What genetic data can tell us about pathogens, species and their evolution

Dr Tanja Stadler

Ecole polytechnique fédérale de Zurich

How do infectious diseases spread in populations? How do the invasion, spread and persistence of infectious diseases depend on the host population structure? How do externally induced changes (such as environmental changes or disease control) affect the spread of infectious diseases?

Which factors determine the speciation and extinction dynamics of a species? To what extent is competition among species for available resources shaping biodiversity? How can we explain the predominance of sexually (compared to asexually) reproducing species despite the two-fold-cost of sex?

These are examples of current key questions in evolution, ecology and epidemiology, both on the slow macroscopic level (species) as well as on the fast microscopic level (infectious pathogens). Answering these questions will have timely implications towards understanding the changes in biodiversity due to current climate change, towards implementing successful public health intervention strategies for emerging epidemics, and towards designing effective vaccines.

My main research interests are to assess these questions by performing *phylogenetic* analyses of genetic sequencing data.

What is phylogenetics?

Evolutionary processes fundamentally shape biological systems. New species evolve from existing species, while others go extinct. Similarly, pathogens such as bacteria and viruses spread in host species populations and disappear again. Thus, the biological world is a highly dynamic system, constantly changing through evolution.

Charles Darwin's influential book, "On the origin of species" [2], where he introduces the concept of evolution by natural selection, contains only one picture, namely one of a phylogenetic tree. A phylogenetic tree represents the evolutionary relationship between species, and thus is central to evolutionary studies. Figure 1 shows Darwin's tree, sketched more than 150 years ago. Today, using large biological datasets, together with sophisticated mathematical and computational tools, we obtain ever more detailed phylogenetic trees. Figure 2 shows a phylogenetic tree of mammalian families, which we reconstructed in 2011. Similar organisms are clustered together in the phylogenetic tree. Each tip in the tree represents a mammalian family, branching points represent

speciation events and a time scale allows us to read off the times of speciation. Based on this tree, we can for example directly read off that placental mammals emerged almost 200 million years ago.

For more than a century the field was data poor, as only morphological (form and structure) and paleontological data could be used to reconstruct phylogenies. The revolution in sequencing technologies led to a revolution in phylogenetics during the last decade. The flood of molecular data allowed the reconstruction of past phylogenetic trees, along with speciation and extinction dynamics, in never before seen detail. As an example, sequence data together with sophisticated mathematical advances finally allowed us to conclude which species our closest relative is: chimpanzees are more closely related to humans than to any other living species [4,6].

