GIO – Genes In Order – A suite of scripts for identification and mapping of SNP markers in GBS data

Journal:	*The Plant Genome*
Manuscript ID	TPG-2015-10-0094
Manuscript Type:	Science Notes
Date Submitted by the Author:	01-Oct-2015
Complete List of Authors:	Skinner, Daniel; USDA-ARS, Crop and Soil Sciences Krishnan, Vandhana; Washington State University, Crop and Soil Sciences See, Deven; USDA-ARS and Washington State University, Plant Pathology
Keywords:	Genome sequencing, Genotyping by sequencing, Next generation sequencing

GIO – *G*enes *I*n *O*rder – A suite of scripts for identification and mapping of SNP markers in GBS data

Daniel Z. Skinner*, Vandhana Krishnan, and Deven R. See

D.Z. Skinner and D.R. See, U.S. Department of Agriculture, Agricultural Research Service, 209 Johnson Hall, Washington State University, Pullman, WA 99164, U.S.A. V. Krishnan and D.Z. Skinner, Department of Crop and Soil Sciences, 115 Johnson Hall, Washington State University, Pullman, WA 99164, U.S.A. D.R. See, Department of Plant Pathology, 345 Johnson Hall, Washington State University, Pullman, WA 99164, U.S.A. *Corresponding author (dan.skinner@ars.usda.gov).

Abbreviations: NGS, Next Generation Sequencing; SNP, Single nucleotide Polymorphism; GIO, Genes In Order.

Abstract

Next generation sequencing (NGS) generates billions of short DNA sequencing "tags" that can be used to identify genomic regions associated with trait determination through inheritance or genome-wide association studies. First, it is necessary to identify pairs of tags that differ by nucleotide base(s), tabulate occurrence among sample lines, and infer genomic locations in mapping populations. GIO provides a simple-to-use tool to carry out these functions through a simple graphical user interface. The output from an association panel study consists of the putative chromosome of origin and the presence/absence of each tag in the sample lines, and an indication of whether the tag was homologous to known expressed sequences. A separate table of tags that differed from a single other tag in the dataset by the user-specified number of SNP(s), and thus may represent alleles, is produced including the genotypes of the sample lines. The output from a biparental study also includes progeny genotypes, separated into low-copy and high copy-number tags. Two draft genetic maps are produced, one based on a user-specified number of genetic distance clusters, and one based on homology to previously-mapped sequences. Thousands of dominant tags typically are identified in biparental studies and GIO provides their putative chromosome of origin, and presence/absence in the progeny. A table of pairs of tags (both co-dominant and dominant) that cosegregate with at least one other tag in at least 65% of the progeny is also provided to facilitate mapping of the dominant tags. GIO is available from the first author.

5585 Guilford Rd., Madison, WI 53711

Keywords: Next generation sequencing, Computational genomics; Genotyping; mapping; Recombination; SNPs

Next generation sequencing has provided a means to quickly generate millions of short DNA sequence reads from study subjects with no required prior knowledge of the genome. The potential benefit of the application of these "tags" in population studies to link genetic variation to phenotypic variation is enormous. Variant forms of stretches of DNA that differ by a single nucleotide base (single nucleotide polymorphism, or "SNP") can be found throughout the genome, theoretically resulting in DNA sequence variation occurring within a short distance of all regulatory and coding regions (Sachidanandam et al., 2001). These SNPs can be used as biological markers, helpful to locating genes that are associated with phenotypic traits of interest (Lander and Botstein, 1989). The tags generated by high-throughput sequencing are a highly efficient means of identifying genomic regions containing SNPs.

In order to identify useful markers, the DNA sequence of each of the DNA tags generated by high-throughput sequencing must be compared to the sequences of the millions of other tags produced. Duplicate tags must be identified and removed, and useful variant forms identified in the appropriate proportion of individuals in the study. By comparing the tags from one parent to tags from the other parent in a biparental population study, tags representing co-dominant alleles may be discovered, as well as tags that do not have a corresponding alternate allele represented. The latter may be treated as dominant markers (Elshire et al., 2011). Both the co-dominant and dominant tags may be genetically mapped using co-segregation and traditional mapping functions (Elshire et al., 2011; Jiang and Zeng, 1997). This mapping effort is greatly assisted by comparing the DNA sequence of the tags to previously mapped sequences. Tags identified through association studies of large collections of individuals can be used for characterizing germplasm collections or disparate populations (Romay et al., 2013), and for defining marker-trait associations, for example, through a conditional probability approach (Skinner et al., 2000).

