References

Jeff (CHI-HSUAN HO)

Genotyping (GT), APT, CPT

Jeff (CHI-HSUAN HO)

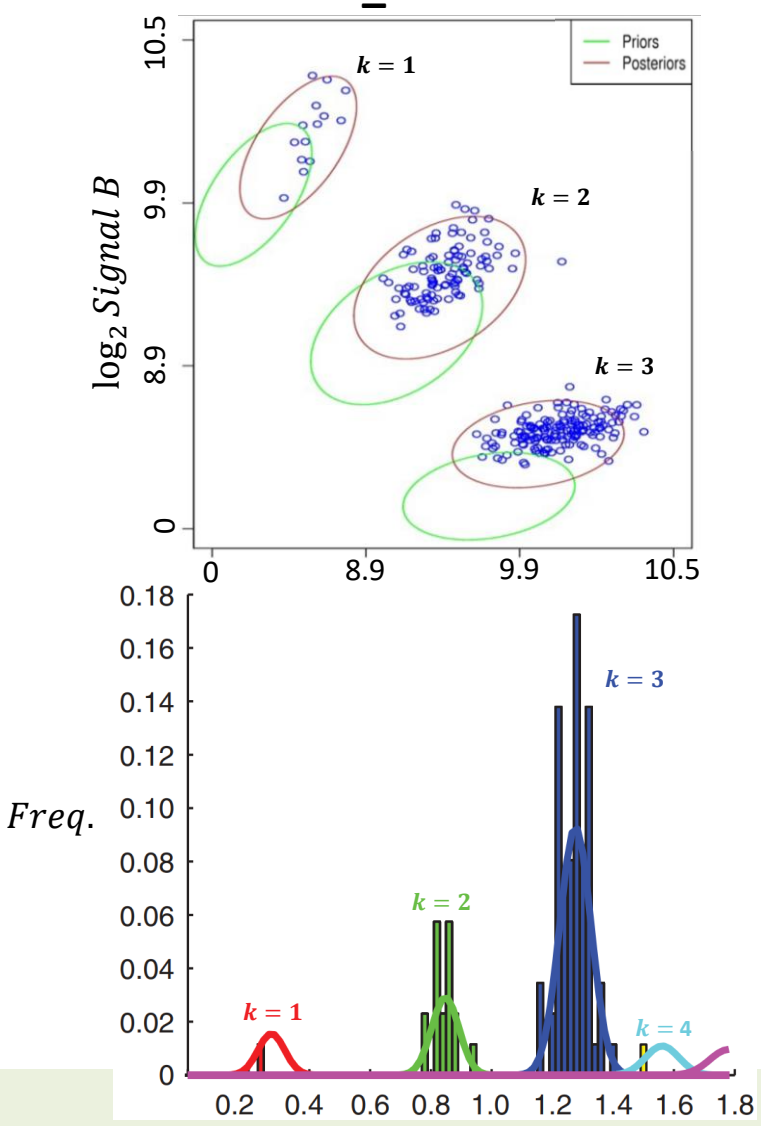
- Genotyping Procedure Documents for Each Chip

Algorithm	Array Type
BRLMM	Human Mapping 100K Array Human Mapping 500K Array
BRLMM-P	Genome-Wide Human SNP Array 5.0 Rat and Mouse Arrays
Birdseed v1 or Birdseed v2	Genome-Wide Human SNP Array 6.0

Axiom GT1 (BRLMM-P)	Axiom Arrays, including:
	• Axiom Human Arrays:
	  • Axiom Genome-Wide Human Arrays
	  • Axiom Genome-Wide CEU 1 Array
	  • Axiom Genome-Wide ASI 1 Array
	  • Axiom Genome-Wide YRI 1 Array set
	  • Axiom myDesign Custom Arrays
	• Axiom Genome-Wide BOS 1 Array

GTC v4.2 P/N 702982 Rev. 3

Gaussian Mixture Model (GMM)

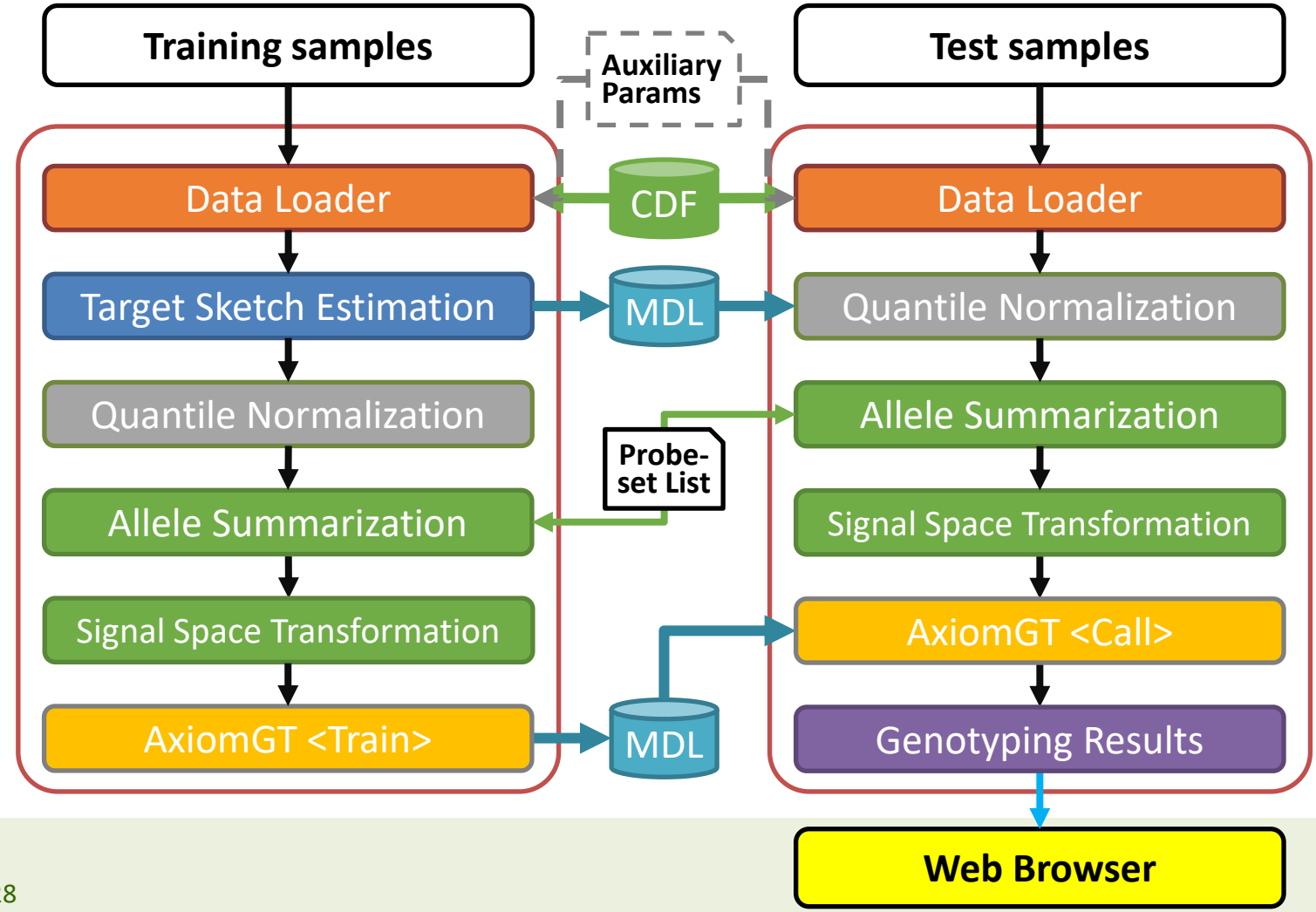


- Bayesian framework clustering model
  - Prior → A guess (e.g. from HapMap)
  - Posterior → A correction of cluster membership
- Applications (Genotyping, CNV analysis):
  - Birdseed (2-D)
  - Brlmm-P (1-D)
  - Canary (1-D)
- Model: $p(\boldsymbol{x}|\Theta) = \sum_{k=1}^{K} w_k \cdot N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$,

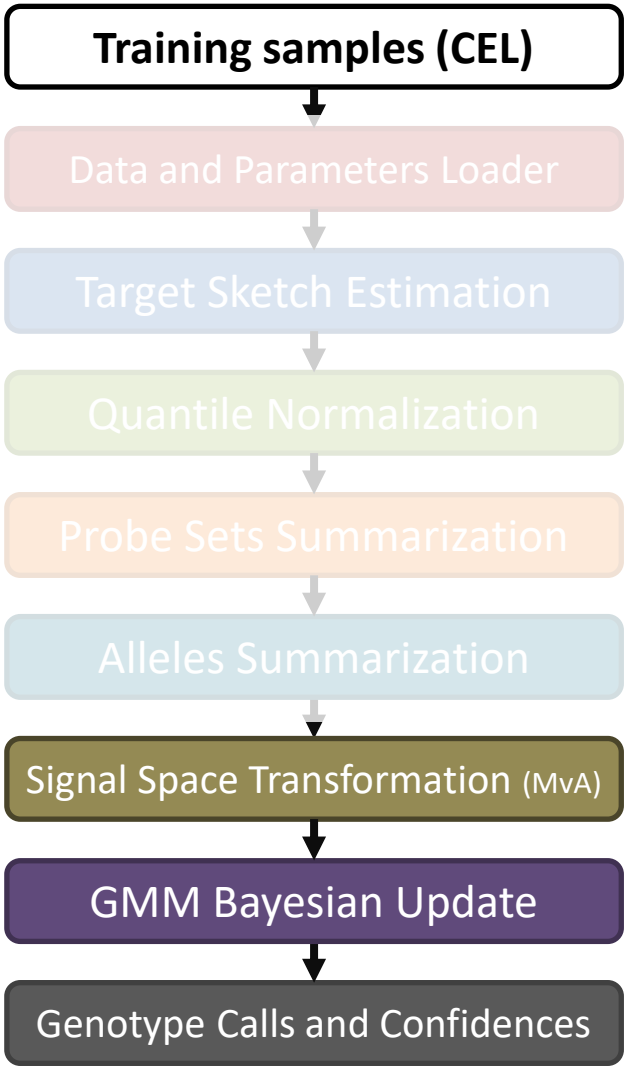
  where $N(\cdot)$ = Gaussian (Normal) dist., w_k= the k_{th} cluster proportion
- Evaluation (Model-based, Domain knowledge):
  - Bayesian Information Criterion (BIC)
  - Resolution of posterior cluster centroids
  - Model reasonability (e.g. outlier cluster)
  - Similarity between posterior and prior (e.g. w_k, $\boldsymbol{\mu}_k$)
  - Biological insight (e.g. Hardy-Weinberg penalty)

Genotyping Analysis Development and Deployment

- AxiomGT Framework

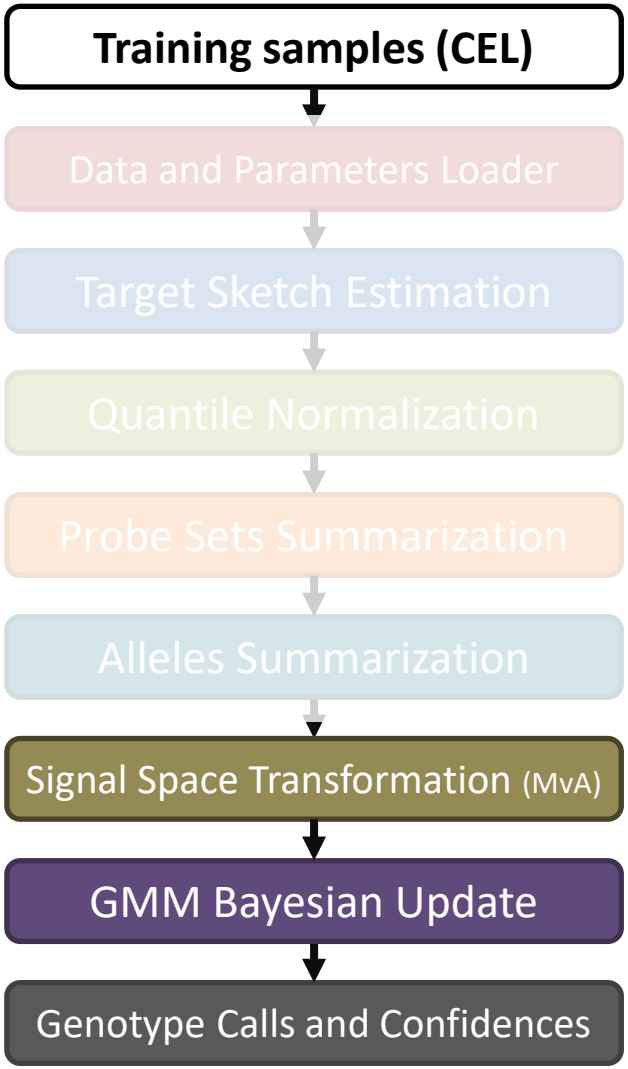


Genotyping Analysis Development and Deployment



- `computeEstimate`
  - AxiomGT1 Algo (`bayes_label`)
    - `initialize_bins`
      `(x -> bins domain)`
    - `Integratebrlmmoverlabelings (bins domain)`
    - `labels_two_posterior (bins domain)`
    - `make_two_calls`
      `(bins domain params results -> x calls)`

Genotyping Analysis Development and Deployment

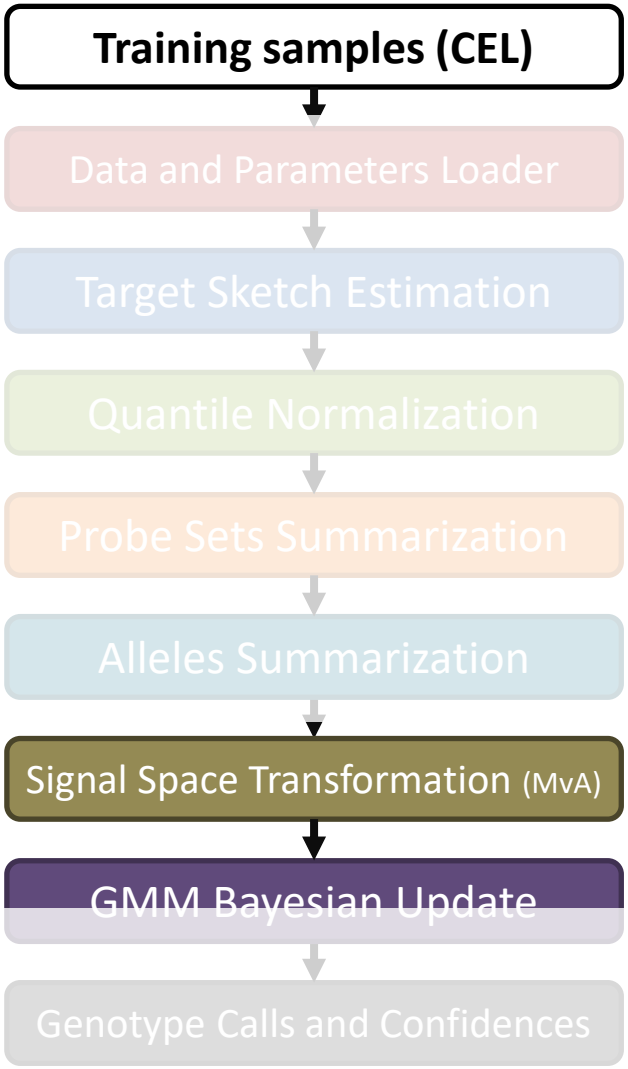


- initialize_bins
  - Sort (x, y) and other related params by x (increasing order).
  - Create bins data (setup_bins (x -> bins domain))

$$delta = \frac{Range(x)}{sp.bins + 1}, \qquad sp.bins = 100 \ (default)$$

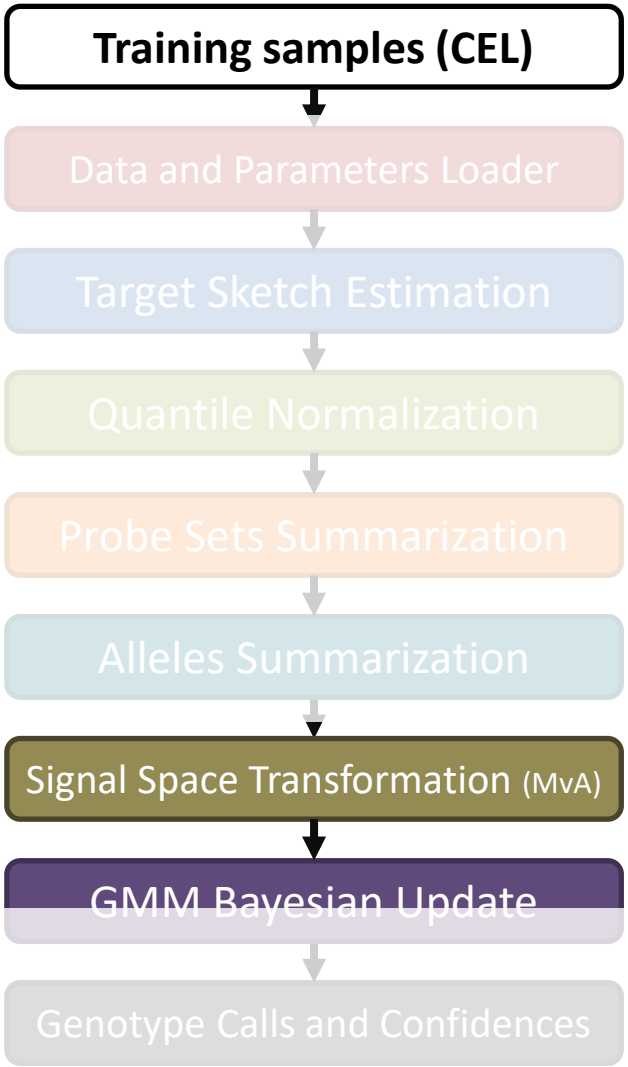
    - Create a new bin and reset collected statistics if
      1. When current data point x lies outside current $boundary$ (previous $x + delta$) $\Rightarrow$ Create new $boundary$ by using current x.
      2. sp.bins = 0 (bins is turned off)
      3. When data points are fewer than bins.
    - Statistics are collected and computed in the same bin:
      1. Data points number
      2. $x, x^2, y, y^2, x \cdot y$
      3. Penalty from each known genotype and inbreeding status (Only used in supervised mode)

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Compute Quality Score (Posterior Analog) (for each partition)**
    - `(sp.mix) mixture_penalty`
    - `(sp.bic) BIC (k*log(N) part)`
    - `1D (x) Log likelihood under posterior params`
    - `1D (x) Log prior probability of posterior params`
    - $\frac{1}{2} *$ `Quality Score Correction`
    - `(sp.CSepPen) Geman-McClure transformed FLD penalty for non-well-separated clusters cases.`
  - **Compute Relative Probability for (Each Partition) & (Each Data Point to Be Each Genotype) under Posterior Information.**

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Quality Score (Posterior Analog) (for each partition)**
    - **(sp.mix) mixture_penalty**

$$-\sum_{g=1}^{3} N_g * \log\left(\frac{(N_g + lambda)}{\sum_g (N_g + lambda)}\right), \qquad g = 1\,(BB), 2\,(AB), 3\,(AA), \qquad lambda = 1$$

    - **(sp.bic) BIC (k*log(N) part)**

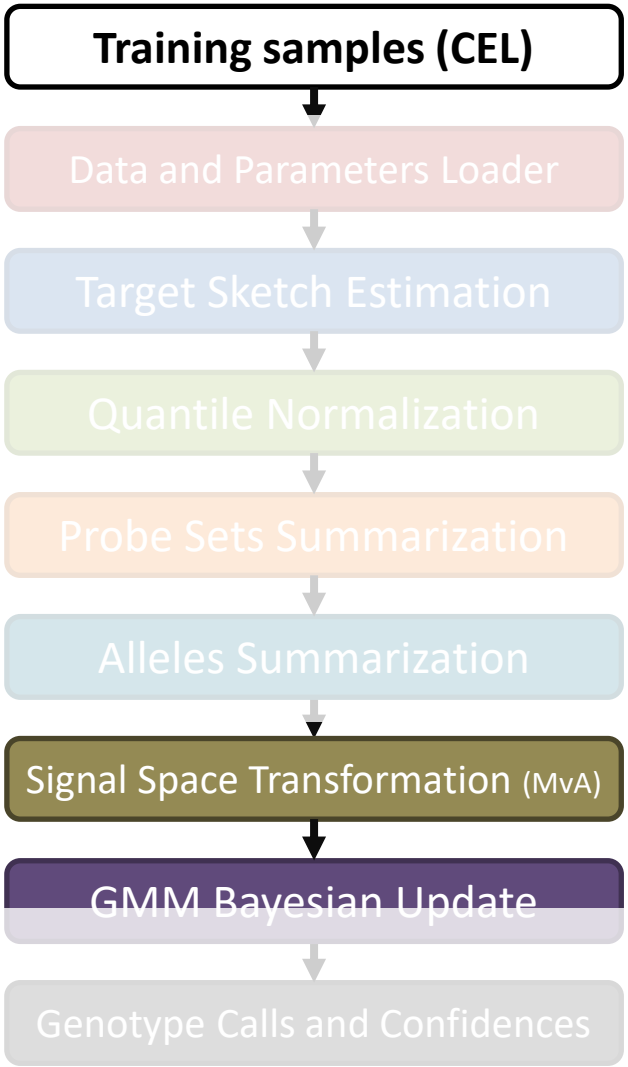
$$c * bic\_k * \log\left(\sum_g N_g\right), \qquad bic\_k = 2 \Rightarrow mean, var, \qquad c = 1,2,3$$

    - **1D (x) Log likelihood under posterior params**
      - $\mathrm{u}'_{3x1} = \left(\boldsymbol{K}_{\mathbf{0}\,3x3}^{-1} + \boldsymbol{N}'_{3x3}\right)^{-1} * \left(\boldsymbol{K}_{0\,3x3}^{-1} * \mathrm{u}_{0\,3x1} + \boldsymbol{N}_{3x3} * \mathrm{m}_{3x1}\right),$

$$\boldsymbol{K}_{0\,3x3}^{-1} = \begin{bmatrix} \frac{k_{10}}{\sigma_{10}^2} & \frac{\sigma_{120}}{\sigma_{10}\sigma_{20}} & \frac{\sigma_{130}}{\sigma_{10}\sigma_{30}} \\ \frac{\sigma_{120}}{\sigma_{10}\sigma_{20}} & \frac{k_{20}}{\sigma_{20}^2} & \frac{\sigma_{230}}{\sigma_{20}\sigma_{30}} \\ \frac{\sigma_{130}}{\sigma_{10}\sigma_{30}} & \frac{\sigma_{230}}{\sigma_{20}\sigma_{30}} & \frac{k_{30}}{\sigma_{30}^2} \end{bmatrix}, \qquad \boldsymbol{N}'_{3x3} = \begin{bmatrix} \frac{N_1}{\sigma_{10}^2} & 0 & 0 \\ 0 & \frac{N_2}{\sigma_{20}^2} & 0 \\ 0 & 0 & \frac{N_3}{\sigma_{30}^2} \end{bmatrix}$$

$$\mathrm{u}_{0\,3x1} = \begin{bmatrix} \mathrm{u}_{10} \\ \mathrm{u}_{20} \\ \mathrm{u}_{30} \end{bmatrix}, \qquad \mathrm{m}_{3x1} = \begin{bmatrix} \sum_i x_{1i} \\ \sum_i x_{2i} \\ \sum_i x_{3i} \end{bmatrix}, \qquad \boldsymbol{N}_{3x3} = \begin{bmatrix} \frac{1}{\sigma_{10}^2} & 0 & 0 \\ 0 & \frac{1}{\sigma_{20}^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_{30}^2} \end{bmatrix}$$

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Quality Score (Posterior Analog) (for each partition)**
    - **`1D (x) Log likelihood under posterior params`**
      - $u'_{3x1} = \left(\boldsymbol{K_0}^{-1}_{3x3} + \boldsymbol{N'}_{3x3}\right)^{-1} * \left(\boldsymbol{K_0}^{-1}_{3x3} * u_{0\,3x1} + \boldsymbol{N}_{3x3} * m_{3x1}\right),$
      - `(sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)).` u'_{3x1}

$$w_g = N_g + k_{g0}, \qquad g = 1,2,3$$

$$gamma = delta * \frac{w_1 - w_3}{w_1 + w_2 + w_3}$$

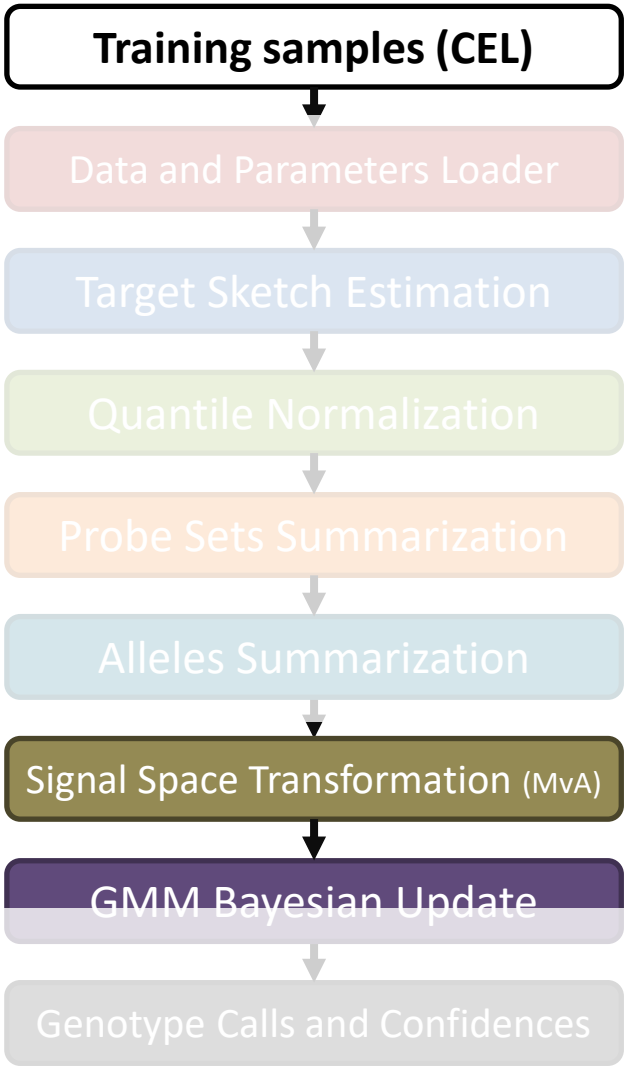
$$u'_{3x1} = \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \end{bmatrix}, \quad \begin{aligned} u'_1 &= u'_1 + delta - gamma \\ u'_2 &= u'_2 - gamma \\ u'_3 &= u'_3 - delta - gamma \end{aligned}$$

