High accuracy base space sequencing: results from error-correction ligation chemistry

Asim Siddiqui, Marcin Sikora, Somalee Datta, Chengyong Yang, Heinz Breu, Dima Brinza, Cisylia Duncan, Ryan Hsu, Srikanth Jandhyala, Brijesh Krishnaswami, Matthew Muller, Vasihnavi Panchapakesa, Daryl Thomas, Vasisht Tadigotla, Sowmi Utiramerur, Arjun Vadapalli, Eric White, Tanya Sokolsky, Yuandan Lou, Amitabh Shukla, Clarence Lee, Alan Blanchard ,Kevin McKernan, Fiona Hyland and Ellen Beasley

ABSTRACT

Sequencing read accuracy is critical for applications such as cancer or environmental sequencing where small subpopulations may play an important role or to improve variant calling in germline sequencing. Throughput cannot overcome lower accuracy sequencing when the variant sought is present at a frequency close to the base error rate of the sequencer. We report redundant sequencing of the DNA template using mathematical principles employed in the error correcting codes found in compact discs (CD's) and telecommunications protocols. Standard '2-base encoding' chemistry utilizes ligation probes that sequence in steps of 5 bases with 5 sequential primers. Probes are pooled such that only 2 of the 5 bases are interrogated. We then interrogate the DNA template with an additional primer in an orthogonal colour space to the 2-base colour determinations i.e. using ligation probes that are pooled and labeled to interrogate different bases. Since the DNA ligase provides specificity over 5 bases (enabling codes as long as 5 bases), we are able to interrogate multiple bases with this single additional primer round – we have chosen to use 4-base encoded probes. The orthogonal coding is a more powerful than simple resequencing. Using synthetic templates, we demonstrate that the sequencing chemistry is accurate to greater than 99.99% accuracy with the majority bases achieving 99.9999% or higher. Higher accuracy data is more valuable mitigating the cost and time of running an additional primer

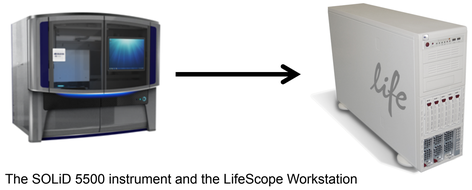
When we apply the ECC chemistry to real samples, we observe errors at the sub 1% level that are clearly induced in the sample preparation steps. ECC chemistry has enabled the optimization of sample preparation to reduce these errors to the sub 0.01% level. We will describe reference-free ECC sequencing of microbes and human germline and introduce reference-assisted ECC sequencing which reduces errors further

INTRODUCTION

High accuracy reads provide an advantage in experiments where the errors cannot be overcome by throughput through redundant sequencing. Such experiments include those where subpopulations are present and their elucidation is critical. For example, searching for circulating tumor cells or a search for a specific variant in a tissue biopsy are both examples from cancer research where subpopulations are important. Metagenome sequencing is another example – teasing apart closely clades is made simpler with higher accuracy reads especially if some populations are present at low frequency. We have developed ECC chemistry as a means of producing higher accuracy reads. The properties of these reads enable both error detection and correction of the sequenced template. This poster describes the chemistry and the bioinformatics analysis of ECC.

The ECC chemistry is available on the SOLiD™ 5500 instrument. This poster also introduces LifeScope™ Software, the new data analysis software package for the SOLiD 5500 instrument.

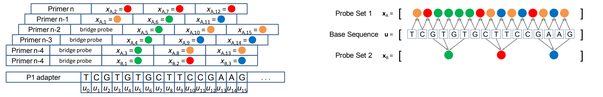
MATERIALS AND METHODS



The SOLiD 5500 instrument and the LifeScope Workstation

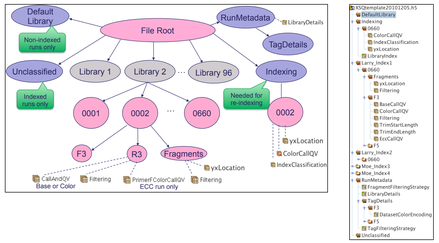
Material and Methods

Figure 1. Example of color encoding for base sequence u = [TCGTGTGCTTCCGAAG]. Colors arranged by A) order of data acquisition, B) base position and probe set.



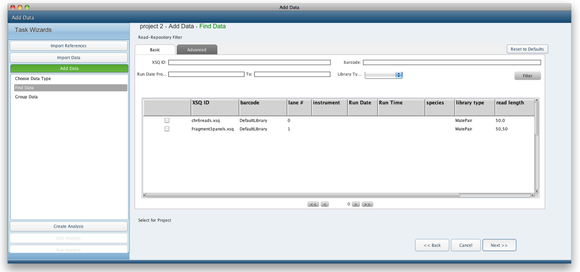
The encoding properties of the ECC chemistry are shown in Figure 1.The first five cycles are identical to the standard 2 base encoded probes used in the original colour space chemistry. ECC adds a 6th cycle which produces redundant information spanning a series of 4 bases and represents a 4 base code.. This redundant information provides the means for detecting and correcting errors. Although, we have chosen a combination of 2 base and 4 base encoded probes, other combinations are possible and SOLiD chemistry allows up to 5 base codes.

