Gene Identification via Phenotype Sequencing

Version 1.5

User manual

Zhu Z, Wang WT, Zhu JH, and Chen X

2015-08-01

Table of contents

Chapter 1. Algorithm

1.1 Overview

The advent of next generation sequencing has enabled fast and cost-effective genotyping, which may significantly accelerate the process of gene identification in forward genetics studies. Sequencing-based forward genetics studies are frequently referred to 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 to 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 procedure with an integral consideration of all factors that affect its chance of success. In particular, after the sequencing results are ready, 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 measurements to help optimize an analysis procedure. The four measurements 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**. Their interpretations are discussed in more detail in section 3.2.4 (Example discussion: strategy to optimize an analysis procedure).

As previously described, a range of experimental and analysis parameters can affect these four measurements of study effectiveness. In general, their impacts on these four study-wise effectiveness measurements can be summarized into their impacts on two sample-wise effectiveness measurements: the “sensitivity” and “specificity” for each sample, with which an analysis procedure under evaluation is expected to detect the phenotype-causing variants. Note that in this framework, it is permitted that different analysis procedures can be 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 measurements 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 to reflect the belief of the investigator.

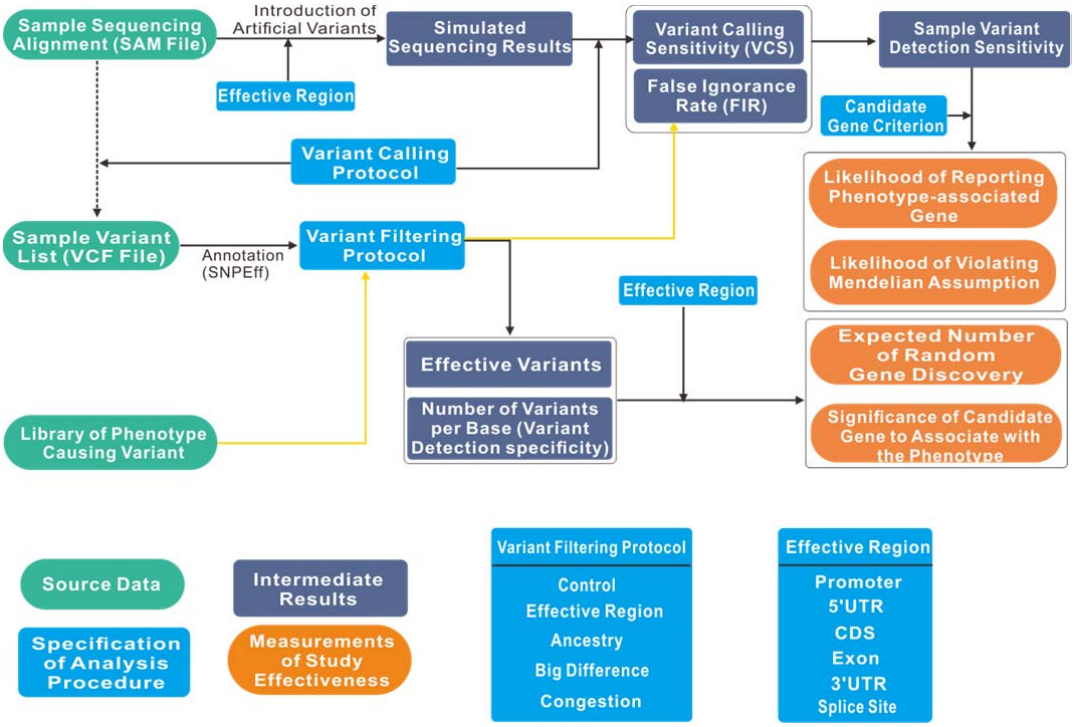


Figure 1. The general workflow of GIPS. Details are give in later sections.

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 to 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 these artificial SNPs can be detected.

GIPS measures the sample-wise specificities of an analysis procedure by computing the frequency of detected variants per base in the effective genomic region (after all variant filtering steps). Assuming that all detected variants are unrelated to the phenotype, these frequencies are 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 procedure 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 procedure under evaluation, before any filtering. It depends on the quality and depth distribution of the sequencing data, the approach to align sequencing reads, and the approach to 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 actually 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 (this filter is applicable only if all samples are known unrelated)**; **3) the big difference filter, which removes small-difference non-synonymous mutations that are unlikely to change protein function**; and **4) the congestion filter, which removes potentially spurious variants that are results of ambiguous alignments of sequencing reads**.

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 procedures.

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 these artificial SNPs that can be detected with the user-supplied variant calling procedure (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 the user-supplied 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 their scripts to call variants from the original alignment results and supply the resultant VCF files to GIPS. This design is also to ensure that the user-supplied variant calling scripts are working.

