

# Bi5444 Analysis of sequencing data

Lesson 1 - General information and introduction

# General introduction

# Teachers



- Dr. Eva Budinska – guarantor, teacher – RECETOX, MU
- Assoc. prof. Marek Mraz, MD – teacher - Univ. hosp. Brno & Fac. of Medicine & CEITEC MU
- Ing, Vojtěch Bartoň – teacher, RECETOX MU

Each will take care of different part of the course and at the end everything will be “merged” together

# Lets get to know each other...

- Are you **familiar** with the field?
- Do you **work** with the **data** somehow?
- Do you **receive** any **outputs** from the analysis?
- Do you **plan** to **work** with it?
- Do you want to **understand** the **outputs**?

The course itself

# After the course you should...

- **Know the latest NGS methods** (next and third generation sequencing), their use and the type of data they produce
- **Be able to distinguish the type** of method based on the data
- Know the **basic scheme** of data analysis
- Able to work with **Linux, Bash** and **R** at a level sufficient for analysis of NGS data – partially
- Know how to **select tools** for data processing and apply them to real data
- **Be able to analyze NGS data** starting from quality control over alignment to the detection of differentially expressed genes (in RNA-Seq), variants (CNV with SNP), genome assembly, etc.

# Main objective:

- The course is **addressed** to everyone who **already works** with or **plans to work** with the NGS data, wants to learn something new from the field or wants to understand the data/outputs
- The **lectures will explain** you **the basics** and show you **examples**
- You will need to **study/exercise some extra** to get **better understanding** of the process
- During **the course**, there will be possibility to **you can discuss your own data analysis** (if any)

# Content of the course

- We will **not cover all topics** in the NGS field –simply there is **not enough time**
- We will provide you with **solid basics of NGS data analysis** that will allow you to **easily extend** to almost **any NGS application** and data types and to **work** with the **data**
- We will give you **hints** where to **look**, what to **look for**, what to **study** and how to **think** about the data
- At the beginning we will cover the **biology background** and also do the **revision of your knowledge** in the **biology/molecular biology** field – necessary for correct understanding of the NGS



# Let's check the prerequisites

- Knowledge of **molecular biology**
- At least a **basic knowledge** of work with **Linux** system
- **Basic** knowledge of **R and statistics** is an **advantage**
- **Basic programming** knowledge is an **advantage**

# Study materials

- There are plenty of **study materials available online**
- But the whole **field changes** very **quickly** – try to **look** for the **latest information**
- **Few years old** materials are most likely **not very useful** any more or they are already surpassed
- **Presentations** from the lectures **will be** available **online**
- There will be always a **link** to some interesting **papers** during the **course** where possible
- **It is never a bad idea to ask!**

# Other recommended courses

- **C2110 UNIX and programming**
- **Bi7560 Introduction to R**
- **Bi7420 Modern methods for genome analysis**
- **Bi7528 Analysis of genomic and proteomic data**
- **Bi7527 Data Analysis in R**
- **Bi7492 DNA Sequence Analysis**
- **Bi5010 Detection of biomarkers from omics experiments**

You can also see the study catalogues of Mathematical Biology and Biomedicine (direction Biomedical Bioinformatics) or Chemoinformatics and bioinformatics degrees for the recommended courses

# Online courses - examples

## Linux/Unix

- <http://www.ee.surrey.ac.uk/Teaching/Unix/>, ...
- BioLinux
- [http://nebc.nerc.ac.uk/nebc\\_website\\_frozen/nebc.nerc.ac.uk//support/training/course-notes/past-notes/intro-bl7](http://nebc.nerc.ac.uk/nebc_website_frozen/nebc.nerc.ac.uk//support/training/course-notes/past-notes/intro-bl7)
- , ...
- R
- <http://www.r-tutor.com/r-introduction>,  
[http://ww2.coastal.edu/kingw/statistics/R-tutorials/text/quick&dirty\\_R.txt](http://ww2.coastal.edu/kingw/statistics/R-tutorials/text/quick&dirty_R.txt), ...
- Other interesting courses
- <https://www.coursera.org/>, <http://online.stanford.edu/courses>, <https://www.edx.org/>,  
<http://www.codecademy.com/>, <http://ocw.mit.edu/index.htm>,  
<http://www.rna-seqblog.com/>, ...
- Questions & Answers
- <http://seqanswers.com/>, <https://www.biostars.org/>, <http://stackoverflow.com/>, ...
- Blogs & Other
- [www.linkedin.com](http://www.linkedin.com), [www.researchgate.net](http://www.researchgate.net), <http://core-genomics.blogspot.cz/>,  
<http://nextgenseek.com/>, <https://twitter.com/>, ...
- Introduction to Next Generation Sequencing
- <https://www.ebi.ac.uk/training/course/introduction-next-generation-sequencing>, ...

