All Your Base: a fast and accurate probabilistic approach to base calling

Tim Massingham*[1], Nick Goldman[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Email: Tim Massingham*- tim.massingham@ebi.ac.uk; Nick Goldman - goldman@ebi.ac.uk;

*Corresponding author

Abstract

The accuracy of base calls produced by the Illumina Genome Analyser is affected by several processes during sequencing, with laser cross-talk and cluster phasing being prominent. We introduce an explicit statistical model of the sequencing process that reduces the rate of miscalled bases. The novel algorithms presented are considerable quicker than competitive base-calling methods, do not require training data and are designed to be robust to gross errors, producing sensible results where other techniques fail.

Background

There can little doubt that the vastly increased throughput of Next-Generation Sequencing (NGS) machines has revolutionised DNA sequencing but the reads produced are both shorter and less accurate than those from capillary sequencing and discoveries from NGS are often verified using traditional sequencing [1]. The challenges to be overcome to improve the accuracy and read length of NGS platforms are different from those that were faced by capillary sequencing [2] and require different strategies to tackle them. In particular, the phasing process — individual molecules of DNA getting out-of-step with others in the same cluster — is complex and ultimately limits the length of reads which can be obtained from

cluster-based sequencing-by-synthesis methods [2]. Here we develop an explicit statistical model of the sequencing process, including phasing and other signal-degrading processes. By implementing a base calling algorithm based on this model, our AYB software is able to produce more accurate reads.

Our statistical model is quite generic and so applicable to all sequencing-by-synthesis and similar platforms (sequencing-by-ligation, pyrosequencing; see Metzker [3] for a comparison) that rely on large numbers of clusters consisting of many homogeneous DNA molecules. We concentrate on the Illumina Genome Analyser (GA-II), both to provide a concrete foundation to aid exposition of the methods and because of local availability of data for testing and comparison. The mechanics of sequencing-by-synthesis on the GA-II platform have been described elsewhere in detail [3]. Here we present a brief overview to establish context and terminology for the rest of this paper.

Fragmented single-stranded DNA is washed through the lanes of a slide, where it attaches and is amplified to form a sequence-homogenous cluster of molecules. Sequencing progresses in steps, referred to as cycles, with each cycle conceptually sequencing one position of DNA. For each cycle, Fluorophore-Labeled Nucleotides (FLNs) are washed through the lanes of the slide and attach to the molecules in each cluster; attachment of more than one nucleotide in a given cycle is prevented by the presence of a reversible terminator element on each FLN. After the attachment process has run to completion, the intensity of fluorescence from each cluster is recorded in four channels, each channel being a combination of illumination with a specific laser and imaging through a specific filter. Clusters are artificially grouped into tiles, regions of the lane consistent over cycles, whose size is constrained by the capacity of the imaging equipment. The terminator elements and fluorophores are then cleaved from the FLNs, setting up each cluster so that in the next cycle of sequencing appropriate FLNs should attach to the next position of sequence.

After processing the images to pick out individual clusters, the output of the sequencing machine is many $channel \times cycle$ matrices of intensities, one matrix for each cluster. In principle the bases could be called straight from these intensities but there are several complicating factors that must be dealt with [4] with cross talk, phasing and dimming being of particular importance.

Cross talk between the channels occurs because, although they are chosen to help distinguish the fluorophores, the emission spectra for the fluorophores overlap and so light from each may is recorded in several channels. There is not a one-to-one correspondence between channels and FLNs and the relationship between the emission of each fluorophore and the intensity observed in each channel needs to be ascertained and corrected for.

Phasing refers to the deterioration in relationship between sequencing cycle and sequence position as the

cluster loses coherence: on a given cycle, FLNs may be attaching to different positions on different molecules within the cluster. There are many possible explanations for phasing: for example, a FLN might have a defective reversible terminator element leading to the attachment of two FLNs to a molecule on a single cycle, allowing the molecule to get ahead in the sequencing process ('pre-phased'), or the cleaving of the reversible terminator might fail for a cycle so the molecule lags behind when the element is finally removed ('post-phased'). Finally, molecules within a cluster may stop contributing to the total signal, laser damage to the individual molecules or problems reversing the terminator element being possible causes, leading to a decrease (dimming) in the overall emission observed from each cluster in later cycles of sequencing.

The cross talk is a consequence of the physics of fluorophore excitation and methods for estimating it have already been developed for dye-terminated capillary electrophoresis sequencing platforms [5]. Phasing and dimming are more specific to NGS methods and the Illumina platform in particular, and have been approached in a variety of ways. The Illumina base caller (Bustard) assumes a constant rate of post-phasing and pre-phasing for all cycles [6], as do the Alta-Cyclic caller [7] and Rolexa [8], whereas BayesCall [6] allows the phasing at each position of the sequence to depend on several of the neighbouring bases and Ibis [9] assumes that all information about the phasing at a given cycle is contained in intensities of the cycles either side. In contrast our method uses a complete empirical model of phasing, allowing all aspects of this process to be determined by the data on a run-by-run, indeed tile-by-tile basis.

Results and Discussion

Phasing

Phased molecules eventually dominate the observed emission from each cluster, becoming the major component of the signal in later cycles. Understanding this process is important for accurate base calling. The empirical phasing matrix produced by AYB gives insight into how phasing varies as the number of cycles increases and assumptions about how the phasing process progresses can be tested against this. Empirical phasing from a 151 cycle run of Human sequence on an Illumina GA-II$_x$ machine is shown in figure 1 and, as expected, the proportion of the signal coming from the in-phase position of the read decreases as the cycle number increases, contributing less than half of the total signal for the final 61 cycles of the run. AYB still makes accurate base calls over these cycles, its mean per-mapped-base error rate being 0.010 compared to 0.015 for (quality filtered) Bustard calls, and the improvement is most noticeable for later cycles where the error rate is greatest. The improved accuracy for later cycles has a marked effect

on the number of error-free reads, AYB producing 19.7% more than Bustard.

