Research achievements (2008.01.01 – 2012.07.31)

Hsin-Chou Yang

With the support from Institute of Statistical Science Academia Sinica and grants of the Academia Sinica Career Development Award, National Science Council and National Research Program of Genomic Medicine, we devoted to developing novel statistical/bioinformatics methodologies and tools for genomic data analysis as well as to answering practical biomedical issues via close collaborations with epidemiologists, clinicians and geneticists. My research achievements in the recent years are outlined as follows:

- **Microarray pooled DNA analyzer** (Yang et al., 2008, *BMC Bioinformatics*)
- **Genome-wide association tests** (Yang et al., 2008, *Genetics*; Yang et al., 2009, *BMC Proceedings*; Yang and Chen, 2011, *BMC Proceedings*)
- **Allele frequency estimation and applications** (Yang et al., 2010, *BMC Genomics*)
- **Loss-of-heterozygosity detection** (Huggins et al., 2008, *Journal of Human Genetics*; Yang et al., 2011, *Genetic Epidemiology*; Yang et al., 2011, *PLoS ONE*)
- **SNP array quality control** (Yang et al., 2011, *BMC Bioinformatics*)
- **Integrative analysis of SNPs and gene expression** (Yang et al., 2012, *BMC Genomics*)
- **Genome-wide association study of hypertension in the Han Chinese population of Taiwan** (Yang et al., 2009, *PLoS One*; Yang et al., 2012, *PLoS One*)

The research achievements are briefly introduced below.

(1) **Microarray pooled DNA analyzer:** We have developed analysis tools for microarray-based pooled DNA analysis. We proposed a generalized concept of pooled DNA, and developed a user-friendly tool named Microarray Pooled DNA Analyzer (MPDA) (Yang et al., 2008, *BMC Bioinformatics*) that we developed to analyze hybridization intensity data from microarray-based pooled DNA experiments. MPDA enables whole-genome DNA preferential amplification/hybridization analysis, allele frequency estimation, association mapping, allelic imbalance detection, and permits integration with shared data resources online. Graphic and numerical outputs from MPDA support global and detailed inspection of large amounts of genomic data. The software, user manual and illustrated examples are available at the MPDA website (http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm).

(2) **Genome-wide association tests:** We have developed three types of genome-wide association mapping methods: (1) kernel-based association test (KBAT), (2) gene-based association test (GBAT), and (3) pathway-based association test (PBAT). First, we proposed a kernel-based association test (KBAT) (Yang et al., 2008, *Genetics*), which is a composite function of ‘‘P-values of single-locus association tests’’ and ‘‘kernel weights related to intermarker distances and/or linkage disequilibria.’’ The KBAT is a general form of some current test statistics. This method can be applied to the study of candidate genes and can scan each chromosome using a moving average procedure. Compared with some existing single-locus and multilocus methods the proposed kernel-based method has the following merits: KBAT (a) is robust against the inclusion of nuisance markers, (b) is invariant to the map scale, and (c) accommodates different types of genomic data, study designs,

and study purposes. The software, user manual and illustrated examples are freely available online at the KBAT website (http://www.stat.sinica.edu.tw/hsinchou/genetics/association/KBAT.htm). The work is selected to be one of significant research achievements of Academia Sinica. Second, we proposed a two-stage genome-wide association scanning procedure, consisting of a single-locus association scan for the first stage and a gene-based association scan for the second stage (Yang et al., 2009, *BMC Proceedings*). Marginal effects of SNPs were examined by using the exact Armitage trend test or logistic regression, and gene effects were examined by using a p-value combination method. Compared with some existing single-locus and multilocus methods the proposed gene-based method has the following merits: GBAT is (a) convenient for definition of biologically meaningful regions, (b) powerful for detection of minor-effect genes, (c) helpful for alleviation of a multiple-testing problem, and (d) convenient for result interpretation. Finally, the proposed gene-based method was further extended to examine association between biological pathways and study diseases (Yang and Chen, 2011, *BMC Proceedings*). We are developing Omnibus P-value Association Tests (OPATs) software to integrate KBAT, GBAT and PBAT into an analysis system for genomic association studies.

