



Quality control (QC) metrics for WGS data

Sept 25, 2018

Kim Ng

Martin Basterrechea
Salido

(Statens Serum Institut)

What does it mean to QC data?

- Someone comes to you excited to do a project and says their data is good quality.
 - What does this mean?
 - Do you trust the data?
 - Does it depend on who it is?
 - Is there a platform for standardized QC? Why do you think this is the case?

So what does QC mean to you?

- It depends on the type of question or questions you are asking.
 - I want to know the Genus?
 - I want to know the Species?
 - I want to know the Strain?
 - I want to know if a gene is present?
 - I want to know if this mutation is present on the gene?
- Each of these asks a question requiring different levels of data. More data is required for the later questions to answer and each question also answers the questions above it. Some questions you can never be completely certain to answer, but you can be reasonably sure.

Some answers

- If you're asking the question
 - I want to know the Genus?
 - (a handful of reads is enough for this)
 - I want to know the Species?
 - (more reads are needed to identify non shared components)
 - I want to know the Strain?
 - (more reads which also differentiate the different strains, typically full genes of typing scheme)
 - I want to know if a gene is present?
 - (similar to knowing the strain but certain considerations need to be made for certain genes (core vs accessory))
 - I want to know if this mutation is present on the gene?
 - (more reads to find the full gene in question, and enough depth to identify the variation for that gene)
- What is your goal for the data? (multi faceted)

QC at SSI

- Statens Serum Institut provides national bacterial surveillance for the country of Denmark.
- I want to be able to compare samples to some level which is appropriate for surveillance so that we can use old and new samples without rechecking quality on each project
- Samples can come from multiple sources so having a similar platform makes data comparable without rerunning softwares
- Due to health aspect a sample is often time sensitive so favour sequencing at higher quality once than more samples with some being re-sequenced
- Need to track lots of (outbreak, project, surveillance) samples a year

Goal for my samples

- I want sequences which are contaminant free and enough depth to reliably perform SNP based bacterial surveillance from a bacterial isolate.
- Single isolate samples which are contaminant free from other species and of enough depth to perform surveillance (SNP based) on the samples.
- When in doubt prefer higher quality data vs resequencing samples (due to time constraints, this is also a cost per sample trade-off)
- This comes down to a minimum coverage across the genome and free of contaminants

Important things I'm not going to go over

- Many aspects of how successful your run is is related to Lab prep and Illumina equipment. This stuff can be detected if you already have a baseline of data to compare against. If you do not I recommend contacting Illumina to ensure your metrics are within their operating parameters.
- How many did you remove? Does this correlate with the Reagent kit expectations (<https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html>), if not can try contacting Illumina for suggestions
- Using sequence analysis viewer, checking run based quality and machine quality of lab setup
- Issues which are cross run and not per sample

Data to usable reads/contigs

- Demultiplex the data into it's associated samples (bcl2fastq).
- Check number of reads you get (fastqc)
- Check and filter based on read quality (trim, filter, adapter removal)
- (De-novo) Assemble your contigs from raw reads (SPAdes, CLC, skesa, etc.)
- Contaminant checks (read based: kraken, blast: contig based)
- (Mapping) Map your reads against your reference sequences (BWA-Mem, Minimap 2, etc.)
- Pull your metrics of interest and compare to your values