GIO Features

GIO is provided as a flexible tool to carry out the identification of dominant and co-dominant tags and their corresponding alleles within user-specified length, occurrence frequency, and read depth limits; preliminary mapping of codominant markers in bi-parental studies is provided. Association panel studies are supported and result in all of the unique tags found within user-

specified length and frequency limits. Chromosomes of origin of the tags are provided in both kinds of studies. The GIO scripts were developed using the PERL scripting language (http://www.perl.org/) and numerous tools available as part of the Linux operating system as well as additional open-source tools described below. GIO was developed using NGS results from wheat (*Triticum aestivum* L.), but should be applicable to any organism.

Implementation of GIO

1) Preprocessing of the data: GIO provides an optional tag-trimming step in which any bases 5' to a user-specified series of bases are removed and only tags starting with the specified sequence are retained. The intent is to remove bases 5' of the restriction site used to ligate adapters and barcodes in multiplexed sequencing, and to discard any tags that resulted from contamination or sequencing error at the start of the sequencing read. Also provided is an option to delete tags that contain regions of identical bases (homopolymers) of user-specified length. Homopolymeric regions have been shown to be potentially problematic in next-generation sequencing results (Quail et al., 2012).

2) Specification of population analysis parameters: All parameters are specified through a simple graphical user interface window (Fig. 1). Default values are provided and all settings are retained from each run of the scripts, such that user-defined parameters become the default settings for the next run. Basic help information is available from the "Help" button. Data files are assigned a numerical identifier for analysis purposes and for biparental populations, GIO must be told which data files represent the parents. A button is provided to allow the user to view the assigned numerical identifiers, thereby providing the identifiers of the parent files. This button also provides a check that the working directory and file type were specified correctly before starting the analysis. Options available to be specified are: length of tags, minimum frequency among lines, maximum frequency among lines, minimum and maximum number of SNPs per tag, and minimum read depth for a tag to be counted as present. The minimum and maximum frequency among lines parameter is provided to allow the user to screen out very common or very uncommon tags, especially useful when analyzing a segregating population with a 1:1 expected segregation ratio. Once the parameters have been specified, the analysis is initiated by pressing the "Submit" button (Fig. 1).

3) Analysis of a segregating population.

5585 Guilford Rd., Madison, WI 53711

Co-dominant tags: All unique DNA sequence tags meeting the specified parameters are identified; all tags retained for segregation analysis must occur in only one parent (zero frequency in the second parent), before the minimum read depth parameter is applied. Using SOAP2 (Li et al, 2009), all of these tags are then compared to reference sequence databases. The scripts were developed using a reference database from the International Wheat Genome Sequencing Consortium (IWGSC, 2014) including over 25,000,000 mapped sequences from all of the chromosome arms of wheat (*Triticum aestivum* L.), over 190,000 unique transcript sequences compiled by the first author (http://cereals.bioinfo.wsu.edu/wa_wheat_transcriptome), DNA sequences from the TREP database of transposable elements (http://wheat.pw.usda.gov/ITMI/Repeats), as well as chloroplast, mitochondria, and, ribosomal genes sequences. Tags homologous to the organellar, ribosomal, and transposon sequences, within a user-specified number of mismatches, are discarded. The putative chromosome arm of origin is then assigned to all remaining tags if significant homology to the reference sequences was found. If no significant homology was found, the tags are identified as being of "unknown" origin.

Next, of the tags that were retained, those that occurred in the first parent but not the second are compared to all tags from the second parent to identify pairs of tags that differ by the user-specified number of SNP(s), thereby identifying co-dominant markers. The presence/absence of pairs of tags identified in this way are then scored across all progeny lines and genotypes are recorded as homozygous for the first parent genotype, as homozygous for the second parent genotype, as heterozygous, or as unknown if neither parent tag was found.

Once the genotypes of the parents and progeny are identified, the tags are separated into two files. The first file holds "repetitive" tags; these are groups of tags in which one tag from one parent paired with multiple tags from the second parent, differing by the specified number of SNP(s). The second file is identified as holding "nonrepetitive" tags, pairs of tags in which a tag from one parent matched a single tag from the second parent, differing by the specified number of SNP(s). A table of read depth in the parents and all progeny is generated for each of the nonrepetitive tags. The nonrepetitive tags are then genetically mapped in two ways as follows.

