Databáza variantov v genóme viniča

Školiteľka: doc. Mgr. Bronislava Brejová, PhD.

Veronika Tordová

Genetický variant

- Genetický variant
  - Špecifická časť genómu rozdielna medzi genómami dvoch jedincov toho istého druhu
  - SNV, inzercia, delécia
- Ako sa varianty hľadajú
  - Sekvenovanie, zarovnanie, hľadanie variantov
- Prečo varianty hľadáme
  - Evolúcia, medicína, šľachtiteľstvo

Vitis vinifera (Vinič hroznorodý)

- *V. vinifera ssp. sylvestris*
- Štúdie
  - Evolutionary genomics of grape (*V. vinifera ssp. vinifera*) domestication (2017)
  - Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses (2019)
- Databázy
  - VitisGDB

Cieľ práce

- Vytvoriť databázu variantov v genóme viniča z verejne dostupných dát
- Porovnať so vzorkami skúmanými na Prírodovedeckej fakulte UK

Prvotný prístup

- Výber podmnožiny vzoriek
- Kontrola a predspracovanie dát
  - FastQC, Trimmomatic
- Zarovnanie čítaní
  - BWA-MEM, Hisat2
- Označenie duplikátov
  - Samtools markdup
- Štatistika a vyhodnotenie nástrojov
  - Samtools stats, samtools flagstat

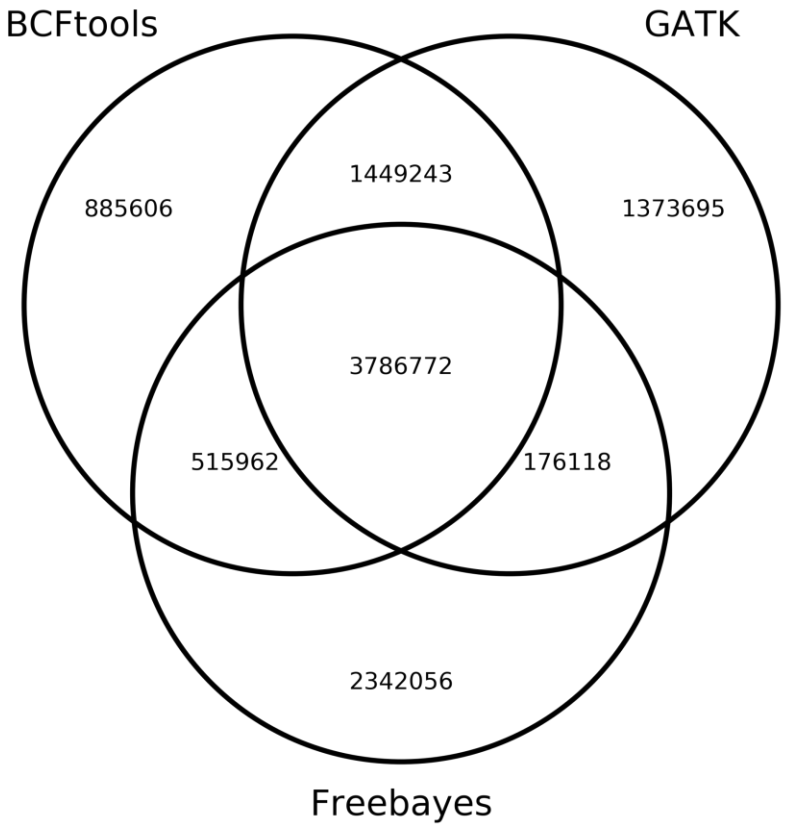
Porovnanie zarovnaných čítaní v percentách nástrojmi BWA-MEM, Hisat2

	Genóm 12X		Genóm 12X.v2	
Vzorka	BWA-MEM	HISAT2	BWA-MEM	HISAT2
V. v. sylvestris Liang2019-TA-6201	98,31 %	74,65 %	97,77 %	67,73 %
Kultivar Liang2019-TA-238	98,86 %	68,65 %	98,38 %	64,62 %
V. v. sylvestris SK-41A13	93,73 %	75,09 %	93,28 %	71,10 %
V. labrusca SK-41A49	88,66 %	59,44 %	88,29 %	56,53 %

