Microbiome analysis using amplicon sequencing on Illumina platform

Ing. Juraj Medo PhD., juraj.medo@uniag.sk
Department of Microbiology
Faculty of biotechnology and food science
Slovak University of Agriculture in Nitra

Metagenomics

- Study of genetic material obtained directly from environmental samples
- Shotgun – eDNA is fragmented and sequenced In data analysis we can found „**who is there“** and „**what it is able to do“** presence of genes.

Hard bioinformatic task | Very expensive | Complex information

- Amplicons – using eDNA, only single gene (identification marker) is amplified using PCR
- **We can find only „who is there“**

Less expensive | Easy to do | Only taxonomic information

Metatranscriptomics

- Study RNA from environmental samples
- Shotgun – eRNA is fragmented and sequenced In data analysis we can found „**who is active there“** and „**what they are doing“**
- Amplicons – using eRNA, only single gene is amplified using PCR
- **We can find only „who is active there“**

Amplicon sequencing

- Starting material is environmental DNA
- 16S rRNA gene (bacteria) or ITS (fungi) and other
- Sequencing of PCR products
- Uses high throughput sequencers (Roche, **Illumina**, Ion-torrent, PacBio?)
- Acquire thousands of sequences from each sample
- Each sequence can be identified
- Result is a spectrum of microorganisms in sample

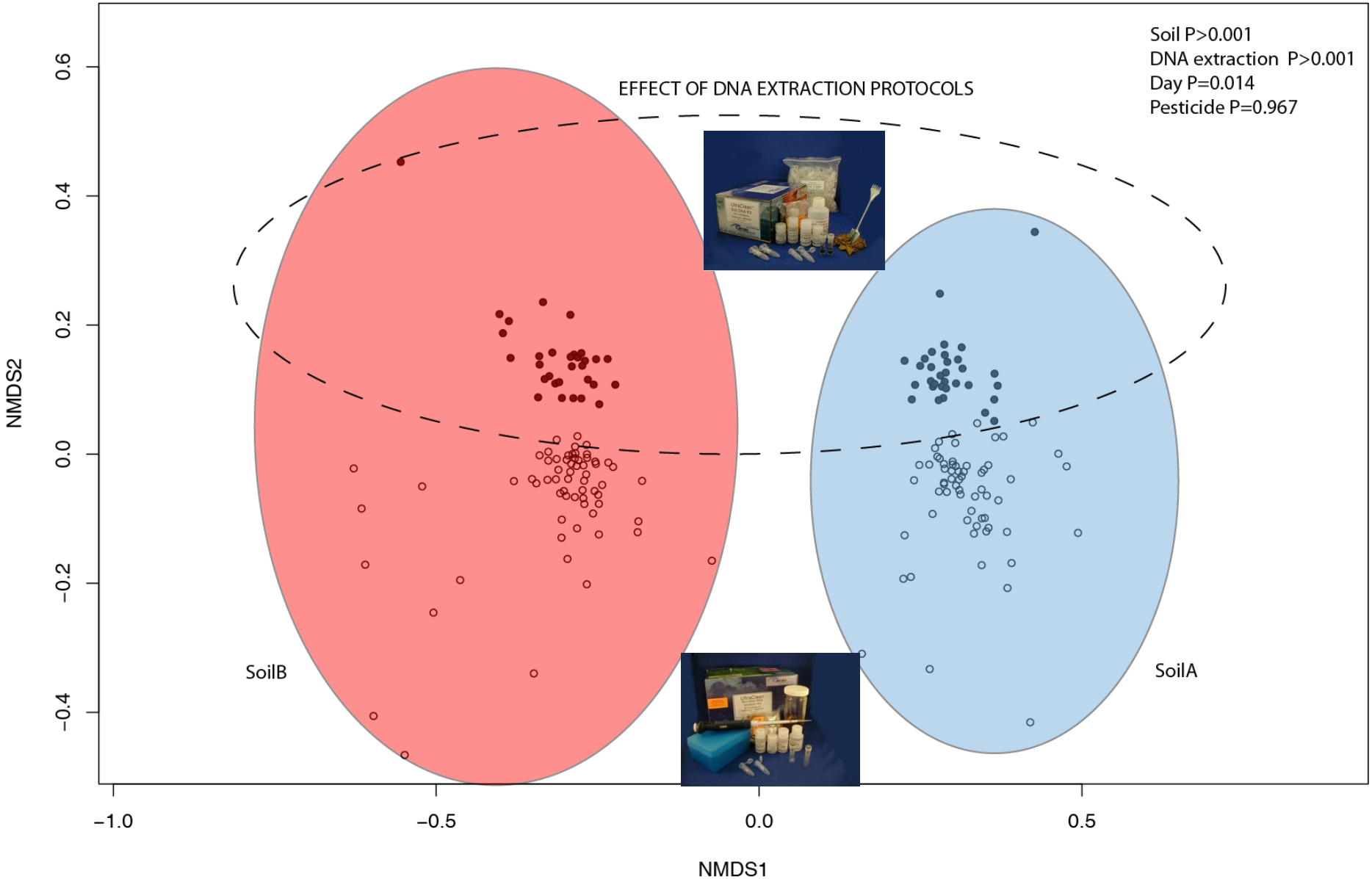
Basic workflow

1. Preaparation of sample
2. DNA extraction
3. PCR
4. Sample pooling and library preparation
5. Sequencing
6. Data processing
7. Statistical evaluation

Preparation of sample and DNA extraction

- Sterilised or un- sterilised sample
- Bulk soil, rhizosphere, rhizoplane, phyloplane, plant tissue?
- Sterilisation with NaHClO 3 - 5 min followed by rinsing with sterile water
- Sterilised tissue contains low numbers of microbes –large scale protocols?
- DNA extraction protocol according sample type (soil or plant kit?)
- **Use the same protocol for all samples!!!**

