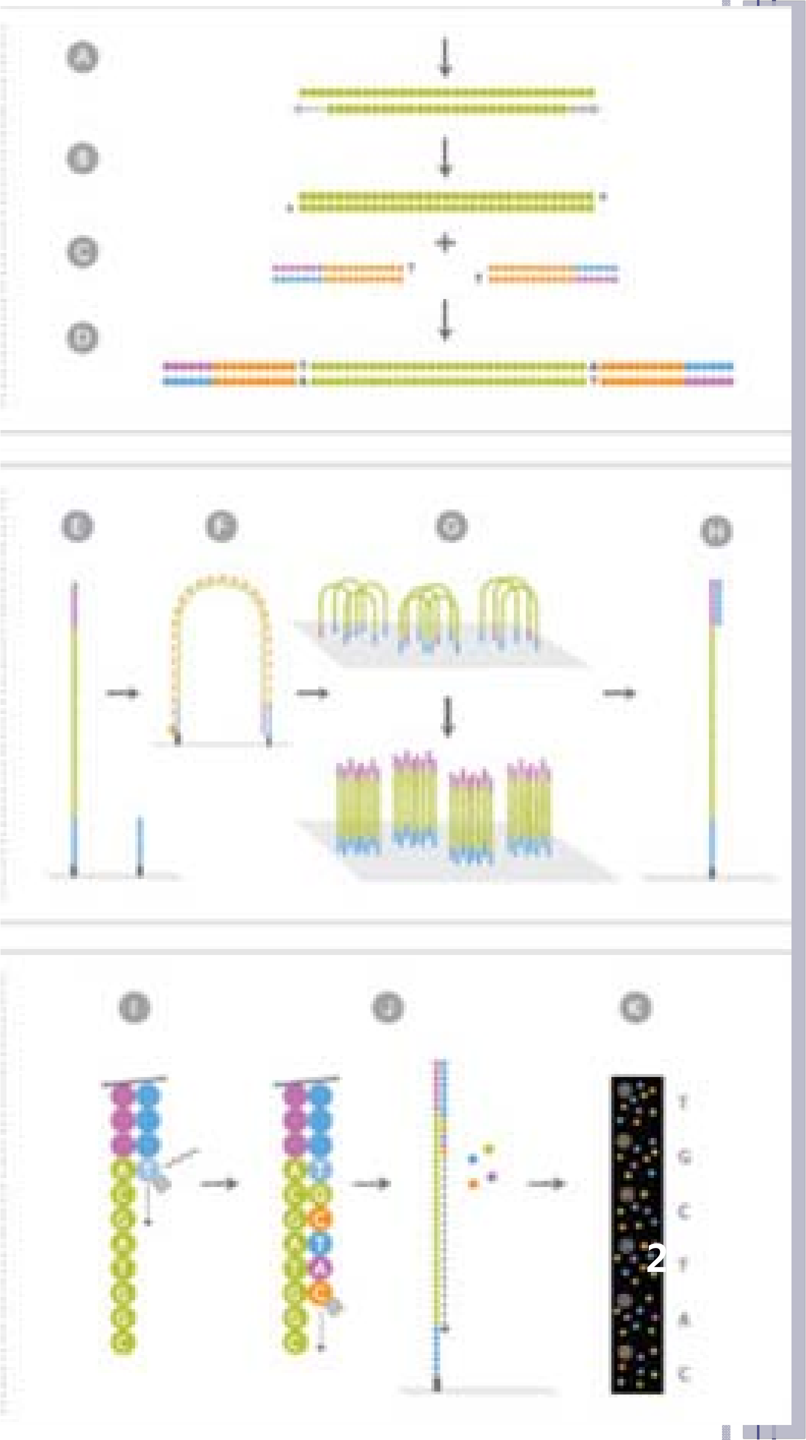
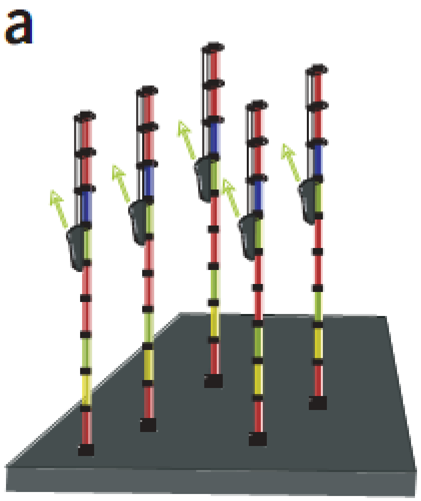
ALTA-CYCLIC: A SELF-OPTIMIZING BASE CALLER FOR NEXT-GENERATION SEQUENCING

