CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

1 Introduction

For those of us interested in the scientific study of language and speech (whom I will call here, for want of a better term, *language scientists*), keeping abreast with the relevant knowledge and scientific disciplines is becoming more and more difficult. Besides linguistics "proper" (and its classic sub-disciplines phonetics, phonology, morpho-syntax, semantics and pragmatics) now we must know some psychology, we have to be conversant in the cognitive neurosciences, understand something about language disorders and, lately, be on friendly terms with genetics. Familiarity with this latter discipline is increasingly necessary for meaningful discussions about *language origins and evolution*, its *acquisition* by children and the design of individually tailored effective *second-language learning* curricula, the structure of our *capacity for language*, and to address language and speech *impediments*, to mention just a few.

Unfortunately, fundamental notions of genetics are not yet part of the standard training in the language sciences, and this means that the interested language scientist must either ignore it at his/her own peril, acquire it piecemeal from heterogeneous popularization sources (with their associated uneven quality, reliability and relevance) or plunge head-on into the dense, confusing and exponentially growing primary literature. Another possibility would be to read one or more of the existing excellent introductions to genetics, genomics, biochemistry, population genetics, evolutionary theory, etc., but these are in general too broad, they address a very different audience and cover much too much material, most of it uninteresting and not directly relevant for the language scientist.

This book aims to fill this gap by offering an introduction to selected aspects of modern genetics and genomics, tailored for scientists involved in the study of language and speech. It tries to provide a condensed selection of relevant topics, briefly introducing the needed concepts, methods and results, and using – as far as possible – examples directly related to language, speech and hearing, while constantly pointing the interested reader towards important papers and recent developments and trends in these areas. I hope that after finishing this relatively small book you will have a deeper appreciation of what genetics is, how it can be used in your work, and how to interpret findings that have a genetic component. Moreover, you should be comfortable addressing

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

the primary literature and navigating the new developments and findings that will keep arriving at an ever increasing rate.

But most important for me is that you should be able to actively participate in inter-disciplinary research involving genetics: to properly study this enormously complex phenomenon – human language – our varied expertise is essential, as essential as that brought in by geneticists, neuroscientists, statisticians and psychometricians, among others, as equal partners in a dynamic and creative dialogue. To make this work, each participant needs to grasp to an acceptable degree what the others are thinking and doing and, more generally, to try to see the world through their eyes.

Thus, this book aims to invite us, the language scientists, to see the world through genetic lenses as it were, making us capable not only of judging the relevance of genetic findings and methods, but of actively participating in the adaptation of existing methods – and the invention of new ones – appropriate to the questions we are interested in answering.

Writing this book has been particularly difficult for a number of reasons. Genetics is an extremely broad, complex and very rapidly evolving field, where quite a sizeable proportion of publications from 10 years ago are literally old and their assumptions, methods and findings were amended or even invalidated by newer publications. A field where new directions and research questions continually pop up, where more often than not there is a real race between multiple teams to publish similar results in high-ranking journals such as *Science* and *Nature*, where hair-raising ethical issues are combined with tremendous pressures emanating from the health industry and political agendas. Where the lone genius is more and more a rarity being replaced by labs of tens of people and networks of tens of such labs producing papers with tens or even hundreds of authors. Where the funds required for a single project go beyond the wildest dreams of most social scientists, not to mention those working in the humanities...

On the other hand, as the title tries to convey, there probably is no unified "Science of Language" to speak of but a plethora of fields of research springing from different historical roots, using quite different methods and having different goals and standards of explanation, which I assume that you, the reader, are painfully familiar with. Thus, besides the several schools within theoretical linguistics proper, there are typologists, historical linguists, sociolinguists, morphologists, syntacticians, semanticists, dialectologists, psycholinguists, phoneticians, phonologists, cognitive scientists, neuroscientists, speech pathologists, engineers working on speech comprehension and synthesis, neurologists and psychiatrists dealing with speech and language, and philosophers of language, to mention just a few and glossing over the differences hidden behind such convenient labels.

