Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Integrating genomics into evolutionary medicine

Juan Antonio Rodríguez[1], Urko M Marigorta[2] and Arcadi Navarro[1,3,4,5]

The application of the principles of evolutionary biology into medicine was suggested long ago and is already providing insight into the ultimate causes of disease. However, a full systematic integration of medical genomics and evolutionary medicine is still missing. Here, we briefly review some cases where the combination of the two fields has proven profitable and highlight two of the main issues hindering the development of evolutionary genomic medicine as a mature field, namely the dissociation between fitness and health and the still considerable difficulties in predicting phenotypes from genotypes. We use publicly available data to illustrate both problems and conclude that new approaches are needed for evolutionary genomic medicine to overcome these obstacles.

Addresses
[1] Institute of Evolutionary Biology (UPF-CSIC-PRBB), Barcelona, Catalonia, Spain
[2] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA
[3] Center for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
[4] National Institute for Bioinformatics (INB), Barcelona, Catalonia, Spain
[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Catalonia, Spain

Corresponding author: Navarro, Arcadi (arcadi.navarro@upf.edu)

Current Opinion in Genetics & Development 2014, **29**:97–102

This review comes from a themed issue on **Genetics of human origin**

Edited by **Aida Andrés** and **Katja Nowick**

http://dx.doi.org/10.1016/j.gde.2014.08.009

Introduction. Two views of medicine

The field of evolutionary medicine, also called Darwinian medicine, was established in the seminal papers by Paul Ewald (1980) [1] and by George C. Williams and Randolph Nesse (1991) [2••], who first advocated the idea that natural selection and, in a wider sense, evolutionary biology, could help understanding the origins and causes of disease in our species. However, the links between evolutionary and medical thought are older than that. For example, evolutionary principles had been unwittingly applied by slave traders, who would lick the skin of African slaves to ascertain their chances of surviving the lengthy and arduous journey to the New World. Individuals tasting less salty were less prone to experience dehydration and thus more likely to survive the trip [3]. In perhaps one of the first uses of evolutionary thought, Muller, in 1948, attempted to explain *why* an ailment existed rather than focusing on *how* it appears and *how* to alleviate it — suggesting that fevers could be an adaptation in response to bacterial toxins. This idea was proven correct almost 40 years later [4,5] and, since then, Darwinian medicine has been providing insight into the evolutionary causes of complex diseases, such as cancer [6] and processes like ageing [7,8••].

In contrast, the field of medical genomics focuses on immediate questions about *how* diseases appear and how they advance within an organism [9,10]. Over the last 50 years, genotype-phenotype studies aimed to identify genetic variants responsible for disease susceptibility and elucidate their molecular mechanisms. As early as the mid-1960s, an HLA haplotype had been associated to Hodgkin's disease [11,12], and by the early 1970s, several other HLA loci were linked to autoimmune conditions, like type 1 diabetes [13]. Thanks to these and other studies, some of the molecular mechanisms behind many diseases were unraveled prior to the genomics era. Two notable cases are the mutations associated with Huntington's disease and cystic fibrosis. The first caused by the expansion of the simple repeat "CAG" in the *HTT* (huntingtin) gene [14] and the second due to the deletion of a phenylalanine in the *CFTR* gene [15]. Progress in this area accelerated once the human genome was completed in 2001 [16], and continues to advance as high-throughput-omics technologies become more accessible [17]. Many of these advances are already resulting in new diagnostic and therapeutic tools that are improving human health world-wide [18].

Unfortunately, these two views of medicine have not yet fully converged. The potential benefits of an evolutionary approach are not widely recognized within medical genomics, and much less within clinical practices. Although many efforts are currently under way to raise awareness about evolutionary thought [19,20], most medical schools still lack an evolutionary biology course [21••]. This state of affairs is somewhat surprising, as a combined formulation of the two views of medicine presented above would result in a much deeper understanding of disease. This combined field could be called *evolutionary genomic medicine* or EGM, even if other names emphasizing the genomic, rather than the medical, aspect have been proposed [22]. EGM studies disease at different levels:

from its ultimate evolutionary origin to its immediate molecular mechanisms. EGM research is gathering momentum and should eventually become a burgeoning area. This type of research has already proved constructive but two main blockers hinder its full-fledged application. We review them below.

