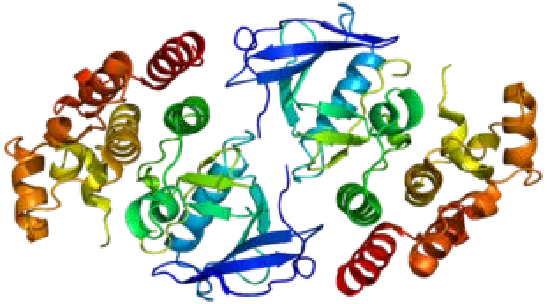
Automatic Analysis of Dual-Channel Droplet Digital PCR Experiments to Detect BRAF-V600 Mutations

Dean Attali

http://deanattali.com

Jennifer Bryan Lab @ MSL, UBC



MSc exit seminar

April 15, 2016

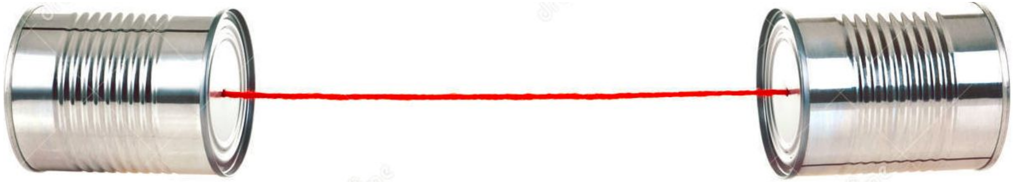
a place of mind

THE UNIVERSITY OF BRITISH COLUMBIA

Summer 2014

Jenny, can you help automate our analysis?

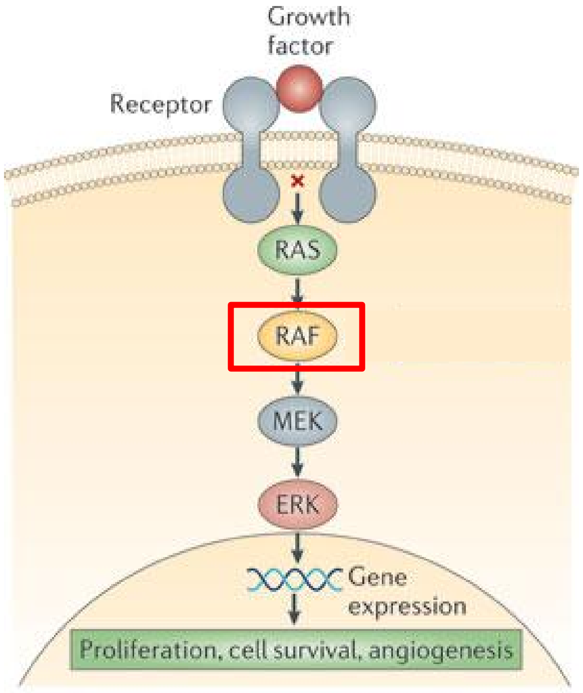
Dean, it’s for you!



BRAF Gene / MAPK Pathway

- B-Raf protein kinase

- Normal conditions:
  Growth factor binds ⇒
  Ras protein activated ⇒
  B-Raf protein activated ⇒
  More phosphorylations ⇒
  Signal for **cell** to **divide**

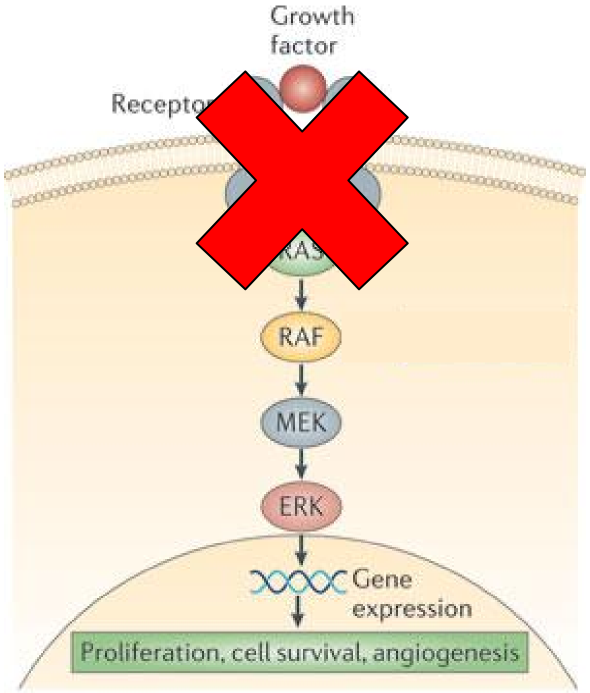


Flaherty et al. 2011.Nature Reviews Drug Discovery 10:811-812

BRAF-V600E Mutation

- V600 **mutation** ⇒ Constitutively active ⇒ **Uncontrolled** cell **growth** ⇒ **Tumour**
- 50% of melanoma tumours, 10% of colorectal cancers

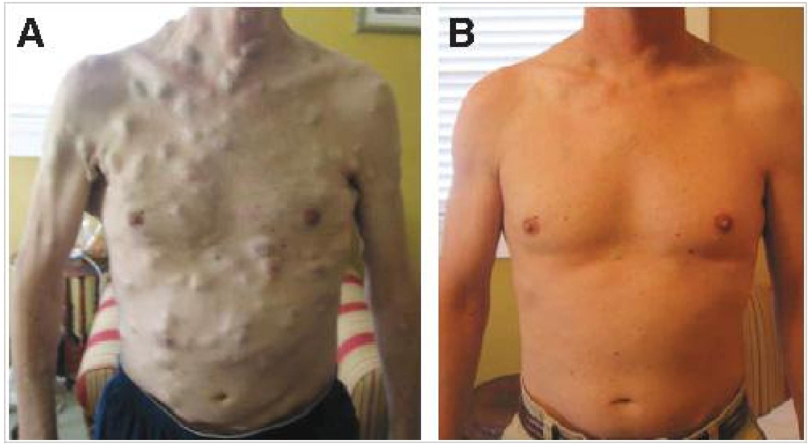
BRAF codon	599			600			601		
Wild type	A	C	A	G	T	G	A	A	A
V600E	A	C	A	G	A	G	A	A	A
V600K	A	C	A	A	A	G	A	A	A
V600D	A	C	A	G	A	T	A	A	A
V600R	A	C	A	A	G	G	A	A	A
V600G	A	C	A	G	G	G	A	A	A
V600M	A	C	A	A	T	G	A	A	A



Flaherty et al. 2011.Nature Reviews Drug Discovery 10:811-812

BRAF-V600 Mutation Tests

- Presence/absence of MT-*BRAF* affects treatment
  - Melanoma patients with mutation can take vemurafenib - a BRAF inhibitor



Wagle et al. “Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling”

BRAF-V600 Mutation Tests

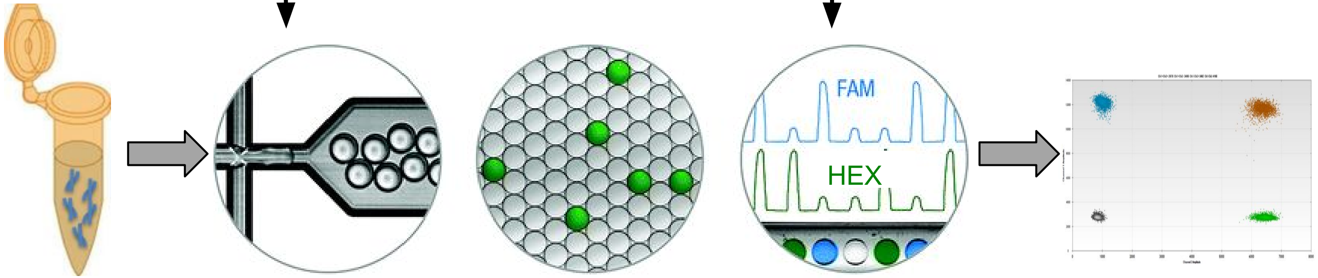
- Cobas® 4800 BRAF V600 Mutation Test (Roche)
- Only looks for V600E
- Requires > 5% mutation level
- Output is qualitative: yes/no

Cobas® 4800
BRAF
V600 Mutation Test

Droplet Digital PCR (ddPCR)

DNA randomly partitioned into 20,000 tiny droplets

Amplified droplets emit specific fluorescence
Fluorescence in two channels measured for each drop

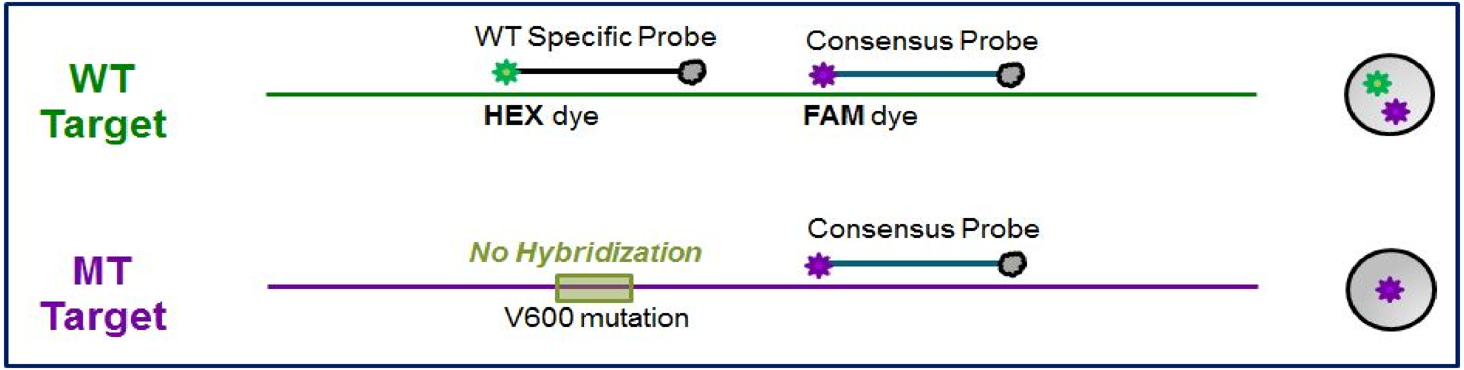


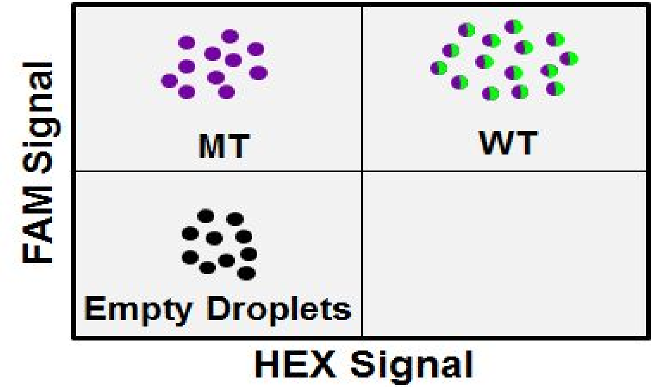
1. **Load sample**
2. **Droplet Generation**
3. **PCR Cycling**
4. **Droplet Reading**
5. **2D plot with 20,000 points**

20 ng (sample DNA, primers, fluorescent probes, PCR mix)

Droplets with target get amplified, droplets without target are not amplified

BRAF-V600 Mutation ddPCR Assay





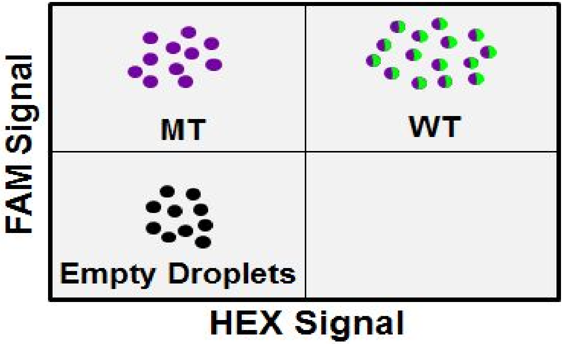
$$\textit{BRAF}\text{ mutation frequency} = \frac{\#\text{ MT droplets}}{\#\text{ MT droplets} + \#\text{ WT droplets}}$$

