Published online 5 June 2018

Nucleic Acids Research, 2018, Vol. 46, No. 15 ***e92***
doi: 10.1093/nar/gky445

Development of the variant calling algorithm, ADIScan, and its use to estimate discordant sequences between monozygotic twins

Yangrae Cho[1,2,†], Sunho Lee[1,3,†], Jong Hui Hong[1,4], Byong Joon Kim[1], Woon-Young Hong[1], Jongcheol Jung[1], Hyang Burm Lee[2], Joohon Sung[5], Han-Na Kim[6], Hyung-Lae Kim[6] and Jongsun Jung[1,*]

[1]Syntekabio Incorporated, Techno-2ro B-512, Yuseong-gu, Daejeon 34025, Republic of Korea, [2]DFTBA, CALS, Chonnam National University, Gwangju 61186, Republic of Korea, [3]School of Computer Science and Engineering, Seoul National University, Seoul, 151-742, Republic of Korea, [4]Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, Seoul 08826, Republic of Korea, [5]Complex Disease and Genome Epidemiology Branch, Department of Epidemiology, School of Public Health, Seoul National University, Seoul 08826, Republic of Korea and [6]Department of Biochemistry, School of Medicine, Ewha Woman's University, Seoul 07985, Republic of Korea

Received June 16, 2017; Revised April 02, 2018; Editorial Decision May 07, 2018; Accepted May 15, 2018

ABSTRACT

Calling variants from next-generation sequencing (NGS) data or discovering discordant sequences between two NGS data sets is challenging. We developed a computer algorithm, ADIScan1, to call variants by comparing the fractions of allelic reads in a tester to the universal reference genome. We then created ADIScan2 by modifying the algorithm to directly compare two sets of NGS data and predict discordant sequences between two testers. ADIScan1 detected >99.7% of variants called by GATK with an additional 724 393 SNVs. ADIScan2 identified ∼500 candidates of discordant sequences in each of two pairs of the monozygotic twins. About 200 of these candidates were included in the ∼2800 predicted by VarScan2. We verified 66 true discordant sequences among the candidates that ADIScan2 and VarScan2 exclusively predicted. ADIScan2 detected many discordant sequences overlooked by VarScan2 and Mutect, which specialize in detecting low frequency mutations in genetically heterogeneous cancerous tissues. Numbers of verified sequences alone were >5 times more than expected based on recently estimated mutation rates from whole genome sequences. Estimated post-zygotic mutation rates were 1.68×10^{-7} in this study. ADIScan1 and 2 would complement existing tools in screening causative mutations of diverse genetic diseases and comparing two sets of genome sequences, respectively.

INTRODUCTION

The genomes of many individuals have been sequenced during the past decade. The whole genome (1), whole exome (2), or a small number of targeted genes (3) were sequenced, primarily for medical genetics. Prominent examples include: The 1000 Genomes (4), The Cancer Genome Atlas (5) and whole exome sequencing projects (2,6,7). The sequence information generated by these projects has clarified the genetic causes of many human diseases and increased our understanding of genetic diversity among human races. The potential to discover causative genes and mutations associated with non-cancerous Mendelian diseases is evidenced by successful examples (7,8). Unbiased analysis of whole genomes from properly selected groups would allow searches of comprehensive lists of genes and known variations that cause complex diseases, such as schizophrenia (9) and Parkinson's disease (10).

Large-scale sequencing of the human genome has been performed by next-generation sequencing (NGS) technologies and usually produces short sequence reads under 150 nucleotides (11). *De novo* assembly of these short reads to obtain whole genome or comprehensive exome sequences is presently beyond our computational capacity. An alternative approach is to align these short reads to a reference sequence and then assemble them into a genome. This process is known as referenced assembly. Discovery of a comprehensive, accurate list of variations requires many aptitudes,

*To whom correspondence should be addressed. Tel: +82 070 7663 0910; Fax: +82 2 6280 0984; Email: jung@syntekabio.com
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

including mapping the proper reads to the reference genome (12,13). Misalignment is unavoidable due to the shortfalls of currently available techniques (12,14,15). Assembling an accurate sequence is only one step among several critical computational challenges. The identification of sequence variation from NGS data is an art and a compound puzzle.

There are several open-access bioinformatics tools or web servers that perform batch annotations of genetic variants from whole genome sequence (WGS) or whole exome sequence (WES) data. Different algorithms use alternative information and statistical models, so have their own strengths and weaknesses (16). Unbiased, consistent variant calling is a major concern regarding the automation of NGS pipelines. Several variant calling tools are effective and widely used: GNUmap (17), GATK (12), SOAPsnp (15), SAMTools (18) and SNVer (19). Their consistency rates were unimpressive, however, as only 57.4% of variants for single nucleotide variations (SNVs) present in dbSNP135 were identified by all five methods (20). Their ability to call variants was even lower for novel SNVs and the concordance rate among the five tools was just 11.4%.

Several approaches use an improved alignment of short reads and various statistical models to call variants with an extreme sensitivity and without sacrificing specificity. Bayesian models were implemented in GATK (12), STRELKA (21), EBcall (22), MuTect (14) and SomaticSniper (23). Probabilistic models, Fisher's Exact Test, and string graphs were respectively used in JointSNVMix (24), VarScan2 (25) and Fermi (26). These algorithms specialize in calling discordant sequences in the ever-changing heterogeneous somatic cells in tumors, and are named somatic variant callers. They use various parameters to identify low-frequency mutations in heterogeneous cancer tissues compared to relatively homogenous healthy tissues. Discordant sequences predicted by several somatic variant callers showed very low concordance rates (27,28). A deep read depth >1000× is preferred to discern true variants from mechanistic artifacts when a probabilistic model is implemented. The current paradigm for read depth, however, is 30–70× for WGS and 100–150× for WES.

The polishing steps for NGS raw data were standard, while the alignment step of short reads to a reference sequence was coordinately developed with variant calling algorithms. We took advantage of well-established polishing and alignment tools in the GATK package and developed a variant calling method using a premise that tissue samples of the genome were homogeneous. This method was designed to rank the extent of heterogeneity of the position based on pair allelic fractions, so we named it **A**llelic **D**epth and **I**mbalance **S**canning, or ADIScan1. We initially evaluated its accuracy by comparing variants detected by this algorithm to those detected by GATK. In addition, we modified the variant calling of ADIScan1 to simultaneously compare a set of NGS sequences and called it ADIScan2. We then evaluated the ability of ADIScan2 to detect discordant sequences between the individuals in each of two sets of monozygotic twins. The results of this study provide a reason to reevaluate the extremely low post-zygotic mutation rates recently estimated based on whole genome sequences.

MATERIALS AND METHODS

Determination and alignment of the whole genome sequence

All samples of total DNA were extracted from whole blood containing WBCs using the Blood DNA Extraction Kit (Qiagen, Palo Alto, CA, USA). The whole genome sequence with 150-bp paired-end reads was determined using the HiSeq X10 system following manufacturer's protocol in the TruSeq DNA PCR-free library (Illumina, San Diego, CA, USA). One microgram of genomic DNA was fragmented by Covaris systems and the double-stranded DNA fragments with 3′ or 5′ overhangs were repaired with an exonuclease and polymerase mix. The appropriate library size was selected using different ratios of Sample Purification Beads. Multiple indexing adapters were ligated to the ends of the DNA fragments to prepare them for hybridization onto a flow cell. The enriched DNA library was further amplified by polymerase chain reaction prior to sequencing. The libraries were sequenced with an Illumina HiSeq X10 sequencer. We verified the quality of each read using the software (FastQC version 0.10.1) in the HiSeq X10 sequencer and generated a Fastq file for each tester sample before sequence alignment and further analysis (Supplementary Figure S1). The Burrows-Wheeler aligner (BWA-MEM; version 0.7.10 (13)) was used to align the sequence reads to the human reference genome sequence GRCh37 with default parameters. We converted the alignments in sequence alignment/map (SAM) format to binary alignment map (BAM) files implemented in SAM tools (SAM tools version 0.1.10 (18)). We then used the Picard tool (version 1.119; http://picard.sourceforge.net) to remove duplicate reads and sort sequence reads in order based on their start position. BAM files were realigned to the reference sequence Ch37d.5.fa with GATK Realigner Target Creator (version 113.3-0). Local alignment was fine-tuned with GATK Indel Realigner. Base quality scores were recalibrated by the GATK base-quality recalibration tool, GATK Base Recalibrator. Subsequently, a fine-tuned BAM file generated by GATK PrintReads was used for variant callings by all four algorithms.

