SOL Project
Sequencing and Bioinformatics
Standards and Guidelines

Version of 13 December 2005

SOL Project

Sequencing and Bioinformatics Standards and Guidelines

Contents

Note: Feedback is welcome. Please email your suggestions and and comments to Lukas Mueller (LAM87@cornell.edu).

Acknowledgments

Based on documents from the Medicao Sequencing Project kindly provided by Nevin Young, and a document obtained from Gramene, kindly provided by Pankaj Jaiswal, and the GMOD project (http://www.gmod.org).

Introduction

This document is an attempt to develop a series of standards and guidelines for the SOL tomato sequencing project to ensure maximum consistency of tomato genome assembly and annotation across different centers. It addresses sequencing quality standards, BAC assembly, SOL data release policy, gene naming (traditional gene names, BAC based names, chromosome based names and gene classification), gene structural and functional annotation, data exchange formats and guidelines for software systems and platforms, with some of the sections still a bit patchy. It should represent a realistic and reasonable compromise that all project members can agree on, so the contributions of all involved with SOL are highly welcome. The standards and guidelines are based mostly on those developed for other genomes, where applicable, as noted in the text. This should make the results of this project more readily comparable to other projects using similar standards.

I. Data Archiving

Data will be archived on SGN and made available through the SGN FTP server (ftp://ftp.sgn.cornell.edu/).

All primary data must to be archived for future reference. For BAC sequences, all the data should be collected in a directory named after the BAC using the Arizona style BAC name (example: C01Hba0001A01). The data to be archived includes the following:

- Sequence reads as chromatogram files (.scf or .ab1) in a subdirectory named chromat_dir
- The original sequence reads and phred quality values in fasta format in a subdirectory names seq_dir
- The assembly files for all BAC assemblies (in .ace format) in a subdirectory called edit_dir
- The BAC sequence in FASTA format, in a file with the BAC name and the suffix .seq. The sequence id in the FASTA file should be the BAC name.
- The parameters used in gene calling software, such as for which species the parameters were calibrated, in a file called parameters.txt
- The original output of the gene calling software in a file called original_output.txt.
- The final annotations in GAME XML in a file named after the BAC with the suffix .xml

I. 1. File naming conventions

BAC standard naming:
C+the chromosome number (always two digits) + the library name (3 chars) + the bac plate number (always 4 digits) + the bac plate well coordinates (always two digits) + "." + shotgun library plate number (two digits)

Example: C01HBa0001A01

The name of the top level folder corresponds to the standard name of the BAC. This folder contains a FASTA file named after the BAC with the extension .seq, like so:

C01Hba0001A01.seq

Containing the assembled fasta sequence. Multiple sequence entries are possible for unfinished BACs. The identifiers for the sequences should correspond to the BAC name. If multiple sequences are present, they should be numbered with dashes.

The GAME xml file containing the annotation is named as follows:

C01HBa0001A01.xml

Recommendation for the trace file naming:
The trace file names are generated as follows: C+the chromosome number (always two digits) + the library name (3 chars) + the bac plate number (always 4 digits) + the bac plate well coordinates (always two digits) + "." + shotgun library plate number (two digits) + "." + read direction (R or F) + "." + the file type (ab1 or scf).

Note: the bac plate numbers and well coordinates refer to the original bac location, as in the Cornell BAC name, and NOT to the re-arrayed plate that is sent out to sequencing centers.

Example: C01HBa0001A01.01A01.R.ab1

The ace file is named as follows:

C01Hba0001A01.ace

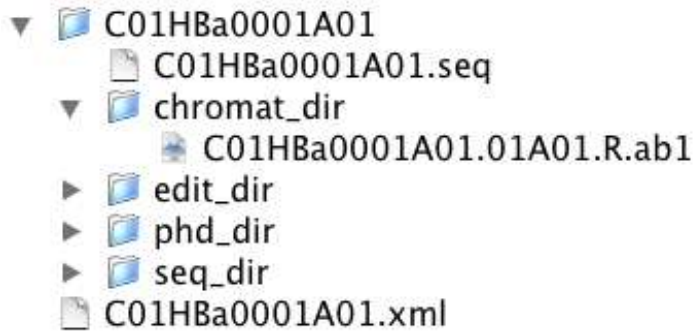

Figure I.1: Directory structure of the submission data before tar and gz.

