Whitepaper: Einschätzung der ethnischen Herkunft (2018)

Überblick:

Das Forscherteam von AncestryDNA hat eine schnelle, ausgeklügelte und akkurate Methode entwickelt, die es erlaubt, Einschätzungen darüber zu geben, woher die DNA eines Kunden stammt. Die Einschätzung der Herkunft kann von mehreren Hundert bis über 1.000 Jahre in die Vergangenheit reichen. Wir haben die Methode vor Kurzem verbessert. Mit dem Ergebnis, dass wir unsere Kunden nun noch mehr Regionen als Referenzpunkte zuweisen können. Durch das Update verbessert sich auch noch einmal die Präzision der zugewiesenen Regionen sowie der für die betreffenden Regionen ermittelten Prozentsätze. Viele der Verbesserungen beruhen auf aktualisierten Analyseprozessen. Es werden fortan ganze DNA-Segmente statt nur einzelne Stellen untersucht. Wir denken, dass wir dank dieses hochmodernen wissenschaftlichen Ansatzes unsere Methoden immer weiter verbessern können. So werden wir mit fortschreitenden technischen Möglichkeiten auch immer genauere Ergebnisse erzielen.

Zur Einschätzung der ethnischen Herkunft eines Kunden wird seine DNA mit einem DNA-Panel anderer Menschen verglichen, deren Herkunft bekannt ist. (Diese Panels werden auch als Referenzpanels bezeichnet.) Dabei wird ermittelt, welche Teile der Kunden-DNA Ähnlichkeiten mit der DNA der Menschen aufweist, die im Referenzpanel in Gruppen eingeteilt sind. Gleicht ein Teil einer Kunden-DNA beispielsweise vor allem der DNA im Referenzpanel der schwedischen Gruppe, wird dieser Teil der Kunden-DNA dem Land Schweden zugewiesen. So gelangt der Test Schritt für Schritt zu einem DNA-Porträt des Kunden, das zu verschiedenen Prozentsätzen den 43 ethnischen Gruppen im Referenzpanel entspricht.

In den folgenden Zeilen geben wir Ihnen einen Überblick darüber, wie AncestryDNA die ethnische Herkunft eines Mitglieds bestimmt. Der Rest dieses Whitepapers geht näher auf die folgenden Sachverhalte ein:

1. Auswahl der Proben für das Referenzpanel sowie Zusammensetzung und Validierung der Proben
2. Funktionsweise des Algorithmus, anhand dessen die ethnische Abstammung eines Kunden bestimmt wird, und Validierung des Algorithmus

1. Einleitung

Die Einschätzung der ethnischen Abstammung bildet einen wichtigen Teil der DNA Story, die AncestryDNA bereitstellt. Wie der Name bereits vermuten lässt, gibt die DNA Story dem Kunden durch Analyse seiner DNA einen Einblick in die eigene Vergangenheit.

AncestryDNA beschäftigt ein Team aus hoch qualifizierten Wissenschaftlern mit hervorragenden Kenntnissen in den Bereichen Populationsgenetik, Statistik, maschinelle Lernprozesse und Computer-Biologie. Auf diesem Fundament haben wir eine schnelle, ausgeklügelte und präzise Methode entwickelt, mit deren Hilfe wir die genetische Abstammung unserer Kunden ableiten können. In diesem Dokument beschreiben wir den Ansatz, den wir verwenden, um die genetische Abstammung unserer Kunden einzuschätzen. Außerdem gehen wir auf die Entwicklung des Referenzpanels ein, mit dem wir die Proben unserer Kunden abgleichen. Ein weiterer Diskussionspunkt ist unsere Methode zur Ableitung der genetischen Abstammung. Zu guter Letzt kommen noch die ausgiebigen Testprotokolle zur Sprache, anhand derer wir die Qualität unserer Einstufungen bewerten.

Glossar

Allel – Eine Variante in der DNA-Sequenz. So können beispielsweise in einem SNP (Definition siehe unten) zwei Allele vorkommen: A oder C.
Centimorgan (cM) – Eine Hilfsmaßeinheit der Genetik. Liegen zwei Positionen im Genom ein Centimorgan auseinander, besteht bei jeder Meiose eine einprozentige Wahrscheinlichkeit, dass es zur Rekombination kommt. (Die Meiose beschreibt die Zellteilung, bei der Eizellen oder Spermien entstehen.)
Chromosom – Ein großes vererbtes Stück DNA. Beim Menschen finden sich in der Regel 23 Chromosomenpaare, wobei jeder Elternteil eine Kopie seiner Gene weitervererbt hat.
Genom – Die Gesamtheit der genetischen Information eines Menschen, mit anderen Worten: die gesamte DNA aller Chromosomen.
Genotyp – Ein Oberbegriff für die zu beobachtenden genetischen Variationen entweder an einer einzelnen Stelle oder im gesamten Genom.
Haplotyp – Ein Teil der DNA-Sequenz auf einem Chromosom.
Hidden Markov Model (HMM) – Ein stochastisches Modell, anhand dessen sich auf Basis unterschiedlicher Beobachtungen verschiedene verborgene Zustände bestimmen lassen.
Locus – Eine Stelle im Genom. Es kann sich dabei um eine ganz spezifische Position oder um einen größeren Teil der DNA handeln.
Microarray – Ein DNA-Microarray ist eine Möglichkeit, Hunderttausende von DNA-Markern gleichzeitig zu analysieren.
Nukleotide – Die DNA besteht aus Molekülsträngen, die auch als Nukleotide (oder auch als Basen) bezeichnet werden. Die vier unterschiedlichen Typen werden in der Regel durch ihre Anfangsbuchstaben gekennzeichnet: A, C, G, T.
Population – Eine Gruppe von Menschen.
Rekombination – Bevor die Chromosomen von den Eltern an die Kinder weitergegeben werden, tauschen die einzelnen Chromosomenpaare in der Regel lange DNA-Segmente miteinander aus, die dann bei der sogenannten Rekombination wieder miteinander verknüpft werden.
Einzelnukleotid-Polymorphismus (SNP) – Eine bestimmte Nukleotid-Position des Genoms, an der bei verschiedenen Menschen unterschiedliche Varianten (Allele) zu beobachten sind.

2. Referenzpanels

2.1 Berechnung der ethnischen Herkunft

Zwei Chromosomen ein und derselben geografischen Region oder derselben Populationen sind von ihrer DNA her ähnlicher als zwei Chromosomen aus unterschiedlichen Regionen oder Gruppen. Wenn also

zwei DNA-Teile von ihrer genetischen Abstammung her auf Schweden zurückgehen, werden sie mehr Ähnlichkeiten in ihrer DNA aufweisen als ein Stück DNA aus Korea und ein Stück DNA aus Schweden. Auf dieser Grundlage schätzt AncestryDNA die ethnische Abstammung seiner Mitglieder ein.

Zur Einschätzung der ethnischen Herkunft eines Kunden wird seine DNA mit einem DNA-Panel anderer Menschen verglichen, deren Herkunft bekannt ist. (Diese Panels werden auch als Referenzpanels bezeichnet.) Dabei wird ermittelt, welche Teile der Kunden-DNA Ähnlichkeiten mit der DNA der Menschen aufweist, die im Referenzpanel in Gruppen eingeteilt sind. Entspricht ein Abschnitt einer Kunden-DNA beispielsweise am ehesten den Proben aus dem senegalesischen Referenzpanel, ordnen wir den betreffenden Abschnitt der Kunden-DNA dieser genetischen Abstammung zu.

Wie genau die Beurteilung der ethnischen Herkunft ist, hängt von der Qualität unseres Referenzpanels ab. Genau deshalb hat AncestryDNA auch viel Arbeit in die Entwicklung der bestmöglichen Referenzproben gesteckt.

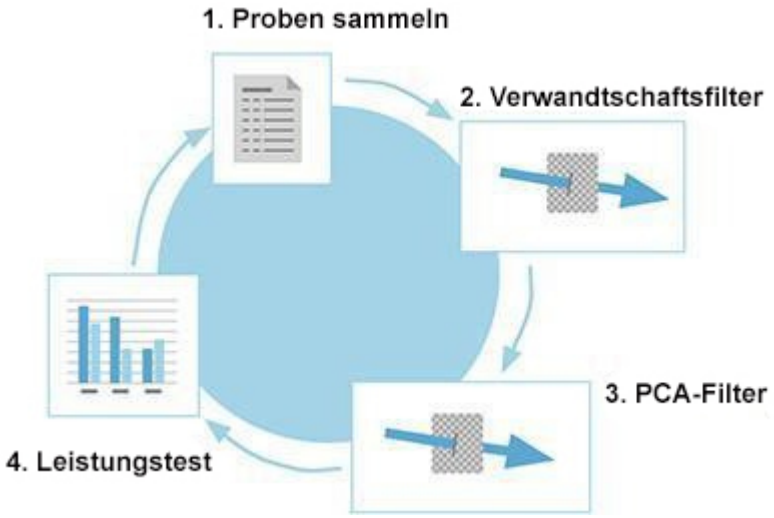


Abbildung 2.1: Zyklus zur Verfeinerung der Referenzpanels. *Eine schematische Darstellung unseres Zyklus zur Verfeinerung der Referenzpanels. In* ***Schritt 1*** *sammeln wir Referenzproben von Kandidaten. Hierzu verwenden wir veröffentlichte Daten, die AncestryDNA-Kundenliste und die AncestryDNA-interne Referenzsammlung. Bei den AncestryDNA-Proben stützen wir uns auf Herkunftsdaten. Auf deren Basis wählen wir die Kandidaten aus, die tief in einer bestimmten Population tief verwurzelt sind. In* ***Schritt 2*** *filtern wir Teile der DNA von eng miteinander verwandten Proben aus der Kandidatenliste heraus. In* ***Schritt 3*** *entfernen wir anhand einer Hauptkomponentenanalyse (engl.: principal component analysis, kurz: PCA) diejenigen Proben, bei denen sich hinsichtlich Abstammung und genetischer Herkunft Diskrepanzen ergeben. Die PCA nutzen wir auch zur Orientierung, um Populationsgruppen zu identifizieren. In* ***Schritt 4*** *folgt der Panel-Leistungstest auf Basis zahlreicher Daten. Dazu kommt ein Vergleich mit der Vorversion. Auf diesem Weg erstellen wir ein hochwertiges und gut getestetes Referenzpanel. Der gesamte Prozess ist zyklisch angelegt. AncestryDNA verbessert ständig das Panel. Das Ziel besteht darin, mit den verfügbaren Daten die größtmögliche Präzision bei der Bewertung der genetischen Herkunft zu erreichen.*

An unserem aktuellen Referenzpanel haben wir verschiedene wichtige Updates vorgenommen. Der Rest des vorliegenden Abschnitts 2 beschäftigt sich mit den Schritten zur Weiterentwicklung unseres aktuellen

Referenzpanels. Hierzu zählen die Probenauswahl, die Qualitätskontrollen sowie die Tests. Im beschriebenen Update wurde unter anderem die Zahl der Populationen im Referenzpanel auf 43 erhöht.

2.2 Auswahl der Kandidaten für das Referenzpanel

Die Identifikation der am besten geeigneten Kandidaten für das Referenzpanel ist ein ausschlaggebender Faktor, wenn es darum geht, so präzise wie möglich die ethnische Herkunft eines Kunden anhand seiner DNA-Probe zu ermitteln. Ideal wäre, wenn das Referenzpanel auf DNA-Proben von Menschen basieren würde, die vor Hunderten von Jahren gelebt haben. Leider ist es noch nicht möglich, verlässliche Daten zu Populationen der Menschheitsgeschichte zu sammeln. Wir müssen daher mit den DNA-Proben von Menschen arbeiten, die noch am Leben sind und ihre Herkunft auf einen einzigen geografischen Bereich oder eine Populationsgruppe eingrenzen können.

Verlässliche Aussagen zum eigenen Stammbaum reichen meist nur eine bis fünf Generationen in die Vergangenheit. Es ist schwer, Menschen zu finden, die noch tiefergehende Kenntnisse der eigenen Familiengeschichte haben. Das liegt einerseits daran, dass es bei der Rückverfolgung mit jeder Generation schwieriger wird, historische Aufzeichnungen zu finden. Andererseits verdoppelt sich mit jeder Generation die Zahl der Vorfahren.

Glücklicherweise reicht es oft, wenn jemand weiß, wo seine nächsten Vorfahren geboren wurden. Das erlaubt meist bereits ganz gute Rückschlüsse auf weit tiefer liegende familiäre Wurzeln. In der jüngeren Menschheitsgeschichte waren Migrationsbewegungen über größere Distanzen nämlich weitaus schwieriger zu bewerkstelligen und dadurch auch weniger üblich. Der Geburtsort der nächsten Vorfahren eines Menschen ist auch oft der Standort, an dem sich die Ursprünge seiner DNA finden.

Kandidaten für das AncestryDNA-Referenzpanel

Bei der Entwicklung des neuesten AncestryDNA-Referenzpanels starteten wir mit einem Set von fast 34.000 Proben. Zunächst einmal untersuchten wir 1.000 Proben von 52 Populationen aus aller Welt. Die Proben stammen aus einem öffentlichen Projekt namens Human Genome Diversity Project (HGDP) (Cann *et al*. 2002, Cavalli-Sforza 2005). Dazu kamen über 1.800 Proben aus 20 Populationen aus dem 1000 Genomes Project (McVean *et al.* 2012). Im zweiten Schritt untersuchten wir die Proben einer unternehmenseigenen AncestryDNA-Referenzdatenbank sowie die AncestryDNA-Proben von Kunden. Die meisten der Kandidaten aus den beiden letztgenannten Gruppen wurden nur berücksichtigt, wenn ihr Stammbaum belegte, dass sie in einer bestimmten Region oder Gruppe tief reichende familiäre Wurzeln hatten. Es wurde euch eine Reihe an Kandidaten ausgewählt, die keinen weit in die Vergangenheit zurückreichenden Stammbaum vorweisen konnten. Diese Gruppe hatte allerdings den unten beschriebenen strengen Sicherheitstest bestanden. Bei den Proben des HGDP und des 1000 Genomes Project war es nicht möglich, die Stammbäume zu verifizieren. Allerdings zielen diese Datenbanken speziell darauf ab, von großen genau umrissenen Populationsgruppen Proben zu entnehmen und dadurch eine Art Weltatlas humangenetischer Variationen zu erstellen.

2.3 Qualitätskontrollen für das Referenzpanel

Für jede Probe analysierten wir einen Satz aus etwa 300.000 SNPs. Die Daten verteilten sich auf die beiden Plattformen Illumina OmniExpress und Illumina HumanHap 650Y und kamen auch bei der Genotyp-Bestimmung der HGDP-Proben zum Einsatz. Nachdem wir Proben mit großen Datenlücken entfernt hatten, filterten wir diejenigen Proben heraus, die mit großer Wahrscheinlichkeit die Leistung des Referenzpanels beeinträchtigt hätten. Entfernt wurden vor allem diejenigen Proben, die eine große Ähnlichkeit zu anderen Referenzproben aufwiesen. Ein weiteres K.-o.-Kriterium waren Diskrepanzen zwischen den zugrundeliegenden genetischen Daten und dem Stammbaum.

