České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra kybernetiky

# Kompenzace vlivu matoucích faktorů při klasifikaci RNASeq dat

BAKALÁŘSKÁ PRÁCE

Vypracovala: Karina Balagazova
Vedoucí práce: doc. Ing. Jiří Kléma, Ph.D.
Rok: 2022

# ZADÁNÍ BAKALÁŘSKÉ PRÁCE

## I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Balagazova** Jméno: **Karina** Osobní číslo: **483729**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra kybernetiky**

Studijní program: **Otevřená informatika**

Specializace: **Základy umělé inteligence a počítačových věd**

## II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

**Kompenzace vlivu matoucích faktorů při klasifikaci RNASeq dat**

Název bakalářské práce anglicky:

**Adjustment for the Confounding Factor Influence in the RNASeq Data Classification**

Pokyny pro vypracování:

1. Seznamte se s problematikou tvorby medicínských molekulárních klasifikátorů vychazejících z RNASeq dat (povaha RNASeq dat, problém přeučení plynoucí z malého počtu vzorků a velkého počtu příznaků).
2. Proveďte rešerši statistických metod kompenzace vlivu matoucích faktorů. Zaměřte se především na studie, ve kterých není možné design řídit.
3. Navrhněte metodu vhodnou pro klasifikaci RNASeq dat dodaných vedoucím práce. Zaměřte se na ověření možností kompenzaci vlivu matoucích faktorů, v daném případě podávání léků.
4. Navrhněte generátor umělých RNASeq dat umožňující nastavit klíčové parametry (počet vzorků, jejich distribuci mezi třídami, počet transkriptů, velikosti efektu u jednotlivých transkriptů pro různé faktory).
5. Vyhodnoťte použitelnost navržené metody, tj. porovnejte klasifikaci s a bez kompenzace vlivu léků. Vyhodnocení proveďte na reálných i umělých datech.

Seznam doporučené literatury:

[1] Brookhart, M. Alan, et al. "Confounding control in healthcare database research: challenges and potential approaches." Medical care 48.6 0 (2010): S114.
[2] Christakoudi, Sofia, et al. "Development and validation of the first consensus gene-expression signature of operational tolerance in kidney transplantation, incorporating adjustment for immunosuppressive drug therapy." EBioMedicine 58 (2020): 102899.
[3] Goksuluk, Dincer, et al. "Mlseq: Machine learning interface for rna-sequencing data." Computer methods and programs in biomedicine 175 (2019): 223-231.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

**doc. Ing. Jiří Kléma, Ph.D., Intelligent Data Analysis FEL**

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **07.01.2021** Termín odevzdání bakalářské práce: **04.01.2022**

Platnost zadání bakalářské práce: **30.09.2022**

| doc. Ing. Jiří Kléma, Ph.D. | prof. Ing. Tomáš Svoboda, Ph.D. | prof. Mgr. Petr Páta, Ph.D. |
|---|---|---|
| podpis vedoucí(ho) práce | podpis vedoucí(ho) ústavu/katedry | podpis děkana(ky) |

## III. PŘEVZETÍ ZADÁNÍ

Studentka bere na vědomí, že je povinna vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. _____
Datum převzetí zadání

_____
Podpis studentky

**Prohlášení**

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 4.1.2022 .....

Karina Balagazova

**Poděkování**

Zde bych ráda poděkovala všem, kdo mě při psání této práce podporoval. Z akademické půdy je to v první řadě vedoucí práce doc. Ing. Jiří Kléma, Ph.D., kterému děkuji za neocenitelné rady, připomínky a trpělivost při vedení práce. Dále bych chtěla poděkovat svým rodičům a sestrám, jejichž podpora mi nesmírně pomáhala během celého studia.

# Abstrakt/Abstract

Analýza dat genové exprese slouží významným zdrojem informace o chování biologických systémů. Jednou z cest jak tuto informaci získat je kvantifikování a měření genové exprese pomocí technologie RNA sekvenování (RNA-Seq). V lékařských studiích však často existují omezení, spojené například s etickými důvody nebo se vzácnosti některých onemocnění. Není proto vždy možné ve studiích zkoumat randomizované skupiny s dostatečným počtem pacientů pro řádné posouzení sledovaného vztahu. Vznikají tak tzv. *matoucí faktory*, které mohou výsledky studií zkreslovat.

Tato práce stručně seznamuje s technologií RNA sekvenování a popisuje základní vlastnosti získávaných touto technologií dat. Dále práce pojednává o problematice matoucích faktorů a představuje rešerši metod pro kompenzaci jejich vlivu. Zvolená metoda pak bude aplikována při klasifikaci poskytnutých a uměle generovaných RNA-Seq dat, na základě čehož bude vyhodnocena její použitelnost.

**Klíčová slova:** matoucí faktor, RNA sekvenování, strojové učení, genová exprese

Analysis of gene expression data serves as an essential source of information about the behavior of biological systems. One way to obtain this information is to quantify and measure gene expression using RNA sequencing (RNA-Seq) technology. However, there can often be limitations in medical studies due to ethical considerations or the rarity of certain diseases. Therefore, it is sometimes impossible to examine well-randomized groups with sufficient patients to assess the observed relationship properly. It causes the appearance of *confounding factors* that may bias the results of studies.

This thesis briefly introduces RNA sequencing technology and describes the basic properties of the RNA-Seq data. Furthermore, the thesis discusses the issue of confounding factors and presents a survey of methods to adjust for their influence. The selected method will then be applied in the classification of real and artificially generated RNA-Seq data, which will evaluate applicability of suggested method.

**Keywords:** confounder, RNA sequencing, machine learning, gene expression

# Obsah

# Seznam použitých zkratek

**DNA** Deoxyribonukleová kyselina (Deoxyribonucleic Acid)

**RNA** Ribonukleová kyselina (Ribonucleic Acid)

**RNA-Seq** RNA sekvenování (RNA sequencing)

**CR** Chronická rejekce (Chronical Rejection)

**STA** Stabilní stav

**IS** Imunosupresivní léky

**PS** Propensity Score

**PSM** Propensity Score Matching

**ROC** Receiver Operating Characteristic

**AUC** Plocha pod křivkou (Area Under Curve)

**OT** Operační tolerance (Operational Tolerance)

**SVM** Metoda podpůrných vektorů (Support Vector Machine)

# Seznam obrázků

# Úvod

Analýza lékařských dat slouží jako významný zdroj informací o chování biologických systémů. Jednou z cest, jak tuto informaci získat, je měření **genové exprese**, která je vyjádřením informace uložené v genu (DNA) do struktury molekul. Zásadní úlohou, kde se tyto informace využívají, je klasifikace zdravotního stavu pacienta a detekce signatur (biomarkerů) – malých podmnožin genů, umožnující rozlišit mezi třídami vzorků (například mezi zdravými a nemocnými lidmi). Kvalita těchto signatur a z nich odvozených klasifikátorů je ale ovlivněna celou řadou okolností – rozsahem naměřených dat, kvalitou použité technologie měření nebo způsobem návrhu klinického experimentu. Tato práce se zaměřuje na situace, kdy možnosti ovlivnit klinický experiment jsou omezené.

V oblasti zpracování lékařských dat v praxi není možné pracovat s jinak častým randomizovaným designem klinických studií, který minimalizuje vliv všech proměnných kromě třídy samotné, tj. není možné pracovat s ideálním statistickým vzorkem. Z etických důvodů nelze nemocným nepodávat léky a naopak není vhodné podávat léky prokazatelně zdravým pro jejich vedlejší účinky. Ve studiích se tedy přirozeně vyskytuje nežádoucí šum formou **matoucích faktorů**, které na jedné straně souvisí s vlastnostmi třídy, ke které vzorek patří, avšak současně ovlivňují i samotná data genové exprese. Tato práce se soustředí na to, jak zmíněné matoucí faktory kompenzovat.

Data, která se v práci prozkoumávají, jsou poskytnuta z výzkumného projektu, jehož cílem je vyšetřit transkriptomický profil příjemců ledvin. Po transplantaci ledviny může docházet k odmítnutí orgánu organismem pacienta, neboť imunitní systém identifikuje transplantační štěp jako cizí těleso a atakuje ho. K potlačení reakce organismu se používají imunosupresivní léky, které omezují nebo blokují reakce imunitního systému, zároveň ale ovlivňují genovou expresi a nesou velké riziko vedlejších účinků.

Jednou ze současných úloh v dané oblasti je naučit se predikovat pacienty s operační tolerancí pro zjednodušení řízení imunosupresivní terapie. Operační tolerance je vzácný jev, vyskytující se u některých pacientů, kteří z různých důvodů přestali užívat imunosupresivní léky poté, co podstoupili transplantaci ledvin. Přes vysazení léků u nich nedochází k odmítnutí transplantačního štěpu a transplantovaný orgán pracuje správně. Detekce operační tolerance z dat genové exprese je ale nesnadnou úlohou. Imunosupresivní léky se v daném případě dají považovat za ma-

toucí faktory. Operačně toleranční pacienti imunosupresiva nedostávají a jejich data genové exprese proto zůstávají neovlivněná, naopak v datech pacientů, kteří tyto léky dostávali, se vliv matoucích faktorů vyskytuje. Z toho vyplývá, že v takových datech je těžké vypozorovat skutečný vztah mezi skupinami pacientů a rozlišit je. Aby bylo možné odlišit operačně toleranční pacienty, je nutné nejprve odfiltrovat vliv imunosupresiv, tj. zaměřit se na kompenzaci nežádoucích matoucích faktorů.

Práce na začátku stručně seznamuje s použitou technologií měření. **RNA sekvenování** je vysoce výkonná sekvenační technologie nové generace, která se používá pro kvantifikování a měření genové exprese. Získáváme kompletní informaci o transkriptomu, což znamená, že výstupem procesu je objemná datová sada, která obsahuje až desítky tisíc naměřených příznaků.

Vzhledem k velikosti dat, kterou člověk není schopen zpracovat sám, hraje v medicině použití statistiky a umělé inteligence velkou roli. Nové technologické pokroky umožňují pomocí metod strojového učení klasifikovat onemocnění, předpovídat zdravotní stav pacientů po operaci a určovat nejvhodnější možnosti léčby. Pro klasifikaci poskytnutých dat se v práci budou používat klasické metody strojového učení, jako je logistická regrese a SVM. Aby se dalo vyhnout problému přeučení, práce se nezbytně zaměřuje i na předzpracování dat a použití metod pro omezení velkého počtu příznaků.

Dále v práci budou představené současné metody pro řešení problému matoucích faktorů. Z důvodu omezení popsaných výše jsou vhodnou volbou statistické metody kompenzace. Jednou z takových metod je naučení se vlivu faktorů pomocí **multivariačního regresního modelu**. Při použití této metody ale mohou vzniknout podstatné problémy, konkrétně problém překompenzace, který vyplývá z nedostatečného počtu vzorků a má za následek nesprávný odhad vlivu jednotlivých faktorů.

Klíčovou otázkou této práce je rozpoznat, za jakých podmínek k této překompenzaci dochází. Abychom mohli simulovat širokou škálu experimentálních podmínek, navrhneme generátor umělých dat, který je založen na analýze skutečné datové sady. Každý transkript se generuje na základě odhadů z reálných dat distribucí středních hodnot a jejich vztahu s rozptylem, jakož i na základě vybraných hodnot vlivu tříd a vlivu samotného matoucího faktoru. Volba takových parametrů umožní podrobněji prostudovat funkčnost navržené metody, protože se skutečnými daty nemáme možnost dané vlivy mezi sebou oddělit. Tento generátor také umožňuje volit takové parametry, jako je počet vzorků, jejich rozdělení mezi třídami, počet transkriptů a velikost efektů jednotlivých transkriptů.

Dalším důležitým cílem této práce je upozornit na problémy klasifikace operační tolerance z dat genové exprese plynoucí z vlivu matoucích faktorů. Práce směřuje ke zjištění prahů navržené metody pro kompenzaci matoucích faktor a k důkladnému posouzení funkčnosti kompenzace. Jedním z klíčových testovaných parametrů je vhodný počet vzorků potřebný pro úspěšné zmenšení vlivu daných faktorů pro další výzkumy. Tyto cíle jsou naplněny v experimentální části práce a shrnuty v závěru.

# Kapitola 1

# Genová exprese a RNA-Seq data

Tato kapitola obsahuje základní pojmy, které slouží k lepšímu pochopení vzniku RNA-Seq dat. Kapitola začíná úvodem do molekulární biologie a vysvětlením pojmu genová exprese. Tato část především vychází z knihy [1]. Dále následuje část, seznamující se samotnou technologii RNA-Seq a také vznikem i povahou RNA-Seq dat. Druhá polovina kapitoly vychází z článku [2].

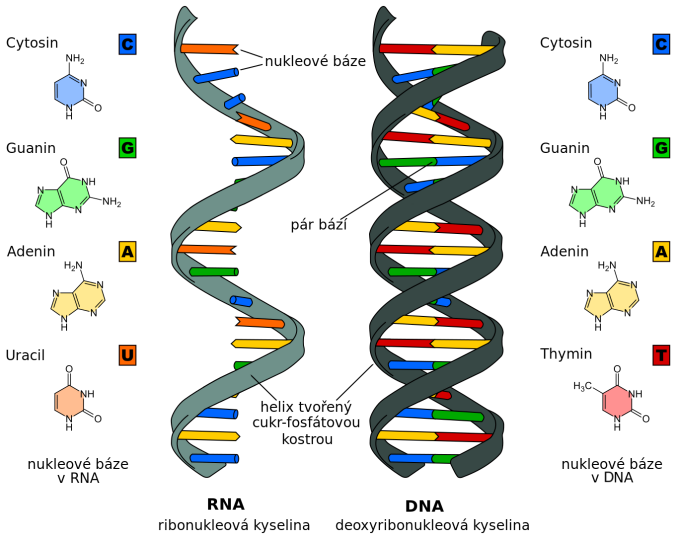
## 1.1 Nukleové kyseliny

Základní stavební jednotkou všech buněk je nukleotid. Nukleotidy tvoří nukleové kyseliny, mezi které patří známé deoxyribonukleová (DNA) a ribonukleová (RNA).

Tyto kyseliny jsou lineární polymery, složené ze sekvence čtyř různých nukleotidů, spojených fosfodiesterovými vazbami. Nukleové báze se dělí na báze purinové (A - adenin, G - guanin) a báze pyrimidinové (C- cytosin, T - thymin, U - uracil).

Struktura DNA se skládá ze dvou vláken a tvoří dvoušroubovici. Každé vlákno je sestaveno z nukleotidů (A, C, G, T), připojených k deoxyribóze a fosfátové skupině. Důležitou vlastností vláken DNA je komplementarita – purinové báze jsou vždy komplementární s pyrimidinovými – adenin se páruje s thyminem a guanin s cytosinem. Jak ukazuje obrázek 1.1, na rozdíl od DNA, molekula RNA se v buňkách vyskytuje jako jednořetězcová šroubovice. Chemicky se RNA od DNA liší hlavně ve dvou ohledech. Ribonukleotidy v RNA obsahují cukr ribózu, namísto deoxyribózy v DNA. Stejně jako DNA, molekula RNA je sestavena z nukleotidů, ovšem místo thyminu (T) obsahuje bázi uracil (U).

Proces přenosu genetického materiálu z nukleových kyselin na proteiny popíšeme v následující sekci.

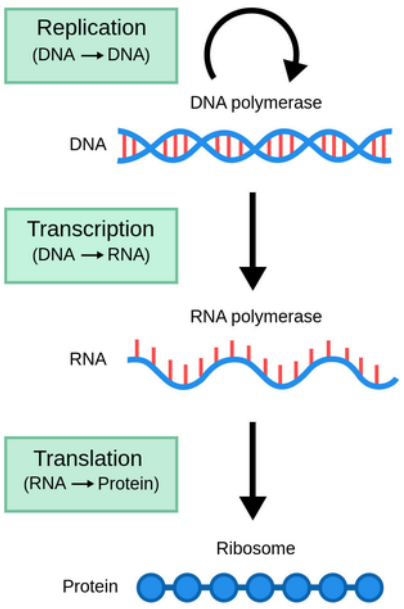



**Obrázek 1.1:** Struktura DNA a RNA molekul [3]

## 1.2 Centrální dogma

Centrální dogma molekulární biologie je zobecňující pravidlo, které vysvětluje proces přenosu genetické informace v biologických systémech. Podle daného konceptu platí, že se genetická informace může přenášet mezi nukleovými kyselinami a z nukleových kyselin na protein, ale ne v opačném směru.

Na obr. 1.2 je ukázán přenos genetického materiálu v biologických systémech. Obecně existují 3 typy přenosu informace – replikace, transkripce a translace.




**Obrázek 1.2:** Centralni dogma molekulární biologie (DNA → RNA → protein) [4]

Replikaci DNA používají všechny organismy pro přesné duplikování své DNA před každým buněčným dělením pro předání genetického materiálu svým potomkům. DNA se replikuje tak, že se přeruší vazba mezi jejími dvěma vlákny a každé vlákno vytvoří komplementárním párováním odpovídající vlákno, znovu se spojí a znovu stočí. V následující sekci spolu s zavedením pojmu genová exprese popíšeme transkripci a translaci.

## 1.3 Genová exprese

Terminologie:

- **Gen** je určitý úsek DNA (sekvence nukleotidů).
- **Genom** je kompletní sada všech genů tj. souhrn veškeré genetické informace zapsané v DNA uvnitř buněk.

Genová exprese je biologický proces, ve kterém se instrukce uložené v genu převádějí na funkční produkt. Funkčním (konečným) produktem genové exprese slouží buď protein (vzniklý z kódující RNA), nebo občas i samotná RNA. Kódující RNA se nazývá mRNA (messenger RNA) a obsahuje v sobě informaci pro řízení syntézy proteinů. Z nekódujících RNA protein nevzniká, avšak jsou taky důležité a slouží jako komponenty pro nejrůznější procesy v buňce. Mezi nekódující RNA patří např. rRNA (ribozomální RNA), tRNA(transferová RNA), snRNA a jiné.

DNA v genomech neřídí syntézu proteinu sama, ale využívá RNA jako zprostředkující molekulu. Obecně tak genová exprese obsahuje 2 klíčové fáze – transkripce a translace – procesy, kterými buňky čtou (exprimují) genetické instrukce ve svých genech.

### 1.3.1 Transkripce (DNA → RNA)

Ve chvíli, když buňka potřebuje protein, sekvence nukleotidů příslušné části DNA se nejprve kopíruje do RNA. Daný proces se nazývá transkripce a je prvním krokem v expresi genu. Informace do RNA se zapisuje stejným jazykem sekvence nukleotidů jako v DNA.

Transkripce začíná rozvinutím části dvoušroubovice DNA. Jedno z řetězců dvoušroubovice DNA pak slouží jako šablona pro syntézu molekuly RNA. Dále komplementárním párováním bází se sekvence z řetězce DNA přepisuje do řetězce RNA. Pokud shoda při párování vstupujícího nukleotidu a šablony DNA bude správná, příchozí ribonukleotid se spojuje s rostoucím řetězcem RNA. Řetězec RNA vzniklý transkripcí - transkript - se tedy prodlužuje po jednom nukleotidu a má sekvenci nukleotidů, jež je přesně komplementární k řetězci DNA použitému jako šablona.

