JARQ 55 (4), 333-339 (2021) https://www.jircas.go.jp

General Utilities for Genotyping Study (GUGS): A Comprehensive Application in Genotype and Sequence Data Manipulation

Tokurou SHIMIZU*

Division of Citrus Research, Institute of Fruit Tree and Tea Science, National Agriculture and Food Research Organization, Shimizu, Japan

Abstract

General Utilities for Genotyping Study (GUGS) is a toolbox for aiding the analysis of DNA marker data and its design in Microsoft Excel (MS Excel). GUGS provides more than 100 flexible functions for the manipulation, evaluation, and conversion of genotype data. It also provides functionalities for genotype format conversion to support linkage analysis using JoinMap software, frequency analysis for population genetics, parentage analysis, and statistical genetic analysis. Its functionality for the manipulation of nucleotide or amino acid sequences also assists DNA marker design. These GUGS features enable users to conduct all steps from DNA marker design to preliminary evaluation, data analysis, and format conversion for advanced study in a single environment without having to export/import data. GUGS is freely available at https://github.com/tokurou/GUGS under the GPL v3 license.

Discipline: Crop Science
Additional key words: bioinformatics, breeding, DNA marker, genetics, linkage analysis

Introduction

DNA marker analysis is a basis of modern genetic studies covering linkage analysis, parentage analysis, personal identification or forensic genetics, phylogenetic analysis, population genetics, and marker-assisted selection in breeding (Singh & Singh 2015, Goodwin et al. 2011, Nei 1987, Shimizu 2020). A set of genotype data is sorted in a two-dimensional marker-to-sample format, and this two dimenstional format works well with spreadsheet software. Many studies use Microsoft Excel (MS Excel) for managing genotype data or for exchanging data sets; however, MS Excel itself has no functionality for manipulating genotype or nucleotide sequence data. Some applications that enable the handling of genotype data in MS Excel have been developed (Chen et al. 2009, Peakall & Smouse 2012), but either focus on a particular analysis or are now outdated. There are excellent applications for data format conversion or genetic data analysis (Glaubitz 2004, Lischer & Excoffier 2012), and many R packages are now available for advanced analysis (Zhao & Tan 2006), but such products often require a different data set format for data importation and exportation, thereby imposing a time-consuming step and often hampering overall analysis performance. In this study, we developed GUGS to support a seamless manipulation of genotype data and DNA marker design in a single MS Excel environment. As a result, GUGS minimizes overall operation time and consequently enhances the performance of genotype data management.

General description

GUGS is a toolbox developed for MS Excel to achieve the seamless manipulation and analysis of genotype data, while also assisting in DNA marker design. More than 100 functionalities of GUGS have been implemented as functions. Users can perform any action by combining those functions with the built-in functions of Excel in any cell. The functions of GUGS are grouped under seven categories: data conversion, basic analysis, linkage analysis, data set analysis, frequency analysis, genetic data analysis, and sequence manipulation (Table 1). Each function performs a simple task, but in combination these functions cover wide fundamental analysis.

*Corresponding author: tshimizu@affrc.go.jp
Received 28 August 2020; accepted 7 December 2020.

Table 1. Categorized summary of General Utilities for Genotyping Study (GUGS) functions

Function category	Allowed data type	Functional classes
Data conversion	Genotype data (SSR, SNP, M, allele)	Normalization; conversion; allele size difference; SNP formatting
Basic analysis	Genotype data (SSR, SNP)	Allele selection; homozygosity test; genotype identity test; allele inclusion test; ploidy estimation; split genotype to allele; find the shared allele
Linkage analysis	Genotype data (SSR, SNP, M)	Estimates segregation mode for CP, BC1, or F2; converts genotype to the format for CP, BC1 or F2 segregation
Data set analysis	Genotype data (SSR, SNP, M, allele)	Counts the number of unique genotypes/alleles; separates a set of unique genotype or allele; ratio of matched genotypes/alleles by pairwise comparison of two data sets
Frequency analysis	Genotype data (SSR, SNP, M, allele)	Estimates frequency of a genotype/allele in a data set; observed (Ho) or expected heterozygosity (He); polymorphic information content (PIC); match probability (PM); the power of discrimination (PD); unbiased estimator of expected heterozygosity (GD, GD2); the probability of genotype match by Ukai
Genetic data analysis	Genotype data (SSR, SNP, M)	Allele sharing test; trio test; estimates the probability of obtaining a particular offspring from alleged parents, random mating, or a combination of an alleged parent with a given population according to Marshall (1998) and Jones & Ardren (2003)
Sequence manipulation	Nucleotide or nucleotide sequence (DNA, RNA)	Complementary, reverse or reverse complementary of sequence/nucleotide; splitting, formatting nucleotide sequence; splitting or extracting a nucleotide sequence; convert DNA to RNA or RNA to DNA; translates nucleotide sequence to amino acid sequence; counts nucleotide composition; GC ratio; motif search; matching score analysis

