出國報告（出國類別：研習）

農委會農業菁英培訓計畫-
表型體分析技術應用於重要種苗作物研究及臺灣種苗產業之研習

服務機關：行政院農業委員會農業試驗所
姓名職稱：杜元凱 助理研究員
赴派國家：澳洲
出國期間：中華民國 107 年 6 月 24 日至 12 月 23 日
報告日期：中華民國 108 年 2 月 23 日

目次

一、目的

本培訓計畫旨於學習技術領先國家如何利用表型體 (Phenomics) 儀器、設施進行大量、精準的作物外表性狀調查 (Phenotyping)，以及隨後的大量數據與影像資料的擷取、儲存、轉換與統計分析，希望透過實地操作、研習如何有效率地進行作物各項重要外表型資料獲取，包含RGB、立體RGB、熱影像、葉綠素螢光以及高光譜影像等表型體資料。另外，本培訓計畫亦希望透過研習作物基因體定序資料分析流程，學習如何執行基因型分析 (Genotyping) 流程以獲取 SNP (Single nucleotide polymorphism)資料，並利用GWAS (Genome-wide association study) 分析串聯基因型與外表型資料(G to P)分析結果。

二、研習行程

1.行程摘要說明

本培訓計畫執行地點依研習目的分別於高解析植物表型體中心 (High Resolution Plant Phenomics Centre, HRPPC) 與澳洲國立大學 (Australian National University, ANU) 進行。表型體相關研習工作主要於HRPPC進行，植物基因體序列資料分析則主要於ANU進行，研習期間並參加第五屆國際植物表型研討會 (5th International Plant Phenotyping Symposium, IPPS)，摘要說明培訓計畫行程如下表1。

日期	工作內容	交通位置
06/24-06/25	出發	台北-坎培拉
06/26-10/01	開始研習工作	坎培拉
10/02-10/05	參加第五屆國際植物表型研討會	阿德雷德
10/06-12/21	研習工作	坎培拉
12/22-12/23	返國	坎培拉-台北

表1. 培訓計畫行程安排摘要說明表。

2.培訓單位簡介

HRPPC隸屬於澳洲表型體中心 (Australian Plant Phenomics Facility, APPF)，中心主持人為Xavier Sirault博士，APPF轄下共設立三個分支中心，其他兩個分別為位於阿德雷德之植物加速中心 (The Plant Accelerator, PA) 以及位於坎培拉之澳洲國立大學分支中心 (ANU node)，APPF任務為提升澳洲於表型體學相關研究社群於國際研究的影響力，其中，HRPPC專注於提供客製化表型體設備開發與表型體學研究相關服務，配置軟、硬體專業團隊與完備的後勤管理、服務系統。HRPPC營運預算主要由澳洲政府補助，部分營運預算來自承接私人公司與外部研究單位之表型體學應用、服務案件。

作物基因體序列資料分析流程研習主要在 ANU 之生物研究系(所) (Research School of Biology, RSB) 進行，指導研習工作之實驗室主持人為 Justin Borevitz 教授，其專長領域為植物族群遺傳學，近年研究利用高通量基因體定序技術、序列分析方法以及 GWAS 等統計應用方法探討植物於自然界之演變、適應，並發表許多論文於國際知名期刊，此外，Justin Borevitz 教授於 G to P 研究領域亦有相當琢磨。

3. 國際植物表型研討會

國際植物表型研討會為每年舉辦一次之國際研討會，第五屆於澳洲阿德雷德舉行，與會人員為來自各國公、私立研究單位植物表型體研究人員、表型體設備開發商、公司等。研討會內容由與會人員分享、討論項目分列如下:

(1).使用表型體學剖析植物基因型與環境因子的交互作用。

(2).應用表型體感應器、鏡頭高通量量測作物於特定環境條件下的生理反應。

(3).大量植物外表型資料之擷取、儲存、轉換與分析。

(4).植物外表型資料之品質管控、分享與資料庫建置、註解。

(5).生物統計學、電腦科學、資訊工程與表型體學的整合應用。

三、培訓內容

1. 表型體學相關影像擷取系統

HRPPC 具有之表型體學儀器、設備包含可於環境控制狀態下運作之 PlantScan、TrayScan，上述兩項為整合各項感測器、影像鏡頭之自動化數據擷取系統。另外，可於田間、戶外操作的設備則有空拍直升機以及 Phenomobile® 等。無論是何種形式、樣態的表型體學設備，均以獲取具有意義的植物生理反應數據或影像資料為基礎，由此設計可高通量資料擷取之自動化或半自動化之機械設備，在設備開發過程需考慮使用作物生長方法、環境條件限制、慣行農耕方式，故並非每一項表型體學儀器設備可應用於不同地理、氣候與農耕方式，因此，於此部分不特別個別介紹儀器設備，內容著重於各項重要鏡頭、感測器之功能以及使用限制，分述如下:

(1).RGB

RGB (red–green–blue) 影像是基於可見光 (400-700 nm) 的一種植物外觀型態研究工具，可應用植物地上部乃至於冠層 (Canopy) 的外表型分析，簡言之，一般肉眼可見的外觀變化，可利用 RGB 影像進行記錄、分析。RGB 的設置已廣泛應用於 2 維及 3 維的影像分析研究，包含田間試驗的近端偵測或配置到無人空拍機(Unmanned aerial vehicle, UAV)的遠端遙測。RGB 影像系統可在短時間內產生大量的影像，已有許多免付費開源軟體 (PlantCV, OpenCV, Bioimagetools)可以

執行影像處理，研究人員可進行影像之特徵擷取並與植物生理反應進行關聯分析外，所收集 RGB 之波長範圍可用來計算各項植被指數 (Normalized difference vegetation index, NDVI)，用以描述植物所處的生理狀態。RGB 影像分析主要問題點在於，如同其他類型影像分析之先天缺陷，由於生長過程植物器官發生重疊或植物之間的相互重疊，導致無法精確收集影像，另外，燈源(照明)變化的影響亦會影響影像收集的品質與再現性，實務操作上應注意物距、對焦與亮度設定是否合適。

(2).Stereo RGB

Stereo RGB 使用 2 個 RGB 鏡頭模擬人類的視覺行為，從 2 個單視覺系統中找到單點後，可以計算視差(disparity)並生成之物件的深度圖 (depth map)，經 3D 重建可於獲取植物立體結構模型。Stereo RGB 系統的主要優點是簡易性，2 個像機即可獲取足夠深度影像資料。目前立體視覺已經發展至多視角立體視覺 (Multi-view stereo, MVS)，並在植物表型體學研究中，與其他技術相比，如以光學雷達 (Light Detection And Ranging, LiDAR) 以及 Phenospex 公司開發之 PlantEye 系統，MVS 亦具有低成本的優勢。但 Stereo RGB 同樣易受照明變化與重疊的影響，同時需要合適的演算法進行立體匹配 (stereo-matching)，需要較高規格的電腦設備進行運算，田間或溫室使用 Stereo RGB 進行影像擷取將增加這些限制。

(3).LiDAR

LiDAR 使用雷射脈衝作為光源，藉由記錄雷射往返物件時間差，計算光源與物件間的距離，由此建構點雲 (point cloud) 資料以重建物件立體結構。LiDAR 始於遙測領域的應用，衛星型、機載型 LiDAR 系統已應用於量測植被冠層高度、面積、體積或生物量等，目前已有龍門系統或陸地型移動裝置搭載 LiDAR，進行全自動或半自動植物表型體研究。LiDAR 缺點為無法辨識顏色，雲點的計算需要時間。實務上當使用 LiDAR 時，因植物在不同生長時期的結構不同，如進入生殖生長或較形成較為複雜的結構時，雷射無穿透至植物結構底層，低層點雲資料量不足，重建立體結構將產生誤差，另外，氣候因子、昆蟲及空氣微粒造成訊號雜訊，亦需要校正的要求。

(4).熱影像 Thermal image

熱影像相機可擷取物體所放射出波長範圍從 3,000 到 14,000 nm 的輻射訊號，約介於中波紅外線 (middle wave infra-red, MWIR)、長波紅外線 (long wave infra-red, LWRI) 範圍，熱影像相機可將物件幅射能量轉換為溫度值。水分逆境和灌溉管理是熱影像的兩個應用領域，植物根據所處的環境條件，可藉由打開或關閉氣孔調節水分含量，氣孔開啟蒸散作用旺盛，氣控周遭溫度及降低，研究人員即可使用熱影像相機進行溫度量化，並依據量化結果推論植物相關生理反應，

如冠層溫度及氣孔導度 (stomatal conductance) 等。使用熱影像最主要需注意鏡頭感應器本身除接收目標物件所放射之紅外線輻射,同時也接收非目標物件分析如土壤、盛具等,環境濕度、放射率 (emissivity) 等參數設定以及植物本身短時間內之生理變化皆會影響熱影像數值的準確率,需要嚴謹的測量步驟及背景噪值的校正。

(5).葉綠素螢光

植物吸收光能進行光合作用,首先於光系統 II (Photosystem II, PSII)中進行光化學反應 (Photochemistry reaction),此步驟由葉綠素吸收光能驅動電子從反應中心 P680 傳遞到主要的電子接受者 Q_A,再由 Q_A 接續將電子傳遞至光系統 I (Photosystem I, PSI)而完成整個光合作用,部分光能在 PSII 中會以葉綠素螢光與熱的形式散失,雖然葉綠素螢光僅佔光能轉換之 1-2%,光化學反應、葉綠素螢光與熱三者競合 PSII 中總吸收光能,由於葉綠素螢光具有簡易量測的優勢,已為研究人員常應用於非破壞性量測植物光合作用效率的指標。當植物遭逢逆境時,造成類囊體膜系不完整或影響光合作用電子傳遞鏈正常運作,將會改變 PSII 光化學進行效率進而影響葉綠素螢光強度,因此,利用量測葉綠素螢光變化如 Fv/Fm 比值 (最大螢光參數),ΦPSII (有效螢光參數) 與 NPQ (non-photochemical quenching, qN) 等數值,可應用描述植物於特定環境下所處之生理狀態。擷取葉綠素螢光影像可於光照與預先進行暗適應 (dark-adaptation),特別需注意不同植物進行暗適應的時間長短不同,一般需要數十分鐘。另外,不同植物需要不同飽和光強度以中止光化學反應,通常需要 3,000 umol $m^{-2}s^{-1}$ 以上。實驗前需先確定影像擷取設備設定是否合宜植物所需,確保所計算暗適應葉片最小螢光值(F_0)、照射飽和光源後最大螢光(Fm)以及 Fv ($Fm-F_0=Fv$)等相關葉綠素螢光參數貼近植物光合作用真實狀態。

(6).高光譜

高光譜影像鏡頭依據等級不同,可接收之反射光波段不一,常見高光譜鏡頭可擷取完整的可見光及、紅外線區以及短波紅外線 (short wave infra-red, SWIR) 的連續光譜影像,解析度通常可達到數個奈米波長。不同顏色、水分含量、生化組成具特定反射圖譜,配合合適的統計分析方法如多變數分析、雙線性檢量法或者機器學習方式如人工神經網絡 (artificial neural network, ANN),可妥善分析圖譜資料與植物生理反應、品質及產量之關聯性。光譜的反射率受照明及冠層結構的影響,依據實驗需求通常需要客製化的燈源與影像擷取角度、距離的設定,另外,龐大高光譜影像的數據負載與影像處理流程,以及高光譜影像系統本身昂貴價格,皆限制了研究人員的使用。

2.表型體學設備操作-TrayScan

TrayScan 為整合 RGB、Stereo RGB、熱影像與葉綠素螢光影像擷取設備的

之自動化系統，配置環境控制、自動澆水與秤重系統，每一個盛盤 (Tray) 可裝載 20 個植株，每次 12 個盛盤，每盛盤影像擷取時間依設定約 10-40 分鐘，每日可擷取植株影像最多可達960株，因盛盤大小設計限制僅適合中、小型植物使用。TrayScan 操作流程以及後續分析軟體詳述於**附錄 1**。

如先前所提，使用各類型表型體學相關儀器、設備，均需考慮硬體設備先天空間與技術性的限制，此外，植物生長特性、試驗設計與使用資材亦會影響擷取影像的品質，因此操作人員對於硬體參數設定、各種鏡頭的使用特性與植物本身之生物性因素需要有足夠的了解，以獲取具生物解釋意義的影像資料與數據。計畫執行人於培訓期間與指導人員進行討論，以 TrayScan 為例，探討如何獲取具有足夠品質影像資料收集標準以及相關參數設定建議，結果詳述於**附錄 2**。

3.植物基因體序列資料分析-SNPs calling 分析流程

當研究人員已具有植物外表性狀資料，通常希望與其遺傳背景資料進行關連分析，期望找出與特定性狀具有高度連結度的分子標誌，配合分子標誌輔助育種，加速育種時效。因此，如何進行植物分子標誌的確認，也為相當重要的一環。隨次世代定序平台 (Next generation sequencing, NGS) 的快速發展，已可在短時間內獲取高通量的序列資料，配合合適的生物資訊分析後，即可得到足夠品質、數量的分子標誌如 SNP (Single nucleotide polymorphism)，以便外表型與基因型資料關聯分析工作的進行。

於 ANU 的研習工作主要學習如何以生物資訊分析流程進行植物 SNPs 的確認，稱 SNPs calling。我們以開源、免付費的 GATK toolkit 建議的分析流程為主，細部介紹請參考 https://software.broadinstitute.org/gatk/網址，我們使用的分析流程共分 9 個分析步驟(如下圖 1)。由於 GATK 需在 Linux 操作系統以指令方式使用，故每一分析步驟完成後，再接續下一步進行，當分析流程已經確立，我們希望調整個別分析步驟之細部參數在重新分析時，或者，未來其它序列資料欲套用同一套分析流程時，必須重頭一步一步進行，曠日廢時。因此，於此研習培訓工作我們使用另一個軟體 Snakemake 協助將我們所建立的分析流程自動化進行，細部介紹請參考 https://snakemake.readthedocs.io/en/stable/index.html#網址，簡言之，只要將我們 9 個分析指令在 Snakemake 進行編譯，未來執行一個 Snakemake 檔案，即可 9 個分析流程即會自動化接續執行，相當方便序列資料的再分析工作與再應用性。

SNPs calling 研習工作所用序列資料使用 96 個番茄收集系之 ddRAD (double digest Restriction enzyme Associated DNA, ddRAD) 建立文庫 (Libary)，並利用 Hiseq 2500 平台進行 125bp paired end 定序。利用 Snakemake 將 GATK 之 9 個分析步驟之指令檔案與細部說明，詳如**附錄 3**。

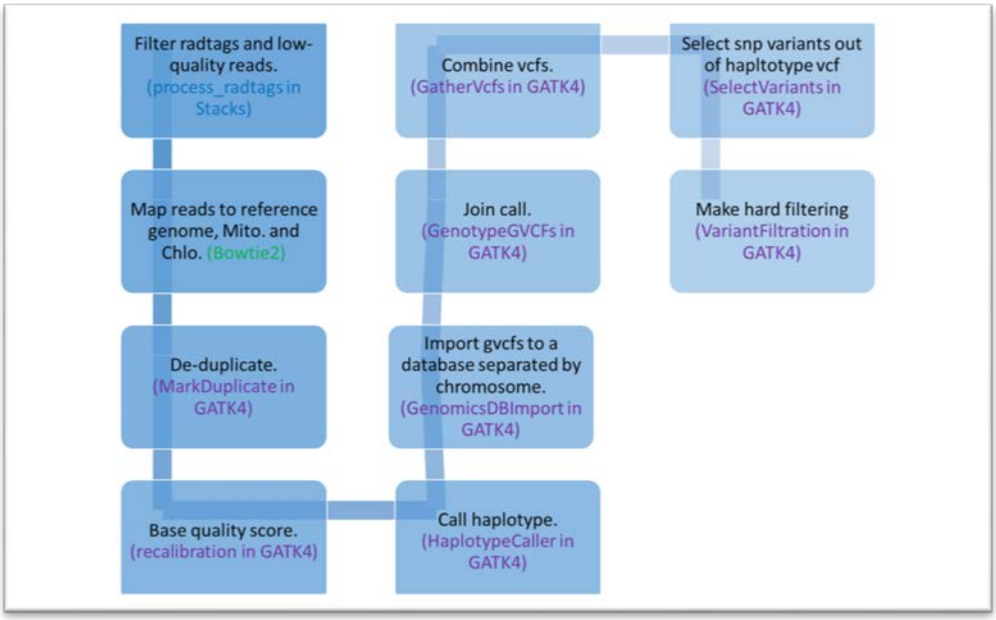

Filter radtags and low-quality reads. (process_radtags in Stacks)
Map reads to reference genome, Mito. and Chlo. (Bowtie2)
De-duplicate. (MarkDuplicate in GATK4)
Base quality score. (recalibration in GATK4)
Call haplotype. (HaplotypeCaller in GATK4)
Import gvcfs to a database separated by chromosome. (GenomicsDBImport in GATK4)
Join call. (GenotypeGVCFs in GATK4)
Combine vcfs. (GatherVcfs in GATK4)
Select snp variants out of hapltotype vcf (SelectVariants in GATK4)
Make hard filtering (VariantFiltration in GATK4)

