Alta-Cyclic - an Improved Solexa Base caller

For more accurate and longer reads

Yaniv Erlich, Partha P. Mitra, Melissa delaBastide, W. Richard McCombie & Gregory J. Hannon
Watson School Of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Poster in a nutshell

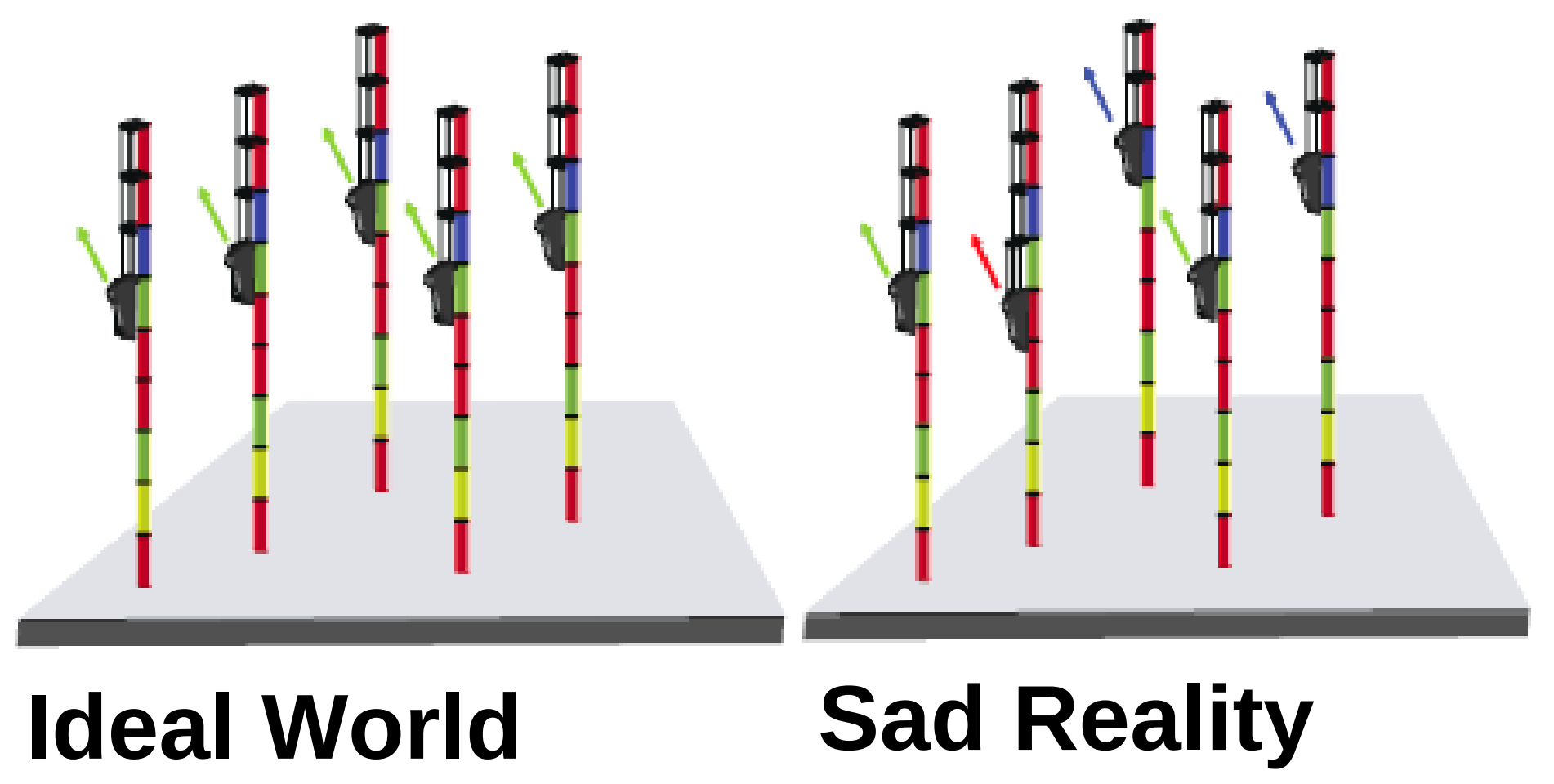
1. A comprehensive model for the noise factors in Illumina 1G was built
2. Alta-Cyclic learns the noise factors and finds optimal parameters
3. Alta-Cyclic pushes the technology beyond its current limitation

