Gene Identification via Phenotype Sequencing

Version 1.2

User manual

Zhu, Z and Chen X

2015-06-01

Table of contents

Chapter 1. Algorithm

1.1 Overview

The advent of next generation sequencing has enabled fast and cost-effective genotyping, which can significantly accelerate the process of gene identification in forward genetics studies. Sequencing-based forward genetics investigations are frequently referred as phenotype-sequencing studies.

The success of a typical phenotype-sequencing study depends on a range of experimental and analysis choices, including the number of phenotype-exhibiting samples being sequenced, the genomic region being sequenced, the quality and depth distribution of sequencing data, the approach to map sequencing reads onto genome, the variant calling method, the approach to filter unlikely-causal variants, and the criterion to report candidate genes. Some previous studies have characterized the impacts of some of these choices (Ratan, Miller et al. 2013, Chilamakuri, Lorenz et al. 2014, Lelieveld, Spielmann et al. 2015). However, for investigators, it is still difficult to design an optimal analysis with an integral consideration of all factors that affect its chance of success. In particular, after the sequencing results are obtained, how to design an effective analysis procedure that fits the quality of this particular set of sequencing data.

To meet this analytical need, we developed the Gene Identification via Phenotype Sequencing (GIPS) tool. GIPS estimates four likelihoods to help optimize an analysis procedure. The four likelihoods are, **1) its chance to report candidate genes that are truly associated with the phenotype**; **2) the expected number of random genes that may be reported**; **3) the significance of each candidate gene to associate with the phenotype**; and **4) the significance of violating Mendelian assumption, if no gene is reported or all candidate genes have failed validation**.

As previously described, a range of experimental and analysis parameters can affect these four likelihoods of study effectiveness. In general, the impacts of experimental and analysis parameters on the four study-wise effectiveness indicators can be summarized into their impacts on two sample-wise effectiveness indicators: the “sensitivity” and “specificity” for each sample, with which an analysis procedure under evaluation is able to detect the phenotype-causing variants. Note that in this framework, it is permitted that different analysis procedures are applied on different samples, which allows fine-tuning of sample-specific analysis procedures based on sample-specific data qualities. With the sample-wise variant detection sensitivities and specificities, the combined study-wise likelihoods of analysis effectiveness can be computed in a recursive form, which is detailed in later sections.

Figure 1 illustrates the general workflow of GIPS, which formally considers seven aspects of an analysis procedure that affect its likelihoods of success in gene identification. These aspects include, **1) the number of phenotype-exhibiting samples being sequenced**, **2) the genomic region being**

sequenced, **3) the quality and depth distribution of sequencing data**, **4) the choice of software and parameters to align sequencing reads**, **5) the choice of software and parameters to call variants**, **6) the choice of strategies to filter variants that are unlikely to associate with the phenotype**, and **7) the criterion to report candidate genes**. The impacts of these choices can be estimated from real data, or customarily specified.

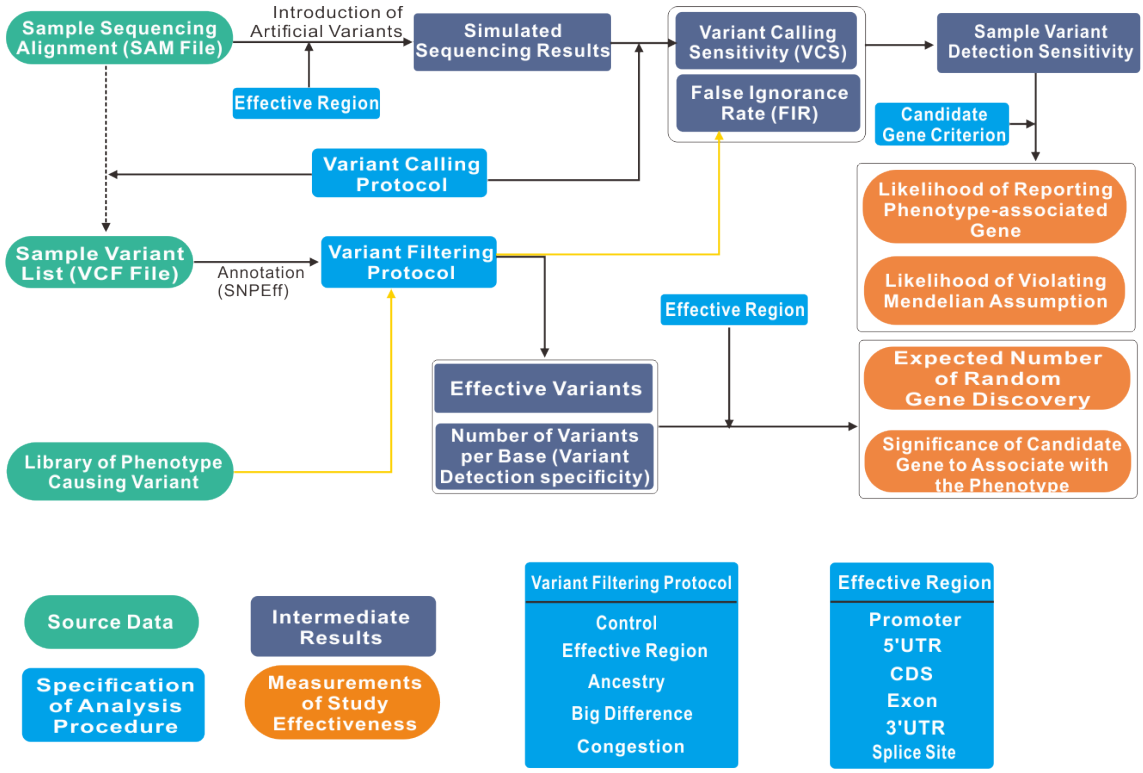


Figure 1. The general workflow of GIPS.

In general, impacts of "the genomic region being sequenced" and "the approach to filter unlikely-causal variants", on the variant detection sensitivity of each sample, can be estimated by using the same approach to filter a library of known phenotype-causing variants, and compute how much proportion of the phenotype-causing variants will be discarded. For human, the ClinVar database collected 19,334 Mendelian phenotype-associated variants (Landrum, Lee et al. 2014). Because the genome structures in higher organisms are similar, ClinVar may also be used as a reference library for other higher organisms, in case no appropriate library exists.

Impacts of "the quality and depth distribution of sequencing data", "the approach to map sequencing reads onto genome", and "the variant calling method", on the variant detection sensitivity of each sample, can be estimated by simulating a set of sequencing data with the same quality and depth distribution. The simulated sequencing reads are from a genome containing random artificial (simulated) variants. Therefore, the combined impact of these factors on variant detection sensitivity

can be estimated through computing how much proportion of the artificial SNPs is discovered.

GIPS measures the sample-wise specificity of an analysis procedure by computing the frequency of detected (after all variant filtering steps) number of variants per base in the effective genomic region. Assuming all detected variants are phenotype-unrelated, this frequency is used to compute how many genes are expected by chance to accumulate random mutations in multiple samples and pass the candidate gene criterion.

1.2 Variant detection sensitivity

Variant detection sensitivity (VDS) is computed for a specific analysis protocol that is applied on a specific sample. It has two components, variant calling sensitivity and false ignorance rate.

Variant calling sensitivity (VCS) characterizes how much proportion of the real variants can be observed with the analysis protocol under evaluation, before any filtering. It depends on the quality and depth distribution of the sequencing data and the approach to align sequencing reads and call variants.

It is not expected all variants are phenotype-related, e.g., synonymous mutations are unlikely to produce phenotype. In most phenotype-sequencing studies, filters are applied to remove observed variants that are unlikely to produce the target phenotype. In GIPS, false ignorance rate (FIR) is used to characterize the likelihood of filters to discard variants that may produce the phenotype.

Focusing on an "effective genomic region" (e.g., using exome sequencing) is considered a filter, because this action risks ignoring phenotype-associated variants that are located outside the effective genomic region. In addition, GIPS supports the use of four types of other filters. They are, **1) the control filter, which removes variants observed in phenotype-absent control samples**; **2) the ancestry filter, which removes variants replicated in multiple phenotype-exhibiting samples (applicable if all samples are known unrelated)**; **3) the big difference filter, which removes non-synonymous mutations that are unlikely to change protein function**; and **4) the congestion filter, which removes potentially spurious variants that are results of ambiguity in sequence alignment**.

Let:

D_{S_i} be the VDS of sample S_i;

H_{S_i} be the VCS of sample S_i;

Φ_{S_i} be the FIR of sample S_i;

We have:

$$D_{S_i} = H_{S_i} \times (1 - \Phi_{S_i}).$$

For each sample S_i, H_{S_i} and Φ_{S_i} can be estimated from real data or customarily specified. Note

that GIPS allows the usage of sample-specific sequence alignment/variant calling protocols.

1.2.1 Variant calling sensitivity

The VCS H_{S_i} of sample S_i can be estimated with real sequencing data. To estimate H_{S_i} for homozygous variants, the raw sequencing reads from sample S_i are first mapped to the reference genome. GIPS will then simulate an alignment result from a hypothetical genome of sample S_i, which harbors artificial SNPs. A number (the "*NUM_SIM_SNPS*" parameter in "*PROJECT.ini*", defaults to 5,000) of artificial SNPs, distributed uniformly over the "effective genomic region", are hypothetically introduced to the genome of S_i. The aligned sequencing reads are updated to reflect the introduction of artificial SNPs. To avoid complications, the artificial SNPs are only introduced at locations where the genome of S_i matches the reference genome. The artificial SNPs randomly change the original nucleotide bases to other ones. The mapped reads in the original alignment result (SAM file) are updated to reflect this change. For each mapped read covering an artificial SNP, its corresponding base is changed as follows. If the base is a correct read (i.e., matching the reference genome), it is changed to match the artificial SNP. If the base is an incorrect read and it does not match the artificial SNP, it is left unchanged. If the base is an incorrect read and it matches the artificial SNP, it is changed to match the reference genome. After this update, the simulated mapping result and the original mapping result will have the same reads quality, mapping quality, and depth distribution. The VCS H_{S_i} of sample S_i is therefore estimated as the proportion of the artificial SNPs that can be discovered with the user-supplied variant calling protocol (script).

Users may specify any variant calling protocol with a custom script, as long as it takes a sequence alignment result in SAM format (Li, Handsaker et al. 2009) as input and produces a variant call file in VCF format. Because GIPS will invoke these variant-calling scripts to call variants from the simulated alignment results, these scripts have to implement a convention of command line arguments. See 2.2.2 for details. Note that GIPS will not invoke these scripts to call variants from the original alignment results. Instead, users should use these scripts to call variants from the original alignment results, and supply the resultant VCF files to GIPS. This design is to ensure that the user-supplied variant calling scripts are all working.

To estimate H_{S_i} for heterozygous variants, GIPS uses the same procedure except that only a random half of the mapped reads in the original sequence alignment result (SAM file) are updated to reflect the introduction of artificial SNPs.

1.2.2 False ignorance rate

The FIR Φ_{S_i} of sample S_i does not depend on the sequencing result of S_i. It depends only on the variant filters that are applied on S_i. The same filtering protocol will always have the same FIR, however, it is allowed to use different filtering protocols for different samples. GIPS supports the use of five types of configurable filters (Figure 2). Two filters, the effective region filter and the big difference filter, are based on predicted variant functions (i.e., variant annotations), which aim to discard variants that are unlikely to produce any phenotype. The other three filters, the ancestry filter, the control filter, and the congestion filter, are study-based filters, which aim to discard variants that are unlikely to produce the target phenotype. Details about these filters are given below.

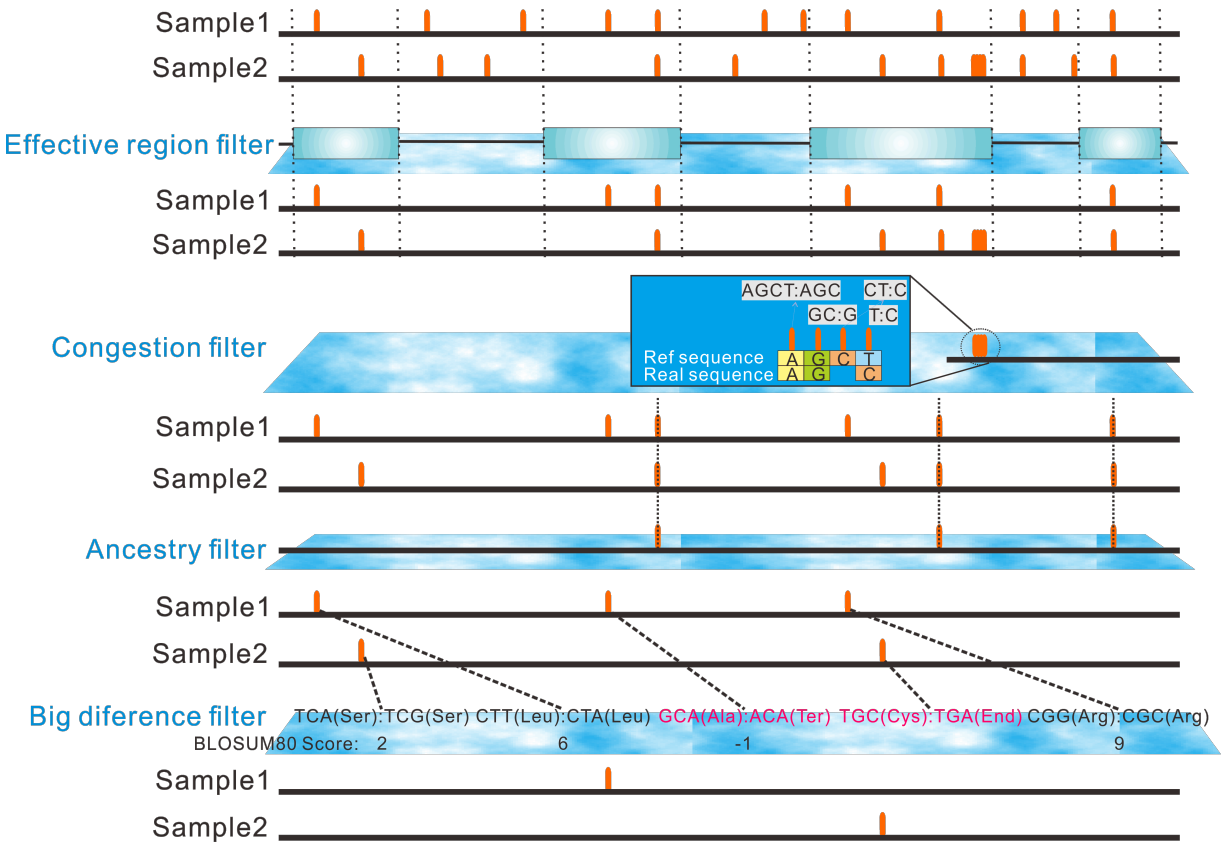


Figure 2. Variant filters.