***De novo construction of genetic maps of co-dominant tags*:** Usually, several thousand co-dominant, nonrepetitive tags are found. Attempting to genetically map the entire dataset is time-

consuming and requires human intervention to generate meaningful linkage groups. To provide a starting point for these linkage groups, GIO first utilizes clustering algorithms included in R (R Core Team, 2014) to cluster the markers on the basis of cosegregation. The user provides an estimate of the number of clusters that may be found, usually the number of chromosomes or chromosome arms in the organism; R uses this estimate as the clustering basis. By default, GIO uses Euclidean distance measures and clustering using Ward's method. GIO then reformats the segregation data from each cluster to MSTMap (Wu et al., 2008) input format and executes MSTMap to attempt to generate map(s) for each cluster using the Haldane mapping function. Three files are saved from each cluster; the original segregation data, a list of tag pairs and the map position within the cluster, a list of tag pairs that were included in the cluster on the basis of segregation, but were not genetically linked to the other tags in the cluster according to the MSTMap results. A combined file with linkage group and map position of all mapped markers also is provided to facilitate map image creation in other software.

Next, GIO compiles a dataset for each index that was provided to SOAP2 to search for homology to known genomic sequences. In our studies of the wheat genome, each chromosome arm is represented in a separate index so GIO builds separate datasets containing only tags with significant homology to one chromosome arm and tags with no significant homology to previously-mapped sequences. MSTMap is again called to generate map(s) from each of these data sets. From the output, GIO combines linkage groups that contain at least one previously-mapped tag into a single file to facilitate inspection of the linkage groups relative to each other.

Dominant tags: Tags found in only one of the parents with the user-specified read depth and frequency among the progeny lines, and for which no tags differing by the specified number of SNPs were found in the other parent, are considered to be dominant markers. GIO provides two output files, one with the sequences, chromosome assignments and occurrence among progeny of the dominant markers from the first parent, designated "Parent A", and a file with this information for the dominant markers from the second parent, designated "Parent B." It may be possible to include some of these tags on the co-dominant marker-based maps on the basis of co-segregation. To facilitate this kind of mapping, a file of "neighbor tags" is generated.

***Neighbor tags*:** Ideally, genetic maps would contain map locations for all available tags, both co-dominant and dominant. However, in our experience with wheat, the great majority of pairs of

5585 Guilford Rd., Madison, WI 53711

dominant and co-dominant tags do not co-occur in more than 50% of the progeny lines, and therefore show no evidence of being genetically linked. To facilitate identification and mapping of those markers that may be linked, GIO provides two files of "neighbor" tags, those that co-occur in at least 65% of the lines in the data set. The first, "MarkerNeighbors.txt," holds the genotypes of the markers in each of the lines designated as A, B, H or U for the co-dominant markers with genotypes of the first parent, second parent, both parents (heterozygous), or neither parent (unknown), respectively. The genotypes of the dominant markers are designated as C, D, or U, indicating the dominant marker from the first parent, second parent, or neither parent, respectively.

Progeny-only tags: In our experience with biparental, segregating populations, the majority of the unique tags found occur in the specified frequency of sample lines, but are not found in either parent. Clearly, these tags must occur in one or both parents, but were simply missed by the sequencing process. Optionally, GIO will tally the occurrence of these tags among the sample lines and generate a separate output file with the presence-absence data.

4) Analysis of an association panel.

Association panels may consist of a number (usually hundreds) of lines that may or may not be related. From a GBS perspective, the objective of association panel analysis is to identify DNA tags that are associated with trait(s) of interest. GIO provides the sequences, chromosome assignments, and frequency of occurrence of all of the tags found in the data set that meet the user-specified criteria of length, frequency and read depth. In addition, a separate data file listing tags that differ by the user-specified number of SNPs is produced. However, in our experience with association panels of hexaploid wheat, a given tag may differ from dozens of other tags by the allowed number of SNPs, all occurring with the allowed frequency among the plant lines. These differences probably are valid, but greatly complicate defining alternate alleles. Therefore, GIO provides a separate output file with pairs of tags that occur with only one "mate" in the data set, i.e. tags that differ by the allowed number of SNPs from exactly one other tag. Identifying the one-mate tags requires the tool agrep (https://github.com/Wikinaut/agrep) to be installed. These pairs of tags potentially represent alternate alleles at a single locus.

