requirements

Title: Immediate Quality Control Method for Next Generation Sequencing Instruments

Date: 30th August 2015

# Introduction and background

The project was initiated as a collaborative venture between the All Wales Medical Genetics Service (AWMGS), a regional diagnostic genetics laboratory (NHS), and the Wales Gene Park (WGP) of the University of Cardiff.

A need was identified for a rapid read-out of quality after Next-Generation Sequencing (NGS) to rapidly and automatically identify failed samples or runs, enabling repeats to be initiated as quickly as possible. Currently post-run quality checking requires the operator to manually check quality metrics generated by the sequencing instrument using a manufacturer-supplied software application (Sequencing Analysis Viewer)(1). Operators are often rushed to complete other tasks and this step is sometimes forgotten, which can lead to poor quality sequencing not being discovered for some time. Downstream analyses may also have been performed on the data prior to identification of low-quality sequencing, wasting time and resources. Fast identification of failed sequencing is of particular importance in the diagnostic setting as delays to obtaining quality sequencing lead to delays in reporting and in the patient receiving their result. This can have important implications not only in clinical management but also in patient well-being, as uncertainty associated with waiting for a test result can be detrimental to mental health(2–4). As the sequencers are shared use between the AWMGS and the WGP the project was initiated as a collaborative venture led by the NHS, with the intention of sharing results with the WGP.

A further need was identified for a method to monitor instrument performance over time, particularly to identify any troubling patterns which may indicate underlying instrument defects or failure. Although this falls outside of the project remit at this time, the process of data gathering which will be undertaken as part of this project should facilitate development of such a method in the future.

Project Lead: Sara Rey

Advisers: NHS bioinformatician, Genetic Technologists, University bioinformaticians.

Main Stakeholders: Lead Genetic Technologist for Next Generation Sequencing section of the laboratory, Genetic Technologists working within this section, University laboratory technicians, Bioinformaticians, for secondary uses of the data.

# scope and purpose

The purpose of the project is to provide a rapid and automated read-out of the quality of sequencing. This is intended to enable immediate assessment of the performance of a run, identify any problems with particular samples and provide this information to the relevant people in an accessible and easy-to-understand format.

A secondary goal of the project is to develop a method to store metrics which may be useful indicators of quality for later mining and development of further methods of quality control (QC).

The scope of this project restricts starting data to metrics output by the sequencing instrument in the InterOp folder and the runParameters.xml file (the data required for the SAV application which was previously being used for manual quality control). This is because these files are relatively small (~2MB, compared to a FASTQ which on the MiSeq are typically between 1 and 4GB), which enables data to be obtained and copied rapidly, and easily stored. An additional advantage is that there is no sequencing data with the files, so there is no potentially personally identifiable data and therefore no information governance issues associated with this data. To enable some additional required data to be captured, the sample sheet (SampleSheet.csv) was also included as part of the starting dataset. As this includes a laboratory sample identifier and the sequencing operator’s initials, data are now pseudonymous (linked-anonymous) rather than anonymous. However, as the rest of the data captured relates only to instrument and sequencing performance and not to the patient, and it is not possible to identify samples or operators without access to NHS databases, the data are still able to be considered as non-personally-identifiable(5). This will facilitate later sharing of information and methodologies developed during this project between the WGP and the AWMGS and potentially between AWMGS and other NHS diagnostic genetics laboratories.

# methods and instrumentation

Prior to generation of this document, information was gathered to identify the requirements of the project and determine the order of priority of implementation of these requirements.

Initially a QC seminar was scheduled, with the topic of “Quality control in NGS”. The chosen format was a series of presentations regarding quality control in NGS were given, to enable all present to understand the steps involved in QC in NGS and importance of quality information, followed by a directed discussion of current issues faced, including ideas for potential solutions and general comments on the existing process of quality control. The directed group discussion, led by bioinformaticians from the AWMGS and the WGP was used to generate some ideas and decide on pre-requisites which would need to be in place for the project to be a success. An example of a pre-requisite identified during this step was the need to share anonymised data for quality control between the University and the NHS; responsibility for this was assumed by one of the WGP bioinformaticians. From this discussion, notes were used to generate a series of bullet points for refinement into requirements.