Po zastavení transkripce se jak šablona DNA, tak hotová molekula mRNA uvolňují. Díky rychlému uvolnění při syntéze řetězce RNA z DNA lze ze stejného genu v

relativně krátké době vytvořit mnoho kopií RNA (u eukaryotů může být z jednoho genu za hodinu syntetizováno více než tisíc transkriptů).

### 1.3.2 Translace (RNA → protein)

RNA transkripty se dále používají jako "šablona" pro vytváření proteinů, které vykonávají buněčné funkce. Proces přenosu genetické informace z RNA do sekvencí aminokyselin proteinů se nazývá translace. Jakmile je protein vytvořen, říká se, že je gen exprimován.

### 1.3.3 Různé úrovně exprese

Každý gen může exprimovat (procházet fází transkripce a translace) s různou účinností, což umožňuje buňce produkovat rozdíly v množství vytvořených proteinů v buňce. To znamená, že se ve výsledku vytvoří velké množství proteinů jednoho typu a malé množství jiného. Navíc buňka může měnit (nebo regulovat) expresi každého ze svých genů podle momentálních potřeb - nejzřejměji řízením produkce své RNA.

Pokud buňky během procesu buněčného dělení trpí nemocemi (např. rakovinou), které způsobují mutace v genech, nekontrolovatelné chování genu se přenese i na dceřiné buňky. Dalším příkladem je to, že při podávání pacientovi léků, které vyvolávají změny v genech, dochází k ovlivnění určitých hodnot exprese genů, které lze zjistit pomocí sledování RNA.

Měření mRNA namísto proteinů je standardní technikou měření exprese genů. Důvodem použití sekvencí mRNA je to, že hybridizují se svými komplementárními sekvencemi RNA nebo DNA, zatímco u proteinů tato vlastnost chybí. Úroveň genové exprese představuje množství RNA produkované v buňce za různých biologických stavů.

### 1.3.4 DGE

Analýza zaměřená na identifikaci genů s různými úrovněmi exprese mezi několika skupinami nebo biologickými stavy se nazývá analýza diferenciální genové exprese (DGE). Pro tento problém byla vyvinuta řada nástrojů, které se pomocí vhodných statistických testů specializují na rozhodování zda geny nejsou odlišné.

### 1.3.5 Použití dat genové exprese

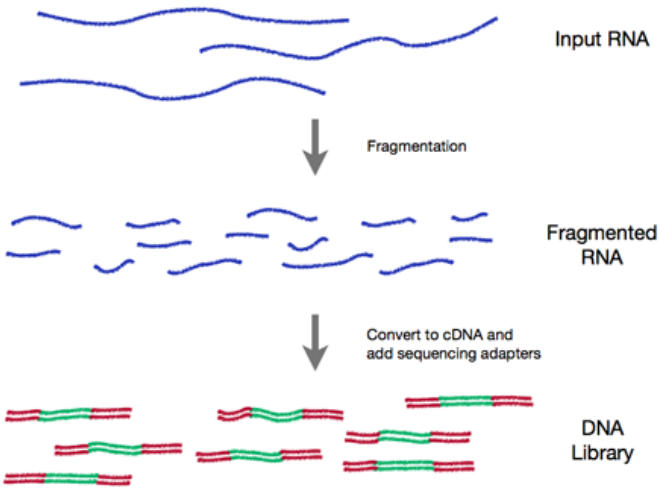
Oblast použití dat genové exprese je velmi rozsáhlá a je možné z nich získat celé spektrum cenných informací o funkci buněk a organismu. Genová exprese je v každém buněčném stadiu vývoje a za různých podmínek odlišná a může být ovlivněna

řadou okolností, např. léčbou nebo zdravotním stavem pacienta. Lze například pozorovat a měřit reakci jednotlivých buněk na léčbu nebo předpovídat reakci pacienta na léky. Je proto důležité umět kvantifikovat a měřit genovou expresi. Jednou z možných technologií pro měření a kvantifikování genové exprese je RNA sekvenování, zkráceně RNA-Seq.

Jednou ze zásadních úloh s využitím dat genové exprese je identifikace signatur (malé podmnožiny genů) a klasifikace dat pro různé účely.

## 1.4 RNA-Seq

RNA sekvenování (RNA-Seq) [5] je vysoce výkonná sekvenační technologie, která se používá pro kvantifikování a měření genové exprese. Hlavním cílem daného procesu je dozvědět se, jaké množství jednotlivého genu/transkriptu obsahuje vzorek. Pro dosažení výsledků sekvenování prochází několika fázemi (obr. 1.3).




**Obrázek 1.3:** RNA-Seq workflow [6]

Před začátkem samotného procesu probíhá příprava dat a návrh klinického experimentu. Jedná se o poměrně důležitý krok, neboť pokud není návrh klinické studií dobře promyšlený nebo v studii vyskytují nějaká omezení, mohou se objevit nežádoucí proměnné, např. matoucí faktory, které popíšeme v kapitole 2.

Samotný proces RNA-Seq začíná odebráním RNA vzorků ze zkoumané buňky nebo tkáně. Každý ze vzorků obsahuje informaci o transkriptomu (kompletní sada všech RNA transkriptů v buňce v daném stádiu vývoje a za daných podmínek). Vzhledem k ochotě získat kvalitní a neporušené RNA, součástí tohoto procesu je izolace a purifikace (přečištění) RNA.

Většina současných sekvenačních platforem je schopna poskytovat pouze relativně krátké sekvenační čtení. Proto proces zahrnuje fragmentační krok, který slouží

ke zlepšeni pokrytí transkriptomu sekvencí. Fragmentované transkripty se pak konvertují do cDNA (komplementární DNA) – molekula DNA vytvořená jako kopie mRNA. Existuje také možnost reverzního přepisu transkriptu RNA a fragmentace výsledné cDNA. Dalším krokem je příprava sekvenační knihovny, což zahrnuje nezbytné přidání speciálních DNA adaptérů (konstantní sekvence, které obklopují cDNA) a amplifikaci DNA pro sekvenování.

Knihovna je poté sekvenována, čímž dostáváme tzv. čtení (angl. read) – přečtená sekvence nukleotidů. Čtení se následně mapují na referenční genom/transkriptom, kvantifikují se a dostáváme tím RNA-Seq data [7], které určují kolikrát daný transkript byl sekvenován. Výsledky se pak dají interpretovat pomocí matice čtení.

### 1.4.1 Matice čtení (read count matrix)

Matice čtení (read count matrix) (obr. 1.4) je matice, jež shrnuje expresi na genové úrovni v každém vzorku datové sady. Taková matice má dimenzi $p \times n$, kde $p$ je počet příznaků (geny, transkripty) a $n$ je počet vzorků. Elementy této matice jsou nezáporné celočíselné hodnoty, tj. diskrétní proměnné. Každá taková hodnota představuje celkový počet přečtených sekvencí, které pocházejí z konkrétního genu ve vzorku. Geny a transkripty se obvykle označují pomocí Ensembl ID [8].

| | OT_05 | OT_17 | OT_18 | OT_21 | OT_23 | OT_33 | OT_34 |
|---|---|---|---|---|---|---|---|
| ENST00000390396.1 | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| ENST00000390400.2 | 6 | 12 | 4 | 2 | 8 | 0 | 2 |
| ENST00000390468.1 | 4 | 7 | 4 | 4 | 5 | 5 | 2 |
| ENST00000390424.2 | 0 | 5 | 3 | 0 | 8 | 6 | 0 |
| ENST00000390463.3 | 0 | 3 | 2 | 1 | 2 | 1 | 0 |
| ENST00000390440.2 | 2 | 5 | 4 | 1 | 5 | 8 | 2 |
| ENST00000390436.2 | 4 | 5 | 12 | 4 | 6 | 11 | 7 |
| ENST00000390435.1 | 3 | 12 | 12 | 0 | 14 | 7 | 1 |
| ENST00000633466.1 | 3 | 12 | 16 | 3 | 16 | 7 | 3 |
| ENST00000611462.1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 |
| ENST00000612787.1 | 9 | 8 | 11 | 2 | 5 | 5 | 8 |
| ENST00000390451.2 | 7 | 0 | 6 | 2 | 0 | 5 | 3 |

**Obrázek 1.4:** Příklad read count matice

## 1.5 Povaha RNA-Seq dat

Hlavním cílem úspěšného experimentu je schopnost správně odhadnout a parametrizovat rozdělení dat, což pomůže identifikovat rozdíly genové exprese mezi dvěma pozorovanými biologickými stavy. V případě RNA-Seq dat rozdělení vypovídá o tom, jaké množství transkriptů obsahuje jeden vzorek.

Jak již bylo zmíněno v sekci 1.4, výstupem sekvenování jsou diskrétní hodnoty. V dané sekci ukážeme typy rozdělení, kterými by se daly dané hodnoty popsat.

### 1.5.1 Poissonovo rozdělení

Nejjednodušší způsob pro popsání RNA-Seq dat je Poissonovo rozdělení [9, 10]. Ve statistice Poissonovo rozdělení vyjadřuje počet výskytů událostí v určitém intervalu. Parametr rozdělení $\lambda > 0$ udává střední počet událostí za časovou jednotku. Střední hodnota daného rozdělení se rovná odchylce.

Pravděpodobnostní funkce je tvaru:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{1.1}$$

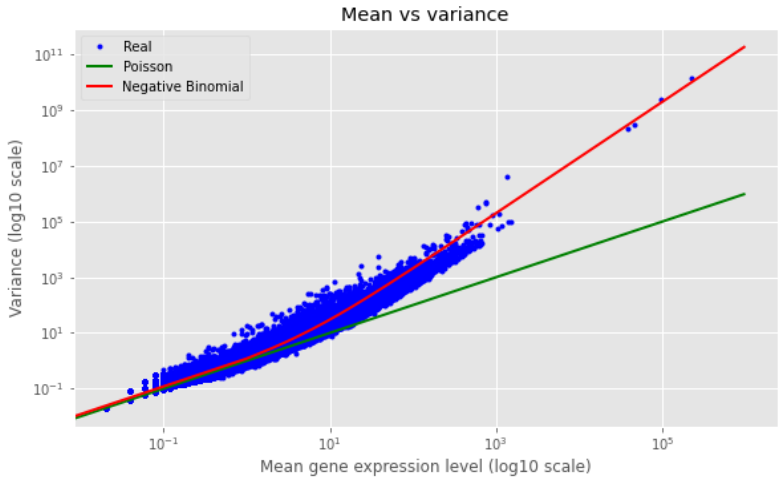
kde $k$ je počet výskytů událostí, $\lambda$ je střední počet událostí.

Často dané rozdělení slouží k popisu rozdělení řídkých jevů ve velké populaci, což je přesně případ RNA-Seq dat. Jednotlivá čtení by v tomto případě hrála roli události, transkripty - roli časového/prostorového intervalu. Každý RNA-Seq vzorek by se pak bral jako pozorování náhodné veličiny.

Ukazuje se ale, že daná distribuce je vhodná pouze pro variabilitu dat spojenou se vzorkováním stejné populace, tj. pro data se stejnou variabilitou přes všechny transkripty. Pro případ kdy mezi replikáty existují rozdíly, Poissonův model bude mít tendenci podceňovat rozptyl a veškeré rozdíly, které jsou pozorované, budou přeceněné.

### 1.5.2 Negativně binomické rozdělení

Některé geny mohou kolísat více či méně kvůli své přirozené povaze. Vzhledem k biologické variabilitě dat, kterou Poissonovo rozdělení není schopné popsat, lepší volbou pro modelování RNA-Seq dat bude negativně binomické rozdělení [11]. Dané rozdělení je obecnějším případem Poissonova rozdělení. Jak je vidět na obrázku 1.5, u negativně binomického rozdělení je patrna overdisperze (tj. rozptyl je větší než střední hodnota).




**Obrázek 1.5:** Vztah mezi střední hodnotou a rozptylem u RNA-Seq dat.

Obecně negativně binomické rozdělení popisuje kolik úspěšných opakování se musí provést, aby se dosáhlo k-té chyby.

Pravděpodobnostní funkce je tvaru:

$$f(k; n, p) = \binom{k+n-1}{n-1}(1-p)^k p^n, \tag{1.2}$$

kde $n > 0$ je počet úspěchů, $k > 0$ je počet neúspěchů a $0 < p < 1$ je pravděpodobnost úspěchu.

Vztah pro střední hodnotu $\mu$ a rozptyl $\sigma^2$ se vyjadřuje jako:

$$\mu = \frac{pn}{1-p}, \tag{1.3}$$

$$\sigma^2 = \frac{pn}{(1-p)^2}. \tag{1.4}$$

Další běžnou parametrizací negativně binomického rozdělení je parametrizace, popisující průměrný počet neúspěchů $\mu$ potřebných k dosažení $r$-tého úspěšného opakování:

$$p = \frac{\mu}{n+\mu}. \tag{1.5}$$

Počet úspěchů $r$ se taky dá nahradit parametrem $\alpha$, který vypovídá o variabilitě dat a popisuje vztah mezi střední hodnotou $\mu$ a rozptylem $\sigma^2 = \mu + \alpha\mu^2$ [12].

Dále lze odvodit, že

$$p = \frac{\mu}{\sigma^2}, \tag{1.6}$$

$$n = \frac{\mu^2}{(\sigma^2 - \mu)} = \frac{\mu \cdot p}{(1-p)}. \tag{1.7}$$

Tato parametrizace se v dané práci použije později v sekci 4.1.4 při generování umělých dat.

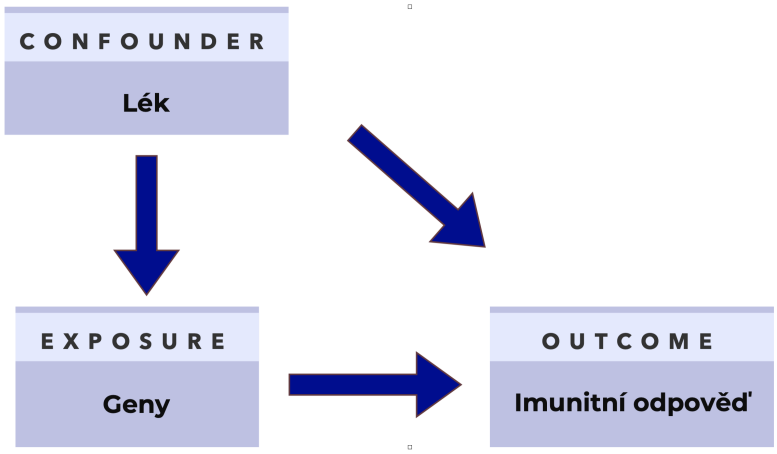
# Kapitola 2

# Matoucí faktory

Zpravidla při interpretaci výzkumných studií před tím, než prohlásíme, že studovaná proměnná, neboli expozice (angl. exposure) a výstup (angl. outcome) jsou mezi sebou v přímém vztahu, je nutné se ujistit, že neexistuje žádný jiný způsob, jakým by se dal daný vztah popsat. Ovlivnit studovanou asociaci může mnoho faktorů, např. náhoda při sběru dat, systematické chyby nebo matoucí faktory, a je velmi důležité umět je rozlišovat [13]. V dané sekci se seznámíme s pojmem matoucí faktor a následně popíšeme jak se s tím dá zacházet. Tato část hlavně vychází z literatury [14, 15].

## 2.1 Úvod

Ve statistice, **matoucí faktor** (angl. confounder, confounding factor) je proměnná, která svou přítomností ovlivňuje expozici a výsledek, přičemž neleží v kauzální cestě mezi nimi, a způsobuje tak falešnou asociaci. Skutečný vztah mezi studovanou proměnnou a výsledkem tedy zůstává neznámý.




**Obrázek 2.1:** Diagram, ukazující vztah mezi expozicí, výslednou proměnnou a matoucím faktorem

Příklad takové asociace je vidět na obrázku 2.1. Představme si, že chceme najít souvislost mezi genovou expresí konkrétního člověka a tím, jak se jeho organismus vypořádá s nemocí. Pro rychlejší zotavení se mu předepisují léky, které pomáhají organismu vyléčit se (tj. ovlivňují outcome), ale současně mají schopnost měnit genovou expresi (tj. ovlivňují expozici). Vzhledem k tomu, že obě studované proměnné jsou ovlivněny matoucím faktorem (léky), nemůžeme tvrdit, že výsledný vztah bude vztahem skutečným.

Jelikož rozlišujeme nemalé množství faktorů a biasů, nikdy se stoprocentní jistotou nevíme, které z těchto proměnných můžeme považovat za matoucí faktory. Doporučuje se při analýze dat experimentovat s různými nastaveními statistických modelů a výsledky porovnávat [13].

Po úspěšném odhalení potenciálního zmatení lze kompenzaci jeho vlivu na studovanou asociaci provést několika způsoby:

- buď před sběrem dat pomocí navržení vhodného designu studie,
- nebo po sběru dat pomocí nástrojů statistické analýzy.

Cílem každého z těchto principů je dosáhnutí stejnorodosti mezi studovanými skupinami. Tyto metody budou podrobněji popsány v následujících dvou sekcích.

## 2.2 Kompenzace při plánování studie

V současné době je téměř nemožné potkat studii s dokonalým designem. Výzkumy zpravidla mají řadu omezení, které mohou být spojené např. s etickými důvody nebo se vzácnosti některých nemocí. I přesto je plánování studie důležitým krokem každého průzkumu, protože správně navržená studie minimalizuje výskyt nežádoucích faktorů. V rámci této sekci budou krátce popsány způsoby, které efektivně minimalizují vliv matoucích faktorů při plánování studií. Mezi nimi hlavně patří randomizace, restrikce a porovnávání (matching).

Při **randomizaci** se pacienti (vzorky) náhodně přiřazují buď do experimentální, nebo do kontrolní skupiny a v ideálním případě by to mělo vést k rovnovážnému rozdělení matoucích faktorů v obojí skupinách. Pro tento způsob se ale předpokládá co největší počet vzorků, neboť s rostoucím počtem vzorků roste i pravděpodobnost toho, že se matoucí faktory ve skupinách rovnoměrně zamíchají. Užitečnou vlastností randomizace je to, že tento způsob umožňuje minimalizovat vliv jak známých, tak neznámých matoucích faktorů.

**Restrikce** je omezení skupiny podle konkrétního rizikového faktoru, např. podle věku. Vybraný faktor bude fixní po dobu studii pro všechny vzorky a proto na výsledek nebude mít žádný vliv. Tento přístup však neřeší to, že výsledky mohou být zmateny i dalšími faktory. Nevýhodou taky je, že tímto omezením se snižuje celkový počet vzorků a bráníme tím extrapolaci výsledků pro jiné skupiny.

**Matching** se typicky používá v case-control studiích a je vhodný např. pro

vyšetřování nemocí. Vytváří se case-control dvojice, kde case je pacient, který je nemocný, a control je pacient se stejnými parametry, který nemoc nemá. Tímto způsobem se matoucí faktor jednoduše kontroluje.

## 2.3 Kompenzace v průběhu analýzy dat

Vzhledem k tomu, že v některých případech nemáme možnost design studie řídit, druhou variantou pro zmenšení vlivu matoucích faktorů jsou statistické metody.

### 2.3.1 Stratifikace

Jednou z nejstarších statistických metod pro kompenzaci matoucích faktorů je stratifikace, která je analogem restrikce. Hlavní myšlenkou této metody je rozdělení dat na několik menších vrstev (tzv. strata) na základě potenciálních matoucích faktorů. Dále se uvnitř každé vrstvy provádí ohodnocení, a jelikož se mezi skupinami uchovává rovnoměrnost, žádný třetí faktor by už nemohl výsledek ovlivnit.

