Novelli et al. Top Italian Scientists Journal; 2024;1(1)
https://doi.org/10.62684/YEWJ9912

Top Italian Scientists Journal

Open Access | Article | Biomedical Sciences

ISSN 3033-5132

The "Black Swan Principle" and the Genetics of Complex Diseases

Giuseppe Novelli[a] , **Juergen K V Reichardt**[b]

[a]Department of Biomedicine and Prevention, School of Medicine and Surgery, Tor Vergata University of Rome, Via Montpellier 1, 00133, Rome, Italy and Department of Pharmacology, School of Medicine, University of Nevada, 89557, Reno, NV, USA.

[b]Australian Institute of Tropical Health and Medicine, James Cook University, Smithfield, QLD, 4878, Australia.

Correspondence to: novelli@med.uniroma2.it

Published: 13 Jan 2024

Abstract

The black swan principle is a philosophy theory created by Nassim Nicholas Taleb that seeks to explain rare and unpredictable events, appearances that seem to defy logic or rational explanation (1). These events, termed "Black Swans," have been observed in various domains, including finance, public administration, infectious diseases, and ecology (2-4). The concept of Black Swans has gained recently, significant attention in academia and practice due to its relevance in understanding extreme and rare occurrences (5-7). The "black swan" concept has been used in genetics for the unexpected developments that genome sequencing would reveal and which could have consequences for

healthcare systems (e.g., increase in often unnecessary and inappropriate diagnostic investigations, increase in non-patients, etc.) (8-10).

Keywords: Black Swan Principle; Genetics of Complex Diseases

Complex Diseases

Identifying a genetic mutation linked to a disease like osteoporosis could be the first step toward developing a treatment. However, the next step is to demonstrate that blocking a receptor or antagonizing an enzyme generates a phenotypic and possibly even medical effect. But this is not yet possible for the many complex diseases that affect humans. Complex diseases are multifactorial conditions influenced by a combination of genetic, environmental, and lifestyle factors. These diseases often involve intricate interactions among multiple genes and environmental elements, making their etiology and pathogenesis challenging to unravel (11-13). The term "complex disease" encompasses a wide range of conditions, including autoimmune diseases, psychiatric disorders, metabolic diseases, and various chronic illnesses (13, 14). The genetic basis of complex diseases is often characterized by the involvement of multiple intermediate phenotypes, each with a component of quantitative inheritance, contributing to the overall pathogenesis (13). Complex diseases, unlike inherited or Mendelian which are generally single gene defects, are characterized by a quantitative phenotypic distribution, based on the interactive action of multiple genes, each of which would act with a small additive effect (polygenic inheritance) (15). This quantitative inheritance model is widely accepted to explain the transmission mechanism of many diseases common, which contribute significantly to the morbidity and mortality of the population (including various congenital defects and adult-onset pathologies, such as diabetes, hypertension, stroke, Alzheimer's disease). The current idea is that different genes, implicated in susceptibility, trigger the triggering effect of some harmful environmental factors. This idea has found confirmation in recent years in thousands of genomic studies, which have allowed to identify over 100,000 genetic variations associated with diseases and complex traits, which in recent years have been used to develop the so-called Polygenic Risk Scores (PRS), which measure the effects and therefore the weight of genomic variations on the phenotype, i.e. the susceptibility to developing a certain trait or disease(16). Variants needed to define a PRS are identified by genome-wide association studies (GWAS) (17). GWAS have significantly contributed to understanding the genetic basis of complex diseases. These studies have identified numerous susceptibility loci for various complex diseases, shedding light on the genetic architecture of

multifactorial disorders (18). GWAS have identified genomic biomarkers associated to diseases such as Crohn's disease, bipolar disorder, pulmonary diseases, schizophrenia, and age-related macular degeneration (19-22). However, it is important to note that the identified loci often account for only a small portion of the heritability of these diseases, leaving a substantial portion of the genetic contribution unexplained ("missing heritability") (18). The "missing heritability" problem has been a subject of extensive research, and it has been attributed to the limited power of traditional linkage studies in detecting variants of modest effect (18, 23). In addition, the majority of existing PRS were developed from European data with limited transferability to other populations (e.g. African populations) (24). Some populations such as African ones have different genetic backgrounds and a genomic architecture and organization which have often led to non-replicable GWAS results or to false or erroneous associations. It is no coincidence, in fact, that extensive biobanks are being developed such as the "All of US Research Program" in the USA which collects biological samples and clinical information from different ethnic groups (25). Human genome is proving to be much more complex and the genetic relationships between different ethnic groups are greatly affected by interactions between groups of people which can significantly modify the allelic frequencies of genes, favoring the selection of some alleles. Recently, it has been shown that people with European ancestry, who were previously treated as a genetically homogeneous group, have clear evidence of mixed genetic lineages, known as "admixture." Therefore, many GWAS-based association studies should be reviewed based on "mixing"(26).

