creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사학위논문

한국인 유전체 변이 표준 데이터베이스 (KOVA) 구축

Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population

2018년 8월

서울대학교 대학원

생명과학부

박 진 만

ABSTRACT

Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population

Jinman Park
School of Biological Sciences
The Graduate School
Seoul National University

Previous efforts in identifying and understanding human genome variations have been systematically preferred towards the populations of Caucasian descendants and have resulted in loss of accuracy and detection power when analyzing genome variations in non-Caucasian populations. To improve accuracy and precision in studying genome variations in the Korean population, we have compiled and curated a large-scale database of coding variants from 1,055 healthy Korean samples. These samples were sequenced to an average depth of 75x with 101 singleton variants per individual. Through analysis of our healthy Korean cohort, we found distinct ethnic group characteristics comparable to that of other ethnic groups in Africa and Europe,

indicating that the Korean population are a discrete ethnic group that can benefit from an independent reference database. Through population genetics analyses, we found that our Korean Variant Archive (KOVA) increased variant filtering power when examining Korean exomes in concert with Exome Aggregation Consortium. Interestingly, we provide a cohort of potential germline variants that may be associated with susceptibility to tumorigenesis. These rare germline variants were validated using independent datasets from The Cancer Genome Atlas and 1000 Genome Project. To our knowledge, KOVA represents the first database curated using high quality whole exome sequencing-based variants specifically from Korean individuals that can serve as a valuable resource for future studies in the study of human genome variations in the Korean population.

Keywords: Coding variants, population genetics, tumor susceptibility, whole exome sequencing

Student ID: 2014-25015

Disclosure: The materials of this thesis paper are from previously published work by the author as a co-first auther entitled Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. Sci. Reports. 7, 4287 (2017).

CONTENTS

INTRODUCTION

Through the advancements and affordability of next-generation sequencing, there has been an explosive expansion of sequencing efforts and curation of databases in various fields of genetics. These include disease-based databases such as The Cancer Genome Atlas (TCGA), which provides an invaluable resource for the study of genomic variations in cancers. Although such large-scale databases have presented an unprecedented opportunity to study disease-specific variants in the human genome that may be associated with disease susceptibility or progression, an over-accumulation of variants in the human genome that do not elicit any functional consequences, also known as germline or passenger variants, have presented a difficult challenge for population and disease genetics studies. To address this challenge, massive-scale databases that have cataloged enormous variant information from healthy populations have been established such as the 1000 Genome Project (1000GP), Exome Aggregation Consortium (ExAC), UK10K project, and the 1KJPN database (1-5). These databases have provided researchers with in-depth genomic profiles of healthy individuals that have greatly aided in providing insights into population genetics and evolutionary studies. However, the databases have been curated largely using the genomic information from individuals of Caucasians populations. Although East Asian populations are included, they represent mainly samples from the

Japanese or Chinese populations and do not cover the vast diversity present in the East Asian region of the world. More importantly, there is a distinct lack of a reference database cataloging high-quality variants specifically tailored to the Korean population. The absence of a Korean-specific database has led to unintended loss of accuracy and detection power for population genetics or disease-based studies focused on the Korean population. The curation of such a database is expected to serve as a valuable resource for future studies that focus on population and disease genetics as well as evolutionary surveys into the Korean population.

As a collaborative effort from multiple research groups, we obtained and performed whole-exome sequencing (WES) of tumor-paired normal samples and samples from healthy individuals with no previously reported clinical conditions. The samples were sequenced to an average depth of 75x and processed using an in-house pipeline with extensive filtering steps to yield 101 singletons per sample and a total of 293,049 high-quality exonic variants. We performed population genetics analyses to compare our cohort with those from the previously reported large-scale population studies such as the 1000GP, ExAC, the UK10K project, and the 1KJPN database from Japan (1-5). We found that the Korean population is a distinct ethnic group from not only the European and African population but also from the East Asian populations such as the diverse ethnic groups in China and Japan. The findings demonstrated the need for a tailored population-specific reference database that can improve

accuracy and power for a diverse array of studies into human genome variations from Korean individuals. In addition, we assessed the potential biological consequence of the coding variants in KOVA through functional annotation and evaluation of potential tumor susceptibility. In doing so, we found a group of rare functional variants that may potentially be associated with cancer susceptibility. These variants were validated using independent databases.

To our knowledge, our collaborative effort represents the first WES-based variant database specifically curated using only Korean individuals. Through population genetics analyses, we found that the Korean population is a distinct ethnic group, clearly distinguishable from the African, European, or other Southeast Asian populations. With our high-quality catalogue of Korean-specific variants, we performed various functional assessment and population genetics analyses revealing insights into the Korean population. Interestingly, our cohort consisted of both tumor-paired samples and samples from healthy individuals with no previous clinical record of disease. The diversity of our database presented an opportunity for closely examining any potential predisposition to cancers elicited by rare germline variants. In doing so, we were able to determine rare germline variant candidates that are potentially associated with tumor susceptibility. Our catalog of Korean variants can serve as a valuable resource for the purpose of disease-based and population genetics studies of Korean individuals and provide a critical addition to the pre-existing databases for improving future genetics studies.

RESULTS

Sample preparation, variant calling and quality control

Whole exome sequencing was conducted on a total of 676 normal tissue samples obtained from cancer patients in which 472 samples were obtained from blood and 204 samples were collected from the adjacent normal tissues. Likewise, 380 blood samples were obtained from healthy individuals without previous clinical conditions (Table 1). Extensive examination and quality-control of the samples were conducted to ensure high quality of the dataset by removing duplicated samples and potential relatives among samples. In doing so, a final cohort of 1,055 samples was curated and established for further analysis. From the curated sample cohort, variant calling was conducted using an in-house pipeline utilizing publicly available tools.

To accurately identify high-quality germline variants from our large-scale sample cohort, a highly optimized and efficient variant calling pipeline was established. Widely validated and publicly available tools such as VCFtools and the Genome Analysis Toolkit (GATK) package was utilized in conjunction with in-house python scripts to develop and optimize the KOVA analysis pipeline (11,12). To control for potential noise introduced by using varying exome capture kits across the research groups, the large sample data collected for the KOVA were all processed simultaneously in a

massively-parallel process that provided high accuracy and time-efficiency (Figure 1). Variants were called in a multi-sample mode and through rigorous filtering, we obtained a final total of 293,199 high-quality variants to establish the KOVA database consisting of only Korean individuals. Overall, the depth coverage of the variants was 41x on average and the minimun genotype quality was determined to be 66, indicating high variant calling quality of our cohort (Figure 2). Furthermore, we examined the transition-to-transversion and hetero-to-homozygosity ratios of the variants and found that the ratios were consistent with previous reports (6).

Careful considerations were taken to exclude any potential systematic bias that may have been introduced by combining samples from five independent research groups using varying capture kits and next-generation sequencing methods. The statistics and variant profiles from each research group was examined. We found on average 42,000 variants and 101 singletons per sample. The samples that deviated excessively according to their variant profiles were excluded from further analysis to ensure that the cohort exhibited stable variant and singleton profiles across all samples from the different research groups.

