**Bioinformatics analysis of human genes associated with diseases at higher rates in African Americans.**

Student Applicants: Mislie C. Jean-Baptiste

Faculty Advisor: Hong Qin

We seek support to initiate an innovative line of research on health disparity. Our long-term goal is to develop functional assays in the model organism of *Saccharomyces cerevisiae* for human genes associated with diseases at higher rates in African Americans (DHRAA). To achieve this goal, Miss Jean-Baptistest will conduct a large scale bioinformatics analysis.

Polymorphisms in human genomes are known to be associated with DHRAAs. For example, sickle-cell disease is caused by semi-recessive mutations in the hemoglobin gene [[1](#_ENREF_1" \o "Serjeant, 2010 #744)]. This disease has a high-incidence in AADPs because the mutant alleles of the hemoglobin genes exist at high frequencies in AADPs. The high frequency of the mutant hemoglobin alleles can be attributed to resistance to malaria, an adaptive advantage found in heterozygous carriers [[1](#_ENREF_1" \o "Serjeant, 2010 #744)]. Kidney diseases occurs disproportionally high in AADPs, and have been linked to several genetic factors [[2](#_ENREF_2" \o "Price, 2002 #743)]. Recently, one form of kidney disease, a spectrum of nondiabetic end stage kidney disease (ESKD) that is a DFHAA, was linked with missense mutations in *APOL1* gene [[3](#_ENREF_3" \o "Tzur, 2010 #745)]. These disease variants occur at higher frequencies in two western African populations and have a role in fighting parasitic trypanosome [[3](#_ENREF_3" \o "Tzur, 2010 #745)].

Because most gene variants associated with DFHAAs likely exist in high frequencies in AADPs, these gene variants could be under selection and are outcomes of trade-off during evolution. This reasoning lead us to hypothesize that recently selected gene variants in AADPs may account for some DFHAAs, and will be addressed by the proposed activities.

## Aim: Perform large-scale computational searches to identify yeast homologs for candidate human genes that are under recent-selection in the in AADPs, and evaluate the population variation of these genes in yeast. We are especially interested in genes with copy-number variations (CNVs), climate-mediated selection, and exceptional longevity in humans.

Rationale: CNVs have been associated with autisms, mental retardation, and schizophrenia [[4](#_ENREF_4" \o "McCarroll, 2007 #772)]. Recent studies have found CNVs with high occurrences in AADPs [[5](#_ENREF_5" \o "Hancock, 2011 #713)], and loci with these CNVs may contain genes involved in DHRAAs. A recent study identified a list of SNPs associated with climate-mediate selection in human populations [[5](#_ENREF_5" \o "Hancock, 2011 #713)]. Some of the candidate loci are found to be enriched in a sub-Saharan African population. Climatic variation is an environmental factor that has strong impact on human physiology. It is plausible that some gene variants selected under hot climate offer fitness advantages to young individuals but lead to age-related diseases in elders. Another recent study found 281 SNPs associated with exceptional longevity in humans [[6-8](#_ENREF_6" \o "Sebastiani, 2010 #826)]. Some of the SNPs are located in genes that are known to be associated with aging, such as LMNA, SOD2, and WRN.

Because of evolutionary trade-off and the fact that natural selection mostly acts upon young individuals, gene variants in recent adaptive selection may bear cost in later life, and hence lead to age-related diseases. By investigating potential selection of these genes in yeast and studying their possible phenotypic changes in life history traits, we may gain insights on these functional changes associated with allelic variations of these genes.

Materials and Methods. The key data sources are the list of CNV loci with high occurrences in AADPs, the list of SNPs under climate-mediated selection [[5](#_ENREF_5" \o "Hancock, 2011 #713), [9](#_ENREF_9" \o "McElroy, 2009 #706)] and are associated with exceptional longevity [[7](#_ENREF_7" \o "Sebastiani, 2012 #821)]. We will identify the genes located in those loci, retrieve their coding sequences from UCSC human genome database (http://genome.ucsc.edu/) or the Ensemble human genome database (http://www.ensembl.org/). The yeast resequenced genomes will be obtained from the Saccharomyces Genome Resequencing project at the Wellcome Trust Sanger Institute and the short reads of resequenced yeast genomes deposited the Sequence Read Archive by the Genome Institute at Washington University. The Qin lab has obtained these yeast strains with completely sequenced genomes from the Sanger Institute and the Washington University, and is in the process of phenotyping life history traits of these strains.

Sequence similarity searches will be performed by pairwise BLASTP run on both the human proteins and yeast proteins. Gene families will be identified by Markov clustering based on similarity scores. Reciprocal best hits will be designated as orthologs. The list of human genes will be searched against the Genetic Association Database (http://ncbi.nlm.nih.gov/dbGaP) to identify genes with known association to human diseases. In addition, a search against the database of aging genes (http://www.uwaging.org/genesdb/) will also be conducted.

