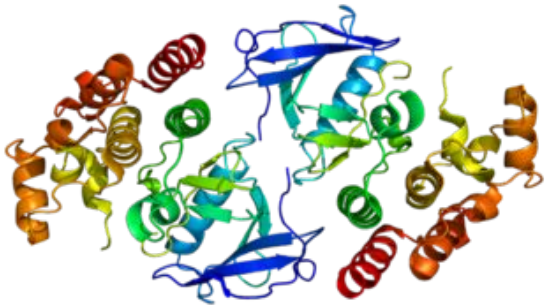
Automatic Analysis of Dual-Channel Droplet Digital PCR Experiments to Detect BRAF-V600 Mutations

Dean Attali

http://deanattali.com

Jennifer Bryan Lab @ MSL, UBC



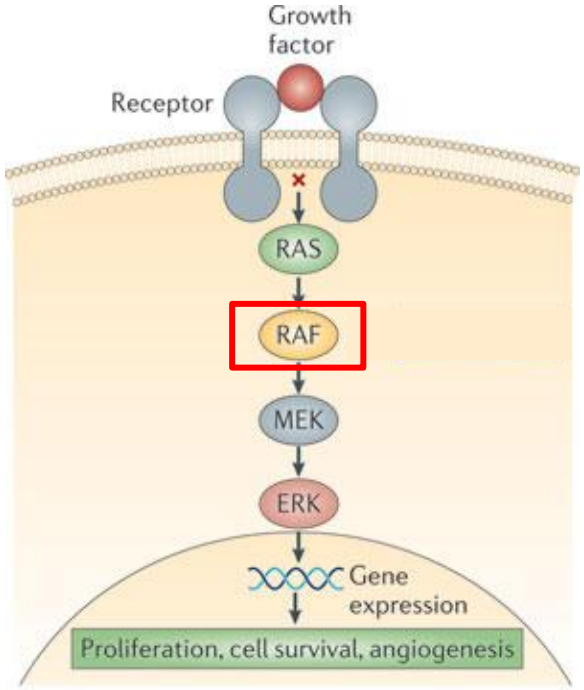
MSc Bioinformatics Defense

April 21, 2016

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**

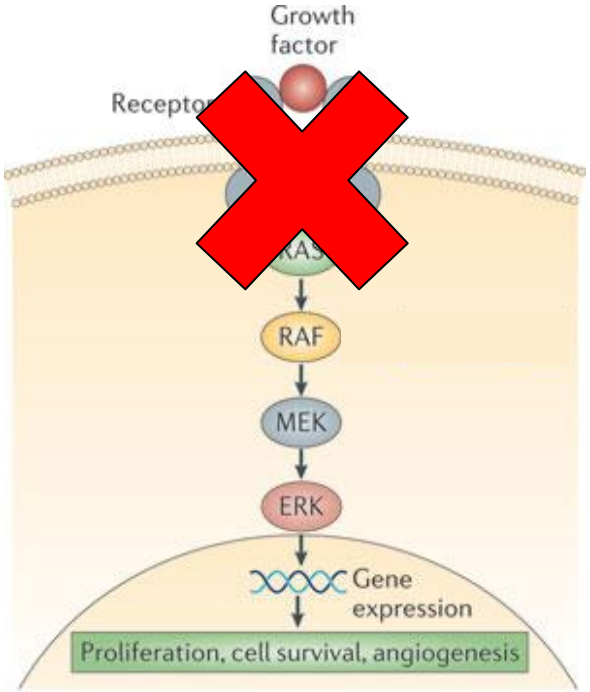


Flaherty et al. 2011.Nature Reviews Drug Discovery 10:811-812

BRAF-V600 Mutations

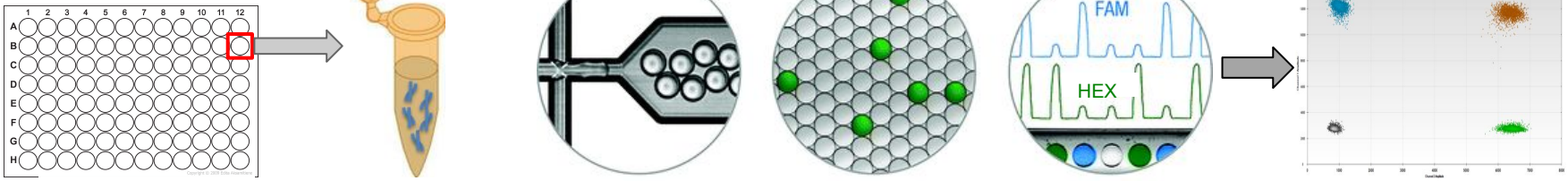
- V600 **mutation** ⇒ Constitutively active ⇒ **Uncontrolled** cell **growth** ⇒ **Tumour**
- 50% of melanoma tumours, 10% of colorectal cancers
- Presence/absence of V600 mutation affects treatment

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

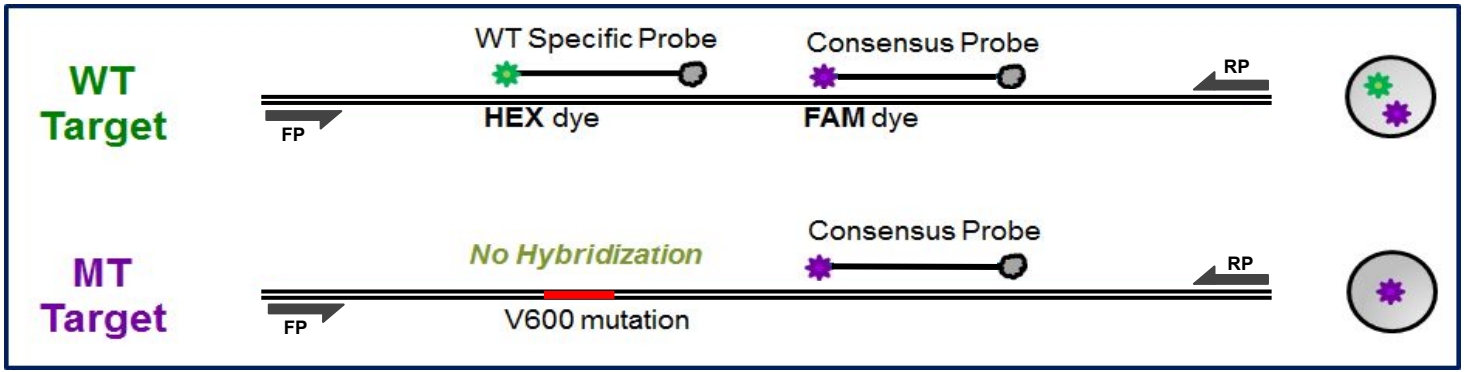


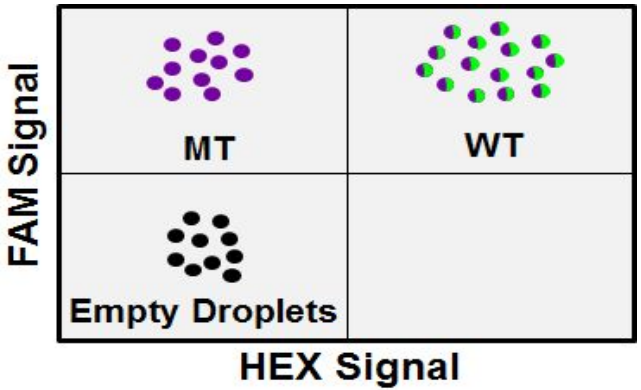
Flaherty et al. 2011.Nature Reviews Drug Discovery 10:811-812

Droplet Digital PCR (ddPCR)



BRAF-V600 Mutation ddPCR Assay

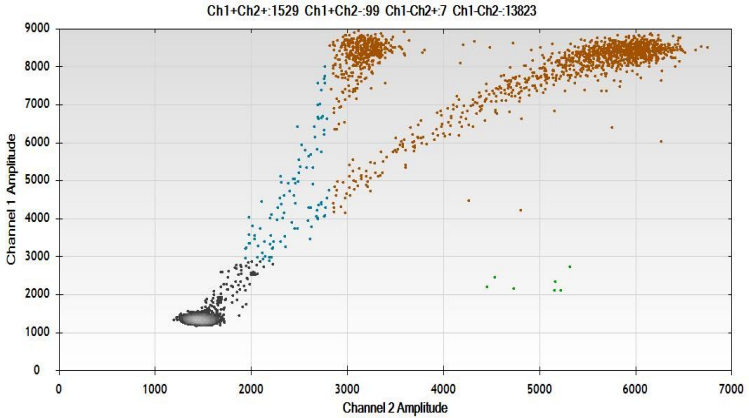


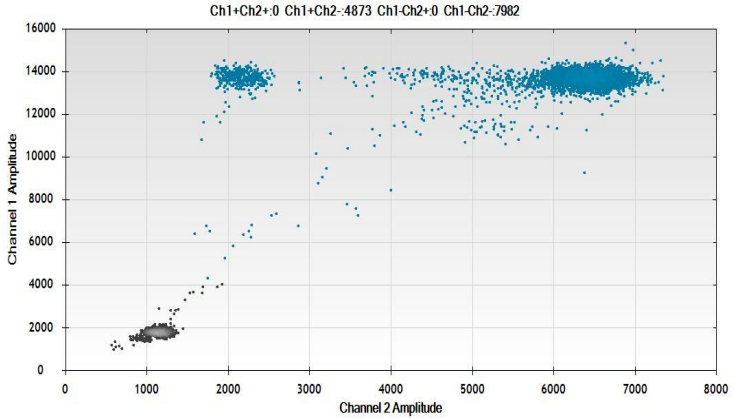


BRAF mutation frequency = $\frac{\text{\# MT droplets}}{\text{\# MT droplets + \# WT droplets}}$

Gating ddPCR Data

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





Goal 1 - Gating ddPCR Automatically

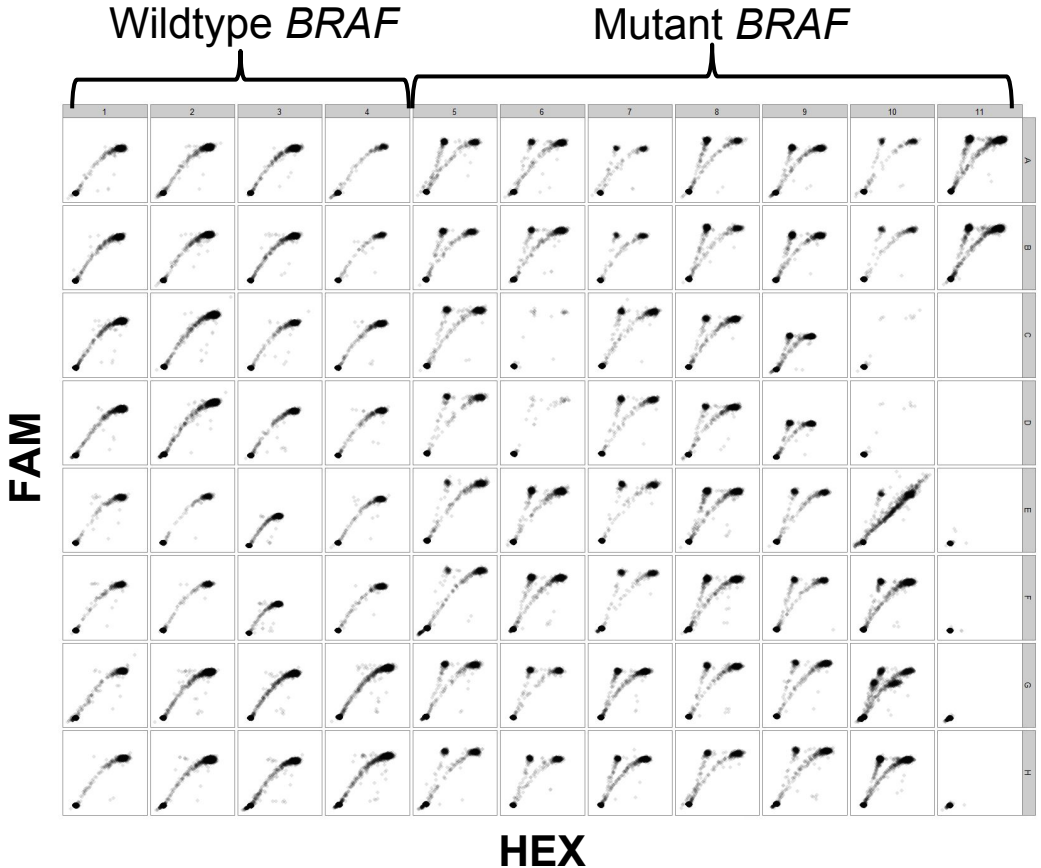
- Given ddPCR output ⇒ calculate mut*BRAF* frequency
- **Automated**
- Objective
- Reproducible
- Better gating than QuantaSoft

Goal 2 - Make it easily accessible

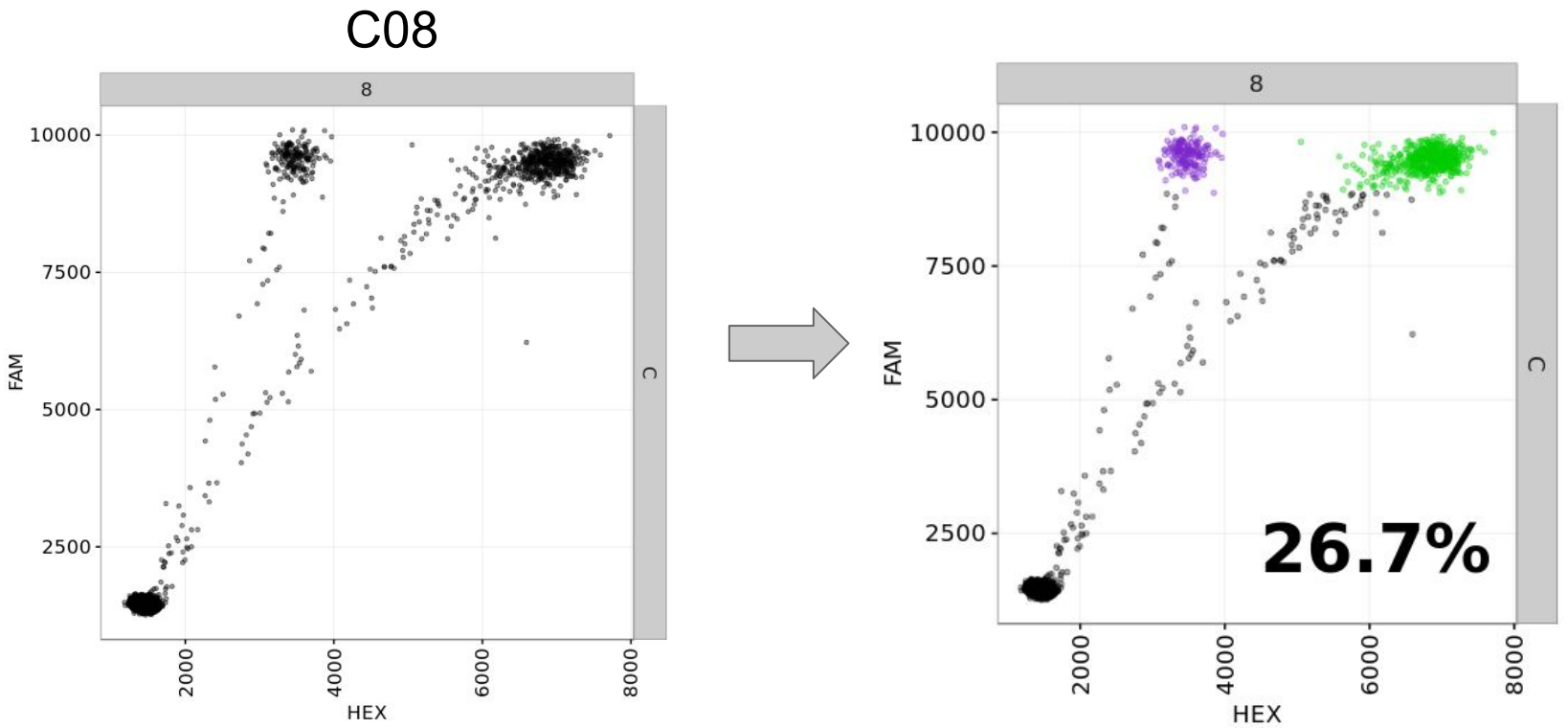
- Make R package
  - For people comfortable with R
  - *ddpcr* (on CRAN)