Stratifikace je velmi jednoduchá v používání metoda, ale bohužel má svoje nedostatky. Její hlavním nedostatkem je omezení počtu faktorů, které lze stratifikovat. Nejlépe tedy metoda bude fungovat pouze s malým počtem vrstev a 1-2 matoucí faktory. Pro případy s větším počtem vrstev byla navržena některá zlepšení [16], která ale taky mají své nedostatky.

### 2.3.2 Multivariační regresní modely

Nejčastěji se pro kontrolu vlivu matoucích faktorů používají metody založené na multivariačních modelech. V zasadě se tyto metody neliší od stratifikace, tj. stále chceme zkoumat asociaci faktorů tak, aby potenciální zmatení zůstávalo stejné přes všechny skupiny. Jistou výhodou těchto metod je to, že na rozdíl od stratifikace dokážou ovládat i větší počty potenciálních matoucích faktorů. Mezi tyto modely patří například lineární a logistická regrese.

V dané sekci pro všechny vzorce se bude používat následující konvence: $\beta_1 ... \beta_n$ jsou koeficienty, $y$ je závislá proměnná, $x_1 ... x_n$ jsou nezávislé proměnné.

**Lineární regresní model**
Nejjednodušší formou regrese je klasická lineární regrese, která předpokládá lineární vztah mezi dvěma veličinami. Rovnici regresního modelu lze vyjádřit ve tvaru:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{2.1}$$

**Zobecněné lineární modely (GLM)**
Zobecněné lineární modely jsou rozšířením lineárních modelů s vetší tolerancí pro

různé distribuční charakteristiky proměnných. Předpoklad na linearitu tady už neplatí nutně a rozdělení nemusí být normální. Závislost hodnot vysvětlované proměnné na hodnotách prediktorů popisuje linkovací funkce, která je většinou nelineární.

$$g(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{2.2}$$

kde $g(y)$ je linkovací funkce.

Kanonickou linkovací funkcí pro Poissonův regresní model slouží logaritmus:

$$ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon, \tag{2.3}$$

z čehož odvozíme proměnnou $y$:

$$y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon}. \tag{2.4}$$

Hlavní myšlenka při odstranění vlivu matoucích faktorů výše popsanou metodou spočívá v tom, že se pomocí regresního modelu chceme jejich vliv naučit a pak jednoduše odečtením od původních dat odstranit.

Metody založené na multivariačních modelech mají i několik nevýhod, například to, že pro umožnění používání regresních metod by měly být splněny předpoklady o datech a o regresním modelu [17] (např. předpoklad o linearitě dat).

### 2.3.3 Propensity score matching (PSM)

Propensity score matching (PSM) je metoda, jejíž popularita za posledních několik let mírně roste [18]. Metoda je založená na Propensity Score (PS) - hodnota, ukazující pravděpodobnost toho, že dané pozorování bude ovlivněno studovaným faktorem (exposure). Koncepčně algoritmus funguje tak, že nejdřív se vyberou potenciální matoucí faktory, jejíž vliv chceme odstranit, dále se spočítá PS, na základě čehož se provede kontrola vyváženosti mezi hlavní a kontrolní skupinou (nevyváženost mezi skupinami by pak mohla vést k špatnému přepočítání PS). Daná kontrola se provádí např. pomocí stratifikace nebo metody nejbližších sousedů [18]. Následujícím a posledním krokem je výpočet samotného vlivu faktoru.

Velkou nevýhodou PSM je, že daná metoda vyžaduje velké vzorky s dostatečným překrytím mezi hlavní a kontrolní skupinou pro umožnění párování vzorků se stejným PS, čehož často nelze ve studiích dosáhnout. Dalším nedostatkem je to, že metoda umí pracovat jenom s těmi proměnnými, které se dají pozorovat. To znamená, že pokud máme informaci jenom o tom, jestli konkrétní vzorek mohl být ovlivněn potenciálním matoucím faktorem, ale nemáme ani přístup k pozorování dané proměnné, ani doplňující informaci o jiných možných proměnných, tak daný faktor nemůže být vyhodnocen pomocí PSM.

## 2.4 Porovnávání metod statistické analýzy

V literatuře se vyskytuje velké množství porovnávání jednotlivých statistických metod. Při porovnávání se obvykle klade důraz na zkreslení (bias), robustnost, přesnost a jiné parametry. Ukazuje se, že PSM se obvykle chová stejně efektivně jako multivariační modely [19]. Stratifikace je zase známá svou jednoduchosti.

V tabulce 2.1 jsou stručně představeny hlavní výhody a nevýhody popsaných metod.

| Metoda | Výhody | Nevýhody |
|---|---|---|
| Stratifikace | Jednoduchost. Málo předpokladů. Pracuje na víc detailní úrovní. | Pouze malý počet confounderů. Citlivý k řídkým datům. |
| Multivariační regresní modely | Zvládá větší počet confounderů. Výsledek může být libovolné formy (nejen binární). | Hodně předpokladů (např. linearita), možnost ztrátu základních znalostí dat. |
| Propensity Score | Efektiví při malém počtu EPV (events per variable). | Doporučuje se pouze v případě velké datové sady (obzvlášť pro kontrolní skupinu). |

**Tabulka 2.1:** Porovnání metod statistické analýzy

## 2.5 Chyby v odstranění vlivu matoucích faktorů

I po kompenzaci matoucího faktoru můžeme narazit na problém zůstatkového zmatku (residual confounding) [20]. Ukazuje se to jako častý problém v observačních studiích. Vzniknout to může například přeučením modelu [21]. V case-control studiích se takový zmatek může objevit kvůli samotné proceduře výběru, i když na první pohled vypadá, že odpovídá všem kritériím pro odhalení matoucího faktoru [21]. Studie ukazují, že i když jsou známé všechny matoucí faktory a jsou kontrolovány pomocí multivariačních metod, je bohužel stále možné dostat nepřesné výsledky a získat tím další zmatení [20].

# Kapitola 3

# Materiály a metody

V této kapitole jsou popsána poskytnutá data. Dále jsou popsány metody pro předzpracování dat a metody strojového učení pro klasifikaci dat. Zaměříme se také na navržení metody pro kompenzaci matoucích faktorů. Pro experimentování s danou metodou navrhneme generátor umělých dat, který umožňuje nasimulovat data s širokou škálou klíčových parametrů.

## 3.1 Poskytnutá data

Reálná data jsou poskytnuta z výzkumného projektu, jehož cílem je vyšetřit transkriptomický profil příjemců ledvin a naučit se predikovat pacienty s operační tolerantností pro zjednodušení řízení imunosupresivní terapie.

Po transplantaci ledviny imunitní systém pacienta bude pravděpodobně čelit odmítnutí orgánu. Aby se tomu zabránilo, pooperační normou je předepisování léků (imunosupresivní terapie), které omezují nebo blokují reakce imunitního systému. Zároveň ale tyto léky ovlivňují genovou expresi a mohou vést k negativním vedlejším účinkům.

U některých pacientů, kteří z nějakého důvodu přestali imunosupresivní léky užívat, k odmítnutí ledviny nedošlo. Takový vzácný jev nazýváme operační tolerance (OT) [22]. Společně s tím také rozlišujeme skupinu pacientů s chronickou rejekcí (CR) a skupinu stabilních pacientů (STA).

Součástí datové sady je:

- **Matice počtů čtení (read counts)**. Obsahuje celkem 80 vzorků a 187626 transkriptů. U vzorků rozlišujeme 5 skupin, z nichž se zaměříme jenom na skupiny OT, CR a STA. Transkripty jsou popsané pomocí databáze Ensembl ID.
- **Data imunosuprese**. Obsahují informaci o skupině, ke které vzorek patří, třech různých imunosupresivách (IS) a jejich podávání pro každého pacienta (0 v případě že pacient lék nedostával, 1 v případě že dostával).

## 3.2 Cíl a popis problému

Na začátku je třeba uvést třídy vzorků a jejich vlastnosti:

- OT: Pacienti ve skupině OT neužívali léky, takže jejich genová exprese není tímto faktorem ovlivněna.
- CR, STA: Pacienti třídy CR a STA léky užívali. Jinými slovy, léky měly nějaký účinek na geny daných pacientů a jejich data jsou tím pádem "zašuměné".

Cílem je naučit se oddělovat pacienty s operační tolerancí (OT) od pacientů s chronickou rejekcí (CR) podle dat genové exprese. Úrovně exprese genů/transkriptů zde slouží jako číselné příznaky pro oddělení skupin pacientů.

Je zřejmé, že vzhledem k tomu, že vzorky třídy CR jsou ovlivněny matoucími faktory a vzorky OT nikoliv, může být oddělení těchto tříd falešné. Hlavním příznakem, který by odděloval skupiny od sebe, by bylo užívání léků a ne úrovně genové exprese. Proto odstranění vlivu léků v datech je nezbytným krokem pro určení skutečného vztahu mezi třídami.

Data však obsahují řadu problémů, které musíme v této práci vyřešit. Prvním problémem je malý počet vzorků, který brání přesné detekci a může být kritický pro algoritmy strojového učení. Dalším nezanedbatelným problémem jsou matoucí faktory (imunosupresiva), respektive jejich odstranění. Tento problém je významný tím, že není snadné v datech oddělit roli třídy, do které daný vzorek patří, od role matoucích faktorů. Vzhledem k těmto problémům bude na konci kapitoly představen generátor umělých dat, který pomůže provést více experimentů s kontrolou vlivu jednotlivých faktorů.

## 3.3 Předzpracování dat a výběr příznaků

Při zpracováni biologických dat často vzniká problém přeučení, způsobený velkým počtem příznaků a malým počtem vzorků. Data mohou obsahovat až stovky tisíc měřených příznaků, což vede k vytváření složitých modelů, které mohou být přeučené na trénovacích datech a nelze je úspěšně klasifikovat na základě testovacích dat. Aby se tomu zabránilo, používá se předzpracování dat a metody výběru příznaků (feature selection), což umožňuje vynechat méně důležité pro analýzu data a tímto způsobem zjednodušit model a zvýšit přesnost klasifikace.

Základním krokem předzpracování RNA-Seq dat je odstraňování příznaků s nízkou (nulovou) genovou expresi, tzn. odstranění takových transkriptů, které v každém vzorku mají (skoro) nulový počet čtení. Taky běžnou praktikou je odstranění transkriptu s extrémně velkou výchylkou a s tzv. low mean počty čtení [23].

Dalším krokem k zjednodušení modelu je zvolení optimálního počtu příznaků (feature selection) [24]. Účelem tohoto procesu je snížit počet příznaků tak, aby zbývající hodnoty mohly co nejlépe rozdělit data na jednotlivé třídy. Pomocí tako-

vého výběru je možné dosáhnout zvýšení efektivity a přesnosti klasifikátoru, protože vyřazení nepotřebných příznaků, případně nahrazení příznaků menším počtem nových, urychluje proces učení. Selekce příznaků může být provedená pomocí algoritmů strojového učení, nebo pomocí analýzy diferenciální exprese.

## 3.4 Klasifikace

V rámci této práce budeme používat algoritmy učení s učitelem, což znamená, že k učení využívají trénovací množinu, která se skládá z informací o třídě a příznacích každého vzorku. V této sekci budou stručně popsány klasifikační algoritmy, jako je logistická regrese, SVM a k-nejbližších sousedů.

### 3.4.1 Logistická regrese

Logistická regrese [25] je lineárním regresním modelem, používaným zejména pro klasifikaci dat. Slouží hlavně k predikci a výpočtu pravděpodobnosti možných výsledků. Normalizace dat vetšinou výkonnost modelu výrazně zlepšuje. Dané pravděpodobnosti jsou modelovány pomocí logistické funkce:

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}, \tag{3.1}$$

kde $\beta_0$ a $\beta_1$ jsou koeficienty, $x_i$ jsou nezávislé proměnné.

Logaritmická ztrátová funkce pak vypadá následujícím způsobem:

$$L_{log} = -\ln(L) = -\sum_{i=1}^{N}\Big[-\ln(1 + e^{(\beta_0+\beta_1 x_i)} + y_i(\beta_0 + \beta_1 x_i))\Big], \tag{3.2}$$

kde $\beta_0$ a $\beta_1$ jsou koeficienty, $x_i$ jsou nezávislé proměnné, L je log-likelihood funkce, $N$ je počet pozorování, $y_i$ je binární výsledek (0 nebo 1).

Pro tvorbu jednoduššího a efektivnějšího modelu při velkém počtu příznaků se používají tzv. regularizační techniky [26], které omezují flexibilitu modelu. Dané techniky řeší problém přeučení a pomáhají s výběrem příznaků. Mezi regularizační techniky patří L1, L2 a ElasticNet regularizace. Klíčovým rozdílem mezi nimi je přidaní různé penalizace ke ztrátové funkci.

#### L2 regularizace (Ridge)

Jednou z regularizačních technik je L2, neboli Hřebenová regrese (angl. ridge regrese). L2 regularizace zmenšuje koeficienty v regresi směrem k nule, malokdy ale nebudou přesně nulové. Vypočítává se penalizační člen jako součet čtverců hodnot vektorů. Nová ztrátová funkce pak vypadá ve tvaru:

$$L_{log} + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{3.3}$$

kde $\lambda$ je síla regularizace, $L_{log}$ je popsána vzorcem 3.2.

#### L1 regularizace (Lasso)

Pomocí drobné úpravy v penalizačním členu vzniká L1 regularizace. Na rozdíl od L2, používáme absolutní hodnoty $\beta_j$ místo čtvercových:

$$L_{log} + \lambda \sum_{j=1}^{p} ||\beta_j|| \tag{3.4}$$

#### Elastic-Net regularizace

Regularizace Elastic-Net v sobě lineárně kombinuje L1 a L2 a tím pádem minimalizuje následující ztrátovou funkci:

$$L_{log} + \lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)||\beta_j||), \tag{3.5}$$

kde $0 < \alpha < 1$ je koeficient řídicí váhu L1 nebo L2 penalizace.

### 3.4.2 SVM

SVM [27] je jedna z nejpopulárnějších metod strojového učení, jejíž výhodou jsou silné matematické základy. Ve své základní formě SVM vyžaduje, aby třídy byly lineárně separabilní a jedná se o binární klasifikátor. Algoritmus se v zásadě snaží dosáhnout maximálního oddělení tříd, proto se hledá optimální nadrovina, která je nejvzdálenější od nejbližších bodů každé třídy. K hledání takové nadroviny používá metoda tzv. podpůrné vektory (angl. support vectors), což jsou takové body, které leží na hranici rozdělující nadroviny.

SVM může pracovat nejen s lineární klasifikací (danou rozdělující nadrovinou), ale také s nelineární pomocí tzv. kernel triku, který mapuje vstupní vektory do vyšší dimenze. Dalším rozšířením je soft-margin klasifikace, jejíž myšlenkou je dovolit SVM tolerovat určitý počet chyb a udržet co nejširší hranici, aby ostatní body mohly být stále klasifikovány správně.

### 3.4.3 kNN

Algoritmus $k$ nejbližších sousedů (kNN)[28] je neparametrická metoda, jež se používá při klasifikaci i regresi. Při klasifikaci je výsledek založen na většině $k$ nejbližších sousedů a bude přiřazen k nejčastější třídě mezi nimi. Tato metoda je závislá na vzdálenosti příznakových vektorů v prostoru. Normalizace dat může výrazně zlepšit výkonnost modelu. Sousedům lze přiřadit váhy udávající důležitost souseda, takže

čím je soused blíže, tím je důležitější. Funkce $k$-NN je citlivá na šum v trénovacích datech, protože by mohla vytvořit lokální shluk špatně klasifikovaných vektorů příznaků.

Metrikou vzdálenosti může být např. Hammingova vzdálenost (pro diskrétní data), euklidovská vzdálenost (pro spojitá data) nebo jiné. Parametr $k$ se vybírá na základě vstupních dat a lze jej vhodně zvolit pomocí heuristických funkcí. Čím vyšší je $k$, tím menší je významnost šumu, ale hranice mezi třídami jsou méně zřetelné. Při binární klasifikaci by $k$ mělo být liché číslo, aby se předešlo problémům s rovností hlasů.

## 3.5 Hodnocení úspěšnosti klasifikační metody

### 3.5.1 Křížová validace

V případech kdy je třeba vyhodnotit úspěšnost klasifikátoru a nejsou dostupná testovací data, používá se křížová validace (angl. cross validation) [29]. V daném procesu se soubor dat náhodně rozděluje do dvou podmnožin, z nichž první (trénovací sada) se používá k učení modelu a druhá (testovací sada) se používá k testování a předpovídání skóre pro vyhodnocení kvality modelu. Jinak řečeno, to umožňuje testování klasifikátoru na datech, které nesloužily k jeho natrénování. Proto se tato metoda často používá k zabránění přeučení klasifikátorů a je vhodná pro malé množství vstupních dat.

Pro $k$-násobnou validaci se vstupní množina dat rozdělí na $k$ stejně velkých podmnožin, a v každém z $k$ opakování $k - 1$ podmnožin se využívá pro trénování modelu, zbylá jedna část pro testování. Speciální případ křížové validace, kdy $k$ je rovno počtu vzorků, se nazývá leave-one-out.

### 3.5.2 Plocha pod ROC křivkou

ROC křivka [30] je nástroj, který se používá k vyhodnocení kvality binárního klasifikátoru. Vypovídá o schopnosti modelu správně predikovat vzorky do jednotlivých tříd a kvantifikuje úspěšnost klasifikátoru.

Při klasifikaci vzorků mohou nastat 4 případné výstupy:

- TP (true positive) správně klasifikovaný pozitivní výsledek
- FP (false positive) negativní výsledek, který byl nesprávně klasifikován jako pozitivní
- TN (true negative) správně klasifikovaný negativní výsledek
- FN (false negative) pozitivní výsledek, který byl nesprávně klasifikován jako negativní

Dané hodnoty pak představují matici záměn (angl. confusion matrix) (tabulka 3.1).

| | | Skutečné | |
|---|---|---|---|
| | | + | − |
| Predikované | + | $TP$ | $FP$ |
| | − | $FN$ | $TN$ |

**Tabulka 3.1:** Matice záměn

ROC křivka pak shrnuje informaci o všech maticích záměn, které každá prahová hodnota metody vytvořila. Svislá osa grafu ukazuje TPR (true positive rate) (formule 3.6), což je podíl pozitivních vzorků, které byly správně klasifikovány a jsou tedy skutečně pozitivní. Na vodorovné ose se vynáší FPR (false positive rate) (formule 3.7), který vypovídá o podílu negativních vzorků, které byly nesprávně klasifikovány a jsou tedy falešně pozitivní.

$$TPR = \frac{TP}{TP + FN} \tag{3.6}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.7}$$

Standardním kvantitativním vyjádřením ROC křivky je AUC (angl. area under curve) – plocha oblasti ohraničená ROC křivkou. Vypovídá o pravděpodobnosti toho, že náhodně vybraný pozitivní vzorek bude hodnocen lépe než náhodně vybraný negativní vzorek. Hodnoty AUC se pohybují od 0 do 1 – čím víc je tato hodnota, tím přesnější je klasifikátor. Je-li plocha pod křivkou rovná 1, klasifikátor se považuje za ideální. Pokud je plocha pod křivkou 0.5 pak takový klasifikátor se dá přirovnat k náhodnému odhadu.

