DETEKCJA POLIMORFIZMÓW - OMÓWIENIE STANDARDOWYCH KROKÓW ANALIZY DANYCH NGS

Analiza danych NGS
Wykład 3

GŁÓWNE ZASTOSOWANIA NGS

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

ANALIZA ZMIENNOŚCI GENETYCZNEJ ZA POMOCĄ WGS

Detection of genomic variation among individuals of a population is among the most frequent applications of next-generation sequencing (NGS).

Locating genomic sequence variations that correlate with disease predisposition or drug response, and esablishing a genotypic basis of various phonotypes become common focuses of many NGS studies in biomedical and life sciences reserach.

Besides variations carried through the germline for generations, NGS has also been applied to identify *de novo* germline and somatic mutations, which occur more frequently than previously expected and underlie numerous human diseases including various types of cancer.

Xinkun Wang. Next Generation Sequencing Data Analysis. 2016, CRC PRESS

ANALIZA ZMIENNOŚCI GENETYCZNEJ ZA POMOCĄ WGS

Detecting the various forms of genetic variations/mutations from NGS (…) is not an easy task. The primary challange is to differentiate true sequence variations/mutations from false positives caused by sequencing errors and artifacts generated in (…) sequnce alignment.

It is, therefore, important to generate high-quality sequence data data before performing data analysis. Equally important, sensitive and yet specific variant/mutant calling algorithms are required to achieve high accuracy in genomic ariation and mutation discovery.

Xinkun Wang. Next Generation Sequencing Data Analysis. 2016, CRC PRESS

ANALIZA ZMIENNOŚCI GENETYCZNEJ ZA POMOCĄ WGS POPULARNE PLATFORMY

Illumina

- ~100-250 bp

Pacific Biosciences (PacBio)

- 10–25 kb

Oxford Nanopore Technologies

- 500 bp - 2.5 Mb

… i inne

PIPELINE

Pipeline = łańcuch przetwarzania danych

Kontrola jakości, czyszczenie danych oraz przyrównanie do genomu referencyjnego bardzo często są podstawą w innych typach analiz, nie tylko związanych z analizą danych DNA-seq.

Uproszczony schemat

Surowe Dane
↓
Kontrola jakości
↓
Przyrównanie do genomu referencyjnego
↓
Detekcja polimorfizmów
↓
Sens biologiczny

SUROWE DANE

Read 1

```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDDHIIIJJDDIIIGDDJGDDGD!
(…)
```

Single-end

Read 1

AGAAATG…

Surowe Dane → Kontrola jakości → Przyrównanie do genomu referencyjnego → Detekcja polimorfizmów → Sens biologiczny

SUROWE DANE

Read 1

```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDDHIIIJJDDIIIGDDJGDDGD!
(…)
```

Read 2

```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 2:N:0:ACAGTG
TTAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGCT
+
DD@FF@@FGHGHH01DDGHHHDDDDHIIIJJJDIIIGDDJGDDGDD
(…)
```

Paired-end

Read 1 AGAAATG… →

← … GGCTGAA **Read 2**

Surowe Dane → Kontrola jakości → Przyrównanie do genomu referencyjnego → Detekcja polimorfizmów → Sens biologiczny

SRR988073_R1.fastq

```
@SRR988073.1.1 HWI-ST833:124:6:1101:1412:2027 length=101
NGCCACATCACATTGTTGTTTGCTTAGCCACAATTCATGAAAATTGTTCTTTAATTACTTCCTGTGGCAGNGTGTGACGCCTGCGAAATGAAGTCGTT
+SRR988073.1.1 HWI-ST833:124:6:1101:1412:2027 length=101
#1=DDFFFHHHHHIIFHIFHIJJJIIIHIIJIGGHGIJJGGHIIJJIIEIIIHCHDHIJIDFEEDC>@AC#-5;;;;AC<<AB<?33'9A:>@((580884
@SRR988073.2.1 HWI-ST833:124:6:1101:1335:2052 length=101
AAAAATTACATTTAAATATGGTAGAAGGTGTCTTACATAATTAGTTGGACAAAAGTTAAGCGCACATGTTTGACAAACTAAAGAAGTTTTACCAAATGTCC
+SRR988073.2.1 HWI-ST833:124:6:1101:1335:2052 length=101
CCCFFFFFHHHHHJJGHHJJJHJJJIJICGFHEGEGIDIIJIIIFEHDEHJIIDGBDHHIBEGBEGEG==@DECEHCACD;?>;>C.;>6;;>@;2;:>(5
@SRR98807
GAACCAGCG
+SRR98807
<@@FFFDFH
@SRR98807
GATAGAGGG
+SRR98807
??@DDD;D:
@SRR98807
GTTTAGCAG
+SRR98807
CCCFFFFFH
@SRR98807
AGCACAGGT
+SRR98807
CCCFFFFFF
```

SRR988073_R2.fastq

```
@SRR988073.1.2 HWI-ST833:124:6:1101:1412:2027 length=101
CGNNTATGGTGCATTTTTGATTGGCAAACTTGGCGACAGATGCGATTGATTAATTGCCTCTTGGCTCTCGTCGCCGCCGCCGAATAAGCACGTTTTATCTG
+SRR988073.1.2 HWI-ST833:124:6:1101:1412:2027 length=101
B@##4=BDH:CFHJJJJJJJJJJJJJJJJJIJJIJIJJGGIJJJGIIJIJFHIIJIIJIJJHGHHHFFFFDDDDDDDDDBBDDDDDDD
@SRR988073.2.2 HWI-ST833:124:6:1101:1335:2052 length=101
ACATAGCAAATGGTGCTTTACGATTTTTTTATCAAAGTGTAAAATATTGTAACACGGTTCGAAGAAAATTGGGTCAAAAGCTAATTG
+SRR988073.2.2 HWI-ST833:124:6:1101:1335:2052 length=101
CCCFFFFFHHHHGCGHJJJJJJHJJJJJJJJIJJJJJJDHGIJJJJJJJJIIIIJJJJHHHFFFFDEEEEDDDDBDDDDDDDDDDDEDDDDCDDDDDDDDD
@SRR988073.3.2 HWI-ST833:124:6:1101:1298:2057 length=101
AGCTGCCTCTCTTGTACAAAAACTTAATGAGAAACTAATTAACTGTGTAAATATCTAATATTAGATATGAATAACTTGAATAGTTTCGAATTTTAATCGGA
+SRR988073.3.2 HWI-ST833:124:6:1101:1298:2057 length=101
@C@FFFFFGDHHHJFHEGIEIIIDHIII@FCHGIJEHI@IGIJJJFHDFDGIIIHIHJJJJIGIJJJIIIGG>HHGGH@CGHIEEEHHHFFEFF@CC;?@D
@SRR988073.4.2 HWI-ST833:124:6:1101:1475:2065 length=101
CGGCCCCACCCACTCGACCCCAGGCCACCGCCATCCCCGCAGGACATCCGGCGGCTCATCGCAGTCCTTGTGGCTAATATTGACATTCCTAATCTGCGCTC
+SRR988073.4.2 HWI-ST833:124:6:1101:1475:2065 length=101
B@?DDFFBHFFFHIGBBHIGGGIGIJ8FFGBHGHIIJJGGHDDDCDCAA>9>@8<-7:>C9??BBCCAC@(:<A?CCCCAAA4>4@C:B@(:34::(&5>B
@SRR988073.5.2 HWI-ST833:124:6:1101:1456:2094 length=101
ATCGACACGTCGACGTTCTTCATCAAGTTCGTCTGGGTGCCGAGATCGCATGTGAACACCGAGTCCTCTCTTGCTGTTAAAGGACCGGCCACATACTGTGC
+SRR988073.5.2 HWI-ST833:124:6:1101:1456:2094 length=101
BC?DFFFFGHHHHIJIJJJJJIJJJJJIJJIFHGIIJFGGIJJIGIHHHHBFCEEEEEDDDDD8ACDDDDDDDDDDACDEDADDBBDDDDDDDDDDCDCCC
@SRR988073.6.2 HWI-ST833:124:6:1101:1424:2119 length=101
CGAATCCGAAGAGTTTAGGCTCCTTCTCCCTACTGCGTTTTAGTCCCGATTTCCATGTGCTCTTCTGCGCCTCCTGCGATGCGGGCGCCAGTCTTTGTCAA
+SRR988073.6.2 HWI-ST833:124:6:1101:1424:2119 length=101
CCCFFFFFHGHHHHIJJJJJIJJJJIJJJJIJJJJJGIJJIJJJJIHHIFHIJJJJJJJHHHHHHFFFDDDDDDDD?DDDDDDDDDDDDACCEDDCADE
```

```
 14G Dec  6 14:35 wgs_S4505Nr1.1.fastq.gz
 15G Dec  6 14:56 wgs_S4505Nr1.2.fastq.gz
 21G Dec  6 15:27 wgs_S4505Nr10.1.fastq.gz
 21G Dec  6 15:59 wgs_S4505Nr10.2.fastq.gz
9.9G Dec  6 16:14 wgs_S4505Nr11.1.fastq.gz
 11G Dec  6 16:30 wgs_S4505Nr11.2.fastq.gz
 24G Dec  6 17:06 wgs_S4505Nr12.1.fastq.gz
 26G Dec  6 17:45 wgs_S4505Nr12.2.fastq.gz
 18G Dec  6 18:12 wgs_S4505Nr13.1.fastq.gz
 19G Dec  6 18:40 wgs_S4505Nr13.2.fastq.gz
 22G Dec  6 19:12 wgs_S4505Nr14.1.fastq.gz
 23G Dec  6 19:46 wgs_S4505Nr14.2.fastq.gz
 19G Dec  6 20:14 wgs_S4505Nr15.1.fastq.gz
 20G Dec  6 20:44 wgs_S4505Nr15.2.fastq.gz
 30G Dec  6 21:28 wgs_S4505Nr16.1.fastq.gz
 31G Dec  6 22:15 wgs_S4505Nr16.2.fastq.gz
 13G Dec  6 22:34 wgs_S4505Nr17.1.fastq.gz
 13G Dec  6 22:53 wgs_S4505Nr17.2.fastq.gz
 23G Dec  6 23:27 wgs_S4505Nr18.1.fastq.gz
 24G Dec  7 00:02 wgs_S4505Nr18.2.fastq.gz
 28G Dec  7 00:44 wgs_S4505Nr19.1.fastq.gz
 29G Dec  7 01:28 wgs_S4505Nr19.2.fastq.gz
 17G Dec  7 01:53 wgs_S4505Nr2.1.fastq.gz
 18G Dec  7 02:20 wgs_S4505Nr2.2.fastq.gz
 18G Dec  7 02:46 wgs_S4505Nr20.1.fastq.gz
 19G Dec  7 03:14 wgs_S4505Nr20.2.fastq.gz
 13G Dec  7 03:33 wgs_S4505Nr3.1.fastq.gz
 14G Dec  7 03:54 wgs_S4505Nr3.2.fastq.gz
 26G Dec  7 04:33 wgs_S4505Nr4.1.fastq.gz
 28G Dec  7 05:15 wgs_S4505Nr4.2.fastq.gz
```

```
417M Jan 16  2014 lane2_NoIndex_L002_R1_039.fastq.gz
395M Jan 16  2014 lane2_NoIndex_L002_R1_044.fastq.gz
393M Jan 16  2014 lane2_NoIndex_L002_R1_025.fastq.gz
404M Jan 16  2014 lane2_NoIndex_L002_R1_008.fastq.gz
390M Jan 16  2014 lane2_NoIndex_L002_R2_014.fastq.gz
391M Jan 16  2014 lane2_NoIndex_L002_R2_046.fastq.gz
405M Jan 16  2014 lane2_NoIndex_L002_R1_028.fastq.gz
395M Jan 16  2014 lane2_NoIndex_L002_R1_024.fastq.gz
402M Jan 16  2014 lane2_NoIndex_L002_R1_027.fastq.gz
406M Jan 16  2014 lane2_NoIndex_L002_R1_058.fastq.gz
407M Jan 16  2014 lane2_NoIndex_L002_R1_029.fastq.gz
392M Jan 16  2014 lane2_NoIndex_L002_R1_015.fastq.gz
390M Jan 16  2014 lane2_NoIndex_L002_R2_004.fastq.gz
392M Jan 16  2014 lane2_NoIndex_L002_R1_052.fastq.gz
389M Jan 16  2014 lane2_NoIndex_L002_R2_023.fastq.gz
404M Jan 16  2014 lane2_NoIndex_L002_R1_040.fastq.gz
391M Jan 16  2014 lane2_NoIndex_L002_R2_006.fastq.gz
394M Jan 16  2014 lane2_NoIndex_L002_R1_054.fastq.gz
391M Jan 16  2014 lane2_NoIndex_L002_R2_003.fastq.gz
399M Jan 16  2014 lane2_NoIndex_L002_R1_006.fastq.gz
393M Jan 16  2014 lane2_NoIndex_L002_R1_023.fastq.gz
384M Jan 16  2014 lane2_NoIndex_L002_R2_045.fastq.gz
391M Jan 16  2014 lane2_NoIndex_L002_R2_013.fastq.gz
410M Jan 16  2014 lane2_NoIndex_L002_R1_030.fastq.gz
387M Jan 16  2014 lane2_NoIndex_L002_R2_021.fastq.gz
397M Jan 16  2014 lane2_NoIndex_L002_R1_003.fastq.gz
```