KOVA population genetics and basic statistics

To provide KOVA database as a critical addition to the existing databases, we compared the variants specific to the Korean

population with the existing database. Comparison with the dbSNP database (build 142) identified that 59.0% variants were previously reported and 41.0% of the variants were novel. When examining the minor allele frequencies (MAF) of the variants, we found that the distribution of rare variants with <0.01 MAF exhibited higher enrichment with novel variants. Consistently, more common variants with >0.05 MAF demonstrated higher enrichment with previously reported variants on the dbSNP database (Figure 3). Our findings demonstrate that KOVA database can provide additional power to future analyses by providing a catalogue of our high-quality variants.

We evaluated the KOVA database with other large-scale WGS or WES-based studies to better understand how the variants profiles differed in the Korean population when compared to Caucasian, African, or other East Asian populations. We compared the KOVA database against the 1000GP, ExAC, the UK10K project, and the 1KJPN database (1-5) and found that only ~37% of the variants catalogued in KOVA were also catalogued in the other populations, indicating the clear distinction of the Korean population among other ethnic groups (Figure 3). As expected, conducting a principal component analysis (PCA) demonstrated that KOVA variants were more closely related to other East Asian populations when compared to European or African populations. However, the Korean population was also an independently distinct population. Interestingly, we found that the PCA clusters were not unlike the geographical distribution of the populations.

Functional analysis of coding variants

Using the 64,469 coding variants in KOVA, we conducted a functional survey of the potential biological impact of the KOVA variants (Table 2). Accordingly, we found that the nonsynonymous-to-synonymous SNVs (N/S) ratio was relatively high in rare variants (Figure 2). As nonsynonymous variants, or those affecting the amino acid sequence of the resultant protein product, can be more detrimental to the cell than synonymous variants, we found that these rarer variants were predicted to be as such according to the lower SIFT scores and higher scaled C-scores (CADD) (Figure 4) (7-10). These findings indicated that rarer variants tended to exhibit higher potential for functional implications. On the other hand, common variants exhibited prediction patterns that indicated minimal functional implications, consistent with the notion that these common variants escaped the functional restrictions opposed by purifying selection. Accordingly, we compared the KOVA variants with those from the ClinVar database which provides a large catalogue of variants from clinical studies implicated in disease progression and predisposition. Despite the samples used for the curation the KOVA database being healthy, we found that certain overlapping variants deemed pathogenic by the ClinVar database in the KOVA database. This suggested that there may be a potential mislabeling of pathogenic variants in the pre-existing databases and indicated that the KOVA database can further serve as an resourceful addition to the current collection of large-scale variant archives.

Potential role of rare germline variants on tumor susceptibility

Although all of the samples used in the curation of the KOVA database were derived from healthy tissues or blood samples, the samples in the KOVA database consisted of tumor-paired normal samples derived from healthy adjacent tissues or blood samples of cancer patients and healthy normal samples from individuals without any previously reported clinical conditions. We hypothesized that by comparing the germline variants from the tumor-paired normal to healthy normal samples, certain variant candidates exhibiting enrichment in the tumor-paired normal samples could be identified as potential variants associated with tumor susceptibility. The high-quality variants in KOVA provided an opportunity to explore the potential role of the rare germline variants within the context of cancer development.

From the final archive of called variants, the variants were identified and separated according to the source of the samples, consisting of 364 lung and 77 stomach samples for the tumor-paired normal samples and 134 samples for the healthy normal samples. Selection criteria for the variant frequency was applied to identify variant candidates that were enriched in the tumor-paired normal samples. The arbitrary criteria selected for candidate variants that exhibited a MAF value of at least 0.01 with a minimum 1.5-fold larger MAF detected in the tumor-paired normal samples than the healthy normal samples. In doing so, we found 54 and 72 rare

variants that exhibited higher prevalence in tumor-paired samples for lung and stomach cancers, respectively.

To validate our findings using independent datasets, we obtained variant data available on the TCGA and 1000GP databases. From the TCGA, we obtained cancer-derived variants from the corresponding cancers types, lung and stomach adenocarcinomas. From the 1000GP database, we obtained two independent sets of variants from healthy normal samples obtained from the European and East Asian populations. The ethnicity of sample was taken into consideration to reduce potential loss of power caused by combining variant data from various sample populations. We identified the potential variant candidates associated with cancer predisposition from KOVA within these independent datasets and compared the MAFs accordingly. In doing so, we observed a consistent behavior and statistically significant difference in the allele frequency of the variant candidates observed in the tumor-paired samples from TCGA against that of healthy samples from the 1000GP datasets (P = 0.018 and P = 0.003 for lung and stomach adenocarcinomas, respectively; Figure 5). Our findings suggested that certain germline variants in the tumor-derived samples whose frequency was significantly greater when compared those from healthy samples could potentially elicit biological consequences in increasing tumor susceptibility.

Furthermore, we examined the genes in which these variant candidates potentially associated with tumor susceptibility and validated using independent dataset were located. We investigated the

genes and their association with potentially interesting molecular pathways such as cell proliferation or cell cycle regulation using the Ingenuity® Pathway Analysis. In doing so, we found that a pathway responsible for epithelial junction signaling was among the top ten predicted pathways in which the variant candidates were associated (Table 3 and 4). Within this pathway, the KOVA variants with potential association with cancer predisposition were located on Notch1, which has been known to be associated with cancer development and tumorigenesis (9). The association are preliminary predictions and require a more thorough investigation to dissect their potential biological impact in the future. Taken together, our findings demonstrate rare germline variants that exhibit a high prevalence in tumor-derived samples compared to healthy normal samples, indicating that certain KOVA variants may potentially function as a predisposition for tumorigenesis.

DISCUSSION

As a collaborative effort from multiple research groups, we obtained and performed whole-exome sequencing of tumor-paired normal samples and samples from healthy individuals with no known clinical conditions. To our knowledge, a high-quality genomic variant catalog specific to the Korean population of this scale is unprecedented. We also show that such a population-specific resource for the Korean population is necessary according to population genetics and comparative analyses with other ethnic groups provided by previous reports (1-5). Our population genetics analyses demonstrate the rationale and the need for cataloguing a Korean-specific variant archive as our findings have indicated the unique ethnicity of the Korean population when compared other ethnic groups. We developed a highly optimized analysis pipeline for the processing and evaluation of 1,055 WES samples. The WES-based samples were sequenced to an average depth of 75x and processed using the KOVA pipeline with extensive filtering steps to yield 101 singletons per sample and a total of 293,049 high-quality coding variants. The KOVA analysis pipeline is robust and can easily accommodate potential addition of datasets as well as expand to whole-genome sequencing data in the future. However, there are plenty of areas that the platform can be improved. Numerous updates and improvements have been made to the implemented tools and available packages since the KOVA analysis pipeline was developed

and optimized for the study. By improving the KOVA analysis with the latest versions of the publicly available tools and further optimized to accommodate for the massive increase in scale presented by whole-genome sequencing data, our analysis platform may serve as a valuable asset to future endeavors that seek to curate a high-quality database like that of KOVA and those that have been previously reported.

