Characterization of intermediate-sized insertions using whole genome sequencing data and analysis of their functional impacts on gene expression

Saeideh Ashouri

University of Tokyo: Tokyo Daigaku

Jing Hao Wong

University of Tokyo: Tokyo Daigaku

Hidewaki Nakagawa

RIKEN: Rikagaku Kenkyujo

Mihoko Shimada

University of Tokyo: Tokyo Daigaku

Katsushi Tokunaga

NCGM: Kokuritsu Kenkyu Kaihatsu Hojin Kokuritsu Kokusai Iryo Kenkyu Center

Akihiro Fujimoto (✉ afujimoto@m.u-tokyo.ac.jp)

University of Tokyo: Tokyo Daigaku <https://orcid.org/0000-0002-0075-0800>

Original Article

Keywords: Structural variants, Intermediate-sized insertions, Whole-genome sequencing, Whole-genome imputation, Expression quantitative trait loci (eQTL), Long-read sequencing

Posted Date: February 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-203007/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Intermediate-sized insertions are one of the structural variants contributing to genome diversity. However, due to technical difficulties in identifying them, their importance in disease pathogenicity and gene expression regulation remains unclear. We used whole-genome sequencing data of 174 Japanese samples to characterize intermediate-sized insertions using a highly-accurate insertion calling method (IMSindel software and joint-call recovery) and obtained a catalogue of 4,254 insertions. We constructed an imputation panel comprising of insertions and SNVs from all samples, and conducted imputation of intermediate-sized insertions for 82 publicly-available Japanese samples. Imputation accuracy, evaluated using Nanopore long-read sequencing data, was 97%. Subsequent eQTL analysis predicted 128 (~ 3.0%) insertions as causative for gene expression level changes. Enrichment analysis of causal insertions for genome regulatory elements showed significant associations with CTCF-binding sites, super-enhancers, and promoters. Among 17 causal insertions found in the same causal set with GWAS hits, there were insertions associated with changes in expression of cancer-related genes such as *BRCA1*, *ZNF222*, and *ABCB10*. Analysis of insertions sequences revealed that 461 insertions were short tandem duplications frequently found in early replicating regions of genome. Furthermore, comparison of functional importance of intermediate-sized insertions with that of intermediate-sized deletions detected in the same sample set in our previous study showed that insertions were more frequent in genic regions, and proportion of functional candidates was smaller in insertions. Here, we characterize a high-confidence set of intermediate-sized insertions and indicate their importance in gene expression regulation. Our results emphasize the importance of considering intermediate-sized insertions in trait association studies.

Introduction

The rapid advances in DNA sequencing technology in the past 10 years have significantly enabled comprehensive characterization of a diverse number of genome variations. Genome variations range in length from as short as 1bp, called single nucleotide variants (SNVs), to several mega-bases, called structural variants (SVs). SVs include insertions, deletions, inversions, duplications, translocations, and other rearrangements (Weischenfeldt et al. 2013; Guan and Sung 2016). SVs are larger than SNVs; therefore, compared with SNVs, they contribute to a higher nucleotide sequence difference between individuals in populations (0.5–1% for structural variants versus 0.1% for SNVs) (Pang et al. 2010). Recent human genetic studies have shown that the presence of SVs in the genome, whether intersecting with genes structures or being in intergenic regions, has phenotypic consequences (Weischenfeldt et al. 2013). The association of SVs with hereditary and rare, sporadic conditions as well as common and complex diseases such as type I and type II diabetes, Crohn's disease, rheumatoid arthritis, autism, and schizophrenia has been previously implicated (The International Schizophrenia Consortium 2008; The Wellcome Trust Case Control Consortium 2010; Psychosis Endophenotypes International Consortium et al. 2017; Brandler et al. 2018). Despite the importance of SVs in both contributing to the genome structure and functional consequences, they are understudied due to the difficulty in their discovery and genotyping (Mahmoud et al. 2019). In addition, lack of high-quality publicly available variant maps

obtained from large populations has made it challenging to interpret SVs in whole-genome sequencing based studies.

Among different forms of SVs, thousands of insertions have been reported to exist in the human genome (The 1000 Genomes Project Consortium et al. 2015). In addition to the comprehensive picture of intermediate-sized insertions, their underlying causative mechanisms and functional effects are still poorly understood. Some insertions originate from transposable elements (TEs) - DNA mobile elements that replicate and insert themselves into different locations throughout the host genome - with Alu, L1, and SVA (SINE-R, VNTR, and Alu) families, still active in the human genome (Stewart et al. 2011; Tang et al. 2018). Some other insertions, especially shorter ones, are caused by polymerase slippage (Taylor 2004; Montgomery et al. 2013). However, detailed mechanisms have yet been analyzed and there may be a new class of insertions in the human genome.

Our understanding about intermediate-sized insertions and deletions (30 – 5,000bp) is far from complete. In our previous study, we focused on intermediate-sized deletions, and established a highly accurate method for identifying them from whole genome sequencing (WGS) data (Wong et al. 2019). The catalogue of identified deletions was subsequently used for imputation in a separate sample set, and the imputation results were used for an eQTL association analysis. The previous study identified 4,378 high-confidence deletions in 174 Japanese individuals and revealed that 4.1% (181/4,378 deletions) of them can change gene expression levels. It indicated the importance of considering intermediate-sized deletions in human genetics studies. However, the previous study especially focused on deletions and we lack the comprehensive understanding of insertions.

In the current study, we sought to identify and characterize the intermediate-sized insertions from the same dataset used in the previous study (Wong et al. 2019). The purposes of the current study are establishment of an intermediate-sized insertion calling method and evaluation of their functional importance. We especially focused on the functional differences between intermediate-sized insertions and intermediate-sized deletions. We identified 4,254 polymorphic intermediate-sized insertions in a Japanese population. Imputation and the subsequent eQTL analysis showed that 128 (~ 3%) of them can affect gene expression levels. The accuracies of insertion calling as well as imputation at the whole-genome level were evaluated using Oxford Nanopore long-read sequencing data. Our study also indicated that 17 insertions could be causative variants of GWAS traits. This study shows an accurate way for identifying intermediate-sized insertions from WGS data as well as implication of insertions in regulation of gene expression.

Methods

Samples and data

Whole-genome sequencing data from 174 ICGC Japanese samples (Fujimoto et al. 2016a), which were used for calling intermediate-sized deletions in our previous study (Wong et al. 2019), was used for

insertion calling in the current study. The procedures of the current study were approved by the institutional review boards (IRBs) at RIKEN and all participating groups.

Calling insertion candidates by IMSindel software

We identified insertions as described previously (Wong et al. 2019). In brief, short reads obtained from WGS data of 174 Japanese samples were mapped to the reference genome. Reads were then classified into three types, mapped reads, unmapped reads, and soft-clipped reads. We used the IMSindel software (Shigemizu et al. 2018) for calling insertions from soft-clipped reads with the options “–indelsize 10000 –alt-read-depth 2 –support-reads 2”. Very short insertion candidates with length less than 10bp were excluded from the SV calling results. We then merged remaining insertion candidates if their positions were located within 10bp from each other. In order to make sure that short indels are removed from the data, we excluded insertions with average length less than 20bp.

Filtering insertion candidates based on their annotations

Annotation of insertion candidates was conducted according to the position of insertions relative to different regions of the genome. Annotation of genes, exons, telomeres, centromeres, simple repeats, and repeat-masked regions was conducted using the list of positions of the genome downloaded from UCSC Table Browser (Casper et al. 2018). Annotation of UTRs was performed using ANNOVAR (Wang et al. 2010). We also annotated insertions for micro-satellites using an in-house comprehensive list of micro-satellites (Fujimoto et al. 2020). For filtering out insertions which were possibly false positives due to misalignments of short reads, we applied a filtering strategy similar to the one used in our previous study (Wong et al. 2019). As identification of insertions in regions such as telomeres, centromeres, simple repeats, micro-satellites, satellites, and low-complexity DNA is difficult due to alignment errors caused by their repetitive identity, insertions located in these genomic regions were removed from further analysis. We also excluded insertions with read depth over 50 and variant allele frequency (VAF) less than 0.1.