Bei der Abschätzung der genetischen Herkunft geht es um eine computergestützte Wahrscheinlichkeitsrechnung. Die Frage lautet: Wie hoch ist die Wahrscheinlichkeit, dass ein bestimmtes DNA-Segment beziehungsweise ein ermittelter Haplotyp aus einer bestimmten Population des Referenzpanels stammt (siehe Abschnitt 4 unten)? Konkret ausgedrückt: Wie hoch ist die Wahrscheinlichkeit, dass ein bestimmter DNA-Abschnitt aus Schweden stammt? Oder aus Frankreich? Oder aus einer der anderen Regionen, die wir abprüfen?

Für diese Berechnung müssen wir einschätzen, wie häufig ein Haplotyp in den einzelnen Populationen vorkommt. Voraussetzung dafür ist, dass keine engen Verwandtschaften zwischen den Menschen im Referenzpanel bestehen. Ähnlichkeiten in DNA-Segmenten, die auf gemeinsame nahe Vorfahren zurückgehen, wie sie durch Identität nach Abstammung (IBD) ermittelt werden können, stellen nämlich keine unabhängigen Haplotypen innerhalb einer Population dar. Werden derartige DNA-Segmente beibehalten, verzerrt dies bei der Einschätzung der Häufigkeit, mit der ein Haplotyp innerhalb einer Population vorkommt, das Gesamtbild. Aus diesem Grund streichen wir bei Kandidaten ab einer gewissen Obergrenze an gemeinsamer IBD-DNA (20 cM) die entsprechenden Segmente. Weitere Informationen zu unserem Ansatz bei der Ermittlung gemeinsamer IBD-DNA-Segmente finden Sie in unserem AncestryDNA Whitepaper zum Matching (https://www.ancestry.com/corporate/sites/default/files/AncestryDNA- Matching-White-Paper.pdf).

Als Nächstes streichen wir die Proben aus dem Kandidatensatz des Referenzpanels, wenn die genetischen Daten zur Abstammung nicht mit den Angaben der betreffenden Person zum eigenen Stammbaum übereinstimmen. Diese Fehleinträge identifizieren wir mithilfe der Hauptkomponentenanalyse (PCA).
Die PCA wird in der Populationsgenetik häufig zur explorativen Datenanalyse verwendet (Jackson 2003). Bei korrekter Anwendung auf die Genotyp-Daten lassen sich mithilfe der PCA genetische Variationen zwischen unterschiedlichen Populationen ermitteln (Patterson 2006).

Wir wenden die PCA auf die Proben an, die den vorherigen Screening-Prozess durchlaufen haben.
Die ersten Stufen der Analyse – die ersten vier Hauptkomponenten – stellen wir in einer Reihe von Streudiagrammen dar. Die einzelnen Proben werden nach Ursprungspopulation eingefärbt.
Bei AncestryDNA-Proben dient der Stammbaum als Ausgangsbasis, bei öffentlichen Proben die Probenbezeichnung (siehe Abbildung 3.3).

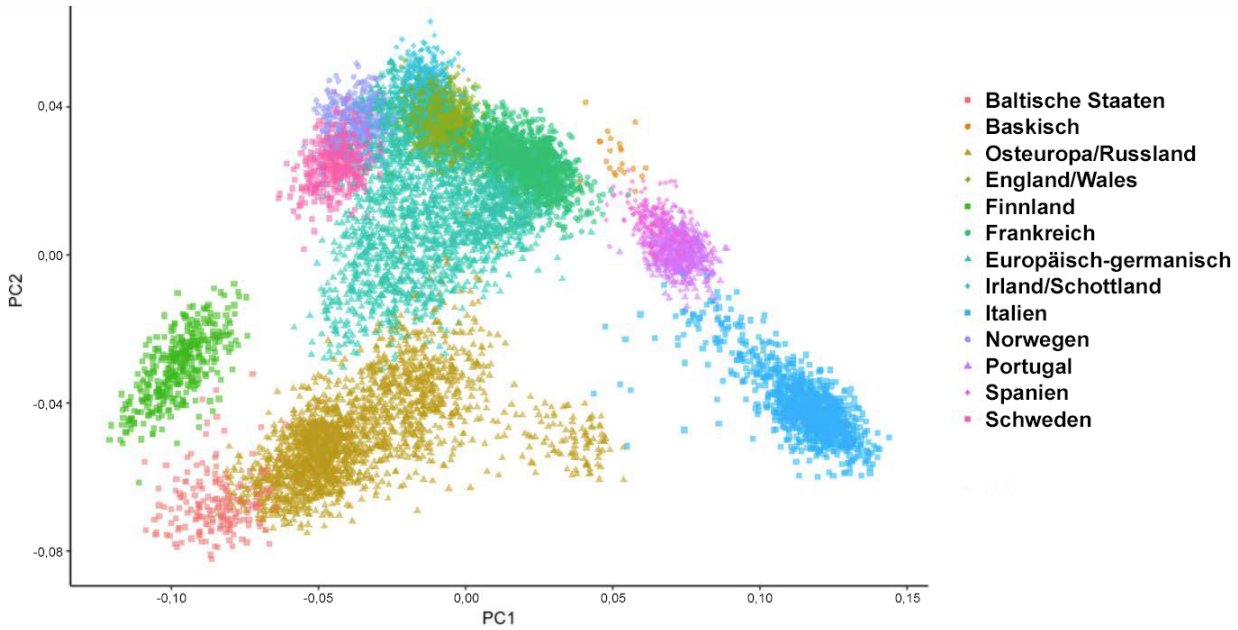


Abbildung 2.2: PCA-Analyse zu Kandidaten des europäischen Panels. *Streudiagramm mit den ersten beiden Komponenten der Hauptkomponentenanalyse (PCA) auf Basis europäischer Kandidatenproben für das AncestryDNA-Referenzpanel. Die visuelle Begutachtung der PCA ist aus unterschiedlichen Gründen der Daten-QS sinnvoll. Zunächst einmal lassen sich damit einzelne Abweichungen wie etwa die Proben für Osteuropa/Russland (gelbe Dreiecke) identifizieren, die mitten im Cluster des germanischen Europa auftauchen. Außerdem kann sich dieser zusätzliche Schritt lohnen, wenn es darum geht, mangelhafte Gruppierungen von Proben zu identifizieren. Ein weiterer Effekt: Es können Regionen mit verschwommenen genetischen Trennlinien ausfindig gemacht werden, wo es zur Überlappung der Cluster kommt. Beispiele hierfür wären Überlappungen zwischen Irland/Schottland sowie England, Wales und Nordwesteuropa. Umgekehrt können hierbei aber auch Regionen entdeckt werden, bei denen eine weitere Unterteilung angezeigt ist.*

Jede Population bildet im Streudiagramm ein Punkte-Cluster, wobei jeder Punkt einer Probe entspricht. Die Cluster sind darauf zurückzuführen, dass genetisch ähnlichere Punkte im PCA-Raum enger beieinanderliegen. Hilfreich ist dabei, dass die Punkte-Cluster auch auf eine geografische Nähe hindeuten, da die meisten Menschen in einer Region genetische Ähnlichkeiten zu anderen Personen umliegender Regionen aufweisen. Darüber hinaus sind in solchen Diagrammen Abweichungen bei den Proben leicht zu erkennen, da sie auch räumlich von den Proben einer zusammengehörigen Population getrennt sind. So deuten beispielsweise die gelben Dreiecke innerhalb des Clusters aus grünen Dreiecken auf Proben hin, deren Stammbaum auf Osteuropa zurückgeht, deren DNA aber eher der deutschen Population entspricht. In diesen Beispielen weicht die angegebene Ursprungspopulation von der im PCA-Raum dargestellten genetischen Herkunft ab.

Bei der visuellen Überprüfung halten wir Ausschau nach Streichkandidaten wie im Streudiagramm der Abbildung 2.2. Durch unterschiedliche Probensammlungen ergeben sich unterschiedliche Verteilungen bei der Populationsstruktur. Deshalb wiederholen wir die PCA und die Streichung von Abweichungen für unterschiedliche Teil-Datensätze. Zunächst einmal streichen wir die Abweichungen auf globaler Ebene. Hier werden alle Proben zusammen dargestellt. Danach geht es weiter auf kontinentaler Ebene, etwa mit Abweichungen in einer PCA, die ausschließlich auf europäischen Proben basiert, anschließend auf regionaler Ebene, etwa mit Abweichungen in einer PCA, die sich nur auf skandinavische Proben konzentriert. Zu guter Letzt folgt die Untersuchung auf Populationsebene, etwa mit Abweichungen in einer PCA, die rein auf norwegischen Proben basiert.

2.4 Verbesserung des iterativen Referenzpanels

Nach dem Entfernen der PCA-Abweichungen teilen wir das globale Referenzpanel in verschiedene Populationen ein, entsprechend der spezifischen genetischen Cluster in den PCA-Diagrammen. Bevor wir mit dem Referenz-Datensatz die Abstammung der AncestryDNA-Kunden einschätzen, ermitteln wir zunächst einmal die Qualität des Datensatzes. Dazu bewerten wir, wie gut sich mittels unseres Referenz-Datensatzes die ethnische Herkunft bestimmen lässt. Dazu ermitteln wir, wie unsere Stammbaum-Analyse bei Proben abschneidet, die per Definition zu 100 Prozent einer spezifischen ethnischen Gruppe zugehören.

Dabei entfernen wir 5 % der Proben aus dem Referenzpanel und bewerten die Abstammung mithilfe der übrigen 95 % der Proben als neuem Referenzpanel. Diesen Vorgang wiederholen wir insgesamt 20 Mal. Dabei entfernen wir jedes Mal andere 5 % des Panels, um anschließend bei den übrigen 95 % die genetische Abstammung zu bestimmen. Im Anschluss sehen wir uns die durchschnittliche vorhergesagte Abstammung für Proben der einzelnen Regionen im Referenz-Datensatz an. Dafür verwenden wir die Ergebnisse aus diesen Kreuzvalidierungsverfahren. In Abbildung 2.3 sind die Ergebnisse des Experiments als Kastengrafik dargestellt.

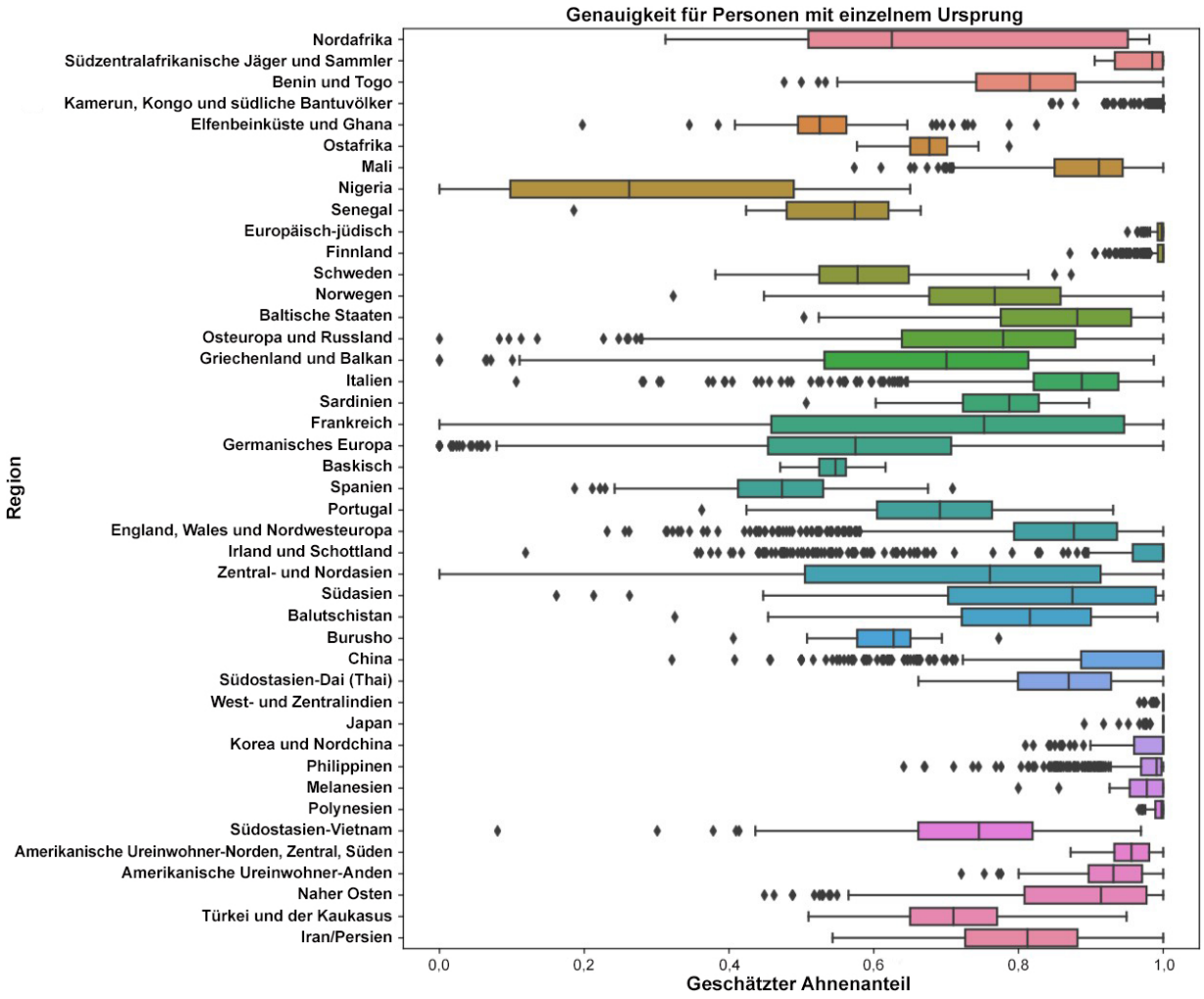


Abbildung 2.3: Zwanzigfaches Kreuzvalidierungsverfahren mit dem Referenzpanel V3. *In diesem Diagramm sind die Ergebnisse eines Experiments zu sehen, bei dem 5 % der Proben vom Referenzpanel entfernt wurden. Die übrigen Proben (95 %) des Panels wurden hinsichtlich ihrer ethnischen Abstammung beurteilt. Die einzelnen Kästen stehen für die Verteilung der angenommenen ethnischen Abstammung für alle Proben einer spezifischen Region (Wahrscheinlichkeiten von 75 %, 50 % und 25 % der angenommenen Abstammung). Für die Mehrzahl der Proben einer jeder Region ist nach unserer Vorhersage die genetische Abstammung zu durchschnittlich 78,9 % korrekt. Die anderen 21,1 % verteilen sich größtenteils auf nahegelegene Regionen. Allerdings gibt es auch Ausnahmen. Vor allem die durchschnittliche Präzision der Vorhersage für Proben aus Nigeria, Spanien und dem Baskenland ist nicht ganz so hoch. Es gibt viele Faktoren, welche die Präzision dieser Zahlen beeinträchtigen, allen voran die Zahl der Referenzproben im Panel für die einzelnen Regionen sowie die genetischen Besonderheiten der einzelnen Regionen.*

Diese Analyse verfolgt zwei Ziele. Einerseits sollen Proben des Referenzpanels mit extrem schwacher Leistungsfähigkeit im Kreuzvalidierungsverfahren eliminiert werden, da sie die betreffende ethnische Abstammung möglicherweise nur unzureichend abbilden. Andererseits belegt das Kreuzvalidierungsverfahren, wie präzise wir die ethnische Abstammung der Proben aus unserem Referenzpanel mithilfe unserer Methode einschätzen können (siehe Abschnitt 3). Das hilft uns,

die Populationsgrenzen genauer zu ziehen. So können wir beispielsweise zwei Populationen zusammenlegen, wenn die Leistung im Kreuzvalidierungsverfahren für beide Gruppen schwach ausfällt, die Ergebnisse in einer zusammengelegten Gruppe aber stichhaltiger sind.

