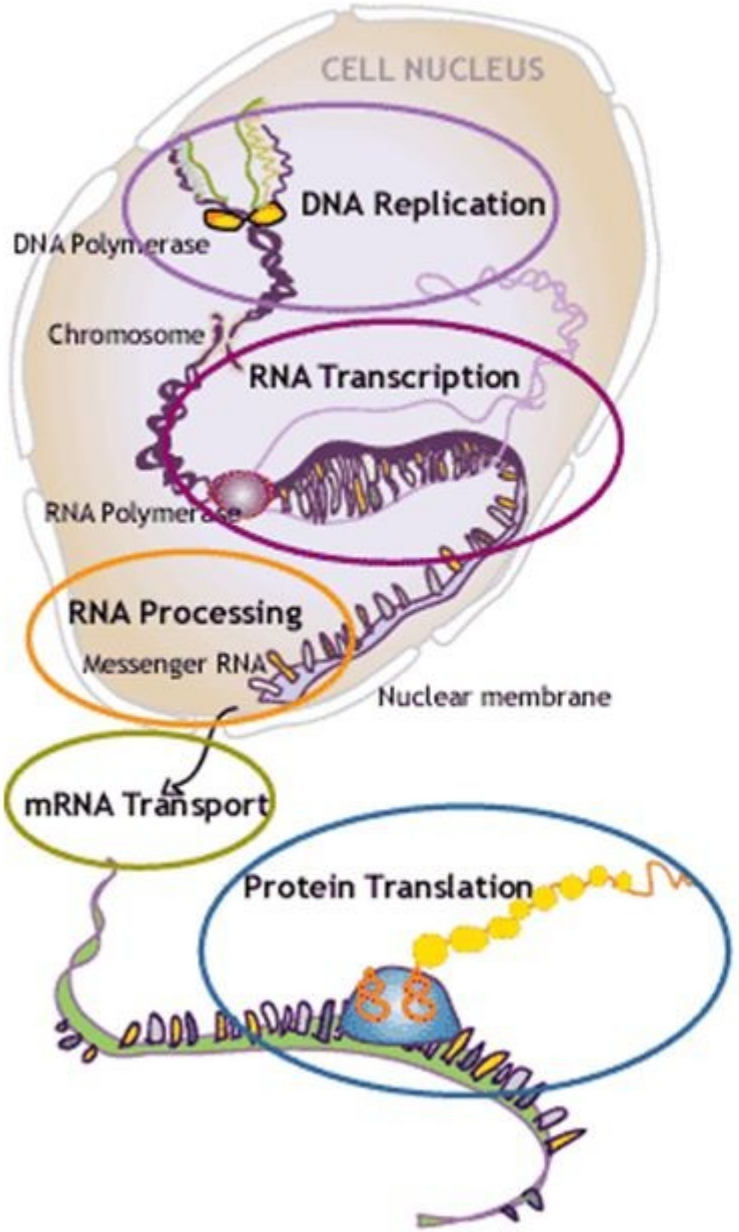
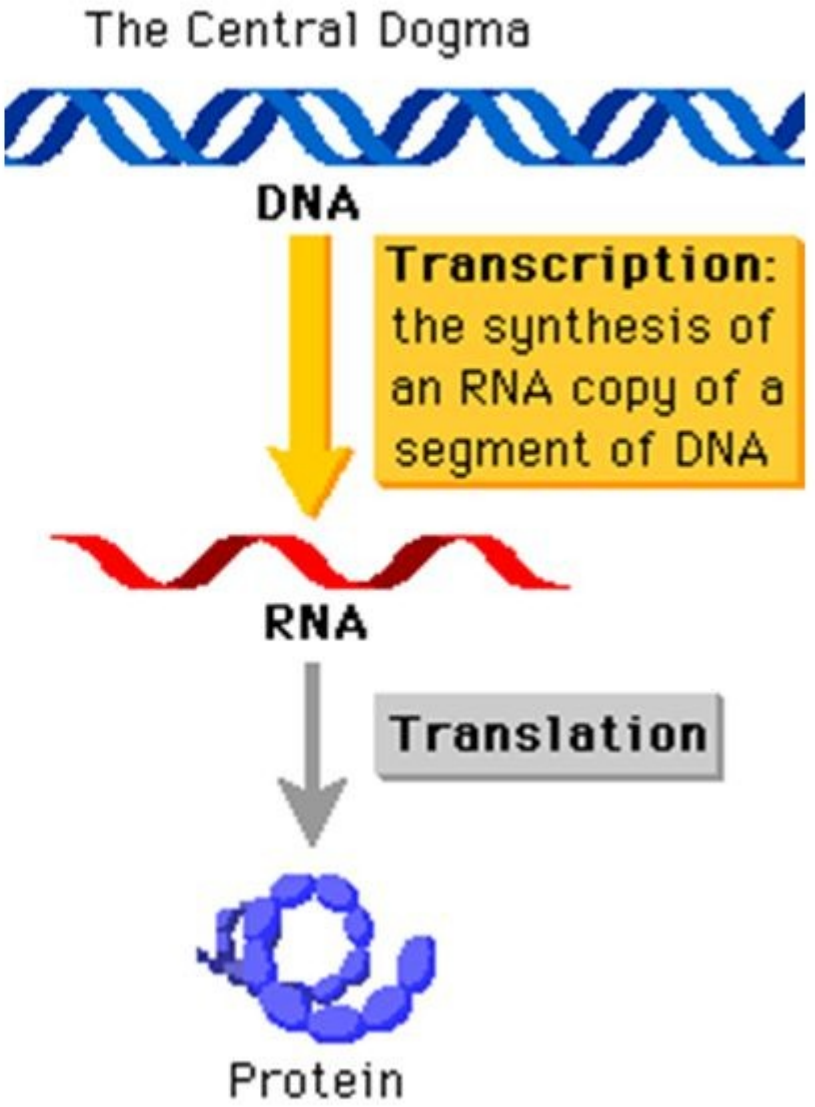
NGS – analýza dat

kroužek, 16.12.2016

Alena Musilová

DNA → RNA → Protein



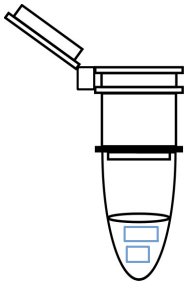


Typy NGS experimentů

Název	Materiál	Cílí na ..?	Cíl experimentu?
amplikon	DNA	malý počet vybraných genů	hledání variant
exom	DNA	všechny geny	hledání variant
transkriptom (RNA-seq)	RNA	dle typu RNA, nejčastěji protein-kódující	hledání exprimovaných variant, exprese RNA
ChIP-seq	DNA	sekvence DNA, které váže vybraný protein	hledání vazebných míst proteinů
CLIP-seq	RNA	sekvence RNA, které váže vybraný protein	interakce proteinů s RNA
...	...	...	...

