The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly

Mark Pinese et al.#

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 first release of the Medical Genome Reference Bank (MGRB), comprising whole genome sequence and phenotype of 2570 elderly Australians depleted for cancer, cardiovascular disease, and dementia. We analyse the MGRB for single-nucleotide, indel and structural variation in the nuclear and mitochondrial genomes. MGRB individuals have fewer disease-associated common and rare germline variants, relative to both cancer cases and the gnomAD and UK Biobank cohorts, consistent with risk depletion. Age-related somatic changes are 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.

#A full list of authors and their affiliations appears at the end of the paper.
Most developed nations face crises in health care associated with population ageing. Healthy ageing is a complex phenotype, influenced by both environmental and genetic factors. 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.

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. 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, common variation at only a relatively small number of loci has been consistently associated with lifespan or parental longevity. Rare pathogenic variants that hasten the onset of common disease 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, no evidence was found for depletion of rare genetic 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 haemato poiesis is known to correlate with both age and susceptibility to disease. 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. The missing heritability arguably resides in the total burden of both common and rare variation, structural variation untagged by simple polymorphisms, and their interactions. WGS enables more comprehensive characterisation 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 maximises statistical power by comparing the extremes of phenotypes of interest. 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 are undertaking WGS of over 4000 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). Here we describe the first release of this resource, comprising WGS data and phenotype for 2570 well elderly individuals. For comparison, we also perform WGS of 344 young subjects, and 273 elderly individuals with cancer. We subject these cohorts to a broad spectrum, systematic analysis of germline and somatic variation within the nuclear and mitochondrial genomes, which we link to both chronicologic age as well as frailty measures.

### Results

**Cohort characteristics and sequencing.** The MGRB consists of over 4000 individuals from the ASPREE study, and Sax Institute’s 45 and Up Study, 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. We sequenced blood of 2926 MGRB individuals 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, 2570 samples remained, forming the first release of the MGRB cohort genomic data (Table 1).

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. MGRB participants were primarily of non-Finnish European genetic background (Table 1, Supplementary Fig. 1). Consistent with previous studies, 51.8% of small variants were singletons and 4.6% of loci were multi-allelic.

In addition to small scale variants, an average of 4036 structural variants (SVs) per individual were observed, most commonly deletions (Supplementary Table 1, Supplementary Fig. 2). In contrast to small variants, only 17% of SVs were unique (Supplementary Table 2). Each individual carried an average of 4264 mobile element insertions (MEI), predominantly of the ALU and L1 classes, and most MEIs were copy number polymorphisms at known loci. However, on average 1535 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.

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. 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. To this end, we identified pathogenic variants that are considered clinically reportable as incidental findings under current American College of Medical Genetics
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 (Fig. 1a, odds ratio 0.060, 95% confidence interval 0.0065–0.27, two-sided p < 0.001, n = 986, 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. 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 (Fig. 1a, 1 of 717 MGRB with pathogenic variants, vs 2 of 40 cancer cases, odds ratio 0.027, 95% CI 0.001–0.53, two-sided p = 0.008, n = 757, Fisher’s exact test).

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 Fig. 3), the MGRB cohort was significantly depleted for alleles specifically associated with risk of cancer, cardiovascular disease, and neurodegenerative disease (Supplementary Data 2, sheet 3, 698 loci, vs gnomAD odds ratio 0.38, 95% CI 0.27–0.52, p = 2.6 × 10^{-9}; vs UKBB odds ratio 0.47, 95% CI 0.34–0.64, two-sided p = 1.1 × 10^{-6}; Fisher’s exact test). This enrichment of protective alleles was specific to the clinical phenotypes excluded from MGRB (cancer, cardiovascular disease, and dementia, see Methods), 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 between MGRB, the gnomAD non-Finnish European reference cohort, the UK Biobank (UKBB), and the 45 and Up Study cancer cohort. Since the MGRB is an Australian cohort in which allele frequencies differ slightly from the white British sample represented by the UKBB, we controlled for expected subtle differences in allele frequencies genome-wide by simulating the reference population as a derivative from the UKBB that differs due to drift by an amount typical of the MGRB-UKBB comparison. Similar comparisons were made for the gnomAD and 45 and Up comparison cohorts, and 100,000 bootstraps were performed (Supplementary Fig. 4). We observed significant depletion of PS in MGRB for 7 of the 12 scores tested (Fig. 1b, Supplementary Data 2, Sheet 9). Notably, a PS associated with short parental lifespan was significantly depleted in MGRB relative to UKBB, consistent with the MGRB healthy elderly phenotype (bootstrap test, 0/100,000 rounds with a test statistic more extreme than observed, p < 0.001 after Holm correction for multiple tests). MGRB individuals were also depleted for prostate cancer risk relative to both UKBB (bootstrap test, 146/100,000 rounds more extreme than observed, Holm-corrected p = 0.055) and prostate cancer cases (two-sample t-test, n = 71 cases, n = 278 MGRB gender- and cohort-matched controls, t = 3.96, df = 126, two-sided p < 0.001), indicating that MGRB is an extreme depletion cohort for prostate cancer polygenic risk (Fig. 1c). 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 UKBB (Fig. 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; height PS was significantly depleted in the MGRB relative to UKBB, consistent with the MGRB healthy elderly phenotype (bootstrap test, 0/100,000 rounds with a test statistic more extreme than observed, p < 0.001 after Holm correction for multiple tests), the MGRB cohort was approximately 4.5 cm taller than UKBB (Fig. 5). We then compared the distribution of PS for prostate, colorectal, and melanoma skin cancer between the 45 and Up

