A scalable variational inference approach for increased mixed-model association power(2025) --Nature Genetics

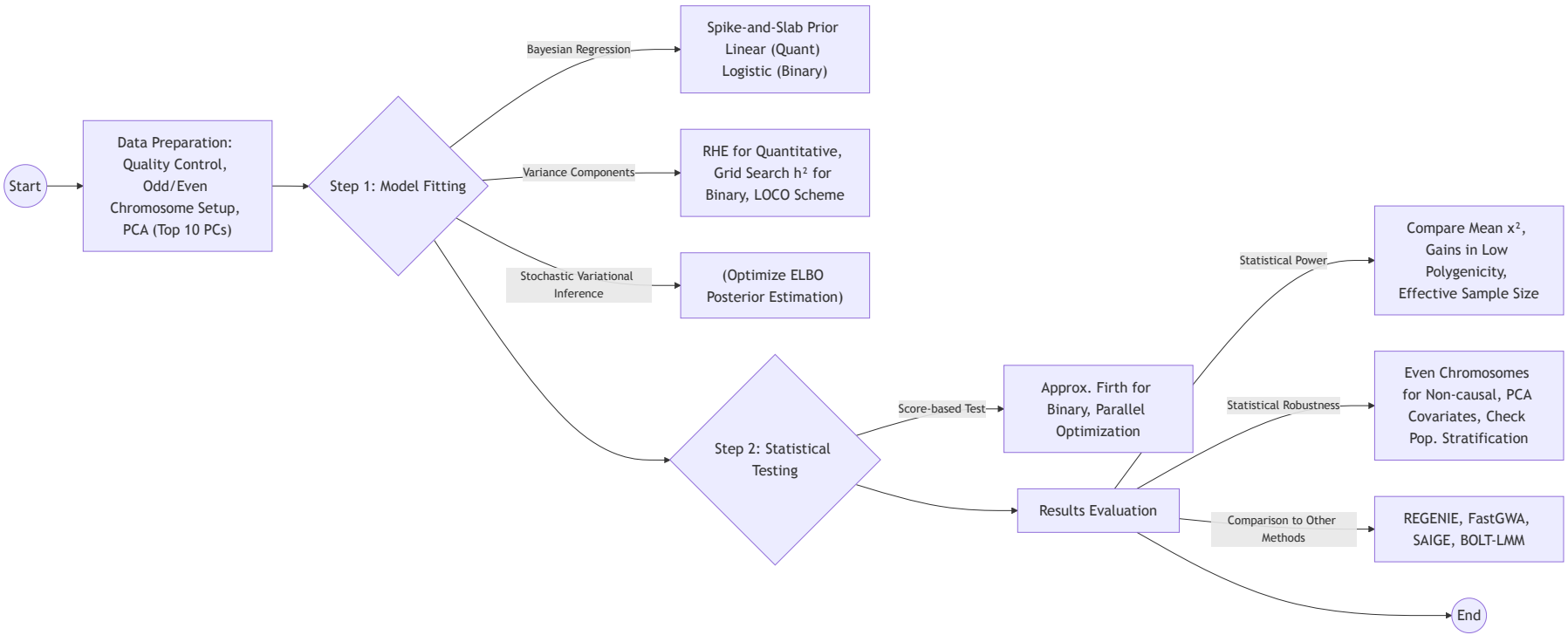
Yang Chen

March 12, 2025

Quickdraws →

Table of contents

Summary of the Quickdraws algorithm



What problem does it solve?

- 📝 **BOLT-LMM** - uses a Bayesian mixture prior to provide state-of-the-art association power but is computationally demanding, especially for multiple traits, and has limited applicability to binary traits
- 🎨 **SAIGE, FastGWA** - rely on **modeling approximations**, such as sparse genetic matrices, which account only for close genetic relationships.
- 👷 **REGENIE** - rely on modeling approximations, such as **block-wise ridge regression**, which assumes that genetic effects are normally distributed.
- 🛠️ **Current GWAS methods** - either highly scalable and resource efficient or highly powered, but not both.
- 🛋️ **Quickdraws** - uses machine learning to simultaneously achieve state-of-the-art association power and computational efficiency for both quantitative and binary traits. Quickdraws uses a Bayesian regression model with **a spike-and-slab prior** on variant effects, efficiently trained using **stochastic variational inference, transfer learning and graphics processing unit (GPU)** matrix operations.

What problem does it solve?

- 📝 **BOLT-LMM** - uses a Bayesian mixture prior to provide state-of-the-art association power but is computationally demanding, especially for multiple traits, and has limited applicability to binary traits
- 🎨 **SAIGE, FastGWA** - rely on **modeling approximations**, such as sparse genetic matrices, which account only for close genetic relationships.
- 👷 **REGENIE** - rely on modeling approximations, such as **block-wise ridge regression**, which assumes that genetic effects are normally distributed.
- 🛠️ **Current GWAS methods** - either highly scalable and resource efficient or highly powered, but not both.
- 🛋️ **Quickdraws** - uses machine learning to simultaneously achieve state-of-the-art association power and computational efficiency for both quantitative and binary traits. Quickdraws uses a Bayesian regression model with **a spike-and-slab prior** on variant effects, efficiently trained using **stochastic variational inference, transfer learning and graphics processing unit (GPU)** matrix operations.

