A generalizable hypothesis for the genetic architecture of disease: pleomorphic risk loci

Andrew Singleton\(^1\) and John Hardy\(^2,\)\(^*\)

\(^1\)Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA and \(^2\)Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, London, UK

The dominant and sometimes competing theories for the aetiology of complex human disease have been the common disease, common variant (CDCV) hypothesis, and the multiple rare variant (MRV) hypothesis. With the advent of genome wide association studies and of second-generation sequencing, we are fortunate in being able to test these ideas. The results to date suggest that these hypotheses are not mutually exclusive. Further, initial evidence suggests that both MRV and CDCV can be true at the same loci, and that other disease-related genetic mechanisms also exist at some of these loci. We propose calling these, pleomorphic risk loci, and discuss here how such loci not only offer understanding of the genetic basis of disease, but also provide mechanistic biological insight into disease processes.

The common disease, common variant (CDCV) hypothesis centres on the notion that the genetic risk for complex, common diseases is mediated by numerous common variants (1). Because such variants by definition have become common, a corollary of this hypothesis is that such risk variants are likely to exert little net negative selective pressure, either because they have quite small biological effect, there is balancing selection, or because the variants are associated with post-reproductive diseases. Genome wide association (GWA) studies have made it possible to explicitly test the CDCV hypothesis and have shown that there is substantive truth to this idea (http://www.genome.gov/26525384) (2). In contrast to the CDCV hypothesis, the multiple rare variant (MRV) hypothesis ascribes the genetic risk component of common, complex diseases to MRVs (3). Unlike common variants, low-frequency alleles are not fixed in the population, and thus, there has been little opportunity for selective pressure to limit the effect size of such variants.

\*To whom correspondence should be addressed. Email: j.hardy@ion.ucl.ac.uk

Published by Oxford University Press 2011.
copy number mutations that increase risk for disease substantially and in a dose-dependent manner.

In summarizing this model: at each locus identified as contributing towards the pathogenesis of disease, we predict that there will be many alleles that contribute to the risk profile for that disease. If these are coding changes, their effects will be dependent on the precise details of their biochemistry. If they affect expression or splicing, we would expect their effects on disease risk to be predictable by the size and nature of their effects on this expression and splicing. Disease risk encoded by gene duplications will also be increased by high expressing alleles: disease risk encoded by gene deletions will also be increased by low expressing alleles and nonsense mutations and are likely to be potentiated by low expressing trans alleles.

CONSEQUENCES AND PREDICTIONS OF THE PRL MODEL

We can begin to predict which types of disease-related genes are more likely to be PRL and what predictions can be made about PRL.

(1) We suspect that these loci are more likely to be associated with later onset diseases. One major form of disease risk for late onset disease, we predict, will be heterozygous loss of function mutations at loci where homozygous loss is either lethal or associated with a related recessive disease, and evidence for such phenomena already exist: first, glucosecerebrosidase mutations that when homozygous cause Gaucher’s disease and when heterozygous are risk loci for Parkinson’s disease (PD) (10); second, loss-of-function mutations as the cause of ichthyosis vulgaris when homozygous and as risk loci for atopic dermatitis and asthma when heterozygous (11,12).

(2) Genes that have been previously linked to monogenic disease through a quantitative pathobiological mechanism are likely to be PRL. One would suspect that in such cases, where monogenic disease is driven by large relative effects on splicing or expression, risk will be affected by more subtle effects on splicing expression mediated by rare or common non-coding variation. An example of this is the microtubule associated protein tau (MAPT) locus in the related syndromes of frontotemporal dementia and progressive supranuclear palsy in which dominant mutations cause a dramatic effect on alternate splicing of a single exon cause the former (13), whereas the MAPT haplotype which has a similar but more subtle effect, we predict that common alleles affecting expression are also likely to modulate risk conferred by rare mutations and common coding variants. (B) The LPR model as it applies to genetic variability at SNCA in PD. Notably common non-coding variants that alter SNCA expression are linked to increased risk for disease; whole gene duplication and triplication mutations of SNCA are a known cause of disease (and copy number is correlated with severity) and rare protein-coding mutations have been identified. (C) The LPR model as it applies to LRRK2 in PD. Rare protein-coding disease-causing mutations have been identified, common protein-coding risk variants have been found and, recently, common non-coding variability 5′ to LRRK2 has been implicated as a risk factor for PD.
predisposes to the latter (14). Clear examples of this come from the comparatively well understood study of variability in blood lipids where ~20% of the loci identified as being important in normal variability had previously been identified as Mendelian loci for lipid disorders (15).

(3) Variants at PRL associated with the same disease will likely have a unifying biological mechanism/substrate. Linking the immediate biological consequences of these divergent types of risk variant will be a unique opportunity toward understanding the underlying disease process.

(4) Within PRL, the risk associated with a rare mutation will be modulated by genetic variants in cis to the mutation. Two examples of this are the association of complement factor H with macular degeneration in which coding changes have a major effect on disease risk, but the haplotype also contributes to the risk (6–9) and apolipoprotein E and Alzheimer’s disease (16) where the same phenomenon operates and the different risk haplotypes (17) are associated with different levels of expression (18). Furthermore, risk variability at the same locus but in trans to the mutation should exert a major effect on expressivity or penetrance of the disease particularly in cases of where loss of function is critical to pathogenicity. There are already a few examples where variability, both cis and trans to pathogenic changes contribute to the variability in disease phenotypes (19,20). As more high-risk variants are identified, we would expect that this type of effect is likely to be almost a general phenomenon where high expressing trans alleles can partially compensate for loss of function alleles or low expressing cis alleles can mitigate the phenotypes of gain of function alleles.