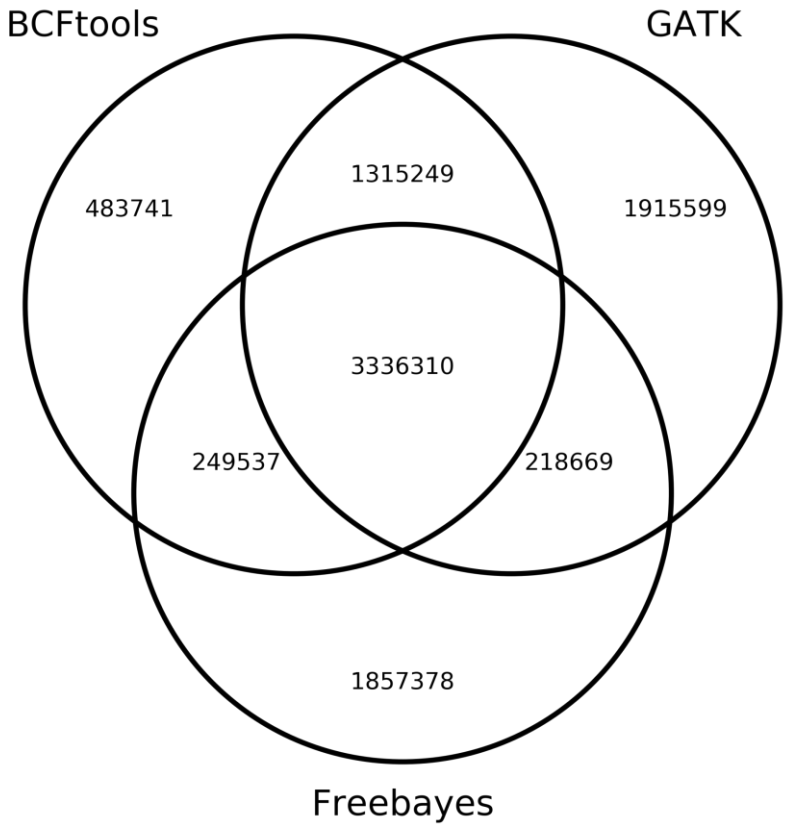
Prvotný prístup

- Identifikácia variantov
  - GATK, Freebayes a Bcftools
- Filtrácia
- Štatistika a vyhodnotenie nástrojov

