Genomics and successful aging: grounds for renewed optimism?

Pilling LC [1*], Harries LW [2*], Powell J [1], Llewellyn DJ [1],Ferrucci L[3], Melzer D [1]

1. *Epidemiology and Public Health Group, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter EX1 2LU, UK*
2. *Institute of Biomedical and Clinical Sciences, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter EX1 2LU, UK*
3. *National Institute on Aging, Baltimore, Maryland, USA*

* - *THESE AUTHORS CONTRIBUTED EQUALLY TO THIS PUBLICATION*

Correspondence: Professor David Melzer**,** Peninsula Medical School, Barrack Road, Exeter EX2 5DW, United Kingdom. Email david.melzer@pms.ac.uk

Abstract

Successful aging depends in part on delaying age-related disease (ARD) onsets until later in life. Conditions including coronary artery disease, Alzheimer's disease, prostate cancer and type 2 diabetes are moderately heritable. Genome Wide Association Studies (GWAS) have identified many risk associated single nucleotide polymorphisms (SNPs) for these conditions, but much heritability remains unaccounted for. Nevertheless a great deal is being learned: here we review ARD associated SNPs and identify key underlying pathways including lipid handling, specific immune processes, early tissue development and cell cycle control. Most ARD associated SNPs do not affect coding regions of genes or protein makeup, but instead influence regulation of gene expression. Recent evidence indicates that evolution of gene regulatory sites is fundamental to interspecies differences. Animal models relevant to human aging may therefore need to focus more on gene regulation rather than testing major disruptions to fundamental pathway genes.

Recent larger scale human studies of in-vivo genome wide expression (notably from the InCHIANTI aging study) have identified changes in splicing, the 'fine tuning' of protein sequences, as a potentially important factor in decline of cellular function with age. Studies of expression with muscle strength and cognition have shown striking concordance with certain mice models of muscle repair and beta-amyloid phagocytosis respectively.

The emerging clearer picture of the genetic architecture of ARDs in humans is providing new insights into the underlying pathophysiological pathways involved. Translation of genomics into new approaches to prevention, tests and treatments to extend successful aging is therefore likely in the coming decades.

1 Introduction

While there is much debate about the processes driving human aging, there is little doubt that genetic influences play a significant role [1]. Humans clearly live very much longer than the currently favored laboratory models of aging, and such interspecies differences in reproductively 'fit' lifespan must have an inherited genetic foundation. Within human populations, environmental and behavioral exposures are important but at least a quarter of life expectancy variation in twin or family studies is attributable to inherited genetic or epigenetic factors [2]. Age related conditions such as type 2 diabetes, myocardial infarction, common cancers and Alzheimer's disease typically have onsets after the fourth decade of life; 'successful' agers delay these onsets until relatively late in life [3]. Many aging traits and diseases show moderate heritability, including cardiovascular disease [4] and impaired physical functioning [5], independent of known environmental risk factors.

Inherited genetic variations in DNA sequences come in several forms, including single-nucleotide polymorphisms (SNPs), insertions/deletions and copy-number variations. Of the 54 million variants documented in dbSNP [6] >88% are single base-pair substitutions or 'SNPs' [7]. SNPs located close to one-another are more likely to be inherited together (called linkage disequilibrium) [8] and this tendency coupled with new array technology allows the assaying of millions of SNPs per sample, capturing most common variation.

The effect a variant has on a given phenotype depends on its position in the genome. Within each gene only a small proportion of the DNA sequence is used as a template for assembling amino-acids into specific proteins [9] (Fig.1); these DNA sections are termed

exons. Exon sequences are 'transcribed' (in part or in full) from DNA into messenger RNA (mRNA), which undergoes 'processing' before forming the template for protein production. If the DNA sequence within an exon is changed, this could have profound consequences to the final structure of the protein synthesized from the gene. However, the protein-coding regions of genes are highly conserved, even between species [10]. This highlights an intriguing question; if highly similar genes are used in the growth and development of different organs, tissues and cell-types in individuals, how can so much phenotypic variation occur? The answer lies in the complex machinery that regulate gene transcription and protein production in different cells types and in response to specific stimuli.

Transcription can vary in amount (i.e. particular mRNAs can be up or down regulated) but can also produce specific versions or 'isoforms' of proteins, in a fine tuning system that contributes to specialization of cell types and tissues. This involves the removal, or 'splicing', of introns - sequences of DNA that separate the exons from one-another - allowing many protein products to be derived from a single gene. Transcription and RNA processing require extensive regulation by transcription factors binding to sites in proximity to the gene, which can be many thousands of base pairs away. DNA variants located within these regulatory binding sites can alter when, where and for how long a gene is expressed by altering the affinity of the various transcription factors to that particular regulatory sequence (Fig.1).

Differential regulation of genes – both in location and timing – between species likely contributes to the divergence and isolation of different species [11]. Human studies are lacking, but work in mice has shown regulatory differences between sub-species [12]. Affordable genomic array technologies that measure the expression of thousands of

genes are now being used in human population studies. Progress in this field is likely to receive a boost in the coming years as studies use large sample collections to investigate the differential regulation of gene expression in humans, and how these differences influence disease and longevity [13].

Epigenetic changes, such as DNA methylation and histone modifications, are also heritable and affect mRNA expression [14]. These changes to the structure of DNA do not affect the sequence, and can change with advancing age [15]. It appears that genetic and epigenetic variations exert their effects by altering either the amount of RNA transcribed from a gene, or the relative proportion of alternatively expressed isoforms produced by the alternative splicing mechanisms. These ultimately affect other downstream elements of the pathway, such as binding partners or inhibitors, resulting in a change in phenotype. It will therefore be necessary for future research programs to integrate genetic variation, epigenetics and associated gene expression profiles to understand the origins of heritable traits and diseases. Such mechanistic understanding may contribute to the discovery of new therapeutic targets for aging pathologies [16].