Nachdem wir das Referenzpanel mithilfe mehrfacher Kreuzvalidierungsverfahren mehrmals optimiert haben, entschlossen wir uns dazu, das neueste globale Referenzpanel in 43 Regionen zu unterteilen. Die Regionen wollen wir in der folgenden Liste genauer benennen.

2.5 Aktualisiertes Referenzpanel

Das aktualisierte AncestryDNA-Referenzpanel zur Einschätzung der ethnischen Abstammung enthält 16.638 sorgfältig ausgewählte Proben (siehe Beschreibung oben). Die Welt wird dabei in 43 einander überlappende Regionen eingeteilt (Tabelle 2.1), die jeweils einzigartige genetische Profile aufweisen. Zum Vergleich: Unser vorheriges globales Panel bestand aus 3.000 Proben und nur 26 Regionen.

Region	**Anzahl der Proben**
Nordafrika	41
Südzentralafrikanische Jäger und Sammler	34
Benin und Togo	224
Kamerun, Kongo und südliche Bantu-Völker	579
Elfenbeinküste und Ghana	124
Ostafrika	82
Mali	169
Nigeria	111
Senegal	31
Amerikanische Ureinwohner Norden/Zentrum/Süden	146
Amerikanische Ureinwohner Anden	63
Asien Zentrum/Norden	186
Asien Süden	600
Belutschistan	53
Hunzukuc	23

China	620
Asien Südosten (Thai)	80
Indien Westen/Zentrum	65
Japan	592
Korea und Nordchina	261
Philippinen	538
Südostasien/Vietnam	159
England, Wales und Nordwesteuropa	1519
Baltische Staaten	194
Baskenland	22
Irland und Schottland	500
Europäisch-jüdisch	200
Frankreich	1407
Europäisch-germanisch	2072
Griechenland und Balkan	242
Italien	1000
Norwegen	367
Portugal	404
Sardinien	30
Osteuropa und Russland	1959
Spanien	270
Schweden	372
Finnland	361
Naher Osten	271
Iran/Persien	459
Türkei und Kaukasus	101
Melanesien	49

Polynesien	58
Insgesamt	**16.638**

Tabelle 2.1: Das abschließende Referenzpanel für AncestryDNA V3

Ausführlichere Tests zur Leistungsfähigkeit des Referenzpanels zur Einschätzung der ethnischen Abstammung stellen wir in Abschnitt 4 vor. Weitere Informationen zur AncestryDNA-Methode, welche wir zur Einschätzung der ethnischen Abstammung verwenden, finden Sie in Abschnitt 3.

3. Einschätzung der ethnischen Abstammung mittels AncestryDNA

3.1 Einleitung

Nachdem das Referenzpanel eingerichtet und validiert wurde, besteht der nächste Schritt darin, die ethnische Zugehörigkeit eines Kunden zu ermitteln. Hierzu werden über 300.000 Einzelnukleotid-Polymorphismen (SNPs) der Kunden-DNA mit SNPs des Referenzpanels verglichen. Wir gehen davon aus, dass die DNA eines Individuums aus einer DNA-Mischung besteht, die in den 43 im Referenzpanel dargestellten Populationen enthalten ist. Diese Annahme veranschaulichen wir in Abbildung 3.1: Ein Kunde erbt durch Rekombination lange DNA-Segmente seiner vier Großeltern, die in diesem Beispiel aus vier unterschiedlichen Referenz-Populationen stammen.

Die DNA wird in langen Segmenten von einer Generation an die nächste weitergegeben. Daher ist es wahrscheinlich, dass die DNA an zwei nahegelegenen SNPs oder Positionen des Genoms von ein und demselben Vorfahren stammt und damit auf die identische Population zurückgeht. Weitere Informationen zur DNA-Vererbung finden Sie in unserem Whitepaper zum DNA-Matching (http://dna.ancestry.com/resource/whitePaper/AncestryDNA-Matching-White-Paper). Wir können also präzisere Ergebnisse erzielen, indem wir uns mehrere nahegelegene SNPs zusammen oder als Gruppe (oder Haplotyp) ansehen, statt die einzelnen SNPs isoliert zu betrachten. Unsere neueste Methode nutzt dieses Prinzip, um die Präzision der Einschätzungen stark zu verbessern.

Bei unserer Einschätzung zur genetischen Abstammung eines Kunden setzen wir voraus, dass jedes Segment seines Genoms auf eine der 43 Populationen im Referenzpanel zurückgeht. Bei der Analyse unterteilen wir das Genom des Kunden in 1.001 sogenannte Fenster. Wir gehen davon aus, dass jedes Fenster klein genug ist, das jeder der beiden im Fenster befindlichen Haplotypen der Eltern aus genau einer Population stammt. Im nächsten Schritt kombinieren wir die Informationen aller Fenster, um einschätzen zu können, welcher Gesamtanteil des Kunden-Genoms von den unterschiedlichen Populationen im Referenzpanel stammt. Dabei hilft uns ein sogenanntes Hidden Markov model (HMM).

In Abbildung 3.1 ist ersichtlich, dass nicht jedem Fenster eine eindeutige ethnische Herkunft zugeordnet sein muss. Stattdessen kann ein Fenster ein Teil vom einen und ein Teil vom anderen Vorfahren enthalten. Das erste Fenster beinhaltet beispielsweise zwei unterschiedliche Abstammungen, gekennzeichnet durch die Farben Grün und Rot. Ein System zur Einschätzung der Abstammung auf Basis der AncestryDNA-Technologie muss die Möglichkeit berücksichtigen, dass jedes Fenster zwei unterschiedliche Abstammungen enthalten kann. Mit anderen Worten: Das System muss einen Ansatz nutzen, anhand dessen die DNA begutachtet und als Mischung aus Rot und Grün identifiziert wird. Es darf sich nicht auf ein einfaches Rot/Rot, Grün/Grün, Rot/Grün oder Grün/Rot beschränken.

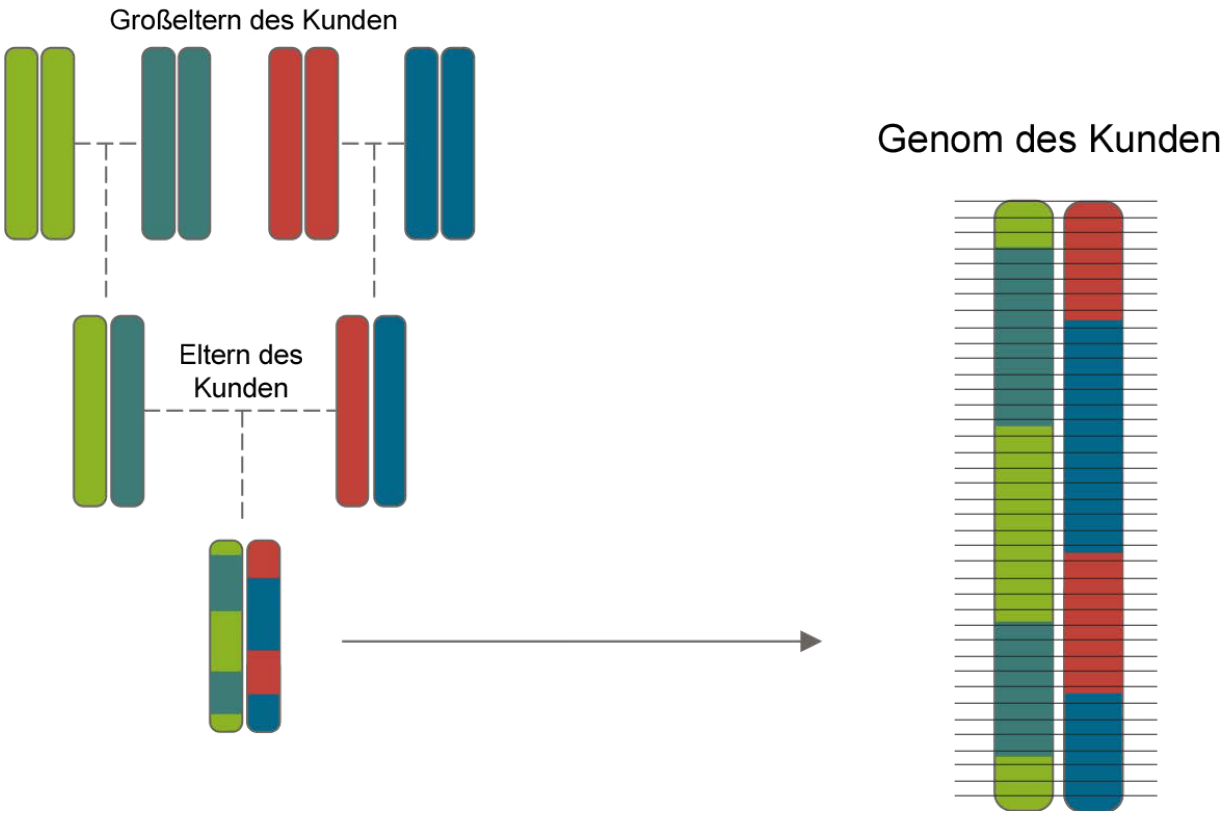


Abbildung 3.1: Vererbung von DNA unterschiedlicher Populationen. *Auf der linken Seite sehen Sie einen genetischen Stammbaum, der sich über drei Generationen erstreckt. Die einzelnen Individuen sind als zwei senkrechte Balken dargestellt. Sie stehen für die zwei Kopien eines einzelnen Chromosoms, wie sie jeder Mensch in sich trägt. Die Balken sind in den Farben der Referenzpopulation dargestellt, von der die DNA geerbt wurde. Alle vier Großeltern (die einfarbigen Balken in der oberen Zeile) haben 100 % ihrer DNA von einer einzelnen Population geerbt, die sich von den anderen drei Populationen unterscheidet. Die DNA wird erst an die Eltern des Kunden und letztendlich an den Kunden selbst weitergegeben. Durch die Rekombination und Neuordnung des Erbguts entsteht ein neues Chromosomenpaar mit DNA beider Eltern. Wie die Abbildung zeigt, besteht die DNA des Kunden aus einer Mischung der von den vier Großeltern geerbten DNA. Die Chromosomen enthalten jeweils längere DNA-Segmente, die von ein und demselben Vorfahren stammen. In der Abbildung rechts haben wir das Genom des Kunden zur Ermittlung seiner ethnischen Herkunft in kleine Fenster unterteilt, die durch schwarze horizontale Linien dargestellt werden. In jedem dieser Fenster weisen wir der Kunden-DNA eine spezifische Population zu, die auf die beiden Eltern zurückgehen. Es gibt also eine Population für jeden der beiden geerbten Haplotypen. Bei der Zuweisung der Populationen zu den Fenstern ziehen wir die Höhe der Übereinstimmung mit den Genomen im Referenzpanel heran.*

3.2 Prinzipien des Hidden Markov Models

Bei der Analyse von DNA-Daten wissen wir vorher noch nicht, auf welche Population die DNA zurückgeht. Stattdessen betrachten wir die Allelpaare (oft auch als Genotyp bezeichnet) an den einzelnen Positionen (SNPs) der DNA. Eines der Allele stammt von der Mutter, das andere vom Vater.

Die Wahrscheinlichkeit des Auftretens spezifischer Alleenpaare an den unterschiedlichen Positionen der DNA variiert in Abhängigkeit von den 43 genannten Regionen. Auf dieser Grundlage können wir einschätzen, aus welcher Region ein DNA-Strang am wahrscheinlichsten stammt. Kommt beispielsweise das Paar AA an einer bestimmten Position bei Spaniern häufiger vor, ist bei Personen mit dem Allelpaar AA an der betreffenden Position die Wahrscheinlichkeit einer spanischen Abstammung erhöht.

Wichtig dabei ist: Das AA-Paar an der betreffenden Position *erhöht nur die Wahrscheinlichkeit*, dass die DNA spanischer Abstammung ist. Das AA-Paar kann auch bei vielen Menschen aus Portugal, Frankreich oder sogar Korea vorkommen. Bei der Einschätzung der genetischen Abstammung werden die Wahrscheinlichkeiten für alle Positionen innerhalb eines Fensters untersucht. So wird festgestellt, aus welcher Region die DNA mit der höchsten Wahrscheinlichkeit stammt. Die genetische Herkunft sämtlicher Positionen leiten wir mithilfe eines statistischen Hilfsmittels ab, das auch als Hidden Markov Model (HMM) bekannt ist (Rabiner 1989).

Das Genom eines jeden Kunden besteht aus einer Sequenz der Nukleotide A, T, C und G, die zusammen eine Kette bilden. Die Positionen der Nukleotide hängen von der Population ab, auf die das betreffende DNA-Segment zurückgeht. Diese Information ist uns anfangs allerdings verschlossen. Zur Bestimmung der genetischen Herkunft bestimmt das HMM statistisch die wahrscheinlichste Referenzpopulation, aus der die betreffende DNA stammt (den wahrscheinlichsten verborgenen Zustand), anhand einer Observationsreihe. In unserem Fall werden dazu die Genotypen oder bestimmte Kombinationen von SNPs herangezogen.

Wir unterteilen das Genom des Kunden in 1.001 DNA-Abschnitte, die wir als Fenster bezeichnen. Anschließend versuchen wir, die verborgenen Zustände der einzelnen Fenster zu bestimmen. Jedes Fenster besteht aus zwei DNA-Abschnitten: der eine von der Mutter, der andere vom Vater. Die beiden Abschnitte können identischer oder unterschiedlicher ethnischer Herkunft sein. Der „verborgene Zustand“ betrifft also in diesem Fall die ethnische Herkunft der einzelnen DNA-Abschnitte im Fenster.

