BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Quality Control of Next Generation Sequence Data

January 17, 2018

Kane Tse, Assistant Bioinformatics Coordinator
Canada’s Michael Smith Genome Sciences Centre
BC Cancer Agency

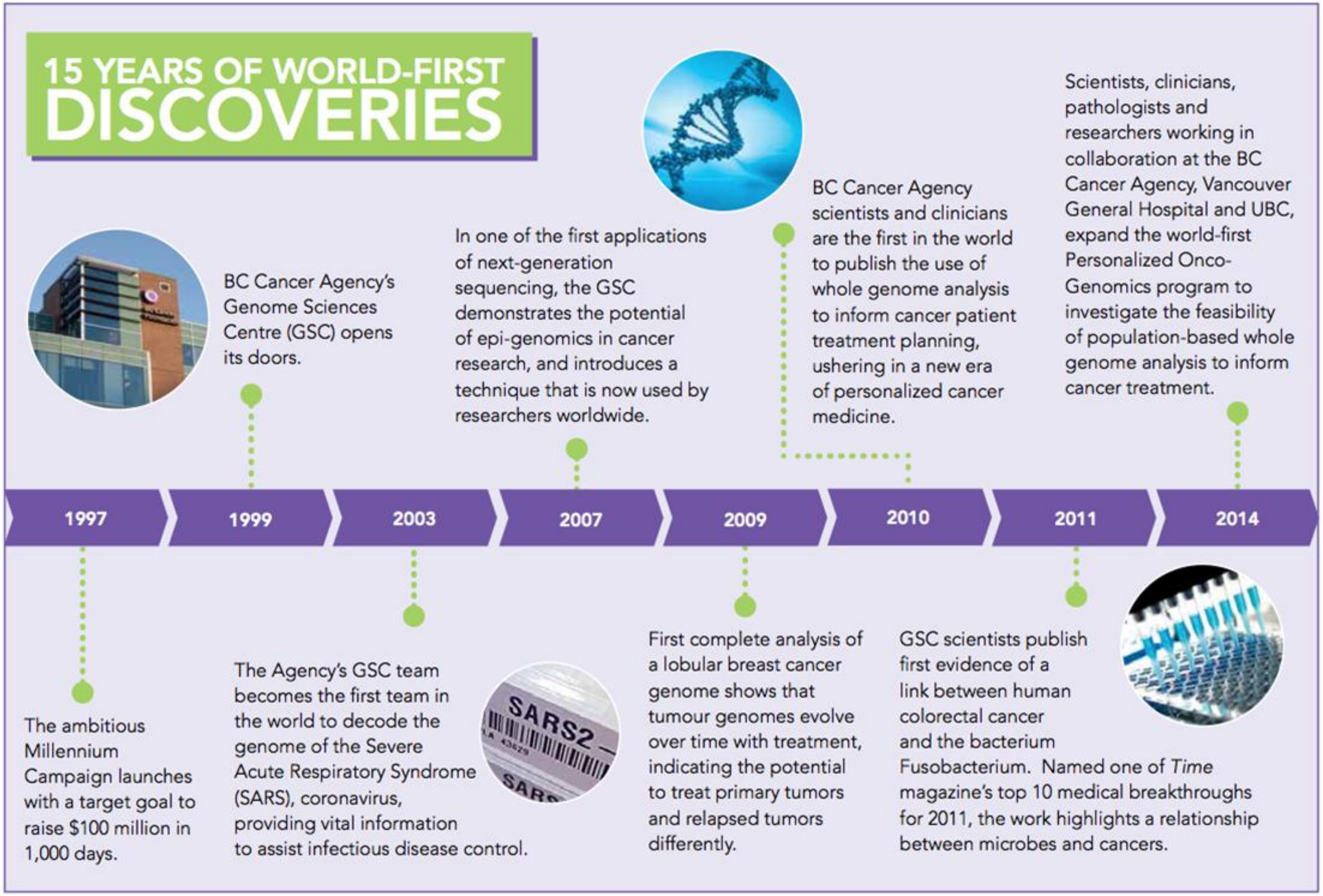
Canada's Michael Smith Genome Sciences Centre

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

BCGSC Overview

Genomics & Bioinformatics Research Centre

- Part of the Cancer Research Centre of the BC Cancer Agency



15 YEARS OF WORLD-FIRST DISCOVERIES

1997

The ambitious Millennium Campaign launches with a target goal to raise $100 million in 1,000 days.

