A draft proposal for the required Minimum Information about a high-throughput Nucleotide SeQuencing Experiment – MINSEQE

(April 1, 2008)

Nucleotide sequencing is now a mainstay technology in modern biology. Increasingly High-Throughput Sequencing (HTS) technology is gaining popularity and is being used not only for traditional applications in genomic and metagenomic but also for novel applications in the fields of transcriptomics, metatranscriptomics, epigenomics, and studies of genome variation. These latter types of experiments measure DNA or RNA in a particular biological state, or compare levels across several different biological states, such as cell types or disease states. Many of these applications are directly comparable to microarray experiments, for example ChIP-chip and ChIP-Seq are for all intents and purposes the same, other than the technology used to identify the immunoprecipitated sequences. Furthermore, both technologies generate vast quantities of experimental data, whose interpretation requires a solid understanding of the biological samples used, as well as the analyses carried out on the data.

Here we propose guidelines for the minimum information that should be reported about HTS experiments to enable unambiguous interpretation. Since describing such experiments is not fundamentally different from analogous microarray experiments, we can benefit from the well-developed and broadly accepted MIAME guidelines [1,2]. We propose that the following six elements of experiment description are considered essential for making available data supporting HTS based publications. In this document we do not distinguish between those sequencing platforms that generate short read sequences and those that generate longer reads. While the applications that they are being used for tend to differ, we contend that sufficient experiment annotation is necessary for both.

1. The description of the biological system and the particular states that are studied

Essential sample annotation, including the experimental factors and their values must be given. Experimental factors are the key experimental variables, for instance "time" in a time series experiment or “antibody” in a ChIP-Seq experiment. In addition to experimental factor values, essential information about the biological system from which samples were taken must be given, for instance, the organism (if known) and organism part, and what treatments have been applied.

2. The sequence read data for each assay

The complete set of read sequences for each assay as generated by the sequencing instrument, in a recognized format, with quality scores, raw intensities and processing parameters for the instrument.

3. The ‘final’ processed (or summary) data for the set of assays in the study

The final processed data are defined as the data on which the conclusions in the related publication are based. In many cases these data can be presented as a matrix with each row representing a genomic region (such as a gene or exon) and each column representing a particular biological state (e.g., a time point in a time course experiment), and each element in the matrix representing a measurement of the particular genomic region in the particular biological state. For instance, in transcriptomics experiments the final processed data are typically presented in a matrix characterizing the transcript abundance of each transcribed genomic region under the particular condition (analogous to normalised gene expression data matrix for gene microarray data). The identifiers used to annotate these processed data files should be traceable via the use of publicly available identifiers or chromosome coordinates along with the genome assembly build and version on which the data are based.

4. The experiment design including sample data relationships

The main purpose of the experimental design description is to specify the essential relationships between different biomaterials, such as samples and data files. For simple experiment designs this may be a table listing all samples and the respective raw data files and references to the respective objects in the summary data (e.g., columns in the gene expression data matrices). If relevant, it is important to show which assays in the experiment are replicates, and which are technical and which are biological replicates. More generally, the description of the experimental design can be thought of as a graph of the sample-sequencing run-data relationships.

5. General information about the experiment

General information about the overall study must be given. This includes a summary of the experiment and its goals, contact information, and any associated publication.

6. Essential experimental and data processing protocols

Data processing and analysis protocols must be described in sufficient detail to enable unambiguous interpretation of how the summarized data have been obtained from the raw data, and to enable scientists to reproduce the steps. For example, the algorithm used to align the reads against the genome should be named and the associated software's version number, run parameters, and genome assembly version captured. The description should also include details of a stepwise mapping strategy if one is used (e.g. a common strategy is to first map all reads to known transcripts and then to map remaining reads to the genome). If the creation of summary data requires additional algorithms (e.g. to transform mapped reads into transcript counts), these must also be described.

While HTS technology is still maturing, it is recommended that detailed information about all relevant technology parameters used in a given experiment be provided. The protocols used to isolate and/or amplify the template library (clones) used for sequencing must be adequately described. Specifically, protocols for any and all rounds of PCR template amplification as well as any techniques used for genomic partitioning or genomic enrichment should be given. Partitioning techniques might include multiplex long-range PCR, solid-phase pull-down (using either beads or microarrays), molecular inversion probes, rolling-loop amplification, or other strategies. When partitioning protocols employ normalized pools of primers or hybridization probes, both the identity and the stoichiometry of the probes/primers used should be described or referenced. When a microarray is used for pull-down or when probes are cleaved from or otherwise processed on a microarray, an array design file along with a description of the array processing steps should be provided or referenced. Finally, if and when an indexing scheme is used (where synthetic nucleotide sequences are appended to templates for the purpose of disambiguating pooled samples), the complete indexing strategy should be described. This description should include the sequences of the indices, the protocol used to attach indices to the templates, and the sequencing strategy used to read the indices (i.e. whether a single sequencing primer was used to read both the template and the index or whether a second sequencing primer was used to sequence the index). As the HTS technologies mature and become more standardised, the description of some of these protocol parameters may become redundant.

References.

[1] Brazma et al, Minimum information about a microarray experiment (MIAME)—toward standards for microarray data, *Nature Genetics*, 29, 365 - 371 (2001).
[2] http://www.mged.org/miame