Chapter 17
Bioinformatics-Based Assessment of the Relevance of Candidate Genes for Mutation Discovery

Michal Slota, Miroslaw Maluszynski, and Iwona Szarejko

Abstract The bioinformatics resources provide a wide range of tools that can be applied in different areas of mutation screening. The enormous and constantly increasing amount of genomic data obtained in plant-oriented molecular studies requires the development of efficient techniques for its processing. There is a wide range of bioinformatics tools which can aid in the course of mutation discovery. The following chapter focuses mainly on the application of different tools and resources to facilitate a Targeting-Induced Local Lesions in Genomes (TILLING) analysis. TILLING is a technique of reverse genetics that applies a traditional mutagenesis to create DNA libraries of mutagenised individuals that are then subjected to high-throughput screening for the identification of mutations. The bioinformatics tools have shown to be useful in supporting the process of candidate gene selection for mutation screening. The availability of bioinformatics software and experimental data repositories provides a powerful tool which enables a process of multi-database mining. The existing raw experimental data (genomics-related information, expression data, annotated ontologies) can be interpreted in terms of a new biological context. This may help in selecting the proper candidate gene for mutation discovery that is controlling the target phenotype. The mutation screening using a TILLING strategy requires a former knowledge of the full genomic sequence of the gene which is of interest. Depending on whether a fully sequenced genome of a particular species is available, different bioinformatics tools can facilitate this process. Specific tools can be also useful for the identification of possible gene paralogs which may mask the effect of mutated gene. Bioinformatics resources can also support the selection of gene fragments most prone to acquire a deleterious nucleotide change. Finally, there are available tools enabling a proper design of oligonucleotide primers for the amplification of a gene fragment for the purpose of mutation screening.

M. Slota • M. Maluszynski • I. Szarejko (✉)
Department of Genetics, Faculty of Biology and Environmental Protection, University of Silesia, Katowice, Poland
e-mail: iwona.szarejko@us.edu.pl

J. Jankowicz-Cieslak et al. (eds.), *Biotechnologies for Plant Mutation Breeding*,
DOI 10.1007/978-3-319-45021-6_17

Keywords Candidate genes • Identification of gene paralogs • Gene expression repositories • TILLING

17.1 Introduction

17.1.1 The Selection of Candidate Gene

The proper selection of a suitable candidate gene for mutational analysis is a fundamental step which determines the chances of success. The identification and prioritisation of candidate genes for TILLING analysis employs different types of accessible in silico resources:

1. *Web-based tools for searching literature-derived data* for the general annotation and characterisation of the biological and molecular function of preliminary candidates
2. *Relevant repositories of genomics-related information* which can be used for retrieving of specific homologous genes from many species
3. *Gene expression repositories* which gather gene expression data: hybridisation arrays, chips, microarrays and RNA-seq data
4. *Gene ontology (GO) databases* which allow searching for genes according to their functional annotation, molecular characteristics or protein localisation

The selection strategy of suitable candidate genes involved in a course of specific developmental process or environmental stress reaction can be carried in different manners. The conventional approach is based on a presumptive knowledge or assumption about associated biological processes. This consists of a literature mining process that employs biological-oriented databases such as PubMed [www.ncbi.nlm.nih.gov/pubmed/], BioText [http://biosearch.berkeley.edu/] or other available services for general text mining as Google Scholar [http://scholar.google.pl/]. This attempt may be effective for such purpose if there is a known biological context associated with well-studied biological processes that were carried out on one of the model species. In the studies of more complex or under-examined traits for non-model species, there may be a lack of such presumptive knowledge. Limited knowledge about the genetic background of a complex trait can be overcome by collecting additional experimental data, as well as performing a wide range of in silico analyses. Gene expression databases such as Gene Expression Omnibus (GEO) [http://ncbi.nlm.nih.gov/geo/], ArrayExpress [http://www.ebi.ac.uk/arrayexpress/] or Genevestigator [https://www.genevestigator.com/] offer a publically available repository of expression data for multiple gene sets. Browsing of gene expression database can lead to the identification of gene subsets that are differentially expressed across different experimental conditions. The identification of gene expression profiles can significantly reinforce the process of the selection of candidate genes associated with a distinct biological process or related to the reaction to environmental stimuli. Meta-analysis carried on database-derived data

can lead to the assortment of gene accessions displaying elevated or lowered expression levels in defined experimental conditions. Such genes may appear to be suitable candidates for the functional analysis using TILLING strategy. The proper description and classification of the gene expression and phenotype data is also crucial to enable meaningful comparisons. Gene ontology (GO) is a bioinformatics initiative aiming to introduce a unified vocabulary system for the biological terms. These terms are applied for the characterisation of the biological properties. GO can be applied for the hierarchical classification of gene products that can be divided into three main categories: cellular component, molecular function and biological process involvement (Carbon et al. 2009). Each annotation of GO term in database has a reference to associated information on related gene involvement. There are several different types of accessible web-based repositories that can be used for plant GO browsing, e.g. agriGO [http://bioinfo.cau.edu.cn/agriGO/], AmiGO Gene Ontology [http://amigo.geneontology.org/], Plant Ontology browser [http://www.plantontology.org/amigo/go.cgi/] and QuickGO browser [http://www.ebi.ac.uk/QuickGO/].