The data should be uploaded to SGN for archival using special SGN uploading accounts. Please contact Lukas (LAM87@cornell.edu) for more information on how to use the upload accounts.

I.2. BAC statuses

To update BAC statuses follow these directions:

(1) Log on to SGN

Point your browser to http://sgn.cornell.edu and click on Login on the top right corner of the screen. Enter your username and password. A page is displayed that allows you to change your user information and other things.

(2) Go to the BAC search page under the search menu (http://sgn.cornell.edu/search/direct_search.pl?search=bacs)

Search for the BAC whose attribution or status needs to be changed. On the result page, click on the appropriate BAC to get to the BAC detail page. The attribution can be adjusted using the attribution link provided. If no attribution link appears, it means that you are not logged in or that you are not a user of type “sequencer”. Please contact SGN to correct the user type of your account.

(3) Modify the BAC status

On the BAC detail page, click on the appropriate link, such as "in progress", "completed", etc, to set

the BAC status to the new status.

IMPORTANT: The BAC status information is used to calculate overall statistics for the sequencing project. The statistics summary on the tomato sequencing overview page (http://sgn.cornell.edu/help/about/tomato_sequencing.html) is generated from the database and as such immediately updated when BAC statuses are changed. All BACs that are in the status "complete" need to be uploaded into the upload accounts (see below).

I.3. SGN upload accounts

The BAC data should be uploaded via the SGN upload accounts. These are unix style accounts that allow data upload via "scp". scp can be used as a command line tool in common unix variants such as Linux, Solaris, and MacOS. There are also graphical interfaces available, such as WinSCP (http://winscp.sourceforge.net/eng/index.php) for Windows. The following instructions are assuming that you use a computer running a variant of Unix. Note: The files have to be uploaded in the standard file format defined in section I.1.

(1) Package and compress the data

Before the files can be sent, they need to be packaged and compressed, using the tar command. In a terminal, assuming a BAC name of C01HBa0001A01, type

tar zcvf C01HBa0001P01.tar.gz C01HBa0001A01*

The tar file should contain the original input/output files from phred/phrap/consed (directories such as chromat_dir, seq_dir, edit_dir, phd_dir and so forth) plus a fasta file with the BAC sequence, with the BAC name and the extension .seq).

(2) Send the data

Using the scp command in a Linux terminal, one would type, for the country "country" account:

scp C01HBa0001A01.tar.gz country@upload.sgn.cornell.edu:

which would upload the file C01HBa0001A01.tar.gz to the sgn upload account. Note that these accounts are special accounts which do not allow you to log in using an interactive session (telnet or ssh).

(3) Generate and send the checksum data

To verify that the upload went through correctly, a checksum is needed. The checksum should be calculated as follows:

md5sum C01HBa0001A01.tar.gz > C01HBa0001A01.tar.gz.md5

This will create a file with a md5 suffix, which should also be copied to the SGN server:

scp C01HBa0001A01.tar.gz.md5 country@upload.sgn.cornell.edu:

II. Standards for BAC Closure/Finishing

The standards for BAC closure/finishing have been adapted from the Medicago sequencing project. As new genomes are sequenced, it is clear that it is neither cost-effective in terms of gene discovery nor biological impact to finish every BAC to HTGS Phase 3. Guided by newer finishing techniques and evolving standards recently adopted by the NHGRI, NSF, and other funding agencies, we propose similar minimum finishing standards for the Tomato genome project. We will use the following suite of standard techniques to close as many gaps in BAC clones remaining after sequencing and assembly, with a goal of reducing overall error rate to fewer than one uncertain base per 10,000 bases.

- Re-sequencing short reads if they extend a contig.
- Re-sequencing failed mates if they fall into gaps.
- Re-sequencing and editing individual reads to higher accuracy.
- Additional reads, if deemed necessary, to improve the sequence quality of low accuracy regions
- Re-assembly at higher stringency.
- Primer walking directly off shotgun sub-clones or PCR amplified region.
- Modified dye primers.
- PCR across putative gaps to order and orient assemblies and, if possible, obtain the sequence of gaps spanned by these PCR products.
- Other techniques, including the use of modified nucleotides, enzymes, and chemicals, to read through GC-rich regions/hard stops may be attempted, but not repeatedly.

