**Bioinformatics Advance Access published August 16, 2016**

*Phylogenetics*

# DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

Cristina Frías-López[1,2,†], José F. Sánchez-Herrero[1,†], Sara Guirao-Rico[1,3], Elisa Mora[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,‡,*], and Julio Rozas[1,‡,*]

[1]Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.[2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.[3]Current affiliation: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors

Associate Editor: Prof. Alfonso Valencia

## Abstract

**Motivation:** The development of molecular markers is one of the most important challenges in phylogenetic and genome wide population genetics studies, especially in studies with non-model organisms. Highly promising approach for obtaining suitable markers are the utilization of genomic partitioning strategies for the simultaneous discovery and genotyping of a large number of markers. Unfortunately, not all markers obtained from these approaches provide enough information for solving multiple evolutionary questions at a reasonable taxonomic resolution.

**Results:** We have developed DOMINO, a bioinformatics tool for informative marker development from both NGS data and pre-computed sequence alignments. The application implements popular NGS tools with new utilities in a highly versatile pipeline specifically designed to discover or select personalized markers at different levels of taxonomic resolution. These markers can be directly used to study the taxa surveyed for their design, utilized for further downstream PCR amplification in a broader set taxonomic scope, or exploited as suitable templates to bait design for target DNA enrichment techniques. We conducted an exhaustive evaluation of the performance of DOMINO via computer simulations and illustrate its utility to find informative markers in an empirical dataset.

**Availability:** DOMINO is freely available from www.ub.edu/softevol/domino.

**Contact:** elsanchez@ub.edu, jrozas@ub.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

It is well known that phylogenetic inferences based on a single or very few genetic markers can lead to systematic errors and reach invalid conclusions (Maddison *et al.*, 1997; Brito and Edwards, 2009). Next generation sequencing (NGS) has become a feasible and cost-effective way of obtaining large amounts of genetic markers suitable for addressing ecological and evolutionary questions. Among current methodologies, the hybrid enrichment and the reduction representation sequencing methods (for a review see Lemmon and Lemmon, 2013) are particularly promising approaches for studies in non-model organisms. Markers developed with these methodologies, however, may not be informative

enough to resolve multiple evolutionary questions across a reasonable taxonomic range; indeed, some markers may be inefficient for a particular study in a specific taxonomic group, or can be useful only for limited phylogenetic ranges. These problems make often necessary to accomplish various cost-intensive enrichment or reduction representation experiments to further obtain makers suitable to be applicable across a wide range of species.

Recently, some optimizing approaches have been developed to try to overcome this limited marker informativeness. For instance, the *MarkerMiner* 1.0 pipeline (Chamala *et al.*, 2015), outputs different types of multiple sequence alignments (MSA) files, some of them including reference coding sequences containing introns, which facilitates the downstream evaluation of the phylogenetic utility of each marker or the prediction of intron-exon boundaries and intron sizes, very useful for primer or probe of development. Nevertheless, the pipeline does not perform these assessments by itself and the application is specifically devised to work only with transcriptome assemblies and with a set of plant reference genomes. Indeed, the possibility of selecting particular markers with a specific number of samples has been recently implemented in the RAD-Seq data processing pipeline *RADIS* (Cruaud *et al.*). However, this application does not include other key options and parameter combinations, such as the selection of a specific nucleotide variation range across a set of pre-defined taxa, options that can be very useful for a plethora of studies. *BaitFisher* (Mayer *et al.*, 2016) also implements a novel approach to optimize the design of target enrichment baits to be applicable across a wide range of taxa. This software includes an algorithm to infer target DNA enrichment baits from multiple taxa by exploiting user-provided nucleotide sequence information of target loci in a representative set of species and can handle both genomic and cDNA data. Nevertheless, this software works on the basis of MSA of already known target loci that directly serves as templates for bait design (i.e., it cannot be used with raw NGS data or for *de novo* marker discovery).