Mobio UltraClean Soil DNA vs Mobio Microbial DNA kit

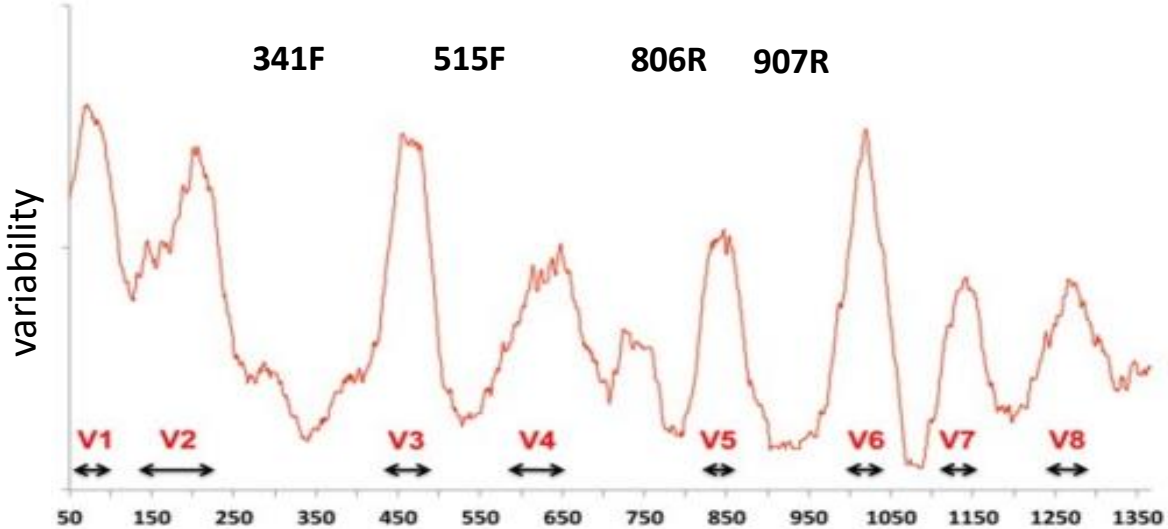


Primer selection for 16S rRNA gene

- Approx 1500 bp, contains 9 variable and 10 conservative regions
- Multiple copies of 16S rRNA gene in single genome
  - variation in copy number
  - some can be different
- Max length of fragment for illumina 500-550 bp
- Larger fragments – better resolution but more errors

Most used combinations:

V3-V4	341F	806R
V4	515F	806R
V4-V5	515F	907R
V1-V3	27F	518R

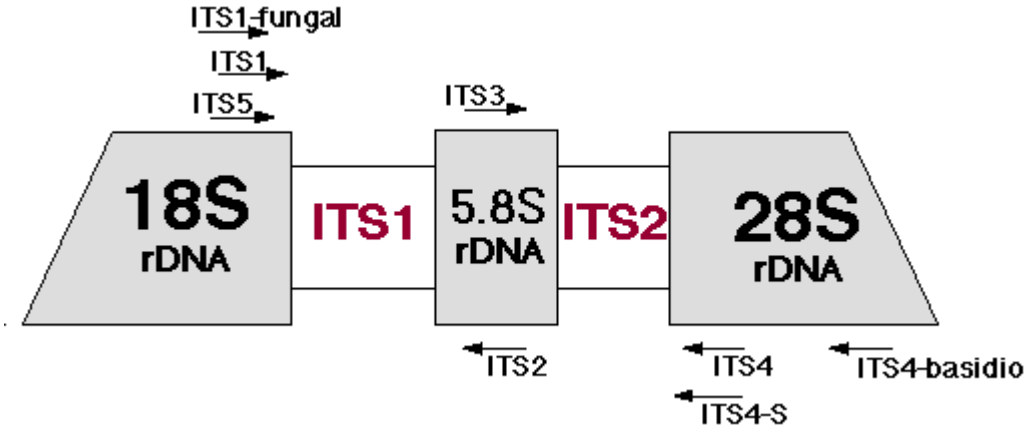


Primer selection for ITS

- Great variation in lenght , ITS1 and ITS 2 variable regions
- Multiple copies of ITS in single genome (may be up to 100)
- Max length of fragment for illumina 500-550 bp – need to choose ITS1 or ITS2
- ITS2 seems to be better

Most used combinations:

ITS1	ITS1 ITS2
ITS1	ITS1F ITS2
ITS2	gITS7 ITS4



PCR

- Use of High fidelity polymerase
- Polymerase without AAA overhang
- Use low number of cycles
- Amplify in 3 replications and pool together

- After PCR, product should be pooled to be sequenced
- Method depends on index strategy

Sequencing

Illumina sequencers
www.illumina.com
20140120

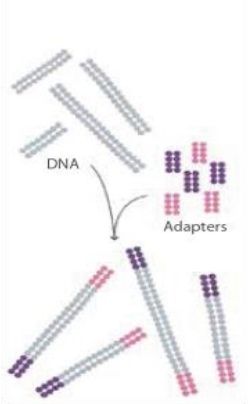
	MiSeq	NextSeq 500	HiSeq 2500	HiSeq X
No. of flow cells (FC)	1 FC	1 FC	2 FC	2 FC
Lanes/FC	1 lane/FC	4 lanes/FC	2* or 8 lanes/FC	8 lanes/FC
Maxm. clusters/run	25M	400M	600M* or 4,000M	6000M
Maxm. read length	Up to 2x300 bp	Up to 2x150 bp	Up to 2x150* or 2x125bp	Up to 2x150 bp
Maxm. Gb/run	15Gb/run	120Gb/run	180Gb* or 1000Gb/run	1,800Gb/run
Maxm. Hrs/run	< 56 hrs	< 30 hrs	40 hrs* or 6 days	3 days
Samples per run				
Human WGS (100Gb/sample)		1 Hu WG/run	≤ 10 HuWG's/run	≤ 18 HuWG's/run
Exomes (25M clusters/sample)	1 exome/run	≤ 16 exomes/run	≤ 160 exomes/run	
Bacterial WGS (0.5Gb/sample)	≤ 30 Bact WG/run	≤ 240 Bact WG/run	≤ 2000 Bact WG/run	
Gene expression (20M clusters/sample)	1 RNA-seq/run	≤ 20 RNA-seq/run	≤ 200 RNA-seq/run	
TF ChIP-Seq (5M clusters/sample)	≤ 5 ChIPSeq/run	≤ 80 ChIPSeq/run	≤ 800 ChIPSeq/run	
Gene panels (2M clusters/sample)	≤ 12 gene panels/run	≤ 200 gene panels/run		
Dimensions (WxHxD)	68.6 x 56.5 x 52.3 cm	58.5 x 53.4 x 63.5 cm	119 x 76 x 94 cm	119 x 76 x 94 cm
Weight	54.5 Kg	83 Kg	225 Kg	226 Kg

*Lower output modes are not described above

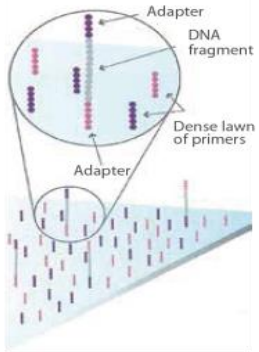
* Applies to Rapid Run mode

Best one for amplicons

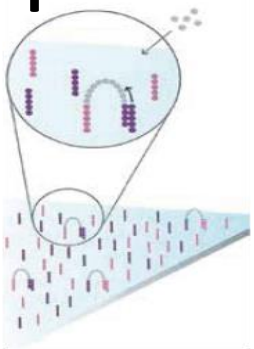
Illumina sequencing



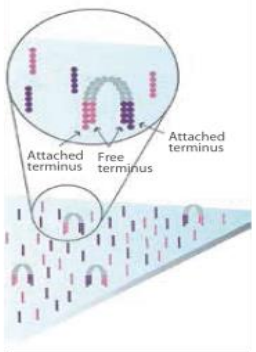
Input fragments with **adapters**



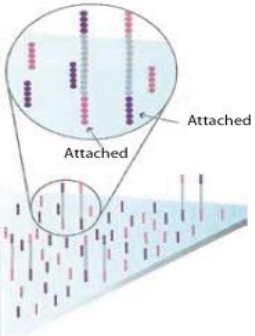
Anealing of fragments to **flow cell**



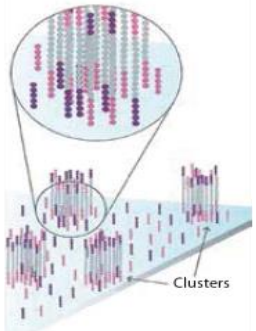
Anealing of second end to near primer: Making **bridge**



