BioS

DIGITAL SKILLS ON COMPUTATIONAL BIOLOGY FOR HEALTH PROFESSIONALS

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

R3.3. BioS trainer's handbook

WP3: Development of the BioS Curricula Syllabi and Educational Material (OER)

Responsible Partner: P5 BIB

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

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

Authoring Partner(s):	BIB

BioS
DIGITAL SKILLS ON COMPUTATIONAL BIOLOGY FOR HEALTH PROFESSIONALS

Table of contents

BioS
DIGITAL SKILLS ON COMPUTATIONAL BIOLOGY FOR HEALTH PROFESSIONALS

1. Introduction to BioS Courses

The BioS courses aim to provide physicians with knowledge, skills and competencies in the field of bioinformatics that will enable them to improve their clinical practice. Therefore, they include both the basic concepts of bioinformatics and computational biology, as well as the most recent advances in this field that can have a clear application for physicians and transform their practice. The courses have a modular and flexible structure. Four modules are offered, which can be taken sequentially or not, depending on the level of knowledge in bioinformatics of the participants. However, it is advisable to take module 1 (introduction to bioinformatics) to make the best use of the following three modules. The courses are based on an active and contextualised learning model, in which students are the centre of learning and learn through diverse authentic activities, close to their professional reality.

BioS courses have been designed so that their implementation does not require the figure of an expert instructor, but it is recommended to have the figure of a facilitator to achieve a successful implementation.

The teaching materials have been developed with a scientific team with expertise in the field of bioinformatics, the advice of an advisory committee formed with different stakeholders (different profiles of medical doctors, bioinformaticians, researchers and e-learning experts) and technical and pedagogical support.

Table 1. The teaching team of BioS modules

Modules	Coordinador	Teachers
Module 1	Dr.Cedric Notredam	Dr.Cedric Notredam Group leader of Notredam Lab- Comparative genomics at Centre for Genomic Regulation-CRG in Barcelona, Spain.
Module 2	Dr. Hafid Laayouni	Dr. Hafid Laayouni Associate teacher and researcher at Universitat Pompeu Fabra. Coordinator of the Bachelor degree of Bioinformatics in Barcelona, Spain. Dra. Laura Serra Associate teacher od Statistics at Universitat de Girona (UdG) in Girona, Spain.

		Dr. Oscar Lao Group leader of Population Genomics research at the CNAG-CRG in Barcelona, Spain.
Module 3	Dr. Ferran Casals	Dr. Ferran Casals Head of the Genomics Core Facility at Universitat Pompeu Fabra, associate teacher and researcher at the same University in Barcelona, Spain. Dr. Gerard Muntaner Postdoctoral scientist at the Universitat Rovira i Virgili (URV) in Tarragona, Spain Dra. Clara Serra Genetic counselor at the Vall d'Hebron University Hospital in Barcelona, Spain. Dra. Ivon Cusco Researcher at the genètic diagnosis laboratory at the Vall d'Hebron University Hospital in Barcelona, Spain.
Module 4	Dra. Clara Serra	Dr. Ferran Casals Dr. Gerard Muntaner Dra. Clara Serra Dra. Ivon Cusco

2. Description of the modules

The BioS curriculum is composed of the following modules: 1) Introduction to bioinformatics, 2) Computational statistics for clinical doctors, 3) Personalized genomics in patient care, and 4) Quality improvement in medical care.

The key learning outcomes of the 4 modules and the dedication of workload expected for each module is described below. The knowledge, skills and competences to be developed in each module are specified in Appendix 1.

Table 2. Modules Learning outcomes and workload

Modules	Learning outcomes (LO)	Weeks/ Workload
1. Introduction to bioinformatics	- Retrieve information and data regarding specific genes and proteins which could be chosen as candidate genes for a disease, e.g. functional information and sequence variant information - Perform analyses and comparisons to identify essential and non-essential parts in a gene or a protein, e.g. multiple sequence alignments using publicly available, web-based tools - Understand how applying such skills can lead to quick and cost-effective diagnoses of diseases and disorders with a genetic component	4 weeks 35-40h
2. Computational statistics for clinical doctors	- Understanding of descriptive statistics, elements of probability, hypothesis testing, nonparametric methods, correlation analysis, and linear regression - Understanding of how to choose appropriate statistical tests and how to assess statistical significance - Understanding of how to visualize data and carry out statistical testing - Understanding of R, a powerful programming language for statistical computing and graphics - Understanding selected kinds of analyses of biomedical data that a professional can carry out easily using e.g. the package of R for the analysis of sequencing data from a patient	4 weeks 35-40h

3. Personalized genomics in patient care	- Explain single nucleotide polymorphisms (SNPs) and different types of SNPs - Describe some example roles of SNP's in disease development - Describe the basic principles of variant effect prediction for genetic variants in protein-coding genes and in micro-RNA genes - Use reports in patient care from commercial personal genomics services - Interpret SNP-related increased and decreased risks in selected diseases - Interpret commercial reports and translate them to actions in appropriate health care segments	4 weeks 35-40h
4. Quality improvement in medical care	- More efficient communication of disease risks related to genetic testing	2 weeks 5-10h

The modules are structured by weeks and contain the following activities:

- **Interviews with experts**: each module is introduced with an interview with a renowned researcher in the field. The objective of the interview is to provide an overview of the state of the art of the topic, its relevance and applicability for the health professionals who take it. These are stimulating interviews that aim to motivate participants to get involved in the course activities.

- **Video lessons**: every week there are two lessons taught by of expert teachers, where they explain in a synthetic way the basic concepts that will be worked during the week. They last about 10 minutes. It is important to watch the videos and understand them in order to be able to do the rest of the activities proposed for the week.

- **Short problems**: every week there is 1 short problem that has the purpose of applying the basic concepts that have been taught in the videos and consolidate them.

- **Learning activity**: every week there is a practical case in which the student must put into practice the knowledge acquired and expand it.

- **Readings of articles or reference materials**: each week there are proposed readings to go deeper into the knowledge of the topic. In some cases, it is

necessary to read the documents in order to carry out the practical cases. When this is the case, it is already specified in the case studies.

- **Quizzes**: each week there is a self-correcting quiz of 10 questions so that the student can check whether he or she has acquired the basic knowledge of the week.

- **Forum**: in each module there is a forum activity to encourage discussion between the different course participants.

3. The training materials

3.1 Module 1: Introduction to bioinformatics

This module aims to introduce the basic knowledge of how molecular data available in modern biomedicine can be used to promote genomic/personalised medicine. The course introduces the student to understanding, managing and analysing different types of data, from genomes, to DNA sequences and proteins or protein structures. The main biological databases are introduced and basic methods for their analysis are taught. The detailed learning outcomes for this module can be found in Appendix 1.

The module is introduced through an interview with Roderic Guigó, a reputable researcher in the field of bioinformatics. He is the group leader of Computational Biology of RNA Processing research group in Centre for Genomic Regulation (CRG) in Barcelona, Spain.

The course offers a variety of activities to achieve the different types of learning outcomes: knowledge acquisition, knowledge application and competence.

The basic knowledge of the module can be obtained through the video classes and can be self-assessed through the weekly quiz. For the students who want to go deeper into the contents, some complementary readings are proposed. The following table shows the contents worked on in the video classes for each week.

Table 3. Video lectures for Module 1. Introduction to bioinformatics

Week	Topic	Video	Title	Teacher
1	Genes within Genomes	111	Overview of bioinformatics with respect to molecular biology and genetics concepts	Cedric Notredam
		112	Navigating genomes	Cedric Notredam
2	Searching Databases for Functional Information	121	Gathering information in medical genetic databases	Cedric Notredam
		122	Gathering information in Generic Bioinformatics databases	Cedric Notredam
3	Exploring RNA and Protein products of a gene	131	Finding out about the level of expression of a gene	Cedric Notredam
		132	Finding out about the protein	Cedric Notredam
4	Healthy and Unhealthy proteins	141	Homology searches and multiple alignments in Biology	Cedric Notredam
		142	Metabolic pathways and integrated cellular function	Cedric Notredam

In order for the student to acquire the skills and competences of the module, he/she will need to put into practice the contents worked on in the video classes. This is why different types of activities are proposed, from the simplest to the most complex.

The key technical learning outcome “Retrieve information and data regarding specific genes and proteins which could be chosen as candidate genes for a disease” can be practiced by the activities of week 2 where user will learn how to use OMIM for information gathering and the activities of week 3 where user gathers data about the domains of BRCA1 protein.

The key technical learning outcome “Perform analyses and comparisons to identify essential and non-essential parts in a gene or a protein, e.g. multiple sequence alignments using publicly available, web-based tools” can be practiced in the

activities of week 4 where users use BLST to identify BRCA1 homologous and estimate the level of conservation of different homologous.

To achieve the learning outcome "Understand how applying such skills can lead to quick and cost-effective diagnoses of diseases and disorders with a genetic component", users can perform the activities of week 1 where users have to propose strategies for data analysis and discuss with the peers in the forum activity.

The complete learning activities with the solutions and the recommended readings are in appendix 2.

3.2. Module 2: Computational Statistics for clinical doctors

The aim of this module is to provide a practical introduction to the analysis of Big data biomedical field, in order to acquire a critical understanding of the reliability of analysis results. In this module, medical doctors will be able to understand and become familiar with the statistical environment of R and apply it to the analysis of biological data in an efficient manner. The detailed learning outcomes for this module can be found in Appendix 1.

The module is introduced through an interview with researcher Antonio Barbadilla, who is the research leader of the Bioinformatics of Genomic Diversity group in the Institute of Biotechnology and Biomedicine at the University Autònoma of Barcelona (Spain).

The course offers a variety of activities to achieve the different types of learning outcomes: knowledge acquisition, knowledge application and competence development.

The basic knowledge of the module can be obtained through the video classes and can be self-assessed through the weekly quiz. For the students who want to go deeper into the contents, some complementary readings are proposed. The following table shows the contents worked on in the video classes for each week.

Table 4. Video lectures for Module 2. Computational Statistics

Week	Topic	Video	Title	Teacher
1	Descriptive statistics	211	Overview of type of variables (qualitative and quantitative) and how to deal with them	Laura Serra
		212	Introduction to R	Laura Serra
2	Statistical inference	221	Overview of statistical significance	Oscar Lao
		222	Hypothesis testing in medical practice	Oscar Lao
3	Linear regression	231	Introduction to linear regression	Laura Serra
		232	Linear regression models using R	Laura Serra
4	Categorical Data	241	Introduction to Categorical Data analysis	Hafid Laayouni
		242	Summary on the statistical inference interpretation	Hafid Laayouni

In order for the student to acquire the competences of the module, he/she will need to put into practice the contents worked on in the video-lessons. This is why different types of activities are proposed, from the simplest to the most complex.

In the learning activities of week 1, users can apply basic concepts of descriptive statistics to analyse different type of medical data, such as those from a flu outbreak or from patients with a gastric cancer. In the learning activities of week 2, users can practice how to design good experiments for medical research, how to choose the most appropriate statistical tests and how to assess statistical significance. Through the learning activities of week 3, users can learn how to use R, a powerful programming language for statistical computing and graphics and they can put in practice by using R for different kinds of analyses of biomedical data. In the learning activities of week 4, students will be able to perform appropriate statistical tests to determine the relationship between a mutation and the development of a disease and learn how to analyse their results.

The complete learning activities with the solutions and the recommended readings are in appendix 2.

3.3. 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 genetic analyses and commercial personalized genomics services, like 23andMe, deCODE, Gene by Gene, etc. This module facilitates integrating these services into their patient care activities. The detailed learning outcomes for this module can be found in Appendix 1.

The module is introduced through an interview with researcher Nuria Lopez-Bigas, which is the group leader of Biomedical Genomics Research Group in Institute for Research in Biomedicine in Barcelona (Spain).

The course offers a variety of activities to achieve the different types of learning outcomes: knowledge acquisition, knowledge application and competence development.

The basic knowledge of the module can be obtained through the video classes and can be self-assessed through the weekly quiz. For the students who want to go deeper into the contents, some complementary readings are proposed. The following table shows the contents worked on in the video classes for each week, as well as the lecturers who teach them.

Table 5. Video lectures of Module 3. Personalized genomics in patient care

Week	Topic	Video	Title	Teacher
1	Inheritance Model of Diseases	311	Analysis of pedigrees. Mendelian models	Clara Serra
		312	Other models of inheritance	Ivon Cusco
2	Human Genetics sources of variability	321	Genetic variation	Ferran Casals
		322	Population variation or Disease causing variants	Gerard Muntaner
3	Diagnostic tools: How to select the	331	Cytogenetics and molecular cytogenetics	Clara Serra

	correct test. Cytogenetics	332	Cytogenetics and molecular cytogenetics. Data analysis and interpretation	Ivon Cusco
4	Diagnostic tools: How to select the correct test. Sequencing	341	Nucleotide variants	Ferran Casals
		342	Interpretation of genomic analysis	Gerard Muntaner

In order for the student to acquire the competences of the module, he/she will need to put into practice the contents worked on in the video-lessons. This is why different types of activities are proposed, from the simplest to the most complex.

The key technical learning outcome “Understanding the nature and role of single nucleotide polymorphisms (SNPs) and other genetic variants” can be practiced by the activities of week 1, to put in context the different models of inheritance in genetic disease, and week 2 where user will learn to classify the different types of genetic variants according to different criteria, as well as understand the different effect that they might have in the protein and their possible relationship with the disease. Week 2 also covers the learning outcomes “Explain single nucleotide polymorphisms (SNPs) and different types of SNPs”, “Describe some example roles of SNP's in disease development” and “Describe the basic principles of variant effect prediction for genetic variants in protein-coding genes and in micro-RNA genes”

The key technical learning outcome “Using variant analyses” can be practiced with the activities of week 3, where the different technologies to generate genetic data are presented, and week 4 where the users will use real data from genetic analyses to annotate the genetic variants and use molecular and biological knowledge to identify the mutations more prone to be related with genetic disease. These practical exercises are also related to the learning outcomes “Use reports in patient care from commercial personal genomics services”, “Interpret SNP-related increased and decreased risks in selected diseases” and “Interpret commercial reports and translate them to actions in appropriate health care segments”.

The complete learning activities with the solutions and the recommended readings are in appendix 2.

3.4. 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. This module is focused especially to the communication, ethics and risks associated with genetic testing and disease risk assessment.

The module is introduced through an interview with the researcher Jordi Surrelles, group leader of Genomic Instability and DNA Repair Research Group in Universitat Autònoma de Barcelona (Spain).

The basic knowledge of the module can be obtained through the video classes and can be self-assessed through the weekly quiz. For the students who want to go deeper into the contents, some complementary readings about the risk estimation for genetic disorders, personalized genetic tests and ethics in genetic counseling are recommended. The following table shows the contents worked on in the video classes for each week, as well as the lecturers who teach them.

Table 6. Video lectures of Module 4. Quality improvement in Healtcare

Week	Topic	Video	Title	Teacher
1	Risk assessment	411	Genetic risk assessment. Bayesian analysis	Ferran Casals
		412	Genetic risk prediction in complex disorders	Gerard Muntaner
2	Ethical and Communication skills	421	Ethical and Communication skills	Clara Serra

The key technical learning outcome "More efficient communication of disease risks related to genetic testing" is practiced with the activities of weeks 1 and 2.

In week 1, the participants will become familiar with the concept of genetic risk, and the methods to calculate it according to the type of genetic disease. In week 2, the participants will discuss on the how to decide the optimal genetic approach for each situation, as well as how to deal with very important aspects as data protection, confidentiality, informed consents, or results communication to patient.

The complete learning activities with the solutions and the recommended readings are in appendix 2.

3. Recommendations for the implementation and dynamization of BioS Courses

Although the BioS materials have been designed so that they can be used without the need for an instructor, we believe that their implementation can be much more successful if they are implemented with an instructor. Ideally, instructors should be experts in the subject matter to provide more in-depth feedback on the course content, but this is not essential.

Below are some general recommendations for an effective course implementation and dynamization:

- **Meet the group of students**. It is important that the instructor knows the learners and that they get to know each other so that they can feel that they are part of a learning community, so it is recommended to start with a round of presentations, in which they can be asked to explain their professional experience and what are their expectations of the course. The presentations can be through a forum with written messages or they can hang a mini video. This will help create a climate of trust that will encourage involvement in the course.

- **Take advantage of the diversity of the group**. Be aware that the group of students can be very heterogeneous in terms of their previous knowledge of bioinformatics, their previous professional experiences, their motivations and

interests for taking the course, their availability to complete it, their cultural background, etc. Therefore, it is very important to make clear from the beginning of the course that this is a modular and flexible course in which each student can shape his or her own itinerary. For those who do not have any knowledge of bioinformatics, it is recommended that they start with module 1, as it provides an overview of bioinformatics, as well as the basic knowledge to follow the other three modules. The other modules can be taken independently. In each module, there is the possibility to supplement the knowledge through recommended readings. Students should be guided to find what is most useful for their professional development.

On the other hand, it is suggested to take advantage of the group diversity as a source of knowledge, promoting peer learning, exchanging of knowledge, experiences and opinions. We believe that for physicians who participate in the course, to discuss with other colleagues the opportunities and the new challenges that bioinformatics poses for clinical practice can be very interesting.

- **Provide different pathways to learners**. Students can take the course for different reasons and they can adapt it to their needs and preferences. Students who aim for levels of recall and understanding of the course content should focus on the videos and quizzes. Those students who are interested in the level of application and analysis should do the short problems and learning activities. Those, who aim for even higher levels, should also use the reading material and participate in the forum discussion activities.

- **Engage the students**. It is crucial to encourage students to get involved in the course and to carry out the activities as this is where they will learn the most. Therefore, it is recommended to start each week with a message encouraging participation in the different activities of the week, explaining the objectives of each activity and highlighting how they can be useful for their learning and professional development. As you will have seen from the materials, each module has a discussion activity (forum) that aims to involve students from a more personal point of view, this can be present throughout the module and can serve as a hook with the module topics.

- **Feedback is crucial for learning**. The most important role of the trainer is to follow the course and give feedback on the students' development. If students feel that someone is supervising their work and receive feedback, their motivation and

involvement will increase. If the group is very large and it is not possible to comment on the work of all students, you can make a general comment on the group's contributions and take one or two as concrete examples. As many of the activities are self-evaluating, you can analyse the group's activity of the week (viewing the videos, carrying out the activities, results of the quizzes and contributions in the forum) and at the end of the week, make a comment on the activity that has taken place.

- **Social learning is more effective**. That is why it is important to encourage collaborative learning. For this purpose, the materials include both forum and peer-assessment activities. In addition, learning activities can easily be transformed into collaborative activities, if the group of learners is appropriate. It should be noted that for collaborative learning activities to work well in an online environment, requires a high commitment of students to course follow-up and a very low dropout rate, otherwise it can generate frustration to students who want to get involved and do the group work.

Appendix 1. Learning outcomes for the four modules

Table 1: Module 1: Learning Outcomes of Introduction to Bioinformatics

5	INTRODUCTION TO BIOINFORMATICS	40	4
KEY TECHNICAL OUTCOME	**KNOWLEDGE**	**SKILLS**	**COMPETENCE**
1. Retrieve information and data regarding specific genes and proteins which could be chosen as candidate genes for a disease.	The Learner knows and understands: ● The relevance of biological sequences for health and diseases ● Basic terms related to sequence handling ● Medical relevance of sequence annotation	The Learner is able to: ● Gather information on selected genes and proteins using tools such as BLAST, UniProt, and PDB ● Compare biological sequences through multiple sequence alignment ● Identify the active site of an HIV Serine protease (or of any other structurally characterized enzyme)	The Learner: ● Is aware of the power of modern high-throughput sequencing methods and applies sequencing data to improve diagnostics of diseases with a suspected genetic component ● Uses up-to-date knowledge from databases of genes and proteins to support their estimates of the significance of genes suggested as contributors in a genetic disease ● When needed, considers protein structure in the
	PERFORMANCE CRITERIA		
	● Uses sequence and annotation files to access up-to-date, in-depth knowledge of medically relevant genes (with the help of the demonstrated Internet databases and tools) ● Uses the discovered information to make more informed clinical decisions based on genetic variant data		
KEY TECHNICAL OUTCOME	**KNOWLEDGE**	**SKILLS**	**COMPETENCE**
2. Visualize genomic features and perform simple analysis on them using Internet based tools.	The Learner knows and understands: ● List of available databases and other Internet resources ● Functional Genome Annotation and Metabolic Pathways	The Learner is able to: ● Visualize genomic data in a genome browser ● Find information of known genome variants associated to desired genes ● Find and visualize functional genome annotations and metabolic pathway information	The Learner: ● Uses genomic tools routinely to get a quick, integrated view of data related to any gene as necessary ● Gains a knowledge-based, data driven view of every new disease-related gene they encounter
	PERFORMANCE CRITERIA		
	● Uses the demonstrated tools to visualize annotations of their candidate gene lists; ● Uses Internet resources including Genome Browsers and KEGG; ● Interprets gene information critically in the light of up-to-date information.		
OUTPUTS			
• **More confidence in data from modern-day high-throughput sequencing;** • **Making more use of such data in clinical decision-making;** • **More informed interpretation of personal genome data;** • **Quicker and more relevant diagnoses of diseases with an assumed genetic component.**			

Table 2: Module 2: Learning Outcomes of Computational Statistics for Clinical Doctors

EQF LEVEL	UNIT OF LEARNING OUTCOME	HOURS	CREDITS
5	COMPUTATIONAL STATISTICS FOR CLINICAL DOCTORS	40	4

KEY TECHNICAL OUTCOME	KNOWLEDGE	SKILLS	COMPETENCE
1. Understanding key elements of modern statistical analysis.	The Learner knows and understands: ● Descriptive statistics, elements of probability, hypothesis testing, nonparametric methods, correlation analysis, and linear regression ● Elements of statistical reports ● Elements of the visualization of statistical data	The Learner is able to: ● Choose appropriate statistical tests ● Perform simple statistical analyses using software tools ● Assess statistical significance ● Evaluate if appropriate statistical test is used in an analysis ● Interpret statistical graphs	The Learner: ● Pays attention to statistical values associated with reports of sequencing data and its analysis reports; ● Weights their decision-making in the light of the statistical uncertainty of each finding.
	PERFORMANCE CRITERIA		
	● Operates with elements of statistical analysis to interpret analysis results; ● Reads analysis reports and research articles with a statistically critical approach.		

KEY TECHNICAL OUTCOME	KNOWLEDGE	SKILLS	COMPETENCE
2. Using free software for statistical analysis of data from case studies.	The Learner knows and understands: ● The role of R software in statistical analysis ● CRAN and Bioconductor	The Learner is able to: ● Install R, RStudio, and selected related statistical software packages and use them ● Organize and perform a data analysis project of biomedical data ● Create and handle graphs from an analysis	The Learner: ● Is aware of the power and limitations of statistical analyses of biomedical data, especially of high-throughput sequencing data ● Requests custom-made analyses from professionals, knowing that efficient, simple, and free tools exist enabling almost any analysis they can think of, in an easy and cost-efficient way
	PERFORMANCE CRITERIA		
	● Makes more use of statistical parameters in decision-making, especially related to analysis of sequencing data from a patient ● Designs ideas for custom analyses if needed in complex situations		

OUTPUTS
• **More efficient use of biomedical research results;** • **Incorporating sequencing data and the use of associated statistical data more efficiently into diagnostic practice.**

Table 3: Module 3: Learning Outcomes of Commercial personalized genomics services in patient care

EQF LEVEL	UNIT OF LEARNING OUTCOME	HOURS	CREDITS
5	COMMERCIAL PERSONALIZED GENOMICS SERVICES IN PATIENT CARE	40	4

KEY TECHNICAL OUTCOME	KNOWLEDGE	SKILLS	COMPETENCE
1. Understanding the nature and role of single nucleotide polymorphisms (SNPs) and other genetic variants.	The Learner knows and understands: ● Different levels of genetic and genomic variants; ● Variant terminology; ● Theoretical aspects of human genetics related to genomic variations; ● Variant analyses within populations; ● Diagnostic tools used for variant detection and analysis.	The Learner is able to: ● Access and study genetic variant data from on-line databases; ● Read reports of patient sequence variant analyses.	The Learner: ● Considers carefully the contribution of genetic variants for a given patient case
	PERFORMANCE CRITERIA		
	● Uses this learning for more informed reading of reports from genomic sequencing services ● Uses variant databases for accessing more information on disease gene candidates		

KEY TECHNICAL OUTCOME	KNOWLEDGE	SKILLS	COMPETENCE
2. Using variant analyses	The Learner knows and understands: ● Types of available commercial personalized genomics services and other genetic tests ● Differences between their analysis methodology ● Principles of variant effect predictions	The Learner is able to: ● Select appropriate genetic tests for a given clinical situation ● Interpret reports from different genomics services ● Assess the reliability of information sources used in different genomic services	The Learner: ● Evaluates always if genetic variant analyses are needed and appropriate; ● Uses personal genomic reports to gain a better understanding of patient's health status; ● Applies available research knowledge for personalizing patient treatment and/or preventive measures.
	PERFORMANCE CRITERIA		
	● Uses different genomics services to support clinical work when a genetic component is assumed ● Assesses individual patient status taking into account information from genomic sequencing reports		

OUTPUTS
• **Integration of the available commercial personalized genomics services into patient care practice**
• **More timely and more cost-efficient clinical decisions**
• **Better choices in treatments and in disease prevention**
• **Improved health**

Table 4: Module 4: Learning Outcomes of Quality Improvement in Healthcare

EQF LEVEL	UNIT OF LEARNING OUTCOME	HOURS	CREDITS
5	QUALITY IMPROVEMENT IN HEALTHCARE	10	2

KEY TECHNICAL OUTCOME	KNOWLEDGE	SKILLS	COMPETENCE
More efficient communication of disease risks related to genetic testing	The Learner knows and understands: ● Concepts related to disease risk assessment ● Ethical issues related to genetic testing ● Tools to support communication	The Learner is able to: ● Communicate risk information to patients in understandable and compassionate ways ● Help health professionals make informed decisions of treatment or non-treatment of discovered diseases ● Give genetic counselling	The Learner: ● Identifies and addresses the differences in patients' values, preferences and expressed needs ● Aims at a coaching culture that supports consistent exceptional care and service
	PERFORMANCE CRITERIA		
	● Implements strategies and tools to support patient-centred care; ● Improves the patient experience by better communication.		

OUTPUTS

- **Attracting and engaging customer-focused employees who are passionate about providing the best and most compassionate, yet efficient, care to the patient**
- **Effective clinical relationships**
- **Patient experience framework that better meets and exceeds the patient's need**

Appendix 2. Learning activities

MODULE 1. INTRODUCTION TO BIOINFORMATICS

Learning activities for Week 1. Genes within Genomes

Short problem

Propose an integrated strategy to figure out the level of differential expression of proteins between two cell lines whose genome is available. You have to analyse as many datasets as possible with your funds to acquire experimental data.

1- Which type of data do you need to acquire
2- Propose a generic pipeline

Solution

1. In theory you should prefer proteomics data, but in practice RNAseq data is much easier to obtain and it has been shown to correlate well with the protein abundance. You should therefore do an RNASeq analysis

2. There exists a wide range of ways of dealing with RNASeq data, but in practice, most pipelines include the following steps:

- Mapping of the reads onto the reference genome
- Transcript quantification
- Quantification of alternatively spliced isoform

A complete overview is available in this recent publication:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8?optIn=false

Learning activity

You want to process the transcriptome of 100 patients. Assuming you have a perfect rate of ribosomal RNA removal, propose and budget a strategy to sequence, store and analyse your data.

Solution

The main purpose of this exercise is to figure out how much sequencing is needed considering that the genes whose differential expression you are interested in have levels of expression that varies around the median of the whole transcriptome expression level. Try to work out how much sequencing you may need using the suggested reading, then use the Amazon services to work out the cost of storage and computation. Discuss the potential trade-off between speed and cost.

Note that there is no real canonical depth when sequencing transcriptomes. The depth depends on relative level of expression of the genes you are interested in. The interest of this exercise is

therefore not to find the *right* numbers, but rather to find numbers that are well justified and to associate a cost. For instance,

Sample Solution:

For instance, in the Encode project (https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972/@@download/attachment/ENCODE%20Best%20Practices%20for%20RNA_v2.pdf)

The recommended number of reads is 30 million paired-end/experiments with total read length typically 150 nucleotides long.

Looking online, we can see (https://www.abmgood.com/RNA-Sequencing-Service.html) that the current cost for 20M paired-end sequencing is $645/20 M reads + $145 for rRNA depletion + $50 total RNA isolation for NGS

This makes a rough total of: $1162/sample and therefore about $116,250 for your whole analysis

Now let us see the storage needs/cost: each experiment should generate 30M*150=4,500 Millions of nucleotides to be stored. It is commonly accepted that without compression the storage requires about 100 Megabytes/million nucleotides, which makes roughly 450 Gb/experiment and therefore about 45 Tb of storage.

Assuming we are budgeting to keep the data for 10 years, the cost on Amazon would be: (https://aws.amazon.com/s3/pricing/?nc1=h_ls), $0.023/Gb/month, therefore: $124,242.

Of course the true cost of storage will be significantly smaller. If you consider that storage cost drops by half every 14 months (Moore's law), the real cost should be about half.