Joint-call recovery of insertion candidates

After mapping short reads, IMSindel uses soft-clipped reads, unmapped reads, and mapped reads to construct consensus sequences, and detects the insertion sequence by comparing the consensus sequence to the reference genome (Shigemizu et al. 2018). Because of local GC contents or random fluctuation, the quality and read depth is not the same in all samples so insertions detected in the same position in each sample can be called by soft-clipped reads in the forward-oriented reads (forward-type), reverse-oriented reads (reverse-type), or by both-types reads. Insertions detected by reads of both types are probably from regions with higher read depth or quality. To ensure the quality of detected insertions, we applied a joint-call recovery method, similar to the one used in our previous study (Wong et al. 2019), on all candidate insertions so that insertions detected by both-type reads in all samples, those detected by both-type reads in some samples and by forward-type and/or reverse-type reads in others, and those detected by forward-type reads in some samples and by reverse-type reads in others were retained, and others were excluded. After applying joint-call recovery, we removed insertions with allele frequency < 0.05 and Hardy-Weinberg Equilibrium (HWE) test p-value < 0.0001 (Fisher’s exact test).

Imputation of intermediate-sized insertions to the study subjects

After all filtering steps, a highly-accurate catalogue of intermediate-sized insertions was obtained. We combined this catalogue with 5,239,921 biallelic variants (SNVs and short indels) obtained from the same sample set (Fujimoto et al. 2016a) to construct an imputation panel. Before combining the data, WGS data was filtered for minor allele frequency (MAF) < 0.05 and HWE test p-value < 0.0001 . Then, this reference panel was used for imputation of intermediate-sized insertions into genomic data of 82 Japanese samples obtained from the 1000 Genomes Project database (The 1000 Genomes Project Consortium 2015). The imputation was conducted using IMPUTE2 software (Howie et al. 2009) with the options “-phase -iter 40 -Ne 20000”.

Evaluation of insertion detection and imputation accuracy

In order to assess the accuracy of insertion calling by IMSindel software and imputation, we compared insertion calls from short-reads with those from Nanopore long-read sequencer as done previously (Wong et al. 2019). Accuracy of IMSindel calling was evaluated by using long-read sequencing data of RK067 (one of ICGC samples). Accuracy of imputation was estimated by using long-read sequencing data of NA18943 (one of 1000 Genomes project samples). Variants predicted to be homozygote or heterozygote for insertion in NA18943 by imputation, and detected as insertions by Nanopore long-read sequencer were considered as true positives and those predicted to be homozygote for reference allele in NA18943 by imputation and detected as insertion by Nanopore long-read sequencer were considered as false negatives. True positive and false negative rates were then calculated. In addition to using Nanopore long-read sequencing data of NA18943, we evaluated the effectiveness and accuracy of imputation results by PCR and gel electrophoresis. We selected 17 insertion candidates and checked for their presence in twenty-one 1000 Genomes JPT individuals.

eQTL association analysis and estimation of causal insertions

In order to evaluate the potential functional impact of the imputed insertion candidates in 82 Japanese samples, the gene expression data of same sample set (Stranger et al. 2012) obtained from the EMBL-EBI ArrayExpress database was used for eQTL association analysis. The eQTL analysis was conducted using the MatrixEQTL R package (Shabalin 2012) for common insertions and SNVs with MAF > 0.05 and HWE test p-value > 0.0001 . To perform analysis for local (*cis*) eQTLs, a *cis*-window of 1 Mb around the transcription start site (TSS) of each gene was defined. After obtaining the results of analysis, we conducted multiple testing correction for each *cis*-window with Benjamini-Hochberg (BH) method at 1% false discovery rate (FDR). We made a Quantile-Quantile (Q-Q) plot of observed p-values versus expected p-values on a-log 10 scale and also calculated the genomic inflation factor. To predict the insertions probably causal for gene expression level changes, we selected the most significant insertions as well as the 100 most significant SNVs or indels (based on nominal p-value) in the *cis* window and estimated their

pairwise LD using PLINK (ver. 1.9). The summary statistics (z-scores) of eQTL association analysis for all insertions and SNVs and pair-wise correlations were given to CAVIAR software (Hormozdiari et al. 2014), and predicted causal insertions were contained within the set of suggested causal variants by CAVIAR.

Annotation of insertion candidates for regulatory features and enrichment analysis

We next sought to examine the genomic context of insertion-eQTLs to search for hints into their possible molecular mechanisms. We hypothesized that causal insertions would be enriched for genome regulatory elements such as enhancers, promoters, and repressors. The candidate insertions, including all causal, non-causal, and non-significant insertions in the eQTL association results were annotated according to their positions relative to different regulatory features of the genome such as super-enhancers, promoters, transcription factor binding sites (TFBS), microRNA binding sites, and predicted chromatin states. As the gene expression data used for eQTL association analysis was obtained from GM12878 B-lymphocyte cell line, we also annotated insertions for the regulatory features specific to this cell line. We then sought to determine if there was any enrichment of regulatory features in causal insertions compared to other insertions. We also compared the causal insertion candidates causing gene expression increase against those causing gene expression decrease for enrichment in regulatory features. All enrichment analyses were conducted using Fisher's exact test. We also compared the allele frequency of causal insertion candidates to other insertions using the Wilcoxon rank-sum test.

To determine whether there is a significant difference in average conservation score between causal and other insertions, we obtained conservation scores for 45 vertebrate genomes compared to the human genome from the UCSC genome browser (Siepel et al. 2005). We calculated the average conservation scores for the 100bp flanking regions of either side of all insertions positions, and compared the average conservation scores for causal insertion candidates to those of other insertions using the Wilcoxon rank-sum test.

We also sought to investigate the contribution of causal insertions to trait-associated loci. We first downloaded GWAS SNPs with significant disease association ($p\text{-value} \leq 5.0 \cdot 10^{-8}$) from the NHGRI-EBI Catalogue of published GWASs (MacArthur et al. 2017) and checked the presence of previously reported GWAS SNPs in CAVIAR causal set lists. The candidate eQTL insertion included in the same causal set with a GWAS SNP was considered to be linked with that particular GWAS SNP.

Investigating the identity and causative mechanisms of insertions

Furthermore, we sought to examine the identity of insertion sequences. Insertions sequences were obtained from the output of IMSindel software and available Oxford Nanopore long-read sequencing data of 2 samples. We used RepeatMasker software (Hubley et al. 2013) to check insertions sequences for interspersed repeats and low complexity DNA sequences. Additionally, the BLAT alignment tool (Kent 2002) was used on insertion sequences to gain insights into their origin.

Comparison of features as well as functional impacts between intermediate-sized insertions and deletions

We compared intermediate-sized insertions against intermediate-sized deletions characterized in our previous study (Wong et al. 2019) for different features such as the number of causal variants and their location relative to genes and gene parts. Additionally, in order to evaluate the relative contribution of insertions and deletions in gene expression regulation, we compared them for enrichment in genome regulatory elements.

Results

Generation of a high-confidence catalogue of insertions by filtering and joint-call recovery

After calling insertion from WGS data of 174 ICGC Japanese samples (Fujimoto et al. 2016a) using the IMSindel software, 49,877,765 insertions were detected. After excluding very short insertions (length < 10bp), merging insertion candidates based on at most 10bp difference in their positions and filtering out merged insertions for average length (< 20bp), 93,900 insertions remained (Fig. 1). Excluding insertions located in repetitive regions of the genome led to 55,238 insertion candidates. After applying the joint-call recovery method similar to the one used in our previous study (Wong et al. 2019), the number of insertions decreased to 11,591. Excluding insertion using population filters such as HWE test (p-value < 0.0001) and insertion allele frequency (< 0.05) resulted in 4,254 high-confidence insertion candidates for further analysis (Fig. 1 and Additional file 2; Table S1). From 4,254 remaining insertions, 1,944 (45.7%) were located inside genes of which 78 (4%) were located in exons, 56 (2.9%) were located in 3' and 5' UTRs, and 1,810 (93.1%) were located in introns (Fig. 2).

Evaluation of insertion detection accuracy using Nanopore long-read sequencing technology

Comparison of insertions called by IMSindel with those detected by Oxford Nanopore long-read sequencer in sample RK067 showed that before all filtering steps, the consistency rate of detected insertion candidates was 45%. After filtering for annotation, the true positive rate increased to 82%, and after applying the joint-call recovery method and population filters, the true positive rate increased to 98% (Fig. 3), indicating a highly-accurate catalogue of intermediate-sized insertions ready to be used for imputation. Insertion candidates detected by Oxford Nanopore long-read sequencer were considered as true positives. Examples of true positive and false positive insertions are provided in Additional file 1, Figure S1.