- 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

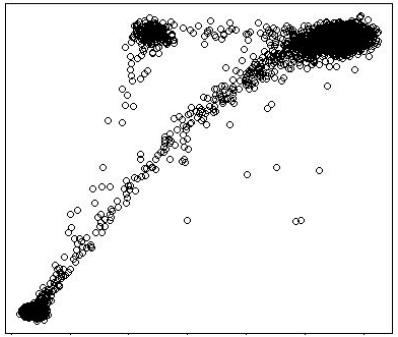


Goal

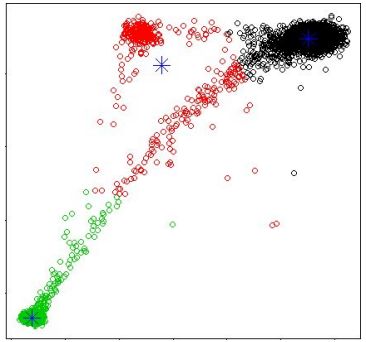

C08
8
10000
7500
5000
2500
FAM
2000
4000
6000
8000
HEX
C
8
10000
7500
5000
2500
FAM
26.7%
2000
4000
6000
8000
HEX
C

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

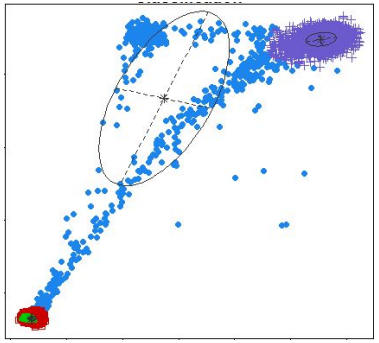
Raw data



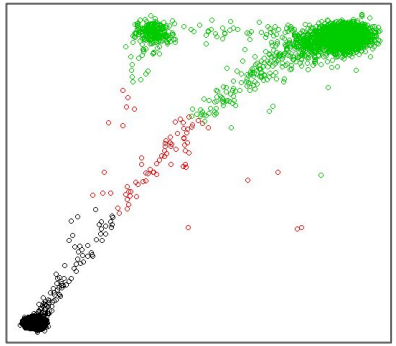
K-means



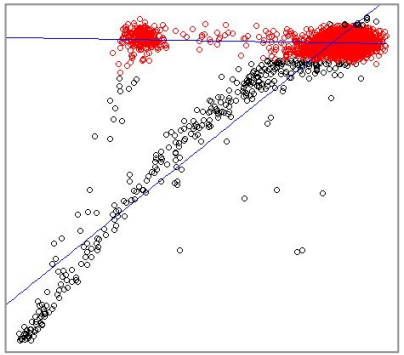
Mixture of gaussians



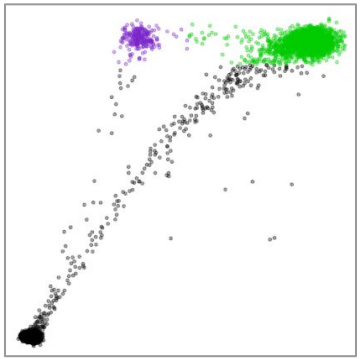
Hierarchical clustering



Mixture of regression models

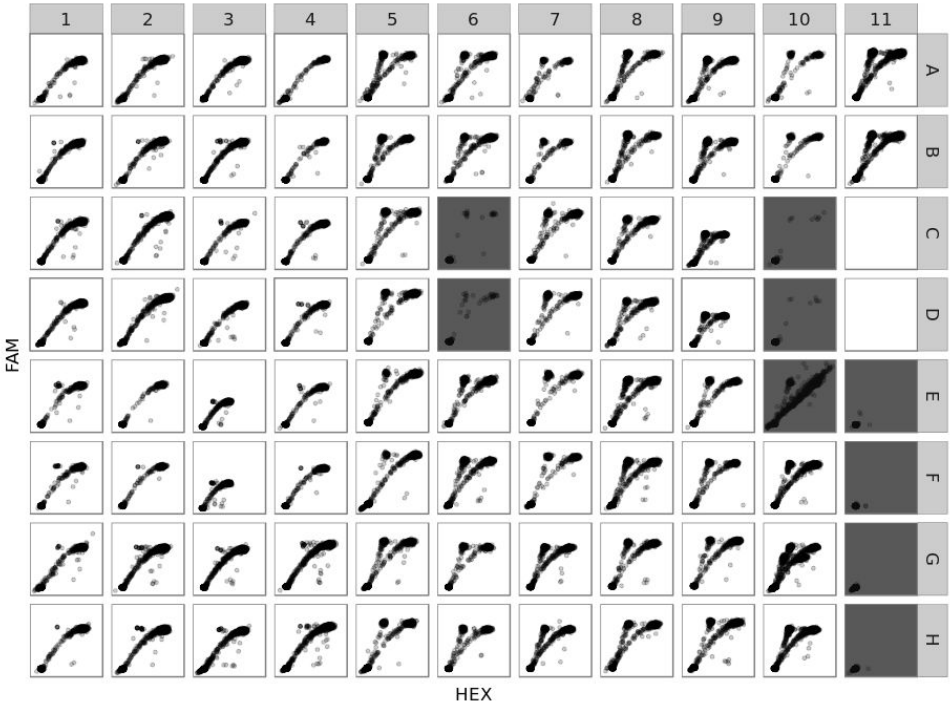


My tool

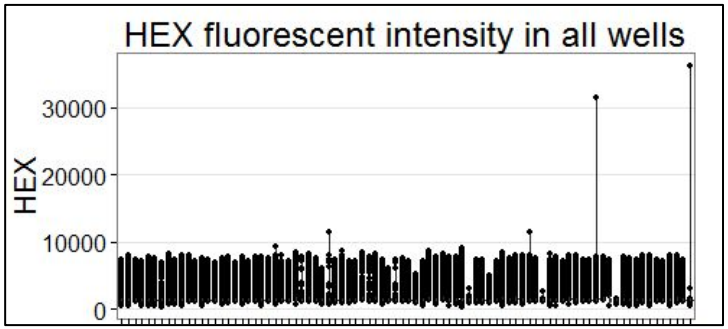


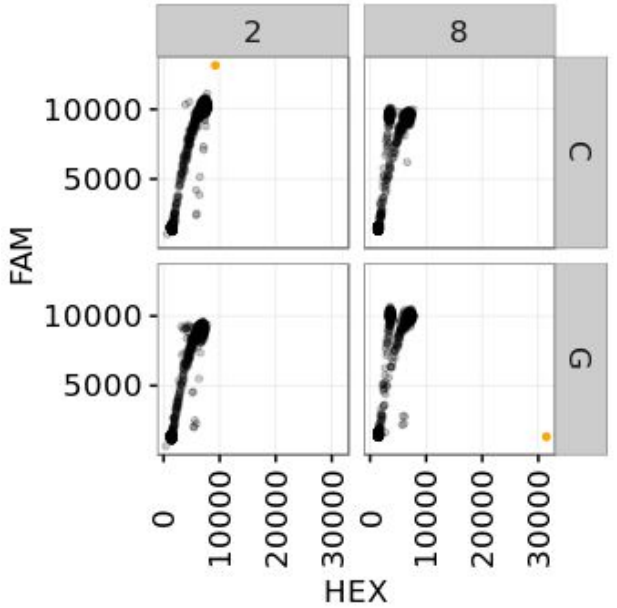
Step 1: Identify failed experiments

- Use QC metrics to ensure enough data in well

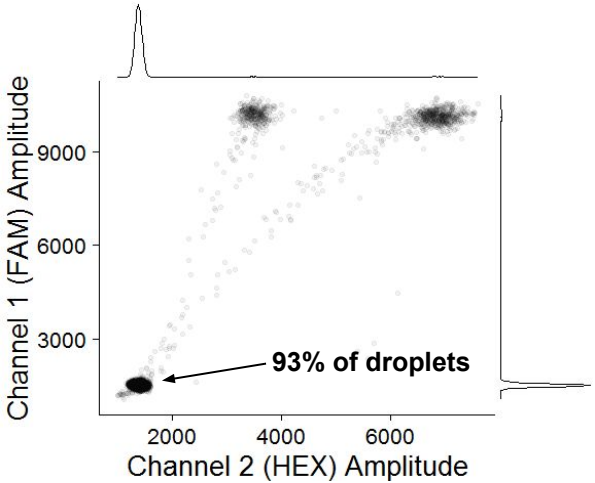


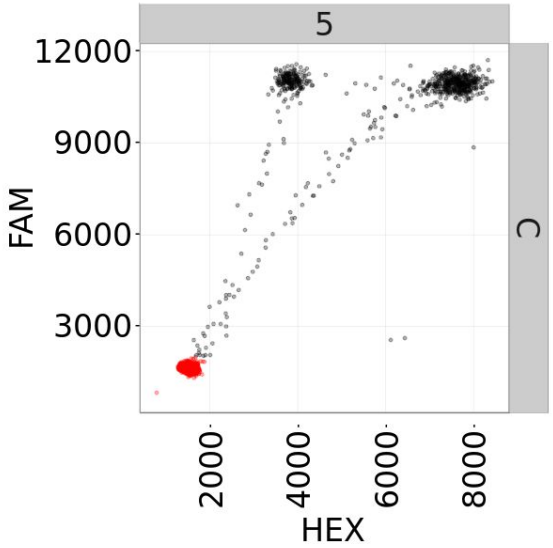
Step 2: Identify outlier droplets

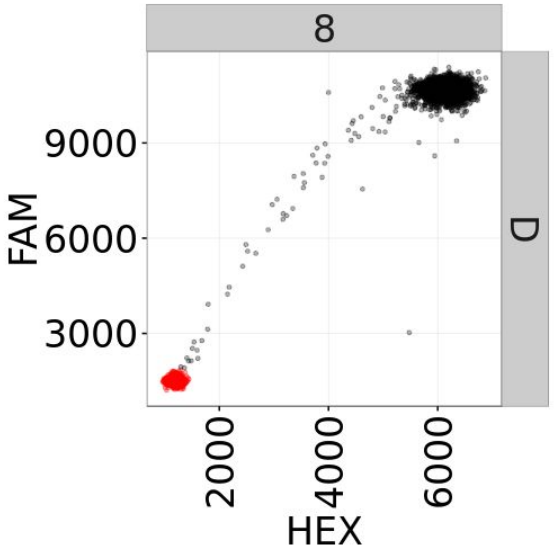




Step 3: Identify empty droplets



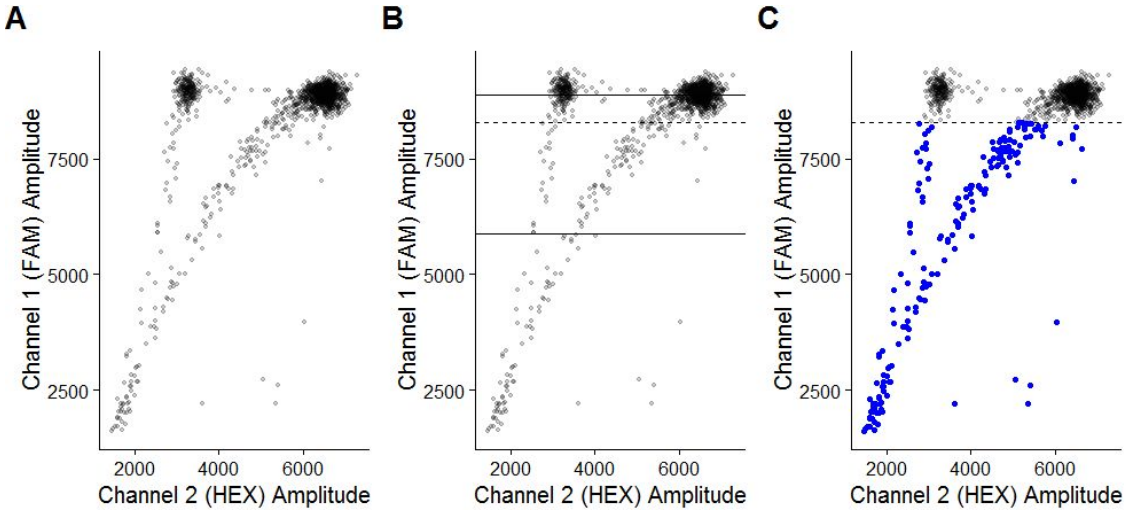




Step 4: Gate droplets (rain/MT/WT)

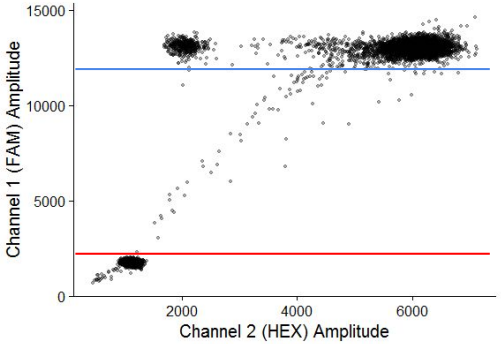
First substep: remove the rain

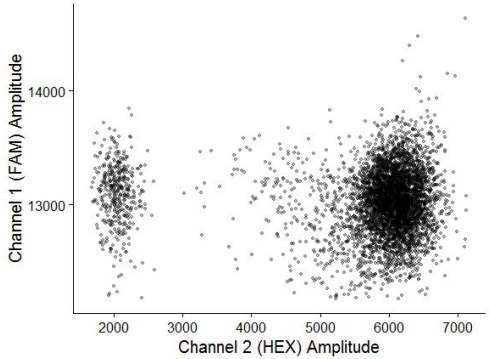
Fit two-component GMM to FAM, threshold = μ - 3σ



