Swift: Primary Data Analysis for the Illumina Solexa Sequencing Platform

Citation

Published Version
doi:10.1093/bioinformatics/btp383

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:4621712

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Sequence analysis

Swift: primary data analysis for the Illumina Solexa sequencing platform

Nava Whiteford1,∗,†, Tom Skelly1, Christina Curtis2, Matt E. Ritchie3, Andrea Löhr4, Alexander Wait Zaranek5, Irina Abnizova1 and Clive Brown1,1

1Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, 2Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way Cambridge CB2 0RE, UK. 3Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, 4Harvard-Smithsonian Center for Astrophysics, 60 Garden Street Cambridge, MA 02138 and 5Harvard Medical School, Genetics, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Received on February 22, 2009; revised on May 4, 2009; accepted on June 18, 2009
Advance Access publication June 23, 2009
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Primary data analysis methods are of critical importance in second generation DNA sequencing. Improved methods have the potential to increase yield and reduce the error rates. Openly documented analysis tools enable the user to understand the primary data, this is important for the optimization and validity of their scientific work.

Results: In this article, we describe Swift, a new tool for performing primary data analysis on the Illumina Solexa Sequencing Platform. Swift is the first tool, outside of the vendor’s own software, which completes the full analysis process, from raw images through to base calls. As such it provides an alternative to, and independent validation of, the vendor supplied tool. Our results show that Swift is able to increase yield by 13.8%, at comparable error rate.

Availability and Implementation: Swift is implemented in C++ and supported under Linux. It is supplied under an open source license (LGPL3), allowing researchers to build upon the platform. Swift is available from http://swiftng.sourceforge.net.

Contact: nwh@genomics.org; nava.whiteford@nanoporetech.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Second generation sequencing technologies such as the Genome Analyzer (Illumina, San Diego, USA), 454-FLX (Roche, Basel, Switzerland) and SOLID (Applied Biosystems, California, USA) have increased the production of sequence data by several orders of magnitude (Quail et al., 2008). Along with an increase in throughput these technologies bring an increased primary data analysis problem, both in terms of the complexity of the analysis and the volume of data that needs to be processed. Second generation devices image a surface to which clusters of DNA grown from a template or beads have been attached. The data analysis problem, therefore, first presents itself as one of image analysis, the result of which is a sequence of intensities which require further signal corrections to produce base calls (Brown et al., 2006). A number of new base-calling methods have been developed including Altacyclic (Erlich et al., 2008) and Rolexa (Rougemont et al., 2008). However, to our knowledge no image analysis methods have been developed outside of the proprietary implementation provided by the vendor.

In this article, we present Swift, an open source primary data analysis package for the Illumina Solexa sequencing platform which performs image analysis and base calling. Swift is the first freely available solution to the primary data analysis problem ‘from images to base calls’ and is available under the LGPL3 at http://swiftng.sourceforge.net. We perform validation against a φX 174 dataset.

The use of benchmark datasets is well established for the validation of analysis methods for gene expression (Cope et al., 2004; Holloway et al., 2006) and genotyping (Lin et al., 2008) microarrays. To facilitate algorithm development and comparison for the Illumina Solexa platform, we have made the raw files from a φX 174 dataset available from our web site. This dataset, which provides a stable reference against which sequencing errors may be determined, is used to assess the performance of our approach.

2 METHODS

In Illumina Solexa sequencing template, DNA sequences are attached to a flowcell (shown in Fig. 1). ‘Clusters’ of single-stranded DNA are grown from these single molecules and the prepared flowcell is placed in the sequencing device for imaging. Sequencing occurs as a cyclic process. A cycle of chemistry is performed which synthesizes a single fluorescently labelled complementary base to each DNA molecule (Bentley et al., 2008). The clusters are then imaged four times per cycle using two different lasers and two filters to detect the excitation of the four labelled nucleotides. The primary data analysis problem is therefore to take sets of images from the device and extract base calls from them. In an ideal scenario, a cluster would fluoresce in a single channel and the sequence of the template could be readily determined. However, the intensity vectors are not purely responsive to one distinct base and there are several signal artefacts present which must be corrected for in order to achieve accurate base calls.