Quiz

1. The assembled sequence of a genome is
    a. a prediction
    b. an experimental model
    c. a hypothesis
    d. a validation
    e. all the four above

2. Bioinformatics methods can be used
    a. only online on remote servers
    b. often both online remotely or on your own computer
    c. only locally on your own computer
    d. only in super-computers
    e. only on smart-phones
3. Given two protein sequences, sequence comparison sometimes makes it possible to:
    a. infer functional similarity
    b. infer common evolutionary origin
    c. infer structural similarity

d. a, b and c
e. a and b

4. The largest Human Chromosome is Chromosome #1. It contains
a. about 250 million base pairs
b. about 25 million base pairs
c. About one billion base pairs
d. about one million base pairs
e. about 10 billion base pairs

5. The significance of a measurement carried out using bioinformatics depends:
a. on its absolute value
b. on the difference with the average
c. the possibility to reproduce it on the same dataset
d. on the distribution of comparable observations
e. none of the above

6. The number of base pairs in a human genome is
a. 60 billion nucleotides
b. twice 3 billion nucleotide base-pairs
c. 100 million nucleotides
d. 10 million nucleotides
e. It can between 10 million and 100 millions

7. The most convenient technique to decode the epi-genome of a cell is:
a. ChIP-Seq sequencing
b. RNAseq
c. DNASeq
d. Proteomics
e. all of the above

answer: ChIP-Seq is the most widely used technique, provided the right antibodies are available.

8. In the human genome, the majority of the genes are
a. protein coding
b. non-coding RNA genes
c. of unknown nature
d. equally split among coding and non-coding
e. pseudo-genes

answer: It appears that the majority of genes are non-coding though it remains unknown if these numerous non-coding genes are functionally as important as their protein coding counterparts.

9. The most efficient method to predict the function of a protein is
a. functional genomics
b. proteomics

c. RNASeq
d. sequence comparison
e. crystallography

10. If your protein of interest has no homologue with a know structure, what would be the most efficient way to determine its structure

a. X-Ray crystallography
b. Nuclear Magnetic Resonance
c. Cryo-Electro Microscopy
d. Molecular Modelling
e. a, b and c

Learning activities for Week 2. Searching Databases for Functional Information

Short problem

Use online resources such as the Genetics Home Reference and OMIM to identify all the genes whose inherited mutations have been shown to be associated with a higher risk of breast cancer.

Solution

1- Open the Genetic Reference Home page : https://ghr.nlm.nih.gov/

Type "Breast Cancer" in the search window. In the filter box click on "Genes". The return page will list all the genes whose mutations have been shown to be associated with an increase in Breast cancer.

https://ghr.nlm.nih.gov/search?query=Breast+Cancer&tab=gene&start=460&rows=10

As you can see there are 469 such genes. All of them are not equally prevalent and in the list the genes are returned starting with those most commonly associated with breast cancer. Note also that in this return list you will find Mendelian mutations that are inherited and Somatic mutations that are acquired during life and cannot be passed to the progeny.

3- Open the OMIM https://omim.org/

Type "Breast Cancer". OMIM will return a list of all the entries featuring all the entries containing this term.

https://omim.org/search/?index=entry&start=1&limit=10&sort=score+desc%2C+prefix_sort+desc&search=breast+cancer

If you click on the Gene Map Table button, near the top right, you will then display a table containing all the genes that have been shown associated with breast cancer. This table provides much more complete information tan the Genetics Reference Home, but the information is much less prioritized (i.e. genes are classified by chromosome rather than frequency of cancer inducing mutations).

Learning activity

Identify SNPs associated with BRCA1 that are likely to increase the risk of cancer. Use OMIM, dbVar and dbSNP

Solution

1- Go into OMIM and search for BRCA1.

This search will return a list of entry featuring BRCA1.

https://omim.org/search/?index=entry&start=1&limit=10&sort=chromosome_number+asc%2C+chromosome_sort+asc&search=BRCA1

Select the first 1

https://omim.org/entry/113705?search=BRCA1&highlight=brca1

This first entry is a high-quality summary of all known information on BRCA1. The part of interest with respect to variability is to be found in the “allelic-Variant” section. You can access it directly by using the summary on the left. This lists all the description of alleles associated with increased breast cancer frequency. Note that these 44 do not constitute an exhaustive catalogue. They are merely OMIM selected representative mutations.

2- Go into dbSNP. https://www.ncbi.nlm.nih.gov/snp/

dbSNP contains all the mutations involving one nucleotide only. These mutations can be silent when they do not change the amino acid, or they can also be non-synonymous when they change the amino acid.

Type BRCA1. The page will return a list of the 24,509 mutations involving BRCA1.

https://www.ncbi.nlm.nih.gov/snp/?term=BRCA1

Note that most of these mutations are expected to be benign. In order to figure out the ones that could be pathogenic, you can refine your search and use the following query

(BRCA1) AND pathogenic[Clinical significance]

In which case the return page:

https://www.ncbi.nlm.nih.gov/snp/?term=(BRCA1)+AND+pathogenic%5BClinical+significance%5D

Features about ten time less entries, all of which are associated with a likely pathological effect.

3- Note that SNPs are not the only mutations affecting our genomes. Structural variants that include copy number variation, translocations, inversions and deletions are also very common and are now considered an important source of genetic variation. In order to explore this aspect of BRCA1, we will use a different resource named dbVAR

Go into dbVAR: https://www.ncbi.nlm.nih.gov/dbvar/

Type BRCA1

The search will return a total of 7194 variants. Note that many of these variants are listed as "Uncertain Significance", which means that they have not been seen associated with a pathology.

If we want to make a more refined search, we can type in the search window:

(BRCA1) AND pathogenic[Clincal Interpretation]

This will return a list of all the 506 BRCA1 variants reported associated with a pathogeny. Contrarily to OMIM, this list is as exhaustive as possible given the data available to the NCBI at the time you make the search.

In summary, we have seen here that all resources do not provide the same amount of information and that in order to be exhaustive, one should navigate across databases.

Quiz

1. The largest databases describing the most common mutations in the human population and their Medical consequences are
   a. **free (for now)**
   b. come at a cost
   c. relatively cheap
   d. not available to the general public
   e. None

2. The Genetic Home Reference.
   a. **Is mostly directed at the general public**
   b. Is mostly directed at professionals
   c. Is not a public resource
   d. Contains commercial data only
   e. Does not exist

3. The names of human genes recently discovered
   a. are chosen by the group that have discovered the gene
   b. are chosen by a national committee
   c. are chosen by an international committee named SHAKESPEARE
   d. **are chosen by the HUGO international committee**

e. are set as numbers reflecting the discovery date

4. The most authoritative online resource for single gene Mendelian diseases is?
    a. SwissPror
    b. Mendelum
    c. GREGOR
    d. OMIM
    e. DB-snp

5. A Mendelian disease is a disease that
    a. Is caused by one or more mutations in a single-gene
    b. Is caused by mutations in several genes
    c. Results from mutations acquired in somatic cells
    d. Results from mutations passed from the parents
    e. a and d

6. OMIM is
    a. Automatically updated as new data is being published
    b. Manually curated and therefore not always up to date
    c. An essential start that must be completed with pubmed searches
    d. the automated combination of many databases
    e. b and c

7. An SNP is a
    a. Single Nucleotide Polymorphism
    b. Short Natural Polymorphism
    c. Significant Neutral Polymorphism
    d. Splicing Nucleotide Polymorphism
    e. Single Nucleotide Pair

8. Primary databases contain
    a. Curated data of high accuracy
    b. All data generated experimentally
    c. High quality models such as predicted genes
    d. Data and models
    e. Only predicted models

9. The database containing all sequenced genes and genomes published in the scientific literature is
    a. Genebank
    b. ENA
    c. DDBJ
    d. a, b and c
    e. a and c

10. In TrEMBL, the largest database of protein coding genes
    a. Every entry corresponds to a protein whose existence has been proven
    b. Every entry corresponds to a predicted protein
    c. Entries correspond to proven and predicted proteins
    d. Entries correspond to proteins with a known 3D structure
    e. All entries correspond to enzymes

Learning activities for Week 3. Exploring RNA and Protein products of a gene

Short problem

What is the Tissue in which BRCA1 is the most expressed, with which proteins does it interact?

Solution

1-In order to identify the tissues in which BRCA1 is the most expressed, we will use the GTex resource that repertories the quantified levels of expression of human genes as measured in various tissues obtained from recently diseased donors

Gtex can be accessed from: https://gtexportal.org/home/

In there we will simply need to type in BRCA1 in the small search window on the top right corner

The result is displayed in : https://gtexportal.org/home/gene/BRCA1

Interestingly, BRCA1 is not especially expressed in breast or in ovaries.

2- In order to find out about the partners of BRCA1, we will now go to the STRING database that reports all proven and predicted interactions between proteins.

open: https://string-db.org/

Click on Search and enter BRCA1

STRING relies on a literature scan to collect all the reported protein-protein interactions. It will ask you to start from a specific source. In this precise case, you can select the first one. This will take us to a page displaying all the known interactors.

https://string-db.org/cgi/network.pl?taskId=teUxvqvJkkPs

As BRCA1 interacts with both DNA repair proteins (RAD51) and with proteins whose alteration is well known in cancer (p53).

Learning activity

Identify the domains that BRCA1 is made of. How many of these domains have a 3D structure? Visualize the BRCT domain structure and find out a way to reveal the position affected by the M1775K mutant.

background: https://www.ncbi.nlm.nih.gov/pubmed/18285836

Solution

1-let us start by gathering the BRCA1 gene from Uniprot: https://www.uniprot.org/

The protein sequence can be obtained here: https://www.uniprot.org/uniprot/P38398.fasta

We can now go to Interpro that will allow us to scan BRCA1 against databases of known protein domains. In order to do so, you will need to

open interPro: https://www.ebi.ac.uk/interpro/

Cut and paste your sequence

Click on Search and wait a little bit, while your search is taking place at the European Institute of Bioinformatics computers. It should not take more than a couple of minutes.

The result is a display of all the domains contained in all domain databases, PFAM being the most important. Note that the Interpro database is merely a meta-database that combines the information of all existing databases

On this figure we can see all the known domains contained in BRCA1. The two last ones are the BRCT domains. They are available in both Pfam and SMART. Note that a domain is defined by specialists selecting representative sequences on the basis of their interpretation of the literature. For this reason, the same domain can be described in slightly different ways across several databases.

For this specific exercise, let us go to the SMART database

http://smart.embl-heidelberg.de/smart/do_annotation.pl?DOMAIN=SM00292

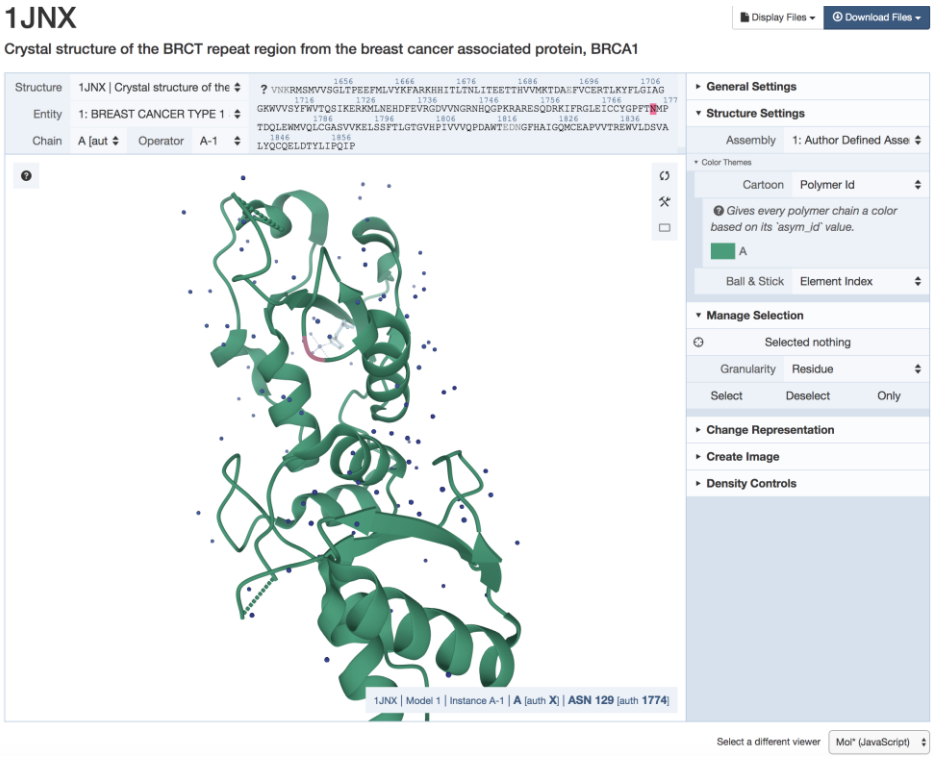
As we can see, this BRCT domain is associated with DNA repair mechanism

If we scroll down a bit, we will find a list of all the structures associated with this domain. One of them is the human BRCA1 sequence whose BRCT domain is available in the PDB database under the name 1jnx

http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1jnx

It is possible to visualize this structure in 3D using one of the provided plugins, but on this site none of them allow us to visualize residues. We will therefore go to the PDB website:

https://www.rcsb.org/3d-view/molstar/1jnx



and use the Ml plugin (selector near the bottom). Note that there exist many different ways to do these visualizations.

Now that the structure is displayed, we can select the Methionine 1775 – highlighted in pink on the figure – as you can see, this methionine stands in a loop and it is interacting with a phosphate. This means that this amino acid is most likely to be interacting with phosphorylated proteins.

Quiz

1. There are two types of large scale sequencing techniques currently available
    a. Sequencing requiring PCR amplification
    b. Sequencing that does not require PCR amplification
    c. RNA sequencing
    d. a and b
    e. a, b and c

2. Which of the following is NOT a high-throughput sequencing method for DNA or RNA
    a. Solexa
    b. PacBio

c. Illumina
d. Nanopore
e. MegaSeq

3. The most common strategy to sequence a new genome involves
   a. Cloning the different parts and sequencing them one after the other
   b. Randomly sequencing pieces of genome and later assembling them
   c. Sequencing each chromosome individually.
   d. Combining RNA and DNA information
   e. Using mass spectrometry

4. When using shotgun sequencing and the illumina technology with reads of about 100 nucleotides to determine a human genome, one needs to sequence
   a. Exactly once every nucleotide in the genome
   b. Exactly twice each nucleotide so as to consider the two haplotype
   c. About 10 times the amount of DNA contained in the genome
   d. About 30 times the amount of DNA contained in the genome
   e. More than 1000 time the amount of DNA contained in the genome

5. The most relevant type of sequencing for a patient suffering from cancer is?
   a. Exome sequencing
   b. Full genome sequencing
   c. Transcriptome sequencing
   d. Hi-C sequencing
   e. a, b, c, and d

6. Which of technique would you use in order to figure out which genes drive blood cell differentiation
   a. RNAseq
   b. Single Cell RNASeq
   c. Single Cell genomic sequencing
   d. Exome Sequencing
   e. metabolomics

7. A PTM is.
   a. a protein post-translational modification
   b. a protein trans-membrane tag
   c. a post transcription motif
   d. a pre transcription motif
   e. a special type of mutation

8. The database containing all proteins with an experimentally characterized 3D structure is
   a. Swissprot

b. TrEMBL
c. PDB
d. uniprot
e. Genebank

9. In order to be able to infer that two proteins may have the same 3D structure, they must have at least
   a. One residue in common
   b. 10 residue in common
   c. 30% identical residues over their entire length when aligned
   d. 100% identical residues over their entire length when aligned
   e. the same length

10. The fraction of proteins with a known 3D structure is closer to:
    a. About half
    b. most of them
    c. 1 in a 100
    d. 1 in a 1,000
    e. 1 in a million

Learning activities for Week 4. Healthy and Unhealthy proteins

Short problem

Use Blast to identify homologues of BRCA1. Use the same program to identify portions of BRCA1 with a known 3D structure. Where is located BRCA1 in the human genome?

Solution

1- Let us start by gathering the BRCA1 gene from Uniprot: https://www.uniprot.org/

The protein sequence can be obtained here: https://www.uniprot.org/uniprot/P38398.fasta

2- We will now use one of the BLAST programs. We could use many different servers but we will use the original BLAST server at the NCBI:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

In order to get homologous sequences, we can simply c/p our sequence and press run. When doing so, BLAST will look for homologous sequences within the Non-Redundant (NR) protein database.

This will take a couple of minutes while the search runs on the NCBI computer. When it is done, the search will report all the homologues contained in NR. These sequences come along with

both the percent of identity and also an E-value. This value tells us the probability that the identity may result from chance. A value of 0 indicates a very high similarity

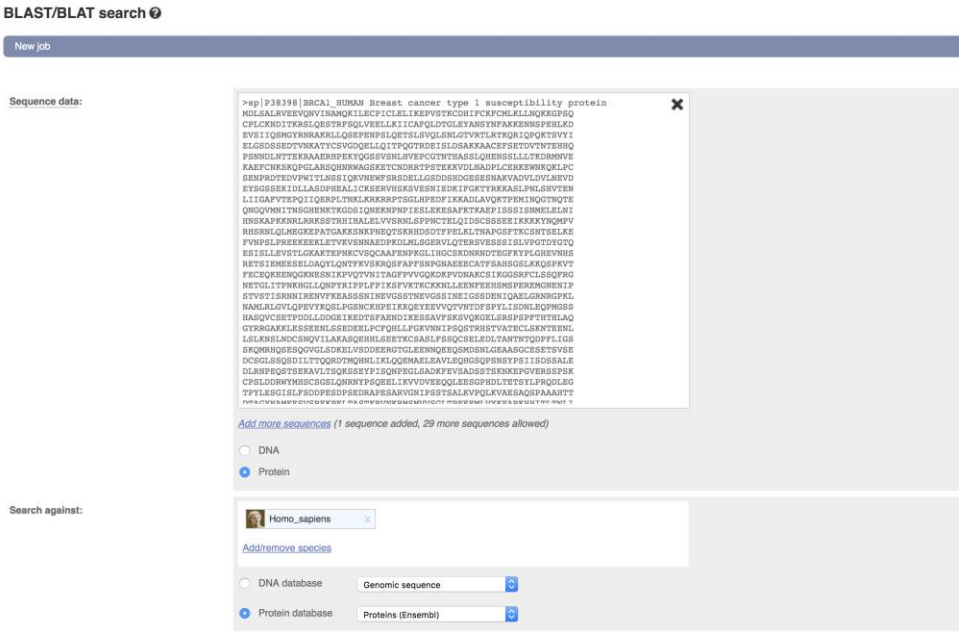
3- If we want to know where on the human genome BRCA1 is located, we need to run BLAST against the human genome. The best place to do so is the ENSEMBL browser, a web service allowing us to explore the human genome

https://www.ensembl.org/index.html

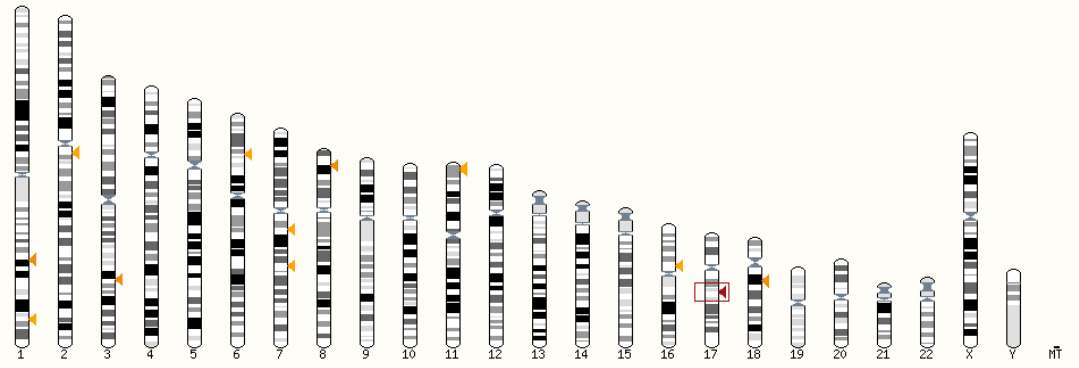
The Blast search is available from:

https://www.ensembl.org/Homo_sapiens/Tools/Blast

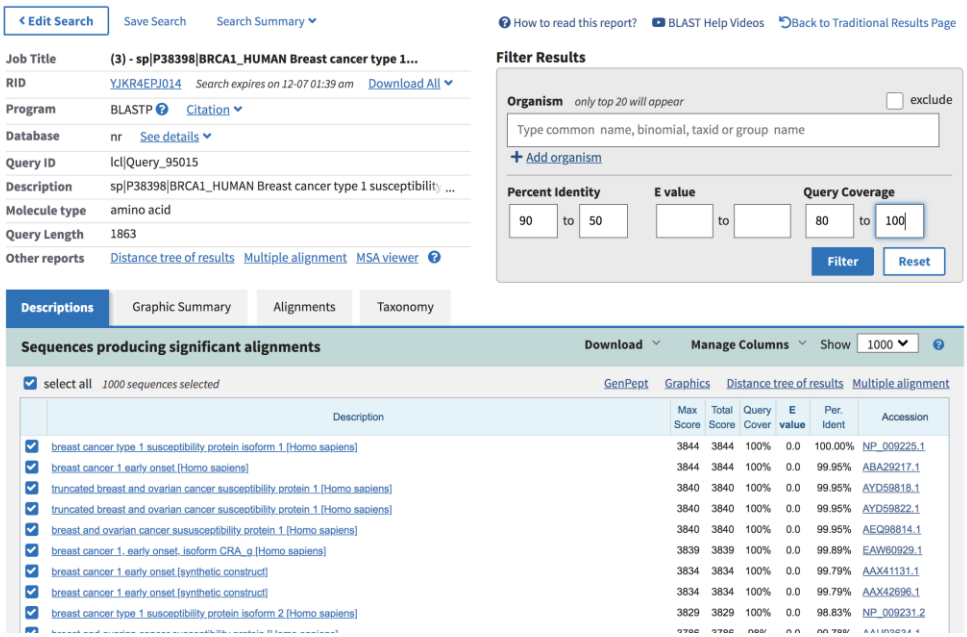
The search runs on the European Bioinformatics Institute computer. To launch the search, c/p your sequence and select the protein database. This will run your search against predicted proteins rather than the whole genome. For a well characterized protein like BRCA1, this option is much more rapid.



The most interesting part of the result is to be found near the bottom that represents a karyotype of the human genome. On this karyotype, the red arrow represents the location of the original BRCA1 gene and the orange ones represent the many homologues of this gene



Thanks to this analysis we now know that BRCA1 is near the centromere of chromosome 17.



Learning activity

Use multiple sequence alignment to estimate the level of conservation of the M1775K (2ING-accesion code in PDB).

Solution

1- In order to identify this level of conservation we will need to align BRCA1 homologues. If we were taking the ones returned by the preview BLAST, they would be too similar to reveal

anything meaningful. We need more distantly related sequences. In order to obtain them, we will run BLAST once again, but we will make sure we filter out the sequences that are too closely related. In order to do so, we will have to start by asking many hits, at least 1,000

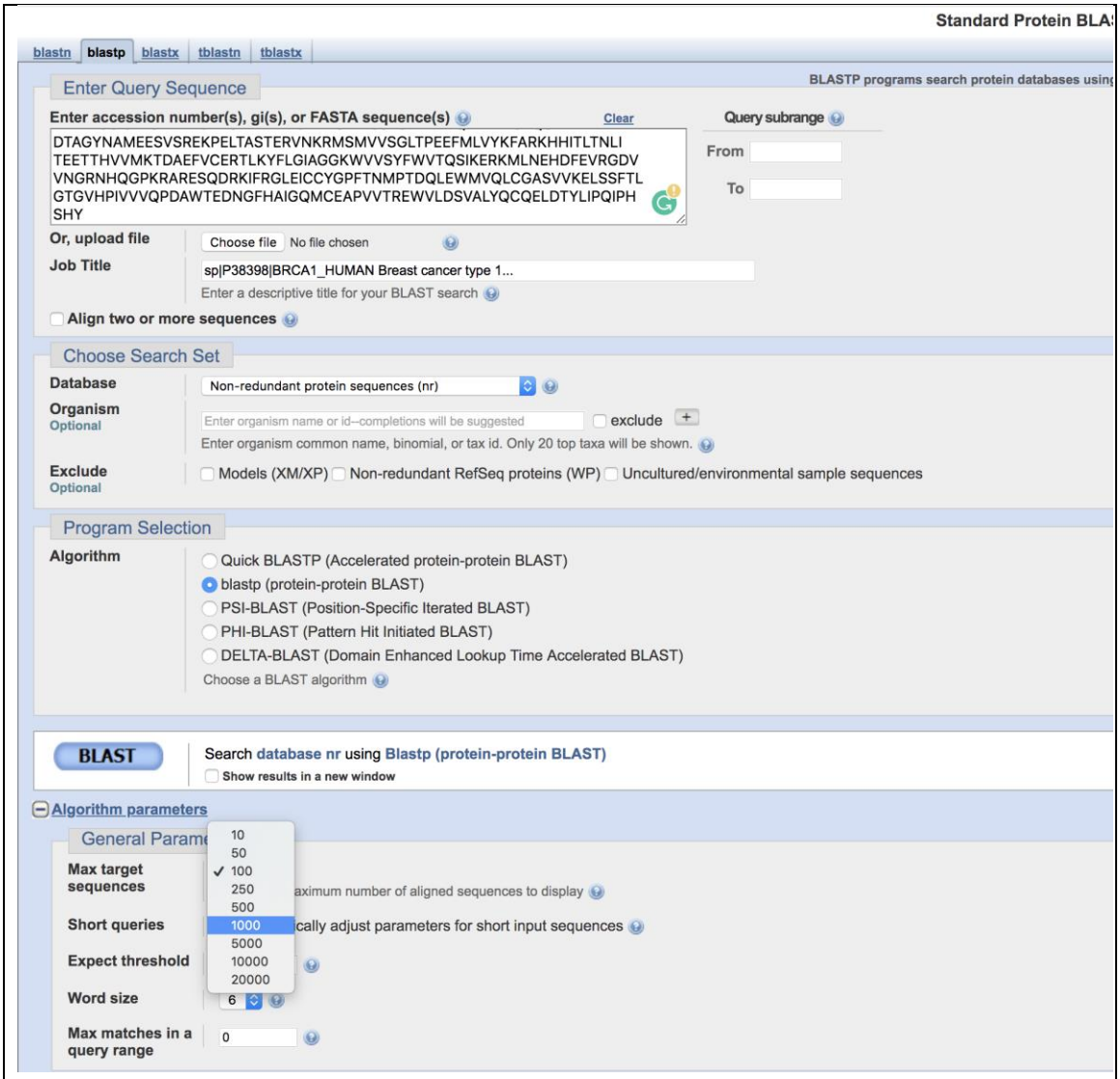
1-let us start by gathering the BRCA1 gene from Uniprot: https://www.uniprot.org/

The protein sequence can be obtained here: https://www.uniprot.org/uniprot/P38398.fasta

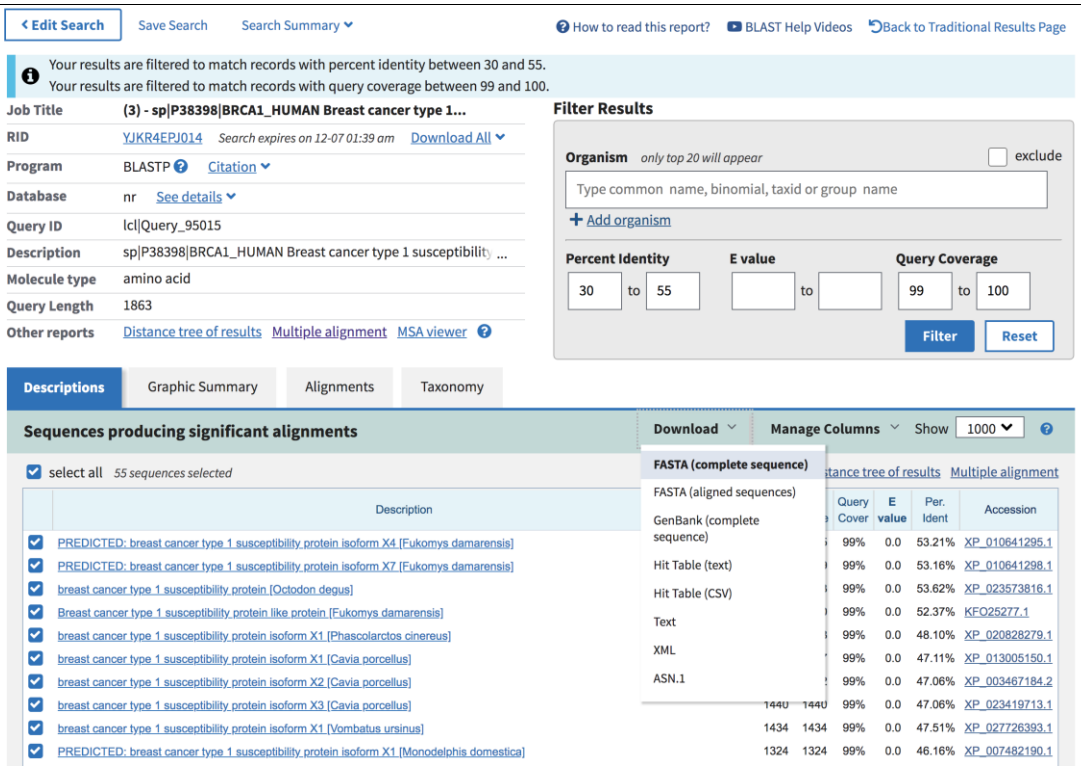
2- We will now use one of the BLAST. We could use many different server but we will use the original BLAST server at the NCBI:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

We will request 1,000 hits to be returned (this is done in the algorithm parameter section)



In order to get an informative multiple sequence alignment, we will select sequences having between 30 and 55% identity with BRCA1, whose alignment covers at least 99% of BRCA1. This coverage is meant to insure we do not select sequences that are only partly homologous to BRCA1.