Our findings suggest a group of rare germline variants in KOVA that may be associated with cancer susceptibility. We validated our observation with these potentially interesting variants exhibiting greater prevalence in tumor-paired normal samples over healthy normal samples by examining their behavior in independent datasets. Available variant data from corresponding cancer types on the TCGA database served as the cancer-derived cohort while the European and East Asian variants on the 1000GP database were used as the healthy normal cohort. Accordingly, we found consistent patterns in prevalence in the independent datasets that were statistically significant. We further examined the variant candidates and found that a cancer-related gene within a biologically-relevant pathway was implicated with the variant candidates. Such findings indicate that there is substantial evidence for delving deeper into these candidates for a more thorough functional assessment of their role in cancer development and susceptibility to tumors. However, these findings are preliminary and albeit, interesting, initial predictions into their potential biological role in the context of cancer

development. Further studies into these susceptibility candidates is required using larger independent datasets followed by more in-depth functional analyses into their potential biological implications. Nonetheless, the KOVA database presents a valuable resource in utilizing a large, high-quality variant information available for investigation into disease progression or in this case, predisposition. In addition, the KOVA database is limited by the sequencing platform used for the samples. Although KOVA serves as a high-quality database of Korean-specific within the coding region, further study into the noncoding region based on large cohort samples using whole-genome sequencing may present even further insights into the variant profiles of the Korean population and provide a more comprehensive database for studies in the future. As next-generation sequencing efforts become more advanced and affordable, KOVA will be improved and expanded to serve as a resourceful addition to the variant databases currently available and further aid future research endeavors in various fields pursuing disease-based, population, or evolutionary genetic studies of the Korean population.

METHODS

Cohorts and sample preparation

All samples collected for the curation of the KOVA database were obtained with written informed consent and following the guidelines and standards outlined by the Institutional Review Board of each respective research group, Seoul National University, Ewha Womans University, Asan Medical Center, and Samsung Medical Center (Table 1).

Data processing and variant calling

Samples were processed using an in-house pipeline as illustrated in Figure 1. Raw sequenced reads were quality trimmed using Sickle (version 1.33, https://github.com/najoshi/sickle) and aligned to the hg19 build of the human genome using the Burrows-Wheeler Aligner (BWA version 0.7.5a). Duplicate reads were marked for removal using Picard Tools (version 1.93). GATK packages (version 2.4-7) were implemented for local realignment and recalibration after the preprocessing steps. The data was combined from all research groups across varying exome capture platforms and variants were called in the highly parallel multi-sample calling mode available on GATK's UnifiedGenotyper. In doing so, we were able to identify any variant calls with homozygous reference or missed calls due to low coverage. The high level of variant quality was achieved

by applying rigorous filtering criteria. Variants that did not exhibit a minimum genotype quality of 30 or depth coverage of 10 were excluded. A max-missing filter of 30% was applied across all variants in which variants missing genotypes in more than 30% of the samples were excluded. Additionally, variants that did not satisfy the Hardy-Weinberg equilibrium allelic frequency of $P < 10^{-6}$ were excluded from further analysis. In-depth description of the data processing and variant calling pipeline including step-by-step instructions and parameters is detailed in Appendix.

Principal component analysis

Principal component analysis (PCA) was performed to show relationship between KOVA and the various previously reported databases. From the 1000GP database, African (AFR) populations excluding the ASW (Americans of African Ancestry in SW USA) and ACB (African Caribbeans in Barbados) and East Asian (EAS) populations including CDX (Chinese Dai in Xishuangbanna, China) and KHV (Kinh in Ho Chi Minh City, Vietnam) were obtained. Overlapping variants between KOVA, AFR, and EAS populations were extracted and combined into a single collective VCF file using VCFtools (version 0.1.15) (12). Common variants with MAF >5% were then filtered from the collective VCF file and used as the input data for the PCA. The R package SNPRelate (version 1.4.2.) (13) was used for clustering and FST was approximated with VCFtools under the –wir-fst-pop option (12, 14). Using the FST values, gene level

analysis was conducted where the FST value was assigned to each variant and variants were assigned to the position information of annotated genes on the GENCODE database (15).

Copy number variation analysis

Copy number variants (CNVs) were called by applying the default parameters and guidelines on the publicly available CODEX software (16). To control for the potential deviations and noise introduced by using varying exome capture kits across the research groups, we performed the CNV calling algorithm to samples from each research group independently and combined the called CNVs for post-analysis. From the Database of Genomic Variants database (DGV, http://dgv.tcag.ca), previously reported CNV information was obtained. The CNVs from studies using WGS or SNP array platforms since 2009 were collected and compared with the CNVs detected from the KOVA samples. The intersecting DGV and KOVA CNVs were determined using BEDtools using the -r 0.5 option which sets the minimum overlap between each CNV to 50% (17). Interesting CNV candidates were visually inspected using the Integrative Genomics Viewer. The ID:gssvL59302 from the DGV database was obtained for the global profiling of the SIGLEC14 CNVs.

Assessing roles of rare germline variants on tumor susceptibility

The variants from the tumor-paired normal samples (364 and 77 samples for lung and stomach adenocarcinoma, respectively) and the healthy samples with no previous clinical history (134 samples) were filtered from the single collective VCF file using the VCFtools –keep option (12), which discard any variants that did not result from the samples indicated by the option. The variant MAFs from tumor-paired samples for lung and stomach cancers were compared to that of variants from the healthy samples independently. The KOVA variants that exhibited a MAF greater than 0.01 and exhibited 1.5 folds greater MAF than that of the variants from heathy variants were selected for further analysis. Furthermore, the variants were annotated using ANNOVAR to include prediction scores from four functional prediction parameters (SIFT, PolyPhen-2, MutationAssessor and GERP++) (19, 20). As an additional filtering conditions, the variants that were predicted to be deleterious in two out of the four prediction tools were selected as potentially interesting candidates of tumor susceptibility. By applying these rigorous filtering conditions, we obtained 54 and 72 variants for lung and stomach cancers, respectively. For validation, the tumor-paired samples from lung (n = 229) and stomach (n = 137) adenocarcinomas from TCGA database were obtained after excluding non-whites. These independent datasets representing cancer-derived samples were processed and variants were called using identical parameters as the KOVA samples using

the KOVA analysis pipeline. The WES samples for the 504 East Asian (EAS) and 503 European (EUR) individuals in the 1000GP database were obtained to maintain consistency among ethnic demographics and processed according to the KOVA pipeline parameters. The variant profiles from these databases were filtered and quality controlled using the procedure described previously for the KOVA variants. The MAF of the variants from the TCGA and 1000GP databases were compared with those from corresponding cancer types and healthy normal samples from KOVA, respectively. Statistical significance was assessed using Wilcoxon's rank sum test.