### Table 2 Counts of clinically significant small variation in the MGRB for all genes in the ACMG SF 2.0 set.

| 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        |
|                   | MYBPC3     | 2        |
|                   | MYL3       | 1        |
| Hypercholesterolaemia | APOB     | 5        |
| Long QT, VA        | KCNH2      | 1        |
|                    | SCN5A      | 1        |
| Marfan syndrome    | MYH7I      | 1        |
| Malignant hyperthermia | RYR1    | 1        |
| Total              |            | 28       |

ARVC arhythmogenic right ventricular cardiomyopathy, CPVT catecholaminergic polymorphic ventricular tachycardia, HCM hypertrophic cardiomyopathy, DCM dilated cardiomyopathy, VA ventricular arrhythmia.

(ACMG) guidelines. Forty pathogenic or likely pathogenic heterozygous small variants were identified, with 28/2570 (1.1%) individuals carrying dominantly acting variants linked to disease (Table 2, Supplementary Data 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; 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. 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.
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 (as above, \( n = 71 \) cases, \( n = 278 \) MGRB gender- and cohort-matched controls, \( t = 3.96, \text{df} = 126, \text{two-sided} \ p < 0.001, \) Fig. 2a) and colorectal cancer (two-sample \( t\)-test, \( n = 41 \) cases, \( n = 690 \) MGRB controls, \( t = 2.46, \text{df} = 44.7, \text{two-sided} \ p = 0.018, \) Fig. 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%.

**Fig. 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 the difference in PS between MGRB and the reference cohorts (UKBB or gnomAD); higher values indicate a higher polygenic score in UKBB or gnomAD. q-Values represent false discovery rate estimates by the Benjamini–Hochberg method70. 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 UKBB, demonstrated here for prostate cancer. AU arbitrary units.
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, as evidenced by the presence of detectable somatic SNVs in blood DNA. Previous studies have largely used deep whole-exome or targeted sequencing to identify CHIP, but these methods lack sensitivity to detect the copy number variation (CNV) commonly observed in myelodysplasia and leukaemia. Low-depth WGS is a powerful tool for detecting SVs that has been applied to the identification of CHIP-associated SNVs, but not CNV. Here we estimate the burden of cancer-associated somatic variation, at both SNV and CNV level, in peripheral blood DNA using whole-genome data in the MGRB cohort.

In total, 184/2570 (7.2%) of MGRB individuals displayed evidence of CHIP, with SNVs associated with overgrowth and neoplasia observed in more than 10% of reads (Supplementary Data 3). Predominantly nonsense mutations (96%), these variants were most commonly seen in TET2 (47 individuals), DNMT3A (23), or ASXL1 (11) (Supplementary Fig. 6). 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) and a putative loss-of-function variant in TP53, C275Y (1). JAK2 V617F is a recognised driver of myeloproliferative disorders, which are also associated with ASXL1 loss, and TET2 and DNMT3A loss-of-function variants are frequent in CHIP. In total, the blood of 91/2570 (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 CNV. Of 2570 MGRB individuals, 1975 were successfully fitted 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 Fig. 7). In total, 9.2% (95% CI 7.9–10.5%) of MGRB samples demonstrated evidence of CHIP by either SNV or CNV, consistent with results from deep WES.