(3) **Allele frequency estimation and applications:** We have proposed a unified procedure to estimate individual-level and population-level allele frequencies based on hybridization intensity data. Allele frequency is one of the most important population indices and has been broadly applied to genetic/genomic studies. Estimation of allele frequency using genotypes is convenient but may lose data information and be sensitive to genotyping errors. We utilized a unified intensity-measuring approach to estimating

individual-level allele frequencies for 1,104 and 1,270 samples genotyped with the single-nucleotide-polymorphism arrays of the Affymetrix Human Mapping 100K and 500K Sets, respectively. Allele frequencies of all samples were estimated and adjusted by coefficients of preferential amplification/hybridization (CPA), and large ethnicity-specific and cross-ethnicity databases of CPA and allele frequency were established. The results showed that using the CPA significantly improves the accuracy of allele frequency estimates; moreover, this paramount factor is insensitive to the time of data acquisition, effect of laboratory site, type of gene chip, and phenotypic status. Based on accurate allele frequency estimates, analytic methods based on individual-level allele frequencies were developed and successfully applied to discover genomic patterns of allele frequencies, detect chromosomal abnormalities, classify sample groups, identify outlier samples, and estimate the purity of tumor samples. The methods were packaged into a new analysis tool, ALOHA (Allele-frequency/Loss-of-heterozygosity/Allele-imbalance). This is the first time that these important genetic/genomic applications have been simultaneously conducted by the analyses of individual-level allele frequencies estimated by a unified intensity-measuring approach. ALOHA and the user manual and illustrated examples are available at http://www.stat.sinica.edu.tw/hsinchou/genetics/aloha/ALOHA.htm. The developed databases of allele frequency and CPA are accessible at http://140.109.72.48/index.htm.

(4) **Loss-of-heterozygosity detection:** We have developed statistical methods and tools for mapping loss of heterozygosity (LOH) (Huggins et al., 2008, *Journal of Human Genetics*; Yang et al., 2011, *Genetic Epidemiology*). LOH occurs when genotypes change from a heterozygous state to a hemizygous or

homozygous state, where an allele or haplotype from one parent is lost. In the first step we determined which chromosomes significantly differed between the unpaired tumor and normal samples by using a nonparametric procedure. We then used a biplot data visualization technique and homozygosity intensity estimates to determine the regions of these chromosomes that required further examination (Huggins et al., 2008, *Journal of Human Genetics*; Yang et al., 2011, *Genetic Epidemiology*). Recently, we found that the pattern of homozygosity is critically important not only in cancer research but also in population genetics and genetics of complex diseases (Yang et al., 2011, *Genetic Epidemiology*; Yang et al., 2012, *PLoS One*). We examined LOH in cancer patients, detected long contiguous stretches of homozygosity (LCSH) in general populations, and located recessive acting susceptibility genes for complex diseases. Powerful LOHAS software was established as the first tool, which can simultaneously depict genomic profiling of LOH/LCSH, detect regions of LOH/LCSH, cluster samples with close LOH/LCSH structures, and identify samples with unusual LOH/LCSH patterns. Tumor suppressor genes of leukaemia, recessive acting genes of rheumatoid arthritis, and genes under selective pressure were identified by LOHAS successfully (Yang et al., 2011, *Genetic Epidemiology*; Yang et al., 2012, *PLoS One*). The software and illustrated examples are freely available online at the LOHAS website (http://www.stat.sinica.edu.tw/hsinchou/genetics/loh/LOHAS.htm).

(5) **SNP array quality control:** We have developed a useful tool for quality control of SNP microarrays (Yang et al., 2011, *BMC Bioinformatics*). Data quality of SNP arrays plays a key role on the accuracy and precision of down-stream data analyses. However, good quality indices still await development. We introduced new quality indices, established references of

allele frequency and quality indices, investigated statistical properties of quality indices, and developed a detector for poor SNP chips and/or DNA samples. SNP Array Quality Control (SAQC) written in R and R-GUI was developed as a user-friendly tool for visualization and evaluation of data quality of genome-wide SNP chips. In practice, SAQC has been used to successfully solve disputes about genotyping quality between National Genotyping Center in Taiwan and its users. The software and illustrated examples are available at http://www.stat.sinica.edu.tw/hsinchou/genetics/quality/SAQC.htm.

