國立臺灣大學生命科學院基因體與系統生物學學位學程

碩士論文

Genome and System Biology Degree Program

College of Life Science

National Taiwan University

Master's Thesis

AlphaThaICNV: 通過次世代定序技術增強 α 地中海型貧血拷貝數變異檢測能力

AlphaThalCNV: Enhanced Detection of CNVs in Alpha Thalassemia via Next-Generation Sequencing.

Nguyen Vuong Thao Vy

指導教授 : 李妮鍾 博士

Advisor: Ni-Chung Lee, M.D Ph.D

中華民國 114 年 1 月

January 2025

doi:10.6342/NTU202500218

Acknowledgements

This journey was both challenging but rewarding, and it would have been impossible and incomplete without the guidance, support, and love of so many extraordinary individuals in my life.

To my professor, Dr. Ni-Chung Lee, I extend my sincere gratitude for your mentorship, patience, and wisdom. Your dedication as a doctor, a researcher, and a teacher has been an inspiration to me in this journey of becoming a master student in medical genetics and becoming a bioinformaticians. Thank you for accepting me into your lab when I was transferred from a different lab that allowed me to pursue my research of interest. Thank you for your understanding and your mentorship that guide me and teach me a lot of knowledge. Without you, I would not be able to achieve my potential and learn so many things. The experience I gained working in your lab at National Taiwan University Children's Hospital will accelerate my pursuit of bioinformatics

To my lab mates, Wanda and Hsu-Ching, thank you for your support and your presence that helped me overcome difficulties. Your experiences and suggestions are an important part of this thesis, without you, I would never have been able to complete this journey.

To my family, thank you for your encouragement, financial and emotional support. Without you, I would never be able to come to Taiwan to pursue this master's degree in the first place. This journey is full of ups and downs, but I truly appreciated that I have grown a lot through challenges with people I love. To my fiancé, thank you for your presence during my worst and best. Thank you for believing in me even though I doubted myself.

doi:10.6342/NTU202500218

摘要

引言

α-地中海貧血是一種常見的血紅蛋白病，由 α-珠蛋白基因（HBA）的缺失引起，導致 α-珠蛋白的生產減少或完全缺乏。這會導致小紅細胞增多症、貧血，以及在嚴重情況下引發如肝腫大和需依賴輸血等併發症。鑒於其對全球健康的負擔，準確、高效且具規模化的分子診斷方法對於 α-地中海貧血攜帶者的篩查與診斷至關重要。傳統方法如單管多重 PCR 僅限於檢測已知的缺失，而全外顯子組測序（Whole-Exome Sequencing）因缺乏靈敏度，難以準確檢測幾乎相同的 HBA1 和 HBA2 基因中的拷貝數變異（CNVs）。

材料與方法

AlphaThalCNV 流程利用 GATK-gCNV 工具，並通過優化參數（區塊大小：50；填充：400），提升 HBA 基因簇內的 CNV 檢測能力。該流程以 100 個隨機樣本的群體訓練模型，實現群體內的 CNV 檢測。隨後，對每個 α-地中海貧血病例進行單獨分析，並將結果與建立的模型比較。結果使用 IGV 可視化並導出為 VCF 文件，根據被刪除或重複的 α-珠蛋白基因數量進行分類。

為驗證 GATK-gCNV 結果的準確性，我們採用了 Samuel S. Chong 等人改編的引物進行多重 PCR，可在單次實驗中檢測多個基因缺失。群體分析納入了 415 名來自台灣的個體，統計分析 α-地中海貧血攜帶者的盛行率。

結果

AlphaThalCNV 成功檢測到三個此前未檢出的 α-地中海貧血病例，包括兩個--SEA 缺

doi:10.6342/NTU202500218

失和一個-3.7 缺失。此外，五名最初因其他疾病住院的 α-地中海貧血攜帶者也被準確檢測並通過多重 PCR 確認為真正陽性。

在台灣群體中，AlphaThalCNV 檢測到--SEA 型兩拷貝缺失的攜帶者盛行率為 5.3%，HBA2 基因中的一拷貝缺失（-4.2 型或-3.7 型 I/II）的盛行率為 2.2%。這些發現與使用其他方法（Hsu 等，2023 年）進行的大規模研究結果一致，證實了該流程的穩健性與準確性。

結論

AlphaThalCNV 流程顯著提升了 α-地中海貧血 CNV 檢測的準確性，特別是在標準流程遺漏的病例中。通過整合精確的分類與可視化工具，該流程為大規模群體分析提供了一種可靠的方法。其適應性使其成為未來醫學診斷和流行病學研究中的重要工具。

關鍵詞：α-地中海貧血，GATK-gCNV，次世代定序（NGS），全外顯子定序（WES），AlphaThalCNV

doi:10.6342/NTU202500218

Abstract

Introduction

Alpha-thalassemia is a prevalent hemoglobinopathy caused by deletions in the α-globin genes (HBA), resulting in reduced or absent α-globin production. This leads to microcytosis, anemia, and in severe cases, complications like hepatomegaly and blood transfusion dependency. Given its global health burden, an accurate, efficient, and scalable molecular diagnostic approach is critical for alpha-thalassemia carrier screening and diagnosis. Traditional methods like Single-Tube Multiplex PCR are limited to detecting known deletions, while Whole-Exome Sequencing lacks sensitivity to accurately detect copy number variations (CNVs) in the nearly identical HBA1 and HBA2 genes.

Materials and Methods

The AlphaThalCNV pipeline leverages the GATK-gCNV tool with optimized parameters (bin size: 50, padding: 400) to enhance CNV detection in the HBA gene cluster. The pipeline was trained using a cohort of 100 random samples, enabling CNV calling within the cohort. Subsequently, α-thalassemia cases were analyzed individually against the constructed model. Results were visualized with IGV and exported as VCF files for further classification based on the number of alpha-globin genes deleted or duplicated.

To confirm the accuracy of GATK-gCNV results, we employed Multiplex PCR with primers adapted from Samuel S. Chong et al., enabling detection of multiple deletions in a single run. The cohort analysis included 415 individuals from Taiwan to statistically analyze α-thalassemia carrier prevalence.

doi:10.6342/NTU202500218

Results

AlphaThalCNV successfully identified CNVs in three previously undetected α-thalassemia cases: two with --SEA deletions and one with -3.7 deletion. Additionally, five carriers of α-thalassemia, initially hospitalized for other diseases, were accurately detected and confirmed as true-positive through Multiplex PCR.

Among the Taiwanese cohort, AlphaThalCNV revealed a carrier prevalence of 5.3% for two-copy deletions (--SEA type) and 2.2% for one-copy deletions in the HBA2 gene (-4.2 type or -3.7 type I/II). These findings align with larger-scale studies conducted via alternate methodologies (Hsu et al., 2023), demonstrating the pipeline's robustness and accuracy.

Conclusion

The AlphaThalCNV pipeline has significantly improved the accuracy of α-thalassemia CNV detection, particularly in cases missed by standard pipelines. By integrating precise classification and visualization tools, the pipeline offers a reliable method for large-scale cohort analysis. Its adaptability makes it a valuable tool for future medical diagnostics and epidemiological studies.

Keywords: α-thalassemia, GATK-gCNV, Next-generation Sequencing (NGS), Whole-exome sequencing (WES), AlphaThalCNV.

doi:10.6342/NTU202500218

Table of Contents

doi:10.6342/NTU202500218

doi:10.6342/NTU202500218

doi:10.6342/NTU202500218

List of Figure

doi:10.6342/NTU202500218

List of Table

doi:10.6342/NTU202500218

I) Introduction

1.1) Alpha-thalassemia

1.1.1) Alpha-thalassemia molecular basis

α-thalassemia is one of the most common hemoglobin genetic abnormalities. This defect is the reduced or absent production of the α-globin chains, which is the constitution of several hemoglobin (Hb) types, including adult hemoglobin HbA (α2 beta2), fetal hemoglobin (α2 gamma2), and minor component of HbA_2 (α2 delta2). α-Globin chain is a subunit essential for both fetal and adult hemoglobin, which constitutes gas-transportation molecules in blood. Adult red blood cell is composed of two α globin chains and two beta globin chains which express from the α-like and β-like globin chains regulated by genes on chromosomes 16 and 11, respectively (Higgs & Wood, 2008).

The α-thalassemia gene is unique due to several factors that distinguish it in the field of genetics and hemoglobin disorders. Humans typically have duplication of α-globin gene (HBA1 and HBA2) on each chromosome 16, resulting in total four copies in total. This allows flexibility for the α-globin production (Farashi & Harteveld, 2018). Interestingly, HBA2 gene encodes 2-3-fold more mRNA translated into more protein than HBA1, this shows that human α-globin gen gene cluster contains a major and a minor locus. This phenomenon also suggests that mutation on HBA2 gene has more impact than HBA1 gene (Liebhaber et al., 1986). Additionally, unlike many genetic disorders caused mainly by point mutations, α-thalassemia is mainly caused by deletion of α-globin gene. Thus, this hemoglobinopathies has wide clinical spectrum from asymptom to lethal (Higgs et al., 1989a). α-thalassemia can be co-inherited with β-thalassemia, this event is rare but still

doi:10.6342/NTU202500218

recorded, or co-existed in high prevalence region with thalassemia disease, such as China (J. Li et al., 2014).

1.1.2) Alpha-thalassemia clinical forms:

α-thalassemia is caused commonly by structural variations in the structural genes which can be chromosomal translocations and duplication affecting the α-globin cluster, deletion removing structural genes, large deletion extending beyond the α-globin cluster, deletion removing the upstream regulatory elements or rare mutation cause α-thalassemia via an antisense RNA (Higgs, 2013a). The less common cause of nondeletion defects in alpha thalassemia is due to point mutations or single nucleotide polymorphisms (SNPs) in the genes responsible for the synthesis of alpha-globin chains. These mutations result in the reduced or absence of α-globin chains synthesis (Aydinok, 2012) thus could leads to imbalance in red blood cell production or anemia.

The clinical condition of α-thalassemia increases its severity with the number of nonfunctional α globin genes (HBA) (Table 1). This deletion leads to the reduction in production of α-globin chains needed for the formation of hemoglobin, imbalance between α-globin chains and β-globin chains cause abnormal red blood cells, resulting in microcytosis. Hb level in α-thalassemia patient varies from normal to severe anemia that depend on blood-transfusion or died from fetus state (Shang & Xu, 2017). The major clinical syndrome resulting from α-thalassemia was first described and published in the 1967 in Thailand with Bart's hydrops fetalis (POOTRAKUL et al., 1967). There are 4 clinical conditions including: two carrier states (cause by the deletion or dysfunction of one to two normal α globin genes) and two clinically relevant forms (HbH disease due to only

doi:10.6342/NTU202500218

one functioning α gene and Hb Bart hydrops fetalis cause by no functioning α gene) (Piel & Weatherall, 2014a). At the phenotypic level, the carrier stated are divided into silent carrier and α-thalassemia trait. Silent carriers are characterized by the deletion or dysfunction of one α globin gene thus carriers are clinically and hematological normal, in some cases, moderate microcytosis and hypochromia can be observed in α-thalassemia carrier (Higgs, 2013a). People with two affected α-globin genes, either affected in the same or different locus, result in microcytic hypochromic anemia. Patients with three dysfunctional α-globin genes result in low production of α-globin chains thus imbalance β-globin chains within red blood cells, resulting in moderate anemia and marked microcytosis. This is also called Hb H disease. Hb Bart's hydrops fetalis is the most severe form of α-thalassemia cause by the deletion of four α-globin gene and often cause intrauterine death (Harteveld & Higgs, 2010; Higgs, 2013b). Infants with Hb Bart's hydrops fetalis have no α-globin produce, resulting in the formation of gamma-globin tetramers in the fetus. This tetramer has extremely high affinity for oxygen thus cannot transfer the oxygen to the fetus tissues effectively, leading to severe hypoxia, edema, plural, and pericardial effusion. Fetus with Hb Bart's hydrops usually die shortly after birth and mother carrying fetus with Hb Bart's increase the risk of maternal complication, miscarriage (Barts, n.d.; Curran et al., 2020; Vichinsky, 2009). In some rare cases, α-thalassemia association with myelodysplastic syndrome, and mental retardation, often association with other abnormalities development (Gibbons, 2012).

