Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders

Peter A. Holmans, Thomas H. Massey and Lesley Jones*

MRC Centre for Neuropsychiatric Genetics and Genomics

School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK

*Corresponding author:

phone: +44 2920 688069

email: Jonesl1@cf.ac.uk

Abstract

In the decades since the genes and mutations associated with the commoner Mendelian disorders were first discovered, technological advances in genetic analysis have made finding genomic variation a much less onerous task. Recently, the global efforts to collect subjects with Mendelian disorders, to better define the disorders and to empower appropriate clinical trials, along with improved genetic technologies, have allowed the identification of genetic variation that does not cause disease, but substantially modifies disease presentation. The advantage of this is it identifies biological pathways and molecules, that, if modified in people, might alter disease presentation. In Huntington's disease (HD), caused by an expanded CAG repeat tract in *HTT*, genetic variation has been uncovered that is associated with change in the onset or progression of disease. Some of this variation lies in genes that are part of the DNA damage response, previously suggested to be important in modulating expansion of the repeat tract in germline and somatic cells. The genetic

evidence implicates a DNA damage response-related pathway in modulating the pathogenicity of the repeat tracts in HD, and possibly, in other trinucleotide repeat disorders. These findings offer new targets for drug development in these currently intractable disorders.

Introduction

Why look for genetic modifiers?

Diseases caused by Mendelian mutations tend to be rare, but together form a substantial cause of morbidity and mortality. Most rare diseases are genetic and life-limiting. By definition a disease must have prevalence lower than 1/2000 to be accounted rare, though this varies across different countries (the International Rare Diseases Consortium: www.irdirc.org; Rare Diseases Clinical Research Network in the USA www.rarediseasesnetwork.org). The causative genetic lesions are known for many such diseases, often for several decades (1). However, finding novel disease modifying therapies for such diseases has been slow, partly because in many cases identifying genetic lesions did not give obvious clues about the underlying disease biology, and partly because targeting drugs to specific tissues, particularly within the central nervous system (CNS), is challenging. New antisense oligonucleotide technologies, that target the mutated mRNA products of the relevant genes directly, have shown early promise. For example, the antisense oligonucleotide (ASO) drug nusinersen has recently been approved by the FDA and EMA in spinal muscular atrophy (2, 3). Trials using similar ASO technologies are ongoing in HD and other currently intractable disorders. However, such direct therapies are not appropriate in all diseases, are currently difficult to deliver, especially to the brain, expensive, may not be allele specific and may well need supplementing with therapeutics addressing other aspects of disease.

In the more common of these rare Mendelian diseases, where substantial numbers of subjects can be assembled, it is possible to gain novel insights into disease biology by looking for variation in the rest of the genome that modifies aspects of the disease phenotype. The advantage of such approaches are their unbiased nature – no *a priori* knowledge of disease biology is required - and the ability to shed light on the biology of inaccessible tissues such as those of the CNS. If such variation identifies specific biological pathways then this highlights relevant pathophysiological processes in people with the disease that are by default target pathways for therapeutics, since their modulation can alter the way the disease presents in people. In a disease like HD, where many pathways have been implicated in the pathogenesis caused by the expanded CAG tract (4), this is likely to assist in decisions about which of those pathways are critical in manifestation of disease symptoms and therefore could form ideal avenues for the development of new therapeutics (or repurposing of existing therapeutics). It is important in this context to note that drug targets underpinned by genetic evidence have a higher chance of progressing into clinical use (5).

Finding genetic modifiers

Before conducting any genetic study it is necessary to test whether the modifier phenotype has a genetic component by performing a segregation analysis in families. For example: Wexler et al. (6) found that about 40% of the variation in age at onset of Huntington's disease was heritable, after correcting for CAG repeat length. The age at onset of motor symptoms in HD is inversely correlated with the CAG repeat length, but this is strongest between 41 and 56 CAG and the exact relationship outside these repeat numbers remains to be determined (7). It is likely that there is incomplete penetrance and therefore under-ascertainment of people below these repeat numbers as they do not come to clinical attention (8). Above 56 CAG repeats the disease onsets below the age of 20 years and there are many fewer cases (Figure 1). This demonstrates that understanding the distribution of the phenotypic variables is critical, and this in turn implies that a minimum sample

size is likely to be necessary. One of the issues in Mendelian diseases is that they are by their nature rare, and therefore collecting the necessary numbers of subjects for genetic modifier studies in such diseases can be difficult. In HD the substantial efforts over the past several decades of studies such as COHORT (9), PREDICT (10), REGISTRY (11), TRACK (12), ENROLL (13)(https://www.enroll-hd.org/) have provided large cohorts of subjects at risk of disease and with manifest disease, with the relevant systematically collected clinical information and DNA samples.

Initial studies of genetic modifiers used linkage analysis on pedigrees containing multiple affected individuals (e.g., (14)). However, linkage studies have low power to detect common modifier variants of small effect (15). Therefore, studies testing association between individual genetic variants and modifier phenotypes have become more widespread. Initially, technical limitations restricted these to studies of candidate genes in relatively small samples, resulting in lack of replicability (16), as in other areas of human genetics (17, 18).

Recently, genotyping advances have made it possible to perform genome-wide association studies (GWAS) for genetic modifiers in large samples (19), an approach that has been used successfully in studies of complex genetic disorders (20). It should be noted that in all cases GWAS have limited power to detect rare variation, which requires the use of sequencing (see (21) for a review of the design and analysis of sequencing studies).