FIGURES and TABLES

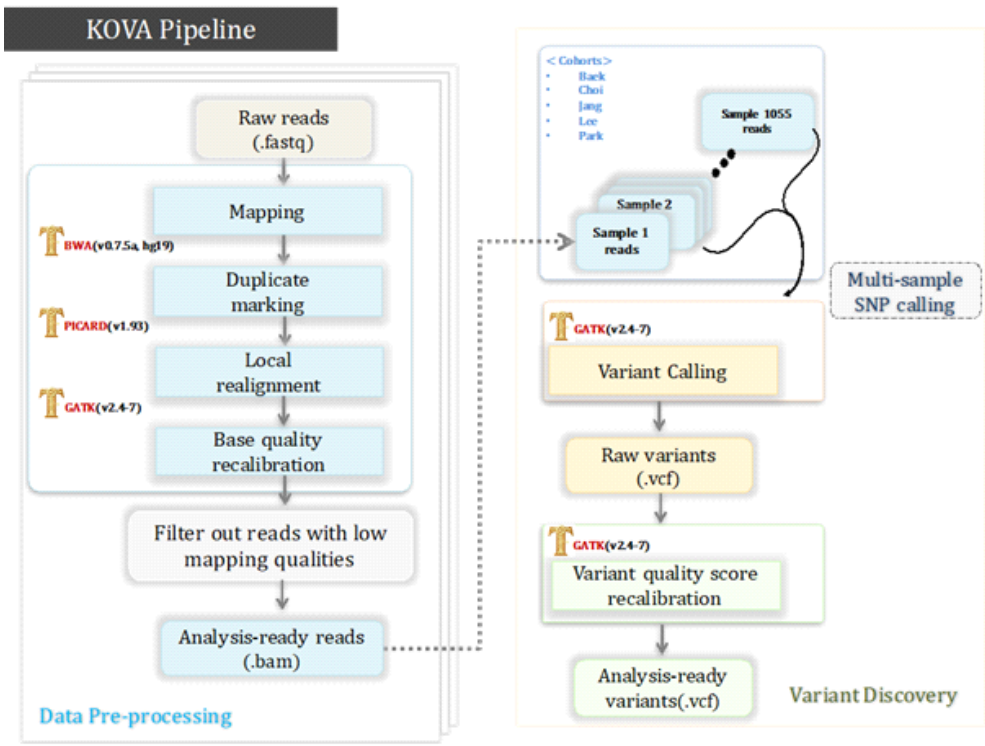

Figure 1: Overview of the KOVA Pipeline.

Schematic of the analysis pipeline outlining the sequential steps for the preprocessing of the raw sequenced reads to the calling of the variants. Variant calling is performed simultaneously in a multi-sample calling mode across all samples. For step-by-step instructions and details on processing WES-based data on the KOVA analysis pipeline, please see Appendix. (Figure obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

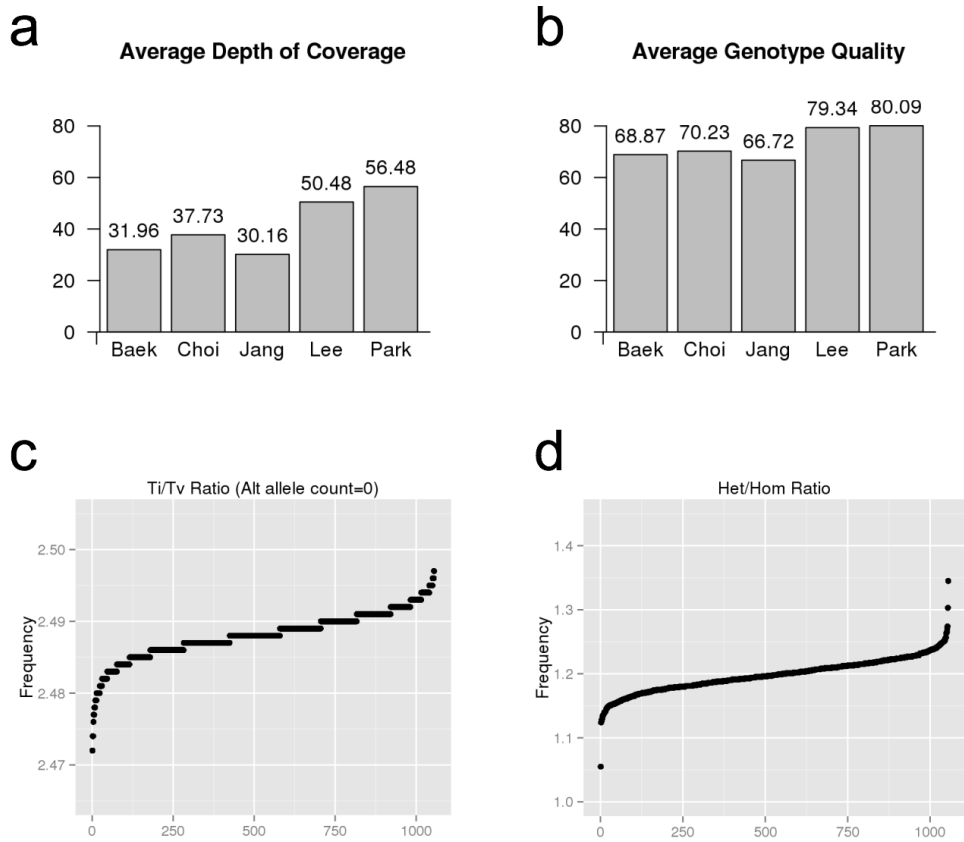

Figure 2: Quality control of KOVA variants.

Basic statistics of the variants profiles from KOVA is shown for samples from each research group. The high-quality variant profiles in KOVA was achieved by examining each data sample and excluding potential outliers. (a) Average depth of coverage of variants. (b) Average genotype quality. (c) Transition/transversion ratio of variants. (d) Heterozygous/homozygous ratio of variants. (Figure obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

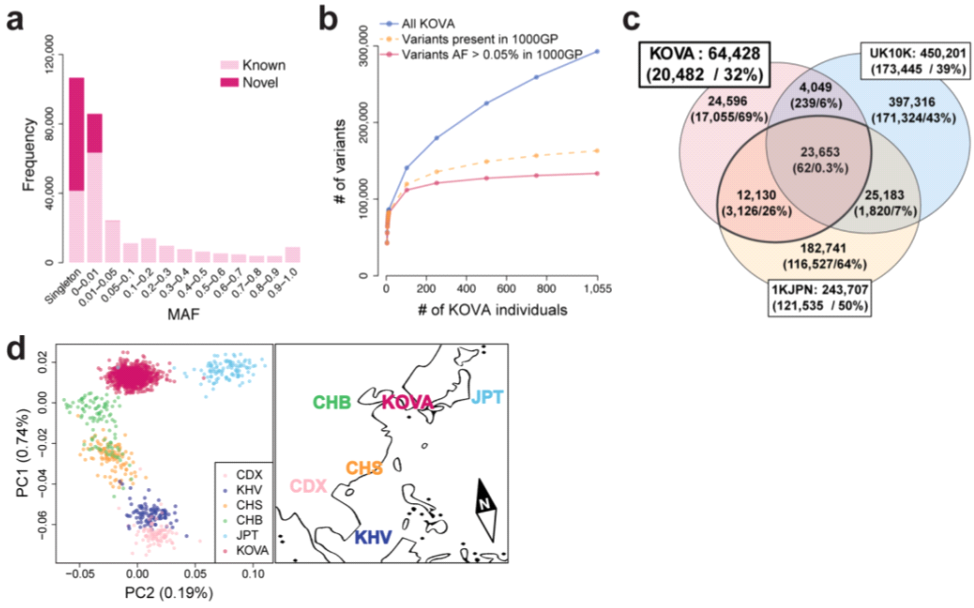

Figure 3: Population genetics and features of KOVA.

