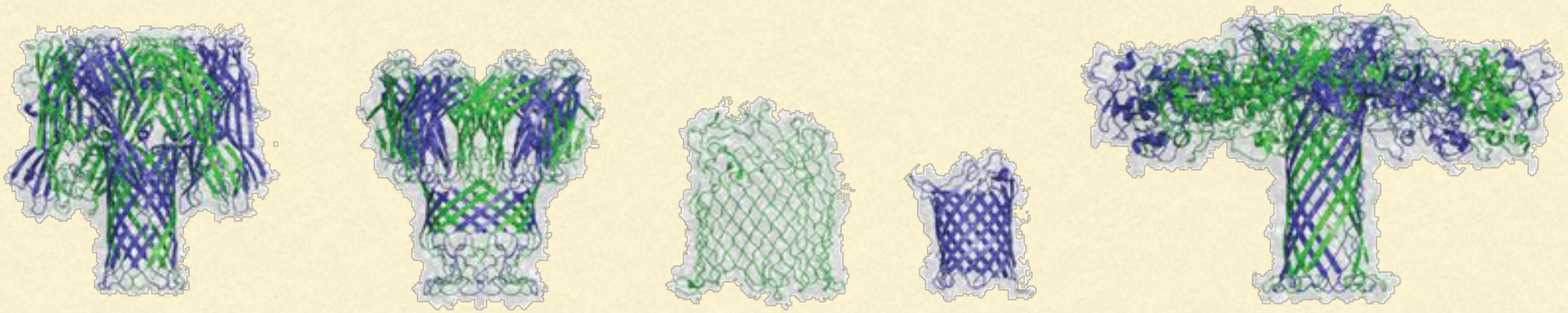
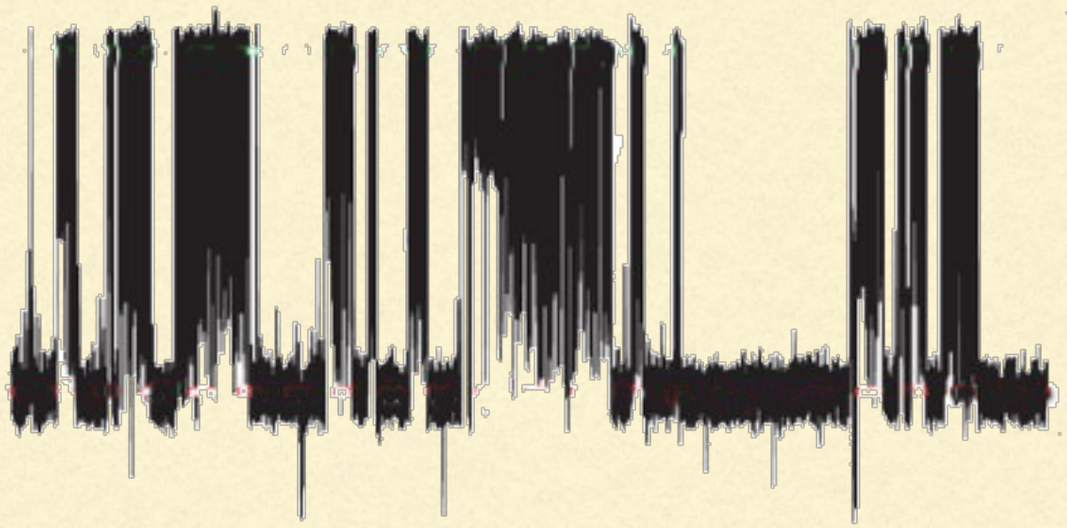


NANOPORE SEQUENCING BIOINFORMATICS RESOURCES

Wojciech Makałowski
Institute of Bioinformatics, University of Muenster, Germany
Department of Computational Biology, University of Tokyo, Japan

http://bioinformatics.uni-muenster.de

TWO TYPE OF COMPUTATIONAL CHALLENGES



```
CTGGGAAACTGCCTGATGGAGGGGGATAACT
||||||||||||||||||||||||||||||
CTGGGAAACTGCCTGATGGAGGGGGATAACT
```

- Turning electric signal into a DNA sequence
- Applying long reads to answer biological questions