Example Results

Biparental population

To test GIO analysis of a segregating population, a dataset comprised of short reads generated by the Ion Proton sequencing system from the parents and 186 progeny from a cross of two winter wheat (*Triticum aestivum,* 2N=6X=42) cultivars was analyzed. There were a total of 90,137,056 raw sequence reads. The options specified to GIO were: Length of tags: 75, Minimum frequency among lines: 0.25, Maximum frequency among lines: 0.75, minimum number of SNPs per tag: 1, maximum number of SNPs per tag: 2, minimum read depth: 1, initial estimate of cosegregation clusters: 30. Any tags with homopolymers of five or more consecutive bases were deleted.

Co-dominant markers. The html summary report generated by GIO is shown in Table 1. Of the 19,837 pairs of tags with 1 or 2 SNPs, 4,326 were nonrepetitive, i.e. pairs of tags that differed from each other by one or two SNPs and by more than two from all other tags found. Of these 4,326 tag pairs, 289 were homologous to known genes. Working with these 4,326 tag pairs (each tag pair comprises a "marker") and the clusters found on the basis of cosegregation, GIO used MSTMap to find 58 linkage groups containing 3,354 markers. Next, using the datasets based on homology to previously-mapped markers, GIO found 2,564 markers in 428 linkage groups (about 10 linkage groups per chromosome arm), where each linkage group contained at least one marker homologous to a previously-mapped marker. These 2,564 markers were mapped to 8,282 locations, meaning that each marker mapped to an average of 3.2 chromosome arms. Because this mapping was in hexaploid wheat, a map location on three chromosome arms, one in each of the three diploid genomes, is not surprising. An example of the linkage groups found by GIO from wheat chromosome 1A is shown in Supplementary Figure S1.

Dominant tags. There were 45,982 apparently dominant tags found from the first parent, and 27,479 from the second. In the comparison of dominant tags to the 2,363 nonrepetitive codominant markers (reported by GIO as "neighbor" tags), 1,441,909 combinations were found that occurred simultaneously in at least 65% of the plant lines. These combinations involved all 2,363 co-dominant markers and 4,419 dominant tags. Most co-dominant markers were associated with a relatively small number of dominant tags. The number ranged from one to 2,993; the first quartile was 18, the median was 220, and the third quartile was 1,041.

Similarly, the dominant "neighbor" tags were each associated with relatively few co-dominant markers; the first quartile was 13, the median was 195, the third quartile was 587, and the

5585 Guilford Rd., Madison, WI 53711

maximum was 1,319. While the segregation data of all of the dominant tags are provided in the GIO output (files from the two parents are named ParentAdominantSNPS.txt and ParentBdominantSNPS.txt, respectively), our experience has shown that generating maps from the raw segregation data is essentially impossible with the mapping software we have tried. Thus, the "neighbor" tags are provided to help augment the genetic maps based on co-dominant markers; especially useful in regions of interest found through marker-trait association studies. For example, in the biparental mapping population described above, 335 co-dominant markers were mapped to chromosome 1A. Of those 335 markers, 211 were associated with at least one dominant tag in at least 65% of the plant lines.

Association Panel

To test GIO analysis of an association panel, a genotyping by sequencing approach was applied to a panel of 402 unique winter wheat entries with 607,325,655 individual sequence reads generated by the Ion Proton system. With the frequencies of accepted tags limited to 5-95% (i.e. excluding the very rare and the very common tags), 611,456 unique tags were found. Of these, there were 54,551 pairs of tags that differed by a single SNP from only one other tag in the dataset. The genotypes of each of the plant lines for these 54,551 tag pairs were reported as homozygous for the first or second tag, as heterozygous, or as unknown. The read depth, i.e. the number of times each tag occurred in each plant line, was determined and reported in a separate file. These results are reported as text files which may be used with other software for determining marker-trait associations.

Conclusions

GIO provides a highly customizable, simple-to-use tool that starts with raw sequence read files, cleans up the read data as specified by the user, and generates genotype, frequency and linkage group information with very little user intervention. Segregating populations and association panels are supported, resulting in comprehensive tabulation of tag occurrence data in simple text format, enabling downstream processing by other software to discover marker-trait associations.

Acknowledgements

This project was supported by USDA–ARS project 5348-21000-030-00D. Mention of product names does not represent an endorsement of any product or company but is given only to clarify the methodology; other products may be equally effective.