Increasing the power of genome-wide analyses for modifier detection

There are a number of ways that the power of genetic modifier studies can be increased. Most of these also apply to common complex genetic disorders and include: increasing sample size; collecting more accurate and more directly genetically encoded phenotypes; using the underlying biology to resolve non-genome-wide significant signal from noise and using information from other diseases. The power to detect genetic modifiers will, of course, depend upon their genetic architecture. Typically, susceptibility to complex genetic traits is largely due to common variants of

small effect (e.g. (22)), but this may not be the case for genetic modifiers of a single-gene disorder, which require a particular genetic background to operate and are therefore not subject to evolutionary constraint in the general population (23).

Genetic and statistical considerations

The usual considerations that apply to any GWAS also apply to modifier studies. GWAS typically test association between the phenotype and the number of copies of the minor (rarer) allele via regression (linear for quantitative phenotypes, logistic for binary phenotypes), although other tests are available to measure deviations from additive effects of the alleles. The resulting (-log) p-values for each variant can then be plotted against genomic position – a "Manhattan plot" (Figure 2). There are many software packages available for performing GWAS, notably PLINK (24, 25). For a more detailed review of the design, analysis and interpretation of GWAS see (26) .

Given that GWAS usually contain several million variants, stringent multiple testing correction must be applied to minimise the chance of false-positive associations. This procedure is complicated by non-independence of individual association tests due to linkage disequilibrium (LD) between SNPs. A p-value criterion of $5x10^{-8}$ is often used to determine genome-wide significance (27). However, this was derived for European populations – African populations show less LD and thus require a more stringent criterion. If both common and rare variants are analysed (for example, in a whole-genome sequencing study), the criterion for genome-wide significance is even more stringent – about $1x10^{-8}$ (28).

Since a stringent p-value is required to declare significance, it follows that large sample sizes are needed to achieve power. Chapman et al. (29) showed that the parameter determining power for additive association to a quantitative trait is equal to $(N-1)r^2h^2$, where N is the sample size, r is the correlation between the trait and test alleles, and h^2 is equal to the narrow-sense heritability. Figure 3 shows how power varies with sample size (and h^2). Narrow-sense heritability is equal to $V_a/(V_g+V_e)$,

where V_a, V_g, V_e are the components of trait variance attributable to additive effects at the test locus (that is, the effect of the locus on the phenotype due to the sum of the individual effects of each of the two alleles), other genetic effects, and environmental factors unrelated to genetics. To maximise h^2, it is necessary (as far as possible) to minimise V_g and V_e, and this is typically done by regressing off known effects on the phenotype, both genetic (e.g. the CAG repeat for age at onset of Huntington's disease) and environmental. V_e may also be reduced by more accurate phenotyping as exemplified in our recent study using a composite prospective HD progression score as a phenotype (30).

Power may also be improved by increasing the correlation r between test variant and the untyped causal variant. This may be achieved by using a densely-genotyped reference panel, such as the large Haplotype Reference Consortium dataset (31) to estimate the correlation structure between the untyped variant and nearby variants from the GWAS SNP panel (see (32) for a review) or sequenced subjects (33). This structure can then be applied to the GWAS dataset to impute the missing genotypes, which can then be tested for association with the phenotype, as described by de Bakker et al (34). However, the ability of imputation methods to capture rare variation (frequency <0.1%) is limited, and sequencing is preferred.

The power to detect associations to individual rare variants is generally low. For these, it is usual to combine variants across a region (typically a gene). For example, the relationship between phenotype and the total burden of rare variants may be tested (35). This method assumes that all rare variants act in the same direction on phenotype, which is reasonable for a disease, but less so for a modifier acting on a disease, which will not be under the same evolutionary constraint. To avoid this assumption, a commonly used alternative is SKAT-O (36, 37). A cost-effective way to increase the power of rare-variant studies is to sequence people from the extremes of the phenotype distribution (38). Higher orders of analysis unit, such as biological pathways, can also be used in such analyses, as outlined below.

There are a number of issues affecting the power of GWAS for modifiers of Mendelian disease. Relatively rare diseases may require subjects from geographically widespread populations and unless suitably accounted for, population stratification can cause false-positive associations in GWAS. Typically, this manifests itself as a systematic inflation of association test statistics when plotted against their expected value (Figure 4). The degree of inflation can be quantified by the genomic inflation factor λ, defined as the median of the observed test statistics divided by its expected value (39). Genomic inflation due to stratification can be reduced by including principal components that capture the genotypic variation across the sample as covariates in the association analysis (40). It should be noted that the value of λ depends on sample size (41). Inflated values of λ may also arise if the phenotype has a polygenic component – this can be disentangled from stratification using LDscore (42). In HD analysis was restricted to those with European ancestry to avoid this problem though the principal components that covaried with population were even then taken into account in analysis (19). In rare diseases where subjects from multiple different populations are genotyped this will be more difficult to account for.

Even in the absence of stratification, failure to account for relatedness among individuals will result in false positive associations. Studies in most Mendelian disorders are likely to include related people. If the relatedness is known in advance, family-based association methods can be used (e.g. (43)): relatedness will not always be obvious to the researchers but can be inferred from the genotyping using linear mixed models (LMMs). LMMs are becoming the preferred method for GWAS, since they correct for population stratification in addition to cryptic relatedness. LMMs, and their software implementation, are reviewed in (44). It should be noted that LMMs can lose power relative to standard methods when applied to a binary phenotype.

Using biology to improve power

Pathway analyses are often used to infer disease-relevant biology in genome-wide studies. These involve testing whether pre-specified sets of biologically-related genes ("pathways") are more significantly associated with the phenotype than other genes, and can overcome issues of heterogeneity in associated genes. Pathway analysis methods for GWAS are reviewed in more detail in (45) and (46). Currently, the preferred pathway analysis for GWAS data is MAGMA, since this has superior statistical properties to other methods (47).

