



References

Jeff (CHI-HSUAN HO)



Bioinformatics Analysis – Genotyping (GT), APT, CPT

Jeff (CHI-HSUAN HO)

Genotyping Analysis Development and Deployment



- 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: <ul style="list-style-type: none">• Axiom Human Arrays:<ul style="list-style-type: none">• 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

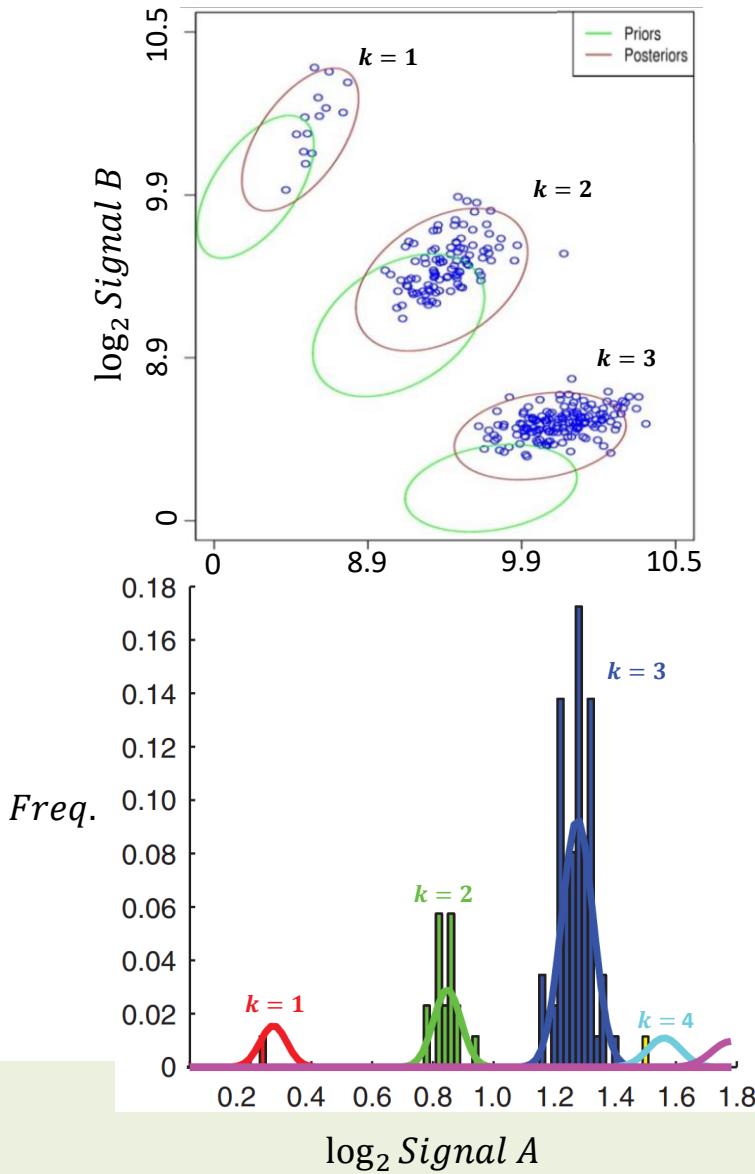
GTC v4.2 P/N 702982 Rev. 3

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

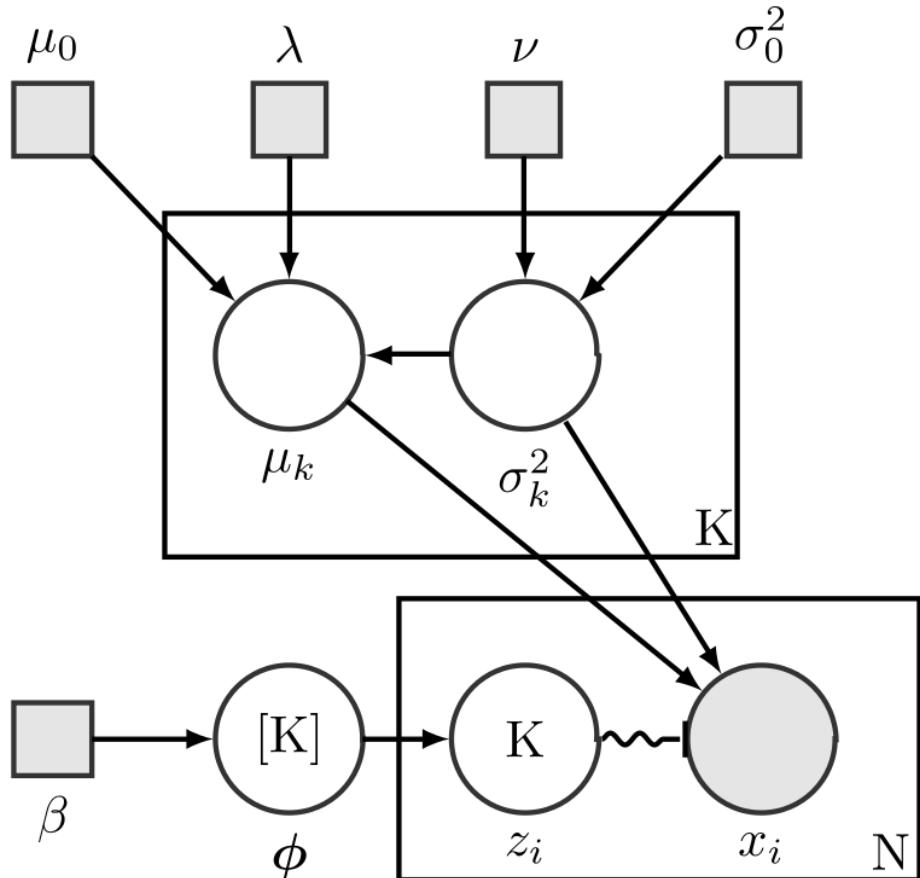
Gaussian Mixture Model (GMM)

SNP_A-2131259



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

Bayesian Gaussian Mixture Model

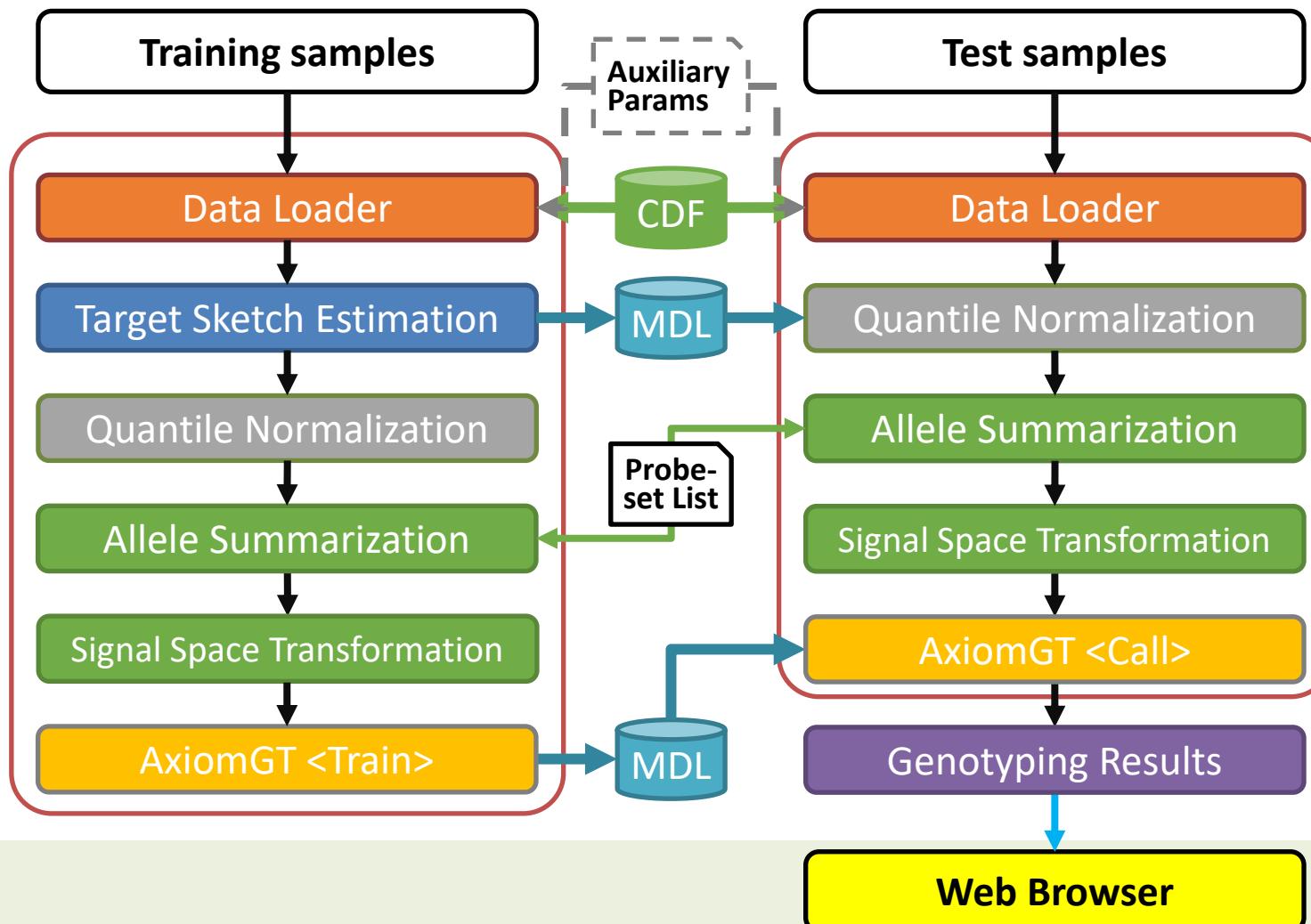


- Model: $\sum_{k=1}^K \phi_k N(\mu_k, \sigma_k^2)$
- K = Number of mixture components
- σ_k^2 = Variance of component k ,
 $\sigma_k^2 \sim \text{Inverse-Wishart}(\nu, \sigma_0^2)$
- μ_k = Mean of component k ,
 $\mu_k \sim \text{Normal}(\mu_0, \lambda \sigma_k^2)$
- N = Number of observations
- z_i = Component (category) of observation i ,
 $z_i \sim \text{Categorical}(\phi), z_i \in \{1, 2, \dots, k\}$
- ϕ_k = Mixture weight, i.e. prior probability of
a particular component k ,
 $\phi \sim \text{Symmetric-Dirichlet}(\beta),$
 $\sum_k \phi_k = 1$
- x_i = Observation i ,
 $x_i \sim \text{Normal}(\mu_{z_i}, \sigma_{z_i}^2)$