The proportion of molecules lagging behind ("post-phased") increases with cycle number at an apparently constant rate whereas the proportion of pre-phased molecules increases at a slower rate and tails off in later cycles. This is consistent with the rate of pre-phasing being lower than post-phasing, so pre-phased molecules tend to become in-phase and ultimately post-phased. The post-phasing and pre-phasing on the final two cycles show a strong artefact, the amount of post-phasing sharply increasing whereas pre-phasing disappears. A possible explanation is that the current and previous positions are being used to predict future sequence and so compensate for the pre-phased position which can no longer be estimated from future cycles. This effect is also apparent in the spread of weight in the phasing matrix on the last two cycles, where many more positions have large weight compared with nearby cycles (not shown).

The contribution to the observed signal made by pre-phased and post-phased molecules can be subdivided into the contributions from molecules pre- or post-phased by multiple positions to reveal further asymmetry between the two types of phasing: the contribution from pre-phased molecules is predominantly from one position ahead, whereas the contribution from post-phased molecules is spread out over many lagging positions so, after the first 40 cycles, the immediately preceding position is no longer dominant.

Quality calibration

The accuracy of a base caller can be easily tested by sequencing DNA from a known reference genome, mapping the called reads and then tabulating the differences. The same procedure is also useful in calibrating the base call quality scores and for comparing the performance of different base callers. The BWA short read aligner [10] was used to map reads back to the appropriate reference genome (edit distance of five) for all comparisons in this paper, having been chosen for its speed and its ability to deal with insertions and deletions. The quality calibration results below are based on control lanes of an Illumina GA-II sequencing run, analysing 9 tiles of 76-cycle ϕX174 data from each of three lanes differing in cluster density (in total 680k, 1300k and 900k clusters for lanes 2, 4 and 6 respectively). The genome of ϕX174 is only 5,386 bases long and so the few tiles analysed were sufficient to cover the genome to a depth of many tens of thousands, enabling accurate SNP correction (variants called using MAQ [11]).

The estimation and quality scoring methods used by AYB are robust and compensate for unusual clusters to produce a reasonable measure of confidence for each base call but their accuracy will deviate due to discordance between AYB's assumptions and how the machines actually operate, the distribution of emission spectra not being Gaussian for example. The relationship between the AYB predicted quality

scores (from equation 9, transformed into a quality score) and the empirical quality scores for base calls from the 9 tiles of 76-cycle ϕX174 data is shown in figure 2 (left; uncalibrated). The relationship is close to linear for the majority of the bases but with a slope and intercept that differs from what would be expected for a perfect fit (slope one and intercept zero).

Quality scores can be calibrated using either of two approaches: applying some parametric function to the raw scores to improve the correspondence with empirical scores, or discretising (e.g. rounding to integers) and using a table to convert to an empirical score. Many methods of calibration have been described in the literature [12] but the strong linear relationship displayed in figure 2 suggests that a simple method will suffice for AYB; here we fit a linear model to the qualities with a constant correction for each triplet representing the local base composition of the sequence (the previous, current and next base). The calibrated quality $Q_{\mathrm{cal}}(b_k)$ of the base b_k at cycle k is related to its quality $Q(b_k)$ (from equation 9, appropriately transformed) by

$$Q_{\mathrm{cal}}(b_k) = \alpha_{b_{k-1},b_k,b_{k+1}} + \beta Q(b_k) \tag{1}$$

where the $base \times base \times base$ table α and the constant β are chosen to agree with previous observation. The parameters α and β in equation 1 were found using the same mapping to the reference genome as used to assess the quality of the 9 tiles of ϕX174 data. A further 18 tiles of ϕX174 data, 9 each from two other lanes of the same run, were also available. Figure 2 (right; calibrated) shows the results of calibrating the first lane using each of the three sets of constants; the agreement between observed and expected is much closer than for the uncalibrated data.

Root Mean Squared (RMS) error has been used as a measure for the accuracy of calibration [13] and AYB achieves an RMS of 0.79 when the data set shown in figure 2 is calibrated against itself, comparable to the best of the within-lane results reported by Abnizova et al. [13]. In contrast, the RMS scores from calibrating using the other two lanes are 2.57 and 1.28, highlighting the observed deviation from perfect fit. The RMS without calibration is 8.33. As a criterion for assessing the accuracy of calibration, the RMS is not ideal since it penalises poor calibration of low quality bases equally as much as the more important high quality bases, and over-estimation of quality is penalised the same as under-estimation despite these errors having different consequences for down-stream analysis. The alternative 'S_p' criterion, inspired by information theoretic considerations, has been suggested (Richard Durbin, personal communication), and the $\bar{S}_p$ criterion defined by equation 2 is a modified form of this. The $\bar{S}_p$ criterion for a set of reads with

qualities q_a assigned to n_a bases of which e_a are errors, is

$$\bar{S}_p = \frac{\sum_a n_a q_a + \frac{10}{\ln 10}\left(n_a - e_a 10^{q_a/10}\right)}{\sum_a n_a} \tag{2}$$

which is maximised if the assigned qualities equal the empirical qualities; this maximum is equal to the average base quality.

Applying the $\bar{S}_p$ criterion shows that calibrating the ϕX174 lane 6 data using constants derived from each of the three lanes achieves close to maximal performance (given in brackets): same lane, 29.3 (29.3); other lanes, 28.7 (29.2) & 29.1 (29.3). Uncalibrated data achieved $\bar{S}_p$ of 10.9 (28.2). The closeness of the $\bar{S}_p$ statistics for the calibrated qualities to their maxima is to be expected given how well all three calibrations perform in the region of quality values where most of the bases are concentrated (roughly quality 25–35). Comparing the maximal possible values of $\bar{S}_p$ for the calibrated data to the maximum value for the uncalibrated data shows that the recalibration is actually improving the quality of the calls by about a unit, suggesting that the local context (the triple of bases) is important to the quality of calls. The RMS and $\bar{S}_p$ results for calibrating the other two lanes are similar to those reported.

The constants in equation 1 are expected to vary depending on the specific machine, sequencing chemistry and protocol, and many other factors. The default calibration for AYB is taken from the ϕX174 data described, although a tool is distributed with the software to enable users to derive calibration constants for their own machines.