Linear mixed Model

Linear mixed models (LMMs) extend linear models to analyze non-independent data, particularly useful in genomic data analysis involving population structure or cryptic relatedness.

$$y = x_{test}\beta_{test} + \alpha C + g + \epsilon$$

- The genetic and environmental effects are modeled as:

$$g = \beta_{GRM} X_{GRM}$$
$$\epsilon \sim \mathcal{N}(0, \sigma_e^2)$$

where X_{GRM} is the standardized genotype matrix and σ_e^2 is the environmental variance.

Genetic Relatedness Matrix

The random genetic effect g can be considered as a **sample** from $\mathcal{N}(0, \sigma_g^2 K)$, where:

$$K = \frac{1}{M} X_{GRM} X_{GRM}^T$$

- Where $ K $ is the genetic relatedness matrix (GRM)

Step 1: Model Fitting - Bayesian Regression Overview

- **Evidence Lower Bound (ELBO)**

Since directly minimizing the KL divergence is often intractable, variational inference instead **maximizes the Evidence Lower Bound (ELBO)**:

$$\log(P(X)) \geq \mathbb{E}\left[\log(P(X \mid \theta)\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

The ELBO $(P(X))$ can be written as:

$$\mathcal{L}_{VI}(\omega) = \mathbb{E}\left[\log(P(X \mid \theta)\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

In the context of Bayesian regression, the ELBO is:

$$\mathcal{L}_{VI}(\omega) = \mathbb{E}\left[\log(P(\mathbf{y} \mid \theta, \mathbf{X})\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

Step 1: Model Fitting - Bayesian Regression Overview

- **Variational Inference in Bayesian Models**

Variational inference is a technique used to **approximate posterior distributions** when exact inference is computationally intractable. It formulates posterior inference as an **optimization problem**.

The Kullback-Leibler (KL) divergence between the approximate posterior $q_\omega(\theta)$ and the true posterior $ P(\theta \mid X) $ is given by:

$$KL\left(q_\omega(\theta) \| P(\theta \mid X)\right) = \int q_\omega(\theta) \log\left(\frac{q_\omega(\theta)}{P(\theta \mid X)}\right) d\theta$$

The goal of variational inference is to find the parameters $ \omega^* $ that minimize this KL divergence:

$$\omega^* = \arg\min_\omega KL\left(q_\omega(\theta) \| P(\theta \mid X)\right)$$

Step 1: Model Fitting - Bayesian Regression Overview

- **Evidence Lower Bound (ELBO)**

Since directly minimizing the KL divergence is often intractable, variational inference instead **maximizes the Evidence Lower Bound (ELBO)**:

$$\log(P(X)) \geq \mathbb{E}\left[\log(P(X \mid \theta)\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

The ELBO $(P(X))$ can be written as:

$$\mathcal{L}_{VI}(\omega) = \mathbb{E}\left[\log(P(X \mid \theta)\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

In the context of Bayesian regression, the ELBO is:

$$\mathcal{L}_{VI}(\omega) = \mathbb{E}\left[\log(P(\mathbf{y} \mid \theta, \mathbf{X})\right] - KL\left(q_\omega(\theta) \| P(\theta)\right)$$

Step 1: Quantitative Traits Model

- **Evidence Lower Bound (ELBO) Optimization:**

$$L_{\mathrm{VI}}^{Q}(\boldsymbol{\psi}, \boldsymbol{\mu}, \sigma) \approx -\frac{1}{B} \sum_{b=1}^{B} \left(\sum_{s=1}^{S} \frac{(\mathbf{y}_b - X_b \boldsymbol{\beta}(s))^2}{2\sigma_e^2} \right)$$

$$+\frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right).$$

  where $\beta(s)$ are effect estimates sampled from the approximate posterior distribution $q(\beta)$, and ψ_j, μ_j, σ_j are variational parameters.

Optimization Approach To reduce the variance of this objective function and accelerate convergence, we employ:

- **Local reparameterization trick**: This technique helps in *decreasing the variance of gradient estimates*.
- **Antithetic variates**: By generating pairs of negatively correlated random samples, this method further reduces variance.

Step 1: Quantitative Traits Model

- **Evidence Lower Bound (ELBO) Optimization:**

$$L_{\mathrm{VI}}^{Q}(\boldsymbol{\psi}, \boldsymbol{\mu}, \sigma) \approx -\frac{1}{B} \sum_{b=1}^{B} \left(\sum_{s=1}^{S} \frac{(\mathbf{y}_b - X_b \boldsymbol{\beta}(s))^2}{2\sigma_e^2} \right)$$

$$+\frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right).$$

  where $\beta(s)$ are effect estimates sampled from the approximate posterior distribution $q(\beta)$, and ψ_j, μ_j, σ_j are variational parameters.

Optimization Approach To reduce the variance of this objective function and accelerate convergence, we employ:

- **Local reparameterization trick**: This technique helps in *decreasing the variance of gradient estimates*.
- **Antithetic variates**: By generating pairs of negatively correlated random samples, this method further reduces variance.

Step 1: Binary Traits Model

- **Evidence Lower Bound (ELBO) for Binary Traits:**

$$L_{\mathrm{VI}}^{\mathrm{Bi}}(\boldsymbol{\psi}, \boldsymbol{\mu}, \sigma) \approx -\frac{1}{B} \sum_{b=1}^{B} \left(\sum_{s=1}^{S} \mathbf{y}_b \log(\sigma(X_b \boldsymbol{\beta}(s))) + (1 - \mathbf{y}_b) \log(1 - \sigma(X_b \boldsymbol{\beta}(s))) \right)$$

$$+ \frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right),$$

- **Optimization Method**:
  - Optimized using **stochastic variational inference (SVI)**.
  - Enables scalable parameter updates via mini-batch sampling and gradient-based methods.

