The Medical Genome Reference Bank: Whole genomes and phenotype of 2,570 healthy elderly

Mark Pinese[1,2,21], Paul Lacaze[3,21], Emma M. Rath[1], Andrew Stone[1,2,24], Marie-Jo Brion[1], Adam Ameur[3-5], Sini Nagpal[23], Clare Puttick[1], Shane Husson[1], Dmitry Degrave[1], Tina Navin Cristina[6], Vivian F. Silva Kahl[22], Aaron L. Statham[1], Robyn L. Woods[3], John J. McNeil[3], Moeen Riaz[3], Margo Barr[7], Mark R. Nelson[3,8], Christopher M. Reid[3,9], Anne M. Murray[10,11], Raj C. Shah[12], Rory Wolfe[3], Joshua R. Atkins[13,14], Chantel Fitzsimmons[13,14], Heath M. Cairns[13,14], Melissa J. Green[15,16], Vaughan J. Carr[15–17], Mark J. Cowley[1,2,24], Hilda A. Pickett[22], Paul A. James[18,20], Joseph E. Powell[1,2,19], Warren Kaplan[1,2], Greg Gibson[23], Ulf Gyllensten[4,5], Murray J. Cairns[13,14], Martin McNamara[6], Marcel E. Dinger[1,2,21], and David M. Thomas[1,2,21,*]

[1] Garvan Institute of Medical Research, Sydney, Australia
[2] St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, Australia
[3] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
[4] Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
[5] National Genomics Infrastructure, Science for Life Laboratory, Sweden
[6] Sax Institute, Sydney, Australia
[7] Centre for Primary Health Care and Equity, University of New South Wales, Sydney, Australia
[8] Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia
[9] School of Public Health, Curtin University, Perth, Australia
[10] Berman Center for Outcomes and Clinical Research, Hennepin Healthcare Research Institute, Hennepin Healthcare, Minneapolis, USA
[11] Division of Geriatrics, Department of Medicine, Hennepin County Medical Center and University of Minnesota, Minneapolis, USA
[12] Department of Family Medicine and Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, USA
[13] School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, Australia
[14] Centre for Brain and Mental Health Research, Hunter Medical Research Institute, Newcastle, Australia
[15] School of Psychiatry, University of New South Wales, Sydney, Australia
[16] Neuroscience Research Australia, Sydney, Australia
[17] Department of Psychiatry, School of Clinical Sciences, Monash University, Melbourne, Australia
[18] Parkville Familial Cancer Centre, Peter MacCallum Cancer Centre, Melbourne, Australia
[19] Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
[20] Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia

[21] These authors contributed equally.
[22] Children's Medical Research Institute, Faculty of Medicine and Health, University of Sydney, Westmead, Australia
[23] Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, USA
[24] Children's Cancer Institute, University of New South Wales, Sydney, Australia
[*] Lead contact. Correspondence: d.thomas@garvan.org.au

Summary

Population health research is increasingly focused on the genetic determinants of healthy ageing, but there is no public resource of whole genome sequences and phenotype data from healthy elderly individuals. Here we describe the Medical Genome Reference Bank (MGRB), comprising whole genome sequence and phenotype of 2,570 elderly Australians depleted for cancer, cardiovascular disease, and dementia. We analysed the MGRB for single-nucleotide, indel and structural variation in the nuclear and mitochondrial genomes. Individuals in the MGRB had fewer disease-associated common and rare germline variants, relative to both cancer cases and the gnomAD and UK BioBank cohorts, consistent with risk depletion. Pervasive age-related somatic changes were correlated with grip strength in men, suggesting blood-derived whole genomes may also provide a biologic measure of age-related functional deterioration. The MGRB provides a broadly applicable reference cohort for clinical genetics and genomic association studies, and for understanding the genetics of healthy ageing. This research has been conducted using the UK Biobank Resource under Application Number 17984.

Introduction

Most developed nations face crises in health care associated with population ageing (Prince et al., 2015). Healthy ageing is a complex phenotype, influenced by both environmental and genetic factors (Lowsky et al., 2014). Healthy ageing – the absence of clinically significant, non-communicable disease or morbidity in old age – is distinct from longevity, which disregards quality of life. Healthy ageing captures the critical distinction between a long life with minimal impairment, and one bearing significant, costly, and potentially prolonged morbidity (Brooks-Wilson, 2013).

Relatively little is known about the genetic determinants of ageing that account for the broad spectrum of health states observed in older people. Genetic variation contributes to healthy ageing through pleiotropic effects on many diseases, immune responses, anthropomorphic, and behavioural phenotypes (Tobacco and Genetics Consortium, 2010; Wray et al., 2018). For example, alleles associated with behavioural phenotypes that contribute to a healthy lifestyle, such as avoidance of smoking, or propensity for regular exercise, might be anticipated to have an effect on healthy ageing. However, to date, Genome Wide Association Studies (GWAS) of common variants have only consistently associated the *APOE* and *FOXO3A* loci with lifespan and longevity (Brooks-Wilson, 2013), with *TERT* also implicated in ageing rate (Lu et al., 2018). Rare pathogenic variants that hasten the onset of common diseases associated with age, such as cancer, cardiovascular disease, and

neurodegenerative disorders, might also be expected to be depleted in healthy aged individuals. While the single study using whole genome sequencing (WGS) in 511 healthy aged individuals confirmed the link with the *APOE* locus, and suggested depletion of polygenic risk for Alzheimer's disease and coronary artery disease (Erikson et al., 2016), no evidence was found for depletion of rare pathogenic variation. Limited by sample size, these studies have focussed on single nucleotide variants and indels, while large scale structural variation remains unexplored. In addition, somatic variation such as clonal haematopoiesis (CH) is known to correlate with both age and susceptibility to disease (Genovese et al., 2014; Jaiswal et al., 2014). A synthesis of all forms of somatic and germline genomic variation is needed to inform our understanding of healthy ageing and disease susceptibility.

The advent of WGS is driving intense interest in mapping the genetic basis of disease, less than 50% of which is currently understood (Visscher et al., 2017). The missing heritability arguably resides in the total burden of both common and rare variation, structural variation untagged by simple polymorphisms, and their interactions (Zuk et al., 2012, 2014). WGS enables more comprehensive characterization of common, rare, and complex variation in human cohorts. The next few years will see the release of large-scale WGS studies in rare diseases and cancer, such as the 100,000 Genomes Project, and population studies like the UK Biobank. Maximising the analytic power of whole genome association studies using these cohorts will require well-phenotyped and high-quality control data. The concept of extreme phenotype sampling maximizes statistical power by comparing the extremes of phenotypes of interest (Barnett et al., 2013; Li et al., 2011; Zhou et al., 2016). We postulate that an elderly cohort depleted of the major common diseases constitutes a powerful and broadly applicable tool for genome-wide association studies of disease.

With this background, we undertook WGS of 2,570 elderly individuals with no reported history of cancer, cardiovascular disease, or neurodegenerative diseases up to age 70, to create the Medical Genome Reference Bank (MGRB) (Lacaze et al., 2018). For comparison, we have also undertaken whole genome sequencing of 344 young subjects, and 273 elderly individuals with cancer. These cohorts have been subjected to a broad spectrum, systematic analysis of germline and somatic variation within the nuclear and mitochondrial genomes, which we have linked to both chronologic age as well as frailty measures.

Results

Cohort characteristics and sequencing

We identified 2,926 individuals from the ASPREE study (McNeil et al., 2017), and Sax Institute's 45 and Up Study (45 and Up Study Collaborators et al., 2008), who lived to at least 70 years of age without any history of cancer, cardiovascular disease, or dementia, confirmed either at baseline entry or study follow-ups (Lacaze et al., 2018). We sequenced all samples by WGS, mapping to build 37 of the human reference genome, and calling variants following GATK best practices. After exclusion of 356 samples that failed quality control and relatedness checks, 2,570 samples remained, forming the MGRB cohort (Table 1).

Table 1: Summary metrics for the MGRB well elderly cohort, sourced from the ASPREE or 45 and Up studies. Aggregate statistics are medians, with ranges in parentheses. Genetic background (ancestry) was determined from genotype data. Although blood was occasionally drawn at younger than 70 years, all individuals lived to at least 70 years without known cancer, cardiovascular disease, or dementia.

Measure	*ASPREE*	*45 and Up*
Individuals	1,853	717
(percent female)	(48.2%)	(59.3%)
Age at blood draw (years)	79	70
	(75 – 95)	(64 – 91)
Height (m)	1.65	1.66
	(1.33 – 1.91)	(1.37 – 1.91)
Mass (kg)	74.5	72.0
	(33.4 – 127.1)	(36.0 – 147.0)
Mean sequencing depth	38.0	39.0
(genome-wide)	(26.8 – 46.0)	(27.3 – 45.5)
Genetic background		
Non-Finnish European	1,805	695
South and Central American	23	5
South Asian	14	6
Finnish European	10	7
East Asian	1	4

A broad diversity of genetic variation was found in the MGRB cohort. We identified 69,996,670 small variant loci in canonical chromosome contigs, with a call rate of 99.5%. Our small variant detection sensitivity was 99.3% and false positive rate 4.84 / Mbp, as assessed by comparing an internal RM 8398 sample against a gold standard (Zook et al., 2014). MGRB participants were primarily of non-Finnish European genetic background (Table 1, Supplementary Figure 1). Consistent with previous studies (Lek et al., 2016; Telenti et al., 2016), 51.8% of small variants were singletons, and 4.6% of loci were multi-allelic.

In addition to small scale variants, an average of 4,036 structural variants (SVs) per individual were observed, most commonly deletions (Supplementary Table 1, Supplementary Figure 2). In contrast to small variants, only 17% of structural variants were unique (Supplementary Table 2). Each individual carried an average of 4,264 mobile element insertions (MEI), predominantly of the ALU and L1 classes, consistent with a previous report (Chaisson et al., 2015), and most MEIs were copy number polymorphisms at known loci. However, on average 1,535 MEI events per individual were in regions of the reference genome not currently described as containing mobile elements. In summary, while small variants comprise the majority of genetic diversity in the MGRB, structural and mobile elements constitute a rich and understudied source of potentially disease-related variation.

The well elderly carry clinically reportable genetic variation

Population genomic studies are contributing to the substantial revision of clinical interpretation of genetic variation thought to drive disease in some cases (Walsh et al., 2017). It is therefore clinically important to understand the frequency of variants currently considered pathogenic in a clinical context, but which are observed in well elderly individuals (Lacaze et al., 2017). To this end, we identified pathogenic variants that are considered clinically reportable as incidental findings under current American College of Medical Genetics (ACMG) guidelines (Kalia et al., 2016; Richards et al., 2015). Forty pathogenic or likely pathogenic heterozygous small variants were identified, with 28/2,570 (1.1%) individuals carrying dominantly acting variants linked to disease (Table 2, Supplementary Data File 1). We sought further evidence of disease phenotypes in individuals carrying relevant pathogenic variants from the ASPREE cohort. We did not identify personal histories of breast or colorectal cancer in individuals harbouring *BRCA2*, *MSH2,* or *PMS2* mutations; cardiac arrest or strokes in individuals harbouring *DSG2*, *DSP*, *KCNH2*, *KCNQ1*, *MYBPC3*, *MYL3*, and *SCN5A* mutations; or elevated blood lipid levels in *APOB* carriers. Cancer-associated genotypes are dependent on stochastic factors which may account for variable penetrance, while anaesthetic-associated malignant hyperthermia linked to loss of function variation in *RYR1* is contingent on environmental exposure. We specifically sought, but did not find, evidence of cardiovascular disease history or related clinical phenotypes in carriers of variants linked to atrial fibrillation, cardiomyopathy and hypertension. Notably, no genotypes predicted to cause severe childhood-onset diseases were identified (Chen et al., 2016); the single *RYR2* variant detected was a truncation not expected to cause autosomal dominant catecholaminergic polymorphic ventricular tachycardia. In five individuals, variants were noted in *PCSK9* that are predicted to be protective against high blood cholesterol (Langsted et al., 2016), comparable to the rate observed in the gnomAD non-Finnish European cohort (Fisher exact test, p = 0.37). Four SVs were found that may disrupt the coding sequence of genes associated with cancer and cardiovascular health (Supplementary Table 3), comprising 10% of potentially pathogenic variation in genes considered reportable by the ACMG.

Table 2: Counts of clinically significant small variation in the MGRB for all genes in the ACMG SF 2.0 set. Abbreviations: ARVC, Arrhythmogenic right ventricular cardiomyopathy; CPVT, Catecholaminergic polymorphic ventricular tachycardia; HCM, Hypertrophic cardiomyopathy; DCM, Dilated cardiomyopathy; VA, Ventricular arrhythmia.

Condition	*Gene*	*Carriers*
Cancer	*BRCA2*	4 (2 female)
	MSH2	1
	MSH6	1
	PMS2	3
Neurofibromatosis	*NF2*	1
ARVC	*DSG2*	1

	DSP	3
CPVT	*RYR2*	1
HCM, DCM	*MYBPC3*	2
	MYL3	1
	TNNI3	1
Hypercholesterolemia	*APOB*	5
Long QT, VA	*KCNH2*	1
	SCN5A	1
Marfan syndrome	*MYH11*	1
Malignant hyperthermia	*RYR1*	1
TOTAL		28

Rare and common risk variants are depleted in the well elderly

One of the primary purposes of the MGRB is to serve as a genetic risk-depleted control cohort for studies of the common causes of morbidity and mortality. To test its utility, we compared the rates of pathogenic variants in tumour suppressor genes between the 717 MGRB individuals from the 45 and Up Study, with 269 demographically-matched cancer cases from the same study (45 and Up Study; Supplementary Table 4). Considering all cancers in aggregate, the MGRB samples were significantly depleted for pathogenic alleles in tumour suppressor genes relative to cancer cases, with 2 of 717 controls carrying pathogenic tumour suppressor variants compared to 12 of 269 cancer cases (Figure 1a, odds ratio 0.060, 95% confidence interval 0.0065 to 0.27, $p < 0.001$, Fisher's exact test). In addition to all cancers, we specifically examined colorectal cancer due to its high incidence in our case set, and well-defined genetic risk (De Rosa et al., 2015). The MGRB samples were significantly depleted for rare pathogenic variation in the *APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *SMAD4* genes, relative to colorectal cancer cases (Figure 1a, 1 of 717 MGRB with pathogenic variants, versus 2 of 40 cancer cases, odds ratio 0.027, 95% CI 0.001 to 0.53, $p = 0.008$). We did not detect a difference in the rate of rare coding loss of function variants in the MGRB relative to the gnomAD non-Finnish European cohort, either genome-wide, or in genes associated with cardiovascular disease or cancer, potentially due to technical factors dominating differences in rare variant patterns between cohorts (8 tests, all $p > 0.08$).