`Pool Adjacent-Violators (PAV) algo.`

$$\begin{cases} u'_g, u'_{g+1}\ , & u'_g \le u'_{g+1} \\ u'_g, u'_{g+1} = \dfrac{\sum_{g \in A} w_g * u'_g}{\sum_{g \in A} w_g}, & A = \{g | u'_g > u'_{g+1}\} \end{cases}, g = 1,2,3$$

$$u'_{3x1} = \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \end{bmatrix}, \quad \begin{aligned} u'_1 &= u'_1 - delta + gamma \\ u'_2 &= u'_2 + gamma \\ u'_3 &= u'_3 + delta + gamma \end{aligned}$$

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Quality Score (Posterior Analog) (for each partition)**
    - **`1D (x) Log likelihood under posterior params`**
      - $u'_{3x1} = \left(\boldsymbol{K_0}^{-1}_{3x3} + \boldsymbol{N'}_{3x3}\right)^{-1} * \left(\boldsymbol{K_0}^{-1}_{3x3} * u_{0\,3x1} + \boldsymbol{N}_{3x3} * m_{3x1}\right),$
      - `(sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)).` u'_{3x1}
      - $\sigma'^{2}_{g} = \frac{v_{g0} * \sigma^2_{g0} + \sum_i (x_{gi} - \bar{x}_g)^2 + \frac{k_g * N_g}{k_g + N_g} * (u'_g - u_{g0})^2}{v_{g0} + N_g} \Rightarrow$
        $\frac{v_{g0} * \sigma^2_{g0} + \sum_i x^2_{gi} - \sum_i x_{gi} * \sum_i x_{gi} * \frac{1}{N_g + 0.0001} + \frac{k_g * N_g}{k_g + N_g} * (u'_g - u_{g0})^2}{v_{g0} + N_g}, \quad g = 1, 2, 3$
      - `(sp.comvar) Ad-hoc shrinkage for` σ'^{2}_{g} `of each cluster (controlled by mixing proportion (lambda) (1)).`

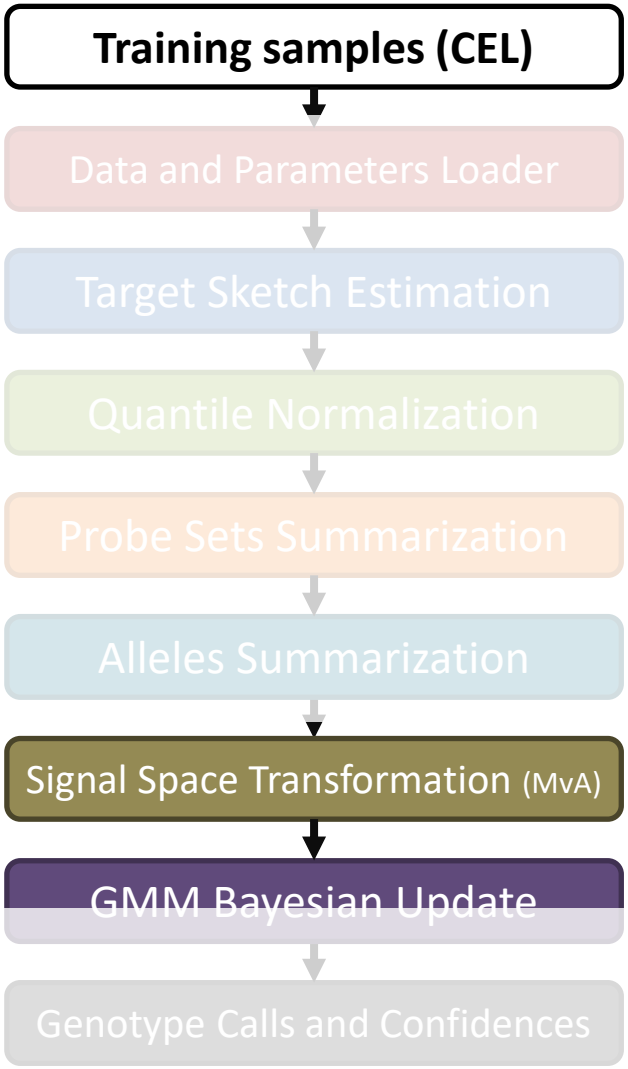
        `Adjusted Pooled Variance.`

        $w_g = N_g + v_{g0}, \quad g = 1, 2, 3$

        $\sigma'^{2}_{t} = \frac{\sum_g w_g * \sigma'^{2}_{g}}{\sum_g w_g}, \quad t = 1, 2, 3,$

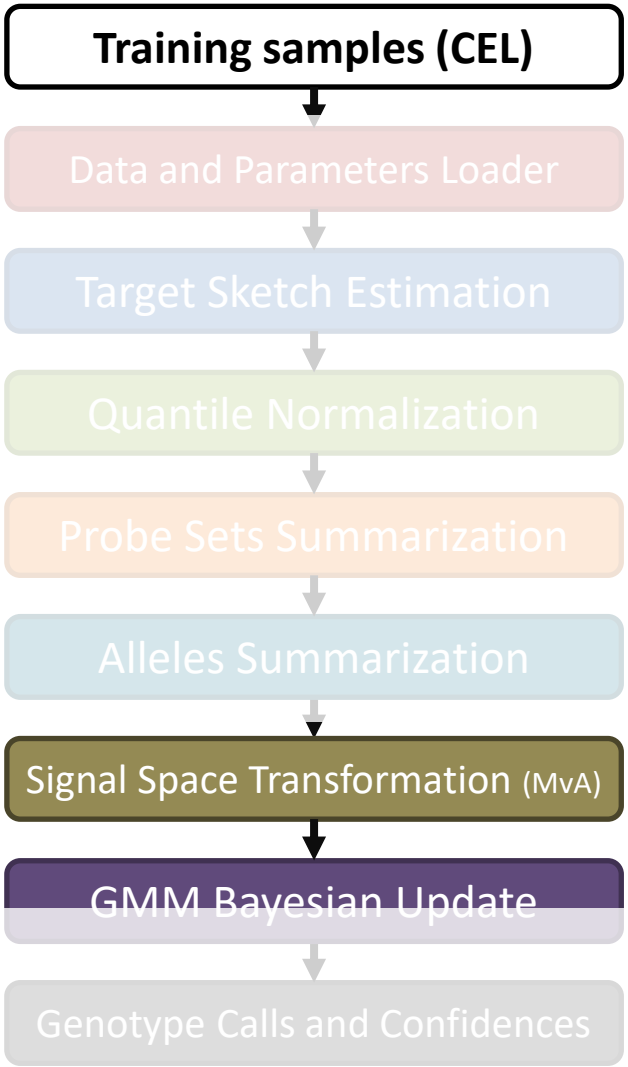
        $\Rightarrow \frac{(3 - 2 * lambda) * w_t * \sigma'^{2}_{t} + \sum_{g \neq t} lambda * w_g * \sigma'^{2}_{g}}{(3 - 2 * lambda) * w_t + \sum_{g \neq t} lambda * w_g}$

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Quality Score (Posterior Analog) (for each partition)**
    - **1D (x) Log likelihood under posterior params**
      - $\mathrm{u}'_{3\mathrm{x}1} = \left(\boldsymbol{K}_{\mathbf{0}\,3\mathrm{x}3}^{-1} + \boldsymbol{N}'_{3\mathrm{x}3}\right)^{-1} * \left(\boldsymbol{K}_{0\,3\mathrm{x}3}^{-1} * \mathrm{u}_{0\,3\mathrm{x}1} + \boldsymbol{N}_{3\mathrm{x}3} * \mathrm{m}_{3\mathrm{x}1}\right),$
      - `(sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)).` $\mathrm{u}'_{3\mathrm{x}1}$
      - $\sigma'^{2}_{g} = \frac{v_0 * \sigma_{g0}^2 + \sum_i (x_{gi} - \bar{x}_g)^2 + \frac{k_g * N_g}{k_g + N_g} * (\mathrm{u}'_g - \mathrm{u}_{g0})^2}{v_0 + N_g}, \quad g = 1, 2, 3$
      - `(sp.comvar) Ad-hoc shrinkage for` σ'^{2}_{g} `of each cluster (controlled by mixing proportion (lambda) (1)).`
      - $\ell = \log \prod_g \prod_{i=1}^{N_g} N\left(\mathrm{u}'_g,\ \sigma'^{2}_{g}\right) = \sum_g \sum_{i=1}^{N_g} \log\left(N\left(\mathrm{u}'_g,\ \sigma'^{2}_{g}\right)\right) \Rightarrow$
        $-\frac{1}{2}\left[\sum_g N_g \log\left(\sigma'^{2}_{g}\right) + \frac{1}{\sigma'^{2}_{g}}\left(\sum_{i=1}^{N_g} x_i^2 - 2\mathrm{u}'_g \sum_{i=1}^{N_g} x_i + N_g \mathrm{u}'^{2}_{g}\right)\right]$
        $\Rightarrow -2 * \ell$
        $= \sum_g N_g \log\left(\sigma'^{2}_{g}\right) + \frac{1}{\sigma'^{2}_{g}}\left(\sum_{i=1}^{N_g} x_i^2 - 2\mathrm{u}'_g \sum_{i=1}^{N_g} x_i + N_g \mathrm{u}'^{2}_{g}\right)$

Genotyping Analysis Development and Deployment



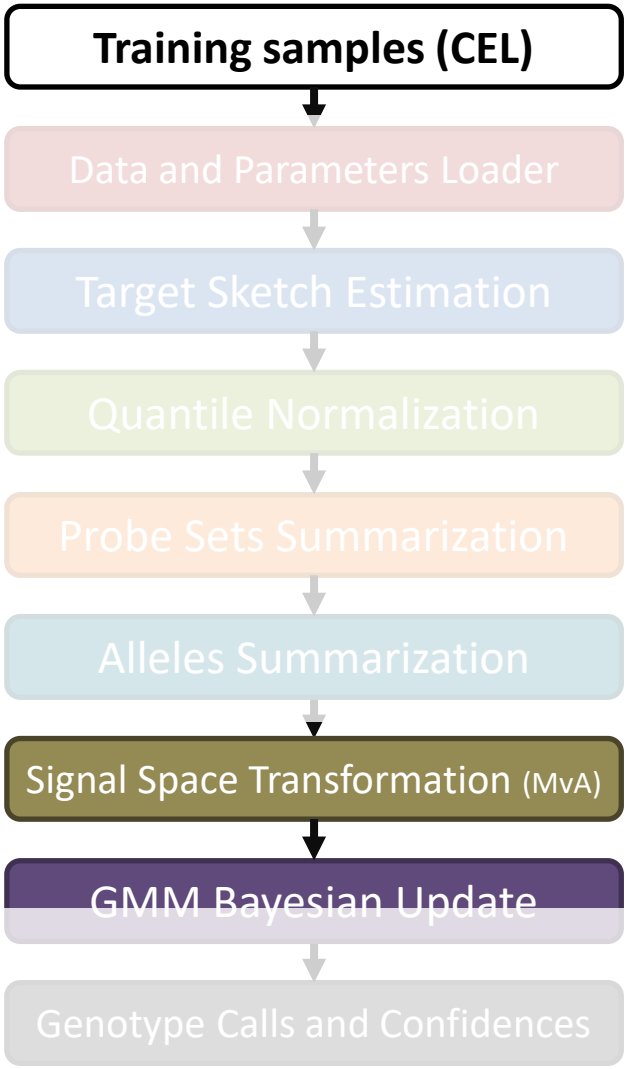
- `Integratebrlmmoverlabelings`
  - **Quality Score (Posterior Analog) (for each partition)**
    - **1D (x) Log prior probability of posterior params**
      - $\ell = \log \prod_g N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right) = \sum_g \log\left(N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right)\right) \Rightarrow -\frac{1}{2}\left[\sum_g \log\left(\frac{\sigma_{g0}^2}{k_{g0}}\right) + \frac{k_{g0}}{\sigma_{g0}^2}\left(u'_g - u_{g0}\right)^2\right]$

        $\Rightarrow -2 * \ell = \sum_g \log\left(\frac{\sigma_{g0}^2}{k_{g0}}\right) + \frac{k_{g0}}{\sigma_{g0}^2}\left(u'_g - u_{g0}\right)^2$
      - $\ell = \log \prod_g IG\left(v_0, \sigma_{g0}^2\right) \Rightarrow -\ell = \sum_g \frac{\sigma_{g0}^2}{\sigma'^2_g} + (v_0 + 1) * \log\left(\sigma'^2_g\right)$
    - $\frac{1}{2} *$ **Quality Score**
    - **(sp.CSepPen) Geman-McClure transformed FLD penalty for non-well-separated clusters.**
      - $-CSepPen * \sum_{i,j,i\neq j} FLD'_{ij}, \; i,j \in g = \{1,2,3\},$

        $$FLD'_{ij} = \begin{cases} \frac{FLD_{ij}}{1 + \frac{FLD_{ij}}{CSepThr}} * (N_i + N_j), & other \\ \frac{FLD_{ij}}{1 + \frac{FLD_{ij}}{2 * CSepThr}} * (N_1 + N_3), & i = 1, j = 3 \end{cases},$$

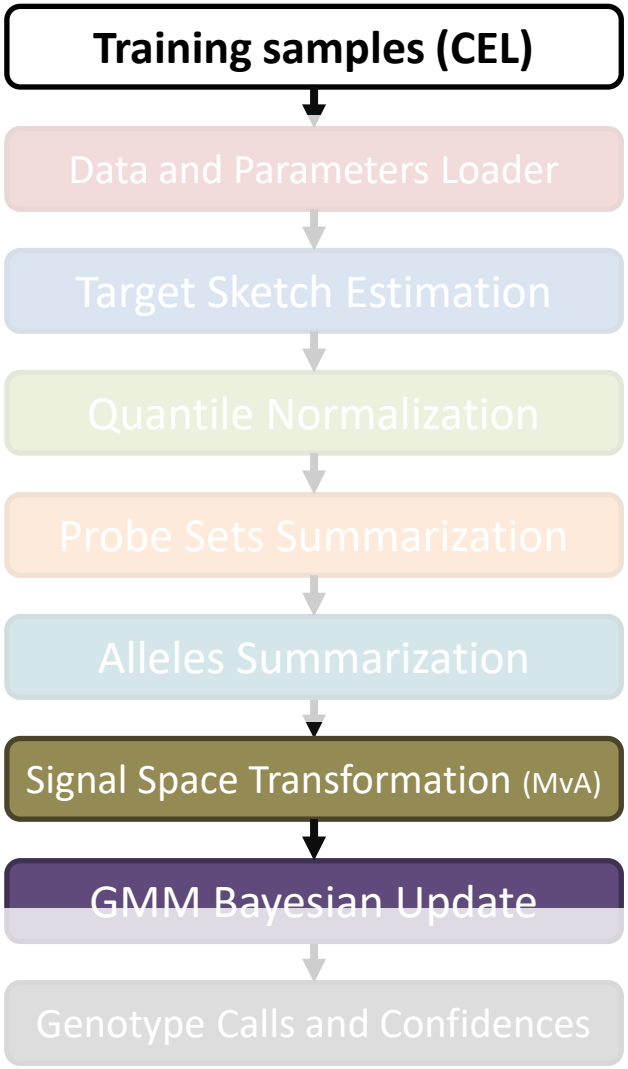
        $$FLD_{ij} = FLD_{ji} = \frac{\left(u'_i - u'_j\right)^2}{\sigma'^2_i + \sigma'^2_j}, \quad CSepPen = 0.1, \quad CSepThr = 4$$

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Relative Probability for Each Partition under Posterior Information.**
    - Quality Score $Q_{i,j}$ for partition i, j
    - Relative Probability $q_{i,j} = \frac{\exp(-Q_{i,j})}{\exp(-\min Q_{i,j})} = \exp\left(\min Q_{i,j} - Q_{i,j}\right)$
      $= \frac{Posterior\ Probability\ of\ Specified\ Partition\ (i,j)}{Maximal\ Posterior\ Probability}$

Genotyping Analysis Development and Deployment



- `Integratebrlmmoverlabelings`
  - **Relative Probability for Each Data Point to Be Each Genotype under Posterior Information after dividing q…**

$q_{i,j}$

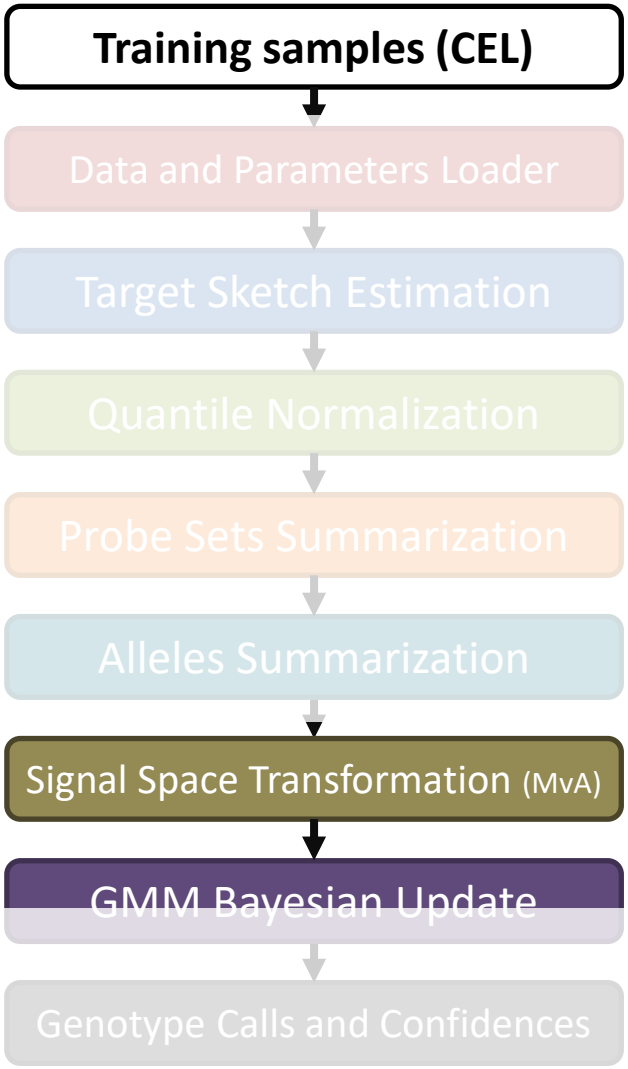
i \ j	0	1	2	3	4	$q_{i\cdot}$	$\sum_i q_{i\cdot}$	$q_{\cdot\cdot} - \sum_i q_{i\cdot}$
0	cccc	bccc	bbcc	bbbc	bbbb	$q_{0\cdot}$	$\sum_{i=0}^{0} q_{i\cdot}$	$\sum_{i=1}^{4} q_{i\cdot}$
1		accc	abcc	abbc	abbb	$q_{1\cdot}$	$\sum_{i=0}^{1} q_{i\cdot}$	$\sum_{i=2}^{4} q_{i\cdot}$
2			aacc	aabc	aabb	$q_{2\cdot}$	$\sum_{i=0}^{2} q_{i\cdot}$	$\sum_{i=3}^{4} q_{i\cdot}$
3				aaac	aaab	$q_{3\cdot}$	$\sum_{i=0}^{3} q_{i\cdot}$	$\sum_{i=4}^{4} q_{i\cdot}$
4					aaaa	$q_{4\cdot}$		
$q_{\cdot j}$	$q_{\cdot 0}$	$q_{\cdot 1}$	$q_{\cdot 2}$	$q_{\cdot 3}$	$q_{\cdot 4}$	$q_{\cdot\cdot}$		
$\sum_j q_{\cdot j}$	$\sum_{j=0}^{0} q_{\cdot j}$	$\sum_{j=0}^{1} q_{\cdot j}$	$\sum_{j=0}^{2} q_{\cdot j}$	$\sum_{j=0}^{3} q_{\cdot j}$				

The relative counts of Sth data point being genotype "a" after observing all data.

The relative counts of Sth data point being genotype "c" after observing all data.

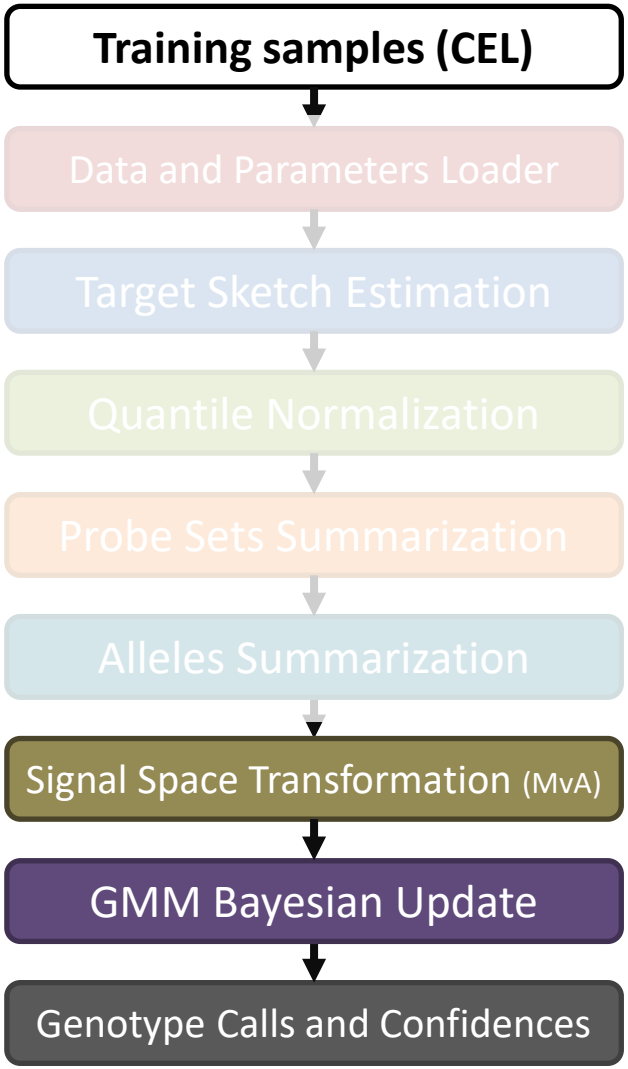
Let "a" ≡ BB genotype, "b" ≡ AB genotype, "c" ≡ AA genotype.