Step 1: Binary Traits Model

- **Evidence Lower Bound (ELBO) for Binary Traits:**

$$L_{\mathrm{VI}}^{\mathrm{Bi}}(\boldsymbol{\psi}, \boldsymbol{\mu}, \sigma) \approx -\frac{1}{B} \sum_{b=1}^{B} \left(\sum_{s=1}^{S} \mathbf{y}_b \log(\sigma(X_b \boldsymbol{\beta}(s))) + (1 - \mathbf{y}_b) \log(1 - \sigma(X_b \boldsymbol{\beta}(s))) \right)$$

$$+ \frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right),$$

- **Optimization Method**:
  - Optimized using **stochastic variational inference (SVI)**.
  - Enables scalable parameter updates via mini-batch sampling and gradient-based methods.

variance components

- **Genetic and environmental variance components:**
  - For quantitative traits, Quickdraws estimates genetic and environmental variance components using **randomized Haseman-Elston regression (RHE)**, allows the simultaneous estimation of variance components for **multiple traits**.
  - For binary traits, Quickdraws performed **a grid search** over a set of heritability values, $h^2 \in \{0.01, 0.25, 0.5, 0.75\}$,running the Bayesian regression for each value and selecting the heritability corresponding to the highest likelihood.
- **A leave-one-chromosome-out (LOCO) scheme is used for phenotype prediction**: Due to **Proximal Contamination**, Construct X_{GRM} using **all variants except those on the same chromosome** as the test variant.
- **Comparison to other methods:**
  - Similar to BOLT-LMM, Quickdraws uses a Gaussian prior for variant effects but focuses on **nonpolygenic trait** architectures by using a spike-and-slab prior. Other scalable methods like REGENIE and FastGWA rely on fully polygenic trait assumptions.

Step 2: Testing - Overview

Technical advancements

- **Scalability:**
  - Quickdraws incorporates stochastic variational inference and uses first-order optimizers for **linear scaling with sample size**, whereas methods like BOLT-LMM require $O(N^{1.5})$ computation.
- **GPU Optimization:**
  - By offloading matrix multiplications and gradient evaluations to GPUs, Quickdraws achieves substantial speedups over CPU-based computation.
- **LOCO Acceleration:**
  - Quickdraws accelerates LOCO by initializing effect estimates from the whole-genome model, improving speed without compromising accuracy.

Step 2: Testing - Overview

Quantitative Traits Testing Model

$$\mathbf{y} = C\boldsymbol{\alpha} + \mathbf{x}_{\text{test}}\beta_{\text{test}} + g + \epsilon,$$

The goal is to compute the χ^2 association statistic:

$$\frac{\left(\mathbf{x}_{\text{test}}^T \hat{V}^{-1}\mathbf{y}\right)^2}{\mathbf{x}_{\text{test}}^T \hat{V}^{-1}\mathbf{x}_{\text{test}}} \sim \chi_1^2.$$

Where $\hat{V}$ is the estimated variance matrix, X_{GRM} is the genotype matrix used for model fitting, defined as:

$$\hat{V} = \frac{\hat{\sigma}_g^2}{M} X_{GRM} X_{GRM}^T + \hat{\sigma}_e^2 I_N$$

To reduce computational costs, according to **GRAMMAR-Gamma**, we approximate the test statistic using the residual phenotype $\tilde{\mathbf{y}}_{\text{LOCO}}$ from Bayesian regression:

$$\chi_Q^2 \propto \frac{(\mathbf{x}_{\text{test}}^T \tilde{\mathbf{y}}_{\text{LOCO}})^2}{\mathbf{x}_{\text{test}}^T \mathbf{x}_{\text{test}}}.$$

Step 2: Testing - Overview

Binary Traits Testing Model

$$\text{logit}(p_i) = C\boldsymbol{\alpha} + \mathbf{x}_{\text{test}}\beta_{\text{test}} + g + \epsilon,$$

where $p_i = P(y_i = 1|\mathbf{x}_{\text{test}}, g, C)$ is the probability of the ith individual being a case. The score test statistic for $H_0 : \beta_{\text{test}} = 0$ is:

$$T = \frac{\mathbf{x}_{\text{test}}^T(\mathbf{y} - \hat{\mathbf{p}})}{\sqrt{\mathbf{x}_{\text{test}}^T\hat{\rho}\mathbf{x}_{\text{test}}}},$$

with $\hat{\rho} = \hat{V}^{-1} - \hat{V}^{-1}C(C^T\hat{V}^{-1}C)^{-1}C^T\hat{V}^{-1}$, $\hat{V} = \frac{\hat{\sigma}_g^2}{M}X_{\text{GRM}}X_{\text{GRM}}^T + \hat{W}^{-1}$, and $\hat{W} = \text{diag}\{\hat{p}(1-\hat{p})\}$. This can be approximated as:

$$T_B \propto \frac{\mathbf{x}_{\text{test}}^T(\mathbf{y} - \hat{\mathbf{p}})}{\sqrt{\mathbf{x}_{\text{test}}^T\hat{W}\mathbf{x}_{\text{test}}}}.$$

Step 2: Testing - Overview

Calibration of Test Statistics

Calibrate the summary statistics by estimating the effective sample size (ESS) increase compared to running linear regression on a homogeneous subset of unrelated individuals. This involves:

1. **Estimating ESS Increase**: Linked to the use of non-infinitesimal Bayesian linear regression (γ_{blr}).
2. **Estimating ESS Reduction**: Due to the presence of close relatives (γ_{rel}).

