Nutzung der Imputation für den Übergang von der Mikrosatelliten-basierten Abstammungsüberprüfung zur SNP-Genotypisierung

W. Nolte[1], E. Kalm[2], N. Krattenmacher[2], S. Lehner[3], R. Reents[4], K.F. Stock[4,5], J. Tetens[6], G. Thaller[2], M. von Depka Prondzinski[3], S. Vosgerau[2], M. Wobbe[4,5], C. Kühn[1,7]

[1]Leibniz-Institut für Nutztierbiologie (FBN), Institut für Genombiologie, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, [2]Christian-Albrechts-Universität Kiel, Institut für Tierzucht und Tierhaltung, 24098 Kiel, [3]Werlhof-Institut MVZ, 30159 Hannover, [4]Vereinigte Informationssysteme Tierhaltung w.V. (vit), 27283 Verden (Aller), [5]Stiftung Tierärztliche Hochschule Hannover, Institut für Tierzucht und Vererbungsforschung, 30559 Hannover, [6]Georg-August-Universität Göttingen, Department für Nutztierwissenschaften, 37077 Göttingen, [7]Universität Rostock, Agrar- und Umweltwissenschaftliche Fakultät, 18059 Rostock

Einleitung

Seit über zwei Jahrzehnten wird die Abstammungskontrolle für Pferde in Deutschland über das von der ISAG (International Society for Animal Genetics, https://www.isag.us/) empfohlene Abstammungspanel mit Mikrosatelliten durchgeführt (Bowling *et al.*, 1997). Im Rahmen der geplanten Einführung der genomischen Selektion beim deutschen Warmblutpferd durch die 2017 gegründete International Association for Future Horse Breeding (IAFH) mit den Gründungsverbänden Verband der Züchter des Oldenburger Pferdes e.V. (OL), Springpferdezuchtverband Oldenburg-International e.V. (OS), Westfälisches Pferdestammbuch e.V. (WESTF), Trakehner Verband e.V. (TRAK), Verband der Züchter des Holsteiner Pferdes e.V. (HOL) soll in naher Zukunft die Umstellung auf ein SNP-basiertes (Single Nucleotide Polymorphism) Panel erfolgen. Der Vorteil der Nutzung eines SNP-Chips besteht im vielfältigen Anwendungsspektrum, weil die Informationen unter der Prämisse einer entsprechenden Anzahl und Auswahl der SNP-Marker, neben der Abstammungskontrolle auch für die Ermittlung genomischer Zuchtwerte für verschiedene, züchterisch relevante Merkmale genutzt werden können. Im Vergleich zu Mikrosatelliten weisen SNPs eine bessere Standardisierbarkeit bei der Genotypisierung und eine geringere Mutationsrate auf. Sie zeichnen sich gleichzeitig durch geringere Kosten pro Genotyp, einen schnelleren Durchsatz bei der Datengewinnung und einen leichter zu automatisierenden Prozess im Labor aus (McClure *et al.*, 2013).

Bei der Abstammungskontrolle ist es unerlässlich, dass Elterntiere und Fohlen auf derselben Informationsebene genotypisiert sind, d.h. dass für beide Generationen entweder Mikrosatelliten- oder SNP-Daten vorliegen müssen. Um Doppeltypisierungen zu vermeiden und damit verbundenen Kosten-, Zeit- und Arbeitsaufwand zu vermeiden, sollen Fohlen zukünftig nur mit einem SNP-Chip genotypisiert werden. Dies erfordert, dass Mikrosatelliteninformationen durch Imputation ergänzt werden, um einen Abgleich mit der Elterngeneration zu ermöglichen.
Bei der Imputation wird auf Basis von SNP-Daten statistisch ermittelt, welche Mikrosatellitengenotypen ein Tier höchstwahrscheinlich trägt. Zu diesem Zweck ist das Anlegen einer sowohl für SNP als auch Mikrosatelliten genotypisierten, umfangreichen Lernstichprobe nötig. Beim Rind wurde der Übergang vom alten Mikrosatellitensystem zu einem SNP-Panel so bereits erfolgreich mit 98-prozentiger Genauigkeit demonstriert (McClure *et al.*, 2012) und umgesetzt.
Ziel ist es, nun auch beim Pferd eine Imputation von Mikrosatellitendaten mit höchstmöglicher Genauigkeit zu erreichen. Um die Fehlerquote abzuschätzen und nachfolgend in der Praxis zu minimieren, wurden im Folgenden drei Optionen getestet: (A) eine Imputation jeweils innerhalb der fünf genannten Rassen, (B) eine Imputation über Rassen hinweg, und (C) eine Imputation innerhalb von Kohorten großer genetischer Ähnlichkeit, die durch eine SNP-basierte Hauptkomponentenanalyse bestimmt wurden. Schlussendlich soll das Szenario mit der geringsten Fehlerquote für die endgültige Imputation und somit Anwendung in der Praxis ausgewählt werden.

