DIGITAL SKILLS ON COMPUTATIONAL BIOLOGY FOR HEALTH PROFESSIONALS

**Project No: 591945-EPP-1-2017-1-DE-EPPKA2-SSA**

# R3.2. BioS assessment exams

**WP3: BioS Educational material**

Authoring Partner: P11 HiDucator

Co-funded by the Erasmus+ Programme of the European Union

**Project information**

| | |
|---|---|
| Project acronym: | BioS |
| Project title: | Digital Skills on Computational Biology for Health Professionals |
| Agreement number: | 2017-3424/591945-EPP-1-2017-1-DE-EPPKA2-SSA |
| EU programme: | Erasmus+ KA2-Sector Skills Alliances (SSA) |
| Project website: | bios-project.eu |

**Prepared by**


| | |
|---|---|
| Name: | Csaba Ortutay, Martti Tolvanen, Laura Dorn, Hartmut Schröder |
| Authoring Partner(s): | HiDucator |
| Position: | CEO |
| Date: | 10/December/2019 |

# Executive Summary

Assessment material is offered for all the four modules covered by the BioS program as follows:

1. Four (4) self assessment practice
2. Twenty (20) peer-assessment assignments
3. Two (2) final exam topics
4. Ten (10) multiple choice question for automated final exam

All the assessment material is suitable to be implemented in virtual environments and are ready to complement the educational material from the four modules of the BioS educational program.

| Dissemination Level | | |
|---|---|---|
| **PU** | Public | |
| **PP** | Restricted to other programme participants (including Commission services and project reviewers) | X |
| **CO** | Confidential, only for members of the consortium (including EACEA and Commission services and project reviewers) | |

# Table of Contents

# List of Tables

# 1. Introduction

The purpose of this document is to offer assessment material to both students and educators taking part in the BioS program titled “Digital Skills on Computational Biology for Health Professionals”. The idea behind this material is to help the learning process, to assist the evaluation of how far the student reached in understanding the offered skills and concepts, and, finally, to help to arrange classic essay type of exams if the program is organized in a traditional educational framework.

# 2. Modes of assessment

Three different modes of assessment are offered for all the four modules in this program (see table 1). These modes are targeting three different phases of the learning process, therefore they are targeting different audiences. All the three assessment modes are suitable to be implemented in both a virtual, or classroom based teaching solutions.

*Table 1: Modes of assessment in BioS training program*

| Mode of assessment | Target audience | Purpose | Usage |
|---|---|---|---|
| Self-assessment practices | Student | Learner can evaluate the depth of knowledge covered in the module | After watching the videos and reading through the learning material, the student is offered a practice where the learned concepts should be applied. The student can use this practice to check possible gaps in their knowledge or to learn about further details of the covered topic. This mode is part of the learning process. |
| Peer-assessment assignments | Student/Peer/Educator | Student can get external feedback about their knowledge, peers can get alternative points of views in selected topics, educators can arrange semi-automated evaluation of large number of students. | After finishing studying a module, the student answers focused topic related questions. Answers are evaluated either by peers or optionally by educators, offering direct feedback to students. This mode of study is suitable for implementing in MOOC environments. |
| Final exam | Educator | Student has to give an essay type of answer about a complex topic covering several aspects of the material covered in a module. | Student gets a topic and write a moderate-length essay. Educator evaluates the answer and offer feedback to the student. This evaluation mode is suitable for more in-depth evaluation with small group implementation of the BioS program. |
| Final exam quiz | Student | The students' basic knowledge in the material can be evaluated automatically. | If the program is organized as automatically run MOOC, these questions can replace essay type of student evaluation. |

## Self-assessment practices

These practices target the learners themselves. The idea behind the practices is that they should assist the students to understand the practical relevance of the concepts covered by the modules. By doing these practices, the student can understand whether the depth of their knowledge is satisfactory for the practical applications of the covered concepts.

## Peer-assessment assignments

These are focused questions which can be answered in a few sentences. It is important to have questions targeting both the lower and higher levels of understanding of the material to give good insights how well the student understood the material. This mode of assignment is suitable for MOOC implementation, where the student supplied answers are evaluated by multiple peers.

## Final exam

Traditional, essay type of assignments are suitable for student evaluation in more traditional frameworks, such as classroom based or supervised small-group on-line course implementation of the BioS program. Here, students are writing a moderate-length essay of a topic related to the module, where the student has an opportunity to demonstrate the depth of knowledge on the covered material.