Evolutionary genomic medicine: successes thus far and challenges ahead

Examples of case studies for EGM are piling up. Perhaps the better known instances of a successful application of this perspective are the text-book example of sickle-cell anemia [23,24] and the identification of several mutations associated to lactase persistence [25], whose celebrated explanation is the co-evolution of dairy farming cultures and lactose tolerance in adults [26,27]. The consequences of the artificial selection imposed by slave licking have also been understood thanks to EGM. Since genetic variants favoring salt and liquid retention were positively selected before and during the ocean trip, current African Americans have increased odds of developing hypertension [3]. Another example is the impact that the Black Death possibly had on gene frequency variation in Europeans. It has been hypothesized that this epidemic shaped variation at the CCR5 locus that now provides resistance to other infectious diseases, such as AIDS [28].

In spite of these cases illustrating the value of EGM, evolutionary approaches are far from being commonplace. The slow advance of EGM has many causes [21••], but we believe that two of them are particularly challenging since they highlight two glaring gaps in our knowledge: the twin dissociations between health and fitness and between genotypes and phenotypes.

Dissociation between fitness and health

Natural selection favors reproduction over health. So, in taking an evolutionary standpoint it is crucial to enquire about the reproductive consequences of any ''disease'' or ''condition'' since, in the end, what we call a ''disease'' may have no consequences in terms of natural selection or evolution [29]. It has been postulated that certain diseases may be the result of adaptations to ancient environments that would have lost their advantage today [30]. For example, the thrifty genotype hypothesis [31•] follows this line of reasoning by posing that alleles conferring risk for certain ''affluence diseases'', such as type 2 diabetes, are common today because they were advantageous in the past. During situations when food resources were scarce those individuals with a more efficient or *thrifty* metabolism would be more likely to survive and pass on their now disadvantageous alleles [32]. Recently, a consortium of type 2 diabetes provided functional evidence for the idea that the Hispanic Mexican population presents higher frequency for risk alleles of type 2 diabetes as an adaptation to a harsher past environment [33]. Already in his 1962 paper, Neel had foreseen this result ''*[...] diabetes mellitus as an untoward aspect of a thriftiness genotype, which is less of an asset now than in the feast-or-famine of hunting and gathering cultures*'' [31•].

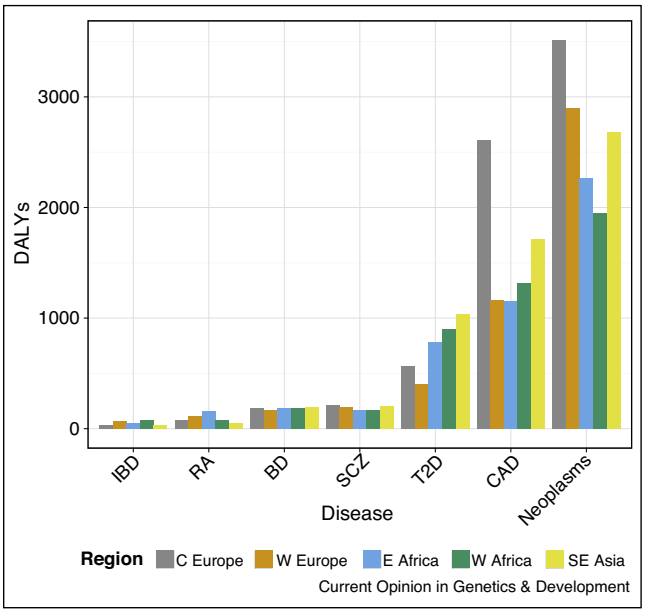
Not all diseases have the same relation to fitness. Rather than past adaptations rendered useless in modern times, some conditions are more likely to represent complex trade-offs arising from adaptive pressures toward different directions. Consider, for instance, elevated testosterone levels. They are known to be beneficial in increasing reproductive success, but it has recently been suggested that they may decrease resistance to infections, since the immune system reallocates to perform further tasks in situations where testosterone and stress hormones are released [34,35].