We will then download our sequences. They come in a simple format named FASTA. In order to make sure our query is included we will have to open this file and c/p the human BRCA1 near the top. If you do so with Microsoft words, make sure you save your file in .txt format.

We will then need to carry out a multiple sequence alignment. For this purpose, we will use Clustal Omega, a very popular aligner. Clustal Omega is not the most accurate aligner but it provides a very good trade off between accuracy and efficiency.

It is available from: https://www.ebi.ac.uk/Tools/msa/clustalo/

This one will take a few minutes to run

When this is done, search for the motif FTNMP. It is the stretch of sequence that contains the methionine we are interested in.

```
XP_025910283.1              WVIQSFKEGRILDEGNFEVRGDVINGRNHQGPRRARQSLTEKIFKDFEICCYGPFTDMTT 1744
XP_025910282.1              WVIQSFKEGRILDEGNFEVRGDVINGRNHQGPRRARQSLTEKIFKDFEICCYGPFTDMTT 1745
XP_025969358.1              WVVQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1783
XP_025969356.1              WVVQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1784
XP_025969357.1              WVVQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1784
XP_025969355.1              WVVQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1784
XP_025969353.1              WVVQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1785
XP_013806921.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1784
XP_025927061.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1783
XP_025927059.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1784
XP_025927060.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1784
XP_025927058.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1784
XP_025927057.1              WVIQSFKEGRILDEDNFEVRGDVINGRNHQGPKRARQSLTERIFKDFEICCYGPFTDMTT 1785
XP_026720501.1              WVIQSFKEGRILDEEDFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1762
XP_026720500.1              WVIQSFKEGRILDEEDFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1771
XP_026720499.1              WVIQSFKEGRILDEEDFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1774
XP_026720498.1              WVIQSFKEGRILDEEDFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1773
XP_027644513.1              WVIQSFTEGRILNEEDFEVRGDVINGRNHQGPRRARQSLTEKIFKDFEICCYGPFIDMTT 1781
XP_027661569.1              WVIQSFTEGRILNEEDFEVRGDVINGRNHQGPRRARQSLTEKIFKDFEICCYGPFIDMTT 1781
XP_027661568.1              WVIQSFTEGRILNEEDFEVRGDVINGRNHQGPRRARQSLTEKIFKDFEICCYGPFIDMTT 1780
XP_010141210.1              WIIQSFKEGRILDEETFEVRGDVINGRNHQGPKRARQALTEKVFKDFEICCYGPFTDMTT 1737
XP_009490367.1              WVIQSFKEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1782
XP_009808601.1              WVIQSFTEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1777
XP_019328559.1              WVIQSFKEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1788
XP_029879002.1              WVIQSFKEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1788
XP_029879003.1              WVIQSFKEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1787
XP_011592497.1              WVIQSFKEGRILDEENFEVRGDVINGRNHQGPKRARQSLTEKIFKDFEICCYGPFTDMTT 1787
XP_019355852.1              WVIQSFKERRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFMDMNT 1795
XP_014464710.1              WVIQSFKERRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFMDMNT 1796
XP_019355858.1              WVIQSFKERRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFMDMNT 1795
XP_019360926.1              WVIQSFKEGRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFTDMNT 1795
XP_019360910.1              WVIQSFKEGRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFTDMNT 1796
XP_019360934.1              WVIQSFKEGRILYESDFEVRGDVINGRNHQGPKRARQSQAGKLFKDFEICCYGPFTDMNT 1795
XP_006125762.1              WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPRKARRSQTGTLFKDFEICCYGPFIDMHT 1809
XP_027690803.1              WIVQSFKEGKILEEYDFEVRGDVINGRNHQGPKRARRSQTGKLFKDFEICCYGPFTDMRT 1828
XP_030396995.1              WIVQSFKEGRILDEYNFEVRGDVINGRNHEGPKRARRSQARKLFKDFEICCYGPFTDMRT 1821
XP_030396992.1              WIVQSFKEGRILDEYNFEVRGDVINGRNHEGPKRARRSQARKLFKDFEICCYGPFTDMRT 1823
TFK10704.1                  WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPKRARRSQAGKLFKDFEICCYGPFTDMRT 1823
XP_024080107.2              WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPKKARRSQAGKLFKDFEICCYGPFTDMRT 1823
XP_023967139.1              WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPKRARRSQAGKLFKDFEICCYGPFTDMRT 1821
XP_023967138.1              WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPKRARRSQAGKLFKDFEICCYGPFTDMRT 1824
XP_023967135.1              WIVQSFKEGRILDEYDFEVRGDVINGRNHQGPKRARRSQAGKLFKDFEICCYGPFTDMRT 1826
AAX92675.1                  WVVQSFKEGKMLPECDFEVRGDVINGRNHRGPERARESQGMKIFRGLEICCYGPFTDMST 1758
NP_001029141.1              WVVQSFKEGKMLPECDFEVRGDVINGRNHRGPERARESQGMKIFRGLEICCYGPFTDMST 1754
XP_007482190.1              WVVQSFKEGKMLPECDFEVRGDVINGRNHRGPERARESQGMKIFRGLEICCYGPFTDMST 1754
XP_007482191.1              WVVQSFKEGKMLPECDFEVRGDVINGRNHRGPERARESQGMKIFRGLEICCYGPFTDMST 1752
XP_020828279.1              WVVQSFKEGKVLPECDFEVRGDVINGRNHRGPERARESQGVKIFKGLEICCYGPFTDMPT 1771
XP_027726393.1              WVVQSFKEGKVLPECDFEVRGDVINGRNHRGPERARESQRVKIFKGLEICCYGPFTDMPT 1766
XP_023419713.1              WVTQSMKAGRVLDEHDFEVRGDVINGRSHQGPRRARESQDRQIFRGLDVYCYGPFTNMPT 1667
XP_013005150.1              WVTQSMKAGRVLDEHDFEVRGDVINGRSHQGPRRARESQDRQIFRGLDVYCYGPFTNMPT 1668
XP_003467184.2              WVTQSMKAGRVLDEHDFEVRGDVINGRSHQGPRRARESQDRQIFRGLDVYCYGPFTNMPT 1667
sp|P38398|BRCA1_HUMAN       WVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPFTNMPT 1777
XP_023573816.1              WVTQSMKERRVLDEHDFEVRGDVINGRSHHGPRRARESQDRSIFAGLAVCCCGPFTNMTT 1748
KFO25277.1                  WVTQSMKERRMLDERDFEVRGDVVNGRNHQSPKRARESRDQKIFGGLDVCCCGPFTNMPT 1675
XP_010641295.1              WVTQSMKERRMLDERDFEVRGDVVNGRNHQSPKRARESRDQKIFGGLDVCCCGPFTNMPT 1651
XP_010641298.1              WVTQSMKERRMLDERDFEVRGDVVNGRNHQSPKRARESRDQKIFGGLDVCCCGPFTNMPT 1650
                            *: **:.  ::* *  *******:***.*..*.:**.:    :* .: : * *** :* *
```

This will highlight the following block. On this representation, completely conserved columns are marked with a "*", well conserved columns with a ":" and less but significantly conserved with a ".". Columns that are not very well conserved are not marked.

As we can see, our methionine of interest is highly conserved, and in general, we may speculate that any mutation affecting a column with a "*" may be detrimental to the function of this protein.

In reality things are a bit more complex and we would need a more sophisticated substitution model in order to predict how severe a mutation may be. For instance, a mutation turning a leucine into an isoleucine will usually not be as serious as a mutation turning a methionine into a lysine.

Quiz

1. Which of the following DNA alteration is not considered a mutation?
    a. Single point mutation (SNP)
    b. Copy number variation (CNV)
    c. methylation
    d. insertion
    e. deletion

2. In a genome, sites evolving under purifying selection
    a. evolve faster than one would expect
    b. evolve at random pace
    c. evolve at a lower pace than one would expect
    d. evolve at the expected pace given a neutral model
    e. are the most abundant

3. The most convenient way to quantify the evolutionary cost for substituting one amino acid into another is:
    a. the genetic code
    b. a substitution matrix like PAM
    c. analyzing the predicted 3D structure of both proteins
    d. an alignment of the two proteins
    e. a stochastic evolutionary model

4. When two residues are in the same column in an alignment. This can be considered a statement that these two residues?
    a. Correspond to the same residue in the common ancestor of the two sequences
    b. Play an equivalent role in their respective structures
    c. Are homologous to one another
    d. Are not mutable
    e. a,b and c

5. The P-Value reported by BLAST is
    a. The probability that two sequences may have a similarity or higher score than the one reported, as a consequence of chance
    b. The probability that the two sequences may be homologous
    c. The probability of the reported score being the result of chance
    d. The probability of finding such a good alignment by chance
    e. The number of alignments equally good that could have been generated by chance

6. When comparing DNA sequences, BLAST

a. Is more accurate than when comparing protein sequences
b. Is less accurate than when comparing protein sequences
c. Is just as accurate as it is when comparing protein sequences
d. Is faster than when comparing protein sequences
e. Is slower than when comparing protein sequences

7. In an alignment a '-' symbol, represents a gap, this means that
   a. an insertion occurred while the sequences were diverging from the common ancestor sequence
   b. A deletion occurred while the sequences were diverging from the common ancestor
   c. Either an insertion or a deletion occurred but we cannot tell which one
   d. A cross over took place between haplotypes
   e. The two sequences cannot be homologous

8. In a multiple sequence alignment, the most likely explanation for a highly conserved column surrounded by more variable positions is.
   a. The total absence of mutations on the considered site
   b. The decreased capacity of any organism featuring a mutation on this site to survive
   c. The presence of an active site on this position
   d. Frequent reversions
   e. It is not known

9. Within proteins, domains are?
   a. Active sites
   b. Regions of the protein that tend to be conserved across proteins
   c. Long regions featuring the same amino acid
   d. Regions of the protein that can often be mapped to specific functions
   e. b and d

10. In order to figure out the importance of a somatic mutation in a disease an efficient strategy involves:
    a. Sequencing the entire genome
    b. Sequencing the coding parts of the genome
    c. Sequencing specific genes involved in cancer
    d. Looking for genomic duplications
    e. Sequencing the transcriptome of the tissues

Forum activity

Imagine you have applied for some funding to do sequencing and subsequent analysis. By the time you get your money, the cost of sequencing will have substantially dropped. How should you use this extra money. Should you do more sequencing? If so more samples, or deeper sequencing, or would the money be better spent on more powerful computers. If so should you use the most powerful CPU, or many more less powerful CPUs, or machines with more memory. What do you think will be the criteria for all these decisions, is there such thing as a one size fits all? Can you propose concrete example for which you think a specific strategy will be clearly more suited than a generic one.

Readings recommended for the module

Kahn S. On the future of genomics data. *Science* 331, 728-729 (2011).
https://science.sciencemag.org/content/331/6018/728.long

Yang et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N Engl J Med* 2013; 369:1502-1511.
DOI: 10.1056/NEJMoa1306555
https://www.nejm.org/doi/full/10.1056/NEJMoa1306555

Berger, B., Peng, J., & Singh, M. (2013). Computational solutions for omics data. *Nature reviews. Genetics*, *14*(5), 333–346. doi:10.1038/nrg3433.
http://ncbi.nlm.nih.gov/pmc/articles/PMC3966295/

MODULE 2. COMPUTATIONAL STATISTICS FOR CLINICAL DOCTORS

Learning activities for Week 1. Descriptive statistics

Short problem

1. 50 applicants for a specific job were tested. The scores obtained were:

4	4	2	10	1	9	5	3	4	5
6	6	7	6	8	7	6	8	7	6
5	4	4	4	5	6	6	7	5	6
6	7	5	6	6	7	5	6	4	3
2	6	6	7	7	8	8	9	8	7

a) Determine the type of variable to be analyzed.

b) Build the frequency table and the corresponding graphic representation by hand and also using the RStudio software. Interpret the results.

c) Find the cut-off score that selects approximately 20% of the best candidates

2. Consider the following data corresponding to the waiting time in hours, in a medical center, during period of flu outbreak:

2,29	2,67	2,84	2,65	2,52	1,75	2,12	1,54	1,95	1,82
1,95	1,75	1,92	1,92	1,46	1,15	1,70	1,86	1,04	1,06

a) Calculate and interpret the mean, median, standard deviation and quartiles.

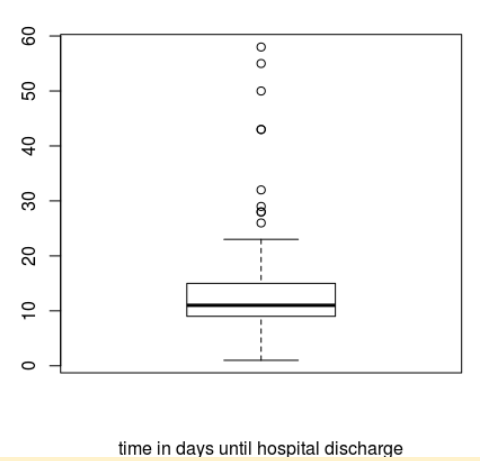
b) Represent the boxplot diagram.

c) Repeat the above calculations using the statistical package RStudio. First you will have to introduce the data by yourself and then apply the corresponding instructions.

3. In a hospital, a tumor has been removed in 86 patients with gastric cancer. Information about the time in days from the day of the surgical intervention until hospital discharge is available. Perform the maximum possible interpretation based on the information from the available data, explain the reasons why you cannot further interpret the available data and indicate the additional required data which should be provided, in order to correctly interpret all the provided information.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	9.00	11.00	14.35	15.00	58.00

sd: 10.53595 var: 111.0063 IQR: 6 length: 86



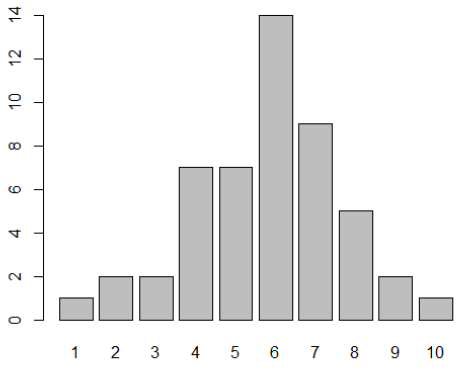
Solution

1.

a) This is a discrete numerical variable

x

	Frequency	Percent
1	1	2
2	2	4
3	2	4
4	7	14
5	7	14
6	14	28
7	9	18
8	5	10
9	2	4
10	1	2
Total	50	100



b) Here you would have to calculate the accumulated percentage and then choose the score that leaves the best marks. 7 selected the best 34% and 8 the best 16%.

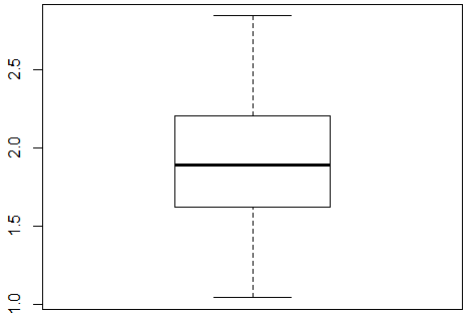
```
> dist
 scores      f     rf     rf(%)  cf    cf(%)
  1          1   0.02      2      1     2
  2          2   0.04      4      3     6
  3          2   0.04      4      5    10
  4          7   0.14     14     12    24
  5          7   0.14     14     19    38
  6         14   0.28     28     33    66
  7          9   0.18     18     42    84
  8          5   0.10     10     47    94
  9          2   0.04      4     49    98
  10         1   0.02      2     50   100
```

```
#f= Absolute frecuency; #rf= Relative frecuency; #rf(%) Percentage rel
ative frecuency; #cf=cumulative frequency; #cf(%)=Percentage cumulativ
e frequency
```

2. a)

Mean: 1.898
Median: 1.890
Standard deviation: 0.5157274
Quartiles: Q1=1.660 Q2=Median Q3=2.163

b)



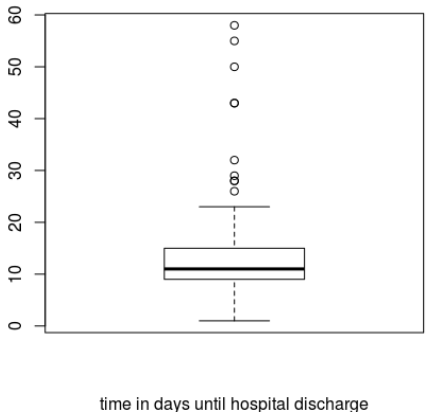
c)

```
y<-c(2.29,2.67,2.84,2.65,2.52,1.75,2.12,1.54,1.95,1.82,1.95,1.75,1.92,
1.92,1.46,1.15,1.70,1.86,1.04,1.06)
summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.040   1.660   1.890   1.898   2.163   2.840
sd(y,na.rm=TRUE)
[1] 0.5157274
```

3.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	9.00	11.00	14.35	15.00	58.00

sd: 10.53595 var: 111.0063 IQR: 6 length: 86



Learning activity

Laparoscopy and operative laparoscopy both play important roles in treating health issues. Laparoscopy refers to using multiple small incisions with ports to perform surgery with specialized instruments. Laparotomy is an older technique that relies on a single large incision, through which a surgeon uses his or her hands to directly perform the procedure. Recovery from laparoscopy is generally faster, because there is less pain; for some operations, the outcomes are also better. However, many operations still must be performed through a laparotomy incision.

The database that will be used in this activity is laparoscopia_vs_laparatomia.dat. This is a database that contains information on 86 patients with gastric cancer who have had a tumor removed using one or another surgical technique. The collected data include: patient **age**, **sex**, **surgeon technique** used (Surg_tecn: 1 (laparoscopy); 2 (laparotomy)), **duration of intervention**

in minutes (dur_intervention), **number of nodes removed** (n_ganglia), **time to hospital discharge in days** (time_to_hospital_discharge), and **survival after 1 year** (exitus_1_year).

1. Given the following database it is requested:

a. Classify the variables in the database according to discrete numerical variable, continuous numerical variable, ordinal categoric variable or nominal categoric variable indicating in a summary way how we can approach the exploratory analysis of each of them.

Variable	Classification and Information processing
age	
sex	
surgeontechnique	
duration of intervention	
number of nodes removed	
time to hospital discharge	
survival after 1 year	

b. For the three variables that are in bold type, you are asked to make a summary: frequency tables for discrete qualitative or numerical variables with few different values and descriptive values (mean, median, variance, standard deviation, quartile value, minimum value and value maximum), for continuous numerical variables. For each of the three variables you have to make a brief comment interpreting the numerical results obtained.

c. Make the appropriate graph for each of the three variables in the previous section (bar or sector diagram for the qualitative or discrete variables with few different values, histogram or box diagram for the continuous variables. Add a brief comment to each of the graphics.

d. Analyze and interpret the boxplots for the variables: age; duration of intervention, number of nodes removed and time to hospital discharge. What do you observe in each case in terms of symmetry, outliers, data dispersion?

e. Consider the variable "number of nodes removed". Remove the outlier and recalculate the statistics. Compare the result with the statistics calculated in section (b). What do you observe? Justify the answer.

Solution

1. a)

Variable	Classification and Information processing
age	continuous numerical variable as its frequency table has a lot of values with a very low percentage. Histogram or boxplot. In addition we can calculate descriptive measures
sex	nominal categoric variable. Frequency table.
surgeon technique	nominal categoric variable. Frequency table.
duration of intervention	continuous numerical variable as its frequency table has a lot of values with a very low percentage. Histogram or boxplot. In addition, we can calculate descriptive measures
number of nodes removed	continuous numerical variable as its frequency table has a lot of values with a very low percentage. Histogram or boxplot. In addition, we can calculate descriptive measures
time to hospital discharge	continuous numerical variable as its frequency table has a lot of values with a very low percentage. Histogram or boxplot. In addition, we can calculate descriptive measures
survival after 1 year	nominal categoric variable. Frequency table.

b)

```
Sex
      Frequency   Percent
1            50    58.14
2            36    41.86
Total        86   100.
```

Knowing that 1 represents men and 2 represents women, we observe that there is a higher percentage of men than women (58.14% vs 41.86%)

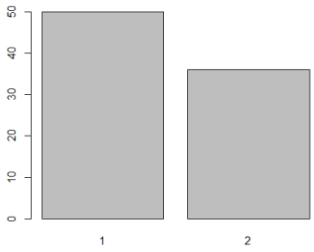
```
> summary(dur_intervention)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   90.0   191.2   255.0   262.9   330.0   480.0       2
> sd(dur_intervention,na.rm=TRUE)
[1] 87.72659
> var(dur_intervention,na.rm=TRUE)
[1] 7695.955
> IQR(dur_intervention,na.rm=TRUE)
[1] 138.75
> length(dur_intervention)
[1] 86
```

The minimum number of minutes of the intervention are 90 and the maximum 480. The mean is 262.9 and 50% of the patients stay less than 255 minutes and the other 50% stay more than 255 min. The mean and the median are quite similar, so a rather symmetric distribution can be assumed.

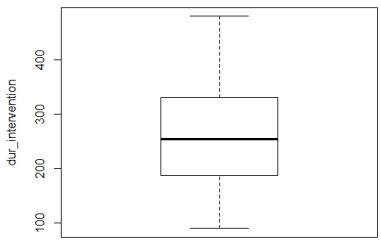
```
> summary(n_ganglia)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   11.00   19.50   22.14   29.75   62.00
> sd(n_ganglia,na.rm=TRUE)
[1] 14.03036
> var(n_ganglia,na.rm=TRUE)
[1] 196.8509
> IQR(n_ganglia,na.rm=TRUE)
[1] 18.75
> length(n_ganglia)
[1] 86
```

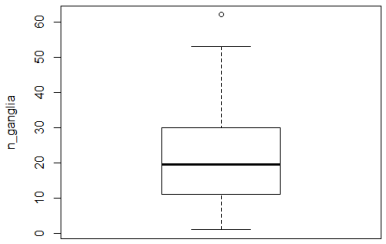
The minimum number of extirpated ganglia are 1 and the maximum 62. However, the mean is 19.5 and 25% of the patients have less than 11 extirpated ganglia and another 25% have more than 29.75 extirpated ganglia.

c)



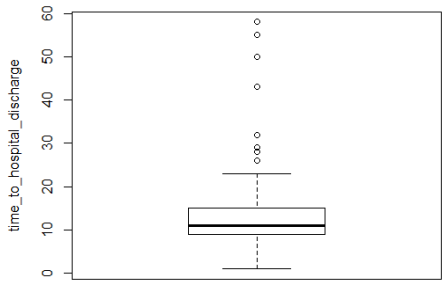
The variable sex can be represented by a bar diagram. As we observed in the frequency table the bar of men is higher than women.

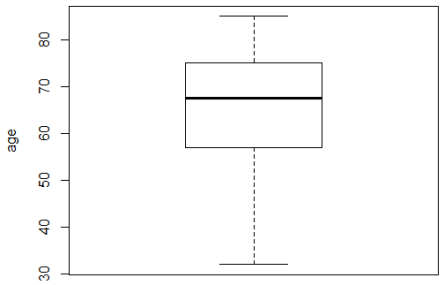




These two variables can be represented by a boxplot. The duration graph has no outliers and is quite symmetric. However, the mean is greater than the median, so this indicates a certain positive asymmetry in the data. The other, which represents the number of ganglia, shows an outlier and shows positive asymmetry, which is also confirmed because the mean is greater than the median.

d) In the previous section two of the variables have already been described: duration of intervention, number of nodes removed.
Regarding the time to hospital discharge (graph on the left) we observe some outliers and a certain positive asymmetry. However, removing outliers the dispersion of the data is not very remarkable.





On the other hand, the boxplot for age does not show any outliers and shows negative asymmetry. In this case the dispersion of the data is more remarkable.

e)

With outliers

```
> summary(n_ganglia)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   11.00   19.50   22.14   29.75   62.00
> sd(n_ganglia,na.rm=TRUE)
[1] 14.03036
> var(n_ganglia,na.rm=TRUE)
[1] 196.8509
> IQR(n_ganglia,na.rm=TRUE)
[1] 18.75
> length(n_ganglia)
[1] 86
```

Without outliers

```
> summary(n_ganglia)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
   1.00   10.50   19.00   21.67   29.50   53.00
> sd(n_ganglia,na.rm=TRUE)
[1] 13,419
> var(n_ganglia,na.rm=TRUE)
[1] 180,057
> IQR(n_ganglia,na.rm=TRUE)
[1] 19
> length(n_ganglia)
[1] 85
```

The statistics that change the most are the mean, the maximum, the standard deviation and the variance, since they are those that are not robust to outliers.

Quiz

1. The measure of dispersion more difficult to interpret in the sense that it is not expressed in the same units of measurement as the mean is:
   a. Standard deviation
   b. Interquartile range.
   c. Mode
   d. **Variance**
   e. Median.

2. What is the graph that allows to represent more directly and visually different position measurements at the same time:
   a. Column bar graph.
   b. Boxplot.
   c. Histogram.
   d. Pie chart.
   e. Scatter plot.

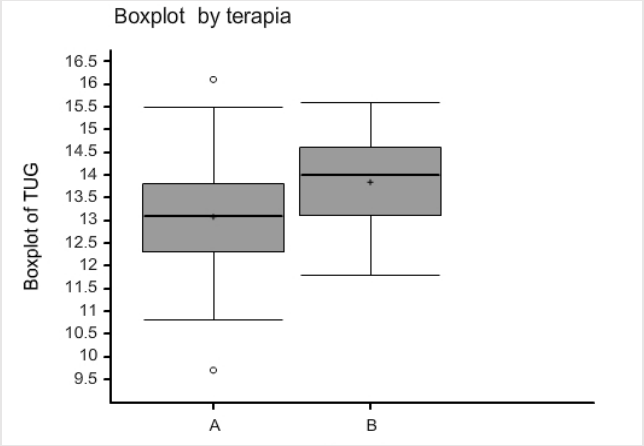
3. Indicate the correct answer
   a. Inferential statistics are used to extract conclusions about the population as a whole to which the individuals in the sample belong.
   b. Inferential statistics are used to extract conclusions about the sample as a whole.
   c. Descriptive statistics are used to extract conclusions about the population as a whole to which the individuals in the sample belong.
   d. Inferential statistics are used to describe and summarize the information obtained.
   e. None of the above statements are correct.

4. Which of the following statements is false?
   a. The accumulated frequencies are very useful to describe nominal qualitative variables.
   b. The accumulated frequencies are very useful to describe ordinal qualitative variables.

c. The accumulated frequencies are very useful to describe discrete quantitative variables.
d. The accumulated frequencies are very useful to describe continuous quantitative variables.
e. The relative frequencies are very useful to describe ordinal qualitative variables

5. Select the correct statement:
a. Nominal variables are those that can be counted or enumerated.
b. Nominal variables are those variables that have a quality or attribute of the population and are measured in some order.
c. Nominal variables are those that can only take two values.
d. Nominal variables are those that come from the name and are characterized by not having a measure of distance between their values.
e. Nominal variables are those that can take whole values or decimals.

6. The following graph represents the value distribution of the index "timed up and go" (TUG), that evaluates the dynamic balance, differentiating between two groups of therapy (A i B)



Which of the following statements is false?

a. The median TUG in therapy A is lower than that of the B therapy.
b. The interquartile range in therapy A match the interquartile range in therapy B and has a value of 1.5.
c. We only observe TUG values above the median plus 1.5 times the interquartile range in the therapeutic group A.
d. The therapeutic group B presents a negative asymmetry.
e. **The 75% lower TUG values in the therapeutic group B are between values 11,8 and 14.**

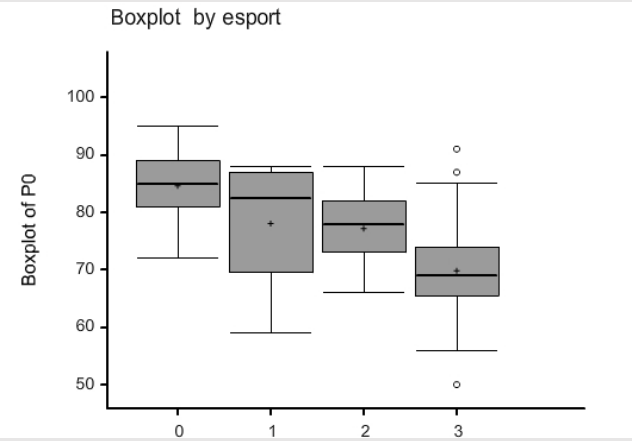
7. Select the option that only includes measures that are not affected by outliers:
a. Mean, Variance and Quartiles
b. Median; Quartiles and interquartile range
c. Median and mean

d. Median, mínimum and maximum
e. Quartiles, range and interquartile range.

8. Which of the following statements is true?

a. Whenever a perfectly symmetrical distribution of values is presented, the variance and the typical deviation must necessarily coincide.
b. Whenever a perfectly symmetrical distribution of values is presented, the interquartile range contains 50% of the central value of the data.
c. Whenever a perfectly symmetrical distribution of values is presented, the average is greater than the median.
d. Whenever a perfectly symmetrical distribution of values is presented, the variance is less than the standard deviation.
e. Whenever a perfectly symmetrical distribution of values is presented, the mean and the median must necessarily coincide.

9. The following graph represents the distribution of values (P0) of the arterial pressure variable at rest (P0) in 4 groups of different students; those who never do sport (0), those who do it sporadically (1), those who do it between 1 and 3 times a week (2) and those who practice it more than 3 times a week (3):



Which of the following statements is false?

a. The median of P0 has a decreasing tendency with the assiduity of the practice of the sport.
b. The group with the lowest variability in values of P0 in terms of interquartile range is group 1.
c. The only group with positive asymmetry in the values of P0 is group 3.
d. Groups 0, 1 and 2 do not have atypical values of P0.
e. The 25% lower values of P0 in group 3 have values lower than 70.