The correction term c is computed as:

$$c = \frac{\gamma_{rel}\gamma_{blr}\frac{N_{gd}}{N_{lr}}\left(\langle\chi^2_{lr}\rangle - 1\right) + 1}{\langle\chi^2_{qd}\rangle}$$

Variables Explanation

- $\langle\chi^2_{qd}\rangle$: Mean χ^2 test statistic from Quickdraws.
- $\langle\chi^2_{lr}\rangle$: Mean χ^2 test statistic from linear/logistic regression on a homogeneous unrelated subset.
- $\frac{N_{gd}}{N_{lr}}$: Ratio of total samples to the number of homogeneous unrelated samples used for linear regression.
- γ_{rel} and γ_{blr}: Correction terms for relatedness and Bayesian linear regression usage, respectively.

Step 2: Testing - Overview

Testing Optimizations

- **Adjustment for Case-Control Imbalance:**
  - Quickdraws adjusts for potential instability in score-based test statistics, especially for binary traits
  - Uses **approximate Firth's logistic regression** for rare variants and low-prevalence traits
  - Applied to variants with p-values < 0.05, focusing on rare variants (MAF < 5%) and rare traits (prevalence < 5%)
- **Parallelization and Technical Optimizations:**
  - The calculation of test statistics is optimized using `Numba`
  - Operations are parallelized across **multiple cores** for efficient computation
  - Test statistics calculated by **streaming genotype blocks** for memory efficiency

Step 2: Testing - Overview

Technical advancements

- **Scalability:**
  - Quickdraws incorporates stochastic variational inference and uses first-order optimizers for **linear scaling with sample size**, whereas methods like BOLT-LMM require $O(N^{1.5})$ computation.
- **GPU Optimization:**
  - By offloading matrix multiplications and gradient evaluations to GPUs, Quickdraws achieves substantial speedups over CPU-based computation.
- **LOCO Acceleration:**
  - Quickdraws accelerates LOCO by initializing effect estimates from the whole-genome model, improving speed without compromising accuracy.

Performance in simulated data

To assess the statistical power and robustness of Quickdraws, we performed extensive simulations using 50,000 samples from the UK Biobank dataset, genotyped at 512,828 variants. Among these, 54,568 were rare variants with a minor allele frequency (MAF) between 10^{-4} and 10^{-2}.

Statistical power and robustness

- **Statistical Power:** Simulated **causal variants** were sampled from odd chromosomes to evaluate statistical power.
- **Statistical Robustness: Non-causal variants** from even chromosomes were used to assess robustness to population relatedness and stratification.

Principal component analysis (PCA) was performed using the top ten PCs as covariates for Quickdraws and other models.

Performance in simulated data

To assess the statistical power and robustness of Quickdraws, we performed extensive simulations using 50,000 samples from the UK Biobank dataset, genotyped at 512,828 variants. Among these, 54,568 were rare variants with a minor allele frequency (MAF) between 10^{-4} and 10^{-2}.

Simulation details

- **Heritability:** Simulate 50 heritable traits with a narrow-sense heritability of $h_g^2 = 0.4$.
- **Polygenicity:** The polygenicity (proportion of variants with non-zero effects) ranged from 0.25% to 10%.
- **MAF-dependent Architecture:** The architecture was parameterized with $\alpha = -0.3$, determining the relationship between MAF and effect sizes.

Performance in simulated data

To assess the statistical power and robustness of Quickdraws, we performed extensive simulations using 50,000 samples from the UK Biobank dataset, genotyped at 512,828 variants. Among these, 54,568 were rare variants with a minor allele frequency (MAF) between 10^{-4} and 10^{-2}.

Sample composition

We created three groups of 50,000 samples each, using ancestry and relatedness information:

1. **GB-unrel:** A set of unrelated, self-reported white British individuals.
2. **GB-rel:** A set of white British individuals with more first- to third-degree relatives (3.4× more compared to the full UK Biobank subset).
3. **EUR:** A set of 50% British and 50% non-British European individuals with a similar level of relatedness to the full white British subset.

These groups were used to simulate shared environmental components among close relatives and ancestry-based population stratification.

Performance in simulated data

To assess the statistical power and robustness of Quickdraws, we performed extensive simulations using 50,000 samples from the UK Biobank dataset, genotyped at 512,828 variants. Among these, 54,568 were rare variants with a minor allele frequency (MAF) between 10^{-4} and 10^{-2}.

Statistical power and robustness

- **Statistical Power:** Simulated **causal variants** were sampled from odd chromosomes to evaluate statistical power.
- **Statistical Robustness: Non-causal variants** from even chromosomes were used to assess robustness to population relatedness and stratification.

Principal component analysis (PCA) was performed using the top ten PCs as covariates for Quickdraws and other models.

Statistical power

Testing with non-spike-and-slab distributions

Simulations were also performed with causal effects sampled from **Gaussian, mixture of Gaussians, or Laplace distributions**. In the mixture of Gaussian setting, Quickdraws and BOLT-LMM exhibited higher power than other models. When traits were fully infinitesimal (with Laplace and Gaussian effects), Quickdraws and BOLT-LMM showed results similar to other infinitesimal methods.