Table 1: Alpha-thalassemia allele genotypes

doi:10.6342/NTU202500218

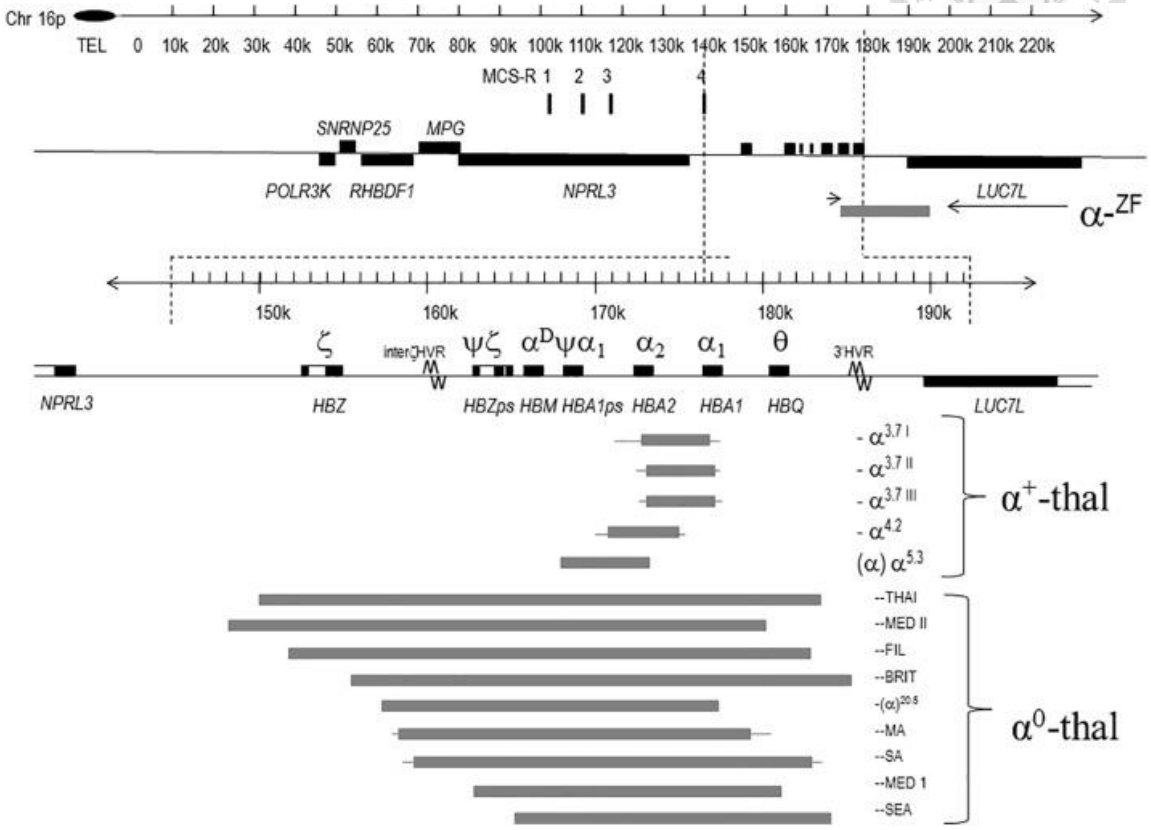
α-Thalassemia phenotypes	Number of α1 and α2 genes functional	Severity
Normal	4 functional α-globin αα/αα	No
Silent carrier	1 nonfunctional gene α-/αα	Asymptomatic
α-thalassemia trait	2 nonfunctional genes α-/α- or αα/--	Mild anemia, microcytosis
Hb H disease (deletional and non-deletional)	3 nonfunctional genes α-/--	Anemia, splenomegaly
Hb Bart's hydrops fetalis	4 nonfunctional genes --/--	Hydrops fetalis

1.1.3) Variance of α-thalassemia:

HBA1 gene has 349 reported small sequence variants (also called non-deletion variants) reported in the globin mutation database and the HBA2 gene has 455, large deletions are responsible for about 85% of patients with α-thalassemia, collected by HbVar database (Giardine et al., 2007). The range of α-thalassemia deletion is varied, could range from $(\alpha\alpha)^{RA}$ mutants (a deletion in the upper region of α-cluster which remain both α-cluster intact), varies length of α^0 deletion (deletion on both HBA1 and HBA2 gene) and varies length of α^+ (deletion on either gene). Often α-thalassemia is classified based on their size and specific breakpoints and named based on their geographic prevalence (Figure 1).

doi:10.6342/NTU202500218

Figure 1:: Summarizes deletion that give rise to α^0-thalassemia and α^+-thalassemia that is common in Southeast Asia region (Farashi & Harteveld, 2018).



As seen above, the solid boxes represent the abolished α-globin complex. Among of which, $--^{SEA}$, $--^{THAI}$, $--^{FIL}$ is very common in Southeast Asia countries (Higgs et al., 1989a), $--^{MED}$, $--(\alpha)^{20.5}$, $--^{SPAN}$, $-(\alpha)^{5.2}$ is more common in Mediterranean. Some other deletions are characterized found in black American, Italian (Cardiero et al., 2023), European (Villegas et al., 1994), Palestinian patients (Brieghel et al., 2015).

1.1.4) The importance of screening α-thalassemia

α-thalassemia mutations affect up to 5% of the world population (Vichinsky, 2013). α-thalassemia is most common in Southeast Asia, Mediterranean, as well as among those with Middle Eastern ancestry (Piel & Weatherall, 2014b) and increase with malaria

doi:10.6342/NTU202500218

infection (Piel & Weatherall, 2014c). The prevalence of α-thalassemia in Vietnam is 7.76% (Doan et al., 2022), Taiwan is 6.88% (Hsu et al., 2023). Due to the high prevalence of α-thalassemia among the population poses a health risk towards thalassemia patients. Additionally, α-thalassemia is at times confused with iron deficiency anemia. Women with thalassemia intermediate or major are posed risks with various maternal complication due to increase iron burden, such as cardiac failure, alloimmunization, viral infection, thrombosis, osteoporosis, new endocrinopathies, diabetes mellitus, hypothyroidism, etc.

An appropriate method for screening and detecting thalassemia is essential for couples with high risk, especially ones with carrier variants or living in high-frequency area (Old J et al., 2012). It helps to prevent the severe maternal complication in case of Hb Bart's fetalis or provide an accuracy detection of thalassemia with other hemoglobinopathies or anemia iron deficiency, or coinherited with Hb S or β-thalassemia. Also, it helps reduce the risk of having a child inheriting a combination of abnormal genes, resulting in more severe forms of thalassemia.

1.2) Method to detect α-thalassemia

Patients suspected of thalassemia will undergo several blood examinations typically involves a combination of several laboratory tests to accurately diagnose the condition. The initial screening includes a complete blood count (CBC), which assesses hemoglobin levels and red blood cell indices, such as mean corpuscular volume (MCV) to assess the average size of red blood cells and mean corpuscular hemoglobin (MCH) to quantify the average amount of red blood cells. These parameters help validate or identify microcytic and hypochromic anemia, common indicators for α-thalassemia. Following the CBC, a blood smear is examined microscopically to look for specific morphological change in red blood

doi:10.6342/NTU202500218

cells such as inclusion bodies or target cell, which are consistent with α-thalassemia (Vijian et al., 2021).

1.2.1) Traditional method to detect α-thalassemia:

a) Gap-polymerase Chain Reaction (Gap-PCR)

The Gap-polymerase chain reaction (Gap-PCR) technique is currently used to detect common deletion types of α-thalassemia. This method is widely used due to being less time-consuming and do not require labor intensive. However, the amplification of the α-globin cluster gene is challenging due to the high homology within the α-globin gene cluster, the unusually high GC content (with peaks reaching 70-80%), and the tendency to form secondary structures that interfere with primers. (Maria Domenica Cappellini, Alan Cohen, John Porter, Ali Taher, 2014). The Gap PCR method amplify the DNA sequences using designed primers to detect deleted regions and only detect known, several common types of deletion cause α-thalassemia including —THAI, —SEA, —FIL, —MED, $-\alpha^{20.5}$, $-\alpha^{3.7}$ and $-\alpha^{4.2}$. Gap PCR also faces the challenges while requiring post-PCR work, not suitable for large screening and detecting new variants or expand variants to other diseases (Vijian et al., 2021).

1.2.2) Next-generation sequencing to detect α-thalassemia:

DNA sequencing is a gold procedure to detect unknown variations (Molchanova et al., 1994). Next-generation sequencing (NGS) has accelerated the era of gene sequencing technique due to the capacity of whole genome sequencing of entire human genome. The capacity is increased thanks to the parallel sequencing (Vijian et al., 2021).

doi:10.6342/NTU202500218

It is important to use NGS to detect copy number variants (CNV). NGS, in advancement of genetic testing of human genetics, has been applied to detect genetic disorder, including α-thalassemia. Recently, when applied in detecting α-thalassemia, NGS was established to cover entire α-globin gene, regulatory regions and modifier genes (Munkongdee et al., 2020). NGS has many advantages in α-thalassemia diagnosis including detect rare and common mutations, not limited to point or structural variation. Additionally, NGS techniques are used as a screening method in areas with high prevalence of α-thalassemia population (He et al., 2017). Thus, NGS could be a powerful tool to assess in α-thalassemia carrier in the population.

There are many bioinformatics tools developed to detect CNV for NGS datasets including WES, WGS, targeted-genome sequencing. There are four main methods for CNV detection including read-pair (RP), split-read (SR), read-depth (RD) and assembly (AS) (Abel & Duncavage, 2013). Each algorithm composes of different strengths and weaknesses, for instance, RP method can detect medium-size CNV but not applicable for small indel events and low-complexity regions. RD, as the name suggests, measures the depth of coverage among genomic regions to detect the change in read-depth for CNV detection. First reads are aligned to reference genome and predefined windows are used to count RD(Janevski et al., 2012). RD methods can detect exacts amount of CNV, and better detection on large size CNV but have limitation on detect exact breakpoint of CNV. SR method is based on the mechanism that mapped reads on reference genome, reads that cannot align continuously or partially aligned are used to determine the breakpoint of structural variation. This gives SR an advantages on detecting exact breakpoint but only achieve high accuracy with NGS reads shorted than 1kb and reliable on unique regions of

doi:10.6342/NTU202500218

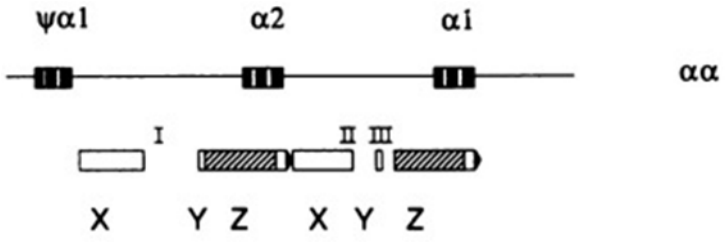
the genome (Zhang et al., 2011). Both RP and SR methods rely on the accuracy of mapping of each ends so these method cannot used to detect CNV on homologous regions such as HBA1 and HBA2 gene (W. Li & Olivier, 2013; Yoon et al., 2009). AS method in contrast generate contig or scaffold then compare against the reference genome to determine structural variation (Nijkamp et al., 2012). However, AS is less common than other methods because this method requires larger computation resources. Additionally, AS method is unable to handle haplotype sequences, only diploid type CNV can be detected, and perform poorly on complex regions due to the generation of contig/scaffold contains segmental duplication and repeats (Xi et al., 2012). Understanding the advantages and disadvantages of each computational method to choose the appropriate method to determine CNV on duplication genome regions such as α-globin cluster HBA1 and HBA2.

