



Reference (1)

Jeff (CHI-HSUAN HO)

The subsequent slides only showcase a few parts of the work materials and results.



Bioinformatics Analyses – 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

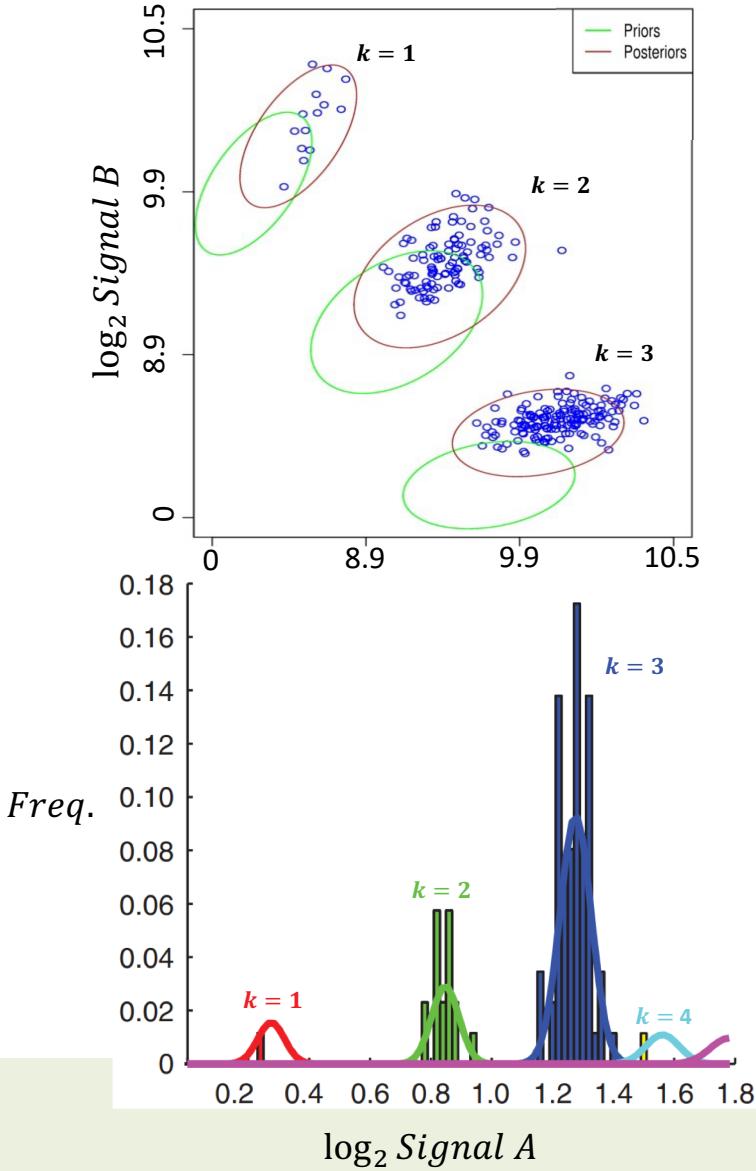
GTC v4.2 P/N 702982 Rev. 3

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.

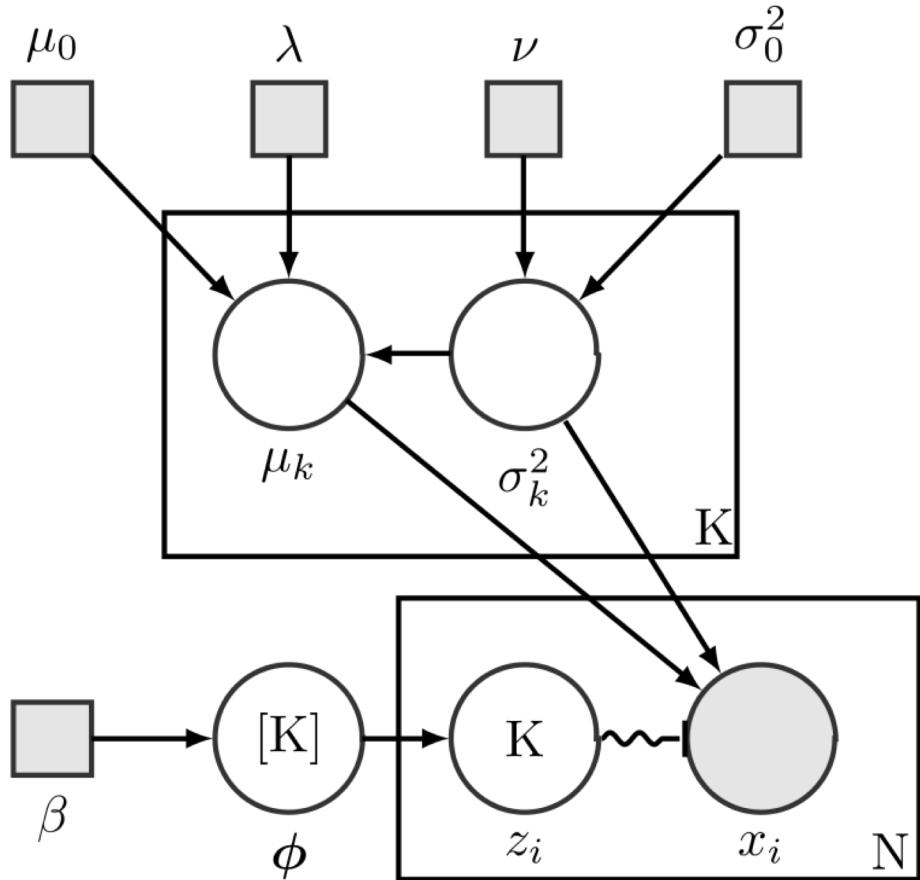
Gaussian Mixture Model (GMM)

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

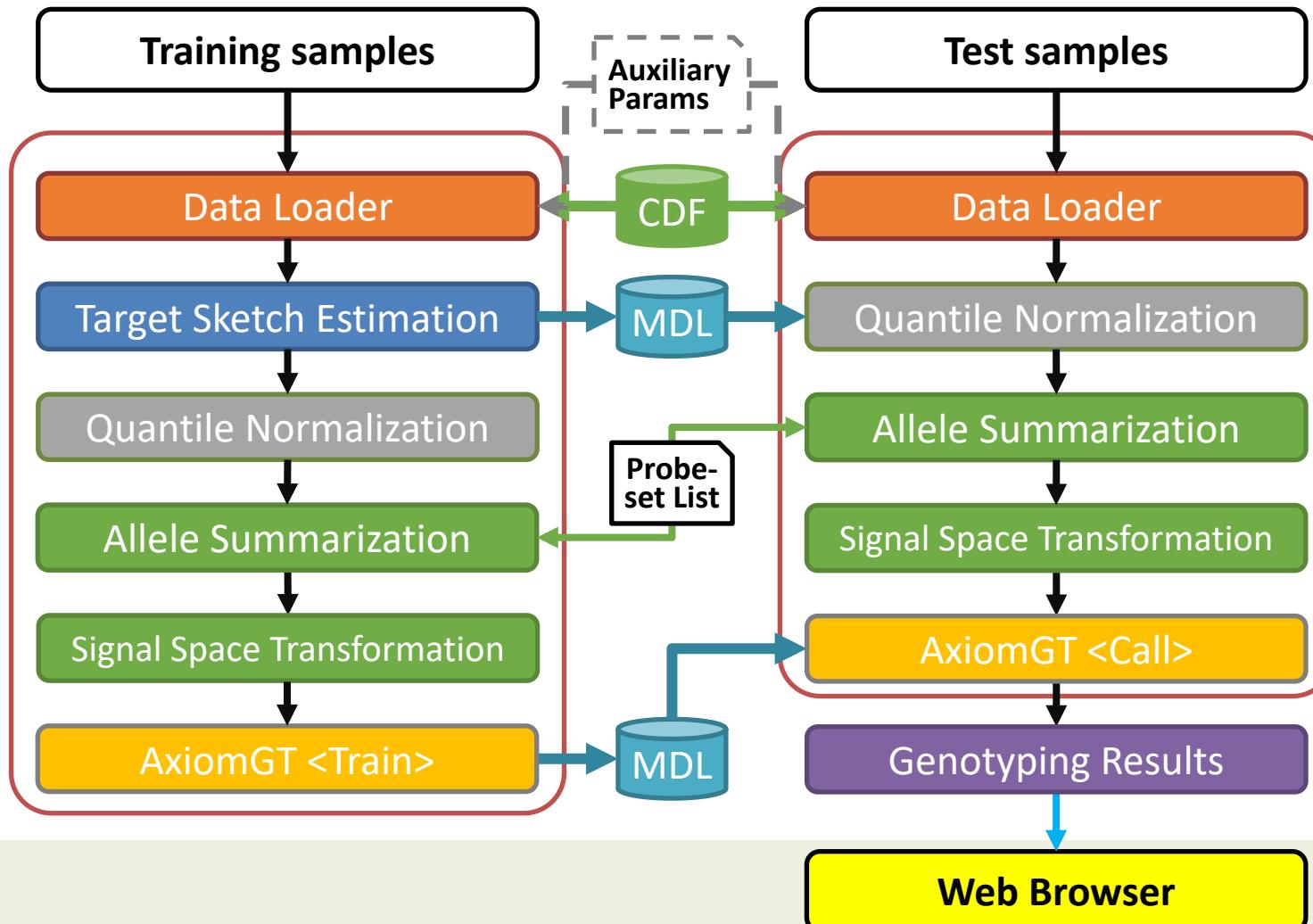


- Model: $\sum_{k=1}^K \phi_k N(\mu_k, \sigma_k^2)$
- K = Number of mixture components
- σ_k^2 = Variance of component k ,
 $\sigma_k^2 \sim \text{Inverse-Wishart}(\nu, \sigma_0^2)$
- μ_k = Mean of component k ,
 $\mu_k \sim \text{Normal}(\mu_0, \lambda \sigma_k^2)$
- N = Number of observations
- z_i = 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>
```

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

Special SNPs

