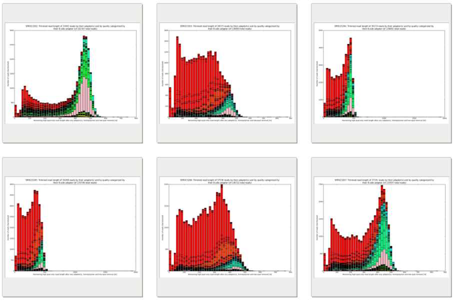
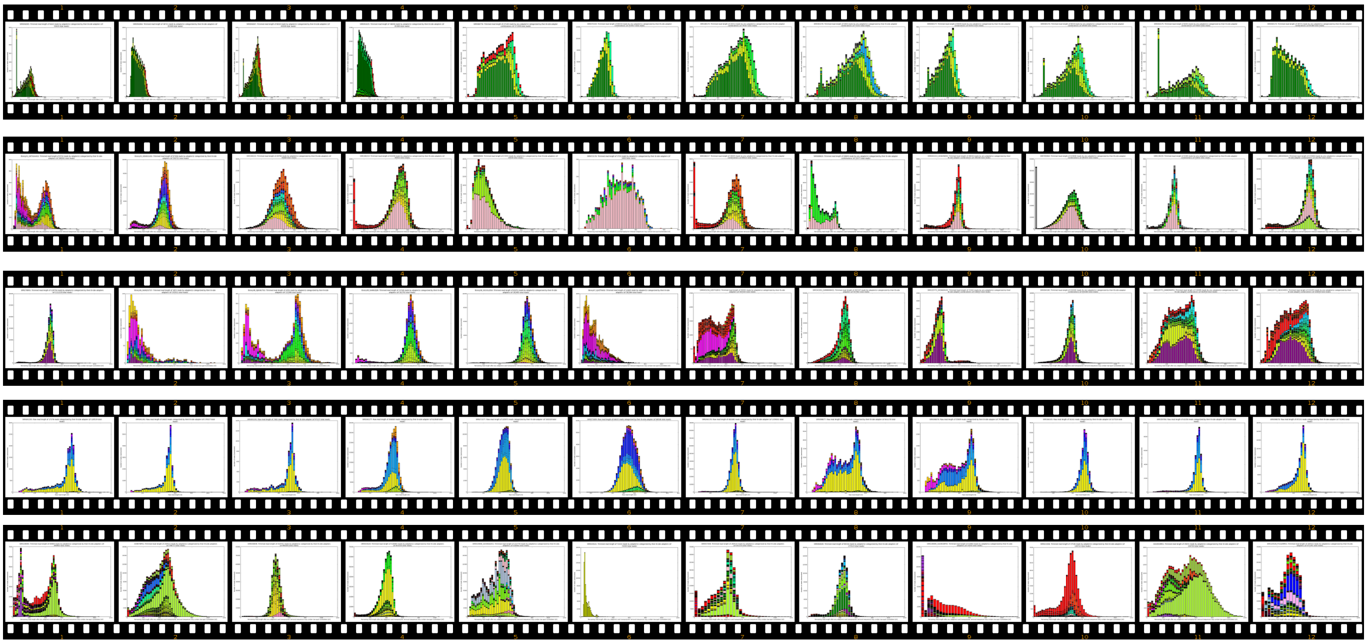
Fixing raw data from a pyrosequencer (primer/adapter/MID/artifact removal and quality control)

Although pyrosequencing technology is available at least since 2005 there are still a few tools to inspect and manipulate output data. Most regretfully, **there are just no tools to perform meaningful quality control**. While a number of sample preparation protocols being an alternative to Roche sample processing are available from third parties, a thorough evaluation of each of them and a robust comparison is missing. The technology has remained a black box, the alternative protocols make the reality more difficult to encompass and during last years more and more users felt they cannot cope with deemed defects in the sample preparation procedure(s) or sequencing or both and because of the lack of proper and in depth service support they rather migrated to competing technologies. While other technologies are a black box as well and must suffer from same type of errors and artifacts they are at least said to be cheaper in terms of nucleotides output per $. Provided the calculations are meaningful then it should hurt less if a sequencing experiment fails.