1.3) Difficulties in detecting α-thalassemia

HBA1 and HBA2 genes are paralogous genes because they arose from α gene duplication event, located on the short-arm telomere of chromosome 16. The human α-gene duplication occurs around 72 to 52 million years ago (Hardison, 2012; Higgs et al., 1989b). Although there was divergence occurring along the evolutionary pathway, HBA1 and HBA2 genes are highly similar. Alpha-globin genes are embedded in highly homologous 4kb duplication, divided into homologous segments (X, Y, and Z) by nonhomologous elements (I, II, and III)

Figure 2: High similarity of alpha-globin cluster gene

doi:10.6342/NTU202500218



This high similarity poses challenging to detect point mutations or structural due to several instinct and technical factors, such as ambiguous reads alignment (Figure 2).

Ambiguities between HBA1 and HBA2 genes created from duplication, in alignment or assembly can produce errors and bias in interpreting result (Treangen & Salzberg, 2011). NGS, featuring by short reads and small genomic fragments size, posed difficulties in alignment issues (for instance, unique strategy, all match strategy, best match strategy) (Treangen & Salzberg, 2011).

1.3.1) Overcome the struggle when detecting α-thalassemia

There are different approaches trying to solve this problem, for name, NGS4THAL is developed aimed to detect both point/short indels mutation and structural (Cao et al., n.d.) NGS4THAL realigns ambiguous mapped reads derived from the homologous Hb gene cluster for an accurate detection of α-thalassemia point and short indel mutation. Additionally, this pipeline uses a combination of CNV detection tools to increase the detection of complex structural variants (SV) and matches those variants with HbVar Database. These steps allow for the identification of pathogenic variants. At first, NGS4THAL was also tried to install to detect α-thalassemia causative variants in NTUCH, but the installation process encountered a lot of difficulties. Not to mention NGS4THAL pipeline is not regularly maintained, thus tools in that pipeline might not be updated, not

doi:10.6342/NTU202500218

user-friendly nor practically a “one-stop” tool. It is a combination of existing “black-box” tool, due to the difficulty in implementing module of NGS4THAL pipeline (GitHub - Jielab/Pigeon: Practical Investigation of Genomic Errors by Observation and Notification, n.d.).

This led to a requirement to establish another pipeline that is more comprehensive and maintainable in detecting α-thalassemia. AlphaThalCNV is established to address this need by integrating advanced methods to detect CNVs with high precision, reliability, and repeatability.

1.3.2) GATK gCNV-caller can be used to detect CNV that causes α-thalassemia

Detecting CNVs has an important clinical role in detecting several cancers and diseases, especially α-thalassemia. However, detecting CNV are sometimes face difficulties when some commons structural variations are complex, relies on large part of human genome (Gabrielaite et al., 2021a). Especially α-thalassemia is caused mostly by deletions in high homology HBA1 and HBA2 (Figure 1) when performing the nucleotide blasting by NCBI. Due to this homology, reads mapped to these regions can result in ambiguous read alignments thus ineffectively detect causative variants.

Thus, it is essential to optimize our in-house pipeline for α-thalassemia detection, GATK-gCNV caller is used for detection copy-number variants. GATK-gCNV caller is used because it is one of the best tools among 50 others CNV calling tools benchmarked for the best CNV callings. To be specific, GATK-gCNV is benchmarked to call more deletions and duplications compared to other tools and achieve higher precision and recall rate more than other tools in both WES and WGS data (Gabrielaite et al., 2021b). In addition, GATK-

doi:10.6342/NTU202500218

gCNV has been used for acute discovery of rare and common copy-number variations (CNVs) from read-depth data obtained either from WGS, WES, or custom panel sequencing. GATK-gcNV calling is composed of two main workflows: model creation from cohort mode and individual sample calling in case mode (Babadi et al., 2023).

1.4) Research aim

It is necessary and crucial to have a pipeline that detects copy number variants (CNVs) on α-globin gene that cause α-thalassemia. AlphaThalCNV is a pipeline that focuses on detecting deletion on homologous α-globin regions focus on accurately detecting true length of α-thalassemia deletion.

II) Method

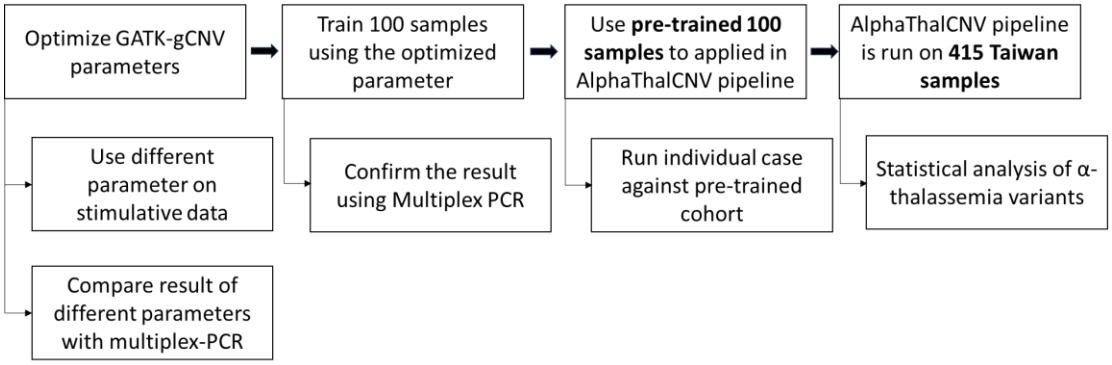
2.1) Study design

The study design is composed of four main steps: optimize GATK-gCNV parameters, then use those parameters to train 100 samples to obtain a cohort. This cohort then integrated into AlphaThalCNV pipeline then ran on 415 Taiwan sample (Table 3). To be specific demonstrate, GATK-gCNV parameters are optimized using different parameters combination on three labeled data (whole-exome sequencing data with confirmed multiplex-PCR result) and other samples. Smaller bin size and larger padding region (400 bp) combination are used compared with the previous pipeline that use smaller padding regions (250 bp) and disable binning. Reducing bin size increases the resolution in detecting breakpoint and detecting CNV but increases the noise and computational resources. In this case, GATK-gCNV is run in the cohort mode focusing on alpha-globin cluster gene only. After having the combination of bin size and padding regions that prove

doi:10.6342/NTU202500218

to be optimized for α-thalassemia, these parameters are used to train a cohort of 100 samples and validate the result by multiple-PCR. The cohort then integrated into AlphaThalCNV which used the cohort as a pretrained model to run against individual cases. AlphaThalCNV then runs on 415 samples from NTUCH for statistical analysis.

Figure 3: Summary of study design



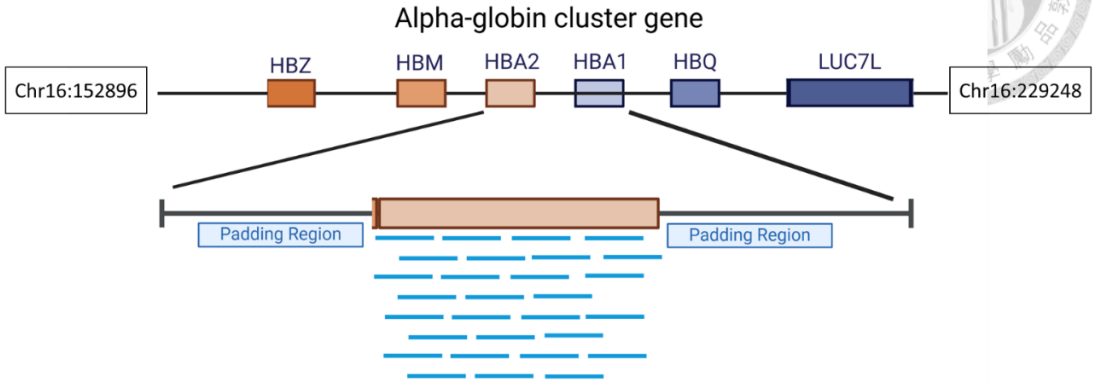
2.2) Data Collection and Preparation

2.2.1) Optimizing the model

The PreprocessInterval is performed to prepare bin for coverage collection in the first phase of creating intervals list and counts read alignments overlapping the intervals. This step pads exome target regions and bins intervals. The term binning refers to creating equally sized interval across the reference. For instance, 100-base binning would define ch1:1-100 as the first bin to mapped to human genome. In this optimize model, the focus is to increase the resolution of detecting deletion spans over α-globin cluster, the bin is cut down into smaller size instead of default parameter that set –bin-length 0 to disable binning. The intervals padding in increased higher than default –padding 250 to target intervals (Figure 4).

doi:10.6342/NTU202500218

Figure 4: Illustration of the PreprocessInterval step when preparing bin size and padding regions.



In this part, different combinations of bin size and padding are used to find the best parameter to detect α-thalassemia deletion on α-globin cluster gene (Table 2). The combination is listed in the table below:

Table 2: Different parameters combination used to determine the most optimum to detect α-thalassemia

Mode	Bin size	Padding regions
WES	5	400
WES	20	400
WES	50	400
WGS	30	400
WGS	100	0

The result is compared with the multiplex PCR to see what parameter would be most optimized.

doi:10.6342/NTU202500218

2.2.2) Training Dataset

A cohort of 100 anonymized samples are randomly selected from NTUCH's genetic repository. Fastq data by WES of 100 samples are first underwent a preprocessing step including FASTQ preprocessing to trim off adapter sequences and remove poor quality reads by fastp (Chen et al., 2018). The parameters are set to disable quality filtering and disable length filtering to retain the reads, disable the duplication and trim polyG read tails by minimum length is 4. Processed reads then aligned to the reference genome hg38 using Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2009) to generate alignment *bam* file. The pipeline processed each sample individually, tagging the output with the sample name for tracking. A read group identified (RG) was added to the alignment files to specify for sample name and sequencing platform. Aligned reads were then processed using GATK MarkDuplicatesSpark (McKenna et al., 2010) to identify and mark the duplicate which results from library preparation. This step ensures the accuracy variant calling and detection of CNV by read-depth algorithms using GATK-gCNV (Babadi et al., 2023). Additionally, GATK BaseRecalibratorSpark (McKenna et al., 2010) is used to improve base quality score accuracy and minimize systematic bias made by sequencing machines. The recalibrated base quality scores generated then applied to the aligned reads (bam files) using GATK ApplyBQSR to recalibrate. These preprocessing steps to avoid the bias and systematic error toward reads for downstream application.

After raw data preprocessing, the recalibrated bam files are used to train the model by GATK-gCNV *cohort-mode* follow GATK instruction (*(How to) Call Rare Germline Copy Number Variants – GATK*, n.d.). First, raw counts data is collected with PreprocessIntervals and CollectReadCounts. PreprocessIntervals pads exome target regions, padding parameter

doi:10.6342/NTU202500218

equal to 400 is chosen to increase the region outside targeted genome. Additionally, bin length is increased to 50 instead of default disable bin length to increase resolution in detecting CNV in targeted region. This produces a Picard-style interval list of exome regions padded by 400 bases on either side of each targeted region, with smaller bin size is 50. However, by increasing bin size will also consume computational memory resources and time inefficiency, the target region focus on Chr16:152,896 – 229,248, which cover α-globin cluster gene from HBZ, HBM, HBA1, HBA2, HBQ and LUC7L promoter. CollectReadCounts then applied to calculate the read depth across targeted genome intervals to generate output files in HDF5 format. The read count data HDF5 file the used for ploidy genomic contigs determined by DetermineGermlineContigPLoidy , calling contig level ploidies for both autosomal (e.g. human chr20) and allosomal contigs (e.g. human chrX). This step determines contig ploidies using total read count per contig to create a contig model of a cohort set thus will be used against individual cases for more efficient CNV detection. The last step is identification of germline CNVs, performed using GATK's GermlineCNVCaller. GATK's GermlineCNVCaller's denoising model ensures consistent CNV detection and is highly sensitive of both rare and common CNVs.

