Statistical Genomics

ANIM_SCI 545/BIOLOGY 545/CROP_SCI 545/HORT 545/PL_P 545

3 credit, Spring Semester 2021

Professor: Zhiwu Zhang

Office: 105 Johnson Hall

Phone: 509-335-2899

Email: Zhiwu.Zhang@wsu.edu

Teaching Assistant: Zhou Tang

Office: 130 Johnson Hall

Phone: 509-335-4551

Email: zhou.tang@wsu.edu

Lecture room: Zoom (Link provided via email).

Class Website: http://zzlab.net/teaching

Anonymous Feedback Form: https://forms.gle/ExUmS2vB94e3XXW1A

Class schedule: MF 3:10-4:00 PM (Lecture) and W 3:10-5:40 PM (Lab)

Office hour: W 5:40-6:40 PM

Lecture: Pre-recorded lectures will be provided. Lecture time will be used to answer questions and host discussion.

Lab: An assignment will be given, but not graded, for each lab class to enhance the understanding of the theory and help the completion of homework.

Course Objective: Develop concepts and analytical skills for modern breeding by using Genome-Wide Association Study and genomic prediction in a framework of mixed linear models and Bayesian approaches.

Course Description: This is a graduate student course for the concepts and applications of statistical methods and computing tools in genomics. The course includes three sections: Fundamental, Genome Wide Association Study (GWAS) and Genomic Prediction/Selection (GS). The fundamental section covers the essential knowledge and skills of statistics, computer programming (R) and genomics. GWAS and GS sections cover the mechanisms, methods, and computing tools in GWAS and GS, respectively. We start from genotypes and pick up some of them as genes to simulate phenotypes. Then we examine how well we can map the genes and predict the phenotypes starting with very intuitive methods such as correlation and regression. Then we vary relevant factors to evaluate their strength and pitfalls. Methods will progress from basic statistical models and computational tools to include mixed models and Bayesian methods. Students will learn key concepts to guide experimental design, map genes controlling complex traits, and predict their underlying genetic potential among individuals. Analytical skills, critical thinking and hand-on operations are emphasized throughout the course.

Textbook: There is no required textbook. Each lecture will be accompanied by a handout that covers all of the in-class material and more in-depth material that is beyond this course. For

students who would like to have a general reference book, I recommend a free book (academia): Genome-Wide Association Studies and Genomic Prediction

http://link.springer.com/book/10.1007%2F978-1-62703-447-0

Prerequisites: General linear model, mixed linear model, Bayesian theory, computer programming in R, genetics, or permission by instructor.

Assessments: Homework (50%), attendance (5%), participation (5%), pre-class quizzes (10%), midterm exam (10%), and final exam (20%).

Homework: Most of homework require to propose hypotheses, prove or disapprove the hypotheses by analyzing data, and write final reports.

Attendance: Attendance in each lecture and lab is expected with camera on. Webcams may be rented through the University (https://scheduling.wsu.edu/Content/equipment.aspx). In accordance with Academic Regulation 73, absences impede a student's academic progress and should be avoided. Those students who must miss a lecture for illness or university-sponsored activities such as field trips, judging teams, sports, conferences, etc. should obtain an official Class Absence Request form from the doctor, or faculty/staff member supervising the off-campus activities. Scheduling conflicts with employment and non-university activities will be considered unexcused absences.

Participation: Students are expected to participate in class discussions. Both questions and answers count as participation.

Pre-Class Quizzes: Students are expected to watch a pre-recorded lecture and complete a comprehension quiz before each synchronous lecture. Quizzes are due at the start of class time.

Exams: Midterm takes 1.5 hours. Final exam is cumulative and takes 3 hours. All exams are synchronous.

Late Policy: The total points for late homework will decrease by 50% per late day unless the delay is due to an excused absence. Late quizzes are not accepted.

Grade Scale: A (93%-100%); A- (90%-93%); B+ (87%-90%); B (83%-87%) B- (80%-83%); C+ (77%-80%); C (73%-77%); C- (70%-73%) D+ (66%-70%); D (60%-66%); F(0%-60%). Note: The upper grade will be assigned to a score without rounding. For examples, a score of 93.0% receives "A" and a score of 92.9% receives "A-".

Student Learning Outcomes: Upon completion of the course, students will be able to:

1) Apply quantitative and scientific reasoning to solve problems in statistical genomics;
2) Understand the development of the statistical methods for gene mapping, molecular breeding and health management;
3) Integrate concepts, principles, methods, and skills in statistics, genetics and computer programming to conduct in a variety of genomic research;
4) Communicate effectively using emerging graphics and graphic media.

All the outcomes will be evaluated by the four assessments (participant, midterm exam, final exam and homework).

WSU Work Statement: For each hour of lecture, students should expect to invest a minimum of two hours of work outside class.

WSU Safety Statement: Classroom and campus safety are of paramount importance at Washington State University, and are the shared responsibility of the entire campus population. WSU urges students to follow the "Alert, Assess, Act" protocol for all types of emergencies and the "Run, Hide, Fight" response for an active shooter incident. Remain ALERT (through direct observation or emergency notification), ASSESS your specific situation, and ACT in the most appropriate way to assure your own safety (and the safety of others if you are able).