Extraordinary, labor-intensive techniques, such as transposon-based sequencing and micro (shatter) libraries may be employed, but not extensively. Thus, while we will make every reasonable effort to close all BACs, we recognize that not all BACs will reach HTGS 3. In those instances, regions approaching HTGS 3 will be fully annotated as such.

We propose the following minimal standards for "finished" sequence — With exceptions noted in annotation comments:
1. A single contig is generated.
2. The bulk of the sequence should be derived from multiple subclones sequenced from both strands. Less than 3% of the sequence should be derived from multiple subclones sequenced from the same strand with the same chemistry. These regions must pass manual inspection for any sequence problems, but do not need to be annotated unless the sequence quality falls below phred 30. Less than 1% of the sequence should be derived from a single subclone. In the case of a region covered by a single subclone, the clone must be sequenced either on both strands or with two different chemistries, and the region must be annotated.
3. More than 99% of the sequence has less than one error in 10,000 bp as reported by phrap or other sequence assembly consensus scores. Exceptions must be manually checked and have passed inspection for possible problems. These areas must be annotated.
4. The assembled sequence is confirmed by restriction enzyme digestion.
5. Where gap closure/finishing is difficult to complete, sequences should be submitted to GenBank as HTGS phase 2 (contigs ordered and oriented) with exceptions noted.
6. Estimates of the size(s) of any un-sequencable gaps should be noted.

Repeats

- In cases of simple repeat sequences, including single nucleotide repeats where the number of repeats can not be determined, the length of the repeat region should be estimated by restriction enzyme digestion or PCR.

- If large repeats cannot be resolved, the size of the repeat region, confirmed by restriction enzyme digestion or PCR, the nature of the repeats, the size of repeats, and the finishing problem should be indicated.
- Sequences of bacterial transposons and other obvious contaminants should be screened and deleted from the finished sequence; the size, sequence, and position of the deleted region should be indicated.
- Sequence should also be screened against chloroplast and mitochondrial sequence. Arabidopsis sequences may be used for that purpose. Criteria for determining whether the sequences represent real insertions into the chromosomal sequence have to be established. [Suggestions welcome].
- For BAC overlaps, we will aim for overlaps of 5-10 kb and will plan to finish to high quality (phase 3) at least 5 kb in the BAC that forms the overlap (ie joins second). We will also accept smaller overlaps (down to ~100 bases). However, such overlaps should be supported by fingerprints from other overlapping BACs and by PCR across the junctions.
- We will develop protocols and XML/DTD to document evidence supporting BAC overlaps. This will include noting the region of high quality overlap (100% identity) plus any discrepancies.
- Since one BAC will be finished before the other, in cases where the BAC overlap is between BACs finished by two different groups, the "owner" of the second BAC will be responsible for generating the data describing the BAC overlap. These data will be maintained both in association with the individual BAC annotation and in the centralized project database at Cornell, where it will be immediately available to both sequencing centers and the community at large.

Finishing Chromosomes and Defining Gaps on Chromosomes

We will be sequencing chromosomes starting from contigs that contain marker-anchored seed BACs. Contigs will be extended by a combination of BAC-end and fingerprint data. Gaps will be apparent when no clones extending contigs or spanning gaps between contigs appear to exist. When this occurs, we will:

- Search for extending BACs by hybridization of unique end probes (overgos or PCR products) to filters of other available libraries.
- Estimate gap sizes by fiber-FISH and attempt to span small (<10 kb) gaps with PCR.
- Order and orient contigs using marker, FISH, and other information.
- For large gaps recalcitrant to these approaches, use marker sequence information and comparative genomics to define likely sequences in the gap, and then use them as hybridization probes to identify candidate BACs for sequencing and gap filling. BACs that can be placed in gaps can become new seeds for contig extension.

BAC and chromosome Assembly pipelines

The phred/phrap/consed pipeline is recommended for assembly. The data formats for the archiving of the data corresponds to the output format of that pipeline.

III. Sequence release policy and submission to Genbank

The sequence release policy is conducted in the spirit of the Bermuda agreement (http://www.hugo-international.org/hugo/bermuda.htm), which states that sequence assemblies of greater than 1 Kb should be released to the public on a daily basis on a server such as a public FTP server.