By training 100 samples, the output included the gCNV interval list, cohort ploidy model, cohort CNV model and preprocessed annotated. These files are integrated into gATKalphaThal in-house pipeline to target the detection of α -thalassemia. The other sequence files (.seq) are used to visualize CNV results and alignment files.

2.2.3) Test Dataset

After having cohort trained and obtained the .seq files, multiplex-PCR is used to confirm the results of trained cohort. As most of the α-thalassemia mainly cause by

doi:10.6342/NTU202500218

common deleterious variants, for example: --$\alpha^{3.7}$, --$\alpha^{4.2}$, -- ^{SEA}, --^{FIL}, --^{THAI}, multiplex-PCR are developed to detect of this common deletion. The problem in amplification of GC-rich α-locus is overcome by utilizing betadine and dimethyl sulphoxide (DMSO), together with optimize thermocycle to the robust and reproducible PCR. Multiplex PCR is separated into two reactions, one to detect rightward deletion include --$\alpha^{3.7}$, other reaction to detect –$\alpha^{4.2}$, -- ^{SEA}, --^{FIL}, --^{THAI}. This reaction uses multiple sets of primers, each specific primers design to amplify region flanking the deletions (Liu et al., 2000). The primers list is adapted from (Liu et al., 2000)

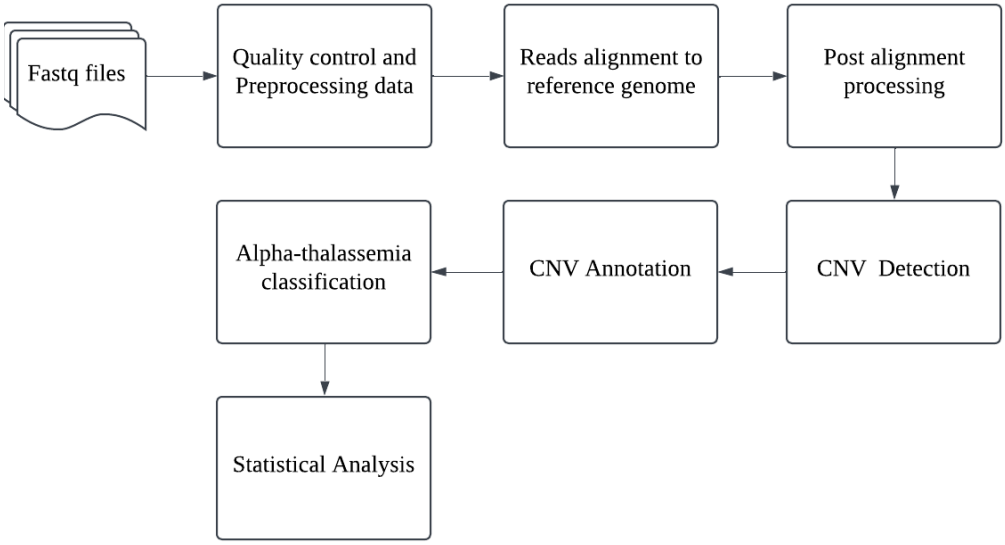
2.3) Pipeline Development and Implemetation

2.3.1) Pipeline implementation:

After having trained model using cohort-mode from GATK, the model is used to detect CNV in the in-house pipeline to detect CNVs cause α-thalassemia in separated individuals. The process is described as figure above, as the preprocessing, alignment and post alignment processing is conducted same as processing step in training dataset. Next, CNV analysis was performed using GATK-gCNV in case mode to detect CNV with GermlineCNVCaller, which models systematic biases and identifies deletions and duplication. The detected CNV then annotated using AnnotSV (Geoffroy et al., 2018), which integrated genomic coordinates with functional annotation to classify and interpret the clinical significance of the CNV. AnnotSV compiles functional, regulatory and clinical relevance information that provide helpful information to interpret structural variants pathogenicity and filter in true positive variants (Figure 5).

Figure 5: Overview of AlphaThalCNV pipeline data analysis in detecting α-thalassemia variants from fastq file.

doi:10.6342/NTU202500218



2.3.2) Alpha-thalassemia classification:

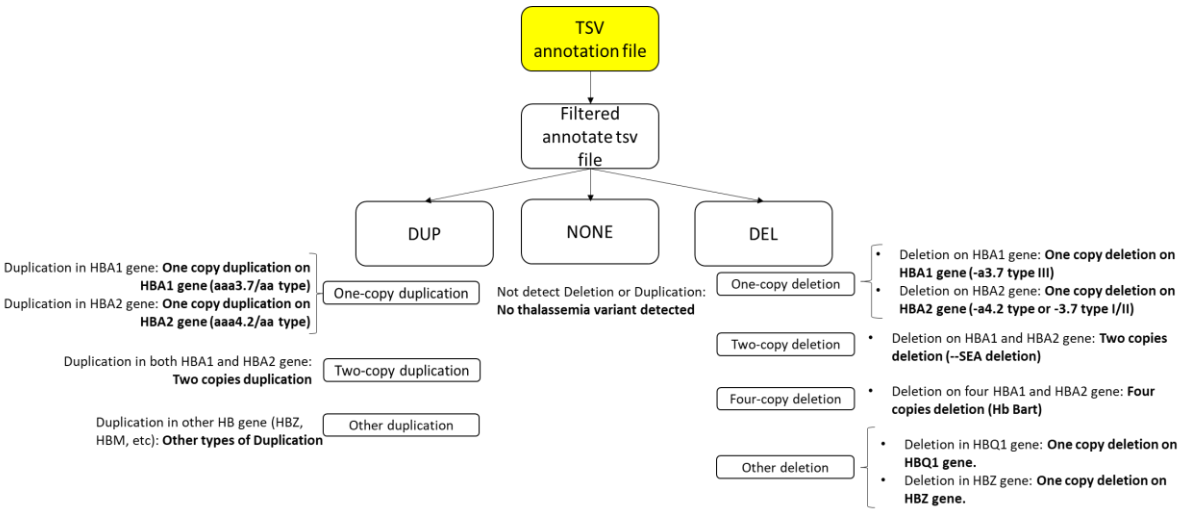
Alpha-thalassemia classification was performed by analyzing the number of deletion and locations on HBA cluster gene. This process was implemented using a custom-developed pipeline, designed with the following steps.

First, variant classification. The pipeline initially identifies and categorizes structural variants as either deletions ("DEL") or duplications ("DUP"). The next step is targeted filtering which variants are filtered to isolate those specific to the HBA gene cluster. This involves retaining only variants located on chromosome 16 and those with identifiers containing "HB.". The next step is quantitative analysis, the algorithm then determines the number of HBA genes affected by deletions or duplications and the gene name affected. For instance, if one gene is deleted, the algorithm will check if the deletion is located on HBA1 or HBA2 gene. Deletion on HBA2 gene will generate classification output “One copy deletion on HBA2 gene (-4.2 or -3.7 type I/II). If two genes deleted over HBA1 and HBA2, the output is “Two copies deletion (-SEA type). If deletion is detected on other HB genes

doi:10.6342/NTU202500218

such as HBZ, HBM, the classification will output accordingly. In summary, the classification output provides as below for each sample (Figure 6).

Figure 6: Overview of α-thalassemia classification workflow.



2.3.3) Cohort statistical analysis

This pipeline was applied in a cohort of 415 individuals, data obtained from WES analysis from National Taiwan University's Children Hospital, enabling a comprehensive analysis α-thalassemia carrier. Statistical methods were employed to determine carrier frequencies, provide valuable insights into the Taiwan's population prevalence of α-thalassemia.

2.4) Multiplex-PCR confirms α-thalassemia variants

Multiplex-PCR is a widely used molecular method for detecting common α-thalassemia deletions, involving one or both α-globin genes (HBA1 and/or HBA2) located on chromosome 16. Multiplex-PCR enables the amplification of multiple target DNA in a single PCR reaction, allowing for the simultaneous detection of multiple deletion associated with α-thalassemia. In this method, primers are designed to target and amplify

doi:10.6342/NTU202500218

the junction fragment, flank the breakpoint, identifying by size. The primer list for specific deletion is described below. A positive PCR amplification indicates that there is deletion in the genomic segments (Baysal & Huisman, n.d.; Chong et al., 2000; De Mare et al., 2010b).

Multiplex PCR is conducted in Dr. Tsang Ming Ko's lab. Twelve samples from test data set were used to run Multiplex PCR to confirm the accuracy of the optimized parameter on detecting α-thalassemia variants (Table 3). A total of 7 samples with no α-thalassemia deletion are used to determine true-negative detection and 5 samples with α-thalassemia deletion are used to determine true-positive detection.

Table 3: Primers sequence target 5 most common α-thalassemia (α^0 thalassemia, --SEA, --Thai, --Fil; α^+ thalassemia -3.7, -4.2) and amplicon size.

Primer	Sequence(s) 5' -> 3'	Amplicon (size)
SEA-F	CGATCTGGGCTCTGTGTTCTC	$--^{SEA}$ junction fragment (1349 bp)
SEA-R	AGCCCACGTTGTGTTCATGGC	
THAI-F	GACCATTCCTCAGCGTGGGTG	$--^{Thai}$ junction fragment (1155 bp)
THAI-R	CAAGTGGGCTGAGCCCTTGAG	
FIL-F	TTTAAATGGGCAAAACAGGCCAGG	$--^{Fil}$ junction fragment (546 bp)
FIL-R	ATAACCTTTATCTGCCACATGTAGC	
3.7-F	CCCCTCGCCAAGTCCACCC	$--^{3.7}$ junction fragment (2100 bp)
3.7-R	AAAGCACTCTAGGGTCCAGCG	

doi:10.6342/NTU202500218

4.2-F	GGTTTACCCATGTGGTGCCTC	--$^{4.2}$ junction fragment (1628 bp)
4.2-R	CCCGTTGGATCTTCTCATTTCCC	

The detection is divided into two sets of reaction. The first set is aimed to detect leftward deletion type including –^{SEA} type, --^{Fil} type, --^{Thai} type, -$^{4.2}$ types, while the second set aimed to detect rightward deletion including αα/-$\alpha^{3.7}$. Different sizes of amplification products will show the type of deletion. For instance, in the first set, PCR product of a normal sample is 302 bp length, --^{SEA} type is 1349 bp, --^{Fil} type is 546 bp, --^{Thai} type is 1155 bp and –$^{4.2}$ type is 1628 bp. In the second set, the positive result with rightward deletion -$\alpha^{3.7}$ has the PCR amplicon length 1900 bp while the normal case's amplicon length is 2100 bp.

Running multiplex PCR, a total 25 µl reagent is used for each sample include 2.50µl RBC 10X Reaction Buffer with 15 mM $MgCl_2$, 2.00 µl dNTP(2.5 mM), 3.90 µl Primer mix (10 µM each reverse/forward primer) , 0.20 µl RBC SensiZyme DNA Polymerase, 15.60 µl ddH_2O, 1.00 µl Genomic DNA (concentration lower than 500 ng). The cycling program comprises of initial denaturation at 95^{0}C for 10 mins, followed by 35 thermal cycles consist of denaturation at 94^{0}C for 1 min, DNA primers and DNA polymerase annealing at 65^{0}C for 1 min, then sequence elongation at 72^{0}C for 2 mins. A final elongation step is conducted at 72^{0}C for 10 mins after finishing 35 thermal cycles, subsequently, the amplicon is stored at 25^{0}C or lower until used for electrophoresis. Ten microliters of each amplification product is analyzed using 2% gel electrophoresis for the first set and 1% agarose gel for the second set in 1X Tris-Borate-EDTA buffer solution in one hour. The agarose gel then visualized under UV light to check the amplicon length (Chong et al., 2000; De Mare et al., 2010b; Tan et al., 2001).