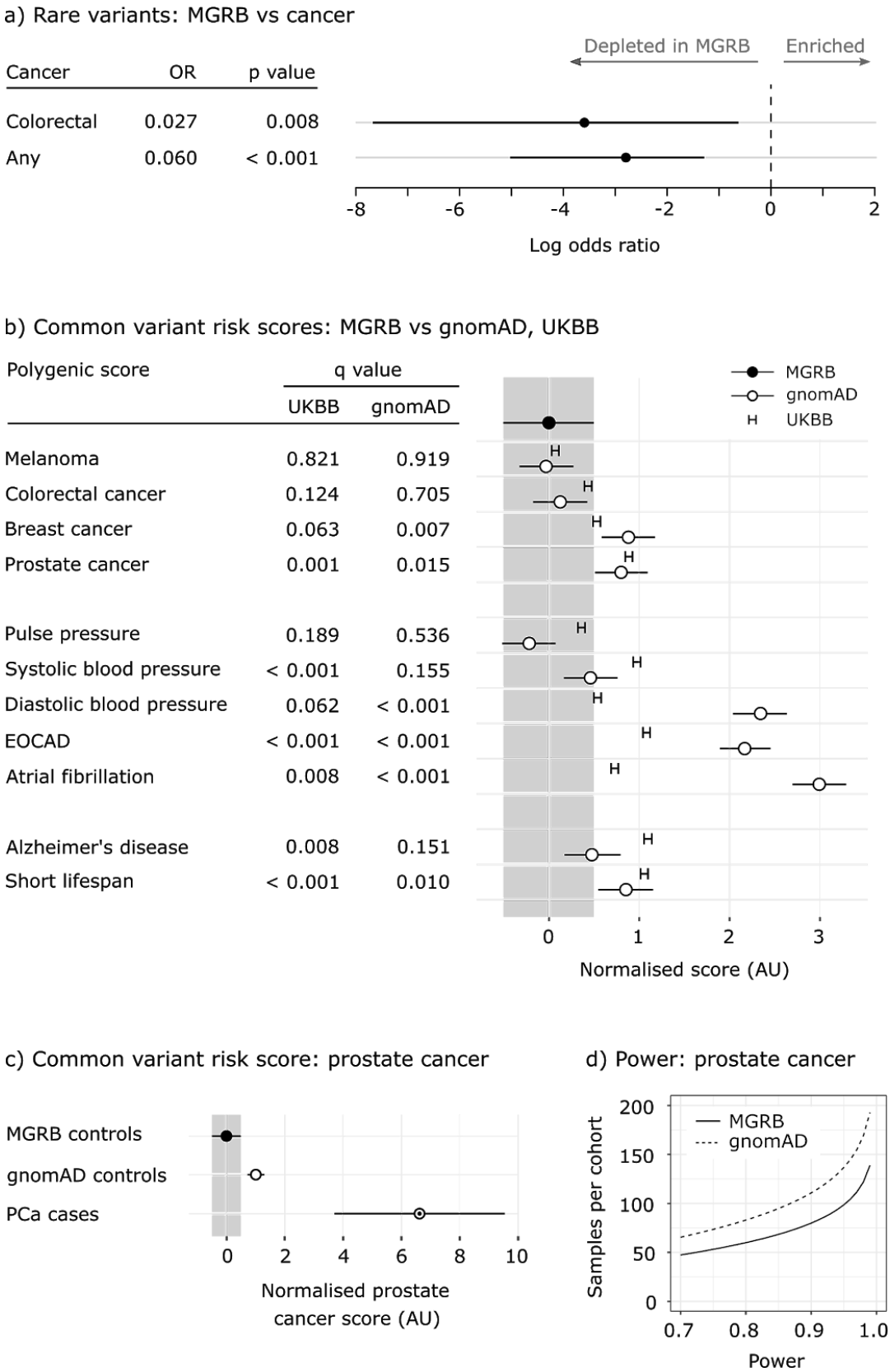


Figure 1: The MGRB is depleted for genomic risk relative to reference and disease cohorts. a) the rate of rare pathogenic variants in tumour suppressor genes is lower in MGRB than in a cohort of cancer cases (log odds for an individual to carry a pathogenic tumour suppressor variant shown). b) the MGRB also has lower polygenic score (PS) estimates for a range of phenotypes, when compared to the gnomAD non-Finnish European population and the UK BioBank samples. MGRB is the

reference in (b), with PS mean set at zero; bootstrap 95% confidence intervals are shown for each PS. q-values represent false discovery rate estimates by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). c) the MGRB has lower PS compared to prostate cancer cases, here considering only samples from the 45 and Up Study. d) for any given sample size, the MGRB has greater statistical power to detect PS difference from a case cohort than to gnomAD, demonstrated here for prostate cancer. AU: arbitrary units.

We next sought evidence for depletion of common disease-associated variation in the MGRB, relative to the gnomAD and UKBB datasets. Although SNP allele frequencies were highly concordant across all three cohorts (Supplementary Figure 3), the MGRB cohort was significantly depleted for alleles specifically associated with risk of cancer, cardiovascular disease, and neurodegenerative disease (Supplementary File 2, 698 loci, odds ratios 0.38 vs gnomAD, 0.47 vs UKBB, both $p < 1.1x10^{-6}$, Fisher's exact test). This enrichment of protective alleles was specific to the clinical phenotypes excluded from MGRB, and was not observed for negative control loci linked to anthropometric (449 loci, both $p > 0.69$) or behavioural (575 loci, both $p > 0.55$) traits.

The aggregate burden of common disease-related variants within individuals can be summarised in a polygenic score (PS). We constructed polygenic predictors for a range of phenotypes measured or depleted in the MGRB, and compared PS distributions between MGRB, the gnomAD non-Finnish European reference cohort, the UK BioBank (UKBB), and the 45 and Up Study cancer cohort. Significant depletion in PS was observed in MGRB for eight of the eleven scores tested (Figure 2b). Notably, a PS associated with short lifespan (Deelen et al., 2014) was significantly depleted in MGRB relative to gnomAD and UKBB, consistent with the MGRB healthy elderly phenotype. MGRB individuals were significantly depleted for prostate cancer risk relative to both gnomAD and prostate cancer cases, indicating that MGRB is an extreme depletion cohort for prostate cancer polygenic risk (Figure 2c). Critically, for the extreme phenotype sampling hypothesis, the use of the MGRB as a control cohort reduced the sample size required to reach a given target power by approximately 25% by comparison with the widely-used gnomAD dataset (Figure 1d).

In addition to the allele frequency-based comparisons above, the availability of individual genotypes for the MGRB and 45 and Up Study cancer cohorts enabled the direct evaluation of the influence of PS on cancer risk. We first confirmed that our polygenic scoring method estimated individual height using published loci (Wood et al., 2014): height PS was significantly predictive of measured height, with a slope of 4.5 cm per polygenic score unit, complete model $R^2 = 0.62$, polygenic score partial $R^2 = 0.14$, n = 2,537 (Supplementary Figure 5). We then compared the distribution of PS for prostate, colorectal, and melanoma skin cancer between the 45 and Up cancer-free cases in the MGRB, and individuals from the 45 and Up cohort with these cancers (Supplementary Table 4). Consistent with the relative depletion of rare cancer variants in the MGRB observed above, MGRB individuals had significantly lower polygenic risk scores than cases for prostate cancer ($p < 0.001$, Figure 2a) and colorectal cancer ($p = 0.022$, Figure 2b), but not melanoma. The contribution of PS to cancer-specific risk was significant: by age 70, individuals with a cancer PS in the top 5% of MGRB had a 7.7-fold increased odds for prostate cancer, and a 3.6-fold increased odds for colorectal cancer, relative to individuals with a score in the bottom 5%.

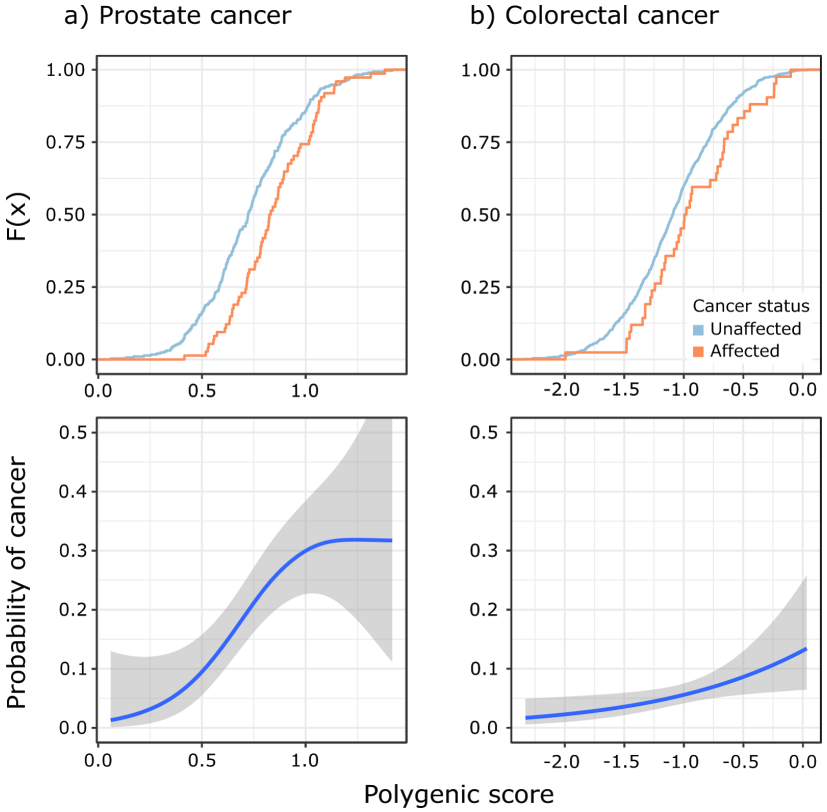


Figure 2: Polygenic risk is strongly related to cancer diagnosis risk. Cumulative distribution functions (top panels) and associated probability of cancer diagnosis by age 70 (bottom panels) are shown for both prostate cancer (a) and colorectal cancer (b). Unaffected individuals are MGRB men (prostate), or all MGRB individuals (colorectal) and were completely cancer-free up to age 70; affected individuals were sourced from the 45 and Up Study cancer cohort and had recorded evidence of the relevant cancer diagnosis prior to age 70. Polygenic scores were computed based on reported loci and model coefficients (Hoffmann et al., 2015; Schumacher et al., 2015). Fits are from logistic regression using a GCV-penalised thin plate spline smooth; bands denote 95% confidence intervals around the mean.

Clonal somatic variation is detectable by WGS

In addition to its use as a surrogate for the germline, peripheral blood DNA carries somatic variation reflecting the life history and health state of the donor. Clonal haematopoiesis of Indeterminate Potential (CHIP) occurs in at least 10% of individuals over the age of 65 years, evidenced by somatically acquired SNVs (Genovese et al., 2014; Young et al., 2016). Previous studies have used deep whole-exome or targeted sequencing to identify CHIP, which lacks sensitivity to detect the SVs commonly observed in myelodysplasia and leukemia. Low depth WGS, a powerful tool for measuring SV, has not been applied to the detection of CHIP. Here we estimate the burden of cancer-associated somatic variation in peripheral blood DNA using whole genome data in the MGRB cohort.

In total, 184/2,570 (7.2%) of MGRB individuals displayed evidence of CHIP, with SNVs associated with overgrowth and neoplasia observed in more than 10% of reads. Predominantly nonsense mutations (96%), these variants were most commonly seen in *TET2* (47 individuals), *DNMT3A* (23), or *ASXL1* (11). We also observed known

gain-of-function missense variants in *JAK2* V617F (9 individuals), *NRAS* G12D (1), a dominant negative allele in *DNMT3A,* R882H (1) (Russler-Germain et al., 2014), and a putative loss-of-function variant in *TP53*, C275Y (1). *JAK2* V617F is a recognised driver of myeloproliferative disorders (Percy and McMullin, 2005), which are also associated with *ASXL1* loss (Gelsi-Boyer et al., 2012), and *TET2* and *DNMT3A* loss-of-function variants are frequent in CHIP (Buscarlet et al., 2017). In total, the blood of 91/2,570 (3.5%) cancer-free MGRB individuals carried deleterious small variation in at least one of these four genes, and 13 individuals had multiple deleterious mutations in this gene set. We next sought evidence for subclonal copy number variation (CNV). 1975 of 2570 MGRB individuals were successfully fit to a subclonal CNV model; of these 55 (2.8%) showed evidence of subclonal CNV, as determined by the presence of an aneuploid lineage representing more than 10% of nucleated blood cells (Supplementary Figure 6). In total, 9.2% (95% CI 7.9 to 10.5%) of MGRB samples demonstrated evidence of CHIP by either SNV or CNV, consistent with results from deep WES (Jaiswal et al., 2014). In sum, subclonal blood DNA changes are detectable from WGS at routine read depths used for germline purposes, providing a quantitative fingerprint of age-related somatic events.

Age-related mitochondrial load is associated with grip strength

As well as CHIP, ageing is associated with telomere shortening (von Zglinicki and Martin-Ruiz, 2005), somatic Y chromosome loss, decreased mitochondrial copy number, and increased mitochondrial heteroplasmy (Kennedy et al., 2013; Wachsmuth et al., 2016). We therefore studied the relationship of age to telomere length, mitochondrial copy number and variation, Y copy number in males, a somatic mutation signature linked to ageing (Alexandrov et al., 2015), and CHIP. Using standard depth WGS data from multiple cohorts, consistent patterns of change with age were observed across all six somatic metrics (Figure 3a-f). Compared to a population of younger individuals (the ASRB cohort, median age 40, Supplementary Figure 4), the MGRB, despite being ascertained on the basis of healthy ageing, was still associated with shorter telomere lengths, increased somatic mutation burden, and decreased Y chromosome and mitochondrial copy number (Table 3). Interestingly, there were differences between each cohort in the relationship with age, with apparent stabilisation of telomere length in the elderly cohorts past approximately 70 years, compared to the expected progressive shortening with increasing age observed in the younger ASRB cohort. In addition, while mitochondrial copy number/nuclear genome was stable up to age 60, significant declines were observed in the older age groups. The rate of change was significantly different between the young (ASRB) and aged (MGRB) cohorts (5 likelihood ratio tests on linear fits, Holm correction, all $p < 0.003$), while the rate of change of the two aged cohorts was consistent across all measures (5 likelihood ratio tests, Holm correction, all $p > 0.28$). Taken together, these results are suggestive of altered kinetics of age-related somatic mutation in the elderly compared to younger populations, although we note longitudinal measurements will be necessary to definitively establish this relationship.

Table 3: The rates of somatic measure change with age are different between middle-aged and old individuals. Numbers show the rate of change of each somatic measure with age in the middle-aged ASRB cohort (median age 40), and the older MGRB cohorts (median age 70 or older). Changes are significantly different between the younger ASRB and older

MGRB cohorts, and consistent within the two older MGRB cohorts. Linear model slopes as change per decade are reported for each of five somatic measures in each cohort, with 95% Wald confidence intervals. Values significantly different from zero are represented in bold. Note that somatic burden and mitochondrial count per nucleus are reported on the natural logarithm scale. N.D.: Not determined due to data use agreement constraints.

	Cohort		
Measure	*ASRB*	*45 and Up Study*	*ASPREE*
Individuals	344	717	1853
Percent female	N.D.	59.3%	48.2%
Median age (range)	40 (18 - 65)	70 (64 - 91)	79 (75 - 95)
Telomere length (AU/decade)	**-0.115** [-0.157, -0.073]	0.040 [-0.010, 0.090]	**0.115** [0.035, 0.196]
Mitochondria count (log_{10} mt/nucleus/decade)	-0.004 [-0.018, 0.010]	**-0.046** [-0.065, -0.027]	**-0.038** [-0.059, -0.017]
Y copy number in males (Y chromosomes/nucleus/decade)	-0.011 [-0.022, 0.001]	**-0.050** [-0.068, -0.033]	**-0.043** [-0.065, -0.021]
Somatic variant burden (log_{10} variants/Mb/decade)	0.038 [-0.002, 0.079]	**0.207** [0.167, 0.247]	**0.228** [0.173, 0.282]
Mitochondrial variants (mt variants/decade)	0.051 [-0.177, 0.278]	**1.665** [1.315, 2.015]	**0.893** [0.195, 1.591]

We next considered whether individual age-related genomic measures may reflect physical function status, independently of chronologic age. To address this question, somatic changes in MGRB samples from the ASPREE cohort were studied in the context of age, grip strength, and gait speed, all representing key predictors of age-related morbidity (Chainani et al., 2016; Dudzińska-Griszek et al., 2017; Syddall et al., 2003). As expected, grip strength and gait speed both consistently decreased with age in both genders (Figure 3g,h). To correct for the strong influence of age on all measures, a conditional analysis was performed to explore whether any somatic measures were associated with physical function even when age is taken into account. Intriguingly, we found that grip strength was positively correlated with the count of mitochondria per nuclear genome, but only in males (two-stage test, first stage p = 0.051, validation p = 0.036).

To illustrate the magnitude of effect of mtDNA copy number on grip strength in men, we modelled the change in “effective age” as determined by grip strength, as a function of mtDNA copy number. This revealed that men with an mtDNA copy number in the lowest 5% for their age have the same grip strength as men with average mtDNA levels, but who are 2.5 years older (Figure 3i).