The Black Swan and Rare Variants

GWAS are a valuable tool for understanding the biology of complex human traits and diseases, but associated variants rarely point directly to causal genes. The variants shown to be associated with a specific disorder, are very common in populations and are unlikely to demonstrate significant biochemical effects. GWAS, in fact, include variants that are shown to have only additive effects, excluding other types of genetic variations (e.g. rare variants, copy number variants). The single or combined effects (PRS) of the common variants used in GWAS are quite small, typically with odds ratios less than 1.5 and often up to 1.1 (27). It is possible as suggested by Greg Gibson (28) that a substantial portion of the variance for complex diseases is due to relatively highly penetrant rare variants, whose allele frequency is typically less than 1%, most of which are recently derived alleles in the human population. Rare variants with very low allele frequencies not included in GWAS could also have large effect sizes (28). The allotment of rare alleles in a population, precisely their peculiarities and characteristics, does not follow the distribution of the classical normal curve, used

by Falconer's model (29) which assumes the existence of alleles with additive effects (15, 23, 28). For this reason, bell curves are in my opinion imperfect in assessing genomic risk in complex and multifactorial diseases. The normal distribution ignores the impact of rare alleles, which are considered infrequent and therefore unlikely and therefore should not be used to predict predictive risks of disease. Similarly to what Taleb Nassim proposed for economic phenomena (30), the bell curve ignores large deviations, and rare mutations can be considered large deviations with significant biological effects. Rare variants have larger effect sizes and are more susceptible to population dynamics and genetic drift. However, identifying true associations of rare variants with a complex disease, is difficult due to small effect sizes, the presence of technical artifacts, and heterogeneity in population structure. The cost-effective sequencing of the human genome and exome has allowed in recent years the identification of many rare genetic variations associated with complex and multifactorial diseases as well as quantitative and continuous traits such as height, lipid levels, left-handedness, sleep-related traits (Table 1) (31-35). Compared to common variants, rare genetic variants are more likely to be functional (36) (37)and therefore can more easily lead to new biological and clinical insights. In many cases, the identification of rare alleles has led to understanding the pathogenesis of the disease and discovering new therapeutic targets. Very interesting was the discovery of rare alleles of genes associated with congenital defects of immunity in COVID-19 from SARS-CoV-2 infection (32, 38-41). The International Covid Human Genetic Effort Consortium (https://www.covidhge.com/) has demonstrated for the first time an enrichment in rare loss-of-function (38) variants at 13 human loci known to govern the production and regulation of interferon molecules and how these mutations, underlying life-threatening COVID-19 pneumonia in patients without prior serious infection(40). The presence of rare alleles in these genes are at the basis of the serious multisystem inflammatory disease of children which during the SARS-CoV-2 pandemic unfortunately led to the death of healthy children such as Zyrin Foots, 10, who died after a two-week battle with COVID-19 (https://www.newsweek.com/10-year-old-covid-dies-after-mom-given-choice-amputate-limbs-let-him-go-1639366)(42). The case of Zyrin Foots can be considered a black swan, a certainly unpredictable event due to the simultaneous presence of a rare genotype in a healthy boy who becomes infected with a new virus (SARS-CoV-2) which quickly leads to failure respiratory and death! The discovery of inborn errors and mechanisms underlying rare infections, which led to the identification of rare monogenic determinants in related common infections, allowed Casanova and Abel (31) to contrive the definition “rare to common” which demonstrates the direct link of rare alleles to complex diseases such as infectious diseases. It is therefore important to focus on the rare variants that make the difference and not on the common variants.

