久留米大学
バイオ統計センター
公開セミナー

日時：2016年7月7日(木) 15:00-17:00

場所：久留米大学バイオ統計センター
コンピュータ室（医学部B棟7階）

講師：**Dr. John B. Cologne**

（放射線影響研究所 統計部）

演題：**Experiences with Genomic Analysis: Cryptic Relatedness, Validation, and SNP Sets (Pathways)**

参加無料、申込不要です。
皆様、奮ってご参加ください。

http://www.biostat-kurume-u.jp/

EXPERIENCES WITH GENOMIC ANALYSIS: CRYPTIC RELATEDNESS, VALIDATION, AND SNP SETS (PATHWAYS)

It was recognized during the GWAS era that much of the expected genetic contribution to complex diseases would not reflected in common SNPs (typically variants with minor allele frequency >5%). Thus, there has been a shift towards assessing gene association with low-frequency (☐0.5%) and rare (<0.5%) variants, or with combinations of common, low-frequency, and rare variants (these cutoffs are merely for classification; in fact, variant frequency follows a continuous spectrum and there is no discrete functional distinction between rare, low-frequency, and common, although loss-of-function variants are more likely to be rare). It also has become apparent that individual-SNP analyses lack power for detecting gene-disease associations, but that power can be improved by combining SNPs that contribute to common pathways or mechanisms (gene sets; Lee et al, Am J Hum Genet 2014; 95:5-23). This is especially true with rare variants, except with very large studies and/or very large variant risks. One well-known tool for performing SNP-set analyses is the sequence kernel association test (SKAT; Wu et al, Am J Hum Genet 2010; 86:929-942). I will briefly describe kernel regression and then illustrate it with an application of SKAT to SNP-set analyses of various gene sets and colon cancer in the RERF Immuno-genome (IMG) Cohort Study. The IMG study, based on sampling from the RERF Adult Health Study (AHS) cohort, which in turn was selected from the RERF Life Span Study (LSS) cohort, was designed to investigate cancer risk for gene-radiation interactions involving genomic variants in immune-function, inflammation-related, and DNA-repair genes. The original study involved collecting peripheral blood lymphocytes from 7,144 AHS participants for an immunological study, beginning in 1981. After exclusion of participants with unknown radiation dose, who had cancer prior to lymphocyte collection, who were aged 80 or older at time of lymphocyte collection, who were in utero survivors, or whose sample did not have extractable DNA, 4,683 participants remained; the "IMG cohort" comprises these individuals. The TaqMan-allelic discrimination method was used for the detection of 343 SNPs. No rare variants were included among the selected loci, so we will not illustrate rare-variant analyses per se, but future work at RERF should include low-frequency and rare variants. Because of the cost of whole-exome or whole-genome sequencing, we implemented a stratified case-cohort design (with stratification on radiation dose) for future studies in the IMG cohort (Cologne et al, Int J Epidemiol 2012; 41:1174-1186). We will briefly describe the results of those simulations.

Prior to conducting genomic analysis, various quality control steps are required to ensure valid data for analysis. In population genomic studies, in addition to filtering SNPs based on genotype calling rate and Hardy-Weinberg equilibrium, as well as identifying and removing erroneously duplicated genotype data, it is necessary to assess (and, when present, adjust for) population stratification (slightly inbred sub-populations) and cryptic relatedness (closely related individuals in the sample). Cryptic relatedness may be a particularly important problem in RERF studies, because the original LSS cohort selection process (which included as many proximally exposed individuals as possible) would have resulted in many family members being included, and these would have been carried forward into the AHS cohort because it included a large fraction of proximally exposed persons from the LSS. It is therefore likely that some family members carried forward into the IMG cohort; investigation of family relatedness in the IMG cohort via the genotype matrix will therefore be briefly described. Due to ethical and privacy considerations, we cannot store or reveal results on what relationships or which individuals are involved, but we can summarize the extent of relatedness and we can use the fact of relatedness to make appropriate adjustments in the analysis. Relatedness can be dealt with in the analysis using subject exclusion, genomic control (adjustment of the test statistics for variance inflation; Devlin and Roeder, Biometrics 1999; 55:997-1004), or a random-effects (variance-component) model (Kang et al, Nature Genetics 2010; 348-354). All of these approaches are conducted blind and non-linkable with regard to subject identity (unlinkably anonymized).

A major criticism of GWAS results in the open literature has been a frequent lack of reproducibility, despite the strict measures taken to reduce the likelihood of declaring false positive results (usually based on Bonferroni adjustment of P values). Although gene-set analyses circumvent this problem to a large extent by drastically reducing the number of tests performed, it is still scientifically prudent to validate findings if possible. We will therefore describe comparisons between the IMG cohort and the Japanese sub-population of the Multi-Ethnic Cohort (MEC) study conducted in Los Angeles and Honolulu. Comparisons were limited to SNPs that were genotyped in both cohorts; because the available SNP data in the MEC cohort are from the MetaboChip, which was designed for examining metabolic- and cardiovascular-disease-related genomic variants but not cancer, overlap is limited to only 33 SNPs. Also, because there is no radiation exposure in the MEC study, comparisons are limited to main effects of variants.

Finally, results for pathway analysis of immune-function, inflammation-related, and DNA-repair gene sets in the IMG cohort will be briefly described. Following analysis by SKAT, individual SNPs in promising gene sets can be examined in greater detail via PLINK using standard association analysis and logistic regression. Gene-radiation interaction can also be assessed using various modifications of SKAT, including iSKAT (Lin et al, Biometrics 2016; 72:156-64, DOI:10.1111/biom.12368) and fastKM ("fast Kernel Machine"; Marceau et al, Genet Epidemiol 2015; 39:456-468).

In concluding, future needs to improve these analyses and potential areas of methodological collaboration between RERF and Kurume University biostatisticians will be discussed.