Given these uncertainties, one of the major challenges of EGM is coming up with an adequate proxy of fitness that adequately reflects the reproductive impact of a disease. The difficulty of this endeavor can be grasped by considering current estimates of the burden of disease in terms of a standardized measure: Disability-Adjusted Life Years, or DALYs [36]. The number of lost DALYs is a unit used by the Institute for Health Metrics and Evaluation [37], to measure how many life years are lost due to sickness, living with a disability or premature death. Some conditions score very low in the DALY scale. For instance, no deaths and nearly no DALYs are lost due to psoriasis, an autoimmune disease with around 2–3% prevalence in populations of European ancestry. Interestingly, the prevalence of psoriasis in Africans is about half of that proportion [38], which may be suggestive of an adaptation to different out-of-Africa conditions. Other diseases, such as child cancers or prenatal disorders, are far more burdensome.

DALYs lost due to six conditions in five world superregions are presented in Figure 1. There are remarkable differences in lost DALYs even between bordering regions within the same continent, such as between Western & Central Europe in DALYs lost to coronary artery disease or between Western & Eastern Africa in DALYs lost to rheumatoid arthritis. These striking variations in the present impact of disease are good indicators of the difficulties of inferring the past fitness impact of disease. Consider, for example, the late Pleistocene, when living circumstances were radically different from now. Even if we can be sure that infection was a basic component of health in these times [39], the field still needs much effort on quantifying the prevalence of many other conditions, including fatal diseases such as childhood cancers.

Difficulties are increased in the likely scenario that most genetic variants causing complex disease are shared across human populations [40], and that, therefore, most of the differences in DALYs are due to environmental and lifestyle causes. In short; disease must be sought in

Figure 1



Disability-Adjusted Life Years (DALYs) across the World. DALYs lost to a certain disease, per 100 000 DALYs lost in the population in five super-regions of the World for 7 conditions, including those analyzed in the genetic risk estimation exercise in the main text (see Section 'Dissociation between fitness and health').

intricate combinations of at least three elements: the evolutionary history of our lineage, the changing environments we have faced in the past and the current evolutionary forces to which we are exposed [2••,41••,42]. Considering all this simultaneously may be fascinating, promising and indeed unavoidable if we want the field to progress, but it is also extremely difficult.

Dissociation between genotypes and phenotypes

A classical problem in evolutionary biology is that the patterns and modes of selection that were obvious at the organism level, including many obvious examples of adaptation, proved difficult to observe at the molecular level [43]. Moreover, the success of modern evolutionary studies in detecting cases of natural selection [44] has not resulted in the unveiling of the genetic architecture of known adaptations. Rather, researchers have been able to identify and sometimes even to date adaptive events while knowledge about their phenotypic effects is scarce [45,46]. In other words: the molecular mechanisms leading to even the most obvious phenotypic adaptation, such as the textbook examples of our opposable thumb and capacity for language, remain largely unknown; and, likewise, most known examples of adaptation at the molecular level still lack a mechanistic explanation and a link to relevant phenotypes.

Linking genotypes and phenotypes is anything but straightforward. Since the burst of Genome-Wide Association Studies (GWAS) in 2007 [47••] researchers are trying to bridge this gap. One initial goal was to predict individual disease risks under the light of the known disease-associated loci. Although large collections of data are available today [48], successes in that field are still meager. However success in genotype-phenotype prediction is a condition for EGM to achieve its full potential. To be able to use the tools and methods of molecular

Box 1 AUC calculations

SNP selection

We started by downloading the NHGRI GWAS Catalog [52]. We obtained all the relevant information, including publication date, for the SNPs that had been reported for the five traits studied here, namely: bipolar disorder, coronary heart disease, Crohn's disease, rheumatoid arthritis and type 2 diabetes. We considered exclusively studies performed and markers validated in samples of Western European origin. After that, we classified the associated SNPs by year of publication of the original study. Sets of 2 consecutive years were made, starting in 2008, up to 2014. For markers that mapped in the same region than a previously discovered one, we applied a linkage disequilibrium threshold of $R^2 > 0.2$ to considering them potentially as tagging the same locus. In these potentially redundant cases, we kept the SNPs that come from the studies with the largest sample sizes.

