The handle http://hdl.handle.net/1887/32015 holds various files of this Leiden University dissertation.

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 Akker¹,², Steven J. Pitts³, Joris Deelen¹,⁴, Matthijs H. Moed¹, Shobha Potluri³, H. Eka D. Suchiman¹, Nico Lakenberg¹, Wesley J. de Dijcker¹, Anton J.M. de Craen⁵, Jeanine J. Houwing-Duistermaat⁶, Genome of the Netherlands Consortium⁷, David R. Cox³†, Marian Beekman¹,⁴, Marcel J.T. Reinders², P. Eline Slagboom¹,⁴

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, yet striking variations in life span are observed among the population at large. 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. The propensity to become long-lived nevertheless clearly runs in families 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, 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), nor did GWASs for longevity revealed sufficient loci to explain the heritability of longevity. 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. 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. Secondly, a targeted knockdown of a single gene in model organisms can already give rise to a long-lived species. 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. 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. Using a more targeted approach, 988 candidate longevity genes were sequenced in 6 centenarians to identify novel non-synonymous SNVs, 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. 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. 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.

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, \(N_{\text{male}} = 82\) (37.6%)) and 98 population controls of the BBMRI biobanking initiative (median age 57, \(N_{\text{male}} = 39\) (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
Germ line and somatic characteristics of the long-lived genome

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: $p=5.09 \times 10^{-5}$, DEL: $p=4.15 \times 10^{-15}$ and INS: $p=1.62 \times 10^{-5}$; 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.

When applying the testing to the more fine-grained annotations of the coding sequence, as provided by Complete Genomics\textsuperscript{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 (\( p = 1.84\times10^{-26} \)) and INSs (\( p = 2.60\times10^{-09} \)) and SNVs residing in splice donor sites

---

**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.
(\(p = 4.51 \times 10^{-04}\)) 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.](image)

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 (\(p = 1.84 \times 10^{-26}\)) and INSs (\(p = 2.60 \times 10^{-09}\)) and SNVs residing in splice donor sites (\(p = 4.51 \times 10^{-04}\)) 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
**Gene Symbol** | **EntrezGeneID** | **DEL** | **INS** | **DEL + INS** | **p.perm** |
--- | --- | --- | --- | --- | --- |
**TTN** | 7273 | 7 (2/5/0) | 2 (0/2/0) | 9 (2/7/0) | <1.00 × 10⁻⁶ |
**DNAH14** | 127602 | 8 (4/2/2) | 0 (0/0/0) | 8 (4/2/2) | <1.00 × 10⁻⁶ |
**FSIP2** | 401024 | 1 (1/0/0) | 6 (3/2/1) | 7 (4/2/1) | <1.00 × 10⁻⁶ |
**LOC100506072** | 100506072 | 3 (2/0/1) | 4 (1/2/1) | 7 (3/2/2) | <1.00 × 10⁻⁶ |
**TET2** | 54790 | 4 (4/0/0) | 2 (2/0/0) | 6 (6/0/0) | 5.00 × 10⁻⁵ |
**SSPO** | 23145 | 5 (2/0/3) | 1 (1/0/0) | 6 (3/0/3) | 6.00 × 10⁻⁵ |
**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 × 10⁻⁶ |
**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 |
**PIK3CG2** | 5288 | 3 (3/0/0) | 1 (0/1/0) | 4 (3/1/0) | 1.91 × 10⁻⁴ |
**TNRC18** | 84629 | 3 (2/1/0) | 1 (0/1/0) | 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).

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Type</th>
<th>Ref</th>
<th>Alt</th>
<th>Impact</th>
<th>LLS</th>
<th>BBMRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106155736</td>
<td>106155737</td>
<td>DEL</td>
<td>T</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106155765</td>
<td>106155766</td>
<td>DEL</td>
<td>G</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106156685</td>
<td>106156686</td>
<td>SNV</td>
<td>C</td>
<td>A</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106156758</td>
<td>106156758</td>
<td>INS</td>
<td>-</td>
<td>C</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106157246</td>
<td>106157246</td>
<td>INS</td>
<td>-</td>
<td>A</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106157781</td>
<td>106157782</td>
<td>DEL</td>
<td>G</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106158107</td>
<td>106158108</td>
<td>SNV</td>
<td>G</td>
<td>A</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106196212</td>
<td>106196213</td>
<td>SNV</td>
<td>C</td>
<td>T</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106196221</td>
<td>106196222</td>
<td>SNV</td>
<td>G</td>
<td>T</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{TET2}</td>
<td>chr4</td>
<td>106197352</td>
<td>106197353</td>
<td>DEL</td>
<td>A</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25463181</td>
<td>25463182</td>
<td>SNV</td>
<td>G</td>
<td>A</td>
<td>NONSENSE</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25463296</td>
<td>25463296</td>
<td>INS</td>
<td>-</td>
<td>A</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25468153</td>
<td>25468154</td>
<td>DEL</td>
<td>G</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25468921</td>
<td>25468923</td>
<td>DEL</td>
<td>AC</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25469921</td>
<td>25469922</td>
<td>SNV</td>
<td>A</td>
<td>G</td>
<td>NONSENSE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25469990</td>
<td>25469991</td>
<td>DEL</td>
<td>A</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\textit{DNMT3A}</td>
<td>chr2</td>
<td>25470930</td>
<td>25470931</td>
<td>DEL</td>
<td>G</td>
<td>-</td>
<td>FRAMESHIFT</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