Für einen HMM sind zwei Faktoren ausschlaggebend: die Emissionen und die Übergangswahrscheinlichkeiten. Anhand der Emissionswahrscheinlichkeiten können wir ablesen, wie wahrscheinlich es auf Basis der analysierten Sequenz ist, dass ein DNA-Strang aus einer der 43 Populationen stammt. Die Übergangswahrscheinlichkeit gibt an, wie wahrscheinlich es ist, dass sich die Populationszugehörigkeit von einem Fenster zum nächsten verschiebt. Wurde beispielsweise die DNA im aktuellen Fenster ausschließlich der schwedischen Population zugeordnet, wird festgestellt, wie hoch die Wahrscheinlichkeit ist, dass die DNA des nächsten Fensters auch aus Schweden stammt.

Das ist ein sinnvolles Modell zur Analyse menschlicher DNA, da das menschliche Genom entlang der Chromosomen linear angeordnet ist. Die natürlichen genetischen Abläufe haben einen weiteren Vorteil, den wir uns zunutze machen: So weisen ganze Genom-Abschnitte und damit auch die auf den Chromosomen aneinandergereihten und beim Ancestry-Test analysierten Nukleotide bei entsprechender Herkunft identische genetische Strukturen auf.

3.3 Ableitung der Herkunftswahrscheinlichkeit mithilfe eines genomumspannenden HMM

Bei AncestryDNA verwenden wir Microarrays, um aus Kundenproben DNA-Daten zu ermitteln. Wir sehen uns über 700.000 einzelne Positionen auf der DNA (sogenannte SNPs) an und bestimmen die Nukleotide an jeder Position. So können sich zum Beispiel an Position 1 ein A und ein T, an Position 2 zwei Gs befinden usw. Bei der Einschätzung der genetischen Herkunft verwenden wir etwa 300.000 dieser SNPs.

Bei der Arbeit mit Array-Daten Ist es wichtig, dass jeder Mensch über zwei Kopien der jeweils 22 Chromosomen verfügt, zu denen AncestryDNA Daten ausgibt. Ein Chromosomensatz stammt von der Mutter, der andere vom Vater. Mit anderen Worten: Es gibt zwei Ergebnisse für jede Position, die wir mit AncestryDNA analysieren. Die Ergebnisse müssen entsprechend interpretiert werden, um herauszufinden, welche DNA von welchem Chromosomensatz stammt. Dieser Prozess wird auch als Phasing bezeichnet. AncestryDNA muss also die möglichen Kombinationen berücksichtigen, die es für die ethnische Herkunft einer Person gibt. Wurde bei einem Kunden beispielsweise ein DNA-Abschnitt identifiziert, der mütterlicherseits auf schwedische und väterlicherseits auf japanische Vorfahren hinweist, muss der Algorithmus diesen Abschnitt von einer anderen Sequenz unterscheiden können, die auf schwedische und nigerianische Vorfahren zurückgeht.

Wir erstellen dafür ein genomumspannendes HMM (dargestellt in Abbildung 3.2), das alle möglichen ethnischen Kombinationen (oder verborgenen Zustände) durch Populationspaare in den Fenstern des Genoms abbildet, wobei sich in benachbarten Fenstern mit großer Wahrscheinlichkeit keine Zustandsänderung ergibt. Mit anderen Worten: Wenn im aktuellen Fenster die DNA der Mutter als auch des Vaters aus Nigeria stammt, steigt die Wahrscheinlichkeit, dass das nächste Fenster auf dieselbe ethnische Herkunft hinweist.

Doch auch die Möglichkeit einer sich ändernden Populationszuweisung zwischen benachbarten Fenstern gilt es zu berücksichtigen (sogenannte Übergangswahrscheinlichkeit). Deutet ein Fenster auf eine schwedische Herkunft väter- und mütterlicherseits hin, müssen die analysierten DNA-Daten sehr stichhaltig sein, damit das benachbarte Fenster eine andere Populationszuweisung erhält. Durch Anwendung dieser Wahrscheinlichkeiten auf das komplette Genom ergibt sich eine komplette Abfolge von Populationszuweisungen über das gesamte Genom eines Kunden hinweg.

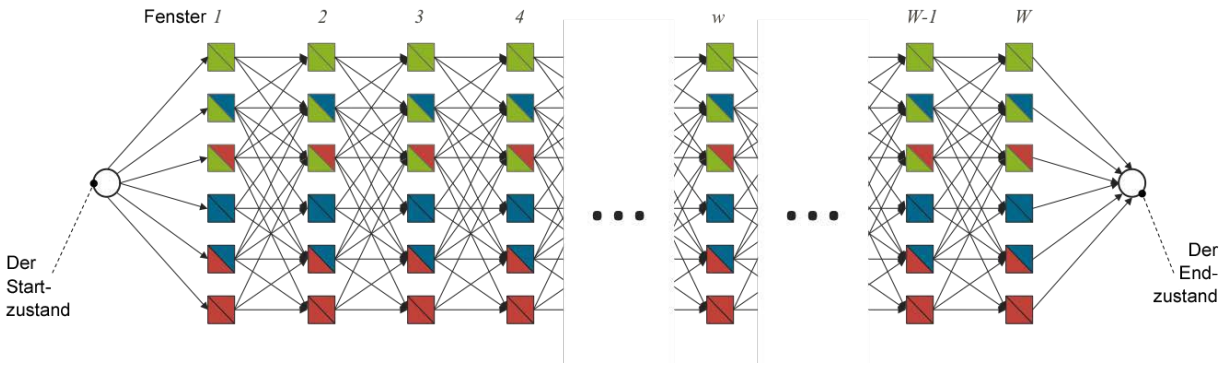


Abbildung 3.2: Illustration des genomumspannenden HMM für drei Populationen. Das Genom wird auf *W* Fenster aufgeteilt. Anschließend skizzieren wir die Übergangswahrscheinlichkeiten von einem Fenster *w* zum benachbarten Fenster *w+1.* Die möglichen (verborgenen) Zustände eines Fensters werden hier durch farbige Quadrate dargestellt. Die unterschiedlichen Farben stehen dabei für verschiedene Populationszuweisungen. Zweifarbige Kästen geben im betreffenden Fenster eine gemischte Herkunft an. Die Pfeile zwischen den Kästen repräsentieren die Übergänge. Jeder Kasten/Zustand gibt den beobachteten Genotyp des Kunden aus, wobei die Wahrscheinlichkeit jeweils im vorherigen Schritt im Voraus berechnet wird. Übergänge von einem Fenster zum anderen, die zu keiner Veränderung des Zustands/der Farbe führen, weisen eine höhere Wahrscheinlichkeit auf als Übergänge mit Zustands-/Farbwechsel. Es sind nur Übergänge erlaubt, bei denen sich nicht mehr als eine Farbe/Population ändert.

3.4 Übergangswahrscheinlichkeiten

Die Übergangswahrscheinlichkeit beschreibt die Wahrscheinlichkeit, dass sich die Herkunft von einem Fenster zum nächsten ändert. AncestryDNA berücksichtigt nur Übergänge zu direkt angrenzenden Fenstern. Die einzigen Einflussfaktoren der Übergangswahrscheinlichkeit sind also der Zustand des aktuellen Fensters sowie des nächsten Fensters. Diese sogenannte Gedächtnislosigkeit ist zentraler Bestandteil eines HMM.

Es sind keine Übergänge erlaubt, bei denen sich beide Populationen verändern. Aus biologischer Sicht ist es nämlich sehr unwahrscheinlich, dass sich innerhalb ein und desselben Fensters die Chromosomensätze von Vater UND Mutter ändern. Mit diesem Wissen können wir die Zahl möglicher Übergänge und die Komplexität des HMM stark einschränken.

Die einzige Ausnahme bildet der Übergang vom Ende eines Chromosoms zum Anfang des nächsten. In diesen Fällen ist die Wahrscheinlichkeit für eine Veränderung weitaus größer als innerhalb ein und desselben Chromosoms, da die DNA-Stränge unterschiedlicher Chromosomen ja nicht miteinander verknüpft sind. Deshalb erzwingen wir zwischen den Chromosomen einen sogenannten Silent State. Das ist durchaus sinnvoll, da es zwischen zwei Chromosomen keine direkte Verbindung gibt. Die Übergangswahrscheinlichkeit von einem Silent State zu einem beliebigen Paar von Populationszuweisungen wird einfach durch die genomumspannende Wahrscheinlichkeit sämtlicher Positionen des Genoms bestimmt, an denen die entsprechende Populationszuweisung vorliegt. Mit anderen Worten: Es gibt keine Daten vorheriger Fenster, welche die Interpretation eines Fensters beeinträchtigen würden, das direkt auf einen Silent State folgt. Den betreffenden Wert schätzen wir im Zuge des HMM ein. Anfangs ist der Wert für jedes Populationspaar

identisch. Er wird während der Iterationen des HMM erlernt, während die einzelnen Proben verarbeitet werden.

3.5 Emissionswahrscheinlichkeiten

Die Emissionswahrscheinlichkeit beschreibt, wie hoch die Wahrscheinlichkeit ist, dass die DNA innerhalb eines Fensters von einer bestimmten Population stammt. Es handelt sich dabei um einen komplizierten Vorgang, der im Anhang genauer beschrieben wird.

Kurz zusammengefasst beinhaltet unser Ansatz die folgenden Schritte:

I. **Definition der Fenster.** Die DNA wird innerhalb der Chromosomen in langen zusammenhängenden DNA-Strängen vererbt, die wir als Haplotypen bezeichnen. Die Arbeit mit diesen DNA-Blöcken ist oftmals ergiebiger als die Analyse einzelner Positionen innerhalb der DNA. Dies ist auch eine der großen Fortschritte im Vergleich zu den vorherigen Algorithmen, die bei AncestryDNA zum Einsatz kamen. Die genauen Haplotyp-Grenzen sind nicht bekannt. Sie unterscheiden sich von einem Menschen zu anderen. Allerdings erreichen wir eine gute Annäherung, indem wir das Genom in 1.001 kleine Fenster aufteilen. Jedes dieser Fenster deckt einen Abschnitt eines einzelnen Chromosoms ab. Die Fenster sind so klein (*beispielsweise* 3–10 Centimorgan), dass der Haplotyp (die DNA) von Vater und Mutter in jedem Fenster mit hoher Wahrscheinlichkeit aus einer einzelnen (wenngleich nicht notwendigerweise derselben) Population stammt.

II. **Erstellen der Haplotyp-Modelle.** Als Nächstes müssen wir den Haplotyp des Kunden innerhalb eines Fensters mit dem Referenzpanel abgleichen. Dabei ermitteln wir für unterschiedliche Populationen die Wahrscheinlichkeit, dass die DNA aus diesen Populationen stammt. Dabei spielen wir alle Wahrscheinlichkeiten durch: die Wahrscheinlichkeiten, dass beide DNA-Segmente eines Haplotyps aus Schweden stammen, dass einer aus Schweden und der andere aus Frankreich stammt usw. Hierfür brauchen wir aber zunächst ein Haplotyp-Modell. Deshalb bauen wir für jedes Fenster aus Hunderttausenden von Haplotypen ein *BEAGLE* -Cluster-Modell auf (Browning 2007). (Mehr Informationen dazu finden Sie im Whitepaper zum Matching.)
Der Genotyp des Kunden hat kein Phasing durchlaufen. Mit anderen Worten: Die väterlichen und mütterlichen Haplotypen sind nicht voneinander differenziert. Daher muss das Modell alle möglichen Haplotypen auf Basis des Genotyp-Satzes berücksichtigen, wobei jeder Zustand in einem Haplotyp-Cluster-Modell für ein Cluster ähnlicher Haplotypen steht.

III. **Zuweisen des Referenzpanels.** Wir möchten die Haplotyp-Cluster innerhalb unseres Modells identifizieren, die mit den einzelnen Populationen des Referenzpanels zusammenhängen.
Die geografische Herkunft der Mitglieder des Referenzpanels ist für uns ein verlässlicher Referenzpunkt. Auf dieser Grundlage können wir also die Wahrscheinlichkeit berechnen, dass der Haplotyp einer spezifischen Population durch einen bestimmten Haplotyp-Cluster dargestellt wird. Aufgrund dieser Werte errechnen wir die Emissionswahrscheinlichkeiten im genomumspannenden HMM, das der Zuweisung der ethnischen Herkunft dient.

IV. **Abgleich der Testproben mit dem Referenzpanel zur Zuweisung von Populationsbezeichnungen mittels HMM.** Hierbei berechnen wir die Wahrscheinlichkeit, dass die Haplotyp-Paare in den unterschiedlichen Fenstern der Testproben aus den Populationen des Referenzpanels stammen. In jedem Fenster können die Haplotypen aus identischen oder unterschiedlichen Populationen stammen. Die darauf basierenden Emissionswahrscheinlichkeiten werden für alle möglichen Kombinationen errechnet.

Die HMMs werden in einer Reihe bestehender Ansätze zur Einschätzung ethnischer Anteile verwendet (Maples 2013). Der Kernpunkt unseres Ansatzes besteht in Schritt III. Hierbei setzen wir in den einzelnen Fenstern besonders stichhaltige Haplotyp-Modelle ein, die wir mit Populationsbezeichnungen der Haplotypen in unserem Referenzpanel versehen. Dabei geht es uns darum, den Haplotypen in unseren Testproben eine Wahrscheinlichkeit für sämtliche Populationsbezeichnungen zuzuweisen. Erwähnenswert ist auch die Tatsache, dass sich mit unserer Methode bei der Einschätzung der ethnischen Abstammung ein hoher Durchsatz erreichen lässt. Die obigen Schritte (I) bis (III) vom Erlernen der Haplotyp-Modelle mithilfe eines großen Trainings-Datensatzes bis zur Zuweisung zu den Populationen des Referenzpanels müssen nur einmal durchgeführt werden.

3.6 HMM-Modell

Wir verwenden HMMs, weil sie effizient und effektiv alle möglichen ethnischen Verknüpfungen in allen Fenstern des Genoms berücksichtigen. Bei AncestryDNA wenden wir das HMM auf die DNA eines Kunden an, um die wahrscheinlichste Sequenz ethnischer Abstammungen entlang der DNA zu ermitteln. Technisch gesprochen nimmt der Algorithmus den Viterbi-Pfad – die Abfolge verborgener Zustände, welche die höchste Wahrscheinlichkeit ausgeben. Für die abschließende ethnische Einschätzung des Kunden werden die Anteile des Viterbi-Pfades berechnet (nach Rekombinationsdistanz gewichtet), die im Referenzpanel einer bestimmten Population zugewiesen sind. Nehmen wir einmal an, das Genom eines Kunden bestünde nur aus den Sequenzen Schweden/Schweden, Schweden/Schweden, Schweden/Schweden, Frankreich/Schweden und Frankreich/Schweden. Dann wäre der Betreffende zu 20 % Franzose und zu 80 % Schwede – vorausgesetzt, die fünf Fenster weisen dieselbe Größe auf.