## Final exam quiz

If the BioS program is organized as a MOOC, essay type final exams are not possible to be used. In these cases assessment should be offered in the form of multiple choice questions with one or more correct answers. The advantage of this version of final exam that it can be completely automated, therefore they are usable with very large number of students or in situations tutoring or professional supervision is not available.

# 3. Assignments for BioS modules

## Module 1: Introduction to Bioinformatics

This module provides basic knowledge of how molecular data connects to modern biomedicine in order to give an understanding of background of genomic/personalized medicine. The course introduces the student to genomic data, DNA and protein sequence data and protein structures, major biological databases and teaches basic methods for their analyses.

## Self-assessment practices

1. Collect a list of at least 5 websites which offers basic bioinformatics tools. List the URLs of the sites, the web tools they offer, and what those tools can be used for.
2. Assemble a table of at least 10 on-line databases which offer molecular biology related sequence data. Provide the database URL, what kind of sequences they collect and what is their target audience.
3. Collect 5 methods offered by bioinformatics which are frequently used in Omics research. Explain which Omics they are targeting and what kind of insights they are able to offer.
4. Compile a list of 10 questions you would ask about the training material if you would be a teacher checking the knowledge of your students. Remember, you are expected to ask questions at all levels of understanding the material: starting from simple 'remember the fact' questions up to the levels where students should be able to judge and compare critically the component of the covered material.

## Peer-assessment assignments

All these questions should be answered in no more than 350 characters.

1. Mention some application of next-generation sequencing methods.
2. What is the biggest informatics related challenge when analysing experimental data from next-generation sequencing methods?
3. Mention three Omics fields and explain what they are doing.
4. Name three on-line databases which provides biological sequences. What kind of sequences can you find there?
5. Mention three on-line databases which provides molecular medicine related biological data, but other than sequences only.
6. Explain briefly what sequence alignments can be used for in bioinformatics research.
7. Explain briefly how genome browsers are useful in molecular biology related research.
8. How the analysis of transcriptomes is useful in molecular medicine related research?

9. What experimental methods can you name which can analyse gene expression levels in patient samples?
10. How can high-throughput sequencing methods be used in transcriptmics studies?
11. What kind of information about a human gene can be found from on-line databases? Name at least 5 distinct data items.
12. How can high-throughput sequencing methods help disease diagnostics?
13. How have genomics data opened up new ways of the research of disease mechanisms?
14. How have genomics data opened up new ways of disease diagnostics?
15. Explain what is the depth of sequencing in an RNAseq experiment.
16. How should be the depth of sequencing of an RNAseq experiment determined. What factors should be taken into account?
17. What is the size of the human genome and how it was measured?
18. Briefly explain the purpose of epi-genomics experiments.
19. Name three experimental methods which can be used in epi-genomics studies. What are their main purpose?
20. How can high-throughput sequencing methods be used in epi-genomics studies?

## Final exam

Choose on from the following topics and write an essay of about 1000 words.

1. Explain a case where bioinformatics can provide insights into the mechanisms of diseases. It does not have to be an actual historical case, but you should be able to explain how bioinformatics can contribute to the research of that mechanism.
2. Explain why on-line available sequence and associated data is useful for exploring the molecular biology of immune responses. Assuming that you are tasked to study the reaction of immune system when administering a new drug, explain how on-line databases can assist your task.

## Final exam quiz

The following assignments are multiple choice questions where students should find the correct answer(s) among the offered alternatives. Correct choices are underlined here.

1. What is the main focus of bioinformatics?

1. It is a domain of science dedicated to the computational handling of biological data.
2. It is using biological entities for large scale calculations.
3. It is the development of genetic algorithms.
4. It is a collection of modern sequencing experiments.
5. It is IT support for the computers of biology researchers.

2. Why bioinformatics relies so much on statistics?
    1. Because p-values are demanded by most biomedical journals.
    2. Because bioinformatics methods were developed mostly by mathematicians who like statistics.
    3. Because modern biological experiemnts generate so much data that even extreme values can be found. The significance of a given observation can be calculated only using statistical inference.
    4. Because software developers creating tools for biological data analysis are very good with numbers and statistics is about numbers.
    5. Because big data is a subdomain of statistics and bioinformatics is a subdomain of big data.
3. Which of the following are not sub-domains of bioinformatics?
    1. RNAseq data analysis
    2. Systems biology
    3. X-Ray crystallography
    4. Epigenomics analysis with antibodies
    5. Reconstructing and analysing the 3D structure of proteins
4. Which of these are bioinformatics methods?
    1. Nucleotide and protein sequence analysis
    2. Aligning sequences
    3. Gene prediction in genome sequences
    4. Protein-protein interaction prediction
    5. All of these
5. What kind of information can be found both in the Online Mendelian Inheritance in Man and the Genetic Home Reference databases?
    1. How to diagnose selected diseases
    2. Disease descriptions in lay-man's term
    3. Genetic variations behind the inheritance of selected diseases
    4. The detailed molecular mechanism behind diseases