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

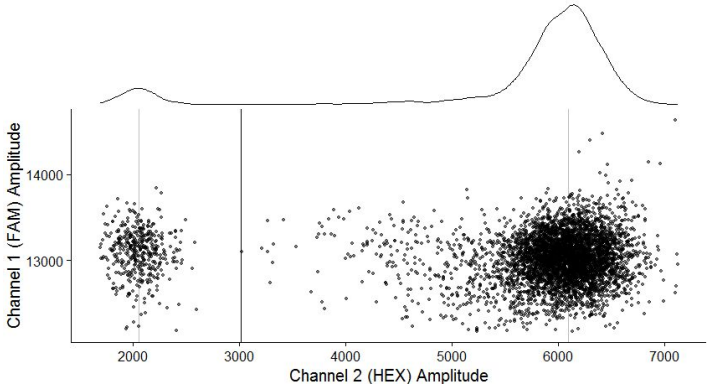
1. Raw data

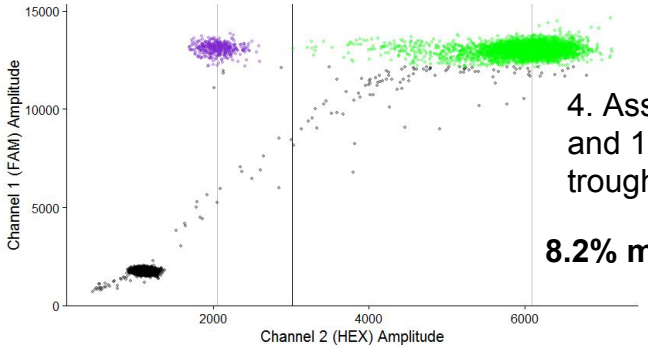




2. Remove empty and rain

3. Kernel density estimation (KDE) of HEX values

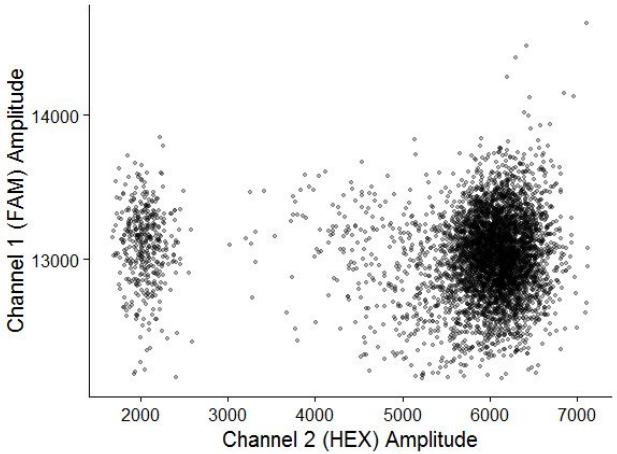




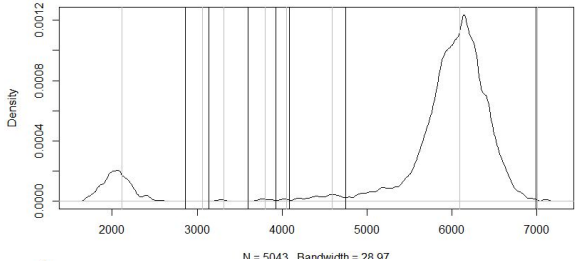
4. Assuming 2 peaks and 1 trough, use trough as gate

8.2% mut*BRAF*

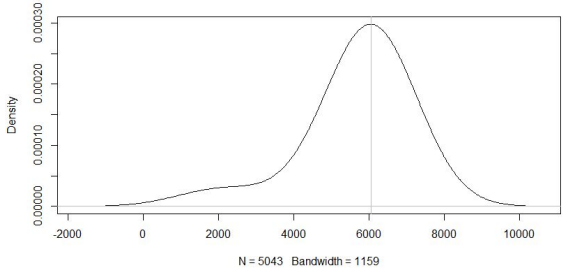
KDE bandwidth selection



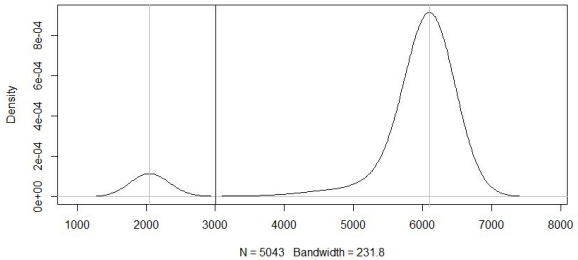
Bandwidth too small



Bandwidth too big

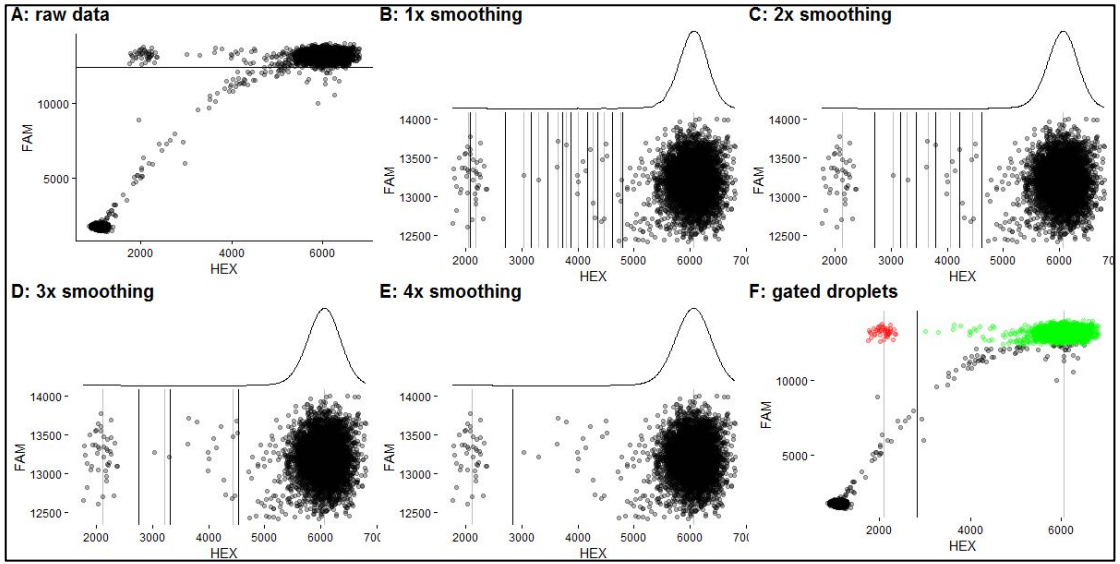


Bandwidth just right



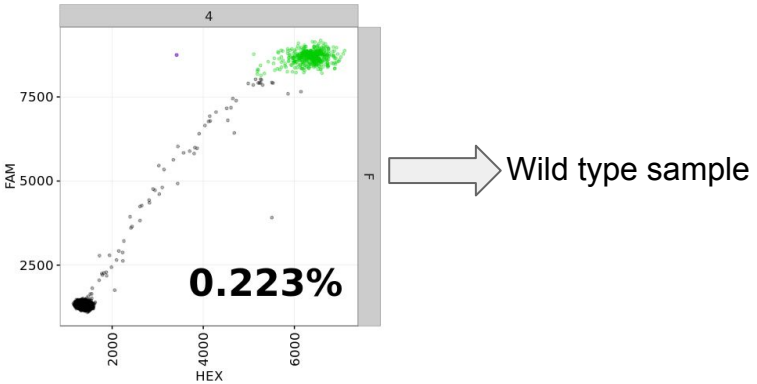
KDE bandwidth selection

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

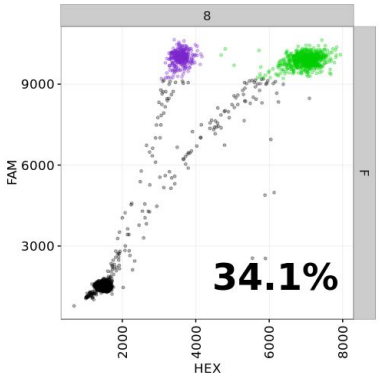


Step 5: Classify sample as MT/WT

- Mutation frequency statistically significantly > p (with significance level α) ⇒ Mutant
- In this analysis: use p = 0.01, α = 0.01
- Use binomial test: What's probability of observing at least N mutant droplets if the true mutant freq is 1% (p)?



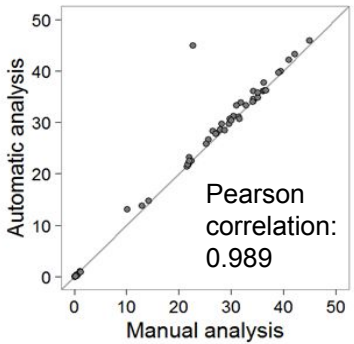
Wild type sample

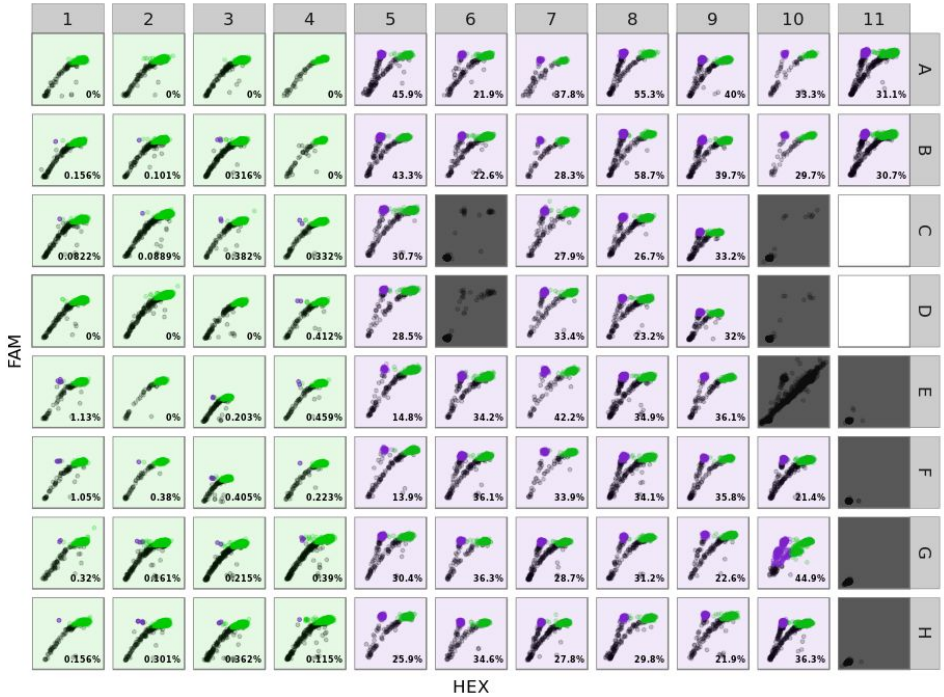


Mutant sample

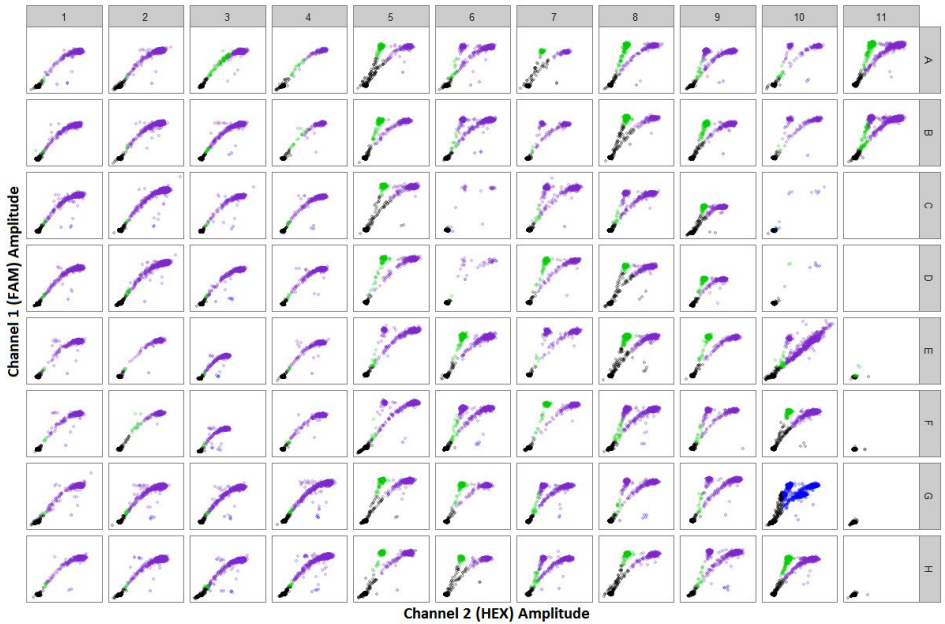
Results: 41 CRC dataset

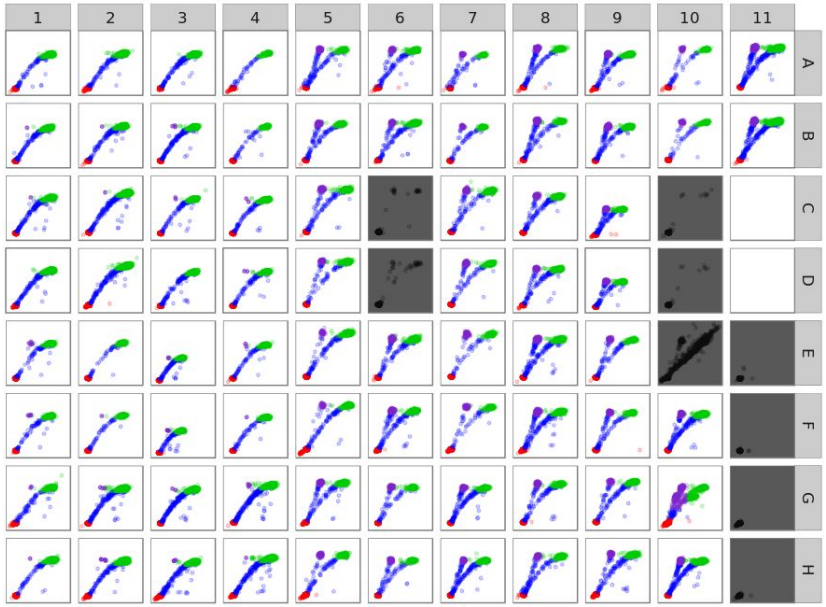
- All MT/WT classifications agree with pathologist
- Excellent agreement between MT frequencies calculated manually vs automatically
  - Mean difference: 0.39%
  - Two one-sided test (for bioequivalence) pvalue: 4.3x10^-8





QuantaSoft vs *ddpcr*





Acknowledgements

CIHR IRSC

Jennifer Bryan

Charles Haynes

Ryan Brinkman

Roza Bidshahri

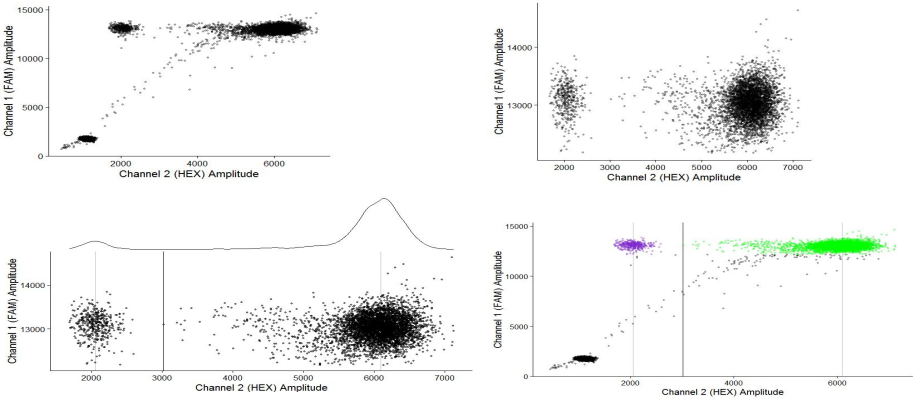
PavLab (Paul Pavlidis)

