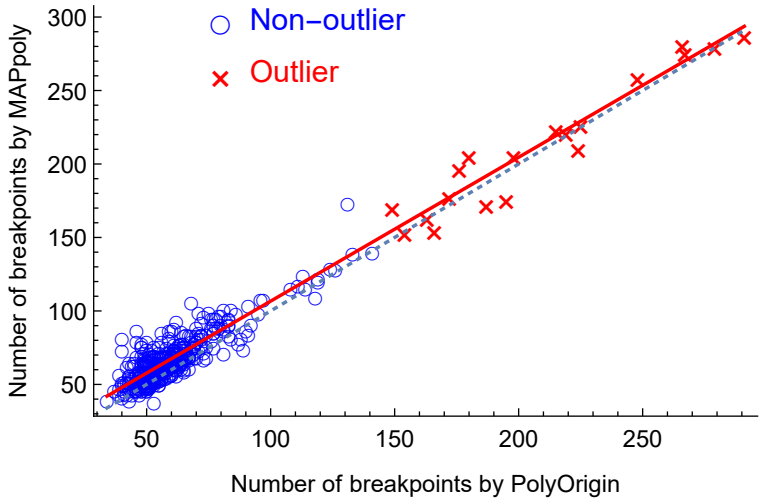

Figure S7: Comparison of PolyOrigin with MAPpoly in terms of the number of haplotype breakpoints for each offspring. Red crosses denote outlier offspring labeled by PolyOrigin, and blue circles denote non-outliers. Dotted line denotes $y = x$, and solid red line denotes the regression line.

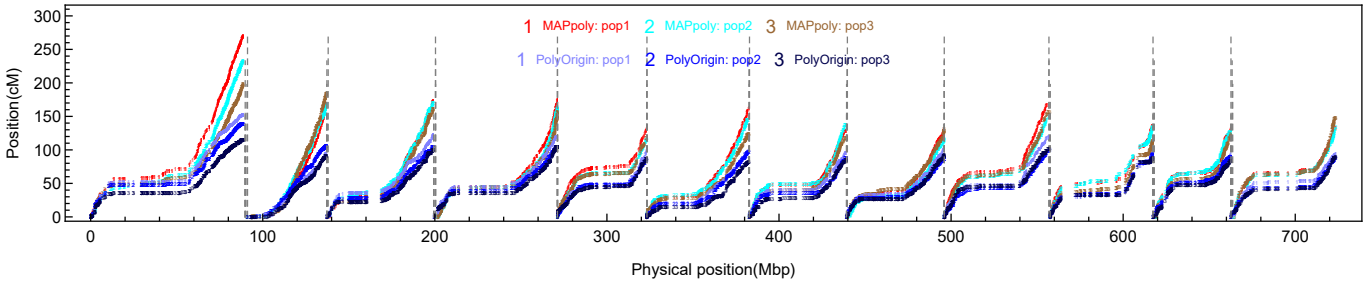

Figure S8: Comparison of PolyOrigin with MAPpoly for each of the three F1 populations in the real 3x3 potato diallel population.

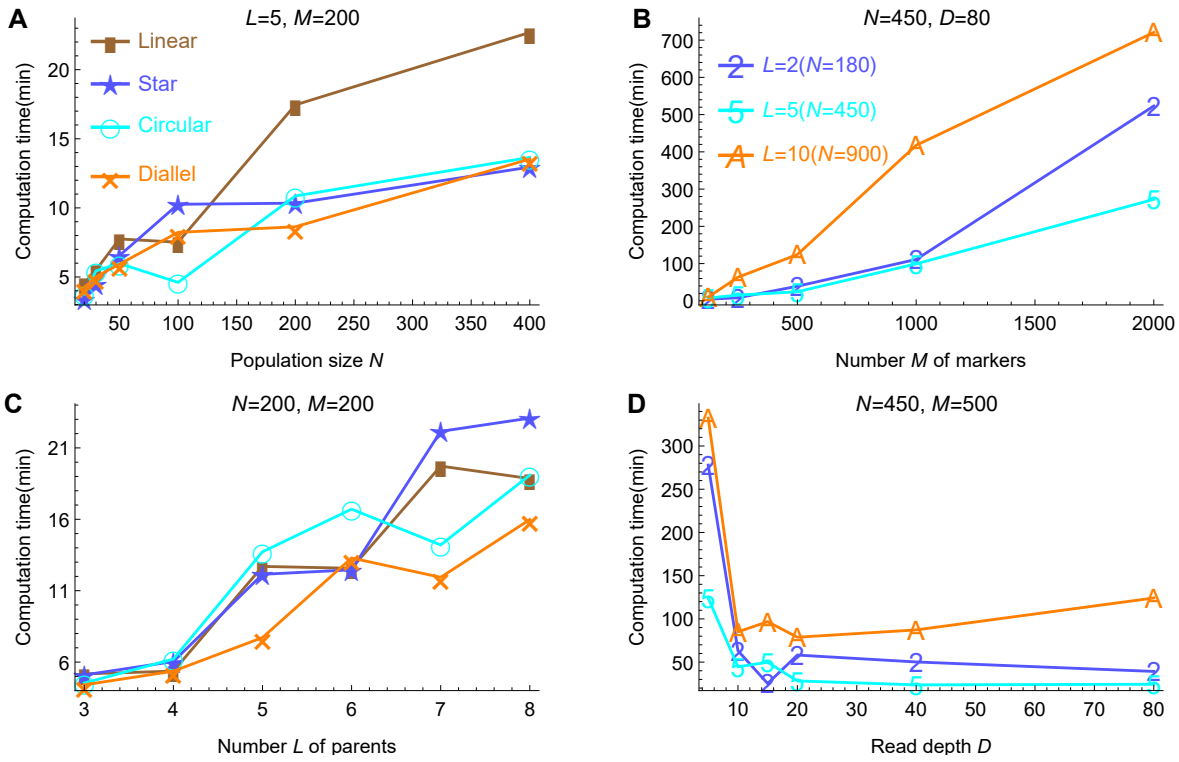

Figure S9: Effect of population design and genotyping design on computational time (in minutes). (**A**&**C**) Computational time used in analyzing the simulated SNP array data in the four mating designs. (**B**&**D**) Computational time used in analyzing the simulated GBS data in the diallel design with $L = 2$, 5, and 10 parents, respectively.