Author: Akker, Erik Ben van den
Title: Computational biology in human aging : an omics data integration approach
Issue Date: 2015-02-18

Chapter 4:

Germ line and Somatic Characteristics of the Long-Lived Genome

Erik B. van den Akker[1,2,], Steven J. Pitts[3], Joris Deelen[1,4], Matthijs H. Moed[1], Shobha Potluri[3], H. Eka D. Suchiman[1], Nico Lakenberg[1], Wesley J. de Dijcker[1], Anton J.M. de Craen[5], Jeanine J. Houwing-Duistermaat[6], Genome of the Netherlands Consortium[7], David R. Cox[3†], Marian Beekman[1,4], Marcel J.T. Reinders[2], P. Eline Slagboom[1,4]

[1.] Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[2.] The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[3.] Rinat-Pfizer Inc, South San Francisco, United States of America

[4.] Netherlands Consortium of Healthy Ageing

[5.] Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands

[6.] Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

[7.] Genome of the Netherlands Consortium members are listed in the Supplemental Materials.

† In memoriam

Part of this work was submitted and is currently under review

1. Abstract

Human longevity has an estimated heritability of approximately 25% in the population at large, which remains largely unexplained by known common genetic variation. The missing heritability in the longevity phenotype might be explained by rare disruptive variants that can be readily measured by the current sequencing techniques. Here we report the results of a whole genome sequencing study into familial longevity comparing the genomes of 218 independent nonagenarians originating from families with a multi-generational history of extended survival into old age and 98 ethnicity-matched random population controls. An exome-wide comparison did not reveal any robust differences in the overall prevalence of rare disruptive variants between the genomes of long-lived cases and random population controls. In contrast, recurrent rare disruptive variants were identified in two key epigenetic genes, e.g. *TET2* and *DNMT3A*, in long-lived cases exclusively, which suggests that a reduced functionality in these genes relates to longevity. Read depth evidence and Sanger re-sequencing data, however, indicated that the variants identified in *TET2* and *DNMT3A* were in general of somatic origin, and should therefore be discarded as potential heritable factors underlying familial longevity. Somatic variation in these genes is generally regarded as an indicator of age-associated outgrowth of myeloid progenitor cells, a pre-malignant phase, that marks the aging hematopoietic stem cell compartment and an increased susceptibility to leukemia. Although nonagenarian carriers of somatic disruptive variants in *TET2* and *DNMT3A* may exhibit signs of a shift in blood cell composition, they did not display a significantly compromised survival during a 10-year follow up. To conclude we found no robust evidence for the long-lived genome to carry either an overall excess or depletion of germ line rare disruptive variants. We do observe an increased prevalence of somatic variation in specific loci likely to stimulate clonal outgrowth.

2. Introduction

In western societies, life expectancy has been steadily growing over the past two centuries[1], yet striking variations in life span are observed among the population at large[2]. Human life span regulation is an extraordinary complex outcome and is largely determined by chance and factors from the environment, though a modest contribution of heritable components (~25%) is also expected in the general population[3]. The propensity to become long-lived nevertheless clearly runs in families[4-6] and seems to relate to the capacity to delay or evade age-associated disease. Offspring of nonagenarians, centenarians and super centenarians display a lower prevalence of cardiovascular disease, type II diabetes and cancer[4-6], as compared to the general population, thus suggesting that human longevity is caused by genetic factors modifying risk of age-associated disease. However, compared to the general population, the genomes of nonagenarians do not show a depletion of common disease susceptibility alleles identified by genome-wide association studies (GWASs)[7], nor did GWASs for longevity revealed sufficient loci to explain the heritability of longevity[8]. Since GWASs predominantly focus on analysing common variants (Minor Allele Frequency>=1%), we hypothesize that the missing heritability of the longevity phenotype might be explained by rare coding variants with disruptive impact on the gene's functioning.

Rare disruptive variants can modify disease risk, like common variants, by affecting the expression or structure of translated proteins, which may contribute to longevity in two ways. First, the genome is reported to contain on average about 100 rare disruptive variants per individual that severely limit or totally negate the functionality of the associated proteins[9]. Hence a genome-wide depletion of such rare disruptive variants might implicate a more complete or better functioning proteome, improving the capacity to maintain the bodily homeostasis. Moreover, such a genome-wide depletion of variants might also point to an improved fidelity of the DNA repair system as compared to the general population[10,11]. Secondly, a targeted knockdown of a single gene in model organisms can already give rise to a long-lived species[12]. Hence, a local enrichment of rare disruptive variants in the genomes of long-lived individuals might implicate that a similar loss of function of the gene originating from that particular locus promotes longevity in humans. Though both genetic mechanisms are plausible, little evidence exists to date whether the genetic propensity for human longevity relates more closely to a fitter proteome or the targeted disruption of particular gene functions.

The first NGS efforts to study rare variants in longevity involve study designs with few extreme cases. The genomes of super-centenarians and centenarians were sequenced in order to describe genetic features of exceptional longevity[13-17]. Obviously, these analyses have a very limited statistical power for revealing evidence in favour of any of the two proposed genetic mechanisms for longevity mentioned above. However, also these very extreme cases do not show a depletion of common disease susceptibility alleles as identified by genome-wide association studies (GWASs), in line with work of Beekman *et al.*[7]. Using a more targeted approach, 988 candidate longevity genes were sequenced in 6 centenarians to identify novel non-synonymous SNVs[18], which were subsequently tested in larger case control studies and suggested *PMS2* and *GABRR3* as novel candidate longevity genes. These initial studies provide some first insights into genetic backgrounds that are conductive to exceptional longevity.

To investigate potential genetic mechanisms for human longevity involving rare disruptive variants, whole-genome sequencing was performed by Complete Genomics on DNA derived of 218 nonagenarian participants of the Leiden Longevity Study (LLS). The Leiden Longevity Study consists of sib pairs of which female members reached at least 91 years of age and male members 89 years of age. First-degree family members of these nonagenarian siblings show a 30% survival advantage as compared to their birth cohort[19]. Moreover, offspring of these nonagenarians exhibit a propensity for healthy aging already at middle age, as indicated by their significantly lowered incidence of hypertension, type II diabetes and use of cardiovascular medication, as compared to population controls[4]. We therefore hypothesize that LLS families show healthy aging and longevity by their genetic predisposition. To further identify genetic variation that predisposes to familial longevity, we compared the genomes of these 218 unrelated long-lived cases with those of 98 younger population controls of the Biobanking and Biomolecular Resources Research Infrastructure of the Netherlands (BBMRI-NL) consortium[20,21].

3. Results

3.1 Study design and variant detection

We explored the human genome for rare variants contributing to human longevity using whole genome sequencing data of 218 independent long-lived cases from the LLS (median age 93.7, N_{male} = 82 (37.6%)) and 98 population controls of the BBMRI biobanking initiative (median age 57, N_{male}= 39 (39.6%)) (Experimental Procedures 5.1). DNA sequencing and subsequent variant calling was performed by Complete Genomics (Complete Genomics Inc., Mountain View California) (median read depth >30x) on genetic material isolated from peripheral blood. Sequencing data were subjected to a stringent quality control prior to performing the analyses. For the following analysis we considered Single Nucleotide Variants (SNVs), small deletions (DELs) and insertions (INSs) called at high quality and with a minimal call rate of 95% in both long-lived cases and population controls. For a more detailed description of variant detection and quality control see Experimental Procedures 5.2.

3.2 Depletion of coding variation in longevity genomes

The genome-wide burden of disruptive genetic variants in long-lived cases compared to the population controls was investigated for all variants in the coding sequence (CDS) jointly and for variants

