Nghiên cứu kiểm soát chất lượng bộ mẫu phân tích tương quan toàn bộ hệ gen

Quality control of samples used for genome-wide association study

Nguyễn Ngọc Trung[*], Lê Gia Hoàng Linh[**], Trần Quang Nam[**], Mai Phương Thảo[**], Hoàng Anh Vũ[**], Đỗ Đức Minh[**]

**Đại học Quốc gia Thành phố Hồ Chí Minh,*
***Đại học Y Dược Thành phố Hồ Chí Minh*

Tóm tắt

Mục tiêu: Các nghiên cứu tương quan toàn bộ hệ gen (GWAS: Genome-wide association study) là một công cụ rất hiệu quả để nghiên cứu vai trò của yếu tố di truyền trong các bệnh lý đa nguyên nhân phức tạp. Tuy nhiên, với số lượng các điểm đa hình đơn nucleotide rất lớn được sử dụng trong các chip microarray, việc kiểm soát chất lượng dữ liệu từ các mẫu nghiên cứu là hết sức cần thiết. Thông qua nghiên cứu này, chúng tôi đã sử dụng các kỹ thuật sinh tin học để kiểm soát chất lượng các mẫu được phân tích toàn bộ hệ gen trên 494 người bình thường và 503 bệnh nhân đái tháo đường típ 2. *Đối tượng và phương pháp:* 997 đối tượng nghiên cứu (bao gồm 494 người bình thường và 503 bệnh nhân đái tháo đường típ 2) được phân tích toàn bộ hệ gen (khảo sát 644.303 điểm đa hình) bằng bộ kit Infinium Global Screening Array (GSA). Bằng cách sử dụng phần mềm GenomeStudio và PLINK, chúng tôi đã kiểm soát chất lượng cho các mẫu nghiên cứu theo chất lượng mẫu, chất lượng gọi điểm đa hình, sự phù hợp giới tính, mức độ dị hợp tử, mức độ quan hệ họ hàng. *Kết quả:* Với ngưỡng kiểm soát chất lượng cho mẫu là tỉ lệ gọi được biến thể (CallRate) ≥ 0,98, cho các điểm đa hình là điểm GenTrain ≥ 0,7, điểm Cluster Sep Score ≥ 0,3 và điểm Call Freq ≥ 0,95, đồng thời loại trừ các mẫu có giới tính không phù hợp, có mức độ di hợp tử cao và có khả năng có quan hệ họ hàng, chúng tôi đã loại trừ 213 mẫu và 264.390 điểm đa hình không đạt chất lượng. *Kết luận:* Với các ngưỡng khảo sát chất lượng nêu trên, chúng tôi đã áp dụng được các tiêu chuẩn kiểm soát chất lượng đầu vào cho các mẫu dữ liệu phân tích tương quan toàn bộ hệ gen với bộ mẫu bao gồm 494 người bình thường và 503 bệnh nhân đái tháo đường típ 2. Việc kiểm soát chất lượng này rất quan trọng để việc phân tích tương quan toàn bộ hệ gen cũng như ước tính chỉ số nguy cơ di truyền đa gen đạt được tính chính xác.

Từ khóa: Nghiên cứu tương quan toàn bộ hệ gen, GenomeStudio, PLINK, kiểm soát chất lượng.

Summary

Objetive: Genome-wide association study (GWAS) is a very effective tool to investigate the role of genetic contribution to the etiology of complex multifactorial diseases. However, due to the large amount of single nucleotide polymorphisms in microarray bead chip, the quality control process of the samples in GWAS is extremely necessary. In this study, bioinformatic tools were used to assess the quality of microarray samples including 503 type 2 diabetic patients and 494 controls. *Subject and method:* 997 subjects (494 controls and 503 type 2 diabetes cases) were genotyped using Infinium

Ngày nhận bài: 9/2/2023, ngày chấp nhận đăng: 01/3/2023

Người phản hồi: Đỗ Đức Minh, Email: ducminh@ump.edu.vn - Đại học Y Dược Thành phố Hồ Chí Minh