Here we present DOMINO (Development Of Molecular markers In Non-model Organisms) a new bioinformatics tool that facilitates the development of highly informative markers from different data sources, including raw NGS reads and pre-computed MSA in various formats (such as those from RAD data). DOMINO efficiently process NGS data or pre-computed MSA and identifies (i.e., *de novo* discovery) or selects the sequence regions or alignments that meet user-defined criteria. Customizable features include the length of variable and conserved regions (when requested), the minimum levels (or a preferred range) of nucleotide variation, how to manage polymorphic variants, or which taxa (or what fraction of them) should be covered by the marker. All these criteria can be easily defined in a user-friendly GUI (Graphical User Interface) or under a command-line version that implements some extended options and that it is particularly useful for working with large NGS data sets in high performance computers (Supplementary Fig. S2; see also the DOMINO documentation). The regions identified or selected in DOMINO can be i) directly used as markers with a particular depth of taxonomic resolution, ii) utilized for their downstream PCR amplification in a broader taxonomic scope or iii) used as suitable templates to optimized bait design for target DNA enrichment techniques.

## 2 Methods and Implementation

### 2.1 DOMINO workflow

The DOMINO workflow consists in four main phases (Fig. 1) that can be run either using the DOMINO GUI or the extended command-line version (see the DOMINO manual in the DOMINO Web page). In both cases, the most relevant results from each phase are conveniently reported in the appropriate output files.

*Input data and pre-processing phase*. DOMINO accepts input sequence data files in two different formats, the 454 Pyrosequencing Standard Flowgram Format (SFF), and FASTQ format (Cock et al., 2010). These input files can contain 454 or Illumina (single or paired-end) raw reads from *m* taxa (the "taxa panel"). The sequences from each taxon should be properly identified with a specific barcode (aka, tag, MID or index), or loaded in separate files, also appropriately named (see the DOMINO manual in the DOMINO Web site for details). DOMINO is designed to filter low quality, low complexity, contaminant and very short reads using either default or user-specified filtering parameters. Mothur, PRINSEQ, NGS QC toolkit, BLAST, as well as new Perl functions specifically written for DOMINO (DM scripts) are used to perform these tasks (Supplementary Table S1). DOMINO uses Mothur v1.32.0 (Schloss et al., 2009) to extract reads from SFF files and store them in FASTQ format, which are subsequently converted to FASTA and QUAL files. Low quality or very short reads are trimmed, or definitely removed, using NGS QC Toolkit v2.3.1 (Patel and Jain, 2012). PRINSEQ v0.20.3 package (Schmieder and Edwards, 2011) is used to eliminate low complexity reads using the implemented DUST algorithm. Putative contaminant sequences, such as bacterial DNA frequently found in genomic samples (Leese et al. 2012), cloning vectors, adaptors, linkers, and high-throughput library preparation primers, can also be removed using a DOMINO function that performs a BLAST search (BLAST v2.2.28) (Altschul et al., 1990) against UniVec database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) and/or against a user-supplied contaminant database (see the DOMINO manual).

*Assembly phase*. When working with just NGS reads, the program first applies an assembly-based approach; the pipeline is therefore optimized to work with genome partitioning methods in which the length of the size-selected (or enriched) fragments and the sequencing depth are enough to permit the assembly of a set of homologous fragments. For data from restriction-site associated DNA (RAD) sequencing and related methods see the Mapping/Alignment phase section. DOMINO performs separate assemblies, one for each panel taxon, using MIRA v4.0.2 (Chevreux, 1999), either with the pre-processed reads from the previous step or with those supplied by the user. Although the default parameter values vary in function of the particular sequencing technology, the majority of them are shared (see the DOMINO manual). In order to avoid including repetitive and chimeric regions, all contigs (and the corresponding reads) identified as repeats in the MIRA algorithm are discarded from the mapping/alignment phase (Chevreux, 1999). Since MIRA can generate redundant contigs because of polymorphic and paralogous regions, we have implemented a specific DOMINO function that performs a clustering of all contigs based on an all versus all contigs BLAST search to identify and remove such redundancies. The DOMINO command line version (see below) also includes an option to perform a second iterative assembly step using the software CAP3 (Huang, 1999). If selected, this option uses MIRA output sequences (contigs and singletons) as input for CAP3 under a relaxed parameter scheme.