6. What are the most important differences between the Genetic Home Reference and the RefSeq databases
    1. They are maintained by entirelly different organizations
    2. RefSeq targets patients as their audience while Genetic Home Reference mostly for medical doctors
    3. RefSeq is organized around sequences while Genetic Home Reference is focusing on diseases
    4. They are essentially the same, but RefSeq is the US version of Genetic Home Reference database maintained at the National Institute of Health
7. Which methods are not used anymore in biomedical sequencing experiments?
    1. Exome sequencing
    2. Microbiome sequencing
    3. Protein sequencing
    4. RNA-seq
    5. CHIPseq
    6. Single cell sequencing
8. What can we learn about a protein using mass-spectrometry?
    1. Experimentally demonstrate the existence of the protein
    2. If the protein went through post-translational modifications
    3. The 3D structure of the protein
    4. What is the biological function of the protein
    5. If the protein is an enzyme and what are their substrates
9. What kind of process can act at a given genomic site where you can not observe any variation across multiple closely related species?
    1. Purifying selection
    2. Positive selection
    3. Balancing selection
    4. Neutral selection or drifting
    5. None of these
10. In what kind of projects can you utilize multiple sequence alignments of protein sequences?
    1. Observing which parts of those proteins are spared by mutations
    2. Producing evolutionary trees of the hosts of the proteins
    3. Classification of functionally related enzymes
    4. To identify functional amino acids in the active centers of enzymes

5. To identify which parts of those proteins should be protected from mutations

# Module 2: Computational Statistics for clinical doctors

This module will provide a practical introduction to analysis of biological and biomedical Big data, in order to develop a critical understanding of the reliability of analysis results. Clinical doctors will learn to appreciate how the R statistical environment can be applied to biological data analysis in a cost-efficient manner.

## Self-assessment practices

1. You are working with a large patient group within a large medical study. Prepare a list of variables you can use to describe the patients. Pay attention that you use all three qualitative and both quantitative variable types. Collect at least two for all the five variable types. In your list, name each variable, provide an explanation about it, offer some realistic example values, and explain the usual descriptive statistics which can be used to describe such variable in a real life study.
2. Consider the following research paper: Vuik, Sabine I et al. "A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population." Population health metrics vol. 14 44. 25 Nov. 2016, doi:10.1186/s12963-016-0115-z. It is freely available from PubMed Central: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124281/

   Find Table 1 in the paper, and collect the variables they have used to describe patient populations. For each statistical variable, identify which variable type it is (such as nominal, or continuous numeric, etc.), and what kind of statistics are used in the paper to summarize them.
3. Do the following exercise:
   1. Open Excel and create an empty file. (This practice should work the same way with LibreOffice Calc or OpenOffice Calc or Google Docs Spreadsheet.)
   2. Put the following formula into the top left (A1) cell:

      =RAND()

      This should generate a random number into that cell.
   3. Highlight A1 cell and pull it to the right so that the random number function is copied to all cells to column A-J (A1:J1).

4. Highlight the first row (A1:J1) and pull it down so that the row of random number functions is copied all the way down to row 1000. At this point you should have a table of random numbers, 10 columns (A-J) and 1000 rows.
5. Go to cell L1, and insert the following formula:

   =TTEST(A1:E1;F1:J1;2;2)

   This will perform Pearson t-test in row 1 for the first 5 values (columns A-E; A1:E1) against second 5 values (columns F-K; F1:J1). The last two parameters in the function specifies that this is a two tailed t-test on two unrelated samples. (This is the t-test variant most often employed in molecular biology studies.) After inserting this formula, you should see a number in this cell in the range of 0-1. You should interpret as the p-value of the t-test for the first row.
6. Highlight the L1 cell with the t-test formula, and pull it down so that it is copied all the way down to row 1000. This way you have just performed 1000 t-tests.
7. Go to an empty cell on the spreadsheet and insert the following formula:

   =COUNTIF(L1:L1000;"<0.05")

   If your software uses coma instead of decimal point, use this: =COUNTIF(L1:L1000;"<0,05")

   This will count that how many p-values show significant tests. (In Excel and LibreOffice Calc you can hit F9 to regenerate the random numbers and see how the number of significant tests change.)
8. Write down what exactly those t-tests do, and how to interpret the p-values what they provide. Explain what significant p-values means in medical studies where t-tests are used. Provide some examples. Explain what the number of significant p-values mean in this exercise. Search for the term “multiple hypothesis testing” and explain the term's relation to this exercise. Provide examples from biomedical research where thousands of t-tests (or similar hypothesis testing method) are employed in a single study. Write down what is the solution there to this problem.

