Nitty-gritties Of Handling NGS Data

Look Before You Leap

Rohit Shukla

Co-Founder & Director

Bionivid Technology Pvt Ltd

Bangalore - India

ABOUT BIONIVID – THE FOUNDING TEAM

Biopreneur

‘Believe and take the leap of faith’

What happens when three dynamic trios come together with similar interests and passion coupled with innate entrepreneurial attributes? The result — **A hot emerging start-up!**

Established in 2011, Bionivid Technology, located in the happening start-up hub of Bangalore, is a genomics and informatics start-up, jointly cofounded by Madavan Vasudevan, Hitesh Goswami and Rohit Nandan Shukla.

BioSpectrum the business of bioscience

BioPharma | BioAgri | Bio IT | BioIndustrial | CROs | Healthcare | Specials | Top20 | Programs

Trending :

Home > Biopreneur > Hot Start-ups: ‘Believe and take the leap of faith’

Bangalore | 18 July 2015 | Features | By Raj Gunashekar

Hot Start-ups: ‘Believe and take the leap of faith’

144 Share | 6 Tweet | 6 Share | 1 G+1 | 179 sharethis | Email | Print | 0 Comment

What happens when three dynamic trios come together with similar interests and passion coupled with innate entrepreneurial attributes? The result -- A hot emerging start-up!

Established in 2011, Bionivid Technology, located in the happening start-up hub of Bangalore, is a genomics and informatics start-up, jointly cofounded by Mr Madavan Vasudevan, Mr Rohit Nandan Shukla and Mr Hitesh Goswami.

Mr Madavan is a qualified microbiologist-turned-bioinformatician; His partner Mr Rohit is also a qualified bioinformatician and a certified programmer; and Mr Hitesh is a neurobiologist with a fine business acumen. All three of them make a perfect ingredient for raising a hot

ABOUT BIONIVID – SERVICE PORTFOLIO & STRENGTH

Diversification

Smart Analysis

Data Generation | Data Analytics | Learning Programme | SQDB / Database | SQIT / Custom App(s) | Genome Stations

Outsourced | Specialization | Niche Offering | Niche Offering | Niche Offering | High Performance Genome Computing Environment

PIONEERS

Genome "IT" Services
NGS Data Analysis

60+ PUBLICATIONS

Experience|Proven Expertise
Genome Informatics field

TEAM

Blend of Youth & Excellence
Great Scientific Acumen

PURVEYORS

Proven Ability To Craft Specialized Work Force

HOLISTIC SOLUTION

Data Generation – Interpretation – Data Management – Training

EVOLUTION OF DNA SEQUENCING

History

1970s: DNA Sequencing Starts

1990: The “Human Genome Project” Starts

2003: First Human Genome Fully Sequenced

2007: NGS Technology (***Massively Parallel*** & ***Universal Adapters***)

