Figure 6: Comparison of BH, AP, localFDR and LIS in the dominant model and the recessive model with different $t$ values and different nominal FDR $\alpha$ values. Table (1) summarizes results. Subfigures (2a)-(2f) shows ROC and PR curves of LIS (solid red lines) and individual $p$-values (dashed green lines) in the dominant model. Subfigures (3a)-(3f) shows ROC and PR curves of LIS and individual $p$-values in the recessive model.

Li et al., 2007; Bernert et al., 2011). The other cluster includes rs11200014, rs2981579, rs1219648, and rs2420946 on chromosome 10. They are exactly the 4 SNPs reported by Hunter et al. (2007). Their associated gene FGFR2 is also well known to be associated with breast cancer. SNP rs4866929 on chromosome 5 is also very likely to be associated because it is highly correlated ($r^2=0.957$) with SNP rs981782 (not included in our data) which was identified from a much larger dataset (4,398 cases and 4,316 controls and a follow-up confirmation stage on 21,860 cases and 22,578 controls) by Easton et al. (2007).

6 Conclusion

In this paper, we use an MRF-coupled mixture model to leverage the dependence in multiple testing problems, and show the improved numerical performance on a variety of simulations and its applicability in a real-world GWAS problem. A theoretical question of interest is whether this graphical model based procedure is optimal in the sense that it has the smallest FNR among all the valid procedures. The optimality of the oracle procedure can be proved under the compound decision framework (Sun & Cai, 2007, 2009), as long as an exact inference algorithm exists or an approximate inference algorithm can be guaranteed to converge to the correct marginal probabilities. The asymptotic optimality of the data-driven procedure (the FNR yielded by the data-driven procedure approaches the FNR yielded by the oracle procedure as the number of tests $m \to \infty$) requires consistent estimates of the unknown parameters in the graphical models. Parameter learning in undirected models is more complicated than in directed models due to the normalization constant. To the best of our knowledge, asymptotic properties of parameter learning for hidden MRFs and MRF-coupled mixture models have not been investigated. Therefore, we cannot prove the asymptotic optimality of the data-driven procedure so far, although we can observe its close-to-oracle performance in the basic simulations.

Acknowledgements

The authors acknowledge the support of the Wisconsin Genomics Initiative, NCI grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NIGMS grant R01GM097618-01, NLM grant R01LM011028-01, NIEHS grant 5R01ES017400-03, eMERGE grant 1U01HG004608-01, NSF grant DMS-1106586 and the UW Carbone Cancer Center.