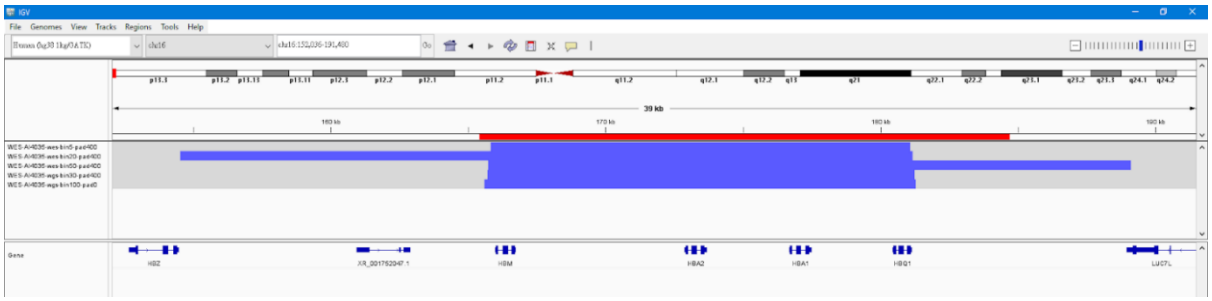
doi:10.6342/NTU202500218

III) Result

3.1) GATK-gCNV with optimized parameter increases accuracy of detection of α-thalassemia's variants length.

Case WES-1: IGV visualization of case WES-1's seg files while optimized with different binning/padding parameters. Blue segment belongs to case WES-1 detected deletion (below) while the red segment (above) shows the true length of the $-^{SEA}$ deletion which around 20.5 kb and span over HBA1 and HBA2 gene. From top to bottom are detected deletion with different binning/padding parameters; Line 1: WES mode, bin 5, padding 400; Line 2: WES mode, bin 20, padding 400; Line 3: WES mode, bin 50, padding 400; Line 4: WGS mode, bin 30, padding 400; Line 5: WGS mode, bin 100, padding 0 (Figure 7).

Figure 7: IGV visualization of case WES-1



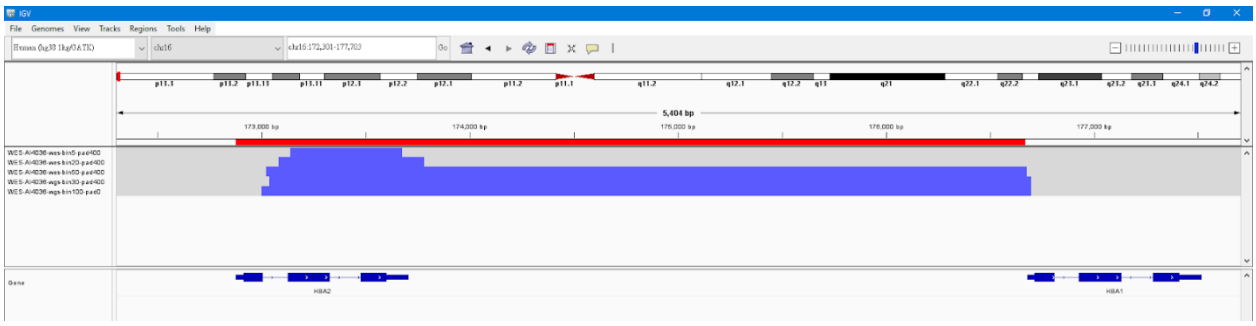
Multiplex PCR result of case WES-1 show that this case has -sea deletion, length around 20.5kb (annotated in red segment). The optimized parameter that increases the length of deletion detection for case WES-1 is WES mode, with a bin size of 50 and a

doi:10.6342/NTU202500218

padding region of 400, which enhances the precision and accuracy of the analysis. Other parameters are not accurately detecting the length of -^{SEA} α-thalassemia deletion.

Case WES-2: IGV visualization of case WES-2's seg files while optimized with different binning/padding parameters. Blue segment belongs to case WES-2 detected deletion (below) while the red segment (above) shows the true length of the –$^{3.7}$ deletion which around 3.7 kb and span over HBA2 gene and a part of HBA1 gene. From top to bottom are detected deletion with different binning/padding parameters; Line 1: WES mode, bin 5, padding 400; Line 2: WES mode, bin 20, padding 400; Line 3: WES mode, bin 50, padding 400; Line 4: WGS mode, bin 30, padding 400; Line 5: WGS mode, bin 100, padding 0 (Figure 8).

Figure 8: IGV visualization of case WES-2



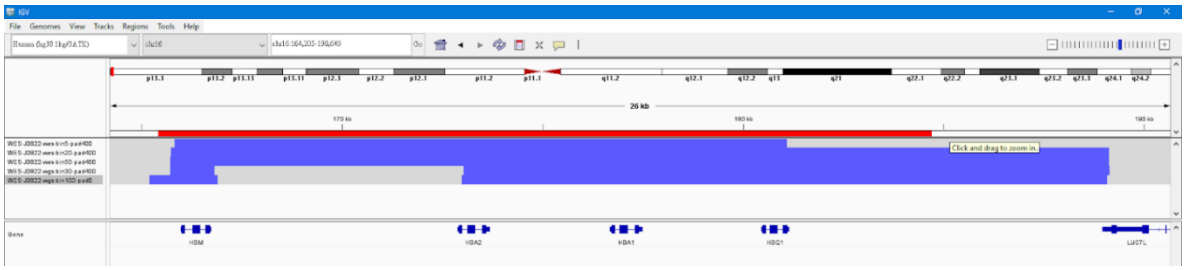
This result shows that parameter WES mode, bin size 50 and padding region 400 obtain the highest accuracy toward -$^{3.7}$ α-thalassemia deletion. The deletion detected by this

doi:10.6342/NTU202500218

parameter is the same length (blue segment) as true -$^{3.7}$ α-thalassemia which is 3700 bp deletion and span over HBA2 gene and a part of HBA1 gene.

Case WES-3: IGV visualization of case WES-1's seg files while optimized with different binning/padding parameters. Blue segment belongs to case WES-3 detected deletion (below) while the red segment (above) shows the true length of the –^{SEA} deletion which around 20.5 kb and span over HBA1 and HBA2 gene. From top to bottom are detected deletion with different binning/padding parameters; Line 1: WES mode, bin 5, padding 400; Line 2: WES mode, bin 20, padding 400; Line 3: WES mode, bin 50, padding 400; Line 4: WGS mode, bin 30, padding 400; Line 5: WGS mode, bin 100, padding 0 (Figure 9).

Figure 9: IGV visualization of case WES-3



Multiplex PCR result of case WES-3 show that this case has -sea deletion, length around 20.5kb (annotated in red segment). The optimized parameter that increases the length of deletion detection for case WES-3 is WES mode, with a bin size of 50 and a padding region of 400, which enhances the precision and accuracy of the analysis. Other parameters are not accurately detecting the length of -^{SEA} α-thalassemia deletion or creating a gap inside deletion.

Summary of different bin/pad parameters (Table 4)

doi:10.6342/NTU202500218

Table 4: Summary of each parameter in detecting α-thalassemia

Parameter	Case WES-1	Case WES-2	Case WES-3
WES-bin5-pad400	Del ~15500bp (HBM -> HBQ1)	Del ~700 bp (HBA2)	Del ~ 15,280 bp (HBM -> HBQ1)
WES-bin20-pad400	Del ~26650 bp (HBZ -> HBQ1)	Del ~800 bp (HBA2)	Del ~23409 bp (HBA1 -> Luc7L)
WES-bin50-pad400	Del ~23400 bp (HBA1 -> Luc7L)	Del ~3700 bp (HBA2 -> HBA1)	Del ~23409 bp (HBA1 -> Luc7L)
WGS-bin30-pad400	Del ~15500bp (HBM -> HBQ1)	Del ~3700 bp (HBA2 -> HBA1)	Del ~23409 bp (Gap inside)
WGS-bin100-pad0	Del ~15500bp (HBM -> HBQ1)	Del ~3700 bp (HBA2 -> HBA1)	Del ~23409 bp (Gap inside)
Multiplex PCR Result	$--^{SEA}$ thalassemia (deletion ~20 500 bp)	$--^{3.7}$ thalassemia (deletion ~3 700 bp)	$--^{SEA}$ thalassemia (deletion ~20 500 bp)

In conclusion, bining size 50 and padding 400 has the highest accuracy when detecting α-thalassemia CNV because the CNV result is the most similar to multiplex PCR result. Thus, this parameter is chosen to train a cohort of 100 samples for further applied in the AlphaThalCNV pipeline.

doi:10.6342/NTU202500218

3.2) AlphaThalCNV detects more cases carrying thalassemia variants compared to previous pipelines.

After using 100 samples from NTUCH as a training model with optimized parameters WES mode, bin 50 and padding 400, the result reveals new cases have CNV on α-globin gene that potentially cause α-thalassemia. Some of these cases are not detected or not clearly detect for α-thalassemia using our normal pipeline (Table 5). These cases are further confirmed by the presence of α-thalassemia by multiplex-PCR.

Table 5: Comparison of mutation detection between optimized parameters GATK-gCNV and our previous pipeline.

Case ID	Mutation detected by optimized parameters	CNV length detected by optimized parameters	Previous pipeline report	Remarks
WES-4	Chr16:168600-176680 (span over HBA2 and a part of HBA1 gene)	Deletion 8,080 bp	Chr16:172824-173711 820 bp (span only on HBA2 gene)	Improved detection
WES-5	Chr16:165186-189036	Deletion 23,85 kb	Chr16:154277-181104 15.12 kb	Improved detection

doi:10.6342/NTU202500218

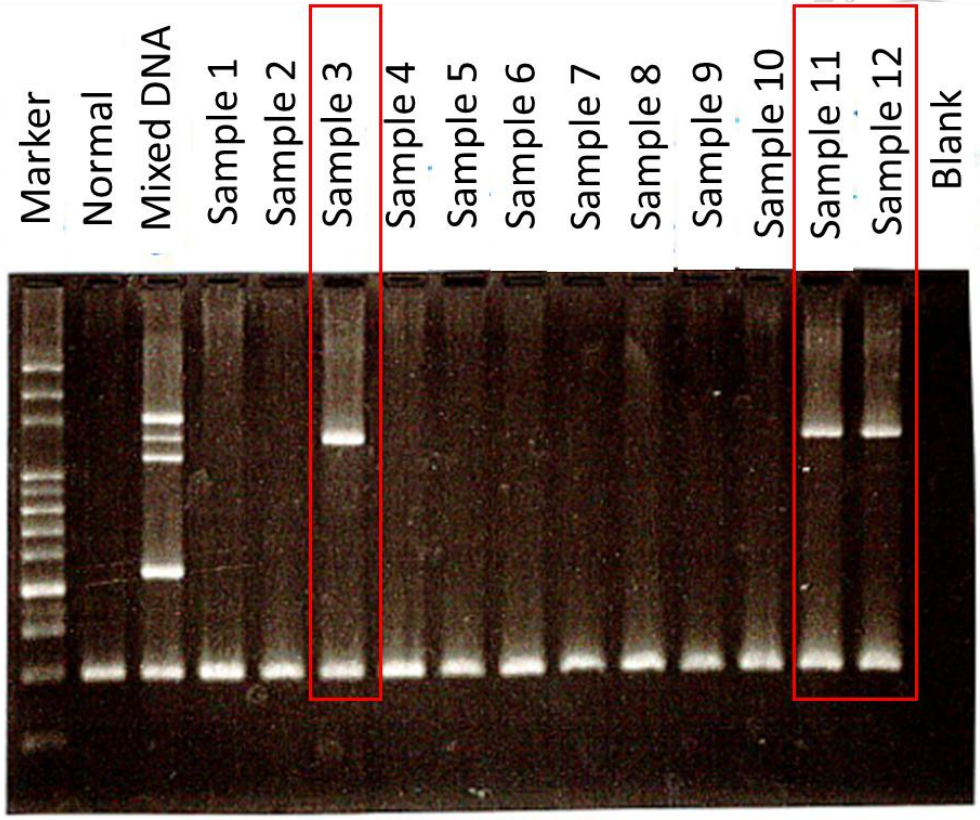
WES-6	Chr16: 173013-172879 (span only over HBA2 gene)	Deletion 866bp	Not detected	Improved detection of false positive result
WES-7	Chr16:165186-189036	Deletion 23,85 kb	Chr16:165984-181104 15.12 kb	Improved detection
WES-8	Chr16:165186-189036	Deletion 23,85 kb	Chr16:165984-181104 15.12 kb	Improved detection
WES-9	Chr16:165186-189036	Deletion 23,85 kb	Chr16:165984-181104 15.12 kb	Improved detection

3.3) Multiplex-PCR confirms the presence of α-thalassemia variants detected by AlphaThalCNV pipeline.