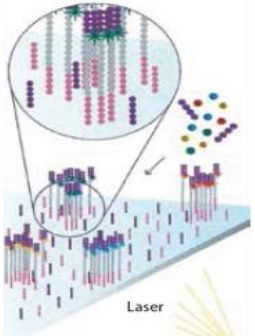
Synthesis of complementary strain: **Bridge amplification**



Denaturation of bridge

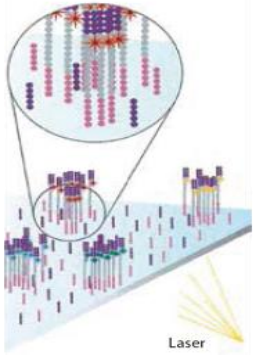


Clusters with thousands of fragments



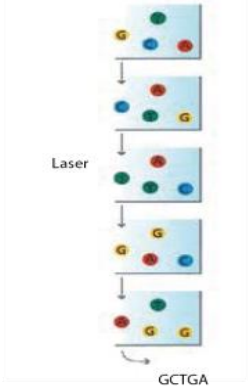
Flooding with **fluorescently** labeled nucleotides

Microscopic **photo of flowcell**

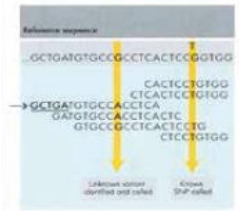


Flooding with fluorescence labeled nucleotides 2nd time

2nd photo of flow cell



Analysis of photos

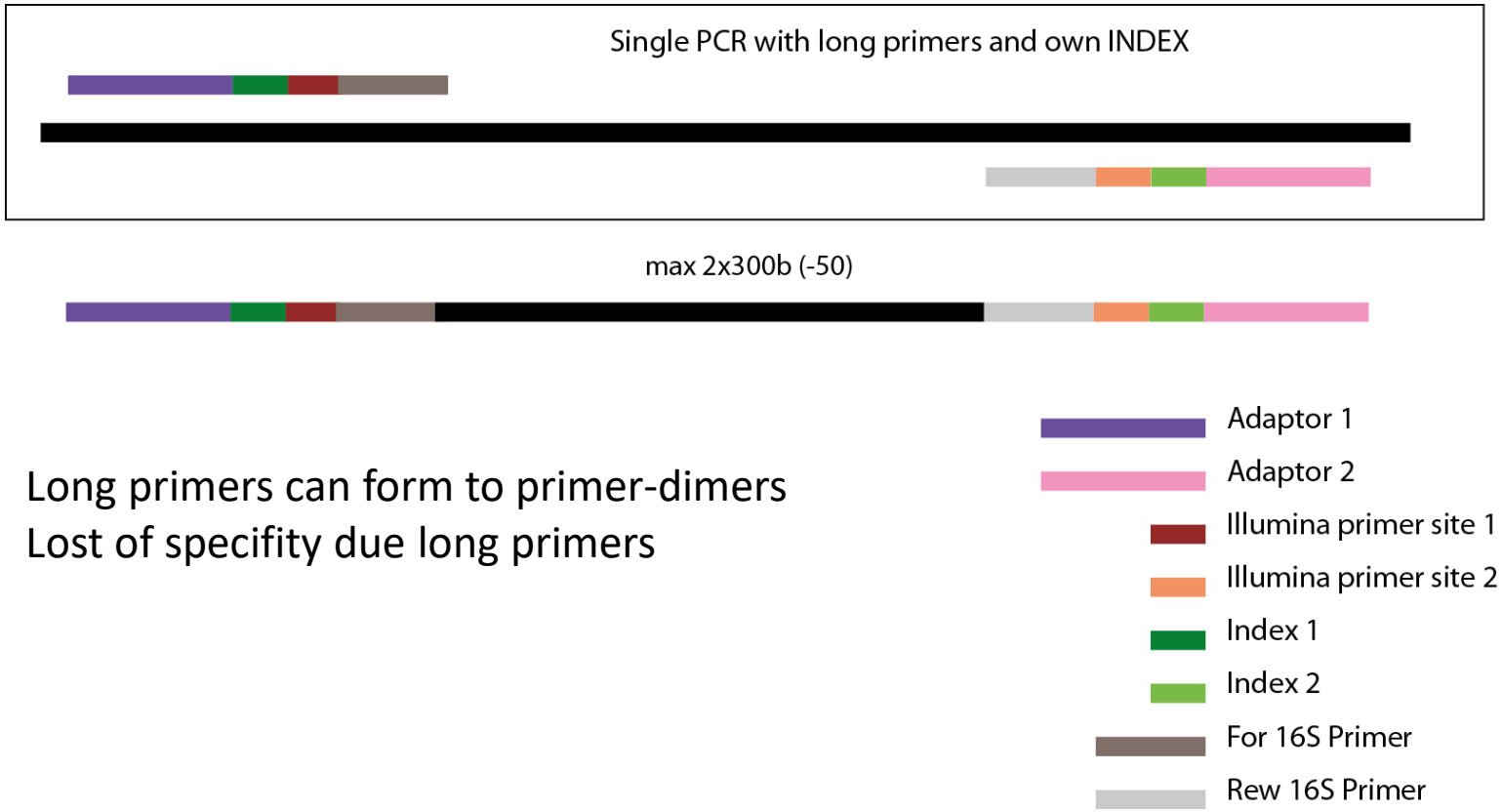


Resulting sequences

Library preparation

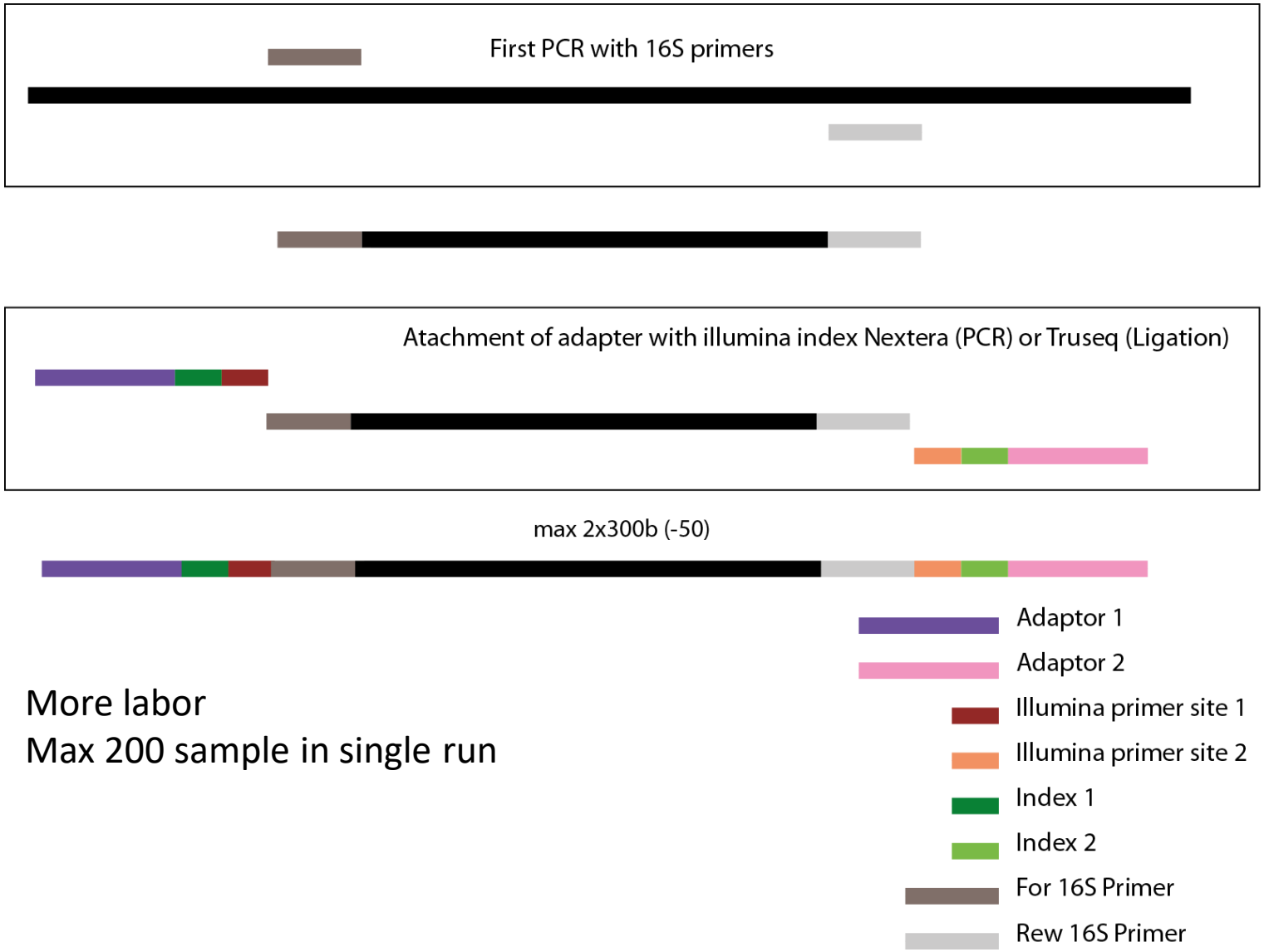
- DNA fragments need to have attached ADAPTORs
- Ligation – TruSeq kits
- PCR – Nextera kits
- During attachment of adaptors INDEXes can be added