(a) Distribution of variant minor allele frequencies (MAFs). (b) Variant increment patterns as the number of the participants increases. (c) Venn diagram of coding variant comparisons among KOVA, Japanese population, and UK10K. Numbers and proportion of novel variants in each area are shown in the parentheses. (d) Principal component analysis of KOVA and East Asian populations from 1000GP database and corresponding geographical locations. (Figure obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

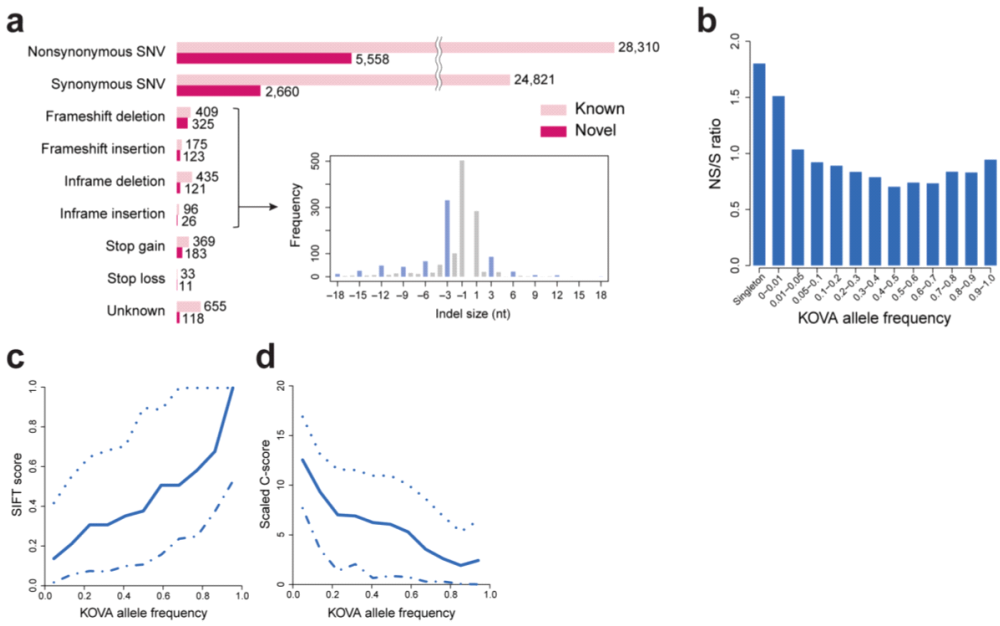

Figure 4: Functional analysis of KOVA coding variants.

(a) Numbers and ratios of novel and known variants categorized by function. The overlaid plot shows size distribution of indels, blue bar indicating multiples of three bases. (b) Nonsynonymous to synonymous SNV (NS/S) ratio by variant allele frequencies. (c) SIFT score and (d) Scaled C-score (CADD) by allele frequencies. (Figure obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

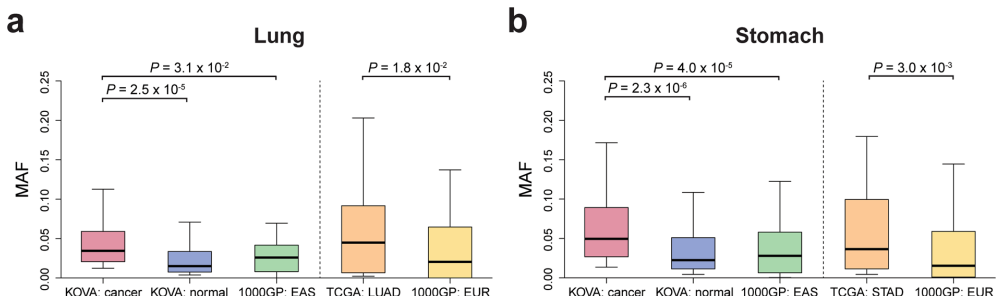

Figure 5: Evaluation of KOVA variants potentially associated with tumor susceptibility.

Rare KOVA variants potentially associated with tumor susceptibility and their MAFs determined from (a) lung adenocarcinoma (n = 54) and (b) stomach adenocarcinoma (n = 76) tumor-paired cohorts and for validation, variants from TCGA and 1000GP are shown. In both cancer types, the variants identified as exhibiting higher prevalence in tumor-paired samples demonstrate similar pattern in the independent datasets, indicating that they may be associated with tumor predisposition. LUAD: lung adenocarcinoma; STAD: stomach adenocarcinoma; EAS: East Asian; EUR: European. *P* values were calculated using Wilcoxon rank sum test. (Figure obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

Group	Sample Count	Sample Information	Capture Platform	Mean Depth of Coverage
Baek	237	Lung adenocarcinoma normal	Agilent SureSelect 50Mb	54X
Choi	190	Parent of a rare disease (166) Cancer normal (24)	Roche V2	71X
Jang	123	Healthy	Agilent SureSelect 50Mb	51X
Lee	214	Lung adenocarcinoma normal (127) Stomach cancer normal (76) Healthy (11)	Agilent SureSelect 50Mb Agilent SureSelect V4	93X
Park	291	Breast cancer normal (137) Glioblastoma normal (62) Colorectal cancer normal (13) Parent of a rare disease (77) Family of hearing loss (2)	Illumina TruSeq Agilent SureSelect V4 Agilent SureSelect V5	109X
Total	1055			75X

Table 1: Summary of sample cohorts.

The number of samples and the summary of sample information among the samples as well as the capture platforms used for the whole exome sequencing and average coverage depths are noted. (Table obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

Types	SNVs	Known	Novel
Nonsynonymous SNV	33,893	21,721	12,172
Synonymous SNV	27,494	21,008	6,486
Frameshift deletion	735	129	606
Frameshift insertion	298	74	224
Inframe deletion	557	178	379
Inframe insertion	122	60	62
Stop gain	553	229	324
Stop loss	44	20	24
Unknown	773	538	235
Total Coding	64,469	43,957	20,512

Table 2: Summary of exonic variants in KOVA.

