

## 단일 염기 다형성 분석을 통한 미생물 출처의 법의학적 조사

임승현<sup>#,1)</sup> · 윤형석<sup>#,1)</sup> · 이승호<sup>1)</sup> · 정주환<sup>1)</sup> · 구세훈<sup>1)</sup> · 이대상<sup>1)</sup> · 송동현<sup>\*,1)</sup>

<sup>1)</sup> 국방과학연구소 국방첨단기술연구원 Chem-Bio센터

## Microbial Forensic Investigations of Microbial Sources through Single Nucleotide Polymorphism Analysis

Seunghyun Lim<sup>#,1)</sup> · Hyeongseok Yun<sup>#,1)</sup> · Seungho Lee<sup>1)</sup> · Juhwan Jung<sup>1)</sup> ·  
Sehun Gu<sup>1)</sup> · Daesang Lee<sup>1)</sup> · Donghyun Song<sup>\*,1)</sup>

<sup>1)</sup> Chem-Bio Technology Center, Agency for Defense Development, Korea

(Received 17 May 2024 / Revised 27 August 2024 / Accepted 10 September 2024)

### Abstract

*Bacillus anthracis*, a potential biological agent for terrorism, has been actively investigated for its underlying property and phylogenetic origin in the field of Microbial forensics. With the advancement of next generation sequencing(NGS) technology, *in silico* analysis becomes feasible at the whole genome sequence level, reducing the time and cost. In this paper, we suggested a methodology for identifying unknown samples from the field, which simulate real forensic evidence rather than highly purified samples, utilizing two *in silico* methods: k-mer analysis and whole genome single-nucleotide polymorphism(wgSNP). We performed prefix-based k-mer analysis using 964 NGS raw data obtained from the NCBI database, along with the NGS data from the unknown samples, by mapping the reads to Ames Ancestor and obtaining the consensus sequence. When analyzed together with 844 assembled sequences obtained from the NCBI database, it was determined that the unknown samples belong to the Injectional anthrax group, which was an infectious group identified among heroin users in Norway in 2000. wgSNP analysis has categorized the sample into discrete low-SNP group-I and high-SNP group-II, with a difference of up to 9 SNPs within each group. We observed 30 SNP positions in group-I, which includes the unknown samples, and confirmed that the SNP of A4564 was identical to that of the unknown samples. These results demonstrate that prefix-based k-mer, and wgSNP analysis can be effectively used for the collection of microbial forensic evidence from field samples.

Key Words : *Bacillus Anthracis*(탄저균), Microbial Forensics(미생물법의학), Single Nucleotide Polymorphism(단일염기 다형성)

<sup>#</sup> Contributed equally to this work

<sup>\*</sup> Corresponding author, E-mail: swpia@add.re.kr

Copyright © The Korea Institute of Military Science and Technology

## 1. 서론

미생물법의학(Microbial forensics)은 생물 범죄 및 누출 예방 차원에서 생물학작용제의 기본 특성을 파악하고 출처를 추정하는 학문이다. 2001년 탄저균(*Bacillus anthracis*)을 이용한 테러 이후 미생물법의학 연구의 필요성이 크게 증가하였으며 특히 해당 종에 대한 계통 및 발생 추적이 활발히 이루어져 왔다<sup>[1]</sup>. USAMRIID (United States Army Medical Research Institute of Infectious Diseases)는 테러에 사용된 생물학작용제가 Ames 아류형(Subtype)으로 포자 형성 유전자의 부위의 인위적 변이가 있음을 밝혀냈으며, 2016년에는 1979년 Sverdlovsk 유출 사건의 샘플을 재분석하여 유전자 조작의 증거가 없음이 밝혀졌다<sup>[2]</sup>.