Da es sich bei den Anteilen um Schätzwerte handelt, brauchen wir eine Möglichkeit, die Zuverlässigkeit der Werte zu überprüfen. Hierfür wird eine willkürliche Auswahl von 1.000 Pfaden herangezogen, die keinem Viterbi-Pfad beziehungsweise nicht der höchsten Wahrscheinlichkeitsstufe entsprechen (aber trotzdem noch eine hohe Wahrscheinlichkeit aufweisen). In jedem der 1.000 Durchläufe wird ein bestimmtes Fenster einem Populationspaar zugewiesen. Die Wahrscheinlichkeit hängt dabei von der Zuweisung des vorherigen Fensters im selben Durchlauf ab, sowie von den zuvor bestimmten Übergangs- und Emissionswahrscheinlichkeiten. Auf Basis dieser 1.000 Werte wird eine Verlässlichkeitsskala für die genannte Viterbi-Schätzung erstellt.

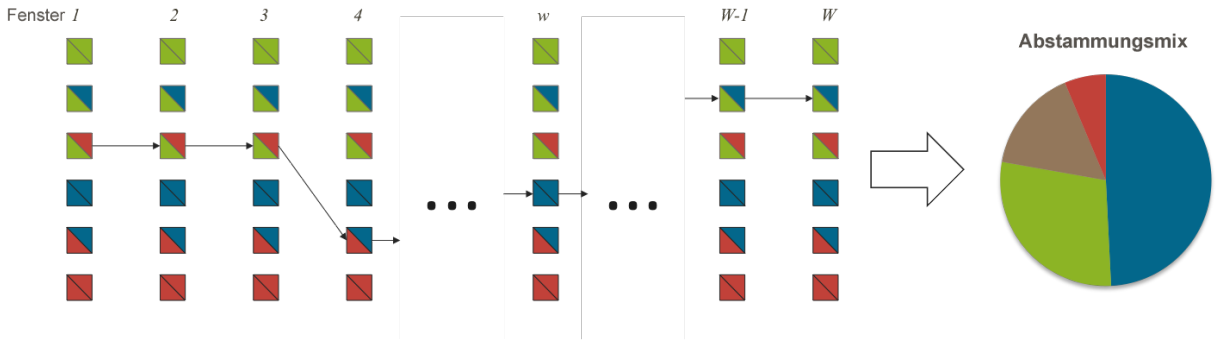


Abbildung 4.6: Darstellung des Viterbi-Pfades mittels Pfeilen anhand des HMM zur Erstellung einer ethnischen Einschätzung.

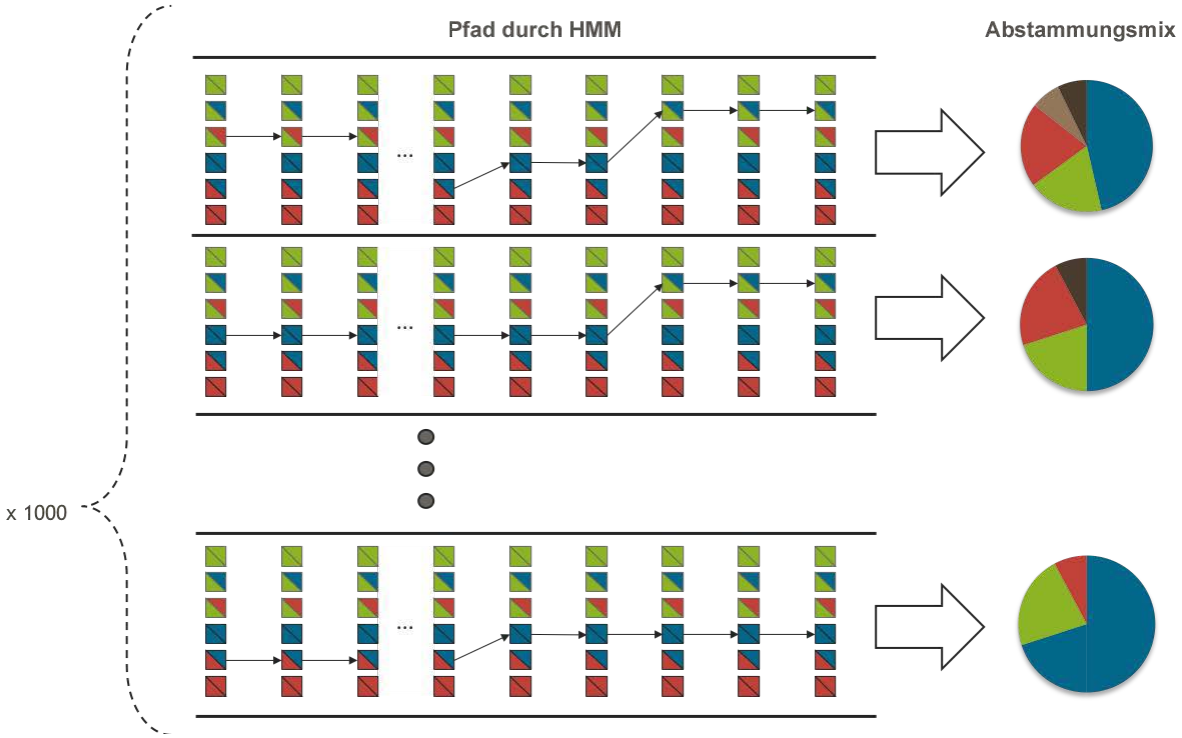


Abbildung 4.7: Darstellung unseres stochastischen Prozesses zur Identifizierung von Pfaden.

4. Bewertung der Leistungsfähigkeit bei der Einschätzung der ethnischen Abstammung

Nachdem wir sowohl den Bewertungsprozess als auch das Referenzpanel entwickelt und optimiert haben, besteht der letzte Schritt darin, zu bestimmen, wie verlässlich sich mit den beiden Hilfsmitteln im Zusammenspiel die ethnische Abstammung zuweisen lässt. Grundsätzlich lässt sich anhand strenger Tests und einer großen Bandbreite an Fallbeispielen mit bekannter ethnischer Abstammung erkennen, wie nahe wir mit unserer Methode der Wahrheit kommen.

4.1 Kreuzvalidierungsverfahren

Den Beleg für die Leistungsfähigkeit des Prozesses zur Einschätzung der ethnischen Abstammung liefern zwei unterschiedliche Testfälle mit bekanntem Resultat. Hierfür ziehen wir zwei Einzelpersonen mit nur einer genetischen Herkunft aus dem Referenzpanel heran, sowie künstlich erstellte Proben mit gemischter Herkunft. So bringen wir in Erfahrung, wie nahe wir mit dem Test an die tatsächliche ethnische Abstammung herankommen.

Einzelpersonen aus dem Referenzpanel: Die Personen, aus denen das Referenzpanel besteht, weisen alle per Definition nur eine einzige genetische Herkunft auf. Wir bewerten unseren Prozess mithilfe von 20 Durchläufen eines Kreuzvalidierungsverfahrens, bei dem wir die Einzelpersonen aus unserem Referenzpanel analysieren. Wie im Abschnitt 2.4 beschrieben suchen wir dafür 5 % der Einzelpersonen aus jeder der 43 Regionen innerhalb des Referenzpanels aus. Die übrigen 95 % aus diesen 43 Regionen nutzen wir als Referenzpanel. Diesen Vorgang wiederholen wir 20 Mal, sodass am Ende jede Einzelperson im Referenzpanel getestet wurde.

Besteht beispielsweise jede Referenzpanel-Gruppe aus 100 Personen, suchen wir uns jeweils 5 Personen aus, die wir mit unserem Algorithmus analysieren. Die übrigen 4.085 Personen dienen dabei als Referenzgruppe. Im nächsten Schritt werden weitere 5 Proben aus jeder Gruppe ausgewählt und demselben Prozess unterzogen.

Wie in Abbildung 4.1 zu sehen ist unser aktualisierter Prozess zur Einschätzung der ethnischen Abstammung in neun europäischen Regionen deutlich leistungsfähiger als unser vorheriger Prozess. Da wir Menschen mit nur einer genetischen Herkunft untersuchen, müsste ein perfekter Algorithmus in all diesen Fällen 100-prozentige Werte ausgeben. Der aktualisierte Algorithmus ist zwar nicht zu 100 % perfekt. Er kommt aber näher an die 100 % heran als die vorherige Methode. Für die meisten anderen Regionen zeichnet sich ein ähnlicher Trend ab (Daten nicht im Diagramm enthalten).

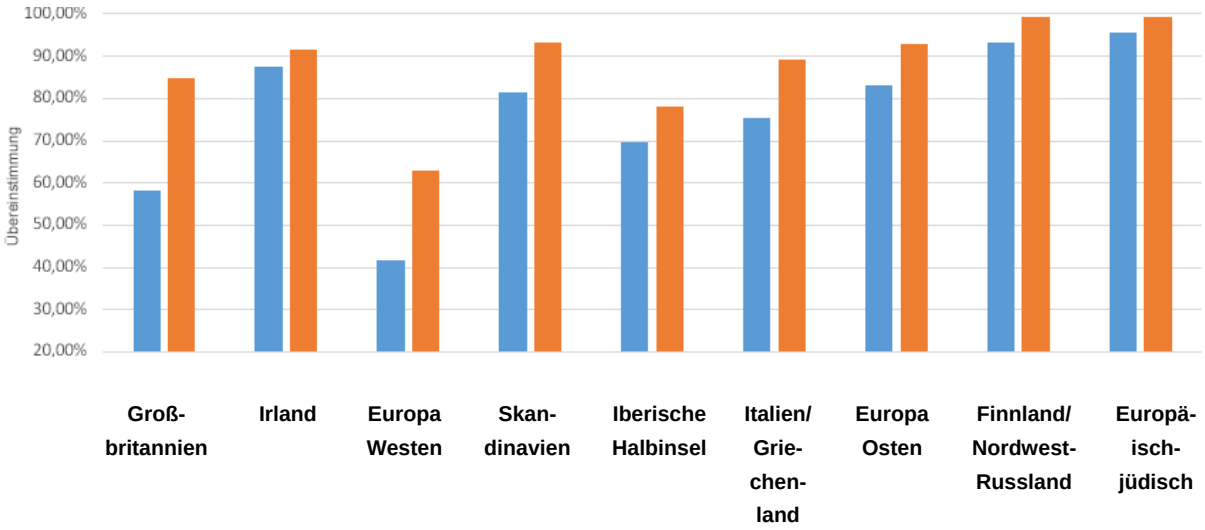


Abbildung 4.1: Der Vergleich zweier Algorithmen zur Einschätzung der genetischen Abstammung bei Personen mit nur einer Herkunft aus neun europäischen Regionen.

Hier sehen Sie einen direkten Vergleich zwischen unserem vorherigen Algorithmus zur Einschätzung der genetischen Herkunft (blau) und der aktualisierten Version (orange). Je näher der Algorithmus an die 100 % herankommt, umso effektiver lässt sich damit die ethnische Herkunft bestimmen. In diesen neun Fällen erweist sich der aktualisierte Algorithmus als im Vergleich zum vorherigen Algorithmus leistungsstärker. Wir haben die 26 Regionen der vorherigen Analyse den 43 Regionen aus der aktualisierten Analyse gegenübergestellt, um einen direkten Vergleich anzustellen.

Bei einer Analyse von Einzelpersonen unseres Referenzpanels ist beobachten, dass beim neuen Prozess die genetische Herkunft insgesamt zu durchschnittlich 78,9 % der korrekten Region zugewiesen wird (Abbildung 2.3). Einen Wert von nahezu 100 % erreichten wir bei der ethnischen Evaluierung der folgenden Gruppen:

- Europäisch-jüdisch
- Japan
- Indien Westen/Zentrum
- Kamerun, Kongo und südliche Bantu-Völker
- Polynesien
- Finnland
- Philippinen
- Südzentralafrikanische Jäger und Sammler

Für einige Regionen wie Nigeria, Spanien und das Baskenland ergeben sich weniger hohe Wahrscheinlichkeiten. Die durchschnittliche Genauigkeit der Schätzung liegt hier bei 28 %, 46 % und 54 %. Selbst in Fällen, in denen die Vorhersage nicht an die 100 % heranreicht, wird die übrige ethnische Zugehörigkeit korrekt nahegelegenen Regionen zugewiesen. So kann es vorkommen, dass Kandidaten aus Spanien die Regionen Frankreich und Portugal zugewiesen bekommen, während bei Personen aus Norwegen und Schweden die Zuweisung unter Umständen vertauscht wird (siehe Abbildung 4.2).

