NanoPipe – interactive tool for MinION sequencing analysis

Poznan Summer School of Bioinformatics 2017

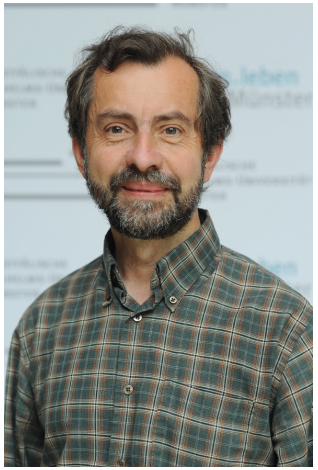
Poznan, September 8

Victoria Shabardina

Nanopipe is a web-based tool for fast and easy processing and analysis of the MinIon sequencing data.

Created in the Institute of Bioinformatics, University of Münster, Germany

Wojciech Makałowski

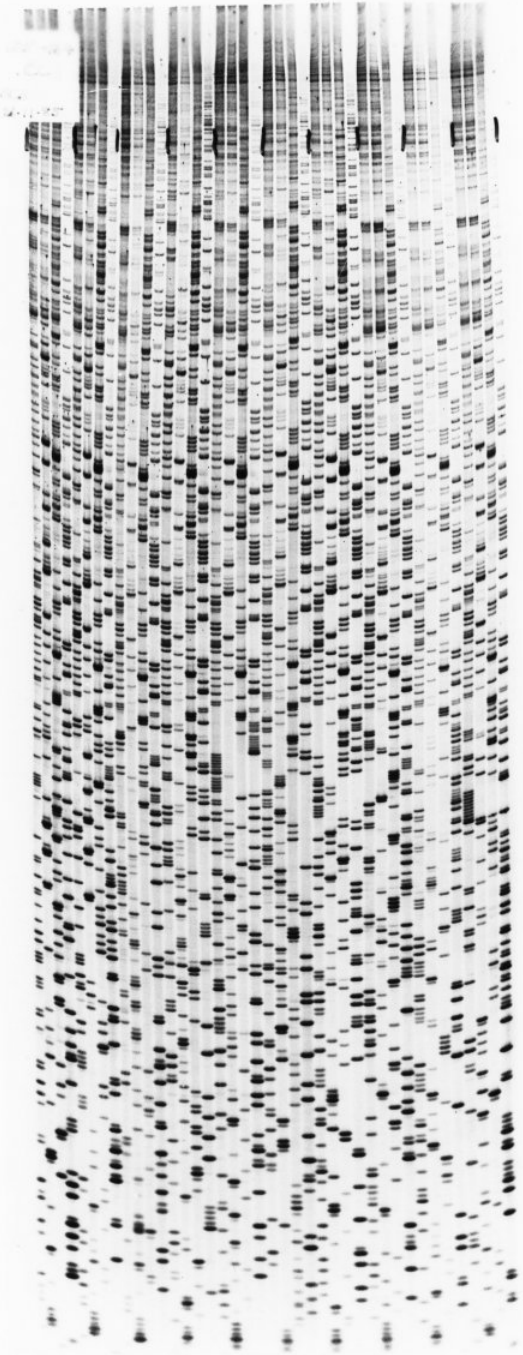


Norbert Grundmann

Tabea Kischka

Victoria Shabardina

1977



Elegance of the technical progress ;)

2017

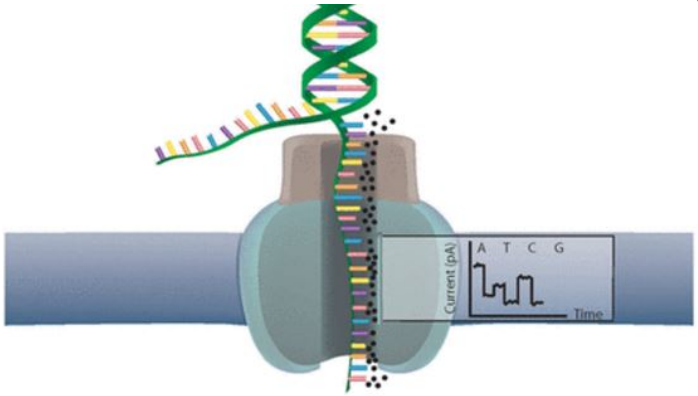
Electric signal → **Basecalling** → .fast5 file → – – –

Real-time MinKnow
Albacore basecalling

HDF viewer

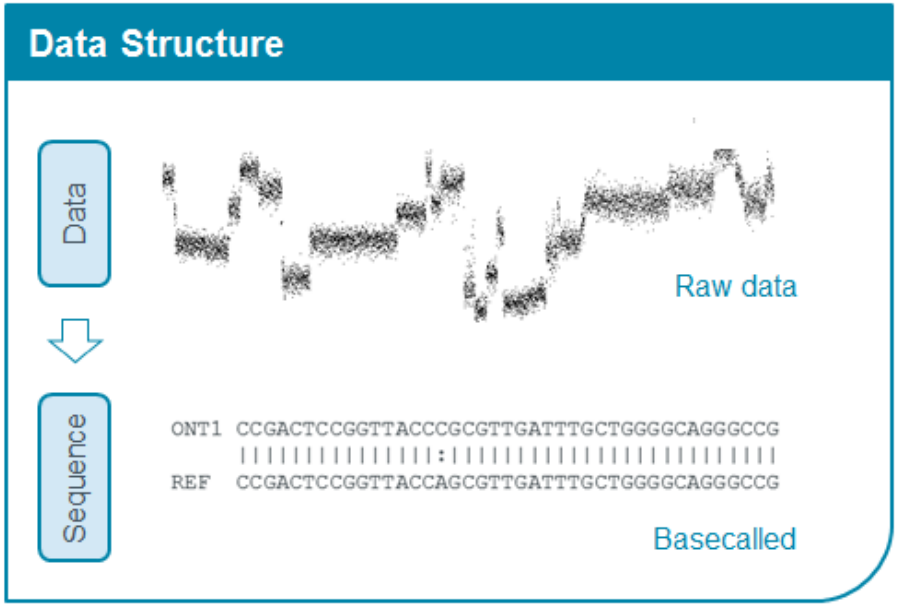
– – – – → .fastq file
.fasta file → Bioinformatics analysis

Basecalling for MinION sequencing reads



→ MinKNOW detects events that are measurements of 5 subsequent nucleotides →

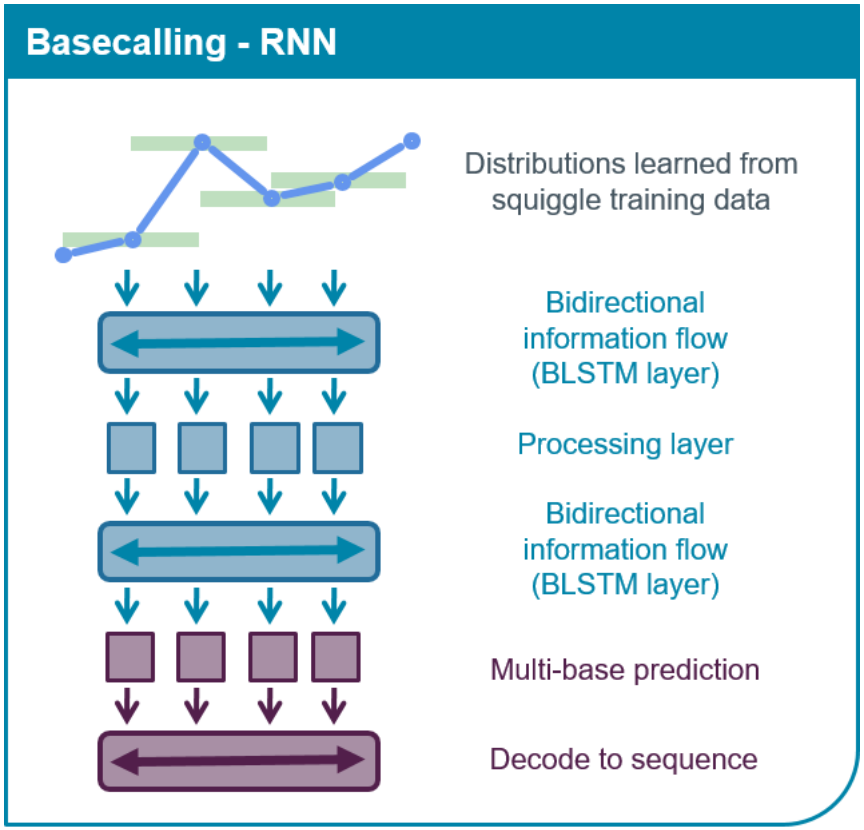
→ Albacore does basecalling:



*change in current
duration of the signal
noise signal

Basecalling for MinION sequencing reads

MinKNOW, Albacore use Recurrent Neural Network mechanism: it can learn and is trained on multiple data from MinIon sequencing



one .fast5/.fastq file per one read

.fast5 file format

.fast5 is a type of the Hierarchical Data Format (HDF) and designed for the storage of big datasets

It is binary – not readable by human

HDFview – the tool to see HGF files:

.fastq file format

Each entry consists of 4 lines:

@header
sequence
+
quality (coded by ASCII symbols: https://en.wikipedia.org/wiki/FASTQ_format#Encoding)

Poretools is the tool developed for MinIon to convert .fasta5 files to .fastq or .fasta files

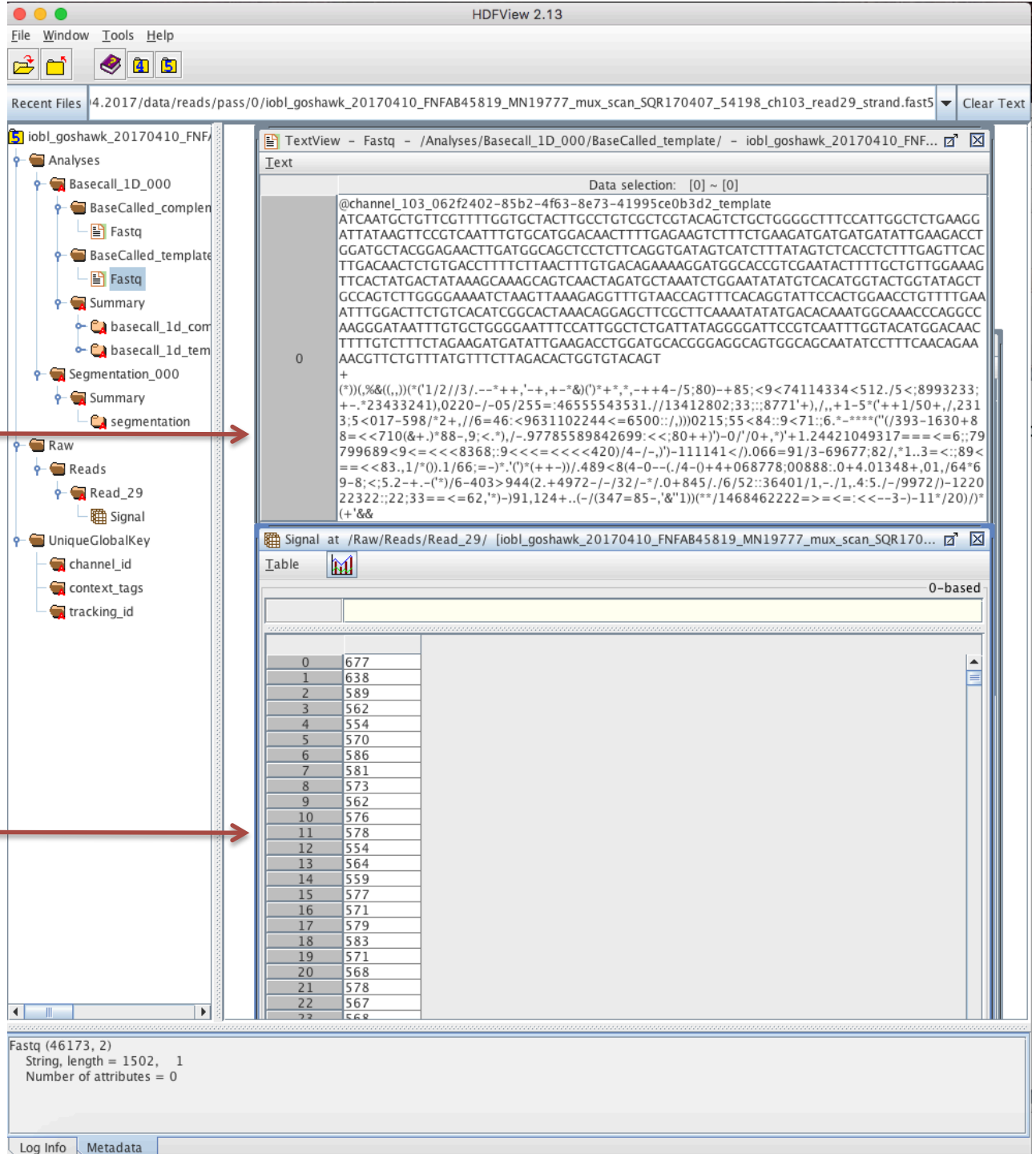
Recently Albacore basecaller has allowed to create directly .fast5 or .fastq files ☺

How HDFview looks like:

https://support.hdfgroup.org

.fastq file
(contains base quality info)

Raw signal
(current from a MinIon channel)



.fasta file format

contains 2 lines: >header and sequence

>ff6c98dd-bce9-4a2b-bf13-081841413c94_Basecall_2D_minion_20170511_Ae_aegypti
GCGCTGGTTCAGTTACATATTGCTAGGGTTAAGCAGTGGTGACCACAGATTTTTATGATTTATGGATT
CTTTTCTTCTGGCTACATTACTGGAACAGAGCCTGCTTCTCAACAGTGTTCTTATGAACGCTTCAGCTTA
GTATAAAGGC
>4c8a2487-0e13-41a6-ac7b-6cd2fbf5eb95_Basecall_2D_minion_20170511_Ae_aegypti
TACGCGGTGACAAAAACGTGCGTACCGGCAACCGCATGTTGAAACAGGAAAACGTACAAAGGACCC
TCGCAAAATGCGCGACAAAATCTGCAACGTACAACATGCGATAAACGTGCGTGAGGAGATC
>…..