∗To whom correspondence should be addressed.
†Present Address: Oxford Nanopore Technologies, Begbroke Science Park, Sandy Lane, Kidlington, OX5 1PF, UK.
Analyzer 2 is 100 tiles per lane, where there are eight lanes per flowcell. Tiles are separated by a margin to prevent clusters being imaged multiple times. The flowcell is moved under the camera in order to image each tile in each cycle. Therefore, during each cycle 800 tiles are imaged in four channels. Runs are typically around 37 cycles producing a total of 118400 images. Each GA2 image is a 2048 $\times$ 1794 pixel 16 bit grey-scale TIFF (though only 12 bits contain data). Each pixel covers $\sim0.14 \mu m^2$ of the flowcell and $\sim9$ pixels comprise one cluster object.

In Swift, the unit of analysis is a set of images covering a single tile. These undergo image analysis the result of which is an intensity vector, for each cluster (a total of 148 intensity values, per cluster, for a 37 cycle run).

2.1.1 Background subtraction The removal of non-specific ‘background’ intensity is desirable in order to obtain a less-biased measure of the true signal. For microarrays, it has been shown that subtracting a conservative estimate of the local background performs well, and avoids negative intensities (Ritchie et al., 2007), which are undesirable in downstream analysis as, and as described below, possible in the GAPipeline (Brown et al., 2006). We therefore take the conservative approach of morphological erosion (Serra, 1983) in Swift. In this method the minimal pixel value within a window around each pixel is subtracted from the central pixel’s value. For efficiency, we use a square structuring element and implement the process using a FIFO queue.

2.1.2 Image correlation In each cycle, the stage supporting the flowcell is moved to each tile position in turn. This repositioning is not entirely accurate and between cycles images may be several pixels out of alignment. We must therefore bring the images within each channel into alignment to compensate for this. A reference cycle is selected (typically the first cycle) to which subsequent images are aligned. Images are thresholded for alignment, discarding noise, using the method described in the following section. Alignment is performed by cross-correlating the images at the pixel level. Registering images to subpixel resolution was found unnecessary. This step may be performed efficiently using a Fast Fourier Transform (FFT) (Castro and Morandi, 1987). Swift uses the FFTW library (Frigo and Johnson, 1998) to implement this process.

To allow for a variation in offsets across the image (due perhaps to incorrect focusing, or warping of the flowcell due to temperature variations), we divide the image into regular regions and calculate and apply offsets in each region independently. After this cross-cycle registration step, images within each channel should be in alignment. In order to make this process robust, the median channel offsets are used, as all imaging channels should be subject to the same stage movement.

Though each channel is now in alignment from cycle to cycle, an offset still exists between channels. This offset is due to the different dye emission frequencies, and therefore differences in the optical path. In order to compensate for this offset an aggregate image is constructed for each channel, which simply sums the intensities from all cycles creating a reference image which should contain all clusters. The reference image for each channel is correlated against the other channels and from this cross-channel offsets are determined. Once these offsets are applied all images should be in alignment and we can begin to identify objects and extract their intensities.

2.1.3 Object identification and intensity extraction Images are thresholded in order to identify foreground (cluster) pixels. In order to threshold the image, we create a morphologically dilated (Serra, 1983) image and threshold those pixels in the original image that are within a given fraction of their dilated value. These parameters may be adjusted by the user. The process is implemented efficiently using a FIFO queue and square structuring element. This thresholding scheme is invariant to the differences in illumination commonly seen in Genome Analyzer image data. Reference contours are produced for each cluster, formed from groups of four-connected foreground pixels. Reference contours are created from non-overlapping objects in first cycle images by default.

Once we have a set of reference contours for each cluster, the maximum pixel intensity within the contour is extracted from the background subtracted, aligned images. This process is performed for each cluster across cycle. The result is a sequence of four intensities per cycle across N cycles.