Ethics approval and consent to participate

The institutional review board at Seoul Samsung Hospital and ethics committee at Seoul National University approved this study. All procedures were executed in accordance with the relevant guidelines and regulations as described below. Two pairs of monozytotic twins, females and males, donated the samples used in this study and provided their written, informed consent. All sample names were randomly number-coded and processed in accordance with institutional guidelines, from library construction to subsequent sequencing and analysis.

Determination of HLA types

We determined HLA types of class I and class II using HLAscan as described elsewhere (29). We performed the initial alignment using BWA-MEM v0.7.10-r789 with default options (15). Based on the initial alignment, we selected reads that aligned to the HLA regions and subse-

quently realigned them to reference HLA alleles obtained from the IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/). Alignments were performed against exons 2, 3 and 4 of class I HLA genes, and exons 2 and 3 of class II HLA genes. We used a score function in HLAscan to evaluate the distribution of aligned reads on the target region, and then determined the closest matches of alleles. This whole process was performed in one step, semi-automatically, by a local computing system at Syntekabio (Daejeon, Republic of Korea).

ADIScan1 for variant calling in NGS

We based development of the variant calling algorithm, allelic depth and imbalance scanning (ADIScan1), on three suppositions: the genome in a tissue was homogenous; the reference sequence was a homozygote at each position; and for all testers, the proportion of an allele at each position was either 0.5 for a heterozygote or 1.0 for a homozygote. We ran a NGS procedure and produced a BAM file for the whole genome sequence of the reference human DNA material NA12878 (National Institute of Standards and Technology, Gaithersburg, MD). We used the Curve Fitting Tool method in MATLAB_R2015a to generate six constants for the variant calling algorithm in ADIScan1. The NA12878 included 3 117 120 variants regarded as gold standard. All other positions with reasonable sequence depth were regarded as non-variants. We used 2.9×10^6 of the 3 117 120 variants as a pool from which training set variants were selected. Allelic frequency information from the NA12878 reference DNA was used in directing ADIScan1 to call variants from the BAM file. The training involved two steps, as described below.

We calculated the variant score, *Adiscore 1*, as a tangential function to distinguish variants over non-variants at each genome position between the homozygotic reference and a comparing tester as follows:

$$Adiscore\,1 = 1/(1 + e \wedge (-S(i)), \tag{1}$$

In Equation (1),

$$(i) = tan\,(\mathrm{D}(i) - 0.5) \times 31 \\ +\log_5(\mathrm{Max}(5,\ \mathrm{Min}(\mathrm{DP}_{\mathrm{ref}}(i),\ \mathrm{DP}_{\mathrm{alt}}(i)))) + 6.964 \tag{2}$$

In Equation (2), 31, 5 and 6.964 were constants derived from the curve-fitting process by Matlab (Figure 1A and B). The curve-fitting process was performed 1000 times using 52 500 randomly selected positions, including 35 000 of the 2.9×10^6 verified variants and 17 500 non-variants. The constants selected were the averages of the lower and upper 95% confidence limits; DP stands for read depth of the NGS sequence where

$$\mathrm{D}(i) = \mathrm{Min}(1,\ (\mathrm{DP}_{alt}(i) + 1)/(\mathrm{DP}_{\mathrm{ref}}(i) + 1)) \tag{3}$$

After determining the variants by Equation (1), we used the following equations to established whether each variant was a heterozygote or alternative homozygote.

$$Heteroscore = 1/(1 + e \wedge (-\mathrm{V}(i)), \tag{4}$$

In Equation (4),

$$V(i) = tan\,(\mathrm{E}(i) - 0.5) \times 99.74 \\ +\log_{4.997}(\mathrm{Max}(4.997,\ \mathrm{Min}(\mathrm{DP}_{\mathrm{ref}}(i),\ DP_{\mathrm{alt}}(i)))) + 29.6 \tag{5}$$

In Equation (5), three constants were derived from the curving-fitting process described for *Adiscore* 1 (Figure 1C and D). We randomly selected 35,000 of the 2.9×10^6 verified true variants without non-variants for each reiterative fitting process.

Where

$$\mathrm{E}(i) = \mathrm{Min}(1,\ (\mathrm{DP}_{\mathrm{ref}}(i) + 1)/(\mathrm{DP}_{\mathrm{alt}}(i) + 1)) \tag{6}$$

Adiscore 1 ranged from 0 to 1, with 0.5 as the default setting to call variants and allelic status. The output file was in VCF format and Ti/Tv ratios were calculated by SnpSift (Ver. 4.2).

ADIScan2 for calling discordant sequences between a set of NGS

We modified the variant calling algorithm to call discordant sequences between a set of next generation sequences and named it ADIScan2. Before sequence comparisons, we classified the types of pair allele fractions between two testers into three groups and eight subgroups based on the differential read depth of two alleles in each tester (Supplementary Table S1). We assigned a differential weight to each subgroup for score calculation. This calculation was the distance of pair allelic fractions at each genome position between two comparing testers as a tangential function, as follows:

$$\mathrm{a}_i = (\mathrm{A} + 1)/(\mathrm{B} + 1) \text{ for tester 1, } \mathrm{x}_i = \mathrm{bin}\,(\mathrm{a}_i) \text{ and} \\ \mathrm{b}_i = (\mathrm{A'} + 1)/(\mathrm{B'} + 1) \text{ for tester 2, } \mathrm{y}_i = \mathrm{bin}\,(\mathrm{b}_i) \tag{7}$$

In Equation (7), A, A' and B, B' are respectively the depth of reads for the minor and the major allele in the position *i*. The a_i or the b_i represents a ratio of total reads between the minor and major alleles. We added 1 to avoid a 0-denominator in the formula. The ratios were allocated to one of 21 bin numbers in order from 1, the lowest, to 21, the highest. The ratio for the first group ranged from 0 to 0.0075 and the ratios for the subsequent bins were increased by adding 0.05 to the previous number each time except for the last step, where 0.075 was added. All other groups were evenly divided into intervals of 0.05. Each of x_i and y_i is a bin number.

$$\mathrm{t}_i = 1/\tan(\mathrm{x}_i/\mathrm{y}_i)\mathrm{x}\ 1/\tan([22 - \mathrm{y}_i][22 - \mathrm{x}_i]) \\ \text{where } \mathrm{y}_i > \mathrm{x}_i \text{ and } [22 - \mathrm{x}_i] > [22 - \mathrm{y}_i] \text{ or,} \\ \mathrm{t}_i = 1/\tan(\mathrm{y}_i/\mathrm{x}_i)\mathrm{x}\ 1/\tan([22 - \mathrm{x}_i]/[22 - \mathrm{y}_i]) \\ \text{where } \mathrm{x}_i > \mathrm{y}_i \text{ and } [22 - \mathrm{y}_i] > [22 - \mathrm{x}_i] \tag{8}$$

In Equation (8), 22 was a constant number generated by adding 1 to the largest bin number, 21, and t_i was the output of the tangent function of the ratio of allelic differences between two comparing testers.

