THÜNEN

Exploring and validating statistical reliability in forensic conservation genetics

Jutta Buschbom

Thünen Report 63

Bibliografische Information:
Die Deutsche Nationalbibliothek verzeichnet diese Publikationen in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information:
The Deutsche Nationalbibliothek (German National Library) lists this publication in the German National Bibliography; detailed bibliographic data is available on the Internet at www.dnb.de

Bereits in dieser Reihe erschienene Bände finden Sie im Internet unter www.thuenen.de

Volumes already published in this series are available on the Internet at www.thuenen.de

Zitationsvorschlag – *Suggested source citation:*
Buschbom J (2018)
Exploring and validating statistical reliability in forensic conservation genetics. Braunschweig: Johann Heinrich von Thünen-Institut, 104 p, Thünen Rep 63,
DOI:10.3220/REP1539879578000

THÜNEN

Thünen Report 63

Herausgeber/Redaktionsanschrift – *Editor/address*

Johann Heinrich von Thünen-Institut
Bundesallee 50
38116 Braunschweig
Germany

thuenen-report@thuenen.de
www.thuenen.de

ISSN 2196-2324
ISBN 978-3-86576-191-0
DOI:10.3220/REP1539879578000
urn:nbn:de:gbv:253-201810-dn060172-6

Exploring and validating statistical reliability in forensic conservation genetics

Jutta Buschbom

Thünen Report 63

VERANTWORTLICHE AUTORIN:

Jutta Buschbom
bis 30.09.2018:
Thünen-Institute of Forest Genetics
Sieker Landstrasse 2
22927 Grosshansdorf, Germany
Phone: +49 (0)4102-696-0
Fax: +49 (0)4102 696 200

Email: buschbom@posteo.de
fg@thuenen.de

Thünen Report 63

Braunschweig/Germany, October 2018

Abstract

Safeguarding biological diversity from evolutionary lineages to ecosystems is a major undertaking for humanity. Forensic conservation genetics for the protection of wild flora and fauna aims to provide statistical inference tools and services for the enforcement of local to global conservation and management strategies. This paper reviews statistical criteria that provide insight into and assess the reliability of conclusions drawn from statistical inference. The translation of these fundamental criteria into practice is illustrated with applications from evolutionary and forensic genetics, specifically focusing on the inference of geographic origin using population assignment approaches.

The key concept that end-users of statistical results (e. g., conservation activists, certification participants, courts and juries evaluating a proposed expert conclusion) need to understand and take into account, is the statistical reliability of an inferred estimate. Its reliability sets a result into an appropriate context, defines the scope (space of applicability) for which it is valid, and thus provides perspective. In statistical terms, this context corresponds to the probability density surfaces of sample, parameter and result spaces. As measures of dispersion (e. g., size, range, variance), the reviewed ancillary statistical criteria convey the structure and characteristics of these surfaces. They make accessible the more insight into reliability, the more their scope is as continuous and wide as possible. Validity (convergence and consistency, measured by precision and accuracy, supplemented by robustness and congruence), efficiency ("speed") and sufficiency ("power"), as well as, model specification (definition, selection and assessment) and hypothesis falsification quantify how confident one can be in the parameter values, support measures, predictive data and test decisions returned by statistical methods. In this way, statistical reliability forms the subject, process and goal of the statistical validation of reference datasets and inference approaches.

This review introduces each ancillary criterion, and discusses general strategies for practical application, as well as, available implementations towards genetic population assignment. It points out the fundamental importance of genome-wide sequence information for reference samples from across the distribution range in non-model organisms. Such reference datasets provide the information-rich genomic data that is required for the development of sufficient and versatile statistical inference approaches. Together they form the prerequisites for arriving at accurate, decisive and reliable tools for conservation, management and law enforcement. The validation of their quality characteristics lays the basis for a widespread practical acceptance and use of such tools also in non-model organisms.

Keywords: conservation, management and enforcement; forensic genetics for wild flora and fauna; statistical reliability; ancillary statistical criteria; statistical validation

Zusammenfassung

Der Schutz biologischer Vielfalt von evolutionären Linien bis hin zu Ökosystemen ist eines der wichtigsten Unterfangen der Menschheit. Forensische Erhaltungsgenetik zum Schutz wilder Flora und Fauna hat das Ziel Werkzeuge und Dienstleistungen für statistische Schlussfolgerungen zur Ausübung und Durchsetzung von lokalen bis globalen Erhaltungs- und Bewirtschaftungsstrategien bereitzustellen. Die vorliegende Arbeit gibt einen Überblick über statistische Kriterien, welche Einblick in und eine Bewertungsgrundlage für die Verlässlichkeit von Schlussfolgerungen geben, welche auf Grund von statistischen Inferenzverfahren gezogen wurden. Die Überführung dieser grundlegenden Kriterien in die Anwendung wird durch ihren Einsatz in der Evolutions- und Erhaltungsgenetik illustriert, wobei der Fokus speziell auf der Rekonstruktion der geographischen Herkunft durch Zuordnungsverfahren auf Populationsebene liegt.

Das Hauptkonzept, welches Endnutzer statistischer Ergebnisse (z. B., eine Expertenaussage abzuwägen habende Naturschutz-AktivistInnen, ZertifizierungsteilnehmerInnen, Gerichte oder die Öffentlichkeit) verstehen und beachten müssen, ist die statistische Verlässlichkeit eines gezogenen Rückschlusses. Die Verlässlichkeit setzt ein Ergebnis in seinen angemessenen Kontext, definiert seinen Gültigkeitsbereich (seinen Anwendungsraum) und gibt ihm somit Perspektive. In statistischen Begriffen ausgedrückt, entspricht dieser Kontext den Wahrscheinlichkeits-oberflächen von Stichproben-, Parameter- und Ergebnisräumen. Als Streuungsmaße (z. B. Stichprobenumfang, Spannweite, Varianz) vermitteln die betrachteten „ancillary“ statistischen Kriterien die Struktur und Charakteristiken dieser Oberflächen. Die Kriterien geben umso mehr Einblick in die Verlässlichkeit, je kontinuierlicher und weiter ihr Gültigkeitsbereich ist. Validität (Konvergenz und Konsistenz, gemessen durch Präzision und Genauigkeit, ergänzt durch Robustheit und Kongruenz), Effizienz („Geschwindigkeit“) und Suffizienz („Power“, Aussagekraft), ebenso wie Modelspezifikation (Modeldefinition, -selektion und -prüfung) und Hypothesen-falsifizierung quantifizieren wie überzeugt man von Parameterwerten, Unterstützungsmaßen, Vorhersagen und Testentscheidungen sein kann, welche durch statistische Verfahren erhalten wurden. Auf diese Art und Weise bildet statistische Verlässlichkeit das Subjekt, den Prozess und das Ziel statistischer Validierung von Referenzdatensätzen und Inferenz-Herangehensweisen.

Dieser Überblick stellt jedes der „ancillary“ statistischen Kriterien vor und diskutiert allgemeine Strategien für die praktische Anwendung, wie auch vorhandene Umsetzungen für die genetische Zuordnung von Stichproben zu Populationen. Die grundlegende Bedeutung genomweiter Sequenzinformationen für Referenzstichproben aus dem gesamten Verbreitungsgebiet in Nichtmodellorganismen wird aufgezeigt. Entsprechende Referenzdatensätze stellen die informationsreichen genomischen Daten bereit, welche die Grundlage für die Entwicklung suffizienter und vielseitiger statistischer Inferenzverfahren sind. Zusammen bilden solche Daten und Verfahren die Voraussetzungen dafür, dass genaue, entscheidungskräftige und verlässliche statistische Werkzeuge für die Erhaltung, Bewirtschaftung und die Strafverfolgung erreicht

werden können. Die Validierung dieser Qualitätsmerkmale legt die Grundlage für die breite praktische Akzeptanz und den Einsatz solcher Werkzeuge auch für Nichtmodellorganismen.

Schlüsselwörter: Erhaltung, Bewirtschaftung und Strafverfolgung; forensische Genetik für wilde Flora und Fauna; statistische Verlässlichkeit; „ancillary“ statistische Kriterien; statistische Validierung

Table of Contents

List of Figures

List of Tables

List of Information Boxes

"… we rely on the recovered representation from a statistical procedure to be meaningfully connected to the true genetic structure that has emerged from a complex evolutionary process. It is essential that we assess the strength of this connection." (Mimno et al. 2015)

"… methods of phylogenetic analysis can be (and often are) highly accurate for problems as diverse as life itself, given sufficient sampling, sufficient attention to rigorous analysis, and sufficient computational power. Just don't expect the final answer anytime soon." (Hillis 1995)

A INTRODUCTION

1 Forensic genetics for biodiversity conservation

Forensic genetic science applied to wild flora and fauna seeks to protect endangered evolutionary lineages, their natural populations and associated communities from overexploitation and habitat destruction (Iyengar 2014, Johnson et al. 2014, Ogden & Linacre 2015). Its objective is to develop and provide reliable and decisive inference tools and services for the enforcement of national to international laws and agreements, as well as for certification systems. In this way, it is an essential part of conservation efforts and management strategies for biodiversity (Waples et al. 2008, Ogden et al. 2009, Alacs et al. 2010, Wilson-Wilde 2010, Iyengar 2014, UNODC 2016b). At the same time that it safeguards the biological environments and resources human societies depend on, it supports societies' ethical values for economic interactions by protecting producers' integrities and customers' rights (Mignone & Howlett 2012, Konnert et al. 2015, Lowe et al. 2016).

Forensic conservation genetic statistics

Forensic conservation genetics is abstract and highly mathematical-statistical. Nevertheless, as disembodied as the specifics and procedures of statistical inference approaches might appear, they can provide indispensable support and services for the executive and judiciary (UNODC 2014, CITES 2015, UNODC 2016a). In this commitment, forensic conservation genetics directly complements hands-on ecological and organism-level conservation projects and campaigns. Law enforcement, NGO (non-government organization) and activist interventions against (often internationally organized) crime associated with the illegal destruction of and trade with protected biological entities aim for actions of the legal system. Here, forensic conservation genetics has the goal to provide statistically sound highly-supported and conclusive results that form the foundation of intelligence and evidence accepted by and of impact in court trials (Johnson et al. 2014, Ogden et al. 2016).

However, robust and reliable statistical inference results are at least as necessary for the establishment and regular routine operation of voluntary DNA-based control and certification systems (Nielsen et al. 2012a, Bekkevold et al. 2015, Lowe et al. 2016). In this application, they are even more important for long-term successes in protecting, conserving and managing biological entities in their natural environments. A well-functioning voluntary survey, control or certification system with an established history of success regarding its decisions as measured in accuracy and resolution might not only efficiently deter illegal exploitation and boost certification initiatives, but also provide additional weight to credibility in arising court trails (cp. Johnson et al. 2014, Ogden & Linacre 2015, UNODC 2016a).

All these civil processes and structures are needed to work in concert to protect and preserve the natural environment and its species, as well as the rights and livelihoods of involved humans and their societies.

1.1 Assignment to evolutionary lineage

The core task in forensic genetics is the assignment of a sample to its original evolutionary lineage from the species- (or higher classification) and the population- to the individual-level (Iyengar 2014, Johnson et al. 2014, Dormontt et al. 2015, Ogden & Linacre 2015, Ogden et al. 2016, for examples see also UNODC 2016b). Assignment tasks at the ends of the evolutionary relationship spectrum work comparably well (fig. 1.1). These include the identification of individuals (individuation; e. g. Nowakowska 2011, Lowe et al. 2016 Box 2, Tereba et al. 2017) and species determination, that is, assignment to independently evolving species (barcoding; e. g. Hebert et al. 2003, Dawnay et al. 2007, Tripathi et al. 2013, Ferri et al. 2015, Mallo & Posada

2016). Here, existing inference approaches provide sufficiently high and robust support or exclusion probabilities in many cases. However, at intermediate evolutionary scales the identification and quantification of relationships are highly context dependent. Currently, development efforts at the population-level towards improving inference methodology focus at achieving reliable, conclusive assignment and exclusion probabilities for these intermediate levels.

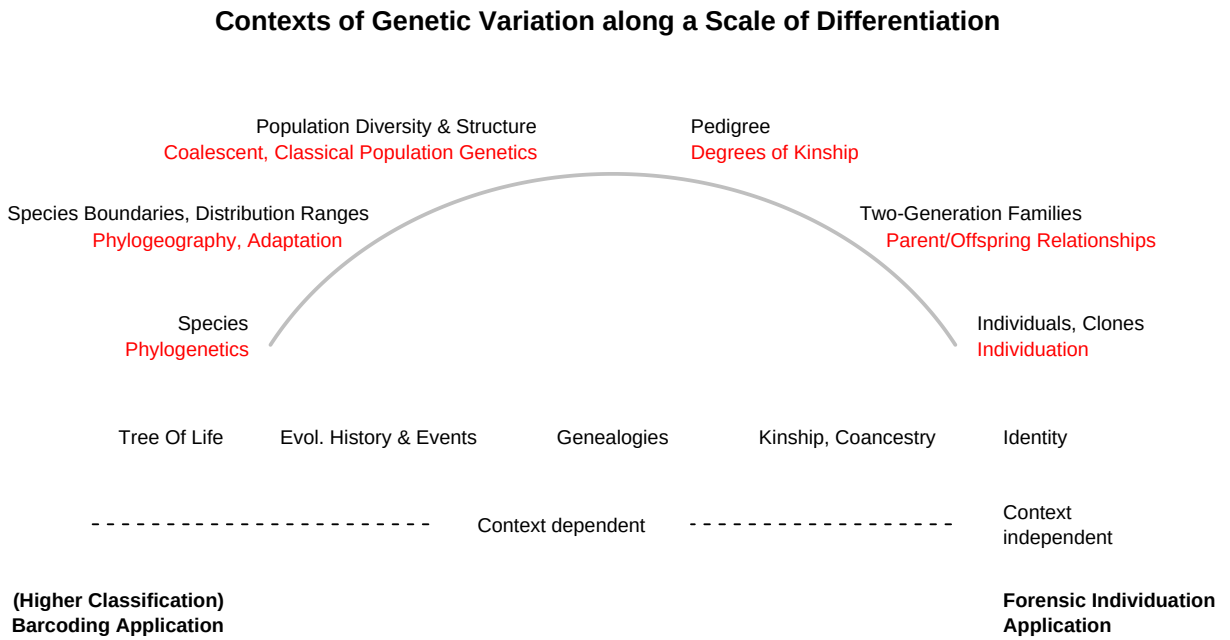


fig. 1.1: The spectrum of evolutionary relationships and their associated inference tasks (in red) along the scale of genetic differentiation. The gradient starts on the left with old, clearly distinct evolutionary lineages, which represent species and higher taxonomic groups. Towards the right, evolutionary lineages become less distinct and more closely related, ending with the identification of individuals and clones. Inference is more or less context independent only at the ends of the spectrum, while all intermediate population-level inference results are context dependent. To date, versatile routine applications are only found at the ends of the spectrum in the form of forensic individuation and barcoding casework.

This review focuses in the applications it presents on geographical population assignment, for which sufficient predictive accuracy and resolution often is difficult to achieve, though many successful applications exist (Ogden 2008, Iyengar 2014, Ogden & Linacre 2015, Dreifus 2016, Millar 2016, Ogden et al. 2016). Geographic population assignment is crucially needed in fishery and forestry for certification and forensic applications (e. g. Withler et al. 2004, Beacham et al.

2005, Van Doornik et al. 2007, Tnah et al. 2009, Degen et al. 2010, Tnah et al. 2010, Jolivet & Degen 2012, Nielsen et al. 2012a, Degen et al. 2013, Bekkevold et al. 2015, Konnert et al. 2015, Lowe et al. 2016, UNODC 2016a, Bernatchez et al. 2017). Here, existing experiences show that for many widely distributed and often common, but intensively harvested taxa it currently is not possible to achieve sufficient support and predictive accuracy for assignment at all required levels (Ogden 2008, Dormontt et al. 2015, Barber & Parker-Forney 2017), cp. also Wasser et al. (2004). These lineages of commercial interest might be regionally rare with fragmented occurrences due to marginal habitats or overexploitation (e. g. Waples et al. 2008, Nielsen et al. 2012a, Degen et al. 2013, UNODC 2016b). Conversely, locally cultivated or managed populations might have been established. A situation that subsequently can be complicated by resulting gene flow into surrounding protected populations (e. g. Glover 2010, Glover et al. 2013, Bylemans et al. 2016). In addition, some of the taxa have threatened sublineages, or are part of species complexes with unclear species boundaries (Waples et al. 2008, Hartvig et al. 2015, Schroeder et al. 2016). In all of these cases, required population assignment to legal entities might show insufficient statistical characteristics. This can be the case at least at some geographic and evolutionary scales or in parts of the distribution range. In this context, legal entities can refer to, for example, countries or ocean boundaries, but also to specific sublineages, with different protection statuses in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES).

1.2 Complexity and the many approaches

The analytical goal in forensic conservation genetics is to establish an open and diverse set of well-assessed statistical approaches that provide the capacity for statistically sound reliable and conclusive results. Development and validation of these approaches are the prerequisites for arriving at effective executive analytical tools that are widely accepted and used in routine application (Bekkevold et al. 2015, Dormontt et al. 2015, Ogden et al. 2016).

Between forensic disciplines, but also within forensic conservation genetics, there might not be the one decisive tool or inference approach in every case, due to the complexities of evolutionary histories and processes (Dormontt et al. 2015). Instead, several different approaches and their results might need to be combined. Hereby, the efforts taken to counter international organized crime against biodiversity, as for example described in the World Wildlife Crime Report (UNODC 2016b), see also UNODC (2012) and Hoelzel (2015), can be considered similar to the approach taken by the Intergovernmental Panel on Climate Change (IPCC 2014). The IPCC and scientists investigating global climate change processes to infer their causes and predict their impacts are drawing their conclusions by combining and evaluating a very diverse and divergent set of evidence (e. g. recently Hansen et al. 2016, Praetorius 2018).

Nevertheless, it is of fundamental importance to arrive at effective stand-alone executive forensic analytical tools for routine application that provide sufficient specificity and discriminating power, while being robust and reliable. Yet, the complexities of evolution make

many statistical inference tasks for genetic forensic cases connected to natural populations and lineages notoriously difficult. This context requires one to understand the species systems well; to collect a set of reference samples representing the genetic diversity across the distribution range; to develop and employ well-fitting (evolutionary process or pattern-recognition) models for the investigated evolutionary lineages; and to have at hand inference approaches that are able to detect and differentiate the multitude of signals recorded in the genome and, thus, present in genomic data. The research community in evolutionary biology has developed a wealth of inference approaches that pertain to objectives of interest in forensic conservation genetics. For example, currently there exist over 70 published algorithms and software implementations for population assignment, clustering and the evaluation of coancestry. Biodiversity research projects with objectives relevant to forensics have shown the potential and feasibility of these inference concepts, are building extensive reference collections and are assembling genomic background information (Ogden et al. 2016, UNODC 2016a).

1.3 Transition to executive applications and new genomic opportunities

Today forensic conservation genetics faces the challenge to navigate and achieve the transition from research to forensic application (Ogden 2008, SWGWILD 2012, 2015, Ogden et al. 2016). This step requires validation of employed methods. These extensive assessments, as far as possible across a range of scales and tasks, form the basis for acquiring an in-depth understanding of the characteristics, strengths and limits of large reference datasets and promising inference methods. As before, also the validation step poses a challenge, due to the widely divergent species and evolutionary histories that are encountered in non-model organisms from across the Tree Of Life, as well as, the lack of preexisting knowledge for the many insufficiently known and studied endangered taxa.

Nonetheless, contrary to the necessity for comprehensive validation and the general reality of evolutionary inference as described in the initial quote by Hillis (1995), conservation scientists and managers, law enforcement, as well as activists are often running out of time with regard to protecting evolutionary lineages from extinction. In the face of the pressures of present day global dynamics, recent technical and methodological advances are opening up new possibilities and the chance for accelerated progress. Currently, evolutionary biology and with it forensic genetics are experiencing a rapid development of statistical theory and inference methods (Weir et al. 2006, Thompson 2013, Schraiber & Akey 2015, Speed & Balding 2015, Cussens & Sheehan 2016). This progress is due to the rich and deeply structured genomic information content that is becoming accessible through genome-wide sequencing (GWS) strategies. Different GWS-approaches today provide alternative options for a, as far as currently possible, representative and unbiased subsampling of the genome or for whole genome sequencing (Ekblom & Wolf 2014, Ellegren 2014, Andrews et al. 2016, da Fonseca et al. 2016, Harvey et al. 2016, Collins & Hrbek 2018).

These methods are starting to allow the production of large GWS- and derived ancestry informative marker (AIM) datasets also for non-model organisms, comprehensively covering their distribution ranges. The developments are providing new opportunities for forensic genetic investigations in evolutionary lineages representing the whole spectrum of natural systems, evolutionary histories and genomic architectures.

2 Statistical inference and ancillary criteria

Statistical methodology aims to infer from a limited set of observed data the characteristics of the underlying generating processes (or, of general data patterns) by a) estimating the values and their support of involved parameters, and b) assessing the support and overall accuracy for results from predictive analyses and hypothesis testing (Fisher 1922, 1935, Gelman et al. 2014, Romeijn 2017). The overarching goals are, first, to arrive at accurate, decisive and reliable distributions of parameter estimates and predicted data and, second, to have available test statistics that provide the specificity and discriminating power necessary for sound decisions in real-life application.

2.1 Statistical inference in non-model organisms

Statistical inference is an iterative process, alternating between inductive and deductive methods (Fisher 1934 p. 9, Box 1976). Hypotheses are inductively conceived, their associated models are specified and logical consequences calculated by deductive analysis. The resulting distributions are compared to empirical observations, with the outcomes forming the basis for practical deductive application or further inductive conception, refinement and optimization of models and hypotheses (Platt 1964, Box 1979), see also the section on “Robustness”.

In non-model organisms, forensic conservation genetics requires an overall two-step process, since fundamental background information and preceding experiences are often missing. In a first step dominated by inductive reasoning, appropriate reference datasets, well-fitting and useful models and effective inference methods need to be iteratively identified and optimized to reflect and take into account the evolutionary patterns and processes shaping the taxon under investigation (**biodiversity research**; cp. Box 3.2 on empirical reference datasets). Subsequently, in the step of, mostly deductive, practical application, these specified evolutionary or pattern-recognition models and inference approaches form the basis for statistical tests of case-specific hypotheses representing prosecution and defense scenarios or certification controls (**concrete forensic casework**).

Ronald Aylmer Fisher (Fisher 1922, 1925, 1934, 1935) in his contribution to the development of the foundations of statistics laid out the required criteria for "**satisfactory**" **statistics** (Fisher 1925 p. 701, 1934 pp. 11-16). These criteria are validity, comprising convergence and consistency, measured by precision and accuracy, and supplemented by robustness and congruence; efficiency (colloquially "speed") and sufficiency (colloquially "power"); as well as model specification, including model definition, selection and assessment, and hypothesis falsification (fig. 2.1). In inference from a complex evolutionary reality, which involves computationally difficult tasks, the criteria are imperative for arriving at reliable conclusions. The criteria's importance lies in their function as **ancillary statistics** (Fisher 1925, 1935). Ancillary statistics, while saying nothing about the value of an inferred parameter itself, characterize the dispersion of estimates. "Their function is, in fact, analogous to the part which the *size* of our sample is always expected to play, in telling us *what reliance* to place on the result." (emphasis as in the original, Fisher 1935 p. 48). They provide, thus, insight into the structure and characteristics of probability density surfaces that are spanning sample, parameter and result spaces. Examples of ancillary functions are the sample size and an estimate's range and variance. Ancillary statistics convey the more information about statistical reliance, the more their scope (space of applicability) is continuous and wide. Statistical reliance, subsequently linguistically modernized to statistical reliability, is subject, process and goal of the statistical validation of reference sets and inference approaches.

In practical application, statistical reliability forms the key concept that end-users of statistical inference (e. g., conservation managers and activists, certification providers, producers, consumers, courts and juries) need to understand and take into account. The myriad of existing empirical data and inference approaches is difficult to overview and comprehend even for experts, with some of the concepts and statistics being exceedingly intricate and convoluted. Yet, the specifics of the inference system are secondary to understanding the statistical reliability of an inferred result. The statistical reliability associated with an estimate establishes an appropriate context for the estimate, defines the scope for which it is valid, and thus provides perspective. In this way, it forms the basis for evaluating a proposed conclusion.

In molecular evolutionary biology, the ancillary criteria for statistical approaches were summarized and discussed in a phylogenetic context twenty-five years ago (Penny et al. 1992, Hillis & Huelsenbeck 1994, Hillis 1995, Huelsenbeck 1995). They were considered at a time when inference approaches for phylogenetic evolutionary relationships had exploded and diversified with the onset and wider accessibility of Sanger DNA-sequencing. Today, the strengths and limits of the different general classes of phylogenetic reconstruction methods are thoroughly explored, understood and robustly established. More recently, several reviews in molecular systems biology incorporated the criteria in their consideration of inference in complex systems for the development of predictive models towards medical and bioengineering applications (Jaqaman & Danuser 2006, Ashyraliyev et al. 2009, Bruggeman 2009, Cedersund & Roll 2009, Kreutz & Timmer 2009, Srinath & Gunawan 2010).

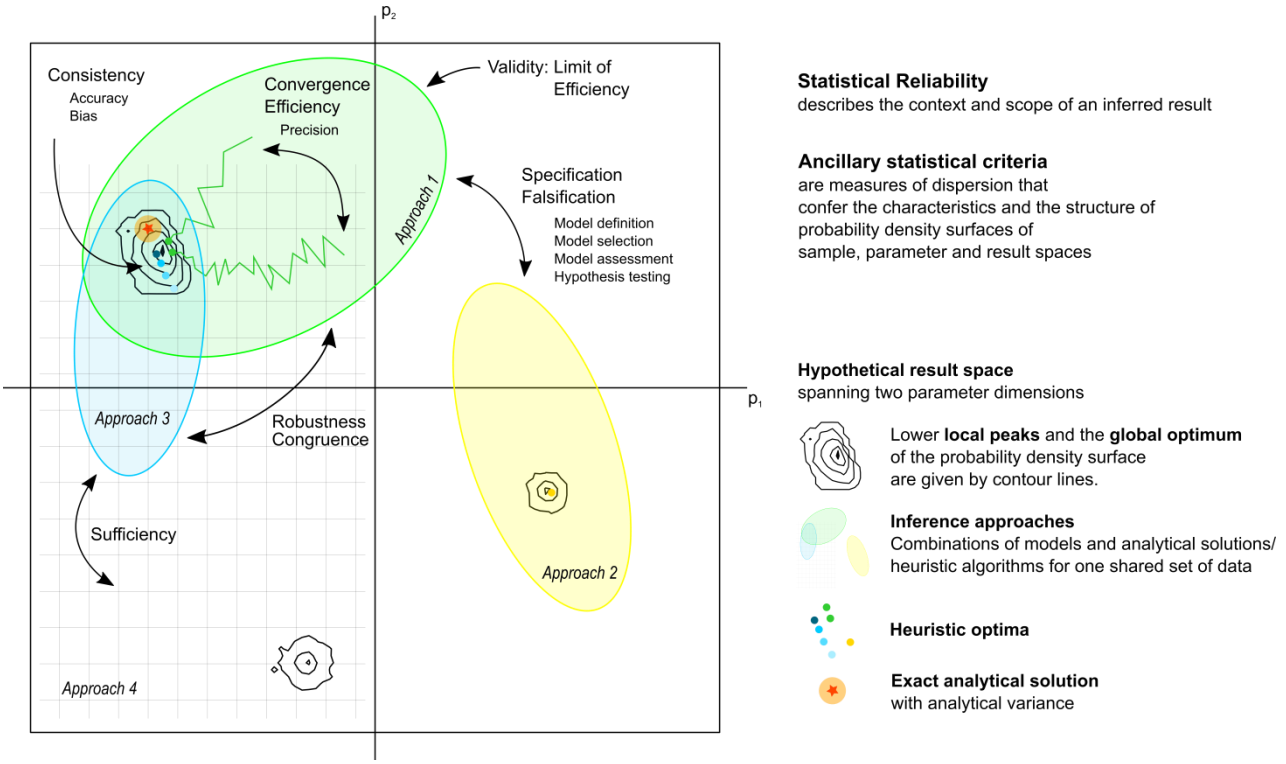


fig. 2.1: A schematic representation of the ancillary statistical criteria in a hypothetical result space spanning two parameter dimensions. The black contour lines of the background landscape show the "true" probability density surface of the result space. The global maximum lies in the upper left quadrant, with two local maxima in the lower half of the graph. The ellipses represent areas of validity of different model and inference approach combinations. The grid arises from an inference approach, which accesses less information in the data, leading to a coarser result space compared to the continuous fill of the ellipses. The orange star represents an analytical point estimator with its associated variance. Dots signify "best" estimators for different model and inference combinations. The series of lighter to darker blue dots associated with model 3 represents consistency with increasing amounts of data. Zigzag-lines suggest the paths of repeated runs of a heuristic algorithm (e. g., MCMC runs) showing convergence.

The present review describes the statistical framework that allows quantification of reliability, and the differentiation of its qualitative components. It introduces each ancillary criterion and discusses general strategies for practical application, as well as, available implementations towards population assignment. In today's explosion and intense exploration of GWS-datasets and newly developed statistical approaches, the reviewed criteria can suggest avenues for finding, solidifying and validating inference procedures with improved resolution and support. The aim is to develop well-understood, reliable, decisive and versatile forensic genetic tools applicable across biodiversity, to be employed in the course of routine control and certification systems, or for case evidence in court trials.

B ANCILLARY STATISTICAL CRITERIA

Ancillary statistical criteria in practical inference processes for forensic conservation genetics

The statistical criteria are presented in the order they likely become of importance in a first practical approach to a new inference objective in evolutionary biology and forensic conservation genetics. They provide the scaffolding for the development and fine-tuning of the inference process.

One starts out by applying and exploring the **validity** of a first inference method by asking if the approach is applicable at all to the available dataset and provides reasonable results (**accuracy** and **precision**). Subsequently, one widens one's consideration and includes further parameter settings and inference approaches in the project, comparing their results with previous ones (**robustness** and **congruence**).

While the inference project grows, questions of **efficiency** and **sufficiency** arise. These include a reevaluation of the initially selected evolutionary model used for inference; the encounter with the practical resource demands of computational complexity; an optimization of the balance between efficiency and sufficiency; and following from this the adjustment of the necessary information content of the data, which by now has become clearer.