2.2 Base calling If no artefacts were present in the signal, we could now simply ‘call’ the maximum intensity as the base call. However, a number of artefacts are present making the post-image analysis signal correction problem significant. Several tools exist for solving this post-image analysis, or base-calling, problem (Erlich et al., 2008; Rougemont et al., 2008). Swift is able to output a GAPipeline-style intensity file, allowing it to be used with any of these tools. However, Swift also provides its own base-calling algorithm allowing complete end to end operation using a simple but efficient and robust approach. In contrast to the existing tools, Swift does not employ machine learning or statistical modelling but rather applies a series of corrections to the signal a methodology similar to that employed in the GAPipeline.

2.2.1 Crosstalk correction The first artefact we must correct for is crosstalk. This is caused by an overlap in the dye emission frequencies. Simply put, the C channel illumination overlaps with the A, causing a C label to produce a small amount of illumination in the A channel, and similarly the G and T dye responses also overlap. Figure 2 shows crosstalk plots typical of those produced by a Genome Analyzer 2. In order to correct for this we use a method similar to (Li and Speed, 1998) with a few minor differences. The basic method puts regression lines through the arms of the plots. The arms are identified by placing bins along the X- and Y-axis and detecting the minimum values in these bins. Linear regression is performed on these values and the slope is used to derive a correction matrix. This process is performed iteratively until almost no slope remains. In Swift, we also perform some basic outlier detection, by placing a number of bins across the crosstalk plot and removing those containing few points.

Some Genome Analyzer datasets produce deviations in crosstalk, which appear intensity dependent. This produces slightly ‘bowed’ plots as seen in Figure 2. Placing a linear regression though the minimal values produces poor results and we therefore adopt a second strategy in addition to the method described to cope with this scenario. In this method, we identify a set of clusters (using the chartney metric described below) where it is likely that the correct base call can easily be determined, and then perform a linear regression. However, we must account for the fact that the base call is not binary (A, C, G, T) but a continuous value. The solution is to use interpolation.

Once we have a set of reference contours for each cluster, the maximum pixel intensity within the contour is extracted from the background subtracted, aligned images. This process is performed for each cluster across cycle. The result is a sequence of four intensities per cycle across N cycles.

...
each from a single DNA template. However, mixed clusters can be grown

2.2.3 Chastity filtering

In the ideal scenario discrete clusters are grown, subsequent cycle's phasing. is applied iteratively starting with the first cycle, then correcting for each estimation to ensure that unreasonable values are not used. The correction cycles and added to the current cycle. A limit is placed upon the phasing of the current cycle's intensity is subtracted from subsequent and previous cycles. From these we calculate which are brightest in the current channel, but whose maximal intensity is not this channel in previous and subsequent cycles. From these we calculate the fraction of the intensity that has leaked between cycles and use this as our phasing estimate. This phasing estimation is then applied as a correction where this fraction of the current cycle's intensity is subtracted from subsequent and previous cycles and added to the current cycle. A limit is placed upon the phasing estimation to ensure that unreasonable values are not used. The correction is applied iteratively starting with the first cycle, then correcting for each subsequent cycle's phasing.

2.2.3 Chastity filtering

In the ideal scenario discrete clusters are grown, each from a single DNA template. However, mixed clusters can be grown regression on these values and derive a correction matrix as before. Figure 3 shows the pairwise intensity plots after correction.

After crosstalk correction negative values, which represent an over-correction, are removed. The resultant signal is re-expressed as a deviation from the median intensity value in each channel. This normalization provides a fixed reference point for signals. We believe this may help compensate for any build up in background intensity, such as the common ‘sticky-T’ phenomenon, due to incomplete cleavage of the T-dye.

To correct this, we rank clusters by ‘Chastity’ (described in the following section) and use the top 400 clusters to estimate phasing.

2.2.2 Phasing correction

As previously stated each cluster contains many identical copies of a template sequence. During each cycle, labelled nucleotides are incorporated into these molecules. However, this is driven by stochastic chemical processes, so some molecules may fail to incorporate a labelled nucleotide, or may fail to block and incorporate >1 nt. This manifests itself as a leakage in intensity between cycles. For example, if a G base is in present in cycle 2, we will see a small amount of G intensity leaking into cycles 1 and 3. Figure 4 illustrates forward phasing.