From the whole genome, frameshift variants identified within \textit{TET2} and \textit{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 \textit{TET2} and \textit{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 five nonsense SNVs and \textit{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 \textit{TET2} and \textit{DNMT3A} were hit with unique frameshift indels or nonsense SNVs with a significantly lower frequency (\textit{TET2}: \( N_{\text{disrupt,EVS}} = 9 \), OR: 24.2 95% CI: 9.0-67.0, \( p = 4.5 \times 10^{-10} \); \textit{DNMT3A}: \( N_{\text{disrupt,EVS}} = 7 \), OR: 19.5 95% CI: 5.8-65.6, \( p = 1.9 \times 10^{-6} \), Fisher’s Exact tests, Supplemental Table 4).
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* (\(\beta=1.29\), 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 (\(\beta=1.29\), 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* (\(N_{\text{tot}}=214, N_{\text{death}}=190, \text{HR}=1.30, 95\% \text{ CI } 0.68-2.47, p=0.424\)) or *DNMT3A* (\(N_{\text{tot}}=214, N_{\text{death}}=190, \text{HR}=0.37, 95\% \text{ CI } 0.15-0.91, p=0.031\), 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.
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 \textit{TET2} and \textit{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 \textit{TET2} and \textit{DNMT3A}. Somatic mutations in \textit{TET2} and \textit{DNMT3A} were first reported in patients suffering from myeloid malignancies\cite{26,27}, but also appear in elderly exhibiting myelodysplasia without overt hematopoietic malignancies\cite{23}. This suggests that somatic mutations in \textit{TET2} and \textit{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 \textit{DNMT3A}. Surprisingly, neither the carriers of somatic disruptive mutations in \textit{DNMT3A}, nor in \textit{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)\cite{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 \textit{TET2} and \textit{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 \textit{TET2} and \textit{DNMT3A} are factors for epigenetic control\cite{28,29} and are thought to silence hematopoietic stem cell self-renewal to permit efficient hematopoietic differentiation\cite{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\cite{32}. Hence on condition that the enhanced self-renewal, instigated by somatic disruptive mutations in \textit{TET2} and \textit{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 \textit{TET2} and \textit{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\cite{33,34} and poor prognosis of AML\cite{27,35}. In the long-lived cases of our study, however, disruptive somatic variants in \textit{TET2} and \textit{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 \textit{TET2} and \textit{DNMT3A} might not be accurate for the oldest old. Secondly, elderly carrying the somatic disruptive mutations in \textit{TET2} and \textit{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 \textit{TET2} or \textit{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 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. 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 (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_male = 82 (37.6%)) and 98 population controls (median age 57, N_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_D unique frameshift deletions and k_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, D+ k_j, I | p_D, p_I) we repeatedly resampled (Z=1,000,000) k_D deletions and k_I insertions with prior probabilities p_D and p_I and counted the number of times where the resampled numbers of frameshift variants k_S equaled or exceeded the number of observed frameshift variants k_j, yielding k^S_j. The estimated p-value is then obtained using:

\[
\hat{p}(K_j > k_{j,0} + k_{j,1}, p_{0}, p_{1}) = \frac{\sum_{i}(k_j^i)+1}{(Z+1)}
\]