SUROWE DANE

PRZECHOWYWANIE DANYCH

1 genom

```
total 44G
-rw-rw---- 1            oclaw 383M Jan 16  2014 lane1_NoIndex_L001_R1_029.fastq.gz
-rw-rw---- 1 szyda upwroclaw 376M Jan 16  2014 lane1_NoIndex_L001_R2_055.fastq.gz
-rw-rw---- 1 szyda upwroclaw 385M Jan 16  2014 lane1_NoIndex_L001_R2_038.fastq.gz
-rw-rw---- 1 szyda upwroclaw 375M Jan 16  2014 lane1_NoIndex_L001_R1_046.fastq.gz
-rw-rw---- 1 szyda upwroclaw 387M Jan 16  2014 lane1_NoIndex_L001_R1_059.fastq.gz
-rw-rw---- 1 szyda upwroclaw 383M Jan 16  2014 lane1_NoIndex_L001_R2_048.fastq.gz
-rw-rw---- 1 szyda upwroclaw 385M Jan 16  2014 lane1_NoIndex_L001_R1_010.fastq.gz
-rw-rw---- 1 szyda upwroclaw 380M Jan 16  2014 lane1_NoIndex_L001_R2_050.fastq.gz
-rw-rw---- 1 szyda upwroclaw 378M Jan 16  2014 lane1_NoIndex_L001_R1_057.fastq.gz
```

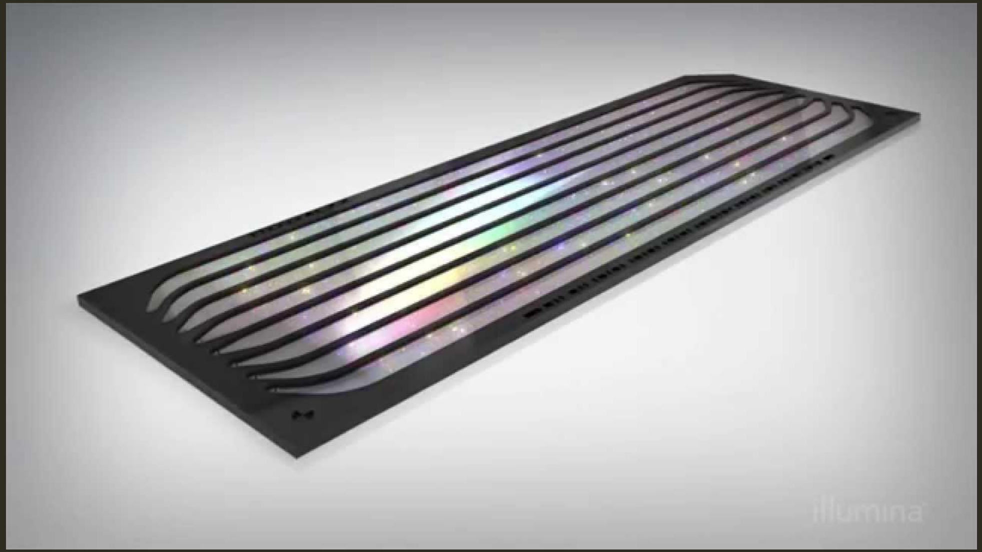
+ dane dodatkowe

http://en.wikipedia.org/wiki/FASTQ_format

```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDDHIIIJJDDIIIGDDJGDDGD!
(…)
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence



www.illumina.com

$$Q = -10 \log_{10} P \Longrightarrow P = 10^{\frac{-Q}{10}}$$

KODOWANIE JAKOŚCI

@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDDHIIIJJDDIIIGDDJGDDGD!
(…)

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN...........................................
 PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0.2......................26...31........41
 0..................20........30........40........50
 0..................20........30........40........50..........................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
N - Nanopore      Phred+33,  Duplex reads typically (0, 50)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

http://en.wikipedia.org/wiki/FASTQ_format

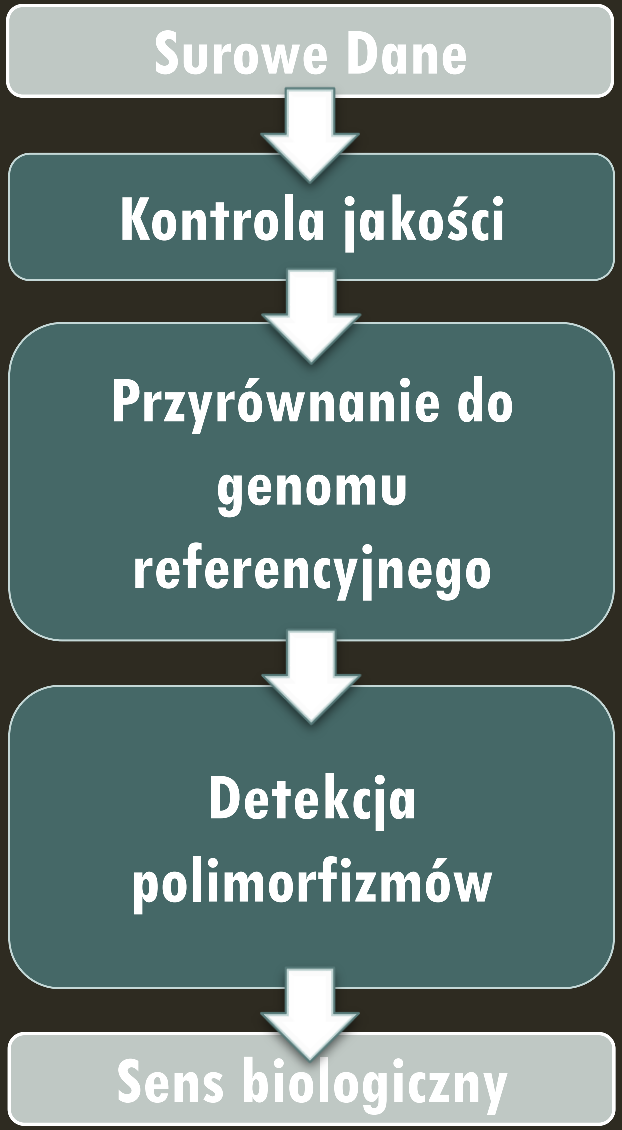
KONTROLA JAKOŚCI

Read 1

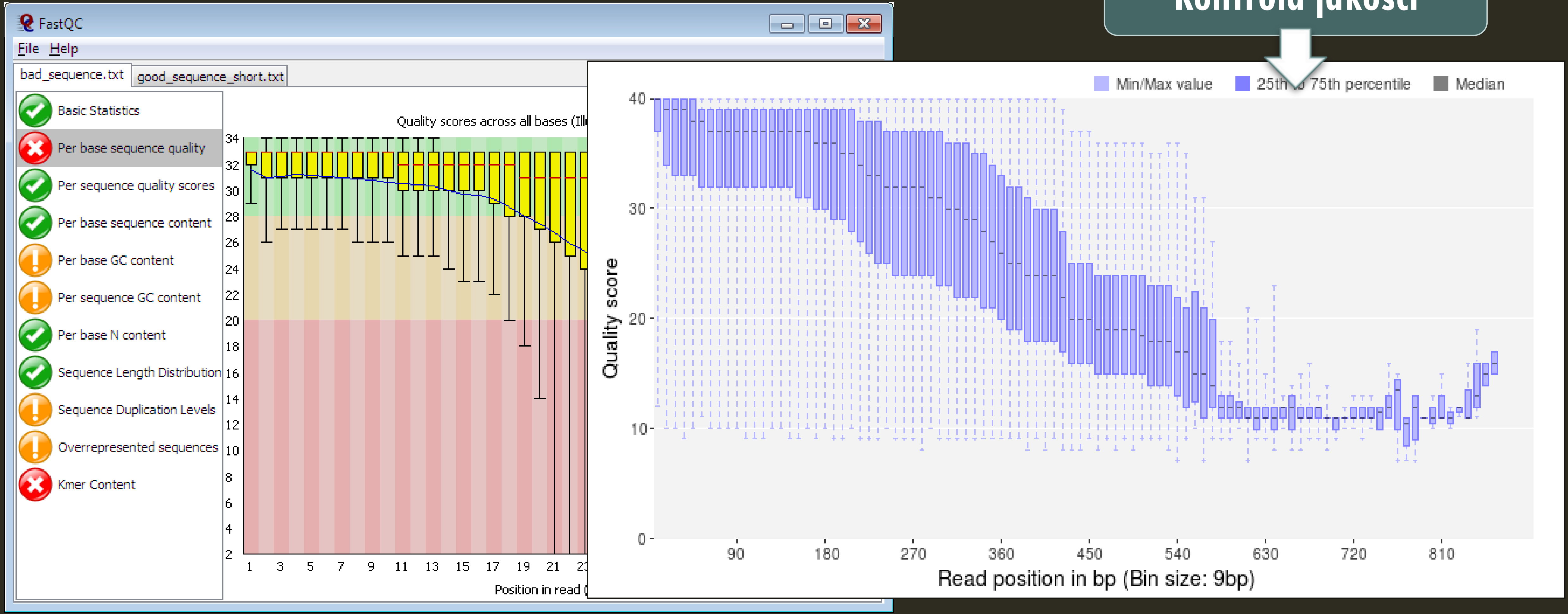
```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 1:N:0:ACAGTG
AGAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGC
+
@@@FFDFFGHGHHFDDDGHHHDDDDHIIIJJDDIIIGDDJGDDGD!
(…)
```

Read 2