From the total of 64,428 exonic variants in KOVA, the distribution of variants according to their type and annotation by the dbSNP database (build 147) have been detailed. (Table obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017).

Ingenuity Canonical Pathways	Cellular Function	Q value
Autoimmune Thyroid Disease Signaling	Endocrine System Disorders	1.74E-05
Graft-versus-Host Disease Signaling	Cell Death and Survival	2.14E-05
Antigen Presentation Pathway	Hematological System Development and Function	1.07E-04
B Cell Development	Cellular Growth and Proliferation	3.31E-04
IL-4 Signaling	Cellular Immune Response	3.09E-03
Protein Kinase A Signaling	Cellular Growth and Proliferation	3.31E-03
Crosstalk between Dendritic Cells and Natural Killer Cells	Cell-To-Cell Signaling and Interaction	3.63E-03
Epithelial Adherens Junction Signaling	Cellular Growth, Proliferation and Development	7.41E-03
Allograft Rejection Signaling	Cell-To-Cell Signaling and Interaction	7.76E-03
Nur77 Signaling in T Lymphocytes	Cell Death and Survival	8.71E-03

Table 3: IPA results for lung adenocarcinoma

The top ten significantly associated pathways using the tumor-paired normal samples from lung adenocarcinoma is shown. The highlighted pathway indicates potentially interesting pathway found to be common in both lung and stomach cancers. Q values were determined the Fisher ¯s exact test and corrected using the false discovery method. (Table obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017, unpublished data).

Ingenuity Canonical Pathways	Cellular Function	*Q* value
Chondroitin Sulfate Degradation (Metazoa)	Carbohydrate Metabolism	5.89E-03
Dermatan Sulfate Degradation (Metazoa)	Cellular Assembly and Organization	6.92E-03
Retinoate Biosynthesis I	Lipid Metabolism	8.13E-03
TREM1 Signaling	Cell-To-Cell Signaling and Interaction	8.71E-03
The Visual Cycle	Endocrine System Development and Function	1.12E-02
JAK/Stat Signaling	Cancer, Cell Cycle	1.38E-02
Retinol Biosynthesis	Vitamin A Biosynthesis	1.45E-02
Clathrin-mediated Endocytosis Signaling	Cellular Function and Maintenance	1.91E-02
Role of JAK family kinases in IL-6-type Cytokine Signaling	Cellular Growth and Proliferation	2.04E-02
Epithelial Adherens Junction Signaling	Cellular Growth, Proliferation and Development	2.29E-02

Table 4: IPA results for stomach adenocarcinoma

The top ten significantly associated pathways using the tumor-paired normal samples from stomach adenocarcinoma is shown. The highlighted pathway indicates potentially interesting pathway found to be common in both lung and stomach cancers. Q values were determined the Fisher ¯s exact test and corrected using the false discovery method. (Table obtained from Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Sci. Reports 2017, unpublished data).

REFERENCES

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. Nature 526, 68-74 (2015).
2. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291 (2016).
3. Huang, J. *et al.* The UK10K project identifies rare variants in health and disease. Nature 526, 82-90 (2015).
4. Nagasaki, M. *et al* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat. Commun. 6, 8018 (2015).
5. Higasa, K. *et al.* Human genetic variation database, a reference database of genetic variations in the Japanese population. J. Hum. Genet. 61, 547-553 (2016).
6. Wang, J. *et al.* Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31, 318-323 (2015).
7. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. Nat. Methods 7, 248-249 (2010).
8. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310-315 (2014).
9. Kumar, P. *et al.* Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073-1081 (2009).
10. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901-913 (2005).
11. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).
12. Danecek, P. *et al.* The variant call format and VCFtools. Bioinformatics 27, 2156-2158 (2011).
13. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 13, 3326-3328 (2012).
14. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the

analysis of population structure. Evolution 38, 1358-1370 (1984).
15. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760-1774 (2012).
16. Jiang, Y. *et al.* CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 43, e39 (2015).
17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).
18. Thorvaldsdóttir, H. *et al.* Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics 14, 178-192 (2013).
19. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164-e164 (2010).
20. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. 34, E2393-2402 (2013).

APPENDIX

Data processing using KOVA Analysis Pipeline

Step 1: Sickle - Quality Trimming

Input: Raw sequencing reads

Output: Quality trimmed reads

Prerequisite: None

Description:

The input files are in FASTQ format and listed on a text file that is used as a constant reference source throughout the analysis. Each FASTQ file is processed individually and outputted as indicated by the user. The number or size of the unpaired output can be an indicator of the sequence quality or in certain cases, mislabeling of the forward and reverse sequencing data.

Options:

-pe paired end

-t quality type

-f forward reads input

-r reverse reads input

-o forward reads output

-p reverse reads output

-s unpaired reads output

Example:

```
sScript  = '<path_to_sickle> pe '
sScript += '-t sangar '
sScript += '-f example.fw.fastq -r example.rv.fastq '
sScript += '-o example.fw.qual.fastq -p example.rv.qual.fastq '
sScript += '-s unpaired.out.fastq '
```

Step 2: BWA - Alignment 1/2

Input: Quality trimmed reads

Output: Intermediate reads for alignment

Prerequisite: Reference genome file

Description:

Quality trimmed files are preprocessed for alignment. Each forward or reverse files are processed individually and outputs an intermediate file that is used for alignment in the next step.

Options:

-aln alignment

-f intermediate output

-t CPU thread count

Example:

```
sScript  = '<path_to_BWA> aln hg19.fa example.fw.fastq '
sScript += '-f forward.qual.sai ' #arbitrary file tag
sScript += '-t 4 '                #depends on system
```

Step 3: BWA - Alignment 2/2

Input: Quality trimmed reads and intermediate alignment files

Output: Aligned reads

Prerequisite: Reference genome file

Description:

Quality trimmed files and the intermediate files are processed together to align the reads to the indicated version of the human genome. Not indicated as an input but the reference genome file needs to have the indexed file as well in the same directory.

Options:

-sampe combine forward and reverse intermediates to SAM output

-r read group labeling

Example:

```
sScript  = '<path_to_BWA> sampe hg19.fa '
sScript += '-r \'@RG\\tID:<File ID>\\tPL:<'Illumina'>\ '
sScript += example.fw.qual.sai example.rv.qual.sai '
sScript += example.fw.qual.fastq example.rv.qual.fastq '
sScript += '> example.sam '
```

Step 4: SAMtools – SAM to BAM, sort and index BAM

Input: Aligned SAM file

Output: Sorted and indexed BAM file

Prerequisite: Reference genome index file

Description:

The resultant aligned file in SAM format is converted to BAM (binary version of SAM), sorted, and indexed according to default parameters using SAMtools.