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 linked to grip strength. As well as CHIP, ageing is associated with telomere shortening, somatic Y chromosome loss, decreased mitochondrial copy number, and increased mitochondrial heteroplasmy. 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, and CHIP. Using standard-depth WGS data from multiple cohorts, consistent patterns of change with age were observed across all six

**Fig. 2 Polygenic risk is strongly related to cancer diagnosis.** 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. Fits are from logistic regression using a GCV-penalised thin plate spline smooth; bands denote 95% confidence intervals around the mean.
Compared to a population of younger individuals (the ASRB cohort, median age 40; Supplementary Fig. 8), 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

Fig. 3 Age-related somatic changes are associated with 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. Sample numbers were 1853 for the ASPREE cohort, 717 for the 45 and Up Study, and 344 for the ASRB cohort.

somatic metrics (Fig. 3a–f). Compared to a population of younger individuals (the ASRB cohort, median age 40; Supplementary Fig. 8), 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 will grow in time to contain sequence and phenotype data for deeply phenotyped aged individuals. This report describes the stage bootstrap test, which found that grip strength was positively correlated with the count function even when age is taken into account. Intriguingly, we investigated whether any somatic measures were associated with physical function in each cohort, with 95% Wald confidence intervals. Values significantly different from zero are represented in bold. Note that somatic burden and mitochondrial copy number are reported on the natural logarithm scale. ND, not determined due to data use agreement constraints.

Table 3 The rates of somatic measure change with age are different between middle-aged and old individuals.

| Measure | ASRB | 45 and Up Study | ASPREE |
|---------|------|----------------|--------|
| Individuals | 344 | 717 | 1853 |
| Percent female | ND | 59.3% | 48.2% |
| Median age (range) | 40 (18-65) | 70 (64-1) | 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 $[-0.177, 0.278]$ | $1.665 [1.315, 2.015]$ | $0.893 [0.195, 1.591]$ |

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 copy number per nucleus are reported on the natural logarithm scale. ND, not determined due to data use agreement constraints.

(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.

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. As expected, grip strength and gait speed both consistently decreased with age in both genders (Fig. 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 bootstrap test, first stage $p = 0.051$, validation $p = 0.036$; Supplementary Table 5).

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 (Fig. 3i).

### 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 programmes, 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 datasets of healthy controls will be essential to identify genetic variation unrelated to disease. To this end we created the MGRB, a WGS resource of deeply phenotyped aged individuals. This report describes the first data release of the MGRB comprising 2570 individuals; the MGRB will grow in time to contain sequence and phenotype data of 4000 healthy elderly Australians.

Although depletion of some common disease-related alleles has been reported in the healthy aged, the MGRB reveals a general 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 casecontrol genetic studies.

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 in the healthy elderly. 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 SVs were identified that may abrogate function of clinically reportable genes. Future studies using whole-genome-based data will benefit from the MGRB in quantitating the contribution of structural variation to ageing and disease.

The rate of pathogenic variation observed in the MGRB appeared lower than that previously reported in European cohorts that were not selected on the basis of healthy ageing. However, the comparison of rare variation rates between cohorts is affected to a greater extent than common variants by interplatform variation and subpopulation stratification. Further study will be required to definitively test whether the well elderly have a significantly lower burden of pathogenic rare variation to the general population.

The ageing process is accompanied by the emergence of somatic mutations in tissues other than blood, mitochondrial depletions and heteroplasmy, and progressive telomere shortening. We developed a suite of methods to detect these age-related changes using 30X WGS, and applied it to the elderly MGRB and a younger cohort. Telomere shortening itself may directly increase the likelihood of neoplasia, while oncogenic mutations in genes such as TP53 may rescue the effects of telomere loss. Telomere dysfunction has been associated with impaired mitochondrial function, linking these genomic features of ageing. Many of the somatic changes of ageing observed in the MGRB are consistent with marrow stem cell depletion, as observed in telomerase-deficient mice.

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. It is even possible that neoplastic events, such as telomerase stabilisation, loss of tumour...
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, comes from mice carrying a hyper-morphic form of Trp53, in which protection from neoplasia was accompanied by accelerated hematopoietic ageing and diminished marrow reserve31. If true, these findings suggest that strategies that suppress tumour 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 leucocyte 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 males28. We note that our power to detect such an effect is low when using a wellelderly cohort, but believe there will be great interest in deriving quantitative measures of biological ageing from standard-depth WGS.

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. 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, 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.

Methods

Experimental model and subjects. Participants of the MGRB were consented through the biobank programmes of the ASPREE and 45 and Up studies11,13. 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 the 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 diagnosis of dementia or a Modified Mini Mental State Examination (3MSE) score ≤ 77, and no severe difficulty or an inability to perform any one of the 6 Katz basic activities of daily living.

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 were required to correctly self-complete a health survey at enrolment, and therefore there will be great interest in deriving quantitative measures of biological ageing from standard-depth WGS.

Sample collection and processing. For ASPREE participants of the MGRB, peripheral blood samples were processed to buffy coat within 4 h 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 h 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. WGS 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. Each sample was sequenced on one HiSeq X lane.