- Indexes allows to distinguish samples
- They are on both sides of fragments
  - TruSeq kits – single or double index 6 bp
  - Nextera kits – double index 8bp

Library preparation option A



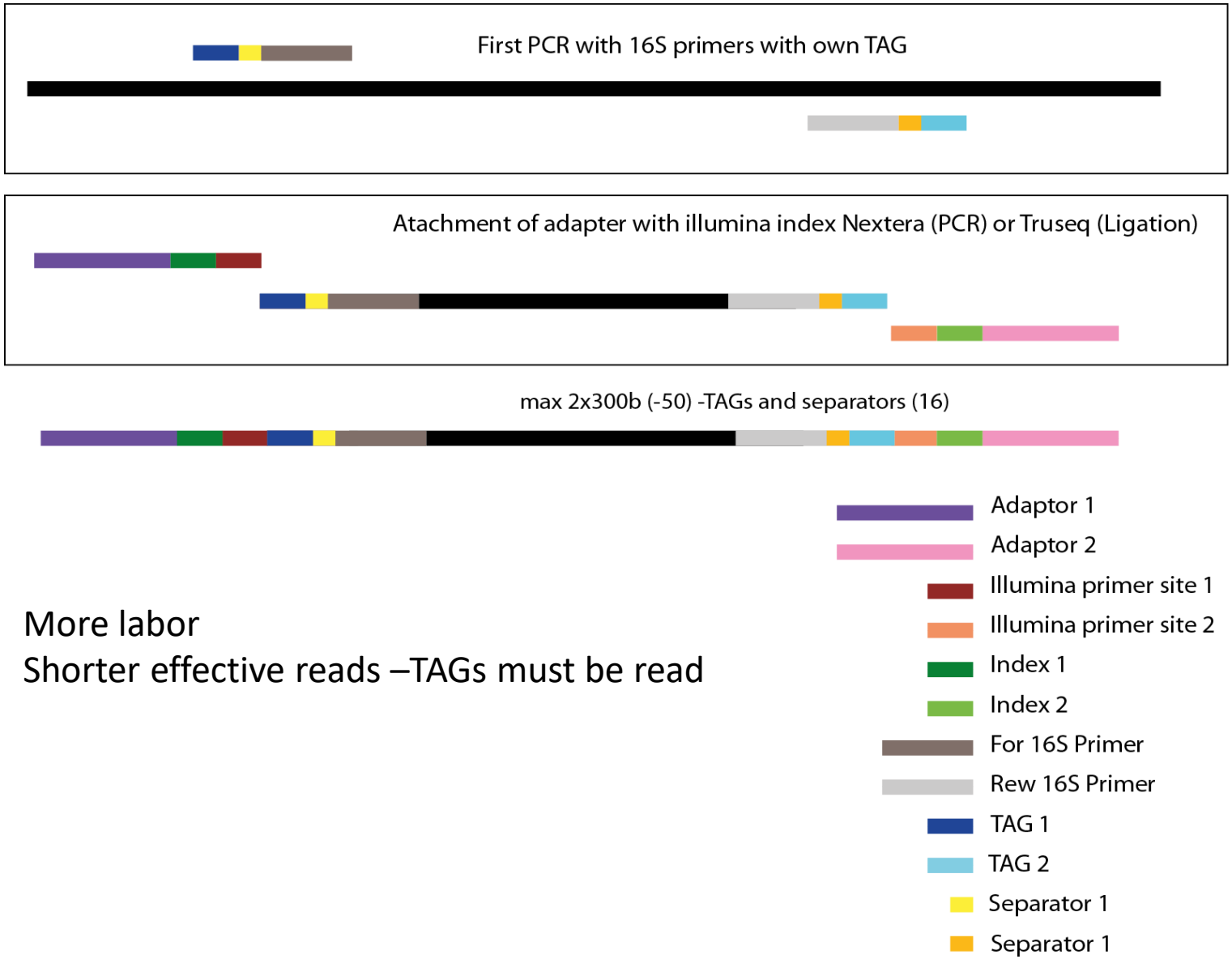
Long primers can form to primer-dimers
Lost of specifity due long primers

Library preparation option B



More labor
Max 200 sample in single run

Library preparation option C



More labor

Shorter effective reads –TAGs must be read

Protocol for TruSeq and own tags

- PCR with special designed primers
- Each primer contain short 6bp identification sequence (tag)
- Each sample need to be PCR with unique combination of primer
- Sample are pooled together sequenced and then identified according their tags