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

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

Genders

Sex Probes

Genotyping Analysis Development and Distribution



Genotyping Results

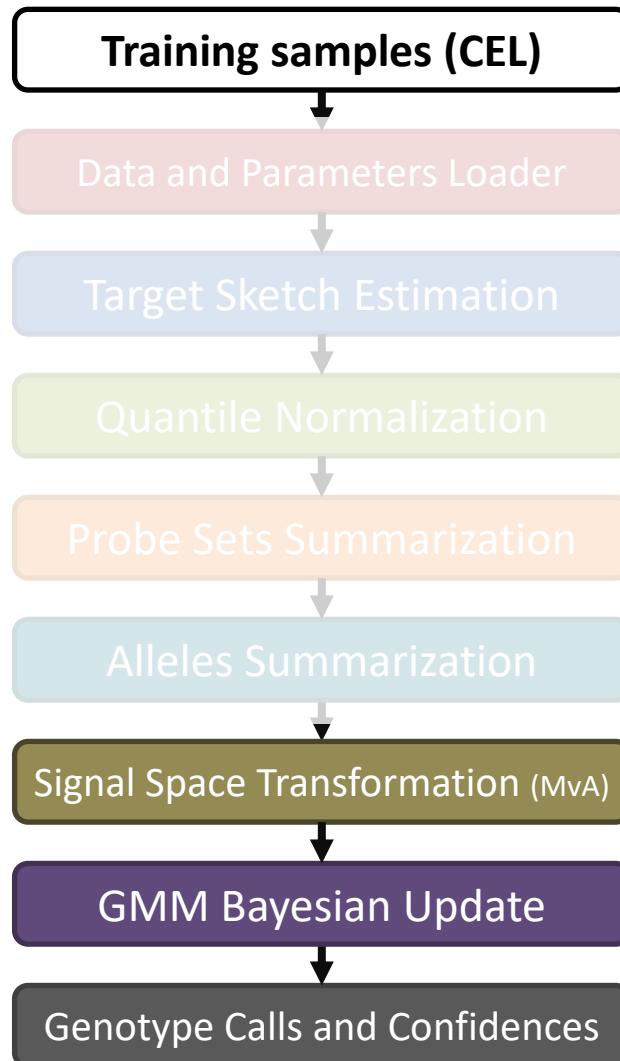
Call Rate

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

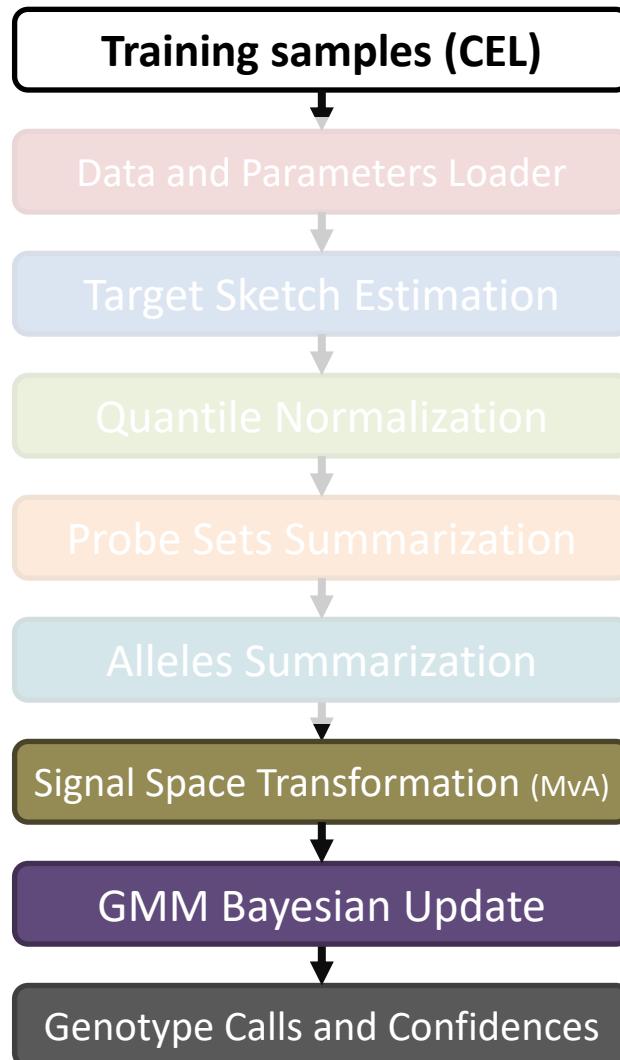


- AxiomGT1 Algo (`bayes_label`)
 - `initialize_bins`
($x \rightarrow \text{bins domain}$)
 - `Integratebrlmmoverlabelings`
(bins domain)
 - `labels_two_posterior`
(bins domain)
 - `make_two_calls`
($\text{bins domain params results} \rightarrow x \text{ calls}$)

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.

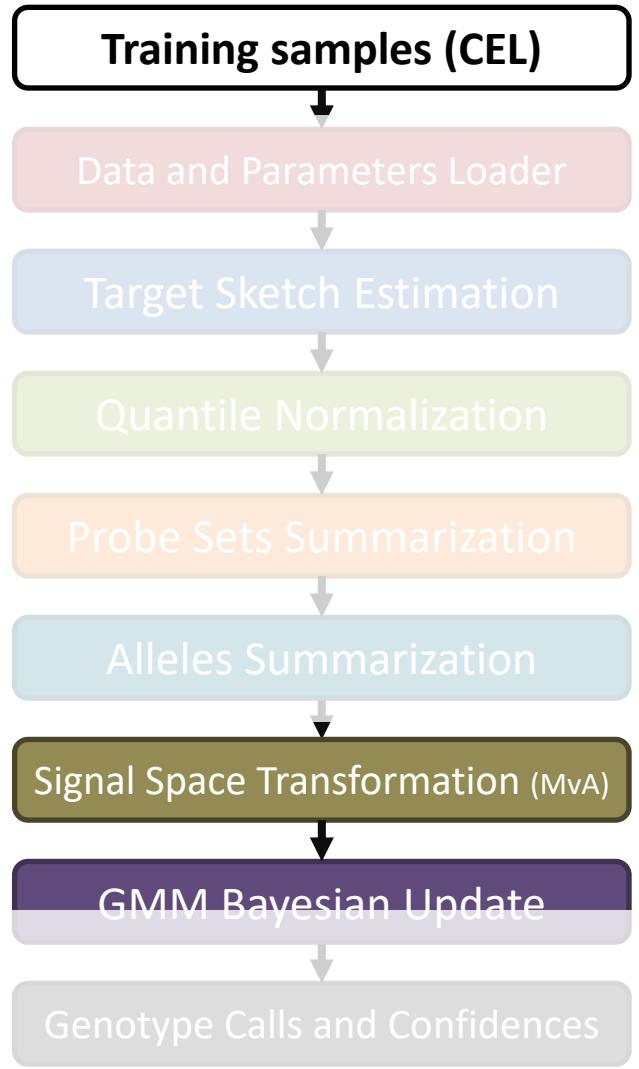
Genotyping Analysis Development and Distribution



- 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
 1. When current data point x lies outside current boundary (previous $x + \text{delta}$) \Rightarrow Create new boundary by using current x .