Fig. 4. Schematic representation of phasing in Illumina Solexa sequencing. The figure shows intensities detected across cycle where the true sequence reads CTACG... Cross-cycle signal leakage can be seen, for example, in cycle 2 which contains a significant fraction of the intensity present in cycle 1. By the final cycle, the signals are almost fully convolved and it becomes difficult to determine the true signal. Phasing is caused by the non-incorporation of a labelled nucleotide in a given cycle. The reverse phenomenon (data not shown) is caused by the multiple incorporation of nucleotides and manifests itself as signal leakage between the current and previous cycles.

In the GAPipeline, a limit is placed on the second lowest value of chastity over the first 25 bases. By default this is set as 0.6. As this is a robust and well-validated metric it is also used in Swift, with a user-selectable threshold.

2.2.4 Base calling

After correction, the base with the maximum intensity is chosen as the called base. Swift can also optionally produce a Fastq file. This file contains the raw probabilities for all four bases. This additional information may prove useful in alignment. Base probabilities are simply calculated as a fraction of the total intensity. Fastq files may also be created. The value used is derived from the probability of the called base scaled over the range Q6–Q30. As with all quality scores this value requires calibration in order to provide a useful metric.

2.2.5 Comparison with the GAPipeline

GAPipeline (the vendor-supplied analysis tool) performs image analysis, base calling and alignment (optional) and is composed of a number of independent programs. The UNIX ‘make’ utility is used to manage workflow and job control. Figure 5 shows the processing steps common to Swift and the GAPipeline. In this section, we briefly discuss our investigation of the GAPipeline’s methods, based on the release prior to GAPipeline 1.0. A detailed description of our analysis is also available (http://igenomics.org/mediawiki/upload/8/80/Pipeline.pdf).

Processing begins with Firecrest, the GAPipeline image analysis tool. Firecrest operates in two passes. In the first pass a full image analysis is performed but only the offsets between imaging channels are retained. The second pass then applies these offsets between channels. This differs from Swift which aligns images within each channel and then performs alignment on aggregate reference images in a single pass. Firecrest begins by applying a Mexican hat filter to each image. This attenuation of high and low frequencies smooths the image and strengthens the edges. The image is broken into an even number of regions >125 pixel square for noise and background estimation. A smoothed, filtered histogram of the

Fig. 2. Pairwise intensity plots from cycle 1 of a Genome Analyzer 2 run. Unlike pairwise plots produced by capillary sequencing (Li and Speed, 1998) many Genome Analyzer plots have a distinctive ‘bowed’ appearance.

Fig. 3. Pairwise intensity plots from cycle 1 of a Genome Analyzer 2 run after crosstalk correction.
In this section, we discuss the validation of Swift against five tiles of \( \phi X174 \) control from a 37 cycle single-end Genome Analyzer 2 run.

\[ \text{Table 1. Comparison of Swift and GApipeline 1.3.2 error rates for given tiles on Sanger Institute run 1851 lane 4} \]

<table>
<thead>
<tr>
<th>Tile no.</th>
<th>Total reads</th>
<th>Dups. Rem.</th>
<th>PF</th>
<th>Err. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>333 945</td>
<td>192 103</td>
<td>100 697</td>
<td>1.20</td>
</tr>
<tr>
<td>21</td>
<td>288 498</td>
<td>158 540</td>
<td>96 923</td>
<td>0.97</td>
</tr>
<tr>
<td>41</td>
<td>286 302</td>
<td>150 220</td>
<td>88 374</td>
<td>1.05</td>
</tr>
<tr>
<td>61</td>
<td>276 433</td>
<td>146 407</td>
<td>92 247</td>
<td>1.00</td>
</tr>
<tr>
<td>81</td>
<td>280 969</td>
<td>135 752</td>
<td>98 872</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Swift yield

