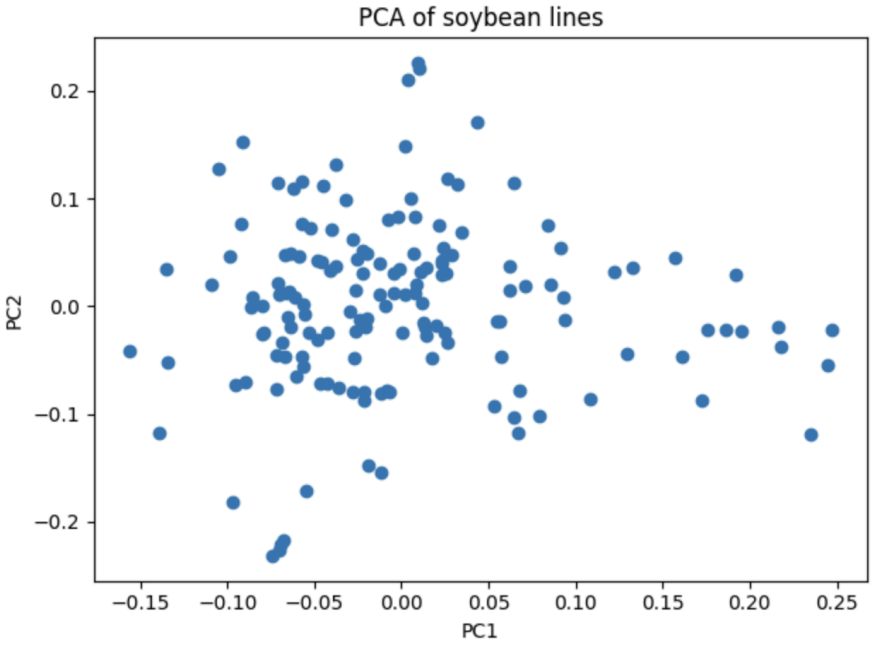
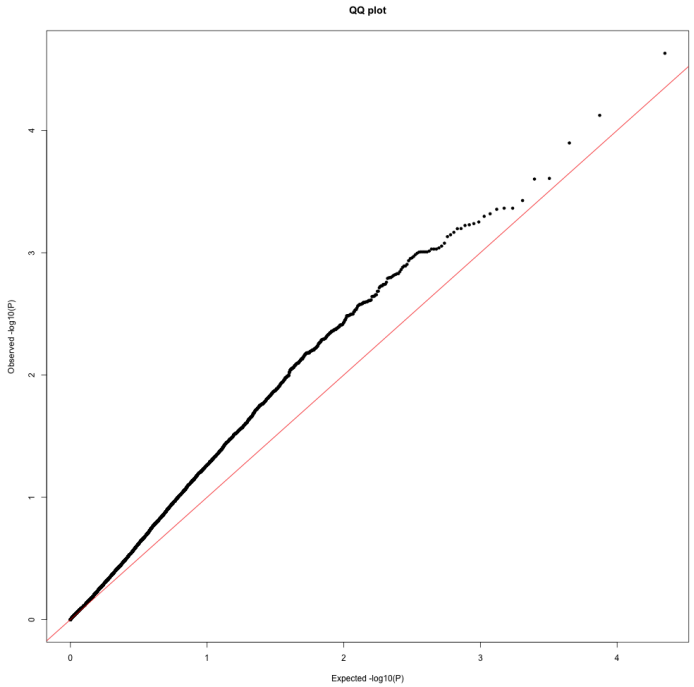
PCA



Проведённая фильтрация VCF-файлов с выбранными порогами дала облако точек без чётких групп. Явного разделения линий на кластеры не наблюдается — большинство образцов образует единый кластер “размазанный по графику”.Возможно, при использовании других фильтров (например, более строгих порогов по пропущенным генотипам или по MAF) структура стала бы чуть яснее, но если например образцы близкородственный, то такая картина PCA может быть характерной.

QQ plot



Точки примерно до уровня ~ 2-2.5 по ожидаемому -log10(P) хорошо лежат на диагональной линии. Это означает, что распределение p-value близко к ожидаемому, и глобальных перекосов нет.

В верхней части графика есть небольшой подъём точек выше линии — такие отклонения обычно соответствуют SNP, которые действительно могут быть связаны с признаком или просто выделяются как «кандидаты».

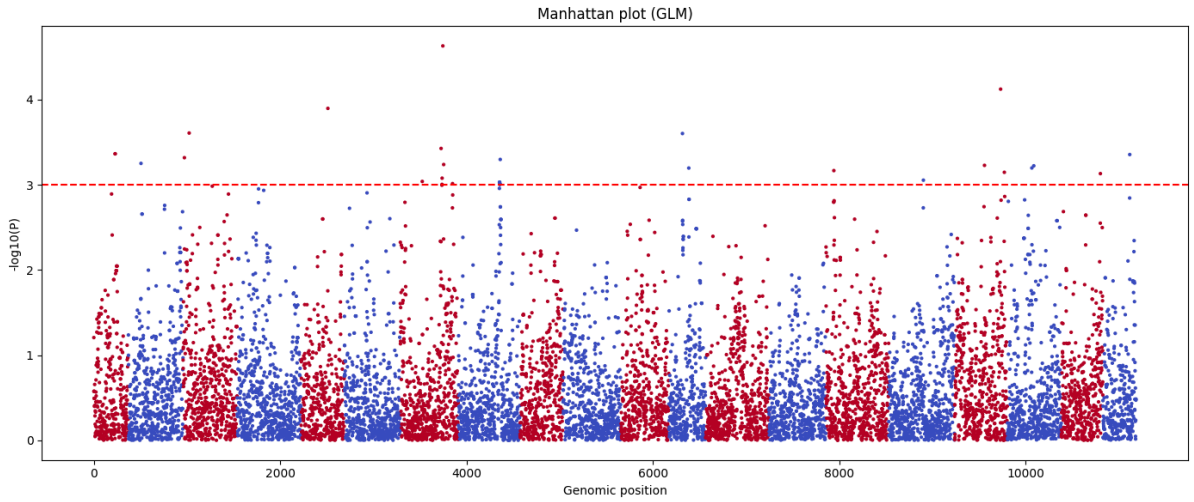
λGC ≈ 1.04 говорит о том, что инфляция p-value минимальна. Инфляция в данном контексте — это когда p-value становятся слишком маленькими из-за скрытой популяционной структуры или технических артефактов. Здесь значения λGC близко к 1, значит, серьёзных искажений нет.

Manhattan plot

На Manhattan-графике есть SNP, которые превышают мягкий порог значимости ($P < 0.001$), и их можно рассматривать как кандидатные маркеры.

Однако если применять более строгую коррекцию на множественные тесты (например, критерий Бонферрони, который при ~10 000 SNP даёт порог около $P \approx 4.5 \times 10^{-6}$), то ни один SNP не достигает этого уровня.

Это означает, что сильных, ярко выраженных QTL нет, а наблюдаемые пики — это скорее умеренные ассоциации, типичные для полигенных признаков вроде урожайности.



По итогу, для более уверенного выявления SNP, связанных с урожайностью, требуется более детальный анализ. Желательно использовать более строгие модели, учитывающие родственные связи и скрытую структуру популяции (например, смешанные модели с матрицей родства).