Base Calling: methods, problems and alternatives

Tim Massingham
tim.massingham@ebi.ac.uk

EMBL Advanced Course in Analysis of Short Read Sequencing Data
8th June 2009 -- 10th June 2009

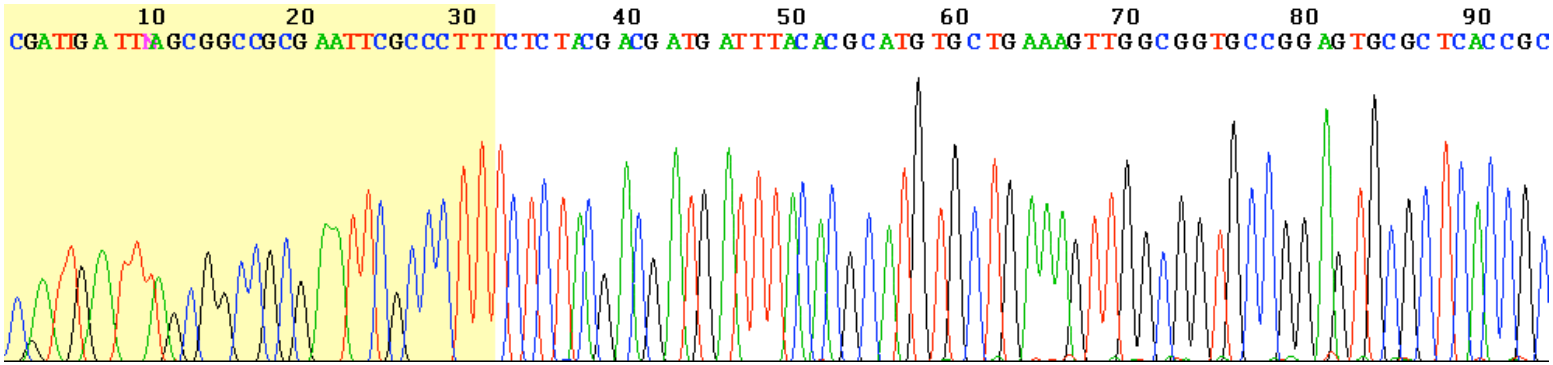
Outline

New sequencing technologies
Image analysis
Lasers and cross talk
Chemistry and phasing
AYB -- a new base calling algorithm
Quality calibration and sequencing errors

Capillary sequencing

AB 3730xl Cutting edge of capillary technology
96 capillaries in parallel

Rapid	2100 kb/day	550 base reads
Accurate	690 kb/day	900 base reads



Images: http://www.appliedbiosystems.com/
http://en.wikipedia.org/wiki/File:Sanger_sequencing_read_display.gif

454 Life Sciences (Roche)

454 GS FLX titanium

Current performance
1 million reads per run
400bp meanlength, at Q20

Future (454 lab demonstration)
650bp mean length (750 modal)

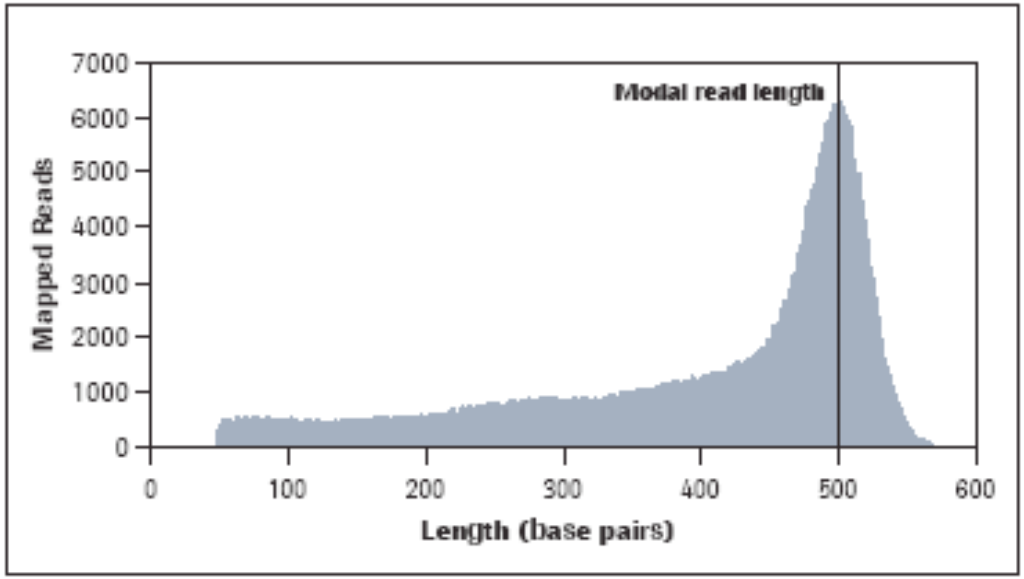


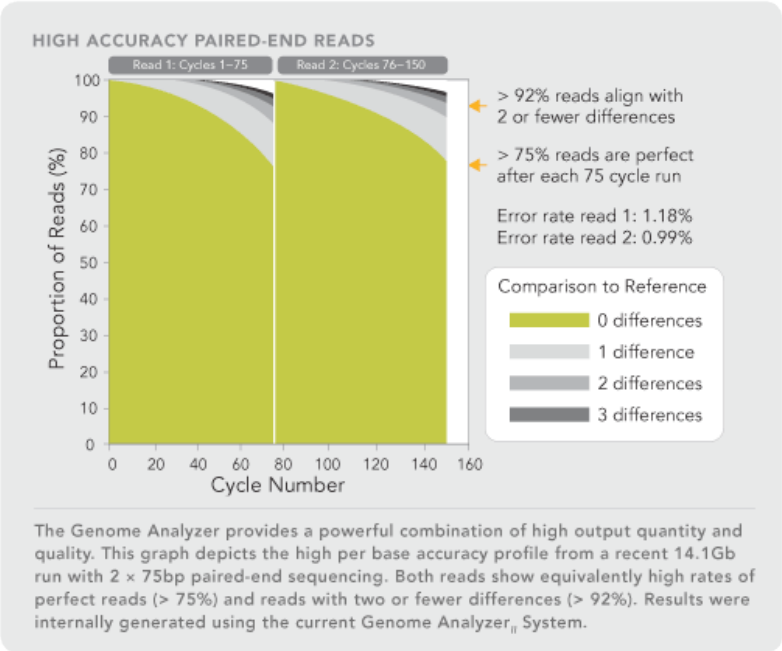
Figure 1: Example Read Length Distribution of 629,643 reads from *E. coli* K-12 (Genome size ~4.5 Mb) with a modal read length of 504 bases.

Images: http://www.454.com/

Illumina GA II

Current performance
14 Gb per run 2 x 75bp
20 Gb per run GA II_x software upgrade
150bp reads demonstrated outside Illumina

Future (2010) with new array technology
"*sub-micro semi-ordered array*"
55 Gb -- 100 Gb 2 x 125bp,

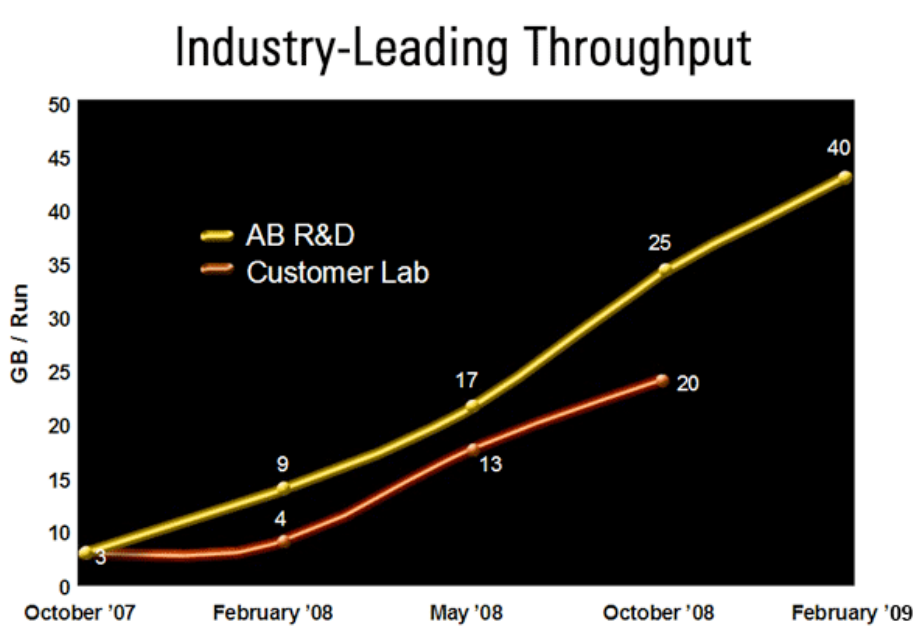


The Genome Analyzer provides a powerful combination of high output quantity and quality. This graph depicts the high per base accuracy profile from a recent 14.1Gb run with 2 × 75bp paired-end sequencing. Both reads show equivalently high rates of perfect reads (> 75%) and reads with two or fewer differences (> 92%). Results were internally generated using the current Genome Analyzer$_{II}$ System.

Images from http://www.illumina.com/

Life Technologies' SOLiD

20-30 Gb 2 x 50bp (~ 2 weeks)

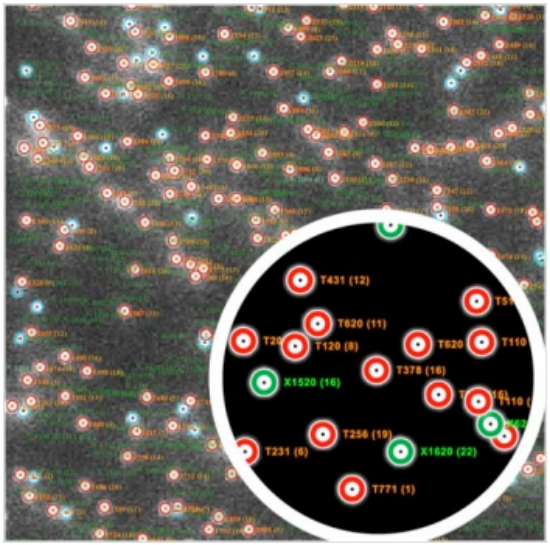


Helicos

Single molecule sequencing -- no amplification

21-28 Gb per run, 105-140 Mb per hour
Read length 20 to 55bp, 30-35bp average

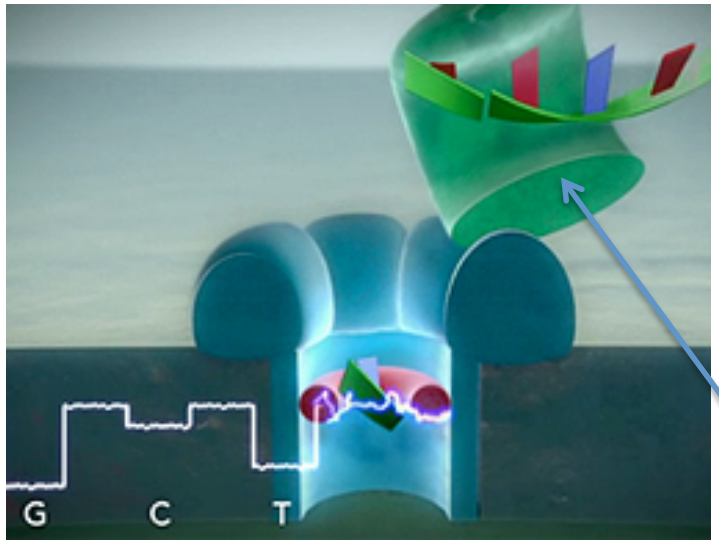
Asynchronous: separate steps for A, C, G, T
- strands get sequenced at different rates
- base composition bias in length of read



Images: http://www.helicosbio.com/

Coming technology

Oxford Nanopore



Nanopore sequencing

No fluorophores
• use electrical properties of base passing through pore
Can detect methyl-cytosine

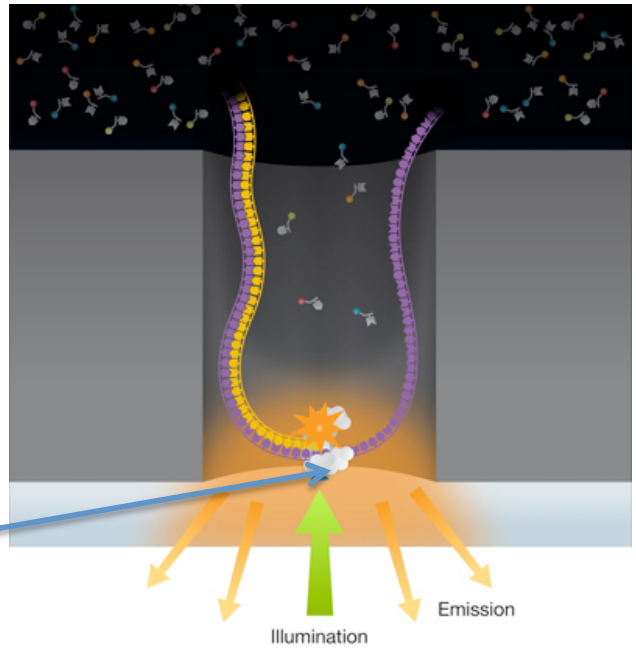
Aims

50bp/sec per read
Kb length reads

100 Gb per hour

modified polymerase

Pacific Biosciences

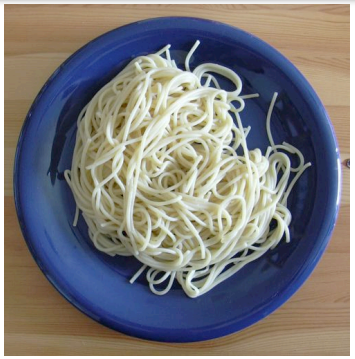


Zero-mode wave guide

Tiny illuminated volume
• only bound fluorophores contribute
• watch incorporation in real-time, including errors

Images: http://www.nanoporetech.com/
http://www.pacificbiosciences.com/

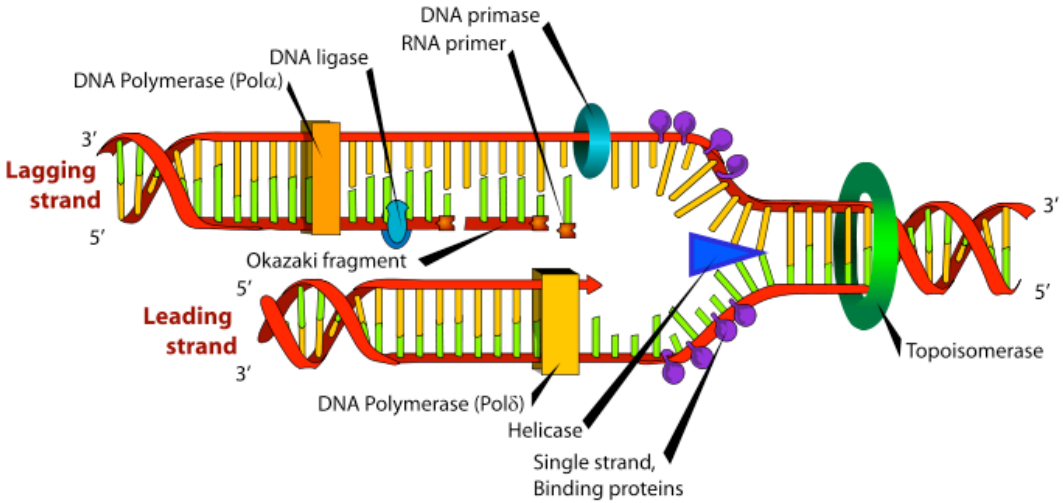
Limits on read length



DNA spaghetti
- knots
- snaps if tugged
- sticks to walls when cooked

Use chromatin to organise?

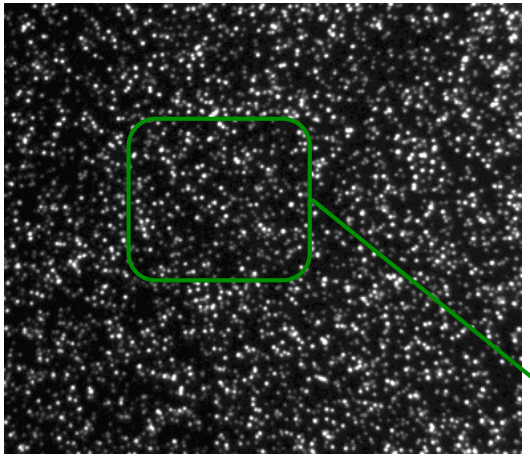
DNA replication



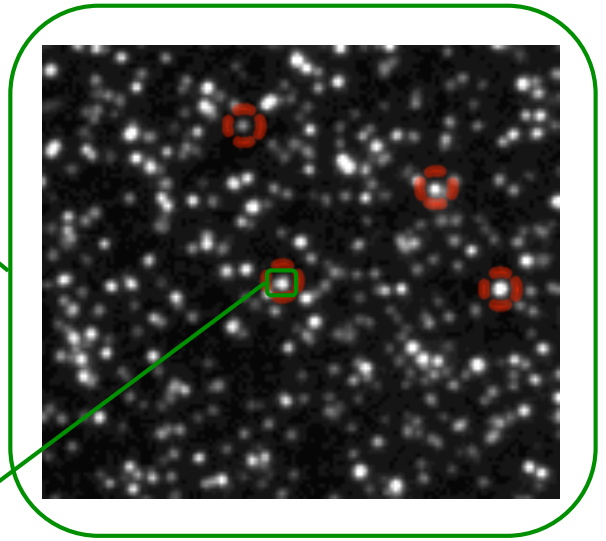
Images http://en.wikipedia.org/wiki/File:Spaghetti.jpg
http://upload.wikimedia.org/wikipedia/commons/6/6a/30nm_Chromatin_Structures.png
http://en.wikipedia.org/wiki/File:DNA_replication.svg

Analysis pipeline

Raw images – – –

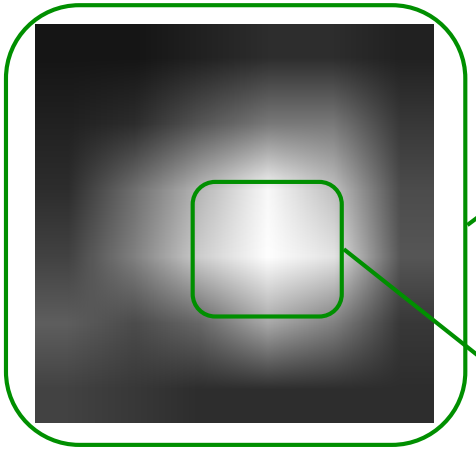


One image per channel, so four images per cycle



Find clusters – – – – – – – – – – – – – – – – – –

Read intensities – – –



```
#CH4:OBJ13808
3210.3 3617.2   242.0   519.3
  62.1   128.3 -101.5 2556.5
 125.8   145.9  -34.9 2685.2
1968.5 1053.9     4.5   59.6
1732.1  995.9     5.0   48.1
```

Estimate cross-talk and phasing, then call bases

Image analysis

Registration

Filtering

- Sharpen clusters
- Edge detection

Normalization

- subtract background
- noise

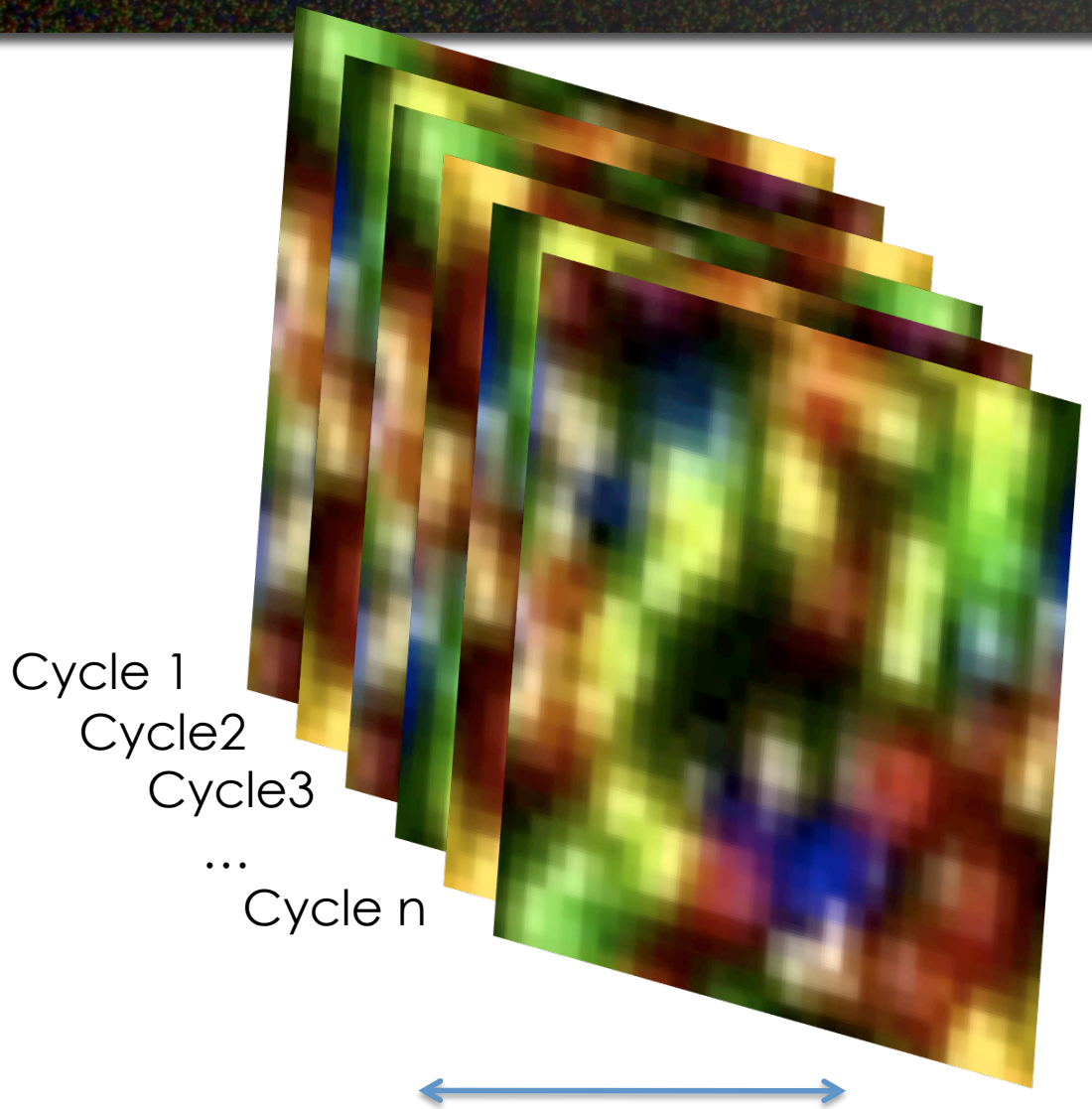
Cluster identification

- deblend (split large clusters)
- remove local background

Warning: out of date, describes older version of pipeline

Based on notes prepared by Nava Whiteford,
http://sgenomics.org/mediawiki/upload/8/80/Pipeline.pdf

Image Registration



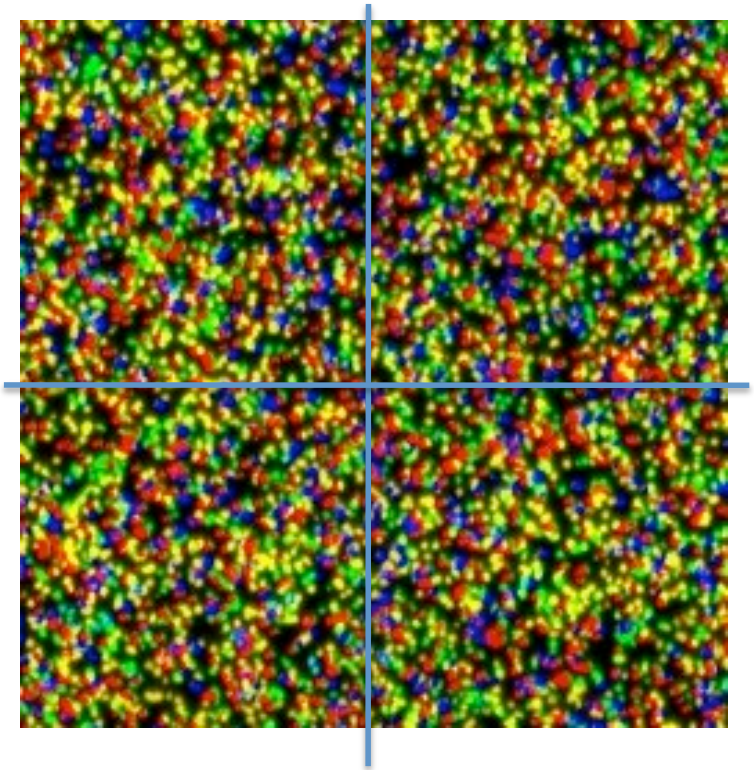
Clusters "move" between cycles
- instrument jitter
- focal changes

Must track clusters between cycles and align images

Image Registration

Three effects to compensate for
- Translation
- Rotation
- Scaling

} Restricted form of affine transformation (procrustes transformation)



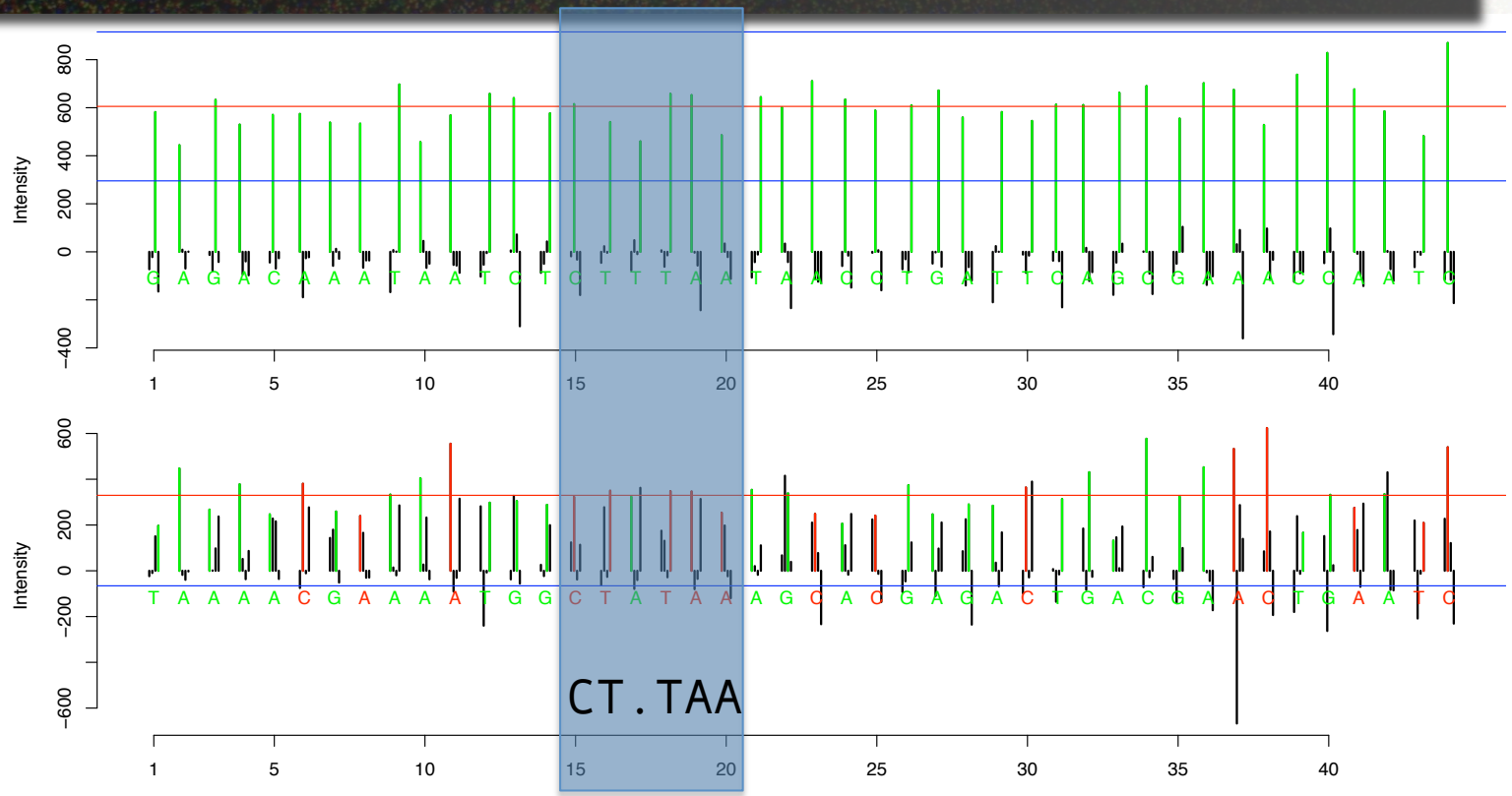
Basically a dynamic programming problem

Split image into regions
Estimate transformation for each region
Take consensus

Image Registration

Read 1 (214,714)

Read 2 (214,715)