圖 1. 利用 GATK 進行 SNPs calling 流程示意圖

4.串聯外表型資料與基因型分析結果-GWAS 分析流程

當已獲取作物基因型資料如 SNPs 以及外表型資料或性狀資料，在合適的族群大小前提下，可利用廣義線性模型 (General linear model ,GLM) 或指混合線性模型 (Mixed linear model, MLM) 進行關聯分析，由此將作物基因型 (G) 與外表型 (P) 資料進行串聯。此項研習工作使用免付費軟體TASSEL進行GWAS分析，一般而言，使用 GLM 進行關聯分析可滿足大部分研究人員需求，然而，若參試作物材料具有複雜的譜系 (Pedigree) 或明顯的創始者效應 (Founder effect) 現象，將導致型 I 誤差 (Type I error) 的膨脹，此時應嘗試使用 MLM 降低型 I 誤差以提升模型的檢力 (Power)。建議進行 GWAS 應同時選用 GLM 與 MLM 進行分析，並藉由繪製 QQ 圖 (Quantile-Quantile plot) 評估 GML 與 MLM 的合理性。分析流程與說明詳如**附錄 4**。

四、心得與建議

本計畫為期間 6 個月培訓研習計畫，計畫執行人獲益良多亦拓展眼界，植物表型體學實為未來農業研究重要之應用型研究項目，後續工作首要希望持續維持與 HRPPC 與 ANU 之研究夥伴關係，同時嘗試研提科技計畫，希望未來我國可初步可導入人力、預算等相關資源，開啟我國表型體學之研究工作。於此感謝本會計畫補助進行培訓計畫，關於農業菁英培訓計畫，本人建議，針對特定研究領域，現行培訓計畫最長為期 3 年(博士班進修)，短期培訓最長為僅半年，且 5 年內不得再申請，以農業研究領域而言，若培訓人員參與培訓內容涉及作物栽培、生長調查，或者是需要跨領域之研習培訓項目，通常無法深入地參與整個研究題

目，且缺乏完整性、延續性，建議可將短期研究延長至少為 1 年，並取消 5 年內不得再申請之限制。

計畫研習項目除學習如何使用各項非破壞性方法直接或間接量測植物各項生理反應，以及學習利用各項統計方法、模型以及相關演算法以描述外觀型態與推估作物生理狀態、生物量等，目前於國內尚未有大專院校或研究機構投入表型體學研究領域，希望將本次研習所得相關經驗、知識，以及所了解之國外該領域研究發展現況與經驗以及所現狀所遇到的問題帶回我國，盡量降低未來欲開展表型體研究過程中不必要的人力、金錢支出。表型體基礎設施設備建置需可觀經費、人力，也需要時間累積經驗，雖國內尚未具有規模地投入表型體學研究，起步較晚，然而基於我國厚實之軟體開發、AI 應用與農業研究能量，個人認為若未來有實際人力、資源投入表型體研究，應當具有相當研究創新能力。

我國尚未投注資源於表型體學研究領域，雖然於研究領域未佔有先機，然而，各國對於發展表型體學研究仍有不同看法，因此也影響投入研究之相關資源，例如日本，目前亦尚未投注經費或計畫進行具規模性的植物表型體研究。計畫執行人認為，表型體學研究雖然相當仰賴於儀器、設備進行高通量、非破壞性植物外表型調查、收集，而儀器、設備的設計使用依據不同作物與目的需要相當程度客製化，因此，進行表型體研究首先更應思考的是要以何種作物進行研究，接續再思考應使用或設計合適的硬體設備協助研究的進行，計畫執行人建議，應先有策略性地針對我國農業發展進行盤點，挑選具國際競爭力的潛在標的作物投入表型體研究。由於我國耕地破碎，無法以大田作物之研究與他國競爭，然而種苗研發、作物育種必然為我國固有之優勢，因此若欲投入資源於表型體研究，可以活化種苗、種原為主，配合表型體學、基因體學之串連，強化、加速我國在育種上之優勢，建議首先可以室內型以標的作物進行相關表型體學之規劃、設計，據此研究結果可直接扣合產業需求，是較為實際之作法。例如，對於國家種原庫長久收集與保存之眾多作物收集系，建立基因型資料後除可應用於移除多餘(Redundant)或重複之收集系，將可有效降低種原庫之內各項作物保存與管理工作。所保留下來之作物收集系，由於數量減少，研究人員可資源集中於此核心收集系 (Core collection)，針對基因型與外表型進行更深入的研究。另一部分，待釐清、彙整種原庫內各收集係之詳細基因型與外表型資料後，未來亦可據此策略性布局我國作物育種方向與經費分配重點，協助育種者與產業選擇合適親本進行作物育種，生產多樣具競爭性之作物品種，鞏固與提升我國農業研究與研發之競爭能力。

Appendix 1 - Operation guideline of TrayScan for a beginner

1. **Running the TrayScan.**

There are four steps to operate the TrayScan:

a. **Engage all cameras and test the camera setting.**

a). Open PlantScreen (a default operation software for TrayScan)

b). Connect Database (click Database tab)

c). Connect FLIR (select add and click, wait for ip add appear)

d). Connect Infrared (Infrared tab) (Put a tray in thermal image station to test imaging)

e). Check light setting (Main tab for brightness test)

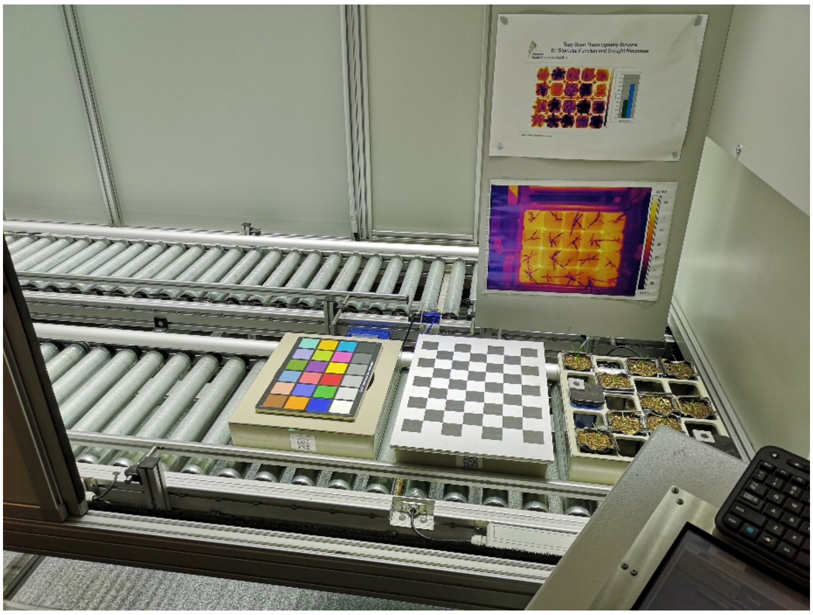
f). Put a tray in stereo RGB station and take a snapshot to check RGB camera are working.

g). Check a blue light on the Fluorcam computer as indicated it is on.

h). Load fluorescence protocol (file.p) in PlantScreen. (Before loading protocol, you could open the protocol in Fluorcam 7.0 and put a tray in fluorescence station for test the image)

i). In PlantScreen and Fluorcam tab, send to device.

j). Check any error messages in the System Messages tab



In this step, the tray should still be in loading position.

b. **Loading trays.**

a). When Plant Screen software show “Fill the tray then close door”.

b). Load only 4 trays at a time (A maximum of 15 rays can be loaded per batch).

c). Load chessboard calibration target and the colour checker calibration target with the barcode facing you. Then load experimental trays and close the door.

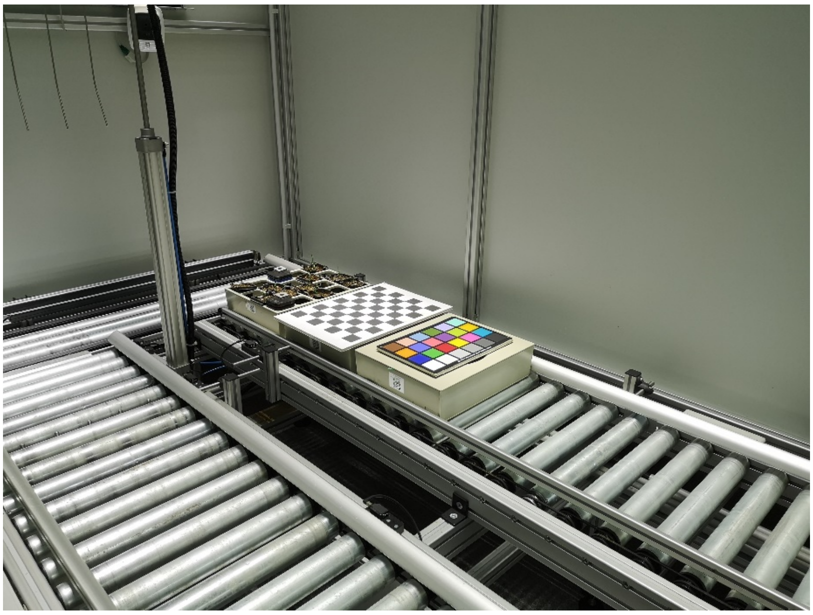
d). Push the white OK button on the PLC to load these trays into the machine.

e). When these trays are loaded, the PlantScreen message will again show "Fill the tray then close door".

f). Load the next set of trays, and repeat until the last trays are loaded. Push the white OK button on the PLC.

g). Once all trays are loaded push the blue Measuring Begin/End button on the PLC.

h). Check the graphic on PlantScreen it should show the position and number of the trays in the system and present the message "Waiting for start protocol".



After loading, the tray should be in acclimation position.

c. Load experiment protocol and start images collection.

a). In PlantScreen, go to the Experiment tab, where existing protocols are loaded and saved. You could create a new protocol or load previous protocol, it can be identified as a ‘‘file.ptc’’.

b). In the experiment tab, fill in the experiment name and responsible person.

c). In the experiment tab the Start button will be highlighted green when all the steps have been completed, and is clicked to initiate the run.

d). Turn off room light, shut the door and change the sign on the door to "in use".

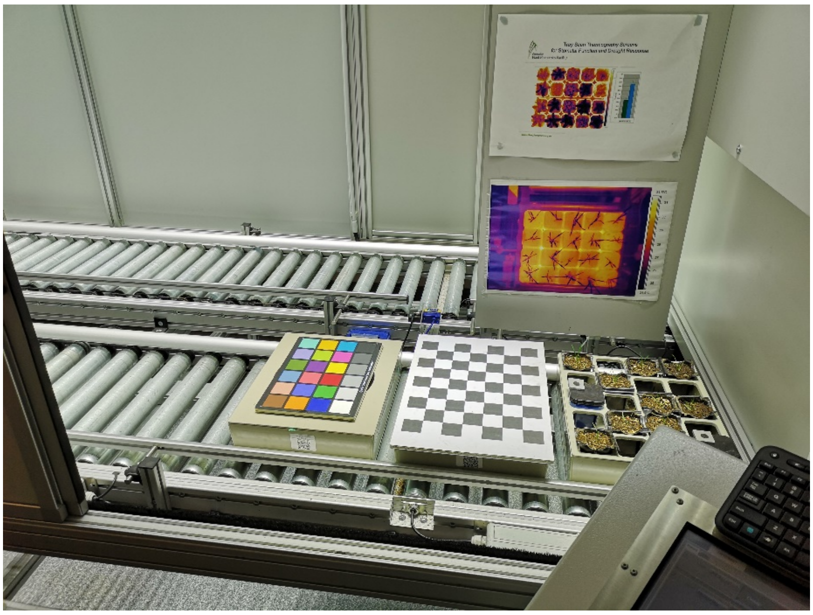
e). In PlantScreen and main tab will display the message "Measuring in Progress", and display the tray ID at each measuring station.

d. Unload the trays.

a). When the experiment has finished the PlantScreen display will show ‘‘Measuring End’’.

b). Push the Measuring Begin/End and OK button on the PLC and wait for PlantScan to display "take out tray". Open the door, remove the trays, and then close the door.

c). Press the the OK button again for the next set of trays to be forwarded. TrayScan may be reloaded to measure further trays.



While all measuring process finish, the tray will return to loading position and you can unload the tray.

※ A whole cycle of TrayScan operation strictly follows a → b → c → d procedure. If there are any unexpected events make the process stuck or you need to abort the procedure. Strongly recommend you just try to let the machine finish the a → b → c → d procedure before you try to reboot the system.

※ Unless requested by HRPPC technical staff DO NOT shutdown the computers.

※ If you have manually turned the lights on before the experiment you will also need to manually turn them off before exiting PlanScreen.

※ To exit the PlantScreen use the "Close client" button on PlantScreen. Close the PlantServer window, and you’ re done.

2. Design your own protocols.

There are two protocols that user can customize according to specific requirements. The file.p and file.ptc files are used to obtain important parameters of photosynthesis system II (PSII) and set the sequential imaging of cameras (thermal, RGB, fluorocam, stereo RGB), respectively.

a. Compile file.p.

User can compile file.p by using FluorCam7 software. Start the Fluorcam7 software and click the Protocols tab, then you are ready to design your file.p and set physical parameters as followed:

a). **TS**: Time at which background signal is measured prior to the fluorescence being sampled during the measurement flash. We suggest to set the TS = 40 ms is a good start.

b). **Shutter**: Duration of opening time for the electric shutter. We suggest to set the Shutter = 2 then adjust after preview the fluorescence image.

c). **Sensitivity**: Sensitivity of the fluorescence camera. Avoid using 80 – 100% unless due to low signal to noise ratio.

d). **Act**: Relative power of actinic light (white light), 0 – 100%, user defined.

e). **Act1**: Relative power of actinic light 1 (red light), 0 – 100%, user defined.

f). **Super**: Amount of light used for saturating pulse, 0 – 100%, user defined.

b. Compile file.ptc.

User can use PlantScreen to decide the sequential imaging of cameras. On the main screen of PlantScreen, click protocol tab. Then arrange your own imaging process.

※ Please see attachment file FluorCam_Instruction_Manual section V.C. to see more detail information of how to compile your file.p.

Software and analysis

1. RGB

There are plenty of free and open-source software that could be used to view, processing RGB file. A simple way to process RGB image is to use Image J. However, we suggest a free OpenCV (Open Source Computer Vision Library: http://opencv.org) which includes several hundreds of computer vision algorithms. If you are familiar with python, we further recommend PlantCV which is an imaging processing package specific for plants that is built upon open-source software platforms OpenCV, NumPy and MatPlotLib

(Donald Danforth Plant Science Center). There are many pipelines already established for regular image processing. Please refer to website below to give you some basic idea how to view and process RGB files using OpenCV and PlantCV.

※ Please browse the HRPPC and PSI website for more information:
https://imagej.nih.gov/ij/index.html
https://docs.opencv.org/3.4/index.html
https://plantcv.readthedocs.io/en/latest/

2. **Thermal image**

Please see attachment file Process FLIR with Fiji to see the demonstration of how to use Image J to analyze thermal image.

3. **Fluorescence image**

Please see attachment file FluorCam_Instruction_Manual section V.D. to V.E. to get more detail how to pre-processing and analyze the results.