Sequence alignment and processing. All sequence data generated at the KCCG were processed following the Genome Analysis Toolkit (GATK) best practices13. 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, with an added contig of NC_001422.1 to act as a decoy for the HiSeq-specific 9X174 sequence spike-in. Reads were aligned to this reference using bwa 0.7.15 mem in paired mode, and duplicates marked with biobambam 2.0.65 bamsormadup, with a minimum optical pixel distance of 2500. 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 Vcf files from all single-lane realigned and recalibrated Bams using recommended parameters. Pipeline steps were accelerated using GNU parallel 20170722 (ref. 44).

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. Typically 15 reads are required for high genotyping performance45; 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 and 11 bp, respectively, to yield a final definition of regions of unusual depth in the MGRB cohort. These regions totaled 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 within the poor-quality subset of the genome de
ded by RepeatMasker, 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 (ref. 47) 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 4).

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 3033 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. A total of 2994/3033 (95.7%) samples passed initial quality thresholds (Table 5). 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, 2890/3033 (95.3%) MGRB and 45 and Up cancer samples passed initial quality control (QC).

Small variant genotyping and final QC. The 2890 MGRB and 45 and Up cancer samples passing initial QC were joint called in a single batch using GATK GenotypeGVCFs, and quality metrics calculated within Hail v0.1 (ref. 48). A total of 2994/3033 (95.7%) samples passed initial quality control thresholds (Table 5). 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, 2890/3033 (95.3%) MGRB and 45 and Up cancer samples passed initial quality control (QC).

Table 4 Quantity in megabases of genomic regions in each locus confidence tier.

| Locus confidence tier | Reference genome | Canonical chromosomes |
|-----------------------|------------------|------------------------|
| 1—highest             | 1212             | 1212                   |
| 2                     | 52               | 52                     |
| 3—lowest              | 1874             | 1832                   |
| Total                 | 3137             | 3096                   |

Table 5 Quality metric conditions for samples to pass quality control.

| 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                |

Two rounds of quality control (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.

ASRB processing and quality control. Sequence processing and quality control for the ASRB cohort proceeded as described for the MGRB. 344/476 ASRB samples passed all QC thresholds and were used for subsequent analysis.

Cohort population structure. The MGRB cohort population structure was determined using principal components analysis (PCA), with reference to the 1000 genomes (1000 G) populations. A merged dataset of all MGRB and 45 and Up cancer genotypes and the 1000G Phase 3 genotypes 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 5–95%, and no evidence of deviation from Hardy–Weinberg equilibrium within any of 17 homogeneous 1000G populations (PHWEx > 0.01/17 for each of population codes BEB, CDX, CEU, CTR, CHS, FIN, GRB, GWD, IBS, ITU, JPT, HUD, HKI, LWK, MSL, STU, TSI, and YRI). Merged variants were LD-pruned in Hail with an r² threshold of 0.1, and PCA performed in Hail on biallelic variants with a combined MGRB and 1000G allele frequency in the range of 5–95%.

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, GRB, 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, and then were applied to predict population source for the MGRB samples. All models were implemented using mlcst v5.3 (ref. 49).

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. Variants were further annotated with 1000 genomes Phase 3, Haplotype reference consortium, GnomAD, and dbSNP allele frequencies, as well as ClinVar, CATO, and Eigen annotations (see Supplementary Data 4 for resource versions).

Germline SV detection. Germline SVs in the MGRB and 45 and Up cancer cohorts were detected using GRIDSS v1.4.1 (ref. 50), excluding regions in the Encode DAC Mappability Consensus Excludable list (Supplementary Data 4). 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, SV 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 MEIs were identified using Mobster v0.2.2 (ref. 51) 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.
Genome-wide common variant frequency comparison. To compare patterns of common variation between the MGRB and other cohorts, we merged the MGRB variant frequencies with the non-Finnish European (NFE) WGS allele frequencies, and allele frequencies from a subset of the UK Biobank genotype set, consistent of participants who self-identified as white British and who had confirmed European ancestry by PCA. To minimise the influence of technical artefacts, variants were restricted to strand-specific biallelic SNPs listed in the EBI GWAS database, 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. In total, 21,033 SNPs remained following this filtering, with very similar allele frequencies across all cohorts (Supplementary Data 2, sheet 1; Supplementary Fig. 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 parameters: the proportion of subjects who self-identified as white and who had confirmed European ancestry by PCA, and the allele frequency difference between the MGRB and gnomAD or UKBB that is specific to the phenotype in question. 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 2, sheets 2–3), and two negative control sets of variants linked to anthropometric traits (Supplementary Data 2, sheets 4–5), or behavioural traits (Supplementary Data 2, sheets 6–7).