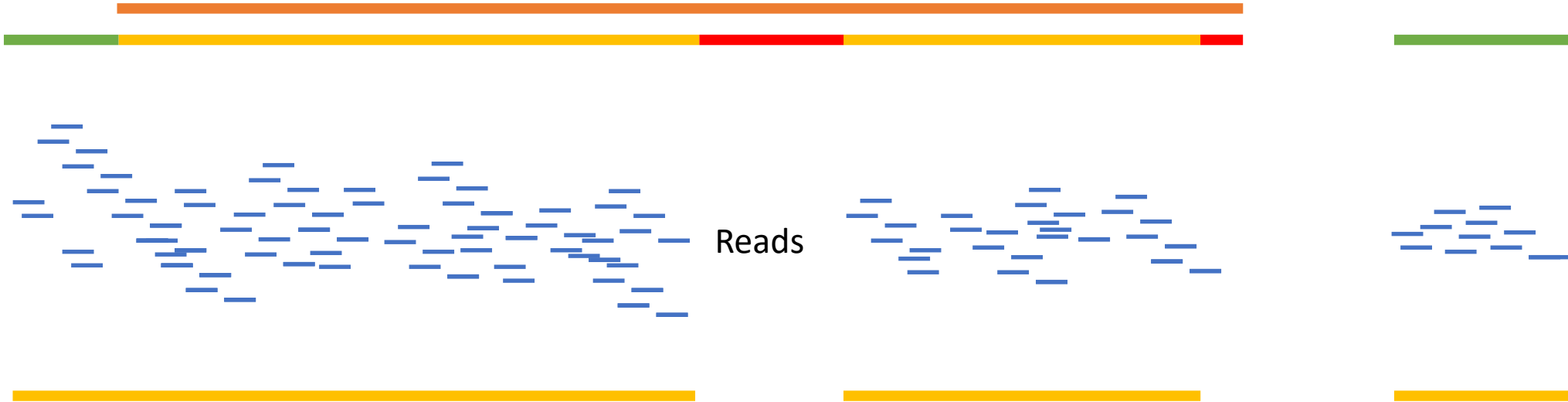
Standard (?) procedure

- Clean data of low quality (either by trimming/filtering or both)
 - Filter reads of low quality
 - Trim data
 - This is not essential just be aware of what you're doing. Can remove bad reads through mapping instead
- Map reads to species of interest and have X (>20) average coverage on this (not high variance)
- Sequence is good to go (generally but doesn't answer a specific question)
- Additional common practice, check contamination (based on user knowledge)

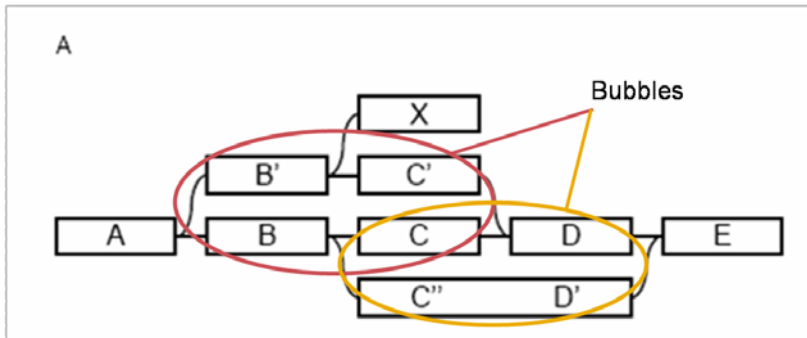
Checking contaminants

- Can only check against what you know meaning well documented Sequences you can be assured on and less documented ones are harder to detect
- What's in the reference database you're using?

Reference based assembly



De-novo based assembly





Reference assembly

- Pros
 - Relative to something you know about, ideally what you're doing surveillance against
 - Higher certainty with results as mapping is typically better than denovo
 - Tells you what reads you have and what you don't
 - Ignores reads that are based on a contaminant
- Cons
 - Doesn't tell you about reads that don't map to reference
 - Doesn't handle events such as recombination or large insertions
 - Reference may not be appropriate
- Suggestion
 - When you want to know what's present and not present against a known reference you should always prefer a reference assembly

