Supplementary material

Ageing as Accelerated Accumulation of Somatic Variants? Whole Genome Sequencing of Centenarian and Middle-Aged Monozygotic Twin Pairs

Kai Ye, Marian Beekman, Eric-Wubbo Lameijer, Yanju Zhang, Matthijs H. Moed, Erik B. van den Akker, Joris Deelen, Jeanine J. Houwing-Duistermaat, Dennis Kremer, Seyed Yahya Anvar, Jeroen F. J. Laros, David Jones, Keiran Raine, Ben Blackburne, Shobha Potluri, Quan Long, Victor Guryev, Ruud van der Breggen, Rudi G. J. Westendorp, Peter A. C. 't Hoen, Johan den Dunnen, Gertjan B. van Ommen, Gonneke Willemsen, Steven J. Pitts, David R. Cox, Zemin Ning, Dorret I. Boomsma, P. Eline Slagboom

Sequencing at Illumina

The Genomic DNA from the twin samples was sent to Illumina for whole-genome sequencing, 100 bp reads and 500 bp insert size. LLS twin 1 had 96.5% of the genome covered at ≥1x, 94.14% of the genome covered at ≥20x and 66.69% covered at ≥ 40x while LLS twin 2 had 96.50%, 94.25% and 72.83%, respectively. NTR twins have similar coverage profiles. Out of the called genome, 4.03 million variants (includes SNPs, indels, structural variants and CNVs) were identified for sample 2749 and 4.09 million variants were identified for LLS twin 2. The breakdown of the variants into corresponding variant types is listed in Table S1.

Table S1

Illumina Data Property

	LLS		NTR	
	Twin 1	Twin 2	Twin 1	Twin 2
Total number of reads	1,250,627,972	1,308,065,757	1,480,010,055	1,197,180,566
Mapped	1,246,271,633 (99.65%)	1,304,820,322 (99.75%)	1,474,782,256 (99.65%)	1,192,686,225 (99.62%)
Paired	1,250,627,972	1,308,065,757	1,480,010,055	1,197,180,566
Read 1	625,258,075	654,048,606	739,986,055	598,815,816
Read 2	625,369,897	654,017,151	740,024,000	598,364,750
Properly paired	1,241,553,576 (99.27%)	1,300,839,952 (99.45%)	1,469,136,792 (99.27%)	1,187,917,342 (99.62%)
With itself and mate mapped	1,243,579,380 (99.44%)	1,303,004,277 (99.61%)	1,471,778,519 (99.44%)	1,189,865,659 (99.39%)
Singletons	2,692,253 (0.22%)	1,816,045 (0.14%)	3,003,737 (0.20%)	2,820,566 (0.24%)
With mate mapped to a different chr	1,390,434 (0.11%)	1,470,041 (0.11%)	1,876,729 (0.12%)	1,253,051 (0.10%)
With mate mapped to a different chr (mapQ ≥ 5)	1,123,127 (0.09%)	1,178,112 (0.09%)	1,550,958 (0.10%)	992,081 (0.83%)

Sequencing at Complete Genomics

Genomic DNA from the twin samples was sent to Complete Genomics for Standard coverage whole-genome sequencing. LLS twin 1 had 97.2% of the genome covered at ≥20x and 97.6% of the genome either fully or partially called by Complete Genomics using their analysis pipeline (Drmanac et al., 2010). LLS twin 1 had 95.2% of the genome covered at ≥20x and 97.4% of the genome either fully or partially called. Regions of the genome that do not have sufficient evidence (Drmanac et al., 2010) to make a reference or a variant call are made no-call.

Out of the called genome, 3.96 million variants (includes SNPs, indels, substitutions) were identified for LLS twin 1 and 3.88 million variants were identified for LLS twin 2. The breakdown of the variants into corresponding variant types per sample is listed in Table S2 (Copy Number Variations [CNV] and Structural Variants [SV] identified are also listed). In order to look for concordance of variant calls between the samples, the callDiff program from cgatools was used which groups variants into 'super loci' and looks for consistency. There were 3.94 million super locus variant sites across the two samples; 86.7% of the super loci were identical between the twins, leading to 525,105 loci that have mismatched calls (either called in both samples and discordant, or called in one sample).

Table S2

Complete Genomics Data Property

	LLS		NTR	
	Twin 1	Twin 2	Twin 1	Twin 2
Gross mapping yield (Gb)	172.877	211.373	172.962	187.974
Both mates mapped yield (Gb)	139.954	169.782	142.261	152.741
Genome coverage ≥ 5	99.4%	99.5%	98.6%	98.9%
Genome coverage ≥ 10	98.7%	99.0%	96.0%	97.3%
Genome coverage ≥ 20	95.2%	97.2%	87.0%	90.9%
Genome coverage ≥ 30	86.5%	92.7%	73.9%	80.3%
Genome coverage ≥ 40	70.5%	83.6%	59.0%	66.5%
Exome coverage ≥ 5	97.4%	98.3%	98.5%	98.5%
Exome coverage ≥ 10x	94.7%	97.1%	97.0%	97.2%
Exome coverage ≥ 20	85.5%	93.0%	91.1%	92.2%
Exome coverage ≥ 30	71.0%	85.8%	81.4%	83.5%
Exome coverage ≥ 40	51.9%	74.0%	69.5%	71.5%

Substitution Variant Calling From Illumina Data

CaVEMan (Cancer Variants through Expectation Maximisation), a bespoke Java application using a simple expectation maximisation algorithm implementation, was used to call single nucleotide substitutions. Through comparison of reads from both twins with the reference genome, CaVEMan calculates a probability for each possible genotype per base. In order to provide more accurate estimates of sequence error rates within the algorithm, thus aid identification of true variants, variables such as base quality, read position, lane, and read orientation are incorporated into the calculations. Once CaVEMan was run, several post processing filters were applied in order to further increase the specificity of somatic mutation calls.