2012: UK Announces sequencing of 100K Genome

2015: USA Announces Sequencing of 1M Genomes

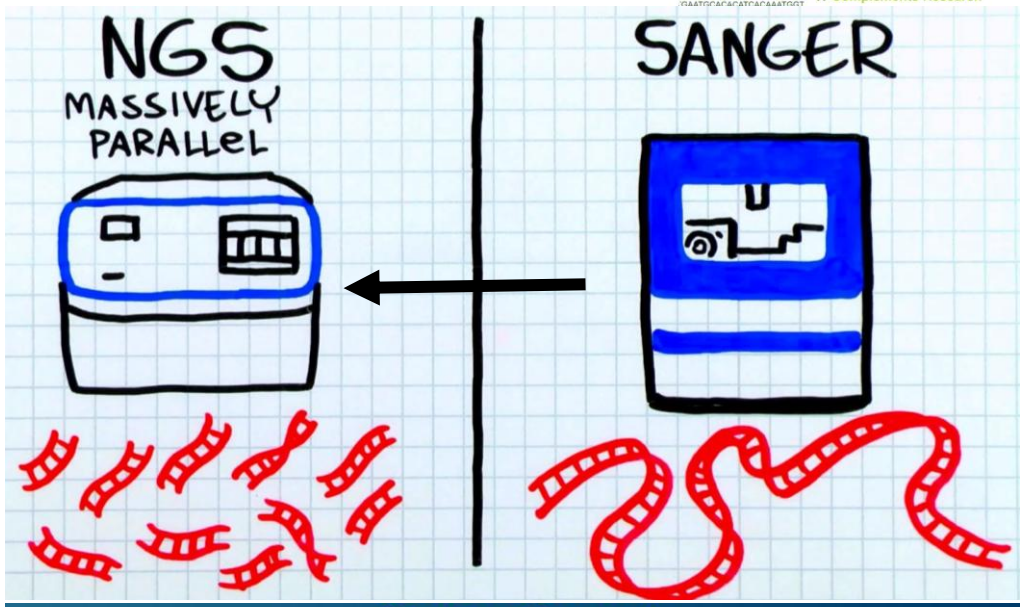
$$$

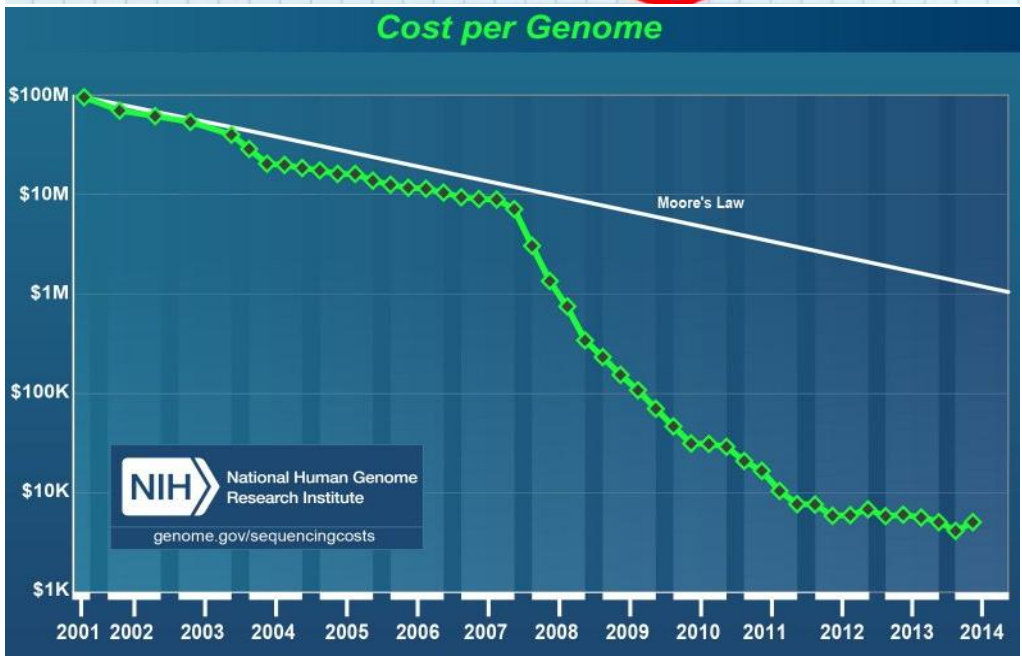
$ 3B : Human Genome Project Cost

$ 250K : Illumina Sequencing Cost (2008)

$ 5K : Complete Genomics (2009), Illumina (2011)

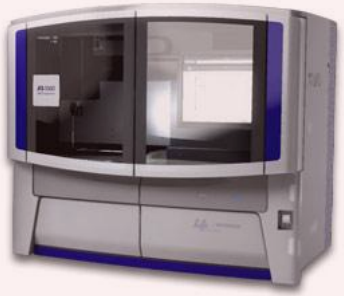
$ 1K : Illumina (2014)