Global Screening Array (GSA) containing 644303 genetic markers. By using GenomeStudio and PLINK softwares, the standard for quality control of these samples was set for sample quality, polymorphism quality, sex-matching, heterozygousity, relationship. *Result:* Samples with any of the specific parameters CallRate < 0.98, GenTrain Score < 0.7, Cluster Sep Score < 0.3, Call Freq < 0.95, sex unmatching, very heterozygous, or potential relatives were considered not qualified. Finally, 213 samples and 264,390 polymorphisms were excluded from our data. *Conclusion:* With the quality threshold described above, we have successfully performed the quality control for GWAS study including 503 type 2 diabetic patients and 494 controls. These quality control steps are crucial for accurate genome analysis as well as polygenic risk score calculation.

Keywords: Genome-wide association study, GenomeStudio, PLINK, quality control.

1. Đặt vấn đề

Đái tháo đường típ 2 là một gánh nặng về sức khỏe lớn trên toàn thế giới với ước tính có khoảng hơn 700 triệu người mắc bệnh vào năm 2045 [1]. Tương tự với các bệnh lý đa yếu tố phức tạp khác, đái tháo đường típ 2 có sự đóng góp quan trọng của yếu tố di truyền và đã được chứng minh qua nhiều các nghiên cứu tương quan toàn bộ hệ gen (GWAS: genome-wide association study) ở nhiều chủng tộc [2, 3]. Tuy nhiên, cho đến nay, các nghiên cứu toàn bộ hệ gen này vẫn chưa được tiến hành ở người Việt Nam, dù cho quốc gia chúng ta có dân số lên đến gần 100 triệu người.

GWAS sử dụng kỹ thuật microarray cung cấp những thông tin về các điểm đa hình (SNP: Single nucleotide polymorphism), kiểm tra hàng trăm nghìn đến hàng triệu biến thể di truyền trên bộ gen để xác định các liên kết giữa kiểu gen và kiểu hình [4, 5]. GWAS ra đời lần đầu tiên vào năm 2005 để nghiên cứu bệnh thoái hóa điểm vàng do tuổi già, hơn 50000 mối liên hệ có ý nghĩa trên toàn bộ hệ gen đã được báo cáo giữa các biến thể di truyền và các bệnh thông thường [5]. Sự kết hợp này dẫn đến những hiểu biết sâu hơn về cấu trúc có khả năng gây bệnh (thông qua việc xác định các gen và cơ chế gây bệnh mới), những cải tiến trong chăm sóc lâm sàng (xác định các mục tiêu thuốc mới [4] và các dấu ấn sinh học gây bệnh) và y học cá thể (dự đoán nguy cơ và tối ưu hóa liệu pháp điều trị dựa vào kiểu gen) [5].

Mục đích của nghiên cứu GWAS là xác định hàng ngàn biến thể di truyền kết hợp với bệnh lý và tính trạng quan trọng ở người [6]. Một quy trình nghiên cứu GWAS thường bao gồm 4 phần chính: tiền xử lý dữ liệu từ dữ liệu thô (gọi kiểu gen), tạo ra dữ liệu mới (kiểm định chất lượng dữ liệu), phân tích thống kê và các phân tích chuyên sâu.

Vì dữ liệu thô xuất ra từ các nghiên cứu GWAS rất lớn nên cần phải có các bước phân tích sinh tin học để đảm bảo chất lượng của các mẫu để các bước phân tích sau đó có kết quả đáng tin cậy. Thông qua nghiên cứu này, chúng tôi mô tả quy trình kiểm soát chất lượng (QC: quality control) dữ liệu từ một bộ dữ liệu thô GWAS bao gồm 494 người bình thường và 503 bệnh nhân đái tháo đường típ 2.

2. Đối tượng và phương pháp

2.1. Đối tượng

Đối tượng của nghiên cứu này là bộ dữ liệu thô GWAS của 997 đối tượng tham gia nghiên cứu, bao gồm 494 người bình thường (nhóm chứng) và 503 bệnh nhân đái tháo đường típ 2 (nhóm bệnh). DNA bộ gen của các đối tượng này được tách và được khảo sát với bộ chip Infinium Global Array (GSA) v2.0. Bộ dữ liệu thô được sao ra dưới dạng file .idat là dữ liệu đầu vào.

