This Month in the *Journal*

Robin E. Williamson[1]

Multi-Locus Meta Analysis

Verzilli et al., 859
By utilizing the data produced by multiple independent-association studies, meta-analyses can increase sample size and the power to detect an association. Difficulties arise when the studies being combined use different platforms and/or genotype different markers. It is possible to exclude data from SNPs that are not common among the studies, but valuable information can be lost in this process. Ideally, the combination of studies would include imputing missing SNPs on the basis of LD relationships with the markers that were typed. Verzilli et al. develop a method that incorporates LD from databases or other studies to generate data from markers across all studies. Additionally, because raw data are not always provided from the independent analyses, the method is able to utilize summary data in its measurements. The authors use their method to combine the data from different studies that look at genetic associations between variants in the C-reactive protein (CRP) region and levels of CRP in plasma. By using all information, they are able to fine map the region and parse out independent effects of the analyzed SNPs.

Human Mitochondrial Substitution Rate

Endicott and Ho, 895
When studying evolutionary events such as selection, population migrations, and population expansions, researchers make timing estimates on the basis of genetic data. A great deal of such work has been done via analysis of mitochondrial DNA (mtDNA), but accurately dating events with mtDNA requires a trustworthy measurement of mtDNA substitution rates. If the substitution rate is underestimated, the timescale to a historical moment will be inflated; likewise, if the substitution rate is too high, history will be condensed. Methods of determining the rates for different regions of mtDNA have been proposed, but there hasn't been a standardized way to establish the substitution rates across the mtDNA. Endicott and Ho developed a Bayesian model-based approach to do just that. Using the new method, the authors assess the importance of calibration in the rate estimations. They find that using an external human-chimpanzee source yields greatly different results than using an internal calibration based on known historical time points. Because the estimation of substitution rate is lower when one uses the external calibration, the corresponding time to the most recent ancestor is higher. The internally calibrated estimates of substitution rates are higher than those previously reported, meaning that divergence times are determined to be more recent than previously thought. These new times correlate well with archaeological evidence.

SNP Arrays in Heterogeneous Tissue Samples

Assie et al., 903
Because of the importance of determining the genetic differences between normal tissue and cancer cells, methods of comparing germline and somatic genotypes have been developed. Genome-wide methods utilizing data from SNP arrays have been implemented, but current analyses can be limited by the need to have matched normal samples for each tumor sample, as well as by the contamination of normal cells mixed into tumor samples. Assie et al. introduce SOMATICs, which uses the normal cells within a tumor sample as an internal control. By taking advantage of tissue heterogeneity, SOMATICs is able not only to overcome several of the problems faced by other methods but also to detect genetic aberrations with increased sensitivity. When applied to 54 samples, SOMATICs identifies small genetic alterations that occur in a small number of cells. The method can also differentiate germline changes from somatic changes and delineate the boundaries of the alterations. Importantly, SOMATICs can also estimate the proportion of affected cells in a sample along with a significance value for this measurement. This determination of tumor content is important for diagnostic predictions but is difficult to determine through other means.

Small RNAs Mapping in Encode Regions

Borel et al., 971
The human genome has been sequenced, but researchers are still trying to figure out what all that sequence does. Increasing evidence suggests that a large number of genomic regions are transcribed, but for the most part not in the traditional gene (ie. coding) sense. The importance of noncoding RNAs has been recognized in work including studies of antisense RNAs and microRNAs. Variants in the regions generating such RNAs are also being evaluated for their involvement in increasing disease risk or in contributing to other phenotypic characteristics. If one wishes to better understand what all these RNAs do, it is useful to

[1]Deputy Editor, *AJHG*
DOI 10.1016/j.ajhg.2008.03.009.

first establish what and where all the RNAs are. Studies have cataloged long RNAs and, here, Borel et al. comprehensively analyze the small RNAs, 19–50 nucleotides in length, that are generated from within the 44 ENCODE regions. An interesting correlation is noted between an increase in the small RNAs near the transcription start site of genes and higher gene expression. This finding is in line with a previously reported observation that an increase in aborted transcripts is related to an increase in gene expression. Additionally, a high proportion of sense/anti-sense RNA pairs are identified, thereby contributing data to the growing evidence of extensive antisense transcription in the genome.

Identification of the SPG15 Gene

Hanein et al., 992

Hereditary spastic paraplegia (HSP) can be uncomplicated or accompanied by other features, ranging from mental retardation and deafness to epilepsy and retinitis pigmentosa. A locus for SPG15, a complicated and autosomal-recessive (AR) form of the disease, was recently mapped to 14q22-q24 in two Irish families. Of the 16 known loci for complicated AR-HSP, SPG15 was predicted to account for 15% of AR-HSP cases. Hanein et al. study eight SPG15 families to further refine the locus and determine which gene is disrupted in this form of AR-HSP. In a first round of screening, three candidate genes in the region are chosen on the basis of their involvement with intracellular trafficking or mitochondrial function, but no disease-causing variants are found. The sequence of *ZFYVE26* is then analyzed because other genes in the same family have been found to be disrupted in related diseases. Truncation mutations in *ZFYVE26* are identified in each of the families. Expression work reveals that the gene is widely expressed and that the localization of the gene product suggests a potential role in endosomal trafficking.