Human Molecular Genetics, 2011, Vol. 20, Review Issue 2 **R158–R162**
doi:10.1093/hmg/ddr358
Advance Access published on August 29, 2011

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

Received June 6, 2011; Revised and Accepted August 10, 2011

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.

CONSOLIDATION AND BEYOND: PLEOMORPHIC RISK LOCI

Critically, both the CDCV and MRV hypotheses are ideas that have been generalized to exclusively encapsulate the total genetic risk in disease. However, that only one of these ideas can be true is a fallacy, these concepts are not mutually exclusive and a more mature view brings these two ideas together as a general hypothesis for disease susceptibility. We believe, however, that it is not only sensible to consolidate these hypotheses generally but that within a single disease, multiple common and rare risk alleles are likely to co-exist at the same locus. Further, we hypothesize that genetic risk at individual loci will comprise a spectrum of both allele frequency and effect size, encompassing low-risk and high-risk variants; we nominate these as pleomorphic risk loci (PRL).

We have previously discussed the concept of graded haplotype risk in the context of common risk alleles (4,5): briefly this argues that GWA studies initially identify only the most significant disease-associated common variants at any one locus, but that it is likely there are several other common risk variants at the same locus, within or out with the same haplotype block. In this model, there is no single haplotype for disease risk, but rather a collection of haplotypes at or across the same locus that are associated with graded risk of disease (Fig. 1A). As we have discussed, there already exists biological support for this idea (4,6–9). The PRL hypothesis is a natural extension of this idea, incorporating other types of genetic variation of varying risk classes within the same loci. In this model, we suggest that at a single locus, there will likely be: (i) common non-coding variants that effect gene expression and/or splicing; (ii) rare coding and highly penetrant mutations; (iii) common coding and low-to-moderate risk alleles; (iv) rare non-coding variants that exert small effects on risk; and (v) rare whole gene

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

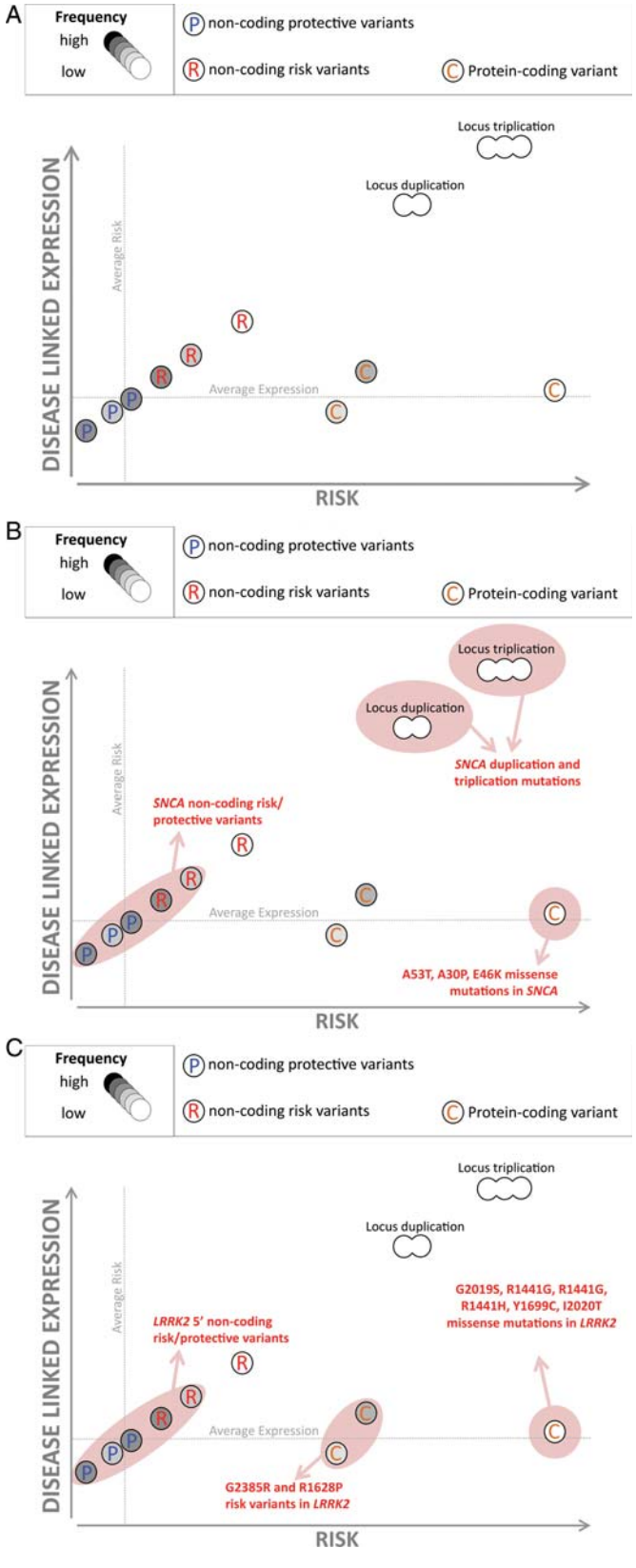

Figure 1. Schematic of PRL. (**A**) This schematic illustrates that at the same locus several disease-related genetic mechanisms may co-exist, each influencing disease through different biological effects on a single gene. In this particular model, expression of a gene is positively correlated with risk. Notably, 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.