17.1.2 The Identification of the Genomic Sequences for Mutation Screening

Depending on the extent of available genomic sequence resources for a species, different approaches can be applied for sequence recovery. If the genome has already been sequenced, the genomic sequence of the gene of interest can be retrieved directly from a database. On the contrary, for a species for which the genome sequence is unknown or the genomic sequence is not available in accessible databases, the orthologous sequence can be identified using a closest relative as a query (e.g. *Arabidopsis* or rice). Basic Local Alignment Search Tool (BLAST) search may be carried, for example, against the rice genome repository to identify monocot orthologs of Arabidopsis genes [http://blast.ncbi.nlm.nih.gov/]. The identification of a complete coding sequence of a selected homologue requires the application of mRNA and/or amino acid sequences as a query in the search of expressed sequence tag (EST) databases. The assembling of identified and amplified sequences allows obtaining a coding sequence of the particular species. In the following step, the genomic sequence of an analysed homologue can be amplified with a use of designed exonic primers. The occurrence of paralogs (homologous genes created through a duplication event) in genomes should be taken into consideration during the selection of genes for TILLING analysis. Paralogous copies of a specific gene with redundant function can mask the effect of the mutation, thus complicating any gene knock-down strategies (Kurowska et al. 2011). In plants, paralogs are widespread among almost all plant lineages (Van Bel et al. 2011). There are accessible plant comparative genomics databases, such as GreenPhylDB [http://www.greenphyl.org/], Ensembl Plants [http://plants.

ensembl.org/] and Phytozome [https://phytozome.jgi.doe.gov/]. These resources group homologous genes into families using clustering useful for identifying possible orthologs. Gene Family Finder tool implemented within PLAZA 2.5 database [http://bioinformatics.psb.ugent.be/plaza/] enables the identification of gene families specific to one or more species and explores genomic information from search of different plants for structural and functional gene annotations (Van Bel et al. 2011). PLAZA package is an effective tool for the verification of the occurrence of possible gene paralogs within genome.

17.1.3 Selection of Gene Fragments for a Mutation Screening

The selection of a suitable gene region provides a higher probability to identify nonsense or useful missense mutations by TILLING screening (Chen et al. 2014). Depending on the applied detection method for mutation screening, a gene fragment length for TILLING should be limited to 1500 bp. The free web-based software Codons Optimised to Detect Deleterious Lesion (CODDLE) [http://blocks.fhcrc.org/proweb/coddle/] is a tool facilitating the choice of gene regions most suitable for TILLING analysis. CODDLE identifies the protein regions that are most likely to contain the largest percentage of deleterious lesions generated by G/C to A/T transitions. Also, the conservation-based sorting intolerant from tolerant (SIFT) software predicts whether an amino acid change is damaging for a protein [http://sift.jcvi.org/]. SIFT performs the multiple alignment for homologous sequences to predict whether the amino acid change is expected to have deleterious effects on the protein stability (Ng and Henikoff 2003). The determination of the region of the gene where the occurrence of point mutation has the highest probability to affect the gene function can be also carried out using a conserved domain prediction method (Till et al. 2007). On the basis of protein sequence obtained by in silico translation of a gene query, the highly conserved domains responsible for the reported catalytic properties can be identified using accessible repositories, such as conserved domain database (CDD) [http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi/].

17.1.4 Designing of Primers for the Amplification of a Gene Fragment

The classical TILLING protocol employs a mismatch-specific and sensitive endonuclease treatment followed by polyacrylamide electrophoresis and visualisation with a use of highly sensitive LI-COR gel analyser system (LI-COR Biosciences). This attempt requires an efficient amplification of a gene fragment selected for

mutational analysis (Till et al. 2006). The oligonucleotide primers flank a fragment of a length of optimally 1–2 kb or less. IR-labeled (labeled with infrared dye) primers which should have high (up to 70 °C) thermal stability can be designed using web-based primer design tools such as Primer3 [http://bioinfo.ut.ee/primer3-0.4.0/primer3/] or Primer-BLAST [www.ncbi.nlm.nih.gov/tools/primer-blast/]. Designed primers should be analysed using BLAST tool against the full genomic DNA sequence to ensure that they target a single sequence. The same primer design and PCR parameters have been used for TILLING by sequencing (*see* Chap. 20).