In later phases of the project, evolutionary models associated with one to several in this way selected inference approaches are defined, specified and fit to the data (**model specification**). This results in models that are fine-tuned to the data. For this purpose, for each approach a set of plausible (to unlikely) evolutionary models is assembled, and each model within the set is defined by decisions about parameters, their interactions and values. The estimates for the parameters of these models are then optimized using the selected inference procedure. Optimized and, with regard to the chosen inference method, completely specified, the evolutionary models now can, on one hand, be tested with the aim of selecting the one best fitting the data, relative to the set of investigated models (**model selection**). On the other hand, they can be explored by model checking describing their strengths and limits. This leads to more resource-intense **model assessment** approaches, which combine aspects of model specification, selection and checking to arrive at one best fitting and most useful, fully specified evolutionary model. By this developmental process the final model has been **validated** and therefore its applicability confirmed.

In the final phase of a forensic conservation genetic task, the actual **forensic hypothesis testing** takes place. It provides the foundation for a decision with real-world consequences. In addition to the null hypothesis of natural evolutionary processes, a set of hypotheses is assembled that represent the specifics of the case. Modifying the previously fitted and optimized evolutionary

model, the alternative forensic models are defined to represent the case hypotheses. This might involve the inclusion of anthropogenic constraints as proposed by the defense and/or prosecution (e. g., the stated origin of the case sample(s)). If necessary, this forensic set of models is optimized as before. Finally, hypothesis testing is carried out. Estimates of its associated quality measures of **overall accuracy**, **error rates** and **predictive values** are quantified based on the available reference dataset. Further exploring and strengthening support for the test conclusion, additional hypothesis testing might be independently performed based on several inference methodologies. This allows an evaluation of the **congruence** of the test results, and might provide additional, supplementary insight, which is contributed by independent, potentially complementary approaches.

3 Validity

3.1 Convergence and consistency: precision and accuracy

The reason to apply statistical methodology is to reliably infer reality (Fisher 1922, 1935, Edwards 1992 pp. 5-6). The validity of an inference approach to do so requires that it is characterized by convergence and consistency, as measured by precision and accuracy (Fisher 1922, there defined in terms of efficiency). Accordingly, a method should return results that converge to a fixed value (within a specified precision; cp. also the definition of efficiency) as the sample size is increased (convergence; cp. Neale 2012). Given infinite data, or more realistically sufficient data, the inferred estimate should equal the true value of the parameter of interest in the underlying population (consistency; Fisher 1922, 1925, Felsenstein 2004 p. 107). In applications to actual empirical data of limited size, the accuracy of an estimate quantifies the degree of closeness to the true value (Fisher 1925). Its precision quantifies the estimate's variability.

In analyses of natural populations (fig. 3.1), adding more data per sample, that is, more genomic characters of the same kind (e. g. neutrally evolving nucleotide sites), such increased genomic sampling should only affect the variability of the result that is due to sampling or random errors (its precision), and not change the estimate itself (its accuracy). A change of the estimate makes bias due to systematic error(s) apparent (Motulsky 2010 p. 60). Nevertheless, inference approaches can be biased in their approach to the truth, but still be consistent and ultimately accurate (Fisher 1925). They might, for example, be skewed towards consistently under- or overestimating the parameter of interest, while asymptotically approaching the true value.

In contrast, adding more samples or datatypes might correctly change the estimate and its precision, since natural populations are not homogeneous theoretical entities and a taxon's diverse characteristics evolve due to different evolutionary processes. Thus, a change in sampling design or marker type might result in distinct representations of evolutionary scale and background, changing the appropriate evolutionary model.

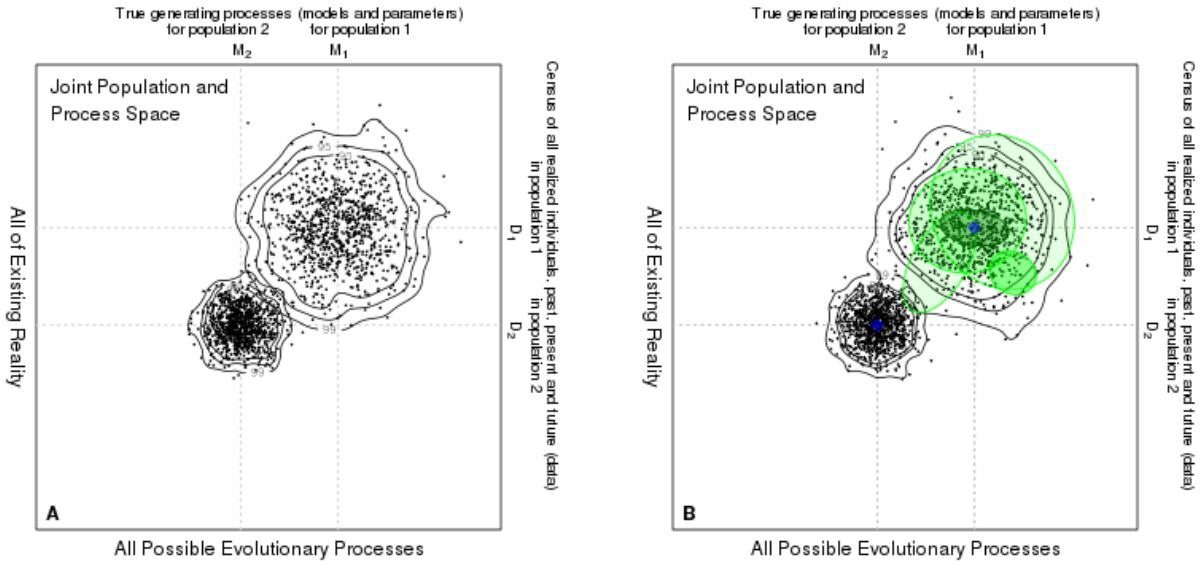


fig. 3.1: Evolutionary reality represented as joint probability space. A: The generation of reality by evolutionary processes (models and parameters) and realized organisms and individuals (data). Evolutionary processes give rise to individuals (samples), which define populations. At the same time, existing individuals in a population (defining its e. g. diversity) shape the evolutionary processes and determine what will be possible in the near future. Depicted are populations in stationary equilibrium over time. Most evolutionary processes are randomly variable over time and space in parameter values and outcomes (data). Contour lines enclose 90, 95 and 99 % of the realized parameter values and data. B: Several random samples of population 1 are shown, representing more or less sampling coverage and bias. Fill densities represent different sampling intensities. The dark green ellipse is an example of one concrete dataset (sample and markers) under investigation. Blue dots are the points of highest density for both model and data.

The validity of an inference method can depend on the parameter space that is represented by the empirical data (see the discussion of intrinsic accuracy and efficiency in small samples in Fisher 1925). One example for this, that has been extensively explored and documented, is the potential for erroneous long-branch attraction in phylogenetics (Felsenstein 1978). Here, several “zones” have been delimited with regard to specific parameter combinations of divergence times, and the speed and heterogeneity of evolutionary rates. The validity of the different tree reconstruction approaches in each of these zones is quite well understood. The underlying reasons and contributing factors are known, and therefore, the limits as well as advantages of the different methods can be explained and predicted in application to empirical datasets presenting specific evolutionary scenarios and sampling schemes.

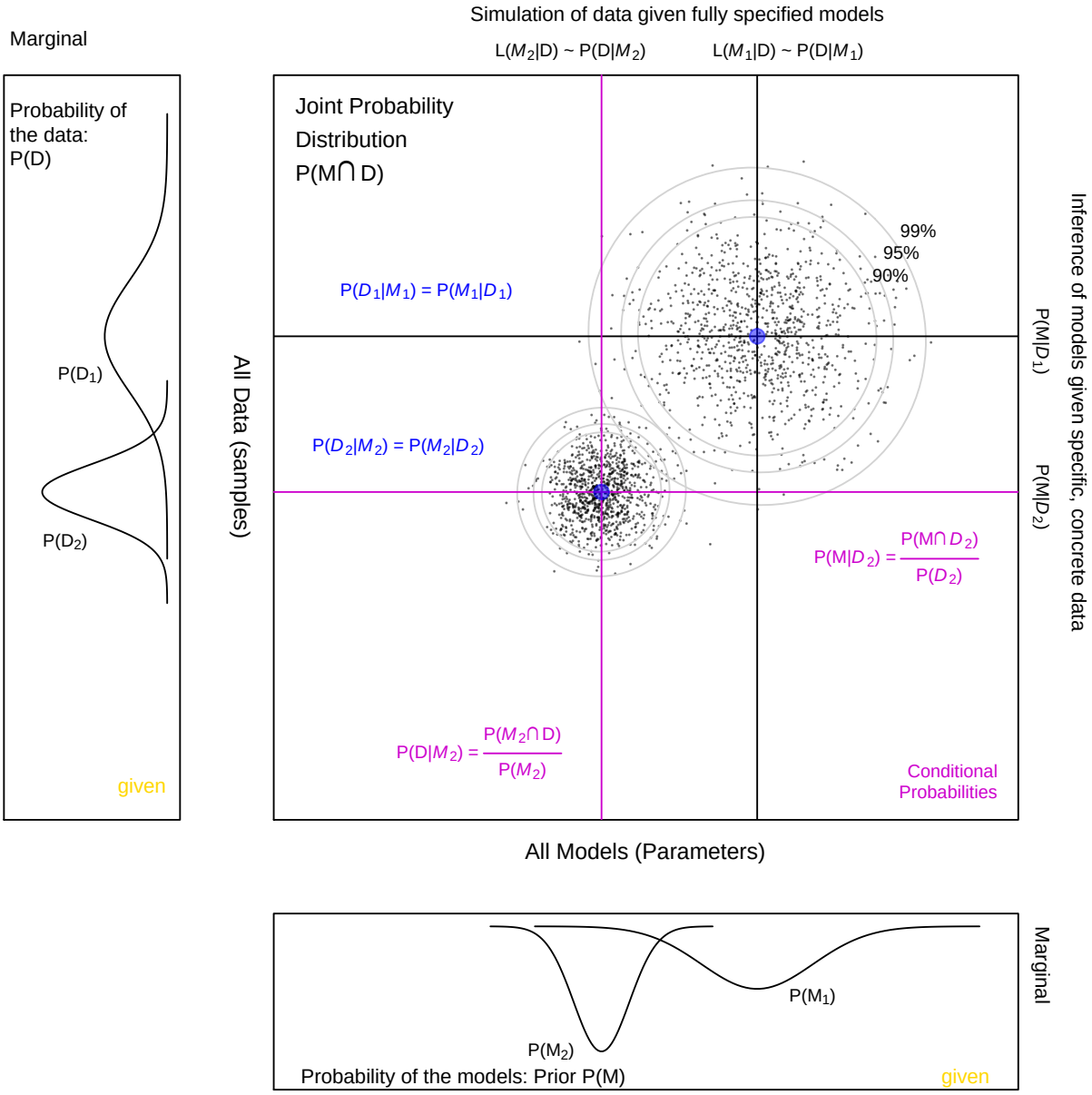


fig. 3.2: Simulation given a model and inference from given data as alternative and complementary approaches. *Italics* denote, if in a reconstruction approach data or model are given, while their opposite (model or data, respectively) is inferred. Gray ellipses denote reconstructed margins of the parameter and sample distributions (90, 95 and 99 % of the inferred parameter values and of the simulated (predicted) samples lie within the ellipses). The marginal distributions of the models (priors) and the data are considered given. They are either proposed (e. g., as “non-informative”), or are estimated by previously inferred posterior (predictive) distributions (Bayesian framework) and bootstrapping (frequentist framework).

Simulation studies are generally the only way to assess the convergence and consistency of results from statistical inference approaches, since the underlying generating process can only be known, when deliberately designed (fig. 3.2). Hence, simulation studies provide the essential means for exploring, assessing and comparing the general properties of newly developed and existing inference approaches applicable to evolutionary and forensic conservation genetics (cp. section 3.2.1 “Sensitivity Analysis”). Simulation runs can be repeated, which is often an intrinsic and automatic part of the simulation (repeatability). These repeated runs correspond to the repeated, independent experiments that form the conceptual basis of the frequentist framework. Hence, they allow the direct estimation of accuracy and precision. Simulation studies also can be reproduced in other, independent settings (reproducibility), e. g. by other working groups, run on different hardware, or by independent programming for the same statistical approach. Again, reproducibility can provide direct estimates of validity.

Controlled sets of experimental evolution in microorganisms and (domestic) model-organisms provide a kind of real-world simulation producing empirical data, for which at least some fundamental generating processes are known (Hillis 1995, Poe 1998). However, such experiments rarely are feasible at the population-level and with regard to natural populations in conservation genetics. An exception can be investigations concerning shorter timespans. Here, breeding programs of zoos or continuously observed small (animal) populations can provide empirical pedigree information, which can be used to validate statistically inferred estimates of kinship.

Evaluating the validity of inference results for data from natural populations

The evolutionary truth for natural populations cannot be known and, moreover, evolution of natural populations in the wild cannot be repeated in place and time. Therefore, in analyses of observed empirical, real-world data, the validity of inferred results can only be approached indirectly. If reality cannot be known, the question becomes if inferred estimates and outcomes of hypothesis tests are at least robust and useful (cp. section 3.2 “Robustness”), and if they make sense (cp. section 3.3 “Congruence”). A qualification and quantification of robustness is approached by exploring the sensitivity of results to assumptions of the inference model and approach (cp. 3.2.1 “Sensitivity analysis”) and the fit of the selected model and approach to concrete data at hand (cp. 3.2.3 “Model checking”). Model checking includes an evaluation of the dependence of results on the particularities of existing data, which it does by resampling both, subsets of the existing data and parameter values of previously inferred (i. e. data-dependent) models. While robustness fundamentally looks for parsimonious ways to explain reality and their usefulness for application, congruence in addition asks if the inferred conclusions make sense in light of independent preexisting results and when considering the wider already known context.

Specific statistical methods and procedures available for the evaluation of robustness and congruence are described in the corresponding sections later on. In the following, the general framework is laid out of how direct and indirect approaches are combined to evaluate the validity of assignment results to natural populations.

Reference datasets assembled for forensic and monitoring objectives provide the possibility to validate the overall accuracy and precision of assignment approaches employing this concrete set of data directly through self-assignment and blinded-sample strategies. These strategies take advantage of the fact that the geographic collection localities, which are part of evolutionary reality, can be known, even while the rest of the generating evolutionary process is inaccessible. The geographic origin is reconstructed for each reference sample employing cross-validation (including blinded samples), predictive distributions (including (non-) parametric bootstrapping) or Markov Chain Monte Carlo (MCMC)-search heuristics as statistical resampling methods. The reconstructed population of origin is compared to the known sampling locality. While cross-validation provides information on accuracy, predictive distributions and MCMC-sampling evaluate precision. Thus, given a specific reference dataset (samples and markers) and a chosen inference approach, the resampling strategies shed light directly onto the validity, as well as the properties of this reference dataset and the selected method for population assignment.

Subsequently, such existing experience with and insight into the validity and usability of a reference dataset and inference approach provides the foundation for evaluating the robustness of a population assignment in concrete forensic genetic casework. The validity of an assignment conclusion for a case sample, for which per definition the geographic origin is in question, can only be assessed indirectly by referring back to the quality characteristics of previous analyses of self-assignment and blinded samples for the reference dataset.

Self-assignment as a core strategy for evaluating the validity of population assignment

The evaluation of specificity, that is, the precision, and more correctly robustness, with which a population can be characterized by its genotypes is already at the heart of the earliest supervised population assignment methods (Paetkau et al. 1995, Rannala & Mountain 1997). To this end, the quality of inference results is assessed by performing predictive resampling from population allele frequencies in combination with cross-validation. Such approaches were implemented and further extended, for example, in the widely used program GeneClass2 (Baudouin et al. 2004, Piry et al. 2004). They have a strong, well-understood and lucid theoretical basis and by now have been thoroughly explored, extensively compared, tested and applied (e. g. Cornuet et al. 1999). They, thus, provide a well-founded and reliable methodological framework for inferring support for population assignment or exclusion decisions. In addition, in pairwise comparisons of populations, they allow the evaluation of discriminating power, hence the estimation of predictive values (Ciampolini et al. 2006).

Yet, one has to keep in mind that most resampling approaches (however not cross-validation) only evaluate precision and not accuracy (Hillis 1995). This has been made explicit by Paetkau et al. (2004) for supervised population assignment and exclusion inference. They found that the implicit assumptions of the integral bootstrapping strategy have a considerable impact on and can result in misleading test results. If the resampling strategy does not fit the generating process and sampling design well, a large number of replicates can result in a highly precise (low random

error), but still incorrect or biased estimate (systematic error) of a parameter distribution and its support.

In forensic genetics for fishery and forestry, existing population assignment results show the robustness and potential of these approaches by returning marked tendencies in the expected direction for tests of self-assignment and exclusion in reference datasets. However, too often assignment tests do not reach sufficiently high overall accuracy, and lack resolution at critical scales of interest or for specific geographic regions. At the current size of reference datasets of hundreds to thousands of samples and well over a hundred pre-selected geographically informative SNPs (AIMs), adding more such AIM markers is not a general solution (Cornuet et al. 1999, Bekkevold et al. 2015). This is probably due to an accumulation of error rates, for example, in estimates of population allele frequencies or already during base calling for genotyping. At this point, it is necessary to improve assignment power by developing and utilizing better fitting models of evolution or improved pattern-recognition algorithms for highly dimensional and very diverse data, as well as, to analyze more informative GWS-data. This is a very active domain of research and development, especially for widely and continuously distributed species, such as humans, that often show a lack of assignment support (Thompson 2013, Novembre & Peter 2016).

Blinded samples

Self-assignment approaches are often complemented by blinded samples. This second approach towards the evaluation of validity allows a more far-reaching characterization and quantification of consistency via accuracy. Here, in addition to the reference set, “case samples” are analyzed. The origin of these blinded samples is known, but only to parties that are not involved in the laboratory and statistical inference analyses. The approach represents, hence, an extension of statistical cross-validation. Blinded samples are very powerful and honest inference procedures. If the blinded samples are chosen appropriately, they provide a random and unbiased test of the reliability of the whole forensic tool, including the reference set, logistics and handling (chain of custody), laboratory work, and the statistical approach. Consequently, blinded samples are generally part of standards and processes defined for accreditation and certification.

3.2 Robustness

The results returned by a (non-) parametric model and inference method should as much as possible be independent of the specific assumptions of the applied statistical approach, including the idiosyncrasies of the available data (Box 1979, Gelman et al. 2014 pp. 141 ff.). They especially should be robust to violations of basic, often unstated assumptions (Hillis et al. 1996 p. 529) and the effects of any violations should preferably be quantifiable (Gelman et al. 1995 p. 162). This criterion of relative insensitivity to assumptions is fundamental, since all assumptions of every

approach are violated to some extent when it comes to real data (Huelsenbeck 1995, Hillis et al. 1996 pp. 526 ff.).

3.2.0.1 "Robustification" as indispensable part of verification

Box (1979) defined robustness as the property of a procedure (comprising model and inference method, as well as, reference data), which renders the answers the procedure gives less sensitive, ideally insensitive, to real-world departures from abstract, ideal assumptions. The process of "robustification" (Box 1979) is a sequential iteration between the specification of a tentative model and a following preliminary inference phase conditional on this tentative model (Box 1979). Diagnostic checks of the thereby inferred results reveal specific deviations from the data and/or from substantial independent knowledge. These diagnostics can point to suitable improvements and extensions of the model, which, changed accordingly, again can be applied to analyze the available dataset. In its observations and decisions at critical development steps, robustification is thus creatively inductive.

A fundamental strategy that is part of the process of robustification is to make underlying, unstated model assumptions explicit. Only in this way, it becomes possible to assess their impacts on the results. Yet, being explicit about assumptions adds parameters to the model, thus leading to models that are more complex and parameter-rich. This, however, proceeds contrary to the insight that robust models, rather, are parsimonious, simple models. The inductive spark and contribution to robustification is to "cunningly choose" approximations that show themselves to be "remarkably useful" and, in addition, might prove to be more generally applicable (Box 1979). The rationale behind the preference for simple, approximate models is that "all models are wrong …" no matter how elaborate (Box 1979). Hence, since models representing the "whole truth" of reality cannot be build, - "… but some are useful" - one might as well concentrate on designing those that are as simple and approximate as possible, while still fitting the data sufficiently well for being "illuminating and useful" given the task at hand (Box 1979).

Box (1979) argues that robustification is best carried out by exploring the model space, not by modifying the inference procedures. His opinion with regard to the inference step is that process-oriented parametric approaches are preferable over nonparametric ones, since the assumptions of these are implicit and unstated. The classical (e. g., maximum likelihood and Bayesian) estimation methods for parametric models should not be modified, but kept - in contrast to the model. The parametric inference methodologies of today are still in accordance with this assessment and fundamentally based in classical maximum likelihood or Bayesian frameworks. Nevertheless, today, one can choose from a plethora of parametric, along with, nonparametric (e. g. multivariate) and mixed inference approaches, both for parameter estimation, as well as, for the prediction of data, that is, of future, so far unobserved samples.

The explosion of computational power created the opportunity and necessity to develop inference approaches, no matter if parametric or nonparametric, that take advantage of and rely on a wide range of heuristic algorithms. These heuristics need to be attuned to each dataset and task, just as models need to be fit to data. They require that search and optimization parameters need to be chosen and specified. Thus, today the inference approaches themselves have become the subjects of inquiry and exploration, requiring robustification and verification.

Especially the Bayesian framework has expanded the available options and methods for a wider range of inference tasks and approaches, as well as, for inquiry into more specific parts of the modeling and inference process. It provides for explicit hierarchical structuring of the inference approach (see section 3.3 “Congruence”), thus enabling the inclusion of quantified priors into the exploration and evaluation of the behavior of models. For the wide range of practical applications encountered, from everyday life to forensic genetics, the Bayesian framework often provides more flexibility for building more elaborate, parameter-rich models to arrive at robust, well-understood and reliable inference approaches. Yet, statistical inquiries and procedures thereby again become more extensive and intricate, and thus time- and expertise-intense.

Hence, today the iterative robustification process involves the development and verification of both, model and inference methodology. A task-specific balance has to be found between all components, which is suitably attuned and sufficiently informative. This encompasses the balance considered by C. E. F. Box between parsimonious, approximate but useful models and more elaborate, parameter-rich models with better fit to the data. Today, in addition, the balance is expanded and necessitates attention to the type of inference approach, which is associated with and required by the selected type of model. This theme, the optimization of model complexity and heuristic algorithm refinement, is taken up in the presentation and discussion of the criteria of efficiency and sufficiency (see chapter 4).

For forensic applications in conservation genetics it is of interest that the exploration and description of robustness can lead to generalization. At least, it can provide dependable, promising starting points for efficient robustification in work with new taxa, inference tasks and reference datasets. The knowledge of the extent and limits of generalization enables versatility, quick response times and reliable conclusions in practical application.

Nevertheless, despite the prospect for generalization, the degree of sensitivity to a certain assumption of a model and inference approach rarely can be sufficiently determined *a priori* and analytically for a specific dataset sampled from a complex evolutionary reality. This is especially the case when met with the high diversity encountered in evolutionary studies of non-model organisms. Here, the transfer of knowledge and experiences from one system to another is often not possible, since no approach is equally independent with regard to all assumptions, and works equally well on data with all kinds of characteristics. The exploration and delimitation of assumption-, model-, inference- and data-specific properties and constraints, as well as the impact of potential error sources, is the core task of inference from each concrete (reference)

dataset. Statistically, this is approached using simulated data for the exploration of model behavior across the parameter space (sensitivity analysis), the integration of similarly well-supported results into one overall result (model averaging) and by model checking, which characterizes the fit of a model to a concrete dataset at hand and similar datasets subsampled from the original data (Table 3.1; Gelman et al. 1995 pp. 161 ff., Hillis et al. 1996 pp. 493 ff.)).

3.2.1 Sensitivity analysis

Sensitivity analysis investigates the reasonableness of an inference procedure for answering a specific objective at large. In evolutionary genetics, simulations are used extensively for sensitivity analysis to investigate the robustness of a model and/or inference method. Simulation-based sensitivity analyses are especially conducted for the evaluation of newly developed inference methods and software implementations. Numerous fundamental simulation studies exist that characterize systematically the robustness of assignment and clustering methodologies and programs (e. g. early studies include Paetkau et al. 1997, Cornuet et al. 1999, Wilson & Rannala 2003, Paetkau et al. 2004, Latch et al. 2006, François & Durand 2010, Safner et al. 2011). Simulations provide important insight into the effects of specific model parameters and the conditions under which an inference method performs well or poorly (Huelsenbeck 1995, Hillis et al. 1996 pp. 526 ff.). However, as these authors pointed out, simulated data always are generated by models that are simplified compared to reality. In addition, simulations are prone to simulation bias and the overgeneralization of their results. It thus is common practice to assess inference approaches not only based on simulated data, but to apply them also to – if possible, already well-known - real datasets (Gelman et al. 2014 pp. 139 ff.).

3.2.2 Model averaging

Model averaging provides a way to combine weighted parameter estimates for the same set of data from divergent models analyzed by the same inference method (Johnson & Omland 2004, Sullivan & Joyce 2005, Stephens & Balding 2009, Grueber et al. 2011). This approach allows information on parameter values returned by different models with substantial (e. g., not significantly different) support to be consolidated. The procedure weights the parameter estimates by the likelihoods or posterior probabilities of their generating models. It thus allows a quantified summary of conclusions drawn from a set of similar, but varying models that were applied to the same empirical dataset.

Sullivan and Joyce (2005) note that reversible-jump Markov chain Monte Carlo techniques provide results that are similar to model averaging. Reversible-jump MCMC has the added advantage to not requiring the evaluated models to assume a particular form of priors.

3.2.3 Model checking

The aim of model checking is to investigate and describe the fit of a model and/or inference approach to concrete empirical data. Model checking evaluates the adequacy of the selected procedure for answering specific objectives given the set of observed data at hand. Gelman et al. (1995 pp. 161 ff.) detail three routes for model checking in the Bayesian context, which have equivalents in non-Bayesian frameworks. These routes differ in their specific combinations of a reliance on independent knowledge and the use of predictive distributions (Table 3.1).

Table 3.1: Robustification: model checking (routes 1-3), sensitivity analysis and model averaging

Based on \ Verification	External Validation	Self-Consistency Checks
Support measure (Model)	*Route 1: Model Selection* Non-supervised global clustering of reference data	*Sensitivity Analysis:* Exploring the impact of model assumptions *Model Averaging:* Combining results with substantive support
Predictive distribution (Data)	*Route 2: Decision Analysis* Supervised self-assignment and assignment of blinded samples	*Route 3: Posterior Predictive Checks* Forensic casework: clustering and assignment

The first two routes employ empirical data that were not used in the analysis, results from independent inference methods or substantive preexisting knowledge for the verification of inferred results (**external validation**; Gelman et al. 1995 p. 162). This information does not enter the prior or the likelihood (Gelman et al. 1995 p. 162). It thus is able to substantiate the adequacy of the fit of the model that generated a result. Basing model checking on data that was not included in the original analysis, provides a connection to cross-validation approaches (Gelman et al. 1995 p. 184). Furthermore, the comparison with results from independent sources and inference procedures as external validation leads to questions of how congruence between the original result and such additional information should be evaluated and quantified (see section 3.3 “Congruence” below).

In studies investigating geographic origin, examples of preexisting knowledge are the known sampling locations for reference samples in self-assignment or the revealed locations for previously blinded samples. In addition, such knowledge in a forensic context could also be reliable documentation, or harvest and transport evidence. It even might be at times intelligence about specific regions currently targeted by illegal logging or fishing, or about illegal trade networks and their sources.

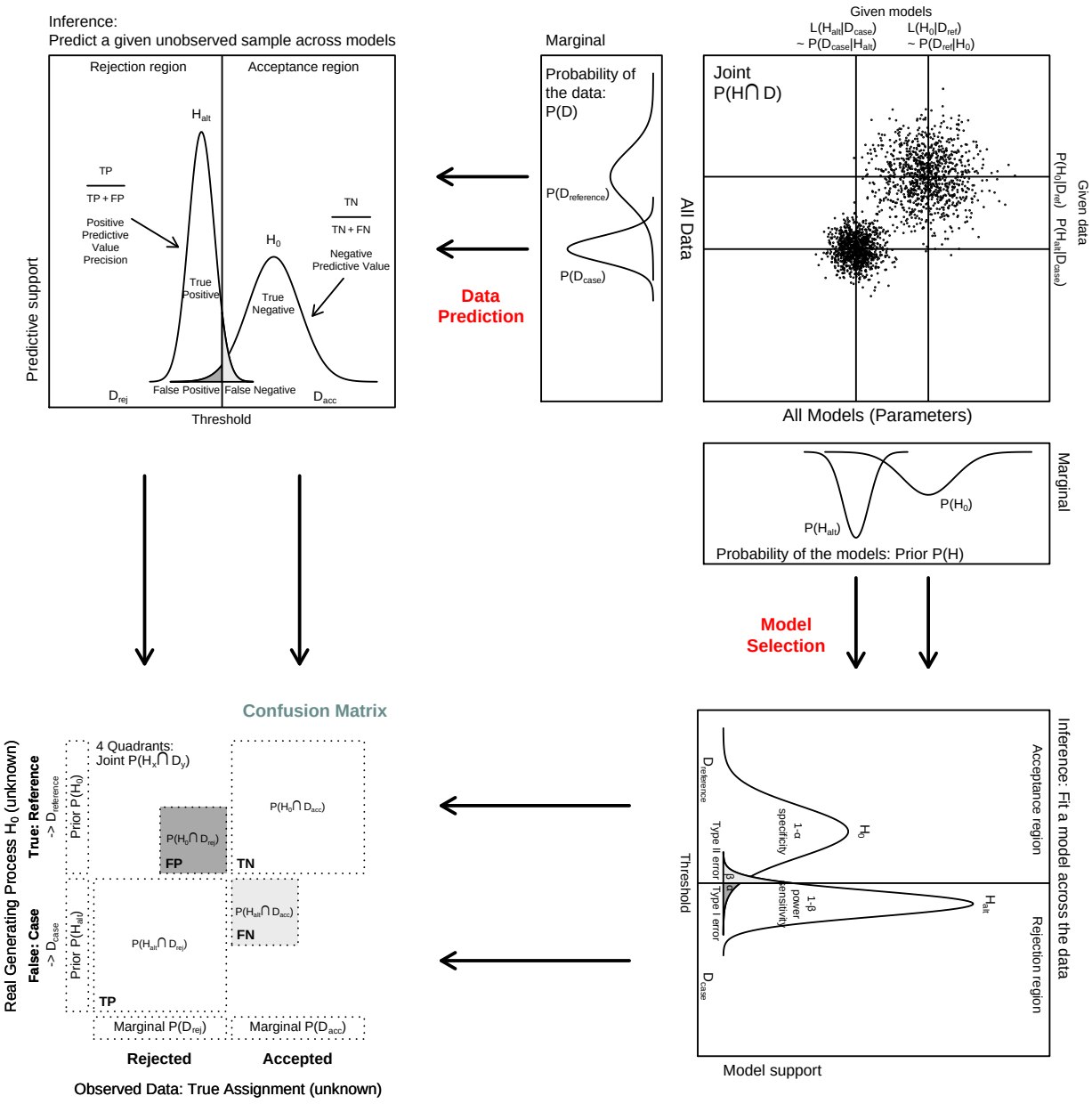


fig. 3.3: Inference approaches focusing on models or data. Starting point is the joint probability density distribution of models and data (samples) representing reality (upper right corner), with the marginal distributions of the two processes (models) and the two populations (data) shown. Moving down (lower right), the best models for the null hypothesis and an alternative hypothesis (e. g. a case-specific hypothesis) are independently inferred and specified (Route 1 of Model Checking). Fully specified, the models then can be tested. Predictive data distributions (upper left) can be generated using the inferred models (from the lower right) or independently pre-specified models to assign given samples (data points) to their populations of origin (Routes 2 and 3 of Model Checking). Both, model- and data-focused approaches have an impact on the reconstruction of the decision matrix ("confusion matrix"; lower left). It forms the basis for estimating overall accuracy, error rates and predictive values, which inform applied decisions and risk assessment. TN: true negative, FP: false positive, TP: true positive, FN: false negative.