De-novo assembly

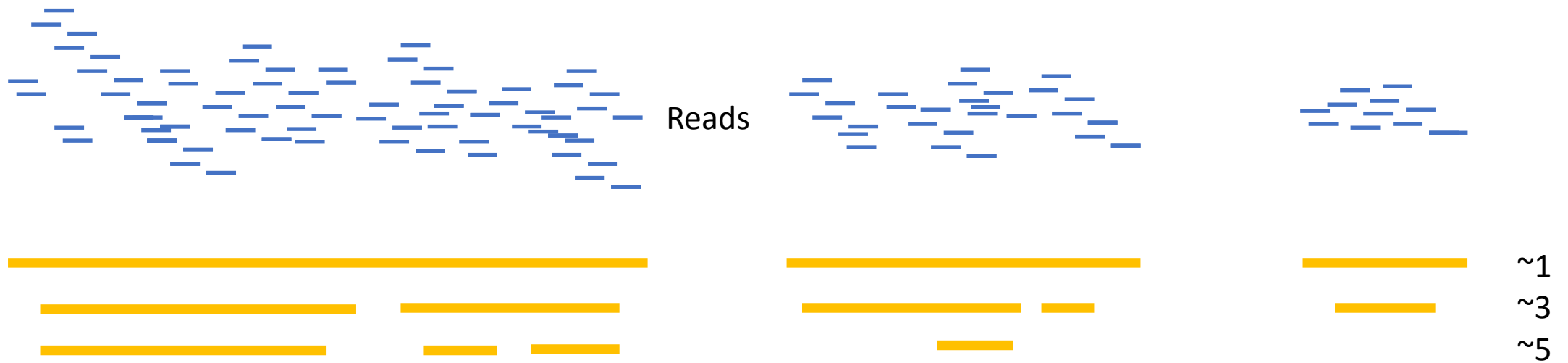
- Pros
 - Provides contigs based on the material you have
 - Results even if no similar reference is available
 - Can provide information on recombination events and incorporated elements
- Cons
 - Can't be as sure as comparing against a reference for presence/a
 - Can be hard to resolve overlapping segments accurately
 - Have to handle contaminants
- Suggestion
 - De-novo assembly provides additional information regarding the material provided which has shortcomings on what's not present but can be applied generally, especially when a reference is not desirable. Useful for finding genes not on reference (e.g. resistance, virulence)

Common metrics

- N50: is the minimum contig length needed to cover 50% of all contigs total length. e.g. if your genome has 4 contigs of 1,2,3 and 5 kb (total length 11kb) you'd need genomes 5kb and 3kb to get past half the genome size. The minimum contig length or N50 is then 3kb. (species and software dependent)
- Number of contigs (species and software dependent)
- Total contig length (species and software dependent)
- Average coverage (Typically $<10x$ the data needs more depth and $>25-30$ should be good)

Useful metrics (for myself)

- Minimum coverage (10x)
- Sizes at various minimum coverages



Sources of error

- Lab bias
- Software bias

Lab bias

- How your lab manages the DNA will determine the quality of the material going in which will have an impact on the result, changing different parameters will cause differences such as less gaps, higher/lower coverage, favouritism to different aspects.
- The isolates you get have their own quirks (e.g. GC) and biases.
- This can make comparisons at start challenging as you need a reference to compare against

Software bias

- Each program while similar in function operates differently. The metrics you pull will be relative to the software's used. In the pipeline. Comparing results from two different pipelines is more challenging because of this.
- Software vary wildly in terms of run time and results. Each makes different decisions on items such as which mapping algorithm to use. Results are similar in the end (though can be challenging to compare) but run time can vary in orders of magnitude.
- Can't compare what you don't track, but data comes at a cost
- Can check against existing data (online)

Other things to be aware of

- Sequencing platform of choice and issues, e.g. repetitive regions
- Easier to be certain something is present than absent
- If possible have some repetition of common samples to check if there's changes either in Lab or Software being applied

Some unmentioned annoyances

- Format converting
 - sam->bam->vcf
 - fastq->fasta
 - Bad headers, ideally (use American [a-zA-Z0-9] if you can, certain special characters will have issues with certain programs so try to avoid them when possible, some programs will work others will not)

What we run (now)

- 2 QC pipelines in parallel, other steps also run on both:
 - Fastqc, sendsketch (bbmap), quast, stats (bbmap), pileup (bbmap), custom scripts (variant summarizing), various file/directory manipulations (samtools)

QCquickie	Assemblatron
1. bbduk Filter reads (Q14) trim on right	1. bbduk Filter reads (Q14) trim on right
2. bbmerge merge overlapping reads for assembly	2. n/a
3. tadpole assemble reads	3. skesa/spades assemble reads
4. bbmap map reads to assemblies	4. minimap2 map reads to assemblies
5. callvariants on mapped reads	5. callvariants on mapped reads
6. n/a	6. prokka annotate contigs