 2. $\text{sp.bins} = 0$ (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)

Centrillion Confidential

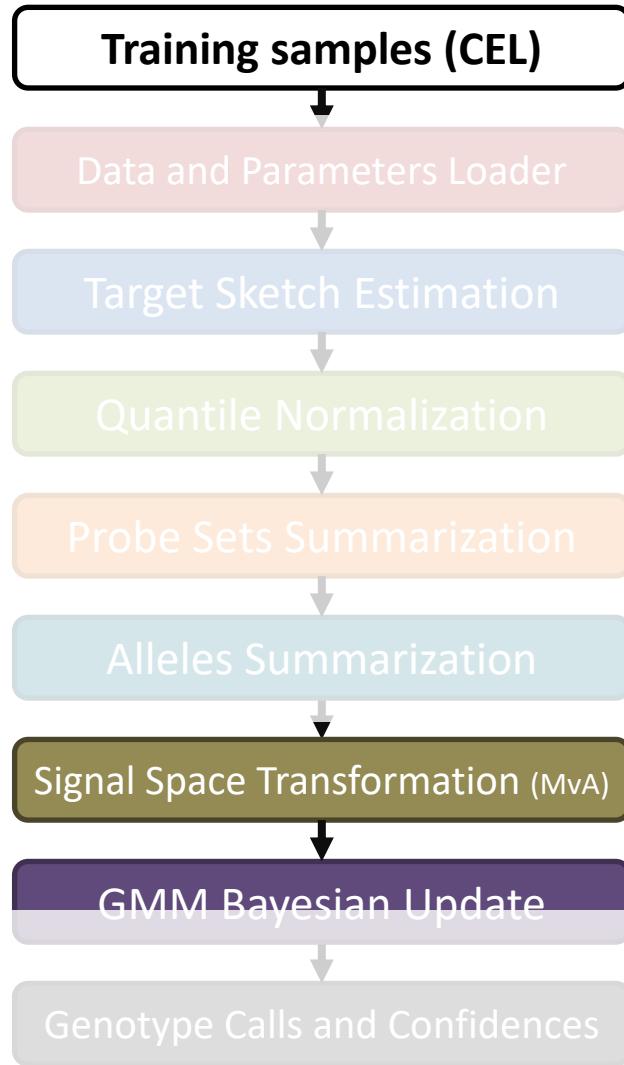
Genotyping Analysis Development and Distribution



- **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

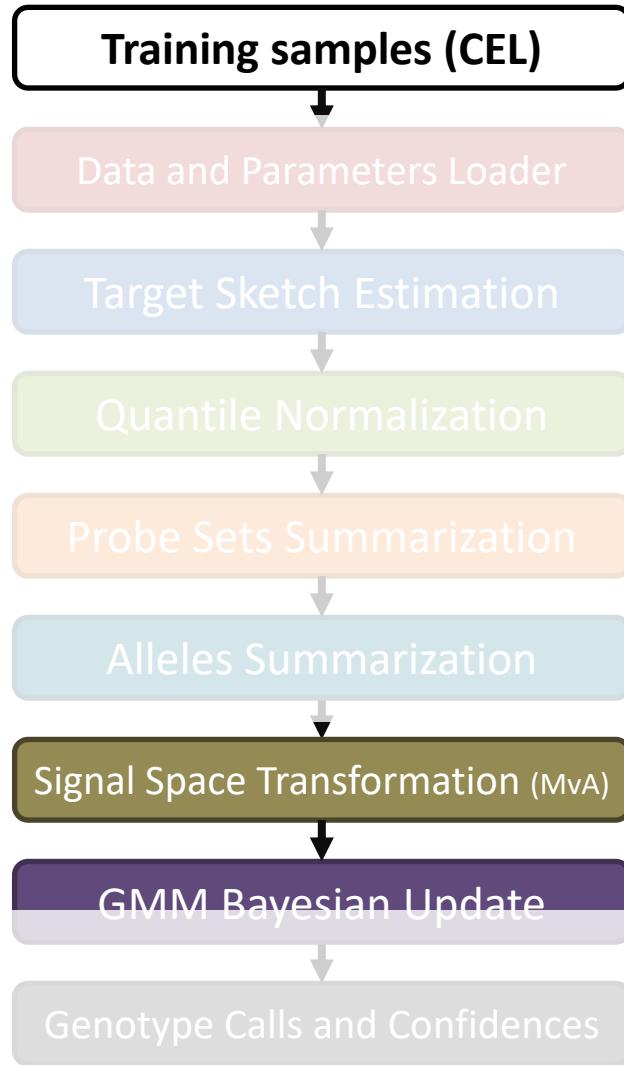


- **Quality Score (Posterior Analog) (for each partition)**
 - **(sp.mix) mixture_penalty**
 - $\sum_{g=1}^3 N_g * \log\left(\frac{(N_g + \text{lambda})}{\sum_g (N_g + \text{lambda})}\right)$, $\text{lambda} \in \{1 (BB), 2 (AB), 3 (AA)\}$
 - **(sp.bic) BIC (k*log(N) part)**
 $c * \text{bic_k} * \log\left(\sum_g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

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}

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

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

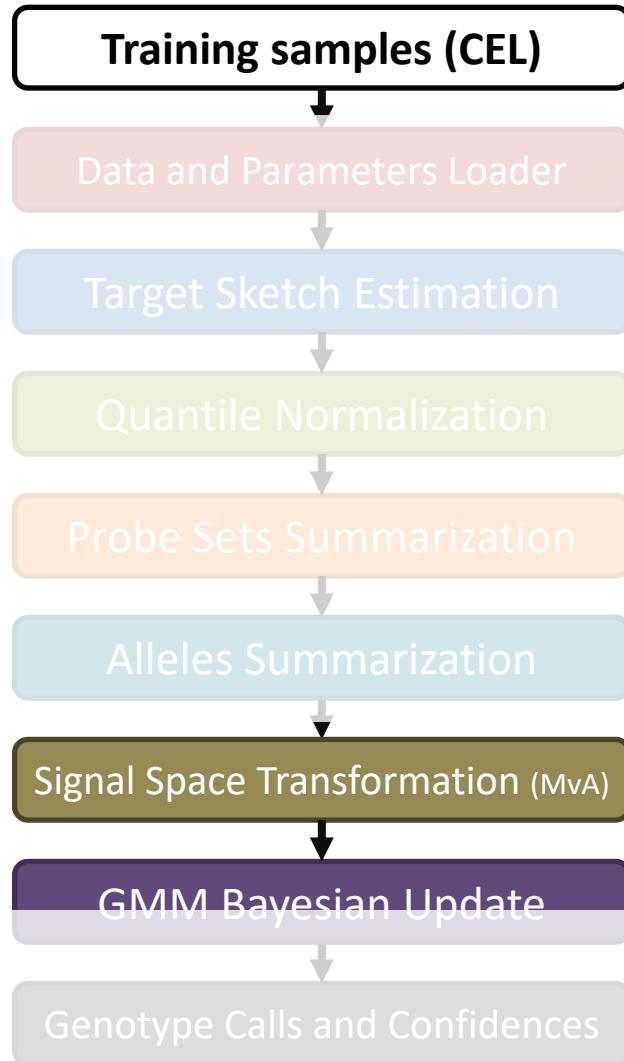
$$u'_{3x1} = \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \end{bmatrix}, \quad \begin{array}{ll} u'_1 = u'_1 + \text{delta} - \gamma \\ u'_2 = u'_2 - \gamma \\ u'_3 = u'_3 - \text{delta} - \gamma \end{array}$$

Pool Adjacent-Violators (PAV) algo.

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

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

Genotyping Analysis Development and Distribution

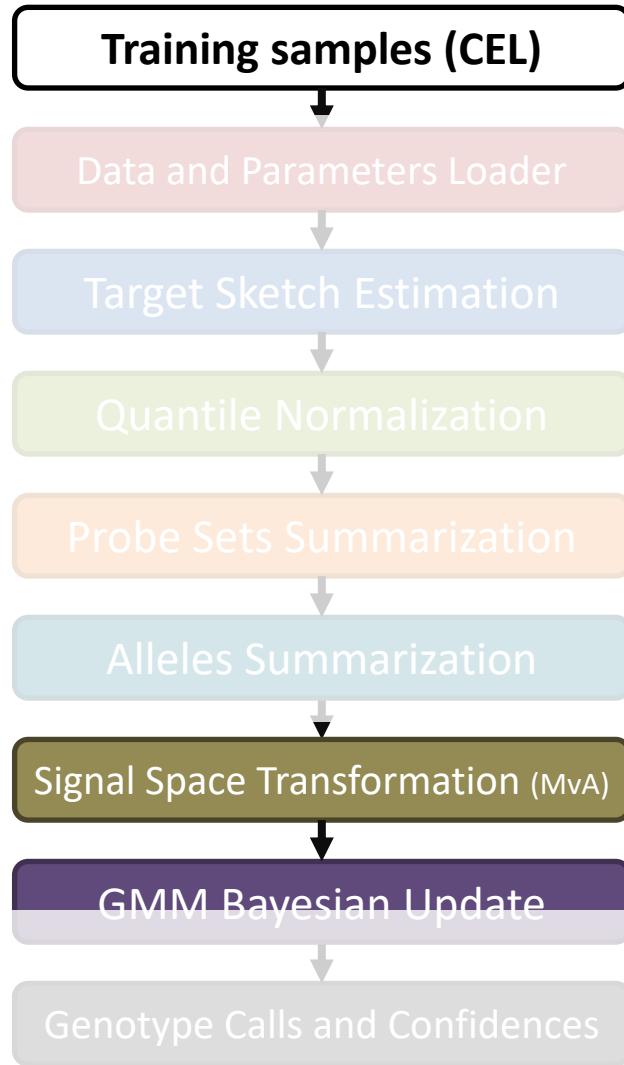


- **Quality Score (Posterior Analog) (for each partition)**
 - **1D (x) Log likelihood under posterior params**
 - $u'_{3x1} = (K_0^{-1}_{3x3} + N'_{3x3})^{-1} * (K_0^{-1}_{3x3} * u_{0,3x1} + N_{3x3} * m_{3x1})$,
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least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

$$\frac{v_{g0} * \sigma^2_{g0} + \sum_i x_{gi}^2 - \sum_i x_{gi} * \sum_i x_{gi} * \frac{1}{N_g + 0.0001} + \frac{k_g * N_g}{k_g + N_g} * (u'_g - u_{g0})^2}{v_{g0} + N_g}, \quad g = 1, 2, 3$$