17.2 Materials

17.2.1 Literature Mining

There are a variety of accessible web-based resources that can be applied for a scientific literature search. Depending on their repository content and applied indexing methods, these resources can be divided into specific categories (Table 17.1):

(a) Search tools for scientific abstracts
(b) Search tools designed to search beyond abstract, allowing access to the full-text articles
(c) Image-oriented browsers which allow searching for figure/table content of specific articles
(d) Databases which aim to enrich a literature search in terms of prioritising the relevance of records, gene ontology annotations and assignment of biological context

Table 17.1 Representative databases which can be applied for scientific literature mining

Category	Database	Website
Literature mining	PubMed	[http://www.ncbi.nlm.nih.gov/pubmed/]
	PubMed Central	[http://www.ncbi.nlm.nih.gov/pmc/]
	Google Scholar	[http://scholar.google.com/]
Figure/table browsers	BioText	[http://biosearch.berkeley.edu/]
	Yale Image Finder	[http://krauthammerlab.med.yale.edu/imagefinder/]
Text records prioritising	RefMed	[http://dm.postech.ac.kr/refmed/]
Ontology annotation	GOPubMed	[http://www.gopubmed.com/]
Context clusterisation	CiteXplore	[http://www.ebi.ac.uk/citexplore/]

17.2.2 Databases of DNA and Protein Sequences

The sequence of a specific gene can be retrieved for many species from publically accessible databases which gather information of nucleotide and protein sequences. As the species with the sequenced genomes are taken into consideration, there is a considerable support resulting from accessible genetic data browsers, such as NCBI GenBank. A repository search can be performed based on sequence query using a Basic Local Alignment Search Tool (BLAST) tool. BLAST searches for CoreNucleotide (main sequence collection), dbEST (expressed sequence tag collection) and dbGSS (genome survey sequences collection of unannotated short single read) can be carried independently. In the case of species for which the amount of genomic data is more limited, there are numerous locally developed search tools available which can be applied. ViroBLAST [http://indra.mullins.microbiol.washington.edu/viroblast/viroblast.php/] is a stand-alone BLAST web server for nucleotide and protein queries of multiple databases and user's defined datasets. ViroBLAST was implemented in various areas of plant research for browsing a raw data of sequenced plant genomes. If the genome of a particular species has not already been sequenced, there are alternative strategies for the identification of the sequence of a gene that is of interest, employing expressed sequence tag (EST) repositories. EST databases are widely used for the identification of a complete coding sequence of a new gene homologue using a related species query. Among different databases of DNA and protein sequences, the ones which are most commonly used for a gene sequence recovery are presented (Table 17.2).

17.2.3 Meta-analysis of Gene Expression Data

The use of gene expression databases can facilitate the process of searching for genes that are differentially expressed in different experimental conditions or tissues. This approach can contribute to the identification of up- or downregulated genes which can be suitable candidates for the functional analysis of their involvement in a particular process. A meta-analysis of gene expression profiles can be conducted with the application of gene expression repositories. Microarray databases aim to store the measurement of expression raw data, to process the content into a searchable index and to provide access to data in other applications for further analyses and interpretation. There are accessible databases which allow analysis of the primary data in order to develop a new data resource (Table 17.3).

Table 17.2 Representative databases which can be applied for a gene sequence mining

Category	Database	Website
Nucleotide databases	NCBI GenBank	[www.ncbi.nlm.nih.gov/genbank/]
	TAIR	[www.arabidopsis.org/]
Search tool	BLAST	[http://blast.ncbi.nlm.nih.gov/blast/]
ViroBLAST search tools	*Hordeum vulgare* ViroBLAST	[http://webblast.ipk-gatersleben.de/barley/viroblast.php/]
	Populus trichocarpa ViroBLAST	[http://popgenie.org/blast/]
	Triticum turgidum ViroBLAST	[http://wheat.pw.usda.gov/GG2/WheatTranscriptome/*viroblast*/]
EST databases	NCBI dbEST	[http://www.ncbi.nlm.nih.gov/dbEST/]
	PlantGDB	[http://www.plantgdb.org/cgi-bin/blast/PlantGDBblast/]
	CR-EST	[http://pgrc.ipk-gatersleben.de/cr-est/]

Table 17.3 Representative databases which can be applied for the meta-analysis of gene expression data

Database	Website	No. of experiments
ArrayExpress	[http://www.ebi.ac.uk/arrayexpress/]	63,450
ArrayTrack	[http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/]	1622
Gene Expression Omnibus (GEO)	[http://www.ncbi.nlm.nih.gov/geo/]	65,361
Genevestigator	[https://www.genevestigator.com/]	2621

17.2.4 Gene Ontology (GO) Analysis

The identification of candidate genes can be aided by using gene ontology-based criteria of semantic similarity. The functional interpretation of a raw experimental data can be accomplished by using the GO, for example, via an enrichment analysis. The annotation of GO terms for target genes can also contribute to achieve a better understanding of the biological context among three domains which are attributed: cellular component, molecular function and biological process. Gene ontology (GO) and plant ontology (PO) data can be browsed using various textual and graphical views and filtered for species-specific repositories. There are several different types of accessible web-based repositories that can be used for plant GO and PO terms browsing, e.g. AmiGO Gene Ontology [http://amigo.geneontology.org/] and Plant Ontology Browser [http://www.plantontology.org/amigo/go.cgi/]. There are numerous publically accessible databases which have implemented GO or PO annotations in order to improve the classification of their inner repositories, such as microarray data, the QTL directories and mutant germplasm (Table 17.4).