10. A good sample:

a. is that which allows us to study the whole population.
b. is that in which we only represent the part of the healthy population.
c. We will never choose individuals from the population at random.
d. We will select the sample based on the results we hope to find.
e. It must be representative or faithful reflection of the population, in such a way that it reproduces its basic characteristics.

Readings recommended for the week

Larson M. G. (2006). Descriptive statistics and graphical displays. Circulation, 114(1), 76–81.

Rodrigues CFS, Lima FJC, Barbosa FT. Importance of using basic statistics adequately in clinical research. Braz J Anesthesiol. 2017;67(6):619-25. doi: https://doi.org/10.1016/j.bjan.2017.01.003.

Arroyo-Borrell, E., Renart, G., Saurina, C. *et al.* Influence maternal background has on children's mental health. *Int J Equity Health***16,** 63 (2017) doi:10.1186/s12939-017-0559-1.

Kaliyadan F, Kulkarni V. Types of Variables, Descriptive Statistics, and SampleSize. Indian Dermatol Online J. 2019;10(1):82–86. doi:10.4103/idoj.IDOJ_468_18.

Learning activities for Week 2. Statistical inference

Short problem

Hypothesis testing involves comparing two competing hypotheses (the null and the alternative). We accept or reject the null hypothesis given the magnitude of a statistic that we compute in the data and a probabilistic model explaining such statistic under the null hypothesis. Then, if the probability of observing such value or a greater/smaller one is smaller than a given threshold of accumulative probability α, we reject the null hypothesis and accept the alternative one. However, remember that accepting the null hypothesis does not necessarily imply that the alternative hypothesis is ultimately wrong and vice-versa. The next table recapitulates the four possible outcomes from a contrast of hypotheses:

	My test says	
The truth is	Accept null hypothesis	Accept alternative hypothesis
Null hypothesis	Correct decision	Error type I (□)
Alternative hypothesis	Error type II (□)	Correct decision

(1-□□ corresponds to the power of the statistic.

Error type I and II can be interpreted in another way. Imagine that a (pharma) company is selling a (medical) product that they say has X measures, which corresponds to the null hypothesis. The company would like to minimize the number of times that the product is rejected because it does not reach the specified X measures, which would be the alternative hypothesis. This is known as the producers' risk, and corresponds to the Error type I. Conversely, the company is interested in accepting as much as possible the null hypothesis. Thus, the test of the company would say: "accept the null hypothesis" (the size of the product is X) but the truth would be "the size is not X" (Error type II). That would be the consumers' risk.

Obviously, we would like to minimize both Error type I and type II so we take the correct decision most of the times. However, this is not easy because (choose the correct one):

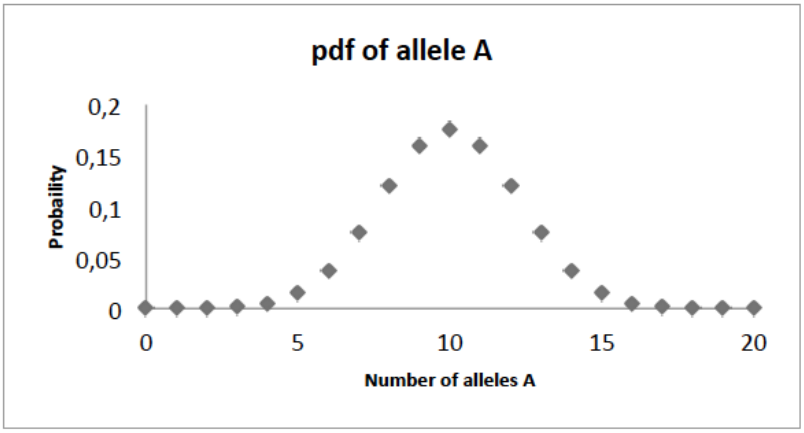
- Error type II depends on threshold α, which is by definition unknown
- We know the shape of the probability distribution of the null distribution, not under the alternative one (CA)[1]
- Error type II is the power, and we must keep it as large as possible

Dr. Smith would like to use a genetic locus as a marker for a given phenotype. Previous reports in other population show the genetic locus has two variants A and B with a frequency of 0.5 each. Dr. Smith took a sample of 10 individuals (20 alleles) and counts the number alleles A as 15 and B as 5 in the samples. Dr. Smith knows that if the statistical distribution of number of alleles if we take 20 alleles follows a binomial distribution with a probability distribution function (pdf) such as this:

[1] CA: Correct Answer

number of heads out of 20 trials	probability under the assumption that the coin is fair
0	9.536743e-07
1	1.907349e-05
2	1.811981e-04
3	1.087189e-03
4	4.620552e-03
5	1.478577e-02
6	3.696442e-02
7	7.392883e-02
8	1.201344e-01
9	1.601791e-01
10	1.761971e-01
11	1.601791e-01
12	1.201344e-01
13	7.392883e-02
14	3.696442e-02
15	1.478577e-02
16	4.620552e-03
17	1.087189e-03
18	1.811981e-04
19	1.907349e-05
20	9.536743e-07

Which visually looks like:



Using this distribution, how likely is to observe exactly 15 alleles A?

- 0.01478577 (CA)
- 0.03696442
- 0.1761971

Which is the contrast of hypothesis we want to conduct (choose one)?

- H_0: frequency of allele A = 0.5; H_1: frequency of allele A = 0.6
- H_0: frequency of allele A = 0.5; H_1: frequency of allele A ≠ 0.5 (CA)
- H_0:frequency of allele A = frequency of allele B; H_1: frequency of allele B = 0.5

What would be the accumulative probability that we would compute in order to conduct the test of hypotheses (i.e. our allele A frequency = 0.5 and observing 15 alleles A out of 20 trials is expected)?

- In fact the right test to do this contrast of hypothesis is a Chi-Square test
- P(allele A ≥15)
- Since we are interested in " frequency A≠ 0.5", the accumulative probability must have into account the lower and upper probability tail. Hence, P(allele A≤ 5) +P(allele A ≥15) (CA)

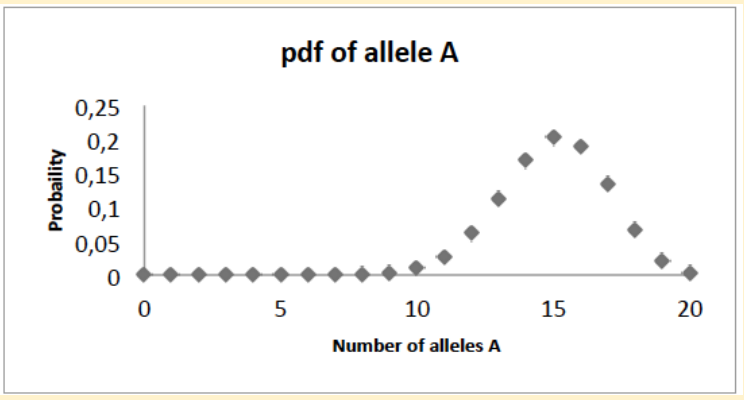
If we set our threshold α to 0.05, shall we accept or reject the null hypothesis?

- Accept the null hypothesis
- Reject the null hypothesis (CA)
- The power does not allow to reach a conclusion

What would be minimum/maximum number of alleles A out of 20 trials so we would reject the null hypothesis at pvalue ≤ 0.05?

- 19 allele A (pvalue = 4.005433e-05)
- 15 allele A (pvalue = 0.04138947) (CA)
- 10 allele A (pvalue= 0.823803)

Usually, the probability distribution under the alternative hypothesis is not known. However, sometimes it is possible to build this distribution using the observed result from the experiment as our best estimate of the parameter. In the case of Dr. Smith experiment, he estimates the frequency of allele A as 0.75. Following is the probability distribution function we could obtain given this result:



number of alleles A out of 20	probability under the assumption that frequency of allele A =0.75
0	9.0949E-13
1	5.457E-11
2	1.5552E-09
3	2.7994E-08
4	3.5693E-07
5	3.4265E-06
6	2.5699E-05
7	0.00015419
8	0.00075169
9	0.00300675
10	0.00992228
11	0.02706075
12	0.06088669
13	0.1124062
14	0.16860929
15	0.20233115
16	0.18968545
17	0.13389562
18	0.06694781
19	0.02114141
20	0.00317121

What would be the power of this contrast of hypothesis given alpha = 0.05?

- Since the number of heads for which we reject the null hypothesis at alpha = 0.05 is 1, the power is 1
- Since the number of heads for which we reject the null hypothesis at alpha = 0.05 is 15 or more, the power is 0.62 (CA)
- There is no power to compute. The distribution under the alternative hypothesis is not really known.

Solution

Solutions of the previous problem are marked as CA (Correct Answer)

Learning activity

A classical test is to check if the mean value that we compute out of a set of samples corresponds to what the maker has specified in the product. That is, we can set a statistical test in the form:

$$H_0: \bar{X} = \mu$$
$$H_1: \bar{X} \neq \mu$$

The statistic:

$$t_s = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

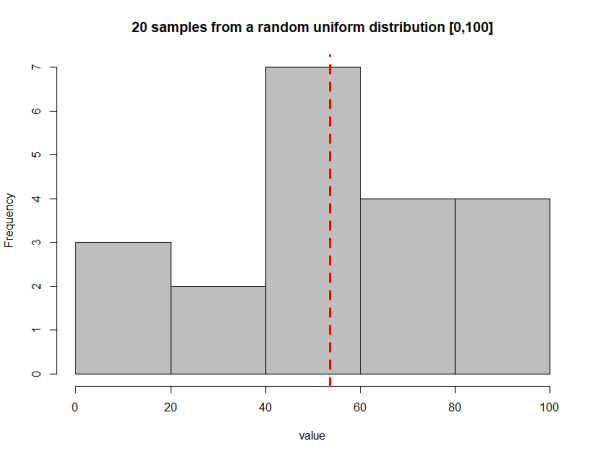
$$\sigma = \sqrt{\frac{\sum(\bar{X} - x)^2}{n-1}}$$

Follows a Student t distribution with *n-1* degrees of freedom under H_0.

A classical misunderstanding is thinking that the assumption of *normality* refers to the distribution from where the samples come.

In this exercise, we will see that the means from a distribution tend to follow normal distributions if *n* is large enough using R-software (https://www.r-project.org/).

Imagine that we conduct an experiment sampling 1,000 times 20 samples from a uniform distribution in the range [0,100]. The histogram of one of these 1,000 resamples (and its mean) could look like (this will change from run to run since we are sampling at random from the uniform distribution):



- Using R software, the code would look like (CODE 1):

  ex <- runif(20, 0, 100); # take 20 samples from a uniform distribution in the range 0 to 100
  hist(ex, main="20 samples from a random uniform distribution [0,100]", col="gray",xlab="value"); # do the histogram
  abline(v = mean(ex), col="red", lwd=3, lty=2); # add the mean to the histogram

- Is it normally distributed?

Now let us see the distribution that comes out of repeating 1,000 times the same experiment and computing at each simulated dataset of 20 individuals the mean.

- In the R console, type (CODE 2):

```
means <- rep(0,1000); # store the means of the simulations
for(m in 1:1000) # repeat 1000 times the sampling of 20 individuals from the uniform distribution
{
  means[m] <- mean(runif(20, 0, 100)); # store the mean of the simulation m
}
hist(means, col = "gray"); # histogram of the means
abline(v = mean(means), col="red", lwd=3, lty=2); # mean of the means
```

- Which is the shape of the means of 20 sampled items from a uniform random distribution [0,100]?

If instead of using *runif(20,0,100)* you use *rpois(20,2)*, you will generate samples from a Poisson distribution (with lambda parameter 2).

- How is the histogram of a single realization of the experiment (CODE 1)?. Does it look normal?
- How is the shape of the means of 1000 realizations of the experiment, each experiment with sample size 20 (CODE 2)?

This behaviour of the means towards a bell shape distribution is due to the *Central limit theorem*. It states that, as *n* tends to be large, the sum of independent variables will tend to *follow* a Normal distribution.

Once we have seen this basic property, we are going to conduct our first t-student test to check if what a product labelling says is true or not. Let us imagine that a maker says the product that we buy has on average 55 units. We do not trust too much this and we decide to construct an experiment to test it. First, we buy 10 items. We measure the variable of interest and we get this:

55.06609,16.41506,75.28631,60.85577,43.05993,71.76336,30.78438,49.65377,79.66662,86.59940

- Compute the sample mean.
- Compute the standard deviation of the sample. In R console
- Compute the t-statistic
- Which is the probability of observing a bigger value than *t* if *t* is positive? Or smaller than *t* if *t* is negative in a student t distribution with n-1 degrees of freedom?

The function *pt(x,n, lower.tail = F)* in R computes the 1- cumulative probability (that is, the probability of observing a value greater than x) of x in a Student's t distribution with *a given* degrees of freedom.

```
Pvalue <- 2*pt(t,9,lower.tail = F);
```

- Why the p-value that we compute is multiplied by two?
- Which is the final p-value that we get?
- What does that mean?
- Shall we accept (the mean is 55) or reject (the mean of the product is different from 55) the null hypothesis?

We do not trust this first dataset too much. We decide to buy more product (up to 100). The dataset is now:

50.83622,41.21380,55.57440,55.85820,52.36584,54.46815,53.10866,49.22832,47.71356,54.92581,43.77760,42.26579,52.20705,64.97459,50.48356,46.52843,46.58054,56.27121,53.69041,51.53339,50.07873,49.78456,40.96808,56.62398,54.97590,49.80877,55.61752,54.30354,49.57232,53.73923,52.62941,49.83051,49.59908,57.87264,55.68242,48.07090,53.54720,51.31824,58.30299,54.39016,47.51912,55.35950,46.19605,58.44408,42.53546,54.58022,49.53793,58.58451,45.56949,50.66142,59.58990,53.34549,57.56043,48.33241,53.87193,49.03302,50.56219,48.15191,48.37849,53.20369,55.20906,54.23983,43.57551,40.82990,50.86912,51.07071,59.89565,55.96597,52.53466,54.82528,54.38412,48.16330,45.61594,53.53410,50.19482,47.10355,55.73813,49.30194,46.78841,47.50694,54.19838,59.46523,50.70361,42.50933,61.03322,54.90206,50.11657,41.32436,50.84496,46.76024,49.75024,54.64130,61.13959,53.86465,48.31012,46.67530,55.30151,49.68629,49.73536,53.77160

- Repeat the same analyses as before, but with this new dataset. Which is the p-value now?
- How can it be such a strong deviation in p-values? We went from saying that H_0 is right to rejecting H_0 with a REALLY strong support.

In fact, the data was generated from a normal distribution with *mean = 53* and *standard deviation = 5*. From that example, we can gather some extra hints about the nature of a statistical test: we will have different power for rejecting the null hypothesis depending on *n*.

- For this particular example (where the distribution that generated the data is known), which is the number of items you should buy to be sure that you can reject the null hypothesis at alpha = 0.05?

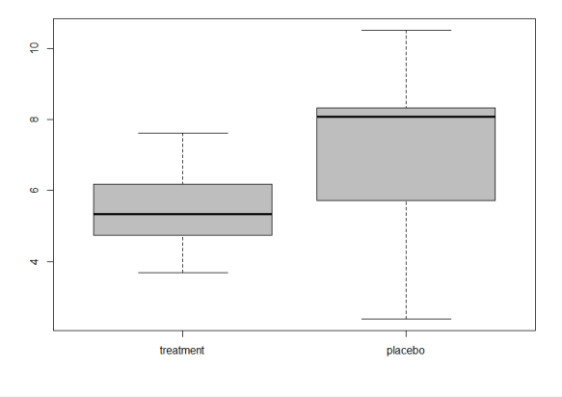
A point of caution must be stressed out at this level. Interpreting the biological meaning of a p-value (rejecting the null hypothesis) is not straightforward.

We can think in a much more complex situation. Imagine that we have a new compound that we think is going to reduce blood pressure. We divide our set of mice in two. We apply the compound to group 1 and a placebo to group 2.

- How would you formally describe the contrast of hypothesis that we want to conduct?

H0:

H1:



The Student's t-test can be modified to account for this new situation. There are several versions of the test depending on whether the sample size of each group is the same or not, whether the variance is the same between the two groups or whether the data is paired or not. Read https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4667138/pdf/kjae-68-540.pdf to see an overview of the different types of t-tests.

Imagine that the outcome from the two groups is the one from this table:

Treatment 1	Placebo
4.524110	2.610371
6.247990	8.401271
4.975337	5.987959
5.896319	8.209667
6.122732	8.032294
3.925834	8.088457
5.390725	8.072965
5.267268	7.284080
5.794629	5.440700
5.290023	4.843819
7.210325	8.065547
4.724606	8.946616
6.741565	8.217999
7.619907	6.339434
4.729632	10.507302
4.772660	2.394933
3.699615	8.239763
6.514663	9.761649
5.718323	8.844672
4.760585	3.555284

A first visualization using a boxplot suggests that the means are quite different.
However, just by eye we see that it looks like the variance is not the same among groups.

- Compute the variance of each group.
- Which test would you apply to see if there is homogeneity of variances among the two groups?

- Based on the outcome of that test, what would be the appropriate classical Student's t test that we would apply?

A more general type of Student's t test that accounts for equal or unequal sample size and equal or unequal variance is known as Welch's t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{s^2}_1}{n_1} + \frac{{s^2}_2}{n_2}}}$$

With degrees of freedom for the Student's t distribution determined as:

$$df = \frac{\left(\frac{{s^2}_1}{n_1} + \frac{{s^2}_2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

- If this test takes into account all the possible situations of a Student's t test, why the others are still used?
- Estimate the t statistic and the degrees of freedom using the Welch's t-test (using R, the command is *t.test(treatment,placebo)*). Is the difference statistically significant at $\alpha = 0.05$?
- Estimate the t statistic assuming equal variance (*t.test(treatment,placebo, var.equal=T)*. Which is the p-value that we get?
- In this case, which test is more conservative (i.e. tends to accept more the null hypothesis)?

Quiz

1. A t test is applied when
   a. We have count data from two experiments
   b. We have data that follows a Normal distribution
   c. We want to compare the means of two groups and we assume that the mean follows a Normal distribution
   d. When the variance of the groups is equal
   e. When the variance of the groups is not equal

2. There are different versions of a t test:
   a. Depending on whether the data is continuous or categorical
   b. Depending on whether the data is paired
   c. Depending on the sample size of the two groups
   d. Depending on whether we assume equal variances or not
   e. B, C and D

3. Increasing the sample size:
   a. Tends to provide more power for a test to reject the null hypothesis
   b. Is independent of the p-value
   c. Increases the power for acccepting the null hypothesis
   d. Implies conducting additional analyses
   e. Should be always conducted

4. If we get a statistically significant p-value we can
   a. Be sure that the differences we observe are not just by chance
   b. Be sure that the differences we observe have a biological meaning
   c. Be sure that there is an error of type one
   d. Be sure that there is an error of type two
   e. None of the above

5. Unequal variance in a t test
   a. Decreases the power for rejecting the null hypothesis
   b. Increases the power for rejecting the null hypothesis
   c. Is irrelevant for the t test
   d. Can be controlled by standardizing the samples
   e. None of the above

6. Error type II
   a. Is associated to the fact that we do not know the distribution under the null hypothesis
   b. Tends to increase when α threshold decreases
   c. Can be controlled by setting a more relaxed experiment design
   d. Is inverselly related to the power of the test
   e. b and d

7. A Student t distribution:
   a. Requires a parameter called degree of freedom.
   b. Is bell-shaped
   c. Models the difference of the sample mean and the parametric mean of the distribution that produced the data
   d. Goes from –infinite to infinite
   e. All of them

8. Pick the right answer.
   a. A one tail comparison implies looking to the upper tail of the distribution
   b. A one tail comparison implies looking at the lower tail of the distribution
   c. A two tail comparison implies looking at the lower tail and the upper tail of the distribution
   d. A two tail comparison implies multiplying by two the accumulative probability of the lower tail.

e. A two tail comparison implies multiplying by two the accumulative probability of the upper tail.

9. Which is true about these sentences?
   a. A Levene test computes the probability that the means of different groups are equal.
   b. A paired t test requires multiple observations for each individual.
   c. A paired t test compare two features in the same individuals
   d. A statistically significant Levene test implies that the variances are the same between groups
   e. Welch's t-test is a type of test for comparing variances among two groups.

10. A false positive:
   a. Implies accepting the null hypothesis when in fact it is true
   b. Implies accepting the alternative hypothesis when in fact it is true
   c. Implies rejecting the alternative hypothesis when in fact it is true
   d. Implies accepting the alternative hypothesis when in fact it is false
   e. None of the avobe.

Learning activities for Week 2. Hypothesis testing in medical practice

Short problem

Imagine that we want to see if the differences that we observe in the sample means estimated from different experiments are just by random sampling. If we are considering two groups (and under some assumptions) we can apply the Student's t test. However, when we are considering more than two groups, the Student's t test is not feasible anymore.

Please, before continuing read the article by Bewick, Cheek and Ball about ANOVA https://www.ncbi.nlm.nih.gov/pmc/articles/PMC420045/pdf/cc2836.pdf

- What is df?
- What is a sum of squares?
- What alternative contrasts of hypothesis can be conducted?

The basic idea of an ANOVA is that the variation that is observed in a dataset can be decomposed into independent factors. The ratio of these factors produces the F statistic, for which the probability distribution is known under the null hypothesis that all the means do not differ. The outcome of a classical generalized one-way ANOVA is a table that looks like this:

Source of variation	df	SS	MS	F statistic	p-value
Between groups	$df_a = g - 1$	$SS_a = \sum_{i=1}^{g} n_i(\overline{Y}_i - \overline{\overline{Y}})^2$	SS_a/df_a	MS_a/MS_w	
Within groups	$df_w = \sum_{i}^{g} n_i - g$	$SS_w = \sum_{i}^{g}\sum_{j}^{n}(y_{j,i} - \overline{Y}_i)^2$	SS_w/df_w		
Total	$\sum_{i}^{g} n_i - 1$	$SS_a + SS_w$			

Where *g* is the number of groups, n_i is the sample size of group *i*, $y_{j,i}$ is the value of individual *j* that belongs to group *i*, $\overline{Y}_i$ is the mean of group *i* and

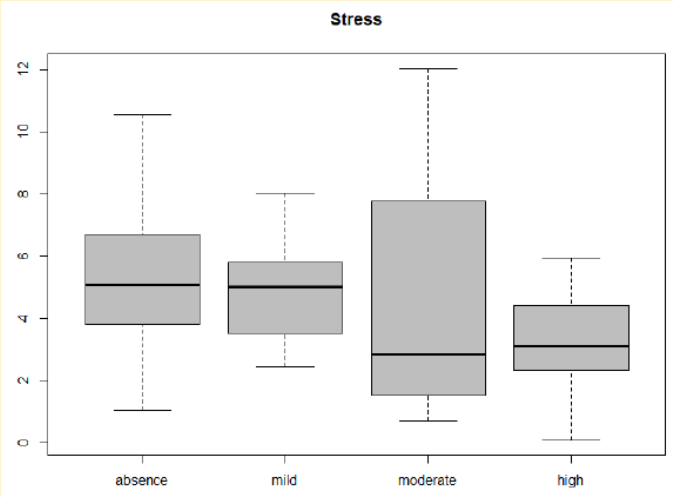
$$\overline{\overline{Y}} = \frac{1}{\sum_i^g n_i}\sum_{i}^{g}\sum_{j}^{n} y_{j,i}$$

Is the grand-mean, or mean independently of which group the individual belongs to. We are going to fill this matrix using the data from the next example.

It is thought that stress may increase susceptibility to illness through suppression of the immune system. In an experiment to investigate this theory, 48 rats were randomly allocated to four treatment groups: no stress, mild stress, moderate stress, and high stress. The stress conditions involved various amounts of restraint and electric shock. The concentration of lymphocytes (cells/ml × 10-6) in the peripheral blood was measured for each rat with the results given in the accompanying table.

Absence of stress	Mild stress	Moderate stress	High stress
5.979725	3.698131	6.6798732	2.6746460
4.078527	4.771929	0.7275278	4.7941649
10.579511	4.588594	8.9137657	1.1062816
3.551018	8.053512	2.2906609	5.9310999
5.636736	5.934523	1.0525576	4.0500735
8.898859	5.281532	1.9908044	3.2356990
1.061627	2.460336	0.9925206	0.1078552
7.439450	6.954691	5.0979815	2.1180197
5.746312	5.548003	3.4339823	2.5775835
4.198445	5.673175	2.2132957	5.9701717
3.398954	3.343923	8.9550667	4.0231196
4.526974	2.661931	12.0358831	2.9717078

A visualization of the data suggests that (*may be)* there are differences between groups:



- Estimate the different statistics that are needed to conduct this anova analysis.
- Which is the value of the F statistic?
- Which is the accumulative probability of observing a value identical or greater (you can use the command pf(MSamong/MSwithin, dfamong, dfwithin, lower.tail=F) function of R to estimate it.
- Is the difference statistically significant at □ = 0.05?

This result is quite strange, because "*it looks like*" there is something. Recall from the previous exercises using the *t* test that the outcome depends not only on the biological magnitude of the differences in the data, but also on how much power we have to detect such magnitude. Increasing the sample size can provide much more confidence about the outcome. For example, imagine that we increase the dataset to 100 individuals per group. The ANOVA table that we would get is now:

Source of variation	df	SS	MS	F statistic	p-value
Between groups	4	645			
Within groups	396	3529			

- Can you fill the missing values of the table?
- According to the F distribution with degrees of freedom df = (4, 396), the F statistic that you obtained is statistically significant different than what you would expect if the differences in mean were due to random sampling?
- How would you present the result?

However, it can also be that we are not taking properly some of the assumptions of the test. In the case of ANOVA, one of the assumptions is that the variance within each group is similar among groups.

If we conduct a Levene test for homogeneity of variances, we will obtain a p-value = 0.08828. This p-value is not statistically significant at the classical α = 0.05, but it is small enough to think that maybe it is blurring our analyses.

Learning activity

Once we have conducted an ANOVA analysis and we have seen that there are differences between the groups, it may be desirable to know the groups that are producing such statistical difference. Please, read https://www.biochemia-medica.com/assets/images/upload/xml_tif/McHugh_ML_-Multiple_comparison_analysis_testing_in_ANOVA.pdf

- Which are the methods available to run multiple comparison analysis testing in ANOVA?
- Which are the pros and cons of each method?

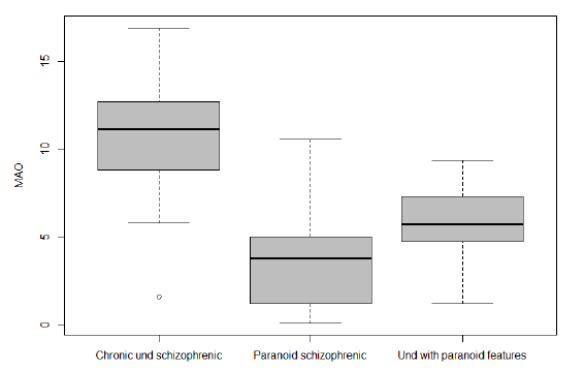
At this point, we need to introduce the concept of planned and unplanned comparisons. Planned comparisons are chosen independently of the results of the experiment and designed *before* conducting the experiment. In contrast, an unplanned comparison comes when, after conducting our test, we see that there are statistically significant differences. By looking at the data, we decide that we want to test whether a particular combination that "looks promising" in fact shows statistically significant differences. The statistical analysis that we apply to each scenario is different. In planned comparisons, as far as these comparisons are orthogonal (i.e. the data used at each comparison do not overlap), the classical ANOVA framework can be used. We can subdivide the sum of squares (SS) and degrees of freedom of the treatments into different boxes or sub-sections without any problem. In the case of unplanned comparisons, a multiple testing correction must be applied in order to prevent statistical inflation.

Consider the next exercise:

Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behaviour. To see whether different categories of schizophrenic patients have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are summarized in the accompanying table. (Values are expressed as nmol benzylaldehyde product/10^8 platelets/hour.)

Chronic undifferentiated schizophrenic	Undifferentiated with paranoid features	Paranoid schizophrenic
11.776811	5.992672	1.84728896
11.718530	1.167285	5.49766655
1.596569	7.572338	0.08741318
11.183117	4.900708	3.54326552
12.633564	8.235566	4.51750692
11.111401	4.708963	0.53349932
10.835040	7.032687	3.99086435
8.728872	4.556668	10.57902270
16.912729	5.414430	
12.695217	6.463861	
9.914097	8.304190	
5.778449	4.732464	
14.972806	4.824537	
14.018889	9.378637	
8.815316	6.576314	
6.527934	2.118641	
14.637435		
10.278222		

The boxplot looks like this:



- Fill the missing values of the ANOVA table:
- Fill the missing values of the ANOVA table:

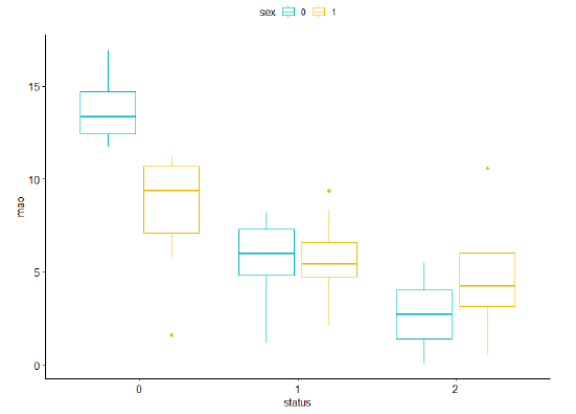
Source of variation	df	SS	MS	F statistic	p-value
Between groups		645	175.67	18.24	2.55e-06
Within groups		3529		NA	NA

- Can we conclude that there are differences in levels of MAO activity between the groups of patients?
- How would you know which is the patient status that is producing such differences between groups?