Environment

GUGS is implemented using Visual Basic for Applications (VBA) in Excel for MS Excel 2010, 2013, 2016, or Office 365 (upward compatible with Excel 2019). It is distributed as an MS Excel file implemented with GUGS VBA (GUGS.xlsm). No prerequisite step is required for its installation or launch, but users are requested to unlock VBA execution when launching GUGS, since MS Excel locks automatic VBA execution by default for security reasons.

Data type

Two codominant-type genotype formats—single-nucleotide polymorphisms (SNP) and simple sequence repeats (SSR)—and their alleles are acceptable for analysis (Table 2). An abbreviated single-letter genotype code (M) used in the popular MapMaker (Lander et al. 1987) and JoinMap software (Stam 1993) is acceptable with the code for segregation mode. Other types of codominant- or dominant-type genotype data are also used after transformation into an authorized genotype. A class of functions to support DNA marker design also accepts a nucleic acid sequence.

Functionalities

1. Data conversion

This supports the preliminary processing of genotype data, which is an essential step in a DNA marker study. The functional class “Norm” (NormSSR, NormSNP, or NormM) formats genotypes of SSR, SNP, or M to eliminate ambiguity through analysis. The function “SSRtoRelSize” converts the SSR genotype to a size relative to the reference genotype. The function “SSRDiff” returns the size difference of two SSR alleles. The function “SNPwithSEP” inserts or replaces the separator with the SNP genotype. The functions “M2SNP” and “HMP2SNP” convert the M genotype to SNP-like code or IUPAC-formatted single-letter code to the SNP genotype. Another function (“interpretSNP”) converts the outputs of GenomeStudio (Illumina) to the SNP genotype.

2. Basic analysis

This covers indispensable steps in most of the genotype data analysis, such as splitting codominant

Table 2. Data types available in General Utilities for Genotyping Study (GUGS)

Data type	Input	Output	Example
SSR genotype	O	O	100/110, 200/200 ...
SSR allele	O	O	100, 110, 200 ...
SNP genotype	O	O	A/G, CG ...
SNP allele	O	O	A, G, C, T
HMP	O	-	A single-letter genotype code for HapMap project
M	O	O	A single-letter genotype for MapMaker/JoinMap
CODE	O	-	A single-letter genotype for simplified SNP used in Illumina Genome Studio or other similar applications
Segregation mode	O	O	Segregation mode: BC1, F2, F1, or CP
Segregation code	O	O	Segregation code for JoinMap: BC1/F2: F2, BC1A, BC1B, F2D CP: <abxcd>, <efxeg>, <hkxhk>, <lmxll>, <nnxnp>
Genetic code	O	O	BC1/F2: a, b, h, - CP: a, b, c, d, e, f, g, l, m, n, p
Numeric	O	O	The allowed difference for SSR matching or returned value of various functions
Boolean	-	O	"TRUE" or "FALSE"
Nucleotide	O	O	Code of the deoxy-ribo nucleotide (A, C, G, T) or ribo nucleotide (A, C, G, U)
Amino acid code	O	O	A single letter code for amino acid
Nucleotide sequence	O	O	Nucleotide sequence of DNA or RNA
Any sequence	O	O	Any types of sequence (DNA, RNA, or amino acid)