Table 17.4 Representative databases that have implemented the gene ontology (GO) or plant ontology (PO) terms for describing and cataloging their internal resources

Database	Website	Annotated data
AgriGO	[http://bioinfo.cau.edu.cn/agriGO/]	Genes
BarleyBase	[http://www.barleybase.org/]	Microarray
GrainGenes	[http://wheat.pw.usda.gov/GG3/]	Genes, QTLs
Gramene (a resource for comparative grass genomics)	[http://www.gramene.org/]	Genes, QTLs, proteins
Genevestigator (microarray database and analysis toolbox)	[http://www.genevestigator.ethz.ch/]	Microarray
International Rice Information System (IRIS)	[http://irri.org/tools-and-databases/international-rice-information-system/]	Mutants
Maize Genetics and Genomics Database (MaizeGDB)	[http://www.maizegdb.org/]	Genes, mutants
Medicago truncatula Gene Expression Atlas (MtGEA)	[http://mtgea.noble.org/v3/]	Microarray
Nottingham Arabidopsis Stock Centre (NASC)	[http://arabidopsis.info/]	Germplasms, genes
Oryzabase	[http://www.shigen.nig.ac.jp/rice/oryzabase/]	Genes, germplasms, microarray, mutants
Oryza Tag Line	[http://oryzatagline.cirad.fr/]	Mutants
Plant Expression Database (PLEXdb)	[http://www.plexdb.org/]	Microarray
Rice Oligonucleotide Array Project	[http://www.ricearray.org/]	Microarray
Solanaceae Genomics Network (SGN)	[https://solgenomics.net/solanaceae-project/]	Genes
SoyBase	[http://soybase.org/]	Genes, microarray, mutants
The Arabidopsis Information Resource (TAIR)	[http://www.arabidopsis.org/]	Genes, germplasms, microarray

17.2.5 Gene Homology Search

A wide variety of bioinformatics tools enables the analyses of sequence homologies between and within plant genomes. Specific databases dedicated for plant comparative genomics studies group homologous genes into families using clustering algorithms. Different packages are available for an effective verification of the occurrence of possible gene paralogs within genome (Table 17.5).

Table 17.5 Representative databases which can be applied for gene homology search

Database	Website	Function
Ensembl Plants	[http://plants.ensembl.org/]	Gene/protein homology search
GreenPhylDB	[http://www.greenphyl.org/]	
Phytozome	[https://phytozome.jgi.doe.gov/]	Gene/protein phylogenetic data
PLAZA 3.0 database	[http://bioinformatics.psb.ugent.be/plaza/]	Identification of gene families

17.2.6 Selection of Gene Fragments for the Mutational Analysis

The design of an amplicon as a crucial step in TILLING analysis can be facilitated using accessible bioinformatics software packages. The selection of a suitable gene region should be performed to provide a higher probability of the identification of nonsense or beneficial missense mutations within the selected gene region. Bioinformatics tools can be applied to facilitate the choice of gene region most suitable for mutation screening and the prediction of mutation effect on protein stability and function (Table 17.6).

17.2.7 PCR Primer Design for Mutation Screening

There is a wide selection of web-based primer design tools which can be applied for the analysis of predesigned primers or entering nucleotide sequence to design primers which match defined properties (Table 17.7). Designed oligonucleotide primers can be additionally verified in terms of their detailed molecular characteristics and possible secondary structure using appropriate software.

17.3 Methods

17.3.1 Biomedical Text Mining

1. Browse a PubMed database [http://www.ncbi.nlm.nih.gov/pubmed/] (*see* **Note 1**) using either a user-provided query associated with a specific topic or a predefined set of publications for text categorisation.
2. Filter obtained results in terms of article type, text availability, publication date or investigated species.
3. Use a 'Related information' menu within abstract content to link to other related NCBI databases for the selected record (e.g. gene and protein sequence).

Table 17.6 Representative databases and bioinformatics tools which can be applied for the selection of gene fragments for the mutational analysis

Database	Website	Function
CODDLE	[http://blocks.fhcrc.org/proweb/coddle/]	Prediction of most suitable gene regions for TILLING analysis
SIFT (Sorting Intolerant From Tolerant)	[http://sift.jcvi.org/]	Prediction of mutation effect on protein stability
PROVEAN	[http://provean.jcvi.org/seq_submit.php/]	
I-Mutant3.0	[http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi/]	
ProtParam	[http://web.expasy.org/protparam/]	Assessment of protein properties
PDBeFold	[http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver]	Modelling of protein secondary structure
SAS	[http://www.ebi.ac.uk/thornton-srv/databases/sas/]	
SOPMA	[http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html/]	
Phyre2	[http://www.sbg.bio.ic.ac.uk/phyre2/]	Modelling of protein tertiary structure
SWISS-MODEL	[http://swissmodel.expasy.org/interactive/]	
DUET	[http://bleoberis.bioc.cam.ac.uk/duet/stability/]	
Conserved Domain Database	[http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi/]	Identification of highly conserved protein domains

Table 17.7 Representative tools which can be applied for oligonucleotide designing

Database	Website	Function
Primer3	[http://primer3.ut.ee/]	Primer design
PrimerQuest	[http://eu.idtdna.com/PrimerQuest/]	
Primer-BLAST	[www.ncbi.nlm.nih.gov/tools/primer-blast/]	
OligoAnalyzer	[http://eu.idtdna.com/calc/analyzer/]	Oligonucleotide analysis

PubMed as a part of NCBI Entrez retrieval system provides a linked access to a set of 39 databases.

