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

Kai Ye,[1] Marian Beekman,[1,2] Eric-Wubbo Lameijer,[1] Yanju Zhang,[1] Matthijs H. Moed,[1] Erik B. van den Akker,[1,3] Joris Deelen,[1,2] Jeanine J. Houwing-Duistermaat,[4] Dennis Kremer,[1] Seyed Yahya Anvar,[5] Jeroen F. J. Laros,[5] David Jones,[6] Keiran Raine,[6] Ben Blackburne,[7] Shobha Potluri,[8] Quan Long,[9] Victor Guryev,[10] Ruud van der Breggen,[1] Rudi G. J. Westendorp,[11] Peter A. C. 't Hoen,[12] Johan den Dunnen,[12] Gert Jan B. van Ommen,[12] Gonneke Willemsen,[13] Steven J. Pitts,[8] David R. Cox,[8] Zemin Ning,[6] Dorret I. Boomsma,[13,*] and P. Eline Slagboom[1,2,*]

[1]Molecular Epidemiology, Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
[2]Netherlands Consortium for Healthy Ageing, Leiden, the Netherlands
[3]The Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands
[4]Statistical Genetics, Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
[5]Center for Human and Clinical Genetics and Leiden Genome Technology Center, Leiden University Medical Center, Leiden, the Netherlands
[6]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[7]Faculty of Life Sciences, University of Manchester, Manchester, UK
[8]Pfizer Inc., San Francisco, CA, USA
[9]Gregor Mendel Institute, Vienna, Austria
[10]Hubrecht Institute, Utrecht, the Netherlands
[11]Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands
[12]Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
[13]Department of Biological Psychology, VU University, Amsterdam, the Netherlands

It has been postulated that aging is the consequence of an accelerated accumulation of somatic DNA mutations and that subsequent errors in the primary structure of proteins ultimately reach levels sufficient to affect organismal functions. The technical limitations of detecting somatic changes and the lack of insight about the minimum level of erroneous proteins to cause an error catastrophe hampered any firm conclusions on these theories. In this study, we sequenced the whole genome of DNA in whole blood of two pairs of monozygotic (MZ) twins, 40 and 100 years old, by two independent next-generation sequencing (NGS) platforms (Illumina and Complete Genomics). Potentially discordant single-base substitutions supported by both platforms were validated extensively by Sanger, Roche 454, and Ion Torrent sequencing. We demonstrate that the genomes of the two twin pairs are germ-line identical between co-twins, and that the genomes of the 100-year-old MZ twins are discerned by eight confirmed somatic single-base substitutions, five of which are within introns. Putative somatic variation between the 40-year-old twins was not confirmed in the validation phase. We conclude from this systematic effort that by using two independent NGS platforms, somatic single nucleotide substitutions can be detected, and that a century of life did not result in a large number of detectable somatic mutations in blood. The low number of somatic variants observed by using two NGS platforms might provide a framework for detecting disease-related somatic variants in phenotypically discordant MZ twins.

■ **Keywords:** aging, next gen sequencing, somatic mutations, twin discordances, human genome

Aging is a natural process of every life form and it intrigues researchers to investigate its driving force. In the 1960s it was suggested that aging is the outcome of accelerated accumulation of somatic mutations at DNA level (Curtis, 1966), hence introducing errors in the primary structure of proteins (Orgel, 1963, 1973). These theories predict that somatic mutations appear stochastically from the beginning

1026

of life onwards, and the accumulation of erroneous proteins and somatic mutations eventually affect key functions of somatic maintenance (synthesis, degradation, repair), leading to a cascade of detrimental consequences (error catastrophe). The lack of high throughput detection techniques, however, hampered any firm conclusions on these theories in humans (Kirkwood, 1977). The question here is whether current high throughput next-generation sequence technology can provide some of the necessary insights, by using a study design in which the whole-genome sequence is compared between co-twins from monozygotic (MZ) twin pairs.

MZ twins generally show remarkable similarity in diverse aspects of behavior, health, and disease (e.g., Martin et al., 1997; van Dongen et al., 2012). However, even for highly heritable traits, discordance within MZ pairs also has been documented (e.g., Zwijnenburg et al., 2010). A remarkable pair of MZ twins discordant for multiple sclerosis (MS) has been investigated, aiming at discordances at genome, epigenome, and transcriptome levels (Baranzini et al., 2010) by the first whole-genome sequencing study of CD4+ T cells. Potentially discordant variants between discordant MZ twins were detected but none of these were reproducible between methods. The result also illustrates the challenges of identification of low rate de novo mutations in next generation sequencing (NGS) data, which contain high rates of sequencing errors.