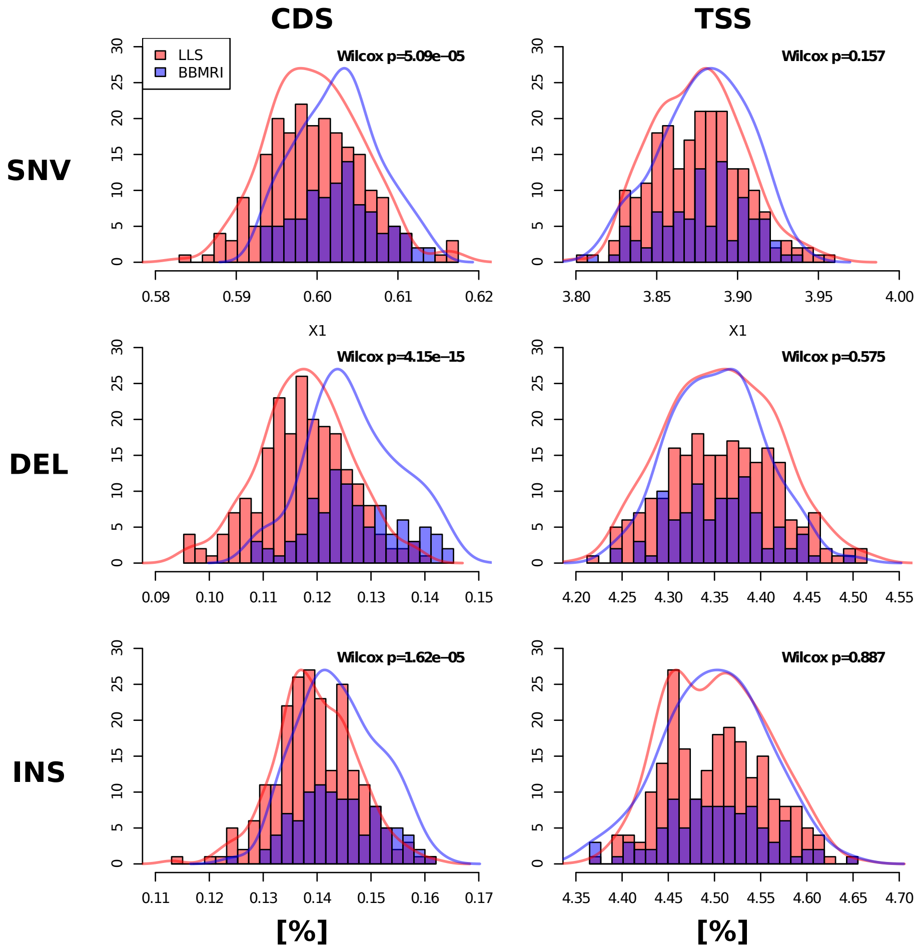
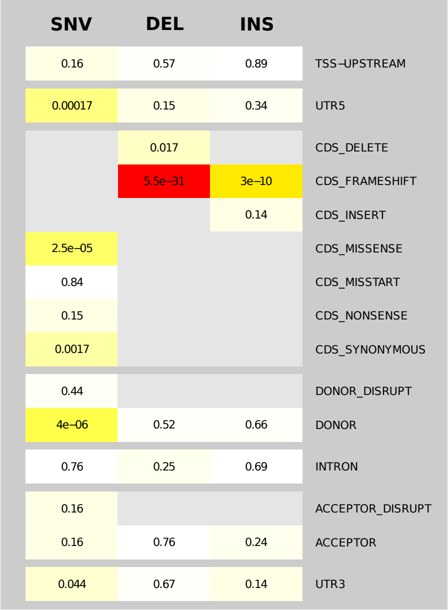


Figure 1: Depletion of coding variants in genomes of long-lived individuals. Distributions of proportions of variants annotated to the CDS (coding sequence) or sequence upstream of the Transcription Start Site (TSS, 0-7.5kb) for each of the three small variant types (SNV, DEL, INS) are displayed for long-lived cases (LLS; red) and random population controls (BBMRI; blue) respectively. Test results for differences in these distributions are reported in the upper right corner (Wilcoxon Rank-Sum test). Whereas a significant depletion of coding variants was observed for all small variant types in long-lived cases (LLS) compared to population controls (BBMRI), no such association was observed for the proportion of variants annotated to TSS.

categorized per impact (e.g. missense or nonsense) and type (single nucleotide variant: SNV, small deletions: DEL or insertions: INS). Counts per thus formed categories were normalized per individual on the totals of variants observed for each variant type to negate biases from overall differences in variant calling between the cohorts. Using this approach, we detect a lowered proportion of variants annotated to the CDS in nonagenarians cases compared to the population controls for all types of variants (Wilcoxon Rank-Sum test: SNV: p=5.09×10^{-5}, DEL: p=4.15×10^{-15} and INS: p=1.62×10^{-5}; Figure 1, left column). As a negative control, we tested for differences in proportions of variants annotated up to 7.5 kb upstream of the Transcription

Start Site (TSS) and did not observe any significant differences (SNV: *p*=0.157, DEL: *p*=0.575 and INS: *p*=0.887, Figure 1, right column). Since total numbers of variants might also reflect the quality of alignment and depth of sequencing, we inspected the correlation between the proportions of variants annotated to the CDS and the total numbers of variants discovered in cases and controls (Supplemental Figure 1), but found no significant biases. Hence, compared to the general population, long-lived cases show a depletion of variation in the coding part of the genome.



SNV	DEL	INS	
0.16	0.57	0.89	TSS–UPSTREAM
0.00017	0.15	0.34	UTR5
	0.017		CDS_DELETE
	5.5e–31	3e–10	CDS_FRAMESHIFT
		0.14	CDS_INSERT
2.5e–05			CDS_MISSENSE
0.84			CDS_MISSTART
0.15			CDS_NONSENSE
0.0017			CDS_SYNONYMOUS
0.44			DONOR_DISRUPT
4e–06	0.52	0.66	DONOR
0.76	0.25	0.69	INTRON
0.16			ACCEPTOR_DISRUPT
0.16	0.76	0.24	ACCEPTOR
0.044	0.67	0.14	UTR3

FIGURE 2: DEPLETION OF DISRUPTIVE VARIANTS IN GENOMES OF LONG-LIVED INDIVIDUALS. A heatmap displaying the results of all variant-categories created by cross tabulating variant-types (columns: SNV, DEL and INS) and variant-impacts (rows: TSS-UPSTREAM (Transcription Start Site and 7.5 kb upstream), UTR5 (UnTranslated Region at 5')), CDS_DELETE (in frame deletion), CDS_FRAMESHIFT (out of frame deletion or insertion), CDS_INSERT (in frame insertion), CDS_MISSENSE (amino acid substitution), CDS_MISSTART (start removed), CDS_NONSENSE (stop created), CDS_NONSYNONYMOUS (no change to protein), DONOR_DISRUPT (2 bp of essential splice donor site), DONOR (12bp of splice donor site), INTRON, ACCEPTOR_DISRUPT (2 bp of essential splice acceptor site), ACCEPTOR (8 bp of splice acceptor site), UTR3 (UnTranslated Region at 3'). The intensity of each cell represents the significance of the Wilcoxon Rank-Sum test computed on the difference in proportions of a particular variant-type annotated to a variant-category between the long-lived cases and the population controls. P-values are displayed in the cells. Cells are empty if no or to little data were available for testing. Note that the frameshift variants are most significantly depleted in the long-lived cases as compared to the random population controls.

When applying the testing to the more fine-grained annotations of the coding sequence, as provided by Complete Genomics[22] we observe that the depletion of CDS variants in long-lived cases compared to population controls can be explained by a few categories in particular. DELs and INSs inducing frameshifts, and missense and synonymous SNVs were present in significantly lower proportions in the long-lived cases as compared to the population controls (Figure 2, Supplemental Table 1). In addition, SNVs residing in splice donor sites and the 5' untranslated regions (5UTR) displayed a similar depletion. Of the depleted variant categories, we expect the most disruptive variant categories to show the highest depletion in long-lived cases. To verify this, counts of frameshift DELs and INSs were re-analyzed, while normalizing for frame preserving DELs and INSs and counts of missense SNVs or SNVs residing in splice donor sites or 5UTR were normalized on counts of synonymous SNVs (Figure 3). Indeed frameshift DELs (p = 1.84×10^{-26}) and INSs (p = 2.60×10^{-09}) and SNVs residing in splice donor sites

(p = 4.51×10^{-04}) displayed an additional significant depletion on top of the general depletion of coding variation in long-lived cases compared to population controls.

High impact variants calls made with short-read sequencing platforms are associated with an increased false positive rate. To investigate the rates of truly reported high impact variants in long-lived cases and random population controls, we randomly selected 15 frameshift variants in each of the two cohorts and validated these using Sanger sequencing. Of the 15 assays for frameshift variants only observed in the long-lived cases, 12 returned good data, which confirmed the presence of seven (58.3%) frameshift variants (Supplemental Table 2). Whereas all of the 15 assays for frameshift variants observed in the population controls that could be successfully designed, only two (13.3%) validated the presence of its targeted variant (Supplemental Table 3). Thus, the ratio of falsely reported variants within the two small samples of high impact variants is considerable, and notably, highest amongst population controls. DNA of long-lived cases and population controls was sequenced on the same platform, be it at two different points in time (within

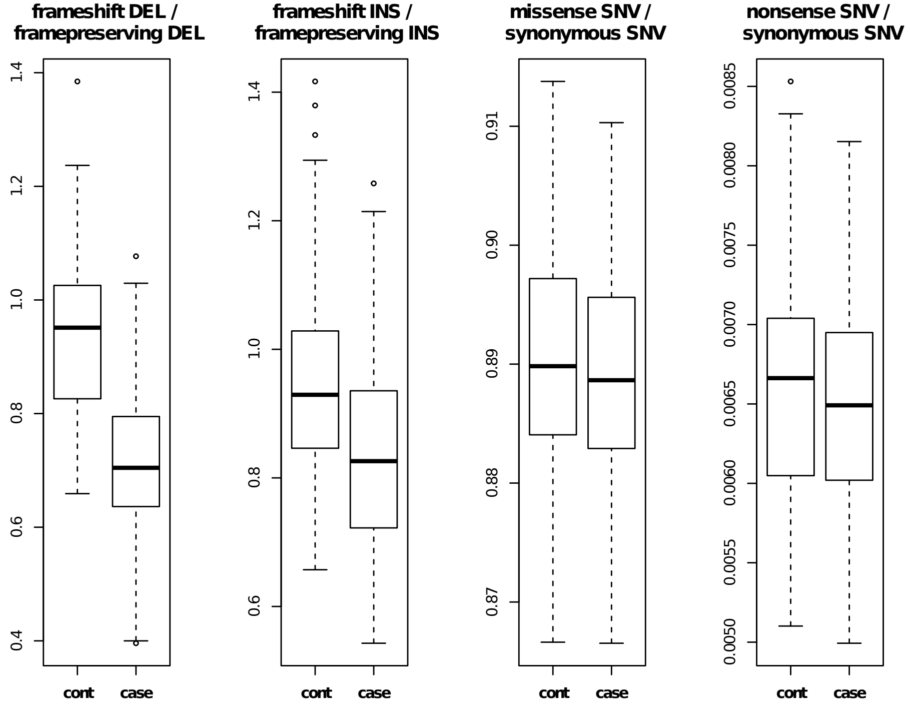


Figure 3: The higher the impact, the more depleted. When normalizing the counts in the more disruptive variant categories on those in the less disruptive variant categories of the same variant type, e.g. by normalizing counts on frame shifting DELs on frame preserving DELs, we confirm our previous findings of a depletion of the most disruptive variants in long-lived cases compared to those population controls. Frameshift DELs (p = 1.84×10^{-26}) and INSs (p = 2.60×10^{-09}) and SNVs residing in splice donor sites (p = 4.51×10^{-04}) displayed an additional significant depletion on top of the general depletion of coding variation in long-lived cases compared to population controls.

