

Niniejsze opracowanie zostało stworzone przez dr Magdę Mielczarek, pracownika Uniwersytetu Przyrodniczego we Wrocławiu w ramach wykonywania obowiązków związanych z kształceniem studentów i jest przeznaczone dla studentów Bioinformatyki (Wydział Biologii i Hodowli Zwierząt) na potrzeby dydaktyczne bez prawa do dalszego rozpowszechniania.

ANALIZA DANYCH NGS - WSTĘP

Analiza danych NGS
Wykład 1

PROWADZĄCY

- Dr Magda Mielczarek
magda.mielczarek@upwr.edu.pl

Katedra Genetyki
Kozuchowska 7
Gabinet 302

Konsultacje:

- środa 11:00 – 12:00
 - czwartek 11:00 – 12:00
 - konieczny jest wcześniejszy kontakt mailowy
-
- Wykłady będą udostępnione online (theta.edu.pl)



PLAN WYKŁADU

- Tematyka przedmiotu
- Charakterystyka zajęć
- Polecane źródła

TEMATYKA PRZEDMIOTU



NEXT GENERATION SEQUENCING (NGS)

- Rewolucja!
- Przejście od analizy fragmentów genu tradycyjnymi metodami do zastosowania technik umożliwiających uzyskanie pełnej sekwencji wielu genów, a nawet genomów

Metoda:

- wysokoprzepustowa, szybka
- tania

www.illumina.com

Accessible Whole-Genome Sequencing

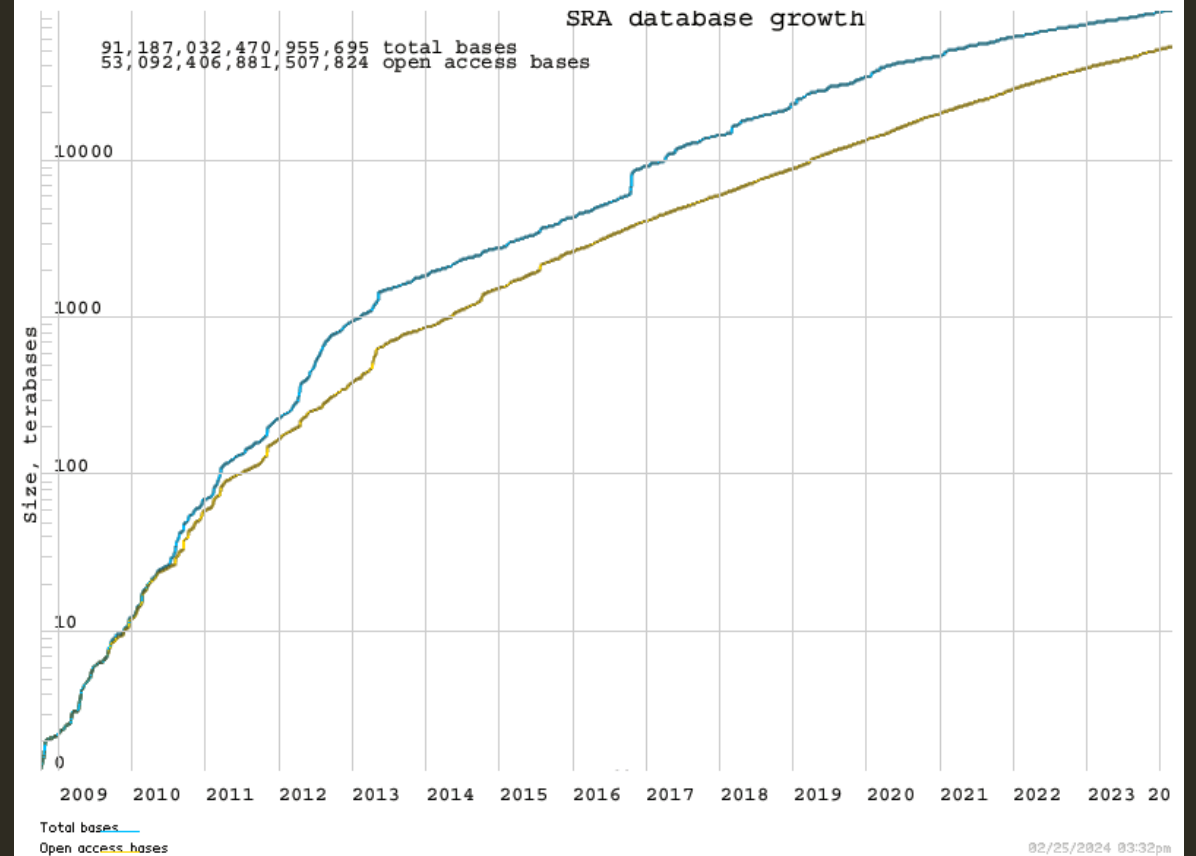
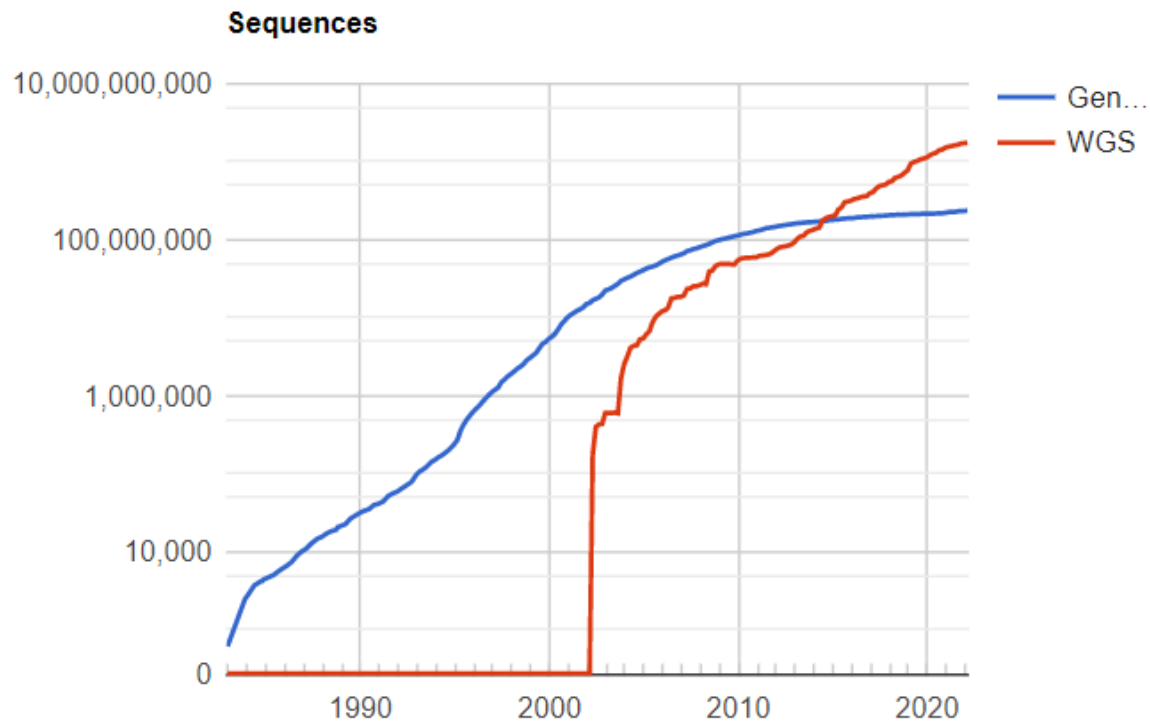
Using capillary electrophoresis-based Sanger sequencing, the Human Genome Project took over 10 years and cost nearly \$3 billion.

Next-generation sequencing, in contrast, makes large-scale whole-genome sequencing (WGS) accessible and practical for the average researcher. It enables scientists to analyze the entire human genome in a single sequencing experiment, or sequence thousands to tens of thousands of genomes in one year.