Let:

Φ_{S_i} be the FIR of sample S_i;

$\Phi_{S_i,F}$ be the FIR of function based variant filters that are applied on S_i;

$\Phi_{S_i,S}$ be the FIR of study based variant filters that are applied on S_i;

We have:

$$\Phi_{S_i} = 1 - (1 - \Phi_{S_i,F}) \times (1 - \Phi_{S_i,S}).$$

The FIR of function-based variant filters $\Phi_{S_i,F}$ is estimated by applying the same filters on a library of known phenotype-causing variants, and compute how much proportion of these variants will be discarded. For human, the ClinVar database lists 19,334 Mendelian phenotype-associated variants (Landrum, Lee et al. 2014). For other higher species, because the genome structures in higher organisms are similar, ClinVar may also be used as a reference library in case no appropriate library for these species exists. Users can specify the reference variant library with the "*LIB_PHENOTYPE_VAR*" parameter in "*PROJECT.ini*". The ClinVar database has been included in the GIPS package. If no reference variant library is specified, ClinVar will be used as default.

The FIR of study based variant filters $\Phi_{S_i,S}$ is estimated as the proportion of "effective genomic region" that is ignored by these filters. Study based variant filters discard variants based on variants detected in other samples. However, variants detected in any sample could be spurious. Therefore, GIPS takes the proportion of "effective genomic region" ignored by study based variant filters as their combined FIR.

Whether to use a specific type of variant filter is specified with the "*VAR_FILTERS*" parameter in the "*[GLOBAL]*" section of the parameter file "*PROJECT.ini*".

VAR_FILTERS: [EBAC]

E: Effective genomic region; B: Big difference filter; A: Ancestry filter; C: Congestion filter.

Note that the "effect genomic region" filter is required. It will take effect even if "E" is not specified here.

In addition, the "control filter" will take effect if control samples are specified. See 1.2.2.5 for details. No specification of control filter is needed with this parameter.

1.2.2.1 Effective region filter

This filter discards variants that are outside the user-specified "effective genomic region".

Effective genomic region is specified according to the reference genome structure. An annotation of the reference genome is needed in generic feature format (GFF3) (Stein 2007). GIPS uses the below gene model to specify the effective genomic region (Figure 3).

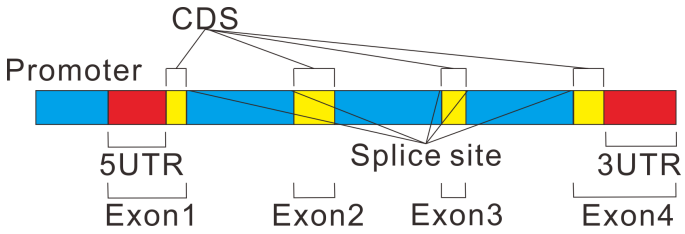


Figure 3. Gene structure model for specification of effective genomic region.

Users can specify whether to include the promoter region (keyword: *Promoter=length*), 5' untranslated region (keyword: *5UTR*), coding sequence (keyword: *CDS*), splice site (keyword: *SpliceSite=length*), exon (keyword: *EXON*), and 3' untranslated region (keyword: *3UTR*) into the effective genomic region. The lengths of promoter and splice site can be customized. The effective region specification applies to all samples, which is required for all GIPS runs. The "*[GLOBAL]*" section of the parameter file "*PROJECT.ini*" contains this specification:

REF_GENOME_ANNOTATION.GFF: /path/to/genome/annotation/file.gff3
EFF_REGION: Keywords separated by "|". The default value is "CDS|SpliceSite=2".

Example:
EFF_REGION: Promoter=1000|5UTR|CDS|3UTR|SpliceSite=2

1.2.2.2 Big difference filter

This filter discards non-synonymous variants that result in similar amino acids. The variants observed in each sample are first annotated by SnpEff (Cingolani, Platts et al. 2012). SnpEff assigns variants into different categories based on their putative effects. If this filter is used, GIPS will keep only variants that are annotated with "*HIGH*" effect and coding region variants that lead to significant amino acid changes. To define "significant amino acid change", users can specify a maximum residue similarity score and an amino acid scoring matrix. GIPS has 87 built-in scoring matrices retrieved from NCBI (ftp://ftp.ncbi.nih.gov//blast/matrices).

Big difference filter can have sample specific parameters. The parameters "*SCORE_MATRIX*" and "*MAX_AA_SCORE*" specified in the "*[GLOBAL]*" section of "*PROJECT.ini*" apply to all samples.

The "*SCORE_MATRIX*" and "*MAX_AA_SCORE*" parameters can also be specified in specific sample sections, which override the corresponding global parameters.

SCORE_MATRIX: Name of a scoring matrix, e.g., BLOSUM80. Defaults to "DEFAULT", which discards only synonymous SNPs. Available scoring matrices are listed in the "*config*" file ("*/path/to/GIPS_folder/config*").

MAX_AA_SCORE: Maximum amino acid similarity score as in the specified scoring matrix. Defaults to 0.

1.2.2.3 Ancestry filter

This filter discards variants that are detected in more than one phenotype-exhibiting samples. If all phenotype-exhibiting samples are known unrelated, the event of random mutations hitting the same genomic position in two or more samples is unlikely. Therefore, the same variant observed in unrelated samples probably indicates shared genetic background (which is different from the reference genome), rather than evidence of phenotype association.

Note that the ancestry filter should only be enabled when all the phenotype-exhibiting samples are known to be unrelated. For example, in plant research, mutants generated by chemical treatments are unrelated. The next version of GIPS will support gene identification with family-based experimental design.

As described above, the FIR of the ancestry filter is estimated as the proportion of "effective genomic region" covered by all detected variants in all phenotype-exhibiting samples. This estimation implements the consideration that mutations in two unrelated samples do have a small chance of hitting the same genomic location, and that some of the observed variants are spurious.

1.2.2.4 Congestion filter

Some genomic regions are highly similar, which makes alignment of sequencing reads difficult. Variants in these regions further complicate the alignment process. As a result, variants observed in similar genomic regions are more frequently false positive calls, especially in cases that many variants appear in a short region. Congestion filter discards a variant if there is another variant in its vicinity (11bp region centering on it).

As described above, the FIR of congestion filter is estimated as the proportion of "effective genome region" covered by all discarded variants and their 11bp vicinity regions.

1.2.2.5 Control filter

This filter discards variants observed in phenotype-absent control samples. The FIR of control filter is estimated as the proportion of "effective genomic region" covered by all observed variants in all control samples. This estimation implements the consideration that some of the observed variants in control samples are spurious.

The "*CONTROL*" parameter specified in the "*[GLOBAL]*" section of "*PROJECT.ini*" applies to all phenotype-exhibiting samples. The "*CONTROL*" parameters specified in specific sample sections override the global parameter for the respective samples.

Control variants are provided with one VCF file.

CONTROL: /path/to/control_sample.vcf

If there are multiple control samples, please merge the variants called from all control samples into one VCF file.

1.3 Variant detection specificity

GIPS measures the specificity of an analysis protocol by computing the frequency of detected variants per base (after all variant filtering steps) in the effective genomic region. This frequency is termed as background variant frequency (BVF) .

Let:

B_{S_i} be the BVF of sample S_i;

Nv_{S_i} be the number of variants detected (after filtering) from sample S_i;

l be the length of the effective genomic region;

We have:

$$B_{S_i} = Nv_{S_i} / \mathrm{l} .$$

1.4 Measurements of study effectiveness

To report a candidate gene, the criterion is specified as a minimal frequency with which the candidate gene harbors variants that are detected (after filtering) in phenotype-exhibiting samples.

Let:

N be the total number of phenotype-exhibiting samples;

M be a user specified threshold.

The reporting criterion is therefore:

In ≥ M out of N phenotype-exhibiting samples, the candidate gene harbors detected variants (after filtering).

GIPS calculates four measurements of study effectiveness, which are detailed below.

1.4.1 Significance of each candidate gene to associate with the phenotype

The significance of a candidate gene to associate with the phenotype is calculated as the likelihood with which this candidate gene may accumulate random variants by chance in ≥ M phenotype-exhibiting samples.

Let:

G_j be the candidate gene;

L_{G_j} be the length of effective genomic region that has been annotated to G_j;

B_{S_i} be the BVF of sample S_i (which is obtained earlier);

PV_{G_j,S_i} be the probability of G_j harboring at least one random variant in sample S_i;

We have:

$$PV_{G_j,S_i} = 1 - (1 - B_{S_i})^{L_{G_j}}.$$

Let:

u, v be variables in the recursive function;

$R_{G_j,u}$ be the number of samples from the first u samples in which G_j harbors at least one random variant;

We have:

$$P\left(R_{G_j,u} = v\right) = \begin{cases} PV_{G_j,S_u} \times P\left(R_{G_j,u-1} = v - 1\right) + \left(1 - PV_{G_j,S_u}\right) \times P\left(R_{G_j,u-1} = v\right) & (0 < v < u) \\ \prod_{i=1}^{u} \left[1 - PV_{G_j,S_u}\right] & (v = 0) \\ \prod_{i=1}^{u} PV_{G_j,S_u} & (v = u) \end{cases}$$

The above recursive function enumerates all possible combinations of v samples from the first u

samples, and sums up the probability of each combination to have G_j harboring at least one random variant in each sample.

Therefore, the likelihood with which G_j accumulates random variants in ≥ M samples from all N samples can be computed as follows.

Let:

Q_{G_j} be the significance of gene G_j to associate with the phenotype;

We have:

$$Q_{G_j} = P\left(R_{G_j,N} \geq M\right) = \sum_{k=M}^{N} P\left(R_{G_j,N} = k\right).$$

1.4.2 Expected number of random genes that may be reported

The likelihood with which a gene can accumulate random variants in ≥ M phenotype-exhibiting samples, $P\left(R_{G_j,N} \geq M\right)$, is calculated in the previous section. Therefore, the number of random genes that are expected to meet the reporting criterion can be calculated as the sum of this likelihood for all genes.

Let:

$E(M,N)$ be the number of random genes that are expected to meet the reporting criterion;

We have:

$$E(M,N) = \sum_{j\in[all\ genes]} P\left(R_{G_j,N} \geq M\right) = \sum_{j\in[all\ genes]} Q_{G_j}.$$

1.4.3 Chance to report candidate genes that are truly associated with the phenotype

The chance that the true phenotype-associated gene will meet the candidate gene criterion is computed as follows.

Let:

u, v be variables in the recursive function;

D_{S_i} be the VDS of sample S_i (which is obtained earlier);

A_u be the number of samples from the first u samples in which the true phenotype-associated variants are detected;

We have:

$$P(A_u = v) = \begin{cases} \mathrm{D_{S_u}} \times P(A_{u-1} = v-1) + \left(1 - \mathrm{D_{S_u}}\right) \times P(A_{u-1} = v) & (0 < v < u) \\ \prod_{i=1}^{u} \left(1 - \mathrm{D_{S_u}}\right) & (v = 0) \\ \prod_{i=1}^{u} \mathrm{D_{S_u}} & (v = u) \end{cases}$$

The above recursive function enumerates all possible combinations of v samples from the first u samples, and sums up the probability of each combination to have the true phenotype-associated variants detected in each sample.

Let:

$H(M, N)$ be the chance of reporting the true phenotype-associated gene;

We have:

$H(M, N) = P(A_N \geq M) = \sum_{\mathrm{K=m}}^{\mathrm{N}} \mathrm{P}(\mathrm{A_N} = \mathrm{k})$.

1.4.4 Significance of violating Mendelian assumption

The significance of violating Mendelian assumption, if no gene is reported or all candidate genes have failed validation, is the likelihood with which true phenotype-associated variants are not detected in ≥ M phenotype-exhibiting samples.

Let:

w' be the number of samples in which the true phenotype-associated variants are detected. $w' < M$.

w be the maximum number of samples in which one gene harbors variants in all these samples, and, this gene has not been proven unrelated to the phenotype. $w' \leq w$.

$Z(w', N)$ be the significance of violating the Mendelian assumption.