2 years), possibly leading to a technical bias. From the validation experiment we conclude that the previously observed difference in prevalence of disruptive variants is most likely due to an elevated false discovery rate in the controls rather than a depletion of rare disruptive variants in the long-lived cases.

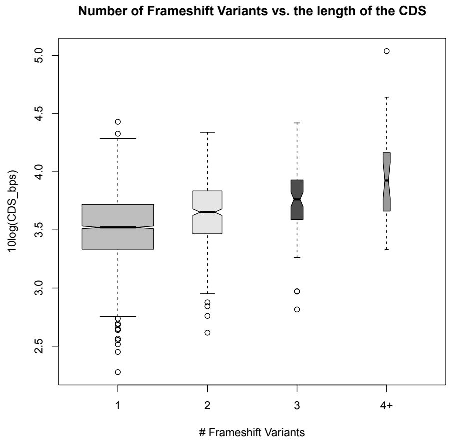


Figure 4: Longer genes are more likely to catch frameshift variants. When plotting the length of the coding sequence as a function of the number of frameshift indels we observe a clear positive correlation.

3.3 Rare disruptive variants cluster at *TET2* and *DNMT3A* in nonagenarian genomes

Moving away from the whole genome depletion of variants, we next investigated whether genes are preferentially hit by disruptive variants as postulated in the second proposed genetic mechanism for human longevity. To investigate which genes are preferentially hit by the disruptive frameshift variants, irrespective of the study, i.e. in long-lived cases and in population controls, we collapsed the deletions and insertions to gene annotations. This yielded a total of 2,193 unique deletions and 1,764 unique insertions in respectively 1,970 and 1,601 genes. Assuming a coding transcriptome of 18,000 independent transcript clusters, we used a resampling approach to assess the significance of the joint presence of the numbers of frameshift deletions and insertions per gene (Experimental Procedures 5.3). The 27 genes hit by at least four unique frameshift mutations are presented in Table 1 and jointly comprise 3.2% of the total number of frameshift variants observed. A strong trend between the length of the coding sequence and the number of frameshift variants present in genes in cases and controls jointly was observed (Figure 4), with the largest gene present in the genome, *TTN*, showing the most significant enrichment of frameshift variants. Hence, few relatively long genes accumulate multiple frameshift variants.

Next we investigated whether any of the 27 genes with four or more frameshift indels was preferentially hit by mutations unique to either the long-lived cases or the population controls. By again using a resampling approach, we assessed the significance of the observed number of private frameshift deletions and insertions present in each of the genes (Experimental Procedures 5.4). Interestingly, we note that the most significant gene-specific accumulations of frameshift variants occur in two genes hit in long-lived cases only: *TET2* and *DNMT3A* (Table 2). Other categories variant types, e.g. nonsense SNVs, confirmed the burden of disruptive variants in *TET2* and *DNMT3A* present in only long-lived cases (Table 3). In total, *TET2* was hit by six frameshift indels and

GeneSymbol	EntrezGeneID	DEL	INS	DEL + INS	p.perm
TTN	7273	7 (2/5/0)	2 (0/2/0)	9 (2/7/0)	$<1.00 \times 10^{-6}$
DNAH14	127602	8 (4/2/2)	0 (0/0/0)	8 (4/2/2)	$<1.00 \times 10^{-6}$
FSIP2	401024	1 (1/0/0)	6 (3/2/1)	7 (4/2/1)	$<1.00 \times 10^{-6}$
LOC100506072	100506072	3 (2/0/1)	4 (1/2/1)	7 (3/2/2)	$<1.00 \times 10^{-6}$
TET2	54790	4 (4/0/0)	2 (2/0/0)	6 (6/0/0)	5.00×10^{-6}
SSPO	23145	5 (2/0/3)	1 (1/0/0)	6 (3/0/3)	6.00×10^{-6}
VPS13C	54832	3 (0/3/0)	2 (1/1/0)	5 (1/4/0)	1.90×10^{-5}
IL3RAY	8218	0 (0/0/0)	4 (3/0/1)	4 (3/0/1)	9.90×10^{-5}
SYNE1	23345	0 (0/0/0)	4 (2/2/0)	4 (2/2/0)	9.90×10^{-5}
UGGT2	55757	0 (0/0/0)	4 (2/2/0)	4 (2/2/0)	9.90×10^{-5}
SLFN12L	100506736	1 (1/0/0)	3 (3/0/0)	4 (4/0/0)	1.27×10^{-4}
ZBTB1	22890	1 (1/0/0)	3 (2/1/0)	4 (3/1/0)	1.27×10^{-4}
SPATA3E1	286234	1 (1/0/0)	3 (2/1/0)	4 (3/1/0)	1.27×10^{-4}
PNPLA7	375775	1 (1/0/0)	3 (2/0/1)	4 (3/0/1)	1.27×10^{-4}
HECTD4	283450	1 (1/0/0)	3 (1/2/0)	4 (2/2/0)	1.27×10^{-4}
PTCHD3	374308	1 (0/0/1)	3 (2/0/1)	4 (2/0/2)	1.27×10^{-4}
POLQ	10721	2 (0/1/1)	2 (0/2/0)	4 (0/3/1)	1.36×10^{-4}
NOTCH3	4854	2 (0/2/0)	2 (1/1/0)	4 (1/3/0)	1.36×10^{-4}
ADAM8	101	2 (0/1/1)	2 (2/0/0)	4 (2/1/1)	1.36×10^{-4}
NIN	51199	2 (0/2/0)	2 (2/0/0)	4 (2/2/0)	1.36×10^{-4}
ZNF469	84627	2 (0/2/0)	2 (2/0/0)	4 (2/2/0)	1.36×10^{-4}
MUC16	94025	2 (0/1/1)	2 (1/1/0)	4 (1/2/1)	1.36×10^{-4}
LMOD2	442721	3 (0/3/0)	1 (0/1/0)	4 (0/4/0)	1.91×10^{-4}
PIK3C2G	5288	3 (3/0/0)	1 (0/1/0)	4 (3/1/0)	1.91×10^{-4}
TNRC18	84629	3 (2/1/0)	1 (1/0/0)	4 (3/1/0)	1.91×10^{-4}
DNMT3A	1788	4 (4/0/0)	0 (0/0/0)	4 (4/0/0)	2.11×10^{-4}
ABCA10	10349	4 (1/1/2)	0 (0/0/0)	4 (1/1/2)	2.11×10^{-4}

Table 1: The 27 genes accumulating at least 4 frameshift variants. Counts of variants are given for DELetions and INSertions separately according to the following format: A (B/C/D) indicate respectively the total (A), private in case (B), private in control (C) and shared number of variants (D).

GeneSymbol	DELs	INSs	Totals	*p*_case	*p*_cont
TET2	4 (4/0/0)	2 (2/0/0)	6 (6/0/0)	0.0049	1
DNMT3A	4 (4/0/0)	0 (0/0/0)	4 (4/0/0)	0.019	1

Table 2: Genes with a private burden in long-lived cases. Within the top 27 genes accumulating at least 4 frameshift variants, *TET2* and *DNMT3A* exhibited a study specific preference. Noteworthy is that both these genes feature frameshift variants in only the long-lived cases.