Genotyping Analysis Development and Deployment



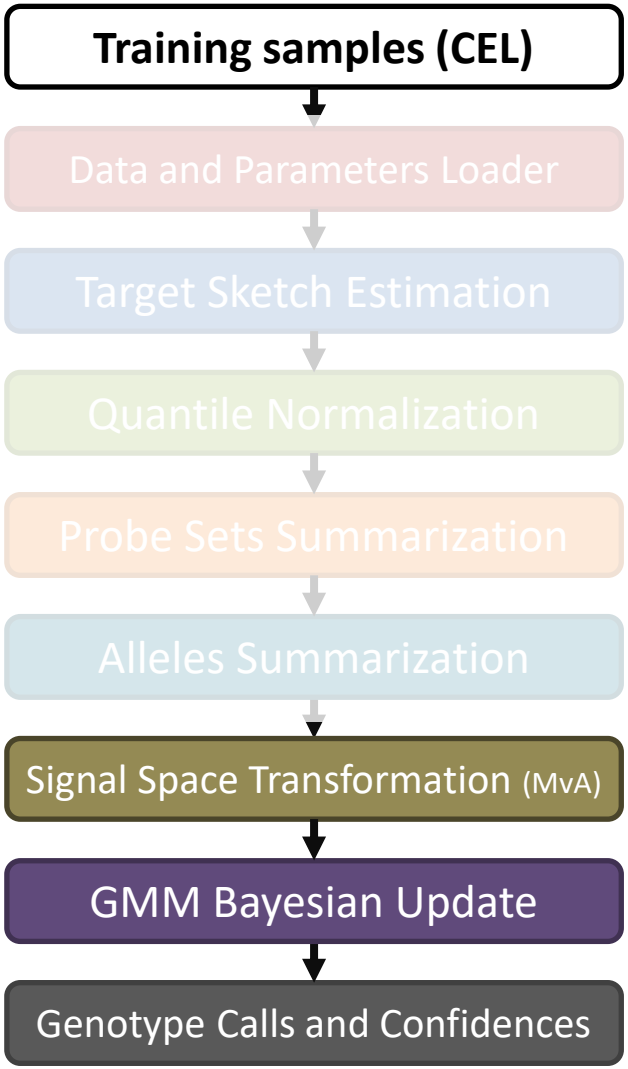
- `Integratebrlmmoverlabelings`
  - **Relative Probability for Each Data Point to Be Each Genotype under Posterior Information.**
    - The relative probability for the S_{th} data point being AA genotype: $\frac{\sum_{j=0}^{S} q_{\cdot j}}{q_{..}}$
    - The relative probability for the S_{th} data point being BB genotype: $1 - \frac{\sum_{i=0}^{S} q_{i\cdot}}{q_{..}}$
    - The relative probability for the S_{th} data point being AB genotype: $1 - \left(1 - \frac{\sum_{i=0}^{S} q_{i\cdot}}{q_{..}}\right) - \frac{\sum_{j=0}^{S} q_{\cdot j}}{q_{..}} = \frac{\sum_{i=0}^{S} q_{i\cdot} - \sum_{j=0}^{S} q_{\cdot j}}{q_{..}}$

Genotyping Analysis Development and Deployment



- labels_two_posterior
  - Update 2D Data model parameters with soft assignment of each data bin
    - $u'_{6x1} = \left(\boldsymbol{K_0}^{-1}_{6x6} + \boldsymbol{N'}_{6x6}\right)^{-1} * \left(\boldsymbol{K_0}^{-1}_{6x6} * u_{0\,6x1} + m_{6x1}\right),$
    - $\boldsymbol{K_0}^{-1}_{6x6} = \begin{bmatrix} k_{10} & \mathbf{0} & \sigma_{xx120} & \mathbf{0} & \sigma_{xx130} & \mathbf{0} \\ \mathbf{0} & k_{10} & \mathbf{0} & \sigma_{yy120} & \mathbf{0} & \sigma_{yy130} \\ \sigma_{xx120} & \mathbf{0} & k_{20} & \mathbf{0} & \sigma_{xx230} & \mathbf{0} \\ \mathbf{0} & \sigma_{yy120} & \mathbf{0} & k_{20} & \mathbf{0} & \sigma_{yy230} \\ \sigma_{xx130} & \mathbf{0} & \sigma_{xx230} & \mathbf{0} & k_{30} & \mathbf{0} \\ \mathbf{0} & \sigma_{yy130} & \mathbf{0} & \sigma_{yy230} & \mathbf{0} & k_{30} \end{bmatrix},$
    - $\boldsymbol{N}_{6x6} = \begin{bmatrix} \sum_i p_{1i} & \mathbf{0} & \cdots & & & \mathbf{0} \\ \mathbf{0} & \sum_i p_{1i} & & & & \\ & & \sum_i p_{2i} & & & \vdots \\ \vdots & & & \sum_i p_{2i} & & \\ & & & & \sum_i p_{3i} & \mathbf{0} \\ \mathbf{0} & & \cdots & & \mathbf{0} & \sum_i p_{31} \end{bmatrix}$
    - $u_{0\,6x1} = \begin{bmatrix} u_{x10} \\ u_{y10} \\ u_{x20} \\ u_{y20} \\ u_{x30} \\ u_{y30} \end{bmatrix}, \quad m_{6x1} = \begin{bmatrix} \sum_i x_{1i} \\ \sum_i y_{1i} \\ \sum_i x_{2i} \\ \sum_i y_{2i} \\ \sum_i x_{3i} \\ \sum_i y_{3i} \end{bmatrix}$

Genotyping Analysis Development and Deployment



- labels_two_posterior
  - Update 2D Data model parameters with soft assignment of each data bin
    - $\mathrm{u}'_{6x1} = \left(\boldsymbol{K_0}^{-1}_{6x6} + \boldsymbol{N'}_{6x6}\right)^{-1} * \left(\boldsymbol{K_0}^{-1}_{6x6} * \mathrm{u}_{0\ 6x1} + \mathrm{m}_{6x1}\right),$
    - $k'_g = k_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$
    - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $\mathrm{u}'_{x\ 3x1}$

      $w_g = k'_g, \quad g = 1,2,3$

      $gamma = delta * \dfrac{w_1 - w_3}{w_1 + w_2 + w_3}$

      $\mathrm{u}'_{x\ 3x1} = \begin{bmatrix} \mathrm{u}'_{x1} \\ \mathrm{u}'_{x2} \\ \mathrm{u}'_{x3} \end{bmatrix}, \quad \begin{aligned} \mathrm{u}'_{x1} &= \mathrm{u}'_{x1} + delta - gamma \\ \mathrm{u}'_{x2} &= \mathrm{u}'_{x2} - gamma \\ \mathrm{u}'_{x3} &= \mathrm{u}'_{x3} - delta - gamma \end{aligned}$

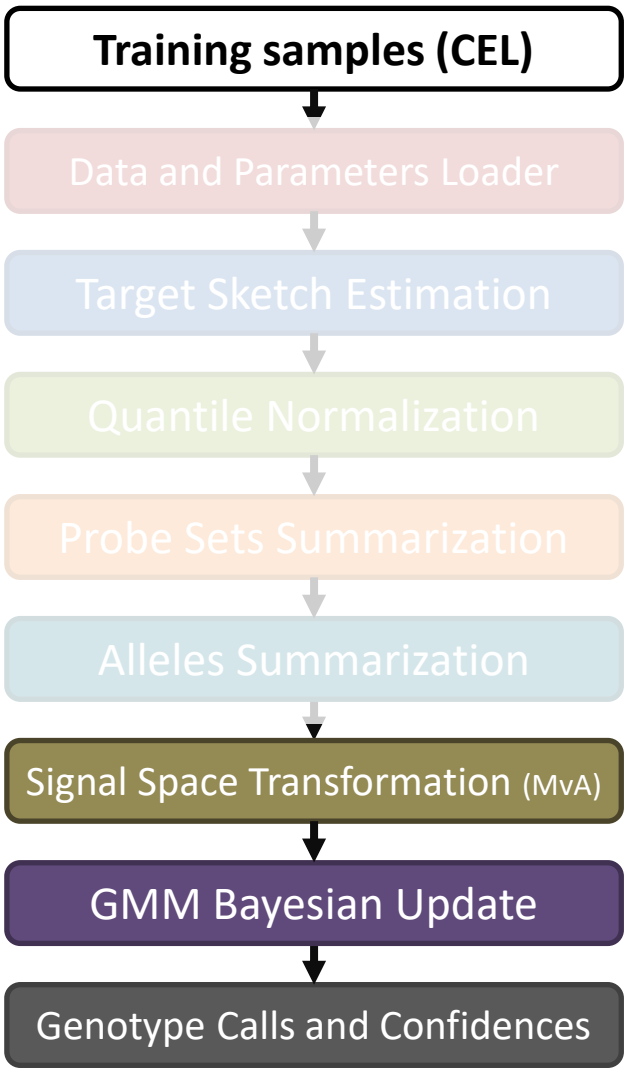
      Pool Adjacent-Violators (PAV) algo.

      $\begin{cases} \mathrm{u}'_{xg}, \mathrm{u}'_{xg+1}\ , & \mathrm{u}'_{xg} \le \mathrm{u}'_{xg+1} \\ \mathrm{u}'_{xg}, \mathrm{u}'_{xg+1} = \dfrac{\sum_{g\in A} w_g * \mathrm{u}'_{xg}}{\sum_{g\in A} w_g}, & A = \{g | \mathrm{u}'_{xg} > \mathrm{u}'_{xg+1}\}, \end{cases}$

      $g = 1,2,3$

      $\mathrm{u}'_{3x1} = \begin{bmatrix} \mathrm{u}'_{x1} \\ \mathrm{u}'_{x2} \\ \mathrm{u}'_{x3} \end{bmatrix}, \quad \begin{aligned} \mathrm{u}'_{x1} &= \mathrm{u}'_{x1} - delta + gamma \\ \mathrm{u}'_{x2} &= \mathrm{u}'_{x2} + gamma \\ \mathrm{u}'_{x3} &= \mathrm{u}'_{x3} + delta + gamma \end{aligned}$

Genotyping Analysis Development and Deployment



- labels_two_posterior
  - Update 2D Data model parameters with soft assignment of each data bin.
    - $u'_{6x1} = \left(\boldsymbol{K_0^{-1}}_{6x6} + \boldsymbol{N'}_{6x6}\right)^{-1} * \left(\boldsymbol{K_0^{-1}}_{6x6} * u_{0\ 6x1} + m_{6x1}\right)$,
    - $k'_g = k_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$
    - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $u'_{x\ 3x1}$
    - $v'_g = v_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$
    - $\sigma'^{2}_{xxg} \Rightarrow \frac{v_{g0} * \sigma^2_{xxg0} + \left(\sum_i p_{gi} x_{gi}^2 - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} x_{gi}\right) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * \left(u'_{xg} - u_{xg0}\right)^2}{v'_g}$
    - $\sigma'^{2}_{yyg} \Rightarrow \frac{v_{g0} * \sigma^2_{yyg0} + \left(\sum_i p_{gi} y_{gi}^2 - \sum_i p_{gi} y_{gi} * \sum_i p_{gi} y_{gi}\right) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * \left(u'_{yg} - u_{yg0}\right)^2}{v'_g}$
    - $\sigma'^{2}_{xyg} \Rightarrow \frac{v_{g0} * \sigma^2_{xyg0} + \left(\sum_i p_{gi} x y_{gi} - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} y_{gi}\right) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * \left(u'_{yg} - u_{yg0}\right) * \left(u'_{xg} - u_{xg0}\right)}{v'_g}$
    - Ad-hoc shrinkage for $\sigma'^{2}_{..g}$ of each cluster (controlled by mixing proportion (lambda) (1)).

      Adjusted Pooled Variance.

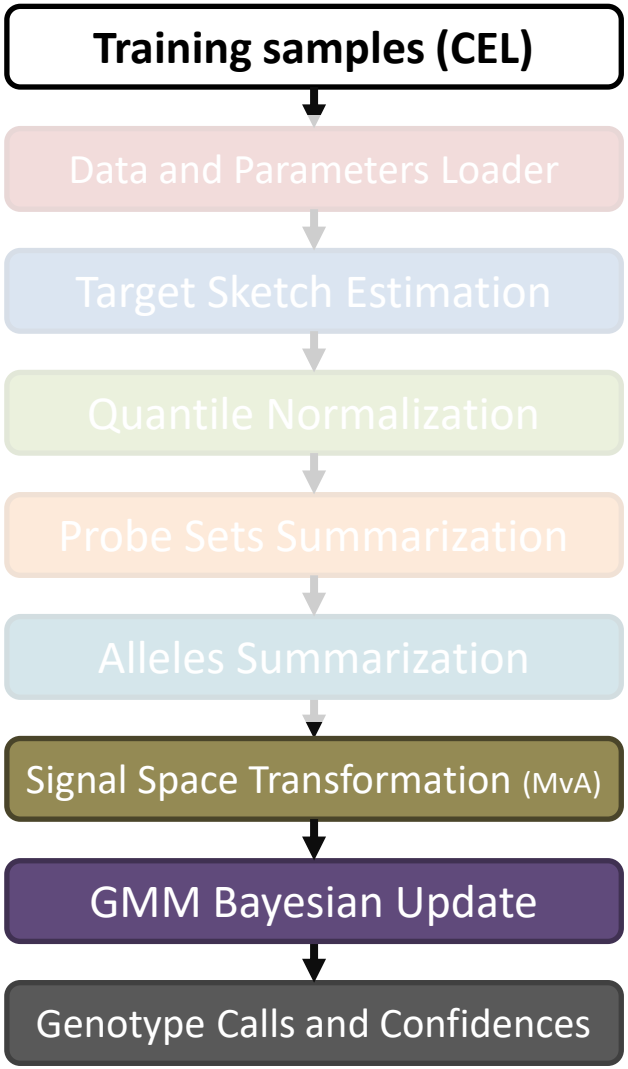
      $w_g = v_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$

      $\sigma'^{2}_{xxt} = \frac{\sum_g w_g * \sigma'^{2}_{xxg}}{\sum_g w_g}\,, \quad \sigma'^{2}_{yyt} = \frac{\sum_g w_g * \sigma'^{2}_{yyg}}{\sum_g w_g}, \quad t = 1,2,3\,,$

      $\Rightarrow \frac{(3 - 2 * lambda) * w_t * \sigma'^{2}_{t} + \sum_{g \neq t} lambda * w_g * \sigma'^{2}_{g}}{(3 - 2 * lambda) * w_t + \sum_{g \neq t} lambda * w_g}$

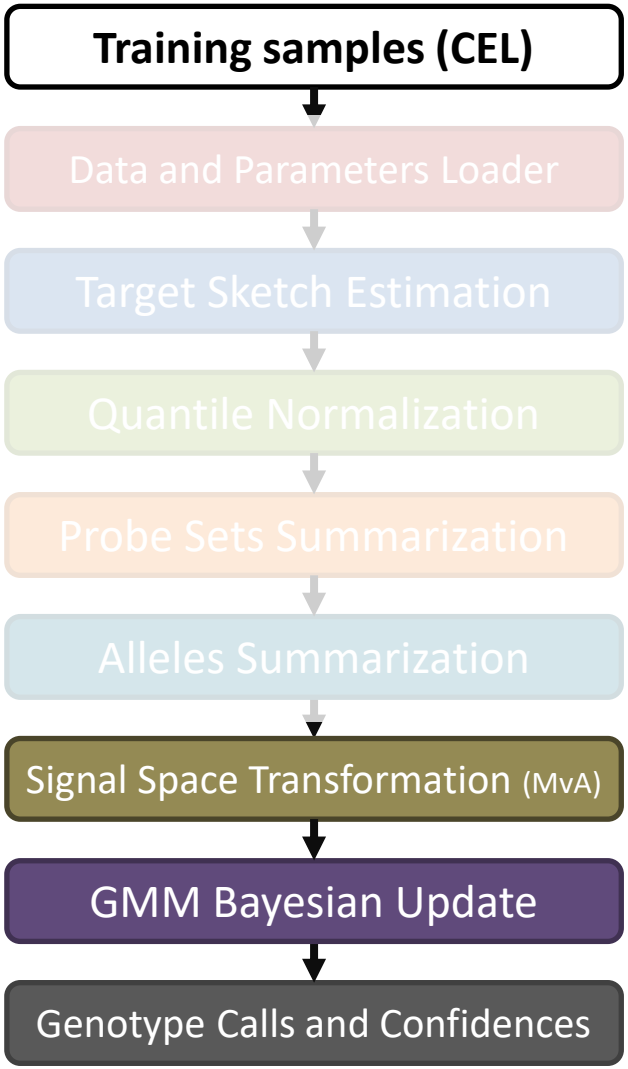
      $\sigma'^{2}_{xyt} = \left(\sigma'_{xxt} * \sigma'_{yyt}\right) * \frac{\sigma'_{xyg}}{\sigma'_{xxg} * \sigma'_{yyg}}, \quad t = 1,2,3\,, \quad g = t,$ means before shrinkage adjustment.

Genotyping Analysis Development and Deployment



- labels_two_posterior
  - Update 2D Data model parameters with soft assignment of each data bin
    - $\mathrm{u}'_{6\mathrm{x}1} = \left(\boldsymbol{K_0}^{-1}_{6\mathrm{x}6} + \boldsymbol{N}'_{6\mathrm{x}6}\right)^{-1} * \left(\boldsymbol{K_0}^{-1}_{6\mathrm{x}6} * \mathrm{u}_{0\ 6\mathrm{x}1} + \mathrm{m}_{6\mathrm{x}1}\right),$
    - $k'_g = k_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$
    - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $\mathrm{u}'_{x\ 3\mathrm{x}1}$
    - $v'_g = v_{g0} + \sum_i p_{gi}\,, \quad g = 1,2,3$
    - $\sigma'^{\,2}_{xxg}, \quad \sigma'^{\,2}_{yyg}, \quad \sigma'^{\,2}_{xyg}, \quad g = 1,2,3$
    - Ad-hoc shrinkage for $\sigma'^{\,2}_{\cdot\cdot g}$ of each cluster (controlled by mixing proportion (lambda) (1)).
    - $\sigma'_{xx12} = \sigma_{xx120}, \quad \sigma'_{xx13} = \sigma_{xx130}, \quad \sigma'_{xx23} = \sigma_{xx230}$
      $\sigma'_{yy12} = \sigma_{yy120}, \quad \sigma'_{yy13} = \sigma_{yy130}, \quad \sigma'_{yy23} = \sigma_{yy230}$

Genotyping Analysis Development and Deployment



- make_two_calls
  - (sp.mix, freqflag) compute the frequency of each cluster (AA, AB, BB)
    - $f_t = \frac{k'_t}{\sum_g k'_g}, t = 1,2,3$

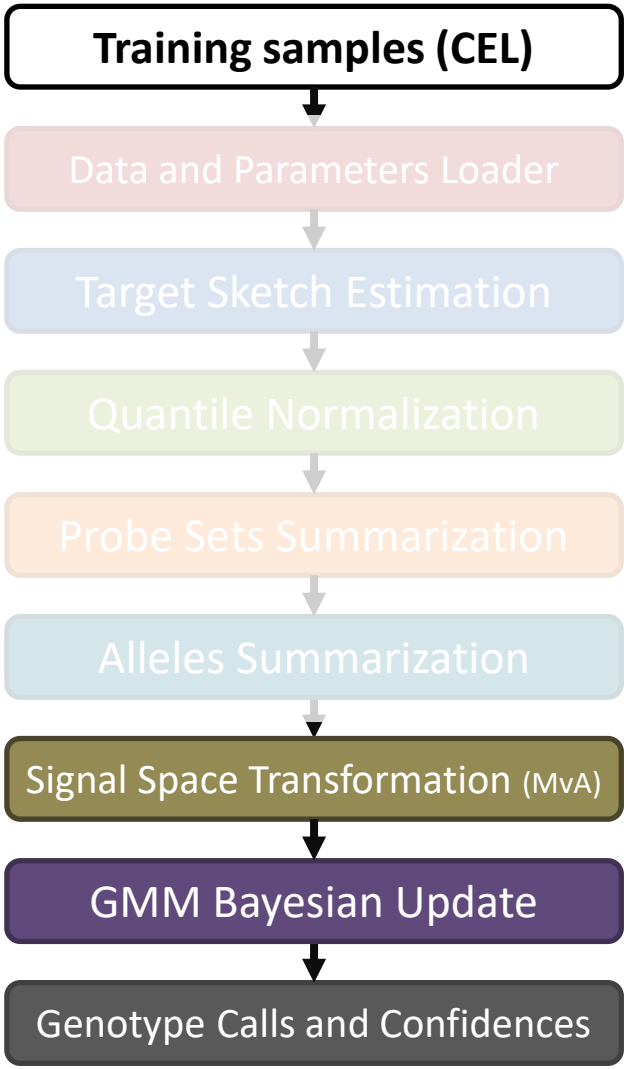
      $\Rightarrow \; logf_t = logk'_t - log\left(\sum_g k'_g\right)$

      $\Rightarrow \; -logf_t = -logk'_t + log\left(\sum_g k'_g\right)$
  - For each point, compute the the probability that a data point X (x, y) belongs to each genotype.
    - $p(X \in t|X) = \frac{p(X \in t, X)}{p(X)} = \frac{p(X \in t)p(X|X \in t)}{ocean + \sum_g p(X \in g)p(X|X \in g)} =$

      $$\frac{f_t \cdot BVN\left(X \middle| \mathbf{u}'_t, \left(1 + \frac{inflatePRA}{k'_t}\right) \cdot \boldsymbol{\sigma}'_t\right)}{ocean + \sum_g f_g \cdot BVN\left(X \middle| \mathbf{u}'_g, \left(1 + \frac{inflatePRA}{k'_g}\right) \cdot \boldsymbol{\sigma}'_g\right)},$$

      $inflatePRA = 0\ (default),\ \ ocean = 0.00001\ (default)$
    - $logp(X \in t)p(X|X \in t) = \log\left(p(X \in t)\right) + \log\left(p(X|X \in t)\right)$
    - If copynumber=1, $p(X \in AB)p(X|X \in AB) = 0$

Genotyping Analysis Development and Deployment



- make_two_calls
  - (sp.mix, freqflag) compute the frequency of each cluster (AA, AB, BB)
    - $f_t = \frac{k'_t}{\sum_g k'_g}, t = 1, 2, 3$
  - For each point, compute the the probability that a data point X (x, y) belongs to each genotype.
    - $p(X \in t|X) = \frac{p(X \in t, X)}{p(X)} = \frac{p(X \in t)p(X|X \in t)}{ocean + \sum_g p(X \in g)p(X|X \in g)}, t = 1, 2, 3$
  - For each point, make a call.
    - $\hat{t} = \underset{t}{\operatorname{argmax}}\, p(X \in t|X)$
    - $confidence = 1 - p(X \in \hat{t}|X)$
    - No call: If $confidence > MS$,
      $MS = 0.15$ (default) @ getGTypeCall()

Genotyping Analysis Development and Deployment

- Evaluation Data: GSE78098
  - Platform: GPL21480 (Axiom_GW_Hu-CHB_SNP)
    - Focus on Chinese (Asian) people.
  - 420 human samples
  - 639653 Probe sets
  - Preprocessing and filter by DQC < 0.82
  - All 419 samples are picked.

- Evaluation Data: Dog Banff
  - Platform: B1C (Banff)
  - 187 dog samples
  - 48283 Probe sets
  - Preprocessing for channel name, vcf allele definition, and change the coordinate system for the Y axis of the heatmap.
  - QC filter by NP probes performance (e.g. NP call rate, NP call slope).
  - 155 dog samples are finally picked and used to build genotyping models.

- Manual, Reports & Results: Project-CPT/CPT.wiki/AxiomGT.md at main · jeff665547/Project-CPT (github.com)

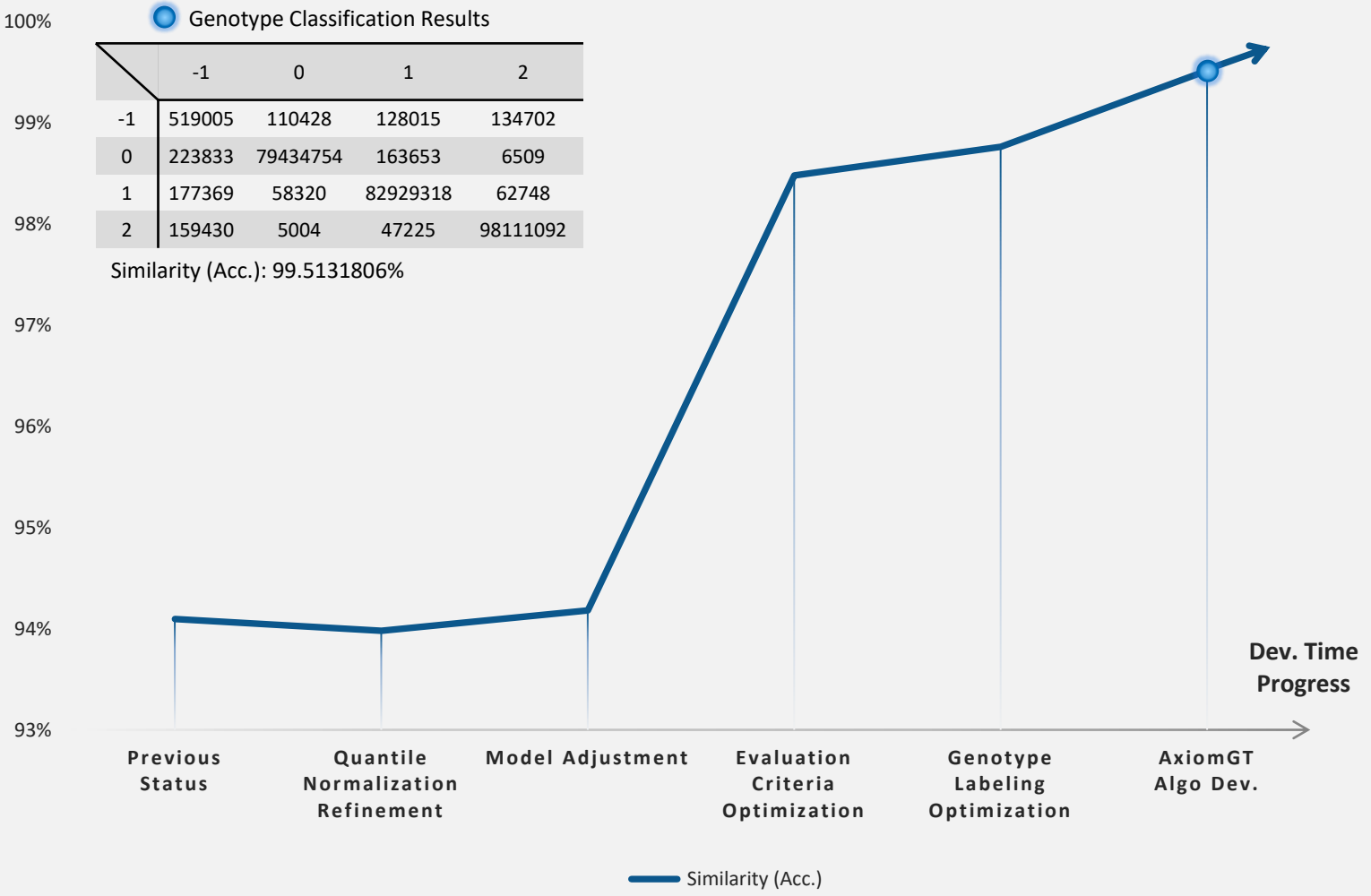
Genotyping Analysis Development and Deployment

GSE78098 **TESTING DATA PERFORMANCE**

Genotype Classification Results

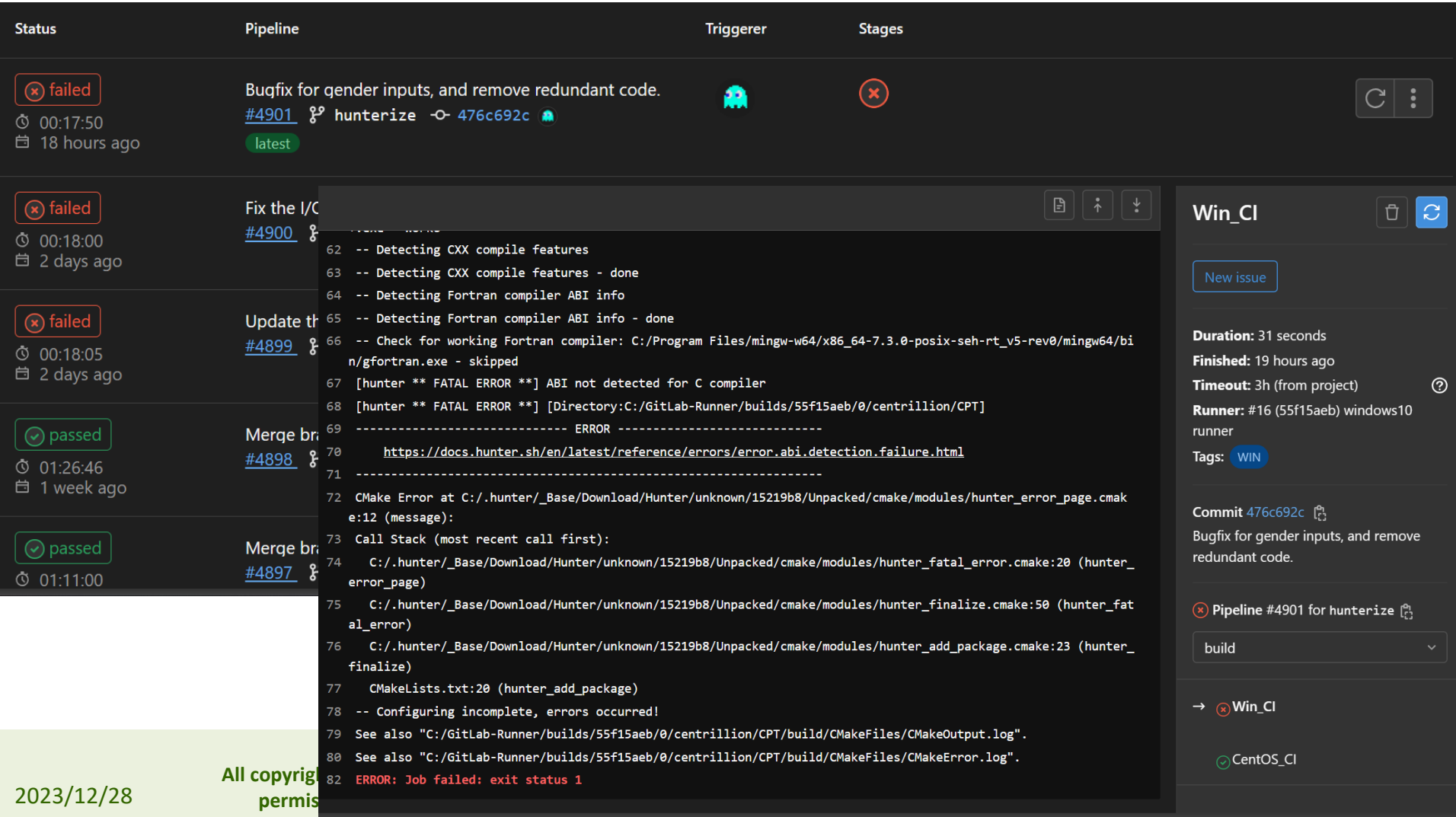
	-1	0	1	2
-1	519005	110428	128015	134702
0	223833	79434754	163653	6509
1	177369	58320	82929318	62748
2	159430	5004	47225	98111092

Similarity (Acc.): 99.5131806%



Genotyping Analysis Development and Deployment

- Bugfix for the CI/CD error when deployment.