## 3.6 Kompenzace vlivu matoucích faktoru

V této části se budeme zabývat odstraněním vlivu matoucích faktorů v datech.

### 3.6.1 Návrh metody

Jak se už probíralo v sekci 2.3, každá ze statistických metod pro kompenzaci vlivu matoucích faktorů má svoje nevýhody, které mohou být pro konkrétní data podstatné.

Tak stratifikace dokáže ovládat jen malý počet matoucích faktorů. PSM potřebuje mít dobře párovanou hlavní a kontrolní skupinu, které mezi sebou balancují. V poskytnutých datech neumíme odhadnout pravděpodobnost léčení, což vede k tomu, že se nepovede PSM ani nastartovat kvůli nemožnosti spočítat jednotlivé skóre.

Proto jako jedinou variantou pro zbavení vlivu matoucích faktorů pro poskytnutá data byl zvolen multivariační regresní model. Stejnou metodu používá studie, která

se zabývá nalezením signatury operační tolerance a která také zahrnuje úpravu pro imunosupresivní terapii [31].

### 3.6.2 Použití navržené metody v podobné studii

Obě studie jsou podobné, avšak se liší některými podmínkami při zpracování vzorků. Zmíněná výše studie používá data genové exprese, získané pomocí techniky real-time PCR, studie, ze které byla poskytnuta data, používá RNA-Seq. Ačkoli se základní principy těchto dvou technik liší, obě mohou z jakéhokoliv typu vzorku poskytnout informace o množství RNA [32].

Základní odlišení technik je v tom, že RNA-Seq poskytuje rozšířenější přehled o celém transkriptomu a hodí se například v případech, kdy ještě není úplně jasné, jaké konkretní geny/transkripty jsou pro studii důležité, nebo taky v případě, kdy je nutné prozkoumat reagování celé tkání nebo organismu, a takový kompletní přehled je žádoucí. Real-time PCR technika se naopak zaměřuje na konkretní geny a je dobrou volbou pro případy kdy přehled o kompletním transkriptomu není věcí zájmu a je známo na které konkrétní geny/transkripty se výzkum zaměřuje.

Dané dvě studii se také liší počtem zkoumaných genů. Ohodnocení u výše popsané studie se provádí na genových signaturách, jejichž počet nepřekračuje 10 příznaků. V našich datech jde o sta tisíce transkriptů.

### 3.6.3 Popis metody

Pro predikování roli IS se bude používat multivariační lineární regresní model v následujícím tvaru:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \tag{3.8}$$

$$\hat{y} \sim NB(\mu, \phi), \tag{3.9}$$

kde $n > 0$ je počet matoucích faktorů (tj. počet druhů imunosupresiv), $\beta_1, \ldots, \beta_n$ jsou koeficienty vyjadřující vliv podávání imunosupresiv.

Dalším krokem po získaní koeficientů je odečtení naučeného vlivu z původních dat. Nezapomínejme přitom, že pracujeme s Poissonovým regresním modelem, který používá logaritmus jako linkovací funkci (2.3):

$$ln(y_{filt}) = ln(\hat{y}) - \beta_1 x_1 - \ldots - \beta_n x_n, \tag{3.10}$$

$$y_{filt} = e^{ln(\hat{y}) - x_1\beta_1 - \ldots - x_n\beta_n} = \hat{y}/e^{x_1\beta_1 + \cdots + x_n\beta_n}, \tag{3.11}$$

kde $\beta$ je koeficient vyjadřující naučený vliv, $x$ je podávání IS (0 nebo 1), $y$ je je vektor vyjadřující původní transkript, $y_{filt}$ je vektor vyjadřující výsledný transkript, / je prvkové dělení.

Obecněji v maticovém tvaru se daný vztah dá zapsat jako:

$$Y_{filt} = Y / e^{(X*B)^T}, \tag{3.12}$$

kde matice X o velikosti *počet CR vzorků* × *počet IS* vyjadřuje podávání imunosupresiv, matice B o velikosti *počet IS* × *počet transkriptů* vyjadřuje naučené koeficienty.

Na poskytnutých datech se model, popsaný vztahem 3.8, učí z STA[1] vzorků. Vzorky třídy OT[2] během celého procesu zůstávají neměnné vzhledem k tomu, že jejich genová exprese není ovlivněna imunosupresivy, a kompenzace se pak provádí jenom u CR[3] vzorků. Výsledkem je pak nová vyfiltrovaná genová exprese, která je rozdílem mezi pozorovanými a predikovanými pomocí modelu hodnotami.

Na obrazcích 3.1 a 3.2 je uveden příklad kompenzace vlivu imunosupresiv na generovaném transkriptu. Učení se provádí pomocí knihovny statsmodels [33]. Funkce *statsmodels.api.formula.glm()* slouží pro fitování obecných lineárních modelů. Jako vstupní parametry, podle kterých se model bude učit, funkce přijímá formulí (3.8), data, a jejich rozdělení.

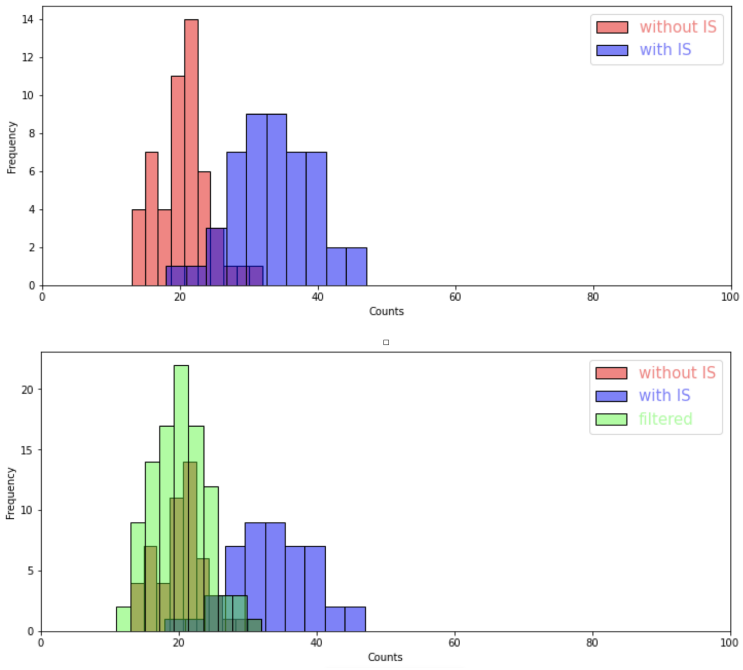
```
# try to learn the parameters from the observations
df = pd.DataFrame({"x": IS1,
                   "y": y})
mp = sm.formula.glm("y ~ x", family=sm.families.Poisson(), data=df).fit()
print(mp.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       98
Model Family:                 Poisson   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -301.84
Date:                Wed, 06 Jan 2021   Deviance:                       95.923
Time:                        22:18:40   Pearson chi2:                     95.3
No. Iterations:                     4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.0082      0.031     97.615      0.000       2.948       3.069
x              0.5065      0.040     12.784      0.000       0.429       0.584
==============================================================================
```

```
# filter the IS1 out from the count data
y_filt = y / np.exp(IS1 * mp.params.x)
```

**Obrázek 3.1:** Kompenzace na příkladě jednoho generovaného transkriptu

[1] Stabilní stav
[2] Operační tolerance (Operational Tolerance)
[3] Chronická rejekce (Chronical Rejection)




**Obrázek 3.2:** Před a po kompenzaci generovaného transkriptu

## 3.7 Generování umělých dat

Jelikož reálná datová sada pro analýzu vlivu imunosuprese není dostatečná, důležitou částí práce je generování umělých RNA-Seq dat. Cílem je na základě reálných dat vytvořit několik umělých datových sad s různými parametry a pomocí provedení experimentů zjistit, jaké parametry pomohou co nejlépe naučit roli IS a efektivně ji snížit.

Generování dat se provádí nastavením klíčových parametrů. Mezi tyto parametry patří:

- Velikost vzorků jednotlivých skupin (OT, CR a STA).
- Počet transkriptů.
- Distribuce mezi třídami.
- Velikost efektu jednotlivých tříd, počet ovlivněných transkriptů.
- Velikost efektu imunosupresivní terapie, počet ovlivněných transkriptů.

### 3.7.1 Obsah dat

Součástí generovaných dat jsou 3 datasety:

1. Původní dataset (origin). Daná data jsou napodobením reálných dat. Obsahují v sobě vliv jak imunosupresiv, tak i třídy.
2. Očekávaný dataset (expected). Data bez vlivu matoucích faktorů, která jsou očekávána po kompenzaci. Vliv třídy obsahují. Budou sloužit výhradně pro ohodnocení úspěšnosti provedené kompenzace.

3. Informace o podávání léků. Hodnoty jsou buď 1, nebo 0 v závislosti na tom, jestli daný vzorek byl, nebo nebyl ovlivněn imunosupresivy.

### 3.7.2 Skupiny vzorků

Generovat se budou vzorky ekvivalentní reálným skupinám pacientů. Každá skupina obsahuje vliv třídy. Pro připomenutí existujících skupin:

- Skupina OT. Operačně toleranční skupina vzorků léky neužívá, tj. vliv imunosupresiv neobsahuje.
- Skupina CR. Skupina vzorků CR léky užívá, vliv imunosupresiv je přítomen.
- Skupina STA. Skupina vzorků STA léky užívá, vliv imunosupresiv je přítomen.

### 3.7.3 Postup

Generátor umělých dat je založen na parametrech převzatých z poskytnutých dat. Odhady distribuce, modelu středních hodnot a rozptylu budou představeny v kapitole 4. V této sekci se seznámíme s postupem použitém při generování dat.

1. **Základní koeficient**. Na začátku se stanoví hodnota pro proměnnou `beta0`. Daná proměnná je odhadnuta střední hodnotou ze skutečných dat.

2. **Vliv třídy**. Vliv třídy je to co pomáhá rozlišit mezi skupinami. Tento efekt se generuje náhodně na základě vstupních parametrů – počet ovlivněných transkriptů a velikost tohoto vlivu. Velikost vlivu se odhaduje normálním rozdělením.

$$\beta_0 = \beta_0 \pm effect$$

3. **Podávání imunosupresiv**. Hodnota náhodně nabývá 0 nebo 1 s pravděpodobností 40%/60%.

4. **Vliv imunosupresiv**. Koeficienty `beta1`, ..., `beta3` popisují efekt jednotlivých podávání imunosupresiv. Generují se náhodně na základě zadaných parametrů – počet ovlivněných transkriptů (`n_IS_effect`) a velikost tohoto vlivu ($0 <$ `IS_effect` $< 1$).

5. **Střední hodnota** $\mu$.

   - Pro generování origin datasetu (1) se z odvozených koeficientů $\beta_0$,..., $\beta_3$ pak generují střední hodnoty $\mu$ pro každý transkript:

$$\mu = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3} \tag{3.13}$$

- Pro případ, kdy genová exprese není ovlivněna matoucími faktory, tj. pro generování expected datasetu (2), proměnné $x_1$,..., $x_3$ jsou nulové, tudíž střední hodnota $\mu$ pro takovou situaci se odhaduje jako

$$\mu = e^{\beta_0} \tag{3.14}$$

6. **Výsledná exprese**.

   Dalším krokem je na základě zvoleného rozdělení negenerovat samotnou expresi. V případě Poissonova rozdělení použít pro parametr $\lambda$ popsanou výše střední hodnotu $\mu$. V případě NB rozdělení použít parametrizaci popsanou pomocí vzorce 1.6 a 1.7. Rozptyl lze získat odhadem vztahu mezi střední hodnotou a rozptylem ze skutečných dat .

# Kapitola 4

# Experimenty

Daná kapitola obsahuje experimenty nad reálnými a synteticky generovanými daty. Na začátku se provede analýza pro lepší představu o datech, na základě čehož budou zvoleny parametry na generováni umělých datasetů. Dále bude ukázána klasifikace reálných dat pomocí vybraných algoritmů. Hlavní častí této kapitoly je kompenzace vlivu matoucích faktorů na reálných a pak i na synteticky generovaných datech.

Veškeré zpracování dat a výpočty byly provedeny ve vysokoúrovňovém interpretovaném jazyku Python v3.8 se základními vědeckými knihovnami a za použitím programovacího prostředí Notebook Jupyter, které umožňuje dokumentovat výpočty ve snadno reprodukovatelné formě.

## 4.1 Analýza poskytnutých dat

### 4.1.1 Předzpracování dat

Před přistoupením k odhadům je třeba nejprve provést předzpracování dat, ve kterém se odeberou transkripty s malou genovou expresí. Dělá se to pro to, aby velké množství nulových genových expresi neovlivnilo výsledek.

Reálná datová sada je představena maticí čtení. Prvním krokem při předzpracování dat je odstranit nepotřebných pro účely dané práce skupin vzorků a nechat pouze vzorky, které nás zajímají, tj. z OT, CR a STA skupin. Tím pádem z 80 vzorků zůstane jenom 50, z nichž 15 vzorků tvoří skupina OT, 12 vzorků - skupina CR a zbylých 23 - skupina STA.

Prvotně data obsahují 187 tisíc transkriptů. Na začátku se odfiltrují transkripty s nulovou expresi, které nenesou žádnou informaci pro další analýzu. Takových transkriptů je v datech 30 tisíc.

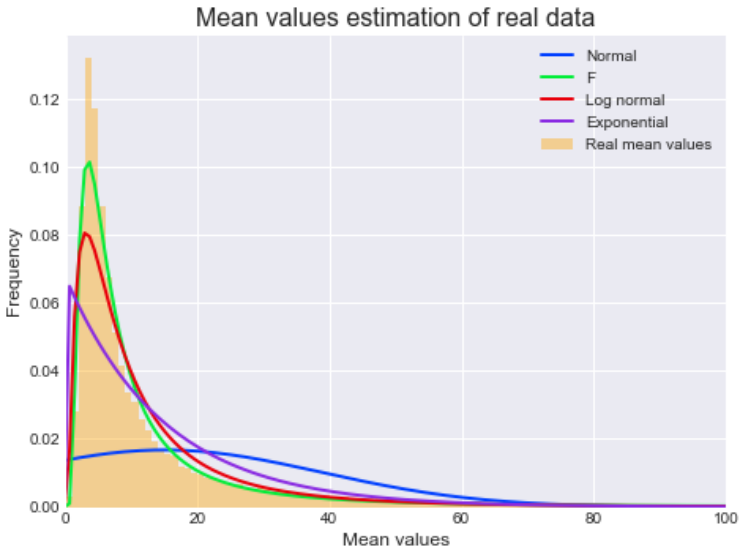
Dále se z dat také odfiltrují transkripty, které mají v sumě 30 a méně čtení. Takových transkriptů je 67 tisíc. Posledním krokem filtrování je nechat pouze takové

transkripty, u nichž aspoň v jednom ze vzorků počet čtení přesahuje 9.

Danými úpravami se podařilo omezit počet transkriptů na 45 tisíc, což je oproti původnímu počtu (187 tisíc) prokazatelnou změnou.

### 4.1.2 Odhad střední hodnoty

Pro účely získaní přehledu o tom, jak rozdělení dat vypadá, je potřeba nakreslit histogram středních hodnot přes všechny transkripty. Odhad střední hodnoty pak bude použit při generování dat (proměnná $\beta_0$ v 3.7.3). Na obrázku 4.1 jsou zobrazeny různé distribuce odhadnuté na základě reálných dat. Pro odhad parametrů byla použita knihovna *scipy*. Srovnaní se provádělo mezi normálním, F, logaritmicko-normálním a exponenciálním rozdělením. Z daných odhadů můžeme pozorovat, že nejlepší fitování má Fisherovo (F) rozdělení.




**Obrázek 4.1:** Odhad středních hodnot z reálných dat pomocí normálního, F, logaritmicko-normálního a exponenciálního rozdělení

Odhadnuté pomocí funkci *scipy.stats.f.fit()* parametry F rozdělení:

$$\begin{aligned} dfn &= 38141.58 \\ dfd &= 3.02 \\ loc &= -0.22 \\ scale &= 5.98 \end{aligned}$$

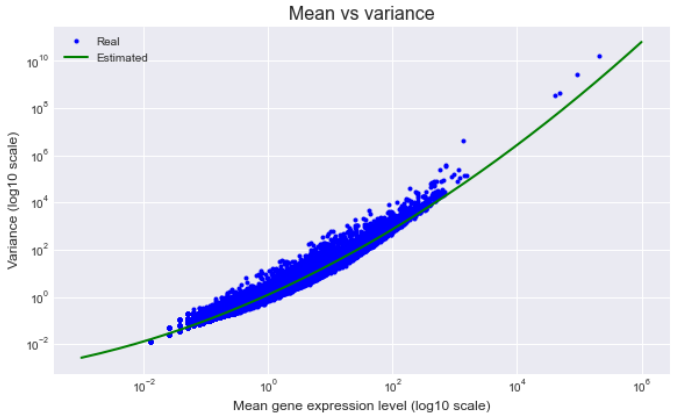
Skoro stejně dobrou aproximaci prokázalo logaritmicko-normální rozdělení s parametry, získanými funkcí *scipy.stats.lognorm.fit()*:

$$\begin{aligned} s &= 1.0 \\ loc &= 0.11 \\ scale &= 8.14 \end{aligned}$$

Z důvodu lepší aproximaci mnou bylo zvoleno F rozdělení. Dobré fitování tohoto rozdělení ale může být způsobeno velkým počtem parametrů a obecně se při modelování středních expresí nepoužívá. V studiích se rozdělení středních expresí často odhaduje pomocí bimodálního modelu [34] (s ohledem na různě regulované transkripty) nebo pomocí lognormálního modelu.

### 4.1.3 Odhad rozptylu

Po zvolení střední hodnoty je třeba se zaměřit na vztah parametrů pro snadný odhad rozptylu ze středních expresi.



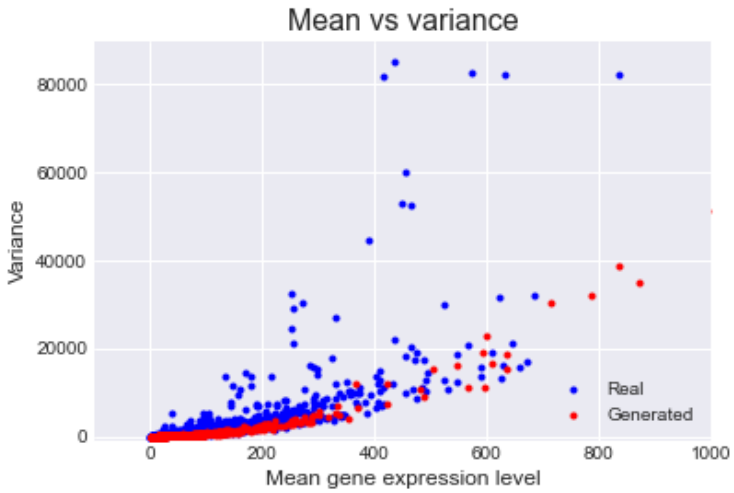

**Obrázek 4.2:** Střední hodnota a rozptyl u reálných dat (označeno modře)

Na obrázku 4.2 je znázorněn graf s logaritmickými osami, kde na vodorovné ose jsou zobrazeny střední hodnoty přes všechny transkripty a na svislé ose – odpovídající rozptyly. Z grafu je velmi dobře pozorovatelná overdisperze u větších středních hodnot. Pro odhad vztahu je nutné využit metody pro optimální proložení množiny bodů křivkou. Mezi nejznámější patří metoda nejmenších čtverců [35], která je jednoduchá, rychlá, spolehlivá a široce se používá v mnoha podobných aplikacích. Daná metoda vyžaduje pouze volbu fitovaného řádu polynomu. Jasnou volbou pro daný případ je parabola (polynom 2. řádu ve formě $ax^2 + bx + c = 0$), která poskytuje dostatečně dobrou aproximaci.