To estimate H_{S_i} for heterozygous variants, GIPS uses the same approach 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 $S_i S_i$. The same filtering procedure will always have the same FIR, however, it is allowed to use different filtering procedures on different samples. GIPS supports the use of five 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.

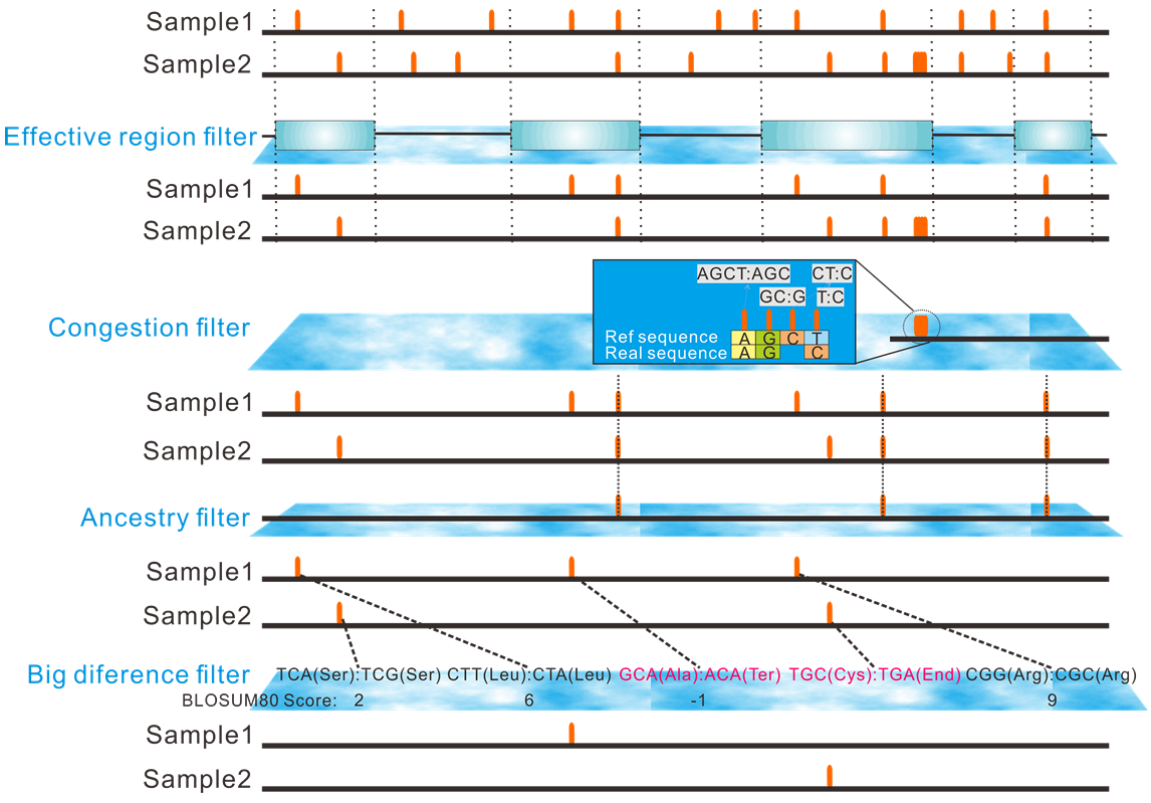


Figure 2. Variant filters.

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 distribution 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 according to variants detected in other samples. However, variants detected in other samples could be spurious. Therefore, GIPS takes the proportion of "effective genomic region" ignored by these study based variant filters as their estimated FIR.

Whether to use a specific 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. "C" here stands for congestion filter.

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 based on 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).

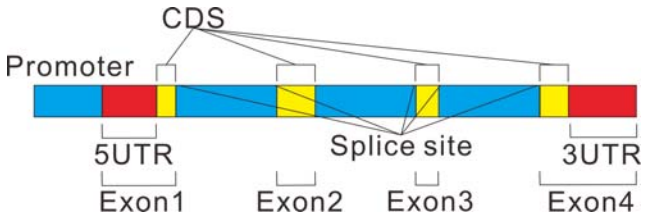


Figure 3. Gene structure model for specification of effective genomic region.

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. 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 "|". Defaults to "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 what is "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, which are 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 to define a significant amino acid change, 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 a shared genetic background, rather than phenotype association.

Note that the ancestry filter should only be used when all the phenotype-exhibiting samples are known 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 these genomic regions are more frequently false positive calls, especially in cases when many variants appear in a short region. Congestion filter discards a variant if there is another variant exists 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.

1.3Variant detection specificity

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

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

1.4Measurements 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).

With this candidate gene criterion, 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 might accumulate random variants in ≥ M phenotype-exhibiting samples by chance.