Genotyping Analysis Development and Deployment

- Bugfix for the CI/CD error when deployment.

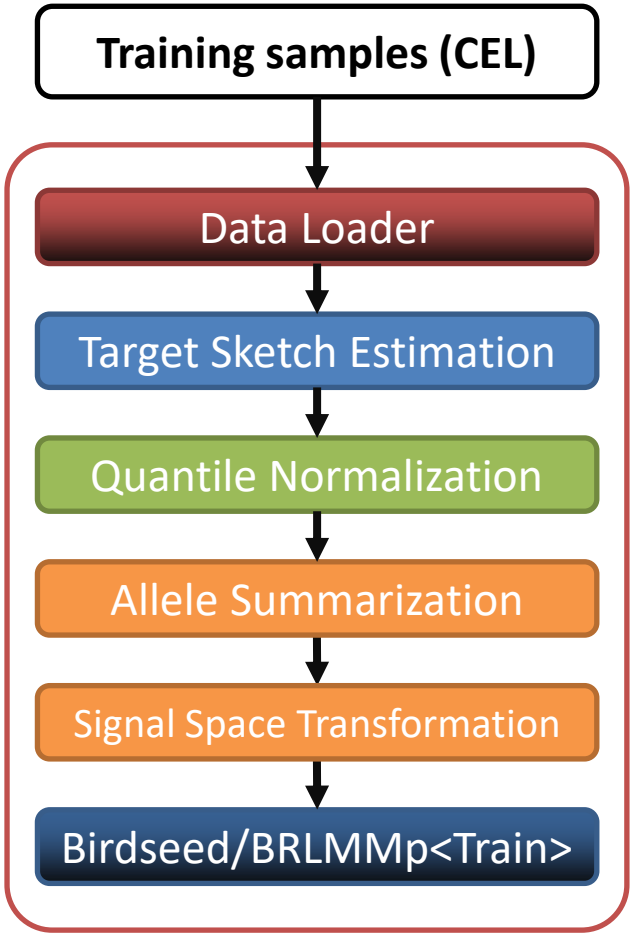
Status	Pipeline	Triggerer	Stages	
passed 01:33:52 15 hours ago	Bugfix for gender inputs, and remove redundant code. #4901 hunterize 476c692c latest			
passed 01:30:03 14 hours ago	Fix the I/O bug for the MvA Transformation. #4900 hunterize 99c7889d			
passed 01:30:22 12 hours ago	Update the logging system. #4899 hunterize 882600dc			
passed 01:26:46 1 week ago	Merge branch 'APT_AxiomGT1' into hunterize #4898 hunterize ba50e116			
passed 01:11:00 1 week ago	Merge branch 'APT_AxiomGT1' into hunterize #4897 hunterize 7a4dfac4			

Other Genotyping Models Research and Development

Jeff (CHI-HSUAN HO)

Progress Report and Future Work

- Birdseed Framework



TECHNICAL REPORTS

nature genetics

Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs

Joshua M Korn[1–5,10], Finny G Kuruvilla[1,4–6,10], Steven A McCarroll[1,4,5], Alec Wysoker[1], James Nemesh[1], Simon Cawley[7], Earl Hubbell[7], Jim Veitch[7], Patrick J Collins[7], Katayoon Darvishi[8], Charles Lee[8], Marcia M Nizzari[1], Stacey B Gabriel[1], Shaun Purcell[1,5], Mark J Daly[1,5,9] & David Altshuler[1,4,5,9]

- Birdseed: K-means training model

$Model_0$: *Gaussian Mixture Model – GMM with K-Means centroid (Existing Model)*

$Model_1$: *Non – prbabilistic Model*

$$BIC = \frac{1}{\sigma^2}\sum_{j=1}^{K}\sum_{i=1}^{N_j} min\left\|\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_j\right\|^2 + Kd \cdot \ln(N)\,, \qquad \hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{N_j}\boldsymbol{X}_i}{N_j}$$

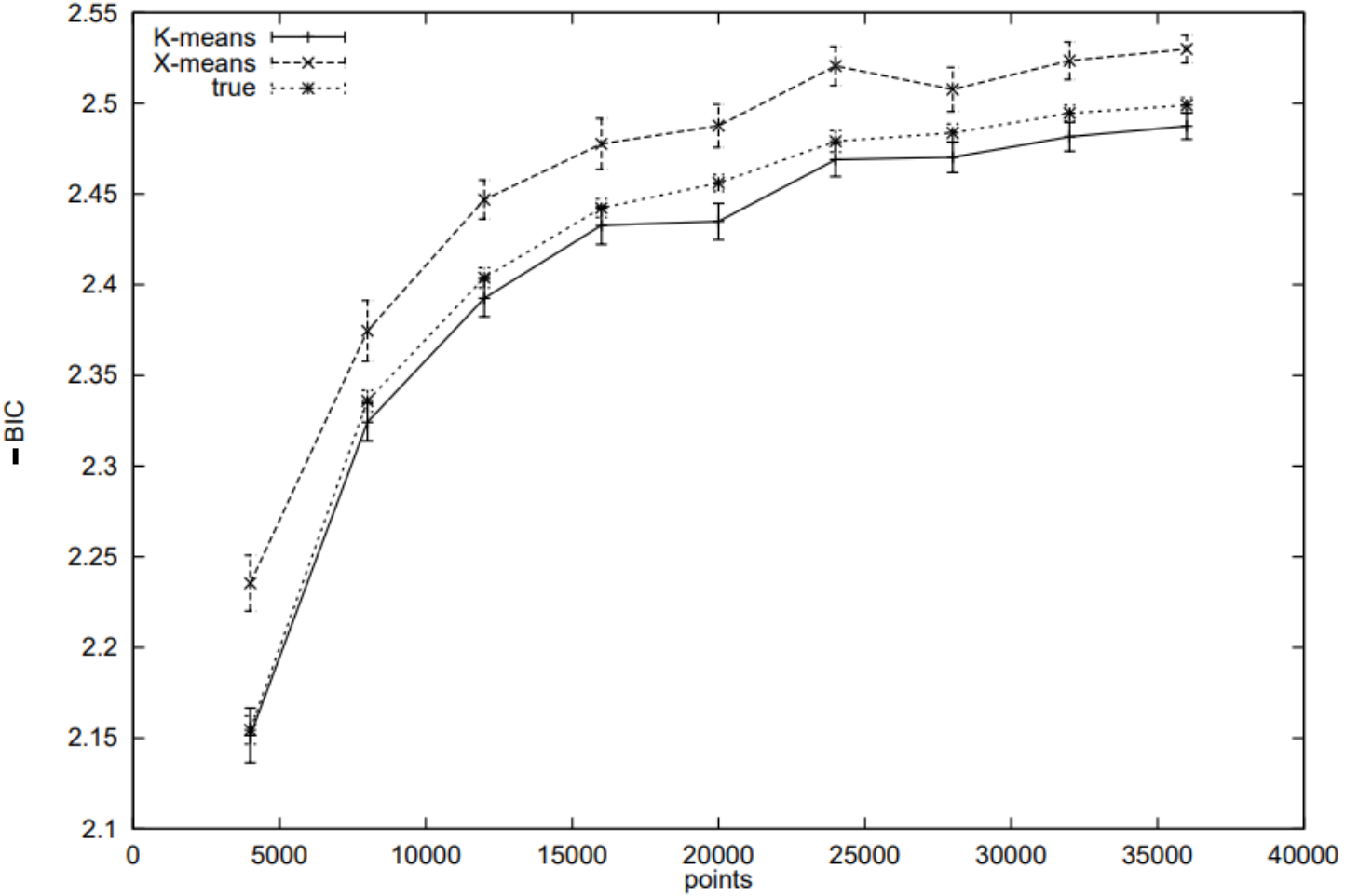
$$Model_2: f(x_i) = \pi_{ij} N\left(\boldsymbol{\mu}_j, \sigma^2 \cdot \boldsymbol{I}_d\right)$$

$$BIC = -2\sum_{j=1}^{K} N_j \ln\left(N_j\right) + 2N\ln(N) + \mathrm{dN}\ln(2\pi\hat{\sigma}^2) + dN + \ln(N)\cdot K(d+1),$$

$$where\ \hat{\sigma}^2 = \frac{1}{dN}\sum_{j=1}^{K}\sum_{i=1}^{N_j}\left\|\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_j\right\|^2, \qquad \hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{N_j}\boldsymbol{X}_i}{N_j}$$

Genotyping Methods Evaluation and Simulation

$Model_2$



- Birdseed: K-means training model

$$Model_3\text{: } ANOVA\text{: } X_{it} = \mu_{it} + E_{it},\ \boldsymbol{\mu} \in \mathbb{R}^{nxd},\ \boldsymbol{E} \in \mathbb{R}^{nxd},\ E_{it} \sim iid\ N(0,\sigma^2),\ i = 1,2,\dots,N,\ t = 1,2,\dots,d$$

$$BIC = Nd \cdot \ln\left(\sum_{j=1}^{K}\sum_{i=1}^{N_j}\left\|\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_j\right\|^2\right) + Nd\left(1 + \ln\left(\frac{2\pi}{Nd}\right)\right)$$

$$+ \ln(Nd) \cdot \left[Kd + \frac{1}{\tilde{\sigma}}\sum_{j=1}^{K'}\sum_{i=1}^{\widetilde{N_j}}\sum_{t=1}^{d}\sum_{l\neq c(i)}\phi\left(\frac{X_{i,t} + \delta_l^{i,t} - \tilde{\mu}_{i,t}}{\tilde{\sigma}}\right) \cdot \lim_{\gamma \rightleftarrows \delta_l^{i,t}} \mathcal{M}\left(\boldsymbol{X} + \gamma \boldsymbol{e}_{i,t}\right)_{i,t}\right],$$

$$where\ \tilde{\sigma}^2 = \frac{1}{dN}\sum_{j=1}^{K'}\sum_{i=1}^{\widetilde{N_j}}\|\boldsymbol{X}_i - \tilde{\boldsymbol{\mu}}_i\|^2,\quad \tilde{\boldsymbol{\mu}} = \mathcal{M}(\boldsymbol{X}; K') \text{ for some } K' > K,$$

$\phi(\cdot)$ is the pdf of Normal (Gaussian) distribution,

$$\lim_{\gamma \rightleftarrows \delta_l^{i,t}} \mathcal{M}\left(\boldsymbol{X} + \gamma \boldsymbol{e}_{i,t}\right)_{i,t} = (-1)^{I_{\{\delta_l^{i,t} > 0\}}}\left(\hat{\mu}_{c(i),t} - \frac{N_l}{N_l+1}\hat{\mu}_{l,t} - \frac{X_{i,t}}{N_l+1} + \delta_l^{i,t}\left(\frac{N_l+1-N_{c(i)}}{(N_l+1)\cdot N_{c(i)}}\right)\right),$$

$$c(i) = argmin_{l\epsilon\{1,2,\dots,K\}}\|\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_l\|^2,$$

$$\left(1 - \left(\frac{N_{c(i)}-1}{N_{c(i)}}\right)^2\right)\delta_l^{i,t\,2} + 2 \cdot \left(\left(X_{i,t} - \hat{\mu}_{l,t}\right) - \left(X_{i,t} - \hat{\mu}_{c(i),t}\right) \cdot \left(\frac{N_{c(i)}-1}{N_{c(i)}}\right)\right)\delta_l^{i,t} + \left(X_{i,t} - \hat{\mu}_{l,t}\right)^2 - \left(X_{i,t} - \hat{\mu}_{c(i),t}\right)^2 = 0$$

Genotyping Methods Evaluation and Simulation

$Model_3$

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Degrees of freedom and model selection for k-means clustering

David P. Hofmeyr

Department of Statistics and Actuarial Science, Stellenbosch University, Cnr. Bosman and Victoria streets, Stellenbosch 7600, South Africa

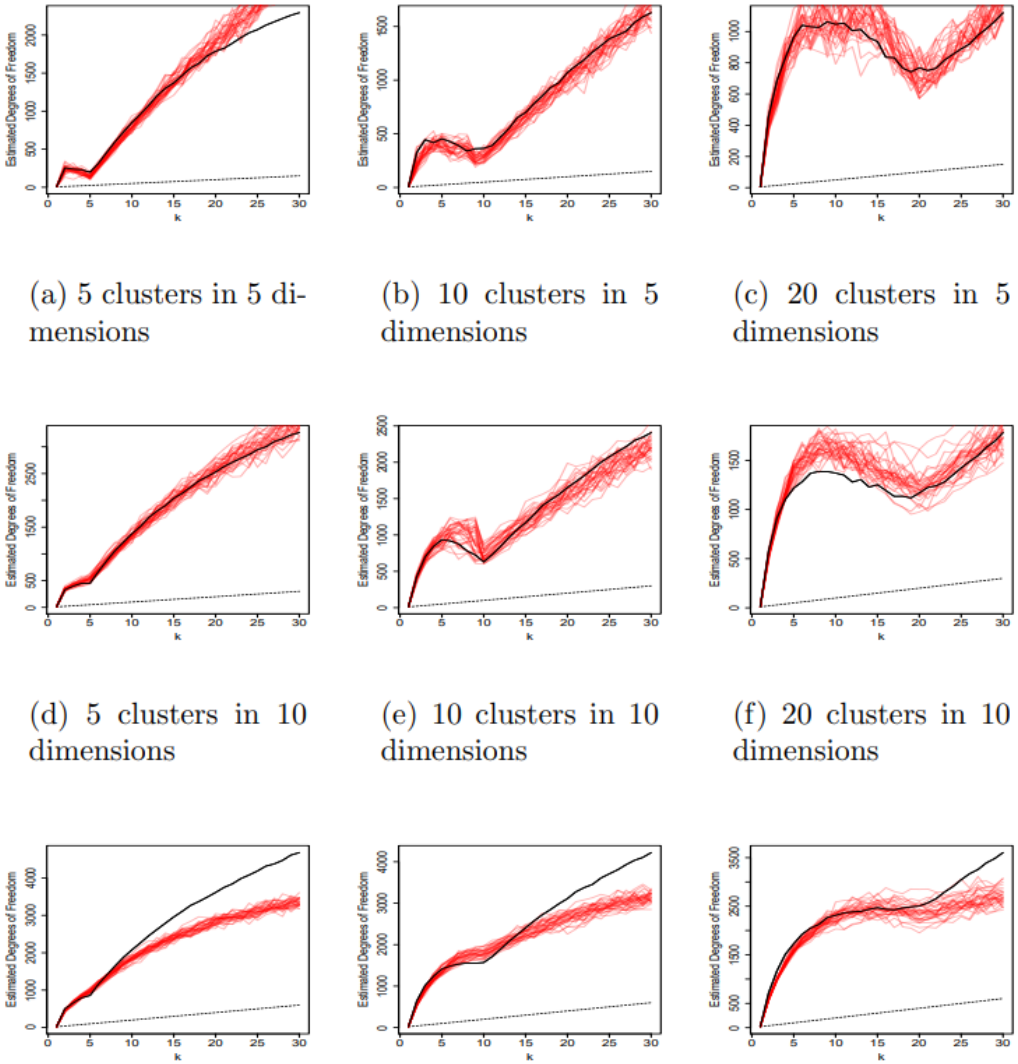
ARTICLE INFO

Article history:
Received 22 November 2019
Received in revised form 30 March 2020
Accepted 1 April 2020
Available online 13 April 2020

Keywords:
Clustering
k-means
Model selection
Cluster number determination
Degrees of freedom
Bayesian Information Criterion
Penalised likelihood

ABSTRACT

A thorough investigation into the model degrees of freedom in k-means clustering is conducted. An extension of Stein's lemma is used to obtain an expression for the effective degrees of freedom in the k-means model. Approximating the degrees of freedom in practice requires simplifications of this expression, however empirical studies evince the appropriateness of the proposed approach. The practical relevance of this new degrees of freedom formulation for k-means is demonstrated through model selection using the Bayesian Information Criterion. The reliability of this method is then validated through experiments on simulated data as well as on a large collection of publicly available benchmark data sets from diverse application areas. Comparisons with popular existing techniques indicate that this approach is extremely competitive for selecting high quality clustering solutions.



(a) 5 clusters in 5 dimensions

(b) 10 clusters in 5 dimensions

(c) 20 clusters in 5 dimensions

(d) 5 clusters in 10 dimensions

(e) 10 clusters in 10 dimensions

(f) 20 clusters in 10 dimensions

Genotyping Methods Evaluation and Simulation

- Birdseed: K-means training model

$Model_0$: $Equal\ weight\ Before\ scale$: 94.1908%
$Model_0$: $But\ differnet\ weight\ Before\ scale$
– $Default\ (trim\ BIC, trim\ Fan)$: 84.6841% $with\ very\ highly\ lose-classified\ rate.$
$Model_0$: $But\ differnet\ weight\ Before\ scale - no\ trim\ BIC$: 90.8778%
$Model_0$: $But\ differnet\ weight\ Before\ scale - no\ trim$ Fan: 0%
$Model_1$: $Before\ scale$
– $Default\ (trim\ BIC, trim\ Fan)$: 98.3863% $with\ very\ highly\ lose-classified\ rate.$
$Model_1$: $Before\ scale\ - no\ trim\ BIC$: 98.6603% → 98.7585% ($BIC\ under\ all\ data$)
$Model_1$: $Before\ scale\ - no\ trim$ Fan: 98.4763%
$Model_1$: $After\ scale$: 94.3179%

```
1   -    -2   -1   0    1    2    cen
2   -2   0    0    0    0    0
3   -1   0    0    68556     134030    87256
4   0    0    0    18744020       278090    4160
5   1    0    0    99968     19637894       134212
6   2    0    0    1485      255610    23149219
7   affy
8   acc : 0.987585
```

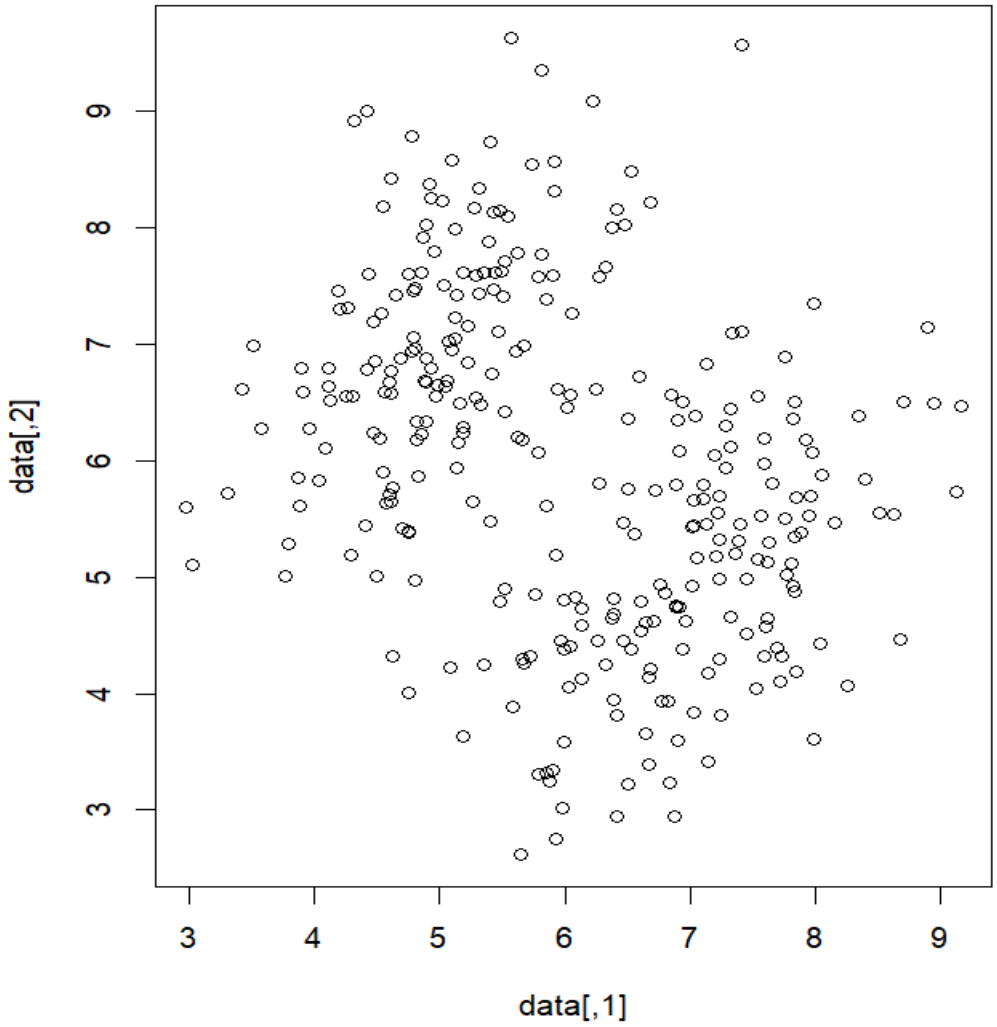
Genotyping Methods Evaluation and Simulation

- Simulation

`sol$k:` $Model_3$

`sol$k_:` $Model_1$: $Before\ scale$

```
> print(sol$k)
[1] 2
> print(sol$k_)
[1] 3
```



