

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=4fLGAnmsjdXzN5oCEgSvLxpSIfoKjD.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 graduate-level course explores the concepts and applications of statistical methods and computational tools in genomics. The course is divided into three main sections: Fundamentals, Genome-Wide Association Studies (GWAS), and Genomic Prediction/Selection (GS). The Fundamentals section introduces essential knowledge and skills in statistics, computer programming (using R), and genomics. The GWAS and GS sections delve into the mechanisms, methodologies, and computational tools specific to each area. The course begins with genotypes, selecting some as genes to simulate phenotypes. We then assess the effectiveness of mapping these genes and predicting phenotypes, starting with intuitive approaches like correlation and regression. Subsequently, we manipulate relevant factors to evaluate their impact and limitations. The methods evolve from basic statistical models and computational tools to more advanced techniques, including mixed models and Bayesian approaches. Students will acquire key concepts to inform experimental design, map genes influencing complex traits, and predict individuals' underlying genetic potential. The course emphasizes analytical skills, critical thinking, and hands-on experience throughout.

Textbook: This course does not require a textbook. Each lecture will be accompanied by a handout that includes all in-class material, as well as additional in-depth content that extends

beyond the scope of the course. For students interested in a general reference book, I recommend a freely available resource from Academia: Genome-Wide Association Studies and Genomic Prediction¹.

<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. This course chooses option D.

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 algebra ¹	
5	1/24/25		Phenotype simulation ^{2,3}	
6	1/27/25	GWAS	Correlation test and multiple tests ^{4,5}	
7	1/31/25		Linkage analysis and linkage disequilibrium ⁶	
8	2/3/25		Power, type I error and False Discovery Rate ⁷	
9	2/7/25		Population structure and PCA ^{8,9}	
10	2/10/25		General Linear Model (GLM)	
11	2/14/25		Kinship ¹⁰	
12	2/17/25		Mixed Linear Model (MLM) ¹¹	
13	2/21/25		Efficient Mixed Model Association (EMMA) ¹²	
14	2/24/25		Compressed MLM ¹³	HW1
15	2/28/25		SUPER GWAS method ^{14,15}	
16	3/3/25		Multiple Loci Mixed Model (MLMM) ¹⁶	
17	3/7/25		FarmCPU ¹⁷	
18	3/17/25		BLINK ¹⁸	
19	3/21/25	GS	MAS and gBLUP ¹⁹⁻²²	
20	3/24/25		Ridge regression (rrBLUP) ^{23,24}	
21	3/28/25		Model fit and cross validation accuracy ²⁵	
22	3/31/25		Bayesian theory	HW2
23	4/4/25		Bayesian methods ²³	
24	4/7/25		Bayesian tools ²⁶	
25	4/11/25		BLUP alphabet ²⁷⁻²⁹	
26	4/14/25		Minning the Maximum Accuracy of Prediction ³⁰	
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	
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)	
9	3/5/25		EMMA, CMLM, and MLMM	
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	
13	4/9/25		Ridge regression (rrBLUP)	
14	4/16/25		Bayesian methods	
15	4/23/25		MMAP	

Reference

1. Gondro, C., der Werf, J. & Hayes, B. *Genome-Wide Association Studies and Genomic Prediction*. vol. 1019 (Springer, 2013).
2. Tang, Y. & Liu, X. G2P: a genome-wide-association-study simulation tool for genotype simulation, phenotype simulation and power evaluation. *Bioinformatics* **35**, 3852–3854 (2019).
3. Fernandes, S. B. & Lipka, A. E. simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic phenotypes. *BMC Bioinformatics* **21**, 1–10 (2020).
4. Perneger, T. V. What’s wrong with Bonferroni adjustments. *Bmj* **316**, 1236–1238 (1998).
5. Armstrong, R. A. When to use the Bonferroni correction. *Ophthalmic and physiological optics* **34**, 502–508 (2014).
6. Lynch, M., Walsh, B. & others. *Genetics and Analysis of Quantitative Traits*. vol. 1 (Sinauer Sunderland, MA, 1998).
7. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
8. Zhao, K. *et al.* An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
9. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
10. Hardy, O. J. & Vekemans, X. spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**, 618–620 (2002).
11. 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).
12. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
13. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355–360 (2010).
14. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* vol. 8 833–835 Preprint at <https://doi.org/10.1038/nmeth.1681> (2011).
15. Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS One* **9**, (2014).
16. Segura, V. *et al.* An Efficient Multi-Locus Mixed-Model Approach for Genome-Wide Association Studies in Structured Populations. *Nature Genetics* vol. 44 825–830 (2012).
17. Kusmec, A. & Schnable, P. S. FarmCPUpp: Efficient large-scale genomewide association studies. *Plant Direct* **2**, 1–6 (2018).
18. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* **8**, (2019).
19. C.Y, C. B. & J, M. D. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 557–572 (2008).
20. Bernardo, R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci* **34**, 20–25 (1994).
21. VanRaden, P. M. Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414–4423 (2008).

22. 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).
23. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819 (2001).
24. Endelman, J. B. Ridge regression and other kernels for genomic selection in the R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
25. Zhou, Y., Isabel Vales, M., Wang, A. & Zhang, Z. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. *Briefings in Bioinformatics* (2016).
26. Pérez, P. & De Los Campos, G. BGLR: A Statistical Package for Whole Genome Regression and Prediction. <https://cran.r-project.org/web/packages/BGLR/vignettes/BGLR-extdoc.pdf> Version 1.0.8 (2008).
27. Zhang, Z. *et al.* Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. **9**, 1–12 (2014).
28. Forni, S., Aguilar, I. & Misztal, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* **43**, 1 (2011).
29. Wang, J. *et al.* Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity (Edinb)* **121**, 648–662 (2018).
30. Wei Huang *et al.* MMAP : a cloud computing platform for mining the maximum accuracy of predicting phenotypes from genotypes. *BIOINFORMATICS btaa824*, 11–14 (2020).