> sing is convenient marker to brows through the .fasta files

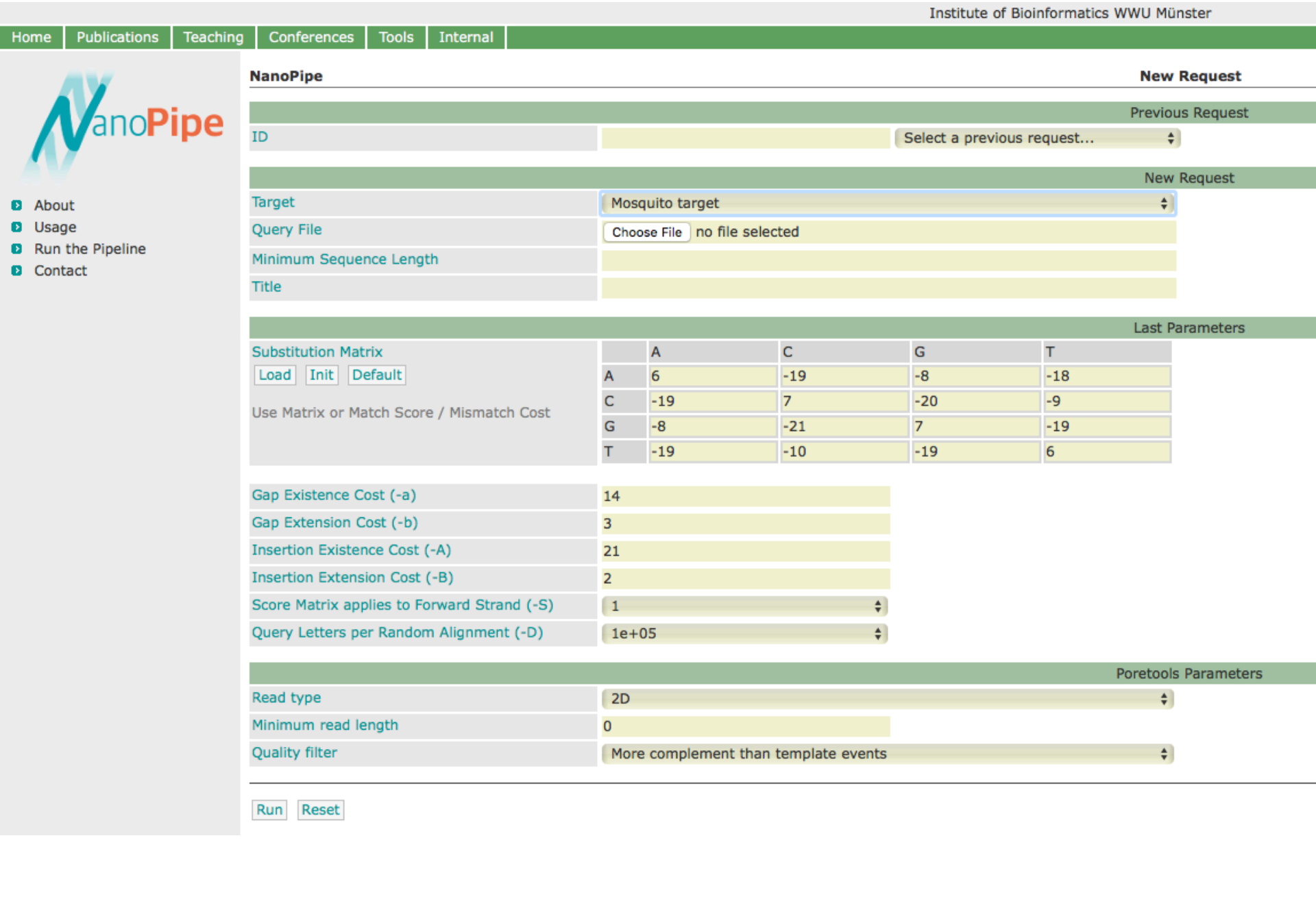
Nanopipe from Münster: why it is a pipe?