미생물법의학은 과거 DNA 지문 분석(DNA fingerprinting)이나 다중 유전자좌 가변수일렬반복 분석(Multiple Locus Variable number of tandem repeat Analysis, MLVA) 등 짧은 서열을 중심으로 이루어졌다<sup>[3]</sup>. 이는 Sanger sequencing을 통해 얻을 수 있는 서열의 길이가 최대 1 Kb이며, megabase(Mb)당 \$500의 비용이 소요됐기 때문이다<sup>[4]</sup>. 차세대염기서열분석법(Next Generation Sequencing, NGS)은 대량으로 전체 유전체(Whole-Genome Sequence, WGS) 정보를 얻음으로써 소요 비용과 시간을 혁신적으로 단축했으며, 전통적인 분석을 *in silico* 수준에서 가능케 했다. Illumina사의 short read 기반 Genome Analyzer의 경우 최대 17 gigabases(Gb)의 서열 확보가 가능하며 Mb당 소요 비용은 \$6 수준으로 감소하였다<sup>[4]</sup>.

k-mer 분석법은 *in silico* 분석 중 하나로, 전체 서열을 길이가 k인 단편 서열로 나눈 후 자카드 계수(Jaccard similarity index)를 계산하여 전체 유전체 간 Distance를 판단하는 방법이다<sup>[5]</sup>. 일반적인 정렬 기법의 경우 서열 길이에 따라 계산 시간이 기하급수적으로 늘어나지만, k-mer 분석은 연구자의 k 값과 prefix(단편 서열의 첫 부분) 선택에 따라 소요 시간 단축이 가능하다. 박테리아 유전자에서 서열 ATGAC는 코딩영역에 분포해 있으며, 기존의 문헌에서 다른 서열에 비해 높은 Average Nucleotide Identity(ANI)와의 상관계수를 가져 유의미한 그룹 분리가 가능함을 입증하였다<sup>[6]</sup>. k-mer 분석법은 이와 같은 prefix를 가진 단편 서열로 분석대상을 제한함으로써 소요 시간을 줄일 수 있으며, 넓은 분해능을 가져 역 단계(superkingdom domain; 고균역, 세균 역, 진핵생물 역)

부터 *Bacillus* 속의 계통 추정까지 가능하다고 알려져 있다<sup>[5,7]</sup>.

k-mer 기반 분석은 연산량을 조정할 수 있고 넓은 범위의 분류군을 구분하기 좋지만, 삽입, 삭제 및 재배열과 같은 변이가 유사도로만 판단되는 단점이 있다. 단일염기다형성(Single Nucleotide Polymorphism, SNP)은 유전적 다양성을 나타내는 또 다른 척도 중 하나로 대상 집단 내 통상 1% 이상의 빈도로 발생하는 염기서열 변이이다<sup>[8]</sup>. 참조서열에 샘플 서열을 정렬한 후 Nucleotide Difference(ND) 값을 계산하여 신뢰구간 안에 있는 SNP를 판별한 후 Neighbor-joining 알고리즘을 통해 트리를 완성한다<sup>[9]</sup>. 전체 유전체 SNP 분석(whole genome SNP analysis; wgSNP)은 정렬 및 매핑 과정에 많은 소요 시간을 요구하지만, 진화적 패턴을 감지하고 특정 종에 대하여 계통군(clades)이나 하위분기군(subclade) 수준의 정밀한 분류가 가능하다<sup>[9,10]</sup>.

본 논문에서는 젤상 고형 배지에 혼합되어있는 *B. anthracis*의 핵산을 분리하여 NGS 방법으로 전장유전체서열을 획득했다. NCBI(National Center for Biotechnology Information) 데이터베이스로부터 획득한 844개의 *B. anthracis* 서열 및 964개의 NGS 미가공 데이터(Sequence Read Archive, SRA)를 이용한 k-mer 분석 결과, 미지의 핵산은 유럽에서 2000년부터 2013년 사이 헤로인 투여 후 집단 감염된 Injectional Anthrax (IA)에 속하는 것으로 확인되었다. k-mer 분석 결과를 바탕으로 NCBI 데이터베이스로부터 획득한 IA 그룹의 SRA 데이터 58개에 대한 wgSNP 분석을 진행한 결과, 미지의 핵산은 *B. anthracis*의 Ames Ancestor [txid: 261594]의 염색체에서 236078, 507075, 3770887이 각각 G, A, T로 치환된 특징을 갖는 A4564임을 특정했다.

