Artificial Intelligence in the Life Sciences 1 (2021) 100003

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Opinion

Combinatorial analytics: An essential tool for the delivery of precision medicine and precision agriculture ☆

Steve Gardner

PrecisionLife Ltd, United Kingdom

1. Introduction

Feb 2021 marked the 20th anniversary of the publication of two seminal papers that revealed the first findings of the Human Genome Project [1–3]. This provides an opportunity to reflect on how knowledge of the human genome and its use in developing more effective treatments for diseases has progressed (or not) [4,5], and which challenges remain to be overcome, potentially enabled by a new generation of AI methods.

Genomic and precision medicine have been revolutionary in improving treatment options in cancer and rare diseases. Optimism for future progress is however tempered by unexpected failures to achieve the same level of insights and impact across all diseases. Outside of cancer and rare disease, the impact of genomic medicine has been lower than originally anticipated. This is true even in disorders like Alzheimer's disease where key genetic risk factors such as *ApoE* have long been identified [6,7], but which has nonetheless suffered from an almost total% clinical trials failure rate [54] .

This failure has significantly limited the extension of precision medicine approaches into more complex chronic diseases such as neurodegenerative, metabolic, immunological, respiratory, cardiovascular and neuropsychiatric disorders. These diseases have large pockets of unmet medical need, impose a huge socioeconomic burden, and account for around 85% of healthcare costs, but have so far been impacted relatively little by genomic insights.

In large part, this is a consequence of the biological assumptions underpinning some of the most ubiquitous analytical methods that we use, including Genome Wide Association Studies (GWAS). GWAS analyze the distribution of one mutation at a time to find those overrepresented in a case vs a control population [9]. GWAS are useful for finding the low-hanging fruit of the simpler forms of disease-variant associations [10]. They work relatively well for cancer and rare diseases as these conditions are often caused by single mutations occurring in protein coding regions, which produce a large effect size that is relatively easy to spot in a case:control study.

GWAS and techniques based on their outputs such as polygenic risk scores [11] and gBLUP methods [12] are however inherently incapable of fully reflecting the polygenic and pleiotropic complexity of chronic disease endotypes and phenotypic traits. Their analytical approaches and the assumptions underpinning them (e.g., the linear additivity of features, the relative importance of rare variants, principles of Mendelian inheritance and the presumed accuracy of a single disease diagnosis across a heterogeneous patient population) impose a detection level limit that we need to overcome to deliver the benefits of genomics more widely in precision medicine and selective breeding applications in crops and livestock.

2. Respecting the biological complexity of chronic diseases

Chronic diseases are heterogenous and polygenic; they exhibit a wide spectrum of symptoms, their diagnoses contain multiple patient sub-groups with different disease aetiologies, severities and therapy responses, and many are relatively late onset conditions with significant clinical and epidemiological confounders.

This complexity means that a simple 'single feature at a time' approach like GWAS is unhelpfully divorced from the biological causes of chronic diseases. It makes it hard to accurately identify a set of causal variants/genes, and the results of GWAS may be incomplete or even suspect in chronic diseases.

For example, many of the 'causal' genes originally found by GWAS in cardiovascular disease (CVD) have subsequently been re-evaluated and their disease relevance repudiated [13–15]. Diseases like CVD and Alzheimer's are now thought to be caused both by ultra-rare Mendelian variants with high effect sizes affecting a small minority of patients and by multiple other genetic and non-genetic factors that modulate most patients' endotypes via complex mechanisms including incomplete dominance, variable expressivity, epistasis and other combinatorial non-linear interactions [16].

This is not surprising – the iconic IUBMB-Nicholson Metabolic Pathways Chart [17] reinforces the inherent complexity of cellular biology. This beautiful and crowded chart reveals some of the complex interplay of intersecting and interacting networks of genetic, epigenetic and metabolic control factors that control both normal cellular function and disease processes.