4. Use a 'Related information' menu for an article record to follow a link to PubMed Central database [http://www.ncbi.nlm.nih.gov/pmc/] for retrieving a full-text paper content.
5. The further extraction of biological context of identified text records can be performed using additional resources. The predefined types of information can be automatically extracted in terms of gene ontology annotation using GOPubMed [http://www.gopubmed.com/] tool. The entire biological context

for a specific list of records can be clustered using CiteXplore [http://www.ebi.ac.uk/citexplore/] tool.

17.3.2 Retrieving the DNA and Protein Sequences

1. Browse an Ensembl Plant database (*see* **Note 2**) by using gene name, symbol or description for all species repositories.
2. Select a most relevant record and explore a 'Pan-taxonomic Compara' menu within an Orthologues repository. The information about the number of identified splice variants, orthologs and paralogs is provided in a 'Gene table' section. A blue square on a generated graph depicts a speciation event (separating orthologs), whereas a red square stands for a duplication event (separating paralogs).
3. Analyse identified gene orthologs which are presented in results table with specified type of homology, Ensembl identifiers, location and query coverage. The number of species for each type of ortholog class is shown in the Type column in the table. The types of orthologs are assigned for each hit and are as follows: 1-to-1 orthologs indicates that only one copy is found in each species, 1-to-many orthologues indicate that one gene in one species is orthologous to multiple genes in the other species and many-to-many orthologs—multiple orthologs can be found in both species. The identified single-copy (1-to-1) orthologs are the most desirable targets.
4. Check the dN/dS ratio—a ratio of dN (non-synonymous substitutions) to dS (synonymous substitutions) of lower than 1 indicates negative (or purifying) selection, while a ratio of higher than 1 indicates positive selection. The information on the location of the ortholog (either on a chromosome or on a scaffold sequence), and the percentage of identical amino acids in the ortholog compared with the sequence of the target gene (target %ID), is also provided. A low identity for amino acid sequences is below 60 %, whereas high identity is up to 98 % which would be a good characteristic of a true ortholog.
5. Perform a DNA or protein alignment of identified orthologs using a built-in Ensembl Plants tool or directly using a BLAST [http://blast.ncbi.nlm.nih.gov/blast/] search.

17.3.3 Meta-analysis of Gene Expression Data Using Gene Expression Repositories

1. Browse a gene expression datasets at GEO DataSets toolbox [http://www.ncbi.nlm.nih.gov/gds/] (*see* **Note 3**) using specific search terms of experiment description to locate desired experiments.

2. Select a GEO DataSet results page which contains the information concerning the experiment summary, type of experimental data, experiment variables and the number and characters of samples. A single result description page contains a comprehensive description of experiment, tested variables and all samples. Pay special attention to the experimental conditions and the number of replicates which were tested.
3. Use a reference in GEO Accession Display menu to external data analysis tools. GEO2R [http://www.ncbi.nlm.nih.gov/geo/geo2r/] is an implemented interactive web tool that allows to compare two or more groups of samples deposited in a GEO database in order to identify genes that are differentially expressed across different experimental conditions.
4. Perform an analysis of selected data set (characterised by specific accession and platform number) by defining the corresponding groups (Fig. 17.1). Up to ten groups can be defined at once. Define at least two opposite groups in the 'samples' panel and select independent analyses within them.
5. Analyse the obtained results which are presented in the browser in a table containing the top 250 genes ranked by *P*-value and characterised by different parameters. Use the select columns option to modify which data and annotation columns are shown in the result table. The data column contains the information on the following parameters: *P*.Val—*P*-value after adjustment for multiple