Options:

-view bt	convert SAM to BAM as standard output
-sort	sort SAM or BAM
-index	index BAM
-T	temporary log file
-o	output file

Example:

```
sScript  = '<path_to_samtools> view -bt hg19.fai '
sScript += example.sam -o example.bam; '
sScript += '<path_to_samtools> sort -T temp.log example.bam '
sScript += '-o example.sorted.bam; '
sScript += '<path_to_samtools> index example.sorted.bam '
```

Step 5: PicardTools - Check Read Group Header Labels

Input: Sorted and indexed BAM file

Output: BAM file with formatted read groups labels for GATK pipeline

Prerequisite: None

Description:

For processing aligned files using the GATK, certain measures have to be checked so that there are not issues with downstream processes. Each aligned BAM file requires proper heading and labels so that they can be properly processed by the GATK tools. Each labeling convention is arbitrary and can be adjusted according to user's needs.

Options:

-I input file

-O output file

-RGLB read group label

-RBPL read group platform

-RBPU read group platform unit (eg. barcode ID)

-RBSM read group sample name

-VALIDATION_STRINGENCY check read group format (null, lenient, or strict)

-SORT_ORDER criteria to sort based on (queryname, coordinate, duplicate)

Example:

```
sScript     = 'java -jar <path_to_AddOrReplaceReadGroups.jar> '
sScript    += 'I=example.sorted.bam '
sScript    += 'O=example.ar.bam '
sScript    += 'RGLB=TCGA '
sScript    += 'RGPL=illumina '
sScript    += 'RGPU=LungCancer '
sScript    += 'RGSM=Patient1 '
sScript    += 'VALIDATION_STRINGENCY=LENIENT ' # optimal
sScript    += 'SORT_ORDER=coordinate ' # genomic coordinate
```

Step 6: PicardTools - Merge SAM/BAM files

Input: BAM file with formatted read groups labels for GATK pipeline
Output: Merged BAM file
Prerequisite: None

Description:
This procedure can be used to merge SAM or BAM files that require combinations or editing. Without the assume-sorted option, the files are resorted which can demand significant computational resources.

Options:

-I	input file
-O	output file
-VALIDATION_STRINGENCY	check read group format
-ASSUMED_SORTED	true or false

Example:

```
sScript      = 'java -jar <path_to_MergeSamFiles.jar> '
sScript     += 'I=example.ar.bam '
sScript     += 'O=example.ms.bam '
sScript     += 'VALIDATION_STRINGENCY=LENIENT ' #optimal
sScript     += 'ASSUME_SORTED=true '
```

Step 7: PicardTools - Mark Duplicates

Input: Merged BAM file
Output: BAM file with duplicated removed
Prerequisite: None

Description:
Duplicate reads can occur albeit rarely during the alignment or sorting procedures from previous steps. These must be marked for removal so they do not incur any issues in downstream processes. The remove-duplicates option can be left false which will just mark them within the output file.

Options:

-I	input file
-O	output file
-M	intermediate file
-VALIDATION_STRINGENCY	check read group format
-ASSUMED_SORTED	true or false
-REMOVE_DUPLICATES	true or false

Example:

```
sScript     = 'java -jar <path_to_MergeSamFiles.jar> '
sScript    += 'I=example.ms.bam '
sScript    += 'O=example.markdups.bam '
sScript    += 'M=example.markdups.temp '
sScript    += 'VALIDATION_STRINGENCY=LENIENT ' # optimal
sScript    += 'ASSUME_SORTED=true '
sScript    += 'REMOVE_DUPLICATES=true; '
sScript    += '<samtools path> index example.markdups.bam; ' #reindex
```

Step 8: GATK - Realigner Target Creator

Input: BAM file with duplicated removed

Output: Interval files for used for local realignment

Prerequisite: Reference genome file

Known indels reference files:

Mills__and_1000G_gold_standard.indels.hg19.sites.vcf

1000G_phase1.indels.hg19.sites.vcf

Description:

The first step in the GATK tool package requires the listed reference files in VCF format. They can be obtained from the GATK homepage in various human genome versions. Outputs intermediate files for realignment in the next step.

Options:

-T	GATK tool name
-R	reference file
-I	input file
-o	output file
-log	log file
-known	known indel reference files

Example:

```
sScript     = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript    += '-T RealignerTargetCreator '
sScript    += '-R hg19.fa '
sScript    += '-I example.markdups.bam '
sScript    += '-o example.realn.intervals'
sScript    += '-log example.log'
sScript    += '-known Mills__and_1000G_gold_standard.indels.hg19.sites.vcf'
sScript    += '-known 1000G_phase1.indels.hg19.sites.vcf'
```

Step 9: GATK - Indel Realigner

Input: BAM file with duplicated removed and indel intervals

Output: BAM file with realigned indels

Prerequisite: Reference genome file

Description:

Realignment takes the interval files with the previous BAM file with duplicates marked and removed. The output can take longer than other procedures and if possible, it can be helpful use a high memory volume machine, for example, R730 machines with 256GB of memory are recommended.

Options:

-T	GATK tool name
-R	reference file

-I input file
-o output file
-log log file
-targetIntervals indel intervals

Example:

```
sScript      = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript     += '-T IndelRealigner '
sScript     += '-R hg19.fa '
sScript     += '-I example.markdups.bam '
sScript     += '-o example.realn.bam '
sScript     += '-log example.log'
sScript     += '-targetIntervals example.realn.intervals '
```

Step 10: GATK - Base Recalibration 1/2

Input: BAM file with realigned indels
Output: Base Recalibration Intermediate File
Prerequisite: Reference genome file
Known SNP reference file (dbSNP)
Known indels reference files:
Mills__and_1000G_gold_standard.indels.hg19.sites.vcf
1000G_phase1.indels.hg19.sites.vcf

Description:
After realignment, the base quality score are recalibrated using the intermediate files generated by this step. The dbSNP reference file in VCF format is used as input along with the previously used reference files available on the GATK homepage.

Options:
-T GATK tool name
-R reference file
-I input file

-o output file
-log log file
- knownSites known references files

Example:

```
sScript = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript += '-T BaseRecalibrator '
sScript += '-R hg19.fa '
sScript += '-I example.realn.bam '
sScript += '-o example.recal.grp ' #arbitrary file tag
sScript += '-log example.log'
sScript += '-knownSites dbsnp_147.hg19.vcf'
sScript += '-knownSites Mills_and_1000G_gold_standard.indels.hg19.sites.vcf'
sScript += '-knownSites 1000G_phase1.indels.hg19.sites.vcf'
```

Step 11: GATK - Print Reads

Input: Base Recalibrated BAM File
Output: Finalize BAM file for variant calling
Prerequisite: Reference genome file

Description:
This is a simple step that generates the recalibrated base score and prints the output as a new BAM file. After the new BAM file is generated, the BAM file is sorted and indexed as before using SAMtools with default parameters.