# Work with the computer

- We will try to cover the basics of work with Linux, bash, R, ... BUT the course is **not** directly meant to be focused on programming and/or work with Linux system, bash, R, etc.
- It will be very helpful (for you) to look into some basics on your own
- There are numerous tutorial available online for everything (uncle Google can help you very well)
- There are also several very helpful online courses organized by top universities all over the world

# Evaluation and grading

- Written test – 10 questions, 20 points
- Activity on the course – 5 points
- To successfully pass you need at least 15 points
- The presence is compulsory, 1 absence is allowed

Computational resources

# Access to the computers/resources

Access to the resources -C4/1.18

- You need to get access to **WOLF** computers (here)  
<http://wolf.ncbr.muni.cz/>
- Apply for the account -now:

<https://einfra.ncbr.muni.cz/whitezone/root/index.php?lang=en&action=ncbr&show=wolf>

- To the description what you want to do please put:  
*"Student of Bi5444 -fall semester 2022"*
- **Temporary login: guestXX, NCBR@1.18**



# Access to the resources - advanced

If you know some Linux we recommend to use MetaCentrumresources

- <https://metavo.metacentrum.cz/en/index.html>
- Apply for the account online “Getting an account -> Registration form”
- Login with MU identification number & secondary password and ask for account creation
- To the description what you want to do please put: “*Analysis of the sequencing data*” or similar
- You can work with you laptop but you would have to install all required tools on your own –contact us if it is your case

# NGS introduction

# Next-generation sequencing introduction

- Deciphering DNA sequence is essential for all the branches of “biological” research
- It has become widely adopted in numerous laboratories all over the world
- **Next-generation sequencing (NGS)** is a new (almost) technology in the sequencing
- It helps to overcome the limitations of older techniques such as speed, scalability, throughput and resolution
- In this course you will get familiar with NGS as itself, its use and basic data processing

# Before we proceed any further

- **Bioinformatics** (and especially the sequencing bioinformatics) is a **very new field**
- No good books, no standards, nothing lasts forever, ... **almost everything** is old and **outdated!**
- **Bioinformaticians** have to be **always** looking for **new methods**, tools, algorithms, ... it's the same when wet-lab people must search for novel methods which for decrease bias, are faster, require less input material, ...
- The good thing is that there is **still a space for improvement** – for you!
- However, the data **analysis is never trivial**
- **Garbage in -garbage out**
- If you **do not understand** the whole process you **don't know** what the **results** mean

# History

- **Maxam-Gilbert** sequencing 1977 -complex, very radioactive, ...
- **Sanger** sequencing 1977 -widely used, dideoxy method, “golden standard” (??), slow, low throughput, ...
- **Next-generation sequencing** since 2001
- Started with pyrosequencing (1999 in Sweden) – later “rented” by 454 -> Roche, now discontinued
- Big leap forward thanks to the **Human Genome Project**
- HGP was launched in 1990 and finished in 2003 by publishing first complete human genome (**\$2.7 billion**) – classic Sanger
- They had competition – Celera genomics founded in 1998 and finished in 2003 (\$300 million) –shotgun sequencing
- But Celera cheated a bit