 - (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

Genotyping Analysis Development and Distribution

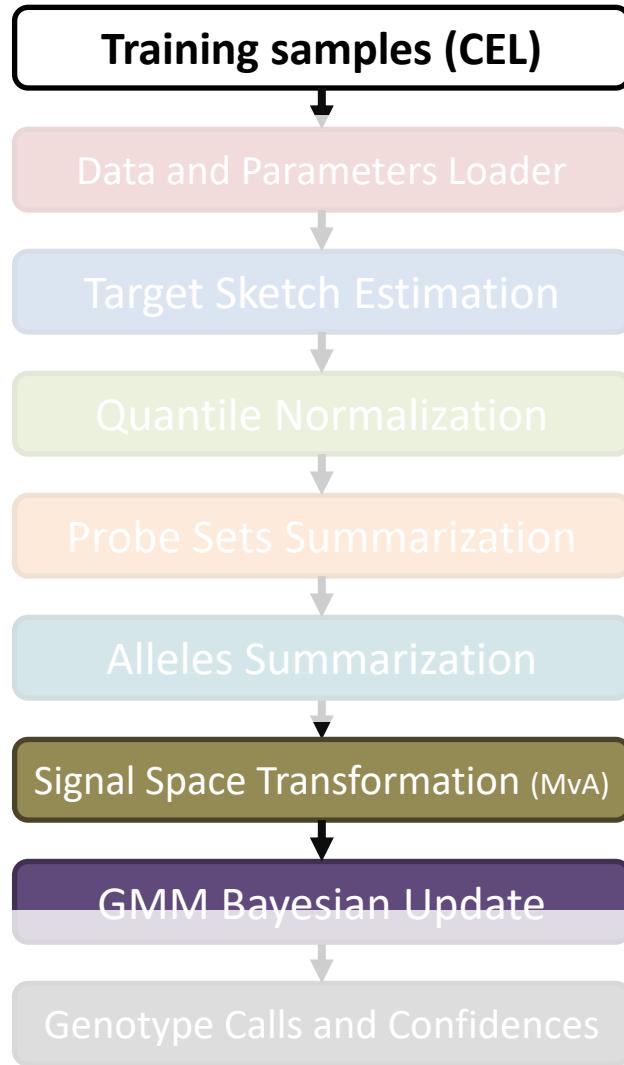


- Quality Score (Posterior Analog) (for each partition)
 - 1D (x) Log likelihood under posterior params
 - $u'_{3 \times 1} = (K_0^{-1}_{3 \times 3} + N'_{3 \times 3})^{-1} * (K_0^{-1}_{3 \times 3} * u_{0 \times 3 \times 1} + N_{3 \times 3} * m_{3 \times 1})$,
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least delta (0.75)). $u'_{3 \times 1}$
 - $\sigma'^2_g = \frac{v_0 * \sigma_{g0}^2 + \sum_i (x_{gi} - x_g)^2 + \frac{k_g * N_g}{k_g + N_g} * (u'_g - u_{g0})^2}{v_0 + N_g}, \quad g = 1, 2, 3$
 - (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

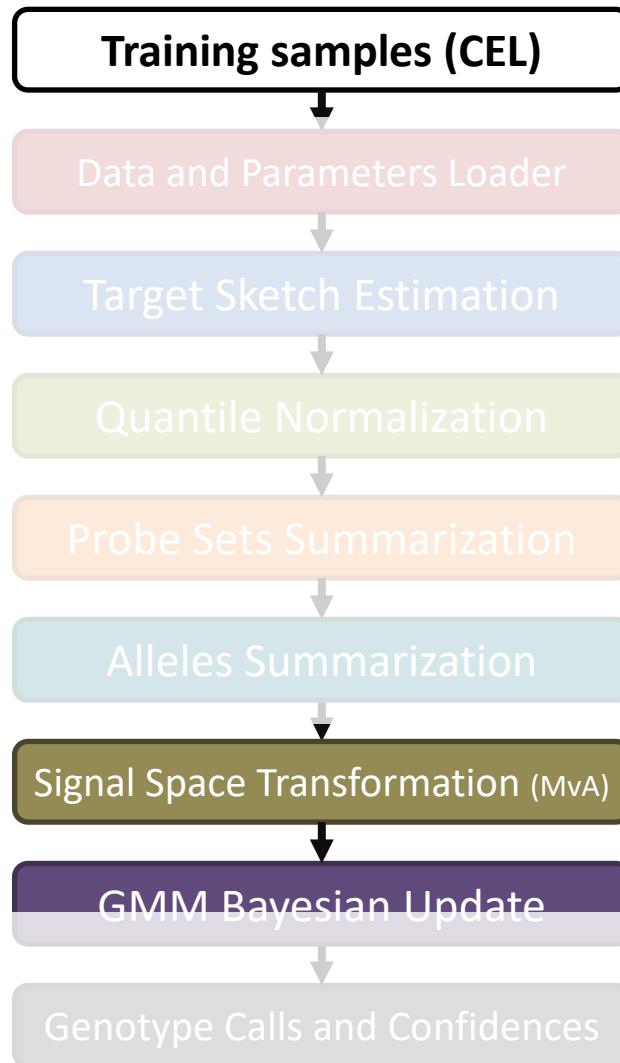


- **Quality Score (Posterior Analog) (for each partition)**
 - **1D (x) Log prior probability of posterior params**
 - $\ell = \log \prod_g N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right) = \sum_g \log\left(N\left(u_{g0}, \frac{\sigma_{g0}^2}{k_{g0}}\right)\right) \Rightarrow -\frac{1}{2}\left[\sum_g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

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

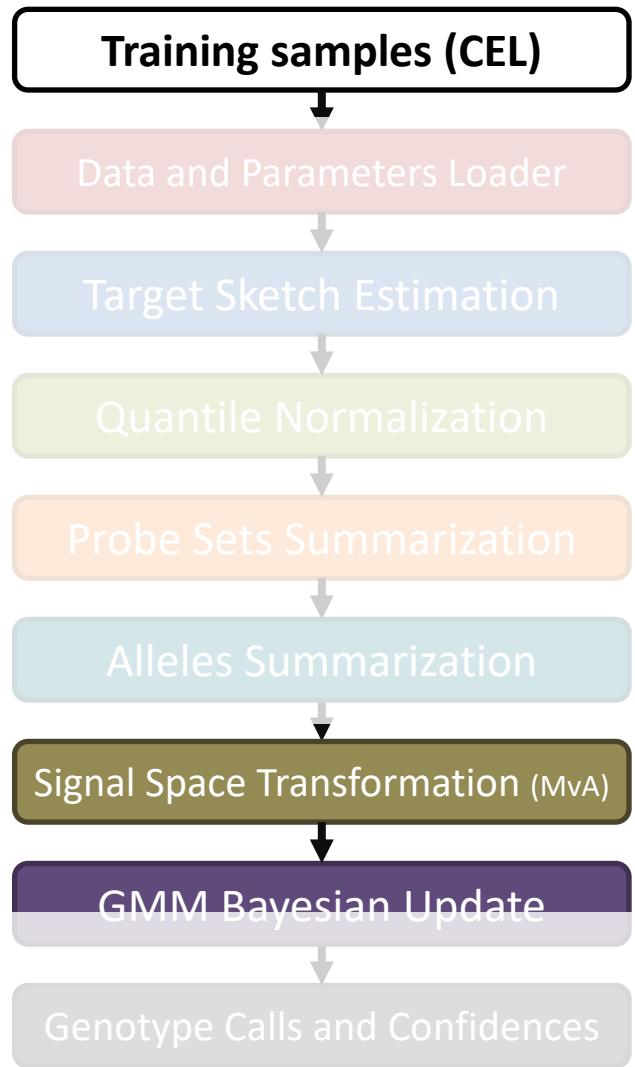


- **Relative Probability for Each Partition under Posterior Information.**
 - Quality Score $Q_{i,j}$ for partition i,j
 - Relative Probability $q_{i,j} = \frac{\exp(-Q_{i,j})}{\exp(-\min Q_{i,j})} = \exp(\min Q_{i,j} - Q_{i,j})$
 $= \frac{\text{Posterior Probability of Specified Partition } (i,j)}{\text{Maximal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

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 $q_{..}$**

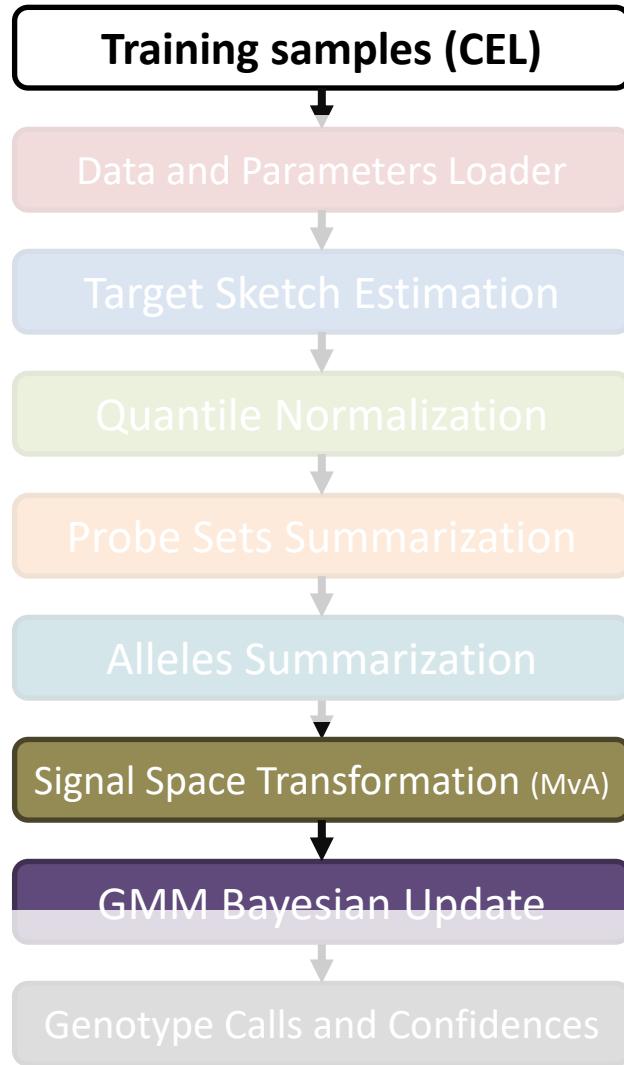