Abstract

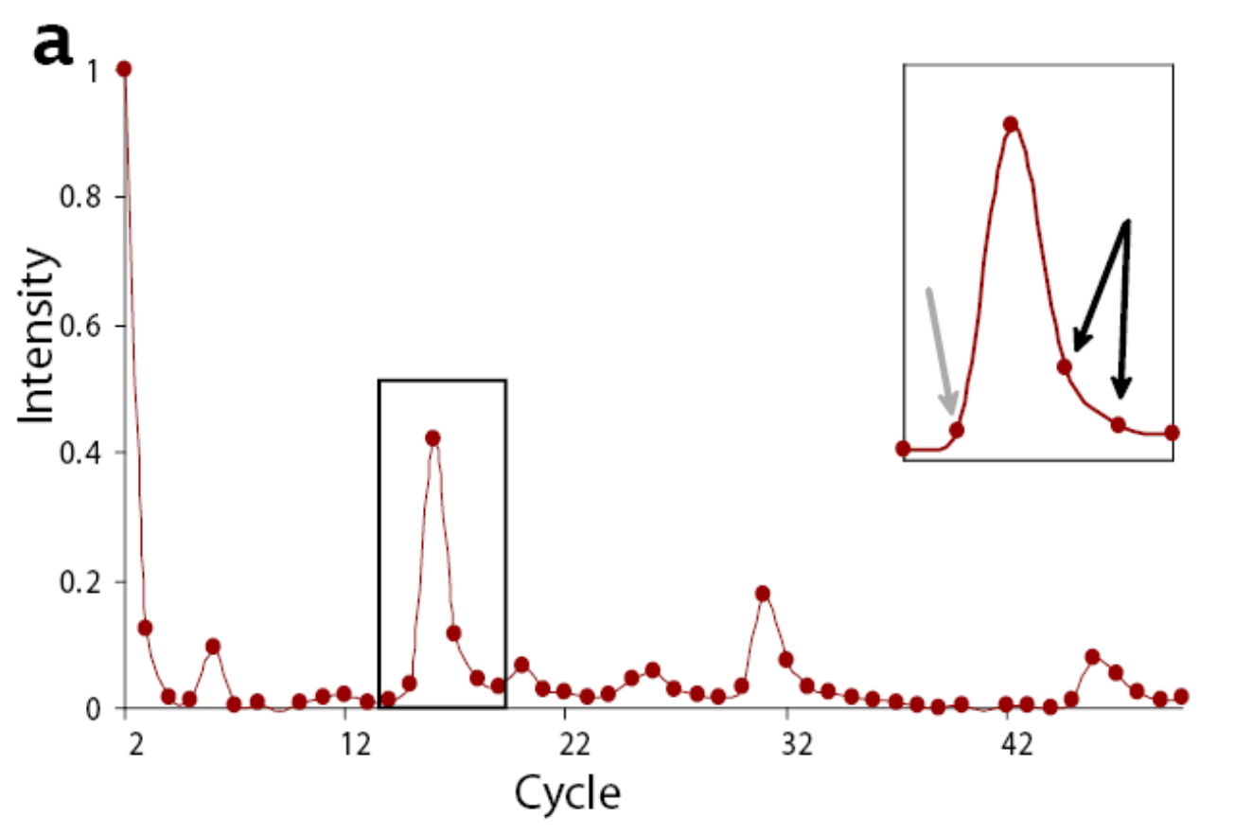
The power of next-generation sequencing is limited by high error rates, as compared to conventional sequencing, and short read lengths. Here, we sought improvements in sequence determination from the Illumina 1G instrument. We systematically analyzed sources of noise that cause error rates to climb with cycle number. We found de-synchronization of polymerases (phasing), progressive loss of signal (fading) and cycle-depending increases in the fluorescent crosstalk among individual nucleotides. We developed a novel base caller, termed Alta-Cyclic, that allows compensation of these noise sources and is based on machine learning approach. This allows valuable 78-base reads, increases the number of accurate reads by more than 4 folds, and reduces systematic biases that degrade the ability to confidently identify sequence variants. Though the analysis presented here is specific to the Illumina 1G machine, the general strategies may also be applicable to other next generation-platforms.

Part I: Revealing the noise sources

(1) Polymerase spread(Phasing):

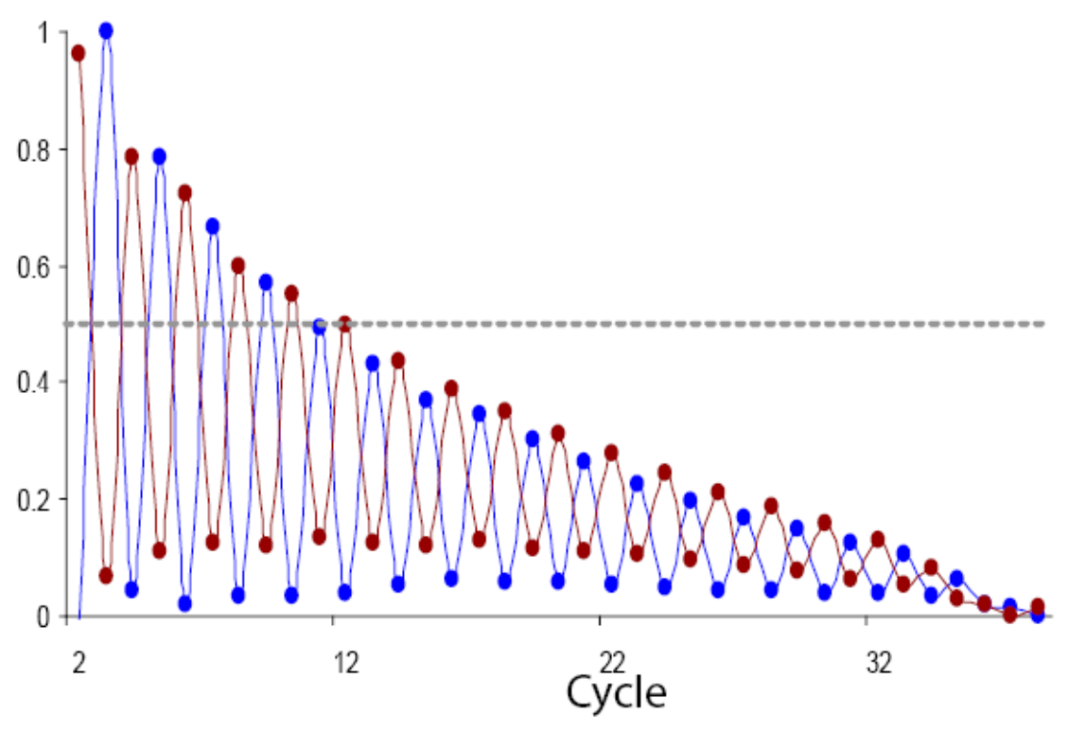


Schematic representation of polymerase spread ("phasing"). A DNA cluster is comprised of identical DNA fragments (colored rods) that are attached to the flow cell. In the ideal situation **(left)** the DNA polymerases (black ovals) are all at the same position and incorporate the same type of nucleotide, which leads to a coherent signal **(right)** The actual situation includes lagging (blue arrows) and leading (red arrow) polymerases. The polymerases spread across several positions, and transmit a mixture of signals.