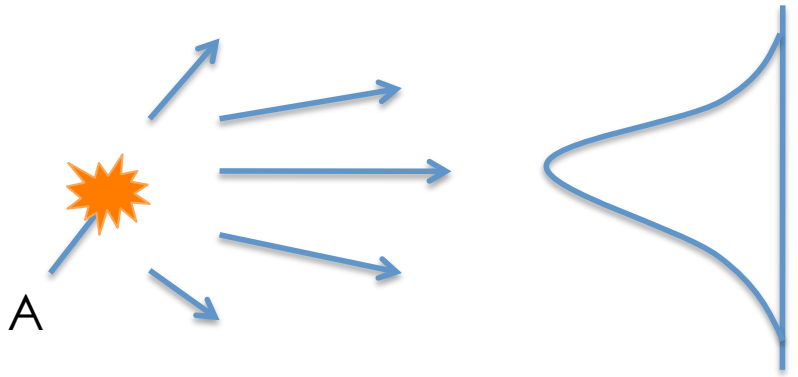
Do errors in read match their neighbour? p-value ~ 1e-4 (0.83)

	Match	Mismatch
Observed	12 (8)	4 (20)
Expected	4 (7)	12 (21)

(Corresponding numbers for positions correct in read 2 are in brackets)

Filtering

Clusters become blurred
- emitted light not entirely coherent
- focal problems



Flurophore

Incident intensity

Need to correct for blurring to find position of cluster and emitted light
- sharpening
- edge detection

Original cluster

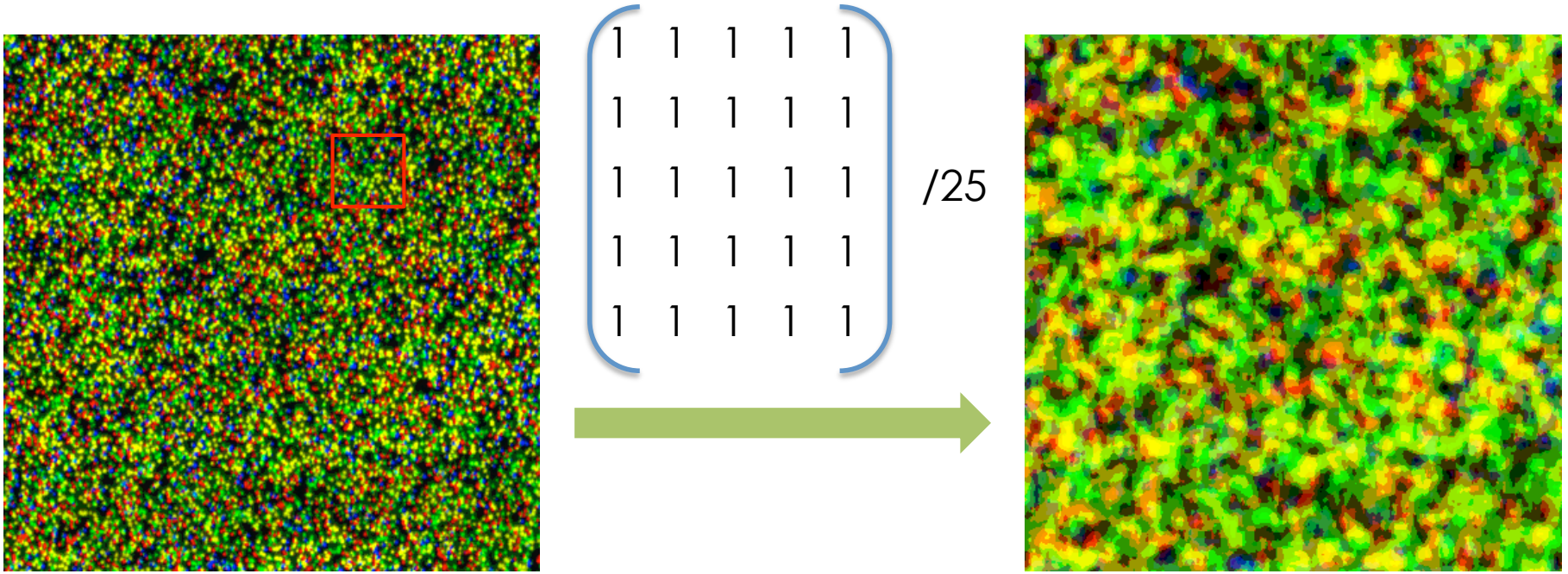
Gaussian blur

Convolution filters

Suppose we want to smooth an image

Replace each pixel by the mean of the surrounding pixels

Represent by matrix giving weights
for each pixel in neighbourhood

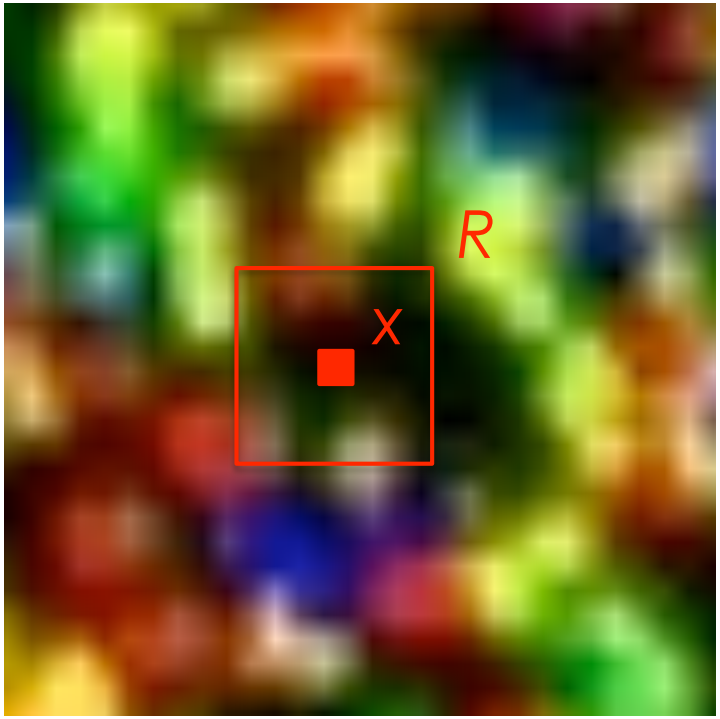


This is an example of a convolution filter
Create new filters by changing the values in the matrix

Convolution filters

Take a region around pixel
Multiply every pixel in region by corresponding value in filter *F*
Sum

$$x_{new} = 1^T (R \circ F) 1$$



E.g. $F = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$, $R = \begin{pmatrix} 0.33 & 1.73 & 2.56 \\ 1.18 & 4.70 & 7.36 \\ 2.17 & 6.76 & 10.1 \end{pmatrix}$

$$x_{new} = \text{sum} \begin{pmatrix} -1 \times 0.33 & -1 \times 1.73 & -1 \times 2.56 \\ -1 \times 1.18 & 8 \times 4.70 & -1 \times 7.36 \\ -1 \times 2.17 & -1 \times 6.76 & -1 \times 10.1 \end{pmatrix}$$

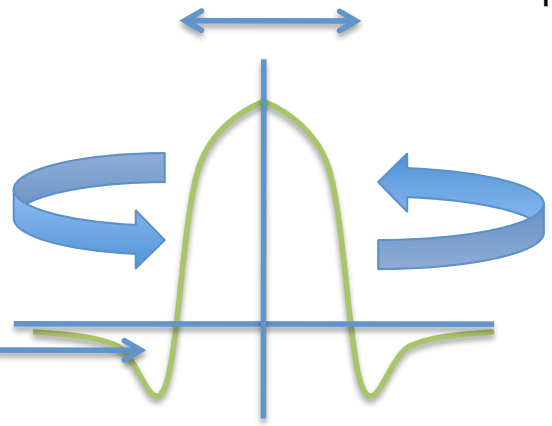
$$= 5.41$$

Normally do calculations in Fourier space - more efficient

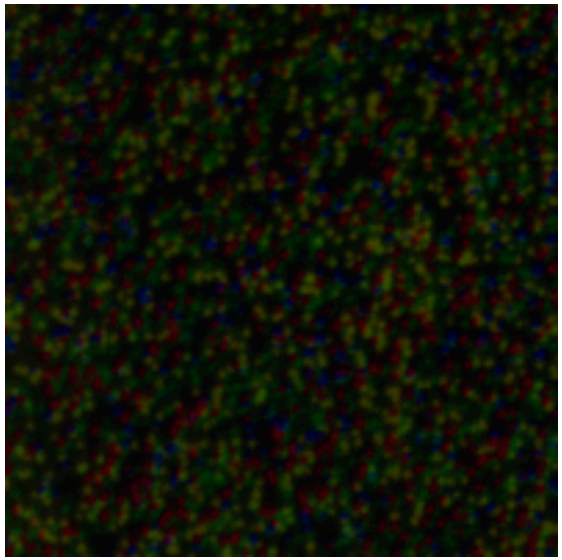
Filtering

Mexican hat = smoothing + edge detection

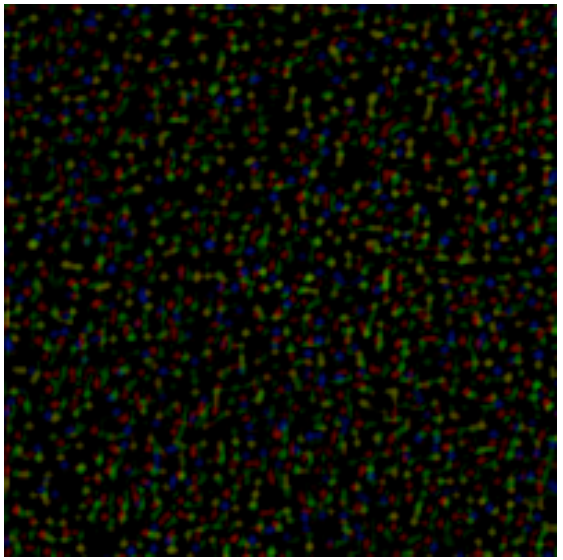
Smoothes around central pixel



Emphasises edges

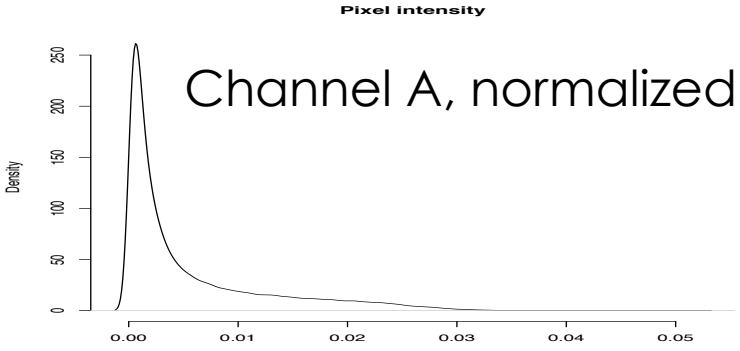


Apply filter
to each channel

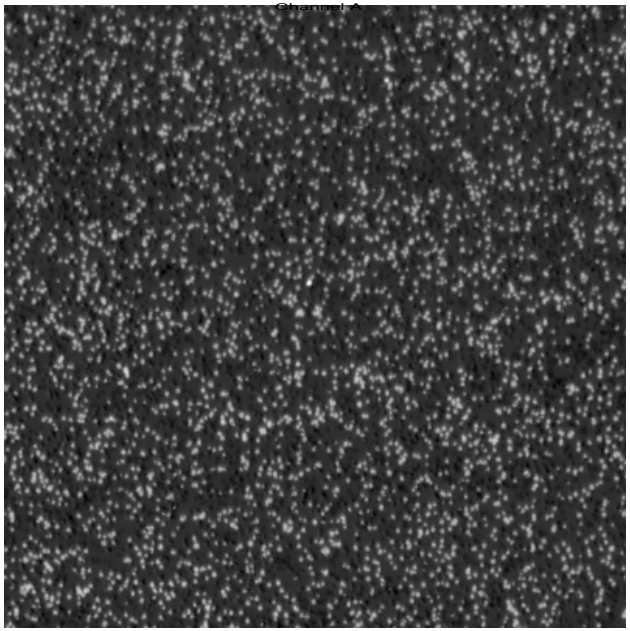


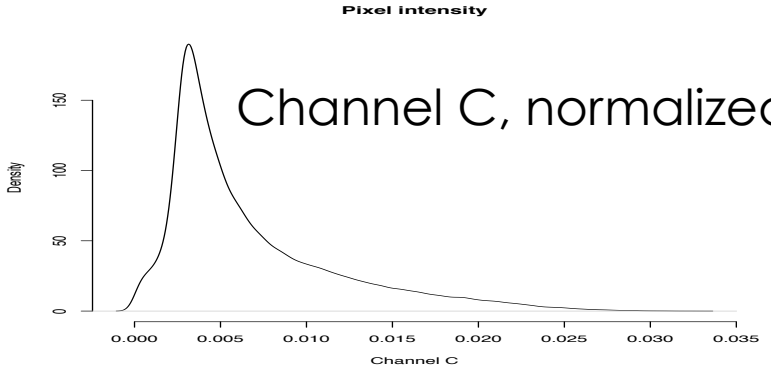
Normalization - background and noise

Background fluorescence: flow cell, unincorporated dyes, etc

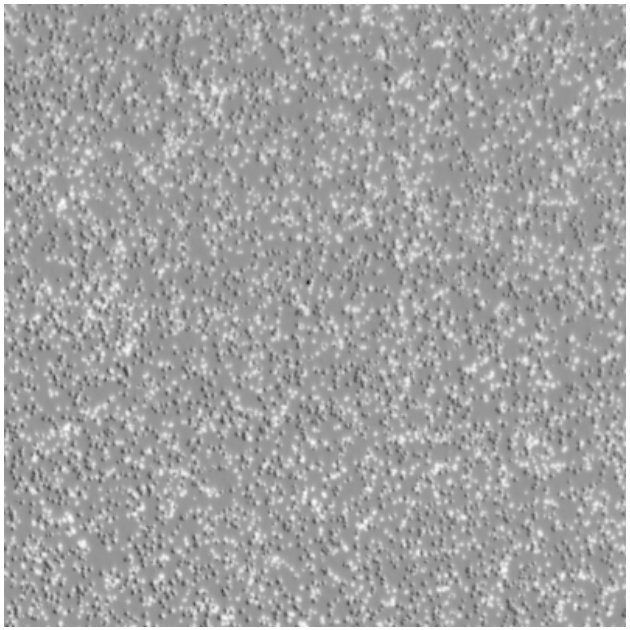


Channel A, normalized



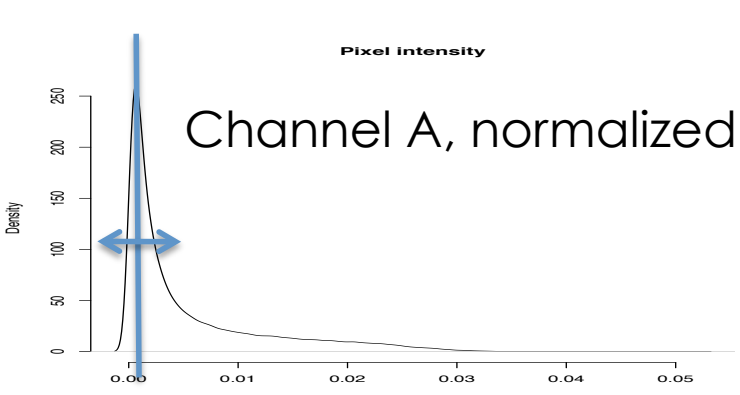


Channel C, normalized

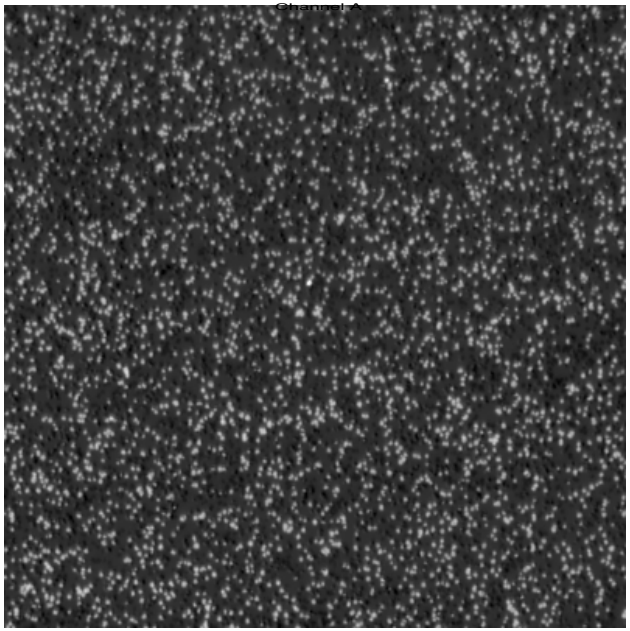


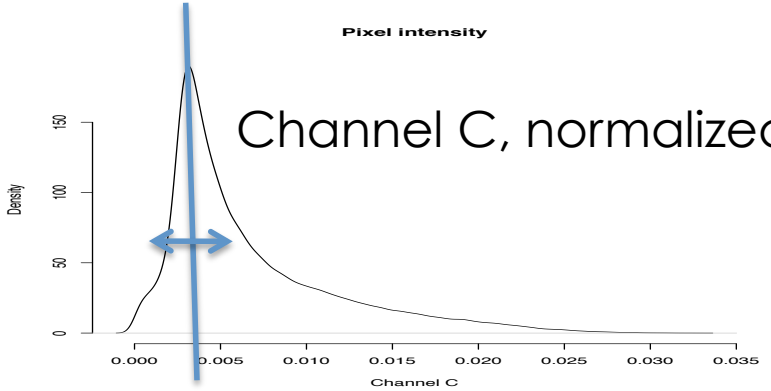
Normalization - background and noise

Robust estimates of mean and standard deviation

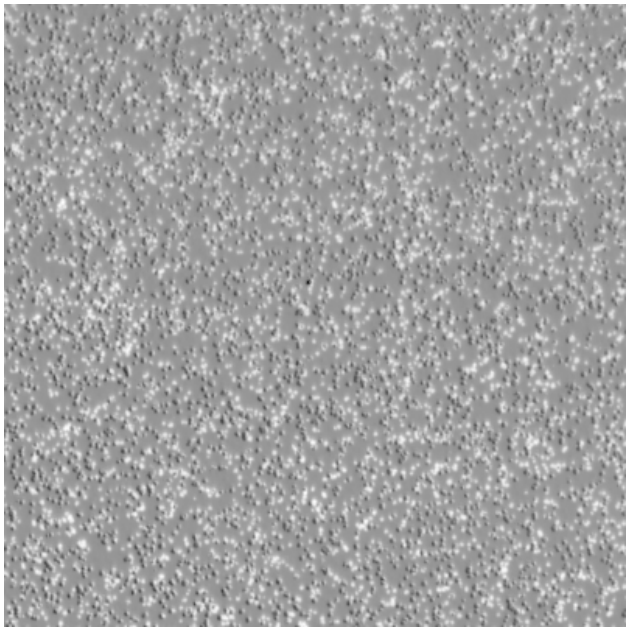


Channel A, normalized





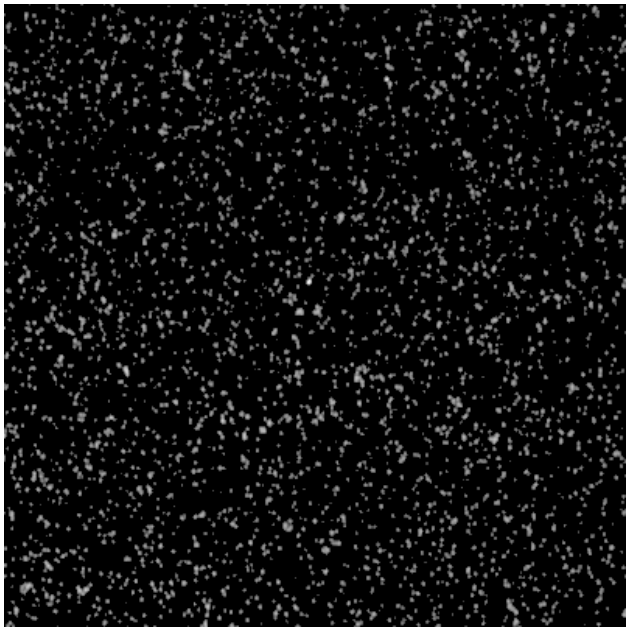
Channel C, normalized



Cluster identification

Warning: based on old version of pipeline;
this bit has probably changed more than any other

Thresholded tile



"Blank slide" model

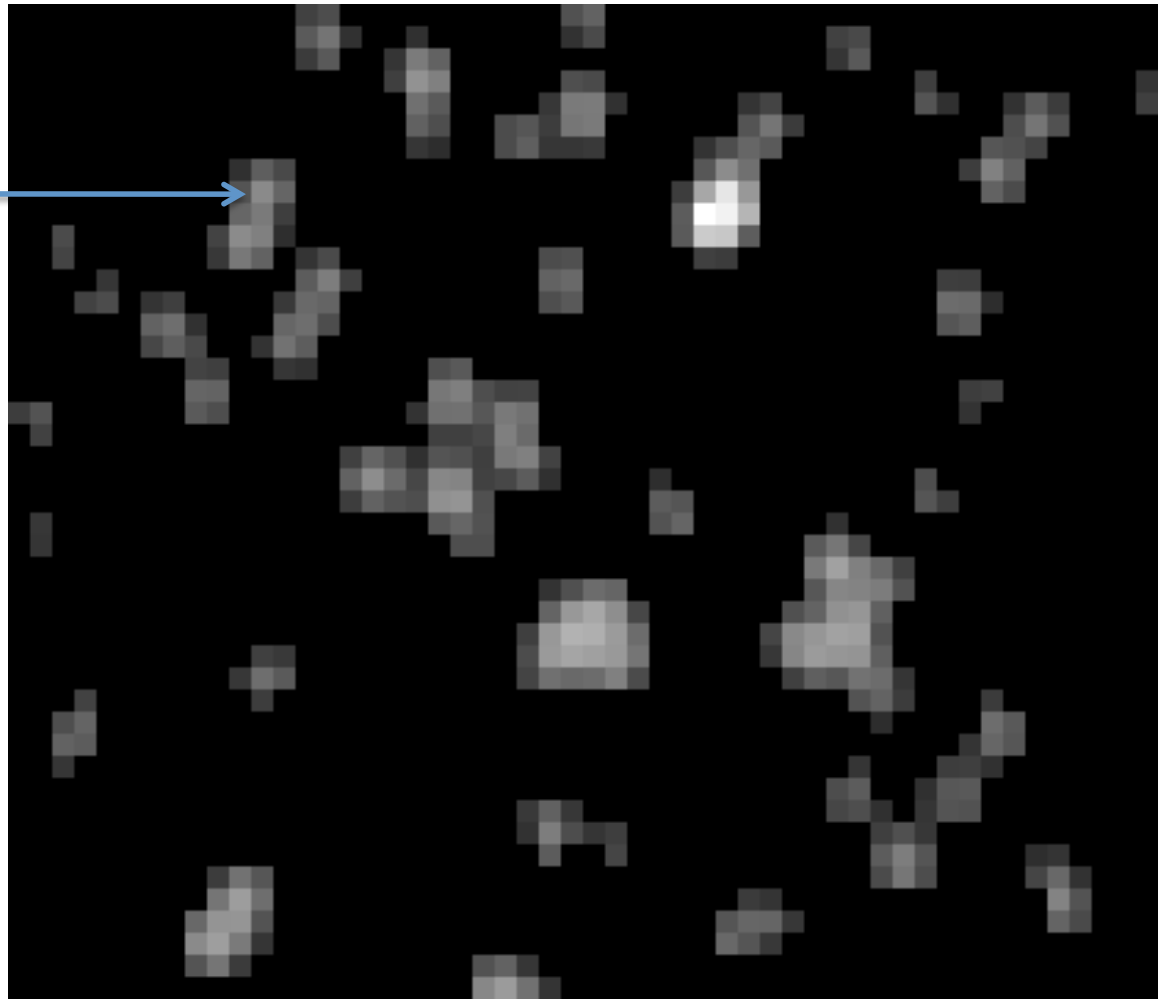
Background fluorescence	mean
Noise estimate	variance

Keep pixels 4 standard deviations above mean
4 sd ~ Q45 (30 errors per 1 million pixels)