A_N be the number of samples from all N samples in which the true phenotype-associated variants are detected. $P(A_N = k)$ is obtained earlier.

We have:

$Z(w', N) \leq Z(w, N) = P(A_N \leq w) = \sum_{k=0}^{w} P(A_N = k)$.

Because w' is difficult to obtain, we compute $Z(w, N)$ as the upper bound of $Z(w', N)$.

Chapter 2. Tool Usage

2.1 General workflow

To initiate a project, GIPS will create a main directory with four sub-folders. Using the "*Test*" project as example (detailed in 3.1), this directory structure is illustrated below (Figure 4). A text file "*PROJECT.ini*" will be placed in the main project folder, and another text file "*RESULT.gips*" will be placed in the "*Working*" sub-folder.

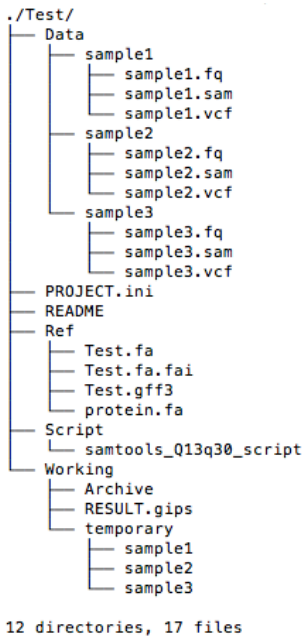


Figure 4. Structure of the GIPS test project directory.

"*PROJECT.ini*" specifies the analysis procedure that will be performed and evaluated by GIPS. A full analysis procedure includes steps to align sequencing reads to reference genome, to call variants from the aligned reads, to filter the observed variants, and finally, to call candidate genes.

Users need to create custom scripts to implement their desired approach to align sequencing reads and call variants. These scripts are recommended to be placed in the "/path/to/project_folder/Script" folder. In addition, it is recommended to use a naming convention for these scripts, which intuitively states the nature of these scripts, for example, "mapper-v1-20150502-q30-nr.sh". Note that GIPS will not directly invoke user scripts to align sequencing reads or call variants from the actual sequencing data. GIPS only uses these scripts to process simulated reads, when estimating the variant calling sensitivity `for each sample` (see 1.2.1 for details). Instead, users should use these scripts to call variants from the `actual` sequencing data, and supply the resultant VCF files to GIPS. This design is to ensure that the user-supplied scripts are all working. Furthermore, it is recommended to use a

naming convention for all intermediate result files, which unambiguously associates a result file to its producing script. For example, the alignment result of sample_1 produced by "mapper-v1-0515-q30-nr.sh" is named "sample_1-mapper-v1-0515-q30-nr.sam". The "/path/to/project_folder/*Data*" folder is designed to store intermediate result files.

To filter variants that are unlikely to associate with the phenotype, users can specify a filtering procedure with five types of configurable filters (see section 1.2. 2 for details). Finally, the candidate gene criterion is specified.

After an analysis procedure has been fully specified in "*PROJECT.ini*", a GIPS run will execute and evaluate this analysis procedure. Candidate genes will be reported and the four measurements of study effectiveness will be computed. The "*RESULT.gips*" file presents the GIPS results, together with the analysis protocol and a summary of intermediate results.

After the first GIPS run, users can optimize the analysis protocol and evaluate how these changes affect candidate genes and study effectiveness. Users can modify "*PROJECT.ini*" and run GIPS in the "update" mode (command line option "-update"). In the "update mode", GIPS will try to reuse previously produced intermediate results whenever possible. Meta data on previous intermediate results are coded and stored in the *"[GIPS TRACEBACK]*" section of the "*RESULT.gips*" file. In each GIPS run, the existing "*RESULT.gips*" file, which contains information of the previous GIPS run, will automatically be archived in the "/path/to/project_folder/Working/Archive" folder. If a user wishes to base further analysis optimization on an earlier analysis protocol (not the immediate previous protocol), he can locate the corresponding "*RESULT.gips*" file in the archive folder and restore this file to "/path/to/project_folder/Working/*RESULT.gips*". This will allow GIPS to reuse related intermediate results in its "update" mode.

2.2 Parameter and result file

2.2.1 Project initialization

Usage: `Java –jar –Xms3g GIPS.jar -init /path/to/project_folder`

GIPS will create the directory "*/path/to/project_folder*". This directory contains four sub-folders:

"*Scripts*", for storing user scripts.

"*Working*", for storing intermediate results and archiving GIPS result files.

"*Ref*", for storing reference data, such as genome annotation files and phenotype-causing variant library.

"*Data*", for storing the user-provided sequence alignment (SAM) files and variant call (VCF)

files for the sequenced samples.

The "*project_folder*" also contains two text files:

"*PROJECT.ini*", which specifies the analysis procedure to be evaluated (detailed below);

"*Working/RESULT.gips*", which presents GIPS results (detailed blow).

2.2.2 Analysis procedure specification ("*PROJECT.ini*")

The "*PROJECT.ini*" file includes one "*[GLOBAL]*" section, which is followed by multiple "*[SAMPLE]*" sections. There is also a section named "*[SAMPLE_LIST]*", which lists all sample names. If the same parameter appears in both the "*[GLOBAL]*" section and a "*[SAMPLE]*" section, the "*[SAMPLE]*" specification overrides the "*[GLOBAL]*" specification, which applies only to that `specific` sample. "#" is used to add comments.

The "*[GLOBAL]*" section contains the following specifications.

PROJECT: Project name, which is automatically filled when a project is initiated.

SNPEFF_GENOME_VERSION: The genome version code used by SnpEff to annotate variants. This is a required parameter with no default value.

REF_GENOME_ANNOTATION.GFF: /path/to/genome_annotation.gff. The genome annotation file that GIPS uses to select the "effective genomic region". This is a required parameter with no default value. If only a file name is specified, GIPS looks for this file in "/path/to/project_folder/Ref/".

SNPEFF: /path/to/program/SnpEff/folder. Note that this is the path to SnpEff program folder, not the path to the SnpEff executable (jar file). This is a required parameter with no default value. SnpEff can be downloaded at http://snpeff.sourceforge.net/index.html.

VAR_CALL_SCRIPT: /path/to/variant_calling_script. GIPS invokes user-supplied scripts to call variants from simulated sequencing reads. User-supplied variant calling scripts need to take two command line parameters. The first one ("$1" in common Linux shell script) specifies a path to the input SAM file, and the second one ("$2" in common Linux shell script) specifies a path to the output VCF file. This is a required parameter with no default value. If only a file name is specified, GIPS looks for this script in "/path/to/project_folder/Script/".

CANDIDATE_CRITERION: The minimal number of phenotype-exhibiting samples in which a candidate gene is expected to harbor variants. Default to the total number of samples (which reports only genes that harbor variants in all samples).

EFF_REGION: Combination of the following regions, separated by "|": promoter=length, CDS,

5UTR, 3UTR, SpliceSite=length. Example: "*EFF_REGION=CDS|PROMOTER=1000*". Defaults to "CDS|SpliceSite=2|PROMOTER=1000". This parameter is case-insensitive.

VAR_FILTERS: Variant filters to be used. "A": Ancestry filter. "B": Big difference filter. "E": Effective region filter. "C": Congestion filter. Example: "*VAR_FILTERS*: EB". Defaults to "*EBA*". Note that the "control filter" will take effect if the "*CONTROL*" parameter is specified. The effective region filter ("E") will always take effect. See 1.2.2 for details.

SCORE_MATRIX: The scoring matrix used for filtering non-synonymous SNPs. GIPS has 87 built-in scoring matrices, which are listed in "*/path/to/GIPS_folder/config*". Defaults to "DEFAULT", which means that no scoring matrix is used and all non-synonymous variants are kept.

MAX_AA_SCORE: The maximum amino acid similarity score as in the user specified scoring matrix for a non-synonymous SNP to be kept by the big difference filter. Its default value is "0".

NUM_SIM_SNPS: Number of artificial SNP to be simulated when estimating the variant calling sensitivity for a sample. Defaults to 5,000.

CONTROL: /path/to/control_sample.vcf. A single file containing all variants that are found in phenotype-absent control samples. If there are multiple control samples, please merge their observed variants into one VCF file.

After the "*[GLOBAL]*" section, in each "*[SAMPLE]*" section, the following parameters are specified.

SAMPLE_NAME: Name of this sample. This is a required parameter with no default value.

READS_ALIGNMENT.SAM: /path/to/sample.sam. The reads alignment result of this sample. Note that the user-supplied variant calling script should be able to take this file as input and produce a variant calling result (which is specified in the "*SAMPLE.VCF*" parameter). This parameter is required if "*SPECIFY_HOMO_VDS*", "*SPECIFY_HETERO_VDS*", and "*SPECIFY_BVF*" are not used. If only a file name is specified, GIPS looks for this file in "/path/to/project_folder/Data/".

SAMPLE.VCF: /path/to/sample.vcf. The variant calling result of this sample. Note that the user-supplied variant calling script should be able to produce this result file from the reads alignment file specified in the "*READS_ALIGNMENT.SAM*" parameter. This parameter is required if "*SPECIFY_HOMO_VDS*", "*SPECIFY_HETERO_VDS*", and "*SPECIFY_BVF*"

are not used. If only a file name is specified, GIPS looks for this file in "/path/to/project_folder/Data/".

SPECIFY_HOMO_VDS: Do not estimate the variant detection sensitivity for homozygous variants from the real sequencing data. Use the specified value instead. This parameter is required if "*READS_ALIGNMENT.SAM*" and "*SAMPLE.VCF"* are not used.

SPECIFY_HETERO_VDS: Do not estimate the variant detection sensitivity for heterozygous variants from the real sequencing data. Use the specified value instead. This parameter is required if "*READS_ALIGNMENT.SAM*" and "*SAMPLE.VCF"* are not used.

SPECIFY_BVF: Do not estimate the background variant frequency from the real sequencing data. Use the specified value instead. This parameter is required if "*READS_ALIGNMENT.SAM*" and "*SAMPLE.VCF"* are not used.

VAR_CALL_SCRIPT: Overrides the same parameter in the"*[GLOBAL]*" section, if specified.

SCORE_MATRIX: Overrides the same parameter in the"*[GLOBAL]*" section, if specified.

MAX_AA_SCORE: Overrides the same parameter in the"*[GLOBAL]*" section, if specified.

CONTROL: Overrides the same parameter in the"*[GLOBAL]*" section, if specified.

NUM_SIM_SNPS: Overrides the same parameter in the"*[GLOBAL]*" section, if specified.

2.2.3 Result file（*"RESULT.gips"*）

"*RESULT.gips*" has five sections, which may not all appear. Whether a section will appear depends on the type of GIPS run.

The "*[Study Effectiveness]*" section shows the chance of reporting the true phenotype-associated gene with the current analysis protocol, and, the significance of violating Mendelian assumption, if no gene is reported or all candidate genes have failed validation.

The "*[Candidate Gene List]*" section shows all candidate genes. Detailed information about the variants in these candidate genes is given in the result file "*Working/CANDIDATE_GENES.txt*".

The "*[Sample Specific Sensitivity and Specificity]*" section shows, for each sample, the variant calling sensitivity (section 1.2.1), false ignorance rate (section 1.2.2), variant detection sensitivity (section 1.2), and background variant frequency (section 1.3).

The "*[Protocol]*" section records the analysis protocol specified in "*PROJECT.ini*".

The "*[GIPS TRACEBACK]*" section encodes the meta-data of intermediate results produced in this GIPS run. These intermediate results can be used to speed up future GIPS runs in the "update"

mode. Users should not modify information in this section.

2.2.4 Running GIPS

Usage: `java -jar GIPS.jar [options]`

Example: `java –Xms3g -jar GIPS.jar -T <tool> -p /path/to/project_folder`

Table 1. GIPS software options

Options:		
-h (-H)		Show help.
-Test		Initiate a new project with the test setup. See the next chapter for the "Test" example.
-init	/path/to/project_folder	Initiate a new project.
-T	<gips|vcs|filter>	Select GIPS function. gips: full workflow; vcs: only estimate the variant calling sensitivity for each sample; filter: only use variant filters to remove likely unrelated variants. Defaults to "gips".
-update		Run GIPS in "update" mode. GIPS will try to reuse intermediate results produced in previous runs.

Chapter 3. Examples

3.1 Test example

This example tests GIPS with synthesized artificial data.

Step 1: Initiate the test project.

Command: `java -jar GIPS.jar –Test`

The above command creates the GIPS project folder, "*Test*". Its directory structure is as follows.

```
./Test/
├── Data
│   ├── sample1
│   │   ├── sample1.fq
│   │   ├── sample1.sam
│   │   └── sample1.vcf
│   ├── sample2
│   │   ├── sample2.fq
│   │   ├── sample2.sam
│   │   └── sample2.vcf
│   └── sample3
│       ├── sample3.fq
│       ├── sample3.sam
│       └── sample3.vcf
├── PROJECT.ini
├── README
├── Ref
│   ├── Test.fa
│   ├── Test.fa.fai
│   ├── Test.gff3
│   └── protein.fa
├── Script
│   └── samtools_Q13q30_script
└── Working
    ├── Archive
    ├── RESULT.gips
    └── temporary
        ├── sample1
        ├── sample2
        └── sample3

12 directories, 17 files
```

In "*Test/Data*", there are three directories, "*sample1*", "*sample2*" and "*sample3*". These directories store the synthesized sequencing reads alignment results and variant calling results from three hypothetical samples. These artificial sequencing reads are generated by the ART simulator (Huang, Li et al. 2012).