4. **Stereo RGB**

We suggest a TrayScan_pipeline which was developed by HRPPC. A semi-global matching algorithm was applied this pipeline to estimate the point clouds of each plants on the tray. By means of surface reconstruction methods, the point clouds of each plants were then used to generate meshes. These meshes were further analyzed to quantify the plant dimensions. Please see the attachment files, Trayscan_guide and Trayscan_stereo_test_run_guide to understand how to install and execute stereo RGB analysis by TrayScan_pipeline.

Appendix 2 - Qualitative recommendations for data collection with TrayScan

It is not possible to provide specific standard operational procedures (SOP) for TrayScan that are suitable for all crop species, growth rates or plant/leaf architectures. For these reasons we provide qualitative recommendations for good data collection practices, these are divided into two aspects, physical and biological parameters. Users are advised to follow the principles as described below.

1. **Physical parameters setting of TrayScan:**

a. **Environment condition**

Environment parameters such as temperature and humidity at acclimation chamber, image stations and the workplace or space should be consistent and stable. Regular inspections of the related environment condition should be conducted to achieve consistent and reproducible measurements. Light intensity at the acclimation chamber is adjustable. We suggest user should check the longevity of the light source and examine the intensity output by light meter as well to prevent any undetectable light decay. We also suggest the sensors for detecting the environment parameters should be calibrated routinely.

b. **Camera calibration**

There are four type of cameras are implemented in TrayScan to obtain thermal image, RGB, stereo RGB and fluorescence image. Those cameras should have regular maintenance and calibration.

c. **Light intensity**

There are two light banks in acclimation chamber and one in the thermal imaging station. The intensity of the light bank is adjustable and ranged from 0 – 1000 μmol m^{-2} s^{-1} at the source. The distance from the light source to the canopy height of plants will influence the intensity of light at the leaf level. We suggest users measure the light intensity with a light meter to assure the setting of light source intensity and distance from light source to plant canopy height fits the experiment requirement.

	Acclimation chamber Light level (μmol m^{-2} s^{-1})		**Thermal imaging station** Light level (μmol m^{-2} s^{-1})
Set level (%)	Tray height (20cm above the conveyor)	**Set level (%)**	Tray height (20cm above the conveyor)
10	116	**10**	54

20	239	**20**	114
30	347	**30**	170
40	447	**40**	226
50	538	**50**	278
60	624	**60**	328
70	706	**70**	376
80	781	**80**	421
90	852	**90**	465
100	914	**100**	503

d. Layout of pots within trays (CRD, RCBD)

We suggest the pot layout on the tray should be complete random design or random complete block design if you find obvious environmental bias or any unexpected environmental variation in TrayScan.

e. Thermal camera (FLIR A600-series)

Default setting of IR camera are listed as followed:

Object emissivity: 1.00
Object distance: 2.00 m
Atmospheric temperature: 20.00 ℃
Reflected temperature: 20.00 ℃
Relative humidity: 50 ℃
Recorded in ℃
Output size: 640 x 480 pixel with resolution 96 x 96 pp.

In the step of Operation guideline for a beginner 1.a.(d), we suggest user check the thermal image first. Make sure the setting of distance, focus and environment temperature are appropriate. The thermal image should cover whole tray and have clear outline of plants and pots.

f. RGB camera (Computar Megapixel, f = 4.5 – 13.2mm)

In the step of Operation guideline for a beginner 1.a.(e) and (f), we suggest user check the RGB first. Make sure the setting of distance, focus and brightness are appropriate. The RGB image should cover whole tray and have apparent, distinct color on soil and plant organs.

g. Fluorescence module (Pentax TV lense 4.8 mm and LED panel)

Regarding the physical parameter setting in file.p, we suggest user to set the shutter on appropriate exposure. The longer the shutter is open, the brighter a measurement will be, and the higher the signal the better the S/N ratio. However, if the measure light is too strong, the measured signal deviated from the true $F_{0.}$ User should prevent the signal received exceeds the maximum capacity of the camera. If the camera receives too much signal, the live screen will show a warning message "Pixels Overflow". In terms of sensitivity setting, unless there is a special requirement for experiment, we suggest user set sensitivity under 80% in case the camera receive exceeding signal than maximum capacity. The live screen will show a warning message "Pixels Overflow" if user set sensitivity too high.

	PAM (Pulse-Amplitude-Modulated) mode		
Electronic shutter number	0	1	2
Shutter opening time (μs)	10	20	33

In order to measure F_0, a pulse of weak Act1 (red light) is required to remove electron from Q_A (oxidized Q_A). We also need low enough Act (white light) which triggers minimal level of fluorescence but does not drive significant PSII chemistry, typically less than 0.1 μmol. To obtain Fm, the saturating pulse has to be sufficiently bright to maximally reduced Q_A. Normally, a short actinic pulse of high PPFD (photosynthetically active photon flux density) less than 1 s at thousands μmol $m^{-2}s^{-1}$ is needed (Please notice the highest PPFD setting of fluorescence module for TrayScan is 3,000 μmol $m^{-2}s^{-1}$). We suggest user preview the fluorescence image with Fluorocam 7.0 software, click the plot tab to assure the Act, Act1 and Super are set appropriately.

Other than those physical parameter setting in file.p as mentioned above. The height from the conveyor belt to the LED light banks has tremendous effect as well. We strongly suggest user should put health plants on the tray to meet the basic requirement for acceptable fluorescence image, that is, under the condition which image covers whole tray, the settings should be able to obtain consistent Fv/Fm near 0.83 for a health plant. As a result, in the step of Operation guideline for a beginner 1.a.(h), user should check the heights and parameters in file.p are set appropriate in order to obtain a 0.83 Fv/Fm for health plants.

Height (mm) on GUI of PlantScreen	**Actual height (cm) to the conveyor belt**
0	30
100	39.5
200	49.5
300	59.5

370 (max)	67.5

h. Stereo RGB (Computar Megapixel, f = 4.5 – 13.2mm)

Due to the stereo camera deployment position and camera type in TrayScan, the distance between camera and objects has to be at least 390 mm. Two other important parameters that may require tuning according to pot size and crop type used in your experiment:

a). green_seg.py: the 'raw_filter' function hard coded distance value to remove underneath tray and pots (420mm atm). Currently top tray:400mm, bottom tray: 482mm.

b). green_seg.py: the 'segment' function in which ExG and ExR (parameter 2 and 1.4) may change due to different lighting condition and plant species.

As for RGB image, user need to assure the setting of distance, focus and brightness are appropriate. Stereo RGB image should cover whole tray and have apparent, distinct color on soil and plant organs as well. We recommend user to process the stereo RGB with TrayScan_pipeline first while obtain your first data set.

2. Biological parameters:

a. Crop type

The physical limitations of TrayScan restrict the size, leaf type of crops that user could apply. The max dimensions for a plant is 59.5 cm (height) x 44 cm (width). In order to obtain image data with acceptable quality, user should prevent using oversized plant and plant with complicated architecture. As mentioned in previous section, we suggest user to preview the test image before performing your experiment.

b. Growth stage

It’s important to have a precise definition of growth stage for your plant. It allows you to clearly delimit the plant growth stage you apply to TrayScan, then user could compare the results between treatment versus control and different batches of experiments based on the same biological foundation.

c. Time of dark adaption

To obtain Fv/Fm value from TrayScan, we need to apply dark adaption to the plants prior to the minimum fluorescence (F0) is measured. Next, saturation pulse is used to close all available reaction centers, then maximum fluorescence (Fm) is measured. The difference between maximum fluorescence and minimum fluorescence is Fv. However, dark adaptation is a complicated process. Depending on crop species, the duration of dark adaptation are different. According to Kalaji et al. (2014), 15 min of dark adaptation is recommended. Generally, we suggest 30 min is reasonable and acceptable for most crop species.

d. Soil, fertilizer and watering

The variation of image results collected from TrayScan arises not only from the individual variance between plants but also from the heterogeneity of the soil, fertilizer and watering method that user apply to plants. Regarding the soil, it could be made of different substances. We recommend user mix the substances though evenly. If necessary, use sieve to filter each substance before mix. As for fertilizer, user should strictly follow the formula to make up the fertilizer. Apply equal amount to each plant in the same way, and notice the expiry date of the fertilizer. Finally, user should make sure each plant acquires enough and balanced water. User should prevent watering too much. Consistent wet soil may lead to moss grow on the surface of soil. Moss could cause some problems while processing image data.