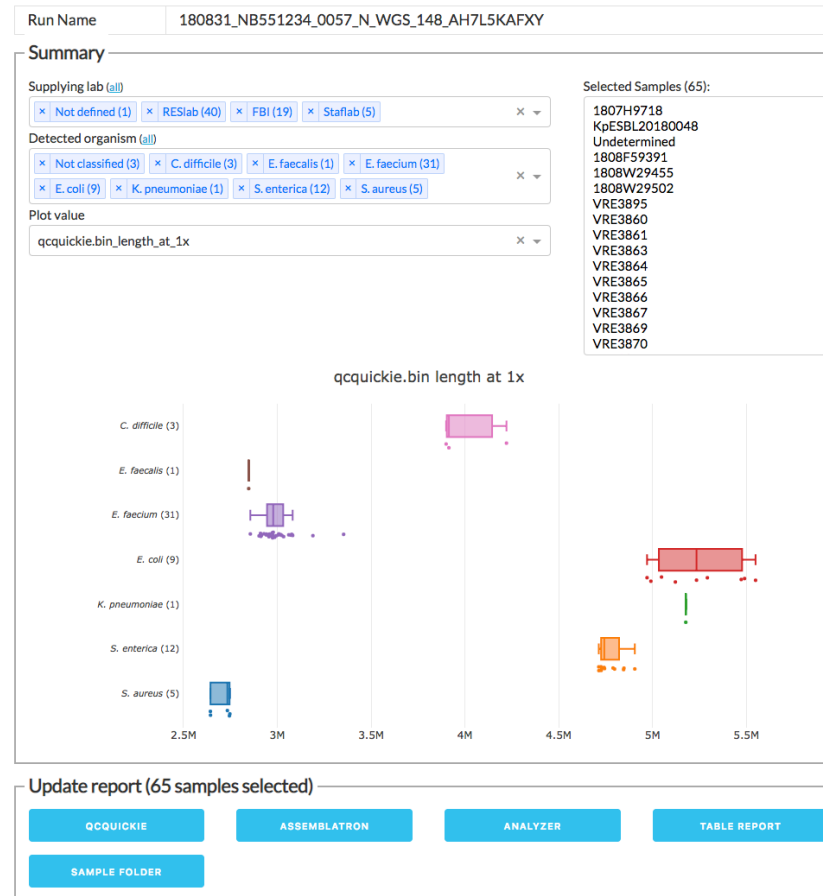
Why run two? (don't worry about this)

- To correlate new data when you replace your pipeline with a different one
- Values for # of contigs, size of contigs, GC, N50, coverage, num of reads, filtered reads, snp filter, indels, some values at different minimum coverages

Extras

- Lots of computationally cheap things to check with data afterwards which are not directly related to QC but can be used to compare
 - MLST
 - Resistance
 - Virulence
 - Your gene of interest with DB
- With contigs this can be done exceptionally quickly (<1 min per sample)

Demo of our existing system



Recommended systems

- Depending on level of expertise (and if you have access to a server)
- Basic
 - CLC Genomics
 - Geneious Prime
- Advanced
 - <https://github.com/tseemann/nullarbor>
 - <https://github.com/ssi-dk/bifrost> (our system, link doesn't work yet)
- Challenge
 - It is likely some manipulation of values is required in order to get everything working. This can be challenging depending on the platform to get everything in an appropriate format to proceed

Relax

- If your data is hitting some common metrics at high levels you're likely to find what you're looking for regardless of process. Its biological data so answers are normally able to be found through multiple approaches
- Be aware though that results are species dependent, more commonly described species are easier to compare as the issues are already known.
- Answers are always harder at the start because you need sequences from your lab to compare against
- Use your expertise on the material to see if it makes sense
- Lots of people have done this stuff before so ask for help
- If something is really wrong you'll know it when you try to do downstream analysis most of the time (you just spend more time to do this then)

Things to think about

- What is the question you're trying to answer with your sample?
Multiple questions can be answered by a single sample.
- How will past samples affect what you're expecting to see.
- If something is “bad” what would you do to adjust it?
- If a better process emerges how will you try and benchmark the differences?
- Keep track of your samples in a meaningful way to how you want to use them