4. Let's practice how to do linear regression with real life data:
   1. The following data is from the US. Department of Agriculture and the US National Science Foundation (see table 2).

*Table 2: Practice data for regression analysis*

| Year | Pounds per capita consumption of mozzarella cheese (US) | Civil engineering doctorates awarded (US) |
|---|---|---|
| 2000 | 9.3 | 480 |
| 2001 | 9.7 | 501 |
| 2002 | 9.7 | 540 |
| 2003 | 9.7 | 552 |
| 2004 | 9.9 | 547 |
| 2005 | 10.2 | 622 |
| 2006 | 10.5 | 655 |
| 2007 | 11 | 701 |
| 2008 | 10.6 | 712 |
| 2009 | 10.6 | 708 |

2. Use any spreadsheet software or R to enter this data. Create a scatterplot (XY plot) to visualize the data with the per-capita cheese consumption on the X axis.
3. Calculate the Pearson correlation between these variables.
4. Fit a linear regression model to these variables. Use this model to predict the number of doctorates in the US for 2012-2015, where the cheese consumption was the following: 10.68, 10.73, 11.16, 11.27 pounds per capita.
5. Write a detailed explanation what this regression model tells us, and how good it is to predict the number of doctorates. (The real numbers for those years are as follows: 714, 806, 817, 882 doctorates.) Try to interpret these results in the light of these terms: common causation, coincidence.

## Peer-assessment assignments

All these questions should be answered in no more than 350 characters.

1. Name the most important differences between descriptive and inferential statistics methods.

2. Explain the relationship between population and sample in statistical inference. Relate these terms to a medical study targeting a disease mechanism. What would be there the population and the sample?
3. What the following measurements tell you about your population? Mean, median, range, standard deviation.
4. Estimate the following measures of the height of people in your country: mean, first and third quartile, minimum and maximum values. Would the median be below or above the mean? Explain why.
5. Explain how R can be useful for statistical analysis in your own field of study/specialisation.
6. What is the power of a statistical test? How it is relevant for sample size and the differences we try to estimate between populations?
7. Describe an example case where t-test is a useful test in you own field of study/specialisation. Remember to check what data types t-test can be used!
8. How to use p-values to evaluate the significance of statistical findings in studies?
9. What are the most important similarities and differences between t-test and ANOVA?
10. Explain the differences between parametric and non-parametric hypothesis tests.
11. Provide 3-3 examples for medical measurements where parametric and non-parametric tests are suitable.
12. Describe a case from medical research where correlation is the result of direct causation.
13. Describe a case from medical research where correlation is NOT the result of causation.
14. Explain how linear regression can be used to predict values in a molecular biology experiment.
15. Explain the relevance of two parameters coming from linear regression calculations. (Slope and intercept.)
16. Mention non-linear regression models together with cases where they are used.
17. Explain how to visualize results from regression modelling.
18. Describe how to assemble contingency tables for categorical data.

19. What are the practical differences between the Chi-square test and Fisher's exact test?
20. Explain the relevance of odds ratio and relative risk in a study which tries to estimate the effect of an SNP on the onset of a medical condition.

## Final exam

Choose on from the following topics and write an essay of about 1000 words. (But read all the topics to get a larger context of the tasks.)

1. You are involved with a project about the relevance of an SNP affecting of a hereditary medical condition. The hospital in the project collected data from patients with and without this SNP, and provides you how many of those are with and without this condition. What kind of statistics can you employ to see if there is association between the SNP and the condition? Describe how would you analysed the data, and what would be your conditions to accept or reject this idea.
2. In the second stage you work with a lab which performs a large screening with 200-200 patients with and without the SNP, and measuring the gene expression levels of all 20000 human genes. Explain the statistics which should be employed in this study (let's assume that gene expression levels show normal distributions of continuous values after appropriate transformations). Explain how you choose candidate genes which you should investigate in further studies.
3. In the third stage of this study you have selected a gene which might be involved with the disease. You measure not only the gene expression levels in the patients, but also the severity of the investigated medical condition, which can be described with a continuous numeric variable. Explain how to use regression modelling with this data, how to evaluate the results, and how to test if measuring gene expression levels in patients can be used to set up a diagnostics procedure.