The ANOVA framework is extremely flexible and can be adapted to a large number of study designs. Two-way ANOVA is a type of ANOVA where we include interactions between variables explaining our observed continuous variable. For example, let us imagine that, in our MAO experiment, we have additional information regarding to the sex of the participants (that is, we have two categorical variables explaining the MAO expression: the patient status and the sex of the patient). An interesting question would be if there is an interaction between a trait and the fact of being from one sex or another. Our table could look like this:

Sex	Chronic undifferentiated schizophrenic	Undifferentiated with paranoid features	Paranoid schizophrenic
Male	11.776811	5.992672	1.84728896
	11.718530	1.167285	5.49766655
	12.695217	7.572338	0.08741318
	14.972806	4.900708	3.54326552
	12.633564	8.235566	
	14.637435	4.708963	
	16.912729	7.032687	
	14.018889		
Female	10.835040	5.414430	4.51750692
	8.728872	6.463861	0.53349932
	9.914097	8.304190	3.99086435
	5.778449	4.732464	10.57902270
	11.183117	4.824537	
	1.596569	9.378637	
	8.815316	6.576314	
	6.527934	2.118641	
	11.111401	4.556668	
	10.278222		

If we look at the boxplot controlling by sex, we observe that not all the patient categories show the same behaviour depending on sex:



In particular, patient *status 0*, corresponding to chronic undifferentiated schizophrenic, seems to show higher MAO activities for males than for females. The other two status seem to show a similar trend independently of the sex. That would suggest a putative interaction between

patient and sex for Chronic undifferentiated schizophrenic patients. The first temptation would be to run a t test within each patient status to see if there are sex differences in MAO activity.

- Why such strategy would be a very bad strategy?

Furthermore, there is no need to do such statistical tricks because the ANOVA design allows for interactions.

Variable	Df	Sum Squares	Mean Squares	F value	Pr(>F)
Between patients	2	351.3	175.67	25.687	1.17e-07
Between sex	1	31.7	31.66	4.630	0.03821
Status by sex	2	97.7	48.84	7.141	0.00244
Within	36	246.2	6.84		

- Can we conclude that there are statistically significant differences between patients?
- Can we conclude that there are statistically significant differences between sex?
- Is there any interaction between both variables?
- How would you explain in a single sentence all the results?

Quiz

1. Multiple p-value comparison implies:
    a. You run an ANOVA comparing multiple means.
    b. You run many planned comparisons in the data
    c. You run many unplanned comparisons in the data
    d. A mixture of p-values.
    e. None of the above

2. The F statistic:
    a. Is the ratio between the means of the groups and the grand-mean
    b. Is a ratio of variances between and within groups
    c. Requires more than two groups to work
    d. Requires unequal variances within groups
    e. None of the above

3. After conducting 10 tests on the same data, we obtain one with a p-value of 0.01. We conclude that
    a. Our test is statistically significant at $\alpha = 0.05$

b. Our test is statistically significant at $\alpha = 0.01$
c. A and B
d. There is a multiple testing problem and we must correct the p-value of each test by the total number of tests before reaching a conclusion.
e. A composite p-value would be significant.
f. None is correct

4. The goal of ANOVA is:
   a. Identify differences in the variance of the groups
   b. Identify differences in the medians of the groups
   c. Identify differences between the means of the groups
   d. Identify outliers
   e. All the above

5. One way-ANOVA
   a. Compares one group against the others, hence the "one way"
   b. Divides the variance present in the dataset into the compartments defined by a categorical variable
   c. Do not consider interactions between variables
   d. Requires multiple testing
   e. B and C

6. Interactions between two variables in an ANOVA design
   a. Can be modeled in a two-way ANOVA design
   b. Require multiple testing
   c. Also include testing each variable independently
   d. Can change the statistical significance of the overall statistical design
   e. A and C are true

7. Two way-ANOVA
   a. Models one categorical variable and two continuous variables.
   b. Models a continuous variable that depends in two categorical variables.
   c. Models a continuous variable that depends in two categorical variables as well as the interaction of these two variables.
   d. Increases the power for identifying associations between categories (hence the two way).
   e. None of the above.

8. The lack of power for an ANOVA design can be explained by:

a. Violation of the assumptions of ANOVA
b. Limited sample size
c. Two non-informative categorical variables
d. A and B
e. None of the above

9. An ANOVA framework:
   a. Only allows one way and two way ANOVA
   b. Is extremely flexible and accommodates many types of models and study designs.
   c. We need more stringent significance thresholds
   d. One of the major problems is the sum of squares within groups
   e. None of them are correct

10. Unplanned comparison means:
   a. Testing a comparison after we observed that there are statistically significant differences among all the groups.
   b. Running a t test among a pair of groups.
   c. Running an ANOVA between a pair of groups
   d. Testing a priori ascertained two groups.
   e. None of the above.

Forum activity

How would you design an experiment to check the performance of a new drug for decreasing blood pressure compared to other three drugs available (and a placebo group) in the market? Which categorical variables would you take into account when making the study design?

Which test(s) would you apply to test the hypothesis that your drug is better for decreasing the blood pressure?

Readings recommended for this week:

Steven F. Sawyer (2009) Analysis of Variance: The Fundamental Concepts, Journal of Manual & Manipulative Therapy, 17:2, 27E-38E, DOI: 10.1179/jmt.2009.17.2.27E

Learning activities for Week 3. Linear regression

Short problem

In renal insufficiency, adequate hemoglobin levels of 12-13 gr / dl are considered. Therefore, lower values indicate kidney failure problems. Urea is a residue of protein breakdown and is therefore directly related to the amount of protein we eat. Normally, the kidneys filter the urea from the blood, but when the kidneys do not work well, the amount of filtered urea is smaller and increases in the blood. The normal blood level is below 40 mg / dl. Serum creatinine is a residue of muscle mass and activity. Its blood level is the most objective and reliable data to know how the kidneys work. High creatinine values indicate kidney failure problems.

The provided database ("name") includes individuals with low levels on hemoglobin and contains information about their levels of creatinine (mg/dl) and ureic nitrogen (mg/dl).

1. Interpret the values of the hemoglobin.

```
> summary(hemoglobin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.200   8.925   9.350   9.730  10.630  11.700
> sd(hemoglobin,na.rm=TRUE)
[1] 1.245481
> IQR(hemoglobin,na.rm=TRUE)
[1] 1.7
> length(hemoglobin)
[1] 10
```

In terms of the values of hemoglobin, what is your conclusion?

2. Interpret the output of the correlation analysis:

```
> cor.test(UN, hemoglobin)

Pearson's product-moment correlation

data:  UN and hemoglobin
t = -2.5416, df = 8, p-value = 0.03462
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.91355903 -0.06693547
sample estimates:
       cor
-0.6683928
```

```
> cor.test(creatinine, hemoglobin)

Pearson's product-moment correlation

data:  creatinine and hemoglobin
t = -3.7193, df = 8, p-value = 0.005877
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9496644 -0.3334958
sample estimates:
      cor
-0.7959833
```

3. Given the estimates of the following models:

Model 0

```
Call:
lm(formula = hemoglobin ~ creatinine)

Residuals:
      Min       1Q   Median       3Q      Max
-1.22740 -0.46507  0.06049  0.62773  0.96817

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0964     0.6846  17.668 1.08e-07 ***
creatinine   -0.4889     0.1315  -3.719  0.00588 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7996 on 8 degrees of freedom
Multiple R-squared:  0.6336,    Adjusted R-squared:  0.5878
F-statistic: 13.83 on 1 and 8 DF,  p-value: 0.005877

> confint(m0, level=0.95)
                2.5 %     97.5 %
(Intercept) 10.517578 13.6751688
creatinine  -0.792053 -0.1857873
```

Model 1

```
Call:
lm(formula = hemoglobin ~ UN)

Residuals:
      Min       1Q   Median       3Q      Max
```

```
-1.33412 -0.65477 -0.04534  0.58679  1.39961

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.95771    1.30739   9.911 9.07e-06 ***
UN          -0.07055    0.02776  -2.542   0.0346 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9826 on 8 degrees of freedom
Multiple R-squared:  0.4467,    Adjusted R-squared:  0.3776
F-statistic:  6.46 on 1 and 8 DF,  p-value: 0.03462


> confint(m1, level=0.95)
                              2.5 %       97.5 %
(Intercept)  9.942867 15.972550773
UN          -0.134561 -0.006541031
```

Model 2

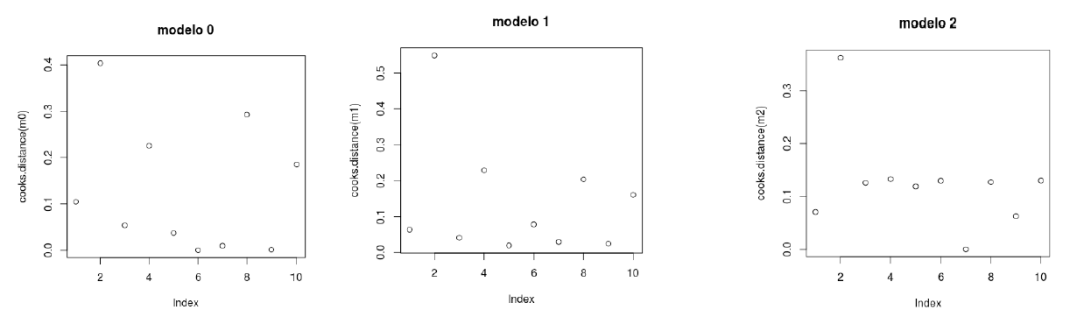
```
Call:
lm(formula = hemoglobin ~ creatinine + UN)

Residuals:
      Min      1Q  Median      3Q     Max
-0.9796 -0.6493  0.1868  0.6242  0.7760

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.65086    1.43189   7.438 0.000145 ***
creatinine  -0.90062    0.38254  -2.354 0.050766 .
UN           0.07515    0.06574   1.143 0.290549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7847 on 7 degrees of freedom
Multiple R-squared:  0.6912,    Adjusted R-squared:  0.603
F-statistic: 7.835 on 2 and 7 DF,  p-value: 0.01636

> confint(m2, level=0.95)
                                2.5 %       97.5 %
(Intercept)  7.26497581 14.036734590
creatinine  -1.80519236  0.003950796
UN          -0.08029552  0.230597137
```

3a. Write the model equation of each of the three estimated models indicating the response variable and the explanatory variables.

3b. Interpret the parameter of model 0

3c. Could you explain what differences you see between models 0 and 1 and model 2? What do you think it can indicate?

3d. What can you say about outliers in all three models? In terms of influential values, what seems to be the best model? Justify your answer.



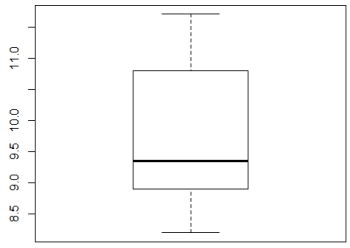
Solution

1. Interpret the values of the hemoglobin.

```
> summary(hemoglobin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.200   8.925   9.350   9.730  10.630  11.700
> sd(hemoglobin,na.rm=TRUE)
[1] 1.245481
> IQR(hemoglobin,na.rm=TRUE)
[1] 1.7
> length(hemoglobin)
[1] 10
```

We observe that there are only 10 values. In addition, the median and the mean seem to be quite similar, so the data would have to be quite symmetrical. However when drawing the boxplot we observe that data presents a positive assymetry. Also, looking at the boxplot we can discard outliers.

Moreover, 25% of individuals have a level on hemoglobin lower than 8.925 and another 25% have a level greater than 10.630. The maximum value is 11.7 and the minimum is 8.2.



In terms of the values of hemoglobin, what is your conclusion?

Analysing the above values we can say that 50% of individuals have a level of hemogloblin lower than 9.350. The other 50% have a higher dispersion taking values from 9.350 to 11.7.

2. Interpret the output of the correlation analysis:

> cor.test(UN, hemoglobin)

Pearson's product-moment correlation

data: UN and hemoglobin
t = -2.5416, df = 8, p-value = 0.03462
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.91355903 -0.06693547
sample estimates:
cor
-0.6683928

Ureic Nitrogen (UN) and hemoglobin are negatively related (cor=-0.668). In addition, looking at the p-value (0.03462<0.05) and assuming a risk level of 5% we can extrapolated the relationship shown in the sample to the entire population. Also, we could confirm this afirmation looking at the confidence interval (-0.91355903, -0.06693547) as it doesn't include the zero.

> cor.test(creatinine, hemoglobin)

Pearson's product-moment correlation

data: creatinine and hemoglobin
t = -3.7193, df = 8, p-value = 0.005877
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9496644 -0.3334958
sample estimates:
cor
-0.7959833

Creatinine and hemoglobin are also negatively related (cor=-0.796). Comparing with the previous case here the two variables are more related since the value of the correlation coefficient is closer to -1.
In addition, looking at the p-value (0.005877<0.05) and assuming a risk level of 5% we can also extrapolated the relationship shown in the sample to the entire population. Also, we could confirm this afirmation looking at the confidence interval (-0.9496644, -0.3334958) as it doesn't include the zero.

3. Given the estimates of the following models:

Model 0

```
Call:
lm(formula = hemoglobin ~ creatinine)

Residuals:
      Min       1Q   Median       3Q      Max
-1.22740 -0.46507  0.06049  0.62773  0.96817

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0964     0.6846  17.668 1.08e-07 ***
creatinine   -0.4889     0.1315  -3.719  0.00588 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7996 on 8 degrees of freedom
Multiple R-squared:  0.6336,    Adjusted R-squared:  0.5878
F-statistic: 13.83 on 1 and 8 DF,  p-value: 0.005877

> confint(m0, level=0.95)
                2.5 %     97.5 %
(Intercept) 10.517578 13.6751688
creatinine  -0.792053 -0.1857873
```

Model 1

```
Call:
lm(formula = hemoglobin ~ UN)

Residuals:
      Min       1Q   Median       3Q      Max
-1.33412 -0.65477 -0.04534  0.58679  1.39961

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.95771    1.30739   9.911 9.07e-06 ***
UN          -0.07055    0.02776  -2.542   0.0346 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9826 on 8 degrees of freedom
Multiple R-squared:  0.4467,    Adjusted R-squared:  0.3776
F-statistic:  6.46 on 1 and 8 DF,  p-value: 0.03462
```

```
> confint(m1, level=0.95)
                 2.5 %      97.5 %
(Intercept)  9.942867 15.972550773
UN          -0.134561 -0.006541031
```

Model 2

```
Call:
lm(formula = hemoglobin ~ creatinine + UN)

Residuals:
      Min      1Q  Median      3Q     Max
-0.9796 -0.6493  0.1868  0.6242  0.7760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.65086    1.43189   7.438 0.000145 ***
creatinine  -0.90062    0.38254  -2.354 0.050766 .
UN           0.07515    0.06574   1.143 0.290549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7847 on 7 degrees of freedom
Multiple R-squared:  0.6912,    Adjusted R-squared:  0.603
F-statistic: 7.835 on 2 and 7 DF,  p-value: 0.01636

> confint(m2, level=0.95)
                  2.5 %      97.5 %
(Intercept)  7.26497581 14.036734590
creatinine  -1.80519236  0.003950796
UN          -0.08029552  0.230597137
```

3a. Write the model equation of each of the three estimated models indicating the response variable and the explanatory variables.

Model 0: hemoglobin =β0+ β1creatinine
Model 1: hemoglobin =β0+ β1UN
Model 2: hemoglobin =β0+ β1creatinine+ β2UN

Explanatory variables are creatinine and UN
Response variable is hemoglobin

3b. Interpret the parameter of model 0

$?_1$=-0.07055.

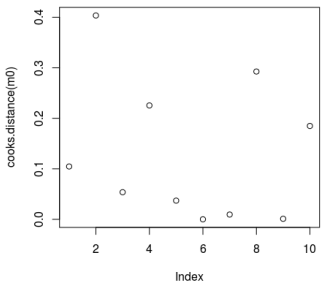
When the creatinine increases 1 unit (mg/dl), the hemoglobin decreases in 0.07055 units.

3c. Could you explain what differences you see between models 0 and 1 and model 2? What do you think it can indicate?

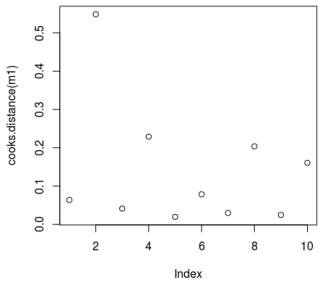
In model 0 and model 1 explanatory variables are statistically significant while when we incorporate both in the model (model 2) then neither is significant. In addition, in the first two models the effect of the explanatory variables is negative. However, in the second model, apart from the fact that variables are not significant, we see that we have a positive and a negative coefficient.
This could be due to the fact that the two explanatory variables are closely related and putting them together will mask the real effect. This is known as confusing variables.

3d. What can you say about outliers in all three models? In terms of influential values, what seems to be the best model? Justify your answer.

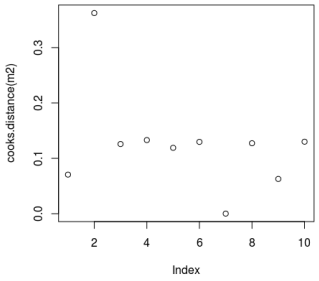
model 1



model 2



model 3



Looking at cook's distance the model with a higher value is model 2. Also model 1 could have an influential value as one of them reaches the value 0.4. Then, in terms of influential values the best model would be the third. However, a more accurate analysis should be done to better determine outliers and influential values.

Learning activity

This exercise aims to 1) Define and estimate a Multiple Linear Regression model, 2) Calculate and interpret the model parameters, 3) Interpret the confidence intervals of the model coefficients 4) Calculate and interpret the coefficient of adjusted determination 5) Validate the model

Then, we use a study that collects the values of systolic blood pressure (SBP), average tobacco consumption (TAB), age (in years) and sex (1 = man, 2 = woman) in a sample of 36 subjects. The objective of the work is to assess the hypothesis of the researchers that smoking could have an

effect on systolic blood pressure. (Source: Doménech JM. Correlación y regresión lineal. Barcelona: Signo; 2004). The data corresponding to this study can be found in the file tobacco_SBP.dat.

QUESTIONS

1) We want to estimate, first of all, the crude effect of tobacco consumption on SBP:

1.1) Write the model equation (model 1) indicating the variables you take as response variable and predictor variable and the model parameters.

1.2) Estimate the model equation using the statistical package RStudio.

1.3) Quantify what effect an increase of 10 cigarettes has on SBP.

1.4) Estimate the degree of association between the variables of the model and the variance of the residual error.

2) It is suspected that age could be associated with tobacco consumption and SBP, so it could be a confusing variable of the crude effect of tobacco on SBP. Consider the appropriate model (model 2) to assess this problem.

2.1) Write the estimates of the coefficients of model 2 and assess their goodness of fit.

2.2) Explain how the effect of tobacco on SBP has been modified.

3) Performs the diagnosis and validation of the linear regression model obtained in the previous section:

3.1) Perform all analyzes associated with the exploratory analysis (univariate and bivariate graphs, correlation analysis)

3.2) Verify if errors follow a normal distribution. What do you conclude?

3.3) Verify if the variance of the errors is constant through a graphic procedure. What do you conclude?

3.4) Verify if errors are independent to each other.

3.5) Check now the other conditions of application of the model: linearity, multicollinearity and existence of influential values.

3.6) What is the percentage of variation of Y that can be explained by the explanatory variables? Which of the two models is better in terms of the coefficient of determination?

Solution (in blue)

1) We want to estimate, first of all, the crude effect of tobacco consumption on SBP:

1.1) Write the model equation (model 1) indicating the variables you take as response variable and predictor variable and the model parameters.

SBP=β_0+$\beta_1$tobacco
SBP: response variable; Tobacco: predictor variable; β_0 and β_1 model parameters

1.2) Estimate the model equation using the statistical package RStudio.

```
> model1<-lm(SBP~TAB)
> summary(model1)

Call:
lm(formula = SBP ~ TAB)

Residuals:
    Min      1Q  Median      3Q     Max
-22.268 -10.136   1.385  10.139  30.139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 127.8609     3.3085  38.646  < 2e-16 ***
TAB           0.9102     0.1527   5.961 9.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 12.98 on 34 degrees of freedom
Multiple R-squared:  0.511,       Adjusted R-squared:  0.4966
F-statistic: 35.53 on 1 and 34 DF,  p-value: 9.688e-07
```

1.3) Quantify what effect an increase of 10 cigarettes has on SBP.

SBP=127.8609+0.9102*tobacco

An increase of 10 cigarettes has an effect of 9.102 on SBP

1.4) Estimate the proportion of the variance in the dependent variable that is predictable from the independent variable.

49.66%

2) It is suspected that age could be associated with tobacco consumption and SBP, so it could be a confusing variable of the crude effect of tobacco on SBP. Consider the appropriate model (model 2) to assess this problem.

2.1) Write the estimates of the coefficients of model 2 and assess their goodness of fit.

```
> model2<-lm(SBP~TAB+age)
> summary(model2)

Call:
```

```
lm(formula = SBP ~ TAB + age)

Residuals:
    Min      1Q  Median      3Q     Max
-23.001  -5.468   1.994   6.696  29.871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.3118     7.7160  13.130 1.18e-14 ***
TAB           0.4056     0.1887   2.149 0.039049 *
age           0.8129     0.2199   3.697 0.000787 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.08 on 33 degrees of freedom
Multiple R-squared:  0.6542,      Adjusted R-squared:  0.6333
F-statistic: 31.22 on 2 and 33 DF,  p-value: 2.455e-08
```

$\beta_{0:}$ 101.3118; $\beta_{1(TAB):}$ 0.4056; $\beta_{2(age):}$ 0.8129

The goodness of fit is given by R-squared. In this case the explanatory variables explain 65.42 % of the dependent variable (SBP). Even better, they explain 63.33% which this is the percent age given by the adjusted R-squared that is not affected by the degrees of freedom.

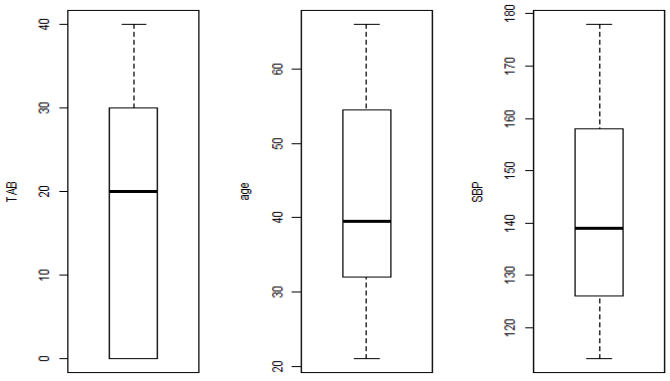
2.2) Explain how the effect of tobacco on SBP has been modified.

The effect of tobacco has decreased. In this model, an increase of 1 cigarrette implies an increase of 0.4056 on SBP instead of the 0.9102 seen in the first model.

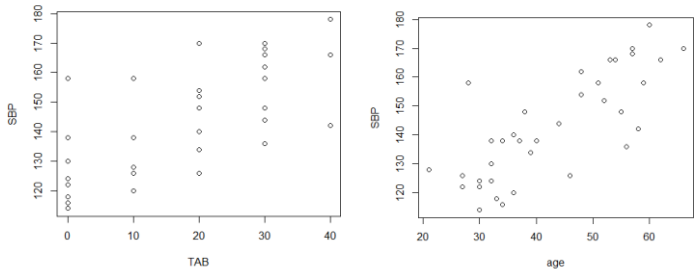
3) Performs the diagnosis and validation of the linear regression model obtained in the previous section:

3.1) Perform all analyzes associated with the exploratory analysis (univariate and bivariate graphs, correlation analysis)

Univariate graphs: We do not observe any influential values.



Bivariate graphs: we observe certain positive relationship between the explanatory variable and the response variable.

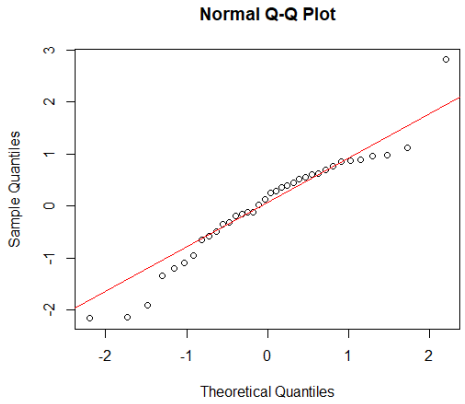



```
Correlation analysis
> cor.test(TAB, SBP)

         Pearson's product-moment correlation

data:  TAB and SBP
t = 5.9606, df = 34, p-value = 9.688e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5048731 0.8449405
sample estimates:
      cor
0.7148393

> cor.test(age, SBP)

         Pearson's product-moment correlation

data:  age and SBP
t = 7.2289, df = 34, p-value = 2.296e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6043609 0.8814789
sample estimates:
      cor
0.7783517
```

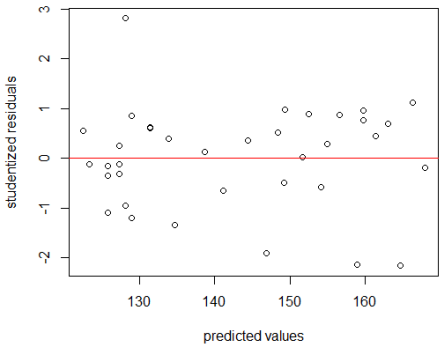
As we saw analysing the bivariate graphs, the correlation coefficient indicates a fairly strong positive correlation between the explanatory variables and the response variable.

3.2) Verify if errors follow a normal distribution. What do you conclude?

Errors do not follow a normal distribution strictly. However, since we have a sufficiently large sample size, we could assume normality in the data.

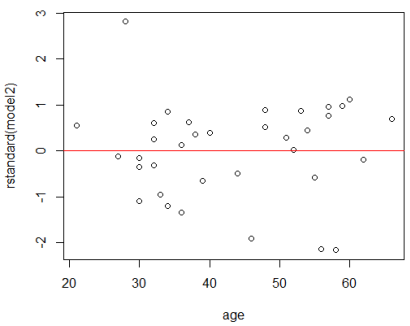


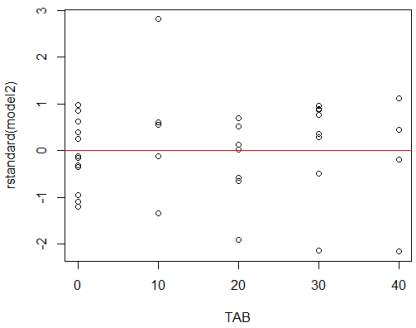
3.3) Verify if the variance of the errors is constant through a graphic procedure. What do you conclude?



We do not observe any pattern so we can assume that the variance of the errors is constant and so the model is homoscedastic.

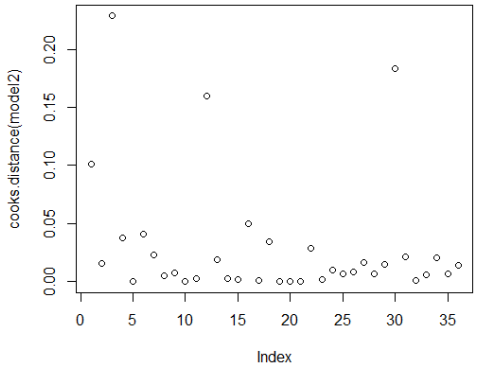
3.4) Verify if errors are independent to each other.





The points are distributed around the red line so we can assume independence between the errors.

3.5) Check now the other conditions of application of the model: linearity, multicollinearity and existence of influential values.




```
vif(model2)
     TAB      age
2.096511 2.096511
```

The Cook distance graph indicates that there are no influential values since we do not have points with very high values. Also the VIF value is lower than 10 so there is no multicollinearity between regressors.
Then, after performing the diagnosis and validation of the linear regression model and with the knowledge we have so far, we can say that model 3 is acceptable as long as we consider a sufficiently large sample. However, we should consider including other explanatory variables that will improve the goodness of the model fit.

3.6) What is the percentage of variation of Y that can be explained by the explanatory variables? Which of the two models is better in terms of the coefficient of determination?

To analyse the goodness of fit it is preferable to consider the adjusted R squared since it is not affected by the degrees of freedom. Then, in model 1 the adjusted R squared is 0.4966 and in the model 2 is 0.633. Therefore the model 2 is better than model 1 because it explains a higher percentage of variation of Y.

Quiz

We want to analyze data on the blood pressure of individuals in the EU collected during 2016. Blood pressure generally is higher in winter and lower in summer. That's because low temperatures cause your blood vessels to narrow — which increases blood pressure because more pressure is needed to force blood through your narrowed veins and arteries.
We are interested in studying the relationship between blood pressure (quantitative response variable "blood_pressure") and 4 explanatory variables: temperature ("temp" in F), age ("age"), number of pieces of fruit per week ("fruit_weekly "), and Body Mass Index (BMI)

In each of the following statements only one option is true. Indicate the correct option by consulting the Results Figures found at the end of this document:

1. The results of the coefficient of determination in Figure 1 (m1) express that:
   a. **18.8% of the percentage of variation in blood pressure can be attributed to the linear relationship with temperature.**
   b. Blood pressure accounts for 18.8% of the temperature variability.
   c. 0.18% of the percentage of temperature variation can be attributed to the linear relationship with blood pressure.
   d. 0.18% of the percentage of the variation in blood pressure can be attributed to the linear relationship with temperature.
   e. 18.8% of the percentage of temperature variation can be attributed to the linear relationship with blood pressure.

2. The models presented in Figure 1 allow us to state that:
   a. The linear association between blood pressure and each of the explanatory variables is inverse.
   b. Blood pressure increases with age.
   c. At a lower age, higher blood pressure.
   d. At a higher age, lower blood pressure.
   e. At a lower age equal blood pressure.