Gene	Chrom	Start	End	Type	Ref	Alt	Impact	LLS	BBMRI
TET2	chr4	106155736	106155737	DEL	T	-	FRAMESHIFT	1	0
TET2	chr4	106155765	106155766	DEL	G	-	FRAMESHIFT	1	0
TET2	chr4	106156685	106156686	SNV	C	A	NONSENSE	1	0
TET2	chr4	106156758	106156758	INS	-	C	FRAMESHIFT	1	0
TET2	chr4	106157246	106157246	INS	-	A	FRAMESHIFT	1	0
TET2	chr4	106157781	106157782	DEL	G	-	FRAMESHIFT	1	0
TET2	chr4	106157913	106157914	SNV	C	T	NONSENSE	1	0
TET2	chr4	106158107	106158108	SNV	G	A	NONSENSE	1	0
TET2	chr4	106196212	106196213	SNV	C	T	NONSENSE	1	0
TET2	chr4	106196221	106196222	SNV	G	T	NONSENSE	1	0
TET2	chr4	106197352	106197353	DEL	A	-	FRAMESHIFT	1	0
DNMT3A	chr2	25463181	25463182	SNV	G	A	NONSENSE	2	0
DNMT3A	chr2	25463296	25463296	INS	-	A	NONSENSE	1	0
DNMT3A	chr2	25468153	25468154	DEL	G	-	FRAMESHIFT	1	0
DNMT3A	chr2	25468921	25468923	DEL	AC	-	FRAMESHIFT	1	0
DNMT3A	chr2	25469921	25469922	SNV	G	A	NONSENSE	1	0
DNMT3A	chr2	25469990	25469991	DEL	A	-	FRAMESHIFT	1	0
DNMT3A	chr2	25470930	25470931	DEL	G	-	FRAMESHIFT	1	0

Table 3: Frameshift and nonsense mutations identified in *TET2* and *DNMT3A*, exclusively present in long-lived cases.

five nonsense SNVs and *DNMT3A* by four frameshift indels, two nonsense SNVs and a single nonsense insertion, all in the 218 genomes of long-lived cases only. Moreover, a look-up on the Exome Variant Server (http://evs.gs.washington.edu/EVS) in exome sequencing results in ~4,125 U.S. participants of European ancestry revealed that *TET2* and *DNMT3A* were hit with unique frameshift indels or nonsense SNVs with a significantly lower frequency (*TET2*: $N_{disrupt\_EVS}$=9, OR: 24.2 95% CI: 9.0-67.0, p=4.5×10^{-10}; *DNMT3A*: $N_{disrupt\_EVS}$=7, OR: 19.5 95% CI: 5.8-65.6, p=1.9×10^{-6}, Fisher's Exact tests, Supplemental Table 4).

Unlike the poor validation rates observed for frameshift variants sampled from the whole genome, frameshift variants identified within *TET2* and *DNMT3A* in the long-lived were generally confirmed using Sanger sequencing (9 out of 10). A closer inspection of these Sanger sequencing results showed in general a much lower signal for the mutant allele as compared to the wild-type allele, an observation supported by the whole genome sequencing results for the frameshifting indels in *TET2* and *DNMT3A* (Table 4). This clear deviation from the 1:1 ratio (Experimental Procedures 5.6), as expected for heterozygous germ line variants, suggests that the identified variants are present in only a part of the measured cells. These results support the impression that the

Gene	Chrom	Start	End	Type	# Ref	# Alt	% Alt	p_{som}
TET2	chr4	106155736	106155737	DEL	30	11	26.8%	0.017
TET2	chr4	106155765	106155766	DEL	60	9	13.4%	2.7×10^{-7}
TET2	chr4	106156758	106156758	INS	33	10	23.3%	0.0047
TET2	chr4	106157246	106157246	INS	24	9	27.3%	0.034
TET2	chr4	106157781	106157782	DEL	36	13	26.5%	0.0083
TET2	chr4	106197352	106197353	DEL	47	22	31.9%	0.016
DNMT3A	chr2	25463296	25463296	INS	47	17	26.6%	0.0028
DNMT3A	chr2	25468153	25468154	DEL	34	10	22.7%	0.0035
DNMT3A	chr2	25468921	25468923	DEL	30	8	21.1%	0.0039
DNMT3A	chr2	25469990	25469991	DEL	34	21	38.2%	0.12
DNMT3A§	chr2	25470930	25470931	DEL	47	4	7.8%	0.0019

Table 4: Number of reads supporting the reference and alternative alleles of frameshift variants in *TET2* and *DNMT3A*. All Frameshift variants identified in the long-lived cases could be confirmed by Sanger sequencing except the variant marked by §. Since this non-confirmed variant had a relatively low % Alt of 7.84% it leaves the possibility that this variant may have gone undetected, as it was not present in a sufficient proportion of the sequenced cells.

long-lived cases, as compared to the younger population controls, have a higher prevalence of somatic frameshifting indels in *TET2* and *DNMT3A*.

Somatic mutations in *TET2* and *DNMT3A* have previously been associated with aging of hematopoietic stem cells (HSCs)[23], which is characterized by a skewing of progenitor cells towards the myeloid fate that compromises immune function and increases the risk for myeloid malignancies[24,25]. Hence, we investigated whether carriership of the identified disruptive variants (Table 3) in long-lived cases was reflected by their blood cell composition. Whereas no signs of skewing in the blood cell composition was observed for the carriers of disruptive variants in *TET2* (β=1.29, 95% CI: -1.04-0.78, *p*=0.78), we observed that carriers with disruptive variants in *DNMT3A* have significantly higher granulocyte counts than non-carriers (β=1.29, 95% CI: 0.24-2.43, *p*=0.016, Experimental Procedures 5.7). Since this may indicate an underlying risk for a compromised immune-capacity or hematopoietic malignancies, we compared the prospective survival of long-lived carriers versus long-lived non-carriers. A prospective survival analysis with a ten years follow-up did not indicate a significantly increased risk on mortality for the carriers of disruptive variants in either *TET2* (N_{tot}=214, N_{death}=190, HR=1.30, 95% CI 0.68-2.47, *p*=0.424) or *DNMT3A* (N_{tot}=214, N_{death}=190, HR=0.37, 95% CI 0.15-0.91, *p*=0.031, Experimental Procedures 5.8). In fact, a modest protective effect was observed for *DNMT3A* mutant carriers (Figure 5) and noteworthy, 4 out of the 9 carriers were still alive at our most recent census of 2012 at ages 99, 100, 104 and 105.

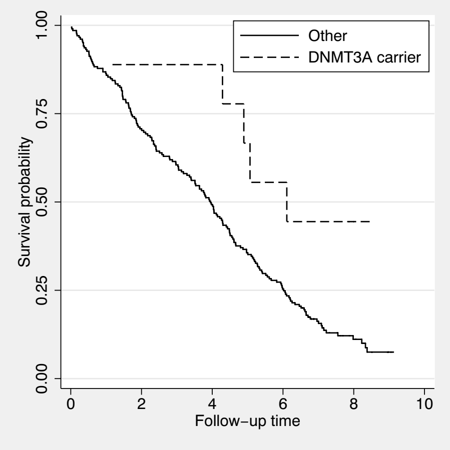


Figure 5: Prospective survival on carriers of disruptive variants in *DNMT3A*. Kaplan-Meier curves for the long-lived cases carrying either a nonsense SNV or frameshift indel in *DNMT3A,* as compared to long-lived non-carriers.

4. Discussion

In the current study we analysed the genome of 218 independent nonagenarians for rare disruptive variants contributing to familial longevity. Although our sequencing study is the largest amongst the oldest old, we found no decisive evidence for either an excess or depletion of rare disruptive germ line variants to contribute to familial longevity. In contrast, we did observe and validate recurrent somatic variants in *TET2* and *DNMT3A,* exclusively present in the genomes of long-lived cases. Hence, we conclude that within this limited sample size, the characteristics most discriminative for the long-lived genome are acquired during life, which, to our current understanding, seem unlikely to constitute a heritable component predisposing to familial longevity.

The genomes of long-lived cases exhibited a gene-specific burden of rare somatic disruptive variants from multiple categories in *TET2* and *DNMT3A*. Somatic mutations in *TET2* and *DNMT3A* were first reported in patients suffering from myeloid malignancies[26,27], but also appear in elderly exhibiting myelodysplasia without overt hematopoietic malignancies[23]. This suggests that somatic mutations in *TET2* and *DNMT3A* in hematopoietic stem cells confer enhanced self-renewal and clonal expansion leading to an age-related myeloid lineage bias. Indeed significantly elevated levels of granulocytes were observed in carriers of somatic mutations in *DNMT3A*. Surprisingly, neither the carriers of somatic disruptive mutations in *DNMT3A*, nor in *TET2*, did exhibit a significantly increased mortality risk over 10 years time, while similar mutations have previously been associated with an increased risk of progression to and poor outcome of acute myeloid leukaemia (AML)[27]. This either suggests that clonal expansion of the myeloid lineage in itself may not necessarily contribute to cancer risk in the highest ages, or alternatively, it may suggest that additional genetic factors, absent in long-lived, may be required for transforming into AML, in which case carriership may accelerate disease progression. Since these somatic mutations are typically found in elderly patients, it is reasonable to assume that the genetic burden at these loci should in effect be interpreted as markers of chronological age, rather than heritable factors underlying human longevity.

Assuming that the disruptive mutations in *TET2* and *DNMT3A* have been acquired during life, in absence of any overt malignancies, the question rises whether these somatic variants in fact could have contributed to the observed extension in lifespan. Both *TET2* and *DNMT3A* are factors for epigenetic control[28,29] and are thought to silence hematopoietic stem cell self-renewal to permit efficient hematopoietic differentiation[30,31]. Therefore, loss of functionality in these genes is likely to underlie an enhanced self-renewal leading to the observed age-related myeloid lineage bias. This skewing towards the myeloid lineage is assumed to have adverse effects on immune functionality in normal healthy individuals, but in the oldest old the increase of the myeloid compartment might be compensative for the age-related decrease in naive T-cells, known as immuno-senescence[32]. Hence on condition that the enhanced self-renewal, instigated by somatic disruptive mutations in *TET2* and *DNMT3A*, leads to increased levels of competent immune cells, be it of the myeloid lineage though, might partly compensate for the age-related loss of immuno-capacity of the lymphoid compartment.