Computations were performed in R 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 \( k^p_{j,\partial} \) private frameshift deletions and \( k^p_{j,i} \) 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 \((p_{\text{del},\text{case}} = 814/2,193 = 0.370)\) or population controls \((p_{\text{del},\text{ctr}} = 1,122/2,193 = 0.512)\) and a frameshift insertion to be exclusively observed in long-lived cases \((p_{\text{ins},\text{case}} = 896/1,764 = 0.508)\) or population controls \((p_{\text{ins},\text{ctr}} = 729/1,764 = 0.413)\). 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 \( k^p_\partial \) and \( k^p_i \) 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 \( p_{\text{obs}}(k^p_j \geq k^p_{j,\partial} + k^p_{j,i} | p_{\text{del,case}}, p_{\text{del,ctr}}, p_{\text{ins,case}}, p_{\text{ins,ctr}}) \) for a given gene \( j \) by repeatedly resampling \((Z=1,000,000)\) \( k^p_\partial \) deletions and \( k^p_i \) insertions with prior probabilities \( p_{\text{del,case}}, p_{\text{del,ctr}}, p_{\text{ins,case}}, \) and \( p_{\text{ins,ctr}} \) for respectively obtaining private deletions \( (k^{\partial}_{j,\partial}) \) and insertions \( (k^{\partial}_{j,i}) \) in cases and controls for each sampling and subsequently counted the number of times the number of sampled private mutations \( k^{\partial}_{j} \) equaled or exceeded the observed number of private mutations \( k^{\partial}_{j} \) \( (k^{\partial}_{j,\partial}) \). The p-value was then estimated by:

\[
\hat{p}(k^p_j > k^p_{j,\partial} + k^p_{j,i} | p_{\text{del,case}}, p_{\text{del,ctr}}, p_{\text{ins,case}}, p_{\text{ins,ctr}}) = \frac{\sum (H(k^{p}_{j}) + 1)}{(Z + 1)}
\]

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 \( \lambda = E \times R_{\text{var}} \) and \( K = R_{\text{var}} \) is used to estimate the probability \( p_{\text{hom}} \) that a variant, called with reads \( R_{\text{var}} \) of which at least \( R_{\text{var}} \), support the variant, is likely to comprise a homozygous reference variant with some noisy reads. Similarly, we employ a Poisson model with mean \( \lambda = 0.5 \times R_{\text{var}} \) and \( K = R_{\text{var}} \) to estimate the probability \( p_{\text{het}} \) that the alternative allele, supported by \( R_{\text{var}} \) 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: \( p_{\text{sam}} = \max(p_{\text{hom}}, p_{\text{het}}) \).

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 \texttt{lm} package of the statistical language R\textsuperscript{35}:

\[
G \sim \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{carrier}
\] (3)

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

5.8 Associations with prospective survival

Associations with prospective survival were performed with the \texttt{Survival} package\textsuperscript{36} of R\textsuperscript{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:

$$\lambda(t) \sim \lambda_0(t) \times \exp(\beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{carrier})$$ (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

Supplemental Materials

Genome of the Netherlands Consortium members:


Cohort collection and sample management group: P. Eline Slagboom, Marian Beekman, Anton J.M. de Craen, H. Eka D. Suchiman, Albert Hofman, Cornelia van Duijn, Dorret I. Boomsma, Gonneke Willemsen, Bruce H. Wolfftenbuttel, Mathieu Platteel, Steven J. Pitts, Shobha Putluri, David R. Cox.

Whole-genome sequencing: Qibin Li, Yingrui Li, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Ning Li, Sujie Cao, Jun Wang, Ning Li, Ethical, Legal, and Social Issues: Jasper A. Bovenberg.

Steering committee: Cisca Wijmenga (Principal Investigator), Morris A. Swertz, Cornelia M. van Duijn, Dorret I. Boomsma, P. Eline Slagboom, Gertjan B. van Ommen, Paul I.W. de Bakker.

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
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 lower 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.
### Supplemental Table 1: Median counts per variant category

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

*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.
### 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.