Primer design

515F_T001	**AAAGCGTGTGCCAGCMGCCGCGGTAA**
515F_T002	**ACGAAGTGTGCCAGCMGCCGCGGTAA**
515F_T048	**AATGCAGTGTGCCAGCMGCCGCGGTAA**
515F_T049	**AATTTAGTGTGCCAGCMGCCGCGGTAA**
515F_T096	**ATTGCGTGTGTGCCAGCMGCCGCGGTAA**
515F_T099	**ATAAAGAGTGTGCCAGCMGCCGCGGTAA**

Tag with shifting nucleotide

Spacer

16S primer sequence

Primer sets

515F_T001	AAAGCGTGTGCCAGCMGCCGCGGTAA
515F_T002	ACGAAGTGTGCCAGCMGCCGCGGTAA
515F_T003	ACCTTGTGTGCCAGCMGCCGCGGTAA
515F_T005	ATAATGTGTGCCAGCMGCCGCGGTAA
515F_T006	AGGGTGTGTGCCAGCMGCCGCGGTAA
515F_T007	AGCCAGTGTGCCAGCMGCCGCGGTAA
515F_T008	AGTTCGTGTGCCAGCMGCCGCGGTAA
515F_T009	ATTGAGTGTGCCAGCMGCCGCGGTAA
515F_T048	AATGCAGTGTGCCAGCMGCCGCGGTAA
515F_T049	AATTTAGTGTGCCAGCMGCCGCGGTAA
515F_T051	AGATTGGTGTGCCAGCMGCCGCGGTAA
515F_T052	ATCCTCGTGTGCCAGCMGCCGCGGTAA
515F_T055	AATAGAGTGTGCCAGCMGCCGCGGTAA
515F_T056	ATCCCTGTGTGCCAGCMGCCGCGGTAA
515F_T058	ACATTTGTGTGCCAGCMGCCGCGGTAA
515F_T059	ATGGAAGTGTGCCAGCMGCCGCGGTAA
515F_T096	ATTGCGTGTGTGCCAGCMGCCGCGGTAA
515F_T099	ATAAAGAGTGTGCCAGCMGCCGCGGTAA
515F_T100	ATGCTGAGTGTGCCAGCMGCCGCGGTAA
515F_T101	ACGGCTCGTGTGCCAGCMGCCGCGGTAA
515F_T102	AGATGATGTGTGCCAGCMGCCGCGGTAA
515F_T103	AATATACGTGTGCCAGCMGCCGCGGTAA
515F_T104	ACTAATCGTGTGCCAGCMGCCGCGGTAA
515F_T105	ACGACATGTGTGCCAGCMGCCGCGGTAA

806R_T003	ACCTTCCGGACTACHVGGGTWTCTAAT
806R_T007	AGCCACCGGACTACHVGGGTWTCTAAT
806R_T011	AACAGCCGGACTACHVGGGTWTCTAAT
806R_T014	ACGGCCCGGACTACHVGGGTWTCTAAT
806R_T020	ACTGGCCGGACTACHVGGGTWTCTAAT
806R_T026	AAGGACCGGACTACHVGGGTWTCTAAT
806R_T027	ACGCTCCGGACTACHVGGGTWTCTAAT
806R_T029	AGCGCCCGGACTACHVGGGTWTCTAAT
806R_T050	AATTAGCCGGACTACHVGGGTWTCTAAT
806R_T052	ATCCTCCCGGACTACHVGGGTWTCTAAT
806R_T054	AACCCGCCGGACTACHVGGGTWTCTAAT
806R_T056	ATCCCTCCGGACTACHVGGGTWTCTAAT
806R_T064	ACTCGCCCGGACTACHVGGGTWTCTAAT
806R_T066	AGGCCTCCGGACTACHVGGGTWTCTAAT
806R_T076	ACCCAGCCGGACTACHVGGGTWTCTAAT
806R_T079	AAAGAGCCGGACTACHVGGGTWTCTAAT
806R_T101	ACGGCTCCCGGACTACHVGGGTWTCTAAT
806R_T105	ACGACATCCGGACTACHVGGGTWTCTAAT
806R_T107	ATCCCGCCCGGACTACHVGGGTWTCTAAT
806R_T111	AGTGCGCCCGGACTACHVGGGTWTCTAAT
806R_T112	AGACTCTCCGGACTACHVGGGTWTCTAAT
806R_T114	AGGGGTTCCGGACTACHVGGGTWTCTAAT

24 X 22 = 528 combinations

Unique combinations

515F_T001	X	806R_T003	=	Sample 1
515F_T002	X	806R_T003	=	Sample 2
515F_T003	X	806R_T003	=	Sample 3
515F_T005	X	806R_T003	=	Sample 4
515F_T006	X	806R_T007	=	Sample 5
515F_T007	X	806R_T007	=	Sample 6
515F_T008	X	806R_T007	=	Sample 7
515F_T009	X	806R_T007	=	Sample 8
515F_T048	X	806R_T011	=	Sample 9
515F_T049	X	806R_T011	=	Sample 10
515F_T051	X	806R_T011	=	Sample 11
515F_T052	X	806R_T011	=	Sample 12
515F_T055	X	806R_T011	=	Sample 13

TruSeq library preparation

- Sample pooling
  - Cleaning (AMPure micro beads)
  - Quantification (qubit)
  - Pooling
- Ligation
  - End Repair
  - Clean-up with AMPure beads – removal of short fragments + enzymes
  - Adenylate 3´ Ends
  - Ligation of adapters
  - Clean-up

Process is succesfull in 5-95%

Library quantification

- Using qubit (fluorimetric quantification)
  - Measure of all DNA (ligated or non-ligated)
- Using qPCR
  - Specific set of primers (compatible to adapters)
  - Measure only ligated fragments
  - Should be more precise

Preparations for sequencing

- Dilution to 4 nM
- Library should be denatured
- 5µl library + 5µl 0,2M NaOH
- Dilution to 20pM
- PhiX (control) should be also denatured
- Dilution to 20pM
- Mix library and PhiX in ratio 99:1
- Final dilution to: 17-20pM for MiSeq kit V3
  15-15pM for MiSeq kit V2

Spustenie sekvenovania

- MiSeq kit:
  - 1X cartridge with chemicals
  - 1X flow cell
  - 1X rinsing solution
- Pipette mixed library and Phix to cartridge and insert to equipment
- Rinse Flow cell by water, dry it and insert to equipment
- Create sample sheet= set parameters (lenght of reading, single/pair end...)
- Set library preparation and indexes
- During sequencing data are logged to basespace cloud

Sequencing

- Forming of clusters (ca 4 hours)
- Optimal cluster density for V3 kit is 1200-1400/mm2 (=18-20pM)
- Reading bases(cca 7 minút/ jedna báza)
- 1. reading from first primer 301 bp
- 2. reading index1
- 3. reading index2 (TruSeq have only 1 index)
- 4. reading from second primer 301 bp