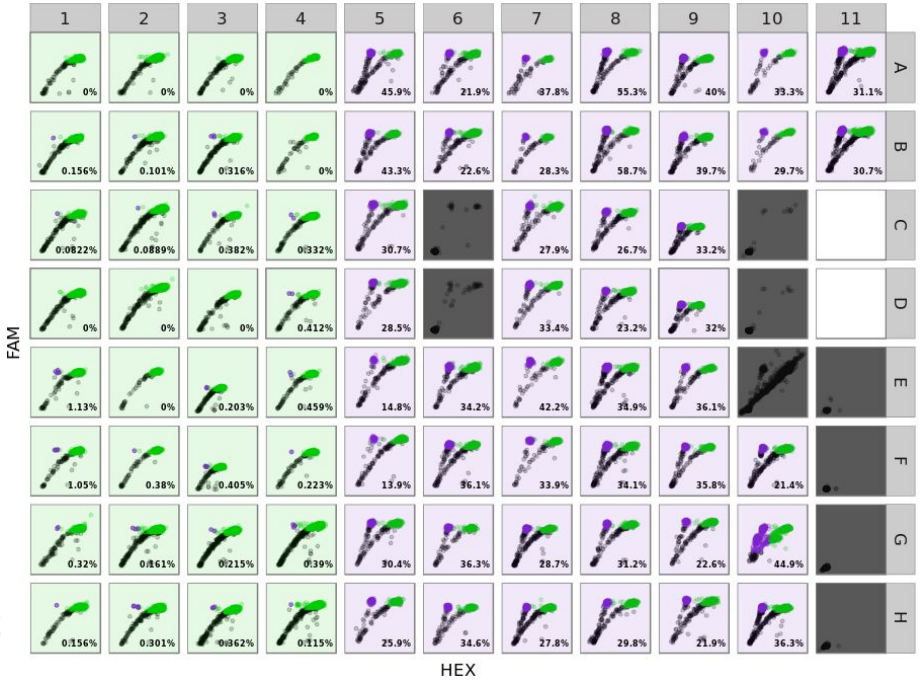
CIHR Strategic Training Program in BIOINFORMATICS

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Summary

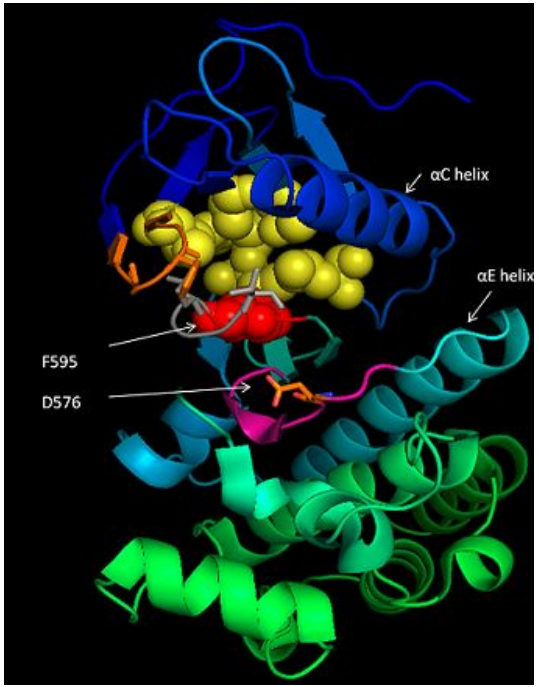
1. Identify failed experiments
2. Identify outlier droplets
3. Identify empty droplets
4. Gate droplets (rain vs MT vs WT)
5. Classify each sample as MT or WT





B-Raf active vs inactive states

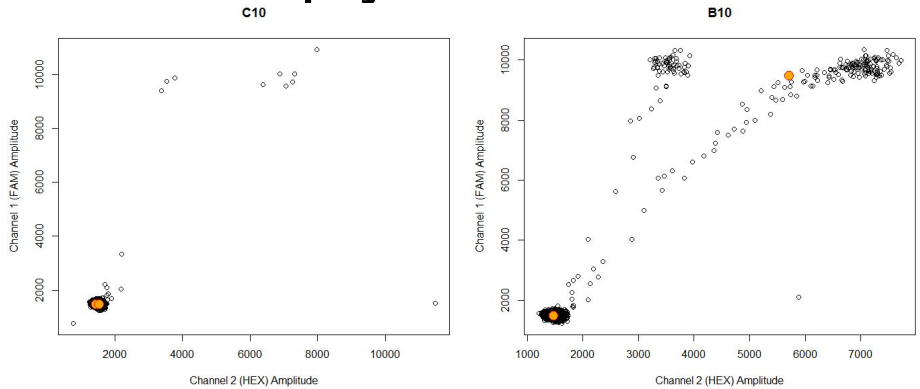
- Activation loop (orange) has strong hydrophobic interactions with P-loop (grey)
- These interactions keep the kinase inactive
- Activation loop gets phosphorylated → kinase becomes active
- Valine (V) hydrophobic, glutamic acid (E) is hydrophilic
- V600E → hydrophobic interactions are lost → kinase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



Wikipedia - BRAF

Step 1: Failed wells conditions

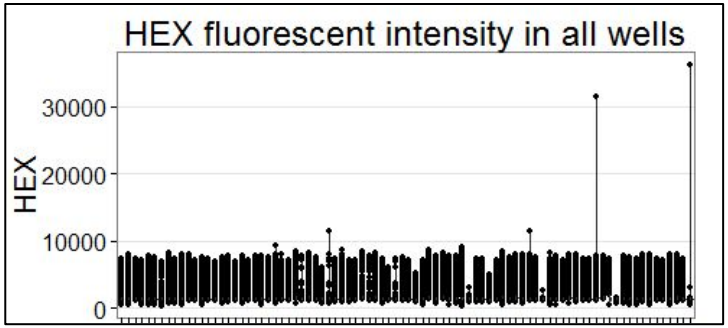
1. # droplets > threshold parameter
2. Empty and non-empty cluster must be well-separated

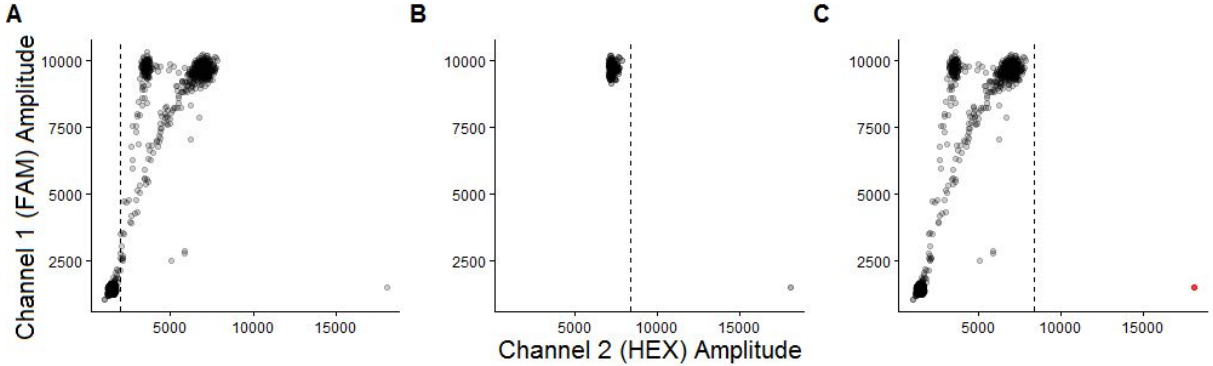


3. Empty cluster must be not too big nor too small

Step 2: Identify outlier droplets

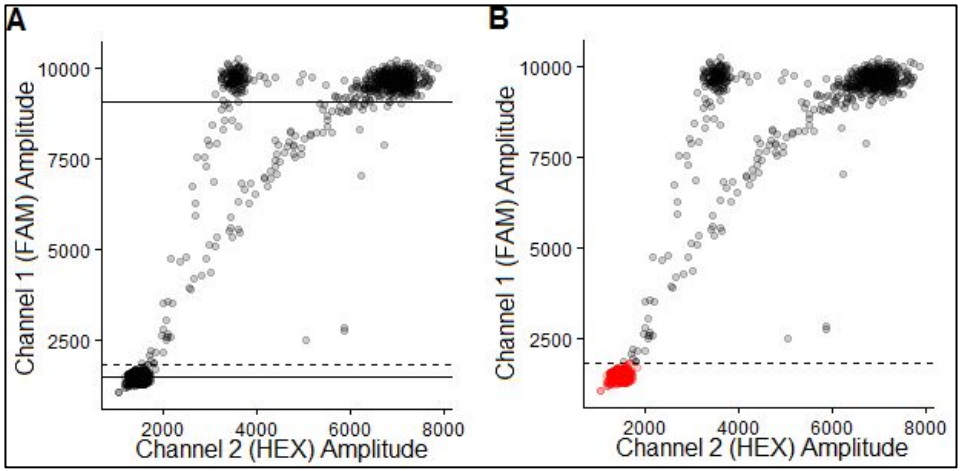
Take top x% of droplets,
define threshold as Q3 + 5IQR

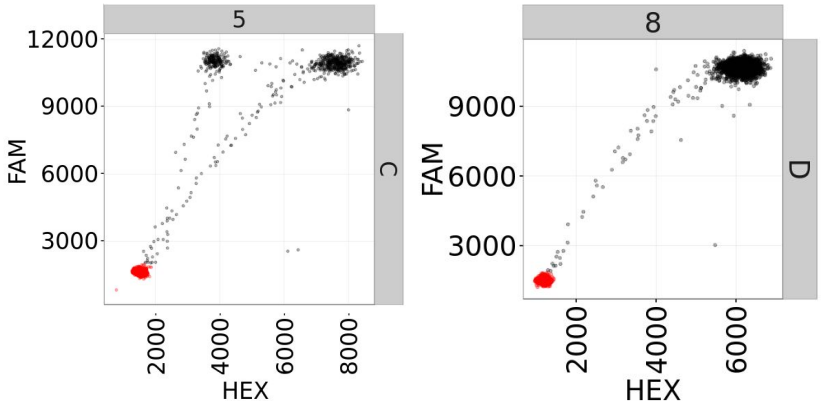




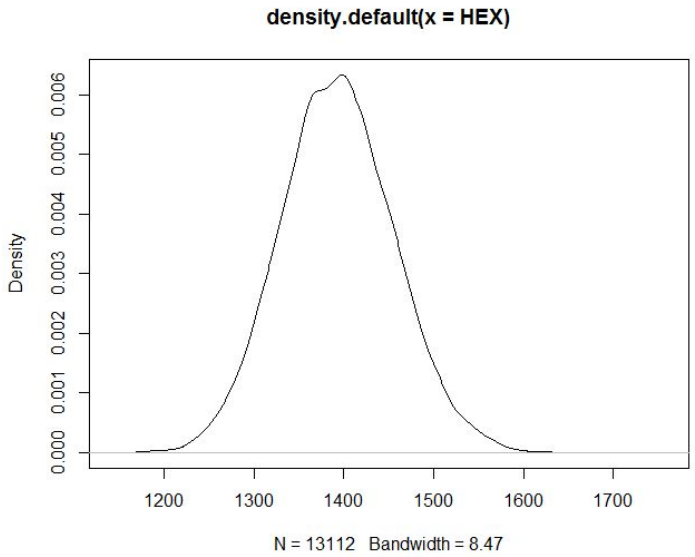
Step 3: Identify empty droplets

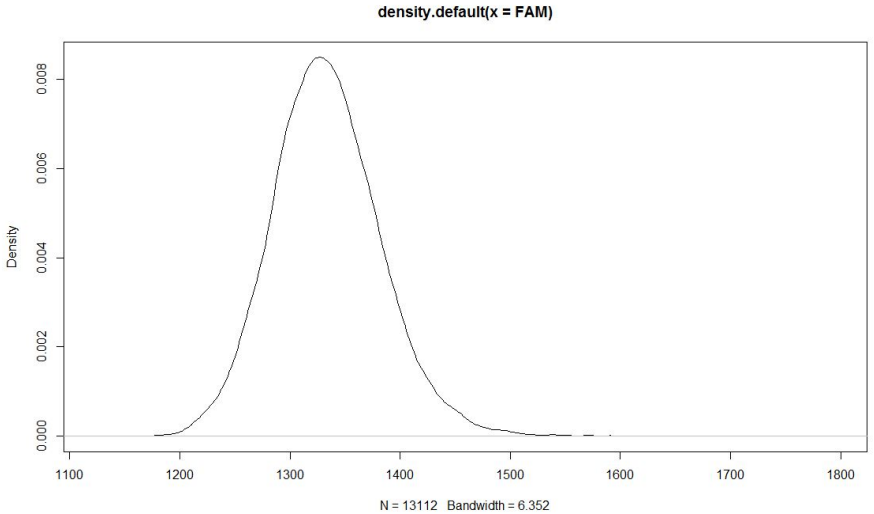
Fit two-component Gaussian mixture model to FAM values →
Lower population is empty droplets, threshold = μ + 5σ



