

PROMOTION RECOMMENDATION

The University of Michigan
School of Public Health
Department of Biostatistics

Hyun Min Kang, assistant professor of biostatistics, Department of Biostatistics, School of Public Health, is recommended for promotion to associate professor of biostatistics, with tenure, Department of Biostatistics, School of Public Health.

Academic Degrees:

Ph.D. (Computer Science)	2009	University of California, San Diego
M.Sc. (Electrical Engineering)	2000	Seoul National University
B.Sc. (Electrical Engineering)	1998	Seoul National University

Professional Record:

2011 – Present	Assistant Professor, Department of Biostatistics, University of Michigan
2009 – 2011	Research Assistant Professor, Department of Biostatistics, University of Michigan

Summary of Evaluation:

Teaching: Professor Kang is an effective, creative and flexible teacher. He has taught a variety of courses in our program. Most recently, he taught our Masters level course on Statistical Inference (BIOSTAT 602, 4 times), which he has gradually modernized to include many continuous evaluation components that have been well received by students. He has also taught Statistical Computing (BIOSTAT 615/815, 3 times). The course was co-listed as a Masters (615) and doctoral (815) level course. The two courses consisted of a shared series of lectures with an additional set of challenging open-ended research problems presented to doctoral level enrollees. Beyond these courses, Professor Kang has taught “Statistical Models and Numerical Methods in Human Genetics” (BIOSTAT 666, once) and “Introduction to Biocomputing” (BIOSTAT 606, also once). Overall, his teaching evaluation scores have been excellent, both for Q1 (“This is an excellent course”) and Q2 (“The instructor is an excellent teacher”).

Professor Kang is currently the dissertation advisor for three doctoral students as well as serving on many dissertation committees. He has co-authored papers with several other doctoral students plus one paper with his current doctoral student. He has also been a key contributor to the first two editions of our Big Data Summer Institute which is aiming to increase our recruitment of under-represented students by providing 30-40 students each summer with an opportunity to participate in research and receive additional training over the summer.

Research: Professor Kang is a world leader in the development of statistical methods and software tools for the analysis of human genetic data. He is extremely creative and his contributions have been important to the field. He has developed statistical methods and related computational approaches that allow geneticists to conveniently model population structure and relatedness in the large genetic association studies that have shaped our view of complex disease genetics, he has developed approaches that enable quality control and variant discovery in large

sequence datasets, and he has developed widely used computational tools. At the same time, he has contributed to the analysis of landmark datasets including some of the largest assemblies of whole genome and whole exome sequence data.

Professor Kang's work first became widely known as the result of a paper entitled "Variance component model to account for sample structure in genome-wide association studies" (Kang et al, *Nature Genetics*, 2010). This paper showed how genetic data could be used to describe relatedness and population structure in a large sample, how these relationships could be summarized in an "empirical relatedness matrix," and how this matrix could be used to define a variance component model that allowed for well calibrated analyses of many genetic association studies, outperforming previous approaches, and proposed efficient strategies for executing the resulting computations. The method opened up a fertile research area and many refinements have now been proposed by researchers. According to Scopus, the paper has received >500 citations and is the 13th most cited paper among the >250 papers published in *Nature Genetics* in 2010. *Nature Genetics* is the top journal in Professor Kang's field and publishes only one to two methodological papers per issue (12 issues per year).

Another example for which Professor Kang is well known is his discovery that low levels (1-3%) of DNA sample contamination were relatively common and were reducing the accuracy of many analyses of DNA sequence data. In two important papers ("Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data," Jun et al, *American Journal of Human Genetics*, 2012; and "Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data," Flickinger et al, *American Journal of Human Genetics*, 2015), Professor Kang described metrics that summarize DNA contamination, methods for estimating these in a variety of settings, and new analysis approaches that are robust to contamination. The metrics are now routinely applied in large scale sequencing projects and contracts, protecting public investments of \$100s of millions of dollars per year in the United States alone.