To examine the level of detectable somatic variants that may occur in a lifetime, whole-genome NGS was done on two randomly chosen MZ twin pairs of 40 and 100 years. The 40-year-old MZ twin pair is participating in the Netherlands Twin Registry (NTR co-twins) and the centenarian MZ twin pair in the Leiden Longevity Study (LLS co-twins). NGS data for these four individuals were generated using two different whole-genome sequencing platforms; by comparing within pair differences across the two platforms we aim to reduce the number of false positive hits, created, for example, by sequencing errors. To clarify whether an accumulation of somatic mutations with age can be detected in human blood, the NGS data were compared for within co-twin DNA sequencing differences in the younger and centenarian co-twins.

Materials and Methods
Study Samples
The twins of 40 years of age were randomly selected from the Netherlands Twin Registry (Willemsen et al., 2013), and the twins of 100 years of age from the Leiden Longevity Study (Westendorp et al., 2009). Genomic DNA was isolated according to standard procedures from blood (Willemsen et al., 2010). The study was approved by Leiden University Medical Center and the Medical Ethics Committee of the VU Medical Centre, Amsterdam. Informed, written

consent was obtained from the twins participating in the study.

Illumina Platform
Illumina Infinium 1M SNP-array and GWAS data indicated that the twins were monozygotic. Paired-end and 100 bp long reads (Bentley et al., 2008) were generated using Illumina GAIIx instruments. Illumina sequences were aligned to the NCBI human reference genome build 37, with Burrows-Wheeler Aligner (BWA; Li & Durbin, 2009). Short indels and structural variants were analyzed with Pindel (Ye et al., 2009). The four nucleotide base counts per genomic location were directly compared between co-twins to first yield a list of potential discordant base substitutions, as in CaVEMan pipeline developed by the Wellcome Trust Sanger Institute (see Supplementary Material available online).

Complete Genomics Platform
Genomic DNA from the twins was sent to Complete Genomics (Drmanac et al., 2010) for whole-genome sequencing. Regions of the genome that do not have sufficient evidence (as determined by the log-likelihood scores assigned to each region) to make a reference or a variant call are made no-call. Discordant variant calling within twins was determined using the Complete Genomics tumor variant calling tool (http://www.completegenomics.com/sequence-data/cgatools/). In order to look for concordance of variant calls between co-twins, 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).

Comparison of Illumina and Complete Genomics Twin Discordances
To identify potential discordant single base substitutions we intersected the within-pair differences detected by Illumina sequencing and those detected by Complete Genomics sequencing. This resulted in a list of potential discordant single base substitutions for each twin pair.

Validation of Potential Discordant Single Base Substitutions
The potential discordant single base substitutions were examined using Sanger, 454 (Margulies et al., 2005), and/or Ion Torrent sequencing, which are documented in the online Supplementary Material.

Results
In order to obtain a first indication, using NGS technology, of whether an accumulation of somatic mutations can be detected with age in humans, we studied NGS whole-genome sequence differences in younger (40-year-old) and

TABLE 1

Phenotypic Characteristics of the LLS and NTR MZ Co-Twins

Measurement	Unit	Normal range	LLS		NTR	
			Twin1	Twin2	Twin1	Twin2
Creatinine	μmol/L	49–90	77	59	83.4	78.60
Gamma GT	U/L	5–40	24	14	33.00	19.00
ASAT [GOT]	U/L	5–30	19	26	24.00	21.00
ALAT [GPT]	U/L	5–34	15	20	12.00	9.00
Glucose	mmol/L	3.1–6.4	7.5	10.3	5.90	5.00
Cholesterol	mmol/L	3.90–7.30	4.65	4.88	5.68	4.63
Triglycerides	mmol/L	0.80–2.30	1.09	0.6	1.23	0.68
HDL-cholesterol	mmol/L	1.16–1.68	1.5	2.43	1.47	1.66
LDL-c (Friedewald)	mmol/L	0.00–3.37	2.65	2.18	3.65	2.65
C-reactive protein	mg/L	0–10	<3	<3	2.43	0.38
Leukocytes	$\times 10^{-9}$/L	4.0–10.0	8.5	6.5	6.50	5.40
Eosinophiles	%	0.0–7.0	3.6	1.7	4.8	1.9
Basophiles	%	0.0–2.0	0.6	0.6	0.3	0.0
Neutrophiles	%	39.0–70.0	54.6	55.4	62.5	53.8
Lymphocytes	%	20.0–40.0	33.2	32.9	24.8	34.7
Monocytes	%	3.0–11.0	8.0	9.4	7.6	9.6
Thrombocytes	$\times 10^{-9}$/L	150–400	361	224	362	262