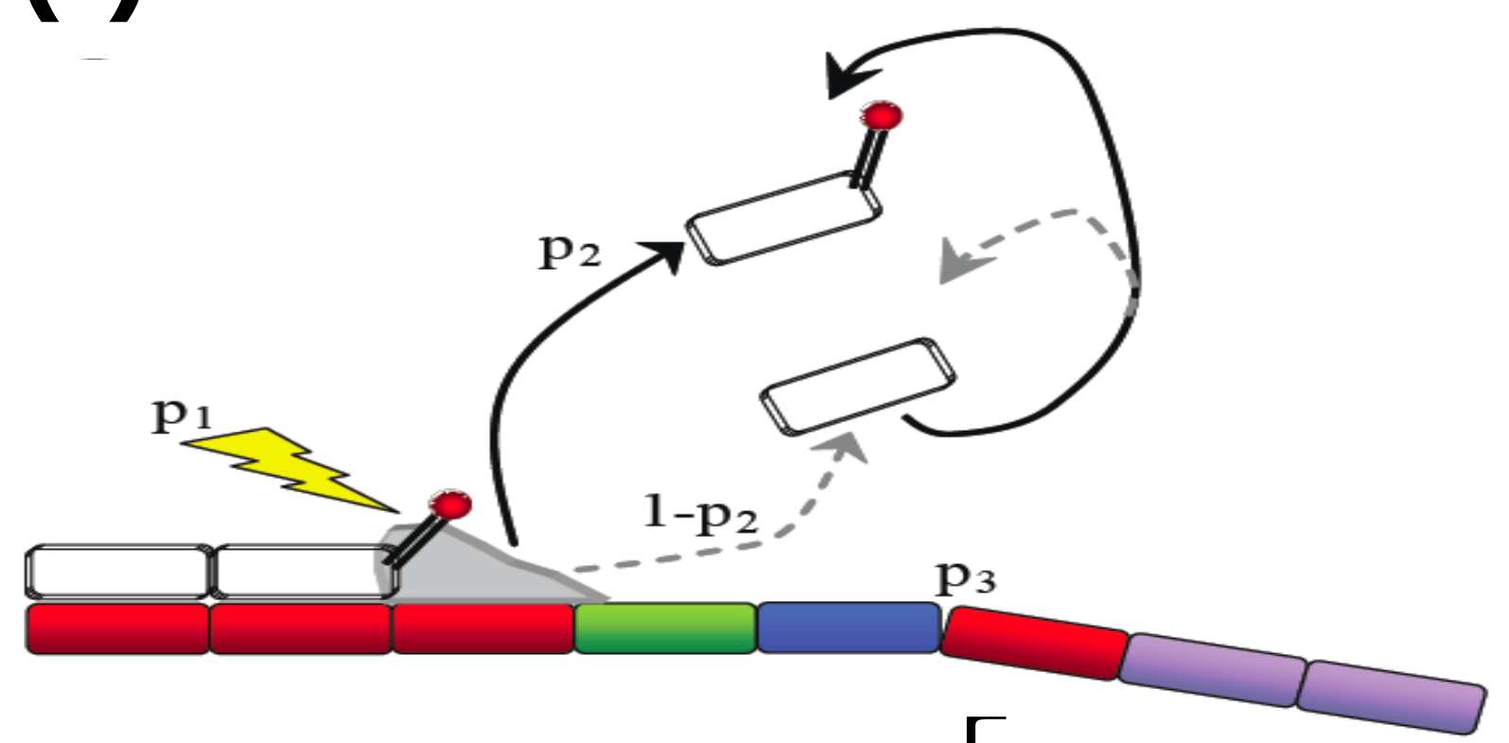
Output of the impulse response test. Shown are the averaged intensities of the cytosine channel from polonies with a delta function sequence. The polymerase spread appears as an anticipation signal that precedes the position of the C in the sequence (gray arrow) and persists in subsequent cycles (black arrows). The diffusive properties of the spread are shown by relative increase in the residual signal in adjacent cycles to the actual C position.

(2)Signal loss (Fading):



Signal decay (fading) is reflected in intensity reads from microsatellite sequences. Shown are the average intensities of the cytosine (red) and adenine (blue) channels from the microsatellite sequence ACAC. In the absence of fading the signals should converge to half of their initial intensities (gray line). Nevertheless, the signal exponentially decays toward zero, which indicates material loss or another mechanism that gradually disrupts the signal.

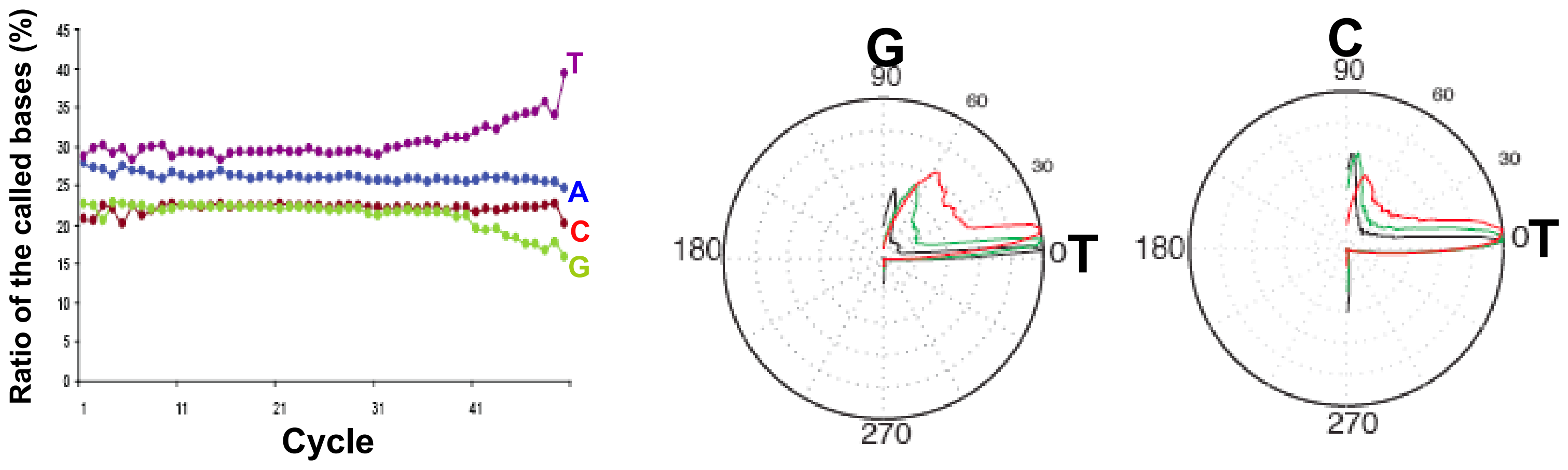
(*)Random walk:



$$P(n=position, t=cycle)=e^{-p_3 t}\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{-i\omega n}\left[1-p_1+\frac{p_1p_2e^{i\omega}}{1-(1-p_2)e^{iw}}\right]^t$$

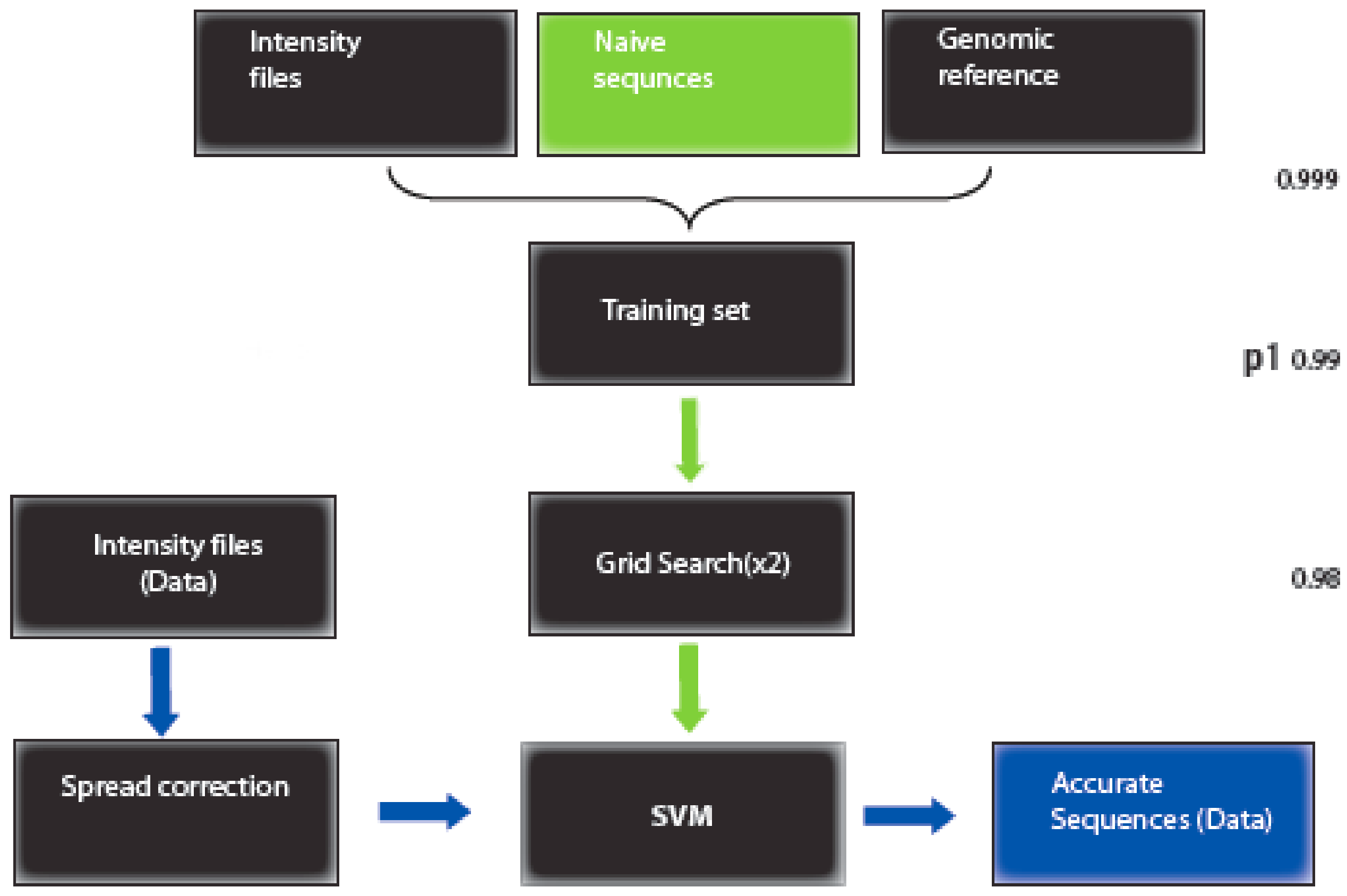
Random walk model. Schematic illustration of the random walk model is shown. In the initial state the last nucleotide of the nascent DNA strand (white rectangles) is blocked and has a fluorophore label (red ball). The block is removed with probability p1, allowing for a nucleotide to be incorporated in the next cycle. A blocked nucleotide is incorporated with probability p2, and the non-blocked with probability (1-p2). If an non-blocked nucleotide is incorporated the process continues (gray arrows), until a blocked nucleotide is incorporated (black arrows). In addition, the template is lost with p3 probability, due to strand breakage or another processes

(3) Fluorophore cross talk changes:

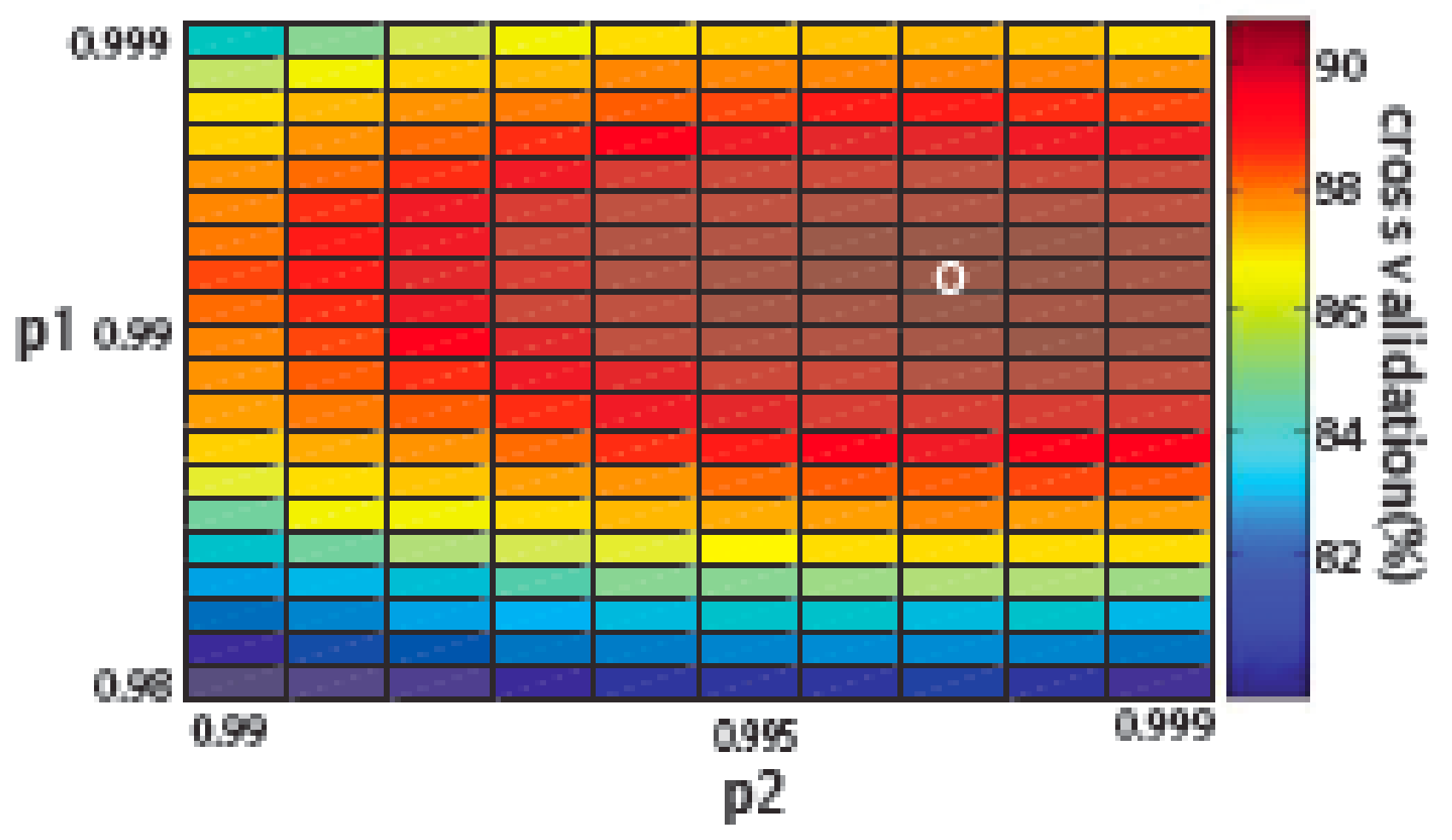


Crosstalk matrix changes. **(left)** The percentage of called bases in the phi X library is plotted as a function of cycle number using the Illumina base caller. The T and the G calls have strong opposite trends. **(middle)** Polar histograms present the ratio between channel intensities correlated with the base preference. In the first cycle (black) the two lobes are orthogonal which indicates correct crosstalk correction. In later cycles (green and red) the G lobe starts to migrate toward the T lobe, which indicates a change in the crosstalk matrix. **(right)** This phenomena is not found between the C and the T channels.