$$\text{Differential score } = \log(\mathrm{t}_i * 40) * \mathrm{w} \\ -\log\,(\mathrm{small}(\mathrm{A},\ \mathrm{B},\ \mathrm{A}\prime,\ \mathrm{B}\prime)) * \mathrm{C1} - \mathrm{C2} \tag{9}$$

In Equation (9), five weights (w = 1.1 or 1.2; and 0.7, 0.8, or 0.9,) were specified. Weights 1.1 and 1.2 rewarded

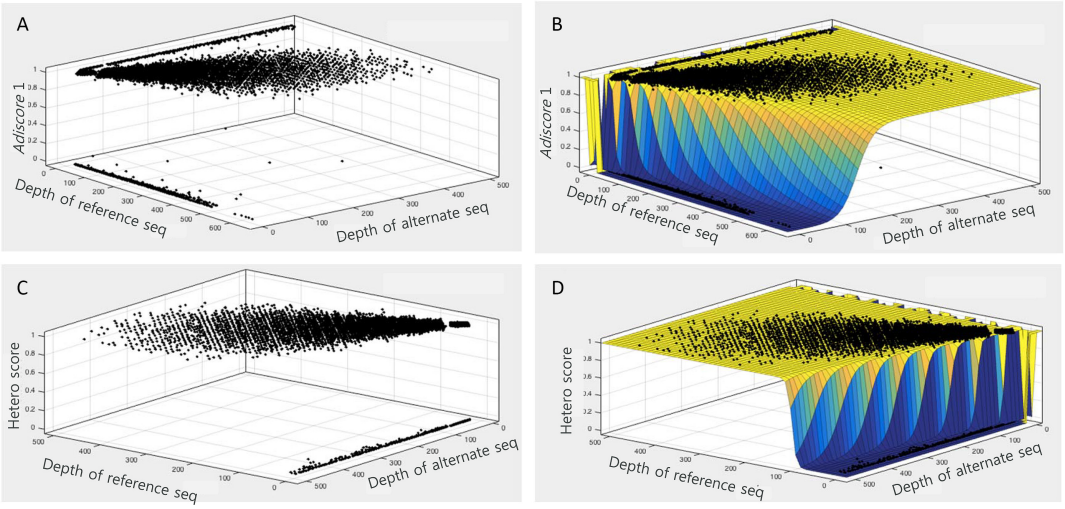

Figure 1. Optimization of the variant calling algorithm, ADIScan1, by tangential conversion of read depth ratios between the reference allele and an alternative allele. (**A**) Three-dimensional chart showing 35,000 variants (mostly at the top) and 17 500 non-variants (mostly at the bottom). Three axes show the information on read depth for reference and alternative alleles, and *Adiscore* 1 to distinguish true variants as 1 and non-variant as 0. (**B**) Fitting process to a general model $f(x,y) = 1/(1 + \exp(-(a*(\tan(\min(1, y/x) - 0.5) + c)))$ to calculate three constants a, b, and c. (**C**) Three-dimensional chart showing heterozygotic variants (top) and homozygotic variants (bottom). (**D**) Fitting process to a general model $f(x,y) = 1 / (1 + \exp(-(a*(\tan(\min(1, y/x) - 0.5) + c)))$ to determine three constants a, b and c.

the cases with few or no sequencing errors in calling a homozygote, while weights 0.7, 0.8 and 0.9 differentially penalized the cases with different extents of sequencing errors. When the ratio of reads for the smaller allele was larger than 27.5%, the position was regarded as a heterozygote. The weights were generally rewarded toward the cases where directions of the pair allelic fraction were opposite between two testers (PAF 2, 3, 4, and 5 in Supplementary Table S1) and their distance from the perfect heterozygotic status was 50 to 50. When the combined read depth of both alleles was zero in either tester due to errors in sequencing or alignment, the case was classified as a 'No read' group and type 1. Subsequent analysis to call discordant sequences between the two testers was terminated. Differential scores ranged from 1 to 50, so we used a score of 25 in calling discordant sequences for further consideration.

BWA-GATK for finding variants in each set of NGS

To compare the variants called by other algorithms, we used the GATK software tools (version 3.3.0; http://www.broadinstitute.org) in alignments and genotype calling and refining, with recommended parameters (12). We used the GATK UnifiedGenotyper to call genotypes and the GATK VariantRecalibrator tool to score variant calls by a machine-learning algorithm and to identify a set of high-quality SNPs using the Variant Quality Score Recalibration (VQSR) procedure. We included only SNVs with a depth of 25 or higher in the final variant calls. We saved the variants in a variant call format (VCF) file for each twin. Ti/Tv ratios were calculated manually or with SnpSift (version 4.2). Discordant bases between monozygotic (MZ) twins included all unique variants in either twin. These unique variants were discovered by subtracting all variants in one twin from the variants of the other paired twin.

BWA-MuTect for finding discordant sequences between a paired NGS

For genotype calling with the Bayesian-based somatic variant caller, MuTect (14), we used the fine-tuned alignment files, cosmic b37_cosmic_v54_120711.vcf, dbSNP_138.b37.vcf, generated by the GATK suite with recommended parameters. The MuTect suite removed low-quality sequence data before variant calling and designated variants as either germline or somatic. When we compared the genomes of identical twins, *de novo* mutations in each individual were like somatic mutations. Thus, we defined one individual as normal and, in a reciprocal manner, the other paired twin as a tumor sample. We combined variants from each twin with a depth of 25 or more and tallied the total number of sequence variants.

BWA-VarScan2 for finding discordant sequences between paired NGS

The somatic variant caller, VarScan2 (30), based on both a heuristic and statistical algorithm, was used to produce a VCF file containing SNVs. The VarScan2 algorithm was designed to discover somatic variants in tumor samples when compared to a normal sample. Thus, we discovered variants in each twin by defining a BAM file for one individual as normal and the other as tumor. After making this distinction, we performed the analysis twice for each twin pair. The status of variants in the output file was classified into germline, somatic, loss of heterozygosity, or unknown. We tallied the final variants as described above in the BWA-MuTect variant calling method.

Exclusion of SNVs in low-complexity regions

Interspersed repeats and low-complexity DNA sequences (hereafter 'low-complexity regions') were excluded from the

variants using RepeatMasker (A.F.A. Smit, R. Hubley & P. Green, unpublished version 4.0.6). Although there was no size limitation in the length of the query sequence or size of batch file, we used fewer than 100 000 selected sequences from 200 bases flanking the testing variants as a batch file. We performed all jobs using the default parameter in a local computing system at Syntekabio (Daejeon, Republic of Korea).

Verification of mutations by Sanger sequencing

To verify the accuracy of variants called by the three algorithms, we selected 216 positions and tried to determine their nucleotide sequences by Sanger sequencing. The 216 positions included 151 positions called exclusively by ADIScan2 and 65 positions called mainly by VarScan2. We performed PCR amplification using 10 ng of genomic DNA from each of the twins. For the PCR reactions, we designed two primers per amplicon by using NCBI/Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). The size of the amplified PCR products was restricted to a relatively small range of 260–572 bp for homogenous amplification, and for clean results from subsequent Sanger sequencing. For PCR amplification of the fragments, we used High-Fidelity Pfu DNA polymerase with an error rate of 1–2 $\times$ 10^{-6}. PCR reactions were performed with 2X H-Star Taq PCR Master mix 1 (BIOFACT Co. Ltd, Daejeon, Republic of Korea), 10 pmol of each primer, 20 ng of genomic DNA in a 30-μl reaction volume by 95°C denaturation for 3 min followed by 35 cycles of 20 s of denaturation at 95°C, 40 s of annealing at 56°C, and 1 min of elongation at 72°C. After the cycling reactions, the final elongation was performed at 72°C for 3 min. PCR products were size fractionated on 1% agarose gel by electrophoresis and visualized by ethidium bromide staining. Scientists at BIOFACT (BIOFACT Co. Ltd, Daejeon, Republic of Korea) purified PCR products with the PCR Cleanup Kit and determined DNA sequences using the Sanger sequencing method. We manually inspected chromatograms to confirm the sequence accuracy of each file. Information on the chromosome positions of variants, sequence of each primer, and amplicon sizes are provided in Supplementary Table S2.