2 SNPs and age related diseases

In the early 2000s, many correlations between SNPs and specific disease were reported from relatively small-scale volunteer and population studies, but unfortunately multiple statistical testing and a bias towards reporting positive results contributed to many of these reports being unreliable, with inflated effect sizes and replication efforts failing to confirm claims [17,18].By 2006 Genome-wide association studies (GWAS) had became

popular, using large numbers of subjects to identify the SNPs most strongly associated with a wide range of diseases and phenotypes. Stringent rules for statistical significance and large collaborations aimed at pooling data into enormous meta-analyses have produced many robust findings: in aging the CHARGE consortium have led the way (http://web.chargeconsortium.com/) [19].

Many striking findings have emerged from GWAS studies to deepen our understanding of successful aging. One of the first breakthroughs was in a GWAS study of age-related macular degeneration (AMD), which identified a SNP related to complement factor H [20], this has since been replicated in larger studies [21,22]. As complement factor H (part of a specialized inflammatory cascade) had previously attracted limited attention in AMD research, GWAS had provided a powerful insight into the underlying mechanisms of the condition; at present 40% of the heritability of AMD has been accounted for by 5 SNPs [23]. Similar insights into the mechanisms underlying menopause, a key aging trait in humans, have very recently emerged from a meta-analysis of numerous studies [24]. Variants related to DNA repair and immune pathways emerged as important.

Using the catalogue of published GWAS results [25] we have re-examined the results for 4 common age-related diseases (ARDs) for this review, to illustrate the lessons being learned from GWAS about the biological architecture of these conditions. We have extracted the SNP results for Alzheimer's disease (AD), cardio-vascular disease (CVD), prostate cancer (PC) and type 2 diabetes mellitus (T2D). We included only robustly associated variants; studies with more than 500 cases/controls were selected, and SNPs with genome wide statistical significance (associations at $p<1*10^{-8}$). At the time of writing (January 2012) 75 SNPs associated with CVD met these criteria, 67 for T2D, 55 for PC and 28 for AZ (Table 1). As some of these SNPs are in linkage with one-another we

have only considered the nearest genes in our analysis: for instance gene *TCF7L2* was only included once, even though several T2D SNPs map to it.

Of the SNPs associated with these 4 ARDs only one is not unique to a single disease (rs2075650 increases risk of both Alzheimer's and cardiovascular disease). A critical feature of the identified SNPs is that very few are in protein-coding regions (4% of these ARD related SNPs are classified as exonic), therefore most do not directly alter the amino acid sequence of the protein product (Fig.2). We therefore identified the nearest genes to each SNP, as the most likely genes affected, although in most cases the mechanisms of effect have yet to be confirmed. Only 5 of the 89 ARD SNP nearest genes are associated with more than one ARD; *MTHFD1L* (AD and CVD), *OASL* (CVD and T2D), *SLC22A3* (CVD and PC), *THADA* (PC and T2D) and *TOMM40* (a proxy for the ApoE haplotype, for AD and CVD). In general therefore, our exemplar ARD appear to be complex polygenic traits (i.e. influenced by many genes of small effect) and these risk traits appear to be largely inherited separately in each individual. We note however that much variation remains unaccounted for, but it is likely that this unexplained heritability will be accounted for by large numbers of very small effect common variants, rare moderate effect variants [26] or perhaps epigenetic effects. Another contributor to the missing heritability might be that heritability estimates are uncertain: twin studies have estimated the genetic heritability of AD anywhere between 36% [27] and 74% [28], partly depending on definitions used.

2.1 Biological pathways implicated in ARD SNPs

As noted for macular degeneration, the results from SNP studies can give a powerful indication of the biological pathways contributing to each trait. Using this principle we have analyzed the GWAS results for 4 common ARDs to identify the affected biological pathways. Using the published GWAS results [25] we extracted the SNPs listed for the 4 common ARDs mentioned above. SNP annotations (mapped genes) were determined using SCAN [29]. We used BiNGO [30] software to determine which biological processes (as defined by Gene Ontology [31]) were statistically overrepresented in the sets of genes derived from the GWAS SNPs for each of the ARDs. The biological processes involved in the ARDs are listed in Table 1 (lists have been summarized to eliminate redundancy - full gene lists and BiNGO results can be found in supplementary tables 1 and 2 respectively).

The findings for type 2 diabetes (T2D) illustrate how this approach can identify the mechanisms underlying a disease. T2D-related pathways included regulation of insulin secretion and glucose homeostasis, clearly confirming the known etiology: if the causes of T2D were unknown the SNPs would have pointed us in the right direction. However, in addition to the known pathways, cell cycle arrest and RNA modification were also prominent. The cell cycle control genes associated with T2D seem to be involved in maintaining or compromising beta cell mass in the pancreas, while recent expression studies have shown RNA processing changes to be strongly age-associated in humans [32].

Biological processes highlighted by the genes with mapped variants associated with CVD are mostly related to lipid digestion, transport, localization and clearance, although the largest single effect variant (at 9p21) is thought to influence repair and regrowth of

the intima of arteries, and is associated with peripheral arterial disease [33] in older populations and excess mortality until late in life [34].

SNPs affecting the risk of developing prostate cancer are largely in development and proliferation related processes. Although some of the processes seem specific to certain organs, the underlying development mechanisms involve similar pro-growth and development genes. Given that cancer is a disease of faulty control over cell proliferation or apoptosis, components of these pathways, or similar, would be expected to be overrepresented in the genes associated with susceptibility to prostate cancer.

Given that the causes of Alzheimer's disease (AD) are still largely mysterious, the GWAS results for AD are of particular interest. Pathways implicated in AD based on the mapped SNPs were mostly related to lipid and cholesterol processing, and specific inflammatory pathways. This may reflect the role of neuro-inflammation, implicated as possibly related to beta-amyloid deposition, a core feature of AD [35]. The best studied loci for AD is the ApoE Epsilon 4 (E4) haplotype, identified as disrupting the normal functioning of ApoE in the distribution of lipids in the central nervous system [36].