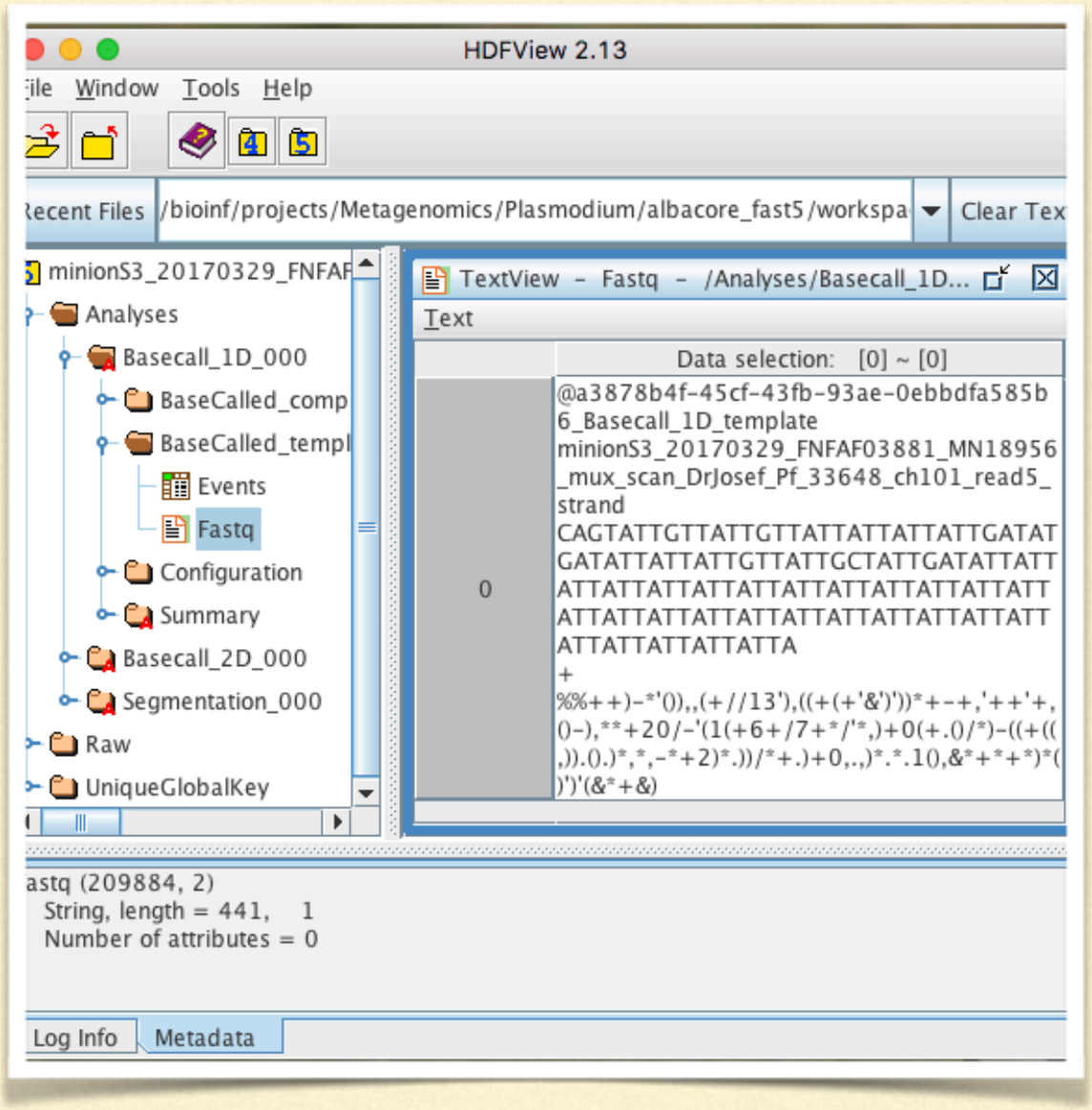
BASE CALLING

Software	Developer	Notes
Albacore	ONT	Stand alone implementation of base caller used in MinKNOW
DeepNano	Boza et al. https://arxiv.org/abs/1603.09195	Recurrent Neural Networks
Nanocall	David et al. Bioinformatics 2017; 33:49-55	HMM approach. Doesn't allow 2D integration.