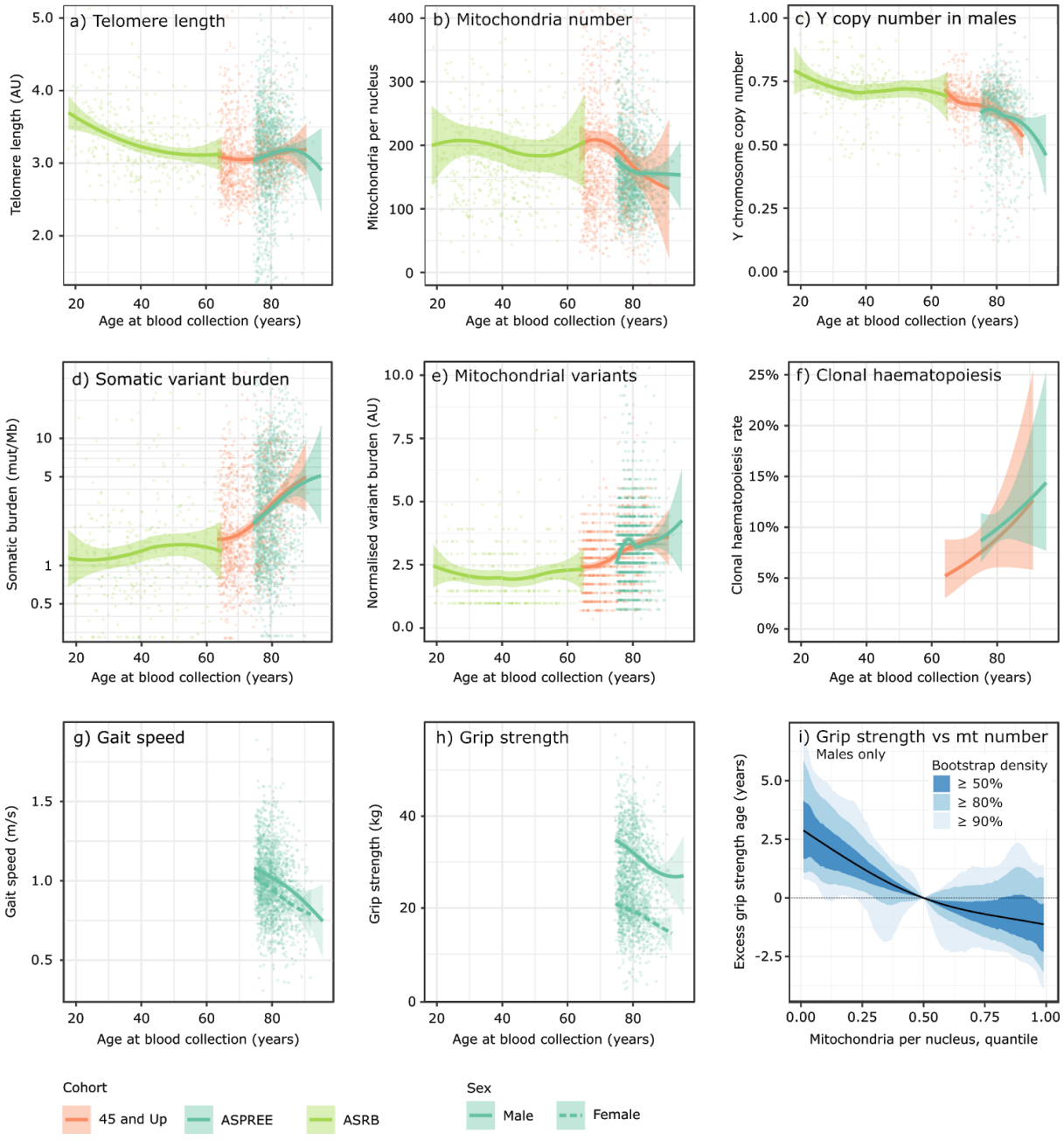


Figure 3: Age-related somatic changes detectable in blood DNA by whole genome sequencing are associated with two measures of physical function. Across multiple cohorts, a consistent decrease with age is observed for telomere length (a), mitochondria per nucleus (b), and Y copy number in males (c). In contrast, advanced age is associated with an increase in somatic mutation burden (d,e) and the fraction of samples with detectable clonal haematopoiesis (f), as well as a decrease in the key physical function measures gait speed (g) and grip strength (h). The count of mitochondria per nucleus is significantly related to grip strength beyond age alone in men, as indicated by the change in effective age as judged by grip strength with varying mitochondria count (i). For (a-c,g,h) individual measurements corrected for cohort batch effect are shown with LOESS smooths, and for (d) a logistic fit was used. Bands around estimates delimit 99% confidence intervals for the mean.

Discussion

Understanding the genetic underpinnings of healthy ageing is as important as, and relevant to, understanding the genetic basis of disease. The next decade will see the fruits of population-scale sequencing programs, much of which will be aimed at understanding the genetic origins of disease. To realise this mission, we need to understand the spectrum of genetic variability in the healthy, and whole genome data sets of healthy controls will be essential to identify genetic variation unrelated to disease (Manrai, Patel, Ioannidis, JAMA, 2018). To this end we created the MGRB, a whole-genome sequencing resource of deeply-phenotyped aged individuals (Lacaze et al., 2018).

Although depletion of some common disease-related alleles has been reported in the healthy aged (Brooks-Wilson, 2013; Deelen et al., 2014; Erikson et al., 2016), the MGRB reveals a striking depletion in disease-associated common and rare variation, relative to both affected cases, as well as datasets frequently used as controls in genetic studies, but not specifically depleted for disease phenotypes. In addition, the MGRB was enriched for protective alleles linked to healthy ageing. Our data also substantiate the premise that extreme phenotype enrichment can enhance statistical power in case:control genetic studies (Li et al., 2011) (Figure 1c,d).

Despite being healthy, over 1% of the MGRB still carry pathogenic small variants that are clinically-reportable under current ACMG guidelines (Table 2), consistent with previous observations (Erikson et al., 2016). A detailed review of individual phenotypes from a subset of mutation carriers excluded even subclinical manifestations of the expected disorders. These data suggest that many apparently pathogenic small variants have variable penetrance, echoing a theme emerging from population genomic studies. Additionally, several rare structural variants were identified that may abrogate function of clinically-reportable genes (Supplementary Table 3). Future studies using whole genome-based data will benefit from the MGRB in quantitating the contribution of structural variation to ageing and disease. These observations suggest the MGRB may provide a filter for rare variants currently thought pathogenic in a clinical context.

The ageing process is accompanied by the emergence of somatic mutations in tissues other than blood, mitochondrial depletion and heteroplasmy, and progressive telomere shortening (Acuna-Hidalgo et al., 2017; Genovese et al., 2014; Jaiswal et al., 2014; Martincorena et al., 2015). We developed a suite of methods to detect these age-related changes using 30X whole-genome sequencing, and applied it to the elderly MGRB and a younger cohort. Telomere shortening itself may directly increase the likelihood of neoplasia (Artandi et al., 2000), while oncogenic mutations in genes such as *TP53* may rescue the effects of telomere loss (Chin et al., 1999). Telomere dysfunction has been associated with impaired mitochondrial function (Sahin et al., 2011), linking these genomic features of ageing. Many of the somatic changes of ageing observed in the MGRB are associated with marrow stem cell depletion (de Haan and Van Zant, 1999), consistent with studies in telomerase-deficient mice (Lee et al., 1998).

Interestingly, we observed a shift in the age trajectory of multiple somatic metrics in the elderly compared to younger individuals, coincident with the emergence of clonal hematopoiesis. It is paradoxical that, in this and other cohorts, somatic clonal expansion driven by oncogenic mutations appears compatible with normal organ function (Genovese et al., 2014; Jaiswal et al., 2014; Martincorena et al., 2015). It is even possible that neoplastic events, such as telomere stabilisation, loss of tumor suppressor genes, or acquisition of oncogenic kinase mutations, might increase clonogenic efficiency of an ageing marrow stem cell compartment. Some support for this concept, reminiscent of antagonistic pleiotropy (Williams, 1957), comes from mice carrying a hypermorphic form of *Trp53,* in which protection from neoplasia was accompanied by accelerated hematopoietic ageing and diminished marrow reserve (Dumble et al., 2007; Tyner et al., 2002). If true, these findings suggest that strategies that suppress tumor formation may accelerate ageing.

We observed an intriguing link between somatic burden and decline in physical function, providing a potential measure of what distinguishes individuals sharing the same age, but different physical function status. The relative depletion of mitochondria per leukocyte appeared to be associated with reduced grip strength in males, after adjustment for age. This finding is consistent with evidence that mitochondrial dynamics are strongly involved in ageing and function, particularly in males (Latorre-Pellicer et al., 2016; Wachsmuth et al., 2016). We note that our power to detect such an effect is low when using a well elderly cohort, but believe there will be great interest in deriving quantitative measures of biological ageing from standard-depth whole genome sequencing.

Although the largest cohort of healthy elderly whole genomes amassed to date, the MGRB is still subject to limitations as a research and clinical tool. The investigation of extremely rare variants is limited by the MGRB's size, and complicated by batch effects in rare variant calls (Tom et al., 2017). Furthermore, the MGRB comprises almost exclusively white Australians, and follow-up studies will be required to assess the spectrum of genetic variation in the healthy elderly from other backgrounds. The MGRB was recruited on the basis of a restricted definition of healthy ageing, being depletion of cancer, cardiovascular disease, and dementia, and MGRB participants do bear other morbidities. However we note that the deep phenotype which accompanies the MGRB enables more focussed participant selection and the construction of for-purpose subset cohorts, making the MGRB of value as a universal control that can be depleted of any measured phenotype. Finally, although we observed associations between somatic measures and age that are suggestive of changes in ageing kinetics, this cannot be definitively established using our cross-sectional study design. Further studies with longitudinal samples will be required to verify our hypothesis of altered ageing kinetics, for which the methodology established here will be valuable.

Quantitative biomarkers of age may provide a summative metric of diverse genetic and environmental effects on health. Interpreted as endophenotypes, such biomarkers show promise to increase our ability to detect genetic patterns associated with ageing rate (Lu et al., 2018), but their true utility may be greater still as clinical tools in their own right. By encoding the aggregate influence of complex and potentially unmeasurable genetic and environmental effects over the life of an individual, biomarkers of age may represent health

and disease risk with greater fidelity than external indicators such as calendar age or functional state.

Particularly with respect to cancer, the DNA-based measures of biological age we have demonstrated here may represent an individual's underlying mutation rate, and therefore true cancer risk, due to combined genetic and environmental factors. This biomarker-centric perspective on cancer risk represents a synthesis and simplification of the traditional genotype- and environment-centric views, and we believe is a promising lens through which to consider disease risk, and differentiate normal compensatory changes associated with ageing, from those that precede malignancy.

Acknowledgements

Whole genome sequencing of the MGRB, ASRB, and 45 and Up cancer cohort was undertaken at the Kinghorn Centre for Clinical Genomics, which is supported by the Kinghorn Foundation. Sequencing of these cohorts was funded through the NSW Genomics Collaborative Grants scheme from the NSW Office for Health and Medical Research.

Processing and warehousing of genomic data employed resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

The authors acknowledge the ASPREE Healthy Ageing Biobank, ASPREE Investigator Group and ASPREE Collaborating Practitioners listed on www.aspree.org. ASPREE was funded by the National Institute on Aging and the National Cancer Institute at the National Institutes of Health (grant number U01AG029824); the National Health and Medical Research Council of Australia (grant numbers 334047, 1127060); the Victorian Cancer Agency and Monash University (Australia). The ASPREE Healthy Ageing Biobank was supported by the Commonwealth Scientific and Industrial Research Organisation (Australia), the Victorian Cancer Agency (Australia) and Monash University (Australia). We acknowledge the dedicated and skilled staff in Australia and the U.S. for the conduct of the ASPREE trial, and the ASPREE participants who willingly volunteered.

This research was completed using data collected through the 45 and Up Study (www.saxinstitute.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council NSW; and partners: the National Heart Foundation of Australia (NSW Division); NSW Ministry of Health; NSW Government Family & Community Services – Ageing, Carers and the Disability Council NSW; and the Australian Red Cross Blood Service. We thank the many thousands of people participating in the 45 and Up Study.

Genomic analysis of the Australian Schizophrenia Research Bank (ASRB) was supported by a New South Wales Health, Collaborative Genomics grant program (MC, MG, VC), a NARSAD Independent Investigator Grant (MC) and National Health and Medical Research Council (NHMRC) project grants (1067137, 1147644, 1051672). Samples were also

collected by the ASRB with the support of the NHMRC, the Pratt Foundation, Ramsay Health Care, and the Viertel Charitable Foundation. The ASRB was supported by the Schizophrenia Research Institute (Australia), utilizing infrastructure funding from NSW Health and the Macquarie Group Foundation.

This research has been conducted using the UK Biobank Resource under Application Number 17984.

D.M.T is the recipient of an NHMRC Principal Research Fellowship (RegKey:1104364). M.J.Cairns was supported by an NHMRC Senior Research Fellowship (#1121474). M.J.Cowley was supported by a NSW Health Early-Mid Career Fellowship. J.R.A. was supported by University of Newcastle RHD and an Emlyn and Jennie Thomas Postgraduate Medical Research Scholarship. M.E.D., C.P., W.K., D.D., and S.H. were supported by the Kinghorn Foundation.

The authors thank Prof. Christopher Goodnow, Prof. Diane Fatkin, Dr Catherine Vacher for helpful comments and discussion, Dr Eleni Giannoulatou for the atrial fibrillation polygenic score coefficients, and Verity Hodgkinson for biospecimen management.

Author Contributions

Conceptualization, M.P., M.E.D., and D.M.T.;
Methodology, M.P., E.M.R., J.E.P., C.P., and D.M.T.;
Software, M.P., E.M.R., and A.L.S;
Formal Analysis, M.P., E.M.R., A.L.S., M.R., C.P., M.J.Cowley, and P.A.J.;
Investigation, V.F.S K, and H.A.P.;
Resources, P.L., A.A., M.B., M. McN., E.B., M.R.N., C.M.R., A.M.M., R.C.S., R.W., M.J.Cairns., H.M.C, J.R.A, C.F., M.J.G., V.J.C., S.N., G.G., R.L.W., U.G., S.R., and J.J.McN.;
Data Curation, M.P., P.L., M.B., M. McN., and T.J.N C;
Writing -- Original Draft, M.P.;
Writing -- Review & Editing, M.P., P.L., and D.M.T.;
Visualization, S.H., D.D., and W.K.;
Project Administration, A.S., and M-J.B.;
Funding Acquisition, M.E.D., and D.M.T.;
Supervision, M.E.D., and D.M.T.

Declaration of Interests

The authors declare no competing interests.

STAR Methods

Experimental Model and Subject Details

Human subjects

Participants of the MGRB were consented through the biobank programs of the ASPREE and 45 and Up studies following protocols described previously (45 and Up Study Collaborators et al., 2008; Lacaze et al., 2018). At the time of blood collection, each participant was aged 60 years or older.

Samples from the ASPREE study were from individuals aged 75 years or older at time of enrolment, with no reported history of any cancer type, no clinical diagnosis of atrial fibrillation, no serious illness likely to cause death within the next 5 years (as assessed by general practitioner), no current or recurrent condition with a high risk of major bleeding, no anaemia (haemoglobin > 12 g/dl males, > 11 g/dl females), no current continuous use of other antiplatelet drug or anticoagulant, no systolic blood pressure ≥180 mm Hg and/or a diastolic blood pressure ≥105 mm Hg, no history of dementia or a Modified Mini-Mental State Examination (3MS) score ≤77 (Teng and Chui, 1987), and no severe difficulty or an inability to perform any one of the 6 Katz basic activities of daily living (Katz and Akpom, 1976).

Samples from the 45 and Up Study were from individuals with no self-reported history of cancer, heart disease, or stroke. Neurological disease was not explicitly excluded, but participants were required to correctly self-complete a health survey at enrolment. We confirmed no record of cancer diagnosis in the NSW Central Cancer Registry, and no record of cancer diagnosis in the NSW Admitted Patient Data Collection, for all 45 and Up Study individuals in the MGRB.

Participants in the Australian Schizophrenia Research Bank (ASRB) were recruited through a national media campaign and consented to data and sample collection genomic analyses as previously described (Loughland et al., 2010).

Ethics

The ASPREE Biobank study was approved by the Monash University Human Research Ethics Committee, and subsequent whole genome sequencing of Australian ASPREE participants was approved by the Alfred Hospital Ethics Committee. The use of 45 and Up Study samples in the MGRB is covered by ethics approvals from the University of New South Wales Human Research Ethics Committee and the NSW Population & Health Services Research Ethics Committee. The use of the ASRB data was approved by the University of Newcastle Human Ethics Research Committee.

Method Details

Sample collection and processing

For ASPREE participants of the MGRB, peripheral blood samples were processed to buffy coat within 4 hours of collection, then stored at −80°C. DNA was later purified from buffy coat via magnetic bead extraction (Qiagen).

For 45 and Up Study participants of the MGRB, peripheral blood samples were refrigerated at 4°C and processed to buffy coat within 48 hours of collection. Buffy coat was stored at −80°C, and DNA purified via column extraction (Qiagen).

ASRB participant PBMCs were extracted from whole blood by centrifugation in Lymphoprep (Vital Diagnostics). Genomic DNA (gDNA) was extracted from PBMCs using salt extraction and quantified by PicoGreen assay (Life Technologies). The integrity of gDNA was determined by agarose gel electrophoresis prior to sequencing.

Sequencing

Whole genome sequencing of the MGRB, 45 and Up cancer, and ASRB cohorts was performed on Illumina HiSeq X sequencers at the Kinghorn Centre for Clinical Genomics (KCCG), Sydney, using paired-end Illumina TruSeq Nano DNA HT libraries and v2.5 clustering and sequencing reagents. 100 / 2,926 MGRB and 85 / 520 ASRB samples were sequenced to high depth (3 HiSeq X lanes per sample, equivalent to approximately 105X human genome), the remainder were sequenced to one lane per sample. Only one lane of data per sample was used to create the MGRB Phase 2 data release.