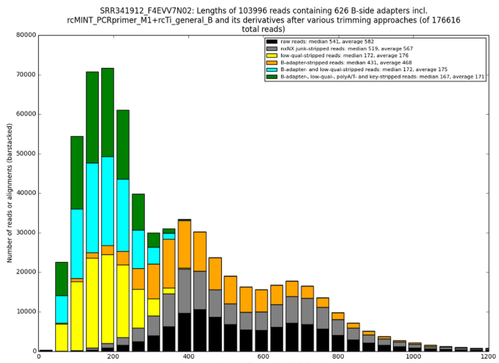


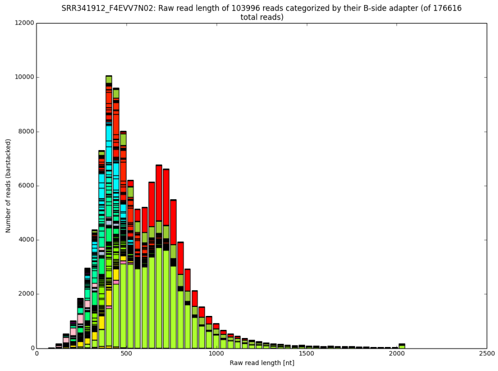
Bioinformatics.cz has done a pioneering work in analysis of publicly available 1800 datasets and as a result of the analyses and of proprietary software development it now offers a service to all users of the Roche/454 technology and IonTorrent users. The offers cover broad range of packages, notably proper removal of adapters/MIDs/artifacts from output sequences but also sharp figures describing quality and performance per every region physically sequenced. The charts enable molecular biologists to evaluate performance of the underlying laboratory protocols and take steps to improve sample processing and/or experimental design. The pipeline does not require original camera images or other intermediate files from Roche sequencers but operates on the Standard Flowgram Format (SFF) files or less preferably, on derived data in FASTQ or FASTA file formats. The cutting edge analytical approach documents vast differences in datasets world-wide and even a brief listing of some figures is a beautiful showcase. In certain aspects it is quite easy to spot a particular problem in sample processing while in other cases at least certain steps can be omitted from the list of suspects. In summary, bioinformatics.cz is the only company in the world offering a broad and independent analytical service, adapter removal and quality control.



Our software pipeline is able to correct data from any Roche 454™ Life Sciences™ sequencing machines (GS20™, GS FLX™, GS FLX+™, GS Junior™), some from IonTorrentPGM and some from Illumina. The core analysis package includes inspection of raw reads, evaluation of adapter localization, orientation followed by precise removal of adapters, MIDs, chimeras, software/virtual junk sequences and physical artifacts. The cleanup results in improved assemblies (less chimeras, less but larger both contigs and scaffolds) and also in some cases in shorter assembly times. Result files available through additional packages allow scientists to evaluate sample preparation quality, sequencing quality and importantly, aid with troubleshooting and problem finding. The key features of the software solution are demonstrated below in respect to Roche technology but some bits also apply to IonTorrent and Illumina data:

- automated recognition of dataset type (adapters, sample barcodes)
- supports all ever reported experimental setups (amplicon, shotgun, transcriptome, all types of MIDs from Roche, certain mixes of samples sequenced within physically same sequencing region)
- detection and removal of MID (Multiplex IDentifier) tags **from both ends** (Roche software still cannot do that), support for custom MID tags
- detection and removal of thousands of Roche/IonTorrentPGM and 3rd-party adapters (queries are dynamically generated specifically for each individual dataset)
- detection and removal of thousands of known chimeras and other artifacts isolated from over 1800 of publicly available datasets
- outputs corrected data in SFF / FASTA+QUAL / FASTQ file formats
  - adapters/MIDs are annotated in the SFF / CSV files and **prevented to interfere with downstream processing**
  - authentic polyA-tails of transcripts are corrected in FASTQ / FASTA+QUAL files and retained to ensure the best assembly results (**reconstruct full-length mRNA** sequences from sequencing reads)
  - **junk sequences** are unleashed in raw data, annotated in output SFF file and omitted from cleaned data files (faster and better assemblies)
  - get **longer reads by rescuing carefully selected portions of the sequence** which Roche sacrificed in the name of low-quality for no good reason (get better assemblies with even less scaffolds, contigs and gaps, get increased scaffold/contig lengths and coverage)
- reports overall sequencing performance in CSV file format for your own Excel spreadsheet-based processing
  - determine your sequencing overhead
  - learn how many nucleotides were wasted in sequencing adapters/artifacts introduced by certain molecular-biology protocol
  - learn how much sequence is sacrificed in unused portion of reads
  - learn your effective sample insert size distribution and compare it with experimentally determined distribution obtained during sample preparation
- output beautiful figures in PNG format for advanced experimental interpretation and troubleshooting