Feedback (and feedforward) regulated processes are subtly balanced in normal metabolism, acting to amplify or inhibit pathways depending on their metabolic context and conditions [18]. They contain rapid feedback mechanisms such as the modulation of protein activity via post-translational modifications or the allosteric binding of small molecule

☆ Opinion Piece for Artificial Intelligence in Life Sciences

E-mail address: steve@precisionlife.com

https://doi.org/10.1016/j.ailsci.2021.100003
Received 26 April 2021; Received in revised form 26 April 2021; Accepted 26 April 2021
Available online 14 May 2021

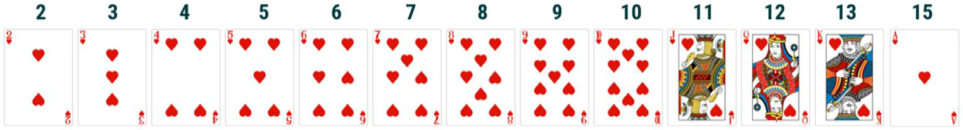


Fig. 1. Face value scores for cards in a deck.

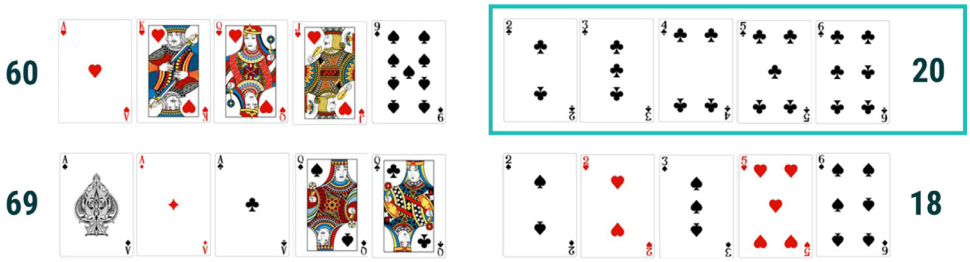


Fig. 2. Four poker hands with their total card face values (winning hand highlighted top right).

ligands, and slower ones such as changes in gene expression. These are however predominantly dynamic and non-linear [19].

This non-linearity of the underlying biology is an inconvenient truth that has been largely overlooked in the field of genomic analysis. In Alzheimer's disease alone 1,000 genetic association studies and 75 GWAS have been performed [20] resulting in just 29 risk genes identified [21] and no new therapies approved. Some practitioners will simultaneously acknowledge the multi-factorial nature of complex diseases such as Alzheimer's or type-II diabetes but ignore the importance of epistatic and other non-linear effects, preferring to search ever larger datasets for ever rarer variants to attempt to improve their models infinitesimally [22,23]. While there will always be new rare variants to be discovered, these are not the only or perhaps even the most important signals missing from our models of disease.

In fact, the non-linear additive effects of feedback loops, epistasis and other interactions are key for understanding chronic disease biology and how it impacts patient populations [24,25]. They enable effective patient stratification, and help overcome issues of poor risk prediction models, missing heritability and unexplained genetic variance, without resorting to an unhelpfully imprecise infinitesimal or omnigenic view of chronic disease or other complex biological traits [26,27].

3. Single feature vs combinatorial analytics

Single feature based methods, however sophisticated, cannot capture or represent the inherent non-linearity of biological interactions. The challenge is analogous to guessing the winning hand in a game of poker. A single feature method might add together the face value of the cards in each hand (similar to the odds ratios for single SNPs) using a scoring scheme like that shown in Fig. 1.

If we dealt hands of five cards to four players (as shown in Fig. 2) and added up their face values we would find that they have face value totals of 60, 69, 20 and 18 respectively. Working from this limited information, and knowing that higher cards win over lower cards in otherwise equivalent hands, you would be justified in predicting one of the highest scores on the left as the winning hand.

In fact, the hand in the top right is a straight flush, the second highest hand in poker and, in spite of a face value score of just 20, the winner. In this analogy, the combination of cards is much more important to (and predictive of) the outcome than either the individual cards in the hand or their summed face values. This signal can only be found using an inherently combinatorial approach.