$q_{i,j}$	0	1	2	3	4	$q_{i..}$	$\sum_i q_{i..}$	$q_{..} - \sum_i q_{i..}$
$i \backslash j$	cccc	bccc	bbcc	bbbcc	bbbb	$q_{0..}$	$\sum_{i=0}^4 q_{i..}$	$\sum_{i=1}^4 q_{i..}$
0						$q_{0..}$	$\sum_{i=0}^4 q_{i..}$	$\sum_{i=1}^4 q_{i..}$
1		accc	abcc	abbc	abbb	$q_{1..}$	$\sum_{i=0}^1 q_{i..}$	$\sum_{i=2}^4 q_{i..}$
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

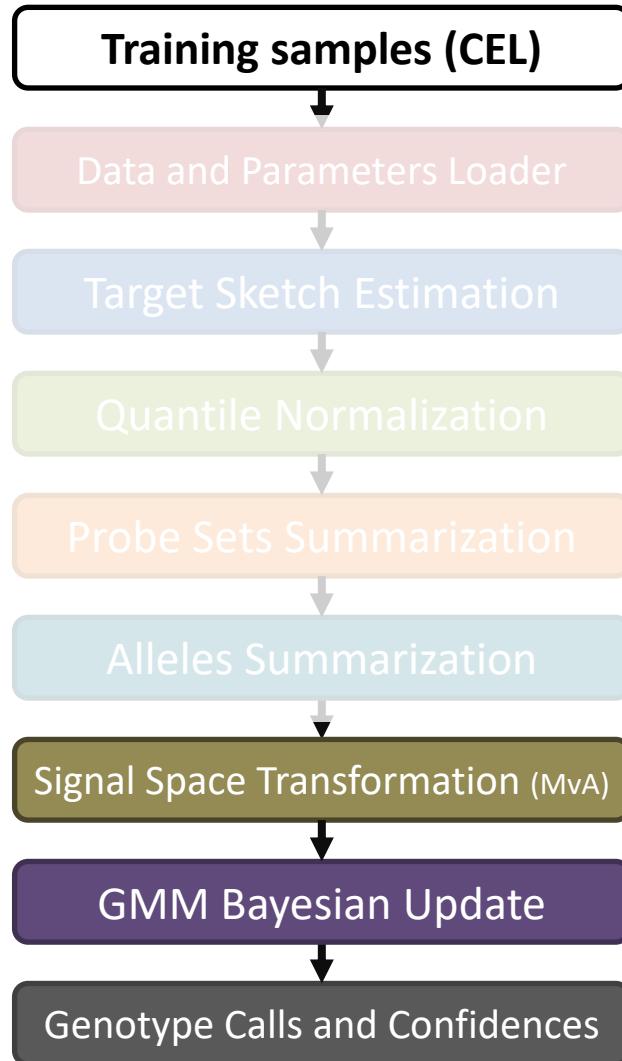


- **Relative Probability for Each Data Point to Be Each Genotype under Posterior Information.**
 - The relative probability for the S_{th} data point being AA genotype: $\frac{\sum_{j=0}^S q_{\cdot j}}{q_{..}}$
 - The relative probability for the S_{th} data point being BB genotype: $1 - \frac{\sum_{i=0}^S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

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

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

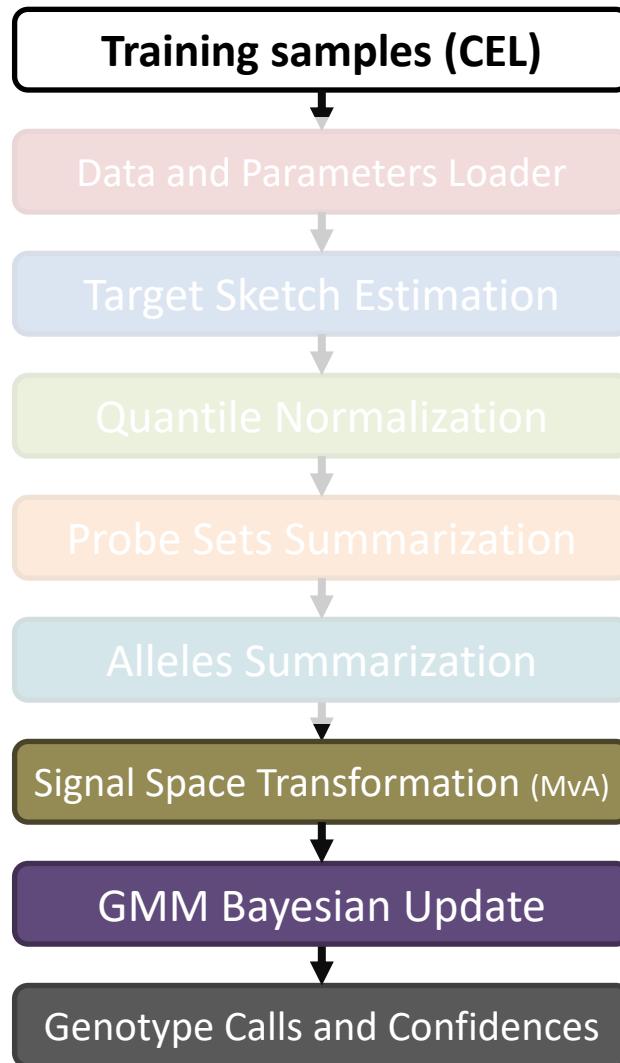
$$- \quad \mathbf{N}_{6x6} = \begin{bmatrix} \sum_i p_{1i} & \mathbf{0} & & & & \mathbf{0} \\ \mathbf{0} & \sum_i p_{1i} & & & & \\ & & \sum_i p_{2i} & & & \\ & & & \sum_i p_{2i} & & \\ \vdots & & & & \ddots & \\ \mathbf{0} & & & & & \mathbf{0} \end{bmatrix}$$

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

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

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

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

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

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

Pool Adjacent-Violators (PAV) algo.

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

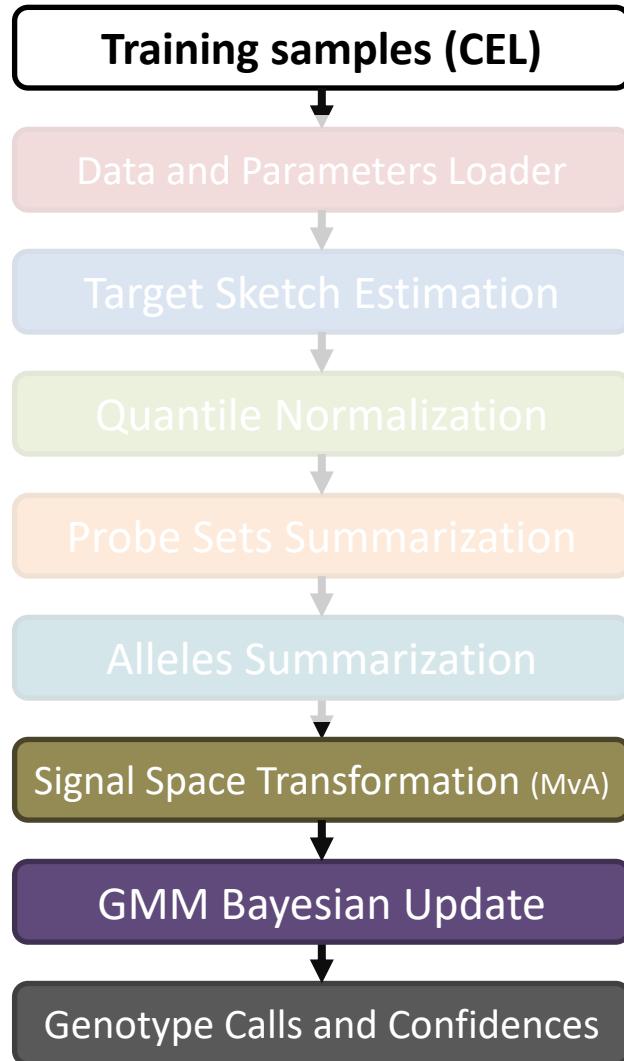
$$g = 1, 2, 3$$

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

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.