2.2 The genetic architecture of age-related conditions

Although a few ARD associated SNPs may involve amino acid changing variants (e.g. the R325W variant in *SLC30A8* gene, the P12A variant of the *PPARG* gene and the *E23K* variant of the *KCNJ11* gene for type 2 diabetes) [37,38], it is striking that over 98% of all variants are located in 'non-coding regions' of the genome [6,7] (Fig.2), only 60% are predicted to alter the final protein sequence (similar proportions in GWAS results: <5% of reported SNPs change the final protein sequence, ~46% intergenic, ~41% intronic, the rest are in other untranslated regions (UTRs)). Although these non-translated regions of

the genome produce no proteins, there is growing evidence that they are far from non-functional [39].

Genetic variants in non-translated regions may affect factors necessary for epigenetic modifications, such as histone modification, chromatin remodeling, miRNA (microRNA) regulation of transcription or DNA methylation. These factors are recognized as important regulators of the activity of many genes, including those involved in glucose metabolism [40,41]. Variants located within a regulatory region of a transcript (for instance the UTRs) may interfere with the efficiency of transcription or translation (Fig.1), as has been recently shown for several human and rodent genes [42–44]. Variants located in introns may be present in uncharacterized exons, or may influence mRNA splicing [45]; the production of aberrant splice products is an important disease mechanism [46,47]. Finally, SNPs located in intergenic regions may act through natural antisense transcripts (NATs) [48] or be located in uncharacterized regulatory regions as was described in 2008 for the *OCA2* gene involved in determination of eye color in humans [49].

Most of the non-coding regulatory regions are involved in the generation of tissue specificity, adaptation or response to the intra- and extracellular environment. Alternative splicing, for example, is a pivotal mechanism in ensuring cellular plasticity. It is noteworthy that organisms such as the dinoflagellates that do not exhibit phenomena such as alternative splicing have a great many more genes (up to 87,688) [50] than do higher organisms such as humans (~23,000 genes). In order to make any further inference about how variation at the DNA level leads to changes in gross phenotype we must now look downstream at changes in gene expression associated with genetic variation, aging and age-related disease.

2.3 Comparison with laboratory models of aging

Laboratory models typically used to study aging, such as *C. elegans* (nematode worm) and *Mus musculus* (mice), have drastically shorter lifespans than our own (~3 weeks [51] and ~3 years [52] respectively, versus a 122 year maximum for humans thus far [53]). In some respects these models are ideal for study in the laboratory (higher turnover, controlled environment), yet applicability to humans can be elusive due to the inherent inter-species differences. For instance, whilst we may share a large portion (over 98.5% identical in protein-coding regions[54]) of our genome with chimpanzees - our closest relative in evolutionary terms - we have evolved distinct phenotypic differences, including aging at a slower rate. These differences will only be more exaggerated in more distantly related species (such as the worm and mouse models). There are however similarities between aged humans and aged model organisms; they all tend to have decreasing overall fitness, and therefore studies using model organisms continue as they may be at least indicative of some aging mechanisms in humans.

Extensions to lifespan in model organisms are mostly associated with disruption to fundamental metabolic pathways, with target genes showing conservation across species. For example, 'knockout' mutations in the IGF1/insulin pathway can double lifespan in *C.elegans* in the laboratory environment [55], and some have argued that this is directly relevant to humans. However, aging traits in higher organisms are likely to be largely influenced by complex and cell type specific regulatory sites, as noted above. Given the barriers against experimental studies of in-vivo regulation and aging in humans, epidemiological studies combining genetic and gene expression data will be crucial in determining the relevance to humans of molecular mechanisms identified in laboratory models.

3 Gene expression arrays

Messenger RNA (mRNA) is a key intermediate between DNA and proteins (Fig.1). Cells tightly regulate the amount and form of new proteins produced in response to intra- and extra-cellular signals. The amount of mRNA being produced at any point in time is a good proxy for the demand within the cell for the corresponding protein. This allows us a unique insight into how, at a cellular level, organisms react to stimuli. By profiling this expression of mRNA we can infer the molecular mechanisms associated with phenotypes such as disease, or genotype. For example Song *et al.* [56] have shown that ~17% of genes are differentially expressed between 2 populations of distinct ethnic decent, which is due to differences in inherited variation and lifestyle (environment) [57].

Microarrays are used to quantify the expression of thousands of mRNA molecules from a single sample. Many gene expression studies in laboratory models of aging have been reported: see the database of age-related genes [58]. Until recently however studies in humans were based on very small sample sizes, risking the same reliability problems that beset candidate SNP studies. As microarray cost has come down drastically in recent years it is now feasible to study large numbers of samples at once.

In humans major early studies focused on, for example, gene expression in muscle biopsy material [59] and identifying genes associated with age in the San Antonio Family Heart Study frozen lymphocyte samples[60]. Comparing results across studies and with the in-vivo situation has been difficult as sample collection, cell separation and storage

processes can alter mRNA concentrations in unpredictable ways. Recent work by the author group and collaborators in the InCHIANTI study of aging [61] have used recent technology to collect *in-vivo* mRNA levels to explore aging and age-related phenotypes in a cohort of 698 individuals that are representative of the general population. Microarrays were used to quantify the expression of 16,571 mRNA transcripts in each individual, from circulating leukocytes in blood. As blood is relatively easily collected in large population samples and is likely to be the main tissue for clinical application of gene expression signatures, blood derived leukocytes are the first choice for in-vivo studies, with research programs to follow up findings in other tissues or in-vitro.