1999

BC Cancer Agency's Genome Sciences Centre (GSC) opens its doors.

2003

The Agency's GSC team becomes the first team in the world to decode the genome of the Severe Acute Respiratory Syndrome (SARS), coronavirus, providing vital information to assist infectious disease control.

2007

In one of the first applications of next-generation sequencing, the GSC demonstrates the potential of epi-genomics in cancer research, and introduces a technique that is now used by researchers worldwide.

2009

First complete analysis of a lobular breast cancer genome shows that tumour genomes evolve over time with treatment, indicating the potential to treat primary tumors and relapsed tumors differently.

2010

BC Cancer Agency scientists and clinicians are the first in the world to publish the use of whole genome analysis to inform cancer patient treatment planning, ushering in a new era of personalized cancer medicine.

2011

GSC scientists publish first evidence of a link between human colorectal cancer and the bacterium Fusobacterium. Named one of *Time* magazine's top 10 medical breakthroughs for 2011, the work highlights a relationship between microbes and cancers.

2014

Scientists, clinicians, pathologists and researchers working in collaboration at the BC Cancer Agency, Vancouver General Hospital and UBC, expand the world-first Personalized Onco-Genomics program to investigate the feasibility of population-based whole genome analysis to inform cancer treatment.

BCGSC Overview

Sequencing Platforms

- Illumina sequence-by-synthesis instruments
  - NextSeq, MiSeq, HiSeq 2500, HiSeqX instruments
- Sanger capillary-based sequencing
  - Life 3730 XL
- Monthly
  - 1,500 libraries constructed
  - >80 terabases sequenced

Bioinformatic Analysis

- 3 large-scale compute clusters
  - 800 nodes, 24,000 hyperthreaded cores, 120TB RAM
- Multiple team-specific clusters
  - Ex - BioQC team: 320 cores, 2.5TB RAM
- 20 Petabytes of storage

Overview

Overview

Description

- What is Quality Control?
- How is Quality Control performed?
- Why is Quality Control important to you?

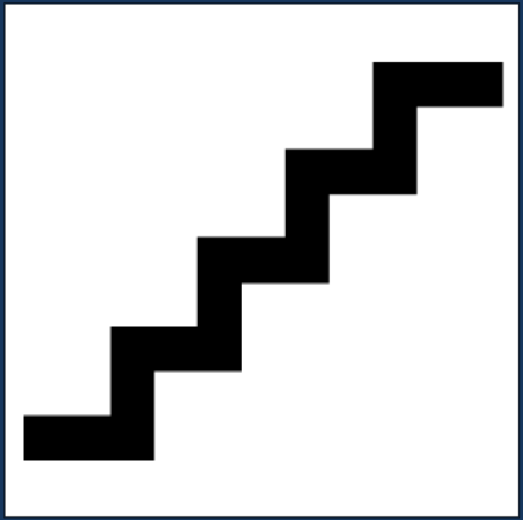
Value

- Quality Control ensures accurate results
- Quality Control can enhance interpretation of results
- Quality Control has scientific merit in publications

Examples

- Unusual cases encountered, and their impact on QC

What is Quality Control



Quality is Important @ BCGSC

If you don't have time to do it right you must have time to do it over. [Unknown]

BCGSC spends time & effort ensuring Quality

- Many teams monitor quality
  - Tend to be manual checks
  - Relies on experience & expertise for detection
- Bioinformatics Quality Control group
  - Automated pipeline to monitor quality and report issues

Why do we care about Quality?

- Identify potential issues before data analysis begins
- Inform collaborators about their experiment
- Improve our laboratory & bioinformatics processes

Definition

qual·i·ty con·trol

/ˈkwälədē kənˈtrōl/

a system of maintaining standards in **manufactured products** by testing a sample of the output against the **specification**.

http://whatis.techtarget.com/definition/quality-control-QC

- **manufactured products** = NGS sequence data
- **specification** = type of experiment (WGS, Capture, miRNA)

Quality Assurance (QA) vs Quality Control (QC)

DETECTION
QUALITY CONTROL
QUALITY ASSURANCE
PREVENTION

ILLUSTRATED BY SEGUE TECHNOLOGIES

Quality Assurance

- Also plays a big role at the BCGSC
  - But not the focus of today’s discussion

Levels of Quality Control

Different Levels of Quality Control

- Level 0: Non-Alignment based metrics
- Level 1: Alignment against a reference genome
- Level 2: Assessment after bioinformatic analysis
  - eg. Variant calling, expression quantification