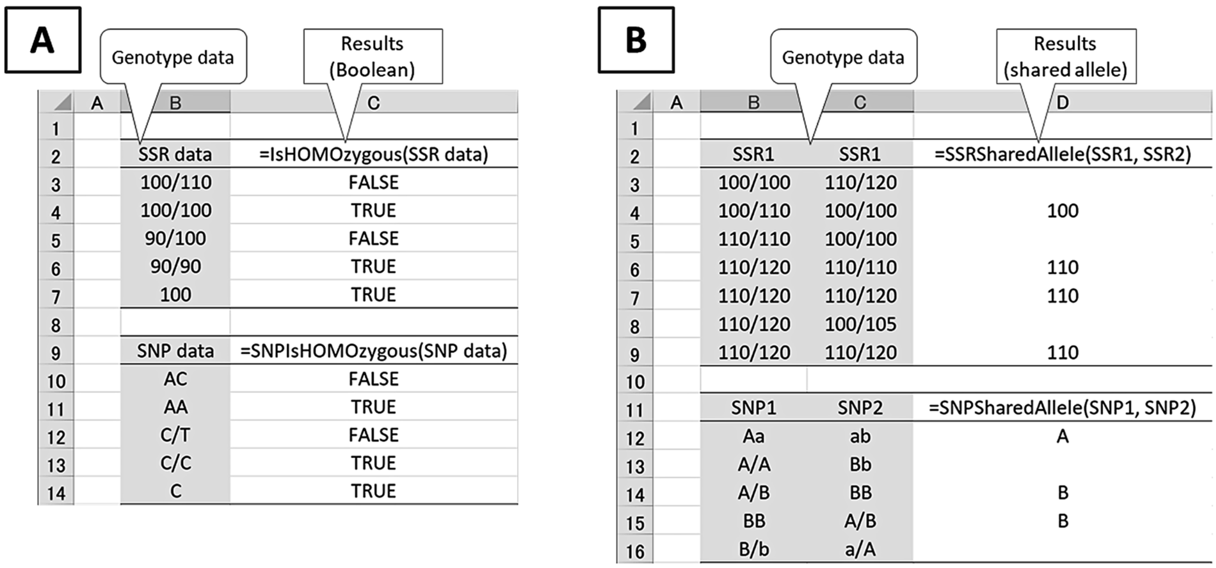
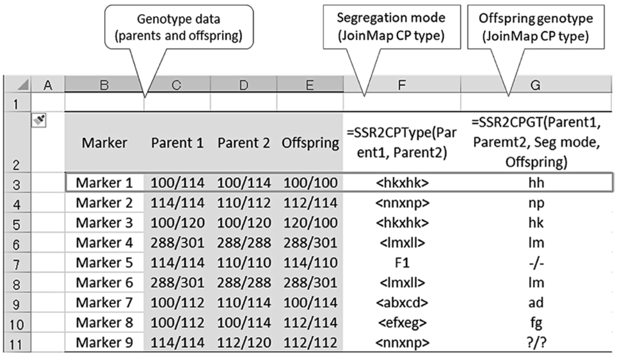


Fig. 1. Functionalities for basic testing of genotype data

A: Homozygosity test. IsHOMOzygous (simple sequence repeats (SSR) genotype) or SNPIsHOMOzygous (single-nucleotide polymorphisms (SNP) genotype) returns TRUE if the given genotype (column B) is homozygous.

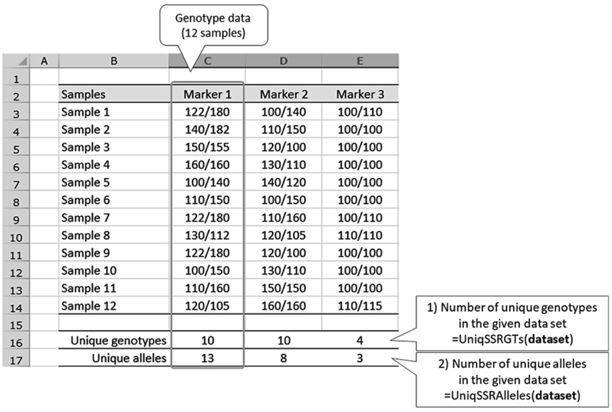
B: Allele sharing test. SSRSharedAllele (SSR genotype) or SNPSharedAllele returns the shared allele (column D) between two given genotype data (columns B and C). A smaller allele will be returned when two alleles are shared.