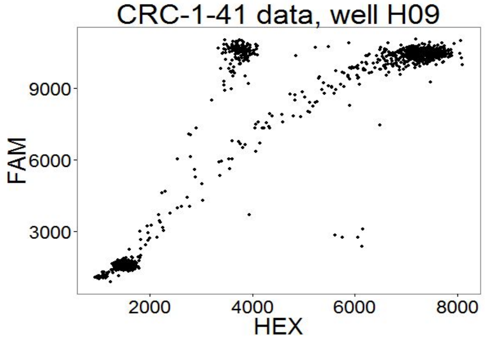
BRAF-V600 Mutation ddPCR Assay

Expectation vs Reality

Big Mac

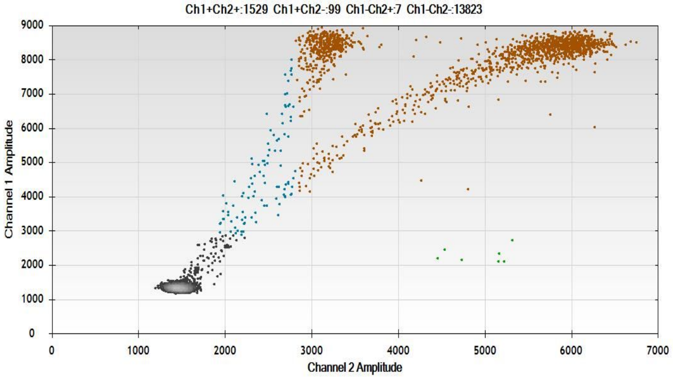
ddPCR

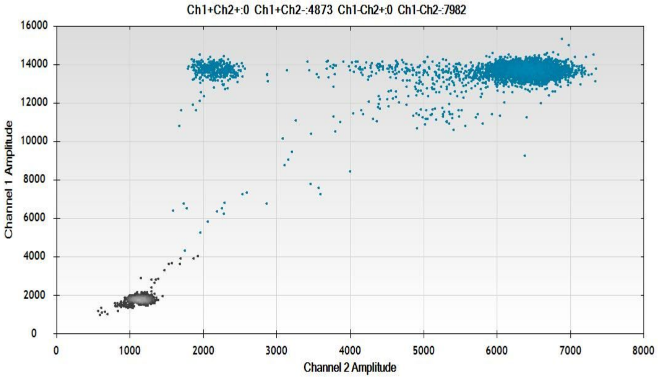




Gating ddPCR Data

- Usually done manually
- QuantaSoft (official analysis software of ddPCR) has auto gating
  - Often wildly inaccurate
- Two tools developed for automatic analysis
  - Both work on single-channel data only
  - Both rely on representative control samples





Gating ddPCR Data

Beaver, Julia A., et al. "Detection of cancer DNA in plasma of patients with early-stage breast cancer." *Clinical Cancer Research* 20.10 (2014): 2643-2650.

> “Droplets were scored as positive or negative based upon their fluorescence intensity which was **determined by gating a threshold** using positive and negative controls as well as no template controls”

Roberts, Chrissy H., et al. "Killer-cell Immunoglobulin-like Receptor gene linkage and copy number variation analysis by droplet digital PCR." *Genome Med* 6 (2014): 20.

> “**Crosshair gating was used** to split the data into four quadrants”

Pretto, Dalyir, et al. "Screening Newborn Blood Spots for 22q11. 2 Deletion Syndrome Using Multiplex Droplet Digital PCR." *Clinical chemistry* 61.1 (2015): 182-190.

“The QuantaSoft software (version 1.4.0.99) includes a **freedraw tool** that enables proper classification of the multiple clusters”

Taly, Valerie, et al. "Multiplex picodroplet digital PCR to detect KRAS mutations in circulating DNA from the plasma of colorectal cancer patients."*Clinical chemistry* 59.12 (2013): 1722-1731.

> “The sizes and locations of the wild-type gate and the mutant gate(s) were **established by manual selection** of the area containing wild-type or mutant clusters”

Milbury, Coren A., et al. "Determining lower limits of detection of digital PCR assays for cancer-related gene mutations." *Biomolecular Detection and Quantification* 1.1 (2014): 8-22.

> “**Objective automated gating** of droplet event clusters is likely **necessary for dPCR practitioners** to take advantage of the full potential sensitivity of the technology for routine applications”

Goal 1 - Gating ddPCR Automatically

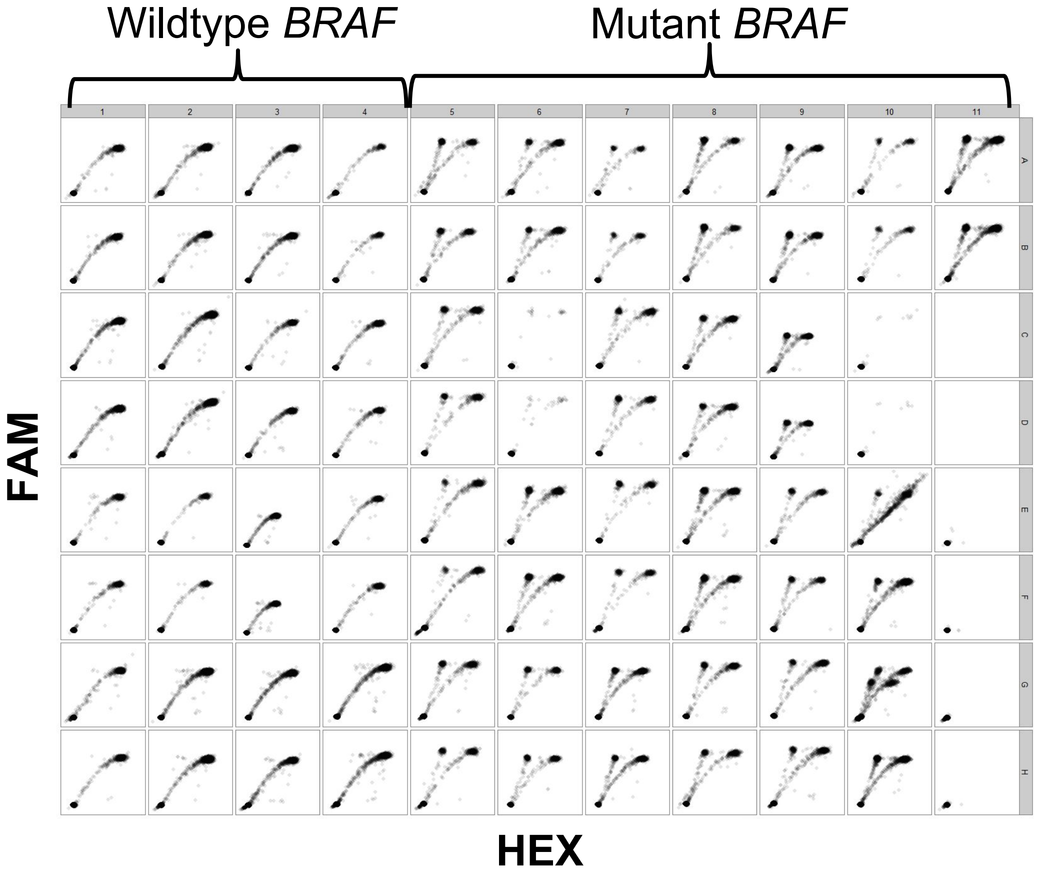
- Given ddPCR output ⇒ calculate mut*BRAF* frequency
- Objective
- Reproducible
- Better gating than QuantaSoft
- No such tools exist

Goal 2 - Make it easily accessible

- Make R package
  - For people comfortable with R
  - *ddpcr* (on CRAN)

- Make web application with visual UI
  - For people who want a point-n-click interface
  - Uses R package under the hood
  - http://daattali.com/shiny/ddpcr

Dataset: FFPE from 41 CRC Patients

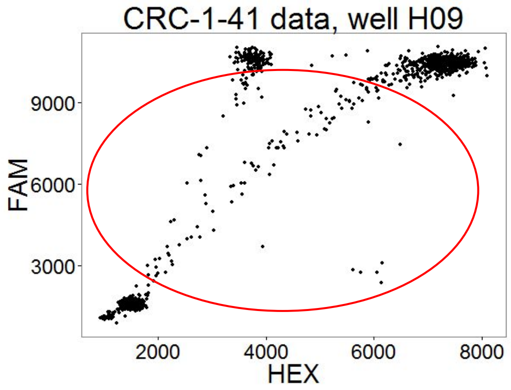


Wildtype *BRAF*

Mutant *BRAF*

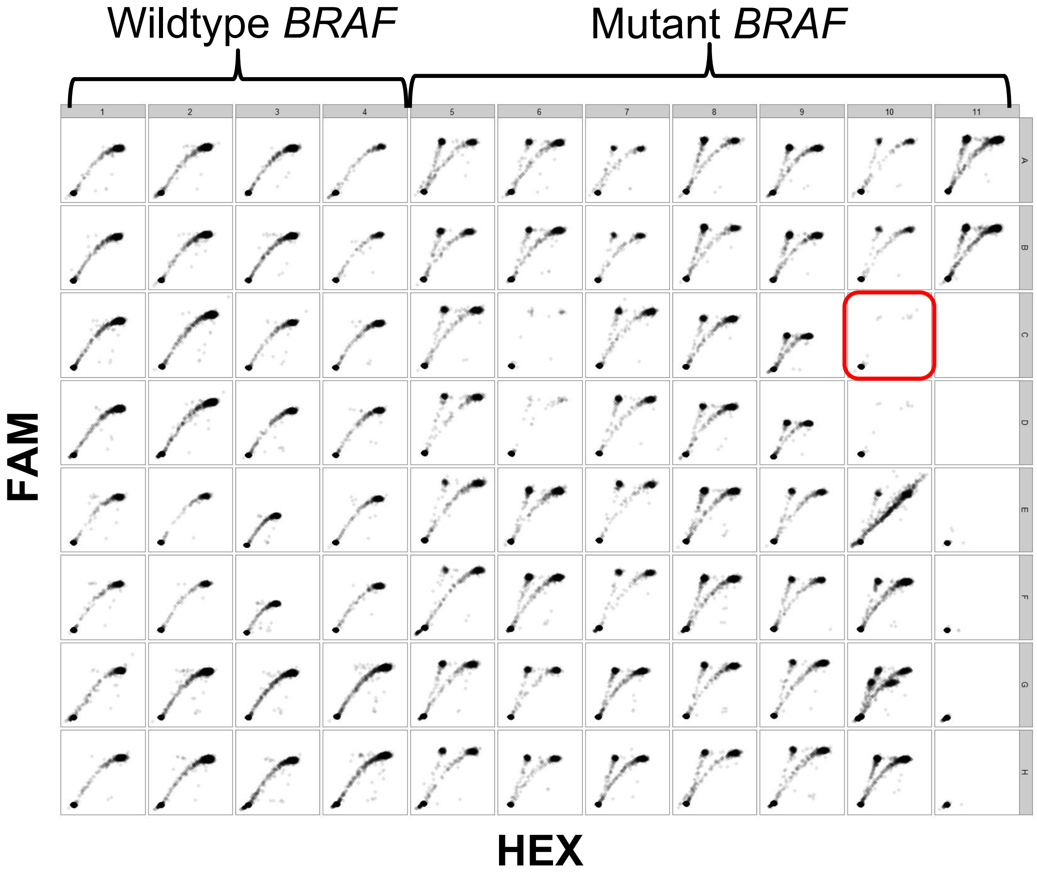
Factors to consider:

- Rain



CRC-1-41 data, well H09

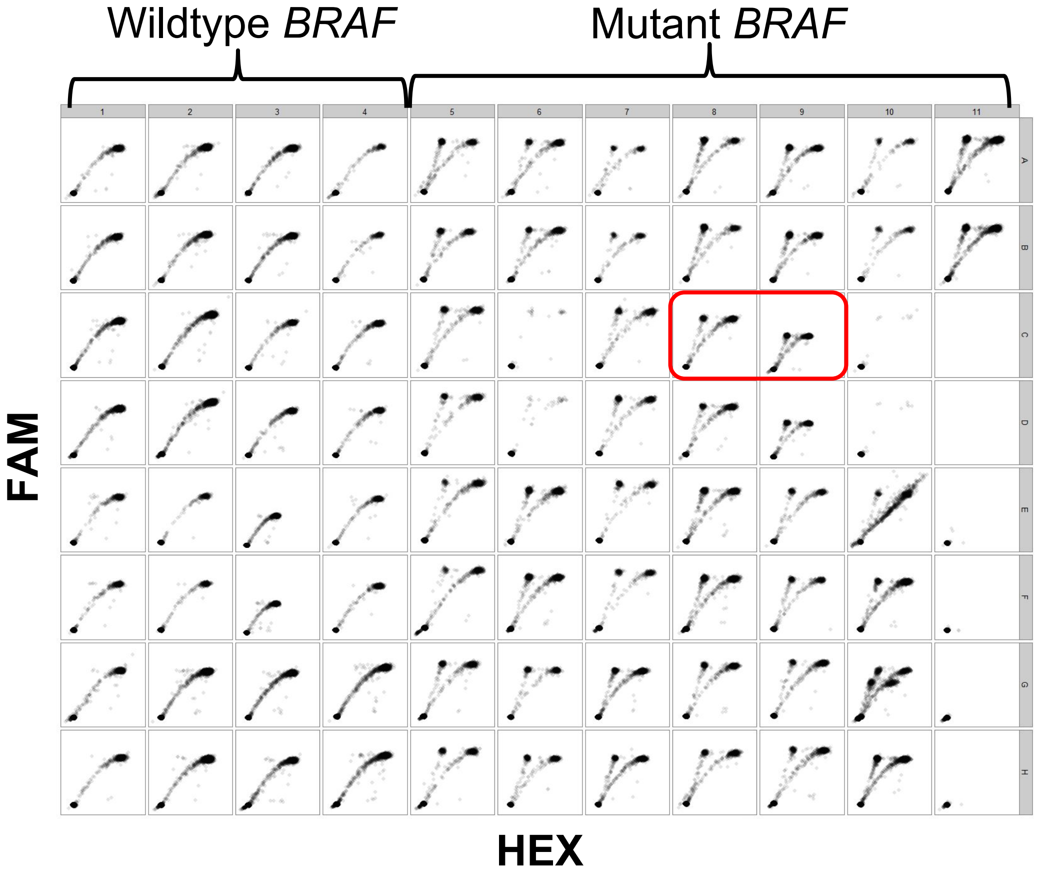
Dataset: FFPE from 41 CRC Patients



Factors to consider:

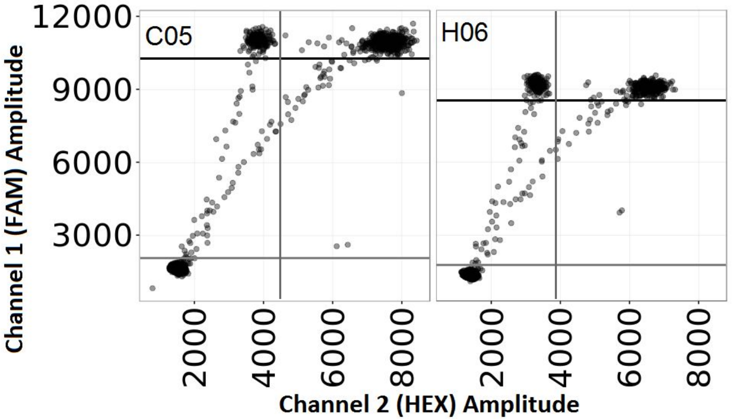
- Failed runs (e.g. C10)

Dataset: FFPE from 41 CRC Patients

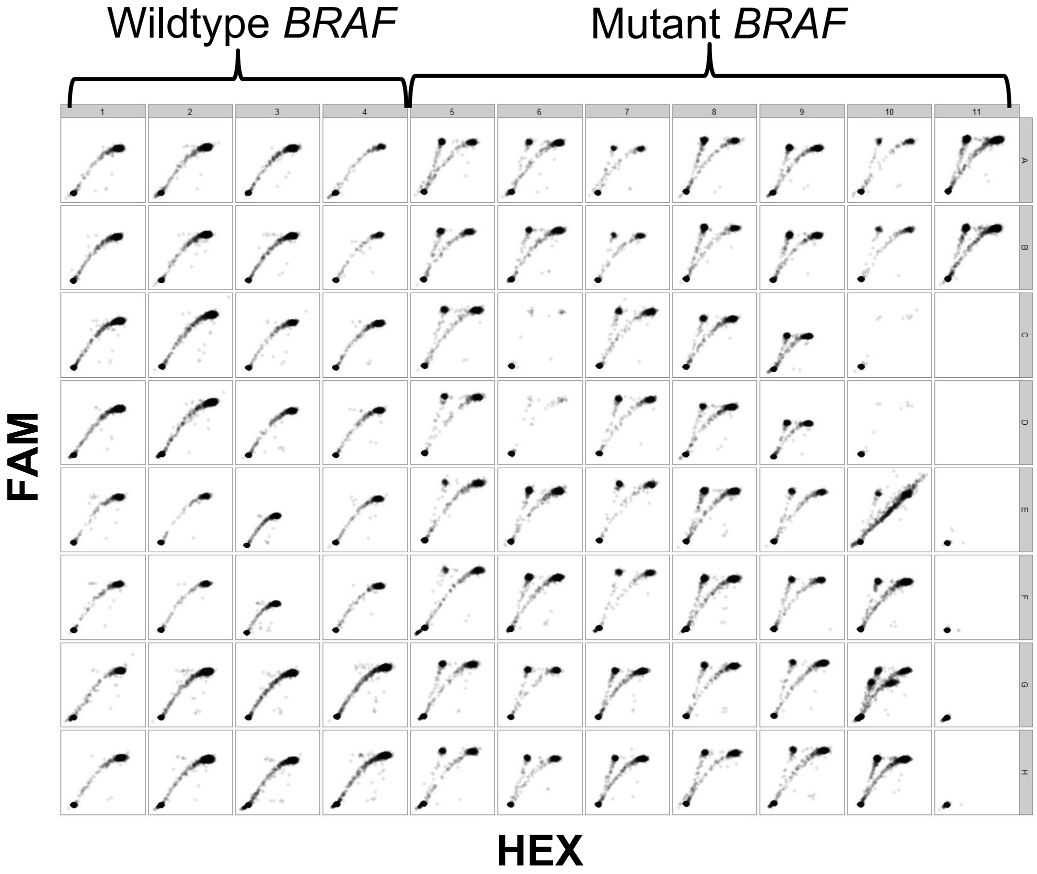


Factors to consider:

- Can’t use same thresholds globally (e.g. C08 vs C09)

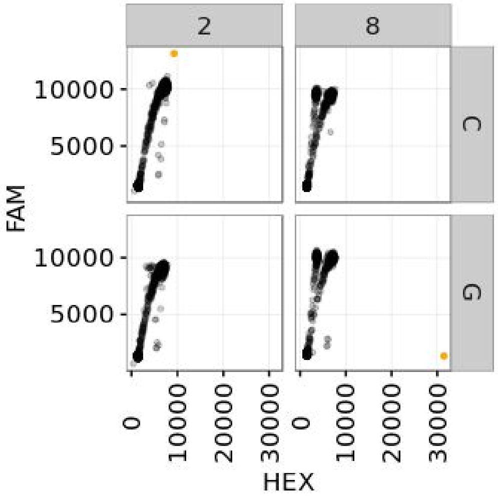


Dataset: FFPE from 41 CRC Patients

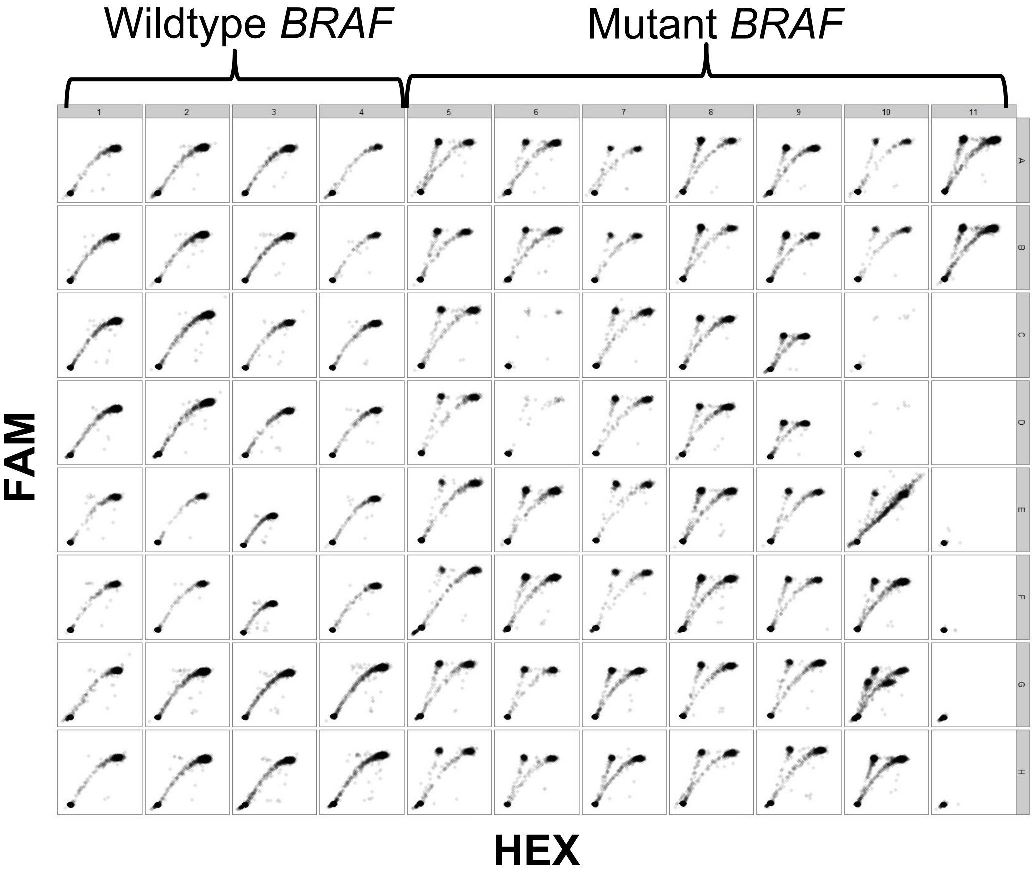


Factors to consider:

- Outliers



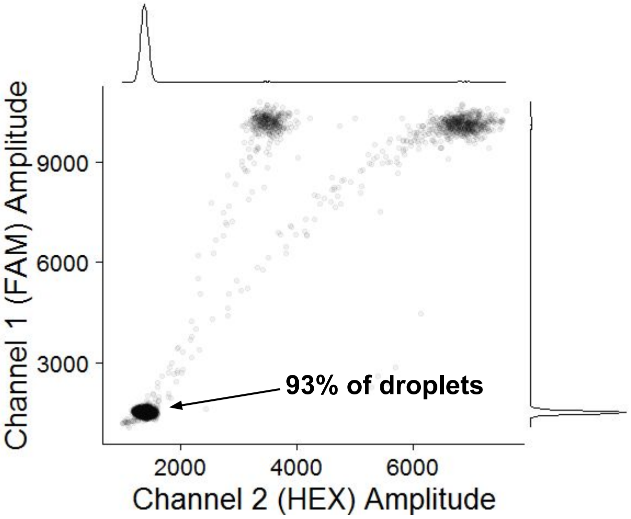
Dataset: FFPE from 41 CRC Patients



Factors to consider:

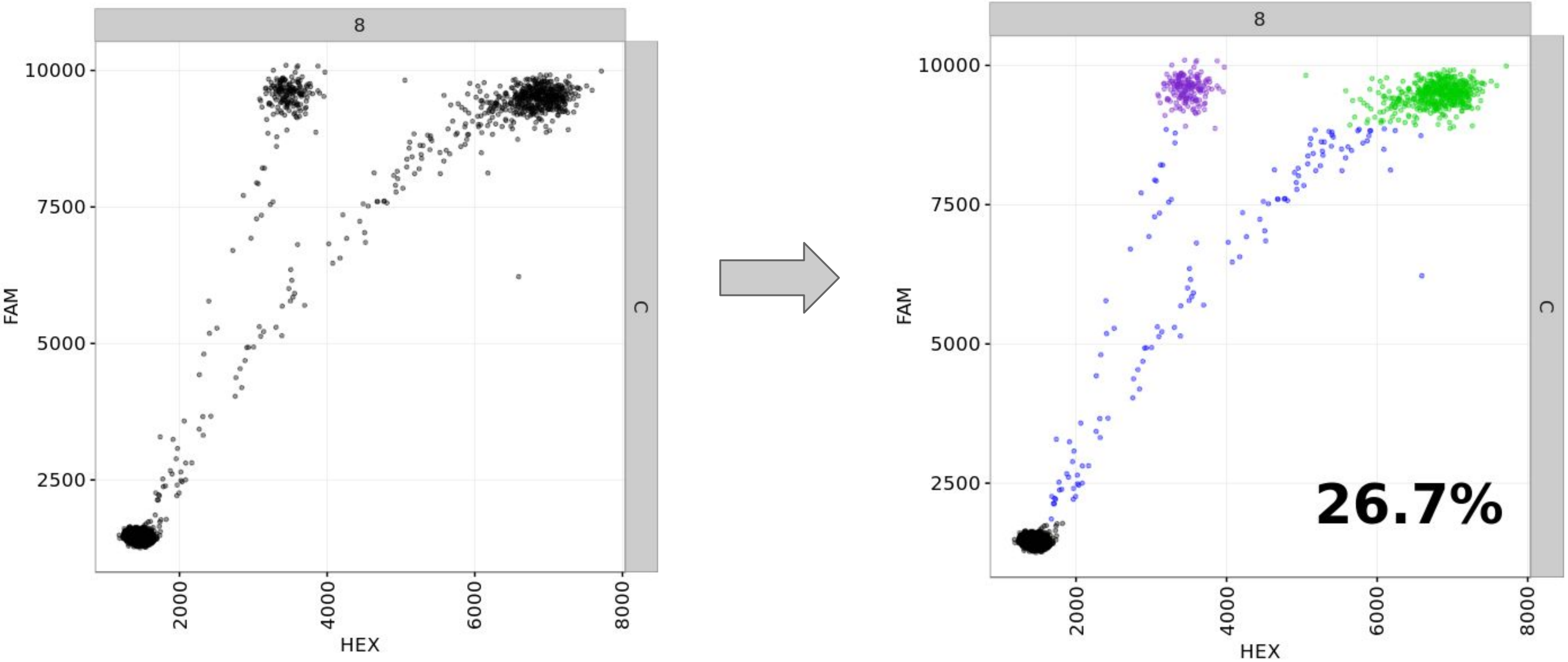
- Most droplets are empty

F06; alpha = 0.05



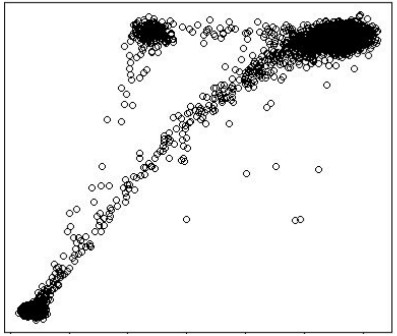
Goal

C08

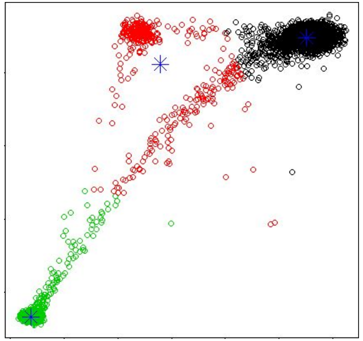

8
10000
7500
5000
2500
FAM
2000
4000
6000
8000
HEX
C
8
10000
7500
5000
2500
FAM
26.7%
2000
4000
6000
8000
HEX
C

First try: off-the-shelf clustering algo’s

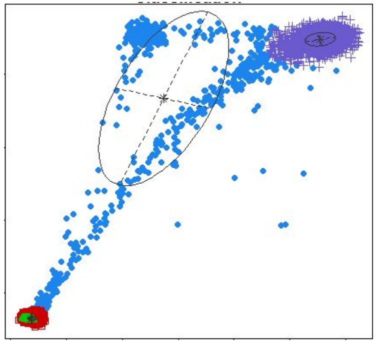
Raw data



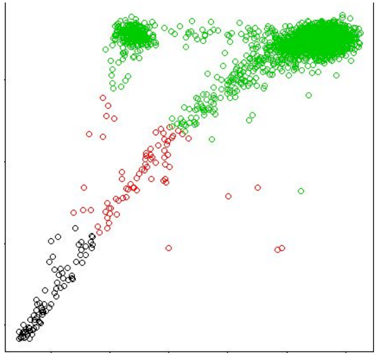
K-means



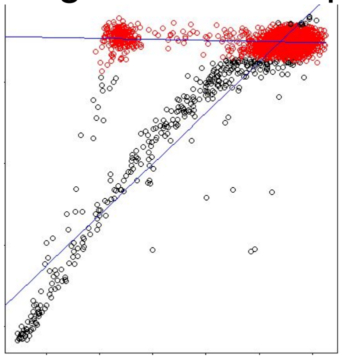
GMM



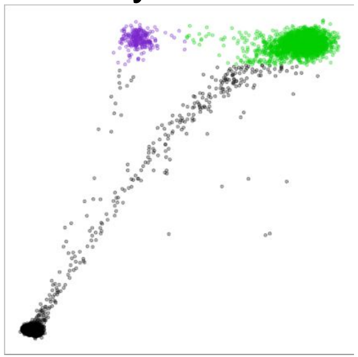
Hierarchical clustering



Finding two linear equations



My tool

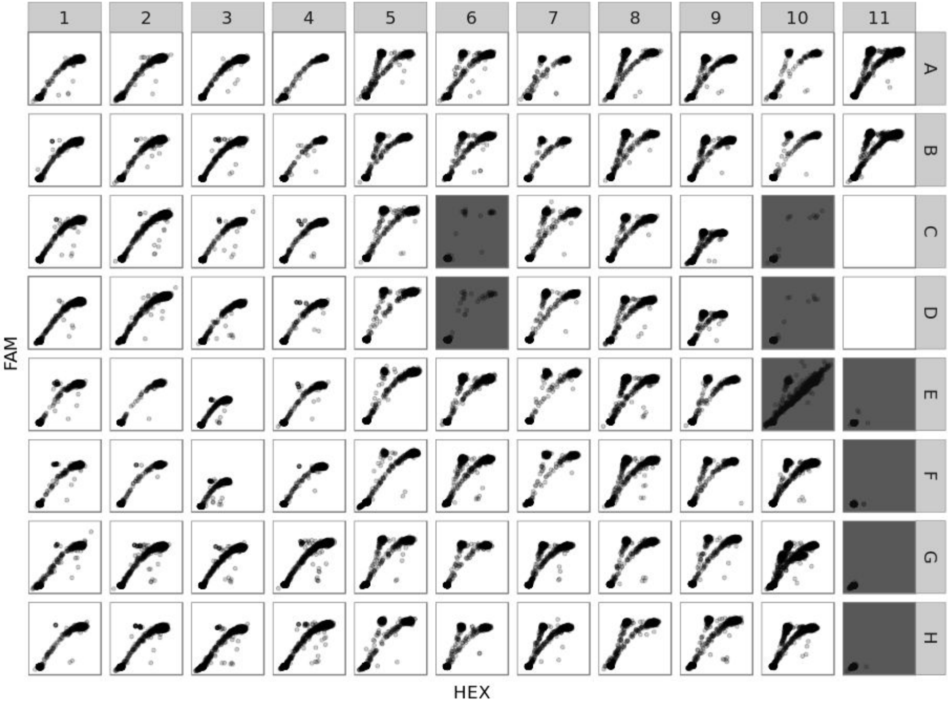


General Pipeline

1. Identify failed experiments
2. Identify outlier droplets
3. Identify empty droplets
4. Gate droplets (rain vs mutant vs wild type)
5. Classify each sample as mutant or wild type
6. (Revisit gating of wild type samples)

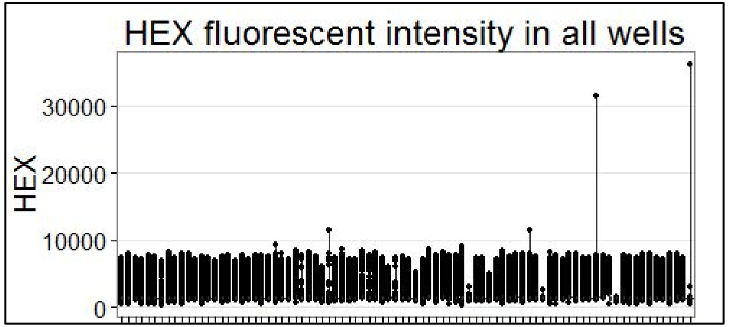
Step 1: Identify failed experiments

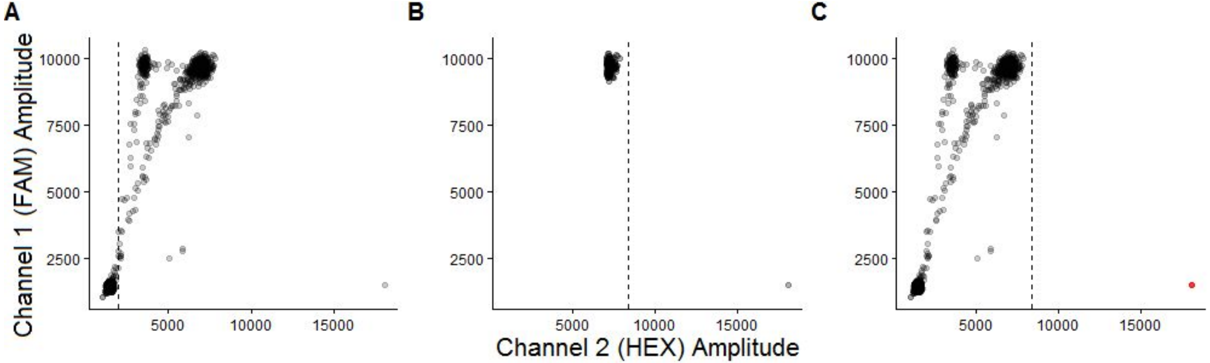
- Use QC metrics to ensure enough data in well



Step 2: Identify outlier droplets

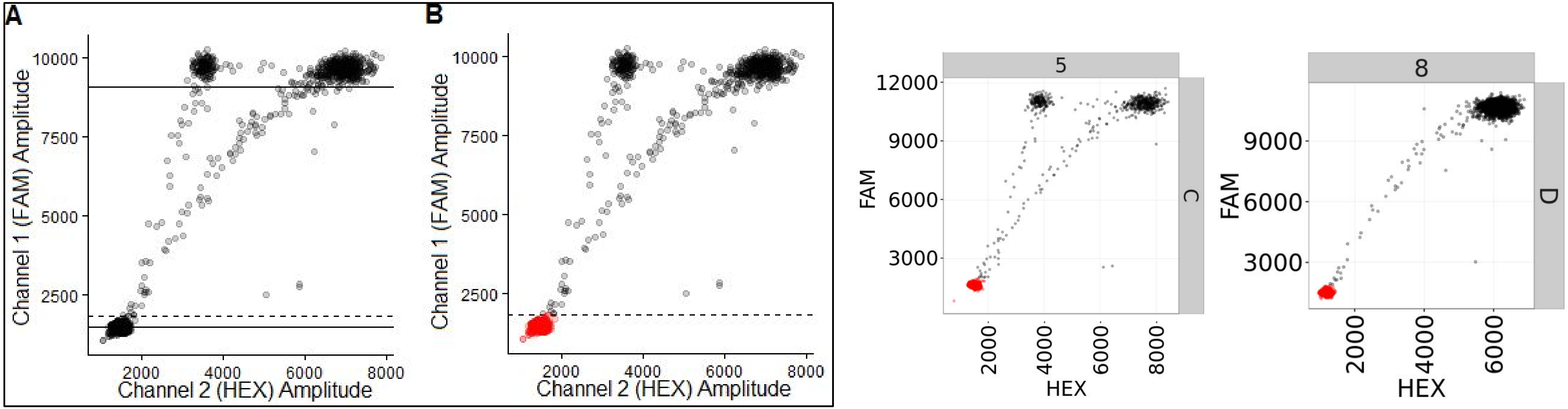
Take top x% of droplets,
define threshold as Q3 + 5IQR