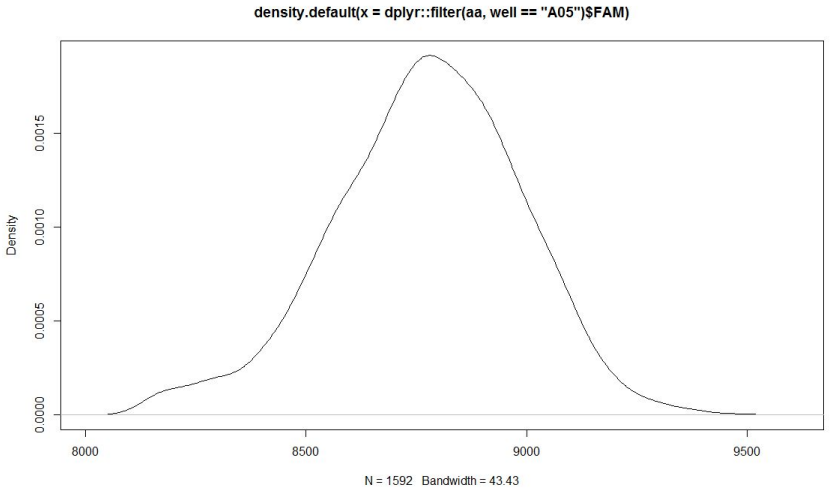


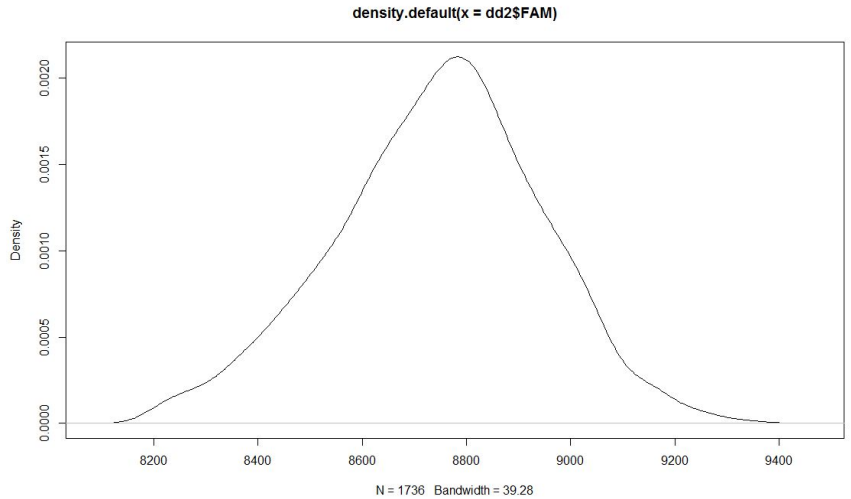
Empty droplets ~ normal





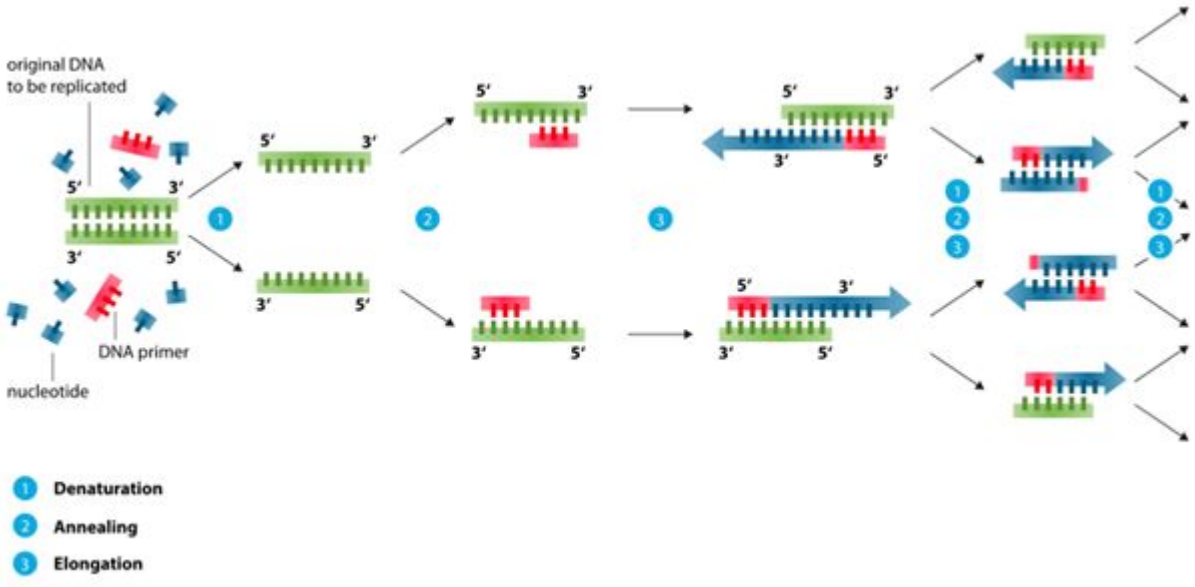
Filled droplets ~ normal





PCR

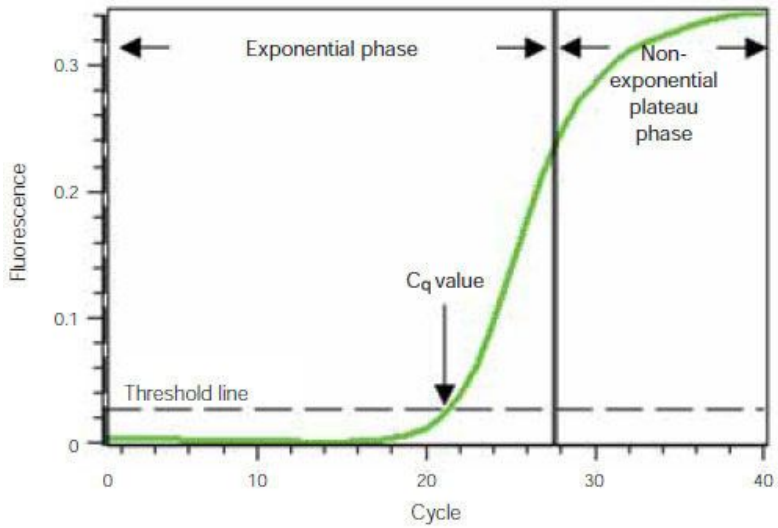
Amplify a specific piece of target DNA



Wikipedia - PCR

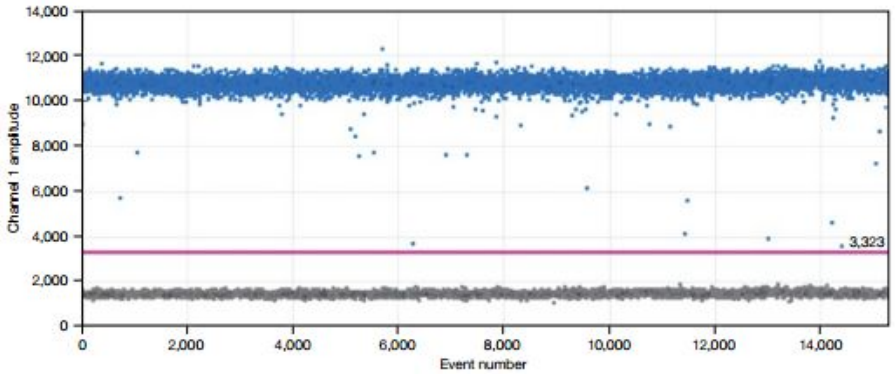
qPCR (real-time)

Monitor fluorescence that gets emitted during amplification to quantify starting amount

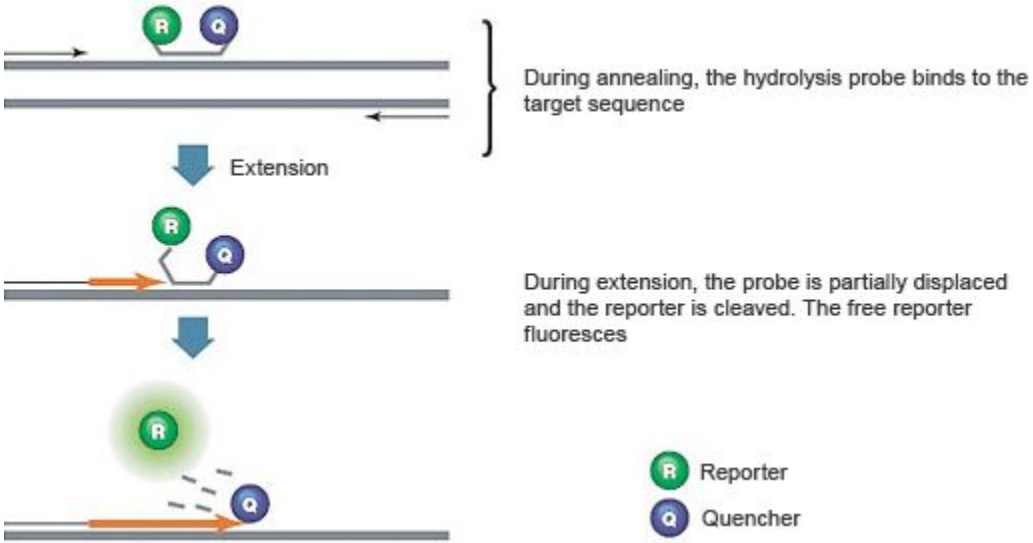


Source: Bio-Rad

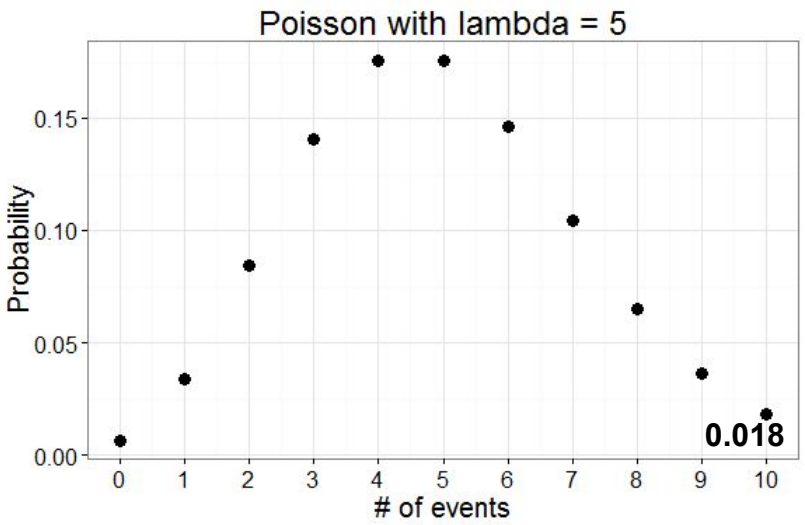
1D ddPCR

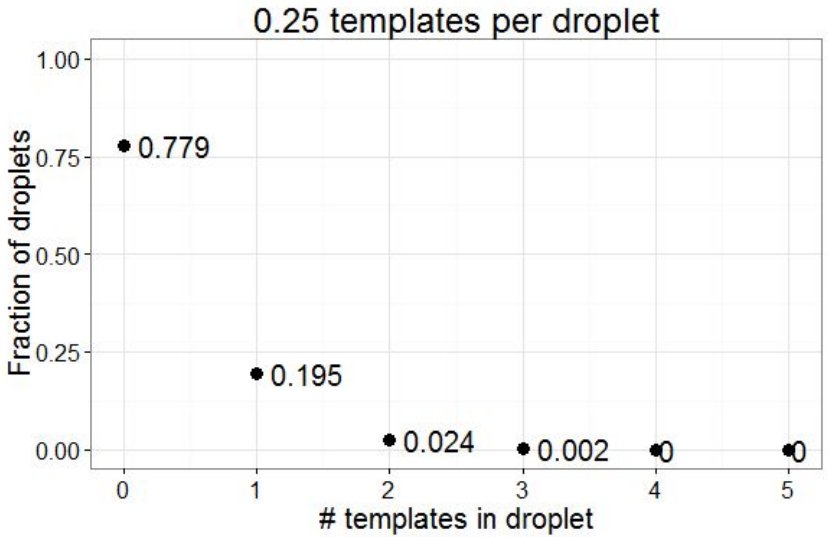

14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Channel 1 amplitude
3,323
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
Event number

Hydrolysis probe



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+

Poisson to calculate concentration

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

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

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

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 → -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

definetherain

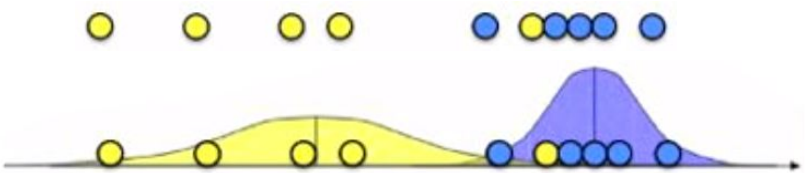
- Upload all positive well
- Use kmeans to define a threshold for positive and threshold for negative (center +- 3 SD)
- Upload negative wells, and it will use the same thresholds to define positive, negative, and rain
- Concentration is calculated without including rain
- Assumes all wells have very 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

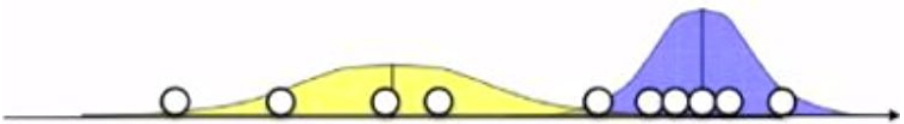
- Use combined data of multiple NTCs to model the extreme values of negative droplets by extreme value theory and set a threshold based on that
- Threshold is defined as the 99.5 percentile of the fitted extreme value distribution & used to classify negative threshold in every well
- Droplets are assigned 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

EM for estimating GMM params

If we know which distribution each observation is from, we can estimate the parameters



If we know the distribution parameters, we can determine which distribution each point is most likely from

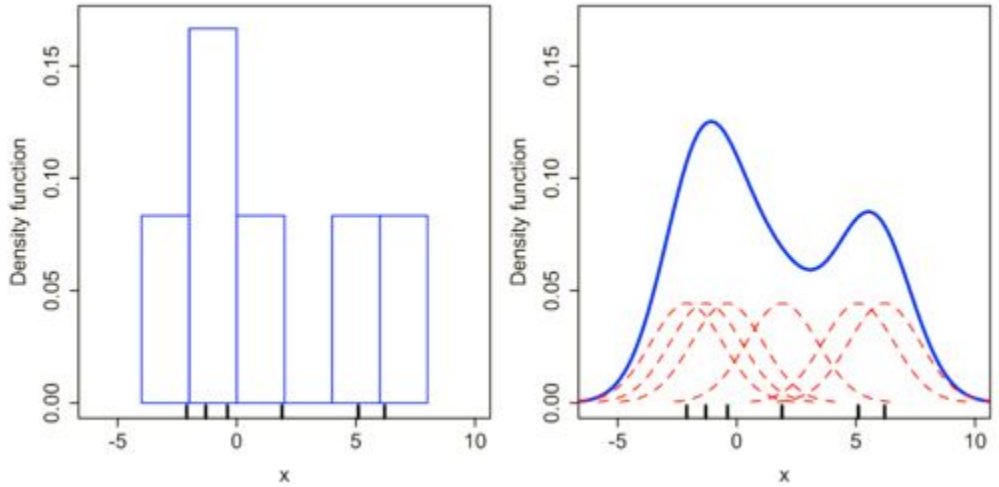


Chicken and egg: if we know one, we can estimate the other, but we know none

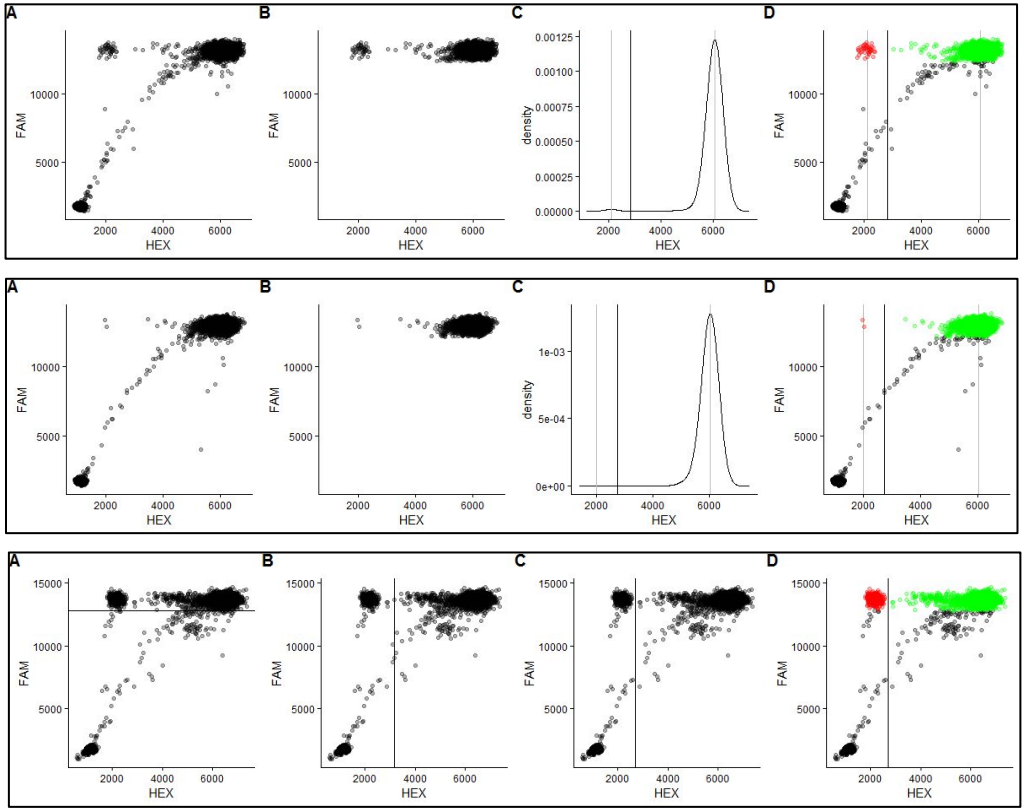
Randomly choose k gaussian parameters
E step: For each point, what's the prob it came from each gaussian?
M step: Use these (weighted) assignments to re-estimate parameters