Part II: Alta-Cyclic architecture



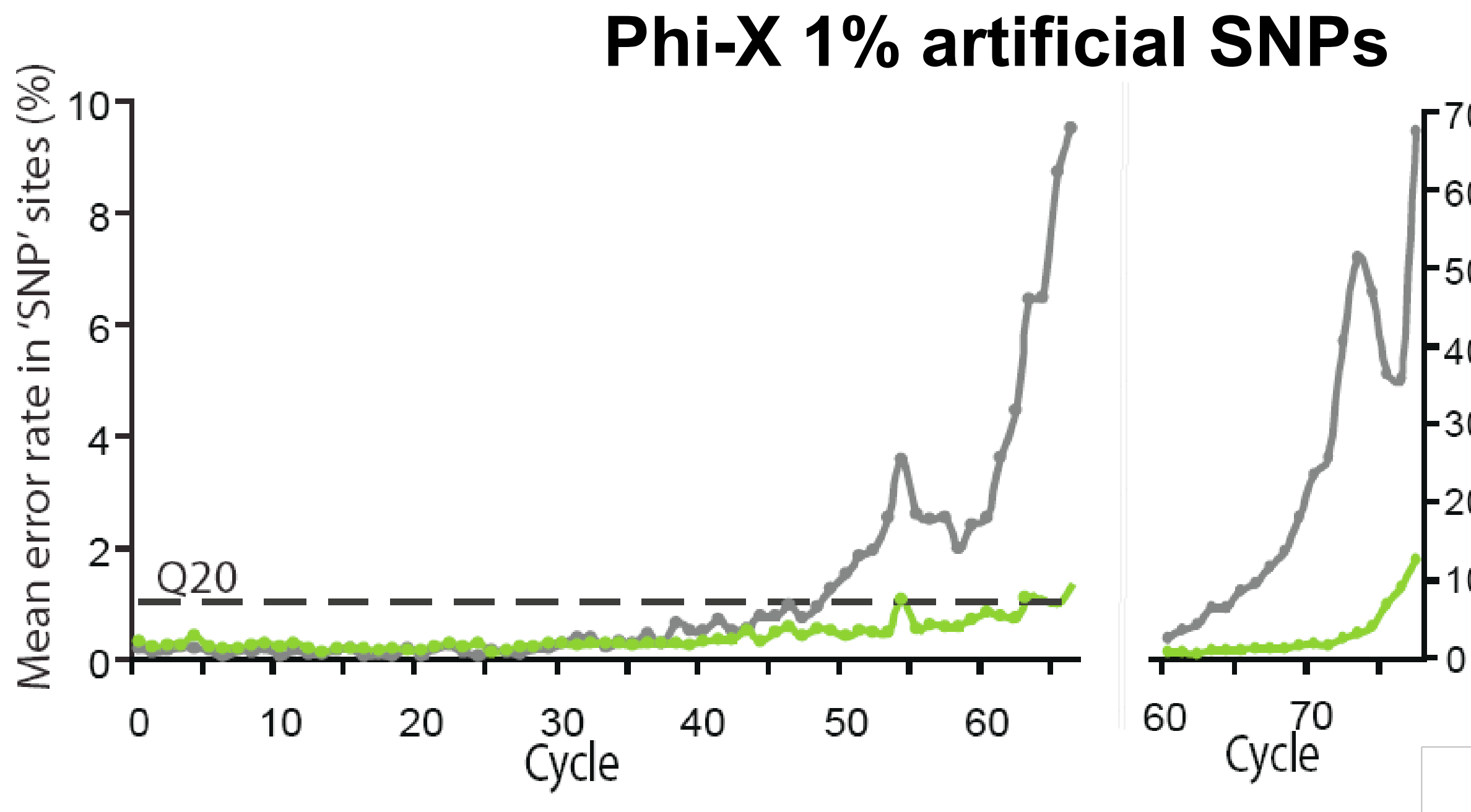
The Alta-Cyclic base caller. For Alta-Cyclic base calling, naïve sequences are aligned to the reference genome and a set of desired output is obtained. This set, with the intensity files that correspond to the naïve sequences, constitute the training set. Parameter optimization of the random walk model and of the SVM machine is obtained by two grid searches. The corrected training data is sent to build a series of SVMs, each of which corresponds to a cycle (green arrows). After the training phase, the intensity files of the desired library undergo deconvolution to correct for the polymerase spread using the values that were found in the grid search. Then, the processed intensities are sent for classification with the SVMs (blue arrows), the output is processed, and sequences and quality scores are reported