3.2.3.1 *Route 1*: Comparing model support to independent information

In the first route (cp. fig. 3.2, right-hand side), the maximum likelihood estimate or the posterior probability distribution of an inferred result is compared to independent information. These measures either represent support for inferred overall model configurations or support for inferred values of specific model parameters of interest. Non-supervised Bayesian clustering approaches that, for example, are based on the Structure-model (Pritchard et al. 2000), follow this route. Such model-based clustering approaches infer the best-supported overall model for the global population structure represented in the dataset, based on the likelihood or the posterior probability.

This "best" overall model - the maximum likelihood estimate (MLE), the maximum a posterior estimate (MAP) or the mean of an MCMC sample of the posterior probability distribution of the global population structure - is then checked by using information available for the individual samples in the dataset. For this purpose, the details of the reconstructed cluster membership or admixture proportions of each individual are compared to known background knowledge recorded for the sample, e. g. the known geographic or taxonomic origins of reference samples. Thus, while these diagnostic checks focus on the correct or erroneous clustering of individual samples, in sum they inform on the fit and appropriateness of the inferred global population structure model, which was selected based on it showing the significantly best support among a set of evaluated model configurations.

3.2.3.2 Predictive distributions: exploring, learning and improving the usefulness and limits of the best fitting model

The just described first route is **model-focused** and uses the support and its variability as measures of confidence in the inferred model and its parameters. In contrast, the remaining two options towards model checking are **data-focused** and evaluate the capability of an inferred model to predict so far unobserved data (fig. 3.3). They characterize the confidence one can have that, for example, a future case sample originated from and thus can be predicted by a reference population, represented by an existing collection of samples (the data) in conjunction with an associated evolutionary process (approximated by a population model). An inferred well-fitting population model that describes a reference sample representative of a population of origin well, at the same time should be able to generate characteristic samples for this population through simulation. Thus, given a case sample that truly originated from this population, an inference approach using this population model should predict the case sample with high reliability.

Statistical approaches that examine the capability of a model to generate "realistic" unobserved or future data are based on **predictive distributions**. These are generated by employing the inferred best fitting model for the observed real dataset at hand to simulate new, predicted data points and datasets. Hence, the predictive distribution builds on the previously carried out

inference of a model with highest support from the available data and the evaluation of its variability (see Route 1). It now takes the next step and evaluates how well this best-supported model actually fits the data.

Model checking employing predictive distributions in evolutionary biology and genetics does not answer the question, if the real data come from the assumed (e. g. inferred population diversity and structure) model and, thus, if the model is "true". Here, the answer is clearly no. Evolutionary history and evolved biodiversity patterns remain too complex to be captured in a (single) model. Their realities and intricacies are only inferred in puzzle pieces. Instead, the goal of employing predictive distributions is, to characterize the fit of the model to the data through evaluation of the model's predicted results. Thereby it becomes possible, first, to obtain more detailed insight into which patterns and processes of reality were not captured by the model, resulting in discrepancies between the data and the inferred model (Gelman et al. 2014 p. 141). Second, these discrepancies can be quantified to assess whether they could have arisen by chance, under the model's own assumptions (Gelman et al. 2014 p. 151). Model checking, thus, is an important tool for learning from the discrepancies between the model and the data, while searching for "specific, consequential errors" (Gelman & Shalizi 2013 p. 21).

For the assessment of an inferred model (and in extension the employed inference approach) using predictive distributions, new replicate datasets are simulated "from" the predictive distribution (Mimno et al. 2015) employing the previously inferred, and thus now fully specified model. In the **frequentist framework**, an inferred maximum likelihood estimate (MLE) for the model and its parameter values can be used to fully specify the simulation model. Parametric bootstrapping then generates all the replicate datasets that form the predictive distribution (Efron 2012, Gelman & Shalizi 2013). The resulting predictive distribution is based on a single MLE value for each model parameter. It thus provides insight into uncertainty present in the data, but does not inform on the uncertainty associated with the specifications of the model parameters. Therefore, the generated replicate datasets might have characteristics that are too narrow or biased compared to the original dataset and only form a limited representation of uncertainty present in the inference approach.

In the **Bayesian framework**, a corresponding posterior predictive distribution (PPD) is generated by using parameter values for the specification of the simulation model that are drawn according to the posterior probability distributions for the parameters, which were inferred from the observed, real dataset. These posterior probability distributions of Route 1 take into account only the uncertainty associated with the inference of the model. The subsequent posterior predictive distribution, in addition, provides also insight into the uncertainty present in the (reference) data. Uncertainty in the data can be present, for example, in the form of lack of signal, noise and conflict. Consequently, the PPD evaluates the joint probability distribution of model and data, including the inference approach (Gelman & Shalizi 2013, Gelman et al. 2014 p. 141). Moreover, this encompasses insight into the appropriateness of and uncertainty associated with the prior. This is the case, since the prior is connected to and can be deduced from the data through their

marginal distribution (Gelman & Shalizi 2013). The PPD thus provides the opportunity to evaluate all components and steps of a statistical inference approach.

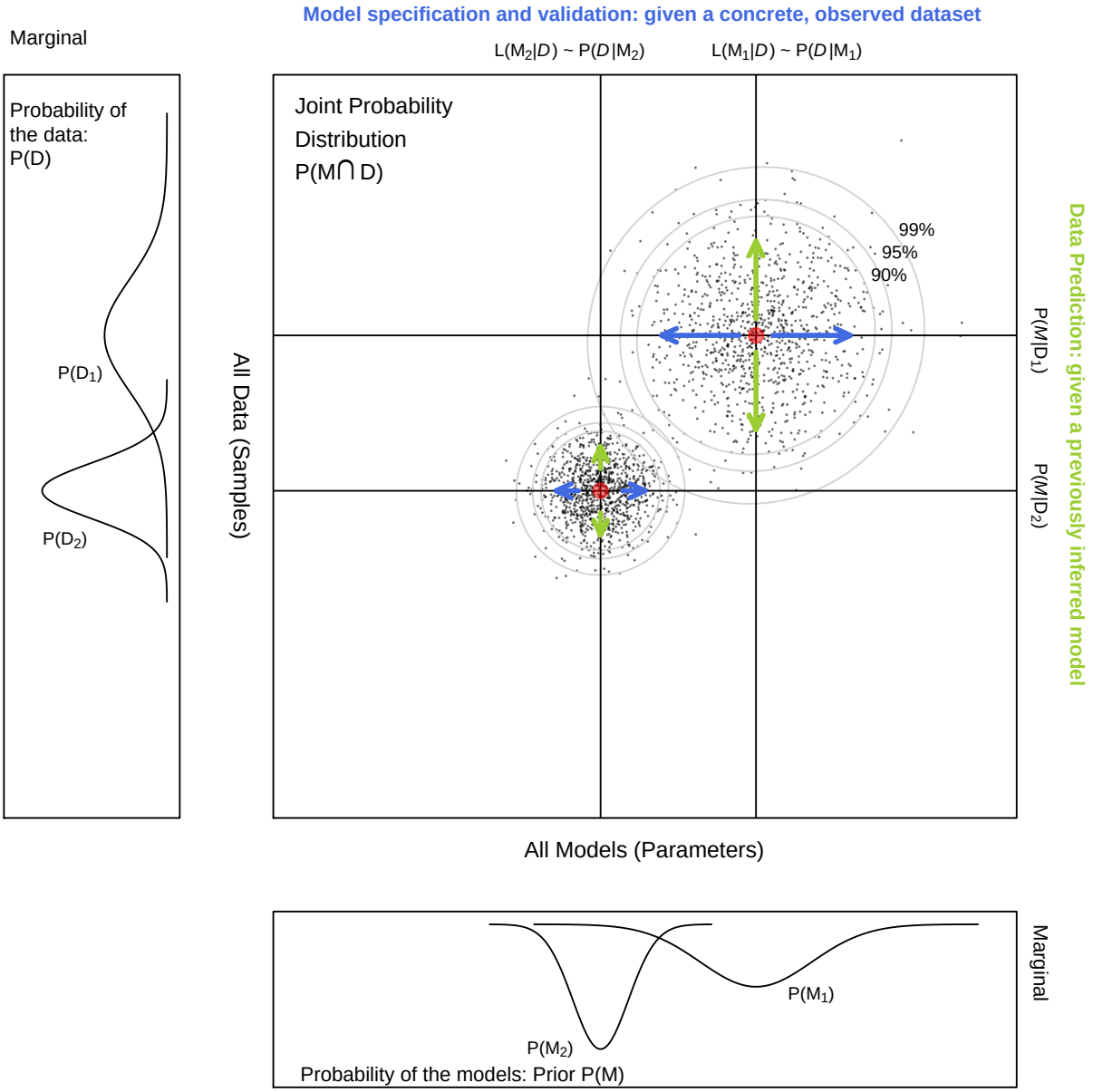


fig. 3.4: Evaluating confidence (uncertainty) in inference of reality. Blue arrows: evaluating uncertainty due to the model. This uncertainty can be represented, for example, by inferred posterior distributions of parameter values (cp. Route 1). Green arrows: evaluating uncertainty due to the data (cp. Route 2 and 3). Uncertainty due to the data can be inferred using resampling approaches, for example, bootstrapping in the frequentist framework. Bayesian posterior predictive distributions provide insight into uncertainty due to both, the model and the data, at the same time. *Italics* denote, if the data or model are given in the evaluation of uncertainty.

Focusing on the model, the PPD characterizes the mode and dispersion of the forecast engendered by an inferred model and thus the extent of its fit to the available data. With regard to the data, the PPD is a representative collection of future data (points, sets, events), which are predicted to occur, based on the knowledge gained and derived from an already available collected sample. These two views are linked, since statistical analysis condenses existing data into an explanatory (process) or descriptive (pattern) model through inference. Thus, taken together, the PPD provides information on the characteristics of data, which are expected to exist or to occur in the future, if the inferred model were true.

From a **practical point of view**, the (posterior) predictive distribution sheds light onto the question of how useful the inferred model is for application, and how it might be further improved. The adequacy of the predictive distribution is checked by comparing it to independent knowledge or back to the original data from which it was derived (self-consistency check). Together the original data and the inferred model are considered to be of use for providing insight into reality, if the original data and/or independent information fall within the distribution of predicted data.

The procedure of this evaluation begins with analyzing the simulated replicate datasets with the same inference approach as was used for the observed, real dataset. Then, one to several **discrepancy measures** are defined to provide insight into specific characteristics of the genetic diversity and its patterns in the real and simulated datasets. The distributions of these measures can be used to evaluate any discrepancy or concordance to independently available information (Route 2). Additionally, or alternatively if no independent information is available, the discrepancy is evaluated only between, first, a measure calculated from the inferred result of the original, observed dataset and, second, the distribution of the measure when it is calculated from each of the results of the simulated datasets of the predictive distributions (Route 3). If (large) discrepancies are found for core measures of importance to the application task or for too many characteristics, then the fit of the model to the data at hand is considered to still be insufficient. The model needs to be further improved to be of use for application. For this purpose, the specific diagnostic measures showing discrepancies provide guidance for a next step in improvement.

The evaluation of usefulness by representing prediction uncertainty in the form of a predictive data distribution can form an integral part of **hypothesis testing**. Hypothesis tests based on predictive distributions are focused on data events, not parameter values. Such tests are the rationale and foundation for the assignment of case samples or pools of case samples to existing reference data and their associated inferred models, that is, evolutionary processes. Predictive distributions in this way are the hallmark and distinguishing functionality of classical assignment and exclusion approaches (Route 2). Yet, the generation of predictive distributions in the procedures has not been explicitly named as such. Thus, predictive distributions have been used only incidentally in genetic population assignment. The inherent application of predictive

distributions differentiates assignment from clustering methodologies. However, this is changing, see e. g. the approach implemented by Mimno et al. (2015) described below (Route 3).

3.2.3.3 *Route 2*: Predictive distributions and independent information: supervised population assignment approaches

The second route compares a (posterior) predictive distribution to preexisting information, which is independent of the current analysis. Hence, the predictive distribution or prognosis engendered by and generated from the inferred model is compared to known reality. In this way, the fit of the inferred model to the data and thus its usefulness are assessed. Classical supervised assignment studies that conduct **self-assignment** of reference samples or assign **blinded samples** follow this route (e. g. Manel et al. 2002, Baudouin et al. 2004). Here, the empirical reality for checking the investigated population models and their delimitations (supervised assignment) is given by the known origins (e. g. sampling localities) of the self-assigned or blinded samples. Thereby, a procedure is available and implemented for external validation of the assignment approach.

Self-assignment as implemented to date, in for example GeneClass2 (Piry et al. 2004), combines cross-validation with the generation of a predictive distribution for the population under consideration. The predictive distribution forms the basis for estimating the within-sample predictive accuracy, while cross-validation, in addition, attempts to quantify the out-of-sample prediction error (Gelman et al. 2014 p. 169-170), which is also called the generalization error. Together they provide an estimate of overall prediction accuracy. The in this way estimated **overall predictive accuracy** forms the context for forensic casework by providing an evaluation of the reliability with which a case sample is assigned or excluded from a proposed population of origin.

In inference approaches for supervised genetic population assignment, **cross-validation** implements the leave-one-out strategy. Here, each sample is in turn removed from the reference set for a population and later used as test sample. At the cross-validation step, the population diversity model is inferred. In this step, the population allele frequencies are the key parameters that need to be estimated. For this purpose, frequentist and Bayesian approaches for population allele frequency estimation have been developed and are available (Paetkau et al. 1995, Rannala & Mountain 1997, Cornuet et al. 1999). The inferred values for the allele frequency parameters are optimized for each evaluated cross-validation repeat. Hence, in leave-one-out cross-validation, the inferred population model and, thus, the predictive distribution generated from it, are specific to the assignment evaluation of each in turn tested reference sample (the one that is "left out").

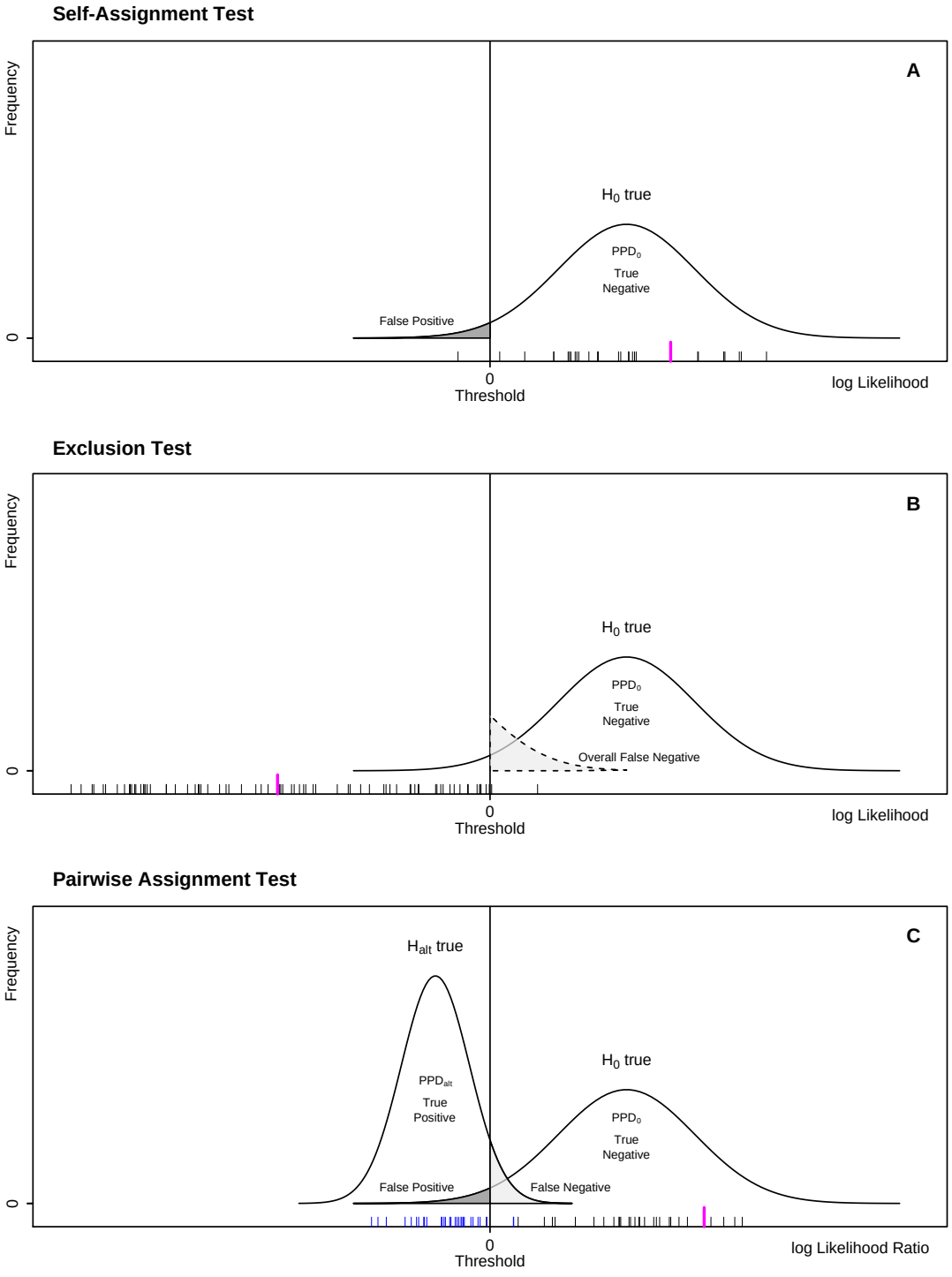


fig. 3.5: Supervised self-assignment and exclusion tests using reference data. Each panel shows the predictive distributions generated within one cross-validation step for the test sample marked in magenta. The tick marks below the distributions represent the likelihoods for all observed, real samples. In panel C, the tick marks of the focal population are in black and magenta, those of the alternative population in blue. PPD: posterior predictive distribution. H_0, H_{alt}: the null and an alternative hypothesis, respectively.

The next step generates the **predictive distribution** (fig. 3.5A) for the assessment of within-sample predictive accuracy by simulation from the now completely specified population diversity model or, alternatively, by bootstrap resampling from the cross-validation sample set (the original reference dataset minus the currently excluded test sample). The currently implemented approaches to supervised population assignment in evolutionary genetics generate the predictive distribution by parametric bootstrapping that often takes the form of non-parametric resampling of markers in implemented applications (cp. Paetkau et al. 2004). In any case, the result is a collection of simulated multilocus genotypes. Each of the simulated genotypes represents a repeat in the predictive distribution.

In the **hypothesis-testing** step, testing can be performed separately for each population of origin and its reference samples, treating each population independently, without requiring or considering an alternative population of origin. At this stage, the tested reference sample is set into relation to the predictive distribution generated by the population diversity model inferred for the current cross-validation repeat. For this purpose, a statistic, the discrepancy measure, is calculated for each sample. The likelihood (for computational reasons calculated as the logarithm of the likelihood) for the population diversity model given a multilocus genotype is calculated as discrepancy measure for both, the genotype of the tested reference sample and each simulated genotype. Subsequently, the placement of the test genotype's log likelihood with regard to the distribution of log likelihoods of the simulated genotypes is noted. Its placement relative to the distribution can be evaluated either graphically allowing visual comparisons and inspections, or by calculating *P*-values.

In the current implementations, the tested sample's assignment to or exclusion from the population is decided by **frequentist significance tests** based on *a priori* chosen significance levels (see section 5.3.1). The evaluation is carried out in a frequentist framework, since the predictive distributions are based on the maximum likelihood point estimators as the basis for specifying the population allele frequency parameters, and not on the Bayesian posterior probability distributions, which would produce a PPD. In the case that the *P*-values derive from the comparison with a Bayesian PPD, they can be interpreted directly as posterior probabilities. However, no fully Bayesian approach is implemented to date. Thus, even if Bayesian model estimation is chosen, in for example GeneClass2, the overall procedure results in only a "partially Bayesian" assignment approach (e. g., Manel et al. 2002 p. 651).

If the population diversity model fits the genotype of the test sample, the discrepancy measure for the tested genotype falls within the distribution of the discrepancy measures calculated for the genotypes predicted by the model. In this case, the population is considered a potential source for the sample. Subsequently, the conclusion from the hypothesis test for each test sample is **verified** using the independently known sampling location or taxonomic origin of this tested reference sample. In this way, self-assignment approaches evaluate the fit of the inferred genetic composition of a given population (the model) to the genotype of an independent test

sample (the data) from this population. In self-assignment or with blinded samples, it is thus possible to check and validate the assignment result against the known origin of the sample.

Finally, this external validation provides the basis for **summing all correct and incorrect assignment conclusions** for the tested reference samples belonging to this population (assignment) or to other populations (exclusion). The sums are expressed as fractions of the total number of tested reference samples for each given population. The fractions for the correctly assigned or excluded samples are the overall predictive assignment and exclusion accuracy, respectively. They characterize as quality measures the employed inference approach, including dataset, population diversity model and inference method.

The fraction of incorrectly excluded samples at a specific significance level α estimates the type I error rate (false positive rate; fig. 3.5A), the one for incorrectly assigned samples provides a general estimate for the type II error rate (β, false negative rate) conditional on the overall reference dataset. A **general exclusion test** can show the overall power of the reference dataset to falsify an incorrect assignment (fig. 3.5B). In this procedure, all non-local reference samples are tested for each of their non-native populations in the reference set. This option allows an evaluation of the case that the true population of origin was not sampled and thus is not present in the reference dataset. Here, the expectancy is that a tested sample is not assigned incorrectly to one of the non-native populations present. A detailed consideration and discussion of five possible cases of sets of error and assignment combinations can be found in Cornuet et al. (1999).

In general, the probabilities of correct assignment and exclusion (as characterized by high overall predictive accuracy combined with low error rates) should be high, if the data are sufficiently informative, the discrepancy measure is effective for evaluating assignment, the population model well-fitting the data, the inference procedure appropriate and the resampling approach within the cross-validation steps adequate for estimating the generalization error.

3.1 Combining global population clustering and supervised assignment

One opportunity to improve assignment results is to incorporate recently developed models into the GeneClass2 resampling framework. Taking advantage of the rich information content in GWS data and new methods to access it, these models and inference approaches calculate and differentiate in more detail the different levels of coancestry, from close kinship to population structure (e. g. Manichaikul et al. 2010, Stevens et al. 2011, Browning & Browning 2012, Moltke & Albrechtsen 2014, Conomos et al. 2016). The strength of the original conceptual assignment framework was from the beginning intended to be its flexibility to accommodate all kinds of distance measures or population-genetic models to compare and test population structure and the origin of samples.

The concept is, thus, similar to the framework that Lawson and Falush (2012) developed for describing and comparing the information content of genetic relationship matrices. Their non-supervised clustering approaches fineStructure (Lawson et al. 2012) and fineRADstructure (Malinsky et al. 2018) quantify genetic population structure and identify populations. This framework for global population clustering could be combined with the supervised assignment framework, since both conceptual environments complement each other. Merged they would form a general framework for the evaluation of population structure and assignment reliability. Such a framework is similar to, but expands the USEPOPINFO and PFROMPOPFLAGONLY options in Structure (Pritchard et al. 2000) and the posterior predictive check approach (see section 3.2.3.4) implemented by Mimno et al. (2015).

Thereby, support for population clusters taken from the posterior probability distribution of fine(RAD)Structure could provide hyperparameters (see section 3.3 “Congruence”) for generating the posterior predictive distributions that form the ideal foundation for supervised population assignment. Having the specific objective of testing the origin of individual samples (or pools of samples), supervised population assignment quantifies error rates and predictive values associated with explicit forensic case hypotheses. The forensic case hypotheses in applications correspond to supervised population definitions. The posterior probabilities of the global clustering approach quantify how likely it is that these a priori assumed population entities or similar (sub-) clusters really exist. Their integration into the population assignment testing routine is made possible through hierarchical Bayesian modelling (see section 3.3 “Congruence”).

Merging both frameworks would expand in this way the point estimators returned by GeneClass2 to more informative credible regions. This is the case, since sampling the clustering posterior probability distribution would include a wider region of the result space with regard to population structure in the calculation of the final result.

Posterior predictive values: extending predictive distributions to pairwise assignment tests

The pairwise self-assignment approach developed by Ciampolini et al. (2006) takes assignment based on the predictive distribution one step further and allows the calculation of power and thus of posterior predictive values – under certain assumptions regarding population priors. The approach takes into account type I and type II error rates estimated in a pairwise self- versus non-self-assignment procedure (fig. 3.5C). This pairwise assignment approach is based on two predictive data distributions that are generated by the population models for a focal population of origin (H_0) and a **specific alternative population** (H_{alt}).

The cross-validation step is performed for all reference samples of both populations under consideration. For each cross-validation test sample, predictive repeats (simulated genotypes) are generated for both populations (fig. 3.6). For each predictive repeat (simulated genotype),

two multilocus-genotype likelihoods are calculated, one based on the population model that generated the repeat (either H_0 or H_{alt}), the other using the opposite population model. The two likelihoods of true versus false assignment are combined into the **log likelihood ratio** (log LR) as discrepancy measure. The log LR is calculated with the focal population of origin (H_0) kept fixed for both predictive data distributions. That is, the log likelihood ratio is in both cases log LR = log $L(H_0|D)$ – log $L(H_{alt}|D)$, with D either being a simulated predictive repeat of the population of origin (true vs. false assignment) or a predictive repeat of the alternative population (false vs. true assignment). Two log LR distributions are the result, one for the predictive repeats generated by H_0, one for those generated by H_{alt}. The expectation is that the resulting log LR distribution for the predictive repeats of H_0 is mostly positive, while for the predictive repeats of H_{alt} the log LR distribution is expected to be mostly negative. The characteristics and quality measures of the pairwise hypothesis tests are based on the overlap between the two log LR distributions and their placement with regard to a given threshold, which represents the chosen significance level α (type I error rate). Verification, again, is based on the placement of the real data genotypes of the collected reference samples with regard to the two distributions and the known geographic origins of the reference samples.

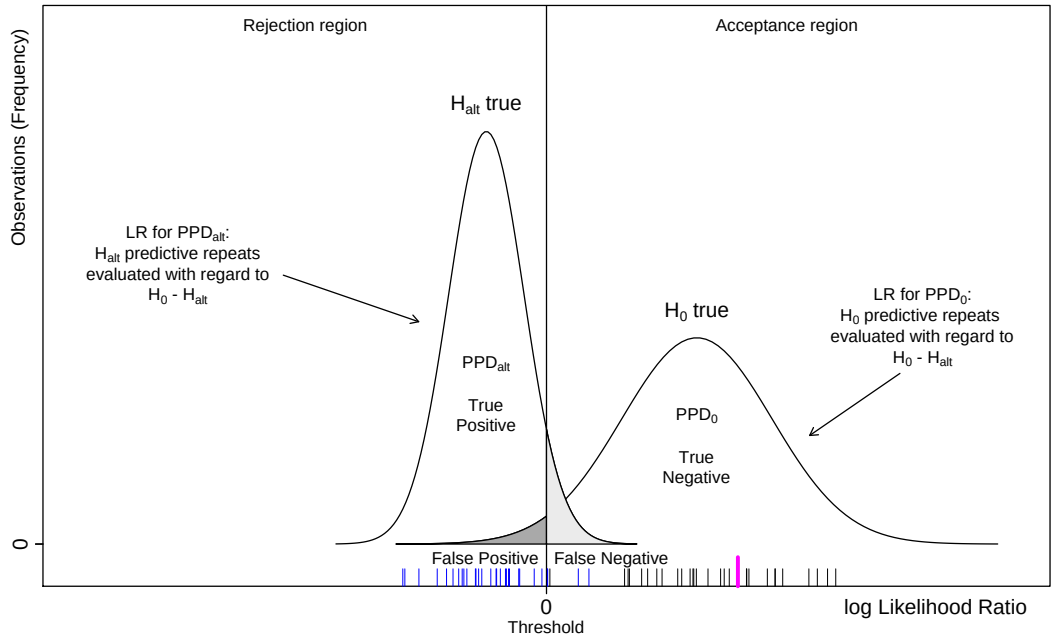


fig. 3.6: Pairwise assignment graph explaining the generated predictive distributions and discrepancy measures. The depicted predictive distributions were generated within one cross-validation step for the test sample marked in magenta. The tick marks below the distributions represent the likelihoods for all observed, real samples. The tick marks of the focal population are in black and magenta, those of the alternative population in blue. PPD_0, PPD_{alt}: posterior predictive distributions generated by the null model (focal population) and an alternative model, respectively. H_0, H_{alt}: the null and an alternative hypothesis, respectively. LR: likelihood ratio.