```
@HWI-1KL157:109:C448WACXX:7:1311:12007:37445 2:N:0:ACAGTG
TTAAATGCCAGGCTAGATGAGTTACAATCTAGTATCAAGATAGGCT
+
DD@FF@@FGHGHH01DDGHHHDDDDHIIIJJJDIIIGDDJGDDGDD
(…)
```

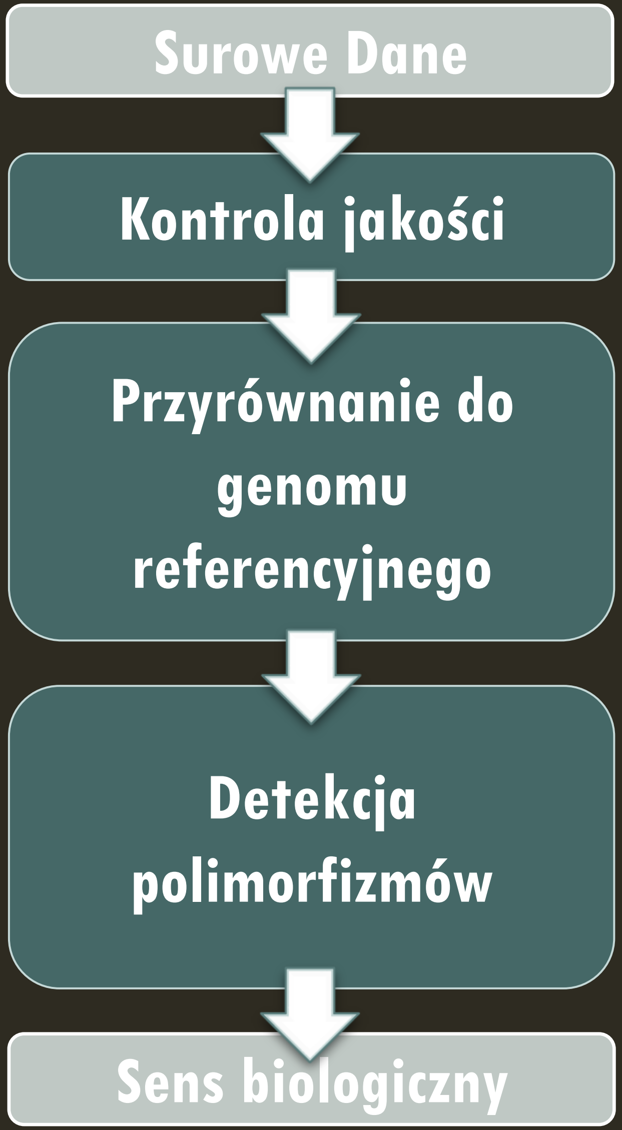


KONTROLA JAKOŚCI



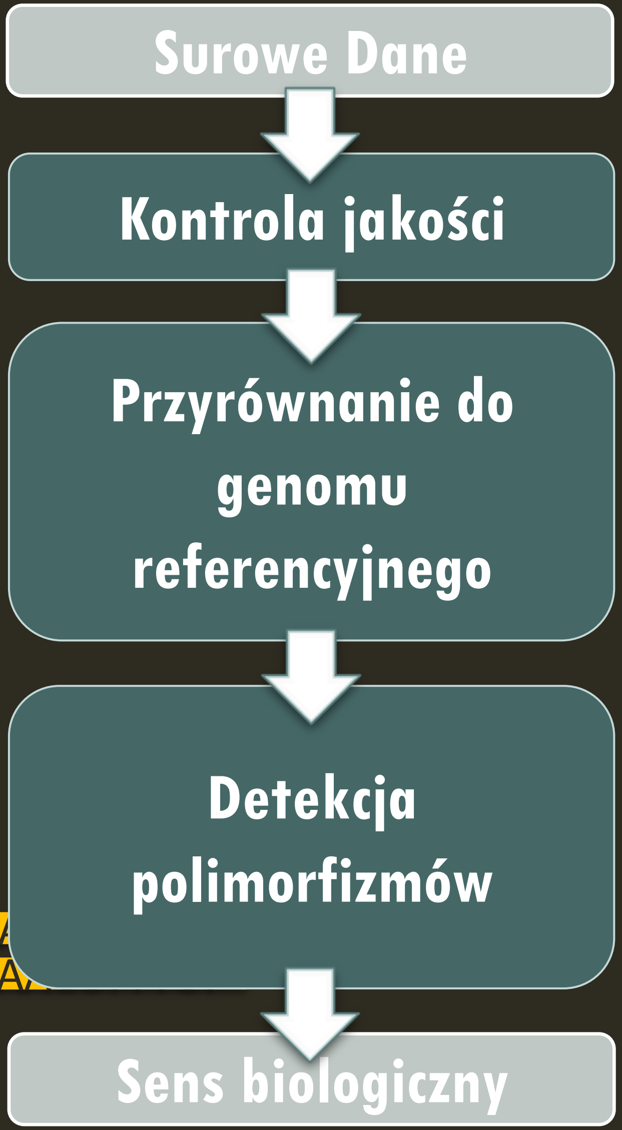
PRZYRÓWNANIE DO GENOMU REFERENCYJNEGO

Genom referencyjny



PRZYRÓWNANIE DO GENOMU REFERENCYJNEGO

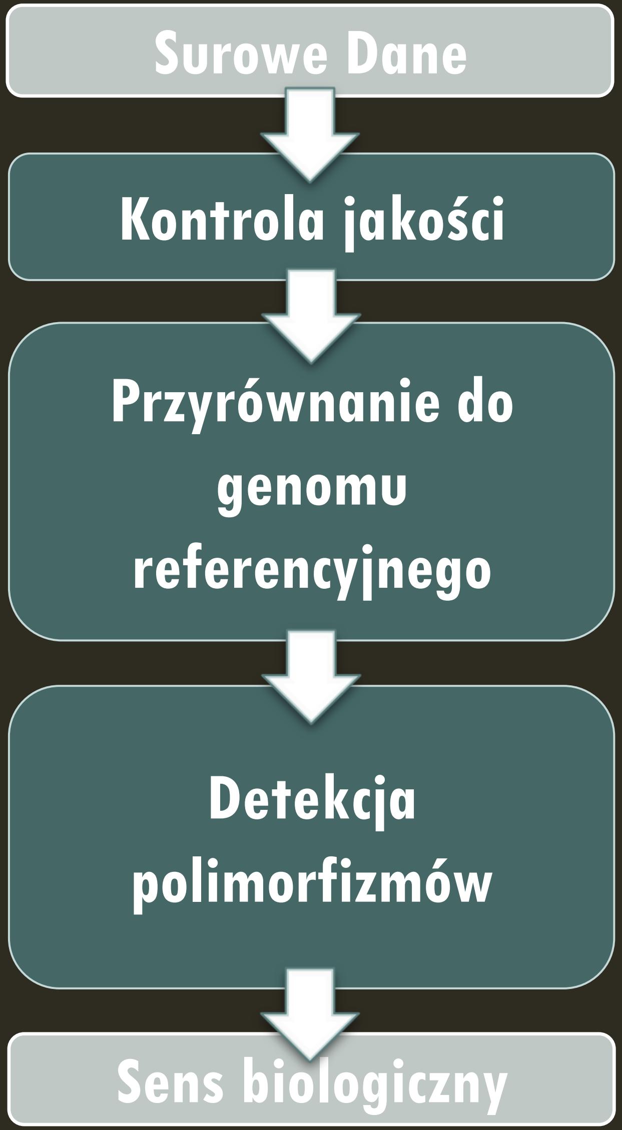
Genom referencyjny



PRZYRÓWNANIE DO GENOMU REFERENCYJNEGO

Table 1 The characteristics of suffix array-based alignment to the reference genome software

Name	Indexing	Output formats	PE mode	Gapped alignment	Supported platforms	Operating system
Bowtie	Genome	SAM	+	−	Illumina, ABI SOLiD	Linux, Macintosh, Windows, Solaris
Bowtie2	Genome	SAM	+	+	Illumina, 454, Ion Torrent	Linux, Macintosh, Windows
BWA	Genome	SAM	+	+	Illumina, 454, Ion Torrent	Linux
Segemehl	Genome	SAM	+	+	Illumina, 454	Linux
SOAP2 (SOAPaligner)	Genome	SOAP	+	+	Illumina	Linux
SOAP3	Genome	Binary/plain text, SAM	+	−	Illumina	Linux
SOAP3-db	Genome	Plain text, SAM, BAM	+	+	Illumina	Linux



SAM (SEQUENCE ALIGNMENT/MAP FORMAT)

header section

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

alignment section

SAM/BAM

```
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
      AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
      NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
      ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<<7<<<< \
      MF:i:18 RG:Z:L2
```

The alignment section consists of multiple TAB-delimited lines with each line describing an alignment. Each line is:

```
<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL> \
  [<TAG>:<VTYPE>:<VALUE> [...]]
```

Field	Regular expression	Range	Description
QNAME	`[^ \t\n\r]+`		Query pair NAME if paired; or Query NAME if unpaired [2]
FLAG	`[0-9]+`	$[0,2^{16}-1]$	bitwise FLAG (Section 2.2.2)
RNAME	`[^ \t\n\r@=]+`		Reference sequence NAME [3]
POS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	`[0-9]+`	$[0,2^{8}-1]$	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) [4]
CIGAR	`([0-9]+[MIDNSHP])+|\*`		extended CIGAR string
MRNM	`[^ \t\n\r@]+`		Mate Reference sequence NaMe; "=" if the same as <RNAME> [3]
MPOS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost Mate POSition of the clipped sequence
ISIZE	`-?[0-9]+`	$[-2^{29},2^{29}]$	inferred Insert SIZE [5]
SEQ	`[acgtnACGTN.=]+|\*`		query SEQuence; "=" for a match to the reference; n/N/. for ambiguity; cases are not maintained [6,7]
QUAL	`[!-~]+|\*`	[0,93]	query QUALity; ASCII-33 gives the Phred base quality [6,7]
TAG	`[A-Z][A-Z0-9]`		TAG
VTYPE	`[AifZH]`		Value TYPE
VALUE	`[^\t\n\r]+`		match <VTYPE> (space allowed)

HWI-1KL157:58:D2FVAACXX:2:2313:3871:71331 147 Chr15 33794413 50 101M = 33794252 -262 GCTCAGCTTTCTTCACAGTCCAACTCTCACATCCATACATGACCACTGGAAAAACCATAGCCTTGACTGGACGGACCTTTGTTAGAGGTTGCTAAAGACTG DBACCDDCDDDDCC>DECDDC@;?3?3HAHGEJHHEFHDAF@CF=EDEEGDGG@IGF@HEHGCGIIIJIGGDIHFGIJIJIEIIJIGG?FHHHFDDDD@@@ NM:i:2 AS:i:91 XS:i:83 RG:Z:D2FVAACXX_2

SAM/BAM

Field	Regular expression	Range	Description
QNAME	`[^ \t\n\r]+`		Query pair NAME if paired; or Query NAME if unpaired [2]
FLAG	`[0-9]+`	$[0,2^{16}-1]$	bitwise FLAG (Section 2.2.2)
RNAME	`[^ \t\n\r@=]+`		Reference sequence NAME [3]
POS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	`[0-9]+`	$[0,2^{8}-1]$	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) [4]
CIGAR	`([0-9]+[MIDNSHP])+|\*`		extended CIGAR string
MRNM	`[^ \t\n\r@]+`		Mate Reference sequence NaMe; "=" if the same as <RNAME> [3]
MPOS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost Mate POSition of the clipped sequence
ISIZE	`-?[0-9]+`	$[-2^{29},2^{29}]$	inferred Insert SIZE [5]
SEQ	`[acgtnACGTN.=]+|\*`		query SEquence; "=" for a match to the reference; n/N/. for ambiguity; cases are not maintained [6,7]
QUAL	`[!-~]+|\*`	[0,93]	query QUALity; ASCII-33 gives the Phred base quality [6,7]
TAG	`[A-Z][A-Z0-9]`		TAG
VTYPE	`[AifZH]`		Value TYPE
VALUE	`[^\t\n\r]+`		match <VTYPE> (space allowed)

SAM/BAM

HWI-1KL157:58:D2FVAACXX:2:2313:3871:71331 147 Chr15 33794413 50 101M = 33794252 -262
GCTCAGCTTTCTTCACAGTCCAACTCTCACATCCATACATGACCACTGGAAAAACCATAGCCTTGACTGGACGGACCT
TTGTTAGAGGTTGCTAAAGACTG
DBACCDDCDDDDCC>DECDDC@;?3?3HAHGEJHHEFHDAF@CF=EDEEGDGG@IGF@HEHGCGIIIJIG
GDIHFGIJIJIEIIJIGG?FHHHFDDDD@@@ NM:i:2 AS:i:91 XS:i:83 RG:Z:D2FVAACXX_2

https://broadinstitute.github.io/picard/explain-flags.html

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: 147 [Explain]

[Switch to mate] Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- [x] read paired
- [x] read mapped in proper pair

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
read reverse strand (0x10)
second in pair (0x80)

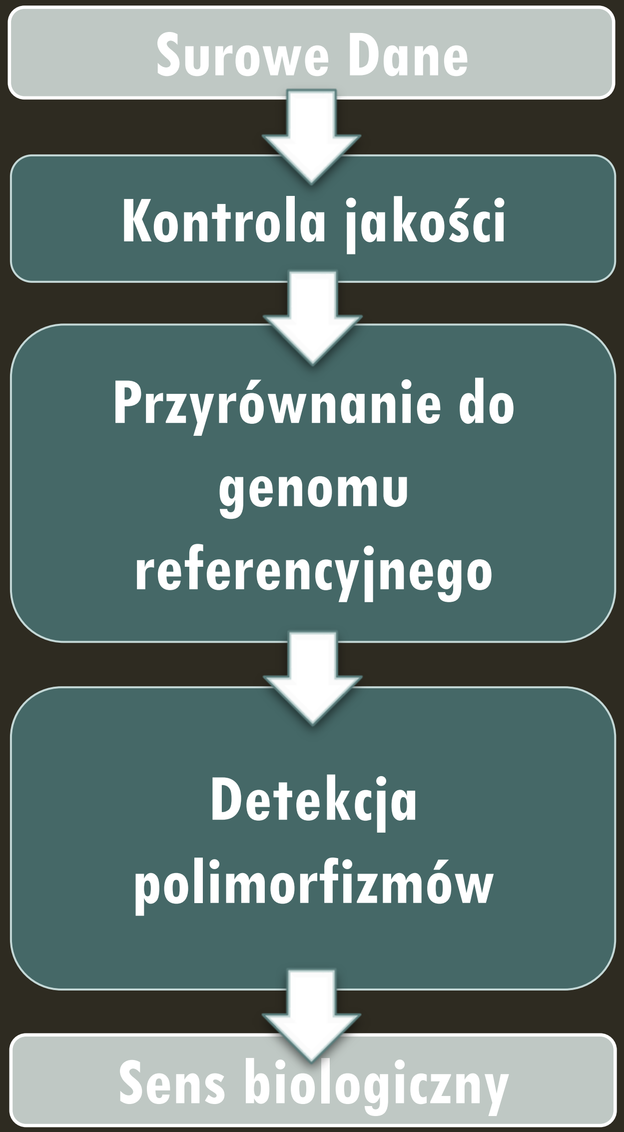
HWI-1KL157:58:D2FVAACXX:2:2313:3871:71331 147 Chr15 33794413 50 101M = 33794252 -262
GCTCAGCTTTCTTCACAGTCCAACTCTCACATCCATACATGACCACTGGAAAAACCATAGCCTTGACTGGACGGACCT TTGTTAGAGGTTGCTAAAGACTG
DBACCDDCDDDDCC>DECDDC@;?3?3HAHGEJHHEFHDAF@CF=EDEEGDGG@IGF@HEHGCGIIIJIG GDIHFGIJIJIEIIJIGG?FHHHFDDDD@@@ NM:i:2 AS:i:91 XS:i:83 RG:Z:D2FVAACXX_2

SAM/BAM

Field	Regular expression	Range	Description
QNAME	`[^ \t\n\r]+`		Query pair NAME if paired; or Query NAME if unpaired [2]
FLAG	`[0-9]+`	$[0,2^{16}-1]$	bitwise FLAG (Section 2.2.2)
RNAME	`[^ \t\n\r@=]+`		Reference sequence NAME [3]
POS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	`[0-9]+`	$[0,2^{8}-1]$	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) [4]
CIGAR	`([0-9]+[MIDNSHP])+|\*`		extended CIGAR string
MRNM	`[^ \t\n\r@]+`		Mate Reference sequence NaMe; "=" if the same as <RNAME> [3]