Presenter: Jian Zhao

Illumina System Workflow

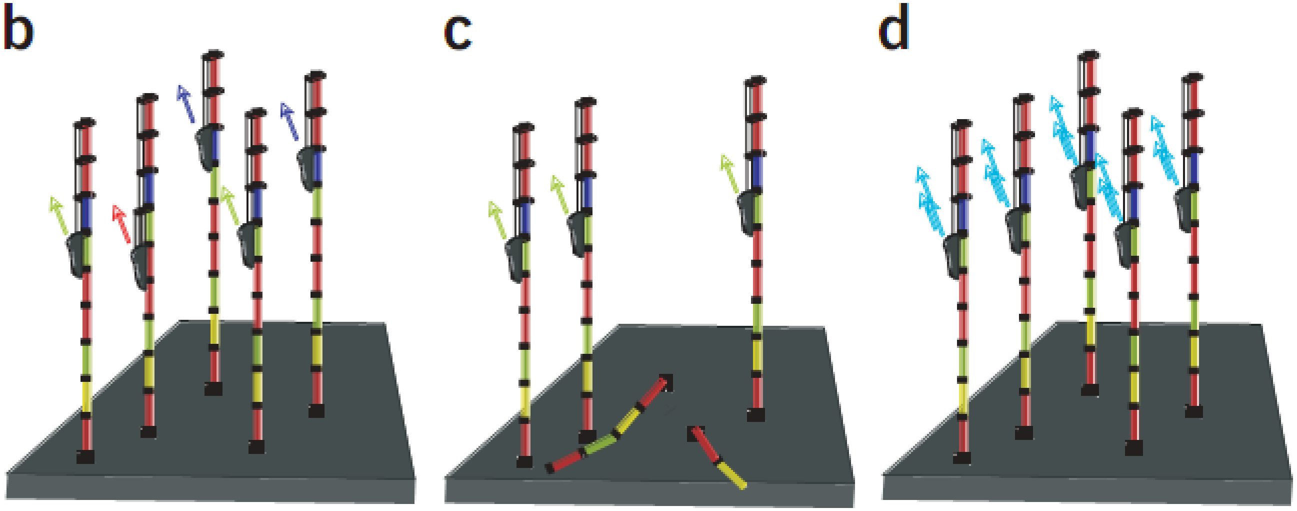
- Library preparation
- Cluster generation
- Sequencing





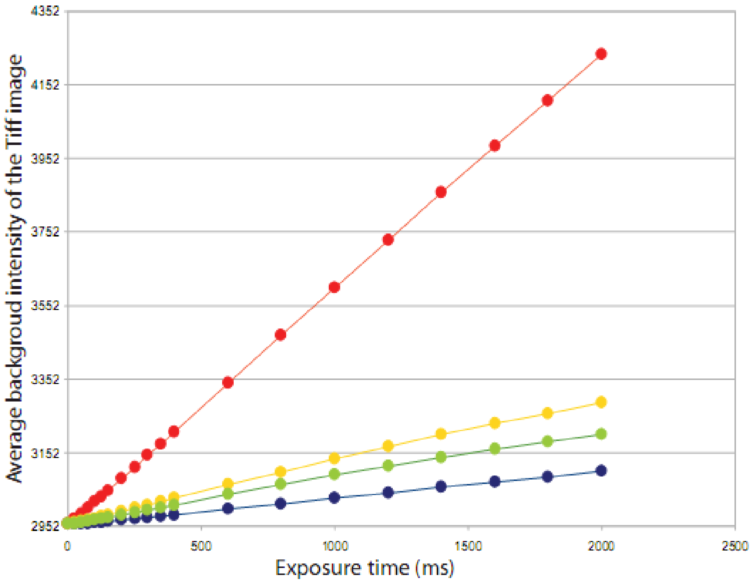
Noise factors

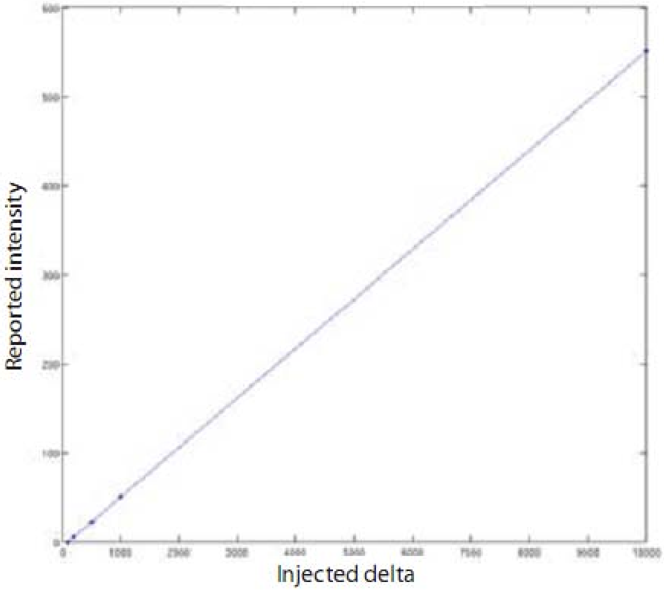
- Phasing noise
  - Leading, lagging
- Fading noise
  - Exponential decay in fluorescent signal
- Cycle-dependent change in fluorophore cross-talk