Quantification and Statistical Analysis

Sequence alignment and processing

All sequence data generated at the KCCG were processed following the Genome Analysis Toolkit (GATK) best practices (Van der Auwera et al., 2013). We first defined a custom reference genome tailored to Illumina HiSeq X sequencers, being the 1000 Genomes Phase 3 decoyed version of build 37 of the human genome (The 1000 Genomes Project Consortium, 2015), with an added contig of NC_001422.1 to act as a decoy for the HiSeq-specific ΦX174 sequence spike-in. Reads were aligned to this reference using bwa 0.7.15 mem in paired mode, and duplicates marked with biobambam2 2.0.65 bamsormadup, with a minimum optical pixel distance of 2,500. All other parameters for both bwa and bamsormadup were left at defaults. For high-depth samples run on multiple sequencing lanes, data merging was performed at this point using samtools 1.5. Indel realignment and base quality score recalibration of mapped reads were performed using GATK 3.7-0 and best practices parameters; unmapped reads were left unmodified. GATK HaplotypeCaller was used to generate g.vcfs from all single-lane realigned and recalibrated BAMs using

recommended parameters. Pipeline steps were accelerated using GNU parallel 20170722 (Tange, 2011).

Locus confidence tiers

We defined locus confidence tiers for WGS genotyping on the basis of prior annotations, sequence complexity, and empirical metrics on our data. Locus tiers ranged from 1 to 3, with 1 indicating the highest confidence in WGS variant detection performance, and 3 the lowest.

To specify the locus confidence tiers, we first identified regions of the genome which empirically had unusual coverage in the MGRB and 45 and Up cancer sequencing data. For each sample we defined bounds on the expected sequence coverage as the 0.001 and 0.999 quantiles of a Poisson distribution, with rate equal to the modal nonzero coverage observed across all autosomal loci within that sample. As typically 15 reads are required for high genotyping performance (Meynert et al., 2014), the lower bound was thresholded to always be at least 15. Within each sample, we defined each autosomal locus as being either in-bound (depth within the sample-specific bounds), or out-of-bound. We then calculated across all samples the rate at which each locus was out-of-bounds, considering the entire MGRB cohort. Regions for which this rate exceeded 5% (in other words, loci which had unusual coverage in at least 5% of MGRB + 45 and Up cancer samples) were marked as problematic. These problematic regions were smoothed by a morphological closing operation followed by an open operation, with structuring elements being centred intervals on the genome of size 131 bp and 11 bp, respectively, to yield a final definition of regions of unusual depth in the MGRB cohort. These regions totalled 409 Mb, 13.0% of the reference genome, 13.2% of the canonical chromosomes (1-22, X, Y), and 14.9% of the CCDS coding sequence (accessed 21 Nov 2017).

We then defined a poor-quality subset of the genome as all loci within 5 bp of the union of: the unusual depth regions, repeat regions identified by RepeatMasker, low complexity regions of the reference genome detected by mdust with default parameters, excludable regions from the ENCODE project, and poorly aligned or non-unique regions from the ENCODE project (Key Resources Table). This poor-quality subset totalled 1,832 Mb in size, 58.4% of the reference genome, 59.0% of the canonical chromosomes, and 18.1% of the CCDS coding sequence.

Variants in non-canonical chromosomes, the pseudoautosomal regions (X: 60001 - 2699520, 154931044 - 155260560; Y: 10001 - 2649520, 59034050 - 59363566), or within the poor-quality subset of the genome defined above, were assigned to the lowest confidence tier 3. For the remaining variants in canonical chromosomes, if the variant overlapped a high-confidence HG001 region identified by the GiaB consortium v3.3.2 (Zook et al., 2014) it was assigned the highest confidence tier 1, else it was assigned an intermediate confidence tier of 2. In total, 81.9% of the CCDS coding genome was in confidence tier 1 or 2 (Table 3).

Table 3: Quantity in megabases of the reference genome, the canonical chromosomes (1-22, X, Y), or the CCDS coding regions in each locus confidence tier.

Locus confidence tier	*Reference genome*	*Canonical chromosomes*	*CCDS*
1 -- highest	1,212	1,212	25.40
2	52	52	1.19
3 -- lowest	1,874	1,832	5.88
Total	3,137	3,096	32.47

Initial sample quality control

Poor quality MGRB and 45 and Up cancer samples were identified on the basis of genotype metrics at a small diagnostic set of loci. All 3,033 single-lane samples were genotyped at SNP loci on the Illumina Infinium QC Array 24 v1.0, using GATK GenotypeGVCFs, and quality metrics calculated within Hail v0.1 (Ganna et al., 2016). 2,904 / 3,033 (95.7%) samples passed initial quality thresholds (Table 4). Of these, 14 (0.5%) had a reported sex that did not match their genetic sex, as determined from the X chromosome inbreeding coefficient; these sex-discordant samples were not considered further. In total, 2,890 / 3,033 (95.3%) MGRB and 45 and Up cancer samples passed initial quality control (QC).

Table 4: Quality metric conditions for samples to pass quality control (QC). Two rounds of QC were performed, with different metric cutoffs: a first round based on genotypes at Illumina Infinium QC Array 24 SNPs only, and a second round based on genotypes called across the whole genome. Only samples passing all cutoffs in both rounds were included in the MGRB Phase 2 release.

Metric	*Initial QC (Infinium SNPs)*	*Final QC (full genotypes)*
Call rate	> 0.98	> 0.98
Depth standard deviation	< 10	< 10
VAF standard deviation at loci called heterozygous	< 1	< 1
Heterozygous : Homozygous variant ratio	< 2	< 2
X chromosome inbreeding coefficient	< 0.2 or > 0.8	Not tested
Singleton rate	< 0.001	Not tested

Small variant genotyping, final QC

The 2,890 MGRB and 45 and Up cancer samples passing initial QC were joint called in a single batch using GATK GenotypeGVCFs, and imported to Hail v0.1 for processing. A second round of QC (Table 4) identified an additional 31 samples with poor quality metrics not revealed by the initial QC round; these were dropped. The PCRELATE component of the

GENESIS 2.8.0 package (Conomos and Thornton, 2016) was used to determine structure-corrected relatedness between the 2,859 samples remaining, using autosomal SNPs LD-pruned with an r^2 threshold of 0.1, KING robust relatedness estimates from SNPrelate 1.12.1 (Zheng et al., 2012), and without a population reference cohort. 14 pairs of individuals related to 2nd degree or closer were identified and excluded from the cohort. MGRB (cancer-free) and 45 and Up cancer samples were split into separate cohorts at this point, and four 45 and Up cancer samples excluded on the basis of incomplete or inconsistent clinical data. In summary 2,841 unrelated samples passed all data quality requirements, comprising 2,570 cancer-free MGRB individuals, 269 45 and Up cancer samples, and the reference materials RM 8391 and RM 8398.

Cohort population structure

The MGRB cohort population structure was determined using principal components analysis (PCA), with reference to the 1000 genomes (1000G) populations. A merged dataset of all MGRB and 45 and Up cancer genotypes and the 1000G Phase 3 genotypes (The 1000 Genomes Project Consortium, 2015) was generated in Hail. To ensure high genotype concordance between platforms, merged variants were restricted to autosomal strand-specific SNPs in Tier 1 regions of the genome (see Locus confidence tiers), with a 1000G allele frequency in the range of 5% to 95%, and no evidence of deviation from Hardy-Weinberg equilibrium within any of 17 homogeneous 1000G populations ($P_{HWE} > 0.01 / 17$ for each of population codes BEB, CDX, CEU, CHB, CHS, FIN, GBR, GWD, IBS, ITU, JPT, KHV, LWK, MSL, STU, TSI, and YRI). Merged variants were LD-pruned in Hail with an r^2 threshold of 0.1, and PCA performed in Hail on biallelic variants with a combined MGRB and 1000G allele frequency in the range of 5% to 95%.

A hierarchical eigenvalue decomposition discriminant analysis classifier was constructed to assign MGRB samples to 1000G populations on the basis of PCA scores. The first classifier layer predicted a sample's 1000G superpopulation (AFR, AMR, EAS, EUR, or SAS), and the second a sample's European population (CEU, FIN, GBR, IBS, or TSI), conditional on EUR being the predicted superpopulation by the first layer. Models were trained on 1000G sample scores only using PC1-4 as predictors, then were applied to predict population source for the MGRB samples. All models were implemented using mclust v5.3 (Scrucca et al., 2016).

Small variant processing and annotation

Small variant processing and annotation was performed within Hail v0.1. Variant consequences were determined using Ensembl VEP 90 with default Ensembl release 90 databases (McLaren et al., 2016). Variants were further annotated with a range of population allele frequencies, database cross-references, and pathogenicity predictions (Table 5)

Table 5: Annotations applied to MGRB small variant data

Annotation	*Version*	*Source or citation*
1000 genomes allele frequencies	Phase 3 (2 May 2013)	(The 1000 Genomes Project Consortium, 2015)

Haplotype reference consortium allele frequencies	1-1	(McCarthy et al., 2016)
GnomAD allele frequencies	2.0.1	http://gnomad.broadinstitute.org/
dbSNP	150	(Sherry et al., 2001)
ClinVar	9 Sep 2017	(Landrum et al., 2014)
CATO	1.1	(Maurano et al., 2015)
Eigen coding	1.1, 9 May 2016	(Ionita-Laza et al., 2016)

Germline structural variant detection

Germline structural variants in the MGRB and 45 and Up cancer cohorts were detected using GRIDSS v1.4.1 (Cameron et al., 2017), excluding regions in the Encode DAC Mappability Consensus Excludable list (Key Resources Table). Where possible, linked sets of breakend calls resulting from a single rearrangement were merged into higher-level structural events. To eliminate overlap with GATK indel calls and enable assessment of cohort frequencies, structural variant events were filtered to be of length at least 50 bp, and those of the same type within a window of 100 bp were merged to the one call.

Germline mobile element insertions (MEIs) were identified using Mobster v0.2.2 (Thung et al., 2014) without blacklisting existing mobile element regions. MEI calls were then processed to remove false positive events in existing mobile element regions and to estimate variant zygosity by local realignment to the reference genome. MEIs occurring in different samples within 100 bp of each other were merged to the one call.

Rare variant burden comparison

To compare rare variant burden between the platform-matched MGRB and 45 and Up cancer cohorts, missense or nonsense variants (as judged by VEP) in ACMG SF 2.0 cancer-associated genes were joint called across both cohorts, and each variant scored for pathogenicity by ACMG criteria, blinded to cohort. The rate of individuals carrying pathogenic variants was then directly compared by Fisher's test. To exclude potential confounding due to source cohort, the 45 and Up component of the MGRB only was compared to the 45 and Up cancer cases.

To compare rare variant burden between the platform-mismatched MGRB and gnomAD non-Finnish European (NFE) WGS cohorts, the following procedure was used. Missense, nonsense, and synonymous variants in protein-coding genes were identified in each cohort using VEP, with identical parameters across both MGRB and gnomAD. Variants were further reduced to a very high-quality set, defined by the intersection of the Genome in a Bottle gold standard regions, and regions sequenced to a depth of at least 15 in at least 98% of samples in both the MGRB and gnomAD WGS cohorts. Variants with alternate alleles present at a frequency of 1% or greater in either cohort were discarded.

Given these high-confidence rare variant alleles, we calculated rare variant burden as $b = \frac{E(b_D)}{E(b_N)+E(b_D)}$, where $E(b_D)$ and $E(b_N)$ are the expected rates of individuals carrying any deleterious or all neutral genotypes in the cohort, respectively, assuming random assortment. Forms for these expectations depend on the genetic model used; for a dominant model $E(b_D) = \sum_{i \in V_D} \frac{AA_i+RA_i}{AA_i+RA_i+RR_i}$, and for a recessive model $E(b_D) = \sum_{i \in V_D} \frac{AA_i}{AA_i+RA_i+RR_i}$, with V_D denoting the set of alleles called deleterious, and AA_i, RA_i, and RR_i the counts of individuals with homozygous alternate, heterozygous, and homozygous reference genotypes for allele i, respectively. Double-alternate heterozygous loci (eg genotype AB) were not considered in this calculation, but given the rarity of the alleles considered these were very uncommon. Expectations for neutral variation $E(b_N)$ are defined analogously, except summed over the set of neutral alleles V_N.

To test the significance of differences in b between cohorts we employed a bootstrap procedure. Bootstrap draws of each source cohort genotypes were created by independent sampling with replacement of observed genotypes at each locus, and used to generate $B = 1000$ bootstrap distributions of b. A two-sided p-value was calculated as $p = \frac{2}{B+1} min\left(\sum_i \left[b_{1,(i)} < b_{2,(i)}\right] + \frac{1}{2}\sum_i \left[b_{1,(i)} = b_{2,(i)}\right] + \frac{1}{2}, \sum_i \left[b_{1,(i)} > b_{2,(i)}\right] + \frac{1}{2}\sum_i \left[b_{1,(i)} = b_{2,(i)}\right] + \frac{1}{2}\right)$, with $b_{c,(i)}$ denoting bootstrap draw i of the statistic for cohort c.

Deleterious alleles were defined as non-synonymous changes with a CONDEL score over 0.7; neutral alleles were defined as synonymous changes. Tests examined variants in all VEP-identified protein coding genes, or variants associated with arrhythmia, cardiomyopathy, or cancer (Supplementary File 5). Eight tests were performed (two models, one neutral/deleterious classification, four gene sets).

Genome-wide common variant frequency comparison

To compare patterns of common variation between the MGRB and other cohorts, we merged the MGRB variants with gnomAD v2.0.1 non-Finnish European (NFE) WGS allele frequencies, and allele frequencies from a homogeneous subset of the UK BioBank genotype set, generated as previously described (Nagpal et al., 2018). To minimise the influence of technical artefacts, variants were restricted to strand-specific biallelic SNPs listed in the EBI GWAS database, that were located in regions of the genome covered by the Genome in a Bottle standard, and were sequenced to a depth of at least 15 in at least 98% of samples in both the MGRB and gnomAD WGS cohorts. Further, variants which were not observed in one or more cohorts, or were genotyped at a rate of less than 97% in any cohort, were excluded. 21,033 SNPs remained following this filtering, with very similar allele frequencies across all cohorts (Supplementary Data File 2, sheet 1; Supplementary Figure 3); these loci and frequencies were used in the following common variant analyses.

We tested for phenotype-linked bias in allele frequency between the cohorts as follows. For a given phenotype-associated set of variants, each variant was scored on two metrics: its variant allele frequency enrichment or depletion in MGRB versus gnomAD or UKBB, and the

positive or negative association of the variant allele with the trait. A Fisher's exact test was then used to test for dependence of variant enriched/depleted status on the trait direction of effect, with deviation from the null indicating an allele frequency bias between MGRB and gnomAD or UKBB that is specific to the phenotype considered.

Three sets of variants were tested by this procedure: a test set of variants reported to be associated with phenotypes depleted in the MGRB (Supplementary Data File 2, sheets 2-3), and two negative control sets of variants linked to anthropometric traits (Supplementary Data File 2, sheets 4-5), or behavioural traits (Supplementary Data File 2, sheets 6-7).

Polygenic score estimation and testing

Polygenic scores were calculated as $S_i = \sum_j \beta_j d_{ij}$ where S_i is the polygenic score for individual i, β_j the GWAS-reported coefficient for a single variant allele at locus j, and d_{ij} is the variant allele dosage for individual i at locus j. We considered only autosomal variants, and if a variant dosage was not available for an individual, it was imputed as $\widehat{d_{ij}} = 2f_j$, with f_j the variant allele frequency reported by the source publication. To reduce bias due to this imputation, variants with a call rate under 97% were excluded from polygenic score calculation in all individuals.

Polygenic score GWAS coefficients were derived from processing of summary statistics for genome-wide significant loci in reporting papers, for colorectal cancer (Schumacher et al., 2015), melanoma (Law et al., 2015), breast cancer (Michailidou et al., 2017), prostate cancer (Hoffmann et al., 2015), blood pressure (Warren et al., 2017), early-onset coronary artery disease (EOCAD) (Thériault et al., 2018), atrial fibrillation (Lubitz et al., 2017), height (Wood et al., 2014), Alzheimer's disease (Lambert et al., 2013), and longevity (Deelen et al., 2014). GWAS coefficients were used as-is for the continuous trait of height. For all other binary traits, coefficients were converted to a log-odds scale. The Alzheimer's disease PRS as originally reported lacked the highly significant *APOE* locus; accordingly this locus was manually added to the PRS using the tag SNP rs10414043, and an estimated $\beta = 1.34$ (Genin et al., 2011). rs10414043 was used in preference to the more conventional rs429358 as the latter was not robustly genotyped on all platforms. All loci, alleles, and coefficients used in the PRS calculations are detailed in Supplementary File 2.