MPOS	`[0-9]+`	$[0,2^{29}-1]$	1-based leftmost Mate POSition of the clipped sequence
ISIZE	`-?[0-9]+`	$[-2^{29},2^{29}]$	inferred Insert SIZE [5]
SEQ	`[acgtnACGTN.=]+|\*`		query SEQuence; "=" for a match to the reference; n/N/. for ambiguity; cases are not maintained [6,7]
QUAL	`[!-~]+|\*`	[0,93]	query QUALity; ASCII-33 gives the Phred base quality [6,7]
TAG	`[A-Z][A-Z0-9]`		TAG
VTYPE	`[AifZH]`		Value TYPE
VALUE	`[^\t\n\r]+`		match <VTYPE> (space allowed)

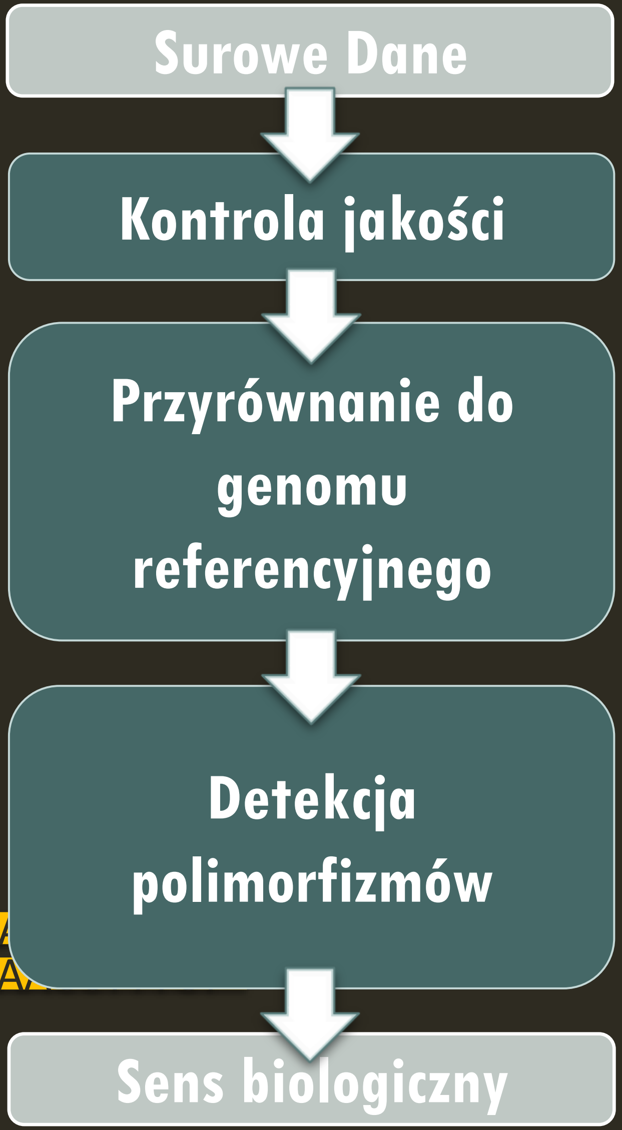
DETEKCJA POLIMORFIZMÓW

• Single Nucleotide Polymorphisms	SNP
• Insertions/Deletions	INDEL
• Copy Number Variations	CNV
• Loss of Heterozygosity	LOH
• Inversions	INV
• Translocations	TRANS



DETEKCJA POLIMORFIZMÓW

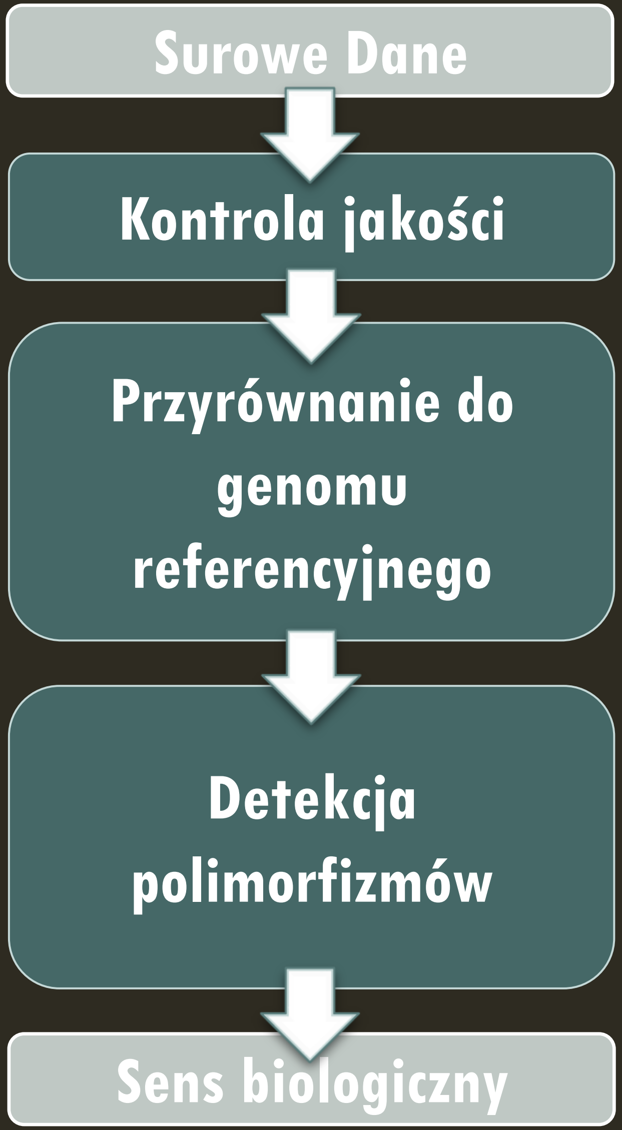
Genom referencyjny



DETEKCJA POLIMORFIZMÓW → SNP

Table 3 The characteristics of SNP calling software

Name	Input formats	Output formats	Realignment	Recalibration	Single-sample	Multi-sample	Called variants	Operating system
Atlas-SNP2	BAM	VCF	−	−	+	−	SNPs	Linux
GATK (UnifiedGenotyper)	BAM	VCF	+	+	+	+	SNPs, InDels	Linux
SAMtools (samtools mpileup)	BAM	VCF	+	+	+	+	SNPs, InDels	Linux
SNVer	BAM	VCF	−	−	+	−	SNPs, InDels	Linux, Macintosh, Windows
SOAPsnp	SOAP	Text format, GLFv2	−	+	+	−	SNPs	Linux
Varscan2	Pileup/mpileup	Text format, VCF	−	−	+	+	SNPs, InDels, CNA	Linux, Macintosh, Windows



http://samtools.github.io/hts-specs/VCFv4.1.pdf / http://samtools.github.io/hts-specs/VCFv4.3.pdf

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",t
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

This example shows in order a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

header section

polimorphisms (body)

VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

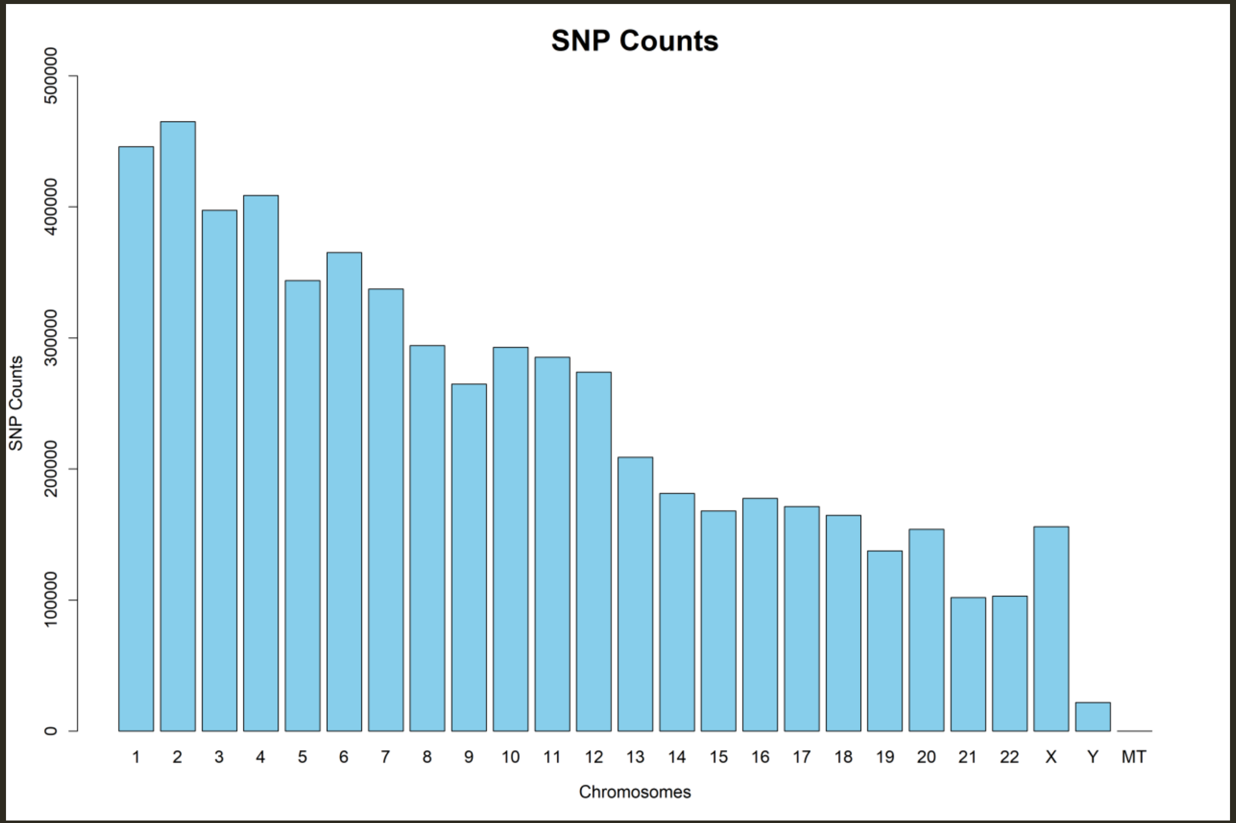
Chr1 238 . C T 48 . DP=8;VDB=4.789490e-02;RPB=-1.551181e+00;AF1=0.5;AC1=1;DP4=4,1,3,0;MQ=44;FQ=51; PV4=1,1,0.16,0.41 GT:PL:GQ 0/1:78,0,135:81

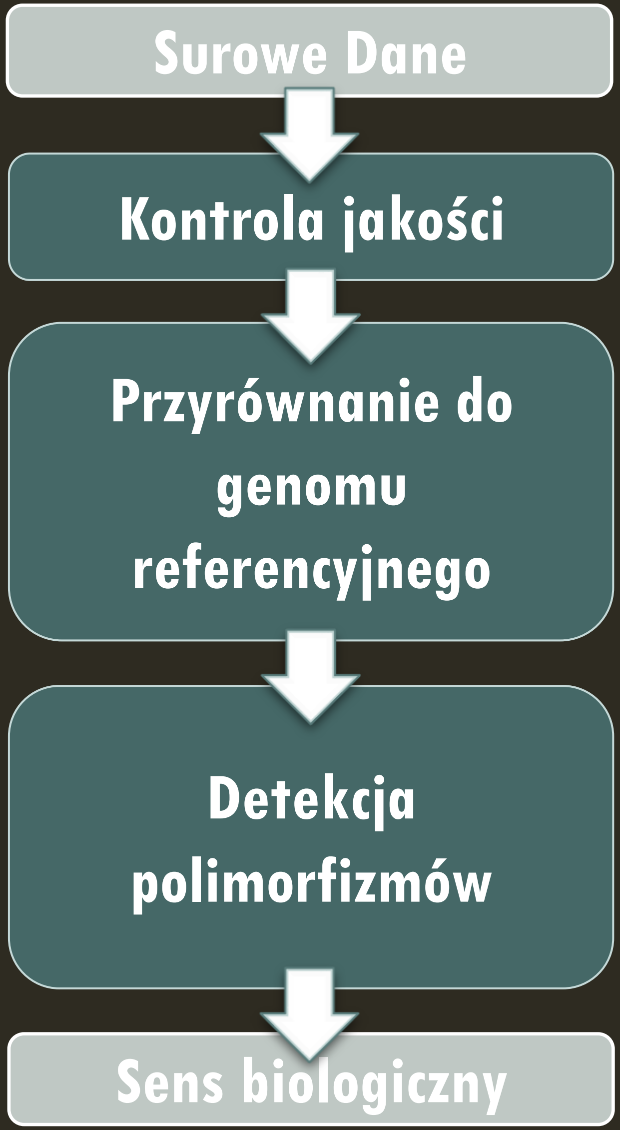
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4

This example shows in order a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

DETEKCJA POLIMORFIZMÓW → SNP

Homo sapiens





SENS BIOLOGICZNY

→FUNKCJONALNA ADNOTACJA

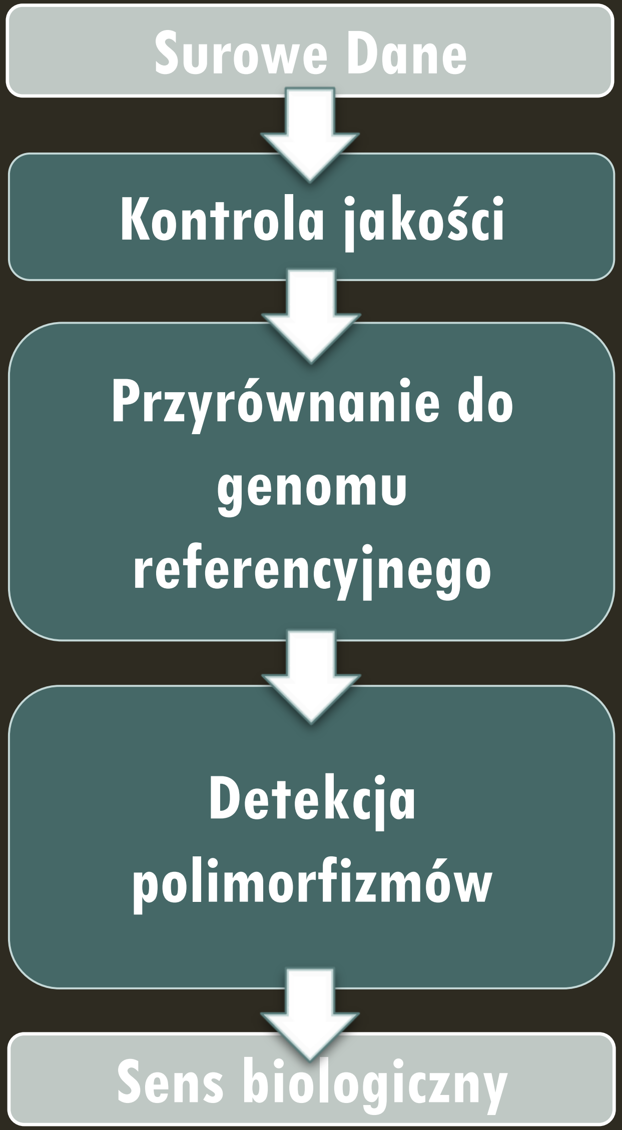
Variant Effect Predictor

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **genes** and **transcripts** affected by the variants
- **location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen** scores for changes to protein sequence
- ... And more!

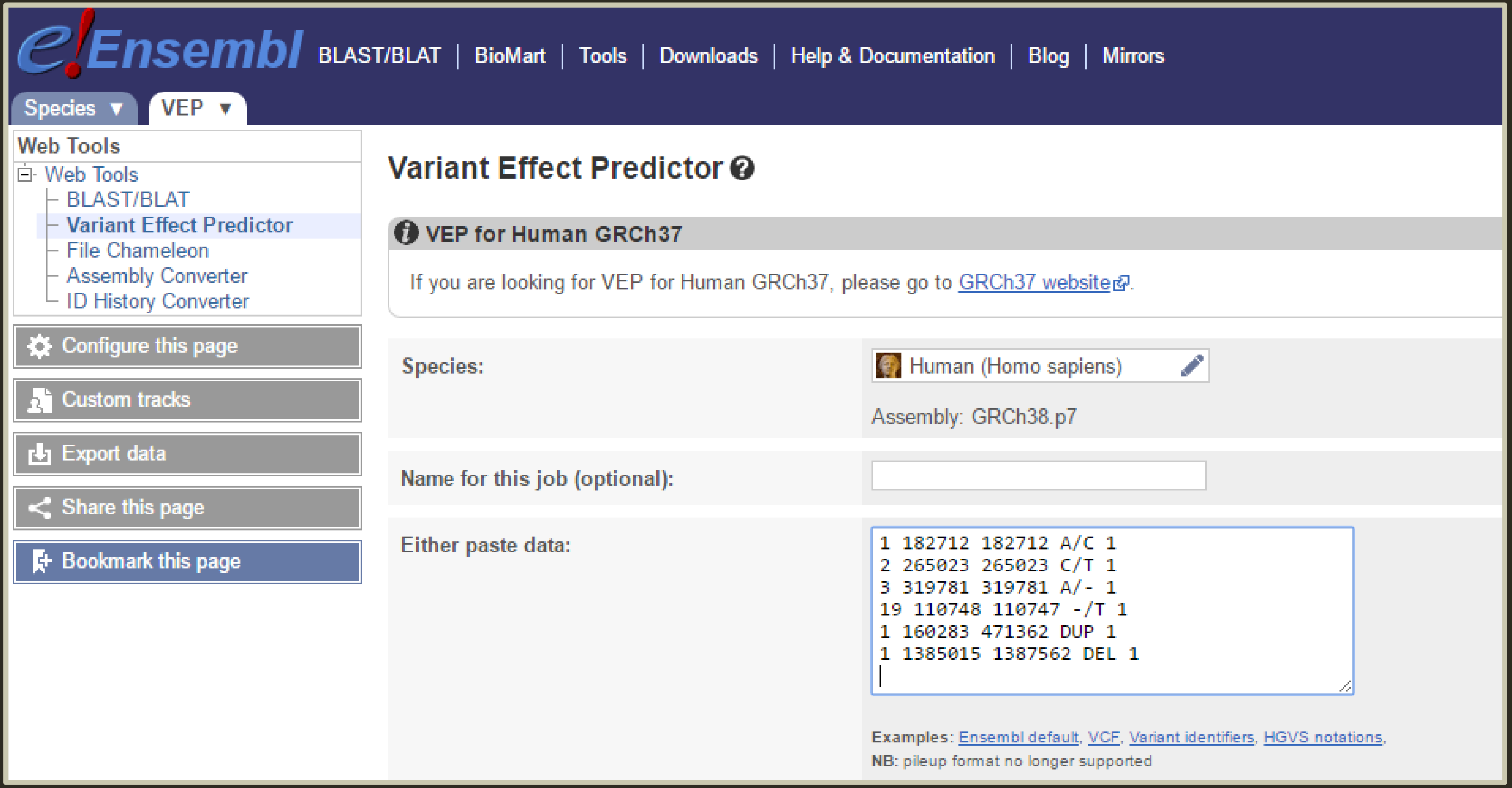
Launch VE!P

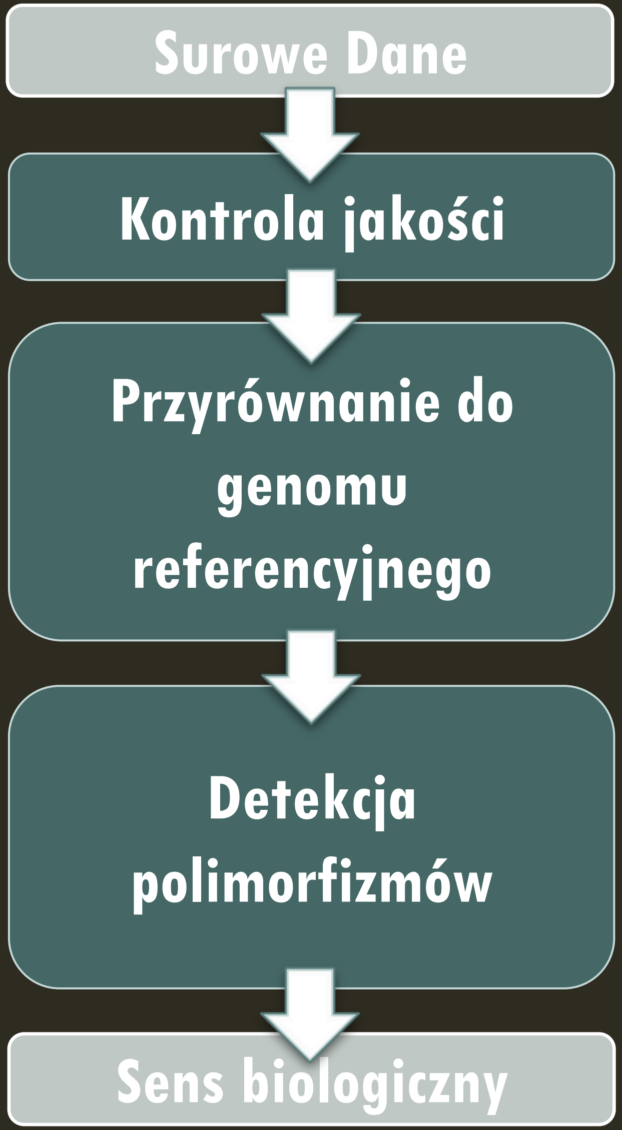
→ www.ensembl.org/Tools/VEP



SENS BIOLOGICZNY

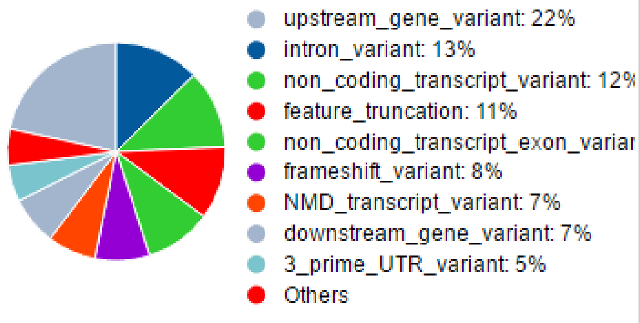
→FUNKCJONALNA ADNOTACJA



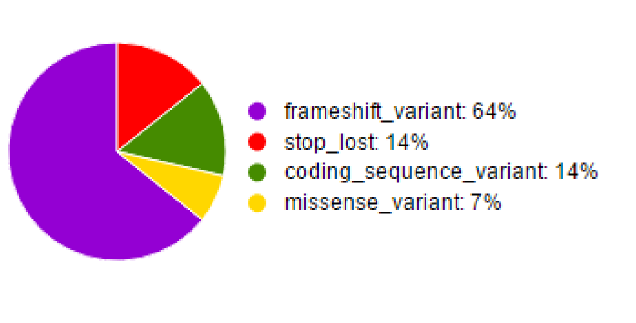


Category	Count
Variants processed	6
Variants filtered out	0
Novel / existing variants	5 (83.3) / 1 (16.7)
Overlapped genes	10
Overlapped transcripts	70
Overlapped regulatory features	1

Consequences (all)



Coding consequences



Results preview

Navigation

Page: 1 of 1 | **Show:** All variants

Filters

Uploaded variant ▾ | is ▾ | defined | **Add**

Download

All: VCF VEP TXT

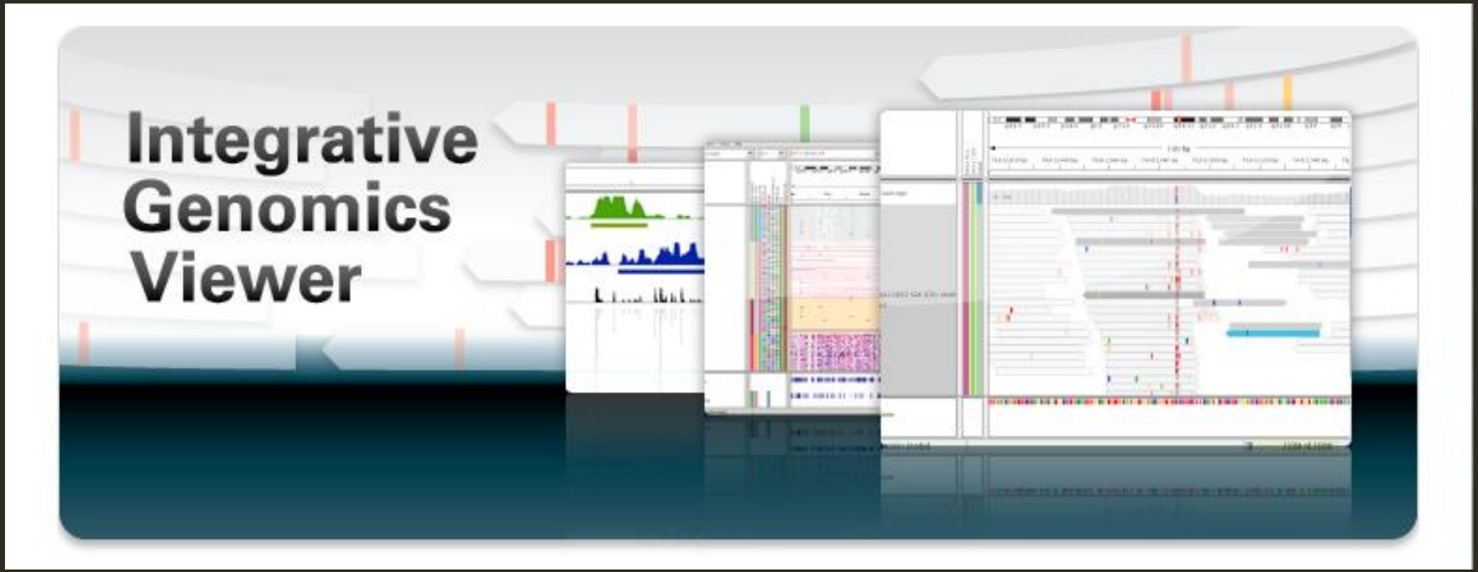
BioMart: Variants Genes

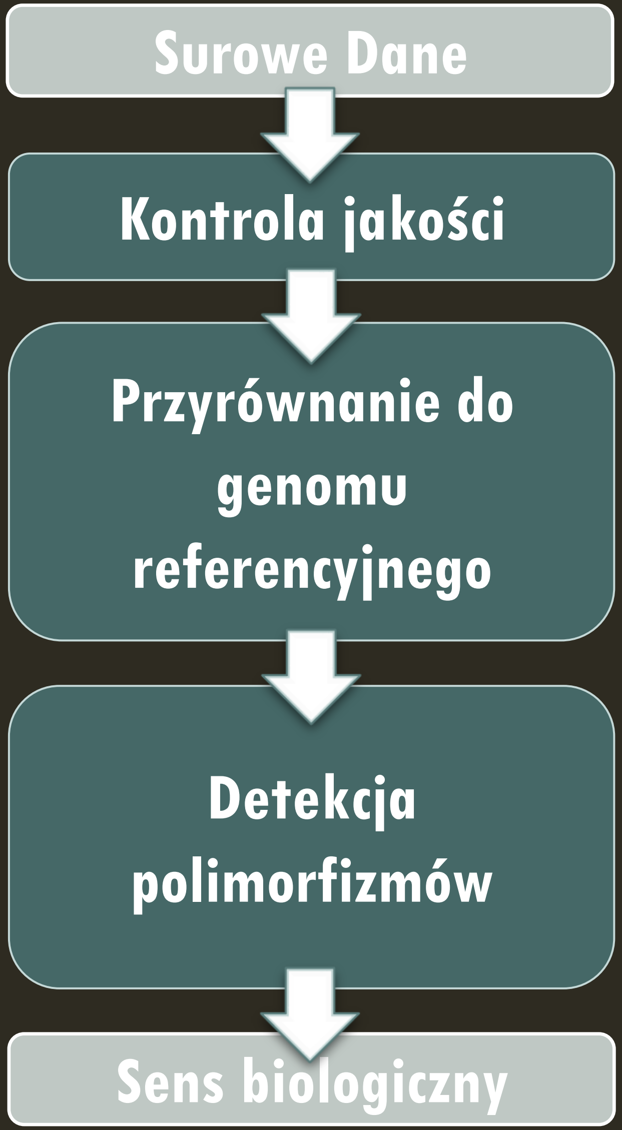
Show/hide columns

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype
1_160283_duplication	1:160282-160282	duplication	upstream_gene_variant	MODIFIER	RNU6-1100P	ENSG00000222623	Transcript	ENST00000410691	snRNA
1_160283_duplication	1:160282-160282	duplication	non_coding_transcript_exon_variant, intron_variant, non_coding_transcript_variant	MODIFIER	RP11-34P13.13	ENSG00000241860	Transcript	ENST00000466557	lincRNA
1_160283_duplication	1:160282-160282	duplication	non_coding_transcript_exon_variant, intron_variant, non_coding_transcript_variant	MODIFIER	RP11-34P13.13	ENSG00000241860	Transcript	ENST00000491962	lincRNA