Linearity of the intensity values

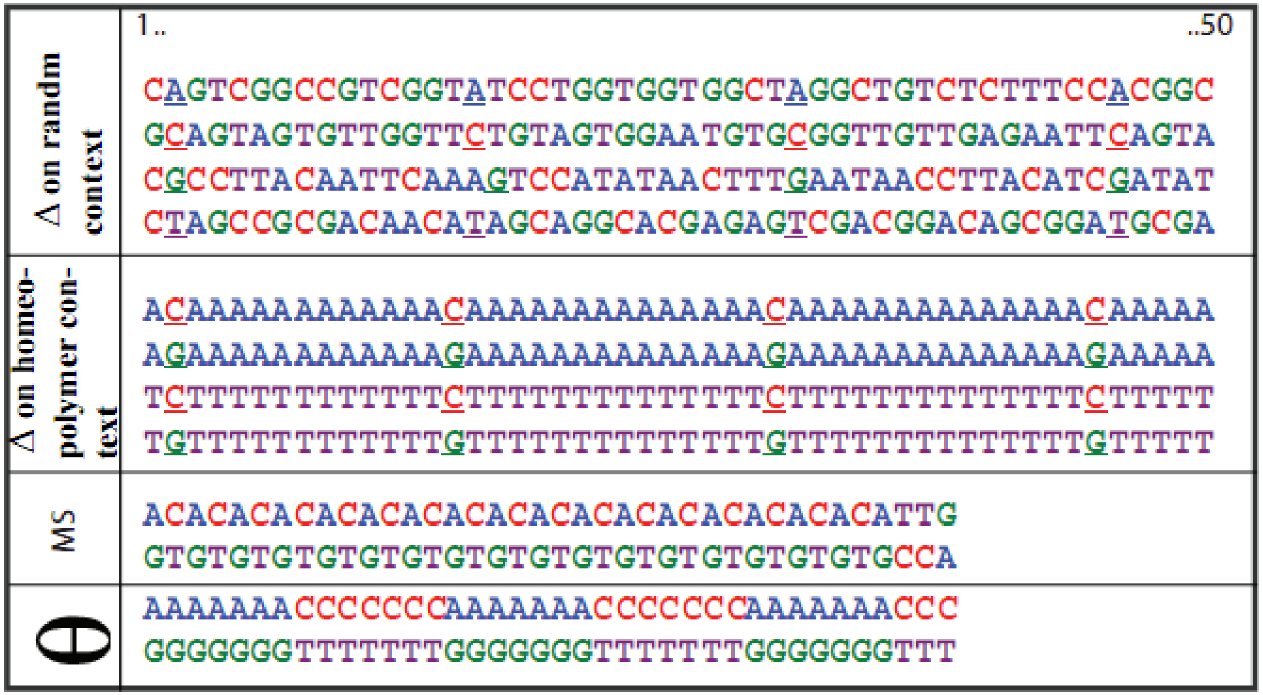
- The optic chain is linear
- Firecrest applies linear transformation to the image





Impulse response analysis

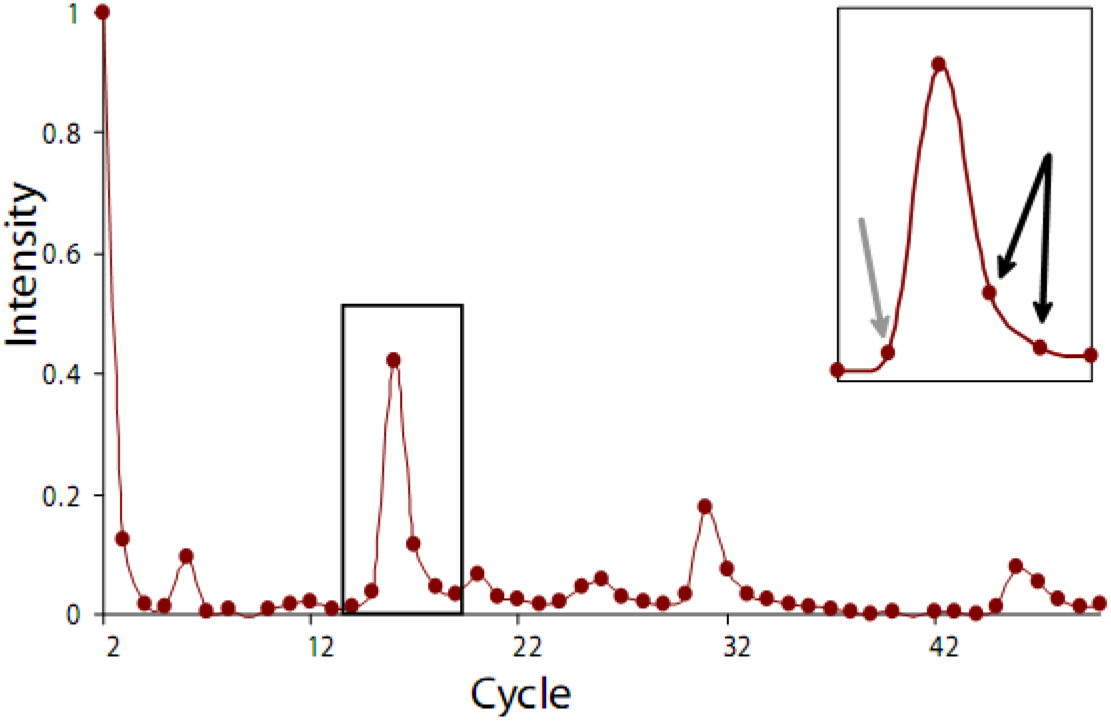
- Synthesized DNA fragments
  - Delta function
  - Dinucleotide microsatellites
  - Theta function