pyra.uni-muenster.de - PuTTY

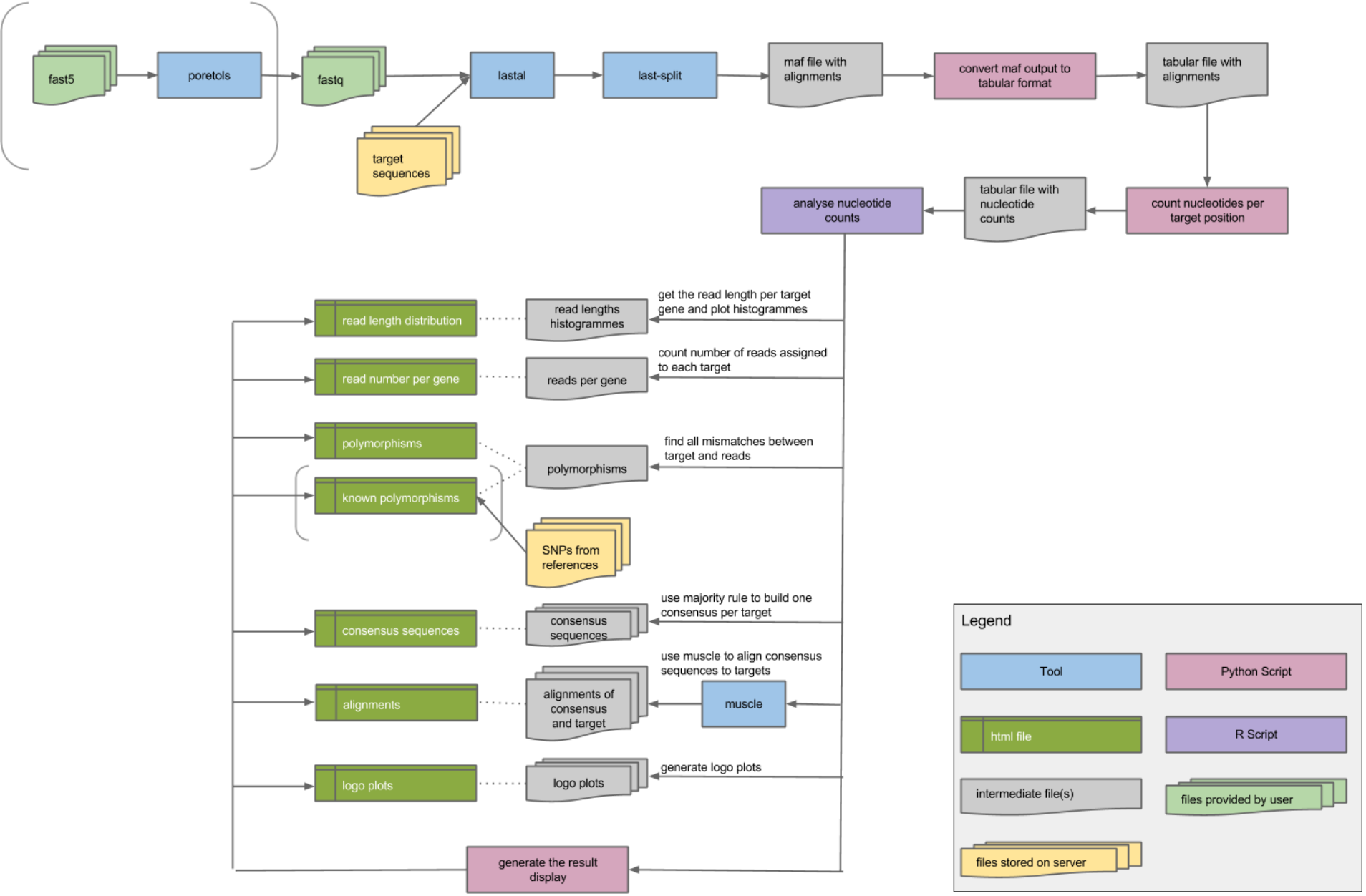
```
CCTGGTCAACGGCCAGCGCGACAGCGCGCTGAACTCGGCGCGAGTGGCAGGCCTCCGCGG
TGAGTTCACCGGACGCAGGCCGAAGTCGCGTGCCCGAGGTGCGGCCGGCCAGCGCGCAGC
CCGGCCAGCTGGTTCGAGCCGAGTTCAATCGCCAGGAAGCAGCCGGGCGCCTTCTCGCCT
GAGCAGCATGTCTGAAGACACCCGCGTACAGTTCGTCCTCGAGCTCCACTCCAGCGAACT
AAGCGCAGAATTTCTGTTGGTGCTGATGTGCTATACTGCGGCCGGGGTGACGAAGTCAGG
TTGATTCGGAGAACCTGAGCAGCTACCACATCAAGGAAGGTAGCAGACTTTAACATGAAG
ATAGAGCGACGGGCAAGTAGCAATGTGCCC
>channel_95_914f9b6b-d481-4a4e-8133-b8c3b3a94619_complement pass_april/11/iobl_goshawk_20170410_FNFAB45819_MN19777_se
TCGAGCATAAACAGACGCGCCATATCTTCATCAAACACGGTAGGGAAAGTGAGGTAGGAA
ATAACTACAAGTGGAAGCCTGACCGACAAATACTACCTCAACGGAAGATCAATAAATTGT
TATATCATATTGATGGTATATACTCTCAATACAATAGTGGCAATTCTGTGGTCTTTCTCC
CAAAACCCTTAACTAGTCTAATCGTAGTAAATATCAGAAAAATTCTCATAAATTTTAAAA
TACCTGATCAGTATTTGCAAAATGTCAAGTCATCAAAACTAAACTCTGGATATTGATGCA
GCCAAGGTTTGAAGATATATAGCCTAAGTAAAGTTTATTAAAGTTAAAGATGGGAAACCA
CAAAGCTAGAAAATATTAGGGTTTTATTTTATAATGGAGAACTTTCTAAGTGTATGACAG
CATATAGAAATCAAAGAAAAGAGTTTATTTTATTTTCTAAAATTAATAAACTTTTGGAGC
AGTTTAGGTTTGAAAAATTGAGTGGCTTCCGTATACTCTCTCACCCACCACTTCCTACTA
ATATGCGCATTGTATAGGCATTGTTACAGCAGATGTGCCAATTGATCCACTATTATTAAC
AAATTTCGGCCAGGCATGGTGTACCTATGTATCAACACTGTAGAGGCCGGAAAGTGGATC
ACGAGGTCAGGAAGTTCAAACCAGACTGGCAAGTGGTGAAACCCCGTACTAAAAATACAA
AATTAGCCGGCATAATGGCAGCAATATCAGCATAATGCCATGCT
>channel_105_3256ac6f-4f29-440f-8bec-a480c1b7a64f_complement pass_april/11/iobl_goshawk_20170410_FNFAB45819_MN19777_s
CTATCAAGCATATGGTAAATATTTGTCCAACCTGGTAGGAAGAATGAGGTCAGACTATGG
TCAGGGCTGCCCAGGAACGAACATGAGGTTGGGGTCCAGGCTGAGCCTGGAGTGAAGGGT
AGTGTGGATAGCAAGCATCGGATCTCTCTCTCTCTGCTGAAGAAAGGGCACTTGCTGAGG
TCTCATTAAACCCAGCTCAGTTATGCCTTTGATGTGTCTGTGTCTTTACGCCAGATTGGC
CTTCGGTTTGAACACGATGCAGATGACTGAACAGATGAGGAAGTTCACCTGCAGGGTGGG
GCCGAAGTTGGGGCACCAGCCCAGCCTATCATAGCTGTATACAGCCGTAATGGCAATATC
CTTTACCAACAAATTTTCAGTGGTGCTTTGATTTCCACATTAGTAGTGTCAAATGCTAAT
CGTGATAATTTTAATAGCATTGTGAGATCAGCTGATAAATTGATAAATTGCTTTCCTGGC
CATGTGTGTGTTGAAAAGGACCTGTTAAAATTGTTTAGTGAAGAATTTGAATCCAGTATC
TGTTAATATAGCTACATGAGTCTTCTCACATGAAGGCAGAAATATACTCTTCATCTATGA
TTTATTTCTTTACTTAAATAAAACATTAGTAAAAATGTATGCAGCCAGAGGCGTAATCCT
TGGCTGTCTTTTTAATATAGAAACTAAAGAAATTAACATGATGTTATTTCTTTTCTCTTT
CTAACTGAAGAATTAGTTAAAATAGCAACACTCTTGAGTTTTGACAGTTTTGGAATTCAT
GCTAGTTACTGATGAAACATTTTGTATTATTCCATATTTCATTTGCTATTGGAAACAAGT
AGTTTGTGTTGACTGTACCCCAGTCTGGCAAGATATAATAGGAGGTAGGAATTTGCCTGC
ATGGATACTTGTGTAGTACATGGGCTTTGATGACCTTCTTTATTTGCAATGTTTACAACT
TCCTCTGGGAAGTGTCCAAATATCTCCCATTCCTGTCCTCGGACCCTTCTTCTTCACGAG
ATTCCTTTCCCTACCGTGTTTGATGAAGATAAGTA
[vickas@pyra-work:KlausLabSequencing] > less readme_analysis_protocol.txt
author: vika
date: 20.04.2017
project: MinIon sequencing of human genes mutated in cancer
file content: decription of the analysis procedure

*pass anf fail folder each were around 4.5 GB; the sequencing protocol was run for 48 hrs; start: 480 pore were avail
flowcell.

1)poretools fastq --type fwd,rev /pass > pass_10.04.2017_fwdrev.fq
the same for the fail
2)Ayako Suzuki: "Using LAST (version 658) with parameters -a13, -A14, -b4 and -B3 which were determined by last-train
all of the 2D reads were mapped to chromosome 7 of the human genome UCSC hg38. The best alignments for each query wer

What I did:
lastdb hg38db /HG38/HG38.fasta
last-train hg38db pass_10.04.2017.fasta #the SDTOUT is saved inder "last-train_21.4.2017.txt" in the directory with t
lastal -a12 -A19 -b3 -B3 -S1 hg38db pass_10.04.2017_fwdrev.fasta | last-split > pass_10.04.2017_hg38_lasttrained.maf
maf-convert sam -d pass_10.04.2017_hg38_lasttrained.maf > pass_10.04.2017_hg38_lasttrained.sam
samtools view -bS -o pass_10.04.2017_hg38_lasttrained.bam pass_10.04.2017_hg38_lasttrained.sam
samtools sort -o pass_10.04.2017_hg38_lasttrained_sorted.bam pass_10.04.2017_hg38_lasttrained.bam
samtools index pass_10.04.2017_hg38_lasttrained_sorted.bam
-->go to IGVviewer
```

```
#
sub skipMinlen {
    return if (!-r "$INPUTPRAEFIX.minlen");
    my $minlen = readFile("$INPUTPRAEFIX.minlen");
    return if ($minlen < 1);

    my $type = readPraefix($QUERYFILE, 1) eq ">" ? 1 : 2;

    rename($QUERYFILE, "$QUERYFILE.tmp");

    open(IN,  "<", "$QUERYFILE.tmp");
    open(OUT, ">", $QUERYFILE);

    if ($type == 1) {
        my ($header, $seq);
        while (my $line = <IN>) {
            if ($line =~ m/^>/) {
                if ($seq) {
                    my $len = $seq =~ tr/A-Za-z//;
                    print OUT $header, $seq if ($len >= $minlen);
                }
                $header = $line;
                $seq    = "";
            }
            else {
                $seq .= $line;
            }
        }
        if ($seq) {
            my $len = $seq =~ tr/A-Za-z//;
            print OUT $header, $seq if ($len >= $minlen);
        }
    }
    else {
        my $index = 0;
        my $firstline;
        my $print = 0;
        while (my $line = <IN>) {
            if ($index == 0) {
                $firstline = $line;
            }
            elsif ($index == 1) {
                my $len = $line =~ tr/A-Za-z//;
                $print = $len >= $minlen;
                print OUT $firstline, $line if ($print);
            }
            else {
                print OUT $line if ($print);
            }
            $index = ($index + 1) % 4;
        }
    }

    close(OUT);
    close(IN);
}
```