Comparison to Bustard and Swift

Swift [14] is an independently developed open-source analysis pipeline for Illumina data aimed at replacing the Illumina software pipeline by implementing similar techniques. The results of base-calling with these two pipelines (referred to as Swift and Bustard for the Swift and Illumina pipelines respectively) provide a useful comparison to the performance of AYB.

The three lanes of bacteriophage ϕX174 data, described above, were used to compare the performance of AYB, Bustard and Swift, the results presented here being the 9 tiles from lane 6 (900k clusters) with the other two lanes being broadly comparable. AYB produces mappable reads from 73.3% of the 900k clusters, of which 74.6% are perfect; Bustard produces 72.1% mappable reads, of which 71.0% are perfect; Swift produces 68.8% mappable reads of which 58.5% are perfect. The per-base error rates are 0.7, 0.8 and 1.1% for AYB, Bustard and Swift respectively, so AYB has the lowest error rate in addition to producing the most mappable reads. Figure 3 shows that AYB has a lower error rate than both Bustard and Swift on all

of the later cycles, the cumulative effect of which becomes quite large by the final couple of cycles. The error-rate for all three methods jumps in the final position and this is probably due to pre-phasing — signal from future positions for which there is no intensity data; here AYB does a much better job than either of the other base-callers at reducing this artifactual spike. The Venn diagram for correct calls made by the three methods shows that AYB has a large overlap with both of the other callers but also makes a lot of additional correct calls.

A second comparison was based on a data set comprising an entire lane (100 tiles) of 76-cycle paired-end reads from *Bordetella pertussis*, using the complete genome of the Tohma I strain as a reference. The tiles from this run showed a large variation in the number of clusters, ranging from 345 to 82,000, with the tiles close to the ends of the flow cell containing fewer clusters. The sequence produced was generally of low quality, the first end of the read-pairs being particularly bad, suggesting problems during the sequencing. Oddities in the cross talk matrix, the channels corresponding to A and C nucleotides being noticeably brighter than those corresponding to G and T, suggest that there may have been illumination problems with one of the lasers. Because of the problems with this run and the presence of polymorphisms relative to the reference genome, it provides a useful comparison between Bustard, Swift and AYB when problems occur. Due to the error characteristics of the *B. pertussis* data and the paucity of reads to train on (few mappable reads and an imperfect reference), it is not possible to use base-callers based on machine learning approaches (see below).

Mapping back to the reference revealed a marked difference in base-caller performance: AYB produced more than four times as many perfect reads as Bustard (1,064k vs. 265k reads) and 2.5 times as many as Swift (407k reads) for the second end of the read-pairs. AYB produces 54% more mapped reads than Bustard and 41% more than Swift. The increased number of perfect and close to perfect reads produced by AYB has real consequences for down-stream analysis, with the genome being covered to a depth of 27.8 for AYB, 9.5 for Bustard and 13.8 for Swift. Greater coverage means more confident SNP and variant detection, which in turn leads to improved mapping of reads. As well as mapping to a reference genome, the length of contigs produced by *de novo* assembly is a useful guide to the quality of reads produced and of relevance in those cases where a reference is not available. Applying Velvet [15] to the second end of the paired-end reads, using a kmer length of 31 and default options, produces an N50 contig length of 996 bases for the AYB reads; the reads from both Bustard and Swift produce much shorter contigs on average, with N50 lengths of 548 and 531 bases respectively. For the first end of the read, AYB produces a greater number of perfect reads (173k reads) than Bustard (108k) or Swift (105k) and 48% more mapped reads

than Bustard and Swift. These results are shown in figure 4, along with the same data trimmed to the first 50 bases to show that AYB still produces more accurate reads even after the worst cycles have been discarded.

Much of the data for the 1000 Genomes project [16] has been archived and is publicly available for reanalysis, allowing for a further comparison between the base-callers and showing that AYB can be usefully applied to improve existing data. Two sequencing runs for NA19240 (Yoruban daughter) were reanalysed: ERR000479 (9.6 million 45bp paired-end reads, part of ERA000013 by the Beijing Genomics Institute) and ERR000610 (14.0 million 51bp paired-end reads, part of ERA000023 by Illumina Inc.). The accuracy of the base-calling was assessed by mapping to the human reference provided by the 1000 Genomes Consortium, based on GRCh37. This genome has variants relative to the the sample sequenced, but their presence penalises all callers equally. The raw intensities submitted to the archive have already been filtered for quality, so a high proportion of the reads map back to the reference: 83% for the BGI run and 95% for the Illumina run and there is little room for improvement. AYB produces 5% more mappable reads for the BGI data and 2% more for the Illumina data compared to the archived reads (produced using Bustard), a small increase but not insignificant. AYB however produces many more perfect reads, 14% and 3% more for the BGI and Illumina runs respectively, showing a gain in accuracy despite this being a situation where AYB would not be expected to do particularly better as the number of cycles, and so phasing, is comfortably small. The read length, vintage and error-rate of the BGI run is consistent with the older "sticky-T" chemistry (incomplete cleavage of the 'T' FLN, leading to an increased concentration in later cycles) and the improvement seen is typical for AYB on similar data. Swift failed to produce meaningful results for many of the tiles in these two runs, probably due to a software issue rather than a flaw in the method; a fair comparison was not possible so the results are not reported.

Comparison to the Ibis base caller

Many of the base callers in the literature reporting improved accuracy use machine learning techniques, in particular Support Vector Machines (SVMs), to analyse intensity data. These include the Alta-Cyclic [7] and Ibis [9] base callers, with the Ibis base caller claimed to be more accurate and requiring less processing power. Unlike AYB and the other previous mentioned base callers, machine learning approaches require training data, i.e. intensities and correct calls, before they can be applied to new data and this training data should be representative of future data (for example: the same machine, sample preparation, genome base composition, etc.). The Ibis base-caller trains a SVM for base calling at a particular cycle based on

the intensities for the previous, current and next cycles, whereas Alta-Cyclic trains on the corrected intensities, choosing the phasing rates to maximise the number of correct calls. Both Ibis and Alta-Cyclic require considerable computational resources to train their models.