Genotyping Analysis Development and Distribution



- AxiomGT Framework



Genotyping Analysis Development and Distribution



Auxiliary Params

Genotype

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<!DOCTYPE boost_serialization>
<boost_serialization signature="serialization::archive" version="15">
<Genohints_table>
    2      -1      0      1      1
    2      2      1      1      2
    0      0      1      1      1
</Genohints_table>
<Probeset_names_for_row class_id="0" tracking_level="0" version="0">
    <count>3</count>
    <item_version>0</item_version>
    <item>AFFX-SNP-000001</item>
    <item>AFFX-SNP-000002</item>
    <item>AFFX-SNP-000003</item>
</Probeset_names>
<Sample_names_for_col>
    <count>5</count>
    <item_version>0</item_version>
    <item>GSM2066668_206-001_CHB</item>
    <item>GSM2066669_206-003_CHB</item>
    <item>GSM2066670_206-004_CHB</item>
    <item>GSM2066671_206-014_CHB</item>
    <item>GSM2066672_206-015_CHB</item>
</Sample_names>
</boost_serialization>
```

Special SNPs

probeset_id	chr	copy_male	copy_female
AX-11086922	X	1	2
AX-11104190	MT	1	1
AX-11106959	Y	1	0
AX-11106974	PAR	2	2
AX-12524149	X	1	2

Meaning of the value: Theoretical copy number for each probeset and gen

- Explanation of the code in the chr column:

code	chr region
X	The non-pseudoautosomal region of the X chromosome
Y	The Y chromosome
MT	Mitochondrial SNPs
PAR	The pseudoautosomal region of the X chromosome

Genders

gender	sample_files
1	Sample01.CEN
1	Sample02.CEN
0	Sample03.CEN
0	Sample04.CEN
1	Sample05.CEN

probe_id	channel_id
288	1
871	1
2014	1

Sex Probes

Genotyping Analysis Development and Distribution



Genotyping Results

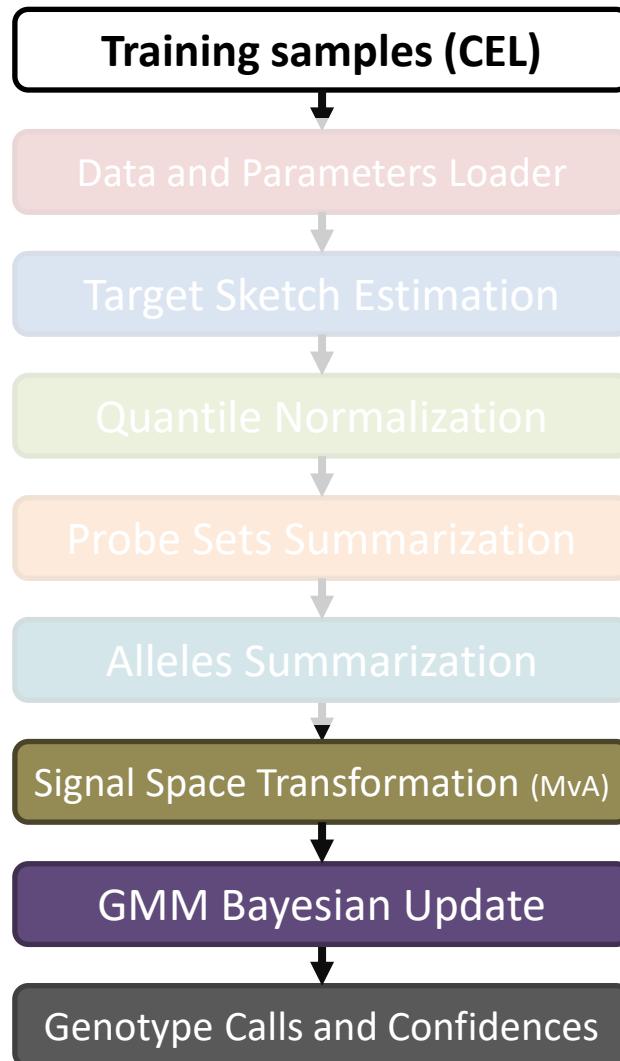
Call Rate

call_rate	sample_name
0.994788	GSM2066668_206-001_CHB
0.994555	GSM2066669_206-003_CHB
0.994841	GSM2066670_206-004_CHB
0.995426	GSM2066671_206-014_CHB
0.994742	GSM2066672_206-015_CHB
0.995266	GSM2066673_206-016_CHB
0.994938	GSM2066674_206-018_CHB
0.995377	GSM2066675_206-019_CHB
0.992347	GSM2066676_206-021_CHB
0.967893	GSM2066677_206-023_CHB
0.972985	GSM2066678_206-028_CHB
0.972951	GSM2066679_206-029_CHB
0.973976	GSM2066680_206-030_CHB
0.973856	GSM2066681_206-032_CHB
0.974752	GSM2066682_206-033_CHB
0.972898	GSM2066683_206-034_CHB

Genotype Inference Report

a_allele	b_allele	genotype	posterior	probeset_name	sample_name
-2.42229	9.69654	2	0.999972	AFFX-SNP-000001	GSM2066668_206-001_CHB
-0.213343	10.0077	1	0.999985	AFFX-SNP-000001	GSM2066669_206-003_CHB
2.62502	9.22009	0	0.999929	AFFX-SNP-000001	GSM2066670_206-004_CHB
-0.312298	9.95922	1	0.99998	AFFX-SNP-000001	GSM2066671_206-014_CHB
-0.264381	9.80488	1	0.999976	AFFX-SNP-000001	GSM2066672_206-015_CHB
0.0327307	10.0347	1	0.999982	AFFX-SNP-000001	GSM2066673_206-016_CHB
-2.92652	9.55797	2	0.999928	AFFX-SNP-000001	GSM2066674_206-018_CHB
-2.49338	9.65748	2	0.999975	AFFX-SNP-000001	GSM2066675_206-019_CHB
0.0407013	9.7349	1	0.999969	AFFX-SNP-000001	GSM2066676_206-021_CHB
-2.94063	9.34759	2	0.999895	AFFX-SNP-000001	GSM2066677_206-023_CHB
3.0863	9.23524	0	0.999049	AFFX-SNP-000001	GSM2066678_206-028_CHB
2.45937	9.3949	0	0.99994	AFFX-SNP-000001	GSM2066679_206-029_CHB
0.0585673	10.002	1	0.99998	AFFX-SNP-000001	GSM2066680_206-030_CHB
-0.260712	10.2414	1	0.999986	AFFX-SNP-000001	GSM2066681_206-032_CHB
-0.468117	9.97722	1	0.999958	AFFX-SNP-000001	GSM2066682_206-033_CHB
-3.29787	9.4166	2	0.996784	AFFX-SNP-000001	GSM2066683_206-034_CHB
-0.125064	10.003	1	0.999986	AFFX-SNP-000001	GSM2066684_206-038_CHB
-3.08476	9.3733	2	0.999697	AFFX-SNP-000001	GSM2066685_206-040_CHB
-0.0546291	9.93013	1	0.999983	AFFX-SNP-000001	GSM2066686_206-041_CHB
2.82779	9.32064	0	0.999878	AFFX-SNP-000001	GSM2066687_206-043_CHB
-2.80064	9.3799	2	0.99995	AFFX-SNP-000001	GSM2066688_206-044_CHB
-0.144338	9.56211	1	0.99996	AFFX-SNP-000001	GSM2066689_206-045_CHB
-2.77376	9.40567	2	0.999957	AFFX-SNP-000001	GSM2066690_206-048_CHB
-0.0830618	9.82269	1	0.99998	AFFX-SNP-000001	GSM2066691_206-049_CHB
-0.286435	10.002	1	0.999982	AFFX-SNP-000001	GSM2066692_206-054_CHB
-0.0575499	9.98275	1	0.999985	AFFX-SNP-000001	GSM2066693_206-056_CHB
-3.01959	9.36143	2	0.999822	AFFX-SNP-000001	GSM2066694_211-001_CHB
-2.95238	9.45029	2	0.999904	AFFX-SNP-000001	GSM2066695_211-003_CHB
-2.94825	9.37206	2	0.999895	AFFX-SNP-000001	GSM2066696_211-006_CHB
-3.12067	9.36538	2	0.999569	AFFX-SNP-000001	GSM2066697_211-009_CHB
-3.72513	9.12557	-1	0.123354	AFFX-SNP-000001	GSM2066698_211-010_CHB

Genotyping Analysis Development and Distribution

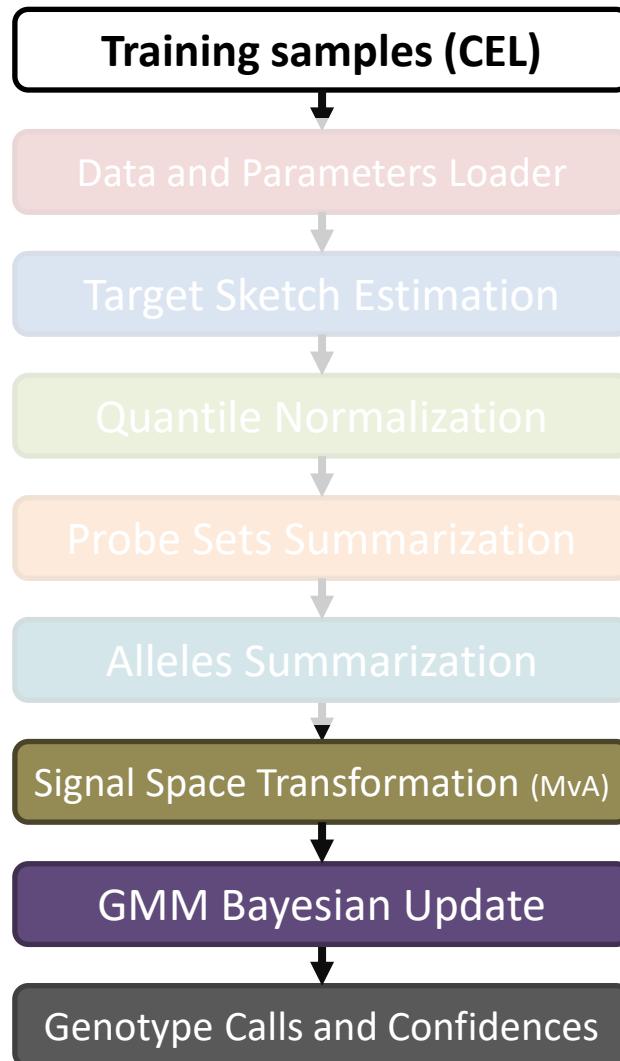


- AxiomGT1 Algo (`bayes_label`)
 - `initialize_bins`
($x \rightarrow \text{bins domain}$)
 - `Integratebrlmmoverlabelings`
(bins domain)
 - `labels_two_posterior`
(bins domain)
 - `make_two_calls`
($\text{bins domain params results} \rightarrow x \text{ calls}$)

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution

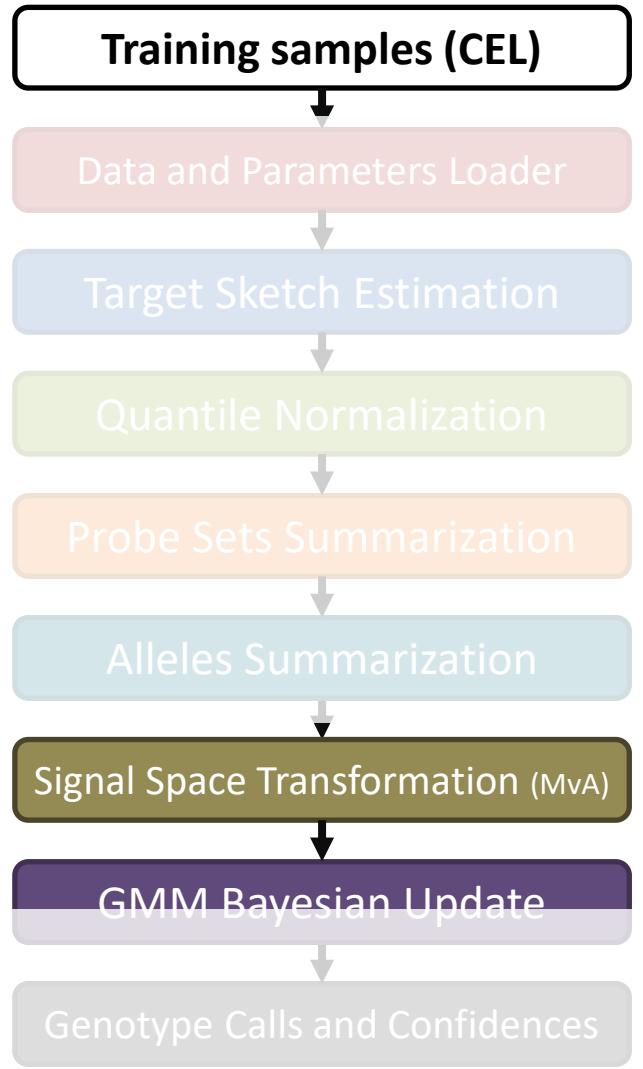


- Sort (x, y) and other related params by x (increasing order).
- Create bins data (setup_bins
 $(x \rightarrow \text{bins domain})$)
 $\text{delta} = \frac{\text{Range}(x)}{\text{sp.bins} + 1}, \quad \text{sp.bins} = 100 \text{ (default)}$
 - Create a new bin and reset collected statistics if
 - When current data point x lies outside current boundary (previous $x + \text{delta}$) \Rightarrow Create new boundary by using current x .
 - $\text{sp.bins} = 0$ (bins is turned off)
 - When data points are fewer than bins.
 - Statistics are collected and computed in the same bin:
 - Data points number
 - $x, x^2, y, y^2, x \cdot y$
 - Penalty from each known genotype and inbreeding status (Only used in supervised mode)

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

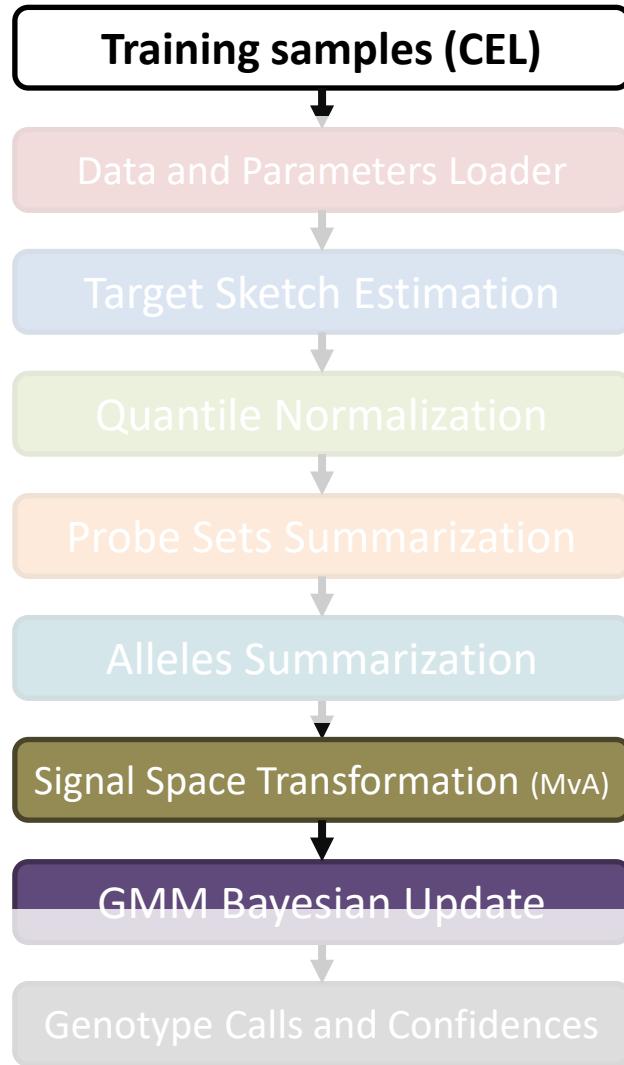
Genotyping Analysis Development and Distribution



- **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.**

Centrillion Confidential

Genotyping Analysis Development and Distribution

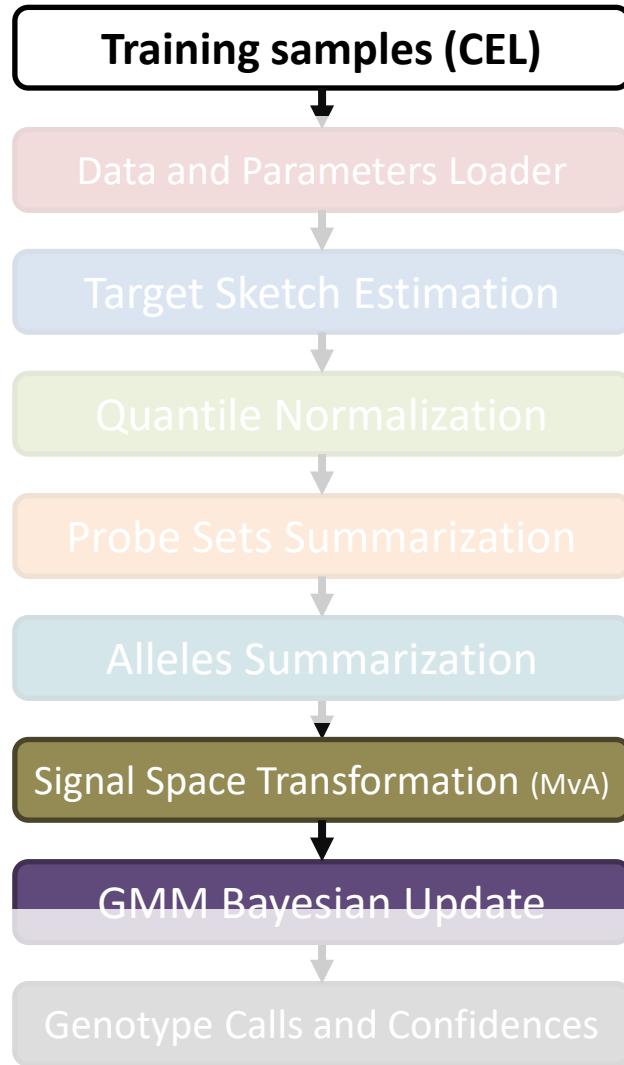


- **Quality Score (Posterior Analog) (for each partition)**
 - **(sp.mix) mixture_penalty**
 - $\sum_{g=1}^3 N_g * \log\left(\frac{(N_g + \text{lambda})}{\sum_g (N_g + \text{lambda})}\right)$, lambda = 1 (BB), 2 (AB), 3 (AA)
 - **(sp.bic) BIC (k*log(N) part)**
 $c * \text{bic_k} * \log\left(\sum_g N_g\right)$, bic_k = 2 \Rightarrow mean, var, c = 1,2,3
 - **1D (x) Log likelihood under posterior params**
 - $u'_{3x1} = (K_0^{-1}_{3x3} + N'_{3x3})^{-1} * (K_0^{-1}_{3x3} * u_{0 3x1} + N_{3x3} * m_{3x1})$,
 - $K_0^{-1}_{3x3} = \begin{bmatrix} \frac{k_{10}}{\sigma_{10}^2} & \frac{\sigma_{120}}{\sigma_{10}\sigma_{20}} & \frac{\sigma_{130}}{\sigma_{10}\sigma_{30}} \\ \frac{\sigma_{120}}{\sigma_{10}\sigma_{20}} & \frac{k_{20}}{\sigma_{20}^2} & \frac{\sigma_{230}}{\sigma_{20}\sigma_{30}} \\ \frac{\sigma_{130}}{\sigma_{10}\sigma_{30}} & \frac{\sigma_{230}}{\sigma_{20}\sigma_{30}} & \frac{k_{30}}{\sigma_{30}^2} \end{bmatrix}$,
 - $N'_{3x3} = \begin{bmatrix} \frac{N_1}{\sigma_{10}^2} & 0 & 0 \\ 0 & \frac{N_2}{\sigma_{20}^2} & 0 \\ 0 & 0 & \frac{N_3}{\sigma_{30}^2} \end{bmatrix}$
 - $u_{0 3x1} = \begin{bmatrix} u_{10} \\ u_{20} \\ u_{30} \end{bmatrix}$,
 - $m_{3x1} = \begin{bmatrix} \sum_i x_{1i} \\ \sum_i x_{2i} \\ \sum_i x_{3i} \end{bmatrix}$,
 - $N_{3x3} = \begin{bmatrix} \frac{1}{\sigma_{10}^2} & 0 & 0 \\ 0 & \frac{1}{\sigma_{20}^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_{30}^2} \end{bmatrix}$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- **Quality Score (Posterior Analog) (for each partition)**
- **1D (x) Log likelihood under posterior params**

■ $u'_{3x1} = (K_0^{-1}_{3x3} + N'_{3x3})^{-1} * (K_0^{-1}_{3x3} * u_{0,3x1} + N_{3x3} * m_{3x1})$

■ (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)). u'_{3x1}

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

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

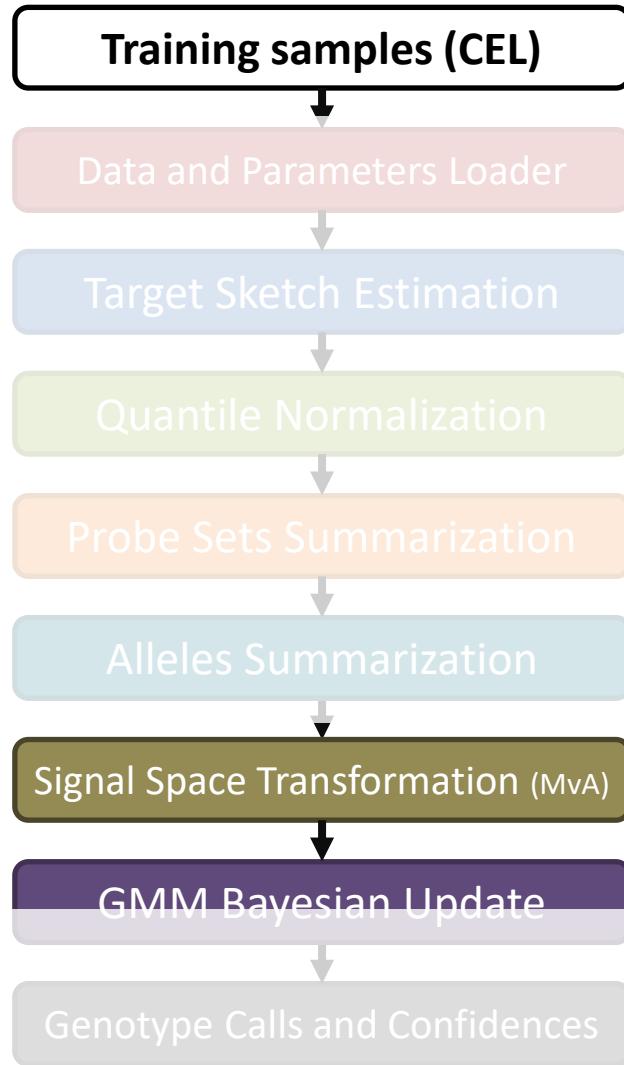
$$u'_{3x1} = \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \end{bmatrix}, \quad \begin{array}{ll} u'_1 = u'_1 + \text{delta} - \text{gamma} \\ u'_2 = u'_2 - \text{gamma} \\ u'_3 = u'_3 - \text{delta} - \text{gamma} \end{array}$$

Pool Adjacent-Violators (PAV) algo.

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

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

Genotyping Analysis Development and Distribution



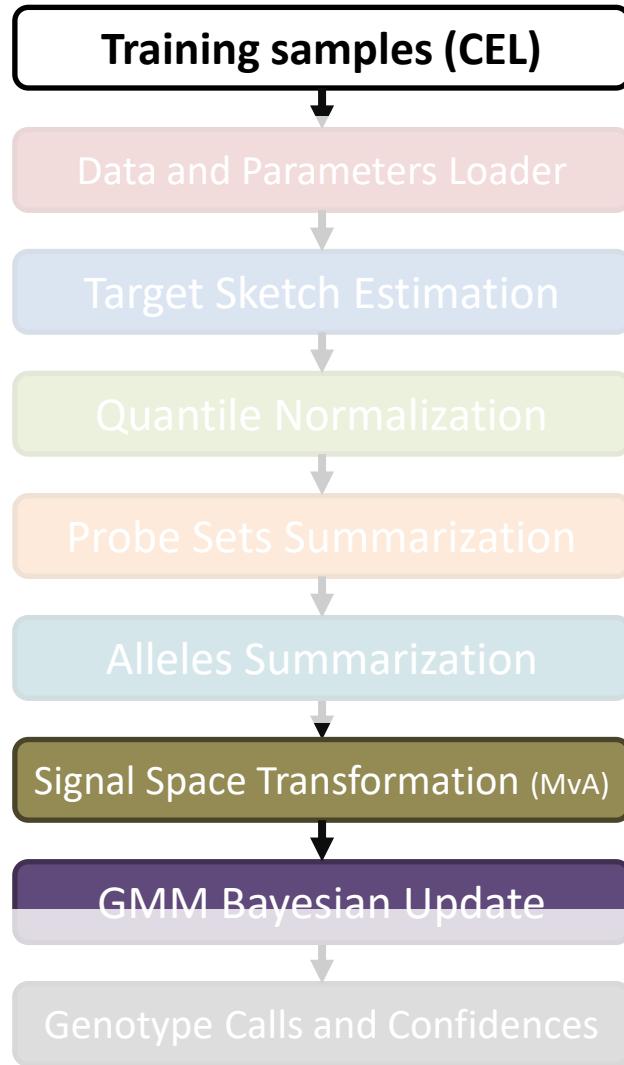
- **Quality Score (Posterior Analog) (for each partition)**
 - **1D (x) Log likelihood under posterior params**
 - $u'_{3x1} = (K_0^{-1}_{3x3} + N'_{3x3})^{-1} * (K_0^{-1}_{3x3} * u_{0,3x1} + N_{3x3} * m_{3x1})$,
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)). u'_{3x1}
 - $\sigma'^2_g = \frac{v_{g0}*\sigma^2_{g0} + \sum_i (x_{gi} - \bar{x}_g)^2 + \frac{k_g*N_g}{k_g+N_g}*(u'_g - u_{g0})^2}{v_{g0} + N_g} \Rightarrow$

$$\frac{v_{g0}*\sigma^2_{g0} + \sum_i x_{gi}^2 - \sum_i x_{gi}*\sum_i x_{gi}*\frac{1}{N_g+0.0001} + \frac{k_g*N_g}{k_g+N_g}*(u'_g - u_{g0})^2}{v_{g0} + N_g}, \quad g = 1, 2, 3$$
 - (sp.comvar) Ad-hoc shrinkage for σ'^2_g of each cluster (controlled by mixing proportion (lambda) (1)).
Adjusted Pooled Variance.
 $w_g = N_g + v_{g0}, \quad g = 1, 2, 3$