	1.. ..50
Δ on randm context	CAGTCGGCCGTCGGTATCCTGGTGGTGGCTAGGCTGTCTCTTTCCACGGC GCAGTAGTGTTGGTTCTGTAGTGGAATGTGCGGTTGTTGAGAATTCAGTA CGCCTTACAATTCAAAGTCCATATAACTTTGAATAACCTTACATCGATAT CTAGCCGCGACAACATAGCAGGCACGAGAGTCGACGGACAGCGGATGCGA
Δ on homeo-polymer con-text	ACAAAAAAAAAAAACAAAAAAAAAAAAAACAAAAAAAAAAAAAACAAAAA AGAAAAAAAAAAAAGAAAAAAAAAAAAAAGAAAAAAAAAAAAAAGAAAAA TCTTTTTTTTTTTTCTTTTTTTTTTTTTTCTTTTTTTTTTTTTTCTTTTT TGTTTTTTTTTTTTGTTTTTTTTTTTTTTGTTTTTTTTTTTTTTGTTTTT
MS	ACACACACACACACACACACACACACACACACACATTG GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCCA
Θ	AAAAAAACCCCCCCAAAAAAACCCCCCCAAAAAAACCC GGGGGGGTTTTTTTGGGGGGGTTTTTTTGGGGGGGTTT

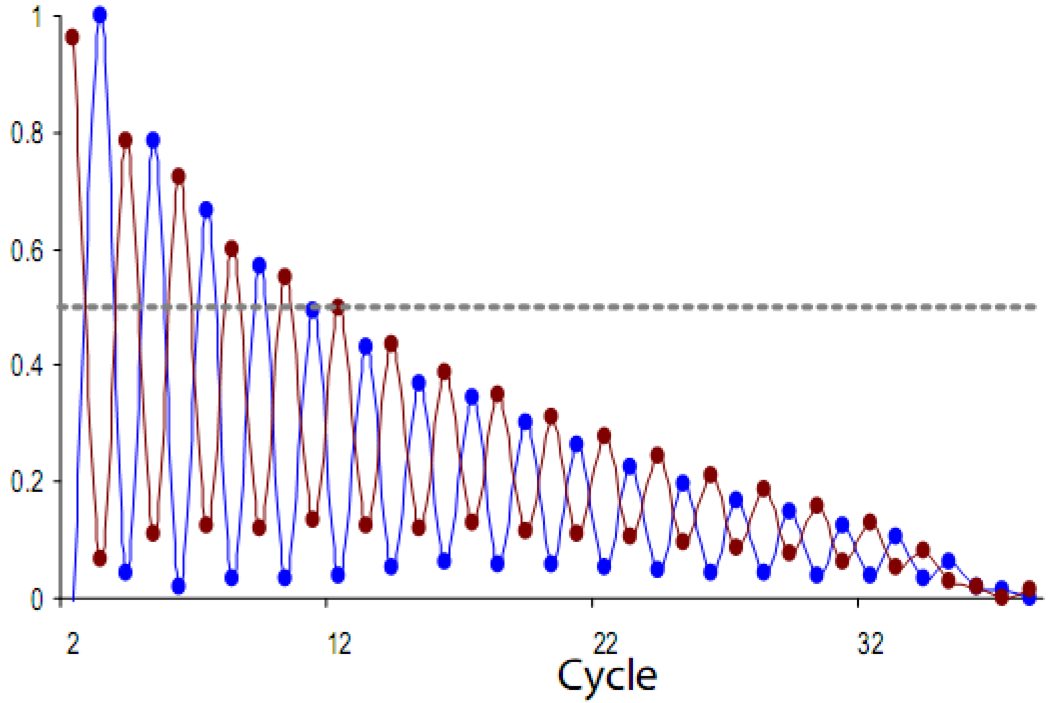
Noise factors - phasing

- Impulse response test of delta function:
GCAGTAGTGTTGGTTCTGTAGTGGAATGTGCGGTT
GTTGAGAATTCAGTA