While pre-specified gene sets, such as those from the Gene Ontology, are a useful initial analytical set, they are limited by prior biological knowledge – poorly studied (but biologically relevant) genes will not be included in the analysis, and poorly annotated and assigned genes will increase noise. One way to extend coverage of these genes is to use other types of genomic data (such as gene-expression) to form networks of correlated genes (48), thereby indicating genes for future study. This approach has been used to show that a co-expression module of immune-related genes were enriched for signal in an AD GWA study (49, 50) and this same module was enriched for commonality in expression signature between HD blood and brain and Alzheimer's disease (51).

While the GeM-HD GWA study revealed three statistically genome-wide significant signals, it also revealed a substantial underlying signal in the DNA repair pathways, implicating the DNA damage response as an important modifier of HD (19). This finding focussed attention on the repeat polymorphism in the DNA as a modifier of disease rather than the huntingtin protein and its downstream effects. If true, then one would expect the same modifiers to operate in other diseases caused by the same trinucleotide repeat expansion mechanism. A candidate gene study (52) demonstrated that this was indeed the case, consistent with modifiers acting on the expanded CAG repeat. Notably, members of the DNA damage response pathways had previously been shown to

affect repeat dynamics and stability in several of these diseases in animal models (53, 54), with hints that they might be important also in people (55).

Other types of genome-wide data such as expression data can also be integrated with GWAS data to indicate relevant genes under association peaks by looking for co-localisation of SNPs associated with the modifier phenotype and expression level (56). Other types of "omics" data, such as Hi-C, can also be used for this purpose (57), and it is likely that many such integrative studies will be used to enhance the power and biological prediction available from such studies.

Enhancing the phenotype

As with common complex disorders, using more accurate phenotypes is likely to enhance the power of genetic analyses (58, 59). In addition, it might well provide endophenotypes that can be measured in everyone at risk of disease, whether currently symptomatic or not. Quantitative phenotypes are ideal - in cystic fibrosis, lung function is a measureable quantitative phenotype and has proven relatively powerful (60). Although the background mutations in *CFTR* vary, they all ablate some or all of CFTR function, so the genotype-phenotype relationship does not all reside in the different mutations in the CFTR gene. Pegoraro et al. (61) examined candidate modifiers of disease severity in in Duchenne muscular dystrophy and detected a variant in the promoter of the SPP1 gene which replicated in a second cohort and further studies have used the objective measure of loss of ambulation as a modifier phenotype to partially confirm these data (62, 63). In HD and other adult onset neurological diseases this is not as straightforward as objective quantitative phenotypes are less easy to capture. In HD, age at onset, even when defined as age at onset of motor symptoms specifically, is not ideal. It is subjective and has often been collected retrospectively. One way to overcome this is to collect prospective multivariate phenotypes and to use these to create quantitative measures that reflect disease burden. Our recent study, generating a multivariate quantitative phenotype using the extensively longitudinally and prospectively clinically

assessed TRACK-HD study (12), was powerful enough to give an almost genome-wide significant signal in 216 subjects, just over half of whom had manifest HD. This signal replicated in the less well phenotyped Registry study using a parallel, but not identical, quantitative phenotype, and gave a genome-wide significant signal on chromosome 5 after meta-analysis (51). This study also highlighted the DNA repair pathways, as the lead SNP was a coding variant in *MSH3*. The interpretation of this finding, in concert with previous functional experimentation provides substantial clues about the nature of one class of modifiers in HD, and by extrapolation, in the other repeat disorders.

Genetic modifiers and trinucleotide repeat disease biology

The GeM-HD GWAS identified three chromosomal loci with genome-wide significance for altered age at motor onset: one on chromosome 8 and tow on chromosome 15. Pathway analyses highlighted DNA repair processes as likely modifiers of phenotype. How do these findings fit together into a model of somatic (non-germline) CAG repeat expansion that might underpin HD pathogenesis? While the genetic evidence is not yet conclusive that *FAN1* is the gene in the chromosome 15 locus, there are a number of pieces of suggestive evidence. FAN1 is a 5'-exo/endo-nuclease involved in interstrand DNA crosslink repair which was identified as interacting with a number of mismatch repair proteins including MLH1, encoded at the chromosome 3 locus in the GeM-HD GWAS (64, 65). In addition, the signal in chromosome 15 has a coding mutation in FAN1, pArg507His (rs150393409; $p = 9.34x10^{-18}$), which is close in significance to the index SNP (rs146353869; $p = 4.30x10^{-20}$) giving 6 years earlier onset of disease (19). The change is predicted to be at the C-terminal end of the DNA binding domain of FAN1 (66). Note that both SNPs were imputed and deciding which of the significant SNPs at a locus is the functional SNP driving the genetic signal is difficult but may be addressed by using larger samples and direct genotyping.

Substantial biological evidence also links DNA mismatch repair with trinucleotide repeat disorders. Mice carrying an expanded CAG repeat in *HTT*/*Htt* crossed with knock-out mice for mismatch repair genes *Msh2* and *Msh3* show no somatic expansion of the repeat locus and improved phenotype (67, 68): this is also seen in mice modelling the non-CAG repeat expansions underlying Fragile X syndrome and Friedreich's ataxia (69, 70). *MSH3*, which has a coding SNP associated with HD progression (30), encodes a protein that forms a heterodimer (MutSβ) with MSH2 that can bind specifically to abnormal DNA structures to direct their repair. Usually this activity is involved with repair of mismatches in the nascent DNA of dividing cells but recent evidence has shown that MMR functions in non-replicating cells, and that MutSβ can bind and stabilise CAG-containing DNA hairpins (71–73). Such hairpin binding could be a precursor to downstream CAG repeat expansion, and hence pathology. *MSH3* is expressed in neurons and upregulated in HD mouse model brain (71, 74), where it is associated with somatic CAG repeat instability (68). In humans, increased somatic expansions of the CAG repeat tract in *HTT* are associated with earlier onset of HD (75).