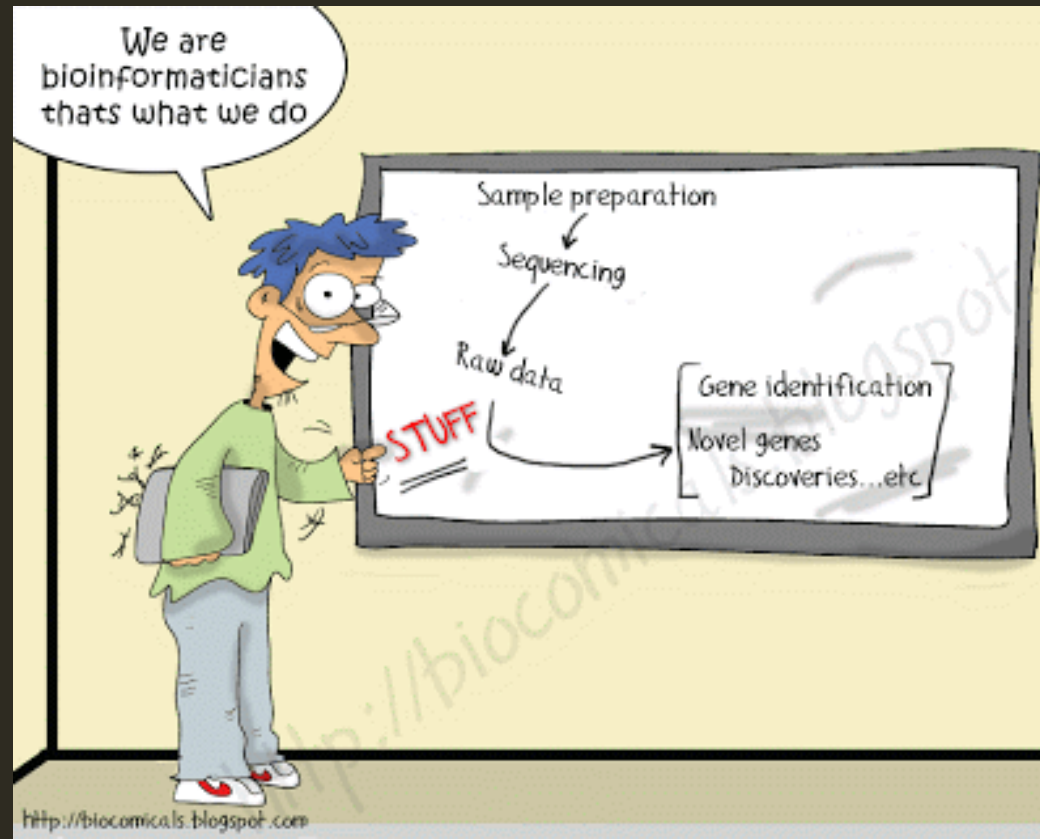
GŁÓWNE ZASTOSOWANIA NGS

- Detekcja polimorfizmów genetycznych
- Poznawanie nowych genomów (*de novo genome assembly*)
- RNA-Seq: profilowanie transkryptomu
- Chip-Seq: Interakcje na linii białko-DNA
- Methyl-Seq: Epigenomika i metylacja DNA
- Metagenomika

BIG DATA - EKSPLOZJA DANYCH ZMIENIA ŚWIAT



TEMATYKA PRZEDMIOTU



PROJEKT 1

The genetic background of clinical mastitis in Holstein Friesian cattle

Sekwencje DNA całych genomów 32 krów HF:

- 16 osobników podatnych na zapalenie wymienia
- 16 osobników zdrowych

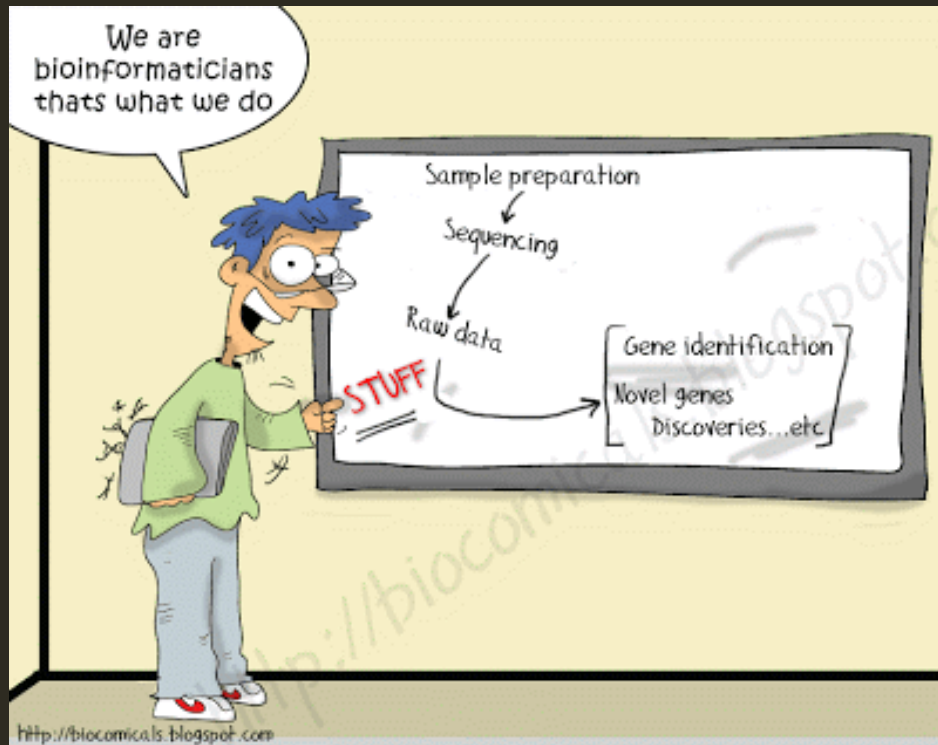
Ujednolicony zbiór danych, różnice odnośnie zapalenia wymienia.

CEL: Znalezienie powiązania pomiędzy polimorfizmami SNP oraz CNV a zapaleniem wymienia