*Mapping/Alignment phase*. DOMINO uses *Bowtie2* (Langmead and Salzberg, 2012) to map the pre-processed reads from each taxon to the assembled contigs of the other *m*-1 taxa from the panel. Thus, in this step, DOMINO builds *m*(*m*-1) sequence alignment/map files (SAM/BAM files. In the case of a panel of *m* = 4 taxa, for example,

DOMINO will build 4 x 3 = 12 SAM/BAM files during this step. The reason behind this particular mapping strategy lies in the dissimilar performance of alignment/mapping algorithms depending on the divergence between the reads and the reference sequences. Immediately after generating BAM files, DOMINO removes all unmapped contigs and multi-mapping reads. This step is critical to avoid alignment artifacts, which can create false positive markers (i.e., sequence regions with misleading high levels of nucleotide diversity). The contigs with an unusually large number of aligned reads, which can correspond to repetitive regions, are also removed (they are not suitable for designing single copy markers) Later, DOMINO will build one pileup file per each BAM/SAM file using the SAMtools v0.1.19 suite (*mpileup* option) (Li et al., 2009).

Since sequencing errors might have a great effect on the marker selection, DOMINO incorporates their own functions for detecting and masking putative sequencing errors, which apply a very conservative criterion for variant calling. First, to avoid the calling of spurious nucleotide variants in low sequencing coverage experiments (i.e. erroneously assigned variants fixed between the taxa from the panel), DOMINO mask the information from positions with only one read mapped to the reference. Furthermore, sequencing errors may also inflate the number of called polymorphisms under the Polymorphic Variants option in the marker identification/selection phase. To avoid such undesirable effect, DOMINO incorporates a similar conservative criterion to use only highly credible polymorphisms. Under the Polymorphic Variant option, DOMINO will assume that each taxon represents a diploid individual; for positions with eight or more reads mapped, DOMINO discards those polymorphic variants in which the frequency of the minor allele is significantly lower than the expected under error free data (hence, in absence of sequencing errors the distribution of observed nucleotide counts at each position would follow a binomial distribution). For lower coverage values, DOMINO will use the information of a polymorphic variant only if the allele with the minor frequency is present in two or more reads. This testing procedure, applied independently for each position within each species, will likely discard some true polymorphic sites; this variant calling approach, however, makes DOMINO highly conservative in detecting true markers when including polymorphisms in the analysis (i.e., DOMINO will use only highly confident within-species segregating variants for the marker Discovery/Selection phase). Ambiguity codes, either introduced by MIRA assembler in contig sequences or present in user-supplied reference sequences or MSA, are also considered by DOMINO to decide whether a position is or not variable.

After applying all the above-mentioned post-mapping filters, DOMINO combines the variation profiles (arrays with the information about the state of each position, conserved or variable between taxa pairs) obtained from each of the *m*-1 pileup files including the same reference sequence (i.e., the same taxon), into a single multiple taxa variation profile (MTVP). Since each of these references will be likely fragmented in *i* contigs, DOMINO will build *i* x *m* MTVP per taxon. Each of these MTVP will be independently scanned for regions containing candidate markers in the next phase. If the user provides reference sequences from a single taxon (e.g. a genome draft), plus the reads from the *m* different taxa, the program builds only one MTVP set (one per contig or scaffold in the supplied reference). On the other hand, if the input includes a single or multiple pre-computed MSA instead of NGS data, DOMINO skips the alignment/mapping phase and directly generates the single MTVP set (one per aligned region). In this point, the program accepts MSA files in FASTA (multiple FASTA files, one per linked region), PHYLIP (multiple PHYLIP files, one per linked region, or one multi PHYLIP file with the alignment of all regions) and pyRAD LOCI (*.loci files generated by the program pyRAD; Eaton, 2014) and STACKS fasta (batch_X.fa output files generated from the population analyses in the program STACKS; Catchen *et al.*, 2011) output files.