## Final exam quiz

The following assignments are multiple choice questions where students should find the correct answer(s) among the offered alternatives. Correct choices are underlined here.

1. Which data visualization technique is usable to explore the relationship between two quantitative variables?
    1. Histograms
    2. Box plots

3. Bar graph
4. Scatter plot
5. Pie chart

2. Which of the following statistics measure the central tendency of a quantitative variable?
    1. Mean
    2. Standard deviation
    3. Mode (modus)
    4. Variance
    5. Median
3. Which of the following statistics measure of the dispersion of data from a quantitative variable?
    1. Interquartile range
    2. Variance
    3. Box-plot
    4. Standard deviation
    5. Ratio of outliers
4. What features of a data distribution is visible from a box-plot?
    1. Mean
    2. Median
    3. Interquartile range
    4. Outliers
    5. Mode (modus)
5. What is the main goal of hypothesis testing in statistics?
    1. To decide if observed data supports our research hypothesis.
    2. To find data which supports our research hypothesis.
    3. To exclude observations from an experiment which goes against our research hypothesis.
    4. To calculate the probability that observed data patterns are results from the stochastics of random sampling.
    5. To calculate the probability that our observations are representing the truth.
6. We compare measurements of a normally distributed variable in two independent groups. We are interested whether the means of this value are different in the compared groups, so we apply hypothesis testing. What is the traditional “null hypothesis” in this situation?
    1. That there are differences between the groups.

2. That there are no differences between the groups.
3. None of these.

7. What is the main goal of performing linear regression in statistics?
    1. To find a mathematical model for predicting a quantitative measure from another quantitative measure.
    2. To draw a line on a scatterplot.
    3. To explain most of the variance in the spread of data.
    4. To maximize the R-squared statistics regarding two normally distributed measurements.
    5. To minimize squared errors of points around a line described by a linear equation.
8. You create a linear regression model using several independent and a single dependent variable. Which statistics tells you that how much of the variance in the dependent variable can be explained by one particular regression model?
    1. The betas associated with the independent variables
    2. The R-squared statistics
    3. The p-values
    4. The confidence intervals of the betas
    5. The slope of the calculated line
9. What is the statistics we can use to demonstrate the dependence between two quantitative variables?
    1. Association
    2. Chi-square test
    3. Confidence intervals
    4. Correlation coefficient
    5. Expected values
10. When you perform hypothesis testing, you calculate p-values to figure the significance of your findings. How the p-value changes with the sample size? (Let's assume that you do e.g. a t-test for two groups of normally distributed values, and that there are real differences between the means across the two groups.)
    1. The smaller the sizes of the two samples, the smaller the p-value is.
    2. The larger the sizes of the two samples, the smaller the p-value is.
    3. There is no relationship between sample size and p-values
    4. P-values are always below 0.05 (5%) in this situation.

5. Sample size should be always 3 in the two groups, to minimize experimental/measurement costs.

# Module 3: Personalized genomics in patient care

The purpose of this module is to provide medical doctors the necessary knowledge and skills to interpret results from personalized genomics services. This module facilitates integrating genomics based methodology into medical patient care activities.

## Self-assessment practices

1. Consider the following article: Hua L, Lin H, Li D, Li L, Liu Z. Mining functional gene modules linked with rheumatoid arthritis using a SNP-SNP network. Genomics Proteomics Bioinformatics. 2012;10(1):23–34. doi:10.1016/S1672-0229(11)60030-2. It is freely available from PubMed Central: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054489/

   This research has identified 15 single nucleotide polymorphisms associated with rheumatoid arthritis. Find the supplementary table showing these SNPs. Go to dbSNP database (https://www.ncbi.nlm.nih.gov/snp/) and find all those 15 polymorphisms. Prepare a table with the following information: the nucleotide variants these SNPs represent, the variants' frequency in the population, and if they are associated with protein coding genes. Can you design a proper diagnostic method based on these results?

2. Record the pedigree tree of a family you know sufficiently well (perhaps your own family). Try to find traits with clear dominant-recessive Mendelian inheritance showing up on the pedigree. You can consider one of the following traits: eye colour, blood type, earlobe attachment, latctase persistance, or any other similar trait. Go to the Online Mendelian Inheritance in Man database (https://www.omim.org/) and try to find information about the trait itself. In reality, is it true that this particular trait is affected by a single gene, or there is a more complex genetic mechanism behind it?