Let:

G_j be the candidate gene;

L_{G_j} be the length of effective genomic region that is 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:

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

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:

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

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 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), the structure of project directory is illustrated below (Figure 4). The GIPS parameter file "*PROJECT.ini*" will be placed in the main project folder, and the GIPS result file "*RESULT.gips*" will be placed in the "*Working*" sub-folder.

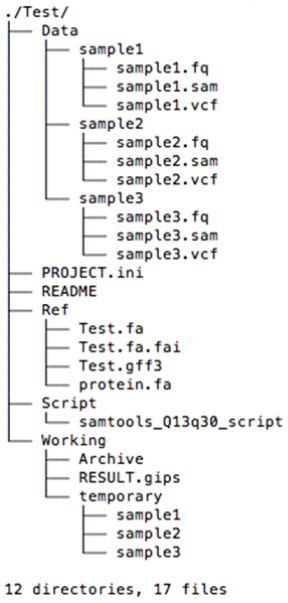


Figure 4. Structure of the GIPS project directory for "test" project.

"*PROJECT.ini*" specifies the analysis procedure that will be 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 to 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 that 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 to call variants from the actual sequencing results. GIPS only uses these scripts to process simulated reads to estimate 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 results, and supply the resultant VCF files to GIPS. Furthermore, it is recommended to use a naming convention for all intermediate result files that 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 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 procedure. Candidate genes will be reported and the four measurements of study effectiveness will be computed. The “*RESULT.gips*” file reports the GIPS results. A summary of the analysis procedure and intermediate results is also included in “*RESULT.gips*”.

After the first GIPS run, users can iteratively 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. 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 procedure (not the procedure of the immediate previous run), 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 library of phenotype-causing variants.

“*Data*”, for storing user-provided sequence alignment (SAM) files and variant call (VCF) files.

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, 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]*" specific 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. Defaults 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 be enabled. 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 the variant calling result that 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*" is 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*" is 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*" is 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.

Please note that the three parameters, SPECIFY_HOMO_VDS, SPECIFY_HETERO_VDS and SPECIFY_BVF, should always be used together. Otherwise, GIPS would report an error, as below.

```
Job start time: 2015-07-31 10:09:24
temporary folder is created
Project: Test
Directory: /nas2/wwt/GIPS/Test
--------------------------------------------------------------------------------
        Gene Identification via Phenotype Sequencing  (Version 1.1.5)
     Copyright(c) 2013-2014, Zhongxu Zhu, Xin Chen. All Rights Reserved.
--------------------------------------------------------------------------------

SAMPLE_NAME:sample1
SPECIFY_HOMO_VDS:1
SPECIFY_BVF:0.03
SAMPLE.VCF:/nas2/wwt/GIPS/Test/Data/sample1/sample1.vcf
08:00:00.67     [sample1] SPECIFY_BVF, SPECIFY_HOMO_VDS and SPECIFY_HETERO_VDS parameters should be specified together
```

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

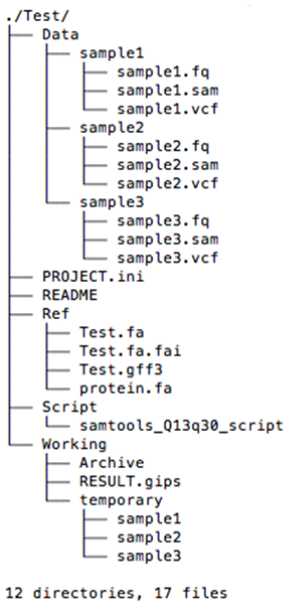
3.1 Test example

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.



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's 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 hypothetical reference genome sequence of the 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 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 executed 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 the "*Test/Script/samtools_Q13q30_script*" accordingly.

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

Unzip the downloaded package and set its folder path to the "*SNPEFF*" parameter in the "*[GLOBAL]*" section of the "*PROJECT.ini*" file and create a new folder named "Test" in "SNPEFF/data".

Then you should put the "*project_folder*/*Ref*/*Test.gff3*" file to the new Test folder and run the following command to create the snpEffectPredictor.bin file of Test.