Ciampolini et al. (2006) compare the assignment quality observed for three different log likelihood ratio thresholds that function as significance levels (log LR = 0, 1 or 2). These accepted type I error rates in turn define the power (1-β, with β being the type II error rate) of the test. The false positive error rate (type I error) and the false negative error rate (type II error) together with the estimates for specificity (correctly assigned) and sensitivity (correctly excluded) fully populate the **confusion matrix** (see fig. 5.1 in section 5.3.1) that forms the foundation for decision analyses and risk assessment.

This confusion matrix forms the basis from which the posterior positive and negative predictive values can be calculated. An important prerequisite of the approach, as mentioned and implemented by Ciamploini et al. (2006), is that it assumes **equal priors** for the populations that are investigated (Ciampolini et al. 2006, Wilkinson et al. 2012). If there is substantive knowledge that the priors differ between the populations, then the population priors need to be factored into the calculations. Different priors can be expected, if the abundance of the two investigated populations is known to be different. This is the case, for example in European white oaks, when the task is to differentiate between core species individuals and the generally much less abundant hybrid individuals between the closely related species in this section of the genus *Quercus*.

In the study of Ciamploini et al. it became apparent by visualizing the predictive distributions that the inferred population models did not fit the investigated Italian cattle breed reference populations well. The log likelihood ratios calculated from the predictive data distributions deviated too much from those based on the real samples. The authors solved this discrepancy to the real data in the following way. Instead of simulating predictive data for inferring the log likelihood ratios, they calculated the log likelihood ratios based on all original, real reference samples of the populations considered in the pairwise population comparisons (not on any simulated predictive genotypes). These distributions (i. e. histograms) of observed log likelihood ratios were tested for normality and could subsequently be assumed to be normally distributed. The means and standard deviations of these **normal distributions** were estimated. They provided approximations for the characteristics of the predictive data distributions. The type I and II error rates were computed based on these means and standard deviations in pairwise comparisons.

Ciamploini et al. used the available independent knowledge with regard to breed origin for the **observed, real data** to specify the two normal distributions that approximated the predictive distributions. They however did not verify the estimates for overall predictive accuracy, the error rates and predictive values using independent samples. Their results, therefore, might have been susceptive to the overfitting of the population models to the available reference datasets. However, this pairwise assignment approach could easily be extended by a **cross-validation** loop for the assessment of the out-of-sample misclassification or prediction error.

Application of predictive distributions to unknown samples with independent information

This second route for model checking generally is possible only for self-assignment or blinded samples. However, sometimes it might be possible to verify an inferred assignment through independent information also in the assignment of individuals of unknown origin. This was possible in an assignment study for fish caught from a mixed natural population (Hauser et al. 2006). Hauser et al. used knowledge from previous fin clippings for the verification of assignment of wild versus captive-bred steelhead (rainbow trout, *Oncorhynchus mykiss* Walbaum) at a location along the North American Northwest coast. They checked the results of a wide range of assignment (and clustering) programs against the independent fin-clipping information that indicated origin.

3.2.3.4 *Route 3*: Posterior predictive checking

In the third route, the model that was inferred from the observed, real data by the Bayesian inference approach of choice, is assessed using its PPD without recourse to an independently known reality (posterior predictive check, PPC; Gelman et al. 1995 pp. 161 ff., cp. also Bollback 2002). In posterior predictive checking, the predictive data points and sets of the PPD are only compared back to the original, real dataset. Posterior predictive checking, thus, is a self-consistency check (Gelman et al. 2014 p. 143).

This third route takes advantage of the possibility that once a forecast was generated based on the inferred model, the adequacy of the predictive procedure can be assessed by going back to the already existing, that is, the original data. The original data that gave rise to the inferred model should still look plausible in comparison to the collection of subsequently generated replicate results from the posterior predictive distribution. If this is the case, the overall fit of the inferred model is judged adequate for the dataset and useful for the task at hand.

Posterior predictive checking only examines if the inferred model is adequate with regard to the original dataset. Its aim is to explore, if the most important characteristics of the investigated and pertinent evolutionary processes and, thus, the most discriminating and informative patterns in the data for the (assignment or clustering) task at hand are captured by the model (Gelman et al. 2014 p.180). In sum, it evaluates that the model's application to the given data itself makes sense.

It might be confusing that the simulated replicate datasets of the PPD are sometimes called "empirical" (e. g. by Mimno et al. 2015) in the context of the PPC. Obviously, these simulated data are not observed, real samples from the existing, empirical reality. The reason for this term is a statistical one. Under the Bayesian paradigm, the PPC represents an "**empirical Bayesian**" approach. Empirical Bayesian strategies can be considered as approximations of fully hierarchical Bayesian inference approaches (Gelman et al. 2014 p. 104; see also section 3.3 "Congruence"). In

hierarchical Bayesian analyses, values for lower-level hyperparameters are fully integrated out across the complete marginal distribution of the model (i. e., independently inferred or known priors) representing the complete (as defined by the researcher) parameter dimension of the joint probability distribution. In empirical Bayesian approximations, these hyperparameters instead are set to their most likely values, often the MLE. In the PPC, the parameter values of the simulation model for the generation of the PPD are chosen according to their previously inferred posterior probability distributions. Therefore, they are integrated out only with regard to these posterior probability distributions. While these posterior probability distributions represent more of the variability of the full prior than a MLE point estimate, they are still limited in that their inference was specific to the concrete, "empirical" dataset at hand. They are dataset-dependent and not integrated out across model posterior probabilities inferred from all possible data (the marginal of the data). Thus, the priors for the simulation approximate P(Model) by using P(Model | Data). They are "empirical", due to being conditional on the given dataset, resulting in PPD repeats that are called "empirical".

In empirical Bayesian analyses, the data therefore are used twice in a seemingly circular way. Gelman (2014) takes the perspective that this is only a problem in the context of the PPC, if the PPC is conducted with the goal of hypothesis testing. In that case, potential overfitting is a possibility and needs to be considered. Circularity is not a problem, if the goal of performing a PPC is to learn about the model's fit to the data, so that it can be explored, improved and its usefulness validated. The inherent circularity of the self-consistency check differentiates Route 3 from Route 2. In supervised self-assignment approaches following Route 2 no repeated use of the data occurs, since fit and out-of-sample error are assessed using independent knowledge and cross-validation, respectively.

Implementation of posterior predictive checking for population structure and admixture inference

Mimno et al. (2015) developed and implemented posterior predictive checking for population structure and admixture inference that employs model-based clustering approaches. Their approach provides not only a method for diagnosing model misspecifications and limited fit arising from specific parameters, but importantly also quantifies the effects of these on inference results. Mimno et al. explicitly refer to the clustering program Structure (Pritchard et al. 2000), which is widely used in evolutionary and conservation genetics. The population structure and admixture estimates produced by Structure can be used as input to their procedures. They implemented the Bayesian PPC approach to assess the fit and usefulness of the originally inferred global model and provide the opportunity to explore in more detail the clustering results. The results of the PPC, thus, can inform further development and refinement of the reference sample set, marker choice, evolutionary model and inference strategy to enhance population differentiation, identification and assignment.

Following the general procedure for generating a PPD, the simulation model for each generated repeat is specified by resampling parameter values from the posterior distribution of the inferred "best-fitting" overall, that is, global, population structure and admixture model for the reference dataset. As implemented by Mimno et al. the approach focuses on population-level discrepancy measures. Its discrepancy measures are population summary statistics that mostly diagnose patterns at the level of inferred clusters. Based on these population-level discrepancy measures the fit of the global population structure model to the total reference dataset is explored and avenues for improvement can be identified.

In contrast to supervised assignment approaches, no discrepancy measure has been implemented by Mimno et al. that evaluates predictive assignment accuracy for individual samples to the inferred clusters. Accordingly, the approach does not check the appropriateness of the global clustering and admixture model and its clustering results at the level of individual samples. An additional evaluation of predictive assignment accuracy at the level of individual samples could be a future extension (cp. Gelman et al. 2014 p. 152-153 and the suggestion for an extension of the fineStructure approach in the box after Route 2 above). Such an extension would provide important insight into the relationship between membership and admixture proportions of individual samples and the predictive accuracy for assigning these samples to clusters. Such knowledge and experience is completely missing to date.

3.2.3.5 Summary: Model checking procedures in forensic conservation genetic casework

In forensic conservation genetic applications based on clustering and population-level assignment, there is no independent ground truth. Without directly applicable independent knowledge as, for example, present in self-assignment, direct verification of specific inferred results is not possible. The reliability of a forensic conclusion for a case sample is indirectly presumed, based on several levels of model checking. First, Routes 1 and 3 characterize and explore the applicability and usefulness of the inferred global population and admixture model for the natural populations represented by the reference dataset. Subsequent validation of the identified clusters or of independently defined populations in the reference dataset can be performed using the supervised self-assignment approaches described for Route 2. At the same time, conducting Route 2 allows the identification and quantification of error rates that then are transferred to casework to provide an indirect evaluation of the assignment results. Finally, the usefulness of a forensic assignment conclusion in casework can be more directly determined using the PPC approach of Route 3 as self-consistency check. Here, the case sample in question is excluded and included, respectively, as part of the dataset for the inference and checking of the global and local (i. e. supervised) population structure and diversity model.

3.3 Congruence

In investigations of empirical data for which the evolutionary truth cannot be known, in addition to evaluations of robustness, the congruence of results provides information about the veracity (validity) of the components of a complex, multi-step inference approach and the reliability of its conclusions (Hillis 1995). While robustness mostly assesses the assumptions of specific parameter settings, generally up to nested models, congruence asks more widely if the results are reproducible when independent, unrelated ways are taken to approach a question. It considers and compares the conclusions drawn from across different evolutionary models, data types and sets, inference approaches, software implementations, and analyses performed by different investigators. The congruence of results is considered to add additional support and credibility to the obtained separate results.

In assignment testing, traditionally, the congruence of results has been approached through the union of assigned and excluded populations of origin into two sets and the subsequent intersection of the two sets (Cornuet et al. 1999, Ciampolini et al. 2006). Furthermore, congruence can be assessed based on (non-) overlapping confidence intervals for parameter estimates from separate analyses (Hillis 1995). For example, in an intuitive, informal way this is applied when admixture proportions are compared between Structure runs based on different settings, e. g. for different numbers of modelled clusters.

3.3.1 Congruence versus total evidence

There exists tension between inference following total evidence versus congruence principles. Both analytical strategies can be pursued, if multiple inference approaches can be applied to the same dataset, or if several data partitions and/or subsets of samples are present in a dataset (see also Efficiency and Sufficiency below). It might seem most straightforward to use the most extensive set of samples and the maximal available genomic dataset with the most inclusive, parameter-rich model, explicitly incorporating and estimating all potential parameters (an exhaustive “super-model”, see Gelman et al. 1995 pp. 161-162). However, it is of principal importance that such a total evidence approach considers and adequately models data heterogeneity, if distinct and variable evolutionary processes had an impact on the data (Grueber et al. 2011, Leigh et al. 2011, Gori et al. 2016). In that case, parameter-rich models that try to capture and adequately represent a multitude of contributing processes need to take into account parameter-dependencies in the form of hierarchical model structures (see below). Parameter-rich approaches that do not represent the generating processes appropriately are prone to overparameterization and overfitting (structural or *a priori* non-identifiability) and to identifiability issues due to insufficient information present in the data (practical or *a posteriori* non-identifiability; Ashyraliyev et al. 2009).

On the contrary, separate analyses can be geared specifically towards and optimized for one type of data, or a specific parameter or process. While sometimes simplifying otherwise, they might be more accurate and precise with regard to the factor in focus, and thus more informative and decisive. The array of specialized phylogenetic and population-genetic approaches that in such a way can contribute to and are informative for forensic conservation genetic questions is huge and very divergent with regard to analyzed parameters and statistical frameworks. However, if and how subsequently best to combine the results of these specialized separate analyses into one integrated meta-result, is not always immediately apparent. One solution often taken is to qualitatively (visually) describe and compare the results from different analyses (including a total evidence approach). In this framework, congruence of independent results provides powerful evidence. Nevertheless, for complex systems, methodological development moves towards the direct integration of multiple specialized analyses and the simultaneous analysis of multiple data partitions for a set of samples (cp. model expansion to continuous parameter definitions, Gelman et al. 2014 pp. 184 ff.).

Integration of multivariate analyses into inferences of coancestry and functional genomic variants

An area of active development is the direct incorporation of results from multivariate analyses into subsequent inference procedures. For example, principal components summarizing and representing coarser levels of genetic diversity (e. g. population structure) are used directly as parameters in inference steps for specific finer-grained parameter estimation (e. g. familial kinship), and vice versa. In this way, it has been shown to be possible to differentiate and quantify concurrent signals of ancient, population-background and recent coancestry contributing to genetic diversity. Applications of such approaches have been implemented for identifying and quantifying simultaneous influences of population structure, admixed ancestry, familial relationships and selective forces (Patterson et al. 2006, Price et al. 2006, Lawson & Falush 2012, Conomos et al. 2015, Conomos et al. 2016, Zheng & Weir 2016).

3.3.2 Hierarchical Bayesian inference models

Hierarchical Bayesian modelling provides a framework for incorporating heterogeneity in processes into explicit, parameter-rich models (Gelman et al. 2014 pp. 101 ff.). It reduces the problem of overparameterization (overfitting) by representing dependencies between parameters. Here, the distributions of finer-level parameters (for example, mutation rates specific to genomic partitions) are themselves assumed to be random draws from an unknown underlying distribution. The thus incorporated dependencies, defined by the underlying distribution, make it possible to “borrow strength” at a quantitatively appropriate level across partitions for overall inference (cp. also Felsenstein 2001, Beaumont & Rannala 2004, Eddy 2004, Felsenstein 2004 pp. 537-538, Ji & Liu 2010, Moore et al. 2014). Such a hierarchical Bayesian approach is implemented, for example, in the assignment program SCAT (Wasser et al. 2004), in

which a geostatistical Gaussian process models spatial dependencies of allele frequencies across all sampled sites. In addition, a hierarchical strategy was used to extend the program by borrowing strength across individual African elephant tusks to sized pools of tusks (Wasser et al. 2007). In a project augmenting the forensic context, hierarchical modelling strengthens the inference of the hybridization processes between the closely related taxa of African elephants and the patterns arising from it, thereby improving the identification of hybrid individuals (Mondol et al. 2015). In sum, fully probabilistic hierarchical models provide the advantages of integrating and appropriately weighing heterogeneous information; of thus stabilizing the overall result; of returning results that can be interpreted and quantified with regard to their associated confidence and support; and to be extensible, that is, further processes can be included as needed (Eddy 2004, Ji & Liu 2010).

Hierarchical Dirichlet models

In assignment studies using model-based clustering approaches, one line of development concerns the incorporation and simultaneous inference of all population structure parameters, including the number of clusters. In the Bayesian framework, implementations of such hierarchical approaches currently exist using parametric Dirichlet models of finite mixtures (Dawson & Belkhir 2001, Corander et al. 2008a, Guillot et al. 2012), and Dirichlet process models modelling infinite mixtures (Pella & Masuda 2006, Onogi et al. 2011, Reich & Bondell 2011, Lawson et al. 2012). The latter approach was further expanded to hierarchical Dirichlet process models for populations with admixed individuals (Teh et al. 2006, Huelsenbeck & Andolfatto 2007, Elliott et al. 2018). Dirichlet processes are Bayesian nonparametric models that do not require a fixed model, but which allow model complexity (here: associated with the number of clusters) to adapt to the data and to grow in complexity as more data are observed (Gershman & Blei 2012). Simultaneously, they avoid label switching problems, are flexible and easy to implement for inference, and make it possible to bypass potentially difficult and cumbersome parts of the model selection process (Elliott 2016, Elliott et al. 2018). Despite these advantages, they have been shown to overestimate the number of clusters, producing small extra clusters, and their model parameters are difficult to translate into biologically meaningful processes. Miller and Harrison (2018) compared nonparametric infinite mixture models and parametric finite mixture models. They found that both approaches share important properties, and suggested that their advantages could be combined in the future.

3.2 Validated empirical reference datasets as foundations for reliable conclusions

Reliable inference from complex evolutionary systems requires that background information is accumulating in the form of independent results and substantive preexisting knowledge (congruence). In addition, it is necessary that the validity and performance of inference approaches has been substantiated in analyses of empirical data (model checking). In human

population genomics and genetics, worldwide reference datasets for genetic diversity have been established over time with increasing genome representation (HGDP (Cann et al. 2002, Li et al. 2008); HapMap (International HapMap Consortium et al. 2010); 1000 Genomes (Sudmant et al. 2015, The 1000 Genomes Project Consortium et al. 2015); SGDP (Mallick et al. 2016)). These population diversity panels are repeatedly reanalyzed to evaluate existing results and to characterize newly developed models and inference approaches. In turn, each of these analyses and methods contributes more and more fine-grained insight into features of these datasets.

In forensic conservation genetics, first steps in this direction exist. For example, for the marine fish Atlantic herring (*Clupea harengus* L.) an extensive sample and marker set was collected and assignment tested at a larger scale (Nielsen et al. 2012a). Parallel, the transcriptome-derived SNP markers were investigated for their selective signatures, the geographic and environmental distribution of alleles, and thus their informativeness for population assignment (Limborg et al. 2012). Subsequently the marker set was employed for assignment at a much finer scale with an extended sample set focusing on questions specific to one region (Bekkevold et al. 2015).

In inventories and monitoring of natural populations, especially (but not only) for taxa requiring protection and forensic tools, versatile, extendable and progressively more well-known reference datasets are essential (see fig. 4.1 in section 4.2.1). They form the foundation for an overall open-ended iterative process towards the optimization and refinement of both, the sample and marker sets, as well as, of inference methods and forensic tools. An extensive context of prior results and insight will build up over time for a once initiated reference set. Such preexisting knowledge guides further (analytical) development and, most importantly, informs conclusions in upcoming casework (cp. Gelman et al. 2014 p. 139; see also section 5.3.4 discussing strong inference and severe testing). Such continual advancement is possible, if the focus is on a single, growing but integrated GWS-represented reference set, which is analyzed and characterized by many inference methods and research groups. In this way, the development of in-depth explored reference datasets that are well understood from a wide range of viewpoints is an indispensable constituent of validation and at the core of reliable practical application.

This underlines the vital importance of reference datasets that are not only representative of the population structure across the distribution range, but also constitute a random, unbiased sample of a taxon's genomic organization (cp. fig. 4.1). A random, sequence-based representation of the genome supports a wide range of inference approaches. In this way, it enables direct comparisons and reciprocal evaluations of methods. GWS-strategies are a logical extension of widely applied approaches sequencing selected gene fragments. In forensic forest genetics, for example, such a gene-based sequencing strategy was carried out for tree species assignment across Malaysia (Tnah et al. 2009, Ng et al. 2017). Sequencing using GWS-strategies expands sequence-locus selection to the whole-genome level, endeavoring improved assignment power.

Reference sets characterized by genome-wide sequencing allow subsequent extensions. First, additional samples can be added, which are covering more of the distribution range, as well as, finer regional to local scales. Second, at the same time, genomic coverage with regard to chromosome regions and sequencing depths can be increased without having to discard existing sequences. DNA-sequence data retain their value and informativeness even with progress leading to new technological inventions and analytical advances. Sequence databases that can vouch for the retained value of sequence data are NCBI Genebank (Benson et al. 2000), with records dating back to the by now ancient history of manual sequencing, and the Barcoding of Life Data Systems (Ratnasingham & Hebert 2007) that have the specific applied aim of assignment at and above the species-level. Subsequently set into the context of reference genomes, the quality and genomic context of preexisting sequence data can be further assessed. For example, it is possible to address questions of homology and gene duplication, sequence functionality, the characteristics of the surrounding chromosome region or the association with genetic hotspots. Such information might even increase the value of archived sequences.

The feasibility to cumulatively build a GWS-based reference dataset was shown in the genus *Quercus* (oaks) by merging two independently created RAD-datasets that were constructed using different protocols. Independent and combined analyses of the two datasets were performed. These analyses validated the congruence of the evidence in the two datasets and reinforced the phylogenetic results (Hipp et al. 2014). Complementing this approach, it has been explored in as far different workflows of RAD-dataset handling and analysis have an impact on biological and evolutionary results (Fitz-Gibbon et al. 2017). More generally, remaining challenges to the use of GWS-data in non-model organisms are continuously addressed (e. g. McCormack et al. 2013, Harvey et al. 2016).

The application of genome-wide sequencing to all individuals of a range-wide representative sample set initially is expensive and resource-intense. This is particularly the case once the sequence data is in, due to the sheer amount and complex structuring of the data. Their handling, exploration and analysis is comparably bioinformatics- and thus manpower-intense, at least until workflows are established and have been validated. As already pointed out, many GWS-strategies return more or less randomly sampled genomic characters. Some of those will not be informative immediately. Especially at the beginning, this seems inefficient, since the marker set is not optimized for the intended assignment application. However, only in this way it is possible over time to adapt the reference set to all kinds of (unexpected) case scenarios and to accumulate experience and insight from a wide range of evolutionary and statistical viewpoints. What is more, at the same time that the full GWS-reference datasets enable versatility, they also allow flexible *in silico* subsampling of markers and thus the construction of marker sets that are optimized towards the resolution of case-specific assignment objectives. Such optimized marker sets will allow for cost-efficient genotyping of case samples and can be geared towards the genotyping of case samples with degraded DNA. Taking all of this together, GWS-based reference datasets provide a maximum of flexibility across evolutionary scales and questions. Hence, in the end, such a sequencing strategy is the most effective and efficient approach towards validation,

for building assignment resolution and reliability, and thus for arriving at user-friendly and adjustable forensic tools in practical everyday application.

This assessment of GWS-based reference datasets strongly argues for augmenting existing range- or country-wide monitoring initiatives in fishery and forestry by a genetic module. The starting point is not necessarily a massive sampling of individuals; rather the focus should be on establishing an appropriate GWS-strategy for each taxon on an initial sample set considered representative for the taxon's global genetic diversity. This dataset should provide a starting point and, as outlined above, the capability to be flexibly extended across all evolutionary and geographic scales and for increasing genomic coverages as practical applications arise, new sequencing options become available and further samples become accessible.

4 Efficiency and sufficiency

4.1 Efficiency

The amount of potential and thus pending conservation- or wildlife-genetic forensic applications and test tasks is huge, and individual test results often need to be achieved on a schedule due to logistic and economic reasons. Hence, forensic conservation genetics has a need for efficient and preferably fast statistical methods. Despite the reality surrounding their application, inference approaches in evolutionary biology are often based on iterative, multi-step procedures and more or less parameter-rich models, which reflect the complexity of evolutionary and biological reality. These complex models require information-rich empirical datasets for compound parameter estimation, event prediction and hypothesis testing. In this overall situation, the fundamental question is, which kind of inference approach – and with it, simplification and approximation – will both be fast while still being sufficient for arriving at a reliable, accurate and precise result.

The efficiency of a consistent inference approach is a function of its variance, that is, its precision. In the large sample limit the variance of an efficient inference approach, its random error, should become as small as possible (tend towards zero). In addition, an inference method's variance preferably should be small already for limited sample sizes and, furthermore, decrease early-on and rapidly with increasing sample sizes. Considered from the point of view of information, efficiency represents the fraction of the available relevant information in the data that is actually utilized by the inference statistic and approach (Fisher 1925).

In real-life application, not all approaches that are theoretically efficient with increasing sample sizes are equally efficient for finite, realistically sized samples and with limited computing

resources. In small samples, efficiency ("speed") is statistically further defined (Fisher 1925) as the inverse of the informativeness (intrinsic accuracy) of the **population parameter** chosen for the investigation, divided by the variance of the **inference approach** (Fisher Information). The denominator representing the variance of the actually employed inference approach is a composite of, first, the specifics of the **estimator** used to inform on the population parameter, second, the particulars of the utilized **inference method** (e. g., its handling of complexity) and, third, the characteristics of the **empirical data** used. This general formulation of efficiency sets the context for the following sections, which outline and discuss the formula's factors for the evolutionary setting and practical forensic application.

Large and continuous distribution ranges challenge efficiency

The widely and continuously distributed, often highly diverse taxa in fishery and forestry pose several challenges for arriving at adequate models and thus efficient inference. Isolation-by-distance processes (Wright 1943, Malécot 1973, Ishida 2009) dominate their populations. Hence, their lineages are shaped by inherently spatial and continuous evolutionary processes, which lead to overall low levels of genetic differentiation over large areas in their distribution ranges. In addition, they likely are in non-equilibrium not only in time, for example due to population demography and evolutionary history, but also across space, due to different migration and selection dynamics within their extensive distribution ranges, which are spanning different adaptive environments and histories. At this point, fundamental challenges persist to encapsulate such complexity with mathematical models to arrive at inference or generative models that are broadly applicable and can be versatilely applied for statistical inference (Novembre & Peter 2016). However, rapid progress is being made with regard to many of the system's contributing characteristics and to implementation obstacles, for a recent review see Novembre & Peter (2016) and references therein.

Nevertheless, for the near future, the development of suitable approximations for processes and patterns at several levels and scales will remain in the focus of assignment inference in this setting. Currently, the most widely applied approximation is the island model (Wright 1931) of classical Haldane-Fisher-Wright population-genetics. It forms the concept underlying the traditional supervised assignment approaches as implemented in GeneClass2 (Baudouin et al. 2004, Piry et al. 2004), the non-supervised global clustering approaches based on the Structure model (Pritchard et al. 2000), as well as, widely applied spatially explicit model-based clustering and assignment approaches (e. g. Wasser et al. 2004, Wasser et al. 2007, Corander et al. 2008b, Caye et al. 2016), reviewed in Guillot et al. (2009) and Francois & Waits (2016)). This approximation works well for time and spatial scales, at which population differentiation is strong compared to mutation-drift-migration(-selection) disequilibrium processes. However, for small F_{st}-values (approx. $F_{st} < 0.05$), assignment reliability seems to plateau under this approximation. Even with large sample sizes of individuals, genomic markers and heuristic resampling iterations, too large overlaps between the distributions of test statistics for different populations or lineages remain, resulting in heterogeneous (high error rates) or indecisive, "non-significant" assignment

and exclusion results. This suggests that, while allowing for comparably fast algorithms, the island model in those cases is not efficient enough. For sufficient resolution, the evolutionary model needs to incorporate more realistic generating processes and use more of the information recorded in the genome.

4.2 Sufficiency

In evolutionary systems, even the ever-growing reference datasets of today represent a small sample setting compared to the length of evolutionary history and the complexities of reality. In this situation, a comparably more efficient inference method is distinguished by returning estimates with a low variance already when applied to smaller datasets. For this, it analyzes population parameters, estimators (statistics) and data that are informative for the objective at hand. Here, an efficient method is called to be also sufficient if it represents all the information present in the data pertaining to the investigated population parameter (Fisher 1925). However, in investigations of a highly complex reality, sufficiency is a theoretical concept that in practice cannot be achieved. Here, a sufficient statistic or approach (in non-statistical settings generally the term “powerful” is used) assesses, detects and analyzes as much as possible of the (noisy and conflicting) information provided by the data and represents this information comprehensively in the results. Accordingly, the more powerful inference approaches are, the more they need to be able to distinguish correctly between multiple processes, between signal and noise, and to allow detection and quantification of errors and systematic bias.

In the context of sample prediction for hypothesis testing in forensic conservation genetics, efficiency and sufficiency denote in consequence that given a set amount of data, a powerful method is able to achieve comparatively high and simultaneously realistic support and accuracy values, with associated correct but preferably low estimates of false-positive and false-negative errors. On the other hand, given a preset required significance and confidence level, likelihood ratio threshold or posterior probability, powerful approaches need less data to reach those thresholds, while converging on the truth. Thus, powerful methods are able to resolve and distinguish well-supported between multiple models, as well as, parameter and predictive distributions for contrasting hypotheses, and they are able to do this already with less data.

Taking all of this together, inference approaches of higher sufficiency offer more insight into different facets of the information present in the data and their associated quality characteristics. Thereby, sufficient inference approaches provide the foundation for the versatility (Hillis & Huelsenbeck 1994) of an inferred estimate with regard to its further usefulness and range of application.

4.1 Sufficiency in reconstructions of evolutionary relationships

Different levels of sufficiency can be illustrated by considering different reconstruction approaches for evolutionary relationships. For example, compare UPGMA hierarchical-clustering dendrograms with character-based tree reconstructions using maximum parsimony, maximum likelihood or Bayesian approaches; or consider K-means partition clustering with clustering based on the Structure model (Pritchard et al. 2000).

First, more sufficient approaches return results with overall scores, thus informing about the probability density surface of the results. Thereby, they allow quantitative comparisons and, accordingly, ranking of better to worse fitting or applicable estimates. They provide information on equally good or next best results and their details. The UPGMA hierarchical clustering algorithm returns only a single solution in the form of a dendrogram. Neither is it possible to evaluate how "good", that is, likely or probable, this solution is, if there are other equally good solutions or how close the next best solutions are, nor are many characteristics of the branching structure available. In contrast, character-based reconstruction approaches (e. g., maximum parsimony, maximum likelihood, Bayesian) return phylograms (or at the population-level coalescents representing genealogies) that provide information not only on the topology, but also on branch lengths and thus substitution rate heterogeneity, branch support, ancestral states, and the distribution, quality and quantity of similarly likely or probable reconstructions (tree space). As part of their inference procedure, they furthermore provide insight into the details of the evolutionary substitution process by optimizing substitution models. These substitution models have been developed and are applicable for a wide range of character types. Similarly, model-based clustering approaches provide more information and detail about their inferred reconstructions of population structure, and of individual cluster membership probabilities or admixture proportions than, for example, K-means partitioning or distance-based hierarchical clustering (however, see "Multivariate approaches" below in section 4.2).

Hence, second, comparably more sufficient approaches for evolutionary relationships and population structure return estimates with more varied and detailed characteristics. They return results not only in the form of summary statistics, but also in highly visual form, allowing access for human pattern recognition capabilities (Nielsen 2016). They thus support intuitive assessments of the details of the results. In addition, visualization facilitates the observation of correlations between the result and different aspects of the data, as well as, the detection and evaluation of parameter correlations.

Finally, while often the truth of empirical data, and thus the power of an approach, is associated with well-resolved relationships and high support, this might not always be the case. On the contrary, a powerful tree-building or clustering approach will return only a low level of resolution for evolutionary lineages above the species-level that retain patterns of incomplete lineage sorting or that are shaped by lateral gene transfer or hybridization. Within species, at the population-level, isolation-by-distance processes and recombination, especially in association

with (long-distance) migration and/or selection, should result in low resolution between lineages, when tree-like branching patterns are assumed in the reconstructions.