I wrote this book trying to keep in mind the varied needs and interests of all of them when it comes to genetics. Some will want to know about the

© in this web service Cambridge University Press www.cambridge.org

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

genetics relevant to language evolution, others to understand how the genes and the environmental factors manage to build competent language users by the billions, while yet others try to find parallels between the patterns of linguistic and genetic diversities. Some will certainly want a conceptual, generic, bird's-eye view of genetics while others will feel disillusioned if actual methodology and mathematics are lacking. And some would prefer cutting-edge research and results while others understandably will want to first build a solid basis on those results that have stood the test of time (of which there are plenty in genetics) from which to confidently start exploring. I hope I have managed to address all these issues and I hope that for each reader there is something useful here, hopefully more than a lone paragraph buried among pages of useless prose, tables, equations and figures.

The book tries to be as modular as possible, but still the best approach is to read it sequentially given that concepts, methods and findings are introduced as needed.

First (Chapter 2), we begin by addressing the various approaches to the nature-nurture question focusing on heritability and the amazing complexity behind seemingly simple concepts such as "innate" and "acquired". This topic, of what is due to "nature" and what to "nurture", is an important one in the language sciences, but unfortunately the manner in which it is sometimes approached feels rooted in the past, disconnected and impervious to recent advances in genetics, developmental and evolutionary biology. This chapter tries to offer an updated view of the concepts, findings and methods, and to ensure a proper understanding. We then (Chapter 3) encounter the actual reality of how genetic information is stored, transmitted and expressed, discussing such processes as replication, transcription and translation, and the structure of genes. Chapter 4 focuses on patterns of inheritance, exemplifying them with some examples relevant to language and speech such as a dominant pathology affecting speech, recessive hearing loss that resulted in the emergence of a new sign language and the sex-linked transmission of colour-perception deficits. We will discover how genes are actually found in Chapter 5 where we encounter association and linkage studies and see some examples of genes discovered using these methods, while the next chapter (6) gives some actual examples of how genes work. This chapter is very important not only because it describes real-world cases of genes affecting phenotypes relevant for speech and language but also because it dispels any simplistic notions about how genes do their jobs. A very short Chapter 7 discusses the promises of whole exome and genome sequencing, but for now this potential has not been used for language and speech. Also here I put together quantitative and molecular genetics and try to illustrate what to expect about the genetic architecture of speech and language. I dedicate a special chapter (8) to population and evolutionary genetics, discussing the forces that shape the genetic structure of human populations and their relevance for understanding human history and

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

the patterns and processes shaping linguistic diversity. This should provide the foundations necessary for discussing in their context the issues surrounding language origins and evolution, on one hand, and the biological background to the patterning of linguistic diversity, on the other. This leads naturally (Chapter 9) to discussing recent advances in the understanding of the fascinating interactions between culture and the biological bases that make it possible. I review several cases of gene–culture co-evolution with a special significance for language and speech such as the spontaneous emergence of new sign languages in communities with a high incidence of hereditary hearing loss and the proposal that our genetic background might bias the process of language change, thus influencing linguistic diversity. Finally, as a guide for the interested reader, the conclusions (Chapter 10) contain a list of further readings and other resources (such as online databases and tools) that can be consulted for further information.

Throughout the book there are references to both reviews and introductory texts, on one hand, and to primary research (fundamental findings, description of methodology or cutting-edge reports) on the other, allowing the interested reader to continue on their own and deepen their expertise. The Appendix provides the actual `R` code implementing some points discussed in the book, while abundant footnotes clarify and give technical detail and actual snippets of code, as needed. Finally, a Glossary provides short definitions of the most important new terms and abbreviations.

> ***Box 1: Boxes with technical details***
>
> Technical discussions and mathematical details are included in boxes such as this one and can be safely skipped. However, they are still recommended for a fuller understanding of the topic under scrutiny.

But before we start this journey of discovery, I must try to answer a fundamental question that I heard several times being explicitly formulated, and many more times lurking implicitly behind comments, suggestions, questions and discussions over a pint of beer: why should I, as a student or scientist interested in language and speech, care about genetics and evolutionary theory?

1.1 Why is genetics relevant for me?

Indeed, why? Why invest precious time and effort in reading this 300+ pages book?

This is a frequent thought (if not a frequently asked question) when genetics is introduced to language scientists. The same might be said about statistics,

© in this web service Cambridge University Press www.cambridge.org

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

evolution or electrical engineering! But while a weak argument can be made for the last one (it's used in some approximations of the acoustics, not to mention building and maintaining experimental equipment), it is much easier to argue for the others. Indeed, statistics is not only useful as a tool for testing the difference between two conditions in an experimental design, but offers a surprising and at times inspiring manner of viewing the world. Likewise, evolutionary thought puts things in a much wider perspective and gives meaning to many otherwise incomprehensible phenomena, the emergence of language being just one of them.