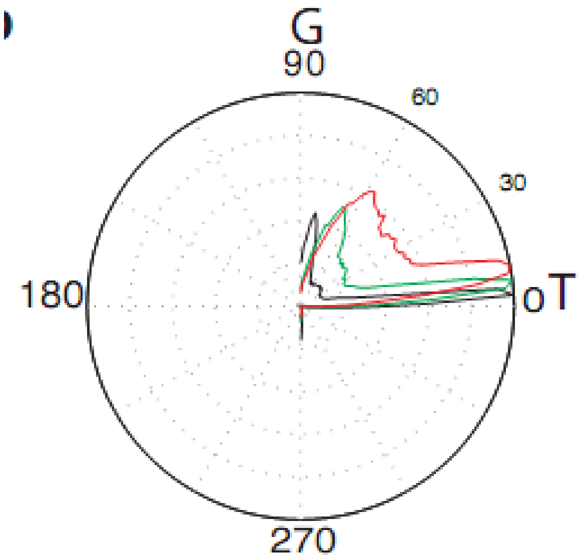
Noise factors - fading

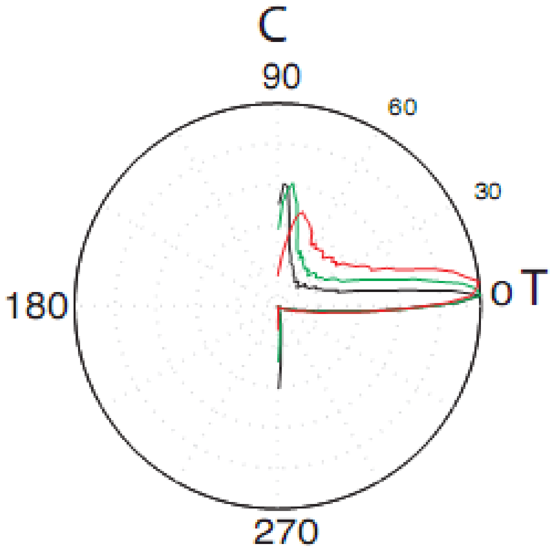
- Output average intensities of microsatellite sequence ACAC…



Noise factors – crosstalk change

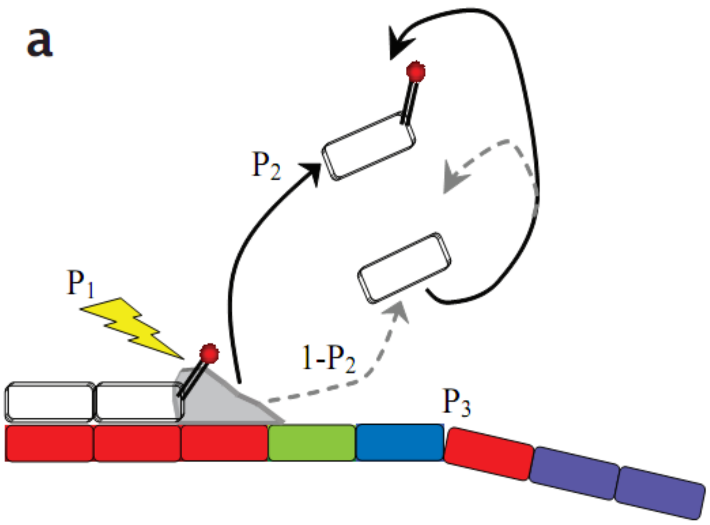
- Polar histograms present the ratio between channel intensities correlated with the base preference (bacteriophage phi-X library)

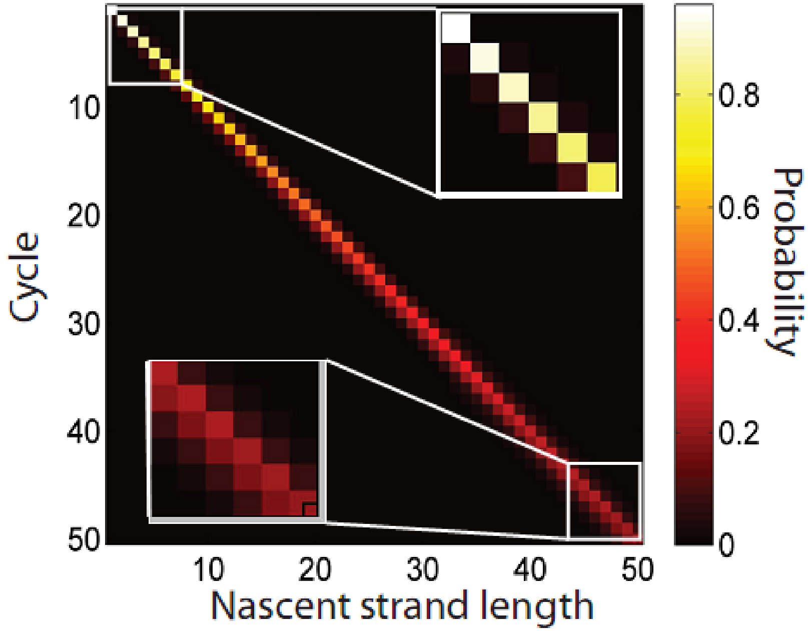




Random walk model of phasing & fading

- P1 – block removal
  - stay the same length with 1-P1
- P2 – incorporation of blocked nucleotide
  - incorporation of non-blocked nucleotide with 1-P2
- P3 – strand loss