3. Use the Online Mendelian Inheritance in Man database (https://www.omim.org/) in this exercise. Prepare a table by collecting information about the following genetic variation types: numeric and structural chromosomal variants, copy number variants, single nucleotide variants, nucleotide insertions/deletions. Find at least one disease which is associated by such variations. Find the corresponding variation. How the presence or absence of the variation affects the probability of onset of the associated disease? Find out what diagnostics is available based on that variation.

4. Compile a list of 10 questions you would ask about the training material if you would be a teacher checking the knowledge of your students. Remember, you are expected to ask questions at all levels of understanding the material: starting from simple 'remember the fact' questions up to the levels where students should be able to judge and compare critically the component of the covered material.

## Peer-assessment assignments

All these questions should be answered in no more than 350 characters.

1. Name the most important features of Mendelian inheritance.
2. What models of inheritance you are aware of in addition to the Mendelian model?
3. Name three human traits with dominant/recessive inheritance.
4. What is the molecular/cellular mechanism causing somatic mosaicism?
5. Name diseases with X-linked inheritance. Describe the gender specific manifestation pattern of the disease.
6. Name the most important genetic variants identified in association with inherited diseases.
7. Describe the mechanism how the presence of a single nucleotide polymorphism can cause a disease. Mention an example.
8. Describe the mechanism how the presence of a copy number variation can cause a disease. Mention an example.
9. Describe the mechanism how the presence of a numeric chromosomal variant can cause a disease. Mention an example.
10. What databases can you use to find information about a given single nucleotide polymorphism and the disease associated with it?
11. What diseases can be diagnosed by using karyotype analysis?
12. What is fluorescence in situ hybridization and how can it be used in diagnostics?
13. How SNP arrays can be used in the diagnostics of inherited traits?
14. What kind of information can be learned from an SNP array analysis of a single patient?
15. What is the polimerase chain reaction and how can it be used for validating disease related SNPs?

16. Name three applications of next-generation sequencing in molecular research.
17. What are the most important differences between the genomic and transcriptomic applications of NGS methods?
18. How commercial genomic services can be used to explore the disease susceptibility of a patient?
19. How commercial genomic services can be used to explore the ancestry of a patient?
20. A patient brings you results from an SNP array based commercial diagnostics service. It says that the patient has two copies of rs1061170 which corresponds to the Y402H mutation in gene CFH. The report says that the likelihood ratio connected to this genotype is 1.64 related to age-related macular degeneration. Help the patient to interpret this information and give advice how to use this information to mitigate the effects of those variants.

## Final exam

Choose on from the following topics and write an essay of about 1000 words.

1. What is the role of cytogenetics methods in the research and diagnostics of inherited diseases? Describe scenarios where the following methods can be used: Karyotype analysis, SNP array.
2. Describe a hypothetical scenario where the genetic determination of an inherited trait is explored by modern cytogenetic methods. Explain the key steps from the identification of inheritance model to the corresponding SNPs, to the design of a diagnostic approach. Use the following technologies: exome-sequencing, SNP array, PCR.

## Final exam quiz

The following assignments are multiple choice questions where students should find the correct answer(s) among the offered alternatives. Correct choices are underlined here.

1. Which of the following is NOT a Mendelian inheritance model?
    1. Dominant
    2. Recessive
    3. X chromosome linked
    4. <u>Mitochondrial</u>
    5. Y chromosome linked

2. Which of the following mutations can NOT cause genetic changes manifesting in subsequent generations?
    1. Single nucleotide insertions
    2. Somatic mutations
    3. Chromosome arm deletions
    4. Mis-sense mutations
    5. Frameshift mutations
3. If an inherited disease is associated with a mutation where an adenine is replaced by a guanine, what term describes this phenomenon most exactly?
    1. It is an SNP
    2. It is a recessive allele
    3. It is a frameshift mutation
    4. It is a transition
    5. It is a disease causing mutation
4. Which of the following human traits and conditions shows simple, Mendelian inheritance model?
    1. Hair color
    2. Eye color
    3. High blood pressure
    4. Susceptibility for cancer
    5. None of the above
5. Which of the following are NOT Chromosomic variants?
    1. Copy number variants of short nucleotides
    2. Aneuploidy
    3. Trisomy 21
    4. Monosomy 22
    5. Robertsonian translocation
6. Which of the following methods are used for diagnosing chromosomal abnormalities?
    1. Fluorescent In-Situ Hybridization
    2. Real-Time PCR
    3. Blood type test
    4. Exome sequencing with NGS
    5. RNAseq