In addition to his creativity, Professor Kang has produced a steady stream of widely used software tools. Tools like EMMAX (for analysis of genomewide association studies -16,060 page views); EPACTS (for analysis of next generation sequence data - 44,616 page views); VerifyBamID (for detection of DNA sample contamination - 18,777 page views) and GotCloud (for DNA sequence analysis and variant calling - 24,997 page views). These tools are best in class and often introduce and apply important new ideas. For example, GotCloud uses machine learning classifiers to identify genetic variants in large pools of sequence data, where true variants must be distinguished from sequencing errors. Another tool, EPACTS, introduces portable containers for association analysis methods that allows distributed analyses to be conveniently deployed across large compute clusters or clouds. It is safe to predict that manuscripts describing these tools will soon add to Professor Kang's roster of highly cited manuscripts.

Professor Kang has more than 70 published articles (including six as first author and nine as last author). Per Google Scholar, his work has garnered >17,000 citations. This is an impressive metric even when assessed relative to other prolific authors in genomics. His work with the 1000 Genomes Project consortium resulted in the 3rd most cited paper in the journal *Nature* over

the past 10 years (1000 Genomes Project, *Nature*, 2010) and his work on analysis of model organism studies is the 3rd most cited paper in the journal *Genetics* over the past 10 years (Kang et al, *Genetics*, 2008).

Professor Kang is a sought-after collaborator. He helped organize the analysis of key large scale projects, such as the 1000 Genomes Project, and regularly funds 80-90% of his salary as well as several students through collaborative research grants – working with investigators at the University of Michigan’s School of Public Health and Medical School and throughout the world. In addition to numerous projects where he is funded as a co-investigator, Professor Kang has also obtained NIH funding as a principal investigator for an R21 grant; this bodes well for the future and puts him ahead of most faculty at the time of promotion to the associate professor level. Professor Kang has received a terrific score on a new R01 that looks likely to help fund his research for years to come.

Recent and Significant Publications:

- Tan A, Abecasis GR, Kang HM. (2015) “Unified representation of genetic variants.” *Bioinformatics*, 31(13):2202-4.
- Jun G, Wing MK, Abecasis GR, Kang HM. (2015) “An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data.” *Genome Res.* 25(6):918-25.
- Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM. (2015) “Correcting for sample contamination in genotype calling of DNA sequence data.” *Am J Hum Genet*, 97(2):284-90.
- Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. (2012) “Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.” *Am J Hum Genet.* 91(5):839-48.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. (2010) “Variance component model to account for sample structure in genome-wide association studies.” *Nat Genet* 42(4):348-54.

Service: Professor Kang has an excellent record of service to the department. He has generally contributed his time to make the department better, serving on committees that require a substantial commitment of time and effort. In particular, he has served on both our admissions committees (we typically review and interview hundreds of applicants each year and admit only a minority of Masters applicants together with a small minority (10-20%) of doctoral applicants) and on the department candidacy committee which defines and administers the qualifying exam each year.

Professor Kang’s record of service beyond the department is also outstanding. Among many highlights, it is noteworthy that, at the university level, Professor Kang serves on the Advanced Research Computing Advisory Team, which provides advice and direction on university high-performance computing efforts, such as the FLUX computing cluster, and high-performance network and storage resources. His appointment in ARC-TS reflects his deep knowledge and expertise in the area of research super-computing. Also at the university level, Professor Kang organized the Genome Science Training Program’s seminar series (the program funds 10-13 trainees each year and typically hosts four to seven distinguished speakers), and served for

several years on the Graduate Admissions Committee for the Bioinformatics Program in the newly established Department of Computational Medicine and Bioinformatics (DCMB).