Příprava knihovny a sekvenace

1) izolace DNA / RNA



2) RNA na cDNA / DNA

3) fragmentace

4) adaptery

5) sekvenování

```
@HWI-7001454:80:C3PEGACXX:5:1207:11848:82728
GCAGAAGAGAAGGCGTGGCATAAGAGAAGACGCGGTTGTTCTGTAGAGAG
+
??<D;?22AA;;;E:1@EDC3CE<9?;@DADDD<DDD<AC@8=C######
```

Základní schéma analýzy NGS dat

1. Primární data

Sekvence (ready) ve formátu fastq

```
@HWI-7001454:80:C3PEGACXX:5:1207:11848:82728
GCAGAAGAGAAGGCGTGGCATAAGAGAAGACGCGGTTGTTCTGTAGAGAG
+
??<D;?22AA;;;E:1@EDC3CE<9?;@DADDD<DDD<AC@8=C######
```

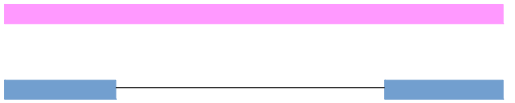
- několik desítek milionů sekvencí (50 – 150 bází), což odpovídá několika GB dat na 1 experiment

→ kontrola kvality primárních dat (např. program FastQC)

FastQC 

single-end vs **pair-end** sekvenování

single-end



pair-end

Základní informace o experimentu

Basic Statistics

Measure	Value
Filename	C7B26ACXX_459D_15s010768-1-1_Martina_lane215s010768_1_sequence.txt.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	33504351
Sequences flagged as poor quality	0
Sequence length	100
%GC	50

vždy

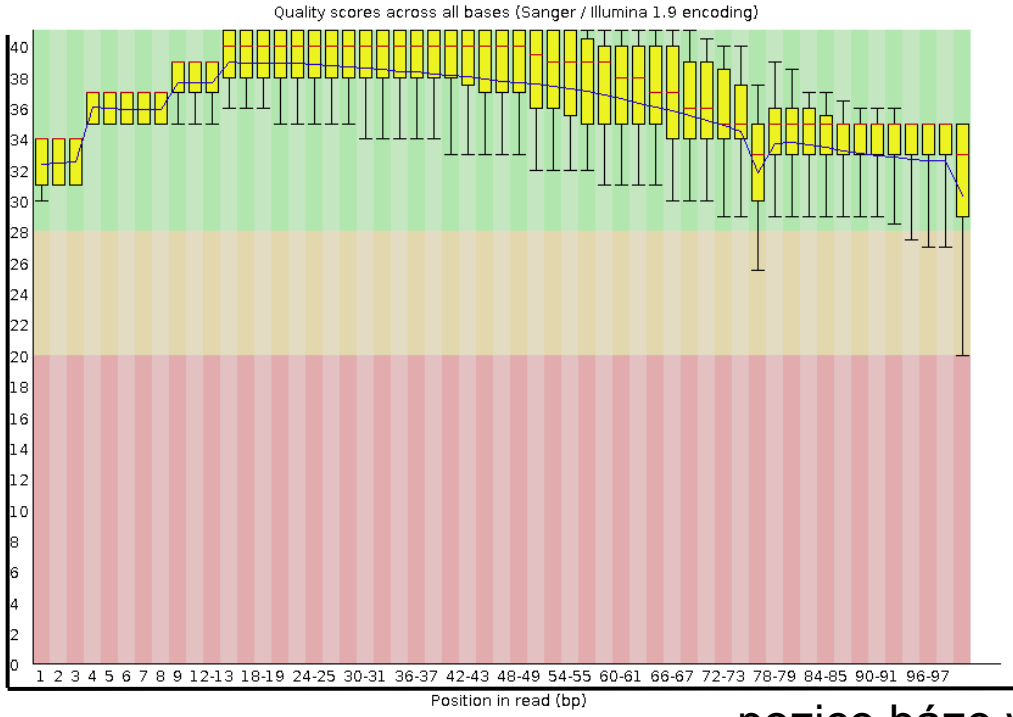
FastQC

Kvalita sekvenace

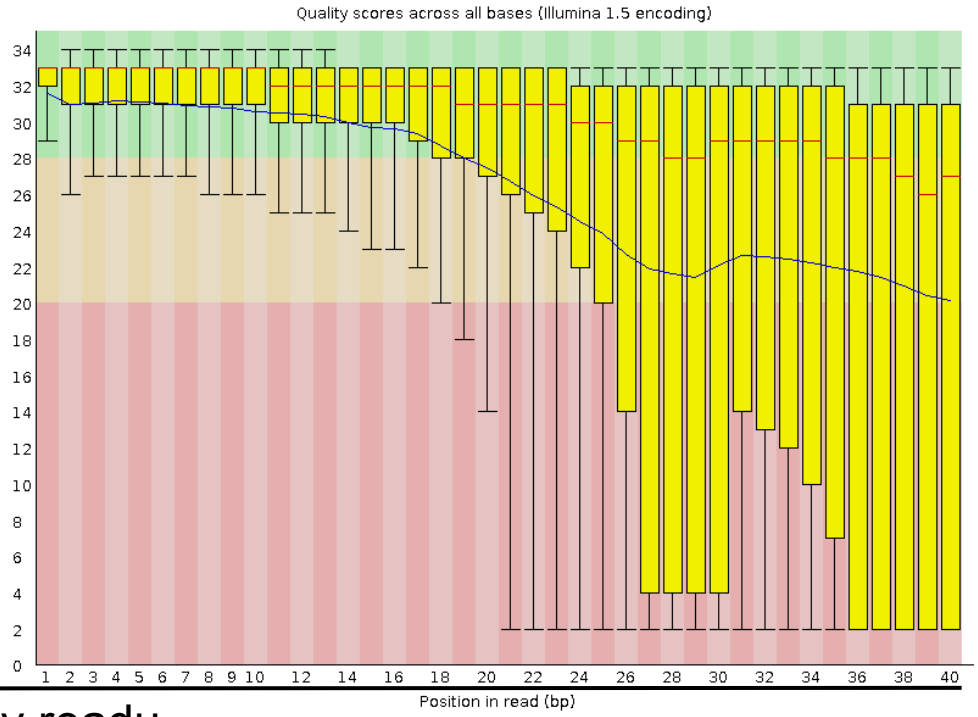
Distribuce Phred skóre jedn. bází

FastQC

DOBŘE



ŠPATNĚ



pozice báze v readu

- varování: Q < 10, median < 25
- chyba: Q < 5, median < 20

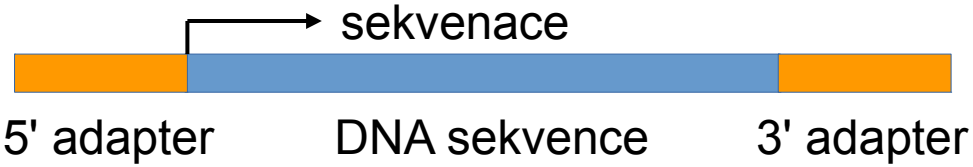
Co s tím?