Pro daný odhad byla použita funkce *polyfit* z knihovny NumPy. NumPy je balíček pro numerické výpočty, který se používá pro definici numerických polí a matic a pro základní operace s nimi [36]. Se získanými koeficienty dostáváme následující kvadratický vztah:

$$\sigma^2 = 0.04324\mu^2 + 1.188\mu + 0.2093 \tag{4.1}$$

Na obrázku 4.3 můžeme pozorovat vztah mezi rozptylem a střední hodnotou u reálných a generovaných pomocí negativně binomického rozdělení dat s využitím vzorce 4.1 pro odhad rozptylu.

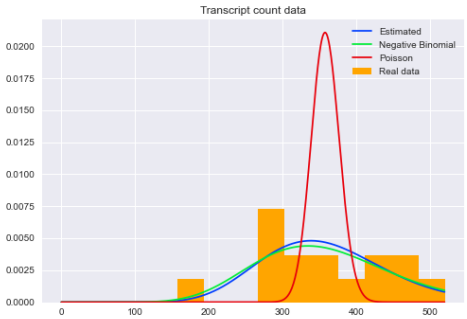



**Obrázek 4.3:** Střední hodnota a rozptyl u reálných (označeno modře) a generovaných (označeno červeně) dat

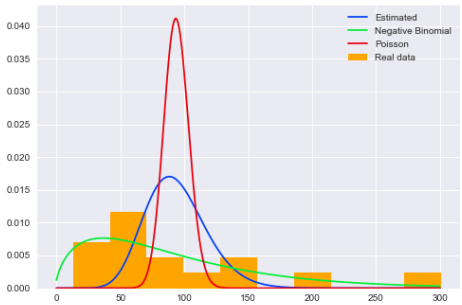
### 4.1.4 Distribuce dat

Jak se probíralo v sekci 1.5, typicky se pro popis count dat používá Poissonovo rozdělení, což pro data s overdisperzí není úplně vhodnou volbou. Na obrazcích 4.4 a 4.5 jsou znázorněny příklady transkriptů s vysokým rozptylem. Červeně je označen odhad pomocí Poissonova rozdělení, který v daném případě vůbec neodpovídá reálným datům. Modře je označen odhad pomocí Negativně binomického rozdělení s využitím rozptylu odhadnutého vzorcem 4.1.

Parametrizace, kterou používá *scipy.stats.nbinom* pro negativně binomické rozděleni, je popsaná vzorci 1.6 a 1.7.




**Obrázek 4.4:** Odhad distribuci transkriptu s velkým rozptylem

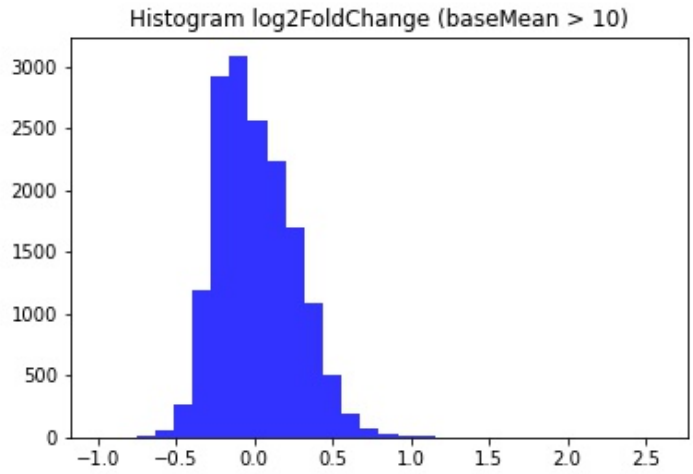



**Obrázek 4.5:** Odhad distribuci transkriptu s velkým rozptylem

### 4.1.5 Odhad vlivu třídy a imunosuprese

Z reálných dat jsme schopni odhadnout celkový počet ovlivněných transkriptů. Pomocí analýzy diferenciální exprese (DGE) stanovíme dolní a horní hranici tohoto

počtu.




**Obrázek 4.6:** Histogram log2 fold change hodnot

Dolní mez byla odhadnuta pomocí adjustované p hodnoty, konkretně $padj < 0.05$. Pro toto nastavení je počet transkriptů roven 5. Počty čtení však u vybraných 5 transkriptů jsou poměrně nízké, což vypovídá o tom, ze tento odhad je málo pravděpodoben. Abychom se vyhnuli nízkým počtům čtení, pro určení horní hranice nejprve vybereme transkripty s průměrnou hodnotou vyšší než 10. Poté na základě histogramu znázorněného na obrázku 4.6 vybereme transkripty s parametry $log2FC > 0.75$. Horní hranice se pak odhaduje jako 81 transkriptů.

Předpokládáme tedy, že počet ovlivněných transkriptů se pohybuje v rozmezí od 5 do 81 transkriptů. Připomeňme, že se jedná o vliv jak třídy, tak imunosuprese.

## 4.2 Klasifikace dat

Všechny výpočty byly provedeny pomocí knihovny *scikit-klearn* [37], mezi jejíž hlavní funkce patří klasifikace, regrese, předzpracování a jiné nástroje strojového učení. I při použití této knihovny je však důležité zadat správné parametry pro metody. Ty byly vybrány na základě vyhodnocení různých parametrů a následně porovnány mezi sebou. Funkce z knihovny přijímá více parametrů, než o kterých se zde píše. Některé z nich byly vynechány, protože jejich výchozí hodnoty jsou již rozumně nastavené nebo jsou příliš specifické na to, aby se daly upravovat a uvádět zde. Pokud některé parametry nebyly zmíněny, předpokládá se, že byly použity výchozí hodnoty poskytované knihovnou.

### 4.2.1 Logistická regrese

Jako penalizační funkce pro logistickou regresi byla použita Elastic-Net regularizace. Logistická regrese jako vstup akceptuje parametry $\alpha$ a *l1_ratio*. Parametr $0 \leq$ *l1_ratio* $\leq 1$ slouží jako speciální parametr pro Elastic-Net, přičemž *l1_ratio* $=0$ odpovídá penalizaci L2, *l1_ratio* $=1$ penalizaci L1. Parametr $\alpha$ je konstanta, kterou se násobí regularizační člen, proto čím je tato hodnota vyšší, tím silnější je regularizace. Výsledky klasifikace jsou znázorněny v tabulce 4.1.

| L1 ratio / $\alpha$ | 0.001 | 0.01 | **0.02** | 0.1 | 1 |
|---|---|---|---|---|---|
| 0.05 | 0.79 | 0.81 | 0.828 | 0.828 | 0.807 |
| **0.1** | 0.81 | 0.8 | **0.829** | 0.79 | 0.8 |
| 0.3 | 0.78 | 0.81 | 0.82 | 0.795 | 0.665 |
| 0.5 | 0.8 | 0.8 | 0.81 | 0.742 | 0.5 |
| 0.7 | 0.8 | 0.81 | 0.82 | 0.79 | 0.5 |
| 0.9 | 0.81 | 0.81 | 0.79 | 0.78 | 0.5 |

**Tabulka 4.1:** Výsledky klasifikace (AUC skóre) pro různé parametry logistické regresi

### 4.2.2 SVM

SVM bylo vyhodnoceno s použitím lineárního kernelu. Lineární kernel pro velký počet příznaku je nejvhodnější volbou, protože na rozdíl od jiných (např. radiální bázové funkce nebo polynomiální) se data netransformují do vyšších dimenzí. SVM využívá L2 regularizaci a akceptuje parametr C jako penalizaci za chybně klasifikované body.

| Kernel / C | 0.01 | 0.1 | **1** | 10 | 100 |
|---|---|---|---|---|---|
| Linear | 0.83 | 0.848 | **0.857** | 0.838 | 0.825 |

**Tabulka 4.2:** Výsledky klasifikace (AUC skóre) pro různé parametry SVM

### 4.2.3 kNN

Optimálním parametrem k je 9.

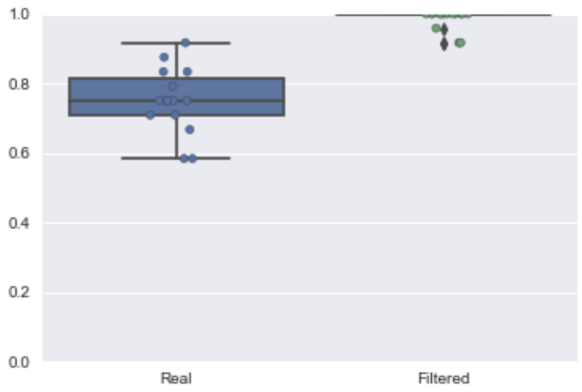
| Metrika / k | 3 | 5 | 7 | **9** | 11 | 13 |
|---|---|---|---|---|---|---|
| euclidean | 0.64 | 0.7 | 0.73 | 0.8 | 0.79 | 0.77 |
| **manhattan** | 0.67 | 0.74 | 0.76 | **0.81** | 0.8 | 0.73 |

**Tabulka 4.3:** Výsledky klasifikace (AUC skóre) pro různé parametry SVM

## 4.3 Kompenzace vlivu matoucích faktorů na reálných datech

Jak již bylo zmíněno, poskytnutá data jsou ovlivněná tím, zda příslušné třídě pacientů byly nebo nebyly podávány léky, které se zde pokládají za matoucí faktory. Výsledky klasifikace proto nelze považovat za skutečné dokud neproběhne úprava dat odstraněním vlivu těchto léků. Pro kompenzaci vlivu léků se používá metoda popsaná v 3.6 a učení probíhá na vzorcích třídy STA. Výpočet je implementován pomocí modulu *statsmodels* [33], který poskytuje funkce pro odhady mnoha různých statistických modelů a provádění statistické analýzy dat. Je nutné podotknout, že pro každý transkript učení probíhá zvlášť a proto je časově náročnější pro data s velkým počtem příznaků.

Po odstranění vlivu probíhá zopakování klasifikace pomocí logistické regrese s parametry $l1\_ratio$=0.1 a $\alpha$=0.02. Obrázek 4.7 ukazuje, že při pokusu kompenzovat vliv léků v reálných datech došlo k překompenzaci. Překompenzace je stav, který je pozorován v případech nedostatečného počtu vzorků, ze kterých se model učí. Ve výsledku nesprávného vytvoření modelu pak vznikají nová zmatení, což ovlivňuje výsledné hodnoty AUC při klasifikaci.




**Obrázek 4.7:** Hodnoty AUC před a po kompenzaci reálných dat

Podle očekávání by se vliv imunosuprese ve skupině CR měl zmenšit, tj. předpokládá se, že hodnota AUC klesne vzhledem k odstranění matoucího faktoru, který klasifikaci falešně zlepšoval. Ve skutečnosti však se hodnota zvedla na 1.0 (ideální klasifikace), což vypovídá o špatném pokusu o odstranění vlivu matoucích faktorů. Takové mylné zlepšení klasifikace se vysvětluje špatným naučením vlivu, které s sebou přináší nová zmatení ve skupině CR pacientů a falešně tak zvyšuje rozdíl mezi třídami vzorků.

## 4.4 Kompenzace vlivu matoucích faktorů na generovaných datech

Vzhledem k problému, se kterým jsme se setkali při pokusu odstranit vliv léků na poskytnutých datech, důležitým krokem je generování a experimentování s umělými datovými sadami. Na začátku každého experimentu je potřeba nadefinovat doménu, ve které se daný experiment bude pohybovat. Pro účely dané práce byly zvoleny následující parametry:

- **Počet STA vzorků** (`n_STA`). Pro hodnoty na experimentování s počtem STA vzorků budeme používat logaritmickou osu: [20, 50, 100, 200, 500, ..., 10000]
- **Počet transkriptů** (`n_transcripts`). Hodnoty se pohybují ve řadech: [10, 100, 1000, 10000]
- **Počet opakování** (`n_repeats`). Aby byl výsledek co nejpřesnější a co nejméně ovlivněn náhodností, musí být každý experiment opakován vícekrát. V rámci teto práci používám počet opakováni = 10, pokud nebude v experimentu označeno jinak.
- **Velikost vlivu imunosuprese**. Pro počet ovlivněných imunosupresivy genů volím [2, 5, 10, 20, 30, 50] a pro samotný efekt – hodnoty [0.3, 0.5, 0.9].
- **Velikost vlivu třídy**. Použiji [0, 2, 5, 10, 20, 30, 50] transkriptů, které budou ovlivněny třídou. Pro vliv se používá Normální rozdělení se střední hodnotou [0.2, 0.3, 0.4] a rozptylem 0.1.

### 4.4.1 Postup

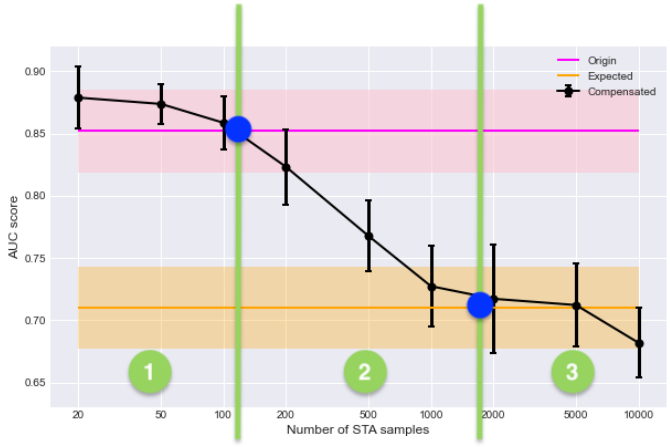
1. Na začátku každého experimentu se vygeneruje rozšířená datová sada s parametry odpovídajícími danému experimentu. Z této datové sady pak budou vznikat nové menší. Generovaná sada obsahuje:
   - Původní data (`origin_counts`). Daná data jsou napodobením reálných dat. Obsahují v sobě vliv jak imunosupresiv, tak i třídy.
   - Očekávaná po kompenzaci matoucích faktorů data (`expected_counts`). Tyto data neobsahují vliv imunosupresiv a budou sloužit výhradně pro vyhodnocení úspěšnosti provedené kompenzace.
   - Informace o podávání léku. Obsahuje hodnoty 1 nebo 0 v závislosti na tom, jestli daný vzorek byl, nebo nebyl ovlivněn imunosupresivy.
2. Z vygenerované datové sady se vyberou `n_STA` vzorků z třídy STA.
3. Provede se kompenzace `origin_counts` dat navrženou v 3.6 metodou. Výsledkem budou odfiltrovaná z vlivu imunosupresiv data `filtered_counts`.
4. Všechny tři sady projdou klasifikací s využitím 10-násobné křížové validace pro přesnější vyhodnocení. Výsledky klasifikace se uloží jako AUC hodnoty.
5. Kroky 2 až 4 se zopakují pro různý počet STA vzorků. Přičemž se dané vzorky pokaždé vybírají z množiny vzorků vybraných v předchozí fázi, tj. předchozí

vybrané vzorky jsou vždycky podmnožinou nově vybraných. Děla se to pro dosáhnutí hladších grafu a z důvodu časové náročnosti.

6. Kroky 2 až 5 se zopakují `n_repeats`-krát s různými kombinacemi vzorků z důvodu eliminace náhodného chování během klasifikace a kompenzace a pro dosáhnutí přesnějších výsledků.

### 4.4.2 Interpretace výsledků pomocí grafu

Výsledky experimentů se interpretují pomocí 2d grafů a heatmap. Jedna kombinace parametrů odpovídá jednomu grafu. Vodorovná osa vyjadřuje počet STA vzorků, svislá osa – AUC hodnoty (výsledky klasifikace).




**Obrázek 4.8:** Příklad grafu pro interpretaci výsledků. Modře jsou označené prahy metody, zeleně – rozdělení na úseky

Na grafu jsou znázorněny tři křivky/přímky:

- Růžová přímka představuje výsledky klasifikace původního (origin) datasetu. Výsledky jsou zprůměrovány ze všech počtů STA, protože jsou na tomto parametru nezávislé. Hodnota je proto konstantní.
- Oranžová přímka odpovídá výsledku klasifikace očekávaného po kompenzaci (expected) datasetu.
- Černá křivka ukazuje výsledky klasifikace po kompenzaci původních dat. Tyto hodnoty jsou závislé na počtu STA, proto se podel vodorovné osy mění.

Výsledky z origin datasetu (růžová) budou vždycky vyšší nebo rovny výsledkům expected datasetu (oranžová). Připomeňme, že u origin datasetu výsledek klasifikace nevypovídá o skutečnem vztahu mezi třídami OT a CR. Klasifikace vykazuje tak dobré výsledky ne proto, že by třídy byly tak dobře oddělitelné, ale zejména proto, že se zde projevuje vliv matoucích faktorů. V umělých datech je možné vliv matoucích faktorů i třídy jasně oddělit, proto expected dataset (pro oranžovou přímku) byl generován bez použití vlivu matoucích faktorů. Oranžová přímka by tím pádem měla

vypovídat o skutečnem vztahu mezi OT a CR. Černá křivka s vetším množstvím STA klesá a blíží se k očekávaným hodnotám (k oranžové přímce).

Důležité jsou dva body (na obrázku 4.8 označené modře):

1. Průsečík růžové a černé křivky – odpovídá minimálnímu počtu STA postačujících pro správné fungování metody. Je to bod kde metoda začíná fungovat, nižší pocty STA vedou k překompenzace.
2. Průsečík (nebo přiblížení) černé křivky a oranžové přímky – bod ukazující počet STA vzorků potřebných pro správné naučení se a odstranění vlivu imunosupresiv.

Graf je těmi body rozdělen do třech intervalů z pohledu správnosti provedené kompenzace.

1. první interval jsou počty STA při kterých dochází k překompenzace, metoda nefunguje.
2. druhý interval se nachází mezi prvním a druhým bodem a to jsou počty STA, pro které metoda funguje, ale nedokonale.
3. třetí interval jsou počty STA, pro které metoda začíná fungovat perfektně, tj. výsledky se ztotožňují s očekávanými.

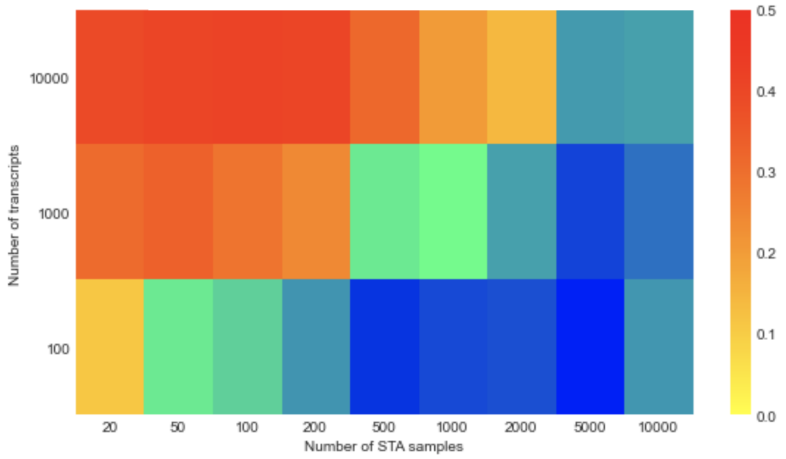
U všech výsledků se vzhledem k počtu opakování experimentů ukazují směrodatné odchylky:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{4.2}$$

kde N je počet opakování.

