Omezení predikčních programů pro určování důsledků genomických variant

Kateřina Břicháčková

Univerzita Karlova

Přírodovědecká fakulta

10. září 2018

Čím se práce zabývá

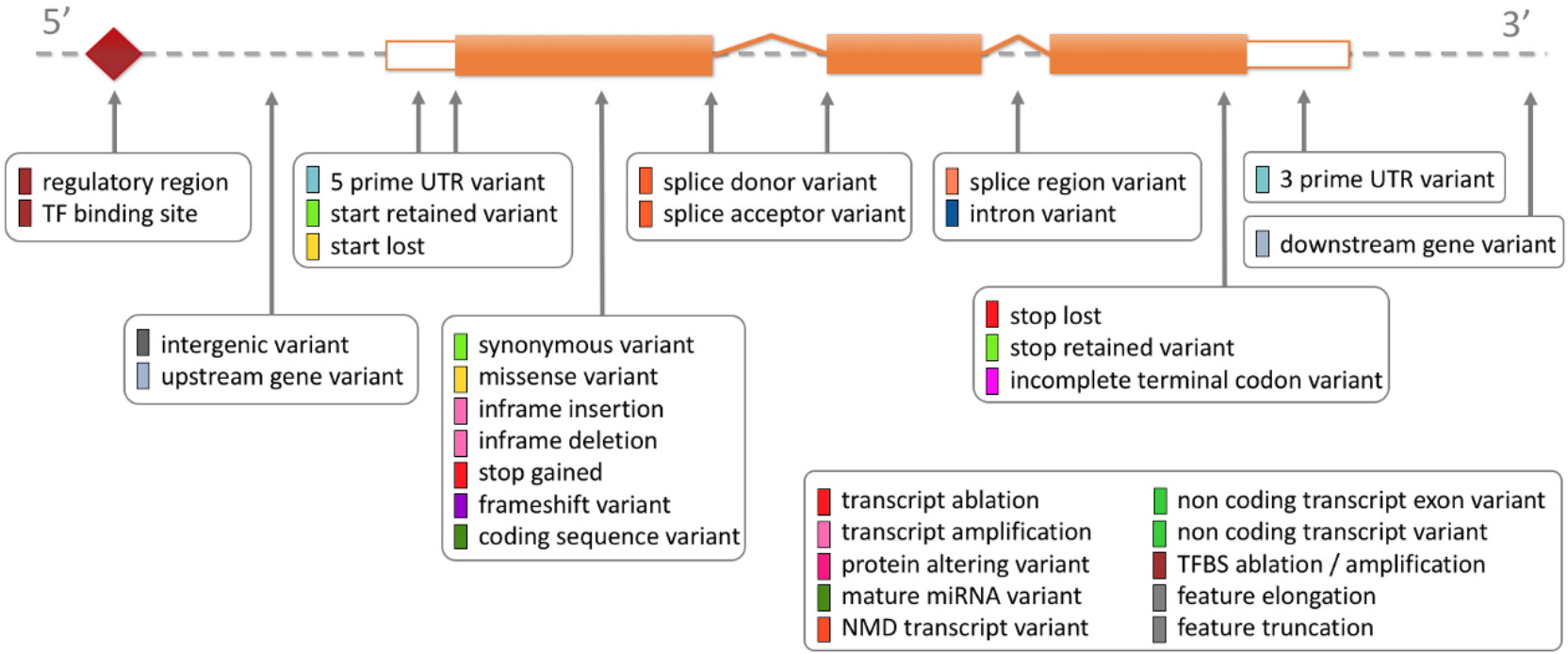
- Predikce důsledků genomických variant
- Reprezentace genomických dat
- Datové formáty (GFF3, FASTA, Fai index, HGVS notace, VCF, Ensembl default, Annovar input)
- Popis programů pro anotaci variant - algoritmy, práce s programy
- Praktická část - sada testů, problematické oblasti a situace

Genomická data

- Referenční genomická sekvence
  - Vytvořena ze sekvence mnoha jedinců
  - Reference genome assembly - GRCh38 (2013)
- Genomické varianty - odlišnosti od referenční sekvence
  - SNP (single nucleotide polymorphism)
  - indely (inzerce + delece)
- Genomové anotace
  - Geny, transkripty, kódující oblasti, regulační oblasti
  - RefSeq (NCBI)
  - Ensembl anotace
  - GENCODE