An approximate bootstrap procedure was used to test for polygenic score shift between MGRB, gnomAD, UKBB, and the 45 and Up cancer cohort. All cohorts were first collapsed to allele-frequency data only, with individual genotypes discarded. For a given polygenic score and a set of cohorts to compare, testing then proceeded as follows. Polygenic score variants were first subset to those called at a rate of at least 97% in each cohort, and with an absolute difference in alternate allele frequency between MGRB, gnomAD, or UKBB of less than 4%. A bootstrap sample of genotypes was then drawn independently for each cohort and locus, polygenic scores calculated for each bootstrapped individual as above, and the mean cohort polygenic scores calculated. This was repeated for 5,000 bootstrap rounds to yield bootstrap distributions of the mean polygenic score in each cohort. To facilitate comparisons between scores of different scales, bootstrap distributions for each score were

normalised by an affine transformation that brought the 0.025 and 0.975 quantiles of the MGRB scores to values of -0.5 and 0.5 respectively. The 95% bootstrap confidence intervals for each cohort were then defined as the 0.025 and 0.975 quantiles of the normalised scores. Approximate two-sided p values for the overlap between two bootstrap distributions were estimated as $p \approx \frac{2}{B+1}(\frac{1}{2} + min(\sum_{k=1}^{B}[s_{1,(k)} < s_{2,(k)}], \sum_{k=1}^{B}[s_{1,(k)} > s_{2,(k)}]))$ where $s_{1,(k)}$ and $s_{2,(k)}$ are samples from the the bootstrap mean values for cohorts 1 and 2, respectively, $B = 100000$ is the number of samples taken with replacement from the bootstrap mean values, and $[\,]$ denotes Iverson brackets.

The statistical power improvement from using the MGRB extreme phenotype cohort as a control versus gnomAD was estimated by asymptotic approximation. Bootstrap distribution means of the mean prostate cancer score difference between the 45 and Up cancer cases, and gnomAD and MGRB controls, were used as mean shift values for statistical power calculation. After correcting for varying cohort sample size, bootstrap distribution variance was highly consistent across all three cohorts, and pooled variance scaled to a sample size of 1 was used as the dispersion parameter. Power was then calculated across a range of sample sizes for both the MGRB vs 45 and Up cancer, and gnomAD vs 45 and Up cancer tests, by direct root finding of the relevant t distributions. Finally, the power vs sample size relationship was inverted by piecewise linear interpolation to yield the sample size vs power curves.

Individual genotypes were available for both MGRB and 45 and Up cancer cohorts. In these cases, a secondary analysis was performed that directly compared the distributions of individual polygenic scores between cohorts. Height prediction was validated by ordinary linear regression of measured individual height against the polygenic height predictor (Wood et al., 2014) with additional additive linear covariates of sex and age at measurement; no evidence for model misspecification was observed. The association between polygenic score and risk of specific cancers was assessed by logistic regression, with the effect of polygenic score on cancer risk modelled by GCV-penalised thin plate splines. Comparisons were restricted to the specific cancers of prostate, colorectal, and melanoma, as other cancers were either poorly sampled in the 45 and Up cancer cases, or did not have polygenic scores defined.

Incidental somatic variant detection

Somatic variants were identified in post-BQSR BAM files using FreeBayes, with options: --pooled-continuous --standard-filters --min-alternate-fraction 0 --min-alternate-count 3 --hwe-priors-off --allele-balance-priors-off --use-mapping-quality. FreeBayes was restricted to detecting variants within 10 kb of RefSeq genes in the COSMIC Cancer Gene Census (Forbes et al., 2017) downloaded 11 December 2017. Variant annotation was performed using the Ensembl VEP (McLaren et al., 2016) release 90, with default options, and variants were notated with COSMIC 83 frequencies.

Annotated variants were filtered to retain only non-synonymous variation (missense, splice donor or acceptor, start lost, stop gained, frameshift, or inframe indel) affecting Cancer Gene

Census Tier 1 genes, with a maximum population allele frequency of less than 0.1%, a variant allele fraction (VAF) of at least 10%, and three or more reads supporting the variant. We then identified likely driver mutations from these filtered variants by the following criteria: either a variant had a HIGH consequence in a canonical tumour suppressor gene transcript, or the variant was observed at least 100 times in the COSMIC database. Consequences and canonical transcripts were as defined by Ensembl VEP; tumour suppressor genes were Tier 1 genes from the COSMIC Cancer Gene Census with a TSG annotation.

Sequence-based measures of age

Telomere length

Telomere lengths were estimated using Telseq v0.0.1 (Ding et al., 2014). To reduce batch effects between the ASRB and MGRB cohorts, ASRB telomere length estimates were calibrated using Deming regression, fit to 85 ASRB samples sequenced both in the original ASRB batch, and contemporaneously with the MGRB.

Telomere length estimation by Telseq was validated by qPCR on a subset of 120 samples from the ASRB and MGRB cohorts (Supplementary Figure 7), as described previously (Cawthon, 2002) with minor modifications. Briefly, qPCR was conducted in triplicate. Reactions included: genomic DNA (5 ng), 2x Rotor-Gene SYBR Green Master Mix (Qiagen), 500 nM Tel forward [5'-CGGTTT(GTTTGG)$_5$GTT-3'] and 500 nM Tel reverse [5'-GGCTTG(CCTTAC)$_5$CCT-3'] or 300 nM 36B4 forward [5'-CAGCAAGTGGGAAGGTGTAATCC-3'] and 500 nM 36B4 reverse [5'-CCCATTCTATCATCAACGGGTACAA-3'] in a 25 µL reaction. Amplification was conducted in a Rotor-Gene Q qPCR cycler (Qiagen) at 95°C for 5 min, followed by 30 cycles of 95°C for 7 sec and 58°C for 10 sec (telomere reaction) or 35 cycles of 95°C for 15 sec and 58°C for 30 sec (single copy gene reaction). Telomere content for each sample was determined by the telomere to single copy gene ratio (T/S ratio) by calculating ΔCt ($Ct_{telomere}$ / $Ct_{single\ copy\ gene}$). The T/S ratio of each sample was normalized to the mean T/S ratio of a reference sample, which was included in each run. The experiment was accepted if the reference sample T/S ratio ranged within 95% variation interval, and if the standard curve had a high correlation factor ($R^2 > 0.95$).

Mitochondria and Y chromosome copy number

Mean mitochondrial genome copy number in each sample was estimated using read counts, as $2 \times (R_{MT} \div S_{MT}) \div (R_A \div S_A)$, where R_Z and S_Z denote the number of reads mapping to contig set Z and the total size of contig set Z, and MT and A denote mitochondrial and autosomal contigs, respectively. Read counts were mapped and aligned reads reported by samtools idxstats, and were not corrected for read duplication. Patch contigs were not included in counts. Y copy number in males was estimated by an analogous procedure, as $2 \times (R_Y \div S_Y) \div (R_A \div S_A)$.

Mitochondrial variants

Variants in the mitochondrial genome were detected using FreeBayes, considering only reads with base quality over 24 and mapping quality over 30; all other parameters were left

at defaults. Variants with fewer than 10 alternate reads, or an alternate allele fraction under 0.001, were discarded. For each variant passing these filters a Phred-like quality score q was calculated as $q = -10log_{10}(1 - F(n;p,N))$, with n the count of alternate allele reads, N the total depth at the variant locus, $p = 0.0025$ a fixed error rate estimate, and $F(n;p,N)$ the cumulative density function of a binomial distribution with N draws and success probability p. Variants with $q < 30$, high depth variants ($n > 15$) with an alternate read strand bias of greater than 0.9, or variants in the highly variable locations MT:302-319 or MT:3105-3109 were discarded. The final metric of mitochondrial variant burden for a sample was defined as the number of low-frequency (variant allele fraction under 0.01) variants passing all above filters in that sample.

Somatic single nucleotide variants

Somatic SNV burden was estimated using a combination of statistical filtering and spectral denoising. Putative somatic SNVs were first identified on the basis of a variant allele frequency that was statistically inconsistent with either machine error or germline variation. The burden of these variants in each sample was then dimensionally reduced by spectral factorization, and per-sample signature scores used as the final somatic variant estimates.

We first developed a statistical filtering procedure to identify likely somatic variants, that uses dynamic thresholds to optimise sensitivity while controlling signal to noise ratio. This procedure calls a variant at a given locus as likely somatic if it satisfies the following criterion:

$$c_E \leq n_A \leq c_H$$

where n_A is the number of non-reference allele reads at the locus, and c_E and c_H are integers which maximise:

$$p_n = r_{RR} \sum_{n_A=c_E}^{c_H} \binom{n}{n_A} p_A{}^{n_A}(1-p_A)^{n-n_A}$$

subject to:

$$\frac{r_c}{1-r_c}\frac{p_n}{r_{RR}\alpha_E + r_H\alpha_H} \geq g_r$$

with $r_{RR} = \frac{1}{2}\left(1 - r_H + \sqrt{1+2r_H}\right)$, and $p_A = \left(1 - \frac{4}{3}\varepsilon\right)f + \varepsilon$. Here n is the sum of reference and alternate allele depths at the locus, r_H is the expected rate of heterozygous variant germline loci, r_C the expected rate of somatic variant loci, ε the base read error rate, g_r the minimum acceptable ratio of true positive calls to false positive, and f is the expected somatic variant allele fraction. α_E and α_H are test sizes corresponding to thresholds c_E and c_H: $c_E \equiv inf\{n_A : Pr(err) < \alpha_E\}$, $c_H \equiv sup\{n_A : Pr(het) < \alpha_H\}$, $Pr(err) = \sum_{i=n_A}^{n} \binom{n}{i}\varepsilon^i(1-\varepsilon)^{n-i}$, $Pr(het) = \sum_{i=0}^{n_A} \binom{n}{i}\left(\frac{1}{2}+\frac{1}{3}\varepsilon\right)^i\left(\frac{1}{2}-\frac{1}{3}\varepsilon\right)^{n-i}$. Informally, this procedure selects variants with an alternate allele count n_A too large to be due to sequencing error ($n_A \geq c_E$), yet too low to be from a poorly-sampled heterozygous germline locus ($n_A \leq c_H$). The derivation of this procedure and further details are available in Supplementary File 4. A notable advantage of this procedure is that it yields per-locus estimates of variant detection sensitivity, as p_n. These estimates are critical in normalization of variant detection rates to account for

differential coverage across variable sequencing runs, which is necessary for the accurate estimation of sample somatic variant burden.

Insufficiencies of the simple error model used above result in incomplete control of the signal to noise ratio, and further filtering is required to reliably quantify somatic variant burden. Assuming that true somatic events and false positive machine noise exhibit differential sequence context bias, we use a spectral dimensionality reduction approach to achieve additional denoising and summarise the total somatic variant burden in each sample. Extending previous cancer somatic signature work (Alexandrov et al., 2013; Gehring et al., 2015), we calculate per-sample sensitivity-normalised somatic variant burden for each of 96 single-nucleotide variant classes, as the total number of detected somatic events of a given class in that sample, divided by the sum of p_n in that sample for all loci corresponding to the given variant class. The resulting $96 \times n$ normalised burden matrix is then reduced by non-negative matrix factorization (Brunet et al., 2004), using 100 random optimisation starting points per cardinality. To select the appropriate factorization cardinality, we reduce the burden matrix by merging groups of 16 age-consecutive samples by summing burden for each variant class, and factorize this reduced matrix with 100 random restarts and cardinality ranging from 2 to 10. The lowest cardinality that gives an inflection point on the plot of explained variance versus cardinality is selected, and applied to the full burden matrix. Per-sample scores are extracted from the best of 100 random runs at this final selected cardinality.

We applied the above procedure to the MGRB and ASRB samples, with parameters adapted to maximise sensitivity with low-depth sequencing data: $g_r = 5$, $f = 0.2$, $\varepsilon = 2.0 \times 10^{-3}$, $r_H = 1.0 \times 10^{-4}$, $r_C = 5.0 \times 10^{-7}$. Our filtering process employed SNVs identified by samtools mpileup, with maximum depth 101, mapping quality adjustment of 50, BAQ recalculation, no indel reporting, and minimum read and mapping qualities of 30, and employed a blacklist of common SNPs observed in either MGRB or dbSNP. The factorization cardinality procedure applied to our data indicated that three signatures best described the mutation patterns observed (Supplementary Figure 8). Signature 3 in this work was quantitatively similar to COSMIC signature 5 (cosine similarity 0.81), previously reported to be associated with age at cancer diagnosis (Alexandrov et al., 2015), and the per-sample scores for this signature were used as the summative somatic burden measure. Signature 1 from this work was very similar to COSMIC signature 1 (cosine similarity 0.95), which has also been associated with spontaneous deamination processes and age. However, we observed substantial inter-cohort differences in score distribution for this signature, suggestive of high technical variability, and did not examine it further.

Somatic copy number variants

We developed a model-based strategy to identify subclonal copy number variants (CNVs), assuming a single genetically homogeneous subclone present on a background of diploid cells.

We first defined a set of autosomal SNPs with highly stable sequencing characteristics on our platform. We selected loci containing autosomal biallelic SNPs in the MGRB cohort, with

a variant allele fraction between 5% and 95%, and a mean GC content in the surrounding 100 bp of between 30% and 55%. These were further filtered to retain only loci with highly consistent coverage in both the MGRB and ASRB cohort data, with mean(DP_{rel}) ∈ [0.9, 1.1] in both cohorts, var(DP_{rel}) ∈ [0.025, 0.033] in the MGRB, and var(DP_{rel}) ∈ [0.025, 0.040] in the ASRB cohort. Here DP_{rel} is locus depth relative to mean sample depth, and statistics are calculated over all samples in each cohort. 1,862,065 loci passed all filters, with a median inter-locus distance of 626 bp, and 5th and 95th percentiles of 30 and 4,904 bp, respectively.

We individually genotyped MGRB, 45 and Up cancer, and ASRB samples at this set of highly reliable loci using GATK HaplotypeCaller with default parameters, except for a variant window size of 100 bp (-ip 100). Within each sample, the depths of reference and variant alleles at all heterozygous SNV target loci were fit to the following subclonal CNV model, to produce estimates of local ploidy and global sample subclonal fraction.

Consider a locus i in a single sample which contains fraction f of aneuploid cells, the remaining $1-f$ being entirely diploid (gonosomes are not modeled). We denote the copy number (ploidy) of the aneuploid cells at i with k_{1i} and k_{2i}, $k_{\bullet i} \in \mathbb{N}$. For example, $k_{1i}, k_{2i} = (1, 1)$ denotes a diploid state (no aneuploidy), $k_{1i}, k_{2i} = (1, 0)$ the deletion of one allele, and $k_{1i}, k_{2i} = (2, 2)$ duplication of both alleles. Our task is to estimate k_{1i}, k_{2i} for all i, and f globally, given reference and non-reference allele depths d_{ri} and d_{ai}.

The extent to which the aneuploid cell ploidies k_{1i} and k_{2i} affect the representation of alleles in the mixed cell population depends on the aneuploid cell fraction f. Let p_{1i} and p_{2i} represent the mean ploidy of each chromatid in the mixed cell DNA pool. As the pool is assumed to consist of only two populations, with $1-f$ of the cells diploid, $p_{1i} = fk_{1i} + (1-f)$ and $p_{2i} = fk_{2i} + (1-f)$.

We assume that the sequencer does not exhibit allelic bias. Then, $E[d_{1i}] = c_i p_{1i}$, $E[d_{2i}] = c_i p_{2i}$, with c_i a normalising constant to account for the depth at locus i. Here d_{1i} and d_{2i} denote the depths of reads from chromatid 1 and 2, respectively. Unfortunately we do not have phased genotypes, and so cannot easily determine the chromatid source of each read. Instead we have unphased reference and non-reference depths d_{ri} and d_{ai}, and must account for the resulting phase uncertainty with a mixture model.
Disregarding allele phasing we model the depths of reference and non-reference reads at i using a mixture: $d_{ri}, d_{ai} \sim D(c_i p_{1i}), D(c_i p_{2i})$ with probability $\frac{1}{2}$, else $d_{ri}, d_{ai} \sim D(c_i p_{2i}), D(c_i p_{1i})$, $D(\mu)$ denoting a distribution function with expected value μ. In our implementation we employ a negative binomial distribution for D, with probability mass function $f_D(x; \mu, s) \equiv \frac{\Gamma(x+s)}{\Gamma(x+1)\Gamma(s)} q^s (1-q)^x$, $q \equiv \frac{s}{s+\mu}$. The size term s captures overdispersion relative to the Poisson distribution, and is optimised per-sample in the model fit.