RESULTS

Whole genome sequences of paired twins

The research organization, Personalized Genomic Medicine 21 in Korea, generated whole genome sequences for two pairs of twins using 150-bp paired-end reads with DNA extracted from their blood. The procedure produced over 1.4 $\times$ 10^9 reads, an average coverage of 75$\times$ for each individual (Table 1). The sequence depth was about twice that of routine NGS of the human genome with HiSeq x10. Of the >2.9 $\times$ 10^9 non-N-bases in target areas of the human reference genome, >2.8 $\times$ 10^9 positions had over 25$\times$ coverage after the final alignment using BWA-MEM (13). Most of the >2.8 $\times$ 10^9 bases were also covered a minimum of 25$\times$ for the individuals in each pair of twins. Each case was now ready for sequence comparison. First, we compared the Human Leukocyte Antigen (HLA) types of each individual. Each paired twin had HLA types identical to its sibling at six loci based on analysis of the NGS data (Supplementary Table S3), supporting the assumption that both pairs of twins were MZ. We used this set of NGS data to evaluate the ability of the new algorithm to call SNVs in each NGS, and then compared it to the GATK algorithm.

Rationale of the variant calling algorithm, ADIScan1

Each locus of the human genome consists of two DNA base molecules, one inherited from each biological parent. The two molecules can be identical or different, producing a homozygote or a heterozygote, respectively. The ratio of two alleles in each heterozygotic locus is generally one in the human genome, with the exception of clonal mosaicism in tissues (31–33). The number of reads for the paired alleles produced by next-generation DNA sequencing of human tissues, however, generates an array of ratios for the two alleles. These results may be due to mosaicism of tissues, technical errors introduced during PCR amplification, NGS, or misalignment of short reads to the reference genome. To call true SNVs using the new algorithm, we hypothesized that each position in the reference sequence was a perfect homozygote and assigned a score for a non-reference allele based on the ratio between the read-numbers of the non-reference allele and a reference allele in each position. We devised the two-step sequential scoring functions by tangential conversion of the ratios using the curve-fitting process in Matlab. We then discovered variants using *Adiscore* 1 (Figure 1A and B), and subsequently distinguished hetero- and alternative homo-variants by Hetero score (Figure 1C and D). *Adiscore* 1, without the prejudice of the Bayesian or other statistical approaches, was the sole criterion for the subsequent call of variants.

SNVs called by GATK and ADIScan1

The average number of SNVs detected by ADIScan1 in the genomes was 3 995 709 (S.D. 16 158, 0.14% of the genome) (Table 2). Compared to GATK, which had high sensitivity and specificity (20), ADIScan1 called 724 393 (18%) more SNVs than GATK and detected >99.74% of the variants (see below). Over 200 000 of the 724 393 variants exclusively called by ADIScan1 were not included in dbSNP (dbSNP_138.b37.vcf). Overall, ADIScan1 appeared to be more liberal than GATK in variant calling. In comparison, GATK (12) detected an average 3 279 887 (S.D. 6729, 0.11% of the genome) of the variants, consistent with the estimated 0.1% diversity in nucleotide sequences among human populations (34). The transition/transversion (Ti/Tv) ratios for SNVs in the whole genome and coding sequences (CDS) regions were respectively 2.1 and 3.1 (Table 2), as expected (1). Only a small portion (0.26%) of the variants detected by GATK escaped detection by ADIScan1. The Ti/Tv ratios for the variants detected by ADIScan1 were slightly lower than GATK in both the whole-genomic and CDS regions, but were within the expected range. More than half of the variants were in interspersed repeats and low-complexity DNA sequences (hereafter 'low-complexity regions') of the genome. Even after we excluded the variants in these low-complexity regions a similar proportion of the variants were

Table 1. Statistics of the whole genome sequencing results for two pairs of monozygotic twins

Sample ID	[a] Total reads ($\times 10^3$)	[b] Ave. length	[c] Ref. seq. ($\times 10^3$)	[d] Mapped reads ($\times 10^3$)	On target rate	# of total nucleotides determined ($\times 10^6$)	Ave. read depth	# of positions with depth >25 ($\times 10^3$)	# of positions with depth >25 in paired twins ($\times 10^3$)	% of ref. genome covered >25	% of ref. genome covered >25 in paired twins
F1	1 725 360	150	2 900 340	1 422 710	82.5%	213 407	74	2 812 382	2 802 747	97%	97%
F2	1 606 073	150	2 900 340	1 352 192	84.2%	202 829	70	2 805 407	2 802 747	97%	97%
M1	1 813 082	150	2 900 340	1 526 924	84.2%	229 039	79	2 821 200	2 812 202	97%	97%
M2	1 760 651	150	2 900 340	1 477 158	83.9%	221 574	76	2 818 852	2 812 202	97%	97%
Ave.	1 726 292	150	2 900 340	1 444 746	83.7%	216 719	75	2 814 460	2 807 474	97%	97%

[a]Total number of reads.
[b]Average length of reads in the nucleotide.
[c]Total number of non-N-bases in the reference sequence.
[d]Total number of mapped reads on the reference sequence. F1 and F2 represent two females in a pair of twins, while M1 and M2 represent another pair of male twins.

common between GATK and ADIScan1 (Table 2 and Supplementary Table S4).

Indirect comparison of NGS for discordant variants

We used variants located in non-repetitive regions (Supplementary Table S4) to estimate discordant sequences in subsequent analyses. This reduced errors originating from the misalignment of short sequences. Among an average 1 408 823 SNVs detected by GATK in the whole genomes of the four MZ twins, >99% (1 382 185/1 408 823) in the whole genome were identical between paired twins. Similarly, ∼98% (1 499 559 of 1 563 932) of SNVs detected by ADIScan1 were also identical between paired twins. We considered the unique variants defined by subtraction (28) in each individual of the paired-twins as potential discordant sequences between the twins. GATK and ADIScan1 respectively detected an average of 26 638 and 64 373 positions (Supplementary Table S5). These numbers were respectively >44× and >107× greater than the 600 expected discordant sequences based on mutation rates of 1.2×10^{-7} in human somatic cells (35). Further, discordant variants between paired twins called by the two methods overlapped by 1249 (<5%) on average (Supplementary Table S5). This contrasted with the >99% of SNVs called by GATK that also were called by ADIScan1 (Table 2 and Supplementary Table S4).

Direct comparison of NGS by ADIScan2

We used ADIScan2 as an alternative to the subtraction method (28) to simultaneously compare NGS reads between paired twins with an assumption that there was no genetic mosaicism in white blood cells (WBCs). We took the fraction of paired alleles between the sets of twins into consideration to avoid the prejudice from prior knowledge on sequence variations in human populations. In this study, we also converted the ratio into tangential functions and treated any allelic sequence with >27.5% of all reads at a position as a candidate for a real allele and reads below this fraction as a technical artifact (Figure 2). We selected this ratio to exclude artificial alleles with allelic fractions unlikely to exceed 10%. These alleles could stem from either NGS procedures or read-alignment processes. Even after the initial cut-off, the final number of variant candidates was 433 107 for female twins and 424 381 for male twins with the differential score ≥5 (Supplementary Table S6). When we used the differential score of 25 routinely used in previous studies, a total of 492 and 474 (5988 and 5353 before using RepeatMasking) positions appeared to be discordant sequences between the paired female and male twins, respectively (Figure 3 and Table 4). The candidates for discordant sequences were >10× fewer than those identified by subtraction, 95% of which were included in the set selected by the subtraction. The average fraction of variant alleles (VAF) was 0.36 for the candidates of discordant sequences.