Note: LLS = Leiden Longevity Study (100-year-old co-twins); NTR = Netherlands Twin Research (40-year-old co-twins).

centenarian female co-twins. The 40-year-old pair participates in the Netherlands Twin Registry and is indicated as NTR (NTR twins), the second pair as LLS (Leiden Longevity Study). Phenotypic characteristics of the NTR and LLS twins are summarized in Table 1.

Sequencing and Germline Variant Calling

Genomic DNA was extracted from whole blood of the two pairs of twins and independently sequenced by Illumina and Complete Genomics. In terms of the number of reads, coverage of the genome, and mapping characteristics, the data from the same platform are comparable among the four samples (see Tables S1 and S2 available online). In addition, all four samples were genotyped for the Illumina 1M array and the concordance rate between these SNP array genotype calls and the calls from the Illumina whole-genome sequence was 99.98%, indicating the high quality of the data and data-processing procedure. On average, 3.2 million SNPs per genome were detected in each genome of the four samples. Using Pindel (Ye et al., 2009) we detected about 0.8 million short insertions and deletions (indels) and 13,000 structural variants per individual.

The mean coverage of the Complete Genomics dataset was 20× for 92.5% of the genome and 40× for 70% of the genome (see Table S2 available online). In comparison to the reference genome, approximately 3.6 million SNPs, 0.4 million short indels, and 5,000 structural variants were observed in the two MZ twin pairs.

Somatic Variant Calling

Next, in a within-pair analysis, we compared the whole-genome sequences of the co-twins from each pair for each platform separately. After mapping Illumina sequence data using BWA (Li & Durbin, 2009), the four nucleotide base counts per genomic location were directly compared between co-twins to first yield a list of 27,529 potentially discordant base substitutions between LLS co-twins and 31,877 between NTR co-twins, using CaVEMan developed by Wellcome Trust Sanger Cancer group for somatic substitution calling from cancer samples. Using Complete Genomics tumor variant calling tool, 11,959 between LLS and 16,007 between NTR co-twins potentially discordant single base substitutions were computed from the variant lists.

The intersection of potential single base substitutions between co-twins supported both by the Illumina and Complete Genomics platform revealed 13 and 17 potential somatic single base substitutions for LLS and NTR co-twins, respectively (Table 2), which did not cluster to particular locations of the genome. After Sanger sequencing these 30 potential single base substitutions, eight were confirmed for both strands. The substitutions identified were present in only a fraction of the reads (substitution ratio Table 2), indicating that these are somatic variations present only in a fraction of the blood cells, as shown in one example Sanger trace for position chr3:4,772,267 (Figures 1 and 2). The eight validated substitutions showed a lower 'background ratio' and contain a higher percentage of reads supporting substituted bases, with an average of 24.7% compared to 12.8% in the five invalidated substitutions (Figure 3).

The intersection may have left much discordance undetected because of platform differences. We therefore examined an additional top 40 potential somatic substitutions for each twin pair from each platform (either Illumina or Complete Genomics) using Roche 454 sequencing. However, none of those 160 (i.e., 4 × 40) sites were confirmed.