WSU Disability Statement: Reasonable accommodations are available for students with a documented disability. If a student has a disability and may need accommodations to fully participate in this class, the student should either visit or call the Access Center (Washington Building 217; 509–335–3417) to schedule an appointment with an Access Advisor. All accommodations MUST be approved through the Access Center.

WSU Academic Honesty Statement: As an institution of higher education, Washington State University is committed to principles of truth and academic honesty. All members of the University community share the responsibility for maintaining and supporting these principles. When a student enrolls in Washington State University, the student assumes an obligation to pursue academic endeavors in a manner consistent with the standards of academic integrity adopted by the University. To maintain the academic integrity of the community, the University cannot tolerate acts of academic dishonesty including any forms of cheating, plagiarism, or fabrication. Academic integrity is the cornerstone of the university and will be strongly enforced in this course. Any student caught cheating on any assignment or exam will be given an F grade for the course, will not have the option to withdraw from the course, and will be reported to the Office of Student Standards and Accountability. Cheating is defined in the Standards for Student Conduct WAC 504-26-010 (3). It is strongly suggested that you read and understand these definitions: http://apps.leg.wa.gov/WAC/default.aspx?cite=504-26-010.

Campus Resources

- Graduate Writing Center, https://writingprogram.wsu.edu/graduate-writing-center
- Library Services, http://www.wsulibs.wsu.edu/
- CACD, Center for Advising and Career Development, https://ascc.wsu.edu
- Office of Student Conduct, http://conduct.wsu.edu
- Counseling and Testing Services, http://counsel.wsu.edu/
- Academic Integrity, http://academicintegrity.wsu.edu

Statistical Genomics

(Lecture on Mondays and Fridays)

No.	Date	Section	Title	HW Due
1	1/20/21	Fundamental	Syllabus/course overview and introduction and R	
2	1/22/21		Random variables and distribution	
3	1/25/21		Statistical inference	
4	1/29/21		Linear algebra[1]	
5	2/1/21		Genotyping By Sequencing (GBS)[2]	
6	2/5/21		Missing genotype imputation[3]	
7	2/8/21		Phenotype simulation	HW1
8	2/12/21		Linkage analysis	
9	2/15/21		Linkage disequilibrium	
10	2/19/21	GWAS	GWAS by correlation	
11	2/22/21		Power, type I error and False Discovery Rate	
12	2/26/21		Population structure and PCA	
13	3/1/20		General Linear Model (GLM)	HW2
14	3/5/20		Kinship	
15	3/8/20		Mixed Linear Model (MLM)[4]	Midterm
16	3/12/20		Efficient Mixed Model Association (EMMA)[5]	
17	3/15/20		Compressed MLM[6]	HW3
18	3/19/20		SUPER GWAS method[8,9]	
19	3/22/20		Multiple Loci Mixed Model (MLMM)[10]	
20	3/26/20		FarmCPU	
21	3/29/20		BLINK	HW4
22	4/2/20	GS	Marker Assisted Selection (MAS)	
23	4/5/20		Model fit and cross validation accuracy	
24	4/9/20		gBLUP[11,12]	
25	4/12/20		Ridge regression (rrBLUP)	HW5
26	4/16/20		Bayesian theory	
27	4/19/20		Bayesian methods[13]	
28	4/23/20		Bayesian tools	
29	4/26/20		BLUP alphabet	
30	4/30/21		Machine learning	HW6

Statistical Genomics

(Lab on Wednesdays)

Lab	Date	Section	Title	Remark
1	1/20/21	Fundamental	R and Documentation	
2	1/27/21		Distribution and Statistical inference	
3	2/3/21		Linear algebra[1]	
4	2/10/21		Missing genotype imputation[3]	
5	2/17/21		Phenotype simulation and GWAS by correlation	
6	2/24/21	GWAS	Power, type I error and False Discovery Rate	
7	3/3/21		PCA and General Linear Model (GLM)	
8	3/10/21		Mixed Linear Model (MLM)[4]	
9	3/17/21		No class (class holiday)	
10	3/24/21		EMMA, CMLM, and MLMM[10]	
11	3/31/21		FarmCPU and BLINK	
12	4/7/21	GS	Model fit and cross validation accuracy	
13	4/14/21		MAS and gBLUP[11,12]	
14	4/21/21		Ridge regression (rrBLUP)	
15	4/28/21		Bayesian methods[13]	

Reference

1. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. *Genetics and analysis of quantitative traits.* (1998).

2. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6,** e19379 (2011).

3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11,** 499–511 (2010).

4. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).

5. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178,** 1709–1723 (2008).

6. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42,** 355–360 (2010).

7. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42,** 348–354 (2010).

8. Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9,** e107684 (2014).

9. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8,** 833–835 (2011).

10. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* **44,** 825–830 (2012).

11. Zhang, Z., Todhunter, R. J., Buckler, E. S. & Van Vleck, L. D. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.* **85,** 881–885 (2007).

12. VanRaden, P. M. Efficient methods to compute genomic predictions. *J Dairy Sci* **91,** 4414–4423 (2008).

13. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157,** 1819–1829 (2001).