Decomposition of phasing & fading

$$DP = R$$

- R(t, n) – probability of a nascent strand to be n nucleotides long after t cycles
- D – fading matrix (t by t diagonal)
- P – phasing matrix (t by n)
  - P(t, n) – probability of finding a nascent strand with length n after t cycles

Intensity of DNA cluster

$$\eta_j \cdot DPS_jG^T = I_j$$

- η_j – size of j-th DNA cluster (scalar)
- S_j – DNA sequence of j-th cluster (n by 4)
- G – crosstalk matrix (4 by 4)
- I_j – intensity signal of j-th cluster (t by 4)

$$\eta_j \cdot (PD)^+DPS_jG^TG^{-T} = (PD)^+I_jG^{-T}$$

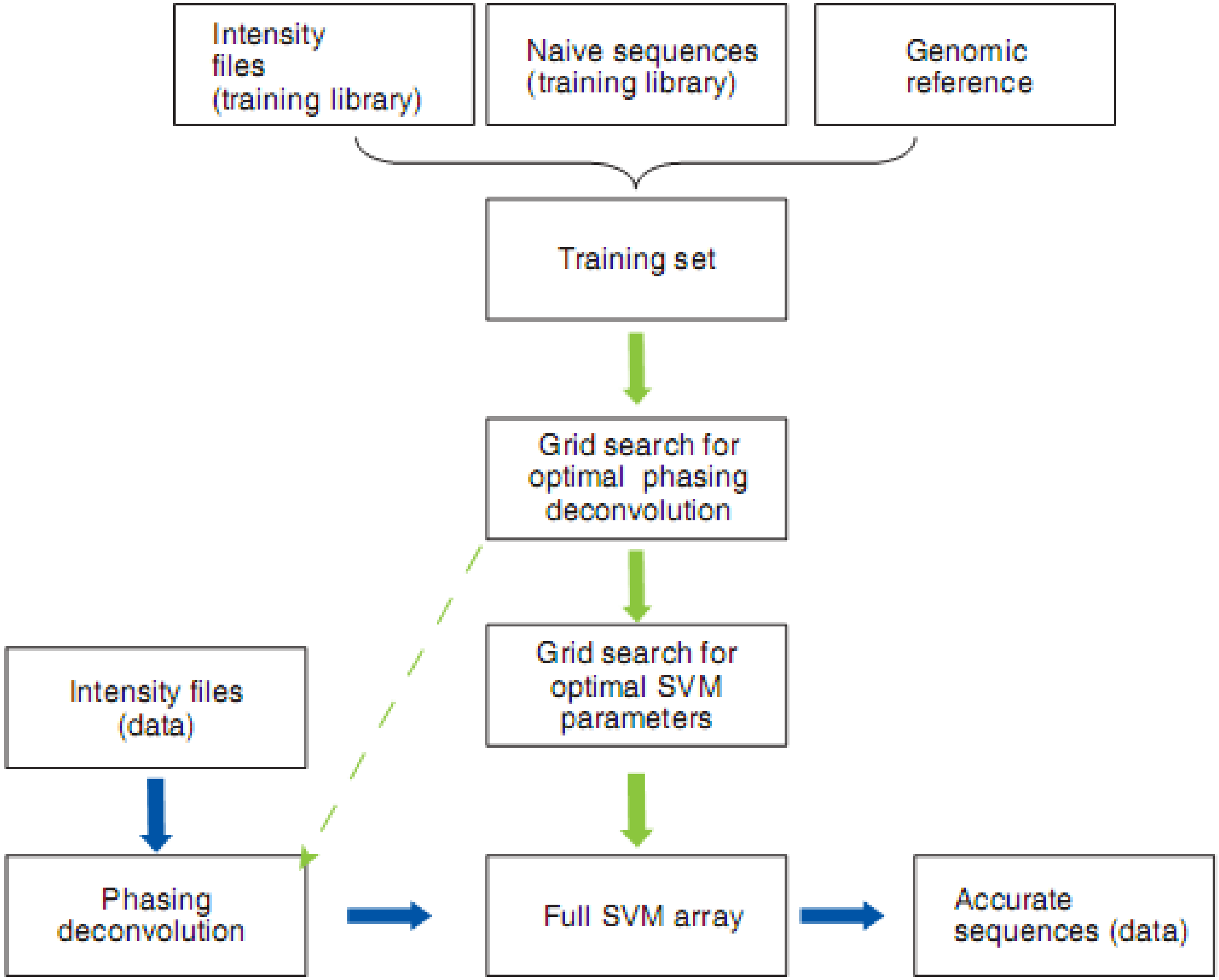
$$\eta_j\Sigma S_j = Y$$

Note the crosstalk matrix G is cycle-dependent!!!