 $\sigma'^2_t = \frac{\sum_g w_g * \sigma'^2_g}{\sum_g w_g}, \quad t = 1, 2, 3,$
 $\Rightarrow \frac{(3 - 2 * lambda) * w_t * \sigma'^2_t + \sum_{g \neq t} lambda * w_g * \sigma'^2_g}{(3 - 2 * lambda) * w_t + \sum_{g \neq t} lambda * w_g}$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution

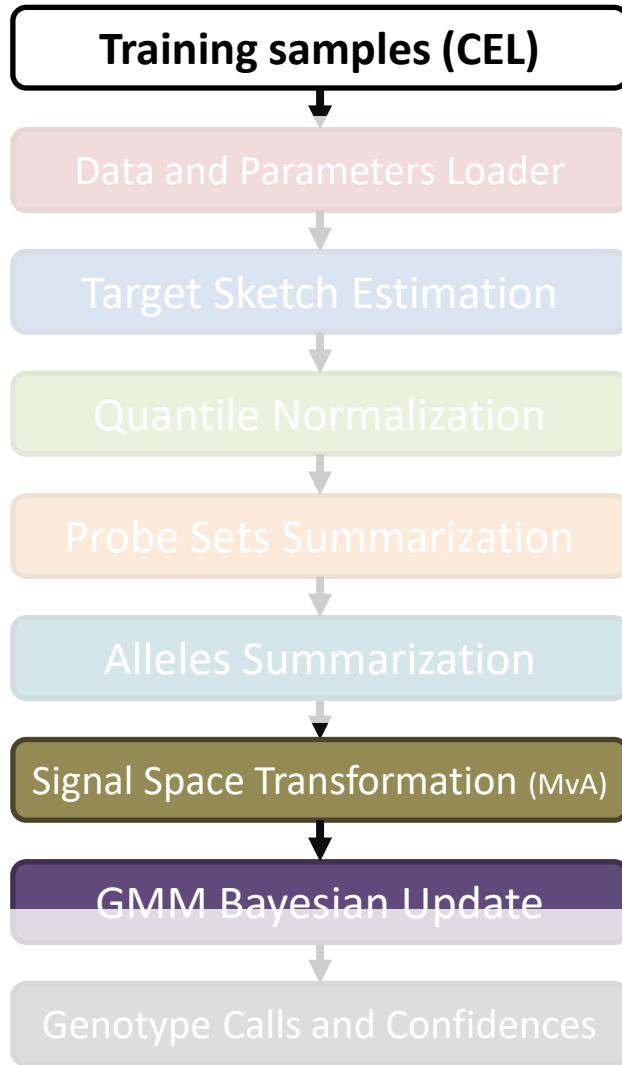


- **Quality Score (Posterior Analog) (for each partition)**
 - **1D (x) Log likelihood under posterior params**
 - $u'_{3 \times 1} = (K_0^{-1}_{3 \times 3} + N'_{3 \times 3})^{-1} * (K_0^{-1}_{3 \times 3} * u_{0 \times 3 \times 1} + N_{3 \times 3} * m_{3 \times 1})$,
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)). $u'_{3 \times 1}$
 - $\sigma'^2_g = \frac{v_0 * \sigma_{g0}^2 + \sum_i (x_{gi} - \bar{x}_g)^2 + \frac{k_g * N_g}{k_g + N_g} * (u'_g - u_{g0})^2}{v_0 + N_g}, \quad g = 1, 2, 3$
 - (sp.comvar) Ad-hoc shrinkage for σ'^2_g of each cluster (controlled by mixing proportion (lambda) (1)).
 - $\ell = \log \prod_g \prod_{i=1}^{N_g} N(u'_g, \sigma'^2_g) = \sum_g \sum_{i=1}^{N_g} \log(N(u'_g, \sigma'^2_g)) \Rightarrow$
$$-\frac{1}{2} \left[\sum_g N_g \log(\sigma'^2_g) + \frac{1}{\sigma'^2_g} \left(\sum_{i=1}^{N_g} x_i^2 - 2u'_g \sum_{i=1}^{N_g} x_i + N_g u'^2_g \right) \right]$$
$$\Rightarrow -2 * \ell$$
$$= \sum_g N_g \log(\sigma'^2_g) + \frac{1}{\sigma'^2_g} \left(\sum_{i=1}^{N_g} x_i^2 - 2u'_g \sum_{i=1}^{N_g} x_i + N_g u'^2_g \right)$$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution

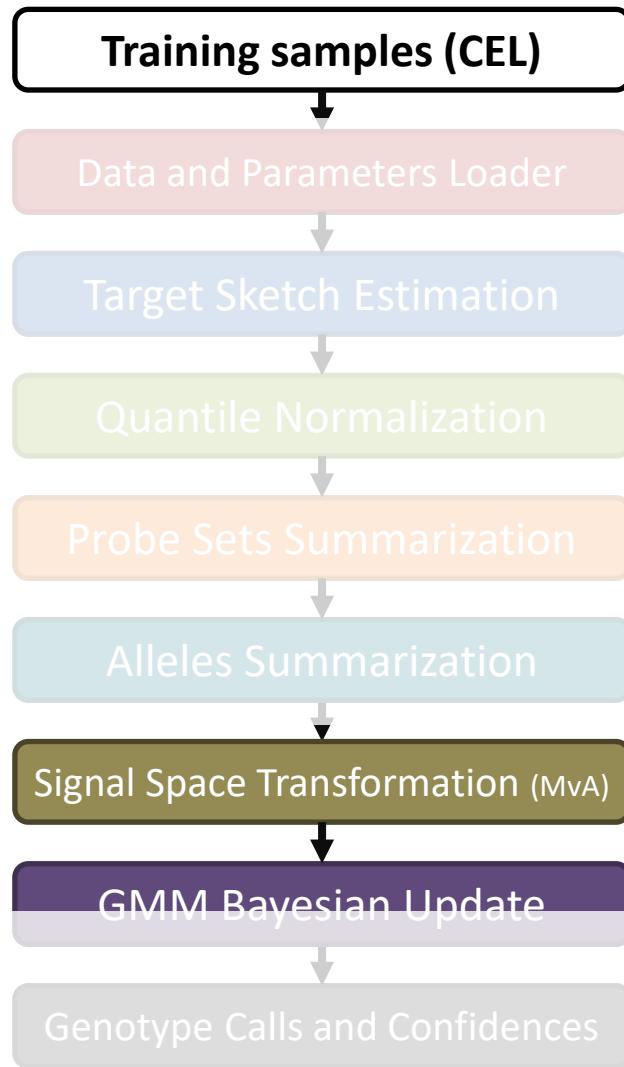


- **Quality Score (Posterior Analog) (for each partition)**
 - **1D (x) Log prior probability of posterior params**
 - $\ell = \log \prod_g N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right) = \sum_g \log\left(N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right)\right) \Rightarrow -\frac{1}{2}\left[\sum_g \log\left(\frac{\sigma_{g0}^2}{k_{g0}}\right) + \frac{k_{g0}}{\sigma_{g0}^2}(u'_g - u_{g0})^2\right]$
 - $\Rightarrow -2 * \ell = \sum_g \log\left(\frac{\sigma_{g0}^2}{k_{g0}}\right) + \frac{k_{g0}}{\sigma_{g0}^2}(u'_g - u_{g0})^2$
 - $\ell = \log \prod_g IG(v_0, \sigma_{g0}^2) \Rightarrow -\ell = \sum_g \frac{\sigma_{g0}^2}{\sigma'^2_g} + (v_0 + 1) * \log(\sigma'^2_g)$
 - $\frac{1}{2} * \text{Quality Score}$
 - **(sp.CSepPen) Geman-McClure transformed FLD penalty for non-well-separated clusters.**
 - $-CSepPen * \sum_{i,j,i \neq j} FLD'_{ij}, i, j \in g = \{1, 2, 3\},$
 - $FLD'_{ij} = \begin{cases} \frac{FLD_{ij}}{1 + \frac{FLD_{ij}}{CSepThr}} * (N_i + N_j), & \text{other} \\ \frac{FLD_{ij}}{1 + \frac{FLD_{ij}}{2 * CSepThr}} * (N_1 + N_3), & i = 1, j = 3 \end{cases},$
 - $FLD_{ij} = FLD_{ji} = \frac{(u'_i - u'_j)^2}{\sigma'^2_i + \sigma'^2_j}, \quad CSepPen = 0.1, \quad CSepThr = 4$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution

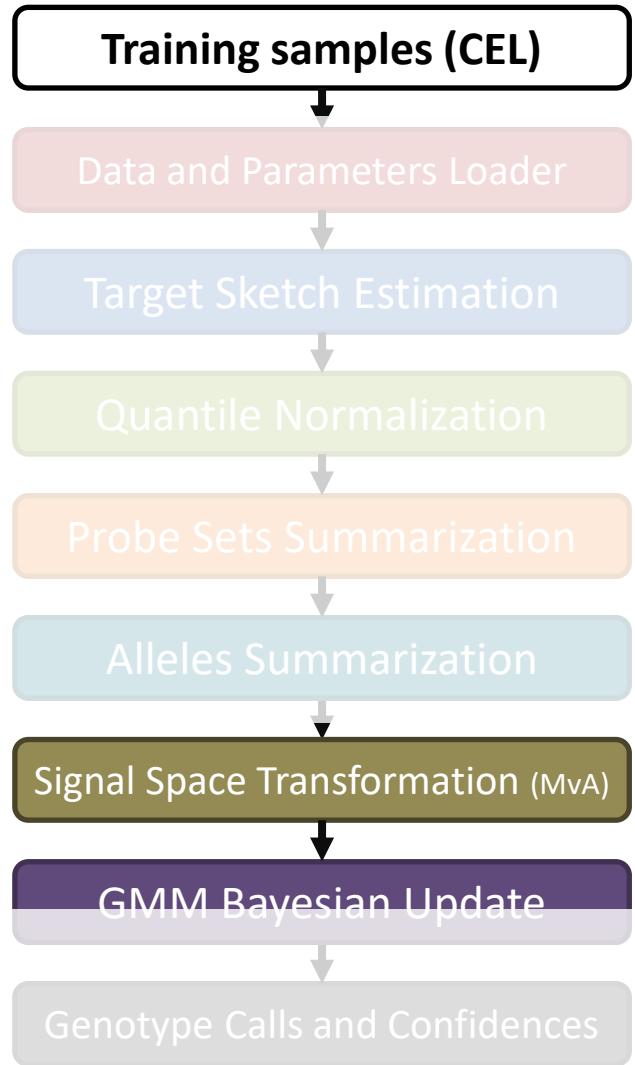


- **Relative Probability for Each Partition under Posterior Information.**
 - Quality Score $Q_{i,j}$ for partition i,j
 - Relative Probability $q_{i,j} = \frac{\exp(-Q_{i,j})}{\exp(-\min Q_{i,j})} = \exp(\min Q_{i,j} - Q_{i,j})$
 $= \frac{\text{Posterior Probability of Specified Partition } (i,j)}{\text{Maximal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

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- Relative Probability for Each Data Point to Be Each Genotype under Posterior Information after dividing $q_{i,j}$

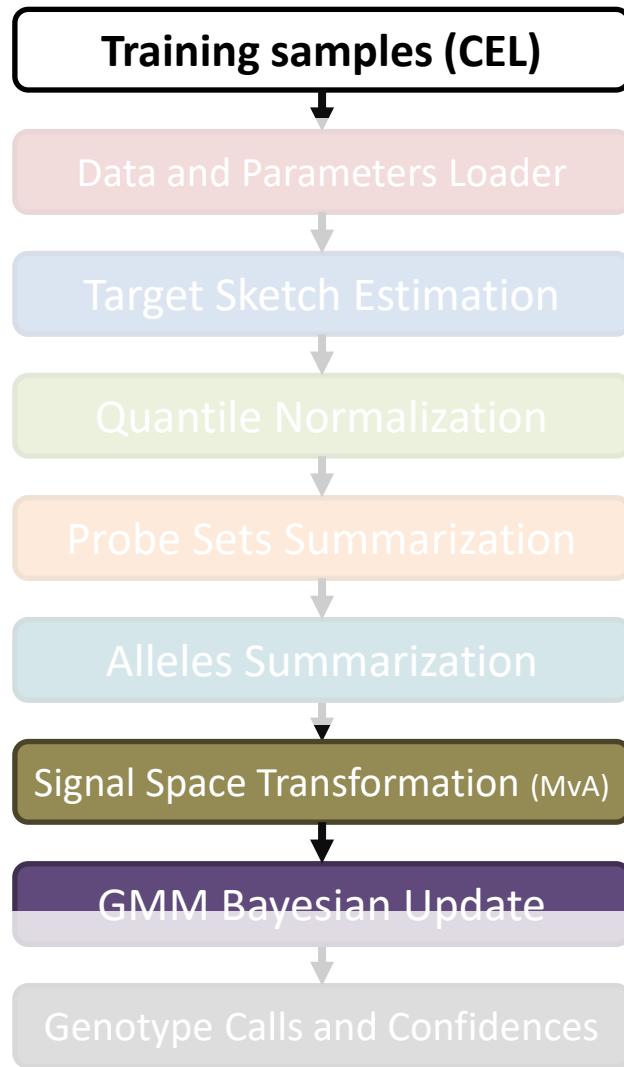
$i \backslash j$	0	1	2	3	4	$q_{i..}$	$\sum_i q_{i..}$	$q_{..} - \sum_i q_{i..}$
0	cccc	bccc	bbcc	bbb	bbb	$q_{0..}$	$\sum_{i=0}^4 q_{i..}$	$\sum_{i=1}^4 q_{i..}$
1		accc	abcc	abbc	abbb	$q_{1..}$	$\sum_{i=0}^1 q_{i..}$	$\sum_{i=2}^4 q_{i..}$
2			aacc	aabc	aabb	$q_{2..}$	$\sum_{i=0}^2 q_{i..}$	$\sum_{i=3}^4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

The relative counts of S^{th} data point being genotype "a" after observing all data.

Let "a" \equiv BB genotype, "b" \equiv AB genotype, "c" \equiv AA genotype.

Centrillion Confidential

Genotyping Analysis Development and Distribution

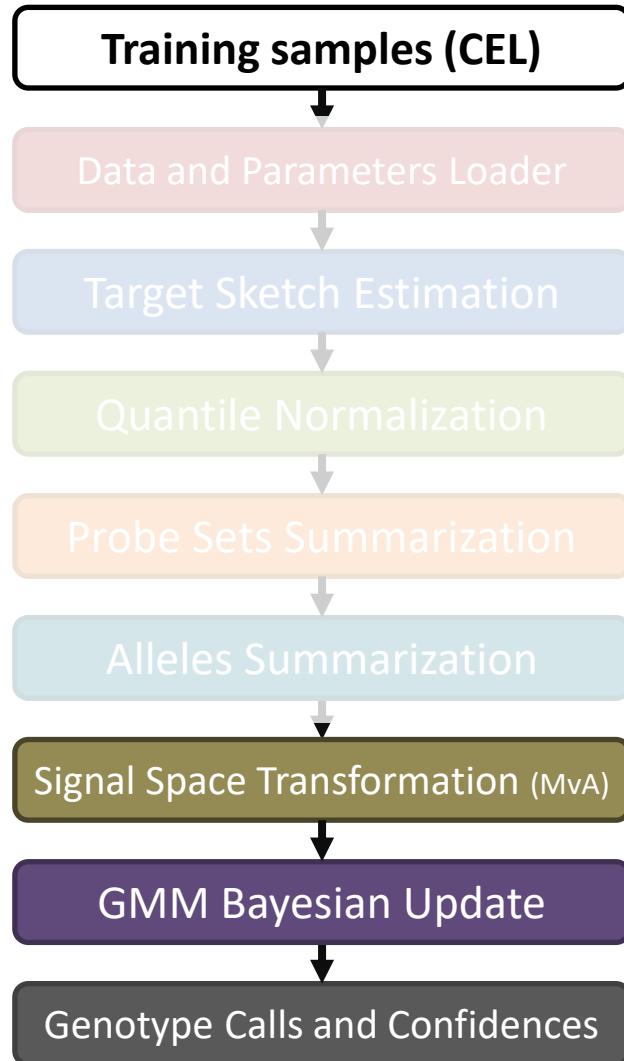


- **Relative Probability for Each Data Point to Be Each Genotype under Posterior Information.**
 - The relative probability for the S_{th} data point being AA genotype: $\frac{\sum_{j=0}^S q_{\cdot j}}{q_{..}}$
 - The relative probability for the S_{th} data point being BB genotype: $1 - \frac{\sum_{i=0}^S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

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- Update 2D Data model parameters with soft assignment of each data bin

$$- \quad u'_{6x1} = (\mathbf{K}_{0\ 6x6}^{-1} + N'_{6x6})^{-1} * (\mathbf{K}_{0\ 6x6}^{-1} * u_{0\ 6x1} + m_{6x1}),$$

$$- \quad \mathbf{K}_{0\ 6x6}^{-1} = \begin{bmatrix} k_{10} & \mathbf{0} & \sigma_{xx120} & \mathbf{0} & \sigma_{xx130} & \mathbf{0} \\ \mathbf{0} & k_{10} & \mathbf{0} & \sigma_{yy120} & \mathbf{0} & \sigma_{yy130} \\ \sigma_{xx120} & \mathbf{0} & k_{20} & \mathbf{0} & \sigma_{xx230} & \mathbf{0} \\ \mathbf{0} & \sigma_{yy120} & \mathbf{0} & k_{20} & \mathbf{0} & \sigma_{yy230} \\ \sigma_{xx130} & \mathbf{0} & \sigma_{xx230} & \mathbf{0} & k_{30} & \mathbf{0} \\ \mathbf{0} & \sigma_{yy130} & \mathbf{0} & \sigma_{yy230} & \mathbf{0} & k_{30} \end{bmatrix},$$

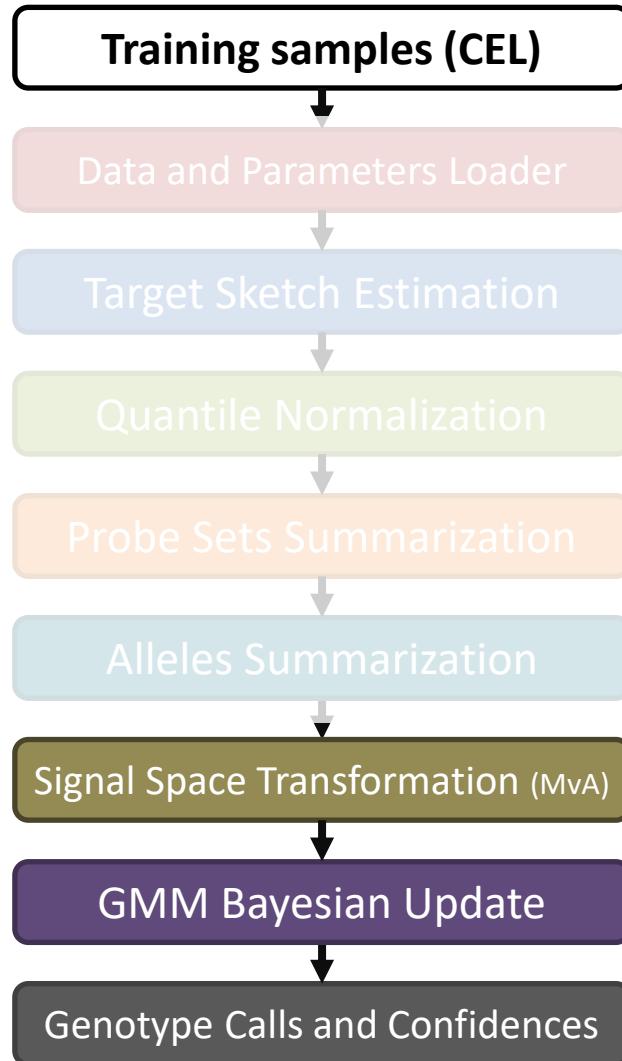
$$- \quad N_{6x6} = \begin{bmatrix} \sum_i p_{1i} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \sum_i p_{1i} & \dots & \vdots \\ \vdots & \sum_i p_{2i} & \dots & \vdots \\ \mathbf{0} & \sum_i p_{2i} & \dots & \sum_i p_{3i} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}$$

$$- \quad u_{0\ 6x1} = \begin{bmatrix} u_{x10} \\ u_{y10} \\ u_{x20} \\ u_{y20} \\ u_{x30} \\ u_{y30} \end{bmatrix}, \quad m_{6x1} = \begin{bmatrix} \sum_i x_{1i} \\ \sum_i y_{1i} \\ \sum_i x_{2i} \\ \sum_i y_{2i} \\ \sum_i x_{3i} \\ \sum_i y_{3i} \end{bmatrix}$$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- Update 2D Data model parameters with soft assignment of each data bin

- $u'_{6x1} = (\mathbf{K}_0^{-1}_{6x6} + \mathbf{N}'_{6x6})^{-1} * (\mathbf{K}_0^{-1}_{6x6} * u_{0\ 6x1} + m_{6x1}),$
- $k'_g = k_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
- (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $u'_{x\ 3x1}$

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

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

$$u'_{x\ 3x1} = \begin{bmatrix} u'_{x1} \\ u'_{x2} \\ u'_{x3} \end{bmatrix}, \quad \begin{array}{l} u'_{x1} = u'_{x1} + \text{delta} - \text{gamma} \\ u'_{x2} = u'_{x2} - \text{gamma} \\ u'_{x3} = u'_{x3} - \text{delta} - \text{gamma} \end{array}$$

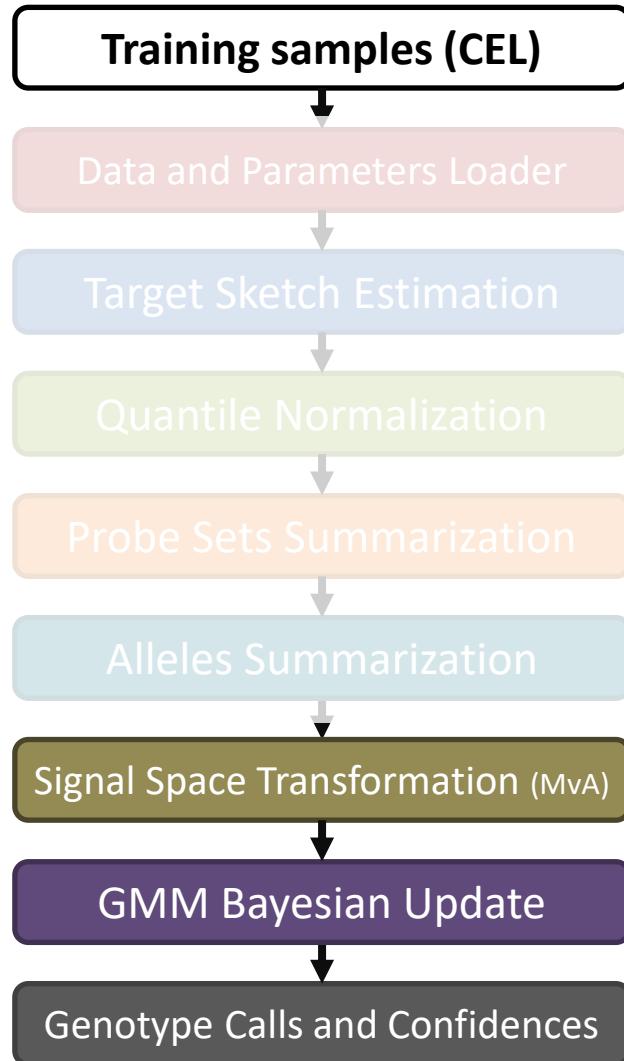
Pool Adjacent-Violators (PAV) algo.

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

$$g = 1, 2, 3$$

$$u'_{3x1} = \begin{bmatrix} u'_{x1} \\ u'_{x2} \\ u'_{x3} \end{bmatrix}, \quad \begin{array}{l} u'_{x1} = u'_{x1} - \text{delta} + \text{gamma} \\ u'_{x2} = u'_{x2} + \text{gamma} \\ u'_{x3} = u'_{x3} + \text{delta} + \text{gamma} \end{array}$$

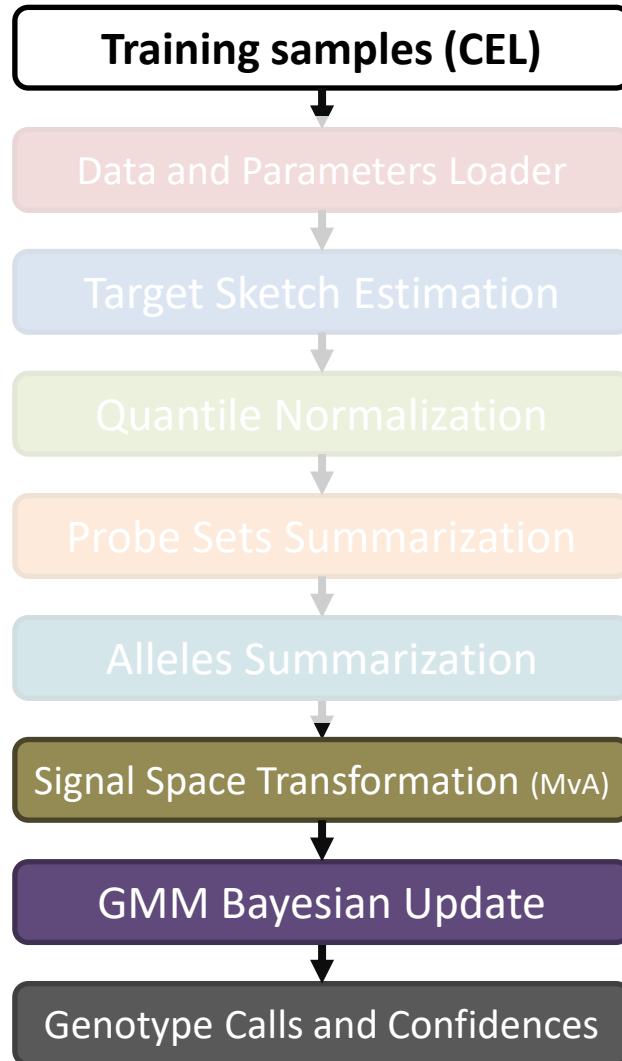
Genotyping Analysis Development and Distribution



- Update 2D Data model parameters with soft assignment of each data bin.
 - $u'_{6x1} = (K_0^{-1}_{6x6} + N'_{6x6})^{-1} * (K_0^{-1}_{6x6} * u_{0,6x1} + m_{6x1})$,
 - $k'_g = k_{g0} + \sum_i p_{gi}, g = 1, 2, 3$
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $u'_{x,3x1}$
 - $v'_g = v_{g0} + \sum_i p_{gi}, g = 1, 2, 3$
 - $\sigma'^2_{xxg} \Rightarrow \frac{v_{g0} * \sigma^2_{xxg0} + (\sum_i p_{gi} x_{gi}^2 - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} x_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{xg} - u_{xg0})^2}{v'_g}$
 - $\sigma'^2_{yyg} \Rightarrow \frac{v_{g0} * \sigma^2_{yyg0} + (\sum_i p_{gi} y_{gi}^2 - \sum_i p_{gi} y_{gi} * \sum_i p_{gi} y_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{yg} - u_{yg0})^2}{v'_g}$
 - $\sigma'^2_{xyg} \Rightarrow \frac{v_{g0} * \sigma^2_{xyg0} + (\sum_i p_{gi} x_{gi} y_{gi} - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} y_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{yg} - u_{yg0}) * (u'_{xg} - u_{xg0})}{v'_g}$
 - Ad-hoc shrinkage for $\sigma'^2_{..g}$ of each cluster (controlled by mixing proportion (lambda) (1)).
 - Adjusted Pooled Variance.
 - $w_g = v_{g0} + \sum_i p_{gi}, g = 1, 2, 3$
 - $\sigma'^2_{xxt} = \frac{\sum_g w_g * \sigma'^2_{xxg}}{\sum_g w_g}, \sigma'^2_{yyt} = \frac{\sum_g w_g * \sigma'^2_{yyg}}{\sum_g w_g}, t = 1, 2, 3,$
 - $\Rightarrow \frac{(3 - 2 * lambda) * w_t * \sigma'^2_t + \sum_{g \neq t} lambda * w_g * \sigma'^2_g}{(3 - 2 * lambda) * w_t + \sum_{g \neq t} lambda * w_g}$
 - $\sigma'^2_{xyt} = (\sigma'^2_{xxt} * \sigma'^2_{yyt}) * \frac{\sigma'^2_{xyg}}{\sigma'^2_{xxg} * \sigma'^2_{yyg}}, t = 1, 2, 3, g = t, \text{ means before shrinkage adjustment.}$

Centrillion Confidential

Genotyping Analysis Development and Distribution

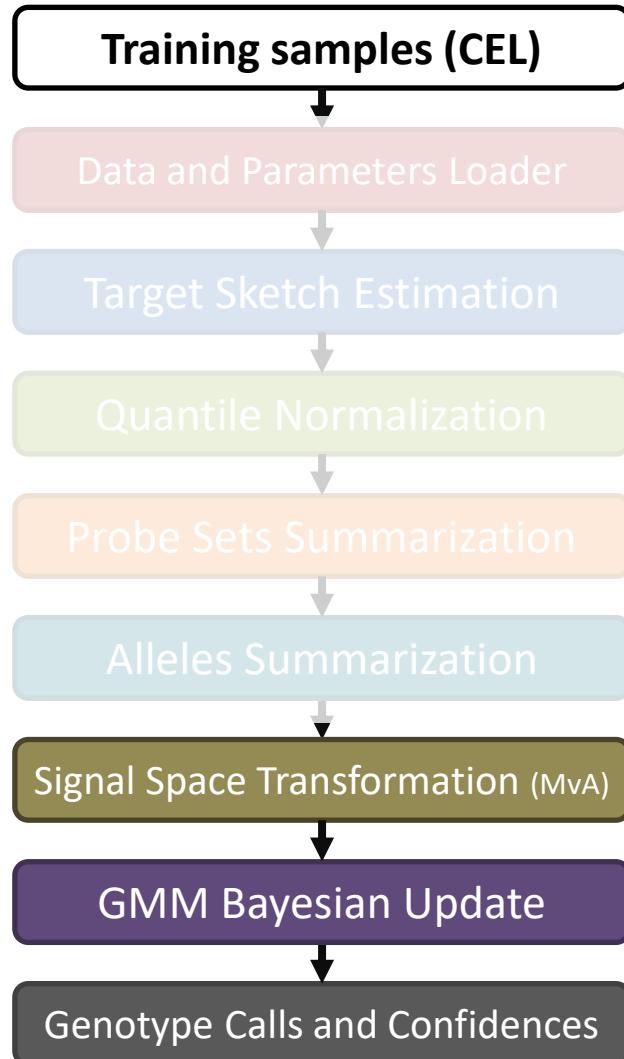


- Update 2D Data model parameters with soft assignment of each data bin
 - $\mathbf{u}'_{6x1} = (\mathbf{K}_{0\ 6x6}^{-1} + \mathbf{N}'_{6x6})^{-1} * (\mathbf{K}_{0\ 6x6}^{-1} * \mathbf{u}_{0\ 6x1} + \mathbf{m}_{6x1}),$
 - $k'_g = k_{g0} + \sum_i p_{gi}, \ g = 1, 2, 3$