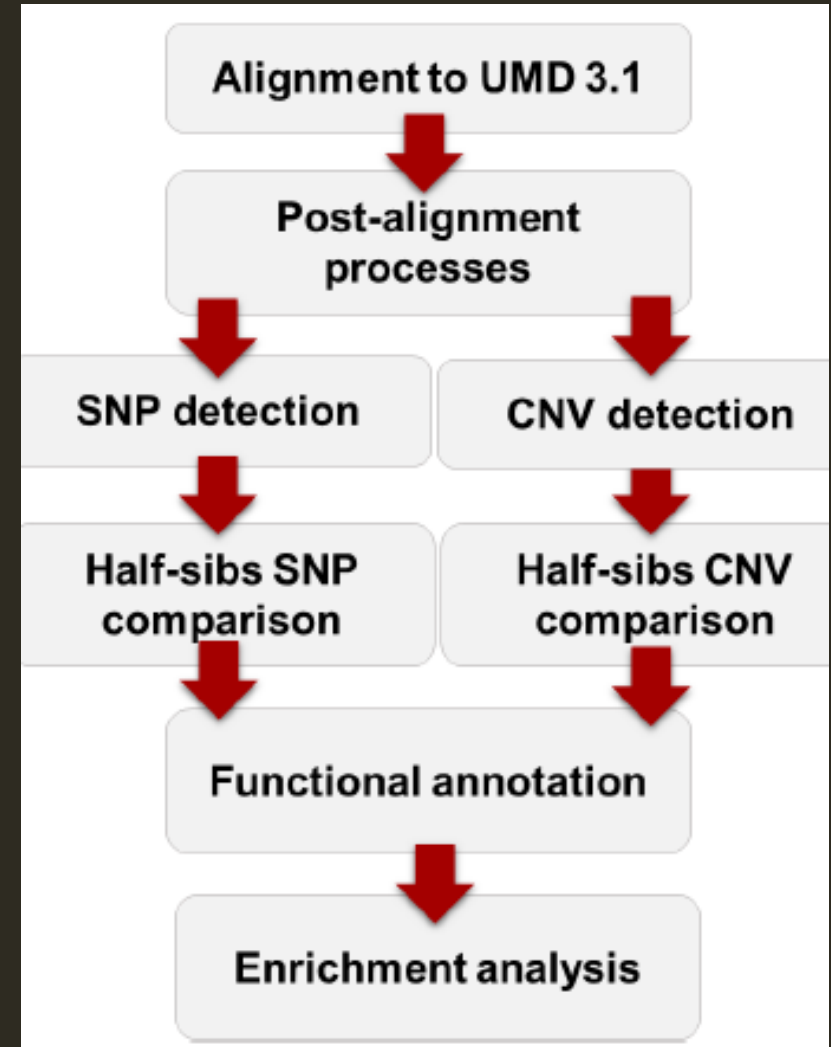


PROJEKT 1

Surowe dane NGS → polimorfizmy?



„STUFF”



The genetic background of clinical mastitis in Holstein-Friesian cattle

J. Szyda^{1,2†}, M. Mielczarek^{1,2}, M. Frąszczak¹, G. Minozzi³, J. L. Williams⁴ and K. Wojdak-Maksymiec⁵

¹Biostatistics Group, Department of Genetics, Wrocław University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wrocław, Poland; ²Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland; ³Department of Veterinary Medicine, Università degli Studi di Milano, Via Giovanni Celoria 10, 20133 Milano, Italy;

⁴Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia; ⁵Department of Animal Genetics, West Pomeranian University of Technology, Długości 26, 71-466 Szczecin, Poland

(Received 15 September 2018; Accepted 28 January 2019; First published online 5 March 2019)

APP: amyloid beta precursor

- bactericidal and antifungal activities in human
- molecular markers for SCM in ruminants

FOXL2: forkhead box L2

- role in inflammation

SNPs

Introns of genes with differences in SNP genotypes between sick and healthy cows

RefSeq coding ring finger protein 122 (10/16 half-sibs)

- cell viability and immune response

SSFA2: sperm specific antigen 2

- associated with SCS = indicator of CM
- overlaps with QTL for bovine immunoglobulin G

CNV DELETIONS

Exons deleted in a sick cow, but present in its healthy half-sib exhibits potential causal influence CM (at least 7 half-sibs pairs)

NDUFS6: NADH:ubiquinone oxidoreductase subunit S6

- QTL for SCS = CM indicator

DGVa: Boussaha *et al.* 2015

MET proto-oncogene receptor tyrosine kinase (9/16 half-sibs)

- processes related to inflammation, cancerogenesis

WRN gene (8/16 half-sibs)

- premature aging in humans, increased susceptibility to infection

ADORA2A: adenosine A2a receptor

- modulating tissue response to inflammation
- in mice highly expressed in mammary gland

DGVa: Hou *et al.* 2011

TXNRD2: thioredoxin reductase 2

- candidate for influencing susceptibility to *S.aureus*

DGVa: Hou *et al.* 2011

OTUD3: deubiquitinase 3

- associated with inflammatory bowel disease in humans

DGVa: Keel *et al.* 2017

PROJEKT 2

The effect of transcriptomic annotations in breast cancer differential gene expression study

ANALIZA SEKWENCJI CAŁEGO TRANSKRYPTOMU

- Technologia NGS (Next Generation Sequencing)
- Sekwencje 4 próbek: RNA-seq: (transkryptom)
- Czynniki różnicujące pomiędzy grupą kontrolną i grupą badaną: nowotwór piersi (analiza różnicowa ekspresji genów; DGE)