Background and noise estimated in regions

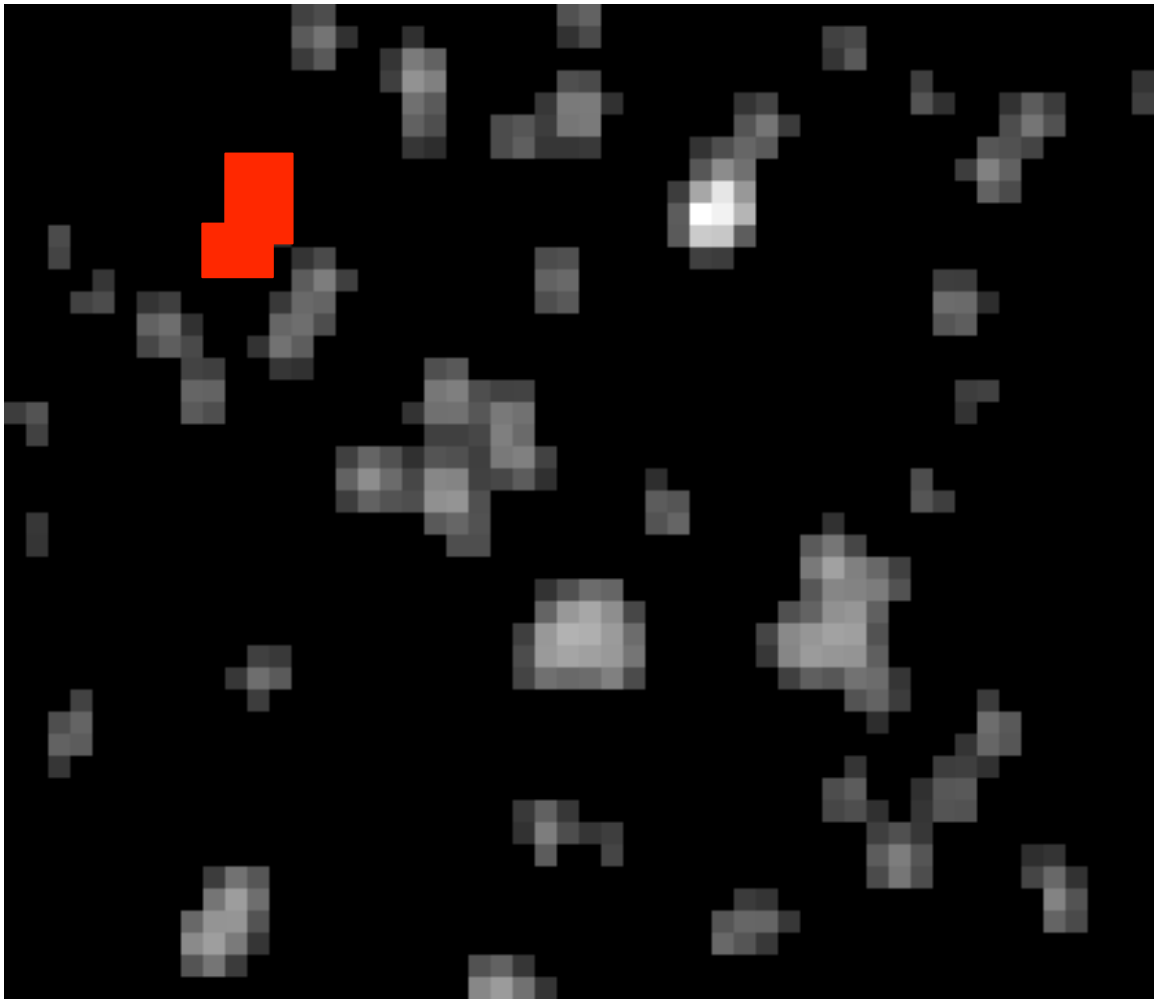
Cluster identification

- Find cluster



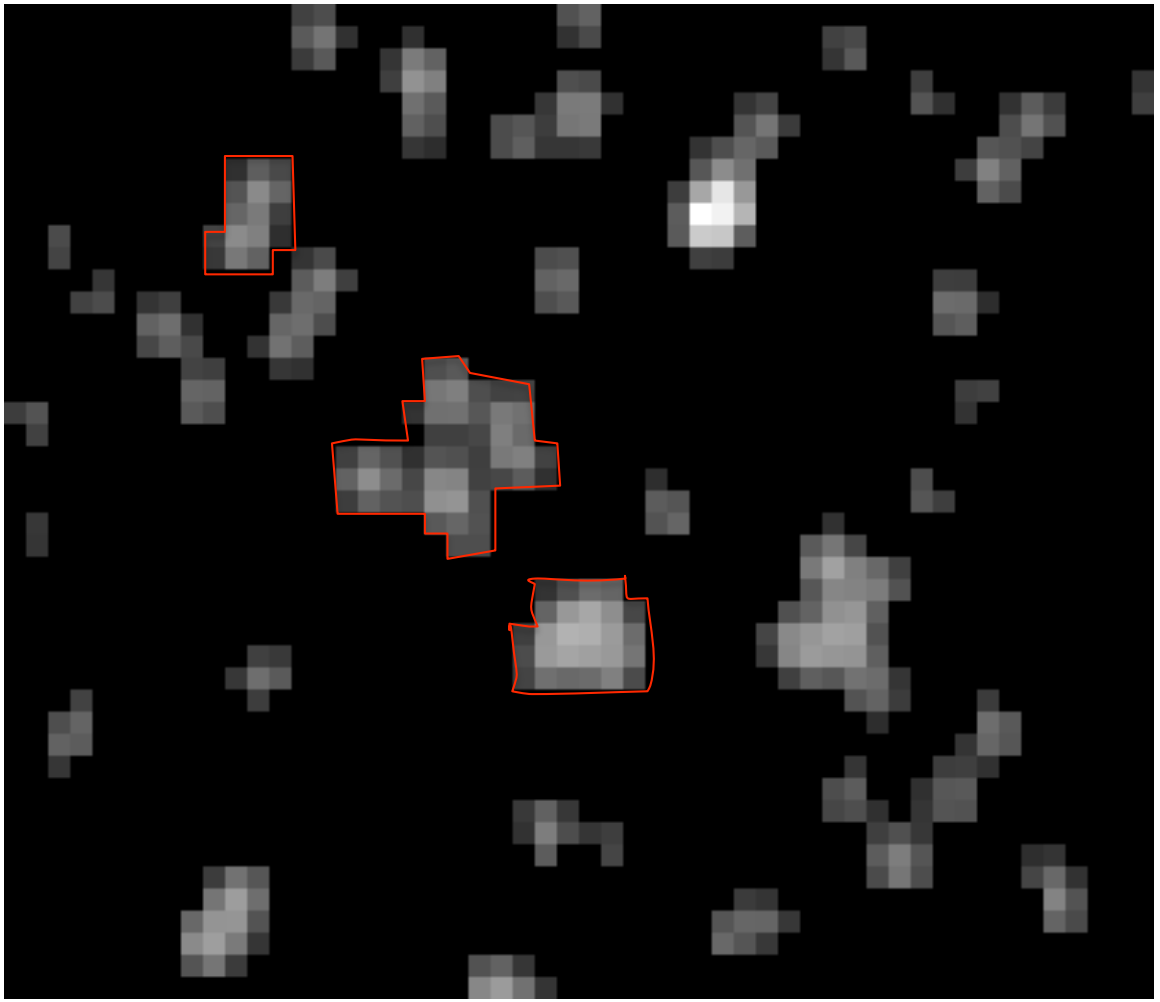
Cluster identification

- Find cluster
- Expand



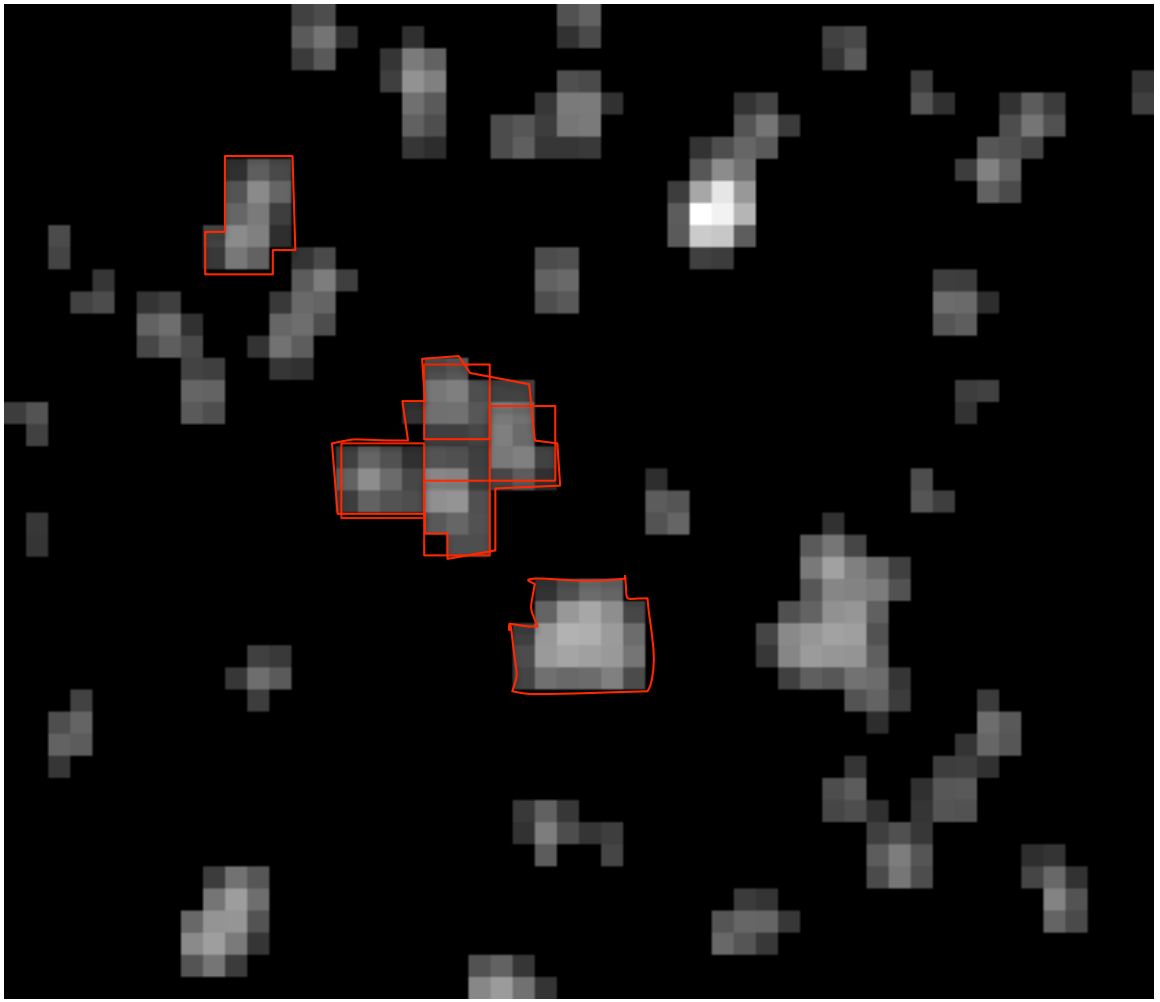
Cluster identification

- Find cluster
- Expand
- Find border



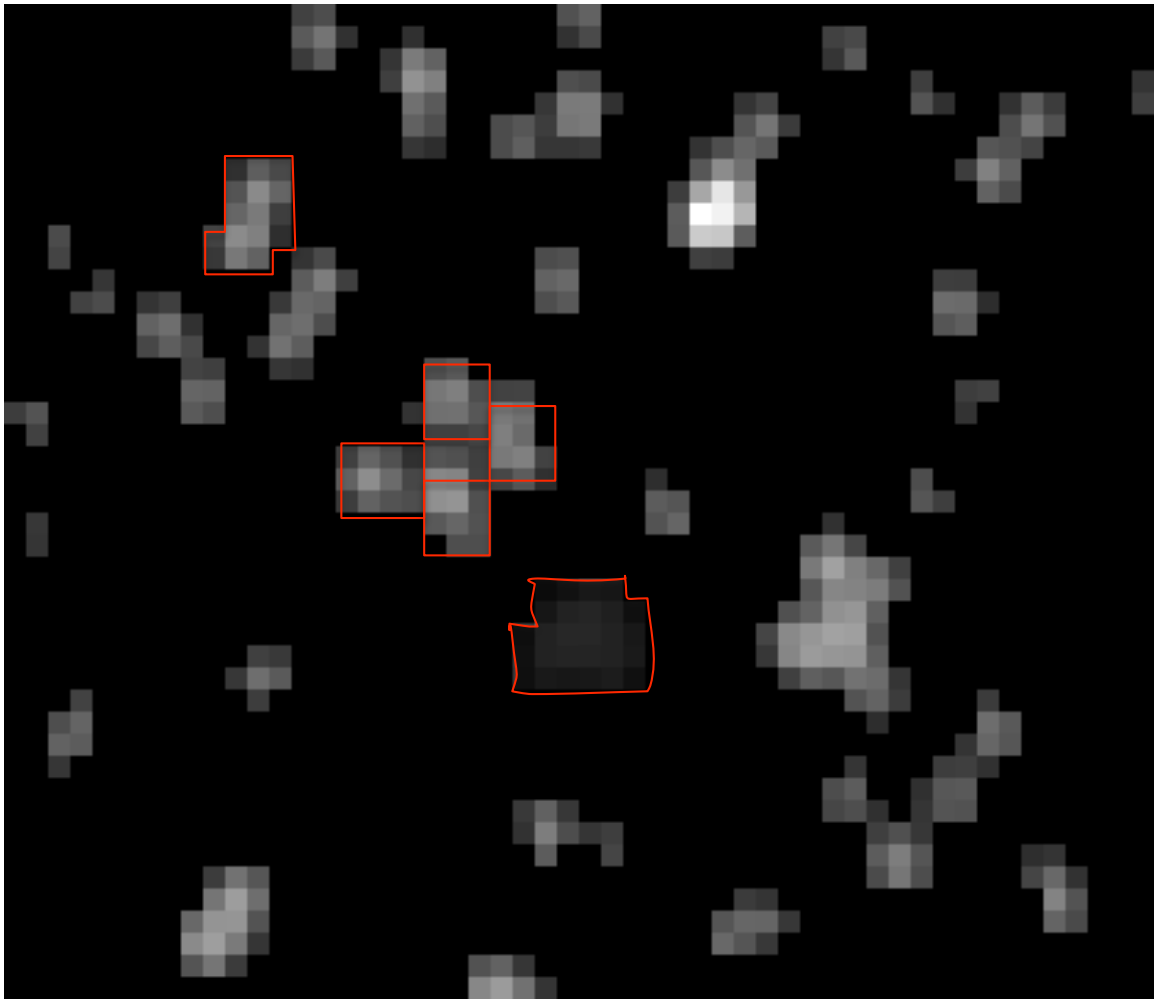
Cluster identification

- Find cluster
- Expand
- Find border
- Deblend (split) large clusters



Cluster identification

- Find cluster
- Expand
- Find border
- Deblend (split) large clusters
- Discard extremely large (probably contamination)



Local background

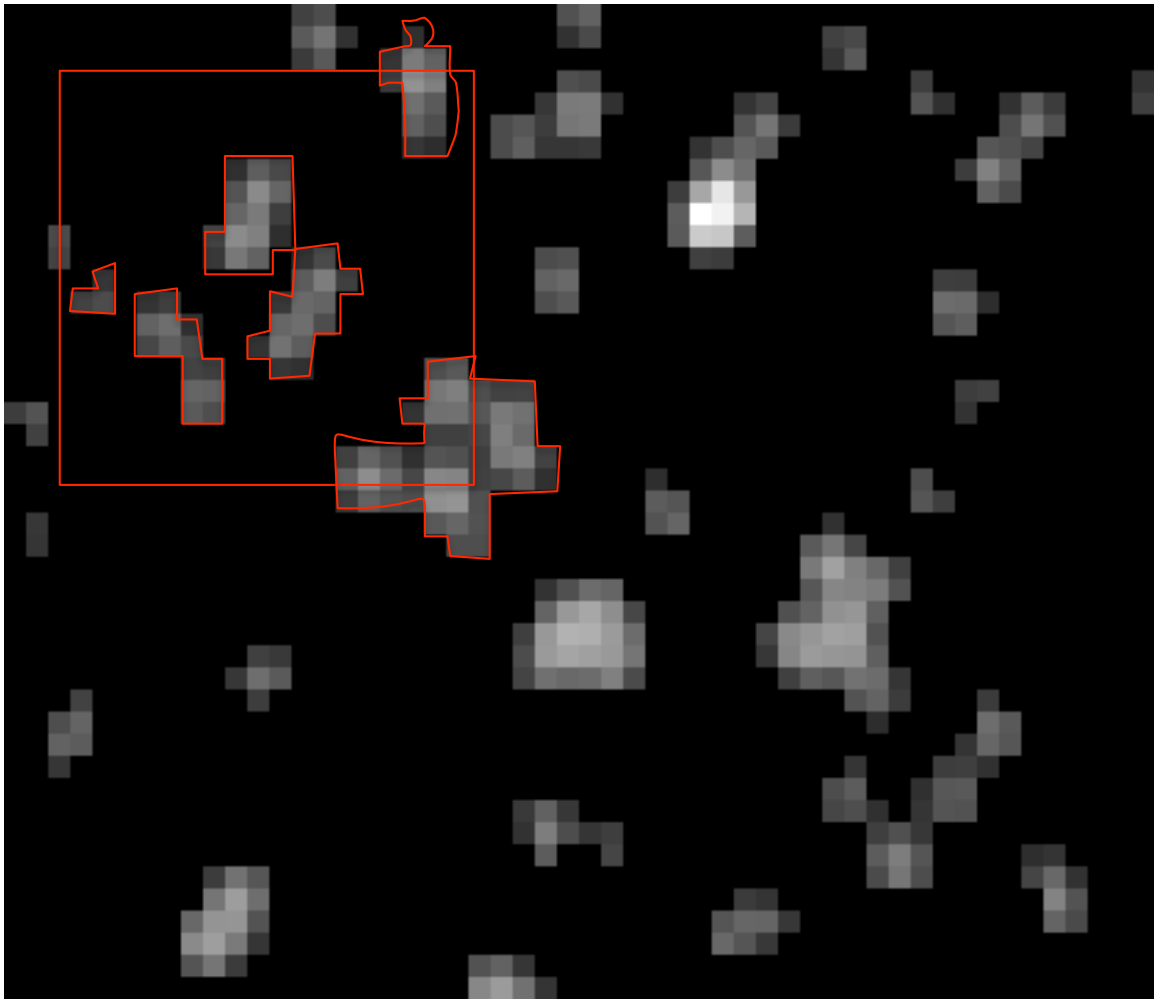
10x10 window around cluster

Take pixels not part of any cluster

Calculate new background noise

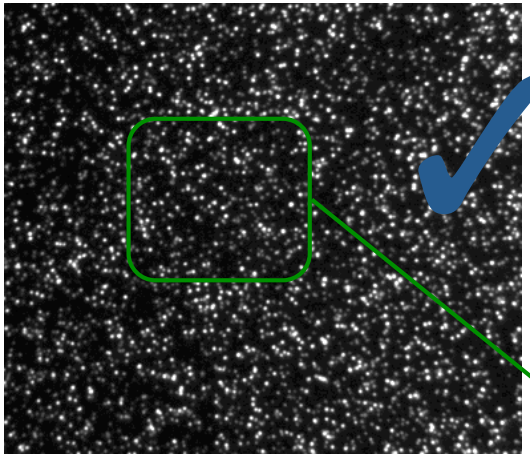
Correct cluster

Find brightest pixel for base caller



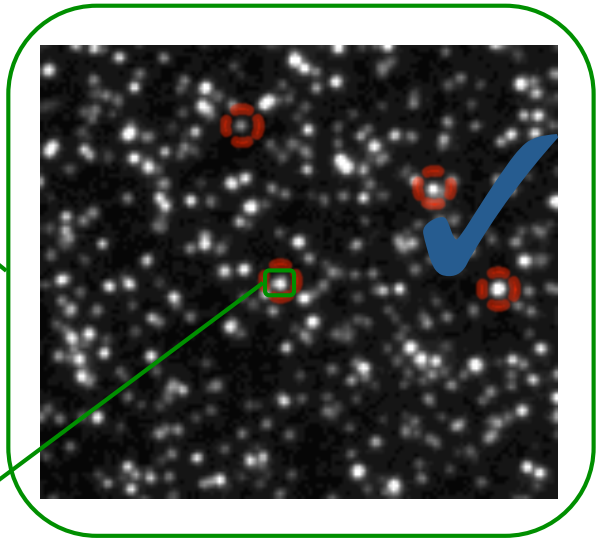
Analysis pipeline

Raw images

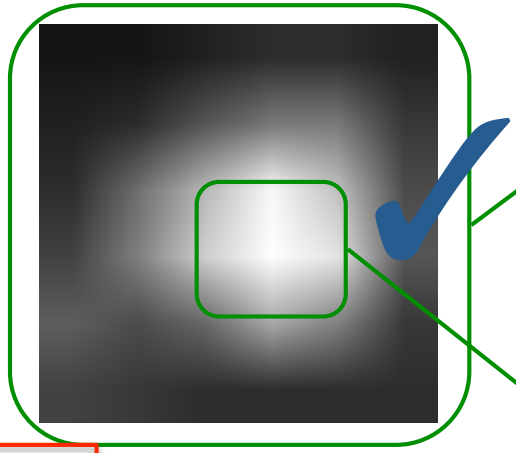


One image per channel, so four images per cycle

Find clusters



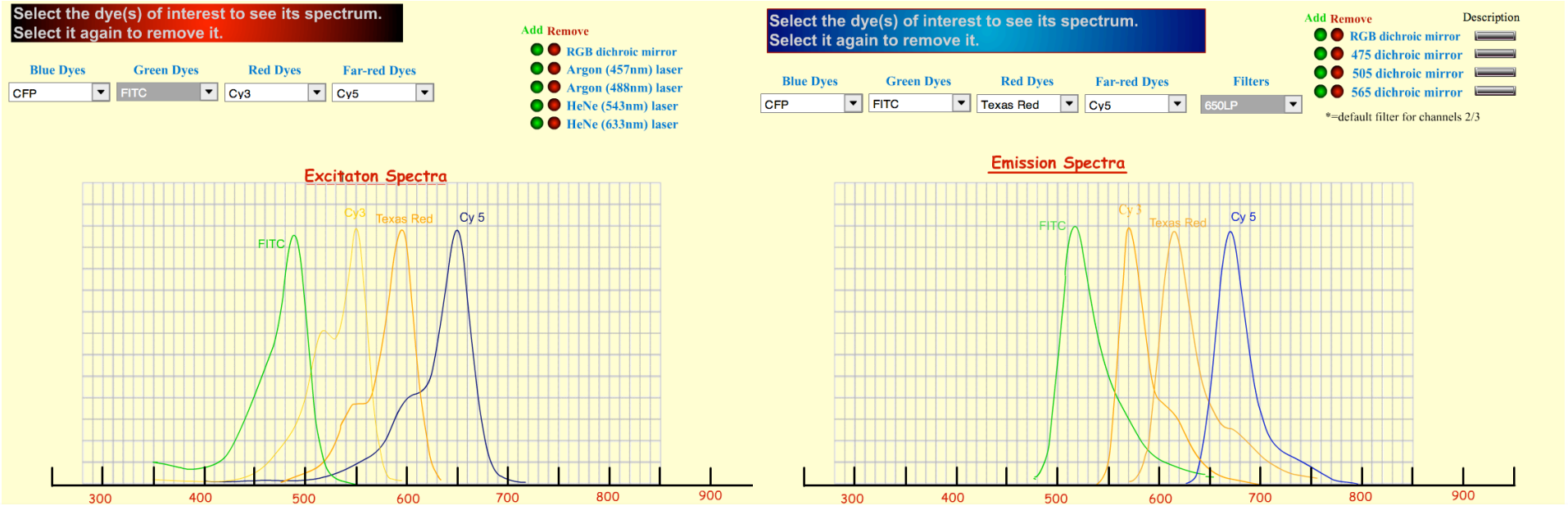
Read intensities



```
#CH4:OBJ13808
3210.3 3617.2   242.0   519.3
  62.1   128.3 -101.5 2556.5
 125.8   145.9  -34.9 2685.2
1968.5 1053.9     4.5   59.6
1732.1   995.9     5.0   48.1
```

Estimate cross-talk and phasing, then call bases

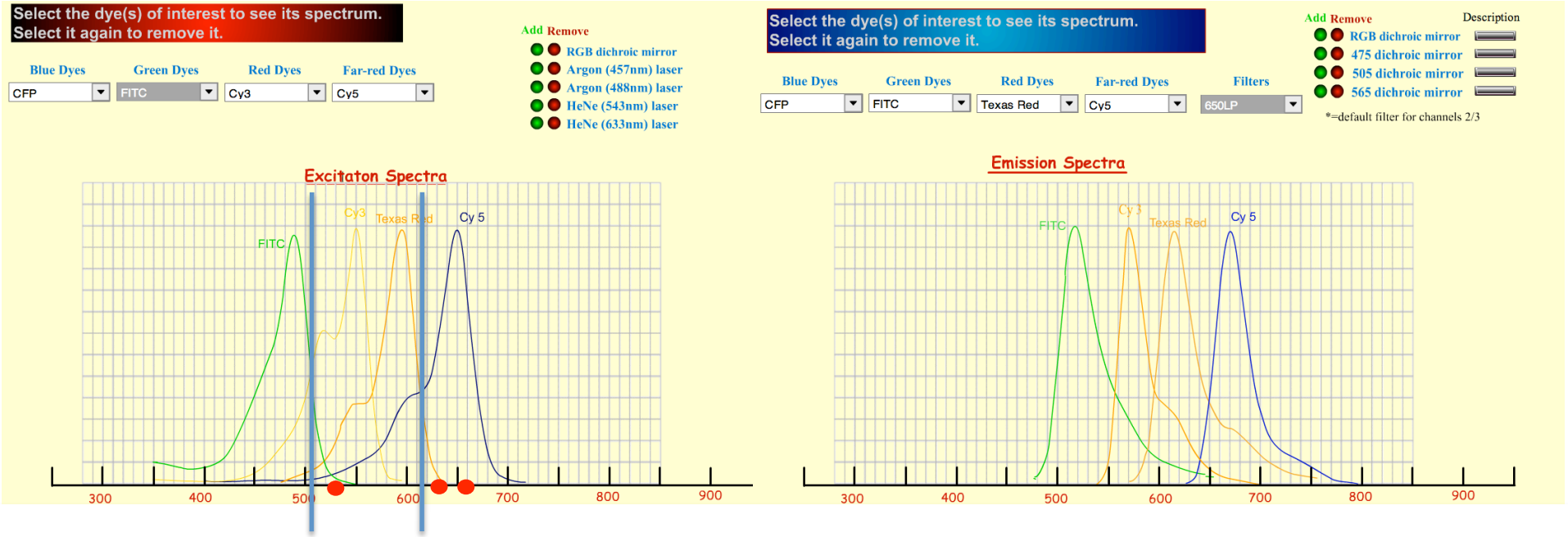
Cross Talk



Excitation spectra shows efficiency of wave-length absorption
Emission spectra shows wave-length of emitted light
Wave-length of emission ~ independent of absorption

Dyes taken from SOLiD marketing material, with FAM replaced by FITC (excitation and emission spectra not available).
Spectra from http://www.mcb.arizona.edu/IPC/spectra_page.htm

Cross Talk



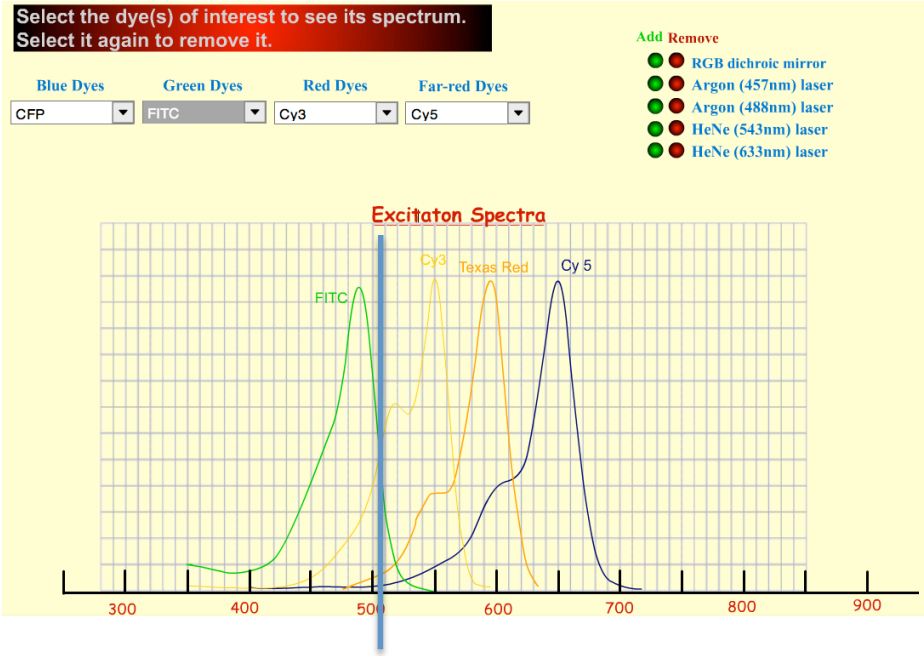
Laser 1 Laser 2 (guesses)

Pick lasers to excite as few fluorophores as possible

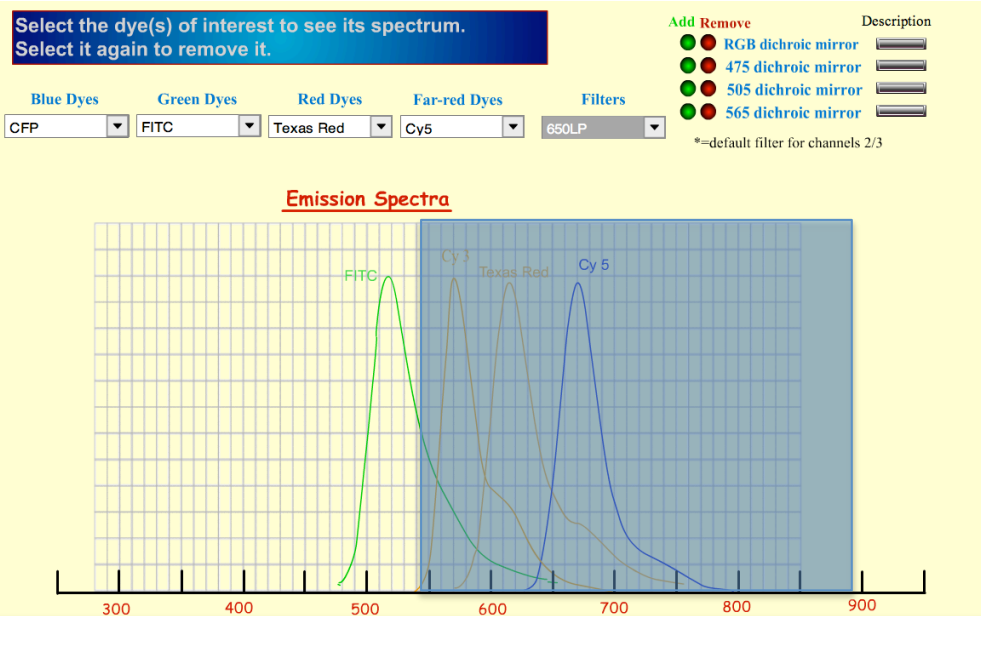
- Each putative laser excites two fluorophores
- Laser 1 excites Texas Red and Cy5 a small amount

Illumina uses a three laser system with wave length 532nm, 635nm and 660nm, two of which are used for imaging and one for focus. Shown by red dots

Cross Talk

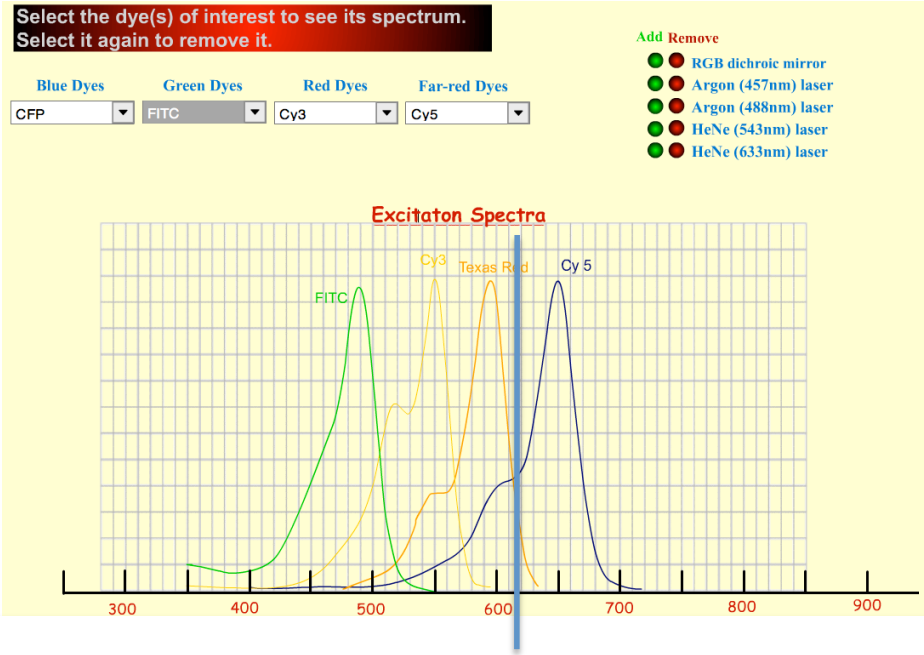


Exciting both FITC and Cy3 with laser -- mixed emission

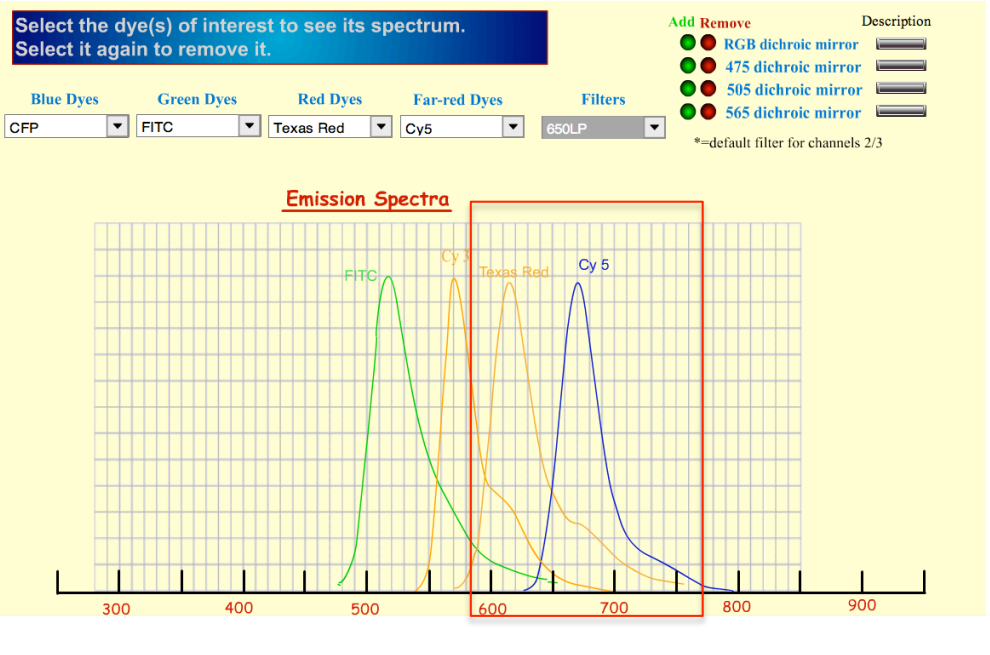


Use a filter to block Cy3 wave lengths, so observed signal is pure

Cross Talk



Exciting both Texas Read and Cy5 with laser -- mixed emission



Emission spectra have strong overlap, hard to construct filter to only allow one through