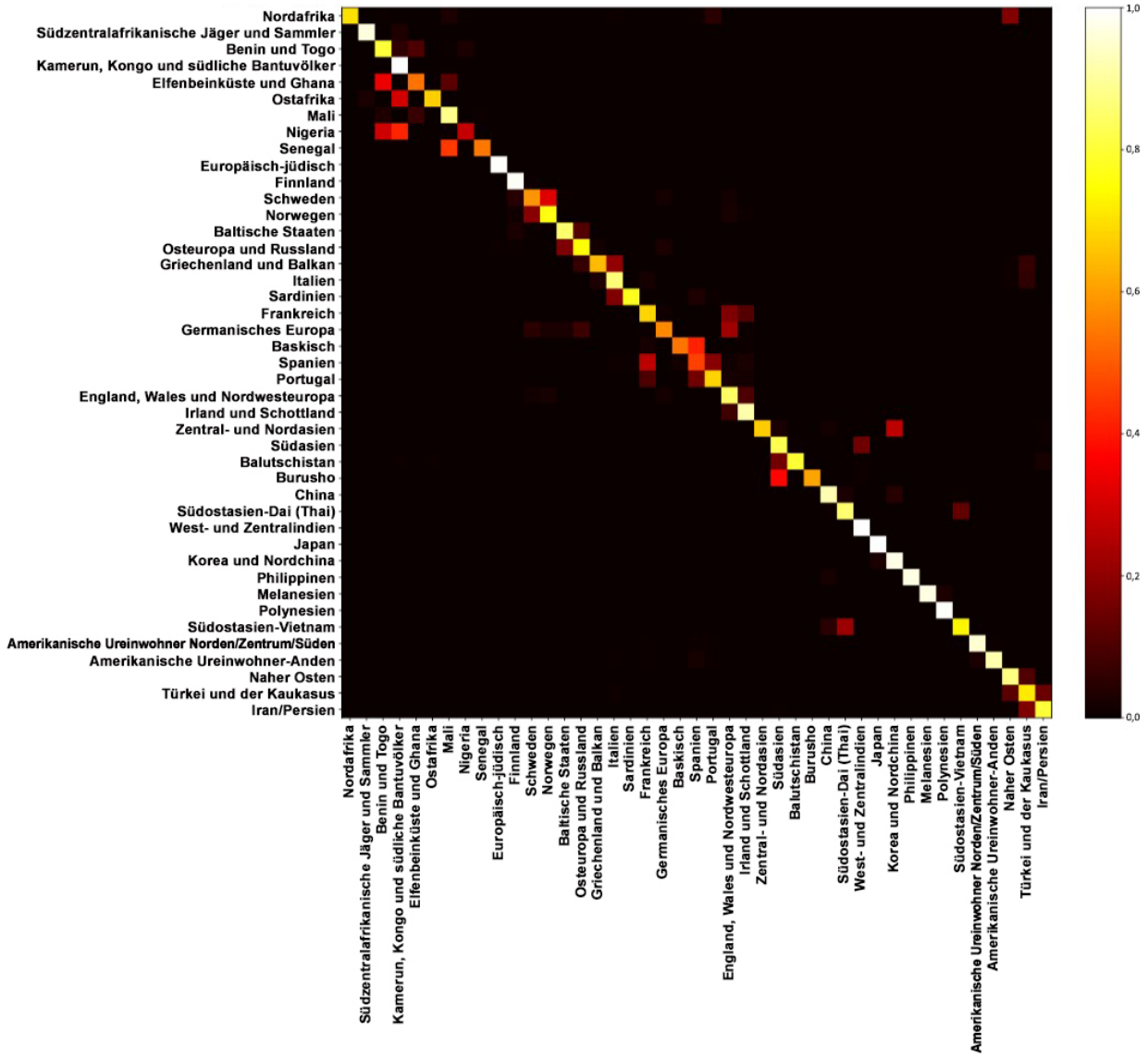


Abbildung 4.2: Durchschnittswerte für die geschätzte ethnische Herkunft von Einzelpersonen aus allen Populationen. In dieser Grafik steht jede Reihe für eine Einzelperson der aufgeführten Population. Die Spalten repräsentieren dabei die 43 möglichen ethnischen Gruppen, denen die Einzelperson zugewiesen werden kann. Das Diagramm ist so angelegt, dass sich die Übereinstimmungen zwischen Individuum und ethnischer Abstammung anhand der diagonalen Linie ablesen lassen. Wäre der Algorithmus zu 100 % perfekt, gäbe es nur weiße Kästchen auf dieser diagonalen Linie. Die weiße Farbe steht für eine 100-prozentige Herkunftswahrscheinlichkeit in Bezug auf die jeweilige Population. Alle Kästchen abseits der diagonalen Linie stehen für falsch zugewiesene Populationen. Diese Grafik zeigt außerdem, dass einzelne ethnische Gruppen miteinander verwechselt werden können. So bekommen durchschnittlich 10 % Personen mit einer zu 100 % nordafrikanischen Abstammung den Nahen Osten als ethnische Herkunft zugewiesen. Bei Menschen mit einer zu 100 % spanischen Abstammung kommt es wiederum vor, dass sie den ethnischen Gruppen Frankreichs oder Portugals zugewiesen werden.

Künstliche Profile mit gemischter ethnischer Herkunft: Wir haben die Präzision der Herkunftsbestimmung auch durch künstliche Profile mit gemischter ethnischer Herkunft geprüft. Wir haben dafür simulierte Testbeispiele aus bekannten ethnischen Mischungen zusammengestellt. Die künstlichen Mischprofile waren zu jeweils unterschiedlichen Anteilen aus zwei bis zu maximal 20 ethnischen Regionen zusammengesetzt. Da uns bei diesen Profilen die

tatsächlichen ethnischen Anteile bekannt sind, können wir damit gut die Präzision und Trefferquote für die einzelnen ethnischen Gruppen berechnen. Die Präzision und die Trefferquote sind zwei wichtige Faktoren für die Evaluierung des Analysevorgangs.

Bei der Präzision geht es darum, welcher Anteil der angegebenen ethnischen Abstammung der Wahrheit entspricht. Wird beim Analyseprozess vorhergesagt, dass eine Einzelperson zu 40 % aus Nordafrika stammt, während der tatsächliche Wert bei 30 % liegt, können wir von einer Präzision von 0,75 für die nordafrikanische ethnische Gruppe ausgehen. Mathematisch gesehen beschreibt die Präzision den Anteil der korrekt identifizierten Abstammung geteilt durch den geschätzten Wert für die betreffende Region.

Die Trefferquote beschreibt, welchen Anteil der wahren ethnischen Zugehörigkeit der Prozess identifizieren kann. Bleiben wir bei unserem Beispiel. Stellen wir uns vor, eine Einzelperson hätte zu 50 % nordafrikanische Vorfahren, während der Algorithmus aber nur 40 % vorhersagt. In diesem Fall hat der Prozess eine Trefferquote von 0,8 für die nordafrikanische Abstammung.

Tabelle 4.1: Die Präzision/Trefferquote für die einzelnen Regionen, berechnet auf Basis der abgeleiteten ethnischen Abstammung künstlicher Profile mit gemischter Herkunft.

Region	**Präzision**	**Trefferquote**
Nordafrika	0,90	0,67
Südzentralafrikanische Jäger und Sammler	0,91	0,98
Benin und Togo	0,73	0,89
Kamerun, Kongo und südliche Bantu-Völker	0,87	0,99
Elfenbeinküste und Ghana	0,79	0,61
Ostafrika	0,97	0,71
Mali	0,82	0,93
Nigeria	0,81	0,26
Senegal	0,89	0,53

Europäisch-jüdisch	0,93	0,97
Finnland	0,84	0,97
Schweden	0,55	0,64
Norwegen	0,62	0,80
Baltische Staaten	0,38	0,90
Osteuropa und Russland	0,86	0,79
Griechenland und Balkan	0,54	0,53
Italien	0,82	0,68
Sardinien	0,63	0,77
Frankreich	0,76	0,63
Europäisch-germanisch	0,79	0,59
Baskenland	0,69	0,55
Spanien	0,65	0,30
Portugal	0,86	0,44
England, Wales und Nordwesteuropa	0,58	0,82
Irland und Schottland	0,55	0,91
Asien Zentrum/Norden	0,98	0,61

Asien Süden	0,95	0,88
Belutschistan	0,85	0,73
Hunzukuc	0,91	0,57
China	0,93	0,88
Asien Südosten – (Thai)	0,69	0,86
Indien Westen/Zentrum	0,60	0,99
Japan	0,92	0,99
Korea und Nordchina	0,71	0,91
Philippinen	0,99	0,96
Melanesien	0,98	0,98
Polynesien	0,98	0,99
Südostasien – Vietnam	0,87	0,76
Amerikanische Ureinwohner – Norden/Zentrum/Süden	0,98	0,96
Amerikanische Ureinwohner – Anden	0,96	0,95
Naher Osten	0,87	0,72
Türkei und Kaukasus	0,29	0,76
Iran/Persien	0,89	0,67

Wir haben festgestellt, dass die Präzision und die Trefferquote bei den meisten Herkunftsregionen über 60 % liegt. Insbesondere für die folgenden Regionen sind die Ergebnisse sehr stichhaltig:

- Philippinen
- Polynesien
- Japan
- Amerikanische Ureinwohner Anden
- Amerikanische Ureinwohner Norden/Zentrum/Süden
- Europäisch-jüdisch
- Südzentralafrikanische Jäger und Sammler
- Kamerun, Kongo und südliche Bantu-Völker

Für manche Regionen wie Nigeria, Spanien und Portugal ist die Trefferquote mit jeweils 26, 30 und 44 % relativ niedrig. Bei anderen Regionen wie der Türkei und dem Kaukasus sowie den baltischen Staaten liegt die Präzision mit jeweils 29 und 38 % eher am unteren Ende des Spektrums. Die niedrige Trefferquote mancher Regionen ist größtenteils darauf zurückzuführen, dass die ethnische Abstammung dieser Regionen teilweise angrenzenden Regionen zugeordnet wird. Dadurch ergeben sich zu niedrige Wahrscheinlichkeitswerte und somit auch eine schlechte Trefferquote. Die niedrige Präzision mancher Regionen ist mit hoher Wahrscheinlichkeit darauf zurückzuführen, dass Teile angrenzender Regionen falsch zugewiesen werden. Dadurch ergeben sich zu hohe Wahrscheinlichkeitswerte und eine niedrige Präzision.

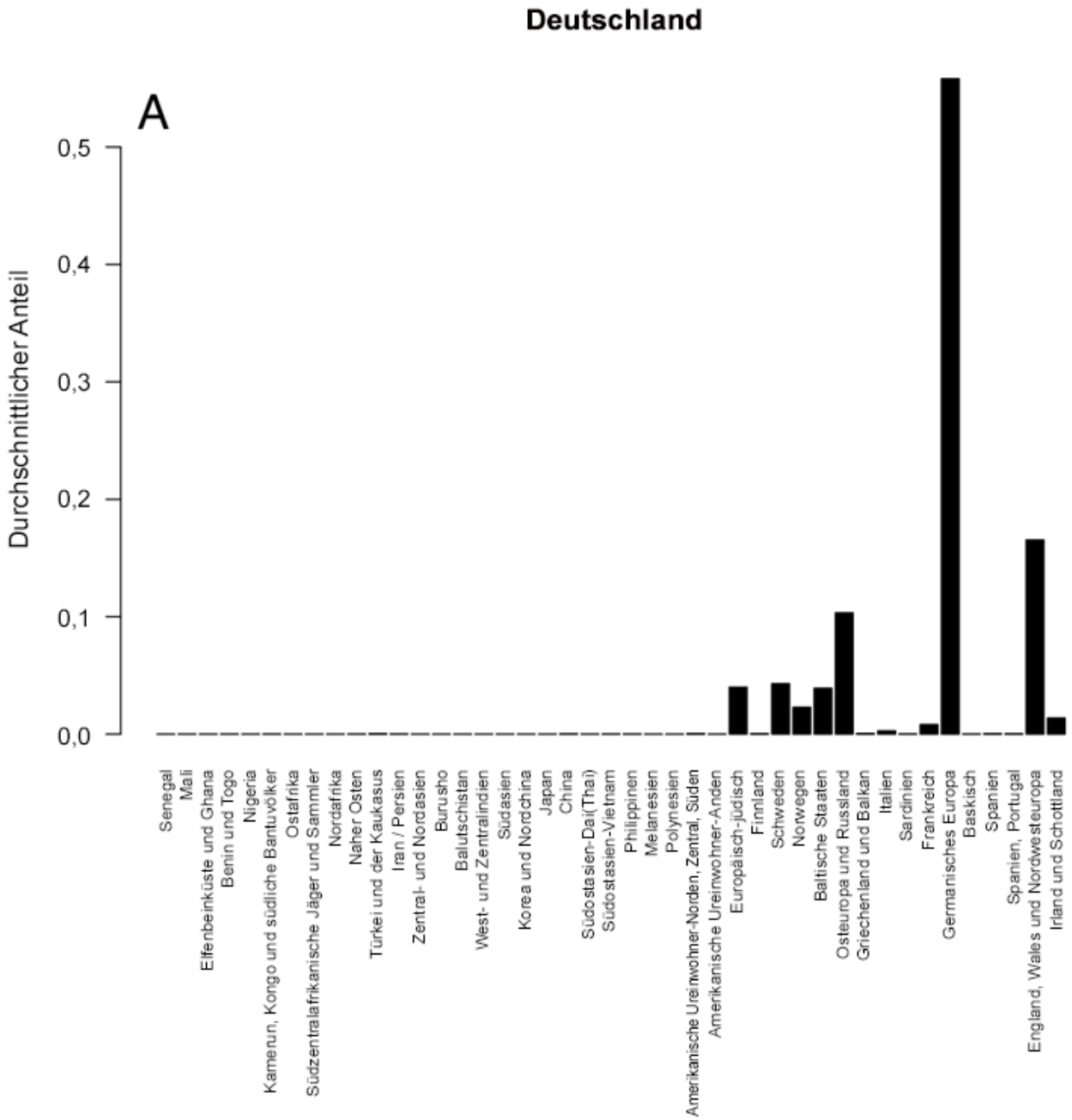
4.2 Einschätzung von Regionen

Indem wir Proben von Einzelpersonen analysieren, deren bekannte Vorfahren aus nur einer der durch uns definierten Regionen stammen, können wir Überlappungen zwischen verschiedenen Regionen bestimmen und dadurch unseren Kunden bei der Interpretation ihrer Ergebnisse helfen. Um betreffende Einzelpersonen ausfindig zu machen, verwenden wir durch Kunden erstellte Stammbäume. Wir halten Ausschau nach Personen, deren Vorfahren ausschließlich aus ein und demselben Land stammen. Im Idealfall handelt es sich dabei um Menschen, deren Großeltern ausschließlich aus einem Land kommen. Aufgrund der niedrigen Zahlen einiger Länder ziehen wir manchmal auch die Eltern oder sogar den Geburtsort des Kunden heran.

Bei Kunden, deren Stammbäume innerhalb eines Landes weit in die Vergangenheit reichen, sollten sich auch hohe Übereinstimmungen mit der ethnischen Gruppe des betreffenden Landes ergeben – was sich in der Praxis allgemein bewahrheitet. Abbildung 4.3A zeigt beispielsweise die durchschnittlichen genetischen Zuordnungen von 1.911 Kunden, bei denen alle vier Großeltern in Deutschland geboren sind. Wie Sie sehen, werden die Kunden von ihrer genetischen Abstammung her größtenteils Deutschland zugeordnet. Allerdings kommen auch kleine aber doch signifikante Anteile anderer Regionen vor. Durch diese Analysen können wir unsere Einschätzungen der ethnischen Herkunft leichter mit den Erwartungen zusammenbringen.

Allerdings erhält nicht jeder Kunde eine Einschätzung, die den Durchschnittswerten eines Landes entspricht. Daher lohnt sich oft ein Blick auf die einzelnen Resultate, um die Gründe dafür zu ermitteln. Abbildung 4.3B

zeigt die durchschnittliche Einschätzung der ethnischen Abstammung bei Menschen, deren vier Großeltern alle aus Japan stammen. Interessanterweise entfallen ein paar Prozent der Einschätzung auf die Regionen England, Wales und Nordwesteuropa sowie Irland und Schottland. Diese wenigen Prozent verteilen sich unter den Kandidaten der Analyse nicht gleichmäßig. Stattdessen gibt es einige Einzelpersonen, bei denen in der Einschätzung eine größtenteils europäische Abstammung ermittelt wird. Das deutet darauf hin, dass die betreffenden Personen von europäischen Immigranten abstammen, die nach Japan ausgewandert sind.