### 4.4.3 Interpretace výsledků pomocí heatmapy

Dalším způsobem interpretace výsledků je heatmapa. Tento způsob umožňuje rozšířenější přehled o chování metody kompenzace a tím umožňuje jednodušší určení vztahů mezi parametry. Příklad takové heatmapy je vidět na obr. 4.9. Vodorovná osa zde vyjadřuje počet STA vzorků, svislá osa ukazuje počty transkriptů. Barvou se označuje rozdíl ve výsledcích klasifikace mezi filtrovanými a očekávanými daty.




**Obrázek 4.9:** Příklad heatmapy pro interpretaci výsledků

Výsledná heatmapa na rozdíl od grafu 4.8 neukazuje konkrétní rezultáty klasifikace, ale pouze to, jak daleko od očekávaného výsledku se kompenzace v jednotlivém bodě nachází.

Charakteristickou vlastností tohoto přístupu je kombinace dvou barevných palet v jedné heatmapě, což usnadňuje rozpoznání stavů provedené kompenzace:

- Červenožlutá paleta označuje stav překompenzace. Překompenzace je situace, kdy rozdíl mezi očekávanou a filtrovanou křivkou vetší než rozdíl mezi původní a očekávanou. Tento stav je analogický intervalu 1 na obrázku 4.8.
- Modrozelená paleta označuje stav analogický intervalům 2 a 3 - tedy intervalům, u nichž lze předpokládat, že metoda funguje.

Takto lze z heatmapy, stejně jako z prvního typu grafu, okamžitě rozpoznat, při jakém počtu vzorků začíná metoda fungovat. Důležitou poznámkou je, ze heatmapy neznázorňují a celkově nepočítají s odchylkou při vypočítání origin a expected hodnot, proto hranici překompenzačního stavu (hranici mezi dvěma barevnými paletami) nelze pokládat za úplně přesnou a slouží spis pro přibližné znázornění situace.

### 4.4.4 Výsledky

Následující sekce obsahuje výsledky získané v rámci experimentů této práce a diskusi k nim. V příslušných podsekcích jsou představeny tabulky a ukázkové grafy. V rámci této sekce jsou především uvedeny výsledky, popisující experimenty s negativně binomickým rozdělením dat. Další výsledky jsou pak uvedeny v příloze B.

Vzhledem k tomu, ze každý experiment obsahuje několik opakování a samotná kompenzace není rychlý proces, je třeba poznamenat, ze experimenty byly časově náročné. Tak například jeden experiment pro 10 tisíc transkriptů trval 6 hodin. Na základě výsledků pak bylo možné rozpoznat některé vzory chování pomocí heatmap

a pokusit se odhadnout parametry skutečných dat pomocí 2d grafů. To všechno je představeno v rámci této a následující sekce.

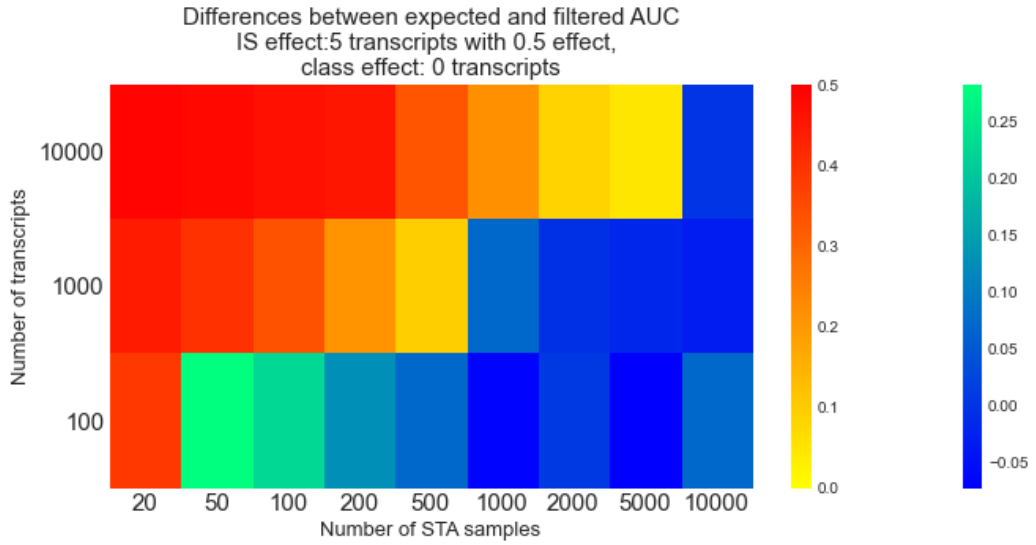
Je nutné podotknout, že u některých grafů dochází k nekontrolovanému chování, které se vysvětluje náhodným šumem. Může se tak stát, že AUC hodnoty compensated datasetu budou nižší, než hodnoty klasifikace expected datasetu, na heatmapě se tak zobrazují záporné hodnoty, které by nastat neměly.

1. **Víc transkriptů = složitější odstranění matoucích faktorů**

   Prvním zjištěním je skutečnost, že s rostoucím počtem transkriptů roste i počet STA vzorků potřebných pro správný odhad modelu.

   Obrázek 4.10 ukazuje rozdíly AUC hodnot mezi filtrovanými a očekávanými daty.

   Na tomto obrázku je vidět, že čím méně transkriptů je v datech přítomno, tím rychleji je metoda schopna překonat překompenzaci. Tato situace byla pozorována ve všech provedených experimentech. Zatímco pro 100 transkriptů začíná metoda pracovat již při 50 vzorcích STA, pro 1000 transkriptů se počet potřebných vzorků STA zvyšuje o 1000 a pro 10000 transkriptů se počet prudce zvyšuje na 10000 STA vzorků.

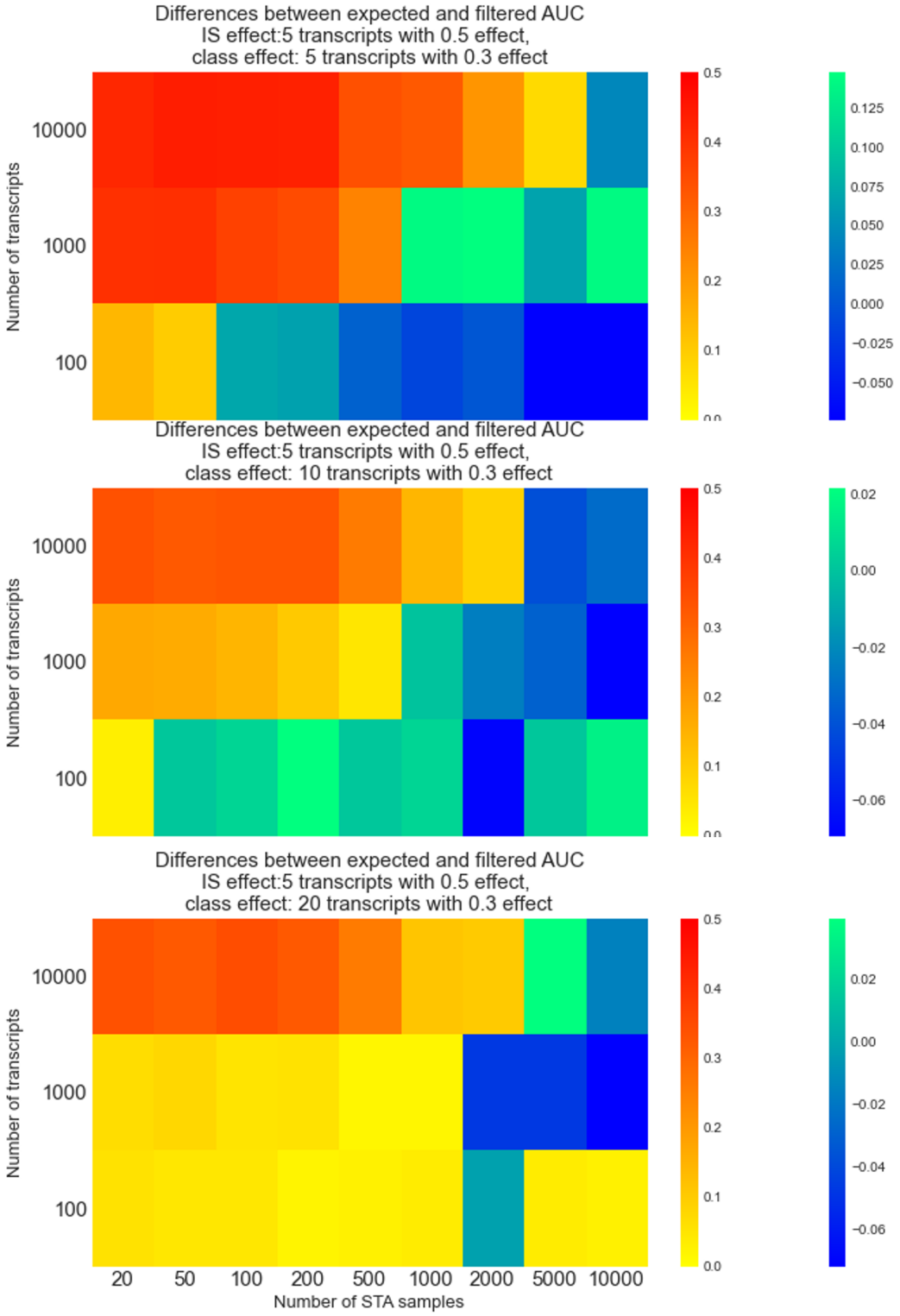



**Obrázek 4.10:** Rozdíly AUC hodnot mezi filtrovanými a očekávanými daty ve tvaru heatmapy

2. **Větší vliv třídy = lepší odstranění matoucích faktorů**

   Na obr. 4.11 všechny parametry zůstávají stejné, mění se jen počet ovlivněných třídou transkriptů. Na horním obrázku je třídou ovlivněno 5 transkriptů, uprostřed — 10, dolů — 20. Je vidět, že čím větší je počet ovlivněných třídou transkriptů, tím dřív dochází ke stavu definovaným intervalem 2 (4.4.2).

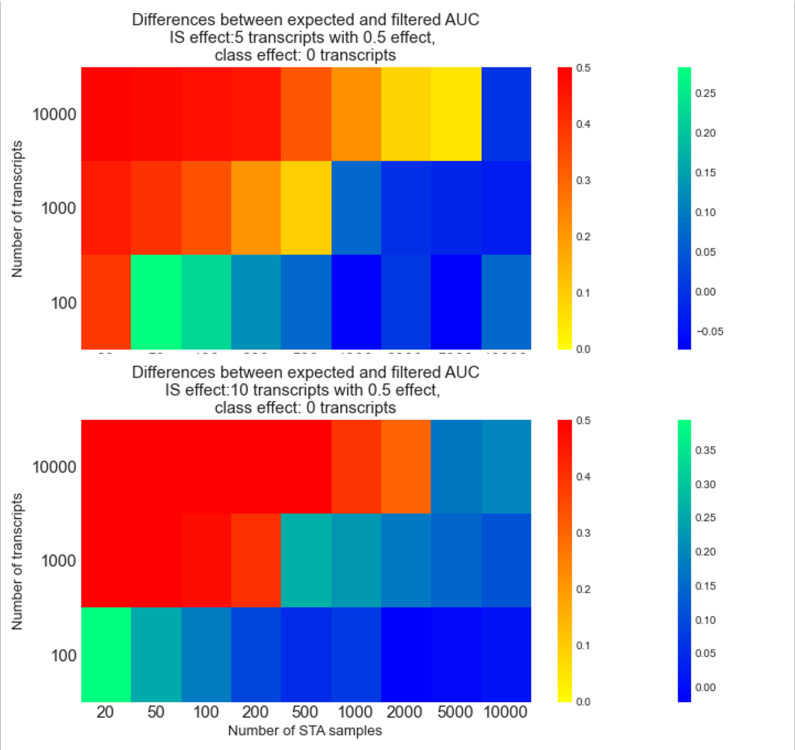
Nestabilní výsledky dolního obrázku pro 100 a 1000 transkriptů jsou způsobeny tím, že pro tento počet transkriptů generovaný vliv třídy byl příliš vysoký a vliv IS naopak malým, což téměř vede k ztotožnění origin a expected výsledků.



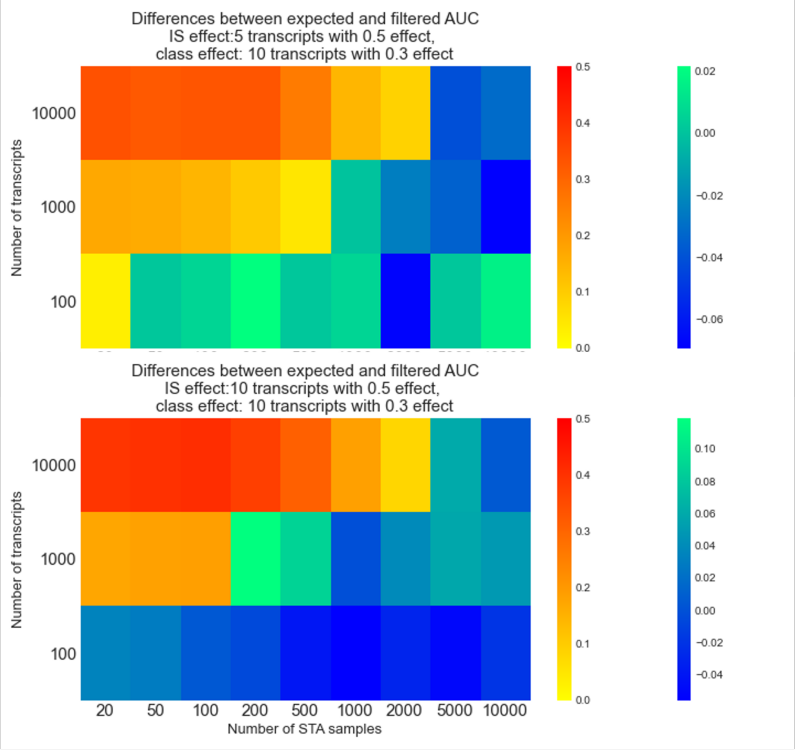

**Obrázek 4.11:** Rozdíly AUC hodnot mezi filtrovanými a očekávanými daty ve tvaru heat mapy. Mění se počet transkriptů ovlivněných třídou.

### 3. Chování při změně vlivu matoucích faktorů

Hodnoty ve stavu překompenzace jsou větší pro data s větším vlivem IS (dolní heatmapa má výraznější, "červenější" barvy), ale přechod do modrozeleného stavu nastává dříve, respektive u horního grafu přibližně stejně, i když na první pohled to tak nevypadá. Připomeňme proto, že heatmapy nepostihují odchylky AUC hodnot a proto pokud bod leží na horní části přímky, ale je zachycen odchylkou, heatmapa to neukáže.




**Obrázek 4.12:** Rozdíly AUC hodnot mezi filtrovanými a očekávanými daty ve tvaru heat mapy. Mění se počet transkriptů ovlivněných IS.




**Obrázek 4.13:** Rozdíly AUC hodnot mezi filtrovanými a očekávanými daty ve tvaru heat mapy. Mění se počet transkriptů ovlivněných IS.

### 4.4.5 Odhad parametrů reálných dat

V rámci této podkapitoly se pokusíme odhadnout parametry efektu třídy a IS v reálných datech. Cílem je odhadnout, které umělé nastavení lze považovat za nejrealističtější z pohledu reálných dat. Důležitým faktorem, na kterém bude odhad založen, je porovnání výsledků klasifikace a kompenzace provedených experimentů. Z 4.1.5 je znám odhad dolní a horní hranice celkového počtu ovlivněných transkriptů, z 4.2 a 4.3 výsledky klasifikace a kompenzace.

Cílem je tedy vyhledat a vybrat grafy podle následujících parametrů:

- celkový počet ovlivněných transkriptů = 5 až 81
- AUC origin[1] = 0.82
- AUC filtered[2]= 0.97
- celkový počet transkriptů = desítky tisíc.

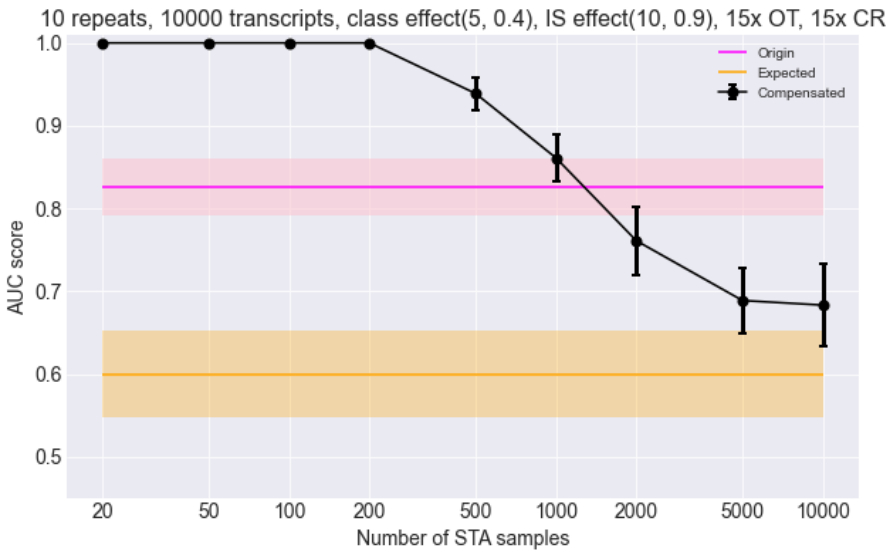
Mezi všemi výsledky se podařilo najít takové, které byly z hlediska popsaných výše parametrů podobné reálným datům. Tyto vybrané experimenty jsou shrnuty v tabulce 4.4 a následně znázorněny na obr. 4.14 - 4.17. Parametry $t_{class}$ a $t_{IS}$ v této tabulce označují počet transkriptů ovlivněných třídou, resp. léky.