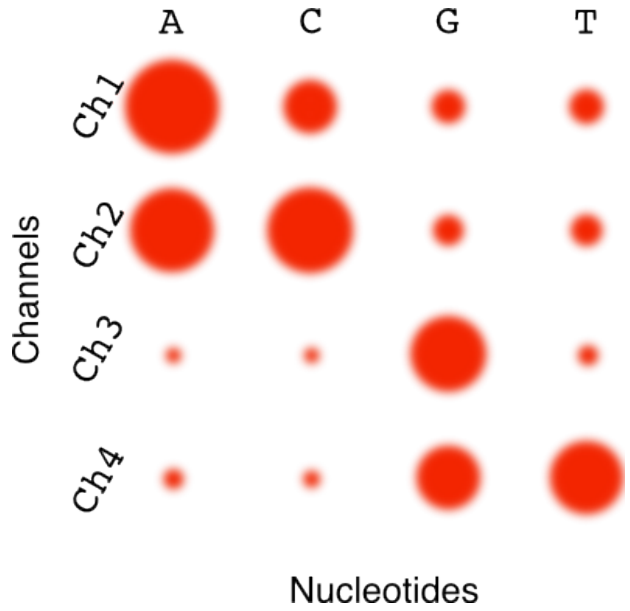
Cross talk

Channel = specific combination of laser and filter

Observe channels rather than nucleotides

Represent cross talk by a matrix

Entries represent how bright each fluorophore appears in each channel



Cross Talk

Laser = coherent light,
Regular patterns of light and dark depending on wavelength
Use a mode scrambler to even out

Mode scrambler problems,
bright and dark patches

Bubble in flow cell,
All clusters lost here for
this cycle

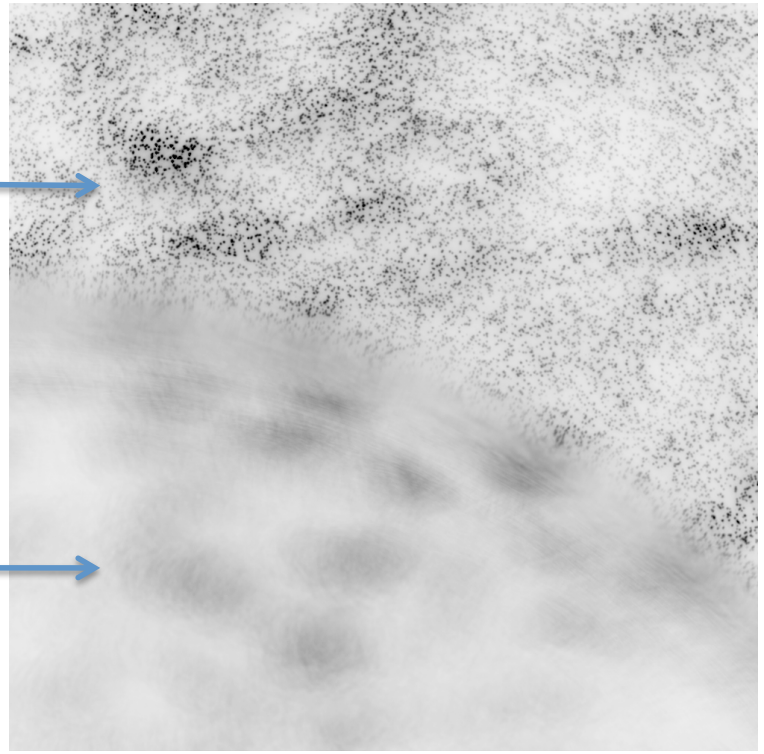
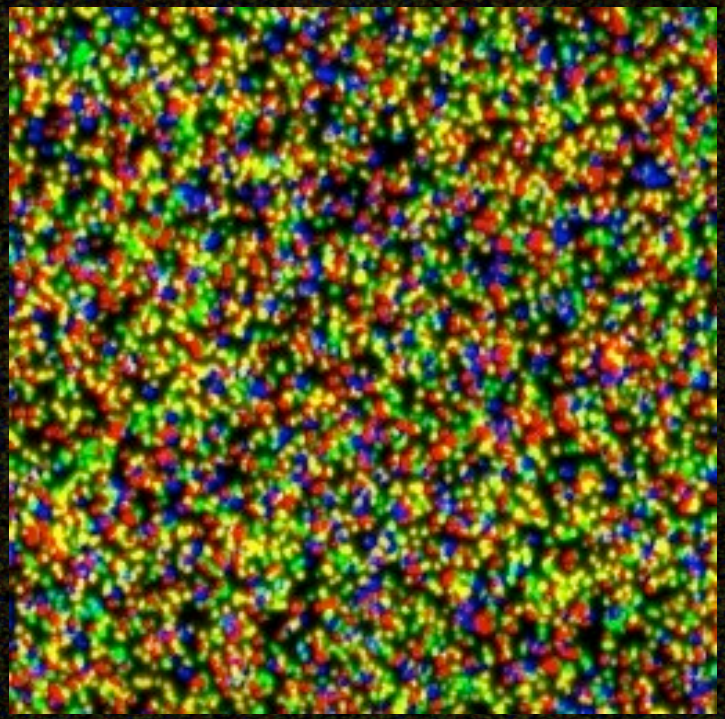
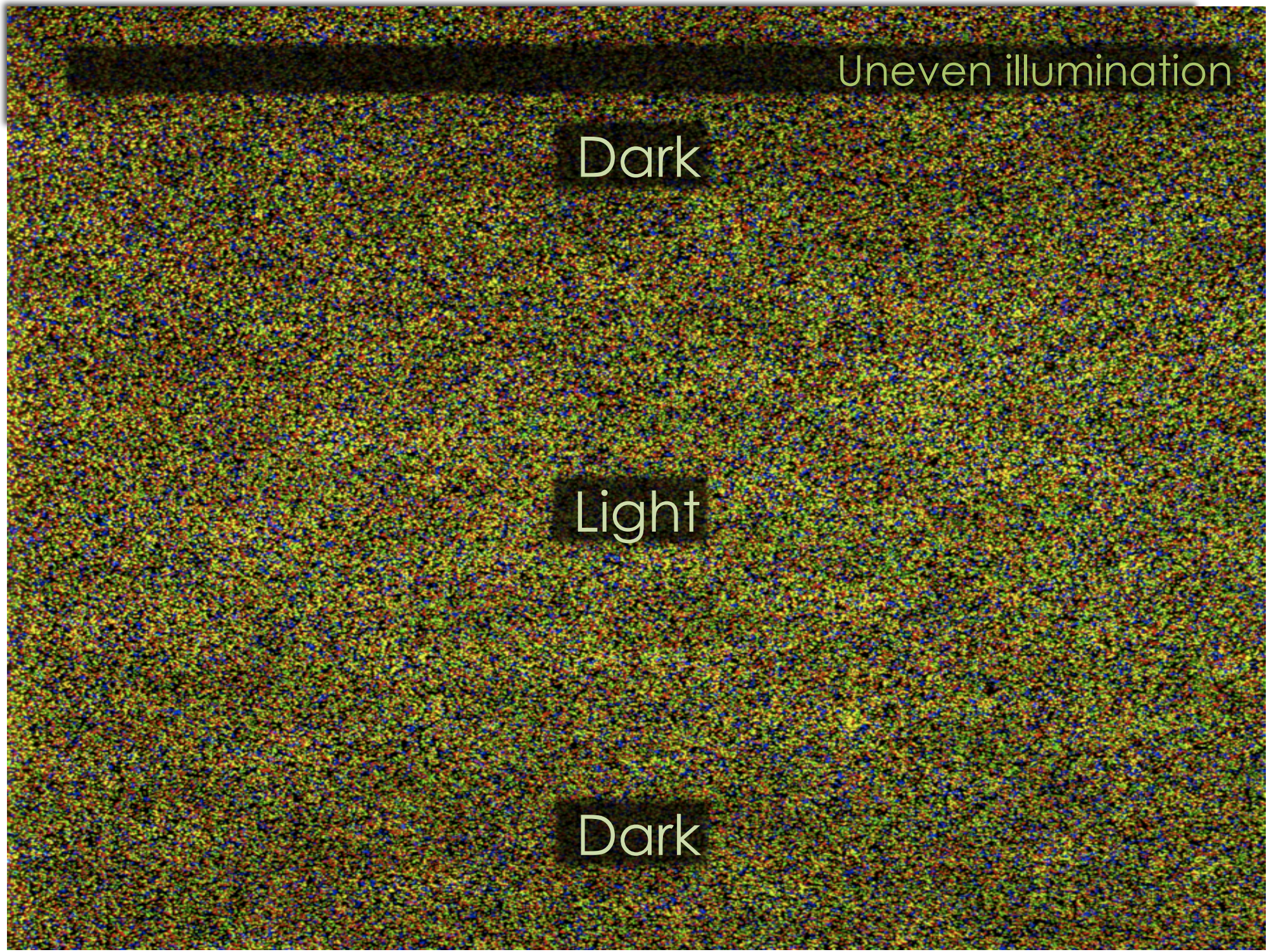


Image courtesy of James Bonfield, Sanger Institute

False colour image of first cycle, crosstalk corrected

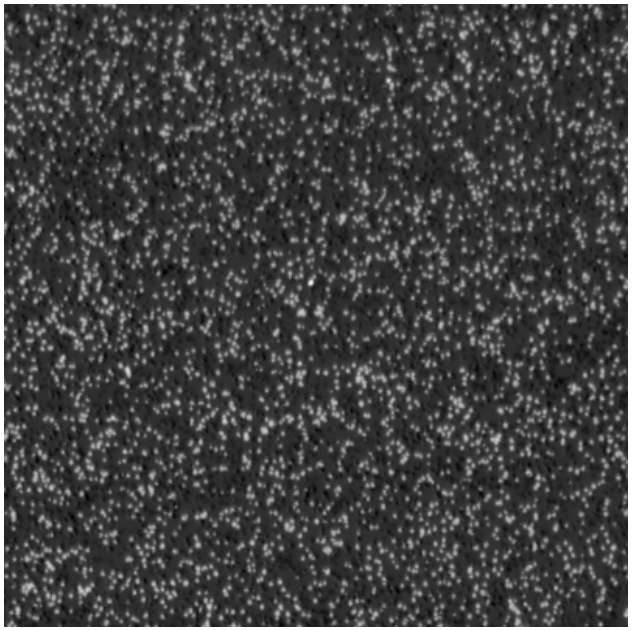



Uneven illumination
Dark
Light
Dark


Wave-like ripples in the illumination
Dark
Light
Dark

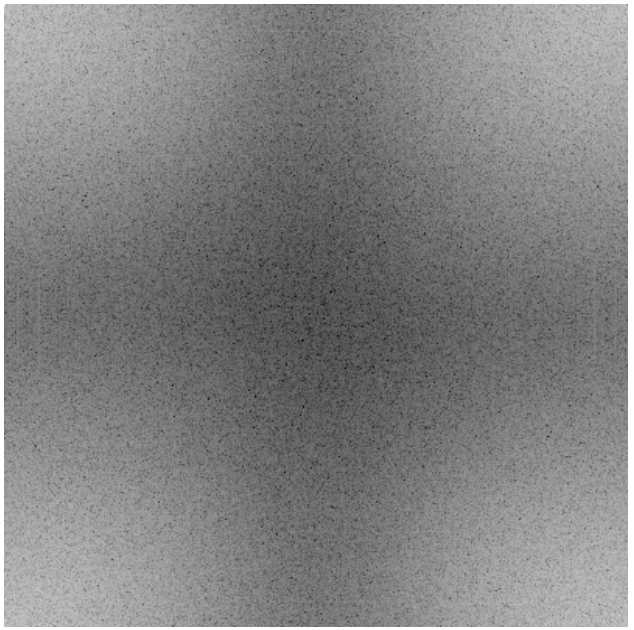
Variation in luminescence

Original image



Fourier transform

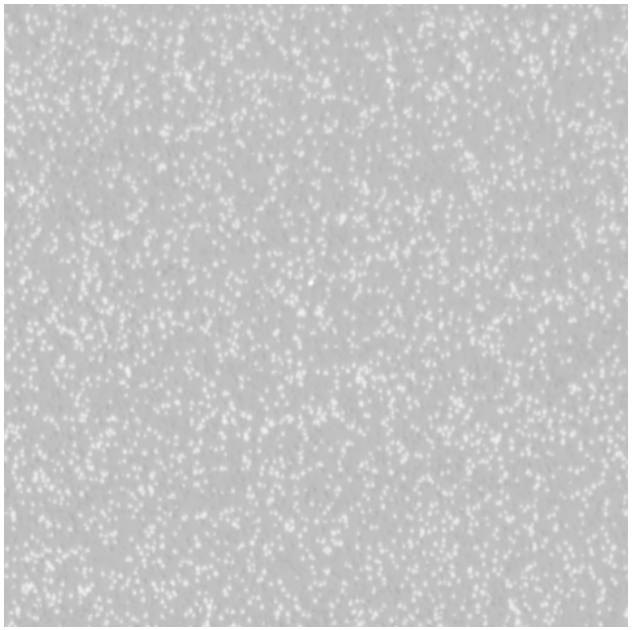
log Mod FT



Variation in luminescence

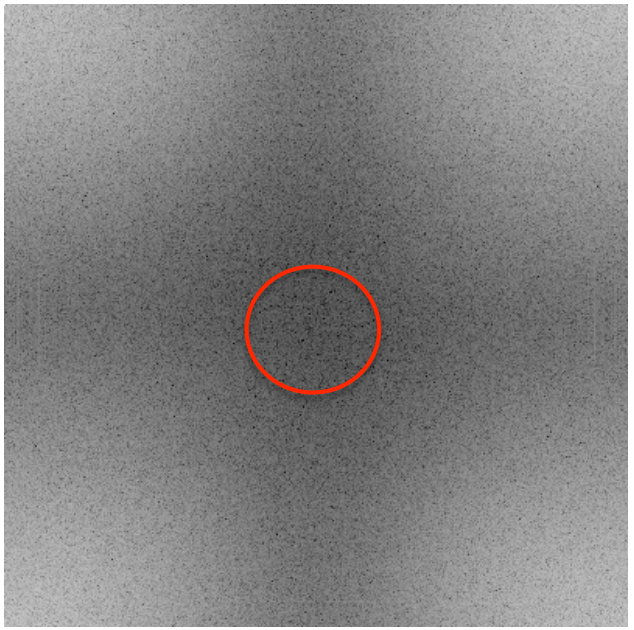
Intensity changes slowly compared to presence / absence of cluster

Original image



Low pass filter
Keep only slowly varying changes

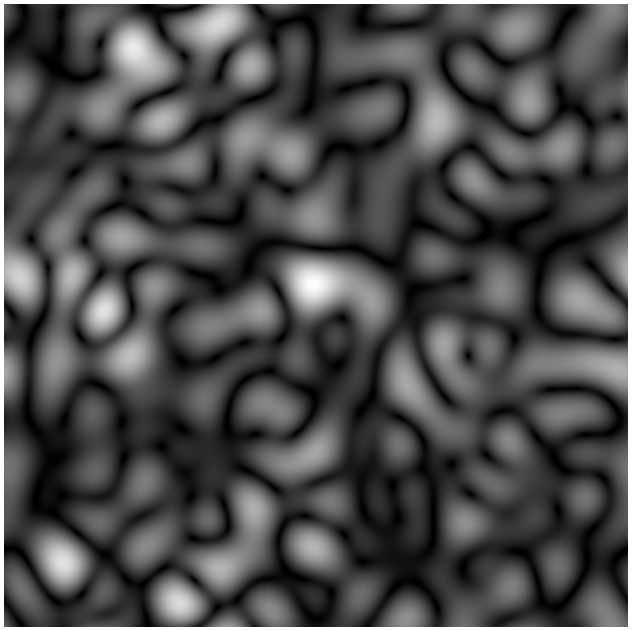
log Mod FT



"optimal filtering"
- a step function

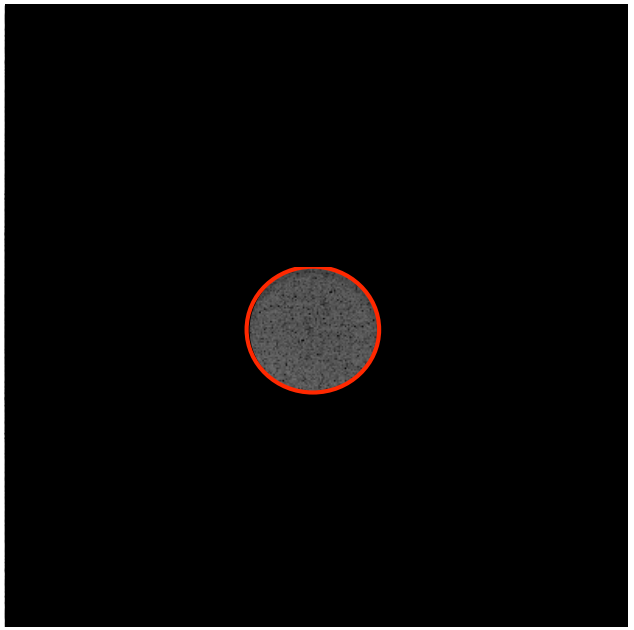
Variation in luminescence

Filtered image



Fourier transform

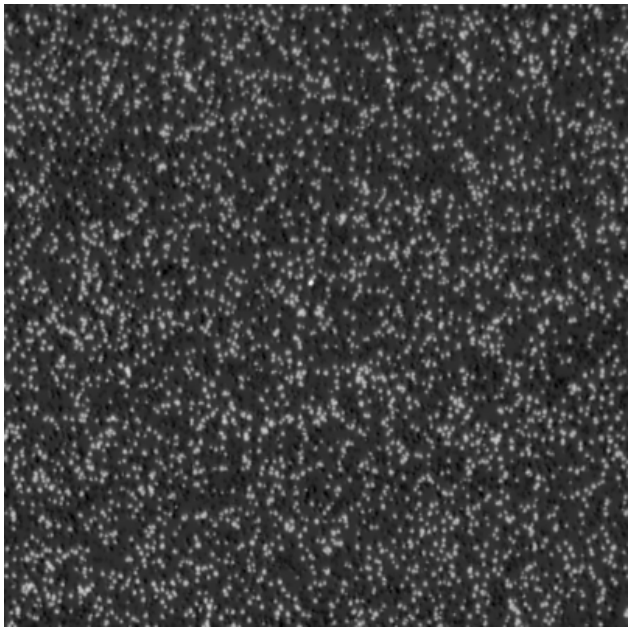
log Mod FT



Variation in luminescence

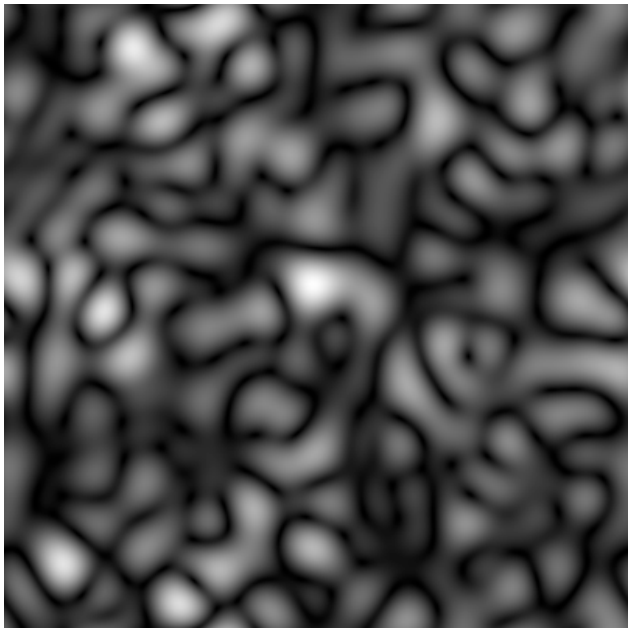
Channel A

IQR: -3.5×10^{-5} -- 4.8×10^{-3}



Filtered, normalized

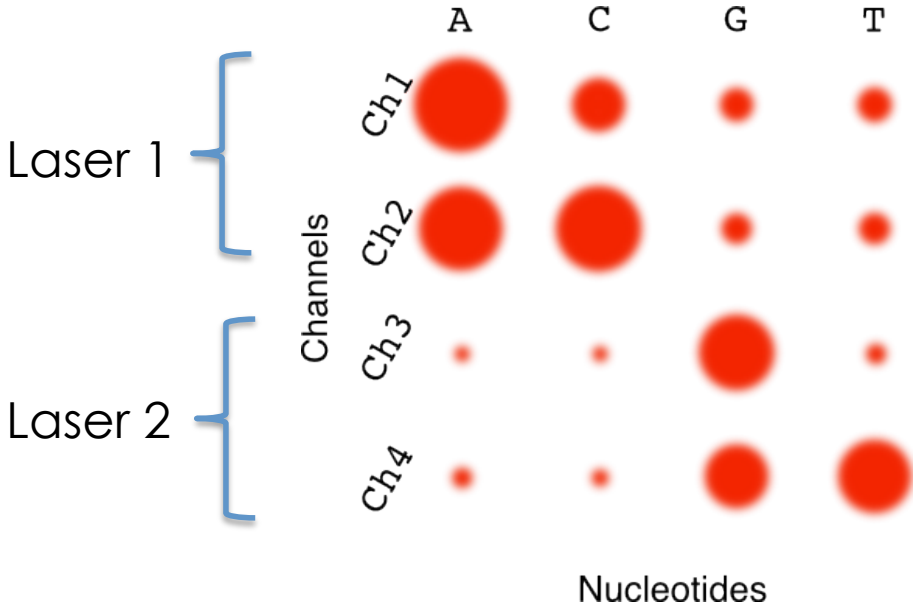
IQR: 3.1×10^{-6} -- 8.2×10^{-6}



Normalized -
accentuate differences

Cross Talk

Variation in laser intensity across flow cell
- three different lasers, different variation in intensity
- variation in cross talk



Variation between cycles/tiles

Laser warming up, becomes more efficient
Changes in focus
Changes in mode scrambler
Background fluorescence

Effects mostly ignored

Phasing

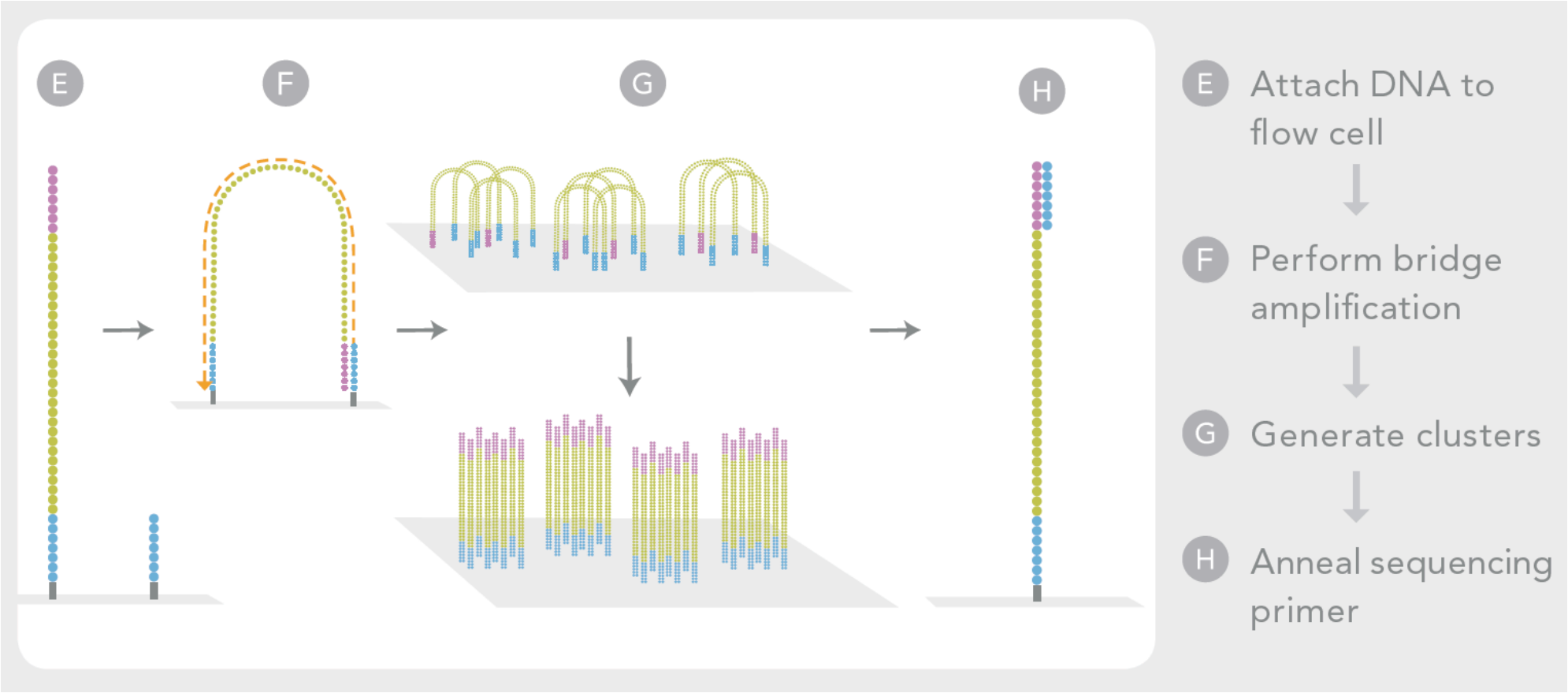
Tendency for molecules to get out of step with others in cluster