Evaluation of the imputation accuracy

We merged the highly-accurate catalogue of insertions with other variants and constructed an imputation panel with 5,244,175 variants. After conducting imputation of intermediate-sized insertions into 82 Japanese samples from the 1000 Genomes project, we filtered imputation results for HWE test p-value < 0.0001 and MAF < 0.05. We then used the long-read sequencing data of sample NA18943 for evaluating the accuracy of imputation. The true positive rate of imputed insertions in NA18943 was 97% and the false negative rate was 7%. Comparison of allele frequency distribution of true positive insertions versus false negatives showed lower allele frequency for false negative insertions (Additional file 1, Figure S2) which is expected as variants with lower allele frequency suffer from lower imputation quality.

We also conducted PCR and gel electrophoresis for experimental validation of imputation results for 17 insertions in 21 samples. From 357 genotypes tested, the zygosity of 351 genotypes was consistent between imputation results and PCR results which shows a high consistency rate of 98.3%. The list of validated insertions and the results of experimental validation are provided in Additional file 2, Table S2.

eQTL association analysis and prediction of causal insertions

We conducted an eQTL association analysis using imputation results along with the gene expression data from the same sample set (Additional file 1, Figure S3). The association analysis determined 184 significant eQTL-insertions (q-value < 0.01) after correcting the association results for multiple testing in each 1Mb cis-window using BH correction method. There was no systemic inflation in the test statistics as the genomic inflation factor (λ) was 1.045 (Additional file 1, Figure S4). CAVIAR predicted 128 insertions to be possibly causal for gene expression level changes (Additional file 1, Figure S5). Among causal insertions, there were some functionally interesting insertion candidates affecting the expression of multiple genes (Additional file 2, Table S3). For example, an insertion on chromosome 19 (chr19:9254308) was found to be associated with decreased expression of 5 different genes (*ANGPTL6*, *FBXL12*, *ZNF558*, *ZNF414*, and *ZNF426*). Another insertion on the same chromosome (chr19:44176872) was associated with decreased expression of 5 members of the clustered ZNF gene family on chromosome 19q13.3 (*ZNF180*, *ZNF222*, *ZNF227*, *ZNF428*, and *ZNF576*), encoding functionally important Krüppel-type (or C2H2) zinc finger transcription factor (ZNF) proteins. There were also some insertions with both increasing and decreasing effects on gene expression levels. For example, an insertion on chromosome 5 (chr5:139595135), located in the gene *CYSTM1*, was found to be causative for increased expression of the genes *HARS2*, *ANKHD1*, and *APBB3*, and decreased expression of the gene *CD14*. In another case, a causal insertion on chromosome 20 (chr20:44380443) was associated with decreased expression of the genes *NEURL2* and *SLC12A5* and with increased expression of the gene *SPINT4*. These results reemphasize the important functional role of intermediate-sized insertions in the regulation of gene expression in the human genome.

Comparison of causal versus non-causal insertion candidates for different genomic features

From 128 causal insertion candidates, 65 of them (51%) were located inside genes and others (49%) were located in inter-genic regions. From those located inside genes, 2 insertions were located inside 5' UTRs of the genes *VAMP4* and *DUSP4*, and 1 insertion was located inside 3' UTR of the gene *SMARCE1*. Despite being located in UTRs, none of them were associated with any changes in the expression of their harboring genes. This led us consider the possibility of enrichment of causal insertions for other regulatory elements of the genome, affecting the gene expression levels through regulatory mechanisms.

Annotation of intermediate-sized insertions for regulatory features and predicted chromatin states is provided in Additional file 2, Table S3. The results of enrichment analysis of causal versus other insertion candidates for regulatory elements of the genome revealed highly significant enrichment of causal insertions for CTCF binding sites ($p\text{-value} = 5.42 \cdot 10^{-5}$, OR = 2.09) (Fig. 4a and Additional file 2, Table S4). Positive enrichment was also observed for super-enhancers ($p\text{-value} = 0.02$, OR = 1.6) and promoters ($p\text{-value} = 0.03$, OR = 4.92) as well as some predicted chromatin states such as transcription elongation ($p\text{-value} = 4.8 \cdot 10^{-4}$, OR = 2.47), and transcription transition ($p\text{-value} = 3.7 \cdot 10^{-3}$, OR = 3.5), while negative enrichment was observed for heterochromatin ($p\text{-value} = 1.51 \cdot 10^{-5}$, OR = 0.39) (Additional file 2, Table S4). Analysis of enrichment of genome regulatory elements specific to GM12878 cell line also showed enrichment of causal insertion candidates in some features such as TFBS ($p\text{-value} = 0.01$, OR = 2.18), as well as some predicted chromatin states such as weak transcription ($p\text{-value} = 0.04$, OR = 1.65), active promoter ($p\text{-value} = 0.03$, OR = 3.37), and transcription elongation ($p\text{-value} = 0.01$, OR = 2.15) (Fig. 4b and Additional file 2, Table S5). Negative enrichment of heterochromatin ($p\text{-value} = 5.83 \cdot 10^{-6}$, OR = 0.42) was observed in GM12878 cell line specific regulatory elements too. Comparison of causal insertions causing a decrease in gene expression versus those causing an increase in gene expression only showed negative enrichment of causal insertion candidates with gene decreasing effects in repressed chromatin ($p\text{-value} = 0.02$, OR = 0.35) (Additional file 2, Table S6).

We also compared causal insertions against other insertion candidates for allele frequency. Causal insertions showed a significantly higher allele frequency ($p\text{-value} = 0.015$) compared to that of other insertions (Additional file 1, Figure S6). Finally, comparison of average conservation score of the flanking 100bp regions of insertions positions between causal and other insertions indicated that causal insertions were located in regions with significantly higher conservation scores (Wilcoxon rank-sum test $p\text{-value} = 3.2 \cdot 10^{-3}$; Additional file 1, Figure S7).

GWAS associated insertion-eQTLs

In an attempt to identify insertions that likely explain prior GWAS results, we found 17 causal intermediate-sized insertions in the same set of predicted causal variants with GWAS hits (Additional file 2, Table S3). The GWAS-associated variants have been reported to be significantly associated with traits such as LDL measurement, breast cancer, homocysteine measurement, platelet count, basal cell carcinoma, allergy, osteoarthritis, Non-small-cell lung carcinoma, etc.

Investigating the identity and causative mechanisms of insertions

The full sequence for a total of 3,639 insertions (20bp-8.6kb) were obtained from IMSindel output and Nanopore long-read sequencing data. Analysis of insertions sequences by RepeatMasker software (Hubley et al. 2013) revealed that 946 of the 3,639 insertions (26%) belonged to different families of TEs and repetitive elements such as SINEs, LINEs, LTRs, SVA, low complexity DNA, and simple repeats.

Our analysis also identified a previously understudied class of insertions. Alignment of insertion sequences to the human genome using BLAT showed that for 461 of the 3,639 insertions (12.7%) with length from 20 to 51bp, the inserted sequence was identical to the sequence immediately adjacent to the site of the insertion event. We considered them as short tandem duplications. From these 461 short tandem duplications, 6 were observed to have 2 copies of the sequence adjacent to the insertion position (Additional file 2, Table S7). We selected 5 insertions for validation, and all of them were successfully detected by Sanger sequencing (Additional file 2, Table S8). We next examined enrichment of short tandem duplications for genes and repetitive elements of the genome. We observed that short tandem duplications frequently occurred inside genes (p-value = 0.003, OR = 1.34) and exons (p-value = 0.009, OR = 2.16). However, a negative enrichment was observed for their occurrence inside repeat-masked regions (p-value = 0.003, OR = 0.74) and in the vicinity of repeat-masked regions and microsatellites (p-value = $3.45 \cdot 10^{-5}$, OR = 0.65 for the vicinity of repeat masked regions and p-value = $8.75 \cdot 10^{-7}$, OR = 0.52 for the vicinity of microsatellites). Since previous studies reported association of SVs frequency and replication timing (De and Michor 2011; Fujimoto et al. 2016b), we sought to examine the association between DNA replication timing and occurrence of short tandem duplications. Data for replication timing in different cell types was downloaded from the ENCODE Project (Dunham et al. 2012), and replication timing for short tandem duplications was compared to that of other insertions. We observed that short tandem duplications frequently occurred in early-replicating regions of the genome (Fisher's exact test p-value = 0.017; Fig. 5, Additional file 1, Figure S8), but other insertions did not.