Predikce funkčních důsledků variant

- Velké množství dat
- Anotační programy pro predikci důsledků variant
- Filtrování, kategorizace a prioritizace variant
- Nejdůležitější v kódujících oblastech



Programy pro určování důsledků genomických variant

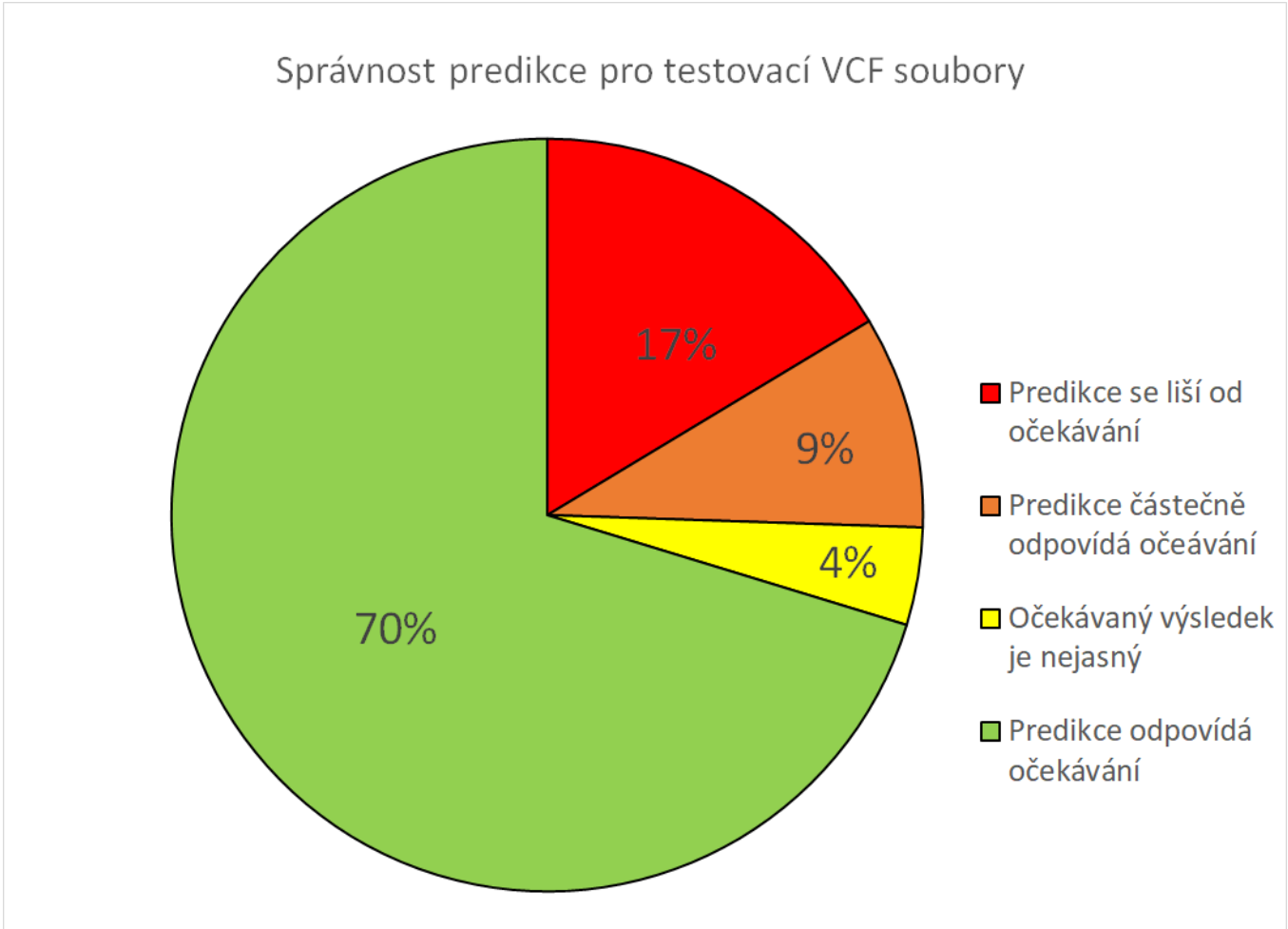
- Variant Effect Predictor (VEP)
- Haplosaurus
- ANNOVAR
- BCFtools/csq
- SnpEff