Although the evidence from human genetics and cellular/animal models of HD implicates mismatch repair processes most strongly in pathogenesis, there are also numerous studies showing that other pathways within the DNA damage response are involved. For example, knockout of base-excision repair or transcription-coupled repair pathways in specific animal or cellular models of CAG repeat diseases can inhibit repeat expansions and ameliorate phenotype (76–79). In reality, the DNA damage response consists of multiple overlapping pathways which are at least partially redundant so that genomic integrity is preserved (80). Given the evidence emerging from studies of genetic modifiers in the trinucleotide repeat diseases, we propose that there is a repeat expansion DNA damage response (REDD) pathway that acts to prevent repeat expansions in the genome (Table 1)(81, 82). Expansions may arise at susceptible genomic loci (e.g. a *HTT* gene with >35 CAG repeats)

through aberrant processing of repeats by DNA repair pathways such as mismatch repair but the REDD pathway could act either to prevent expansion in the first place, or to repair expansions after they have occurred. Similar homeostatic mechanisms exist in cells to maintain repeat structures at crucial genomic loci such as centromeres and telomeres (79, 83, 84). The precise mechanism and implications for repeat disorders and normal cellular function require further work.

Future work

Analysis techniques for complex traits (e.g. large-scale GWAS) are showing promise for detecting modifying loci for single-gene disorders. To improve the power of such studies, as in common complex disorders, collecting larger sample sizes is necessary and in HD and the other trinucleotide repeat disorders, quite feasible. Including other diseases with similar mutational mechanisms, such as the spinocerebellar ataxias (SCAs) (52), myotonic dystrophy, Friedreich's ataxia and Fragile X, in meta-analyses, may also increase detection power. Challenges for the future are to collect more subjects, to develop more informative phenotypes and to efficiently integrate different types of genome wide data: the latter is likely to be led by studies in common disease and large international consortia generating such data.

Such phenotype improvements will also augment sequencing studies, targeted, exome and whole genome, that will be increasingly used to highlight rare loss of function and coding variation, as in common diseases. Here sample sizes are currently an important consideration in which type of sequencing to deploy. In most of these diseases the sample sizes that can be achieved are well below 100,000. Hence rare variants will need to show substantial effect sizes to be detected (21). This in turn means that most such effects will ablate gene function by substantially altering gene expression, splicing or causing functionally important amino acid changes.

These genetic studies highlight new hypotheses of disease causation. This can generate targets for drug development with the advantage that they are based on disease modification in people. However, to understand the detailed mechanisms through which such modifiers and potential drug targets operate, the genetic information will need to be used to power novel biological studies in appropriate model systems in cells and animals. This may well allow the definition of new mechanisms and pathways operating in disease and in turn, the generation of assays with disease-relevant outcomes. Much of this is in the future - but relatively feasible - and these initial studies demonstrate the potential power of searching for genetic modifiers in Mendelian disease to shed light on fundamental disease biology and open up new pathways for therapeutic intervention.

Acknowledgements

We thank Branduff McAllister for preparing Figure 1.

References

1. McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

2. Scoto,M., Finkel,R.S., Mercuri,E. and Muntoni,F. (2017) Therapeutic approaches for spinal muscular atrophy (SMA). *Gene Ther.*, 10.1038/gt.2017.45.

3. Finkel,R.S., Chiriboga,C.A., Vajsar,J., Day,J.W., Montes,J., De Vivo,D.C., Yamashita,M., Rigo,F., Hung,G., Schneider,E., *et al.* (2016) Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. *Lancet (London, England)*, **388**, 3017–3026.

4. Bates,G.P., Dorsey,R., Gusella,J.F., Hayden,M.R., Kay,C., Leavitt,B.R., Nance,M., Ross,C.A., Scahill,R.I., Wetzel,R., *et al.* (2015) Huntington disease. *Nat. Rev. Dis. Prim.*, **1**, 15005.

5. Nelson,M.R., Tipney,H., Painter,J.L., Shen,J., Nicoletti,P., Shen,Y., Floratos,A., Sham,P.C., Li,M.J., Wang,J., *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.*, **47**, 856–860.

6. Wexler,N.S., Lorimer,J., Porter,J., Gomez,F., Moskowitz,C., Shackell,E., Marder,K., Penchaszadeh,G., Roberts,S.A., Gayan,J., *et al.* (2004) Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl. Acad. Sci. U.*

S. A., **101**, 3498–3503.

7. Langbehn,D.R., Brinkman,R.R., Falush,D., Paulsen,J.S., Hayden,M.R. and International Huntington's Disease Collaborative Group (2004) A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.*, **65**, 267–77.

8. Kay,C., Collins,J.A., Miedzybrodzka,Z., Madore,S.J., Gordon,E.S., Gerry,N., Davidson,M., Slama,R.A. and Hayden,M.R. (2016) Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology*, **87**, 282–288.

9. Dorsey,E.R. (2012) Characterization of a large group of individuals with huntington disease and their relatives enrolled in the COHORT study. *PLoS One*, **7**, e29522.

10. Paulsen,J.S., Hayden,M., Stout,J.C., Langbehn,D.R., Aylward,E., Ross,C.A., Guttman,M., Nance,M., Kieburtz,K., Oakes,D., *et al.* (2006) Preparing for preventive clinical trials: the Predict-HD study. *Arch. Neurol.*, **63**, 883–890.