Future perspectives

Extended genome sequencing will allow us to continuously identify rare new variants in individuals and this will change the approach to complex and multifactorial diseases in the coming years. It is possible that the weight of rare variants on genomic analyzes will be combined with the PRS predictive score based on millions of SNPs to constitute a new and more precise associated genetic risk as hypothesized by Lali et al. (43). This new predictive tool will certainly be able to help us better understand the pathogenesis of complex diseases and produce new innovative drugs such as *evolocumab*, a monoclonal antibody that inhibits the expression of the *PCSK9* gene by reducing cholesterol levels in subjects with familial hypercholesterolemia. This extraordinary result was possible thanks to the presence of two rare mutations (p.Tyr142Ter and p.Cys679Ter) of PCSK9, in some normal black subjects, which led to a reduction in average LDL cholesterol and the risk of coronary heart disease(44).

The unique biological effects of rare alleles have now been found in the coding regions of analyzed genes. However, there is evidence that rare variants in non-coding regions could have a large impact on gene expression and disease (45, 46). Only the combined study between rare genetic variants and multi-omics data, including data on transcriptome, post-transcriptional regulation, epigenome, post-translational protein modification, metabolome, and microbiome, will contribute in the future to improve our understanding of biological burden and effects "black swans" of our genome. But how many "black swans" exist in our genome? Reading a person's DNA, we can find millions of common variants and at least 25,000-50,000 rare ones; of these at least 70-80 are new mutations, that is, they are not inherited from the parents (*de novo*). Furthermore, we can find hundreds of lost or duplicated DNA segments, of often unknown significance. For this reason, the American Society of Genetics and Genomics (47) has developed guidelines indicating around seventy genes to be looked at with greater attention (actionable genes), i.e. genes that have clinical interest and are susceptible to possible therapeutic actions (47). A recent Icelandic study and based on the analysis of approximately 60,000 individual genomes has made it possible to identify "actionable" genes in 4% of the people analyzed (48). An interesting aspect of this new study is the correlation with the average lifespan of carriers of these genes compared to non-carriers. On average, they observed that carriers of “actionable” genes associated with cancer had a shorter survival than non-carriers, by about three years. This study, although important, will have to be confirmed on other populations (Iceland is a genetically homogeneous population) and appropriately validated through the integration of family and behavioral data of the people analyzed (lifestyle, drugs used) before being able to use this information

for population screening. This confirms the decisive role of rare variants in determining or in any case directing a phenotype.

Is it therefore absurd to think of characterizing them all and providing accurate diagnoses and predictions of future disease risk? Why not consider developing AI systems that can consider all genetic variants across the entire genome, including structural variants such as copy number variations, insertions, inversions, and translocations along with the functional impact of each variant? Of course, the calculation must also consider electronic health records, including digital images, data from health monitoring devices and other environmental exposures. Is it science fiction? No, advances in AI software and hardware, particularly deep learning algorithms and the graphics processing units (GPUs) that power their training, mean a specific type of AI algorithm is possible soon. artificial intelligence known as deep learning is used to process large and complex genomic datasets [103]. But managing this wealth of information requires the development of new training programs based on innovation and application of knowledge.

Declarations

Acknowledgements

The studies of G.N. on complex diseases are supported by grants of HORIZON-HLTH-2021-DISEASE-04 program under grant agreement 01057100 (UNDINE) and HEAL ITALIA
Health Extended ALliance for Innovative Therapies, Advanced Lab-research, and Integrated Approaches of Precision Medicine, PNRR MUR, Mission 4 Component 2.

Conflict of Interest

The Authors declare that there is no conflict of interest.

Table 1. Complex Diseases and Continues Traits Associated with the Enrichment of Rare Variants

Phenotype	Gene
Human Handedness	***TUBB4B***
Neurodevelopmental disorders	***CUL3***
Severe adult-onset obesity	***BSN, APBA1***
Hidradenitis suppurativa	***PSTPIP1***
Nicotine addiction	***CHRNB2***