Run JuarajMix17: Run & Lane Metrics

SUMMARY | SAMPLES | CHARTS | METRICS | INDEXING QC | SAMPLE SHEET | FILES

	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%≥Q30
Read 1	301	8.77 Gbp	8.77 Gbp	0.97	4.87	240	63.98
Read 2 (I)	6	146.22 Mbp	146.22 Mbp	0.00	0.00	563	92.98
Read 3	301	8.77 Gbp	8.77 Gbp	0.95	8.20	195	49.77
Non-Index Reads Total	602	17.55 Gbp	17.55 Gbp	0.96	6.54	217	56.87
Totals	608	17.69 Gbp	17.69 Gbp	0.96	6.54	333	57.17

LANE	READ	TILES	DENSITY (K / MM^2)	CLUSTER PF (%)	PHAS/PREPHAS (%)	READS	READS PF	%≥Q30	YIELD	CYCLES ERR RATED	ALIGNED (%)	ERROR RATE (%)	ERROR RATE 35 CYCLES (%)	ERROR RATE 75 CYCLES (%)	ERROR RATE 100 CYCLES (%)	INTENSITY CYCLE 1	COMMENTS	STATUS
1	1	38	1,460 ±26	87.00 ±0...	0.162 / ...	33,611,7...	29,243,0...	63.98	8.77 Gbp	300	0.97 ±0.03	4.87 ±0.32	0.15 ±0.02	0.30 ±0.03	0.43 ±0.05	240 ±15		QC Passe...
	2 (I)	38	1,460 ±26	87.00 ±0...	0.000 / ...	33,611,7...	29,243,0...	92.98	146.22 ...	0	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	563 ±36		
	3	38	1,460 ±26	87.00 ±0...	0.153 / ...	33,611,7...	29,243,0...	49.77	8.77 Gbp	300	0.95 ±0.03	8.20 ±0.45	0.27 ±0.03	0.43 ±0.04	0.55 ±0.05	195 ±13		

Number of sequences

% of PhiX control

% of errors after 100 cycles (according PhiX)