Praktická část

- Cíle
  - Vytvořit sadu testů (VCF - Variant Call Format)
  - Porovnat výsledky pěti programů
  - Diskutovat výsledky, zdůraznit situace, ve kterých je predikce problematická
- Ensembl transkripty (release 92, 5.4.2018)
- GRCh38.p12

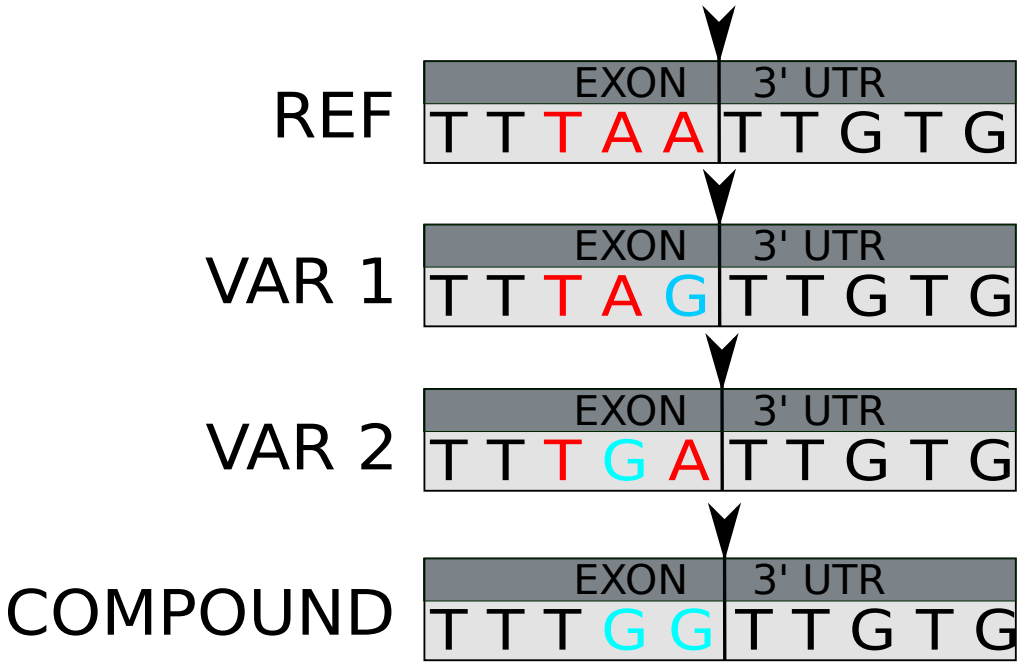
Výsledky

- 105 VCF testovacích souborů



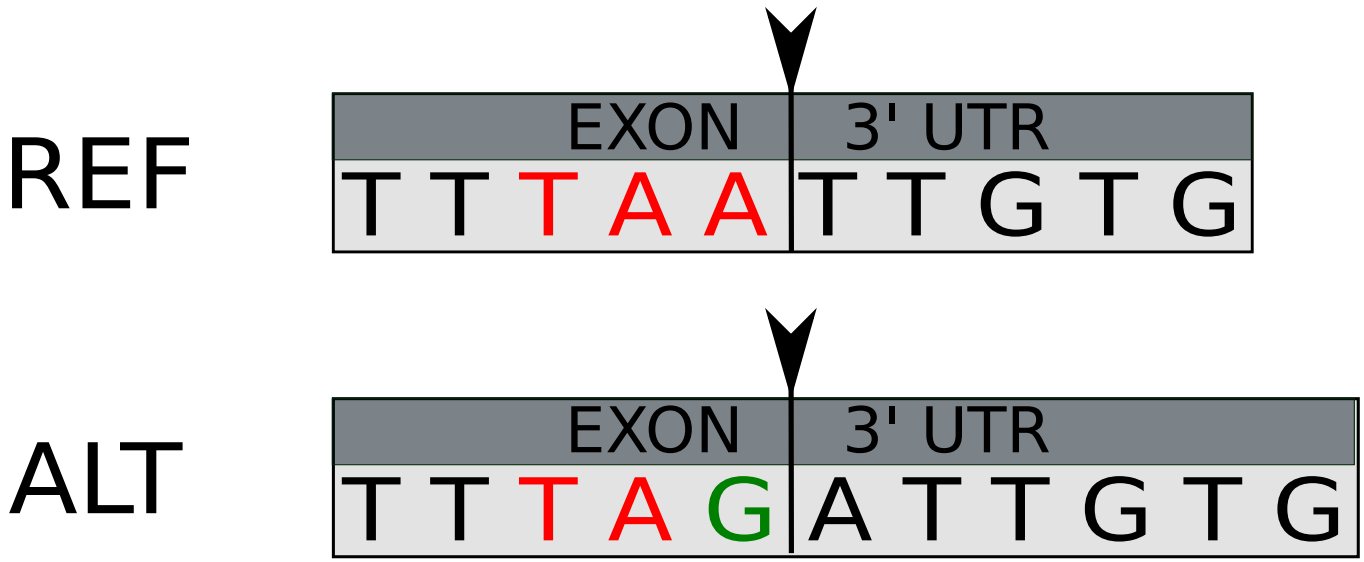
“Haplotype-aware consequence calling”

