Statistical Genomics

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

3 credit, Spring Semester 2025

Professor: Zhiwu Zhang
Office: 403C Plant Science Building
Phone: 509-335-2899
Email: Zhiwu.Zhang@wsu.edu

Teaching Assistant: Meijing Liang
Office: 403 Plant Science Building
Phone: 509-338-5185
Email: Meijing.Liang@wsu.edu

Class room: Wilson Short 5 (Lecture) and Clark 149 (Lab)

Class schedule: MF 3:10-4:00 PM (Lecture) and W 3:10-5:40 PM (Lab)
ZOOM (Lec): https://wsu.zoom.us/j/91745964556?pwd=1kVAMq9YCaro9diYkAnBaqRpRX4nZt.1
ZOOM (Lab): https://wsu.zoom.us/j/98345956521?pwd=4fLGAnmsjdXzN5oCEgSvLxpS1fokjD.1

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

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

Office hour: Upon Request

Lecture: To explain why and how the concepts work in statistical genomics.

Lab: to enhance the understanding of the theory lecture 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: Statistical inference, General linear model, mixed linear model, Bayesian theory, computer programming in R, genetics, or permission by instructor.

Assessments: Attendance (10%), Participation (10% by student, 10% by TA, and 10% by instructor), and homework (60%).

Homework: Most of homework require to propose hypotheses, prove or disapprove the hypotheses by analyzing data, and write final reports. Teamwork is encouraged for debugging and discussion on solving strategies and result interpretation. Homework will be evaluated by TA and instructor.

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. Attendance will be recorded by TA.

Participation: Students are expected to participate in class discussions. Both questions and answers count as participation. Participation will be evaluated by students, TA and instructor.

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.

Use of generative AI

It is important to remember that the unauthorized use of generative AI in a course is a violation of WSU's Community Standards and can be reported to the Center for Community Standards (CCS). As a reminder, any expectations about AI usage should be communicated in writing. WSU recognizes that additional policies may be preferred by some instructors, including explicit encouragement of students using AI. Instructors can consider them as starting points in the exploration of what AI policy will be. <u>This course chooses option D</u>.

A. AI Use prohibited
Students are not allowed to use advanced automated tools (artificial intelligence or machine learning tools such as ChatGPT, Co-Pilot, or Dall-E) on assignments in this course. Each student is expected to complete each assignment without substantive assistance from others, including automated tools.

B. AI Use only with prior permission
Students are allowed to use advanced automated tools (artificial intelligence or machine learning tools such as ChatGPT, Co-Pilot, or Dall-E) on assignments in this course if instructor permission is obtained in advance. Unless given permission to use those tools, each student is expected to complete each assignment without substantive assistance from others, including automated tools.

C. AI Use only with acknowledgement
Students are allowed to use advanced automated tools (artificial intelligence or machine learning tools such as ChatGPT, Co-Pilot, or Dall-E) on assignments in this course if that use is properly documented and credited. For example, text generated using ChatGPT-3 should include a citation such as: "Chat-GPT-3. (YYYY, Month DD of query). "Text of your query." Generated using OpenAI. https://chat.openai.com/" Material generated using other tools should follow a similar citation convention.

D. AI Use is freely permitted with no acknowledgement
Students are allowed to use advanced automated tools (artificial intelligence or machine learning tools such as ChatGPT, Co-Pilot, or Dall-E) on assignments in this course; no special documentation or citation is required.

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	Remark
1	1/6/25	Fundamental	Syllabus and course overview	
2	1/10/25		Random variables and distribution	
3	1/13/25		Statistical inference	
4	1/17/25		Linear algebra1	
5	1/24/25		Phenotype simulation	
6	1/27/25	GWAS	Correlation test and multiple tests	
7	1/31/25		Linkage analysis and linkage disequilibrium	
8	2/3/25		Power, type I error and False Discovery Rate	
9	2/7/25		Population structure and PCA	
10	2/10/25		General Linear Model (GLM)	
11	2/14/25		Kinship	
12	2/17/25		Mixed Linear Model (MLM)4	
13	2/21/25		Efficient Mixed Model Association (EMMA)5	
14	2/24/25		Compressed MLM6	HW1
15	2/28/25		SUPER GWAS method8,9	
16	3/3/25		Multiple Loci Mixed Model (MLMM)10	
17	3/7/25		FarmCPU	
18	3/17/25		BLINK	
19	3/21/25	GS	MAS and gBLUP11,12	
20	3/24/25		Ridge regression (rrBLUP)	
21	3/28/25		Model fit and cross validation accuracy	
22	3/31/25		Bayesian theory	HW2
23	4/4/25		Bayesian methods13	
24	4/7/25		Bayesian tools	
25	4/11/25		BLUP alphabet	
26	4/14/25		Minning the Maximum Accuracy of Prediction	
27	4/18/25		Machine learning	
28	4/21/25		Deep learning for GWAS	
29	4/25/25		Deep learning for GS	HW3 (May 2)

Statistical Genomics

(Lab on Wednesdays)

Lab	Date	Section	Title	Remark
1	1/8/25	Fundamental	R and Documentation	
2	1/15/25		Distribution and Statistical inference	
3	1/22/25		Linear algebra[1]	
4	1/29/25		Phenotype simulation and GWAS	
5	2/5/25		GWAS by correlation	
6	2/12/25	GWAS	Power, type I error and False Discovery Rate	
7	2/19/25		PCA and General Linear Model (GLM)	
8	2/26/25		Mixed Linear Model (MLM)[4]	
9	3/5/25		EMMA, CMLM, and MLMM[10]	
10	3/19/25		FarmCPU and BLINK	
11	3/26/25		Model fit and cross validation accuracy	
12	4/2/25	GS	MAS and gBLUP[11,12]	
13	4/9/25		Ridge regression (rrBLUP)	
14	4/16/25		Bayesian methods[13]	
15	4/23/25		MMAP	

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).