Alzheimer	***RELN, ABCA7***
Rheumatoid arthritis	***IL2RA, IL2RB. TYK2***
Age-related macular degeneration	***CF1,CFB, CEPT***
Schizophrenia	***SETD1A***
Orofacial clefts (OFCs)	***SEC24D***
Amyotrophic lateral sclerosis (ALS)	***SOD1, TARDBP, TBK1***
Hyperlipidemia	***LDLR, PCSK9, APOC3, ANGPLT3, ABCG5, NPC1L1***
Lupus	***TNFAIP3, STAT4, IL10, TRAF3IP2, HCP5***
Blood pressure	***KIF3B***
Diabetes	***GIGYF1***
Human height	***HMGA1, MIR497HG***
COVID-19	***IFNGR1, IFNGR2, IFNAR1, IFNAR2, IL12RB1, IRAK4, MYD88, STAT1 GOF, CXCR4, TBK1, TLR3, TLR7, IRF3, IRF7, IRF9***
Attention-deficit/hyperactivity disorder (ADHD)	***ASXL3, DOT1L, DIP2C, KDM2A, KDM1A, KMT2B, SETDB1, SLC22A23, COL4A3BP, DET1***
Sleep-related traits (sleep duration, insomnia symptoms, chronotype, daytime sleepiness, daytime napping, ease of getting up in the morning, snoring and sleep apnea)	***ST3GAL1, ANKRD12, PLEKHM1, ZBTB21, WDR59***

References

1. Hannabuss S. The Black Swan: The Impact of the Highly Improbable. Library Review. 2008.

2. Ponkin IV. "Black Swan" Event as Manifestation of Uncertainties in Public Administration. Mediterranean Journal of Social Sciences. 2019.

3. Velappan N, Davis-Anderson K, Deshpande A. Warning Signs of Potential Black Swan Outbreaks in Infectious Disease. Frontiers in Microbiology. 2022.

4. Wind TR, Rijkeboer MM, Andersson G, Riper H. The COVID-19 Pandemic: The 'Black Swan' for Mental Health Care and a Turning Point for E-Health. Internet Interventions. 2020.

5. Parameswar N, Chaubey A, Dhir S. Black Swan: Bibliometric Analysis and Development of Research Agenda. Benchmarking an International Journal. 2021.

6. Simianer H, Reimer C. COVID-19: a "black swan" and what animal breeding can learn from it. Anim Front. 2021;11(1):57-9.

7. Vacante M, D'Agata V, Motta M, Malaguarnera G, Biondi A, Basile F, et al. Centenarians and supercentenarians: a black swan. Emerging social, medical and surgical problems. BMC Surgery. 2012;12(1):S36.

8. Doyle S. Waiting for medicine's black swans. CMAJ. 2012;184(5):E246-7.

9. Jonsen AR, Durfy SJ, Burke W, Motulsky AG. The advent of the "unpatients'. Nat Med. 1996;2(6):622-4.

10. Kish LJ, Topol EJ. Unpatients-why patients should own their medical data. Nat Biotechnol. 2015;33(9):921-4.

11. Novelli G, Biancolella M, Latini A, Spallone A, Borgiani P, Papaluca M. Precision Medicine in Non-Communicable Diseases. High Throughput. 2020;9(1).

12. Novelli G, Cassadonte C, Sbraccia P, Biancolella M. Genetics: A Starting Point for the Prevention and the Treatment of Obesity. Nutrients. 2023;15(12).

13. Blanco-Gómez A, Castillo-Lluva S, Sáez-Freire MdM, Hontecillas-Prieto L, Mao JH, Castellanos A, et al. Missing Heritability of Complex Diseases: Enlightenment by Genetic Variants From Intermediate Phenotypes. Bioessays. 2016.

14. Kim YH, Kim SI, Park B, Lee ES. Clinical Characteristics of Psoriasis for Initiation of Biologic Therapy: A Cluster Analysis. Annals of Dermatology. 2023.

15. Levitan M. Textbook of human genetics. 3rd. ed. Oxford U.P1988.

16. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28(R2):R133-r42.

17. Blumberg RS, Dittel BN, Hafler DA, Herrath Mv, Nestle FO. Unraveling the Autoimmune Translational Research Process Layer by Layer. Nature Medicine. 2012.

18. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the Missing Heritability of Complex Diseases. Nature. 2009.

19. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-Wide Association Study Identifies New Susceptibility Loci for Crohn Disease and Implicates Autophagy in Disease Pathogenesis. Nature Genetics. 2007.

20. Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landén M, et al. Genome-Wide Association Study of 40,000 Individuals Identifies Two Novel Loci Associated With Bipolar Disorder. 2016.