e. Treatment

Similar to Quality recommendation for data collection 1 a, the treatment such as heat, drought, cold to the plants should be consistent and stable. Make sure the treatment is effective compared to the control plants. If user use growth chamber or green house, regular inspections of the related environment condition and setting should be conducted. We also suggest the sensors for detecting the environment parameters should be calibrated routinely.

```
## Environment setting and softwares
# 1. Snakemake (follow installation instruction and establish a python env with default environmental.yaml)
# 2. Stacks-2.2
# 3. gatk-4.0.11.0
# 4. picard.jar
# 5. rtg-tools-3.10

## If you use this Snakefile in different PC, remember to change the path first before you run the Snakefile in yourown PC.
# 1. Put processing file and picard.jar in the same directoryself.
# 2. Changing path in each steps like  /home/usr/working_directory/cleaned_reads/
#                                                    compared to /home/yktu/Try_process_radtags/cleaned_reads/
# (Optional: 3. Don't forget to change the path in cohor.sample_map.)

## Prerequisite files:
# 1. unmapped_reads/R1.fastq, unmapped_reads/R2.fastq. The paired reads are debarding and demultiplexing by process_rapture.py which is not including in this analysis pipeline.
# 2. ref_genome/ref_genome.fa. Remember to index the ref genome by bwa index first.
# 3. vcf_base_recali/SNPdb.vcf for BaseRecalibrator step. If it's a VCF file, please index it using the bundled tool IndexFeatureFile to make .tib file.
# (Optional: 4. cohort.sample_map, a Tab-delimited text file with sample_name--tab--path_to_sample_vcf per line.)
#   ( Optional: Don't forget to change the path in cohor.sample_map.)

## Prerequisite variables setting:
# 1. SAMPLES = ["r01", "r02"]  Accroding to your sample numbers, could be extend to r96.
# 2. CHRNUM = ["ch00","ch01", "ch02", "ch03", "ch04", "ch05", "ch06", "ch07", "ch08", "ch09", "ch10", "ch11", "ch12"] for combining the variants information from each by chromosome.


## We go data pre-processing first according to Best practice workflow | Created 20180109 | Last update 20180306
```

```
# Debarding and demultiplexing paired end reads (read1 and read2 for each sample) by process_rapture.py which is not including in this analysis pipeline.
# 1. Clean paired readsself.
# 2. Make unmapped BAM from cleaned paired reads. At the time, align cleaned paired read to ref genome then make aligned BAMself.
# 3. Merge uBAM and BAM.
# 4. Sorted merged BAM.
# 5. Mark the duplicated of merged BAM.
# 6. Make the recalibration table for merged BAM.
# 7. Apply the recalibration talbe to the merged BAM.
# 8. Calling variants and generate gVCF filesself.
# 9. Combining gVCF of each sameples by chromosomeself as chromosome.db.
# 10. Convert chromosome.db to VCF.
# 11. Integrate all VCF filesself.
# 12. Create index for the integrated VCFself.
# 13. Filter variants by user's criteria.
# 14. Select SNP by user's demand.

SAMPLES = ["r01", "r02", "r03", "r04", "r05", "r06", "r07", "r08", "r09", "r10", "r11", "r12", "r13", "r14",
                "r15", "r16", "r17", "r18", "r19","r20", "r21", "r22", "r23", "r24", "r25", "r26", "r27", "r28",
                "r29", "r30", "r31", "r32", "r33", "r34", "r35", "r36", "r37", "r38", "r39", "r40", "r41", "r42",
                "r43", "r44", "r45", "r46", "r47", "r48", "r49", "r50", "r51", "r52", "r53", "r54", "r55", "r56",
                "r57", "r58", "r59", "r60", "r61", "r62", "r63", "r64", "r65", "r66", "r67", "r68", "r69", "r70",
                "r71", "r72", "r73", "r74", "r75", "r76", "r77", "r78", "r79", "r80", "r81", "r82", "r83", "r84",
                "r85", "r86", "r87", "r88", "r89", "r90", "r91", "r92", "r93", "r94", "r95", "r96"]

CHRNUM = ["ch00","ch01", "ch02", "ch03", "ch04", "ch05", "ch06", "ch07", "ch08", "ch09", "ch10", "ch11", "ch12"]

rule all:
    input:
        "Filter/filtered.all.vcf.gz",
        "Sel_Var/selected_all.vcf",
        "Gather_vcf/all.vcf.gz",
        expand("Joint_vcf/{chr_num}.vcf.gz", chr_num = CHRNUM),
        #expand(directory("{chr_num}_database"), chr_num = CHRNUM),
        expand("interGVCF/{sample}.g.vcf.gz", sample = SAMPLES),
        #expand("appliedBQSR/{sample}.bam", sample = SAMPLES),
        #expand("recal_dataTable/{sample}.table", sample = SAMPLES),
        #expand(["MarkDuplicate/{sample}.bam",
"MarkDuplicate/{sample}_metrics.txt"], sample = SAMPLES),
        #expand("Sorted_mergeBAM/{sample}.bam", sample = SAMPLES),
        #expand("Merge_uBAM_alBAM/{sample}.bam", sample = SAMPLES),
        #expand("mapped_reads/{sample}.bam", sample = SAMPLES),
        #expand("unmapped_BAM/{sample}.bam", sample = SAMPLES),
        #expand("cleaned_reads/{sample}_R1.1.fq", sample = SAMPLES),
        #expand("cleaned_reads/{sample}_R2.2.fq", sample = SAMPLES),
        #temp("cleaned_reads/{sample}_R1.fastq.rem.1.fq"),
        #temp("cleaned_reads/{sample}_R2.fastq.rem.2.fq")


rule clean_reads:# Clean the reads.fastq by precess_radtags tool of Stacks after
debarding and demultiplexing by process_rapture.py
    input:
        R1 = "unmapped_reads/{sample}_R1.fastq",
        R2 = "unmapped_reads/{sample}_R2.fastq"
    output:
        "cleaned_reads/{sample}_R1.1.fq",
        "cleaned_reads/{sample}_R2.2.fq",
        temp("cleaned_reads/{sample}_R1.fastq.rem.1.fq"),
        temp("cleaned_reads/{sample}_R2.fastq.rem.2.fq")
    shell:
        "process_radtags -1 {input.R1} -2 {input.R2} -o
/home/poweruser/96_snakemake/cleaned_reads/ -e pstI -r -c -q"

rule uBAM:# Creat unmapped BAM to keep important information originate from
raw reads directly from cleaned reads.fastq.
    input:
        cleaned1 = "cleaned_reads/{sample}_R1.1.fq",
        cleaned2 = "cleaned_reads/{sample}_R2.2.fq"
    output:
        uBAM = "unmapped_BAM/{sample}.bam"
    shell:
        "java -jar picard.jar FastqToSam F1={input.cleaned1} F2={input.cleaned2}
O={output.uBAM} SM={wildcards.sample} PL=illumina" # Remember to add PL
argument.

rule bwa_map:# Creat mapped BAM from cleaned reads. # This step is slow, try add
threads!!
    input:
        "ref_genome/S_lycopersicum_chromosomes.2.50.fa",#Remember to index
the ref genome by bwa index first.
        "cleaned_reads/{sample}_R1.1.fq",         #And need to create dict by
picard.jar CreatSequence Dictionary.
        "cleaned_reads/{sample}_R2.2.fq"          #Use 2.5 instead of 3.0 is
because on rule MarkDuplicates step, we could only obtain vcf of 2.5 version.
    output: # output for shell in this rule.
        alBAM = "mapped_reads/{sample}.bam"
    shell:# Running with shelll script
        "bwa mem {input} | samtools view -Sb - > {output.alBAM}"

rule Merge_uBAM_alBAM:# Merge uBAM and mapped BAM to keep alignment and
unmapped information together.
    input:
        alBAM = "mapped_reads/{sample}.bam",
        uBAM = "unmapped_BAM/{sample}.bam"
    output:
        "Merge_uBAM_alBAM/{sample}.bam"
    shell:
        "java -jar picard.jar MergeBamAlignment ALIGNED={input.alBAM}
UNMAPPED={input.uBAM} O={output}
R=ref_genome/S_lycopersicum_chromosomes.2.50.fa"

rule Sort_mergeBAM:# Sort merged BAM.
    input:
        "Merge_uBAM_alBAM/{sample}.bam"
    output:
        "Sorted_mergeBAM/{sample}.bam"
    shell:
        "java -jar picard.jar SortSam I={input} O={output}
SORT_ORDER=coordinate"

rule MarkDuplicates:# To mark read pairs that are likely to have orininated from
duplicates of the same original DNA fragments through some artifactual processes.
    input:
        "Sorted_mergeBAM/{sample}.bam"
    output:
        markBAM = "MarkDuplicate/{sample}.bam",
        METRIC = "MarkDuplicate/{sample}_metrics.txt"
    shell:#Pls notice, for the java script there should be no sapce between I, = and {}.
        "java -jar picard.jar MarkDuplicates I={input} O={output.markBAM}
M={output.METRIC}"

rule Base_recali:# Make a recalibration talbe based on dbSNP.
    input:
        markBAM = "MarkDuplicate/{sample}.bam"
    output:
        calTab = "recal_dataTable/{sample}.table"
    shell: # Remember to  index the SNPdp using the bundled tool IndexFeatureFile
        "/home/poweruser/gatk-4.0.11.0/./gatk BaseRecalibrator -I
{input.markBAM} -R ref_genome/S_lycopersicum_chromosomes.2.50.fa
--known-sites vcf_base_recali/360_merged_2.50.vcf -O {output.calTab}"

rule Apply_BQSR:# Apply the recalibration talbe to the markduplicated reads.
    input:
        markBAM = "MarkDuplicate/{sample}.bam",
        calTab = "recal_dataTable/{sample}.table"
    output:
        "appliedBQSR/{sample}.bam"
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk ApplyBQSR -R
ref_genome/S_lycopersicum_chromosomes.2.50.fa -I {input.markBAM}
--bqsr-recal-file {input.calTab} -O {output}"

rule haplotypecaller:# Generate a intermediate GVCF with likeihoods determination
```

```
for the haplotypes given the read data and asign genotypes (SNPs and Indels) to sample.
    input:           # This step is pretty slow, try add threads   or other way to accelerate.
        "appliedBQSR/{sample}.bam"
    output:
        "interGVCF/{sample}.g.vcf.gz"
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk --java-options '-Xmx8g' HaplotypeCaller -R ref_genome/S_lycopersicum_chromosomes.2.50.fa -I {input} -O {output} -ERC GVCF "

rule combine_gVCF_byChromosome:# We combine gVCF of each sample by chromosome as seperated database. Then, databases will be integrated in following step.
    input:
         I1  = "interGVCF/r01.g.vcf.gz",
         I2  = "interGVCF/r02.g.vcf.gz",
         I3  = "interGVCF/r03.g.vcf.gz",
         I4  = "interGVCF/r04.g.vcf.gz",
         I5  = "interGVCF/r05.g.vcf.gz",
         I6  = "interGVCF/r06.g.vcf.gz",
         I7  = "interGVCF/r07.g.vcf.gz",
         I8  = "interGVCF/r08.g.vcf.gz",
         I9  = "interGVCF/r09.g.vcf.gz",
         I10  = "interGVCF/r10.g.vcf.gz",
         I11  = "interGVCF/r11.g.vcf.gz",
         I12  = "interGVCF/r12.g.vcf.gz",
         I13  = "interGVCF/r13.g.vcf.gz",
         I14  = "interGVCF/r14.g.vcf.gz",
         I15  = "interGVCF/r15.g.vcf.gz",
         I16  = "interGVCF/r16.g.vcf.gz",
         I17  = "interGVCF/r17.g.vcf.gz",
         I18  = "interGVCF/r18.g.vcf.gz",
         I19  = "interGVCF/r19.g.vcf.gz",
         I20  = "interGVCF/r20.g.vcf.gz",
         I21  = "interGVCF/r21.g.vcf.gz",
         I22  = "interGVCF/r22.g.vcf.gz",
         I23    = "interGVCF/r23.g.vcf.gz",
         I24    = "interGVCF/r24.g.vcf.gz",
         I25    = "interGVCF/r25.g.vcf.gz",
         I26    = "interGVCF/r26.g.vcf.gz",
         I27    = "interGVCF/r27.g.vcf.gz",
         I28    = "interGVCF/r28.g.vcf.gz",
         I29    = "interGVCF/r29.g.vcf.gz",
         I30    = "interGVCF/r30.g.vcf.gz",
         I31    = "interGVCF/r31.g.vcf.gz",
         I32    = "interGVCF/r32.g.vcf.gz",
         I33    = "interGVCF/r33.g.vcf.gz",
         I34    = "interGVCF/r34.g.vcf.gz",
         I35    = "interGVCF/r35.g.vcf.gz",
         I36    = "interGVCF/r36.g.vcf.gz",
         I37    = "interGVCF/r37.g.vcf.gz",
         I38    = "interGVCF/r38.g.vcf.gz",
         I39    = "interGVCF/r39.g.vcf.gz",
         I40    = "interGVCF/r40.g.vcf.gz",
         I41    = "interGVCF/r41.g.vcf.gz",
         I42    = "interGVCF/r42.g.vcf.gz",
         I43    = "interGVCF/r43.g.vcf.gz",
         I44    = "interGVCF/r44.g.vcf.gz",
         I45    = "interGVCF/r45.g.vcf.gz",
         I46    = "interGVCF/r46.g.vcf.gz",
         I47    = "interGVCF/r47.g.vcf.gz",
         I48    = "interGVCF/r48.g.vcf.gz",
         I49    = "interGVCF/r49.g.vcf.gz",
         I50    = "interGVCF/r50.g.vcf.gz",
         I51    = "interGVCF/r51.g.vcf.gz",
         I52    = "interGVCF/r52.g.vcf.gz",
         I53    = "interGVCF/r53.g.vcf.gz",
         I54    = "interGVCF/r54.g.vcf.gz",
         I55    = "interGVCF/r55.g.vcf.gz",
         I56    = "interGVCF/r56.g.vcf.gz",
         I57    = "interGVCF/r57.g.vcf.gz",
         I58    = "interGVCF/r58.g.vcf.gz",
         I59    = "interGVCF/r59.g.vcf.gz",
         I60    = "interGVCF/r60.g.vcf.gz",
         I61  = "interGVCF/r61.g.vcf.gz",
         I62  = "interGVCF/r62.g.vcf.gz",
         I63  = "interGVCF/r63.g.vcf.gz",
         I64  = "interGVCF/r64.g.vcf.gz",
         I65  = "interGVCF/r65.g.vcf.gz",
         I66  = "interGVCF/r66.g.vcf.gz",
         I67  = "interGVCF/r67.g.vcf.gz",
         I68  = "interGVCF/r68.g.vcf.gz",
         I69  = "interGVCF/r69.g.vcf.gz",
         I70  = "interGVCF/r70.g.vcf.gz",
         I71  = "interGVCF/r71.g.vcf.gz",
         I72  = "interGVCF/r72.g.vcf.gz",
         I73  = "interGVCF/r73.g.vcf.gz",
         I74  = "interGVCF/r74.g.vcf.gz",
         I75  = "interGVCF/r75.g.vcf.gz",
         I76  = "interGVCF/r76.g.vcf.gz",
         I77  = "interGVCF/r77.g.vcf.gz",
         I78  = "interGVCF/r78.g.vcf.gz",
         I79  = "interGVCF/r79.g.vcf.gz",
         I80  = "interGVCF/r80.g.vcf.gz",
         I81  = "interGVCF/r81.g.vcf.gz",
         I82  = "interGVCF/r82.g.vcf.gz",
         I83  = "interGVCF/r83.g.vcf.gz",
         I84  = "interGVCF/r84.g.vcf.gz",
         I85  = "interGVCF/r85.g.vcf.gz",
         I86  = "interGVCF/r86.g.vcf.gz",
         I87  = "interGVCF/r87.g.vcf.gz",
         I88  = "interGVCF/r88.g.vcf.gz",
         I89  = "interGVCF/r89.g.vcf.gz",
         I90  = "interGVCF/r90.g.vcf.gz",
         I91  = "interGVCF/r91.g.vcf.gz",
         I92  = "interGVCF/r92.g.vcf.gz",
         I93  = "interGVCF/r93.g.vcf.gz",
         I94  = "interGVCF/r94.g.vcf.gz",
         I95  = "interGVCF/r95.g.vcf.gz",
         I96  = "interGVCF/r96.g.vcf.gz"          # Or by using a Tab-delimited text file with sample_name--tab--path_to_sample_vcf per line.
     output:
         directory("{chr_num}_database")# Function directory() is used to mark
"{chr_num_databae}" as a directory.
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk --java-options '-Xmx8g -Xms8g'
GenomicsDBImport -V {input.I1} -V {input.I2} -V {input.I3} -V {input.I4} -V
{input.I5} -V {input.I6} -V {input.I7} -V {input.I8} -V {input.I9} -V {input.I10} -V
{input.I11} -V {input.I12} -V {input.I13} -V {input.I14} -V {input.I15} -V
{input.I16} -V {input.I17} -V {input.I18} -V {input.I19} -V {input.I20} -V
{input.I21} -V {input.I22} -V {input.I23} -V {input.I24} \
        -V {input.I25} -V {input.I26} -V {input.I27} -V {input.I28} -V {input.I29}
-V {input.I30} -V {input.I31} -V {input.I32} -V {input.I33} -V {input.I34} -V
{input.I35} -V {input.I36} -V {input.I37} -V {input.I38} -V {input.I39} -V
{input.I40} -V {input.I41} -V {input.I42} -V {input.I43} -V {input.I44} -V
{input.I45} -V {input.I46} -V {input.I47} \
        -V {input.I48} -V {input.I49} -V {input.I50} -V {input.I51} -V {input.I52}
-V {input.I53} -V {input.I54} -V {input.I55} -V {input.I56} -V {input.I57} -V
{input.I58} -V {input.I59} -V {input.I60} -V {input.I61} -V {input.I62} -V
{input.I63} -V {input.I64} -V {input.I65} -V {input.I66} -V {input.I67} -V
{input.I68} -V {input.I69} -V {input.I70} \
        -V {input.I71} -V {input.I72} -V {input.I73} -V {input.I74} -V {input.I75}
-V {input.I76} -V {input.I77} -V {input.I78} -V {input.I79} -V {input.I80} -V
{input.I81} -V {input.I82} -V {input.I83} -V {input.I84} -V {input.I85} -V
{input.I86} -V {input.I87} -V {input.I88} -V {input.I89} -V {input.I90} -V
{input.I91} -V {input.I92} -V {input.I93} \
        -V {input.I94} -V {input.I95} -V {input.I96} --genomicsdb-workspace-path
{output} --batch-size 96 -L SL2.50{wildcards.chr_num} --reader-threads 6"

rule chr_db_to_vcf:
    input:
        directory("{chr_num}_database")
    output:
        "Joint_vcf/{chr_num}.vcf.gz"
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk --java-options '-Xmx8g'
GenotypeGVCFs -R ref_genome/S_lycopersicum_chromosomes.2.50.fa -V
gendb://{input} -O {output}"

rule integrate_all_chr_vcf:
    input:
        chro00 = "Joint_vcf/ch00.vcf.gz",
        chro01 = "Joint_vcf/ch01.vcf.gz",
        chro02 = "Joint_vcf/ch02.vcf.gz",
        chro03 = "Joint_vcf/ch03.vcf.gz",
        chro04 = "Joint_vcf/ch04.vcf.gz",
        chro05 = "Joint_vcf/ch05.vcf.gz",
        chro06 = "Joint_vcf/ch06.vcf.gz",
        chro07 = "Joint_vcf/ch07.vcf.gz",
        chro08 = "Joint_vcf/ch08.vcf.gz",
        chro09 = "Joint_vcf/ch09.vcf.gz",
        chro10 = "Joint_vcf/ch10.vcf.gz",
        chro11 = "Joint_vcf/ch11.vcf.gz",
        chro12 = "Joint_vcf/ch12.vcf.gz"
    output:
        "Gather_vcf/all.vcf.gz"
    run:
        shell("java -jar picard.jar GatherVcfs I={input.chro00} I={input.chro01}
I={input.chro02} I={input.chro03} I={input.chro04} I={input.chro05}
I={input.chro06} I={input.chro07} I={input.chro08} I={input.chro09}
I={input.chro10} I={input.chro11} I={input.chro12} O={output}")
        shell("/home/poweruser/gatk-4.0.11.0/./gatk IndexFeatureFile -F {output}")
#Additional step for index the all.vcf.gz.

rule select_vcf:
    input:
        "Gather_vcf/all.vcf.gz"
    output:
        "Sel_Var/selected_all.vcf"
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk SelectVariants -R
ref_genome/S_lycopersicum_chromosomes.2.50.fa -V {input}
--select-type-to-include SNP -O {output}"

rule filter_all_vcf:
    input:
        "Sel_Var/selected_all.vcf"
    output:
        "Filter/filtered.all.vcf.gz"
    shell:
        "/home/poweruser/gatk-4.0.11.0/./gatk VariantFiltration -R
ref_genome/S_lycopersicum_chromosomes.2.50.fa -V {input} -O {output}
--filter-expression 'AB < 0.2 || MQ0 > 50' --filter-name 'my_filters'"
```

Appendix 4 - GWAS (Genome-wide association study)

In this documentation, we would like to demonstrate the process GWAS (Genome-wide association study) analysis using TASSEL 5.0. TASSEL 5.0 is a software to evaluate traits associations, evolutionary patterns, and linkage disequilibrium. User could select a general linear model (GLM) or mixed linear model (MLM) for associate analysis. While addressing complex pedigrees, families, founding effects and population structure, MLM is capable of reducing the Type I error in association mapping. We suggest user should engage all models to GWAS and evaluate the adequacy and performance of each models by plotting QQ-plot (Quantile-Quantile plot). After deciding a reasonable model for GWAS, then it is more confident for you to choose the markers that highly associate to specific trait based on Manhattan plot.

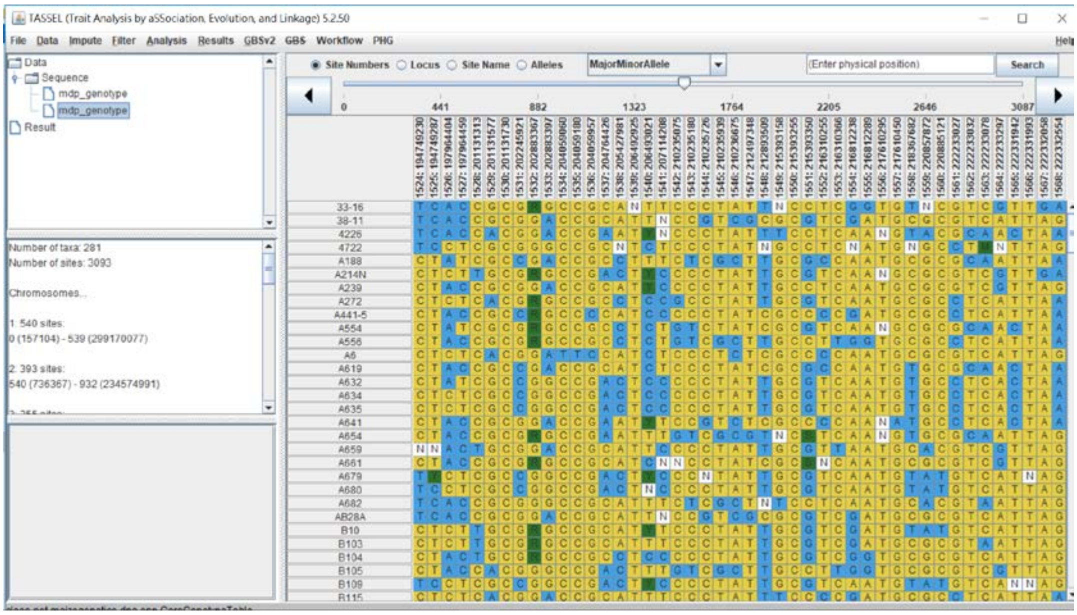
1. GLM

Step 1:

Go to http://www.maizegenetics.net/tassel. Download TASSEL 5.0 and Tassel Tutorial Data.