- In fully infinitesimal traits, Quickdraws and BOLT-LMM performed similarly to methods assuming infinitesimal trait architecture.

Power evaluation in large datasets

We evaluated Quickdraws on larger datasets with 405,000 white British individuals and 460,000 European individuals from the UK Biobank.

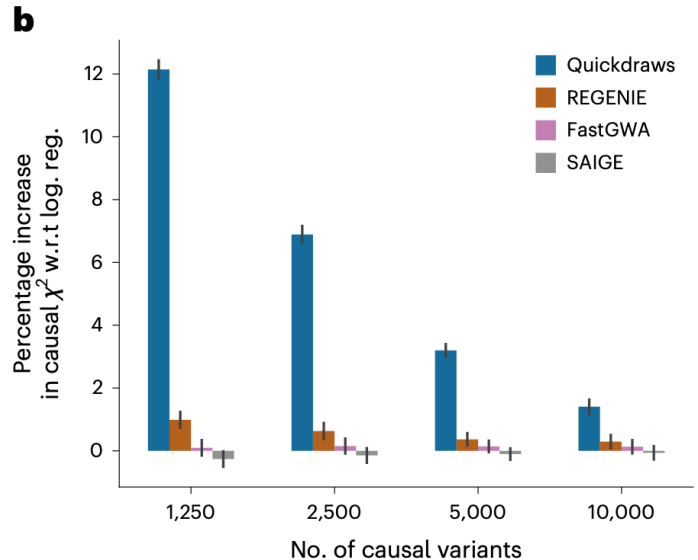
- Quickdraws exhibited the highest association power, surpassing BOLT-LMM, especially when polygenicity was low.

Statistical power

Binary trait association

Phenotypes were simulated under **a liability threshold model**, using a default cutoff of $p < 0.05$, with varying numbers of causal variants.

- Quickdraws outperformed SAIGE, REGENIE, and FastGWA-GLMM in terms of statistical power, particularly for traits with low polygenicity (0.25%).
- For traits with low polygenicity, Quickdraws achieved 11.5% and 12.04% higher average χ^2 compared to REGENIE and FastGWA-GLMM, respectively.



Statistical power

Testing with non-spike-and-slab distributions

Simulations were also performed with causal effects sampled from **Gaussian, mixture of Gaussians, or Laplace distributions**. In the mixture of Gaussian setting, Quickdraws and BOLT-LMM exhibited higher power than other models. When traits were fully infinitesimal (with Laplace and Gaussian effects), Quickdraws and BOLT-LMM showed results similar to other infinitesimal methods.

- In fully infinitesimal traits, Quickdraws and BOLT-LMM performed similarly to methods assuming infinitesimal trait architecture.

Power evaluation in large datasets

We evaluated Quickdraws on larger datasets with 405,000 white British individuals and 460,000 European individuals from the UK Biobank.

- Quickdraws exhibited the highest association power, surpassing BOLT-LMM, especially when polygenicity was low.

Calibration and statistical robustness of Quickdraws

Overall results

- **Statistical Power:** Quickdraws outperformed or matched BOLT-LMM for quantitative traits and achieved higher power than existing methods for binary traits.
- **FPR Control:** Quickdraws provided controlled FPRs across all simulated scenarios, including those with population structure, relatedness, and low-prevalence binary traits.

Calibration and statistical robustness of Quickdraws

Calibration in binary traits

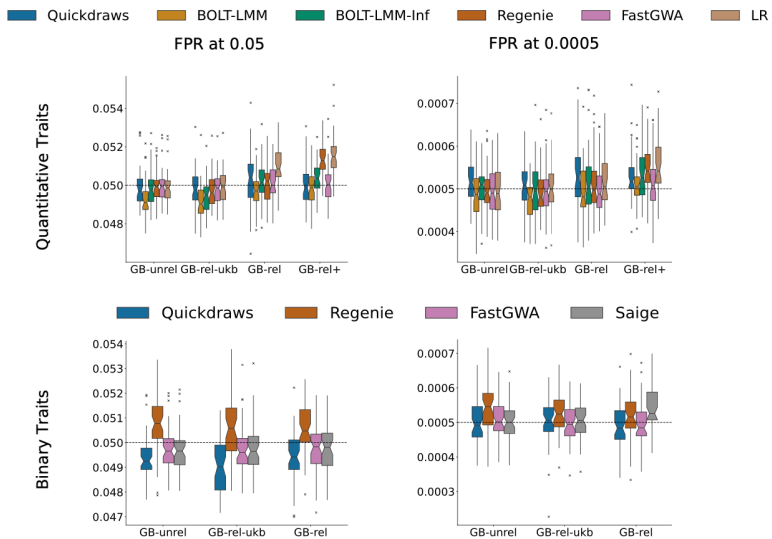
Calibration and statistical robustness of Quickdraws

Robustness across different conditions

Quickdraws was evaluated across varying **levels of population structure**, relatedness, and causal effect-size distributions.

Valuate FPRs in simulations involving all white British individuals (N ≈ 405,000) and all self-identified European individuals from the UK Biobank (N ≈ 460,000), varying levels of polygenicity in quantitative traits (1–10%) and varying levels of prevalence (0.3–0.001) in binary traits.

- Quickdraws maintained controlled FPRs in all tested conditions, while methods like FastGWA and REGENIE showed inflation in some cases due to residual population stratification.