Comparison of features as well as functional effects of deletions and insertions

The total number of insertions after all filtering steps (4,254 insertions) in the current study is comparable to that of our previous study (Wong et al. 2019) on intermediate-sized deletions (4,378 deletions). The proportion of insertions located inside genes (1,944/4,254) was significantly higher (p-value = $2.99 \cdot 10^{-10}$, OR = 1.31, Fisher's exact test) than that of deletions (1,707/4,378) (Additional file 2, Table S9). Similar patterns were observed for insertions and deletions located inside gene parts such as UTRs and exons (Fig. 6). A significantly higher number of insertions (56/4,254) was observed to be located in UTRs (p-value = $1.56 \cdot 10^{-5}$, OR = 2.9) than that of deletions (20/4,378). Similarly, the number of insertions located inside exons was significantly higher (p-value = $6.1 \cdot 10^{-4}$, OR = 1.92) than that of deletions (78/4,254 for insertions versus 42/4,378 for deletions) (Additional file 2, Table S9). Even after excluding the short tandem duplications from the insertions list and conducting the tests, the proportion of insertions located inside genes, exons, and UTRs was still higher than that of deletions (inside genes: p-value = $6.72 \cdot 10^{-8}$, OR = 1.27; inside exons: p-value = $7.4 \cdot 10^{-3}$, OR = 1.72; inside UTRs: p-value = $1.07 \cdot 10^{-4}$, OR = 2.73). The

difference in occurrence of these variants in exons might be explained by more disruptive effects of deletions on protein functions which is less tolerated by natural selection.

The number of insertions causative for gene expression level changes in the current study (128/4,254) was significantly lower ($p\text{-value} = 5.3 \cdot 10^{-3}$, $OR = 0.72$) compared to the intermediate-sized deletions (181 causative deletions) (Additional file 2, Table S9) in the previous study (Wong et al. 2019) which could be possibly explained by more deleterious role of deletions in the human genome. We found no significant differences in the number of causal deletions causing increase and decrease in gene expression level with that of insertions (Additional file 2, Table S10). In addition, no significant differences were observed for causal insertions and causal deletions located inside genes, exons, and UTRs, and linked to the GWAS hits (Additional file 2, Table S10).

We also compared enrichment of causal insertions and causal deletions for regulatory elements of the genome as well as predicted chromatin states (Additional file 2, Table S10). There was a significant enrichment of insertions in CTCF-binding sites ($p\text{-value} = 0.03$, $OR = 1.63$). No significant enrichment was observed for other elements and chromatin states. Comparison of enrichment of insertions and deletions for regulatory elements specific to GM12878 cell line also showed that insertions were negatively enriched for CTCF-binding sites ($p\text{-value} = 0.024$, $OR = 0.29$), repetitive/CNV state ($p\text{-value} = 9.8 \cdot 10^{-3}$, $OR = 0.1$), and weak enhancers ($p\text{-value} = 1.1 \cdot 10^{-3}$, $OR = 0.12$) (Fig. 6, Additional file 2, Table S11).

Discussion

In the current study a catalogue of highly-accurate intermediate-sized insertions was successfully identified using short-read sequencing data obtained from 174 Japanese samples and imputed into a separate sample set comprising 82 Japanese samples. Using short reads for characterizing different types of SVs, especially intermediate-sized insertions, is challenging. However, we demonstrated that using our filtering and processing steps, intermediate-sized insertions could be accurately identified. The abundance, biological roles, and clinical implications of insertions in the human genome (Hancks and Kazazian 2016) makes the need for accurate and efficient way of detecting them inevitable. Validation of insertion calls using Oxford Nanopore long-read sequencing data with 98% true positive rate, validation of imputation results using Oxford Nanopore long-read sequencing data with 97% true positive rate and by PCR with 98.3% true positive rate showed that our method is a reliable way for characterizing and subsequent analysis of insertions in the human genome.

Imputation results were used for eQTL analysis. From 184 significant eQTL-insertions detected, 128 of them (3% of all imputed insertions) were predicted to be possibly causal for gene expression level changes. Interestingly, we observed that some causal insertions were associated with changes in the expression of several genes. As only 2 causal insertions were located inside UTRs, we showed enrichment of the insertions for regulatory elements of the genome such as enhancers, promoters, CTCF binding sites, etc. These results indicate that insertions may also operate as common variants affecting gene expression levels and subsequently the risk of different diseases. Our findings are consistent with those

of a study on the impact of different SVs on human gene expression by GTEx Consortium (Chiang et al. 2017), that showed the majority (88.3%) of SVs predicted to directly alter gene expression levels were located outside coding regions of the genome and enriched for enhancers and other regulatory elements.

We also observed that 17 causal insertion candidates (13.3% of all causal insertions) were in the same causal set with GWAS hits involved in diseases such as breast carcinoma, type I and type II diabetes, schizophrenia, allergy, basal cell carcinoma, and non-small cell lung cancer. Functional studies and experimental validations are needed to confirm the causal relationships between any of these insertions and GWAS traits. However, there are some promising insertion candidates among these 17 loci. In one case, an insertion located in the intron of *BRCA1* gene at chr17:41223538 was associated with increased expression of the gene *BRCA1* (corrected p-value = $9.25 \cdot 10^{-9}$), known to be associated with the risk of breast cancer, and in the same causal set with rs1799949 reported to be an eQTL for *BRCA1* gene in GTEx portal database (Lonsdale et al. 2013) and a susceptibility locus for age at menopause (Day et al. 2015). In another case, an insertion located at chr19:44176872 was found to be associated with decreased expression of the gene *ZNF222* (corrected p-value = $4.27 \cdot 10^{-3}$) and linked to rs3760982, a SNP found to be associated with the expression of *ZNF222* (Michailidou et al. 2013) and the risk of breast cancer (Michailidou et al. 2013; Michailidou et al. 2015; Michailidou et al. 2017). In another case, an insertion at chr1:228921863, located in a super-enhancer, was observed to be associated with increased expression of *ABCB10* gene (corrected p-value = $3.7 \cdot 10^{-3}$) and in the same causal set with rs61824911, a GWAS susceptibility locus for basal cell carcinoma (BCC) (Chahal et al. 2016). Multiple studies have reported the gene *ABCB10* as a susceptibility locus for different types of cancers, such as breast cancer, clear cell renal cell carcinoma, hepatocellular carcinoma, and epithelial ovarian cancer (Liang et al. 2017; Chen et al. 2019; Huang et al. 2019; Fu et al. 2019), and its increased expression has been reported in a retinoblastoma cell line (Hendig et al. 2009). Future studies on SVs can contribute to deeper understanding on mechanism of disease.

By aligning the insertions sequences to the human reference genome, we observed that sequences of 461 candidate insertions were identical to the sequence adjacent to the insertion position (short tandem duplications). Similar patterns have been observed for short indels, previously (Taylor 2004; Montgomery et al. 2013). We also found a significant association between short tandem duplications and replication timing. Short tandem duplications were frequent in early-replicating regions consistent with previous studies on SVs in cancer (De and Michor 2011; Fujimoto et al. 2016b). They reported that large amplifications were frequent in early replicating regions (De and Michor 2011; Fujimoto et al. 2016b) and our study showed a similar pattern in short tandem duplications. These observations might offer insights into the possible mechanisms underlying their occurrences. It is suggested that formation of higher-order DNA structures due to single-stranded DNA and spontaneous DNA breaks during replication in early replicating regions may play a role in the occurrence of such tandem duplications (Zhang et al. 2009; De and Michor 2011).

We compared insertions of the current study with deletions from our previous study (Wong et al. 2019) for features and enrichment of regulatory elements. We observed that the number of insertions located

inside genes, exons, and UTRs is higher than that of deletions. This could indicate that compared to deletions, insertions are more tolerated in genic regions. Although higher number of insertions was observed inside genes, the number of deletions predicted to be causative for gene expression level changes was significantly higher than insertions, possibly indicating the more deleterious role of deletions in the human genome.

There are some limitations to the current study. Using short reads for calling and determining the insertions enabled us to detect only insertions positions, so determining the sequence of functionally interesting insertions for further analysis needs direct sequencings of the surrounding region. In addition, due to ambiguity of insertion calling in repeat regions and in order to obtain a reliable catalogue of insertions, we excluded insertions located inside simple repeats, satellites, micro-satellites, and low complexity regions. This might have removed truly existing insertions. It is expected that with using rapidly evolving long-read sequencing technologies, many more SVs in a wide range of genomic regions will be identified and used in the human reference genome. Additionally, we removed insertions with allele frequency less than 0.05 and those with HWE test p-value < 1E-04. This led to removal of low frequency variants with potentially functional effects in the regulation of gene expression. We expect that using larger sample sets for identifying the insertions will address this issue too.

Conclusions

In the current study, we identified a catalogue of accurate, highly-confident intermediate-sized insertions from whole genome sequencing data of a Japanese sample set. Our results indicate that intermediate-sized insertions can be important gene expression-altering genetic variants with the potential to affect expression of multiple genes. With advances in long-read sequencing technologies, it will be possible to include these SVs along with other genomic variants in typical association studies.