A test data set of 200k clusters is distributed with Ibis, representing 51 cycles of ϕX174 sequence taken from a single lane of a single run. After training on the Bustard base calls for these clusters mapped back to the reference genome to assess correctness, Ibis produces 176,894 (88.4%) reads which map back to the genome (again using BWA with five or fewer edits relative to the reference) with 132,259 (66.1%) error-free reads and a total per-mapped-base error rate of 0.92%. In contrast, AYB produces 174,830 (87.4%) mappable reads with 133,137 (66.6%) error-free reads and a per-mapped-base error rate of 0.88%.

An immediate criticism of this analysis of the Ibis training data is that Ibis was trained on the same same set of data that it was calling and so would be expected to have an advantage. For a fairer comparison between AYB and Ibis we analysed our ϕX174 data, 9 tiles from each of three lanes described previously, which contain many more clusters than the Ibis test set. For each of the three lanes, Ibis was trained on that lane and then used to call the two other lanes and the results averaged; the effects of over-training should be reduced since Ibis is trained on data independent from that being called but still from the same run and the correct genome. Training and base calling for Ibis took 8, 10.5 and 10 hours for lanes 2, 4, and 6 respectively; the number of mappable reads, using the same criteria as above and averaged over the calls from each of the two possible training lanes, is 530,674 (78.3%), 849,245 (65.4%) and 666,091 (74.0%) for lanes 2, 4 and 6. Similarly, the numbers of error-free reads were 398,957 (58.9%), 549,843 (42.3%) and 478,454 (53.1%), and the per-mapped-base error rates were 0.85%, 1.04% and 0.78% respectively. The results from training Ibis on the data to be called are broadly in line with those trained on independent data, so there is no evidence that Ibis is overtraining.

AYB was substantially quicker than Ibis, taking approximately 2 minutes to analyse each tile or less than half an hour for all 9 tiles in each lane, estimating both the cross talk and phasing separately for each tile. The number of mappable reads was 524,387 (77.4%), 833,930 (64.2%) and 660,862 (73.4%) for lanes 2, 4 and 6 respectively, with 400,524 (59.1%), 558,112 (43.0%) and 493,235 (54.8%) error-free reads and per-mapped-base error rates of 0.61%, 0.96% and 0.70%. If a more stringent mapping criterion is used, only accepting those reads which map with one or no errors, then AYB produces more mapped reads than Ibis for two of the three lanes and always has a lower per-mapped-base error rate. The difference in performance between the two mapping criteria suggests that the majority of the small excess of mappable reads produced by Ibis over AYB are those with many errors.

AYB and Ibis are broadly similar in performance, despite taking different approaches to the base calling problem, but their performance on individual lanes diverges, with Ibis generally producing slightly more mapped reads and AYB producing more perfect reads and having a lower per-mapped-base error rate. The two methods have noticeable differences in the types of error they make, with AYB being more accurate on later cycles when phasing becomes the major source of errors and Ibis making more accurate predictions on the first 30 cycles or so (see figure 5). A possible explanation for the superiority of Ibis on the initial cycles is that it may be adjusting for differences in the cross talk between cycles whereas AYB assumes this to be constant.

Conclusions

A particular focus when developing AYB was to make the algorithms robust to problems that might arise during normal use, so it can be used confidently in cases where other base-callers require manual intervention to get the best results. The *B. pertussis* example was presented as such a case; another is data produced using the TraDIS technique [17] where the first few cycles of every cluster consist of known identical sequence, causing algorithms that estimate cross talk from a single early cycle to fail; AYB however estimates the cross talk over all cycles and produces reliable results without additional changes.
The statistical model underlying AYB has several weakness that could be addressed in future work. The model assumes that the descriptive parameters M (cross talk) and N (systematic noise) are constant across a tile but this is only going to be approximately true in practice: differences in illumination (e.g. mode scrambler problems) and laser intensity will affect the cross talk and background noise; also the expected amount of phasing might be affected by fluctuations in the chemistry. The phasing matrix represents an average over many clusters and the actual amount of phasing at a particular cluster is subject to stochastic variation; the fewer molecules contained in the cluster, the further from the average it is likely to deviate and this can lead to counterintuitive consequences as a small cluster that has, by chance, undergone little phasing will fit the average model as poorly as one that has undergone a lot of phasing — clusters can be penalised despite giving clear signal.
Sequence-like errors, for example mutations introduced during sample preparation, short fragments ligating together or adapter sequence, are essentially invisible to the base-caller and render it impossible to call the original sequence accurately. Other sources of error may not appear sequence-like: for example, microscopic particles of dust can get entangled in a cluster and produce bright artefacts for one or more cycles. Since very bright peaks deviate from the average brightness of the read, AYB penalises these calls

heavily and they rarely contribute to the higher quality base-calls but they also reduce the quality of the surrounding calls due to over-correction for pre- and post-phasing. Ideally over-bright peaks would be removed prior to analysis and treated as missing data, the actual intensity and base-call imputed from the remaining three intensities and the position has on the neighbour cycles through the phasing correction. A similar idea could be used to deal with clusters where intensities are missing (i.e. unrecorded, perhaps due to image registration problems) for some cycles, producing low quality calls rather than arbitrarily treating them as a cycle with four exactly zero intensities.

A final issue that AYB fails to account for is that of heterogeneous clusters of sequence, a common cause of which is two clusters merging into each other during the amplification step, since there is an implicit assumption that each cluster only contains fragments from one particular sequence. The intensities from such clusters appear to be extremely noisy, far above the stochastic background, and AYB's criteria to assess model fit are misled since the effects of both constituent sequences need to be be removed to get the residual noise. Failure to do so means that the calls from the strongest sequence get penalised for badly fitting the model; in particular, cycles where the two constituent clusters have the same base appear much brighter than expected given the intensities from other positions and are thus penalised despite the fact we should be more confident about these calls. In principle heterogeneous sequence could be explicitly estimated for each cluster, the relative brightness being used to separate contributions, but this will result in a loss of power in the majority of cases where the cluster is homogeneous and may not result in high-quality calls otherwise.