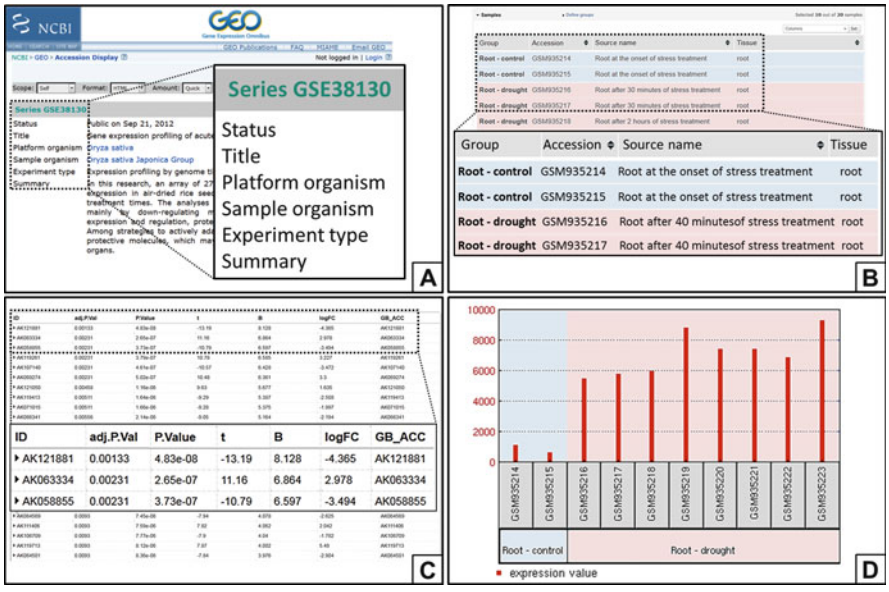

Fig. 17.1 Steps of gene expression data browsing using a Gene Expression Omnibus (GEO) database. The experiment selected for the analysis was 'Gene expression profiling of acute drought response in leaf and root of rice', accession: GSE38130 and platform: GPL15594 (**a**). In the samples panel, two groups were defined consisting of root and shoot tissues at the onset of stress treatment compared with four drought treatments in two replicates (**b**). Results are presented in the browser as a table of the top 250 genes ranked by *P*-value (**c**). Gene expression profile graph can be displayed by clicking on a specific row of the gene list (**d**)

testing, *P*.Value—raw *P*-value, *t*—moderated t-statistic, B—B-statistic or log odds that the gene is differentially expressed, logFC—log2 fold change between two experimental conditions and F—moderated F-statistic.

6. The gene expression profile graphs can be displayed for each row of the gene list.
7. To identify and catalogue the differently expressed genes, the dataset can be exported as a .txt file as a sample-probe matrix with label information.
8. The sorting of differently expressed genes can be accomplished using spreadsheet software (e.g. Microsoft Excel). Genes with the smallest *P*-values will be the most reliable targets. Presort the genes with $p < 0.05$ and fold change of >1.5 and above.
9. After the identification of specific gene expression profiles, there are several links on the profile records that help to identify additional genes that is of interest, including similarly expressed genes or genes within close proximity on the chromosome. Full information about these links is provided on the About GEO Profiles page.

17.3.4 Gene Ontology (GO) Annotation

1. Search for a specific GO annotation of a gene that is of interest using QuickGO browser for gene ontology [http://www.ebi.ac.uk/QuickGO/]. GO search can be conducted using a gene/protein name, symbol, accession number or description word as a query.
2. Obtained GO annotation data consists of three main aspects: molecular process involvement (P), molecular function (F) and localisation in specific component (C).
3. Analyse the QuickGO result table which contains a specific GO terms and identifiers, type of evidence and the affiliation to a database which created the annotation.
4. You can visualise and/or download annotation statistics in .xls format. Statistics reports contain the information on the frequency of covered ontology aspect: molecular process involvement/molecular function/localisation in specific component, the percentage of different types of evidence—experimental (EXP), direct Assay (IDA), physical interaction (IPI), mutant phenotype (IMP), genetic interaction (IGI), expression pattern (IEP) and the frequency of annotation to distinct taxon.
5. Additionally, a created gene list can be classified in terms of specific GO annotation for different ontology categories.
6. Perform a singular enrichment analysis (SEA) using an agriGO [http://bioinfo.cau.edu.cn/agriGO/] database.
7. Paste a gene symbols list compatible to an input format as well as the reference (user may choose between the genomic background and customised annotated reference which consist of a second group of GO annotation data).

8. Browse a Detail information table to check the enrichment of input gene targets for different GO terms. GO term can be considered significantly enriched, if the parameter of false discovery rate (FDR) is <0.05 and p-value is <0.01 when compared to all the gene transcripts annotated in the selected genomic background or the set of genes applied as the background.
9. AgriGO tool uses a gene list input to create GO terms abundance chart for the provided accessions (Fig. 17.2).
10. Analyse a provided chart which summarises the gene's affinity to different biological processes in comparison to common genomic reference.
11. If required you may complement the obtained results by the statistical analysis. To test the statistical differences between the enrichments of the input list to the previously computed background or a subset of reference list, select a hypergeometric or a fisher method. If the number of queries is larger than a few, chi-square test is more suitable.