*Marker Discovery/Selection phase.* Each MTVP generated in the previous step is either scanned for the presence of candidate marker regions using a sliding window approach (DOMINO marker discovery module), or used to select markers (with the desired features) among the MSA loaded in the previous tab (DOMINO marker selection module). In the first case, a specific DOMINO function searches for sequence regions of desired length (Variable region Length, VL), showing the minimum level of variation indicated by the user (Variable region Divergence, VD). DOMINO can also restrict that this variable region was flanked (or not) by highly conserved regions (Conserved region Divergence, CD) of a predefined length (Conserved region Length, CL); an information useful to further design PCR primers. Moreover, DOMINO can strictly restrict the search to a particular set of taxa (from the panel), or just specify the minimum number of taxa required to be covered by the marker (by changing the Minimum number of Covering Taxa parameter; MCT < *m*). As indicated, DOMINO can use or not the information from polymorphic sites. An appropriated combination of selected taxa and MCT and VD parameter values will allow the user select a large set of informative markers suitable to be applicable across a wide range of taxa. In the second case, the DOMINO selection module allows directly selecting the most informative markers among the loaded by the user in the same way and with the same personalized features described above. For RAD loci, a particular range of variable positions (VP) between the closest taxa (instead of the VD parameter) must be specified. This option allows selecting informative RAD loci while excluding those exhibiting anomalous high levels of variation, which might reflect RAD tag clustering errors. The specific selection of a set of loci/MSA that meet some specific phylogenetic criteria using the DOMINO selection module can be very helpful to further design probes for different target enrichment techniques, including the enrichment of specific RAD segments using hyRAD (Suchan *et al.*, 2016).

After the last phase, DOMINO reports the list the genomic regions (and their coordinates) or MSA that meet the selection criteria, along with the corresponding MSA of these regions for the selected taxa. Since DOMINO can work with more than one MTVP set (*m* in a full DOMINO run), some of the markers found in MTVP based on different reference taxa may be redundant (they can cover the same genomic region, although with different coordinates; see Mapping/Alignment phase section), while other can be found only in one particular profile. To avoid reporting redundant information, we have implemented a BLAST-based function to collapse these maker sequences, only reporting unique markers. To maximize the probability of finding informative markers, the final list of candidates under the DOMINO marker discovery module can include overlapped regions that fulfill the specified characteristics. Operationally, all regions that meet the criteria for being considered a candidate marker (after moving the scanning window five or more base pairs) are listed as different markers in the final output. In this way, users can choose the best marker to be used directly for further analyses or the more appropriated region of each contig to be PCR amplified and sequenced in additional focal species (i.e., the best marker from each linked block).

### 2.2 DOMINO GUI

DOMINO can be run either in the command prompt, by setting a large set of command line options, or using the GUI specifically developed to facilitate its use to non-experts in NGS bioinformatics tools. (Fig. 2; see also the DOMINO manual for details). The DOMINO GUI is a cross-platform application that allows the user to interactively set marker selection criteria by tuning the most important parameters and options available in the command prompt version. It should be noted that for huge NGS data sets (which require substantial amounts of computational resources) a full DOMINO run using the GUI version is not recommendable. In this case, the user can either run DOMINO under the command line version using high performance computer clusters or, take advantage of the custom run options available in the GUI version to enter in DOMINO partially processed data, e.g. pre-processed reads, assemblies or alignment files (SAM/BAM) obtained with other memory-efficient software (Supplementary Table S2).