MinION DATA FORMATS AND HANDLING

- FAST5
- FASTQ

MinION DATA FORMATS FAST5



- The raw data is stored as binary files in HDF5 standard
- HDFView allows quick look at the raw data files
- https://support.hdfgroup.org/products/java/release/download.html#bin

MinION DATA FORMATS
FASTQ

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

MinION DATA FORMATS FASTQ

$$Q = -10 \log_{10} p$$

p = probability that the corresponding base call is incorrect

ASCII	p	Q
!	1	0
)	0.1	10
3	0.01	20
=	0.001	30
H	0.0001	40
~		93

!”#$%&'() *+,-./0123 456789:;<= >?@ABCDEFGH I

MinION DATA FORMATS
FASTA

Very simple format but it may contain quite a bit in formation on the sequence.
Used by many software including BLAST and NanoPipe

```
>SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
```

DATA HANDLING TOOLS

Software	Output format	Availability
HPG Pore	FASTA FASTQ	https://github.com/opencb/hpg-pore
minoTour	a real time analysis of minION reads	http://minotour.nottingham.ac.uk/index.php
NanoOK	FASTA FASTQ	https://github.com/TGAC/NanoOK
npReader	real time FASTA FASTQ	https://github.com/mdcao/npReader
R_poRe	FASTA FASTQ	https://sourceforge.net/projects/rpore/
PoreTools	FASTA FASTQ	https://github.com/arq5x/poretools
seqtk	FASTA	https://github.com/lh3/seqtk