Because Illumina and Complete Genomics use very different techniques, we expect that generally the intersection represents true positive discordances among many sequencing artifacts, which are assumed to randomly distribute over the genome and largely not overlapping between the

TABLE 2

Potential Single-Base Substitutions Supported by Both Illumina and Complete Genomics Data and Subsequent Validation by Sanger Sequencing, Roche 454, and Ion Torrent Sequencing (Only for LLS Twins)

Genomic information				Twin	Predicted genotype		Confirmation			Substitution ratio		
chr	Pos	Ref	Gene		Twin 1	Twin 2	Sanger	454	Ion Torrent	Read[1]	454[2]	Ion Torrent[3]
X	49458866	T	PAGE1	LLS	T:T	T:C	Yes	Yes	Yes	0.2632	0.1705	0.2422
X	103014545	C			C:T	C:C	No	No	Yes	0.1667	—	0.2172
3	4772267	A	ITPR1		A:C	A:A	Yes	Yes	Yes	0.2174	0.1925	0.2319
4	3600567	G			G:T	G:G	No	No	No	0.1429	—	—
5	134136335	G	DDX46		G:T	G:G	Yes	Yes	Yes	0.1905	0.1505	0.2373
5	165483058	C			C:T	C:C	No	Yes	No	0.125	0.1635	—
6	164758167	G			G:G	G:A	Yes	Yes	Yes	0.1905	0.1542	0.3664
11	114423117	T	FAM55A		T:T	T:A	No	No	No	0.0943	—	—
12	8357028	A	FAM66C		A:A	A:T	No	No	No	0.1087	—	0.0144
14	81921345	G			G:G	G:A	Yes	Yes	Yes	0.2333	0.1719	0.2175
16	83303332	C	CDH13		C:T	C:C	Yes	Yes	Yes	0.2206	0.1574	0.2042
20	31939491	C			C:C	C:T	Yes	Yes	Yes	0.3704	0.1908	0.4164
20	58727637	A	AX747739		A:G	A:A	Yes	Yes	Yes	0.2885	0.1711	0.2372
2	238496884	A		NTR	A:A	A:G	No	No	—	0.20	—	—
2	47995854	A	RAB17		A:A	A:G	No	No	—	0.1923	—	—
6	150412374	A			A:A	A:C	No	No	—	0.2857	—	—
6	29905621	A	HLA-H		A:G	A:A	No	No	—	0.1667	—	—
6	31313556	G	HLA-B		G:G	A:G	No	No	—	0.1786	—	—
6	57402971	G	PRIM2		C:G	G:G	No	No	—	0.102	—	—
7	57284841	G			G:T	G:G	No	No	—	0.1795	—	—
7	61531167	C			A:C	C:C	No	No	—	0.093	—	—
13	19642122	A	PHF2P1		A:G	A:A	No	No	—	0.1184	—	—
14	30525248	T			A:T	T:T	No	No	—	0.1304	—	—
16	32473534	A			A:A	A:G	No	No	—	0.2083	—	—
16	33932012	A			A:A	A:G	No	No	—	0.1220	—	—
17	78660018	G	RPTOR		A:G	G:G	No	No	—	0.2272	—	—
18	28954823	T			T:T	G:T	No	No	—	0.2083	—	—
19	56676264	T			A:T	T:T	No	No	—	0.1737	—	—
20	57786120	A	ZNF831		A:C	A:A	No	No	—	0.4783	—	—
20	60370175	A	CDH4		A:G	A:A	No	No	—	0.975	—	—

Note: LLS = Leiden Longevity Study (100-year-old co-twins); NTR = Netherlands Twin Research (40-year-old co-twins).

[1]The ratio between the number of sequence reads with a substitution variant and the total coverage in the sample with somatic substitution.

[2]The ratio between 454 reads supporting a substitution and the total number of 454 reads covering the position.

[3]The ratio between the number of k-mers in Ion Torrent reads with a substitution variant and the total number of k-mers with or without the substitution variant in the sample with the predicted somatic substitution.
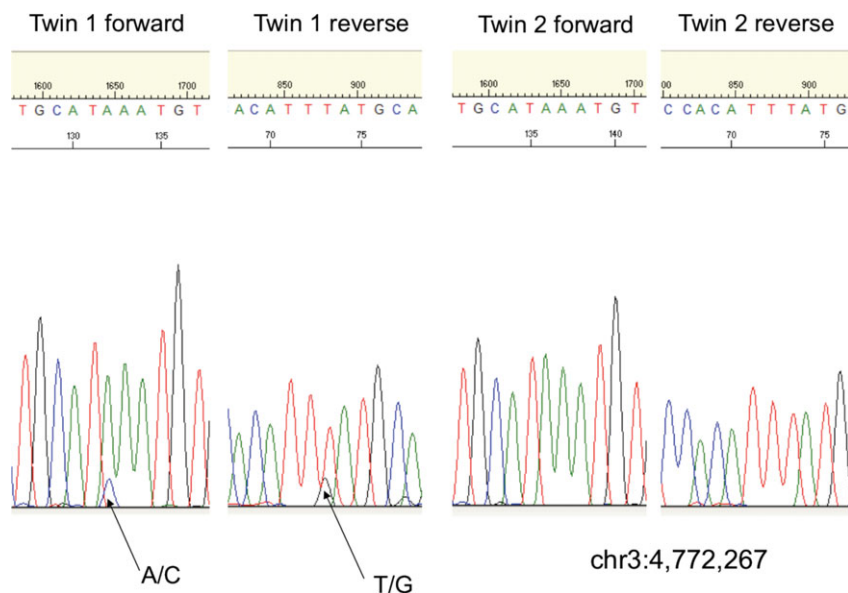
FIGURE 1