11. Orth,M. and Network,T.E.H. 's D. (2011) Observing Huntington's disease: the European Huntington's Disease Network's REGISTRY. *J. Neurol. Neurosurg. Psychiatry* , **82**, 1409–1412.

12. Tabrizi,S.J., Scahill,R.I., Owen,G., Durr,A., Leavitt,B.R., Roos,R.A., Borowsky,B., Landwehrmeyer,B., Frost,C., Johnson,H., *et al.* (2013) Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet. Neurol.*, **12**, 637–649.

13. Walker,T., Ghosh,B. and Kipps,C. (2017) Assessing Decline: Visualising Progression in Huntington's Disease using a Clinical Dashboard with Enroll-HD Data. *J. Huntingtons. Dis.*, 10.3233/JHD-170234.

14. Li,J.-L., Hayden,M.R., Almqvist,E.W., Brinkman,R.R., Durr,A., Dode,C., Morrison,P.J., Suchowersky,O., Ross,C.A., Margolis,R.L., *et al.* (2003) A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study. *Am. J. Hum. Genet.*, **73**, 682–687.

15. Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

16. Gusella,J.F., MacDonald,M.E. and Lee,J.-M. (2014) Genetic modifiers of Huntington's disease. *Mov. Disord.*, **29**, 1359–65.

17. Button,K.S., Ioannidis,J.P.A., Mokrysz,C., Nosek,B.A., Flint,J., Robinson,E.S.J. and Munafo,M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.*, **14**, 365–376.

18. Ioannidis,J.P.A., Tarone,R. and McLaughlin,J.K. (2011) The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, **22**, 450–456.

19. Huntington's,G.M. of and Disease Consortium,(GeM-HD) (2015) Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*, **162**, 516–526.

20. Visscher,P.M., Brown,M.A., McCarthy,M.I. and Yang,J. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.

21. Goldstein,D.B., Allen,A., Keebler,J., Margulies,E.H., Petrou,S., Petrovski,S. and Sunyaev,S. (2013) Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.*, **14**, 460–

470.

22. Gratten,J., Wray,N.R., Keller,M.C. and Visscher,P.M. (2014) Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.*, **17**, 782–790.

23. Karlin,S. and McGregor,J. (1972) The evolutionary development of modifier genes. *Proc. Natl. Acad. Sci. U. S. A.*, **69**, 3611–3614.

24. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

25. Chang,C.C., Chow,C.C., Tellier,L.C., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.

26. McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P.A. and Hirschhorn,J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

27. Pe'er,I., Yelensky,R., Altshuler,D. and Daly,M.J. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, **32**, 381–385.

28. Xu,C., Tachmazidou,I., Walter,K., Ciampi,A., Zeggini,E. and Greenwood,C.M.T. (2014) Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.*, **38**, 281–290.

29. Chapman,J.M., Cooper,J.D., Todd,J.A. and Clayton,D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.*, **56**, 18–31.

30. Moss,D.J.H., Pardiñas,A.F., Langbehn,D., Lo,K., Leavitt,B.R., Roos,R., Durr,A., Mead,S., Coleman,A., Santos,R.D., *et al.* (2017) Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.*, 10.1016/S1474-4422(17)30161-8.

31. McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A., Kang,H.M., Fuchsberger,C., Danecek,P., Sharp,K., *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.

32. Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.

33. Lek,M., Karczewski,K.J., Minikel,E. V, Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

34. de Bakker,P.I.W., Ferreira,M.A.R., Jia,X., Neale,B.M., Raychaudhuri,S. and Voight,B.F. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, R122-8.

35. Morris,A.P. and Zeggini,E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.

36. Wu,M.C., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

37. Lee,S., Wu,M.C. and Lin,X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

38. Peloso,G.M., Rader,D.J., Gabriel,S., Kathiresan,S., Daly,M.J. and Neale,B.M. (2016) Phenotypic extremes in rare variant study designs. *Eur. J. Hum. Genet.*, **24**, 924–930.

39. Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

40. Price,A.L., Patterson,N.J., Plenge,R.M., Weinblatt,M.E., Shadick,N.A. and Reich,D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

41. Freedman,M.L., Reich,D., Penney,K.L., McDonald,G.J., Mignault,A.A., Patterson,N., Gabriel,S.B., Topol,E.J., Smoller,J.W., Pato,C.N., *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, **36**, 388–393.

42. Bulik-Sullivan,B.K., Loh,P.-R., Finucane,H.K., Ripke,S., Yang,J., Patterson,N., Daly,M.J., Price,A.L. and Neale,B.M. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

43. Chen,W.-M. and Abecasis,G.R. (2007) Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.*, **81**, 913–926.

44. Yang,J., Zaitlen,N.A., Goddard,M.E., Visscher,P.M. and Price,A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

45. Holmans,P. (2010) Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.*, **72**, 141–179.

46. Wang,K., Li,M. and Hakonarson,H. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 843–854.

47. de Leeuw,C.A., Mooij,J.M., Heskes,T. and Posthuma,D. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.*, **11**, e1004219.

48. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

49. Lambert,J.C., Ibrahim-Verbaas,C.A., Harold,D., Naj,A.C., Sims,R., Bellenguez,C., DeStafano,A.L., Bis,J.C., Beecham,G.W., Grenier-Boley,B., *et al.* (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.

50. Convergent genetic and expression data implicate immunity in Alzheimer's disease. (2015) *Alzheimers. Dement.*, **11**, 658–671.

51. Hensman Moss,D.J., Flower,M.D., Lo,K.K., Miller,J.R.C., van Ommen,G.-J.B., 't Hoen,P.A.C., Stone,T.C., Guinee,A., Langbehn,D.R., Jones,L., *et al.* (2017) Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Sci. Rep.*, **7**, 44849.