Porovnanie počtu nájdených variantov nástrojmi Bcftools , GATK a Freebayes



Pred filtráciou



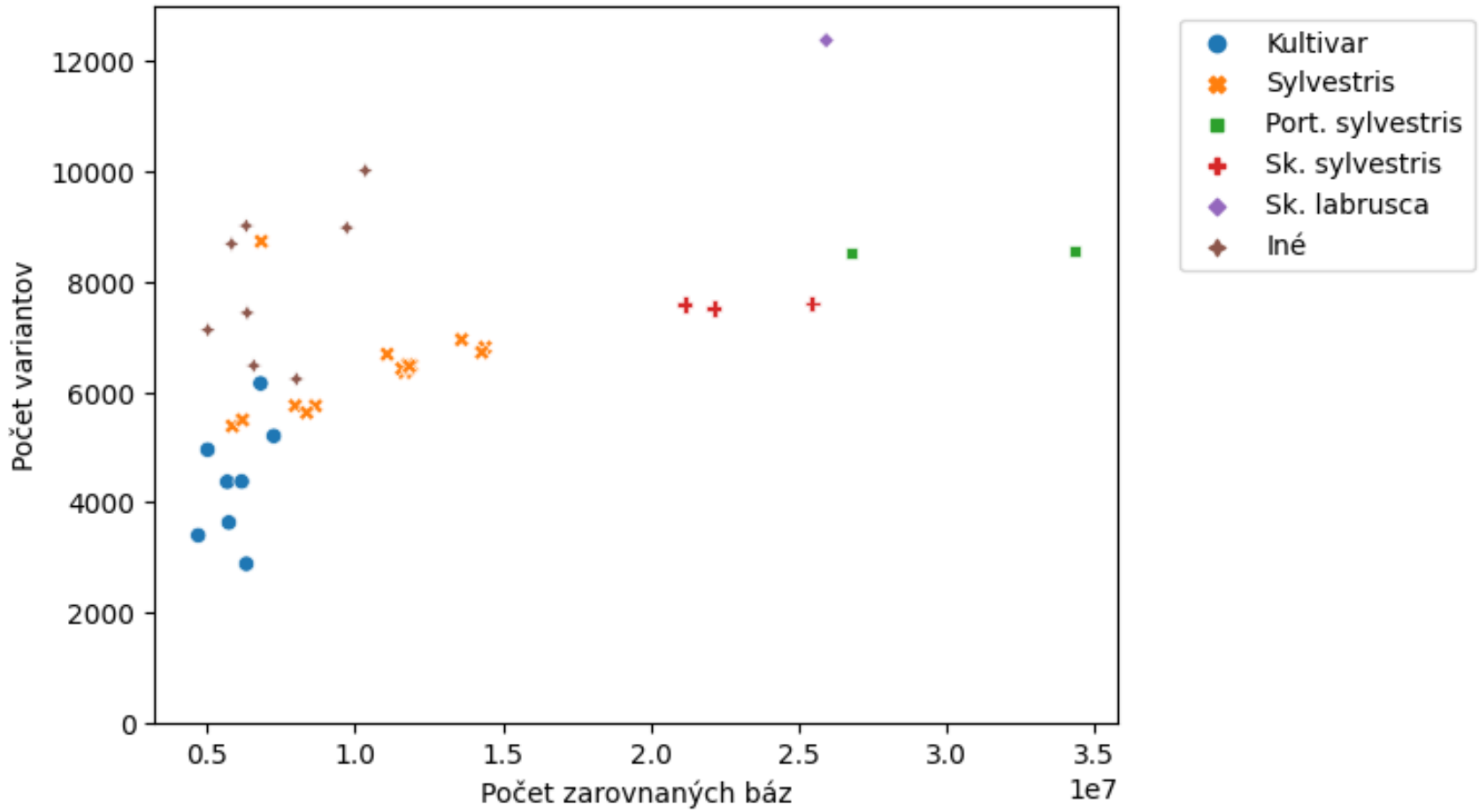
Po filtrácii

Hromadné spracovanie