Discordant sequences detected by VarScan2 and MuTect

To evaluate the accuracy of ADIScan2, we compared discordant sequences called by this method to two somatic variant callers, the statistics-based VarScan2 and Bayesian-based MuTect (14,30). The numbers of variant candidates identified by VarScan2 and MuTect from data for the two sets of MZ twins (female/male) were respectively 19 675/13 226 and 5534/5242 for all positions in the whole genomes of the twin pairs (Table 4). These numbers were 2860/2749 and 1237/1037 after excluding variants in low-complexity sequence regions (Table 4 and Figure 3). The number of discordant sequences called by the three methods was different, and the concordance rate among them was only ∼10% for all but one possible combination (Figure 3). This finding of the low concordance rate among different algorithms was comparable to the results generated by other research groups (20,27). The exception was that 41% of the discordant variants called by ADIScan2 were also called by VarScan2. The VAF among the discordant sequences predicted by MuTect was exceptionally low (<0.12), compared to 0.33 among those called by ADIScan2 and VarScan2 (Table 4).

Verification of variants called by VarScan2 and ADIScan2

It was of note that ADIScan called the least number of discordant sequences and over 40% of them were also called by VarScan2, while the concordance rate among discordant sequences called by the four algorithms was generally low. To estimate the accuracy of callings, we selected 217 positions

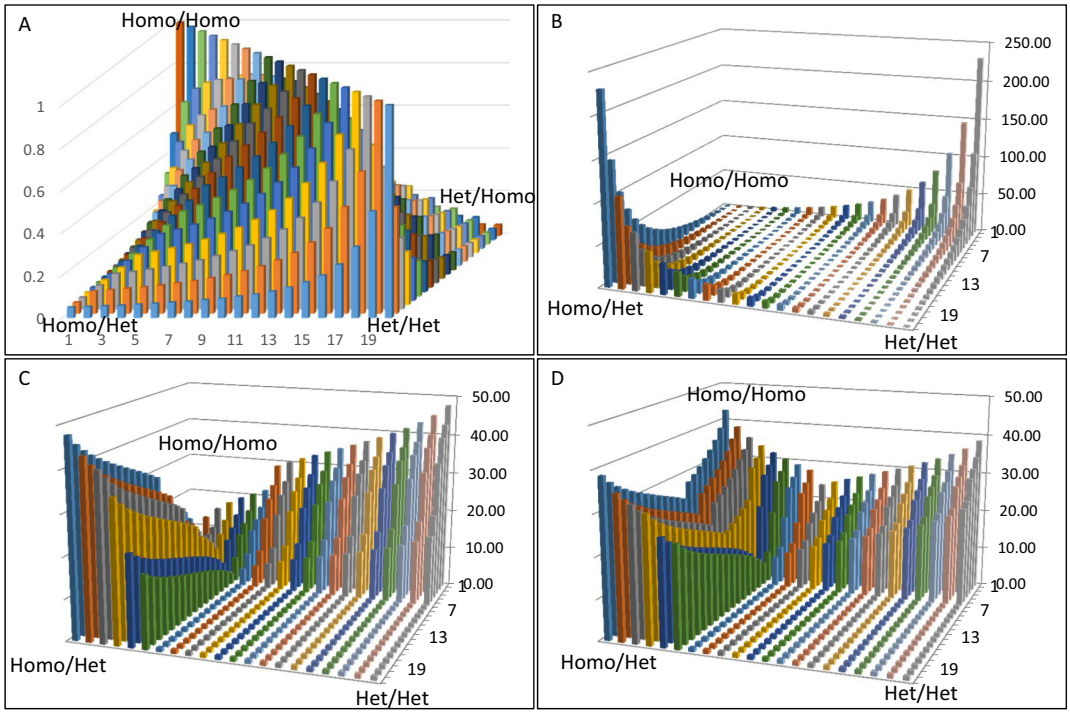

Figure 2. Process of Differential score generation. X- and Y-axes show bin numbers assigned by the ratios of two allelic sequence reads. An extreme, homozygote (Homo), has all reads with only one allelic sequence, and an assigned bin number of 1. The alternative extreme, heterozygote (Het), has an equal number of the two alleles and an assigned bin number of 21. Paired allelic fractions are in the same direction and homo/homo represents identical sequences in both testers (A–C). (**A**) Ratio of bin numbers in two testers. (**B**) Inverse tangential conversion of the bin-number ratios. Value of Homo/Het and Het/Homo increased. (**C**) Differential score calculated by logarithmic conversion of the tangential values with applying predefined weights and subsequent adjustment of the score based on the allelic ratios. (**D**) Differential score calculated for cases in which each tester has different sequences although both were homozygotes and represented homo/alternative homo (Homo/Homo).

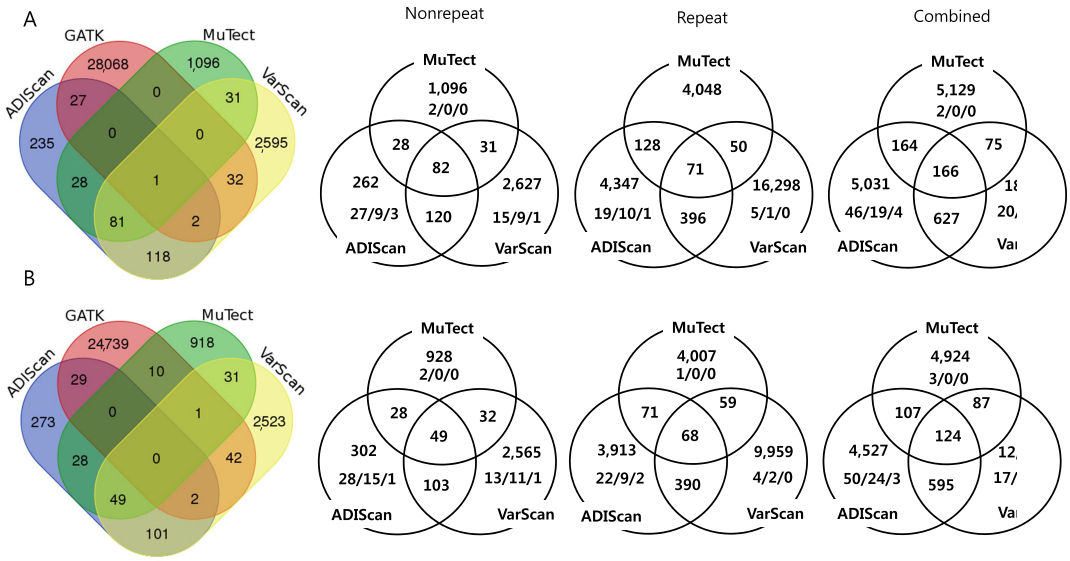

Figure 3. Comparison of single-nucleotide variant (SNV) concordance called by four different algorithms. (**A**) Female twins. B. Male twins. Numbers in the Venn diagrams at the far left side indicate variants in non-repetitive regions of the genome called by ADIScan2, GATK, MuTect and VarScan based on the BAM files generated by BME-GATK alignment. Following diagrams show SNVs in non-repetitive regions, repetitive-regions, and both regions combined, in order. Three numbers separated by slashes are PCR-failure cases and correct variants, followed by incorrect variants. Concordant SNVs among the four methods were identified by matching genomic coordinates and custom Venn diagrams drawn in the website http://bioinformatics.psb.ugent.be/webtools/Venn/.