Figure 2. XSQ file format



New data requires a new format. We have built the XSQ (eXtensible SeQuence) file format upon the open HDF format [1]. The format supports base space data and colour space data. Hence, it supports both classic 2 base encoded data and the new ECC colour data. It can also store the base space sequence derived from the ECC codes. XSQ is a binary format and hence offers compact storage.

Figure 3. LifeScope Software

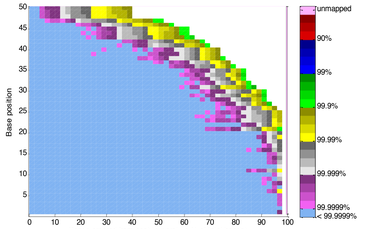


A screenshot from the LifeScope Software.

Results

Figure 4. Majority of bases from synthetic template are sequenced at > 99.9999% accuracy

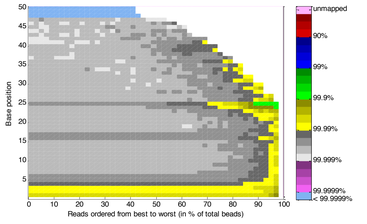
Sample Description	
Sample	Synthetic template beads
Read length	50bp
Mapped Reads	99.6%
Perfect Reads	99.7%
Raw Accuracy	99.9983%



Synthetic template beads are preloaded with DNA of known sequence. As such they are not created through the standard sample preparation method, but they are run on the instrument as a standard run. The resulting FASTQ file was filtered and end trimmed using a base QVs prior to mapping. Instrument throughput is reported on filtered and trimmed reads.

Figure 5. Including sample preparation, average accuracy over all bases is > 99.99%

Sample Description	
Sample	DH10b
Read length	50bp
Mapped Reads	99.7%
Perfect Reads	98.3%
Raw Accuracy	99.9905%



Unlike the data shown in Figure 4, this sample is derived from DH10b strain of E. coli and processed through the standard sample preparation steps (sufficient DNA was used to remove the need for amplification cycles of PCR). Here the error rate is higher than in Figure 5. We see a higher error rate in the first few bases and this has been associated with the chemistry of attaching the adaptor sequence. The perfect matches in the last 3 bases are an artifact arising from the mapping parameters – in order for the alignment to extend the full length of the read, there must be no mismatches in the last three bases, otherwise the alignment is terminated before the end of the read. The last 3 bases are excluded from accuracy calculation.

Figures 4 and 5 demonstrate the error introduced in the sample preparation step. While with present methods, this places a bound on the achievable accuracy, ECC provides a unique tool for understanding the source of sample preparation errors and hence modifying the protocol to lower the error beyond what is achievable today.

The Importance of Accurate QVs

Most bioinformatics tools make extensive use of reported QVs to make accurate SNP calls. Inaccurate QVs will result in higher false positives and false negatives and hence a lower validation rate for the reported results. We can further boost accuracy by using the reference as a prior. Figure 7 shows that the reported QV values have the expected behaviour.

Figure 6. 90% of bases have QV >= 40

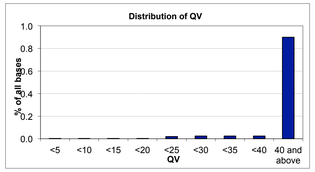
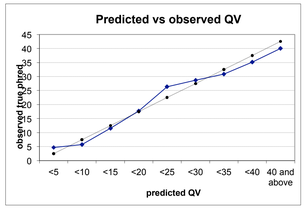


Figure 7. QV accuracy is an important factor in accuracy



With ECC and reference mediated base space translation, 90% of the bases have a reported accuracy > 99.99%. This predicted or reported accuracy is validated by mapping to the genome as shown in Figure 5. Figure 7 also attests to the accuracy of reported QVs. The reported QVs behave well when compared to measured QV rates over the entire range of QVs

CONCLUSIONS

This poster presented an analysis of base space data derived using ECC. Analysis of this data will be supported by the LifeScope Software. We have demonstrated the potential of achieving individual base accuracy of > 99.9999% with synthetic template beads sequenced on the platform. Using standard sample preparation methods, we achieve > 99.99% accuracy. This method provides a platform for further improvements to sample preparation improving accuracy beyond what is achievable today. Finally, accurate QVs are important for bioinformatics tools to work well. We show that the reported QVs behave as expected over the full range of values.

REFERENCES

1. HDF – www.hdfgroup.org

ACKNOWLEDGEMENTS

We thank members of the sample preparation team, Beverly sequencing team, software team and project management teams of SOLiD 5500 and LifeScope Software.

TRADEMARKS/LICENSING