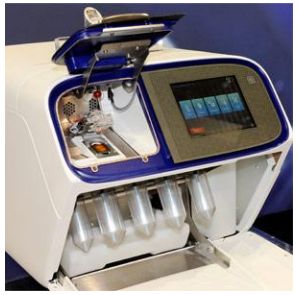
NEXT GENERATION SEQUENCING PLATFORMS

Roche FLX Titanium Plus



Life Tech - SOLiD 5000 XL

Illumina HiSEQ 2500

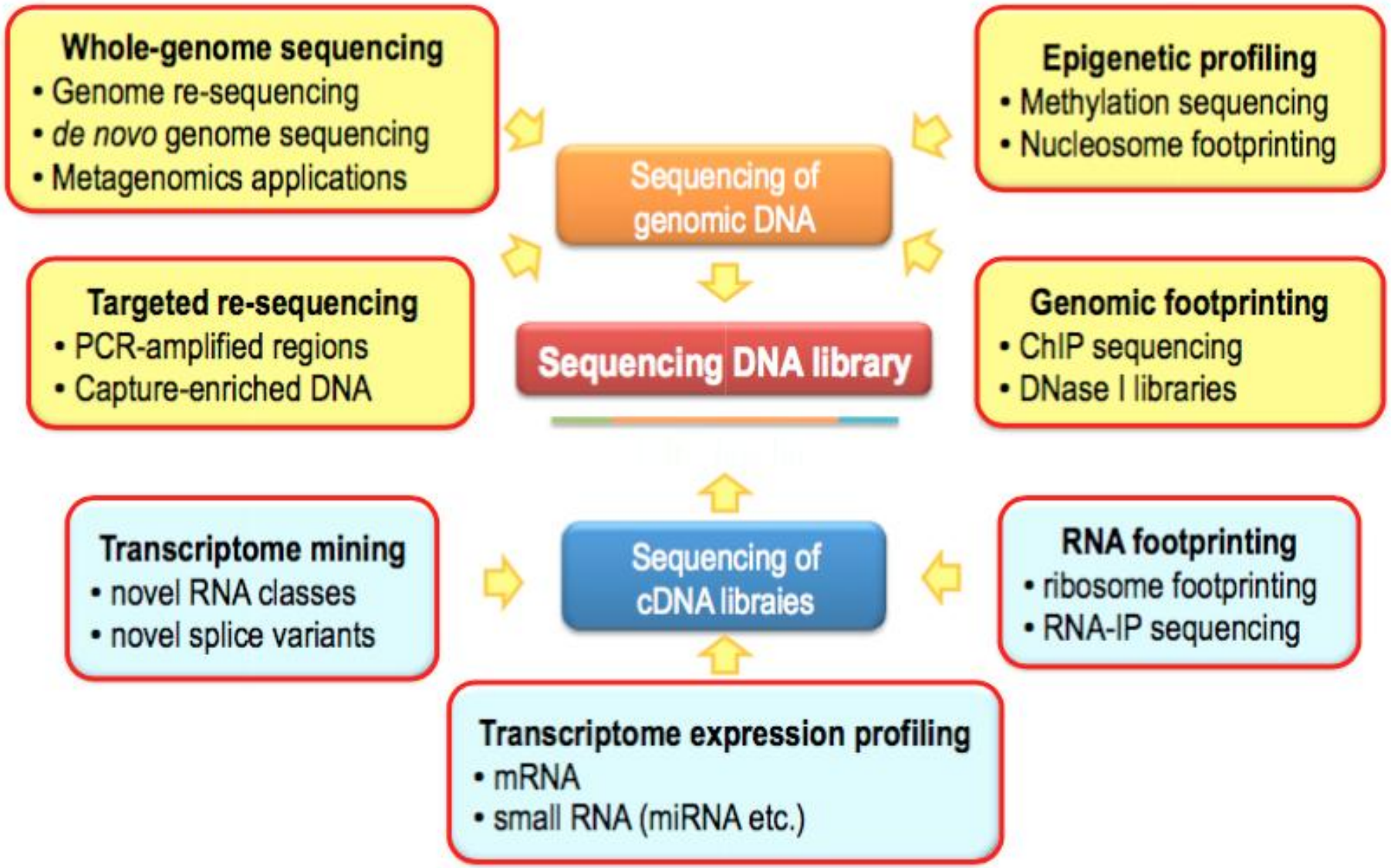


Life Tech - Ion Proton

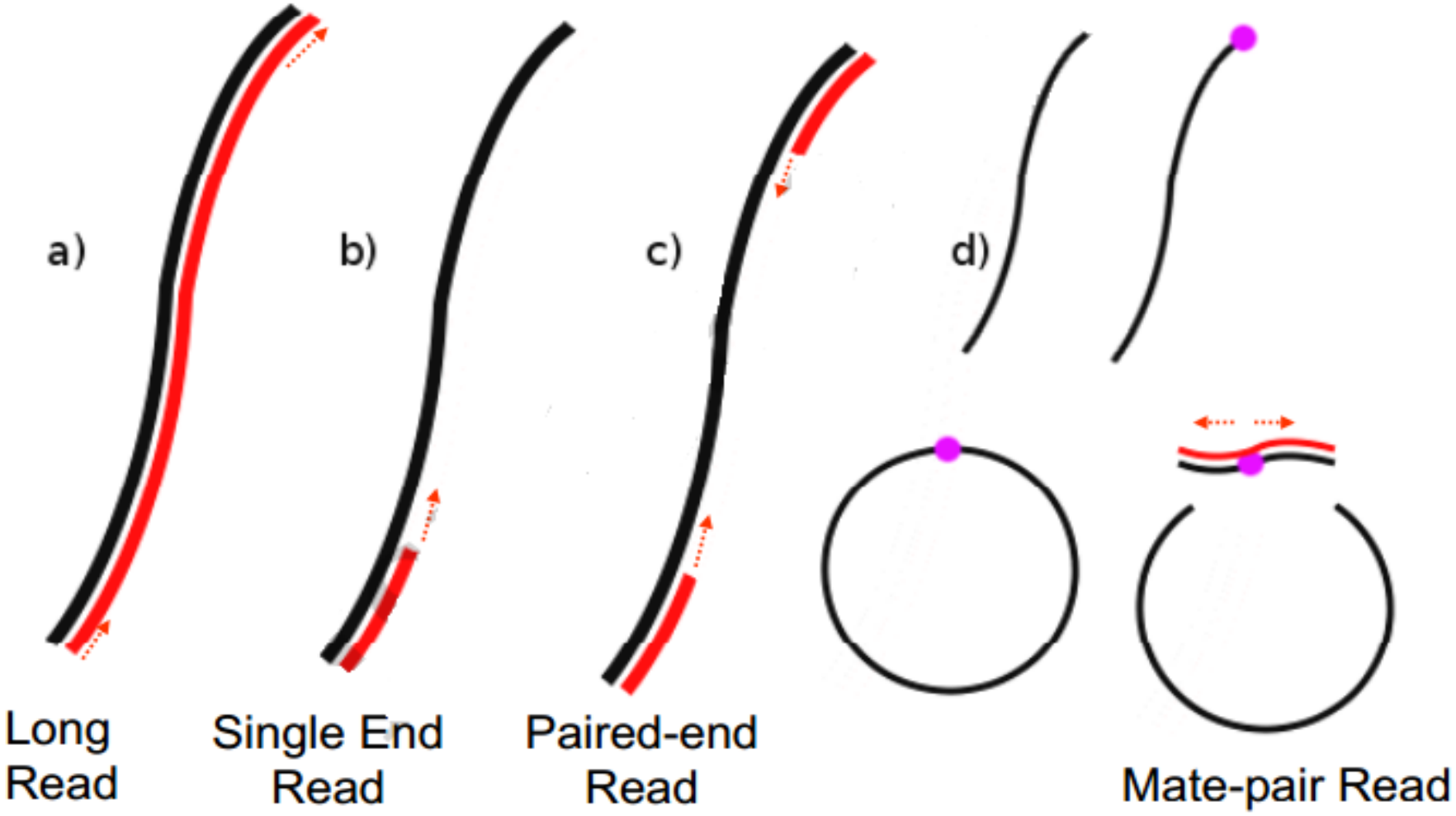
PacBIO RS

Oxford Nanopore

NEXT GENERATION SEQUENCING - APPLICATIONS

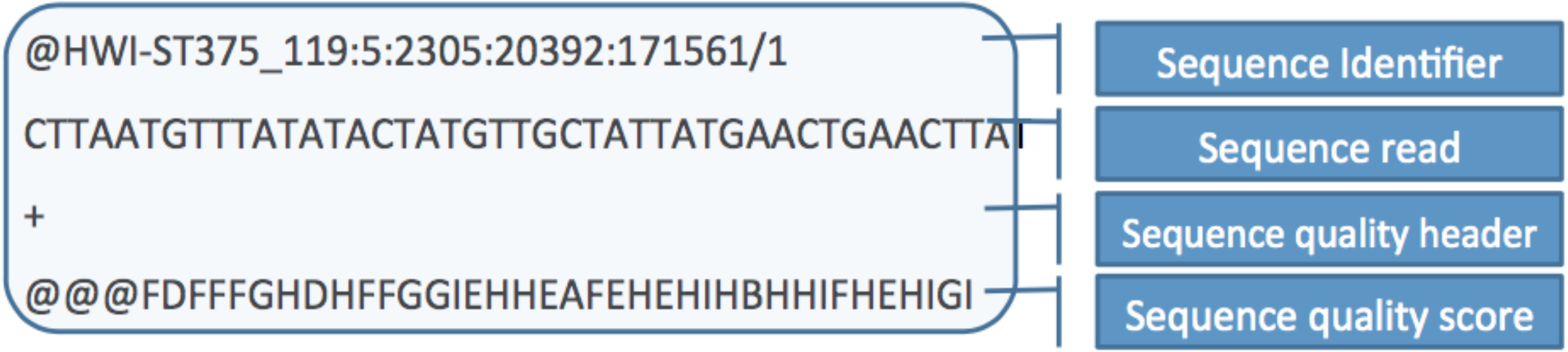


SEQUENCING MODULES IN PRACTICE



UNDERSTANDING NGS RAW DATA

FASTQ format is a text-based format for storing a biological sequence and its corresponding quality scores. It has become