## 2. 실험방법

분석 시료는 RKI(Robert Koch Institute)로부터 전달 받았으며 0.85% NaCl(#23-1-2)와 고형배지(#23-3-1)에 핵산이 혼합된 형태였다. Qiagen사의 DNeasy Blood & Tissue Kits를 이용해 핵산추출을 하였고, 농도가 낮은 샘플은 효과적인 유전체 염기서열 확보를 위하여 TruPrime Whole Genome Amplification Kit<sup>®</sup> (4basebio<sup>™</sup>)을 이용하여 핵산 증폭을 수행하였다.

증폭된 핵산은 Covaris sonicator를 사용하여 Illumina sequencing에 맞는 크기(300 bp)로 분절화하였다. 그 후에는 Apollo324 라이브러리 자동화 시스템(TaKaRa)과 Illumina 호환 PrepX DNA Library Kit(TaKaRa)를 사용하여 shotgun 라이브러리를 생성했다. 라이브러리 생성에는 Illumina sequencing에 적용되는 dual index adaptor를 사용하였고, DNA 서열 데이터는 Illumina 사의 양방향 서열 분석을 사용하여 생성했다. Illumina 사의 TruSeq Library Prep Kit를 이용해 시료 전처리 과정을 수행하였다. 모든 cleanup은 Beckman Coulter의 AMPure Xp bead를 이용했으며, 생성된 라이브러리에 대한 농도 검증은 일차적으로 Agilent Technologies의 TapeStation 4200과 D1000 ScreenTape, 이차적으로 KAPA Biosystems의 KAPA Library Quantification Kit와 Thermo Fisher Scientific의 QuantStudio 6 Flex Real-time PCR system을 이용하였다. Illumina 사의 10 % PhiX control v3와 함께 NextSeq 500 benchtop platform을 통해 염기서열분석을 진행했다.

NCBI로부터 획득한 *B. anthracis*[txid: 1392] SRA 데이터 964개는 CLC genomics workbench 24.0을 이용하여 adaptor trimming 및 quality trimming(>0.05)을 진행했으며 Ames Ancestor를 참조서열로 삼아 공통서열(Consensus sequence)을 확보했다<sup>[12]</sup>. 마찬가지로의 방법으로 미지 시료 Unknown Biological Sample(UBS) #23-1-2 (#23-1)과 #23-3-1(#23-3)로부터 분리된 핵산의 NGS 결과에서 공통서열을 확보했다. 이와 함께 조립(assembly)이 완료된 GenBank 데이터 844개를 사용하여 k-mer 분석을 수행했으며 k는 21 mer로, 참고 문헌에 따라 prefix는 ATGAC를 사용했다<sup>[6]</sup>.

IA 그룹의 NGS 미가공 데이터 58개에 대한 wgSNP 분석은 Ames Ancestor를 참조서열로 삼아 기존의 문헌을 참고하여 수행되었으며, SNP calling은 SAMTools v0.1.18을 사용하여 정렬 후 위치당 coverage 10 %, prune distance 10을 기준값으로 잡아 필터링했다<sup>[9,13]</sup>. 매핑된 reads에 대해 z-score>=1.96 이상만 SNP로 인정했으며, UPGMA(Unweighted Pair Group Method with Arithmetic mean)을 사용해 bootstrap을 1000 회 시행했다. 각 시료의 wgSNP 분석에 사용된 유전체 내 염기서열의 위치(position)는 Ames Ancestor의 유전체를 기준으로 분석했다. 염기서열 분석 결과로부터 k-mer 분석 및 SNP 분석, 그리고 strain 특정의 전체 과정은 Fig. 1의 순서로 수행되었다.