Genetics is relevant on many levels to those studying language and speech. A useful distinction can be made here between three such levels of complexity and associated time scales (Smith and Binder, 2013): the *individual*, the *population* and the whole *species*. The individual level and the associated **ontogenetic** processes concern the build-up of the machinery necessary for learning and using language. The population level and the **glossogenetic** timescale (Hurford, 1990) refer to the supra-individual processes acting over longer periods of time and involved in language change and the patterning of linguistic diversity. Lastly, the species level involves **phylogenetic** processes shaping the emergence and evolution of language on even longer time scales.

Necessarily, these distinctions are artificial and all these levels, processes and time scales are continuously interacting, defining each other in the process. Thus, it makes no sense to discuss isolated individuals (where would they acquire language from?) or groups abstracted away from the people composing them (who is doing the talking?) without the evolutionary context that produced both the capacity for language and the actual languages we speak (which is most probably not a coincidence). Nevertheless, these three levels offer a useful first approximation and should be kept in mind as we think about the genetic foundations of language.

To reiterate, an understanding of genetics is relevant to all of them. First, it should be obvious that the development and maintenance of a language user are rooted in genetic mechanisms that, in intimate and continuous interaction with environmental factors including the general cultural and the linguistic, ensure the development of the organs and systems necessary for perceiving, producing, processing and learning language. As we will see, this is not an encapsulated phase of "development" where the genome is "read" (as a recipe for making bread would be) and then archived and forgotten until the next generation needs it in the womb; quite the opposite, our genome is a dynamic and active thing continuously being expressed and involved in complex regulatory cycles, reacting to changes within and without our bodies on the level of the millisecond, allowing us to adapt to our continuously changing environment and to learn.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

Likewise, nobody would deny that changes in our genomes were necessary to make our species develop (invent?) and use language, but the exact nature of these changes, the reasons they happened, and when they did are very contentious issues. Understanding the structure of our genome and how it is expressed helps constrain the range of plausible accounts for the emergence and evolution of language and current breathtaking advances quickly transform armchair speculation into testable hypotheses. Famously, the Société de Linguistique in Paris banned discussion about the origins of language in 1866, arguing, convincingly, that the evidence simply is not there to test the proposals, but we are quickly reaching an age where after some 150 years, thanks mostly to genetics, this ban can be safely lifted.

Probably the hardest to justify is why genetics would be relevant at the population, glossogenetic level: what would a typologist, historical linguist or field linguist working in the Amazon gain from understanding genetics? One answer is that, for language as for biology or any other entity shaped by historical processes, the past is the key for understanding the present and more and more understanding the past of language means understanding the past of its speakers, where a major role is played by genetics (Cavalli-Sforza et al., 1994; Jobling et al., 2013). We will only touch on this fascinating topic here as there are very good introductory works available, with Jobling et al. (2013) being highly recommended, but the key insight is that this type of correlation (or lack thereof) between languages and genes is purely accidental, being caused by a shared causal factor, namely historical processes affecting populations. Thus, there's nothing in the genes of the populations speaking Chinese languages (such as Mandarin, Cantonese, or Wu) that makes them speak such languages; any correlations there might be between their genes and their languages are purely an accident of history.

But there might also be causal links between a population's genetic make-up and the language(s) it speaks (Dediu, 2011a, 2013) in the sense that a genetic background (dis)favours the presence of certain structural (or typological) linguistic features, such as the use of variations in voice pitch to convey not only intonation but also distinctions between words or grammatical information (what is called **linguistic tone**; see for example Yip, 2002). For example, there could be something in the distribution of genetic diversity within South-East Asia that makes the presence of tone languages (such as the Chinese languages but also Vietnamese and Thai) much more probable than say in Europe (Dediu and Ladd, 2007). Such *genetic biases* are very weak at the individual level but get amplified through language use and transmission, such that they influence the trajectory of language change and, ultimately, the distribution of linguistic diversity (Ladd et al., 2008).