Command: `java -jar /path/to/SNPEFF/snpeff.jar build -gff3 -v Test`

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 to estimate 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 by 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: c8ad42c6ee5bb33346a2035eaf08b6aa
# Completion time:2015-07-31 15:26:50
# Feature option '-T vcs'
#================================================================

[Sample Specific Variant Calling Sensitivity (VCS)]
SAMPLE_NAME     HOMO_VCS        HETERO_VCS
sample1 0.661   0.263
sample2 0.638   0.351
sample3 0.647   0.383

#----------------------------------------------------------------
# [Protocol]
#----------------------------------------------------------------
# [GLOBAL]
# PROJECT        : Test
# REF_GENOME_ANNOTATION.GFF: /nas2/wwt/GIPS/Test/Ref/Test.gff3
# SNPEFF_GENOME_VERSION    : Test
```

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 –jar 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: 138ef913f56f1a44e9b83ca8415abdda
# Completion time:2015-07-31 15:31:46
# Feature option '-T gips'
#============================================================

[Study Effectiveness]
Chance to report true phenotype-associated genes
If phenotype is recessive: 0.256
If phenotype is  dominant: 0.0394

Significance of violating Mendelian phenotype assumption
If phenotype is recessive: 0.743
If phenotype is  dominant: 1.0

Expected number of random genes that may be reported: 0.861

[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     0.03706 8       3

[Sample Specific Sensitivity and Specificity]
Variant Calling Sensitivity (VCS)
SAMPLE_NAME     HOMO_VCS        HETERO_VCS
sample1 0.647   0.276
sample2 0.631   0.387
sample3 0.685   0.402

False Ignorance Rate (FIR)
SAMPLE_NAME     FIR_FUNCTION    FIR_STUDY       FIR
sample1 0.0263  0.0     0.0263
sample2 0.0263  7.094E-4         0.027
sample3 0.0263  7.094E-4         0.027

Sample Variant Detection Sensitivity (VDS) (Measurement of sensitivity)
SAMPLE_NAME     HOMO_VDS        HETERO_VDS
sample1 0.629   0.268
sample2 0.613   0.376
sample3 0.666   0.391
```

In addition, a text file "*Test/Working/CANDIDATE_GENES.txt*" will be created, which provides detailed information on 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 enabled. The raw sequencing results can be retrieved from the Sequence Read Archive Database with accession SRP058881.

3.2.2 Run with empirical parameters

Step 1: Download raw sequencing data and convert to the FASTQ format.

Table 2. The Pho2 suppressor identification 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 data in SRA format need to be converted to FASTQ format using the SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software).

```
vdb-validate.2 /path/to/sample.sra
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), using the Bowtie2 software (Langmead and Salzberg 2012). The reference genome FASTA file (all.chrs.con.fa) and annotation file (all.gff3) can be downloaded from

ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotat

ion_dbs/pseudomolecules/version_7.0/all.dir/

There is a blank line in the reference genome file (all.chrs.con.fa) at the bottom of Chr2. This blank line needs to be removed, or it will cause an error when building Bowtie2 index. This blank line looks like below:

```
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

>Chr3
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Suppose that the reference genome and annotation files are store in "/path/to/pho2/reference/", and 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 commands are:

```
# bowtie2 builds index
/path/to/bowtie2-build /path/to/pho2/reference/all.chrs.con.fa rice7

# bowtie2 alignment
/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

# sort and add read groups
java -Xms8g -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"

# remove duplicates
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

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

# samtools call variants
/path/to/samtools mpileup -ugf /path/to/all.chrs.con.fa /path/to/pho2/dataset/
sample_markdup.bam -go - | /path/to/bcftools call -vmO 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 –init 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 : /path/to/pho2/reference/all.gff3
SNPEFF_GENOME_VERSION     : rice7
SNPEFF  : /path/to/snpEff
```

Move the reads alignment results and variant calling results obtained in Step 2 into the GIPS project folder, "*/path/to/GIPS-Rice/Data*". Provide the variant calling steps to GIPS as a script. See below for an example. This script takes two command line parameters. The first specifies the input sam file and the second specifies the output vcf file. Here, the below example script is named "*default-Q13-q0.sh*". Move the variant calling script to "*/path/to/rice/Script*" and set its path in "*PROJECT.ini*".

```
#! /bin/bash
# $1 /path/to/sam $2 /path/to/vcf. Intermediate files are recommended to store in /path/
to/project/Working/temporary. samtools version: 1.2

# sort and add read groups
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 duplicates
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

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

# samtools call variants
/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 -vm0 z -o - | /path/to/bcftools
filter -O v -o "$2"
```