The speed at which tiles can be analysed is extremely important given the vast amount of data produced by current and future platforms. AYB is much quicker than many competitive base callers, taking only a few minutes to analyse each tile on ordinary single-core computing hardware, but even this could be prohibitive if computing resources are limited. There are, however, several possibilities to increase the speed of AYB. As noted previously, the estimates for the phasing and systematic noise are analytic given the cross talk and so they can be estimated without iteration; since cross talk is a well understood problem and good estimators already exist [5] this should result in little reduction in accuracy. AYB assumes that the cross talk and phasing are constant within a tile and this could be strengthened by assuming they are constant across tiles or across lanes, a similar assumption to that which Bustard and other base calling programs make. The relevant matrices could be estimated from a subset of data and then held fixed so AYB needs only to call bases for the majority of the data.

The phasing solution (equation 5) requires the multiplication of large matrices, an operation that is cubic

in the number of cycles and so scales badly. Therefore, a third way to speed up the software is to change the form of the phasing matrix, replacing the current nonparametric matrix with a more parametric model. A simpler model with single or cycle-wise values for the phasing and pre-phasing would be soluble with much less computation effort; both Bustard and Alta-Cyclic use phasing models that have single phasing and pre-phasing parameters.

AYB is considerably quicker than other methods of base-calling with comparable accuracy comparable. As the yield from the sequencing machines increases, speed of analysis becomes important and our base-calling method offers a unique combination of speed and accuracy. In addition AYB has two other desirable properties, not requiring training data so calls can be made where a reference sequence is unknown and using robust methods to limit undesirable consequences of gross errors in a few clusters.

The AYB base-calling software is written in C and available under the GPL v. 3 licence from http://www.ebi.ac.uk/goldman-srv/AYB/ along with the quality-score calibration tool. A set of utilities for extracting and manipulating CIF format intensity data files, under the same licence as AYB, is also available from http://www.ebi.ac.uk/goldman-srv/ciftools/.

Methods

The two major differences between AYB and other base-callers are its empirical model of the phasing process, potentially allowing the intensities at a given cycle to depend on the entire sequence rather than just a few neighbouring cycles, and its focus on robust algorithms so that sensible base calls are still made even when problems have occurred during a run. Here we describe the underlying statistical model used by AYB, the method of estimation and the techniques used to make the procedure robust.

The foundation of AYB is a mechanistic model of the sequencing process, relating what is observed at each cycle to the underlying sequence of nucleotides. Clusters are analysed in groups, the natural such group being a tile, with various parameters such as the phasing and cross talk matrices assumed to be constant and common to all clusters within each group. Other parameters such as the luminescence and the sequence are specific to each cluster.

Each cluster (indexed by i) is considered to contain homogeneous sequence, represented by the $base \times position$ matrix S_i whose (b, j) entry is one if the base at the j^{th} position of the sequence is base 'b' or zero otherwise[1]. Each column of S_i therefore contains exactly one non-zero entry. The amount of light

[1] So i takes values $1 \ldots \mathcal{C}$, where $\mathcal{C}$ is the number of clusters analysed, typically of the order of 100K per tile; $b \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$; $j \in \{1 \ldots \mathcal{P}\}$, where $\mathcal{P}$ is the number of cycles, i.e. sequence positions for which base calls will be made.

emitted by a cluster in a given cycle is proportional to the number of FLNs bound to the cluster, which in turn is proportional to the number of molecules in the cluster; this cluster-specific scaling is represented by the scalar λ_i, referred to as the luminescence since it also incorporates a factor representing the intensity of light incident on the cluster.

Due to phasing, the molecules within a cluster lose synchronicity with each other and the relationship between position and cycle becomes blurred; the procession from one cycle to the next of an average cluster is described by the *position*×*cycle* phasing matrix P. Each column of P corresponds to one cycle and describes the distribution of sequence positions where the FLNs bind, so the (j, k) entry is the relative proportion of FLNs bound to position j of the sequence on cycle k of the sequencing process[2]. As sequencing progresses, the signal decreases as molecules randomly become inactive and stop contributing (dimming) and this is incorporated into P by scaling its columns so each sums to the proportion of molecules in the cluster expected to be still active. An ideal P would have ones down its leading diagonal with all other elements being zero; a good P will be dominated by its diagonal and each column sum will be close to one. Elements of P are non-negative, and its column sums are ≤ 1.

Finally, the emissions from each cluster are observed via the four channels and the cross talk, the relationship between fluorophore emission and what is observed in each channel, is represented by a *channel*×*base* (4×4) matrix M. Column b of M describes the strength of signal in each of the four channels for a unit emission of the FLN b. In principle the cross talk is determined by the physics of system, and so is assumed fixed throughout the run.

Putting together all the components of the sequencing process model described above, the observed intensities I_i (a *channel*×*cycle* matrix) for cluster i is related to the underlying sequence by the relationship

$$I_i = \lambda_i M S_i P + N + \epsilon_i \tag{3}$$

where N is systematic background noise for all clusters and ϵ_i is the residual error for the fit to the intensities, an observation of a random variable with expectation zero. Both N and ϵ_i are *channel*×*cycle* matrices.

The statistical model described by equation 3 could be fitted to the raw intensity data using a variety of criteria (maximum likelihood, Bayesian techniques, etc.) but we chose a least squares criterion using an iterative approach. The major reasons for the use of least squares are that analytic solutions exist for many of the steps of the iteration, making it computationally efficient, and that the simple Iteratively reWeighted

[2]So $j, k \in \{1 \ldots \mathcal{P}\}$.

Least Squares (IWLS) technique can be used to fit the model in a manner robust to contamination [21]. The IWLS approach is very similar to Ordinary Least Squares (OLS), seeking to minimise the sum of squared errors over all the clusters, except that the squared error for each cluster is weighted and the algorithm proceeds iteratively with the weights being updated between iterations; each iteration is equivalent to the Weighted Least Squares (WLS) criterion, which has an analytic solutions. The weights are defined by a function of how well each cluster fits the model relative to the other clusters, so badly fitting clusters (high residual error) get progressively down-weighted. AYB uses the Cauchy function for weighting but many alternatives have been described and are summarised in the subsection on 'Robust Estimation' in *Numerical Recipes* [22].