ŘEŠENÍ: ořez koncových bází

Phred skóre = -log10 * P (pravděpodobnost nesprávného určení báze)

Kontaminace

Nadměrně zastoupené sekvence



1) Adaptery

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATC	838817	20.970425000000002	Illumina Paired End PCR Primer 2 (100% over 40bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA	317338	7.933450000000001	Illumina Paired End PCR Primer 2 (96% over 33bp)
GCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT	74507	1.862675	Illumina Paired End PCR Primer 2 (100% over 39bp)

Co s tím?

ŘEŠENÍ: programy, které umí adaptery z dat odstranit

2) Technické artefakty

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAGGCGTTCGT	5773	0.2785442744507882	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAGGCGTTCGT	3715	0.17924683519568302	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTAAAGTGTTG	2836	0.1368355382543626	No Hit

Co s tím?

ŘEŠENÍ: nedojde k alignmentu na referenční genom

FastQC

Základní schéma analýzy NGS dat

2. Alignment (mapování) **x** **assembly**

- mapování sekvencí na referenční genom

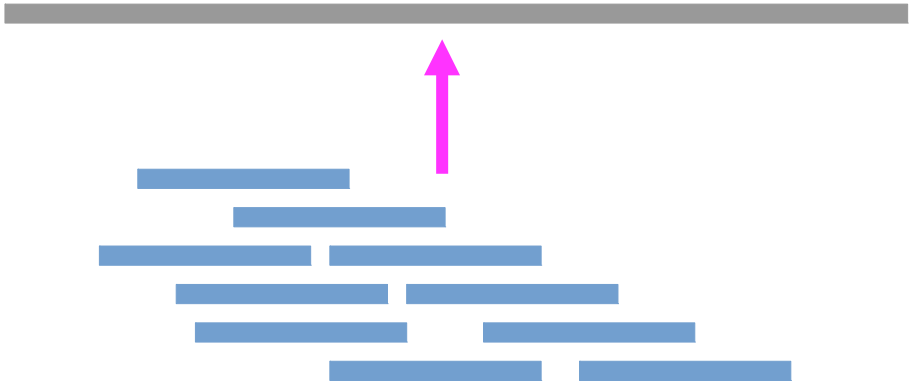
- de novo skládání genomu

Základní schéma analýzy NGS dat

2. Alignment (mapování) **x** **assembly**

- mapování sekvencí na referenční genom

- de novo skládání genomu



Základní schéma analýzy NGS dat

2. Alignment (mapování) x assembly

- mapování sekvencí na referenční genom

- výhodný, když je dostupný referenční genom

- rychlý, provede i normální stolní počítač

- problém u repetitivních sekvencí

- alignery (BWA, STAR, ..)

- de novo skládání genomu

- pokud není dostupný referenční genom nebo pokud chceme individuální referenční genom

- velmi náročný: čas, výkon stroje

- velké problémy u repetitivních sekvencí

Alignment

reference: **AAATGCCGTACCGTCGTCGCCTACGAGAGGATTACTCGGTTTACCGTATCG**

read 1 **TACCGTCGCTA**

Alignment

reference: **AAATGCCGTACCGTCGTCGCCTACGAGAGGATTACTCGGTTTACCGTATCG**

read 1 **TACCGTCGTCG**

Alignment

reference: **AAATGCCGTACCGTCGTCGCCTACGAGAGGATTACTCGGTTTACCGTATCG**

read 1 **TACCGTCGTCG**

read 2 **CGCCTACCAGA**

Alignment

reference: **AAATGCCGTACCGTCGTCGCCTACGAGAGGATTACTCGGTTTACCGTATCG**

read 1	**TACCGTCGTCG**
read 2	**CGCCTACCAGA**
read 3	**ACCAGAGGATT**
read 4	**CAGAGGATTAC**
read 5	**GTCGCCTACCA**
read 6	**CTACCAGAGG**

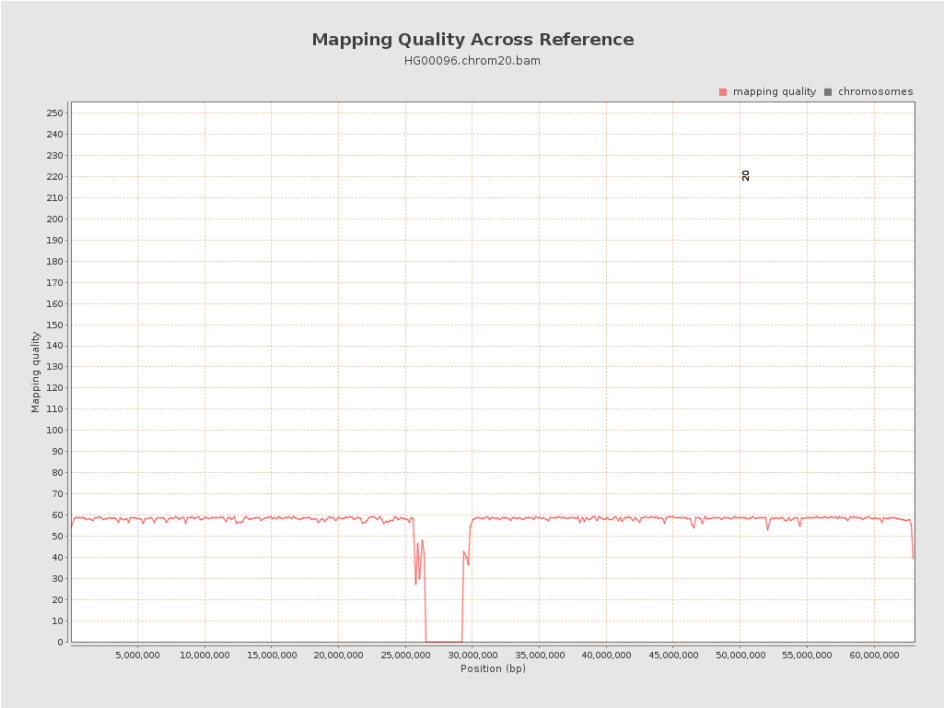
Kontrola kvality alignmentu

mapping quality = -10 * $\log_{10}$Pr

(Pr = pravděpodobnost, že je read nesprávně namapován)