PS estimation and testing. PSs were calculated as $S_i = \sum_j \beta_i \delta_j$ where $S_i$ is the PS for individual $i$; $\beta_i$ the GWAS-reported coefficient for a single variant allele at locus $j$, and $\delta_j$ 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 $\delta_j = 2f_j$ 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 PS calculation in all individuals.

PS GWAS coefficients were derived from each source publication by the following procedure. Firstly, all loci that were reported to be genome-wide significant in replication were selected, along with the effect allele’s association coefficient in the phenotype cohort. If replication results were not reported, the derivation cohort loci and coefficients were used. Repression coefficients for binary traits were then transformed to a log-odds scale; coefficients for the continuous traits of height and parental lifespan were used as-is. Loci were converted to GRCh37-centric coordinates, and where possible any loci falling within tier 3 regions (see section Locust confidence tiers) were imputed to an alternative locus falling in a tier 1 or 2 region, chosen so that the original and alternative loci had an $R^2$ of at least 0.9 as assessed in the GBR population of 1000 genomes by NCI LDLink. If no alternative loci fulfilled this criterion, the original locus was dropped from the PS. PSs were derived by this procedure for colorectal cancer (Supplementary Data 2, sheets 4–5), breast cancer (Supplementary Data 2, sheets 6–7), coronary artery disease (Supplementary Data 2, sheets 8–9), atrial fibrillation (Supplementary Data 2, sheets 10–11), attention deficit hyperactivity disorder (Supplementary Data 2, sheets 12–13), blood pressure (Supplementary Data 2, sheets 14–15), early-onset Alzheimer’s disease (Supplementary Data 2, sheets 16–17), early-onset coronary artery disease (Supplementary Data 2, sheets 19–20), melanoma (Supplementary Data 2, sheets 21–22), breast cancer (Supplementary Data 2, sheets 23–24), prostate cancer (Supplementary Data 2, sheets 25–26), blood pressure (Supplementary Data 2, sheets 27–28), blood pressure (Supplementary Data 2, sheets 29–30), and coronary artery disease (Supplementary Data 2, sheets 31–32).

The association between PS and risk of specific cancers was assessed by logistic regression, with the effect of PS 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 PSs defined.

Incidental somatic variant detection. Somatic variants were identified in post-BQSR BAM files using FreeBayes, with options: --binned-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 REseq genes in the COSMIC Cancer Gene Census downloaded 11 December 2017. Variant annotation was performed using the Ensembl VEP release 98, 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 T/S ratio greater than 3 in any cohort. 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 Catalogue with a TSG annotation.

**Telomere length.** Telomere lengths were estimated using Telseq v0.0.1 (ref. 66). 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 Fig. 10). Briefly, oligo-dT primers were added to gDNA (5 ng), 2x Rotor-Gene SYBR Green Master Mix (Qiagen), 500 nM Tel forward [5′-CGGTTT(GTTTG)(GT), 500 nM Tel reverse [5′-GGTCCTG(CTTC)3]-CCT] or 300 nM 36B4 forward [5′-CAGCAATGGGAAGGTGTAATCC]-36B4 reverse [5′-CCACTTCTATCTACCAAGCTGACA]-35 cycles of 95 °C for 1 s; 56 °C for 10 s (telomere reaction) followed by 30 cycles of 95 °C for 1 s, 58 °C for 10 s (single copy gene reaction). The experiment was accepted if the reference sample T/S ratio was $\geq 1.34$ (ref. 63), $R^2 > 0.95$ and the distribution of $\Delta$Ct from different samples was consistent with individual batch effects. Telomere length estimation by Telseq was validated by qPCR on a subset of 120 samples from the ASRB and MGRB cohorts (Supplementary Fig. 10). 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.
Mitochondria and Y chromosome copy number. Mean mitochondrial genome copy number in each sample was estimated using read counts, as $2 \times (R_N \times S + \Sigma R_i) / (S + \Sigma S_i)$, where $R_N$ and $S_N$ denote the number of reads mapping to contig set $Z$ and the total size of contig set $Z$, and $M$ and $A$ denote mitochondrial and autosomal contigs, respectively. Read counts were mapped and aligned reads reported by samtools idxtats, and were not corrected for read duplication. Patch contigs were not included in counts. In this way $\sum R_i$ was estimated by an analogous procedure, as $2 \times (R_N \times S + \Sigma R_i) / (S + \Sigma S_i)$.

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 $q = -10 \log_{10}[(1 - P(n, N)) / (n + P(n, N))]$, where $P$ is the count of alternate allele reads, $N$ the total depth at the variant locus, $p = 0.0025$ a fixed error rate estimate, and $P(n, p, N)$ 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 E and MT and A denote mitochondrial and autosomal contigs, respectively. The burden of these variants in each sample was then dimensionally reduced by spectral factorisation, and per-sample signature scores used as the final mitochondrial variant burden.