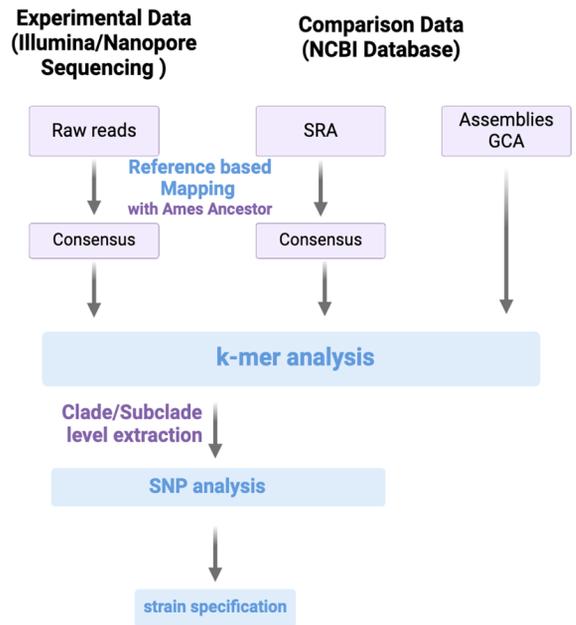


Fig. 1. Workflow of whole analysis

### 3. 결과

미지시료에 포함된 *B. anthracis* 핵산의 출처를 밝히기 위하여 시료로부터 분리한 핵산의 NGS 결과 두 건(#23-1, #23-3)과 NCBI 데이터베이스에서 확보한 964개의 SRA 데이터로부터 Ames Ancestor를 참조서열로 하여 공통서열을 확보했다(Fig. 1). 이와 함께 NCBI GenBank로부터 획득한 조립된 서열 844개를 포함한 총 1,809개 데이터에 대해 k-mer 분석을 시행했다. 전체 데이터에 대하여 k = 21로 k-mer 분석을 진행했을 때 연산 자원 상의 문제로 Out of Memory (OOM) 에러가 발생했으며, prefix 값을 ATGAC로 지정해 이를 해결하였다.

k-mer 분석 결과 그룹 군은 7개의 그룹으로 분류되었으며 이는 기준에 알려진 아류형과 유사성을 보였다(Fig. 2). *B. Anthracis* 분류는 크게 clade A(~90 %), clade B(~10 %), clade C(~1 %)로 이루어져 있으며<sup>[2]</sup>, clade B에 속하는 B branch가 하나의 그룹으로 묶이는 것을 확인했다(Fig. 2, green). 이외의 6개 그룹은 모두 clade A에 속한 subclade로 이루어져 있었으며 여기에는 Ancient A, TransEurAsia(TEA), V770, Ames, Australia94(Aust94), Sterne 등이 속해있다<sup>[2]</sup>. 이들 중