Therefore, understanding how our genome is structured and how it works is indeed relevant for most kinds of language scientists, and opens up new perspectives on the nature of language, its evolution and change.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

2 Nature, nurture, and heritability

In this chapter we approach, at a fairly abstract level, the fundamental questions concerning the relationships between the phenotype (the observable properties of individuals), the genotype and the environment. We discuss the paramount importance of variation in studying these relationships and we define, estimate and discuss the meanings and misinterpretations of heritability. Far from being a simple concept, heritability will turn out to have some non-intuitive properties that make the interpretation of heritability estimates quite a tricky exercise. Likewise, we will discover that, in fact, all the related concepts and distinctions, such as innate and acquired, or nature and nurture, are fuzzy and far from their apparent clarity in everyday discourse. We will end with a very brief survey of heritability studies in speech and language. This chapter also introduces several fundamental concepts of statistics that are necessary for a proper understanding of many topics covered in this book.

2.1 Phenotype, genotype and environment

It is unquestionable that both "nature" and "nurture" are required for the development of a linguistic human being. Lacking "nature" will limit language development no matter how much "nurture" there might be, as many a pet owner can easily confirm. This is seemingly supported by studies of chimps (such as Nim Chimpsky and Washoe) reared in conditions similar to those experienced by human babies and infants, but which nevertheless fail to go beyond a rather limited level of language usage. On the other hand, having "nature" but lacking "nurture" is equally devastating, as shown by the cases of children who, for various reasons, have not been exposed to language during the so-called **critical period** for language acquisition (a well-known case being Genie) and who fail to develop full-blown language despite considerable efforts.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

Thus, if we denote, in a highly abstract manner, the "nature" as G (from genetics) and the "nurture" as E (for environment), then we can attempt to write down a symbolic equation describing how these two factors relate and interact in producing the phenomenon of interest, P (for **phenotype**). For us, P will usually mean some aspect of language and speech or some other relevant cognitive process, but it can mean virtually any feature an organism possesses. Thus, it can refer to what we may loosely conceptualize as individual features, such as a person's height (as measured with a meter from the top of the head to the feet while standing), to the eye colour subjectively placed in categories such as "blue", "green", "brown" or "black", to molecular aspects such as the speed with which a certain enzyme breaks down a given molecule in the body, or to relational phenomena such as pair-bonding or the use of language and speech. Of course, these levels are far from clear-cut and fixed, but they prove useful in understanding complex phenomena such as those of interest here. Moreover, a major enterprise in modern science is to be able to understand how lower-level phenomena interact in order to produce higher-level ones which, in this context, means that we would like to provide a "full story" ranging from molecules to social networks and language change, without necessarily implying a strong a priori reductionism.

The simplest form of such an equation would be (i) $P = G$ or (ii) $P = E$, which would represent the cases where a phenotype is purely the product of the genes or of the environment. What would qualify here, though? Maybe (i) could describe those aspects of an individual which are "purely" biological while (ii) would be applicable to those which are shaped by the environment alone. Like having a heart, which all humans do, or dyeing your hair, which clearly depends on someone's culture. In what sense is it meaningful to think that (i) or (ii) would hold? What is the basis for the intuition that having a heart might be an instance of (i) while dyeing one's hair is an instance of (ii)? It seems that we think G is behind humans having a heart because all humans do and there are other things which do not and this seems quite stable across environments, cultures, historical periods, etc. Analogously, we think that E is behind hair dyeing because some do and others don't and this critically depends on one's local culture, available technology, etc. Thus, all these judgements essentially rest on *the patterns of variation* in P: while hearts seem to follow biology and disregard the environment, hair dyeing seems to do the opposite.

However, taking a closer look at those phenotypes which might seem fully determined by "nature", on one hand, and at those fully determined by "nurture", it soon becomes clear that these cases are pure abstractions, lacking any reality or meaningfulness. There are people without a heart (not only in the metaphorical sense!), but they are simply aborted at an early stage of development. Moreover, there are other milder conditions compatible with

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

intra-uterine development, birth and even life to maturity, characterized by various defects of the heart[1] and, while for some there is a clear genetic component, for others there is an as clear environmental causation, including maternal pre-gestational diabetes, infection with rubella or exposure to the drug Thalidomide during gestation, among others (Jenkins et al., 2007; van der Bom et al., 2010). On the other hand, hair dyeing basically needs some hair to start with, hair development, its patterning on the body, the right biochemical properties, etc., which all involve a lot of genetics (Shimomura and Christiano, 2010; Törnqvist et al., 2010). Moreover, one's personality, gender and other factors, all having some genetic component, have a role to play in the actual behaviour under consideration. Thus, we can safely conclude that the equations (i) and (ii) cannot hold and we need to consider more complex ones involving both G and E.