References

1. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011 May 4;6(5):e19379.
2. IWGSC: The International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science, 345(6194). http://doi.org/10.1126/science.1251788
3. Jiang, C., and Zeng, Z-B. 1997. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica. 101(1):47–58.
4. Lander, E.S., and Botstein, D.1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989 121(1):185–99.
5. Li, R., Yu, C., Li, Y., Lam, T-W., Yiu, S-M, Kristiansen, K., et al. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. Aug 1;25(15):1966–1967.
6. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., et al.2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 24;13(1):341.
7. R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
8. Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, T.M., et al. 2013.Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology. 11:14(6):R55.
9. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409(6822):928–33.
10. Skinner, D.Z., Loughin, T., Obert, D.E. 2000. Segregation and conditional probability association of molecular markers with traits in autotetraploid alfalfa. Molecular Breeding. 6(3):295–306.

5585 Guilford Rd., Madison, WI 53711

11. Wu, Y., Bhat, P.R., Close, T.J., Lonardi, S. 2008. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet. 2008 Oct 10;4(10):e1000212.

5585 Guilford Rd., Madison, WI 53711

Figure Caption

Figure 1. The graphical user interface (GUI) used to interact with GIO. All options are specified through this one GUI.

Supplemental Figure S1. Example of the linkage groups found by GIO from wheat chromosome 1A. The genetic distances were computed by MSTMap. The underlined numbers are the genetic positions described in the IWGSC mapping effort of Chinese Spring for markers that were homologous to the markers found by GIO and identified on the map. The chromosome arm designation following a marker name indicates significant homology of the corresponding marker to sequence(s) from the indicated chromosome was found in comparisons to the Chinese Spring data from the IWGCS mapping effort. The images were drawn using MapChart (Voorrips, R.E., 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. The Journal of Heredity 93 (1): 77-78.)

Table 1. Summary of tag, polymorphism, genotype information, and runtime specifics as output by GIO.

Type of tag	Tag Count	SNP	SNP Frequency	Genotypes among progeny	Genotype Frequency
Initial Unique Tags of 75 bases	18,263,477	[K] (G/T)	8,063	A (1st parent)	927,972
Tags meeting frequency criteria	189,740	[M] (A/C)	9,561	B (2nd parent)	884,747
Tags from 1st parent with SNP(s) relative to 2nd parent	9,405	[R] (A/G)	16,957	H (heterozygous)	534,889
Tags from 2nd parent with SNP(s) relative to 1st parent	7,745	[S] (C/G)	10,645	U (unknown)	1,381,748
Tag pairs with SNP(s)*	19,837	[W] (A/T)	6,477	**Codominant**	
Dominant tags,1st parent	45,982	[Y] (C/T)	15,913	**Polymorphism**	**Count**
Dominant tags, 2nd parent	27,479			SNPs	18,478
Tags in progeny only**	99,129			INDELs	1,359

*Tags from one parent may pair with more than one tag from the second parent.

**Tags that met the length and frequency criteria but were not found in either parent.

Runtime specifics
The total number of sequence read files found was 188.
The parents were represented in files numbered 1 and 2.
The runtime options were:
Length of tags: 75
Minimum frequency among lines: .25
Maximum frequency among lines: .75
Minimum number of snps per tag: 1

5585 Guilford Rd., Madison, WI 53711

Maximum number of snps per tag: 2
Minimum number of times a tag had to occur in a sequence file to be included (accepted as "real"):1

The number of independent processes (approximately the number of cores used) was: 30
SOAP was used to search for homology to the sequence databases in: **/home/dzs/GBS/Kim/SNPscripts/GIO/Filters** allowing 1 mismatch(es) per tag.
Tags homologous to sequences in Repeats.fa were discarded.
Tags homologous to sequences in wwt.fas were identified as putative genes.
Adjacent base transposition was counted as a single SNP.