[1]hodnota klasifikace původních dat
[2]hodnota klasifikace dat po kompenzaci

| Číslo | $t_{class}$ | Vliv třídy[3] | $t_{IS}$ | Vliv léků[4] | AUC origin | AUC expected | AUC filtered[5] | Bod č.1[6] |
|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 0.4 | 10 | 0.9 | 0.82 | 0.6 | 1.0 | 1000+ |
| 2 | 20 | 0.3 | 10 | 0.5 | 0.79 | 0.61 | 0.96 | 2000 |
| 3 | 20 | 0.3 | 20 | 0.5 | 0.84 | 0.7? | 0.96 | 1000 |
| 4 | 20 | 0.3 | 30 | 0.5 | 0.86 | 0.65 | 0.97 | 500+ |

**Tabulka 4.4:** Výsledky experimentů s parametry podobnými skutečným datům.

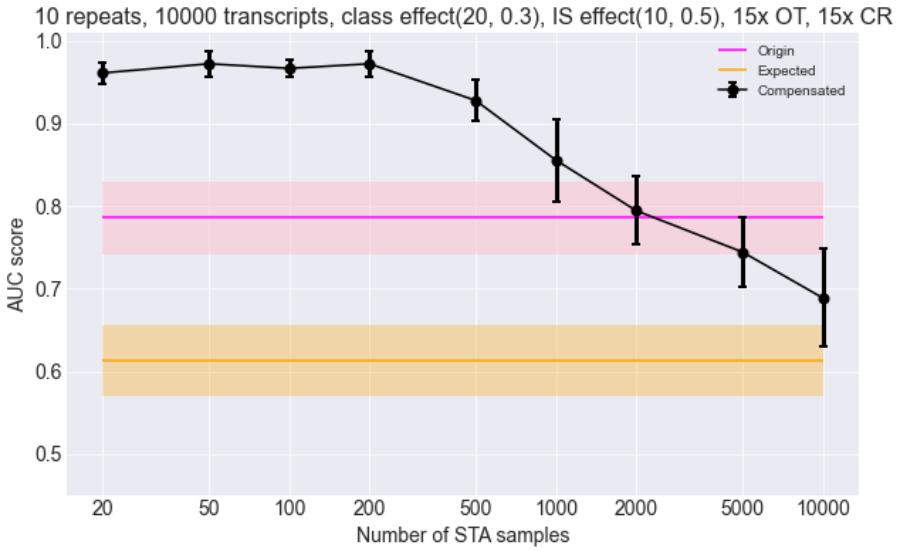
Graf č.1 (4.14) má stejné výsledky klasifikace (AUC origin) jako reálná data, proto byl mezi tyto vybrané grafy zařazen. Zajímavostí je, že zde k ideální kompenzace (tj. přiblížení k oranžové přímce) nedochází. Bod 1 (konec stavu prekompenzace) se nachází mezi 1000 a 2000 STA vzorků.




**Obrázek 4.14:** Graf č.1

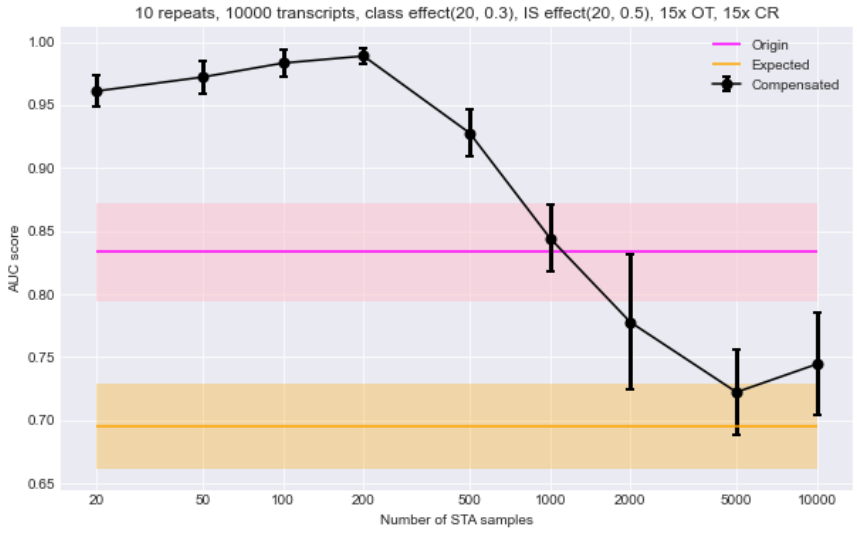
U dalšího obrázku (graf č.2 – 4.15) k ideální kompenzaci nedochází také. Výsledky klasifikace (AUC origin) tu jsou o trochu nižší, než u reálných dat (0.79 oproti 0.82). Bod 1 je zde větší než u výše popsaného grafu a tvoří 2000 STA vzorků.

[3]vliv třídy je představen normálním rozdělením s rozptylem 0.1 a danou střední hodnotou
[4]vliv léků je reprezentován náhodným rozdělením s danou hodnotou
[5]hodnota AUC po kompenzaci v bodě STA = 27 vzorků
[6]bod (představený počtem STA vzorků), ve kterém kompenzace začíná fungovat



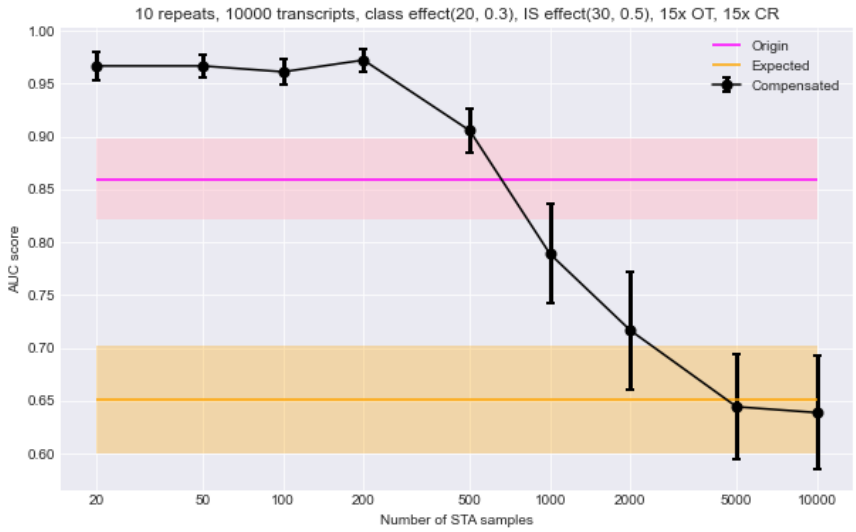

**Obrázek 4.15:** Graf č.2

Graf č.3, popsaný na obr. 4.16 má interval v bodech 1000 - 5000 (1000 STA vzorků pro fungování metody, 5000 – pro ideální kompenzaci). Vzestup na konci grafu se nejspíš vysvětluje náhodou.




**Obrázek 4.16:** Graf č.3

Počet STA vzorků, potřebných k fungování metody u grafu č.4 (obr. 4.17) je 500. K ideální kompenzaci zde dochází při STA = 5000.




**Obrázek 4.17:** Graf č.4

Graf č.4 ze všech předložených grafů vykazuje nejlepší výsledky z hlediska intervalu, popisujícího fungování metody. Hodnoty klasifikace se však nejvíc podobají skutečným u grafů č. 1 a č. 2. Zde se ale jedná o nejméně podobné skutečným výsledky kompenzace ze všech čtyř vybraných grafů.

# Závěr

V této bakalářské práci jsme se nejdříve seznámili s RNA-Seq technologií, od stručného úvodu molekulární biologii po popsaní povahy produkovaných RNA-Seq dat. Byla popsána interpretace daných dat a bylo potvrzeno, že se k poskytnutým datům nejlépe sedí Negativně binomické rozdělení.

Dále byl vysvětlen problém výskytu matoucích faktorů v datech, provedena byla rešerše existujících metod pro odstranění vlivu těchto faktorů a bylo předloženo porovnání daných metod. V rámci této bakalářské práce byla pro poskytnutá data zvolena metoda naučení se vlivu pomocí multivariačního regresního modelu. Na reálných datech kompenzace neproběhla úspěšně z důvodu malého počtu vzorků, proto byl představen generátor umělých RNA-Seq dat, který umožnil provést širokou škálu experimentů.

Cílem experimentů bylo odhadnout počet vzorků nutných ke správné kompenzaci vlivu matoucích faktorů a také odhadnout klíčové parametry (vliv třídy vzorku, vliv matoucích faktoru) reálných dat. Každý experiment na umělých datech obsahoval tři fázi: klasifikace originálních dat, jejich kompenzace a klasifikace filtrovaných dat. Z výsledků experimentů bylo posouzeno, že úspěšnost kompenzace matoucích faktorů ovlivňují hlavně tři faktory:

1. Celkový počet transkriptů v datech (víc transkriptů = složitější odstranění matoucích faktorů)
2. Vliv třídy vzorku (větší vliv = lepší odstranění matoucích faktorů)
3. Vliv matoucích faktoru (větší vliv = těžší odstranění matoucích faktorů).

Experimenty odhalily, že dolní hranici, u které metoda začíná fungovat, je 500 STA vzorků. Takový efekt se pozoruje u experimentu s nastavením $t_{class} = 20$, vliv třídy $\sim N(0.3, 0.1)$, $t_{IS} = 30$, vliv IS = 0.5. Pro toto nastavení se k ideální kompenzaci přibližuje při 5000 STA vzorků.

Práce nabízí mnoho prostoru pro budoucí pokračování, především rozšíření škály parametrů, např. počtu vzorků OT a CR třídy, počtu imunosupresivních léků.

Další možností by bylo použití jiných typů klasifikátorů, zejména SVM, který ukazoval nejlepší výsledky na poskytnutých datech. Vzhledem k časové náročnosti každého experimentu, klasifikace se prováděla pouze jedním klasifikátorem, a to pomocí logistického regresního modelu. Provedené experimenty také obsahují šum způsobený náhodou – proto je nezbytné se zaměřit na eliminaci tohoto šumu např.

zvýšením počtu opakování (nynější experimenty byly použity s 10 opakování).

Možné pokračování práce zahrnuje také rozšíření generátoru a použití jiných statistických modelů pro simulaci dat. Nyní je například vliv třídy reprezentován normálním rozdělením a vliv imunosupresiv – rovnoměrným. V budoucnu by bylo možné použít složitější modely.

# Bibliografie


1. ALBERTS, B; JOHNSON, A; LEWIS, J. *Molecular Biology of the Cell. 4th edition.* New York: Garland Science, 2002. Dostupné také z: https://www.ncbi.nlm.nih.gov/books/NBK21054/.
2. GOKSULUK, Dincer et al. MLSeq: Machine learning interface for RNA sequencing data. *Computer Methods and Programs in Biomedicine*. 2019, roč. 175, s. 223–231. ISSN 0169-2607. Dostupné z DOI: doi:10.1016/j.cmpb.2019.04.007.
3. SPONK. *Rozdíly mezi DNA a RNA* [https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-CS.svg]. 2014. Cit. 09.11.2021.
4. *Labster theory* [https://theory.labster.com/central_dogma_molecular_biology_pre/]. [N.d.]. Cit. 08.11.2021.
5. STARK, R.; GRZELAK, M.; HADFIELD, J. RNA sequencing: the teenage years. *Nature reviews. Genetics*. 2019, roč. 20, s. 631–656. Dostupné z DOI: doi:10.1038/s41576-019-0150-2.
6. *RNA-seqlopedia* [https://rnaseq.uoregon.edu/#exp-design]. [N.d.]. Cit. 08.08.2021.
7. WANG, Zhong; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*. 2009, roč. 10, č. 1, s. 57–63. ISSN 1471-0064. Dostupné z DOI: doi:10.1038/nrg2484.
8. HOWE, Kevin L et al. Ensembl 2021. *Nucleic Acids Res*. 2021, roč. 49, č. 1, s. 884–891. Dostupné z DOI: doi:10.1093/nar/gkaa942.
9. WITTEN, Daniela M. Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*. 2011, roč. 5, č. 4, s. 2493–2518. Dostupné z DOI: doi:10.1214/11-AOAS493.
10. MARIONI, J. C. et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008, roč. 18, č. 9, s. 1509–1517. Dostupné z DOI: doi:10.1101/gr.079558.108.
11. ROBINSON, Mark D.; SMYTH, Gordon K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007, roč. 23, č. 21, s. 2881–2887. Dostupné z DOI: doi:10.1093/bioinformatics/btm453.
12. *SciPy documentation* [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.nbinom.html]. [N.d.]. Cit. 08.08.2021.

13. GRIMES, David A; SCHULZ, Kenneth F. Bias and causal associations in observational research. *The Lancet*. 2002, roč. 359, č. 9302, s. 248–252. ISSN 0140-6736. Dostupné z DOI: doi:10.1016/S0140-6736(02)07451-2.

14. BROOKHART, M. Alan et al. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*. 2010, roč. 48, s. 114–120. Dostupné z DOI: doi:10.1097/MLR.0b013e3181dbebe3.

15. POURHOSEINGHOLI, M. A.; BAGHESTANI, A. R.; VAHEDI, M. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*. 2012, roč. 5, č. 2, s. 79–83. ISSN 2008-2258. Dostupné také z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017459/.

16. MIETTINEN, Olli S. Stratification by a multivariate confounder score. *American Journal of Epidemiology*. 1976, roč. 104, č. 6, s. 609–620. ISSN 0002-9262. Dostupné z DOI: doi:10.1093/oxfordjournals.aje.a112339.

17. MCNAMEE, R. Regression modelling and other methods to control confounding. *Occupational and environmental medicine*. 2005, roč. 62, č. 7, s. 500–506. ISSN 1351-0711. Dostupné z DOI: doi:10.1136/oem.2002.001115.

18. GRJIBOVSKI, Andrej et al. Propensity score matching as a modern statistical method for bias control in observational studies with binary outcome. *Human Ecology*. 2016, roč. 44, s. 50–64. Dostupné z DOI: doi:10.33396/1728-0869-2016-5-50-64.

19. KAHLERT, J. et al. Control of confounding in the analysis phase – an overview for clinicians. *Clinical epidemiology*. 2017, roč. 9, s. 195–204. Dostupné z DOI: doi:10.2147/CLEP.S129886.

20. LIANG, Wenbin; ZHAO, Yuejen; LEE, Andy H. An Investigation of the Significance of Residual Confounding Effect. *BioMed Research International*. 2014, roč. 2014, s. 658056. Dostupné z DOI: doi:10.1155/2014/658056.

21. JAGER, K.J. et al. Confounding: What it is and how to deal with it. *Kidney International*. 2008, roč. 73, č. 3, s. 256–260. ISSN 0085-2538. Dostupné z DOI: doi:10.1038/sj.ki.5002650.

22. MASSART, A. et al. Operational tolerance in kidney transplantation and associated biomarkers. *Clinical and experimental immunology*. 2017, roč. 189, č. 2, s. 138–157. Dostupné z DOI: doi:10.1111/cei.12981.

23. *Differential gene expression (DGE) analysis* [https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html]. [N.d.]. Cit. 15.1.2021.

24. YANG, Sheng et al. A Systematic Evaluation of Feature Selection and Classification Algorithms Using Simulated and Real miRNA Sequencing Data. *Computational and mathematical methods in medicine*. 2015, roč. 2015. Dostupné z DOI: doi:10.1155/2015/178572.

25. JAMES, Gareth et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370. Dostupné z DOI: doi:10.5555/2517747.

26. FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*. 2010, roč. 33, č. 1, s. 1–22. Dostupné také z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/.

27. NOBLE, William S. What is a support vector machine? *Nature biotechnology*. 2006, roč. 24, č. 12, s. 1565–1567.

28. PETERSON, Leif E. K-nearest neighbor. *Scholarpedia*. 2009, roč. 4, č. 2, s. 1883.

29. BERRAR, Daniel. Cross-Validation. In: 2018. ISBN 9780128096338. Dostupné z DOI: 10.1016/B978-0-12-809633-8.20349-X.

30. FAWCETT, Tom. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006, roč. 27, č. 8, s. 861–874. ISSN 0167-8655. Dostupné z DOI: doi:10.1016/j.patrec.2005.10.010.

31. CHRISTAKOUDI, Sofia et al. Development and validation of the first consensus gene-expression signature of operational tolerance in kidney transplantation, incorporating adjustment for immunosuppressive drug therapy. *EBioMedicine*. 2020, roč. 58, s. 102899. ISSN 2352-3964. Dostupné z DOI: doi:10.1016/j.ebiom.2020.102899.

32. *Analýza genové exprese: RNA-Seq nebo kvantitativní PCR?* [https://www.baria.cz/blog/analyza-genove-exprese-rna-seq-nebo-kvantitativni-pcr/]. [N.d.]. Cit. 23.06.2021.

33. SEABOLD, Skipper; PERKTOLD, Josef. Statsmodels: Econometric and statistical modeling with python. In: *9th Python in Science Conference*. 2010.

34. HEBENSTREIT, Daniel et al. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*. 2011, č. 7, s. 497. Dostupné z DOI: https://doi.org/10.1038/msb.2011.28.

35. MILLER, Steven J. Chapter 24. The Method of Least Squares. In: *The Probability Lifesaver*. Princeton University Press, 2017, s. 625–635. Dostupné z DOI: doi:10.1515/9781400885381-026.

36. HARRIS, Charles R. et al. Array programming with NumPy. *Nature*. 2020, roč. 585, č. 7825, s. 357–362. Dostupné z DOI: 10.1038/s41586-020-2649-2.

37. PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011, roč. 12, s. 2825–2830.

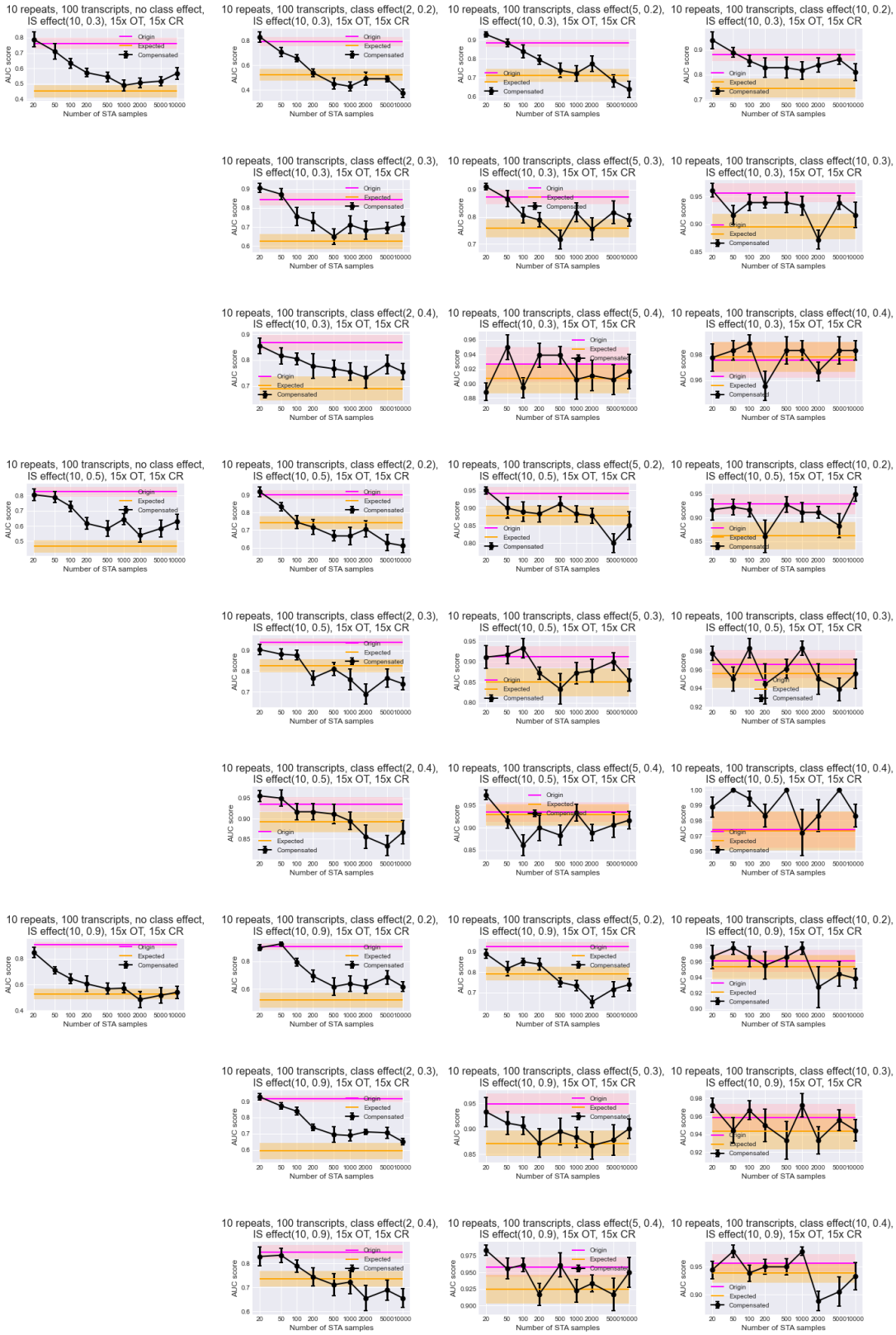
# Přílohy

## A Obsah přiloženého média

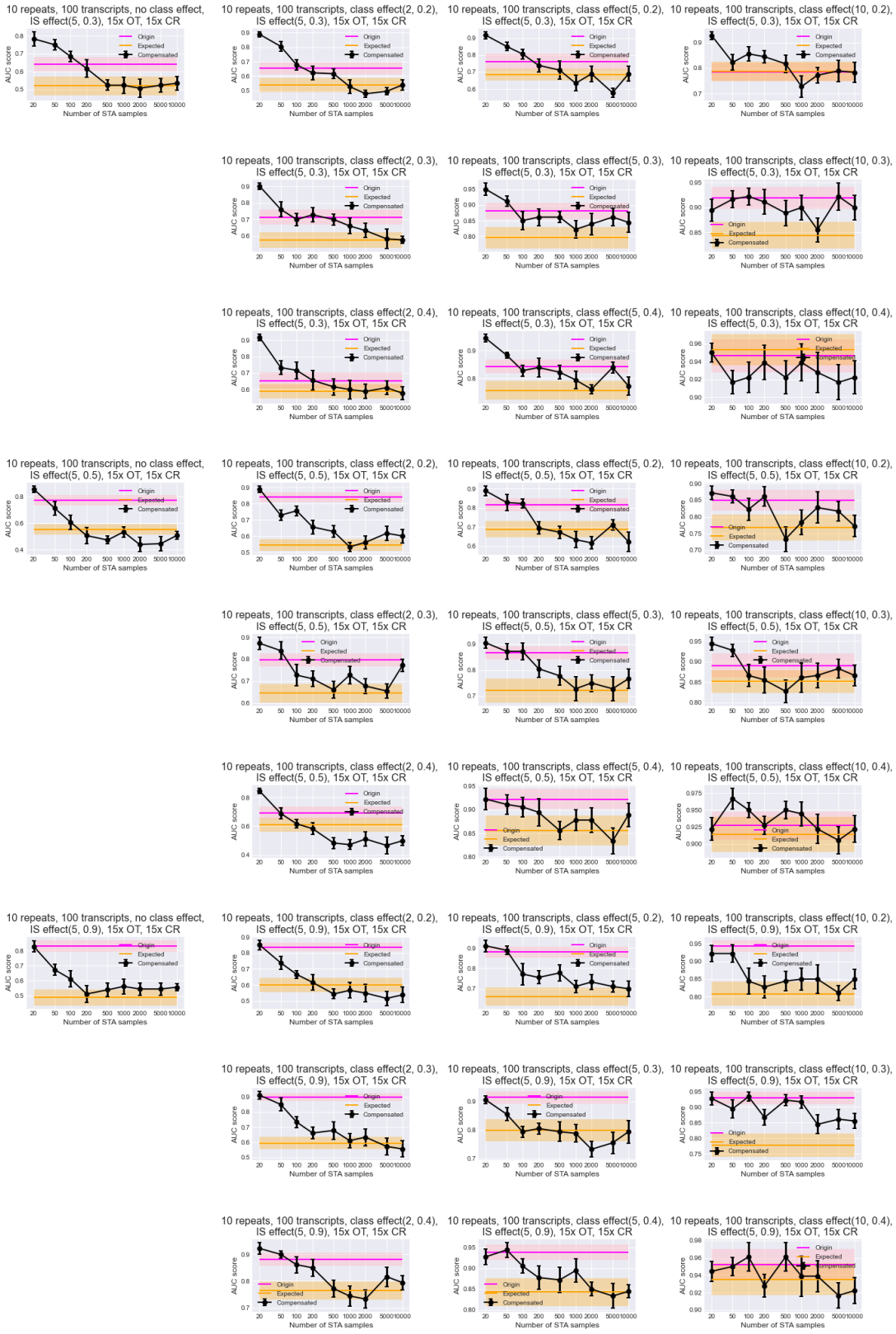
```
src.zip ..... zdrojové kódy implementace
thesis.zip ..... zdrojová forma práce ve formátu LaTeX
thesis.pdf ..... text práce ve formátu PDF
```

# B Výsledky experimentů

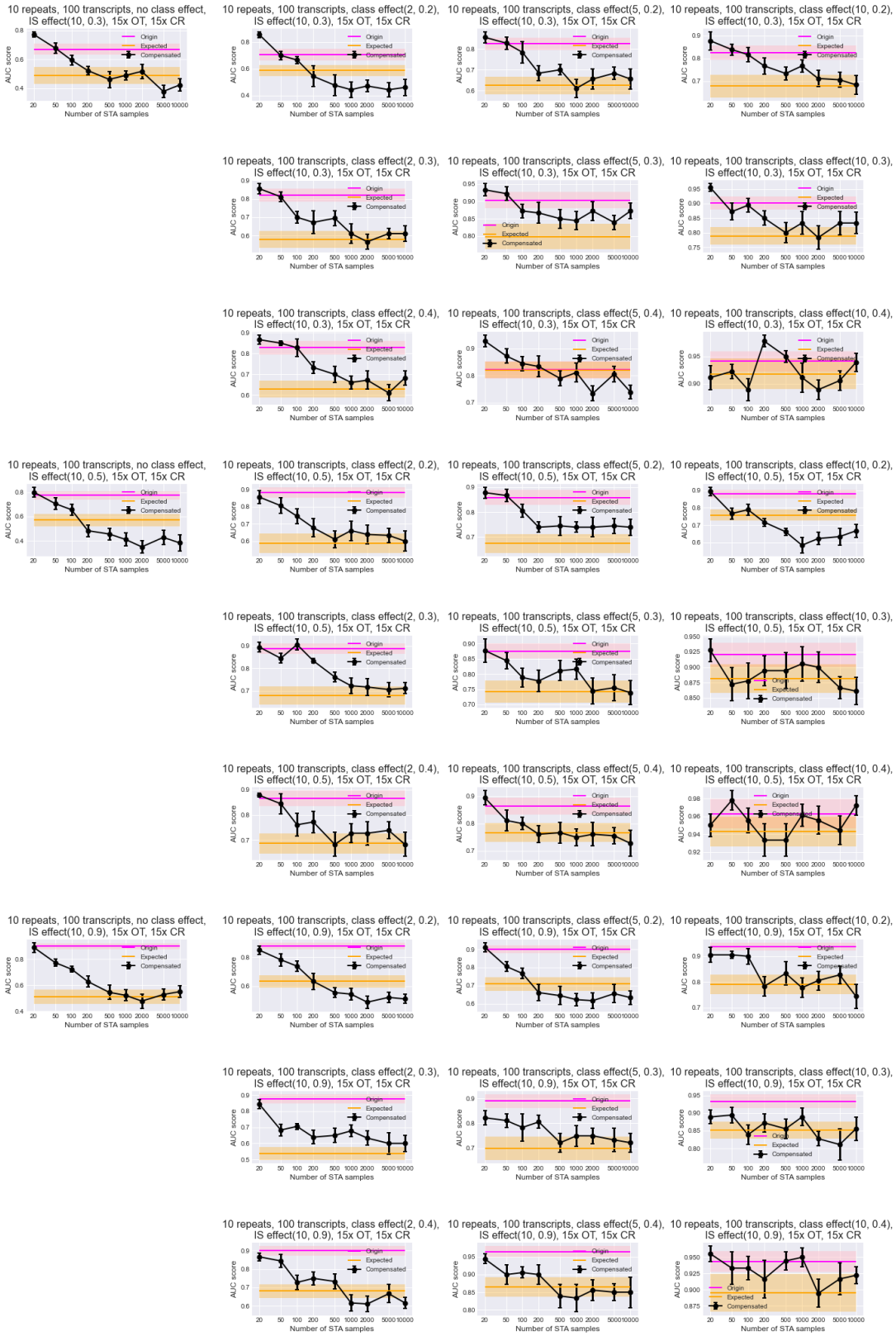

**Obrázek 18:** 100 transkriptu. Poissonovo rozdělení. 5 ovlivněných imunosupresivy



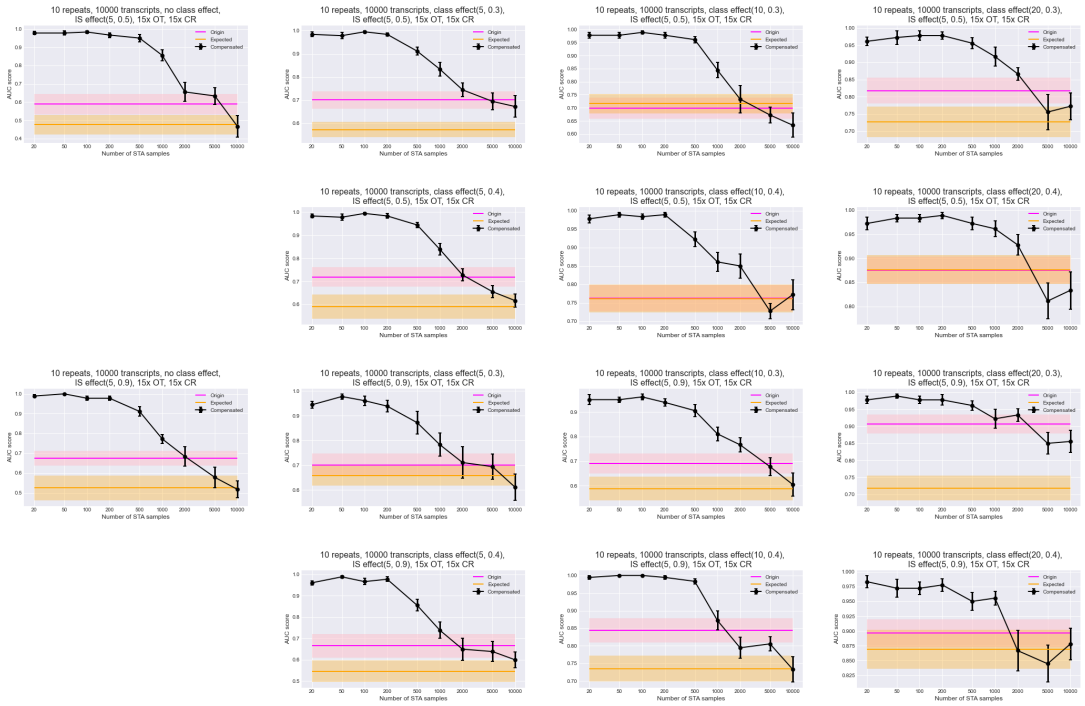

**Obrázek 19:** 100 transkriptu. Poissonovo rozdělení. 10 ovlivněných imunosupresivy



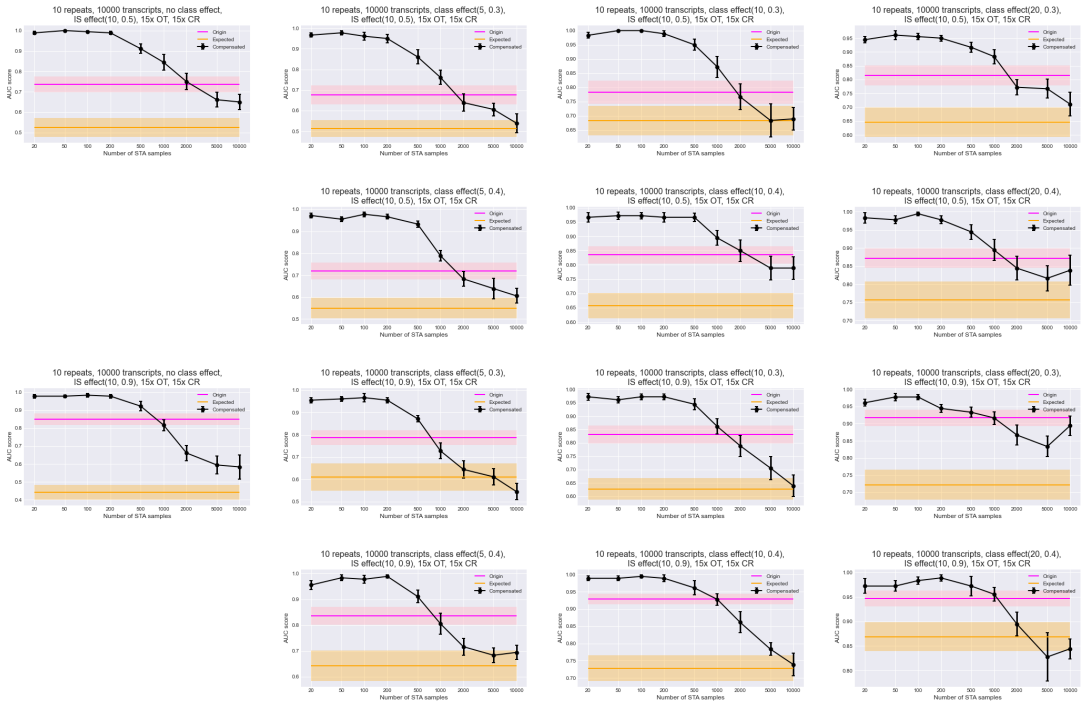

**Obrázek 20:** 100 transkriptu. NB rozdělení. 5 ovlivněných imunosupresivy



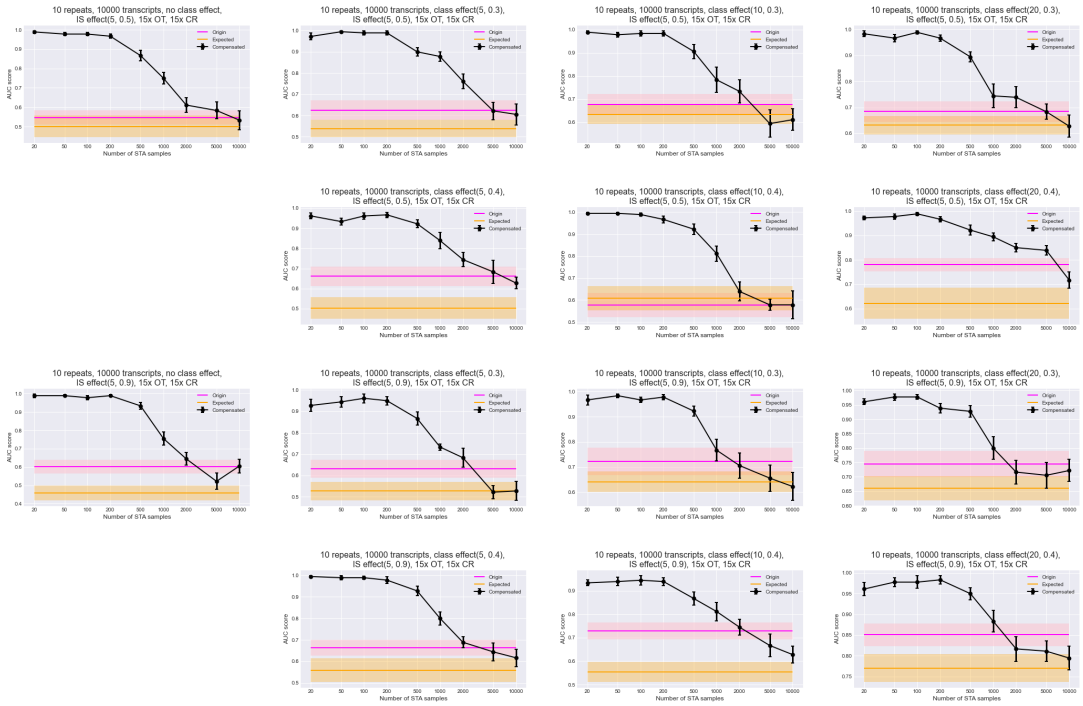

**Obrázek 21:** 100 transkriptu. NB rozdělení. 10 ovlivněných imunosupresivy



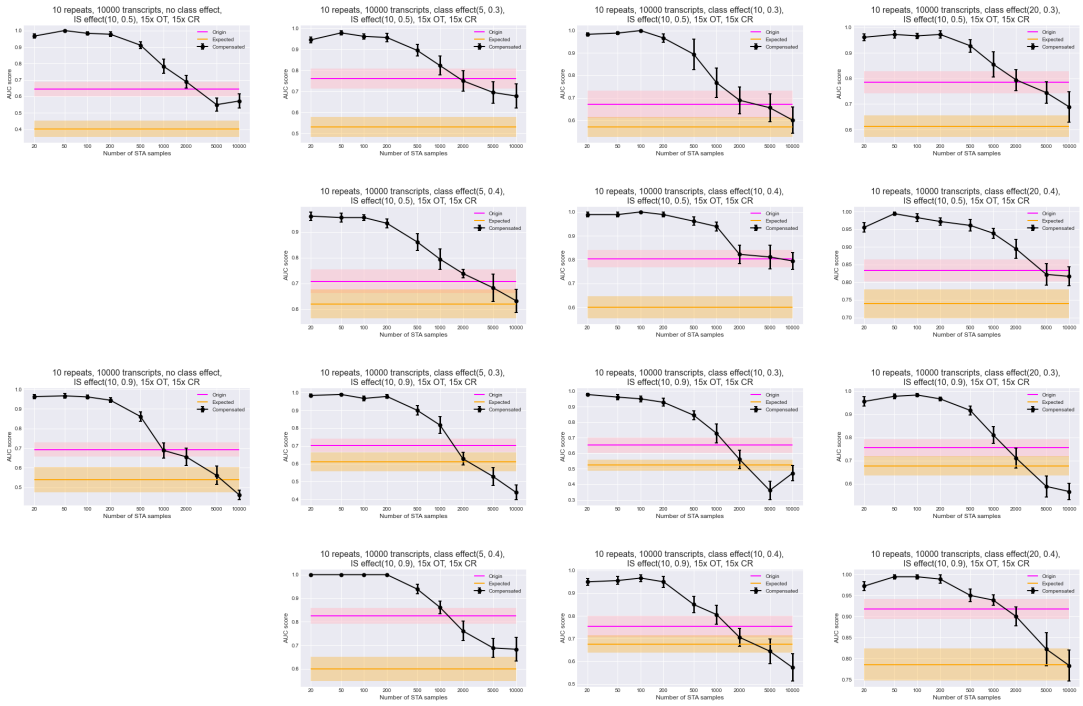

**Obrázek 22:** 10 tisic transkriptu. Poissonovo rozdělení. 5 ovlivněných imunosupresivy




**Obrázek 23:** 10 tisic transkriptu. Poissonovo rozdělení. 10 ovlivněných imunosupresivy




**Obrázek 24:** 10 tisic transkriptu. NB rozdeleni. 5 ovlivnenych imunosupresivy




**Obrázek 25:** 10 tisic transkriptu. NB rozdeleni. 10 ovlivnenych imunosupresivy