3. With the results of the model
Blood_pressure = $\beta 0 + \beta 1$ temp + ϵ,
which are presented in Figure 1, it can be stated that:
   a. An increase of 14 F in temperature increases, on average, 10 units the blood pressure.
   b. An increase of 14 F in temperature lowers, on average, 10 units blood pressure.
   c. An increase of 10 F in temperature lowers, on average, 14 units blood pressure.
   d. An increase of 1 F in temperature increases, on average, 14 units blood pressure.

e. An increase of 1 F in temperature increases, on average, 108 units blood pressure.

4. In terms of goodness of fit and model validity, looking at Figure 1 and according to the coefficient of determination, what is the model that has a higher percentage variation in Y explained by the explanatory variables?

a. Model 1.
b. Model 2.
c. Model 3.
d. Model 4.
e. All the models explain the same percentage of variation in Y.

5. The multiple linear regression model
Blood_pressure = β0 + β1 temp + β2 age + β3 fruit_weekly + β4 BMI + Є
(m5) which is presented in Figure 2:

a. It was not globally statistically significant (F = 12.16, p <0.001).
b. It was not globally statistically significant because some of the variables in the model were not significant (p = 0.123).
c. It was globally statistically significant (F = 20.27, p <0.001).
d. The significance of the model cannot be assessed with the results presented.
e. It was globally statistically significant (F = 12.16, p <0.001).

6. The multiple linear regression model
Blood_pressure = β0 + β1 temp + β2 age + β3 fruit_weekly + β4 BMI + Є
(m5) which is presented in Figure 2:

a. It shows only two variables statistically significant at 5% risk.
b. It shows only two variables statistically significant at 1% risk.
c. It shows only one variable statistically significant at 10% risk.
d. The model is not globally significant.
e. Assuming a 5% risk we can interpret only one variable

7. A diagnosis of the m5 is shown in Figure 3 (Blood_pressure β0 + β1 temp + β2 age + β3 fruit_weekly + β4 BMI + Є) in regards to some of the hypotheses of the multiple linear regression model? What can we conclude?

a. The normal probabilistic graphs of the studentized residuals verify the assumption of normality
b. Homocedasticity hypothesis is fulfilled.
c. Residuals are distributed around zero
d. There is no correlation between the dependent variable and the regressors
e. Residuals have clear outliers.

8. What is the Pearson correlation coefficient?

a. It is an index that takes all values less than 1.
b. It is an index that takes values between -1 and 0.
c. It is an index that takes values greater than 1.
d. It is an index that can take any value.
e. It is an index that can only take values between -1 and 1.

9. A certain interval of confidence will increase if:

a. Increase the size of the sample.
b. We increase the level of risk.
c. Increase accuracy.
d. Increase the level of confidence.
e. We reduce the margin of error.

10. In a study with the 4 examples of data represented in Figure 4, Pearson's correlation coefficient is reported to be 0.816, the 95% CI (0.424, 0.951), with a significance level of p = 0.002, in all cases. In view of these results you conclude that:

a. Reported results may not be correct.
b. There is only a linear relationship between the variables of the 4 examples.
c. It was not correct to calculate Pearson's linear correlation coefficient in any of the cases.
d. It was correct to calculate Pearson's linear correlation coefficient in all cases.
e. This is a very significant result that indicates that 81% of the variation of each variable is linearly explained by another variable.

Figures for the quizz

Figure 1. Simple linear regression models with response variable: blood pressure and explanatory variable: temp (m1), age (m2), weekly fruit consumption (m3) and BMI (m4).

```
> m1<-lm(blood_pressure~temp)
> summary(m1)

Call:
lm(formula = blood_pressure ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max
-31.248 -11.830  -3.305   4.456  72.680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.5711    26.3437   4.121  0.00019 ***
temp         -1.4081     0.4686  -3.005  0.00462 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.42 on 39 degrees of freedom
Multiple R-squared:  0.188,      Adjusted R-squared:  0.1672
F-statistic:  9.03 on 1 and 39 DF,  p-value: 0.004624
```

```
> m2<-lm(blood_pressure~age)
> summary(m2)

Call:
lm(formula = blood_pressure ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-26.976 -12.968  -3.495   6.710  67.177

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.610574   3.691587   4.770 2.58e-05 ***
age          0.026859   0.005099   5.268 5.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.17 on 39 degrees of freedom
Multiple R-squared:  0.4157,     Adjusted R-squared:  0.4007
F-statistic: 27.75 on 1 and 39 DF,  p-value: 5.363e-06

> m3<-lm(blood_pressure~fruit_weekly)
> summary(m3)

Call:
lm(formula = blood_pressure ~ fruit_weekly)

Residuals:
    Min      1Q  Median      3Q     Max
-27.114 -16.914  -2.847   5.531  78.464

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.357     25.009   0.614    0.543
fruit_weekly    1.556      2.619   0.594    0.556

Residual standard error: 23.66 on 39 degrees of freedom
Multiple R-squared:  0.008966,   Adjusted R-squared:  -0.01644
F-statistic: 0.3528 on 1 and 39 DF,  p-value: 0.5559

> m4<-lm(blood_pressure~BMI)
> summary(m4)

Call:
lm(formula = blood_pressure ~ BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-36.098 -12.499  -4.408   5.919  77.301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.2270    15.3991  -0.469   0.6415
BMI           0.3273     0.1318   2.484   0.0174 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.09 on 39 degrees of freedom
Multiple R-squared:  0.1366,     Adjusted R-squared:  0.1144
F-statistic: 6.169 on 1 and 39 DF,  p-value: 0.0174
```

Figure 2. Multiple linear regression models with response variable: blood pressure and explanatory: temp, age, fruit_weekly and BMI (m5)

```
> m5<-lm(blood_pressure~temp+age+fruit_weekly+BMI)
> summary(m5)

Call:
lm(formula = blood_pressure ~ temp + age + fruit_weekly + BMI)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-25.443  -9.321  -1.161   3.647  63.011

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.668675  38.504163   2.147   0.0386 *
temp         -0.979324   0.413323  -2.369   0.0233 *
age           0.025273   0.004701   5.376 4.74e-06 ***
fruit_weekly -3.062354   1.940282  -1.578   0.1232
BMI           0.168628   0.106799   1.579   0.1231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.14 on 36 degrees of freedom
Multiple R-squared:  0.5747,     Adjusted R-squared:  0.5274
F-statistic: 12.16 on 4 and 36 DF,  p-value: 2.354e-06
```

Figure 3. Validation of the multiple regression model m5

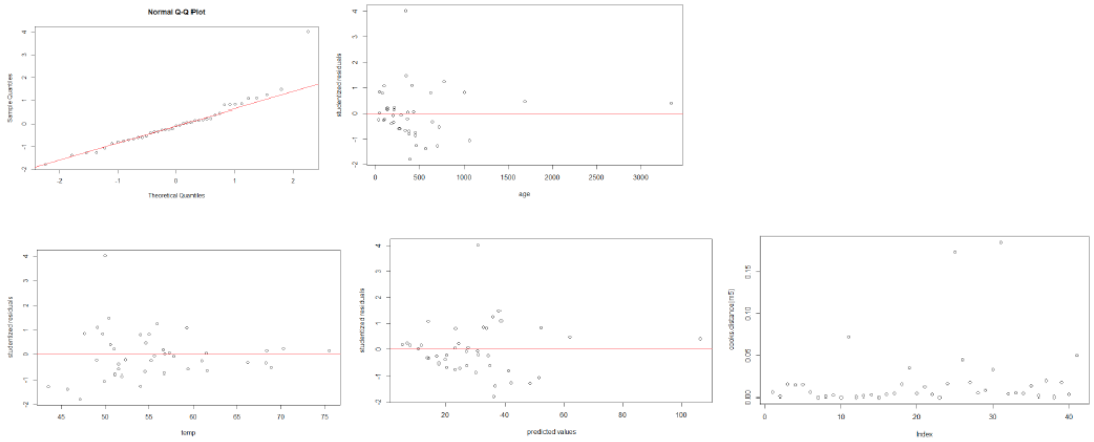
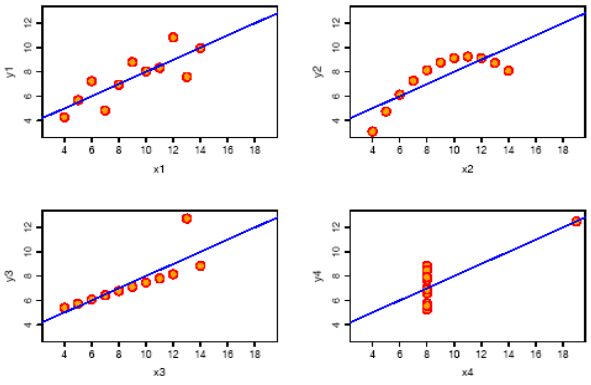


Figure 4. Anscombe data



Readings recommended for this week:

Benítez Brito N, Suárez Llanos JP, Fuentes Ferrer M, Oliva García JG, Delgado Brito I, Pereyra-García Castro F, et al. (2016) Relationship between Mid-Upper Arm Circumference and Body Mass Index in Inpatients. PLoS ONE 11(8): e016.

Bewick, V., Cheek, L. & Ball, J. Statistics review 7: Correlation and regression. *Crit Care* **7,** 451 (2003) doi:10.1186/cc2401.

Aggarwal, R., & Ranganathan, P. (2017). Common pitfalls in statistical analysis: Linear regression analysis. *Perspectives in clinical research*, *8*(2), 100–102. doi:10.4103/2229-3485.203040.

Aggarwal, R., & Ranganathan, P. (2016). Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in clinical research*, *7*(4), 187–190. doi:10.4103/2229-3485.192046.

Learning activities for Week 4. Categorical Data

Short problem

An association study is an observational study of a set of genetic variants in different individuals to see if any variant is associated with a trait. The following table presents the results of an association study for one genetic variant and a disease (M and N refer to two alleles)

	MM	MN	NN	
Cases	10	19	22	51
Controls	17	22	6	45
	27	41	28	96

1. A risk allele is the genetic variant increasing the probability of having the disease in comparison to the normal allele. In our example, which allele seems to be the risk allele?
2. In an association study, researchers are able to count how many individuals carry a given genotype. In our example 17 individuals from the control group are homozygote MM (carrying 2 alleles M each of them)
    a. How many alleles M do we have in control group?
    b. Transform the table of genotypes in a contingency table for the alleles. Calculate the relative risk associated with this allele.

3. Calculate the odds ratio associated with this allele and Interpret the OR in this context
4. Calculate the confidence interval of the OR associated with the risk allele at 95%, is this OR significant?
5. Perform the tests of association between the disease and alleles
6. Perform the tests of association between the disease and genotype (Genotypic association)

Solution

1. Allele N frequency increase in cases. Allele looks like the risk allele

2a. Homozygote MM carry 34 alleles M (they are diploid) and heterozygote carry 22 alleles M. The total number of allele M in control group is 56 alleles.

2b.

	Allele M	Allele N	
case	39	63	102
control	56	34	90
	95	97	192

The relative risk (RR) can be calculated according to the formula

$$RR = \frac{A/(A+B)}{C/(C+D)} = \frac{63/(63+34)}{39/(39+56)} = \frac{63/97}{34/95} = \frac{0.65}{0.36} = 1.81$$

3. Calculate the odds ratio associated with this allele

$$OR = \frac{A/B}{C/D} = \frac{63/34}{39/56} = \frac{1.85}{0.70} = 2.64$$

Interpret the OR in this context

According to this data, carrying allele N increase the risk of having the disease by a factor of 2.64 in relation to those who carry allele M.

4. Confidence intervals are calculated using the formula shown below

$SE(OR) = \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}} = \sqrt{\frac{1}{39}+\frac{1}{63}+\frac{1}{56}+\frac{1}{34}}$ =0.298

$$CI_{OR}^{95\%} = e^{\ln(OR)\pm 1.96\times SE(OR)}$$

Lower 95% CI is the value 1.47 and the upper 95% CI is the value 4.73

Since the 95% CI of 1.47 to 4.73 does not span 1.0, the increased odds (OR of 2.64) of having the disease do reach statistical significance.

5. To perform a statistical test, we should realize we have 2 categorical variables and the goal is to look for possible association between both. The appropriate statistical test is then the Chi-squared test
The null and alternative hypotheses are

H_0: *There is no association between the disease and the genetic variant*
H_1: *There is an association between the disease and the genetic variant*

Expected values can be calculated using the formula of expected values of 2 independent events is the product of their individual probabilities. The results will be the following

	Allele M	Allele NB	
Cases	50.47	51.53	102
Controls	44.53	45.47	90
	95	97	192

The Chi-squared value is $X^2 = 2.61 + 2.55 + 2.95 + 2.89 = 11.01$

There is 1 df in this test and the critical value looking to a Chi-square table is 3.84. Then the p-value is small then 0.05 and we can reject the null hypothesis. We have enough evidence an association exist between the disease and the genetic variation in the considered locus.

6. In a similar way then before:
The null and alternative hypotheses are

H_0: *There is no association between the disease and the genetic variant*
H_1: *There is an association between the disease and the genetic variant*

The expected values are

	MM	MN	NN	
Cases	14.34	21.78	14.88	51
Controls	12.66	19.22	13.13	45
	27	41	28	96

The Chi-squared value is $X^2 = 1.32 + 0.36 + 3.41 + 1.49 + 0.40 + 3.87 = 10.84$
There are 2 df in this test ((Raw -1)x(Columns –1)) and the critical value looking to a Chi-square table is 5.99. Then the p-value is small then 0.05 and we can reject the null hypothesis. We have enough evidence an association exist between the disease and the genetic variation in the considered locus.

Learning activity

A pilot project explores the possible relationship between a disease and a mutation. Researchers were able to recruit 33 volunteers. From a group of 15 people carrying a certain mutation, 9 have a disease. In another group of 18 people (without the mutation), only 5 have the disease. Our question would be: Are the people carrying the mutation more likely to have the disease?

We have 2 variables, the variable disease may be considered as a dependent variable. It can take only 2 possibilities, sick or healthy. The variable "Carrier" can also take 2 possibilities and may be considered as independent variable. The appropriate statistical test will be a chi-squared test.

Let's write first the information in R as a contingency table. To do so we use the built in function matrix. The names of rows and columns could be stated using the argument dimnames

```
patients_info<- matrix(c(9, 6, 5, 13), nrow=2, ncol=2, byrow=TRUE, dimnames = list(c("Carrier", "NotCarrier"), c("sick", "Healthy")))
```

Look to the table typing the name of the object "patients_info"

```
patients_info
```

To apply the chi-square test, is enough to call the function “chisq.test”

chisq.test(patients_info)

Interpret the result. *Is there any association between the disease and the mutation? Could you know if this OR is significant by looking only to the confidence interval? What is the effect size of this experiment? Do you think is worth to repeat the experiment with more sample size?*

Solution

The test statistic is 2.28 and the probability associated to this values according to chi-square distribution with 1 degree of freedom is 0.1307. This probability is high then 5% and those a similar result or more extreme is likely to be observed by chance. No evidence against the null hypothesis is obtained using this data.

Be careful! A correct use of a chi-squared test needs all the expected values in each cell to be above 5. If one or more of the cells are below, then you must use the **Fisher's exact test**

To check that, just run:
chisq.test(patients_info)$expected

The result of this command shows all expected values are above 5; even though we can run a Fisher exact test if we wish. To do so we call the function fisher.test

fisher.test(patients_info)

The result is the following
Fisher's Exact Test for Count Data

data: patients_info
p-value = 0.08527
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7372816 21.6160354
sample estimates:
odds ratio
 3.729192

You can see the exact p-values obtained is quite the same than the one obtained using the fisher exact test. The conclusion is the same.

The default result of the fisher exact test gives the estimated OR and the associated 95% confidence interval.
The OR is 3.73 which can be interpreted as the mutation increase the risk of having the disease by a factor of 3.73.

Could you know if this OR is significant by looking only to the confidence interval?
Yes, given the confidence interval contains the value 1, it shows the OR is statistically not different than 1. The value 1 is the expected value according to the null hypothesis of no association.

What is the effect size of this experiment?
The effect size is measure of the effect of the independent variable on the dependent one; here the OR of 3.73 is the effect size

Do you think is worth to repeat the experiment with more sample size?
Yes, in this situation, is highly recommended to repeat the experiment with more sample size. Effect size is usually stable when we repeat experiments and the p-value will be smaller with higher sample size.

Quiz

1. In the chi-square test for independence, the null hypothesis states that…
   a. there is no difference between the two variables being examined.
   b. there is no relationship between the two variables being examined.
   c. there is a difference between the two variables
   d. there is a relationship between the two variables being examined.
   e. All the above

2. Which of the statements is correct?
   a. The chi-square test compares the observed frequency distribution of the data to the frequency distribution that would be expected if the null hypothesis were true
   b. The chi-square test compares the observed frequency distribution of the data to the frequency distribution that would be expected if the null hypothesis were false
   c. The chi-square test compares the expected frequency distribution of the data to the frequency distribution that would be observed if the null hypothesis were true
   d. The chi-square test compares the expected frequency distribution of the data to the frequency distribution that would be observed if the null hypothesis were false

e. None of the above

3. We would like to know how a population of patients is distributed across categories. To do so we calculated the expected numbers of each category according to a given model. A chi-square test for goodness of fit is used to evaluate if observed number in each category follows expectation.What is the value of df for a test with five categories and a sample of n = 150?

a. 5
b. 4
c. 149
d. 150
e. 600

4. Two variables are classified into categories. A chi-square test for independence is used to evaluate the relationship between the two variables. If one variable is classified into 3 categories and the other variable is classified into 2 categories, then the chi-square statistic will have the following numbers of degrees of freedom

a. 6
b. 5
c. 4
d. 2
e. The total sample size is needed to determine the value of degrees of freedom

5. The sample data for a chi-square test are called…

a. observed proportions
b. observed frequencies
c. expected proportions
d. expected frequencies
e. chi-square values

6. For the following statements, which one may be related to the effect size concept

a. The variable under study
b. The sample size
c. The p value
d. The test statistic value
e. The magnitude of the relationship between variables

7. In an association study the aim is to explore possible relationship between a disease and genetic loci. The Odd Ratio (OR) is used to estimate the effect size of the genetic variant on the disease. If the relationship between the disease and the loci is statistically significant, the OR is…

a. different from 0
b. different from 1
c. higher than 1
d. higher than 0
e. Will be 1

8. A medical doctor is informed the result of clinical study were exciting. Researcher were able to show the drug is effective with high significance (p-value < 0.00001). The medical doctor should interpret

a. The result is statistically significant and of important practical relevance
b. The probability of the drug being not effective is unlikely by chance
c. The effect of the drug is big and is unlikely by chance
d. The effect of the drug may be big or small, and is unlikely by chance
e. None of the above

9. Comparing two groups (treated by a drug and placebo) with a t-student test, we get a significant result (p value less than 0.001). Given the mean of treated groups is 12.5; the mean the control is 10; the standard deviation of each of the groups is 10, and the calculated t statistic is 5.6; the effect size is

a. 5.6
b. 2.5
c. 0.79
d. 0.25
e. We need the sample size to calculate the effect size

10. In a clinical trial, if the null hypothesis is true (a treatment is not effective for instance), then in these conditions,

a. Increasing the sample size increase the statistical power
b. Increasing the sample size decrease the statistical power
c. The probability to reject the null hypothesis is not function of the sample size
d. The power of the test is higher with high sample size
e. The power of the test is lower with high sample size

11. In a clinical trial, if the null hypothesis is true (a treatment is not effective for instance), then

a. If the null hypothesis is true, increasing sample size decreases false positives
b. If the null hypothesis is true, increasing sample size decreases false negatives
c. If the null hypothesis is true, increasing sample size decreases false positives and false negative

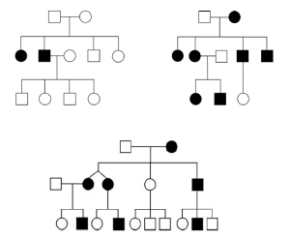
d. If the null hypothesis is true, increasing sample size increases false positives and false negative
e. **If the null hypothesis is true, false positive and False negative rates are not affected by sample size**

MODULE 3. PERSONALIZED GENOMICS IN PATIENT CARE

Learning activities for Week 1. Inheritance Model of Diseases

Short problem

Analyze the following pedigrees and explain the type of inheritance that you think is the most likely in each case. Explain why you discard the other inheritance models.



Solution

Upper left. Autosomal recessive. Both males and females are affected, discarding sex linked inheritance. The disease is present only in the second generation and is not transmitted to the third generation by an affected homozygous individual, discarding dominant inheritance.

Upper right. MItochondrial. All the daughters of an affected woman have the disease, whereas affected men do not transmit the disease. The presence in the three generations is not compatible with a recessive inheritance, and the fact that 100 % of the offspring of affected woman have the disease is far from the expected 50 % in case of autosomal dominant transmission.

Below. Autosomal dominant, with equal proportion of affected males and females, presence in the three generations and inherited by about 50 % of the offspring of affected individuals.

Learning activity

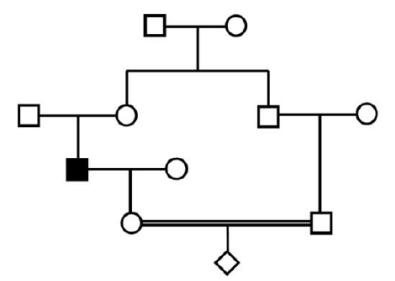
1. A family comes to your hospital for genetic counseling because their two sons (Max and Mark) present muscle weakness. The parents (Peter and Mary) are healthy and do not report muscle problems. Peter has a brother and two sisters, they are all healthy. The brother has two healthy daughters. The older sister has a healthy son and the younger sister has no children. Paternal grandparents do not refer medical problems either.

Mary has a brother and a sister, and they have two sons each. Mary's brother has muscle weakness. However, their two children are healthy. Mary's sister is healthy but it seems that one of her children begins to present symptoms similar to Mark and Max. When asked about their extended family, Mary tells you that her maternal uncle died at age 34 due to neurological problems and that he had been diagnosed with multiple sclerosis. Mary's uncle had two children, male and female, both healthy. Mary and Peter deny consanguinity in the family and do not refer any other history of neurological or muscular problems running in the family.

a) Draw the pedigree and discuss which type of inheritance would be compatible with the family history and why. Explain also why you discard the other types of inheritance.

b) If the family history (uncle with muscle weakness, cousin with muscle weakness) came from the paternal branch, would you propose the same type of inheritance? Why?

2. Next, we show a pedigree. Write a complete family history compatible with it



Solution

1.

a) The disease is compatible with X-linked recessive inheritance. It affects only males and is transmitted by heterozygous (healthy carriers) women in Mary's family. In cannot be dominant since both parents are healthy.

b) No, since Peter is not affected he cannot be a healthy carrier.

2. The pedigree is compatible with autosomal recessive inheritance, with an affected individual from two healthy parents. He does not transmit the disease to the next generation, since his woman is not carrier, as most frequently happens with rare Mendelian disorders.

Quiz

1. In an autosomal dominant inherited disease, only 80% of mutation carriers develop the disease, and of these, only a small percentage has the most severe symptoms. The gene responsible for this disease presents:

a. incomplete penetrance because of X-chromosome
b. full penetrance and incomplete expressivity
c. variable penetrance and incomplete expressivity
d. incomplete penetrance and variable expressivity
e. Modifications probably due to mitochondrial factors

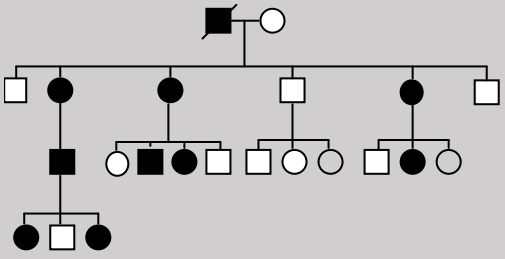
2. When providing the reproductive counseling in Marfan syndrome, all of the following points must be taken into account, except:

a. The risk of transmission to the offspring for an affected subject is 50%
b. Affected women have health risks during pregnancy, which can be often predicted
c. The phenotype is a bit variable in different families but also between relatives from the same family
d. Preimplantation genetic diagnosis is an option to prevent transmission
e. In de novo mutations, the recurrence risk for future pregnancies is almost 15%

3. A woman has a disease caused by a homoplasmic mutation in the mitochondrial genome. Which proportion of her offspring can suffer from the disease?

a. 100% of her children
b. 50% of her sons
c. 25% of her children
d. 50% of her children
e. 50% of her daughters

4. In the family tree depicted below, there are individuals affected by Vitamin D resistant rickets. Which is the inheritance of this disease?



a. autosomal dominant inheritance with variable penetrance
b. Mitochondrial inheritance
c. Genetic heterogeneity

d. Dominant X-linked inheritance

e. Recessive X-linked inheritance

5. If you detect a homozygous mutation in an autosomal recessive disease:

a. The clinical manifestations will depend on the gender of the proband (the individual with the mutation)

b. We should confirm the carrier status of the parents and counsel relatives from both branches of the family

c. The healthy siblings of the proband have a risk of ½ to be carriers of the mutation

d. The patient must be an adult

e. The parents must be consanguineous

6. What should be offered to a woman carrying a F508 mutation (cystic fibrosis)?

a. Genetic counseling and the study of her partner

b. Nothing since she is only a carrier and not an affected subject

c. A prenatal diagnosis for all pregnancies if desired

d. Preventive treatment for the disease

e. Nothing as she is a woman and this disease is X-linked

7. Which of the following statements is true regarding the phenomenon of imprinting?

a. It is an abnormal phenomenon always associated with disease

b. There is no known syndrome that can be caused by imprinting abnormalities in two different chromosomes

c. The transient neonatal diabetes can be due to imprinting mutations in chromosome 6

d. "Cri du Chat" syndrome is due to imprinting disorders in 5p

e. Silver-Russell syndrome is not caused by imprinting disorders

8. Which type of inheritance is the most likely in this pedigree?

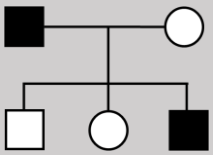
a. Mitochondrial inheritance

b. Y-linked inheritance

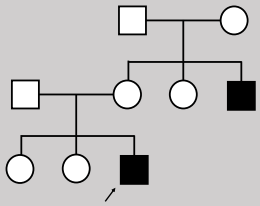
c. Autosomal recessive inheritance

d. Dominant X-linked inheritance

e. Recessive X-linked inheritance

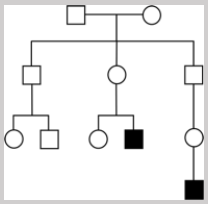


9. Considering the family tree shown below, corresponding to a recessive disease linked to the X chromosome, what is the risk of suffering the disease for the male offspring of the individual marked with the arrow?



a. 50 %.
b. 0 %.
c. 100 %.
d. 25 %.
e. 12 %.

10. Indicate which type (s) of inheritance could be discarded considering the family tree shown below:



1. Autosomal dominant inheritance with almost complete penetrance
2. Recessive X-linked inheritance
3. Dominant X-linked inheritance
4. Autosomal dominant inheritance in an imprinted locus

a. 1, 2 and 3
b. 1 and 3
c. 2 and 4
d. 4
e. All them

Learning activities for Week 2. Human Genetics sources of variability

Short problem

Genetic variants can be classified into different types according to their size, cell of origin, frequency, genome position or effect on the protein. In this exercise we will become familiar with the main types of genetic variants. We will also practice using two genome browsers.

In Genomics we call genetic variant any difference in the DNA sequence of an individual compared to the reference genome.

2. This is the sequence of a patient in a 10 nucleotides region (positions 103,271,300 to 103,271,310 in chromosome twelve in hg19): *CTCTTGAATGG*

   Use a genome browser (for example https://genome.ucsc.edu/index.html) to visualize this region (**WARNING**, always verify that you are using the same genome assembly, in this case hg19)

Get this sequence from the reference genome, and by comparing the two sequences determine if there is any difference between them. In case there is a difference (that is a genetic variant), determine if this is:

a. A single-nucleotide variant (SNV)
b. A short insertion or deletion (indel)
c. Structural change

3. This is the sequence of the patient in a different region (positions 10,000,040 to 10,000,050 in chromosome seven, hg19): *TTTAATATCTA*. Do you see any mutation following the same procedure as in the first question? What type of mutation is in this case?
4. Are the two genetic variants above located in a gene? (is this gene related to any genetic disease?)
5. Can you say if the variant in chromosome 19 is located in an exon or in an intron? (zoom out in the browser)
6. Looking at the information in the genome browser, can you determine if this gene is coded in the negative or the positive strand? (type the name of gene names in the browser and zoom in the coding regions).
7. Look at the enclosed file. We have the whole sequence of *CYP21A2*, a gene responsible for congenital adrenal hyperplasia, for two patients.

   To see if the patient harbors any mutation in this gene, compare it to the reference sequence. For this, download the whole gene sequence from Ensembl (https://www.ensembl.org/index.html). Once in the gene page, click in the coordinates and then Export Data. Save the sequence in a text file, that will include both the exons and the gene introns (if you export data directly from the gene page you will get the different transcripts) (SCREEN VIDEO SHOWING HOW TO DOWNLOAD THE SEQUENCE).

   Use an online alignment tool for comparing the sequence from each patient to the reference (for example MUSCLE at https://www.ebi.ac.uk/Tools/msa/muscle/)

   Is the mutation(s) described in each patient homozygous or heterozygous?