After adjustment for potentially confounding factors, only 2% (295) of the 16,571 probes tested in InCHIANTI were robustly associated with age [32], and the majority of these were down-regulated. Using gene ontology information in a pathways-based analysis we found that the most deregulated biological processes across the age range are related to RNA processing [31,62]. mRNA splicing is the process that allows a single gene to code for more than one transcript ('isoform'), which may have a different function, temporal or spatial expression pattern to an alternative isoform from the same gene. The implications of this finding are wide-ranging; a decrease in the production of mRNA splicing factors is indicative of a general loss of specificity or 'fine tuning'– the alternative forms of genes will be less common in the older individuals. This may mediate some of the decreased 'fitness' and declining ability of cells to respond to stresses with advancing age.

Using only 6 of the mRNA probes studied, it was possible to very efficiently distinguish between old and young individuals in a independent confirmatory subset of study respondents (for methods and detailed results see the paper [32]), suggesting that gene expression biomarkers of human biological aging may soon be validated.

3.1 Gene expression and age-related traits

Age-related pathologies themselves can also reveal mechanisms intrinsic to the aging process. Loss of muscle strength is a characteristic feature of aging, and can lead to sarcopenia and frailty. We found that the strongest gene expression association with muscle strength in our human subjects pinpointed a highly plausible mechanism. After adjustment for confounding factors, a transcript of CCAAT/enhancer-binding protein-beta (*CEBPB*) showed the strongest association with human strength, across a wide age range [63]. *CEBPB* is a gene involved in the maintenance and repair of damaged muscle fibers by macrophages. A mouse model of disrupted *CEBPB* signaling produced myocyte loss and fibrosis, reminiscent of sarcopenia [64]. Finding, for the first time, that *CEBPB* expression is so closely associated with human muscle strength suggests that the macrophage mediated repair pathways should be a priority for investigations of age related declines in muscle, particularly in light of recent findings that strength is predictive of mobility decline in older adults [65].

We have also examined expression associations with cognitive function, using the mini mental state examination (MMSE) score [66]. It has been argued that chronic inflammation contributes to impaired cognitive function and dementia [67,68], so we expected to find many related expression correlates. In fact, we found that the strongest association was with a single chemokine receptor (*CCR2*), with little other robust evidence of gene expression deregulation (Fig.3). *CCR2* is involved in the macrophage response to atherosclerotic plaque formation and in brain microglia. Blocking *CCR2* signaling in mice models of Alzheimer's disease resulted in accelerated Alzheimer's-like pathologies [69]

and transplantation of 'normal' hematotpoietic *CCR2* competent monocytes restored memory capacities [70]. This is another strong indication of concordance between human in-vivo leukocyte gene expression and one of the many laboratory models of Alzheimer's disease.

While gene expression studies in humans therefore appear promising, there remain major challenges. As noted above, gene expression can be affected by confounders such as lifestyle and ethnicity[57]. RNA is an unstable molecule and specimen handling is a challenge. Perhaps the greatest challenge in studying human aging is the limited access possible to many tissues of interest. Nevertheless, expression arrays using RNA from available tissues such as blood are helping to identify novel age related expression signatures and have pinpointed concordance with specific laboratory models of key traits.

4 Could genomics be part of ARD risk or diagnostic testing?

In the excitement of the human genome sequencing project, many argued that genomics would predict common diseases and personalize treatment. However, this optimism waned as it became clear that most common diseases are influenced by very large numbers of variants of small effect (plus strong environmental influences), and that much variation is still unaccounted for [71].

One of the first SNPs to be identified for type 2 diabetes was the *TCF7L2* variant, which was associated with a significant increase in risk (OR = 1.56, 95% CIs = 1.41 to 1.73, p = 4.7×10^{-18}) [72]. Could this marker be useful for disease prediction? Rates of type 2

diabetes increase with age, and studying older populations provides the best indication of how effective the prediction might be. In fact, we found that most (>80%) of those carrying the 'risk' *TCF7L2* alleles in the InCHIANTI population study of aging had not developed diabetes, and many of those with diabetes did not have the risk variant [73]. If the larger effect variants are not predictive, could panels of variants be useful? Potentially they could: a set of lipid altering variants do predict whether older people are over treatment thresholds for HDL and LDL cholesterol [74], although the prediction is never likely to be precise. Adding gene variant panels into screening program algorithms may also be of some utility, for example in breast cancer screening [75].

The key to a useful test is, of course, having an effective intervention to offer if the test is 'positive'. The success in identifying a good proportion of the heritability in age-related macular degeneration is likely to be translated into a useful predictive test when a preventive treatment becomes available. Each clinical application however needs to be properly investigated using a systematic approach to test evaluation, as initial data in preselected case and control groups tends to seriously exaggerate test efficiency [76].

In addition to risk prediction, there is some optimism about the potential for ARD variants to identify new drug targets: for example, six type 2 diabetes variants are in or near targets for existing drugs, and there is therefore optimism that novel SNP loci might lead to new drug targets [77].

In contrast to SNP based tests, gene expression markers are already being used to guide disease sub-typing and treatment selection, notably in breast [78] and other cancers. The tests involved use mainly tumor tissue itself as the source of mRNA, but extension

of this approach to identifying gene expression 'signatures' early in disease processes seems likely. In the case of age related changes we have shown that a panel of 6 markers related to immunity and inflammation, maintenance/development of muscle tissue and vascularization are highly predictive of age in a confirmatory subset [32], but such expression gene signatures of aging will need extensive and independent validation.

5 Conclusions

Aging is a heterogeneous process that includes rapidly rising risks of age-related diseases (ARDs). Delaying onset of common ARDs (including cardiovascular disease, Alzheimer's disease, prostate cancer and type 2 diabetes) is a major element in 'successful' aging. GWA studies have identified many risk genetic variants for ARDs, but a lot of variability remains to be identified. Despite this, risk SNPs are shedding new light on the underlying biological pathways involved in ARDs, with notable new insights in age related macular degeneration, Alzheimer's disease and several age related cancers.