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 icthyosis 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,

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 *IFIH1* 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 locus hypothesis, can be used to explain much of the risk for common disease. Of course, we recognize that there will be interactions between disease loci and between, more generally, 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

1. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
2. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
3. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
4. Hardy, J. and Singleton, A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.
5. Singleton, A., Myers, A. and Hardy, J. (2004) The law of mass action applied to neurodegenerative disease: a hypothesis concerning the etiology and pathogenesis of complex diseases. *Hum. Mol. Genet.*, **13**(Spec No. 1), R123–R126.
6. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., San Giovanni, J.P., Mane, S.M., Mayne, S.T. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.
7. Haines, J.L., Hauser, M.A., Schmidt, S., Scott, W.K., Olson, L.M., Gallins, P., Spencer, K.L., Kwan, S.Y., Noureddine, M., Gilbert, J.R. *et al.* (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science*, **308**, 419–421.
8. Edwards, A.O., Ritter, R. 3rd, Abel, K.J., Manning, A., Panhuysen, C. and Farrer, L.A. (2005) Complement factor H polymorphism and age-related macular degeneration. *Science*, **308**, 421–424.
9. Li, M., Atmaca-Sonmez, P., Othman, M., Branham, K.E., Khanna, R., Wade, M.S., Li, Y., Liang, L., Zareparsi, S., Swaroop, A. and Abecasis, G.R. (2006) CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet.*, **38**, 1049–1054.
10. Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A. *et al.* (2009) Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.*, **361**, 1651–1661.
11. Smith, F.J., Irvine, A.D., Terron-Kwiatkowski, A., Sandilands, A., Campbell, L.E., Zhao, Y., Liao, H., Evans, A.T., Goudie, D.R., Lewis-Jones, S. *et al.* (2006) Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nat. Genet.*, **38**, 337–342.
12. Palmer, C.N., Irvine, A.D., Terron-Kwiatkowski, A., Zhao, Y., Liao, H., Lee, S.P., Goudie, D.R., Sandilands, A., Campbell, L.E., Smith, F.J. *et al.* (2006) Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.*, **38**, 441–446.
13. Hutton, M., Lendon, C.L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A. *et al.* (1998) Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, **393**, 702–705.
14. Myers, A.J., Pittman, A.M., Zhao, A.S., Rohrer, K., Kaleem, M., Marlowe, L., Lees, A., Leung, D., McKeith, I.G., Perry, R.H. *et al.* (2007) The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol. Dis.*, **25**, 561–570.
15. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.
16. Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L. and Pericak-Vance, M.A. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923.
17. Chartier-Harlin, M.C., Parfitt, M., Legrain, S., Perez-Tur, J., Brousseau, T., Evans, A., Berr, C., Vidal, O., Roques, P., Gourlet, V. *et al.* (1994) Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. *Hum. Mol. Genet.*, **3**, 569–574.
18. Lambert, J.C., Perez-Tur, J., Dupire, M.J., Galasko, D., Mann, D., Amouyel, P., Hardy, J., Delacourte, A. and Chartier-Harlin, M.C. (1997) Distortion of allelic expression of apolipoprotein E in Alzheimer's disease. *Hum. Mol. Genet.*, **6**, 2151–2154.
19. Risch, N.J., Bressman, S.B., Senthil, G. and Ozelius, L.J. (2007) Intragenic cis and trans modification of genetic susceptibility in DYT1 torsion dystonia. *Am. J. Hum. Genet.*, **80**, 1188–1193.
20. Goldfarb, L.G., Petersen, R.B., Tabaton, M., Brown, P., LeBlanc, A.C., Montagna, P., Cortelli, P., Julien, J., Vital, C., Pendelbury, W.W. and Gajdusek, D.C. (1992) Fatal familial insomnia and familial Creutzfeldt-Jakob disease: disease phenotype determined by a DNA polymorphism. *Science*, **258**, 806–808.
21. Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
22. Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., Sveinbjörnsdóttir, S. *et al.* International Parkinson Disease Genomics. (2000) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, **377**, 641–649.
23. Spencer, C.C., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., Barker, R.A., Bellenguez, C., Bhatia, K., Blackburn, H. *et al.* UK Parkinson's Disease Consortium; Wellcome Trust Case Control Consortium 2. (2011) Wood NW (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.*, **20**, 345–353.
24. Paisan-Ruiz, C., Jain, S., Evans, E.W., Gilks, W.P., Simon, J., van der Brug, M., Lopez de Munain, A., Aparicio, S., Gil, A.M., Khan, N. *et al.*

(2004) Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron*, **44**, 595–600.
25. Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., Hulihan, M., Uitti, R.J., Calne, D.B. *et al.* (2004) Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*, **44**, 601–607.
26. Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R. *et al.* (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science*, **276**, 2045–2047.
27. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. *et al.* (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science*, **302**, 841.
28. Chartier-Harlin, M.C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M. *et al.* (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet*, **364**, 1167–1169.
29. Ibanez, P., Bonnet, A.M., Debarges, B., Lohmann, E., Tison, F., Pollak, P., Agid, Y., Durr, A. and Brice, A. (2004) Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet*, **364**, 1169–1171.
30. Singleton, A. and Gwinn-Hardy, K. (2004) Parkinson's disease and dementia with Lewy bodies: a difference in dose? *Lancet*, **364**, 1105–1107.
31. Zabetian, C.P., Yamamoto, M., Lopez, A.N., Ujike, H., Mata, I.F., Izumi, Y., Kaji, R., Maruyama, H., Morino, H., Oda, M. *et al.* (2009) LRRK2 mutations and risk variants in Japanese patients with Parkinson's disease. *Mov. Disord.*, **24**, 1034–1041.
32. Smyth, D.J., Cooper, J.D., Bailey, R., Field, S., Burren, O., Smink, L.J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D.B. *et al.* (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.*, **38**, 617–619.
33. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
34. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.