Description of Sample Sheet for CRISPR-DAV Pipeline

The CRISPR sample sheet describes all samples in a project. The sample sheet is required for CRISPR pipeline. It’s used by prepare\_run.pl to create input files and wrapper script run.sh for the pipeline.

N-to-N relationship of the fields:

* The sample sheet contains only one project ID.
* A project contains 1 or more amplicons.
* One amplicon can have:
  + more than one gene;
  + only one reference genome;
  + 1 or more guide sequences.
  + 1 or more samples
* Guide field can have 0, 1 or more guide sequences
  + 0: The guide sequence field is empty. This is for control sample (although the guide sequences can be entered as well for control sample). All guide sequences in the amplicon will be applied to the sample. This is the same as: “guideA,guideB”.
  + 1 guide sequence: consists of ~20 bases (ATCG). Low or upper case is OK.
  + 2 or more guides:
    - guideA,guideB: guide A and B will be analyzed separately and go into separate result pages (index.html).
* Sample: each guide field is associated with 1 sample. Within one amplicon, the sample names must be unique.
* Guide vs HDR: a guide can have 0 or 1 HDR. If one sample has HDR, all samples that have the same guide will be analyzed with the same HDR; if no sample has HDR, then none of the samples will be analyzed with HDR.

Format and requirements for the fields:

0a. There are two header rows. The first row is instruction. The second is column header. The names of the column names are not used by pipeline, but the positions and order of the columns are. Inserting or deleting a column would cause error when running the prepare\_run.pl.

0b. Data rows starts after the 2 header rows. Spaces are allowed in all fields. However, they will be stripped off by the program.

1. Column A - Gene symbol: One word. For standard genome, if a RefSeq gene table is provided, and the amplicon coordinates can match the coordinates of one of the genes in the table, the symbol in the RefSeq table will be used. If that gene is not found, the gene symbol provided by user will be used.

1. Column B - Genome: The allowed genome name would be checked against the conf.txt. Case sensitive. When the genome name is not found in the conf.txt, error occurs.
2. Column C - Amplicon range: format <chr>:<start>-<end>. The chr format follows UCSC format, i.e. it starts with alphabet and is not just a number as in GRC. It’s ok if the start is greater than end, or if the start or end contains “,”, or if chr11 is written as Chr11. However, for non-standard chromosome names, like the ones used in criGri1, the chromosome name must match exactly.
   * Make sure the rows with the same amplicon range should have the same combination of gene symbol + genome. This prevents typo in gene symbol and genome.
3. Column D - sgRNA sequence:
   * Case insensitive. Allowed characters: ATCGatcg and comma.
   * It’s OK for a cell to be empty or has only spaces. Usually that is suitable for control samples. In this case, the sample will be analyzed together with all other samples.
   * It’s OK to have multiple sgRNA sequences separated by comma, e.g. CTAGGAGTCAGCGACATATG, AAAGGAGGTCACGACATATG. However, the guide will be analyzed separately, and their results go into separate html pages.
   * Make sure the rows with the same sgRNA sequence have the same value for the combination of genesymbol + Genome + Amplicon Range. This can prevent a lot of typos, such as the situation where the genome or amplicon ranges increment by one caused by dragging in Excel.
4. Column E (HDR Base Changes):
   * It’s OK for the cell to be empty, if there is no HDR to study.
   * Base Changes should be dependent on sgRNA sequence. If a string value is present, it should have this pattern:  pos1Base1, pos2Base2… like 709734A,709734C,709752C. The position is genomic coordinate. The base letter is the intended base after repair. The position and base refer to the positive strand only. If an HDR is entered in one sample, all other samples of the same guide sequence will use the same HDR.
5. Column F (Sample Name):
   * Sample name should have some meaning so that it can help tell what the sample is about. Sample name must be unique within the same amplicon range.
   * Allowed chars: letters, numbers, - and \_
6. Column G (Sample ID):
   * Sample ID is used to identify fastq file names. Sample ID is the initial part in fastq file that unambiguously identifies read 1 (and read2) fastq files that belong to the sample. Error will occur because there are more than 2 files per sample are included.
7. Column H (Project ID): OK to be empty.

The column must be the same for all samples. The project ID is optional. If specified, the program will save it for downstream use if any.

1. Column I (Fastq Path): The paths can be all the same or different for the samples. The actual fastq files must end with .gz. If no path is provided here, a path must be provided via command line.