   Can you determine the type of mutation in A and B patients?

   a. Insertion
   b. Deletion
   c. Translocation
   d. Inversion

7. Can you predict the effect on the protein of these mutations? You can take the nucleotides encompassing your mutation and use the Blat tool in the UCSC Genome browser to obtain the

location of the sequences. By zooming in the gene region you can visualize the nucleotide and amino acid sequences (VIDEO).

Solution

1. A single-nucleotide variant (SNV)

2. An indel, since the patients's sequence lacks one nucleotide

3. The first region is located in the PAH region . You can check in the OMIM database that these gene is related to phenylketonuria.The second region is located in a intergenic region (zoom out to see genes around).

4. Is located in an exon, you can see the translation into amino. If you zoom out you will see the exons and introns of the gene.

5. By zooming out, you will see that this is gene is coded in the negative strand (the direction of the arrows in the introns indicates to the left)

6. Patient A presents a deletion of 8 nucleotides and Patient B has a single nucleotide variant (T to A).

7. The deletion in patient A is a frameshift indel. By deleting the two last nucleotides of the GGA codon (Glycine) and two other codons (GAC, Asparagine; TAC, Tyrosine). Thus, the first nucleotide of the first codon will be combined with the first and second nucleotide of the next codon, altering the reading frame of the protein. This variant is reported in OMIM as related to congenital adrenal hyperplasia (**OMIM** 613815.0015).

The SNV in patient B is located in the second base of an ATC coding for Isoleucine. The new codon AAC codes for an Asparagine. This is a missense change (rs6475, also related to adrenal hyperplasia: **OMIM** 613815.0001).

Learning activity

Genetic analysis for diagnosis and research requires the integration of information from different sources. Understanding the molecular effect of a particular mutation or the pattern of inheritance of a syndrome are necessary steps, but often still not enough to reach conclusions. We might need to know which genes have been related to a certain disease or phenotype, which genetic variants have been proven to originate a syndrome, or retrieve more information on a particular set of genes or mutations to better understand their possible role in a genetic disease.

In this activity we will practice through examples and learn how to retrieve useful information from the main publicly accessible databases in human genetics.

Let's start with **Online Mendelian Inheritance in Man (OMIM)**, a continuously updated catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationship.

Open a browser and go to **OMIM (https://www.omim.org/)**. Let's choose a human disease. For instance, **Cystic Fibrosis (CF)**.

- Have a look to the OMIM FAQS (question 1.3, https://www.omim.org/help/faq) to understand the use of the different symbols.

Read the text. Then, try to answer the following questions:

1. What is the affected gene? CFTR
2. What is the gene doing? The CFTR gene encodes an ATP-binding cassette (ABC) transporter that functions as a low conductance Cl(-)-selective channel gated by cycles of ATP binding and hydrolysis at its nucleotide-binding domains (NBDs) and regulated tightly by an intrinsically disordered protein segment distinguished by multiple consensus phosphorylation sites termed the regulatory domain.
3. What is the most common mutation for CF? delF508
4. Describe the molecular abnormality associated with CF? Free text
5. Can you explain the occurrence of infertility in males with CF? It is due to congenital bilateral absence of the vas deferens. Mutations in the CFTR gene disrupt the function of the chloride channels, preventing them from regulating the flow of chloride ions and water across cell membranes. As a result, cells in the male genital tract produce mucus that is abnormally thick and sticky. This mucus clogs the vas deferens as they are forming, causing them to deteriorate before birth.

Let's now try with other diseases. Please check if you can retrieve this information from OMIM.

1. Which is the pattern of inheritance of Thyroid cancer? Autosomal dominant

2. List the genes reported to be related to all non-medullary thryoid cancers.

NKX2-1
HABP2
NRAS
MINPP1

HRAS
SRGAP1
FOXE1

3. Which diseases have been related to the CF gene: only CFTR?

No. Actually, Bronchiectasis with or without elevated sweat chloride 1; neonatal Hypertrypsinemia, hereditary Pancreatitis, Congenital bilateral absence of vas deferens, Cystic fibrosis and Sweat chloride elevation without Cystic Fibrosis.

As you can see, the OMIM database is giving us useful information on genes and mutations that have been already related to a particular syndrome. Other times we may need more specific information about a particular mutation. Imagine that you discover a genetic variant in one gene that you suspect could originate a genetic syndrome in a patient. If you are lucky, this variant will have been reported as causal in OMIM, but this is often not the case.

In that sense, during the last years, scientists are identifying, cataloguing, and studying genetic variations among humans. A large number of reliable public databases have been developed and form an important resource for variant interpretation (dbSNP, ClinVar, CADD, ExAC, Decipher,etc.).

Let's continue with the example of Cystic Fibrosis. Imagine that we have sequenced the whole gene in some healthy individuals, just to know if some of them are carriers of the disease. We have not found any of the genetic variants listed in OMIM, but there are other genetic variants in this gene.

- 7:117120163 T / C
- 7:117120179 G / A
- 7:117199644 A / G (rs113993960)
- 7:117120217 A / T
- 7:117120229 A / C
- 7:117144299 T / A
- 7:117144400 A / G 7:117149101 G / T (rs77284892)
- 7:117267694 C / G (rs121908763)
- 7:117171096 C / G

Some of the genetic variants found in our patients have been already identified and have an SNP ID. You can check these variants directly in the **dbSNP** database (https://www.ncbi.nlm.nih.gov/snp).

The Single Nucleotide Polymorphism database (dbSNP) is a variation database at the National Center for Biotechnology Information (NCBI). It is a public archive of all short sequence variations, not just single nucleotide substitutions, dbSNP includes also a broad collection of

simple genetic variations such as single-base nucleotide substitutions, small-scale multi-base deletions or insertions, and microsatellite repeats.

For those positions with an SNP id, just write the name of the SNP in the search engine. You will be directed to the page with information for this specific SNP. Here, you will find a lot of information for this genetic variant.

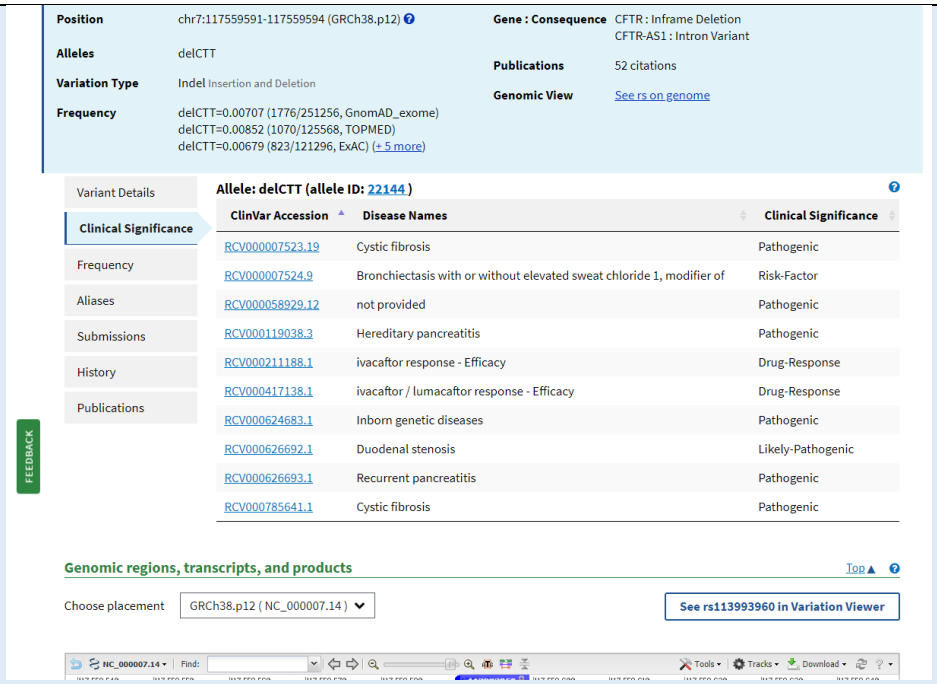
These SNP explain a large amount of CF cases. Can you retrieve this infomartion for the three SNPs with previous SNP id?

- What are the type of mutation?
- What is the frequency in all populations?
- In what population is this variant most frequent?

rs113993960 (is a deletion, freq delCTT=0.004. Regarding 1000Genomes is more frequent in Americans.)
rs77284892 (is a stop gained, freq T=0.00003. Most frequent in Europeans)
rs121908763 (is a stop gained mutation, freq G=0.00002. Most frequent in Europeans)

Among others, you can also find information about its clinical significance, position in the genome, and publications reporting it.

- Are all the reported clinical significances related to specific diseases? Not uniquely.
- What else do you find? Hereditary pancreatitis, ivacaftor response – Efficacy, Inborn genetic diseases, Duodenal stenosis, etc.



To study the rest of obtained variants we can take profit of **gnomAD**, a database of exome sequencing data from a wide variety of large-scale sequencing projects. The data set provided on this website spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies.

Go to gnomAD (http://exac.broadinstitute.org/) and check whether the obtained genetic variants have been already described previously in any individual.

- Which of your variants are present in the database?
- 7:117120163 T / C
- 7:117120179 G / A
- 7:117120217 A / T
- 7:117120229 A / C (is not present)
- 7:117144299 T / A
- 7:117171096 C / G (is not present, the change C /T is the one present in the database)

You can also obtain from here information on its frequency, type of mutation it produces, position in the gene, etc.

- Is there any mutation you did not find anywhere? 7:117120229 A / C and 7:117171096 C / G

Help: if you click in a variant nearby you are directed to a new page with information for this specific variant. At the bottom, you move to the specific position to see the sequence and help predicting new mutations.

Sometimes, it is difficult to know whether the genetic changes observed are really involved in the disease and might be that some of your variants have never been described before. Thus, prioritizing variants for further experimental investigation is a key challenge in current sequencing studies for exploring complex diseases. In this sense, pathogenicity indexes are bioinformatics algorithms that inform us about whether an amino acid substitution is tolerated or damaging. A large number of in silico tools have been employed for this task, including **PolyPhen-2, SIFT, FatHMM, MutationTaster-2, MutationAssessor, CADD, GERP and phyloP**. As an example, CADD is typically used for deleteriousness prediction of single nucleotide variants as well as insertion/deletions.

CADD score is a powerful tool to predict the effect of mutation that integrates multiple databases in one single metric.

If the amino acid change is predicted to be damaging, that is that affects protein function, the CADD score is higher (usually >30). On the other hand, lower values of CADD score predict tolerated changes.

- Please read this paper and we will do an activity based on that.

CADD scores. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992975/

Are you able to predict the effect of the mutations found using CADD?

(**WARNING**, always verify that you are using the same genome assembly, in this case hg19)

You can leave empty the reference and alternative alleles and you will obtain all possible combinations in this position. Pay attention only at the variant we are interested in.

Go to Single nucleotide variant (SNV) lookup (https://cadd.gs.washington.edu/snv). Here you can access the score of a single nucleotide variant. Let's try with some of the ones we obtained before:

Which reference genome were we using?

Which are the CADD scores for all the variants we have identified?

- 7:117120163 T / C → 13.5
- 7:117120179 G / A → 18.47
- 7:117120217 A / T → 11.70

- 7:117120229 A / C → 13.71
- 7:117144299 T / A → 11.35
- 7:117171096 C / G → 15.59

Which you think are the two most harmful and would be the best candidates for further studies? 7:117120179 G / A and 7:117171096 C / G.

Solution

Solutions are written in blue in the learning activity

Quiz

1. Which of the following genes is in the positive strand of the human genome?
   a. LRBA
   b. ABO
   c. SLC24A5
   d. HGD
   e. None

2. This is the sequence for an individual for positions chr8:55,540,510 to 55,540,520: GGATTCAAACTG. Is there any genetic variant when compared to the reference genome (hg19).
   a. No
   b. Yes, a single-nucleotide variant (SNV)
   c. Yes, a insertion of one nucleotide
   d. Yes, a deletion of two nucleotides
   e. Yes, a deletion of ten nucleotides

3. For the same region as in the previous question, this is the sequence for an individual with a single nucleotide variant (SNV): GTATTCAACTG. This variant causes:
   a. A synonymous change
   b. A new stop codon that will disrupt the protein
   c. A missense change of aspartic acid to tyrosine.
   d. A missense change of asparagine to a cysteine
   e. A missense change of glycine to a valine

4. Sickle-cell disease is caused by the replacement of A by T at the 17th nucleotide of the gene for the beta chain of hemoglobin, which changes the codon GAG to GTG. Can you tell what kind of mutation causes the disease?
    a. Silent mutation
    b. Copy number variant
    c. Frameshift mutation
    d. nonsene mutation
    e. Missense mutation

5. What kind of mutation introduces a premature stop codon into a gene?
    a. Nonsense
    b. Silent
    c. Missense
    d. Frameshift
    e. Inversion

6. Which of the following mutations likely result in frameshifts?
    a. synonymous
    b. missense
    c. insertions
    d. deletions
    e. c and d

7. The ultimate source of genetic variation is:
    a. Mutation
    b. The combination of parental chromosomes during sexual reproduction
    c. Chromosomal variants
    d. Splicing
    e. All of them

8. New human mutations typically occur at a rate of about ____ per individual among normal, healthy people.
    a. 2-3
    b. 1000-2000
    c. 0
    d. Millions
    e. It is not known

9. Which is true about random mutations?

a. They are always negative
b. They are always positive
c. They remove variation from a population
d. They introduce new variation into a population
e. They make measuring variation impossible

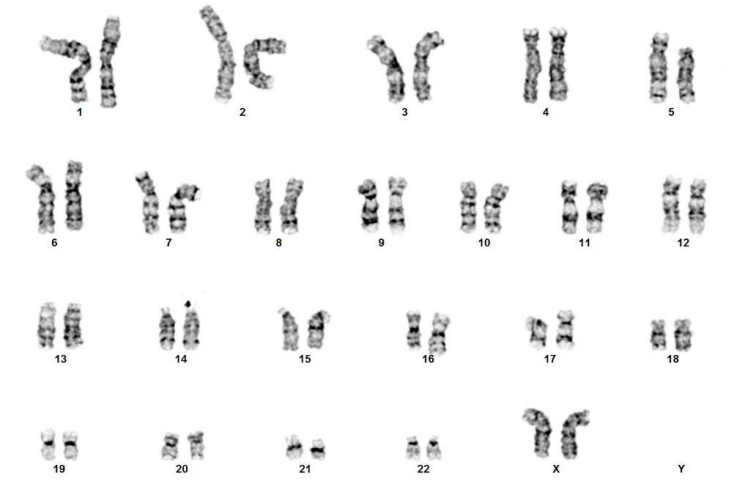
10. The gene defect for both Huntington's Disease and Fragile-X syndrome consists of:
a. a series of repeated nucleotide sequences
b. a mispairing of base pairs
c. a major deletion of an important segment of a gene
d. a metabolic block
e. a missense genetic variant

Learning activities for week 3. Diagnostic tools: How to select the correct test. Cytogenetics

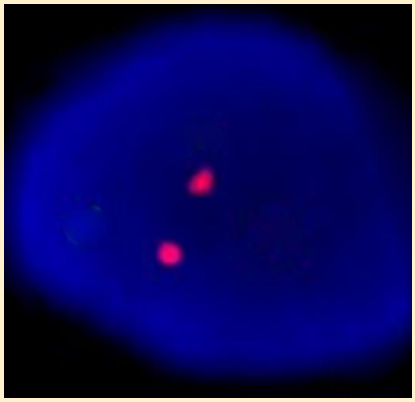
Short problem

Analyze and make the interpretation of the three cytogenetic tests performed in a sample from a 1-year old child with development delay, microcephaly, hypotonia, congenital heart defect and dysmorphic face.

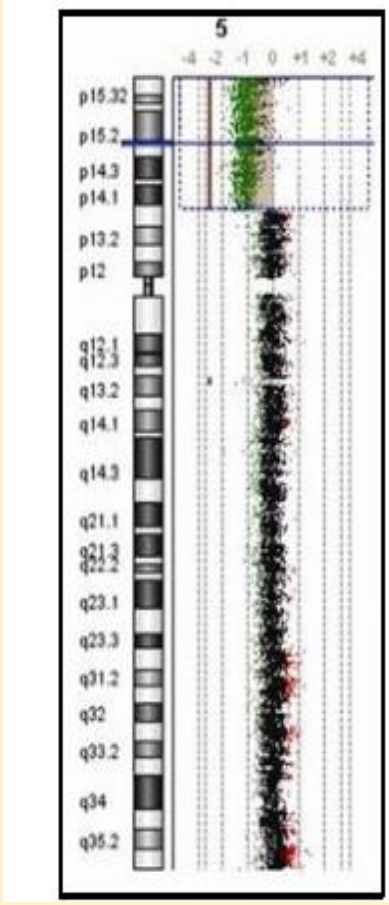
Test 1: Karyotype



Test 2: *FISH* (using centrometric probe of chromosome 5)



Test 3: Molecular Karyotype (CGH-Array)



1. Explain in detail the interpretation of each test and make an integrative final report.
2. Which is the molecular diagnostic of the patient?

3. Which is the most sensitive test (of the three used) to detect the observed alteration? Is any of those three test non informative?

Solution

1.

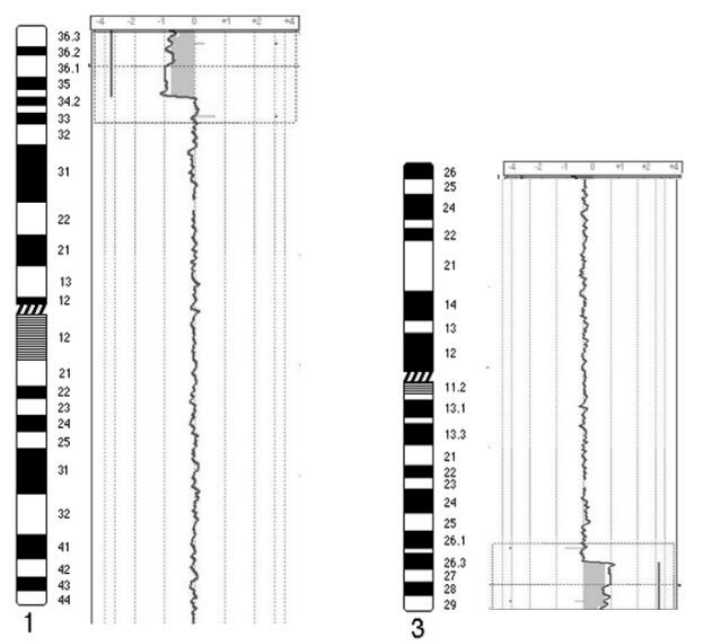
- Test1: Karyotype = 46,XX,del (5) (p-ter)
- Test2: FISH = Disomic for centromeric probes of chromosome 5
- Test3: Molecular Karyotype = arr (reference genome) (5p15.32-p14.1) x1

2. Molecular diagnosis of Cri du chat syndrome (5p minus syndrome). The phenotype reported by the clinician is compatible with the molecular diagnostic obtained.

3. The most sensitive test for the identification of this deletion is the molecular karyotype, although the standard karyotype is also able to detect this large rearrangement. The FISH study with centromeric probes of chromosome 5 is not indicated to molecular diagnose Cri du Chat syndrome.

Learning activity

A clinical geneticist has a consultation of a family with an affected 3 years old child with severe intellectual disability, deafness and dysmorphic features, and a healthy 5 years old girl. The couple wants to know the risk of disease for a future pregnancy.
The genetic test suggested by the geneticist was a molecular karyotype, and the result showed the following pattern:



1. Explain the finding, justify if you think it could be the cause of the pathology and what would be the risk of recurrence in other family members.
2. Explain which genetic tests you would perform to determine the risk of recurrence and in which individuals.

Solution

1. The molecular karyotype shows a p-terminal deletion of chromosome 1 and q-terminal duplication of chromosome 5. This complex rearrangement suggest the presence of a balanced translocation t(1,5) in one of the progenitors. If one of the progenitors was carrier of this balanced translocation the recurrence of this event in a new gestation depends on the specific breakpoints of the translocation and the family history, but is around 30-40%.

2. In order to determine if one of the progenitors is carrier of the balanced translocation t(1,5) the standard karyotype of the progenitors must be performed. In case that the rearrangement is the novo the recurrence risk diminishes to 1%.

Quiz

1. Select the most appropriate technique to identify:

A. Trisomy 18	1. FISH
B. Point mutation in *GRIA1* gene	2. Molecular Karyotype
C. Recurrent mircroduplication/microdeletion	3. Sanger sequencing
D. Submicroscopic balanced translocation	4. Standard Karyotype

a. A1-B2-C3-D4
b. A4-B2-C3-D1
c. A2-B1-C3-D4
d. A4-B3-C2-D1
e. A1-B3-C2-D4

2. Establish the correct relationships between the interpretation of the findings and the following karyotypes:

1. Triploidy	A. 46,XX,del(5)(p13)
2. Aneuploidy	B. 69,XXX
3. Balanced translocation	C. 47,XXY
4. Subtelomeric Cryptic rearrangement	D. 46,XX,t(14;18)

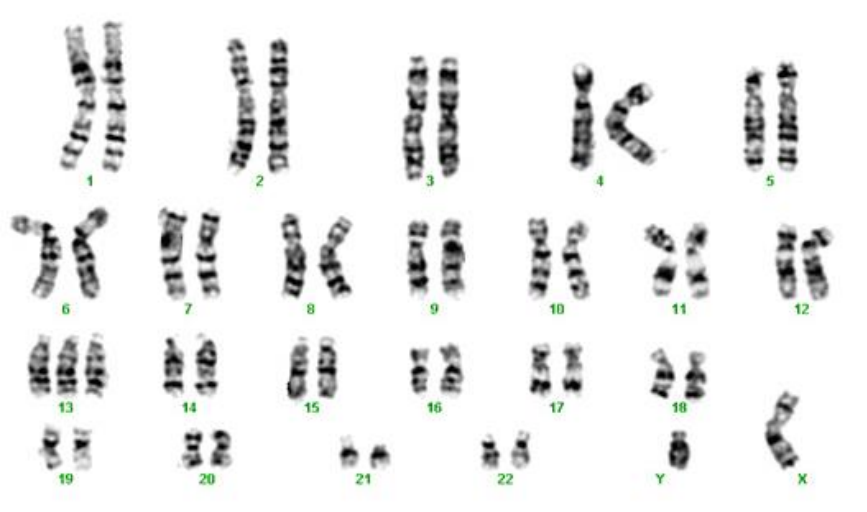
a. 1B-2C-3D-4A
b. 1C-2B-3D-4A
c. 1B-2C-3A-4D
d. 1A-2D-3C-4B
e. 1C-2B-3A-4D

3. In which of the following cases a CGH array would be suitable?
a. Suspicion of uniparental disomy of chromosome 7
b. Study of carriers in a family with point mutations identified in *CFTR* (cystic fibrosis)
c. Suspicion of a balanced translocation in a couple with recurrent spontaneous abortions
d. Association study by SNPs in patients with schizophrenia and controls
e. Suspicion of cryptic rearrangement in patients with intellectual disability and congenital malformations.

4. Point out which of the following statements does NOT correspond to an advantage of molecular karyotyping by Comparative Genomic Hybridization (aCGH) over the conventional karyotype:
a. Decreased response time since it does not require cell culture

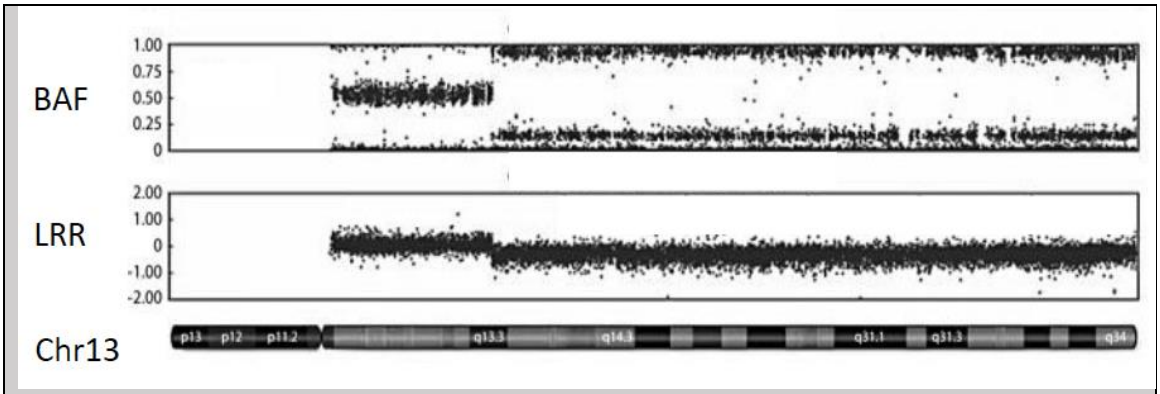
b. Much higher resolution
c. Greater diagnostic capacity as it inspects a higher number of alterations
d. Detection of balanced rearrangement
e. More automated analysis

5. What is the CORRECT result in the interpretation of the following karyotype?



a. Male with karyotype: 46,XY,rob(13;14),+13
b. Male with Patau syndrome and karyotype: 47,XY,+13
c. Phenotypically normal male carrier of a balanced translocation t (4;13)
d. Male with Edwards syndrome and trisomy 13
e. Male with 47 chromosomes, translocation t(13;X) and phenotype of Klinefelter syndrome

6. In a molecular karyotype by SNP array we identify the following rearrangement on chromosome 13 (values of BAF and log2 with respect to the idiogram shown below). We can affirm that this individual presents:



a. A 13q13.3-q terminal deletion in heterozygosity
b. A segmental uniparental isodysomy in 13q (13q13.3-qter)
c. A cen-13q13.3 duplication in homozygosity (tetrasomy)
d. A p-terminal deletion in homozygosity
e. A balanced translocation that affects the q-terminal arm of chromosome 13

7. Identify the most appropriate molecular technique (A-E) for the study of the following situations (1-5):

1) Carrier of Spinal Muscular Atrophy (SMN1 gene deletion)
2) Triploidy
3) Translocation between BCL2 and IgH in patients with lymphoma
4) Uniparental isodysomy of chromosome 15
5) Newborn with low weight and malformations

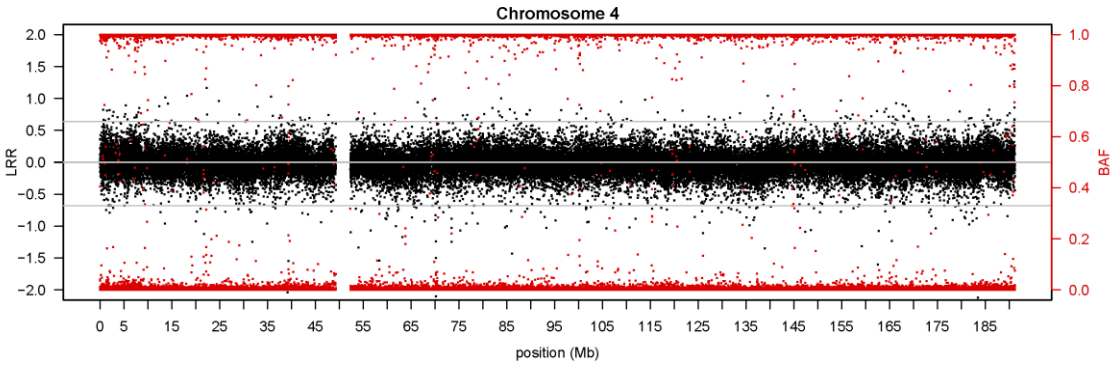
A. Molecular karyotype, aCGH or aSNPs
B. Conventional karyotype
C. MLPA
D. FISH with double fusion probes
E. Microsatellite markers or SNP array

a. 1A, 2B, 3C, 4D, 5E
b. 1B, 2D, 3E, 4A, 5C
c. 1C, 2B, 3D, 4E, 5A
d. 1C, 2E, 3D, 4B, 5A
e. 1D, 2A, 3B, 4C, 5E

8. Regarding the indications of diagnostic genetic tests, indicate the WRONG answer.
a. In a patient with congenital malformations, intellectual disability and dysmorphic features, a molecular karyotyping would be indicated (aCGH / aSNP)
b. In a patient in whom a disease caused by a single gene is suspected, a directed sequencing of the gene in question would be indicated

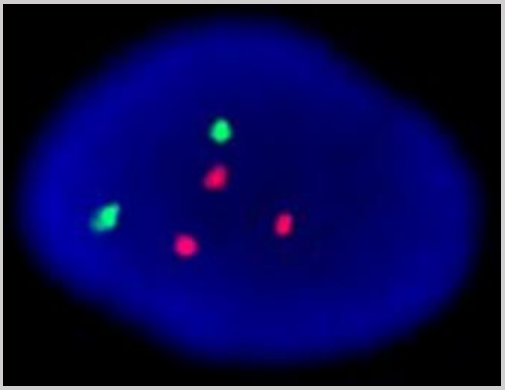
c. In a patient in whom a disease is suspected that can be caused by at least 20 different genes, an exome or genome sequencing would be indicated
d. In a couple with repeat abortions, a molecular karyotyping would be indicated (aCGH /aSNP)
e. In a patient with suspected Down syndrome, a conventional karyotype would be indicated.

