The analysis of complex traits in livestock using dense genomic information - tools and applications

M.Sc. Claas Heuer 1. Berichterstatter: Prof. Dr. Georg Thaller

Die Verfügbarkeit von zehntausenden Markern basierend auf Einzelbasenaustauschen (*single nucleotide polymorphisms*, SNP) hat zu einem Paradigmenwechsel in der Tierzucht geführt und eröffneten neue Möglichkeiten für die Forschung und Zuchtpraxis. Das Ziel dieser Dissertation war es, neue statistische Methoden und Rechenwerkzeuge für die Auswertung komplexer Merkmale zu entwickeln und diese auf Basis von hochdichten Markerdaten anzuwenden. Ein aktueller Forschungsschwerpunkt stellt die Aufteilung genetischer Varianz in additive und nicht additive Komponenten dar. Es wurde eine Methode entwickelt, welche es erlaubt alle phänotypisierten Tiere in einer Population zur Schätzung von Dominanzeffekten an Markern heranzuziehen, indem Genotypwahrscheinlichkeiten für diese Tiere berechnet werden. Hierzu wurde ein Datensatz von ca. 470.000 deutschen Holstein Kühen verwendet, für die Erstlaktationsdaten für Milchleistungsmerkmale zu Verfügung standen und von denen genotypisierte Väter und Großväter bekannt waren. In einer genomweiten Assoziationsstudie konnten signifikante Dominanzeffekte für die Merkmale Milch kg, Fett kg und Eiweiß kg auf verschiedenen Chromosomen geschätzt werden

Die grundlegende Annahme solcher genomweiten Assoziationsstudien ist es, dass die verwendeten Marker im Kopplungsungleichgewicht (*linkage disequilibrium*, LD) mit sogenannten *quantitative trait loci* (QTL) stehen. Das Ausmaß des LD kann variieren und daher wurde untersucht, welchen Einfluss unvollständiges LD auf schätzbare QTL-Effekte an Markern hat. Mittels Simulationen und deterministischen Berechnungen konnte gezeigt werden, dass unvollständiges LD zwischen Marker und QTL zu Verzerrung der schätzbaren additiven Effekte am Marker führt, falls Dominanz am QTL vorliegt. Um diesen Effekt zu quantifizieren, wurden Formeln abgeleitet, welche die Berechnung der schätzbaren genetischen Effekte an Markern bei gegebenen LD und QTL-Effekten erlauben.

Die fortschreitende Genotypisierung von Zuchttieren und das Imputieren nicht genotypisierter Tiere erzeugt rechentechnische Herausforderungen. Ein möglicher Lösungsansatz stellt die Verwendung von parallelen Rechenmethoden dar. Es wurde ein Programmpaket für die Statistikumgebung R entwickelt, das es erlaubt mehrere Rechenkerne für genomweite Assoziatonsstudien, genomische Leistungsvorhersage und gemischte lineare Modelle mittels Gibbs Sampling zu verwenden. Um die Eigenschaften des Pakets zu analysieren, wurden Datensätze unterschiedlicher Größe und Struktur simuliert und die Anzahl der verwendeten Rechenkerne variiert. Grundsätzlich hängt die Verringerung der Rechenzeit durch die Verwendung mehrerer Rechenkerne sehr stark von der Größe des Datensatzes ab. Für 10,000 Marker und eine Million Beobachtungen, ergab sich für ein Ridge Regression Markermodell eine Verringerung um den Faktor acht, wenn die Anzahl der Rechenkerne von eins auf über zwölf erhöht wurde. Die absolute Rechenzeit betrug 17 Stunden bei 30.000 Iterationen des Gibbs Samplers. Die Rechenzeiten bei der Berechnung einer genomischen Verwandtschaftsmatrix für 2.000 Tiere und 50.000 Markern, sowie bei Kreuzvalidierung und genomweiten Assoziationsstudien verringerte sich linear mit der Anzahl der Rechenkerne.

Viele Merkmale von tierzüchterischem Interesse sind nicht kontinuierlich verteilt. Sofern die verschiedenen Klassen eines solchen Merkmals hierarchisch angeordnet werden können, sind ordinale Schwellenwertmodelle die Methode der Wahl, andernfalls müssen multinomiale Modelle verwendet werden. Ein Konzept zur genomischen Vorhersage von nicht geordneten kategoriellen Merkmalen wurde entwickelt und dieses auf die Vorhersage von Subpopulationszuweisungen in deutschen Warmblut Pferderassen angewendet. Dem multinomialen Problem wurde durch paarweise binäre Kontraste mittels Schwellenwertmodellen und Support Vector Machines begegnet. Multinomiale Vorhersagewahrscheinlichkeiten aus beiden Schemata wurden durch Normalisieren oder *pairwise coupling* aus den binären Wahrscheinlichkeiten berechnet. Für jedes Tier wurde diejenige Klasse zugewiesen, welche die höchste Vorhersagewahrscheinlichkeit hatte. In einer *leave-one-out* Kreuzvalidierung unter Verwendung von 917 Hengsten aus den Rassen Hannoveraner, Holsteiner, Oldenburger und Trakehner, wurde die Zuweisung zu den Subpopulationen genomisch vorhergesagt. Die Genauigkeit der Zuordnung wurde als der Quotient zwischen korrekten Klassifizierungen und allen Klassifizierungen berechnet und lag im Gesamtdatensatz zwischen 0,74 und 0,84. Support Vector Machines erreichten dabei ähnliche Vorhersagegenauigkeiten wie Ridge Regression Schwellenwertmodelle.