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). \mathbf{u}'_{3x1}
 - $v'_g = v_{g0} + \sum_i p_{gi}, \ g = 1, 2, 3$
 - $\sigma'^2_{xxg}, \ \sigma'^2_{yyg}, \ \sigma'^2_{xyg}, \ g = 1, 2, 3$
 - Ad-hoc shrinkage for $\sigma'^2_{..g}$ of each cluster (controlled by mixing proportion (lambda) (1)).
 - $\sigma'_{xx12} = \sigma_{xx120}, \ \sigma'_{xx13} = \sigma_{xx130}, \ \sigma'_{xx23} = \sigma_{xx230}$
 $\sigma'_{yy12} = \sigma_{yy120}, \ \sigma'_{yy13} = \sigma_{yy130}, \ \sigma'_{yy23} = \sigma_{yy230}$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution

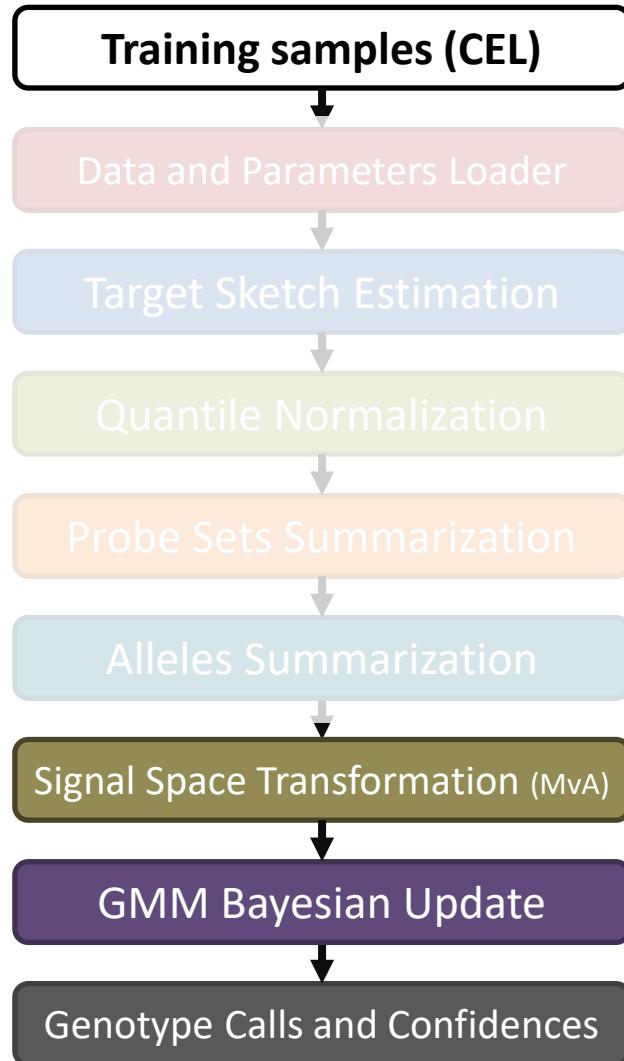


- (sp.mix, freqflag) compute the frequency of each cluster (AA, AB, BB)
 - $f_t = \frac{k'_t}{\sum_g k'_g}, t = 1, 2, 3$
 $\Rightarrow \log f_t = \log k'_t - \log(\sum_g k'_g)$
 $\Rightarrow -\log f_t = -\log k'_t + \log(\sum_g k'_g)$
- For each point, compute the probability that a data point X (x, y) belongs to each genotype.
 - $p(X \in t | X) = \frac{p(X \in t, X)}{p(X)} = \frac{p(X \in t)p(X|X \in t)}{ocean + \sum_g p(X \in g)p(X|X \in g)} =$
$$\frac{f_t \cdot BVN(X | \mathbf{u}'_t, \left(1 + \frac{\text{inflatePRA}}{k'_t}\right) \cdot \boldsymbol{\sigma}'_t)}{ocean + \sum_g f_g \cdot BVN(X | \mathbf{u}'_g, \left(1 + \frac{\text{inflatePRA}}{k'_g}\right) \cdot \boldsymbol{\sigma}'_g)}$$
 $\text{inflatePRA} = 0 \text{ (default)}, \ ocean = 0.00001 \text{ (default)}$
 - $\log p(X \in t)p(X|X \in t) = \log(p(X \in t)) + \log(p(X|X \in t))$
 - If `copynumber=1`, $p(X \in AB)p(X|X \in AB) = 0$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- (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 probability that a data point X (x, y) belongs to each genotype.
 - $p(X \in t|X) = \frac{p(X \in t, X)}{p(X)} = \frac{p(X \in t)p(X|X \in t)}{\text{ocean} + \sum_g p(X \in g)p(X|X \in g)}, t = 1, 2, 3$
- For each point, make a call.
 - $\hat{t} = \operatorname{argmax}_t p(X \in t|X)$
 - $confidence = 1 - p(X \in \hat{t}|X)$
 - No call: If $confidence > MS$,
 $MS = 0.15$ (default) @ getGTypeCall()

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



- 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 \(github.com\)](https://Project-CPT/CPT.wiki/AxiomGT.md at main · jeff665547/Project-CPT (github.com))

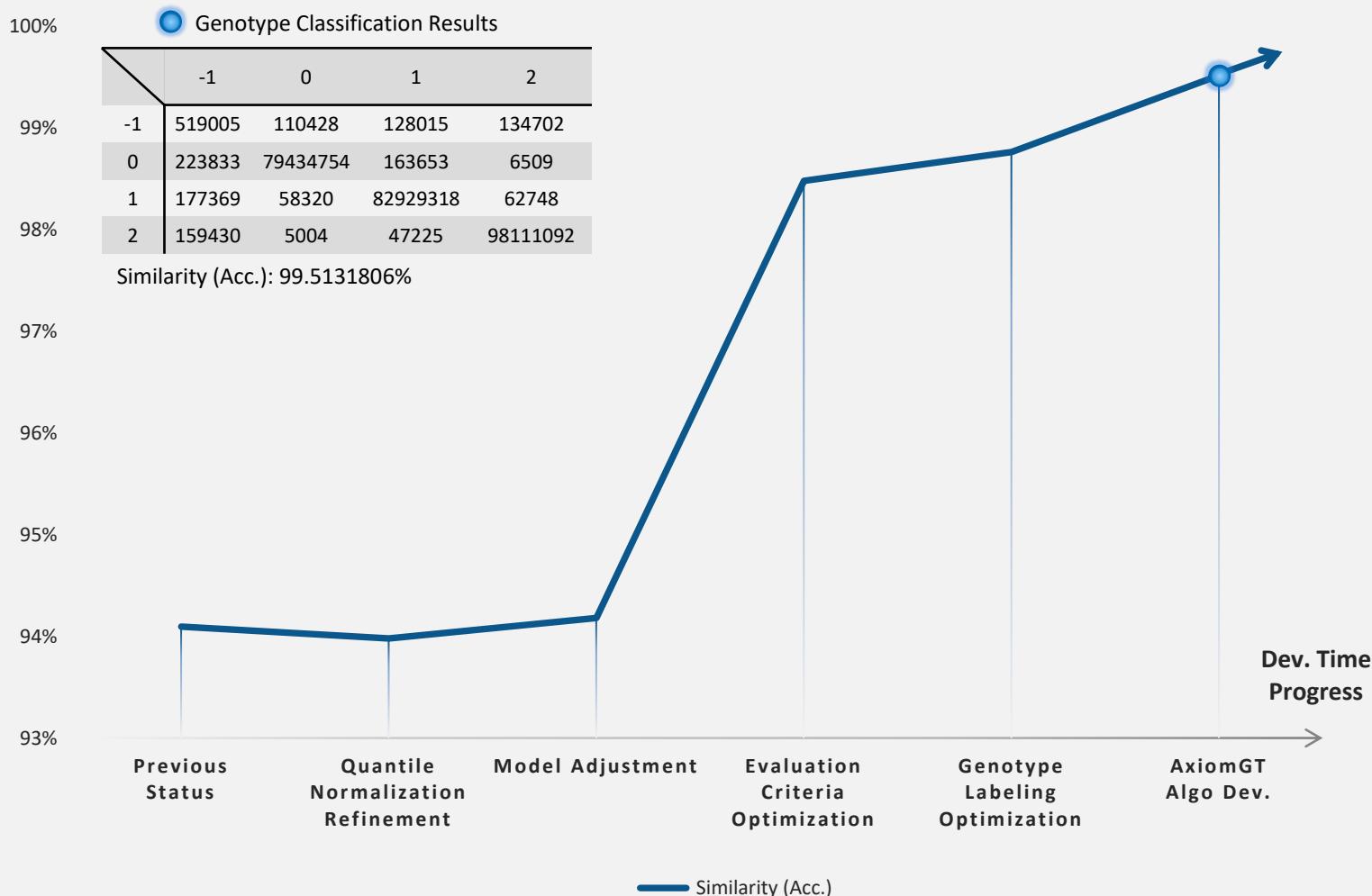
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Analysis Development and Distribution



GSE78098 TESTING DATA PERFORMANCE



Genotyping Analysis Development and Distribution



- Bugfix for the CI/CD error when deployment and distribution.

Status	Pipeline	Triggerer	Stages	
X failed ⌚ 00:17:50 🕒 18 hours ago	Bugfix for gender inputs, and remove redundant code. #4901 ↗ hunterize -O 476c692c 🐛 latest		X	C ⋮
X failed ⌚ 00:18:00 🕒 2 days ago	Fix the I/C #4900 ↗		X	C ↑ ↓
X failed ⌚ 00:18:05 🕒 2 days ago	Update th... #4899 ↗			C ⋮
✓ passed ⌚ 01:26:46 🕒 1 week ago	Merge br... #4898 ↗			C ⋮
✓ passed ⌚ 01:11:00	Merge br... #4897 ↗			C ⋮
All copyright reserved	<pre>62 -- Detecting CXX compile features 63 -- Detecting CXX compile features - done 64 -- Detecting Fortran compiler ABI info 65 -- Detecting Fortran compiler ABI info - done 66 -- Check for working Fortran compiler: C:/Program Files/mingw-w64/x86_64-7.3.0-posix-seh-rt_v5-rev0/mingw64/bin/gfortran.exe - skipped 67 [hunter ** FATAL ERROR **] ABI not detected for C compiler 68 [hunter ** FATAL ERROR **] [Directory:C:/GitLab-Runner/builds/55f15aeb/0/centrillion/CPT] 69 ----- 70 ----- ERROR ----- 71 ----- 72 CMake Error at C:/./hunter/_Base/Download/Hunter/unknown/15219b8/Unpacked/cmake/modules/hunter_error_page.cmake:12 (message): 73 Call Stack (most recent call first): 74 C:/./Base/Download/Hunter/unknown/15219b8/Unpacked/cmake/modules/hunter_fatal_error.cmake:20 (hunter_error_page) 75 C:/./hunter/_Base/Download/Hunter/unknown/15219b8/Unpacked/cmake/modules/hunter_finalize.cmake:50 (hunter_fatal_error) 76 C:/./hunter/_Base/Download/Hunter/unknown/15219b8/Unpacked/cmake/modules/hunter_add_package.cmake:23 (hunter_finalize) 77 CMakeLists.txt:20 (hunter_add_package) 78 -- Configuring incomplete, errors occurred! 79 See also "C:/GitLab-Runner/builds/55f15aeb/0/centrillion/CPT/build/CMakeFiles/CMakeOutput.log". 80 See also "C:/GitLab-Runner/builds/55f15aeb/0/centrillion/CPT/build/CMakeFiles/CMakeError.log". 82 ERROR: Job failed: exit status 1</pre>			

Win_CI

New issue

Duration: 31 seconds

Finished: 19 hours ago

Timeout: 3h (from project)

Runner: #16 (55f15aeb) windows10 runner

Tags: WIN

Commit 476c692c ↗

Bugfix for gender inputs, and remove redundant code.

→ X Win_CI

CentOS_CI

Genotyping Analysis Development and Distribution



- Successful deployment and distribution.

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

Centrillion Confidential

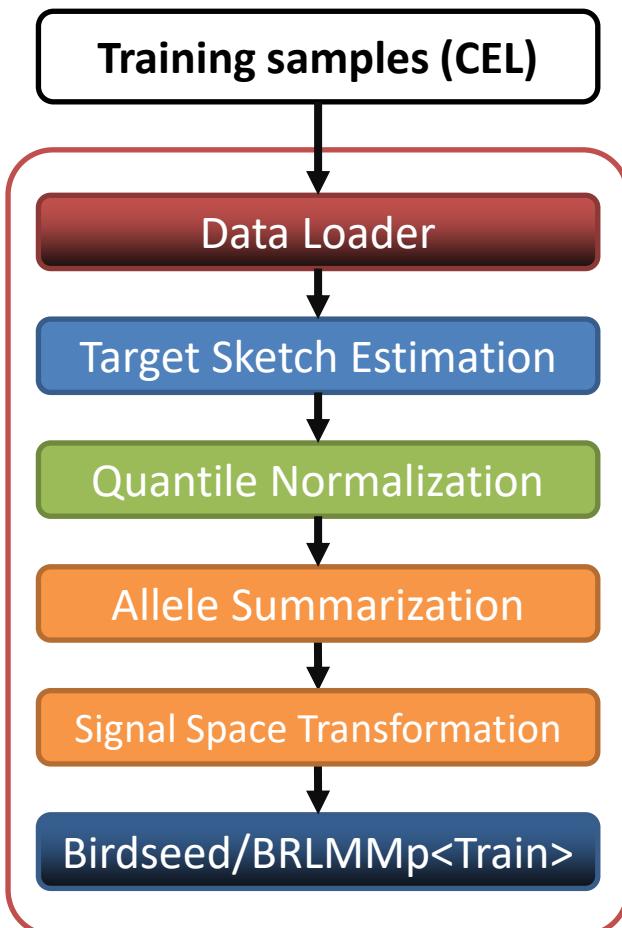
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.



Other Genotyping Models Research and Development

Jeff (CHI-HSUAN HO)

- Birdseed Framework



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 Wysoker¹, James Nemesh¹, Simon Cawley⁷, Earl Hubbell⁷, Jim Veitch⁷, Patrick J Collins⁷, Katayoon Darvishi⁸, Charles Lee⁸, Marcia M Nizzari¹, Stacey B Gabriel¹, Shaun Purcell^{1,5}, Mark J Daly^{1,5,9} & David Altshuler^{1,4,5,9}

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Methods Evaluation and Simulation



- Birdseed: K-means training model

Model₀: Gaussian Mixture Model – GMM with K-Means centroid (Existing Model)

Model₁: Non – probabilistic Model

$$BIC = \frac{1}{\sigma^2} \sum_{j=1}^K \sum_{i=1}^{N_j} \min \|X_i - \hat{\mu}_j\|^2 + Kd \cdot \ln(N), \quad \hat{\mu}_j = \frac{\sum_{i=1}^{N_j} X_i}{N_j}$$

Model₂: $f(x_i) = \pi_{ij} N(\mu_j, \sigma^2 \cdot I_d)$

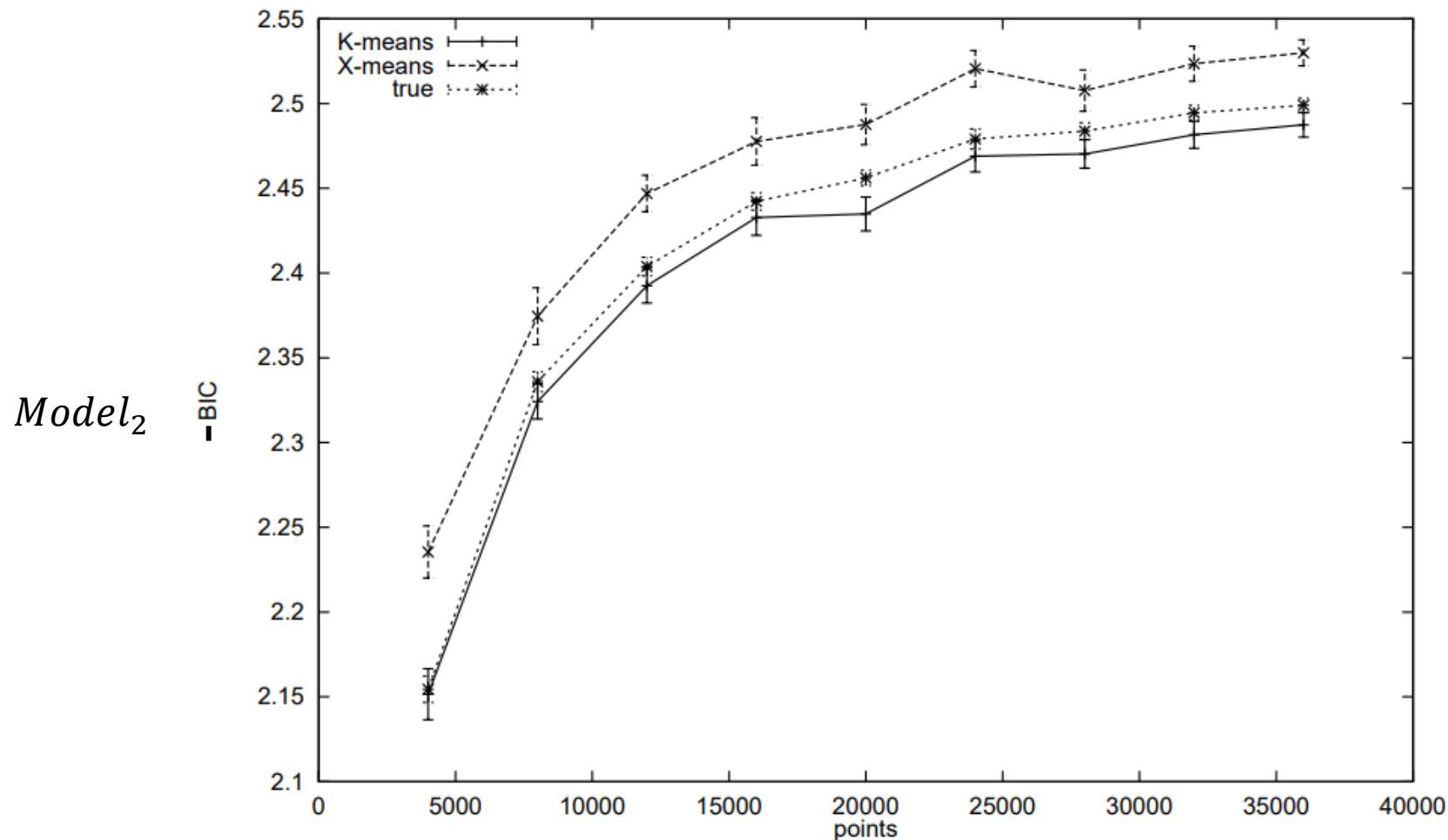
$$BIC = -2 \sum_{j=1}^K N_j \ln(N_j) + 2N \ln(N) + dN \ln(2\pi\hat{\sigma}^2) + dN + \ln(N) \cdot K(d+1),$$

$$\text{where } \hat{\sigma}^2 = \frac{1}{dN} \sum_{j=1}^K \sum_{i=1}^{N_j} \|X_i - \hat{\mu}_j\|^2, \quad \hat{\mu}_j = \frac{\sum_{i=1}^{N_j} X_i}{N_j}$$

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Methods Evaluation and Simulation



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Methods Evaluation and Simulation



- Birdseed: K-means training model

Model₃: ANOVA: $X_{it} = \mu_{it} + E_{it}$, $\boldsymbol{\mu} \in \mathbb{R}^{nxd}$, $\mathbf{E} \in \mathbb{R}^{nxd}$, $E_{it} \sim iid N(0, \sigma^2)$, $i = 1, 2, \dots, N$, $t = 1, 2, \dots, d$

$$BIC = Nd \cdot \ln \left(\sum_{j=1}^K \sum_{i=1}^{N_j} \|X_i - \hat{\boldsymbol{\mu}}_j\|^2 \right) + Nd \left(1 + \ln \left(\frac{2\pi}{Nd} \right) \right) \\ + \ln(Nd) \cdot \left[Kd + \frac{1}{\tilde{\sigma}} \sum_{j=1}^{K'} \sum_{i=1}^{\widetilde{N}_j} \sum_{t=1}^d \sum_{l \neq c(i)} \phi \left(\frac{X_{i,t} + \delta_l^{i,t} - \tilde{\mu}_{i,t}}{\tilde{\sigma}} \right) \cdot \lim_{\gamma \rightarrow \delta_l^{i,t}} \mathcal{M}(\mathbf{X} + \gamma \mathbf{e}_{i,t})_{i,t} \right],$$

where $\tilde{\sigma}^2 = \frac{1}{dN} \sum_{j=1}^{K'} \sum_{i=1}^{\widetilde{N}_j} \|X_i - \tilde{\boldsymbol{\mu}}_i\|^2$, $\tilde{\boldsymbol{\mu}} = \mathcal{M}(\mathbf{X}; K')$ for some $K' > K$,
 $\phi(\cdot)$ is the pdf of Normal (Gaussian) distribution,

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

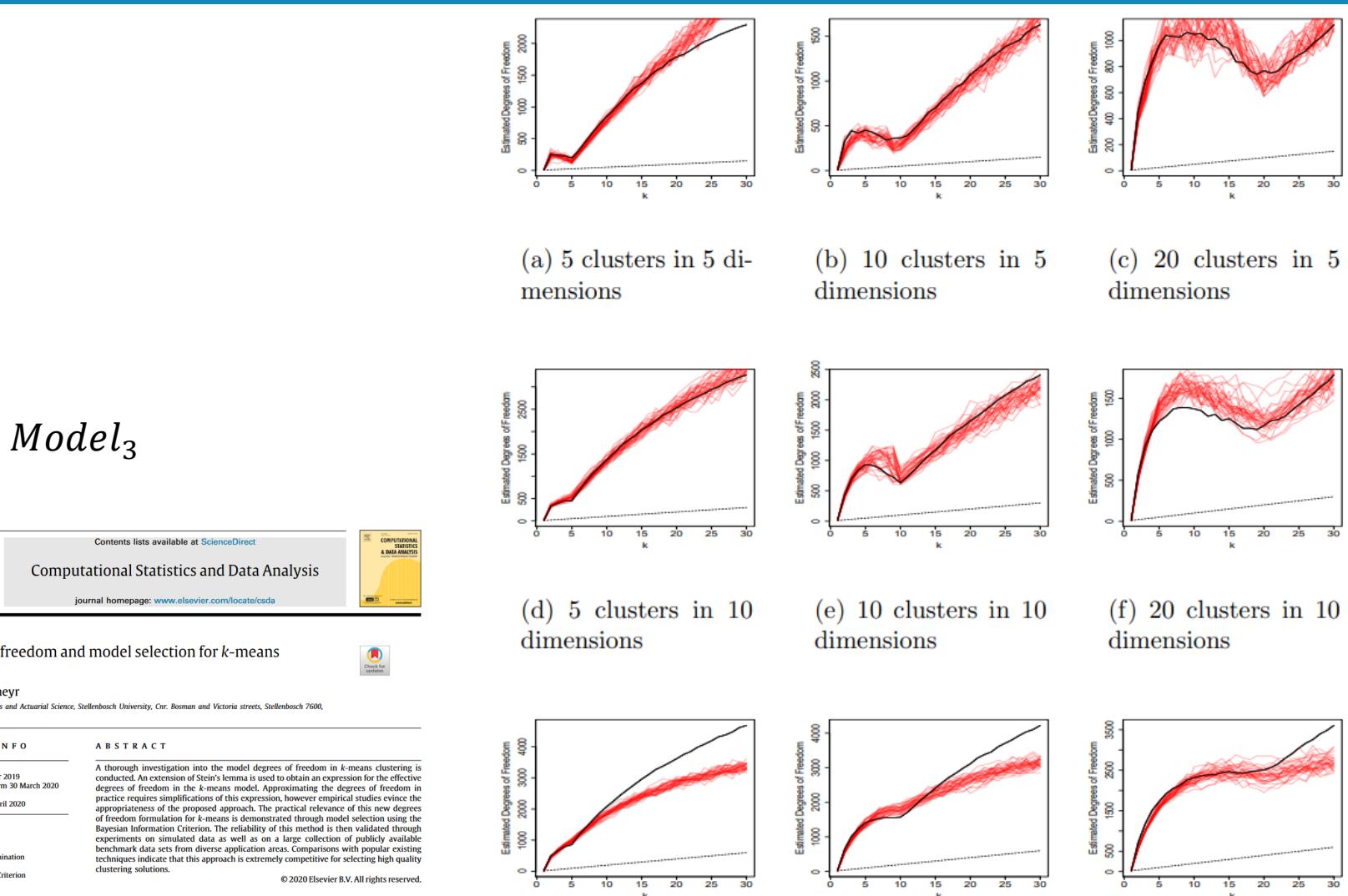
$$c(i) = argmin_{l \in \{1, 2, \dots, K\}} \|X_i - \hat{\boldsymbol{\mu}}_l\|^2,$$

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

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Methods Evaluation and Simulation



Model₃



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 2009
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



© 2020 Elsevier B.V. All rights reserved.

(d) 5 clusters in 10 dimensions

(e) 10 clusters in 10 dimensions

(f) 20 clusters in 10 dimensions

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Genotyping Methods Evaluation and Simulation



- Birdseed: K-means training model

Model₀: Equal weight Before scale: 94.1908%

Model₀: But differnet weight Before scale

– Default (trim BIC, trim Fan): 84.6841% with very highly lose – classified rate.

Model₀: But differnet weight Before scale – no trim BIC: 90.8778%

Model₀: But differnet weight Before scale – no trim Fan: 0%

Model₁: Before scale

– Default (trim BIC, trim Fan): 98.3863% with very highly lose – classified rate.

Model₁: Before scale – no trim BIC: 98.6603% → 98.7585% (BIC under all data)

Model₁: Before scale – no trim Fan: 98.4763%

Model₁: 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					

Centrillion Confidential

Genotyping Methods Evaluation and Simulation



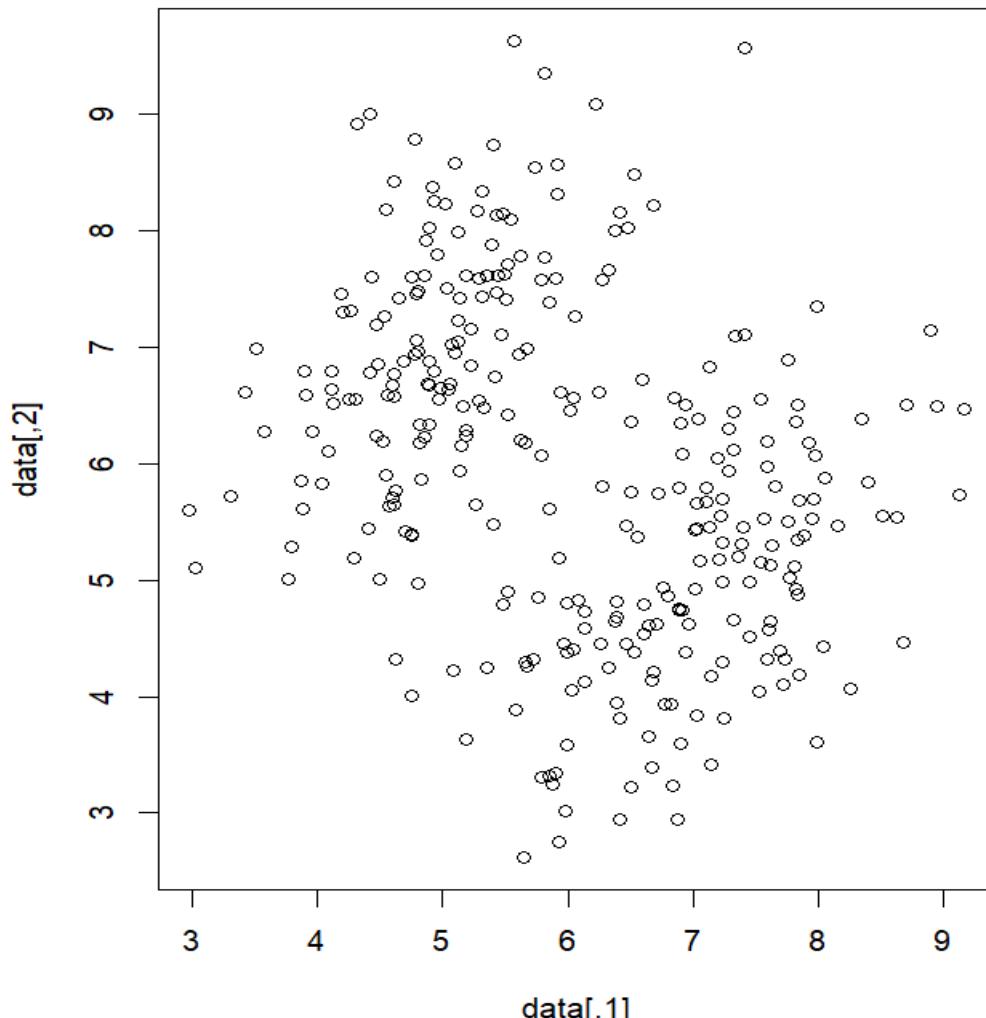
- Simulation

`sol$k:` *Model₃*

`sol$k_:` *Model₁: Before scale*

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

Ans: $k = 2$



Genotyping Methods Evaluation and Simulation



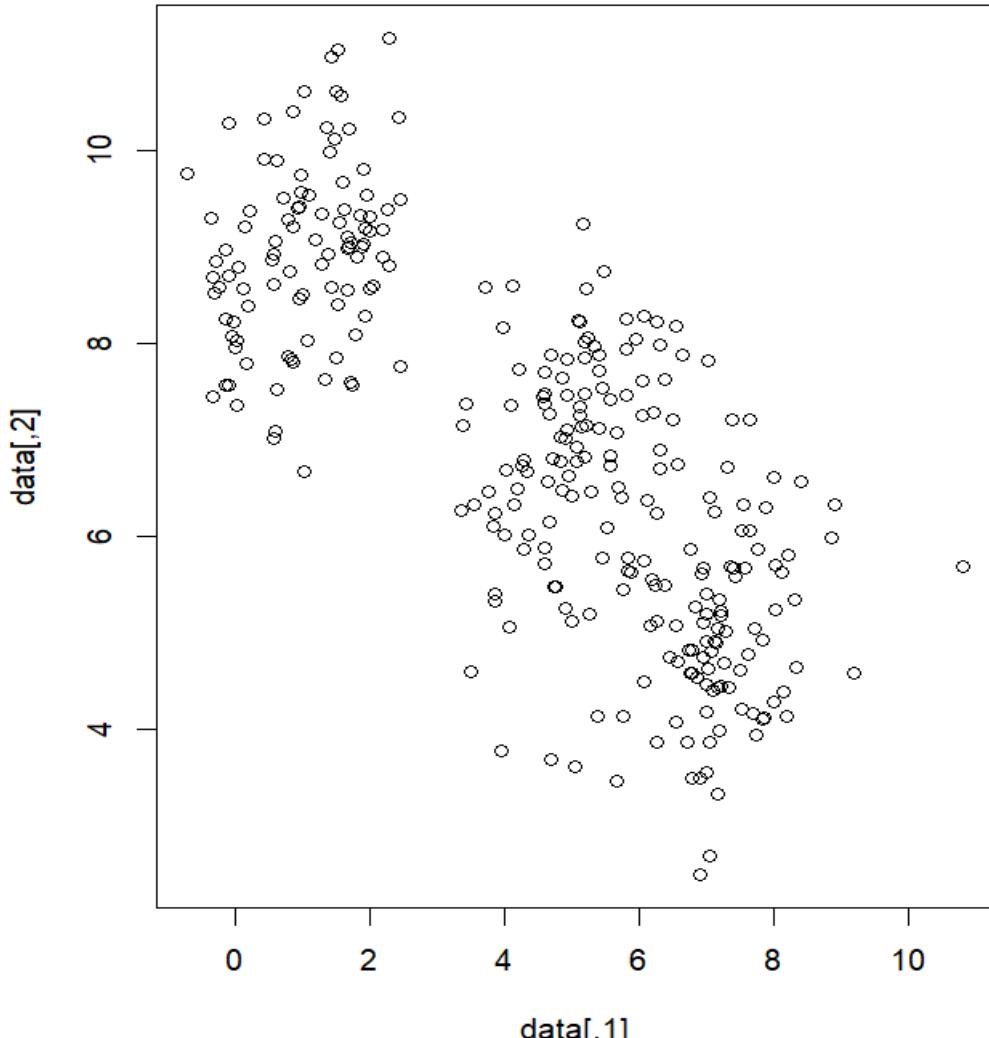
- Simulation

`sol$k:` *Model₃*

`sol$k_:` *Model₁: Before scale*

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

Ans: $k = 3$





Intensity Extraction and Data collecting Software – Summit Grid Review, Improve & Evaluate

Jeff (CHI-HSUAN HO)

- **Summit.Grid**

- Chip Spec Parameters Setup