KDE

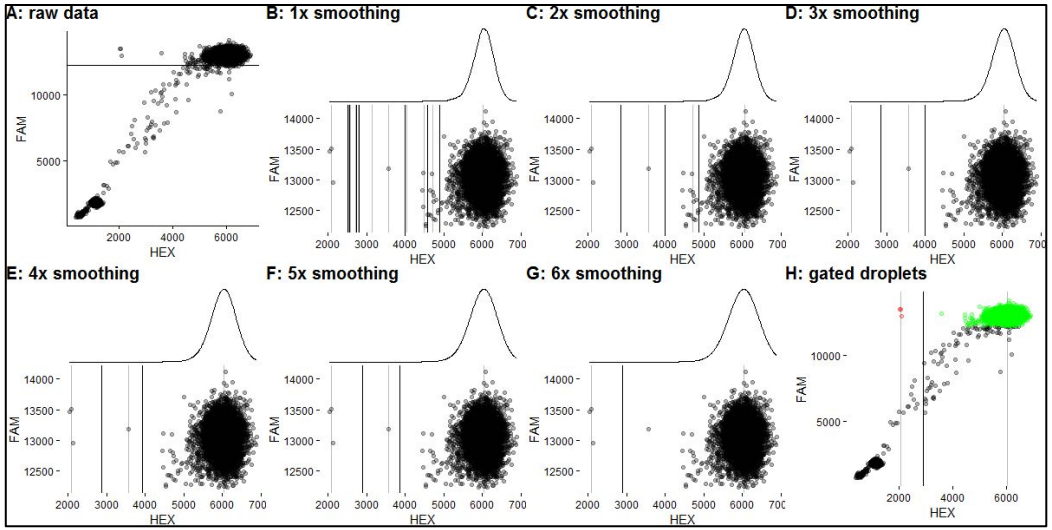
- Similar to histogram, but instead of each observation contributing to a bin, it contributes to a gaussian to make it smooth



Step 4: More examples

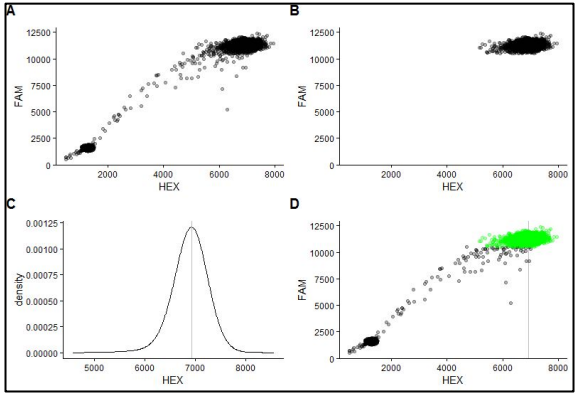


KDE bandwidth selection

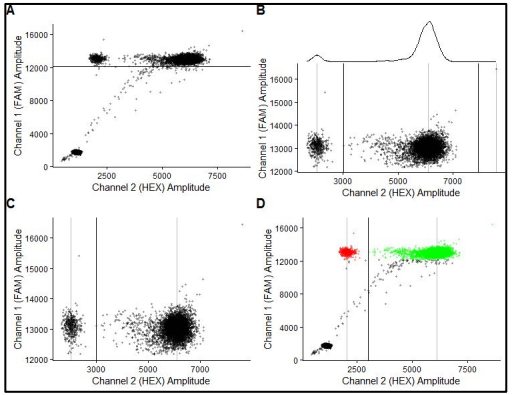


Step 4: Heuristics

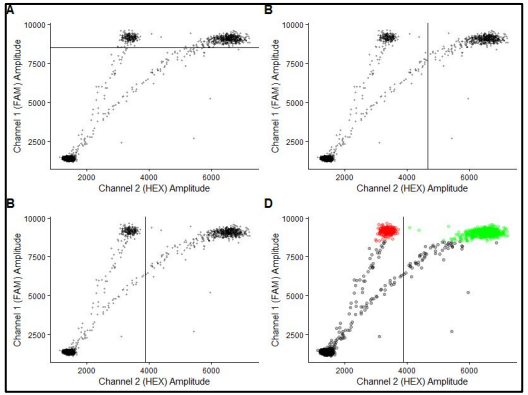
If only one peak initially, assume all droplets are WT



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



Step 5: Classify sample as MT/WT

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

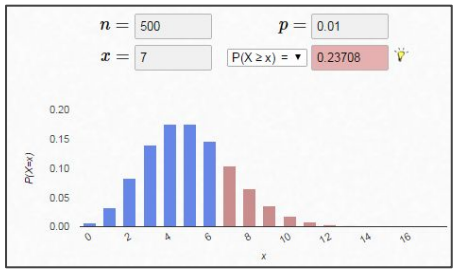
> α

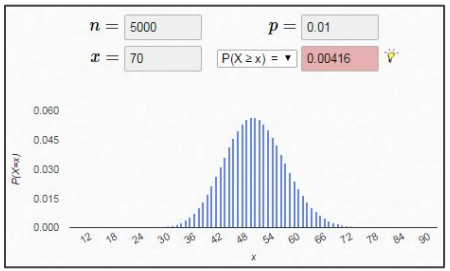
Well is classified as WT even though mut*BRAF* = 1.4%

P(X=*r*)

= $_nC_r \cdot p^r \cdot q^{n-r}$

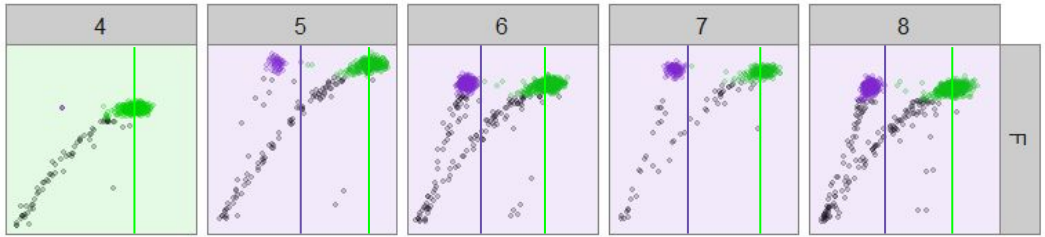
where *n* = total droplets,
r = mutant droplets,
p = 0.01 (1%)





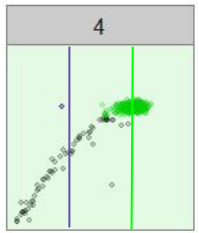
Step 6: Revisit gating of WT samples

- 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

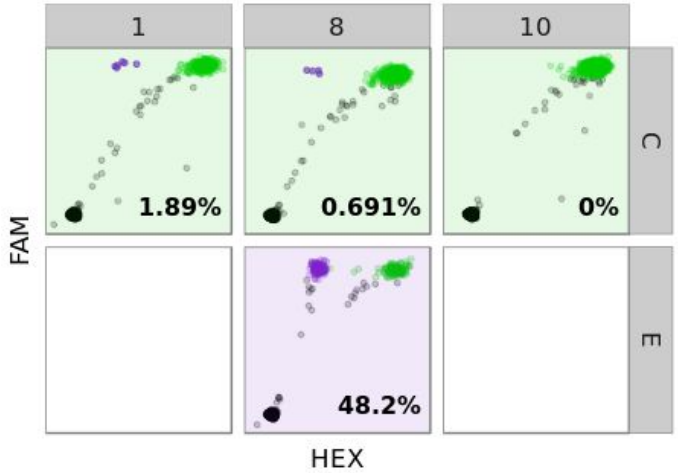


$\frac{\text{median(WT)}}{\text{max(MT)}}$ = 0.64 0.59 0.58 0.60

75-percentile of ratios: 0.60



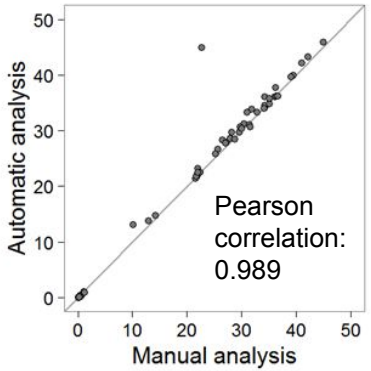
Results: Horizon samples



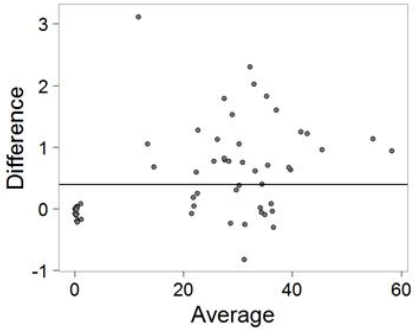
True MT freq	Calculate MT freq
1.4	1.89
0.8	0.691
0	0
50	48.2

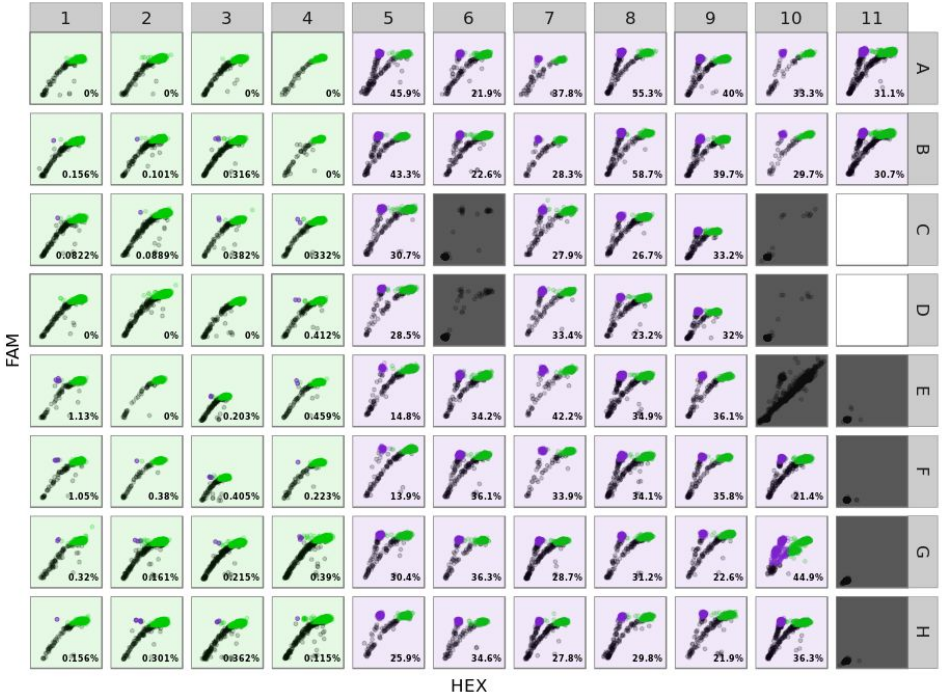
Well	Sample	Drops	Success	Drops outlier	Drops empty	Drops non empty	Drops empty fraction	Concentration	Significant mutant cluster	Mutant num	Wildtype num	Mutant freq
C01	brafv600e	12,863	TRUE	0	12,518	345	0.973	30	FALSE	6	311	1.89
C08	brafv600k	11,952	TRUE	0	11,340	612	0.949	57	FALSE	4	575	0.691
C10	WT	14,627	TRUE	0	13,485	1,142	0.922	89	FALSE	0	1,113	0
E08	brafv600r	10,979	TRUE	0	10,643	336	0.969	34	TRUE	151	162	48.2

Results: CRC-41



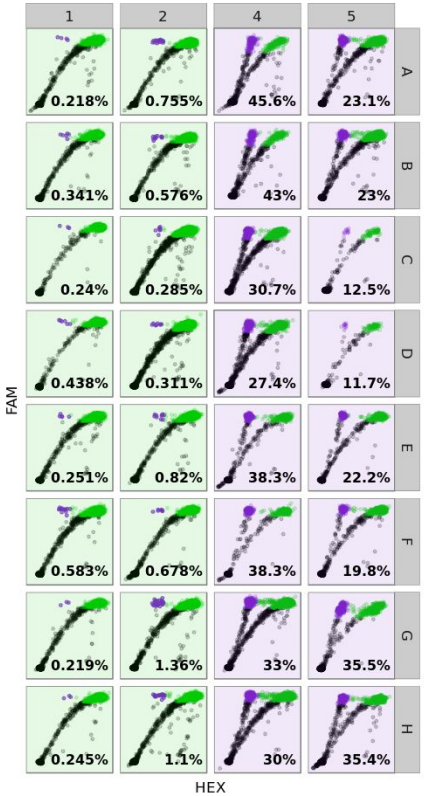
Mean difference: 0.39
TOST pvalue: 4.3x10^-8

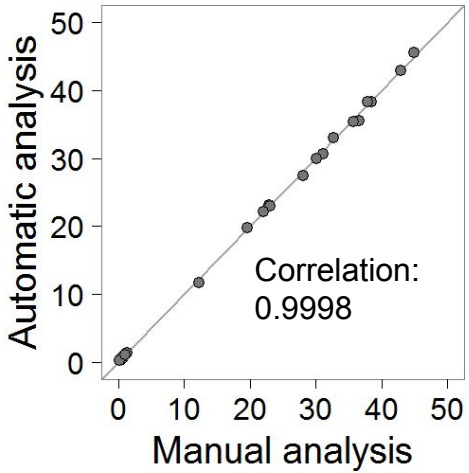


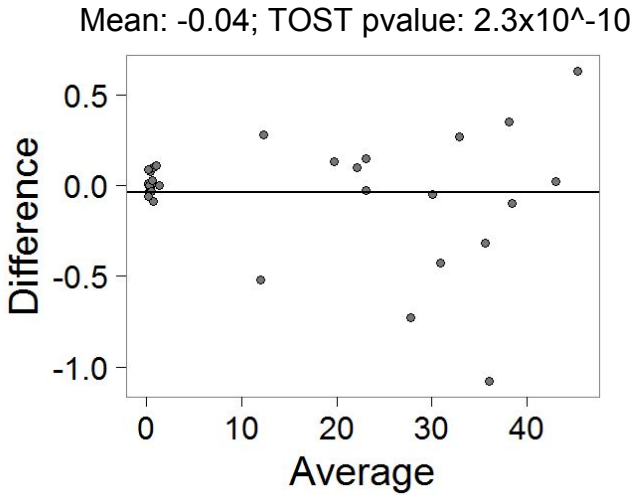


*Bottom right corner of plate was not sealed properly, hence samples there have more rain and lower quality