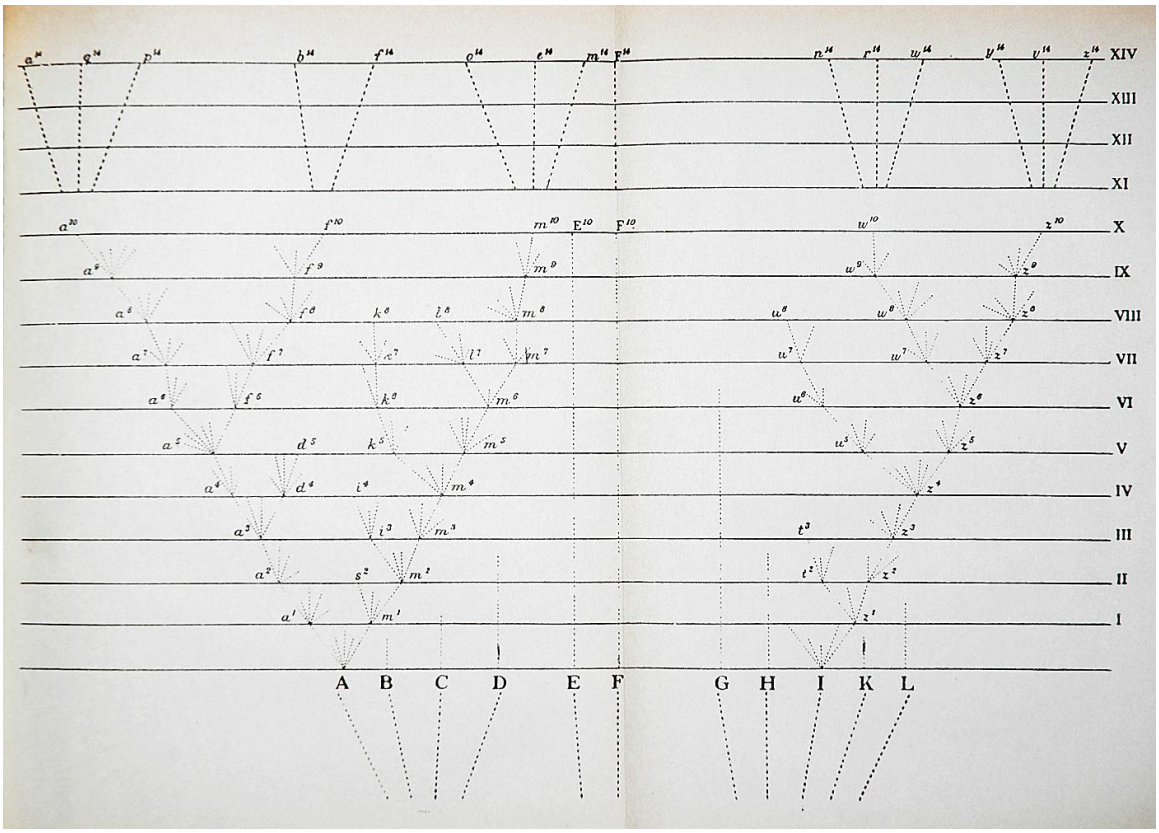

Figure 1: The only figure in Charles Darwin's "On the origin of species" [2]: a phylogenetic tree.

The field of phylogenetics is undergoing an important change over recent time. For a long time phylogenetics focused on describing the species tree as accuratly as possible. The increasing amount of sequence data offers the

unique opportunity to use phylogenetic methods for not only describing the past, i.e. infer the phylogenetic trees, but also inferring the underlying dynamics. Thus more recently there is a shift in the field of phylogenetics towards a functional inference of the evolutionary process itself, i.e. researchers attempt to infer what processes are most likely responsible for the particular pattern of speciation observed in a given phylogeny.

How I got into phylogenetics and where I am now

My background is in mathematics. In the last year of my Master's degree I happened to read a book with the simple title: "Phylogenetics". I became so fascinated as to how mathematics can be used in this field that I decided to move to New Zealand to work with one of the authors, Mike Steel, on my Master's and PhD theses.

Today I lead the Computational Evolution group at the Department of Biosystems Science and Engineering at the ETH Zürich. The Computational Evolution group develops phylogenetic methods in order to understand evolutionary processes. Using our phylogenetic methods, we address questions in a number of fields, focusing on macroevolution (i.e. the classical application of phylogenetics), as well as epidemiology, public health & medicine, and language evolution. In our daily work, we define and analyze stochastic models, implement computational methods, analyze empirical data, and discuss our insights with clinicians and public health policy makers.

In the area of macroevolution, we use genetic sequencing data together with morphological data, fossil occurrences and ecological information in order to determine the main drivers of macroevolution. In order to determine the evolution and spread of pathogens, we combine host and pathogen genetic data with clinical data. With these insights, the goal is to predict the future fate of a pathogen, allowing us to suggest new public health interventions. Insight into human language evolution is obtained through phylogenetic methods by interpreting the spelling of words as the genetic code. In the next two sections I will present some of our findings in these fields in more detail.