The same is true in genetics, where we have based most of our target discovery analyses and disease/trait prediction models, such as polygenic risk scores (PRS), on the odds ratios of single mutations' disease associations calculated individually in isolation. PRS attempt to synthesize a patient's full genotype data into a single score that predicts genetic risk for a disorder or trait [28]. gBLUPs do a similar job for the heritability of phenotypic traits in livestock and crop species accounting for covariance based on genetic factors in a mixed model framework [29].

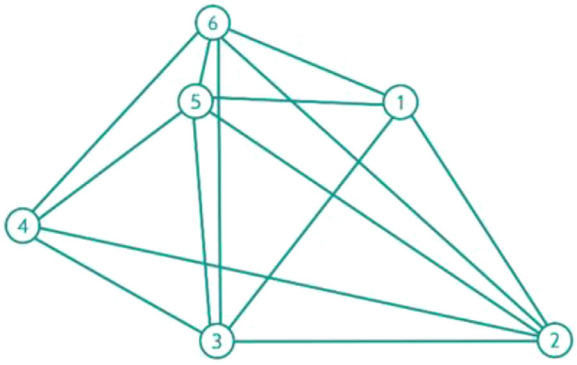


Fig. 3. Conceptual graph of a disease associated combination of 6 SNP genotypes. SNP genotypes labelled 1-6 co-occur in multiple patients, indicated by the interconnections between the SNPs.

PRS are usually calculated using a selected set of the most significantly disease associated SNPs (from a GWAS), summing an individual's risk alleles using Bayesian or linear regression methods usually weighted by allele frequency and effect size. As the odds ratios of GWAS derived SNPs are assumed to be linearly additive in most polygenic risk scores for complex diseases, the limitations of GWAS discussed above, and the inability to stratify the specific disease associated factors relevant to different patient sub-groups also adversely affect the predictive accuracy of such scores.

4. Combinatorial analytics reveals chronic disease biology

Combinatorial analysis on the other hand attempts to identify combinations of features that when they occur together are highly associated with a phenotype (occurring in many cases and relatively few controls). The features may include SNPs, CNVs and/or other clinical, transcriptomic, epigenetic, epidemiological, environmental or other factors. The structure of a sixth order phenotype associated combination can be seen in Fig. 3.

Using combinatorial analysis we can find additional signal in patient datasets that is invisible to existing GWAS and other genetic analysis methods [30–32]. This power to find this 'missing' signal is illustrated by a meta-analysis of the various large-scale studies that have been performed into the genetic factors underpinning COVID-19 host response as it relates to disease susceptibility and severity.

COVID-19 surprised the medical community by the range of its symptoms. Rather than a pure viral infection and inflammasome/cytokine response, the disease has had widespread, potentially long-term effects across a range of tissues [33,34]. As data became available at the beginning of the pandemic, GWAS analysis of very large patient populations were run by various groups.

A GWAS study involving 1,131 severe patients and 15,434 mild controls identified one locus associated with high risk of developing severe COVID-19 around the ABO blood group gene and another locus on chromosome 3 [35]. This study was then extended in a global effort that ran GWAS on genomic data from 13,641 severe disease patients and over 2 million controls, identifying four genome-wide significant loci that are associated with SARS-CoV-2 infection and 11 associated with severe manifestations of COVID-19 [55].

Most of these correspond to previously documented genes with associations to lung or autoimmune and inflammatory diseases. The symptomology of COVID-19 however extends much further than can be explained by these findings into neurological, coagulation and cardiovascular, renal and other consequences beyond inflammation driven disease.

In contrast, a previous study using a combinatorial analytics approach had been run several months earlier on the very first COVID-19 datasets available from UK Biobank [38,39], with just a few hundred severe patients, while controlling for all the existing known predisposing co-morbidities [40]. This study, working off much smaller datasets than the GWAS studies, identified 156 severe disease associated loci that mapped to 68 protein coding genes, spread across a range of mechanisms.