Calibration and statistical robustness of Quickdraws

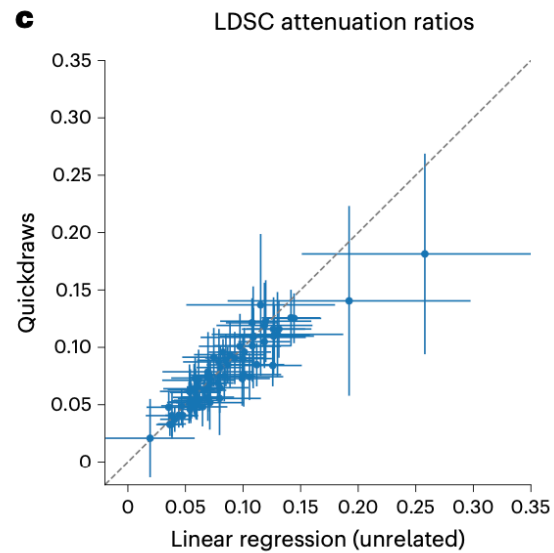
Overall results

- **Statistical Power:** Quickdraws outperformed or matched BOLT-LMM for quantitative traits and achieved higher power than existing methods for binary traits.
- **FPR Control:** Quickdraws provided controlled FPRs across all simulated scenarios, including those with population structure, relatedness, and low-prevalence binary traits.

UK Biobank analysis

Calibration and Functional Annotation

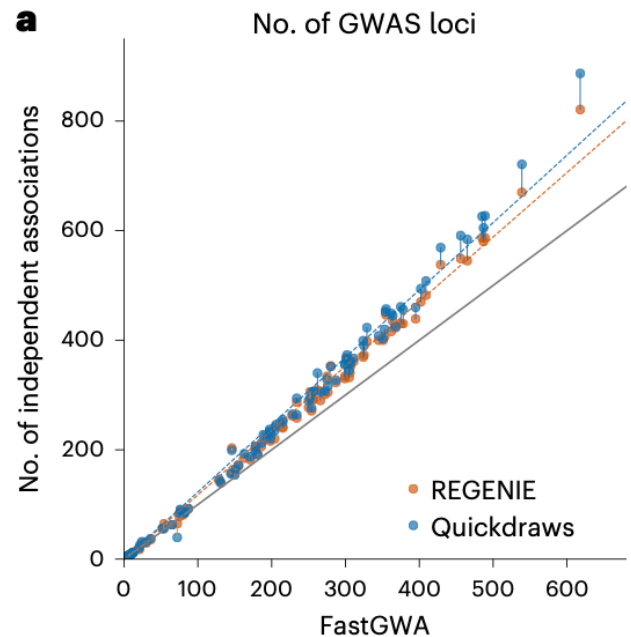
- **Assessed calibration by comparing linkage disequilibrium (LD) score regression attenuation ratios**
  - Quickdraws produced attenuation ratios close to those of linear regression in unrelated British samples
    - Quickdraws: 0.0832 (s.e. = 0.008)
    - Linear regression: 0.0892 (s.e. = 0.008)
  - For low-prevalence binary traits, Quickdraws remained calibrated
    - Did not produce signatures of false-positive associations
- **Evaluated functional annotation of regions with variants found using Quickdraws but not REGENIE**
  - Similar enrichments compared with variants detected using both REGENIE and Quickdraws
  - Indicating no major differences in the functional profile of variants exclusively detected by Quickdraws



UK Biobank analysis

Number of Independent Associations Detected

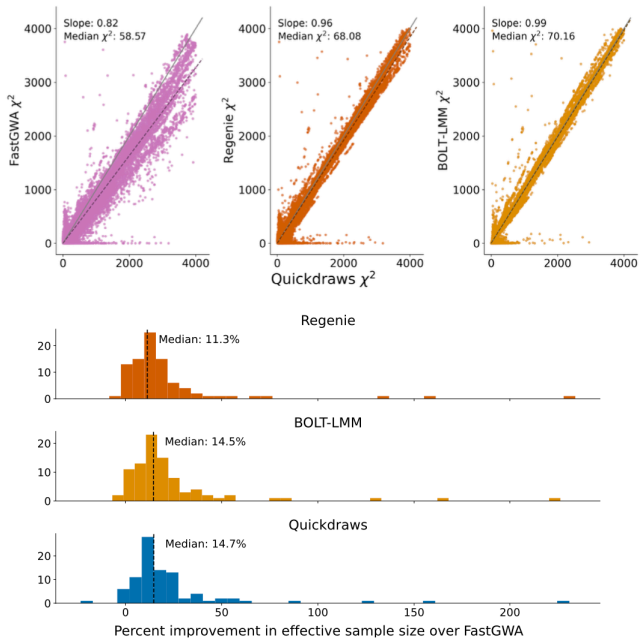
- Quickdraws **detected significantly more independent associations** than REGENIE and FastGWA for both quantitative and disease traits (binomial test $P < 1.9 \times 10^{-3}$)
- For quantitative traits:
  - 4.97% more independent associations than REGENIE
  - 22.71% more independent associations than FastGWA
  - Similar to BOLT-LMM (Quickdraws: 26,236, BOLT-LMM: 26,368)
- For disease traits:
  - 3.25% more independent associations than REGENIE
  - 7.07% more independent associations than FastGWA-GLMM
- Larger gains in traits with high estimated heritability or low polygenicity (e.g., 8.04% increase over REGENIE for mean platelet volume and 26.1% increase for standing height)