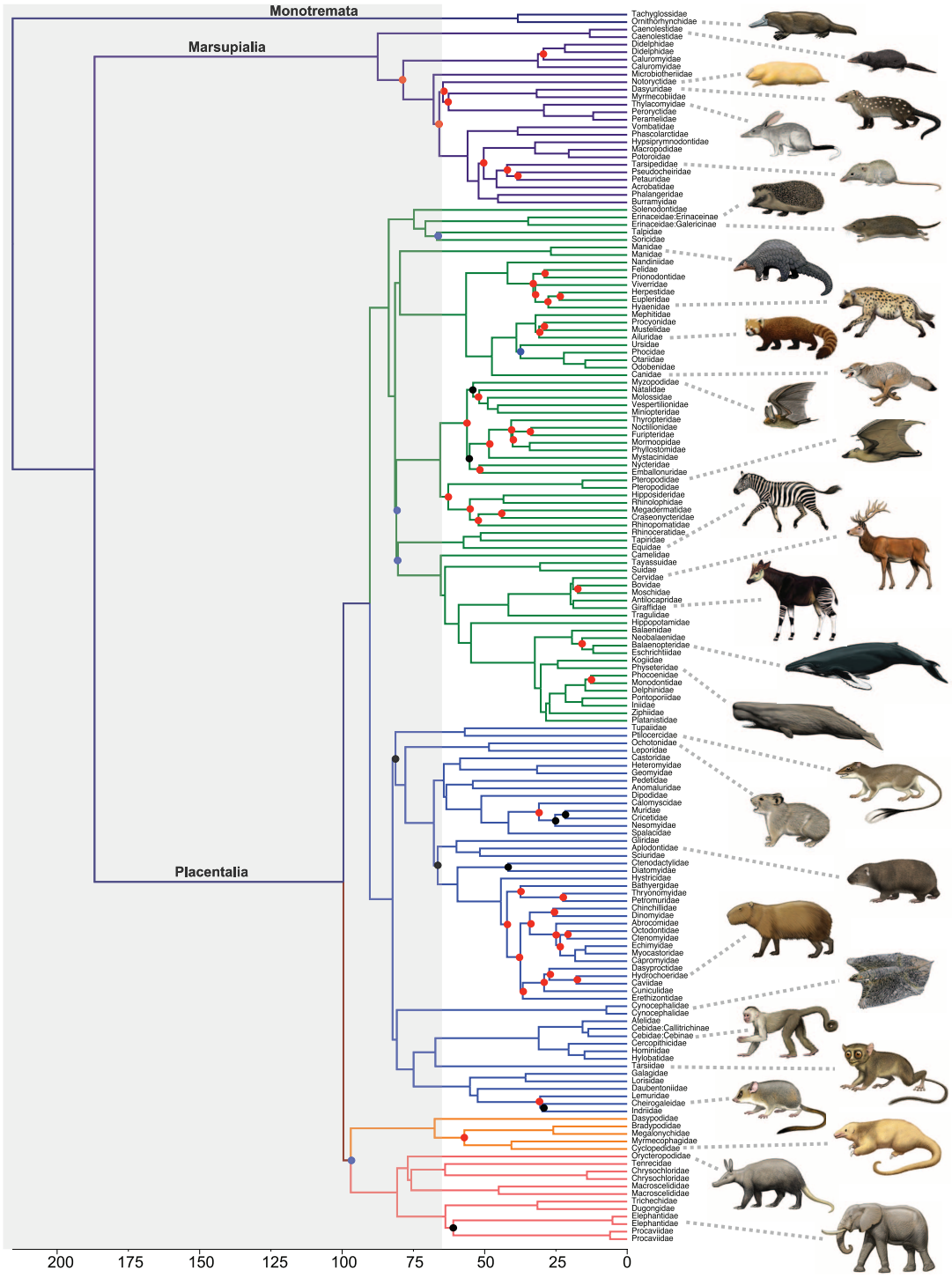

Figure 2: Phylogenetic tree of the mammalian families [5]. Each tip represents a mammalian family. Branching points are speciation events and branch lengths represent time in million of years.

Understanding macroevolution through phylogenetic analysis

I aim to develop statistical tools and computational methods to estimate, based on a species phylogeny (such as the one shown in Figure 2), the dynamics of speciation and extinction. The idea is to use information about the topology of the phylogeny and the time between branching events to gain information about past macroevolutionary dynamics.

My first major contribution to the field was to test a common hypothesis about the mammalian past. Did mammals only begin to diversify rapidly after the extinction of dinosaurs? The hypothesis of late mammalian diversification is mainly based on fossil evidence and thus put forward by paleontologists. I showed, by developing novel statistical tools, that, based on molecular data, mammals actually already diversified millions of years prior to the dinosaur extinction, meaning I reject the paleontological hypothesis [5,7].

I went on and further investigated more distant clades to humans. I looked at birds, showing that Warblers are competing across species and concluded that the overall species number is bound [3]. A study focusing on plants revealed that the sapindaceous clade (including species such as maple, horse chestnut, and lychee) underwent increased diversification around 30 million years ago [1]. My future goal is to determine the global rules of speciation, i.e. the major factors determining speciation and extinction, such as climate or tectonics, as well as reproduction modes meaning sexual vs. asexual reproduction, a major puzzle across the tree of life.

Looking for dinosaurs and other events happening millions of years ago is in principle a very pure science — it is driven by our wish to understand where we come from. However, once we have a good understanding of past events, we may be able to use it for conservation biology strategies.

Understanding epidemic spread through phylogenetic analysis

On a much faster and more immediate timescale, I use the same methods for uncovering the threats posed to human society by pathogens.

Unlike species, pathogens do not evolve on a timescale of million of years, but within days and months. Thus, e.g. every HIV infected individual has a different pathogen population within their body. Looking at the pathogens within each host we reconstruct the phylogeny where the branching points are transmission events (see Figure 3). Sequencing data of human pathogens are widely available, partly because for some pathogens, such as HIV, sequencing is routinely

performed by medical doctors for detecting drug resistance mutations in the virus, and thus providing the patient with the optimal drug cocktail based on the HIV genotypes present in the patient's pathogen population.