Levels of Quality Control

	Industry Definitions	Bioinformatic Context
Level 0	• Raw unprocessed data • Directly observed on the instrument • Absolute measurements	• Input = fastq files from sequencer • Indifferent to protocols, regardless of pipeline (WGS, RNAseq, etc.)
Level 1	• Quality Controlled data • Associated with metadata • Compared with calibrations	• Using aligner (BWA or Novoalign) to compare against "standards" (human, mouse reference genomes, etc.) • Mapping rate, dup-rate, paired
Level 2	• Derived products that require scientific & technical interpretation • Standards defined by the community that collects or utilizes the data	• Assembly • Expression levels • Variant calling • On-Target Rate

How is Quality Control performed?

QC Across the GSC

Laboratory QC

- DNA Quantification, Agilent traces, Cluster density, intensity, focus scores, PF rate, Q30/Q20, index splitting

Bioinformatic Level 0 QC

- 60 metrics
  - total_reads, contamination, reagent_leftover, miRNA_adapter…

Bioinformatic Level 1 QC

- Alignment (3):
  - % aligned to genome, % properly paired reads, % duplicate rate…
- ChIP-seq (6):
  - Fraction of reads in peaks (FRiP), domain reads as % of mapped reads..
- Bisulfite-seq (4):
  - Lambda bisulfite conversion rate, human bisulfite conversion rate…
- RNAseq (10)
  - Num Genes Covered @ 1X/10X, Percent reads mitochrondrial, intergenic reads…
- miRNA (2)
  - Num. miRNA reads, Diversity of miRNA species

QC in the Lab

Levels of Quality Control

Pre-Sequencing

- DNA quantification
  - Determine how much DNA is in a sample
- qPCR
  - Determine how many fragments contain Illumina adapters

On Instrument

- First base report
  - Try to detect library issues or machine issues
  - Look for biased libraries from basecalls
  - Review cluster density
- Post-run QC
  - Q30/Q20 scores – contamination of cleavage mix, temperature of instrument
  - Index splitting – uneven pooling, unknown indices

Level 0 QC

Why level 0?

It’s Fast

- QC all lanes within 24 hours of sequencing
- Rapid feedback to the lab on go/no-go for subsequent lanes

It’s Universal

- Works regardless of protocol or sequencing method
- Detects reagents, spike-ins
- Scan & optionally remove microbial genomes