Material und Methoden

Die erstellte Lernstichprobe beinhaltete bis dato 1.984 Stuten, die auf SNP-Chips mittlerer Dichte (GGP Equine 70k und GGP Equine Plus Beadchip, Neogen / Illumina) genotypisiert wurden. Nach Anwendung von Qualitäts-filtern für Hardy-Weinberg-Gleichgewicht ($P > 0{,}001$), Call Frequency je SNP-Marker (90 Prozent), Genotypisierungsrate je Tier (95 Prozent), GC-Score (0,6) sowie einer Minor Allele Frequency (MAF, 1 Prozent) verblieben 1.745 Stuten (436 HOL, 716 OL, 140 OS, 278 TRAK, 175 WESTF) mit Genotypdaten für je 59.933 SNP-Marker und zusätzlich für 17 Mikrosatelliten (AHT4, AHT5, ASB17, ASB2, ASB23, CA425, HMS1, HMS2, HMS3, HMS6, HMS7, HTG10, HTG4, HTG6, HTG7, LEX3, VHL20) im Auswertungsdatensatz. Nach der händischen Erstellung einer VCF-Datei für Mikrosatellitengenotypdaten wurde der Datensatz mit den SNP-Genotypisierungen fusioniert.

Zur Prüfung der Imputationsgenauigkeit wurden 10 Prozent der Tiere jeder Rasse randomisiert ausgewählt und für die Mikrosatellitendaten maskiert, d.h. auf unbekannt gesetzt. Mit Beagle 5.0 (Browning *et al.*, 2018) wurden die fehlenden Mikrosatelliten- (und SNP-) Genotypen dann imputiert und mit den Originaldaten verglichen. Der Randomisierungs- und Maskierungsprozess wurde 10-fach wiederholt, wobei Tiere wiederholt in der randomisierten Auswahl der zehn Runden vorkommen konnten. Die drei oben genannten Imputationsoptionen wurden hinsichtlich ihrer Genauigkeit verglichen:

(A) Imputation innerhalb der jeweiligen Rasse

(B) Imputation über Rassen hinweg, d.h. alle 1.745 Stuten gemeinsam

(C) Imputation innerhalb von 3 Clustergruppen (siehe Abbildung 1), die mit der Funktion kmeans in R nach einer SNP-basierten Hauptkomponentenanalyse, basierend auf einer genomischen Verwandtschaftsmatrix, in GCTA (Yang *et al.*, 2011) ermittelt wurden.