In "*Test/Script*", there is a variant calling script, which invokes SAMTools to call variants from mapped sequencing reads. Please modify this script accordingly to fit the user computing system. This script was used to call variants from the mapped reads of each hypothetical sample, and will be invoked by GIPS to estimate variant calling sensitivity (section 1.2.1). This script takes two command line parameters. The first one specifies a path to the input SAM file, and the second one specifies a

path to the output VCF file.

In "*project_folder*/*Ref*", there are two files. "*Test.fa*" provides the reference genome sequence of the hypothetical samples. "*Test.gff3*" annotates the hypothetical reference genome.

The "Working" directory has stored a sample GIPS result file, "*RESULT.gips*", and some intermediate results. It is recommended that users also follow this directory convention to store intermediate results when developing their custom scripts to work with GIPS.

The parameter file "*PROJECT.ini*" specifies the analysis procedure to be performed and evaluated by GIPS (section 2.2.2).

Step 2: Setup SAMTools and SnpEff.

SAMTools can be downloaded at http://samtools.sourceforge.net.

Please make sure that "*samtools*", "bcftools", "vcfutils.pl" are all placed in the executable search path of the user system ($PATH). Otherwise, please modify "*Test/Script/samtools_Q13q30_script*" accordingly.

SnpEff can be downloaded at http://snpeff.sourceforge.net/download.html.

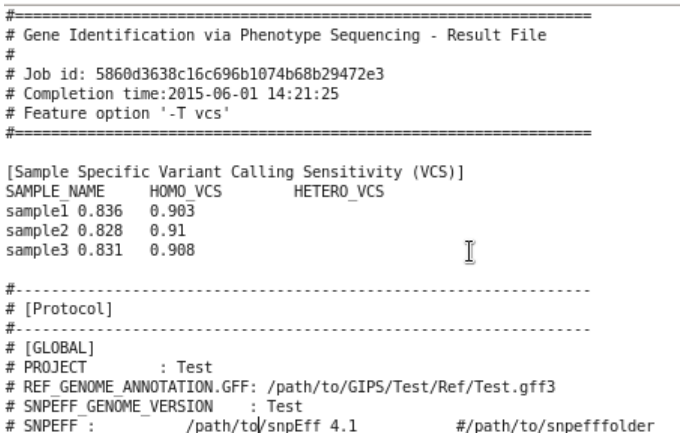
Unzip the package and set its folder path to the "*SNPEFF*" parameter in the "*[GLOBAL]*" section of the "*PROJECT.ini*" parameter file.

Step 3: Run GIPS to estimate variant calling sensitivity.

Command: `java -Xms3G -jar GIPS.jar -p Test -T vcs`

This command tests the GIPS function of estimating sample-specific variant calling sensitivities. The script "*Test/Script/samtools_Q13q30_script*" will be invoked to call variants from simulated reads alignment result, and GIPS will compute how much proportion of the simulated SNPs is recovered with this script (section 1.2.1).

The GIPS result file, "*Test/Working/RESULT.gips*", will show contents like below.



```
#=============================================================
# Gene Identification via Phenotype Sequencing - Result File
#
# Job id: 5860d3638c16c696b1074b68b29472e3
# Completion time:2015-06-01 14:21:25
# Feature option '-T vcs'
#=============================================================

[Sample Specific Variant Calling Sensitivity (VCS)]
SAMPLE_NAME     HOMO_VCS        HETERO_VCS
sample1 0.836   0.903
sample2 0.828   0.91
sample3 0.831   0.908

#-------------------------------------------------------------
# [Protocol]
#-------------------------------------------------------------
# [GLOBAL]
# PROJECT        : Test
# REF_GENOME_ANNOTATION.GFF: /path/to/GIPS/Test/Ref/Test.gff3
# SNPEFF_GENOME_VERSION    : Test
# SNPEFF :          /path/to/snpEff 4.1          #/path/to/snpefffolder
```

Step 4: Run GIPS to filter variants.

Command: `java –Xms3G -jar GIPS.jar –T filter –p Test`

This command tests the GIPS function to filter variants that are unlikely to associate with the target phenotype. The filtering procedure is specified in "*Test/PROJECT.ini*". If successful, users will find three result files in "*Test/Working*", with the names "sample1.filter.gips", "sample2.filter.gips", and "sample3.filter.gips".

Step 5: Run a full GIPS workflow.

Command: `java –Xms3G -GIPS.jar –p Test –T gips`

This command tests the full GIPS workflow to call candidate genes from sequencing data. The hypothetical gene "*gene1_gips*" is expected to appear in the "*[Candidate Gene List]*" section of the GIPS result file "*Test/Working/RESULT.gips*". The GIPS result file will show contents like below.

```
#============================================================
# Gene Identification via Phenotype Sequencing - Result File
#
# Job id: b0c43909cb687db82c3ec075cfe8a2a3
# Completion time:2015-06-01 14:20:49
# Feature option '-T gips'
#============================================================

[Study Effectiveness]
Chance to report true phenotype-associated genes
If phenotype is recessive: 0.549
If phenotype is  dominant: 0.738

Significance of violating Mendelian phenotype assumption
If phenotype is recessive: 0.45
If phenotype is  dominant: 1.0

Expected number of random genes that may be reported: 0.956

[Candidate Gene List]
#Details on the variants they harbor can be found in CANDIDATE_GENES.txt
GENE_ID GENE_NAME       Chr     EFF_REGION_LENGHT       SIGNIFICANCE    NUM_VARIANT     NUM_SAMPLE
gene1   gene1_gips      chrI    315     0.055805        11      3

[Sample Specific Sensitivity and Specificity]
Variant Calling Sensitivity (VCS)
SAMPLE_NAME     HOMO_VCS        HETERO_VCS
sample1 0.833   0.928
sample2 0.851   0.92
sample3 0.822   0.918

False Ignorance Rate (FIR)
SAMPLE_NAME     FIR_FUNCTION    FIR_STUDY       FIR
sample1 0.0184  0.0     0.0184
sample2 0.0184  7.094E-4        0.0191
sample3 0.0184  7.094E-4        0.0191

Sample Variant Detection Sensitivity (VDS) (Measurement of sensitivity)
SAMPLE_NAME     HOMO_VDS        HETERO_VDS
sample1 0.817   0.91
sample2 0.834   0.902
sample3 0.806   0.9
```

In addition, a text file "*Test/Working/CANDIDATE_GENES.txt*" will be created, which provides detailed information regarding the variants detected in "*gene1_gips*" in each sample.

3.2 Identification of a Pho2 suppressor gene in Oryza Sativa

3.2.1 Overview

Inorganic phosphorus (Pi) is an essential nutrient for plant growth and development, and Pi limitation is widely a constraint for crop yield in cultivated soils (Raghothama 1999). Understanding the molecular mechanisms of Pi uptake and utilization is important for breeding practices to improve nutrient efficiency in crops. The mutation of *PHOSPHATE2* (*PHO2*) was first described in Arabidopsis. The phenotype is over-accumulation of Pi in shoot tissues (Delhaize and Randall, 1995). Arabidopsis PHO2 was later characterized as an ubiquitin-conjugating E2 enzyme (Liu, Huang et al. 2012). OsPHO2, the PHO2 homolog in rice (LOC_Os05g48390), was also identified as an important regulator in phosphate translocation and homeostasis, which showed similar functions like AtPHO2. *Ospho2* mutant shows leaf tip necrosis and Pi accumulation largely in mature leaves (Wang, Ying et al. 2009, Hu, Zhu et al. 2011). In this study, we performed a forward genetic study to identify its potential suppressor genes in rice. A *pho2* Tos17 insertion mutant was obtained from the Rice Genome

Resource Center, Japan (http://tos.nias.affrc.go.jp/, Accession: NE8536). We derived a homozygous *pho2* mutant (HNE8536) from this line (Wang, Ying et al. 2009).

An EMS (ethyl methanesulfonate) induced mutant library was generated from the HNE8536 homozygous *pho2* mutant line. From the M2 population of ~15,000 lines grown in soil, three partial suppressor mutants showing an identical phenotype of Pi tolerance were obtained (M28, M29 and M249, Figure 5 (1A)). Because these mutants are produced by EMS treatment, they are considered unrelated and the ancestry filter (section 1.2.2.3) is applicable. The raw sequencing results are retrievable from the Sequence Read Archive Database (Accession: SRP058881).

3.2.2 Run with empirical parameters

Step 1: Download raw sequencisng data.

Table 2. Pho2 suppressor dataset.

SAMPLE NAME	SAMPLE ID	NCBI SRA database link
M28	SRS949736	http://www.ncbi.nlm.nih.gov/sra/?term=SRS949736
M29	SRS949738	http://www.ncbi.nlm.nih.gov/sra/?term=SRS949738
M249	SRS949741	http://www.ncbi.nlm.nih.gov/sra/?term=SRS949741

Sequencing reads files are in SRA format and need to be converted to FASTQ format using SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software):

```
fastq-dump.2 --split-3 /Path/to/sample.sra
```

Step2: Align sequencing reads to reference genome and call variants.

Paired-end sequencing reads are aligned to the Oryza Sativa Nipponbare reference genome version 7 (Kawahara, de la Bastide et al. 2013), with using (Langmead and Salzberg 2012). Suppose that the raw sequencing reads are store in "/path/to/pho2/dataset/".

The SAMTools package version 1.2 (Li, Handsaker et al. 2009) is used to call variants from the aligned reads. PCR duplicates are marked and removed by Picard. The core commands are:

```
/path/to/bowtie2 -p 6 -x rice7 -1 /path/to/pho2/dataset/sample_1.fastq -2 /path/to/
pho2/dataset/sample_2.fastq -S /path/to/pho2/dataset/sample.sam

java -jar /path/to/picard.jar AddOrReplaceReadGroups INPUT=/path/to/pho2/dataset/
sample.sam OUTPUT=/path/to/pho2/dataset/sample.bam SORT_ORDER=coordinate RGPL=illumina
RGLB="GIPS" RGPU=temp RGSM="GIPS"

java -Xms8g -jar /path/to/picard.jar MarkDuplicates I=/path/to/pho2/dataset/sample.bam
O=/path/to/pho2/dataset/sample_markdup.bam M=/path/to/pho2/dataset/duplication_metrics
REMOVE_DUPLICATES=true AS=true

/path/to/samtools index /path/to/pho2/dataset/sample_markdup.bam

/path/to/smalls index /path/to/pho2/dataset/sample_markdup.bam

/path/to/samtools mpileup -ugf /path/to/ref.fa /path/to/pho2/dataset/
sample_markdup.bam -go -| /path/to/bcftools call -vm0 z -o - | /path/to/bcftools
filter -O v -o /path/to/pho2/dataset/samtools_Q13_q0_noDPfilter.vcf
```

Step 3: Set up the GIPS project.

First initiate the GIPS project with the following command:

```
java -jar GIPS.jar -p GIPS-Rice
```

Then modify "*PROJECT.ini*":

Set "*REF_GENOME_ANNOTATION.GFF*", "*SNPEFF*", and "*SNPEFF_GENOME_VERSION*" appropriately. In our system, these parameters are set to:

```
[GLOBAL]
PROJECT :GIPS-Rice
REF_GENOME_ANNOTATION.GFF : /nas1/zzx/pho2/Ref/all(rice).gff3
SNPEFF_GENOME_VERSION     :  rice7
SNPEFF  :    /nas1/zzx/pho2/software/snpEff_4.1
```

Move the reads alignment results and variant calling results obtained in Step 2 into the GIPS project folder, "*/path/to/GIPS-Rice/Data*". Write a variant calling script according to GIPS requirement (case shown in below) in "*/path/to/GIPS-Rice/Script*". This script is named "*2015-04-26-default-Q13-q0.sh*" in the project. There are two command line parameters in this script. The first one ("$1" in common Linux shell script) specifies a path to the input SAM file, and the second one ("$2" in common Linux shell script) specifies a path to the output VCF file. Move the variant calling script to "*/path/to/rice/Script*" and set its path in "*PROJECT.ini*".

```
#! /bin/bash
# $1 sam $2 vcf. Intermediate files are recommended to store in /path/to/Project/Work/
intermediate directory. Samtools version: 1.2

#sort and add read group
java -Xms8g -jar /path/to/picard.jar AddOrReplaceReadGroups INPUT="$1" OUTPUT=/path/
to/GIPS-Rice/Working/temporary/btout.bam SORT_ORDER=coordinate RGPL=illumina
RGLB="GIPS" RGPU=temp RGSM="GIPS"

# remove duplicate
java -Xms8g -jar /path/to/picard.jar MarkDuplicates I=/path/to/GIPS-Rice/Working/
temporary/btout.bam O=/path/to/GIPS-Rice/Working/temporary/bt_markdup.bam M=/path/to/
GIPS-Rice/Working/temporary/duplication_metrics REMOVE_DUPLICATES=true AS=true

# index
/path/to/samtools index /path/to/GIPS-Rice/Working/temporary/bt_markdup.bam

# samtool call
/path/to/samtools mpileup -ugf /path/to/all.chrs.con.fa /path/to/GIPS-Rice/Working/
temporary/bt_markdup.bam -go -| /path/to/bcftools call -vmO z  -o - | /path/to/
bcftools filter -O v -o "$2"
```

After the above parameter specifications, the parameter file "*PROJECT.ini*" should look like below.

```
[GLOBAL]
PROJECT :GIPS-Rice
REF_GENOME_ANNOTATION.GFF : /path/to/GIPS-Rice/Ref/all(rice).gff3
SNPEFF_GENOME_VERSION     :  rice7
SNPEFF  :    /path/to/snpEff_4.1
CANDIDATE_CRITERIA :  3
VAR_CALL_SCRIPT :   /path/to/GIPS-Rice/Script/2015-04-26-default-Q13-q0.sh
EFF_REGION  : CDS|splicesite=2|Promoter=1000
VAR_FILTERS :   EAB
SCORE_MATRIX: DEFAULT
MAX_AA_SCORE:
NUM_SIM_SNPS : 10000
MAX_VAR_DENSITY:

[SAMPLE_LIST]
M28
M29
M249

[SAMPLE]
SAMPLE_NAME : M28
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M28/samtools_Q13_q0_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M28/btout.sam

[SAMPLE]
SAMPLE_NAME : M29
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M29/samtools_Q13_q0_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /nas1/zzx/RiceData/out/M29/btout.sam


[SAMPLE]
SAMPLE_NAME : M249
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M249/samtools_Q13_q0_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M249/btout.sam
```

Step 4: Call candidate genes.