### 2.3 System and Availability

The GUI was built using the cross-platform library and user interface framework Qt (https://www.qt-project.org/) based on C++ scripting language. Since most of the functions specifically developed for this work are implemented in Perl scripting language, users need to install first a recent version of Perl (version 5.12 or higher; http://www.perl.org/). The source code, the documentation and some example data files are freely distributed under the GNU GPL software license at: http://www.ub.edu/softevol/domino.

## 3 Results and Conclusions

### 3.1 Computer simulations

We conducted an exhaustive computer simulation study to assess the performance of DOMINO in detecting informative markers (i.e., simulated regions that meet specific marker selection criteria) from NGS data. For that, we emulated an RRL-like experiment of four closely related species exhibiting different levels of nucleotide divergence among them and incorporating substitution rate heterogeneity across sites to create genuine informative markers. The topology of the species tree used for the simulations was fixed (Supplementary Fig. S1). In each replicate, we generated an independent RRL-like data set of 100 fragments, of different length (3 kb or 10 kb) each. The nucleotide sequences were simulated with the program evolver, included in the PAML v4.7 package (Yang, 1997, 2007), using 0.1, 0.15, 0.20 or 0.30 substitutions per site between the two most divergent sequences, under the Jukes and Cantor (1969) substitution model with substitution rate heterogeneity across sites (modeled as a discrete gamma with 10 categories and $\alpha$=0.01). For each replicate, we simulated a complete NGS experiment in the Roche-454 (reads with an average length of ~400 bp), and the Illumina HiSeq2000 platforms (average length of 101 bp; single and paired-ends) using the ART v2.5.8 program (Huang et al., 2012) with default parameters and three different sequencing coverage values (5X, 10X and 20X). We generated 500 simulation replicates for each of the 48 possible scenarios (i.e., for each combination RRLs fragment length, divergence, sequencing platform and coverage values), resulting in a total 27,000 DOMINO runs, which took roughly 80,000 CPU hours.

Using the DOMINO marker discovery module under the command line version, we first traced the number and the location of the regions that meet the selection criteria present in each simulated fragment previous to emulate their NGS sequencing (true markers; TNM). Subsequently, for each data set, we execute a full run of our program using the simulated NGS reads to obtain the list of candidate markers (detected markers; DNM) for each scenario. For this experiment, we define an informative marker as a variable region of 600 bp (VL = 600), present in all four species (MCT = 4), showing at least 0.01 nucleotide substitutions per site between any pair of species (VD = 0.01), and flanked by two highly conserved regions of 60 or more bp long (CL = 60; only one substitution across species was permitted; CD = 1). We assessed the performance of DOMINO in detecting the TNM by measuring the sensitivity and precision in each replicate and plotting their distribution across the 500 replicates (Fig. 3; Supplementary Information).

We found that DOMINO pipeline has a high sensitivity in detecting the existing TNM, yielding averages of true positive rates values > 0.9 for Illumina reads and when coverage values are equal or higher than 10X (Fig. 3). As expected, lower coverage values (5X) result in a reduction of the sensitivity estimates; in this case, DOMINO runs using 454 long reads outperforms those using Illumina short reads (e.g. average sensitivities close to 0.8 for the 454 under all tested nucleotide divergences in 3kb fragments). Noticeably, we found that DOMINO show high sensitivities even for relative high divergence levels (up to 0.3 substitutions per site between the two more diverged taxa); in this case, the program performs slightly better when using short reads as input. In the light of this high sensitivity, precision becomes a critical aspect to be considered for further successful marker discovery. We found that DOMINO also detects TNM regions with high precision (most values are close to 1 regardless of the condition), yielding very few number of false positives. The performance of DOMINO when using reads from larger library fragments (10 kb) is very similar to that of the observed for 3kb (Supplementary Fig. S2).