CEL:

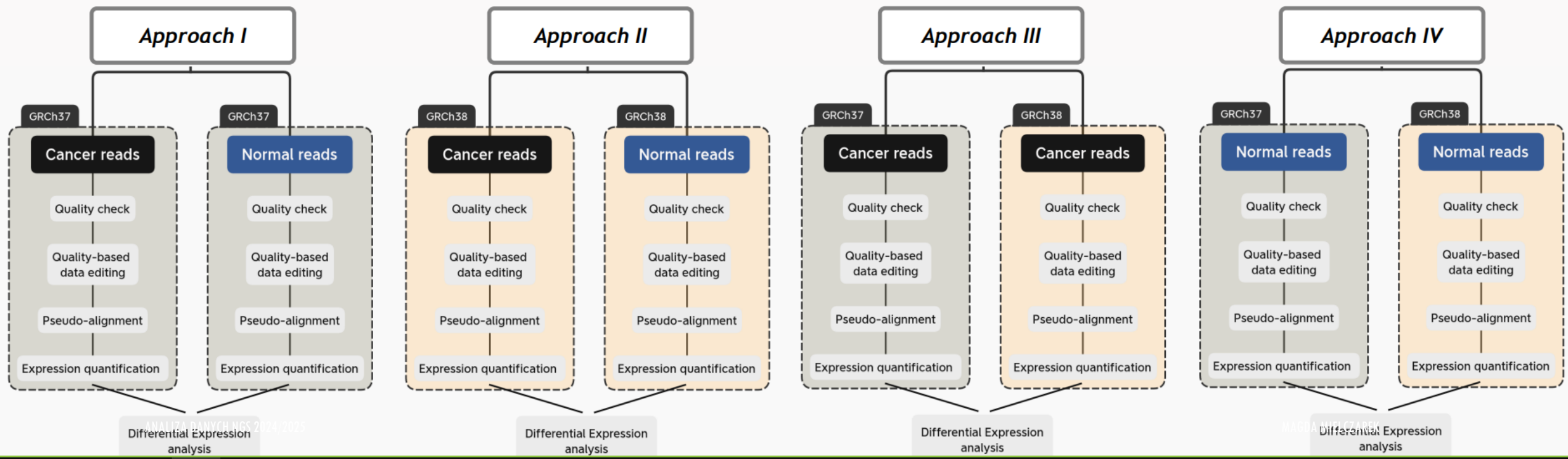
- Zbadanie wpływu wersji transkryptomu referencyjnego na wyniki analizy różnicowej ekspresji genów



PROJEKT 2

The effect of transcriptomic annotations in breast cancer differential gene expression study

„STUFF”



PROJEKT 2

The effect of transcriptomic annotations in breast cancer differential gene expression study

Wnioski:

1. Wyniki DGE nie były identyczne pomiędzy adnotacjami GRCh37 a GRCh38
2. adnotacja transkryptomu nie miała wpływu na geny związane z predykcją przeżycia w raku piersi

PROJEKT 3

- GENOM (DNA-seq)
- TRANSKRYPTOM (RNA-seq)