Finished BAC sequences will be submitted to Genbank by each sequencing project. A tomato sequencing consortium will be defined at Genbank allowing every consortium member to update the sequences and the anntotations. Therefore, annotations can safely be submitted without the fear that they will not be updated later on. The initial annotations should satisfy the minimum annotation quality specified in the annotation chapter of this document.

All primary data is also archived on SGN. See chapter I, DataArchiving, for more information.

Refer to http://www.ncbi.nlm.nih.gov/HTGS/ for more information on HTGS Phases 0 - III.

Status	Location	Definition
Phase 0	HTG division	single-few pass reads of a single clone (not contigs).
Phase 1	HTG division	Unfinished, may be unordered, unoriented contigs, with gaps.
Phase 2	HTG division	Unfinished, ordered, oriented contigs, with or without gaps.
Phase 3	Primary division	Finished, no gaps (with or without annotations).

During the sequencing phase, the status of each BAC should be reported to SGN's BAC registry database, to prevent BACs from being sequenced twice and get an overview of the entire project status. The BAC registry database is currently being implemented.

IV. Gene Nomenclature Conventions

The biological community moves towards a unified system for naming genes. Earlier sequencing projects have already defined nomenclatures that should be adopted by other model organisms in order to move towards a more common genetic language that will facilitate structural, functional and evolutionary comparisons of genes and genetic variation among organisms. Recently, the rice sequencing groups have adopted, and developed further, many of the conventions introduced by the Arabidopsis and yeast sequencing projects. This document is based on the rice project gene nomenclature document, adapted for the Solanaceae.

The rules for tomato gene nomenclature used to date can be found on the web (http://gcrec.ifas.ufl.edu/tgc/newsletters/vol42/v42p6.html) and in the TGC 20:3-5 and TGC 23:3. These documents spell out the rules for naming chromosomes, linkage groups, genes, alleles, mutants and genetic rearrangements. The conventions given in this document should largely supersede most of the conventions given in the TGC.

Systematic Locus Ids

Term definitions:
Terms are used as defined by the Sequence Ontology (http://cvs.sourceforge.net/viewcvs.py/song/ontology/so.ontology?rev=1.45&view=auto).

Locus: a segment of chromosomal DNA that has been identified as containing a gene or pseudogene. The locus starts at the transcription start site and ends at the transcription end site. A locus name is a 'bag' that comprises all alleles and alternatively spliced transcripts. In this definition, the promoter is not part of the locus.
Gene or Gene Model: a gene model is the structure of a transcript mapped back onto the genomic sequence and that specific sequence. It corresponds to a specific transcript, e.g., a specific splice variant.
Pseudogenes: A sequence in the genome that has gene like characteristics but cannot be transcribed or translated (e.g, due to stop codons). [A more precise definition of pseudogene would be desirable].

Nuclear genes

BAC-based names: During the sequencing project, BAC-based systematic names will be assigned for predicted or experimentally verified genes on BACs. The BAC based names consist of the BAC name and a number, separated by a dot, for example "F12H11.20". Numbering is incremented by 10 for each gene and starts at 5' end of the finished BAC sequence and is irrespective of strand. The direction of the BAC in the assembled chromosome is not taken into account.

Chromosome-based names: After pseudochromosomes are built, a chromosome-based name will be assigned to each locus. The name will consist of the following: A two letter code for the species: SL for *Solanum lycopersicon* [Another possibility would be to use *Lycopersicon esculentum*, LE, but SL is preferred as the new name], followed by two digits indicating the chromosome (01-12). The next letter is a "G" for RNA and protein coding genes. It is a "T" for transposons and "R" for repeats. The transposon and repeat names use their own number space. The letter can also be used to denote other parts of the gene: "P" is used to denote the promoter of the corresponding gene (SL08p00010 is the promoter of gene SL08g00010).

[Another possibility would be to use the letter to designate the strand of the gene, as was done in yeast: W (Watson) for the forward strand in C (Crick) for the reverse strand].

This is followed by a five-digit number that corresponds to a sequential numbering from the top of the chromosome to the bottom independent of orientation. Initially, the numbers will be assigned in increments of 10 to leave room for further genes that may be identified later. [For repeats and transposons five digits may not be enough. The size of that number space will have to be defined at a later point.]