Run JuarajMix17: Charts

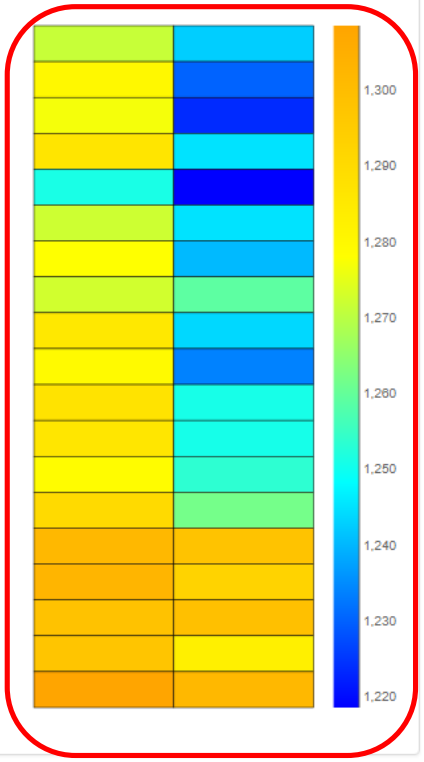
SUMMARY | SAMPLES | CHARTS | METRICS | INDEXING QC | SAMPLE SHEET | FILES

Flow Cell: 000000000-AKF85 Extracted: 608 Called: 608 Scored: 608

Quality of sequences

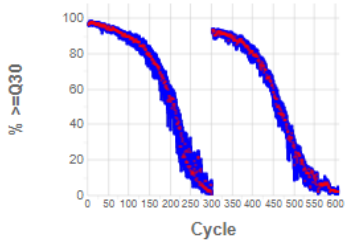
Flow Cell Chart

Chart: Density PF
Surface: Both Surfa
Fix Scale



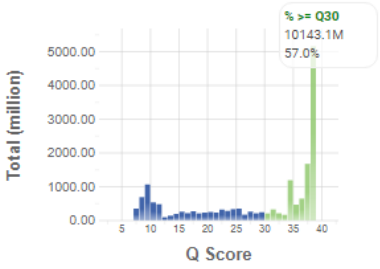
Data By Cycle

Chart: % >=Q30
Fix Scale
Accum



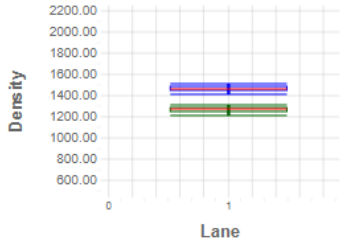
QScore Distribution

Surface: Both Surfa
Read: All Reads
Cycle: All Cycles
Fix Scale



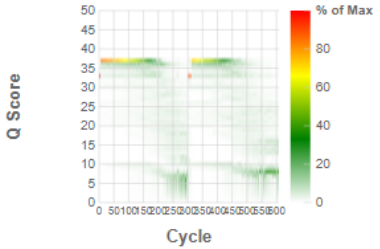
Data By Lane

Show: Density
Surface: Both Surfa



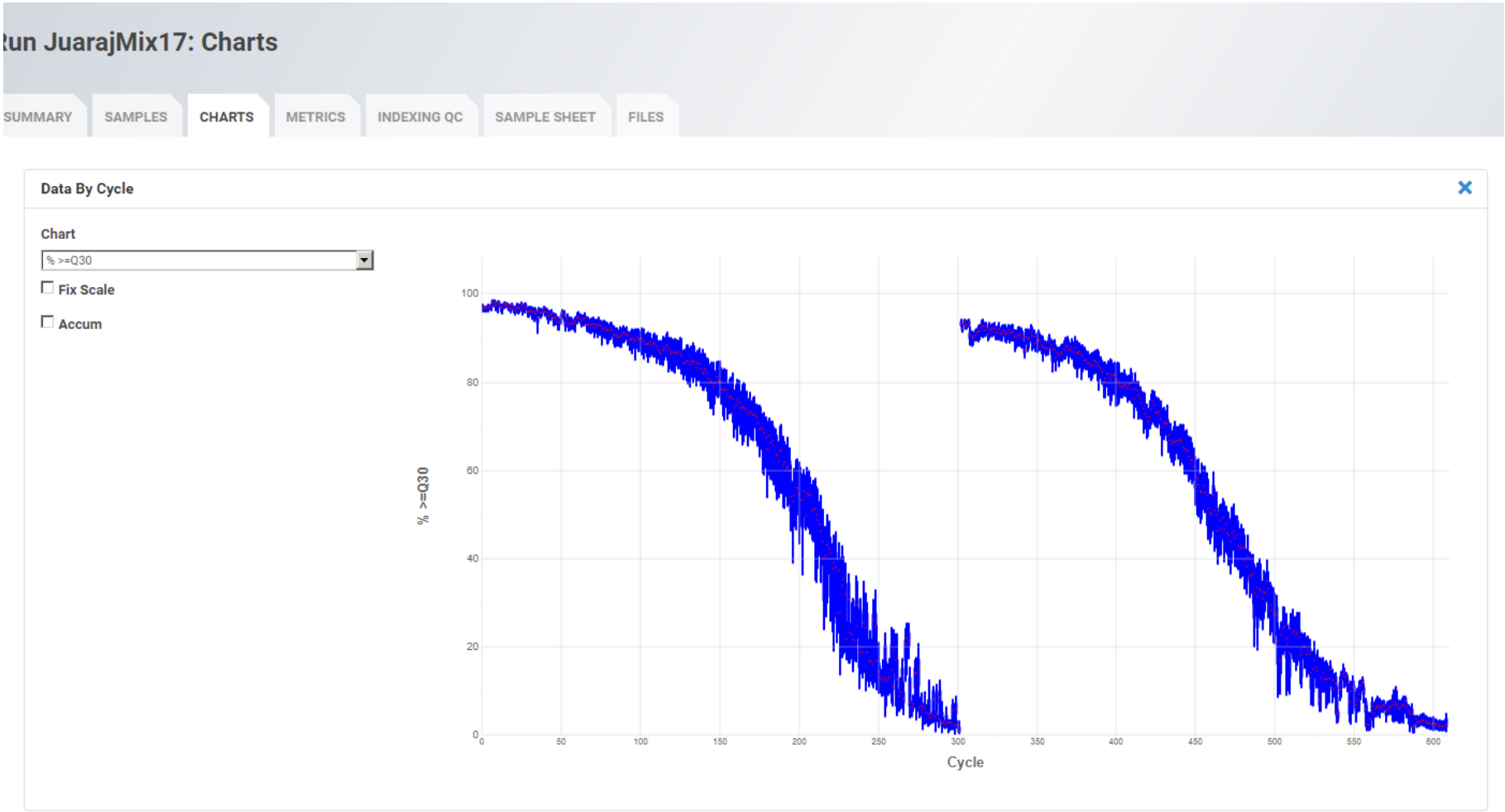
QScore Heatmap

Surface: Both Surfa

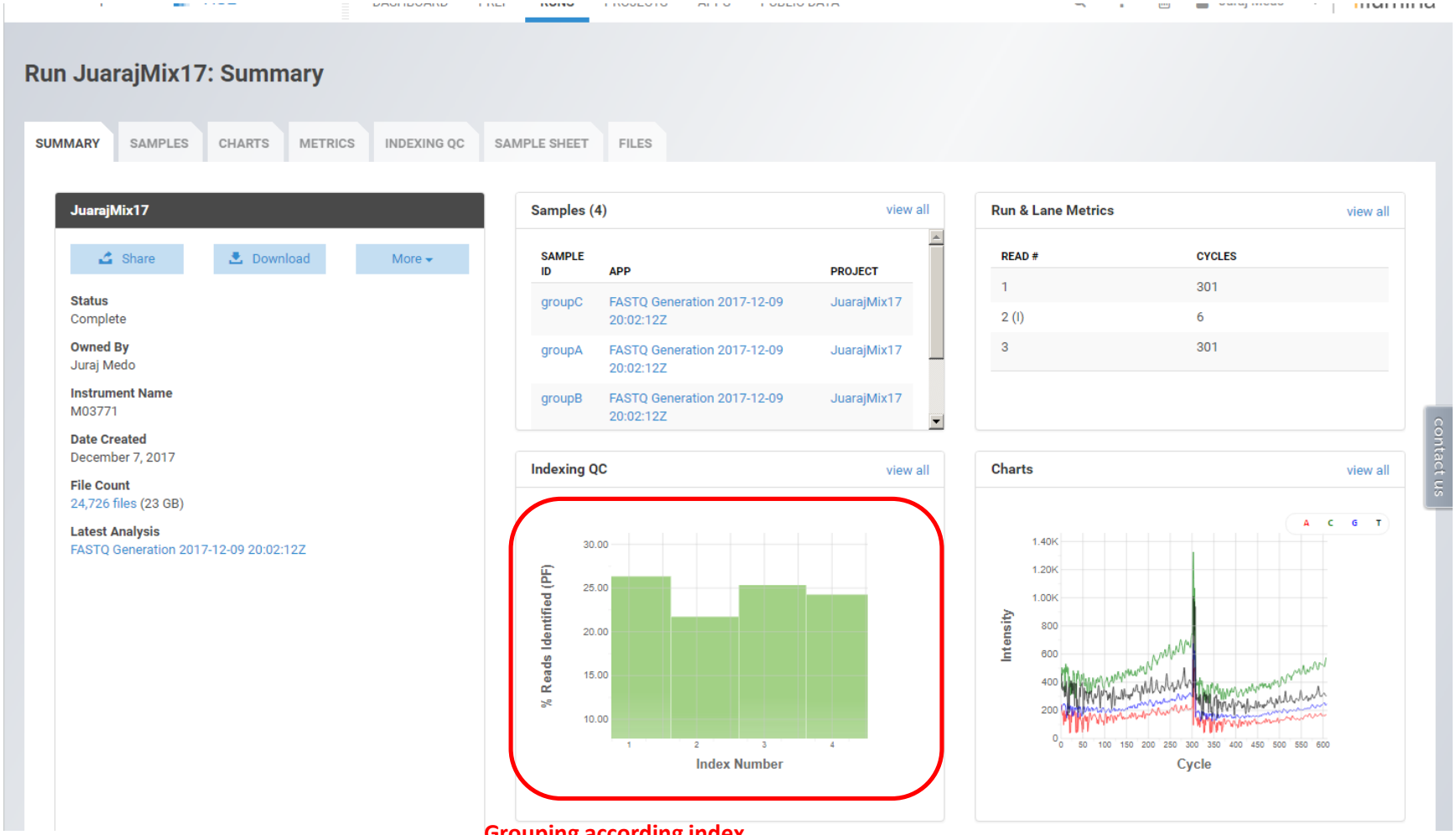


Cluster density on flowcell

Quality of basecalling



Rozdelenie indexov



Grouping according index

Sequencing data

- Automatically sorted according index
- For each index 2 files (forward and reverse read)
- Format FASTQ – containing sequences and quality of bases
- Big files cca 15 GB for single read whole run
- Each sequence have unique name (based on cluster position on flow cell)