The analysis of these pathogen phylogenies sheds light on the dynamics of how pathogens spread and which public health interventions are necessary. I focus on human pathogens, an immediate threat to all of us. I recently showed how to quantify the severity of the spread of an emerging pathogen using genetic data [9]. In particular, we could identify a peak in past Hepatitis C virus (HCV) transmission in Egypt, co-occurring with antischistosomal injection therapy which is suspected to have been performed with HCV-contaminated needles. Further, phylogenetic tools can uncover population structure in epidemics [8], a factor of highest importance for predicting future spread. We showed that the heterosexual HIV epidemic in Latvia would disappear without the constant inflow of infections via intravenous drug users. Moreover, the epidemic within men-having-sex-with-men is mainly driven by 10% of the population (so-called superspreaders) contributing 90% of the transmissions. We communicate our findings widely and hope that these insights can help the design of effective public health interventions.

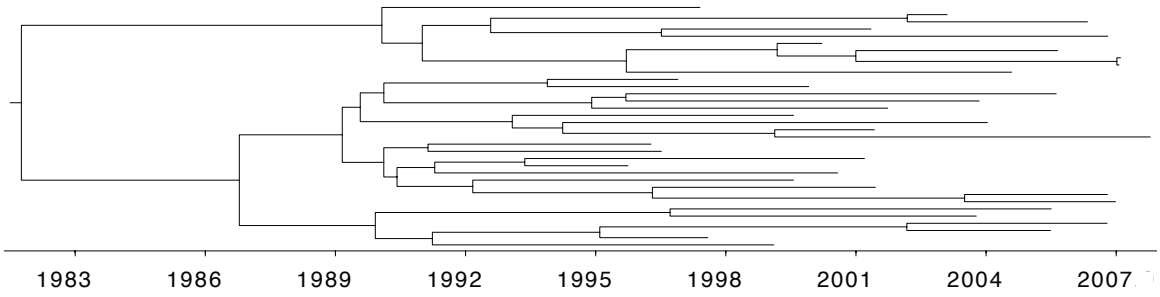

Figure 3: Phylogenetic tree of an HIV epidemic. Each tip represents HIV sequencing data from one infected host. Branching points are transmission events and branch lengths represent time.

The same methods are applied not only for human pathogens, but also for domesticated animal and plant as well as wildlife pathogens. Epidemics in domesticated animals cause great economic burdens as seen during the 2001 Foot-and-Mouth outbreak in the UK, which resulted in the killing of over 10 million sheep and cattle, and was estimated to have cost 8 billion Pounds. Epidemics in wildlife threaten biodiversity, e.g. a recent fungus already drove several amphibian species to extinction. Animal epidemics furthermore have highest relevance to humans, as most novel human pathogens are pathogens that

jumped from animal hosts to humans. Prominent examples of such zoonoses are HIV, SARS and Influenza. I recently started working on methods to understand the dynamics of these host jumps.

Summary

In summary, by obtaining a more complete understanding of how the living world around us changes, I want to continue to contribute to pure research, but in particular also use the acquired knowledge about speciation and extinction to inform conservation biology interventions to maintain biodiversity, a major challenge for evolutionary biologists, ecologists and conservation biologists. In the area of infectious diseases, I want to use our acquired knowledge to inform public health policy makers to improve strategies fighting epidemic spread in humans, animals and plants.

References

[1] S. Buerki, F. Forest, T. Stadler, N. Alvarez. The abrupt climate change at the Eocene-Oligocene boundary and the emergence of Southeast Asia triggered the diversification of sapindaceous lineages. Annals of Botany, 112(1): 151-160, 2013.

[2] C. Darwin. On the origin of species. 1859.

[3] R.S. Etienne+, B. Haegeman, T. Stadler, T. Aze, P.N. Pearson, A. Purvis, A.B. Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. Roy. Soc. B, 279: 1300-1309, 2012.

[4] R. Gibbs, J. Rogers. Genomics: Gorilla gorilla gorilla. Nature, 483: 164–165.

[5] R.W. Meredith, J.E. Janecka, J.Gatesy, O.A. Ryder, C.A. Fisher, E.C. Teeling, A. Goodbla, E. Eizirik, T. Stadler, D.L. Rabosky, R.L. Honeycutt, J.J. Flynn, C. Steiner, T. Williams, T. Robinson, A. Burk, N.A. Ayoub, M.S. Springer, W.J. Murphy. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Extant Mammal Diversification. Science, 334(6055): 521-524, 2011.

[6] M. Ruvolo. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol., 14(3):248-65.

[7] T. Stadler. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Nat. Acad. Sci., 108(15): 6187-6192, 2011.

[8] T. Stadler, S. Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Phil. Trans. Roy. Soc. B, 368 (1614): 20120198, 2013.

[9] T. Stadler, D. Kühnert, S. Bonhoeffer, A. Drummond. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Nat. Acad. Sci., 110(1): 228-233, 2013.