Table 2. Comparisons of single nucleotide variants (SNVs) in the whole genome called by ADIScan1 and GATK

	GATK				ADIScan1				Common in GATK and ADIScan1				Differences between GATK and ADIScan1			
	Number of SNVs in WGS	Ti/Tv ratio	Number of SNVs in CDS	Ti/Tv ratio	Number of SNVs in WGS	Ti/Tv ratio	Number of SNVs in CDS	Ti/Tv ratio	WGS	Ti/Tv ratio	CDS	Ti/Tv ratio	GATK-ADIScan1[a]		ADIScan1-GATK[b]	
Sample ID													WGS	CDS	WGS	CDS
F1	3 291 076	2.09	20 392	3.13	4 023 501	1.90	24 124	2.77	3 282 483	2.09	20 375	3.13	8593	17	741 018	3749
F2	3 278 042	2.09	20 416	3.12	3 983 656	1.92	24 016	2.80	3 269 723	2.09	20 401	3.12	8 319	15	713 933	3615
M1	3 273 111	2.09	20 442	3.12	3 989 027	1.90	23 971	2.78	3 264 321	2.09	20 416	3.12	8790	26	724 706	3555
M2	3 277 320	2.09	20 465	3.11	3 986 653	1.90	24 039	2.76	3 268 738	2.09	20 431	3.12	8582	34	717 915	3608
Average	3 279 887	2.09	20 429	3.12	3 995 709	1.90	24 038	2.78	3 271 316	2.09	20 406	3.13	8571	23	724 393	3632

[a]GATK-ADIScan1: difference set of positions in GATK after the same positions in ADIScan1 were subtracted.
[b]ADIScan1-GATK: difference set of positions in ADIScan1 after the same positions in GATK were subtracted. WGS = whole genome sequence, CDS = coding sequence region, Ti = transition; Tv = transversion.

for Sanger sequencing. We included only five candidates called by MuTect in the verification set because their VAF was so low (<0.12) that they were likely technical errors instead of true discordant sequences between MZ twins. Only 85 of 217 (39%) of the PCR amplifications were successful for subsequent use with the Sanger method. Rather than perfecting the PCR procedure, we performed direct PCR sequencing with the amplified products. Of the 85 positions, 60 were called by ADIScan2 and 25 by VarScan2. Of the 60 positions exclusively called by ADIScan, 43 were true discordant sequences, while 23 of the 25 positions exclusively called by VarScan were true discordant sequences. These missed sequences were included in the group with differential scores of ADIScan2 between 10 and 20. Forty-four of the 66 verified positions were within non-repetitive regions and the other 22 positions were in repetitive and low-complexity regions. Based on the ∼44% accuracy of VarScan2 calls in both non-repetitive regions plus numbers of missed sequences, we extrapolated an average of 1012 discordant sequences between both members of each pair of MZ twins. Mutation could happen in either individual of the MZ twins. In our analysis there were 530 mutations in an individual, an estimated mutation rate of 1.68×10^{-7}.

DISCUSSION

Comprehensive variant calling by ADIScan1

Very low frequency mutations in any one of 3743 genes can cause 5,991 rare genetic diseases (36) (OMIM, https://www.omim.org/statistics/geneMap). With the advancement of NGS technology, whole-genome or whole-exome sequencing makes it possible to discover causative genetic mutations by massive parallel sequencing instead of the gene-by-gene approach used by the Sanger method. It is not unusual to find new mutations in causative genes among patients for previously known diseases. The accuracy of NGS results depends on several factors including library preparation, the sequencing technology platform, and read depth. Relatively high error rates are associated with calling false positives or negatives and variant calling remains a challenge. Intuitively, an effort to decrease false negatives is likely to increase false positives and *vice versa*. Discovering all positives without false negatives is nearly impossible with NGS analysis. In the process of discovering causative variants for rare genetic diseases, therefore, false negatives in variant calling are a critical problem. Since each rare genetic disease is generally caused by a single-gene mutation, missing even one variant can results in failure to discover a disease-causing mutation (16).

For a comprehensive detection of mutations, the full-length sequence of a whole genome by NGS is necessary. The next step requires powerful variant calling algorithms to detect all true variants. Notably, ADIScan1 called over 99.7% of the variants called by GATK (Table 2), unlike other programs that found fewer (20). Notably, ADIScan1 called >18% more variants than the total called by GATK (Table 2). Over 200 000 of these additional variants were not listed in the dbSNP (dbSNP_138.b37.vcf), which was probably part of the reason they were not selected by GATK. Although GATK was very sensitive in variant calling, its sensitivity was not perfect. A significant portion of the additional variants called by ADIScan1 may have been true variants, as GATK missed a substantially significant fraction of true positives that were detected by SOAPsnp (20). Like most variant calling algorithms, GATK prefers variants listed in dbSNP and discriminates against new variants (12), although each algorithm uses different parameters (20). Alternatively, most of the extra variants were false positives. ADIScan1 ignored the chemical specificity of the NGS method and used only VAF to call variants. We used an *Adiscore* 1 of 0.5 in a 0–1 scoring scale to call variant candidates in WBCs, which might have been too generous. There were at least two additional signs that ADIScan1 was overly sensitive and had room for further improvement. ADIScan1 called >60 000 discordant sequences by the subtraction method, compared to <25 000 by GATK, and Ti/Tv ratios for the variant candidates called by ADIScan1 were lower than the ratios called by GATK. We can adjust ADIScan1 for a higher or lower score depending on the nature of further downstream work. A high score will select variants with high confidence, but unavoidably select false negatives. A low score produces more comprehensive variant calling with increased false positives. The latter setting would be better for detecting variants in cancerous tissues that contain low-frequency clones due to relatively recent, progressive mutations. The high-score method was sensitive enough to detect all causative mutation in 103 cohorts in a study of rare genetic diseases (37). Although ADIScan may need further improvement, we believe that ADIScan1 is a reliable detection method of variants at genic regions and is compatible with and complementary to existing variant calling algorithms, including GATK. It was also adequate in detecting variants in NGS sequences with a <100× read depth. Under special circumstance where false positives are

not an issue, ADIScan1 is a better tool than GATK for the comprehensive detection of mutations. For example, discovering novel *de novo* mutations for rare genetic diseases from a single generation requires sensitive algorithms to avoid false negative rates.

Discovery of discordant sequences using two different methods

We mined discordant sequences using two different approaches: the subtraction method of variants and simultaneous comparisons of two NGS sets. Surprisingly, ∼95% of mostly true discordant sequences called by direct comparison were also called by the subtraction method (Table 3). Unlike other studies (28), this result indicated that the subtraction method between variant candidates called by ADIScan1 recognized the candidates of discordant sequences. The subtraction method was liberal, however, predicting >20 times more candidates of discordant sequences than direct comparisons of NGS between each pair of twins.