```
{
    "name": "banff",
    "w": 2480,
    "h": 2480,
    "w_cl": 496,
    "h_cl": 496,
    "cell_w_um": 4,
    "cell_h_um": 4,
    "space_um": 1,
    "location_marker": {
        "template": "resource/banff/pat_white.tif",
        "mask" : "resource/banff/pat_white_mask.tif",
        "w": 60,
        "h": 60
    },
    "shooting_marker": {
        "origin_desc": "center of the chip",
        "type": "regular_matrix",
        "mk_pats": [
            {
                "filter": 0,
                "w_um": 60,
                "h_um": 60,
                "path": "resource/banff/pat_white.tif",
                "mask": "resource/banff/pat_white_mask.tif" // module parameters
            },
            {
                "um2px_r": 2.68,
                "path": "resource/banff/pat_2_68.tif"
            },
            {
                "um2px_r": 2.41,
                "path": "resource/banff/pat_2_41.tif"
            }
        ],
        "view": {
            "offset": [ 0, 0 ],
            "layout": [ 1, 1 ],
            "stride": [ 810, 810 ]
        }
    },
    "chip": "banff33"
}
```

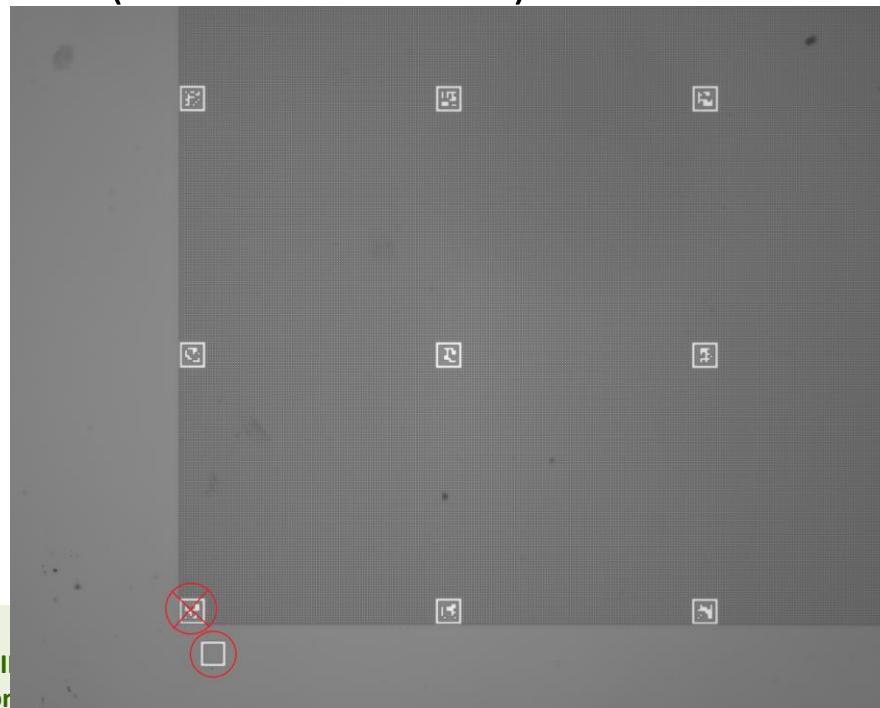
```
{
    "name": "banff33",
    "spec": "banff",
    "fov": {
        "mask_src_type": "au",
        "mask": "resource/ba",
        "x_i" : -810,
        "y_i" : -810,
        "rows": 3,
        "cols": 3,
        "w_d" : 810,
        "h_d" : 810,
        "seq": [
            [[0, 0]], [[1, 0]], [[2, 0]],
            [[2, 1]], [[1, 1]], [[0, 1]],
            [[0, 2]], [[1, 2]], [[2, 2]]
        ],
        "gain": 0,
        "exposure_time_abs": 250000,
        "camera_delay_time": 1,
        "filter": 2,
        "marker_type": "AM3"
    },
    "origin_infer": {
        "algo": "aruco_detection",
        "pyramid_level": 3,
        "nms_count": 9,
        "cell_size_px": 5,
        "loc_mk_layout_pt": [3,3]
    },
    "af": {
        "search_range" : 120,
        "step_size" : 30,
        "margin" : 0,
        "kernel_scale" : 0.0,
        "length_scale" : 45.0,
        "noise_level" : 1e-3,
        "maxiter" : 10000
    },
    "qaqc": [
        "Merge branch 'master' into LAB",
        "Merge branch 'master' into FAB",
        "[Feature] Update the scanning parameter for S2B",
        "QAQC Shasta chip (S2B) support.",
        "Merge branch 'master' into LAB",
        "Merge branch 'master' into FAB",
        "Support for scanning and gridding Shasta (S2B)",
        "QAQC Jasper chip (J1C, J2C) support.",
        "Merge branch 'master' into LAB",
        "Merge branch 'master' into FAB",
        "Support for scanning and gridding Jasper (J1C)",
        "QAQC Jasper chip (J1C) support.",
        "Merge branch 'master' into FAB"
    ]
}
```

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

- Summit.Grid
 - Bugfix for wrong nms_count (WARNING for S1C)
 - Original
 - Noise influence (WARNING for Y2B)

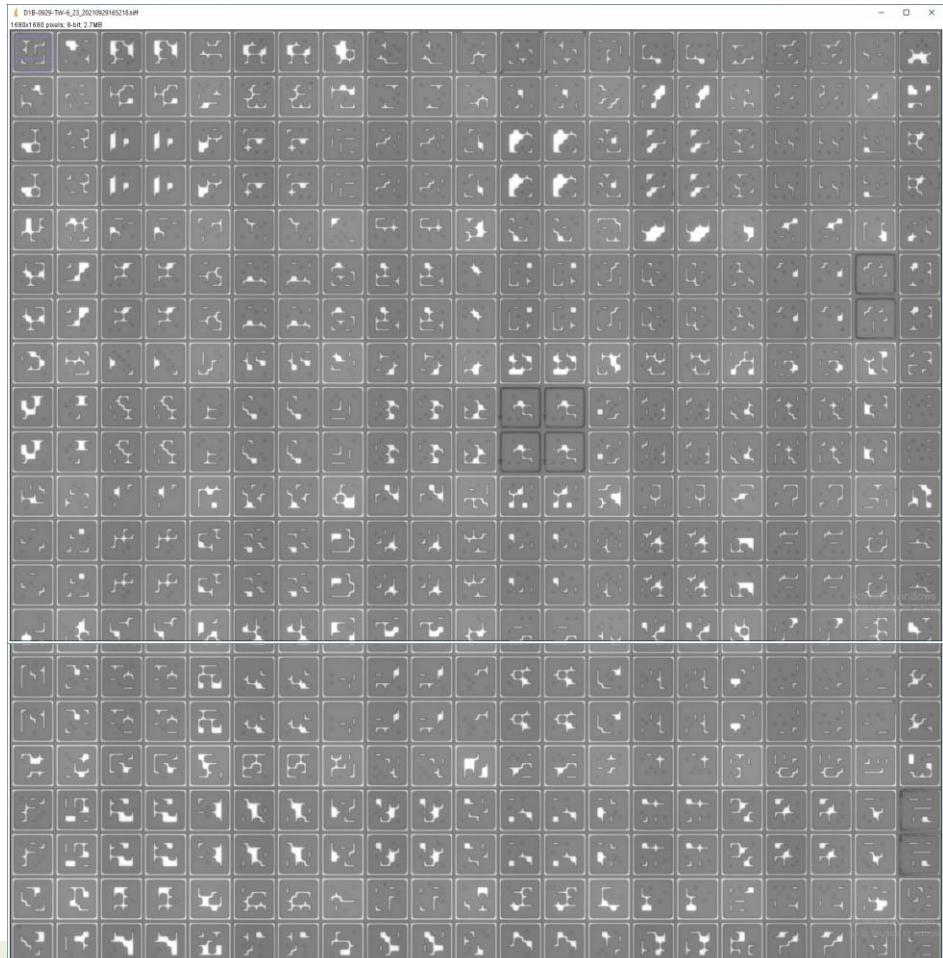
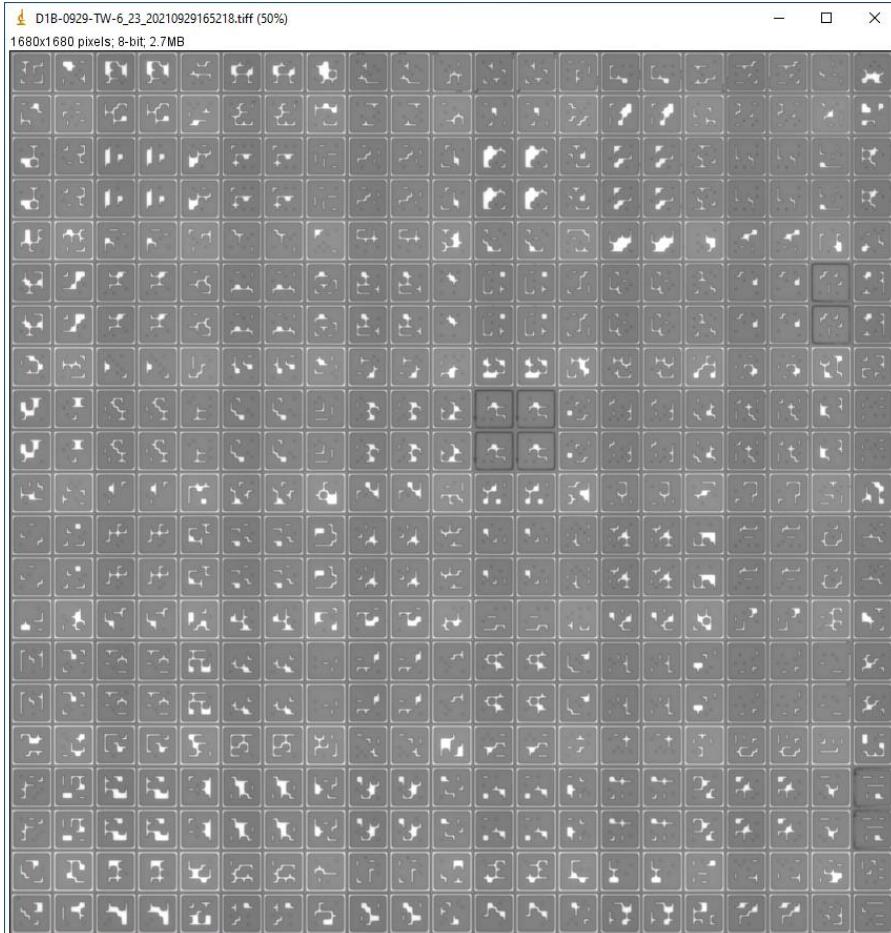
```
// detection parameters
nms_count_ = (fov_wd_ / mk_wd_cl_ + 1) * (fov_hd_ / mk_hd_cl_ + 1);    Alex, 2 years ago • support new aruco recognition ...
nms_radius_ = aruco_marker_->at("nms_radius");
```



Gridding Development



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



Centrillion Confidential

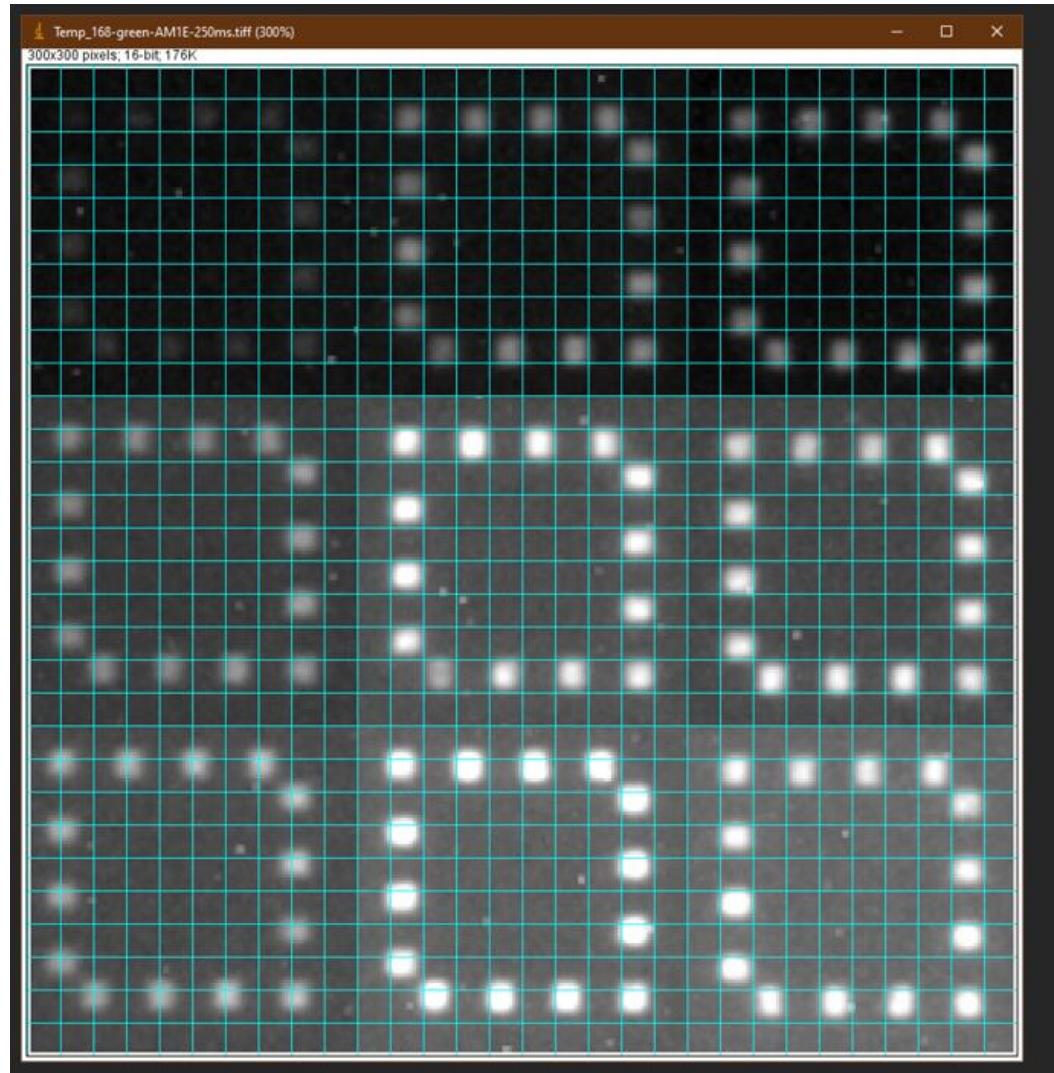
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Gridding Development



- Summit.Grid
 - PGD Images Processing.

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



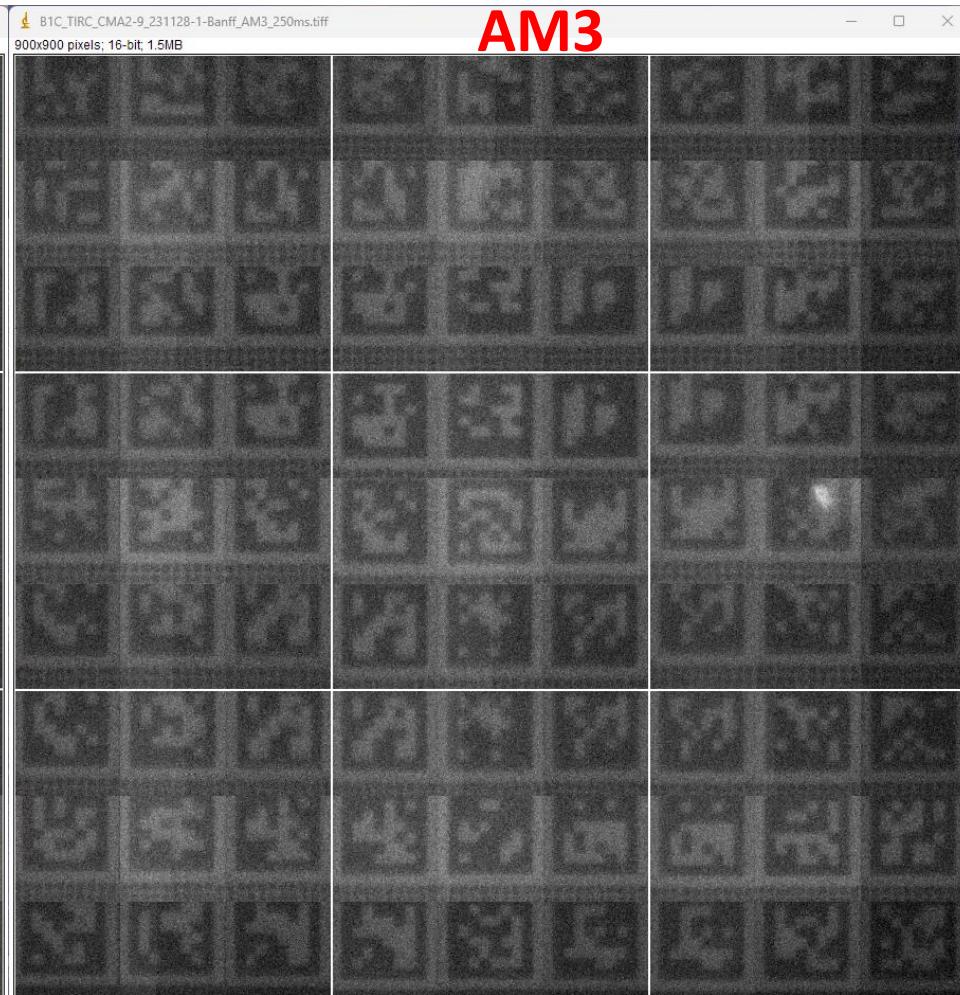
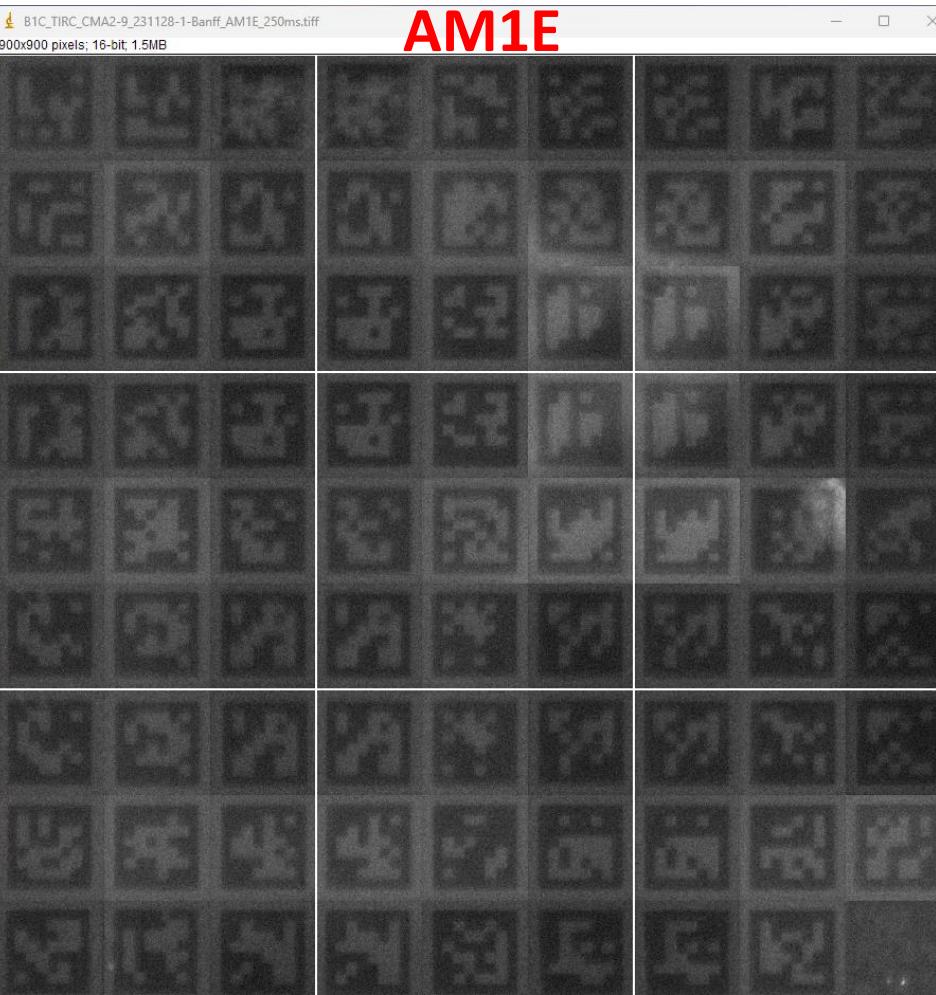
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Gridding Development



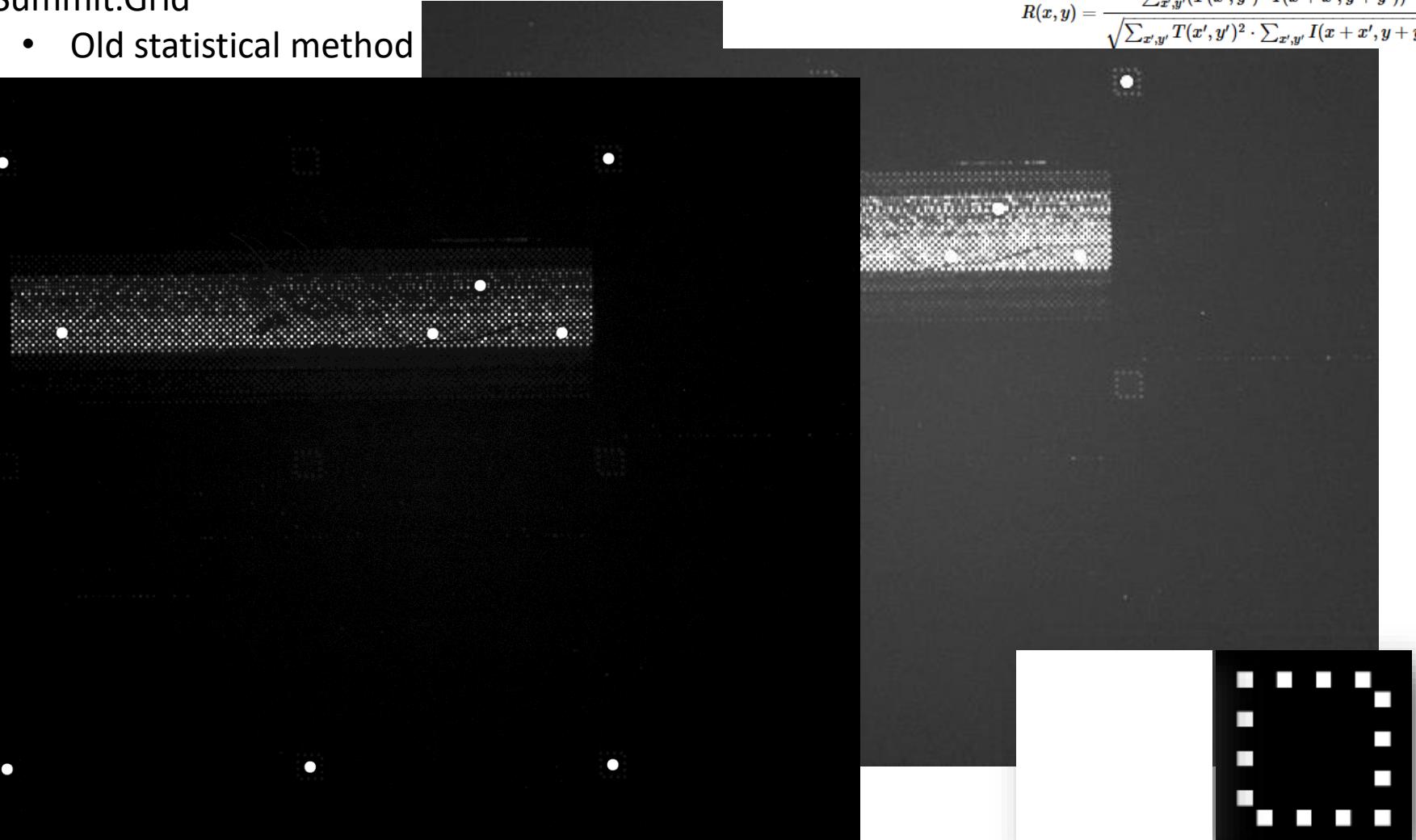
- Summit Grid checking support.



Performance for New Gridding Software



- Summit.Grid
 - Old statistical method



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Performance for New Gridding Software



- Summit.Grid
 - New statistical method

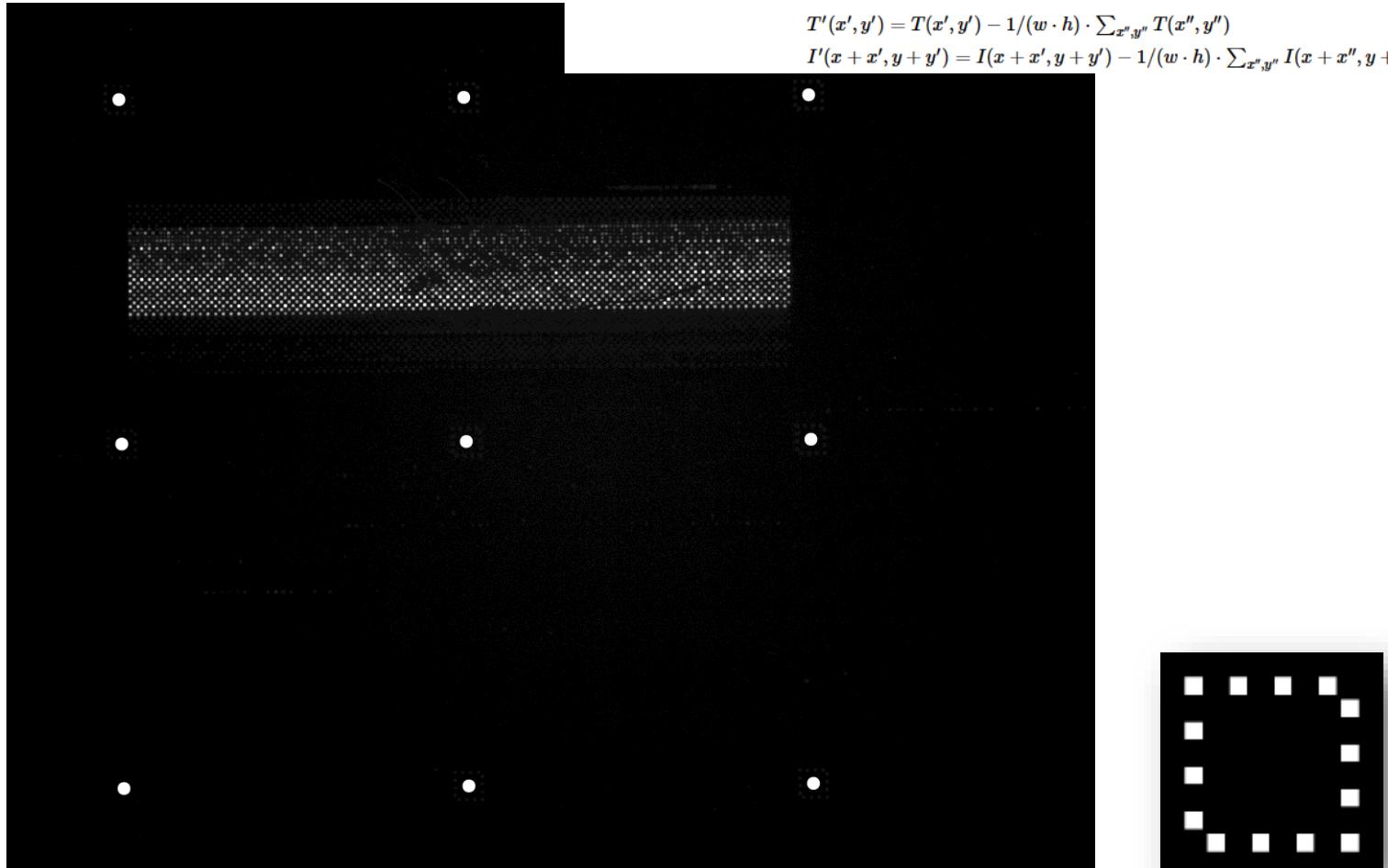
method=TM_CCOEFF_NORMED

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

where

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

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