21. Ratnapriya R. Transcriptomics Insights Into Interpreting AMD-GWAS Discoveries for Biological and Clinical Applications. Journal of Translational Genetics and Genomics. 2022.

22. Hindorff LA, Sethupathy P, Junkins H, Ramos EM, Mehta JP, Collins FS, et al. Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits. Proceedings of the National Academy of Sciences. 2009.

23. Colona VL, Biancolella M, Novelli A, Novelli G. Will GWAS eventually allow the identification of genomic biomarkers for COVID-19 severity and mortality? J Clin Invest. 2021;131(23).

24. Fatumo S, Sathan D, Samtal C, Isewon I, Tamuhla T, Soremekun C, et al. Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. Genome Med. 2023;15(1):87.

25. Ginsburg GS, Denny JC, Schully SD. Data-driven science and diversity in the All of Us Research Program. Sci Transl Med. 2023;15(726):eade9214.

26. Gouveia MH, Bentley AR, Leal TP, Tarazona-Santos E, Bustamante CD, Adeyemo AA, et al. Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. Nat Commun. 2023;14(1):6802.

27. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev. 2008;18(3):257-63.

28. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135-45.

29. Falconer DS. Introduction to quantitative genetics: Pearson Education India; 1996.

30. Taleb N. The black swan : the impact of the highly improbable. Penguin Books ed. London: Penguin; 2008. xxviii, 366 p. p.

31. Casanova JL, Abel L. From rare disorders of immunity to common determinants of infection: Following the mechanistic thread. Cell. 2022;185(17):3086-103.

32. Cobat A, Zhang Q, Abel L, Casanova JL, Fellay J. Human Genomics of COVID-19 Pneumonia: Contributions of Rare and Common Variants. Annu Rev Biomed Data Sci. 2023;6:465-86.

33. Fei C-J, Li Z-Y, Ning J, Yang L, Wu B-S, Kang J-J, et al. Exome sequencing identifies genes associated with sleep-related traits. Nature Human Behaviour. 2024.

34. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biology. 2017;18(1):77.

35. Pounraja VK, Girirajan S. A general framework for identifying oligogenic combinations of rare variants in complex disorders. Genome Res. 2022;32(5):904-15.

36. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216-20.

37. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. J Hum Genet. 2021;66(1):11-23.

38. Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 2020;370(6515):eabd4570.

39. Notarangelo LD, Bacchetta R, Casanova JL, Su HC. Human inborn errors of immunity: An expanding universe. Sci Immunol. 2020;5(49).

40. Matuozzo D, Talouarn E, Marchal A, Zhang P, Manry J, Seeleuthner Y, et al. Rare predicted loss-of-function variants of type I IFN immunity genes are associated with life-threatening COVID-19. Genome Med. 2023;15(1):22.

41. Biancolella M, Colona VL, Luzzatto L, Watt JL, Mattiuz G, Conticello SG, et al. COVID-19 annual update: a narrative review. Hum Genomics. 2023;17(1):68.

42. Lee D, Le Pen J, Yatim A, Dong B, Aquino Y, Ogishi M, et al. Inborn errors of OAS-RNase L in SARS-CoV-2-related multisystem inflammatory syndrome in children. Science. 2023;379(6632):eabo3627.

43. Lali R, Chong M, Omidi A, Mohammadi-Shemirani P, Le A, Cui E, et al. Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. Nature Communications. 2021;12(1):5852.

44. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006;354(12):1264-72.

45. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. Nat Genet. 2019;51(9):1349-55.

46. Wakeling MN, Owens NDL, Hopkinson JR, Johnson MB, Houghton JAL, Dastamani A, et al. Non-coding variants disrupting a tissue-specific regulatory element in HK1 cause congenital hyperinsulinism. Nat Genet. 2022;54(11):1615-20.

47. Miller DT, Lee K, Abul-Husn NS, Amendola LM, Brothers K, Chung WK, et al. ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023;25(8):100866.

48. Jensson BO, Arnadottir GA, Katrinardottir H, Fridriksdottir R, Helgason H, Oddsson A, et al. Actionable Genotypes and Their Association with Life Span in Iceland. New England Journal of Medicine. 2023;389(19):1741-52.