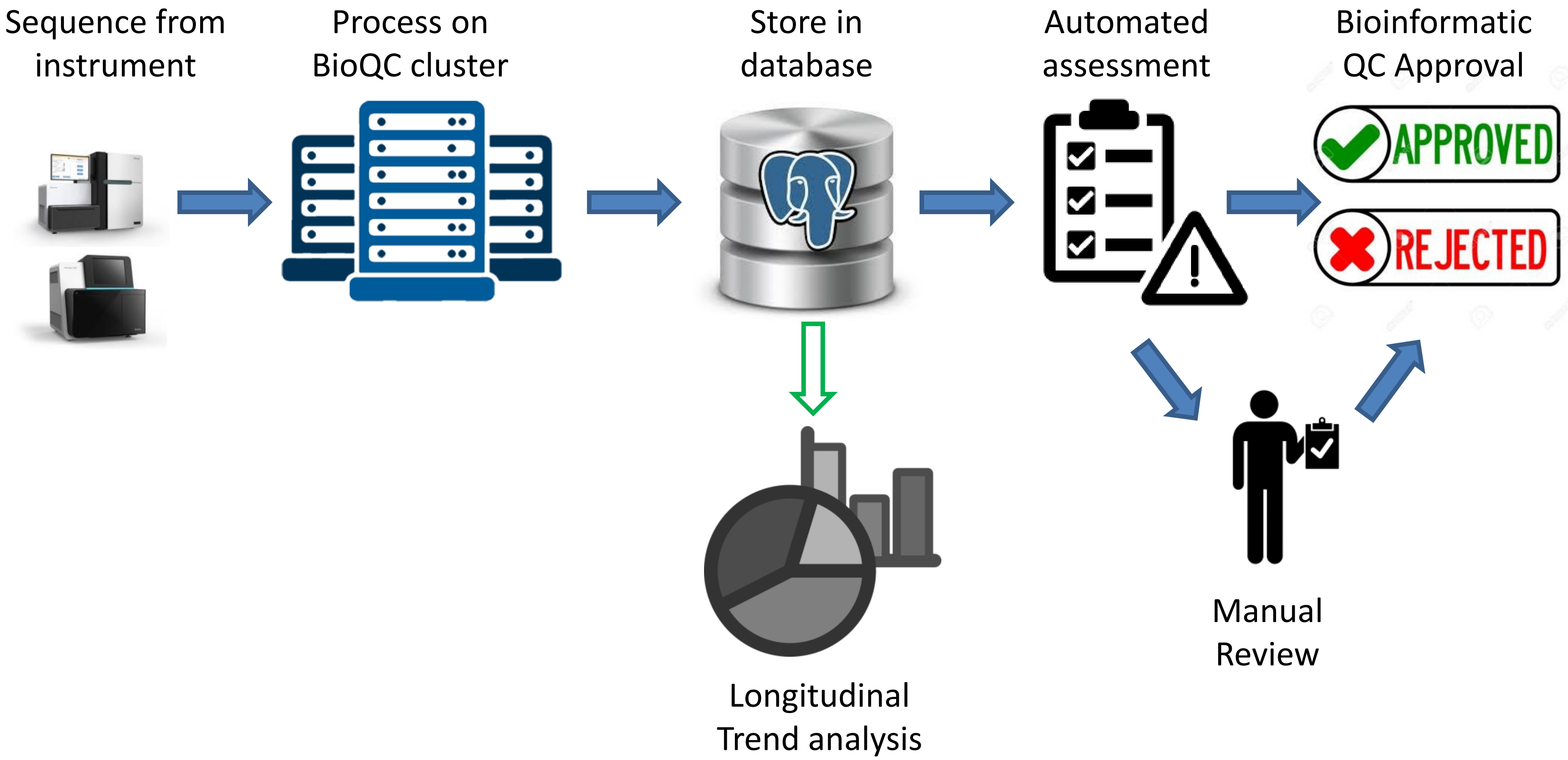
It’s Consistent

- Metrics are generated and loaded automatically into a DB
- Forms a basis for historical comparison & trend analysis

BioQC Pipeline

Every lane analyzed for a standard set of metrics

- Some metrics used for pass/fail assessment
- All metrics stored in a database for historical comparison



Sequence from instrument

Process on BioQC cluster

Store in database

Automated assessment

Bioinformatic QC Approval

Longitudinal Trend analysis

Manual Review

What can you look for without alignment?

Reagent content

- Detect sequences that contain adapters, vectors, standards, ladders

Microbial Contamination

- Use read classification tools like BioBloomTool (BBT) to detect specific microbial contaminants (45 species)

Index splitting & Pooling problems

- Check if the index no-match bin contains a large number of reads
- Check for expected indexes that are missing reads

Sample Swap

- Compare variant calls between samples of same individual
- Look for spike-ins (PhiX or a GSC-specific spike-in)
- Check that the distribution of indices matches what was pooled

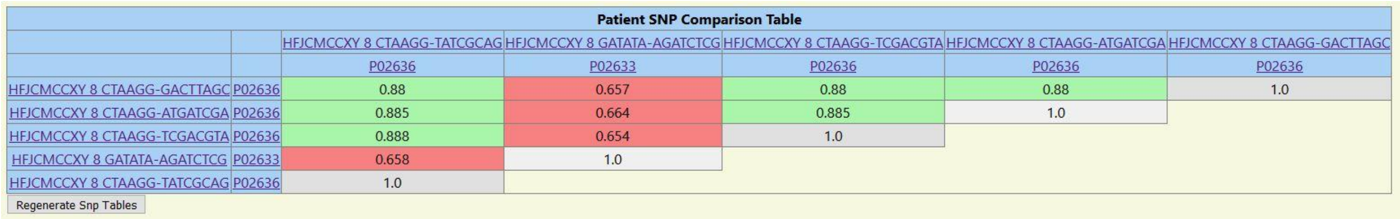
Sample Swaps

Multiple methods of detection

- SNP Concordance (human libraries only)
- Customized spike ins (WGS, RNAseq, amplicon, WGBS)
- Index splitting (for pooled libraries)