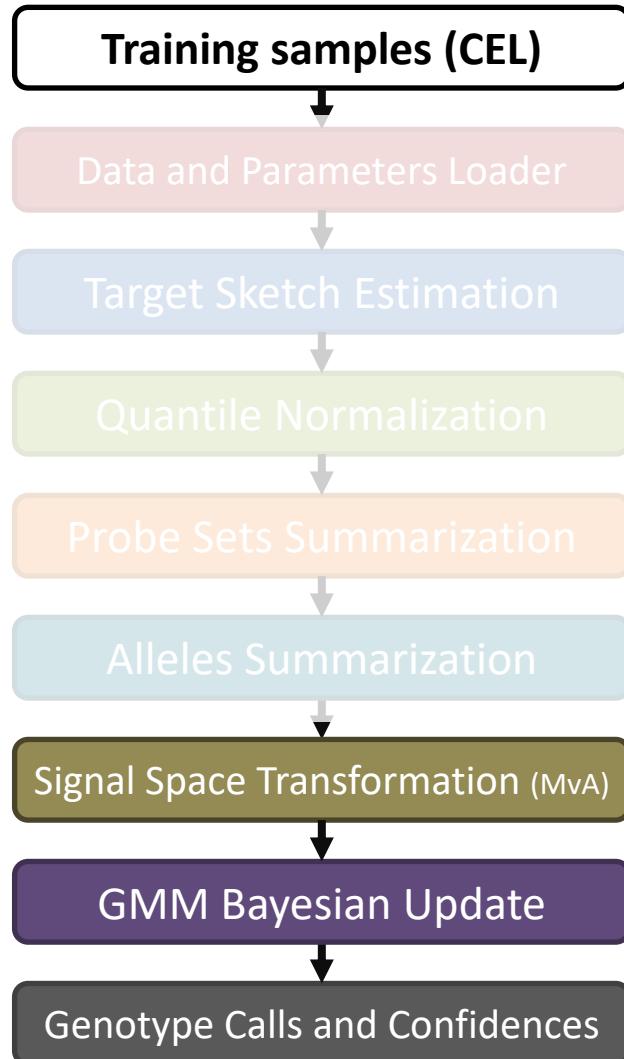
Genotyping Analysis Development and Distribution



- Update 2D Data model parameters with soft assignment of each data bin.
 - $u'_{6x1} = (\mathbf{K}_0^{-1}_{6x6} + N'_{6x6})^{-1} * (\mathbf{K}_0^{-1}_{6x6} * u_{0,6x1} + m_{6x1}),$
 - $k'_g = k_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $u'_{x,3x1}$
 - $v'_g = v_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
 - $\sigma'^2_{xxg} \Rightarrow \frac{v_{g0} * \sigma^2_{xxg0} + (\sum_i p_{gi} x_{gi}^2 - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} x_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{xg} - u_{xg0})^2}{v'_g}$
 - $\sigma'^2_{yyg} \Rightarrow \frac{v_{g0} * \sigma^2_{yyg0} + (\sum_i p_{gi} y_{gi}^2 - \sum_i p_{gi} y_{gi} * \sum_i p_{gi} y_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{yg} - u_{yg0})^2}{v'_g}$
 - $\sigma'^2_{xyg} \Rightarrow \frac{v_{g0} * \sigma^2_{xyg0} + (\sum_i p_{gi} x_{gi} y_{gi} - \sum_i p_{gi} x_{gi} * \sum_i p_{gi} y_{gi}) * \frac{1}{\sum_i p_{gi} + 0.001} + \frac{k_{g0} * \sum_i p_{gi}}{k_{g0} + \sum_i p_{gi}} * (u'_{yg} - u_{yg0}) * (u'_{xg} - u_{xg0})}{v'_g}$
 - Ad-hoc shrinkage for $\sigma'^2_{..g}$ of each cluster (controlled by mixing proportion (lambda) (1)).
Adjusted Pooled Variance.
 - $w_g = v_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
 - $\sigma'^2_{xxt} = \frac{\sum_g w_g * \sigma'^2_{xxg}}{\sum_g w_g}, \quad \sigma'^2_{yyt} = \frac{\sum_g w_g * \sigma'^2_{yyg}}{\sum_g w_g}, \quad t = 1, 2, 3,$
 - $\Rightarrow \frac{(3 - 2 * \text{lambda}) * w_t * \sigma'^2_t + \sum_{g \neq t} \text{lambda} * w_g * \sigma'^2_g}{(3 - 2 * \text{lambda}) * w_t + \sum_{g \neq t} \text{lambda} * w_g}$
 - $\sigma'^2_{xyt} = (\sigma'^2_{xxt} * \sigma'^2_{yyt}) * \frac{\sigma'^2_{xyg}}{\sigma'^2_{xxg} * \sigma'^2_{yyg}}, \quad t = 1, 2, 3, \quad g = t, \text{ means before shrinkage adjustment.}$

Centrillion Confidential

Genotyping Analysis Development and Distribution

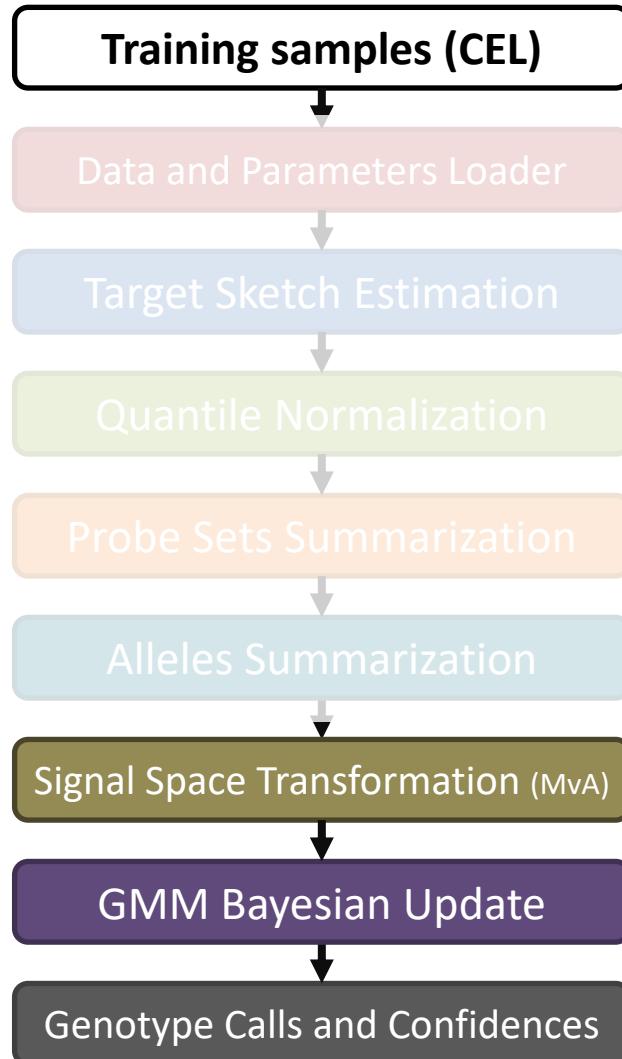


- Update 2D Data model parameters with soft assignment of each data bin
 - $u'_{6x1} = (K_0^{-1}_{6x6} + N'_{6x6})^{-1} * (K_0^{-1}_{6x6} * u_{0,6x1} + m_{6x1}),$
 - $k'_g = k_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
 - (sp.hardshell) Isotonic Regression adjustment for cluster centers (separated by at least sp.shellbarrier (delta, 0.75)). $u'_{x,3x1}$
 - $v'_g = v_{g0} + \sum_i p_{gi}, \quad g = 1, 2, 3$
 - $\sigma'^2_{xxg}, \sigma'^2_{yyg}, \sigma'^2_{xyg}, \quad g = 1, 2, 3$
 - Ad-hoc shrinkage for $\sigma'^2_{..g}$ of each cluster (controlled by mixing proportion (lambda) (1)).