- Společný efekt více variant
- BCFtools/csq, Haplosaurus



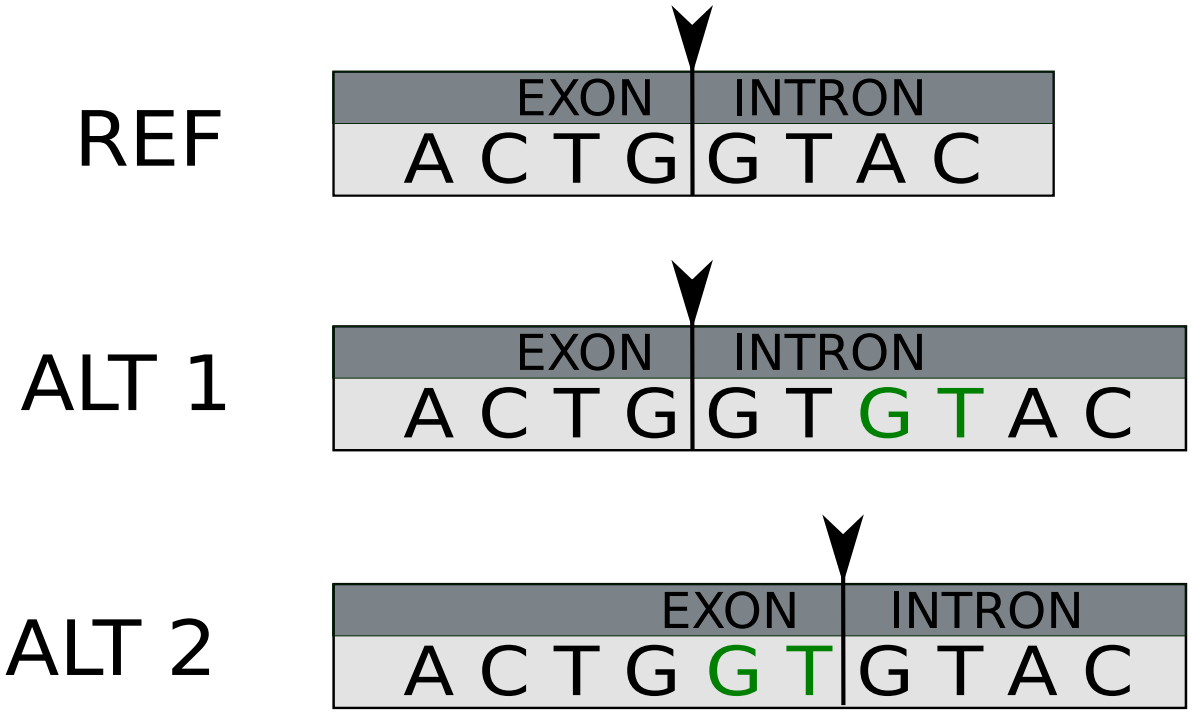
Problematické oblasti - *stop_retained_insertion.vcf*

- Stop kodon zachován
- VEP a BCFtools/csq správně *stop_retained_variant*
- Všech pět chybně *frameshift_variant*



Problematické oblasti - *splice_region_insertion.vcf*

- Varianty reprezentovány nejednoznačně
- Očekávaný důsledek nejednoznačný - frameshift a splice region?