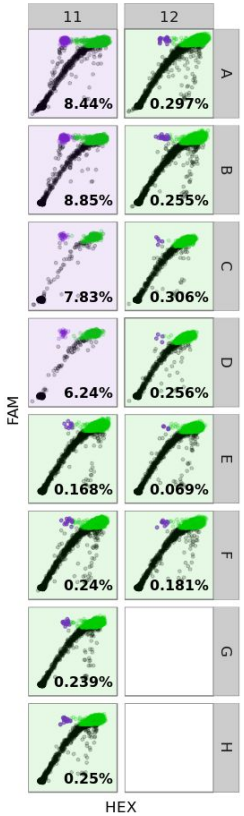
Results: CRC repeat

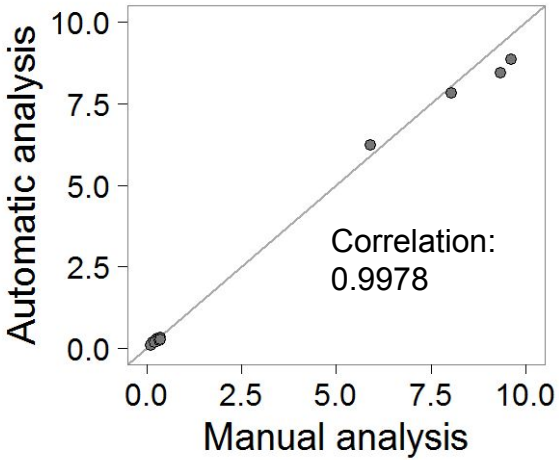


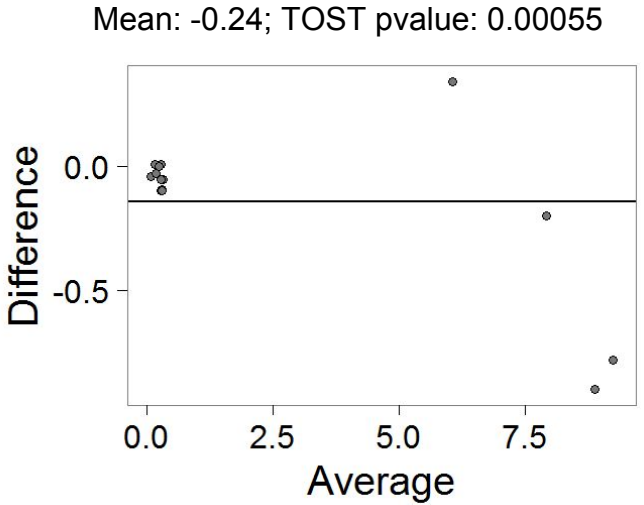




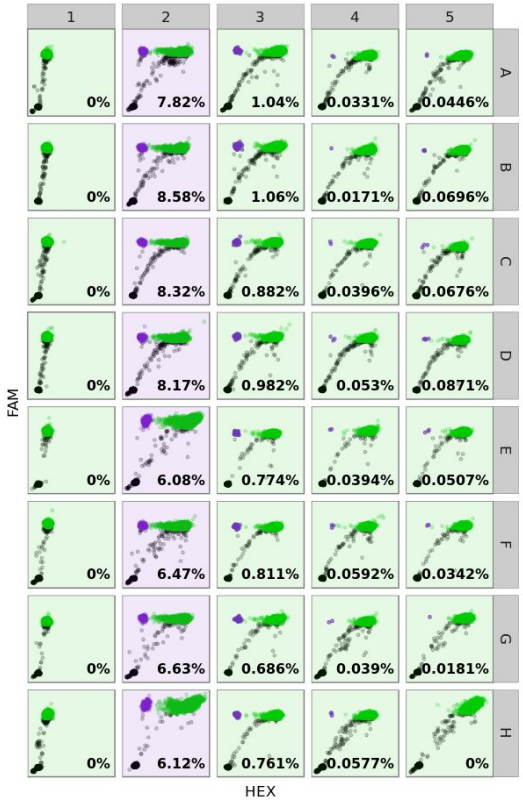
Results: Thyroid cancer dataset







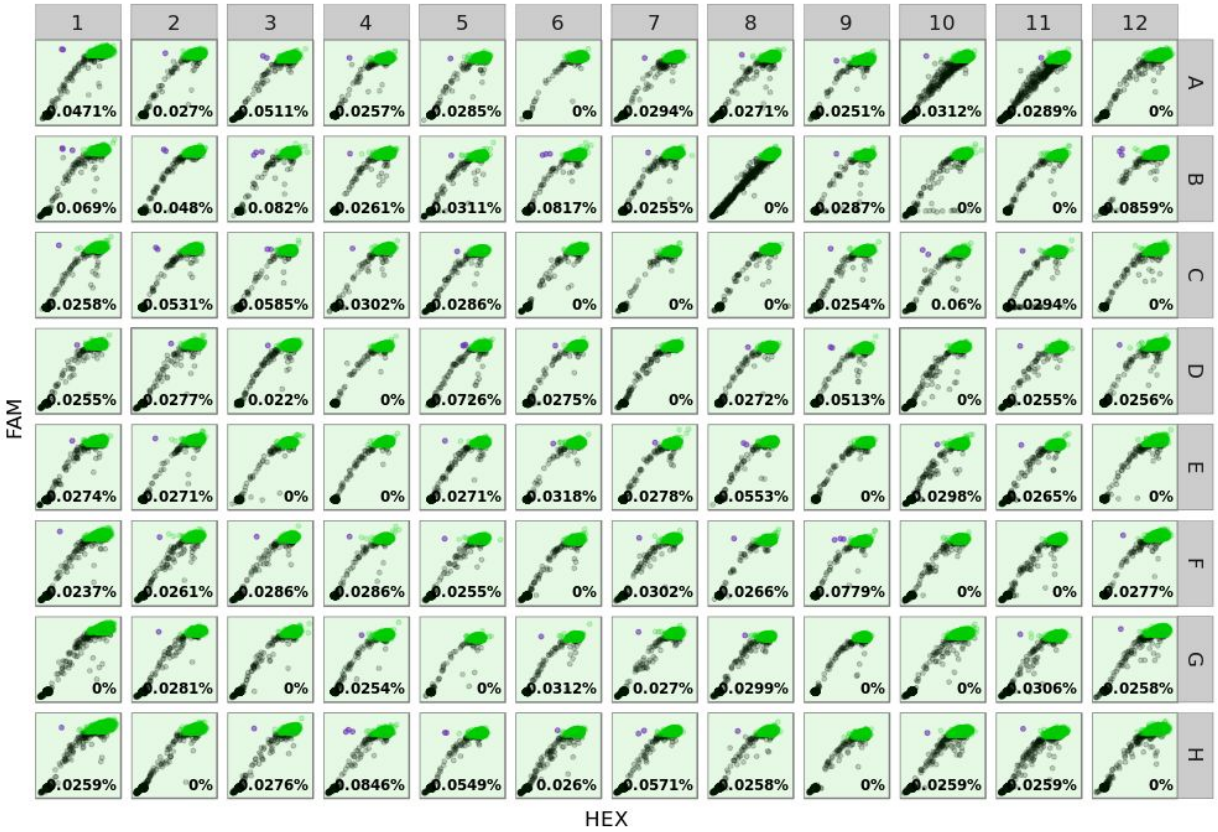
Results: Plasmid & YUMAC cell line



sample	replicates	wells	mean	StdErr
10%YUMAC	4	E02,F02,G02,H02	6.325	0.134
10%K	4	A02,B02,C02,D02	8.223	0.159
1%YUMAC	4	E03,F03,G03,H03	0.758	0.026
1%K	4	A03,B03,C03,D03	0.991	0.04
0.1%YUMAC	4	E04,F04,G04,H04	0.049	0.006
0.1%K	4	A04,B04,C04,D04	0.036	0.007
0.05%YUMAC	4	E05,F05,G05,H05	0.026	0.011
0.05%K	4	A05,B05,C05,D05	0.067	0.009

*K = dilutions from plasmid with 100% BRAF-V600K
**Dilutions in both plasmid and cell line were not accurate

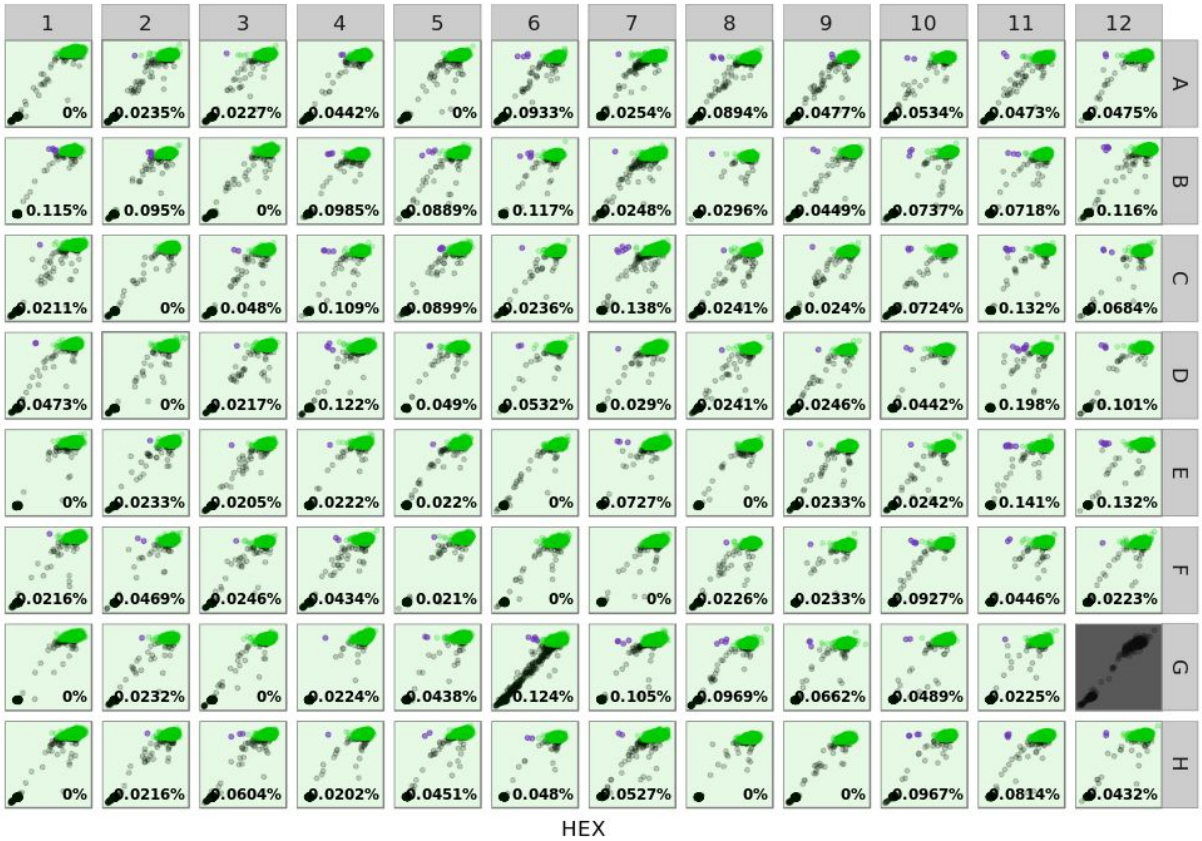
Results: plasmid brafv600k



sample	reps	wells	mean	StdErr
K-0.025%	48	A01:H06	0.03	0.003
0.01%K	48	A07:H12	0.022	0.003

*K = dilutions from plasmid with 100% BRAF-V600K

Results: YUMAC

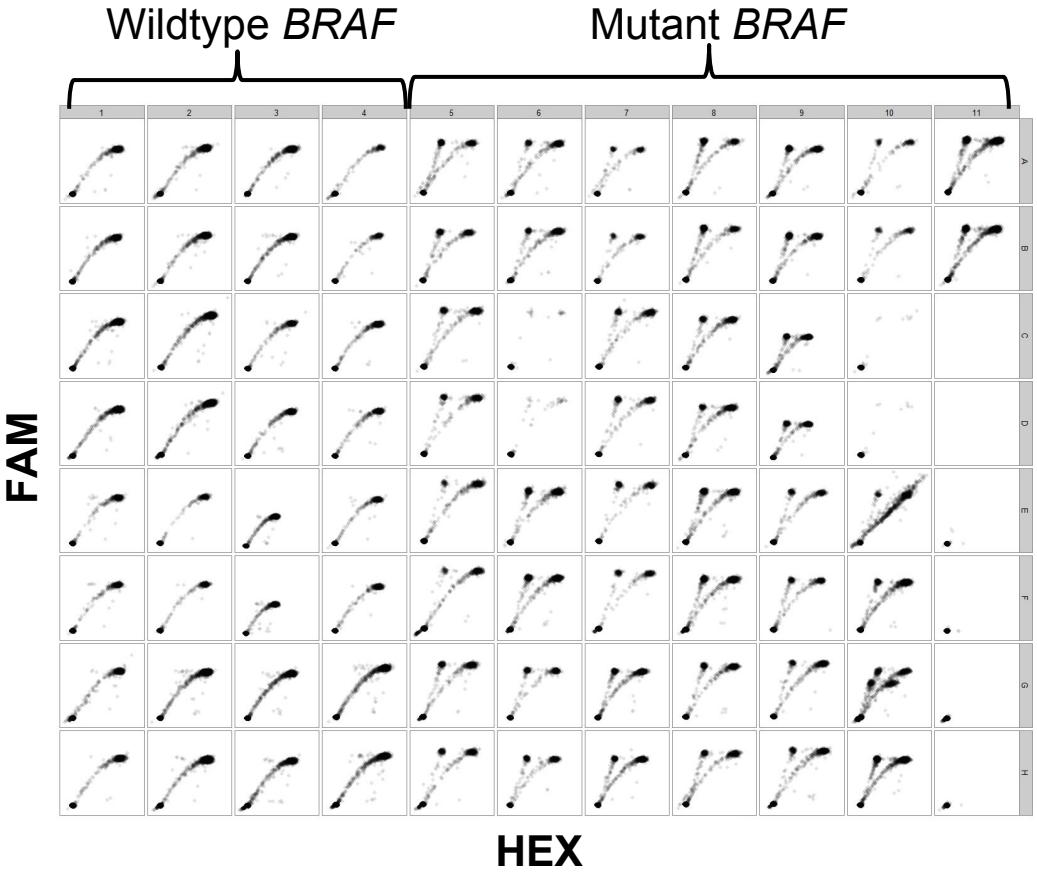


sample	reps	wells	mean	StdErr
WTOnly-Yumac-LOD	24	A01:H03	0.027	0.006
0.05%-Yumac-LOD	24	A10:H12	0.077	0.009
0.025%-Yumac-LOD	24	A07:H09	0.041	0.007
0.01%-Yumac-LOD	24	H04:H06	0.054	0.008