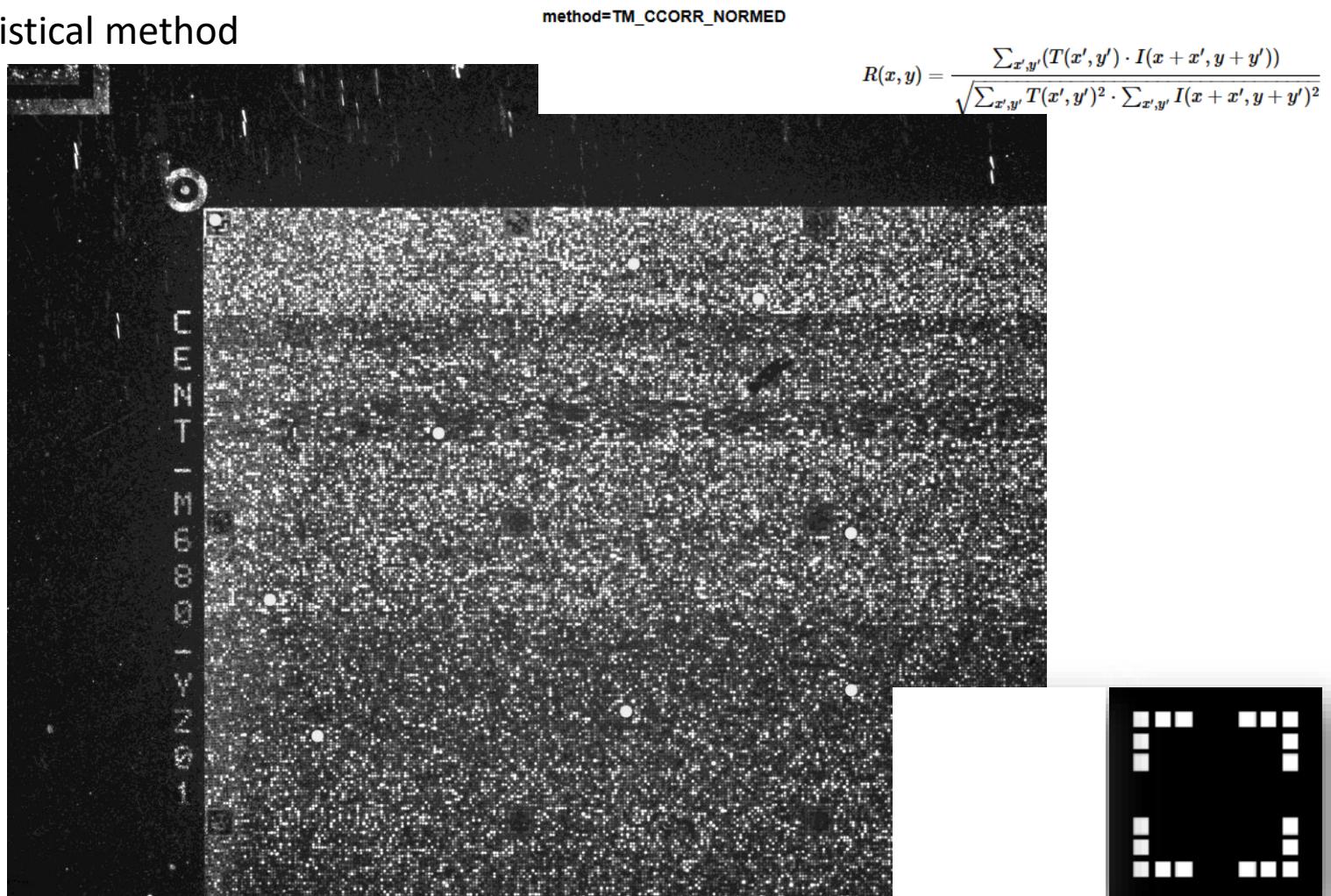
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Performance for New Gridding Software



- Summit.Grid
 - Old statistical method



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Performance for New Gridding Software



- Summit.Grid
 - New statistical method



Centrillion Confidential

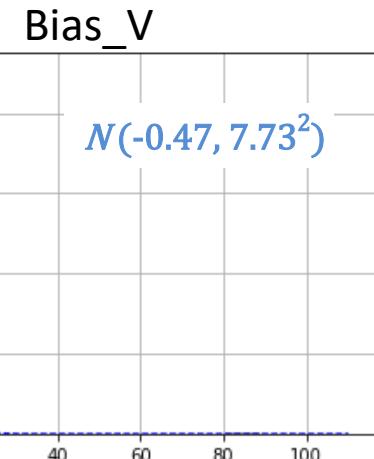
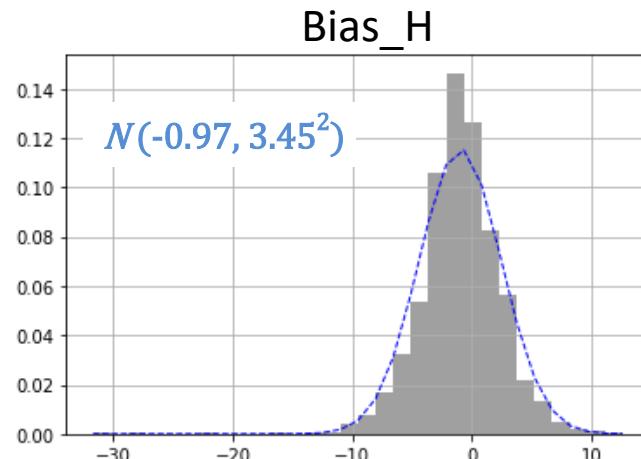
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Sampling Distribution for Different Chip Scan Mode

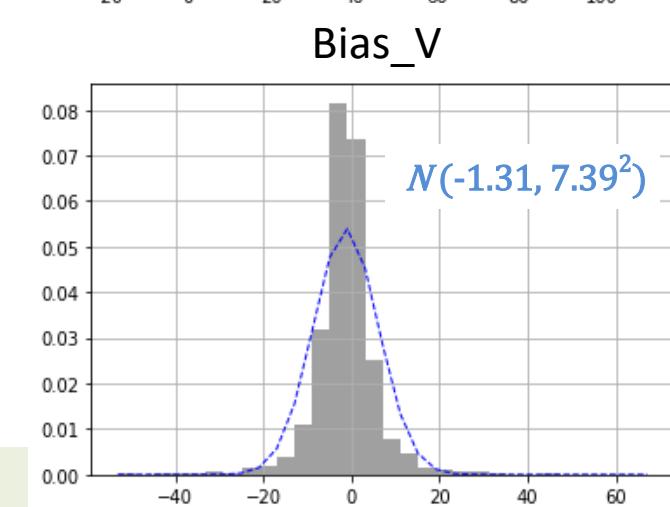
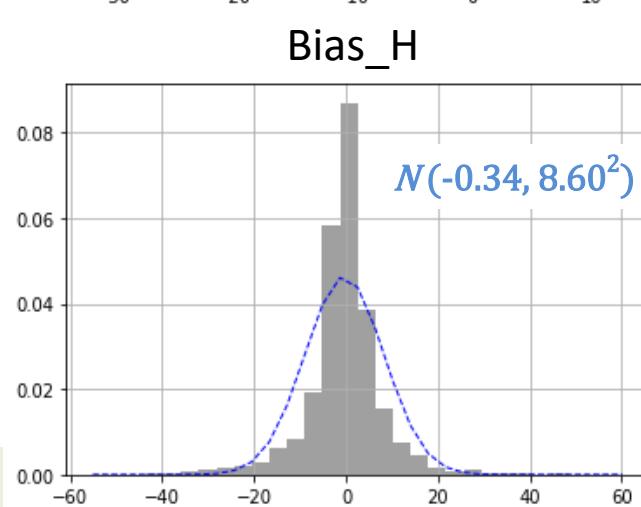


- Quick scan mode experiments
 - Sample: 5 YZ01 chips (7x7 FOVs) x 10 runs => 2450 FOVs
 - Estimation: $\text{Var}(X+Y)$

SUMMIT Test 2



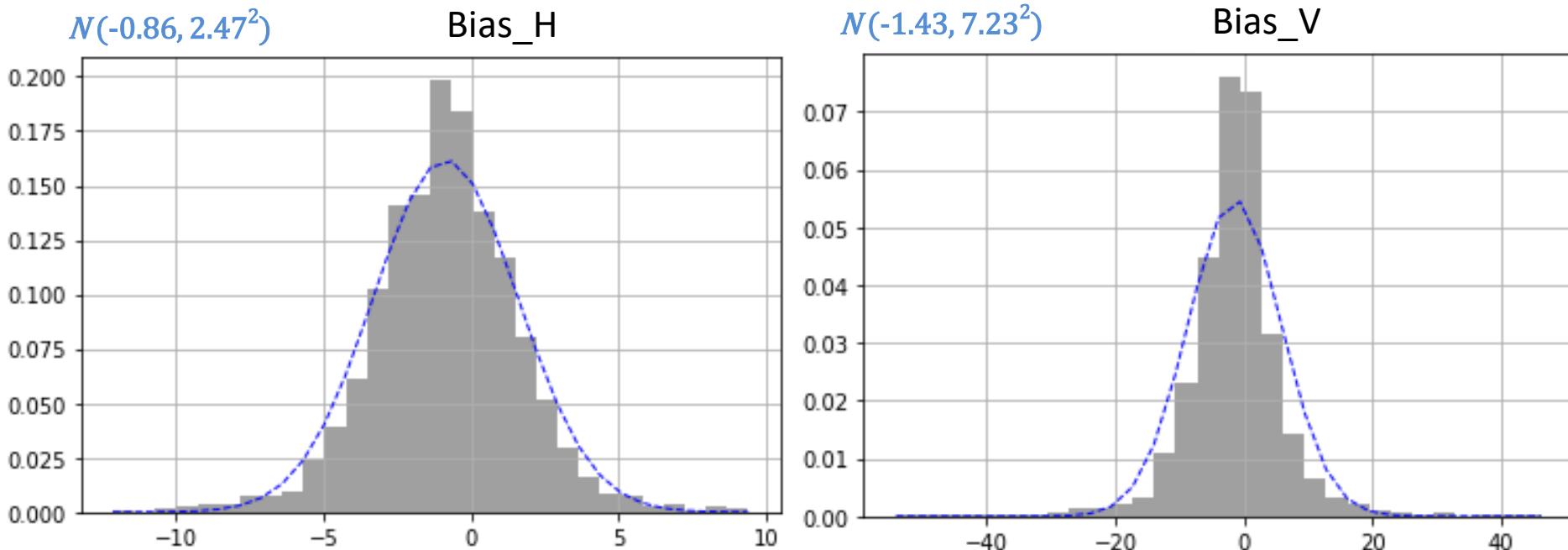
SUMMIT Test 3



Sampling Distribution for Different Chip Scan Mode



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



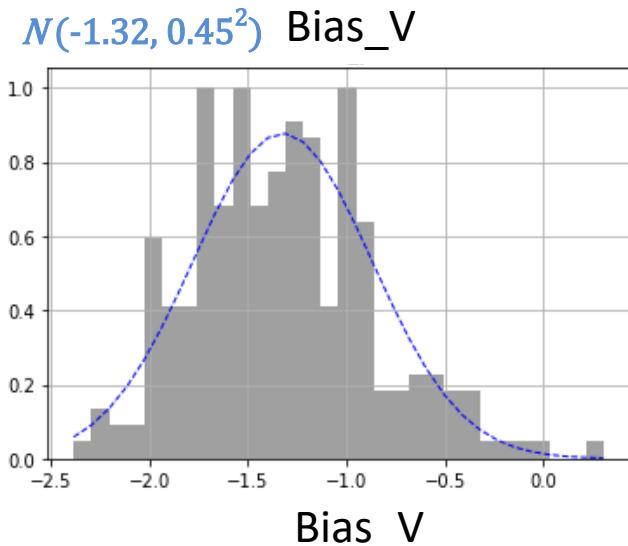
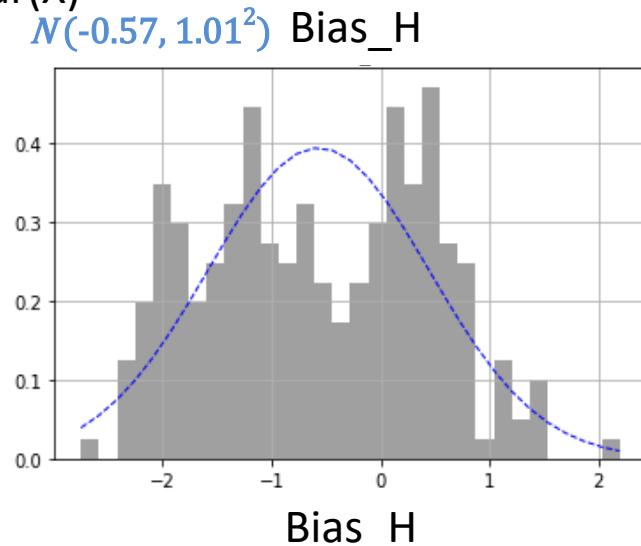
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

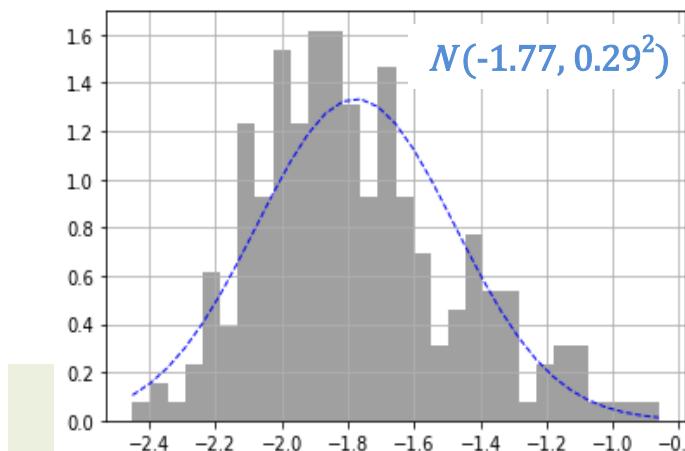
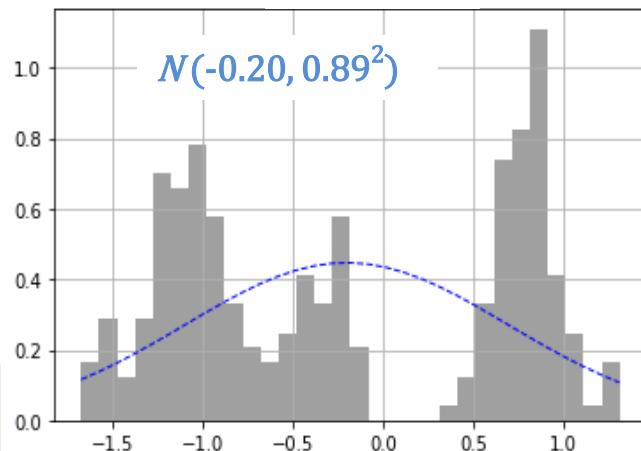
Sampling Distribution for Different Chip Scan Mode

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

SUMMIT Test 2



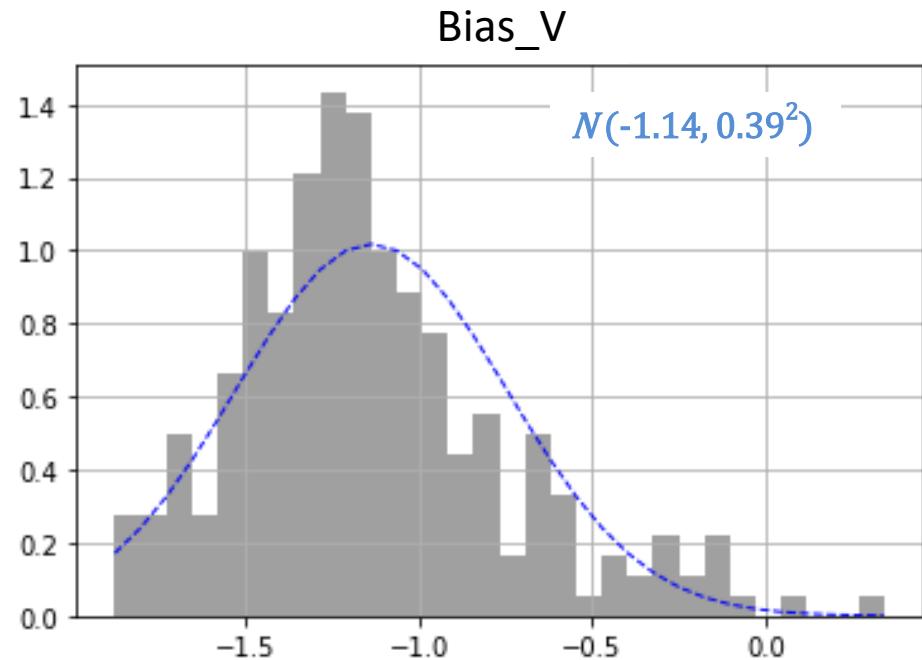
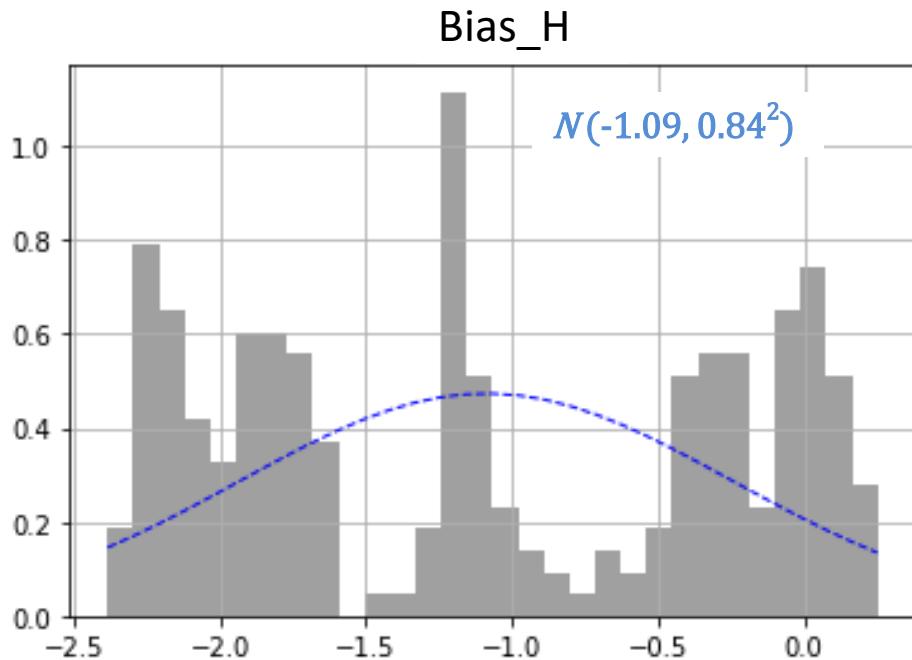
SUMMIT Test 3



Sampling Distribution for Different Chip Scan Mode



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



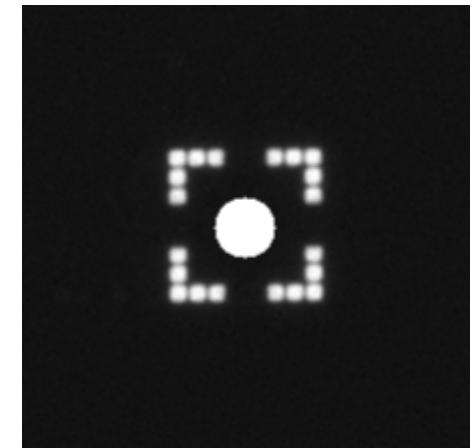
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Statistical Conclusion from Experiments Results



- Gridding
 - SUMMIT Parameters Estimation
 - Let $X \equiv r. \nu.$ of the displacement from changing the filter (BF -> fluorescent).
 - Let $Y \equiv r. \nu.$ of the displacement from relocating the plate to the same position.
 - In the quick scan mode,
Estimate $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
 - In the regular high precision mode,
Estimate $\text{Var}(X)$ Only



Statistical Conclusion from Experiments Results



- Gridding
 - 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.
 - $2.3 \times BF_mark_size$, 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
 - Normal Distribution
 - $P(|T - \mu| < Z_{0.0025} \cdot \sigma) = 99.5\%, Z_{0.0025} = 2.807$, 99.5% confidence.
 - A. Quick scan for estimating $Var(X + Y)$
 - $\sigma = 7.74$, Cover radius: 21.7 (pixels)
 - B. Regular scan for estimating $Var(X)$
 - $\sigma = 1.01$, Cover radius: 2.8 (pixels)
 - Overall Performance - successfully recognized rate: nearly 100%.

Signal Intensity Extraction Techniques Comparison



- Raw Data (NPcall Analyzer)

Grid2 (current)	No.	Data	Grid1	Grid2_subpix	Grid2_subpix_cvfix	Ranking		
85.245%	1	85_46_20210324123447	85.944%	86.131%	86.131%	3	1	1
86.113%	2	85_68_20210303140400	87.582%	87.506%	87.506%	1	2	2
84.400%	3	85_69_20210303142117	87.605%	87.261%	87.261%	1	2	2
85.746%	4	85_76_20210303143808	86.719%	87.244%	87.238%	3	1	2
87.552%	5	85_77_20210303145506	87.751%	88.974%	88.817%	3	1	2
93.293%	6	90_21_20210325132601	93.916%	93.893%	93.945%	2	3	1
92.110%	7	90_54_20210324125308	92.512%	92.587%	92.634%	3	2	1
92.255%	8	90_62_20210324131130	93.036%	92.978%	92.984%	1	3	2
91.748%	9	90_70_20210324132953	92.657%	92.611%	92.611%	1	2	2
90.653%	10	90_77_20210325142116	91.492%	91.533%	91.550%	3	2	1
93.467%	11	95_38_20210324121430	94.015%	94.079%	94.161%	3	2	1
95.798%	12	95_76_20210324164230	96.096%	96.294%	96.311%	3	2	1
94.837%	13	95_77_20210221134715_94_8	95.402%	95.367%	95.379%	1	3	2
95.868%	14	95_77_20210324165927	96.230%	96.317%	96.270%	3	1	2
94.965%	15	95_78_20210221140401	95.688%	95.641%	95.624%	1	2	3

Signal Intensity Extraction Techniques Comparison



- Raw Data (GT Caller)

No.	Data	Grid1	Grid2_subpix	Grid2_subpix_cvfix	Ranking		
1	85_46_20210324123447	92.225%	93.225%	93.358%	3	2	1
2	85_68_20210303140400	93.333%	92.650%	92.450%	1	2	3
3	85_69_20210303142117	93.325%	92.991%	92.958%	1	2	3
4	85_76_20210303143808	92.492%	93.266%	93.050%	3	1	2
5	85_77_20210303145506	93.058%	94.208%	94.192%	3	1	2
6	90_21_20210325132601	98.667%	99.600%	98.825%	3	1	2