UK Biobank analysis

Protein Trait Analysis

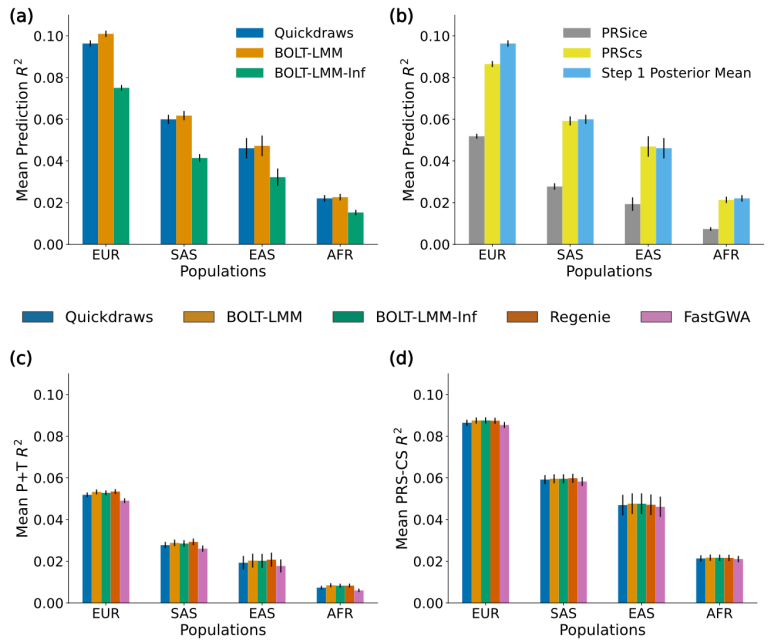
- We analyzed 250 plasma protein traits (N ≈ 43,000)
  - These traits typically have lower polygenicity
  - Quickdraws identified 5.54% more loci than REGENIE (P = 6.6×10^{-3})
- Effective sample size comparison:
  - 14.7% higher than FastGWA (P = 7.6×10^{-4})
  - 3.4% higher than REGENIE (P = 0.197)
  - Similar to BOLT-LMM (P = 0.46)



UK Biobank analysis

Polygenic Prediction Accuracy

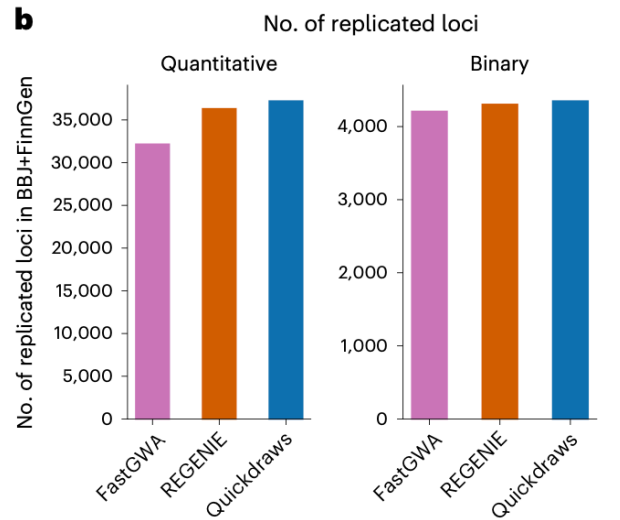
- **Evaluating accuracy of non-infinitesimal modeling:**
  - Used predictors from step 1, trained on 405,000 white British individuals to predict traits for remaining samples
  - Mean correlations across 79 quantitative traits:
    - Quickdraws: 0.307 (s.e. = 0.0061)
    - BOLT-LMM: 0.313 (s.e. = 0.0061)
    - BOLT-LMM-Inf: 0.271 (s.e. = 0.0061), notably lower
  - Consistent with previously observed improvements from non-infinitesimal trait architecture modeling
- **Comparison with polygenic scores (PGS) built using recent methods (PRS-CS and P+T):**
  - Quickdraws' step 1 predictors were significantly more accurate in European held-out set (P <



UK Biobank analysis

Replication Analysis Validation

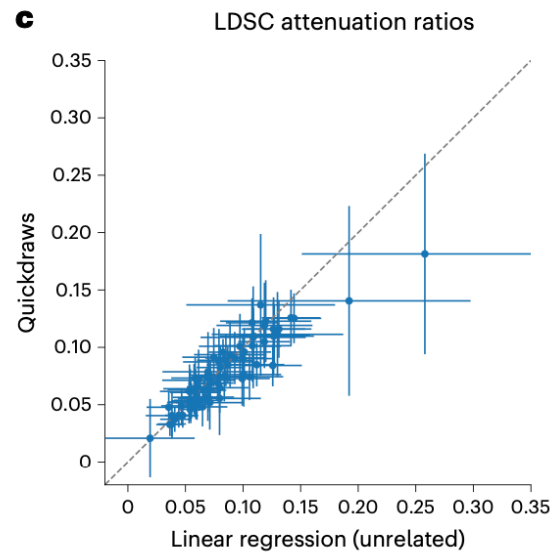
- Replication analysis based on GWAS summary statistics from Biobank Japan, FinnGen, and other large-scale studies
- Across 53 traits (comprising 40 approximately independent traits):
  - Quickdraws yielded a higher number of replicated loci than REGENIE and FastGWA
    - Binomial test P = 0.014 (vs. REGENIE)
    - Binomial test P = 7×10^{-4} (vs. FastGWA)
- For 30 quantitative traits:
  - 2.5% more replicated loci than REGENIE
  - 15.72% more replicated loci than FastGWA
  - Similar to BOLT-LMM (Quickdraws: 37,210, BOLT-LMM: 37,072)
- For 23 disease traits:
  - 1.07% more replicated loci than REGENIE