Sanger trace around position chr3:4,772,267.

Note: The complete list of eight substitution sites are depicted in Figure 2.
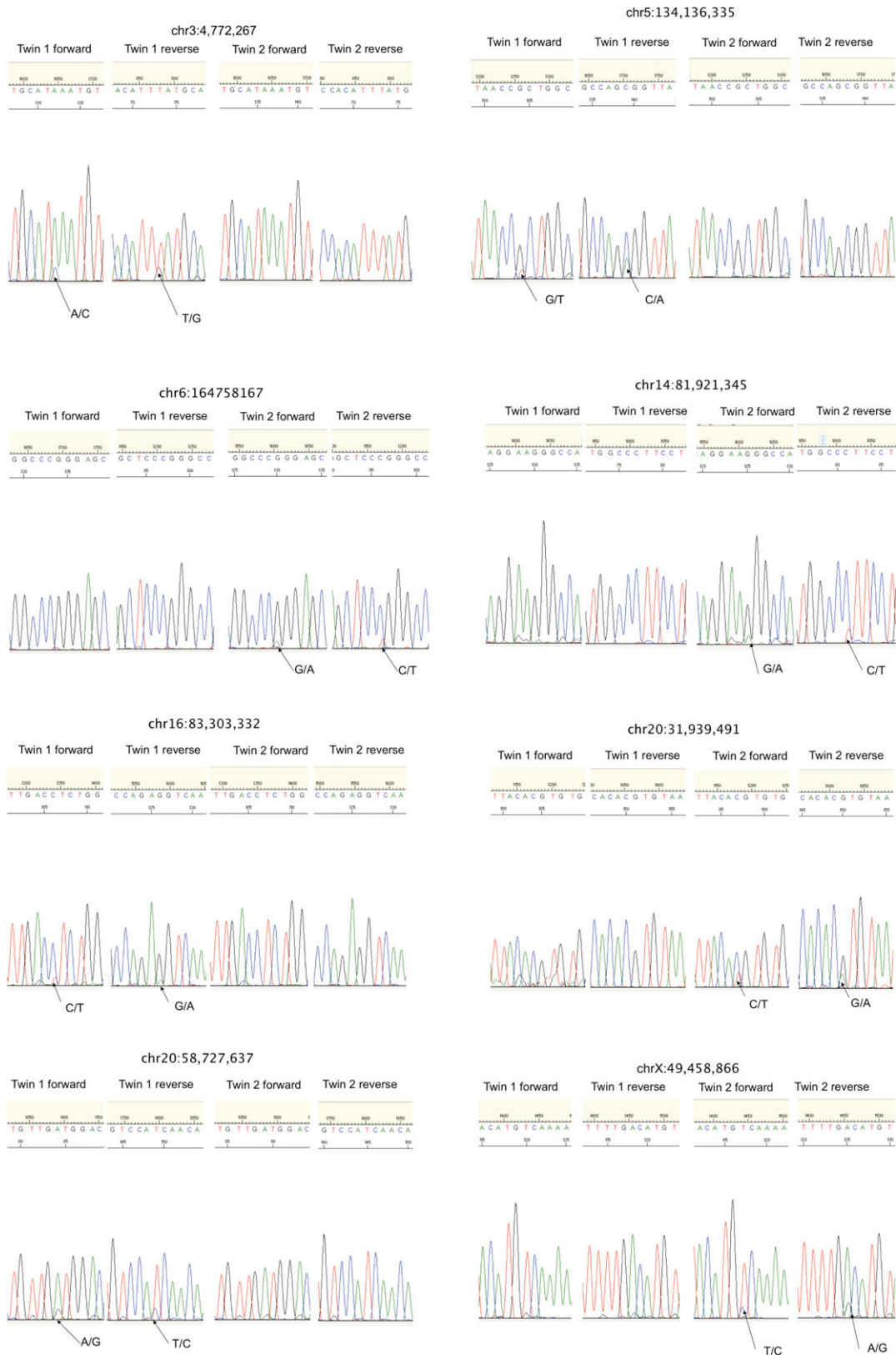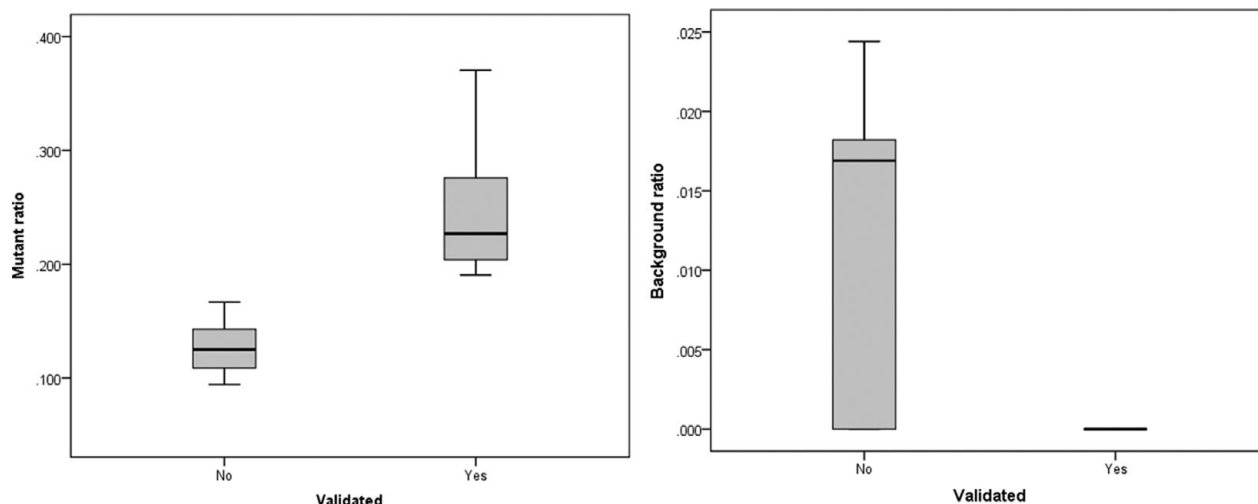
FIGURE 2