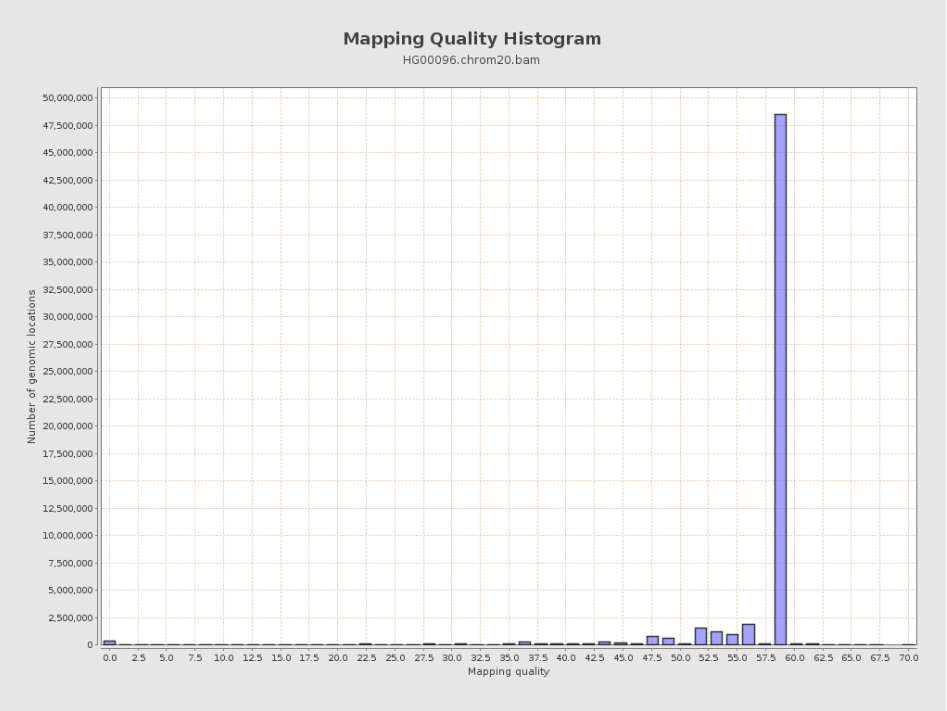
QualiMap

Mapping kvalita



reference

Počet bp



mapping kvalita

Co s tím?

ŘEŠENÍ: ready s nízkou map. kvalitou nejsou ve většině analýz brány v potaz

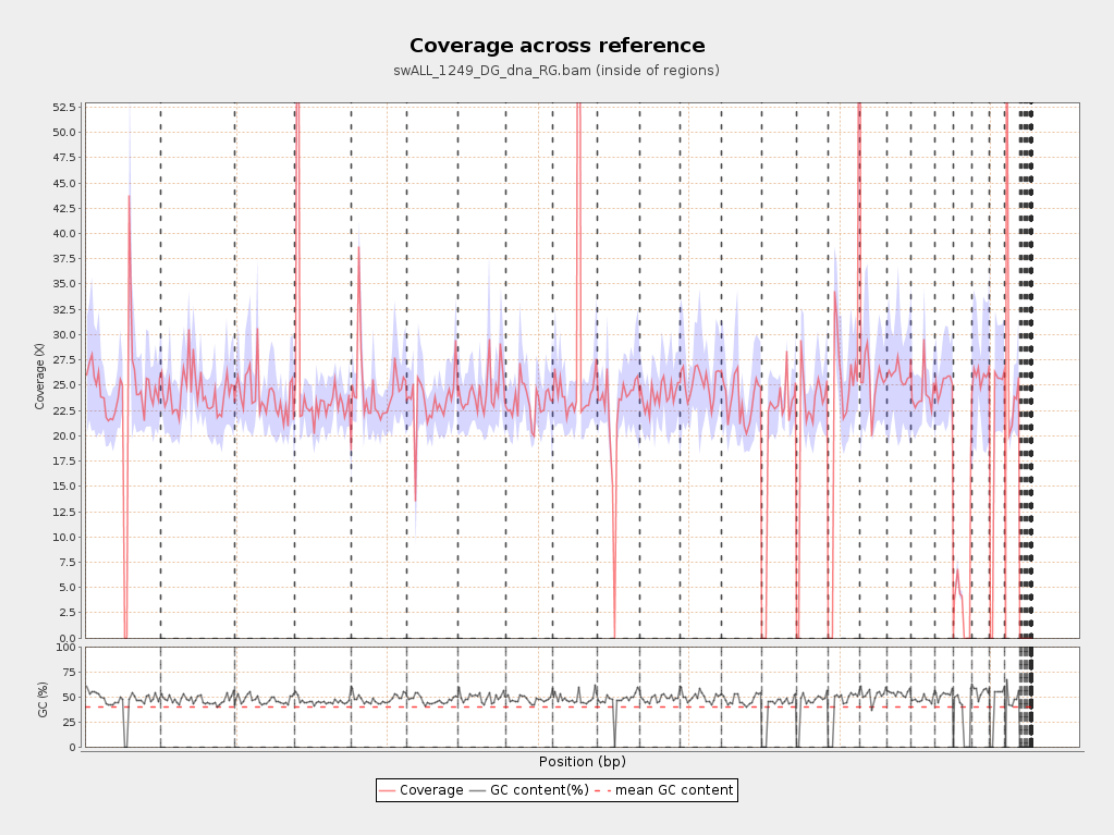
Kontrola kvality experimentu

Pokrytí

Je třeba vždy **zvolit vhodnou referenci**
(např. **exom**: omezení pouze na vybírané regiony při přípravě sekvenační knihovny)

QualiMap

pokrytí



reference

Co s tím?

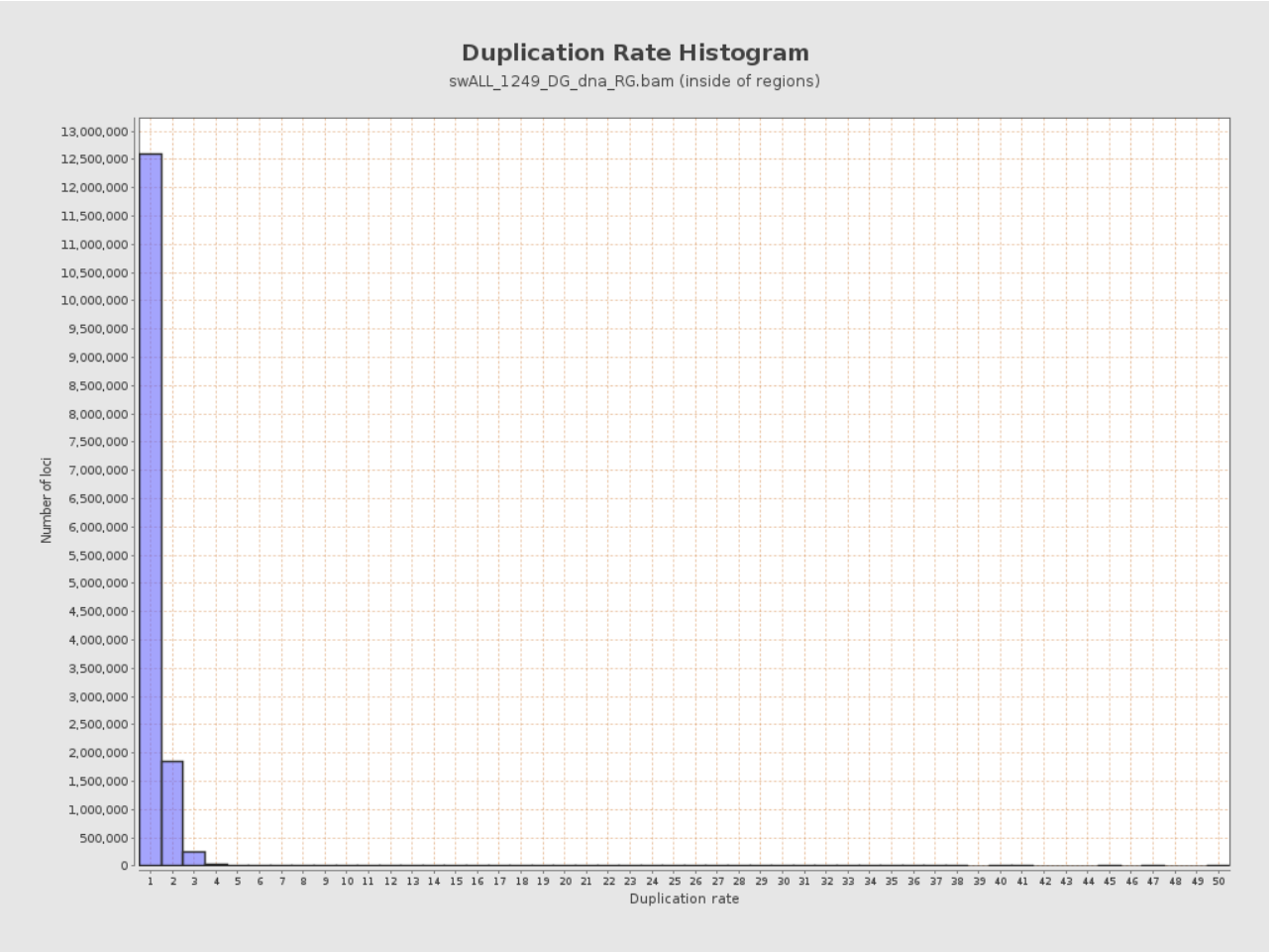
ŘEŠENÍ: např. exom: k označení SNVs je potřeba určité pokrytí dané báze