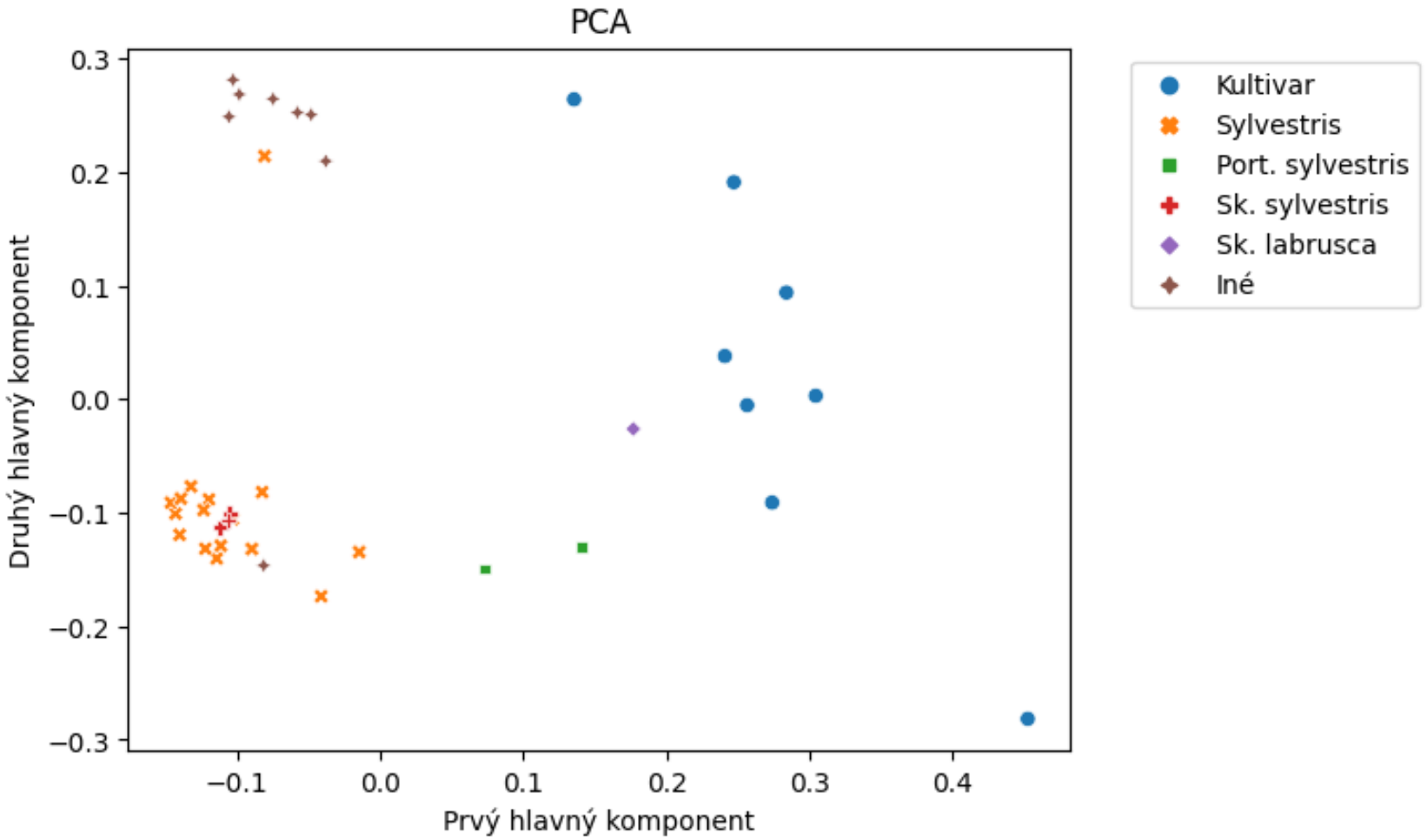
- Systém Snakemake
- Podklad pre automatizáciu
- Priečinková štruktúra
- Zoznam variantov (VCF) pre každú skúmanú vzorku
- Analýza výsledkov