Many novel targets were identified that are involved in key severe COVID-19 pathology and disease mechanisms, including production of pro-inflammatory cytokines, endothelial cell dysfunction, lipid droplets, neurodegeneration and viral susceptibility factors. The novel disease associated mechanisms identified in this genetics based study were replicated and validated by subsequent combinatorial analysis of de-identified COVID-19 patient health records (non-genetic data comprising longitudinal diagnosis, claims and lab data) in the UnitedHealth Group COVID-19 Data Suite [41]. Several of the novel targets have also been subsequently validated in collaborative studies into drug repurposing using viral plaque assays and other disease models [42,43].

5. Patient stratification in chronic disease

Chronic diseases often present a spectrum of symptoms with varied treatment responses across patients who share a diagnosis. There may be multiple pathways leading to the key diagnostic phenotypes – for example bronchoconstriction and wheezing in asthmatics, or cycles of mania and depression in bipolar patients, and variations in response may reflect different disease aetiologies and influences. Because of this diversity within a disease population, GWAS (whose signals must be significant across the whole population) often cannot find all of the causes of a disease, or accurately stratify patient subgroups within the whole population.

Combinatorial analysis can help with patient stratification, distinguishing the different disease causes for multiple subpopulations (endotypes) even in highly heterogenous diseases such as asthma [44]. Asthma patients can be categorized into two molecular phenotypes; those with high T-helper cell type 2 (eosinophilic/T2) expression and those without (non-T2). The latest developments in treatment options for severe, uncontrolled asthma have focused on the T2 cytokine pathway, but these exhibit no benefit to patients with non-T2 disease [45]. It was however unclear whether the sub-groups represented different disease populations (endotypes) or simply different stages of disease, epidemiological sensitization factors or environmental triggers.

Combinatorial analysis of 7,094 T2 cases and 15,071 non-T2 cases enabled the identification of combinations (up to 5th order) of features associated with the patients' asthma diagnosis. By tracking which patients contain the SNPs identified in these combinations, the SNPs can be clustered to show how they are distributed across the patient population and how they correlate with the molecular phenotypes of the clearly distinguishable patient subgroups.

This analysis generates detailed disease insights showing mechanistically explanatory differences between the two endotypes (Fig. 4) and several novel druggable targets for the non-T2 sub-group [46]. Many of the significant genes found in the T2 subtype are involved in well-known processes such as Type 2 immune response, positive regulation of IL-5 and IL-13 production and lymphocyte differentiation, whereas genes associated with the non-T2 population, are involved in non-immune pathways, including fatty acid synthesis, LDL oxidation and modulators of GABA, purinergic and glutamate signalling.

6. Combinatorial analytics for complex health, production and fertility traits in livestock and crops

The same limitations we find in GWAS also affect our ability to identify features with significant predictive value for complex health, productivity and fertility traits in selective breeding applications in livestock and plant genetics. In animal and crop applications the infinitesimal model has been central to selective breeding applications since the late 19th century. Fisher showed in 1919 [47] that phenotypic trait values and their (co)variances can be broken down into components within pedigree structures, and that some level of phenotypic variance is consistent with multiple Mendelian factors acting additively. This limited signal has been good enough to improve the phenotype within heavily in-bred pedigrees across multiple species.

Classical quantitative genetics predicts traits as a linear combination of a set of variance components, with their coefficients determined by allele frequency. Selection – for example within selective breeding programmes – can change the trait mean, at a rate proportional to the additive genetic variance. However, when the trait depends on large numbers of genes, each of which makes a small contribution, selection has a negligible effect on the variance contributed by any individual locus. This is a clear indication of the non-linearity of the underlying biology.