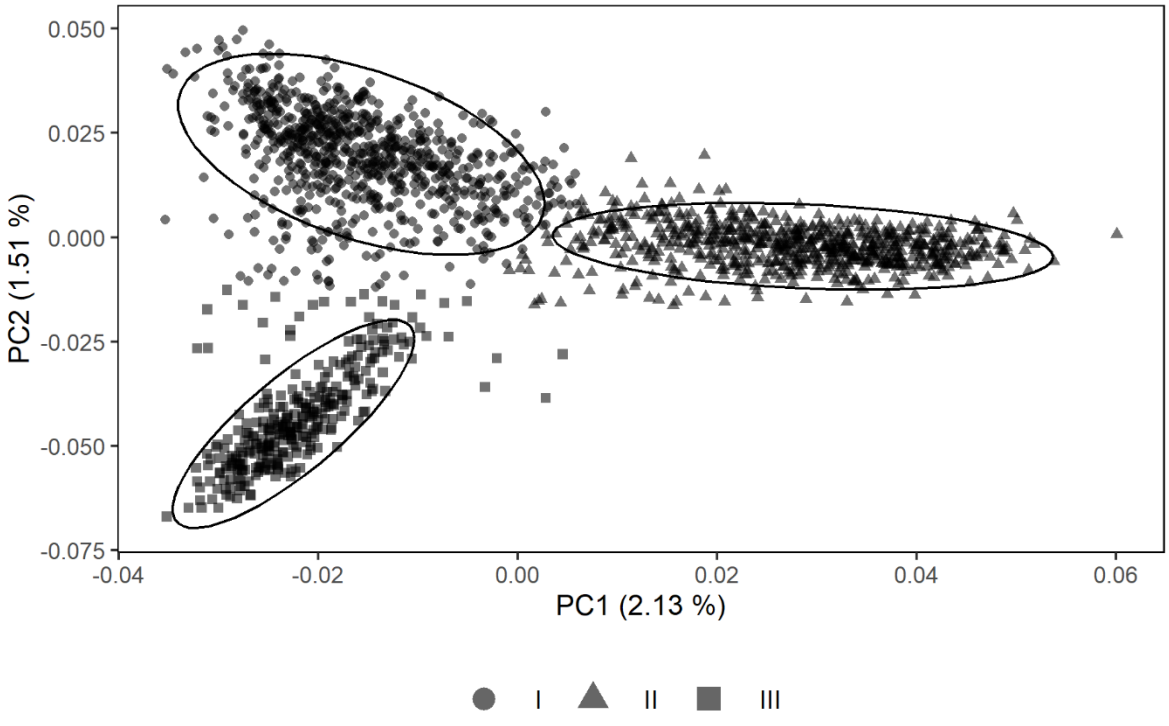

Abbildung 1: Bildung dreier Clustergruppen nach genotypischer Ähnlichkeit, nach einer Hauptkomponentenanalyse basierend auf SNP-Daten von 1.745 Stuten aus 5 Warmblutrassen.

Um eine Vergleichbarkeit der drei Optionen innerhalb einer Testrunde zu ermöglichen, wurden immer dieselben Tiere je Runde maskiert und imputiert. Für die Genauigkeitsprüfung wurde für jedes Tier für jeden Mikrosatelliten eine Punktzahl vergeben: 0 (beide Allele falsch imputiert im Vergleich zum Original), 0,5 (ein Allele korrekt imputiert) und 1 (beide Allele korrekt imputiert). Im Anschluss wurden durchschnittliche Genauigkeitswerte je Runde, Marker und Rasse ermittelt.

Ergebnisse und Diskussion

Vergleich der Optionen A, B und C

Der Vergleich der Imputation nach den drei Optionen (A, B, C) ergab, dass sich die höchste Genauigkeit von 99,37 Prozent bei einer Imputation innerhalb der fünf Rassen erzielen lässt (Option A). Die Optionen B und C wiesen mit 0,41 und 0,63 Prozentpunkten weniger jeweils eine minimal geringere Genauigkeit auf (siehe Tabelle 1).

Tabelle 1: Durchschnittliche Genauigkeit der Mikrosatellitenimputation anhand von SNPs je Option

Option	Methode	Genauigkeit (%)
A	innerhalb der Rassen	99,37
B	alle Rassen zusammen	98,96
C	Clustergruppen der PCA	98,74

Weiterhin ließ sich beobachten, dass jede der fünf Rassen durchgängig in Option A das beste Imputationsergebnis erzielte (siehe Tabelle 2). Die Rassen OS und WESTF erzielten im Rassevergleich in den drei Optionen jeweils niedrigere Imputationsgenauigkeiten, was vermutlich auf die geringere Stichprobengröße zurückzuführen ist (140 OS, 175 WESTF). Die Fehlerquote lag bei OL als mit 716 Stuten zahlmäßig am stärksten vertretenem Verband durchweg unter der Einprozentmarke.

Tabelle 2: Genauigkeiten (%) der Mikrosatellitenimputation anhand von SNPs je Option und Rasse

Option	HOL	OL	OS	TRAK	WESTF	alle
A	99,67	99,39	98,87	99,62	98,51	99,37
B	98,37	99,38	98,76	99,61	97,88	98,96
C	99,46	99,17	97,84	99,61	97,61	98,74

Die geringsten Fehlerquoten ergaben sich für TRAK und damit die Rasse, die zwar hinsichtlich der Stichprobengröße zwischen OL und WEST

rangierte, sich aber als klassische Reinzucht deutlich von den anderen Rassen absetzte. Dies spiegelte sich auch in der Hauptkomponentengrafik wider. So bestand die Clustergruppe III (siehe Abbildung 1) nahezu ausschließlich aus TRAK, wohingegen sich die Clustergruppen I und II in der Mehrzahl aus Tieren verschiedener Rassen zusammensetzen.

Weitere Datenanalyse Option A

Aus dem Vergleich der drei Optionen ergab sich, dass die höchste Imputationsgenauigkeit in Option A erreicht werden konnte. Ein Vergleich der zehn Testrunden zeigte, dass Runde 1 bezüglich der Genauigkeit deutlich niedriger lag als die anderen Runden (siehe Abbildung 2). Ein Trainingseffekt des Imputationsprogramms durch Wiederholung konnte durch weitere Tests ausgeschlossen werden. Weiterhin ergaben zusätzliche Untersuchungen, dass sich die Reihenfolge der Runden nicht auf deren Genauigkeit auswirkte, sondern die (randomisierte) Tierauswahl selbst ausschlaggebend war. Runde 1 beinhaltete demnach in jeder Rasse überdurchschnittlich viele Stuten mit seltenen Genotypen, deren Imputation erschwert war, wenn keine oder sehr wenige Tiere mit demselben Genotyp unmaskiert in der Referenzgruppe vorlagen.