genotype to allele, extracting an allele, testing for homozygosity, genotype identity, allele inclusion, and finding a shared allele between two given genotypes (Fig. 1). The functional class "RightAllele" (RightAllele or SNPRightAllele) or "LeftAllele" (LeftAllele or SNPLeftAllele) returns the allele on either side. The functional class "IsHomozygous" (IsHOMOZygous or SNPIsHOMOZygous) examines homozygosity. The functions "IsSameSSR" and "SNPMatch" examine the identity of two given genotypes. The functional class "IsIncluded" (IsIncluded or SNPIsIncluded) examines whether the given genotype includes a designated allele.



Marker	Parent 1	Parent 2	Offspring	=SSR2CPType(Parent1, Parent2)	=SSR2CPGT(Parent1, Paremt2, Seg mode, Offspring)
Marker 1	100/114	100/114	100/100	<hkxhk>	hh
Marker 2	114/114	110/112	112/114	<nnxnp>	np
Marker 3	100/120	100/120	120/100	<hkxhk>	hk
Marker 4	288/301	288/288	288/301	<lmxll>	lm
Marker 5	114/114	110/110	114/110	F1	-/-
Marker 6	288/301	288/288	288/301	<lmxll>	lm
Marker 7	100/112	110/114	100/114	<abxcd>	ad
Marker 8	100/112	100/114	112/114	<efxeg>	fg
Marker 9	114/114	112/120	112/112	<nnxnp>	?/?

Fig. 2. Functionalities for linkage analysis of genetic data
Columns C-E: A set of trio genotype data for nine SSR markers. Column F: segregation mode according to CP mode of JoinMap software as estimated by the SSR2CPType function. Column G: genotype code of the offspring as converted by SSR2CPGT according to the segregation mode for CP mode. The gray box represents the data set and results for evaluation.



Samples	Marker 1	Marker 2	Marker 3
Sample 1	122/180	100/140	100/110
Sample 2	140/182	110/150	100/100
Sample 3	150/155	120/100	100/100
Sample 4	160/160	130/110	100/100
Sample 5	100/140	140/120	100/100
Sample 6	110/150	100/150	100/100
Sample 7	122/180	110/160	100/110
Sample 8	130/112	120/105	110/110
Sample 9	122/180	120/100	100/100
Sample 10	100/150	130/110	100/100
Sample 11	110/160	150/150	100/100
Sample 12	120/105	160/160	110/115
Unique genotypes	10	10	4
Unique alleles	13	8	3

Fig. 3. Sample of functionalities for data set analysis
Columns C-E: genotype data of 12 samples for three SSR markers. Row 16: number of unique genotypes obtained using the UniqSSRGTs function. Row 17: number of unique alleles obtained using UniqSSRAlleles. The gray box represents the data set and results for evaluation.

The functional class "SSRAllele" or "SNPAllele" splits all alleles in the given genotype as an array formula. The functional class "SharedAllele" (SSRSharedAllele or SNPSharedAllele) examines whether two given genotypes share the same alleles. Another function ("SSRPloidy") counts the ploidy from the SSR genotype that enables fast detection of genotyping error.

3. Linkage analysis

This converts raw genotype data to the genotype code for the linkage analysis software, and entails a simple yet laborious and confusing process. The functional class "CPType" (SSR2CPType or SNP2CPType) estimates the segregation mode as CP mode in JoinMap. Similarly, the functional class "SegType" (SSR2SegType or SNP2SegType) determines the segregation mode as BC1 or F2 mode in JoinMap (Fig. 2). Functional classes "CPGT" (SSR2CPGT or SNP2CPGT), "BC1GT" (SSR2BC1GT or SNP2BC1GT), or "F2GT" (SSR2F2GT or SNP2F2GT) individually convert the offspring genotype to the code according to CP, BC1, or F2 mode of JoinMap. For a single-letter genotype, the function "MWillSegregate" examines whether the parent genotypes will segregate, and "MSegregateType" determines the segregation mode useful to validate the converted genotype code. These functionalities help linkage analysis by automating segregation mode automation and genotype conversion.