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

```
[GLOBAL]
PROJECT :GIPS-Rice
REF_GENOME_ANNOTATION.GFF : /path/to/pho2/reference/all.gff3
SNPEFF_GENOME_VERSION     : rice7
SNPEFF  : /path/to/snpEff
CANDIDATE_CRITERION : 3
VAR_CALL_SCRIPT : /path/to/GIPS-Rice/Script/default-Q13-q0.sh
EFF_REGION  : CDS|SpliceSite=2|Promoter=1000
VAR_FILTERS :  EAB
SCORE_MATRIX: DEFAULT
MAX_AA_SCORE:
NUM_SIM_SNPS : 10000

[SAMPLE_LIST]
M28
M29
M249


[SAMPLE]
SAMPLE_NAME : M28
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M28/samtools_Q13_q0_noDPfilter.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_noDPfilter.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M29/btout.sam

[SAMPLE]
SAMPLE_NAME : M249
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M249/samtools_Q13_q0_noDPfilter.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 considers only 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*", this default analysis procedure has ~91% chance to detect homozygous variant and ~81% chance to detect heterozygous variant in each sample. The combined likelihood of the true phenotype-associated gene harboring detected variants in all three samples is high (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	0.934	0.837	0.0167	0.918	0.822	1.51E-5
M29	0.883	0.812	0.0167	0.868	0.798	8.32E-6
M249	0.952	0.837	0.0167	0.936	0.822	8.46E-6

The default analysis procedure identified 28 candidate genes ("*/path/to/GIPS-Rice/Working/RESULT.gips*. Detailed information on the variants that these candidate genes harbor is provided in the result file "*/path/to/GIPS-Rice/Working/CANDIDATE_GENES.txt*" "), which looks like below.

```
[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    1.91128E-6      5       3
LOC_Os12g26100  LOC_Os12g26100  Chr12   1272    2.14343E-6      18      3
LOC_Os03g64310  LOC_Os03g64310  Chr3    1312    2.35057E-6      6       3
LOC_Os04g02550  LOC_Os04g02550  Chr4    1446    3.14017E-6      16      3
LOC_Os08g05950  LOC_Os08g05950  Chr8    1470    3.29787E-6      6       3
LOC_Os04g08690  LOC_Os04g08690  Chr4    1489    3.42638E-6      4       3
LOC_Os06g39744  LOC_Os06g39744  Chr6    1558    3.92082E-6      5       3
LOC_Os01g67710  LOC_Os01g67710  Chr1    1658    4.71779E-6      8       3
LOC_Os10g07040  LOC_Os10g07040  Chr10   2201    1.09421E-5      16      3
LOC_Os04g35000  LOC_Os04g35000  Chr4    2571    1.73379E-5      6       3
LOC_Os11g12710  LOC_Os11g12710  Chr11   2717    2.04153E-5      4       3
LOC_Os06g38270  LOC_Os06g38270  Chr6    3166    3.20722E-5      13      3
LOC_Os04g51840  LOC_Os04g51840  Chr4    3180    3.24924E-5      7       3
LOC_Os10g15210  LOC_Os10g15210  Chr10   3194    3.29161E-5      8       3
LOC_Os10g15170  LOC_Os10g15170  Chr10   3356    3.80851E-5      5       3
LOC_Os10g13550  LOC_Os10g13550  Chr10   3447    4.12086E-5      3       3
LOC_Os02g56510  LOC_Os02g56510  Chr2    3504    4.32479E-5      3       3
LOC_Os08g06280  LOC_Os08g06280  Chr8    3529    4.41627E-5      10      3
LOC_Os12g11570  LOC_Os12g11570  Chr12   3975    6.26682E-5      3       3
LOC_Os12g08110  LOC_Os12g08110  Chr12   4141    7.06662E-5      7       3
LOC_Os04g02520  LOC_Os04g02520  Chr4    4484    8.92358E-5      12      3
LOC_Os05g51770  LOC_Os05g51770  Chr5    4780    1.07595E-4      5       3
LOC_Os03g64320  LOC_Os03g64320  Chr3    4819    1.10183E-4      6       3
LOC_Os03g23760  LOC_Os03g23760  Chr3    4838    1.11458E-4      3       3
LOC_Os04g05104  LOC_Os04g05104  Chr4    5084    1.28838E-4      6       3
LOC_Os02g32690  LOC_Os02g32690  Chr2    5330    1.47884E-4      12      3
LOC_Os04g49890  LOC_Os04g49890  Chr4    5346    1.49183E-4      6       3
LOC_Os04g13510  LOC_Os04g13510  Chr4    5578    1.68841E-4      23      3
```

Rough running times for the above analysis procedures are listed in Table 4.

Table 4. Rough running times on a single CPU Intel Xeon E3-1230 server.

Procedure	Runtime
SRA Toolkit	~12min
Bowtie2 build index	~9min
Bowtie2 align	~2.5h
Picard sort and add groups	~45min
Picard rmdup	~40min
Samtools index	~2.5min
Samtools call variants	~3.5h
Total	**~ 8h**
GIPS	**1h**

3.2.3 Run with optimized parameters

The default analysis procedure identified 28 candidate genes, which are too many for experimental validation. Such a big number of identified candidates indicates the lack of specificity of the default analysis procedure. Therefore, next, we modify the default analysis procedure to improve its specificity.

Step 5: Modify the analysis procedure.

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

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 procedure is as below.