Options:
-T GATK tool name
-R reference file
-I input file
-o output file
-BQSR base recalibration intermediate file

Example:

```
sScript      = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript     += '-T PrintReads '
sScript     += '-R hg19.fa '
sScript     += '-I example.realn.bam '
sScript     += '-o example.recal.bam '
sScript     += '-BQST example.recal.grp; '
sScript     += 'sort example.recal.bam example.recal.sorted.bam;' # re-sort
sScript     += 'index example.recal.sorted.bam;' # reidnex BAM
```

Step 12: GATK - Unified Genotyper

Input: Finalize BAM file for variant calling

Output: VCF file according to chromosome window

Prerequisite: Reference genome file
Known SNP reference file (dbSNP)
For multiplexed calling:
List of BAM files
List of chromosome windows (eg. chr1:1-4001 #window size =4000)

Description:

This is the most time-consuming step and required a higher processor machine rather than memory-heavy machine as the algorithm itself is optimized to use minimum memory resources, ex) optiplex nodes are recommended. The full list of all the input BAM files is indicated using the -I options with -I prefixing each BAM file. The command is iterated over all predetermined chromosome IDs and shifting window positions with the -L option. It is recommended that the shifting windows have at least a 200bps of overlapping region. The two master lists must remain constant throughout the analysis from this step and on.

Options:

-T GATK tool name

-R	reference file
-I	input file
-o	output file
-L	chromosome window
-glm	SNP or INDELS or BOTH
-stand_call_conf set	threshold for calling
-stand_emit_confset	threshold for emitting (discarding)
--dbsnp	known SNPs

Example:

```
sScript      = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript     += '-T UnifiedGenotyper '
sScript     += '-R hg19.fa '
sScript     += '-glm BOTH '
sScript     += '-stand_call_conf 30.0 '
sScript     += '-stand_emit_conf 10.0 '
sScript     += '-I <list of BAM files> '
sScript     += '-o example.chr1:1-4001.vcf '
sScript     += '-L chr1:1-4001 '
```

Step 13: GATK - Concatenate Window VCF files

Input: VCF file according to chromosome window

Output: VCF file according to chromosome

Prerequisite: Reference genome file

Description:

To reduce I/O lag and as an additional validity checkpoint, the individually processed genomic window files are combined according to chromosome IDs. This requires a GATK version 3.1-1 or above as the previous versions were found to be bugged. The resultant chromosome-specific variants are sorted according to coordinates using VCFtools.

Options:

-V	window files from each chromosome
-out	output file
-assumeSorted	true by default

Example:

```
sScript      = '<path_to_GenomeAnalysisTK.jar> ' #GATK version 3.1-1 or above
sScript     += ' org.broadinstitute.sting.tools.CatVariants '
sScript     += '-R hg19.fa '
sScript     += '-V <list of VCF files> '
                  #example.chr1-1-4001.vcf example.chr400-8001.vcf…
sScript     += '-out example.chr1.vcf ' #all VCFs for chr1 combined
sScript     += '-assumeSorted; '
sScript     += '<path_to_VCFtools> vcf-sort example.chr1.vcf ' #re-sort VCF
```

Step 14: GATK - Combined Variants

Input: VCF file according to chromosome window

Output: VCF file according to chromosome

Prerequisite: Reference genome file

Description:

The chromosome-specific VCF files are combined into a full VCF file. This step requires massive memory resources, typically >100GB to process and combine all the VCF files in a timely manner. The priority list allows the VCF files to be in a desired order of chromosomes. The UNIQUIFY option is used when each of the files being combined are from independent datasets and there are no overlapping regions in the windows.

Options:

-T	GATK tool name
--variant	input VCF files
-R	reference file

-o output file
-genotypeMergeOptions PRIORITIZE or UNIQUIFY (same or different callset)
-priority sort criteria

Example:

```
sScript      = 'java -jar <path_to_GenomeAnalysisTK.jar> '
sScript     += '-T CombineVariants '
sScript     += '-R hg19.fa '
sScript     += '--variants:chr1 example.chr1.vcf' #chr2 example.chr2.vcf etc..
sScript     += '-o example.full.vcf '
sScript     += '-genotypeMergeOptions PRIORITIZE '
sScript     += '-priority <list of chromosome IDs> '#chr1 chr2 chr3 ch4..
```

The above procedure for massively parallel calling of variants followed by the combining the VCF files has been thoroughly tested to ensure that there is zero loss of variant information when compared to calling variants in a single, genome-wide manner. Accordingly, parallel calling can reduce the total analysis time by 10-folds or more depending on computational resources when compared to a single, genome-wide calling method.

국문 초록

한국인 유전체 변이 표준 데이터베이스 (KOVA) 구축

박진만

자연과학대학 생명과학부

서울대학교 대학원

2001년 인간게놈 프로젝트를 통해 한 개인의 전체유전체가 판독되어 후속 유전체 연구의 규모가 전장규모로 확대되었다. 하지만 인간게놈 프로젝트와 그 후 보고된 대규모 데이터 베이스들의 구축은 대부분 코카서스 후손의 샘플들을 기반으로 진행되었으며 코카서스 후손의 유전적 특성에 의해 바이어스 되었다. 따라서 지금까지 보고된 데이터베이스들을 이용한 인간 게놈 변이의 식별 및 기능 규명을 위한 연구들은 비 코카서스 인종에 대한 분석 정확성과 검출력을 상실했다. 우리는 한국인의 게놈 변이에 대한 연구의 정확성을 높이기 위해 1,055개의 건강한 한국인 샘플들을 기반으로 대규모 코딩 변이체 데이터베이스를 확립하고 큐레이팅 했다. 이 샘플들은 평균 75x의 깊이로 시퀀싱 되었으며, 샘플 당 101개의 싱글톤 변이(singleton variant)들을 찾아내었다. 우리는 건강한 한국인 집단 분석을 통해 아프리카와 유럽의 다른 소수 민족 집단과 비교할 수 있는 명확한 민족 집단 특성을 발견했고 이는 한국인 인구가 독립적인 인구 집단이고 한국인의 특정 게놈 변이 데이터베이스로부터 많은 이

익을 받을 수 있는 개별 민족 집단임을 나타냈다. 우리는 인구 유전학 분석을 통해 Korean Variant Archive (KOVA)와 Exome Aggregation Consortium 데이터베이스를 함께 사용하여 한국인의 exome들을 검토할 때 유전자 변이 필터링 능력이 향상됨을 확인했다. 또한 우리는 종양 형성의 감수성과 관련된 잠재적 생식 세포 내 변이를 가진 집단에 대한 정보를 제공하였으며, 이를 The Cancer Genome Atlas (TCGA)와 1000 Genome Project 데이터베이스를 이용해 검증했다. KOVA는 고품질 exome 시퀀싱으로부터 얻은 한국인 특이적 유전체 변이 정보를 이용하여 큐레이팅된 최초의 데이터베이스이며 한국인의 유전체 변이 연구를 위한 중요한 자원으로 사용될 것을 기대한다.

주요어: Coding variants, population genetics, tumor susceptibility, whole exome sequencing

학번: 2014-25015

비고: 내용의 많은 부분은 제출자 본인이 관련된 연구 결과인, Lee S. M., Seo J. H., Park J. M., and Nam J. Y. *et al.* Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. Sci. Reports. 7, 4287 (2017). 로부터 차용하였음을 밝힙니다.