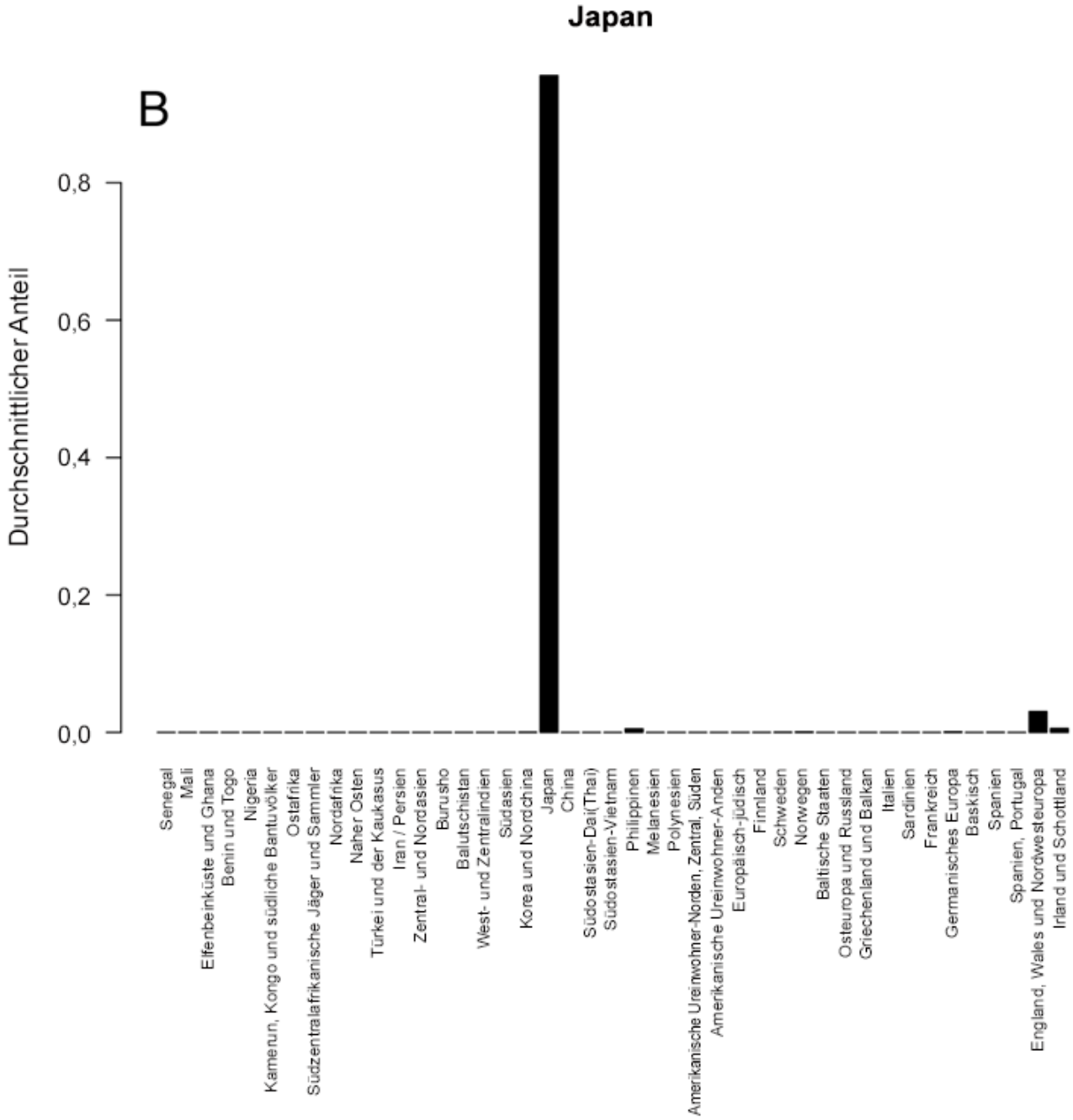


Abbildung 4.3 Die durchschnittliche Einschätzung der ethnischen Abstammung auf Grundlage des Geburtsortes der Großeltern. *Die durchschnittliche Einschätzung der ethnischen Abstammung bei Menschen, deren vier Großeltern alle im selben Land geboren sind. (A) Deutschland, (B) Japan*

Karten wie die in Abbildung 4.4 verwenden wir unter anderem auch, um sicherzustellen, dass die Einschätzung der ethnischen Abstammung geografisch sinnvoll ist. Wenn wir sehen, wie die Einschätzungen der ethnischen Abstammung geografisch verteilt sind, kann uns das in vielen Fällen helfen, auf den ersten Blick überraschende Ergebnisse richtig einzuordnen. So zeigt beispielsweise Abbildung 4.4,

dass die französische Bretagne einen hohen Anteil an Vorfahren aus Irland und Schottland aufweist. Das ergibt auch Sinn. Schließlich deutet die Zuweisung Irland und Schottland auf keltische Vorfahren hin, die sowohl in der Bretagne als auch in Irland und Schottland verbreitet waren. Dieser Sachverhalt wird auch in der bretonischen Sprache deutlich, die in der Region verbreitet ist. Es handelt sich dabei ebenfalls um eine keltische Sprache. Die höheren ethnischen Anteile an irischen und schottischen Vorfahren in Wales sind wahrscheinlich ebenfalls das Ergebnis keltischer Migrationsbewegungen in diese Region.

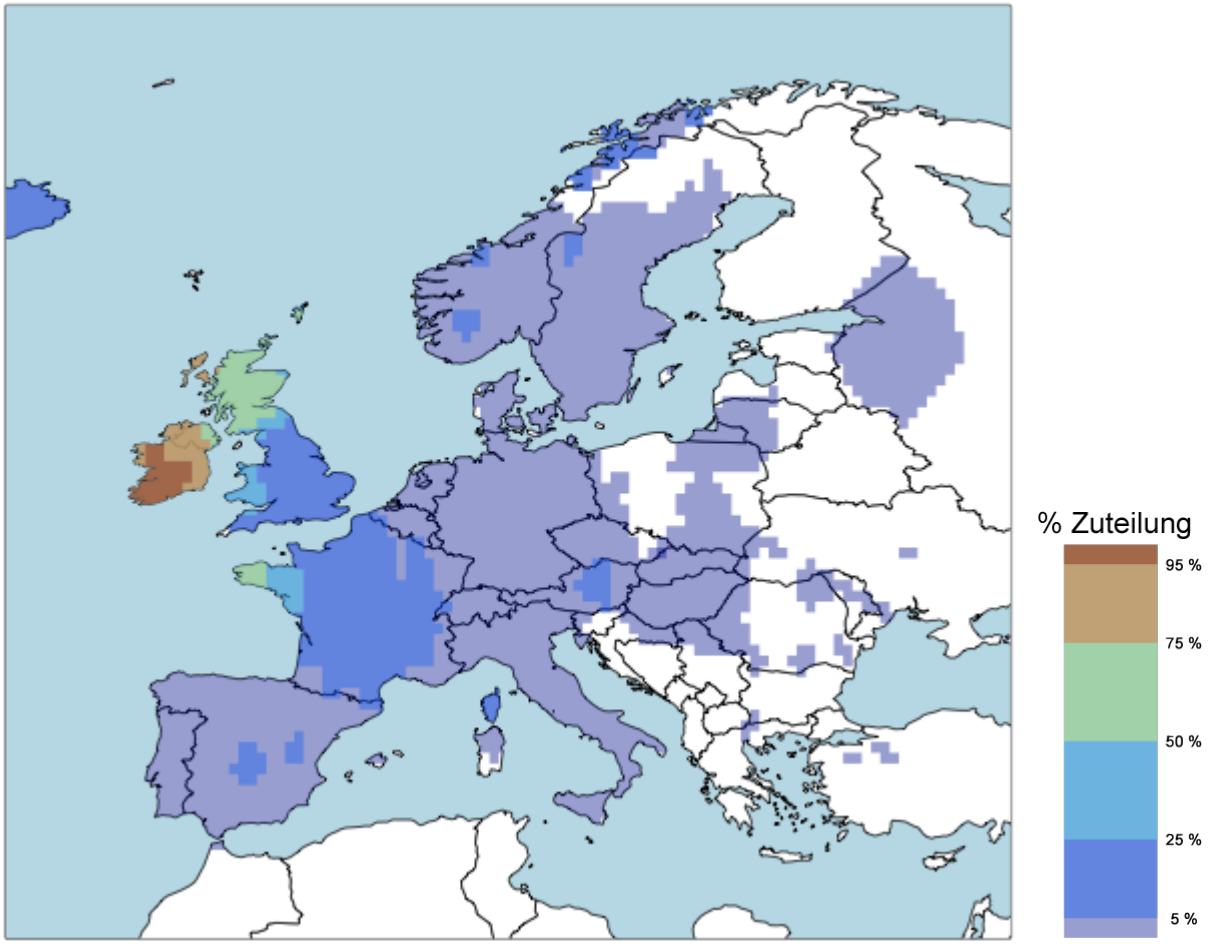
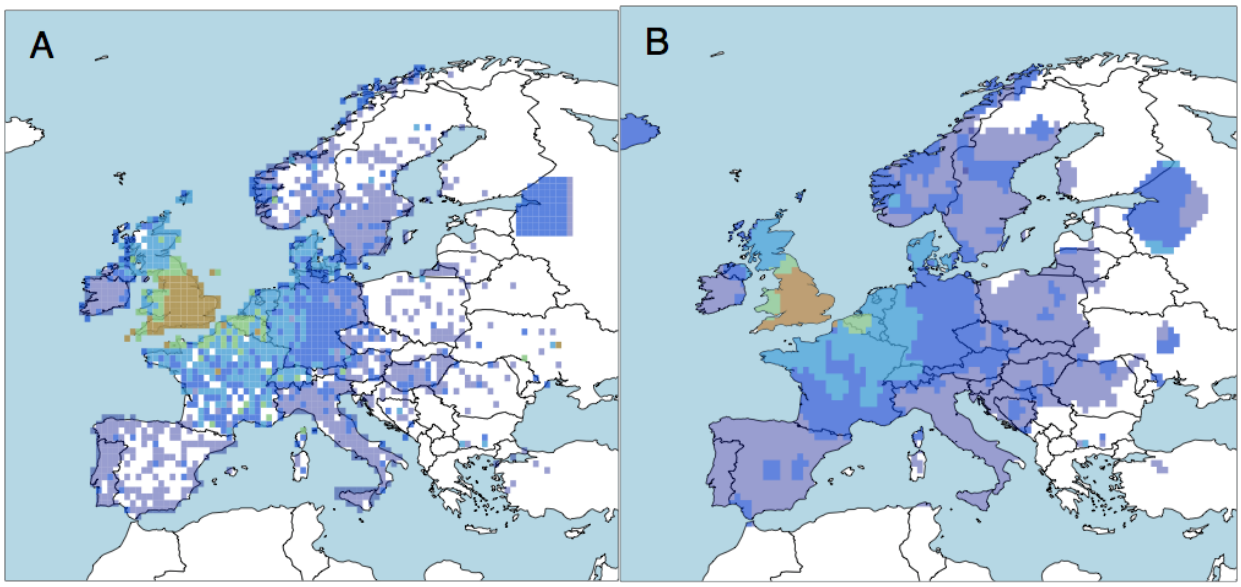


Abbildung 4.4 ***Die Karte mit den durchschnittlichen Einschätzungen für Irland und Schottland.*** *Die hohen Werte außerhalb von Irland und Schottland, Wales und der Bretagne (in Hellblau und Grün dargestellt) sind ein plausibler Indikator für Migrationsbewegungen des keltischen Volkes in der Menschheitsgeschichte.*

Diese Analysen helfen uns, die genetische Vielfalt dieser Regionen besser zu verstehen. Auf dieser Grundlage können wir die Ergebnisse unseren Kunden besser erklären. Kommen beispielsweise die Vorfahren eines Kunden aus Deutschland, ist auch ein gewisser Prozentsatz an ethnischen Wurzeln in angrenzenden Regionen wahrscheinlich. Diese Analysen unterstützen uns darüber hinaus bei der Planung der nächsten Updates, mit denen wir die Einschätzung der ethnischen Herkunft optimieren.

4.3 Regionaler Polygon-Aufbau

Wir verwenden 43 globale Populationen für unser Referenzpanel. Deshalb unterteilen wir den Erdball in 43 sich gegenseitig überlappende geografische Regionen/Gruppen. Jede Region steht für eine Population mit einem einzigartigen genetischen Profil. Zur Grenzziehung zwischen den Regionen verwenden wir wo immer möglich die bekannten geografischen Standorte unserer Proben. Abbildung 4.5 zeigt exemplarisch, wie anhand der Daten die regionalen Polygone definiert werden.



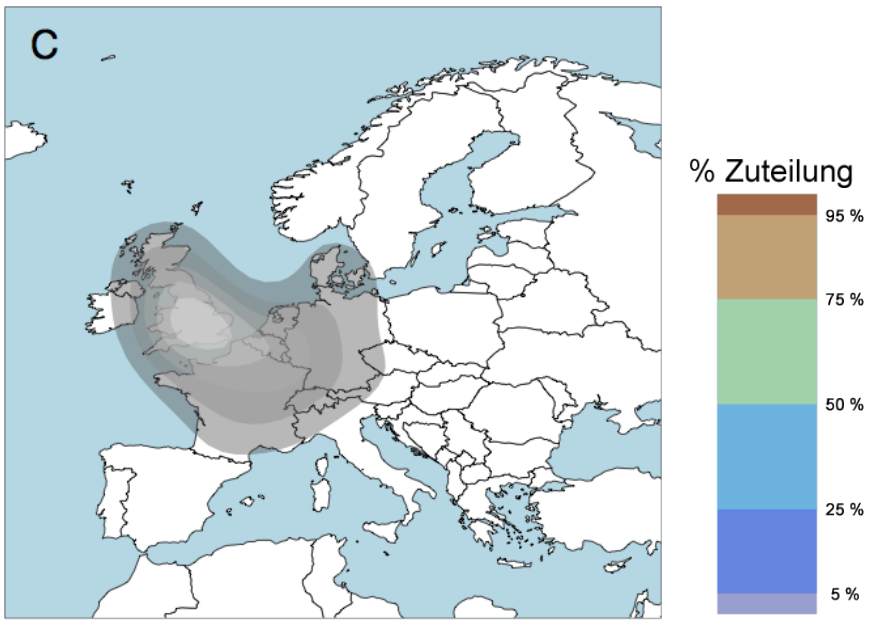
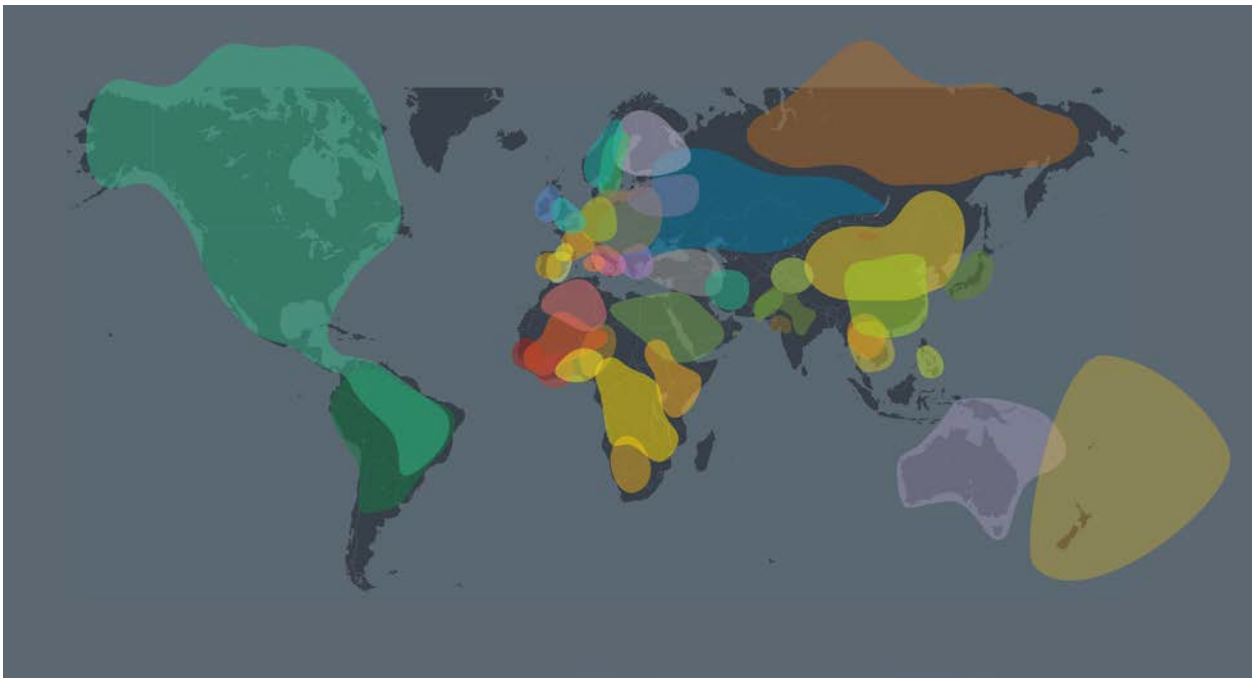