Declarations

Funding

This research was supported by Japan Agency for Medical Research and Development (AMED) under Grant Number 19km0405207h0004 (to AF), 20km04052 (to AF), and 20km0405205 (to KT).

Competing interests

The authors declare that they have no competing interests.

Availability of data and material

The data used in the current study are as follows:

Nanopore MinION long-read sequencing data for sample RK067 is available from the Japanese Genotype-phenotype Archive (JGA) database (<https://ddbj.nig.ac.jp/jga/viewer/view/studies>) under

accession number JGAS00000000180.

Nanopore MinION long-read sequencing data for sample NA18943 is available from the DNA Data Bank of Japan (DDBJ) Sequence Read Archive (DRA) (<https://www.ddbj.nig.ac.jp/dra/index-e.html>) under accession number DRA008482.

The reference panel used for the imputation in the current study is not publicly available due to the risk of compromise of participant privacy but are available from the corresponding author on reasonable request.

Other web resources or tools used in this study are as follows:

IMSindel software: <https://github.com/NCGG-MGC/IMSindel>

PLINK: <https://www.cog-genomics.org/plink2>

IMPUTE2: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

MatrIXeQTL: http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

CAVIAR: <http://genetics.cs.ucla.edu/caviar/>

RepeatMasker software developed by A.F.A. Smit, R. Hubley, and P. Green; see <http://www.repeatmasker.org/>

All statistical tests were conducted in R (ver3.5.0): <https://www.r-project.org/>

GTEx portal database: <http://gtexportal.org/home/>

Gene expression data for 82 Japanese HapMap samples used for the eQTL association analysis were obtained from EMBL-EBI ArrayExpress under accession E-MTAB-264 (Montgomery).

Published GWAS association results are available from NHGRI-EBI Catalogue of published GWAS: <https://www.ebi.ac.uk/gwas/home>

Super-enhancer region data is available from dbSUPER (<http://asntech.org/dbsuper/>)

CTCF binding site data is available from CTCFBSDB2.0 (<http://insulatordb.uthsc.edu/>)

Ensembl regulatory region data was obtained using BioMart for Ensembl Release 93 (<https://grch37.ensembl.org/info/index.html>)

UCSC phastCons genome conservation scores are available from the UCSC genome browser (<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons46way/>)

Code availability

Source code is available from <https://github.com/Saeideh-Ashouri/Intermediate-Sized-Insertions>

Authors' contributions

All authors contributed to the study conception and design. Akihiro Fujimoto designed the study. Saeideh Ashouri performed the computational analyses and the experiments. Saeideh Ashouri, Jing Hao Wong, Hidewaki Nakagawa, Mihoko Shimada, Katsushi Tokunaga, and Akihiro Fujimoto interpreted the results. Saeideh Ashouri, Jing Hao Wong, and Akihiro Fujimoto wrote the first draft of manuscript and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

This study was approved by institutional review boards (IRBs) at RIKEN (H20-11) and Kyoto University (G1031).

Consent to participate

This study was conducted in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo (<http://sc.hgc.jp/shirokane.html>).

References

Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, Hong O, Chapman TR, Tan S, Tandon P, Pang T, Tang SC, Vaux KK, Yang Y, Harrington E, Juul S, Turner DJ, Thiruvahindrapuram B, Kaur G, Wang Z, Kingsmore SF, Gleeson JG, Bisson D, Kakaradov B, Telenti A, Venter JC, Corominas R, Toma C, Cormand B, Rueda I, Guijarro S, Messer KS, Nievergelt CM, Arranz MJ, Courchesne E, Pierce K, Muotri AR, Iakoucheva LM, Hervas A, Scherer SW, Corsello C, Sebat J (2018) Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 360(6386):327–331. <https://doi.org/10.1126/science.aan2261>

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haeussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D, Fiddes IT, Eisenhart C, Diekhans M, Clawson H, Barber GP, Armstrong J, Haussler D, Kuhn RM, Kent WJ (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* 46(D1):D762–D769. <https://doi.org/10.1093/nar/gkx1020>

Chahal HS, Wu W, Ransohoff KJ, Yang L, Hedlin H, Desai M, Lin Y, Dai H-J, Qureshi AA, Li W-Q, Kraft P, Hinds DA, Tang JY, Han J, Sarin KY (2016) Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nat Commun* 7. <https://doi.org/10.1038/ncomms12510>

Chen Y, Ye X, Xia X, Lin X (2019) Circular RNA ABCB10 correlates with advanced clinicopathological features and unfavorable survival, and promotes cell proliferation while reduces cell apoptosis in epithelial ovarian cancer. *Cancer Biomark* 26(2):151–161. <https://doi.org/10.3233/CBM-190064>

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM (2017) The impact of structural variation on human gene expression. *Nat Genet* 49(5):692–699. <https://doi.org/10.1038/ng.3834>

Day FR, Ruth KS, Thompson DJ, Lunetta KL, Pervjakova N, Chasman DI, Stolk L, Finucane HK, Sulem P, Bulik-Sullivan B, Esko T, Johnson AD, Elks CE, Franceschini N, He C, Altmaier E, Brody JA, Franke LL, Huffman JE, Keller MF, McArdle PF, Nutile T, Porcu E, Robino A, Rose LM, Schick UM, Smith JA, Teumer A, Traglia M, Vuckovic D, Yao J, Zhao W, Albrecht E, Amin N, Corre T, Hottenga J-J, Mangino M, Smith AV, Tanaka T, Abecasis G, Andrulis IL, Anton-Culver H, Antoniou AC, Arndt V, Arnold AM, Barbieri C, Beckmann MW, Beeghly-Fadiel A, Benitez J, Bernstein L, Bielinski SJ, Blomqvist C, Boerwinkle E, Bogdanova NV, Bojesen SE, Bolla MK, Borresen-Dale A-L, Boutin TS, Brauch H, Brenner H, Brüning T, Burwinkel B, Campbell A, Campbell H, Chanock SJ, Chapman JR, Chen Y-DI, Chenevix-Trench G, Couch FJ, Coviello AD, Cox A, Czene K, Darabi H, De Vivo I, Demerath EW, Dennis J, Devilee P, Dörk T, dos-Santos-Silva I, Dunning AM, Eicher JD, Fasching PA, Faul JD, Figueroa J, Flesch-Janys D, Gandin I, Garcia ME, García-Closas M, Giles GG, Girotto GG, Goldberg MS, González-Neira A, Goodarzi MO, Grove ML, Gudbjartsson DF, Guénel P, Guo X, Haiman CA, Hall P, Hamann U, Henderson BE, Hocking LJ, Hofman A, Homuth G, Hooning MJ, Hopper JL, Hu FB, Huang J, Humphreys K, Hunter DJ, Jakubowska A, Jones SE, Kabisch M, Karasik D, Knight JA, Kolcic I, Kooperberg C, Kosma V-M, Kriebel J, Kristensen V, Lambrechts D, Langenberg C, Li J, Li X, Lindström S, Liu Y, Luan J, Lubinski J, Mägi R, Mannermaa A, Manz J, Margolin S, Marten J, Martin NG, Masciullo C, Meindl A, Michailidou K, Mihailov E, Milani L, Milne RL, Müller-Nurasyid M, Nalls M, Neale BM, Nevanlinna H, Neven P, Newman AB, Nordestgaard BG, Olson JE, Padmanabhan S, Peterlongo P, Peters U, Petersmann A, Peto J, Pharoah PDP, Pirastu NN, Pirie A, Pistis G, Polasek O, Porteous D, Psaty BM, Pylkäs K, Radice P, Raffel LJ, Rivadeneira F, Rudan I, Rudolph A, Ruggiero D, Sala CF, Sanna S, Sawyer EJ, Schlessinger D, Schmidt MK, Schmidt F, Schmutzler RK, Schoemaker MJ, Scott RA, Seynaeve CM, Simard J, Sorice R, Southey MC, Stöckl D, Strauch K, Swerdlow A, Taylor KD, Thorsteinsdottir U, Toland AE, Tomlinson I, Truong T, Tryggvadottir L, Turner ST, Vozzi D, Wang Q, Wellons M, Willemsen G, Wilson JF, Winqvist R, Wolffenbuttel BBHR, Wright AF, Yannoukakos D, Zemunik T, Zheng W, Zygmont M, Bergmann S, Boomsma DI, Buring JE, Ferrucci L, Montgomery GW, Gudnason V, Spector TD, van Duijn