*Dilutions from YUMAC cell line with 100% BRAF-V600K

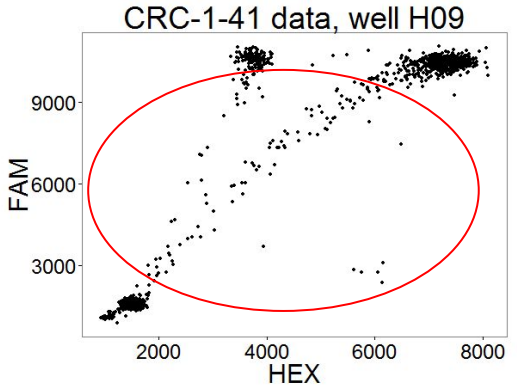
**Cell line data seems much cleaner than plasmid, perhaps due to supercoiling of DNA in plasmid

Dataset: FFPE from 41 CRC Patients

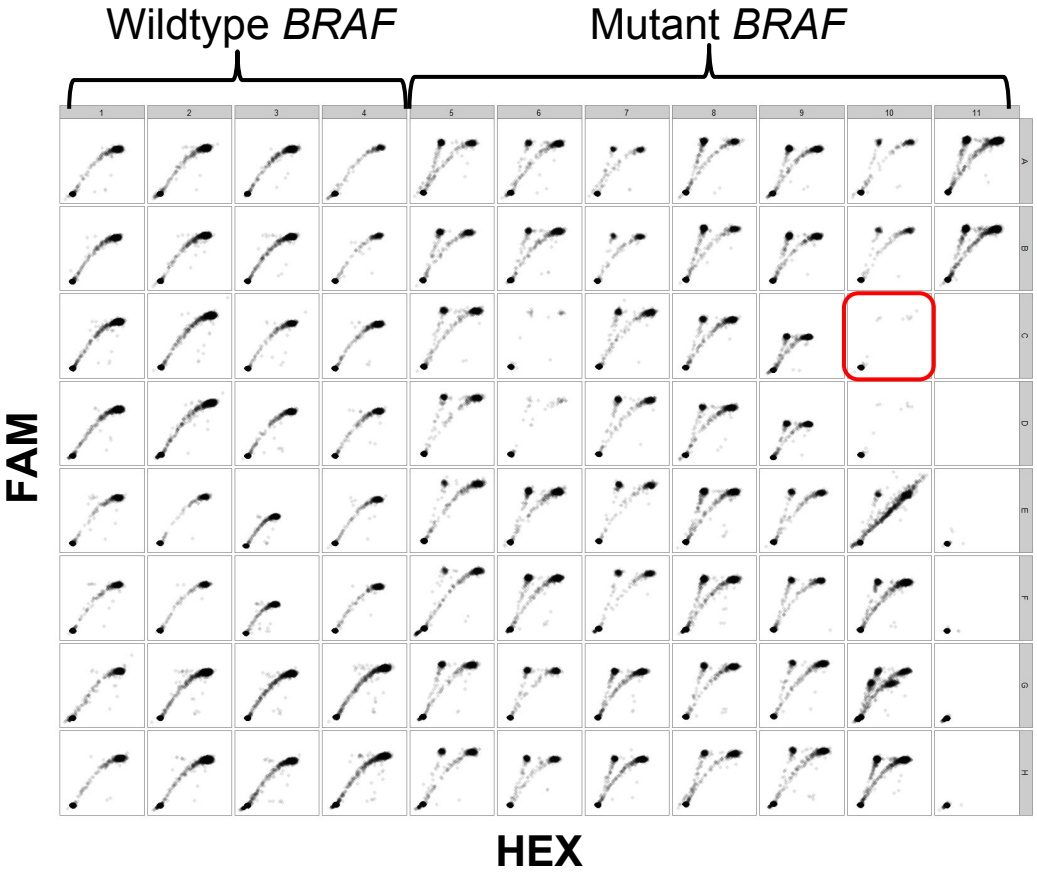


Factors to consider:

- Rain



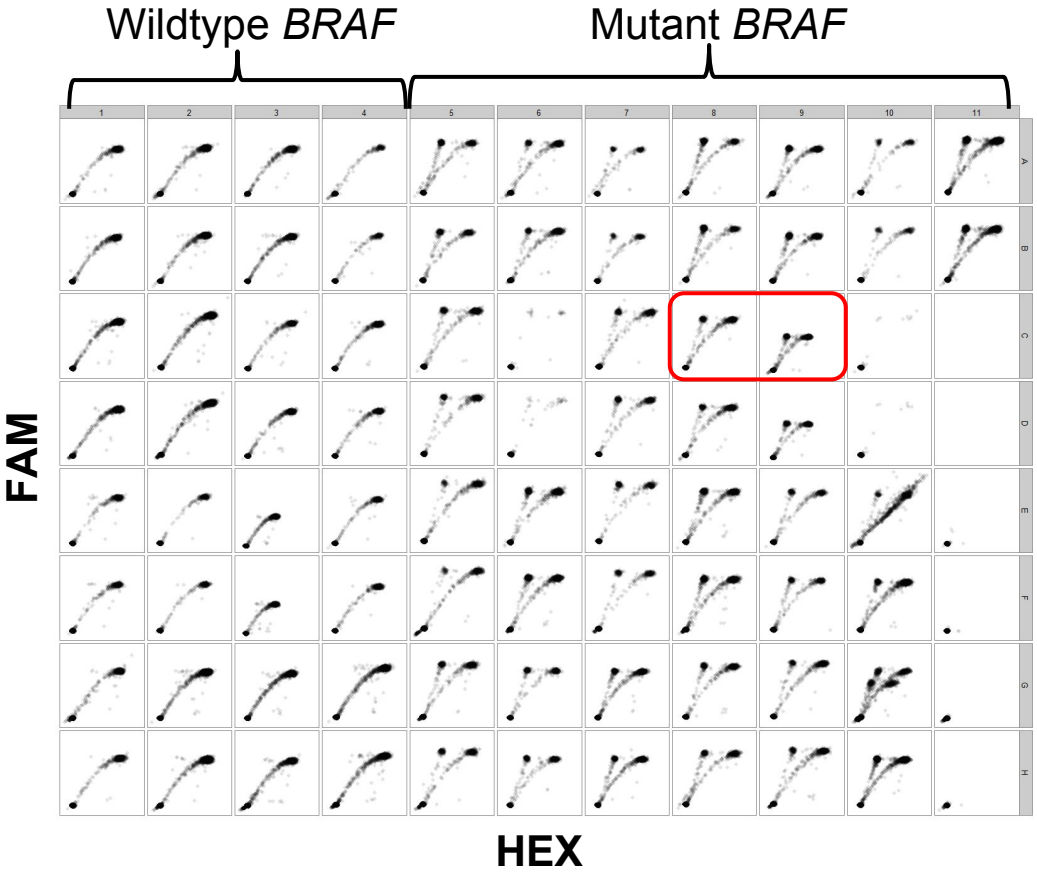
Dataset: FFPE from 41 CRC Patients



Factors to consider:

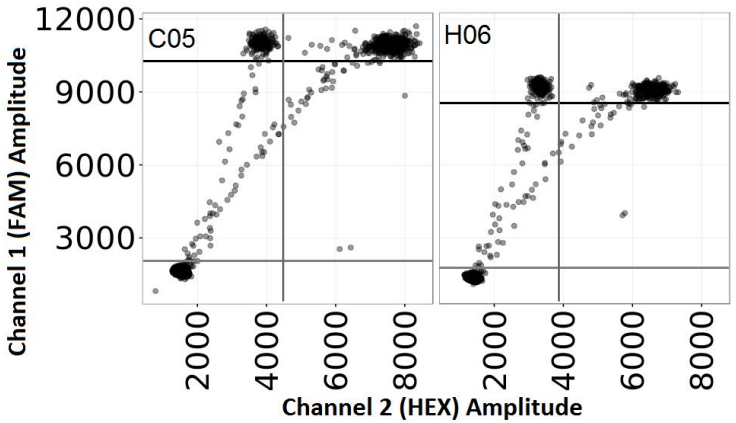
- Failed runs (e.g. C10)

Dataset: FFPE from 41 CRC Patients

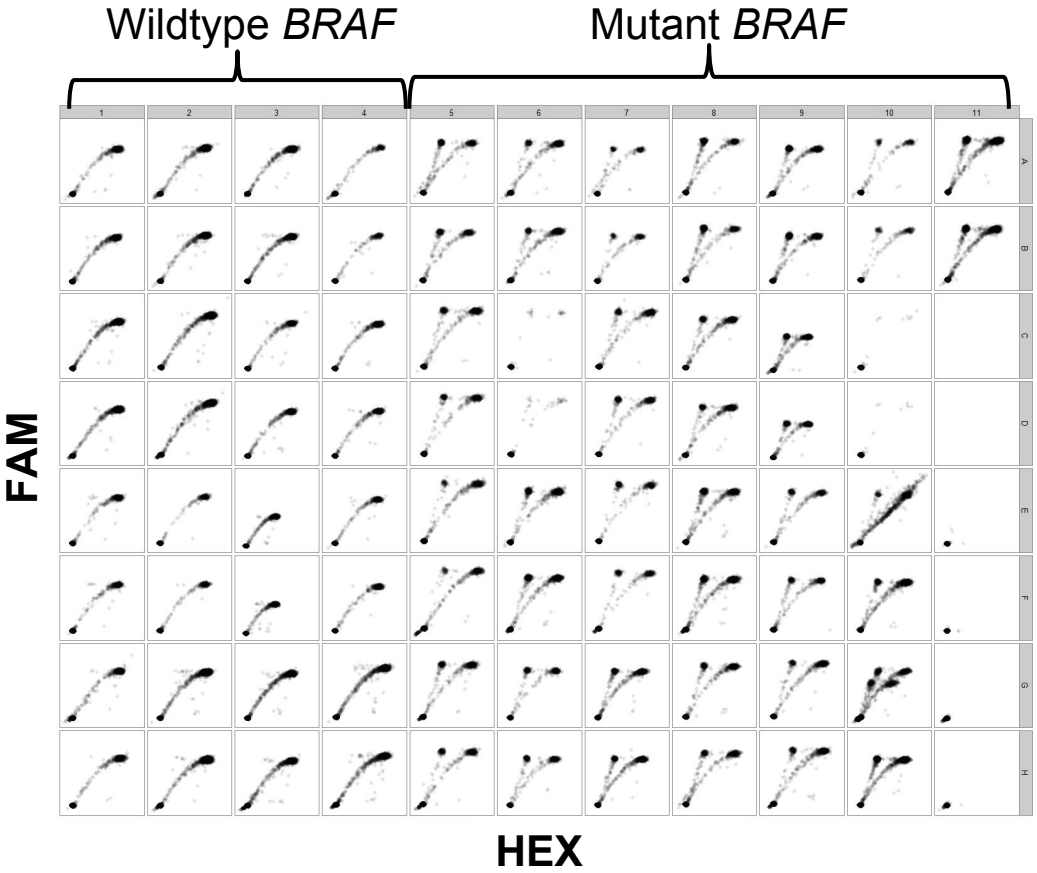


Factors to consider:

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

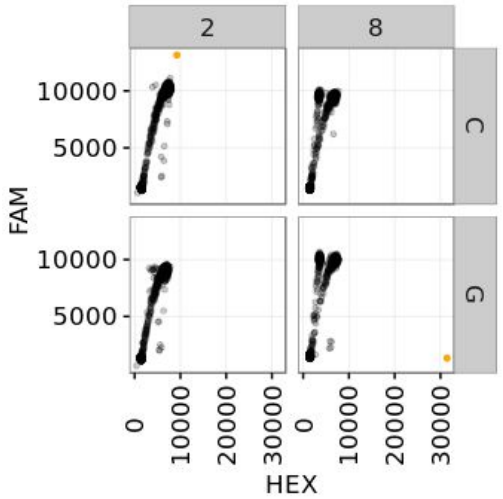


Dataset: FFPE from 41 CRC Patients

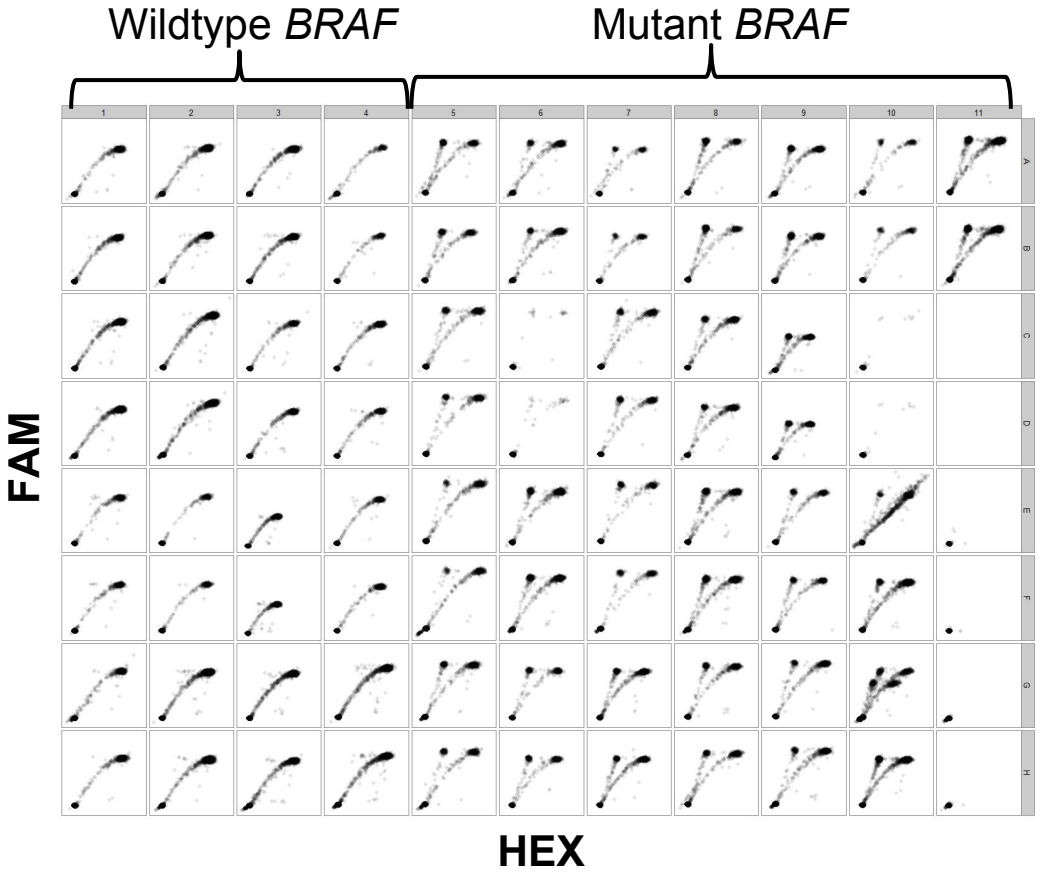


Factors to consider:

- Outliers



Dataset: FFPE from 41 CRC Patients



Factors to consider:

- Most droplets are empty

F06; alpha = 0.05