

## บทคัดย่อ

รหัสโครงการ TRG5180005

ชื่อโครงการ การหาสถิติที่สำคัญสำหรับการจำแนกประชากรโดยวิธีแบบชื่อย่อย

ชื่อนักวิจัย อภิชาติ อินทรพานิชย์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

E-mail address: [apichart.intarapanich@nectec.or.th](mailto:apichart.intarapanich@nectec.or.th)

ระยะเวลาโครงการ 1 ปี

ปัจจุบันข้อมูลของจีโนมไทป์มีขนาดใหญ่มากขึ้น ทำให้การอนุมานจำนวนกลุ่มประชากรย่อยและการจำแนกคนเข้าในกลุ่มประชากรย่อยเป็นเรื่องยาก วิธีการที่เป็นที่นิยมและมีประสิทธิภาพในการคำนวณข้อมูลที่มีขนาดใหญ่คือ Principal Components Analysis (PCA) ในปัจจุบันนี้ PCA สามารถตรวจจับหาโครงสร้างของประชากรได้เท่านั้นไม่สามารถที่จะหาจำนวนกลุ่มของประชากรย่อยได้ ดังนั้นคณะวิจัยจึงทำการพัฒนาระเบียบวิธีโดยปรับเปลี่ยนเทคนิคของ PCA เพื่อทำการหาจำนวนกลุ่มประชากรย่อยและสามารถตรวจสอบหากกลุ่มประชากรที่เหมาะสมกับบุคคลได้ ระบบวิธีที่ทำการพัฒนาขึ้นนี้เรียกว่า iterative pruning PCA (ipPCA) สามารถใช้กับข้อมูลจีโนมไทป์มีขนาดใหญ่ได้อย่างมีประสิทธิภาพ โดยระเบียบวิธี ipPCA ได้ถูกทดสอบประสิทธิภาพกับข้อมูลจำลอง ผลการทดลองแสดงให้เห็นว่าสำหรับข้อมูลกลุ่มประชากรที่ไม่ซับซ้อนผลที่ได้สอดคล้องกับระเบียบวิธี STRUCTURE และ AWclust อย่างไรก็ตาม สำหรับข้อมูลของกลุ่มประชากรที่ซับซ้อน มีเพียงระเบียบวิธี ipPCA เท่านั้นที่สามารถจำแนกกลุ่มประชากรได้อย่างถูกต้อง นอกจากนี้ระเบียบวิธี ipPCA ถูกทดสอบกับข้อมูลจริงที่มีขนาดของข้อมูลจีโนมไทป์ใหญ่มากซึ่งไม่สามารถวิเคราะห์ได้โดยระเบียบวิธี STRUCTURE และ AWclust

## Abstract

Project Code : TRG5180005

Project Title :

Investigator : Apichart.intarapanich National Electronics and Computer Technology Center

E-mail address: [apichart.intarapanich@nectec.or.th](mailto:apichart.intarapanich@nectec.or.th)

Project Period : 1 year

As genotypic datasets become ever larger, it is increasingly difficult to correctly estimate the number of subpopulations and assigning individuals to them. The computationally efficient non-parametric, chiefly Principal Components Analysis (PCA)-based methods are thus becoming increasingly relied upon for population structure analysis. Current PCA-based methods can accurately detect structure; however, the accuracy in resolving subpopulations and assigning individuals to them is wanting. When subpopulations are closely related to one another, they overlap in PCA space and appear as a conglomerate. This problem is exacerbated when some subpopulations in the dataset are genetically far removed from others. We propose a novel PCA-based framework which addresses this shortcoming.

A novel population clustering program called iterative pruning PCA (ipPCA) was developed which assigns individuals to subpopulations and infers the total number of subpopulations present. Genotypic data from simulated and real population datasets with different degrees of structure were analyzed by the proposed method. For datasets with simple structures, the subpopulation assignments of individuals made by ipPCA were consistent with the STRUCTURE and AWclust algorithms. On the other hand, highly structured populations containing many closely related subpopulations could be accurately resolved only by ipPCA, and not by other methods. The algorithm is computationally efficient and not constrained by the dataset complexity. This systematic clustering approach removes the need for prior population labels, which could be advantageous when cryptic stratification is encountered in datasets of individuals otherwise assumed to be homogenous.