Modulation of gene expression by CNVs

ANALIZA SEKWENCJI DNA CAŁEGO GENOMU I TRANSKRYPTOMU

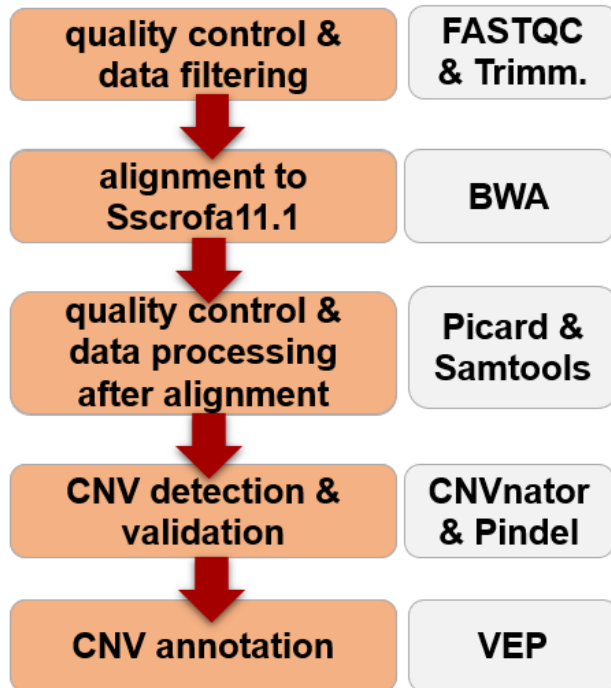
- Technologia NGS (Next Generation Sequencing)
- Zbiór danych zawierających sekwencje DNA (genom) i RNA (transkryptom)



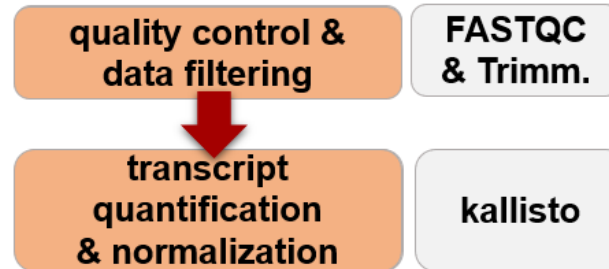
PROJEKT 3

„STUFF”

1. CNVs detection (DNA)



2. Transcript quantification (RNA)



3. CNV impact on gene expression investigation



Wnioski:

1. Delecje powodują spadek ekspresji genów
2. Duplikacje nie zmieniają poziomu ekspresji genów
3. Nie ma korelacji między wielkością delecji/duplikacji a poziomem ekspresji genów

PROJEKTY NGS (DNA-SEQ)

WSPÓŁPRACA THETA + SKN + TUTORING

1. **M. Kołomański**, J. Szyda, M. Frąszczak, M. Mielczarek.
DNA sequence features underlying large-scale duplications and deletions in human.



DNA sequence features underlying large-scale duplications and deletions in human

Mateusz Kołomański¹  · Joanna Szyda¹  · Magdalena Frąszczak¹  · Magda Mielczarek¹ 

Received: 28 December 2021 / Revised: 22 March 2022 / Accepted: 5 May 2022 / Published online: 20 May 2022
© The Author(s) 2022

Abstract

Copy number variants (CNVs) may cover up to 12% of the whole genome and have substantial impact on phenotypes. We used 5867 duplications and 33,181 deletions available from the *1000 Genomes Project* to characterise genomic regions vulnerable to CNV formation and to identify sequence features characteristic for those regions. The GC content for deletions was lower and for duplications was higher than for randomly selected regions. In regions flanking deletions and downstream of duplications, content was higher than in the random sequences, but upstream of duplication content was lower. In duplications and downstream of deletion regions, the percentage of low-complexity sequences was not different from the randomised data. In deletions and upstream of CNVs, it was higher, while for downstream of duplications, it was lower as compared to random sequences. The majority of CNVs intersected with genic regions — mainly with introns. GC content

PROJEKTY NGS (RNA-SEQ)

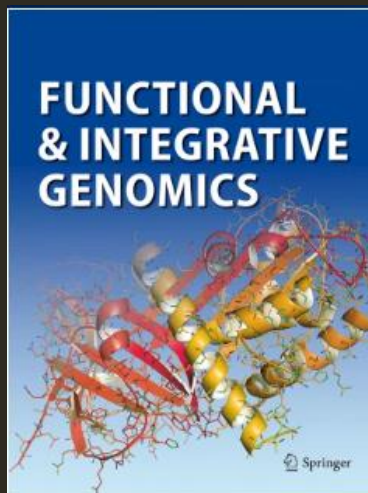
WSPÓŁPRACA THETA + SKN + TUTORING

1. **B. Hofman**, J. Szyda, M. Frąszczak, M. Mielczarek.
LncRNAs Variability in Skeletal Muscle of Polish Landrace Boars.
2. **M. Jarosz**, M. Mielczarek, **W. Wilman**, M. Wołoszyńska.
Analiza różnicowa ekspresji genów u *A. thaliana*.
3. **R. Stępień**, J. Szyda, B. Czech, M. Mielczarek.
The effect of transcriptomic annotations in breast cancer differential transcript expression study.