```
/path/to/samtools mpileup -ugf /path/to/all.chrs.con.fa /path/to/pho2/dataset/
sample_markdup.bam -Q 15 -q 20 -go - | /path/to/bcftools call -vmO z -o - | /path/to/
bcftools filter -O v -o /path/to/pho2/dataset/samtools_Q15_q20_noDPfilter.vcf
```

The updated procedure is then used to call variants from the actual sequencing reads. The resultant VCF files (i.e., "/*path/to/pho2/Dataset/sample/samtools_Q15_q20_noDPfilter.vcf*") should be moved to the "*/path/to/GIPS-Rice/Data/sample/*" folder. Meanwhile, create a new script to supply this updated procedure to GIPS. We name this script "*samtools_Q15_q20.sh*", and put it in "/*path/to/GIPS-Rice/Script*". This script is a modified version of "*default-Q13-q0.sh*" (see above), with the samtools parameters changed. This script takes two command line parameters. The first specifies the input sam file and the second specifies the output vcf file.

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

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

SCORE_MATRIX: PAM120

This change instructs 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 CDS regions and splice sites.

It is arguable that most phenotype-related de novo mutations happen in the CDS regions and splice sites. By considering only the CDS regions and splice sites, we reduce the number of detected variants that may confound the gene identification process.

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

EFF_REGION: CDS|SpliceSite=2

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

```
[GLOBAL]
PROJECT :GIPS-Rice
REF_GENOME_ANNOTATION.GFF : /path/to/pho2/reference/all.gff3
SNPEFF_GENOME_VERSION     : rice7
SNPEFF  : /path/to/snpEff
CANDIDATE_CRITERION : 3
VAR_CALL_SCRIPT : /path/to/GIPS-Rice/Script/samtools-Q15-q20.sh
EFF_REGION  : CDS|SpliceSite=2
VAR_FILTERS :  EAB
SCORE_MATRIX: PAM120
MAX_AA_SCORE:
NUM_SIM_SNPS : 10000

[SAMPLE_LIST]
M28
M29
M249

[SAMPLE]
SAMPLE_NAME : M28
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M28/samtools_Q15_q20_noDPfilter.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_noDPfilter.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M29/btout.sam