4. Data set analysis

Counting the number of genotypes or alleles in a data set is the initial step for estimating the genetic distance, diversity, or selection process. The functional classes "UniqAlleles" (UniqSSRAlleles, UniqSNPAlleles, UniqMAlleles, and UniqAlleles) and "GetUniqAllele" (GetUniqSSRAllele, GetUniqSNPAllele, GetUniqMAllele, and GetUniqAllele) count the number of unique alleles in a given data set or return a unique allele in a data set individually (Fig. 3). Two other functional classes, "UniqGTs" (UniqSSRGTs, UniqSNPGTs, and UniqMGTs) and "GetUniqGTs" (GetUniqSSRGT, GetUniqSNPGT, and GetUniqMGT), count the number of unique genotypes or return a unique genotype in a data set individually. The functional class "MatchedRatio" (SSRMatchedRatio, SNPMatchedRatio, MMatchedRatio, or MatchedRatio) by pairwise comparison counts the ratio of matched genotypes between two data sets of the same size. And the functional class "SharedRatio" (SSRSharedRatio, SNPSharedRatio, or MSharedRatio) counts the ratio of shared alleles between two data sets of the same size by pairwise comparison.

5. Frequency analysis

Evaluating the frequency of an allele or a genotype is an essential step for detailed genetic analysis. Two functional classes, "AlleleFreq" (SSRAlleleFreq, SNPAlleleFreq, MAlleleFreq, or AlleleFreq) and "GTFreq" (SSRGTFreq, SNPGTFreq, or MGTFreq), evaluate the allele frequency or genotype frequency of the given data set (Fig. 4). This frequency analysis is also the basis to provide an overview of a data set or the performance of the DNA marker. GUGS also provides wide measures to obtain the scores for those evaluations (Fig. 4). Two functional classes, "Ho" (SSRHo, SNPHo, or MHo) and "HZ" (SSRHZ, SNPHZ, or MHZ), calculate

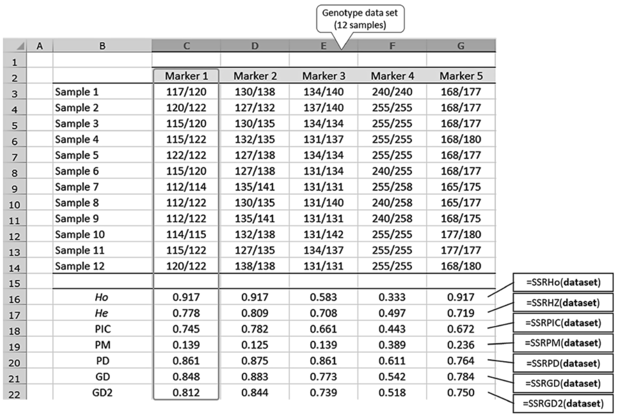


Fig. 4. Sample of functionalities for frequency analysis
Columns C-G: genotype data for frequency analysis obtained from 12 samples with five SSR markers. Rows 16-22: *Ho* (observed heterozygosity), *He* (expected heterozygosity), PIC (polymorphic information content), PM (match probability), PD (power of discrimination), GD (unbiased estimator of expected heterozygosity for a random population), GD2 (unbiased estimator of expected heterozygosity for a selfed population) as estimated individually by functions SSRHo, SSRHZ, SSRPIC, SSRPM, SSRPD, SSRGD, and SSRGD2.
The gray box represents the data set and results for evaluation.

observed heterozygosity (*H*o) and expected heterozygosity (*H*e) individually. Similarly, the functional class "PIC" (SSRPIC, SNPPIC, or MPIC) evaluates the polymorphic information content (PIC) of the DNA marker, and class "PM" (SSRPM, SNPPM, or MPM) evaluates the match probability (PM); class "PD" (SSRPD, SNPPD, or MPD) evaluates the power of discrimination (PD), and classes "GD" (SSRGD, SNPGD, MGD, or GD) and "GD2" (SSRGD2, SNPGD2, MGD2, or GD2) have functions to evaluate an unbiased estimator of expected heterozygosity for a random or selfed population (Goodwin et al. 2011, Nei 1987). A set of functions "Ukaif0" and "UkaiP1" estimates the probability of a particular individual occurring, and shows an identical genotype in a population according to Ukai (Ukai 2004).