Set 1: Multiplex result for $-^{SEA}$ deletion (Figure 10)

Figure 10: Multiplex PCR result of --SEA deletion. (Multiplex PCR is performed by Dr. Tsang-Ming Ko's Lab)

doi:10.6342/NTU202500218



Sample 1,2,4,5,6,7,8,9,10 show negative with deletion; Sample 3,11,12 show positive with heterozygous αα/--^{SEA} deletion. (Normal sample: 302 bp, --^{SEA} type: 1349 bp) (Table 6)

Table 6: Summary of deletion length detected by GATK-gCNV optimized parameters and the multiplex PCR result for cases with –SEA α-thalassemia.

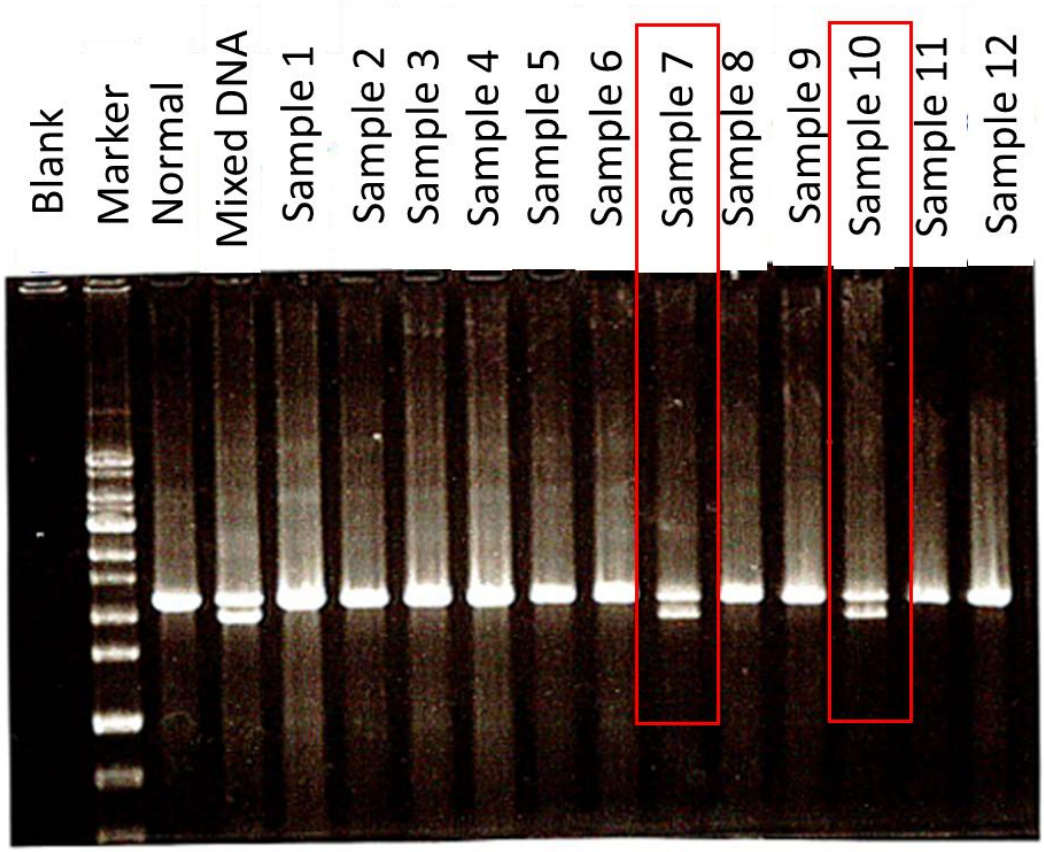
Sample ID	Case ID	Mutation detected	Multiplex PCR result
Sample 3	WES-5	Chr16:165186-189036 Deletion 23,85 kb	--SEA deletion

doi:10.6342/NTU202500218

Sample 11	WES-7	Chr16:165186-189036 Deletion 23,85 kb	--SEA deletion
Sample 12	WES-9	Chr16:165186-189036 Deletion 23,85 kb	--SEA deletion

Set 2: Multiplex result for -$^{3.7}$ deletion (Figure 11)

Figure 11: Multiplex PCR result of -3.7 deletion. (Multiplex PCR is performed by Dr. Tsang-Ming Ko's Lab)



Sample 1,2,3,4,5,6,8,9,11,12 are negative result with αα/--$^{3.7}$ deletion; sample7,10 are positive result with heterozygous αα/--$^{3.7}$ deletion. (Normal: 2100 bp, -$^{3.7}$ deletion: 1800 bp) (Table 7)

doi:10.6342/NTU202500218

Table 7: Summary of deletion length detected by GATK-gCNV optimized parameters and the multiplex PCR result for cases with –3.7 α-thalassemia.

Sample ID	Case ID	Mutation detected	Multiplex PCR result
Sample 7	WES-4	Chr16:168600-176680 8,080 bp (Deletion span over HBA2 and a part of HBA1)	--3.7 deletion
Sample 10	WES-6	Chr16: 173013-172879 866bp (Deletion of HBA2 gene only)	--3.7 deletion

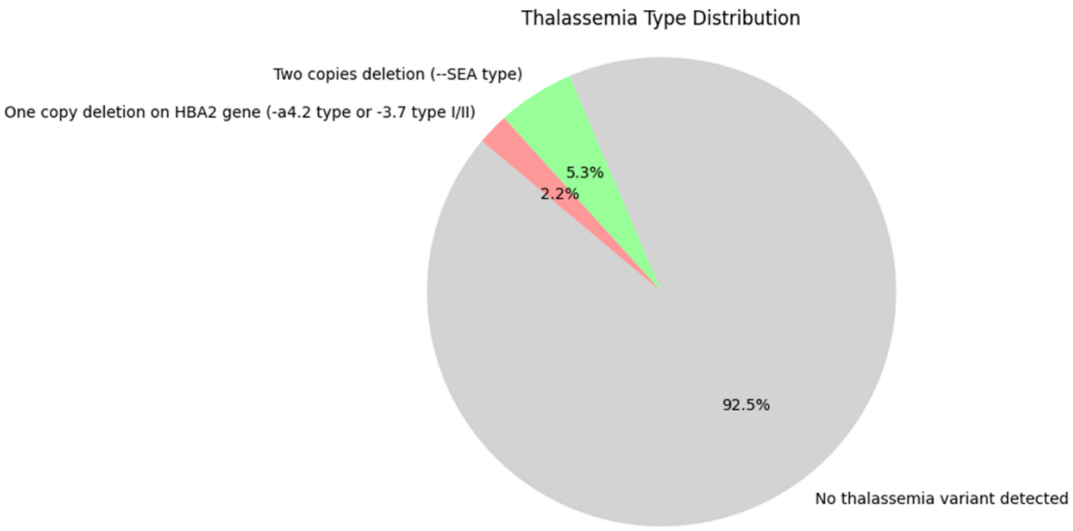
3.4) AlphaThalCNV applied in Taiwan's cohort to find the α-thalassemia statistical carrier in Taiwan.

After obtaining 100 samples trained for the cohort, this cohort is integrated into AlphaThalCNV pipeline to run against individual cases, aiming for a precise detection of α-thalassemia. AlphaThalCNV was applied to run on total 415 samples from National Taiwain University's Children Hospital to find proportional of variants carriers in Taiwan population. The result showed that there are a total of 384 samples that has no α-thalassemia CNV detected, taking up to 93% of whole cohort. Twenty-two samples detected with Two copies deletion (both HBA1 and HBA2 gene) take 5.3 % and 9 samples detected with One copy deletion HBA2 gene ($-\alpha^{4.2}$ types or $-\alpha^{3.7}$ type I/II)(Figure 12). This

doi:10.6342/NTU202500218

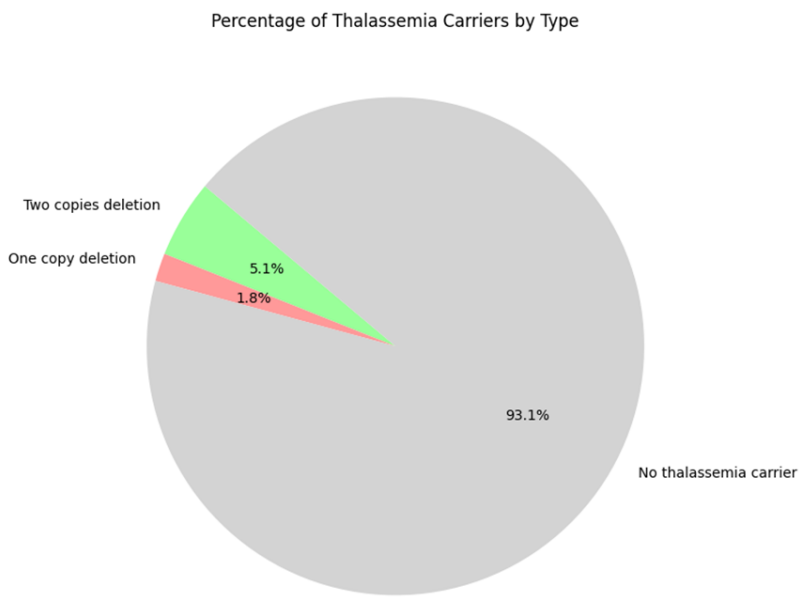
result aligns with previous research of (Hsu et al., 2023) about the α-thalassemia variant carriers of Taiwan populations (Figure 13).

Figure 12: Distribution of α-thalassemia variants in Taiwan cohort, samples obtained from NTUCH, run by AlphaThalCNV pipeline.



doi:10.6342/NTU202500218

Figure 13: Distribution of α-thalassemia variants in Taiwan cohort, data obtained from (Hsu et al., 2023).



IV) Discussion:

4.1) Advantages of this AlphaThalCNV pipeline

4.1.1 AlphaThalCNV pipeline increases the accuracy of detection α-thalassemia length.

AlphaThalCNV increases the accuracy of detection α-thalassemia length. The modified bin size and padding regions have improved the length of the α-thalassemia CNV. The $–^{SEA}$ α-thalassemia is reported to length around 19-20.5 kb and span over HBA2 and HBA1 gene. The optimized parameters have improved the detection length from around 15500 bp to approximately 23400 bp, closer to the true length of the $–^{SEA}$ α-thalassemia. This improvement also implied in detection of $α^+$ deletion as well when deletion spans over one gene copy. The previous pipeline are failed to detect true length deletion of $-^{3.7}$ α-

doi:10.6342/NTU202500218

thalassemia or fail to detect $-^{3.7}$ deletion. For instance, after modification, case WES-2 deletion finding is increased the precision or case WES-6 is detected to have small deletion of HBA2 gene that later confirm by multiplex PCR to have $-^{3.7}$ α-thalassemia while the previous pipeline did not show any alteration.