Step 2:

To perform GWAS base on GLM, we need genotype data (SNPs and their position on chromosome for each taxa) and corresponding phenotype data (result of traits evaluation, EarHT: ear height, dpoII quantification, EarDia: ear diameter). TASSEL could read the genotype data in three different file format, .hmp, .map, .ped. Result of genotype data (mdp_genotype.map, mdp_genotype.ped) input should look like figure below:


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.50
File Data Impute Filter Analysis Results GBSv2 GBS Workflow PHG
Data
Sequence
mdp_genotype
mdp_genotype
Result
● Site Numbers ○ Locus ○ Site Name ○ Alleles
MajorMinorAllele
(Enter physical position)
Search
0 441 882 1323 1764 2205 2646 3087

Number of taxa: 281
Number of sites: 3093

Chromosomes...

1: 540 sites:
0 (157104) - 539 (299170077)

2: 393 sites:
540 (736367) - 932 (234574991)

Result of phenotype data (mdp_trait.text) input should look like figure below:

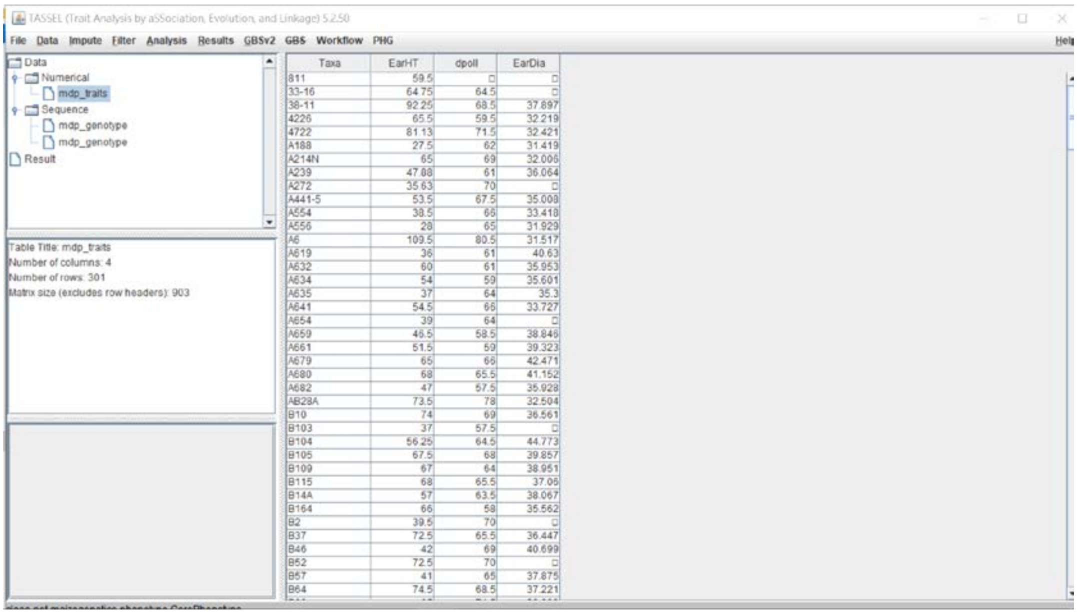
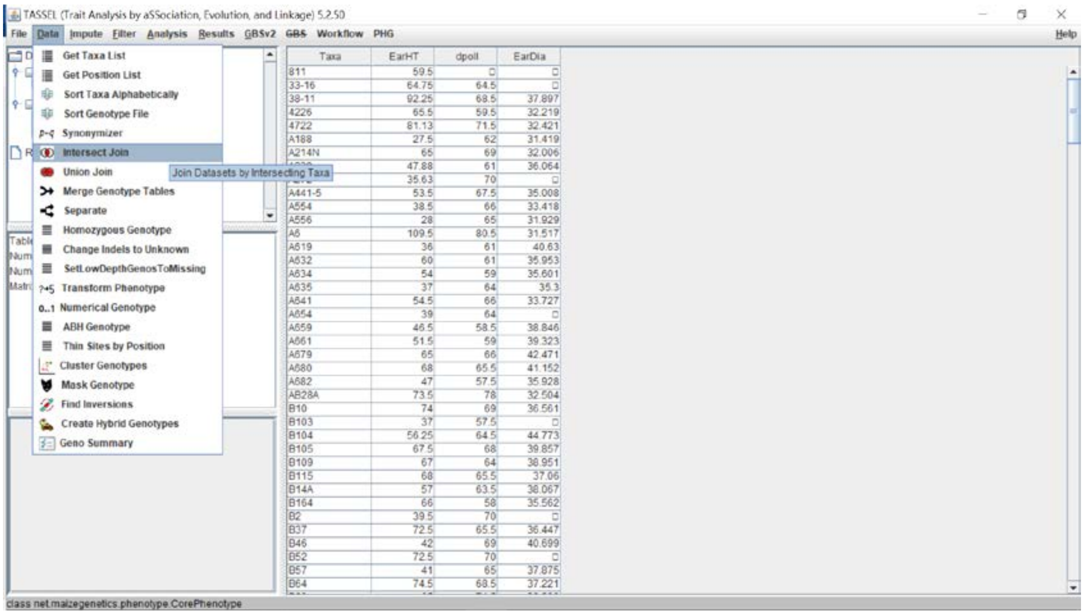

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.50
File Data Impute Filter Analysis Results GBSv2 GBS Workflow PHG
Data
Numerical
mdp_traits
Sequence
mdp_genotype
mdp_genotype
Result

Table Title: mdp_traits
Number of columns: 4
Number of rows: 301
Matrix size (excludes row headers): 903

Taxa	EarHT	dpoll	EarDia
811	59.5	□	□
33-16	64.75	64.5	□
38-11	92.25	68.5	37.897
4226	65.5	59.5	32.219
4722	81.13	71.5	32.421
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A272	35.63	70	□
A441-5	53.5	67.5	35.008
A554	38.5	66	33.418
A556	28	65	31.929
A6	109.5	80.5	31.517
A619	36	61	40.63
A632	60	61	35.953
A634	54	59	35.601
A635	37	64	35.3
A641	54.5	66	33.727
A654	39	64	□
A659	46.5	58.5	38.846
A661	51.5	59	39.323
A679	65	66	42.471
A680	68	65.5	41.152
A682	47	57.5	35.928
AB28A	73.5	78	32.504
B10	74	69	36.561
B103	37	57.5	□
B104	56.25	64.5	44.773
B105	67.5	68	39.857
B109	67	64	38.951
B115	68	65.5	37.06
B14A	57	63.5	38.067
B164	66	58	35.562
B2	39.5	70	□
B37	72.5	65.5	36.447
B46	42	69	40.699
B52	72.5	70	□
B57	41	65	37.875
B64	74.5	68.5	37.221

Step 3:

In this step, we joint genotype and phenotype datasets by taxa then create a file (mdp_trait + mdp_genotype). Select “mdp_traits” and “data_genotype”, and click Data Tab -> Intersect Join.



Taxa	EarHT	dpoll	EarDia
811	59.5	□	□
33-16	64.75	64.5	□
38-11	92.25	68.5	37.897
4226	65.5	59.5	32.219
4722	81.13	71.5	32.421
A188	27.5	62	31.419
A214N	65	69	32.006
[illegible]	47.88	61	36.064
[illegible]	35.63	70	□
A441-5	53.5	67.5	35.008
A554	38.5	66	33.418
A556	28	65	31.929
A6	109.5	80.5	31.517
A619	36	61	40.63
A632	60	61	35.953
A634	54	59	35.601
A635	37	64	35.3
A641	54.5	66	33.727
A654	39	64	□
A659	46.5	58.5	38.846
A661	51.5	59	39.323
A679	65	66	42.471
A680	68	65.5	41.152
A682	47	57.5	35.928
AB28A	73.5	78	32.504
B10	74	69	36.561
B103	37	57.5	□
B104	56.25	64.5	44.773
B105	67.5	68	39.857
B109	67	64	38.951
B115	68	65.5	37.06
B14A	57	63.5	38.067
B164	66	58	35.562
B2	39.5	70	□
B37	72.5	65.5	36.447
B46	42	69	40.699
B52	72.5	70	□
B57	41	65	37.875
B64	74.5	68.5	37.221

class net.maizegenetics.phenotype.CorePhenotype

Step 4:

Fitting a generalized linear model. Select (mdp_trait + mdp_genotype) file, then click Analysis Tab -> Association -> GLM -> OK.

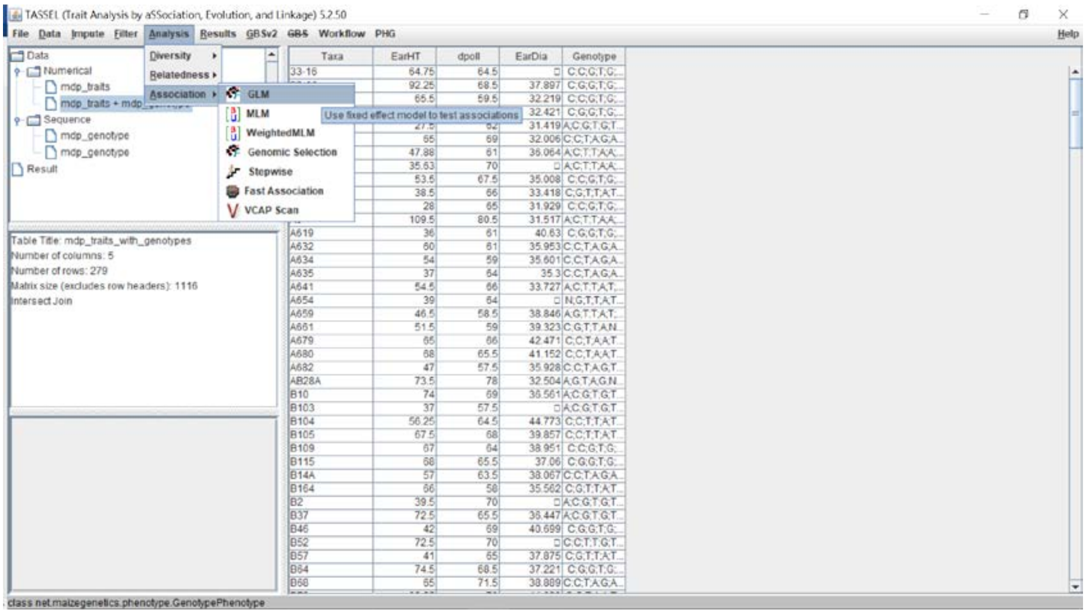


Table Title: mdp_traits_with_genotypes
Number of columns: 5
Number of rows: 279
Matrix size (excludes row headers): 1116
Intersect Join

Taxa	EarHT	dpoll	EarDia	Genotype
33-16	64.75	64.5	□	C;C;G;T;G;...
[illegible]	92.25	68.5	37.897	C;G;G;T;G;...
[illegible]	65.5	59.5	32.219	C;C;G;T;G;...
[illegible]	[illegible]	[illegible]	32.421	C;G;G;T;G;...
[illegible]	27.5	62	31.419	A;C;G;T;G;T...
[illegible]	65	69	32.006	C;C;T;A;G;A...
[illegible]	47.88	61	36.064	A;C;T;T;A;A;...
[illegible]	35.63	70	□	A;C;T;T;A;A;...
[illegible]	53.5	67.5	35.008	C;C;G;T;G;...
[illegible]	38.5	66	33.418	C;G;T;T;A;T...
[illegible]	28	65	31.929	C;C;G;T;G;...
[illegible]	109.5	80.5	31.517	A;C;T;T;A;A;...
A619	36	61	40.63	C;G;G;T;G;...
A632	60	61	35.953	C;C;T;A;G;A...
A634	54	59	35.601	C;C;T;A;G;A...
A635	37	64	35.3	C;C;T;A;G;A...
A641	54.5	66	33.727	A;C;T;T;A;T;...
A654	39	64	□	N;G;T;T;A;T...
A659	46.5	58.5	38.846	A;G;T;T;A;T;...
A661	51.5	59	39.323	C;G;T;T;A;N...
A679	65	66	42.471	C;C;T;A;A;T...
A680	68	65.5	41.152	C;C;T;A;A;T...
A682	47	57.5	35.928	C;C;T;A;G;T...
AB28A	73.5	78	32.504	A;G;T;A;G;N...
B10	74	69	36.561	A;C;G;T;G;T...
B103	37	57.5	□	A;C;G;T;G;T...
B104	56.25	64.5	44.773	C;C;T;T;A;T...
B105	67.5	68	39.857	C;C;T;T;A;T...
B109	67	64	38.951	C;C;G;T;G;...
B115	68	65.5	37.06	C;G;G;T;G;...
B14A	57	63.5	38.067	C;C;T;A;G;A...
B164	66	58	35.562	C;G;T;T;A;T...
B2	39.5	70	□	A;C;G;T;G;T...
B37	72.5	65.5	36.447	A;C;G;T;G;T...
B46	42	69	40.699	C;G;G;T;G;...
B52	72.5	70	□	C;C;T;T;G;T...
B57	41	65	37.875	C;G;T;T;A;T...
B64	74.5	68.5	37.221	C;G;G;T;G;...
B68	65	71.5	38.889	C;C;T;A;G;A...

class net.maizegenetics.phenotype.GenotypePhenotype

Step 5:

Select (GLM_Stats_mdp_traits + mdp_genotype) file generated from step 4, then click Results Tab -> Association -> QQ plot or Manhattan plot.

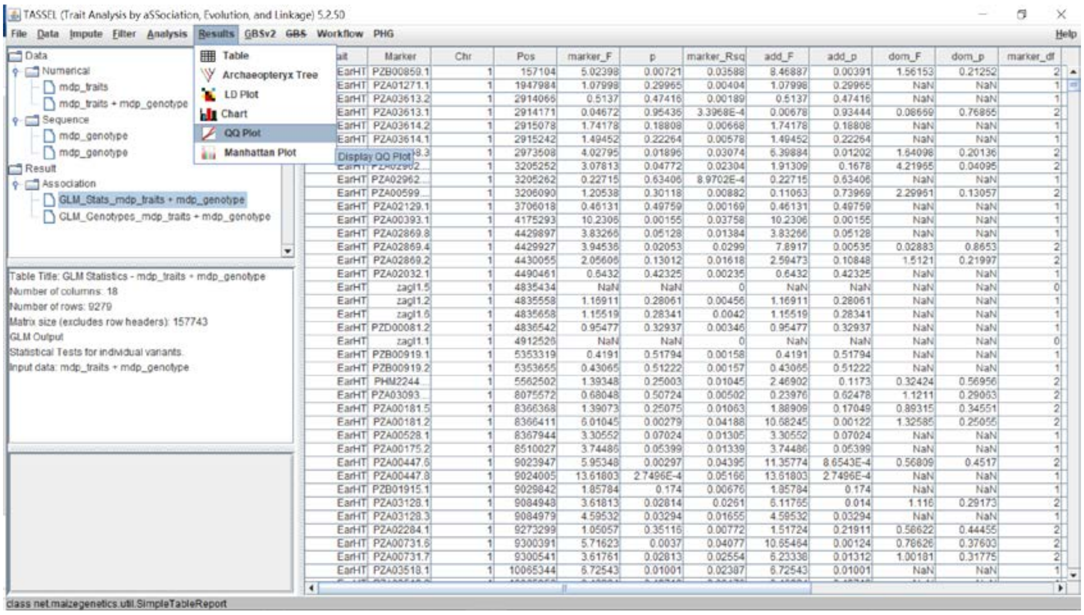
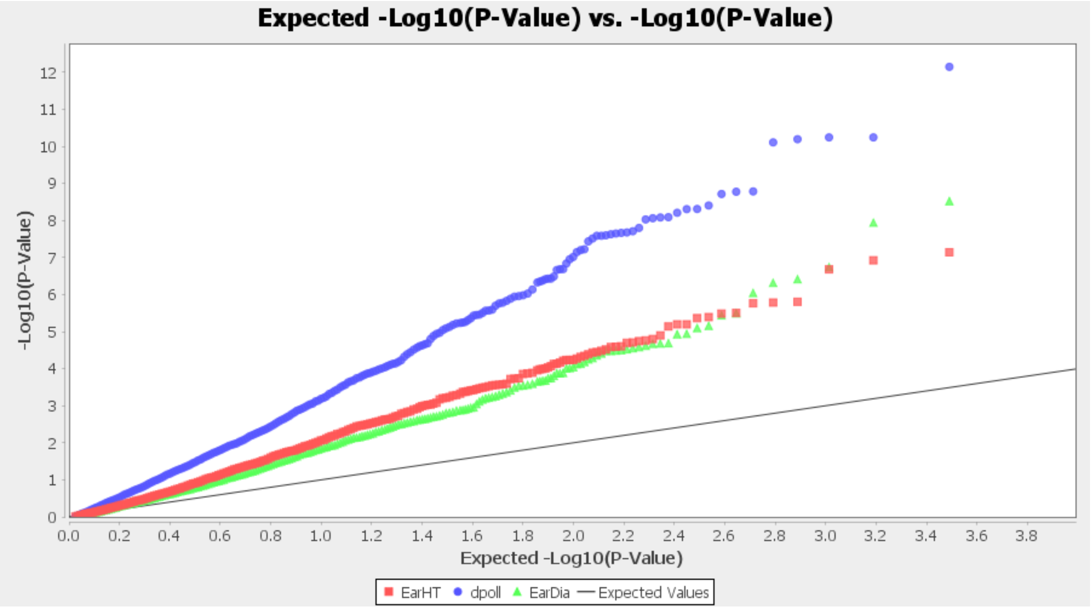