Kontrola kvality experimentu

PCR duplikáty

QualiMap

počet readů



kolikrát je read duplikován

Co s tím?

ŘEŠENÍ: ve většině experimentů nejsou brány v analýze v potaz

Základní schéma analýzy NGS dat

3. Specifické kroky analýzy dle NGS experimentu

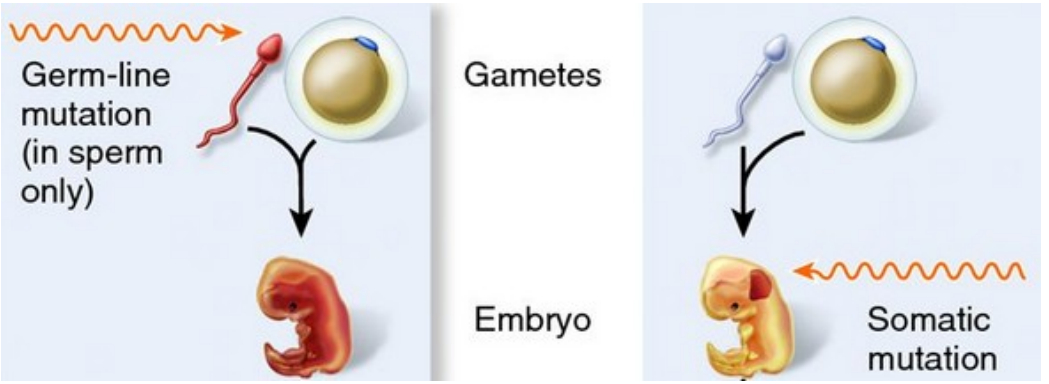
Variant calling (hledání variant) – amplikon, exon, RNA-seq

Typy variant:

CGCCTACGAGA

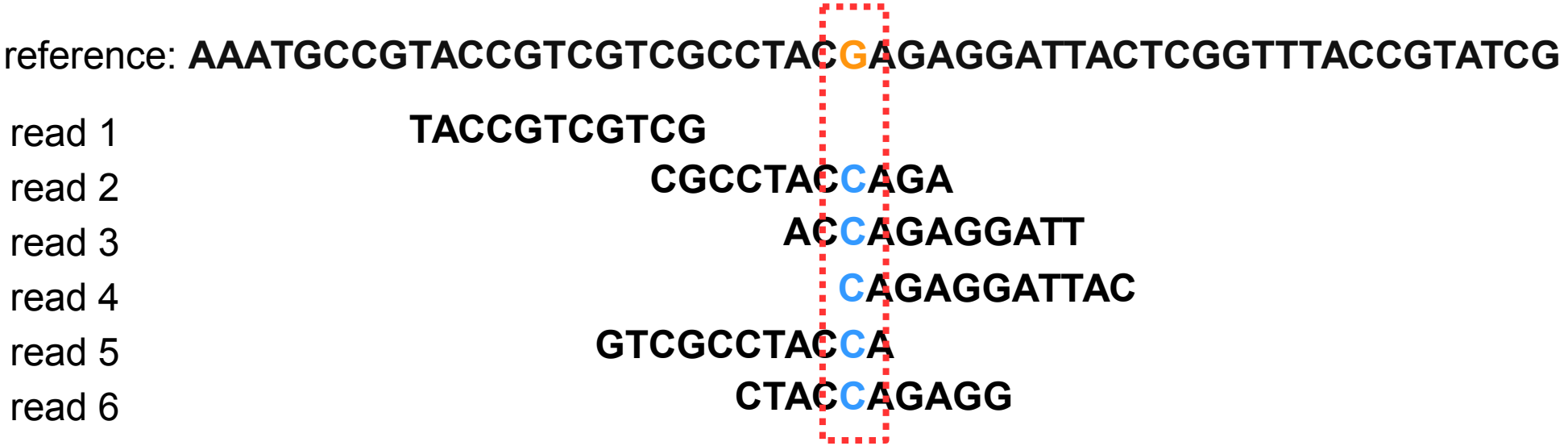
- **SNVs** (single nucleotide variant) = záměna 1 báze **CGCCTACCAGA**
- **indely** = krátké inzerce nebo delece (1 – 100 bází) **CGC----AGA** x **CGCCTATATCCAGA**
- **CNVs** (copy number variants) = duplikace nebo delece velkých oblastí
- **strukturální** = translokace nebo inverze velkých oblastí beze změny počtu těchto oblastí v genomu

- **somatické** (nedědičné)
- **germinální** (dědičné)