 - $\sigma'_{xx12} = \sigma_{xx120}, \quad \sigma'_{xx13} = \sigma_{xx130}, \quad \sigma'_{xx23} = \sigma_{xx230}$
 $\sigma'_{yy12} = \sigma_{yy120}, \quad \sigma'_{yy13} = \sigma_{yy130}, \quad \sigma'_{yy23} = \sigma_{yy230}$

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



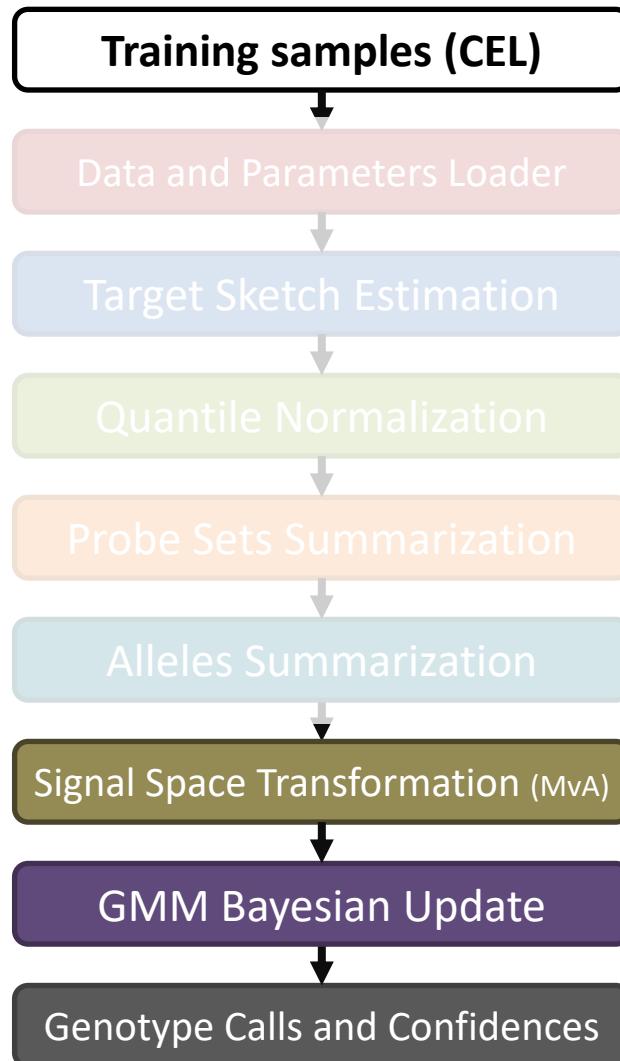
- (sp.mix, freqflag) compute the frequency of each cluster (AA, AB, BB)
 - $f_t = \frac{k'_t}{\sum_g k'_g}, t = 1, 2, 3$
 $\Rightarrow \log f_t = \log k'_t - \log(\sum_g k'_g)$
 $\Rightarrow -\log f_t = -\log k'_t + \log(\sum_g k'_g)$
- For each point, compute the probability that a data point X (x,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

 $\text{inflatePRA} = 0 \text{ (default)}, \text{ocean} = 0.00001 \text{ (default)}$
 - $\log p(X \in t)p(X|X \in t) = \log(p(X \in t)) + \log(p(X|X \in t))$
 - 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))

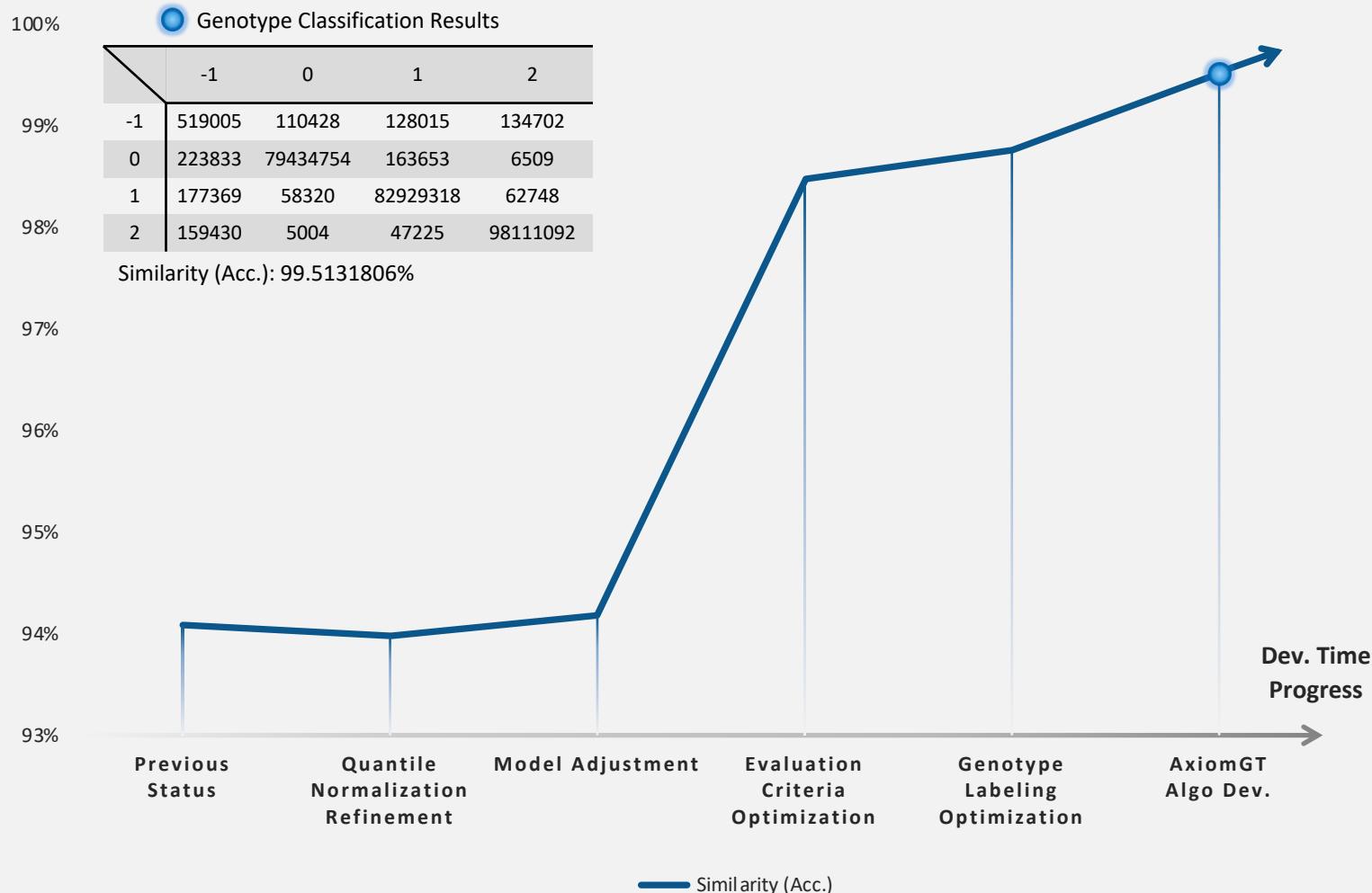
Centrillion Confidential

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	
✖ failed ⌚ 00:17:50 ⌚ 18 hours ago	Bugfix for gender inputs, and remove redundant code. #4901 ↗ hunterize -o 476c692c 🐛 latest		✖	C ⋮
✖ failed ⌚ 00:18:00 ⌚ 2 days ago	Fix the I/O #4900 ↗			D ↑ ↓
✖ failed ⌚ 00:18:05 ⌚ 2 days ago	Update th #4899 ↗			
✔ passed ⌚ 01:26:46 ⌚ 1 week ago	Merge br #4898 ↗			Win_CI New issue
✔ passed ⌚ 01:11:00	Merge br #4897 ↗			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.
				✖ Pipeline #4901 for hunterize ↗ build
				→ ✖ Win_CI
				CentOS_CI

All copyright reserved

2024/1/27

```
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
```

Genotyping Analysis Development and Distribution



- Successful deployment and distribution.

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

Centrillion Confidential

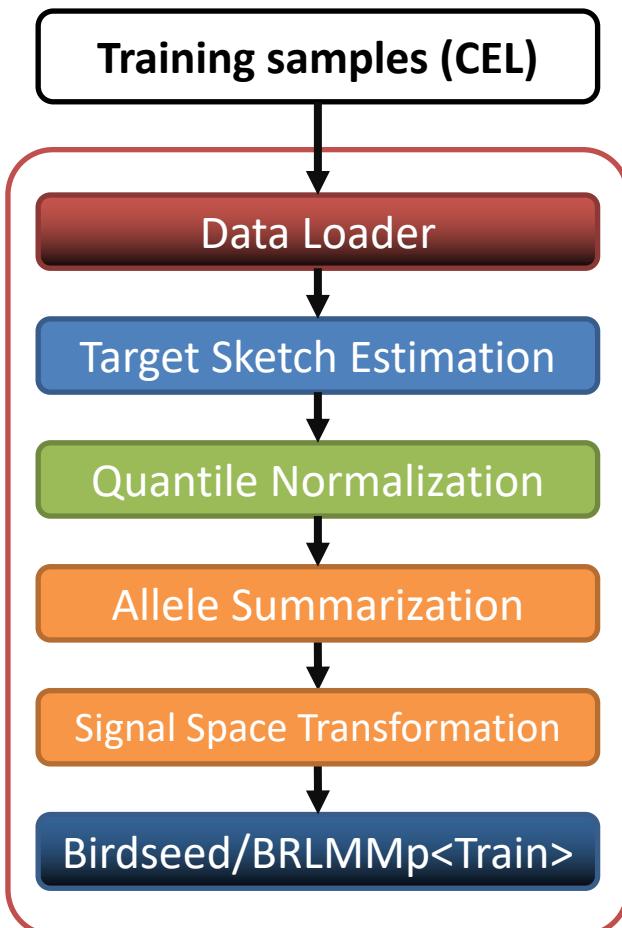
All copyrights and IP belong to Centrillion. For reference only and may not be copied or distributed without written permission from Centrillion. Centrillion shall not be responsible for any party's reliance on these materials.