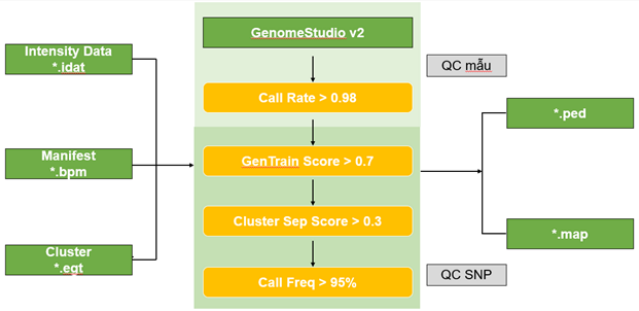
Mục tiêu: Kiểm định chất lượng các dữ liệu microarray với các ngưỡng chất lượng thường được sử dụng cho các nghiên cứu GWAS.

2.2. Phương pháp

2.2.1. Phần mềm GenomeStudio

Module Genotyping của GenomeStudio xử lý từ dữ liệu thô của chip microarray Illumina thành định dạng PLINK (là định dạng chuẩn để lưu trữ dữ liệu kiểu gen) [7]. Quy trình xử lý trên phần mềm GenomeStudio Software v2.0 bao gồm một số bước

như mô tả trong Hình 1 [8, 9]. Phần mềm này chủ yếu để khảo sát chỉ số tỉ lệ gọi được biến thể (Call Rate) trong mẫu; chỉ số GenTrain Score để xác định sự phân tách AA, AB hoặc B của kiểu gen; chỉ số Cluster Sep Score đánh giá khả năng phân cụm chính xác; chỉ số Call Freq cho thấy xác suất một SNP được định danh tại một locus cụ thể ở đa số các mẫu.



Hình 1. Quy trình kiểm soát chất lượng và xác định kiểu gen trên phần mềm GenomeStudio v2

2.2.2. Phần mềm PLINK

Phần mềm PLINK được sử dụng để xác định sự phù hợp về giới tính giữa dữ liệu gen và dữ liệu khai báo ban đầu, mức độ dị hợp tử của mẫu và khả năng họ hàng giữa các mẫu.

2.3. Vấn đề đạo đức trong nghiên cứu

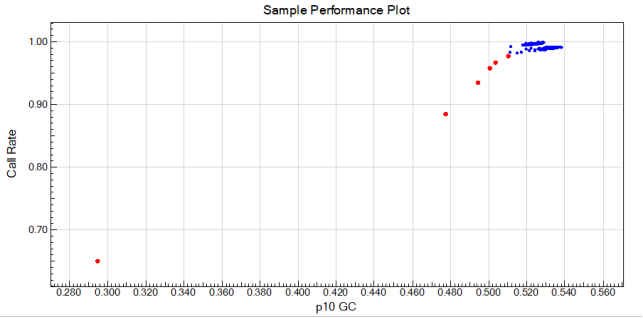
Đề tài nghiên cứu đã được sự chấp thuận của Hội đồng đạo đức trong nghiên cứu Y sinh học Đại học Y Dược Thành phố Hồ Chí Minh với quyết định số 350/HĐĐĐ-ĐHYD ngày 26 tháng 5 năm 2020.

3. Kết quả

3.1. Kiểm định chất lượng mẫu bằng phần mềm GenomeStudio

Dữ liệu microarray thô từ 997 đối tượng tham gia nghiên cứu (494 người bình thường và 503 ca bệnh) được xử lý bằng phần mềm GenomeStudio để xác định kiểu gen. Kết quả được trình bày ở Hình 2 cho thấy, 6 mẫu có tỉ lệ gọi được biến thể (Call Rate) < 0,98 và được loại bỏ khỏi nghiên cứu. Sau khi loại bỏ 6 mẫu có giá trị Call Rate không phù hợp, 991 mẫu còn lại sẽ tiếp tục được đánh giá chất lượng SNP. Dữ liệu Call Rate còn cho thấy nồng độ DNA và chất lượng tách chiết DNA của các mẫu tham gia nghiên cứu đạt chất lượng.