For simultaneous sequence comparisons between two tester sets of NGS, most algorithms adopted either Bayesian approaches or statistical tests (27). The ADIScan2 approach used simple ratios of allele frequencies. ADIScan2 called the least numbers of variants between MZ twins. Concordance rates were low between the results of ADIScan2 and the other two somatic variant calling methods, VarScan2 and MuTect (Figure 3 and Supplementary Figure S2). Although the numbers were similar between ADIScan2 and MuTect, they called very different sets of variants. Most variants called by ADIScan2 had VAFs that were over 0.33, compared to 0.12 for variants called by MuTect. Limited verification experiments suggested that ∼30% of discordant sequences called exclusively by ADIScan2 were true positive, which means these sequences were missed by the other two methods. It suggested that false negative rates in calling discordant sequences between MZ twins were substantial for the two methods. The reason these two highly regarded algorithms had missed discordant sequences deserves explanation. These two algorithms specialize in detecting delicate somatic mutation signals in tumor tissues by comparing them to healthy non-tumorous tissues (14,20,30). The frequencies of a minor allele at the mutated position in cancerous tissue can be similar to or lower than the error rates of NGS technologies. VarScan2 and MuTect were designed to detect low-frequency alleles of clinical relevance in cancer tissues because these are of potential clinical importance and might have originated from a subclone of cancer cells. Allele frequencies of variants called by MuTect averaged 0.12 (0.08–0.12, 90%), which was lower than variants called by ADIScan and VarScan. Heterozygote allele distribution cannot be expected in mosaic tissues composed of several clones of cells. Most tumor tissues are mixed with healthy cells, and clonal transforming and transformed cancerous cells. Further, cell aneuploidy, large genomic deletions, duplication, and subclonality within cancer-cell populations can reduce the frequency of mutated alleles to <1% (33). The ADIScan2 method also missed true discordant sequences when the differential score 25 was used. They were detected by a score of 10, however, with a dramatic increase in candidate positions to >300 000 (Supplementary Table S6). Accuracy in detecting discordant variants in this group with a differential score below 25 is expected to be low, but not zero (Figure 3). Further optimization of the algorithm is justified to detect more true variants without dramatically increasing false positive calls. We believe this method is better than subtraction methods and is useful in discovering the causative mutations of rare genetic diseases by simultaneously comparing NGS data from a patient and healthy family members.

Estimation of discordant sequences between MZ twins

ADIScan detected additional discordant sequences that VarScan2 missed. ADIScan2 appeared to be a considerable complementary method for discovering discordant sequences between MZ twins where tissue heterogeneity was low. We verified 66 true discordant sequences in two pairs of MZ twins using a Sanger sequencing method. It was possible to estimate the number of discordant sequences between MZ twins from the verification results. Based on 23 of 65 (35%) accuracy for 2860 variants called by VarScan2 in non-repetitive regions (Tables 4 and 5, Figure 3, Supplementary Figure S2), we extrapolated an average of 1012 (893 for female and 1,082 for male twins) discordant sequences between paired MZ twins in 97% (2.80×10^9 of 2.81×10^9) of the positions in the whole genome. Sequence differences could have been caused by mutations in either twin member, so the average number of mutations per individual was 506 (446 for female and 541 for male). The somatic mutation rate estimated based on this study was 1.68×10^{-7}. The number of discordant sequences between monozygotic twins was at least 100× higher than other estimations based on whole genome analyses (38,39) and ∼2× more than the approximately 300 postzygotic mutations that each individual would carry in the nuclear genome of their WBCs (35). The estimate of 300 mutations was based on the hypothesis that genetic mutations occur during DNA replication just before the human blastocyst splits into two embryos to produce MZ twins. These mutations are carried into somatic tissues, including WBCs and the germline. Mutation rates in non-malignant tissues were estimated as 1.18×10^{-8} and 2.52×10^{-7} per nucleotide, or within that range (40–43).

The primary goal of this study was to estimate the accuracy of a new algorithm for calling variants in individual genomes and discordant sequences between two genomes. We validated the mutations using the Sanger method, which detected *de novo* mutations rather than mosaicism. The estimate of mutations in an individual was an imprecise by-product. Nonetheless, it appeared that somatic point mutations arising in early development were more frequent than estimates based on the mutation rates per generation or extrapolations based on microarray experiments (35,40,42). Alternatively, sequence differences between MZ twins did not suggest that mutations occurred once during DNA replication, just before the blastocyst split, but accumulated during many cycles of cell division during embryonic and post-embryonic development (31–33). The discordant sequences between MZ twins revealed in this study underscore the need for additional research. This work could focus on the frequency and accumulation of mutations in human somatic cells in the context of genetic diseases (7,44),

Table 3. Discordant sequences in the whole genome sequence (WGS) detected by ADIScan1 between paired twins

	Before application of RepeatMasker		After application of RepeatMasker			
Sample ID	Union of difference set	Direct comparison set	Union of difference set	Direct comparison set	Common set	% of common variants in direct com. set
F1 vs F2	ND	5988	67 124	1046	990	94.6%
M1 vs M2	ND	5353	61 621	893	851	95.3%

Comparisons between data derived from the union of difference sets of variants and direct comparisons of sequence-read depth. F1 and F2 represent a pair of twin females, while M1 and M2 represent a pair of male twins. ND = not done.

Table 4. Discordant sequences identified by ADIScan2, VarScan and MuTect

	Before RepeatMasker application			After Repeatmasker application		
	ADIScan2 (VAF)	VarScan (VAF)	MuTect (VAF)	ADIScan2 (VAF)	VarScan (VAF)	MuTect (VAF)
F1 vs F2	5988 (0.33)	19 675 (0.33)	5534 (0.12)	492 (0.36)	2860 (0.29)	1237 (0.12)
M1 vs M2	5353 (0.33)	13 226 (0.32)	5242 (0.11)	474 (0.36)	2749 (0.29)	1037 (0.08)

F1 and F2 represent a pair of female twins, while M1 and M2 represent a pair of male twins. VAF = variant allele fractions.

Table 5. Validation of variants by the Sanger sequencing method

		PCR amplification		Sequencing validation	
	Positions selected for verification	Failure[a]	Success	Correct	Incorrect
VarScan2 (M)	33	19 (57%)	14 (43%)	13 (39.4%)	1
ADIScan2 (M)	78	51 (65%)	27 (35%)	24 (30.8%)	3
VarScan2 (F)	32	21 (65%)	11 (35%)	10 (31.2%)	1
ADIScan2 (F)	73	47 (64%)	26 (36%)	21 (28.8%)	5
Sum	216	138 (64%)	78 (36%)	68 (31.5%)	10

[a]The main reason for the failed validation was unsuccessful PCR amplification. M and F are the genders of the twins, male and female.

cancer research, forensic sciences (45), and human genome evolution in general (40,46).

DATA AVAILABILITY

The whole-genome sequencing data used in this study are available from Clinical and Omics data archives in the Korea National Institute of Health (http://coda.nih.go.kr/ Accession number R000132). Detailed instruction to download the data is available at the website (http://coda.nih.go.kr/coda/introduction/2/selectIntroductionView.do). Two versions of the ADIScan program, one for calling variants (ADIScan1, version 01.01) and the other for discordant sequences (ADIScan2, version 02.01) were deposited in http://genomekorea.com/display/tools/ADIscan. The files are in adiscan-test-set.tgz. A simple user manual is available on the same site and includes several options for genome analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the PGM21 genome repository and members of the data user's group for discussions related to debugging and validation, and ICGC members for their helpful feedback on early ADIScan results. We thank Fred Brooks for his critical reading and revision of the manuscript, Eungoo Jung and Haejung Yeom for the Sanger verification experiment, Minho Kim for HLA-related work, and the Syntekabio Informatics Team for assisting in data analysis.

FUNDING

Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea [HI14C0072 to J.J. and H.L.K.]; Post-Genome Technology Development Program, Developing Korean Reference Genome, funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) [10050164 to J.S.] (in part); Cheongju-si industry-academia collaboration (to Y.C.); National Institute of Biological Resources (NIBR), funded by the Ministry of Environment (MOE) of the Republic of Korea under the NIBR [2017-02-001 to H.B.L.]; INNOPOLIS Foundation [A2014DD101]; Institute for Information & Communications Technology Promotion funded by the Ministry of Science, ICT and Future Planning [B0101-15-0104 to J.J.]. Funding for open access charge: NIBR [2017-02-001].

Conflict of interest statement. We do not anticipate direct financial benefits from this publication, though indirect effects may occur due to publicity for Syntekabio, Inc., a public company. Six of the eleven co-authors are employees of Syntekabio, a company that develops systems for genetic analysis.