Výsledky

- Bioinformatický postup
- Spracovanie 38 vzoriek
- 3 mld. čítaní a 483 mld. nukleotidov
- 44 miliónov variantov s kvalitou aspoň 30
- Ďalšie analýzy nájdených variantov
  - Závislosť medzi počtom zarovnaných báz a počtom variantov
  - Analýza hlavných komponentov (PCA)
  - Jaccardova miera podobnosti



Závislosť počtu zarovnaných báz a počtu variantov

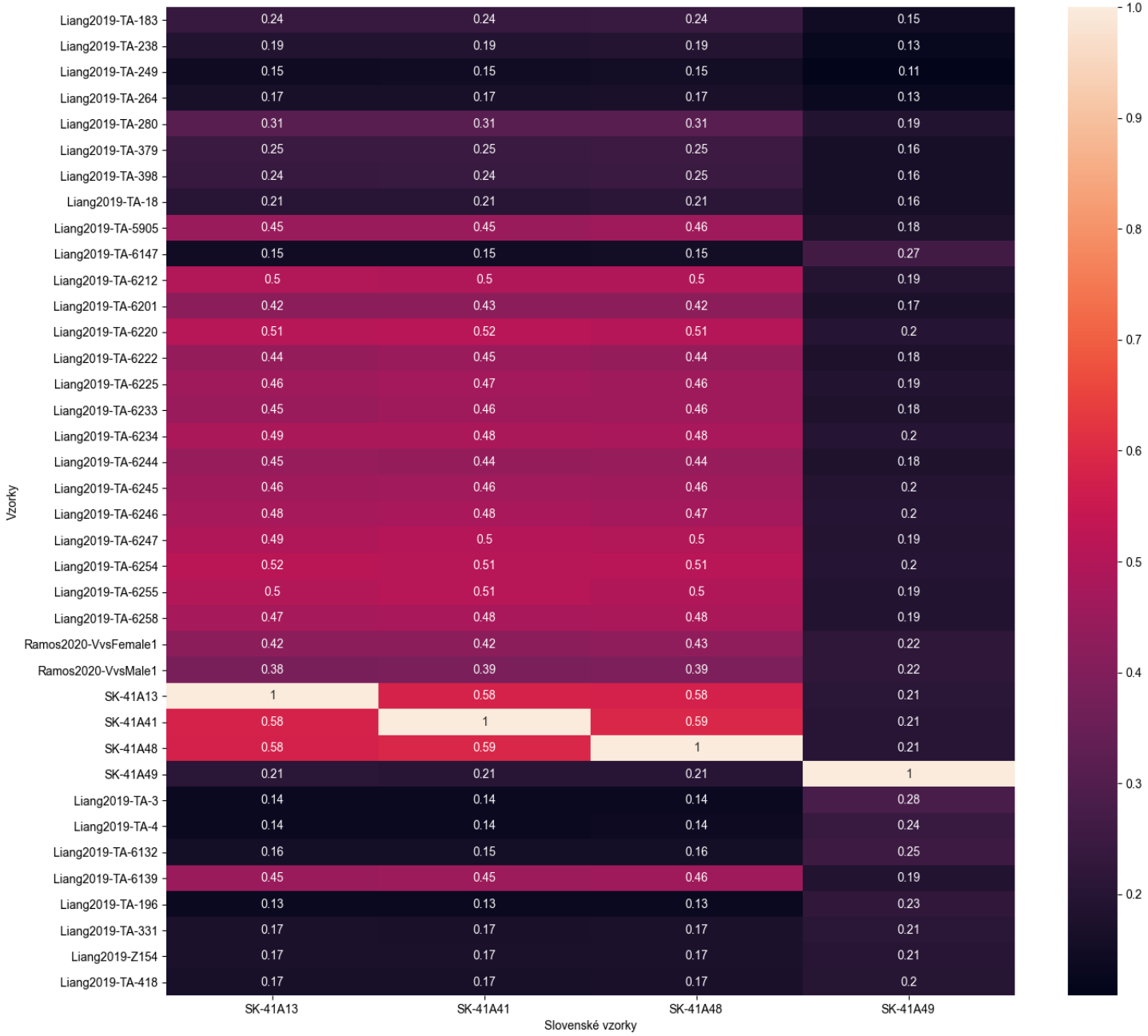


Analýza hlavných komponentov

Jaccardova miera podobnosti

- Vyjadrenie podobnosti dvoch množín
- Podiel prieniku a zjednotenia množín
- Množiny variantov – VCF súborov
- Medzi každou slovenskou a každou skúmanou vzorkou

Jaccardova miera podobnosti



Návrhy do budúcna

- Doplnenie databázy ďalšími vzorkami
- Štúdium fylogenetických vzťahov
- Voľba iných nástrojov, nastavení

Ďakujem za pozornosť

Otázky

Pri porovnaní zarovnaní vybraných súborov voči referenčnému genómu Pinot Noir genóm 12X a genóm 12X.v2 bolo v staršej verzii (12X) zarovnaných o niečo viac čítaní ako v novšej (12Xv2). Ako si vysvetľujete tento rozdiel?

	Genóm 12X	Genóm 12X.v2
Vzorka	BWA-MEM	
V. v. sylvestris Liang2019-TA-6201	98,31 %	97,77 %

Odpoveď:

- Vybrať čítania, ktoré sa nezarovnali a pozrieť, kam sa čítania v druhom genóme zarovnali

Otázky

Najvyššie percento zarovnaných čítaní k referenčnému genómu bolo pozorované u vzorky Liang2019-TA-264. Označený kultivar Pinot Noir Liang2019-TA-379 totožný s referenčnou odrodou vykazoval až druhé najvyššie percento čítaní k referenčnému genómu. Najmenej variantov bolo nájdených vo vzorke Liang2019-TA-249. U ktorej zo vzoriek (Liang2019-TA-264 *vs.* Liang2019-TA-249) je vyššia pravdepodobnosť usudzovať, že sa jedná o odrodu Pinot Noir?

Otázky

Vzorka	Čítania	Zar. Čítania	Chyb.	Var.
Liang2019-TA-249	46743	99,03 %	2,10 %	2890
Liang2019-TA-264	42253	99,21 %	2,03 %	3638
Liang2019-TA-379	43838	99,16 %	1,75 %	4381

Odpoveď:

- Nevieme s istotou povedať na základe štatistík
- PCA analýza
- Ďalšie metódy a parametre

Otázky

Najvyšší počet všetkých variantov bol nájdený u vzorky SK-41A49 *V. labrusca*. Tento druh viniča je zaujímavý najmä pre odolnosť voči fyloxére. Na druhej strane, senzorické vlastnosti vína vyrobeného z tejto odrody nie sú najlepšie hodnotené. Aké nástroje/prístup by ste navrhovali použiť na odhalenie najviac variabilných oblastí genómu *V. labrusca* v porovnaní s *V. v. sylvestris*?

Odpoveď:

- Vyfiltrovanie rozdielnych variantov v oboch vzorkách
- Skúmanie rozdielov v génových oblastiach
- Pozrieť sa na nesynonymné mutácie

Otázky

Mohla by autorka stručne zhrnúť metódy, ktoré sa používajú na odlíšenie biologických variantov od technických artefaktov?

Odpoveď:

- Predspracovanie dát – orezanie čítaní, odfiltrovanie čítaní s nízkou kvalitou, odstránenie adaptérov
- Odstránenie duplikátov
- Filtrovanie variantov – odlišnosť kvality referenčnej a alternatívnej bázy, podiel kvality a pokrytia genómu