Coalescent approaches

One avenue to arrive at more powerful (sufficient) inference results for taxa with large, continuous distribution ranges is to incorporate at more levels and for more processes sample- and thus individual-based coalescence models (Kingman 1982, Tajima 1983, Hudson 1990). However, until recently, inference employing the structured coalescent (Notohara 1990), underlying model-based phylogeographic inference (e. g. Nielsen & Beaumont 2009, Bloomquist et al. 2010), required *a priori* (supervised) knowledge of and decisions on population entities or lineages. Here the spatially continuous coalescent is opening promising new options (Guindon et al. 2016, Joseph et al. 2016). Equally important is the fact that the coalescent concept provides a basis to represent non-equilibrium situations. It thus extends in applicability from the very recent past (Wilton et al. 2017) further into phylogenetic time scales (Yang & Rannala 2014). Thereby it provides a natural connection to phylogenetic concepts and approaches (Cutter 2013). Originally developed in a phylogenetic background, network reconstruction approaches (see Novembre & Peter 2016) can provide reconstructions that approximate the, so far, for most applications computationally intractable ancestral recombination graph, an extension of the haplotype-based coalescent to recombining lineages. Moving further towards longer time scales, the multispecies coalescent provides an approach to reconstruct evolutionary relationships for comparatively young species or at long-term unclear species boundaries by taking into account incomplete lineage sorting and conflicting gene (locus) trees (Heled & Drummond 2010, Rannala 2015). Finally, the connection of the coalescent to phylogenetics provides an opportunity to employ ancestral character state reconstructions (e. g. of geographic origin) on evolutionary trees and genealogies (Stone et al. 2011, Landis et al. 2013, De Maio et al. 2015).

Multivariate approaches

An alternative avenue towards sufficiency for very large genomic datasets is provided by “model-free” pattern-recognition algorithms. Patterson et al. (2006), summarized in Novembre & Peter (2016), proposed that above a general threshold in the relationship between F_{st} and dataset size (matrix size > $1 / F_{st}^2$) even weak population differentiation ($F_{st} < 0.01$) can be detected independently of the inference approach taken. Correspondingly, multivariate methods have become increasingly popular for analyses of the large data matrices of GWS data. Already, principal components have been interpreted with regard to relatedness, coalescence, hybridization and spatial coordinates, accompanied by discussions of assumptions and limitations (e. g. Novembre & Stephens 2008, Jombart et al. 2009, McVean 2009, Gompert & Buerkle 2016, Zheng & Weir 2016, Lever et al. 2017).

However, as Patterson et al. (2006) also pointed out, the detection of population structure is not the same as the reliable assignment of samples. Accordingly, pattern-recognition algorithms present an efficient and fast alternative to cumbersome and costly parameter-rich model-based inference for population differentiation. However, in the case of an assignment objective, pattern-recognition algorithms might not sufficiently extract and represent the information present in the data. This is due to the fact that methods based on distance measures use only a progressively diminishing fraction of the information on evolutionary or genealogical relationships present in sequence data as the number of samples increases (Steel et al. 1988, Penny et al. 1992). Nevertheless, it has been suggested that character-based sequence information is sufficiently approximated by the pairwise distance estimates of genetic relationship matrices (Lawson et al. 2012, Weir & Zheng 2015, Novembre & Peter 2016, Sun et al. 2016, Wang 2016). At least, when these distance estimates represent genome-wide averages based on a very large number of genomic sites and regions, assuming that those sites and regions are removed, which are affected by deviating processes.

4.2.1 Stochasticity of evolutionary population parameters

All evolutionary processes are characterized by their inherent stochasticity, that is, the theoretical randomness with regard to specific individuals, genome regions and events in the finite collections of individuals in reality. The underlying causes are mutation, recombination and segregation, migration, selection and, from a genetic point of view, often rapid and unpredictable extrinsic historical events (as for example, climatic changes and fluctuations, meteorite impacts, the arrival of *Homo sapiens* L. on a continent). Accordingly, the pedigrees and genealogies connecting individuals or even only intragenomic partitions are highly variable between individuals and genome regions, thus also between sets of samples and markers. From this follows, that coancestry and coalescent parameters, and hence population-genetic summary parameters, have large variances (Hey & Machado 2003, Thompson 2013, Speed & Balding 2015).

These large theoretical variances can only be reduced by the decisive events and the strong to approaching deterministic processes that are part of and characterize biological reality. Biologically realistic datasets thus often show pervasive signals that give rise to the sufficiency of analytical equations and heuristic search algorithms, enabling the inference of accurate and precise estimates in the face of theoretical stochasticity (Felsenstein 2004 p. 61). For the purposes of forensic conservation genetics, the goal, thus, is to find and focus on processes that leave clear and preferably strong signals in the genomic record for the question at hand. In the next step, it is often necessary to identify the specific characteristics and types of genomic data partitions that most appropriately transport the signal of these selected processes at the evolutionary scale investigated (Lawson et al. 2012, Thompson 2013). Building on the identified signal and its associated reduced variance, the subsequent aim is to keep the amount of any added variances due to sampling, modelling and inference as small as possible.

Linkage patterns and rare variants carry strong signals

Especially for investigations concerning the recent past, sufficient inference approaches require a focus on specific, more informative characteristics of genomic data. One such feature is, to take into account the footprints that past recombination and segregation events left in the genome. Correlated character states at linked and unlinked sites across the genome provide insight into coancestry along a continuum of time depths (Edwards 2003, Falush et al. 2003). For example, ancestry segments along chromosomes represent inherited blocks of distinct ancestries. They are reconstructed or "painted" using local, that is, locus- (sequence site) focused, inference approaches (e. g. Tang et al. 2006, Lawson et al. 2012). Distinct ancestry blocks that extend over the background decay of physical linkage disequilibrium in a genomic region are expected to be the result of more recent admixture events. Analyses of the shape of such long-range admixture linkage disequilibrium distributions provide insight into the timeframe of the admixture (Moorjani et al. 2011, Loh et al. 2013). Thus, the nucleotide compositions and reconstructed lengths of ancestry blocks form the basis for tests of admixture, provide information on ancestral population sources, are informative for the clustering of samples, and moreover, allow the dating of admixture events (e. g. Hellenthal et al. 2014, Busby et al. 2015). An integrated suite of these procedures (fig. 4.1) is implemented in the program Globetrotter (Hellenthal et al. 2014). Focusing directly on population assignment, HaploPOP specifically uses the additional information provided by linkage disequilibrium for increased assignment power and resolution (Duforet-Frebourg et al. 2015). For inference at even more recent time scales, that is, of familial relationships, shared chromosomal segments of identity-by-descent allow the inference of degrees of relatedness between individuals connected by extended pedigrees (reviewed in Browning & Browning 2012, Thompson 2013).

Another highly informative genomic feature that emerged in population genomics are rare alleles. Rare variants are presumed to be recent mutations, whose distributions are still geographically localized (Mathieson & McVean 2012), and for which the potential for persistence has been suggested in isolation-by-distance models (Schneider et al. 2016). They thus could provide important markers for local assignment and reconstruction of close kinship (Novembre & Slatkin 2009, Gravel et al. 2011, Mathieson & McVean 2012, Thompson 2013, Weir & Zheng 2015). Their identification requires high quality GWS reference datasets that allow the validation of rare and low frequency alleles, and the calculation of associated error rates. At the same time, such genomic datasets provide the potential to call rare variants through imputation, using linkage information for the reconstruction of haploblocks (Gravel et al. 2011, Mathieson & McVean 2012).

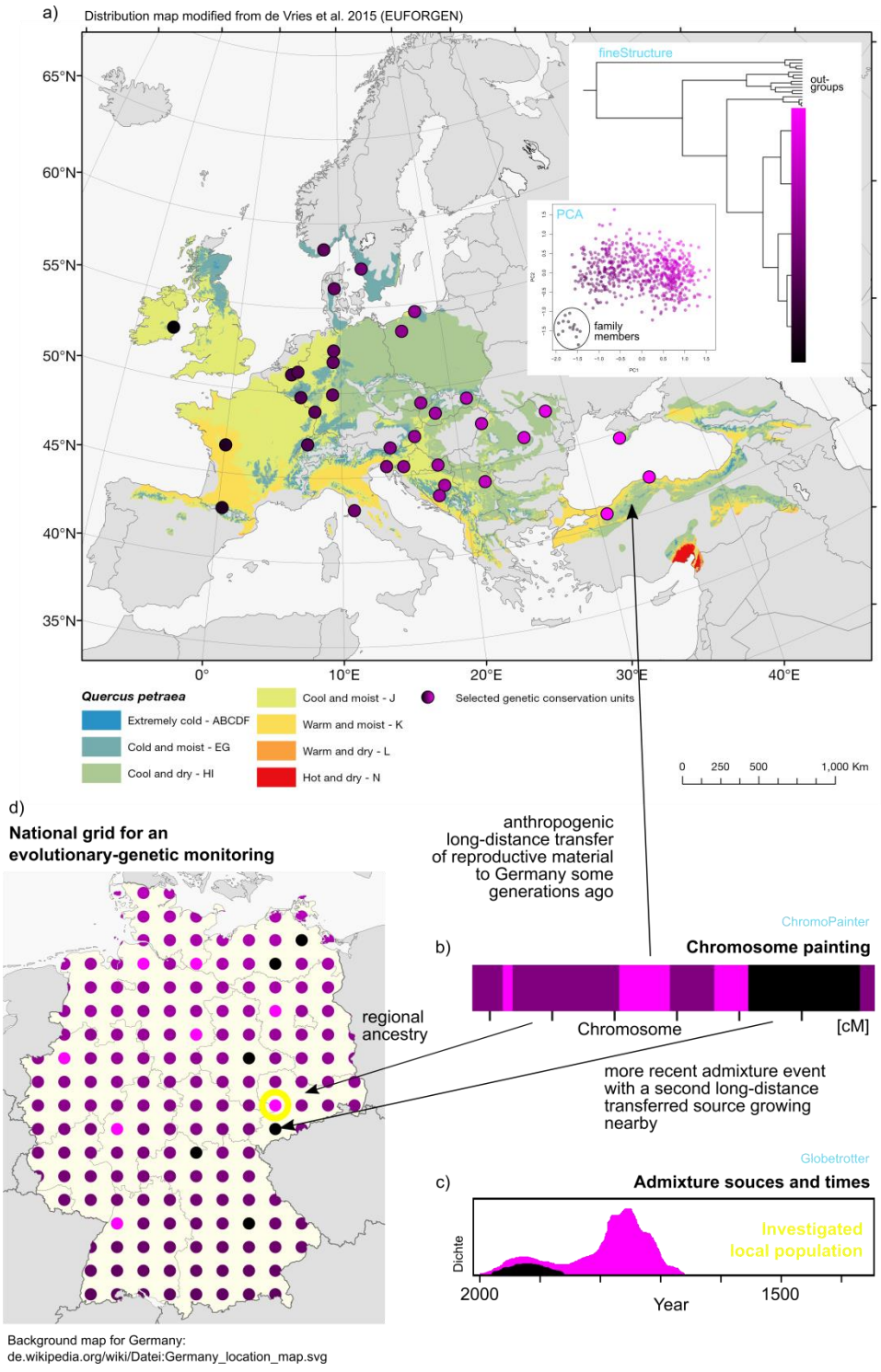


fig. 4.1: Proposed reference dataset design and inference integrating analyses of a strong signal for reliable applications in conservation genetics. Representative sampling (following, e. g., de Vries et al. 2015 for European tree species) of distribution-wide genetic diversity (a) and genome-wide sequence and ancestry block information (b, c) jointly provide insight into the genetic composition and evolutionary, as well as, anthropogenic history of local tree populations and forest stands (d). light blue: examples of inference programs (cp. Novembre & Peter 2016). PCA: principal component analysis. cM: centimorgan.

4.2.2 Computational approaches under complexity

4.2.2.1 Analytical solutions

Exact mathematical estimators exist for many population-genetic parameters. One of their advantages, apart from their calculation being fast, is that they provide analytical solutions for the variances or at least the total variance of the estimator (Hey & Machado 2003). These analytical population- or sample (coalescent)- based estimators, while often being robust to violations of assumptions, nevertheless require a priori knowledge of the applicable evolutionary model and its parameter settings, a requirement that for most non-model organisms is not given. In analytical approaches, simplifications and assumptions have to be made, since only one or a couple of parameters can be analytically evaluated simultaneously. Apart from the necessity to assess potential bias, this also can lead to issues of identifiability, since in many cases distinct processes can cause the same result, even with unlimited data (compare, e. g., the signatures of demography and selection at a single locus). Furthermore, analytical summary statistics are point estimates that generally do not provide sufficient information about the shape of the probability density surface. However, sets of summary statistics that complement each other, as for example, summary statistics representing the characteristics of the site frequency spectrum or a set of hierarchical F-statistics, can optimize the representation of relevant information recorded in the data and provide sufficient inference approaches. In this way summary statistics can allow estimates of population parameters that are of similar quality to those obtained with more sufficient approaches (Hey & Machado 2003).

4.2.2.2 Heuristic algorithms

Analytical calculations and exhaustive methods of enumeration are certain to return the global optimum of an estimate and, thus, are robust with regard to the specifics of the inference strategy. Generally, however, parameter and result spaces are too large and complex to describe analytically or to sample exhaustively. It is thus necessary to develop and employ heuristic approaches implementing and preferably combining optimization (accuracy) and resampling (precision) aspects that are more efficient than exhaustive enumeration. Nevertheless, for most tasks in evolutionary genetics such heuristics still are NP-complete or -hard (see below). Moreover, they give no guarantee of being the most efficient or to even find the global optimum.

Heuristic approaches are algorithms that are employed, when analytical or exhaustive approaches are too slow, or when they are not able to provide any exact or sufficiently informative solution at all. In both cases, heuristic algorithms, while being fallible (their results cannot be proven) and without guaranteeing success, provide plausible and useful strategies (Romanycia & Pelletier 1985, Vesterstrøm 2005, Ibsen-Jensen et al. 2015). They are designed to improve performance, find at least an approximate solution, be able to verify a solution or list

several distinct solutions in a given time and space. Essentially, they provide guidance and insight for searches and decision making (Romanycia & Pelletier 1985).

4.2.2.3 NP-hard efficiency

The efficiency of an algorithm is the amount of resources - time (CPUs) and/or space (memory and storage) - it requires. With regard to dataset sizes, efficient algorithms respond to increases in input size (in samples and/or characters) with comparably low growth rates of their computation requirements. Hence, a search or optimization routine should arrive at a result preferably in linear (e. g. n) or at most polynomial (e. g. n^4) time. Linear time indicates that with increasing input size (n) the computation requirements and complexities grow also only linearly (Penny et al. 1992). However, denoting an inference task as NP-complete or -hard (NP: non-deterministic polynomial), states that all algorithms so far known for the task and similar problems, to date in actual inference applications, have exponential (e. g. 2^n) or worse (e. g. n!) time requirements.

Assignment problems and associated inference tasks, as for example unsupervised model-based clustering or the character-based reconstruction of genealogies and gene trees, are NP-hard (Graham & Foulds , Felsenstein 2004 pp. 59 ff.). Despite this, constantly new or modified heuristic approaches are developed that take different routes to reducing at least the coefficients of resource requirements. Historically, at the within-species, population level, one of the first heuristic approaches to searching the space of possible genealogies for the global optimum was based on importance sampling methodology (Griffiths & Tavaré 1994). Widely employed global heuristic approaches for searching the spaces of genealogies or clustering partitions use Markov Chain Monte Carlo (MCMC) with Metropolis-Hasting (Kuhner et al. 1995) or reversible jump (Green 1995) strategies. The application of Dirichlet process models (see section 3.3.2) in evolutionary genetic inference is a more recent development. Heuristics for assignment-associated tasks can also be cast as explicit optimization problems (local optimization) that then are resolved through approximate Bayesian computation (ABC; Csilléry et al. 2010), Expectation-Maximization (EM; Tang et al. 2005), matrix optimization (Alexander et al. 2009, Caye et al. 2016) and Bayesian variational (Raj et al. 2014) strategies (for a taxonomy of heuristic methods see, for example, Ashyraliyev et al. (2009)).

The parameter-rich models that form the basis for heuristic algorithms transform assumptions into explicit parameters. They thus are capable to more sufficiently access, represent, process and return the information content present in the data. However, in this way they also include more parameters that necessitate exploration, optimization and evaluation. These models and the heuristic algorithms that accompany them can be sensitive to initial parameter and algorithm settings, as well as, model violations in unexpected ways. They require fine-tuning and extensive dataset-specific exploration of their strengths and limits. Hereby, each individual software

implementation of a heuristic algorithm can strongly influence performance and results (e. g. Ashyraliyev et al. 2009).

4.2.2.4 Balancing between model realism and solution optimality

"… it's difficult to obtain a precise solution to a problem because we either have to approximate a model or approximate the solution" (Michalewicz & Fogel 2004 p. 19). Consequently, the optimal balance between sufficiency (power) and speed, including ease of application, of an inference approach lies in the trade-off between model realism (i. e. model complexity) and solution optimality (Vesterstrøm 2005 pp. 8-9). In practical applications, this situation requires that a decision has to be made and over time continuously optimized. On the one hand, there is the option of solving a simplified and thus approximate model exactly, which assesses and represents less information in the data. This guarantees finding the optimum, though potentially associated with larger variances and biases. On the other hand, the option is to employ a complex, parameter-rich and thus precise model for which one searches for plausible, but approximate solutions. Michalewicz and Fogel state that this second approach "is often superior" (Michalewicz & Fogel 2004 p. 19).

One way to combine the strengths of both options is to use the results of analytically derived and fast summary statistics as starting points for more sufficient but approximate algorithms (Hey & Machado 2003). Furthermore, integrated analysis environments can contribute to the effectiveness of multipart workflows. Finally, to take full advantage of powerful approaches, advanced summaries and visualizations allow one to access meaningful facets of large datasets and information-rich, complex results (Nielsen 2016).

4.2 Integration of heuristics into practical applications for non-model organisms

It has been emphasized that, for each new taxon and type of reference dataset, inference based on heuristic algorithms requires experience and several years of preparation before the results are reliable (Hillis 1995, Bloomquist et al. 2010). Only after this initial development phase, forensic validation can commence from sensible null models of natural evolution and parameter settings. One example for such an extended development requirement can be found in the ongoing multi-workgroup process towards the genetic determination of species delimitations and admixture dynamics in European white oaks using unsupervised clustering approaches (e. g. Lepais et al. 2009, Neophytou 2014, Neophytou et al. 2015). Hence, highly informative, powerful inference approaches generally require extensive preparation and optimization for algorithms, which, while requiring run time to arrive at informative and robust results, yet might not provide the final answer. While being no miraculous solve-all's, powerful heuristic approaches are inevitable, since they provide the necessary detailed information for further iterations of

development cycles towards improved models, parameter settings and inference approaches. They support, in this way, more sufficient and versatile inference development and validation.

So far, experiences gained in forensic conservation and human genetics suggest that the application of heuristic algorithms is needed for arriving at tools supporting the successful protection of many endangered taxa. The development, fine-tuning and optimization of inference approaches that employ heuristics for a threatened taxon, accepting the time and further resource requirements, makes only sense if well-established, that is, cleaned, highly informative and well-characterized cumulative GWS-reference datasets with preexisting complementary results and experiences, are available and will be used for years to come. Carstens et al. (Carstens et al. 2009, Carstens et al. 2013, Hickerson 2014, Pelletier & Carstens 2014) describe an integrated framework, in which different advanced heuristic approaches inform on and complement each other (see section 5.2 “Model validation”).

4.2.3 Accessing and representing the information content present in the dataset

The first steps of statistical inference are data preprocessing, cleaning and exploration (Anderson et al. 2010, Laurie et al. 2010). The central aims of these are to remove data errors and thus achieving the clearest and most unbiased representation of the real genetic diversity to be able to detect and extract the targeted signal. Traditionally, data decisions at this stage are made based on experience and informed thresholds. These decisions code characters (e. g. DNA-sequence sites) as one of the allelic states or as missing data. They accept character homology or remove ambiguously aligned data partitions and/or samples with too much missing data completely from subsequent analyses. The signal-to-noise ratio (and remaining error) in such “clean” datasets is explored and then quantified through resampling strategies. The potential impact of the initial data cleaning decisions on the results of subsequent in-depth inference methods is discussed in the final evaluation of the results. However, focusing on uncertainty associated with GWS-sequencing, it has been shown that such informed, but still arbitrary decisions introduce ascertainment bias in many forms into the data and thus the results (Crawford & Lazzaro 2012, Nielsen et al. 2012b, Nevado et al. 2014).

An emerging alternative to this procedure, applicable to sequence alignments of selected gene regions and to GWS sequence data, is to incorporate uncertainty in alignment, and allelic or genotypic state directly into the calculations of inference routines for parameter estimates and their associated likelihoods or posterior probabilities (Fumagalli et al. 2014, Korneliussen et al. 2014, O'Rawe et al. 2015, Vieira et al. 2016). Skotte et al. (2013) implemented this approach for admixture clustering. Data uncertainty is propagated via genotype likelihoods based on

sequencing, coverage and alignment quality values. In this way, conflicting processes and error probabilities can become integral parts of the results, and are appropriately represented in estimates and their support. Such an approach is especially of interest in population-level investigations of non-model organisms, for which there is no prior information, that show high levels of diversity and for which many individuals are sequenced at low coverage with GWS technologies (Crawford & Lazzaro 2012, Nevado et al. 2014, Wang et al. 2016). The importance of accounting for uncertainty in subsequent inference steps has also been discussed with regard to personal medicine that, like forensics, requires very high accuracy and reliability (O'Rawe et al. 2015).

5 Model specification and hypothesis falsification

Our perception and consciousness of the world, that is, the applicability of ideas to reality, is built on the specification (definition) and falsification of (falsifiable) hypotheses. Philosophies of science provide the conceptual frameworks and the axioms of logics the standardized methodology for this process. Together they form the foundation for insight into the world, scientific knowledge and all their derived applications. In concrete application, to answer substantial questions (“Are all ravens black?”), and generally even with apparently unequivocal evidence (“No, I saw a white one.”), hypotheses have to be transformed into mathematical models (“I checked that it is the same species …”), since these can be analyzed, statistically evaluated and compared (“… using phylogenetic and population genetic concepts and inference methods”). Hence, models form the starting point and the foundation blocks of statistical inference. Coming full circle, inferences about models provide the information for drawing conclusions about their corresponding hypotheses and thus to gaining insight into reality.

Once a model has been defined, it is applied to available data, under the assumption that the dataset at hand is representative of reality. Through application, it becomes possible to evaluate the model’s fit to the data based on one to several measures. These measures of fit provide the rationales and facts for rejecting a model and, thus, its associated hypothesis. Lacking fit, a model does not sufficiently summarize and thus explain the information in the data. In extension, comparisons of measures of fit form the basis for choices between several possible models or values of model parameters. Examples of measures of fit that provide a basis for evaluation and comparison are model likelihood and posterior probability, the sum of residuals, discrepancy measures to a predictive distribution, as well as, overall predictive accuracy and error rates.

5.1 Model specification, model validation and hypothesis falsification

Model specification: asks for the model that best represents a hypothesis given the available data

Heuristic algorithms: continually change and evaluate parameter settings within a given model structure during searches and optimizations

Model selection: standardized procedures for choosing among models of different (nested) structures that all are in agreement with the statement of a specific hypothesis. Model selection compares independently specified models using information criteria to determine the relative best model or models in an *a priori* designated set of models corresponding to the hypothesis.

Model Validation: determines the sufficiency and robustness, that is, the usefulness and applicability of the best inferred model and inference approach.

Model assessment: evaluates if the inferred best, fully specified model summarizes the data sufficiently well to be of use to answer the objective at hand. At the same time one gains insight into the types and patterns of information that contradict the investigated model, which might be the causes for and explain the rejection of the associated hypothesis.

Hypothesis falsification: test if the data (reality) significantly contradict a hypothesis, that is, an inferred model, which was shown to have the best fit and to be (generally) useful.

A single hypothesis: evaluate the overall extent of information in the data against a proposed hypothesis (*P*-value, significance level α, specificity)

Two or more alternative hypotheses: test if a hypothesis, which is not rejected by the data, explains the data significantly better than another contrasting hypothesis or several alternative hypotheses (power, sensitivity)

An informed choice among models is a fundamental objective and competence, as well as, a necessary procedure throughout the inference process for a scientific project or applied statistical tool. It is central to fitting a model to the data, that is, inferring and optimizing model specification. Heuristic algorithms continually are calculating and evaluating model fit, while searching for or optimizing the best overall solution for a model given the data (cp. figs. 3.2 and

3.4). Simultaneously, on the way to finding the best overall solution, heuristic algorithms are inferring and specifying the model's individual parameter values. Moreover, heuristically (or analytically) inferred measures of overall best model fit, in subsequent analyses, provide the rationales for choices among (nested) models with different structural definitions and their, for each model independently, inferred parameter specifications. Here, model selection in itself can become a statistical inference approach with standardized procedures to systematically look for, specify a set of applicable, possible models and select from it the model that best represents a specific hypothesis given the available data.

Once a model has been specified that best represents the hypothesis of interest given the data at hand, model assessment asks and evaluates, how likely this best inferred model is overall, or at least, if this best inferred model fits reality sufficiently well to be useful (model validation). If a model is shown to lack applicability and to be of limited use, model assessment provides insight into the causes why the model still deviates from the available data in important ways and how it can be developed further and improved.

At this point of the development and validation process, one has arrived at a fully specified model that sensibly represents the hypothesis of interest. This model was shown, given the data, to be the best model in a set of investigated models, and hence the parameter space the set stands for, for representing the hypothesis. Furthermore, this model was validated to be generally robust and useful in embodying the hypothesis in applications to reality. Thus, the specified model was shown to be fit for explaining the information detected in the data. It is now, finally, possible to present the model to the intended, actual hypothesis testing procedure. Testing a fully specified model employing a specific set of empirical data, the line of reasoning is that one can inductively infer from the test result the status of the proposed hypothesis about reality. The test procedure has the goal to provide a decisive conclusion about the falsity of the *a priori* set hypothesis, while expecting and tolerating a reasonable amount of uncertainty (Fisher's interpretation) or allowing for *a priori* set error margins (Neyman-Pearson framework).

If the hypothesis could not be rejected with the specific data at hand, for all practical purposes it is concluded, again employing inductive reasoning, that its corresponding model, now identified and fully developed, is more generally applicable, for example, as a null model representing the hypothesis (even while one could, on philosophical grounds, without falsification not gain additional insight about the hypothesis, that is, reality). In that case, such a previously established and independently specified model can be utilized in subsequent investigations. On one hand, it can be used to test if this concrete model is rejected by new, additional datasets. Such an evaluation of out-of-sample fit and error, starts a new cycle of model development and hypothesis testing. On the other hand, the independently specified model, reasoned to be robust and of more general applicability, can be employed in data prediction approaches. Here, assignment tasks investigate and test, if a proposed origin (a hypothesis) can be falsified for new data instances (e. g. single events or pools of case samples). Assuming the model to be useful for explaining the original reference set of empirical data sufficiently well, the model is now

employed to test if these new observations are rejected. A rejection falsifies claims that the case sample(s) originated from the same data basis, that is, from the same underlying reality (statistical population), as the one that was used to specify the model. If the sample does not fit the model, the prediction is that it came from a different origin.

The inference tasks of model selection, model assessment and hypothesis testing require that the data provide sufficient information in conjunction with the applied approach. Only in that case will it be possible to quantify the fit of a model and its assumptions precisely enough, and to choose between specific alternative hypotheses or – if appropriate and required – weigh and combine their parameter estimates according to the (relative) support of their hypotheses (see section 3.2.2 “Model averaging”).

Model selection towards specification, model assessment for validation and hypothesis testing endeavoring falsification, are closely related and overlap in their objectives, concrete methods and applications. They often cannot be distinguished in an easy way and be clearly separated from each other.

5.1 Model specification

Model specification is the fundamental step in statistical inference (Fisher 1922, 1935). It transforms hypotheses into clearly stated, well-defined models with accurately and precisely specified parameters. Towards this goal, model specification involves the definition of parameters, their settings and interactions, but also the clarification of any implied simplifications and assumptions.

5.1.1 Multiple working hypotheses

Starting out, the main task in model specification is to become aware of one’s perceptions and conscious about reality. It is necessary to bring together and clarify the processes and events that have or could have shaped the investigated evolutionary lineages and that drive their evolution. Investigating an as complex reality as is the natural world, it is, furthermore, of practical advantage to decide, which facets of the investigated complex system are or might be of vital importance with regard to answering the objective at hand.

The core quality of model specification is to approach reality and its empirical data with an open mind that tries to minimize inescapable preconceptions and allows multiple hypotheses to exist – and even to be true - simultaneously (method of multiple working hypotheses; Chamberlin 1890, Elliott & Brook 2007, Rosen 2016). It is not easy to not have preset expectations influencing the inference process. The need for discovery of and reflection on presumptions and biases is already starting at the very beginning with hypothesis definition and selection, continuing to study design

decisions, model development, data exploration and visualization, and further on to the final step of hypothesis testing and results interpretation. Several strategies have been proposed for avoiding a too quick narrowing to favorite pet hypotheses, and instead for keeping the field of prospective models open and diverse (Symes et al. 2015, Rosen 2016). This can be buoyed by challenging one's preconceptions and expectations through "obviously wrong", absurd and outrageous, to highly unlikely hypotheses; by blinding and/or randomizing data; by applying a diverse set of statistical approaches, including non-(process)-model-based, e. g., multivariate, methods; by adding a constant or noise to results; or by removing critical information (e. g., the axes) from graphical representations for initial discussions. Fundamentally, these approaches allow one to become aware of unconscious basic assumptions about the species, sampling, dataset and presumed model, and to question if these really hold. Assigning a colleague as "devil's advocate" or having "pre-mortem" sessions can provide "space for critical reflection and challenge in a structured way" (Lynsky 2017).