7. Which of the following method has the primary goal of identifying the sequence of a genomic region?
    1. CGH array
    2. Exome sequencing with NGS
    3. Sanger sequencing
    4. Two of the above
    5. Three of the above
8. Which of the following sequencing method is the most cost-efficient for producing genomic sequences of a patient?
    1. Massive parallel seq
    2. Sanger sequencing
    3. Sequencing by tiling arrays
    4. Sequencing using stop-nucleotides
    5. All these methods are the same in efficiency
9. Which of the following databases contain information which can be helpful for interpreting results from genomic analysis of a patient?
    1. OMIM
    2. dbSNP
    3. GWAS
    4. Two of the above
    5. All the three are good databases
10. Genomic services offer a wide variety of results available from their raw data. What kind of information can be accessed via one or more of these services?
    1. Ancestry information
    2. Estimated susceptibility for a panel of inherited diseases
    3. Y-DNA haplogroup
    4. Estimates for phenotypes of inherited traits
    5. All of these

# Module 4: Quality Improvement in Healthcare

This module programme will aim to equip trainees with a range of knowledge and skills, which are relevant and applicable in communications within healthcare contexts. Participants will learn how to build high-performing and engaged healthcare teams, establish and sustain effective clinical relationships, as well as implement strategies and tools to support patient-centered care. Additionally, with patient safety initiatives at the forefront of care, a major goal of this module will be to help health care professionals to

develop the background knowledge and skills necessary for the specialty of risk management.

## Self-assessment practices

1. Read the following article:

   Ormond, Kelly E et al. “Genetic counseling globally: Where are we now?.” American journal of medical genetics. Part C, Seminars in medical genetics vol. 178,1 (2018): 98-107. doi:10.1002/ajmg.c.31607 It is freely available on PubMedCentral: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5947883/

   Based on the aspects in Table 2 of this paper, research the status of genetic counselling in your country?

2. Compile a list of 10 questions you would ask about the training material if you would be a teacher checking the knowledge of your students. Remember, you are expected to ask questions at all levels of understanding the material: starting from simple 'remember the fact' questions up to the levels where students should be able to judge and compare critically the component of the covered material.

## Peer-assessment assignments

All these questions should be answered in no more than 350 characters.

1. There is a family with two sons, one of them is affected by an X-linked rare recessive condition, and and unaffected daughter. Explain, how to estimate the probability that the daughter is a carrier for the disease.
2. There is a family with two sons, who are affected by a Y-linked condition, and and unaffected daughter. Explain, how to estimate the probability that the daughter is a carrier for the disease.
3. There is a family pedigree with a condition which manifests in male members only. It is one particular version of vitilligo. Though vitilligo in general has a very complex genetic background, in this family the inheritance pattern is very clearly X-linked recessive. The maternal grandfather has this condition. There are four children in the family: two sons and two daughters. Both of the sons have the condition. Explain, how to estimate the probability that the sons of the two daughters will have vitilligo.
4. Studies found that in average, 17.8 out of 100 men of European ethnicity are estimated to develop Prostate Cancer between the ages of 35 and 79. At the same time, some studies suggest that in case of people with the genotype AG at

the SNP rs7127900 in average 19.9 people develop prostate cancer. The same average value for people with genotype AA at SNP rs4430796 is 22.9 out of 100 men. A patient has both AG at rs7127900 and AA at rs4430796. Explain how to calculate the odds ratio that he develops Prostate cancer.

5. Explain briefly the goal of polygenic risk score.
6. PRS tests are available for a number of conditions. How should a high risk score of a patient for coronary heart disease be interpreted?
7. What is the relationship between polygenic risk score and genomic population of a patient?
8. Discuss the differences between the association of genetic variants to complex disorders and causation of those diseases.
9. Can polygenic risk score test for susceptibility for venous thromboembolism (VTE) be considered as diagnostics for VTE?
10. What are the most important differences in genetic counseling targeting adult or minor patients and their families?

## Final exam

Choose on from the following topics and write an essay of about 1000 words.