GWAS studies have helped to change our ideas about the genetic architecture of ARDs. The majority of ARD risk variants are not in coding regions and few influence the structure of proteins themselves. Instead most variants appear to affect gene expression and the processing of RNA, often in cell or tissue type specific ways. Emerging evidence supports evolution of regulatory sites as a key driver of interspecies differences. Such evolution of gene regulation allows complex specialization of biological functions while conserving the basic structure of fundamental genes across species. More caution may

therefore be needed in extrapolating from laboratory models of aging involving disruption of apparently conserved metabolic pathway genes. We suggest therefore that animal models of aging may need to focus more on regulatory processes, and that human studies of gene regulation and expression are needed to test the relevance of each lab model finding to humans.

In larger scale gene expression studies, we have shown that RNA processing and alternative splicing, the 'fine tuning' of protein sequences, may be a potentially important factor in loss of cellular function with advancing age in humans. Analyses of expression changes in age-related traits including muscle strength and cognitive function have also pinpointed concordance with specific animal model findings. Such genome wide studies are likely to help focus research in humans on the most relevant mechanisms from laboratory models, although access to a wider range of relevant tissues is a major barrier.

The time taken for new scientific technologies to produce new tests and treatments is often initially underestimated, and this certainly applies to the genomics of ARDs. Nevertheless, the synergy of '-omics' technologies across genetic, epigenetic and transcript variations is already leading to better understanding of the fundamental processes underlying human aging and related disease. New modes of prevention and treatment of age-related diseases are likely to follow in the coming decades, although much remains uncertain.

Funding

Work on this review was supported by an unrestricted research grant from the Dunhill Medical Trust (Drs Melzer and Llewellyn) and by internal University of Exeter funding. Dr Ferrucci is supported by the Intramural Research Program, National Institute on Aging, U.S. National Institutes of Health.

Acknowledgments

We acknowledge the great contribution of the InCHIANTI study investigators and of Dr Andrew Singleton, Laboratory of Neurogenetics, National Institute on Aging to our program on genomics of aging. None of the authors declared a conflict of interest.

References

1. Melzer D, Hurst AJ, Frayling T. Genetic variation and human aging: progress and prospects. *The journals of gerontology. Series A, Biological sciences and medical sciences*. 2007;62(3):301-7.

2. Skytthe A, Pedersen NL, Kaprio J, et al. Longevity studies in GenomEUtwin. *Twin research: the official journal of the International Society for Twin Studies*. 2003;6(5):448-54.

3. Depp CA, Jeste DV. Definitions and predictors of successful aging: a comprehensive review of larger quantitative studies. *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry*. 2006;14(1):6-20.

4. Wienke A, Herskind AM, Christensen K, Skytthe A, Yashin AI. The heritability of CHD mortality in danish twins after controlling for smoking and BMI. *Twin research and human genetics: the official journal of the International Society for Twin Studies*. 2005;8(1):53-9.

5. Carmelli D, Kelly-Hayes M, Wolf PA, et al. The contribution of genetic influences to measures of lower-extremity function in older male twins. *The journals of gerontology. Series A, Biological sciences and medical sciences*. 2000;55(1):B49-53.

6. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001;29(1):308-11.

7. Koboldt DC. The Current State of dbSNP. 2012. Available at: http://www.massgenomics.org/2012/01/the-current-state-of-dbsnp.html.

8. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science*. 2002;296(5576):2225-2229.

9. Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. *In silico biology*. 2004;4(4):387-93.

10. Carroll SB. Evolution at two levels: on genes and form. *PLoS biology*. 2005;3(7):e245.

11. Wray GA. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*. 2007;8(3):206-16.

12. Fraser HB, Babak T, Tsang J, et al. Systematic detection of polygenic cis-regulatory evolution. *PLoS genetics*. 2011;7(3):e1002023.

13. Fraser HB. Genome-wide approaches to the study of adaptive gene expression evolution: systematic studies of evolutionary adaptations involving gene expression will

allow many fundamental questions in evolutionary biology to be addressed. *BioEssays: news and reviews in molecular, cellular and developmental biology*. 2011;33(6):469-77.

14. Hamilton JP. Epigenetics: principles and practice. *Digestive diseases (Basel, Switzerland)*. 2011;29(2):130-5.

15. Sinsheimer JS, Bocklandt S, Lin W, et al. Epigenetic Predictor of Age. *PLoS ONE*. 2011;6(6):1-6.

16. Le Couteur DG, McLachlan AJ, Quinn RJ, Simpson SJ, de Cabo R. Aging biology and novel targets for drug discovery. *The journals of gerontology. Series A, Biological sciences and medical sciences*. 2012;67(2):168-74.

17. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. *PLoS medicine*. 2007;4(4):e168.

18. Zeggini E, Morris A. Interpreting Association Signals. In: *Analysis of Complex Disease Association Studies*.; 2010:261-275.

19. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation. Cardiovascular genetics*. 2009;2(1):73-80.

20. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*. 2005;308(5720):385-9.

21. Yu Y, Bhangale TR, Fagerness J, et al. Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration. *Human molecular genetics*. 2011;20(18):3699-709.

22. Thakkinstian A, McKay GJ, McEvoy M, et al. Systematic review and meta-analysis of the association between complement component 3 and age-related macular degeneration: a HuGE review and meta-analysis. *American journal of epidemiology*. 2011;173(12):1365-79.

23. Sobrin L, Maller JB, Neale BM, et al. Genetic profile for five common variants associated with age-related macular degeneration in densely affected families: a novel analytic approach. *European journal of human genetics: EJHG*. 2010;18(4):496-501.

24. Stolk L, Perry JRB, Chasman DI, et al. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nature genetics*. 2012.

25. Hindorff L, MacArthur J, Wise A, et al. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed January 11, 2012.

26. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.

27. Posner SF, Pedersen NL, Gatz M. Application of life table analysis to the onset of dementia in a genetically informative design. *American journal of medical genetics*. 1999;88(2):207-10.

28. Gatz M, Pedersen NL, Berg S, et al. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *The journals of gerontology. Series A, Biological sciences and medical sciences*. 1997;52(2):M117-25.