the standard for storing the output of high throughput sequencing instruments.



IMPACT OF SEQUENCING ERRORS

The costs of poor data quality

Anders Haug, Frederik Zachariassen, Dennis van Liempd
University of Southern Denmark (DENMARK)

Received August 2010
Accepted January 2011

Haug, A., Zachariassen, F., & van Liempd, D. (2011). The cost of poor data quality. *Journal of Management*, 4(2), 168-193. doi:10.3926/jiem.2011.v4n2.p168-193

ORIGINAL RESEARCH

The Role of Quality Control in Targeted Next-generation Sequencing Library Preparation

Rouven Nietsch[1, a, #], Jan Haas[1, 2, b, #], Alan Lai[1, c], Daniel Oehler[1, 2, d], Stefan Mester[1, 2, e], Karen S. Frese[1, 2, f], Farbod Sedaghat-Hamedani[1, 2, g], Elham Kayvanpour[1, 2, h], Andreas Keller[3, i], Benjamin Meder[1, 2, j]

Rapid evaluation and quality control of next generation sequencing data with FaQCs

Chien-Chi Lo and Patrick S G Chain

BMC Bioinformatics 2014 15:366 | DOI: 10.1186/s12859-014-0366-2 | © Lo and Chain; licensee BioMed Central Ltd. 2014
Received: 3 July 2014 | Accepted: 29 October 2014 | Published: 19 November 2014