Abbildung 4.5: Der Einsatz geografischer Probenstandorte zur Definition regionaler Polygone. *Grafik A zeigt die vorhergesagte Verteilung der ethnischen Gruppe England, Wales und Nordwesteuropa für einen Probensatz aus geografischen Daten. Die Proben werden Feldern zugewiesen, die der Größe eines jeweils halben Längen- und Breitengrades entsprechen und auf dem durchschnittlichen Geburtsstandort der Großeltern basieren. Die Farben der einzelnen Quadrate auf der Karte zeigen die durchschnittlichen ethnischen Anteile der Proben aus den einzelnen Feldern an, bezogen auf die ethnische Gruppe England, Wales und Nordwesteuropa. In Grafik B wurden die fehlenden Regionen aufgefüllt und die Resultate geglättet. Hierfür wurden die Daten mit einem Kernel-Glätter bearbeitet. Auf dieser Basis wurden die Umrisse erstellt, welche für die ermittelten Regionen der Vorfahren stehen, wie sie in Grafik C angezeigt werden.*

In Abbildung 4.5A ist ersichtlich, welcher ethnische Anteil bei einer Teilmenge der Referenzproben mit bekanntem geografischem Standort auf die Region England, Wales und Nordwesteuropa entfällt. Abbildung 4.5B ist eine Darstellung, bei der Werte zugewiesen und Lücken gefüllt wurden. Außerdem wurden hier Glättungsmethoden angewendet, um das Diagramm gleichförmiger zu gestalten. Aus dem Diagramm ist klar ersichtlich, dass es einen Verlauf der ethnischen Zugehörigkeit innerhalb der Region gibt, dessen Ausgangspunkt in England liegt und dessen Konzentrationsstufe zu den umliegenden Regionen hin abnimmt. So liegt die nächstkleinere Konzentrationsstufe – im Bild durch die Farbe Grün angezeigt – in englischen Nachbargebieten wie Wales, Frankreich und Belgien. Die nächste Verlaufsstufe der ethnischen Zugehörigkeit ist Lila dargestellt. Die entsprechenden Grenzen reichen schon bis nach Italien, in die Schweiz, nach Schweden und Irland. Die Daten werden soweit möglich direkt mit den regionalen Grenzen abgestimmt (Abbildung 4.5C), die als Teil der AncestryDNA-Produktpräsentation auf den Karten erscheinen.

Die Polygone dieser Darstellung erscheinen als ineinander verschachtelte Regionen mit unterschiedlich intensiven Schattierungen. Die rotbraunen Regionen stehen für die Gebiete mit der höchsten durchschnittlichen Zuweisung. Dies sind die wahrscheinlichsten tatsächlichen Herkunftsorte der Vorfahren eines Kunden. In den blau/lila eingefärbten Regionen ist das Durchschnittsniveau niedriger. Sie stehen für andere mögliche Herkunftsorte mit geringerer Wahrscheinlichkeit. Zu jedem Polygon-Satz gehört eine detaillierte Beschreibung der Geschichte der betreffenden Region.

Auf der folgenden Karte sind die Polygone für sämtliche Populationen dargestellt, auf Basis der zweiten Polygon-Stufe. Der Aufbau entspricht der oben angegebenen Beschreibung.



4.4 Ausgabe von Ungewissheiten bei den Schätzwerten

Die Bewertung der ethnischen Herkunft ist keine präzise Wissenschaft. Der Prozentsatz, der AncestryDNA einem Kunden ausgibt, ist der wahrscheinlichste Prozentsatz innerhalb eines bestimmten Prozentbereichs. In diesem Abschnitt gehen wir näher darauf ein, wie wir diesen Prozentbereich berechnen. Wir möchten noch einmal daran erinnern, dass wir bei AncestryDNA unser Angebot anhand unserer neusten Arbeiten ausbauen, um unseren Kunden immer präzisere Ergebnisse liefern zu können.

Dazu ein konkretes Beispiel: Nehmen wir an, ein Kunde stammt zu 40 % aus der Region England, Wales und Nordwesteuropa, bei einem Konfidenzintervall von 30–60 %. Das bedeutet, derjenige stammt sehr wahrscheinlich zu 40 % aus England, Wales und Nordwesteuropa. Der Anteil dieser ethnischen Herkunft kann aber zwischen 30 und 60 Prozent schwanken.

Wie in Abschnitt 3 bereits erwähnt lassen wir einen genomumspannenden Viterbi-Algorithmus über die DNA-Probe des Kunden laufen. Das Resultat entspricht unserer Einschätzung der wahrscheinlichsten ethnischen Herkunft eines Kunden. Auf dieser Grundlage können wir Übergangswahrscheinlichkeiten erstellen, anhand derer wir dann neue Einschätzungen der ethnischen Abstammung generieren, die zwar wahrscheinlich sind, aber nicht die größte Wahrscheinlichkeit aufweisen. Die Schwankungsbreite der Wahrscheinlichkeit basiert auf 1.000 derartiger Probe-Pfade. Besteht in einem Fenster beispielsweise eine 80-prozentige Wahrscheinlichkeit für die Herkunftsregion England und Wales, liegt gleichzeitig eine

20-prozentige Wahrscheinlichkeit für eine andere Region vor. Das Konfidenzintervall deckt diese niedrigeren Wahrscheinlichkeiten ab, die sich für die Kunden-DNA ergeben können.

Wir haben einen Weg gefunden, mithilfe der 1.000 Probeschätzungen das Konfidenzintervall der Viterbi-Berechnung zu bestimmen, die wir an den Kunden weitergeben. Bei der Entwicklung dieses Ansatzes bestand unser Ziel darin, eine größtmögliche Wahrscheinlichkeit dafür zu erreichen, dass der ausgegebene Prozentbereich tatsächlichen ethnischen Anteilen entspricht (**Trefferquote**). Gleichzeitig wollten wir damit auch die **Präzision** verbessern, indem wir den Prozentbereich möglichst klein halten.

Wir verwenden die mittlere Abweichung und die Standardabweichung der 1.000 Probeschätzungen und berechnen auf deren Grundlage ein Konfidenzintervall für die Viterbi-Berechnung. Bei der Berechnung des Prozentbereichs berücksichtigen wir den Wert der Viterbi-Berechnung und die Population, für die wir den Prozentbereich ermitteln.

Den Prozess zur Bestimmung des Prozentbereichs können wir anhand derselben gemischten Profile testen, die wir bereits im Kreuzvalidierungsverfahren verwendet haben. So stellen wir fest, wie oft der bekannte ethnische Anteil innerhalb des Bereichs korrekt errechnet wird. Mit anderen Worten: Wir stellen fest, wie oft der Bereich sich mit der bekannten ethnischen Herkunft deckt. Wie wir festgestellt haben, liefert der Algorithmus bei manchen Populationen bessere Ergebnisse als bei anderen. Dadurch, dass wir die tatsächliche ethnische Herkunft kennen, können wir jedoch für jede Population spezifische Korrekturfaktoren mit berücksichtigen. Dadurch maximiert sich die Wahrscheinlichkeit, dass die tatsächliche ethnische Abstammung in den Prozentbereich fällt.

5. Zukünftige Verbesserungen zur Einschätzung der ethnischen Abstammung

Wir bei AncestryDNA sind besonders stolz auf die Verbesserungen, die diese aktualisierte Version zur Einschätzung der ethnischen Abstammung mit sich bringt. Trotzdem arbeiten wir daran, das Produkt stetig weiter zu verbessern. Es stehen immer neue Daten und methodische Ansätze zur Verfügung. Außerdem gibt es ständig neue Entdeckungen zu bestimmten humangenetischen Variationsmustern. All diese Faktoren bergen ein großes Entwicklungspotenzial für die Zukunft.

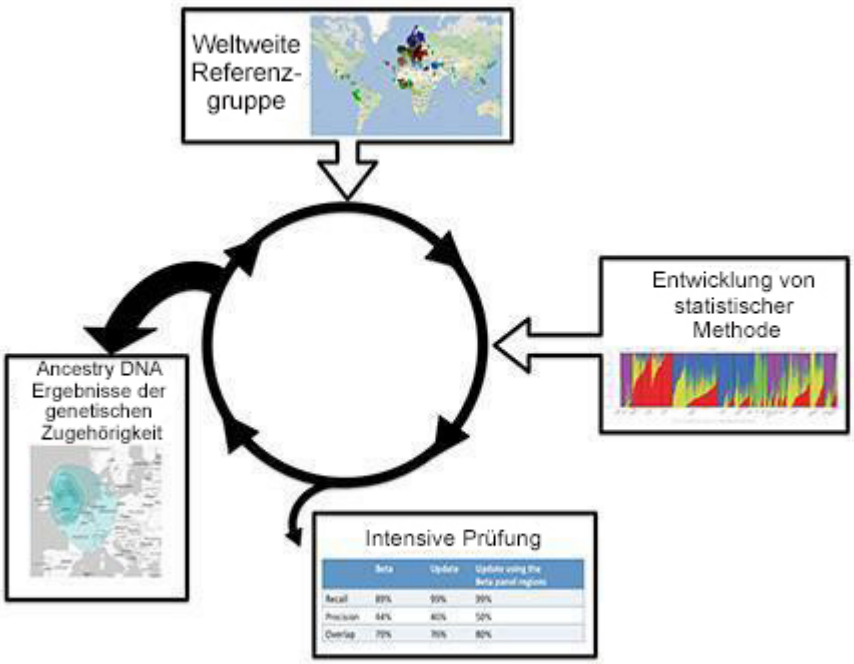


Abbildung 5.1: *Verbesserungszyklus zur Einschätzung der ethnischen Herkunft.*

Jeder der oben geschilderten Schritte stellt einen wesentlichen Teil der Entwicklung unserer Methode zur Einschätzung der ethnischen Herkunft dar. Aktuell arbeiten wir daran, das globale Referenzpanel in zukünftigen Updates zur Einschätzung der ethnischen Abstammung auszuweiten. Wir haben bereits damit begonnen, für ein kommendes Update Proben zu analysieren und die Genotypen zu ermitteln. Wir erwarten, dass die neue Version noch bessere Schätzungen ausgibt. Eine weitere Initiative beschäftigt sich mit der Diversität der Populationen. Wir sammeln DNA-Proben von bislang untervertretenen Regionen in aller Welt, um die Zahl der Gruppen auszuweiten, die wir unseren Kunden anbieten können.

Gleichzeitig arbeiten wir auch an der Verbesserung der Algorithmen, die wir zur Einschätzung der ethnischen Herkunft verwenden. In kommenden Updates werden wir möglicherweise unter anderem auch unseren statistischen Ansatz überarbeiten, um die in genetischen Daten enthaltenen Informationen umfassender nutzen zu können und noch mehr Informationen zur Populationsgeschichte zu enthüllen. Dabei führen wir auch weitere Tests und Analysen durch, bei denen wir genauso gründlich vorgehen wollen wie bei den oben beschriebenen Untersuchungen. Diese Tests bilden die Grundlage für unsere fokussierten Weiterentwicklungen und helfen uns dabei, unsere Methode noch zu verfeinern.

Jedes neue Update zur Einschätzung der genetischen Abstammung ist ein weiterer Schritt in die Zukunft. Unsere Kunden bekommen dadurch ein immer genaueres Bild ihrer ethnischen Wurzeln und ihrer genetischen Abstammung. Wir hoffen, dass Sie als Kunde diese zukünftigen Entwicklungen genauso gespannt mitverfolgen wie das gesamte Team hier bei AncestryDNA.

6. Literaturverzeichnis

- D.H. Alexander, J. Novembre, and K. Lange: Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19: 1655–1664 (2009).
- [Browning, 2007] S. R. Browning and B. L. Browning: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 81: 1084–1096 (2007).
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL: A human genome diversity cell line panel. Science 12; 296 (5566): 261–2 (April 2002).
- Cavalli-Sforza LL: The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 6(4): 333–40 (April 2005).
- International HapMap Consortium: A haplotype map of the human genome. Nature 437 (7063): 1299–1320 (Oktober 2005).
- International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, et al.: A second generation human haplotype map of over 3.1 million SNPs. Nature 449 (7164): 851–61 (Oktober 2007).
- Jackson, J.E.: A User s Guide to Principal Components. John Wiley & Sons, New York, 2003.
- [Maples 2013] Maples, Brian K., et al.: "RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference." *The American Journal of Human Genetics* 93.2: 278–288 (2013).
- [Noto 2014] K. Noto, Y. Wang, M. Barber, J. Granka, J. Byrnes, R. Curtis, N. Myres, C. Ball, and K. Chahine: Underdog: A fully-supervised phasing algorithm that learns from hundreds of thousands of samples and phases in minutes (2014). Invited Talk at the American Society of Human Genetics (ASHG) annual meeting, San Diego, CA (Oktober 2014).
- [Noto 2015] K. Noto, Y. Wang, M Barber, J. Byrnes, P. Carbonetto, R. Curtis, J. Granka, E. Han, A. Kermany, N. Myres, C. Ball, and K. Chahine: *Polly*: A novel approach for estimating local and global admixture proportion based on rich haplotype models. (2015). Invited Talk at the American Society of Human Genetics (ASHG) annual meeting, Baltimore, MD (Oktober 2015).
- Patterson N, Price AL, Reich D: Population Structure and Eigenanalysis. PLoS Genet 2(12): e190 (2006).
- Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 155(2): 945-59 (Juni 2000).
- Purcell, S: PLINK v1.07. http://pngu.mgh.harvard.edu/purcell/plink/
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–75 (September 2007).
- [Rabiner, 1989] Rabiner L.: A tutorial on hidden Markov models and selected applications in speech recognition.*Proceedings of the IEEE*, 77(2):257 286 (1989).