

GENPHEN: Genotype/Phenotype Association with Reference to Phylogeny

Samuel K Handelman^{1,2}, Joseph S Verducci³, Jesse J Kwick⁴, Surender B Kumar SB⁵, Daniel A Janies^{2*}

1. Mathematical Biosciences Institute; 2. Department of Biomedical Informatics, College of Medicine; 3. Department of Statistics; 4. Division of Infectious Diseases, Department of Microbiology, The Center for Microbial Interface Biology and The Center for Retrovirus Research; 5. Department of Veterinary Biosciences, The Ohio State University, Columbus, OH, USA.

* To whom correspondence should be addressed: Daniel.Janies@osumc.edu

When genome sequences are obtained from organisms with different associated phenotypes, it should be possible to identify those sequence properties which confer a given phenotype. However, the evolutionary relationships between organisms lead to non-independence between the sequence properties. For example, the HIV-1 virus has a population structure reflecting both transmission between individuals and evolution of the HIV-1 quasispecies within each patient. This non-independence can introduce interdependence between unrelated mutations giving a false appearance of causation. These evolutionary relationships are an issue even in HIV-1 where recombination is rapid, and are pervasive in humans, where linkage disequilibrium is extensive. In human disease studies, this can sometimes be overcome by comparing siblings: alleles common only in “sick” siblings are likely true causative alleles. GENPHEN identifies, in a phylogenetic reconstruction, sibling lineages where the phenotype varies. Then, GENPHEN uses modified proportional hazard models to identify causal polymorphisms. GENPHEN’s advantages include: speed practical for high-throughput sequence data, estimates of relative strength or speed of different effects, and improved precision even vs. other tree-based methods: 50%-300% improvement in precision at same recall, either to predict experimental correlations (obtained from STRING: <http://string-db.org/>) or in simulations under biologically reasonable parameters on HIV quasispecies sequence trees.