Command: `Java -Xms3G -jar GIPS.jar -T gips -p GIPS-Rice`

This command will call candidate genes with the default analysis procedure. The default analysis procedure focuses on coding sequences and splice sites (effective genomic region), and uses ancestry filter (section 1.2.2.3) and big difference filter (section 1.2.2.2). The default behavior of big difference filter is to remove only synonymous SNPs. Candidate genes are expected to harbor detected variants in

all three samples.

As shown in "*/path/to/GIPS-Rice/Working/RESULT.gips*", the default analysis procedure has ~91% chance to detect homozygous variant and ~80% chance to detect heterozygous variant in each sample. The combined likelihood of the phenotype-associated gene harboring detected variants in all three samples is 74%. These results are summarized in Table 3.

Table 3. Sample-wise study effectiveness measurements of the default analysis procedure. Homo_VCS: homozygous variant calling sensitivity; Hetero_VCS: heterozygous variant calling sensitivity; FIR: false ignorance rate; Homo_VDS: homozygous variant detection sensitivity; Hetero_VDS: Heterozygous variant detection sensitivity; BVF: background variant frequency.

Sample	Homo_VCS	Hetero_VCS	FIR	Homo_VDS	Hetero_VDS	BVF
M28	0.94	0.839	0.0152	0.928	0.826	1.73E-5
M29	0.886	0.801	0.0152	0.872	0.788	9.97E-6
M249	0.952	0.832	0.0152	0.937	0.819	9.20E-6

The default analysis procedure identified 30 candidate genes, which are shown below. Detailed information on the variants that these candidate genes harbor is provided in the result file "*/path/to/GIPS-Rice/Working/CANDIDATE_GENES.txt*".

[Candidate Gene List]
Detailed information on the variants that these candidate genes harbor is provided in CANDIDATE_GENES.txt

GENE_ID	GENE_NAME	Chr	EFF_REGION_LENGHT	SIGNIFICANCE	NUM_VARIANT	NUM_SAMPLE
LOC_Os12g35520	LOC_Os12g35520	Chr12	1224	2.84938E-6	4	3
LOC_Os12g26100	LOC_Os12g26100	Chr12	1272	3.19512E-6	18	3
LOC_Os03g64310	LOC_Os03g64310	Chr3	1312	3.50358E-6	6	3
LOC_Os06g39752	LOC_Os06g39752	Chr6	1402	4.26818E-6	3	3
LOC_Os04g02550	LOC_Os04g02550	Chr4	1446	4.67904E-6	17	3
LOC_Os08g05950	LOC_Os08g05950	Chr8	1470	4.91376E-6	7	3
LOC_Os04g08690	LOC_Os04g08690	Chr4	1489	5.105E-6	4	3
LOC_Os10g31380	LOC_Os10g31380	Chr10	1542	5.6643E-6	4	3
LOC_Os06g39744	LOC_Os06g39744	Chr6	1558	5.84076E-6	5	3
LOC_Os10g07040	LOC_Os10g07040	Chr10	2201	1.62761E-5	30	3
LOC_Os04g35000	LOC_Os04g35000	Chr4	2571	2.5768E-5	6	3
LOC_Os11g12710	LOC_Os11g12710	Chr11	2717	3.03315E-5	4	3
LOC_Os06g38270	LOC_Os06g38270	Chr6	3166	4.76017E-5	23	3
LOC_Os04g51840	LOC_Os04g51840	Chr4	3180	4.82237E-5	7	3
LOC_Os10g15210	LOC_Os10g15210	Chr10	3194	4.88511E-5	11	3
LOC_Os10g13550	LOC_Os10g13550	Chr10	3447	6.11227E-5	3	3
LOC_Os02g56510	LOC_Os02g56510	Chr2	3504	6.41391E-5	4	3
LOC_Os08g06280	LOC_Os08g06280	Chr8	3529	6.54921E-5	10	3
LOC_Os07g30650	LOC_Os07g30650	Chr7	3609	6.99465E-5	6	3
LOC_Os11g17900	LOC_Os11g17900	Chr11	3846	8.42889E-5	3	3
LOC_Os04g02520	LOC_Os04g02520	Chr4	4484	1.32047E-4	19	3
LOC_Os05g51770	LOC_Os05g51770	Chr5	4780	1.59108E-4	6	3
LOC_Os03g64320	LOC_Os03g64320	Chr3	4819	1.6292E-4	6	3
LOC_Os03g23760	LOC_Os03g23760	Chr3	4838	1.64798E-4	5	3
LOC_Os02g32690	LOC_Os02g32690	Chr2	5330	2.18414E-4	14	3
LOC_Os11g22430	LOC_Os11g22430	Chr11	5330	2.18414E-4	6	3
LOC_Os04g49890	LOC_Os04g49890	Chr4	5346	2.20324E-4	6	3
LOC_Os05g06280	LOC_Os05g06280	Chr5	5489	2.37867E-4	3	3
LOC_Os04g13510	LOC_Os04g13510	Chr4	5578	2.49226E-4	22	3
LOC_Os04g20790	LOC_Os04g20790	Chr4	5781	2.76426E-4	5	3

3.2.3 Run with optimized parameters

The default analysis procedure identified 30 candidate genes, which are too many for further experimental validation. The identification of such many candidates indicates the lack of specificity of the default analysis procedure. Therefore, we next modify the default analysis procedure to increase its specificity.

Step 5: Modify the analysis procedure.

1) Use a more strict quality criterion in variant calling.

We increase the minimum base quality score from 13 to 15, and increase the minimum mapping quality score from 0 to 20. The updated variant calling script looks as follows.

```
/path/to/bowtie2 -p 6 -x rice7 -1 /path/to/pho2/dataset/sample_1.fastq -2 /path/to/pho2/
dataset/sample_2.fastq -S /path/to/pho2/dataset/sample.sam

java -jar /path/to/picard.jar AddOrReplaceReadGroups INPUT=/path/to/pho2/dataset/
sample.sam OUTPUT=/path/to/pho2/dataset/sample.bam SORT_ORDER=coordinate RGPL=illumina
RGLB="GIPS" RGPU=temp RGSM="GIPS"

java -Xms8g -jar /path/to/picard.jar MarkDuplicates I=/path/to/pho2/dataset/sample.bam O=/
path/to/pho2/dataset/sample_markdup.bam M=/path/to/pho2/dataset/duplication_metrics
REMOVE_DUPLICATES=true AS=true

/path/to/samtools index /path/to/pho2/dataset/sample_markdup.bam

/path/to/smalls index /path/to/pho2/dataset/sample_markdup.bam

/path/to/samtools mpileup -ugf /path/to/ref.fa /path/to/pho2/dataset/sample_markdup.bam -
go -| /path/to/bcftools call -vm0 z -o - | /path/to/bcftools filter -Q 15 -q 20 0 v -o /
path/to/pho2/dataset/samtools_Q15_q20_noDPfilter.vcf
```

The updated script is then used to call variants from the actual sequencing reads. The resultant VCF files (i.e., "*/path/to/pho2/Dataset/M28/samtools_Q15_q20_noDPfilter.vcf*") are moved to the "*/path/to/GIPS-Rice/Data*" folder. Meanwhile, create a script with GIPS requirement in "*/path/to/GIPS-Rice/Script*" with updated parameters. This script is named "*samtools_Q15_q20.sh*". Set new script path and new variant files to "*VAR_CALL_SCRIPT*".

The "*PROJECT.ini*" file is updated to reflect these changes.

```
[SAMPLE]
SAMPLE_NAME : M28
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M28/samtools_Q15_q20_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M28/btout.sam

[SAMPLE]
SAMPLE_NAME : M29
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M29/samtools_Q15_q20_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /nas1/zzx/RiceData/out/M29/btout.sam


[SAMPLE]
SAMPLE_NAME : M249
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M249/samtools_Q15_q20_noDPFiltered.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M249/btout.sam
```

2) Use a more strict "big difference filter".

In the "*[GLOBAL]*" section of "*PROJECT.ini*", set:

SCORE_MATRIX: PAM120

These changes instruct the big difference filter to remove non-synonymous variants that result in similar amino acid changes with positive similarity scores in the PAM120 scoring matrix.

3) Shrink the effective genomic region to include only the CDS regions.

It is arguable that most phenotype-related *de novo* mutations happen in the CDS regions. Here we take an aggressive approach to focus only on the CDS regions, with the intention to reduce the number of detected variants that may confound gene identification.

In the "*[GLOBAL]*" section of "*PROJECT.ini*", set:

EFF_REGION: CDS|SpliceSite=2

Step 6: Run GIPS in "update" mode to evaluate the modified analysis procedure.

Command: `java -xms3G -jar GIPS.jar -T gips -p GIPS-Rice -update`

This command will call candidate genes with the modified analysis procedure.

As shown in "*/path/to/GIPS-Rice/Working/RESULT.gips*", the modified analysis procedure has ~70% chance to detect homozygous variant and ~67% chance to detect heterozygous variant in each sample. The combined likelihood of the phenotype-associated gene harboring detected variants in all three samples is 29%. These results are summarized in Table 4.

Table 4. Sample-wise study effectiveness measurements of the modified analysis procedure. Homo_VCS: homozygous variant calling sensitivity; Hetero_VCS: heterozygous variant calling sensitivity; FIR: false ignorance rate; Homo_VDS: homozygous variant detection sensitivity; Hetero_VDS: Heterozygous variant detection sensitivity; BVF: background variant frequency.

Sample	Homo_VCS	Hetero_VCS	FIR	Homo_VDS	Hetero_VDS	BVF
M28	0.913	0.881	0.197	0.732	0.707	5.85E-6

M29	0.808	0.766	0.197	0.648	0.614	2.58E-6
M249	0.907	0.85	0.197	0.727	0.682	2.67E-6

If compared to the default analysis procedure, the modified analysis procedure has less chance to detect the true phenotype-associated variant in each sample. This results in a decrease of the likelihood that the true phenotype-related gene harbors detected variants in all three samples. On the other hand, the modified analysis procedure detects much less number of variants in each sample, only ~xx% of those detected with the default analysis procedure. This decrease of confounding variants helps the identification of phenotype-associated gene with increased confidence. As a result, the top candidate significance has increased from xxx to 8.635×10^{-7} using the modified analysis procedure.

The modified analysis procedure reports only two candidate genes, *LOC_Os02g56510* and *LOC_Os04g13510*. The second reported candidate, *LOC_Os04g13510*, is a retrotransposon protein, which tends to accumulate SNPs and is unlikely to perform direct physiological function.

The top candidate *LOC_Os02g56510* (OsPHO1;2) is a functional homology of the Arabidopsis AtPHO1 gene (Ref). AtPHO1 was proposed to function as Pi transporter (Ref). A recent study showed that AtPHO1 is a crucial downstream component of AtPHO2. AtPHO2 modulates the degradation of AtPHO1 in endomembranes to maintain Pi homeostasis in Arabidopsis (Ref). Therefore, it is highly likely that this is the gene that partially suppressed the *pho2* mutant phenotype. The variants detected in OsPHO1;2 are listed in Table 5.

Table 5. Variants detected in gene *OsPHO1;2*.

Sample	Pos	Ref	Alt	GT	Effect
M28	34611907	C	T	1/1	Stop_gained
M29	34614218	G	A	1/1	Missense
M249	3464218	C	T	1/1	Missense

To perform further validation, we sequenced the OsPHO1;2 gene in another EMS mutant (M358) obtained through another round of library screening. M358 exhibits the same phenotype as M28, M29, and M249. Results show that M358 has a non-synonymous mutation in *OsPHO1;2*, which is potentially functional, and is different from those observed in M28, M29, and M249. The phenotypes and genotypes of M28, M29, M249, and M358 are shown in Figure 5.