Web interface



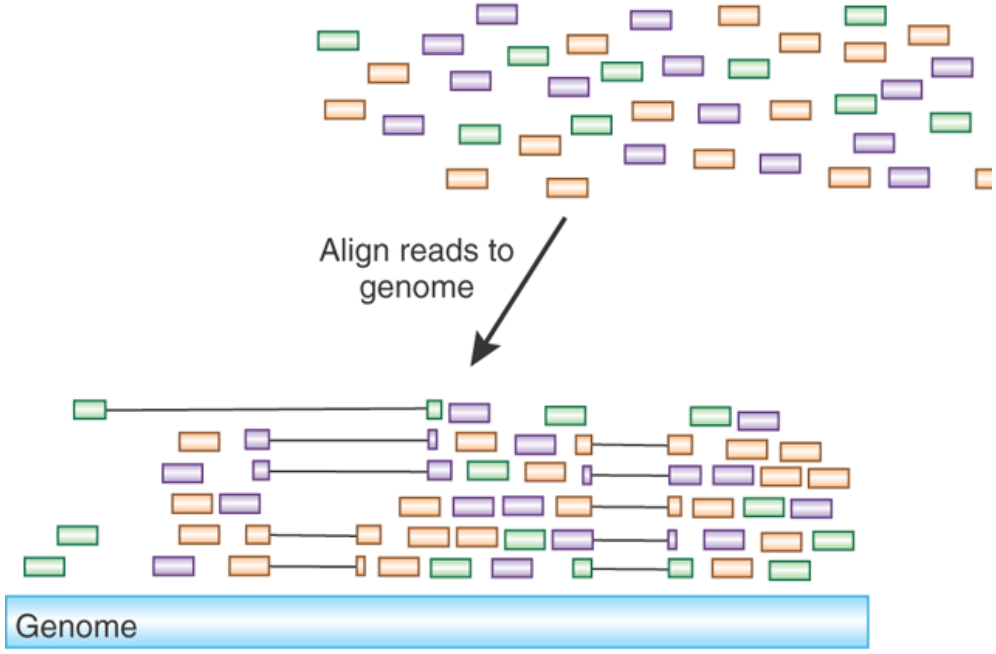
Nanopipe's workflow



Key steps of Nanopipe

1) LAST sequence aligner maps MinIon-produced reads to a target (selected region, exon, gene, genome)

Martin Frith
University of Tokyo,
Division of Biosciences



Other aligners: BLAST (psi-BLAST, delta-BLAST), HMMER, MUSCLE, MAFFT…

File formats that are important to know when working with mapped reads:

.maf
.sam
.bam

.maf file format – shows pairs of aligned sequences with the coordinates

```
# LAST version 828
#
# a=12 b=3 A=19 B=3 e=31 d=23 x=30 y=9 z=30 D=1e+06 E=725.241
# R=10 u=0 s=2 S=1 M=0 T=0 m=10 l=1 n=10 k=1 w=1000 t=0.910239 j=3 Q=0
# readsDB
# Reference sequences=340931 normal letters=689426062
# lambda=1.10061 K=0.337715
#
#    A  C  G  T
# A  1 -1 -1 -1
# C -1  1 -1 -1
# G -1 -1  1 -1
# T -1 -1 -1  1
#
# Coordinates are 0-based.  For - strand matches, coordinates
# in the reverse complement of the 2nd sequence are used.
#
# name start alnSize strand seqSize alignment
#                        Seq.name| Align.start|  Align.length| Seq.length
# batch 0
a score=1405 EG2=0 E=0
s 6e5acca3-80ce-43d6-ac9a-6f63dfe61404 4582 2456 + 7060 TAACGCCATTACCTACAAAAGCCCAGCGCGACAAAAATGCCAGAGAACTGAAGCTGGCGAACAAGC--AGT
GTCGATCAGTGCATTGAAGCCACCACCGCCTCCGGCGTGGATAATGCAACCTCCCCCCCCGACTTGAACTACGCTGAACGGGATTATTTCACCCTCAGAGAGAATTTGATCACTATGCAAAA-CAAC
TTT---ACGAATGTTTGGGTTTCTGTTTAACAACATTTTCTGCGCCGCCACAAATTTTTGGCTGCATCGACAGTTTTCTTCTGCCCAATTTGAAACGAAGAAATGATGAATTGATGTGGTTTCCTTT
GTAGAAAATTAAACAAACCCTAAACAATGAGTTGAAATTTCATATTGTG--TATTTATTAATGTATGTCAGGTGCGATGAATCGTCATTTGTATTCCCGGATTAACTATGTCCACAGCCCTGAAG--
TGTTCTGAATGCTCTCAGTAAATAGTAATGGGT-ATCAAAGGTATAGTAATATCTTTTATGTTCATGGATATTTGTAACCCATCGGAAAACTCCTGCTTTAAGATT-TTTTCCCTGTATTGCTGAAA
TACAGATAGTAAATATAATGTGAGACGTTGTGACGTTTTAGTTCAGAATAAAACAATTCACAGTCTAAATCTTTTCGCACTTGATCGAATATTTCTTTAAAAATGGCAACCTGAGCCATTGGTAAAA
CGGTCCTGCTGGCATTCTGGAGGGAAATACAACCGACAGATGTATGTAAGGCCAACGTGCTCAAATCTTCATACAGAAAGATTTGAAGTAATATTTTAACCGCTATAGATGAAGAGCAGAAAGCGCA
AGCATTCAGAGCAATTGTGAGGCAGCGTTGGTGAAGCACGATAATAATATGAAGGATTATT-CCTGGTGGTTGACTGATCACCATAACTTTGCTAATCATTCAAACTATTTGGTCTGTGACAGAGCC
TGTATGCTCTCTTTTCTGACGTTGGTCTCCGACGGCAGGCT--AATGACCCAGGCTGAGAAATTCGGACCCTTTTTTTGCTCAAGAGCGATGTTAATTTGTTCAATCATTTGGTTTAGGAAAGCGGA
ATTATTTGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGATTGATAATCATTATCACTTTACGGGTCCTTTCCGGTGAAAAAAAGGTACCAAAAAACATCGTC
s DNA_CS                                1072 2465 + 3560 TAACGCCATTACCTACAAA-GCCCAGCGCGACAAAAATGCCAGAGAACTGAAGCTGGCGAACGCGGCAATT
TGTCAGTCAGTGCGTGAAGCCACCACCGCCTCCGGCGTGGATAATGCAGCCTCCCCCCGACTGGCAGACACCGCTGAACGGGATTATTTCACCCTCAGAGAGAGGCTGATCACTATGCAAAAACAAC
TTACGAATGTTTGCTGGGTTTCTGTTTTAACAACATTTTCTGCGCCGCCACAAATTTT-GGCTGCATCGACAGTTTTCTTCTGCCCAATTCCAGA--AACGAAGAAATGATGGGTGATGGTTTCCTT
GTAGAAAATTAAACAAACCCTAAACAATGAGTTGAAATTTCATATTGTTAATATTTATTAATGTATGTCAGGTGCGATGAATCGTCATT-GTATTCCCGGATTAACTATGTCCACAGCCCTGACGGG
TGTTCTGAATGCTCTCAGTAAATAGTAATGAATTATCAAAGGTATAGTAATATCTTTTATGTTCATGGATATTTGTAACCCATCGGAAAACTCCTGCTTTAGCAAGATTTTCCCTGTATTGCTGAAA
GCTAGATAGTAAATATAATGTGAGACGTTGTGACGTTTTAGTTCAGAATAAAACAATTCACAGTCTAAATCTTTTCGCACTTGATCGAATATTTCTTTAAAAATGGCAACCTGAGCCATTGGTAAAA
CGGTCCTGCTGGCATTCTGGAGGGAAATACAACCGACAGATGTATGTAAGGCCAACGTGCTCAAATCTTCATACAGAAAGATTTGAAGTAATATTTTAACCGCTA-----GATGAAGAGCAAGCGCA
AGCATTCAGAGCAAT--TGAGGCAGCGTTGGTGAAGCACGATAATAATATGAAGGATTATTCCCTGGTGGTTGACTGATCACCATAAC--TGCTAATCATTCAAACTATTTAGTCTGTGACAGAGCC
TATATGCTCTCTTTTCTGACGTTAGTCTCCGACGGCAGGCTTCAATGACCCAGGCTGAGAAATTCCCGGACCCTTTTTGCTCAAGAGCGATGTTAATTTGTTCAATCATTTGG-TTAGGAAAGCGGA
ATTATTTGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGA-TGATAATCATTATCACTTTACGGGTCCTTTCCGGTGAAAAAAAAGGTACCAAAAAAAACATC
```