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 factorisation, 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 condition:

$$c_l \leq n_l \leq r_l,$$

where $n_l$ is the number of non-reference allele reads at the locus, and $c_l$ and $r_l$ are integers which maximise:

$$P_e = r_k \sum_{n_i=r_l}^{c_l} \binom{n_i}{n_l} p_e^{n_l} (1 - p_e)^{n_i-n_l}$$

subject to:

$$r_k = r_l \frac{N_{\text{read}}}{f_{\text{read}}} \sum_{n_i=r_l}^{c_l} \binom{n_i}{n_l} p_e^{n_l} (1 - p_e)^{n_i-n_l} \geq \epsilon,$$

with $r_l = \frac{1}{2}, \frac{1}{2} \sqrt{2}, \frac{1}{2} \sqrt{2}, \frac{1}{2}$, and $p_e = \left(1 - \frac{x + y}{2}\right)^{1 + x + y}$. Here $x$ is the sum of reference and alternate allele depths at the locus, $r_l$ is the expected rate of heterozygous variant germline loci, $r_l$ the expected rate of somatic variant loci, $e$ the base read error rate, $g_e$ the minimum acceptable ratio of true positive calls to false positive, and $f$ is the expected somatic variant allele fraction, $n_l$ and $a_l$ are test sizes corresponding to thresholds $c_l$ and $c_l$, $n_l$ is defined as $\sum_{i=r_l}^{c_l} \binom{n_i}{n_l} p_e^{n_l} (1 - p_e)^{n_i-n_l}$. In the context of this discussion, $p_e$ is the fraction of cells with deletion, $p_r$ is the fraction of cells with addition, and $p_l$ is the fraction of cells with loss.

Somatic copy number variants. We developed a model-based strategy to identify somatic copy number variant alleles, assuming a single genetically homogeneous subclone present on a background of diploid cells.

We applied the above procedure to the MGRB and ASRB samples, with parameters adapted to maximise sensitivity with low-depth sequencing data: $g_e = 5, f = 0.2$, $r_e = 0.0025$, $r_l = 0.0025$. Our filtering protocol employed SNVs identified by samtools mpileup, with maximum depth 101, mapping quality adjustment of 50, BAQ recalibration, 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 factorisation cardinality procedure applied to our data indicated that three signatures best described the mutation patterns observed (Supplementary Fig. 11). Signature 3 in this work was quantitatively similar to COSMIC signature 1 (cosine similarity 0.81), previously reported to be associated with age at cancer diagnosis29, 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 3 (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.

---

**Nature Communications** | https://doi.org/10.1038/s41467-019-14079-0 | ARTICLE
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 length 100 kb, and merged consecutive segments of the genome if they were assigned identical ploidy state. The above approach was used to segment all the genomes of the MGRB cohort into regions of differing ploidy state using the same algorithm.

To evaluate whether these segments were biologically meaningful, we performed an analysis of segment length. As whole-genome doubling, a focal length 100 kb, and mitochondrial variant count were long-term current pseudotransformed to modelling; and grip strength in kg was power transformed with exponent 0.7.

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 variant count, and mitochondrial variant count were long-term current pseudotransformed 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.

We used a permutation procedure to test the importance of somatic burden measures in predicting grip strength and grip speed, conditioned on age. For each of 18 possible frailty measure × somatic measure × sex combinations (frailty measures: age, age + (weight + s(BMI) + s(abdomino) + s(somatic), with age in years, weight in kg, BMI in kg/m^2, abdominal circumference (abdomino) in cm, and the somatic measure of interest (transformed if relevant following exploratory analysis)). In this model specification, (age) denotes a GCV-penalised thin plate spline smooth term in x as implemented in R package mgcv 1.8-17 (ref. 68), with Gaussian error and identity link. This model’s deviance p was compared to the deviance d_0 of 10,000 models fit in the same manner but with the somatic variable permuted, and the p value estimated as \( p = \frac{1}{B} \sum \left( d_i \leq d_0 + 0.5 \right) \). 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 vs age. To remove these solely for the purposes of illustration (Fig. 3), for each somatic measure we fit the generalised additive model measure \( \sim \mathcal{N} \left( \text{age}, \text{age} + \text{sex} \right) + \text{cohort} \) with Gaussian error and identity link. In this model specification (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 68. After confirming the suitability of the model fits, cohort-specific effects were removed by calculating the quantity \( y_{i} = y_{i} - \hat{\delta}_{i} + \hat{\gamma}_{i} \) for each individual and measure, where \( y_{i} \) is the cohort-adjusted somatic measure for individual i to be plotted; \( \hat{\delta}_{i} \) is the original measurement for individual i in cohort C and \( \hat{\gamma}_{i} \) is the estimated 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-individual quantile of mitochondrial DNA copy number \( q_{i} \) was defined as \( q_{i} = F_{i}^{-1}(\cdot) \), where \( F_{i} \) 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 \pm 1\text{ 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 79 years, this neighbourhood contained 35 men with ages in the range 78–80 years. Given these local mtDNA copy number quantile estimates \( q_{i} \), a generalised additive model of the form grip strength–age + s(q) was fit using the R package mgcv 68, 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.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Summary variant frequency data for the MGRB cohort are available at the web portal: https://gsc.genavn.org.au/explore. Raw genomic data have been deposited at the European Genome-Phenome Archive under study ID EGAS00001003511. Phenotype data are available upon application to the MGRB Data Access Committee at mgb@garnav.org.au.