7	90_54_20210324125308	98.175%	98.369%	98.183%	3	1	2
8	90_62_20210324131130	98.500%	98.392%	98.458%	1	3	2
9	90_70_20210324132953	98.108%	97.942%	97.950%	1	3	2
10	90_77_20210325142116	97.950%	98.158%	98.158%	3	1	1
11	95_38_20210324121430	99.092%	99.008%	99.025%	1	3	2
12	95_76_20210324164230	99.666%	99.666%	99.683%	2	2	1
13	95_77_20210221134715_94_8	99.392%	99.416%	99.408%	3	1	2
14	95_77_20210324165927	99.583%	99.608%	99.608%	3	1	1
15	95_78_20210221140401	99.367%	99.316%	99.333%	1	3	2

Signal Intensity Extraction Techniques Comparison



- Multiple data NP call results comparison (from finally 38 results)
 - H_0 : Grid1 intensity \geq Intensity extracted from new developed process (Grid2)
 - H_1 : Grid1 intensity $<$ Intensity extracted from new developed process (Grid2)

t-Test: Paired Two sample for Means (NP call Analyzer)

	Grid1	Grid2_subpix
Mean	15307.82185	15322.24077
Variance	2060861.144	2050166.456
Observations	38	38
Pearson Correlation	0.999513654	
Hypothesized Mean Difference	0	
Df	37	
t Stat	-1.980941596	
P(T<=t) one-tail	0.027533779	
t Critical one-tail	1.68709362	
P(T<=t) two-tail	0.055067559	
t Critical two-tail	2.026192463	

	Grid1	Grid2_subpix_cvfix
Mean	15307.82185	15322.34463
Variance	2060861.144	2047607.206
Observations	38	38
Pearson Correlation	0.999572526	
Hypothesized Mean Difference	0	
df	37	
t Stat	-2.123345038	
P(T<=t) one-tail	0.020238129	
t Critical one-tail	1.68709362	
P(T<=t) two-tail	0.040476259	
t Critical two-tail	2.026192463	

t-Test: Paired Two sample for Means (GT Caller)

	Grid1	Grid2_subpix
Mean	16100.97195	16126.25089
Variance	2272377.702	2210252.078
Observations	38	38
Pearson Correlation	0.998958931	
Hypothesized Mean Difference	0	
df	37	
t Stat	-2.182738079	
P(T<=t) one-tail	0.017738267	
t Critical one-tail	1.68709362	
P(T<=t) two-tail	0.035476534	
t Critical two-tail	2.026192463	

	Grid1	Grid2_subpix_cvfix
Mean	16100.97195	16120.53029
Variance	2272377.702	2195827.937
Observations	38	38
Pearson Correlation	0.99898407	
Hypothesized Mean Difference	0	
df	37	
t Stat	-1.672830702	
P(T<=t) one-tail	0.051398391	
t Critical one-tail	1.68709362	
P(T<=t) two-tail	0.102796783	
t Critical two-tail	2.026192463	



Normal Gamma Background Correction & Data Preprocess

Jeff (CHI-HSUAN HO)

- **Model Assumption**

- For each single array:

$$\textcolor{green}{X}_j = \textcolor{orange}{S}_j + \textcolor{blue}{B}_j$$

- $BgC : \textcolor{green}{X}_j \Rightarrow \textcolor{orange}{S}_j$ 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

¹Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway.

²Department 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

Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

- **Models and Notations**

- For each single array j :

$$\textcolor{teal}{X}_j = \textcolor{orange}{S}_j + \textcolor{blue}{B}_j$$

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

- **Models and Notations**

- $X_j = S_j + B_j, \quad X_j \sim f_x(x), \quad S_j \sim f_s(s), \quad B_j \sim f_B(b)$
- 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**

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

$$\Rightarrow X_j \sim f_X(x) \equiv f_{\mu, \sigma, \alpha}^{nexp}(x) = \frac{1}{\alpha} \exp\left\{\frac{\sigma^2}{2\alpha^2} - \frac{x-\mu}{\alpha}\right\} \Phi(\bar{x}), \quad \text{where } \bar{x} = \frac{(x-\mu-\frac{\sigma^2}{\alpha})}{\sigma}$$
$$\Rightarrow \hat{S}^{nexp}(x|\Theta) = \sigma\left(\bar{x} + \frac{\phi(\bar{x})}{\Phi(\bar{x})}\right), \quad \Theta = (\mu, \sigma, \alpha)$$

- If we know $(\hat{\mu}, \hat{\sigma}, \hat{\alpha}) \Rightarrow$ we know $\hat{S}^{nexp}(x)$

- **The Parameter Estimation of normexp Model**

- MLE
- Adapted RMA
- Non-parametric estimation (NP)
- Bayesian estimation

- **The normal-gamma Model**

- $S_j \sim f_s(s) = \begin{cases} \text{Gamma}(k, \theta), & j \in J \\ 0, & j \in J_0 \end{cases}, B_j \sim f_B(b) \Rightarrow N(\mu, \sigma^2)$
 $\Rightarrow X_j \sim f_X(x) \equiv f_{\mu, \sigma, k, \theta}^{ng}(x) = \int f_{k, \theta}^{gam}(t) f_{\mu, \sigma}^{norm}(x - t) dt \Rightarrow fft-based approximation$