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.50
File
Data
Impute
Filter
Analysis
Results
GBSv2
GBS
Workflow
PHG
Help
Data
Numerical
mdp_traits
mdp_traits + mdp_genotype
Sequence
mdp_genotype
mdp_genotype
Result
Association
GLM_Stats_mdp_traits + mdp_genotype
GLM_Genotypes_mdp_traits + mdp_genotype
Table
Archaeopteryx Tree
LD Plot
Chart
QQ Plot
Manhattan Plot
Display QQ Plot
class net.maizegenetics.util.SimpleTableReport

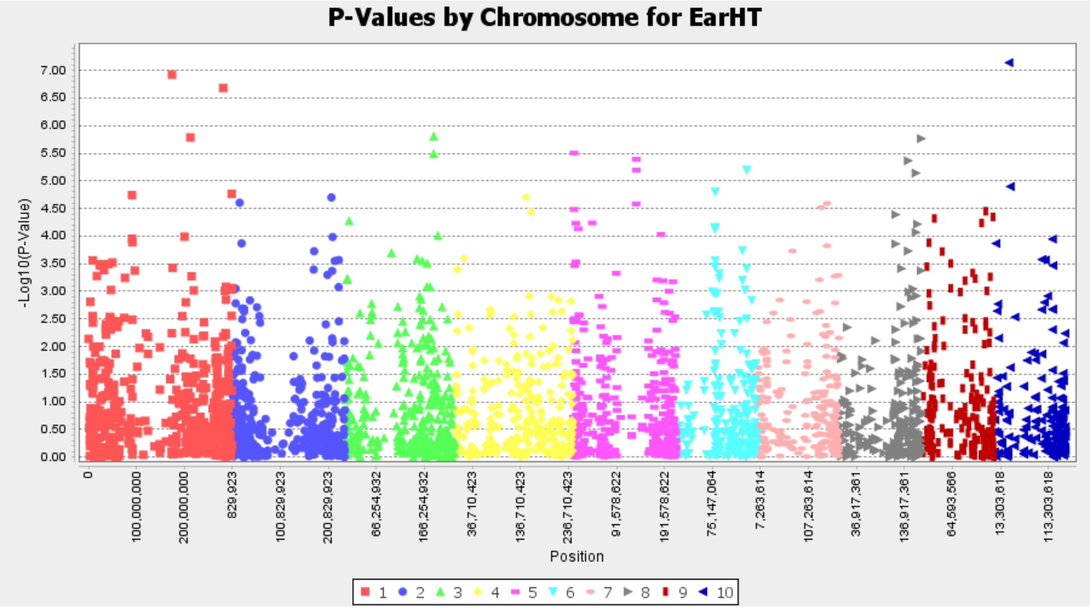
Table Title: GLM Statistics - mdp_traits + mdp_genotype
Number of columns: 18
Number of rows: 9279
Matrix size (excludes row headers): 157743
GLM Output
Statistical Tests for individual variants.
Input data: mdp_traits + mdp_genotype

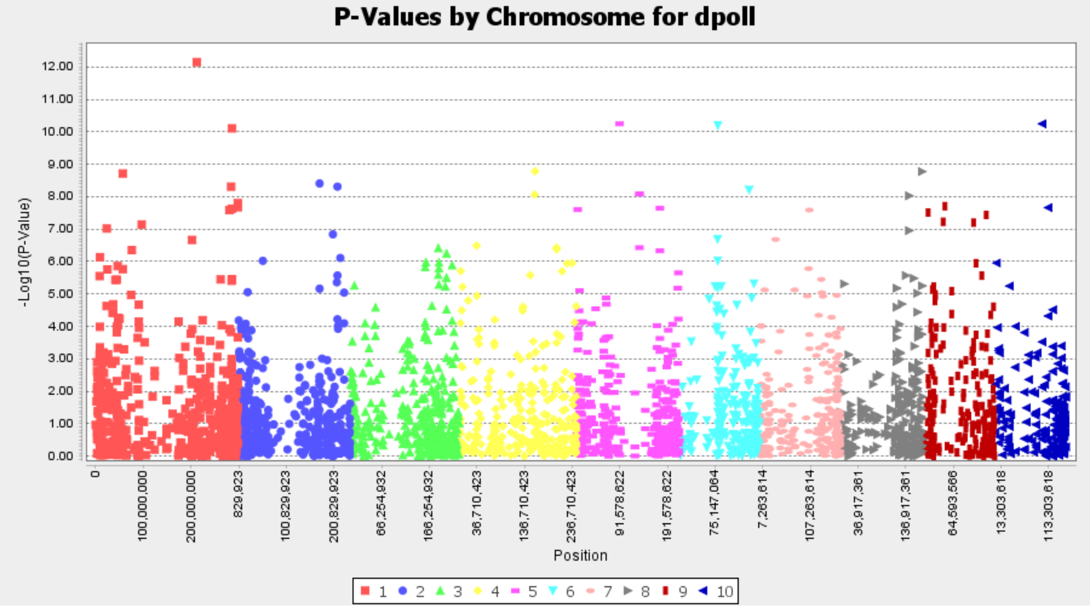
ait	Marker	Chr	Pos	marker_F	p	marker_Rsq	add_F	add_p	dom_F	dom_p	marker_df
EarHT	PZB00859.1	1	157104	5.02398	0.00721	0.03588	8.46887	0.00391	1.56153	0.21252	2
EarHT	PZA01271.1	1	1947984	1.07998	0.29965	0.00404	1.07998	0.29965	NaN	NaN	1
EarHT	PZA03613.2	1	2914066	0.5137	0.47416	0.00189	0.5137	0.47416	NaN	NaN	1
EarHT	PZA03613.1	1	2914171	0.04672	0.95436	3.3968E-4	0.00678	0.93444	0.08659	0.76865	2
EarHT	PZA03614.2	1	2915078	1.74178	0.18808	0.00668	1.74178	0.18808	NaN	NaN	1
EarHT	PZA03614.1	1	2915242	1.49452	0.22264	0.00578	1.49452	0.22264	NaN	NaN	1
	8.3	1	2973508	4.02795	0.01896	0.03074	6.39884	0.01202	1.64098	0.20136	2
EarHT	PZA02962...	1	3205252	3.07813	0.04772	0.02304	1.91309	0.1678	4.21965	0.04095	2
EarHT	PZA02962...	1	3205262	0.22715	0.63406	8.9702E-4	0.22715	0.63406	NaN	NaN	1
EarHT	PZA00599...	1	3206090	1.20538	0.30118	0.00882	0.11063	0.73969	2.29951	0.13057	2
EarHT	PZA02129.1	1	3706018	0.46131	0.49759	0.00169	0.46131	0.49759	NaN	NaN	1
EarHT	PZA00393.1	1	4175293	10.2306	0.00155	0.03758	10.2306	0.00155	NaN	NaN	1
EarHT	PZA02869.8	1	4429897	3.83266	0.05128	0.01384	3.83266	0.05128	NaN	NaN	1
EarHT	PZA02869.4	1	4429927	3.94536	0.02053	0.0299	7.8917	0.00535	0.02883	0.8653	2
EarHT	PZA02869.2	1	4430055	2.05606	0.13012	0.01618	2.59473	0.10848	1.5121	0.21997	2
EarHT	PZA02032.1	1	4490461	0.6432	0.42325	0.00235	0.6432	0.42325	NaN	NaN	1
EarHT	zagl1.5	1	4835434	NaN	NaN	0	NaN	NaN	NaN	NaN	0
EarHT	zagl1.2	1	4835558	1.16911	0.28061	0.00456	1.16911	0.28061	NaN	NaN	1
EarHT	zagl1.6	1	4835658	1.15519	0.28341	0.0042	1.15519	0.28341	NaN	NaN	1
EarHT	PZD00081.2	1	4836542	0.95477	0.32937	0.00346	0.95477	0.32937	NaN	NaN	1
EarHT	zagl1.1	1	4912526	NaN	NaN	0	NaN	NaN	NaN	NaN	0
EarHT	PZB00919.1	1	5353319	0.4191	0.51794	0.00158	0.4191	0.51794	NaN	NaN	1
EarHT	PZB00919.2	1	5353655	0.43065	0.51222	0.00157	0.43065	0.51222	NaN	NaN	1
EarHT	PHM2244...	1	5562502	1.39348	0.25003	0.01045	2.46902	0.1173	0.32424	0.56956	2
EarHT	PZA03093...	1	8075572	0.68048	0.50724	0.00502	0.23976	0.62478	1.1211	0.29063	2
EarHT	PZA00181.5	1	8366368	1.39073	0.25075	0.01063	1.88909	0.17049	0.89315	0.34551	2
EarHT	PZA00181.2	1	8366411	6.01045	0.00279	0.04188	10.58245	0.00122	1.32585	0.25055	2
EarHT	PZA00528.1	1	8367944	3.30552	0.07024	0.01305	3.30552	0.07024	NaN	NaN	1
EarHT	PZA00175.2	1	8510027	3.74486	0.05399	0.01339	3.74486	0.05399	NaN	NaN	1
EarHT	PZA00447.6	1	9023947	5.95348	0.00297	0.04395	11.35774	8.6543E-4	0.56809	0.4517	2
EarHT	PZA00447.8	1	9024005	13.61803	2.7496E-4	0.05166	13.61803	2.7496E-4	NaN	NaN	1
EarHT	PZB01915.1	1	9029842	1.85784	0.174	0.00676	1.85784	0.174	NaN	NaN	1
EarHT	PZA03128.1	1	9084948	3.61813	0.02814	0.0261	6.11765	0.014	1.116	0.29173	2
EarHT	PZA03128.3	1	9084979	4.59532	0.03294	0.01655	4.59532	0.03294	NaN	NaN	1
EarHT	PZA02284.1	1	9273299	1.05057	0.35116	0.00772	1.51724	0.21911	0.58622	0.44455	2
EarHT	PZA00731.6	1	9300391	5.71623	0.0037	0.04077	10.65464	0.00124	0.78626	0.37603	2
EarHT	PZA00731.7	1	9300541	3.61761	0.02813	0.02554	6.23338	0.01312	1.00181	0.31775	2
EarHT	PZA03518.1	1	10065344	6.72543	0.01001	0.02387	6.72543	0.01001	NaN	NaN	1

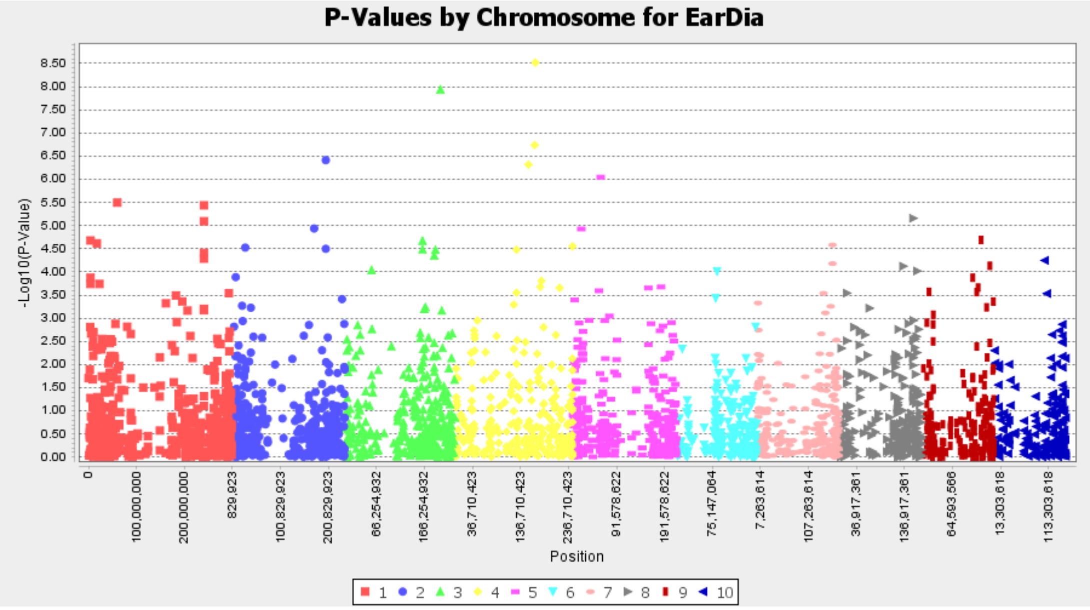
Step 6:

Check the results of QQ plot and Manhattan plots for 3 different traits, EarHT, dpoII and EarDia.




P-Values by Chromosome for EarHT
-Log10(P-Value)
7.00
6.50
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0
100,000,000
200,000,000
829,923
100,829,923
200,829,923
66,254,932
166,254,932
36,710,423
136,710,423
236,710,423
91,578,622
191,578,622
75,147,064
7,263,614
107,263,614
36,917,361
136,917,361
64,593,566
13,303,618
113,303,618
Position
1
2
3
4
5
6
7
8
9
10


P-Values by Chromosome for dpoll
-Log10(P-Value)
12.00
11.00
10.00
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
0
100,000,000
200,000,000
829,923
100,829,923
200,829,923
66,254,932
166,254,932
36,710,423
136,710,423
236,710,423
91,578,622
191,578,622
75,147,064
7,263,614
107,263,614
36,917,361
136,917,361
64,593,566
13,303,618
113,303,618
Position
1
2
3
4
5
6
7
8
9
10



2. MLM

Step 1:

We first create a population structure file but includes only two covariate:

Open and select "mdp_population_structure.text", then Filter Tab > Traits > Deselect Q3 > OK. This step crates a "Filtered_mdp_population_structure" file.