Initial analyses of the whole-genome sequencing data lead us to the false impression that the long-lived genome was characterized by a depletion of coding variation, most evidently present amongst SNVs residing in splice donor sites or indels leading to a frameshift. The prevalence of these disruptive variants per individual is generally very low, which indicates that these types of variants are generally not tolerated. This also explains the increased false positive rate amongst the variant calls of disruptive variants generally observed in sequencing studies, including the current one. Validation experiments indicated that the few disruptive variants observed in the long-lived cases and population controls combined, were almost as likely to be erroneous as to be genuine and notably that the false positive rate was considerably higher amongst population controls. We therefore conclude that a genome-wide depletion of germ line disruptive variants in the genomes of long-lived individuals could not be decisively shown.

We conclude, that nonagenarian members of long-lived families have an increased prevalence of somatic disruptive variants in *TET2* and *DNMT3A.* Given their somatic origin, however, these variants seem unlikely to represent the heritable component of familial longevity. Previously, somatic mutations in these loci have been associated with risk on progression to[33,34] and poor prognosis of AML[27,35]. In the long-lived cases of our study, however, disruptive somatic variants in *TET2* and *DNMT3A* do not seem to compromise the 10-year survival. Implications of this finding are twofold. First, clinical risk assessments based on the mutational status of *TET2* and *DNMT3A* might not be accurate for the oldest old. Secondly, elderly carrying the somatic disruptive mutations in *TET2* and *DNMT3A* in absence of any overt malignancies may provide key insights in the factors most decisive for oncogenic transformation. Hence, the implications of somatic mutations in either *TET2* or *DNMT3A* for health in the oldest old remain illusive and therefore warrant more research into these key epigenetic loci.

5. Experimental Procedures

5.1 Study population

The Leiden Longevity Study[4] is a family based study consisting of 421 Dutch Caucasian nonagenarian sibships and is designed to investigate the genetic determinants of human longevity. To maximally enrich for genetic signal predisposing to human longevity within the sample of sequenced genomes, we selected those sibships (N=218) displaying the most profound family history of excess survival[36]. For each of these sibships, the DNA sequence of the genome of the sib with the highest age at censoring was determined using Next Generation Sequencing (Complete Genomics Inc.). As controls for our study, we employed sequencing data assayed on 100 individuals of Dutch Caucasian origin aged below 65 and collected by the Dutch Biobanking and Biomolecular Resources Research Infrastructure initiative[20,21] (BBMRI). Participants of BBMRI are not selected for particular characteristics other than that they should reflect a random sample of the apparently healthy Dutch population.

5.2 Data preprocessing and quality control

Complete Genomics performed whole genome sequencing (>30x), read alignment and variant calling for both the long-lived cases as population controls, though at different time points. To minimize the technical variance between datasets, raw sequencing data created on the LLS samples was reprocessed by Complete Genomics to match the version of the preprocessing pipeline used for calling variants in the genomes of the BBMRI participants. The quality of the resulting data was re-checked (Supplemental Figures 2-6) per study separately and in combination.

One of the population controls was excluded beforehand for its distant familial relationship with one of the nonagenarian cases. Another population control displayed excessive proportions of unique variants indicating either a potential contamination of the sample before sequencing or a mixed ancestry of one of the BBMRI participants. Multidimensional scaling was performed with 10,000 randomly selected common SNVs (MAF ≥ 5%), and did not indicate the presence of population substructure. In effect, all following comparisons reported in this paper have been performed using 218 nonagenarian cases (median age 93.7, $N_{\mathrm{male}} = 82$ (37.6%)) and 98 population controls (median age 57, $N_{\mathrm{male}} = 39$ (39.6%)).

5.4 Assessing the significance of a genic burden of frameshift indels

To assess the significance of the presence of $k_{j,\mathrm{D}}$ unique frameshift deletions and $k_{j,\mathrm{I}}$ unique frameshift insertions jointly giving rise to k_j unique frameshift mutations in gene j, irrespective whether observed in long-lived cases or population controls, the following resampling approach was used. Assuming a coding transcriptome of 18,000 independent transcript clusters, we determined the prior probabilities of a gene being hit by a frameshift deletion (p_{D} = 2,193/18,000 = 0.122) or a frameshift insertion (p_{I} = 1,764/18,000 = 0.098). To assess the empirical probability $P(K_j > k_{j,\mathrm{D}+} k_{j,\mathrm{I}} | p_{\mathrm{del},} p_{\mathrm{ins}})$ we repeatedly resampled (Z=1,000,000) $k_{j,\mathrm{D}}$ deletions and $k_{j,\mathrm{I}}$ insertions with prior probabilities $p_{\mathrm{D},}$ and p_{I} and counted the number of times where the resampled numbers of frameshift variants k^s_j equaled or exceeded the number of observed frameshift variants k_j, yielding k^S_j. The estimated p-value is then obtained using:

$$\hat{P}\left(K_j > k_{j,D} + k_{j,I} \mid p_D, p_I\right) = \sum\nolimits_S (I(k_j^S)+1) \Big/ (Z+1) \tag{1}$$

Computations were performed in R[37] and repeated with different random seeds to verify the stability of the sampling experiments.

5.5 Assessing the significance of a case or control specific genic burden of frameshift indels

When inspecting the repeatedly hit genes, we noted that some genes were hit by frameshift mutations exclusively present (private) in either the long-lived cases or the population controls. To assess the significance of the preference of a gene for being hit by $k^{p}_{j,\mathrm{D}}$ private frameshift deletions and $k^{p}_{j,\mathrm{I}}$ private frameshift insertions jointly giving rise to k^{p}_{j} unique and exclusive frameshift mutations in gene *j*, all observed in either long-lived cases or population controls, the following resampling approach was used. First we determined the prior probabilities of a frameshift deletion to be exclusively observed in long-lived cases ($p_{\mathrm{D,case}}$ = 814/2,193 = 0.370) or population controls ($p_{\mathrm{D,ctr}}$ = 1,122/2,193 = 0.512) and a frameshift insertion to be exclusively observed in long-lived cases ($p_{\mathrm{I,case}}$ = 896/1,764 = 0.508) or population controls ($p_{I,ctr}$ = 729/1,764 = 0.413). Note that these probabilities do not add up to one as some deletions and insertions are observed in both the long-lived cases as the population controls and thus are not exclusive to any of the two. Furthermore, let $k_{j,D}$ and $k_{j,I}$ respectively be the total numbers of unique frameshift deletions and unique frameshift insertions observed for a particular gene *j*. Then we assess the empirical probability $P_{priv}(k^{p}_{j} >= k^{p}_{j,\mathrm{D}} + k^{p}_{j,\mathrm{I}} \,|p_{\mathrm{D,case}},p_{\mathrm{D,ctr}},p_{\mathrm{I,case}},p_{\mathrm{I,ctr}})$ for a given gene *j* by repeatedly resampling (Z=1,000,000) $k_{i,1}$ deletions and $k_{i,2}$ insertions with prior probabilities $p_{\mathrm{D,case}}$, $p_{\mathrm{D,ctr}}$, $p_{\mathrm{I,case}}$ and $p_{\mathrm{I,ctr}}$ for respectively obtaining private deletions ($k^{p,s}_{i,\mathrm{d}}$) and insertions ($k^{p,s}_{i,\mathrm{i}}$) in cases and controls for each sampling and subsequently counted the number of times the number of sampled private mutations $k^{p,s}_{i}$ equaled or exceeded the observed number of private mutations k^{p}_{i} ($k^{p,S}_{i}$). The p-value was then estimated by:

$$\hat{P}\left(k^{p}_{j} > k^{p}_{j,D} + k^{p}_{j,I} \mid p_{D,case}, p_{D,ctr}, p_{I,ctr}, p_{I,ctr} \right) = \frac{\sum_{S} (I(k^{p,S}_{j}) + 1)}{(Z+1)} \quad (2)$$

5.6 Somatic calls

Heterozygotic variant calls with read evidence deviating from the expected 1:1 ratio might point to the presence of a somatic variant that is present in part of the sequenced DNA. Alternatively, it might comprise either a sequencing error, or an under-sampling of a truly heterozygotic variant, which both can be modeled by employing Poisson distributions. First we model the probability of sequencing errors explaining the observed disbalance in ratio's, by assuming an error rate E = 1% of reads falsely supporting a variant call. Hence, a Poisson model $P(\lambda,K)$ with mean $\lambda = \mathrm{E} \times \mathrm{R}_{\mathrm{tot}}$ and $\mathrm{K} = \mathrm{R}_{\mathrm{var}}$ is used to estimate the probability p_{hom} that a variant, called with reads $\mathrm{R}_{\mathrm{tot}}$ of which at least $\mathrm{R}_{\mathrm{var}}$ support the variant, is likely to comprise a homozygous reference variant with some noisy reads. Similarly, we employ a Poisson model with mean $\lambda = 0.5 \times \mathrm{R}_{\mathrm{tot}}$ and $\mathrm{K} = \mathrm{R}_{\mathrm{var,}}$ to estimate the probability p_{het} that the alternative allele, supported by $\mathrm{R}_{\mathrm{var}}$ or less reads, is likely to be a truly heterozygotic variant of which the alternative allele is under-sampled relative to the reference. In case both these hypotheses are rejected, we may assume that the variant is indeed a somatic variant, thus: $p_{\mathrm{som}} = \max(p_{\mathrm{hom}}, p_{\mathrm{het}})$.