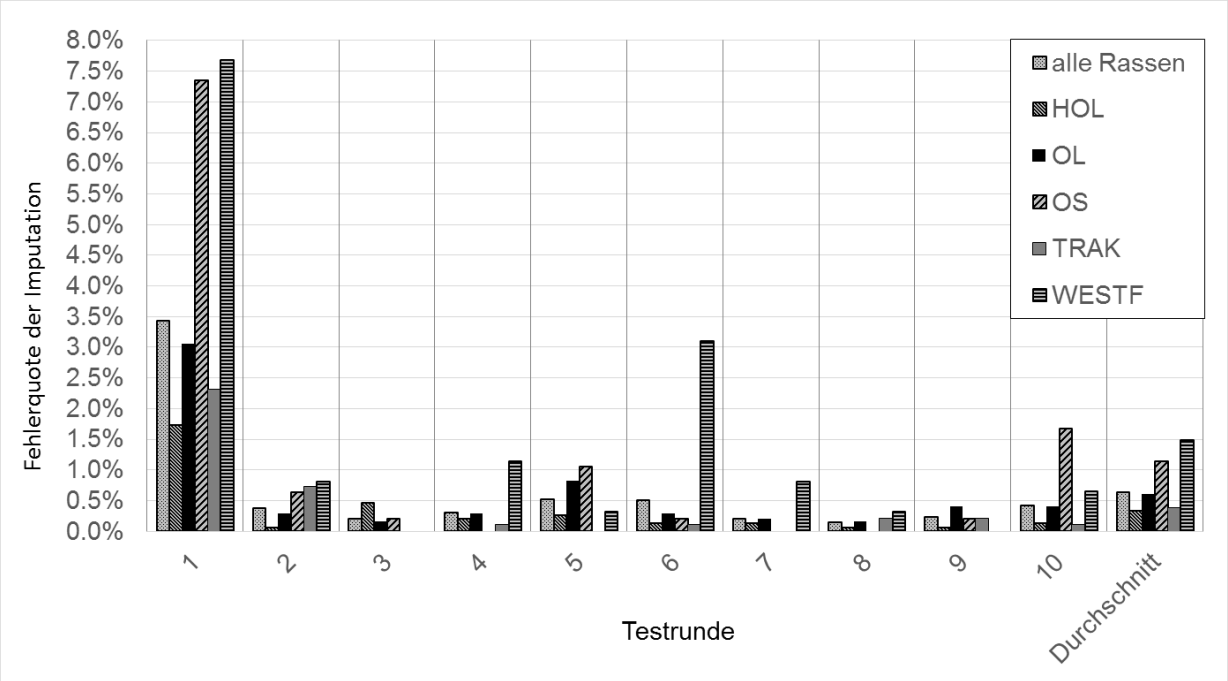

Abbildung 2: Fehlerquoten der Mikrosatellitenimputation anhand von SNPs aus den 10 Testrunden mit randomisierter Tierauswahl (Option A: Imputation innerhalb der Rassen).

Imputationsgenauigkeit je Mikrosatellit (Option A)

Eine Aufschlüsselung der Ergebnisse aus Option A nach Mikrosatelliten ergab eine heterogene Verteilung der Fehlerquoten (siehe Abbildung 3). Auffällig waren die Marker ASB17, ASB23 und im Besonderen AHT5, dessen Imputation sich bei allen fünf Rassen als problematisch erwies (durchschnittliche Fehlerquote > 3,5 Prozent in allen Rassen). Die Marker HMS1, CA425 und LEX3 zeigten zudem erhöhte Fehlerquoten bei WESTF, was sich, wie zuvor erwähnt, mit der vergleichsmäßig kleinen Stichprobengröße in Verbindung bringen lässt.