Step 3: Identify empty droplets

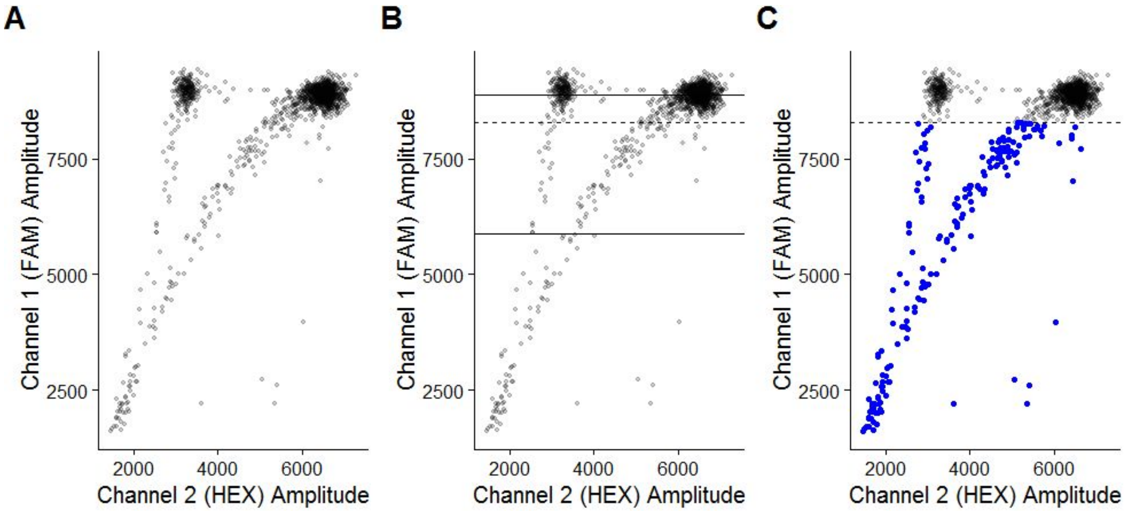
Fit two-component Gaussian mixture model to FAM values → Lower population is empty droplets, threshold = $\mu + 5\sigma$



Step 4: Gate droplets (rain/MT/WT)

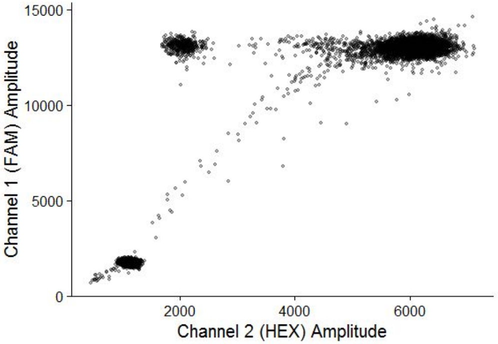
First substep: remove the rain

Fit two-component GMM to FAM, threshold = μ - 3σ

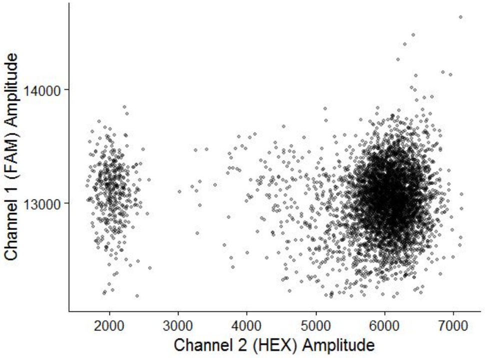


Step 4: Gate droplets (rain/MT/WT)

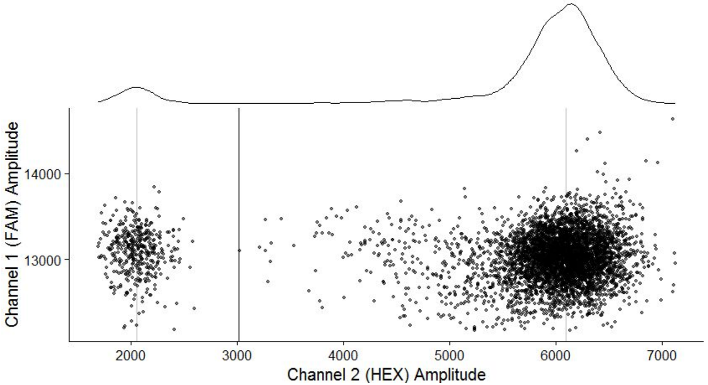
1. Raw data

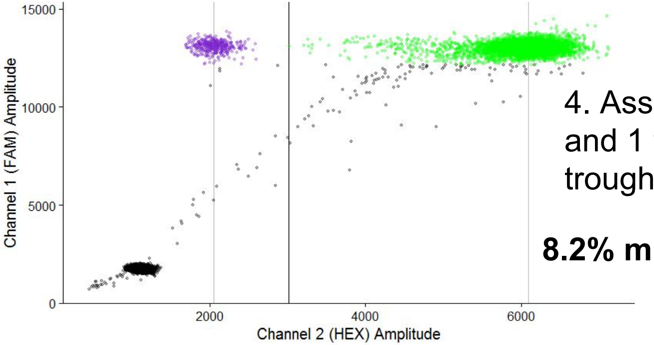


2. Remove empty and rain



3. Kernel density estimation (KDE) of HEX values

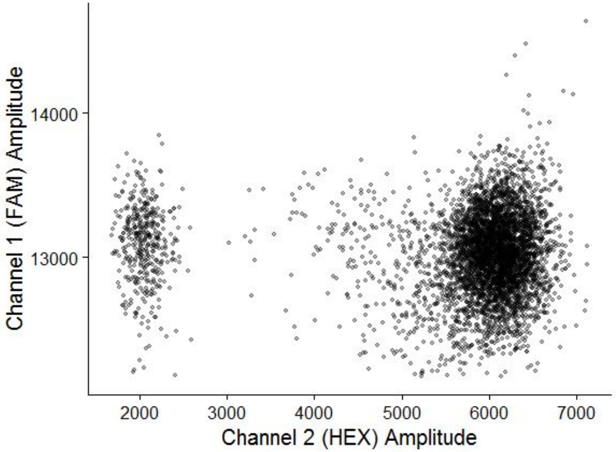




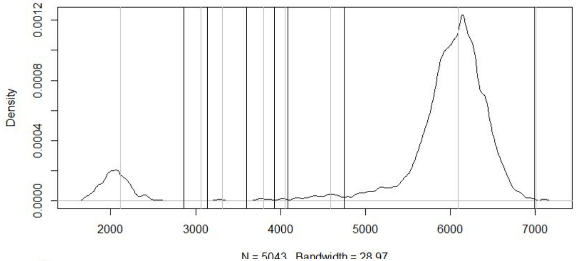
4. Assuming 2 peaks and 1 trough, use trough as gate

8.2% mut*BRAF*

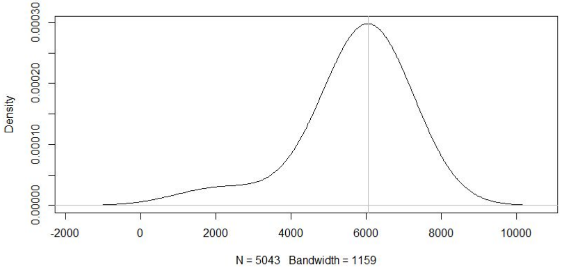
KDE bandwidth selection



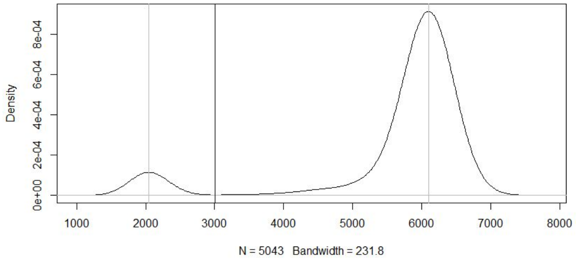
Bandwidth too small



Bandwidth too big

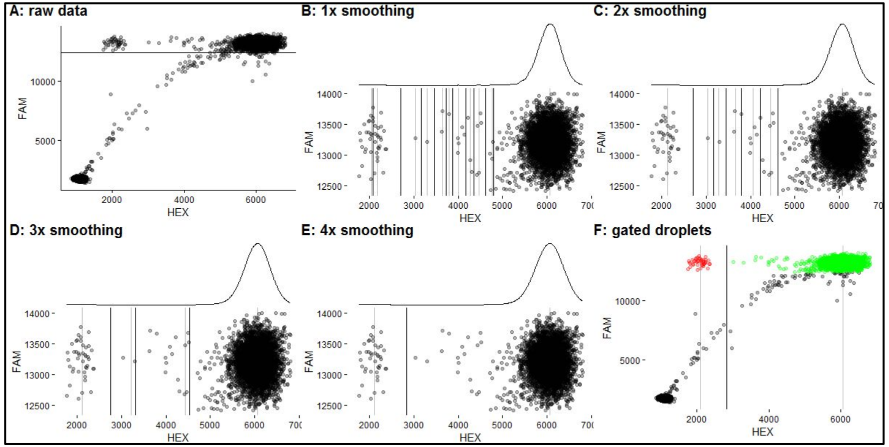


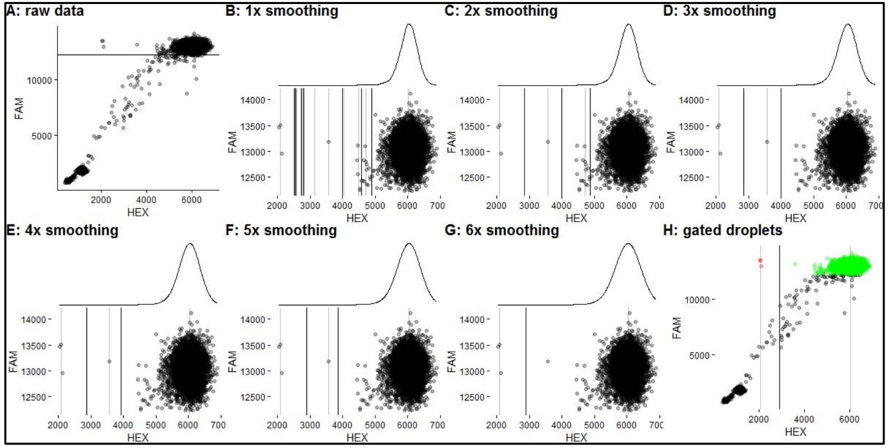
Bandwidth just right



KDE bandwidth selection

Start with low bandwidth, if more than 2 peaks, increase





Step 5: Classify sample as MT/WT

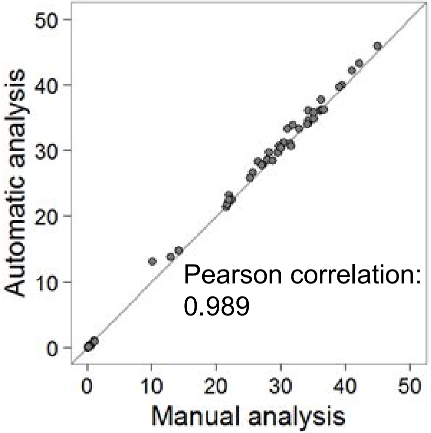
- Mutation frequency statistically significantly > 1% ⇒ Mutant
- Use binomial test: What’s prob. of observing at least N mutant droplets if the true mutant freq is 1%?
- Example: 500 droplets, 7 mutant. H0: freq is < 1%
  Prob observing at least 7 mutants
  = P(X >= 7)
  = 1 - P(X < 7)
  = 1 - [P(X=0)+ P(X=1)+ … + P(X=6)]
  = 0.237
  > pvalue
  Well is classified as WT even though mut*BRAF* = 1.4%