EXISTING PROOF FOR THE PRL HYPOTHESIS

The data surrounding common risk variability have matured considerably over the previous 4 years, and now there are hundreds of replicated risk loci (2). With even more history, linkage and positional cloning efforts have revealed mutations that underlie >1500 monogenic diseases (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html). It is readily apparent that several loci contain both common risk variants and rare, highly penetrant mutations, associated with the same or related diseases. From our own work, common genetic variability at the SNCA, MAPT and LRRK2 loci confers mild risk for PD (21–23); these three genes have previously been shown to contain highly penetrant point mutations that cause familial forms of parkinsonism (13,24–26). We use two of these loci to illustrate the concept of PRL.

First, common and rare variants at the SNCA locus: missense mutations in SNCA were the first-described genetic cause of PD (26); subsequently both whole locus triplication and duplication were shown as a cause of rare familial forms of PD, mediating disease through increased expression of the protein product, α-synuclein. Notably, in these cases, disease severity and age at onset were related to SNCA copy number, and implicitly α-synuclein levels (27–29) in a dose-dependent manner (30). Given our model, one would predict that common non-coding variability, which alters SNCA expression, would be a risk factor for PD. This indeed seems to be the case with the identification of multiple risk haplotypes at SNCA. Furthermore, these risk haplotypes underpin increased SNCA expression (21,22). Thus, in the case of SNCA, rare coding mutations, rare whole gene copy number variants and common expression-related variants, all occur and contribute in different ways to disease risk (Fig. 1B).

Secondly, genetic variability at LRRK2: missense mutations in LRRK2 were identified as a relatively common cause of PD in 2004 (24,25). Subsequent work has shown that common missense variants within the Asian population act as risk factors for PD, increasing risk for disease ~2-fold (31). Further work using GWA testing suggests that in addition to these protein-coding risk variants, there are low-risk, non-coding variants immediately 5′ to LRRK2 (22). Thus, in the case of LRRK2, there exists three known categories of variant associated with PD; high penetrance missense mutations, moderate risk missense variants and low-risk non-coding risk variants (Fig. 1C). Parsimony would suggest that at this locus common, non-coding variability that presumably mediates risk effects by splicing/expression and apparently qualitative changes to the protein-coding sequence, ultimately exert their effects through similar pathobiological mechanisms.

The next stage of investigation is likely to centre on finding rare variants, with much lower risk effects than disease-causing mutations. Because of the technological challenges associated with identifying such variants, particularly those that are non-coding, there is a paucity of data surrounding this mechanism. However, initial work in this area has begun to provide examples of this phenomenon; perhaps most notable has been described in diabetes. Common variability at the IFIHI1 locus was implicated in risk for type 1 diabetes (T1D) by GWA (32). Subsequent resequencing of this locus revealed that MRVs exist within this gene that confer a small protective effect against T1D, and that these variants are independent of the originally identified GWA signal (32,33). This resequencing work centred on exons and immediate regulatory sequences; however, one might expect that as second-generation sequencing affords more ability to perform highly multiplexed sequencing of large genomic regions, the field will begin to identify more non-coding variants associated with disease. The increasing resolution of the 1000 genomes project, and related efforts, coupled with the availability of ever more GWA data for common diseases will afford a complementary approach to this, through the ability to impute even rare variants as significantly increasing disease risk.

HOW TO TEST THE PRL HYPOTHESIS

We predict that as genetic data accumulate for common diseases, the PRL hypothesis will be explicitly proven. The initial testing of this idea will likely result from follow-on experiments to GWA studies, which aim to perform deep resequencing of genomic loci implicated by this method. While we recognize that in terms of our hypothesis, this type of experiment may provide support that is circular in its logic, these types of focused data will be invaluable as we begin to construct methods to understand and delineate benign, risk and protective rare variants. At some point, this type of effort will transition from targeted resequencing, to whole genome
resequencing. Should the PRL hypothesis be generalizable, the pre hoc prioritization for analysis of loci previously implicated in the disease in question, either because of a link to monogenic forms of disease or because of identified common risk variants at the locus, may provide considerable statistical benefit.

CONCLUDING REMARKS

The CDCV and MRV hypotheses were proposed as models for disease pathogenesis largely before systematic data were available. Now that data have become available through the application of array and sequencing technologies, we can see that they both have some elements of truth to them, but that a greater and more interesting truth comes from their synthesis: that disease pathogenesis is initiated by a mixture of common and rare variants which interact in a manner which is, to some extent, predictable at each locus. We suggest that this synthesis, which we encapsulate as the pleomorphic risk is, to some extent, predictable at each locus. We recognize that there will be interactions between disease loci and between, more generally, genes and environment and that stochastic effects may be a consequence of disease loci and between, more gener- ally, genes and environment and that stochastic effects may often have a role. Additionally, as large numbers of loci are identified, as has now been possible for human height (34), it will be possible to elucidate the biochemical and the developmental pathways which the genetic findings underpin. Eventually, these too will need to be incorporated into our models as we develop a complete description of disease risk.

Conflict of Interest statement

None declared.

FUNDING

This work was supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services; project number Z01 AG000957-07 (A.S.) and by the Wellcome Trust and MRC (J.H.).

REFERENCES