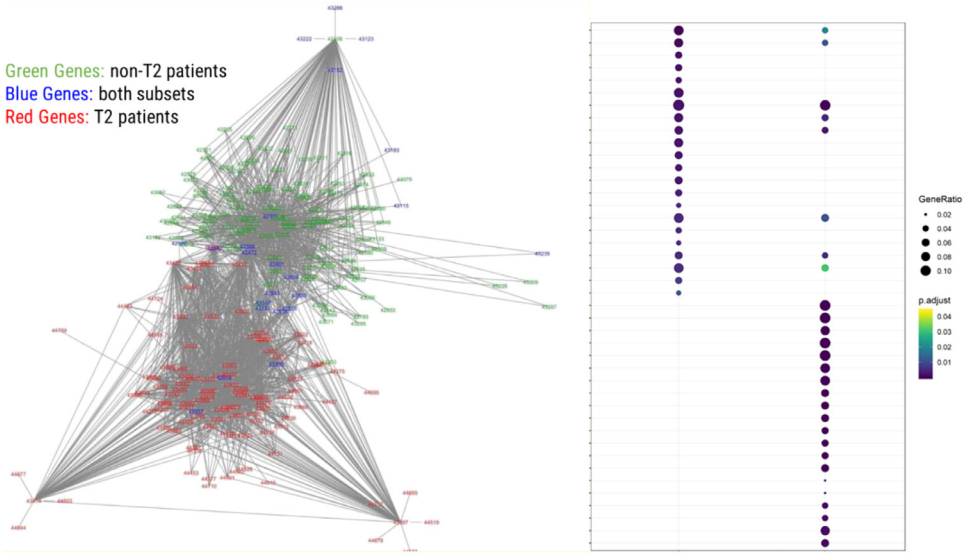
Existing tools obviously have some explanatory power but are failing to find all of the available signal. Even within very heavily inbred populations with highly selected phenotypes, there is considerable room for non-linear interactions to be used to improve genetic prediction models that have previously been based on GWAS results.

For example, egg weight is an economically important trait for laying hens. Egg weight is however thought to be lowly heritable in older birds and better phenotype prediction could lead to yield improvements. A combinatorial analytics approach was used with a publicly available dataset of 1,027 pure bred and highly selected Rhode Island Red hens with 300,000 SNPs per bird [48]. The study identified combinations (up to 5th order) comprised of 2,018 highly predictive SNPs, mapping to 122 genes that correlated with increased egg weight for birds at 56 weeks of age.

Conversion of the SNP combinations into a standard genomic prediction model showed a 49% increase in phenotype prediction accuracy as well as a materially wider estimated breeding value distribution compared with standard breeding models based on GWAS and gBLUPs. Used to power a targeted selective breeding strategy, this can be expected to result in significantly increased genetic gain in laying hen breeding.

7. Computational challenges of combinatorial methods

Evaluating fully and in a hypothesis free manner the non-linearity of interacting endogenous and exogenous factors associated with a phenotype remains computationally challenging. The number of possible combinations to be evaluated expands exponentially as the combinatorial order increases.



(A) (B)

Fig. 4. (A) Disease architecture diagram illustrating the major genes involved in T2 and non-T2 asthma, and (B) very limited overlap of the pathway enrichment results for T2 vs non-T2 asthma sub-groups showing clear segregation of between T2 (left) and non-T2 (right) patients.

In a study with a 3-state SNP genotype model this increases according to $n!\bullet 3^r/r!(n-r)!$ where n = the total number of SNPs and r = the maximum combinatorial order. In a study with 550,000 SNPs, the number of 6-combinations is 1×10^{35}. For such a study, each additional step in combinatorial order increases the number of combinations to be tested by a factor of about 10^6. This becomes even more problematic if a large number of cycles of permutation are used to correct for multiple sampling.

Despite these serious challenges, the past few years have seen development of methods beginning to identify higher-order interactions. Their underlying statistical models can be based on frequentist (regression or simple statistic, e.g., chi-squared), Bayesian, or AI approaches. Typically, regression or AI models can correct for other covariates (such as environmental factors) and can assess SNP interactions not only in binary disease phenotypes, but also in quantitative traits.