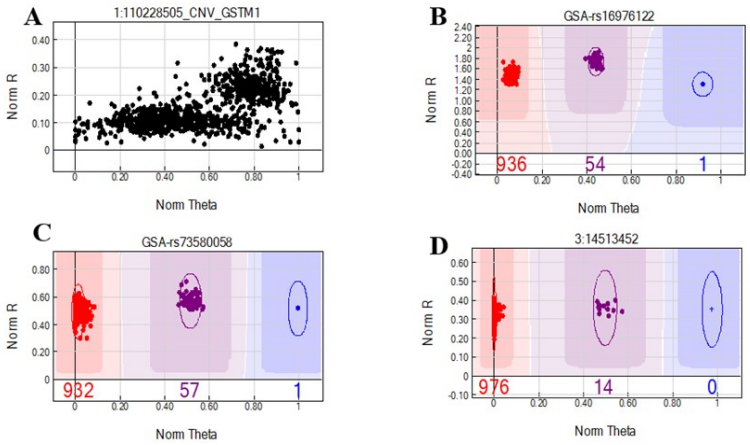
Các SNP của 991 mẫu sẽ được đánh giá dựa trên các chỉ số GenTrain Score, Cluster Sep Score và Call Freq. Đối với các SNP có chỉ số GenTrain Score thấp < 0,7 (Hình 3A), các SNP không được chia thành 2 hoặc 3 cụm rõ ràng trên đồ thị. Vì vậy, các SNP này sẽ không thể được xác định kiểu gen AA, AB hoặc BB và sẽ bị loại khỏi nghiên cứu. Đối với các SNP đó GenTrain Score > 0,7 (Hình 3B, C, D), các cụm kiểu gen được phân chia rõ ràng và đặc hiệu. Các SNP tiếp tục được đánh giá chất lượng dựa vào tiêu chuẩn Cluster Sep Score > 0,3 và Call Freq > 95%. Các SNP có điểm Cluster Sep < 0,3 cho thấy việc phân cụm không được thực hiện chính xác và Call Freq < 95% cho thấy SNP tại một locus cụ thể không được gọi ở đa số các mẫu (Hình 4A).



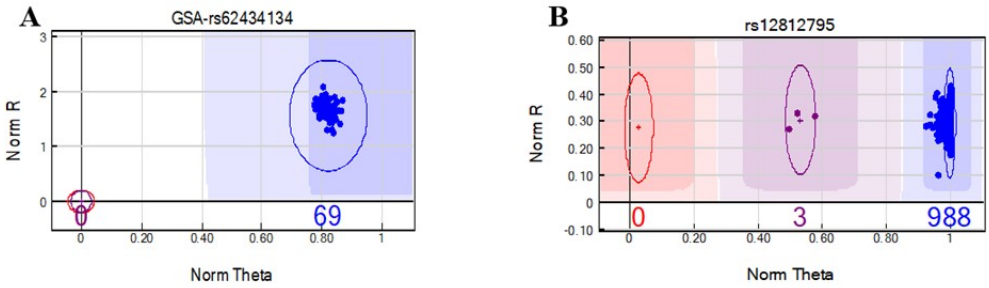
Hình 2. Đánh giá chất lượng và hiệu suất DNA mẫu tham gia nghiên cứu

Sau khi loại trừ các SNP này, những SNP được phân cụm tốt và được gọi ở đa số các mẫu (Hình 4B) sẽ được lọc để tiến hành tạo dữ liệu đầu vào trên phần mềm PLINK (*.ped và *.map).