29. Cox N, Nicolae D, Dolan ME, Gamazon E. SCAN: SNP and CNV Annotation Database. Available at: http://www.scandb.org. Accessed January 17, 2012.

30. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*. 2005;21(16):3448-9.

31. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*. 2000;25(1):25-9.

32. Harries LW, Hernandez D, Henley W, et al. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging cell*. 2011;(May):1-13.

33. Cluett C, McDermott MM, Guralnik J, et al. The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people. *Circulation. Cardiovascular genetics*. 2009;2(4):347-53.

34. Dutta A, Henley W, Lang IA, et al. The Coronary Artery Disease Associated 9p21 Variant and Later Life 20 Year Survival to Cohort Extinction. *Circulation Cardiovascular genetics*. 2011.

35. Tuppo EE, Arias HR. The role of inflammation in Alzheimer's disease. *The international journal of biochemistry & cell biology*. 2005;37(2):289-305.

36. Mahley RW, Weisgraber KH, Huang Y. Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103(15):5644-51.

37. Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*. 2000;26(1):76-80.

38. Gloyn AL, Weedon MN, Owen KR, et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*. 2003;52(2):568-572.

39. Bird CP, Stranger BE, Dermitzakis ET. Functional variation and evolution of non-coding DNA. *Current opinion in genetics development*. 2006;16(6):559-64.

40. Gray SG, De Meyts P. Role of histone and transcription factor acetylation in diabetes pathogenesis. *Diabetesmetabolism research and reviews*. 2005;21(5):416-433.

41. Hales CN, Barker DJ. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia*. 1984;35(1):595-601.

42. Ord T, Ord D, Kõivomägi M, Juhkam K, Ord T. Human TRB3 is upregulated in stressed cells by the induction of translationally efficient mRNA containing a truncated 5'-UTR. *Gene*. 2009;444(1-2):24-32.

43. Ishii H, Sakuma Y. Complex organization of the 5'-untranslated region of the mouse estrogen receptor α gene: identification of numerous mRNA transcripts with distinct 5'-ends. *The Journal of steroid biochemistry and molecular biology*. 2011;125(3-5):211-8.

44. Welham SJ, Clark AJ, Salter AM. A novel liver specific isoform of the rat LAR transcript is expressed as a truncated isoform encoded from a 5'UTR located within intron 11. *BMC Molecular Biology*. 2009;10:30.

45. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*. 2002;3(4):285-298.

46. Fackenthal JD, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. *Disease models mechanisms*. 2008;1(1):37-42.

47. Pajares MJ, Ezponda T, Catena R, et al. Alternative splicing: an emerging topic in molecular and clinical oncology. *The lancet oncology*. 2007;8(4):349-357.

48. Werner A, Carlile M, Swan D. What do natural antisense transcripts regulate? *Rna Biology*. 2009;6(1):43-48.

49. Eiberg H, Troelsen J, Nielsen M, et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics*. 2008;123(2):177-87.

50. Hou Y, Lin S. Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes Redfield RJ, ed. *PLoS ONE*. 2009;4(9):8.

51. Antebi A. Genetics of aging in Caenorhabditis elegans. *PLoS Genetics*. 2007;3(9):1565-1571.

52. Yuan R, Peters LL, Paigen B. Mice as a mammalian model for research on the genetics of aging. *ILAR journal National Research Council Institute of Laboratory Animal Resources*. 2011;52(1):4-15.

53. Robine J-M, Allard M. Validation of Exceptional Longevity - Jeanne Calment: Validation of the Duration of Her Life. 2003. Available at: http://www.demogr.mpg.de/books/odense/6/09.htm. Accessed January 26, 2012.

54. Polavarapu N, Arora G, Mittal VK, McDonald JF. Characterization and potential functional significance of human-chimpanzee large INDEL variation. *Mobile DNA*. 2011;2(1):13.

55. Jenkins NL, McColl G, Lithgow GJ. Fitness cost of extended lifespan in Caenorhabditis elegans. *Proceedings of the Royal Society B Biological Sciences*. 2004;271(1556):2523-2526.

56. Song M-Y, Kim H-E, Kim S, Choi I-H, Lee J-K. SNP-based large-scale identification of allele-specific gene expression in human B cells. *Gene*. 2011.

57. Storey JD, Madeoy J, Strout JL, et al. Gene-expression variation within and among human populations. *American journal of human genetics*. 2007;80(3):502-9.

58. de Magalhaes JP. GenAge: Database of Ageing Genes. Available at: http://genomics.senescence.info/.

59. Zahn JM, Sonu R, Vogel H, et al. Transcriptional Profiling of Aging in Human Muscle Reveals a Common Aging Signature. *PLoS Genetics*. 2006;2(7):e115.

60. Hong M-G, Myers AJ, Magnusson PKE, Prince JA. Transcriptome-wide assessment of human brain and lymphocyte senescence. *PloS one*. 2008;3(8):e3024.

61. Ferrucci L, Bandinelli S, Benvenuti E, et al. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *Journal of the American Geriatrics Society*. 2000;48(12):1618-25.

62. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(43):15545-50.

63. Harries LW, Pilling LC, Hernandez LDG, et al. CCAAT-Enhancer-Binding Protein-Beta (CEBPB) Expression In-Vivo is Associated with Muscle Strength. *Aging cell*. 2012.

64. Ruffell D, Mourkioti F, Gambardella A, et al. A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(41):17475-80.

65. Hicks GE, Shardell M, Alley DE, et al. Absolute Strength and Loss of Strength as Predictors of Mobility Decline in Older Adults: The InCHIANTI Study. *The journals of gerontology. Series A, Biological sciences and medical sciences*. 2012;67(1):66-73.

66. Harries LW, Bradley-Smith RM, Llewellyn DJ, et al. Leukocyte CCR2 Expression is Associated with Mini-Mental-State-Examination (MMSE) Score in Older Adults. *Revujenation Research*. 2012.