6. Genetic data analysis

This provides statistical measures for estimating and assessing the parent-child relationship, which is the basis of forensic analysis, population genetic study, and parentage estimation (Marshall et al. 1998, Jones & Ardren 2003, Goodwin et al. 2011). The functional class "AlleleShared" (AlleleShared or SNPAlleleShared) identifies the common allele between the two given

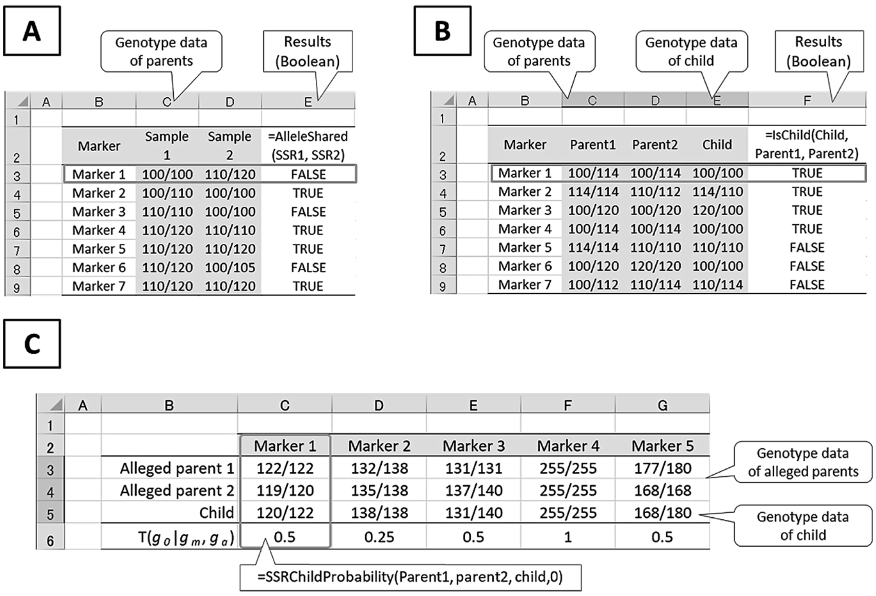


Fig. 5. Functionalities for genetic data analysis
A: AlleleSharing test function. Columns C and D: genotype data set of two samples for seven markers; Column E: AlleleShared function returns TRUE when two samples shared the same allele.
B: Trio test function. Columns C-D: parents genotypes; Column E: child genotype; Column F: IsChild function returns TRUE when parent and child genotypes satisfy as a trio.
C: Columns C-G: five SSR marker genotypes; rows 3-5: SSR genotypes for two alleged parents and child; row 6: probability of obtaining a child's genotype from the genotypes of alleged parents as estimated by the SSRChildProbability function.
The gray box represents the data set and results for evaluation.

genotypes (Fig. 5), which is the initial step for estimating parentage. The functional class "IsChild" (IsChild, SNPIsChild, or MIsChild) examines whether three given genotypes satisfy Mendel's law as a trio. For statistical evaluation of the proposed parentage, the functional class "ChildProbability" (SSRChildProbability, SNPChildProbability, or MChildProbability) returns the probability to obtain an offspring from two parental individuals corresponding to $T(g_0|g_m, g_a)$ of Marshall (1998). The functional class "GTProbability" (SSRGTProbability, SNPGTProbability, or MGTProbability) estimates the probability of obtaining a particular offspring from a random mating of a given population corresponding to $P(g_0)$ of Marshall (1998). And the functional class "ParentageProbability" (SSRParentageProbability or SNPParentageProbability) estimates the probability of obtaining a particular offspring from the mating of an alleged parent and a randomly selected alleged parent in a given population corresponding to $T(g_0|g_m)$ of Marshall (1998).