A meeting between bioinformaticians from the AWMGS, the WGP and other University departments was called at which the ideas generated by the QC seminar and discussion were discussed in detail. From this meeting, further ideas were gathered and ideas generated and documented after the QC seminar were discussed. This meeting also provided a valuable opportunity to discuss potential methods of implementation and the practicalities of this, although this lies outside of the scope of requirements gathering.

Subsequently discussions were held with two genetic technologists and an instrument operator from the University side to gather opinions reflective of the two different environments. The questions asked are outlined below.

* How do you currently check the quality of a sequencing run after the run has finished?
* How soon after a run has finished do you do this step?
* Do you find anything getting in the way of this step, or that doing this is a particular nuisance? If so, why?
* Which markers of quality do you find most useful?
* If you had an automated read-out, would you rather have all the information available or just some? What would you like this read-out to look like?
  + Would you just like to know if a run has failed or would you like to know why it has failed?
  + Would you like the software to tell you what action you need to take?
* Do you think that having an immediate quality control step which is automated will make your work easier?
  + Will it let you get repeats on more quickly?
* Do you think that having all of the quality data in a database would be something that you might want to use in the future?
  + What information would you be most interested in being able to look at?
* Do you have any way of looking over time at how an instrument is performing?
  + Do you think that this would be useful? Why or why not?

Finally, the requirements identified from the above steps were discussed with the lead genetic technologist, after appointment in late 2015, to ensure that the direction that the project was taking was suitable and compatible with the strategic vision for the newly-created NGS section.

It was considered to develop and distribute a questionnaire to all operators of sequencing instruments, and such a document was under development, however, the department then began a process of restructuring, which was predicted to extend into 2016. As the project has a deadline for completion of early 2016, and during the restructuring period there was a series of rapid changes in operators, it was not considered in the best interests of the department to request responses to questionnaires at this time. This is a method which will be explored once the restructuring process is complete, to see if there are any additional requirements which may be added to future projects which build on the results generated by this project.

# limitations, questions and issues

For reasons of practicality it was not possible to involve every user of the sequencers and data from the sequencers in detail, therefore there is the chance that requirements will be limited to opinions from those who were able to attend and willing to comment during the collaborative directed discussion after the QC seminar, to the bioinformaticians who attended the bioinformatician’s meeting and to the sequencing operators who were available to be questioned.

The NHS laboratory restructure resulted in a change to the organisation of the NGS section and regular changes to the genetic technologists who prepared the libraries and ran the sequencing. There were also a few changes to WGP staff during this period. This complicated the requirements gathering process, because, as new sequencing operators came in, they often lacked experience in the process and therefore were not as able to advise on what would be useful to them in terms of quality control. Furthermore, it made consistency in current approaches to QC and future directions difficult to assess during the project planning process.

It has immediately been recognised that there are a huge number of approaches that can be taken to quality control of NGS data, and that there will be many and varied requirements and opinions on what is wanted or needed. Furthermore, it is unlikely that all of these will be achievable in the time allocated for the project. It will therefore be necessary to refine the requirements generated during this process and narrow them down to the most important and most achievable, to maximise the chance of completion within the timeframe of the project. Approaches taken to meeting the requirements will also need to consider the time restrictions on the project.

# requirements

As it was intended to use an agile method of software development, requirements were gathered as User Stories, which are often used with Scrum methodologies(6–8). This method of capturing requirements has a recommended template:

“As a <some user> I want <feature> so that <reason>”