1_160283_duplication	1:160282-160282	duplication	non_coding_transcript_exon_variant, intron_variant, non_coding_transcript_variant	MODIFIER	RP11-34P13.9	ENSG00000241599	Transcript	ENST00000496488	lincRNA
1_182712_A/C	1:182712-182712	C	downstream_gene_variant	MODIFIER	FO538757.1	ENSG00000279457	Transcript	ENST00000623083	protein_coding
1_182712_A/C	1:182712-182712	C	downstream_gene_variant	MODIFIER	FO538757.1	ENSG00000279457	Transcript	ENST00000623834	protein_coding
1_182712_A/C	1:182712-182712	C	missense_variant	MODERATE	FO538757.2	ENSG00000279928	Transcript	ENST00000624431	protein_coding
1_182712_A/C	1:182712-182712	C	downstream_gene_variant	MODIFIER	FO538757.1	ENSG00000279457	Transcript	ENST00000624735	protein_coding

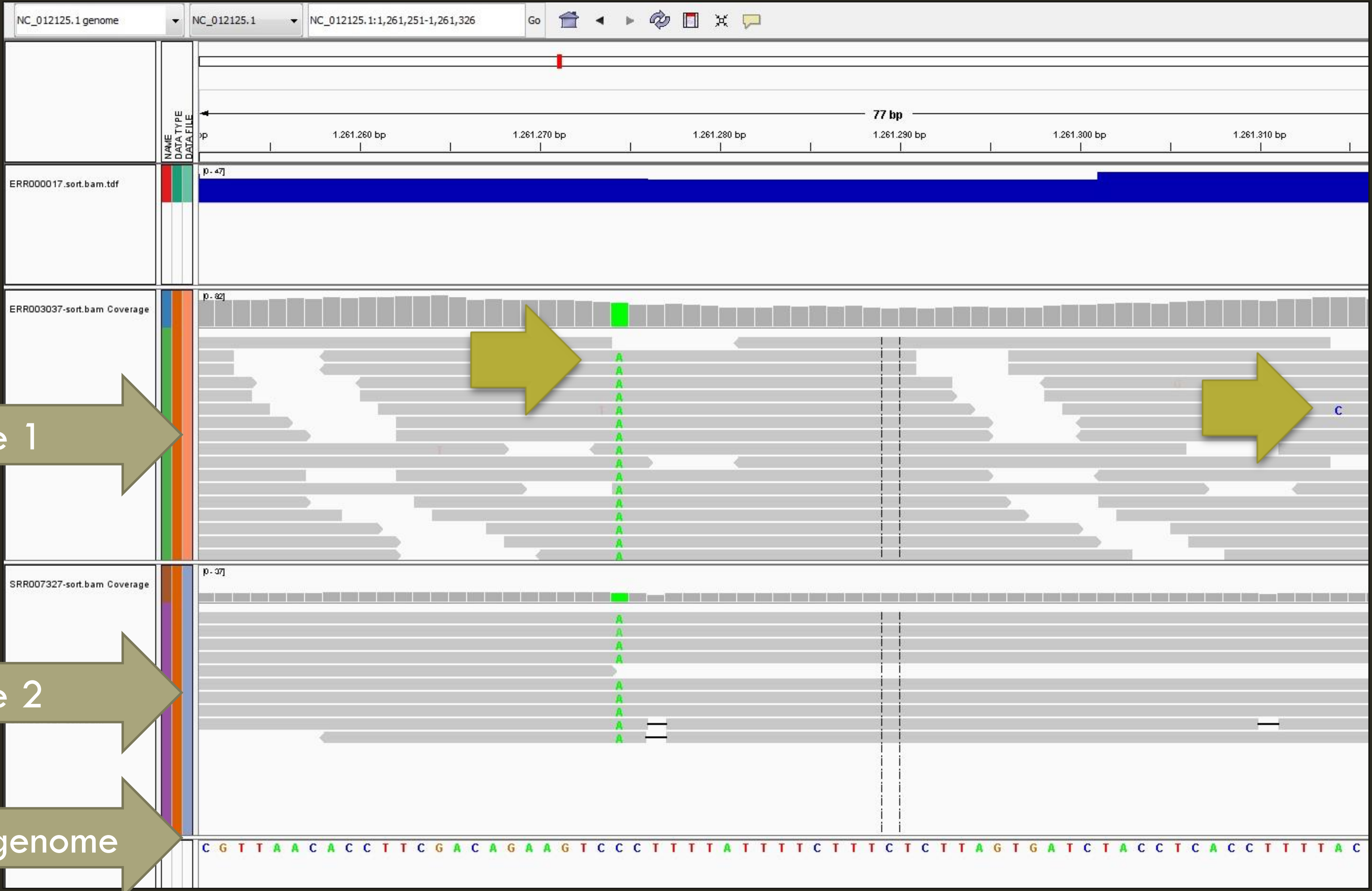
WIZUALIZACJA

Wizualizacja w programie IGV
http://software.broadinstitute.org/software/igv/





IGV

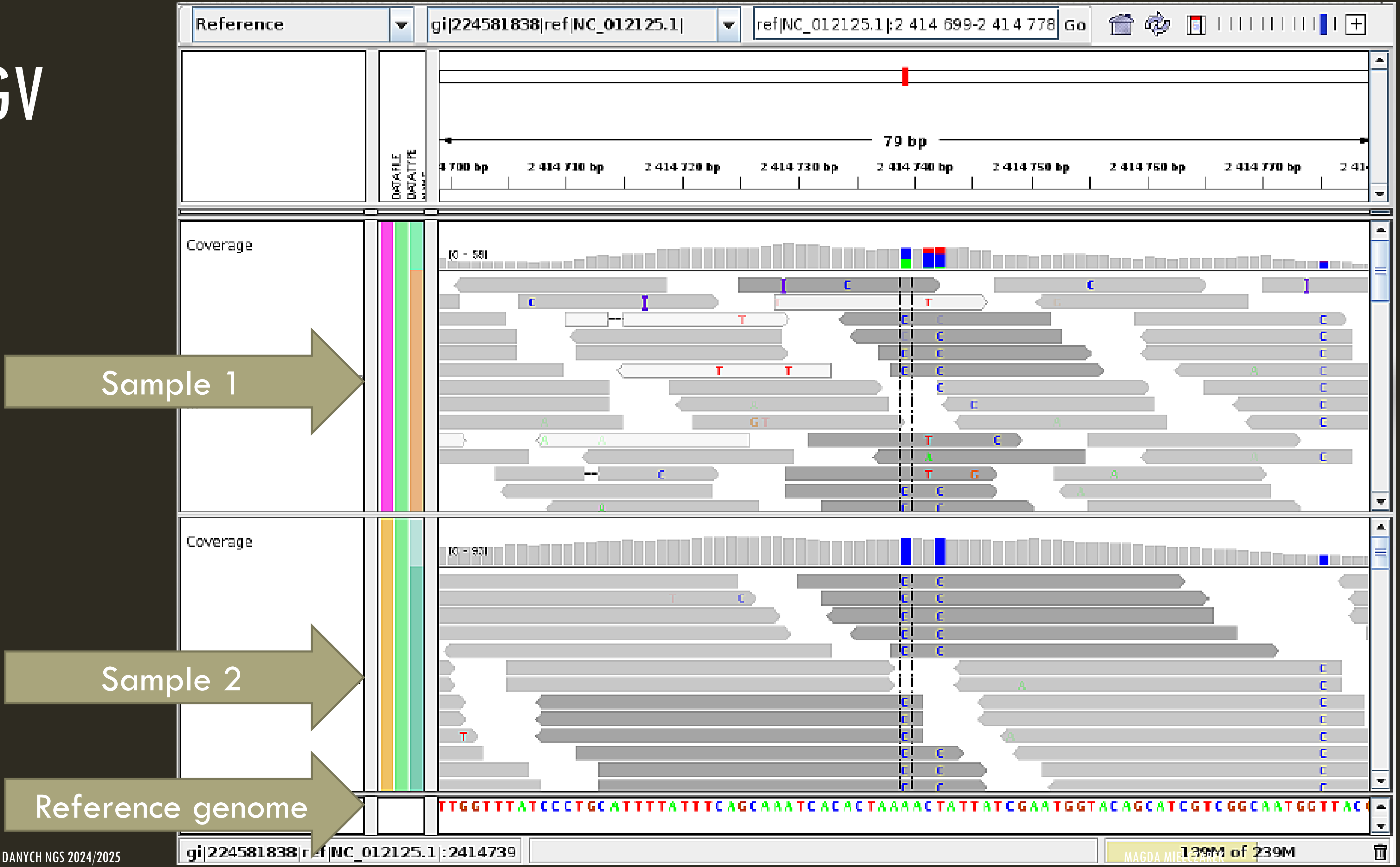


Sample 1

Sample 2

Reference genome

IGV


Reference
gi|224581838|ref|NC_012125.1|
ref|NC_012125.1|:2 414 699-2 414 778
Go
79 bp
Coverage
Coverage
Sample 1
Sample 2
Reference genome
gi|224581838|ref|NC_012125.1|:2414739
139M of 239M

SENS BIOLOGICZNY → MEDYCYNA PERSONALIZOWANA

doi: 10.2217/pme.14.34

The road from next-generation sequencing to personalized medicine

Manuel L. Gonzalez-Garay

Author information ► Copyright and License information ►

Abstract

Go to:

Moving from a traditional medical model of treating pathologies to an individualized predictive and preventive model of personalized medicine promises to reduce the healthcare cost on an overburdened and overwhelmed system. Next-generation sequencing (NGS) has the potential to accelerate the early detection of disorders and the identification of pharmacogenetics markers to customize treatments. This review explains the historical facts that led to the development of NGS along with the strengths and weakness of NGS, with a special emphasis on the analytical aspects used to process NGS data. There are solutions to a the steps necessary for performing NGS in the clinical context where the majority of them are very efficient, but there are some crucial steps in the process that need immediate attention.

Keywords: CADD, functional prediction program, genomics, GWAVA, NGS, personalized medicine, workflow management system

The current medical model focuses on the detection and treatment of pathologies. Treating disorders, especially on advanced states, is very expensive for patients and society in general. Screening for five of the most common disorders in the USA (cardiovascular disorders, stroke, cancer, chronic obstructive pulmonary disease and diabetes) could protect million of lives and reduce the healthcare deficit [1]. Tailoring drug therapies by practicing personali ine (PM) has the potential to improve treatment of cancer and save lives by preventing drug-rel . A new technology, next-generation sequencing (NGS), has the potential to accelerate the e y detection of disorders and to detect pharmacogenetics markers to customize treatments [2].

Use of NGS to diagnose human disorders

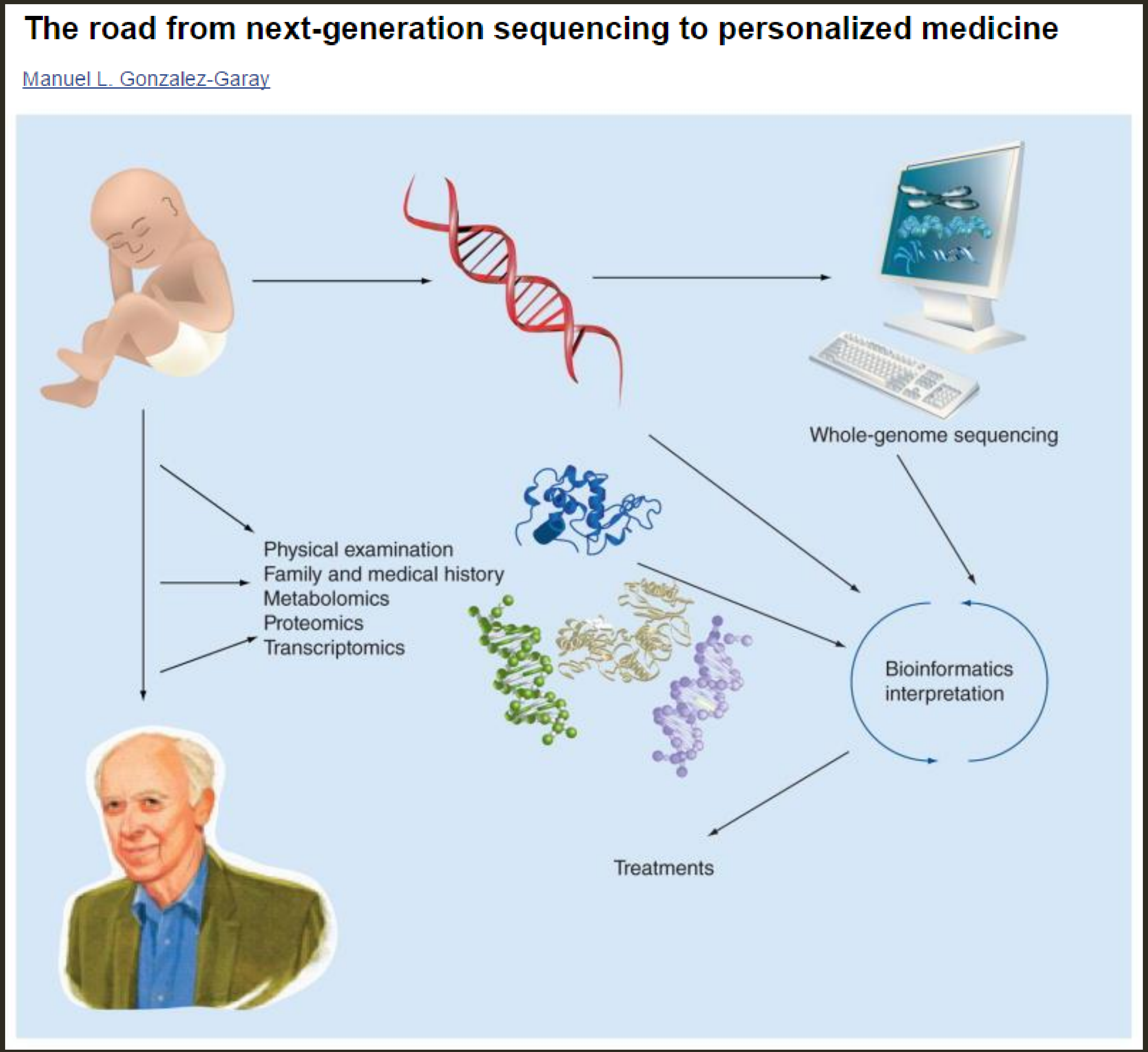
Go to:

O f the major concerns of medical diagnosis is to identify genes and mutations responsible for human . Early identification of causative mutations enables the early detection of a myriad of disorders. W e living in an age of high healthcare cost. Early detection of genetic disorders, carrier status, genetic predispositions for cancer and cardiovascular disease could potentially reduce the healthcare cost.

The first proof of concept that the NGS technology could be used to detect genetic disorders was provided by Shendure's group on September 2009 [225]. A few months later, the same group reported the detection of the first recessive disorder (Miller syndrome) detected by whole-exome sequencing (WES) [226]. These two papers marked a new era where NGS became the preferred tool for rare Mendelian disease gene identification. There are several excellent reviews that describe the exponential growth in disease gene identification that started in 2010 [227–229]. Up to 27 February 2014, the number of genes with phenotype-causing mutations has reached 3162 according to online Mendelian inheritance in man (OMIM) Mgene map statistics [230]. In a recent review, Rabbani *et al.* estimated that from January 2010 to May 2012, over 100 causative genes in various Mendelian disorders have been identified by means of exome sequencing [231].