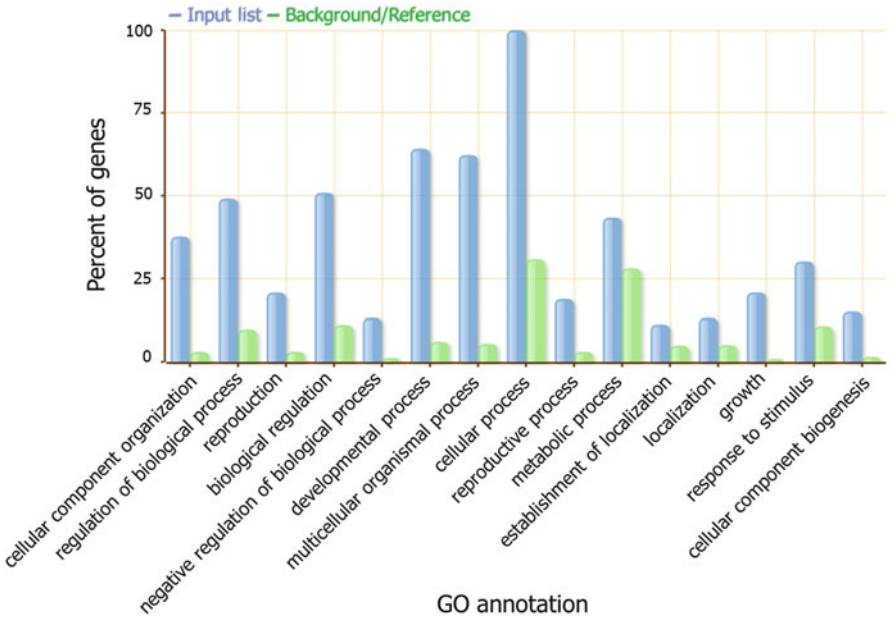

Fig. 17.2 Singular enrichment analysis (SEA) results obtained using agriGO database. Complete preliminary candidate genes list served as a query. *Blue bars* represent the GO term enrichment for the input gene list, whereas *green bars* represent the enrichment for a background genome (*Arabidopsis*)

17.3.5 Identification of Possible Gene Paralogs

1. The identification of the possible paralogs (genes duplicated within a species) of the gene which is of interest can be performed using Ensembl Plants [http://plants.ensembl.org/] database.
2. Browse an Ensembl Plants database by using gene name, symbol or description for all species repositories.
3. Select a most relevant record and explore a 'Plant Compara' menu within a paralog repository. The information about the number, ancestral taxonomy, chromosomal localisation and the percentage of identity to the target gene for identified paralogs is provided in a table.
4. Use the 'Region Comparison' function available in Compare tab. The top panel is similar to the chromosome diagram and gene map from the Location tab. The localisation of gene paralogs in the genome of interest can be displayed graphically in the lower panel. This page displays chromosomes, scaffolds and sequence contigs.
5. Use the 'Alignment (protein)' or Alignment (cDNA) function available in Compare tab. Result page contains the information on the gene-paralog alignment details. A multiple sequence alignment (MSA) for the pair of sequences is also provided in CLUSTAL W format.

17.3.6 Selection of a Gene Fragment for Mutation Screening

1. Submit a complete genomic sequence (FASTA format) and an exon/intron position statement (the positions of the start and end of each protein-coding exon in the submitted sequence) compatible with input GenBank format in a CODDLE [http://blocks.fhcrc.org/proweb/coddle/] (*see* **Note 4**) tool menu.
2. Specify an optional identifier which will be included in the output.
3. Select a mutagenesis method and scoring system.
4. Perform a computation to a search for a target window that maximises the probability of recovering missense mutations and truncations based on the characteristics of the mutagen and the specified organism.
5. Analyse a graphical display of the search results, indicating the region within the amplicon where mismatch detection is prone to be sensitive for the induction of the mutation (Fig. 17.3).

17.3.7 Designing of PCR Primers for Mutation Screening

1. Design oligonucleotide primers for the amplification of selected gene fragments using accessible web-based primer design tools. Primer3 [http://bioinfo.ut.ee/