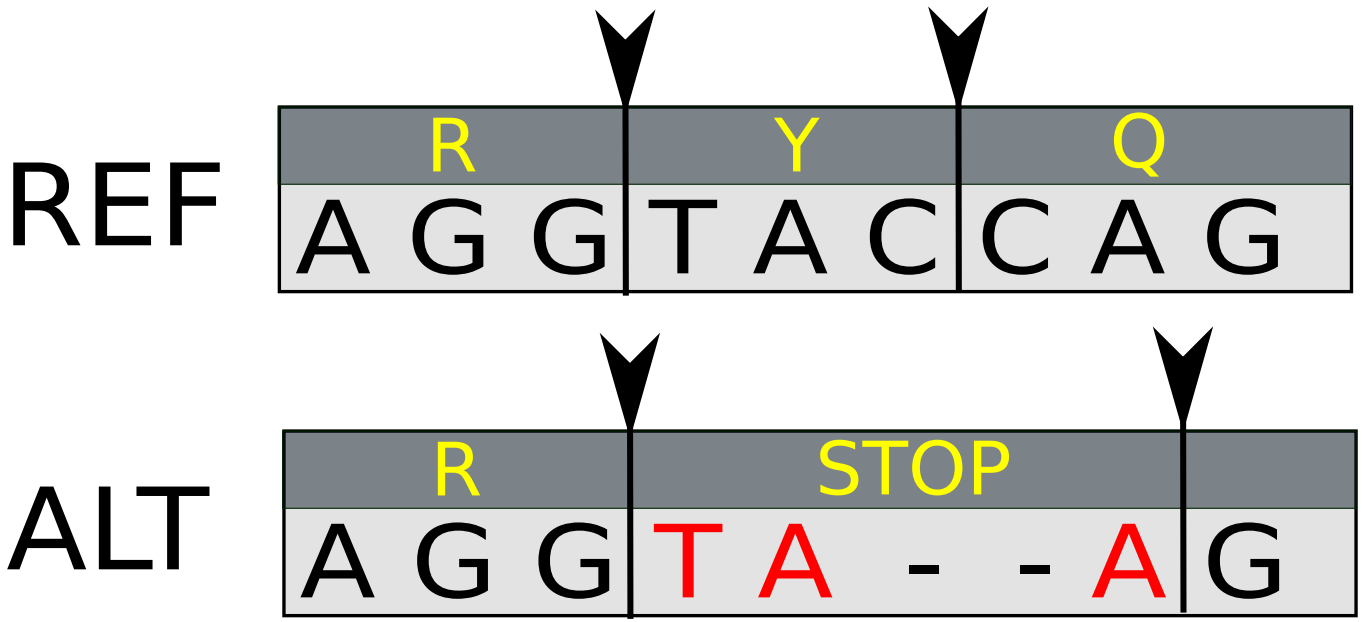


Problematické oblasti - *splice_region_insertion.vcf*

	ALT 1	**ALT 2**
ANNOVAR	splicing & intronic	frameshift insertion
VEP	splice_region_variant & intron_variant	frameshift_variant & splice_region_variant
SnpEff	splice_region_variant & intron_variant	splice_region_variant & intron_variant
BCFtools/csq	splice_region	synonymous & splice_donor
Haplosaurus	XXX	frameshift & indel

Problematické oblasti - *stop_gained_deletion.vcf*

- Vytvořen stop kodon
- VEP správně *stop_gained_variant*
- Všechny ostatní chybně *frameshift_variant*



Závěr

- Důležitost predikce důsledků variant
- Existující programy pro určování důsledků variant nejsou bezchybné
- Použít různé nástroje a porovnat výsledky
- Někdy není očekávaný důsledek zřejmý

Děkuji za pozornost

Poznámky

- Nejnovější verze GRCh38, předchozí GRCh37 (Genome Reference Consortium)
- Genome Reference Consortium
  - Wellcome Sanger Institute
  - The McDonnell Genome Institute at Washington University
  - The European Bioinformatics Institute
  - The National Center for Biotechnology Information
  - The Zebrafish Model Organism Database
- International Human Genome Sequencing Consortium (IHGSC)

UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
hg18	Mar. 2006	NCBI Build 36.1	Available
hg17	May 2004	NCBI Build 35	Available
hg16	Jul. 2003	NCBI Build 34	Available
hg15	Apr. 2003	NCBI Build 33	Archived