CM, Alizadeh BZ, Ciullo M, Crisponi L, Easton DF, Gasparini PP, Gieger C, Harris TB, Hayward C, Kardia SLR, Kraft P, McKnight B, Metspalu A, Morrison AC, Reiner AP, Ridker PM, Rotter JI, Toniolo D, Uitterlinden AG, Ulivi S, Völzke H, Wareham NJ, Weir DR, Yerges-Armstrong LM, Price AL, Stefansson K, Visser JA, Ong KK, Chang-Claude J, Murabito JM, Perry JRB, Murray A (2015) Large-scale genomic analyses link reproductive ageing to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet* 47(11):1294–1303. <https://doi.org/10.1038/ng.3412>

De S, Michor F (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol* 29(12):1103–1108. <https://doi.org/10.1038/nbt.2030>

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakraborty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads

(data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor U of G Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), Duke University E University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), HudsonAlpha Institute C UC Irvine, Stanford group (data production and analysis) (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74. <https://doi.org/10.1038/nature11247>

Fu Y, Cai L, Lei X, Wang D (2019) Circular RNA ABCB10 promotes hepatocellular carcinoma progression by increasing HMG20A expression by sponging miR-670-3p. *Cancer Cell Int* 19(1):338. <https://doi.org/10.1186/s12935-019-1055-z>

Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, Gotoh K, Ariizumi S, Wardell CP, Hayami S, Nakamura T, Aikata H, Arihiro K, Boroevich KA, Abe T, Nakano K, Maejima K, Sasaki-Oku A, Ohsawa A, Shibuya T, Nakamura H, Hama N, Hosoda F, Arai Y, Ohashi S, Urushidate T, Nagae G, Yamamoto S, Ueda H, Tatsuno K, Ojima H, Hiraoka N, Okusaka T, Kubo M, Marubashi S, Yamada T, Hirano S, Yamamoto M, Ohdan H, Shimada K, Ishikawa O, Yamaue H, Chayama K, Miyano S, Aburatani H, Shibata T, Nakagawa H (2016a) Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* 48(5):500–509. <https://doi.org/10.1038/ng.3547>

Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, Gotoh K, Ariizumi S, Wardell CP, Hayami S, Nakamura T, Aikata H, Arihiro K, Boroevich KA, Abe T, Nakano K, Maejima K, Sasaki-Oku A, Ohsawa A, Shibuya T, Nakamura H, Hama N, Hosoda F, Arai Y, Ohashi S, Urushidate T, Nagae G, Yamamoto S, Ueda H, Tatsuno K, Ojima H, Hiraoka N, Okusaka T, Kubo M, Marubashi S, Yamada T, Hirano S, Yamamoto M, Ohdan H, Shimada K, Ishikawa O, Yamaue H, Chayama K, Miyano S, Aburatani H, Shibata T, Nakagawa H (2016b) Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* 48(5):500–509. <https://doi.org/10.1038/ng.3547>

Fujimoto A, Fujita M, Hasegawa T, Wong JH, Maejima K, Oku-Sasaki A, Nakano K, Shiraishi Y, Miyano S, Yamamoto G, Akagi K, Imoto S, Nakagawa H (2020) Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res* 30(3):334–346. <https://doi.org/10.1101/gr.255026.119>

Guan P, Sung W-K (2016) Structural variation detection using next-generation sequencing data A comparative technical review. *Methods* 102:36–49. <https://doi.org/10.1016/j.ymeth.2016.01.020>

- Hancks DC, Kazazian HH (2016) Roles for retrotransposon insertions in human disease. *Mob DNA* 7(1):9. <https://doi.org/10.1186/s13100-016-0065-9>
- Hendig D, Langmann T, Zarbock R, Schmitz G, Kleesiek K, Götting C (2009) Characterization of the ATP-binding cassette transporter gene expression profile in Y79: a retinoblastoma cell line. *Mol Cell Biochem* 328(1):85–92. <https://doi.org/10.1007/s11010-009-0077-6>
- Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014) Identifying Causal Variants at Loci with Multiple Signals of Association. :21
- Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* 5(6):e1000529. <https://doi.org/10.1371/journal.pgen.1000529>
- Huang Y, Zhang Y, Jia L, Liu C, Xu F (2019) Circular RNA ABCB10 promotes tumor progression and correlates with pejorative prognosis in clear cell renal cell carcinoma. *Int J Biol Markers* 34(2):176–183. <https://doi.org/10.1177/1724600819842279>
- Hubley R, Smit A, Green P (2013) RepeatMasker Open-4.0.
- Kent WJ (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res* 12(4):656–664. <https://doi.org/10.1101/gr.229202>
- Liang H-F, Zhang X-Z, Liu B-G, Jia G-T, Li W-L (2017) Circular RNA circ-ABCB10 promotes breast cancer proliferation and progression through sponging miR-1271. *Am J Cancer Res* 7(7):1566–1576
- Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struwing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6):580–585. <https://doi.org/10.1038/ng.2653>

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 45(D1):D896–D901. <https://doi.org/10.1093/nar/gkw1133>

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. *Genome Biol* 20(1):246. <https://doi.org/10.1186/s13059-019-1828-7>

Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, Wang Q, Dicks E, Lee A, Turnbull C, Rahman N, Fletcher O, Peto J, Gibson L, Silva I dos S, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Czene K, Irwanto A, Liu J, Waisfisz Q, Meijers-Heijboer H, Adank M, van der Luijt RB, Hein R, Dahmen N, Beckman L, Meindl A, Schmutzler RK, Müller-Myhsok B, Lichtner P, Hopper JL, Southey MC, Makalic E, Schmidt DF, Uitterlinden AG, Hofman A, Hunter DJ, Chanock SJ, Vincent D, Bacot F, Tessier DC, Canisius S, Wessels LFA, Haiman CA, Shah M, Luben R, Brown J, Luccarini C, Schoof N, Humphreys K, Li J, Nordestgaard BG, Nielsen SF, Flyger H, Couch FJ, Wang X, Vachon C, Stevens KN, Lambrechts D, Moisse M, Paridaens R, Christiaens M-R, Rudolph A, Nickels S, Flesch-Janys D, Johnson N, Aitken Z, Aaltonen K, Heikkinen T, Broeks A, Van't Veer LJ, van der Schoot CE, Guénel P, Truong T, Laurent-Puig P, Menegaux F, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Zamora MP, Perez JIA, Pita G, Alonso MR, Cox A, Brock IW, Cross SS, Reed MWR, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Henderson BE, Schumacher F, Le Marchand L, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Lindblom A, Margolin S, Hoening MJ, Hollestelle A, van den Ouweland AMW, Jager A, Bui QM, Stone J, Dite GS, Apicella C, Tsimiklis H, Giles GG, Severi G, Baglietto L, Fasching PA, Haeberle L, Ekici AB, Beckmann MW, Brenner H, Müller H, Arndt V, Stegmaier C, Swerdlow A, Ashworth A, Orr N, Jones M, Figueroa J, Lissowska J, Brinton L, Goldberg MS, Labrèche F, Dumont M, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Brauch H, Hamann U, Brüning T, Radice P, Peterlongo P, Manoukian S, Bonanni B, Devilee P, Tollenaar RAEM, Seynaeve C, van Asperen CJ, Jakubowska A, Lubinski J, Jaworska K, Durda K, Mannermaa A, Kataja V, Kosma V-M, Hartikainen JM, Bogdanova NV, Antonenkova NN, Dörk T, Kristensen VN, Anton-Culver H, Slager S, Toland AE, Edge S, Fostira F, Kang D, Yoo K-Y, Noh D-Y, Matsuo K, Ito H, Iwata H, Sueta A, Wu AH, Tseng C-C, Van Den Berg D, Stram DO, Shu X-O, Lu W, Gao Y-T, Cai H, Teo SH, Yip CH, Phuah SY, Cornes BK, Hartman M, Miao H, Lim WY, Sng J-H, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsarn P, Shen C-Y, Hsiung C-N, Wu P-E, Ding S-L, Sangrajrang S, Gaborieau V, Brennan P, McKay J, Blot WJ, Signorello LB, Cai Q, Zheng W, Deming-Halverson S, Shrubsole M, Long J, Simard J, Garcia-Closas M, Pharoah PDP, Chenevix-Trench G, Dunning AM, Benitez J, Easton DF (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 45(4):353-361e2. <https://doi.org/10.1038/ng.2563>

Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Maranian MJ, Bolla MK, Wang Q, Shah M, Perkins BJ, Czene K, Eriksson M, Darabi H, Brand JS, Bojesen SE, Nordestgaard BG, Flyger H, Nielsen SF, Rahman N, Turnbull C, BOCS, Fletcher O, Peto J, Gibson L, dos-Santos-Silva I, Chang-Claude J, Flesch-Janys D, Rudolph A, Eilber U, Behrens S, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Khan S, Aaltonen K, Ahsan H, Kibriya MG, Whittemore AS, John EM, Malone KE, Gammon MD, Santella