Examples : SL08g00010 is the first gene on the top arm of chromosome 8.
SL08r00010 is the first repeat on chromosome 8.
SL08t00010 is the first transposon on chromosome 8.

Chromosome-based names are in theory case insensitive. SL should always be capitalized as the lower case L can be confused with the number one.

Organellar genomes

The chromosome digits are replaced by the letter MT for mitochondrial genes and PL for plastid genes (this is a small change from the way rice handles organellar chromosomes. In rice, the letters M and P are used to designate the organellar genome, but I think it is better to always have the format SLCCDNNNNN, where CC always has two placeholders for identifying the chromosome, so that the core name is always 10 characters long).

Pseudogenes

Genomic sequences that have gene-like features but are not functional genes are termed pseudogenes [more explicit definition of pseudogene desirable]. These genes will be followed by .P suffix in the systematic locus name.

Transcripts

A locus can give rise to multiple gene products through alternative splicing and post transcriptional events. Different transcripts that arise through alternative splicing are designated by suffixes of the form: -1. The numbering is arbitrary and will in most cases reflect the order of discovery of alternatively spliced products. Caution is required when assigning alternatively spliced genes. Not every irreconcilable evidence should be taken to indicate alternative splicing.

Example SL08g00010-3 is an alternatively spliced transcript of locus SL08g00010.

Proteins

Proteins have the same identifiers as the corresponding transcripts.

Genes in sequence gaps

It is suggested that a number space be reserved for 200 genes per 100kb of predicted gap (that gives room for 20 genes with a increment of 10). On un-anchored BACs, only BAC-based names will be used until the sequence gap is filled.

Genes on BAC overlaps and incomplete genes at BAC ends

A problem with the BAC-based annotation approach is that there will be incomplete genes at the BAC ends and sometimes duplicated annotations on overlaps. No adjustment to the BAC sequence (such as addition of sequence from another BAC) should be made to include a complete gene on a given BAC, so that the BAC sequences submitted to Genbank will always match the BAC sequences in the databases (this is a problem with some of the Arabidopsis BACs which were changed and now don't correspond to the physical BACs). Using phrap scores the overlap with the better sequence quality should be determined in the chromosome sequence. If sequence qualities are identical, the sequence of the 5' will be used. Genes that are split between BACs will be annotated on the separate BACs as gene fragments belonging to one gene. A 'polishing' step will be required after chromosome assembly for these genes.

Gene Classification

We define five classes of gene annotations based on sequence similarity and availability of expression data.

Gene Class	Description	Annotation
“Known Gene”	The gene is identical to an already known gene.	Report name of gene (e.g. CHS)
“Putative function”	The gene has strong similarity to a known gene, meaning low evalues (e < 1e-100) in blast. Some degree of judgment by the curator is required for different gene families.	Putative + name of gene
“Similar to”	The gene has some similarity with a known gene, for example, evalues are between 1e-10 and 1e-100.	Similar to + name of gene
“Expressed Gene”	No similarity to any genes, (e.g., evalue > 1e-10), but the gene has expression data (associated ESTs, mRNAs etc).	Expressed Gene
“No evidence”	No evidence a part from the gene prediction is available.	No evidence

V. Guidelines for structural and functional gene annotations

The annotation will happen in three phases:

1) In the first phase, the every center is free to use any annotation system it has available. The features will be names on each BAC will be ad hoc identifiers in a namespace unique to each center. Annotations should not be submitted to Genbank.

2) In a second phase, an annotation system will be specified that will be used to annotate all the BACs. "Official" BAC-based names (see nomenclature section) will be used as identifiers. Tracking of merges, splits, etc will have to start.

3) In a third phase, chromosome-based names (see nomenclature section) will be attributed, using the official chromosome-based namespace. An authority will be designated that handles the attribution and tracking of chromosome based names.

1. Gene Structural Annotation

Structural annotations will be based on both intrinisic methods (computational predictions using gene finding programs) and experimental data from the alignment of known mRNA and EST sequences. The predictions and alignments are integrated into gene models.

The annotations will be performed on a BAC by BAC basis and BAC-based names will be assigned to each identified gene according to the gene naming conventions (see section II). Chromosome-based names will be assigned after the chromosome pseudomolecule sequences are assembled.