Exhaustive (i.e., hypothesis-free) methods have been used to identify genetic co-expression networks [49] and pairwise associations between genes (traditional epistasis) [50]. Those based on brute force GWAS methods do not however scale well to higher-order regulatory patterns.

Hypothesis driven methods that test only a reduced, preselected set of features chosen based for example on pathway information have also been proposed [51], but these methods are likely by design to miss a large number of unexpected combinations [52]. These approaches however still require huge compute power and even with their significant compromises can lack statistical power. Consequently, they have been limited to subsets of the lower orders, and the functional interactions of high-order combinations have not yet been systematically explored.

This provides huge opportunities for innovative new methods, especially using massively parallel computational architectures such as GPUs, to explore new and generate new disease insights for complex, chronic diseases such as bipolar disease [53].

8. Remaining challenges in a combinatorial world

Combinatorial methods do not by themselves solve all of the problems associated with analyzing complex patient or animal datasets. The disease heterogeneity and number of samples available in a dataset will limit the combinatorial order of the associations that can be validated statistically, meaning that very small datasets will not be amenable to this approach. The results will also remain inherently rooted in and dependent on the populations under study and will be most relevant to these groups.

Population structure effects and inconsistent diagnosis/phenotype measurements will continue to have an impact on the accuracy even of combinatorial analytics methods and the predictive tools built on their findings. However, rather than being smeared by the population averaging effect of GWAS, these effects will tend to be more obvious, manifesting as clear disease sub-groups with different ancestries and/or phenotype metrics (e.g. smoking, BMI, prescription history, environment etc.) than other subgroups.

Combinatorial methods do nonetheless offer the potential for transformational insight into the complexity of biology underpinning chronic diseases and complex health, production and fertility traits in animal and crop production. The ability to identify and apply the non-linear signals underpinning complex biological processes could lead to much more widespread and accurate use of genomic and precision medicine approaches outside just the oncology and rare disease spaces. It may also lead to much more accurate genetic prediction models that can help improve the efficiency, environmental footprint and security of agriculture and the food production system.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: SG is a shareholder and director of PrecisionLife Ltd and a director of Synomics Ltd.

References

[1] The human genome project; https://www.genome.gov/human-genome-project (accessed 2 April 2021)
[2] International human genome consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[3] Venter JC, Adams MD, Myers EW, Li PW, et al. The sequence of the human genome. Science 2001;291:1304–51.
[4] Denny JC, Collins FS. Precision medicine in 2030—seven ways to transform healthcare. Cell 2021;184(6):1415–19. doi:10.1016/j.cell.2021.01.015.
[5] Venter, J.C. Reflections on the 20th anniversary of the first publication of the Human Genome Scientific American (Feb 2021); https://www.scientificamerican.com/article/reflections-on-the-20th-anniversary-of-the-first-publication-of-the-human-genome/ (accessed 2 April 2021)
[6] Yamazaki Y, Zhao N, Caulfield TR, et al. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. Nat Rev Neurol 2019;15:501–18. doi:10.1038/s41582-019-0228-7.
[7] Sims R, Williams J. Defining the genetic architecture of Alzheimer's disease: where next. Neurodegener Dis 2016;16(1-2):6–11. doi:10.1159/000440841.

[9] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017 Jul 6;101(1):5–22. doi:10.1016/j.ajhg.2017.06.005.

[10] Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun 2020;11:5900. doi:10.1038/s41467-020-19653-5.

[11] Polygenic Risk Scores and Clinical Utility. PHG foundation (2021) downloaded from https://www.phgfoundation.org/documents/polygenic-scores-and-clinical-utility.pdf (accessed 24 April 2021)

[12] Clark SA, van der Werf J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol Biol 2013;1019:321–30. doi:10.1007/978-1-62703-447-0_13.

[13] Hosseini SM, Kim R, Udupa S, Costain G, et al. National Institutes of Health Clinical Genome Resource Consortium. Reappraisal of reported genes for sudden arrhythmic death. Circulation 2018;138:1195–205.