 $\Rightarrow \hat{S}^{ng}(x|\Theta) = \frac{\int s f_{k, \theta}^{gam}(s) f_{\mu, \sigma}^{norm}(x-s) ds}{f_{\mu, \sigma, k, \theta}^{ng}(x)} = \frac{k\theta \left(\int f_{k+1, \theta}^{gam}(s) f_{\mu, \sigma}^{norm}(x-s) ds \right)}{f_{\mu, \sigma, k, \theta}^{ng}(x)}$
 $= \frac{k\theta f_{\mu, \sigma, k+1, \theta}^{ng}(x)}{f_{\mu, \sigma, k, \theta}^{ng}(x)} \Rightarrow fft-based approximation$
- 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

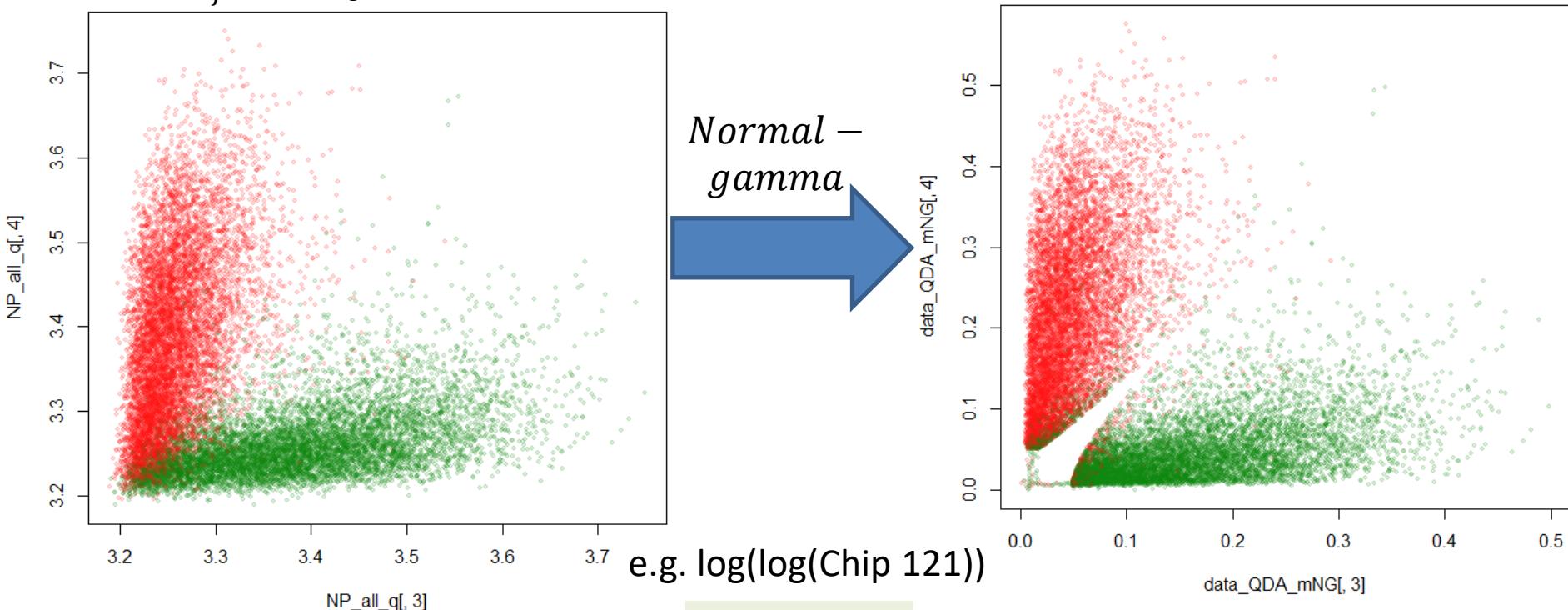


- Normal-gamma BgC.

- For each single probe in each cluster (channel):

$$X_j = S_j + N_j$$

- Normal – gamma* Correction: $X_j \Rightarrow S_j$. Enhance the biological validity of the results.
- N_j represents the noise. \Rightarrow Assumption: Normal distribution.
- Remove N_j and use Gamma distribution to estimate S_j .
- $X_j \Rightarrow$ Normal-gamma distribution.



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

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 ($\log(\log(\cdot))$)	
GMM + EM	94.2308%
QDA	94.5863%
ALL_NG + QDA	94.2774%

Call rate ($\log(\log(\cdot))$)	
GMM + EM	66.8298%
QDA	67.8205%
ALL_NG + QDA	67.7681%

Centrillion Confidential

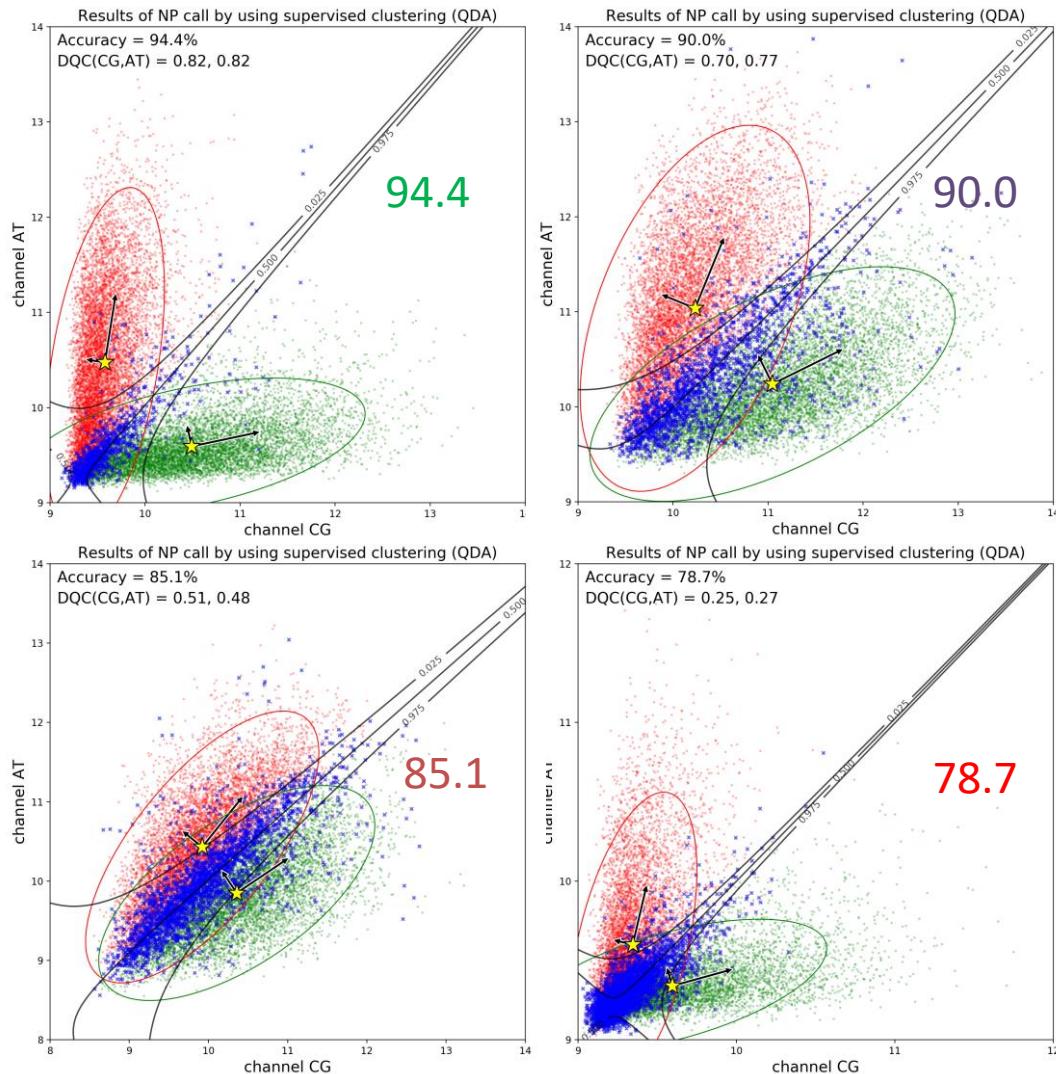
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Data Preprocess



- Data Example

Wafer	Chips	Np call rate (%) (NP call Analyzer)
198-04	121	94.4
197-02	230	90.0
198-15	277	85.1
197-02	233	78.7



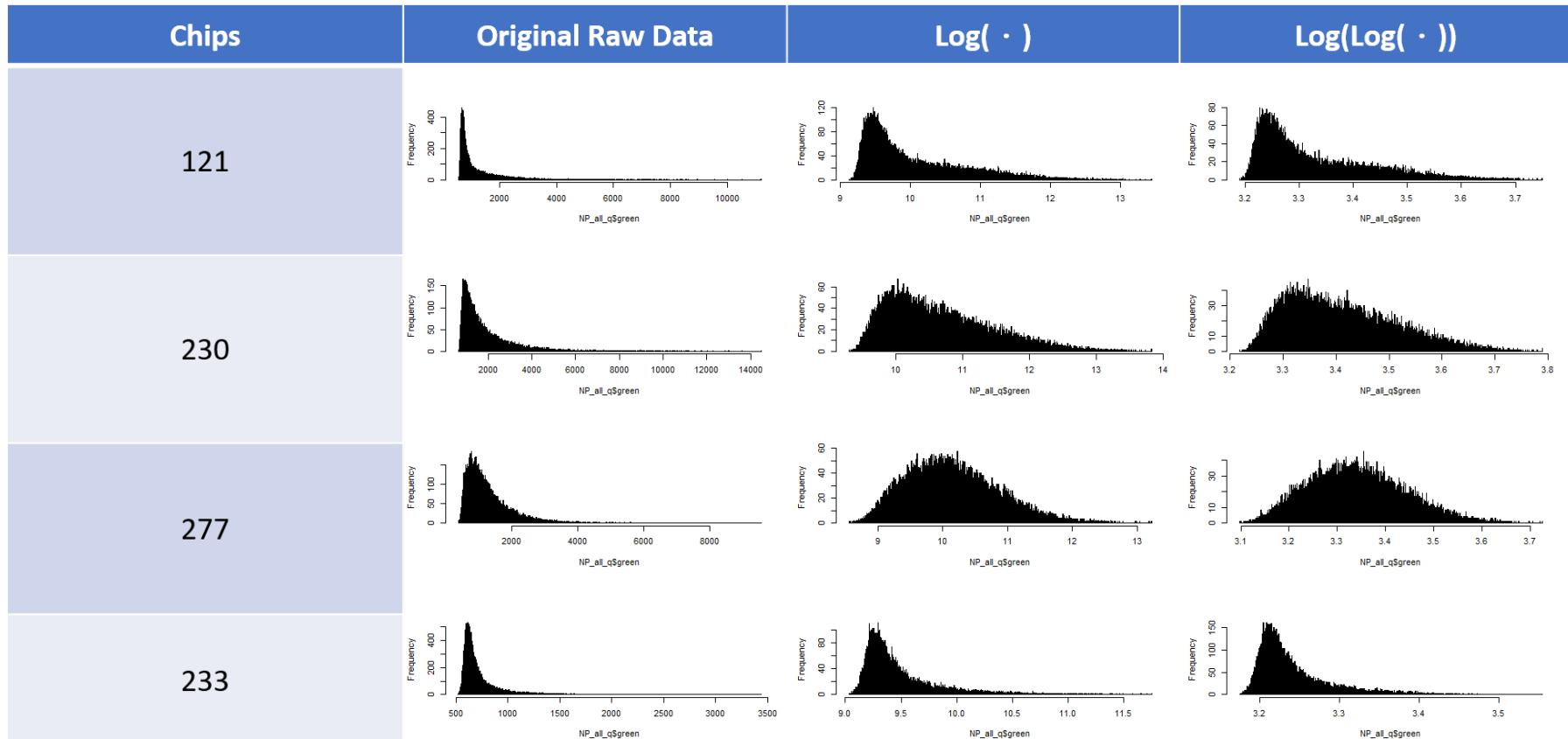
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Data Preprocess



- Logarithm Effect



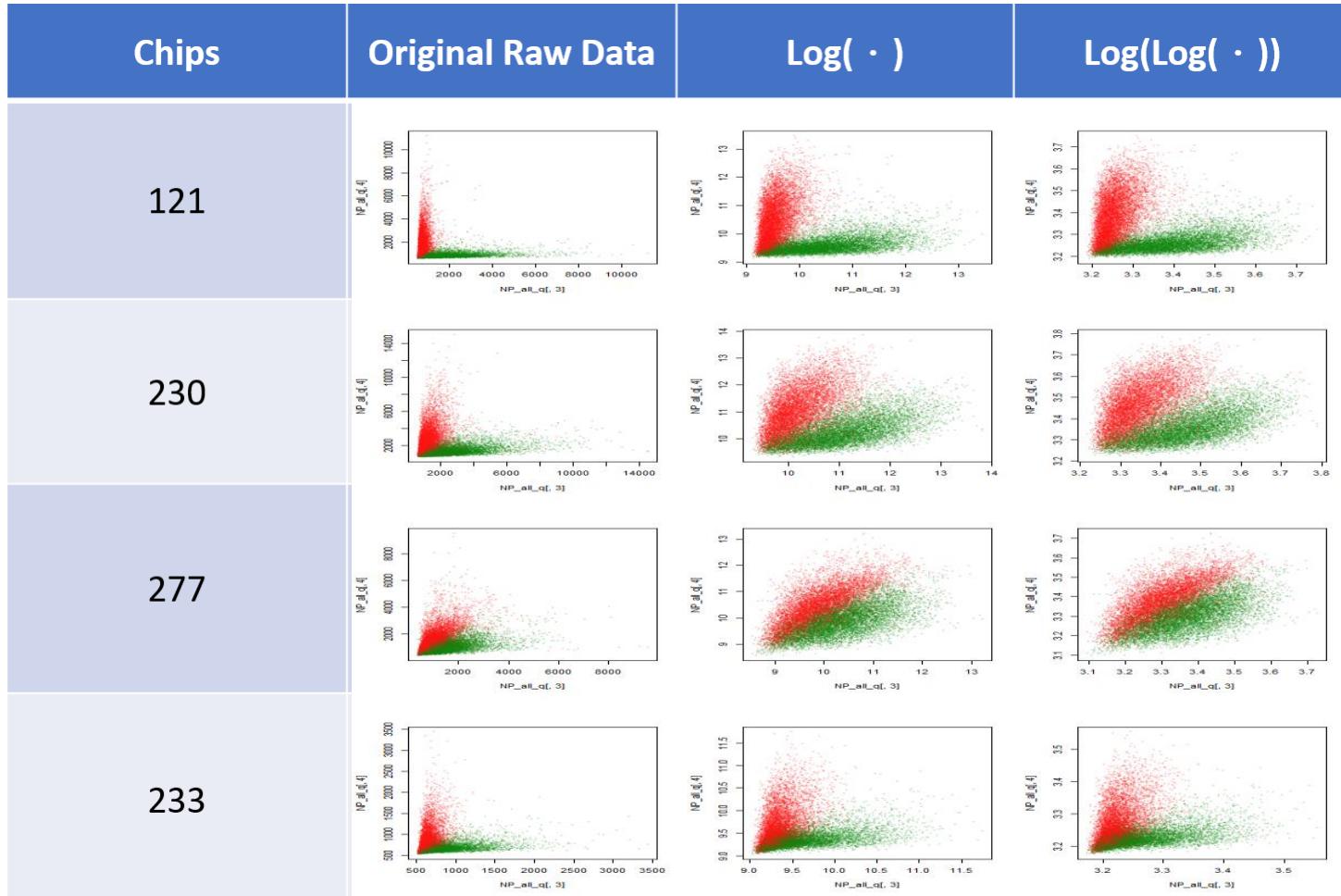
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Data Preprocess



- Logarithm Effect



Centrillion Confidential

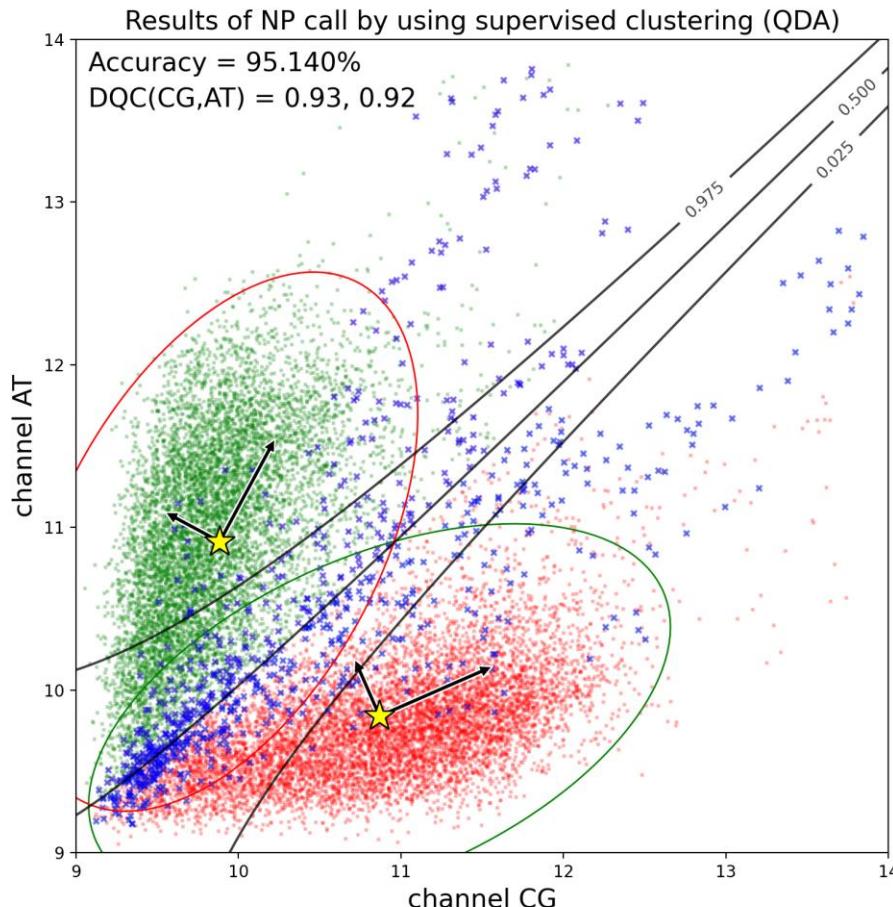
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.



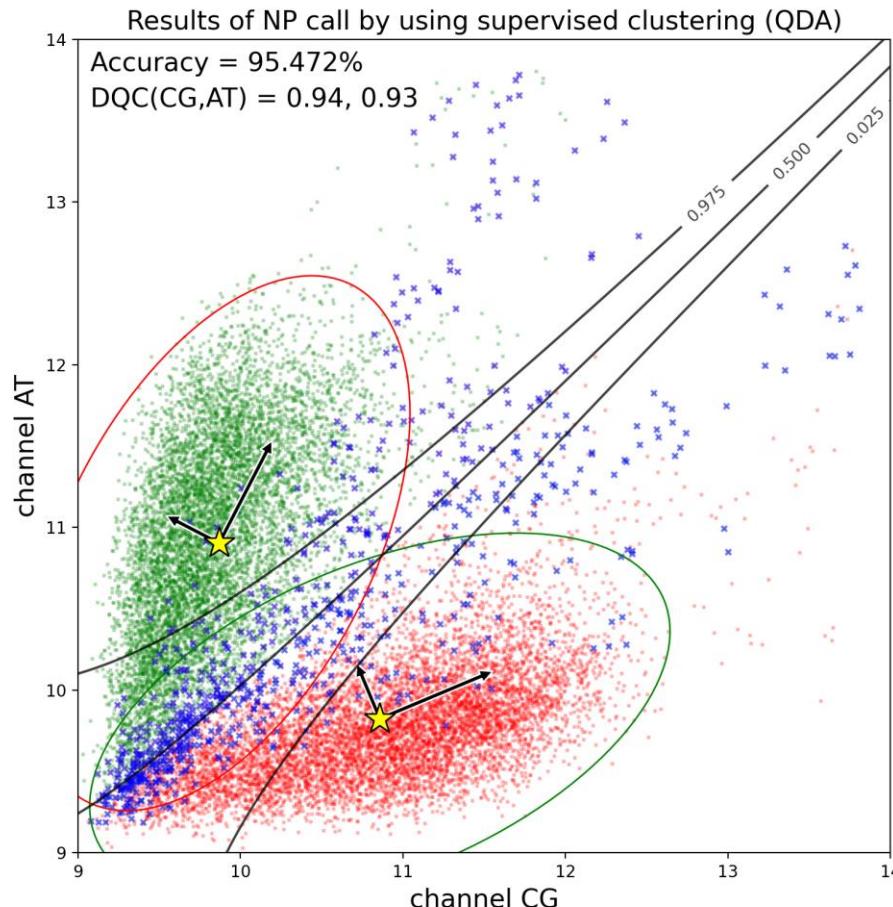
Chip QC and NPcall Analyzer

Jeff (CHI-HSUAN HO)

Chip No.54 Quality Control

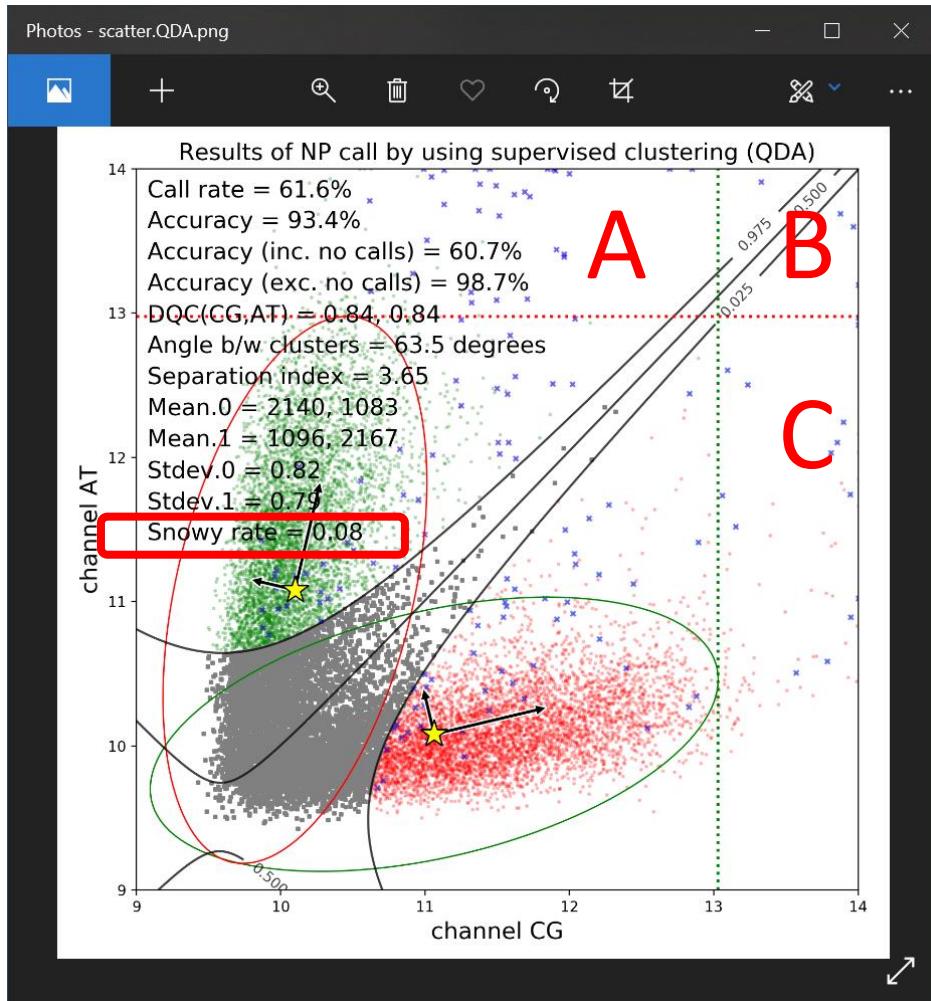


Chip NO.62 Quality Control



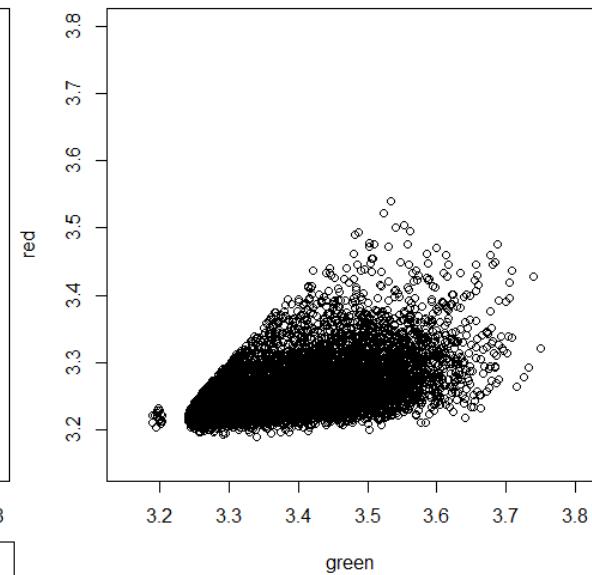
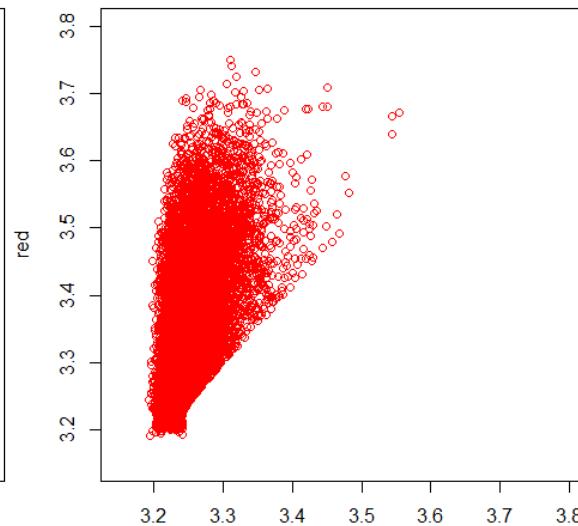
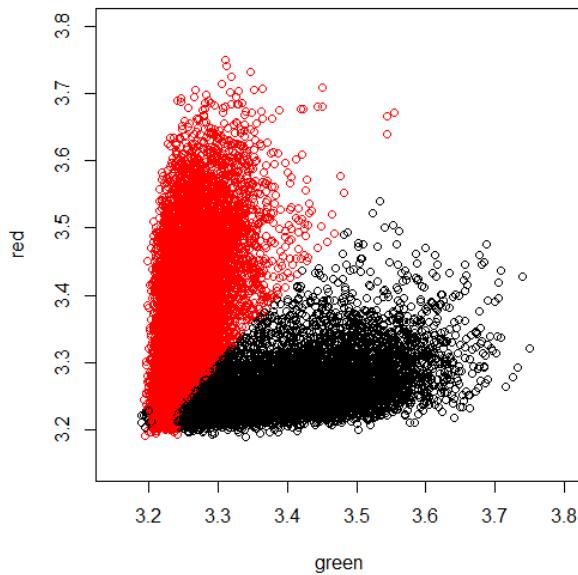
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.



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

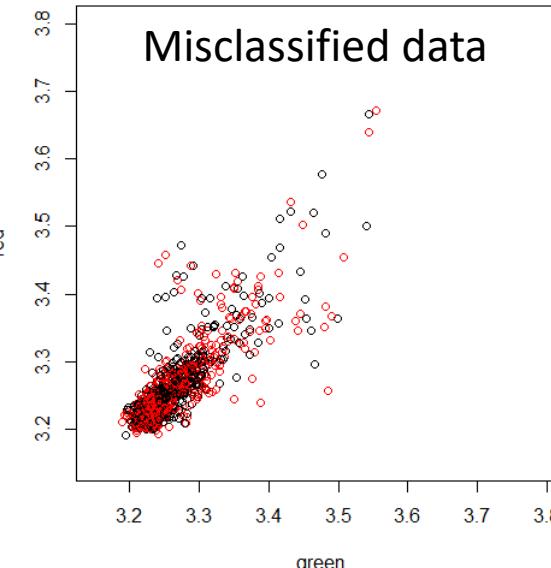
- GMM-EM Results (log(log))



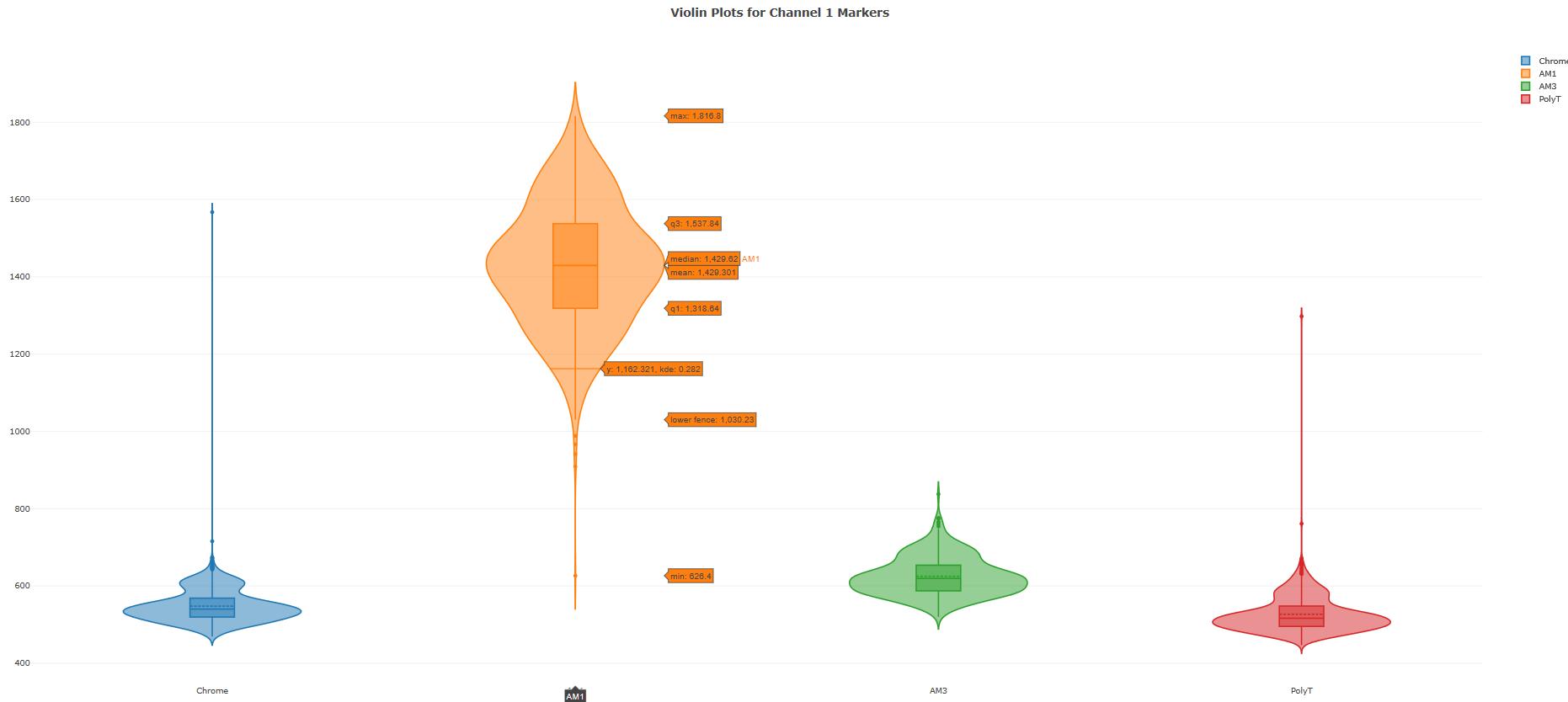
NP call: 94.2308%

No Call Rate: 66.8298%

Misclassified data



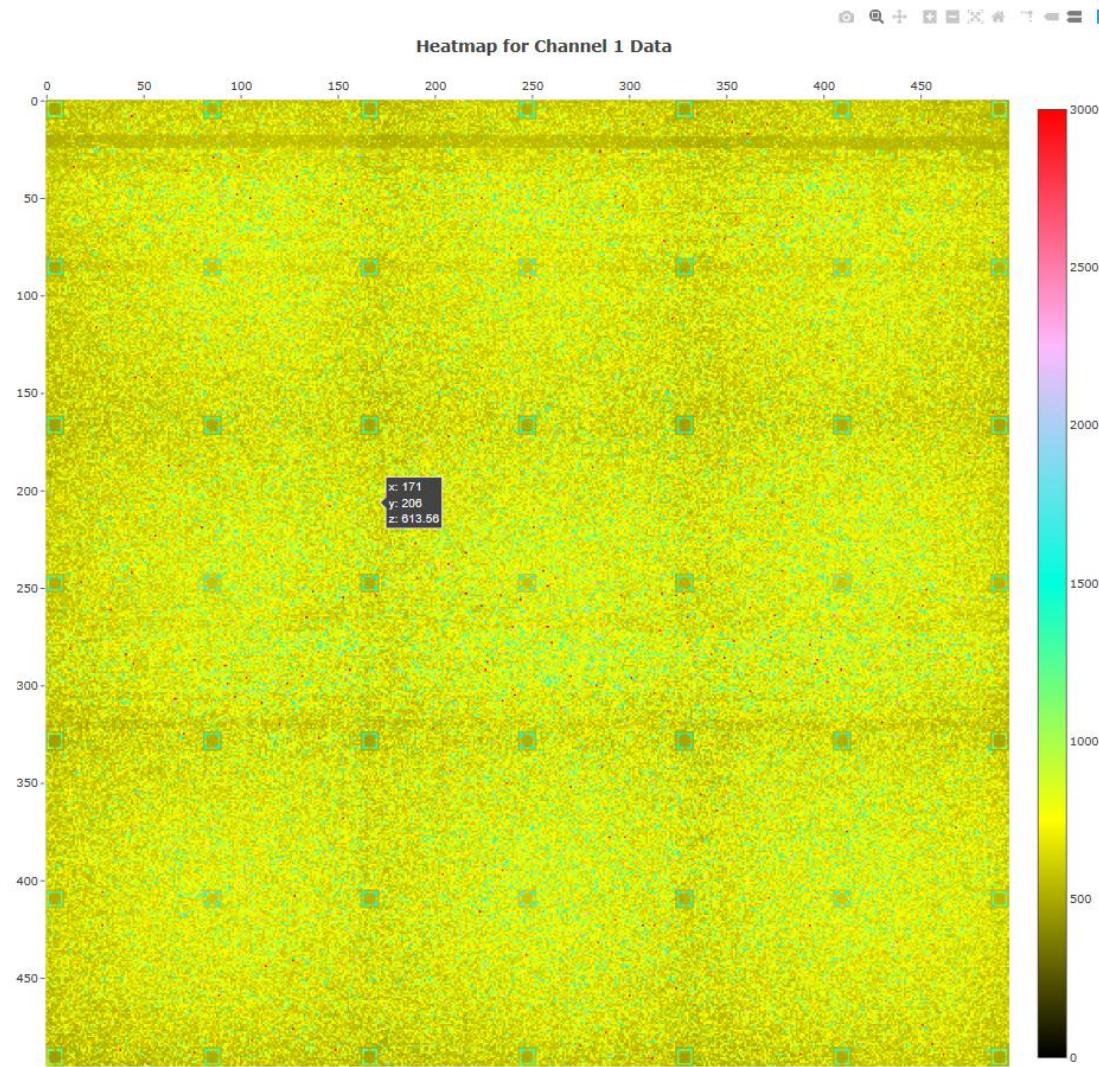
Banff chip QC – Violin Plot



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Banff chip QC – Heatmap



Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Banff chip QC – Marker Raw Data



marker_ch1.csv - Excel

檔案 常用 插入 頁面配置 公式 資料 校閱 檢視 說明 搜尋

自動儲存 (○關閉) | 貼上 | 新細明體 | 12 | A⁺ A⁻ | 自動換行 | 通用格式 | \$ % , | 00 00 | 設定格式化 | 格式化為儲存格的條件 | 表格 | 樣式 | 插入 | 刪除 | 格式 | 儲存格 | 編輯 | 共用 | 註解 | 奇軒 何 | 回 | 一 | □ | × |

剪貼簿 | 字型 | 對齊方式 | 數值 | 樣式 | 儲存格 | 編輯 |

A1 : fx 487.24

	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								
2	492.84	908.88	1176.56	1107.6	604.4	610.56	1087.88	1232.97	1247.84	512.56	nan								
3	505.84	941	503.44	487.92	471.76	481.48	501.96	520.93	1232.72	521.52	nan								
4	512.08	626.4	499.52	493.76	471.12	500.76	500.08	515.4	1231.2	503.76	nan								
5	496.16	561.28	493.92	488.36	489.32	492.52	486.2	503.97	542.04	496.28	nan								
6	501.4	569.76	504.92	534.2	480.52	498.84	485.92	516.13	621.84	524.84	nan								
7	499.88	1113.76	501.52	489.4	500.6	477.44	506.6	490.5	1230.76	521	nan								
8	514.33	1297.9	525.1	479.67	493.3	478.57	508.27	502.22	1192.9	547.2	nan								
9	520.96	1349.44	1247.16	1273.52	683.08	607	1319.56	1300.5	1266.76	526.32	nan								
10	502.52	514.36	528.24	507.64	513.48	503.16	524.92	520.63	537.84	509.72	nan								
11	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	
13	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	
15	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	
17	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	

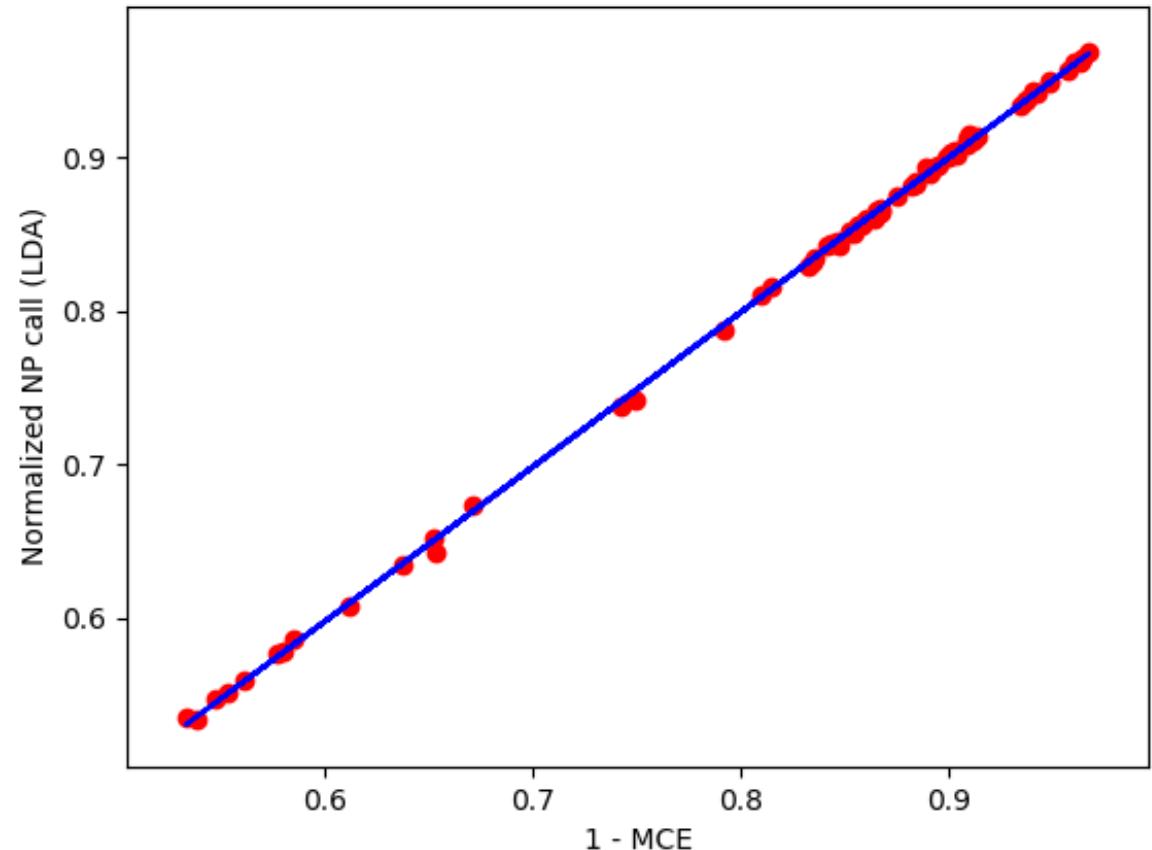
Centrillion Confidential

All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.

Correlation and Regression analysis (MCE vs. NP call)



MCE vs. NPcall (Linear Regression)



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.