[SAMPLE]
SAMPLE_NAME : M249
SAMPLE.VCF  : /path/to/GIPS-Rice/Data/M249/samtools_Q15_q20_noDPfilter.vcf
READS_ALIGNMENT.SAM : /path/to/GIPS-Rice/Data/M249/btout.sam
```

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 variants and ~67% chance to detect heterozygous variants in each sample. The combined likelihood of the true phenotype-associated gene harboring detected variants in all three samples is 29%. These results are summarized in Table 5.

Table 5. 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	0.916	0.879	0.196	0.735	0.705	5.89E-6
M29	0.819	0.773	0.196	0.657	0.62	3.63E-6
M249	0.897	0.836	0.196	0.72	0.671	2.86E-6

In contrast to the default analysis procedure, the modified analysis procedure has a less chance to detect the true phenotype-associated variant in each sample. This results in a decrease of the combined 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 variants in each sample, only ~23% of those detected with the default analysis procedure. This decrease of confounding variants increases the confidence in the identification of phenotype-associated gene. As a result, the significance of the top candidate gene has increased from 1.911×10^{-6} to 9.453×10^{-7}, using the modified analysis procedure.

The modified analysis procedure reports only one candidate gene, *LOC_Os02g56510* (OsPHO1;2). This candidate is a functional homology of the Arabidopsis AtPHO1 gene(Secco, Wang et al. 2012). AtPHO1 was proposed to function as a Pi transporter (Secco, Wang et al. 2012)A recent study has demonstrated that AtPHO1 is a crucial downstream component of AtPHO2. AtPHO2 modulates the degradation of AtPHO1 in endomembranes to maintain Pi homeostasis in Arabidopsis(Liang, Wang et al. 2014). Therefore, it is highly likely that OsPHO1;2 is the gene that suppressed the *pho2* mutant phenotype in rice. The variants detected in OsPHO1;2 are listed in Table 6.

Table 6. Variants detected in *OsPHO1;2*.

Sample	Pos	Ref	Alt	GT	Effect
M28	34611907	C	T	1/1	Missense (A:V)
M29	34614585	C	T	1/1	Missense (H:Y)
M249	34614218	G	A	1/1	Stop_gained

To further validate this gene-phenotype association, we screened another 5000 lines from the EMS-induced mutant library based on the HNE8536 homozygous *Ospho2* mutant line. One additional mutant (M358), which exhibited the same Pi tolerance phenotype, was obtained. As expected, we found a high-effect mutation Ser340Gly in *OsPHO1;2* in M358.

The mutant phenotypes and genotypes of M28, M29, M249, and M358 are shown in Figure 5. The phenotypes of the mutants were further validated by measuring their shoot Pi concentrations, and confirming their *Ospho2* Tos17 insertion (data not shown).

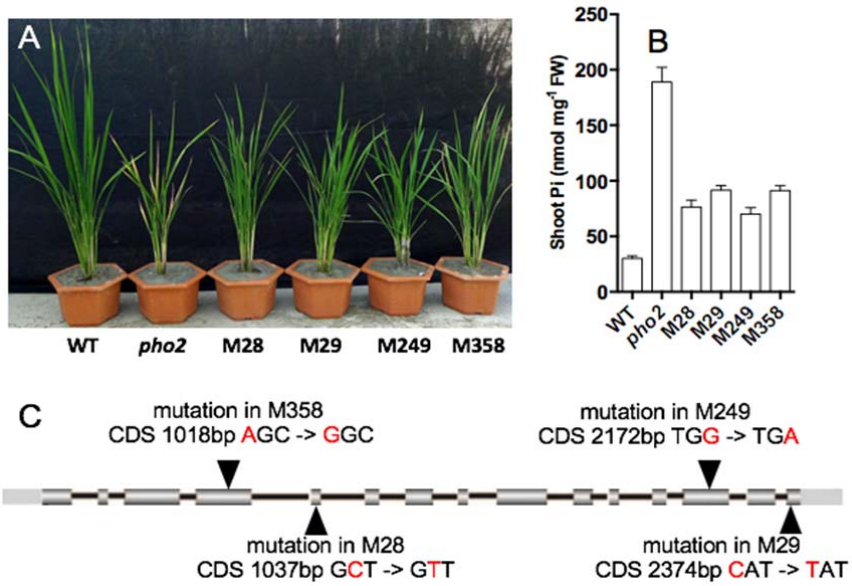


Figure 5. Plant phenotypes, Pi contents, and SNPs detected in the OsPHO1;2 region. (A) Phenotypes of 60-day-old plants grown in soil. WT: wild type Nipponbare; pho2: homozygous pho2 mutant (HNE8536). (B) Shoot Pi contents of 40-day-old plants grown in +Pi (200 μM) hydroponic medium (Error bars represent SE, n = 3). (C) Variations detected in the OsPHO1;2 region.

3.2.4 Example discussion (strategy to optimize an analysis procedure)

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

In the above example, the default analysis procedure reported 28 candidate genes (section 3.2.2), which are too many for validation. A more strict analysis procedure is therefore desired to identify candidates of higher confidence. For this reason, we optimized the analysis procedure toward accuracy and the optimized procedure identified only one candidate gene, which was subsequently confirmed with experiments. Although the optimized analysis procedure has a lower "chance to report the true phenotype-associated gene", it is not informative as long as this procedure can identify a biologically sound candidate that is worthy of validation. The measurement "chance to report the true phenotype-associated gene" is intended to provide guidance on the next steps when an analysis procedure cannot identify any biologically sound candidate or all identified candidates have failed validation.

To identify a phenotype-associated gene, a more "strict" or "accurate" analysis procedure will only consider highly confident variations that are supported by a significant number of high quality sequencing reads. Subsequent filtering process will also keep only variations that are very likely to produce the target phenotype. This "accurate" procedure will result in less number of possible variations that might produce the phenotype in each sample. Consequently, a candidate gene harboring variations in multiple samples will have a higher significance to associate with the phenotype. However, such an "accurate" procedure will also risk ignoring the true phenotype-causing mutations that are not supported by unequivocal evidence. This undesired ignorance can result in the true phenotype-associated gene harboring variations in an insufficient number of samples, which will fail the candidate gene criterion.

On the other hand, a more "permissive" or "comprehensive" analysis procedure will include more false positive variations and/or more less-likely phenotype-causing variations in each sample, which results in a higher chance that random genes may harbor variations in multiple samples. 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 more unlikely to render the true phenotype-associated gene failing the candidate gene criterion, such a procedure is more likely to report phenotype-unrelated candidates, which demands significant extra effort in their validations.

Therefore, to most investigators, it is recommended to use a more accurate analysis procedure whenever possible. Accurate procedures produce highly confident candidates, which minimize the chance of failure in candidate 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 is not discouraged by other evidence (e.g., evidence that some of its variations might be false positive calls), it is advised to validate this candidate first. In this scenario, the study effectiveness measurement "chance to report the true phenotype-associated gene" is not informative. If it is low, it just indicates that this analysis procedure happens to fit the need of identifying this candidate gene very well. On the other hand, if an analysis procedure produces many candidates, this procedure is probably not accurate enough. Investigators are advised to try more aggressive approaches to further increase the confidence of the reported candidates.