7. Sequence manipulation

The typical workflow of DNA marker design requires the trimming of a nucleotide sequence, scoring the sequence, surveying the motif or repeat sequence, and specifying a polymorphic sequence or nucleotide for the target of the DNA marker. Though most scientists manage the data obtained from certain software for individual purposes in MS Excel, such a process becomes complicated and troublesome due to the necessary formatting, exporting of data, and importing of the result in each step. GUGS allows users to directly manipulate nucleic or amino acid sequences in a single MS Excel environment, thereby eliminating the time that would otherwise be spent on formatting and transferring the data. GUGS supports basic and frequently used functions for the manipulation, evaluation, translation, motif analysis, and matching analysis of nucleotide sequence (Fig. 6). For sequence manipulation, GUGS supports returning a complementary sequence ("comp"), reverse sequence ("reverse"), or reverse-complementary sequence ("revcomp"). It also supports formatting of the nucleotide sequence ("splitseq," "fold," or "shrink"), marking a part of the sequence or separating the marked part ("bracket" or "prune"), and clipping the 5′ or 3′ end sequence ("clip5" or "clip3"). GUGS also supports showing the nucleotide composition ("composition") or GC content ("GCratio") for evaluation. In translation analysis, GUGS supports transforming DNA to RNA or vice versa ("DNA2RNA," "RNA2DNA," "toRNA," or "toDNA") and translating the DNA sequence to the amino acid sequence ("nuc2aa"). In motif analysis, GUGS provides a search and marking of the motif sequence ("motifcount," "firstmotif," "findmotif," or "markmotif"), and also supports two functions ("matchseq" and "matchscore") for matching analysis.

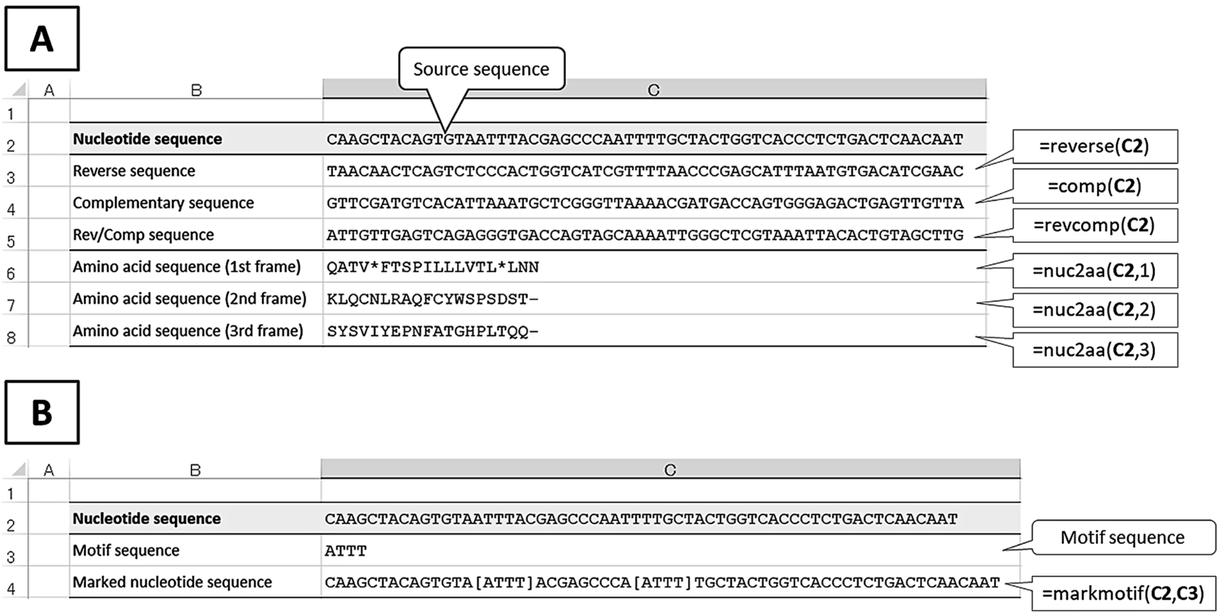


Fig. 6. Functionalities for nucleotide sequence manipulation

A: cell C2, a nucleotide sequence for evaluation. Cells C3-C5: reverse, complementary, or reverse-complementary sequence. Cells C6-8: translated amino acid sequence for first to third reading frame

B: cell C3, a motif sequence as a query. Cell C4: nucleotide sequence marked for the motif sequence with brackets

Application of GUGS

GUGS has been used to verify SNPs (Shimizu et al. 2016a), parentage estimation, and statistical verification, in order to estimate unidentified citrus pedigrees with SSR markers (Shimizu et al. 2016b). And with accurate genotype verification, GUGS has enabled genome-assisted selection by genome-wide association studies and genomic selection analysis (Minamikawa et al. 2017). Goto et al. (2018) also applied linkage analysis functions for developing a linkage map construction. Shimizu et al. (2020) used GUGS to verify the genetic identity of wild tachibana populations. The throughput of GUGS is sufficient for most analysis. For example, GUGS converts 10,000 SNP genotypes to the genotype for CP mode of JoinMap software within a few seconds (3.4 GHz Intel Core i7, Windows 10 PC with 32 GB memory).