An extended outgroup sampling supports more robust and sufficient model specification

Providing support for model decisions and a versatile inference horizon, it is advantageous, if not necessary, to set observations into their evolutionary context from the beginning. Incorporating and increasing an outgroup sampling (including distant lineages and populations) provides (further) support for the monophyly and coherence of the investigated evolutionary lineage. Moreover, it provides information for the question if the reference dataset includes and adequately represents all ancestral populations for admixture analyses. Equally importantly, an adequate outgroup sampling allows the polarization of character states, that is, the identification of the ancestral allele(s) at polymorphic sequence sites. Polarization thus enables the differentiation of ancestral versus derived mutations and, hence, the identification of recent (or ancestral) mutations that are common or rare in a population or lineage. In consequence, such an outgroup sampling can shed light onto the geographical origin of mutations and thus the (potential) origin of samples, In addition, it provides one way to gain insight into the selection pressures specific to an investigated non-model lineage and the selection coefficients of its genotyped sites. Importantly, a wider data basis supports a flexible adjustment of inference approaches, should unexpected patterns and questions arise during the course of application.

5.1.3 Model selection

The ultimate reason for assembling a set of fully specified models for a single hypothesis (given the data) is to evaluate these models. Evaluation has the goal to determine the model or subset of models that best reflects the proposed reality of the hypothesis within the evaluated set. The aim is to identify, for the hypothesis, the model that best fits the data (Bollback 2002, Sullivan & Joyce 2005). At the same time, and of equal interest, insight is gained about support for alternative specifications.

Due to practical and theoretical limitations of defining absolute support, outlined in the following section 5.2 on “Model validation”, the most commonly applied approaches to model selection rely on comparisons. These relative strategies only return that one model fits the data (significantly) better than another does. In evolutionary genetics, several distinct approaches and criteria have been developed for comparing and testing relative model fit in frequentist, likelihood and Bayesian frameworks (e. g. Bollback 2002, Johnson & Omland 2004, Sullivan & Joyce 2005, Beerli & Palczewski 2010, Csilléry et al. 2010, Duchêne et al. 2016). Likelihood approaches are based on the maximum likelihood estimate and are selecting between two or more nested models using likelihood ratio tests or information-theory criteria (e. g., Akaike (AIC) or Bayesian (BIC) information criterion). The information criteria are applicable also to non-nested models, allowing thus more flexibility with regard to permissible model comparisons (Johnson & Omland 2004, Sullivan & Joyce 2005, Grueber et al. 2011).

Within the Bayesian framework relative approaches to model selection can be based on the Bayes factor, which compares the marginal likelihoods of the data across the parameter spaces of the investigated models in pairwise model comparisons (Beerli & Palczewski 2010, Duchêne et al. 2016). Marginal likelihoods of the data for complex evolutionary models (also keeping in mind very large sample spaces) require approximation, either by importance sampling (harmonic means and AICM criteria), or by path sampling (thermodynamic integration and generalized stepping-stone sampling).

An alternative to Bayes factors are cross-validation procedures that try to approximate out-of-sample errors, based on deviations between model specifications that were inferred from different sub-datasets. Cross-validation approaches circumvent the need for (proper) priors and the calculation of marginal probabilities (Zhou et al. 2007, Duchêne et al. 2016). They repeatedly split the data (and/or summary statistics calculated from the data) into training and test sets during maximum likelihood or Bayesian inference. They specify each model (and/or select the set of diagnostic summary statistics) based on the training dataset and subsequently compare the expectations for the model likelihoods (or summary statistics) inferred from the test dataset for all of the proposed models (or model structures) that are tested.

Tests and choices among large sets of *a priori* defined and applicable models, to even exhaustive sets that include all possible model structures across the whole of the relevant parameter space, can be and generally need to be organized in a methodical way. Model evaluations and pairwise tests can be structured by codified decision trees (e. g. evaluations of increasingly complex sequence substitution models, cp. the classic ModelTest by Posada & Crandall (1998)), by using heuristic algorithms (cp. jModelTest 2, Darriba et al. 2012) or organized as systematic grid-based evaluations (e. g. Approximate Bayesian Computation (ABC) approaches, cp. Pelletier & Carstens (2014)). Thus, beyond pairwise comparisons, model selection itself can be standardized as a logically and systematically structured inference approach.

All relativistic model selection approaches pose the problem that significance thresholds cannot be objectively defined. More importantly, even the clearly best model of a set of models might still be incorrectly applied and have a very low absolute probability. Investigation-specific inference of significant relative support without an immediately available option for the assessment of the general validity and usefulness of the finally chosen model, are especially of concern in applications of widely employed admixture clustering and relatedness analyses. They require the presence of an appropriate sample of populations as a reference baseline and, moreover, an appropriate relative representation in the reference dataset (Rousset 2002, Manichaikul et al. 2010, Powell et al. 2010, Thompson 2013, Thornton et al. 2014, Wang 2014). As one of several reasons, such a reference baseline is needed for the precise estimation of population allele frequencies. This is of consequence, because the approaches rely fundamentally on population allele frequencies and are known to be too sensitive and at the same time too robust to violations of their assumptions and to misspecifications. They will not return meaningful results, if adequate ancestral populations are not represented in the dataset, without necessarily providing a way to detect this (Hauser et al. 2006, Thornton et al. 2012). Even more fundamentally, partition clustering methods provide no indication if they should (not) have been applied at all. They will return clusters, even if they are applied to populations that are predominantly structured by isolation-by-distance processes (Wright 1943, Malécot 1973, Ishida 2009), and not the island-model process (Wright 1931) appropriate for admixture clustering algorithms (Guillot et al. 2009, Meirmans 2012).

Setting the model for the evolutionary null hypothesis in forensic casework

In forensic conservation genetics, the process of model specification applies to both stages of an assignment objective. In the first step, the goal is to infer and select the global evolutionary model for the reference dataset, which best represents the natural processes and historical events that shaped and shape the taxon under investigation, or at least represents its evolutionary lineage-wide diversity patterns (null hypothesis). In the second, applied step, the model for the null hypothesis might need to be further focused and re-specified for the objectives of a concrete case. Subsequently, this case-adapted null model is modified to correspond to the alternative hypotheses particular to the certification and forensic test questions at hand.

In applications, the specification of the appropriate modifications to generate the model(-s) of the prosecution and/or defense often are in the focus. However, it is not always obvious which null model (i. e., what evolutionary model for the natural populations under investigation) most appropriately to choose, on which these modifications can be based (Rohlfs et al. 2012, Thornton et al. 2012, Weir & Zheng 2015, Buckleton et al. 2016). That is, it often is not clear, which null model is an adequate starting point for the set of models that reflect the forensic case and will be tested.

The general difficulty to arrive at absolute support, that is, at an absolute probability for a data-hypothesis combination (see section 5.2 “Model validation”) can result in the choice and application of an incorrect basis for model and hypothesis testing in the relative framework. This is the case, when basic model assumptions are not met, but this is not detected or questioned. Model selection and thus hypothesis testing that does not involve a null model or a set of null models, which reflect reality well, was shown to have real-world consequences in practical application. Incorrect conclusions due to a lack of model adequacy were revealed, for example, by Rohlfs et al. (2012) to be a realistic problem in practical application. They showed that a lack of model adequacy in the context of familial searching in national offender databases led to the discrimination of the Native American community in the USA. Model checking, as performed by Rohlfs et al., provides a way to explore the appropriateness of the null model for the data under consideration. They used different subsets of the reference sample, that is, supervised definitions of human ancestral populations representing different evolutionary scales and processes, to evaluate the impact of ancestral background on error rates. Since then, several approaches were developed to ameliorate the shortcomings pointed out be Rohlfs et al. For example, admixture models and inference approaches were developed for heterogeneous populations (cp. many of the ascribed ancestral backgrounds in humans, among them “Native American”). Thereby, investigated heterogeneous human populations were found to be characterized by admixture between very divergent ancestral backgrounds and assortive mating dynamics (e. g. Manichaikul et al. 2010, Thornton et al. 2012).

5.2 Model validation

5.2.1 The posterior probability as measure of absolute model support

The intuitive desire would be to be able to straightforwardly calculate how probable it is that a model is true and applicable. This intuition follows the knowledge that in the reverse situation this approach works. Here, given a model (e. g. a well-balanced die), it is logically possible to stochastically calculate the probability of observed or predicted data (e. g. sets of rolls). As an alternative notion to model selection based on relative comparisons, therefore, it should be possible to try to assess model validity and adequacy based on absolute model probabilities. In the Bayesian framework, the posterior probability seems to provide such a measure for quantifying the absolute validity of an optimized, completely specified model and its parameters given the present data.

Taking into account the insight that all models are wrong by Box (Box 1979), one would not expect a posterior probability of 1.00 for an inferred model. Still, the posterior probability, so is the expectation, should provide an absolute value of how close we are to the truth. However, even this has been refuted, or at least questioned, on philosophical grounds (Gelman & Shalizi 2013). Bayes’ formula requires that a prior distribution is defined, thereby setting the parameter

space pertinent for the posterior probability. Gelman & Shalizi (2013) pointed out that, therefore, the posterior probability will be dependent on the scope that was imagined for the prior distribution. Any models, which were not considered when designing the prior distribution, but which cannot be rejected per se, will not be covered by the estimate for the posterior probability.

"Fundamentally, the Bayesian agent [i. e. the calculation process resulting in the posterior probability] is limited by the fact that its beliefs always remain within the support of its prior. For the Bayesian agent the truth must, so to speak, always already partially believed before it can become known. … To sum up, what Bayesian updating does when the model is false (i. e. in reality, always) is to try to concentrate the posterior on the best attainable approximations to the distribution of the data, 'best' being measured by likelihood. But depending on *how* the model is misspecified, and how ϑ represents the parameters of scientific interest, the impact of the misspecification on inferring the latter can range from non-existent to profound." (emphasis as in the original; Gelman & Shalizi 2013).

Even if the prior is supported by the truth or in support of the truth, respectively, and, thus, includes the most useful model, the posterior probability will always remain tied to the data at hand (Gelman et al. 2014 p.170). This is the case, since per its definition the posterior probability is defined as conditional probability: the probability of the model given the concrete data investigated. Hence, in addition to an adequate scope for the prior, the data need to be sufficiently informative to allow generalization. It is often not (immediately) clear, in as far the data are sufficient and representative enough to allow further application and what the limits of an appropriate generalization are.

Finally, computation of the posterior probability often involves complex heuristic inference approaches, which endeavor to sample the parameter space representatively to obtain the marginal likelihood of the data. These heuristic algorithms can be prohibitively resource intense to impossible to even design.

Hence, returned posterior probabilities are limited in scope of information content, explanatory power and applicability, first, with regard to the parameter space that was imagined for the prior, second, with regard to the representativeness of the concrete dataset the inference is based on and, third, with regard to the sufficiency of the heuristic inference algorithms.

5.2.2 Model assessment

While absolute model validation is limited and difficult based on theoretical and pragmatic grounds, it is always possible to assess models' discrepancies from the truth present in the data, that is, their usefulness, and to determine the limits of their generalization power and applicability. (cp. Hastie et al. 2009 pp.219-223, Gelman et al. 2014 pp. 170-171, 178 ff.). Such model assessment combines model selection (see section 5.1 "Model specification") with

evaluations of model sensitivity and model checking (see section 3.2 “Robustness”) for inference from one concrete, given (reference) dataset (cp. the simultaneous assessment of uncertainty along both axes in fig. 3.4).

Under the umbrella of a single evolutionary (or forensic) hypothesis, it is assessed if a model or set of models, including their assumptions, can be validated to sufficiently fit the data under investigation. Thus validated, the model(s) will provide a reliable basis for the subsequent testing of the corresponding hypothesis and, thus, an inductive drawing of conclusions about reality. Assessment of the model itself and its comparison with additional challenging sub-hypotheses, ideas and evolutionary processes provides insight if, for example, a specific null model remains the best fitting representation of reality and at the same time retains the lowest prediction error. An additional part of model validation investigates, if the selected model remains reliable, when the context is widened by considering more and different data (partitions). The application of the model to independent data allows an assessment of its qualification for generalization. Model assessment for the purpose of model validation, thus, has direct ties to and application in hypothesis testing.

Predictive approaches, such as posterior predictive checks or predictive inference in conjunction with cross-validation, that is, model checking and Approximate Bayesian Computation (ABC) approaches (model selection and sensitivity based on inferences from real data) provide inference approaches for evolutionary data that are often used in the assessment of model adequacy (Bollback 2002, Sullivan & Joyce 2005, Csilléry et al. 2010). Designing and setting up model assessment as a systematically organized procedure involves difficult, often conflicting decisions within extensive, complex and highly dimensional spaces.

The design of ABC simulations can be intricate, since the necessary choices might not be obvious. These choices involve the to be included model structures and parameters, the ranges of their values and priors, as well as, the summary statistics to be used for the evaluation of the discrepancy from the observed, “real” data (reviewed in Beaumont 2010, Bertorelle et al. 2010, Csilléry et al. 2010, Hartig et al. 2011, Sunnaker et al. 2013). The results of ABC-approaches might not always represent consistent approximations of posterior probabilities. This is the case, if important characteristics of prior distributions are unknown, parameterizations of models are too uncertain and the marginal likelihood distribution of the given data across all of a too large and too complex parameter space involving too many dimensions cannot adequately be approximated, that is, representative sampling of the parameter space cannot be achieved.

Model checks based on the generation of predictive distributions have been integrated as an additional step in ABC-inference approaches. In general, the evaluation of model adequacy based on predictive accuracy will always remain dependent on the representativeness of the concrete dataset that formed the basis for the assessment, for example, a reference dataset. This is even the case, if cross-validation is used to approximately quantify generalization error (out-of-sample-error; Gelman et al. 2014 p. 170). Predictive approaches including cross-validation loops for

datasets from evolving, highly diverse and often, to different degrees, heterogeneous natural populations require many data partitions and predictive repeats. Here, small changes in the test data partition can result in large fluctuations of predictive accuracy. Thus, predictive accuracy not only might remain rather limited within the original reference dataset, but moreover might be limited in its informativeness for drawing generalizations beyond it.

Overall, for many parameter-rich models and larger datasets, repeatedly needed steps involving model inference from the original, real data and the generation of replicate data with subsequent inference from these simulated datasets, quickly result in comprehensive model assessment becoming challenging to design and computationally very resource-intense.

An example of an expanding model assessment approach over successive studies

Versatile and well-known reference datasets provide the possibility to go beyond the one-step inference of a single null model. They allow a shift in focus from parameter inference for one (pre-selected default) model, to model development and improvement, leading to comprehensive model assessment. Carstens et al. provide a well thought-out example for such an objective, strategy and workflow (Carstens et al. 2009, Carstens et al. 2013, Hickerson 2014, Pelletier & Carstens 2014). They conducted and discussed a theoretically grounded and practically feasible approach for the testing of multiple working hypotheses. Their objective was to test the pre- to post-glacial population dynamics and recolonization patterns of two NW-North American species, a willow and a salamander. The thereby inferred null models of evolution were to be applied as the basis for the development of conservation and management strategies and in decisions associated with them.

Carstens and colleagues illustratively show, how to progress from preexisting results to the identification of a stepwise expanding set of reasonable, more or less likely alternative models. The resulting set(s) of models at each step, formed the basis for a more systematic and extensive exploration of the model (parameter) space. In their procedure, they complemented inference based on fully parameterized population-genetic models that use the full likelihood function, with more approximate inference approaches that allow a more flexible and adaptable design and scope of analyses. As far as realizable, they analyzed their datasets with inference software using the full likelihood function. They discuss that full likelihood approaches, while desirable, are still rather limited in scope due to restrictions of the implemented programs, feasibility and resource-limitations. Thus, their exploration of the parameter space also encompassed the application of Approximate Bayesian Computation and posterior predictive simulations. These provide flexible approaches for inference from complex population-genetic models that are to date not tractable by full-likelihood inference. They evaluated all models in pairwise comparisons, based on information theoretical criteria, or assessed them within a Bayesian probabilistic framework. Using these procedures, they identified models, and thus evolutionary processes, of better fit to the data. Subsequently, both options provided a basis for model averaging that allowed summarizing the relative contributions of different evolutionary processes

into an overall result. At the same time, model averaging provided additional insight into the sensitivity of the result to different incorporated processes. While none of their inference approaches involved population assignment itself, the detailed inference of models accurately representing population structure, demography and history is a prerequisite for subsequent robust and reliable population-level assignment.

5.3 Hypothesis falsification

On the way towards gaining insight into reality, hypothesis testing forms the last step in any investigation of the world. This step can be implicit and intuitive in truthful descriptions of the world ("fits well enough, looks right and makes sense"). In systems with (quantifiable) uncertainty and (explicable) random variation, it most often is formalized as a stringent statistical procedure. This procedure provides the building blocks for logical arguments that allow conclusions about truth. In highly complex and expansive systems with rugged parameter spaces and many unknowns, due to limited data and a lack of pre-existing knowledge, statistical testing procedures can but provide an assessment of limited scope. For such systems, efforts to arrive at conclusions that can be generalized require additional analyses from independent starting points with regard to models, statistical inference and data. Even such combined efforts in the end often resort, at least in part, to informal, intuitive induction in their overall conclusions. This is the case, until new developments of methodology might expand the scope and power of formalized statistical inference and tests for additional parts of the whole inference process.

The prerequisites for both, informal and formal, hypothesis testing are, first, that the available data were explored and found to be appropriate, because relevant to and sufficiently informative, for decisions concerning the investigated, *a priori* proposed hypothesis or for distinguishing between proposed alternative hypotheses. Second, statistical hypothesis testing requires that in previous inference steps, reliable model structures and fully specified models were already identified for the representation of the hypotheses. Thus, it requires that for each investigated hypothesis the best model or set of models was already selected and found to be adequately useful and robustly valid. To distinguish conceptually and consciously model selection and assessment from the actual hypothesis testing phase, ensures that the model representing the corresponding hypothesis summarizes in the best possible way the information in the data. Fortunately, the reliability of conclusions from hypothesis testing in applications does not require complete fidelity of the (in evolutionary genetics often inherently heuristic) inference approach to the model (Gelman & Shalizi 2013). Thus, applied inference tools do not need to be perfect to work.

These prerequisites and qualities are of importance for the ability to successfully transfer a developed model into a practical inference tool. Such inference tools cumulate in formal hypothesis testing and thus consolidate the basis for reliable reasoning in everyday application.

5.3.1 Hypothesis testing in a behavioral context focusing on predicted error rates

The formal statistical procedures for evaluating the falsity of a hypothesis by means of its associated specified, best model (including the inference approach) in hypothesis testing are the same as for evaluating, comparing and testing models or inference approaches for the purpose of selecting the best model, exploring and checking its robustness and ascertaining its validity in model assessment.

In applications, as for example, certification or forensic case contexts, the preparation, execution and interpretation of hypothesis tests follow a behavioral motivation (Romeijn 2017). Hypothesis testing is conducted with the aim to provide support for a decision that has behavioral consequences, e. g., enabling a certification or legal body to act. In such behaviorally motivated contexts, hypothesis testing often follows the framework developed by Jerzy Neyman and Egon Pearson (Neyman & Pearson 1933, see also section 5.3.3 "Understanding and interpreting test results"). Here, testing involves alternative hypotheses. These can be the hypothesis of the defendant versus the one of the prosecution, or vice versa. The hypotheses can also be a null hypothesis that is compared to a mutually exclusive alternative hypothesis covering exhaustively all other possibilities (i. e. the negation of the null hypothesis). For cases involving natural populations, one of the tested hypotheses might assume the evolutionary null hypothesis, represented by the natural evolutionary model. The alternative hypothesis then is represented by a forensic case model, which might either be restricted in scope to a specific geographic origin or include anthropogenic processes.

The possible outcomes of a hypothesis test can be visualized in a confusion matrix (fig. 5.1), the core tool for designing, optimizing and interpreting diagnostic tests. A range of quality statistics and optimization strategies have been developed to evaluate and improve classification decisions, with the goal to minimize assignment errors (e. g. Hastie et al. 2009 pp. 313 ff., Kelleher et al. 2015 pp. 397 ff.).

In practical applications requiring decisions, the results of hypothesis tests are summarized and interpreted with a focus on the inferred estimates of error rates. Population assignment tasks present classification problems, which are evaluated based on estimates of overall predictive accuracy and predictive errors (fig. 5.1). These provide the rationales for assignment decisions and casework conclusions. Without ground truth for case samples, overall assignment accuracy and error rates are estimated indirectly, by transferring the inferred quality estimates from an appropriate self-assignment procedure conducted with the samples of the employed reference dataset.

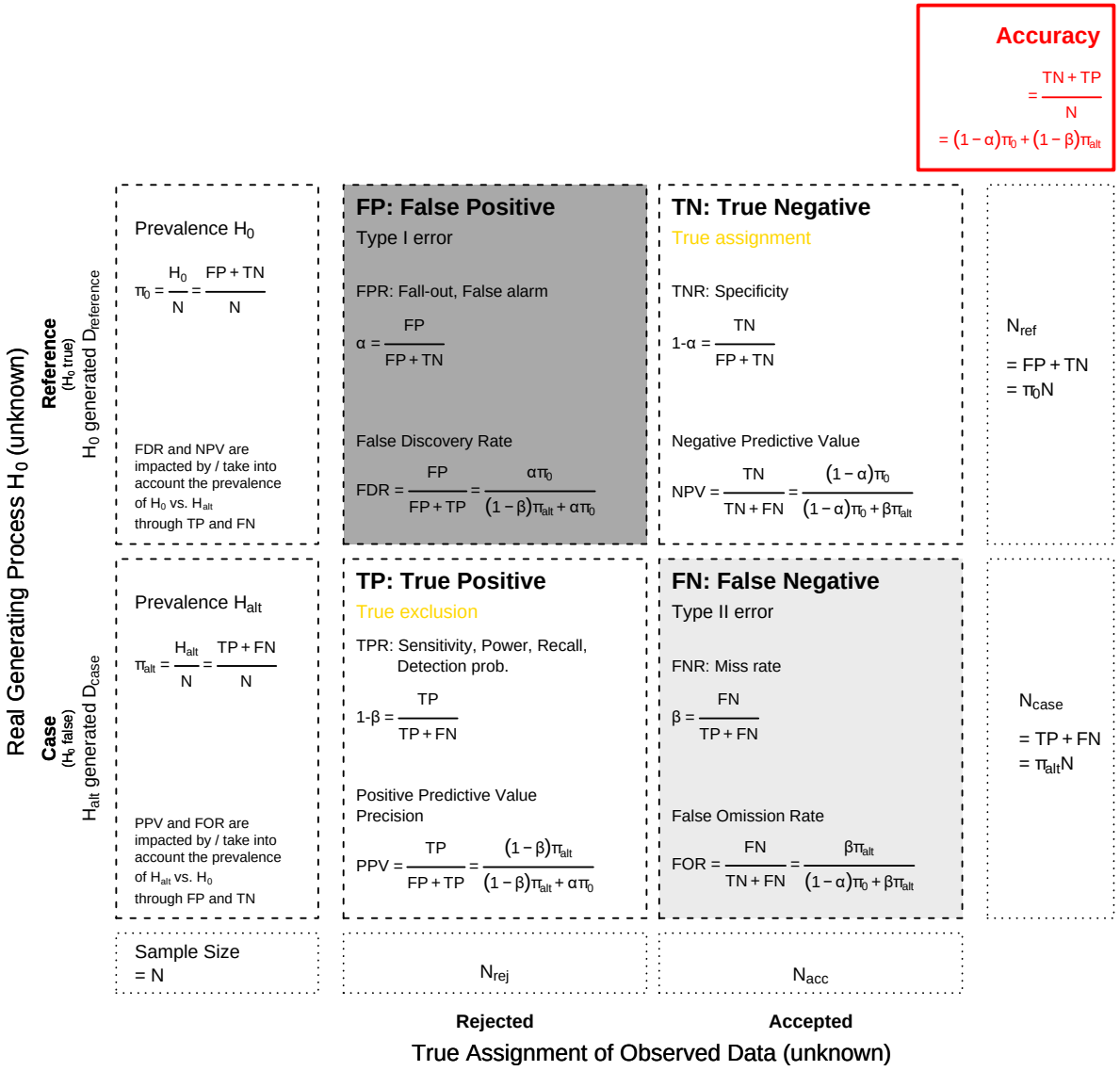


fig. 5.1: Confusion matrix summarizing all possible outcomes of (self-)assignment and exclusion tests. The confusion matrix approach corresponds to the hypothesis testing framework of Neyman-Peason for testing mutually exclusive alternative hypotheses. The equivalences in the formulations present an ideal case in which, first, the sample-based cumulative, overall estimators for the reference dataset provide accurate estimates of the population parameters of, e. g., prevalence, the significance level α and discriminatory power β, and, second, the out-of-sample (generalization error) is negligible.

Following the interpretation of hypothesis testing results developed by Neyman and Pearson (Neyman & Pearson 1933, Lew 2012), the significance level α needs to be decided upon and set in advance, before the testing of the case sample, of blinded samples or self-assignment. In addition, the sample size, required for successfully being able to conduct hypothesis testing at all, needs to be determined beforehand. The sample size determines the power that can be expected for being able to discriminate between the two tested hypotheses at a given effect size, e. g. level of population differentiation.

In practical applications of assignment testing, the setting of the significance threshold starts by subjectively deciding upon a socially acceptable overall false positive error rate (FPR, Type I error). The cumulative (overall) false positive error rate achieved in the self-assignment of all reference samples is determined by the corresponding significance level α (Sokal & Rohlf 1995), which was employed in the testing of each of the leave-one-out test samples within the cross-validation steps of the self-assignment procedure. Hence, the estimate of the FPR for the reference dataset identifies the significance level to be used in the testing of the case sample. This significance level commonly represents a log likelihood ratio threshold. The significance level α and the FPR are the same, if the sample design closely represents reality and the out-of-sample, that is, the generalization error is negligible. Thus, the subjectively deemed acceptable overall false positive error rate provides the rationale for adjusting and setting the significance level α used for all of the sample-specific (individual) hypothesis tests, conducted both, during self-assignment and for testing the case sample.

In the same way, the expected ability of a reference dataset and inference procedure to successfully test a case sample, that is, the power of the hypothesis test, needs to be evaluated and its contributing factors specified in advance. Only if the expected discriminatory power of the test is sufficient, can hypothesis testing be sensibly conducted and a proceeding to case work makes sense. The discriminatory power of the hypothesis test (1-β), the complement of the Type II error (β), cumulatively resulting in the overall false negative error rate (FNR), depends on the choice of the significance threshold α, and, thus, the accepted overall false positive error rate. The ability to achieve a required level of discriminatory power depends on the sample size during the sample-specific hypothesis tests, as well as, the number of cross-validation repeats, which provide an estimate of the out-of-sample errors and therefore the test power. In assignment testing employing (posterior) predictive distributions, the number of generated predictive repeats can be adjusted as required within each cross-validation step. Here, forensic tools should identify the number of required repeats based on the observed level of population differentiation as measure of the expected effect size. Nevertheless, the effect size also depends on the sample sizes representing the investigated populations, which determine the number of cross-validation leave-one-out repeats per investigated population.

In practice, during an initial phase of exploratory self-assignment of the reference samples, the significance level α is modified repeatedly, until the estimates of overall prediction accuracy and

the overall prediction error rates (FPR and FNR) are close to the *a priori* and independently decided upon error margins considered acceptable for hypothesis testing.

One area that still needs more thought, before assignment testing in conservation genetics can be considered to have developed into the level of applied tools, are appropriate estimates of the base-rate frequencies. Base-rate frequencies, also called prevalences, need to be taken into account for the investigated populations under the null hypothesis, as well as, for the occurrence of the alternative, the case hypothesis (e. g. of illegal logging). The prevalences of the two hypotheses set the FPR and FNR error rates into an appropriate relation to the expected true positives and negatives, respectively. This is required for the calculation of the positive and negative predictive values, respectively. In the Bayesian context the base-rate frequency might be estimated by population frequencies or determined based on independent information as "subjective" Bayesian priors. No clear guidelines exist on how to arrive at sensible values for the prevalences in forensic conservation genetic assignment testing to natural populations.

5.3.2 Specificity, discriminating power (sensitivity) and precision in hypothesis testing

All of the reviewed and described statistical criteria have an impact on the quality of the inferred outcome of a hypothesis test. They contribute to arriving at a behavioral decision via successfully, that is reliably, accurately and decisively, testing the hypotheses of the investigated case.

Statistical procedures that are valid, that is, convergent, consistent, robust and congruent, provide accurate and reliable estimates of parameter distributions and predictive data distributions. They therefore can be expected to result in realistic (preferably also lower) false positive error rates (Type I error, α), allowing the use of more stringent significance levels. Thus, despite the limitations of sampling, they reduce the possibility of convicting an innocent, by being specific (specificity, $1 - \alpha$).

Towards the goal of developing efficient and effective forensic decision tools it is necessary to strike a balance between being highly specific, and the chance of acquitting a perpetrator (Type II error β, false negative error rate) and thus being ineffective (high miss rate and false omission rate). It hence is advantageous and desirable to have efficient and sufficient, thus, highly informative statistical methods with low levels of noise contributing to the decision process. Such approaches are able to incorporate, analyze and resolve ever finer and complex genetic patterns and interacting processes. This suggests that these criteria will lead to high(er) discriminating power (sensitivity, $1 - \beta$) for the hypothesis test, as well as, low false discovery rate, high predictive values and thus high overall predictive accuracy.

5.3.3 Understanding and interpreting test results

In the final step of evaluating the original hypothesis or hypotheses, it is necessary to exactly understand the meaning of the outcomes of the hypothesis tests, that is, the support measures and their quantitative values. Only the correct interpretation of *P*-values, likelihoods and thus likelihood ratios, as well as, posterior probabilities (estimated by, e. g., overall predictive accuracy), and therefore Bayes' factors and posterior odds, will allow one to come to a meaningful and reliable conclusion regarding the rejection of a hypothesis.

In a first approach, the definitions of the support measures are well-known. In the frequentist framework, the *P*-value measures the "strength of evidence in the data against the null hypothesis" (Lew 2012 p. 1560). It is "the long-run frequency of obtaining data as extreme as the observed data, or more extreme, given that the null hypothesis is true." (Lew 2012 p. 1560). R. A. Fisher interpreted the *P*-value as a continuous variable, which, together with additional, independent (background) knowledge, is to guide intuition about the next step in an inference process (Lew 2012, Halsey et al. 2015). Contrarily, Jerzey Neyman and Egon Pearson understood the *P*-value as a dichotomous variable, for which, in conjunction with the determination of the expected effect size and the desired power, an *a priori* threshold is set (the significance level α) that decides in a final, absolute way about the rejection of the hypothesis (Lew 2012).