Given the statistical formulation, the core of the AYB method can be described by the following six steps, the solution of which will be described in the following sections:

1. Initialise estimates; set all weights to one.
2. Estimate global parameters (cross talk M, phasing P, systematic noise N)
3. Estimate cluster-specific luminescence λ_i
4. Call bases for each cluster, giving sequence S_i
5. Update weights for all clusters
6. Iterate steps 2–5 to refine estimates
7. Compute quality of base calls

1. Initialisation

Initialising to good values greatly helps the speed of the AYB algorithm, reducing the number of iterations needed until a good solution is found. An initial cross talk matrix M can be found from the intensities of an early cycle of the run [5], making the implicit assumption that phasing does not contribute a significant amount to these observed intensities. Since cross talk is primarily determined by physics, having a similar form on different runs and machines, AYB instead initialises M to a known good value rather than approximating a new starting matrix for every new set of data.

The matrix P is split into two components, the phasing and the dimming. Phasing is set so each site is only dependent on its immediate neighbours at a fixed rate; dimming for a given cycle is set to the average over

clusters of the intensities relative to those for the first cycle. Solving equation 3 for $\lambda_i S_i$, assuming that the systematic and random noise (N and ϵ_i) are zero, gives a set of corrected intensities from which bases and luminescence can be estimated. The initial estimate of the base at each position is that which has the greatest intensity, and the luminescence of the cluster is the mean of the intensities of the called bases.

2. Estimation of cross talk, phasing and noise

After fixing all other other parameters, equation 3 is linear in both the cross talk and phasing (a bilinear model) and the WLS estimate for either can be found analytically. If w_i is the current IWLS weight for cluster i, and defining $B_i = \lambda_i S_i P$ and $W_i = \lambda_i M S_i$, then the WLS estimates are:

$$\widehat{M}^t = \left(\sum_i w_i B_i B_i^t - \frac{1}{\widetilde{w}} \widetilde{B}\, \widetilde{B}^t\right)^{-1} \left(\sum_i w_i B_i I_i^t - \frac{1}{\widetilde{w}} \widetilde{B}\, \widetilde{I}^t\right) \tag{4}$$

$$\widehat{P} = \left(\sum_i w_i W_i^t W_i - \frac{1}{\widetilde{w}} \widetilde{W}^t\, \widetilde{W}\right)^{-1} \left(\sum_i w_i W_i^t I_i - \frac{1}{\widetilde{w}} \widetilde{W}^t\, \widetilde{I}\right) \tag{5}$$

$$\widehat{N} = \frac{1}{\widetilde{w}} \sum_i w_i \left(I_i - \lambda_i M S_i P\right) = \frac{1}{\widetilde{w}} \left(\widetilde{I} - \widetilde{W} P\right) \tag{6}$$

where $\widetilde{w} = \sum_i w_i$, $\widetilde{B} = \sum_i w_i B_i$, $\widetilde{W} = \sum_i w_i W_i$ and $\widetilde{I} = \sum_i w_i I_i$. Note that equations 4 and 5 do not depend on N.

The WLS problem does not have a closed-form solution for $\widehat{M}$ and $\widehat{P}$ simultaneously but the estimates can be found by iteratively applying equations 4 and 5, updating W_i and B_i inbetween. Each iteration decreases the (weighted) least squares error and convergence is guaranteed by convexity. The least squares estimate for $\widehat{N}$ is the matrix that makes the mean weighted residual error zero and can easily determined once the estimates $\widehat{M}$ and $\widehat{P}$ have converged.

The columns of M define the strength of signal observed in each channel for a unit fluorophore emission and the number of units is proportional to the cluster luminescence λ_i; the actual scale is arbitrary and the expected emission $\lambda_i M S_i P$ is unchanged if the elements of M are doubled and every λ_i halved. This ambiguity is resolved by choosing the scaling so the determinant of M is 1, and can be accomplished by scaling $\widehat{M}$ and every λ_i after each iteration. There is a similar ambiguity in the scaling of P, whose columns represent the proportion of molecules in a cluster that are contributing to the signal at a given cycle; the phasing and luminescence are conflated since increasing the proportion of molecules contributing to the signal by some factor increases the emission from the cluster by the same factor. Again, this ambiguity can be resolved by scaling P so its determinant is 1 and incorporating the scaling factor into the luminescence.

The calculations in equations 4, 5 and 6 require several matrix multiplications for every cluster, of which there are generally many more than the number of cycles ($\mathcal{C} \gg \mathcal{P}$), and so are computationally demanding for densely populated tiles. In the additional material we show show these solutions can be rearranged so many of the multiplications can be done in advance of the iteration, meaning each step of the iteration can be performed rapidly.

3. Estimation of luminescence

The estimation of luminescence for each cluster could be integrated into the estimation procedure for the cross talk and phasing but, despite that iteration being based on the robust IWLS procedure, it reduces to OLS for the luminescence since the weights used are per-cluster and so affect all cycles equally. A major weakness of OLS estimation is that it is not very robust in the presence of contamination, outliers strongly influencing the estimate, and such problems do occur in real data: "dust" contaminating the flow cell and producing extremely bright peaks for one cycle of one cluster, for example. Instead, AYB estimates luminescence separately from the intensities, after correcting them for the effects of cross talk and phasing, using a WLS approach with per-position weights $\omega_{i;j}$ (for position j of cluster i) derived using the Cauchy function specific to each cluster independently of the previous weights. The corrected intensities for cluster i, C_i, can be found from the observed intensities by rearranging equation 3:

$$C_i = M^{-1}\left(I_i - N\right)P^{-1} = \lambda_i S_i + \epsilon_i^* \tag{7}$$

where $\epsilon_i^* = M^{-1}\epsilon_i P^{-1}$ is the residual error of the corrected intensities.