.bam and .sam file format – informative about alignment of sequences to a target

.bam is a binary file (humans can not read), it contains a lot of information about the alignment. Important part is the HEADER and FLAGs specifications (https://broadinstitute.github.io/picard/explain-flags.html)

.sam format is the readable version of .bam (some info is casted away):

```
HWI-ST208:1:1208:20889:115208#0 145     1       9995    37      5M1I45M 3       184392705       0       GTATGAATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA     ggc]ea^^e^eba_^Oca^YHYdfhhdggffbedhgcab`^cgcccee___     XT:A:U  NM:i:3  XN:i:6  SM:i:37 AM:i:37 X
0:i:1  X1:i:XM:i:2   XO:i:1  XG:i:1  MD:Z:2C0C46     RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1102:11436:179038#0 163     1       9998    0       51M     =       10235   287     CTAAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC     _[_ccdacgeaggffhhhfhfaefeegfhiihihgfdfh_efg^cfhXbZa     XT:A:R  NM:i:2  XN:i:3  SM:i:0  AM:i:0  X0:i:2  X
1:i:7  XM:i:XO:i:0   XG:i:0  MD:Z:1G1T47     RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2102:2266:180517#0  99      1       9998    0       51M     =       10181   234     GGTTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC     bbbeeeceggggfghhiiiiiiihiiiiihiiiiihiihhiiiiiiiiiii     XT:A:R  NM:i:2  XN:i:3  SM:i:0  AM:i:0  X0:i:6  X
1:i:1  XM:i:XO:i:0   XG:i:0  MD:Z:0C1A48     RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1106:12187:25220#0  97      1       10000   23      51M     6       146315262       0       ATAACCCTAACCCTAACCCTAACCCTAACACTAACCCCAACCCTAACCCTA     PPPQQ^QJ[JQ``JQQQQ^J[`R[JQ^Y`IPIPPYHOHO\Z```BBBBBBB     XT:A:U  NM:i:2  XN:i:1  SM:i:23 AM:i:23 X
0:i:1  X1:i:XM:i:2   XO:i:0  XG:i:0  MD:Z:29C7T13    XA:Z:4,-191043864,51M,3;        RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1107:15842:36208#0  73      1       10000   0       51M     =       10000   0       ATAACCCTAACCCTAACCCTAACCTTAACACTAACCCTAAGCCTAACCCTA     YYZa`cacgcge`dZ^d`J```Jbgdgff$^I^cb`H^a^IORX^ccBBBB     XT:A:R  NM:i:3  XN:i:1  SM:i:0  AM:i:0  X0:i:2  X
1:i:0  XM:i:XO:i:0   XG:i:0  MD:Z:24C4C10C10 XA:Z:12,-95599,51M,3;   RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1108:1899:167769#0  73      1       10000   0       51M     =       10000   0       ATAACCCTAAACCTAACCCTAACCCTAACCCTAACCCTAACCTTAACCCTA     BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB     XT:A:R  NM:i:2  XN:i:1  SM:i:0  AM:i:0  X0:i:2  X
1:i:461     XM:i:2   XO:i:0  XG:i:0  MD:Z:10C31C8    RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1202:12536:33295#0  97      1       10000   0       51M     15      27551735        0       ATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTCACCCTAACCCTA     PJPSSQ`^ccccJ[``d_bddbcdeabddZccccbcH^H^WWacaXcBBBB     XT:A:R  NM:i:1  XN:i:1  SM:i:0  AM:i:0  X
0:i:2  X1:i:459      XM:i:1  XO:i:0  XG:i:0  MD:Z:38A12      RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2101:12428:178559#0 99      1       10000   9       51M     =       10172   223     ACAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAACCCTA     ^^^ccZWJ[`^J[JQQ[YRbbeeh`H^bWcRbBBBBBBBBBBBBBBBBBB     XT:A:U  NM:i:2  XN:i:1  SM:i:0  AM:i:0  X0:i:1  X
1:i:8  XM:i:XO:i:0   XG:i:0  MD:Z:1T41T7     RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2305:4612:97642#0   73      1       10000   0       51M     =       10000   0       ATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACACTAACCCTA     J\_cccc`Zaeaehehabefde^d`Zbf_RcZce^eRYaaHIXIX^cWa\b     XT:A:U  NM:i:1  XN:i:1  SM:i:0  AM:i:0  X0:i:1  X
1:i:461     XM:i:1   XO:i:0  XG:i:0  MD:Z:41C9       RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1106:9058:143709#0  99      1       10001   0       21M1I27M2S      =       10064   114     TAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCCAACCCCCACCCCA     bbbeeeeefgggghiiiiihhdghiiiiiiiicfhiii`g`fffFaFadgB     XT:A:M  NM:i:4  SM:i:0  AM:i:0  XM:i:3  X
O:i:1  XG:i:MD:Z:36T5T0A4    RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1208:19850:40560#0  99      1       10001   0       18M1I32M        =       10088   138     TAACCCTAACCCTAACCCTTAACCCTAGCCTTAGCCCTAGCCCTAGCCCTA     bb_eeeeeggfgiiiiiiiiiiiibgffhiiihhhhhadfgdgiiiiihh     XT:A:M  NM:i:6  SM:i:0  AM:i:0  XM:i:5  X
O:i:1  XG:i:MD:Z:26A2C2A5A5A5       RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1302:8776:114457#0  99      1       10001   29      7M1I43M =       10297   347     TAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA     _^_ccccaegZbcdaedefghgZ``Zefffgaggddg`cc_eeZeeeggaZ     XT:A:M  NM:i:1  SM:i:29 AM:i:29 XM:i:0  XO:i:1  X
G:i:1  MD:Z:50       XA:Z:7,-159128598,51M,0;15,-102521324,51M,0;18,+10154,51M,1;4,+10053,51M,1;1,-249240218,51M,1;1,-249240562,51M,1;4,-191044006,51M,1;10,-135524471,51M,1;      RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2206:8595:101855#0  163     1       10001   17      39M1I9M2S       =       10101   151     TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCA     bbbeeeeefggfghiiiiihiiiidgiihffhhhihgihifdUZPWbe^dT     XT:A:M  NM:i:1  SM:i:17 AM:i:17 XM:i:0  X
O:i:1  XG:i:MD:Z:4B  RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1105:12211:11357#0  163     1       10002   9       35M1I6M1I8M     =       10101   150     AACCCTAACCCTAACCCTAACCCTAATCCTAACCCTTAACCCTTAACCATA     b___eeeeefegggiihhhfgfhffihhgehhh_cgdfedbghhihiff]Ra     XT:A:M  NM:i:4  SM:i:9  AM:i:9  XM:i:2  X
O:i:2  XG:i:MD:Z:26C19C2     XA:Z:18,+10106,35M1I15M,3;4,-191044042,14M1I36M,3;4,+10018,18M1I32M,3;  RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2207:15092:146978#0 137     1       10002   0       51M     =       10002   0       AACCCTAACCCTAAACCCAACCCTAACCCTAACCCTAACCCTAACACTAAC     a_aPSa`Q\`JJJ[JQQJR`JP[JY^HYRH^aH^RYaegH^XWR^IMH\aa     XT:A:R  NM:i:3  SM:i:0  AM:i:0  X0:i:463        X
M:i:3  XO:i:XG:i:0   MD:Z:14C2T27C5  RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1303:8901:125774#0  99      1       10003   17      51M     =       10176   224     ACCCTAACCCTAACCCTAACCCTTTCCCTAATCCTAATCCTAACCCTAACCCTAACC     BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB     XT:A:M  NM:i:4  SM:i:17 AM:i:17 XM:i:4  XO:i:0  X
G:i:0  MD:Z:17A0A6C5C19      RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2101:15111:55510#0  163     1       10003   12      29M1I6M1I14M    =       10073   121     ACCCTAACCCTAACCCTAACCCTAACCCTAAACCCTAAAAACTAACCCTAA     aa_aaa^ccgefcghii]^fag`[JJbWHbX^cegRO^eBBBBBBBBBBB     XT:A:M  NM:i:4  SM:i:12 AM:i:12 XM:i:2  X
O:i:2  XG:i:MD:Z:37C0C10     RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:2305:4599:116299#0  163     1       10003   5       46M5S   =       10137   185     ACCCTCACCCTCACCCTCACCCTCACCCTAACCGTAACCCCAACCCCCACC     ^^\acZYQQ`^a^cY[`b]fffbffd__dcffhffgU^a^UUVbd_BBBBB     XT:A:M  NM:i:6  SM:i:5  AM:i:5  XM:i:6  XO:i:0  X
G:i:0  MD:Z:5A5A5A5A9C6T5    RG:Z:120309_SN208_0274_AD0R6DACXX       SM:Z:TCGA-49-4488-01A-01D-1751
HWI-ST208:1:1303:10882:25078#0  163     1       10004   17      15S9M1D11M16S   =       10103   150     AGCCCAGCCCCGCCCCCCAAACCCAACCCCACCCCCCTCCCCCCCCCCAAC     BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB     XT:A:M  NM:i:4  SM:i:17 AM:i:17 XM:i:3  X
```