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}

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

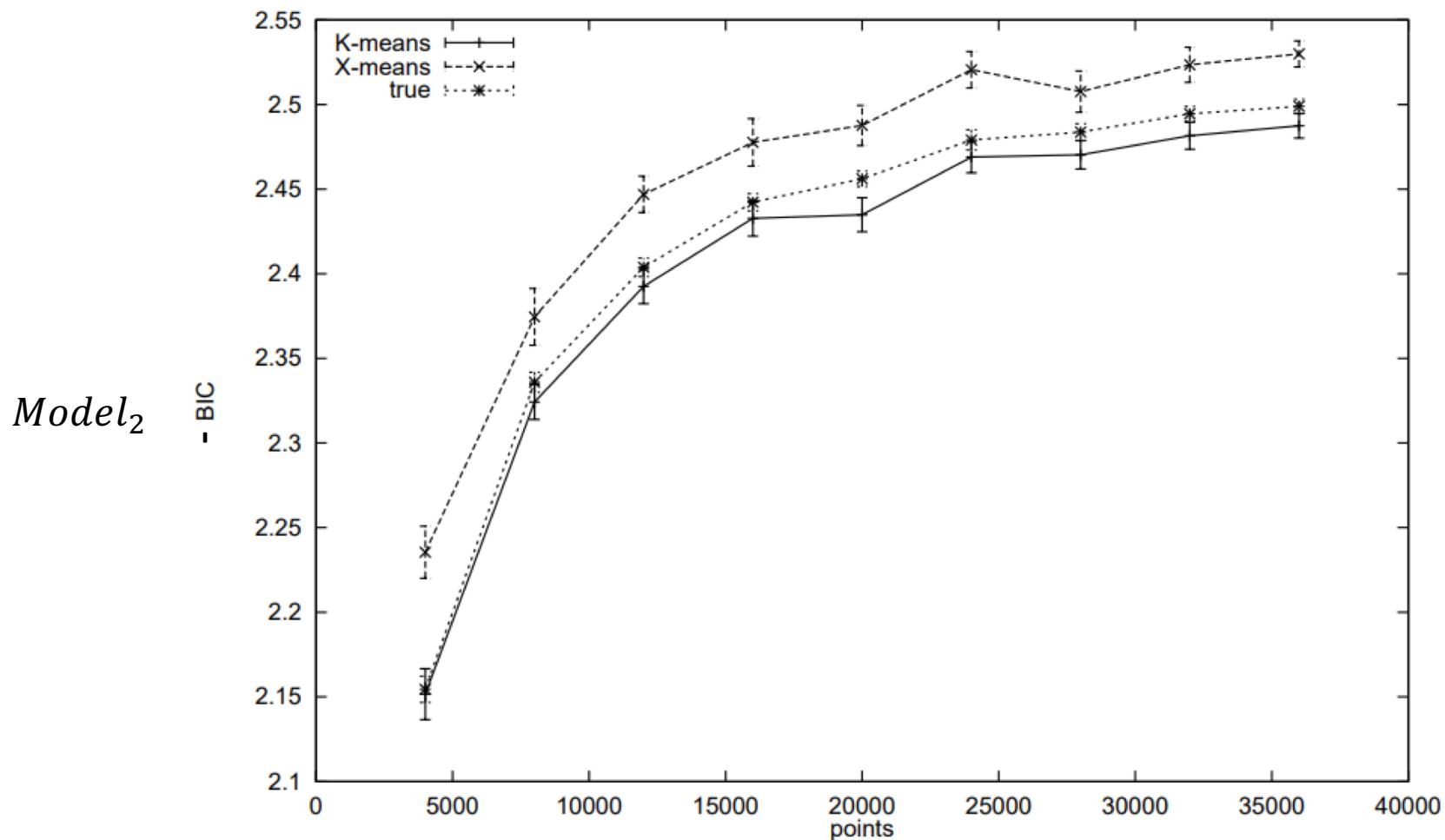
$$BIC = -2 \sum_{j=1}^K N_j \ln(N_j) + 2N \ln(N) + dN \ln(2\pi\hat{\sigma}^2) + dN + \ln(N) \cdot K(d+1),$$

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

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,

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

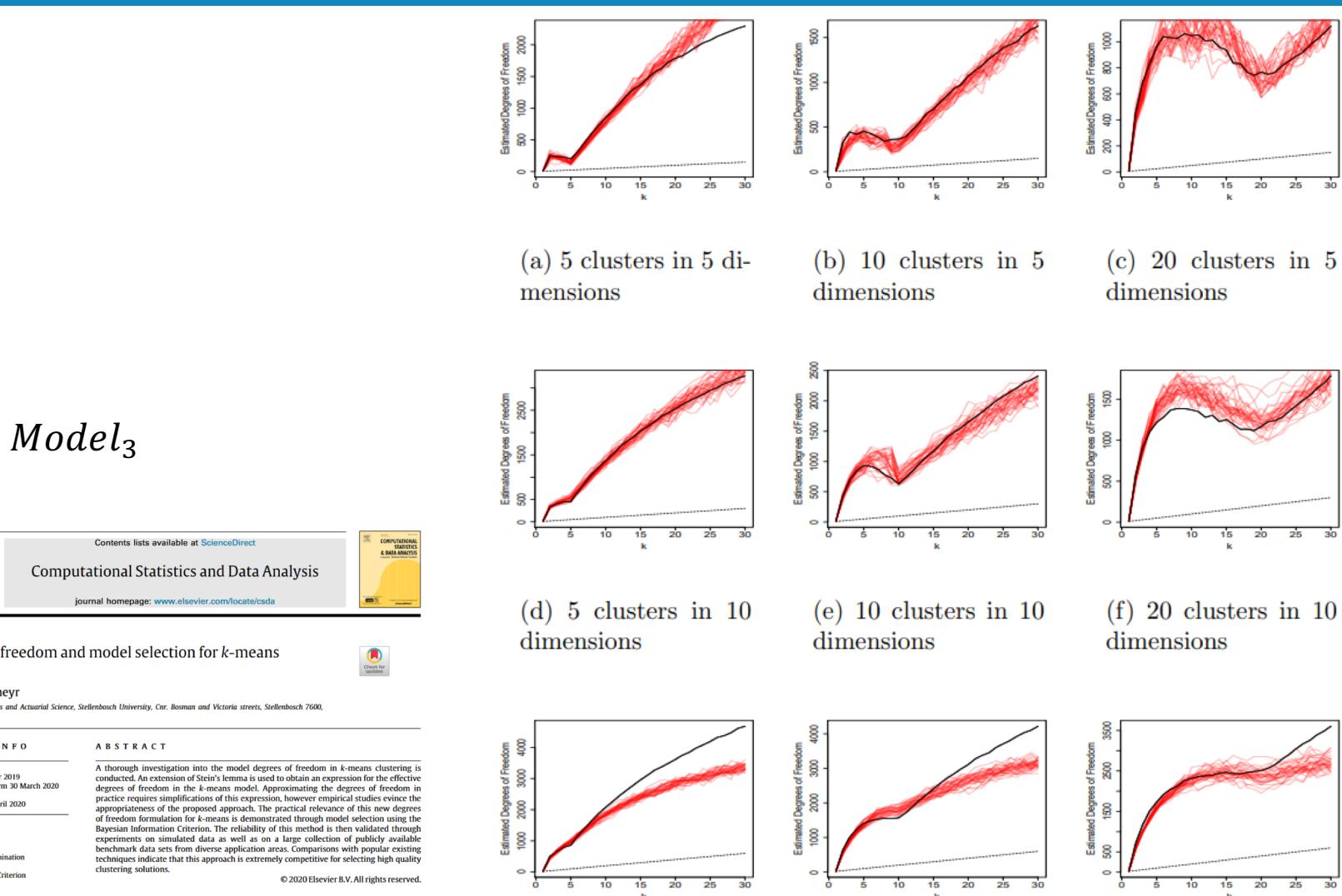
$$c(i) = argmin_{l \in \{1, 2, \dots, K\}} \|X_i - \hat{\boldsymbol{\mu}}_l\|^2,$$

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

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 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.

© 2020 Elsevier B.V. All rights reserved.

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



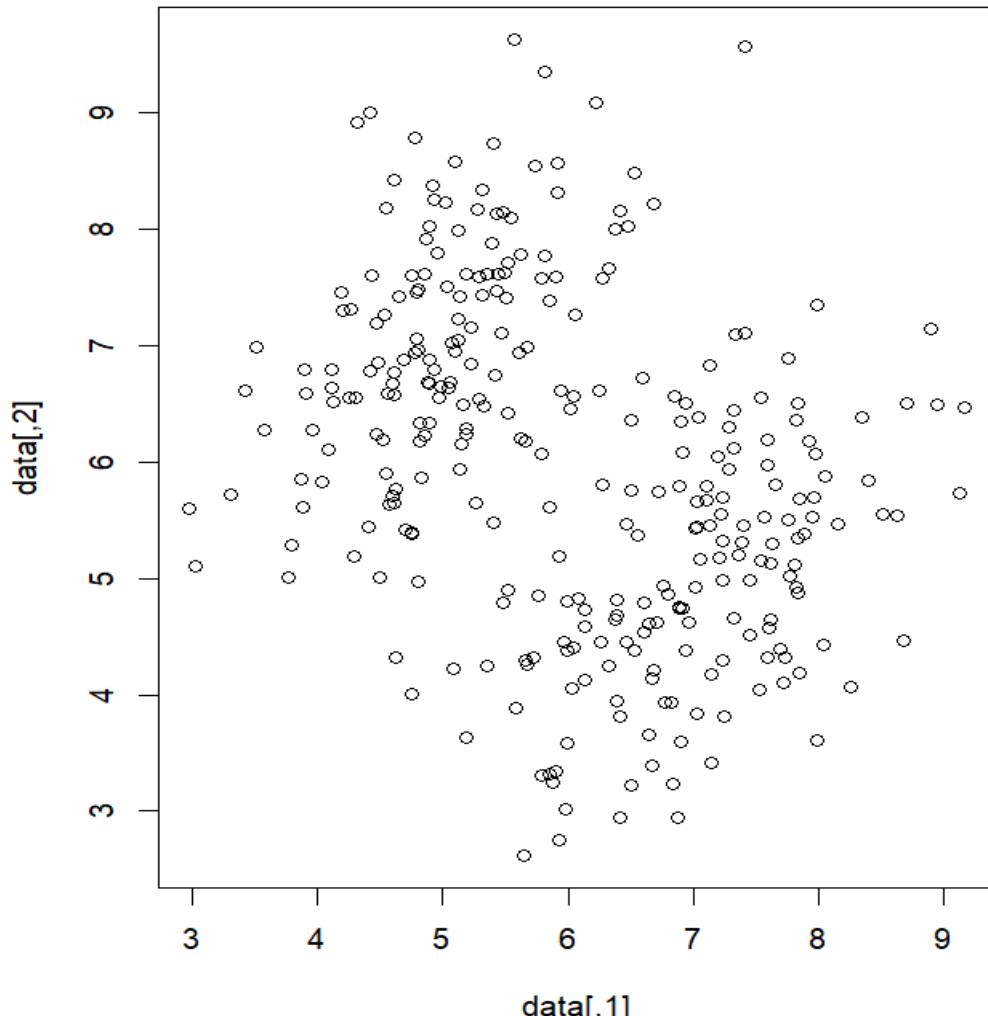
- Simulation

`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: $k = 3$