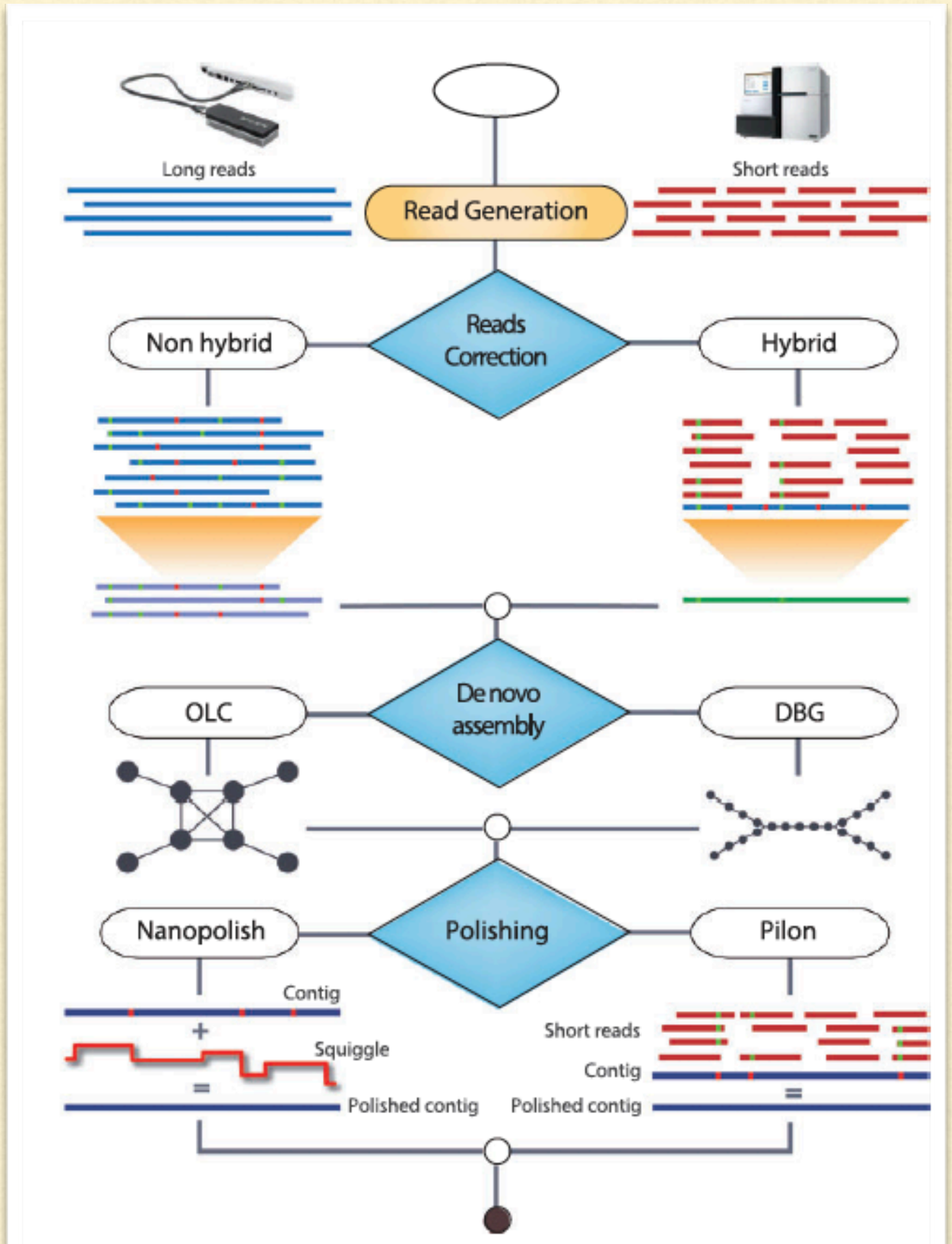
SEQUENCE ANALYSES

- Mapping and resequencing
- *De novo* assembly
- Variant discovery
- Metagenomics

MAPPING AND RESEQUENCING

Software	Input format	Availability
BLASR	FASTA FASTQ	https://github.com/PacificBiosciences/blasr
BWA	FASTA FASTQ	http://bio-bwa.sourceforge.net
GraphMap	FASTA FASTQ	https://github.com/isovic/graphmap
LAST	FASTA FASTQ	http://last.cbrc.jp
marginAlign	BAM	https://github.com/benedictpaten/marginAlign
NanoPipe	FAST5, FASTA, FASTQ	http://bioinformatics.uni-muenster.de/tools/nanopipe/

DE NOVO ASSEMBLY



Magi et al. Briefings in Bioinformatics, 2017, 1–17

DE NOVO ASSEMBLY
error correction (e) and polishing (p)

Software	Alghoritm	Task	Availability
Nanocorr	Hybrid	e	https://github.com/jgurtowski/nanocorr
NaS	Hybrid	e	https://github.com/institut-de-genomique/NaS
Nanocorrect	Non-hybrid	e	https://github.com/jts/nanocorrect
PoreSeq	Non-hybrid	e/p	https://github.com/tszalay/poreseq
Nanopolish	Non-hybrid	p	https://github.com/jts/nanopolish

Magi et al. Briefings in Bioinformatics, 2017, 1–17

DE NOVO ASSEMBLY PIPELINES

Software	Alghoritm	Availability
ABruijn	Non-hybrid DBG	https://github.com/fenderglass/ABruijn
ALLPATHS-LG	Hybrid DBG	https://software.broadinstitute.org/allpaths-lg/blog/
Canu	Non-hybrid OLC	https://github.com/marbl/canu
Falcon	Non-hybrid OLC	https://github.com/PacificBiosciences/FALCON
LQS	OLC-Celera with corrections	https://github.com/jts/nanopore-paper-analysis
MaSuRCA	Hybrid with super-reads	http://masurca.blogspot.com
Miniasm	OLC without corrections	https://github.com/lh3/miniasm
SAMRTdenovo	OLC without corrections	https://github.com/ruanjue/smartdenovo
SPAdes	Hybrid DBG	http://cab.spbu.ru/software/spades/