Trong tổng số 665.608 SNP trên bộ kit Infinium Global Screening Array v2, sau khi lọc với các tiêu chuẩn đánh giá chất lượng có 640.213 SNP (nhóm chứng) và 642.075 SNP (nhóm bệnh) đạt yêu cầu ở 991 mẫu tham gia nghiên cứu. Đối với các SNP hiếm (tần suất biến thể < 0,01 và Call Freq < 0,9999), có tổng cộng 52.457 SNP nhóm bệnh và 55.381 SNP nhóm chứng được lọc từ phần mềm GenomeStudio giúp thu nhận được một danh sách chứa các SNP hiếm không được gọi bằng thuật toán GenCall. Cuối cùng, các mẫu và SNP đạt chất lượng sẽ được phần mềm GenomeStudio xuất dữ liệu dưới định dạng PLINK bao gồm 2 file: *.ped (chứa thông tin lâm sàng, kiểu hình, dữ liệu kiểu gen của đối tượng) và *.map (chứa thông tin về mã số và vị trí của SNP).



Hình 3. Đồ thị mô tả chất lượng SNP dựa vào chỉ số GenTrain Score. A) SNP với điểm GenTrain 0,0. B) SNP với điểm GenTrain 0,7126. C) SNP với điểm GenTrain 0,9552. D) SNP với điểm GenTrain 0,9609



Hình 4. Đồ thị mô tả chất lượng SNP dựa vào chỉ số Cluster Sep Score và Call Freq. A) SNP với Cluster Sep = 0 và Call Freq = 0,0696. B) SNP với Cluster Sep Score = 1 và Call Freq = 1.

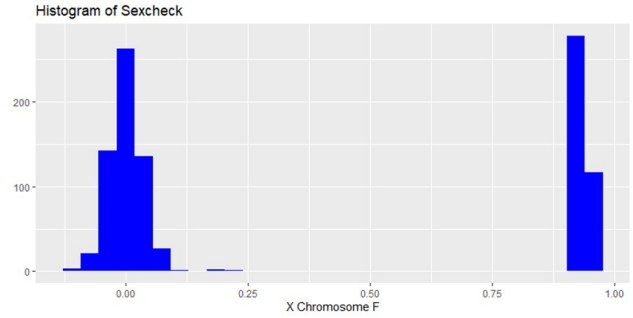
Bảng 1. Kết quả lọc dữ liệu kiểu gen nhóm bệnh và nhóm chứng với phần mềm GenomeStudio

		Chứng		Bệnh	
		Trước QC	Sau QC	Trước QC	Sau QC
QC Mẫu	N (Số mẫu)	494		503	
	Call Rate > 0,98	494	490	503	501
QC SNP	GenTrain Score > 0,7	665.608	640.213	665.608	642.075
	Cluster Sep > 0,3				
	Call Freq > 0,95				
Rare SNP (bao gồm các tiêu chí ở trên)	MAF < 0,01	665.608	55.381	665.608	52.457
	Call Freq < 0,9999				

3.2. Kiểm soát chất lượng dữ liệu bằng phần mềm PLINK

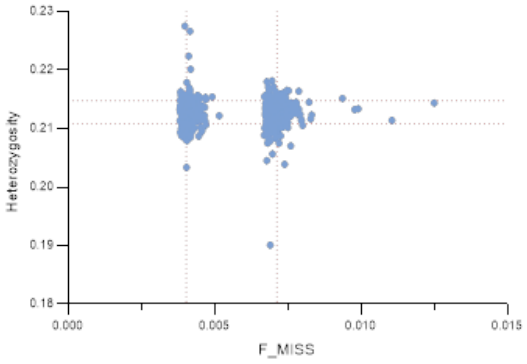
Dữ liệu thô từ file *.ped và *.map sau khi được xuất ra từ phần mềm GenomeStudio được chuyển về định dạng nhị phân bao gồm 3 file *.bim, *.bed và *.map bằng phần mềm PLINK. Tất cả các SNP được tạo ra sẽ được chuyển về mạch xuôi theo bộ gen người Hg19. 991 mẫu sẽ được đánh giá sự khác biệt về giới tính giữa kiểu gen và kiểu hình thông qua chỉ

số F từ phần mềm Plink. Kết quả của nghiên cứu được trình bày trong Hình 5. Kết quả cho thấy trong tổng số 991 mẫu, có 1 mẫu bệnh có giá trị F > 0,2 và < 0,8 (mẫu FID = 402 với F = 0,2184) và 3 mẫu chứng có sự khác biệt về kiểu hình thực tế so với kiểu gen với giá trị F tương ứng là -0,0024629, 0,9363 và 0,949 (FID = 8, 20, 45). Cả 4 mẫu này sẽ bị loại khỏi nghiên cứu để đảm bảo kết quả không bị ảnh hưởng.