Ans: $k = 2$

Genotyping Methods Evaluation and Simulation

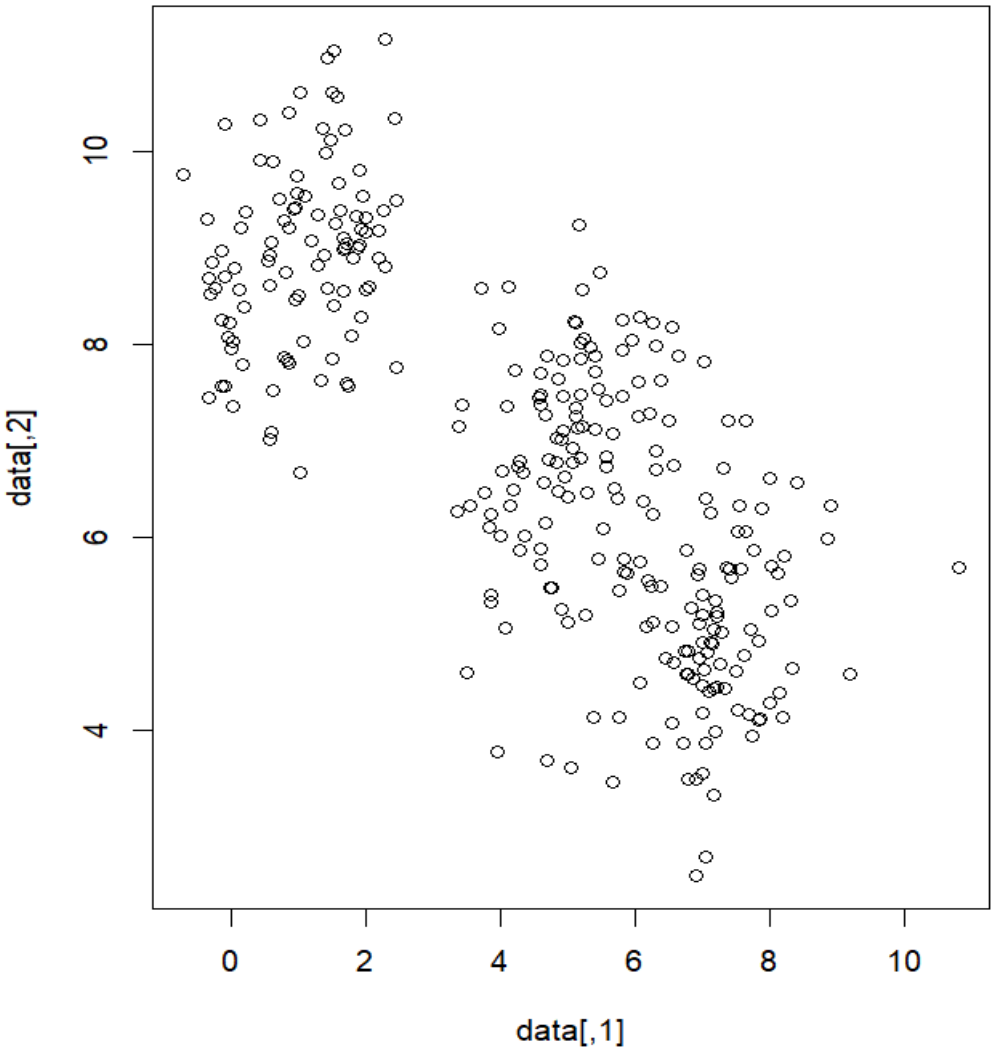
- Simulation

`sol$k:` $Model_3$

`sol$k_:` $Model_1$: $Before\ scale$

```
> print(sol$k)
[1] 3
> print(sol$k_)
[1] 3
```

Ans: $k = 3$



Summit Grid Algorithm Design

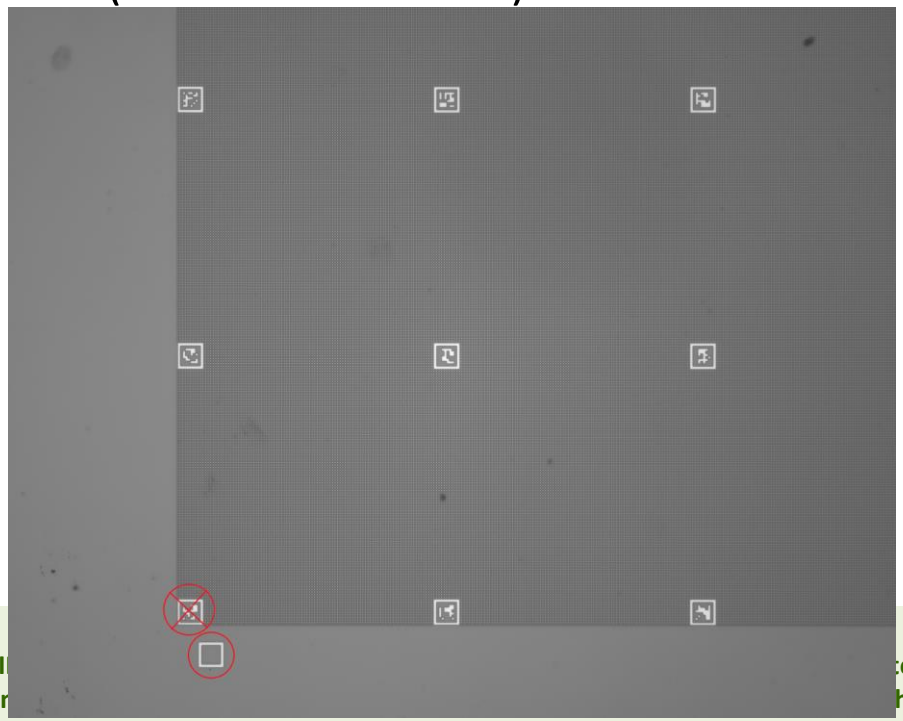
Jeff (CHI-HSUAN HO)

Gridding Algo. Development

- Summit.Grid
  - Bugfix for wrong nms_count (WARNING for S1C)
    - Original

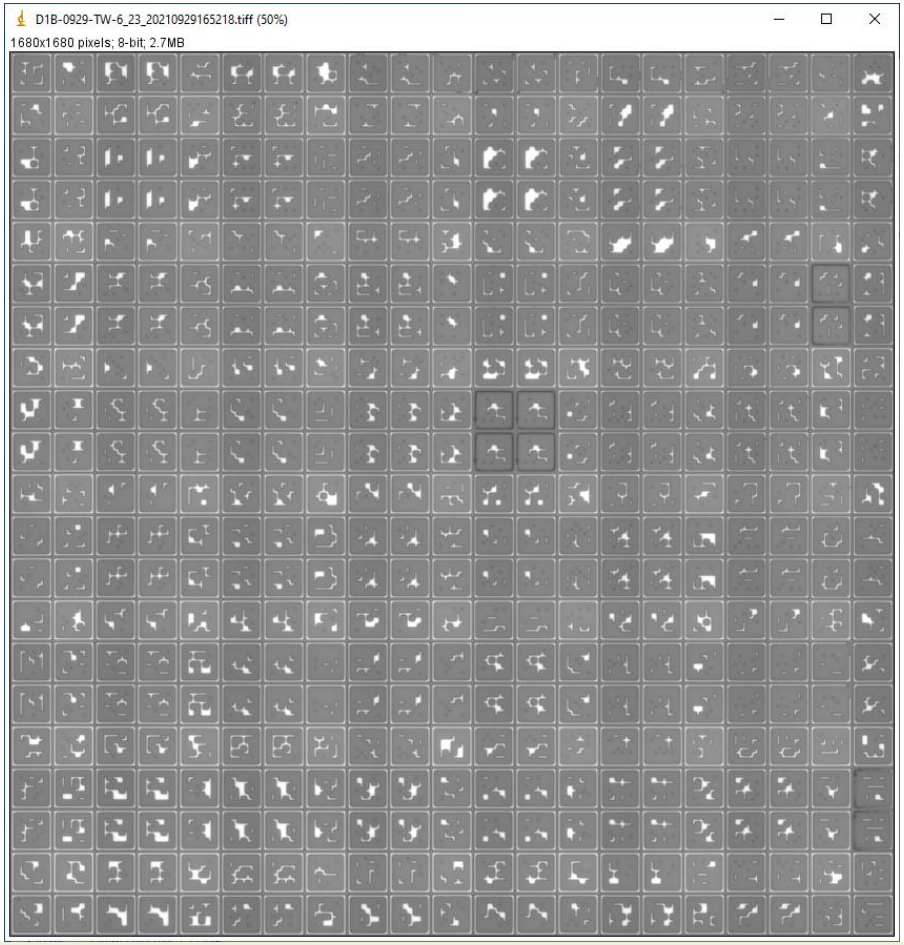
```
// detection parameters
nms_count_      = (fov_wd_ / mk_wd_cl_ + 1) * (fov_hd_ / mk_hd_cl_ + 1);      Alex, 2 years ago • support new aruco recognition …
nms_radius_     = aruco_marker_->at("nms_radius");
```

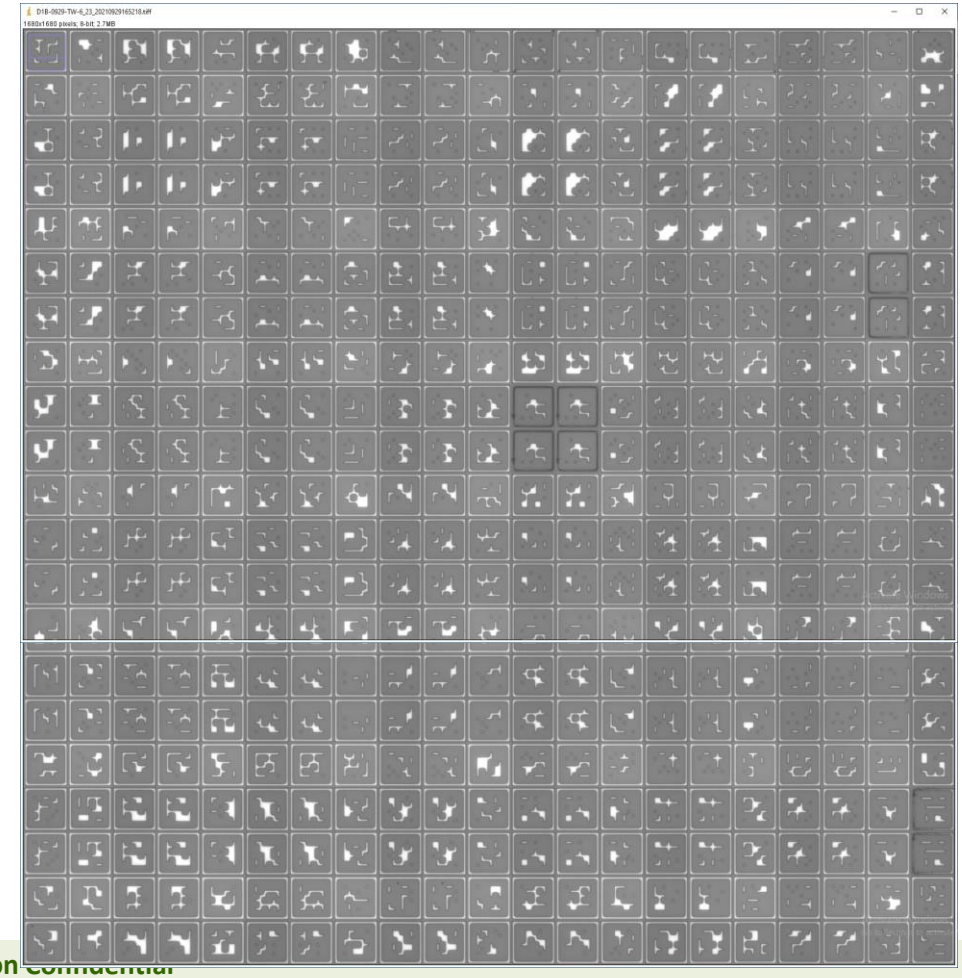
  - Noise influence (WARNING for Y2B)



Gridding Algo. Development

- Summit.Grid
  - Rescue mechanism for gridding bad fov (for erosion).

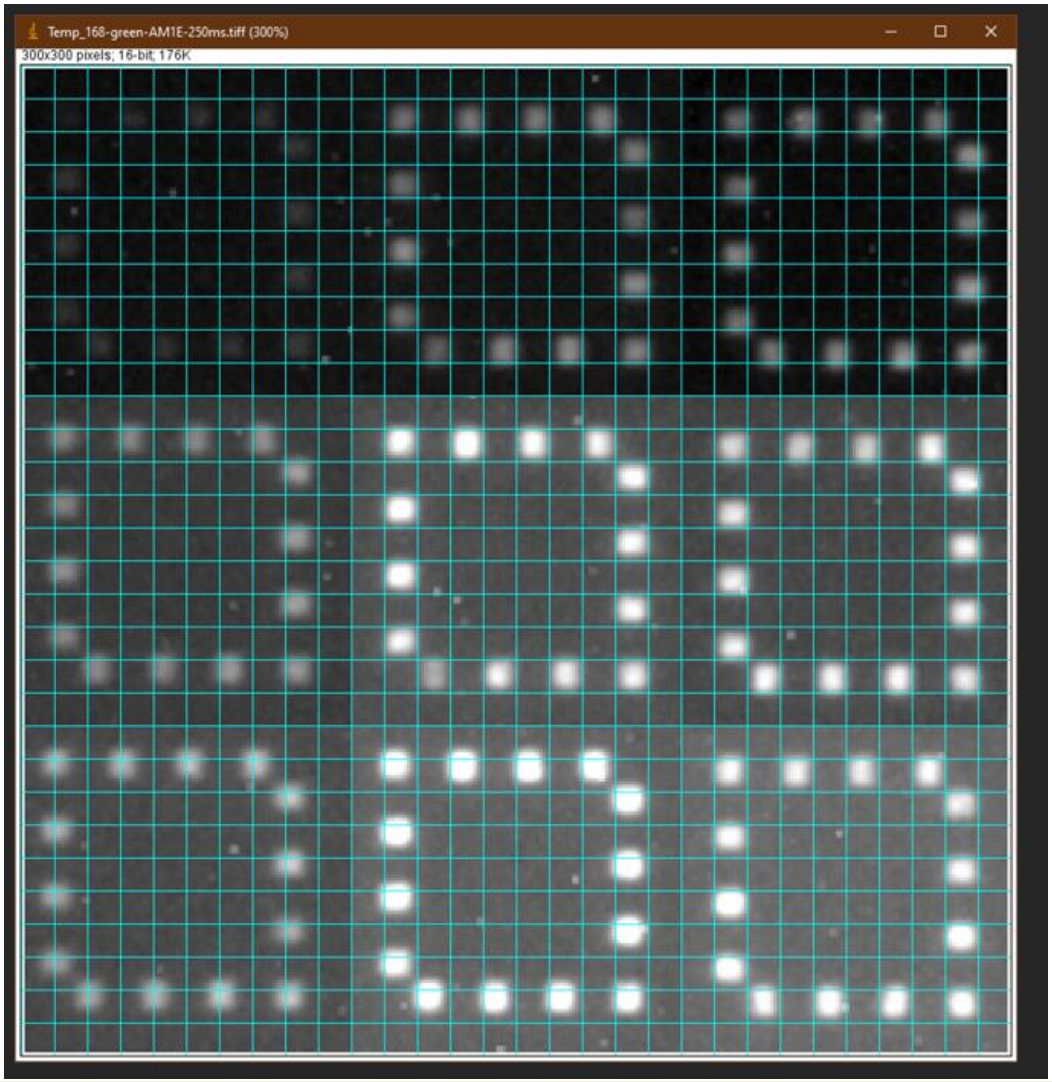




Gridding Algo. Development

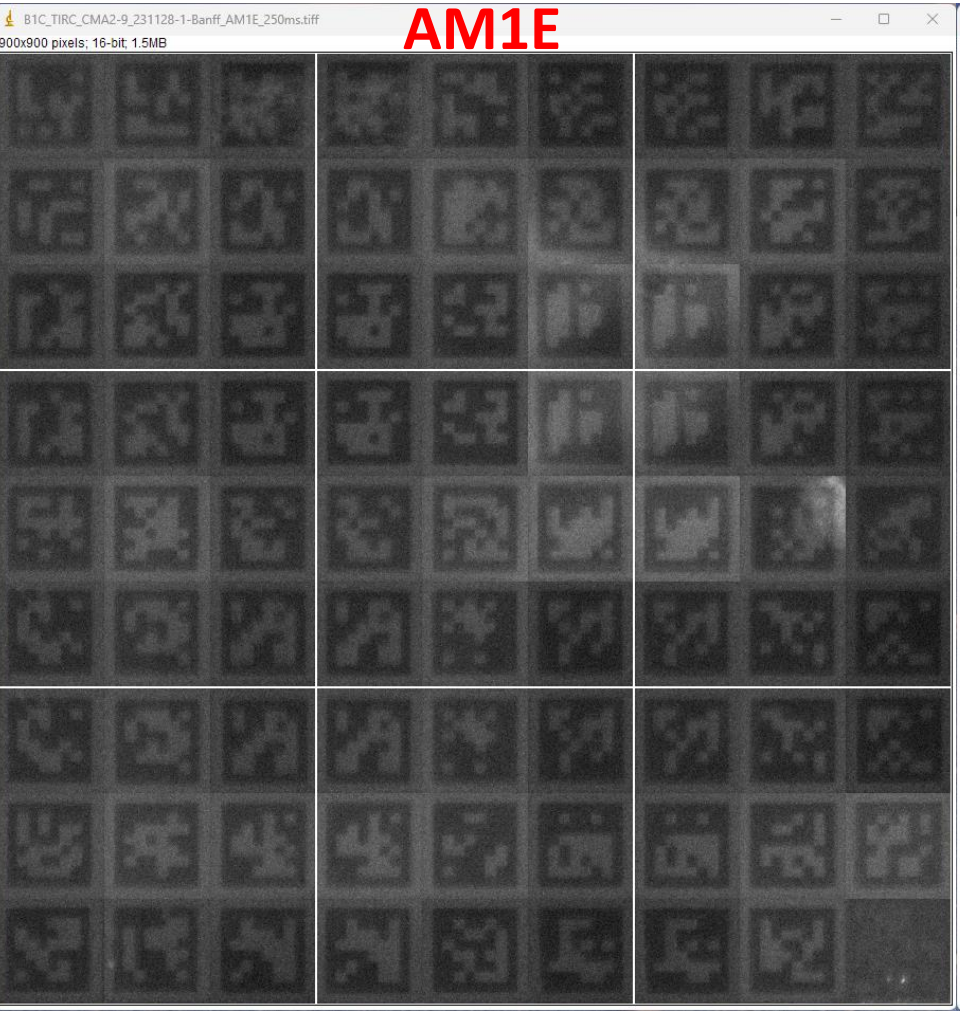
- Summit.Grid
  - PGD Images Processing.

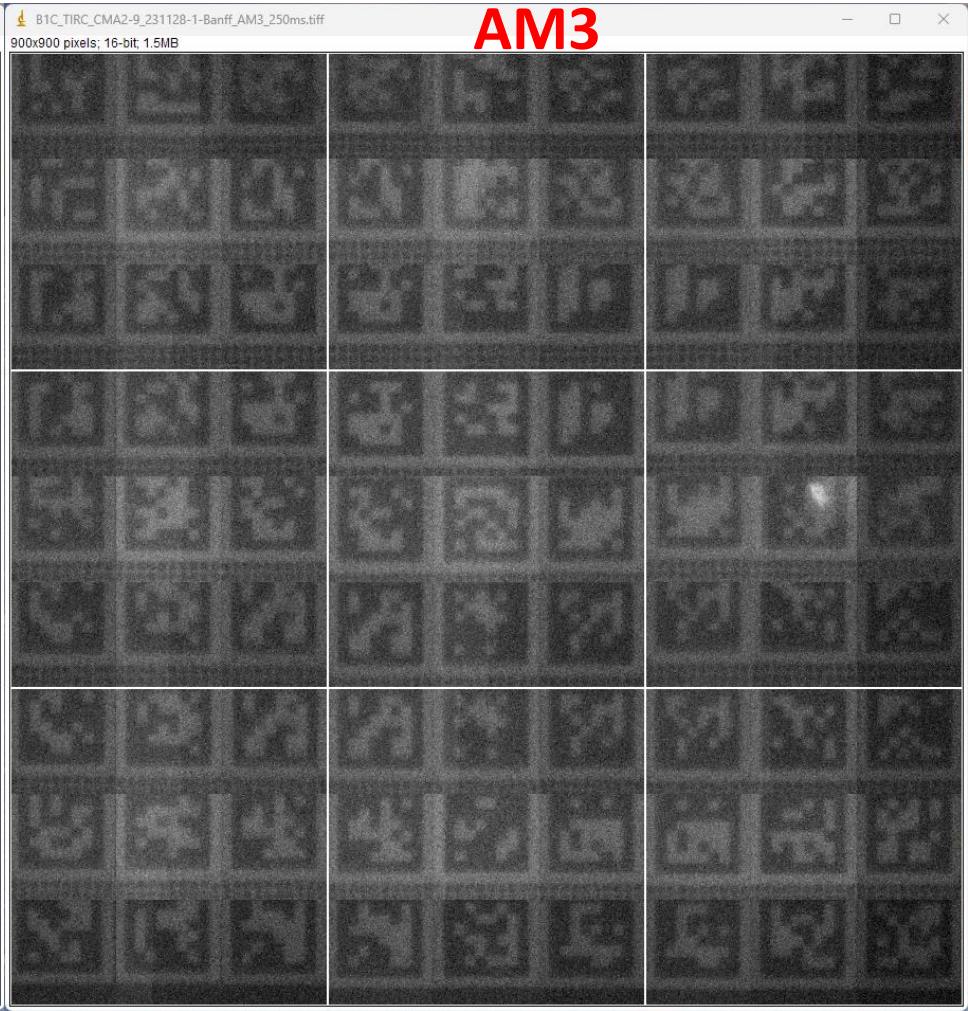

```
  ],
  "warp_mat": [
    [
      1.3189138576779025,
      0.013670411985018604,
      746.4366977969215
    ],
    [
      -0.013857677902621766,
      1.313483146067416,
      216.52305980929015
    ]
  ]
```

Gridding Algo. Development

- Summit Grid checking support.



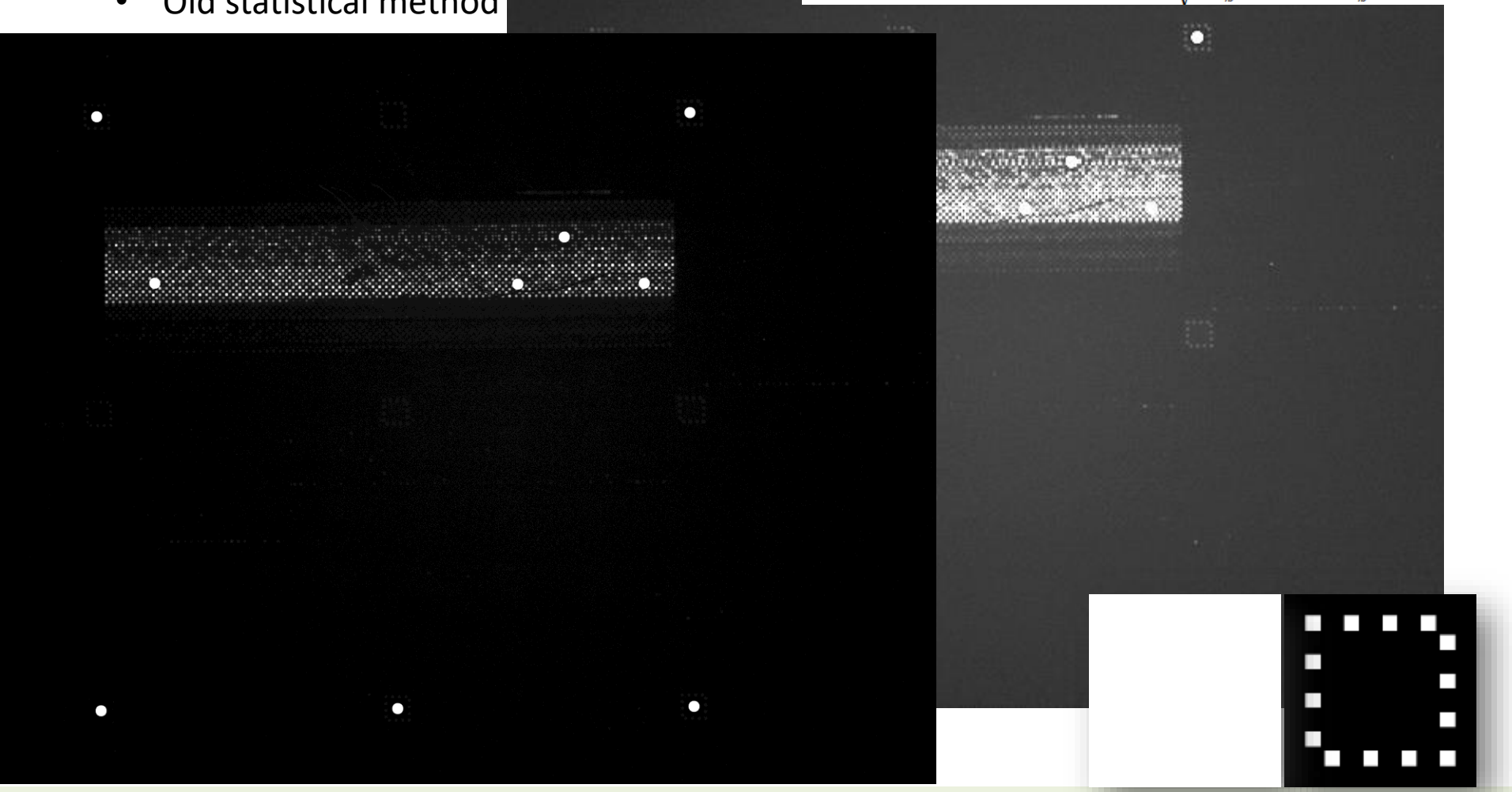


Performance for New Gridding Algo.

- Summit.Grid
  - Old statistical method

method=TM_CCORR_NORMED

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x', y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x', y+y')^2}}$$



Performance for New Gridding Algo.