Summary: files' formats

Files' formats used for storing sequences, sequencing results:

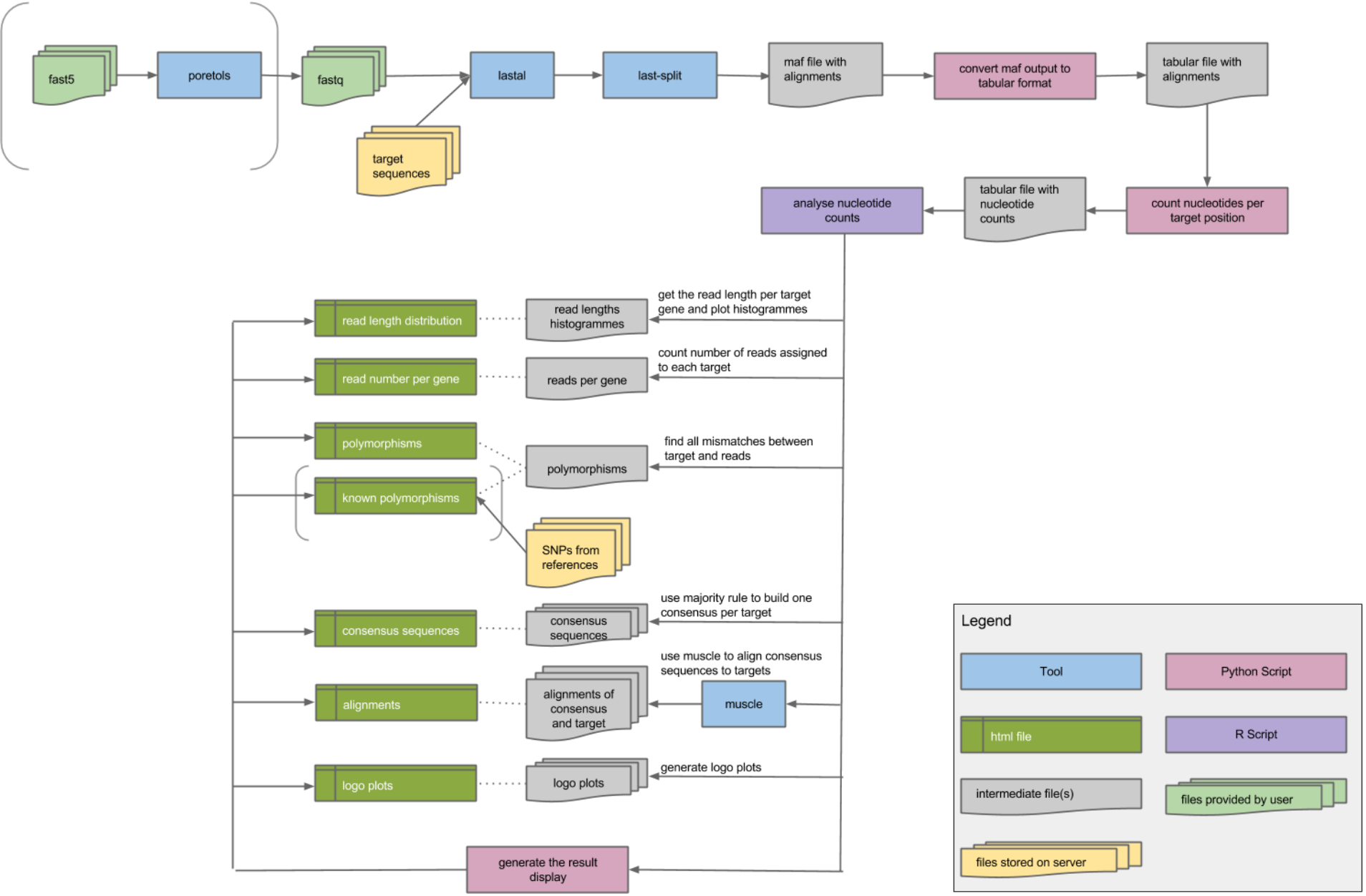
.fast5 – big, binary (cant read), contain a lot of metadata
.fastq – readable by human, contains sequences and sequence quality
.fasta – readable, contains sequences

Files' formats used for storing results of sequence alignment:

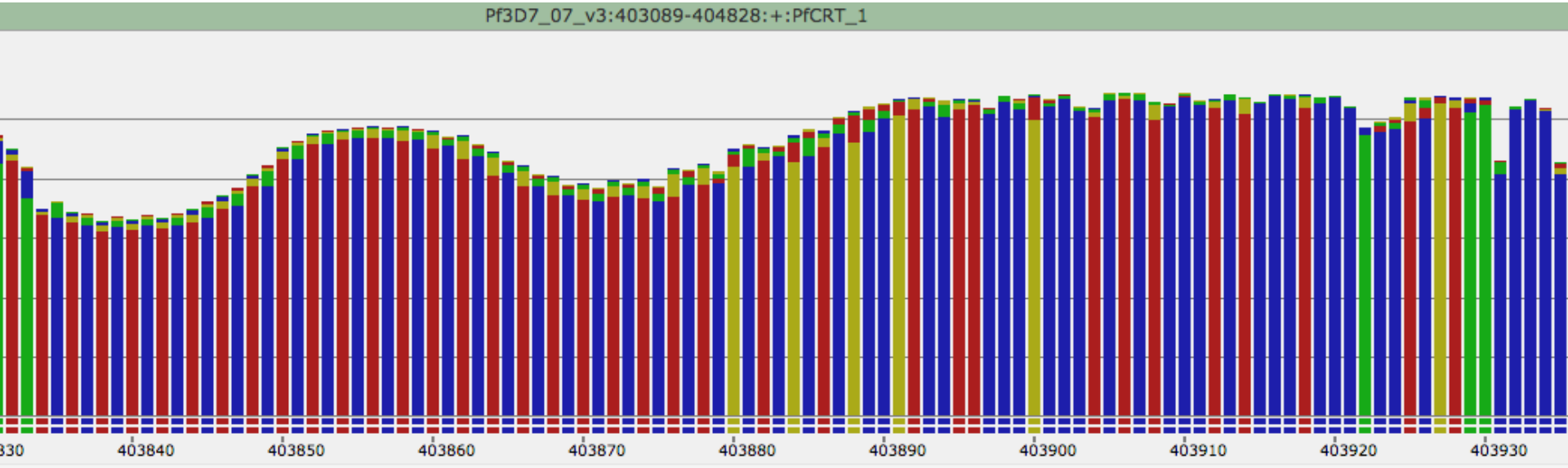
.maf – contains pairs of aligned sequences with the alignment's coordinates; for example, used by LAST
.bam – binary format, includes aligned sequences, coordinates, information about bioinformatics processing, quality, … Was created as a compact version of .bam file to safe space
.sam – human readable version of .bam, much bigger in size

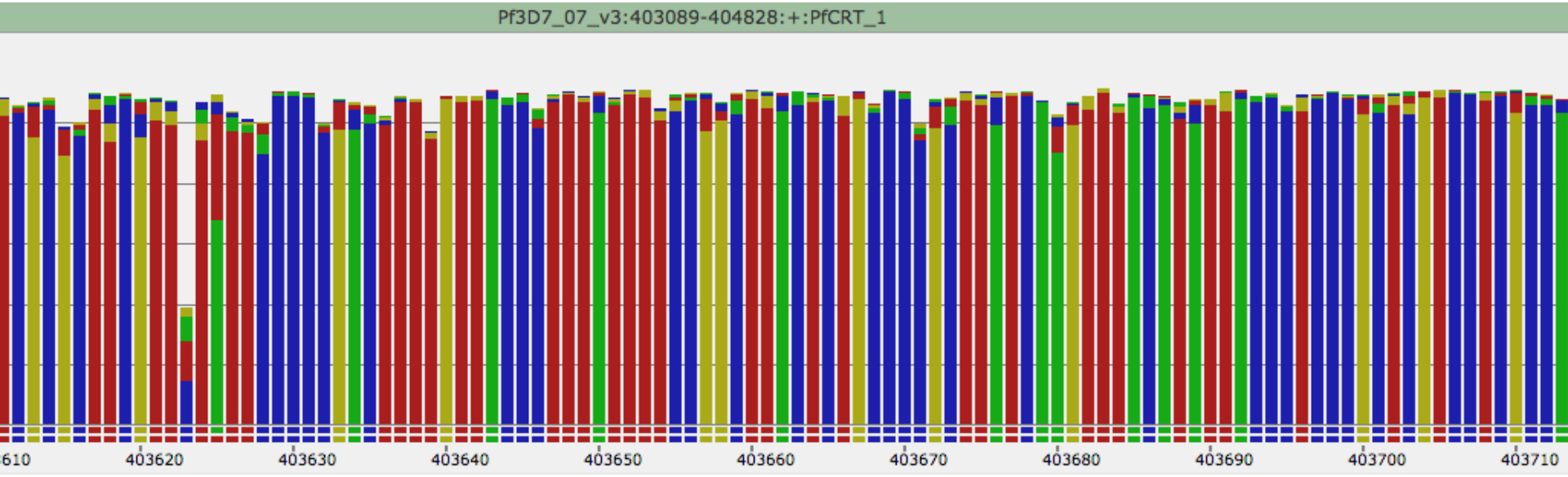
.fastq/.fasta and .bam files are widely used in all DNA/RNA bioinformatics analysis

Nanopipe's workflow



What is Consensus Sequence?





Key steps of Nanopipe

2) Tracing mutations (nucleotide frequency at a particular position) – majority rule, i.e. If alternate nucleotide is >=50% (?)

3) Generating consensus sequence and comparing it to the target sequence – MUSCLE aligner (using IUPAC symbols in ambiguous situation, for example, “R” means “A or G”, “Y” – for “C or T” (see http://www.bioinformatics.org/sms/iupac.html or on Nanopipe help-page: http://bioinformatics.uni-muenster.de/tools/nanopipe/usage/)

Nanopipe helps us to…

- See if our sequencing worked
- Detect insertions/deletions and nucleotide variations
- Visualization of the experiment
- We have the consensus sequence of our sample

Key steps of Nanopipe

2) Tracing mutations (nucleotide frequency at a particular position) – majority rule, i.e. If alternate nucleotide is >=50% (?)

3) Generating consensus sequence and comparing it to the target sequence – MUSCLE aligner (using IUPAC symbols in ambiguous situation, for example, “R” means “A or G”, “Y” – for “C or T” (see http://www.bioinformatics.org/sms/iupac.html)

Nanopipe helps us to…

- See if our sequencing worked
- Detect insertions/deletions and nucleotide variations
- Visualization of the experiment
- We have the consensus sequence of our sample

We can see how our reads map to the target

IGV-viewer software uses .bam files and .fasta file of the target sequence
IGV-viewer is easy to install and use

Let's install IGVview and SEE the results of our MinIon sequencing ☺

1) Go to https://software.broadinstitute.org/software/igv/download

2) Click "Download Windows Package"

3) "unzip" the downloaded folder then open it, open next folder and click "igv" icon.

4) Wait 1 minute, do nothing ☺

5) Voila

*The link to the .fasta files with the target sequences:
http://bioinformatics.uni-muenster.de/share/Bangkok_2017

Tools from Oxford Nanopore

Metagenomics analysis: identifying different taxa in mixed samples (soil, food, human mucus,…)

EPI2ME
Bioinformatics cluster
(amazon cloud based)

→ WIMP ("what is in my pot") : prokaryotes, fungi, archaea

→ S16 : eukaryotes, prokaryotes, fungi, archaea

Find the PDF with the presentations here:

http://www.compgen.uni-muenster.de/home/presentations/

Practical session

Go to http://bioinformatics.uni-muenster.de/tools/nanopipe/

and let it begin

Reads quality

*a hint: when you finally obtain your sequencing reads it is important to know their quality. One of the tools (simple and good ☺) is FASTQC.

Download FASTQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/