Mathematical Formulation: V-Gene–Specific Beta–Binomial Model (Direct α/β Parameterization)

We model sequencing **error counts** y_i out of n_i trials (coverage) in the 260 bp FR1–3 region, grouped by V-gene.

---

1. Data and indexing

- $i = 1, \ldots, N$: index of an observation (e.g., position or small window)
- $g_i \in \{1, \ldots, G\}$: V-gene group index for observation i
- $y_i \in \{0, 1, \ldots, n_i\}$: observed error count
- $n_i > 0$: total coverage/trials

---

2. Likelihood

For each observation i belonging to V-gene group g_i :

$$y_i \mid g_i \sim \text{Beta-Binomial}\left(n_i, \alpha_{g_i}, \beta_{g_i}\right)$$

The **Beta–Binomial** probability mass function is:

$$P(y_i \mid n_i, \alpha, \beta) = \binom{n_i}{y_i} \frac{B(y_i + \alpha,\ n_i - y_i + \beta)}{B(\alpha, \beta)}$$

where $B(\cdot, \cdot)$ is the Beta function.

---

3. Parameter definitions

For each V-gene group g : - $\alpha_g > 0$: first Beta shape parameter - $\beta_g > 0$: second Beta shape parameter

The **mean error rate** and **concentration** for group g are:

$$\mu_g = \frac{\alpha_g}{\alpha_g + \beta_g},$$

$$\phi_g = \alpha_g + \beta_g.$$

The **overdispersion metric** is:

$$\rho_g = \frac{1}{1 + \phi_g} \in (0, 1).$$

4. Priors

We assign Gamma priors directly to α_g and β_g :

$$\alpha_g \overset{i.i.d.}{\sim} \text{Gamma}(a_\alpha, b_\alpha),$$
$$\beta_g \overset{i.i.d.}{\sim} \text{Gamma}(a_\beta, b_\beta).$$

Here $a_\alpha, b_\alpha, a_\beta, b_\beta > 0$ are shape and rate hyperparameters (can be set to weakly informative values, e.g., 1).

---

5. Joint model

The joint distribution of all parameters and data is:

$$P(\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \left[\prod_{g=1}^{G} P(\alpha_g) P(\beta_g)\right] \times \prod_{i=1}^{N} P(y_i \mid n_i, \alpha_{g_i}, \beta_{g_i}).$$

---

6. Posterior inference

The posterior distribution is:

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{n}, \mathbf{g}) \propto \left[\prod_{g=1}^{G} P(\alpha_g) P(\beta_g)\right] \times \prod_{i=1}^{N} P(y_i \mid n_i, \alpha_{g_i}, \beta_{g_i}).$$

We use MCMC (e.g., Stan's NUTS sampler) to draw samples from this posterior, yielding estimates and uncertainty intervals for: - μ_g (mean error rate) - ϕ_g (concentration) - ρ_g (overdispersion)

---

7. Summary of model properties

- **Interpretability:** Parameters μ_g and ϕ_g have clear meanings.
- **Numerical stability:** Avoids logit transforms; works directly with positive α_g, β_g .
- **Flexibility:** Gamma priors can encode domain knowledge about likely error rates.