67. Gorelick PB. Role of inflammation in cognitive impairment: results of observational epidemiological studies and clinical trials. *Annals Of The New York Academy Of Sciences*. 2010;1207:155-162.

68. Philipson O, Lord A, Gumucio A, et al. Animal models of amyloid-beta-related pathologies in Alzheimer's disease. *FEBS Journal*. 2010;277(6):1389-1409.

69. El Khoury J, Toft M, Hickman SE, et al. Ccr2 deficiency impairs microglial accumulation and accelerates progression of Alzheimer-like disease. *Nature medicine*. 2007;13(4):432-8.

70. Naert G, Rivest S. Hematopoietic CC-chemokine receptor 2-(CCR2) competent cells are protective for the cognitive impairments and amyloid pathology in a transgenic mouse model of Alzheimer's disease. *Molecular medicine (Cambridge, Mass.)*. 2011.

71. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753.

72. Grant SFA, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics*. 2006;38(3):320-3.

73. Melzer D, Murray A, Hurst AJ, et al. Effects of the diabetes linked TCF7L2 polymorphism in a representative older population. *BMC medicine*. 2006;4:34.

74. Murray A, Cluett C, Bandinelli S, et al. Common lipid-altering gene variants are associated with therapeutic intervention thresholds of lipid levels in older people. *European heart journal*. 2009;30(14):1711-9.

75. Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. *The New England journal of medicine*. 2010;362(11):986-93.

76. Melzer D, Hogarth S, Liddell K, et al. Genetic tests for common diseases: new insights, old concerns. *BMJ (Clinical research ed.)*. 2008;336(7644):590-3.

77. Collins FS. Reengineering translational science: the time is right. *Science translational medicine*. 2011;3(90):90cm17.

78. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet*. 2011;378(9805):1812-23.

Figures

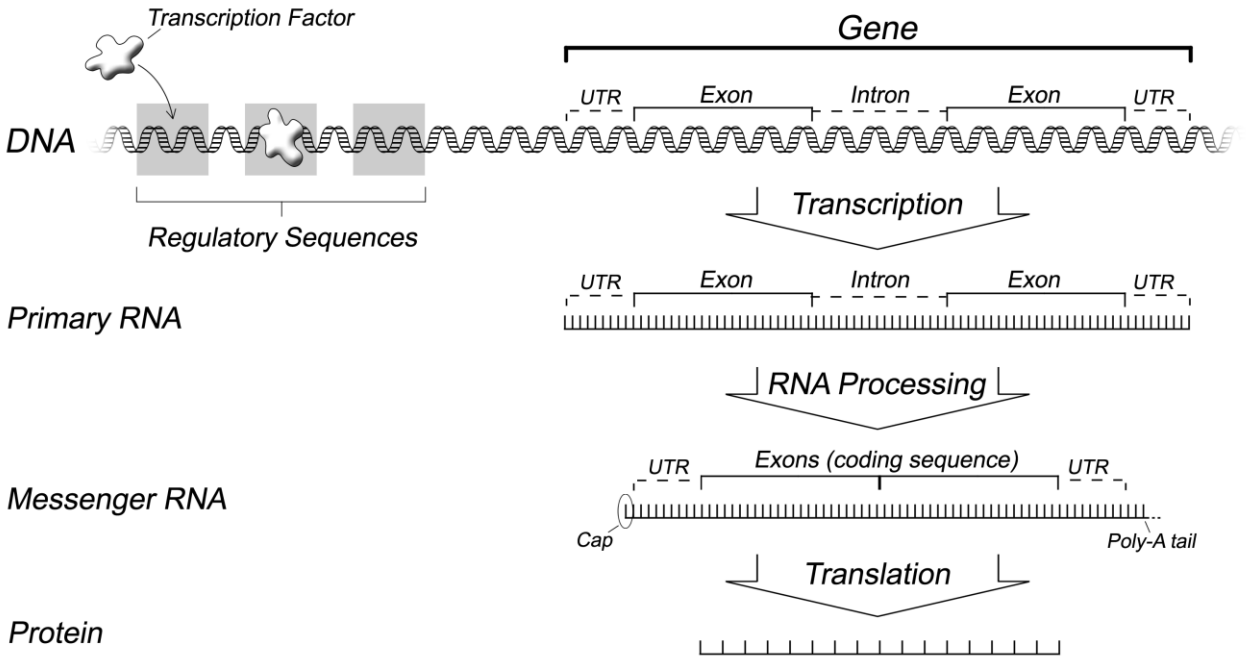

Figure.1 The central dogma – extensive regulation provides specificity

Gene expression in higher organisms requires regulation that is often specific to cell-type and is responsive to the cells environment. Here we show an example of a gene being transcribed, the RNA being processed and spliced, and then finally translated into a sequence of amino-acids: a protein. Changes to the DNA sequence within the regulatory sites could result in transcription factors failing to bind, or an intron to be incorrectly processed, etc, all of which will affect the downstream biological pathway.

Abbreviations: UTR = untranslated region, Poly-A tail = RNA ‘tail’ sequence consisting of many adenosine nucleotides.