If the corrected intensities for position j of cluster i are $C_{i;\bullet j}$ (the j^{th} column of C_i) and the corresponding sequence is $S_{i;\bullet j}$, then the WLS estimate of the cluster luminescence is

$$\widehat{\lambda}_i = \frac{\sum_j \omega_{i;j} C^t_{i;\bullet j} S_{i;\bullet j}}{\sum_j \omega_{i;j}}$$

which is simply the weighted average of the corrected intensities for the called bases.

4. Base calling

As well as being linear in the cross talk and phasing, the observed intensities in equation 3 are also a linear function of the sequence and so calling bases also requires fitting of a linear model. As described, equation 3 assumes that each element of the random error is independent and identically distributed (IID) but this is not found to be the case in real data (results not shown). The violation of the IID assumption is

not a problem when estimating the cross talk and phasing, since these estimates are produced from a large number of independent clusters and so the random error is small, but is much more significant when trying to estimate the sequence since there are many fewer, dependent, observations. Forcing the IID assumption onto the random noise produces poor base calls (results not shown) and so correlation between the elements of ϵ_i^* must be taken into account — a General Linear Model (GLM).

The sequence that minimises the generalised least square error of equation 3 also minimises the generalised least square error of equation 7 and this latter formulation, relating the sequence to the corrected intensities, is more convenient to work with. Finding the minimum generalised least square is a type of constrained binary quadratic programming problem and so difficult to solve exactly. Instead of solving directly, we make the additional assumption that there is no correlation in the noise between different read positions. The covariance V_j of the random error between nucleotides at position j can easily be estimated from the residual errors ϵ_i^*. The estimate of the base $b_{i;j}$ at position j of cluster i is the one that minimises the least squared error

$$L_{i;bj} = (C_{i;\bullet j} - \lambda_i 1_b)^t \, V_j^{-1} \, (C_{i;\bullet j} - \lambda_i 1_b)$$

where 1_x is a vector with all elements zero other than the x^{th} element, which is one. The optimal base is determined exhaustively, independently of the other clusters and positions, so the complexity of calling bases is linear in the length of the read. The sequence matrix for each cluster can then be constructed from the called bases by setting the appropriate entries of S_i as described earlier.

5. Updating weights

The weighting of the clusters plays an important part in making the AYB method robust to contamination and other misleading observations, reducing their influence on the parameter estimates. The weight for each cluster is calculated, after all model parameters have been fitted, using the Cauchy function so $w_i = 1/(1 + L_i^2/2\sigma^2)$ where L_i^2 is the least square error for cluster i and σ^2 is a measure of the variation in least square error (the variance, for example). In contrast to OLS, where every cluster would receive a weight of one, the weighting function means that only perfect observations, those with a least square error of zero, receive full weight whereas worse-fitting observations receive progressively lower weights.

6. Iteration and termination

As the luminescence and individual bases are estimated using a different criterion to that used to estimate the cross talk and phasing, the least squares error when the parameter estimation and base calling steps

are iterated is not guaranteed to decrease. Theoretically this can lead to problems with convergence but this was not found to be the case, a small number of cycles sufficing to get good estimates. Numerical experiments suggest that three to five iterations are sufficient, with little change in accuracy for addition iterations (results not shown).

7. Quality of calls

To differentiate between good and bad reads, each base is assigned a quality score — a measure of the probability that it has been correctly called. Commonly these are reported as Phred scores: $Q_{\text{Phred}} = -10 \log_{10} e$, where e is the probability of a base being incorrect [12]. It is trivial to convert these scores to and from probabilities, so one only needs to assess the probability of each call being incorrect.

A quality score should combine two pieces of information: the relative confidence between bases that they are correct, and whether the model is a good representation of the data and can reliably assess base confidence. We treat base calling as a model selection problem, choosing between the four models 'A', 'C', 'G' and 'T' for each cycle of each cluster, and apply Bayes theorem to get the posterior probability $p_{i;bj}$ that the base at position j in cluster i is b:

$$p_{i;bj} = \frac{\pi_b f_{i;bj}}{\sum_{x \in \{A,C,G,T\}} \pi_x f_{i;xj}} \tag{8}$$

where $f_{i;xj} = \exp(-\frac{1}{2} L_{i;xj})$ is the (scaled) probability density of the observed intensities and π_x is the prior probability of base x. The particular form of $f_{i;bj}$ comes from assuming that the random error is distributed as multivariate normal, as assumed for calling the bases, but other elliptical distributions would also be appropriate.

Equation 8 does not incorporate any information about how well all the four models A, C, G & T fit a particular position; nor does the original probabilistic model allow for rare large deviations. It is possible that all four models could fit the data badly, implying the corrected intensities do not represent sequence, but that one model is favoured relative to the others and a base is called with spurious confidence. A small correction term, τ, is incorporated to avoid this problem, representing the chance that the observation is contaminated and so the model is a poor description of what actually happened. Adding such a constant defines an M-estimator [23] and is related to the robust likelihood approach to estimation (attributed to Colin Mallows [24], pg 59–60). Analogous to the posterior probability of a base being correct, a modified probability of correctness, $p^*_{i;bj}$, is defined by:

$$p^*_{i;bj} = \frac{\pi_b \left(\tau + f_{i;bj}\right)}{\sum_{x \in \{A,C,G,T\}} \pi_x (\tau + f_{i;xj})} \tag{9}$$

If a base b fits well then it has a small least squared error and the value of $f_{i;bj}$ is close to 1 and so dominates τ. When all bases are unlikely and τ dominates, the effect is to pull the posterior probabilities back towards the prior — rather than being over-confident in an ill-fitting base, no effective prediction is made.

Competing interests

None declared.

Authors contributions

TM designed the study, derived the statistical models and implemented them in software, carried out the comparisons and performed the statistical analysis, and drafted the manuscript. NG conceived and helped to design the study and to draft the manuscript. All authors read and approved the final manuscript.

Additional Files

The following additional data are available with the online version of this paper. Additional data file 1 is a document describing and deriving how to calculate least-squares estimates of the phasing and cross-talk matrices iteratively in a manner that does not require a summation over all clusters during each step of the iteration.