SNP concordance

- Bioinformatic implementation of Affy's 500k chip array



Patient SNP Comparison Table

		HFJCMCCXY 8 CTAAGG-TATCGCAG	HFJCMCCXY 8 GATATA-AGATCTCG	HFJCMCCXY 8 CTAAGG-TCGACGTA	HFJCMCCXY 8 CTAAGG-ATGATCGA	HFJCMCCXY 8 CTAAGG-GACTTAGC
		P02636	P02633	P02636	P02636	P02636
HFJCMCCXY 8 CTAAGG-GACTTAGC	P02636	0.88	0.657	0.88	0.88	1.0
HFJCMCCXY 8 CTAAGG-ATGATCGA	P02636	0.885	0.664	0.885	1.0	
HFJCMCCXY 8 CTAAGG-TCGACGTA	P02636	0.888	0.654	1.0		
HFJCMCCXY 8 GATATA-AGATCTCG	P02633	0.658	1.0			
HFJCMCCXY 8 CTAAGG-TATCGCAG	P02636	1.0				

Spike Ins

- Add 200bp oligos into each sample at tiny amounts
- Detect those oligos in sequenced data (~10,000 reads)

Categories of QC metrics

Sequencing Quality

- Adapters, reagents, dimers
- Duplicate rate
- Contamination
- Read quality
- PF Rate (Chastity Passed)
- Coverage

Success of Laboratory Processes

- Bisulfite conversion rate
- Pooling efficiency
- ChIP capture efficiency
- On-target read rate (specific capture)
- Mitochrondrial or rRNA content

Sample Degradation

- RNA degradation
- Fragment size

Historical Comparison

- Lane to lane comparison

Sample Identity

- Plasmid Spike Ins
- SNP concordance
- Index splitting

Gene Complexity and Library Diversity

- miRNA diversity
- # of Genes detected
- Intergenic content
- Intron-Exon Ratio

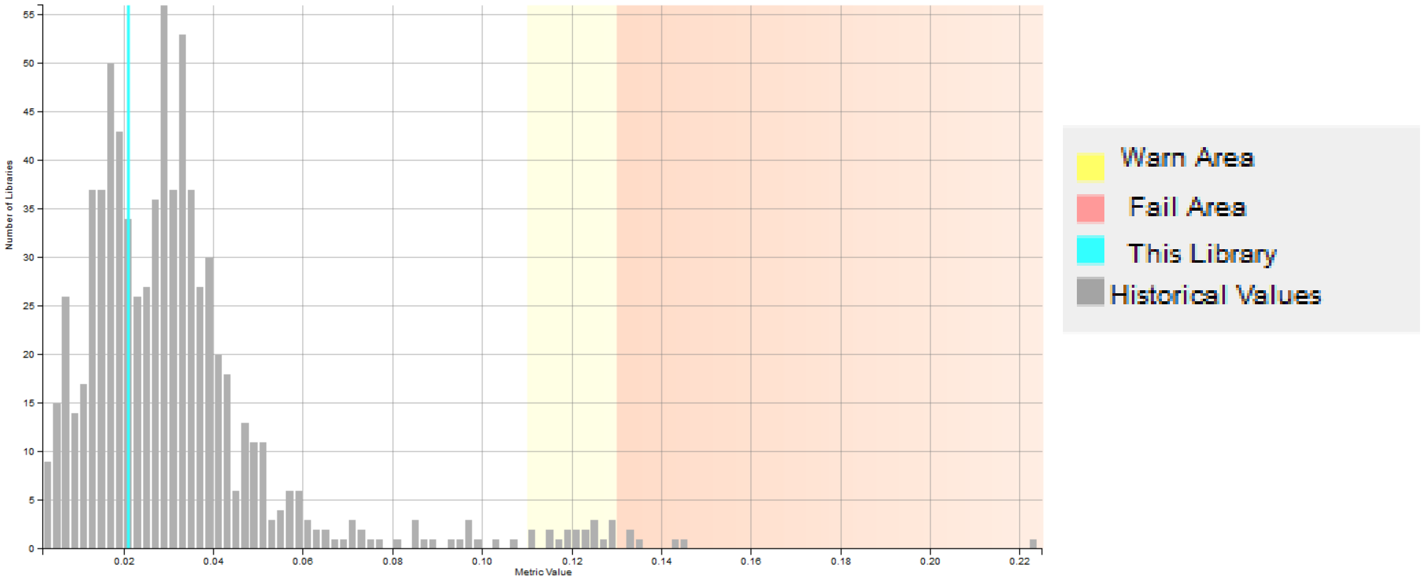
Deciding on a pass/fail

Lab metrics

- Generated on-instrument, manually evaluated based on experience

Bioinformatic metrics

- From a population of libraries (minimum 50 runs)
  - determine 95th (warning) and 99th (fail) percentile