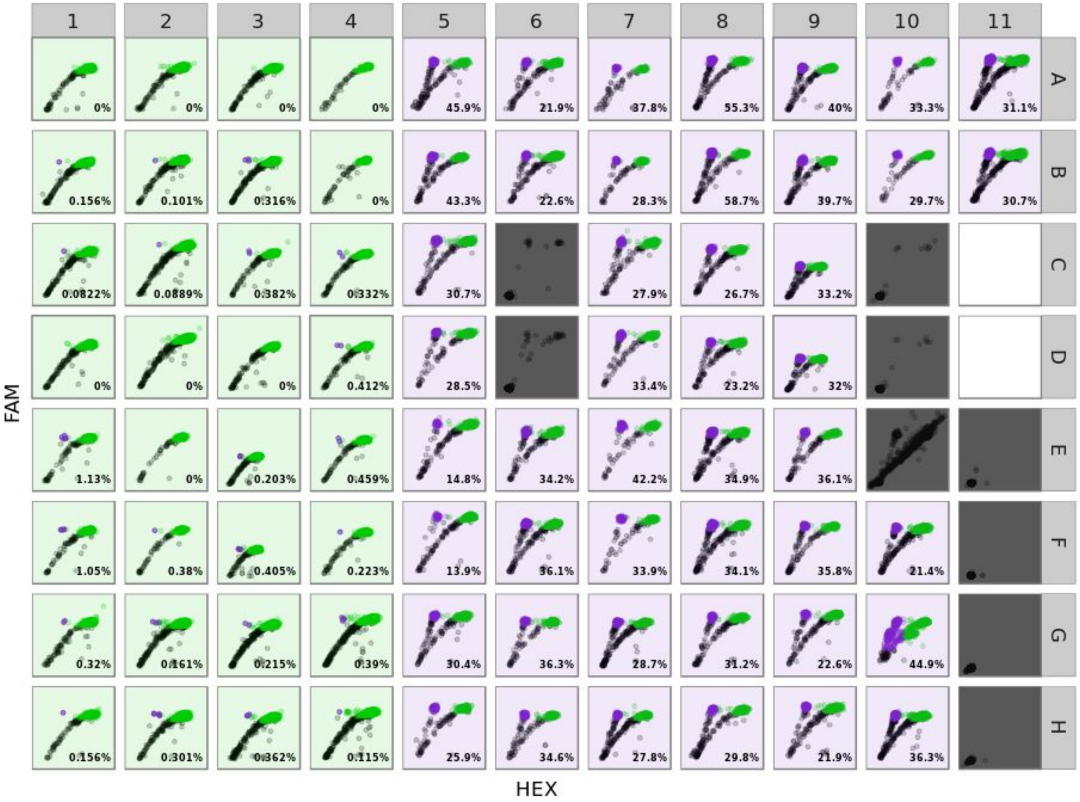
P(X=r)
= ${}_{n}C_{r} \bullet p^{r} \bullet q^{n-r}$

where n = total droplets,
r = mutant droplets,
p = 0.01 (1%)

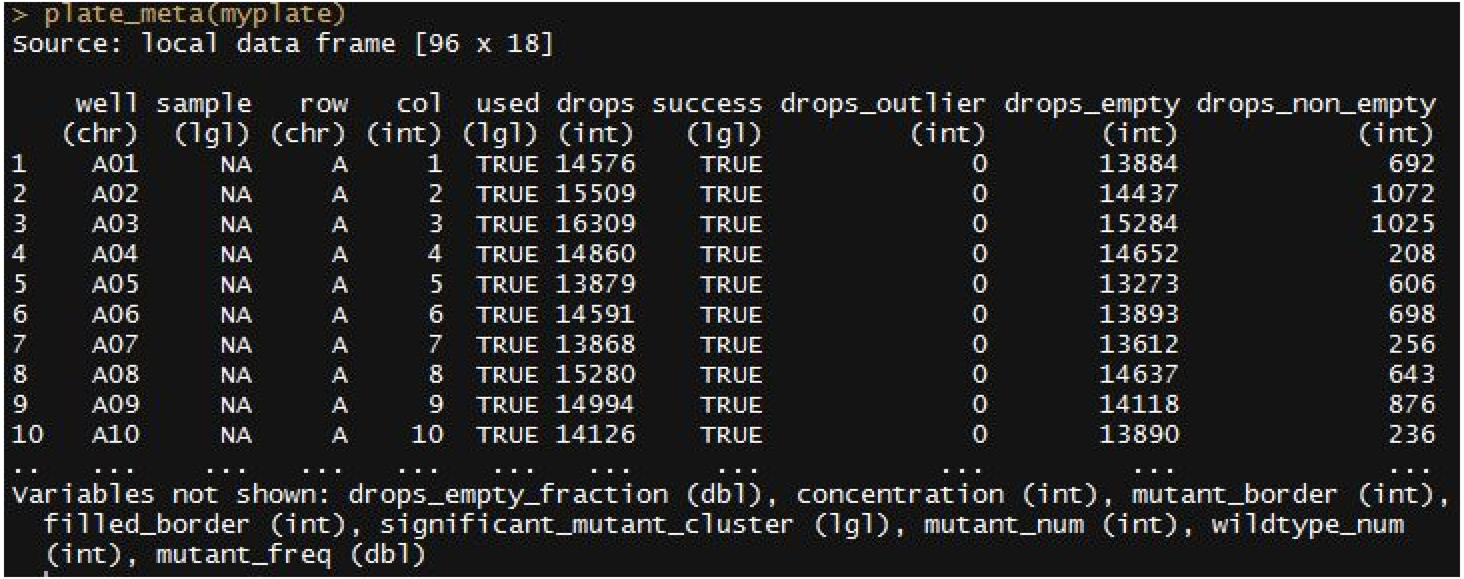
Results: 41 CRC dataset

- All MT/WT classifications agree with pathologist
- Excellent agreement with manual approach (TOST pvalue: 1.8x10^-14)
- 64 seconds on my 3 year old personal laptop





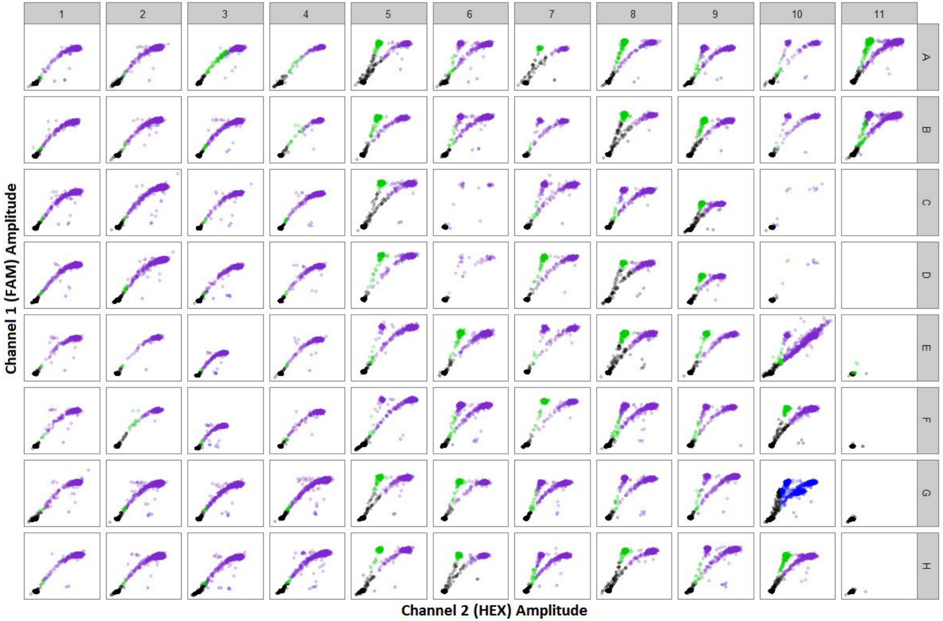
Results: 41 CRC dataset

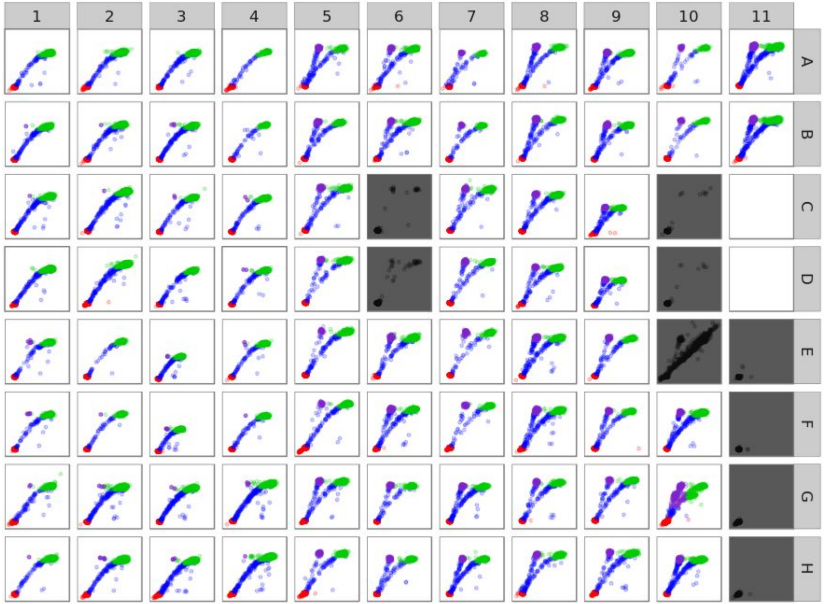


```
> plate_meta(myplate)
Source: local data frame [96 x 18]

    well sample   row   col  used drops success drops_outlier drops_empty drops_non_empty
   (chr)  (lgl) (chr) (int) (lgl) (int)   (lgl)         (int)       (int)           (int)
1    A01     NA     A     1  TRUE 14576    TRUE             0       13884             692
2    A02     NA     A     2  TRUE 15509    TRUE             0       14437            1072
3    A03     NA     A     3  TRUE 16309    TRUE             0       15284            1025
4    A04     NA     A     4  TRUE 14860    TRUE             0       14652             208
5    A05     NA     A     5  TRUE 13879    TRUE             0       13273             606
6    A06     NA     A     6  TRUE 14591    TRUE             0       13893             698
7    A07     NA     A     7  TRUE 13868    TRUE             0       13612             256
8    A08     NA     A     8  TRUE 15280    TRUE             0       14637             643
9    A09     NA     A     9  TRUE 14994    TRUE             0       14118             876
10   A10     NA     A    10  TRUE 14126    TRUE             0       13890             236
..   ...    ...   ...   ...   ...   ...     ...           ...         ...             ...
Variables not shown: drops_empty_fraction (dbl), concentration (int), mutant_border (int),
  filled_border (int), significant_mutant_cluster (lgl), mutant_num (int), wildtype_num
  (int), mutant_freq (dbl)
```

QuantaSoft vs *ddpcr*





Acknowledgements

CIHR IRSC

Jennifer Bryan

Charles Haynes

Ryan Brinkman

Roza Bidshahri

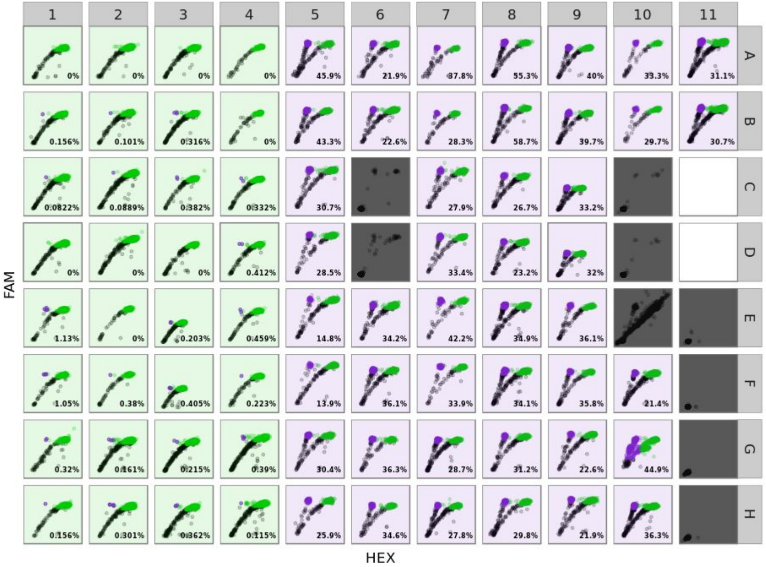
PavLab (Paul Pavlidis)

CIHR Strategic Training Program in BIOINFORMATICS

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Summary

1. Identify failed experiments
2. Identify outlier droplets
3. Identify empty droplets
4. Gate droplets (rain vs mutant vs wild type)
5. Classify each sample as mutant or wild type
6. (Revisit gating of wild type samples)

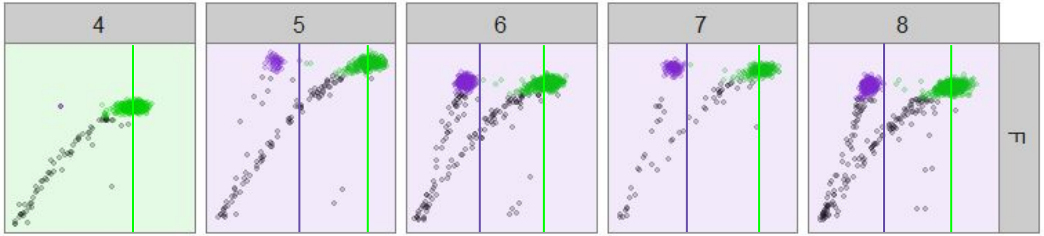


R package: ddpcr

Online: http://daattali.com/shiny/ddpcr/

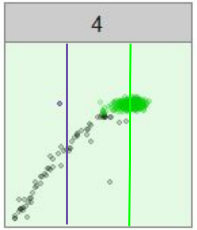
Step 6: Revisit gating of WT samples

- Wells with few MT droplets don’t have enough data to accurately gate
- Look at all mutant samples, we have an idea of where mutant drops are relative to wild type drops