Confirmed somatic mutations supported by both Illumina and Complete Genomics platforms in LLS MZ twins.

FIGURE 3

Substitution ratio in Sanger sequencing-confirmed and not confirmed single base substitutions. Left: *p* value for differences between validated and non-validated is .002; Right: *p* value for differences between validated and non-validated is .077.

platforms. Therefore, the results present a reliable lower bound of somatic single base substitution rate between co-twins.

We detected putative discordant short indels, structural variations, or copy number variations between the co-twins but none of them were supported by both platforms. As for potential heteroplasmy sites of mitchondrial DNA among these four samples, we first computed the variant-reference read count ratio at all heteroplasmy sites discovered in any of the four individuals. The correlation of mutation ratio between co-twins is high (>0.994), indicating no somatic mutation between them (see Table S3 available online). However, there is no correlation (~0) between unrelated individuals.

Discussion

Eight single base substitutions between co-twins were supported by both platforms and eventually validated as somatic variants. More variants in the potential list could be true but as stated above, technical differences between the NGS platforms probably hamper replication. The number of somatic variants may be substantially larger but those present in smaller fractions of cells go undetected. Consistent, detectable somatic variation likely includes somatic mosaicism in blood generated during development or clonal expansion of mutations generated at any point during the lifetime. The frequency of these variants is limited in blood even after 100 years of life. The percentage of confirmed somatic substitutions overlapping with introns is surprisingly high (five out of eight). We need additional samples to investigate whether somatic protein-altering variants are under stronger selection. Our observation of low numbers of somatic variants using two NGS

platforms may provide a framework for detecting disease-related somatic variants in phenotypically discordant MZ twins. This will also facilitate the study of diseases caused by somatic mutations occurring in tissues.