Metrics and Thresholds

Metric

- A measured or calculated characteristic of a library

Threshold

- A value at which a library is to be assessed for quality

Not all metrics have thresholds

- Metrics that do not thresholds:
  - Read count; expected spike-in observed
- Metrics that have thresholds:
  - Reagent leftover, contamination rate, alignment rate

Hard Thresholds vs Outliers

Hard Threshold

- Absolute point at which a library must be failed
- Indicates something has gone severely wrong
- Examples:
  - Very low alignment rate (<60%)
  - Very high contamination (>50%)

Outliers

- Metric beyond the 95th percentile of historical BCGSC data
- Contains usable data, but less than ideal
- Examples:
  - Low quality/low input material
  - Slightly lower genomic coverage
- BCGSC will manually review every library with 3 outlier metrics

Why is QC Important?

Organizational Benefits of QC

How QC is useful to your processes

1. Confirm sample identity
   - Swaps or contamination events

2. Detect problems with laboratory processes
   - Uneven pooling, high ribosomal RNA content

3. To make improvements to protocols
   - How does a new protocol compare to the old version?

4. To compare results to previous experiments
   - Batch effects over time
   - Are additional lanes needed? How many?

5. To reduce costs
   - Avoid analyzing bad data and integrating results into existing data

Scientific Benefits of QC

How QC is useful to your science

- As a QC gate
  - Prevent bad data from being incorporated into an analysis
  - Sample swaps, low library diversity
- To identify outliers
  - Samples that have known issues that may affect analysis results
  - Explains observations in data when publishing results
- To perform trend analysis
  - Look at results over time
  - Provides a baseline by experiment type for comparison
  - Identify areas of optimization in lab & bioinformatic pipelines

Examples

Simple Examples: Index Splitting

Expected Indices	Observed Indices
TCCCGA	22%
ATCACG	26%
CTAGCT	24%
TGACCA	**0%**
No match	28%

Example 1:

- Conclusion – Incorrect 4th index specified

- Additional analysis – Examine no match bin
  - Infer missing index sequence based from most frequently observed index

Simple Examples: Index Splitting

Expected Indices	Library A Lane #1	Library A Lane #2	Library A Lane #3
TCCCGA	24%	**0%**	23%
ATCACG	20%	**3%**	18%
CTAGCT	23%	**35%**	28%
TGACCA	33%	**10%**	31%
No match	0%	**52%**	0%

Example 2:

- Conclusion – Lane #2 has been swapped with some other lane

Biological Exceptions – failing QC but the data is still usable

FFPE Samples

- Degraded DNA means PCR amplification was needed
- Higher duplicate read rate

Amplicon Libraries

•If amplicon sizes are small, high amounts of adapter are detected via read-through of fragment

Metagenomic Studies & Xenograft Libraries

•Alignment rate to a single target species may be low, but doesn’t mean the data is bad

Low Input Libraries

•Frequently see higher background, lower fragment diversity

Safety Checks – when failing QC is a reason to stop

Low alignment rate

- BWA-aln works poorly on reads >125bp, use BWA-MEM
- Aligned to the wrong reference genome

Sample swaps

- Don't want to publish/analyze data for the wrong sample

Low bisulfite conversion rate of lambda phage

- Conversion reaction not done completely in lab

Genomic Contamination

- RNAseq library contains too much genomic DNA
  - Might affect observed expression levels

Conclusion

Conclusion

What is Quality Control

- 3 levels of QC

How QC is Carried out at the BCGSC

- Laboratory
- Automated Bioinformatic QC Pipeline
  - Role of manual review
- Some data that fails QC can sometimes be used

How is QC Useful

- Saves time in data analysis
- Aids in interpretation of data (publication)
- Identifies trends and areas for improvement

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

Acknowledgements

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Genome Sciences Centre

Dr. Marco Marra

Dr. Steven Jones

Dr. Yussanne Ma

Bioinformatics Quality Control

- Eric Chuah
- Irene Li
- Gina Choe
- Dorothy Cheung
- Correy Lim
- Robert Lin

Laboratory Production Teams

- Dr. Andy Mungall
- Dr. Richard Moore
- Michael Mayo

GSC Production Teams

- Library Construction
- Sequencing Group
- LIMS
- BioApps Team
- Software Analysis
- Analysis Pipelines
- Data Analysis
- Systems Group

- Reanne Bowlby

More Information

BCGSC Website: http://www.bcgsc.ca

Email: ktse@bcgsc.ca