For the computational identification of protein coding genes GeneMark.hmm (Borodovsky and McIninch, 1993) and Fgenesh (Solovyev et al., 1994) are expected to give good results. GlimmerM (Salzberg et al., 1999) and Genscan+ (Burge and Karlin, 1997), and Eugene (pending availability for general use) are also recommended. The genefinders will be calibrated for tomato codon usage. The tomato calibrations should be shared with other centers for maximum consistency. tRNA genes will be detected using tRNAscan-SE (Lowe and Eddy, 1997). Other RNA genes, such as microRNAs and snoRNAs, will be identified through sequence similarity to known RNA genes.

For the experimental gene identification, Tomato cDNAs, ESTs and unigenes (and corresponding data from other Solanaceae species) to the tomato genome sequence Geneseqer (Usuka et al., 2000), sim4 or blast can be used.

2. Gene functional annotation pipeline

Several strategies should be combined to obtain the highest quality functional annotation from an automated pipeline: Identifying conserved protein domains, analyses of homology to other proteins, and protein location predictions (membrane domains, subcellular location). Wherever possible, the automated annotations should be converted to Gene Ontology codes for easier comparisons with the other annotated genomes.

For the protein domain identification, identification of Interpro domains using the software and domain databases available from the Interpro consortium (http://www.ebi.ac.uk/interpro) will be

used. Interpro domains are a selection of Pfam, TIGRfam, Prosite, ProDom and PRINTS domains that are grouped according to function and have defined evolutionary relationships. Each group of domains has an Interpro domain identifier in addition to the source database identifier (Pfam etc). The Interpro domain identifier will be used to convert Interpro domain annotations into Gene Ontology annotations using the mapping file provided by Interpro. The main functional categorization will be based on these Gene Ontology terms, which will be mapped to one of the GO slim vocabularies available from the GO consortium to produce a top-level classification.

In addition, protein functional annotations will be generated through analyses of sequence similarity using blast: Predicted tomato proteins will be blasted against Arabidopsis, rice, Medicago (pending availability), Swiss-Prot, and the Genebank non-redundant (nr) dataset.

The location of proteins will be determined using TMHMM (Krogh et al, 2001) for transmembrane domains, and TargetP for subcellular location. These results will be translated to Gene Ontology codes.

3. Resolution of conflicts

Many of the predictions and the experimental gene identifications will result in conflicting evidence for a given gene. Number of exons, lengths of untranslated 5' and 3' sequences, and joined or split genes will have to be resolved to make a gene call. In some cases, the conflicting evidence will be due to alternative transcription start sites and alternative splicing. Programs such as PASA (Haas et al, 2003) and GeneSeqer can be used to automatically determine the best gene model for the given data.

The instances that cannot be resolved using these tools will require manual curation. The main tool for the manual curation will be the open source Apollo genome viewer (http://www.fruitfly.org/annot/apollo). Similarly, conflicting information in functional annotations will have to be resolved. For the functional annotations, the similarity-based classification will be checked for consistency with the interpro derived annotations.

VI. Training Datasets and Quality Control

1. Training datasets for gene finders

A training dataset will be generated after several hundred BACs have been sequenced, and used to train gene prediction programs. [...]

2. Quality control

2.1.Assaying for contamination

The BACs should be screened for contamination from bacterial, chloroplast and mitochondrial sequences. The tomato chloroplast and mitochondrial genome are currently being sequenced by Italy and Argentina. Until the sequences become available, the Nicatiana tabacum chloroplast sequence (Genbank Z00044) can be used to asses for chloroplast contamination /insertion and the Arabidopsis mitochondrial genome sequence (Genbank NC_001284) can be used to screen mitochondrial contamination/insertion events.

Quality control procedures will be implemented in which BAC sequencing results from one center will be made available to other centers, and the assembly and annotation between centers compared.

VII. Data format standards

1. Structural Genome Annotation

Recommended format: GAME XML

GAME is an XML-based format that was developed specifically for Apollo. It covers structural and functional genome annotation and also stores meta information such as evidence. Because it is a rich format and the native format for the Apollo viewer, which most bioinformatics centers will probably use for structural manual annotation, this is the most logical choice for a project-wide standard to exchange genome annotation information.