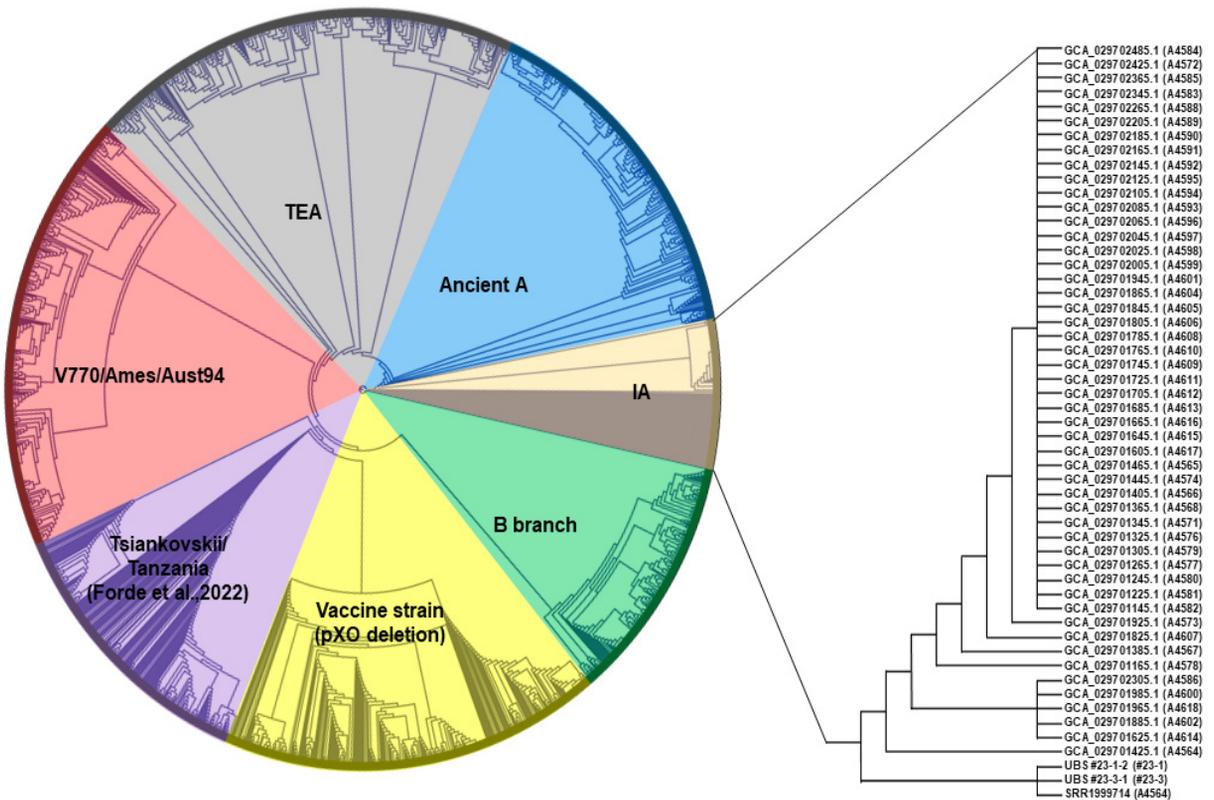


Fig. 2. k-mer tree based on Jaccard-distance of *B. anthracis* and a subtree of IA subclade

TEA subclade는 가장 넓은 다형성(polytomy)을 가지며 Tsiankovskii, Injectional Anthrax(IA) 등 여러 그룹을 포함한다고 알려졌지만 k-mer 분석에서는 분산된 형태를 띠었다(Fig. 2, purple and ivory). 또한 sterne 그룹은 pXO가 결손된 백신 균주로 pXO를 가지고 있지 않은 모든 *B. anthracis*는 k-mer 분석 결과 sterne 그룹에 속하는 것을 관찰했다(Fig. 2, yellow). 미지 시료 #23-1과 #23-3은 이 중 IA 그룹에 속해있었다. IA 그룹은 2000년도 노르웨이에서 헤로인 사용자들로부터 처음 발병이 보고되었으며 공통적으로 사용한 주사기를 통해 감염 및 전파가 일어났다<sup>[14][6]</sup>. 가장 근접한 strain은 NCBI 데이터베이스로부터 획득하여 본 연구에서 공통서열을 추출한 SRR1999714(A4564)였으며, GenBank로부터 획득한 IA 그룹의 조립된 서열과도 유사한 결과를 관찰했다(Fig. 2).

GCA로 시작하는 GenBank 조립서열은 기존 논문에서 SKESA v2.2를 통해 조립된 것이다. 실험을 통해 추출한 raw data를 가공하여 하나의 유전체 혹은 다