WES is now a valid and standard diagnostic approach for the identification of molecular defects in patients with suspected genetic disorders. This fact was demonstrated last year by a publication in the *New England Journal of Medicine* by the Medical Genetics Laboratory group of Baylor College of Medicine. The group reported the WES sequencing of 250 probands referred by physician, 98% of the cases were billed to the insurance. They reported a 25% molecular diagnostic rate (62 cases) [232]. In September 2013, the NIH funded four groups to explore the use of NGS for newborn screening [233]. With the cost per genome getting close to the US$1000, it is becoming affordable to get sequenced at an early age, allowing for reanalysis of our genetic information at multiple intervals during the life of a person (Figure 3). A recent review outlines the approach, challenges, and benefits of such screening for adult genetic disease risks [2]. We a ly published a proof of concept project aimed to evaluate the benefits of screening healthy adul . Our pilot project demonstrated that when WES is combined with medical and family histo ngs are substantial. In a cohort of 81 unrelated individuals, we identified 271 recessive risk alleles (214 genes), 126 dominant risk alleles (101 genes) and three X-recessive risk alleles (three genes). In addition, we linked personal disease histories with causative disease genes in 18 volunteers [234].

SENS BIOLOGICZNY→ MEDYCYNA PERSONALIZOWANA



Surowe Dane

↓

Kontrola jakości

↓

Przyrównanie do genomu referencyjnego

↓

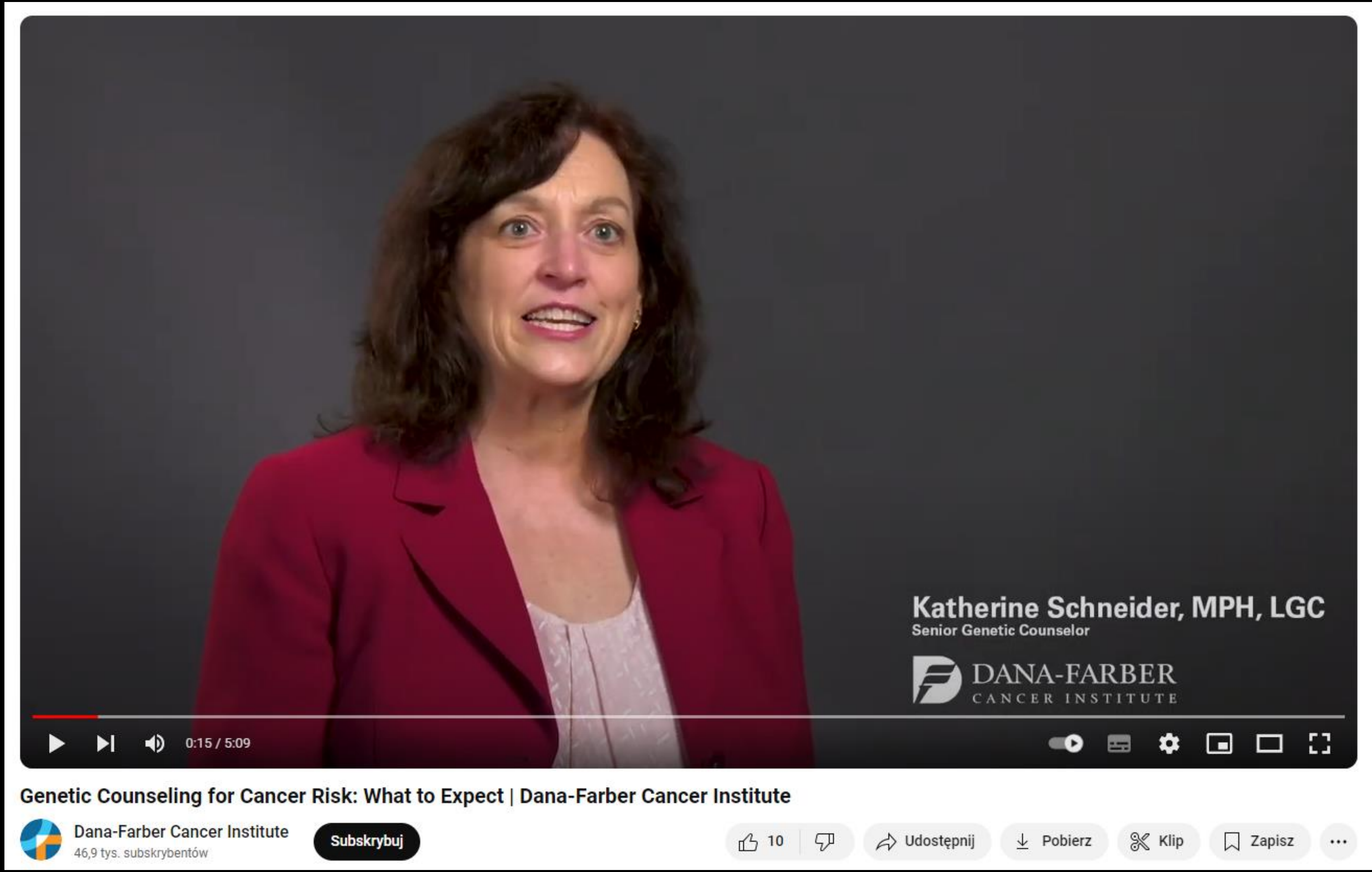
Detekcja polimorfizmów

↓

Sens biologiczny

SENS BIOLOGICZNY → MEDYCYNA PERSONALIZOWANA

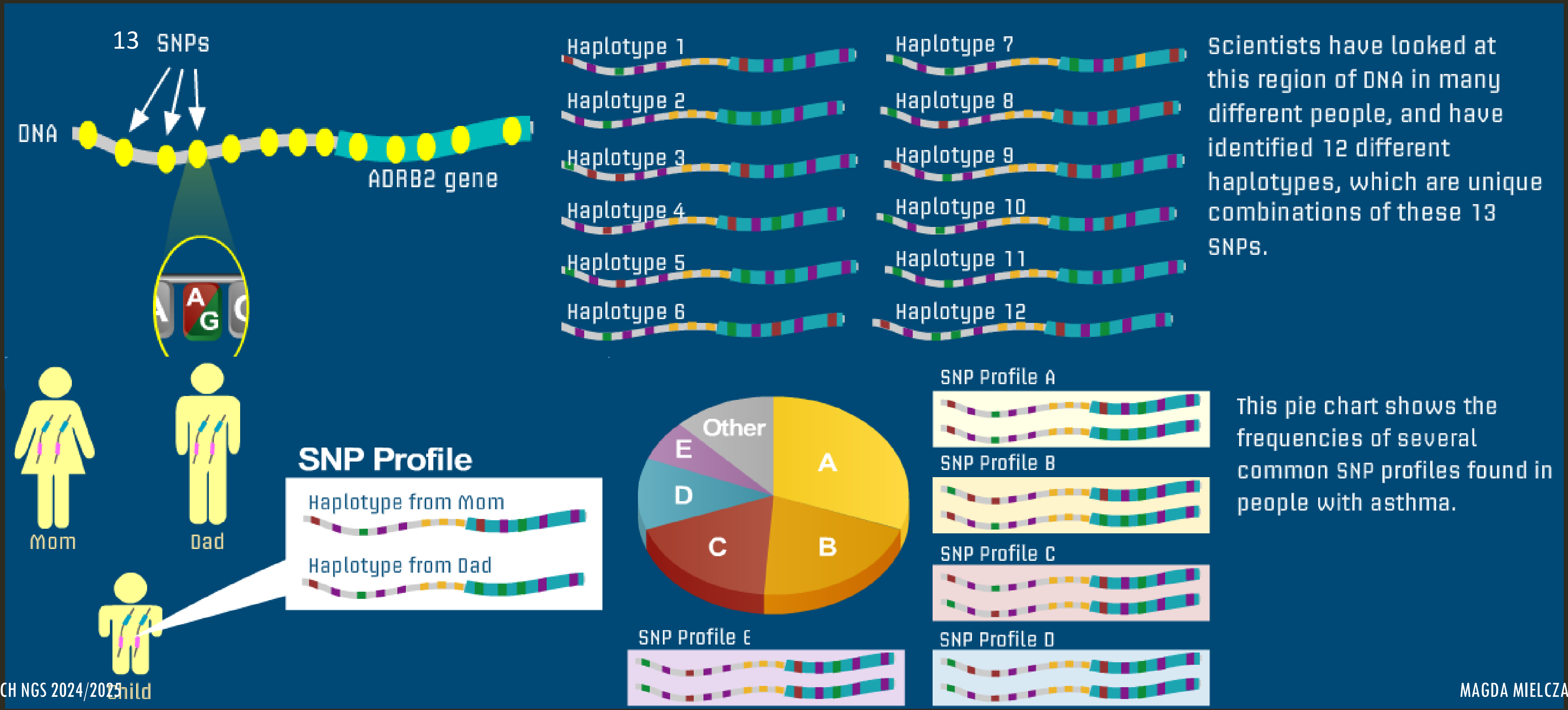
www.youtube.com/watch?v=C34dhHUWp0Y



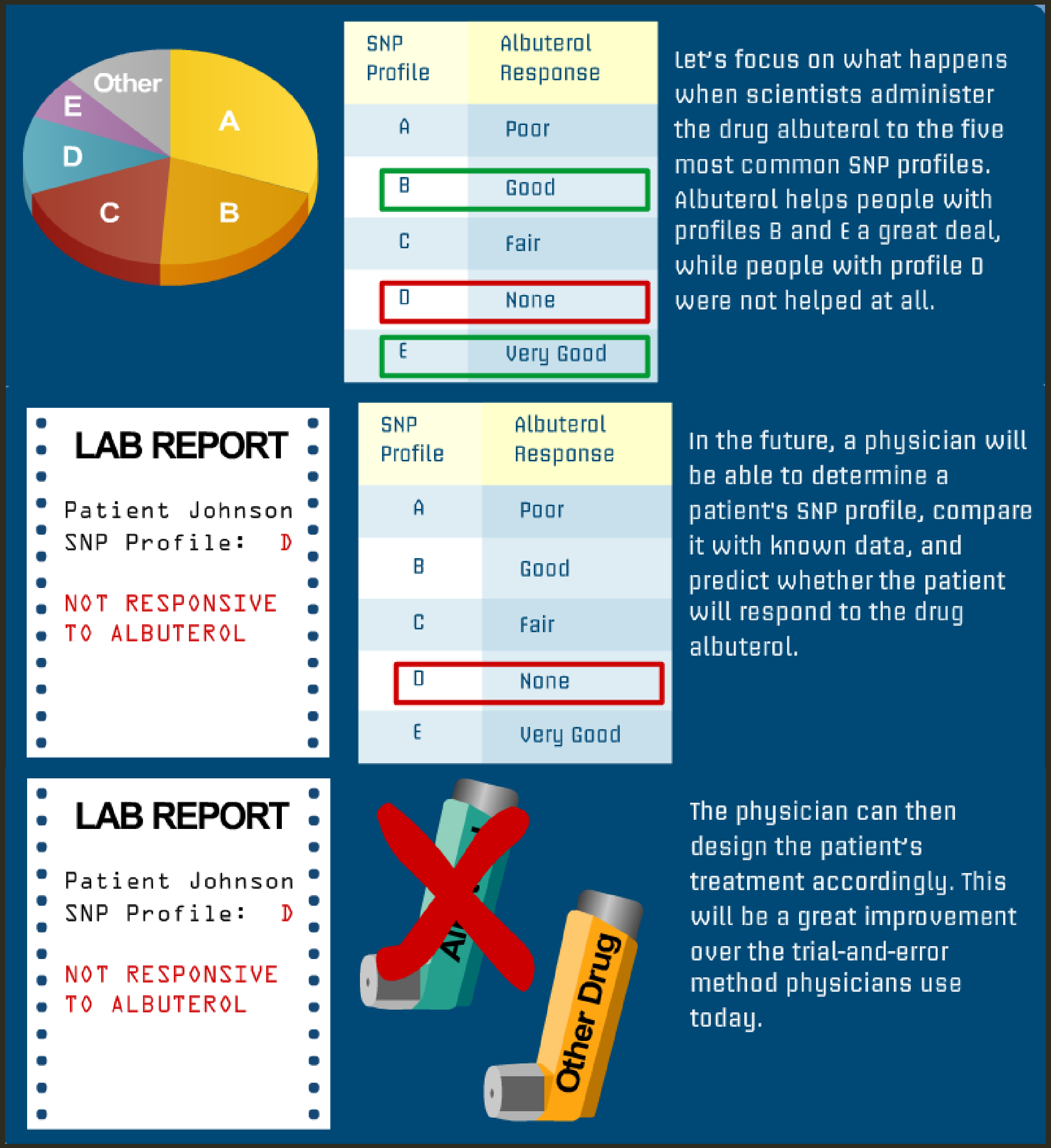
SENS BIOLOGICZNY→ FARMAKOGENETYKA/FARMAKOGENOMIKA

„Applying SNP profiles to drug choices”

- astma ← gen ADRB2
- polimorfizmy punktowe a odpowiedź na leczenie albuterolem
- albuterol łagodzi objawy astmy tylko u niektórych chorych



SENS BIOLOGICZNY→ FARMAKOGENETYKA/FARMAKOGENOMIKA



Review | Open access | Published: 26 October 2020

Best practices for variant calling in clinical sequencing

Daniel C. Koboldt

Genome Medicine **12**, Article number: 91 (2020) | Cite this article

Abstract

Next-generation sequencing technologies have enabled a dramatic expansion of clinical genetic testing both for inherited conditions and diseases such as cancer. Accurate variant calling in NGS data is a critical step upon which virtually all downstream analysis and interpretation processes rely. Just as NGS technologies have evolved considerably over the past 10 years, so too have the software tools and approaches for detecting sequence variants in clinical samples. In this review, I discuss the current best practices for variant calling in clinical sequencing studies, with a particular emphasis on trio sequencing for inherited disorders and somatic mutation detection in cancer patients. I describe the relative strengths and weaknesses of panel, exome, and whole-genome sequencing for variant detection. Recommended tools and strategies for calling variants of different classes are also provided, along with guidance on variant review, validation, and benchmarking to ensure optimal performance. Although NGS technologies are continually evolving, and new capabilities (such as long-read single-molecule sequencing) are emerging, the "best practice" principles in this review should be relevant to clinical variant calling in the long term.