Your lab protocol scenarios supported by our software pipeline

A. General pre-processing protocols (not bound to downstream sequencing technology):

These protocols are independent of the sequencing technology and are mostly related to cDNA preparation, prior to the sequencing. These are enumerated here because their adapters/primers/artefacts must be removed from the sequence of all sequencing reads prior to assembly.

□ [[NONE]] Tick this item when the sample contains only sequencing-technology-specific adapter sequences, no additionals from sample-prep. For example, this is the case of nebulized / sonicated / sheared genomic or mitochondrial DNA, or sample fragmented using dsDNA endonuclease (Fragmentase®) method, NEBNext®, NewEngland BioLabs (Adey et al., 2010).

Contrary to this item, the alternatives listed below introduce additional nucleotide sequences into sequencing reads, not strictly related to the sequencing technology.

Roche protocol:

□ [[a]] Roche transcriptome cDNA preparation (random hexamer was used for cDNA first-strand synthesis)

Evrogen MINT1 protocol:

□ [[b]] non-oriented insert sequences

Evrogen MINT2 protocols:

□ [[c]] non-normalized cDNA, non-oriented inserts
□ [[d]] non-normalized cDNA, oriented inserts

Clontech SMART (Matz el al., 1999), SMARTer and SMARTer II protocols:

□ [[e]] normalized cDNA (Clontech SMART^{TM}), non-oriented inserts
□ [[f]] normalized cDNA (Clontech SMART^{TM}), oriented inserts
□ [[g]] normalized cDNA (Clontech SMARTer II)
□ [[h]] modified Clontech PlugOligo with MmeI (Clontech SMART^{TM}) (Zeng et al., 2011)
□ [[i]] non-normalized DNA, non-oriented inserts, modified MINT_PCRprimer_M1 with MmeI (Brenchley et al., 2012)

Other protocols:

□ [[j]] ncRNA sample (Huttenhofer and Vogel 2006; Mrazek et al., 2007)
□ [[k]] microRNA sample
□ [[l]] T7 RNA polymerase promoter-based methods
Clontech GenomeWalker-based cDNA reads (PT3042-1, PT1116-1),
a method from LGC Genomics, Berlin
□ [[m]] SP6 polymerase-based *in vitro* run-off transcripts
□ [[n]] transposase-based adapter insertion, Nextera, Epicentre® (Adey et al., 2010)
□ [[o]] BAC clone sequencing
□ [[p]] I.M.A.G.E. clone sequencing
□ [[q]] a method from GATC Biotech, Konstanz ("pine tree" cDNA project) (Chang et al., 1993; Pavy et al., 2008)
□ [[r]] Suppression subtractive hybridization (SSH) PCR (Clontech PCR-Select Bacterial genome subtraction kit) (Diatchenko et al., 1996)
□ [[s]] full-length cDNA method (BioS&T uncloned normalized cDNA, possibly the F. Bonaldo method: Soares et al., (1994),
see also Patanjali et al., (1991))
□ [[t]] Cap-Trsa-CV-based normalized cDNA method (modified Clontech SMART by J. Buchanan-Carter, Z. Smith, K. Mockaitis, M. Matz)
□ [[u]] Nimblegen capture method (Bashiardes, 2005)
□ [[v]] custom oligo-dT adapter with GsuI restriction site (Leroux et al., 2010; Le Cam et al., 2012)

B. Roche (454^{TM}) protocols and their alternatives:

Standard Amplicon Protocols:

GS20/GS FLX^{TM} emPCR Kit II (aka amplicon A, Paired End) (Roche Application Notes No. 3 (2006) and No. 8 (2007), Kappa Biosystems #KP0001):