량의 contig를 만드는 조립 과정은 다양한 알고리즘을 사용하지만, 조립 시 선택한 변수 조건(gap cost, reads 간 유사도 등)에 따라 tandem repeat이 반복되는 위치에서 조립오류(assembly error)를 동반하기 때문에 결과가 달라질 수 있다<sup>[18,19]</sup>. 그럼에도 SRR1999714로부터 유래된 GCA\_29701425.1과 미지시료 #23-1/#23-3이 가장 가장 유사한 서열임을 k-mer 분석상에서 관찰할 수 있었다. 이는 첫째로 SKESA 알고리즘이 k = 21을 기본값으로 하는 raw reads의 k-mer 유사도 값을 이용하기 때문이며 prefix를 기반으로 한 k-mer 분석이 coding region에 주목함으로써 조립오류에 영향을 받지 않았으며, SKESA 알고리즘을 사용한 조립의 퀄리티가 충분히 좋았다고 해석할 수 있다. 해당 알고리즘은 SRA 데이터 및 약한 오염 시료에 대해 SPAdes 및 Megahit 같은 타 조립 알고리즘 보다 적은 조립오류를 보였다<sup>[17]</sup>.

k-mer 분석법은 전체 서열을 길이가 k인 단편 서열로 나눈 후 자카드 계수(Jaccard similarity index)를 계

단일 염기 다형성 분석을 통한 미생물 출처의 법의학적 조사

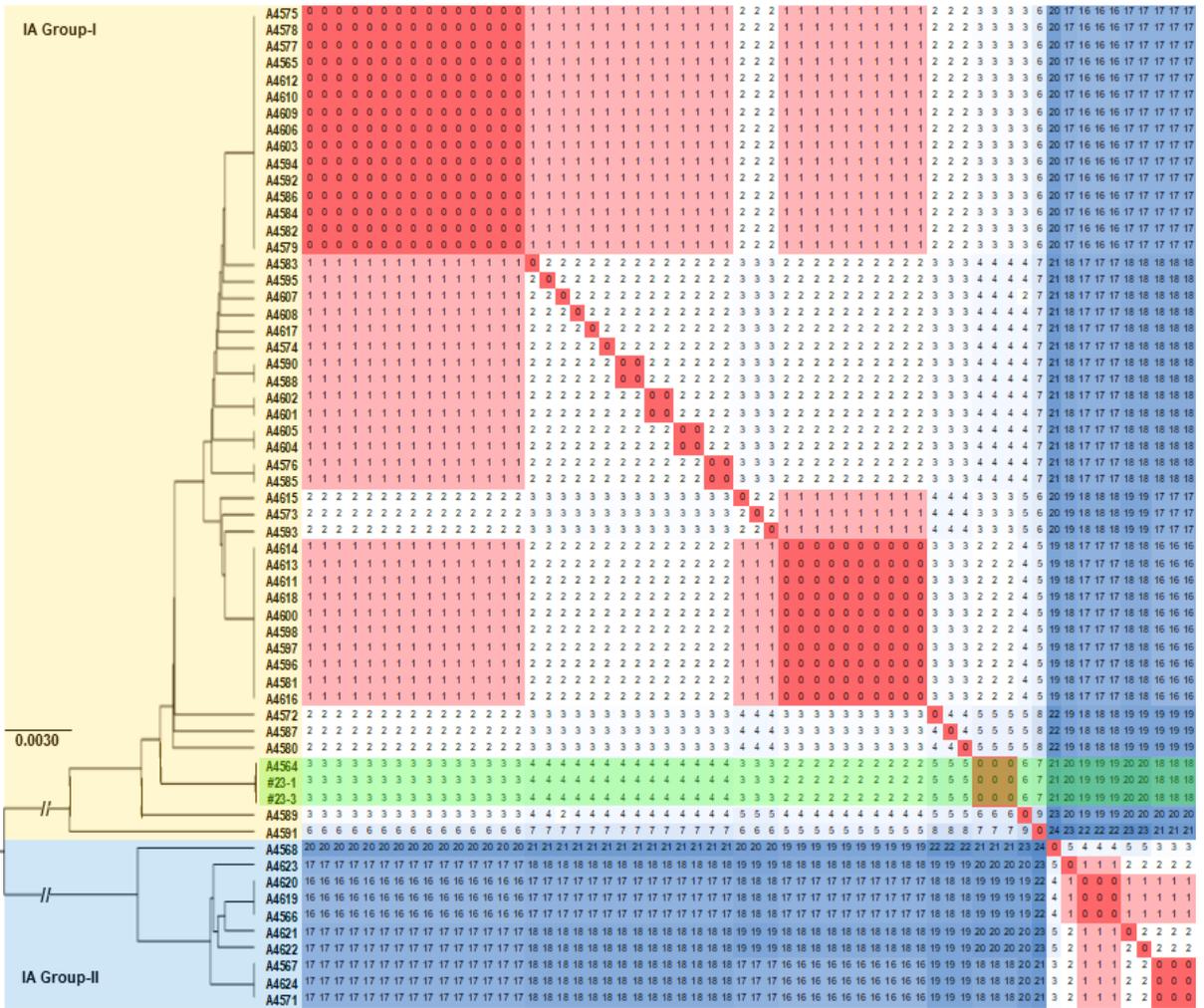


Fig. 3. Phylogenetic tree based on wgSNP alignment with 1000 bootstrap replications and SNP heatmap matrix of IA subclade