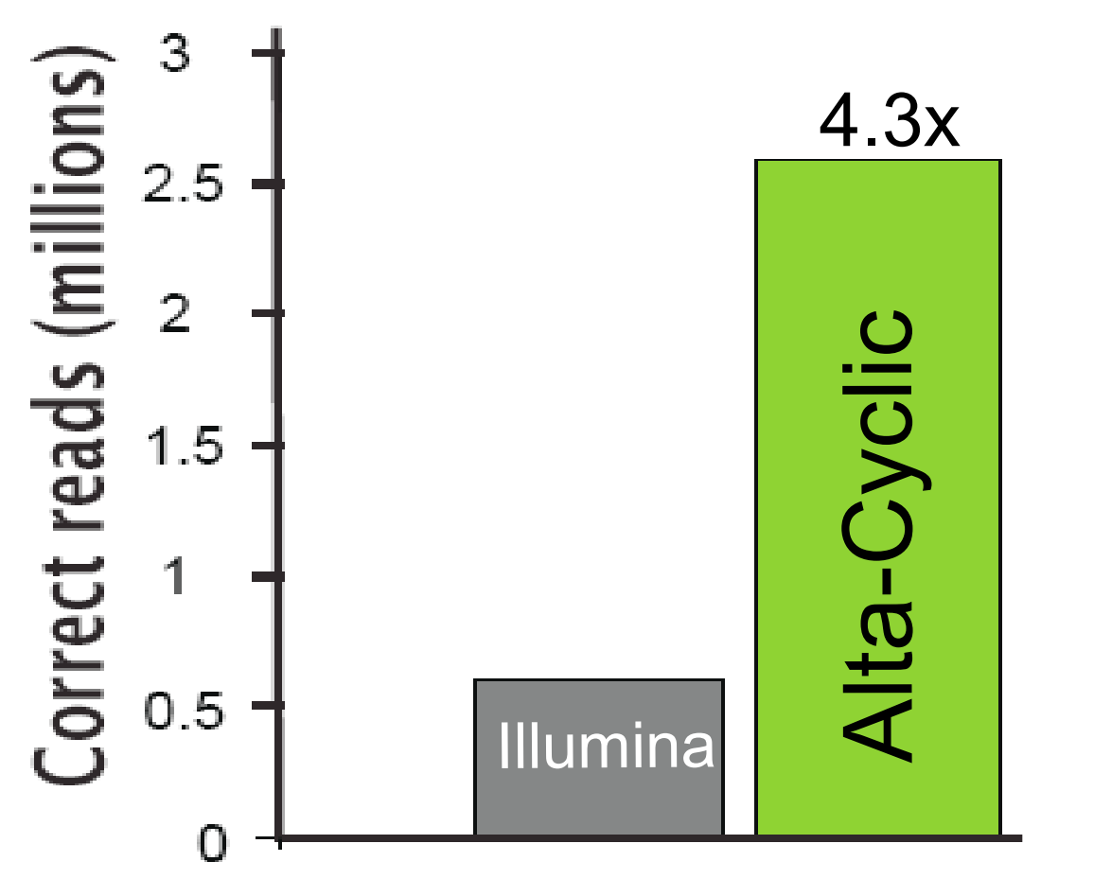


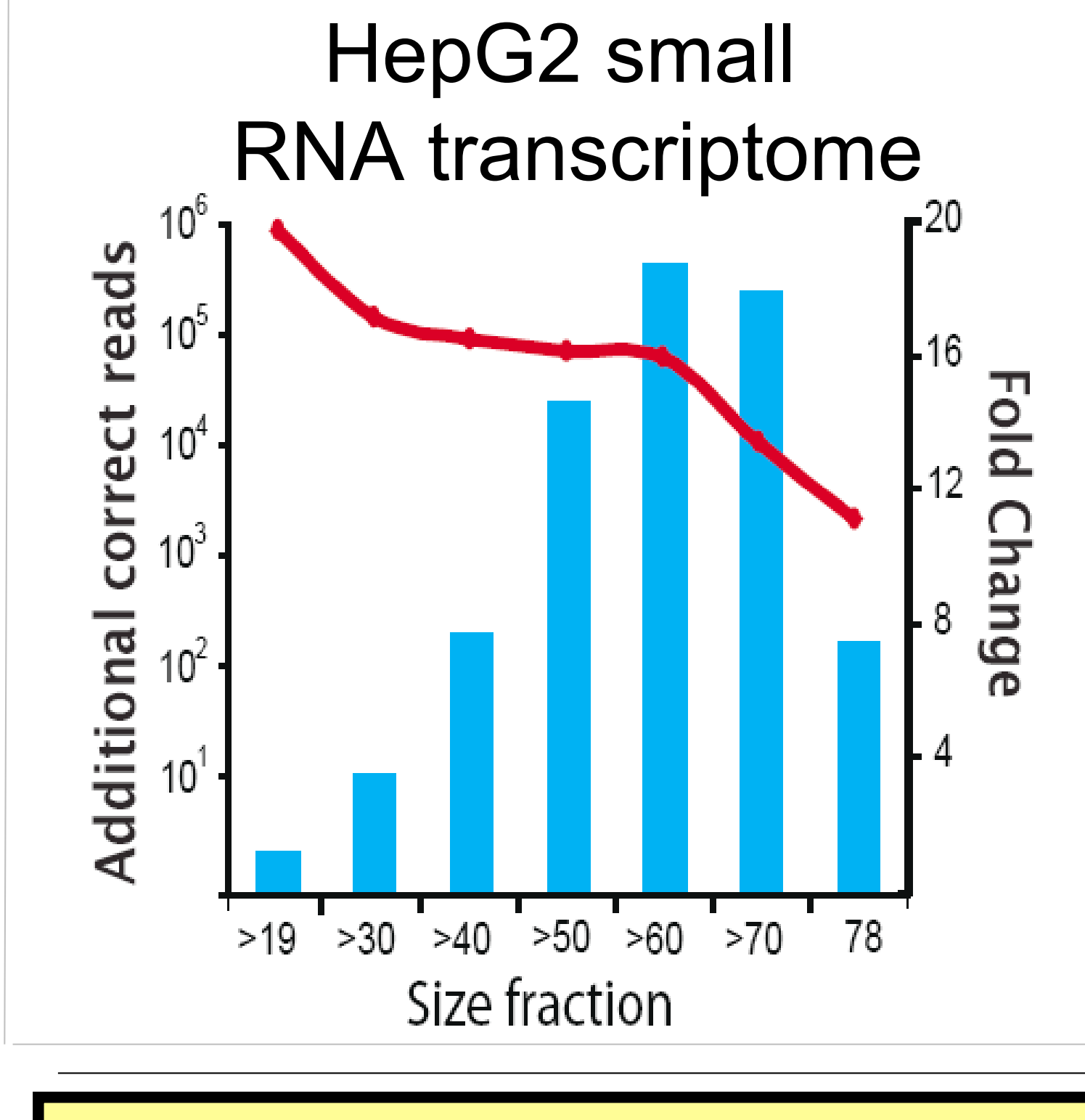
Grid Search Grid search results for optimization of the random walk parameters. The Y-axis corresponds to p1 and the X-axis to p2. The color of each cell indicates the cross-validation rate that was achieved. The white circle shows the values that were chosen (p1 = 0.992, p2 = 0.997).