RM, Ursin G, Makalic E, Schmidt DF, Casey G, Hunter DJ, Gapstur SM, Gaudet MM, Diver WR, Haiman CA, Schumacher F, Henderson BE, Le Marchand L, Berg CD, Chanock SJ, Figueroa J, Hoover RN, Lambrechts D, Neven P, Wildiers H, van Limbergen E, Schmidt MK, Broeks A, Verhoef S, Cornelissen S, Couch FJ, Olson JE, Hallberg E, Vachon C, Waisfisz Q, Meijers-Heijboer H, Adank MA, van der Luijt RB, Li J, Liu J, Humphreys K, Kang D, Choi J-Y, Park SK, Yoo K-Y, Matsuo K, Ito H, Iwata H, Tajima K, Guénel P, Truong T, Mulot C, Sanchez M, Burwinkel B, Marme F, Surowy H, Sohn C, Wu AH, Tseng C, Van Den Berg D, Stram DO, González-Neira A, Benitez J, Zamora MP, Perez JIA, Shu X-O, Lu W, Gao Y-T, Cai H, Cox A, Cross SS, Reed MWR, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, kConFab Investigators, AOCs Group, Lindblom A, Margolin S, Teo SH, Yip CH, Taib NAM, Tan G-H, Hooning MJ, Hollestelle A, Martens JWM, Collée JM, Blot W, Signorello LB, Cai Q, Hopper JL, Southey MC, Tsimiklis H, Apicella C, Shen C-Y, Hsiung C-N, Wu P-E, Hou M-F, Kristensen VN, Nord S, Alnaes GIG, NBCS, Giles GG, Milne RL, McLean C, Canzian F, Trichopoulos D, Peeters P, Lund E, Sund M, Khaw K-T, Gunter MJ, Palli D, Mortensen LM, Dossus L, Huerta J-M, Meindl A, Schmutzler RK, Sutter C, Yang R, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsarn P, Hartman M, Miao H, Chia KS, Chan CW, Fasching PA, Hein A, Beckmann MW, Haeberle L, Brenner H, Dieffenbach AK, Arndt V, Stegmaier C, Ashworth A, Orr N, Schoemaker MJ, Swerdlow AJ, Brinton L, Garcia-Closas M, Zheng W, Halverson SL, Shrubsole M, Long J, Goldberg MS, Labrèche F, Dumont M, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Brauch H, Hamann U, Brüning T, GENICA Network, Radice P, Peterlongo P, Manoukian S, Bernard L, Bogdanova NV, Dörk T, Mannermaa A, Kataja V, Kosma V-M, Hartikainen JM, Devilee P, Tollenaar RAEM, Seynaeve C, Van Asperen CJ, Jakubowska A, Lubinski J, Jaworska K, Huzarski T, Sangrairang S, Gaborieau V, Brennan P, McKay J, Slager S, Toland AE, Ambrosone CB, Yannoukakos D, Kabisch M, Torres D, Neuhausen SL, Anton-Culver H, Luccarini C, Baynes C, Ahmed S, Healey CS, Tessier DC, Vincent D, Bacot F, Pita G, Alonso MR, Álvarez N, Herrero D, Simard J, Pharoah PPDP, Kraft P, Dunning AM, Chenevix-Trench G, Hall P, Easton DF (2015) Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* 47(4):373–380. <https://doi.org/10.1038/ng.3242>

Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, Lemaçon A, Soucy P, Glubb D, Rostamianfar A, Bolla MK, Wang Q, Tyrer J, Dicks E, Lee A, Wang Z, Allen J, Keeman R, Eilber U, French JD, Chen XQ, Fachal L, McCue K, McCart Reed AE, Ghoussaini M, Carroll J, Jiang X, Finucane H, Adams M, Adank MA, Ahsan H, Aittomäki K, Anton-Culver H, Antonenkova NN, Arndt V, Aronson KJ, Arun B, Auer PL, Bacot F, Barrdahl M, Baynes C, Beckmann MW, Behrens S, Benitez J, Bermisheva M, Bernstein L, Blomqvist C, Bogdanova NV, Bojesen SE, Bonanni B, Børresen-Dale A-L, Brand JS, Brauch H, Brennan P, Brenner H, Brinton L, Broberg P, Brock IW, Broeks A, Brooks-Wilson A, Brucker SY, Brüning T, Burwinkel B, Butterbach K, Cai Q, Cai H, Caldés T, Canzian F, Carracedo A, Carter BD, Castela JE, Chan TL, Cheng T-YD, Chia KS, Choi J-Y, Christiansen H, Clarke CL, Collée M, Conroy DM, Cordina-Duverger E, Cornelissen S, Cox DG, Cox A, Cross SS, Cunningham JM, Czene K, Daly MB, Devilee P, Doherty KF, Dörk T, dos-Santos-Silva I, Dumont M, Durcan L, Dwek M, Eccles DM, Ekici AB, Eliassen AH, Ellberg C, Elvira M, Engel C, Eriksson M, Fasching PA, Figueroa J, Flesch-Janys D, Fletcher O, Flyger H, Fritschi L, Gaborieau V, Gabrielson M, Gago-Dominguez M, Gao Y-T, Gapstur SM, García-Sáenz JA, Gaudet MM, Georgoulas V, Giles GG, Glendon G, Goldberg MS, Goldgar DE, González-Neira A, Grenaker Alnæs GI, Grip M, Gronwald J, Grundy A, Guénel P,

Haeberle L, Hahnen E, Haiman CA, Håkansson N, Hamann U, Hamel N, Hankinson S, Harrington P, Hart SN, Hartikainen JM, Hartman M, Hein A, Heyworth J, Hicks B, Hillemanns P, Ho DN, Hollestelle A, Hoening MJ, Hoover RN, Hopper JL, Hou M-F, Hsiung C-N, Huang G, Humphreys K, Ishiguro J, Ito H, Iwasaki M, Iwata H, Jakubowska A, Janni W, John EM, Johnson N, Jones K, Jones M, Jukkola-Vuorinen A, Kaaks R, Kabisch M, Kaczmarek K, Kang D, Kasuga Y, Kerin MJ, Khan S, Khusnutdinova E, Kiiski JI, Kim S-W, Knight JA, Kosma V-M, Kristensen VN, Krüger U, Kwong A, Lambrechts D, Marchand LL, Lee E, Lee MH, Lee JW, Lee CN, Lejbkiewicz F, Li J, Lilyquist J, Lindblom A, Lissowska J, Lo W-Y, Loibl S, Long J, Lophatananon A, Lubinski J, Luccarini C, Lux MP, Ma ESK, MacInnis RJ, Maishman T, Makalic E, Malone KE, Kostovska IM, Mannermaa A, Manoukian S, Manson JE, Margolin S, Mariapun S, Martinez ME, Matsuo K, Mavroudis D, McKay J, McLean C, Meijers-Heijboer H, Meindl A, Menéndez P, Menon U, Meyer J, Miao H, Miller N, Mohd Taib NA, Muir K, Mulligan AM, Mulot C, Neuhausen SL, Nevanlinna H, Neven P, Nielsen SF, Noh D-Y, Nordestgaard BG, Norman A, Olopade OI, Olson JE, Olsson H, Olswold C, Orr N, Pankratz VS, Park SK, Park-Simon T-W, Lloyd R, Perez JIA, Peterlongo P, Peto J, Phillips K-A, Pinchev M, Plaseska-Karanfilska D, Prentice R, Presneau N, Prokofieva D, Pugh E, Pylkäs K, Rack B, Radice P, Rahman N, Rennert G, Rennert HS, Rhenius V, Romero A, Romm J, Ruddy KJ, Rüdiger T, Rudolph A, Ruebner M, Rutgers EJTh, Saloustros E, Sandler DP, Sangrajrang S, Sawyer EJ, Schmidt DF, Schmutzler RK, Schneeweiss A, Schoemaker MJ, Schumacher F, Schürmann P, Scott RJ, Scott C, Seal S, Seynaeve C, Shah M, Sharma P, Shen C-Y, Sheng G, Sherman ME, Shrubsole MJ, Shu X-O, Smeets A, Sohn C, Southey MC, Spinelli JJ, Stegmaier C, Stewart-Brown S, Stone J, Stram DO, Surowy H, Swerdlow A, Tamimi R, Taylor JA, Tengström M, Teo SH, Terry MB, Tessier DC, Thanasiythichai S, Thöne K, Tollenaar RAEM, Tomlinson I, Tong L, Torres D, Truong T, Tseng C, Tsugane S, Ulmer H-U, Ursin G, Untch M, Vachon C, van Asperen CJ, Van Den Berg D, van den Ouweland AMW, van der Kolk L, van der Luijt RB, Vincent D, Vollenweider J, Waisfisz Q, Wang-Gohrke S, Weinberg CR, Wendt C, Whittemore AS, Wildiers H, Willett W, Winqvist R, Wolk A, Wu AH, Xia L, Yamaji T, Yang XR, Yip CH, Yoo K-Y, Yu J-C, Zheng W, Zheng Y, Zhu B, Ziogas A, Ziv E, Lakhani SR, Antoniou AC, Droit A, Andrulis IL, Amos CI, Couch FJ, Pharoah PDP, Chang-Claude J, Hall P, Hunter DJ, Milne RL, García-Closas M, Schmidt MK, Chanock SJ, Dunning AM, Edwards SL, Bader GD, Chenevix-Trench G, Simard J, Kraft P, Easton DF (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature* 551(7678):92–94. <https://doi.org/10.1038/nature24284>