Acknowledgements

The development of AYB has benefited from the help of many people, for both advice and feedback about performance '*in the wild*'. In particular we would like to thank Ewan Birney (EMBL-EBI); Jonathon Blake (EMBL); Gordon Brown (CRI), Kevin Howe (formerly CRI); Tony Cox, Klaus Maisinger, Lisa Murray (Illumina Inc.); Christophe Dessimoz & Christian Ledergerber (ETH Zürich); Richard Durbin, Tom Skelly (Sanger Institute), Nava Whiteford (formerly Sanger Institute); David Van Heel (Blizard Institute).
The 76 cycle ϕX174, 151 cycle *H. sapiens* and 76 cycle *B. pertussis* data sets were provided by the Sanger Institute. Other data as indicated in the text.
The development and maintenance of AYB is supported by a Wellcome Trust Technology Development grant WT088151MA.

References

1. Varela I, Klijn C, Stephens PJ, Mudie LJ, Stebbings L, Galappaththige D, van der Gulden H, Schut E, Klarenbeek S, Campbell PJ, Wessels LFA, Stratton MR, Jonkers J, Futreal PA, Adams DJ: **Somatic structural rearrangements in genetically engineered mouse mammary tumors**. *Genome Biology* 2010, **11**:R100.
2. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV: **The challenges of sequencing by synthesis**. *Nature Biotechnology* 2009, **27**(11):1013–1023.
3. Metzker ML: **Sequencing technologies – the next generation**. *Nature Reviews Genetics* 2010, **11**:31–46.
4. Ledergerber C, Dessimoz C: **Base-calling for next-generation sequencing platforms**. *Briefings in Bioinformatics* 2011. [Epub early access].
5. Li L, Speed T: **An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing**. *Electrophoresis* 1998, **20**:1433–1442.
6. Kao WC, Stevens K, Song YS: **BayesCall: a model-based basecalling algorithm for high-throughput short-read sequencing**. *Genome Research* 2009, **19**(10):1884–1895.
7. Erlich Y, Mitra PP, de la Bastide M, McCombie WR, Hannon GJ: **Alta-Cyclic: a self-optimizing base caller for next-generation sequencing**. *Nature Methods* 2008, **5**:679–682.
8. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F: **Probabilistic base calling of Solexa sequencing data**. *BMC Bioinformatics* 2008, **9**:431.
9. Kircher M, Stenzel U, Kelso J: **Improved base calling for the Illumina Genome Analyzer using machine learning strategies**. *Genome Biology* 2009, **10**:R83.
10. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.
11. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**:1851–1858.
12. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Research* 1998, **8**:186–194.
13. Abnizova I, Skelly T, Naumenko F, Whiteford N, Brown C, Cox T: **Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing**. *J Bioinform Comput Biol.* 2010, **8**(3):579–591.
14. Whiteford N, Skelly T, Curtis C, Ritchie M, Lohr A, Zaranek A, Abnizova I, Brown C: **Swift: primary data analysis for the Illumina Solexa sequencing platform**. *Bioinformatics* 2009, **25**(17):2194.
15. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research* 2008, **18**:821–829.
16. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061–1073.
17. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK: **Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants**. *Genome Research* 2009, **19**:2308–2316.
18. Rothberg J, Hinz W, Johnson K, Bustillo J: **Methods and apparatus for measuring analytes using large scale FET arrays** 2007. [US Patent App. 12/002,291].
19. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al.: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**(5910):133.
20. Clarke J, Wu H, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nature nanotechnology* 2009, **4**(4):265–270.
21. Agresti A: *Categorical Data Analysis*. John Wiley & Sons, second edition 2002.
22. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C*. Cambridge University Press, second edition 1992.

23. Huber PJ: *Robust Statistics*. John Wiley & Sons 1981.

24. Owen A: *Empirical Likelihood*. Chapman & Hall/CRC 2001.

25. Wilson EB: **Probable inference, the law of succession, and statistical inference**. *J. Amer. Stat. Assoc.* 1927, **22**:209–212.

Figures

Figure 1 - Phasing progression across 151 cycles

Progression of phasing for 151 cycles of a GA-II run containing human sequence, split into average contributions from pre-phased, post-phased and in-phase (correct) molecules over all clusters. The bar chart above the phasing graph shows how the proportion of the signal contributed by the correct read position decreases as the cycle number increases, from almost one (no phasing) on the first few cycles to less than 0.5 (phasing dominates) on the last 61 cycles.

Figure 2 - Quality of base calls before and after calibration

Quality of calls for 76 cycles of ϕX174 data before (left) and after calibration (right). The actual error rate is estimated from real data by mapping the reads back to a known reference, with 99% confidence intervals calculated by transforming Wilson's interval for the binomial proportion [25]. The dashed lines are the lines of best fit, minimising the least squared error weighted by the number of bases in each bin, and the solid line $y = x$ represents perfect calibration. The calibrated graph shows calibration curves using three different lanes, the best fitting being calibration of the data onto itself.

Figure 3 - Comparison of error rates between Bustard, Swift and AYB

Comparison of per-cycle error rates at higher cycle numbers between AYB, Bustard and Swift from 900k unfiltered reads of 76-cycle ϕX174 data. The inset graph shows the proportion of mapped reads with a given number of differences compared to the reference and the Venn diagram shows the overlap in successfully mapped reads between the methods. A further 230,290 reads were not mappable using any of these base callers.

Figure 4 - Number of reads mapped to B. pertussis

Number of mapped reads and base substitution errors in 4.2 million reads of 76-cycle paired-end *B. pertussis* data relative to a reference genome. The base callers Bustard, Swift and AYB are compared on ends 1 and 2 of the reads by the number of differences to the reference, to a maximum of five differences,

with the total percentage of mapped reads mapped displayed at the top of each bar. Results for both the full reads and reads trimmed to the first 50 cycles are shown.

Figure 5 - Comparison of error rates between Ibis and AYB

The percentage of mapped reads called by Ibis from 76 cycles of ϕX174 data that contain an error on a given cycle (bars, left axis) and the difference between the percentage error of AYB and Ibis (circles, right axis). On the final cycle, Ibis has an error rate of 4.7% whereas AYB has an error rate of 3.4%, giving a difference of -1.2%.