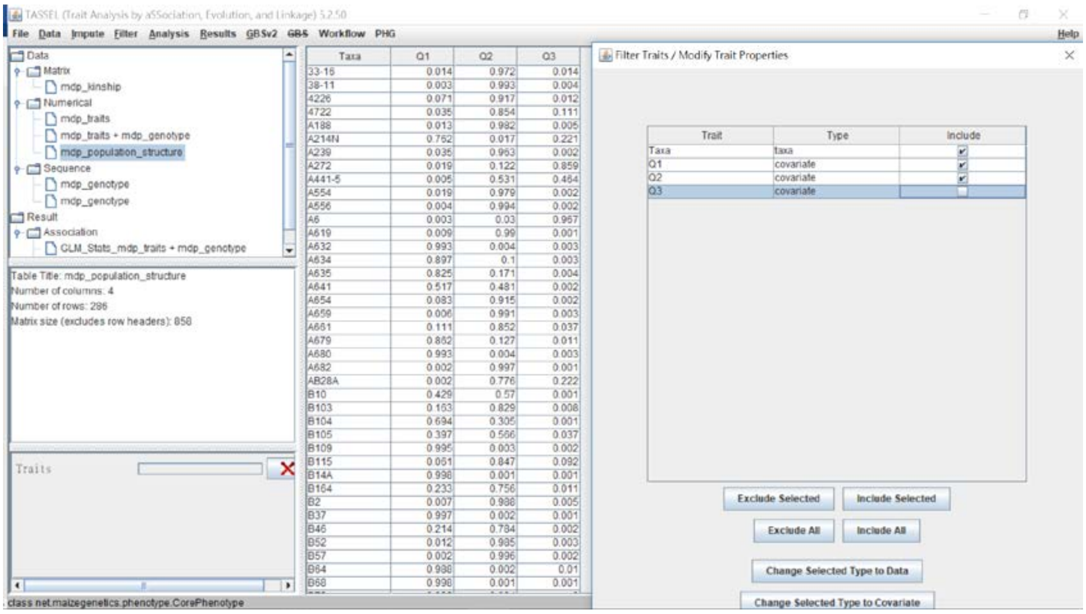


Table Title: mdp_population_structure
Number of columns: 4
Number of rows: 286
Matrix size (excludes row headers): 858

Taxa	Q1	Q2	Q3
33-16	0.014	0.972	0.014
38-11	0.003	0.993	0.004
4226	0.071	0.917	0.012
4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
A214N	0.762	0.017	0.221
A239	0.035	0.963	0.002
A272	0.019	0.122	0.859
A441-5	0.005	0.531	0.464
A554	0.019	0.979	0.002
A556	0.004	0.994	0.002
A6	0.003	0.03	0.967
A619	0.009	0.99	0.001
A632	0.993	0.004	0.003
A634	0.897	0.1	0.003
A635	0.825	0.171	0.004
A641	0.517	0.481	0.002
A654	0.083	0.915	0.002
A659	0.006	0.991	0.003
A661	0.111	0.852	0.037
A679	0.862	0.127	0.011
A680	0.993	0.004	0.003
A682	0.002	0.997	0.001
AB28A	0.002	0.776	0.222
B10	0.429	0.57	0.001
B103	0.163	0.829	0.008
B104	0.694	0.305	0.001
B105	0.397	0.566	0.037
B109	0.995	0.003	0.002
B115	0.061	0.847	0.092
B14A	0.998	0.001	0.001
B164	0.233	0.756	0.011
B2	0.007	0.988	0.005
B37	0.997	0.002	0.001
B46	0.214	0.784	0.002
B52	0.012	0.985	0.003
B57	0.002	0.996	0.002
B64	0.988	0.002	0.01
B68	0.998	0.001	0.001

Trait	Type	Include
Taxa	taxa	☑
Q1	covariate	☑
Q2	covariate	☑
Q3	covariate	□

class net.maizegenetics.phenotype.CorePhenotype

Step 2:

Select "mdp_traits", "data_genotype" and "Filtered_mdp_population_structure", and click Data Tab -> Intersect Join. This step creates a "mdp_traits + data_genotype + Filtered_mdp_population_structure" file.

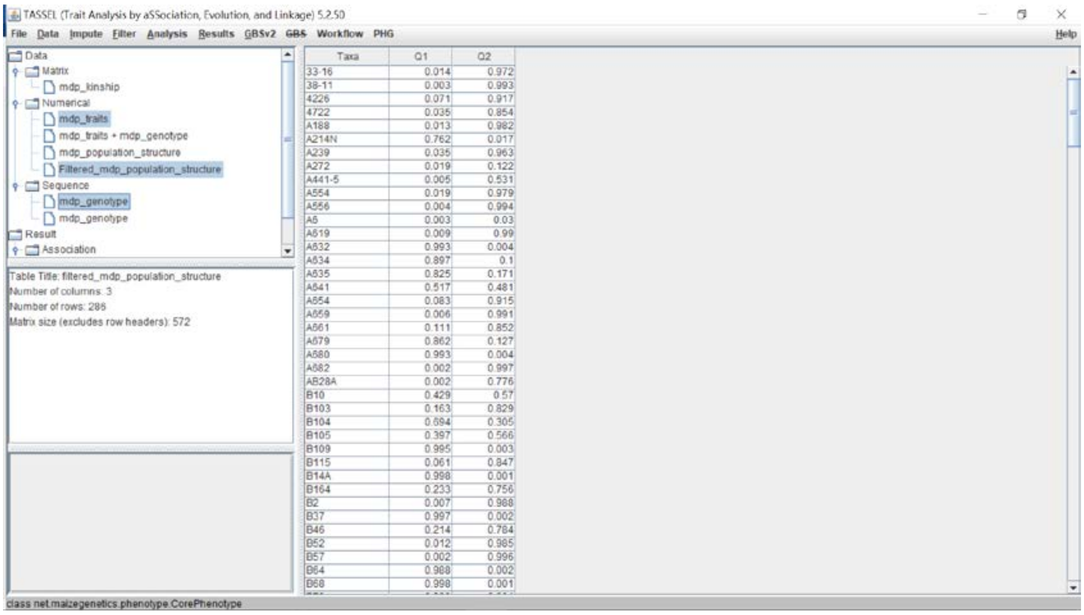


Table Title: filtered_mdp_population_structure
Number of columns: 3
Number of rows: 286
Matrix size (excludes row headers): 572

Taxa	Q1	Q2
33-16	0.014	0.972
38-11	0.003	0.993
4226	0.071	0.917
4722	0.035	0.854
A188	0.013	0.982
A214N	0.762	0.017
A239	0.035	0.963
A272	0.019	0.122
A441-5	0.005	0.531
A554	0.019	0.979
A556	0.004	0.994
A6	0.003	0.03
A619	0.009	0.99
A632	0.993	0.004
A634	0.897	0.1
A635	0.825	0.171
A641	0.517	0.481
A654	0.083	0.915
A659	0.006	0.991
A661	0.111	0.852
A679	0.862	0.127
A680	0.993	0.004
A682	0.002	0.997
AB28A	0.002	0.776
B10	0.429	0.57
B103	0.163	0.829
B104	0.694	0.305
B105	0.397	0.566
B109	0.995	0.003
B115	0.061	0.847
B14A	0.998	0.001
B164	0.233	0.756
B2	0.007	0.988
B37	0.997	0.002
B46	0.214	0.784
B52	0.012	0.985
B57	0.002	0.996
B64	0.988	0.002
B68	0.998	0.001

class net.maizegenetics.phenotype.CorePhenotype

Step 3:

Select the “mdp_traits + data_genotype + Filtered_mdp_population_structure” file and “kin_data_hapmap”, then Analysis -> MLM -> Run

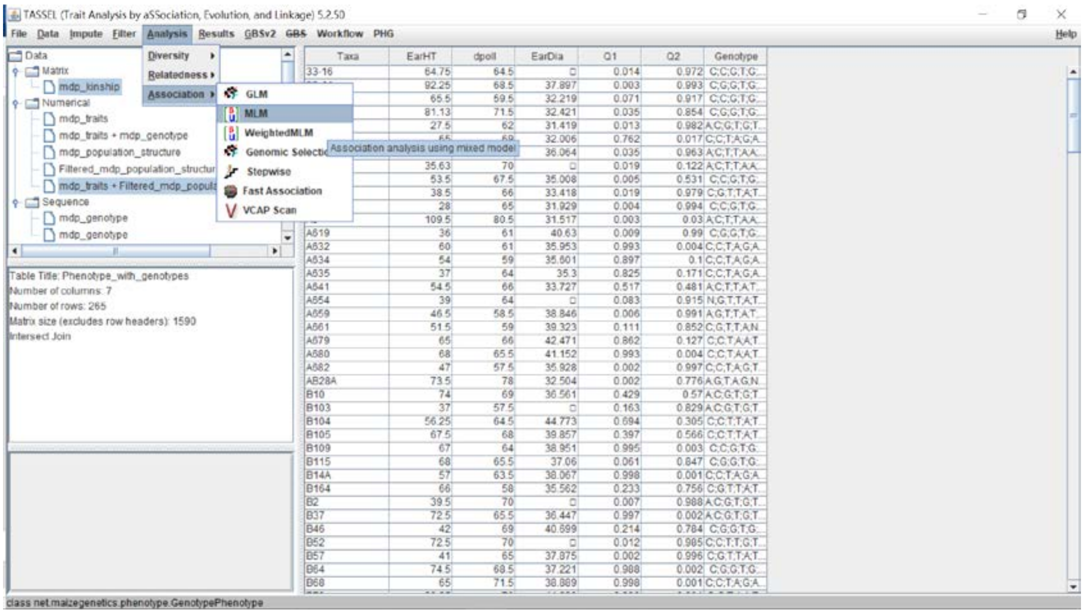


Table Title: Phenotype_with_genotypes
Number of columns: 7
Number of rows: 265
Matrix size (excludes row headers): 1590
Intersect Join

Taxa	EarHT	dpoll	EarDia	Q1	Q2	Genotype
33-16	64.75	64.5	□	0.014	0.972	C;C;G;T;G...
	92.25	68.5	37.897	0.003	0.993	C;G;G;T;G...
	65.5	59.5	32.219	0.071	0.917	C;C;G;T;G...
	81.13	71.5	32.421	0.035	0.854	C;G;G;T;G...
	27.5	62	31.419	0.013	0.982	A;C;G;T;G;T...
	65	69	32.006	0.762	0.017	C;C;T;A;G;A...
			36.064	0.035	0.963	A;C;T;T;A;A...
	35.63	70	□	0.019	0.122	A;C;T;T;A;A...
	53.5	67.5	35.008	0.005	0.531	C;C;G;T;G...
	38.5	66	33.418	0.019	0.979	C;G;T;T;A;T...
	28	65	31.929	0.004	0.994	C;C;G;T;G...
	109.5	80.5	31.517	0.003	0.03	A;C;T;T;A;A...
A619	36	61	40.63	0.009	0.99	C;G;G;T;G...
A632	60	61	35.953	0.993	0.004	C;C;T;A;G;A...
A634	54	59	35.601	0.897	0.1	C;C;T;A;G;A...
A635	37	64	35.3	0.825	0.171	C;C;T;A;G;A...
A641	54.5	66	33.727	0.517	0.481	A;C;T;T;A;T...
A654	39	64	□	0.083	0.915	N;G;T;T;A;T...
A659	46.5	58.5	38.846	0.006	0.991	A;G;T;T;A;T...
A661	51.5	59	39.323	0.111	0.852	C;G;T;T;A;N...
A679	65	66	42.471	0.862	0.127	C;C;T;A;A;T...
A680	68	65.5	41.152	0.993	0.004	C;C;T;A;A;T...
A682	47	57.5	35.928	0.002	0.997	C;C;T;A;G;T...
AB28A	73.5	78	32.504	0.002	0.776	A;G;T;A;G;N...
B10	74	69	36.561	0.429	0.57	A;C;G;T;G;T...
B103	37	57.5	□	0.163	0.829	A;C;G;T;G;T...
B104	56.25	64.5	44.773	0.694	0.305	C;C;T;T;A;T...
B105	67.5	68	39.857	0.397	0.566	C;C;T;T;A;T...
B109	67	64	38.951	0.995	0.003	C;C;G;T;G...
B115	68	65.5	37.06	0.061	0.847	C;G;G;T;G...
B14A	57	63.5	38.067	0.998	0.001	C;C;T;A;G;A...
B164	66	58	35.562	0.233	0.756	C;G;T;T;A;T...
B2	39.5	70	□	0.007	0.988	A;C;G;T;G;T...
B37	72.5	65.5	36.447	0.997	0.002	A;C;G;T;G;T...
B46	42	69	40.699	0.214	0.784	C;G;G;T;G...
B52	72.5	70	□	0.012	0.985	C;C;T;T;G;T...
B57	41	65	37.875	0.002	0.996	C;G;T;T;A;T...
B64	74.5	68.5	37.221	0.988	0.002	C;G;G;T;G...
B68	65	71.5	38.889	0.998	0.001	C;C;T;A;G;A...

class net.maizegenetics.phenotype.GenotypePhenotype

Step 4:

Select (MLM_Stats_for_ mdp_traits + data_genotype + Filtered_mdp_population_structure) file generated from step 4, then click Results Tab -> Association -> QQ plot or Manhattan plot.

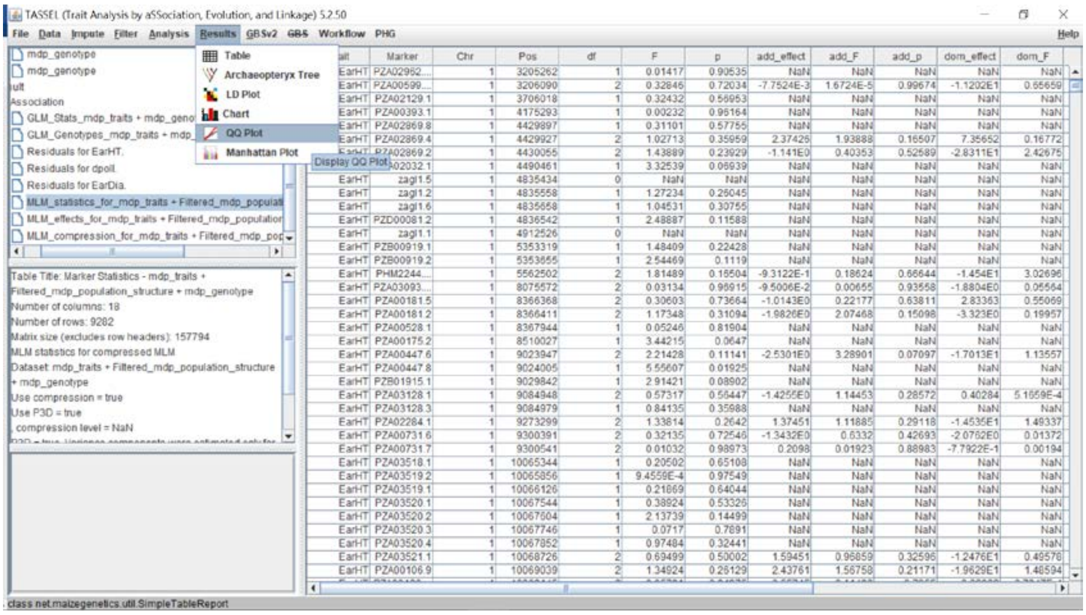
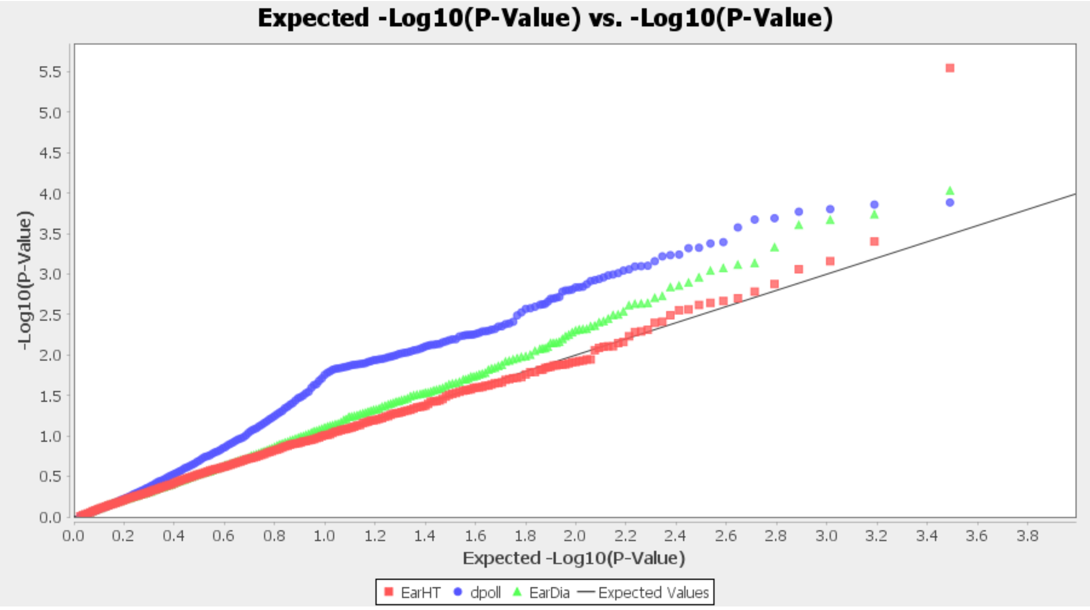