5585 Guilford Rd., Madison, WI 53711

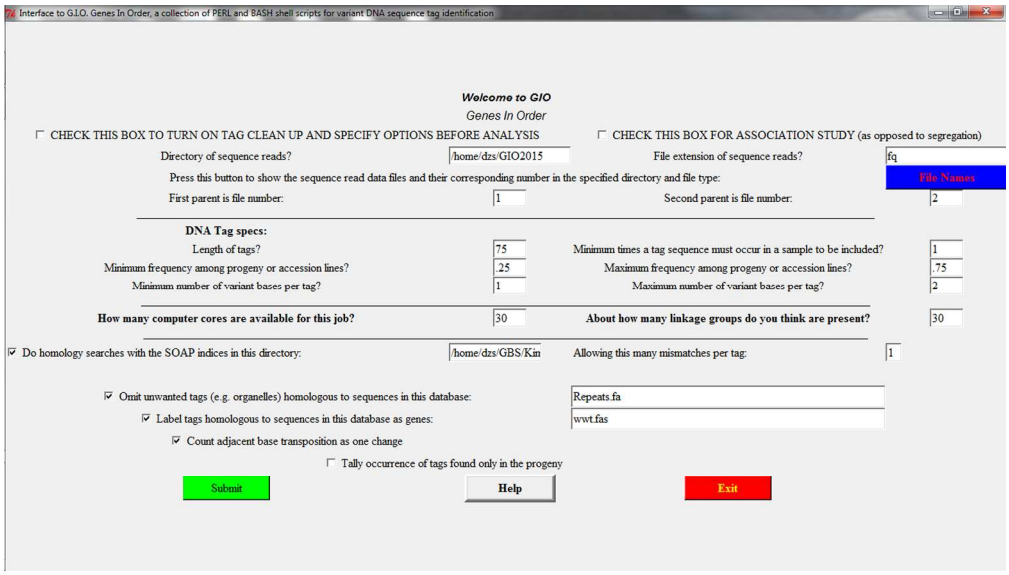
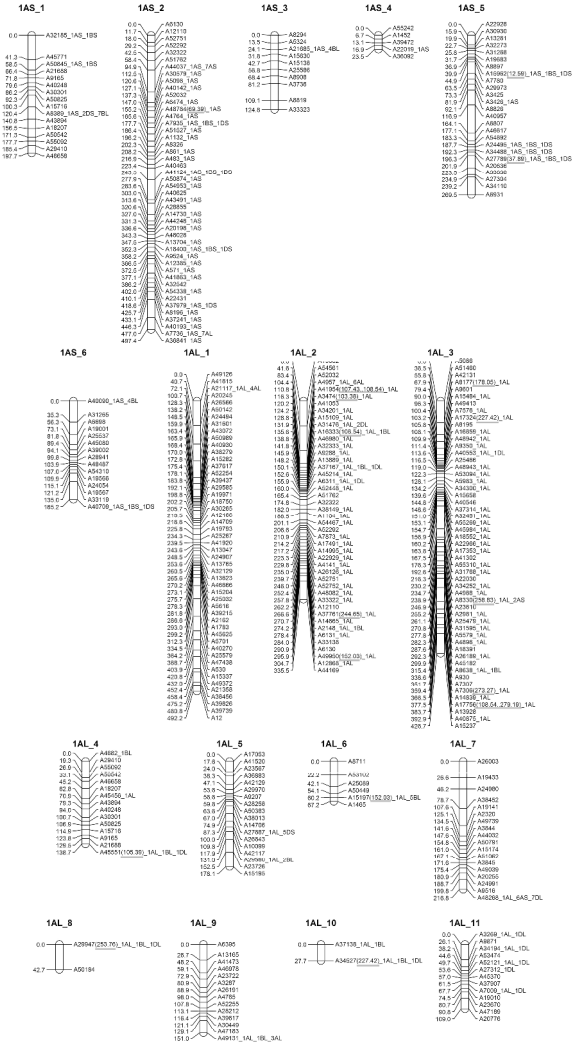


Figure 1. The graphical user interface (GUI) used to interact with GIO. All options are specified through this one GUI.
479x270mm (300 x 300 DPI)



Supplemental Figure S1. Example of the linkage groups found by GIO from wheat chromosome 1A. The genetic distances were computed by MSTMap as part of the GIO process. The underlined numbers are the genetic positions described in the IWGSC mapping effort of Chinese Spring for markers that were homologous to the markers found by GIO and identified on the map. The chromosome arm designation following a marker name indicates significant homology of the corresponding marker to sequence(s) from the indicated chromosome was found in comparisons to the Chinese Spring data from the IWGCS mapping effort. The images were drawn using MapChart (Voorrips, R.E., 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. The Journal of Heredity 93 (1): 77-78.)

304x632mm (300 x 300 DPI)