Magi et al. Briefings in Bioinformatics, 2017, 1–17

VARIANT DISCOVERY

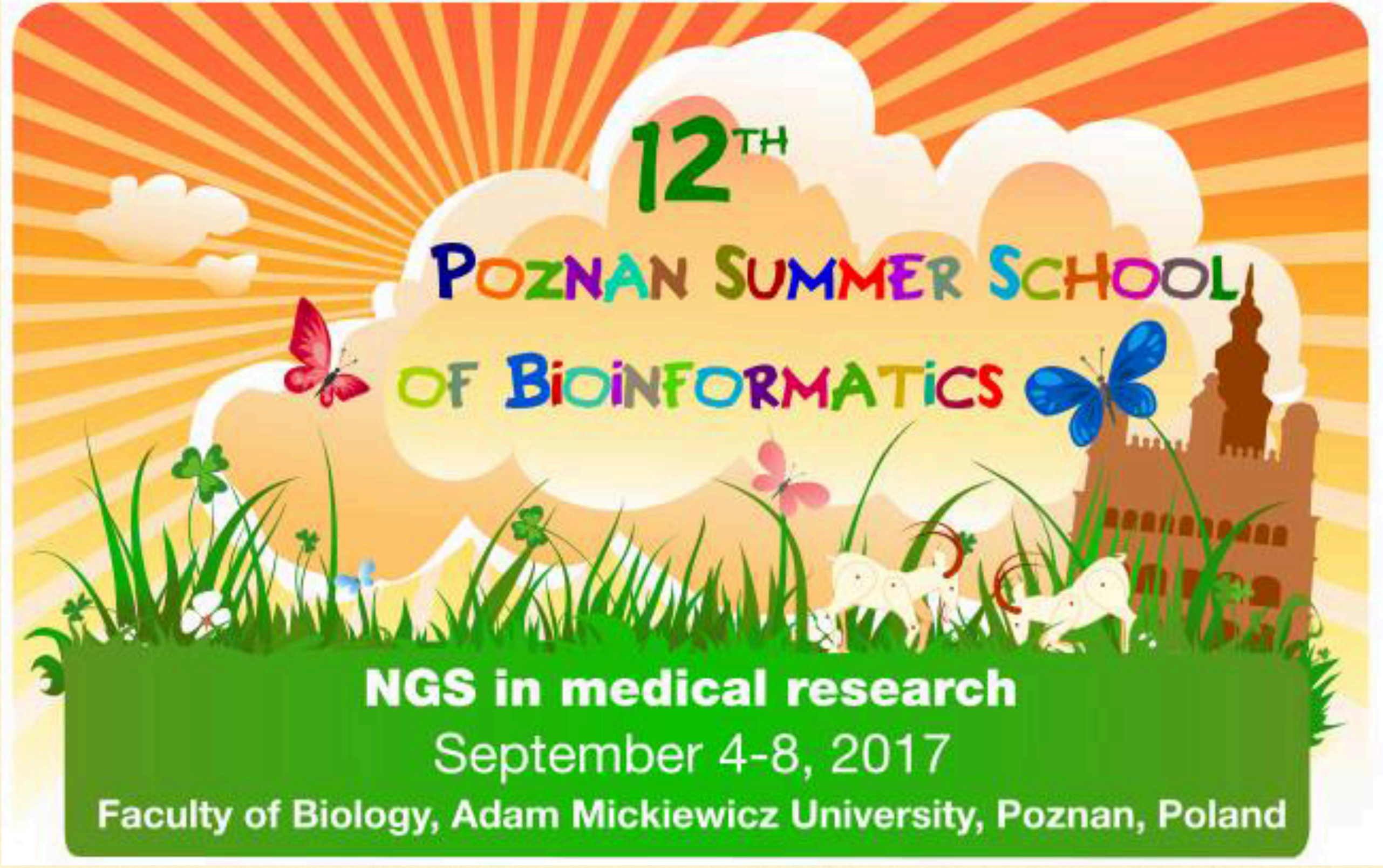
Software	Input format	Availability
marginCaller	BAM	https://github.com/benedictpaten/marginAlign
NanoPipe	FAST5, FASTA, FASTQ	http://bioinformatics.uni-muenster.de/tools/nanopipe/index.hbi?
Nanopolish	BAM FAST5	https://github.com/jts/nanopolish
Nanosv	BAM	https://github.com/mroosmalen/nanosv

METAGENOMICS

Software	Input format	Availability
EPI2ME (16S and WIMP)	FAST5	https://epi2me.nanoporetech.com/workflow
NanoPipe	FAST5, FASTA, FASTQ	http://bioinformatics.uni-muenster.de/tools/nanopipe/
Centrifuge	FASTA	https://ccb.jhu.edu/software/centrifuge/

NANOPORE_TOOLS

https://docs.google.com/spreadsheets/d/15LWXg0mUeNOHVthl8JRX-FzJ9w8jrWogS4YhDcxyAfI/pubhtml?gid=0&single=true



http://bioinformatics-school.pl