1. At least one-third of mutant alleles in reads are of quality ≥25.
2. At least one mutant allele in must fall in the middle third of the read, unless the read depth is less than 10, when a mutant allele in the first third is acceptable.
3. There is no more than one high quality (≥20) mutant allele in a read from the other sample in comparison.

Indel and Structural Variant Calling From Illumina Data

An improved version of Pindel[9] (https://trac.nbic.nl/pindel/) was used to call insertions and deletions. By modifying the input file generation process we were able to increase sensitivity and increase confidence in events detected by BWA, which was used as the initial mapping tool. The accepted approach for generating input for Pindel is to provide all read pairs where one end is unmapped and the other is confidently mapped to the genome, an anchor read. We found that by including read pairs where both ends map to the genome, but allowing for one of the pair to have mismatches, insertions or deletions, we could greatly increase coverage over smaller events (in some cases both ends are used as an anchor, creating two input records). The majority of these small events are detected by the BWA mapping algorithm; however, this increases confidence that the events are worth investigating.

Heteroplasmy of Mitochondria DNA

We tried to identify heteroplasmy by counting the coverage ratio between alternative allele and reference allele and found 18 potential heteroplasmy sites. We observed that the allele frequencies of potential heteroplasmic allele are quite similar between twins (corr = 0.984). However, after we manually examined those sites we consider them as mapping artifact, which may be caused by nuclear-mitochondria (NUMT) insertions.

Table S3

Potential Heteroplasmy Sites of Mitochondrial DNA of Two MZ Twin Pairs

		LLS		NTR	
Location	Alleles (Ref/Alt)	Twin 1	Twin 2	Twin 1	Twin 2
1082.0000	A/C	0.0894	0.1019	0.1014	0.0971
1132.0000	T/C	0.0589	0.0948	0.0908	0.0826
5705.0000	A/C	0.0651	0.0559	0.0801	0.0674
6278.0000	T/G	0.0558	0.0715	0.0753	0.0592
6355.0000	A/C	0.0508	0.0590	0.0536	0.0569
6819.0000	A/C	0.1497	0.2339	0.2443	0.2367
7712.0000	T/C	0.1098	0.1492	0.0013	0.0000
10239.0000	A/C	0.0591	0.0534	0.0711	0.0660
13681.0000	A/C	0.0507	0.0731	0.0755	0.0669
15759.0000	T/C	0.1851	0.1702	0.0000	0.0000
152.0000	T/C	0.9983	1.0000	0.0542	0.0600
316.0000	G/C	0.0367	0.0215	0.0626	0.0723
6822.0000	T/C	0.0368	0.0842	0.1008	0.0876
8274.0000	C/T	0.0003	0.0009	0.0770	0.0674
8275.0000	C/A	0.0006	0.0000	0.1077	0.1011
8278.0000	C/G	0.0000	0.0000	0.3234	0.3040
9765.0000	A/C	0.0403	0.0602	0.0638	0.0544
12812.0000	A/C	0.0450	0.0788	0.0771	0.0899
13937.0000	A/C	0.0281	0.0483	0.0457	0.0546

Validation of Potential Somatic Substitutions Using Ion Torrent Sequencing

Pooled DNA samples were enriched for 13 target regions that carry potential somatic substitutions. Primers were designed for fragments of approximately 300bp in size, harboring a given mutation. The protocol that was used for designing the primers is as follows: Indexed libraries for resequencing were prepared according to the manufacturer's instruction (Life Technologies). Libraries with index 1 and 2 (LLS twin 1 and twin 2, respectively) were mixed and clustered together for target enrichment sequencing. Sequencing was performed on Ion Personal Genome Machine (PGM) semiconductor sequencer. Samples were sequenced using 65 cycles of incorporation on a single 314R chip. Primary analysis and quality control were performed using the Ion Torrent suite 1.5 from Life Technologies. The fastq files were then aligned against HG19 using TMAP version 0.1.3-1. SAMtools version 0.1.16 was used to generate the SAM and BAM output files, mark duplicates in the resulting sample BAM files, and assess the coverage depth. A two-step approach was applied to test for the presence of imputed SNPs at every position. First, a simple likelihood ratio was used for the uniquely defined 22-mers that carry the somatic substitution. These subsequences were designed in such a way that the somatic substitution would be placed at the 11th base. Next, to provide a better estimation of the ratio in which these substitutions occur, custom scripts were used. The latter approach has the advantage of handling larger k-mers and potential mismatches that may fall within flanking sequences. Marker libraries were defined as two 15-mer sub-sequences that surround a targeted 5-mer, within which the SNP is positioned at the third base. We allowed one mismatch within the surrounding 15-mers to compensate for a potential insert, deletion, or substitution. By using a regular expression strategy, all the possible structural compositions of the targeted 5-mer regions and their counts

were reported. Consequently, likelihood ratios were calculated to identify SNPs that are most likely to be real somatic substitutions. In both analyses, strong evidence for somatic substitution was observed for 9 (8 of which was also confirmed by Sanger sequencing) out of initial 13 variants. We observed no other prominent mutations as the remaining imputed SNPs had no occurrence or the estimated ratio.