(6) **Integrative analysis of SNPs and gene expression:** We first proposed to integrate SNPs and gene expression to distinguish samples from closely related population (Yang et al., 2012, *BMC Genomics*). Ancestry informative markers (AIMs) are a type of genetic marker that is informative for tracing the ancestral ethnicity of individuals. Application of AIMs has gained substantial attention in forensic sciences, population genetics, and medical genetics. SNPs, the materials of AIMs, are useful for classifying individuals from distinct continental origins but cannot discriminate individuals with subtle genetic differences from closely related ancestral lineages. Proof-of-principle studies have shown that gene expression (GE) also is a heritable human variation that exhibits differential intensity distributions among ethnic groups. GE supplies ethnic information supplemental to SNPs; this motivated us to integrate SNPs and GE to construct AIM panels with a reduced number of required markers and provide high accuracy in ancestry inference. Few studies in the literature have considered GE in this aspect, and none have integrated GE and SNPs to aid classification of samples from closely related ethnic populations. We integrated a forward selection procedure into flexible discriminant analysis to identify key SNPs and/or GE with the highest cross-validation prediction accuracy. By analyzing

genome-wide SNPs and/or GE markers in 210 independent samples from four ethnic groups in the HapMap II Project, we found that average testing accuracies for a majority of classification analyses were quite high, except for SNP-only analyses that were performed to discern study samples containing individuals from two close Asian populations. The average testing accuracies ranged from 0.53 to 0.79 for SNP-only analyses and increased to around 0.90 when GE markers were integrated together with SNPs for the classification of samples from closely related Asian populations. Compared to GE-only analyses, integrative analyses of SNPs and GE showed comparable testing accuracies and a reduced number of selected markers in AIM panels. Integrative analysis of SNPs and GE provides high-accuracy and/or cost-effective classification results for assigning samples from closely related or distantly related ancestral lineages to their original ancestral populations. User-friendly BIASLESS (**B**iomarkers **I**dentification **a**nd **S**amp**les S**ubdivision) software was developed as an efficient tool for selecting key SNPs and/or GE markers and then building models for sample subdivision. BIASLESS can be downloaded at http://www.stat.sinica.edu.tw/hsinchou/genetics/classification/BIASLESS.htm .

(7) **Genome-wide association study of hypertension in the Han Chinese population of Taiwan:** We have conducted disease gene mappings of hypertension in the Han Chinese population of Taiwan by using Affymetrix 100K SNP chips (Yang et al., 2009, *PLoS One*). This is the first genome-wide hypertension association study in the Han Chinese population. We carried out a two-stage association scan to map young-onset hypertension susceptibility genes. The first-stage analysis, a genome-wide association study, analyzed 175 matched case-control pairs; the second-stage analysis, a confirmatory association study, verified the results at the first stage based on a

total of 1,008 patients and 1,008 controls. Single-locus association tests, multilocus association tests and pair-wise gene-gene interaction tests were performed to identify young-onset hypertension susceptibility genes. Four SNPs from two SNP triplets with strong association signals ($-\log_{10}(p)>7$) and 13 SNPs from 8 interactive SNP pairs with strong interactive signals ($-\log_{10}(p)>8$) were carefully re-examined. The confirmatory study verified the association for a SNP quartet 219 kb and 495 kb downstream of *LOC344371* (a hypothetical gene) and *RASGRP3* on chromosome 2p22.3, respectively. Intrinsic synergy involving *IMPG1* on chromosome 6q14.2-q15 was also verified. The genes are novel hypertension targets identified in this first genome-wide hypertension association study of the Han Chinese population. Recently, we further conducted the first genome-wide gene-based association scan for hypertension in a Han Chinese population (Yang et al., 2012, *PLoS One*). By analyzing genome-wide single-nucleotide-polymorphism data of ~400 matched pairs of young-onset hypertensive patients and normotensive controls genotyped with the Illumina HumanHap550-Duo BeadChip, 100 susceptibility genes for hypertension were identified and also validated with permutation tests. Seventeen of the 100 genes exhibited differential allelic distributions and differentially expressed distributions between patient and control groups. These genes provided a good molecular signature with an accuracy rate of 96% for classifying hypertensive patients and normotensive controls. Among the 17 genes, three genes (*IGF1*, *SLC4A4*, and *WWOX*) were not only identified by our gene-based association scan and gene expression analysis but were also replicated by a gene-based association analysis of the Hong Kong Hypertension Study. Identification of these genes enriches the collection of hypertension susceptibility genes, thereby shedding light on the

etiology of hypertension in Han Chinese populations.

In summary, my studies have been developed with the consideration of both statistical methods and practical applications. In statistical science, we have developed novel statistical tools for quality control of genomic data and detections of disease association and chromosomal aberrations using different types of genomic data. The performance of these tools has been carefully evaluated using large amount of real data and simulation studies. Not only statistical methodologies, but also user-friendly software and online-shared databases are made available to the entire research community. In biomedical science, we have collaborated closely with biologists, clinicians, and epidemiologists to study complex disorders and cancers. The achievements of these studies have led to identifications of important disease genes, and expanded our knowledge of the etiology of complex disorders and mechanisms of tumorigenesis.