In the likelihood framework, the likelihood function provides "a measure of relative plausibility" for different values of the model parameter(s) (Barnard et al. 1962 p. 321). The likelihood principle, moreover, defines that the likelihood contains all the evidence present in the data that is relevant to a given model and its parameters (Birnbaum 1962). It is thus independent from the context and the inference process within which it was calculated.

In the Bayesian framework, the posterior probability most commonly "expresses the strength of belief in the hypothesis" (Romeijn 2017 section 4.2.1). Here the most widely held view is that progress focuses less on tests that eliminate hypotheses from a set of hypotheses (Romeijn 2017), but "[rather] than employing tests and attempted falsification, learning proceeds more smoothly: an accretion of evidence is summarized by a posterior distribution … The goal is to learn about general laws, …" (Gelman & Shalizi 2013 p. 9).

---

fig. 5.2 (following side): Decision tree underlying the "Tree Detectives Rallye" at the open house of the Thünen-Institute of Forest Genetics on September 16, 2018. Contrasting assignment results are obtained based on the model likelihood versus the posterior probability. For example, based on the likelihood of the model, the small wood cubes are more likely to have their origin in population 1. In contrast, the posterior probability supports an origin in population 2, due to consideration of the sampling (i. e., logging) context in the form of the prior. The background map (right side) showing the distribution of chloroplast haplotypes in European white oaks (*Quercus* L. sect. Quercus) across Europe was extracted from Petit et al. (2002).

Station 1:
Species Identification

Qualitative exclusion traits / criteria:
Species Identification

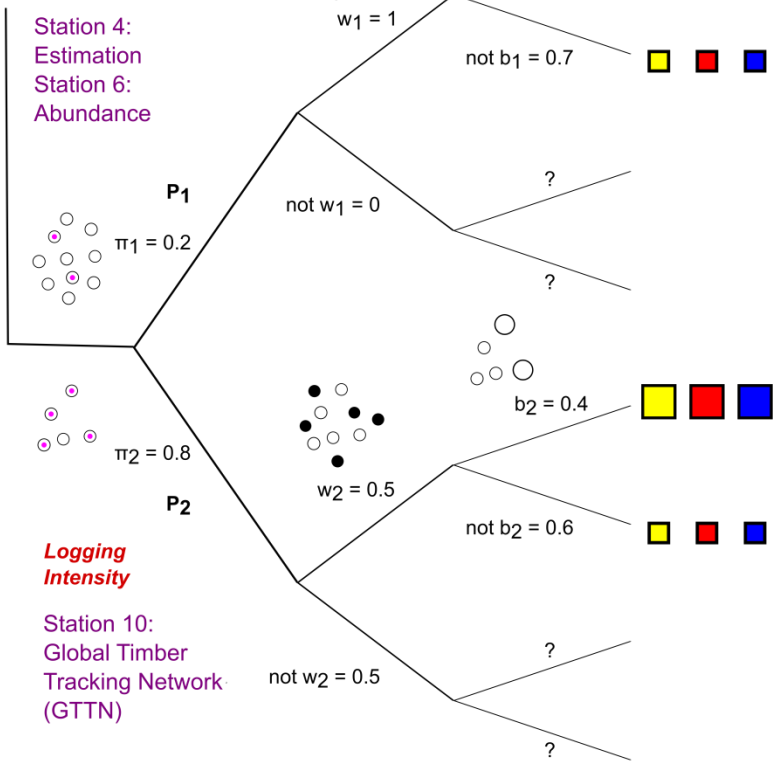
Station 8:
Genotyping

Chloroplast-Haplotype

Station 5:
Measuring Trees

Growth

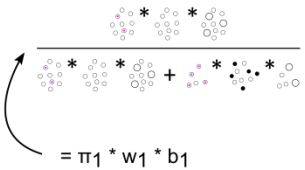
Station 4:
Estimation
Station 6:
Abundance



Logging Intensity

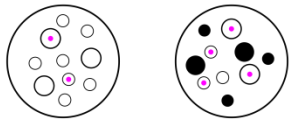
Station 10:
Global Timber Tracking Network (GTTN)

Bayes´ formula:

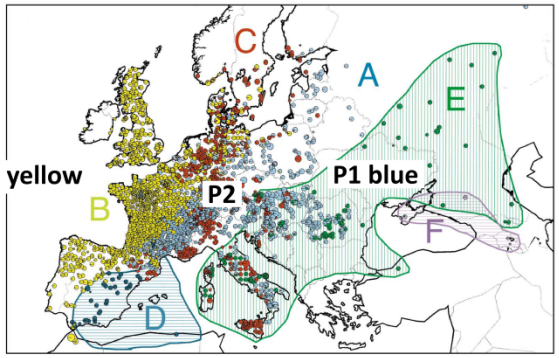


$= w_1 * b_1$ $= 1 * 0.3$ $=0.3$	$= \pi_1 * w_1 * b_1$ $/ 0.16 + 0.06$ $= 0.06 / 0.22$ $= 0.27$
$= w_1 * \text{not } b_1$ $= 1 * 0.7$ $=0.7$	$= \pi_1 * w_1 * \text{not } b_1$ $/ 0.24 + 0.14$ $= 0.14 / 0.38$ $= 0.37$
$= w_2 * b_2$ $= 0.5 * 0.4$ $=0.2$	$= \pi_2 * w_2 * b_2$ $/ 0.16 + 0.06$ $= 0.16 / 0.22$ $= 0.73$
$= w_2 * \text{not } b_2$ $= 0.5 * 0.6$ $=0.3$	$= \pi_2 * w_2 * \text{not } b_2$ $/ 0.24 + 0.14$ $= 0.24 / 0.38$ $= 0.63$

Composition of the two potentail populations of origin
(white: haplotype of the wood piece, black: a different haplotype)



Station 9:
Geographic Reference Data



Petit et al. 2002

Compare the support for both populations, 1 and 2.
Is a clear conclusion about the origin of the wood possible?

$L(\vartheta_1 \mid w, \circ) / L(\vartheta_2 \mid w, \circ)$ $= 0.3 / 0.2 = 1.5$ not significant at LR = 2	$P(\vartheta_2 \mid w, \circ) / P(\vartheta_1 \mid w, \circ)$ $= 0.73 / 0.27 = 2.7$ anecdotal evidence
$L(\vartheta_1 \mid w, \circ) / L(\vartheta_2 \mid w, \circ)$ $= 0.7 / 0.3 = 2.33$ significant at LR = 2	$P(\vartheta_2 \mid w, \circ) / P(\vartheta_1 \mid w, \circ)$ $= 0.63 / 0.37 = 1.7$ anecdotal evidence

Prior knowledge π	Modell ϑ	**Concrete Wood Cube**	L(M | D): Likelihood	P(M | D): Posterior Probability	Likelihood Ratio	Posterior Odds
Independent Information not obtained from the wood cube: Logging intensity	Information from the wood cube: haplotype w und growth b The traits are independent, but their abundancies are dependent on the population of origin	Rallye-participant is holding it in her/his hand	$\sum^4 = 1.5$	$\sum_{\circ} = P(\vartheta 1 \mid w, \circ) + P(\vartheta 2 \mid w, \circ) =$ $\sum_{\circ} = P(\vartheta 1 \mid w, \circ) + P(\vartheta 2 \mid w, \circ) =$ $= 1$		

While these definitions might be considered minutiae, only of interest to philosophers of science and theoretical statisticians, the support measures and their interpretations repeatedly are shown to produce contradicting conclusions in examples. Such examples can be quite close to realistic inference procedures in forensic conservation genetics (fig. 5.2). Several causes have been determined for such contradicting conclusions and additional problems in practical application.

First, one problem is the inadvertent use of a hybrid form of the *P*-value, mixing Fisher's and Neyman-Pearson's interpretations (Lew 2012). This is currently of acute concern regarding studies testing population assignment in forestry and fisheries. These implicitly employ the Newman-Pearson framework, but significance thresholds are not explicitly set *a priori*, rather *P*-values are interpreted in a Fisherian way. In addition, neither effect size, nor power are determined in advance to guide the study design, or reported *a posteriori* in publications, raising concerns about reproducibility (cp. Sham & Purcell 2014, Halsey et al. 2015).

Second, the base-rate fallacy needs to be taken into account. This connects to the question, if the behavioral and error-based application of Neyman-Pearson hypothesis testing in the frequentist framework can be expanded to provide information about the epistemic nature of the tested hypothesis, that is, for the believe in its truth or falsity (Romeijn 2017 section 3.2.1). In the Bayesian framework, Bayes' formula resolves the base-rate fallacy by incorporating estimated population frequencies or independently determined priors (e. g. illegal logging frequencies or hybridization rates) in the calculation of the posterior probability.

Third, in frequentist approaches the problem of optional stopping contexts persists, which points out a violation of the likelihood principle (Romeijn 2017 section 3.2.2). The background here is, that the decision to reject a hypothesis not only depends on the concretely used available dataset, but also on the probability distribution across the sample space (cp. fig. 5.2). For reliable inference and decisions, the investigated scope needs to consider sample events that are possible, but were not observed. Transferred to practical applications, this describes the problem that different "intentions", that is, objectives, sample designs and inference procedures can influence the outcome of a hypothesis test.

Fourth, in the Bayesian framework very influential interpretations of the probability over hypotheses cause problems. These associate the degree of believe in a hypothesis directly with behavioral consequences, i. e. as a willingness to act. Thereby, restraints are put on the definition and scope of "believe" by pragmatic views with unconscious basic assumptions, which solely focus "believe" on and restrict it to navigating the world successfully. Such a pragmatic focus clashes with a scientific and statistical setting that is more concerned with "believe" as a truthful representation of the world. Expressed more formalistic, this concerns assumptions which take for granted that the tested alternative hypotheses and their priors with certainty include the true hypothesis, thus not doing justice to the tenet that "scientific theory must be open to revision at

all times" (Romeijn 2017 section 4.2.1), compare also the discussion of the scope of priors in Gelman & Shalizi (2013).

Fifth, Bertrand's paradox introduces problems. It states that "in some cases, we do not even know what parameters to use to express our ignorance over" (Romeijn 2017 section 4.2.2). This problem concerns the nature and application of "subjective" priors.

All of these open problems contribute to an ongoing and currently very active interest in the foundations of reasoning, the meaning of the support measures and their (potential) correspondence to each other.

5.3.4 Strong inference, severe testing, model checking and expansion

The current, from the outside confusing and conflicting, stage of the development process towards a, potentially, more coherent statistical theory, however, does not preclude hands-on practical application of statistical inference already today. Reliable and decisive conclusions realistically can be achieved for decision processes in conservation, management and law enforcement (e. g. Downes 2010).

Several principles, concepts and procedures have been developed to ascertain and improve reliability and predictive capability. These provide the prerequisite and foundation for a successful transition from academia and basic biodiversity research to conservation applications (fig. 5.3). Among these are the principles of **multiple working hypotheses** (Chamberlin 1890, Elliott & Brook 2007), **strong inference** (Platt 1964), **severe testing** (Mayo 1996), **model checking** (Gelman & Shalizi 2013, Gelman et al. 2014) and **model expansion** (Gelman et al. 2014). Most of these approaches were already introduced and discussed earlier in this review. They stress two general principles: first, to clearly design and formulate the hypothesis that is to be tested and to devise alternative hypotheses and/or test statistics that can detect specific deviations from the data and thus consequential errors of the hypothesis. Second, individual results of hypothesis tests need to be placed into a wider context of additional samples, data types, models and (predictive) tests. The principles taken by heart and implemented will result in growing, complementary knowledge and insight, an accumulation of evidence, the stepwise building, expansion and, thus, practical availability of useful models and of discrepancy probes, as well as, their required inference environments and pipelines, which are capable of thoroughly interrogating for and detecting errors.

Transferred into the forensic conservation genetic context and applied to its practical tasks, these guidelines stress the importance of assembling a reference dataset, which by spanning the whole distribution range and all ecological habitats, as well as, representing the whole genome provides the necessary evolutionary, ecological and anthropogenic contexts for strong inference. Thorough data exploration and cleaning of the resulting dataset will provide a reliable and well

understood data basis. Analyses and severe testing of this cleaned reference dataset, both globally and locally, using unsupervised and supervised inference will provide in-depth experience with and knowledge of the strength and limits of the data. Model checking, model assessment and evaluations of predictive accuracy ensure the development and availability of versatile models and inference approaches attuned towards and optimized for specific tasks. A wide variety of complementary assessment approaches, including predictive evaluations, at the same time, provides insight into the limits of the achieved generalization capability. In casework, self-assignment of the reference samples provides context for the assignment and exclusion probabilities obtained for a single, unknown case sample, while model expansion to Bayesian hierarchical models (cp. also Bayesian network approaches) allows the combination of many different datatypes and data partitions, as well as, independent evidence into a single inference procedure for an overall evaluation of uncertainty.

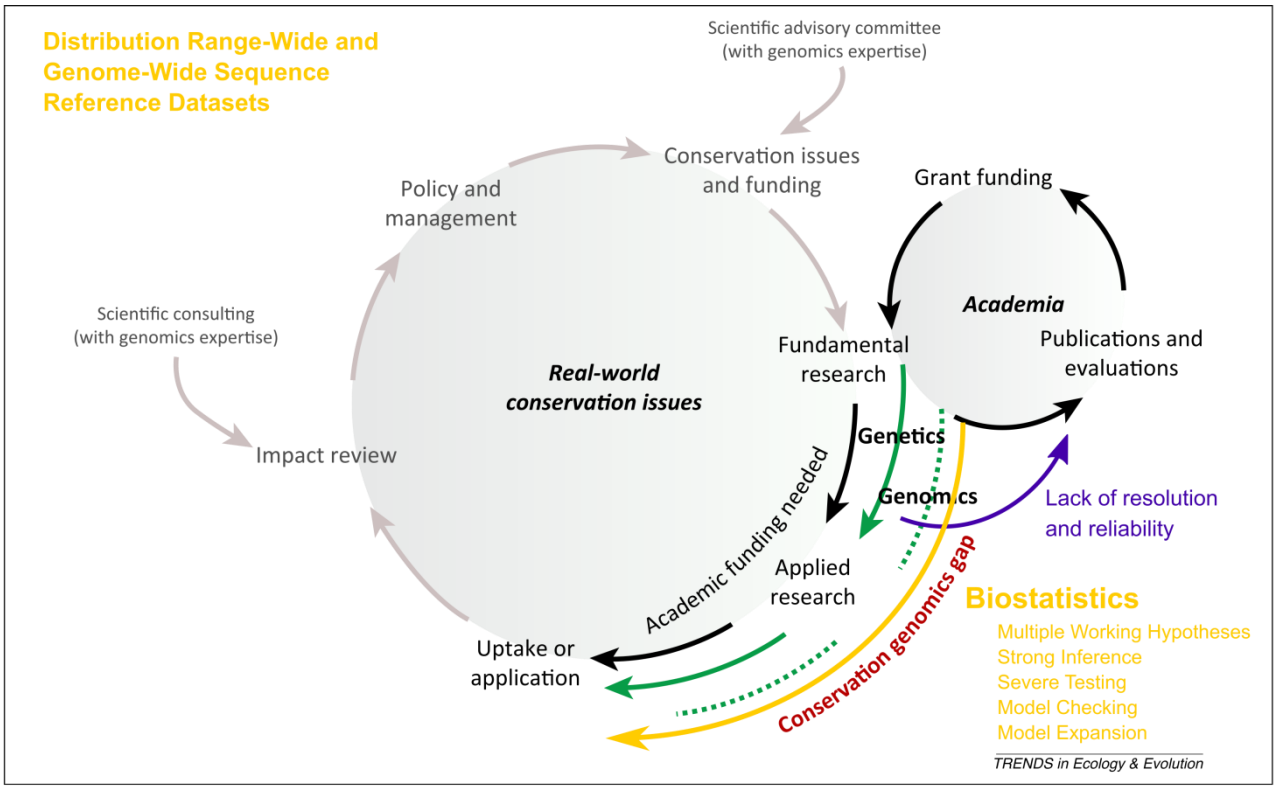


Figure 1. Schematic for the translation of conservation genomics research to conservation policy and action. Currently there are two largely separate spheres of applied and basic research. The black lines represent basic research to applied workflow, with the green lines reflecting the extent to which conservation genetics and genomics currently fit into this scheme – note that conservation genetics has integrated into the applied sphere. The broken green line represents the gap we see between the academic groundwork readily embracing genomic technology and on-site conservation needs. The gray lines are reflective of the larger framework we envision (and that is emerging) where conservation questions directly fund conservation genomics research and feed into management and biodiversity policies. Importantly, the applied component maintains a firm link to the conceptual advances driven by basic research.

Shafer et al.: Trends in Ecology & Evolution February 2015, Vol. 30, No. 2: 78-87 (yellow, purple: modifications J. Buschbom)

fig. 5.3: Effective transfer of basic biodiversity research to applied conservation genetic tools. The background graph and original figure legend were created by Shafer et al. (2015).

C CONCLUSIONS

Forensic conservation genetics aims to protect biodiversity by countering reliably and effectively the destruction and illegal exploitation of evolutionary lineages and natural populations. At the same time, in pursuing this goal, it protects and fosters legal and sustainable economic strategies, such as certification systems.

At the center of its progress towards determining and resolving the geographic origin of samples across all evolutionary scales, lies the versatility of well-developed, integrated and expansible reference datasets. They provide the flexibility to infer the whole spectrum of evolutionary processes that give rise to population diversity and structure. They are able to do so, by allowing the application of accurate and sufficient inference approaches that access, process and represent in depth the information recorded across the whole genome. Iterative in nature, this process provides the basis for the verification of the reference datasets themselves, the verification of models and inference approaches that are continuously expanded, developed and attuned, as well as, the accumulation of necessary evolutionary and biological background knowledge across the distribution range and ecological habitats. In this way, the empirical and statistical context is build that allows the evaluation of the validity and reliability of inference results in actual applied casework. It also creates and promotes expertise and skills in all parties involved, as well as, builds up the resources, the background information and the processes to respond quickly and reliably to the specifics of upcoming case scenarios. In this sense, forensic conservation genetics for the protection of wild flora and fauna becomes at its heart applied biodiversity genomics that is able to counter successfully the extinction of species and lineages, the loss of diversity and the destruction of complex and resilient ecosystems.

The task of building the logistic, collection and data management infrastructure, genomic and statistical resources, as well as, biodiversity knowledge and inference environments for a large number of non-model organisms across the diversity of the Tree of Life is huge. At the same time, the urgency is increasing for enforcing the protection of threatened wild flora and fauna, which are often located in regions facing social, economic and political pressures. In this situation, both in fisheries and forestry, concerted efforts exist, for example see Waples et al. (2008), Dormontt et al. (2015), Lowe et al. (2016), Bernatchez et al. (2017) and the Global Timber Tracking Network (GTTN2; http:www.globaltimbertrackingnetwork.org). They aim to reach out to all involved parties, to integrate their perspectives and expertise, to discuss next steps and new developments, to build resources and infrastructure, and to cooperatively find solutions to remaining obstacles towards the widespread implementation of accurate, effective and reliable forensic genetic tools.

Yet, all the described statistical efforts are still only a technological fix (Borgmann 2012). They will remain limited in their significance and without decisive impact and success, if humanity,

especially the individuals of its first world populations, will not change its cultures and lifestyles to enable a reduction of its ecological footprints to long-term sustainable resource-use levels.

References

Alacs EA, Georges A, FitzSimmons NN, Robertson J. 2010. DNA detective: a review of molecular approaches to wildlife forensics. *Forensic Sci. Med. Pathol.* 6 (3): 180-194

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9): 1655-1664

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. 2010. Data quality control in genetic case-control association studies. *Nat. Protoc.* 5 (9): 1564-1573

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17 (2): 81-92

Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA, Blom JG. 2009. Systems biology: parameter estimation for biochemical models. *FEBS J.* 276 (4): 886-902

Barber CV, Parker-Forney M. 2017. Rainforest CSI: How we're catching illegal loggers with DNA, machine vision and chemistry. In blog *Insights: WRI Blog*. World Resources Institute, Washington, DC USA. Retrieved from http://www.wri.org/blog/2017/04/rainforest-csi-how-were-catching-illegal-loggers-dna-machine-vision-and-chemistry (Accessed August 31, 2018)

Barnard GA, Jenkins GM, Winsten CB. 1962. Likelihood inference and time series. *J. R. Stat. Soc., Ser. A: Stat. Soc.* 125 (3): 321-372

Baudouin L, Piry S, Cornuet JM. 2004. Analytical Bayesian approach for assigning individuals to populations. *J. Hered.* 95 (3): 217-224

Beacham TD, Candy JR, McIntosh B, MacConnachie C, Tabata A, et al. 2005. Estimation of stock composition and individual identification of sockeye salmon on a Pacific Rim basis using microsatellite and major histocompatibility complex variation. *Trans. Am. Fish. Soc.* 134 (5): 1124-1146

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41 (1): 379-406

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* 5 (4): 251-261

Beerli P, Palczewski M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185 (1): 313-326

Bekkevold D, Helyar SJ, Limborg MT, Nielsen EE, Hemmer-Hansen J, et al. 2015. Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES J. Mar. Sci.* 72 (6): 1790-1801

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2000. GenBank. *Nucleic Acids Res.* 28 (1): 15-18

Bernatchez L, Wellenreuther M, Araneda C, Ashton DT, Barth JMI, et al. 2017. Harnessing the power of genomics to secure the future of seafood. *Trends Ecol. Evol.* 32 (9): 665-680

Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* 19 (13): 2609-2625

Birnbaum A. 1962. On the foundations of statistical inference (with discussion). *J. Am. Stat. Assoc.* 57 (298): 269-326

Bloomquist EW, Lemey P, Suchard MA. 2010. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.* 25 (11): 626-632

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19 (7): 1171-1180

Borgmann A. 2012. The setting of the scene: Technological fixes and the design of the good life. In *Engineering the Climate*, ed. C Preston, pp. 189-199. Lexington Books, Lanham, MD USA

Box GEP. 1976. Science and statistics. *J. Am. Stat. Assoc.* 71 (356): 791-799

Box GEP. 1979. Robustness in the strategy of scientific model building. *MRC Technical Summary Report #1954*. University of Wisconsin-Madison, Mathematics Research Center, Madison, WI USA

Browning SR, Browning BL. 2012. Identity by descent between distant relatives: detection and applications. *Annu. Rev. Genet.* 46: 617-633

Bruggeman FJ. 2009. Systems biology: from possible to plausible to actual models. *FEBS J.* 276 (4): 885

Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Population-specific F_{ST} values for forensic STR markers: A worldwide survey. *Forensic Sci. Int. Genet.* 23: 91-100

Busby GBJ, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, et al. 2015. The role of recent admixture in forming the contemporary West Eurasian genomic landscape. *Curr. Biol.* 25 (19): 2518-2526

Bylemans J, Maes GE, Diopere E, Cariani A, Senn H, et al. 2016. Evaluating genetic traceability methods for captive-bred marine fish and their applications in fisheries management and wildlife forensics. *Aquac. Environ. Interact.* 8: 131-145

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, et al. 2002. A human genome diversity cell line panel. *Science* 296 (5566): 261-262

Carstens BC, Brennan RS, Chua V, Duffie CV, Harvey MG, et al. 2013. Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Mol. Ecol.* 22 (15): 4014-4028

Carstens BC, Stoute HN, Reid NM. 2009. An information-theoretical approach to phylogeography. *Mol. Ecol.* 18 (20): 4270-4282

Caye K, Deist TM, Martins H, Michel O, Francois O. 2016. TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* 16 (2): 540-548

Cedersund G, Roll J. 2009. Systems biology: model based evaluation and comparison of potential explanations for given biological data. *FEBS J.* 276 (4): 903-922

Chamberlin TC. 1890. The method of multiple working hypotheses. *Science* 15: 92-96

Ciampolini R, Cetica V, Ciani E, Fosella X, Marroni F, et al. 2006. Statistical analysis of individual assignment tests among four cattle breeds using fifteen STR loci. *J. Anim. Sci.* 84: 11-19

CITES. 2015. Global forensic capacity to address illegal trafficking in wildlife. *Notification to the Parties No. 2015/061*. Convention on International Trade in Endangered Species of Wild Fauna and Flora, Geneva, CH

Collins RA, Hrbek T. 2018. An *in slico* comparison of protocols for dated phylogenomics. *Syst. Biol.* 67 (4): 633-650

Conomos MP, Miller MB, Thornton TA. 2015. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39 (4): 276-293

Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* 98 (1): 127-148

Corander J, Marttinen P, Siren J, Tang J. 2008a. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinform.* 9: 539

Corander J, Sirén J, Arjas E. 2008b. Bayesian spatial modeling of genetic population structure. *Comput. Stat.* 23 (1): 111-129

Cornuet J-M, Piry S, Luikart G, Estoup A, Solignac M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153: 1989-2000

Crawford JE, Lazzaro BP. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front. Genet.* 3: 66

Csilléry K, Blum MGB, Gaggiotti OE, François O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* 25 (7): 410-418

Cussens J, Sheehan NA. 2016. Editorial: Special issue on new developments in relatedness and relationship estimation. *Theor. Popul. Biol.* 107: 1-3

Cutter AD. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. *Mol. Phylogen. Evol.* 69 (3): 1172-1185

da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, et al. 2016. Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Mar. Genom.* 30: 3-13

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9 (8): 772

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS. 2007. Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Sci. Int.* 173 (1): 1-6

Dawson KJ, Belkhir K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* 78 (01): 59-77

De Maio N, Wu C-H, O'Reilly KM, Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genet.* 11 (8): e1005421

de Vries SMG, Murat A, Bozzano M, Burianek V, Collin E, et al. 2015. Pan-European strategy for genetic conservation of forest trees. European Forest Genetic Resources Programme (EUFORGEN). Bioversity International, Rome, Italy

Degen B, Höltken A, Rogge M. 2010. Use of DNA-fingerprints to control the origin of forest reproductive material. *Silvae Genet.* 59 (6): 268-273

Degen B, Ward SE, Lemes MR, Navarro C, Cavers S, Sebbenn AM. 2013. Verifying the geographic origin of mahogany (*Swietenia macrophylla* King) with DNA-fingerprints. *Forensic Sci. Int. Genet.* 7 (1): 55-62

Dormontt EE, Boner M, Braun B, Breulmann G, Degen B, et al. 2015. Forensic timber identification: It's time to integrate disciplines to combat illegal logging. *Biol. Conserv.* 191: 790-798

Downes BJ. 2010. Back to the future: little-used tools and principles of scientific inference can help disentangle effects of multiple stressors on freshwater ecosystems. *Freshw. Biol.* 55: 60-79

Dreifus C. 2016. A conversation with: Samuel K. Wasser, a scientific detective tailing poachers. *The New York Times*. Retrieved from https://www.nytimes.com/2016/06/14/science/samuel-wasser-dna-elephants-ivory.html?_r=0 (Accessed March 21, 2017)

Duchêne S, Duchêne DA, Di Giallonardo F, Eden J-S, Geoghegan JL, et al. 2016. Cross-validation to select Bayesian hierarchical models in phylogenetics. *BMC Evol. Biol.* 16: 115

Duforet-Frebourg N, Gattepaille LM, Blum MGB, Jakobsson M. 2015. HaploPOP: a software that improves population assignment by combining markers into haplotypes. *BMC Bioinform.* 16: 242

Eddy SR. 2004. What is a hidden Markov model? *Nat. Biotechnol.* 22 (10): 1315-1316

Edwards AWF. 1992. *Likelihood. Expanded edition*. Cambridge University Press, Cambridge, UK. 275 pp.

Edwards AWF. 2003. Human genetic diversity: Lewontin's fallacy. *BioEssays* 25 (8): 798-801

Efron B. 2012. Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.* 6 (4): 1971-1997

Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* 7 (9): 1026-1042

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29 (1): 51-63

Elliott LP, Brook BW. 2007. Revisiting Chamberlin: Multiple working hypotheses for the 21st century. *Bioscience* 57 (7): 608

Elliott LT. 2016. *Bayesian nonparametric models of genetic variation*. Dissertation thesis. University College London, London, UK. 134 pp.

Elliott LT, De Iorio M, Favaro S, Adhikari K, Teh YW. 2018. Modeling population structure under hierarchical Dirichlet processes. *Bayesian Analysis* Advance Online Publication, DOI: 10.1214/17-ba1093

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567-1587

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27 (4): 401-410

Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* 53 (4-5): 447-455

Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA USA. 664 pp.

Ferri G, Corradini B, Ferrari F, Santunione AL, Palazzoli F, Alu M. 2015. Forensic botany II, DNA barcode for land plants: Which markers after the international agreement? *Forensic Sci. Int. Genet.* 15: 131-136

Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond., Ser. A* 222: 309-368

Fisher RA. 1925. Theory of statistical estimation. *Proc. Camb. Philos. Soc.* 22: 700-725

Fisher RA. 1934. *Statistical Methods for Research Workers* (5th ed.). Oliver and Boyd LTD., Edinburgh, London, UK. 319 pp.

Fisher RA. 1935. The logic of inductive inference. *J. R. Stat. Soc.* 98 (1): 39-54

Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork VL. 2017. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section Quercus). *Genome* 60 (9): 743-755

François O, Durand E. 2010. Spatially explicit Bayesian clustering models in population genetics. *Mol. Ecol. Resour.* 10 (5): 773-784

Francois O, Waits LP. 2016. Clustering and assignment methods in landscape genetics. In *Landscape Genetics: Concepts, Methods, Applications*, ed. N Balkenhol, SA Cushman, AT Storfer, LP Waits, pp. 114-128. John Wiley & Sons Ltd., Chichester, UK

Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* 30 (10): 1486-1487

Gelman A. 2014. Discussion with Sander Greenland on posterior predictive checks. In blog *Statistical Modeling, Causal Inference, and Social Science*. Retrieved from https://andrewgelman.com/2014/08/11/discussion-sander-greenland-posterior-predictive-checks/ (Accessed August 31, 2018)

Gelman A, Carlin JS, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC, Boca Raton, FL USA. 667 pp.

Gelman A, Shalizi CR. 2013. Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66 (1): 8-38

Gelman AB, Carlin JS, Stern HS, Rubin DB. 1995. *Bayesian Data Analysis* (1st ed.). Chapman & Hall/CRC, Boca Raton, FL USA. 526 pp.