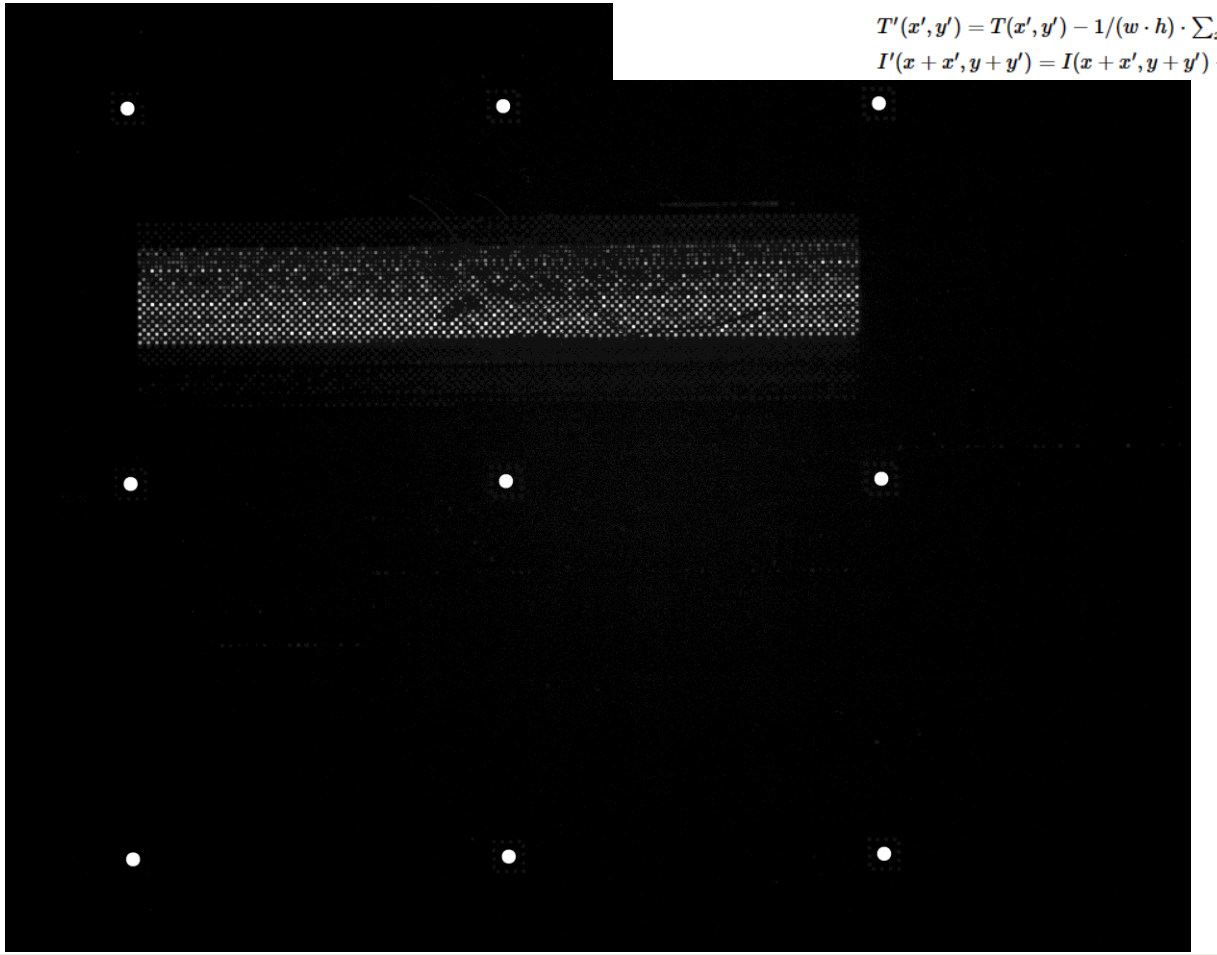
- Summit.Grid
  - New statistical method

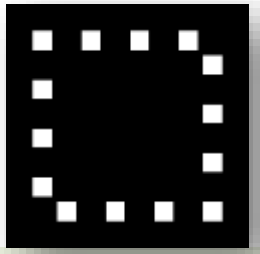
method=TM_CCOEFF_NORMED

$$R(x,y) = \frac{\sum_{x',y'}(T'(x',y') \cdot I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}}$$

where

$$T'(x',y') = T(x',y') - 1/(w \cdot h) \cdot \sum_{x'',y''} T(x'',y'')$$
$$I'(x+x',y+y') = I(x+x',y+y') - 1/(w \cdot h) \cdot \sum_{x'',y''} I(x+x'',y+y'')$$



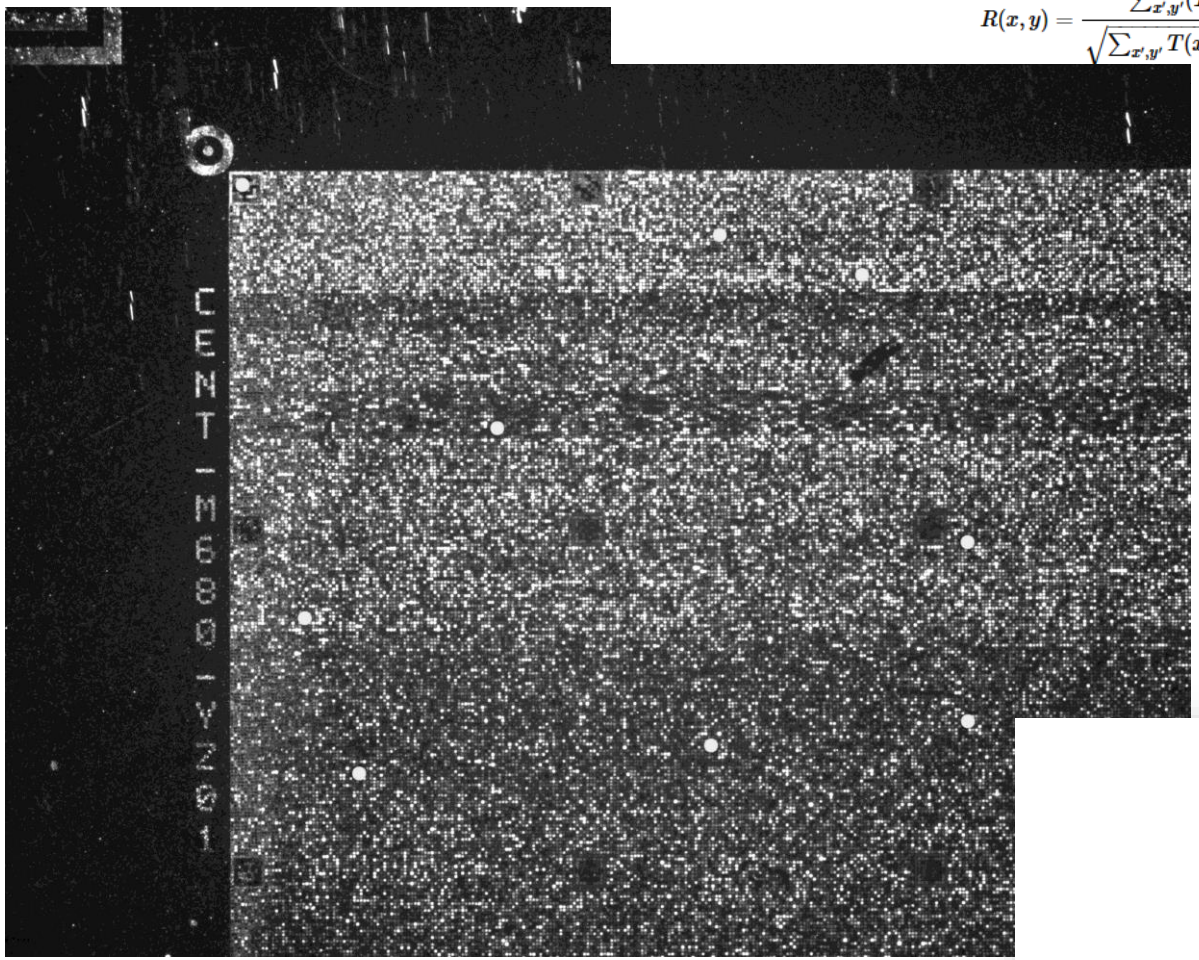


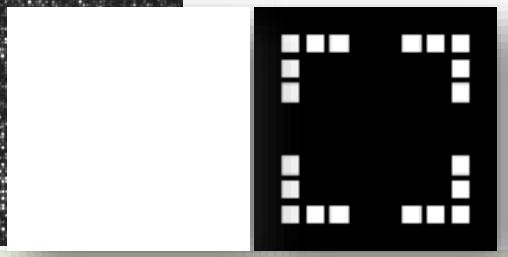
Performance for New Gridding Algo.

- Summit.Grid
  - Old statistical method

method=TM_CCORR_NORMED

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x', y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x', y+y')^2}}$$





Performance for New Gridding Algo.

- Summit.Grid
  - New statistical method

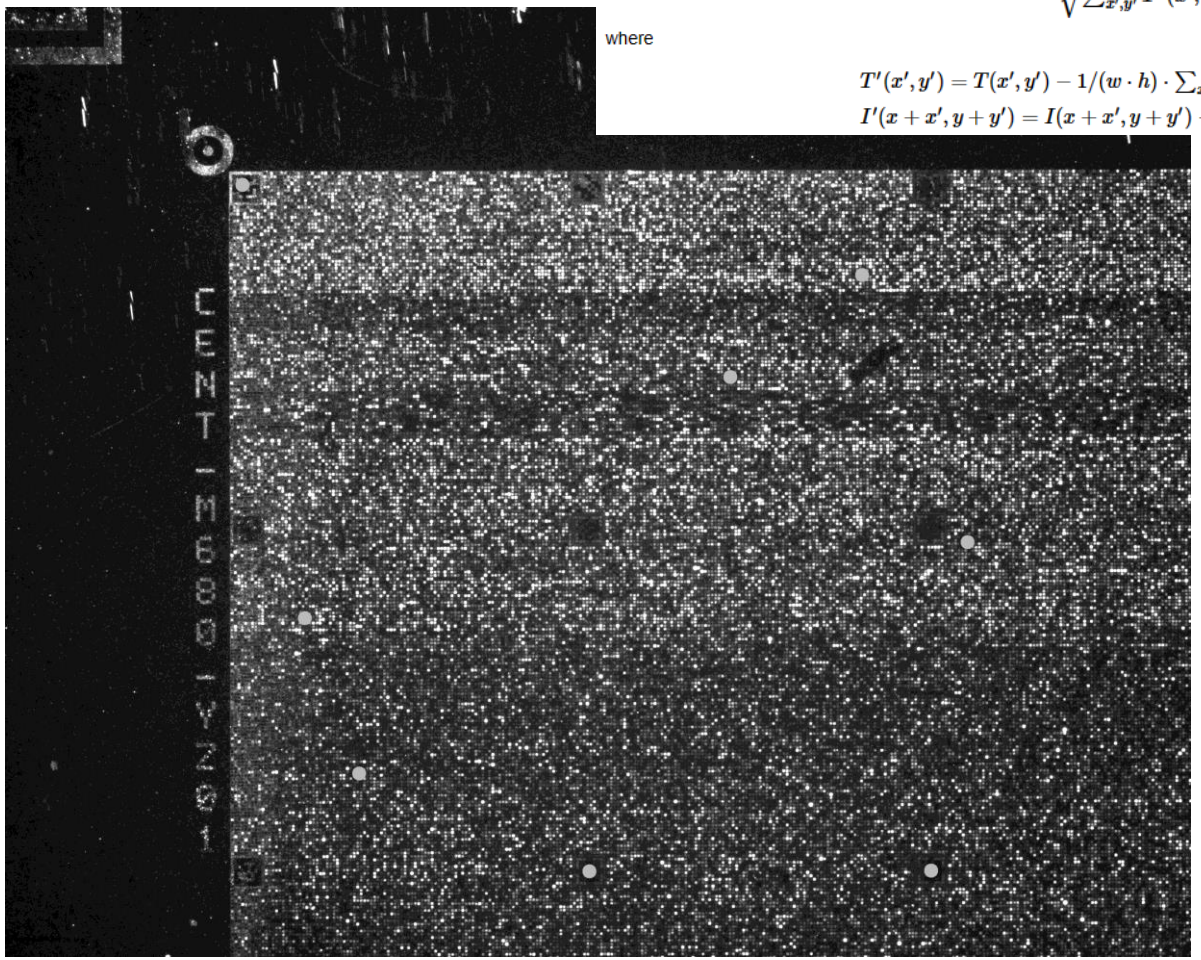
method=TM_CCOEFF_NORMED

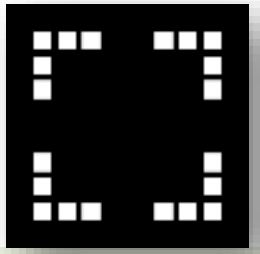
$$R(x,y) = \frac{\sum_{x',y'}(T'(x',y') \cdot I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}}$$

where

$$T'(x',y') = T(x',y') - 1/(w \cdot h) \cdot \sum_{x'',y''} T(x'',y'')$$

$$I'(x+x',y+y') = I(x+x',y+y') - 1/(w \cdot h) \cdot \sum_{x'',y''} I(x+x'',y+y'')$$



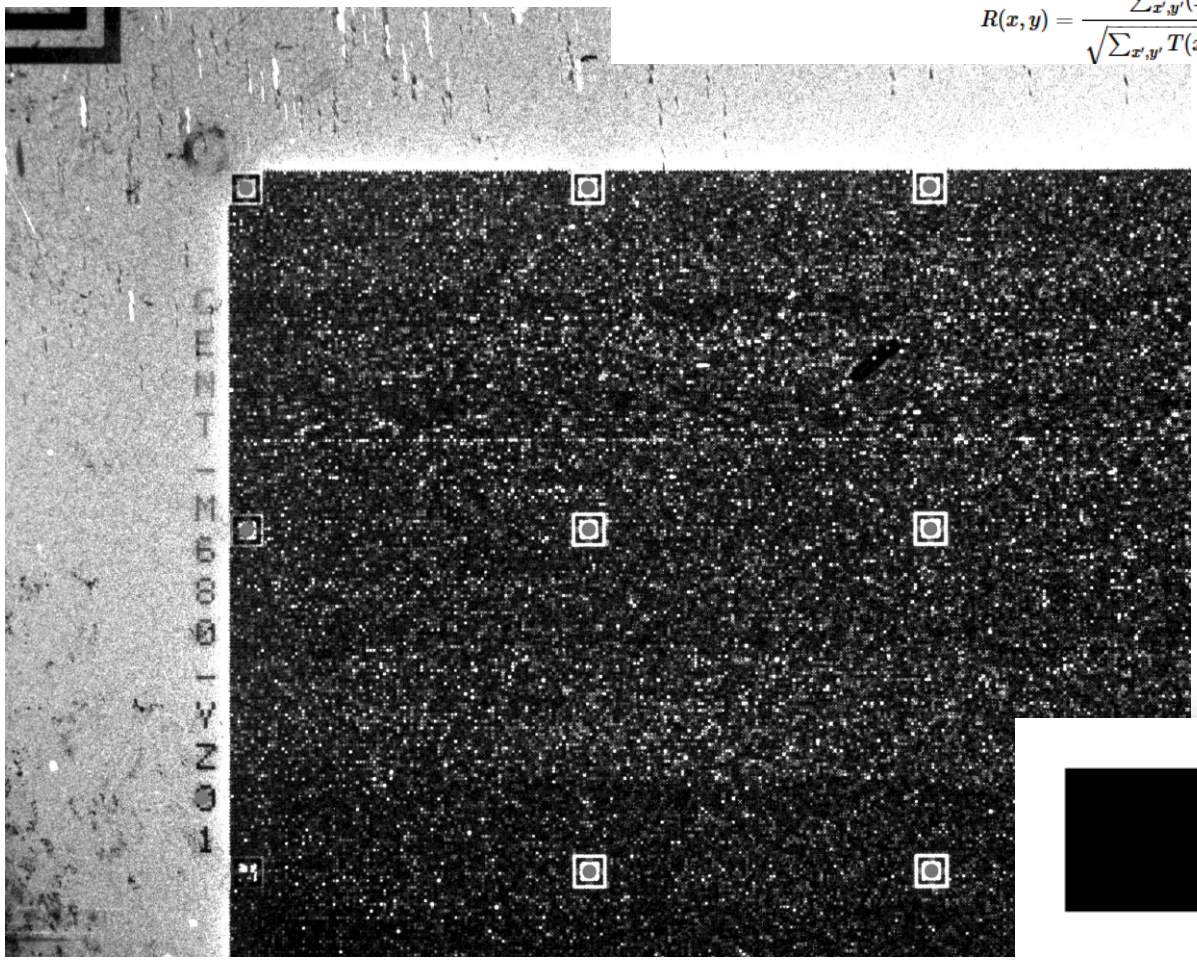


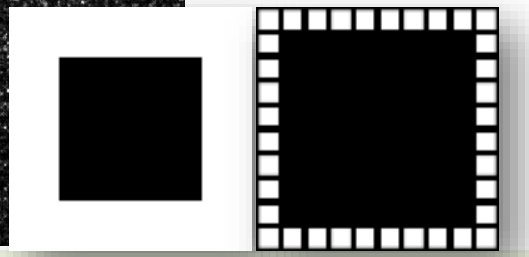
Performance for New Gridding Algo.

- Summit.Grid
  - Old algorithm

method=TM_CCORR_NORMED

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x', y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x', y+y')^2}}$$





Performance for New Gridding Algo.

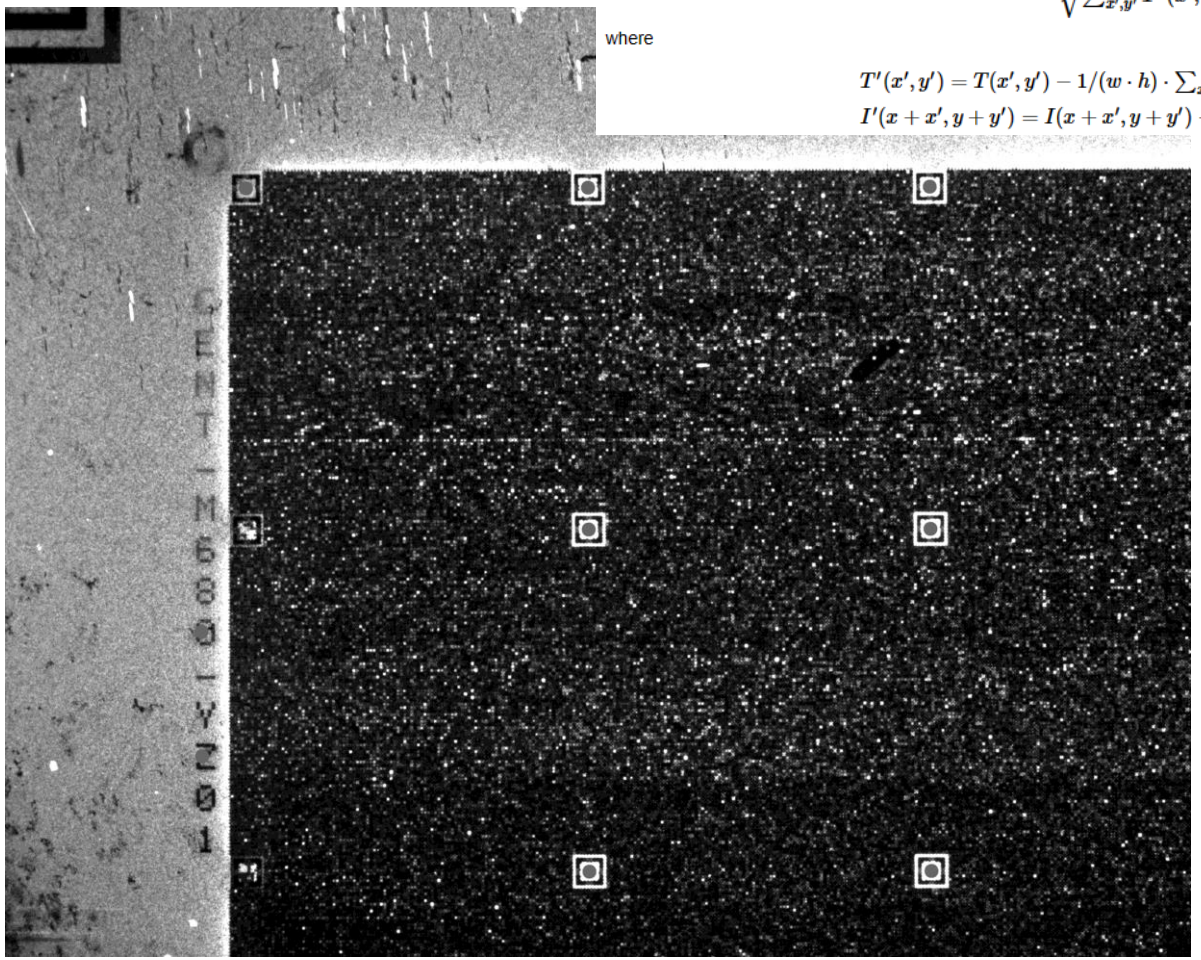
- Summit.Grid
  - New statistical method

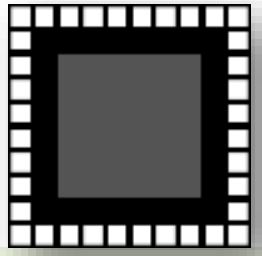
method=TM_CCOEFF_NORMED

$$R(x,y) = \frac{\sum_{x',y'}(T'(x',y') \cdot I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}}$$

where

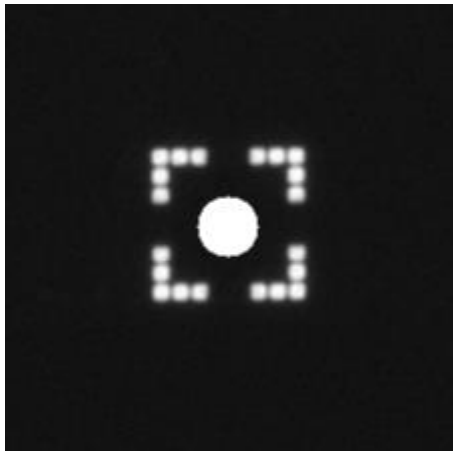
$$T'(x',y') = T(x',y') - 1/(w \cdot h) \cdot \sum_{x'',y''} T(x'',y'')$$
$$I'(x+x',y+y') = I(x+x',y+y') - 1/(w \cdot h) \cdot \sum_{x'',y''} I(x+x'',y+y'')$$





Statistical Conclusion for Experiments Results

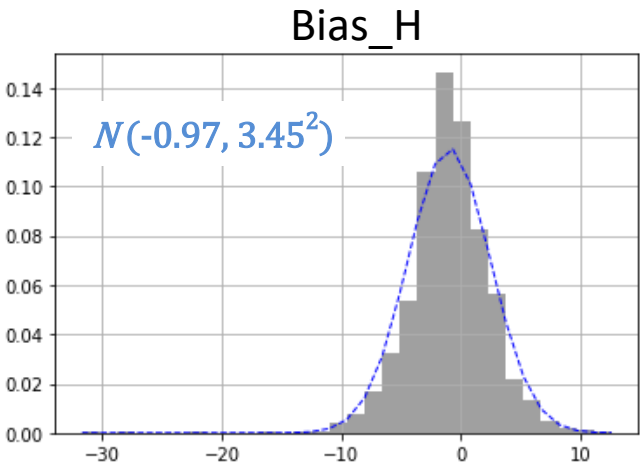
- Gridding
  - SUMMIT Parameters Estimation
    - Let $X \equiv r.v.$ of the displacement from changing the filter (BF -> fluorescent).
    - Let $Y \equiv r.v.$ of the displacement from relocating the plate to the same position.
    - In quick scan mode,
      Estimate $Var(X+Y) = Var(X) + Var(Y)$
    - In regular high precision mode,
      Estimate $Var(X)$ Only
  - Chebyshev's Inequality
    - $P(|Z-\mu| \geq k \cdot \sigma) \leq \frac{1}{k^2}$
    - A. Quick scan for estimating $Var(X+Y)$
      - $k = 14.3, \sigma = 7.74,$ Cover radius: 110.7 (pixels)
      - 99.5% ↑ confidence.
      - 2.3 x BF_mark_size, Lower bound: 110.7 (pixels)
    - B. Regular scan for estimating $Var(X)$
      - $k = 7.2, \sigma = 1.01,$ Cover radius: 7.272 (pixels)
      - 98.06% ↑ confidence.
      - 0.15 x BF_mark_size
  - Performance - successfully recognized rate: nearly 100%.