Two strategies for the identification and validation of extremely rare somatic variants from error-prone short-read sequencing have been reported. Reumers et al. (2012) investigated filtering strategies for Complete Genomics data of a MZ twin pair discordant for schizophrenia and identified a list of 846 putative discordances. Subsequent validation by Sanger sequencing showed that only 2 out of 814 investigated discordances were validated as somatic variants (Lam et al., 2012; Reumers et al., 2012). The alternative strategy was demonstrated by Lam et al. (2012) by a performance comparison of Illumina (Bentley et al., 2008) and Complete Genomics (Drmanac et al., 2010) whole-genome sequence showing that a significantly higher validation rate of SNPs could be obtained among variants detected by both platforms than among the platform-specific ones (Lam et al., 2012). The same was illustrated in our data since 8 out of 30 potentially discordant variants detected in the overlap of Illumina and CG platforms were validated against 0 of the 40 variants that did not overlap. These two independent platforms have become particularly popular in patient- and population-based studies and their difference with respect to data generation, read structure, and variant detection was investigated in a performance comparison study by Lam et al. (2012).

In summary, this study shows that the number of detectable somatic variants in blood by using NGS is very low and that accumulation of somatic mutations is not necessarily a consequence of a century of life. Stochastic somatic variation occurring in less than 20% of cells will go undetected, however. The low background somatic

variation in MZ co-twins illustrated by our study will be an advantage for identification of disease-related mutations in discordant twins. Our systematic approach shows that such mutations may be detected through comparison of whole-genome next generation sequences generated by at least two independent platforms.

Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/thg.2013.73

Acknowledgments

References

Baranzini, S. E., Mudge, J., van Velkinburgh, J. C., Khankhanian, P., Khrebtukova, I., Miller, N. A., . . . Kingsmore, S. F. (2010). Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, *464*, 1351–1356.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*, 53–59.

Curtis, H. J. (1966). *Biological mechanisms of aging*. Springfield, IL: C.C. Thomas.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., . . . Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, *327*, 78–81.

Kirkwood, T. B. (1977). Evolution of ageing. *Nature*, *270*, 301–304.

Lam, H. Y., Clark, M. J., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., . . . Snyder, M. (2012).

Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, *30*, 78–82.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376–380.

Martin, N., Boomsma, D. I., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, *17*, 387–391.

Orgel, L. E. (1963). The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proceedings of the National Academy of Sciences, USA*, *49*, 517–521.

Orgel, L. E. (1973). Ageing of clones of mammalian cells. *Nature*, *243*, 441–445.

Reumers, J. P., De Rijk, H., Zhao, A., Liekens, D., Smeets, J., Cleary, P., . . . Del-Favero, J. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology*, *30*, 61–68.

van Dongen, J., Slagboom, P. E., Draisma, H. H., Martin, N. G., & Boomsma, D. I. (2012). The continuing value of twin studies in the omics era. *Nature Reviews Genetics*, *13*, 640–653.

Westendorp, R. G., van Heemst, D., Rozing, M. P., Frolich, M., Mooijaart, S. P., Blauw, G. J., . . . Slagboom, P. E. (2009). Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *Journal of the American Geriatrics Society*, *57*, 1634–1637.

Willemsen, G., de Geus, E. J., Bartels, M., van Beijsterveldt, C. E., Brooks, A. I., Estourgie-van Burk, G. F., . . . Boomsma, D. I. (2010). The Netherlands Twin Register biobank: A resource for genetic epidemiological studies. *Twin Research and Human Genetics*, *13*, 231–245.

Willemsen, G., Vink, J. M., Abdellaoui, A., den Braber, A., van Beek, J. H., Draisma, H. H., . . . Boomsma, D. I. (2013). The Adult Netherlands Twin Register: Twenty-five years of survey and biological data collection. *Twin Research and Human Genetics*, *16*, 271–281.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*, 2865–2871.

Zwijnenburg, P. J., Meijers-Heijboer, H., & Boomsma, D. I. (2010). Identical but not the same: The value of discordant monozygotic twins in genetic research. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*, 1134–1149.