52. Bettencourt,C., Moss,D.H., Flower,M., Wiethoff,S., Brice,A., Goizet,C., Stevanin,G., Koutsis,G., Karadima,G., Panas,M., *et al.* (2016) DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann. Neurol.*, 10.1002/ana.24656.

53. Jones,L., Houlden,H. and Tabrizi,S.J. (2017) DNA repair in the trinucleotide repeat disorders. *Lancet Neurol.*, **16**, 88–96.

54. Schmidt,M.H.M. and Pearson,C.E. (2016) Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst).*, **38**, 117–26.

55. Morales,F., Vásquez,M., Santamaría,C., Cuenca,P., Corrales,E. and Monckton,D.G. (2016) A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst).*, **40**, 57–66.

56. Giambartolomei,C., Vukcevic,D., Schadt,E.E., Franke,L., Hingorani,A.D., Wallace,C. and Plagnol,V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.

57. Won,H., de la Torre-Ubieta,L., Stein,J.L., Parikshak,N.N., Huang,J., Opland,C.K., Gandal,M.J., Sutton,G.J., Hormozdiari,F., Lu,D., *et al.* (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, **538**, 523–527.

58. Manchia,M., Cullis,J., Turecki,G., Rouleau,G.A., Uher,R. and Alda,M. (2013) The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One*, **8**, e76295.

59. Luo,X., Stavrakakis,N., Penninx,B.W., Bosker,F.J., Nolen,W.A., Boomsma,D.I., de Geus,E.J., Smit,J.H., Snieder,H., Nolte,I.M., *et al.* (2016) Does refining the phenotype improve replication rates? A review and replication of candidate gene studies on Major Depressive Disorder and Chronic Major Depressive Disorder. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, **171B**, 215–236.

60. Corvol,H., Blackman,S.M., Boelle,P.-Y., Gallins,P.J., Pace,R.G., Stonebraker,J.R., Accurso,F.J., Clement,A., Collaco,J.M., Dang,H., *et al.* (2015) Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat. Commun.*, **6**, 8382.

61. Pegoraro,E., Hoffman,E.P., Piva,L., Gavassini,B.F., Cagnin,S., Ermani,M., Bello,L., Soraru,G., Pacchioni,B., Bonifati,M.D., *et al.* (2011) SPP1 genotype is a determinant of disease severity in Duchenne muscular dystrophy. *Neurology*, **76**, 219–226.

62. Bello,L., Flanigan,K.M., Weiss,R.B., Spitali,P., Aartsma-Rus,A., Muntoni,F., Zaharieva,I., Ferlini,A., Mercuri,E., Tuffery-Giraud,S., *et al.* (2016) Association Study of Exon Variants in the NF-kappaB and TGFbeta Pathways Identifies CD40 as a Modifier of Duchenne Muscular Dystrophy. *Am. J. Hum. Genet.*, **99**, 1163–1171.

63. Bello,L., Kesari,A., Gordish-Dressman,H., Cnaan,A., Morgenroth,L.P., Punetha,J., Duong,T., Henricson,E.K., Pegoraro,E., McDonald,C.M., *et al.* (2015) Genetic modifiers of ambulation in the Cooperative International Neuromuscular Research Group Duchenne Natural History Study. *Ann. Neurol.*, **77**, 684–696.

64. Cannavo,E., Gerrits,B., Marra,G., Schlapbach,R. and Jiricny,J. (2007) Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. *J. Biol. Chem.*, **282**,

2976–2986.

65. Smogorzewska,A., Desetty,R., Saito,T.T., Schlabach,M., Lach,F.P., Sowa,M.E., Clark,A.B., Kunkel,T.A., Harper,J.W., Colaiacovo,M.P., *et al.* (2010) A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol. Cell*, **39**, 36–47.

66. Jin,H. and Cho,Y. (2017) Structural and functional relationships of FAN1. *DNA Repair (Amst).*, 10.1016/j.dnarep.2017.06.016.

67. Wheeler,V.C., Lebel,L.-A., Vrbanac,V., Teed,A., te Riele,H. and MacDonald,M.E. (2003) Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. *Hum. Mol. Genet.* , **12**, 273–281.

68. Dragileva,E., Hendricks,A., Teed,A., Gillis,T., Lopez,E.T., Friedberg,E.C., Kucherlapati,R., Edelmann,W., Lunetta,K.L., MacDonald,M.E., *et al.* (2009) Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol. Dis.*, **33**, 37–47.

69. Bourn,R.L., De Biase,I., Pinto,R.M., Sandi,C., Al-Mahdawi,S., Pook,M.A. and Bidichandani,S.I. (2012) Pms2 suppresses large expansions of the (GAA.TTC)n sequence in neuronal tissues. *PLoS One*, **7**, e47085.

70. Zhao,X.-N., Kumari,D., Gupta,S., Wu,D., Evanitsky,M., Yang,W. and Usdin,K. (2015) Mutsbeta generates both expansions and contractions in a mouse model of the Fragile X-associated disorders. *Hum. Mol. Genet.*, **24**, 7087–7096.

71. Tomé,S., Manley,K., Simard,J.P., Clark,G.W., Slean,M.M., Swami,M., Shelbourne,P.F., Tillier,E.R.M., Monckton,D.G., Messer,A., *et al.* (2013) MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS Genet.*, **9**, e1003280.

72. Rodriguez,G.P., Romanova,N. V, Bao,G., Rouf,N.C., Kow,Y.W. and Crouse,G.F. (2012) Mismatch repair-dependent mutagenesis in nondividing cells. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 6153–8.