<table>
<thead>
<tr>
<th>Tile no.</th>
<th>Total reads</th>
<th>PF</th>
<th>Err. (%)</th>
<th>Inc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112 353</td>
<td>79 548</td>
<td>1.10</td>
<td>26.6</td>
</tr>
<tr>
<td>21</td>
<td>108 588</td>
<td>85 330</td>
<td>1.04</td>
<td>13.6</td>
</tr>
<tr>
<td>41</td>
<td>101 574</td>
<td>79 453</td>
<td>1.13</td>
<td>11.2</td>
</tr>
<tr>
<td>61</td>
<td>104 237</td>
<td>84 658</td>
<td>1.16</td>
<td>8.9</td>
</tr>
<tr>
<td>81</td>
<td>112 546</td>
<td>90 996</td>
<td>1.04</td>
<td>8.6</td>
</tr>
</tbody>
</table>

A full set of tile images, intensity files and base calls is available from http://sgenomics.org/swift/paperdataset.html. A comparably small dataset was chosen in order to allow us to make the full dataset available. We identify clusters from the first three cycles. In doing so, we run the risk of producing ‘optical duplicates’, that is, identifying multiple clusters and therefore reads, where there is only really one. In order to mitigate this effect, we apply an optical duplicate filter. This filter removes those duplicate reads within a 6 \( \times \) 6 window with similar sequence (allowing eight mismatches), and retains the read with the highest ‘chastity’. No optical duplicate filtering was performed for the GApipeline. Settings may be tuned to reduce the duplicate rate, however, this will result in a reduction in yield.

In order to determine whether improvements come from image analysis or base calling, we analysed intensity files from the GApipeline using Swift (Table 2). The results show a small increase in yield over the GApipeline in most cases, but no reduction in error rate. This lane contained \( \text{dN} \) 174, a commonly used control genome. Sequences were aligned using PhageAlign from the GApipeline version 1.3.2. The GApipeline performed its own alignment. Default parameters we used for the GApipeline. As can be seen Swift provides an average increase in yield of 13.8% at comparable error rate.

In order to support our validation of Swift, a different and larger genome was analysed. A lane of data from 3 Mb of exonic regions in PGP2 (Zaranek et al., 2009) Lessons from the initial data release of the personal genome project, in preparation.) was used for this purpose. This dataset is available online via Tranche at http://openwetware.org/wiki/PGP_and_Tranche. Analysis was performed in a virtual machine on the Free Factories compute infrastructure (Zaranek et al., 2008).
As previously described, Swift generates a set of four probabilities that are used to determine the quality of each base call. These probabilities reflect the true quality, as to be expected from an uncalibrated quality score. In order to correct for this, we apply a simple calibration (Ewing and Green, 1998) scheme. To generate a calibration table, reads from a training tile are aligned to a reference, those reads with more than seven mismatches are discarded as being most likely generated from contamination. Once the observed quality scores have been calculated, they are used to construct a mapping from predicted to observed values. This mapping may then be applied to another dataset where the true reference may not be known, using a simple lookup table. Figure 7b shows the results of applying this mapping onto a different tile from the same run, the quality scores now largely reflect the true base quality (showing that the method is somewhat transferable). The highest quality assigned is Q32, with 53% of bases being calibrated to Q30 or above. This result shows that the calibration table generated is transferable between tiles. This robust and simple calibration scheme is therefore a reasonable placeholder until more precise methods are developed.

### 3.2 Computational requirements

Processing a single GA2 tile across 37 cycles requires ~1 GB of main memory. When compiled using the Intel C++ compiler version 11, processing took 25 min. The GAPipeline uses GNU C++ and modifying this is non-trivial. The GAPipeline took 23 min to process this tile and 657 MB of memory on our dataset (gcc version 4.2.3 installed). Using this compiler Swift took 30 min. Swift however operates on the tile level; a user may therefore submit 800 jobs and gain maximum utilization of their cluster (as opposed to eight jobs which may be submitted for the GAPipeline). Our benchmarks were performed on a single core of an Intel Core2Duo T8100 at 2.10 GHz.