5.7 Associations with granulocyte counts

Absolute counts of granulocytes in long-lived cases were computed by summing counts of neutrophils, eosinophils and basophils derived from whole blood cell counts. Differences in granulocyte counts between carriers of disruptive variants in *TET2* or *DNMT3A* were tested using a linear model as implemented in the *lm* package of the statistical language R[37]:

$$G \sim \beta_{1} \times age + \beta_{2} \times sex + \beta_{3} \times carrier \quad (3)$$

where the covariates *age* is provided in years, *sex* as either 1 (male) or 2 (female), *carrier* as either 0 or 1 to indicate carriership of a disruptive variant.

5.8 Associations with prospective survival

Associations with prospective survival were performed with the *Survival* package[38] of R[37] using an age at inclusion and sex-adjusted, left-truncated Cox proportional hazards model to adjust for late entry into the dataset according to age. Mortality analyses between carriers and non-carriers of disruptive variants in *TET2* were performed using:

$$\lambda(t) \sim \lambda0(t) \times exp(\beta_1 \times age + \beta_2 \times sex + \beta_3 \times carrier) \quad (4)$$

where the covariates *age* designates age at inclusion and is provided in years, *sex* as either 1 (male) or 2 (female), *carrier* as either 0 or 1 to indicate carriership of a disruptive variant.

6. Acknowledgements

The research leading to these results has received funding from the Medical Delta (COMO), Pfizer Inc, and the European Union's Seventh Framework Programme (FP7/2007-2011) under grant agreement number 259679. This study was financially supported by the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology and the Netherlands Consortium for Healthy Ageing (grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO), by Unilever Colworth and by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

7. References

1. Oeppen, J. & Vaupel, J.W. Demography. Broken limits to life expectancy. *Science* **296**, 1029-31 (2002).
2. Hitt, R., Young-Xu, Y., Silver, M. & Perls, T. Centenarians: the older you get, the healthier you have been. *Lancet* **354**, 652 (1999).
3. Skytthe, A. *et al.* Longevity studies in GenomEUtwin. *Twin Res* **6**, 448-54 (2003).
4. Westendorp, R.G. *et al.* Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *J Am Geriatr Soc* **57**, 1634-7 (2009).
5. Atzmon, G. *et al.* Clinical phenotype of families with longevity. *J Am Geriatr Soc* **52**, 274-7 (2004).
6. Terry, D.F. *et al.* Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *J Am Geriatr Soc* **52**, 2074-6 (2004).
7. Beekman, M. *et al.* Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. *Proc Natl Acad Sci U S A* **107**, 18046-9 (2010).
8. 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* (2014).
9. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
10. Garinis, G.A., van der Horst, G.T., Vijg, J. & Hoeijmakers, J.H. DNA damage and ageing: new-age ideas for an age-old problem. *Nat Cell Biol* **10**, 1241-7 (2008).
11. Hoeijmakers, J.H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-85 (2009).
12. Clancy, D.J. *et al.* Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science* **292**, 104-6 (2001).
13. Sebastiani, P. *et al.* Whole genome sequences of a male and female supercentenarian, ages greater than 114 years. *Front Genet* **2**, 90 (2011).
14. Ye, K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* **16**, 1026-32 (2013).
15. Holstege, H. *et al.* A Longevity Reference Genome Generated From the World's Oldest Woman. *Oral Presentation ASHG 2011* (2011).
16. Gierman, H.J. *et al.* Whole-Genome Sequencing of the World's Oldest People. *PLoS One* **9**, e112430 (2014).

17. Cash, T.P. *et al.* Exome sequencing of three cases of familial exceptional longevity. *Aging Cell* **13**, 1087-90 (2014).

18. Han, J. *et al.* Discovery of novel non-synonymous SNP variants in 988 candidate genes from 6 centenarians by target capture and next-generation sequencing. *Mech Ageing Dev* **134**, 478-85 (2013).

19. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur J Hum Genet* **14**, 79-84 (2006).

20. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221-7 (2014).

21. The Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* (2014).

22. CG_website. http://media.completegenomics.com/documents/DataFileFormats_Standard_Pipeline_2.4.pdf.

23. Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet* **44**, 1179-81 (2012).

24. Beerman, I. *et al.* Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A* **107**, 5465-70 (2010).

25. Beerman, I., Maloney, W.J., Weissmann, I.L. & Rossi, D.J. Stem cells and the aging hematopoietic system. *Curr Opin Immunol* **22**, 500-6 (2010).

26. Delhommeau, F. *et al.* Mutation in TET2 in myeloid cancers. *N Engl J Med* **360**, 2289-301 (2009).

27. Ley, T.J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* **363**, 2424-33 (2010).

28. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-20 (1998).

29. Mohr, F., Dohner, K., Buske, C. & Rawat, V.P. TET genes: new players in DNA demethylation and important determinants for stemness. *Exp Hematol* **39**, 272-81 (2011).

30. Trowbridge, J.J. & Orkin, S.H. Dnmt3a silences hematopoietic stem cell self-renewal. *Nat Genet* **44**, 13-4 (2012).

31. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11-24 (2011).

32. Franceschi, C., Bonafe, M. & Valensin, S. Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity, and the filling of immunological space. *Vaccine* **18**, 1717-20 (2000).

33. Jankowska, A.M. *et al.* Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood* **113**, 6403-10 (2009).

34. Ewalt, M. *et al.* DNMT3a mutations in high-risk myelodysplastic syndrome parallel those found in acute myeloid leukemia. *Blood Cancer J* **1**, e9 (2011).

35. Metzeler, K.H. *et al.* TET2 mutations improve the new European LeukemiaNet risk classification of acute myeloid leukemia: a Cancer and Leukemia Group B study. *J Clin Oncol* **29**, 1373-81 (2011).

36. Rozing, M.P. *et al.* Familial longevity is associated with decreased thyroid function. *J Clin Endocrinol Metab* **95**, 4979-84 (2010).

37. R-Core-Team. R: A Language and Environment for Statistical Computing. (2013).

38. Therneau, T. A Package for Survival Analysis in S. (R package version 2.37-7; http://CRAN.R-project.org/package=survival, 2014).

Supplemental Materials

Genome of the Netherlands Consortium members:

Analysis group: Morris A. Swertz[6,7] (Co-Chair), Laurent C. Francioli[1], Freerk van Dijk[6,7], Androniki Menelaou[1], Pieter B.T. Neerincx[6,7], Sara L. Pulit[1], Patrick Deelen[6,7], Clara C. Elbers[1], Pier Francesco Palamara[2], Itsik Pe'er[2,8], Abdel Abdellaoui[9], Wigard P. Kloosterman[1], Mannis van Oven[10], Martijn Vermaat[11], Mingkun Li[12], Jeroen F.J. Laros[11], Mark Stoneking[12], Peter de Knijff[13], Manfred Kayser[10], Jan H. Veldink[14], Leonard H. van den Berg[14], Heorhiy Byelas[6,7], Johan T. den Dunnen[11], Martijn Dijkstra[6,7], Najaf Amin[15], K. Joeri van der Velde[6,7], Jouke Jan Hottenga[9], Jessica van Setten[1], Elisabeth M. van Leeuwen[15], Alexandros Kanterakis[6,7], Mathijs Kattenberg[9], Lennart C. Karssen[15], Barbera D.C. van Schaik[16], Jan Bot[17], Isaäuc J. Nijman[1], David van Enckevort[18], Hailiang Mei[18], Vyacheslav Koval[19], Kai Ye[20,21], Eric-Wubbo Lameijer[21], Matthijs H. Moed[21], Jayne Y. Hehir-Kwa[22], Robert E. Handsaker[5,23], Shamil R. Sunyaev[4,5], Mashaal Sohail[4,5], Fereydoun Hormozdiari[24], Tobias Marschall[25], Alexander Schönhuth[25], Victor Guryev[26], Paul I.W. de Bakker[1,3–5] (Co-Chair);

Cohort collection and sample management group: P. Eline Slagboom[21], Marian Beekman[21], Anton J.M. de Craen[21], H. Eka D. Suchiman[21], Albert Hofman[15], Cornelia van Duijn[15], Dorret I. Boomsma[9], Gonneke Willemsen[9], Bruce H. Wolffenbuttel[27], Mathieu Platteel[6], Steven J. Pitts[28], Shobha Potluri[28], David R. Cox[28,34];

Whole-genome sequencing: Qibin Li[29], Yingrui Li[29], Yuanping Du[29], Ruoyan Chen[29], Hongzhi Cao[29], Ning Li[30], Sujie Cao[30], Jun Wang[29,31,32]; Ethical, Legal, and Social Issues: Jasper A. Bovenberg[33]

Steering committee: Cisca Wijmenga[6,7] (Principal Investigator), Morris A. Swertz[6,7], Cornelia M. van Duijn[15], Dorret I. Boomsma[9], P. Eline Slagboom[21], Gertjan B. van Ommen[11], Paul I.W. de Bakker[1,3–5]

Affiliations:

1: Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands

2: Department of Computer Science, Columbia University, New York, NY, USA

3: Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands

4: Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

5: Broad Institute of Harvard and MIT, Cambridge, MA, USA

6: Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

7: Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

8: Department of Systems Biology, Columbia University, New York, NY, USA

9: Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands

10: Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands

11: Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

12: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

13: Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

14: Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands

15: Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands

16: Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands

17: SURFsara, Science Park, Amsterdam, The Netherlands

18: Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

19: Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands

20: The Genome Institute, Washington University, St. Louis, MO, USA

21: Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

22: Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

23: Department of Genetics, Harvard Medical School, Boston, MA, USA

24: Department of Genome Sciences, University of Washington, Seattle, WA, USA

25: Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands

26: European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

27: Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands

28: Rinat-Pfizer Inc, South San Francisco, CA, USA

29: BGI-Shenzhen, Shenzhen, China

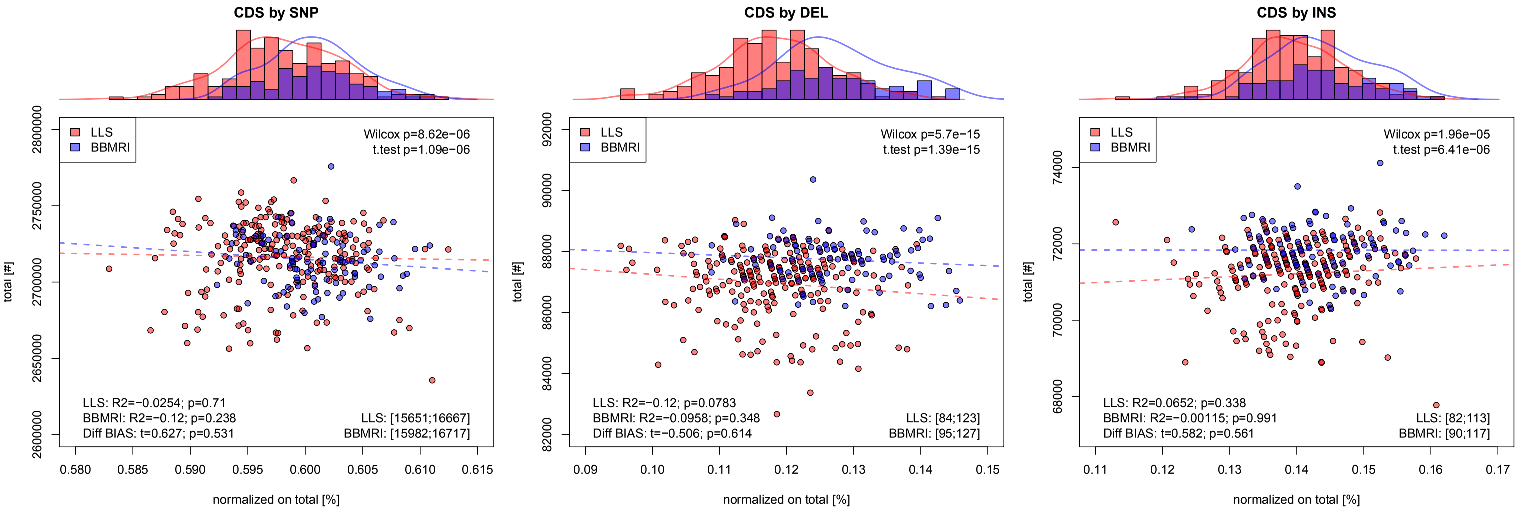
30: BGI-Europe, Copenhagen, Denmark

31: Department of Biology, University of Copenhagen, Copenhagen, Denmark

32: The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark

33: Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands

34: Deceased



Supplemental Figure 1: A sensitivity analysis on the observed differences in proportions of variants annotated to the CDS between long-lived cases and population controls with respect to the overall calling quality, proxied by overall calling rates. The proportions of variants annotated to the CDS are related to the total number of variants discovered within the long-lived cases (red) and random population controls (blue) for the variant types SNV, DEL and INS separately. Distributions of the percentages of variants annotated to CDS are displayed at the top of the figure for each variant type respectively. Test results for differences in these distributions are reported below (Wilcoxon Rank-Sum test and a Welch's t test). Raw data is also plotted in the scatter plot below to visually inspect the relation between proportions of variants annotated to the CDS (x-axis) and the absolute number of variants identified per sample. The correlation between the two, as illustrated by the dotted lines, is assessed per study and is compared between studies, reported in the low left corner. The minimal and maximal numbers of variants annotated to CDS are reported in the right lower corner for both studies separately. No significant biases were observed in these plots.

Type	Location	Impact	BBMRI	LLS	stat.W	*p*
DEL	CDS	FRAMESHIFT	40	32	19378	**5.53E-31**
INS	CDS	FRAMESHIFT	31.5	28	15414.5	**3.00E-10**
SNV	DONOR		358	354	14145	**4.05E-06**
SNV	CDS	MISSENSE	7182	7131	13846	**2.54E-05**
SNV	UTR5		2637.5	2625	13510	**1.67E-04**
SNV	CDS	SYNONYMOUS	8051	8026	13034	1.75E-03
DEL	CDS	DELETE	44	45	8894	0.02
SNV	UTR3		18785.5	18762.5	12194	0.04
SNV	CDS	NONSTOP	10	9	11983	0.08
INS	UTR3		793	791	9560	0.14
INS	CDS	INSERT	34	34	9572.5	0.14
SNV	CDS	NONSENSE	53	52	11766	0.15
DEL	UTR5		73	71	11757	0.15
SNV	TSS-UPSTREAM		104542.5	104644	11746	0.16
SNV	ACCEPTOR	DISRUPT	20	21	9629	0.16
SNV	ACCEPTOR		1733	1730	11734	0.16
INS	ACCEPTOR		60	61	9802	0.24
DEL	INTRON		35412.5	35148.5	9824	0.25
INS	UTR5		74	73	11402	0.34
SNV	DONOR	DISRUPT	34	35	11261	0.44
DEL	DONOR		20	20	11170	0.52
DEL	TSS-UPSTREAM		3787.5	3767	10260	0.57
INS	DONOR		13	13	11015.5	0.66
DEL	UTR3		920	912	10363	0.67
INS	INTRON		28921.5	28725	10381	0.69
SNV	INTRON		1036737	1037931	10449	0.76
DEL	ACCEPTOR		100	99	10454	0.76
SNV	CDS	MISSTART	16	15.5	10838	0.84
INS	TSS-UPSTREAM		3199	3183.5	10575	0.89

Supplemental Table 1: Median counts per variant category. Median counts of variants observed per variant category in long-lived cases (LLS) and population controls (BBMRI). Comparisons with median counts < 10 for both long-lived cases (LLS) as population controls (BBMRI) were not considered (NONSTOP SNV). Differences between distributions of counts normalized on totals per gvarType were tested using the Wilcoxon Rank-Sum test.

AssayID	Chrom	Start	End	Type	Ref	Alt	Carrier
LLS_01	chr2	113479780	113479780	INS		C	FAILED
LLS_02	chr11	120198285	120198287	DEL	CT		Yes
LLS_03	chr19	16976264	16976264	INS		G	FAILED
LLS_04	chr17	62038698	62038698	INS		C	No
LLS_05	chr17	7323945	7323946	SUB	C	AA	Yes
LLS_06	chr1	151372104	151372104	INS		C	No
LLS_7	chr4	95496888	95496888	INS		A	No
LLS_08	chr2	27730169	27730169	INS		A	Yes
LLS_09	chr10	55568865	55568867	DEL	TG		Yes
LLS_10	chr17	5404003	5404004	DEL	A		FAILED
LLS_11	chr9	90500990	90500992	DEL	CT		Yes
LLS_12	chr9	134398412	134398412	INS		G	Yes
LLS_12	chr9	134398412	134398412	INS		G	Yes
LLS_13	chr13	113980131	113980135	DEL	AAAC		Yes
LLS_13	chr13	113980131	113980135	DEL	AAAC		No
LLS_14	chr7	76828864	76828867	SUB	GAC	AGGT	No
LLS_15	chr17	41174273	41174274	SUB	T	AA	No

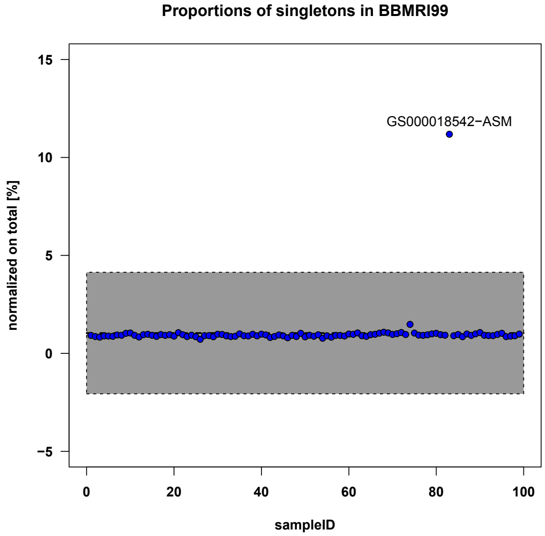
Supplemental Table 2: Sanger sequencing experiments on frameshift variants identified within the long-lived cases. Of the 15 independent assays designed, 12 returned good data, which confirmed the presence of 7 variants.