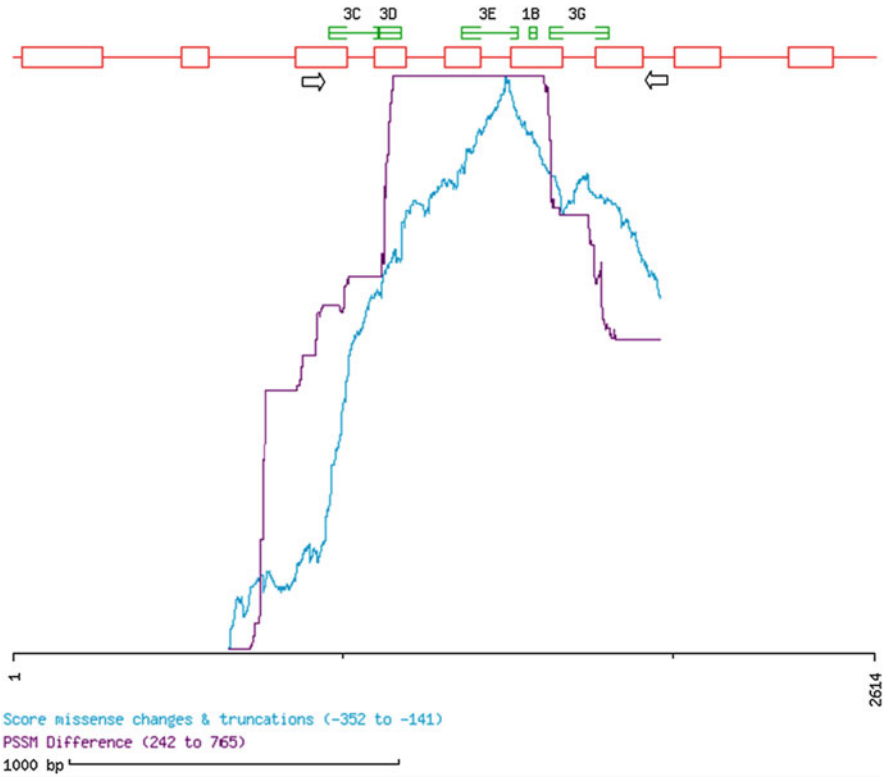

Fig. 17.3 Graphical representation of the results obtained using a CODDLE tool. Exons are indicated as *open boxes* and introns as a *single line*. Nucleotide position numbers are depicted on the *x*-axis at the *bottom* of the plot. The probability to discover a nucleotide change that is disruptive for a gene is indicated by the score on the *y*-axis

primer3-0.4.0/primer3/] tool can be applied due to a user-friendly interface and numerous primer design properties.

2. Paste a raw nucleotide sequence of a gene that is of interest (5′→3′) onto a sequence menu.
3. Predesigned primer sequences can be additionally tested by entering their sequences on appropriate menus.
4. Select a 'General Primer Picking Conditions'. Designed primers for TILLING analysis should match the specified criteria:

   - A length of 18–30 nucleotides.
   - The melting temperature (T_m) of the primers between 65 and 75 °C, and not more than 5 °C difference of each other.
   - The GC content between 40 and 60 %, with the 3′ of a primer ending in C or G to promote binding.
   - Lack of regions of secondary structure.

- A balanced distribution of GC-rich and AT-rich domains.
- Lack of runs of four or more of one base or dinucleotide repeats.
- Lack of intra-primer homology (more than three bases that complement within the primer) or inter-primer homology (forward and reverse primers having complementary sequences).

5. Choose a 'Pick Primers' command to obtain five pairs of designed primers which best match the selected criteria.
6. Verify the properties of designed oligonucleotide primers using appropriate software, e.g. OligoAnalyzer [http://eu.idtdna.com/calc/analyzer/].
7. Check the sequences of the selected pair of primers using a nucleotide BLAST tool for being unique in the genome of the corresponding species (if possible).

17.4 Notes

1. PubMed database was selected to demonstrate the biomedical text mining process as it contains a most extensive and frequently updated repository of scientific papers on the plant-related studies.
2. The sequence of the gene that is of interest can be easily retrieved from NCBI GenBank repository in case of species of which a full genomic sequence is already known. If the genome of a particular species has not been already sequenced or its genome sequence is not deposited in GenBank or other databases (e.g. species-specific ViroBLAST repository), the homologous sequence of a gene that is of interest can be retrieved by homology search. The identification of homologous gene sequences of particular species can be performed using databases such as Ensembl Plants [http://plants.ensembl.org/] for the othrologs search.
3. GEO (Gene Expression Omnibus database) was selected for the meta-analysis of gene expression profiles as it contains a large repository of plant microarray data, which is stored and arranged into a searchable index providing an easy access for further analyses and interpretation.
4. Codons Optimised to Detect Deleterious Lesion (CODDLE) tool is a software which identifies the protein regions that are most likely to contain the largest percentage of deleterious lesions can be generated by G/C to A/T transitions. The software can be downloaded and run locally (http://blocks.fhcrc.org/blocks/uploads/proweb/).

Acknowledgments Funding for this work was provided by the Food and Agriculture Organization of the United Nations and the International Atomic Energy Agency through their Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture through Research Contract No. 15419 of IAEA Coordinated Research Project D24012 and the Polish Ministry of Science and Higher Education (Grant No. 2080/IAEA/2011/0, 2557/FAO/IAEA/2012/0, 2904/FAO/IAEA/2013/0).

References

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288–289

Chen L, Hao L, Parry MA, Phillips AL, Hu YG (2014) Progress in TILLING as a tool for functional genomics and improvement of crops. J Integr Plant Biol 56(5):425–443

Kurowska M, Daszkowska-Golec A, Gruszka D, Marzec M, Szurman M, Szarejko I, Maluszynski M (2011) TILLING—a shortcut in functional genomics. J Appl Genetics 52:371–390

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814

Till BJ, Zerr T, Comai L, Henikoff S (2006) A protocol for TILLING and Ecotilling in plants and animals. Nat Protoc 1(5):2465–2477

Till BJ, Comai L, Henikoff S (2007) TILLING and EcoTILLING for crop improvement. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement. Springer, Dordrecht, pp. 333–349

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2011) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol 158:590–600