9. Indicate which alteration is compatible with the graph shown below, corresponding to the result of a SNP microarray of chromosome 4 (LRR on the left, BAF on the right)



a. Mosaic trisomy
b. Interstitial duplication
c. Monosomy
d. Uniparental disomy (isodisomy)
e. Interstitial deletion

10. Which are the correct result of this FISH study, performed with a centromeric probes of chromosome 3 (green) and 21 (red)?



a. Disomy of chromosome 21 and trisomy of chromosome 3
b. Male with down syndrome
c. Disomy of chromosome 3 and trisomy of chromosome 21
d. Woman with down syndrome
e. Non informative result

Learning activities for week 3. Diagnostic tools: How to select the correct test. Sequencing

Short problem (solutions are written in blue)

In this first activity we will familiarize ourselves with the different output formats and visualization of genetic variants. We will also introduce you to the enrichment analysis in gene expression.
Sanger sequencing has been for many years the canonical technology. Even with the advent of the new generations of sequencers it is still considered the gold standard technology, and often genetic variants discovered with the new technologies are validated with Sanger sequencing. The final output of Sanger sequencing is a chromatogram with different peaks of fluorescence intensity for each position that are translated to nucleotides.
**In case you are not familiar with this, please answer the next questions on chromatograms taken from scientific literature.*

In massive parallel sequencing technologies, DNA reads in FASTQ files are mapped to a reference genome. The results of mapping are stored in a bam file, that includes information on the coordinates for each read in the genome along with quality parameters. Bam files can be visualized using Integrative Genome Browser (IGV). You can install it in your computer (http://software.broadinstitute.org/software/igv/). Now download these files (.bam and .bai files in the BAM/TPP1 folder) and visualize in IGV the alignments. They belong to anonymous individuals analyzed in the 1000 Genomes Project. The three individuals (NA12892, NA12891 and NA12878) are a trio (mum, dad and daughter).

Load the three bam files. They belong to a 60 Kb region in the chromosome 11 (from 6,620,000 to 6,680,000). Note that you only need to load the three .bam files (the .bai file is and index file for .bam files, and just needs to be in the same folder)

How many genes do you see in this region? Write their names

RRP8, ILK, TAF10, TPP1, DCHS1

Go to the position chr11:6,638,385. For this position NA12878 is a carrier of a known loss-of-function variant affecting to the splicing (rs56144125), and related to autosomal recessive neurodegenerative disorders. This genetic variant has been inherited from her father (NA12891).
Go to following nucleotides by writing the coordinates in the box and complete this table with information on the reference nucleotide and the homozygous/heterozygous state in the three samples of the trio.

Position	Reference	NA12878	NA12891	NA12892
chr11:6,638,385	C	CT	CT	CC
chr11:6,643,976	C	CT	CC	CT
chr11:6,655,433	G	AA	AG	AA

Load now the three BAM files for the PELI2 gene region in chromosome 14. Check the position chr14:56,763,353 in NA12878.
What is the reference nucleotide and the genotype for NA12878?
The reference nucleotide is C, and the genotype for this sample is CT.
Is this mutation present in the parents (NA12891 and NA12892)?
No, both parents show the CC genotype.
How can you explain this?
This is a *de novo* mutation, with a most probable origin in the first embryonic divisions or in the gonadal cells of a parent.

Finally, note that in addition to C and T, there are also a few reads in NA12878 with the G nucleotide.
Do you know how to explain this?

Additional nucleotides can be also seen in other positions. They correspond to reads with lower quality (represented with lower intensity) or are a consequence of the substitution error rate of NGS technologies.

Once reads have been aligned to a reference genome, the next steps is generating a list of genetic variants, that is a list of the positions where at least one individual has a difference to the reference genome. The most common file format for this is the vcf file (https://en.wikipedia.org/wiki/Variant_Call_Format). Open the vcf file for the TPP1 gene in five European individuals (In the VCF folder. You can just drag the file in excel, as we do in this video). After several comment lines, you have information on the genetic variants with genotype for each of the individuals

Identify the chr11:6,638,385 position and confirm the heterozygous status in NA12878.
Is there any other individual with this mutation?
No, the rest of individuals are homozygous for the reference allele (0/0)
Do you see any genetic variant present in homozygous state in all five individuals?
Yes, rs7943955 and rs2734718 (all the individuals showing the 1/1 genotype)

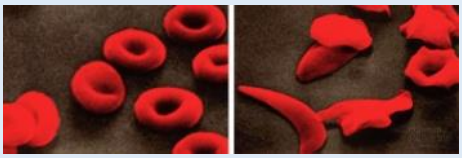
VCF files only contain information on the existing genetic variants, but not on their properties. Adding this information to the vcf file is called annotation process, and can be done using the vcf file. As an example, you can easily do it by using the Variant Effect Predictor (VEP) in Ensembl (https://www.ensembl.org/info/docs/tools/vep/index.html). Go to this tool (set it to GRCh37) and load your file.
A genetic variant will probably has more than one predicted effect, since it can be included in different trasncripts. To simplify the output, activate "Show one selected effect per variant" in the Filtering options.
How many variants are coding? 3
Are they producing a change in the protein sequence? The three variants are synonymous, thus they are not producing an amino acid change
Can you confirm the role in splicing of the genetic variant in chr11: 6,638,385?
Yes, this variant is annotated as splice_acceptor variant
Which are the variants with highest and lowest population frequency?
rs7943955 has an allele frequency of 0.989. rs56144125 has an allele frequency of 0.0002
Is there any genetic variant predicted to be pathogenic?
rs56144125, with the lowest frequency is annotated as pathogenic
Why some genetic variants are annotated as located in a different gene (TAF10)?
This two variants are located in the 5' endo of TAF10, and in the 3' end of TPP1. We have selected the option of only one predicted effect per variant

Learning activity

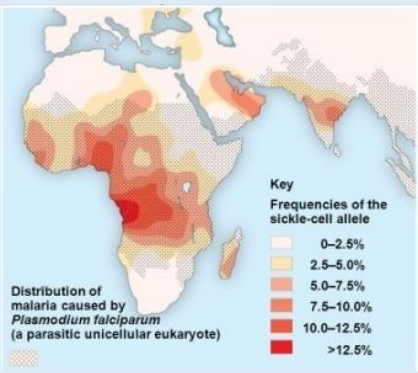
As we have learned within the previous module, humans display geographic variation as well as variation within populations. Geographic variation, or the distinctions in the genetic makeup of different populations, often occurs when populations are geographically separated by environmental barriers or when they are under selection pressures from a different environment.

One example of interplay between geographic variation and mutations is sickle-cell anemia. The sickle-cell anemia trait causes red blood cells to become malformed. These sickle-shaped red blood cells may clog small blood vessels, and stop other red blood cells from delivering oxygen to the body, ultimately starving tissues and organs of oxygen.

Sickle cell disease is the most common inherited blood disorder in the United States, affecting 70,000 to 80,000 Americans. The disease is estimated to occur in 1 in 500 African Americans and 1 in 1,000 to 1,400 Hispanic Americans.



But, individuals who are carriers of the sickle-cell allele have been proven to be resistant to malaria, in fact, 60 percent of sickle-cell carriers survive malaria. This has generated a specific observable geographical pattern of the sickle-cell allele.



Solution

Please, read this study:

- Piel, F.B. et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat Commun. 1:104 doi:10.1038/ncomms1104 (2010).

What is the affected gene in this disease? Beta-globin
Can you provide the genomic position of the gene in hg19 coordinates? 11:5246695-5250626

The sickle-cell disease is passed from generation to generation in a pattern of inheritance called autosomal recessive inheritance. Both the mother and the father must pass on the defective copy of a gene.

- What is the type of mutation that leads to the disease? missense
- Which position and change? 6 Glu→Val

As you can see, codification also follows some rules. You must annotate the position number together with the caused change. In this case, is a glutamate that changes to a valine.

- If the mutation produces such disease, how do you think it is maintained in the population and not removed by natural selection? It is partially beneficial.

Here, you have a vcf file (HBB folder) including all the Yoruba individuals present in the 1000Genomes project. Look at this SNP: rs334. This is the SNP causing the 6 Glu→Val change. As you can see, SNPs have their own nomenclature. Usually, they are called as "rs" followed by a number. Each SNP has its own codification and it represents a specific position in the genome.

- What are the two possible allleles? T and A
- Which is the minor allele frequency of this SNP? 13%

You can corroborate these results in Ensembl and compare the frequencies of the alleles with European populations.

Let's go to gnomAD, a database of human variation. Search the HBB gene.
Now, let's look at this genetic loci: 11-5247781-A-C. Corresponding to SNP rs368604295.

- How many times it is the minor allele present in the population? 893
- Do you see any difference between populations at the level of frequency? Yes.

Similarly, there are diseases that can be both due to a single-gene mutation and a combination of risk variants. For instance Parkinson's disease.
Please, download these 2 files: SNCA_mother and SNCA_patient. These data have been generated in an exome study on a patient with an early-onset Parkinson's Disease and his

mother (without the disease). These are fasta files of the α-synuclein gene (SNCA). Let's see if you are able to detect any new mutation in the patient?

1- Visually inspect the files. The FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. What do our files contain, nucleotides or amino acids?
2- The first line in a FASTA file starts with a ">" (greater-than) symbol. It is used to hold a summary description of the sequence. It gives a name and/or a unique identifier for the sequence, and may also contain additional information. Following the initial line is the actual sequence itself in standard one-letter character string. How many characters our sequences do have?

   SNCA is a relatively short protein, which is found mainly at the neurons.

3- We will use Blastp to align the two sequences (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins).
   Select align two or more sequences and upload the files. Then, run Blast.

   Are the two files identical? No.
   Go to the *Alignments* tab. Change the alignment views to detect differences between the sequences.
4- What is the amino acid change and its specific position? A53T
5- Are you able to find any information on this specific change? Free text

Exactly, the A53T mutation produces Parkinson's Disease. But, it only explains a very small amount of cases. In fact, all familial forms of Parkinson's Disease are rare and explain about 3%–5% of the sporadic cases.

We will get a piece of the data from this study: Nat Genet. 2009 Dec; 41(12):1308-12.
It comes from a Genome-wide Association Study of Parkinson Disease in Samples Sharing Common Caucasian Ancestry. Here, you can find a brief description of the study.
Look at the enclosed file called *PD.xlsx*. It contains the most significant results from this study. In the first sheet you find a brief description of the study and the fields of the results (sheet 2).

- How many variants this study tested? 463,185 SNPs
- Are there any significant results? Yes.

rs4889730
rs3784847
rs2736990
rs3857059
rs415430
rs11931074

- Which are the associated Odds ratios of the associated variants? What is then your risk of having Parkinson's disease if you are a carrier of these variants? Do the tested allele increase or decrease the risk for PD?

SNP ID	P-value	Chr ID	Chr Position	Allele1	Allele2	Odds ratio
rs4889730	2,83E-11	17	21717726	G	T	0,7263
rs3784847	1,66E-09	16	61977448	G	A	1,585
rs2736990	5,69E-09	4	90678540	G	A	1,272
rs3857059	3,60E-08	4	90675237	A	G	1,492
rs415430	4,50E-08	17	44859143	C	T	0,7464
rs1193107 4	4,78E-08	4	90639514	G	T	1,486

Look for the associated variants

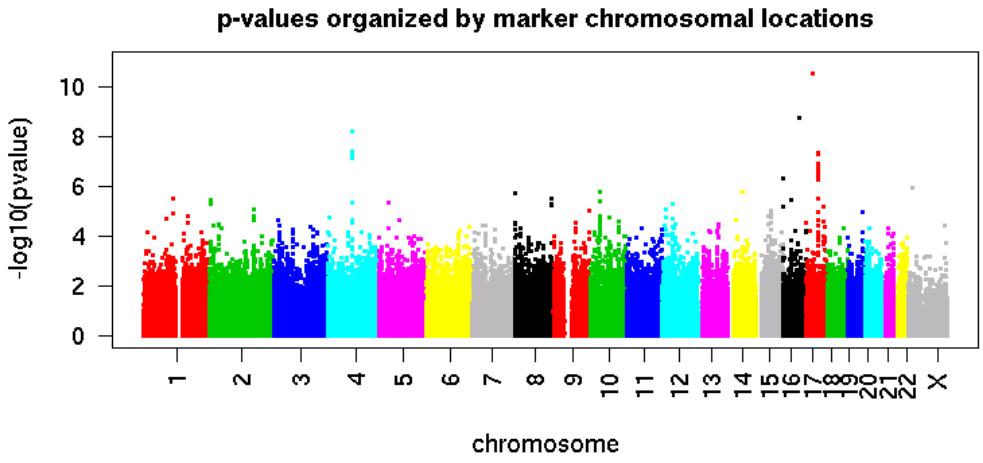
- Are all the associated variants located within genes?

  rs4889730 intron variant
  rs3784847 intron variant
  rs2736990 intron variant
  rs3857059 intron variant
  rs415430 intron variant
  rs11931074 intron variant

- What do you think that intronic and/or intergenic variants are doing?

Functional polymorphisms that can influence the expression of the genes

Usually, results from a GWAS analysis are presented as Manhattan plots. This is the manhattan plot generated from this study. On the x-axis you find all the genomic positions tested ordered in chromosomes. On the y-axis, the –log10 p-value of each genetic variant is plotted.



So, this study discovered some genetic variants putatively associated with Parkinson's Disease in the cohort studied, with very low OR. But this study is from 2009, since then many other variants have been associated to disease. Let's visit the GWAS Catalog website.

- How many variants have been reported to be associated to Parkinson's disease? 276 associations at 01/01/2020

Which is the lowest p-value observed?
At 07/01/2020, 8 x 10^{-240} for young adult-onset parkinsonism (chr21:46916033-?) and 5 x 10-123 for Parkinson's disease (rs356182-G).

Quiz

1. If you want to study variation at the level of genes, what kind of approach will you choose?
    a. **Exome sequencing.**
    b. SNP array
    c. FISH
    d. Whole-genome sequencing.
    e. None of the above

2. A SNP is:
    a. a frame shift mutation
    b. a nucleotide change
    c. a rare genetic variant

d. a coding genetic variant
e. a real risk variant

3. Many SNPs tested for association with disease show no evidence for association. What would be the odds ratio for SNPs with no evidence of association?
   a. One
   b. Zero
   c. Two
   d. Minus one
   e. None is correct

4. The goal of GWAS is to discover genetic factors that contribute to:
   a. Mendelian diseases
   b. Common diseases
   c. Diseases with chromosomal aberrations
   d. Diseases with very low prevalence in the population
   e. All the above

5. Which variants would you prioritize as candidates for a rare and severe Mendelian disorder?
   a. Rare over common
   b. Coding over intergenic
   c. Missense over synonymous
   d. Nonsense over missense
   e. All the above

6. Which of thee following sequences in data generation and analysis after sequencing is correct?
   a. Sequencing (fastq file)->Variant calling (vcf)->Variant annotation ->mapping (bam)
   b. Sequencing (fastq file)-> Variant annotation -> Variant calling (vcf)->mapping (bam)
   c. Sequencing (fastq file)-> mapping (bam) ->Variant annotation -> Variant calling (vcf)
   d. Sequencing (fastq file)-> mapping (bam) -> Variant calling (vcf)-> Variant annotation
   e. Sequencing (fastq file)-> Variant annotation -> mapping (bam) ->Variant calling (vcf)

7. Which of the following sentences on expression analysis is false?

a. RNA is retrotrasncribed into cDNA for library preparation and sequencing.
b. The number of sequence reads of a transcript is proportional to its levels of expression.
c. Both arrays and sequencing technologies are used for gene expression analysis.
d. Increasing the number of samples will confer power to detect smaller expression differences.
e. Whole transcriptome sequencing is also an appropriate and cost-efficient tool to study expression differences in a single gene.

8. GWAS evaluates:
   a. All the genomic positions from all individuals in a study
   b. Only coding positions
   c. Only common variation in the population
   d. Only non-coding positions
   e. Rare variants

9. In a GWAS:
   a. The identified variant may not be the causal one
   b. Many identified variants may correspond to a single causal variant due to linkage disequilibrium
   c. We need more stringent significance thresholds
   d. One of the major confounders is population stratification
   e. All of them are correct

10. The study of how genetic variations influence our responses to medications is called:
    a. GWAS
    b. Pharmacogenomics
    c. Risk Score
    d. FISH
    e. Metabolome

Forum

Now that you know how a good pedigree should be done, how many of you think that you do it in your clinical practice? What aspect do you consider is the one you should improve the most? What do you think are the advantages of drawing a good pedigree?

Recommended readings for this module

Bennett RL1, Steinhaus KA, Uhrich SB, O'Sullivan CK, Resta RG, Lochner-Doyle D, Markel DS, Vincent V, Hamanishi J. **Recommendations for standardized human pedigree nomenclature.** J Genet Couns. 1995 Dec;4(4):267-79.

Bennett RL1, French KS, Resta RG, Doyle DL. **Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors.** J Genet Couns. 2008 Oct;17(5):424-33. doi: 10.1007/s10897-008-9169-9. Epub 2008 Sep 16

Merlin G. Butler **Genomic imprinting disorders in humans: a mini-review.** J Assist Reprod Genet. 2009 Oct; 26(9-10): 477–486.

MODULE 4. QUALITY IMPROVEMENT IN HEALTHCARE

Learning activities for week 1. Risk Assessment

Short problem

In the video we have seen how we can integrate additional information from a pedigree (having three healthy sons) to predict genetic risk.

Now, we propose a new situation with age-dependent penetrance of the disease, where the probability of being affected increases with age. We are going to estimate this in a family with an autosomal dominant syndrome.

1. Draw a pedigree for a couple where the man is affected by a dominant syndrome, and two of their three offspring are also affected. Also, half of the offspring of both affected are also affected.
   Note that since this is an autosomal dominant disorder no matter if you represent them as males or females, but please draw one male and one female per couple!
2. What is the a priori probability for any offspring of the original couple to manifest the disease?

3. Now imagine that the healthy child of the original couple is 35 years old and is planning to have children, and this disease is manifested during adulthood with a probability of manifesting the disease for 35-years old carriers of 80 %.
   What is his/her probability of being a carrier?
   Use a table with the prior, conditional, joint and posterior probabilities, as we did in the video.

Solution

1.

2. 0.5

3.

Hypothesis	Carrier	Non-carrier
Prior probability	0.5	0.5
Conditional probability (of being healthy at 35)	0.2	1
Joint probability	0.1	0.5
Posterior probability	0.1/0.6=0.17	0.5/0.6=0.83

Obviously, when the mutation producing a disease in a family is known, the individual status can be defined with a genetic test interrogating this position. However, in some situations not all the mutations originating the disease are known, so a negative result does not completely exclude the possibility of being a carrier, and Bayesian risk estimations help to better estimate the probabilities of being affected or being a carrier for a certain disease. Let's do it with an example. Imagine a population with a relatively high prevalence for an autosomal recessive disorder of 1/50. An individual without family history of this disease gets a negative result in a genetic test screening only 70 % of the mutations for the disease. What is now his/her probability of being a carrier? Again, to calculate this complete the following table.

Hypothesis	Carrier	Non-carrier
Prior probability	1/50	49/50
Conditional probability (of a negative result if he is carrier/non-carrier)	3/10 (30 %)	1
Joint probability	3/500	49/50
Posterior probability	0.006/0.986=0.0061	0.98/0.986=0.9939

You can also read more on the combination of Bayesian analysis and genomic screening in the first recommended lecture for this week (*Genomic screening and genomic diagnostic testing-two very different kettles of fish*)

Learning activity

There are several factors that contribute to your risk for complex health conditions like diabetes and heart disease.

There are two kinds of risks that are used when talking about a person's risk for developing diseases: absolute risk and relative risk.

Absolute risk is an individual's risk of developing a given disease. For example, if you have 1 in 10 chance of developing diabetes in your lifetime, you are said to have a 10% absolute risk.

Relative risk is used to compare the risk between two groups of people; one group has a certain risk factor, and the other group does not. For example, you could compare the risk for heart disease of a group of people who smoke to a group of people that do not smoke. The risk given to each group would be called their relative risk, meaning their risk compared to another group's risk.

A relative risk of 1 means there is no difference in risk between the two groups. If a person's relative risk is greater than 1, it means their risk for developing that disease is higher than the other group. If a person's relative risk is lower than 1, they have a lower disease risk than the other group.

Genetic variant risk is usually reported using relative risk. GWAS most often report Odds Ratios (OR) relative to the low-risk allele or lowest-risk genotype. To turn this into a meaning risk estimate, the prevalence of the disease and the genotype frequencies must be taken into account.

We have genotyped the SNP rs1333049 in a cohort of 2675 participants. Among them, 775 were controls and 1900 were people affected by Coronary Artery Disease (CAD). The following table shows the genotypic distributions for the SNP.

	Controls	Cases
GG	250	460
GC	375	940
CC	150	500

1. Calculate the Odds Ratio (OR) for developing CAD if you have one copy of the risk variant of the SNP rs1333049.

Help:

OR_{GC}=odds of disease in an individual with the GC genotype/odds of disease in an individual with the GG genotype

2. And what will be your odds ratio if you are homozygote for the risk allele (having two copies of the C allele)?

Having a protective genetic variant can actually decrease the likelihood that you will develop a disease. The protective allele G in the variant rs671 is associated with a lower risk of developing coronary artery disease. Here we screen the SNP rs671 in the same population and the genotypic frequencies are distributed as follows:

	Controls	Cases
GG	280	370
GA	335	910
AA	160	620

3. What is the OR for the GG genotype?

Besides the genotypic OR, we can also calculate the allelic OR. Thus, in spite of using the genotypes, we will use only the alleles.

Here, we will represent the allelic distribution of the SNP rs671 for cases and controls in our cohort.

4. Can you fill out the empty cells?

	Controls	Cases
G	895	
A	655	

For diseases such as CAD, researchers use multiple genetic variants to estimate the genetic risk. While this will allow for a better estimation of genetic risk than when using only a single genetic variant, this multi-variant risk reports will *not* represent your complete genetic risk for a condition.

5. Why do you think that this multigenic variant-risk do not represent completely the risk of developing a particular disease?

    1) there are additional genetic variants that are still not included in the calculations

    2) some genetic variants have yet to be discovered

    3)some genetic variants have an uncertain association with disease.

    4) all the above are correct.

The OR for each SNP can be accounted together to combine effects from multiple SNPs to obtain a straightforward way of calculating a Genetic Risk Score (GRS).

This is, in fact, what commercial tests usually do. They use an SNP array to decipher your variation at several SNP positions and they use previously reported odds of disease in these positions to compute your genetic risks for diseases or how you respond to some drugs.

This weighted method weights each risk allele by the logarithm odds ratio (Log(OR)) for that allele in our dataset. In fact, the wGRS is a linear combination of the number of risk alleles weighted by the Log(OR) as coefficients, which accounts for the effect sizes of the SNPs.

Basically, this can be done as follows log(OR) SNP1 + log(OR) SNP2 + log(OR) SNP3 = Total OR

As an example, we can compute the genetic risk of developing Venous Thromboembolism by using a combination of these 3 SNPs:

The allele T for the SNP rs6025 has an OR of 3.6 of developing VT.

The allele A for the SNP rs1799963 has an OR of 2.8 of developing VT.

And finally, the allele C for the SNP rs505922 has an OR of 1.8 of developing VT.

6. First, use dbSNP to quickly annotate these variants. Are they intronic? Exonic? In which genes are they located?

We have genotyped three unrelated people for the 3 alleles. They showed the following genotypes:

	John	Peter	Joe
rs6025	CT	CC	TT
rs1799963	AG	AG	AA

rs505922	CC	TT	CT

7. Let's compute a GRS for each one:

Hint. use this formula:

N risk alleles x log(OR_{SNP1}) + N risk alleles x log(OR_{SNP2}) + N risk alleles x log(OR_{SNP3})

Those are our weighted GRS.

8. Who are the ones with a higher and lower GRS?

9. If the mean risk in the population for this GRS is 0.5, can you compute how many times more risk to develop the disease each one has?

Of course, for many complex traits, using only three SNPs provide low contributions to the trait predictivity. Most of the times, one would need thousands of SNPs to achieve enough predictive power.

On top of that, these are only genetic risks, and as we now know, genetics and environment play sometimes equal roles in developing complex diseases.

Solution

1.

OR=1.36

This means you are 36% more likely to develop coronary artery disease than someone who does not have any copies of the risk variant.

You can go to GWAS Catalog or SNPedia and search for this SNP. There are many different studies that associated this variant to CAD with similar ORs.

2.

OR= 1.81

3. If you compare the risk of a person that has two copies of this protective variant to a person that has no copies of the protective variant, the individual with two copies of the protective variant is 70% less likely (OR = 0.3) to develop CAD.

Again, you can check SNPedia and GWAS Catalog to contrast our results.

4.

	Controls	Cases
G	895	1650
A	655	2150

5. 4) all the above are correct.

6.

rs6025 is exonic

rs1799963 is a 3 Prime UTR Variant

rs505922 is intronic.

7.

John= 1.51

Peter= 0.45

Joe= 2.26

8.

Joe has the higher GRS and Peter the lower GRS.

9.

John has 3 times more risk to develop the disease than the average

Peter 0.9

And Joe has 4-5 times more risk than the average population.

Quiz

1. After a genetic test, we have confirmed that the first child of a healthy carrier for an autosomal recessive disease is also a carrier. What is the probability of his/her second child being affected?
    a. 0
    b. 0.25
    c. 0.5
    d. 0.75
    e. 1

2. What is the probability of carrying the disease for a 30-year-old man whose mother has a late-onset autosomal dominant disorder, and when the probability for carriers of manifesting the disease at 30 years is 10%?
    a. 0
    b. 0.09
    c. 0.47
    d. 0.53
    e. 1
4. After a genetic test we have determined that both members of a couple are carriers for heterozygous mutations causing the same recessive disorder. The first child is affected for this condition. What is the probability that the second one is affected?
    a. 0
    b. 0.06
    c. 0.25
    d. 0.5
    e. 1

4. 95 % of the mutations originating a rare autosomal recessive disorder with a prevalence of 1/10,000 are located in exons of a given gene. Sanger sequencing of the coding fraction of the gene is commonly performed as a tool for molecular diagnosis. Which of the following statements is true?:
    a. The probability of being a carrier for an individual with a negative result in the test is 5 %.
    b. The probability of transmitting the mutation to the offspring for an individual with a positive result in the test is 95 %.
    c. The probability of being a carrier for an individual that has not been screened is 5 %.
    d. All the above are true.
    e. None is true.

5. A woman's probability of being a carrier for an X-linked disease, given that his mother was a carrier:
    a. A priori is 0.5
    b. Is less than 0.5 if she has had three unaffected sons
    c. a and b
    d. Is less than if she has had three unaffected daughters
    e. All the above
6. In polygenic risk scoring, the number of risk alleles carried by an individual are:
    a. Multiplied

b. Summed, and weighted by the OR from the discovery GWAS
c. Not informative
d. Informative, but only three of them
e. More than three.

7. Polygenic risk scores:
   a. Are widely used in clinics for complex disease prediction
   b. Try to predict common disease risk from DNA
   c. Are free of false-positive results
   d. Are made on genes but not SNPs
   e. Are used in monogenic diseases
8. In complex diseases:
   a. The origin is unknown
   b. The genetics do not play a role
   c. Very little number of genetic variants are associated
   d. Many genetic variants influence the trait
   e. The PRS can be calculated using one genetic variant
9. Having a high polygenic risk score of diabetes:
   a. Means that the person will develop the disease
   b. Means that the person will never develop the disease
   c. Does not mean anything
   d. Means that the person has more risk of developing the disease
   e. Any of the above is correct
10. A discovery GWAS is performed in Africans:
    a. The PRS will perform better in Africans
    b. The PRS will perform well in non-Africans
    c. Is not appropriate since all GWAS need mixed genetic backgrounds
    d. You can only use the genetic variants that overlap in Europeans

It is great, because we all come from Africa and this will work in all human backgrounds.

Learning activities for week 2. Ethical and Communication skills

Learning activity

Mark and Sarah, a non-consanguineous couple of 39 and 35 years old, come to your genetic counseling consultation. Mark has recently been diagnosed of Huntington's disease by genetic testing because of his family history (his mother died of this disease). Currently he has no symptoms.

The reason for the consultation is related to their offspring: they have a 9-year-old daughter (Julie) and Sarah is pregnant again (10 weeks of pregnancy). Both of them want to carry out the genetic test of their daughter Julie to know if she has inherited the genetic alteration that causes the disease in the family since they argue that the uncertainty of not knowing if Julie will suffer the same disease as her grandmother and her father is unsustainable. In addition, they ask for prenatal diagnosis for the ongoing pregnancy to know the status of the fetus in relation to Huntington's disease. Due to the anxiety presented by both members of the couple, they also ask if it is possible to expand the prenatal genetic study to rule out more genetic pathologies since they are unable to assume another genetic diagnosis in the family. They have been reading in the Internet and specifically ask you if it would be possible to perform an exome sequencing in the current pregnancy since, according to the couple's textual words "they want to discard everything".

Identify the ethical dilemmas raised by the clinical case. Do you think that in any of them an adequate informed consent could be useful?

Forum

Read the text “Welcome to Holland” in the website (https://www.ndss.org/resources/a-parents-perspective/). Discuss with your classmates if a family could benefit from receiving genetic counseling in a situation as the one described in the text.

Recommended readings for the module

On the risk estimation for Mendelian disorders

Biesecker, L.G. Genomic screening and genomic diagnostic testing—two very different kettles of fish. *Genome Med* **11,** 75 (2019) doi:10.1186/s13073-019-0696-9
https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0696-9

On personalized genetic tests

Tandy-Connor, S., Guiltinan, J., Krempely, K. *et al.* False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet Med* **20,** 1515–1521 (2018) doi:10.1038/gim.2018.38

https://www.nature.com/articles/gim201838

https://www.sciencenews.org/article/health-dna-genetic-testing-disease

https://www.theguardian.com/commentisfree/2018/aug/10/dna-ancestry-tests-cheap-data-price-companies-23andme

Ethics in genetic counseling

Clarke, A.J., Wallgren-Pettersson, C. Ethics in genetic counselling. *J Community Genet* **10,** 3–33 (2019) doi:10.1007/s12687-018-0371-7**.**
https://link.springer.com/content/pdf/10.1007%2Fs12687-018-0371-7.pdf

Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics.
https://www.nature.com/articles/gim2016190.pdf