UK Biobank analysis

Calibration and Functional Annotation

- **Assessed calibration by comparing linkage disequilibrium (LD) score regression attenuation ratios**
  - Quickdraws produced attenuation ratios close to those of linear regression in unrelated British samples
    - Quickdraws: 0.0832 (s.e. = 0.008)
    - Linear regression: 0.0892 (s.e. = 0.008)
  - For low-prevalence binary traits, Quickdraws remained calibrated
    - Did not produce signatures of false-positive associations
- **Evaluated functional annotation of regions with variants found using Quickdraws but not REGENIE**
  - Similar enrichments compared with variants detected using both REGENIE and Quickdraws
  - Indicating no major differences in the functional profile of variants exclusively detected by Quickdraws



Computational costs

Technical Implementation Details

Hardware and Setup

- **Markers:** 458,464 for Quickdraws, REGENIE, and BOLT-LMM; 89,177 for SAIGE
- **Hardware:** Tested on UK Biobank RAP with up to four machines
- **Quickdraws:** Utilizes **GPU acceleration** for Bayesian regression

Memory Management

- **High-Memory Mode:** Loads genotype matrix in memory
- **Low-Memory Mode:** Streams data from disk (20% slower)
- **RAM usage:** Scales with MN/4 + GPU memory
- **Comparison:** REGENIE/FastGWA <15 GB RAM for 405,000 samples

Computational costs

Quantitative Trait Analysis Performance

Table 2 | Computational efficiency for quantitative trait association

Samples	Method	Step 1	Step 2	Total time	Total memory[a]	Cost on RAP[b]
		(h)	(h)	(h)	(GB)	(£)
50,000	Quickdraws	9.6	9	18.6	48 (16)	7.9
	BOLT-LMM	127	671.5	798.5	<15	158.4
	REGENIE	1.1	19.4	20.5	<15	4.0
	FastGWA	–	20.2	20.2	<15	3.6
405,088	Quickdraws	97.7	51.5	149.3	63 (16)	93.0
	BOLT-LMM	1,150	7,250	8,400	46	7,500.0
	REGENIE	4.7	45.3	50.0	<15	24.2
	FastGWA	–	128.4	128.4	<15	37.3

- **Multi-trait efficiency**: Quickdraws particularly efficient for multiple traits
- **GPU acceleration**: Significant speedups for Bayesian regression step

Computational costs

Technical Implementation Details

Hardware and Setup

- **Markers:** 458,464 for Quickdraws, REGENIE, and BOLT-LMM; 89,177 for SAIGE
- **Hardware:** Tested on UK Biobank RAP with up to four machines
- **Quickdraws:** Utilizes **GPU acceleration** for Bayesian regression

Memory Management

- **High-Memory Mode:** Loads genotype matrix in memory
- **Low-Memory Mode:** Streams data from disk (20% slower)
- **RAM usage:** Scales with MN/4 + GPU memory
- **Comparison:** REGENIE/FastGWA <15 GB RAM for 405,000 samples

Performance optimization

Speed Optimization

1. **GPUs for Bayesian Regression**: Use GPUs with `PyTorch`
2. **Transfer Learning**: LOCO models initialized with effect estimates from whole-genome regression, reducing iterations and speeding up model fitting (2x-2.5x faster).
3. **Numba Optimization**: Vectorized linear regression for multiple traits using Numba's `parallel` and `njit`
4. **Test Statistic Calibration**: Calibration using 458k genotyped variants, reducing runtime and ensuring consistent results with imputed variants.
5. **Data Loading with HDF5 and PySnpTools**: Raw genotypes stored in compressed `HDF5` files, with `PySnpTools` used for efficient variant-wise access to bgen/bed files.
6. **Approximate Firth Logistic Regression**: Applied to variants with p-values < 0.05, focusing on rare variants (MAF < 5%) and rare traits (prevalence < 5%).

Performance optimization

Speed Optimization

1. **GPUs for Bayesian Regression**: Use GPUs with `PyTorch`
2. **Transfer Learning**: LOCO models initialized with effect estimates from whole-genome regression, reducing iterations and speeding up model fitting (2x-2.5x faster).
3. **Numba Optimization**: Vectorized linear regression for multiple traits using Numba's `parallel` and `njit`
4. **Test Statistic Calibration**: Calibration using 458k genotyped variants, reducing runtime and ensuring consistent results with imputed variants.
5. **Data Loading with HDF5 and PySnpTools**: Raw genotypes stored in compressed `HDF5` files, with `PySnpTools` used for efficient variant-wise access to bgen/bed files.
6. **Approximate Firth Logistic Regression**: Applied to variants with p-values < 0.05, focusing on rare variants (MAF < 5%) and rare traits (prevalence < 5%).

Discussion

Future Work and Limitations

- **GPU Dependency**: Slower on CPU hardware; future work could optimize for CPU performance
- **Cross-trait Correlations**: Current version doesn't capture these; future development could improve performance
- **External Cohort Integration**: Using PGSs from larger external cohorts could enhance power
- **Participation Bias**: Quickdraws is susceptible; need to integrate adjustment strategies into initial steps
- **Meta-analysis Strategies**: Develop methods to combine posterior effect estimates across cohorts for federated analyses