Emphasis on *trio* sequencing for inherited disorders and somatic mutation detection in cancer patients

Strengths and weaknesses of:
- panel
- exome
- whole-genome sequencing

for variant detection.

Recommended ***tools*** and ***strategies*** for calling variants of different classes

The "best practice" principles in this review should be relevant to clinical variant calling in the long term.

Review | Open access | Published: 26 October 2020

Best practices for variant calling in clinical sequencing

Daniel C. Koboldt

Genome Medicine **12**, Article number: 91 (2020) | Cite this article

NGS technologies enabled ambitious large-scale genomic sequencing efforts that have transformed our understanding of human health and disease, such The Cancer Genome Atlas, the Centers for Mendelian Genomics.

They have also been widely *adopted for clinical genetic testing*:

→ Targeted panels to interrogate medically relevant subsets of genes have become core components of precision oncology.

→ *Whole-exome sequencing*, which selectively targets the protein-coding regions of known genes, has become a frontline diagnostic tool for inherited disorders

SENS BIOLOGICZNY→ BADANIA PRENATALNE

OPIS TESTU

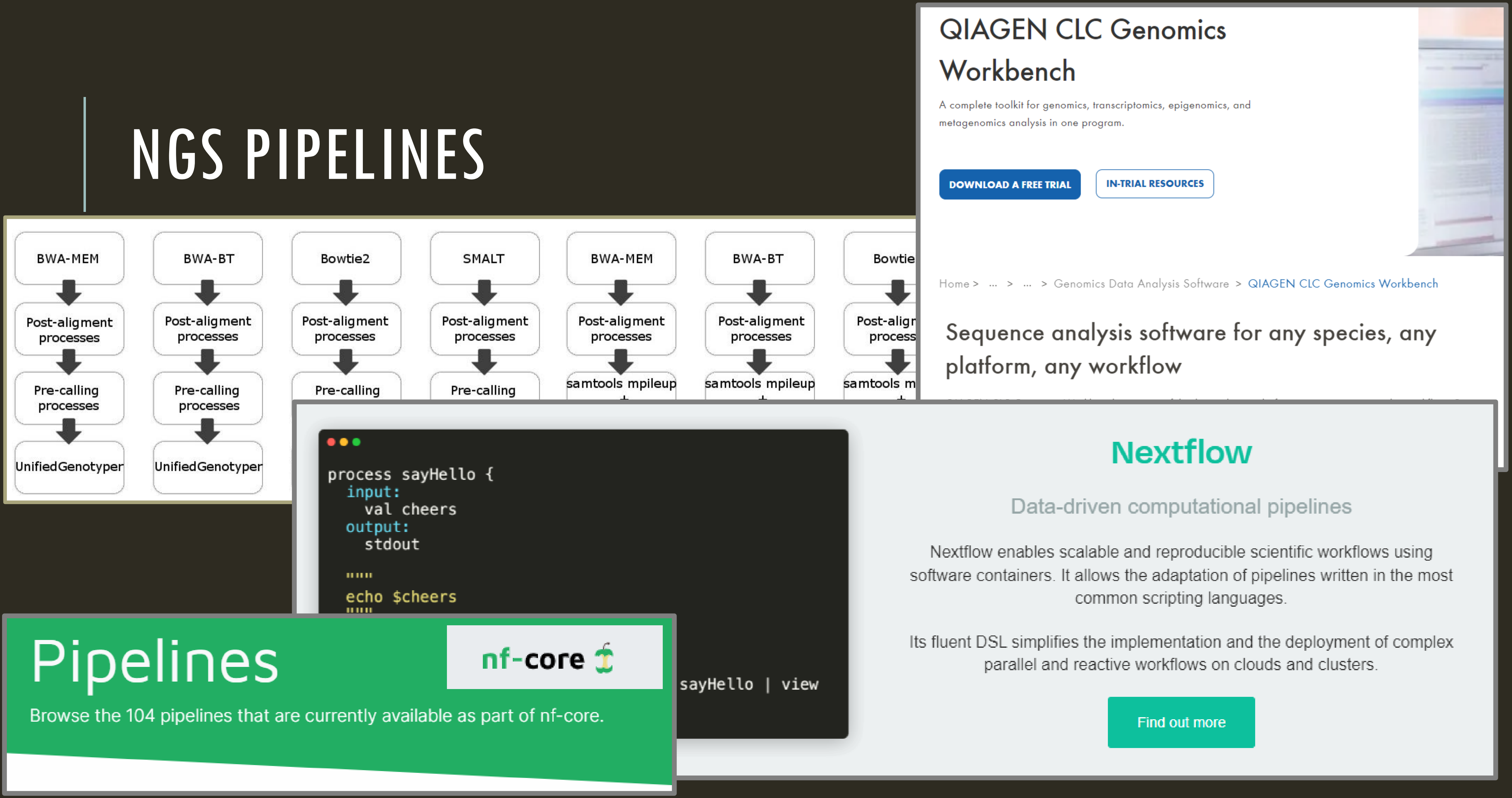
Test NIFTY Pro™ jest testem przesiewowym, nie jest testem diagnostycznym. Jego działanie polega na izolowaniu cfDNA (w tym zarówno matczynego, jak i płodowego DNA) z próbki krwi matki i wykonywaniu sekwencjonowania całego genomu przy użyciu technologii sekwencjonowania nowej generacji. Unikalne odczyty każdego chromosomu są obliczane i porównywane z optymalną próbką kontrolną stanowiącą punkt odniesienia. Dane są analizowane przy użyciu zastrzeżonych algorytmów bioinformatycznych BGI i ocena jest sporządzana tylko dla badanych warunków. Badania powinny być zawsze zamawiane przez wykwalifikowanego pracownika służby zdrowia a wyniki omawiane z pacjentem. Badanie nie może być stosowane jako jedyna podstawa do diagnozy lub innej decyzji w sprawie leczenia ciąży. Identyfikacja płci polega na izolowaniu DNA wolnego od komórek (w tym zarówno matczynego, jak i płodowego DNA) z próbki krwi matki, a następnie molekularnych badań genetycznych w celu określenia względnych ilości chromosomu Y.

ZABURZENIE	CZUŁOŚĆ
Trisomia 21	99,17%
Trisomia 18	98,24%
Trisomia 13	>99,9%
Trisomia 9, 16, 22	NA
CNV	>90%
Płeć	99,53%
ANEUPLOIDIE CHROMOSOMÓW PŁCIOWYCH	**CZUŁOŚĆ**
XYY	>99,9%
XXY	>99,9%
XXX	>99,9%
XO	>99,9%

ZASTRZEŻENIA

Test NIFTY Pro™ NIE jest testem diagnostycznym, wyniki są przeznaczone do wykorzystania informacyjnego, dlatego nie można wykluczyć wyników fałszywie dodatnich i fałszywie ujemnych. Wydajność "Inne ustalenia" nie została w pełni zweryfikowana, ale dane w poniższej tabeli mogą być wykorzystane do celów informacyjnych. W tym teście wykryto 84 rodzaje zespołów del/dup. Niektóre z chorób na liście zespołów del/dup mogą być również spowodowane przez inne czynniki genetyczne, NIFTY ProTM tylko wykrywa i analizuje konkretny fragment zgodnie z autoryzowanymi bazami danych. Wykrywanie chromosomu Y zawarte w niniejszym raporcie nie może być wykorzystywane do diagnozowania płci płodowej lub chorób związanych z płcią i jest wykorzystywane wyłącznie jako dodatkowe informacje do analizy referencyjnej. Potencjalne źródła niedokładnego wyniku badania mogą obejmować między innymi: mozaicyzm matki, płodu i/lub łożyska, niską frakcję płodową, transfuzję krwi, operację przeszczepu, terapię komórkami macierzystymi, terapię heparyną i nieprawidłowy kariotyp biologicznych rodziców lub surogatki. Wynik badania jest specyficzny dla badanej próbki i powinien być zawsze interpretowany przez wykwalifikowanego specjalistę w kontekście danych klinicznych i rodzinnych.

NGS PIPELINES



FORMATY DANYCH - PODSUMOWANIE

Surowe dane

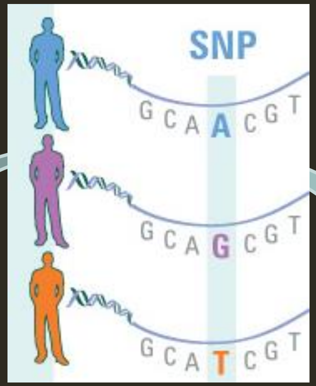
FASTQ

Przyrównanie do genomu referencyjnego

READS
&
REF

SAM/BAM

Detekcja polimorfizmów



VCF