For a complete DTD, see http://www.fruitfly.org/annot/gamexml.dtd.txt.

2. Raw Sequence data

- Chromatograms should be made available in scf format.
- Raw sequence data should be made available in Fasta files, including phrap score files.
- Each sequence should have associated a minimal information set was established that should always be associated with a gene call in Medicago as requested by the advisory committee. This includes originating center and unique id (the namespace:id pair), free text description, BAC seqversion, start/stop coordinates, evidence level, method and last modified date.

The medicago project has the following standardized fasta headers, and it is suggested that we adopt something very similar:

The FASTA header line consists of >, id, whitespace, description, newline. It should not contain more than 255 characters.
For the Medicago gene calls, the id will consist of a namespace:id pair separated by a pipe symbol, so:

>namespace|id

Example: >TIGR|ath1.m000123

It was decided that it will be useful if the id for Medicago gene calls includes "Mt" to designate the species, so

Example: >TIGR|mth1-13.m000123

Next a space is obligatory. The following free text description follows the FASTA conventions,that is it can include any characters except newline, though unnecessary use of special characters is not recommended as it can mess up HTML display. The description will usually contain a common name or human-readable designation of the gene call at hand.

Examples: Clavata1-like LRR-containing receptor-like protein kinase, unknown protein, automatic

gene prediction by FGENESH using the dicot model
The description is followed by an obligatory space and then the accession and version of the GENBANK/EMBL/DDBJ (BAC clone) sequence on which the gene call is based. The format is as in the SV field (seqversion) of EMBL, i.e. accession.version

Example: AC000123.2

After a space, the coordinates of the gene call on the sequence identified by the seqversion are given as start-stop, where start is the first nucleotide of the translation start codon ATG and stop the last nucleotide of the translation stop, e.g. TGA. Separated by -/dash/minus, no padding zeros, no whitespace. Coordinate 1 is always the first nucleotide of the sequence as retrieved using the seqversion from EMBL/GENBANK/DDBJ (reversing the sequence to achieve forward orientation relative to the chromosome is not allowed). Gene calls on Crick/reverse/- strand have stop > start.

Example: 12978-16483

After a space, a single letter code for the level of evidence that underlies the gene call is given. These codes are:

- F : full coverage/FL-cDNA: The complete gene model from translation start to translation stop is covered by expressed Medicago sequence, e.g. FL-cDNA or EST alignments across the full length of the coding sequence.
- E : expressed/EST matches: Expression of the gene is supported by Medicago EST sequence that matches the gene call (partially).
- H : homology/heterologous: the gene call is supported by similarity to Medicago or other ESTs, protein, FL-cDNA, genomic or other sequences with partial or full-length alignments.
- I : intrinsic/ab initio/inferred/hypothetical: the gene call is based only on intrinsic prediction tools such as FGENESH, Genscan or Eugene, and no significant alignments to other sequences are available.

The classification will be done top-down, so any gene call that does not fall under F will fall under E, if it does not satisfy the requirements of E it will be H and all gene calls that do not fulfill H will be called I.

Example: E

A space will be followed by a method abbreviation that mnemnonically shows the method used to generate the gene call. Usually, this will include the software name and the version or species matrix used. The abbreviation may not include spaces or other whitespace and should be 10 characters or less.

Examples: FgenesHdic, FgenHdicot, FgenesHmt, Eugene2, DeC!FR3, Intuition

A space will be followed by the date where the gene call was made or last modified in yyyymmdd format. This must be the last information in the FASTA header and must be followed immediately by newline.

Example: 20040610\n

Complete example:

>TIGR|mth10-123.m000123 expressed protein of unknown function (matched by TIGRGI|

TC006789) AC001234.2 1245-2642 E FgenH(at) 20040427
atgtacggatgcaatatcaacacctataccacatgggatatgtattt
tatagtccttgattactaa

For parsing, split on the spaces counting from the end is recommended, with a check that the individual fields correspond to the format (e.g. accession.seqversion, coordinates, one letter code, date).

2. Functional Annotation data using GO

The GO flat file format is recommended (http://www.geneontology.org/).

3. Microarray Data

MageML is recommended (http://mged.sourceforge.net/software/index.php).

VIII. Standard MOBY Services

Descriptions of standard MOBY services that are the basis for data exchange go here.