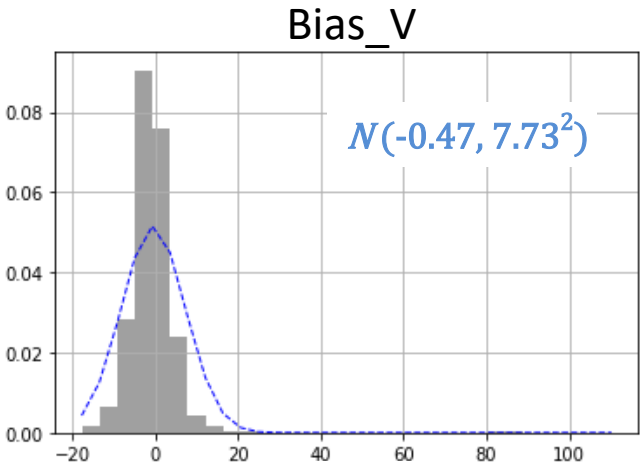


Sampling Distribution for Different Chip Scan Mode

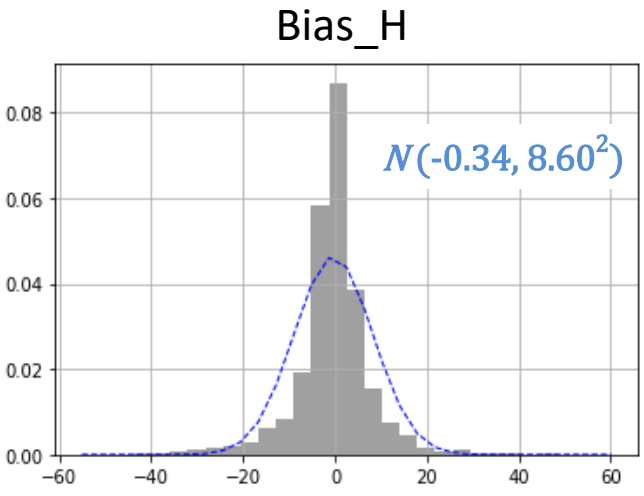
- Quick scan mode
  - Sample: 5 YZ01 chips (7x7 FOVs) x 10 runs => 2450 FOVs
  - Estimation: Var(X+Y)

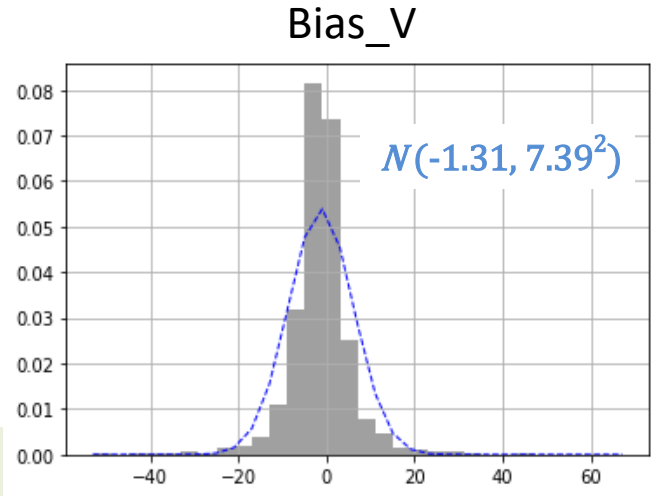
SUMMIT Test 2





SUMMIT Test 3

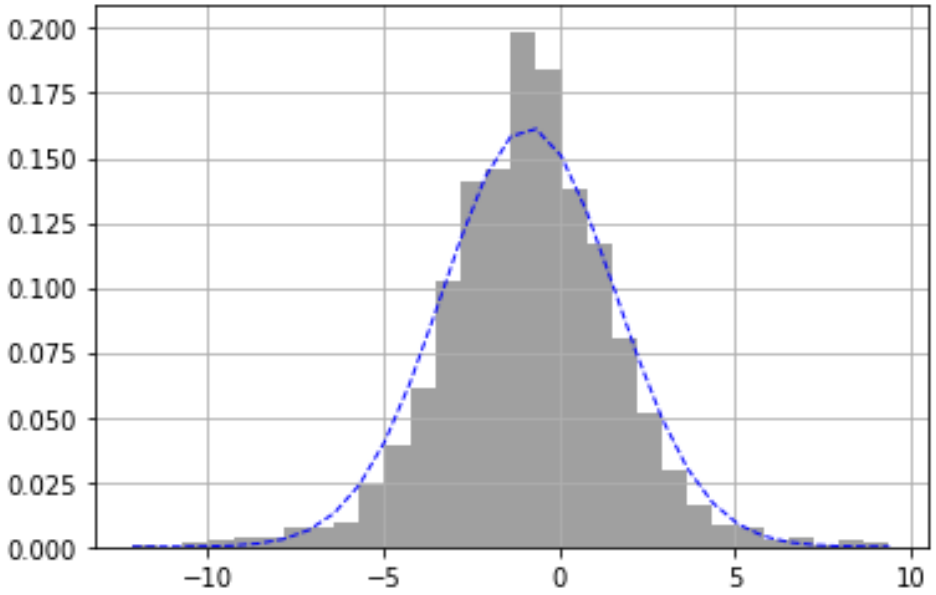




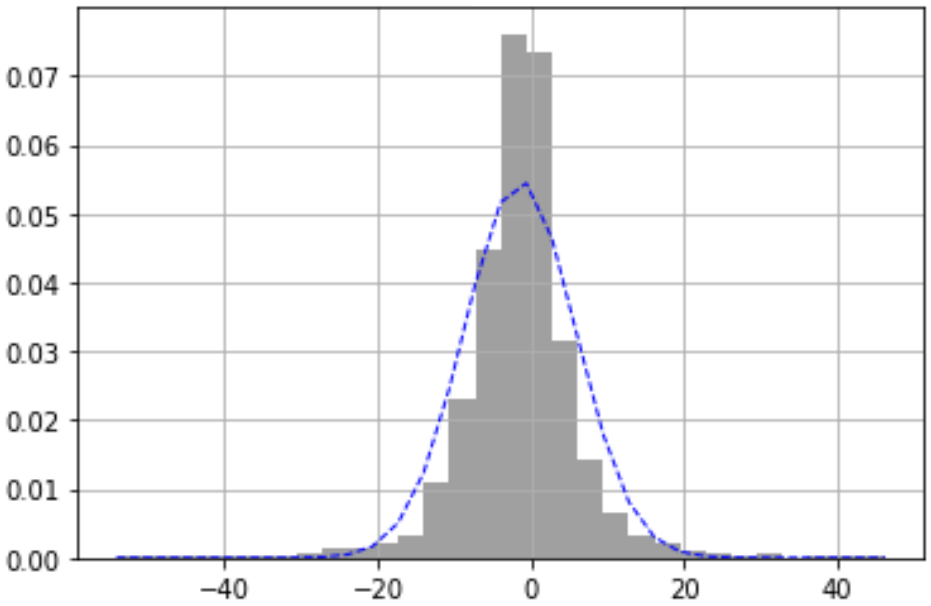
Sampling Distribution for Different Chip Scan Mode

- Quick scan mode
  - Sample: 5 YZ01 chips (7x7 FOVs) x 10 runs => 2450 FOVs
  - Estimation: Var(X+Y)
  - SUMMIT with precise sliding

$N(-0.86, 2.47^2)$ Bias_H



$N(-1.43, 7.23^2)$ Bias_V

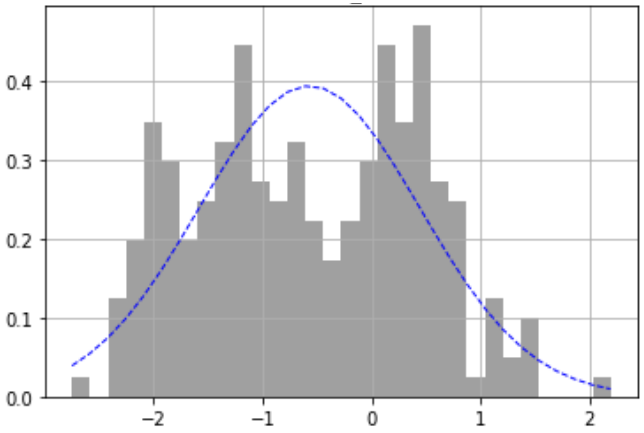


Sampling Distribution for Different Chip Scan Mode

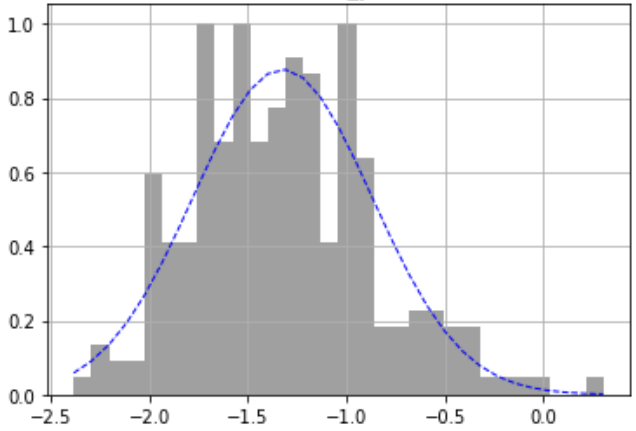
- Regular high precision mode
  - Sample: 5 YZ01 chips (7x7 FOVs) x 1 runs => 245 FOVs
  - Estimation: Var(X)

SUMMIT Test 2

$N(-0.57, 1.01^2)$ Bias_H

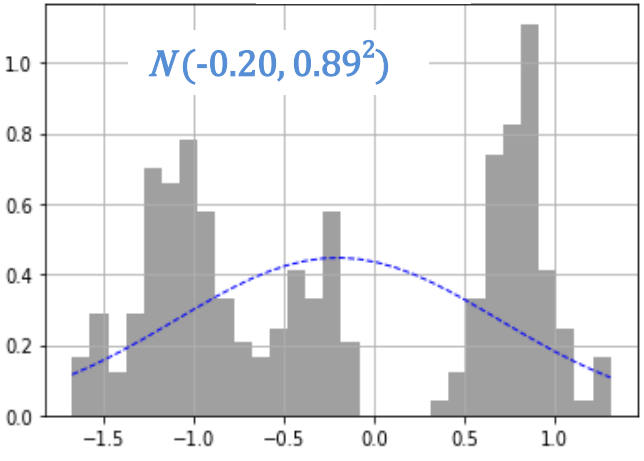


$N(-1.32, 0.45^2)$ Bias_V

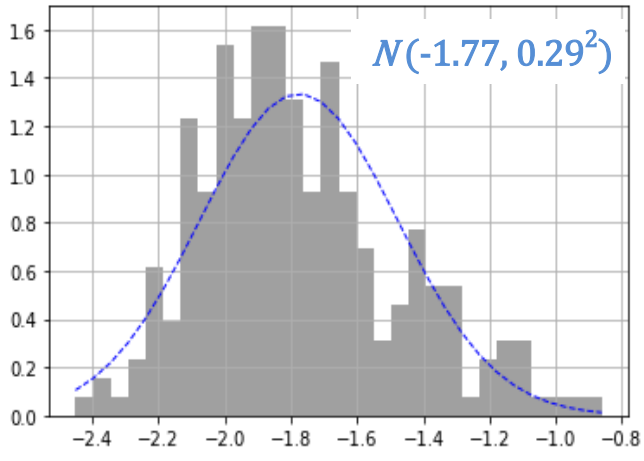


SUMMIT Test 3

Bias_H



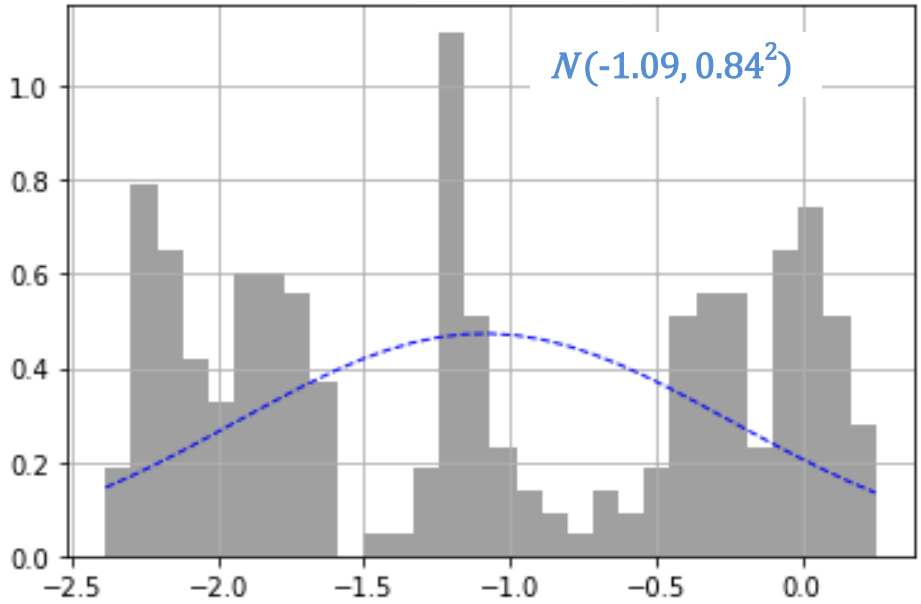
Bias_V



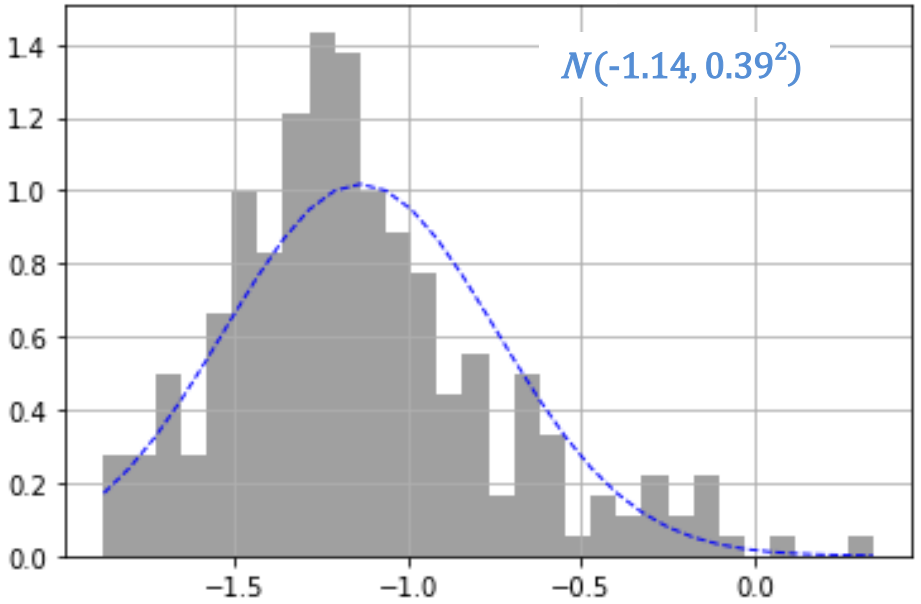
Sampling Distribution for Different Chip Scan Mode

- Regular high precision mode
  - Sample: 5 YZ01 chips (7x7 FOVs) x 1 runs => 245 FOVs
  - Estimation: Var(X)
  - SUMMIT with Precise Sliding

Bias_H



Bias_V



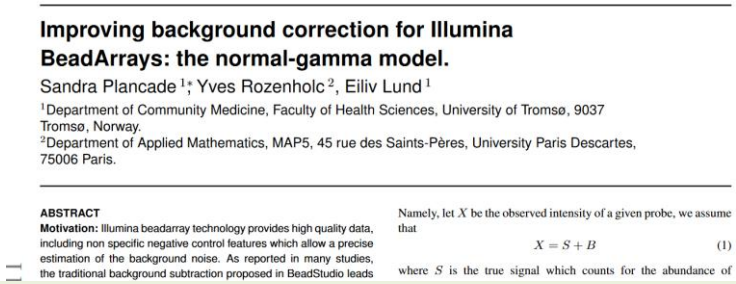
Normal Gamma Background Correction

Jeff (CHI-HSUAN HO)

- **Model Assumption**
  - For each single array:

$$X_j = S_j + B_j$$

  - $BgC: X_j \Rightarrow S_j$ Enhance the biological validity of the results.



Improving background correction for Illumina BeadArrays: the normal-gamma model.

Sandra Plancade [1]*, Yves Rozenholc [2], Eiliv Lund [1]

[1]Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway.
[2]Department of Applied Mathematics, MAP5, 45 rue des Saints-Pères, University Paris Descartes, 75006 Paris.

ABSTRACT

Motivation: Illumina beadarray technology provides high quality data, including non specific negative control features which allow a precise estimation of the background noise. As reported in many studies, the traditional background subtraction proposed in BeadStudio leads

Namely, let X be the observed intensity of a given probe, we assume that

$$X = S + B \quad (1)$$

where S is the true signal which counts for the abundance of

- **Models and Notations**
  - For each single array j:

$$X_j = S_j + B_j$$

  - $X_j = \begin{cases} S_j + B_j, & j \in J \Rightarrow regular\ probes\ set \\ 0 + B_j = B_j, & j \in J_0 \Rightarrow negative\ control\ probes\ set \end{cases}$
  - $X_j \sim f_x(x),\ S_j \sim f_s(s),\ B_j \sim f_B(b),\ S_j\ and\ B_j\ are\ independent.$
  - $N(\mu, \sigma^2) \Rightarrow f_{\mu,\sigma}^{norm}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
  - $\phi(x) \Rightarrow N(0,1),\quad \Phi(t) = \int_{-\infty}^{t} \phi(x)dx$
  - $Gamma(k,\theta) \Rightarrow f_{k,\theta}^{gam}(x) = \frac{\left(\frac{1}{\theta}\right)^k}{\Gamma(k)} x^{k-1} \exp\left\{-\frac{x}{\theta}\right\}, k: shape\ parameter, \theta: scale\ parameter$
    $\xrightarrow[k=1,\theta=\alpha]{} Exp(\alpha) \Rightarrow f_{\alpha}^{exp}(x) = \frac{1}{\alpha} \exp\left\{\frac{x}{\alpha}\right\}$

- **Models and Notations**
  - $X_j = S_j + B_j, \quad X_j \sim f_x(x), \; S_j \sim f_S(s), \; B_j \sim f_B(b)$
  - By the convolution formula, $x_j = s_j + b_j \Rightarrow b_j = x_j - s_j \Rightarrow |J| = \left|\frac{db_j}{dx_j}\right| = 1$

$\Rightarrow X_j \sim f_x(x) = \int_{-\infty}^{\infty} f_{X,S}(x,s)ds = \int_{-\infty}^{\infty} f_{S,B}(s, x-s)|J|ds = \int_{-\infty}^{\infty} f_{S,B}(s, x-s)ds$

$= \int_{-\infty}^{\infty} f_S(s) f_B(x-s)ds$

$\Rightarrow$ Estimated Signal: $\hat{S}(x) = E[S|X=x] = \int_{-\infty}^{\infty} S f_{S|X=x}(s)ds = \int_{-\infty}^{\infty} S \frac{f_{S,X}(s,x)}{f_x(x)} ds$

$= \frac{\int_{-\infty}^{\infty} S f_{S,X}(s,x)ds}{\int_{-\infty}^{\infty} f_{S,X}(s,x)ds} = \frac{\int_{-\infty}^{\infty} S f_S(s) f_B(x-s)ds}{\int_{-\infty}^{\infty} f_S(s) f_B(x-s)ds}$

  - Thus, if $f_x(x)$ is known $\Rightarrow \hat{S}(x)$ is known.
  - No analytic expression $\Rightarrow$ Fast Fourier Transformation-based (fft) approximation.

- **The normexp Model**
  - $S_j \sim f_s(s) = \begin{cases} Exp(\alpha), & j \in J \\ 0, & j \in J_0 \end{cases}$, $B_j \sim f_B(b) \Rightarrow N(\mu, \sigma^2)$

    $\Rightarrow X_j \sim f_X(x) \equiv f^{nexp}_{\mu,\sigma,\alpha}(x) = \frac{1}{\alpha}\exp\left\{\frac{\sigma^2}{2\alpha^2} - \frac{x-\mu}{\alpha}\right\}\Phi(\bar{x}), \quad where\ \bar{x} = \frac{\left(x-\mu-\frac{\sigma^2}{\alpha}\right)}{\sigma}$

    $\Rightarrow \hat{S}^{nexp}(x|\Theta) = \sigma\left(\bar{x} + \frac{\phi(\bar{x})}{\Phi(\bar{x})}\right), \quad \Theta = (\mu, \sigma, \alpha)$
  - If we know $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ $\Rightarrow$ we know $\hat{S}^{nexp}(x)$
- **The Parameter Estimation of normexp Model**
  A. MLE
  B. Adapted RMA
  C. Non-parametric estimation (NP)
  D. Bayesian estimation

- **The normal-gamma Model**
  - $S_j \sim f_s(s) = \begin{cases} Gamma(k,\theta), & j \in J \\ 0, & j \in J_0 \end{cases}$, $B_j \sim f_B(b) \Rightarrow N(\mu,\sigma^2)$

    $\Rightarrow X_j \sim f_X(x) \equiv f^{ng}_{\mu,\sigma,k,\theta}(x) = \int f^{gam}_{k,\theta}(t) f^{norm}_{\mu,\sigma}(x-t)dt \Rightarrow fft-based\ approximation$

    $\Rightarrow \hat{S}^{ng}(x|\Theta) = \frac{\int s f^{gam}_{k,\theta}(s) f^{norm}_{\mu,\sigma}(x-s)ds}{f^{ng}_{\mu,\sigma,k,\theta}(x)} = \frac{k\theta\left(\int f^{gam}_{k+1,\theta}(s) f^{norm}_{\mu,\sigma}(x-s)ds\right)}{f^{ng}_{\mu,\sigma,k,\theta}(x)}$

    $= \frac{k\theta f^{ng}_{\mu,\sigma,k+1,\theta}(x)}{f^{ng}_{\mu,\sigma,k,\theta}(x)} \Rightarrow fft-based\ approximation$
  - If we know $(\hat{\mu},\hat{\sigma},\hat{k},\hat{\theta})$ $\Rightarrow$ we know $\hat{S}^{ng}(x) \Rightarrow \hat{S}_j = \hat{S}^{ng}(x_j)$
- **The Parameter Estimation of normal-gamma Model**

  A. MLE with classical minimization algorithms (L-BFGS-B)

Performance on the Real Data

- GMM-EM + Normal-gamma BgC.
  - Set the related environment in R. (Data Preprocess, NP Probes, QN & log, QDA)
  - Set the corresponding evaluation tools in R. (NP call rate, No call rate)
  - Run the GMM-EM in R. (10 times => max NP call.)
  - Run the normal-gamma correction (all together) before running QDA.
  - Debug for the no call rate in the NP call analyzer.
  - An example: NP call Analyzer: NP call: 94.4%, call: 66.9%

NP call (log(log(·)))	
GMM + EM	94.2308%
QDA	94.5863%
ALL_NG + QDA	94.2774%

Call rate (log(log(·)))	
GMM + EM	66.8298%
QDA	67.8205%
ALL_NG + QDA	67.7681%

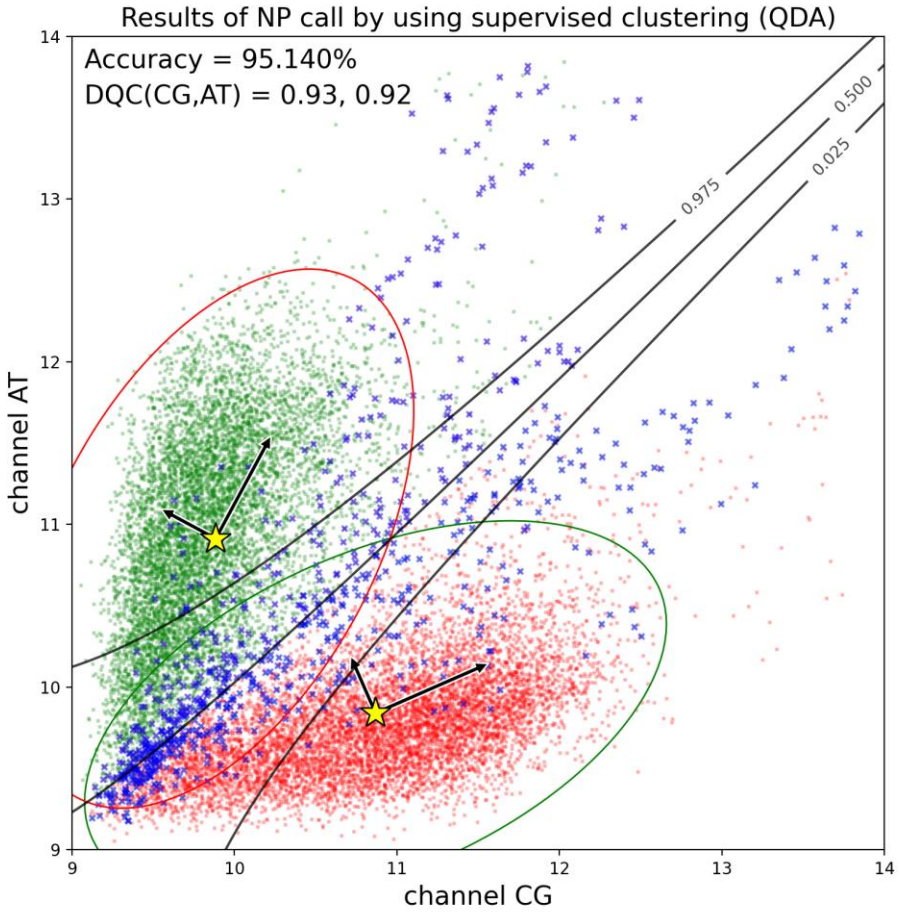
Chip QC and NPcall Analyzer

Jeff (CHI-HSUAN HO)