The normalising constant c_i is half the expected depth at locus i, which is itself a complex function of locus- and sample-specific factors. We model this function at the locus- and sample-level using empirical cohort depth measurements, and a sample-specific GC bias correction. Specifically, we define $c_i \equiv b_i exp(h(g_i))$, where b_i is the mean relative depth of

locus i (where relative depth is defined as $d_i \div \frac{1}{n}\sum_i d_i$, with $d_i \equiv d_{ri} + d_{ai}$ and n the number of target loci, $n = 1862065$), across the sample's cohort (either MGRB or ASRB), h is a smooth function, and g_i is a vector of GC fraction in windows of various size around locus i. For this work, g_i was a 5-vector of GC fraction in windows of size 100, 200, 400, 600, and 800 bp, calculated on the reference sequence centered at locus i. The sample-specific GC correction function h was implemented using a generalised additive model (GAM) with five smooth terms, and fit to all heterozygous loci for each sample as $ln(c_i \div b_i) \sim s(r_{1i}) + s(r_{2i}) + s(r_{3i}) + s(r_{4i}) + s(r_{5i})$, with r_{ji} being the score of the j th principal component of the GC fraction matrix for locus i, and s denoting a penalised regression spline term. GAMs were fit using mgcv 1.8-17 (Wood, 2004) with default parameters.

A greedy agglomerative algorithm was used to segment the genome of each sample into regions of differing ploidy state. Initially the genome was divided into segments of 100 consecutive heterozygous loci, with segment boundaries enforced between chromosomes. Adjacent segments were tested for identical distribution of $d_{ri} \div c_i, d_{ai} \div c_i$ by a 2-sample Kolmogorov-Smirnov test, and the two segments with the highest p value genome-wide were merged. This process was repeated until either no segment pairs remained to merge, or all Kolmogorov-Smirnov test p values were less than 0.01. Segments were never merged between chromosomes.

The above model was fit to the allele counts within each genome segment by maximum likelihood. Ploidies of each segment, k_{1i} and k_{2i}, as well as the global aneuploid fraction f and overdispersion s, were optimised by grid search with local polishing. As very high ploidies coupled with low f result in highly expressive but likely incorrect models, the maximum allowable ploidy k_{max}; $k_{1i}, k_{2i} \leq k_{max}$ was determined in an outer loop through minimisation of the Bayesian Information Criterion (BIC). A final polishing step was applied to the BIC-optimal model, which merged consecutive segments of the genome if they were assigned identical chromatid ploidies by the model. This final polished model yielded global cell fraction f, as well as local ploidies across the genome, for the single aneuploid clone assumed to be present in each sample.

Clonal haematopoiesis

Extending previous work (Jaiswal et al., 2014), clonal haematopoiesis of indeterminate potential (CHIP) was defined in an individual if either: a somatic small variant (see section *Incidental somatic variant detection*) was detected with a variant allele frequency of at least 10%, or somatic copy number variation (see section *Somatic copy number variants*) indicated the presence of a clone comprising at least 10% of nucleated blood cells.

Somatic burden statistical analysis

Exploratory analysis indicated that variable transformation was required for some measures. For the following analyses, telomere length and Y copy number were modelled as-is; somatic variant burden, mitochondrial load, and mitochondrial variant count were

log-transformed prior to modelling; and grip strength in kg was power transformed with exponent 0.7.

Within-cohort trends in somatic measures were estimated by linear regression, with 95% Wald confidence intervals. Likelihood ratio tests of nested models were used to evaluate inter-cohort trend differences, with p-values corrected for multiple testing by Holm's step-up procedure (Holm, 1979).

We used a permutation procedure to test the importance of somatic burden measures in predicting grip strength and gait speed, conditioned on age. For each of eighteen possible frailty measure x somatic measure x sex combinations (Frailty measures: grip strength, gait speed; Somatic measures: Telseq telomere length, nuclear somatic variant burden, mtDNA copy number, mitochondrial variant count, and Y copy number in males only), we calculated the deviance of the following generalised additive model $frailty \sim s(age) + s(weight) + s(BMI) + s(abdocirc) + s(somatic)$, with age in years, weight in kg, BMI in kg/m^2, abdominal circumference (abdocirc) in cm, and the somatic measure of interest (transformed if relevant following exploratory analysis). In this model specification, $s(x)$ denotes a GCV-penalised thin plate spline smooth term in x as implemented in R package mgcv (Wood, 2004), with Gaussian error and identity link. This model's deviance d was compared to the deviance $d_{(i)}$ of 10,000 models fit in the same manner but with the somatic variable permuted, and a p-value estimated as $\hat{p} = \frac{1}{10{,}001}(\sum_i \left[d_{(i)} \leq d \right] + 0.5)$. To address multiple testing concerns we used a two stage process. In the first stage p-values were calculated as above for all 18 tests on a randomly selected subset of 25% of the ASPREE samples. Tests with a p-value less than 0.2 in the first stage were tested in the second validation stage on the remaining 75% of the ASPREE samples, and these second-stage p-values corrected for multiple testing by Holm's method.

We observed cohort differences in intercepts in plots of somatic measures versus age. To remove these solely for the purposes of illustration (Figure 3), for each somatic measure we fit the generalised additive model $measure \sim s(age, by = sex) + cohort$, with Gaussian error and identity link. In this model specification $s(age, by = sex)$ denotes a GCV-penalised thin plate spline with age as the predictor variable, stratified by sex. Model fits were performed using the R package mgcv (Wood, 2004). After confirming the suitability of the model fits, cohort-specific effects were removed by calculating the quantity $y_i' = y_i - \hat{s}_C + \hat{s}_{ASPREE}$ for each individual and measure, where y_i' is the cohort-corrected somatic measure for individual i, to be plotted; y_i is the original measurement for individual i in cohort C; and $\hat{s}_C$ and $\hat{s}_{ASPREE}$ are the model estimates of the cohort intercept term for cohort C and the ASPREE cohort, respectively. In this manner, somatic measurements were transformed to have an intercept matching that fitted to the ASPREE cohort.

We used the following procedure to illustrate the effect of mtDNA copy number on grip strength in males. For each male individual i in the ASPREE cohort, an age-local quantile of mitochondrial DNA copy number c_i was defined as $q_i \equiv \hat{F}_i(c_i)$, where $\hat{F}_i$ is the empirical

cumulative distribution function of c in the neighbourhood of individual i, with the neighbourhood of an individual i defined as all male ASPREE individuals within ± 1 year of age of i. Ages were rounded to the nearest integer for the purposes of neighbourhood definition; for the median ASPREE male age of 80 years, this neighbourhood contained 293 men with ages in $[79, 81]$ years. Given these local mtDNA copy number quantile estimates q, a generalised additive model of the form $gripstrength \sim age + s(q)$ was fit using the R package mgcv (Wood, 2004), with s smooth term as above. Predictions from this model with $age = 80$ and varying q defined the estimated influence of age-local mtDNA copy number on grip strength for an 80 year old man. These grip strength predictions were transformed to effective age estimates assuming typical mtDNA copy number by inversion of the model predictions for $s = 0.5$, and used to calculate an age excess as a function of q. Variability of this relationship was estimated using 100,000 bootstrap samples, and results presented as highest posterior density intervals.

Data and Software Availability

Summary variant frequency data for the MGRB cohort are available at the web portal: https://sgc.garvan.org.au/explore. Complete genotype, phenotype, and raw data are available upon application to Prof. David Thomas (d.thomas@garvan.org.au), or sgc@garvan.org.au. Source code for all analyses is available at https://github.com/mpinese/mgrb-manuscript; source code for the somatic SNV and LoH detection tools can be found at https://github.com/mpinese/soma-snv and https://github.com/mpinese/soma-cnv.

Additional Resources

The Medical Genome Reference Bank web portal: https://sgc.garvan.org.au/explore

Key Resources Table

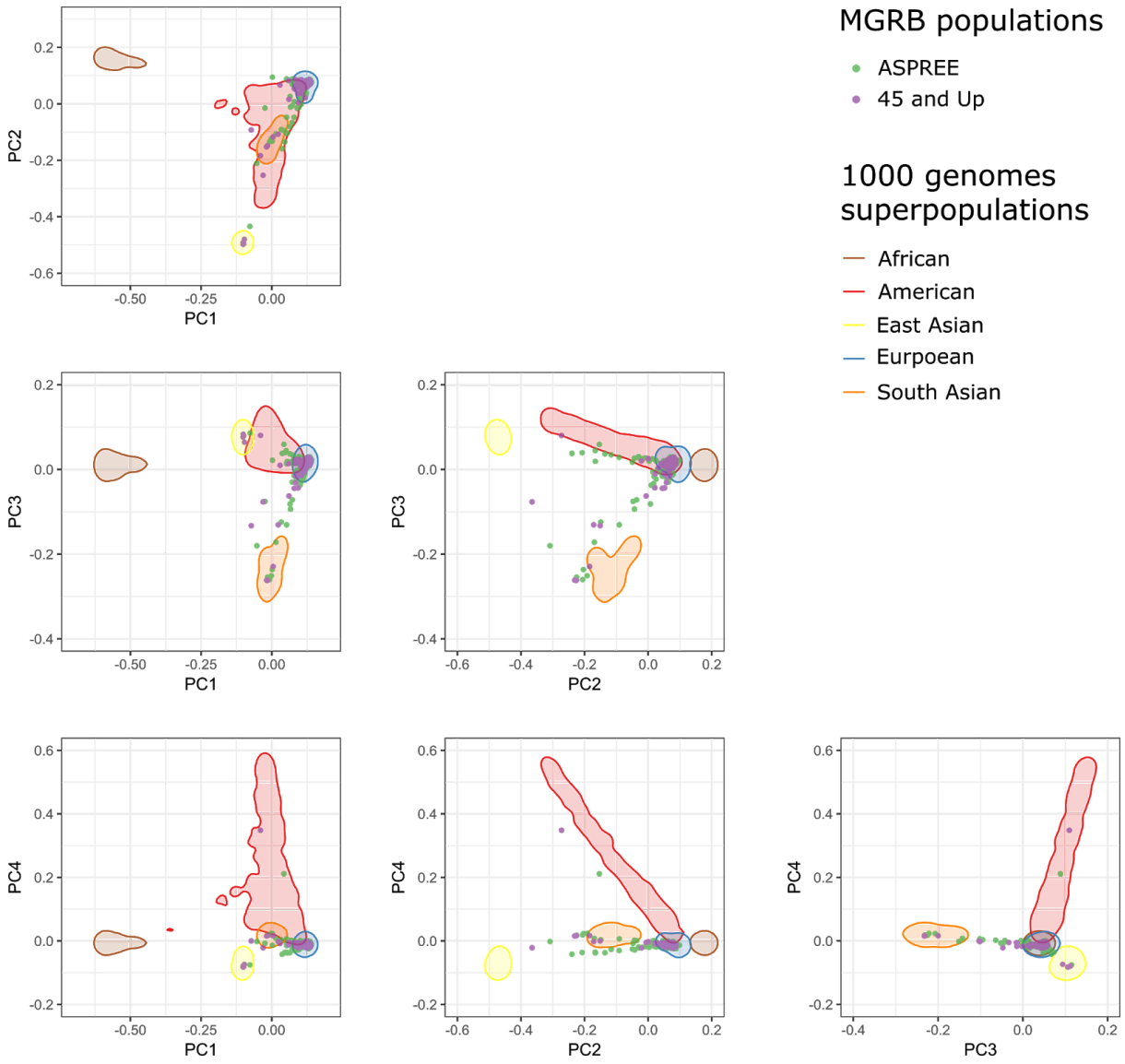
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
TruSeq Nano DNA HT library	Illumina	
HiSeq X clustering and sequencing reagents	Illumina	v2.5
Deposited Data		
Human reference genome, 1000 Genomes Project phase 3 build 37 with decoy	1000 Genomes Project	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz
ΦX174 genome	NCBI RefSeq	NC_001422.1

Consensus CDS coding regions, accessed 21 Nov 2017	UCSC browser table ccdsGene	https://genome.ucsc.edu/cgi-bin/hgTables
Repetitive regions, last updated 27 April 2009	UCSC annotations database	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz
ENCODE excludable regions, DAC, last updated 5 May 2011	UCSC browser table wgEncodeDacMapabilityConsensusExcludable	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgEncodeDacMapabilityConsensusExcludable.txt.gz
ENCODE excludable regions, Duke, last updated 29 March 2011	UCSC browser table wgEncodeDukeMapabilityRegionsExcludable	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgEncodeDukeMapabilityRegionsExcludable.txt.gz
Poor alignment or uniqueness, CRG, last updated 27 April 2010	UCSC browser table wgEncodeCrgMapabilityAlign100mer score < 1	https://genome.ucsc.edu/cgi-bin/hgTables
Poor alignment or uniqueness, Duke, last updated 12 May 2011	UCSC browser table wgEncodeDukeMapabilityUniqueness35bp score < 1	https://genome.ucsc.edu/cgi-bin/hgTables
Infinium QC Array 24 1.0 loci	Illumina	https://support.illumina.com/downloads/infinium-qc-array-24-v1-0-support-files.html
GiaB HG001 high-confidence regions 3.3.2	(Zook et al., 2014)	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_nosomaticdel.bed
1000 Genomes Project phase 3 genotypes, released 2 May 2013	(The 1000 Genomes Project Consortium, 2015)	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

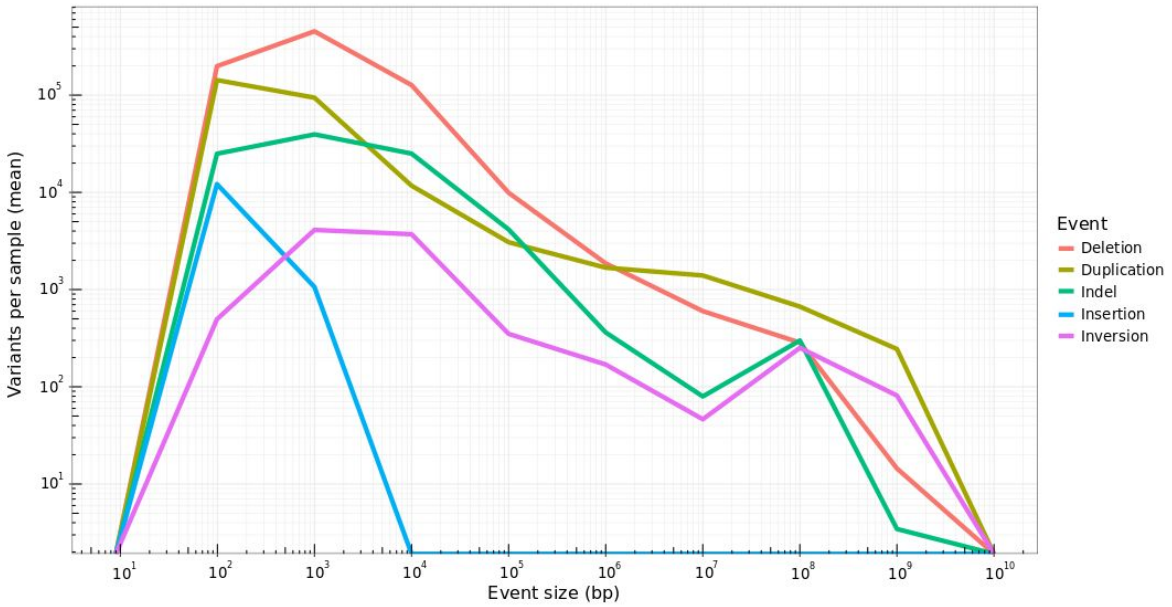
Haplotype reference consortium allele frequencies, 1-1	(McCarthy et al., 2016)	
GnomAD allele frequencies 2.0.1		http://gnomad.broadinstitute.org/
dbSNP 150	(Sherry et al., 2001)	http://www.haplotype-reference-consortium.org/
ClinVar, downloaded 9 September 2017	(Landrum et al., 2014)	https://www.ncbi.nlm.nih.gov/clinvar/
CATO 1.1	(Maurano et al., 2015)	http://www.mauranolab.org/CATO/
Eigen coding 1.1 (9 May 2016)	(Ionita-Laza et al., 2016)	http://www.columbia.edu/~ii2135/eigen.html
COSMIC Cancer Gene Census, downloaded 26 April 2018		http://www.sanger.ac.uk/science/data/cancer-gene-census
UK BioBank	(Sudlow et al., 2015)	https://www.ukbiobank.ac.uk/
Software and Algorithms		
BWA 0.7.15		https://github.com/lh3/bwa
biobambam2 2.0.65-release-20161130121735	(Tischler and Leonard, 2014)	https://github.com/gt1/biobambam2
samtools 1.5	(Li et al., 2009)	https://github.com/samtools
mdust commit 3e3fed8		https://github.com/lh3/mdust
GATK 3.7.0-gcfedb67	(DePristo et al., 2011)	https://software.broadinstitute.org/gatk/
vt 0.5722-60f436c3	(Tan et al., 2015)	https://github.com/atks/vt
Hail 0.1-0320a61	(Ganna et al., 2016)	https://github.com/hail-is/hail
VEP 90-3fcc9dd	(McLaren et al., 2016)	https://github.com/Ensembl/ensembl-vep