Publications（*Corresponding author）:

1. **Yang, H.-C.***, Huang, M.-C., Li, L.-H., Lin, C.-H., Yu, L. T., Diccianni, M. B., Wu, J.-Y., Chen, Y.-T. and Fann, C. S. J.* (2008/04). MPDA: microarray-based pooled DNA analyzer. *BMC Bioinformatics* 9, 196. SCI.
2. **Yang, H.-C.***, Hsieh, H.-Y. and Fann, C. S. J. (2008/06). KBAT: Kernel-based association test. *Genetics* **179**, 1057-1068. SCI.
3. Huggins, R., Li, L.-H., Lin, Y.-C., Yu, A.L.T. and **Yang, H.-C.*** (2008/12). Nonparametric estimation of LOH using Affymetrix SNP genotyping arrays for unpaired samples. *Journal of Human Genetics* **53**, 983-990. SCI.
4. **Yang, H.-C.**, Liang, Y.-J., Wu, Y.-L., Chung, C.-M., Chiang, K.-M., Ho, H.-Y., Ting, C.-T., Lin, T.-H., Sheu, S.-H., Tsai, W.-C., Chen, J.-H., Leu, H.-B., Yin, W.-H., Chiu, T.-Y., Chen, C.-I., Fann, C.S.J., Wu, J.-Y., Lin, T.-N., Lin, S.-J, Chen, Y.-T., Chen,

J.-W.* and Pan, W.-H.* (2009/05). Genome-wide association study of young-onset hypertension in the Han Chinese Population of Taiwan. *PLoS One* **4**, e5459. SCI.

5. **Yang, H.-C.***, Liang, Y.-J., Chung, C.-M., Chen, J.-W. and Pan, W.-H. (2009/12). Genome-wide gene-based association study. *BMC Proceedings* **3**, S135. PubMed.
6. **Yang, H.-C.***, Lin, H.-C., Huang, M.-C., Li, L.-H., Pan, W.-H., Wu, J.-Y. and Chen, Y.-T. (2010/07). A new analysis tool for individual-level allele frequency for genomic studies. *BMC Genomics* **11**, 415. SCI.
7. **Yang, H.-C.***, Lin, H.-C., Kang, M., Chen, C.-H., Lin, C.-W., Li, L.-H., Wu, J.-Y., Chen, Y.-T. and Pan, W.-H. (2011/04). SAQC: SNP array quality control. *BMC Bioinformatics* **12**, 100. SCI.
8. **Yang, H.-C.***, Chang, L.-C., Huggins, R. M., Chen, C.-H. and Mullighan, C. G. (2011/05). LOHAS: Loss-of-heterozygosity analysis suite. *Genetic Epidemiology* **35**, 247-260. SCI
9. **Yang, H.-C.*** and Chen, C.-W. (2011/11). Region-based and pathway-based QTL mapping using a p-value combination method. *BMC Proceedings* **5**, S43. PubMed.
10. **Yang, H.-C.**, Liang, Y.-J, Chen, J.-W., Chiang, K.-M., Chung, C.-M., Ho, H.-Y., Ting, C.-T., Lin, T.-H., Sheu, S.-H., Tsai, W.-C., Chen, J.-H., Leu, H.-B., Yin, W.-H., Chiu, T.-Y., Chern, C.-I., Lin, S.-J., Tomlinson, B., Guo, Y., Sham, P. C., Cherny S. S., Lam, T. H., Thomas, G. N. and Pan, W.-H.* (2012/03). A genome-wide gene-based association study identifies *IGF1*, *SLC4A4*, *WWOX* and *SFMBT1* as hypertension susceptibility genes in a Han Chinese population. *PLoS One* **7**, e32907. SCI.
11. **Yang, H.-C.***, Chang, L.-C., Liang, Y.-J., Lin, C.-H. and Wang, P.-L. (2012/04). A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human Major Histocompatibility Complex. *PLoS One* **7**, e34840. SCI.
12. **Yang, H.-C.***, Wang, P.-L., Lin, C.-W., Chen, C.-H. and Chen, C.-H. (2012/07).

Integrative analysis of single nucleotide polymorphisms and gene expression efficiently distinguishes samples from closely related ethnic populations. *BMC Genomics* **13**, 346. SCI.