Risk estimation

With the set of markers per year and per disease obtained above, we aimed to evaluate the progression of the AUC for each disease. Using data from the original WTCCC GWAS [49] for each of the mentioned diseases, we performed 50 random resamplings of 500 cases and 500 controls out of the ~2000 cases per disease and ~3000 shared controls analyzed in the original paper.

For each of these 50 groups of 1000 individuals per disease we estimated the individual risk, as previously reported in [53]. To set a neutral reference background, we used the genotypic frequencies from 85 European ancestry individuals (CEU population) from the 1000 Genomes Project [54]. Out of these frequencies, and using a custom R [55] script, we simulated 100 000 individuals and estimated their genetic risk scores. Genetic risks were computed assuming the classical additive model by multiplying the number of risk alleles at each locus by the decimal logarithm of the Odds-Ratio (OR) in that locus and adding up over all risk loci [56•]. Doing so, we obtained a distribution of risk scores that could represent the background risk in the CEU population.

The next step was to calculate a risk score for each of the individuals that had been resampled from the WTCCC study. These individual risk scores were compared against the simulated background risk distribution. Individuals were classified as ''cases'' or ''controls'' according to the percentile in which they fell. Then, using the package rocR [57] and the true disease status of these individuals we calculated the AUCs. Similar results to those shown in the main text were achieved when using proportions of 5% cases ($n = 50$), and 95% controls ($n = 950$).

We are aware that there are several methods available [58] that may outperform the extremely simple classification approach followed here [59•]. Moreover, the adequacy of the AUC method is not free of criticism when applied to genetic risk prediction [60]. However, the aim of this study is merely to compare how prediction ability changes over time, rather than to study this ability in itself.

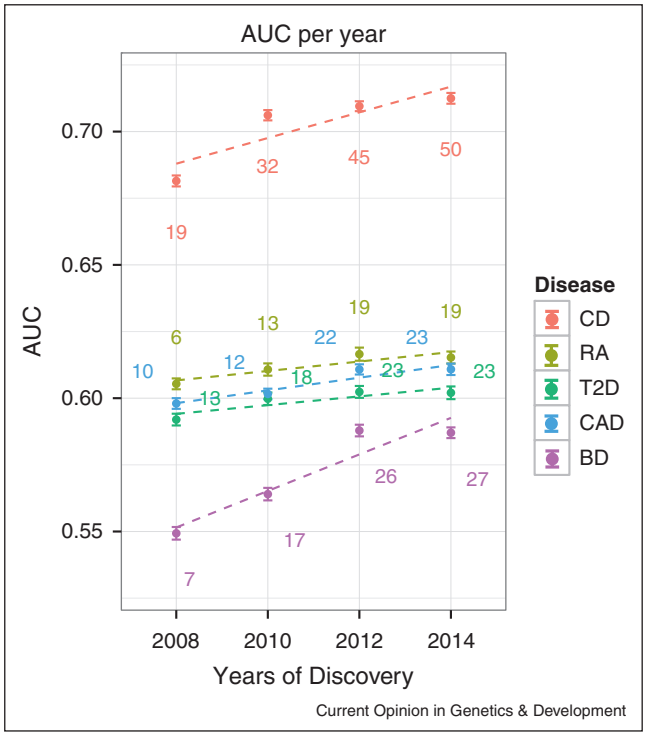
evolution in estimating the times of origin of mutations, their associated selection coefficients and so on, it is necessary that an appreciable proportion of the phenotypic variance for any trait or condition be assigned to observed variation. Are we getting any closer to that goal?