Gershman SJ, Blei DM. 2012. A tutorial on Bayesian nonparametric models. *J. Math. Psychol.* 56 (1): 1-12

Glover KA. 2010. Forensic identification of fish farm escapees: the Norwegian experience. *Aquac. Environ. Interact.* 1: 1-10

Glover KA, Pertoldi C, Besnier F, Wennevik V, Kent M, Skaala Ø. 2013. Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs. *BMC Genet.* 14: 74

Gompert Z, Buerkle CA. 2016. What, if anything, are hybrids: enduring truths and challenges associated with population structure and gene flow. *Evol. Appl.* 9 (7): 909-923

Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C. 2016. Clustering genes of common evolutionary history. *Mol. Biol. Evol.* 33 (6): 1590-1605

Graham RL, Foulds LR. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.* 60 (2): 133-142

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. 2011. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* 108 (29): 11983-11988

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82 (4): 711-732

Griffiths RC, Tavaré S. 1994. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46: 131-159

Grueber CE, Nakagawa S, Laws RJ, Jamieson IG. 2011. Multimodel inference in ecology and evolution: challenges and solutions. *J. Evol. Biol.* 24 (4): 699-711

Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Mol. Ecol.* 18 (23): 4734-4756

Guillot G, Renaud S, Ledevin R, Michaux J, Claude J. 2012. A unifying model for the analysis of phenotypic, genetic, and geographic data. *Syst. Biol.* 61 (6): 897-911

Guindon S, Guo H, Welch D. 2016. Demographic inference under the coalescent in a spatial continuum. *Theor. Popul. Biol.* 111: 43-50

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. 2015. The fickle *P* value generates irreproducible results. *Nat. Methods* 12 (3): 179-185

Hansen J, Sato M, Hearty P, Ruedy R, Kelley M, et al. 2016. Ice melt, sea level rise and superstorms: evidence from paleoclimate data, climate modeling, and modern observations that 2 °C global warming could be dangerous. *Atmos. Chem. Phys.* 16 (6): 3761-3812

Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. 2011. Statistical inference for stochastic simulation models - theory and application. *Ecol. Lett.* 14 (8): 816-827

Hartvig I, Czako M, Kjaer ED, Nielsen LR, Theilade I. 2015. The use of DNA barcoding in identification and conservation of rosewood (*Dalbergia* spp.). *PLoS One* 10 (9): e0138231

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT. 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Syst. Biol.* 65 (5): 910-924

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd (corrected 12th printing 2017) ed.). Springer-Verlag, New York, NY USA. 745 pp.

Hauser L, Seamons TR, Dauer M, Naish KA, Quinn TP. 2006. An empirical verification of population assignment methods by marking and parentage data: hatchery and wild steelhead (*Oncorhynchus mykiss*) in Forks Creek, Washington, USA. *Mol. Ecol.* 15 (11): 3157-3173

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond., Ser. B: Biol. Sci.* 270 (1512): 313-321

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27 (3): 570-580

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, et al. 2014. A genetic atlas of human admixture history. *Science* 343 (6172): 747-751

Hey J, Machado CA. 2003. The study of structured populations - new hope for a difficult and divided science. *Nat. Rev. Genet.* 4 (7): 535-543

Hickerson MJ. 2014. All models are wrong. *Mol. Ecol.* 23: 2887-2889

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. *Syst. Biol.* 44 (1): 3-16

Hillis DM, Huelsenbeck JP. 1994. To tree the truth: biological and numerical simulations in phylogeny. In *Molecular Evolution of Physiological Processes*, ed. DM Fambrough, pp. 55-67. The Rockefeller University Press, New York, NY USA

Hillis DM, Moritz C, Mable BK. 1996. *Molecular Systematics* (2nd ed.). Sinauer Associates, Sunderland, MA USA. 655 pp.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS One* 9 (4): e93975

Hoelzel AR. 2015. Can DNA foil the poachers? Forensic data help to identify elephant poaching hotspots. *Science* 349 (6243): 34-35

Hudson RR. 1990. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, ed. DJ Futuyma, J Antonovics, pp. 1-43. Oxford University Press, Oxford, UK

Huelsenbeck JP. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44 (1): 17-48

Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. *Genetics* 175: 1787-1802

Ibsen-Jensen R, Chatterjee K, Nowak MA. 2015. Computational complexity of ecological and evolutionary spatial dynamics. *Proc. Natl. Acad. Sci. USA* 112 (51): 15636-15641

International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467 (7311): 52-58

IPCC. 2014. Climate change 2014: Synthesis report. *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Core Writing Team, RK Pachauri, LA Meyer. Intergovernmental Panel on Climate Change, Geneva, CH

Ishida Y. 2009. Sewall Wright and Gustave Malécot on isolation by distance. *Philos. Sci.* 76 (5): 784-796

Iyengar A. 2014. Forensic DNA analysis for animal protection and biodiversity conservation: A review. *J. Nat. Conserv.* 22 (3): 195-205

Jaqaman K, Danuser G. 2006. Linking data to models: data regression. *Nat. Rev. Mol. Cell Biol.* 7 (11): 813-819

Ji H, Liu XS. 2010. Analyzing 'omics data using hierarchical models. *Nat. Biotechnol.* 28 (4): 337-340

Johnson JB, Omland KS. 2004. Model selection in ecology and evolution. *Trends Ecol. Evol.* 19 (2): 101-108

Johnson RN, Wilson-Wilde L, Linacre A. 2014. Current and future directions of DNA in wildlife forensic science. *Forensic Sci. Int. Genet.* 10: 1-11

Jolivet C, Degen B. 2012. Use of DNA fingerprints to control the origin of sapelli timber (*Entandrophragma cylindricum*) at the forest concession level in Cameroon. *Forensic Sci. Int. Genet.* 6 (4): 487-493

Jombart T, Pontier D, Dufour AB. 2009. Genetic markers in the playground of multivariate analysis. *Heredity* 102 (4): 330-341

Joseph TA, Hickerson MJ, Alvarado-Serrano DF. 2016. Demographic inference under a spatially continuous coalescent model. *Heredity* 117: 94-99

Kelleher JD, Mac Namee B, D'Arcy A. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies*. The MIT Press, Cambridge, MA USA. 595 pp.

Kingman JFC. 1982. The coalescent. *Stoch. Proc. Appl.* 13: 235-248

Konnert M, Fady B, Gömöry D, A'Hara S, Wolter F, et al. 2015. Use and transfer of forest reproductive material in Europe in the context of climate change. European Forest Genetic Resources Programme (EUFORGEN). Bioversity International, Rome, Italy

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of next generation sequencing data. *BMC Bioinform.* 15: 356

Kreutz C, Timmer J. 2009. Systems biology: experimental design. *FEBS J.* 276 (4): 923-942

Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140: 429-434

Landis MJ, Matzke NJ, Moore BR, Huelsenbeck JP. 2013. Bayesian analysis of biogeography when the number of areas is large. *Syst. Biol.* 62 (6): 789-804

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE. 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conserv. Genet.* 7 (2): 295-302

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, et al. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34 (6): 591-602

Lawson DJ, Falush D. 2012. Population identification using genetic data. *Annu. Rev. Genom. Hum. Genet.* 13: 337-361

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet.* 8 (1): e1002453

Leigh JW, Lapointe FJ, Lopez P, Bapteste E. 2011. Evaluating phylogenetic congruence in the post-genomic era. *Genome Biol. Evol.* 3: 571-587

Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, et al. 2009. Species relative abundance and direction of introgression in oaks. *Mol. Ecol.* 18 (10): 2228-2242

Lever J, Krzywinski M, Altman N. 2017. Points of Significance: Principal component analysis. *Nat. Methods* 14 (7): 641-642

Lew MJ. 2012. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know *P*. *Br. J. Pharmacol.* 166 (5): 1559-1567

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100-1104

Limborg MT, Helyar SJ, De Bruyn M, Taylor MI, Nielsen EE, et al. 2012. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Mol. Ecol.* 21: 3686-3703

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, et al. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193 (4): 1233-1254

Lowe AJ, Dormontt EE, Bowie MJ, Degen B, Gardner S, et al. 2016. Opportunities for improved transparency in the timber trade through scientific verification. *Bioscience* 66 (11): 990-998

Lynsky A. 2017. The Stupidity Paradox: The Power and Pitfalls of Functional Stupidity at Work (Book review). *The Actuary*. Retrieved from http://www.theactuary.com/opinion/2017/03/the-stupidity-paradox-the-power-and-pitfalls-of-functional-stupidity-at-work/ (Accessed July 16, 2018)

Malécot G. 1973. Génetique des populations diploïdes naturelles dans le cas d'un seul locus. III. Parenté, mutations et migration. *Ann. Genet. Sel. Anim.* 5 (3): 333-361

Malinsky M, Trucchi E, Lawson DJ, Falush D. 2018. RADpainter and fineRADstructure: Population inference from RADseq data. *Mol. Biol. Evol.* 35 (5): 1284-1290

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538 (7624): 201-206

Mallo D, Posada D. 2016. Multilocus inference of species trees and DNA barcoding. *Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci.* 371 (1702): 20150335

Manel S, Berthier P, Luikart G. 2002. Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conserv. Biol.* 16 (3): 650-659

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26 (22): 2867-2873

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* 44 (3): 243-246

Mayo DG. 1996. *Error and the Growth of Experimental Knowledge*. Chicago University Press, Chicago, IL USA. 509 pp.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogen. Evol.* 66 (2): 526-538

McVean GAT. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.* 5 (10): e1000686

Meirmans PG. 2012. The trouble with isolation by distance. *Mol. Ecol.* 21: 2839-2846

Michalewicz Z, Fogel DB. 2004. *How to Solve It: Modern Heuristics* (2nd ed.). Springer, Berlin, D. 554 pp.

Mignone A, Howlett M. 2012. From paper trails to DNA barcodes: enhancing traceability in forest and fishery certification. *Nat. Resour. J.* 52: 421-441

Millar H. 2016. How forensics are boosting the battle against wildlife trade. *Yale Environ. 360*. Retrieved from http://e360.yale.edu/feature/wildlife_forensics_boosting_fight_against_illegal_trade_poaching/3051/ (Accessed March 21, 2017)

Miller JW, Harrison MT. 2018. Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.* 113 (521): 340-356

Mimno D, Blei DM, Engelhardt BE. 2015. Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proc. Natl. Acad. Sci. USA* 112 (26): E3441-3450

Moltke I, Albrechtsen A. 2014. RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics* 30 (7): 1027-1028

Mondol S, Moltke I, Hart J, Keigwin M, Brown L, et al. 2015. New evidence for hybrid zones of forest and savanna elephants in Central and West Africa. *Mol. Ecol.* 24 (24): 6134-6147

Moore BR, McGuire J, Ronquist F, Huelsenbeck JP. 2014. Bayesian analysis of partitioned data. *ArXiv preprint* arXiv:1409.0906v1 [q-bio.PE]

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, et al. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* 7 (4): e1001373

Motulsky H. 2010. *Intuitive Biostatistics. A Nonmathematical Guide to Statistical Thinking* (2nd ed.). Oxford University Press, New York, NY USA. 447 pp.

Neale CA. 2012. *Quantifying and enhancing the statistical convergence of equilibrium and non-equilibrium properties in molecular simulations of proteins, peptides, and amino acid side chain analogs embedded in lipid bilayers*. Dissertation thesis. University of Toronto, Toronto, CA. 354 pp.

Neophytou C. 2014. Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. *Tree Genet. Genomes* 10: 273-285

Neophytou C, Gärtner SM, Vargas-Gaete R, Michiels H-G. 2015. Genetic variation of Central European oaks: shaped by evolutionary factors and human intervention? *Tree Genet. Genomes* 11

Nevado B, Ramos-Onsins SE, Perez-Enciso M. 2014. Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Mol. Ecol.* 23 (7): 1764-1779

Neyman J, Pearson EE. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond., Ser. A* 231 (694-706): 289-337

Ng CH, Lee SL, Tnah LH, Ng KKS, Lee CT, et al. 2017. Geographic origin and individual assignment of *Shorea platyclados* (Dipterocarpaceae) for forensic identification. *PLoS One* 12 (4): e0176158

Nielsen CB. 2016. Visualization: A mind-machine interface for discovery. *Trends Genet.* 32 (2): 73-75

Nielsen EE, Cariani A, Mac Aoidh E, Maes GE, Milano I, et al. 2012a. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nat. Commun.* 3: 851

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol. Ecol.* 18 (6): 1034-1047

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012b. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS One* 7 (7): e37558

Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* 29: 59-75

Novembre J, Peter BM. 2016. Recent advances in the study of fine-scale population structure in humans. *Curr. Opin. Genet. Dev.* 41: 98-105

Novembre J, Slatkin M. 2009. Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* 63 (11): 2914-2925

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* 40 (5): 646-649

Nowakowska JA. 2011. Application of DNA markers against illegal logging as a new tool for the Forest Guard Service. *Folia For. Pol. Ser. A For.* 53 (2): 142-149

O'Rawe JA, Ferson S, Lyon GJ. 2015. Accounting for uncertainty in DNA sequencing data. *Trends Genet.* 31 (2): 61-66

Ogden R. 2008. Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish Fish.* 9: 462-472

Ogden R, Dawnay N, McEwing R. 2009. Wildlife DNA forensics - bridging the gap between conservation genetics and law enforcement. *Endanger. Species Res.* 9: 179-195

Ogden R, Linacre A. 2015. Wildlife forensic science: A review of genetic geographic origin assignment. *Forensic Sci. Int. Genet.* 18: 152-159

Ogden R, Mailley J, The Society for Wildlife Forensic Science. 2016. A review of wildlife forensic science and laboratory capacity to support the implementation and enforcement of CITES. *CoP17 Doc. 25 Annex 4*. commissioned by the Secretariat of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), United Nations Office on Drug and Crime (UNODC, Vienna), United Nations, New York, NY USA

Onogi A, Nurimoto M, Morita M. 2011. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinform.* 12: 263

Paetkau D, Calvert W, Stirling I, Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* 4: 347-354

Paetkau D, Slade R, Burden M, Estoup A. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol. Ecol.* 13: 55-65

Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics* 147: 1945-1957

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2 (12): e190

Pella J, Masuda M. 2006. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Can. J. Fish. Aquat. Sci.* 63 (3): 576-596

Pelletier TA, Carstens BC. 2014. Model choice for phylogeographic inference using a large set of models. *Mol. Ecol.* 23 (12): 3028-3043

Penny D, Hendy MD, Steel MA. 1992. Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol.* 7 (3): 73-79

Petit RJ, Csaikl UM, Bordacs S, Burg K, Coart E, et al. 2002. Chloroplast DNA variation in European white oaks - Phylogeography and patterns of diversity based on data from over 2600 populations. *For. Ecol. Manage.* 156: 5-26

Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. 2004. GENECLASS2: a software for genetic assignment and first-generation migrant detection. *J. Hered.* 95 (6): 536-539

Platt JR. 1964. Strong inference. *Science* 146 (3642): 347-353

Poe S. 1998. The effect of taxonomic sampling on accuracy of phylogeny estimation: test case of a known phylogeny. *Mol. Biol. Evol.* 15 (8): 1086-1090

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14 (9): 817-818

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* 11 (11): 800-805

Praetorius SK. 2018. North Atlantic circulation slows down. *Nature* 556: 180-181

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8): 904-909

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197 (2): 573-589

Rannala B. 2015. The art and science of species delimitation. *Curr. Zool.* 61 (5): 846-853

Rannala B, Mountain JL. 1997. Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci. USA* 94: 9197-9201

Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* 7 (3): 355-364

Reich BJ, Bondell HD. 2011. A spatial Dirichlet process mixture model for clustering population genetics data. *Biometrics* 67 (2): 381-390

Rohlfs RV, Fullerton SM, Weir BS. 2012. Familial identification: population structure and relationship distinguishability. *PLoS Genet.* 8 (2): e1002469

Romanycia MHJ, Pelletier FJ. 1985. What is a heuristic? *Comput. Intell.* 1 (1): 47-58

Romeijn J-W. 2017. Philosophy of statistics. In *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Metaphysics Research Lab, Stanford University, Stanford, CA USA

Rosen J. 2016. A forest of hypotheses. *Nature* 536: 239-241

Rousset F. 2002. Inbreeding and relatedness coefficients: what do they measure? *Heredity* 88: 371-380

Safner T, Miller MP, McRae BH, Fortin M-J, Manel S. 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *Int. J. Mol. Sci.* 12 (2): 865-889

Schneider DM, Martins AB, de Aguiar MAM. 2016. The mutation-drift balance in spatially structured populations. *J. Theor. Biol.* 402: 9-17

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16 (12): 727-740

Schroeder H, Cronn R, Yanbaev Y, Jennings T, Mader M, et al. 2016. Development of molecular markers for determining continental origin of wood from white oaks (*Quercus* L. sect. Quercus). *PLoS One* 11 (6): e0158221

Shafer ABA, Wolf JBW, Alves PC, Bergstrom L, Bruford MW, et al. 2015. Genomics and the challenging translation into conservation practice. *Trends Ecol. Evol.* 30 (2): 78-87

Sham PC, Purcell SM. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* 15 (5): 335-346

Skotte L, Korneliussen TS, Albrechtsen A. 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195 (3): 693-702

Sokal RR, Rohlf FJ. 1995. *Biometry* (3rd ed.). W. H. Freeman and Company, New York, NY USA. 887 pp.

Speed D, Balding DJ. 2015. Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* 16 (1): 33-44

Srinath S, Gunawan R. 2010. Parameter identifiability of power-law biochemical system models. *J. Biotechnol.* 149 (3): 132-140

Steel MA, Hendy MD, Penny D. 1988. Loss of information in genetic distances. *Nature* 336: 118

Stephens M, Balding DJ. 2009. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10 (10): 681-690

Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, Pevsner J. 2011. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet.* 7 (9): e1002287

Stone GN, Nee S, Felsenstein J. 2011. Controlling for non-independence in comparative analysis of patterns across populations within species. *Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci.* 366 (1569): 1410-1424

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526 (7571): 75-81

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36 (1): 445-466

Sun M, Jobling MA, Taliun D, Pramstaller PP, Egeland T, Sheehan NA. 2016. On the use of dense SNP marker data for the identification of distant relative pairs. *Theor. Popul. Biol.* 107: 14-25

Sunnaker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. 2013. Approximate Bayesian computation. *PLoS Comp. Biol.* 9 (1): e1002803

SWGWILD. 2012. SWGWILD standards and guidelines. *Standard Version 2.0 - Accepted by SWGWILD December 19, 2012*. Scientific Working Group for Wildlife Forensic Science. Society for Wildlife Forensic Science

SWGWILD. 2015. Standards and guidelines for forensic botany identification. *Standard*. Scientific Working Group for Wildlife Forensic Science. Society for Wildlife Forensic Science

Symes LB, Serrell N, Ayres MP. 2015. A practical guide for mentoring scientific inquiry. *Bull. Ecol. Soc. Am.* 96 (2): 352-367

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* 79: 1-12

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* 28 (4): 289-301

Teh YW, Jordan MI, Beal MJ, Blei DM. 2006. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101 (476): 1566-1581

Tereba A, Woodward S, Konecka A, Borys M, Nowakowska JA. 2017. Analysis of DNA profiles of ash (*Fraxinus excelsior* L.) to provide evidence of illegal logging. *Wood Sci. Technol.* 51 (6): 1377-1387

The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, et al. 2015. A global reference for human genetic variation. *Nature* 526 (7571): 68-74

Thompson EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* 194 (2): 301-326

Thornton T, Conomos MP, Sverdlov S, Blue EM, Cheung CYK, et al. 2014. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. *BMC Proc.* 8 (Suppl 1): S5

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. 2012. Estimating kinship in admixed populations. *Am. J. Hum. Genet.* 91 (1): 122-138

Tnah LH, Lee SL, Ng KKS, Faridah Q-Z, Faridah-Hanum I. 2010. Forensic DNA profiling of tropical timber species in Peninsular Malaysia. *For. Ecol. Manage.* 259 (8): 1436-1446

Tnah LH, Lee SL, Ng KKS, Tani N, Bhassu S, Othman RY. 2009. Geographical traceability of an important tropical timber (*Neobalanocarpus heimii*) inferred from chloroplast DNA. *For. Ecol. Manage.* 258 (9): 1918-1923

Tripathi AM, Tyagi A, Kumar A, Singh A, Singh S, et al. 2013. The internal transcribed spacer (ITS) region and trnH-psbA are suitable candidate loci for DNA barcoding of tropical tree species of India. *PLoS One* 8 (2): e57934

UNODC. 2012. Wildlife and forest crime analytic toolkit - revised edition. United Nations Office on Drug and Crime (Vienna), United Nations, New York, NY USA

UNODC. 2014. Guidelines on methods and procedures for ivory sampling and laboratory analysis. United Nations Office on Drug and Crime (Vienna), United Nations, New York, NY USA

UNODC. 2016a. Best practice guide for forensic timber identification. United Nations Office on Drug and Crime (Vienna), United Nations, New York, NY USA

UNODC. 2016b. World wildlife crime report: Trafficking in protected species. United Nations Office on Drug and Crime (Vienna), United Nations, New York, NY USA

Van Doornik DM, Teel DJ, Kuligowski DR, Morgan CA, Casillas E. 2007. Genetic analyses provide insight into the early ocean stock distribution and survival of juvenile coho salmon off the coasts of Washington and Oregon. *N. Am. J. Fish. Manage.* 27 (1): 220-237

Vesterstrøm J. 2005. *Heuristic algorithms in bioinformatics*. Dissertation thesis. University of Aarhus, Aarhus, DK. 228 pp.

Vieira FG, Albrechtsen A, Nielsen R. 2016. Estimating IBD tracts from low coverage NGS data. *Bioinformatics* 32 (14): 2096-2102

Wang J. 2014. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. *J. Evol. Biol.* 27 (3): 518-530

Wang J. 2016. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.* 107: 4-13

Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics* 202: 1185-1200

Waples RS, Punt AE, Cope JM. 2008. Integrating genetic data into management of marine resources: how can we do it better? *Fish Fish.* 9: 423-449

Wasser SK, Mailand C, Booth R, Mutayoba B, Kisamo E, et al. 2007. Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proc. Natl. Acad. Sci. USA* 104 (10): 4228-4233

Wasser SK, Shedlock AM, Comstock K, Ostrander EA, Mutayoba B, Stephens M. 2004. Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proc. Natl. Acad. Sci. USA* 101 (41): 14847-14852

Weir BS, Anderson AD, Hepler AB. 2006. Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7 (10): 771-780

Weir BS, Zheng X. 2015. SNPs and SNVs in forensic science. *Forensic Sci. Int. Genet. Suppl. Ser.* 5: e267-e268

Wilkinson S, Archibald A, Haley CS, Megens H-J, Crooijmans RP, et al. 2012. Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genom.* 13: 580

Wilson-Wilde L. 2010. Wildlife crime: a global problem. *Forensic Sci. Med. Pathol.* 6 (3): 221-222

Wilson G, Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163: 1177-1191

Wilton PR, Baduel P, Landon MM, Wakeley J. 2017. Population structure and coalescence in pedigrees: Comparisons to the structured coalescent and a framework for inference. *Theor. Popul. Biol.* 115: 1-12

Withler RE, Candy JR, Beacham TD, Miller KM. 2004. Forensic DNA analysis of Pacific salmonid samples for species and stock identification. *Environ. Biol. Fishes* 69 (1): 275-285

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97-159

Wright S. 1943. Isolation by distance. *Genetics* 28: 114-138

Yang Z, Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* 31 (12): 3125-3135

Zheng X, Weir BS. 2016. Eigenanalysis of SNP data with an identity by descent interpretation. *Theor. Popul. Biol.* 107: 65-76

Zhou Y, Rodrigue N, Lartillot N, Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol. Biol.* 7: 206

Thünen Report

Bereits in dieser Reihe erschienene Hefte – *Volumes already published in this series*

1 - 39	siehe http://www.thuenen.de/de/infothek/publikationen/thuenen-report/
40	Frank Offermann, Martin Banse, Claus Deblitz, Alexander Gocht, Aida Gonzalez-Mellado, Peter Kreins, Sandra Marquardt, Bernhard Osterburg, Janine Pelikan, Claus Rösemann, Petra Salamon, Jürn Sanders **Thünen-Baseline 2015 – 2025: Agrarökonomische Projektionen für Deutschland**
41	Stefan Kundolf, Patrick Küpper, Anne Margarian und Christian Wandinger **Koordination, Lernen und Innovation zur Entwicklung peripherer ländlicher Regionen** **Phase II der Begleitforschung zum Modellvorhaben *LandZukunft***
42	Sebastian Rüter, Frank Werner, Nicklas Forsell, Christopher Prins, Estelle Vial, Anne-Laure Levet **ClimWood2030 'Climate benefits of material substitution by forest biomass and harvested wood products: Perspective 2030' Final Report**
43	Nicole Wellbrock, Andreas Bolte, Heinz Flessa (eds) **Dynamik und räumliche Muster forstlicher Standorte in Deutschland – Ergebnisse der Boden-zustandserhebung im Wald 2006 bis 2008**
44	Walter Dirksmeyer, Michael Schulte und Ludwig Theuvsen (eds) **Aktuelle Forschung in der Gartenbauökonomie – Nachhaltigkeit und Regionalität – Chancen und Heraus-forderungen für den Gartenbau** – Tagungsband zum 2. Symposium für Ökonomie im Gartenbau
45	Mirko Liesebach (ed) **Forstgenetik und Naturschutz – 5. Tagung der Sektion Forstgenetik/Forstpflanzenzüchtung am 15./16. Juni 2016 in Chorin – Tagungsband**
46	Claus Rösemann, Hans-Dieter Haenel, Ulrich Dämmgen, Annette Freibauer, Ulrike Döring, Sebastian Wulf, Brigitte Eurich-Menden, Helmut Döhler, Carsten Schreiner, Bernhard Osterburg **Calculations of gaseous and particulate emissions from German agriculture 1990 - 2015** **Berechnung von gas- und partikelförmigen Emissionen aus der deutschen Landwirtschaft 1990 – 2015**
47	Niko Sähn, Stefan Reiser, Reinhold Hanel und Ulfert Focken **Verfügbarkeit umweltrelevanter Daten zur deutschen Süßwasseraquakultur**
48	Markus Ehrmann **Modellgestützte Analyse von Einkommens- und Umweltwirkungen auf Basis von Testbetriebsdaten**
49	Mirko Liesebach, Wolfgang Ahrenhövel, Alwin Janßen, Manuel Karopka, Hans-Martin Rau, Bernd Rose, Randolf Schirmer, Dagmar Schneck, Volker Schneck, Wilfried Steiner, Silvio Schüler, Heino Wolf **Planung, Anlage und Betreuung von Versuchsflächen der Forstpflanzenzüchtung** **Handbuch für die Versuchsanstellung**
50	Tobias Mettenberger **Jugendliche Zukunftsorientierungen in ländlichen Mittelstädten** Zur Rolle des alltäglichen (sozial-)räumlichen Kontexts beim Übergang von der Hauptschule in den weiteren Ausbildungsweg
51	Stefan Neumeier **Modellvorhaben chance.natur – Endbericht der Begleitforschung –**

52	Andreas Tietz **Überregional aktive Kapitaleigentümer in ostdeutschen Agrarunternehmen: Entwicklungen bis 2017**
53	Peter Mehl (ed) **Aufnahme und Integration von Geflüchteten in ländliche Räume:** **Spezifika und (Forschungs-)herausforderungen** Beiträge und Ergebnisse eines Workshops am 6. und 7. März 2017 in Braunschweig
54	G. Rahmann, C. Andres, A.K. Yadav, R. Ardakani, H.B. Babalad, N. Devakumar, S.L. Goel, V. Olowe, N. Ravisankar, J.P. Saini, G. Soto, H. Willer **Innovative Research for Organic 3.0 - Volume 1** Proceedings of the Scientific Track at the Organic World Congress 2017 November 9-11 in Delhi, India
54	G. Rahmann, C. Andres, A.K. Yadav, R. Ardakani, H.B. Babalad, N. Devakumar, S.L. Goel, V. Olowe, N. Ravisankar, J.P. Saini, G. Soto, H. Willer **Innovative Research for Organic 3.0 - Volume 2** Proceedings of the Scientific Track at the Organic World Congress 2017 November 9-11 in Delhi, India
55	Anne Margarian unter Mitarbeit von Matthias Lankau und Alena Lilje **Strategien kleiner und mittlerer Betriebe in angespannten Arbeitsmarktlagen** Eine Untersuchung am Beispiel der niedersächsischen Ernährungswirtschaft
56	Frank Offermann, Martin Banse, Florian Freund, Marlen Haß, Peter Kreins, Verena Laquai, Bernhard Osterburg, Janine Pelikan, Claus Rösemann, Petra Salamon **Thünen-Baseline 2017 – 2027: Agrarökonomische Projektionen für Deutschland**
57	Hans-Dieter Haenel, Claus Rösemann, Ulrich Dämmgen, Ulrike Döring, Sebastian Wulf, Brigitte Eurich-Menden, Annette Freibauer, Helmut Döhler, Carsten Schreiner, Bernhard Osterburg **Calculations of gaseous and particulate emissions from German agriculture 1990 - 2016** **Berechnung von gas- und partikelförmigen Emissionen aus der deutschen Landwirtschaft 1990 – 2016**
58	Anja-Kristina Techen **Reduzierung von landwirtschaftlichen Stickstoffeinträgen in Gewässer: die Wirksamkeit von Beratung am Beispiel der hessischen WRRL-Beratung**
59	Katja Oehmichen, Susann Klatt, Kristin Gerber, Heino Polley, Steffi Röhling, Karsten Dunger **Die alternativen WEHAM-Szenarien: Holzpräferenz, Naturschutzpräferenz und Trendfortschreibung** **Szenarienentwicklung, Ergebnisse und Analyse**
60	Anne Margarian **Strukturwandel in der Wissensökonomie: Eine Analyse von Branchen-, Lage- und Regionseffekten in Deutschland**
61	Meike Hellmich **Nachhaltiges Landmanagement vor dem Hintergrund des Klimawandels als Aufgabe der räumlichen Planung - Eine Evaluation im planerischen Mehrebenensystem an den Beispielen der Altmark und des Landkreises Lüchow-Dannenbergs -**
62	Bernd Degen, Konstantin V. Krutovsky, Mirko Liesebach (eds.) **German Russian Conference on Forest Genetics - Proceedings - Ahrensburg, 2017 November 21-23**
63	Jutta Buschbom **Exploring and validating statistical reliability in forensic conservation genetics**

THÜNEN

Thünen Report 63

Herausgeber/Redaktionsanschrift

Johann Heinrich von Thünen-Institut
Bundesallee 50
38116 Braunschweig
Germany

www.thuenen.de