# Year 2010



Year 2022

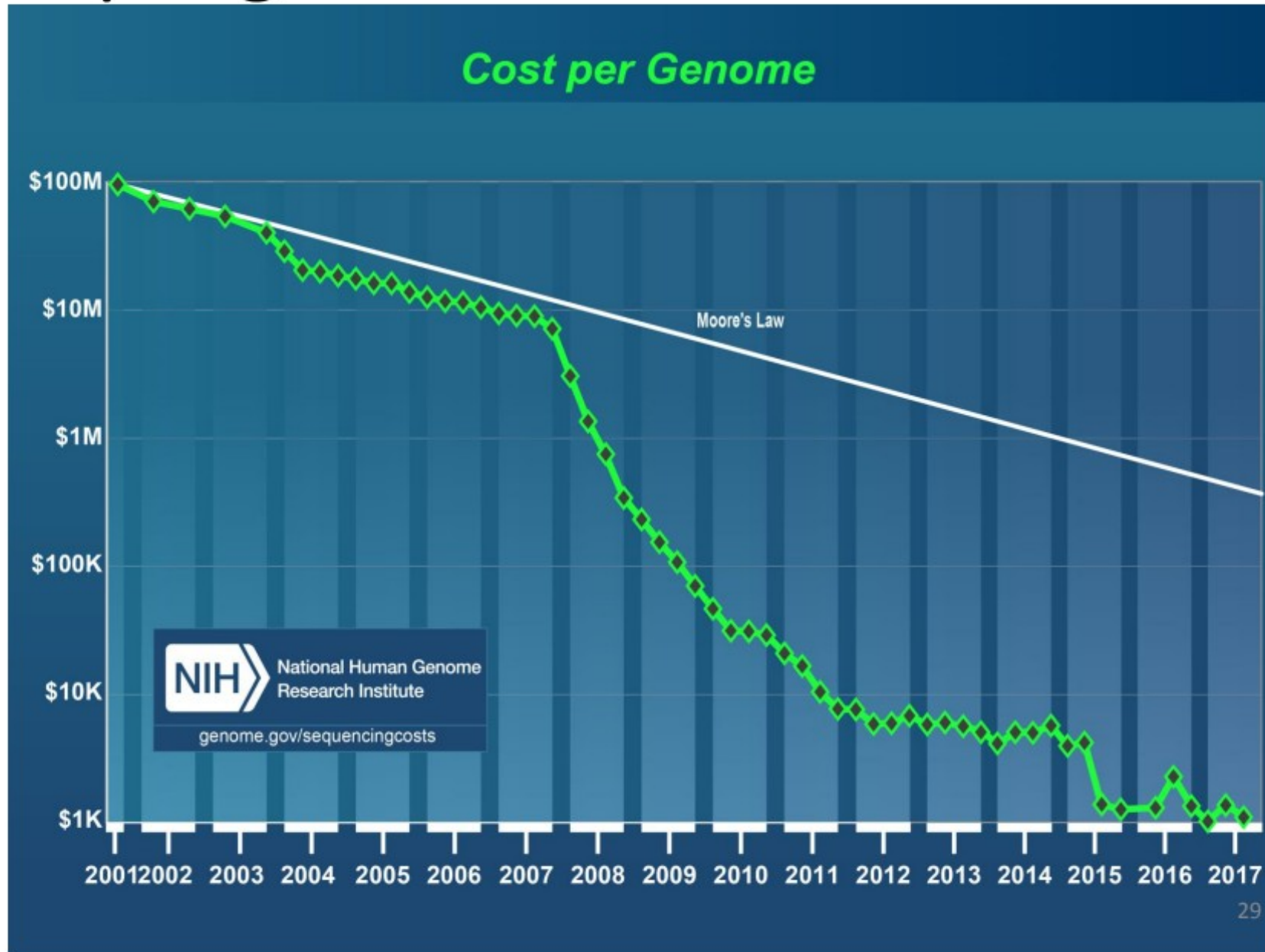


# Comparison of NGS

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US\$)	Advantages	Disadvantages
<b>Single-molecule real-time sequencing (Pacif Biosciences)</b>	30,000 bp <a href="#">(N50)</a> ; maximum read length >100,000 bases <a href="#">[66][67][68]</a>	87% raw-read accuracy <a href="#">[69]</a>	500,000 per Sequel SMRT cell, 10–20 gigabases <a href="#">[66][70][71]</a>	30 minutes to 20 hours <a href="#">[66][72]</a>	\$0.05–\$0.08	Fast. Detects 4mC, 5mC, 6mA. <a href="#">[73]</a>	Moderate throughput. Equipment can be very expensive.
<b>Ion semiconductor (Ion Torrent sequencing)</b>	up to 600 bp <a href="#">[74]</a>	99.6% <a href="#">[75]</a>	up to 80 million	2 hours	\$1	Less expensive equipment. Fast.	Homopolymer errors.
<b>Pyrosequencing (454)</b>	700 bp	99.9%	1 million	24 hours	\$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
<b>Sequencing by synthesis (Illumina)</b>	MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length <a href="#">[76]</a>	\$0.05 to \$0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
<b>Combinatorial probe anchor synthesis (cPAS- BGI/MGI)</b>	BGISEQ-50: 35-50bp, MGISEQ 200: 50-200bp, BGISEQ-500, MGISEQ-2000: 50-300bp <a href="#">[77]</a>	99.9% (Phred30)	BGISEQ-50: 160M, MGISEQ 200: 300M, BGISEQ-500: 1300M per flow cell, MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.	1 to 9 days depending on instrument, read length and number of flow cells run at a time.	\$0.035- \$0.12		
<b>Sequencing by ligation (SOLiD sequencing)</b>	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	\$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences. <a href="#">[78]</a>
<b>Nanopore Sequencing</b>	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)	~92–97% single read	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	\$500–999 per Flow Cell, base cost dependent on expt	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
<b>Chain termination (Sanger sequencing)</b>	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	\$2400	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time



# Cost per genome



# \*Seq things

NGS sequencing has a **wide range of use**

- One of many nice list give you an example of all possible applications
- <http://enseqlopedia.com/enseqlopedia/>
- Approximately (on this list) ~**200 different** techniques...
- Another (simple) list of NGS based techniques
- <https://liorpachter.wordpress.com/seq/>
- <http://nextgenseek.com/2012/11/did-you-know-there-are-at-least-20-applications-of-next-generation-dna-sequencing/>

# BGI



# Some interesting resources

- [https://en.wikibooks.org/wiki/Next\\_Generation\\_Sequencing\\_\(NGS\)](https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS))
- <http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna->
- [https://en.wikipedia.org/wiki/DNA\\_sequencing#Next-generation\\_methods](https://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods)

# Galaxy

Some more information can be found here

- <https://galaxyproject.org/learn/>
- <https://galaxyproject.github.io/training-material/topics/introduction/tutorials/galaxy-intro-strands/tutorial.html>
- <https://galaxyproject.org/tutorials/g101/>

And the main server is hosted here

- <https://usegalaxy.org/>

# See you next week, same place, same time



PacBio Sequel – Pacific Biosciences  
Technologies (SMRT)



MinION - Oxford Nanopore

# Extra



SmidgION: Oxford Nanopore, iPhone-powered sequencing