Montgomery SB E-MTAB-264 - HAPMAP3_EXPRESSION. ArrayExpress, E-MTAB-264. <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264>

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, Consortium T 1000 GP, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G (2013) The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res* 23(5):749–761. <https://doi.org/10.1101/gr.148718.112>

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, Kirkness EF, Levy S, Feuk L, Scherer SW (2010) Towards a comprehensive structural variation map of an individual human genome. :14

Psychosis Endophenotypes International Consortium, CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium, Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, Gujral M, Brandler WM, Malhotra D, Wang Z, Fajardo KVF, Maile MS, Ripke S, Agartz I, Albus M, Alexander M, Amin F, Atkins J, Bacanu SA, Belliveau RA, Bergen SE, Bertalan M, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Bulik-Sullivan B, Byerley W, Cahn W, Cai G, Cairns MJ, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Cheng W, Cloninger CR, Cohen D, Cormican P, Craddock N, Crespo-Facorro B, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, DeLisi LE, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farh K-H, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedman JI, Forstner AJ, Fromer M, Genovese G, Georgieva L, Gershon ES, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Gratten J, de Haan L, Hamshire ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Huang H, Ikeda M, Joa I, Kähler AK, Kahn RS, Kalaydjieva L, Karjalainen J, Kavanagh D, Keller MC, Kelly BJ, Kennedy JL, Kim Y, Knowles JA, Konte B, Laurent C, Lee P, Lee SH, Legge SE, Lerer B, Levy DL, Liang K-Y, Lieberman J, Lönnqvist J, Loughland CM, Magnusson PKE, Maher BS, Maier W, Mallet J, Mattheisen M, Mattingdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melle I, Meshulam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Müller-Myhsok B, Murphy KC, Murray RM, Myin-Germeys I, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, Oh S-Y, Olincy A, Olsen L, O'Neill FA, Van Os J, Pantelis C, Papadimitriou GN, Parkhomenko E, Pato MT, Paunio T, Perkins DO, Pers TH, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Savitz A, Schall U, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Silverman JM, Smoller JW, Söderman E, Spencer CCA, Stahl EA, Strengman E, Strohmaier J, Stroup TS, Suvisaari J, Svrakic DM, Szatkiewicz JP, Thirumalai S, Tooney PA, Veijola J, Visscher PM, Waddington J, Walsh D, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wormley BK, Wray NR, Wu JQ, Zai CC, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Cichon S, Collier DA, Corvin A, Daly MJ, Darvasi A, Domenici E, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jönsson EG, Kendler KS, Kirov G, Knight J, Levinson DF, Li QS, McCarroll SA, McQuillin A, Moran JL, Mowry BJ, Nöthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sklar P, St Clair D, Walters JTR, Werge T, Sullivan PF, O'Donovan MC, Scherer SW, Neale BM, Sebat J (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 49(1):27–35. <https://doi.org/10.1038/ng.3725>

Shabalín AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353–1358. <https://doi.org/10.1093/bioinformatics/bts163>

Shigemizu D, Miya F, Akiyama S, Okuda S, Boroevich KA, Fujimoto A, Nakagawa H, Ozaki K, Niida S, Kanemura Y, Okamoto N, Saitoh S, Kato M, Yamasaki M, Matsunaga T, Mutai H, Kosaki K, Tsunoda T (2018) IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and

gapped global-local alignment with split read analysis. *Sci Rep* 8(1):5608.

<https://doi.org/10.1038/s41598-018-23978-z>

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.

<https://doi.org/10.1101/gr.3715005>

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HYK, Lee W-P, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT, Project 1000 Genomes (2011) A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLOS Genet* 7(8):e1002236. <https://doi.org/10.1371/journal.pgen.1002236>

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET (2012) Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLOS Genet* 8(4):e1002639.

<https://doi.org/10.1371/journal.pgen.1002639>

Tang W, Mun S, Joshi A, Han K, Liang P (2018) Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Res* 25(5):521–533.

<https://doi.org/10.1093/dnares/dsy022>

Taylor MS (2004) Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Res* 14(4):555–566. <https://doi.org/10.1101/gr.1977804>

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74. <https://doi.org/10.1038/nature15393>

The 1000 Genomes Project Consortium, Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lammeijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalina AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571):75–81.

<https://doi.org/10.1038/nature15394>

The International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455(7210):237–241. <https://doi.org/10.1038/nature07239>

The Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464(7289):713–720.

<https://doi.org/10.1038/nature08979>

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164–e164. <https://doi.org/10.1093/nar/gkq603>

Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 14(2):125–138.

<https://doi.org/10.1038/nrg3373>

Wong JH, Shigemizu D, Yoshii Y, Akiyama S, Tanaka A, Nakagawa H, Narumiya S, Fujimoto A (2019) Identification of intermediate-sized deletions and inference of their impact on gene expression in a human population. *Genome Med* 11(1):44. <https://doi.org/10.1186/s13073-019-0656-4>

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 10:451–481.

<https://doi.org/10.1146/annurev.genom.9.081307.164217>

Figures

Figure 1

Workflow of the current study. This figure displays the workflow of the current study. WGS data of 174 Japanese samples were used to call intermediate-sized insertions using IMSindel software. After filtering and processing, the accurate list of insertions was combined with SNVs from the same sample set to construct an imputation panel. Then, insertions were imputed into a separate Japanese sample set. A

subsequent eQTL association analysis identified insertions that were possibly causative for gene expression level changes

Figure 2

Features of final insertion candidates. From 4,254 insertion candidates, 1,956 (45%) of them were located inside genes and others (55%) were found to be located in intergenic regions. From 1,956 insertion located inside genes, only 78 (~4%) were detected to be located inside coding regions (exons) and others (96%) were found to be in non-coding regions

Figure 3

Effectiveness of progressive filtering procedure on the list of intermediate-sized insertions. Estimation of true positive rate of insertion detection by comparing the list of detected insertions with Nanopore long-read sequencing data. Before filtering, the consistency rate was 45%, with annotation filtering increasing the true positive rate to 82% and joint-call recovery and population filters increasing it to the final accuracy rate of 98%

Figure 4

Enrichment of intermediate-sized insertions for annotated regulatory features. a Odds ratios (ORs) for regulatory feature significantly enriched between causal insertion candidates and other insertions in overall set. Ensembl promoters showed the largest positive effect (OR = 4.92) and heterochromatin showed a negative enrichment (OR = 0.39). b ORs for significantly enriched regulatory feature specific to GM12878 cell line between causal insertion candidates and other insertions. From predicted chromatin states, active promoter showed the largest positive effect (OR = 3.38) and heterochromatin showed a negative enrichment (0.42)

Figure 5

Observed proportion of short tandem duplications and other insertions by the replication timing of the regions containing the insertion positions. Replication timing is shown in four groups corresponding to the earliest to latest from left to right. Short tandem duplications tend to occur frequently in early-replicating regions

Figure 6

Comparison of causal intermediate-sized insertions with causal intermediate-sized deletions for features and enrichment in genome regulatory elements. Significant differences were observed for the number of causal variants (OR = 0.72) as well as the number of insertions and deletion located inside genes (OR = 1.31), exons (OR = 1.89), and UTRs (OR = 2.9). There was also significant enrichment of insertions for CTCF-binding sites (OR = 0.29). **: p-value < 0.01. *: p-value < 0.05. N.S: Not significant

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.pdf](#)
- [Additionalfile2.xlsx](#)