To obtain a rough evaluation of how fast the field is moving toward that point, we examined how our ability to predict phenotypes has changed since the first GWAS results became available. We downloaded genotype data from five diseases from the WTCCC study [49•]: Crohn's disease (CD), rheumatoid arthritis (RA), bipolar disorder (BD), coronary artery disease (CAD) and type 2 diabetes (T2D). Our goal was to compare predictions of the phenotypes of these individuals using the information about disease loci available to the scientific community at four different time points: 2008, 2010, 2012 and 2014. Following the procedures described in Box 1, we classified individuals and estimated the AUC (Area Under the ROC Curve). The AUC estimator shows graphically the performance of any binary classifier after it has been applied to a blind set of cases and controls. An AUC close to 1 is indicative of a very good classifier, while an AUC close to 0.5 indicates that the classifier is not faring better than what would be expected by chance.

The ability for risk prediction clearly varies significantly between these diseases (Figure 2). We can distinguish three main patterns that are clearly related to the genetic architecture of each disease. For the case of CD, prediction ability was high to start with and has become better with time. In 2008, our estimated AUC for CD was ~0.65, consistent with predictions by then [50], and with further studies it has increased to ~0.71, also in line with actual estimates [51•]. This is indicative of a few loci of strong effects that were detected by the first studies. A different pattern is observed for BD, with prediction ability increasing remarkably since 2008, when they were very poor, already suggesting that this is a highly polygenic trait affected by many loci of modest effects. Finally, a third group comprising diseases like RA, CAD and T2D are very complex disorders where environmental components may play a substantial role. Nonetheless, prediction power is slowly rising for each of them.

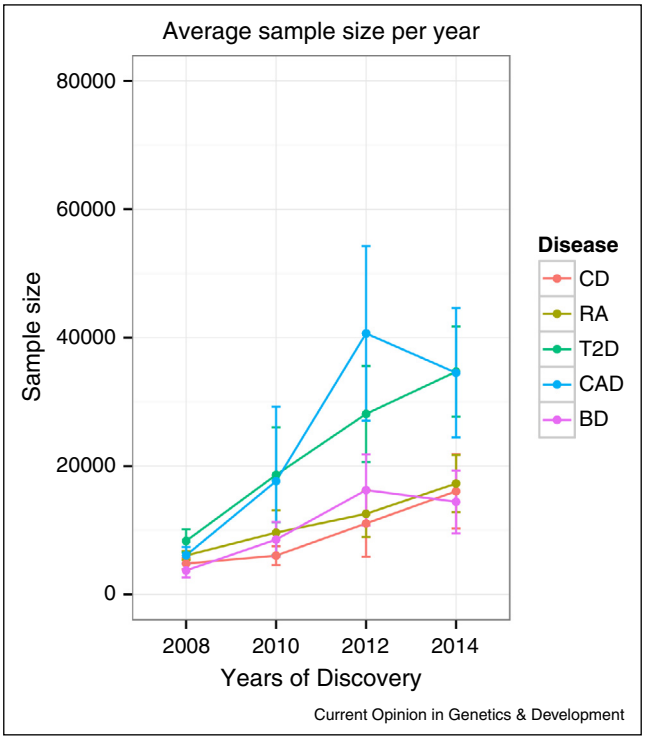
Most of the increase in AUC can be attributed to the larger sample sizes of more recent studies (Figure 3). The initial studies on CD and BD had sample sizes of ~5000 individuals, a figure that has doubled by 2014 resulting in a 4–5% increase in AUC. Contrarily, for T2D and CAD the more than 7-fold increase since 2008 in sample sizes had no major impact on prediction. Interestingly, by

Figure 2



Prediction ability (measured as AUC) across time. Graphical representation of the change in the area under the curve (AUC) for five WTCCC diseases between 2008 and 2014. Error bars represent the standard error of 50 resamplings.

Figure 3



Average sample sizes of GWAS studies across time. Average sample size for the studies in the GWAS Catalog for each of the five diseases. This average is calculated in a cumulative manner, that is, the value for a year is the average of all the studies since 2008 to that year.