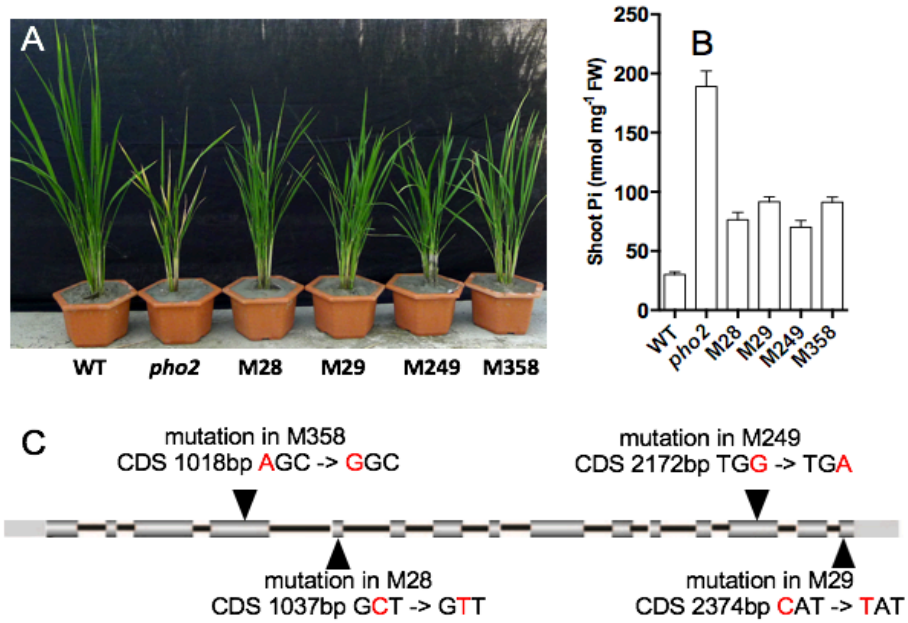


Figure 5. Phenotype, Pi Concentrations and mutations in pho2 supressor mutants. (A) and (B) The phenotype of 60-d-old plants in soil(A, up panel) and the Pi concentrations of the shoot (B) of 40-d-old wild-type (WT), Ospho2 and supressor mutants grown in +Pi (200 μM) hydroponic medium (Error bars represent SE ,n = 3 in [B]). Lower panel in (A) the PCR results of the pho2 Tos17 insertion. F+R, PCR results using gene specific primers across the Tos17 insertion site. T+R, PCR results using gene specific primer combining with primer for Tos17 sequences. Primers were described by Wang et al. (2009). (C) The gene structure of OsPHO1;2. Mutations in M28, M29, M249, M358 were indicated with arrowheads respectively.

3.2.4 Example discussion (strategy of analysis procedure optimization)

This example provides case for discussing how an analysis of phenotype-sequencing dataset can be optimized based on the study effective measurements computed by GIPS, in particular, “the chance to report candidate genes that are truly associated with the phenotype” and “the significance of violating Mendelian assumption, if no gene is reported or all candidate genes have failed validation”.

In a phenotype-sequencing experiment, it is always of the utmost importance to identify the phenotype-associated gene. To this purpose, a more “strict” or “accurate” analysis procedure will only consider highly confident variants that are supported by a significant number of high quality sequencing reads. Subsequent filtering process will also keep only variants that are very likely to produce the target phenotype. This “accurate” procedure will result in less number of possible variants that may produce the phenotype in each sample. Consequently, a candidate gene harboring variants in multiple samples will have a higher significance to associate with the phenotype. However, such an “accurate” procedure will also risk ignoring true phenotype-causing variants that are not supported by unequivocal evidence. This undesired ignorance can result in the true phenotype-associated gene harboring variants in an insufficient number of samples to meet the candidate gene criterion. On the

other hand, a more "permissive" or "comprehensive" analysis procedure will include more false positive variants and more less-likely phenotype-causing variants in each sample, which will result in a higher chance that a random gene harbors variants in multiple samples. As a result, when the candidate gene criterion is met, the significance of a candidate to associate with the phenotype is also lower. Although a comprehensive analysis procedure is unlikely to render the true phenotype-associated gene failing the candidate gene criterion, such a procedure is also more likely to report phenotype-unrelated genes, which demands significant effort in their validation.

Therefore, in most investigations, it is recommended to use a more accurate analysis procedure whenever possible. Accurate procedures produce highly confident candidate genes, which minimize the chance of failure in further experimental validations. GIPS calculates a significance p-value for each candidate gene to associate with the phenotype. If there is at least one candidate gene unconfirmed and not discouraged by any evidence (e.g., some of its variants might be false positive calls), it is advised to validate this candidate gene first. In this case, the study effectiveness measurement "chance to report the true phenotype-associated gene" is not important. If it is low, it just means that this analysis procedure happens to fit the need of identifying this candidate very well.

On the other hand, if an analysis procedure produces many candidates, this procedure is probably not accurate enough. Users may try a more aggressive approach to increase the confidence of reported candidate genes.

As mentioned above, it is always advised to optimize an analysis procedure toward accuracy. In case that a protocol is too strict to identify any candidate gene or all candidate genes have failed validation, the two study effectiveness measurements, "the chance to report candidate genes that are truly associated with the phenotype" and "the significance of violating Mendelian assumption", will provide guidance on the next steps. If the chance to report candidate genes is low, the analysis procedure is likely too strict and needs some relaxation. Users may consider validating more candidates that are reported by a more permissive procedure, or sequencing more phenotype-exhibiting samples to increase the support of the true phenotype-associated gene.

If an investigator have sequenced a large number of phenotype-exhibiting samples or have validated many candidates, and the phenotype-associated gene is still unidentified, another study effectiveness measurement, the significance of violating Mendelian assumption, may advise on the next steps. If this significance is low ($p>0.05$), there is no compelling evidence that the phenotype is controlled by multiple genes and users are still advised to add samples, validate more candidates, or to adjust their analysis procedures. However, if the significance of violating Mendelian assumption is high ($p<0.05$), users are advised to re-examine the phenotype-exhibiting samples included in the study. For a qualitative trait, though it is possible that multiple genes produce similar traits, it is unlikely that

two genes produce exactly the same phenotype. This is because most, if not all, genes are pleiotropic. Disruptions of different genes might produce the same major effect, but each will also exhibit dissimilar "side effects". Therefore, arguably, key to the success of a sequencing-based forward genetic study is the definition of proper phenotype criteria that can identify mutants of the same gene. The stricter the phenotype criteria are, the more likely the included samples harbor variants in the same gene. When GIPS reports a high significance of violating Mendelian assumption, users are advised to re-confirm the phenotypes of the samples included in the study, and if there is no doubt, to consider using a more strict phenotype criteria.

In the above example of *pho2* suppressor gene identification, the default analysis procedure reports 30 candidate genes (section 3.2.2), which are too many for validation. A more strict analysis procedure is therefore desired to identify candidates with higher confidence. For this reason, we optimize the analysis procedure toward accuracy and the final procedure identifies only 2 candidate genes. The first candidate was experimentally confirmed, and the second candidate is known to be highly variable, which has a high chance of being phenotype-unrelated. Although as a more accurate procedure, the optimized analysis procedure has a low "chance to report the true phenotype-associated gene", it does not matter as long as this procedure can still identify a biologically sound candidate gene worthy of further validation. The measurement, "chance to report the true phenotype-associated gene", is intended to provide guidance on the next steps if a procedure cannot identify any biologically sound candidate or all candidate genes have failed validation.

3.3 Reanalysis of the Kabuki sydrom exome sequencing data

3.3.1 Overview

Ng et al. sequenced the exome regions of 10 individuals exhibiting Kabuki syndrome (Ng, Bigham et al. 2010). Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, immunological defects and mild to moderate mental retardation. Originally described in 1981, Kabuki syndrome has an estimated incidence of 1 in 32,000, and approximately 400 cases have been reported worldwide. The vast majority of reported cases have been sporadic. These sequencing results have been deposited in dbGaP (Mailman, Feolo et al. 2007) with accession phs000295.v1.p1. This dataset is available for scientific investigators pursuing (any) research questions. To acquire this dataset, investigators are required to obtain permission from the NIH Data Access Committee. Application can be made within the dbGaP website. Users who do not have access to this dataset may not be able to replicate all results in this section. However, it should be noted that we use this dataset only for the purpose of putting our

discussion into context. No function of GIPS is specific to this dataset. In the original report of Ng, Bigham et al., sequencing reads from individuals exhibiting Kabuki syndrome were aligned to reference genome by MAP (Li, Ruan et al. 2008). Variants were called from mapped reads by MAP, requiring consensus quality >=20 and coverage >=8. Variants were filtered against known SNPs in dbSNP129 and 1000 genomes. Novel variants were further filtered based on their putative functions. Only non-synonymous variants, splice acceptor/donor mutations, and coding indels were kept. Even after such rigorous filtering, there were 26 genes left, which harbored novel variants in at least seven out of ten samples. These genes were considered "candidate genes". According to the authors, there was no obvious way to further prioritize these genes.

To identify genes that are truly associated with the phenotype, Ng, Bigham et al. used a subjective approach to rank the phenotypes of sequenced individuals, and found that only *MLL2* harbored high-effect variants in top four samples. The association between *MLL2* and kabuki syndrome was confirmed within 26 independent validation samples (Ng, Bigham et al. 2010).

A key step in the above analysis was to rank the phenotypes of sequenced individuals. This approach worked in this case. However, as the authors acknowledged, it is not an objective approach, which may not be generalizable to other studies (Ng, Bigham et al. 2010). The kabuki syndrome dataset has been re-analyzed with GIPS.

3.3.2 Run with empirical parameters

Step 1: Download raw sequencing data.

The Kabuki syndrome dataset is re-analyzed by GIPS. This dataset is downloaded from the following URLs with permission.

Table 6. Kabuki syndrome dataset.

SAMPLE ID	NCBI SRA database link
SRS086451	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086451
SRS086452	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086452
SRS086453	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086453
SRS086454	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086454
SRS086455	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086455
SRS086456	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086456
SRS086457	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086457
SRS086458	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086458
SRS086459	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086459

SRS086460	http://www.ncbi.nlm.nih.gov/sra/?term=SRS086460

Sequencing reads downloaded from dbGaP are in SRA format, which needs to be converted to FASTQ format using SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software):

```
fastq-dump.2 --split-3 /Path/to/sample.sra
```

Step2: Align sequencing reads to reference genome and call variants.

Sequencing reads are then aligned to reference human genome version grch37.p10, using bowtie2 (Langmead and Salzberg 2012) with default parameters. The core command is:

```
bowtie2 -p 6 -x /path/to/hs_ref_GRCh37.p10.Chr -U sample.fastq -S sample.sam
bowtie2 -p 6 -x /path/to/hs_ref_GRCh37.p10.Chr -1 sample_1.fastq -2 sample_2.fastq -S sample.sam
```

Then SAM format files are merged by sample. The SAMTools (Version: 0.1.18) package (Li, Handsaker et al. 2009) is used to call variants from aligned reads, requiring base quality >=13, reads map quality >=30, and depth >=10. "rmdup" is used to remove potential PCR duplicates. The core commands are:

```
samtools view -Bs /path/to/sample.sam /path/to/sample.bam

samtools sort /path/to/sample.bam /path/to/sample.sort

samtools index /path/to/sample.sort.bam

samtools rmdup /path/to/sample.sort.bam /path/to/sample.rmdup.bam

samtools index /path/to/sample.rmdup.bam

samtools mpileup -Q 13 -q 30 -ugSf /path/to/hs_ref_GRCh37.p10.Chr /path/to/sample.rmdup.bam |
bcftools view -bvcg - > /path/to/samtoolsQ13q30.bcf

bcftools view /path/to/samtoolsQ13q30.bcf | vcfutils.pl varFilter > /path/to/samtoolsQ13q30.vcf
```

Step 3: Set up the GIPS project

```
java -jar GIPS.jar -p Kabuki
```

Write a variant calling script (named "*default_script*" in this project) that be modified as GIPS required:

```
# $1 sam $2 vcf. Intermediate files are recommended to store in /path/to/Project/Work/intermediate
directory. Samtools version: 0.1.18

samtools view –Bs "$1" /path/to/Kabuki/Working/temporary/sample.bam

samtools sort /path/to/Kabuki/Working/temporary/sample.bam /path/to/Kabuki/Working/temporary/sample.sort

samtools index /path/to/Kabuki/Working/temporary/sample.sort.bam

samtools rmdup /path/to/Kabuki/Working/temporary/sample.sort.bam /path/to/Kabuki/Working/temporary/
sample.rmdup.bam

samtools index /path/to/Kabuki/Working/temporary/sample.rmdup.bam

samtools mpileup -Q 13 -q 30 -ugSf /path/to/hs_ref_GRCh37.p10.Chr /path/to/Kabuki/Working/temporary/
sample.rmdup.bam | bcftools view -bvcg - > /path/to/Kabuki/Working/temporary/samtoolsQ13q30.bcf

bcftools view /path/to/Kabuki/Working/temporary/samtoolsQ13q30.bcf | vcfutils.pl varFilter > "$2"
```

Then modify "*PROJECT.ini*":

Set "*REF_GENOME_ANNOTATION.GFF*", "*HUMAN_GRCH37_ANNOTATION.GFF*", "*SNPEFF*", "*SNPEFF_GENOME_VERSION*" appropriately. Set "*EFF_REGION*" as coding region, for exome sequencing always focuses on coding sequence region. Move the sample read mapping results and sample variant calling results to "*/path/to/Kabuki/Data*". Move the eligible variant calling script to "*/path/to/Kabuki/Script*". Specify the locations of these data/script in "*PROJECT.ini*", like:

```
[GLOBAL]
PROJECT :Kabuki
SNPEFF_GENOME_VERSION:GRCh37.75
SNPEFF  :     /path/to/snpEff_3_6
REF_GENOME_ANNOTATION.GFF :   /path/to/Kabuki/Ref/HomoSapiens.gff3
VAR_CALL_SCRIPT :   /nas2/zzx/PaperProject/Kabuki/Script/default_script
EFF_REGION  : CDS|SpliceSite=2
VAR_FILTERS :  EBA
SCORE_MATRIX: DEFAULT
CANDIDATE_CRITERIA :
NUM_SIM_SNPS : 10000
MAX_VAR_DENSITY   : 0
MAX_AA_SCORE:0

[SAMPLE_LIST]
SRS086451
SRS086452
SRS086453
SRS086454
SRS086455
SRS086456
SRS086457
SRS086458
SRS086459
SRS086460

[SAMPLE]
SAMPLE_NAME :SRS086451
SAMPLE.VCF  : /path/to/Kabuki/Data/SRS086451/samtoolsQ13q30.vcf
READS_ALIGNMENT.SAM :/path/to/Kabuki/Data/SRS086451/SRS086451.sam
```

Step 4: Call candidate genes.