Hình 5. Kiểm soát chất lượng giới tính giữa kiểu gen và kiểu hình

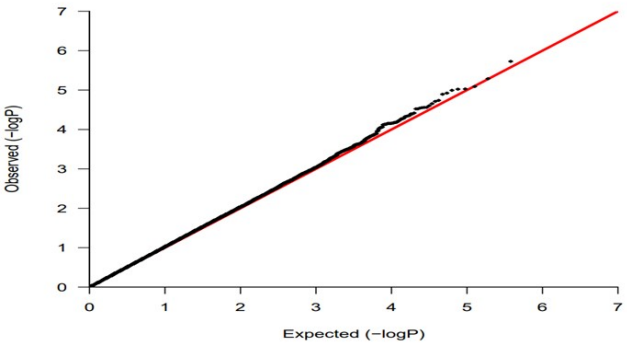
Dữ liệu sau đó tiếp tục được sử dụng để đánh giá mức độ thiếu thông tin đa hình và tỷ lệ dị hợp tử. Kết quả của nghiên cứu được thể hiện ở Hình 6. Kết quả với trục x là tỷ lệ thiếu dữ liệu đa hình với độ lệch chuẩn ±0,001546 và trục y là tỷ lệ dị hợp tử với độ lệch chuẩn ±0,001979284. Kết quả cho thấy có 210 mẫu tham gia nghiên cứu có tỷ lệ dị hợp tử cao. Đối với thiếu dữ liệu các điểm đa hình dựa trên phân tích này, hầu hết các mẫu đã được đánh giá chất lượng về mức độ thiếu dữ liệu đa hình ở mức > 0.98 (mức rất cao). Vì vậy, đối với bước kiểm soát chất lượng này sẽ loại bỏ 210 mẫu có mức độ dị hợp tử cao.



Hình 6. Đồ thị hiển thị mức độ thiếu dữ liệu kiểu gen và dị hợp tử

Các mẫu tiếp tục được đánh giá về mức độ quan hệ họ hàng hoặc mẫu bị lặp lại trong quá trình nghiên cứu. Kết quả cho thấy trong tổng số 991 mẫu tham gia nghiên cứu, có 2 cặp mẫu có chỉ số PI_HAT > 0,185 thuộc nhóm bệnh là mẫu 33 và 82 (với PI_HAT = 1); mẫu 338 và 368 (với PI_HAT = 1). Cả 4 mẫu này sẽ bị loại khỏi nghiên cứu.

Các mẫu còn lại sẽ được tiếp tục đánh giá chất lượng SNP dựa vào ngưỡng tần xuất alen là 0,01, giá trị p-value của cân bằng Hardy-Weinberg (HWE: Hardy-Weinberg equilibrium) là 10-5 và chỉ số phân cụm của tất cả các SNP. Kết quả được trình bày ở Hình 7. Các giá trị $-\log_{10}(p)$ trong khoảng từ 0-2 cho thấy không có sự phân tầng quần thể hoặc các đối tượng tham gia nghiên cứu có quan hệ họ hàng. Giá trị trong khoảng từ 3-6 cho thấy có khả năng xuất hiện các SNP kết hợp với bệnh lý.



Hình 7. Đồ thị Q-Q Plot biểu hiện giá trị p-value mong đợi và p-value thực tế.