Requirements are outlined according to this template in the table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Priority | User | Feature | Reason |
| 1 | 1 | Genetic Technologist | Reagent and flow cell information | Problem batches can be identified.  If they are out of date this can be identified. |
| 2 | 1 | Bioinformatician | Quality information stored in a structured and accessible manner | Quality data can be easily accessed for checking if indicated.  Quality data can be mined to look for patterns in data. |
| 3 | 2 | Genetic Technologist | Storage of sequencing operator initials | Sequencing operator can be identified and problems with a particular operator can be identified. |
| 4 | 1 | Project Lead | Storage of sequencing operator initials | A summary of sequencing quality can be provided to the correct person after quality control. |
| 5 | 5 | Instrument operator | Ability to manually change thresholds of quality control metrics | Thresholds can be tailored to my specific project. |
| 6 | 5 | Bioinformatician | Ability to port over the code to the Java programming language | Future code developed in the AWMGS is to be in the Java programming language. Adherence to a “house-style” is required in the draft best practice guidelines for software development within NHS genetics laboratories(9). |
| 7 | 4 | Lead Genetic Technologist | Graphical user interface to enable querying of data for audit purposes | Data can be easily queried for audit purposes, e.g. all runs from a single operator or all runs from a particular batch number can be inspected. Perhaps inspection of plotted data could also be possible? |
| 8 | 1 | Bioinformatician | Cross-platform compatibility of the software | The software can be run on either the computer attached to the sequencer or on the computer cluster. It is not clear currently where would be best to site it. |
| 9 | 1 | Instrument operator | To see if a run has passed or failed and why it has failed | Appropriate remedial action can be taken. |
| 10 | 3 | Bioinformatician | To provide information in the software on what action the operator should take as a result of failure on particular metrics | People don’t ask the bioinformatician what to do every time. |
| 11 | 4 | Genetic Technologist | To be able to access graphs and plots of the data | Potential patterns in the data can be identified by eye. |
| 12 | 2 | Lead Genetic Technologist | The software to tell me if a sequencer is giving consistently poor performance | It can be investigated and a service booked if needed. |
| 13 | 3 | Instrument operator | Support for the HiSeq | I can use the application with data from the HiSeq. |
| 14 | 1 | Bioinformatician | Good documentation of the code and a series of tests and testing schedule | The application can be maintained and upgraded. Also so that the application follows best practice and is compliant with quality requirements(9–11). |
| 15 | 2 | Project Lead | The results of the tests for each run to be stored in the database | The results can be easily and quickly retrieved for all instruments or during a particular time, which will be useful for instrument control. |

Priorities have been assigned to the project requirements based on the MoSCoW principle;

1-Must have, 2-Should have, 3-Could have, 4-Would have, 5-For later

# references

1. Illumina Technical Assistance. Sequencing Analysis Viewer v1.10 Software Guide. (October 2015).

2. Leenen CHM, Heijer M Den, van der Meer C, Kuipers EJ, van Leerdam ME, Wagner A. Genetic testing for Lynch syndrome: family communication and motivation. Fam Cancer [Internet]. Springer Netherlands; 2015;15(1):63–73. Available from: http://link.springer.com/10.1007/s10689-015-9842-8

3. Scuffham TM, Macmillan JC. Huntington disease: Who seeks presymptomatic genetic testing, Why and what are the outcomes? J Genet Couns. 2014;23(5):754–61.

4. Kim SY, Im K, Park SN, Kwon J, Kim J -a., Lee DS. CALR, JAK2, and MPL Mutation Profiles in Patients With Four Different Subtypes of Myeloproliferative Neoplasms: Primary Myelofibrosis, Essential Thrombocythemia, Polycythemia Vera, and Myeloproliferative Neoplasm, Unclassifiable. Am J Clin Pathol [Internet]. 2015;143(5):635–44. Available from: http://ajcp.ascpjournals.org/cgi/doi/10.1309/AJCPUAAC16LIWZMM

5. NHS Wales Informatics Service. Information Governance and Caldicott [Internet]. 2013 [cited 2016 Feb 20]. Available from: http://www.wales.nhs.uk/sites3/home.cfm?orgid=950

6. James M. Scrum Methodology [Internet]. 2015 [cited 2016 Feb 18]. Available from: http://scrummethodology.com/

7. Cohn M. User Stories [Internet]. Mountain Goat Software. 2016 [cited 2016 Feb 18]. Available from: https://www.mountaingoatsoftware.com/agile/user-stories

8. Martin RC. Agile software development: Principles, Patterns, and Practices. New Jersey, USA: Pearson Education, Inc; 2011. 529 p.

9. Whiffin N, Brugger K, Ahn JW. Draft Guidelines for development and validation of software, with particular focus on bioinformatics pipelines for processing NGS data. 2015.

10. International Organization for Standardization. We’re ISO, the International Organization for Standardization. We develop and publish International Standards [Internet]. [cited 2016 Feb 18]. Available from: http://www.iso.org/iso/home.html

11. ISO 15189:2012. 2012 p. 1–53.