TelSeq 0.0.1-be185ec	(Ding et al., 2014)	https://github.com/zd1/telseq, tagged release 0.0.1, commit be185ec, downloaded Feb 8, 2017
GRIDSS 1.4.1	(Cameron et al., 2017)	https://github.com/PapenfussLab/gridss, version 1.4.1
FreeBayes 1.1.0-54-g49413aa		https://github.com/ekg/freebayes, version 1.4.1
R 3.5.0	(R Core Team, 2017)	https://www.r-project.org/
GENESIS 2.8.0	(Conomos and Thornton, 2016)	https://bioconductor.org/packages/release/bioc/html/GENESIS.html
SNPrelate 1.12.1	(Zheng et al., 2012)	https://bioconductor.org/packages/release/bioc/html/SNPRelate.html
mclust 5.3	(Scrucca et al., 2016)	https://cran.r-project.org/web/packages/mclust/index.html
mgcv 1.8-17	(Wood, 2004)	https://cran.r-project.org/web/packages/mgcv/index.html
SomaticSignatures 2.16.0	(Gehring et al., 2015)	https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html

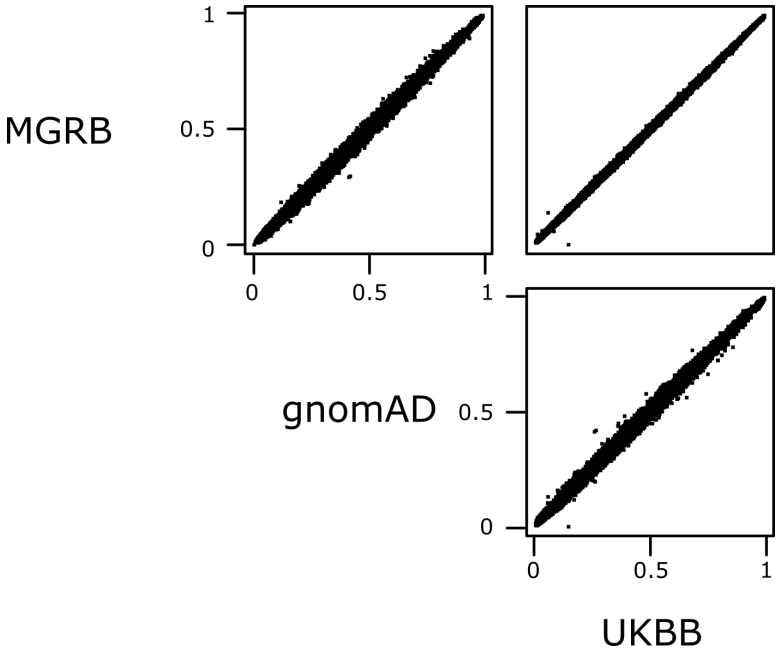
Supplemental Information



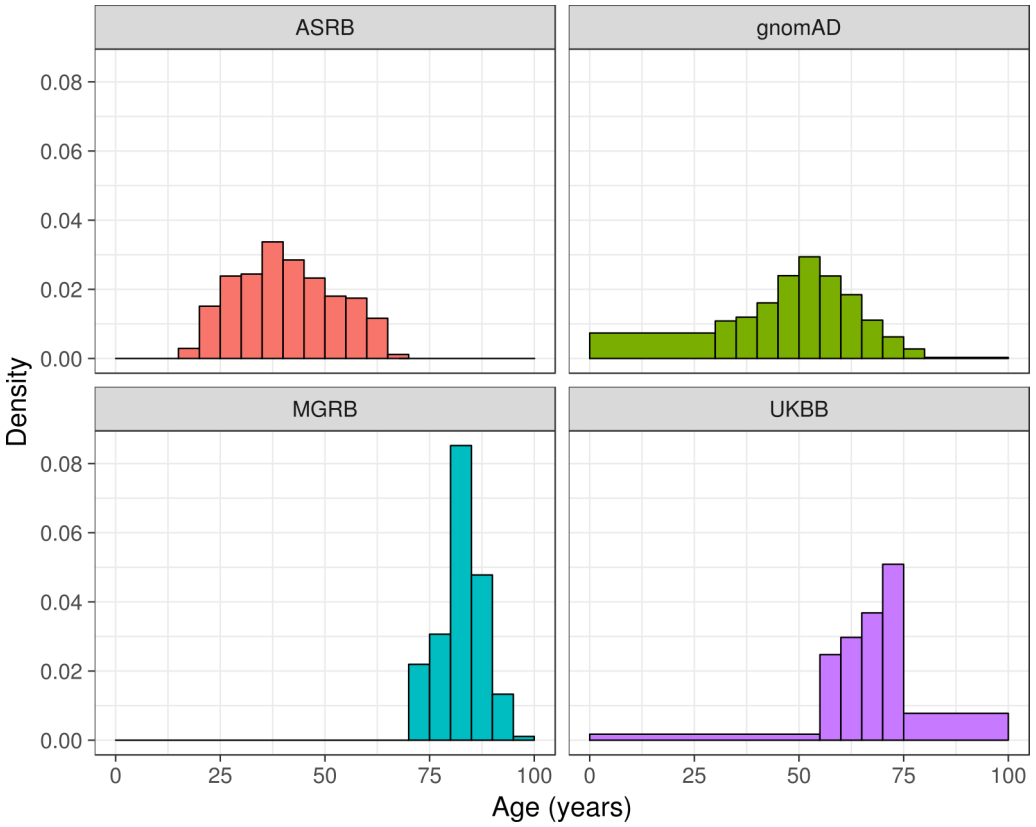
Supplementary Figure 1: Population structure in the MGRB. The MGRB was combined with the 1000 Genomes cohort at high-confidence SNVs, and PCA was performed following LD pruning. Four strong components resulted, scores for which are shown relative to 95% kernel density estimates of the 1000 Genomes superpopulations. The MGRB cohort was largely homogeneous and clustered with the 1000 Genomes European superpopulation.



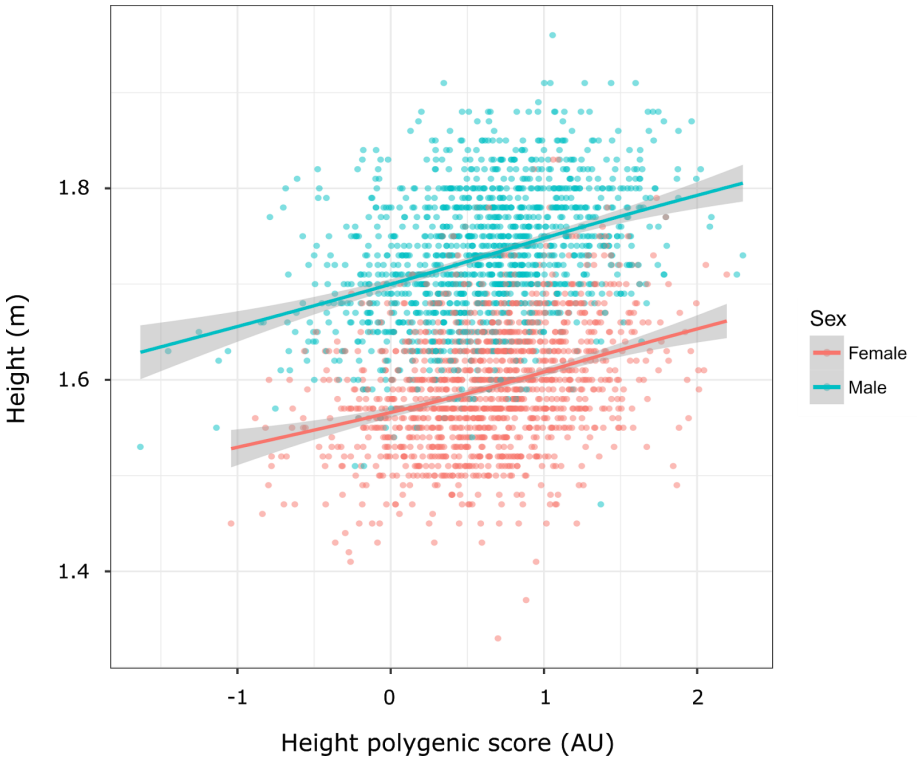
Supplementary Figure 2: Distribution of structural variant event types and sizes detected in the MGRB by GRIDSS.



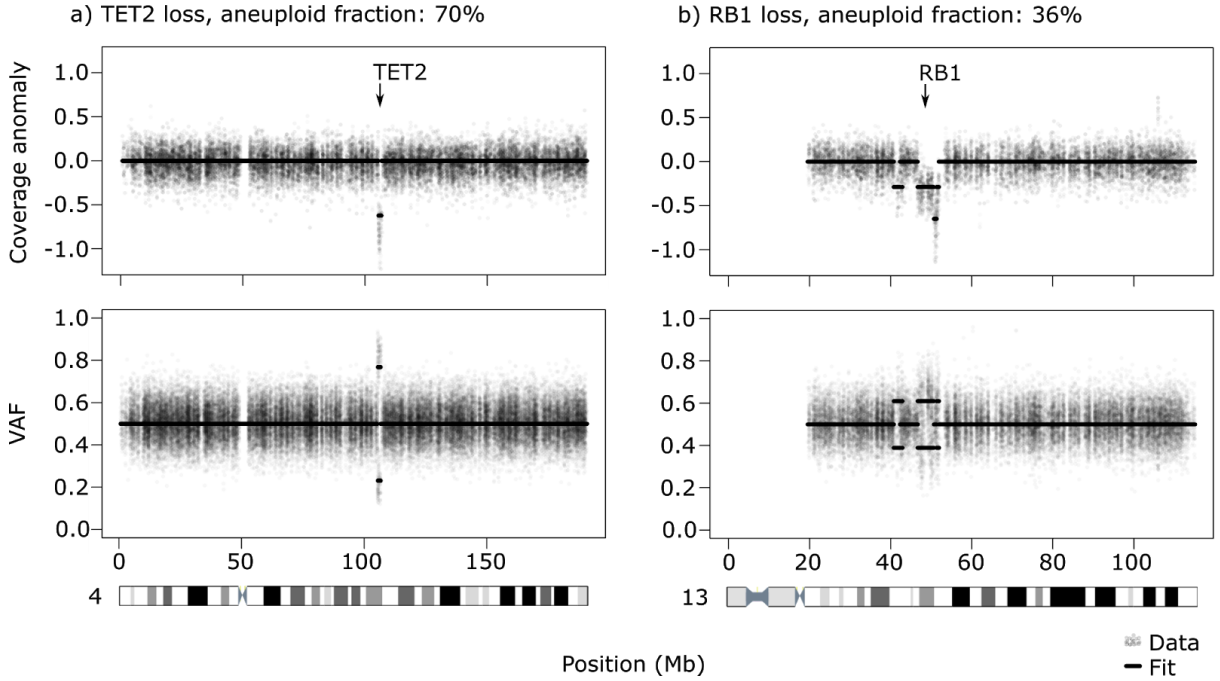
Supplementary Figure 3: SNP alternate allele frequencies compared between MGRB, gnomAD, and UK BioBank cohorts. Strand-specific biallelic SNPs in well-called regions and reported in the EBI GWAS catalogue only shown.



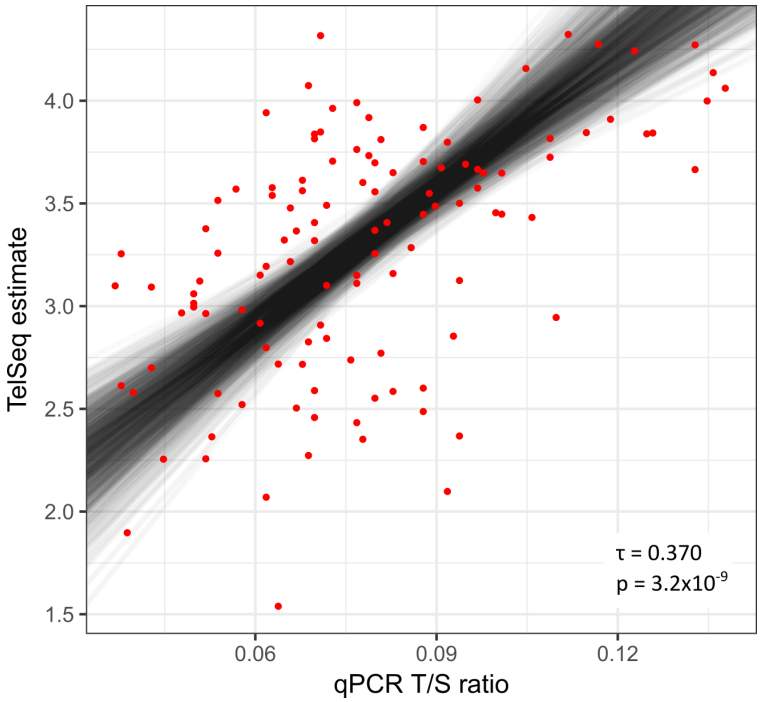
Supplementary Figure 4: Distribution of participant ages in the Australian Schizophrenia Research Bank (ASRB), gnomAD, MGRB, and UK BioBank (UKBB) cohorts. Ages were truncated at 100 years and binned into five year intervals except for terminal bins, which vary in size as shown.



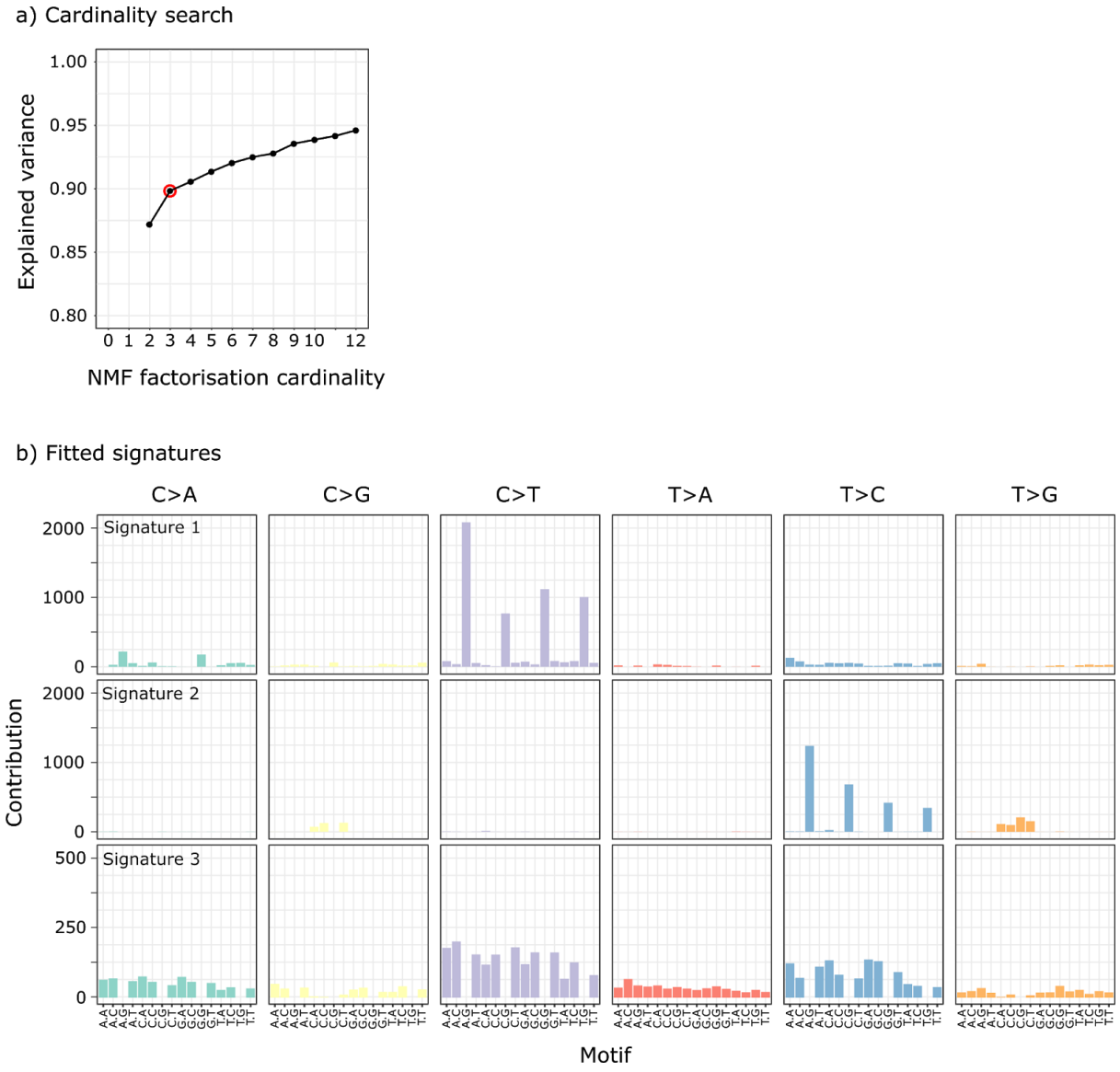
Supplementary Figure 5: Prediction of height in MGRB using a polygenic score (Wood et al., 2014). Each point represents the predicted and observed height of an MGRB individual; lines denote GCV-penalised generalised additive model thin plate spline fits.