2014 studies on T2D or CAD were almost double in sample size than those for BD, but the AUCs by 2014 were similar. These observations suggest that, for some diseases, a sample size increase in one, or perhaps two orders of magnitude may unveil most of the relevant loci and would allow the study of their evolution at the molecular level. For other pathologies, in contrast, it is quite possible that molecular evolutionary tools cannot be deployed unless their power to detect tiny and complex signals of selection is drastically improved.

Concluding remarks

In the immediate future great developments in EGM do not seem easy. Two major conditions for its progress are not met: neither are we able to correctly classify patients as healthy or sick according to their genetic information; nor have we a clear idea of what has been the fitness impact of being sick. As it is often the case, treasures are buried deep, and the promising fruits of EGM will be difficult to reap. Still, as it is also frequent, progress in Science often comes from unexpected sources and it is quite possible that novel, and as yet unidentified, ideas or approximations contribute to the advancement of EGM. Given the stakes, they would be most welcome.

Acknowledgements

This work has been supported by the Spanish Multiple Sclerosis Network (REEM), of the Instituto de Salud Carlos III (RD12/0032/0011) to AN; by the Spanish Government Grant BFU2012-38236 to AN and by FEDER. We thank David Allen Hughes for generously reading and improving the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ewald PW: **Evolutionary biology and the treatment of signs and symptoms of infectious disease**. *J Theor Biol* 1980, **86**:169-176.

2. •• Williams GC, Nesse RM: **The dawn of Darwinian medicine**. *Q Rev Biol* 1991, **66**:1-22.
The seminal paper that introduced the concept of evolutionary, or Darwinian, medicine. The authors argue the benefits of including evolutionary biology in medical school curricula.

3. Wilson TW, Grim CE: **Biohistory of slavery and blood pressure differences in blacks today**. *Hypertension* 1991, **17**:122-128.

4. Kluger MJ: **Is fever beneficial?** *Yale J Biol Med* 1986, **59**:89-95.

5. Kluger MJ, Kozak W, Conn CA, Leon LR, Soszynski D: **The adaptive value of fever**. *Infect Dis Clin North Am* 1996, **10**:1-20.

6. Merlo LMF, Pepper JW, Reid BJ, Maley CC: **Cancer as an evolutionary and ecological process**. *Nat Rev Cancer* 2006, **6**:924-935.

7. Kirkwood TBL: **Understanding the odd science of aging**. *Cell* 2005, **120**:437-447.

8. •• Williams GC: **Pleiotropy, natural selection and the evolution of senescence**. *Evolution (NY)* 1957, **11**:398-411.
The first formal proposal of the pleiotropic theory of senescence which suggests that mutations that increase disease risk late in life may be maintained in populations because they increase individual fitness earlier in life.

9. Guttmacher AE, Collins FS: **Genomic medicine — a primer**. *N Engl J Med* 2002, **347(19)**:1512-1520.

10. Feero WG, Guttmacher AE, Collins FS: **Genomic medicine — an updated primer**. *N Engl J Med* 2010, **362(21)**:2001-2011.

11. Amiel J: **Study of leucocyte phenotypes in Hodgkin's disease**. In *Histocompatibility Testing.* Edited by Curtoni E, Mattiuz P, Munskgaard TR. 1967:79-81.

12. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases**. *Nat Genet* 2008, **40**:695-701.

13. Singal D, Blajchman M: **Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus**. *Diabetes* 1973, **22**:429-432.

14. The Huntington's Disease Collaborative Research Group: **A novel gene containing a trinucleotide that is expanded and unstable on Huntington's disease chromosomes**. *Cell* 1993, **72**:971-983.

15. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou J *et al.*: **Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA**. *Science* 1989:245.

16. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

17. **Applications of next generation sequencing**. *Nat Rev Genet [J Ser]* 2014.

18. Dixon-Salazar TJ, Silhavy JL, Udpa N, Schroth J, Bielas S, Schaffer AE, Olvera J, Bafna V, Zaki MS, Abdel-Salam GH *et al.*: **Exome sequencing can improve diagnosis and alter patient management**. *Sci Transl Med* 2012, **4**:138.