Command: `java –Xms3G –jar GIPS.jar –T gips –p Kabuki`

In "*RESULT.gips*", one candidate gene is listed:

```
#============================================================
# Gene Identification via Phenotype Sequencing - Result File
#
# Job id: 7b850d1ffdc4d8bcba1f0c88c85c594c
# Completion time:2014-07-23 09:40:34
# Feature option '-T gips'
#============================================================

[Study Effectiveness]
Chance to report true phenotype-associated genes
If phenotype is recessive: 0.0717
If phenotype is  dominant: 0.0598

Significance of violating Mendelian phenotype assumption
If phenotype is recessive: 8.151E-6
If phenotype is  dominant: 1.0

Expected number of random genes that may be reported: 0.714

[Candidate Gene List]
#Details on the variants they harbor can be found in CANDIDATE_GENES.txt
GENE_ID GENE_NAME       Chr     EFF_REGION_LENGHT       SIGNIFICANCE    NUM_VARIANT
gene5348        TTN     Chr2    115881  0.66599 47      10
```

GIPS reports the *TTN* gene as candidate. This gene has long effective genomic region spanning 115881 bp. Because random mutations may easily accumulate inside such a long region, it is not very unexpected that all samples harbor SNPs in this genomic region. Consequently, GIPS reports a very weak significance value for this gene to associate with phenotype, 0.666. Furthermore, the expected number of random genes that may pass reporting criterion is 0.714. Also, in previous studies of gene-disease association, *TTN* was also frequently observed to harbor high-effect mutations in a significant proportion of samples, but its association with disease was seldom validated (Hughes and Davies 1994, Digilio, Marino et al. 2001). Therefore, *TTN* is likely a false positive candidate.

Further examination of "*KABUKI/Working/CANDIDATE_GENES.txt*" indicates that *TTN* harbors 47 variants in ten samples. This potentially means that the specificity of our current protocol to call variants is be high enough. Therefore, more stringent variant calling protocols and more rigorous variant filtering strategies are desired.

In "*RESULT.gips*", the change to report true phenotype-associated genes is estimated to be 0.072. As shown in the "*[Sample Specific Sensitivity and Specificity]*" section, the chance to detect a phenotype-related variant in each sample is never over 85%. Requiring phenotype-related variants to be detected in all samples is unrealistic. Therefore, a more relaxed candidate gene criterion is needed to increase the chance of identifying true phenotype-related genes.

```
Sample Variant Detection Sensitivity (VDS) (Measurement of sensitivity)
SAMPLE_NAME     HOMO_VDS        HETERO_VDS
SRS086451       0.8     0.789
SRS086452       0.842   0.823
SRS086453       0.809   0.799
SRS086454       0.833   0.834
SRS086455       0.527   0.509
SRS086456       0.812   0.788
SRS086457       0.763   0.754
SRS086458       0.761   0.745
SRS086459       0.785   0.767
SRS086460       0.81    0.8
```

It is further noted from the sample specific sensitivities and specificities that the sequencing quality of sample SRS086455 is inferior to those of the others. Users may consider using a different protocol to map sequencing reads and call variants for this sample, so that the VDS and BVF of this sample may better contribute to candidate gene identification. This observation may also alert users to treat this sample with caution in subsequent analyses.

3.3.3 Run with optimized parameters

Considering the observations discussed earlier, we make some changes to analysis protocol.

Step 5: Modify the analysis procedure.

1) Relax the criterion of candidate gene

Set candidate gene criterion to 7 (in global section), which increases the chance of reporting true phenotype-related genes.

2) Use a more strict quality criterion in variant calling

Use a more stringent variant calling script to call variants from all samples except SRS086455. This script requires a minimum base quality of 20, a minimum map quality of 30 and minimum sequencing depth of 10. The updated variant calling script looks as follows.

```
# $1 sam $2 vcf. Intermediate files are recommended to store in /path/to/Project/Work/
intermediate directory. Samtools version: 0.1.18

samtools view -Bs "$1" /path/to/Kabuki/Working/temporary/sample.bam

samtools sort /path/to/Kabuki/Working/temporary/sample.bam /path/to/Kabuki/Working/temporary/
sample.sort

samtools index /path/to/Kabuki/Working/temporary/sample.sort.bam

samtools rmdup /path/to/Kabuki/Working/temporary/sample.sort.bam /path/to/Kabuki/Working/
temporary/sample.rmdup.bam

samtools index /path/to/Kabuki/Working/temporary/sample.rmdup.bam

samtools mpileup -Q 20 -q 30 -ugSf /path/to/hs_ref_GRCh37.p10.Chr /path/to/Kabuki/Working/
temporary/sample.rmdup.bam | bcftools view -bvcg - > /path/to/Kabuki/Working/temporary/
samtoolsQ13q30.bcf

bcftools view /path/to/Kabuki/Working/temporary/samtoolsQ20q30.bcf | vcfutils.pl varFilter > "$2"
```

The updated script is then used to call variants from the actual sequencing reads. The resultant VCF files (i.e., "/path/to/Kabuki/Dataset/SRS086451/samtoolsQ13q30d10.vcf") are moved to the "/path/to/Kabuki/Data" folder. Meanwhile, create a script with GIPS requirement in

"/path/to/Kabuki/Script" with updated parameters. This script is named "updated.sh". Set new script path and new variant files to "VAR_CALL_SCRIPT".

3) Use a more strict "big difference filter"

Use the scoring matrix BLSOUM80 for filtering similar non-synonymous SNPs. This change leads to removal of more variants.

Update "*PROJECT.ini*" for the locations of these variant calling results (in sample specific sections) and the calling script (in global section). The "*PROJECT.ini*" file is updated to reflect these changes.

```
[GLOBAL]
PROJECT :Kabuki
SNPEFF_GENOME_VERSION:GRCh37.75
SNPEFF  :     /path/to/snpEff_3_6
REF_GENOME_ANNOTATION.GFF :   /path/to/Kabuki/Ref/HomoSapiens.gff3
VAR_CALL_SCRIPT :   /nas2/zzx/PaperProject/Kabuki/Script/updated.sh
EFF_REGION  : CDS|SpliceSite=2
VAR_FILTERS :  EBA
SCORE_MATRIX: BLOSUM80
CANDIDATE_CRITERIA :
NUM_SIM_SNPS : 10000
MAX_VAR_DENSITY   : 0
MAX_AA_SCORE:0

[SAMPLE_LIST]
SRS086451
SRS086452
SRS086453
SRS086454
SRS086455
SRS086456
SRS086457
SRS086458
SRS086459
SRS086460

[SAMPLE]
SAMPLE_NAME :SRS086451
SAMPLE.VCF  : /path/to/Kabuki/Data/SRS086451/samtoolsQ20q30d10.vcf
READS_ALIGNMENT.SAM :/path/to/Kabuki/Data/SRS086451/SRS086451.sam
```

4) Use a more relaxed variant calling script to call variants from sample SRS086455

Specifically, this script requires a minimum base quality of 13, a minimum map quality 20 and minimum depth 10x. Update "*PROJECT.ini*" for the calling script (in sample specific section for SRS086455).

```
[SAMPLE]
SAMPLE_NAME :SRS086455
SAMPLE.VCF  :/path/to/Kabuki/Data/SRS086455/samtoolsQ13q20d10.vcf
READS_ALIGNMENT.SAM :/path/to/Kabuki/Data/SRS086455/SRS086455.sam
VAR_CALL_SCRIPT :   /path/to/Kabuki/Script/callerscriptqQ13q20_d10-script
```

Step 6: Run GIPS in "update" mode to evaluate the modified analysis procedure

Command: `java -xms3G -jar GIPS.jar -T gips -p Kabuki -update`

In the process of variant filtering, the number of variants in each sample is shown in Table 7. After big difference filter, the remaining variants are defined as effective variants that are used to identify the casual gene.

Table 7. Number of variants in each sample while GIPS running.

Sample id	Origin variants	After ancestry reference filter	After effective region filter	After big difference filter	Specificity (probability to harbor irrelevant variants per base)
SRS086451	34780	3850	1353	471	1.352 E-5
SRS086452	43602	6617	1858	647	1.858 E-5
SRS086453	41450	6726	1562	595	1.709 E -5
SRS086454	44102	10546	1687	583	1.674 E -5
SRS086455	20916	5279	1537	701	2.013 E -5
SRS086456	31231	4077	1577	642	1.844 E -5
SRS086457	28516	3945	1660	614	1.763 E -5
SRS086458	27976	4265	2117	797	2.289 E -5
SRS086459	31955	6050	3113	1084	3.113 E -5
SRS086460	34798	5656	1674	657	1.887 E -5

In "*RESULT.gips*", variant calling sensitivity (VCS) of sample SRS086455 improved from 0.52 to 0.57. The chance to report true phenotype-associated genes improved from 0.056 to 0.52 (if dominant phenotype is assumed). The variant detection specificity improved from 3.03E-5 random variant per base to 1.95E-5 random variant per base. However, the false ignorance rate (chance to discard true phenotype-related variants by variant filters) increased from 0.031 to 0.144. This is because more strict filters are used. Nonetheless, because less random variant are kept, the statistical significance of the reported candidate genes are improved. The top candidate gene significance improved from 0.67 (insignificant) to 2.79E-05. These observations indicate that the new analysis protocol is more effective. Further optimization of this analysis protocol is possible, but is beyond the scope of this section, which is limited to illustrating GIPS usage in context. Main results of Kabuki project are shown in Table 8.

Table 8. Variant detection sensitivity along with variant calling sensitivity and false ignorance rate of each sample

SAMPLE_NAME	HOMO_VCS	HETERO_VCS	FIR	HOMO_VDS	HETERO_VDS
SRS086451	0.82	0.809	1.44E-01	0.701	0.691
SRS086452	0.858	0.859	1.44E-01	0.733	0.734
SRS086453	0.818	0.815	1.44E-01	0.699	0.696

SRS086454	0.857	0.855	1.44E-01	0.732	0.731
SRS086455	0.581	0.579	1.44E-01	0.496	0.495
SRS086456	0.831	0.759	1.44E-01	0.71	0.648
SRS086457	0.804	0.728	1.44E-01	0.687	0.622
SRS086458	0.775	0.707	1.44E-01	0.662	0.604
SRS086459	0.795	0.741	1.44E-01	0.679	0.633
SRS086460	0.837	0.764	0.144	0.715	0.653

3.3.4 Example discussion (advantage of using probabilistic candidate gene significance)

This example provides a case to illustrate the advantage of using probabilistic candidate gene significance. In contrast, most previous phenotype-sequencing experiments rank candidate genes only according to the number of phenotype-exhibiting samples in which they harbor variants.

The simple criterion, number of phenotype-exhibiting samples in which a candidate gene harbors variants, has its validity. However, it does not consider the facts that different variants have unequal chances to produce phenotype, that larger genes have a greater chance to harbor phenotype-unrelated random variants, and that different samples with heterogeneous sequencing qualities do not provide the same level of support to gene-phenotype association. As detailed in the algorithm section, the GIPS probabilistic candidate gene significance considers all these factors. Therefore, the candidate gene significance computed by GIPS provides a more reliable measurement to prioritize multiple candidate genes.

In the case of identification of Kabuki syndrome-associated genes, GIPS reports five candidate genes including the experimentally validated gene *MLL2* (Table 9). *MLL2* ranks the second in all candidate genes reported by GIPS. In contrast, *MLL2* ranked below tenth in the gene list reported by the original investigators. This difference illustrates the advantage of using GIPS probabilistic candidate gene significance. Furthermore, GIPS reports five candidate genes with the expected number of random gene discovery being 1.89. Consistently, it is noted that among the five candidate genes, *ALPK2* and *MLL2* have significant p values to associate with the phenotype, whereas the other three, *OBSCN*, *MUC16*, and *TTN* have insignificant p values. The top significance genes, *ALPK2* and *MLL2*, are not those harbor variants in the biggest proportion of phenotype-exhibiting samples. They harbor variants in 7 samples, but their significances are higher than those of *OBSCN*, *MUC16*, and *TTN*, which harbor variants in 7, 8, and 9 samples respectively. The reason why OBSCN, *MUC16*, and *TTN* have lower significances is that they have longer coding sequences than *ALPK2* and *MLL2*. Therefore, they have higher chances to accumulate phenotype-unrelated random variants.

Table 9. Candidate genes reported by GIPS. Genes are ranked by their significances to associate with the phenotype. The last two columns show the total number of variants detected in the candidate genes and the number of samples in which the candidate genes harbor variants.