## 4 DISCUSSION

We have described a new open source pipeline for the analysis of primary data from the Illumina Solexa sequencing platform. Our analysis provides validation of the vendor-provided analysis tools and is the first openly documented technique for extracting sequence data from images on this platform. We have provided a pipeline which other researchers may use as a platform for further development and which allows them to freely distribute their modifications.

Swift protects the user from potentially undesirable changes to the vendor-supplied analysis tools. It gives users an alternative to running multiple different pipelines with different parameters, which can be confusing and time-consuming. Instead, Swift allows users to easily tune its parameters to lower the optical duplicate rate, while still maintaining high quality base calls. This makes Swift a valuable tool for researchers who want to analyze their sequencing data in a consistent and reproducible manner.

### Table 2. Comparison of Swift and GAPipeline 1.3.2 Image analysis when used in conjunction with the Swift basecaller, as for Table 1

<table>
<thead>
<tr>
<th>Tile no.</th>
<th>Total reads</th>
<th>Dups. Rem.</th>
<th>PF reads</th>
<th>Err (%)</th>
<th>Inc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112,925</td>
<td>112,503</td>
<td>92,345</td>
<td>1.12</td>
<td>9.05</td>
</tr>
<tr>
<td>2</td>
<td>104,237</td>
<td>103,782</td>
<td>87,455</td>
<td>1.24</td>
<td>5.48</td>
</tr>
<tr>
<td>41</td>
<td>101,574</td>
<td>101,171</td>
<td>85,281</td>
<td>1.15</td>
<td>5.63</td>
</tr>
<tr>
<td>61</td>
<td>112,546</td>
<td>112,122</td>
<td>82,588</td>
<td>1.36</td>
<td>19.72</td>
</tr>
</tbody>
</table>

### Table 3. Comparison of Swift and the GAPipeline 1.3.2 for a lane of data from 3 Mb of exonic regions in PGP2 (Zaranek, A. et al. (2009) Lessons from the initial data release of the personal genome project, in preparation.)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Total reads</th>
<th>PF reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swift</td>
<td>8,164,716</td>
<td>4,665,259</td>
</tr>
<tr>
<td>GAPipeline 1.3.2</td>
<td>6,903,576</td>
<td>5,168,531</td>
</tr>
<tr>
<td>GAPipeline 1.3.2 + Swift</td>
<td>6,694,377</td>
<td>4,954,683</td>
</tr>
</tbody>
</table>

For Swift, “total reads” indicates the number of reads after duplicate filtering.
bundled analysis platforms and allows them to manage their own IT infrastructure. It also provides an openly documented analysis tool enabling the user to understand the primary data, and to investigate frequent device changes which can cause unexpected side effects. For example, the original version of the Genome Analyzer did not image the flowcell wall, current revisions do resulting in artificial poly A sequences. Understanding the primary data is critical to the operation of a high-throughput sequencing facility.

We have also shown that there is significant room for improvement in the vendor-supplied analysis showing an increased yield of 13.6% (at the same error rate). We believe that the image analysis methods presented should prove to be transferable to other next-generation sequencing platforms, such as the Applied Biosciences SOLiD™ Sequencer and Roche 454 FLX, though in the latter case, the image analysis problem should be simplified due to the regular arraying of beads. Potentially, this allows a user to maintain a single primary data analysis platform for all second generation systems.

ACKNOWLEDGEMENTS

We thank the following individuals from the Wellcome Trust Sanger Institute: James Bonfield for many illuminating discussions, Andy Brown and Roger Pettett for their suggestions relating to Swift’s reporting module and Tony Cox for his continued support. We also thank Klaus Maisinger of Illumina for discussions relating to the Illumina primary data analysis methods and Tony Cox of Illumina for answering our queries relating to PhageAlign.

Conflict of Interest: Since the inception of this work Nava Whiteford and Clive Brown have moved to 3rd Generation sequencing company, Oxford Nanopore Technologies. The other authors have declared none.

REFERENCES