산하는 방식이고, 본 연구에서는 연산량을 줄이기 위해 prefix를 ATGAC로 설정했다<sup>5,6)</sup>. 따라서 k-mer 분석 결과 미지시료 #23-1/#23-3이 strain A4564와 가장 비슷함을 관찰했다라도 유전체에 대한 wgSNP 분석을 통해 염기서열을 다형성을 관찰해야만 미지시료의 출처 특정이 가능하다. 따라서 우리는 IA 그룹의 전장유전체서열을 획득한 프로젝트 PRJNA282011의 SRA 데이터 58개와 미지시료 #23-1/#23-3 두 시료에 대해 Ames Ancestor를 참조서열로 하여 wgSNP 분석을 수행했다.

SNP 분석 결과, IA 그룹 내에서 참조서열과 비교

했을 때 6개 이하의 적은 SNP를 지닌 그룹과 16개 이상의 많은 SNP를 지닌 두 그룹으로 나누어지는 것을 확인했으며 각 서브 그룹(group-I, group-II) 내에서 SNP는 9개 이내였다(Fig. 3). 이는 IA 그룹이 2009~2010년 스코틀랜드 발병 사건(group-I)과 2000년도 노르웨이로부터 유래한 2012-2013 유럽 전역의 발병 사건(group-II)으로 나누어진다는 기존의 연구에 부합했다<sup>20)</sup>.

미지시료 #23-1과 #23-3이 속한 IA group-I의 48개 표본에 대한 심층분석 결과 확인된 SNP는 총 30개였으며 참조서열인 Ames Ancestor의 유전체를 기준으로

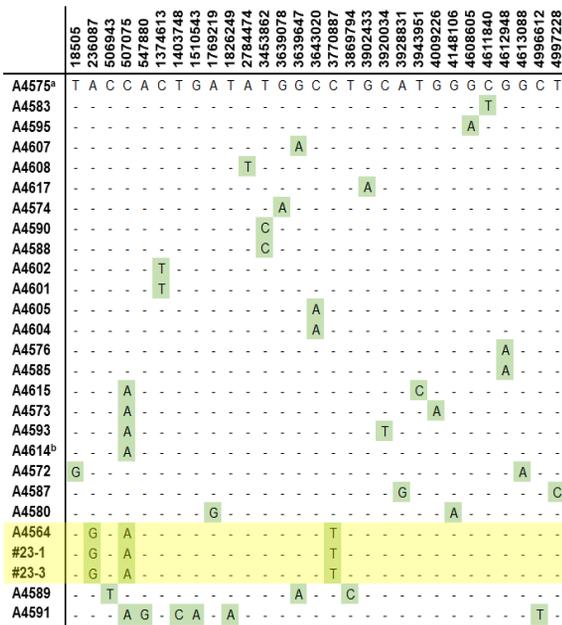


Fig. 4. SNP position and profile. Highlighted position is novel in this study. <sup>a</sup>Contains 14 sequences with the same whole-genome SNP results as A4575. <sup>b</sup>Contains 9 sequences with the same whole-genome SNP results as A4614