Sau khi kiểm soát chất lượng các điểm đa hình và các đối tượng tham gia nghiên cứu, kết quả được trình bày ở Bảng 2 và Bảng 3. Dữ liệu kiểu gen sau khi được kiểm soát chất lượng còn lại 397 ca bệnh và 381 ca chứng bao gồm 311 nam và 467 nữ đạt tiêu chuẩn. Trong tổng số 665.608 SNP, có 264.390 SNP bị loại khỏi nghiên cứu (Bảng 2). Có tổng cộng 213 mẫu bị loại khỏi nghiên cứu, trong đó có 4 mẫu bị loại do khác biệt giới tính ghi nhận trên kiểu gen so với thực tế; Có 210 mẫu bị loại do mức độ dị hợp tử cao; 4 mẫu bị loại do có quan hệ họ hàng hoặc mẫu bị lặp lại trong quá trình xử lý. Trong tổng số 264.390 SNP bị loại có 6.448 SNP bị thiếu kiểu gen ở đa số các mẫu; 6.444 SNP bị loại do tỉ lệ gọi kiểu gen giữa nhóm bệnh và chứng quá chênh lệch. Có

257.880 SNP bị loại do có tần suất alen lặn (MAF) < 0,01 và 61 SNP bị loại do sai quy luật HWE. Sau khi loại bỏ tất cả các mẫu và SNP dưới ngưỡng QC. Toàn bộ dữ liệu còn lại đạt yêu cầu sẽ tiếp tục được sử dụng để xác định mối liên hệ giữa kiểu gen và kiểu hình bệnh lý đái tháo đường típ 2.

Bảng 2. Số lượng đối tượng và SNP trước và sau khi QC

	Bệnh	Chứng	Nam	Nữ	SNP
Trước QC	501	490	394	597	644.303
Sau QC	397	381	311	467	379.913

Bảng 1. Số lượng đối tượng và SNP bị loại bỏ ở các bước QC

		Bệnh	Chứng
QC đối tượng	Giới tính không chính xác	1	3
	Dị hợp tử	102	108
	Quan hệ họ hàng	4	0
	Phân tầng quần thể	0	0
QC SNP	Thiếu kiểu gen	6.448	
	Dữ liệu SNP giữa Case và Control	6.444	
	HWE	61	
	MAF	257.880	

4. Bàn luận

Qua các bước dùng các phần mềm sinh tin học như GenomeStudio và PLINK, chúng tôi đã loại trừ được các mẫu và các điểm đa hình không đảm bảo chất lượng để có thể phân tích tiếp tục ở các bước sau. Từ đó cho thấy tính quan trọng của việc kiểm soát chất lượng mẫu trong các nghiên cứu tương quan toàn bộ hệ gen. Đây là các kết quả nước đầu giúp cho việc phân tích sâu hơn để so sánh về mối liên hệ và sự khác biết giữa bộ gen người đái tháo đường típ 2 so với người khỏe mạnh thông thường ở dân số người Việt Nam.

Ngưỡng lựa chọn giá trị Callrate thông thường được lựa chọn từ 95-98% tùy vào mức độ chặt chẽ của thiết kết nghiên cứu được xem là tiêu chuẩn chung cho các nghiên cứu tương quan toàn bộ hệ gen [9]. Với ngưỡng lựa chọn này, tỉ lệ mẫu thường bị loại trừ khỏi nghiên cứu là từ 1-2% dân số nghiên cứu [7]. Việc lựa chọn giá trị này là 98% trong nghiên cứu của chúng tôi nhằm tạo ra một tiêu chuẩn chặt chẽ cho chất lượng các mẫu tham gia nghiên cứu, với sự lựa chọn này, chúng tôi đã loại trừ 6 mẫu nghiên cứu, dẫn đến tỉ lệ loại trừ do chỉ số Callrate là 6%. Một phần lớn các mẫu sau kiểm định chất lượng cũng bị loại trừ do mức độ di hợp tử. Tính dị hợp tử quá cao trong các mẫu nghiên cứu có thể là nguy cơ của việc ngoại nhiễm mẫu, ngược lại khi tính dị hợp tử quá thấp là nguy cơ của việc các mẫu có quan hệ họ hàng [7]. Để loại trừ hai khả năng này, đã có 210 mẫu bị loại khỏi bước phân tích này để đảm bảo chất lượng của quần thể mẫu nghiên cứu.