Code availability

Source code for all analyses is available at https://github.com/mpine/segrep-manuscript; source code for the somatic SNV and LoH detection tools can be found at https://github.com/mpine/soma-snv and https://github.com/mpine/soma-cnv.

Received: 21 July 2019; Accepted: 13 December 2019; Published online: 23 January 2020

References

1. Timmers, P. R. et al. Genomics of 1 million human lifespans implicates novel pathways and common diseases and distinguishes survival chances. Elife, e39856 (2019).
2. Pilling, L. C. et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. J. Gerontol. A Biol. Sci. Med. Sci., 72, 504–520 (2017).
3. Deelen, J. et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. Nat. Commun. 10, 3669 (2019).
4. Erikson, G. A. et al. Whole-genome sequencing of a healthy aging cohort. Cell 165, 1002–1011 (2016).
5. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487 (2014).
6. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. 371, 2488–2498 (2014).
7. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. 101, 5–22 (2017).
8. Zak, O. et al. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA 111, E453–E464 (2014).
9. Zak, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA 109, 1193–1198 (2012).
10. Li, D., Lewinger, P. J., Gauderman, W. J., Murray, C. E. & Conti, D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet. Epidemiol. 35, 790–799 (2011).
11. Lacaze, P. The Medical Genome Reference Bank: a whole-genome data resource of 4,000 healthy elderly individuals. Rationale and cohort design. Eur. J. Hum. Genet. 27, 308–316 (2018).
12. McNeil, J. J. et al. Baseline characteristics of participants in the ASPREE (ASPrin in Reducing Events in the Elderly) Study. J. Gerontol. A Biol. Sci. Med. Sci. 72, 1586–1593 (2017).
13. 45 and Up Study Investigators et al. Cohort profile: the 45 and up study. Int. J. Epidemiol. 37, 941–947 (2008).
14. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. 32, 246–251 (2014).
15. Telenti, A. et al. Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. USA 113, 11901–11906 (2016).
16. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
17. Walsh, R. et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet. Med. 19, 192–203 (2017).
18. Kalai, S. S. et al. 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 (2017).
19. Chen, R. et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat. Biotechnol. 34, 531–538 (2016).
20. Langsted, A., Nordestgaard, B. G., Benn, M., Tybjærg-Hansen, A. & Kamstrup, P. R. PCSK9 R46L loss-of-function mutation reduces Lipoprotein
Russler-Germain, D. A. et al. The R882H DNMT3A mutation associated with
van den Akker, E. B. et al. Uncompromised 10-year survival of oldest old carrying
NATURE COMMUNICATIONS | (2020) 11:435 | https://doi.org/10.1038/s41467-019-14079-0 | www.nature.com/naturecommunications
Wood, A. R. et al. De
Van der Auwera, G. A. et al. From FastQ data to high-con
Loughland, C. et al. Australian Schizophrenia Research Bank: a database of
dynamics during aging.
Sahin, E. et al. Telomere dysfunction induces non-reciprocal translocations and epithelial cancers in mice. Nature 406, 641–645 (2000).
Cherayil, L. et al. p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. Cell 97, 527–538 (1999).
Sahin, E. et al. Telomere dysfunction induces metabolic and mitochondrial compromise. Nature 470, 559–365 (2011).
Lee, H. W. et al. Essential role of mouse telomerase in highly proliferative organs. Nature 392, 569–574 (1998).
Dumble, M. et al. The impact of altered p53 dosage on hematopoietic stem cell dynamics during aging. Blood 109, 1736–1742 (2007).
Loughland, C. et al. 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 (2010).
Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calling: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013).
Tange, O. Gnu parallel-the command-line power tool. USENIX Magazine 36, 42–47 (2011).
Muertter, A. M., Ansari, M., Fitpatrick, D. R. & Taylor, M. S. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15, 247 (2014).
Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat. Neurosci. 19, 1563–1565 (2016).
Commomos, M. P., Miller, M. B., & Thornton, T. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. 39, 276–293 (2015).
Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328 (2012).
Sobocinska, E. B., Murphy, T. B. & Raftery, A. E. nullS: clustering, classification and density estimation using Gaussian finite mixture models. R. J. 8, 289–317 (2016).