REFERENCES

1. Genomes Project, C., Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
2. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
3. Liu,J.H., Wei,X.X., Li,A., Cui,Y.X., Xia,X.Y., Qin,W.S., Zhang,M.C., Gao,E.Z., Sun,J., Gao,C.L. *et al.* (2017) Novel mutations in COL4A3, COL4A4, and COL4A5 in Chinese patients with Alport Syndrome. *PLoS One*, **12**, e0177685.
4. Genomes Project, C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
5. Ceccarelli,M., Barthel,F.P., Malta,T.M., Sabedot,T.S., Salama,S.R., Murray,B.A., Morozova,O., Newton,Y., Radenbaugh,A., Pagnotta,S.M. *et al.* (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, **164**, 550–563.
6. Yi,X., Liang,Y., Huerta-Sanchez,E., Jin,X., Cuo,Z.X., Pool,J.E., Xu,X., Jiang,H., Vinckenbosch,N., Korneliussen,T.S. *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.
7. Lee,H., Deignan,J.L., Dorrani,N., Strom,S.P., Kantarci,S., Quintero-Rivera,F., Das,K., Toy,T., Harry,B., Yourshaw,M. *et al.* (2014) Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*, **312**, 1880–1887.
8. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.
9. Handel,A.E., Disanto,G. and Ramagopalan,S.V. (2013) Next-generation sequencing in understanding complex neurological disease. *Expert Rev. Neurotherapeut.*, **13**, 215–227.
10. Taghavi,S., Chaouni,R., Tafakhori,A., Azcona,L.J., Firouzabadi,S.G., Omrani,M.D., Jamshidi,J., Emamalizadeh,B., Shahidi,G.A., Ahmadi,M. *et al.* (2018) A Clinical and Molecular Genetic study of 50 families with autosomal recessive parkinsonism revealed known and novel gene mutations. *Mol. Neurobiol.*, **55**, 3477–3489.
11. Quail,M.A., Smith,M., Coupland,P., Otto,T.D., Harris,S.R., Connor,T.R., Bertoni,A., Swerdlow,H.P. and Gu,Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
12. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
13. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
14. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
15. Li,R., Li,Y., Fang,X., Yang,H., Wang,J., Kristiansen,K. and Wang,J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.
16. Lyon,G.J. and Wang,K. (2012) Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Med.*, **4**, 58.
17. Clement,N.L., Snell,Q., Clement,M.J., Hollenhorst,P.C., Purwar,J., Graves,B.J., Cairns,B.R. and Johnson,W.E. (2010) The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, **26**, 38–45.
18. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
19. Wei,Z., Wang,W., Hu,P., Lyon,G.J. and Hakonarson,H. (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, **39**, e132.
20. O'Rawe,J., Jiang,T., Sun,G., Wu,Y., Wang,W., Hu,J., Bodily,P., Tian,L., Hakonarson,H., Johnson,W.E. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.*, **5**, 28.
21. Saunders,C.T., Wong,W.S., Swamy,S., Becq,J., Murray,L.J. and Cheetham,R.K. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, **28**, 1811–1817.
22. Shiraishi,Y., Sato,Y., Chiba,K., Okuno,Y., Nagata,Y., Yoshida,K., Shiba,N., Hayashi,Y., Kume,H., Homma,Y. *et al.* (2013) An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.*, **41**, e89.
23. Larson,D.E., Harris,C.C., Chen,K., Koboldt,D.C., Abbott,T.E., Dooling,D.J., Ley,T.J., Mardis,E.R., Wilson,R.K. and Ding,L. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.
24. Roth,A., Ding,J., Morin,R., Crisan,A., Ha,G., Giuliany,R., Bashashati,A., Hirst,M., Turashvili,G., Oloumi,A. *et al.* (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**, 907–913.
25. Koboldt,D.C., Larson,D.E. and Wilson,R.K. (2013) Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr.Protoc. Bioinformatics*, **44**, doi:10.1002/0471250953.bi1504s44.
26. Li,H. (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, **28**, 1838–1844.
27. Kroigard,A.B., Thomassen,M., Laenkholm,A.V., Kruse,T.A. and Larsen,M.J. (2016) Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One*, **11**, e0151664.
28. Roberts,N.D., Kortschak,R.D., Parker,W.T., Schreiber,A.W., Branford,S., Scott,H.S., Glonek,G. and Adelson,D.L. (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, **29**, 2223–2230.
29. Ka,S., Lee,S., Hong,J., Cho,Y., Sung,J., Kim,H.N., Kim,H.L. and Jung,J. (2017) HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics*, **18**, 258.
30. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
31. Vattathil,S. and Scheet,P. (2016) Extensive hidden genomic mosaicism revealed in normal tissue. *Am. J. Hum. Genet.*, **98**, 571–578.
32. Machiela,M.J. and Chanock,S.J. (2017) The ageing genome, clonal mosaicism and chronic disease. *Curr. Opin. Genet. Dev.*, **42**, 8–13.
33. Abyzov,A., Tomasini,L., Zhou,B., Vasmatzis,N., Coppola,G., Amenduni,M., Pattni,R., Wilson,M., Gerstein,M., Weissman,S. *et al.* (2017) One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. *Genome Res.*, **27**, 512–523.
34. Jorde,L.B. and Wooding,S.P. (2004) Genetic variation, classification and 'race'. *Nat. Genet.*, **36**, S28–S33.
35. Li,R., Montpetit,A., Rousseau,M., Wu,S.Y., Greenwood,C.M., Spector,T.D., Pollak,M., Polychronakos,C. and Richards,J.B. (2014) Somatic point mutations occurring early in development: a monozygotic twin study. *J. Med. Genet.*, **51**, 28–34.
36. Chakravarti,A. (2011) Genomic contributions to Mendelian disease. *Genome Res.*, **21**, 643–644.
37. Cho,Y., Lee,C.-H., Jeong,E.-G., Kim,M.-H., Hong,J.H., Ko,Y., Lee,B., Yun,G., Kim,B.J., Jung,J. *et al.* (2017) Prevalence of rare genetic variations and their implications in NGS-data interpretation. *Scientific Rep.*, **7**, 9810.
38. Ju,Y.S., Martincorena,I., Gerstung,M., Petljak,M., Alexandrov,L.B., Rahbari,R., Wedge,D.C., Davies,H.R., Ramakrishna,M., Fullam,A. *et al.* (2017) Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, **543**, 714–718.
39. Dal,G.M., Erguner,B., Sagiroglu,M.S., Yuksel,B., Onat,O.E., Alkan,C. and Ozcelik,T. (2014) Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J. Med. Genet.*, **51**, 455–459.

40. Conrad,D.F., Keebler,J.E., DePristo,M.A., Lindsay,S.J., Zhang,Y., Casals,F., Idaghdour,Y., Hartl,C.L., Torroja,C., Garimella,K.V. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.
41. Roach,J.C., Glusman,G., Smit,A.F., Huff,C.D., Hubley,R., Shannon,P.T., Rowen,L., Pant,K.P., Goodman,N., Bamshad,M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.
42. Xue,Y., Wang,Q., Long,Q., Ng,B.L., Swerdlow,H., Burton,J., Skuce,C., Taylor,R., Abdellah,Z., Zhao,Y. *et al.* (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.: CB*, **19**, 1453–1457.
43. Nachman,M.W. and Crowell,S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
44. Petersen,B.S., Spehlmann,M.E., Raedler,A., Stade,B., Thomsen,I., Rabionet,R., Rosenstiel,P., Schreiber,S. and Franke,A. (2014) Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics*, **15**, 564.
45. Weber-Lehmann,J., Schilling,E., Gradl,G., Richter,D.C., Wiehler,J. and Rolf,B. (2014) Finding the needle in the haystack: differentiating 'identical' twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Sci. Int. Genet.*, **9**, 42–46.
46. Messer,P.W. (2009) Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics*, **182**, 1219–1232.