Signal from cluster becomes a mixture of previous and future bases

Blurs and becomes harder to tell what current base is

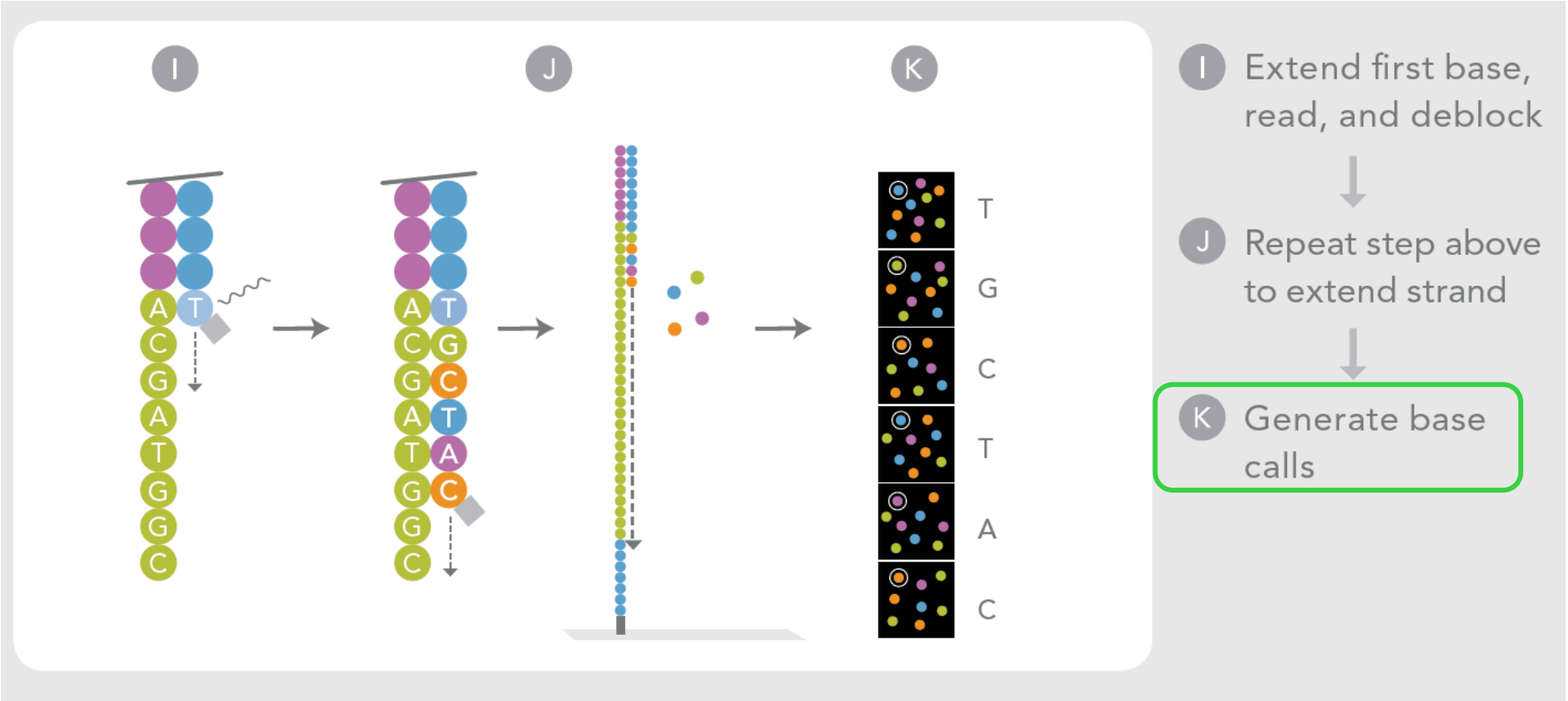
Primarily a chemistry problem

Illumina chemistry



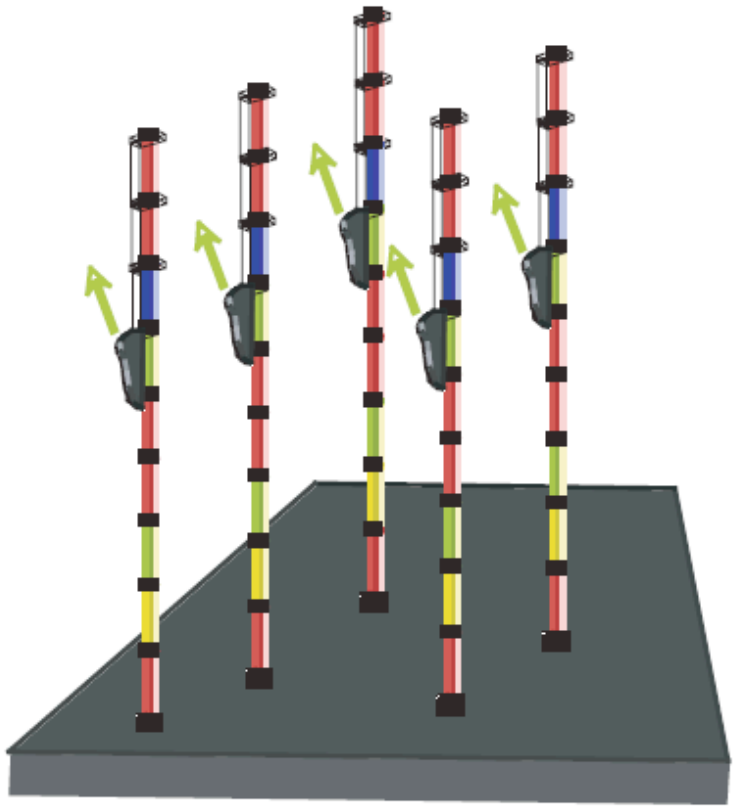
Source: illumina www.illumina.com/sequencing

Chemistry/Physics



Source: illumina www.illumina.com/sequencing

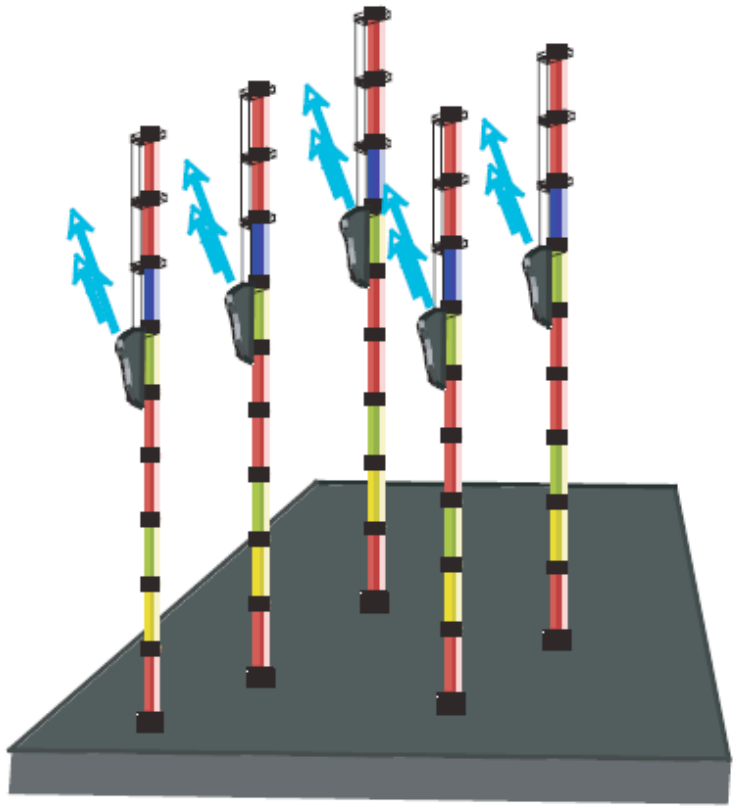
Ideal data



Ideally, signal is strong (green arrows)

Source: Erlich *et al.* (2008) *Nature Methods* **5**:679–682

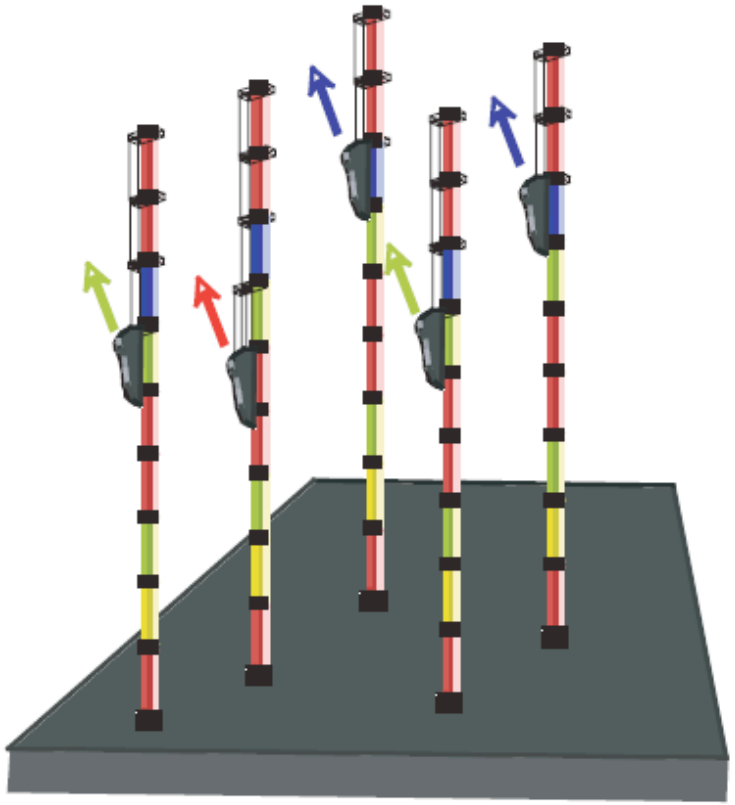
Real data



Laser cross-talk: changes in measured light emissions, leading to distorted signal (blue arrows)

Source: Erlich *et al.* (2008) *Nature Methods* **5**:679–682

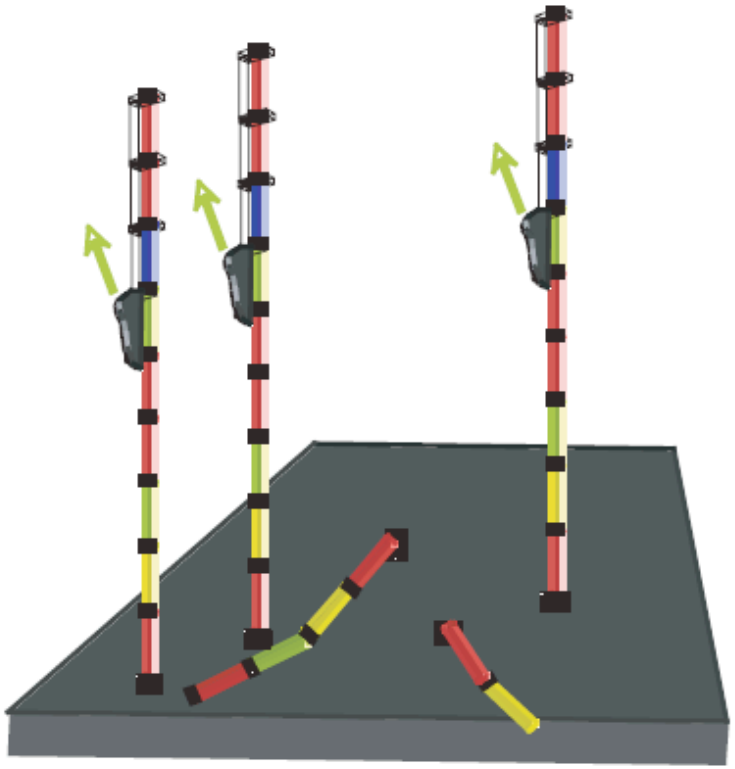
Real data



Phasing:
some strands lead (red) or lag behind (blue), leading to mixed signal

Source: Erlich *et al.* (2008) *Nature Methods* **5**:679–682

Real data

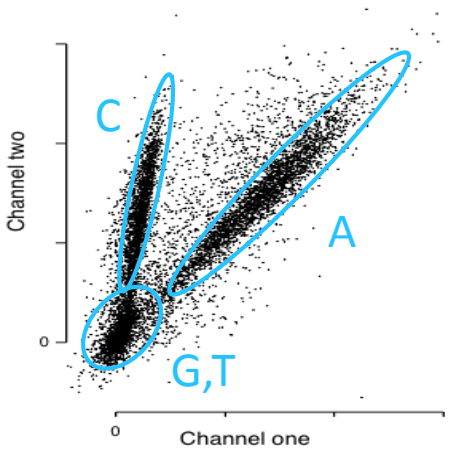


Fading/dimming:
some strands 'die', leading to reduced signal

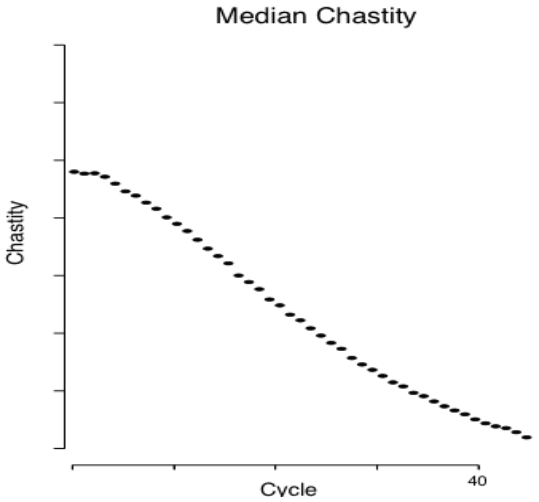
Source: Erlich *et al.* (2008) *Nature Methods* **5**:679–682

Sources of error

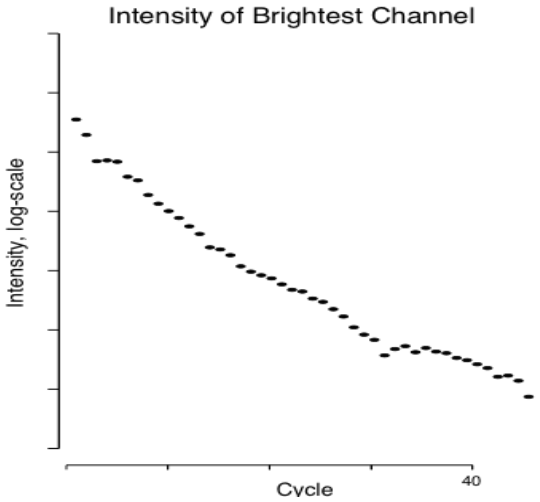
laser cross-talk



phasing

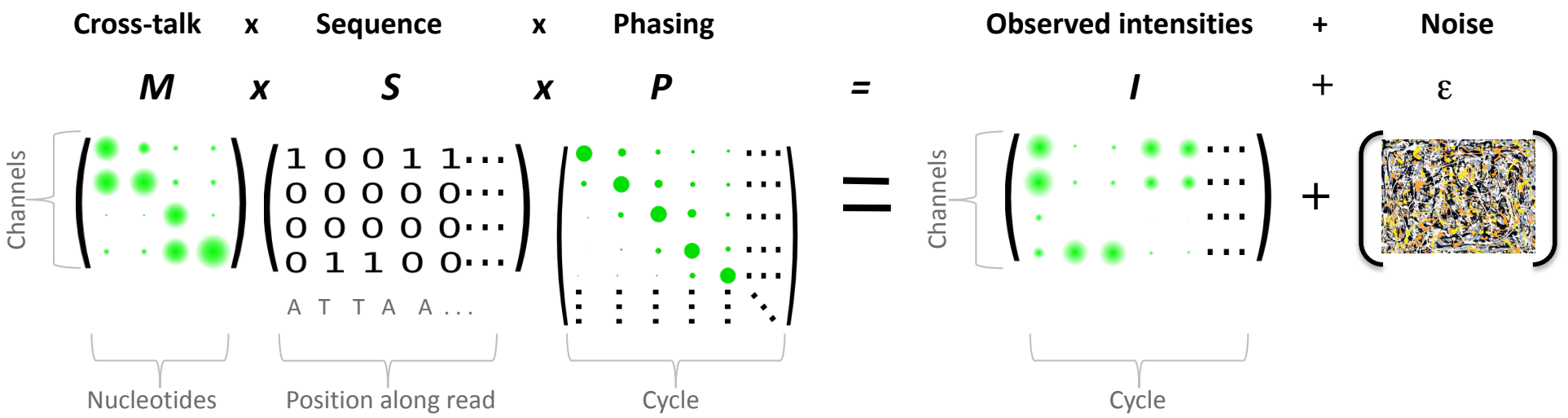


dimming



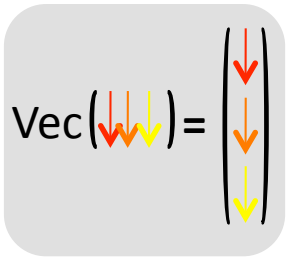
+ contamination
+ flow cell artefacts
+ random error

AYB statistical model

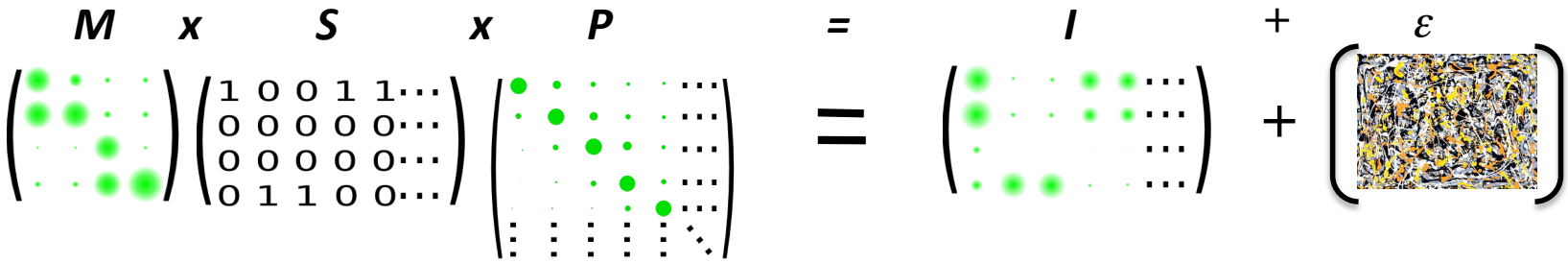


Hidden linear relationship

$$\mathrm{Vec}(I) = (P^T \otimes M)\,\mathrm{Vec}(S) + \mathrm{Vec}(\varepsilon)$$



Sequence inference



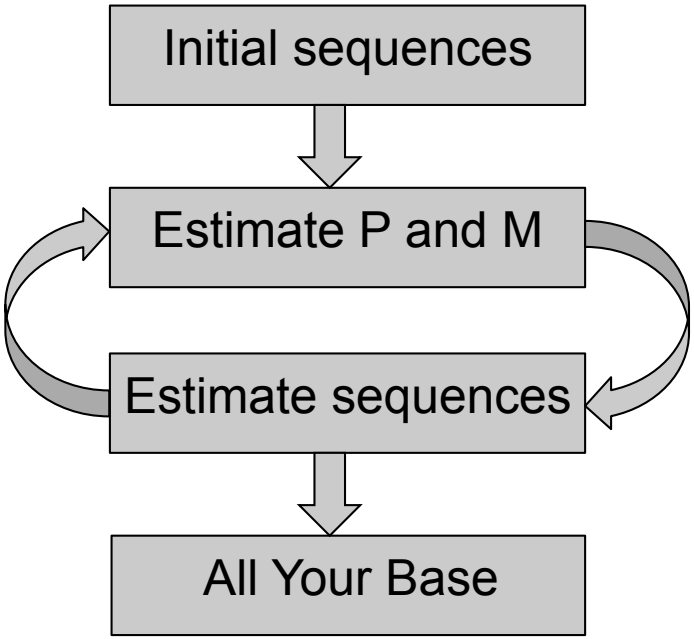
Ordinary Linear Model

noise assumed independent

General Linear Model

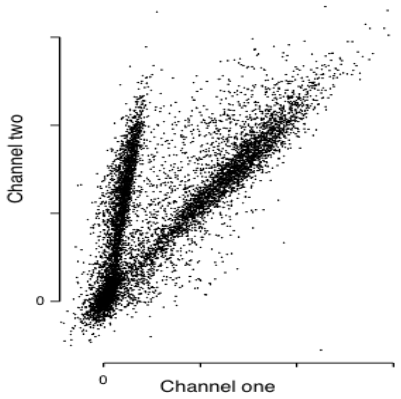
for each cluster

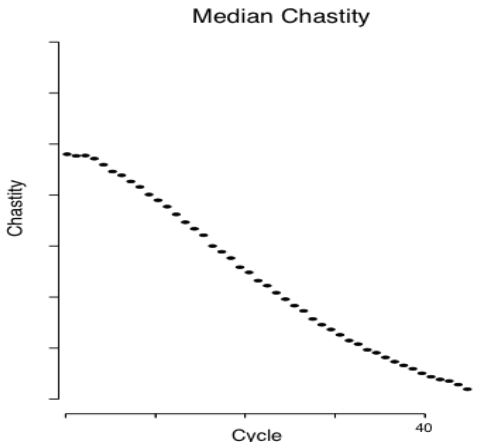
Var(*noise*) estimated using all clusters

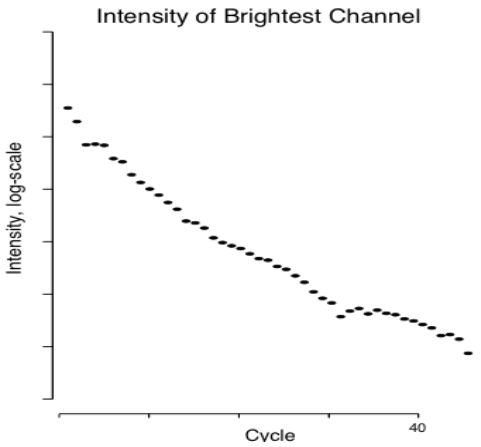


Noise removal

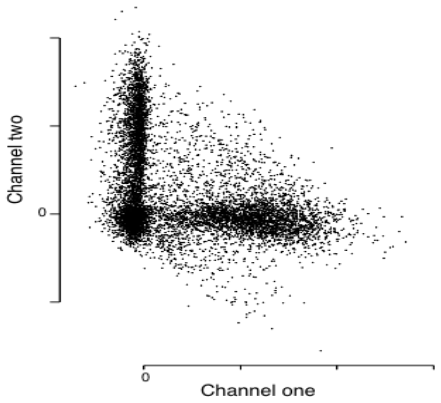
Raw data

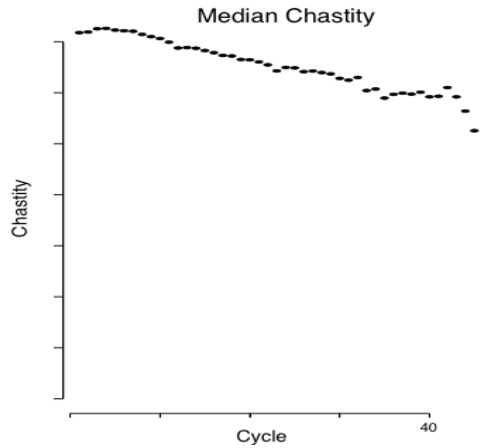


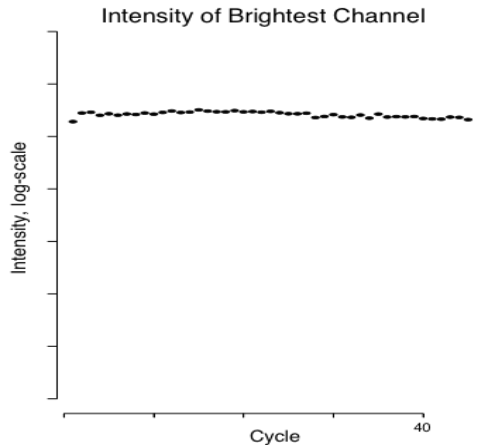




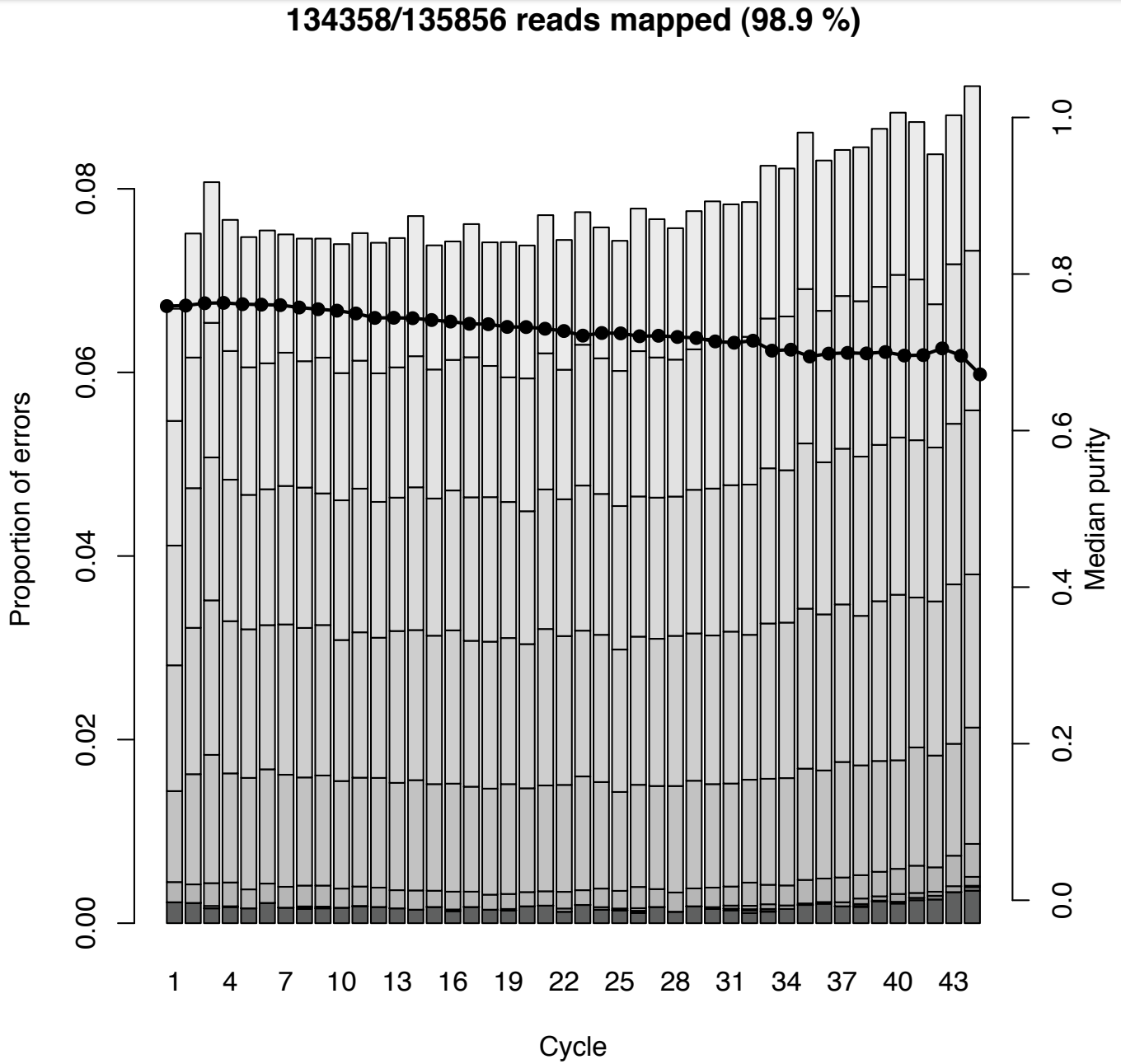
AYB processed



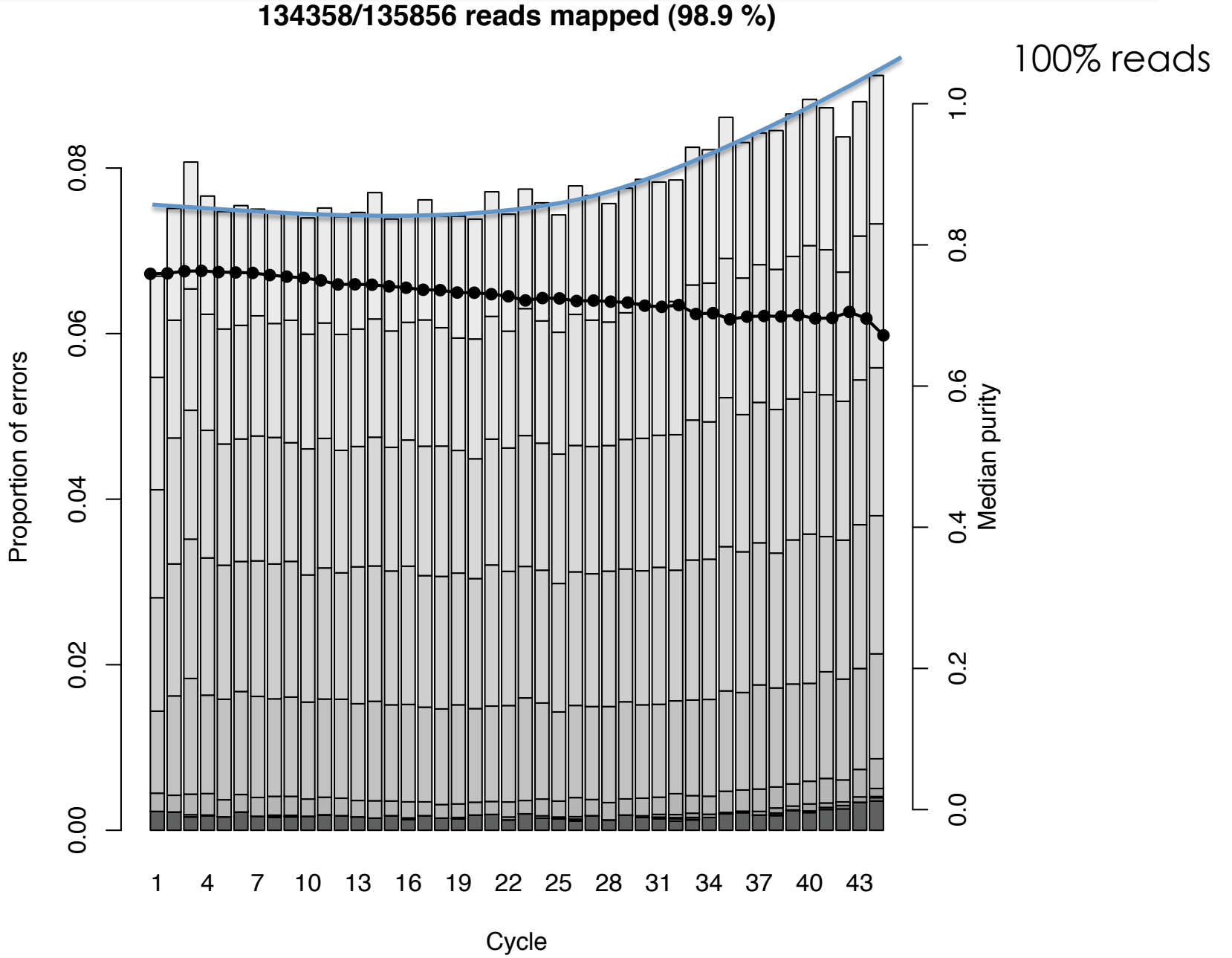




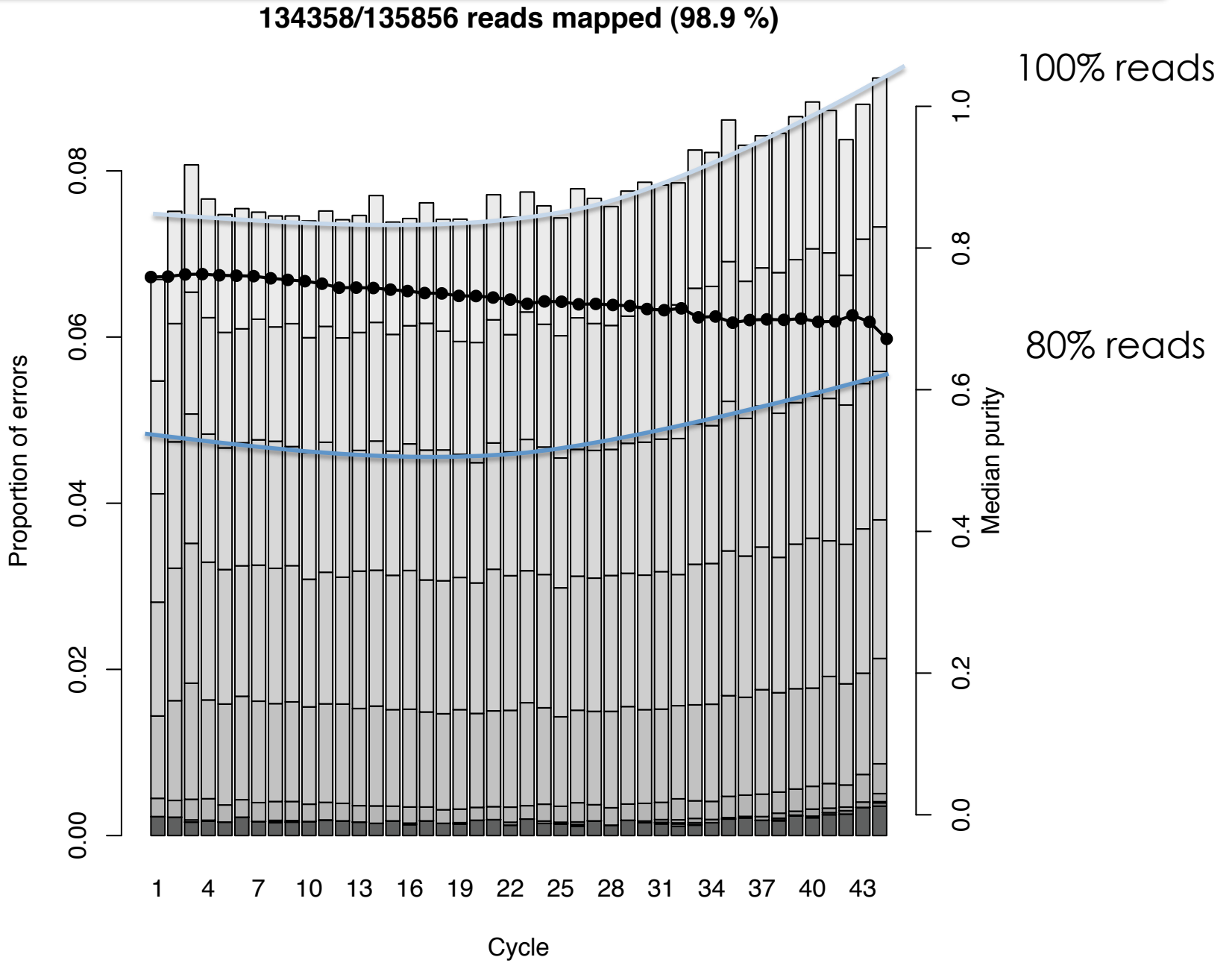
Accuracy


134358/135856 reads mapped (98.9 %)
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle

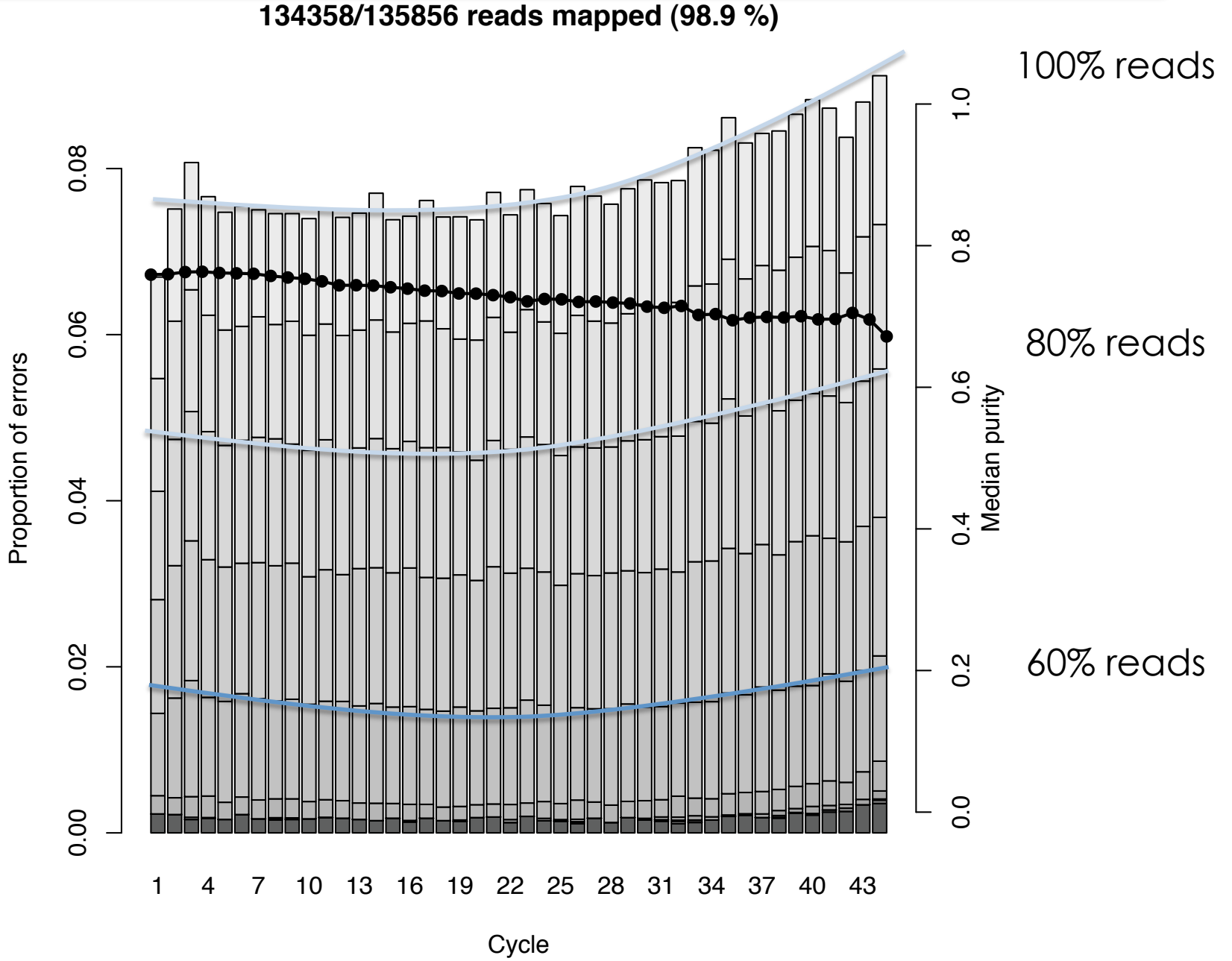
Accuracy


134358/135856 reads mapped (98.9 %)
100% reads
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle

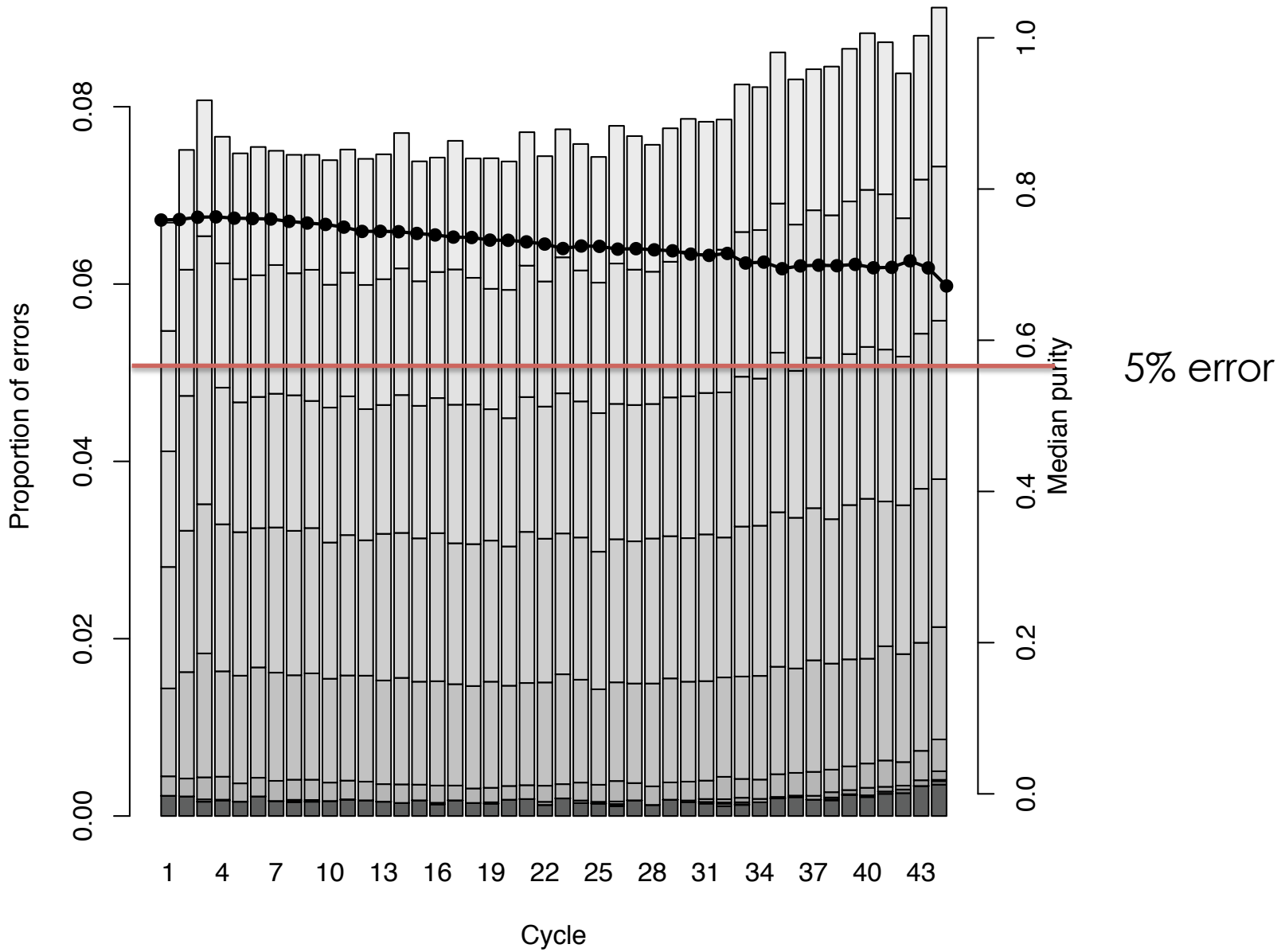
Accuracy


134358/135856 reads mapped (98.9 %)
100% reads
80% reads
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle

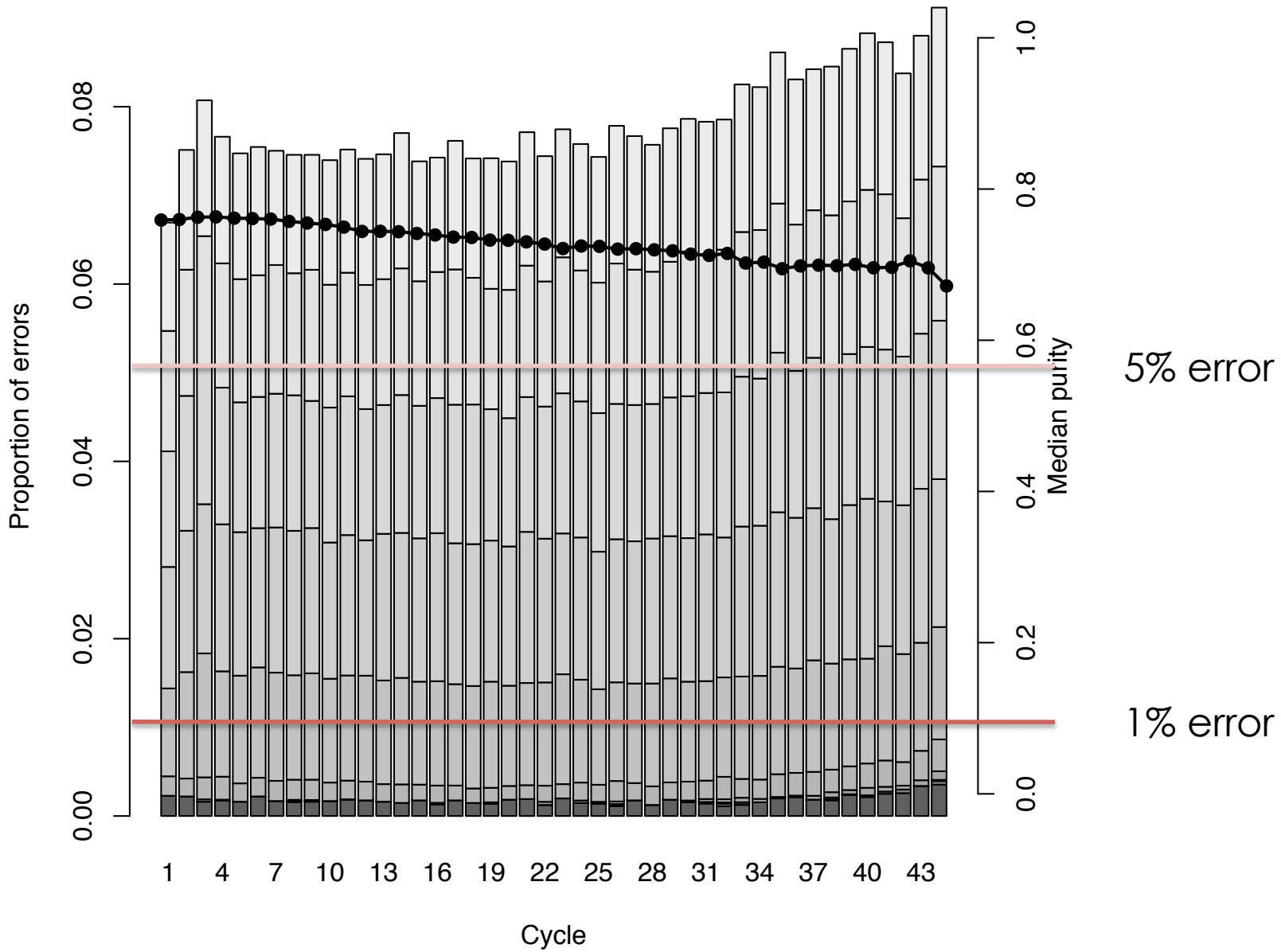
Accuracy


134358/135856 reads mapped (98.9 %)
100% reads
80% reads
60% reads
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle

Accuracy

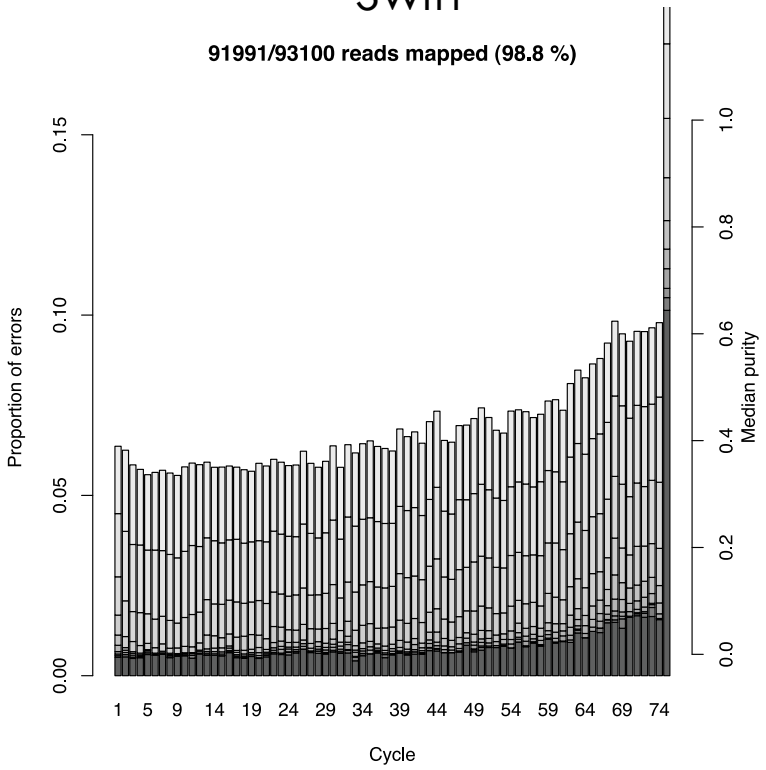

134358/135856 reads mapped (98.9 %)
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle
5% error

Accuracy

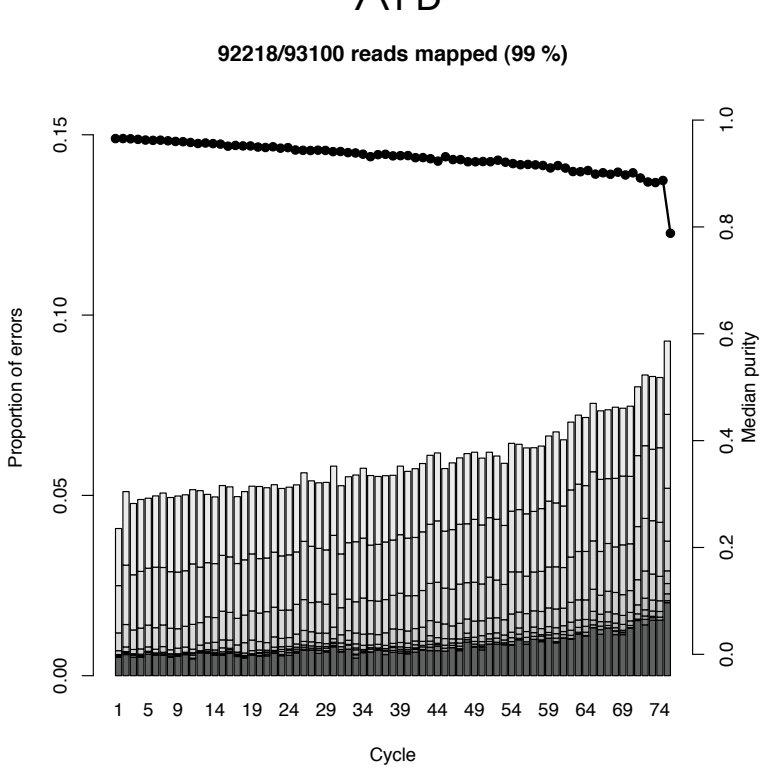

134358/135856 reads mapped (98.9 %)
Proportion of errors
0.00
0.02
0.04
0.06
0.08
Median purity
0.0
0.2
0.4
0.6
0.8
1.0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
Cycle
5% error
1% error

75 cycle comparison

Swift



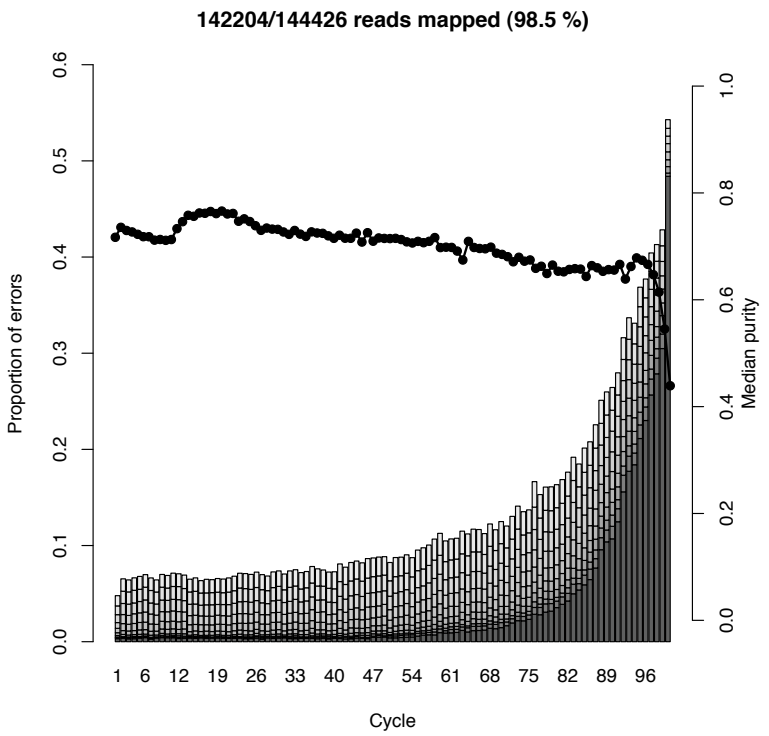
AYB



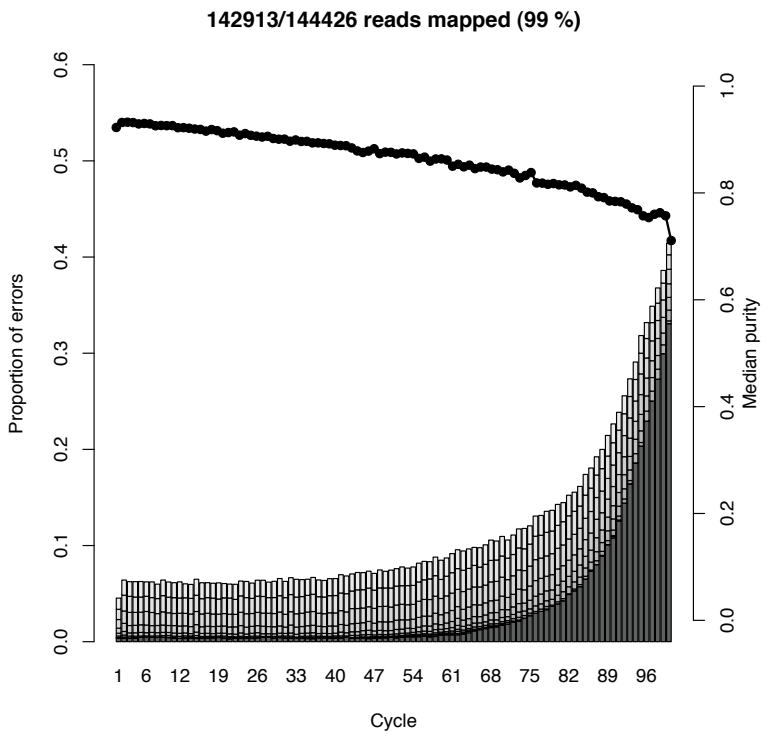
No purity values for Swift because of bug, ordered by purity for AYB

100 cycle accuracy

Swift

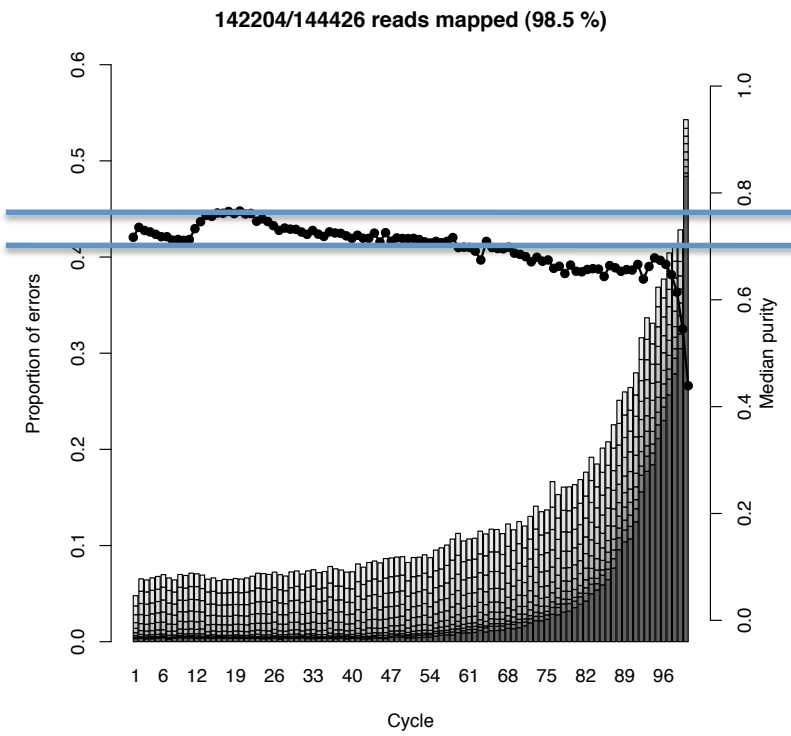


AYB

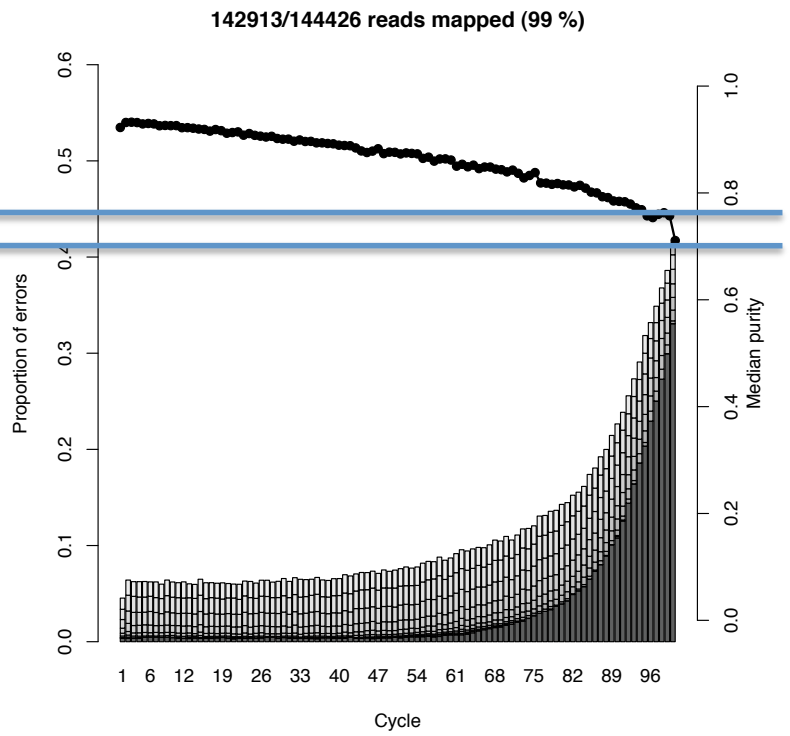


100 cycle accuracy

Swift



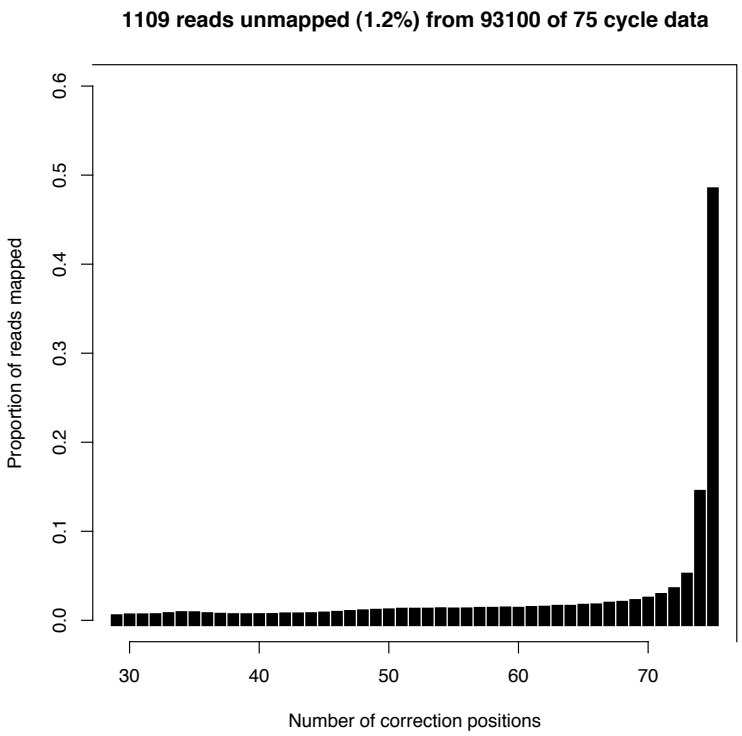
AYB



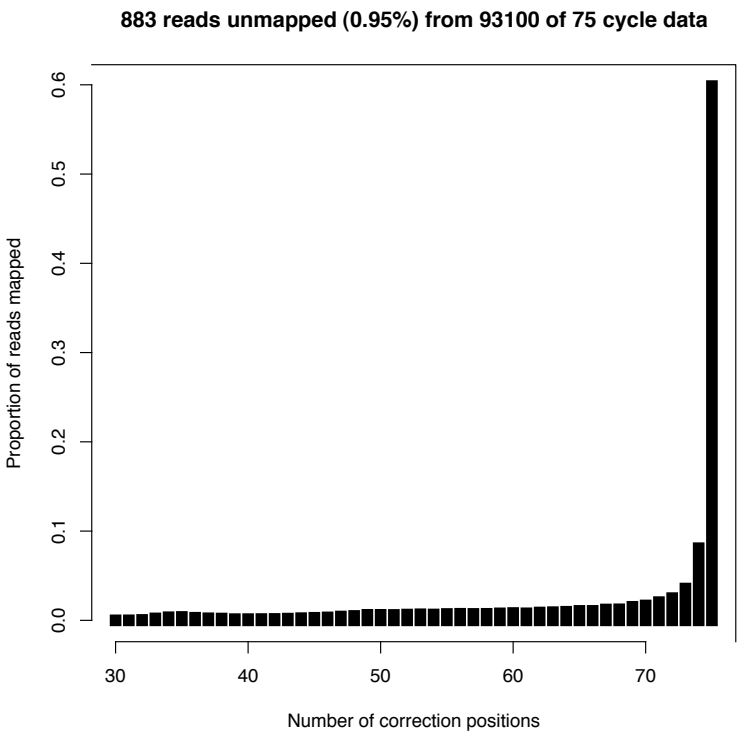
Number of correct reads

75 cycle phiX data

Swift



AYB



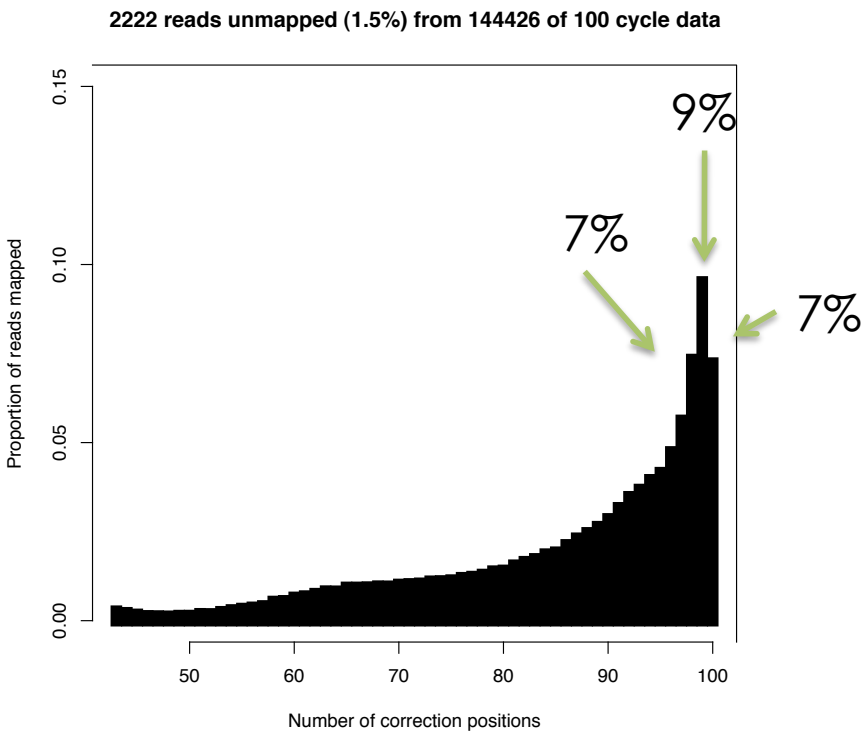
AYB improvement (BWA alignments, edit distance 7)