Gene	Gene length	Significance	Variants	Samples
ALPK2	6561	2.79E-05	7	7
MLL2	16826	0.0063278	9	7
OBSCN	28446	0.07042	14	7
MUC16	43856	0.11724	24	8
TTN	115881	0.67339	24	9
Expected number of random gene discovery: 1.869				

In addition, it is interesting to note that GIPS reports *ALPK2* as the top candidate. This gene was not identified by the original investigators and was not validated for association with Kabuki syndrome. However, it is known that a wide variety of health problems frequently occur with Kabuki syndrome. Among the most reported are heart abnormalities (Hughes and Davies 1994, Digilio, Marino et al. 2001). *ALPK2* is a heart alpha-kinase, which is consistent with the known association between Kabuki syndrome and heart abnormalities. Therefore, this example also produces a new hypothesis of Kabuki syndrome-associated gene, to be further validated.

References

Chilamakuri, C. S. R., S. Lorenz, M.-A. Madoui, D. Vodák, J. Sun, E. Hovig, O. Myklebost and L. A. Meza-Zepeda (2014). "Performance comparison of four exome capture systems for deep sequencing." BMC genomics **15**(1): 449.

Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly **6**(2): 0--1.

Digilio, M. C., B. Marino, A. Toscano, A. Giannotti and B. Dallapiccola (2001). "Congenital heart defects in Kabuki syndrome." American journal of medical genetics **100**(4): 269-274.

Hu, B., C. Zhu, F. Li, J. Tang, Y. Wang, A. Lin, L. Liu, R. Che and C. Chu (2011). "LEAF TIP NECROSIS1 plays a pivotal role in the regulation of multiple phosphate starvation responses in rice." Plant Physiology **156**(3): 1101-1115.

Huang, W., L. Li, J. R. Myers and G. T. Marth (2012). "ART: a next-generation sequencing read simulator." Bioinformatics **28**(4): 593-594.

Hughes, H. E. and S. J. Davies (1994). "Coarctation of the aorta in Kabuki syndrome." Archives of disease in childhood **70**(6): 512-514.

Kawahara, Y., M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu and S. Zhou (2013). "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data." Rice **6**(1): 4.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church and D. R. Maglott (2014). "ClinVar: public archive of relationships among sequence variation and human phenotype." Nucleic acids research **42**(D1): D980-D985.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Lelieveld, S. H., M. Spielmann, S. Mundlos, J. A. Veltman and C. Gilissen (2015). "Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein Coding Regions." Human Mutation.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H., J. Ruan and R. Durbin (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome research **18**(11): 1851-1858.

Liu, T.-Y., T.-K. Huang, C.-Y. Tseng, Y.-S. Lai, S.-I. Lin, W.-Y. Lin, J.-W. Chen and T.-J. Chiou (2012). "PHO2-dependent degradation of PHO1 modulates phosphate homeostasis in Arabidopsis." The Plant Cell Online **24**(5): 2168-2183.

Mailman, M. D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall and L. Phan (2007). "The NCBI dbGaP database of genotypes and phenotypes." Nature genetics **39**(10): 1181-1186.

Ng, S. B., A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper and H. C. Mefford (2010). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." Nature genetics **42**(9): 790-793.

Raghothama, K. G. (1999). "Phosphate acquisition." Annual review of plant biology **50**(1): 665-693.
Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri and S. C. Schuster (2013). "Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample." PloS one **8**(2): e55089.

Stein, L. (2007). Generic feature format version 3.

Wang, C., S. Ying, H. Huang, K. Li, P. Wu and H. Shou (2009). "Involvement of OsSPX1 in phosphate homeostasis in rice." The Plant Journal **57**(5): 895-904.

Term abbreviation and illustration

Term	Abbreviation/Illustration	Section
BVF :	Background variant frequency	1.3
CDS:	Coding sequence	
ClinVar:	Clinical variant database	1.2.2
FIR:	False ignorance rate	1.2.2
GIPS:	Gene identification via phenotype sequencing	
SAM :	Sequence Alignment/Map Format	
VDS:	Variant detection sensitivity	1.2
VCS:	Variant calling sensitivity	1.2.1
VCF:	Variant Call Format	
PROJECT.ini	Specify the analysis procedure that will be performed and evaluated by GIPS	2.2.2
RESULT.gips	GIPS result information	2.2.3
CANDIDATE_GENES.txt	Detailed information about the variants in candidate genes which are given in the result file	2.2.3
Effective region	The specific region which may harbor phenotype related variants	1.2.2.1
Effective region filter	Discards variants that are outside the user-specified “effective genomic region”	1.2.2.1
Ancestry filter	Discards variants that are detected in more than one phenotype-exhibiting samples	1.2.2.3
Big difference filter	Discards non-synonymous variants that result in similar amino acids	1.2.2.2
Congestion filter	Discards a variant if there is another variant in its vicinity (11bp region centering on it)	1.2.2.4
Control filter	Discards variants observed in phenotype-absent control samples	1.2.2.5

Annex I: Kabuki exome sequencing: variants detected in each candidate gene

Gene: *ALPK2*

Sample	Chr	Pos	Ref	Alt	Genotype	Effect
SRS086452	Chr18	56184343	G	A	0/1	NON_SYNONYMOUS_CODING(R:C)
SRS086454	Chr18	56205373	TGGGG	TGGGGG	0/1	FRAME_SHIFT
SRS086455	Chr18	56247525	AG	AGG	0/1	FRAME_SHIFT
SRS086456	Chr18	56203252	AAAGAAG	AAAG	0/1	CODON_DELETION
SRS086457	Chr18	56202470	G	A	0/1	NON_SYNONYMOUS_CODING(A:V)
SRS086458	Chr18	56204391	T	TCTCTC	0/1	FRAME_SHIFT
SRS086459	Chr18	56191053	G	A	0/1	NON_SYNONYMOUS_CODING(R:C)

Gene: *MLL2*

Sample	Chr	Pos	Ref	Alt	Genotype	Effect
SRS086453	Chr12	49443785	GT	GTT	0/1	FRAME_SHIFT
SRS086453	Chr12	49445028	G	A	0/1	NON_SYNONYMOUS_CODING(P:L)
SRS086454	Chr12	49420554	C	T	0/1	STOP_GAINED
SRS086454	Chr12	49426460	A	G	0/1	NON_SYNONYMOUS_CODING(S:P)
SRS086455	Chr12	49432651	G	A	0/1	STOP_GAINED
SRS086456	Chr12	49435971	G	A	0/1	STOP_GAINED
SRS086457	Chr12	49435258	G	A	0/1	STOP_GAINED
SRS086458	Chr12	49420131	A	C	0/1	STOP_GAINED
SRS086460	Chr12	49425791	G	A	0/1	STOP_GAINED

Gene: *OBSCN*

Sample	Chr	Pos	Ref	Alt	Genotype	Effect
SRS086452	Chr1	228482059	G	A	0/1	NON_SYNONYMOUS_CODING(G:R)
SRS086453	Chr1	228412228	G	A	0/1	NON_SYNONYMOUS_CODING(A:T)
SRS086453	Chr1	228432264	A	T	0/1	NON_SYNONYMOUS_CODING(H:L)
SRS086453	Chr1	228434280	C	A	0/1	NON_SYNONYMOUS_CODING(T:N)
SRS086453	Chr1	228479598	G	T	0/1	SPLICE_SITE_ACCEPTOR
SRS086454	Chr1	228505725	A	C	0/1	NON_SYNONYMOUS_CODING(E:A)
SRS086455	Chr1	228444565	T	A	1/1	NON_SYNONYMOUS_CODING(V:D)
SRS086455	Chr1	228495983	G	A	0/1	NON_SYNONYMOUS_CODING(R:H)
SRS086456	Chr1	228504670	C	T	1/1	NON_SYNONYMOUS_CODING(R:W)
SRS086456	Chr1	228469825	C	T	0/1	NON_SYNONYMOUS_CODING(R:W)
SRS086458	Chr1	228412228	G	A,C	1/1	NON_SYNONYMOUS_CODING(A:P),NON_SYNONYMOUS_CODING(A:T)

SRS086458	Chr1	228481897	A	C	0/1	NON_SYNONYMOUS_CODING(T:P)
SRS086459	Chr1	228447432	T	G	0/1	NON_SYNONYMOUS_CODING(C:G)
SRS086459	Chr1	228548197	G	A	0/1	NON_SYNONYMOUS_CODING(R:H)

Gene: *MUC16*

Sample	Chr	Pos	Ref	Alt	Genotype	Effect
SRS086451	Chr19	9087970	G	A	0/1	NON_SYNONYMOUS_CODING(T:I)
SRS086452	Chr19	9050111	G	T	0/1	NON_SYNONYMOUS_CODING(T:N)
SRS086452	Chr19	9057177	G	T	0/1	NON_SYNONYMOUS_CODING(T:N)
SRS086452	Chr19	9083170	G	A	0/1	NON_SYNONYMOUS_CODING(T:I)
SRS086453	Chr19	9048895	A	C	0/1	NON_SYNONYMOUS_CODING(S:R)
SRS086453	Chr19	9057468	T	A	0/1	NON_SYNONYMOUS_CODING(H:L)
SRS086453	Chr19	9059159	T	G	0/1	NON_SYNONYMOUS_CODING(K:N)
SRS086453	Chr19	9059181	G	A	0/1	NON_SYNONYMOUS_CODING(P:L)
SRS086453	Chr19	9059307	C	T	0/1	NON_SYNONYMOUS_CODING(G:D)
SRS086453	Chr19	9060915	C	T	0/1	NON_SYNONYMOUS_CODING(R:H)
SRS086453	Chr19	9065203	T	C	0/1	NON_SYNONYMOUS_CODING(S:G)
SRS086453	Chr19	9086983	G	T	0/1	NON_SYNONYMOUS_CODING(T:K)
SRS086454	Chr19	9009301	G	A	0/1	NON_SYNONYMOUS_CODING(P:S)
SRS086454	Chr19	9071179	C	T	0/1	NON_SYNONYMOUS_CODING(A:T)
SRS086455	Chr19	9064929	G	T	0/1	STOP_GAINED
SRS086455	Chr19	9085050	GGAG	GGAGAG	0/1	FRAME_SHIFT
SRS086457	Chr19	9045821	G	A	0/1	NON_SYNONYMOUS_CODING(T:I)
SRS086457	Chr19	9076911	G	A	0/1	NON_SYNONYMOUS_CODING(T:I)
SRS086457	Chr19	9083791	G	A	0/1	NON_SYNONYMOUS_CODING(S:L)
SRS086457	Chr19	9085958	A	G	0/1	NON_SYNONYMOUS_CODING(S:P)
SRS086457	Chr19	9086145	G	C	0/1	NON_SYNONYMOUS_CODING(N:K)
SRS086457	Chr19	9090182	T	C	0/1	NON_SYNONYMOUS_CODING(T:A)
SRS086458	Chr19	9089870	C	A	0/1	NON_SYNONYMOUS_CODING(V:F)
SRS086459	Chr19	9009641	G	A	0/1	NON_SYNONYMOUS_CODING(R:C)

Gene: *TTN*

Sample	Chr	Pos	Ref	Alt	Genotype	Effect
SRS086451	Chr2	179606538	G	A	0/1	NON_SYNONYMOUS_CODING(P:S)
SRS086452	Chr2	179393111	A	G	0/1	NON_SYNONYMOUS_CODING(V:A)
SRS086452	Chr2	179395560	G	A	0/1	NON_SYNONYMOUS_CODING(P:L)
SRS086452	Chr2	179398509	C	A	0/1	NON_SYNONYMOUS_CODING(G:V)
SRS086452	Chr2	179404628	T	A	0/1	NON_SYNONYMOUS_CODING(I:F)

SRS086452	Chr2	179554549	G	C	0/1	NON_SYNONYMOUS_CODING(P:A)
SRS086452	Chr2	179612383	C	T	0/1	NON_SYNONYMOUS_CODING(R:H)
SRS086453	Chr2	179472319	C	T	0/1	NON_SYNONYMOUS_CODING(R:H)
SRS086453	Chr2	179613049	A	G	0/1	NON_SYNONYMOUS_CODING(I:T)
SRS086454	Chr2	179586604	C	G	0/1	NON_SYNONYMOUS_CODING(D:H)
SRS086455	Chr2	179422448	G	GGA	0/1	FRAME_SHIFT
SRS086455	Chr2	179659912	G	A	0/1	NON_SYNONYMOUS_CODING(R:C)
SRS086456	Chr2	179430889	T	TAG	0/1	FRAME_SHIFT
SRS086456	Chr2	179486037	C	A	0/1	NON_SYNONYMOUS_CODING(K:N)
SRS086458	Chr2	179497018	G	A	1/1	NON_SYNONYMOUS_CODING(R:C)
SRS086459	Chr2	179428061	C	T	0/1	NON_SYNONYMOUS_CODING(A:T)
SRS086459	Chr2	179435887	A	G	0/1	NON_SYNONYMOUS_CODING(I:T)
SRS086459	Chr2	179440820	C	G	0/1	NON_SYNONYMOUS_CODING(D:H)
SRS086459	Chr2	179454686	G	T	0/1	NON_SYNONYMOUS_CODING(S:Y
SRS086459	Chr2	179542435	C	A	0/1	STOP_GAINED
SRS086459	Chr2	179643821	G	A	0/1	NON_SYNONYMOUS_CODING(R:C)
SRS086460	Chr2	179404498	G	C	0/1	NON_SYNONYMOUS_CODING(A:G)
SRS086460	Chr2	179458723	C	G	0/1	NON_SYNONYMOUS_CODING(G:A)
SRS086460	Chr2	179610967	C	T	0/1	NON_SYNONYMOUS_CODING(C:Y)