[14] Walsh R, Buchan R, Wilk A, et al. Defining the genetic architecture of hypertrophic cardiomyopathy: re-evaluating the role of non-sarcomeric genes. Eur Heart J 2017;38:3461–8.

[15] Ingles J, Goldstein J, Thaxton C, et al. Evaluating the clinical validity of hypertrophic cardiomyopathy genes. Circ Genomic Precis Med 2019;12:e002460.

[16] Walsh R, Tadros R, Bezzina CR. When genetic burden reaches threshold. Eur Heart J 2020;41(39):3849–55 14 October. doi:10.1093/eurheartj/ehaa269.

[17] Nicholson DE. IUBMB-Nicholson metabolic pathways charts. Biochem Mol Biol Educ 2001;29:42–4.

[18] Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Nat Rev Microbiol 2014;12(5):327–40 May. doi:10.1038/nrmicro3238.

[19] Chaves M, Oyarzún DA. Dynamics of complex feedback architectures in metabolic pathways. Automatica 2019;99:323–32. doi:10.1016/j.automatica.2018.10.046.

[20] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 2007;39:17–23.

[21] Bertram L, Tanzi RE. Alzheimer disease risk genes: 29 and counting. Nat Rev Neurol 2019;15:191–2.

[22] Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. Nature 2016;536(7614):41–7. doi:10.1038/nature18642.

[23] Prokopenko, D., Morgan, S.L., Mullin, K. et al. Whole-genome sequencing reveals new Alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development medRxiv 2020.11.03.20225540; doi: https://doi.org/10.1101/2020.11.03.20225540

[24] Bullock JM, Medway C, Cortina-Borja M, et al. Discovery by the Epistasis Project of an epistatic interaction between the GSTM3 gene and the HHEX/IDE/KIF11 locus in the risk of Alzheimer's disease. Neurobiol Aging 2013;34(4):1309.e1–1309.e7.

[25] Ebbert MTW, Ridge PG, Wilson AR, et al. Population-based analysis of Alzheimer's disease risk alleles implicates genetic interactions. Biol Psychiatry 2014;75(9):732–7.

[26] Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell 2017;169(7):1177–86. doi:10.1016/j.cell.2017.05.038.

[27] Barton NH, Etheridge AM, Véber A. The infinitesimal model: definition, derivation, and implications. Theor Popul Biol 2017;118:50–73. doi:10.1016/j.tpb.2017.06.001.

[28] Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet 2019;28(R2):R133–42 15 October. doi:10.1093/hmg/ddz187.

[29] Clark SA, van der Werf J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol Biol 2013;1019:321–30. doi:10.1007/978-1-62703-447-0_13.

[30] Mellerup, E., Andreassen, O., Bennike, B., et al. Connection between genetic and clinical data in bipolar disorder PLoS One. 2012;7(9):e44623. doi:10.1371/journal.pone.0044623

[31] Mellerup E, Møller GL. Combinations of genetic variants occurring exclusively in patients. Comput Struct Biotechnol J 2017;15:286–9 Mar 10. doi:10.1016/j.csbj.2017.03.001.

[32] Tam V, et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;20(8):467–84. doi:10.1038/s41576-019-0127-1.

[33] Jain U. Effect of COVID-19 on the organs. Cureus 2020;12(8):e9540 Published 2020 Aug 3. doi:10.7759/cureus.9540.

[34] Rando, H.M., Bennett, T.D., Byrd, J.B., et al. Challenges in defining long COVID: striking differences across literature, electronic health records, and patient-reported information. Preprint. medRxiv. 2021;2021.03.20.21253896. Published 2021 Mar 26. doi:10.1101/2021.03.20.21253896

[35] Shelton JF, Shastri AJ, Ye C, et al. Trans-ancestry analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. Nat Genet 2021. https://doi.org/10.1038/s41588-021-00854-7.

[38] Sudlow C, Gallacher J, Allen N, Beral V, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779 Mar 31. doi:10.1371/journal.pmed.1001779.