4.1.2) AlphaThalCNV increases the robustness of detecting α-thalassemia

AlphaThalCNV is a robust pipeline that can be used to run on multiple samples at a time, that helps researchers to reduce time-consuming on running on large cohort. AlphaThalCNV is written in Nextflow coding language, enabling the efficiently handling multiple samples through parallel execution. This parallelization reduces runtime significantly, making the pipeline suitable for either single sample or large dataset (DI Tommaso et al., 2017). The AlphaThalCNV pipeline also leverages Nextflow's dynamic input management to process multiple effortlessly, eliminating the need for manual intervention in data preparation or workflow execution. Additionally, besides the advantage of being a robust pipeline, AlphaThalCNV is easy customization thanks to the modular design of Nextflow pipeline allows for the integration of new tools or workflows with minimal efforts (Ewels et al., 2020). Thus, in the future research expand, AlphaThalCNV may integrated with more bioinformatics tools that aim to find point mutation that cause α-thalassemia. This might enhance its diagnosis capacity by enabling the detection of both CNVs and single nucleotide variants (SNVs), providing a comprehensive approach to identifying most of the genetic alterations associated with α-thalassemia.

In addition, AlphaThalCNV is designed to focus on the chromosome 16 by modifying bed file only to focus on α-globin locus instead of the whole exome. This modification can

doi:10.6342/NTU202500218

reduce time consumption and avoid computational inefficiency. Each sample only took around 1 hour and a half to finish analysis.

4.2) Limitation

4.2.1) AlphaThalCNV is limited to determine compound α-thalassemia.

The use of GATK-gCNV in AlphaThalCNV for detecting α-thalassemia CNVs is limited in compound cases such as Hongkong α-thalassemia, characterized by a 3.7 kb deletion on one allele and a 4.2 kb duplication on the other allele. There is one case in NTUCH only characterized with $\alpha\alpha\alpha^{4.2}/-\alpha^{3.7}$ using multiplex-PCR but failed to detect CNV by AlphaThalCNV pipeline. This limitation might arise from the compensation of mapped reads between the deletion region and the duplication region. Normally, duplication regions might increase the number of reads mapped while deletion regions might decrease the number of reads mapped. However, in the case of compound α-thalassemia, reads from duplicated regions might be mapped into the deleted region, thus balance out the total reads over regions. Read alignment tools like BWA often struggle with mapping in repetitive or homologous region due to ambiguous mapping. Reads spanning the breakpoints of deletion or duplication may be mapped incorrectly or split between regions or entirely discarded (Musich et al., 2021). GATK-gCNV employed the read-depth count to determine the deletion/duplication therefore could not accurately identify correct deletion and duplication region of this samples. Read depth of this sample also visualized by IGV showed that there is no significant difference in the suspected duplication and deletion region compared to normal sample. Publications also showed that Hongkong α-thalassemia is detected using PCR-based method, MLPA assay or long-read sequencings; this might explain (Third-

doi:10.6342/NTU202500218

generation sequencing), not short-read sequencing (De Mare et al., 2010a; J. Li et al., 2022).

4.2.2) AlphaThalCNV is not designed to detect point mutations.

The use of AlphaThalCNV pipeline also does not focus on detecting point mutations causing α-thalassemia, thus might miss out more than 70 forms of non-deletional mutations (Kalle Kwaifa et al., 2020) The most common α-thalassemia non-deletional mutations reported in Southeast Asia including Hb Constant Spring (Hb CS), Hb Quong Sze (Hb QS), and Hb Adana (Vijian et al., 2023)or reported at other regions such as Hb Icara, Hb Chesapake, Hb Dartmount, Hb Seal Rock, etc. These non-deletional mutations form a termination codon at HBA2 gene, resulting in the imbalance of α-globin chain thus leads to the instability of red blood cells. A combination of non-deletional mutation such as α-thalassemia Hb CS and 2 copies deletion α^0 could result in Haemoglobin H-Constant Spring (--/-α^{CS}). HbH-CS phenotype is wide from mild anemia to very complicated, severe symptoms that depend on blood-transfusion. Since AlphaThalCNV did not design for point mutation detection so it could miss out diagnosis for compound α-thalassemia thus might underestimate the expected symptom for patients.

In conclusion, AlphaThalCNV has shown to improve the accuracy of α-thalassemia length detection using optimized parameters, α-cluster gene focused as well as integrating classification method. AlphaThalCNV is also a robust pipeline that applied on larger cohort and can be applied to future medical diagnosis.

doi:10.6342/NTU202500218

V) Reference:

Abel, H. J., & Duncavage, E. J. (2013). Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. *Cancer Genetics*, *206*(12), 432–440. https://doi.org/10.1016/J.CANCERGEN.2013.11.002

Aydinok, Y. (2012). Thalassemia. *Hematology*, *17*(SUPPL. 1). https://doi.org/10.1179/102453312X13336169155295

Babadi, M., Fu, J. M., Lee, S. K., Smirnov, A. N., Gauthier, L. D., Walker, M., Benjamin, D. I., Zhao, X., Karczewski, K. J., Wong, I., Collins, R. L., Sanchis-Juan, A., Brand, H., Banks, E., & Talkowski, M. E. (2023). GATK-gCNV enables discovery of rare copy number variants from exome sequencing data. *Nature Genetics*, *55*(9), 1589. https://doi.org/10.1038/S41588-023-01449-0

Barts, H. (n.d.). *Arizona Hemoglobin Bart's Fact Sheet for Health Care Providers*.

Baysal, E., & Huisman, T. H. J. (n.d.). Detection of Common Deletional ~Thalassemia-2 Determinants by PCR. *American Journal of Hematology*, *46*, 994. https://doi.org/10.1002/ajh.2830460309

Brieghel, C., Birgens, H., Frederiksen, H., Hertz, J. M., Steenhof, M., & Petersen, J. (2015). Novel 31.2 kb α0 Deletion in a Palestinian Family with α-Thalassemia. *Hemoglobin*, *39*(5), 346–349. https://doi.org/10.3109/03630269.2015.1054512

Cao, Y., Ha, S., So, C.-C., Tony Tong, M., Sze-man Tang, C., Zhang, H., Liang, R., Yang, J., Hon-Yin Chung, B., Chi-Fung Chan, G., Lung Lau, Y., Garcia-Barcelo, M.-M., Shiu-Kwan Ma, E., Sucharitchan, P., Hirankarn, N., & Yang, W. (n.d.). *NGS4THAL, a*

doi:10.6342/NTU202500218

One-Stop Molecular Diagnosis and Carrier Screening Tool for Thalassemia and Other Hemoglobinopathies by Next-Generation Sequencing. https://doi.org/10.1016/j.jmoldx.2022.06.006

Cardiero, G., Musollino, G., Prezioso, R., Nigro, V., & Lacerra, G. (2023). Alpha-Thalassemia in Southern Italy: Characterization of Five New Deletions Removing the Alpha-Globin Gene Cluster. *International Journal of Molecular Sciences*, *24*(3), 2577. https://doi.org/10.3390/IJMS24032577/S1

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/BIOINFORMATICS/BTY560

Chong, S. S., Boehm, C. D., Higgs, D. R., & Cutting, G. R. (2000). Single-tube multiplex-PCR screen for common deletional determinants of α-thalassemia. *Blood*, *95*(1), 360–362. https://doi.org/10.1182/BLOOD.V95.1.360

Curran, M., Mikhael, M., Sun, W. D., Lim, J., Leung, A., Morchi, G., & Chmait, R. H. (2020). Perinatal Management of Bart's Hemoglobinopathy: Paradoxical Effects of Intrauterine, Transplacental, and Partial Exchange Transfusions. *AJP Reports*, *10*(1), e11. https://doi.org/10.1055/S-0039-3401799

De Mare, A., Groeneger, A. H. O., Schuurman, S., Van Den Bergh, F. A. T. J. M., & Slomp, J. (2010a). A rapid single-tube multiplex polymerase chain reaction assay for the seven most prevalent alpha-thalassemia deletions and alphaalphaalpha(anti 3.7) alpha-globin gene triplication. *Hemoglobin*, *34*(2), 184–190. https://doi.org/10.3109/03630261003670259

doi:10.6342/NTU202500218

De Mare, A., Groeneger, A. H. O., Schuurman, S., Van Den Bergh, F. A. T. J. M., & Slomp, J. (2010b). A Rapid Single-Tube Multiplex Polymerase Chain Reaction Assay for the Seven Most Prevalent α-Thalassemia Deletions and αααanti 3.7 α-Globin Gene Triplication. *Hemoglobin*, *34*(2), 184–190. https://doi.org/10.3109/03630261003670259

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology 2017 35:4*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Doan, P. L., Nguyen, D. A., Le, Q. T., Hoang, D. T. T., Nguyen, H. Du, Nguyen, C. C., Doan, K. P. T., Tran, N. T., Ha, T. M. T., Trinh, T. H. N., Nguyen, V. T., Bui, C. T., Lai, N. D. T., Duong, T. H., Mai, H. L., Huynh, P. U. V., Huynh, T. T. T., Le, Q. V., Vo, T. B., … Phan, M. D. (2022). Detection of maternal carriers of common α-thalassemia deletions from cell-free DNA. *Scientific Reports*, *12*(1), 13581. https://doi.org/10.1038/S41598-022-17718-7

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology 2020 38:3*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Farashi, S., & Harteveld, C. L. (2018). Molecular basis of α-thalassemia. *Blood Cells, Molecules, and Diseases*, *70*, 43–53. https://doi.org/10.1016/J.BCMD.2017.09.004

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., Kinalis, S., Madsen, M. B., Kodama, M., Demircan, G. S., Simonyan,

doi:10.6342/NTU202500218

A., Yde, C. W., Olsen, L. R., Marvig, R. L., Østrup, O., Rossing, M., Nielsen, F. C., Winther, O., & Bagger, F. O. (2021a). A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*, *13*(24). https://doi.org/10.3390/CANCERS13246283/S1

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., Kinalis, S., Madsen, M. B., Kodama, M., Demircan, G. S., Simonyan, A., Yde, C. W., Olsen, L. R., Marvig, R. L., Østrup, O., Rossing, M., Nielsen, F. C., Winther, O., & Bagger, F. O. (2021b). A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*, *13*(24). https://doi.org/10.3390/CANCERS13246283/S1

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, *34*(20), 3572–3574. https://doi.org/10.1093/BIOINFORMATICS/BTY304

Giardine, B., van Baal, S., Kaimakis, P., Riemer, C., Miller, W., Samara, M., Kollia, P., Anagnou, N. P., Chui, D. H. K., Wajcman, H., Hardison, R. C., & Patrinos, G. P. (2007). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Human Mutation*, *28*(2), 206. https://doi.org/10.1002/HUMU.9479

Gibbons, R. J. (2012). α-thalassemia, mental retardation, and myelodysplastic syndrome. *Cold Spring Harbor Perspectives in Medicine*, *2*(10). https://doi.org/10.1101/CSHPERSPECT.A011759

doi:10.6342/NTU202500218

GitHub - jielab/pigeon: practical investigation of genomic errors by observation and notification. (n.d.). Retrieved December 26, 2024, from https://github.com/jielab/pigeon

Harteveld, C. L., & Higgs, D. R. (2010). α-thalassaemia. *Orphanet Journal of Rare Diseases*, *5*(1), 13. https://doi.org/10.1186/1750-1172-5-13

He, J., Song, W., Yang, J., Lu, S., Yuan, Y., Guo, J., Zhang, J., Ye, K., Yang, F., Long, F., Peng, Z., Yu, H., Cheng, L., & Zhu, B. (2017). Next-generation sequencing improves thalassemia carrier screening among premarital adults in a high prevalence population: the Dai nationality, China. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *19*(9), 1022–1031. https://doi.org/10.1038/GIM.2016.218