Cameron, D. L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 27, 2050–2060 (2017).
Thung, D. T. et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 15, 488 (2014).
Nagpal, S., Gibson, G. & Marigorta, U. M. Pervasive modulation of obesity risk by the environment and genomic background. Genes 9, E411 (2018).
Machida, J. M. & Chabner, B. A. Uncompromised 10-year survival of oldest old carrying somatic mutations in DNM13A and TET2. Blood 127, 1512–1515 (2016).
Russler-Germain, D. A. et al. The R882H DNM13A mutation associated with AML dominantly inhibits wild-type DNM13A by blocking its ability to form active tetramers. Cancer Cell 25, 442–454 (2014).
Gehi-Boyer, V. et al. Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. J. Hematol. Oncol. 5, 12 (2012).
Wachsmuth, M., Hübner, A., Li, M., Madea, B. & Stoneking, M. Age-related and heteroplasmy-related variation in human mtDNA copy number. PLoS Genet. 12, e1005939 (2016).
Kennedy, S. R., Salk, J. J., Schmitt, M. W. & Loeb, L. A. Ultra-sensitive screening reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. 9, e1003794 (2013).
von Zglinicki, T. & Martin-Ruiz, C. M. Telomeres as biomarkers for ageing and age-related diseases. Curr. Mol. Med. 5, 197–203 (2005).
Alexandrov, L. A. et al. Clock-like mutational processes in human somatic cells. Nat. Genet. 47, 1402–1407 (2015).
Chainani, V. et al. Objective measures of the frailty syndrome (hand grip strength and gait speed) and cardiovascular mortality: a systematic review. Int. J. Cardiol. 215, 487–493 (2016).
Manrai, A. K., Patel, C. J. & Ioannidis, J. P. A. In the era of precision medicine – who is normal? PLoS Med. 319, 1981–1982 (2018).
Deelen, J. et al. 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 (2014).
Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354, eaad8314 (2016).
Amendola, L. M. et al. Actionable exonic incidental findings in 6503 participants: challenges of variant classification. Genome Res. 25, 305–315 (2015).
Dorschner, M. O. et al. Actionable, pathogenic incidental findings in 1,000 participants’ exomes. Am. J. Hum. Genet. 93, 631–640 (2013).
Martincocera, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886 (2015).
Artandi, S. E. et al. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature 406, 641–645 (2000).
Cherayil, L. et al. p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. Cell 97, 527–538 (1999).
Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. Nat. Genet. 45, 1452–1458 (2013).
Genin, E. et al. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol. Psychiatry 16, 903–907 (2011).
McLaren, W. et al. The Ensembl variant effect predictor. Genome Biol. 17, 122 (2016).
Ding, Z. et al. Estimating telomere length from whole genome sequence data. Nucleic Acids Res. 42, e175 (2014).
Cawthon, R. M. Telomere measurement by quantitative PCR. Nucleic Acids Res. 30, e47 (2002).
Alexandrov, L. L. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics 31, 3673–3675 (2015).
Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. 99, 673–686 (2004).
Holm, S. A simple sequentially rejective multiple test procedure. Scand. Stat. Theory Appl. 6, 65–70 (1979).
Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57, 289–300 (1995).
Acknowledgements
Whole genome sequencing of the MGBR, 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 334407, 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 ASRB (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 programme (to M.J.C., M.G.J., V.J.C.), a NARSAD Independent Investigator Grant (to M.J.C.) and National Health and Medical Research Council
Mark Pineau, Paul Lacaze, Emma M. Rath, Joseph E. Powell, Sini Nagpal, Clare Puttick, Shane Husson, Dmitry Degrave, Tina Navin Cristina, Ulf Gyllensten, Murray J. Cairns, Andrew Stone, Marie-Jo Brion, Chantel Fitzsimmons, Heath M. Cairns, Melissa J. Green, Martin McNamara, Moeen Riaz, Margo Barr.

Conceptualisation: M.P., M.E.D., and D.M.T.; methodology: M.P., E.M.R., J.E.P., C.P., G. G., 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., A.M., M.B., M.N., 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.I.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.; visualisation: 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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-14079-0.

Correspondence and requests for materials should be addressed to D.M.T.

Peer review information Nature Communications thanks Joris Deelen and the other anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2020