- 4% more reads aligned
- 25% more perfect reads

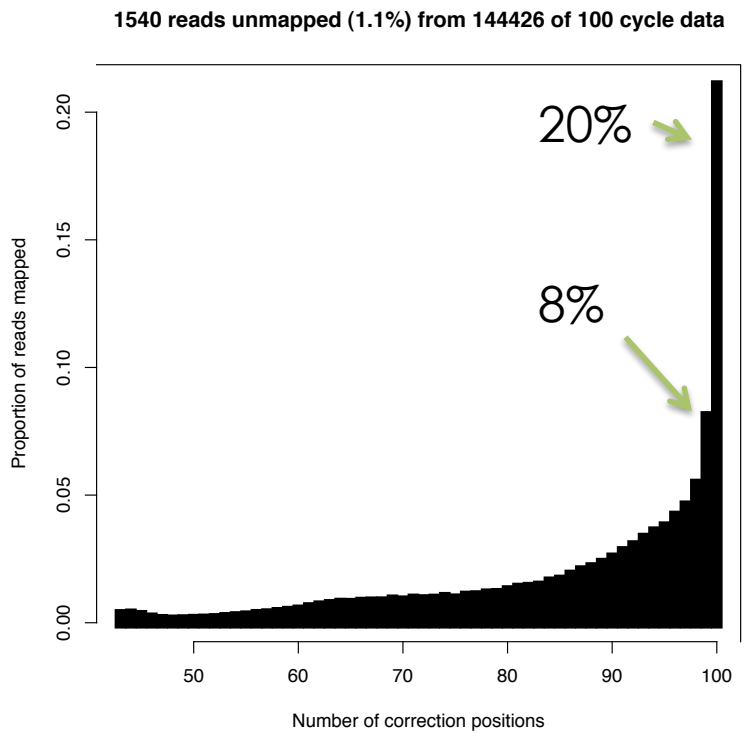
Number of correct reads

100 cycle phiX data

Swift



AYB

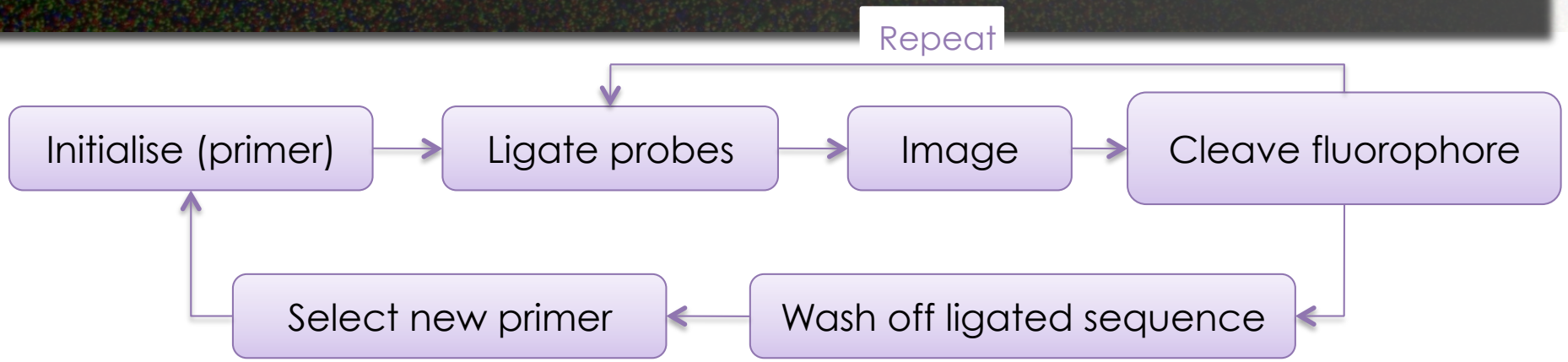


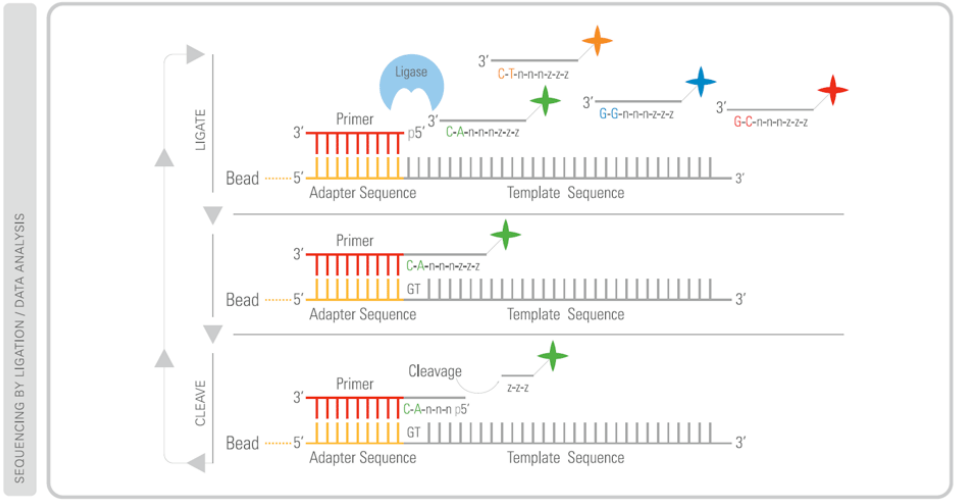
AYB improvement (BWA alignments, edit distance 7)

- 15% more reads aligned
- 180% more perfect reads

Will show later that ~25% of reads in this data set have contamination at final cycle

SOLiD chemistry



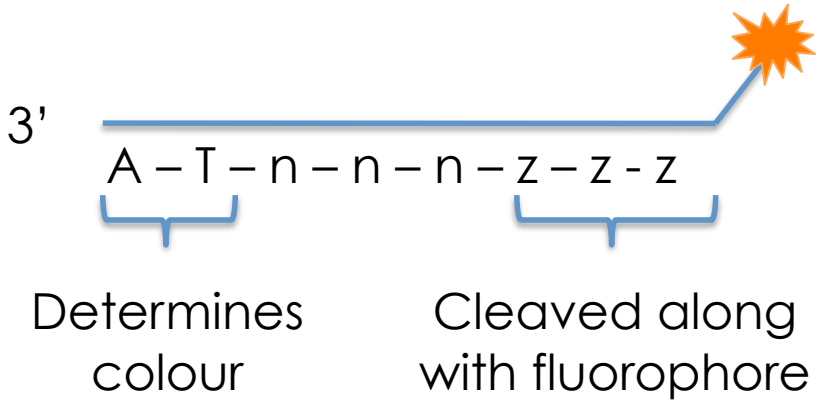


Additional step attaches permanent blockers to reads that where not extended

SOLiD trades phasing for dimming
- less opportunity for correction
- may still be able to improve calls

Source: Applied Biosystems, www.appliedbiosystems.com

SOLiD probes



Errors

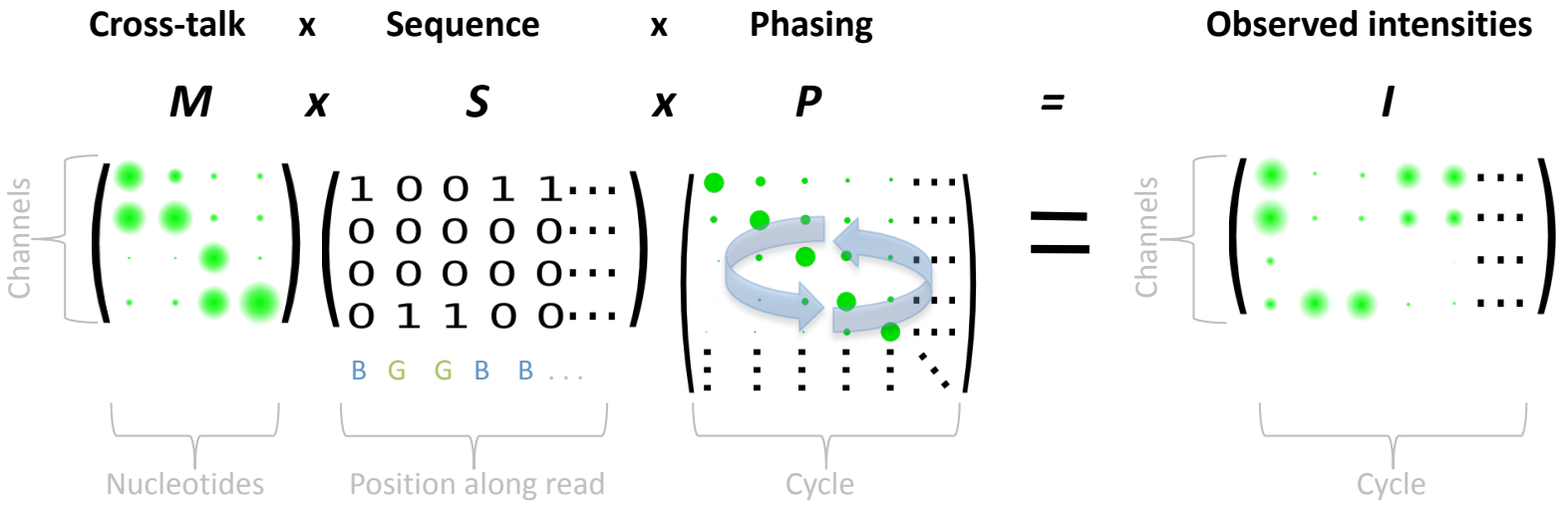
- Failure to ligate probe sequence
- Ligate multiple probe sequences (blocker failure)
- Incorrect cleavage (stop sequencing entirely, or leave additional bases)

All these are captured by phasing matrix

- Bad incorporation (similar but incorrect probe)
  - only problem if first two bases affected
  - transient error

AYB on SOLiD

Same model as before, using colours rather than bases



Permutation of read position to cycle

- permute phasing matrix

Primer reset

- intensity increases each primer round? Less dimming

All these covered since phasing is estimated empirically

Calibration

Measure confidence in each base call

AYB - fit of each possible call to model $\mathbf{P}(\text{data} \mid \text{base is A})$

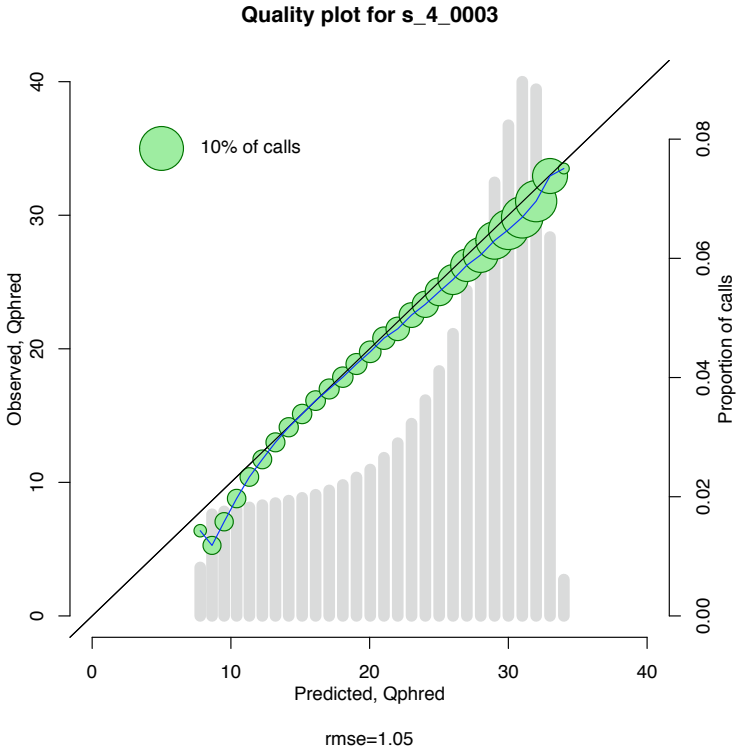
Use robustified Bayesian approach

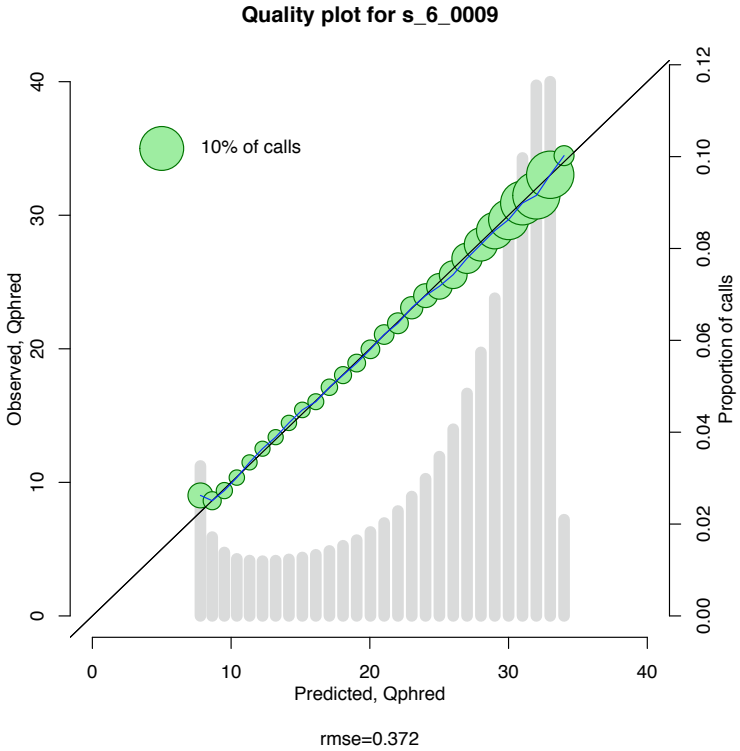
$$\mathbf{P}(\text{base is A}) = \frac{\mathbf{P}(\text{data} \mid \text{base is A}) + \eta}{\sum_{j \in \{A,C,G,T\}} \mathbf{P}(\text{data} \mid \text{base is } j) + \eta}$$

η represents "contamination" from other sources
η ~ Q50 by default

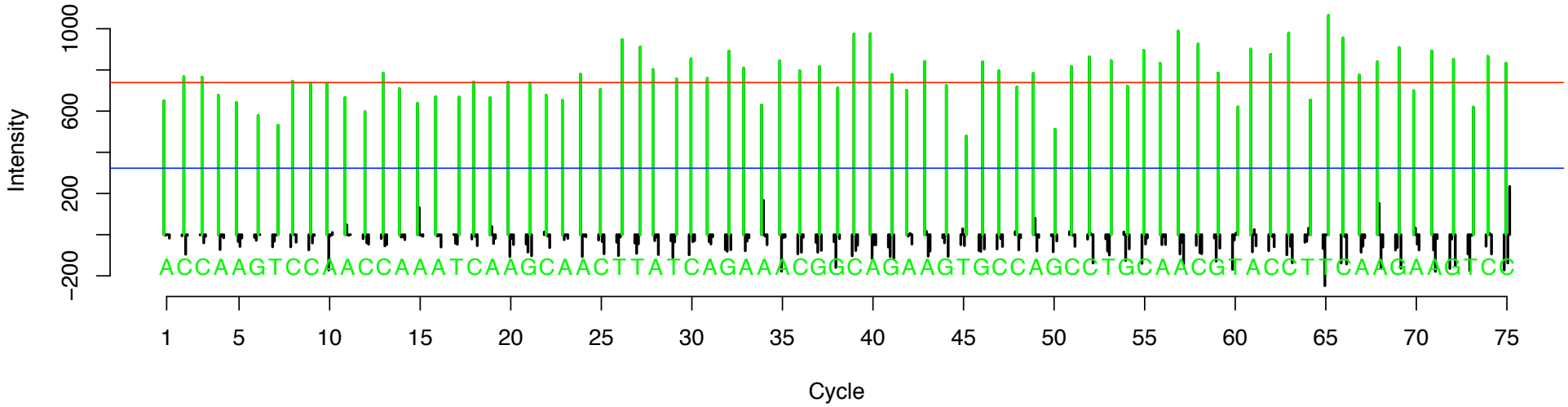
If none of the bases fit well, then posterior probability tends to 0.25
Confidence resets when data does not look like sequence

Calibration





Assessing quality



Unmodelled effects are upper bound on quality

From: $P(B \cup O) = P(B) + P(O) - P(B \cap O)$ $\in [0, \min P(B), P(O)]$

Total error

Base calling error

Other effects

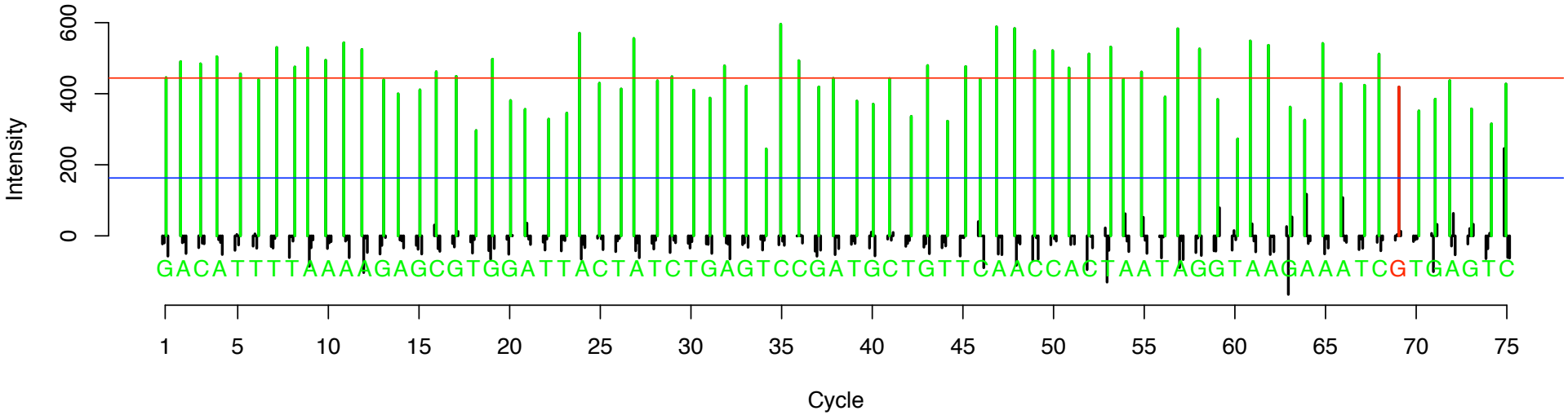
Get: $\min(Q_{\text{Base}}, Q_{\text{Effect}}) \geq Q_{\text{Total}} \geq \min(Q_{\text{Base}}, Q_{\text{Effect}}) - 3$

E.g. polymerase error ~ Q40

Early polymerase errors

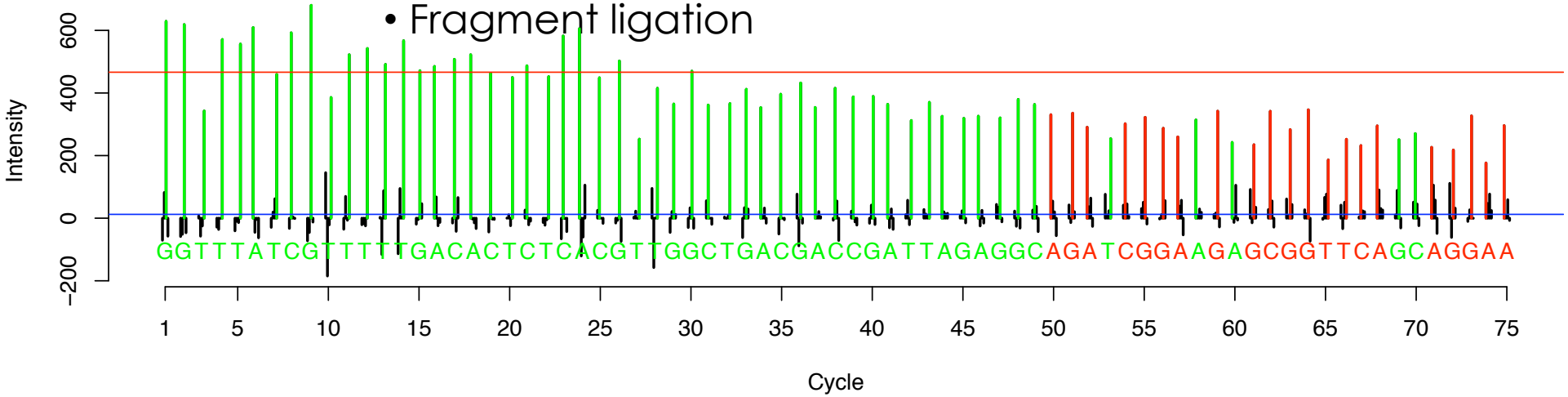
Polymerase error during sample preparation or early amplification

Indistinguishable from a SNP
- sequence with a difference from reference
- call sequence in cluster correctly but get error

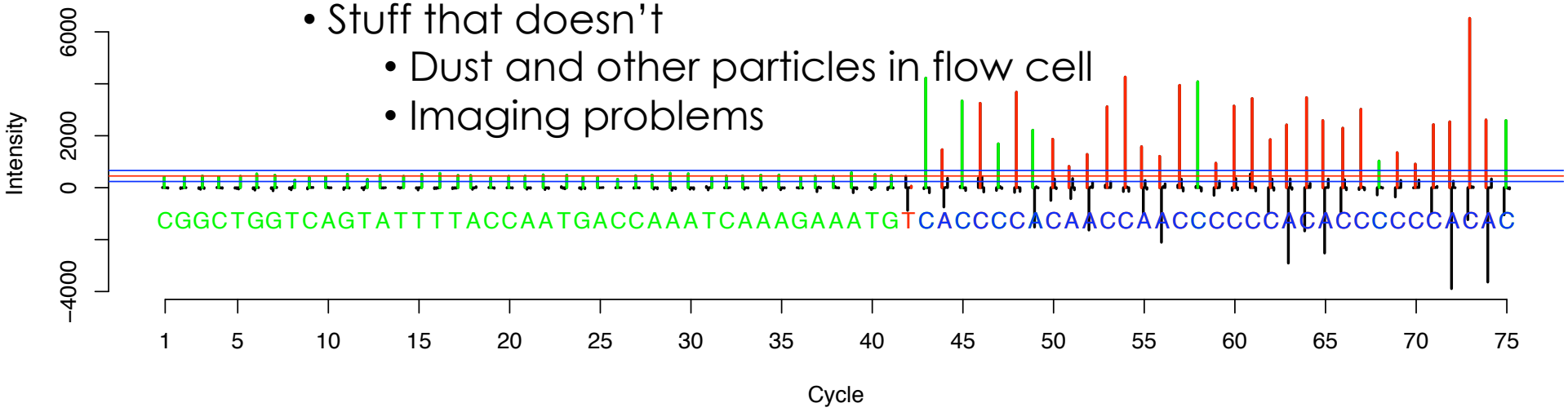


Errors in read

- Stuff that looks like sequence
  - E. coli, H. sapiens etc sequence contamination
  - Bits of replication machinery
  - Adapter sequence
  - Fragment ligation

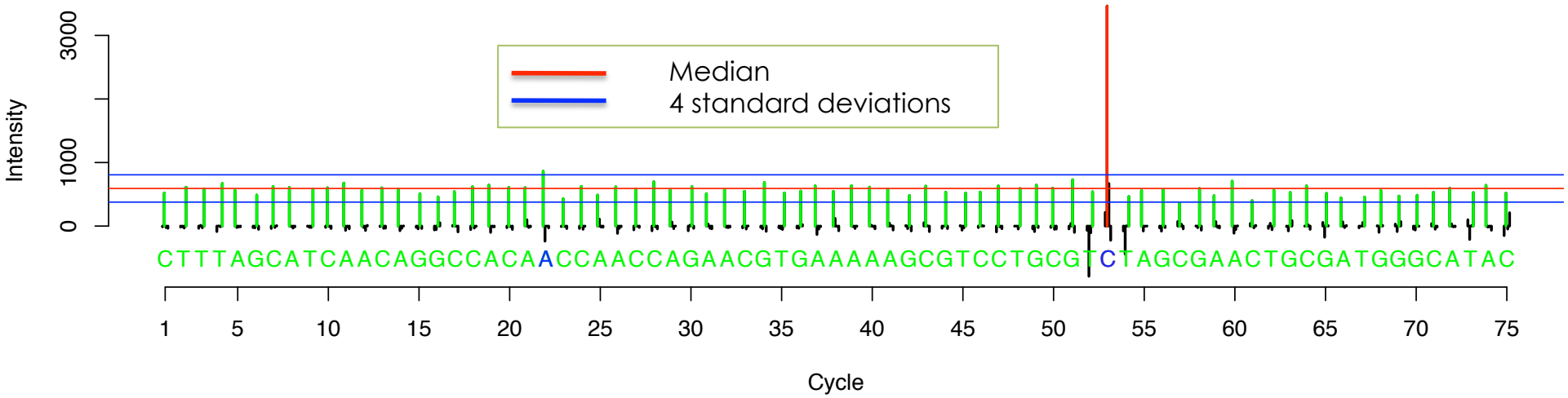


- Stuff that doesn't
  - Dust and other particles in flow cell
  - Imaging problems