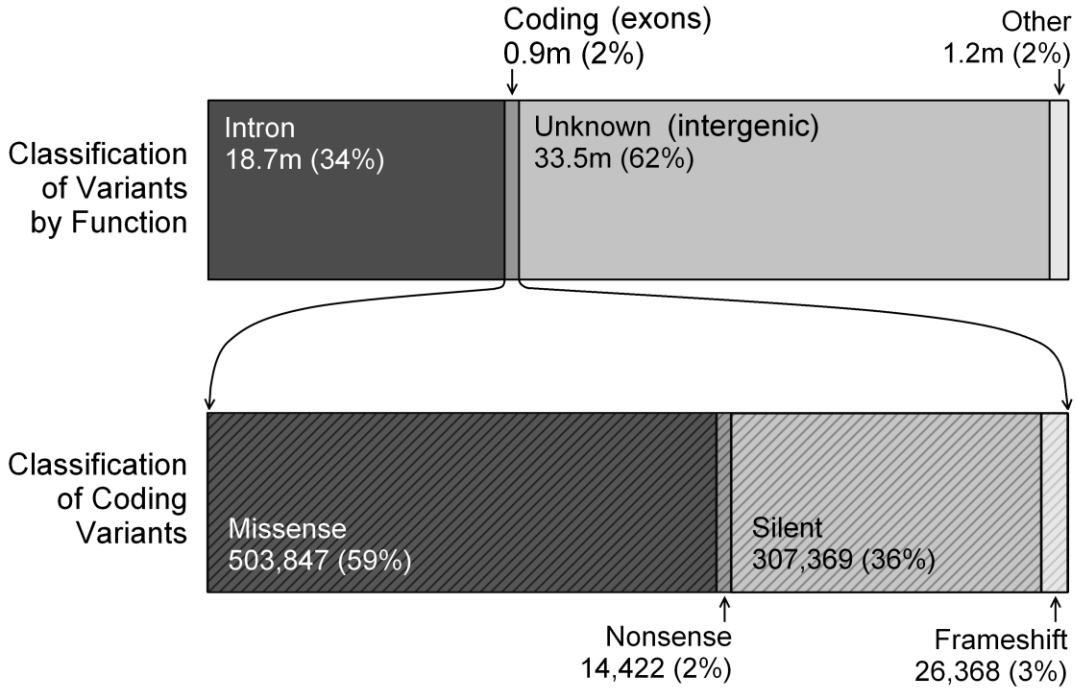

Figure.2 Classification of variants in dbSNP

Types of DNA sequence variation (from dbSNP, build 135[6], summary estimates by Dan Koboldt [7]. Only 2% of all variants are in protein coding regions, the rest are either unknown (intergenic - not located near a protein-coding gene), intronic (within genes, but non-coding) or 'other', which include those in the UTR (un-translated region) and 'near gene'. The protein-coding variants have been further classified by final protein function; they can be missense (causes an amino-acid substitution), nonsense (causes premature termination), silent (no change to amino-acid sequence) or frameshift (changes the translation 'frame') mutations.

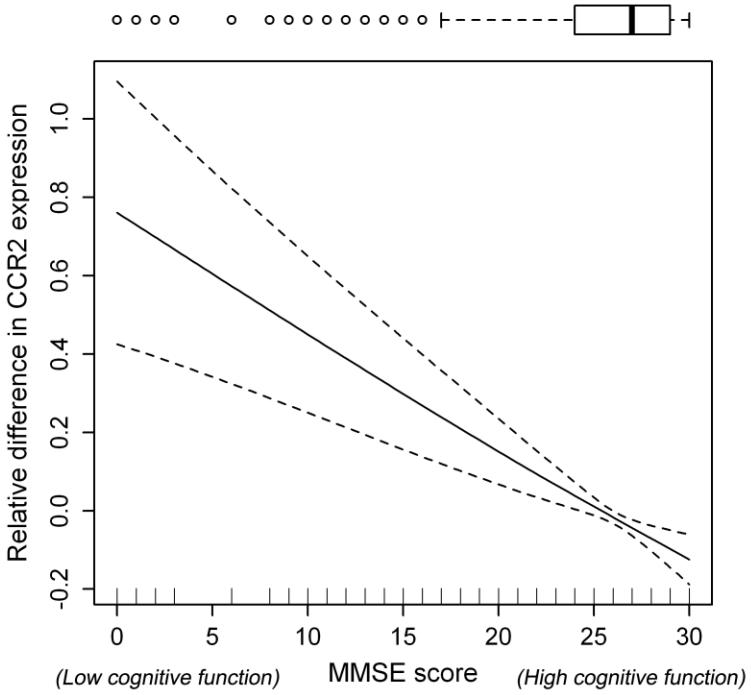

Figure.3 The relationship between CCR2 gene expression and MMSE score

This plot shows the fitted association (adjusted for confounding factors) between standard deviation differences in CCR2 gene expression levels by MMSE scores in the InCHIANTI population, from a penalized cubic spline regression model.

Table 1. Biological pathways in age-related disease: biological processes statistically overrepresented in the GWAS results for 4 common age-related diseases (ARDs): Alzheimer's disease (AD), cardio-vascular disease (CVD), prostate cancer (PC) and type 2 diabetes mellitus (T2D).

Age-related disease	Studies	SNPs	Genes*	Biological Pathways	P-value
Alzheimer's disease	14	28	11	Lipid transport	7.30E-03
				Regulation of receptor-mediated endocytosis	7.30E-03
				Positive regulation of immune response	7.30E-03
				B cell mediated immunity	9.90E-03
				Positive regulation of response to stimulus	9.90E-03
Cardio-vascular disease	10	75	42	Lipoprotein catabolic process	2.30E-02
				Sterol transport	2.30E-02
				Low-density lipoprotein particle clearance	2.30E-02
				Lipid digestion	2.70E-02
				Intestinal absorption	4.60E-02
Prostate cancer	12	55	16	Pancreas development	4.60E-03
				Epithelial cell proliferation	2.80E-02
				Regulation of embryonic development	2.80E-02
				Endocrine system development	4.10E-02
				Kidney development	4.60E-02
Type 2 diabetes	22	67	25	Regulation of insulin secretion	2.40E-04
				Cell cycle arrest	1.90E-02
				Stem cell development	2.00E-02
				RNA modification	3.20E-02
				Glucose homeostasis	3.20E-02

Notes: ARD = Age-related disease. * Number of mapped genes (not all SNPs map to genes – multiple SNPs can map to a single gene). P-values represent the strength of the association between the list of derived genes (from the SNPs list) and the biological process.
Source: gemone-catagloue (www.genome.gov/gwastudies)