73. Owen,B.A.L., Yang,Z., Lai,M., Gajec,M., Gajek,M., Badger,J.D., Hayes,J.J., Edelmann,W., Kucherlapati,R., Wilson,T.M., *et al.* (2005) (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat. Struct. Mol. Biol.*, **12**, 663–70.

74. Gonitel,R., Moffitt,H., Sathasivam,K., Woodman,B., Detloff,P.J., Faull,R.L.M. and Bates,G.P. (2008) DNA instability in postmitotic neurons. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 3467–3472.

75. Swami,M., Hendricks,A.E., Gillis,T., Massood,T., Mysore,J., Myers,R.H. and Wheeler,V.C. (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.*, **18**, 3039–3047.

76. Hubert,L., Lin,Y., Dion,V. and Wilson,J.H. (2011) Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. *Hum. Mol. Genet.*, **20**, 4822–4830.

77. Kovtun,I. V, Liu,Y., Bjoras,M., Klungland,A., Wilson,S.H. and McMurray,C.T. (2007) OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature*, **447**, 447–452.

78. Budworth,H., Harris,F.R., Williams,P., Lee,D.Y., Holt,A., Pahnke,J., Szczesny,B., Acevedo-Torres,K., Ayala-Peña,S. and McMurray,C.T. (2015) Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLoS Genet.*, **11**, e1005267.

79. McMurray,C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet*, **11**, 786–799.

80. Pearl,L.H., Schierz,A.C., Ward,S.E., Al-Lazikani,B. and Pearl,F.M.G. (2015) Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer*, **15**, 166–180.

81. Ceccaldi,R., Sarangi,P. and D'Andrea,A.D. (2016) The Fanconi anaemia pathway: new players and new functions. *Nat. Rev. Mol. Cell Biol.*, **17**, 337–349.

82. Brown,J.S., O'Carrigan,B., Jackson,S.P. and Yap,T.A. (2017) Targeting DNA Repair in Cancer: Beyond PARP Inhibitors. *Cancer Discov.*, **7**, 20–37.

83. Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–40.

84. Kim,J.C., Harris,S.T., Dinter,T., Shah,K.A. and Mirkin,S.M. (2017) The role of break-induced replication in large-scale expansions of (CAG)n/(CTG)n repeats. *Nat. Struct. Mol. Biol.*, **24**, 55–60.

Table 1 Parallels between the mismatch repair pathway and the proposed repeat expansion DNA damage response pathway

Pathway	**Mismatch repair**	**Repeat expansion DNA damage response**
Source of DNA damage or structural change	DNA polymerase proofreading errors	Repeat expansions in DNA
Damage sensors	MSH2, MSH3, MSH6, MLH1, PMS2	**FAN1?, MSH3**, MLH1, MSH6, PMS2, PMS1, **RRM2B**?
Effector proteins	LIG1, EXO1, POLD	**FAN1**?, LIG1, POLD, **RRM2B**?

Proteins implicated by genetics in the Repeat expansion DNA damage response are in bold where they are in genome-wide significant loci and in normal font where implicated by pathway analyses (19, 30, 52). After Brown et al. (82).

Figure Legends

Figure 1. Plot of motor age at onset vs. CAG length in the REGISTRY sample, showing the expected age of onset predicted by (7). Note that below 41 CAG repeats, the model tends to overestimate age at onset, whereas the opposite is true over 56 CAG repeats. These differences are likely due to ascertainment bias.

Figure 2. Manhattan plot of the results of the GeM GWAS of motor age at onset. Physical location is plotted on the x-axis by chromosome and –log (association p-value) on the y-axis. Each data point corresponds to an individual SNP. The horizontal dotted line corresponds to the criterion for genome-wide significance ($p=5x10^{-8}$). From ref 19, © Elsevier Inc., with permission.

Figure 3. Power to detect variants at genome-wide significance ($p=5x10^{-8}$) depends on sample size (N) and the proportion of phenotypic variance accounted for by additive effects of the variant alleles, denoted by h^2 (heritability).

Figure 4. Example of a q-q plot, taken from the GeM GWAS of motor AAO. Observed –log(association p-values) plotted on the y-axis, expected –log(p-values) in the absence of association plotted on the x-axis. The divergence of the observed p-values above their expected values (red line) indicates the presence of true associations. Inflation factor gives the extent of systematic inflation of test statistics (1=no inflation). From ref 19, © Elsevier Inc., with permission.