Though it is always more advisable to use a accurate analysis procedure as long as it can identify biologically sound candidate genes for validation, in cases that the protocol is too strict to identify any candidate or all identified candidates have failed validation, the two study effectiveness measurements, "chance to report the true phenotype-associated gene" and "significance of violating Mendelian assumption", may provide a guidance on the next steps. If the chance to report phenotype-associated gene is low, the analysis procedure is likely too strict and needs relaxation. Investigators may consider

validating more candidates reported by a more permissive analysis procedure, or sequencing more phenotype-exhibiting samples to increase the support of the phenotype-associated gene.

Conversely, if an investigator has sequenced a large number of phenotype-exhibiting samples and/or have validated many candidates, and the phenotype-associated gene is still at large, the study effectiveness measurement, "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 the investigator is still advised to add samples, validate more candidates, or to relax the analysis procedure to identify more candidates. If this significance is high ($p<0.05$), the investigator is advised to re-examine the phenotype-exhibiting samples included in the study.

For a qualitative trait, though it is possible that multiple genes can produce similar phenotypes, 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 produce dissimilar "side effects". Therefore, arguably, key to the success of a sequencing-based direct gene cloning study is the definition of a proper set of phenotype criteria, which can identify mutants of the same gene. The stricter the phenotype criteria are, the more likely that the included samples are mutants of the same gene. When GIPS reports a high significance of violating Mendelian assumption, investigators are advised to re-confirm the phenotypes of the samples included in the study. If there is no doubt, investigators are advised to consider using a more strict set of phenotype criteria for this study, which examine more minor phenotypic traits and can distinguish mutants of different but functionally related genes.

In general, the approach of sequencing based direct cloning in forward genetics study is expected to gain popularity. The reasons are two folds. First, this approach does not require the generation of cross or backcross populations, which significantly accelerates the gene identification process. Although this time advantage may come with more effort spent in screening mutant library to obtain multiple unrelated mutants of the same phenotype, library screenings in typical forward genetics studies require only bare-eye observation. The cost of screening a larger library is usually acceptable. Furthermore, rapid development of automated phenotyping technologies makes it possible to screen large libraries for minor phenotypes that cannot be easily observed(Pardo-Martin, Allalou et al. 2013). Second, because this approach does not require the creation of a population, it is free from the related limitations. This approach is readily applicable in identification of genes that are important in organ development (the disruption of which breaks the plant life cycle), and reproductive development (the disruption of which impedes crossing operation).

In this context, GIPS provides guidance on the effective design and execution of a sequencing based direct cloning study. It is different from other gene prioritization software, such as ANNOVAR (Wang, Li et al. 2010), which scores genes and variants to provide a rank. These priority scores do not advise, when no phenotype-associate gene can be identified, whether an investigator should change the analysis procedure, validate more candidate genes, add more samples, or re-examine the phenotype criteria used in the study. GIPS implements a probabilistic framework that models the entire process of sequencing based direct cloning study. Within this framework, other gene prioritization software focusing on removing genes/variants that are unlikely to associate with the phenotype can be integrated with the GIPS workflow as additional gene/variant filters.

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.

Delhaize, E. and P. J. Randall (1995). "Characterization of a Phosphate-Accumulator Mutant of Arabidopsis thaliana." Plant Physiology **107**(1): 207-213.

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.

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.

Liang, C., J. Wang, J. Zhao, J. Tian and H. Liao (2014). "Control of phosphate homeostasis through gene regulation in crops." Curr Opin Plant Biol **21**: 59-66.

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.

Pardo-Martin, C., A. Allalou, J. Medina, P. M. Eimon, C. Wählby and M. F. Yanik (2013). "High-throughput hyperdimensional vertebrate phenotyping." Nature communications **4**: 1467-1467.

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.

Secco, D., C. Wang, B. A. Arpat, Z. Wang, Y. Poirier, S. D. Tyerman, P. Wu, H. Shou and J. Whelan (2012). "The emerging importance of the SPX domain-containing proteins in phosphate homeostasis." New Phytologist **193**(4): 842-851.
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.

Wang, K., M. Li and H. Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Research **38**(16): e164-e164.

Abbreviations

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	Parameter file to specify the analysis procedure that will be executed and evaluated by GIPS	2.2.2
RESULT.gips	GIPS result file	2.2.3
CANDIDATE_GENES.txt	Detailed information about the variants found in candidate genes	2.2.3
Effective genomic region	Genomic regions that are considered by GIPS in gene identification	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 (when all samples are known unrelated)	1.2.2.3
Big difference filter	Discards non-synonymous variants that result in similar amino acids changes	1.2.2.2
Congestion filter	Discards a variant if there is another variant located 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