Selection in yeast homologs will be evaluated by estimate the joint posterior distribution of multiple parameters, including the number of segregating sites (S), a summary of the allele frequency spectrum (Tajima’s *D*), and a summary of distinct haplotypes and linkage disequilibrium in the sample (Fay and Wu’s H) [[10](#_ENREF_10" \o "Barbash, 2004 #836), [11](#_ENREF_11" \o "Przeworski, 2003 #837)]. We will use Hudson’s ms software to simulate the neutral evolution and generate the posterior distribution for the null hypothesis [[12](#_ENREF_12" \o "Hudson, 1983 #843)]. We will use mbs, a modified version of Hudson’s ms software, to conduct coalescent simulations under selection and estimate posterior distribution of selection under various demographic scenarios (i.e., the alternative hypotheses) [[13](#_ENREF_13" \o "Teshima, 2009 #839)]. This approximate Bayesian approach can take complex demographic structures into account [[14](#_ENREF_14" \o "Csillery, 2010 #840)], especially given the known population structures in yeast strains [[15](#_ENREF_15" \o "Schacherer, 2009 #842)].

Expected Results & Limitations. Some human genes are well-conserved and are expected to have yeast orthologs, but other human genes are not expected to have yeast counterparts. For candidate human DFHAA genes conserved in yeast, nonsynonymous polymorphisms (mutations with amino acid changes) will be searched in both human genomes and yeast genomes. It is likely amino acids changes in the same protein regions may influence protein functions in similar ways. Phenotypic differences associated with yeast nonsynonymous polymorphisms can offer clues on functional consequences for similar changes in the human orthologous proteins. For many of the candidate DHRAA SNPs, we expected to find at least one protein-coding gene at or near its chromosomal location, and we will focus on these protein-coding genes during this project. It is quite plausible that non-protein genes, such non-coding RNAs, may play a role in DHRAAs, but they will be pursued in separate studies.

# References

# 1. Serjeant, G.R., *One hundred years of sickle cell disease.* Br J Haematol, 2010. 151(5): p. 425-9.

# 2. Price, D.A. and E.D. Crook, *Kidney disease in African Americans: genetic considerations.* J Natl Med Assoc, 2002. 94(8 Suppl): p. 16S-27S.

# 3. Tzur, S., S. Rosset, R. Shemer, G. Yudkovsky, S. Selig, A. Tarekegn, E. Bekele, N. Bradman, W.G. Wasser, D.M. Behar, and K. Skorecki, *Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene.* Hum Genet, 2010. 128(3): p. 345-50.

# 4. McCarroll, S.A. and D.M. Altshuler, *Copy-number variation and association studies of human disease.* Nat Genet, 2007. 39(7 Suppl): p. S37-42.

# 5. Hancock, A.M., D.B. Witonsky, G. Alkorta-Aranburu, C.M. Beall, A. Gebremedhin, R. Sukernik, G. Utermann, J.K. Pritchard, G. Coop, and A. Di Rienzo, *Adaptations to climate-mediated selective pressures in humans.* PLoS Genet, 2011. 7(4): p. e1001375.

# 6. Sebastiani, P., N. Solovieff, A. Puca, S.W. Hartley, E. Melista, S. Andersen, D.A. Dworkis, J.B. Wilk, R.H. Myers, M.H. Steinberg, M. Montano, C.T. Baldwin, and T.T. Perls, *Genetic signatures of exceptional longevity in humans.* Science, 2010. 2010.

# 7. Sebastiani, P., N. Solovieff, A.T. Dewan, K.M. Walsh, A. Puca, S.W. Hartley, E. Melista, S. Andersen, D.A. Dworkis, J.B. Wilk, R.H. Myers, M.H. Steinberg, M. Montano, C.T. Baldwin, J. Hoh, and T.T. Perls, *Genetic signatures of exceptional longevity in humans.* PLoS One, 2012. 7(1): p. e29848.

# 8. Sebastiani, P., A. Riva, M. Montano, P. Pham, A. Torkamani, E. Scherba, G. Benson, J.N. Milton, C.T. Baldwin, S. Andersen, N.J. Schork, M.H. Steinberg, and T.T. Perls, *Whole genome sequences of a male and female supercentenarian, ages greater than 114 years.* Front Genet, 2011. 2: p. 90.

# 9. McElroy, J.P., M.R. Nelson, S.J. Caillier, and J.R. Oksenberg, *Copy number variation in African Americans.* BMC Genet, 2009. 10: p. 15.

# 10. Barbash, D.A., P. Awadalla, and A.M. Tarone, *Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus.* PLoS Biol, 2004. 2(6): p. e142.

# 11. Przeworski, M., *Estimating the time since the fixation of a beneficial allele.* Genetics, 2003. 164(4): p. 1667-76.

# 12. Hudson, R.R., *Properties of a neutral allele model with intragenic recombination.* Theor Popul Biol, 1983. 23(2): p. 183-201.

# 13. Teshima, K.M. and H. Innan, *mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection.* BMC Bioinformatics, 2009. 10: p. 166.

# 14. Csillery, K., M.G. Blum, O.E. Gaggiotti, and O. Francois, *Approximate Bayesian Computation (ABC) in practice.* Trends Ecol Evol, 2010. 25(7): p. 410-8.

# 15. Schacherer, J., J.A. Shapiro, D.M. Ruderfer, and L. Kruglyak, *Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae.* Nature, 2009. 458(7236): p. 342-5.