□ [[1]] amplicon A without GSMIDs
□ [[2]] amplicon A with GSMIDs on the left
□ [[3]] amplicon A with GSMIDs on both sides*

GS20/GS FLX^{TM} emPCR Kit III (aka amplicon B) (Jarvie and Harkins (2008), Galan et al. (2010), Kappa Biosystems #KP0001):

□ [[4]] amplicon B no GSMIDs
□ [[5]] amplicon B with GSMIDs on the left
□ [[6]] amplicon B with GSMIDs on both sides*

GS20/GS FLX^{TM} emPCR Kit II+III combined (amplicon A+B):

□ [[7]] amplicon A+B without GSMIDs
□ [[8]] amplicon A+B with GSMIDs on the left
□ [[9]] amplicon A+B with GSMIDs on both sides*

Titanium Amplicon protocols:

GS FLX Titanium^{TM} emPCR Kit II (aka amplicon A, Paired End):

□________[[10]] amplicon A without TiMIDs
□________[[11]] amplicon A with TiMIDs on the left
□________[[12]] amplicon A with TiMIDs on both sides*

GS FLX Titanium^{TM} emPCR Kit III (aka amplicon B):

□________[[13]] amplicon B without TiMIDs
□________[[14]] amplicon B with TiMIDs on the left
□________[[15]] amplicon B with TiMIDs on both sides*

GS FLX Titanium^{TM} emPCR Kit II+III combined (amplicon A+B):

□________[[16]] amplicon A+B without TiMIDs
□________[[17]] amplicon A+B with TiMIDs on the left
□________[[18]] amplicon A+B with TiMIDs on both sides*

Whole Genome Shotgun protocols:

GS20/GS FLX Standard (prepared using emPCR Kit I)

□________[[19]] GS20/GSFLXstd without GSMIDs
□________[[20]] GS20/GSFLXstd with GSMIDs on both sides*

Rapid Library Y-type adapter for Taqman MGB qPCR quantification (Zheng et al., 2010)

□________[[21]] Y-type-GS20-FLXstd-Taqman-MGB_qPCR without GSMIDs
□________[[22]] Y-type-GS20-FLXstd-Taqman-MGB_qPCR with GSMIDs on both sides*

Rapid Library Y-type adapter for Taqman MGB qPCR quantification (Zheng et al., 2011)

□________[[23]] Y-type-RapidLib-Taqman-MGB_qPCR without GSMIDs
□________[[24]] Y-type-RapidLib-Taqman-MGB_qPCR with GSMIDs on both sides*

GS FLX/FLX+ Titanium (emPCR Kit I) (Roche TCB-004 2009)

General Library Preparation Protocol with Roche adapter

□________[[25]] no TiMIDs (IonTorrent shotgun data also match this)
□________[[26]] with TiMIDs on the left
□________[[27]] with TiMIDs on both sides*

Rapid Library Preparation Protocol with Roche adapter

□________[[28]] Y-type-RapidLib without RLMIDs
□________[[29]] Y-type-RapidLib with RLMIDs on both sides

Rapid Library Preparation Protocol with oligonucleotide adapter(s) from http://IDT.dna.com:

□________[[30]] Y-type-RapidLib-IDT without RLMIDs
□________[[31]] Y-type-RapidLib-IDT with RLMIDs on both sides

Crazy datasets:

□________[[32]] mis-carried samples with two Roche adapters surrounding the sample sequence
□________[[33]] mis-carried samples with two Roche adapters surrounding the sample sequence with GSMID/TiMID/RLMID tags to make things worse
□________[[34]] mixed reads with two or more different Roche adapters in same sequencing region
□________[[35]] mixed reads with two or more different Roche adapters in same sequencing region with GSMID/TiMID/RLMID tags to make it worse

Notes:
** Analysis of multiplexed samples is computationally highly expensive. It is getting even worse once you have* ***different*** *MIDs on the left and right end of the read. However, the true performance killers are Evrogen-based transcriptomic setups involving MIDs and datasets with doubled Roche or 3rd-party adapters. For a more thorough pricing calculation please contact us with details of your experimental setup.*

Published by

Martin Mokrejs, PhD.
BIOINFORMATICS.CZ
Donovalska 1658
149 00 Prague
Czech Republic
mmokrejs [a] bioinformatics.cz