**SampleSheet parser question**

**Why are we asking for a programming test?**

* We are interested in your credentials and would like to evaluate your programming skills
* We would like to understand your software design practices and your approach to solve the problem
* Additionally, this will allow you to show off your programming skills!

**Project Overview:**

A SampleSheet file is a plain text, CSV, meta-data file, that represents samples to be sequenced through an Illumina sequencer. This file is a required input file for sample data generation.

The first line in the SampleSheet file is always a header, which represents field names for various meta-data. Records for each sequenced sample follow as rows below the header line (the order of meta-data fields for a given sample are always in the same order as the header).

For instance, “Lane” field defines physical location where sample is sequenced, and can have only two values ‘1’ or ‘2’. If sample (SampleID header field) is considered as a pivot, for each sample meta-data line in “Lane 1”, there has to be a matching sample meta-data line from “Lane 2”. See below, for an example SampleSheet, the color represent same sample entries from “Lane 1” and “Lane 2”. The ordering of columns is also strictly enforced. The figure below shows an example of the correct order of columns: FCID is first, then Lane, then SampleID and so on. Same colors indicate a specific sample repeated in Lane1 and Lane2.

FCID,Lane,SampleID,SampleRef,Index,Description,Control,Recipe,Operator,SampleProject

H9EX,1,123-T,HUMAN,CGATGT,Tumor,N,,AY|John;Doe|M123|Male,IMPACTv3-20140001

H9EX,1,123-N,HUMAN,CGATGT,Normal,N,,AY|John;Doe|M234|Male,IMPACTv3-20140001

H9EX,1,234-T,HUMAN,CGATGT,Tumor,N,,AY|Johnny;D|M321|Male,IMPACTv3-20140001

H9EX,2,123-T,HUMAN,CGATGT,Tumor,N,,AY|John;Doe|M123|Male,IMPACTv3-20140001

H9EX,2,123-N,HUMAN,CGATGT,Normal,N,,AY|John;Doe|M123|Male,IMPACTv3-20140001

H9EX,2,234-T,HUMAN,CGATGT,Tumor,N,,AY|Johnny;D|M123|Male,IMPACTv3-20140001

**Problem statement:**

SampleSheets generated manually by lab techologists can contain errors. A common error is when a sample is listed in “Lane 1”, but is missing in “Lane 2”. A second common error is incorrect ordering of columns in the SampleSheet. In the example below, the SampleID and Lane columns are switched, the Operator and SampleProject columns are switched, both in the header and in the data.

FCID,SampleID,Lane,SampleRef,Index,Description,Control,Recipe,SampleProject,Operator

H9EX,123-T,1,HUMAN,CGATGT,Tumor,N,,IMPACTv3-20140001,AY|John;Doe|M123|Male

H9EX,123-N,1,HUMAN,CGATGT,Normal,N,,IMPACTv3-20140001,AY|John;Doe|M234|Male

H9EX,234-T,1,HUMAN,CGATGT,Tumor,N,,IMPACTv3-20140001,AY|Johnny;D|M321|Male

H9EX,123-T,2,HUMAN,CGATGT,Tumor,N,,IMPACTv3-20140001,AY|John;Doe|M123|Male

H9EX,123-N,2,HUMAN,CGATGT,Normal,N,,IMPACTv3-20140001,AY|John;Doe|M123|Male

H9EX,234-T,2,HUMAN,CGATGT,Tumor,N,,IMPACTv3-20140001,AY|Johnny;D|M123|Male

**Part 1:**

Create a command line tool using an Object Oriented Programming (OOP) approach, that will read in a SampleSheet file, check it for errors and consistency given the SampleSheet requirements above, and output the SampleSheet fields in correct order. Again, the correct order of the fields is:

FCID,Lane,SampleID,SampleRef,Index,Description,Control,Recipe,Operator,SampleProject

**Part 2:**

Assuming that the lab technologists have loaded their created SampleSheets into a MySQL database (or any other database of your choice), extend your code from Part 1 to retrieve the loaded SampleSheet from the database, check it for errors/consistency and output a valid SampleSheet. You may assume:

* The getDBHandle() is provided to you and will connect to the appropriate database.
* The database table name is “samplesheet”, and fields exactly match the header field names.