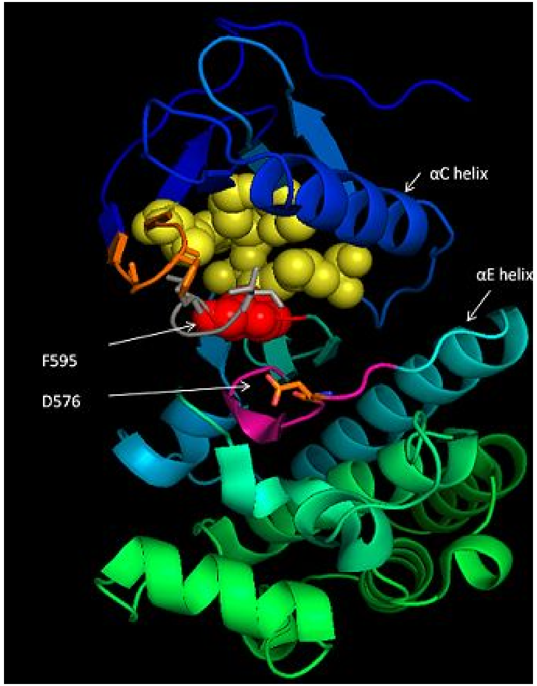
$\frac{\text{median(WT)}}{\text{max(MT)}} =$ 0.64 0.59 0.58 0.60

75-percentile of ratios: 0.60



B-Raf active vs inactive states

- Activation loop (orange) has strong hydrophobic interactions with P-loop (grey)
- These interactions keep the kinase inactive
- Activation loop gets phosphorylated → kinase becomes active
- Valine (V) hydrophobic, glutamic acid (E) is hydrophilic
- V600E → hydrophobic interactions are lost → kinase always active



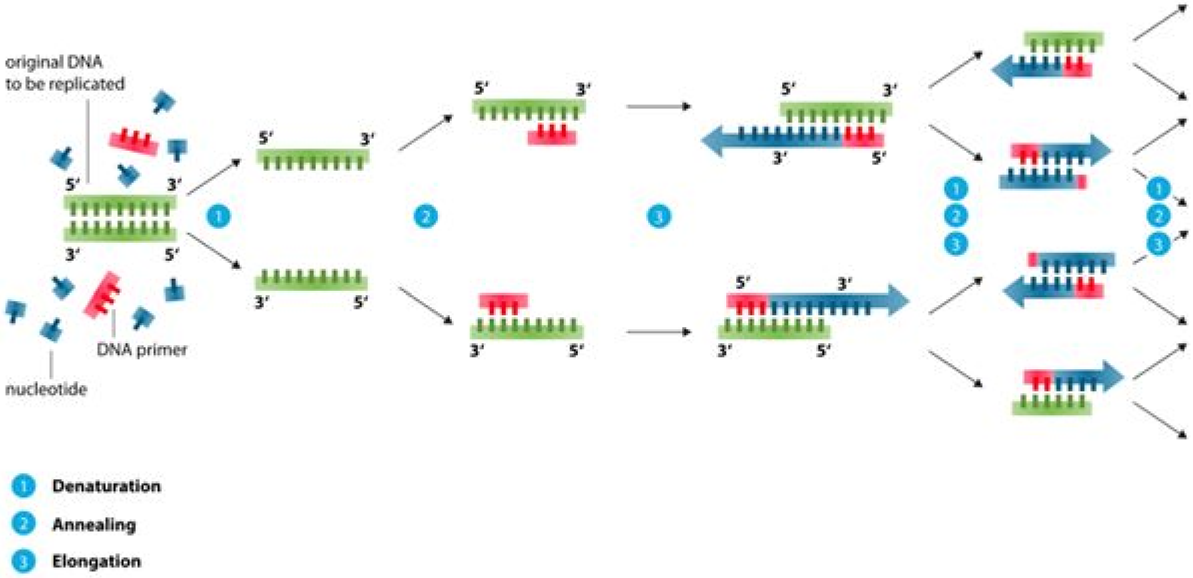
Wikipedia - BRAF

FFPE

- Formalin-fixed, paraffin-embedded
- A way to preserve tissue DNA
- Alternative to freezing
    - Less ideal, but doesn’t take up as much space and more practical
- Treat sample with formalin solution (which contains formaldehyde) to crosslink the DNA and fix it in place, then put it in paraffin wax
- Formalin causes some degradation, and also causes some C > T mutations

PCR

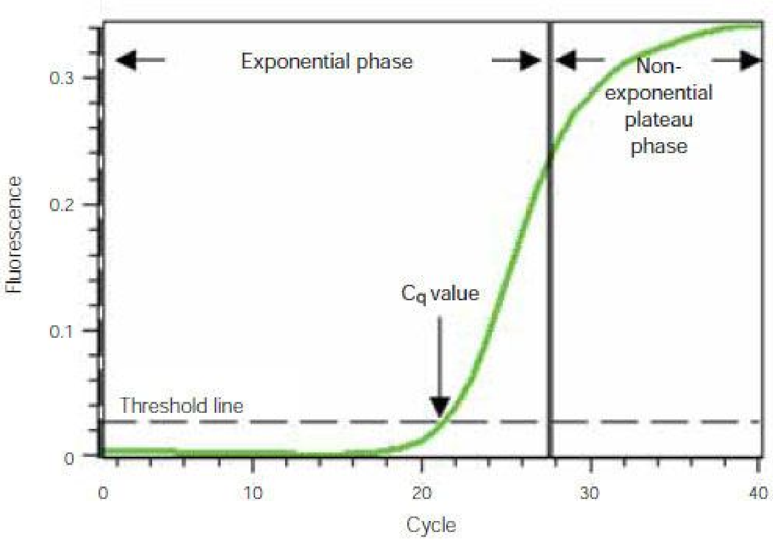
Amplify a specific piece of target DNA



Wikipedia - PCR

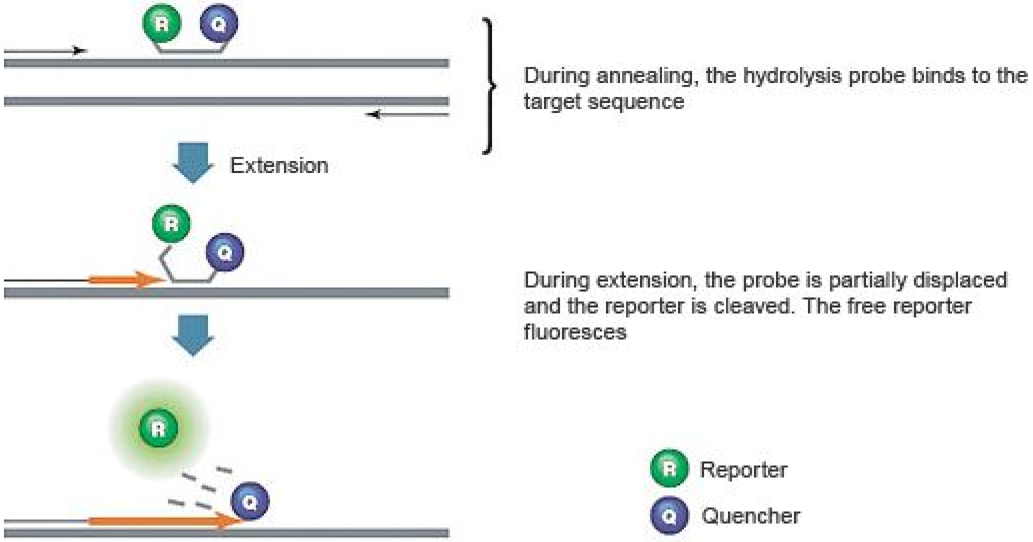
qPCR (real-time)

Monitor fluorescence that gets emitted during during amplification to quantify starting amount



Source: Bio-Rad

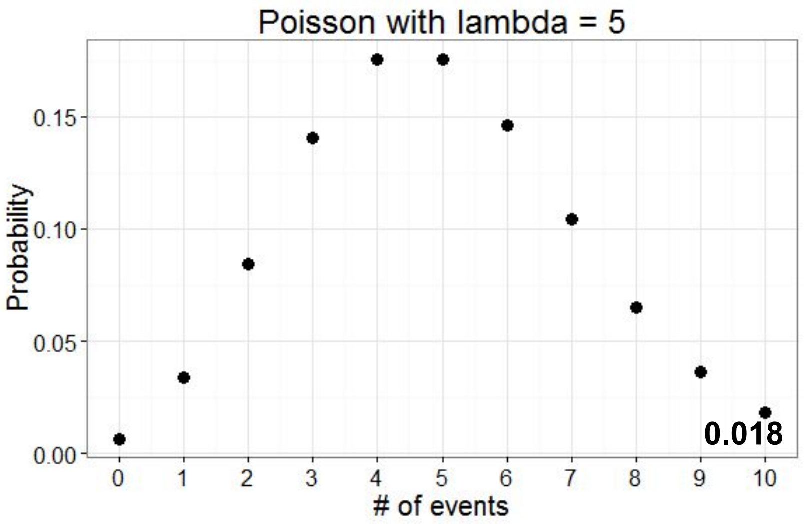
Hydrolysis probe

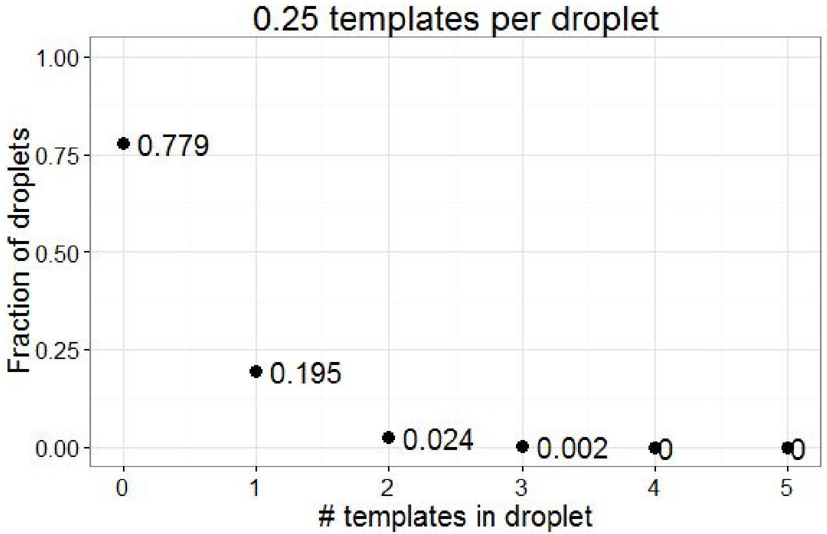


During annealing, the hydrolysis probe binds to the target sequence

During extension, the probe is partially displaced and the reporter is cleaved. The free reporter fluoresces

Copies of target / Droplet ~ Poisson





Example: if **20,000 droplets** and **5000 DNA molecules**, expect 0.25 copies / droplet on average
This means 78% of droplets will be empty, 19.5% will have one template, virtually none will have 4+

definetherain

- Upload all positive well
- Use kmeans to define a threshold for positive and threshold for negative (center +- 3 SD)
- Upload negative wells, and it will use the same thresholds to define positive, negative, and rain
- Concentration is calculated without including rain
- Assumes all wells have very similar distribution
- Still requires manual work of deciding which wells positive and which negative, & upload in two batches
- Kmeans fails if there is lots of rain

ddpcRquant

- Use combined data of multiple NTCs to model the extreme values of negative droplets by extreme value theory and set a threshold based on that
- Threshold is defined as the 99.5 percentile of the fitted extreme value distribution & used to classify negative threshold in every well
- Droplets are assigned to k groups (blocks) → maximum fluorescence intensity in each group (called the block maxima method) is used to estimate parameters for a generalized extreme value distribution → this distribution used to define threshold
- Assume all wells have same distribution of negatives as NTC
- Website claims R package available Nov 2015, still just an R script