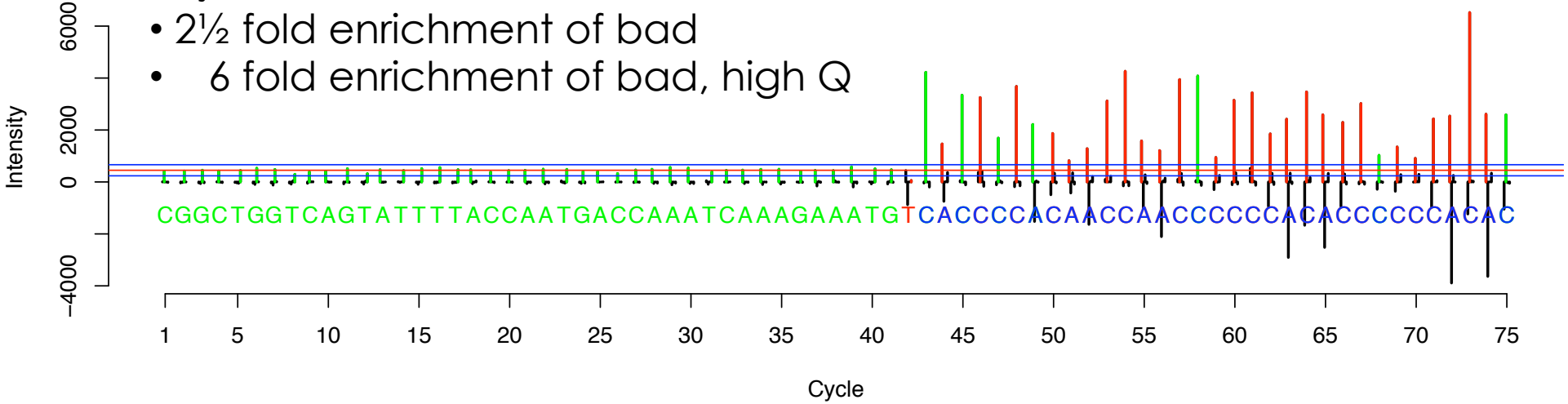
Filtering artefacts

Observation: contaminates are much brighter than ordinary sequence



Reject 0.8% of bases

- 2½ fold enrichment of bad
- 6 fold enrichment of bad, high Q

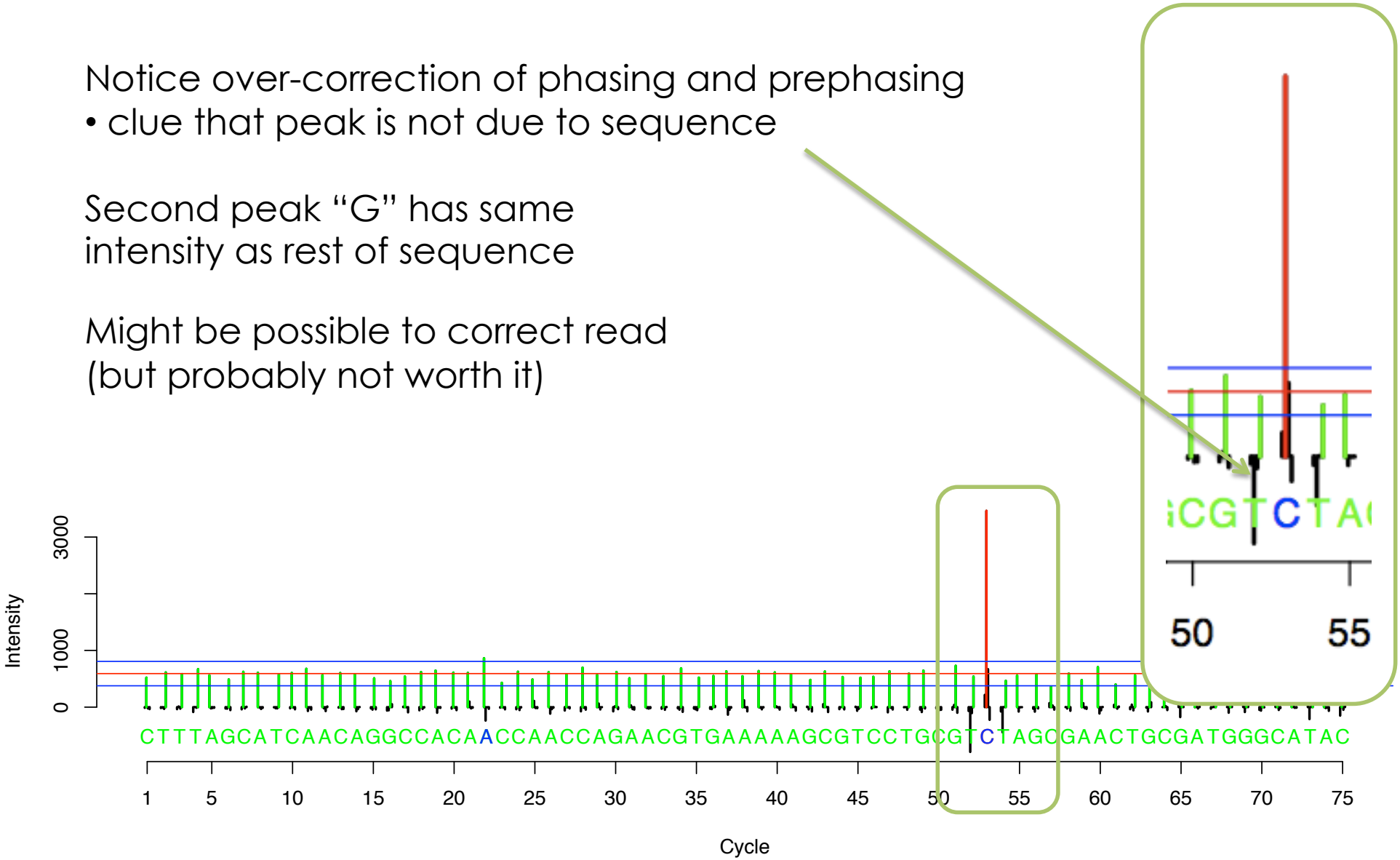


Filtering artefacts

Notice over-correction of phasing and prephasing
- clue that peak is not due to sequence

Second peak "G" has same intensity as rest of sequence

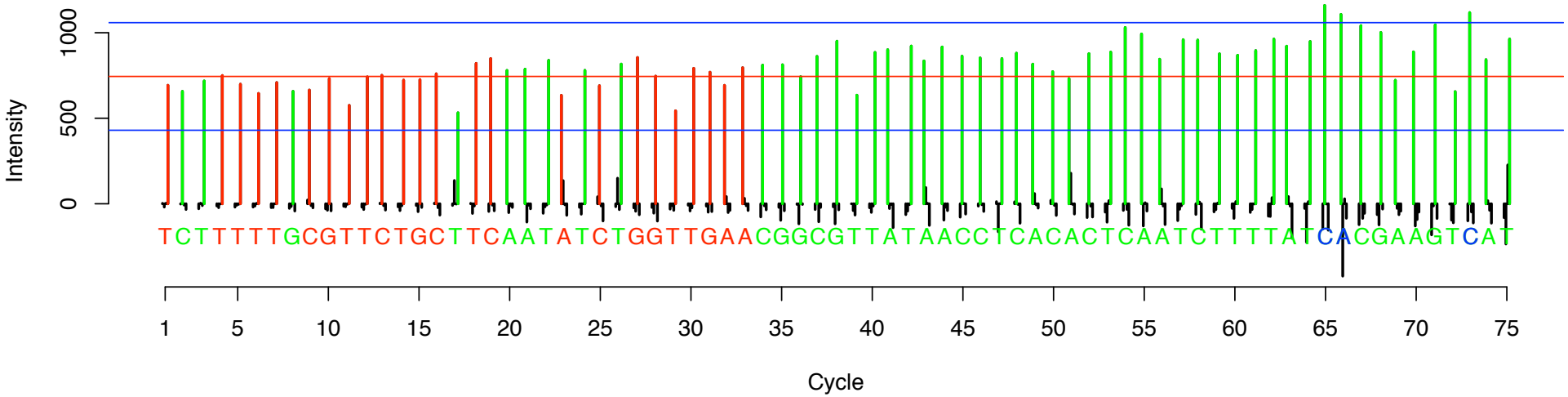
Might be possible to correct read (but probably not worth it)



Fragment ligation

Two fragments of DNA can ligate before sequencing

- Apparently good read
- High error rate
- Rare




```
TCTTTTTGCGTTCTGCTTCAATATCTGGTTGAACGGCGTCGCGTCGTAACCCAGCTTGGTAAGTTGGATTAAGCA PhiX 5190 -ve
|||||||||||||||||||||||||||||||||||||||||||        | || |    |   |   |
TCTTTTTGCGTTCTGCTTCAATATCTGGTTGAACGGCGTTATAACCTCACACTCAATCTTTTATCACGAAGTCAT Read
 ||    |         |  ||| | |       ||||||||||||||||||||||||||||||||||||||||||
CCTCAGCGGCAAAAATTAAAATTTTTACCGCTTCGGCGTTATAACCTCACACTCAATCTTTTATCACGAAGTCAT PhiX 2273 -ve
```

Polymerase slippage

20 base repeat

```
CGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCA   phi-X174
||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CGTCACGTTTATGGTGAACAGTGGATTAAGTTCA                        read  1 to 34
             TGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCA     read 35 to 75
```

Templated sequence

Polymerase slips during replication causing a region to be repeated

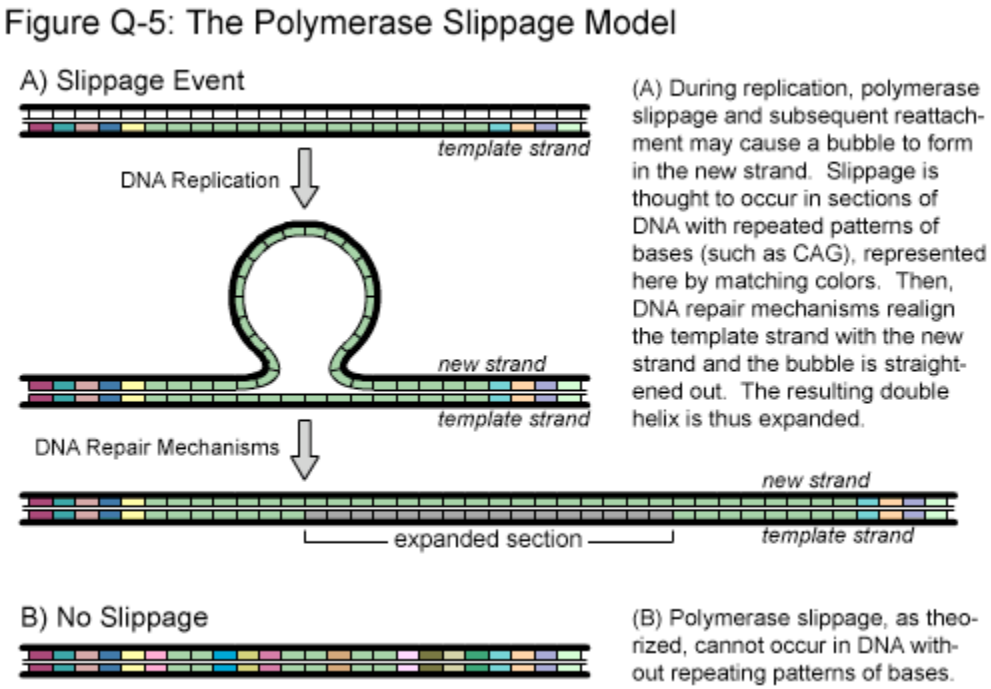
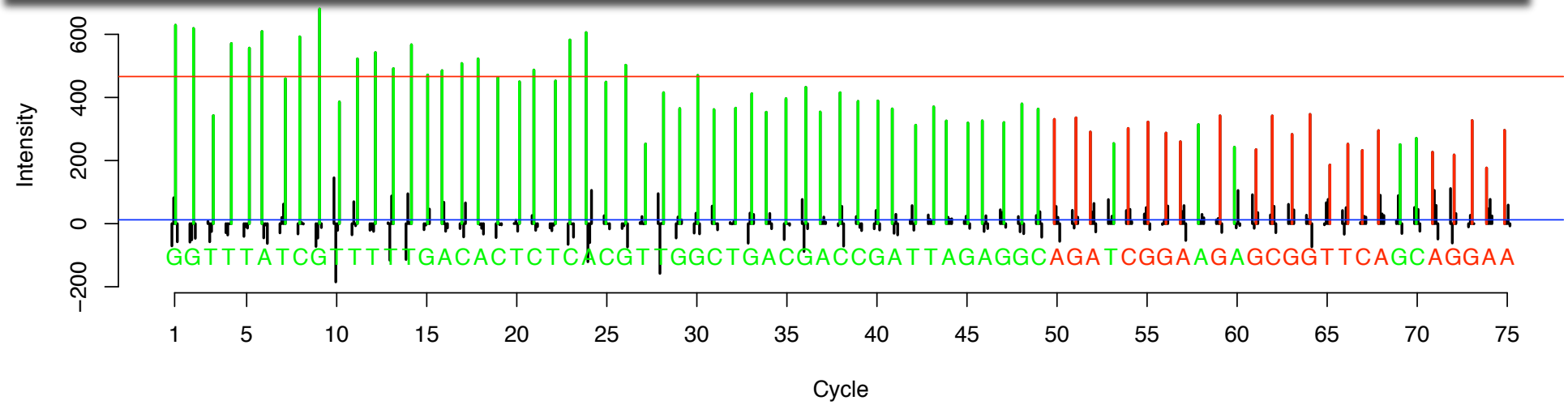


Image: http://hopes.stanford.edu/causes/mutation/q5.html

Adapter Sequence

AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATAT



GGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCAGATCGGAAGAGCGGTTCAGCAGGAA A read

Adapter sequence

||||||||||||||||||||||||||||||||||

AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATAT

A couple of reads

TACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGACTTAGATCGGAAGAGCGGTTCAGC

TCATGAGTCAAGTTACTGAACAATCCGTACGTTTCCAGACCGCTTTGAGATCGGAAGAGCGGTTCAGCAGGAATC

AATATCAGCACCAACAGAAACAACCTGATTAGCGGCGTTGACAGATGTATCCATCTGAAGATCGGAAGAGCGGTT

TCAGAAAGAGATTGCCGAGATGCAAAATGAGACTCAAAAAGAGATTGCTGGAGATCGGAAGAGCGGTTCAGCAGG

AGTAATCACGTTCTTGGTCAATATAACCAGTAGTGTTAACAGTCGGGAAGATCGGAAGAGCGGTTCAGCAGGAAT

TGACTATTCCAATGCAAAAACTGAACGGCCTGGAAACACTGGTCATAATCATGGTGGCGAGATCGGATGAGCGGT

TGAGGTTATAACGCCGAAGCGGTAAAAATTAAAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATAT

AACCAGAACGTGAAAAAGCGTCCTGCGTGTAGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATC

GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATCGTA

ATTCAGTCGGCGACTTCACGCCAGAATACGAAAGACCAGGTATATGCACAAAAGATCGGAAGAGCGGTTCAGCAG

Note: Adapter sequence derived from study of Sanger Institute reads, yours may differ

A crude method to locate adapters

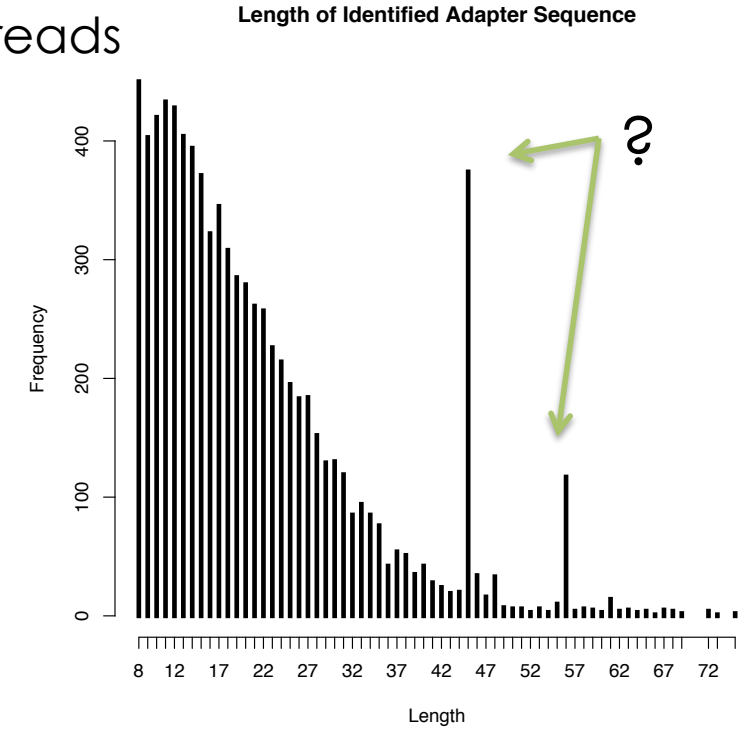
Search for read tails
- Starting with AGAT
- >90% ID with adapter sequence
- Length at least 8 bases

Best ungapped hit of adapter to phiX

```
PhiX      AGAACGAGAAGACGGTTACGCAGTTTTGCCG
          |||.||..|...||||||..||||...|||||
Adapter   AGATCGGAAGAGCGGTTCAGCAGGAATGCCG
```

75 cycle data: about 0.08% of bases, 0.3% of reads

71% bases miscalled for adapter set
c.f. 6% bases miscalled for non-adapter set



Affect on quality

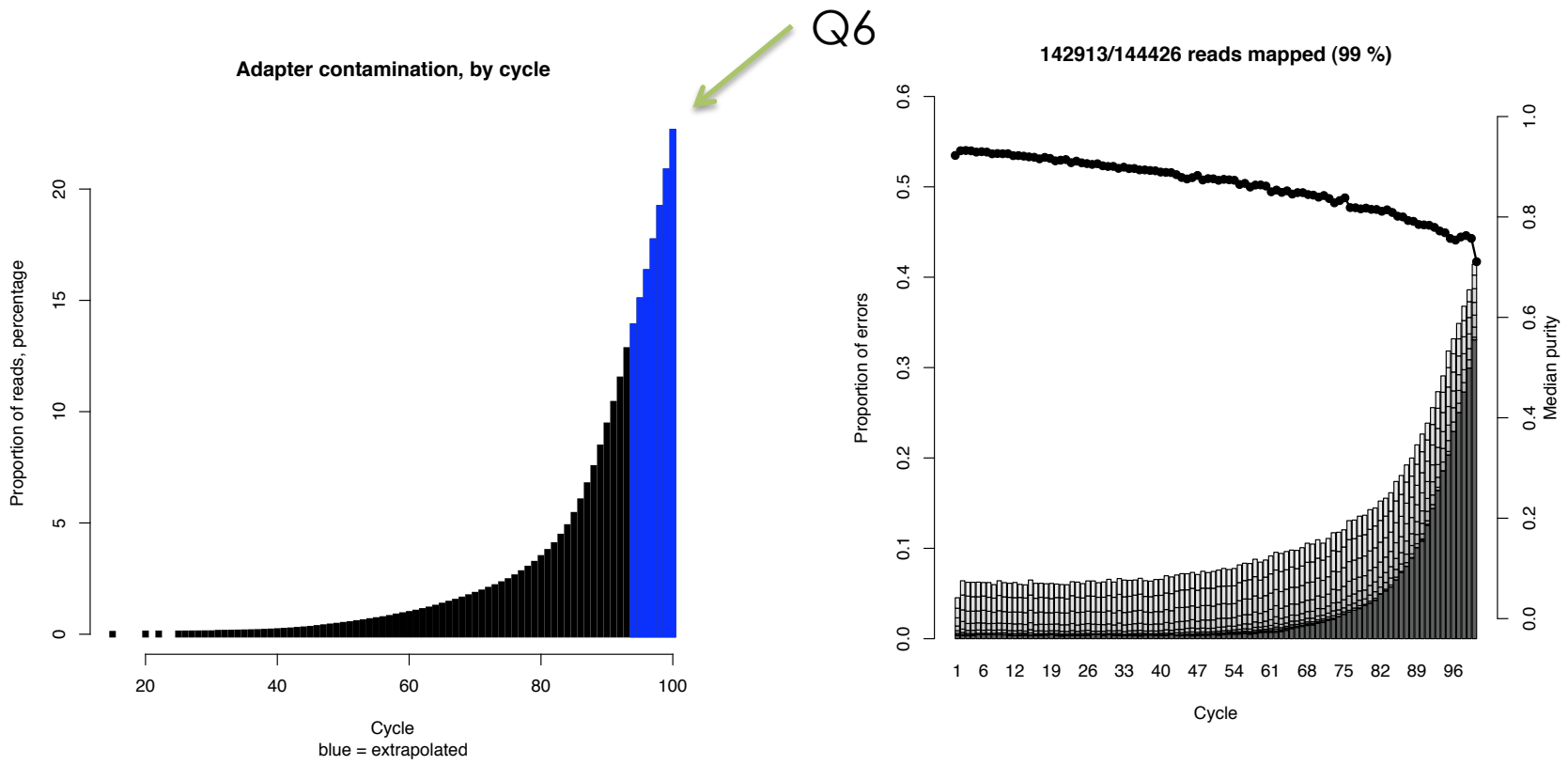
44 cycles	Q40	1 in 10,000
75 cycles	Q31	8 in 10,000
100 cycles	Q16	25 in 1,000

Average. Worse as cycle number increases

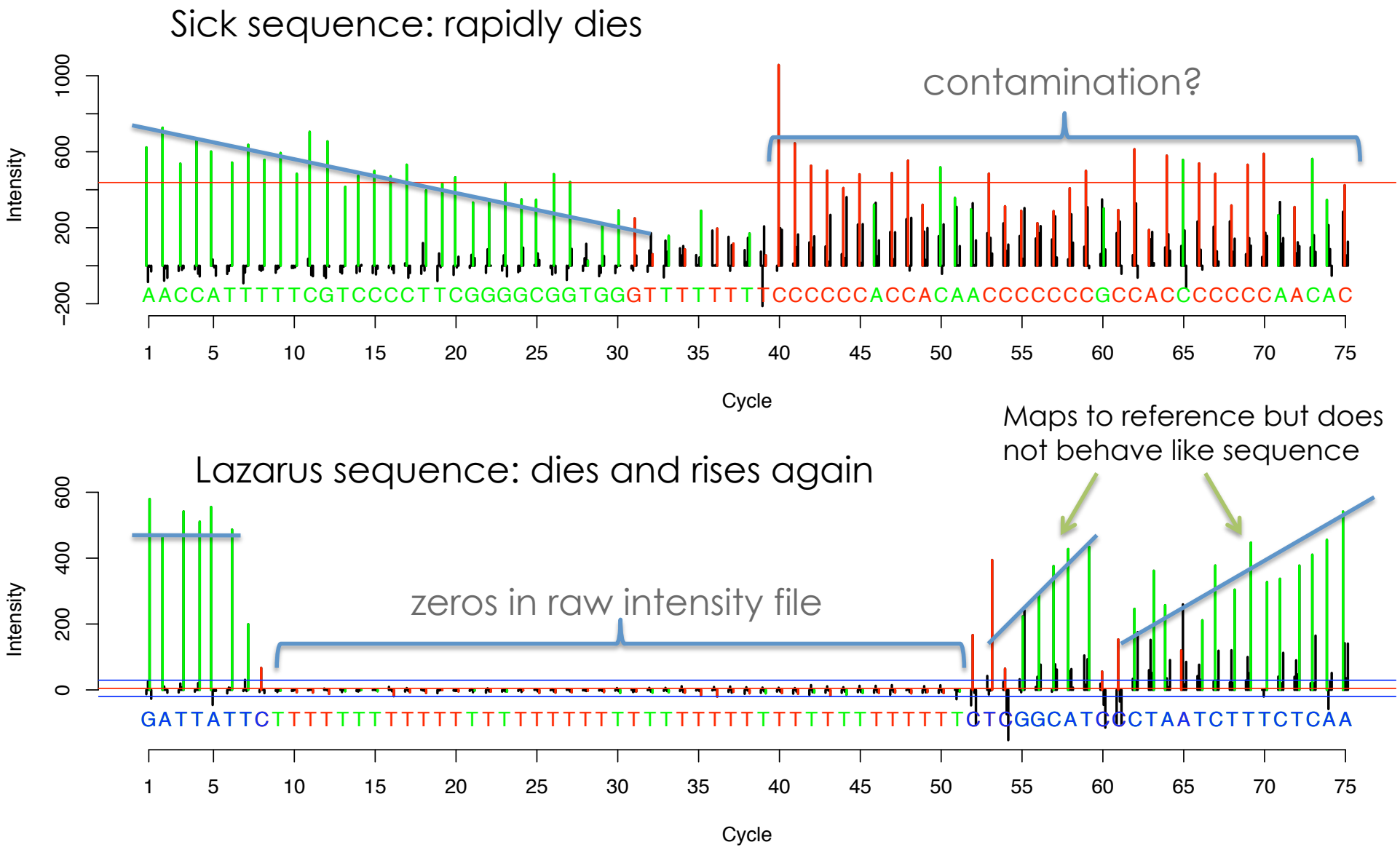
100 cycle data

Adapter contamination is accumulative
- starts rare but total effect can be large

Extrapolate number of adapters to final 7 sites
- ~ 45% of final cycle errors are due adapter sequence
- median purity still high; missing other effects?



Other attractions in the sequence zoo



Error frequencies

Manual look at all errors in 27 tiles of high quality sequence (Q34 bases)
Rates in Qphred

	Good read	Bad read	Indel	Adapter	Ligation	Unknown
Rate	36.9	40.1	47.7	46.3	47.0	45.6
Lower	36.5	39.5	46.3	45.1	45.7	44.5
Upper	37.4	40.8	49.4	47.8	48.5	46.9

Good read: only error is high quality base.
Bad read: otherwise messy, several errors.
Indel: presence of insertion of deletion
Adapter: undetected adapter sequence (after filtering)
Ligation: strong evidence of ligation

Error frequencies

Manual look at all errors in 27 tiles of high quality sequence (Q34 bases)
Rates in Qphred

	Good read	Bad read	Indel	Adapter	Ligation	Unknown
Rate	36.9	40.1	47.7	46.3	47.0	45.6
Lower	36.5	39.5	46.3	45.1	45.7	44.5
Upper	37.4	40.8	49.4	47.8	48.5	46.9

$$\min(Q_{\text{Base}}, Q_{\text{Effect}}) \geq Q_{\text{Actual}} \geq \min(Q_{\text{Base}}, Q_{\text{Effect}}) - 3$$

Q_{34}

$Q_{37} \geq$ $\quad$ $Q_{37-3} \geq$

Implementation and availability

http://www.ebi.ac.uk/goldman/AYB/

Written in R
Licensed under GPL (version 3)
Plug-in replacement for Bustard
• Single change to Makefile

Change this line →

GAPipeline (Illumina) v. 0.3 Makefile

Makefile

```
# Makefile auto-generated by makefile.py (version 1.9.2)
# 26-09-08 15:44:32 CEST

installationdir:=/home/solexa/pipeline/bin/GAPipeline-0.3.0/Goat
#basecall_exe:=$(installationdir)/run_base_caller
basecall_exe:=/home/solexa/pipeline/bin/AYB/run_ayb.sh
phasing1_exe:=$(installationdir)/Phasing/phasing_estimate
phasing2_exe:=$(installationdir)/Phasing/phasing_summary.py
phasing3_exe:=$(installationdir)/Phasing/select_phasing.py

#analysisdir:=/data/ValidationData/
Solexa_Validation_Dataset_v_0_1/061212_SLXA-B7_0037_FC2818/Data/
C1-26_Firecrest1.9.2_26-09-2008_solexa.2
analysisdir:=..
sequencedir:=/data/ValidationData/
Solexa_Validation_Dataset_v_0_1/061212_SLXA-B7_0037_FC2818/Data/
C1-26_Firecrest1.9.2_26-09-2008_solexa.2/Bustard1.9.2_26-09-2008_solexa
sample1:=s
sample2:=s
sample3:=s
sample4:=s
sample5:=s
sample6:=s
sample7:=s
sample8:=s
analysis_subfolder:=Firecrest
sequence_subfolder:=Bustard
lanes:=$(sample1)_1 $(sample2)_2 $(sample3)_3 $(sample4)_4 $(sample5)_5 $
(sample6)_6 $(sample7)_7 $(sample8)_8
analysisdir1:=$(analysis_subfolder)/L001
```

Acceptable performance
2 hours per lane on an 8 core machine(~128 CPU hours per run)

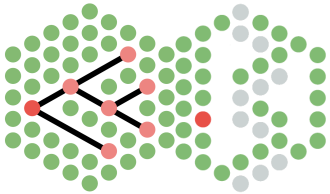
• Faster if phasing and cross-talk assumed to constant (ala Bustard)
• Due to be rewritten with focus on performance and reliability

Thanks to:
Jonathon Blake, EMBL Genomics Core Facilities

Thanks

Thanks to:

- EBI:
  - Ewan Birney
  - Paul Flicek (1000 Genomes Project)
- EMBL Genomics Core Facilities, Heidelberg:
  - Vladimir Benes
  - Jonathon Blake
- Sanger Institute:
  - Nava Whiteford (now at Oxford Nanopore Technologies)
  - Tom Skelly
  - Irini Abnizova
- Illumina:
  - Tony Cox
- CRUK Cambridge Research Institute:
  - Gordon Brown
  - Kevin Howe
- Cambridge Institute for Medical Research:
  - Vincent Plagnol
- Institute of Cell and Molecular Science, Queen Mary University of London:
  - David Van Heel



君達の基地は、全て