Supplementary Figure 6: Examples of subclonal copy number variation observed in the MGRB. Figures show background-corrected coverage (top panels) and heterozygous variant allele frequency (bottom panels) as a function of genomic location. Individual locus measurements are represented by semi-transparent dots, with model fits indicated by horizontal segments. These samples demonstrated loss of a single copy of TET2 in an estimated 70% of nucleated blood cells (a), or loss of a single copy of RB1 in approximately 36% of cells (b). Coverage is background-corrected and on a $\log_2$ scale, with zero indicating diploidy.



Supplementary Figure 7: Comparison of Telseq WGS telomere length estimates to qPCR measurements. Points denote 119 randomly-selected samples from the MGRB and ASRB cohorts; one outlier with a Telseq estimate over 5 was excluded. Telseq estimates are directly as reported by the software; qPCR measurements are telomere / single copy gene copy ratios. Lines represent fits from 1000 bootstrap replicates of Deming regression using within-bootstrap median absolute deviation as an empirical variance estimate. The measures are significantly correlated (Kendall's $\tau = 0.370$, $p = 3.2 \times 10^{-9}$).



Supplementary Figure 8: Somatic variant motif factorization. A cardinality search on age-grouped samples indicated that a cardinality of 3 was appropriate, being the inflection point on the explained variance vs cardinality plot (a, selected cardinality marked with red circle). When the single-sample motif frequencies were factorized at this selected cardinality, the three signatures resulting were well resolved (b), with Signatures 1 and 3 respectively resembling Signatures 1 and 5 as previously reported (Alexandrov et al., 2015).

Supplementary Table 1: Rates of structural variation (SV) detected in the MGRB. Mean event counts per individual are given, with standard deviation in parentheses.

SV class	*Rate*	*(SD)*	*Fraction*
GRIDSS-reported			
Insertion	45	(7)	0.5%
Deletion	2750	(136)	33.1%
Indel	327	(21)	3.9%
Duplication	882	(98)	10.6%
Inversion	32	(5)	0.4%
Total GRIDSS	4036	(249)	48.6%
Mobster-reported			
L1 insertion	1072	(380)	12.9%
ALU insertion	2754	(246)	33.2%
SVA insertion	436	(92)	5.3%
HERV insertion	3	(2)	< 0.1%
Total Mobster	4264	(634)	51.4%
Grand total	8300	(675)	100%

Supplementary Table 2: Singleton and polymorphic rates for structural variants (SVs) identified in the MGRB. Structural variant count and fraction are given as a function of the number of samples identified to share that variant. Sample ranges are inclusive.

Samples with variant	*SV count (fraction)*	
1 (singletons)	155287	(17.1)
2 - 10	589771	(65.0)
11 - 100	139340	(15.3)
101 - 500	13669	(1.5)
501 - 1000	4623	(0.5)
1001 - 1500	2178	(0.2)
1501 - 2000	1367	(0.2)
2001 - 2500	1218	(0.1)
2501 - 2570	486	(0.1)

Supplementary Table 3: Structural variants identified in MGRB that may disrupt an ACMG incidentally-reportable gene.

Gene	*Variant*	*Predicted effect*
PCSK9	1:g.55494888_55509044del	Loss of 5' UTR and exon 1.
SMAD4	18:g.48556989_48573287del	Loss of 5' UTR.
TMEM43	3:g.141047610_14277101del	Deletion of entire *TMEM43* locus.
VHL	3:g.10067411_10421889inv	Inversion of entire *VHL* locus.

Supplementary Table 4: Clinical and demographic characteristics of the 45 and Up cancer cases, compared to the 45 and Up cancer-free individuals included in the MGRB. Cancer cases had some evidence of a cancer diagnosis prior to age 70, either by self report or admission and registry records; cancer-free individuals had no such evidence prior to age 70. Aggregate statistics are medians, with ranges in parentheses. As some individuals had multiple cancers, the sum of cancer types exceeds the number of cancer cases.

Measure	**Cancer cases**	**Cancer-free**
Individuals	269	717
(percent female)	(45.3%)	(59.3%)
Age at collection (years)	71	70
	(64 – 88)	(64 – 91)
Height (m)	1.70	1.66
	(1.47 – 1.96)	(1.37 – 1.91)
Mass (kg)	76.0	72.0
	(44.5 – 120.0)	(36.0 – 147.0)
Cancer type		
Prostate	74	——
Melanoma of skin	58	——
Colorectal	40	——
Breast	26	——
Non-melanoma skin	20	——
Lung	13	——
Bladder	10	——
Other	124	——

References

45 and Up Study Collaborators, Banks, E., Redman, S., Jorm, L., Armstrong, B., Bauman, A., Beard, J., Beral, V., Byles, J., Corbett, S., et al. (2008). Cohort profile: the 45 and up study. Int. J. Epidemiol. *37*, 941–947.

Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S.H., Kiemeney, L.A.L.M., Veltman, J.A., Gilissen, C., and Hoischen, A. (2017). Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. Am. J. Hum. Genet. *101*, 50–64.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Artandi, S.E., Chang, S., Lee, S.L., Alson, S., Gottlieb, G.J., Chin, L., and DePinho, R.A. (2000). Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature *406*, 641–645.

Barnett, I.J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genet. Epidemiol. *37*, 142–151.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 289–300.

Brooks-Wilson, A.R. (2013). Genetics of healthy aging and longevity. Hum. Genet. *132*, 1323–1338.

Brunet, J.-P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A. *101*, 4164–4169.

Buscarlet, M., Provost, S., Zada, Y.F., Barhdadi, A., Bourgoin, V., Lépine, G., Mollica, L., Szuber, N., Dubé, M.-P., and Busque, L. (2017). DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. Blood *130*, 753–762.

Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., and Papenfuss, A.T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res.

Cawthon, R.M. (2002). Telomere measurement by quantitative PCR. Nucleic Acids Res. *30*, e47.

Chainani, V., Shaharyar, S., Dave, K., Choksi, V., Ravindranathan, S., Hanno, R., Jamal, O., Abdo, A., and Abi Rafeh, N. (2016). Objective measures of the frailty syndrome (hand grip strength and gait speed) and cardiovascular mortality: A systematic review. Int. J. Cardiol. *215*, 487–493.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat. Biotechnol. *34*, 531–538.

Chin, L., Artandi, S.E., Shen, Q., Tam, A., Lee, S.L., Gottlieb, G.J., Greider, C.W., and DePinho, R.A. (1999). p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. Cell *97*, 527–538.

Conomos, M.P., and Thornton, T. (2016). GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R Package Version 2.

Deelen, J., Beekman, M., Uh, H.-W., Broer, L., Ayers, K.L., Tan, Q., Kamatani, Y., Bennet, A.M., Tamm, R., Trompet, S., et al. (2014). Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. Hum. Mol. Genet. *23*, 4420–4432.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

De Rosa, M., Pace, U., Rega, D., Costabile, V., Duraturo, F., Izzo, P., and Delrio, P. (2015). Genetics, diagnosis and management of colorectal cancer (Review). Oncol. Rep. *34*, 1087–1096.

Ding, Z., Mangino, M., Aviv, A., Spector, T., Durbin, R., and UK10K Consortium (2014). Estimating telomere length from whole genome sequence data. Nucleic Acids Res. *42*, e75.

Dudzińska-Griszek, J., Szuster, K., and Szewieczek, J. (2017). Grip strength as a frailty diagnostic component in geriatric inpatients. Clin. Interv. Aging *12*, 1151–1157.

Dumble, M., Moore, L., Chambers, S.M., Geiger, H., Van Zant, G., Goodell, M.A., and Donehower, L.A. (2007). The impact of altered p53 dosage on hematopoietic stem cell dynamics during aging. Blood *109*, 1736–1742.

Erikson, G.A., Bodian, D.L., Rueda, M., Molparia, B., Scott, E.R., Scott-Van Zeeland, A.A., Topol, S.E., Wineinger, N.E., Niederhuber, J.E., Topol, E.J., et al. (2016). Whole-Genome Sequencing of a Healthy Aging Cohort. Cell *165*, 1002–1011.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat. Neurosci. *19*, 1563–1565.

Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring

mutational signatures from single-nucleotide variants. Bioinformatics *31*, 3673–3675.

Gelsi-Boyer, V., Brecqueville, M., Devillier, R., Murati, A., Mozziconacci, M.-J., and Birnbaum, D. (2012). Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. J. Hematol. Oncol. *5*, 12.

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M.J., Engelborghs, S., De Deyn, P., Berr, C., et al. (2011). APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol. Psychiatry *16*, 903–907.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

de Haan, G., and Van Zant, G. (1999). Dynamic changes in mouse hematopoietic stem cell numbers during aging. Blood *93*, 3294–3301.

Hoffmann, T.J., Van Den Eeden, S.K., Sakoda, L.C., Jorgenson, E., Habel, L.A., Graff, R.E., Passarelli, M.N., Cario, C.L., Emami, N.C., Chao, C.R., et al. (2015). A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov. *5*, 878–891.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. Stat. Theory Appl. *6*, 65–70.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2016). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2. 0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. *19*, 249–255.

Katz, S., and Akpom, C.A. (1976). A measure of primary sociobiological functions. Int. J. Health Serv. *6*, 493–508.

Kennedy, S.R., Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. *9*, e1003794.

Lacaze, P., Winship, I., and McNeil, J. (2017). Penetrance and the Healthy Elderly. Genet. Test. Mol. Biomarkers *21*, 637–640.

Lacaze, P., Pinese, M., Kaplan, W., Stone, A., Brion, M.-J., Woods, R.L., McNamara, M., McNeil, J.J., Dinger, M.E., and Thomas, D.M. (2018). The Medical Genome Reference Bank: a whole-genome data resource of 4,000 healthy elderly individuals. Rationale and cohort design. Eur. J. Hum. Genet.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

Langsted, A., Nordestgaard, B.G., Benn, M., Tybjærg-Hansen, A., and Kamstrup, P.R. (2016). PCSK9 R46L loss-of-function mutation reduces Lipoprotein (a), LDL cholesterol, and risk of aortic valve stenosis. J. Clin. Endocrinol. Metab. *101*, 3281–3287.

Latorre-Pellicer, A., Moreno-Loshuertos, R., Lechuga-Vieco, A.V., Sánchez-Cabo, F., Torroja, C., Acín-Pérez, R., Calvo, E., Aix, E., González-Guerra, A., Logan, A., et al. (2016). Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. Nature *535*, 561–565.

Lee, H.W., Blasco, M.A., Gottlieb, G.J., Horner, J.W., 2nd, Greider, C.W., and DePinho, R.A. (1998). Essential role of mouse telomerase in highly proliferative organs. Nature *392*, 569–574.

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.

Li, D., Lewinger, J.P., Gauderman, W.J., Murcray, C.E., and Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet. Epidemiol. *35*, 790–799.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Loughland, C., Draganic, D., McCabe, K., Richards, J., Nasir, A., Allen, J., Catts, S., Jablensky, A., Henskens, F., Michie, P., et al. (2010). Australian Schizophrenia Research Bank: a database of comprehensive clinical, endophenotypic and genetic data for aetiological studies of schizophrenia. Aust. N. Z. J. Psychiatry *44*, 1029–1035.

Lowsky, D.J., Olshansky, S.J., Bhattacharya, J., and Goldman, D.P. (2014). Heterogeneity in healthy aging. J. Gerontol. A Biol. Sci. Med. Sci. *69*, 640–649.

Lu, A.T., Xue, L., Salfati, E.L., Chen, B.H., Ferrucci, L., Levy, D., Joehanes, R., Murabito, J.M., Kiel, D.P., Tsai, P.-C., et al. (2018). GWAS of epigenetic aging rates in blood reveals a critical role for TERT. Nat. Commun. *9*, 387.

Lubitz, S.A., Yin, X., Lin, H.J., Kolek, M., Smith, J.G., Trompet, S., Rienstra, M., Rost, N.S., Teixeira, P.L., Almgren, P., et al. (2017). Genetic Risk Prediction of Atrial Fibrillation. Circulation *135*, 1311–1320.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886.

Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet. *47*, 1393–1401.

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.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

McNeil, J.J., Woods, R.L., Nelson, M.R., Murray, A.M., Reid, C.M., Kirpach, B., Storey, E., Shah, R.C., Wolfe, R.S., Tonkin, A.M., et al. (2017). Baseline Characteristics of Participants in the ASPREE (ASPirin in Reducing Events in the Elderly) Study. J. Gerontol. A Biol. Sci. Med. Sci. *72*, 1586–1593.

Meynert, A.M., Ansari, M., FitzPatrick, D.R., and Taylor, M.S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics *15*, 247.

Nagpal, S., Gibson, G., and Marigorta, U.M. (2018). Pervasive Modulation of Obesity Risk by the Environment and Genomic Background. Genes *9*.

Percy, M.J., and McMullin, M.F. (2005). The V617F JAK2 mutation and the myeloproliferative disorders. Hematol. Oncol. *23*, 91–93.

Prince, M.J., Wu, F., Guo, Y., Gutierrez Robledo, L.M., O'Donnell, M., Sullivan, R., and Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. Lancet *385*, 549–562.

R Core Team (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria).

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405.

Russler-Germain, D.A., Spencer, D.H., Young, M.A., Lamprecht, T.L., Miller, C.A., Fulton, R., Meyer, M.R., Erdmann-Gilmore, P., Townsend, R.R., Wilson, R.K., et al. (2014). The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers. Cancer Cell *25*, 442–454.

Sahin, E., Colla, S., Liesa, M., Moslehi, J., Müller, F.L., Guo, M., Cooper, M., Kotton, D., Fabian, A.J., Walkey, C., et al. (2011). Telomere dysfunction induces metabolic and mitochondrial compromise. Nature *470*, 359–365.

Schumacher, F.R., Schmit, S.L., Jiao, S., Edlund, C.K., Wang, H., Zhang, B., Hsu, L., Huang, S.-C., Fischer, C.P., Harju, J.F., et al. (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat. Commun. *6*, 7138.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. *8*, 289–317.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

Syddall, H., Cooper, C., Martin, F., Briggs, R., and Aihie Sayer, A. (2003). Is grip strength a useful single marker of frailty? Age Ageing *32*, 650–656.

Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204.

Tange, O. (2011). Gnu parallel-the command-line power tool. The USENIX Magazine *36*, 42–47.

Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U. S. A. *113*, 11901–11906.

Teng, E.L., and Chui, H.C. (1987). The Modified Mini-Mental State (3MS) examination. J. Clin. Psychiatry *48*, 314–318.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Thung, D.T., de Ligt, J., Vissers, L.E.M., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E.P., Ye, K., Veltman, J.A., and Hehir-Kwa, J.Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. *15*, 488.

Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. *9*, 13.

Tobacco and Genetics Consortium (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat. Genet. *42*, 441–447.

Tom, J.A., Reeder, J., Forrest, W.F., Graham, R.R., Hunkapiller, J., Behrens, T.W., and Bhangale, T.R. (2017). Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinformatics *18*, 351.

Tyner, S.D., Venkatachalam, S., Choi, J., Jones, S., Ghebranious, N., Igelmann, H., Lu, X., Soron, G., Cooper, B., Brayton, C., et al. (2002). p53 mutant mice that display early ageing-associated phenotypes. Nature *415*, 45–53.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 11–10.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Wachsmuth, M., Hübner, A., Li, M., Madea, B., and Stoneking, M. (2016). Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number. PLoS Genet. *12*, e1005939.

Walsh, R., Thomson, K.L., Ware, J.S., Funke, B.H., Woodley, J., McGuire, K.J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J.C., et al. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet. Med. *19*, 192–203.

Williams, G.C. (1957). Pleiotropy, natural selection, and the evolution of senescence. Evolution *11*, 398–411.

Wood, S.N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. J. Am. Stat. Assoc. *99*, 673–686.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J. 'an, Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173.

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.

Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. *7*, 12484.

von Zglinicki, T., and Martin-Ruiz, C.M. (2005). Telomeres as biomarkers for ageing and age-related diseases. Curr. Mol. Med. *5*, 197–203.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

Zhou, Y.-J., Wang, Y., and Chen, L.-L. (2016). Detecting the Common and Individual Effects of Rare Variants on Quantitative Traits by Using Extreme Phenotype Sampling. Genes *7*.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U. S. A. *109*, 1193–1198.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. *111*, E455–E464.