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.50
File
Data
Impute
Filter
Analysis
Results
GBSv2
GBS
Workflow
PHG
Help
Table
Archaeopteryx Tree
LD Plot
Chart
QQ Plot
Manhattan Plot
Display QQ Plot
mdp_genotype
mdp_genotype
ult
Association
GLM_Stats_mdp_traits + mdp_geno
GLM_Genotypes_mdp_traits + mdp_
Residuals for EarHT.
Residuals for dpoll.
Residuals for EarDia.
MLM_statistics_for_mdp_traits + Filtered_mdp_populati
MLM_effects_for_mdp_traits + Filtered_mdp_populatior
MLM_compression_for_mdp_traits + Filtered_mdp_pop
class net.maizegenetics.util.SimpleTableReport

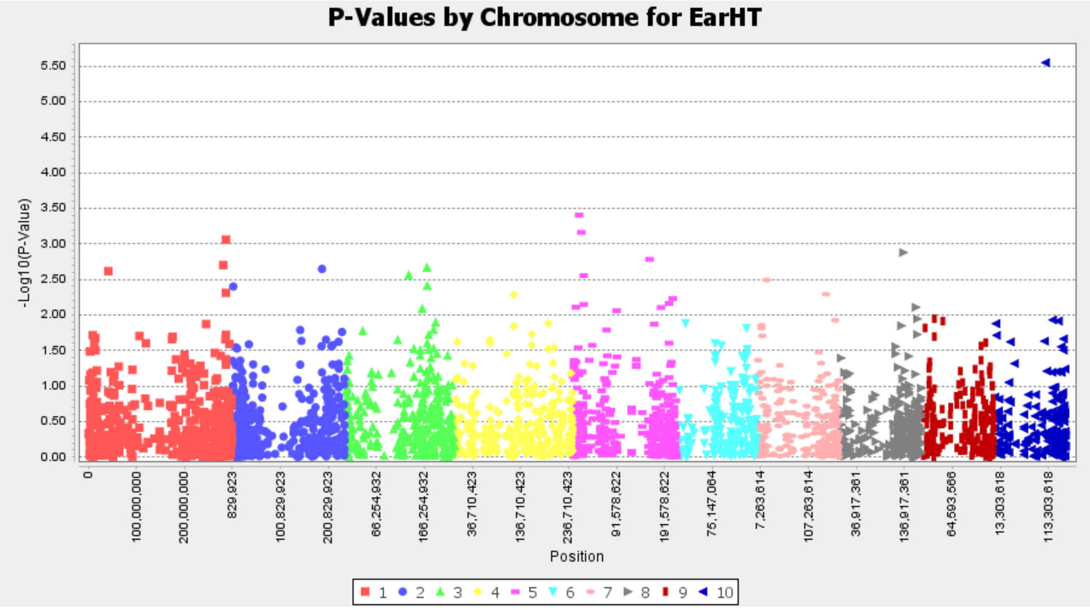
Table Title: Marker Statistics - mdp_traits + Filtered_mdp_population_structure + mdp_genotype
Number of columns: 18
Number of rows: 9282
Matrix size (excludes row headers): 157794
MLM statistics for compressed MLM
Dataset mdp_traits + Filtered_mdp_population_structure + mdp_genotype
Use compression = true
Use P3D = true
, compression level = NaN

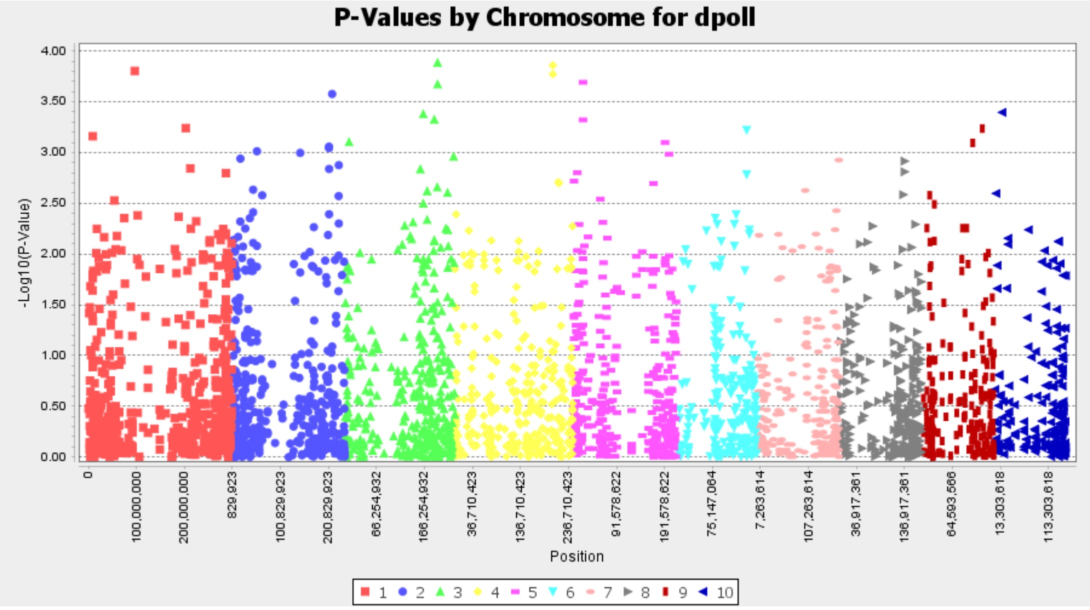
ait	Marker	Chr	Pos	df	F	p	add_effect	add_F	add_p	dom_effect	dom_F
EarHT	PZA02962....	1	3205262	1	0.01417	0.90535	NaN	NaN	NaN	NaN	NaN
EarHT	PZA00599....	1	3206090	2	0.32846	0.72034	-7.7524E-3	1.6724E-5	0.99674	-1.1202E1	0.65659
EarHT	PZA02129.1	1	3706018	1	0.32432	0.56953	NaN	NaN	NaN	NaN	NaN
EarHT	PZA00393.1	1	4175293	1	0.00232	0.96164	NaN	NaN	NaN	NaN	NaN
EarHT	PZA02869.8	1	4429897	1	0.31101	0.57755	NaN	NaN	NaN	NaN	NaN
EarHT	PZA02869.4	1	4429927	2	1.02713	0.35959	2.37426	1.93888	0.16507	7.35652	0.16772
EarHT	PZA02869.2	1	4430055	2	1.43889	0.23929	-1.141E0	0.40353	0.52589	-2.8311E1	2.42675
	02032.1	1	4490461	1	3.32539	0.06939	NaN	NaN	NaN	NaN	NaN
EarHT	zagl1.5	1	4835434	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EarHT	zagl1.2	1	4835558	1	1.27234	0.26045	NaN	NaN	NaN	NaN	NaN
EarHT	zagl1.6	1	4835658	1	1.04531	0.30755	NaN	NaN	NaN	NaN	NaN
EarHT	PZD00081.2	1	4836542	1	2.48887	0.11588	NaN	NaN	NaN	NaN	NaN
EarHT	zagl1.1	1	4912526	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EarHT	PZB00919.1	1	5353319	1	1.48409	0.22428	NaN	NaN	NaN	NaN	NaN
EarHT	PZB00919.2	1	5353655	1	2.54469	0.1119	NaN	NaN	NaN	NaN	NaN
EarHT	PHM2244....	1	5562502	2	1.81489	0.16504	-9.3122E-1	0.18624	0.66644	-1.454E1	3.02696
EarHT	PZA03093....	1	8075572	2	0.03134	0.96915	-9.5006E-2	0.00655	0.93558	-1.8804E0	0.05564
EarHT	PZA00181.5	1	8366368	2	0.30603	0.73664	-1.0143E0	0.22177	0.63811	2.83363	0.55069
EarHT	PZA00181.2	1	8366411	2	1.17348	0.31094	-1.9826E0	2.07468	0.15098	-3.323E0	0.19957
EarHT	PZA00528.1	1	8367944	1	0.05246	0.81904	NaN	NaN	NaN	NaN	NaN
EarHT	PZA00175.2	1	8510027	1	3.44215	0.0647	NaN	NaN	NaN	NaN	NaN
EarHT	PZA00447.6	1	9023947	2	2.21428	0.11141	-2.5301E0	3.28901	0.07097	-1.7013E1	1.13557
EarHT	PZA00447.8	1	9024005	1	5.55607	0.01925	NaN	NaN	NaN	NaN	NaN
EarHT	PZB01915.1	1	9029842	1	2.91421	0.08902	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03128.1	1	9084948	2	0.57317	0.56447	-1.4255E0	1.14453	0.28572	0.40284	5.1659E-4
EarHT	PZA03128.3	1	9084979	1	0.84135	0.35988	NaN	NaN	NaN	NaN	NaN
EarHT	PZA02284.1	1	9273299	2	1.33814	0.2642	1.37451	1.11885	0.29118	-1.4535E1	1.49337
EarHT	PZA00731.6	1	9300391	2	0.32135	0.72546	-1.3432E0	0.6332	0.42693	-2.0762E0	0.01372
EarHT	PZA00731.7	1	9300541	2	0.01032	0.98973	0.2098	0.01923	0.88983	-7.7922E-1	0.00194
EarHT	PZA03518.1	1	10065344	1	0.20502	0.65108	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03519.2	1	10065856	1	9.4559E-4	0.97549	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03519.1	1	10066126	1	0.21869	0.64044	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03520.1	1	10067544	1	0.38924	0.53326	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03520.2	1	10067604	1	2.13739	0.14499	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03520.3	1	10067746	1	0.0717	0.7891	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03520.4	1	10067852	1	0.97484	0.32441	NaN	NaN	NaN	NaN	NaN
EarHT	PZA03521.1	1	10068726	2	0.69499	0.50002	1.59451	0.96859	0.32596	-1.2476E1	0.49576
EarHT	PZA00106.9	1	10069039	2	1.34924	0.26129	2.43761	1.56758	0.21171	-1.9629E1	1.48594

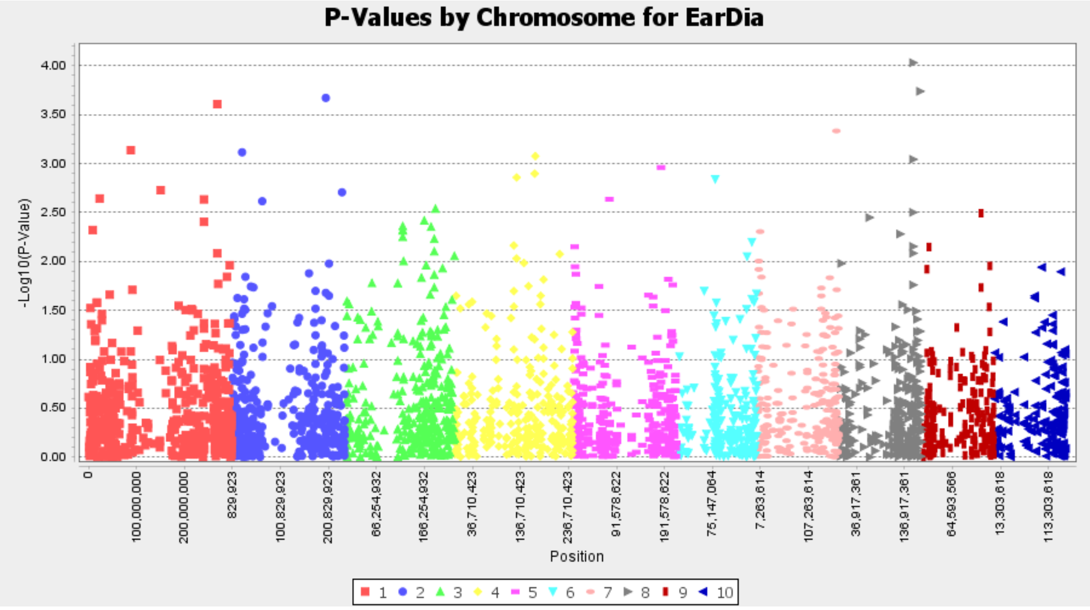
Step 5:

Check the results of QQ plot and Manhattan plots for 3 different traits, EarHT, dpoII and EarDia.




P-Values by Chromosome for EarHT
-Log10(P-Value)
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0
100,000,000
200,000,000
829,923
100,829,923
200,829,923
66,254,932
166,254,932
36,710,423
136,710,423
236,710,423
91,578,622
191,578,622
75,147,064
7,263,614
107,263,614
36,917,361
136,917,361
64,593,566
13,303,618
113,303,618
Position
■ 1 ● 2 ▲ 3 ◆ 4 ▬ 5 ▼ 6 ⬬ 7 ▶ 8 ▮ 9 ◀ 10


P-Values by Chromosome for dpoll
-Log10(P-Value)
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0
100,000,000
200,000,000
829,923
100,829,923
200,829,923
66,254,932
166,254,932
36,710,423
136,710,423
236,710,423
91,578,622
191,578,622
75,147,064
7,263,614
107,263,614
36,917,361
136,917,361
64,593,566
13,303,618
113,303,618
Position
■ 1 ● 2 ▲ 3 ◆ 4 ▬ 5 ▼ 6 ⬬ 7 ▶ 8 ▮ 9 ◀ 10



3. Results comparison

MLM is a model includes both fixed and random effects. MLM incorporates not only the population structure (Q) information but also the information about relationships among individuals, a genetic marker based kinship matrix (K). The "Q+K" approach of MLM improves statistical power compared to "Q" only approach (GLM).

As we can see from the results of the QQ plot for GML (Fig 1) and MLM (Fig 2). Obviously, most of the *p* values were much higher than expected for GLM (Fig 1), which means most of the SNPs are highly associated to traits. This is not reasonable outcomes and we may suspect that the false positive rate is too high and therefore the *p* values are overestimated. For this situation, exception for utilizing other model to perform GWAS, it's necessary to examine the population construction, pedigree and population size. Compared to GML, the QQ plot of MLM showed a more reasonable results. Besides dpoII (blue dots), we could tell the only small amount of *p* values were higher than expected. For the rest of the *p* values, they fitted quite well to the expected. In this case, we may have a conclusion that MLM is ideal model for GWAS, then we could pick up those significant SNPs in the Manhattan plot as potential high associated markers for the traits.

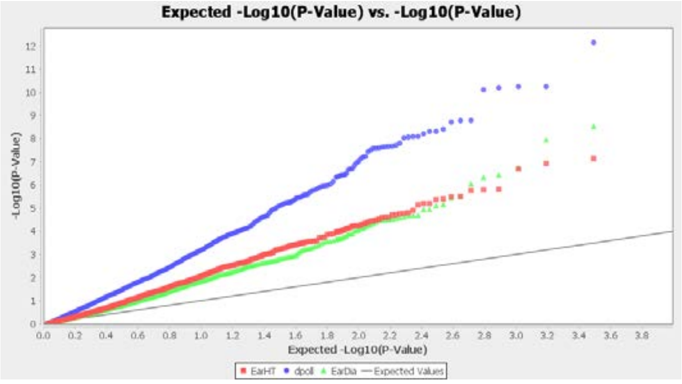


Fig 1. QQ plot of GLM.

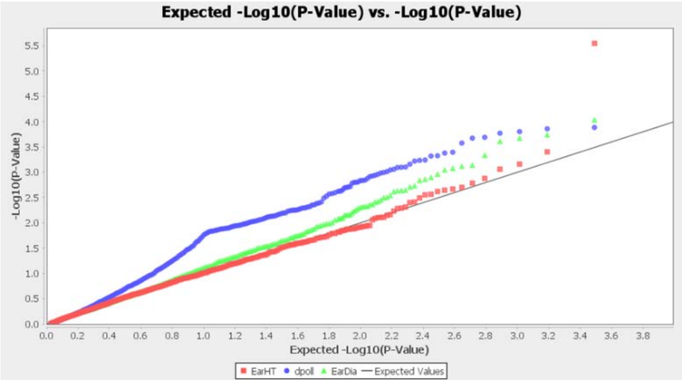


Fig 2. QQ plot of MLM.