Alta-Cyclic

- Treat sequencing as a classification problem, use SVM to learn noise patterns
- Training set: fluorescence intensities and corresponding correct base calls
- Training process
  - Deconvololute the phasing effect of intensities according to grid coordinate
  - Pick the intensities and correct base calls of last few cycles, run SVM for each cycle
  - Average success rate in the cross-validation of SVMs is used as a feedback to the grid search
  - Optimize SVM parameters by using grid search

Steps of training and base calling



Intensity files (training library)

Naive sequences (training library)

Genomic reference

Training set

Grid search for optimal phasing deconvolution

Grid search for optimal SVM parameters

Intensity files (data)

Phasing deconvolution

Full SVM array

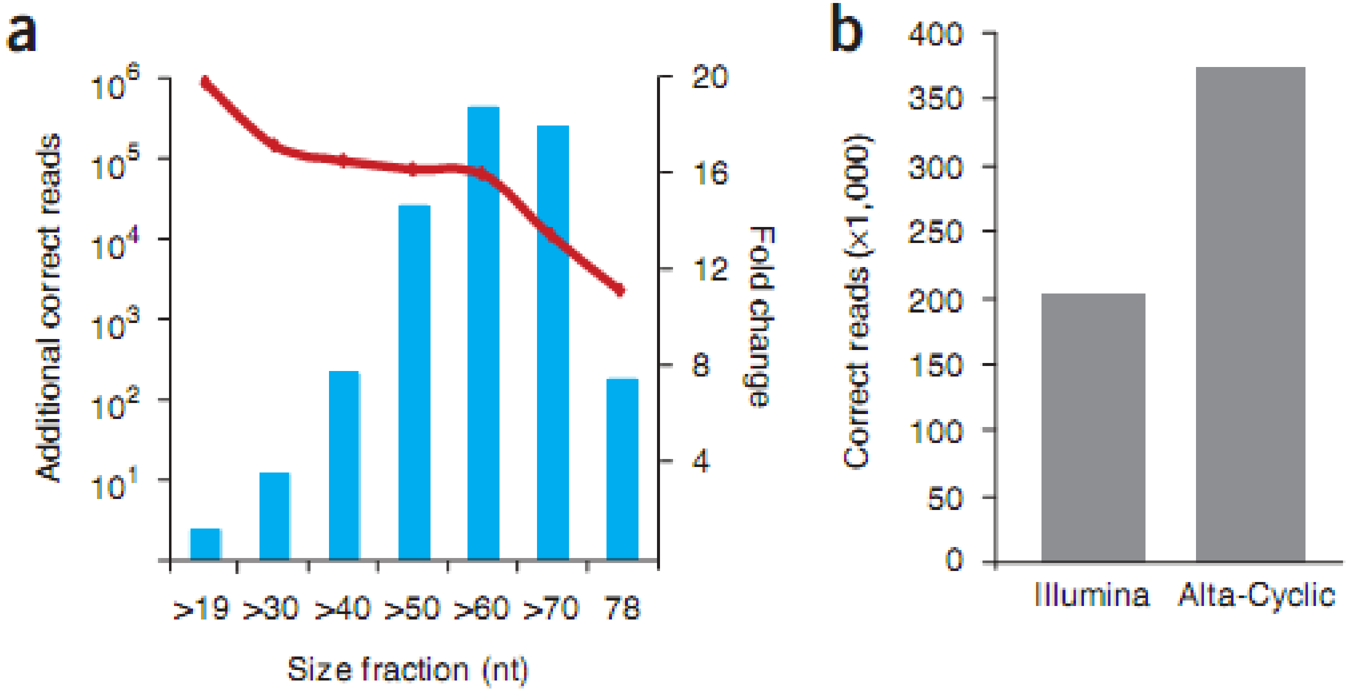
Accurate sequences (data)

Features of Alta-Cyclic

- All the calling parameters are optimized empirically and tested to enhance the accuracy for each run
- Phasing parameters are based on a parametric model and calculated from data of latest cycles
- Dynamically track changes in fluorophore cross-talk