He helped organize many international meetings, generally focusing on areas at the interface of molecular biology and computation. For example, he was a member of the program committee for Annual International Conference on Intelligent Systems for Molecular Biology (ISMB, one of the top meetings in the field, 2010 - 2015), for the European Conference on Computational Biology (2015), for the RECOMB Satellite Workshop on Massively Parallel Sequencing (2012), among others. He has contributed to the organization of several other meetings in other roles, including as a reviewer of conference proceedings and abstracts. As expected, his expertise is sought after and Professor Kang is a regular reviewer for top journals.

External Reviewers:

Reviewer A: “I am writing in strong and enthusiastic support for the tenure application for Dr. Hyun Min Kang. It’s important to stress that this work is critical, creative and unique to Dr. Kang. [His] work on linear mixed models...has become the dominant way to handle GWAS data. ... I would be delighted to have him on my faculty and I think Michigan should be equally excited about his tenure.”

Reviewer B: “In a landmark 2010 paper that has 700 citations, Dr. Kang developed an improved computational method that enabled linear mixed models to be applied to large genetic association studies. [He] is also a leader in developing methods for analysis of next-generation sequence data. Dr. Kang introduced the use of [machine learning techniques] for identifying genomic positions which vary between individuals. [He] has made multiple internationally-recognized contributions to statistical, computational and analytical methods for genomic data. I strongly support Dr. Kang’s promotion to associate professor with tenure.”

Reviewer C: “[Professor Kang] is one of the leading scientists [in his cohort] in the field, and his work has made a tremendous impact on the field. He has a spectacular publication record, including 77 publications in first tier journals. His work on the software tool EMMA is the 3rd most cited paper in the journal *Genetics* (one of the main journals in the field). [Professor Kang] has also been very successful in leading his lab, both in terms of mentoring students and in terms of guaranteeing funding. His achievements are way above the bar for the rank of Associate Professor in my institution. I therefore recommend [promotion] with no hesitation.”

Reviewer D: “His grant support is well diversified. Even more remarkable than his grant support is his publication record. Google Scholar currently credits him with 16,206 citations. His curriculum vitae lists between 74 and 77 published papers, depending on how you count. Mentoring of graduate students and postdoctoral fellows exceeds my expectations for an assistant professor. Prof. Kang’s primary contributions are to the genetic modeling and software development. His production of relevant software is amazing. One of the secrets of Prof. Kang’s success is problem selection. He has a knack for finding major computational bottlenecks. He is a superb computer scientist, an excellent geneticist, and a member of one of the most productive teams in statistical genetics internationally. If I were in your shoes, I would speedily approve his promotion.”

Reviewer E: “[Dr. Kang] is one of the top scientists worldwide in statistical and computational genomics, and has my wholehearted support for this promotion. [He] has achieved national and international recognition for his work in statistical and computational genomics. This record is rarely seen so early. The funding success of the statistical genomics group at the University of Michigan is outstanding. In my view, Dr. Kang is essential to this success. He has my complete support for this [promotion], and would be promoted anywhere!”

Reviewer F: “I am writing to express my strongest enthusiasm for the promotion of Hyun Min Kang. [He] has continuous extramural funding, publications, software support, as well as invited presentations at seminars, national and international meetings. He is obviously a good ‘citizen’ through his service on Departmental and School committees. Dr. Kang is a remarkable and critically important contributor to genomics research at the University of Michigan.”

Summary of Recommendation:

Professor Kang is an outstanding biostatistician and is an extremely valuable contributor to the university and the department. He has made creative and key contributions to the approaches, methods and computational techniques and software used to analyze human genetic studies. He is an outstanding researcher, actively engaged in collaboration with many colleagues throughout the university and the world. More importantly, he combines all this great research productivity with an excellent record of teaching and service. It is with the support of the School of Public Health Executive Committee that I recommend Hyun Min Kang for promotion to associate professor of biostatistics, with tenure, Department of Biostatistics, School of Public Health.



Martin A. Philbert, Ph.D.
Dean, School of Public Health

May 2017