AssayID	Chrom	Start	End	Type	Ref	Alt	Carrier
BBMRI_01	chr3	121208840	121208840	INS		T	No
BBMRI_02	chr2	11696893	11696897	DEL	GAAG		Yes
BBMRI_03	chr10	118305625	118305629	SUB	TCAC	GGACT	No
BBMRI_04	chr1	100207825	100207825	INS		T	No
BBMRI_05	chr22	37964284	37964285	DEL	G		No
BBMRI_06	chr7	134719554	134719554	INS		C	No
BBMRI_07	chr9	35738865	35738865	INS		A	No
BBMRI_08	chr13	97639501	97639501	INS		AAGAAGGTCATCT	Yes
BBMRI_09	chr18	29122734	29122734	INS		G	No
BBMRI_10	chr18	55322554	55322555	SUB	C	AA	No
BBMRI_11	chr1	11008274	11008275	DEL	C		No
BBMRI_12	chr16	46695701	46695702	DEL	G		No
BBMRI_13	chr9	113457714	113457714	INS		A	No
BBMRI_14	chr4	122741747	122741748	DEL	A		No
BBMRI_15	chr17	7369290	7369291	DEL	C		No

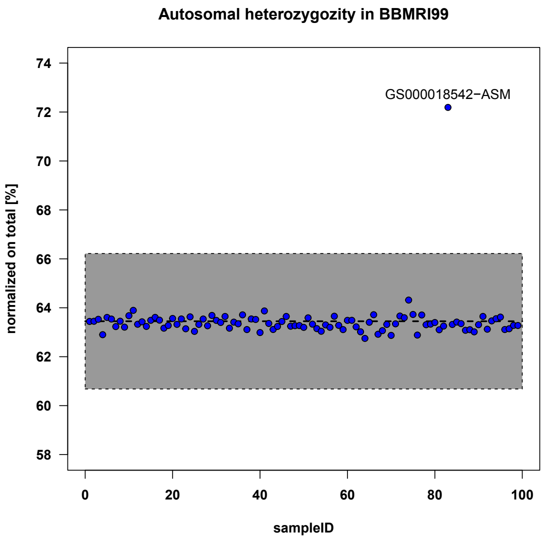
Supplemental Table 3: Sanger sequencing experiments on frameshift variants identified within the population controls. Of the 15 independent assays designed only 2 confirmed the presence of the targeted variant.

Gene	Chrom	Position	Type	Ref	Alt	Impact	Allele Counts (Alt|Ref)
TET2	chr4	106156278	DEL	G		FRAMESHIFT	1|8251
TET2	chr4	106156687	SNV	C	T	NONSENSE	1|8599
TET2	chr4	106157504	DEL	C		FRAMESHIFT	1|8251
TET2	chr4	106157506	SNV	C	T	NONSENSE	1|8597
TET2	chr4	106157653	SNV	G	T	NONSENSE	1|8599
TET2	chr4	106157700	SNV	T	G	NONSENSE	1|8599
TET2	chr4	106157807	DEL	C		FRAMESHIFT	3|8251
TET2	chr4	106158113	DEL	G		FRAMESHIFT	1|8253
TET2	chr4	106158157	SNV	C	T	NONSENSE	1|8599
TET2	chr4	106158441	DEL	C		FRAMESHIFT	21|8233
DNMT3A	chr2	25459834	SNV	C	A	NONSENSE	1|8599
DNMT3A	chr2	25466830	DEL	T		FRAMESHIFT	1|8115
DNMT3A	chr2	25467468	SNV	G	C	NONSENSE	1|8599
DNMT3A	chr2	25468163	SNV	C	A	NONSENSE	1|8599
DNMT3A	chr2	25468917	DEL	TCGTACA		FRAMESHIFT	20|8234
DNMT3A	chr2	25469529	DEL	C		FRAMESHIFT	12|8226
DNMT3A	chr2	25471030	DEL	GGCT		FRAMESHIFT	69|8185

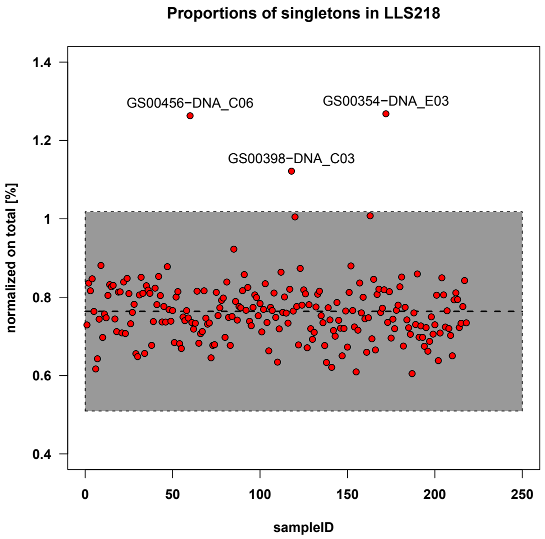
Supplemental Table 4: Frameshift and nonsense variants in *TET2* and *DNMT3A* on Exome Variant Server. Variants were called against the reference transcript NM_017628.4 and NM_022552.4 for *TET2* and *DNMT3A* respectively.



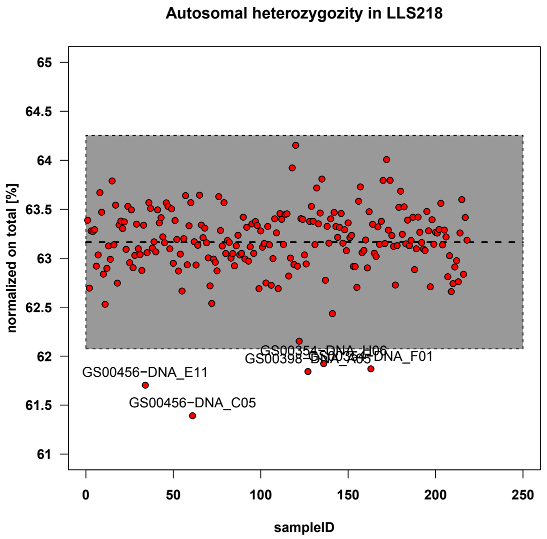
Supplemental Figure 2: Singleton Check BBMRI. Depicted are the proportions of singletons (SNVs unique for one sample) and overall numbers of identified SNVs per sample within the BBMRI study. The grey area marks the 3 SD thresholds, indicating that sample GS0000018542-ASM has a disproportionately high number of variants not observed in the rest of the study, suggesting either a distinct ancestry or a sample contamination.



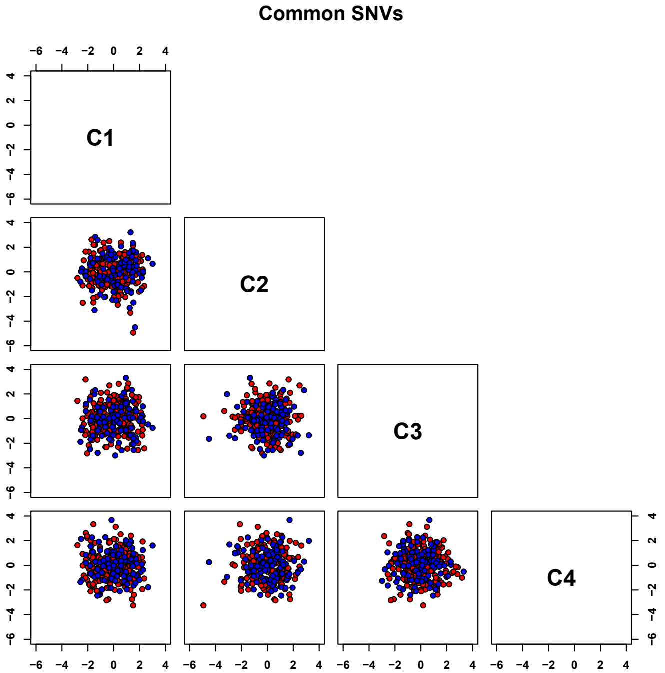
SUPPLEMENTAL FIGURE 3: AUTOSOMAL HETEROZYGOZITY BBMRI. The proportions of heterozygous SNV genotypes and overall numbers of identified SNVs per sample within the BBMRI study. Again the grey area marks the 3 SD thresholds, indicating again that sample GS0000018542-ASM exhibits a genomic make up that is very distinct from the remaining participants of the BBMRI study. Such an elevated heterozygozity again points to either a distinct ancestry or a sample contamination. Due to the consistent appearance of sample GS0000018542-ASM as a major outlier, we decided to remove it from further analyses.



SUPPLEMENTAL FIGURE 4: SINGLETON CHECK LLS. Depicted are the proportions of singletons (SNVs unique for one sample) and overall numbers of identified SNVs per sample within the LLS study. The 3 SD deviation of the expectation is indicated in grey. Slightly elevated proportions of unique SNVs are observed for GS00456-DNA_C06, GS00354-DNA_E03 and GS00398-DNA_CO3.



Supplemental Figure 5: Autosomal Heterozygozity LLS. The proportions of heterozygous SNV genotypes and overall numbers of identified SNVs per sample within the LLS study. The 3 SD deviation of the expectation is indicated in grey. Slightly lowered proportions of heterozygous SNVs are observed for GS00354-DNA_H06, GS00354-DNA_F01, GS00456-DNA_E11, GS00456-DNA_C05 and GS00398-DNA_A05. Noteworthy is that none of the samples overlapped with the outliers that came forward in the singleton check. Various types of artefacts such as mixed ancestry, sample pollution, variation in total read depth or just biological variation might explain slight deviations in both the singleton and heterozygozity proportions. However, since outliers where not consistently picked up in both tests, we decided not to exclude any samples.



Supplemental Figure 6: Multidimensional Scaling. MDS was performed with Plink using 10,000 randomly selected common SNVs (MAF ≥ 5%) to inspect the data for signs of differences in population substructure. Sample space was reduced to four dimensions and all combinations thereof are plotted. Long-lived cases are displayed in red, population controls in blue. No apparent substructure was observed.