<table>
<thead>
<tr>
<th>AssayID</th>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Type</th>
<th>Ref</th>
<th>Alt</th>
<th>Carrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLS_01</td>
<td>chr2</td>
<td>113479780</td>
<td>113479780</td>
<td>INS</td>
<td>C</td>
<td></td>
<td>FAILED</td>
</tr>
<tr>
<td>LLS_02</td>
<td>chr11</td>
<td>120198285</td>
<td>120198287</td>
<td>DEL</td>
<td>CT</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_03</td>
<td>chr19</td>
<td>16976264</td>
<td>16976264</td>
<td>INS</td>
<td>G</td>
<td></td>
<td>FAILED</td>
</tr>
<tr>
<td>LLS_04</td>
<td>chr17</td>
<td>62038698</td>
<td>62038698</td>
<td>INS</td>
<td>C</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>LLS_05</td>
<td>chr17</td>
<td>7323945</td>
<td>7323946</td>
<td>SUB</td>
<td>C</td>
<td>AA</td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_06</td>
<td>chr1</td>
<td>151372104</td>
<td>151372104</td>
<td>INS</td>
<td>C</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>LLS_07</td>
<td>chr4</td>
<td>95496888</td>
<td>95496888</td>
<td>INS</td>
<td>A</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>LLS_08</td>
<td>chr2</td>
<td>27730169</td>
<td>27730169</td>
<td>INS</td>
<td>A</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_09</td>
<td>chr10</td>
<td>55568865</td>
<td>55568867</td>
<td>DEL</td>
<td>TG</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_10</td>
<td>chr17</td>
<td>5404003</td>
<td>5404004</td>
<td>DEL</td>
<td>A</td>
<td></td>
<td>FAILED</td>
</tr>
<tr>
<td>LLS_11</td>
<td>chr9</td>
<td>90500990</td>
<td>90500992</td>
<td>DEL</td>
<td>CT</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_12</td>
<td>chr9</td>
<td>134398412</td>
<td>134398412</td>
<td>INS</td>
<td>G</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_12</td>
<td>chr9</td>
<td>134398412</td>
<td>134398412</td>
<td>INS</td>
<td>G</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_13</td>
<td>chr13</td>
<td>113980131</td>
<td>113980135</td>
<td>DEL</td>
<td>AAAC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>LLS_13</td>
<td>chr13</td>
<td>113980131</td>
<td>113980135</td>
<td>DEL</td>
<td>AAAC</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>LLS_14</td>
<td>chr7</td>
<td>76828864</td>
<td>76828867</td>
<td>SUB</td>
<td>GAC AGGT</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>LLS_15</td>
<td>chr17</td>
<td>41174273</td>
<td>41174274</td>
<td>SUB</td>
<td>T</td>
<td>AA</td>
<td>No</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>AssayID</th>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Type</th>
<th>Ref</th>
<th>Alt</th>
<th>Carrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBMRI_01</td>
<td>chr3</td>
<td>121208840</td>
<td>121208840</td>
<td>INS</td>
<td>T</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_02</td>
<td>chr2</td>
<td>11696893</td>
<td>11696897</td>
<td>DEL</td>
<td>GAAG</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>BBMRI_03</td>
<td>chr10</td>
<td>118305625</td>
<td>118305629</td>
<td>SUB</td>
<td>TCAC GGACT</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_04</td>
<td>chr1</td>
<td>100207825</td>
<td>100207825</td>
<td>INS</td>
<td>T</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_05</td>
<td>chr22</td>
<td>37964284</td>
<td>37964285</td>
<td>DEL</td>
<td>G</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_06</td>
<td>chr7</td>
<td>134719554</td>
<td>134719554</td>
<td>INS</td>
<td>C</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_07</td>
<td>chr9</td>
<td>35738865</td>
<td>35738865</td>
<td>INS</td>
<td>A</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_08</td>
<td>chr13</td>
<td>97639501</td>
<td>97639501</td>
<td>INS</td>
<td>AAGAAGTCATCT</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>BBMRI_09</td>
<td>chr18</td>
<td>29122734</td>
<td>29122734</td>
<td>INS</td>
<td>G</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_10</td>
<td>chr18</td>
<td>55322554</td>
<td>55322555</td>
<td>SUB</td>
<td>C</td>
<td>AA</td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_11</td>
<td>chr1</td>
<td>11008274</td>
<td>11008275</td>
<td>DEL</td>
<td>C</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_12</td>
<td>chr16</td>
<td>46695701</td>
<td>46695702</td>
<td>DEL</td>
<td>G</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_13</td>
<td>chr9</td>
<td>113457714</td>
<td>113457714</td>
<td>INS</td>
<td>A</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_14</td>
<td>chr4</td>
<td>122741747</td>
<td>122741748</td>
<td>DEL</td>
<td>A</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>BBMRI_15</td>
<td>chr17</td>
<td>7369290</td>
<td>7369291</td>
<td>DEL</td>
<td>C</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>
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 GS000018542-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 Table 4: Frameshift and nonsense variants in \textit{TET2} and \textit{DNMT3A} on Exome Variant Server. Variants were called against the reference transcript NM_017628.4 and NM_022552.4 for \textit{TET2} and \textit{DNMT3A} respectively.
**Supplemental Figure 3: Autosomal Heterozygosity BBMRI.** The proportions of heterozygous SNV genotypes and overall numbers of identified SNVs per sample within the BBMRI study. Again in 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 heterozygosity 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_C03.
Supplemental Figure 5: Autosomal Heterozygosity 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 heterozygosity 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.