Characterizing and measuring bias in sequence data

Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum and David B Jaffe

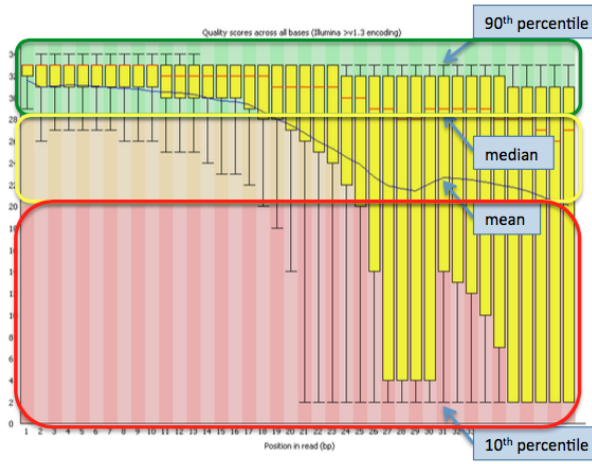
Genome Biology 2013 14:R51 | DOI: 10.1186/gb-2013-14-5-r51 | © Ross et al.; licensee BioMed Central Ltd. 2013
Received: 11 December 2012 | Accepted: 29 May 2013 | Published: 29 May 2013

Error types and rates in DNA sequencing

Errors of the data acquisition process

The DNA sequence gathered through experimental process is gained through an examination of the fluorescent-dye intensity signal that is output by automatic sequencing machines. Even with the newest generation of sequencers, raw sequence data obtained from them is - by all means - everything but trustworthy in its entirety. Inevitable artifacts degrade the quality of the sequences obtained and are caused by experimental as well as systematic factors. Chromatography is a chemical process and thus subject to stochastic and non-stochastic oscillations, which can cause sub-optimal signal quality. Errors in a determined DNA sequence can be caused by flaws in the translation operations of the electrophoresis signal or quirks that arose during the experiment itself. This becomes visible in the wide diversity of data that is obtained even when using a single chemistry type, let alone different ones: under- and over- oscillations of the signals, unseparated curves (compression artefacts), and signal peaks or dropouts are frequent. Incorrect signal analysis raises errors in the base calling process of the signals and constitutes a limiting factor in the automation of assembly processes.