3. Specifické kroky analýzy dle NGS experimentu

Variant calling (hledání variant) – amplikon, exon, RNA-seq

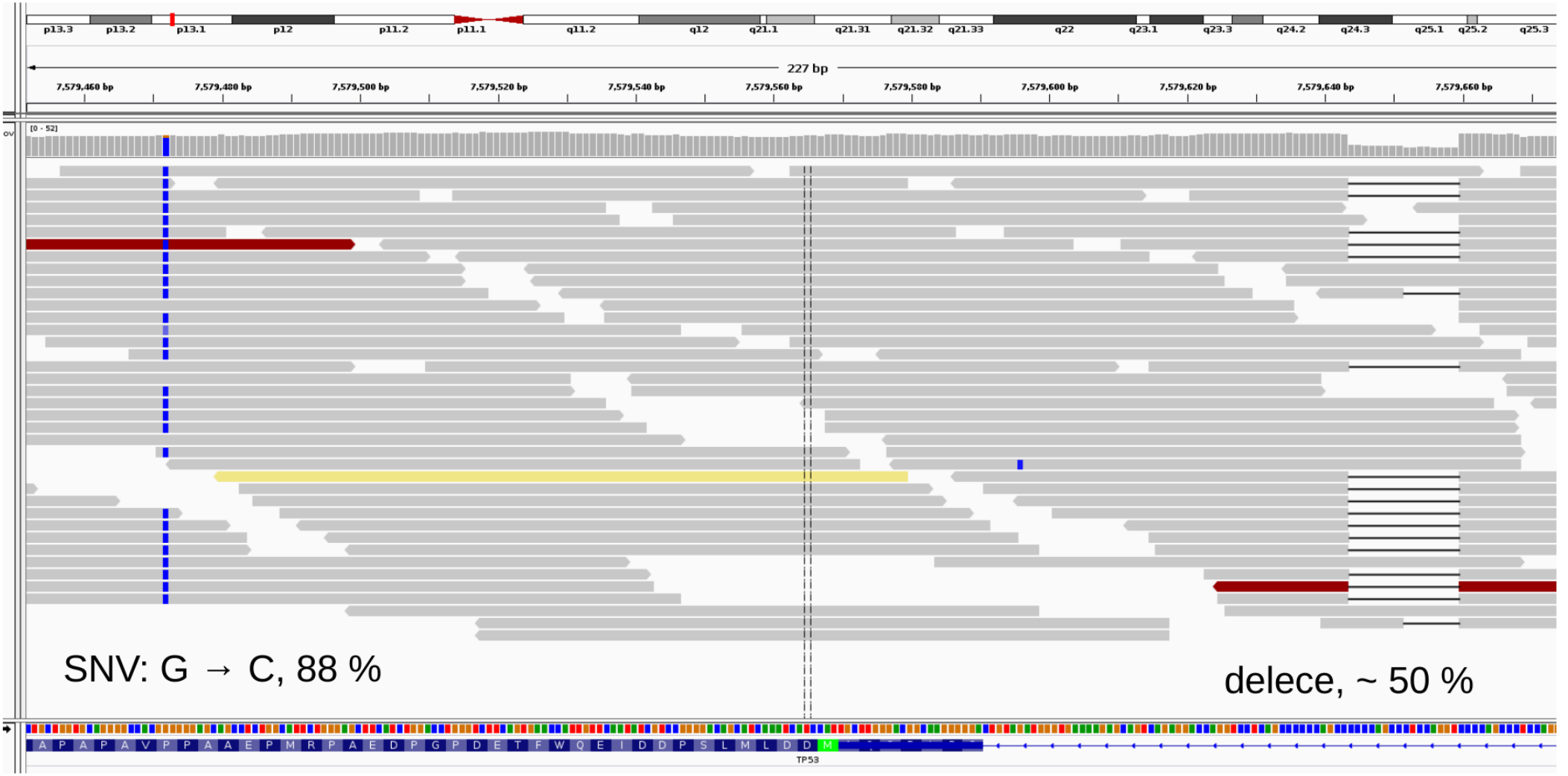


SNVs: G → C, allele frequency 100 %

3. Specifické kroky analýzy dle NGS experimentu

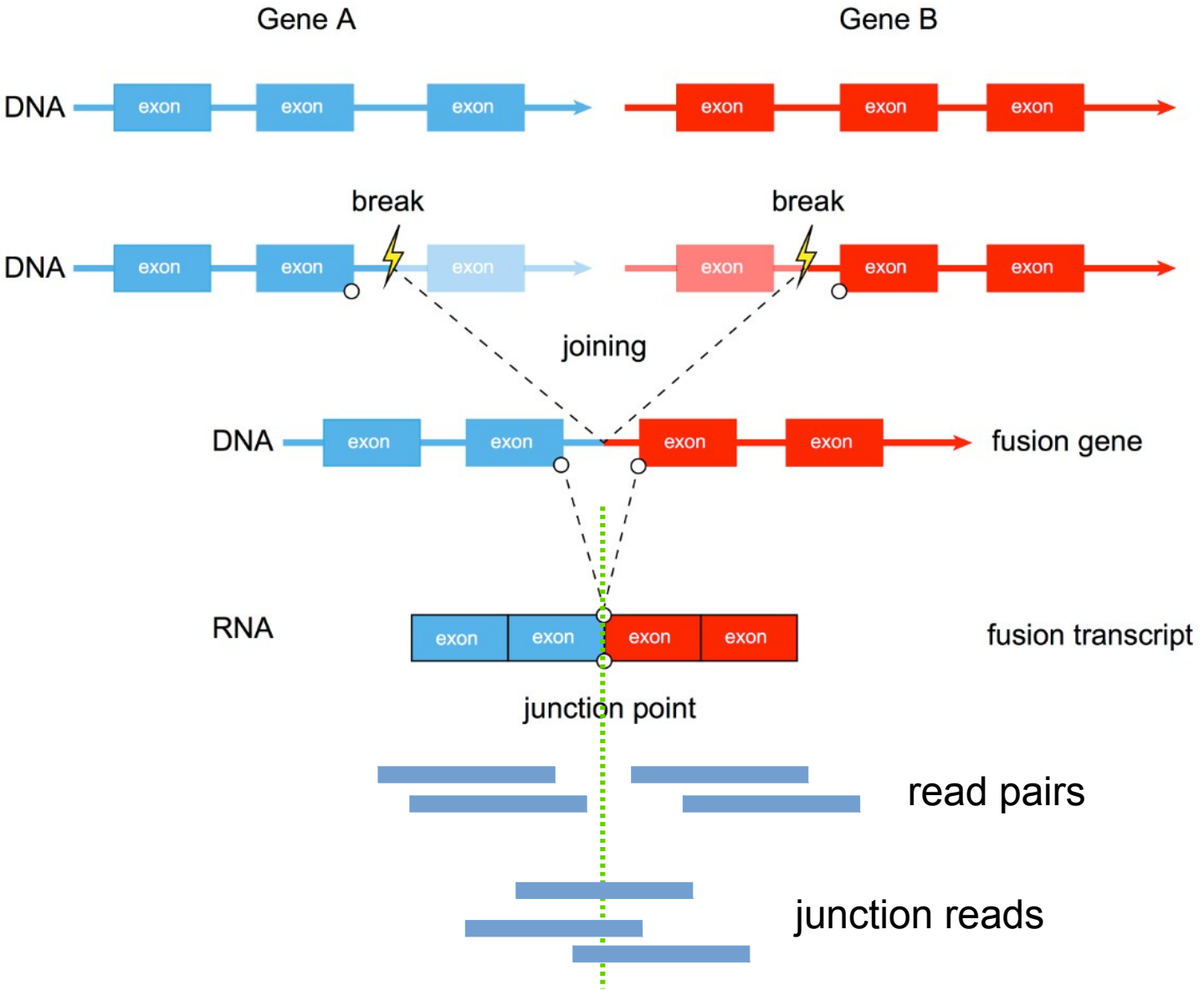
Variant calling (hledání variant) – amplikon, exon, RNA-seq