Processing of results

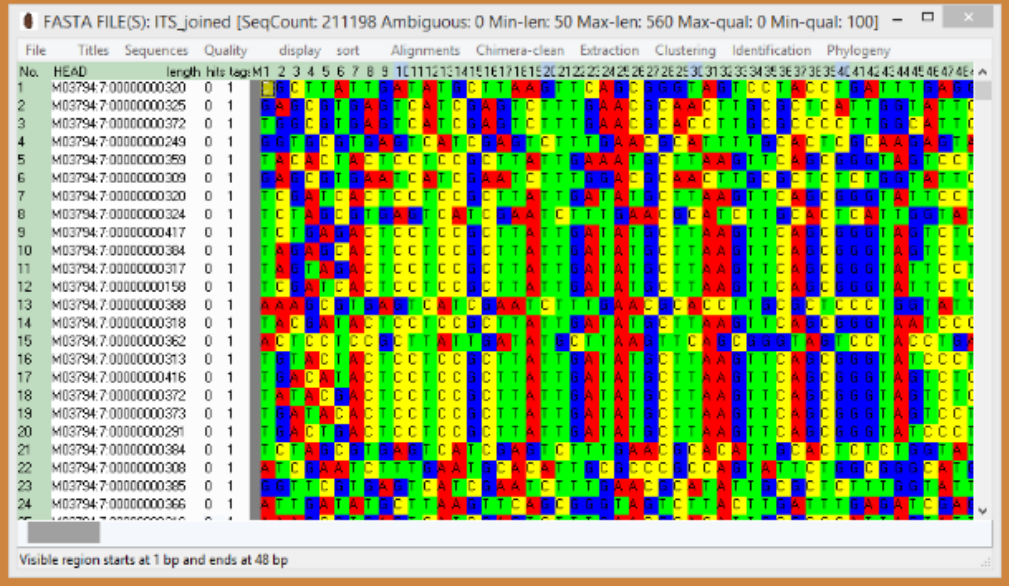
- Many bioinformatic pipelines
- QIIME, MOTHUR, DADA2....(most of them in Linux, command line based)
- We use pipeline SEED2 (Větrovský, Morais, Baldrian, 2018)
  - Windows based
  - Graphically oriented (you see sequences)
  - Contains all procedures required for analysis
  - Excelent for students and beginers
  - http://www.biomed.cas.cz/mbu/lbwrf/seed/

SEED2 - sequence editor and pipeline for Windows XP/7/8/10

- SEED2
- DOWNLOAD SEED
- TUTORIALS
- ARCHIVE
- SEED2 PAPER
- CONTACT

SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses

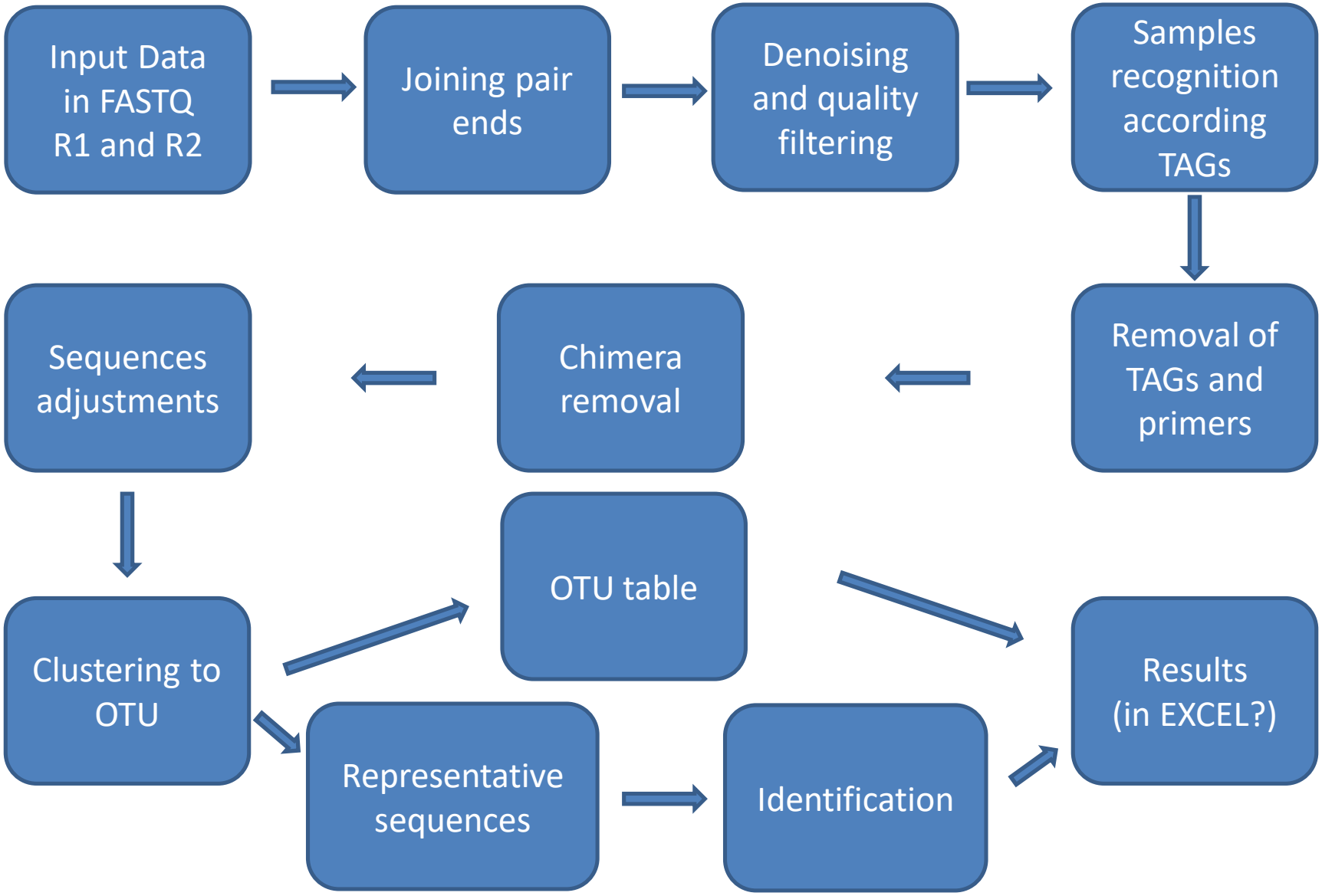
SEED2 is a free to use graphical user interface-based sequence editor and NGS data analysis pipeline for all Windows platforms with 64-bit architecture designed to guide users intuitively through the data analysis workflow. SEED2 was created to provide an intuitive interface for fast bioinformatic analysis of PCR amplicons from NGS (such as the bacterial 16S or fungal ITS amplicons) according to the suggested workflow. SEED2 is fully optimized for Illumina data and has the capacity to process whole Illumina MiSeq run on a standard personal computer with 16 GB RAM. The pipeline has a wide array of functions including the editing of sequences and their titles, sorting, quality trimming, pair-end joining, grouping of sequences based on sequence motifs or titles, batch processing of sequence groups, chimera removal, ITS extraction, sequence alignments and clustering, OTU table construction, construction of consensus sequences, creation of local databases for BLAST and searching within them or in the whole NCBI, retrieval of taxonomical classification from the NCBI, calculation of diversity parameters. Newly, it also covers tools to construct and edit phylogenetic trees. We believe that the pipeline is not only useful for processing of high-throughput amplicon data, but also for fast and easy editing and formatting of any large sequence datasets under the Windows environment. The latest version of SEED2 is Seed2.1_64bit.



Vetrovsky, T., Baldrian, P. and Morais, D. (2018) "SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses." Bioinformatics, bty071, 2018 (https://doi.org/10.1093/bioinformatics/bty071)

C4Sys

Pipeline SEED2



Statistical evaluation

- Diversity indices , rarefactions
- Graphs (barcharts, Krona)
- Heatmaps
- Ordination methods (PCA, PCoA, NMDS)
- Sample distance (similarity) computation
  - According OTU table (bray-curtis, Pearson...)
  - OTU table + phylogenetic tree of OTU (UniFrac)
- Hypothesis testing (PERMANOVA, ANOSIM)
- Difference analysis (simper)