MAJOR QUALITY CONTROL MATRICES



Very good quality

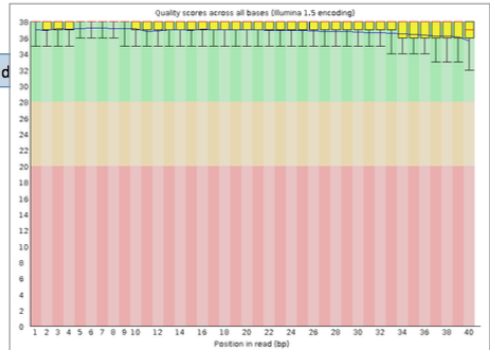
Reasonable quality

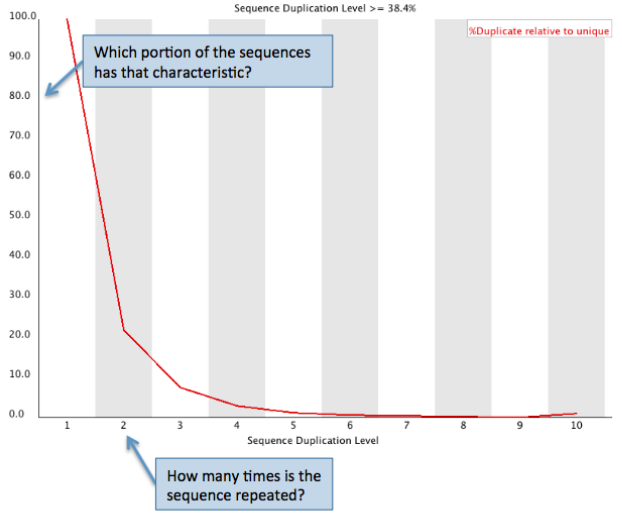
Poor quality

PER BASE SEQUENCE QUALITY

ACCTGGGATCAAACATTCAGGACATATAGCACAATAGGAC

A very good quality sequence

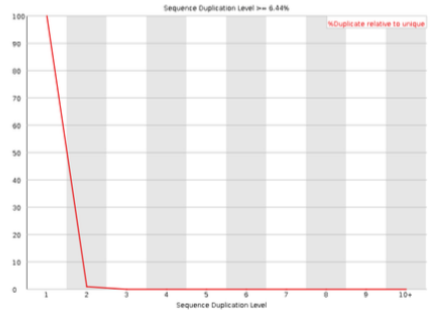




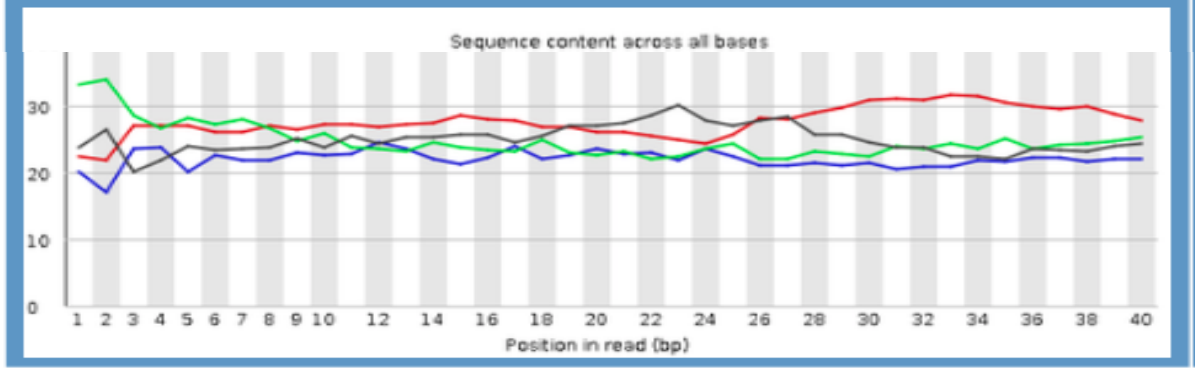
DUPLICATION LEVELS

Computed for the first 200'000 reads gives an **overall impression** for the duplicate levels in the whole file