[39] Armstrong, J., Rudkin, J.K., Allen, N., Crook, D.W., Wilson, D.J., Wyllie, D.H. and O'Connell, A.M. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank (2020) Microbial Genomics doi:10.1099/mgen.0.000397

[40] Taylor, K., Das, S., Pearson, M., Kozubek, J., Pawlowski, M., Jensen, C.E., Skowron, Z., Møller, G.L., Strivens, M.A., Gardner, S.P. Analysis of genetic host response risk factors in severe COVID-19 patients medRxiv 2020.06.17.20134015; doi: https://doi.org/10.1101/2020.06.17.20134015

[41] Das, S., Pearson, M., Taylor, K., Bouchet, V.A., Møller, G.L., Hall, T.O., Strivens, M.A., Tzeng, K.T.H., Gardner, S.P. Combinatorial analysis of phenotypic and clinical risk factors associated with hospitalized COVID-19 patients (in press) medRxiv 2021.02.08.21250899; doi: https://doi.org/10.1101/2021.02.08.21250899

[42] Schultz, B., Zaliani, A., Ebeling, C., et al. The COVID-19 PHARMACOME: a method for the rational selection of drug repurposing candidates from multimodal knowledge harmonization (in press) bioRxiv 2020.09.23.308239; doi: https://doi.org/10.1101/2020.09.23.308239

[43] Sugiyama, M.G., Cui, H., Redka, D.S., Karimzadeh, M. et al. Multiscale interactome analysis coupled with off-target drug predictions reveals drug repurposing candidates for human coronavirus disease bioRxiv 2021.04.13.439274; doi: https://doi.org/10.1101/2021.04.13.439274

[44] Kuruvilla ME, Lee FE, Lee GB. Understanding asthma phenotypes, endotypes, and mechanisms of disease. Clin Rev Allergy Immunol 2019;56(2):219–33 Apr. doi:10.1007/s12016-018-8712-1.

[45] Johnson N, et al. A review of respiratory biologic agents in severe asthma. Cureus 2019;11(9):e5690.

[46] PrecisionLife Genetic Underpinnings of T2 (Eosinophilic) versus non-T2 (non-eosinophilic) asthma https://precisionlife.com/wp-content/uploads/2020/12/T2-vs-non-T2-Asthma-Disease-Study-290121.pdf (accessed 24 April 2021)

[47] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh 1919:399–433. doi:10.1017/S0080456800012163.

[48] Synomics genomic improvement in laying hens https://www.synomics.ai/genomic-improvement-in-laying-hens/ (accessed 24 April 2021)

[49] Kontio JAJ, Rinta-aho MJ, Sillanpää MJ. Estimating linear and nonlinear gene coexpression networks by semiparametric neighborhood selection genetics, 215; 2020. p. 597–607. July 1. doi:10.1534/genetics.120.303186.

[50] Nagel RL. Epistasis and the genetics of human diseases. C. R. Biol. 2005;328.

[51] Guo X, Meng Y, Yu N, Pan Y. Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. BMC Bioinform 2014;15:102 Apr 10. doi:10.1186/1471-2105-15-102.

[52] Tuo S, Zhang J, Yuan X, He Z, Liu Y, Liu Z. Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations. Sci Rep. 2017;7(1):11529 Sep 14Erratum in: Sci Rep. 2018 Apr 17;8(1):6353. doi:10.1038/s41598-017-11064-9.

[53] Mellerup E, Jørgensen MB, Dam H, Møller GL. Combinations of SNP genotypes from the Wellcome Trust Case Control Study of bipolar patients. Acta Neuropsychiatr 2018;30(2):106–10 Apr. doi:10.1017/neu.2017.36.

[54] Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2020. Alzheimers Dement (N Y). 2020;6(1):e12050. doi:10.1002/trc2.12050.

[55] Ganna A, et al., The COVID-19 Host Genetics Initiative Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. Preprint. https://www.medrxiv.org/content/10.1101/2021.03.10.21252820v2. medRxiv 2021. doi:10.1101/2021.03.10.21252820.