### 3.2 Application to empirical data

To illustrate the utility of DOMINO on real biological data, we performed a RRL sequencing experiment (using 454 reads; see Supplementary Information for details), which allow running all phases of the application and the DOMINO marker identification module, from raw reads to marker selection. We used four individuals (panel with four taxa) belonging to the spider family *Nemesiidae* (Araneae) for this analysis (Supplementary Information, Fig. S3). We identified many candidate regions that fulfill the requested marker characteristics (Supplementary Tables S3-S6), and tested the suitability of six of them by PCR amplification and Sanger sequencing in a larger panel that also included other 14 phylogenetically related species (focal species). The obtained phylogenetic tree not only recovered the expected relationships among the taxa from the panel but also demonstrates that the sequenced markers are useful to establish the phylogenetic relationships of the focal ones (Supplementary Fig. S4).

### 3.3 Conclusions

DOMINO will assist researches working with non-model organisms in the development of molecular markers for DNA variation studies. First, it allows obtaining a list of "personalized" markers that meet user specific criteria without the mandatory need of a reference genome, which will improve their application from highly specific taxonomic scopes to more wide phylogenetic ranges. Second, its output alignment files, jointly with the information about markers coordinates and features provided by the program, can be either directly utilized in variation studies, or used as a templates for further downstream PCR amplification or target DNA enrichment probe design. Third, the DOMINO GUI makes this application accessible and easy-to-use to non-experts in the bioinformatics of

NGS data handling and analysis. Finally, DOMINO is open cross-platform software that can be straightforwardly adapted to other pipelines or used in high performance computers. Although current version of the program works with raw reads of a limited number of reduction representation schemes (e.g. DOMINO cannot process raw reads from RAD- or RNA-Seq approaches) and sequencing platforms (Illumina short and 454 long reads), the modular structure of DOMINO will allow easily expanding the software to accept NGS data from other sources.

## Funding

Grants from the Ministerio de Educación y Ciencia of Spain (BFU2010-15484 and CGL2013-45211 to JR, and CGL2012-36863 to MAA).

## References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

Brito,P.H. and Edwards,S. V (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–55.

Catchen,J.M. *et al.* (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda).*, **1**, 171–82.

Chamala,S. *et al.* (2015) MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant Sci.*, **3**, 1400115.

Chevreux,B. *et al.* (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Comput. Sci. Biol. Proc. German Conf. Bioinform.* 99, 45-56.

Cock,P.J.A. *et al.* (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–71.

Cruaud,A. *et al.* RADIS: Analysis of RAD-seq data for InterSpecific phylogeny.

Eaton,D.A.R. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–9.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–4.

Huang,X. (1999) CAP3: A DNA Sequence Assembly Program. *Genome Res.*, **9**, 868–877.

Jukes T H & Cantor C R. (1969). Evolution of protein molecules, pp.21-132. In H. N. Munro (ed.), Mammalian Protein Metabolism. Academic Press, New York. *H. N. Munro (ed.), Mamm. Protein Metab. Acad. Press. New York*, pp.21–132.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.

Leese,F. *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS One*, **7**, e49202.

Lemmon,E.M. and Lemmon,A.R. (2013) High-Throughput Genomic Data in Systematics and Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, **44**, 99–121.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Maddison,W.P. *et al.* (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.

Mayer,C. *et al.* (2016) BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design. *Mol. Biol. Evol.*, **33**, 1875–1886.

Patel,R.K. and Jain,M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–41.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–4.

Suchan,T. *et al.* (2016) Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLoS One*, **11**, e0151651.

Yang,Z. (1997) PAML : a program package for phylogenetic analysis by maximum likelihood. **13**, 555–556.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–91.