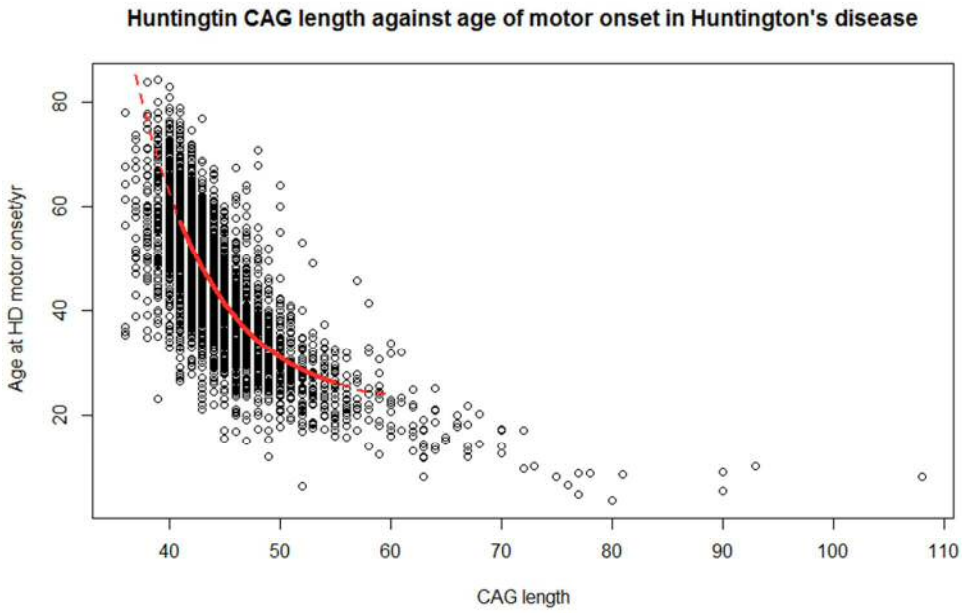
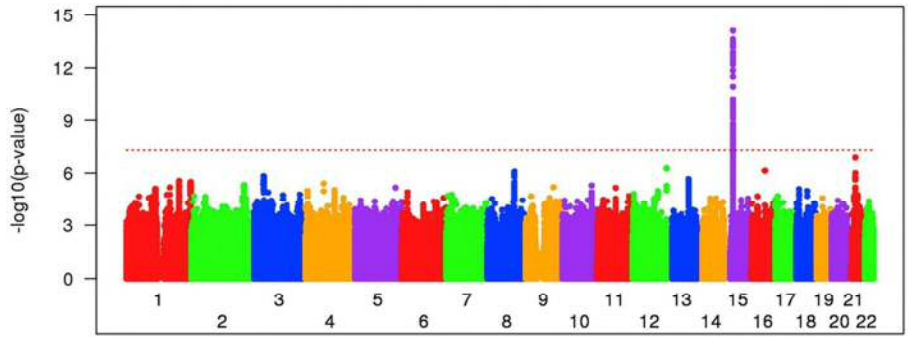


Figure 1. Plot of motor age at onset vs. CAG length in the REGISTRY sample, showing the expected age of onset predicted by (7). Note that below 41 CAG repeats, the model tends to overestimate age at onset, whereas the opposite is true over 56 CAG repeats. These differences are likely due to ascertainment bias.



Cell 2015 162, 516-526DOI: (10.1016/j.cell.2015.07.003)

Figure 2. Manhattan plot of the results of the GeM GWAS of motor age at onset. Physical location is plotted on the x-axis by chromosome and –log (association p-value) on the y-axis. Each data point corresponds to an individual SNP. The horizontal dotted line corresponds to the criterion for genome-wide significance (p=5x10-8). From ref 19, © Elsevier Inc., with permission.

338x190mm (96 x 96 DPI)

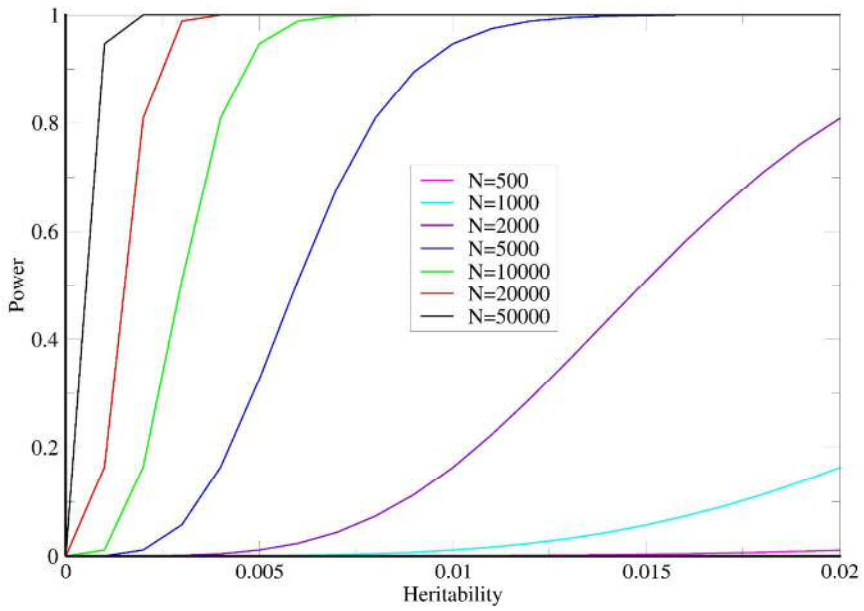


Figure 3. Power to detect variants at genome-wide significance (p=5x10-8) depends on sample size (N) and the proportion of phenotypic variance accounted for by additive effects of the variant alleles, denoted by h2 (heritability).

1047x809mm (96 x 96 DPI)

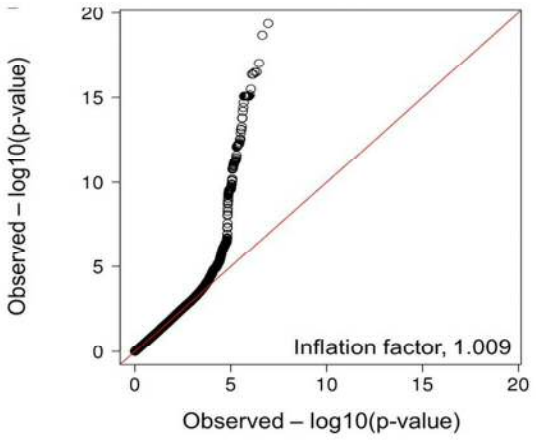


Figure 4. Example of a q-q plot, taken from the GeM GWAS of motor AAO. Observed –log(association p-values) plotted on the y-axis, expected –log(p-values) in the absence of association plotted on the x-axis. The divergence of the observed p-values above their expected values (red line) indicates the presence of true associations. Inflation factor gives the extent of systematic inflation of test statistics (1=no inflation). From ref 19, © Elsevier Inc., with permission.

338x190mm (96 x 96 DPI)