Higgs, D. R. (2013a). The Molecular Basis of α-Thalassemia. *Cold Spring Harbor Perspectives in Medicine*, *3*(1). https://doi.org/10.1101/CSHPERSPECT.A011718

Higgs, D. R. (2013b). The Molecular Basis of α-Thalassemia. *Cold Spring Harbor Perspectives in Medicine*, *3*(1). https://doi.org/10.1101/CSHPERSPECT.A011718

Higgs, D. R., Vickers, M. A., Wilkie, A. O. M., Pretorius, I. M., Jarman, A. P., & Weatherall, D. J. (1989a). A Review of the Molecular Genetics of the Human α-Globin Gene Cluster. *Blood*, *73*(5), 1081–1104. https://doi.org/10.1182/BLOOD.V73.5.1081.1081

Higgs, D. R., Vickers, M. A., Wilkie, A. O. M., Pretorius, I. M., Jarman, A. P., & Weatherall, D. J. (1989b). A Review of the Molecular Genetics of the Human α-Globin

doi:10.6342/NTU202500218

Gene Cluster. *Blood*, *73*(5), 1081–1104. https://doi.org/10.1182/BLOOD.V73.5.1081.1081

Higgs, D. R., & Wood, W. G. (2008). Long-range regulation of α globin gene expression during erythropoiesis. *Current Opinion in Hematology*, *15*(3), 176–183. https://doi.org/10.1097/MOH.0B013E3282F734C4

(How to) Call rare germline copy number variants – GATK. (n.d.). Retrieved November 19, 2024, from https://gatk.broadinstitute.org/hc/en-us/articles/360035531152--How-to-Call-rare-germline-copy-number-variants

Hsu, J. S., Wu, D. C., Shih, S. H., Liu, J. F., Tsai, Y. C., Lee, T. L., Chen, W. A., Tseng, Y. H., Lo, Y. C., Lin, H. Y., Chen, Y. C., Chen, J. Y., Chou, T. H., Chang, D. T. H., Su, M. W., Guo, W. H., Mao, H. H., Chen, C. Y., & Chen, P. L. (2023). Complete genomic profiles of 1496 Taiwanese reveal curated medical insights. *Journal of Advanced Research*. https://doi.org/10.1016/J.JARE.2023.12.018

Janevski, A., Varadan, V., Kamalakaran, S., Banerjee, N., & Dimitrova, N. (2012). Effective normalization for copy number variation detection from whole genome sequencing. *BMC Genomics*, *13 Suppl 6*(6), 1–11. https://doi.org/10.1186/1471-2164-13-S6-S16/FIGURES/5

Kalle Kwaifa, I., Lai, M. I., & Md Noor, S. (2020). Non-deletional alpha thalassaemia: A review. *Orphanet Journal of Rare Diseases*, *15*(1), 1–12. https://doi.org/10.1186/S13023-020-01429-1/FIGURES/3

doi:10.6342/NTU202500218

Kamanzi, N. G. (2024). Hemoglobinopathy for Malaria Protection: A Comprehensive Review. *IDOSR JOURNAL OF BIOCHEMISTRY, BIOTECHNOLOGY AND ALLIED FIELDS*, *9*(3), 30–34. https://doi.org/10.59298/IDOSR/JBBAF/24/93.3034000

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, J., Xie, X. M., Liao, C., & Li, D. Z. (2014). Co-inheritance of α-thalassaemia and β-thalassaemia in a prenatal screening population in mainland China. *Http://Dx.Doi.Org/10.1177/0969141314548203*, *21*(4), 167–171. https://doi.org/10.1177/0969141314548203

Li, J., Ye, G., Zeng, D., Tian, B., Wang, W., Feng, Q., & Zhu, C. (2022). Accurate genotype diagnosis of Hong Kongαα thalassemia based on third-generation sequencing. *Annals of Translational Medicine*, *10*(20), 1113–1113. https://doi.org/10.21037/ATM-22-4309

Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics*, *45*(1), 1–6. https://doi.org/10.1152/PHYSIOLGENOMICS.00082.2012/SUPPL_FILE/SUPPDATA.PDF

Liebhaber, S. A., Cash, F. E., & Ballas, S. K. (1986). Human alpha-globin gene expression. The dominant role of the alpha 2-locus in mRNA and protein synthesis. *Journal of Biological Chemistry*, *261*(32), 15327–15333. https://doi.org/10.1016/S0021-9258(18)66871-1

doi:10.6342/NTU202500218

Liu, Y. T., Old, J. M., Miles, K., Fisher, C. A., Weatherall, D. J., & Clegg, J. B. (2000). Rapid detection of α-thalassaemia deletions and α-globin gene triplication by multiplex polymerase chain reactions. *British Journal of Haematology*, *108*(2), 295–299. https://doi.org/10.1046/j.1365-2141.2000.01870.x

Malaria and Thalassemia in the Mediterranean Basin. (2021).

Maria Domenica Cappellini, Alan Cohen, John Porter, Ali Taher, V. V. (2014). Guidelines for the Management of Transfusion Dependent Thalassaemia (TDT) 3rd Edition. *Thalassaemia International Federation*.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/GR.107524.110

Molchanova, T. P., Pobedimskaya, D. D., & Postnikov, Y. V. (1994). A simplified procedure for sequencing amplified DNA containing the α2- or α1-globin gene. *Hemoglobin*, *18*(3), 251–255. https://doi.org/10.3109/03630269409043628/ASSET//CMS/ASSET/6949543B-FFF7-439F-A5FD-C395B32B748B/03630269409043628.FP.PNG

Munkongdee, T., Chen, P., Winichagoon, P., Fucharoen, S., & Paiboonsukwong, K. (2020). Update in Laboratory Diagnosis of Thalassemia. *Frontiers in Molecular Biosciences*, *7*, 74. https://doi.org/10.3389/FMOLB.2020.00074

doi:10.6342/NTU202500218

Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Frontiers in Plant Science*, *12*, 657240. https://doi.org/10.3389/FPLS.2021.657240/BIBTEX

Nijkamp, J. F., Van Den Broek, M. A., Geertman, J. M. A., Reinders, M. J. T., Daran, J. M. G., & De Ridder, D. (2012). De novo detection of copy number variation by co-assembly. *Bioinformatics*, *28*(24), 3195–3202. https://doi.org/10.1093/BIOINFORMATICS/BTS601

Nucleotide BLAST: Search nucleotide databases using a nucleotide query. (n.d.). Retrieved December 20, 2024, from https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=GeoBlast&PAGE_TYPE=BlastSearch

Old J, Harteveld CL, Traeger-Synodinos J, Petrou M, Angastiniotis M, & Galanello R. (2012). *Prevention of thalassaemias and other haemoglobin disorders. Laboratory protocols. Thalassaemia International Federation 2012*. *2*.

Peng, C. T., Liu, S. C., Peng, Y. C., Lin, T. H., Wang, S. J., Le, C. Y., Shih, M. C., Tien, N., Lu, J. J., & Lin, C. Y. (2013). Distribution of thalassemias and associated hemoglobinopathies identified by prenatal diagnosis in Taiwan. *Blood Cells, Molecules, and Diseases*, *51*(3), 138–141. https://doi.org/10.1016/J.BCMD.2013.04.007

Piel, F. B., & Weatherall, D. J. (2014a). The α-Thalassemias. *New England Journal of Medicine*, *371*(20), 1908–1916. https://doi.org/10.1056/NEJMRA1404415

doi:10.6342/NTU202500218

Piel, F. B., & Weatherall, D. J. (2014b). The α-thalassemias. *The New England Journal of Medicine*, *371*(20), 1908–1916. https://doi.org/10.1056/NEJMRA1404415

Piel, F. B., & Weatherall, D. J. (2014c). The α-Thalassemias. *New England Journal of Medicine*, *371*(20), 1908–1916. https://doi.org/10.1056/NEJMRA1404415/SUPPL_FILE/NEJMRA1404415_DISCLOSURES.PDF

POOTRAKUL, S., WASI, P., & NA-NAKORN, S. (1967). Haemoglobin Bart's hydrops foetalis in Thailand. *Annals of Human Genetics*, *30*(4), 293–308. https://doi.org/10.1111/J.1469-1809.1967.TB00031.X

Shang, X., & Xu, X. (2017). Update in the genetics of thalassemia: What clinicians need to know. *Best Practice and Research: Clinical Obstetrics and Gynaecology*, *39*, 3–15. https://doi.org/10.1016/j.bpobgyn.2016.10.012

Tan, A. S. C., Quah, T. C., Low, P. S., & Chong, S. S. (2001). A rapid and reliable 7-deletion multiplex polymerase chain reaction assay for α-thalassemia. *Blood*, *98*(1), 250–251. https://doi.org/10.1182/BLOOD.V98.1.250

Taylor, S. M., Parobek, C. M., & Fairhurst, R. M. (2012). Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *12*(6), 457–468. https://doi.org/10.1016/S1473-3099(12)70055-5

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, *13*(1), 36. https://doi.org/10.1038/NRG3117

doi:10.6342/NTU202500218

Vichinsky, E. P. (2009). Alpha thalassemia major--new mutations, intrauterine management, and outcomes. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, 35–41. https://doi.org/10.1182/ASHEDUCATION-2009.1.35

Vichinsky, E. P. (2013). Clinical Manifestations of α-Thalassemia. *Cold Spring Harbor Perspectives in Medicine*, *3*(5). https://doi.org/10.1101/CSHPERSPECT.A011742

Vijian, D., Wan Ab Rahman, W. S., Ponnuraj, K. T., Zulkafli, Z., Bahar, R., Yasin, N., Hassan, S., & Esa, E. (2023). Gene Mutation Spectrum among Alpha-Thalassaemia Patients in Northeast Peninsular Malaysia. *Diagnostics*, *13*(5), 894. https://doi.org/10.3390/DIAGNOSTICS13050894/S1

Vijian, D., Wan Ab Rahman, W. S., Ponnuraj, K. T., Zulkafli, Z., & Mohd Noor, N. H. (2021). Molecular Detection of Alpha Thalassemia: A Review of Prevalent Techniques. *Medeniyet Medical Journal*, *36*(3), 257. https://doi.org/10.5222/MMJ.2021.14603

Villegas, A., Sanchez, J., Gonzalez, F. A., Carreño, D. L., & Ropero, P. (1994). α-thalassemia-1 (— −-CAL mutation) in a Spanish family. *American Journal of Hematology*, *46*(4), 367–368. https://doi.org/10.1002/AJH.2830460421

Wang, H. C., Hsieh, L. L., Liu, Y. C., Hsiao, H. H., Lin, S. K., Tsai, W. C., & Liu, T. C. (2017). The epidemiologic transition of thalassemia and associated hemoglobinopathies in southern Taiwan. *Annals of Hematology*, *96*(2), 183–188. https://doi.org/10.1007/S00277-016-2868-7/FIGURES/2

doi:10.6342/NTU202500218

Xi, R., Lee, S., & Park, P. J. (2012). A Survey of Copy-Number Variation Detection Tools Based on High-Throughput Sequencing Data. *Current Protocols in Human Genetics*, *75*(1), 7.19.1-7.19.15. https://doi.org/10.1002/0471142905.HG0719S75

Yoon, S., Xuan, Z., Makarov, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, *19*(9), 1586–1592. https://doi.org/10.1101/GR.092981.109

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. *BMC Genomics*, *12*(1), 1–12. https://doi.org/10.1186/1471-2164-12-375/FIGURES/7

α +-Thalassemia and Protection from Malaria. (2006). *PLOS Medicine*, *3*(5), e221. https://doi.org/10.1371/JOURNAL.PMED.0030221

移民署中文網. (n.d.). Retrieved December 10, 2024, from https://www.immigration.gov.tw/

doi:10.6342/NTU202500218