**Figure 1.** Workflow showing the basic steps used to discover or select molecular markers with the DOMINO software

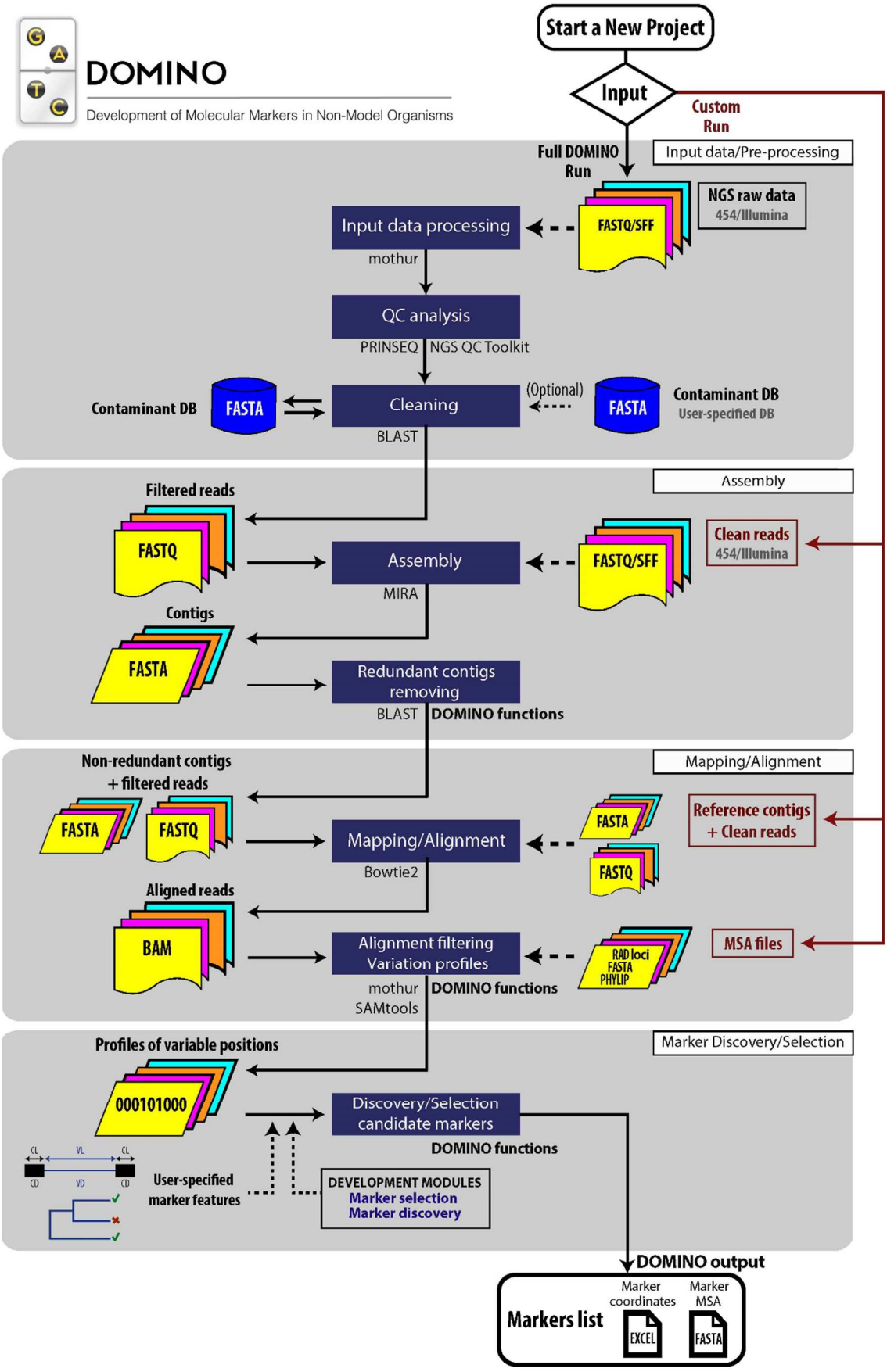

**Figure 2**. Screenshot of Marker Discovery/Selection TAB included in the DOMINO GUI.

**Figure 3.** Sensitivity and precision estimates for simulated data sets of 100 fragments of 3 kb after their *in silico* sequencing with Illumina and Roche-454 technologies.