IGV



3. Specifické kroky analýzy dle NGS experimentu

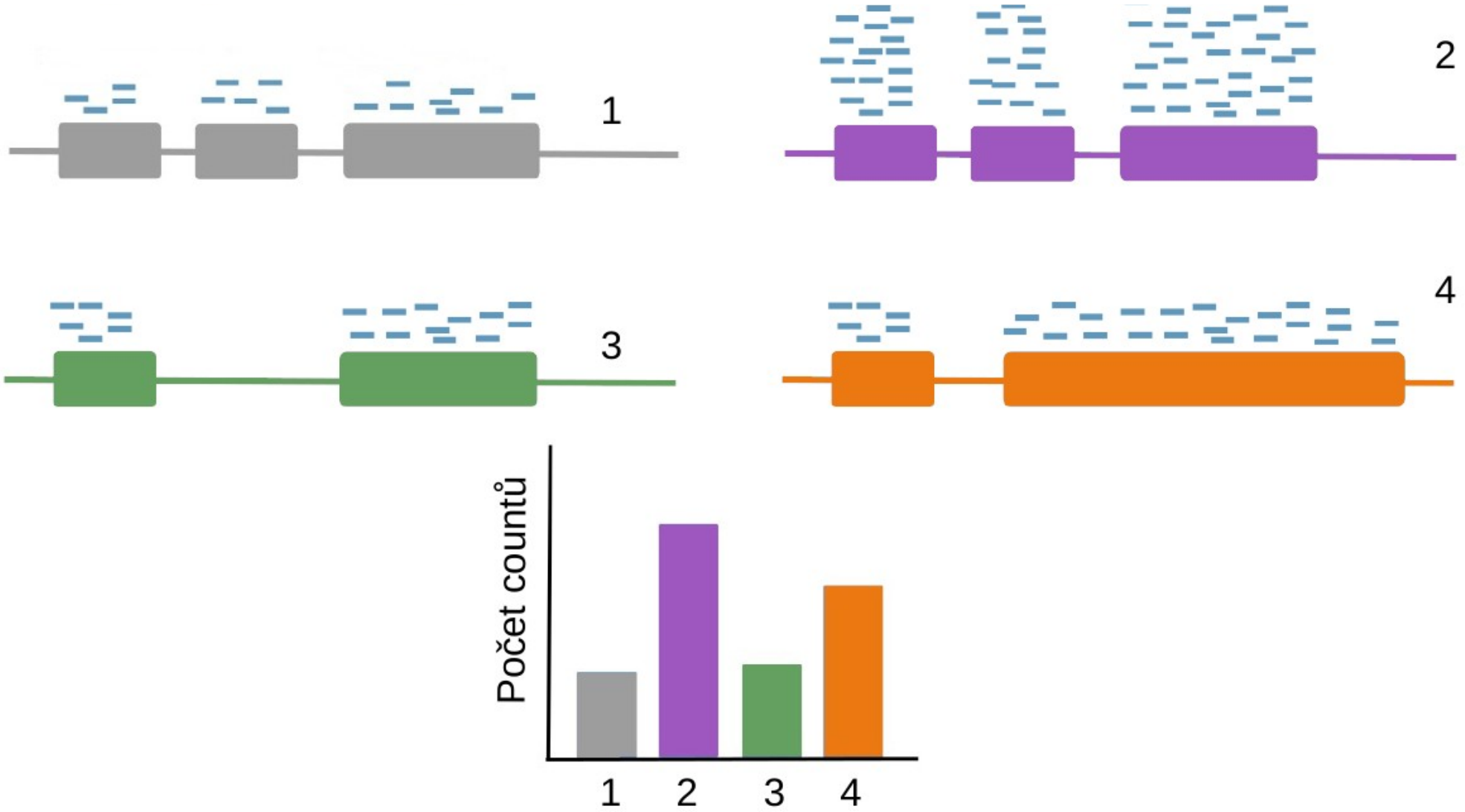
Hledání fúzních genů – RNA-seq



3. Specifické kroky analýzy dle NGS experimentu

Exprese RNA / genů – RNA-seq

- počítání countů (readů)



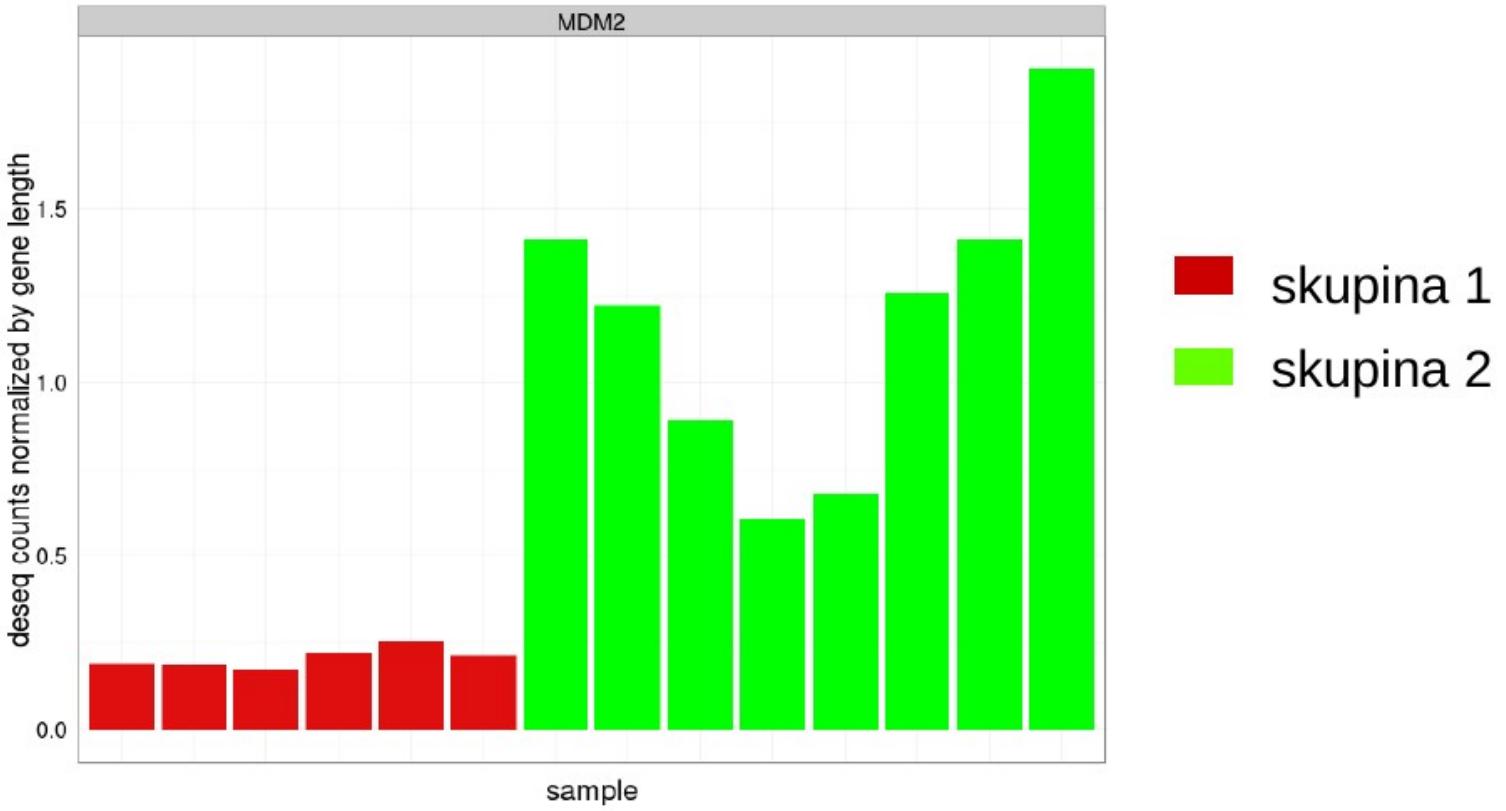
3. Specifické kroky analýzy dle NGS experimentu