PROJEKTY NGS (DNA-SEQ + RNA-SEQ)

WSPÓŁPRACA THETA + SKN + TUTORING

1. M. Mielczarek, M. Frąszczak, A. Zielak-Steciwko, B. Nowak, B. Hofman, J. Pierścińska, J. Szyda. Impact of Copy Number Variation on gene expression.



An effect of large-scale deletions and duplications on transcript expression

Magda Mielczarek¹ · Magdalena Frąszczak¹ · Anna E. Zielak-Steciwko¹ · Błażej Nowak¹ · Bartłomiej Hofman¹ · Jagoda Pierścińska¹ · Wojciech Kruszyński¹ · Joanna Szyda¹

Received: 26 September 2022 / Revised: 14 December 2022 / Accepted: 15 December 2022 / Published online: 23 December 2022
© The Author(s) 2022

Abstract

Since copy number variants (CNVs) have been recognized as an important source of genetic and transcriptomic variation, we aimed to characterize the impact of CNVs located within coding, intergenic, upstream, and downstream gene regions on the expression of transcripts. Regions in which deletions occurred most often were introns, while duplications in coding regions. The transcript expression was lower for deleted coding ($P=0.008$) and intronic regions ($P=1.355 \times 10^{-10}$), but it was not changed in the case of upstream and downstream gene regions ($P=0.085$). Moreover, the expression was decreased if duplication occurred in the coding region ($P=8.318 \times 10^{-5}$). Furthermore, a negative correlation ($r=-0.27$) between transcript length and its expression was observed. The correlation between the percent of deleted/duplicated transcript and transcript expression level was not significant for all concerned genomic regions in five out of six animals. The exceptions were deletions in coding regions ($P=0.004$) and duplications in introns ($P=0.01$) in one individual. CNVs in coding (dele-

CHARAKTERYSTYKA ZAJĘĆ



WYKŁADY (30 H)

1. Analiza danych NGS – wstęp
2. Wybrane bazy danych
3. Detekcja polimorfizmów – omówienie standardowych kroków
4. Kontrola jakości i edycja danych NGS
5. Przyrównanie do genomu referencyjnego, post alignment
6. Detekcja i filtrowanie polimorfizmów
7. Adnotacja polimorfizmów, analiza funkcjonalna
8. Analiza genomu – podsumowanie
9. Transkryptomika - analiza DGE
10. Epigenetyka
11. Prezentacje
12. Prezentacje
13. Prezentacje
14. Prezentacje
15. Podsumowanie tematyki przedmiotu – dyskusja

ĆWICZENIA (30 H)

1. Ćwiczenia organizacyjne
2. Przygotowanie danych i środowiska do pracy nad projektem
- 3.-10. Analiza danych WGS – praca nad projektem

Dwa tygodnie na:

→ ZAPISY

→ Wybranie danych do analiz

ZALICZENIE

Nieobecności: max 20% (6h → 2 zajęcia), dotyczy ćwiczeń

Zaliczenie: Przedmiot zakończony egzaminem ustnym w formie prezentacji

- Na początku semestru zapisujemy się na konkretny dzień zaliczenia
- Praca w parach, każda (!) osoba prezentuje i uczestniczy w dyskusji
- Prezentacja zgodna z wytycznymi

POLECANE ŹRÓDŁA



KSIĄŻKI, PUBLIKACJE NAUKOWE