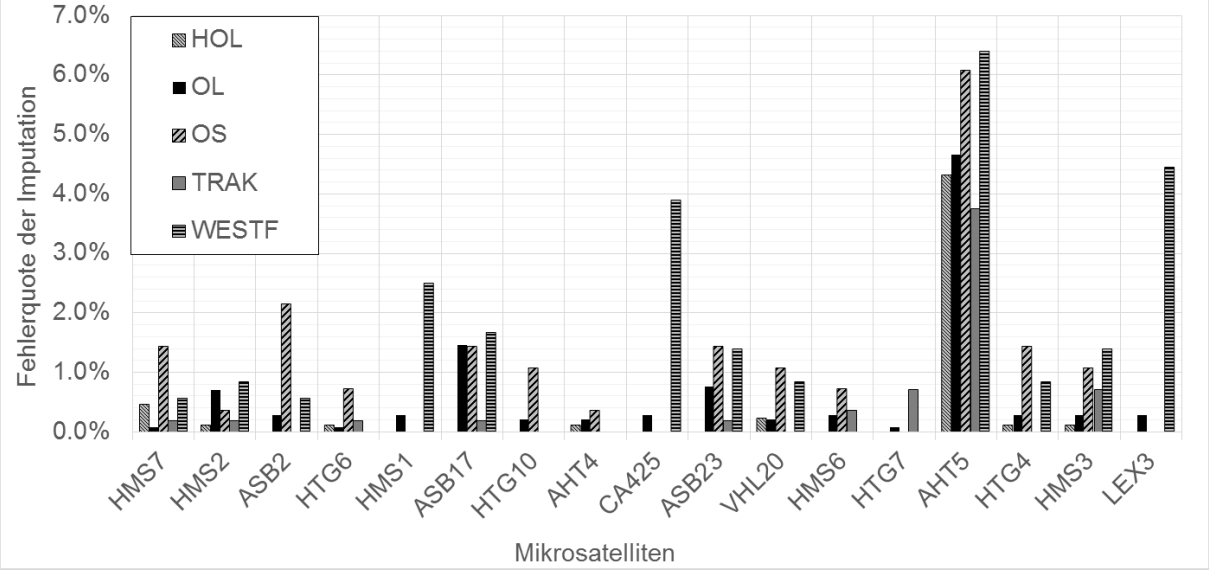

Abbildung 3: Fehlerquoten der Mikrosatellitenimputation anhand von SNPs, aufgeschlüsselt nach Mikrosatelliten (Option A: Imputation innerhalb der Rassen).

Fazit und Ausblick

Zusammenfassend lässt sich sagen, dass sich bei der Mikrosatelliten-imputation anhand von SNPs innerhalb von Rassen die höchste Genauigkeit erreichen ließ und dieses Verfahren zur weiteren Anwendung empfohlen werden kann. Durch Einbindung weiterer Pferde für jede der Rassen sollten sich die Genauigkeiten zudem weiter steigern lassen.

Der Gemeinschaft der IAFH liegen aus einem früheren Verbundprojekt (FUGATOplus-Projekt GENE-FL) noch SNP-Genotypisierungsdaten von knapp 600 Hengsten vor (HOL, OL, OS, TRAK), die auch für Mikrosatelliten genotypisiert wurden. Weitere Testimputationen sollen zeigen, ob sich durch die Einbeziehung dieser Hengste in die Referenzgruppe die Genauigkeit der Imputation erhöhen lässt. Da einige der Hengste sich in den Pedigrees der Stuten der Lernstichprobe wiederfinden, ist dieser Ansatz sehr vielversprechend.

Literatur

Bowling, A.T., Egglestone, M.L., Byrns, G., Clark, R.S., De Leanis, S., Wictum, E. (1997). Validation of microsatellite markers for routine horse parentage testing. *Animal Genetics*, 28, 247-252. doi:10.1111/j.1365-2052.1997.00123.x

Browning, B. L., Zhou, Y., Browning, S. R. (2018). A one-penny imputed genome from next generation reference panels. *American Journal of Human Genetics*, 103(3): 338-348. doi:10.1016/j.ajhg.2018.07.015

McClure, M., Sonstegard, T., Wiggans, G., & Van Tassell, C. P. (2012). Imputation of microsatellite alleles from dense SNP genotypes for parental verification. *Frontiers in Genetics*, 3, 140. doi:10.3389/fgene.2012.00140

McClure, M. C., Sonstegard, T. S., Wiggans, G. R., Van Eenennaam, A. L., Weber, K. L., Penedo, C. T., *et al.* (2013). Imputation of microsatellite alleles from dense SNP genotypes for parentage verification across multiple Bos taurus and Bos indicus breeds. *Frontiers in Genetics*, 4, 176. doi:10.3389/fgene.2013.00176

Yang, J., Lee, S. H., Goddard, M. E., Visscher, P. M. (2011) GCTA: a tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*, 88(1): 76-82. doi:10.1016/j.ajhg.2010.11.011