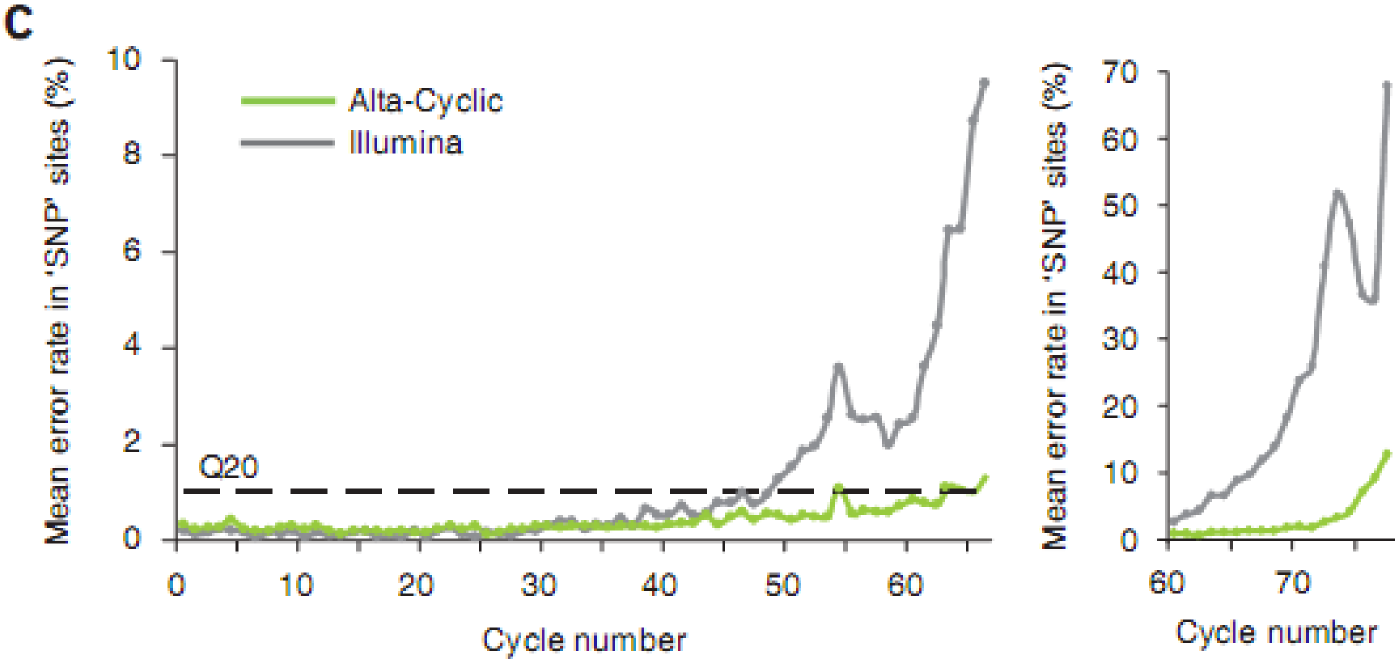
Experiments - for long runs

- A: HepG2 RNA library
- B: Tetrahymena micronuclear library



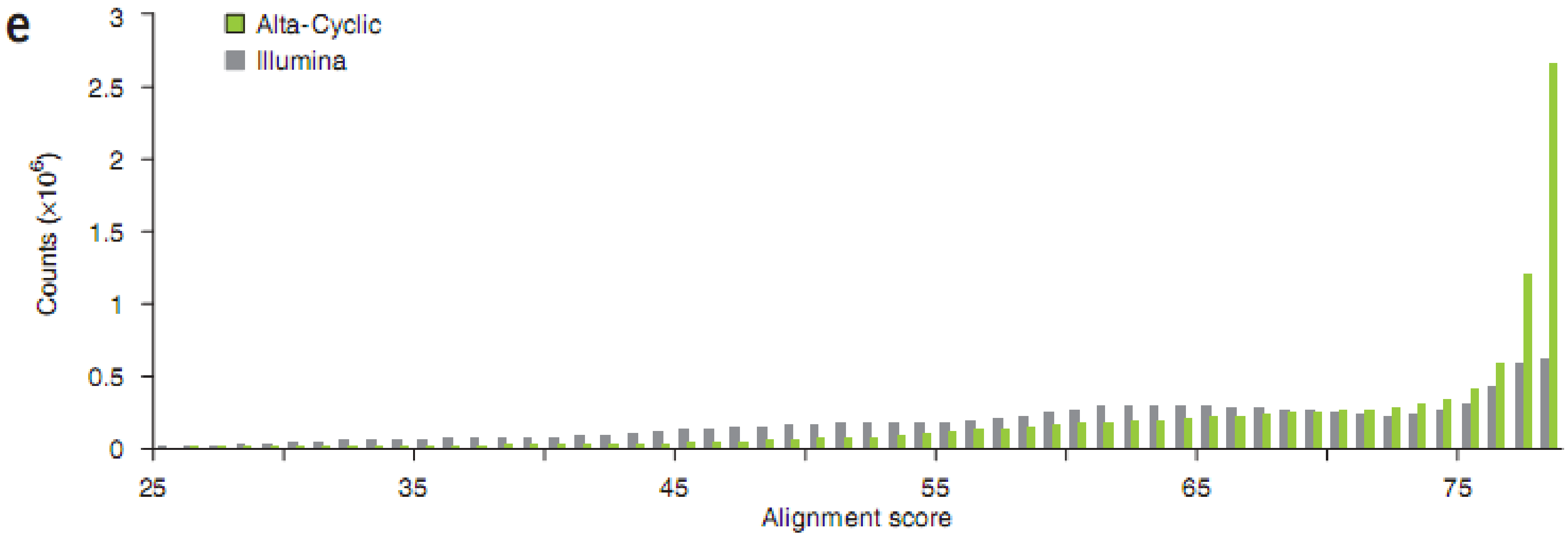
Experiments – identify sequence variants

- Phi X library with 1% artificially single-base changes



Experiments – very noisy reads

- Align output to phi x genome (allowing 53 mismatches out of 78)



Weak points (from my point of view)

- Need more computation time; iterative grid search of parameters is time consuming
- Referencing DNA library must be prepared and extra DNAs must be sequenced for each run
- Training dataset could be noisy
- SVM parameters used in grid search for parameters of random walk model are not mentioned.

Thank you!