Poisson to calculate concentration

$P(x, u) = (e^{-u})(u^{x})/x!$

If we set $x = 0$, then
$P(0, u)$ = prob droplet contains no templates
$= e^{-u}$
= fraction of negative droplets

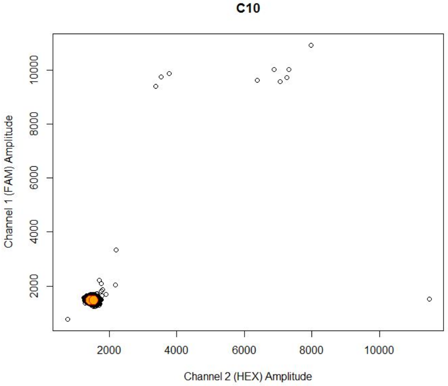
$P(x, u)$ = chance of having x copies in a droplet (where x = number of copies in a droplet, u = CPD)

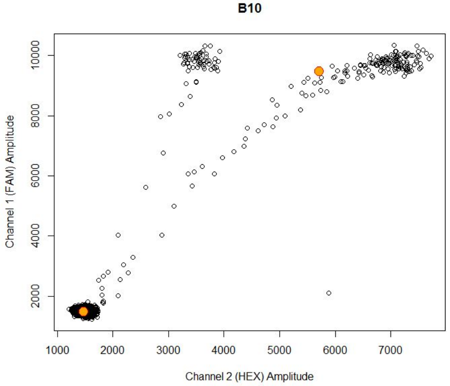
So if we know how many negative/positive droplets we have, we can use poisson equation to figure out the u (average copies per droplet) in the sample

$q = e^{-u} \rightarrow -\ln(q) = u$ (q = fraction of negative droplets)
For example, if 75% of droplets are empty, then CPD is $-\ln(.75) = 0.288$
If the droplet had a total of N droplets, then $0.288*N$ = total copies of target in initial sample

Step 1: Failed wells conditions

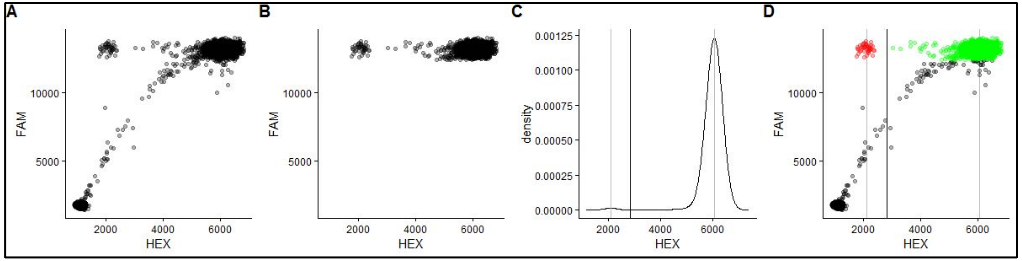
1. # droplets > threshold parameter
2. Empty and non-empty cluster must be well-separated

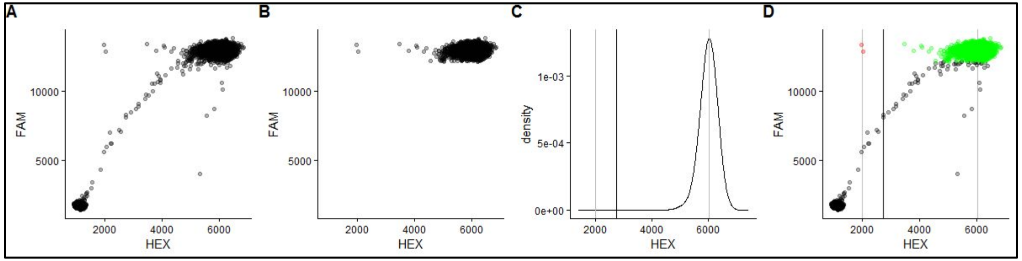


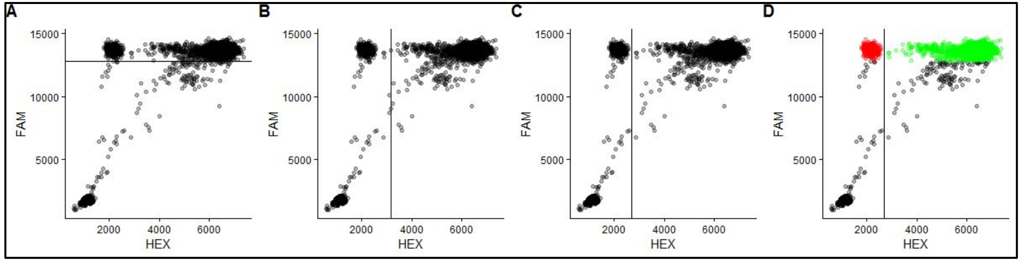


3. Empty cluster must be not too big nor too small

Step 4: More examples

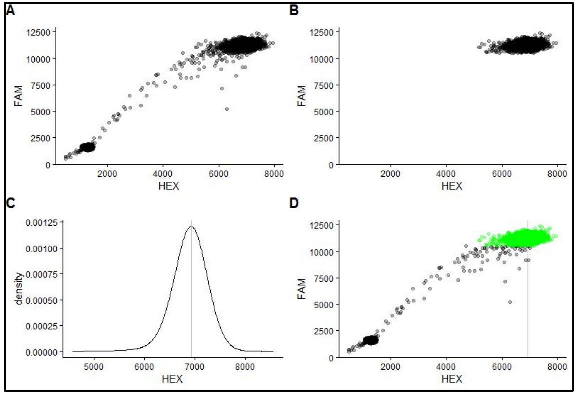




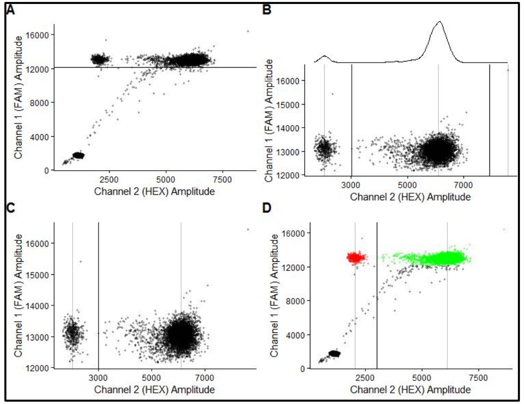


Step 4: Heuristics

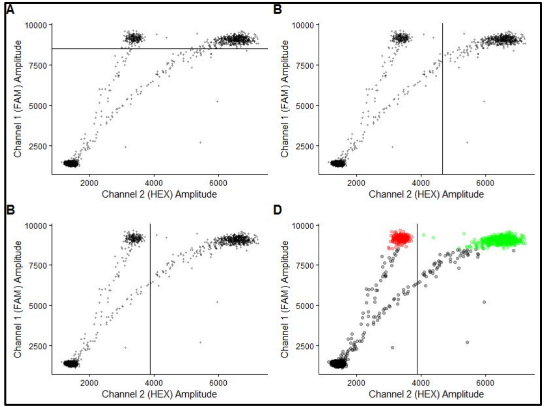
If only one peak initially, assume all droplets are WT



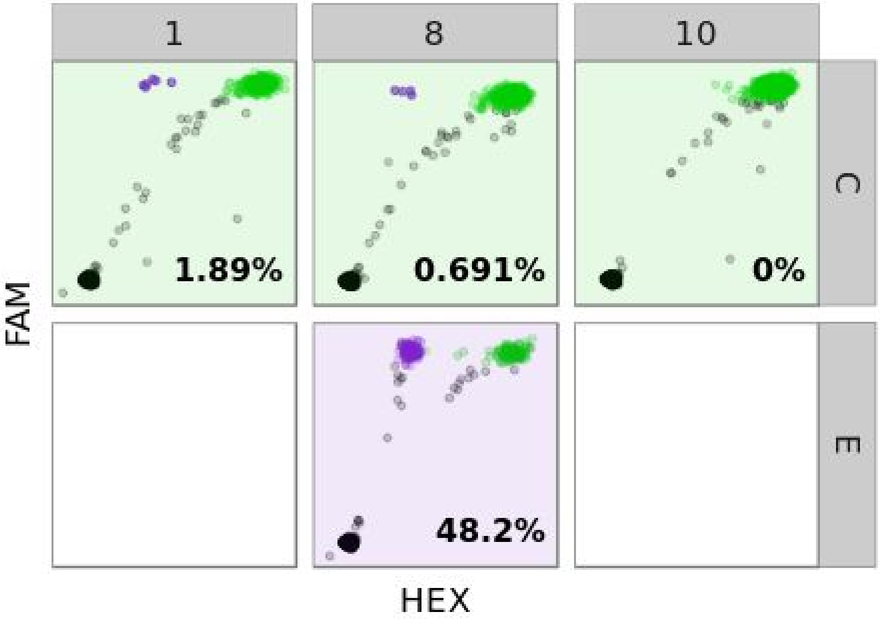
Every iteration, if <10% of droplets are beyond right-most peak, discard it



New gate is calculated as center+3SD of mutants, if it’s closer then use it instead

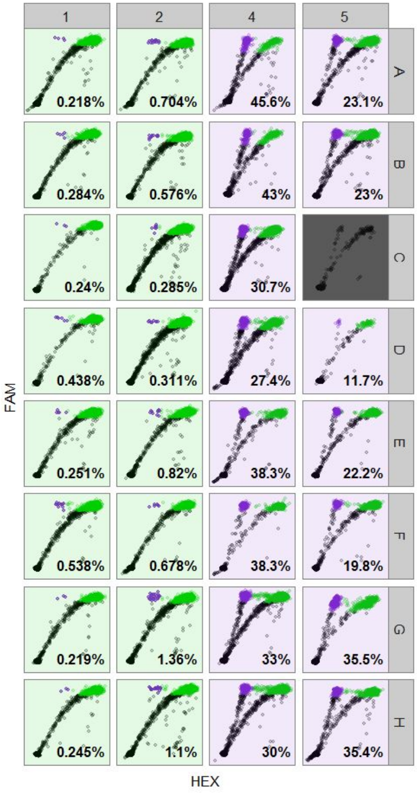


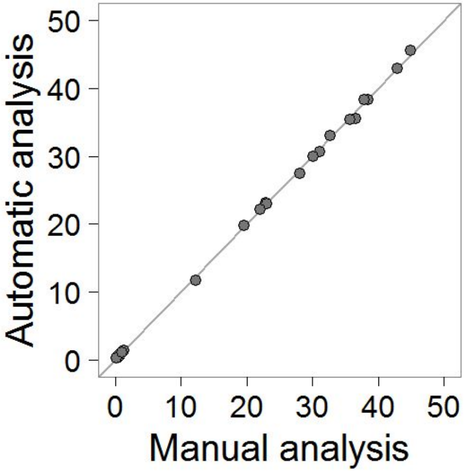
Results: Horizon samples



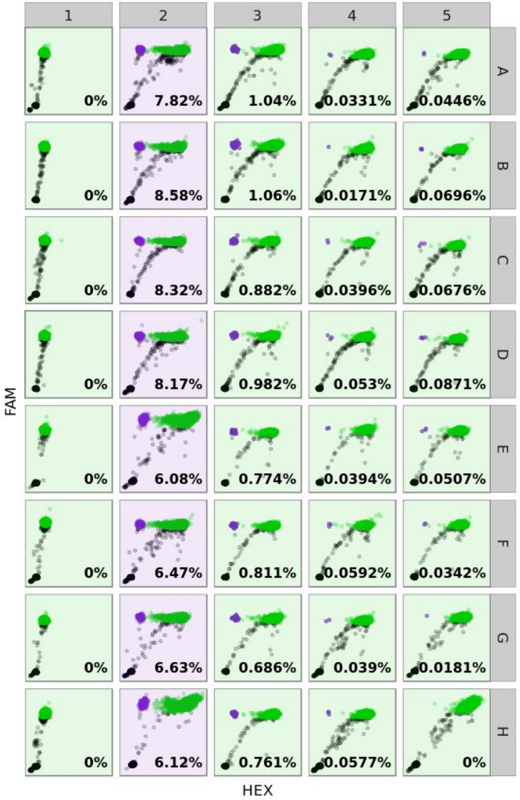
Real MT freq	Calculate MT freq
1.4	1.89
0.8	0.691
0	0
50	48.2

Results: CRC repeat





Results: brafv600k_plasmid_celline


A
B
C
D
E
F
G
H
FAM
HEX