1. Genetic counselling is a communication process in which aspects related to the risk of appearance of a genetic disease in a family are treated. Explain which aspects of a diagnosed disease should be explained to an affected patient in the framework of counselling. Demonstrate how these aspects contribute to the ultimate objective of the counselling.
2. A patient participated to a private genomic screening and has got a report regarding his susceptibility to prostate cancer. The test used 12 reported SNPs and identified the genotype of the patient for all of those markers. The final result from the screening was that the patient has an elevated susceptibility with an odds ratio of 1.56. The study also states that the heritability of these markers is estimated to be 42-57%. Draft a genetic counselling session (or sessions) for this patient covering all the important aspects of the case. What topics will be covered by what kind of professionals? The patient is anxious about how this is affecting him personally plus his future children. Address those aspects with special focus.

## Final exam quiz

The following assignments are multiple choice questions where students should find the correct answer(s) among the offered alternatives. Correct choices are underlined here.

1. What is the first step for predicting an individual's risk for a particular genetic disease?
    1. To collect information about the family pedigree and the genetic variants discovered.
    2. To differentiate between the model of only one gene producing the disease, and complex disorders with many risk factors including all genetic variants.
    3. <u>To define the pattern of inheritance of the disease, and establishing two main groups: Mendelian, and complex disorders.</u>
2. Which information can enhance the accuracy to predict genetic disease?
    1. Different types of Mendelian inheritance
    2. <u>Medical information</u>
    3. Bayesian analysis
    4. Genetic variants
    5. Environmental influences
3. What is a joint probability?
    1. The probability of having two children unaffected by a disorder.
    2. The probability of the conditional and posterior to occur together.
    3. The probability of being or not a carrier.
    4. The probability of two conditional probabilities to occur together.
    5. <u>None of these.</u>
4. How we can obtain the posterior probability?
    1. <u>By dividing the joint probability of an event by the sum of all the joint probabilities.</u>
    2. By dividing the joint probability of an event by the sum of all given genetic variants.
    3. By dividing the conditional probability of an event by the sum of all the joint probabilities.
    4. By dividing the joint probability of an event by the sum of all the conditional probabilities.
    5. None of these.
5. How can we calculate genetic risk for Mendelian disorders?

1. By considering any available information together with Mendelian, and complex disorders.
2. By considering the family pedigree together with the Mendelian inheritance laws.
3. By considering all genetic variants together with the pattern of inheritance.
4. By considering any available information together with the Mendelian inheritance laws.
5. By considering all genetic variants together with the Mendelian inheritance laws.

6. Consider this statement: “Genetic information can be extended to other members of the family, therefore the diagnosis of a genetic disease does not only affects the patient but also their relatives.” Which one is true from the following options?
    1. Genetic information cannot be extended to other members of the family, therefore the statement is false.
    2. While genetic information indeed can be extended to other members of the family, still, the diagnosis of a genetic disease affects only the patient never their relatives; therefore the statement is false.
    3. The statement is true as it stands.
7. Why is genetic information by nature different from the rest of medical information?
    1. Because genetic information can be extended to relatives.
    2. Because genetic conditions are for life.
    3. Because genetic tests sometimes allow us to make pre-symptomatic diagnoses of pathologies.
    4. Because genetics often involves abstract and complicated concepts that make it difficult for the patient and their family members to understand the situation.
    5. All of these.
8. Here is a list of expertise in relationship with inheritable genetic problems. Which of these are NOT part of general genetic counselling?
    1. Medical aspects related to diagnosis, prognosis, and follow-up.
    2. Genetic aspects that allow understanding risk of occurrence or recurrence.
    3. Options to modify the aforementioned risk.
    4. Options to modify the problematic genes.
    5. Psychological support during the process.

9. In the context of any genetic test that is carried out, consent must be obtained from the patient. Which should be (among others) always part of the informed consent? (Choose only one.)
    1. The limitations of the test that will be carried out.
    2. The possible studies that could be derived from the initial test.
    3. The exact methodology how the test will be executed and analyzed.
    4. Two of these.
    5. Three of these.
10. Which one is true for genetic testing in minors? (You can choose more than one.)
    1. Genetic tests are considered appropriate when these studies aim to diagnose a pathology that the patient already has.
    2. Carrier screening tests of diseases with adult onset for which there is no effective treatment are considered always unethical.
    3. Presymptomatic tests are ethical in minors given that their parents provided informed consent on their behalf.
    4. Professionals focused on genetics and ethics try to respect the autonomy of the child when there is no medical benefit of performing the test by offering testing to their parents.
    5. Professionals focused on genetics and ethics try to respect the autonomy of the child when there is no medical benefit of performing the test by not suggesting carrier screening tests in minors.