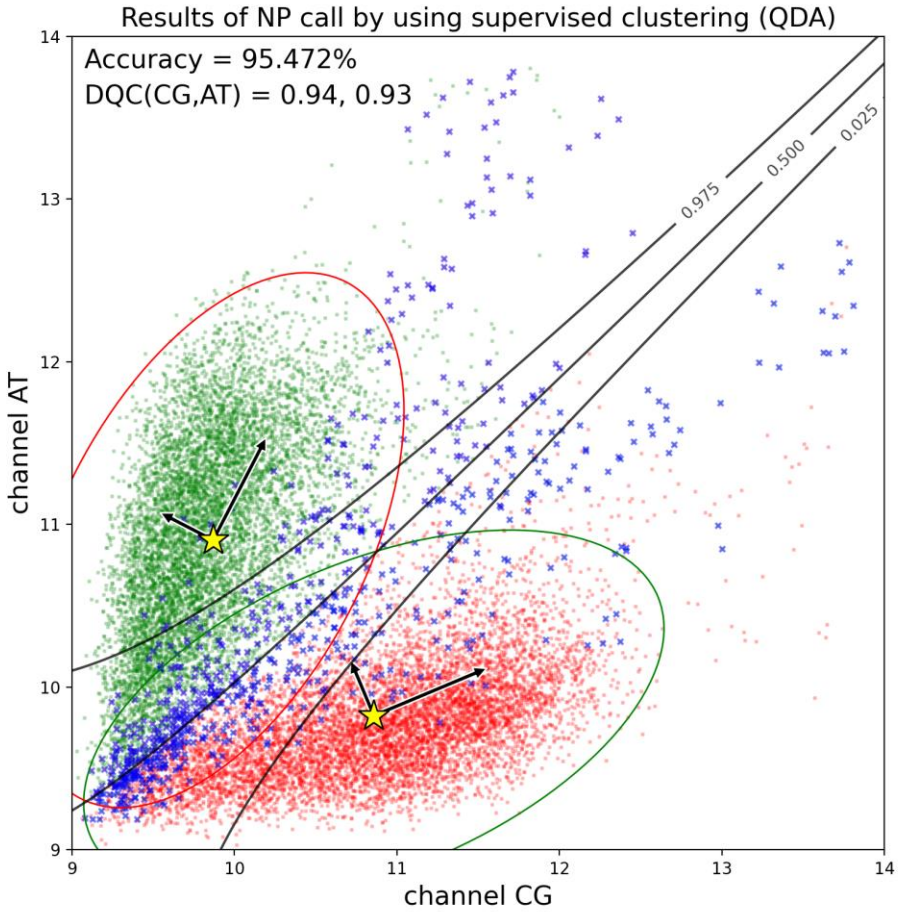
Summit.NPCallAnalyzer

- Summit.Grid

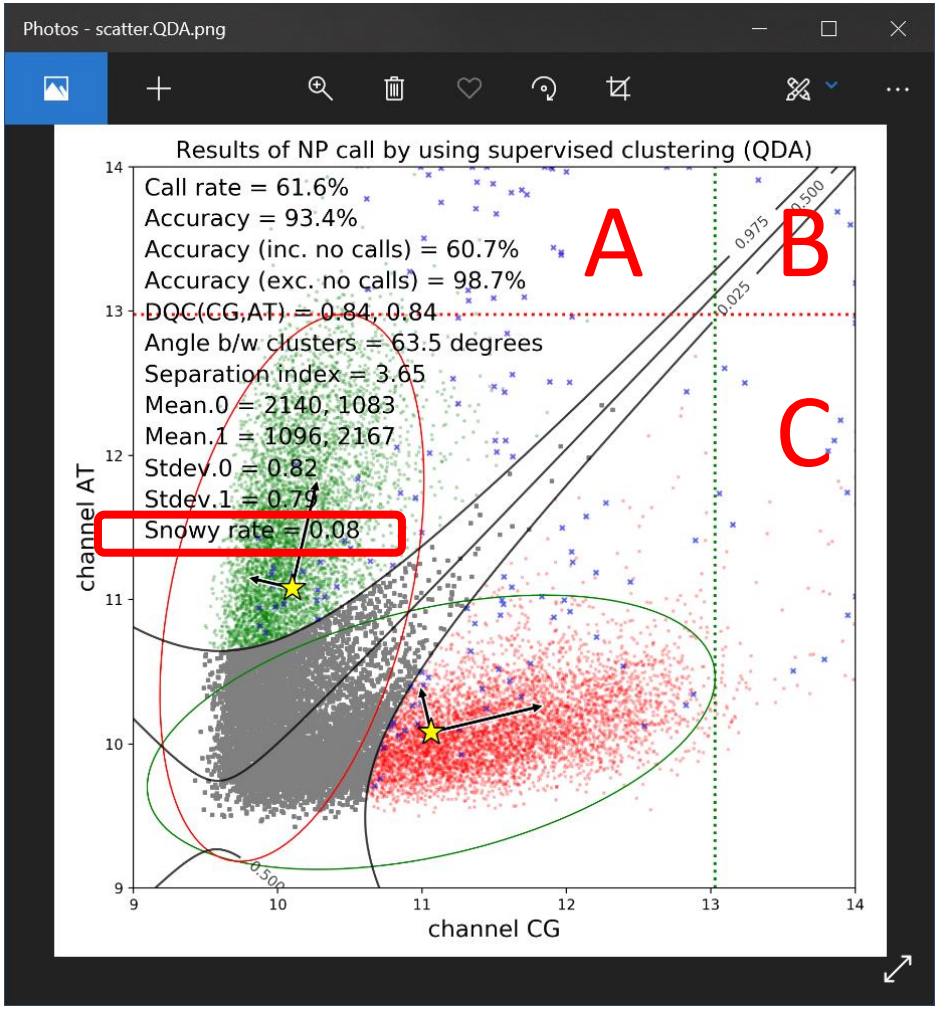
Chip No.54 Quality Control



Chip NO.62 Quality Control



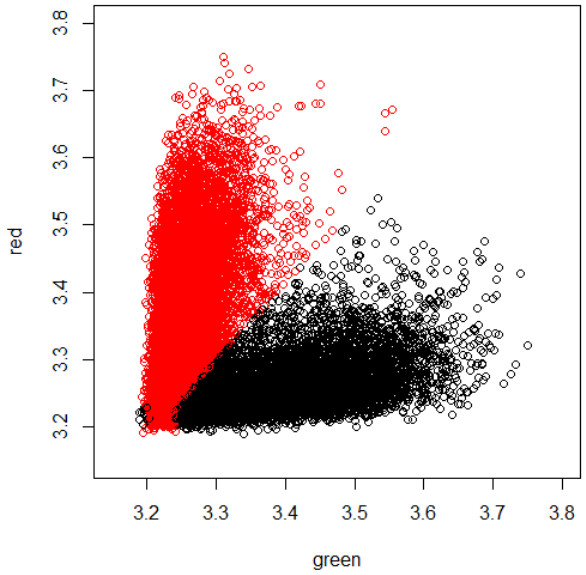
Summit.NPCallAnalyzer

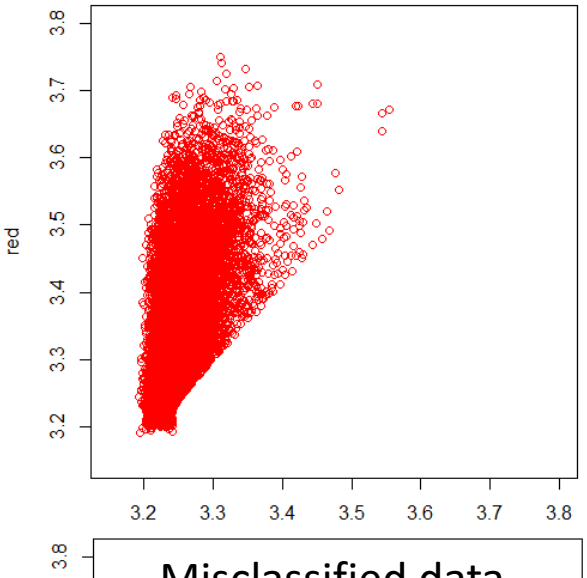


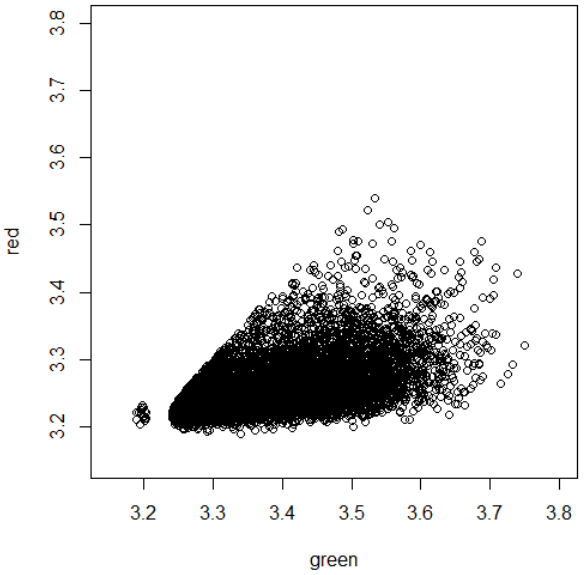
- Snowy rate $= \frac{\text{x } in\ B}{all\ points\ in\ A+B+C}$

Summit.NPCallAnalyzer

- GMM-EM Results (log(log))

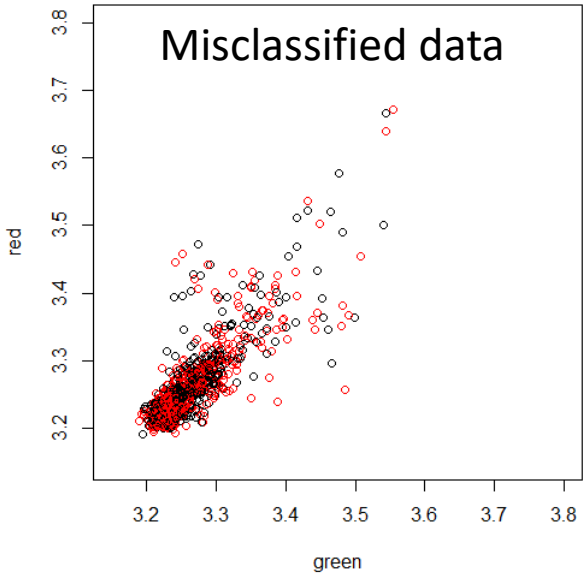




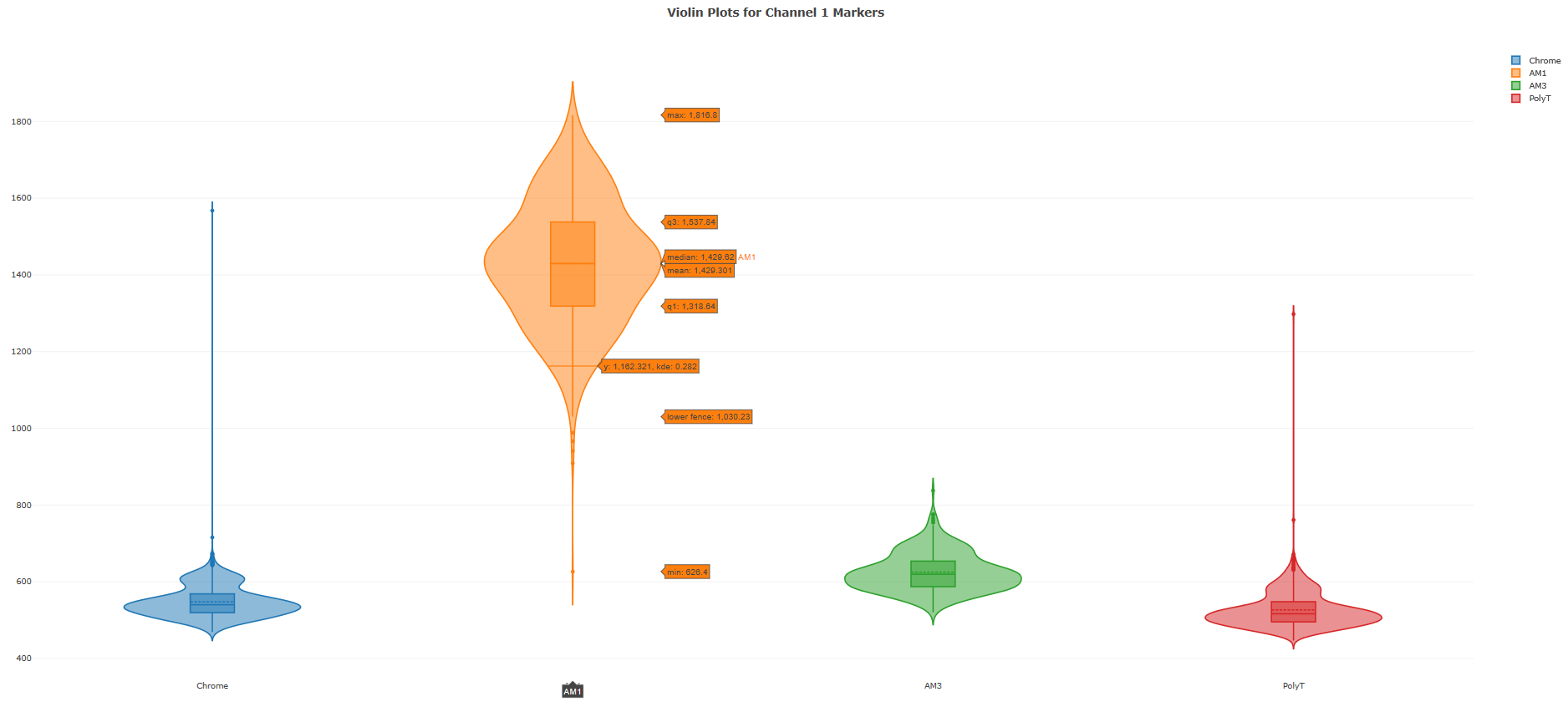


NP call: 94.2308%

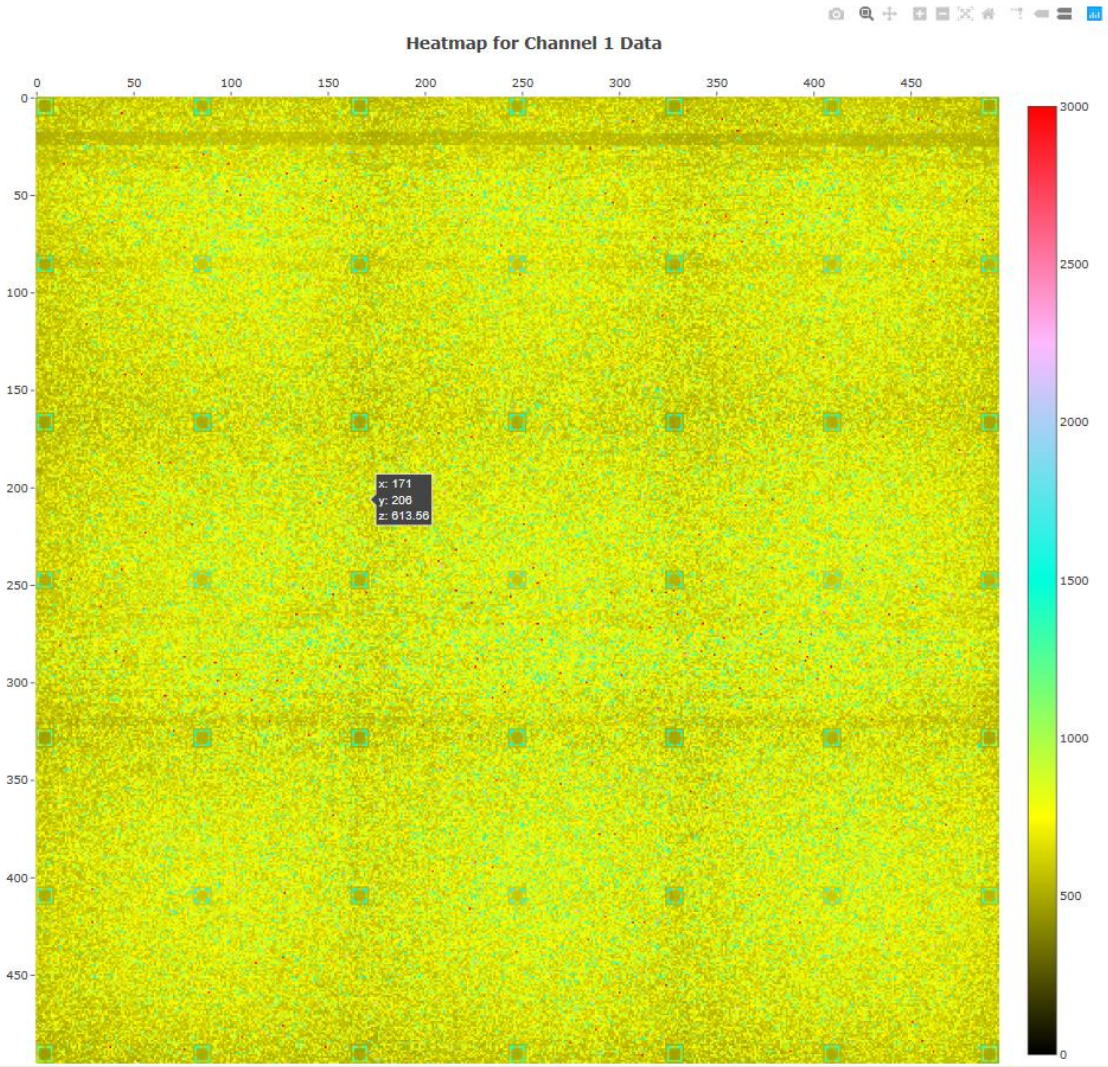
No Call Rate: 66.8298%



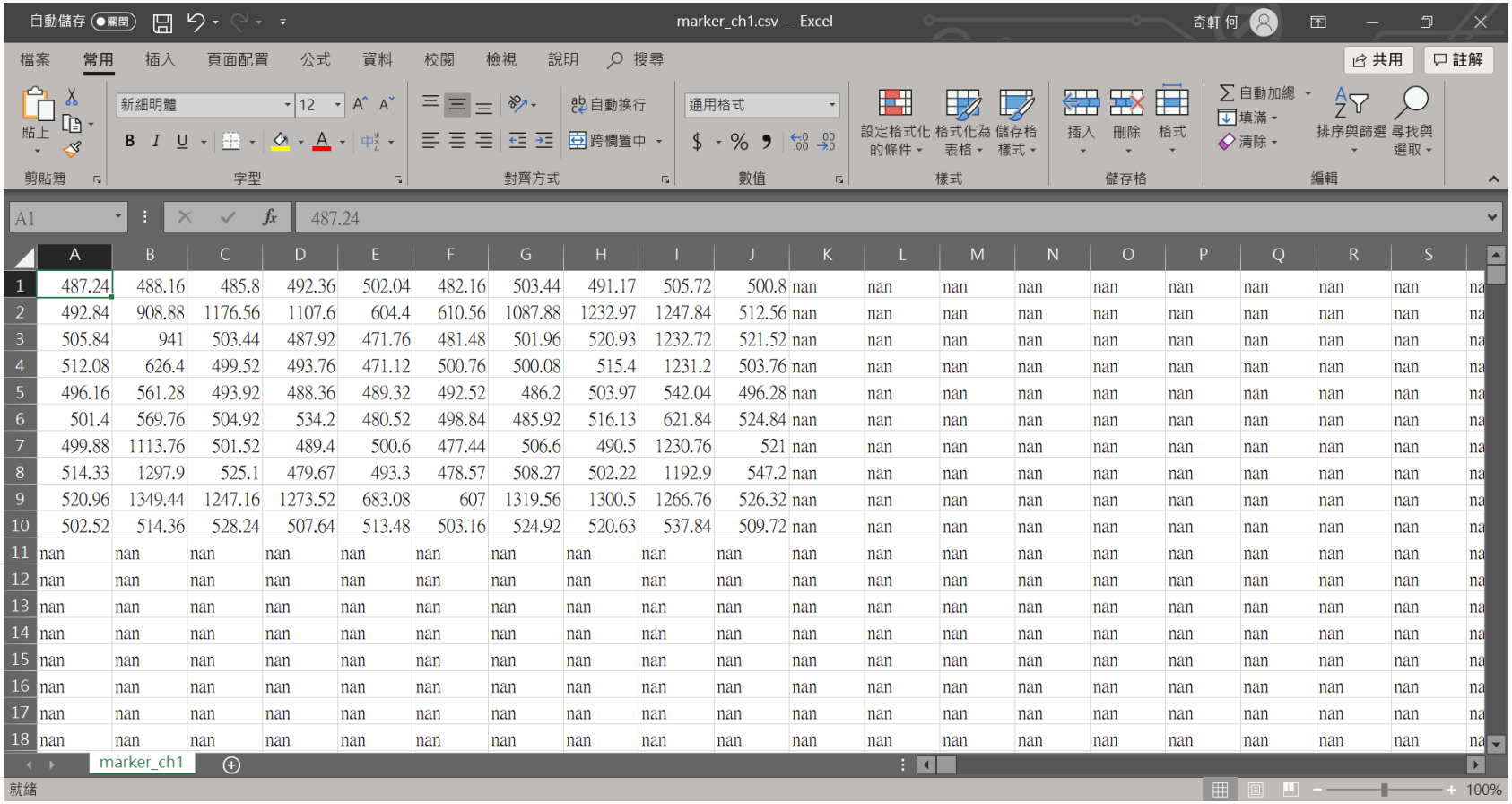
Banff chip QC – Violin Plot



Banff chip QC – Heatmap

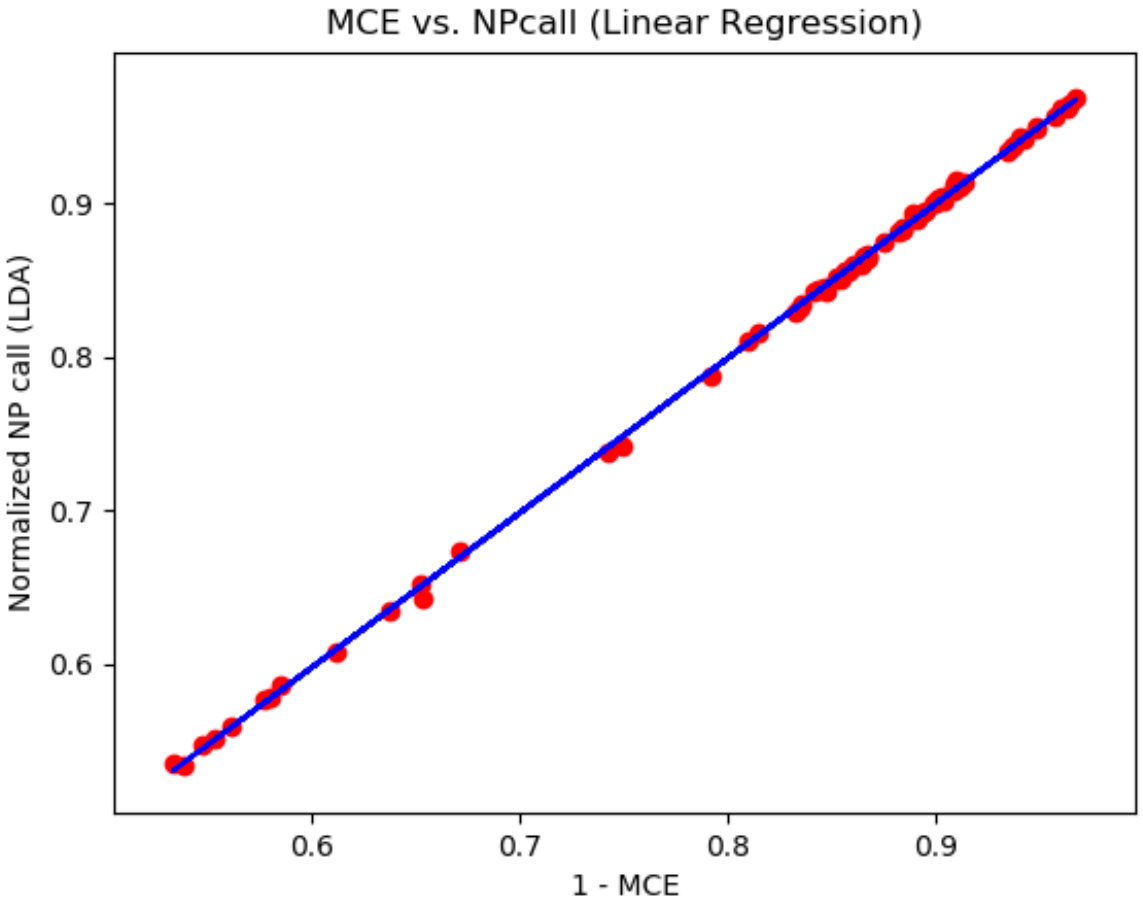


Banff chip QC – Marker Raw Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	487.24	488.16	485.8	492.36	502.04	482.16	503.44	491.17	505.72	500.8	nan	nan	nan	nan	nan	nan	nan	nan	nan
2	492.84	908.88	1176.56	1107.6	604.4	610.56	1087.88	1232.97	1247.84	512.56	nan	nan	nan	nan	nan	nan	nan	nan	nan
3	505.84	941	503.44	487.92	471.76	481.48	501.96	520.93	1232.72	521.52	nan	nan	nan	nan	nan	nan	nan	nan	nan
4	512.08	626.4	499.52	493.76	471.12	500.76	500.08	515.4	1231.2	503.76	nan	nan	nan	nan	nan	nan	nan	nan	nan
5	496.16	561.28	493.92	488.36	489.32	492.52	486.2	503.97	542.04	496.28	nan	nan	nan	nan	nan	nan	nan	nan	nan
6	501.4	569.76	504.92	534.2	480.52	498.84	485.92	516.13	621.84	524.84	nan	nan	nan	nan	nan	nan	nan	nan	nan
7	499.88	1113.76	501.52	489.4	500.6	477.44	506.6	490.5	1230.76	521	nan	nan	nan	nan	nan	nan	nan	nan	nan
8	514.33	1297.9	525.1	479.67	493.3	478.57	508.27	502.22	1192.9	547.2	nan	nan	nan	nan	nan	nan	nan	nan	nan
9	520.96	1349.44	1247.16	1273.52	683.08	607	1319.56	1300.5	1266.76	526.32	nan	nan	nan	nan	nan	nan	nan	nan	nan
10	502.52	514.36	528.24	507.64	513.48	503.16	524.92	520.63	537.84	509.72	nan	nan	nan	nan	nan	nan	nan	nan	nan
11	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
12	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
13	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
14	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
15	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
16	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
17	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
18	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Correlation and Regression analysis (MCE vs. NP call)



Training data: 80% data
Model: $NPcall = -0.005 + 1.005 \cdot (1 - MCE)$
R^2: 99.96%

Testing data: 20% data
MSE: 5.999574703240224e-06

It still need NP data to calculate the MCE.