19. Nesse R, Bergstrom C: **Making evolutionary biology a basic science for medicine**. *Proc Natl Acad Sci U S A* 2010, **107(1)**.

20. Stearns SC, Nesse RM, Govindaraju DR, Ellison PT: **Evolution in health and medicine Sackler colloquium: evolutionary perspectives on health and medicine**. *Proc Natl Acad Sci U S A* 2010, **107**:1691-1695.

21. •• Antolin MF, Jenkins KP, Bergstrom CT, Crespi BJ, De S, Hancock A, Hanley KA, Meagher TR, Moreno-Estrada A, Nesse RM *et al.*: **Evolution and medicine in undergraduate education: a prescription for all biology students**. *Evolution* 2012, **66**:1991-2006.
The paper reviews the key evolutionary biology principles that would benefit medicine courses. Additionally, ten straightforward examples of connections between evolutionary biology and medicine are examined in detail.

22. Crespi BJ: **The emergence of human-evolutionary medical genomics**. *Evol Appl* 2011, **4**:292-314.

23. Luzzato L, Usanga EA, Shunmugam R: **Glucose-6-Phosphate Dehydrogenase deficient red cells: resistance to infection by malarial parasites**. *Science* 1969, **164**:839-842.

24. Allison AC: **Protection afforded by sickle-cell trait against subtertian malarial infection**. *Br Med J* 1954, **1**:290-294.

25. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I: **Identification of a variant associated with adult-type hypolactasia**. *Nat Genet* 2002, **30**:233-237.

26. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M *et al.*: **Convergent adaptation of human lactase persistence in Africa and Europe**. *Nat Genet* 2007, **39**:31-40.

27. Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA *et al.*: **Genetic origins of lactase persistence and the spread of pastoralism in Africa**. *Am J Hum Genet* 2014, **94**:496-510.

28. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L *et al.*: **Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes**. *Am J Hum Genet* 1998, **62**:1507-1515.

29. Nesse RM: **On the difficulty of defining disease: a Darwinian perspective**. *Med Health Care Philos* 2001, **4**:37-46.

30. Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL: **Common risk alleles for inflammatory diseases are targets of recent positive selection**. *Am J Hum Genet* 2013, **92**:517-529.

31. • Neel JV: **Diabetes mellitus: a ''thrifty'' genotype rendered detrimental by ''progress''?** *Am J Hum Genet* 1962, **77**:694-703 [discussion 692–3].

The introduction of the ''thrifty'' genotype hypothesis used explain the high incidence of type 2 diabetes mellitus in human populations. Neel argues that genotypes favouring the storage of carbohydrates might have been beneficial during early human evolution as protection against periods of famine. In modern environments where abundant resources are available, the once thrifty genotype is now deleterious and can result in phenotypes like diabetes.

32. Di Rienzo A, Hudson RR: **An evolutionary framework for common diseases: the ancestral-susceptibility model**. *Trends Genet* 2005:21.

33. Williams AL, Jacobs SBR, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Márquez-Luna C, García-Ortíz H, Gómez-Vázquez MJ, Burtt NP, Aguilar-Salinas CA *et al.*: **Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico**. *Nature* 2014, **506**:97-101.

34. Braude S, Tang-Martinez Z, Taylor GT: **Stress, testosterone and the immunoredistribution hypothesis**. *Behav Ecol* 1995, **10**:333-350.

35. Bribiescas R, Ellison PT: **How hormones mediate trade-offs in human health and disease**. In *Evolution in health and disease*. Edited by Stearns S, Koella J. Oxford Univ. Press; 2007:77-94.

36. Murray CJL: **Quantifying the burden of disease: the technical basis for disability-adjusted life years**. *Bull World Health Organ* 1994:72.