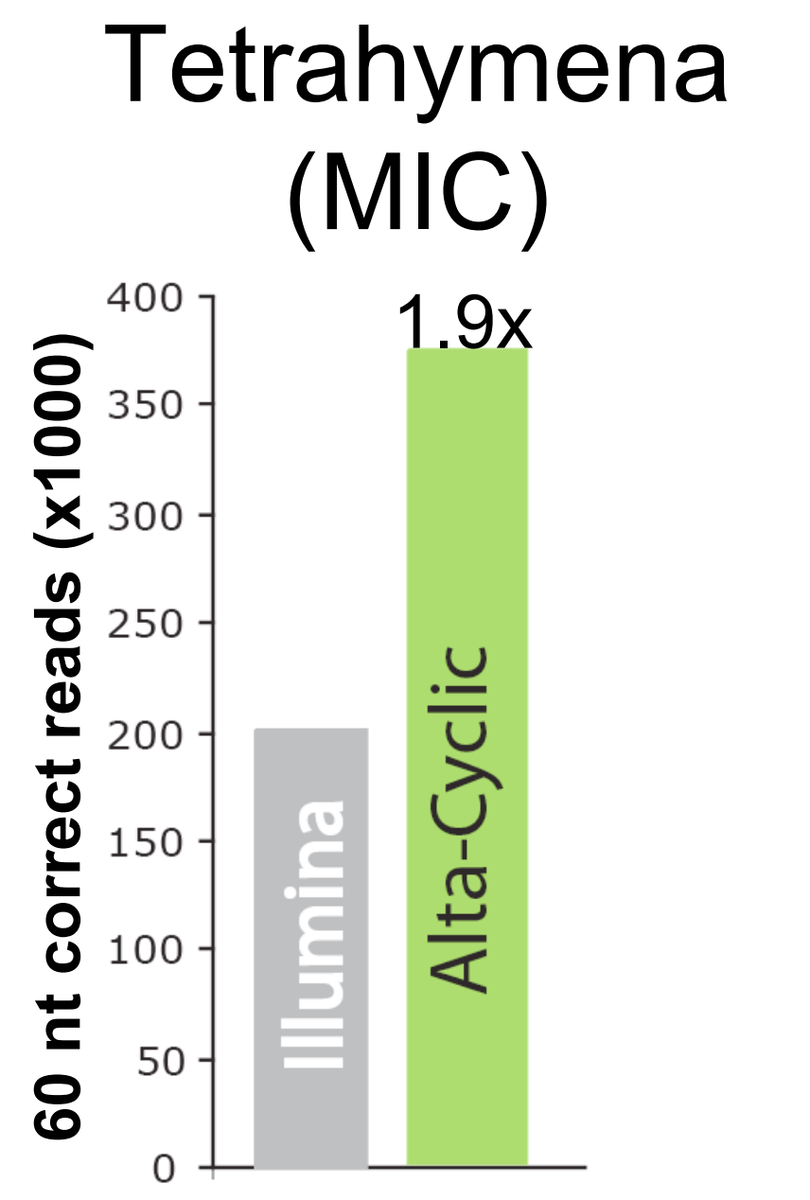
Part III: Results

78 cycles on GAII machines

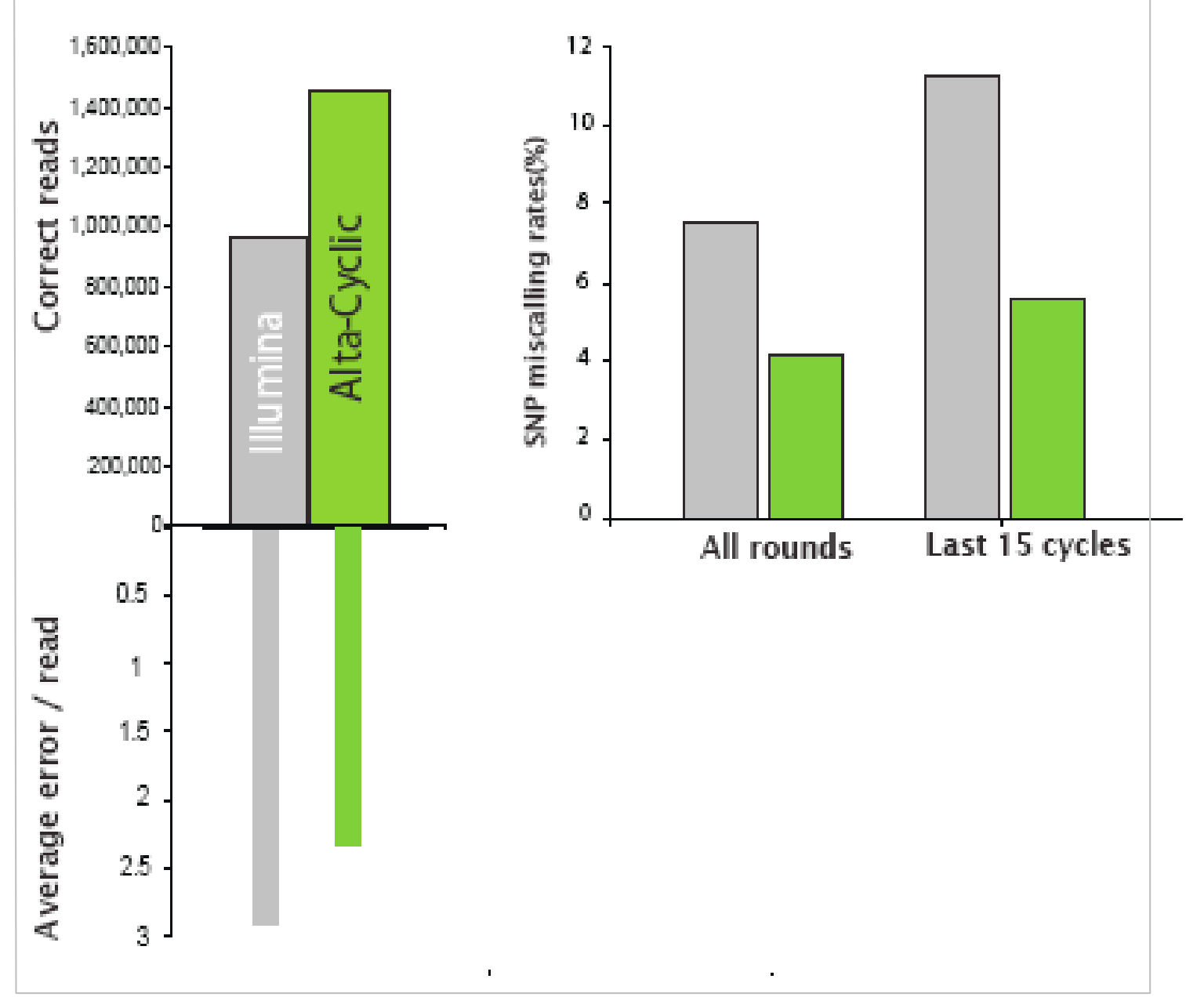








50 cycles on GAI



Conclusions & Future Directions

* Alta-Cyclic provides useful information from 78 cycles Illumina 1G runs
* It improves SNP detection, de-novo sequencing and expression analysis
* The software will be available upon publication.