But before delving into the complexities of heritability it is instructive to briefly overview the fascinating debates surrounding seemingly simple and familiar concepts such as "innate", "acquired", "inborn", "learned", "nature" and "nurture".

2.2 Innateness, a slippery and complex concept

Many fundamental arguments in the language sciences (and cognitive sciences in general) seem to revolve around the twin notions of "innate" and "acquired" (or the related "nature" and "nurture"). To some leading linguists, Noam Chomsky included, various properties of language such as the patterning of linguistic diversity with the apparent existence of **universals** (Newmeyer, 2005; Comrie, 1989; Croft, 2003), the process of language acquisition seemingly capitalizing on information not present in the primary data (the so-called **poverty of the stimulus** argument; Chomsky, 1980), and the computational data structures and machinery postulated for processing language (Chomsky, 1965, 1980) seem to point to "innateness". However, what exactly is "innate" and the exact nature of this "innateness" are extremely vague (see, for example, Pullum and Scholz, 2002; Mameli and Bateson, 2011; Mameli, 2008; Cowie, 1999; Kiikeri and Kokkonen, 2007; or Bateson and Mameli, 2007) and in need of clarification.

In fact, when a proper analysis of these apparently obvious concepts is done in the light of modern data and theory from the biological and cognitive sciences, one is left with a collection of not necessarily consistent proposals and properties. Mameli and Bateson (2006) (see also Mameli and Bateson, 2011; Mameli, 2008; Bateson and Mameli, 2007) conduct such an analysis of "innateness" and identify not less than 26 manners in which this concept has been (explicitly or implicitly) defined and used in the scientific literature

[1] Encompassed under "congenital heart disease" (Hoffman and Kaplan, 2002).

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00129-9 - An Introduction to Genetics for Language Scientists: Current Concepts, Methods and Findings
Dan Dediu
Excerpt
More information

(Table 1 on page 177 in Mameli and Bateson, 2006, gives a summary). For example, one might define an "innate" character as being not acquired, but a quick look at development reveals that basically all characters are acquired given that the just-fertilized egg is quite different from the adult. Likewise, the "obvious" presence at birth fails to capture things like sexual characteristics and parental behaviour, while its refinement requiring the reliable appearance at a well-defined life stage fails to account for learning/cultural influences.

Genetic determinism, influence and encoding are increasingly refined versions of a familiar argument but fail to explain much, given that genes do not, in general, deterministically dictate development but arguably all traits are influenced by genes, while genetic information/encoding is a very difficult notion. In this same vein falls Chomsky's poverty of the stimulus argument which, in a nutshell, proposes that some traits do not extract information from the environment (in the particular case of language acquisition, supposedly some evidence is simply not in the data the child sees). However, while intuitively appealing (if this information is not provided by the environment it must come from somewhere else, namely the genes), as always, nature is much more complex and subtle. For example, scars and calluses do not seem to extract information from the environment for their formation but are not really "innate", while the complexity of the interactions between genes and environmental factors during development seems to rule out such simple dichotomies (we will encounter later in this book several examples of such fascinating interactions). An example used by Mameli and Bateson (2006) is the classic work by Gottlieb (e.g., 1991, 1997) on mallard ducks showing that the exposure of ducklings to their own calls while still inside the egg facilitates the later recognition of the (otherwise quite different) adult species-specific calls.

One final type of definition that we will discuss here is the widespread idea that something is "innate" if it is species-specific or species-typical. However, for almost all interesting characteristics there are some members of the species that lack it (or have "deviant" versions of it), raising the question of how to define the norm and the exceptions. This problem of defining the norm is extremely difficult to solve, as shown by the complex issues faced by medicine and psychiatry, for example. Another issue is that for many approaches individuals suffering from genetic pathologies – such as the developmental speech dyspraxia (**DVD**) experienced by members of the KE family affected by a mutation in the ***FOXP2*** gene (see Sections 4.2 and 6.7 for details) – must be considered atypical, implying that these pathologies are not "innate" on this account. Finally, characteristics that are learned by all normal individuals, with reading, writing and using a mobile phone rapidly becoming valid examples, must be considered as "innate".

The bottom line of this extremely instructive exercise is that no single definition of "innate" (or "acquired", "nature, "nurture") seems to capture

© in this web service Cambridge University Press www.cambridge.org