Một chỉ số quan trọng hàng đầu khi kiểm định chất lượng của các SNP là chỉ số Gentrain, chỉ số đánh giá mức độ phân cụm của các SNP, được biểu thị với giá trị từ 0 đến 1. Việc lựa chọn ngưỡng 0,7 cho các SNP được xem là ngưỡng chung của các nghiên cứu trên thế giới, giúp cho các cụm kiểu gen được phân đinh rõ ràng, tránh trường hợp xuất hiện tình trạng phân tán hoặc số lượng cụm đa hình nhiều hơn 3 (không tuân thủ định luật Mendel) [7, 9]. Tuy nhiên, đa phần các SNP bị loại trong bước kiểm định chất lượng này là do MAF < 0,01, một lý giải có thể cho sự loại trừ này là do sự khác biệt về

chủng tộc. Đối với các chủng tộc chưa có trình tự tham khảo toàn bộ hệ gen với số mẫu lớn như người Việt Nam, việc phân tích các SNP có MAF thấp là một thách thức vì tần suất thấp ở chủng tộc này có thể cao ở chủng tộc khác và ngược lại. Do đó, rất cần có thêm các nghiên cứu giải trình tự toàn bộ hệ gen với cỡ mẫu lớn ở người Việt để xác định được vai trò của các SNP có MAF thấp, tránh việc loại trừ quá mức một số SNP có thể có mối liên quan với tình trạng bệnh lý cần khảo sát.

Một hạn chế của nghiên cứu này là số lượng mẫu chưa nhiều (đối với các nghiên cứu hệ gen, số lượng mẫu có thể cần lên đến vài ngàn mẫu để có thể thấy được các điểm đa hình có liên quan với nguy cơ bệnh lý), tuy nhiên, đây sẽ là một trong các kết quả ban đầu phục vụ cho các phân tích gộp về hệ gen trên người châu Á sau này.

5. Kết luận

Qua nghiên cứu này, chúng tôi đã xác định được các tiêu chuẩn và tiến hành kiểm định chất lượng mẫu của các nghiên cứu GWAS. Các bước thực hiện này có thể được áp dụng rộng rãi cho các nghiên cứu GWAS sau này.

Lời cảm ơn: Nghiên cứu đã được tài trợ bởi Quỹ Phát triển Khoa học và Công nghệ Quốc gia (NAFOSTED) trong đề tài mã số 108.01-2019.319.

Tài liệu tham khảo

1. Saeedi P, Petersohn I, Salpea P et al (2019) *Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition*. Diabetes Res Clin Pract 157: 107843. doi:10.1016/j.diabres.2019.107843.
2. Xue A, Wu Y, Zhu Z et al (2018) *Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.* Nat Commun 9(1): 2941 doi:10.1038/s41467-018-04951-w.
3. Wheeler E, Barroso I (2011) *Genome-wide association studies and type 2 diabetes.* Brief Funct Genomics 10(2):52-60. doi:10.1093/bfgp/elr008.
4. Cao C, Moult J (2014) *GWAS and drug targets*. BMC Genomics 15(4): 5. doi:10.1186/1471-2164-15-S4-S5.
5. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019) *Benefits and limitations of genome-wide association studies*. Nat Rev Genet 20(8): 467-484. doi:10.1038/s41576-019-0127-1.
6. Zhu Z, Zhang F, Hu H et al (2016) *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet 48(5): 48-487. doi:10.1038/ng.3538.
7. Zhao S, Jing W, Samuels DC, Sheng Q, Shyr Y, Guo Y (2018) *Strategies for processing and quality control of Illumina genotyping arrays*. Brief Bioinform 19(5): 765-775. doi:10.1093/bib/bbx012.
8. Fountain ED, Zhou LC, Karklus A et al (2019) *Cross-Species Application of Illumina iScan Microarrays for Cost-Effective, High-Throughput SNP Discovery*. Front Ecol Evol. 2021;9. Accessed June 21, 2022. https://www.frontiersin.org/article/10.3389/fevo.2021.629252.
9. Guo Y, He J, Zhao S et al (2014) *Illumina human exome genotyping array clustering and quality control.* Nat Protoc 9(11): 2643-2662. doi:10.1038/nprot.2014.174.