Exprese RNA / genů – RNA-seq

- počítání countů (readů)

```
> head(deseq_logfc)
                  baseMean log2FoldChange     lfcSE      stat       pvalue         padj
1 ENSG00000135679 6083.6171      -2.485823 0.2594709 -9.580351 9.671586e-22 2.343812e-17
2 ENSG00000180767   73.2637       5.864500 0.6615192  8.865200 7.636700e-19 9.253389e-15
```

← MDM2



3. Specifické kroky analýzy dle NGS experimentu

Příklad: hledání fúzních genů & exprese RNA - RNA-seq

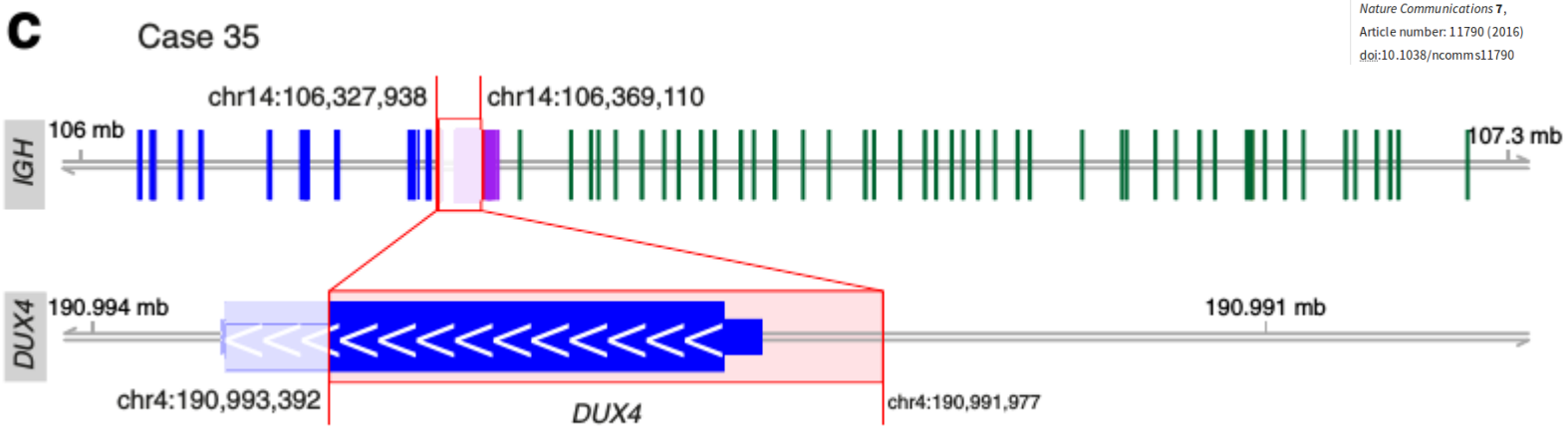
Article | OPEN

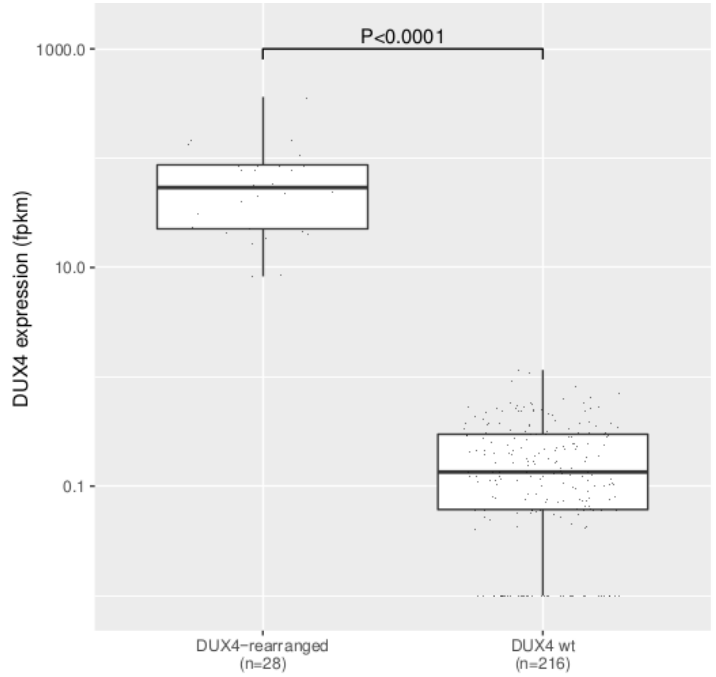
Identification of *ETV6-RUNX1*-like and *DUX4*-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia

Henrik Lilljebjörn, Rasmus Henningsson, Axel Hyrenius-Wittsten, Linda Olsson, Christina Orsmark-Pietras, Sofia von Palffy, Maria Askmyr, Marianne Rissler, Martin Schrappe, Gunnar Cario, Anders Castor, Cornelis J. H. Pronk, Mikael Behrendtz, Felix Mitelman, Bertil Johansson, Kajsa Paulsson, Anna K. Andersson, Magnus Fontes & Thoas Fioretos

Nature Communications **7**,
Article number: 11790 (2016)
doi:10.1038/ncomms11790

Received: 26 February 2016
Revised: 11 April 2016
Accepted: 28 April 2016





preB-ALL s přestavbami v genu DUX4

- 4 % všech preB-ALL, 16 % z B-others

- „objev roku“ 2016

- skupina preB-ALL se specifickým expresním profilem a častými delecemi genu ERG

Děkuji za pozornost!