# Imputieren von SNP Genotypen bei Freiberger Pferden

**M. Frischknecht[1,2,3,4], M. Neuditschko[1,3], V. Jagannathan[2,3], C. Drögemüller[2,3], J. Tetens[5], G. Thaller[5], T. Leeb[2,3], S. Rieder[1,3]**
*[1] Agroscope – Schweizerisches Nationalgestüt SNG, Avenches; [2] Institut für Genetik, Vetsuisse-Fakultät, Universität Bern, Bern; [3] Kompetenzzentrum für Tierzucht der Universität Bern, Berner Fachhochschule HAFL & Agroscope; [4] Graduate School for Cellular and Molecular Biology, Universität Bern, Bern; [5] Institut für Tierzucht und Tierhaltung, Christian-Albrechts-Universität, Kiel, Deutschland*

## Einleitung

Mit Hilfe von Next-Generation Sequencing (NGS) Technologien kann ein Säugergenom nahezu vollständig sequenziert werden, womit eine maximale Markerdichte für genetische Analysen verfügbar wird. Allerdings ist es aus Zeit- und Kostengründen immer noch nicht möglich, eine grosse Anzahl von Individuen komplett zu sequenzieren. Deshalb werden aktuell die Sequenzinformationen von repräsentativen Referenztieren benutzt, um fehlende Genotypen der restlichen Tiere abzuleiten. In der Statistik werden solche Verfahren unter dem Begriff Imputation zusammengefasst. Die resultierende höhere Markerdichte ist vorteilhaft zum Beispiel für genomweite Assoziationsstudien oder genomische Selektion.

## Material und Methoden

Der in dieser Studie verwendete Datensatz umfasste 1'077 Freiberger Pferde (Signer-Hasler et al. 2012). Diese Pferde wurden für 54'602 Single Nucleotide Polymorphisms (SNP) genotypisiert. Für die Re-Sequenzierung wurden 28 repräsentative Freiberger Pferde ausgewählt. Davon waren 26 im vorher beschriebenen Datensatz enthalten. Des weiteren wurden 14 Warmblutpferde sequenziert. Nach dem Sequenzieren und dem „variant calling" wurden etwa 13 Millionen SNPs für die Imputation verwendet. Wir haben die Imputation mit den Programmen Beagle (Browning und Browning 2009), Impute2 (Howie et al. 2009) und FImpute (Sargolzaei et al. 2011) durchgeführt.

## Resultate

Die Evaluierung der verwendeten Programme erfolgte auf Basis der Chromosomen 16 und 31 mittels eines Kreuzvalidierungsverfahren. In diesem Verfahren wurden die 26 sequenzierten Pferde mit 50k Genotypen-Information berücksichtigt. Für die Berechnung der Genauigkeit abgeleiteter SNP Genotypen haben wir imputierte und wahre Genotypen miteinander verglichen und den relativen Anteil von korrekt imputierten SNPs berechnet (Tab. 1). Die Ergebnisse dieser Analyse zeigten einen deutlich erkennbaren Unterschied zwischen den verwendeten Programmen. Der Unterschied zwischen den beiden getesteten Chromosomen war hingegen nur marginal. In dieser Studie lieferte das Programm Impute2 die mit Abstand höchsten Genauigkeiten imputierter SNP Genotypen. Im Vergleich zu anderen Programmen werden bei Impute2 die Haplotypen basierend auf SNP Informationen von Referenz- und Testtieren rekonstruiert. Dieses Verfahren ist vor allem bei kleinen Referenzpopulationen zu bevorzugen.

Mit Impute2 haben wir anschliessend das ganze Genom imputiert. Dabei stellte sich heraus, dass die Genauigkeit zwischen allen Chromosomen sehr ähnlich ist. Nur für Chromosom 12 war die Genauigkeit mit 91.9% deutlich tiefer. Dies hat vermutlich in erster Linie mit der Qualität des Referenzgenoms dieses Chromosoms zu tun.

Um weitere wichtige Einflussfaktoren für das korrekte Imputieren von SNP Genotypen zu überprüfen, haben wir den Anteil korrekt imputierter SNPs für jedes einzelne Pferd berechnet. Die Genauigkeit pro Pferd lag zwischen 85.7% und 99.8%. Die Überprüfung der Abstammung der Pferde zeigte, dass der Warmblutanteil eines Pferdes einen grossen Effekt auf die Genauigkeit der Imputation hatte (Abb. 1). Um diesem Effekt entgegenzuwirken haben wir in einer weiteren Analyse die Sequenzdaten von 14 Warmblutpferden berücksichtigt und somit unsere Referenzpopulation vergrössert. Mit Hilfe der Warmblut Sequenzdaten verbesserte sich der Anteil von korrekt imputierten SNPs für 24 der 26 Pferde. Die Genauigkeit pro Pferd lag bei dieser Analyse zwischen 91.9% und 99.4%.

## Schlussfolgerung

Die Ergebnisse dieser Studie zeigen, dass die Imputation von 50k Genotypen bis zur Stufe der vollständigen Sequenz beim Freiberger mit einer hohen Genauigkeit möglich ist. Die wichtigsten Einflussfaktoren für ein erfolgreiches Imputieren von SNP Genotypen sind die optimale Auswahl des Imputationsverfahrens und die zugrundeliegende Populationsstruktur.

## Literatur

Browning B.L., Browning S.R., 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet, 84:210-223.

Howie B.N., Donnelly P., Marchini J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet, 5(6):e1000529

Sargolzaei M., Chesnais J.P., Schenkel F.S., 2011. FImpute - An efficient imputation algorithm for dairy cattle populations. J Anim Sci: 89, E-Suppl. 1 / J Dairy Sci 94, E-Suppl. 1: 421 (333).

Signer-Hasler H., Flury C., Haase B., Burger D., Simianer H., Leeb T., Rieder S., 2012. A genome-wide association study reveals loci influencing height and other conformation traits in horses.PLoS One, 7(5): e37282

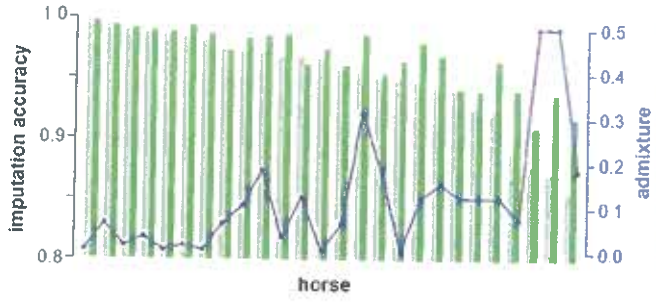


**Abb. 1 Genauigkeit der Imputation in Abhängigkeit vom Warmblutanteil pro Pferd.**
Die grauen Balken zeigen die Genauigkeit pro Pferd wenn nur die 28 FM Pferde als Referenz gebraucht wurden. Die grünen Balken zeigen die Genauigkeit wenn 14 Warmblutpferde zur Referenzpopulation hinzugefügt wurden. Die blauen Punkte zeigen den Warmblutanteil von jedem evaluierten Pferd.

**Tab. 1**
**Genauigkeit der Imputation für 3 verschiedene Programme**

| Programm | ECA 16 | ECA 31 |
|---|---|---|
| Impute2 | 0.958 | 0.954 |
| Beagle | 0.750 | 0.743 |
| FImpute | 0.774 | 0.772 |

Die Werte zeigen die Übereinstimmung zwischen wahren und imputierten Genotypen für 2 verschiedene Chromosomen (ECA)