Conclusion

GUGS provides a set of frequently used functionalities for manipulating genotype data, which must be processed with other applications. Briefly, GUGS reduces the effort of exporting and importing a data set for individual analysis, and automates a long and complicated data transformation process, thereby shortening the operation time and eliminating mistakes during data analysis. It also supports DNA marker design in a single environment. The functionality for linkage analyses is a unique feature of GUGS because no similar applications are available. Therefore, these features would play a major role in advanced studies. Though GUGS will work for typical analysis, it can extend new functionalities upon request. The GUGS source code is freely available under version 3 of the GNU General Public License (GPLv3) at https://github.com/tokurou/GUGS. Users are also encouraged to append new functionalities for individual purposes.

Acknowledgements

We are grateful to Dr. Keisuke Nonaka and Dr. Shingo Goto for evaluation of the preliminary version of GUGS. This work was supported by JSPS KAKENHI grant number 18K05634, and by the Government of Japan's Cabinet Office, under the Cross-ministerial Strategic Innovation Promotion Program (SIP), "Technologies for Smart Bio-industry and Agriculture" (funded by the NARO Bio-oriented Technology Research Advancement Institution; grant number DDB2001).

References

Chen, B. et al. (2009) SNP_tools: a compact tool package for analysis and conversion of genotype data for MS-Excel. *BMC Res. Notes*, **2**, 214.

Glaubitz, J. C. (2004) CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol. Ecol. Notes*, **4**, 309-310.

Goodwin, W. et al. (2011) An introduction to forensic genetics. Wiley, West Sussex, UK.

Goto, S. et al. (2018) QTL mapping of male sterility and transmission pattern in progeny of Satsuma mandarin. *PLoS ONE*, **13**, e0200844.

Jones, A. G. & Ardren, W. R. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.*, **12**, 2511-2523.

Lander, E. S. et al. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **1**, 174-181.

Lischer, H. E. L. & Excoffier, L. (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298-299.

Marshall, T. C. et al. (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**, 639-655.

Minamikawa, M. F. et al. (2017) Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. *Sci. Rep.*, **7**, 4721.

Nei, M. (1987) Molecular evolutionary genetics. Columbia University Press, New York, USA, pp. 514.

Peakall, R. & Smouse, P. E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics*, **28**, 2537-2539.

Shimizu, T. (2020) Genomic breeding. *In* Talon, M. et al. (eds.), The Genus Citrus. Elsevier, Duxford, UK, pp. 149-169.

Shimizu, T. et al. (2016a) A genomic approach to selecting robust and versatile SNP sets from next-generation sequencing data for genome-wide association study in citrus cultivars. *Acta Hortic.*, **1135**, 23-32.

Shimizu, T. et al. (2016b) Hybrid origins of citrus varieties inferred from DNA marker analysis of nuclear and organelle genomes. *PLoS ONE*, **11**, e0166969.

Shimizu, T. et al. (2020) Evaluation of genetic diversity in wild Tachibana population of Heda, Shizuoka, using DNA marker analysis, and stable maintenance of the population. *Hort. Res.*, **19**, 141-149 [In Japanese with English summary].

Singh, B. D. & Singh, A. K. (2015) Marker-assisted plant breeding: principles and practices 1st ed. Springer, India, New Delhi, pp. 451.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.*, **3**, 739-744.

Ukai, Y. (2004) A theory for varietal identification of plant cultivars. *Nougyou Oyobi Engei*, **79**, 194-198 [In Japanese].

Zhao, J. H. & Tan, Q. (2006) Integrated analysis of genetic data with R. *Hum. Genomics*, **2**, 258-265.