37. Institute for Health Metrics and Evaluation: http://www.healthdata.org [accessed on 03.05.14].

38. Leeder R, Farber E: **The variable incidence of psoriasis in sub-Saharan Africa**. *Int J Dermatol* 1997:36.

39. Karlsson EK, Kwiatkowski DP, Sabeti PC: **Natural selection and infectious disease in human populations**. *Nat Rev Genet* 2014, **15**:379-393.

40. Marigorta UM, Navarro A: **High trans-ethnic replicability of GWAS results implies common causal variants**. *PLOS Genet* 2013, **9**:e1003566.

41. •• Nesse RM, Williams GC: **Why we get sick?** *Times Books* 1994.

Must read! This book reviews the concept of Darwinian medicine for both the general public and those in scientific community unfamiliar with some major evolutionary concepts. It also advances some evolutionary explanations for certain ''modern'' diseases.

42. Nesse RM, Stearns SC: **The great opportunity: evolutionary applications to medicine and public health**. *Evol Appl* 2008, **1**:28-48.

43. Lewontin R: *The Genetic Basis of Evolutionary Change*. Columbia University Press; 1974.

44. Culotta E, Pennisi E: **Evolution in action**. *Science* 2005:310.

45. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data**. *Genome Res* 2005, **15**:1566-1575.

46. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O *et al.*: **A composite of multiple signals distinguishes causal variants in regions of positive selection**. *Science* 2010, **327**:883-886.

47. •• Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery**. *Am J Hum Genet* 2012, **90**:7-24.

Since mid-late 2000s, GWAS have been the preferred way to correlate genetic loci to diseases. In this article Peter Visscher and colleagues review what was learned during the first five years of GWAS including the main findings and the major problems.

48. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L *et al.*: **The NCBI dbGaP database of genotypes and phenotypes**. *Nat Genet* 2007, **39**:1181-1186.

49. • WTCCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls**. *Nature* 2007, **447**:661-678.

The first large-scale GWAS published that helped to shape standards in the field. The WTCC Consortium analyzed seven diseases belonging to three categories (mental, autoimmune & inflammatory, and metabolic) and found 50 associated loci. Most of these results are still replicating in nowadays GWAS.

50. Evans DM, Visscher PM, Wray NR: **Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk**. *Hum Mol Genet* 2009, **18**:3525-3531.

51. • Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H: **Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies**. *Nat Genet* 2013, **45**:400-405 405e1–3.

Expectations on how phenotype risk prediction will change as average sample sizes increase with time.

52. Hindorff LA (European Bioinformatics Institute), Macarthur J (European Bioinformatics Institute), Morales J, Junkins H, Hall P, Klemm A. A catalog of published genome-wide association studies [accessed 04.01.14].

53. Olalde I, Sánchez-Quinto F, Datta D, Marigorta UM, Chiang CWK, Rodríguez JA, Fernández-Callejo M, González I, Montfort M, Matas-Lalueza L *et al.*: **Genomic analysis of the blood attributed to Louis XVI (1754–1793), king of France**. *Sci Rep* 2014, **4**:4666.

54. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061-1073.

55. R Development Core Team: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2005, .

56. • Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N: **Estimation of effect size distribution from genome-wide association studies and implications for future discoveries**. *Nat Genet* 2010, **42**:570-575.

This article proposes methods to determine the genomic architecture of a phenotype and estimate the number of undiscovered associated markers for said phenotypes based on the number of loci previously discovered and on the distribution of their effect sizes and frequencies.

57. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R**. *Bioinformatics* 2005, **21**:3940-3941.

58. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mentch F, Van Steen K, Visscher PM *et al.*: **Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease**. *Am J Hum Genet* 2013, **92**:1008-1012.

59. • Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: **Pitfalls of predicting complex traits from SNPs**. *Nat Rev Genet* 2013, **14**:507-515.

A review of biases and possible errors of phenotype predictions when using genotype data. The authors elegantly illustrate each type of bias with real examples from the literature.

60. Wray NR, Yang J, Goddard ME, Visscher PM: **The genetic interpretation of area under the ROC curve in genomic profiling**. *PLOS Genet* 2010, **6**:e1000864.