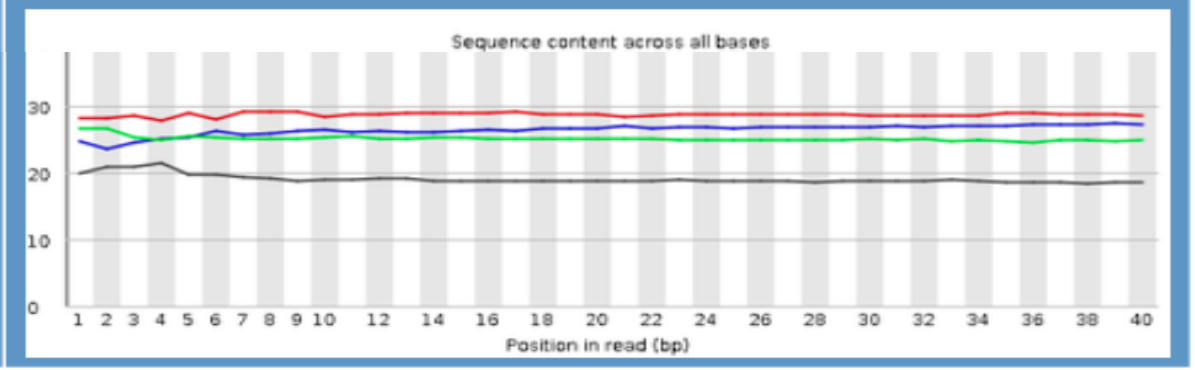
A very good quality sequence



Poor Quality Data Set



Good Quality Data Set



Whole Genome Sequencing – POINTS TO CONSIDER

	Re-Sequencing Approach	De-Novo Sequencing Approach
Sequencing Strategy	Paired End Sequencing	Standard Approach: Paired End Hybrid Approach : Paired End, Mate Paired & Long Read
Sequencing Platform	Illumina Hi-Seq (Deep Sequencing)	Illumina HI-Seq & PacBio RS II
Read Length	100 bp or 150 bp	100 bp & 5kb to 20kb Long Reads
Sequencing Depth	~ 30 x	100 x – 150 x
Expected Coverage	90 – 95 %	80 – 85 %
Supported Analysis	Genetic Variations, Structural Variations, Exome CNV	Draft Genome Construction, Repeat Identification / Classification, Major Genomic Elements Characterization, Coding Gene identification & Characterization

Whole RNAome Sequencing – POINTS TO CONSIDER

	Re-Sequencing Approach	De-Novo Sequencing Approach
Sequencing Strategy	Paired End Sequencing For LongRNA Single End Sequencing For SmallRNA **NOTE:** Replicates are Mandatory	Paired End Sequencing For LongRNA Single End Sequencing For SmallRNA **NOTE:** Replicates are Mandatory
Sequencing Platform	Illumina Hi-Seq	Illumina HI-Seq PacBio Iso-Seq (For Full Length Isoform Sequencing)
Read Length	100 bp or 150 bp For LongRNA 50 bp For Small RNA	100 bp or 150 bp For Long RNA 50 bp For Small RNA ~1Kb - 3Kb Using PacBio Iso Seq
Sequencing Depth	~70x 25-30 Million Reads For LongRNA 8-12 Million Reads For SmallRNA	~100x 70-80 Million Reads / Per Tissue For LongRNA 8-12 Million Reads For SmallRNA **NOTE:** SD between Samples should not be > 20%
Expected Coverage	90 – 95 %	80 – 85 %
Supported Analysis	RNAome Profiling, Expression & Differential Expression, Fusion Gene, Significant Biology Analysis, *mRNA:miRNA Integrome Analysis*	LongRNA Assembly (**De-Bruijn Graph & OLC Based**), RNAome Profiling, Expression & Differential Expression, Fusion Gene, Significant Biology Analysis, *mRNA:miRNA Integrome Analysis*

ChIP Sequencing – POINTS TO CONSIDER

	Re-Sequencing Approach
Sequencing Strategy	Single End Sequencing (Preferred) **NOTE:** Input / Control is MANDATE for every batch of Ip DNA Very low quantity of Ip DNA is always a challenge
Sequencing Platform	Illumina Hi-Seq
Read Length	50 bp (Preferred)
Sequencing Depth	~30 Million Reads For Large genomes ~15 Million Reads Fly / Worm genomes **NOTE:** Biological Replicates are mandatory statistical significance **NOTE:** SD between Input vs Ip DNA Samples should not be > 20%. Reads Input >= Read Ip
Expected Coverage	90 – 95 %
Supported Analysis	Integration with gene expression data, Comparison with other ChIP experiments from ENCODE / MODCODE consortiums etc.

Metagenome (16s RNA) Sequencing – POINTS TO CONSIDER

	Re-Sequencing Approach
Sequencing Strategy	Paired End (For Illumina Plaform) Single End (For Ion Torrent / PacBio Platform)
Sequencing Platform	Illumina Hi-Seq OR PacBio OR Ion Proton
Read Length	150 / 300 bp Paired End, 100 bp – 500 bp Single End
Sequencing Depth	For V3 region: ~1 million read per sample For V3-V4 region: >1 million reads per sample **NOTE:** SD between Samples should not be > 20%
Expected Coverage	90 – 95 %
Supported Analysis	Microbial Biodiversity analysis, Rarefaction curves, Alpha and Beta Diversity Analysis etc

We are here for you !