염기서열의 위치를 나타내었다(Fig. 4). A4575와 같은 SNP를 가진 14개의 strain은 A4575로 대표되었으며 마찬가지로 A4614와 같은 9개의 strain 또한 A4614로 표시했다. 미지 시료 #23-1/#23-3과 A4564는 236087의 아데닌(A)이 구아닌(G)으로, 507075의 시토신(C)이 아데닌(A)으로, 3770887의 시토신(C)이 트레오닌(T)으로 변이된 특징을 가지는 변이주로서 그 특성이 완벽하게 일치함에 따라 A4564로부터 유래한 시료임을 특정했다.

4. 결론

본 연구에서는 k-mer 분석 및 wgSNP 분석을 통하여 특정 매트릭스가 포함된 미량의 미지시료에 대한 법의학적인 분석을 수행했다. 미지시료 #23-1/#23-3의 전장유전체서열 분석 결과로부터 공통서열을 추출하고 NCBI 데이터베이스로부터 확보한 1,808개의 *B. anthracis* 유전체 서열을 k-mer 분석을 통해 하위분기

군 수준의 분류를 수행한 결과, 미지시료는 헤로인 사용자들에서 발병한 IA 그룹으로 분류됨을 확인했다. 이를 토대로 IA 그룹과 미지시료에 대한 wgSNP 분석 결과 미지 시료 #23-1과 #23-3은 A236087G, C507075A, C3770887T 변이를 특징으로 하는 *B. anthracis* strain A4564로부터 유래되었음을 밝혀냈다.

세균 출처의 법의학적인 조사를 위해 NCBI 데이터베이스로부터 획득한 1,808개의 서열에 대한 wgSNP 분석을 수행하는 것은 연산 자원상의 문제가 생길 가능성이 높을 뿐만 아니라, 제출된 공통서열의 조립 방식이나 분석에 사용한 NGS 결과의 품질에 따라 부정확한 SNP가 관찰될 가능성이 있다. 이러한 문제를 해결하기 위해 본 연구에서는 prefix를 설정한 k-mer 분석을 통해 미지시료의 출처를 IA 그룹의 58종으로 선별했으며, 추가적으로 wgSNP 분석을 수행하여 미생물 출처를 특정했다. 또한 고순도의 실험실 시료가 아닌 실제 법의학적인 증거를 모사한 매트릭스가 포함된 미량의 시료로부터 핵산추출 및 NGS를 활용한 wgSNP 분석이 미생물법의학적인 증거 획득에 유의미함을 증명했다.

References

- [1] Keim P, Pearson T, Okinaka R. Microbial forensics: DNA fingerprinting of *Bacillus anthracis*(anthrax). *Anal Chem*. 2008 Jul 1;80(13):4791-9.
- [2] Sahl JW, Pearson T, Okinaka R, Schupp JM, Gillette JD, Heaton H, Birdsell D, Hepp C, Fofanov V, Noseda R, Fasanella A, Hoffmaster A, Wagner DM, Keim P. A *Bacillus anthracis* Genome Sequence from the Sverdlovsk 1979 Autopsy Specimens. *mBio*. 2016 Sep 27;7(5):e01501-16.
- [3] Perez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. *Infect Genet Evol*. 2018 Sep;63:346-359.
- [4] Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet*. 2009 Aug;85(2):142-54.
- [5] Bussi Y, Kapon R, Reich Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS One*. 2021 Oct 14;16(10):e0258693.

- [6] Lumpe J, Gumbleton L, Gorzalski A, Libuit K, Varghese V, Lloyd T, Tadros F, Arsimendi T, Wagner E, Stephens C, Sevinsky J, Hess D, Pandori M. GAMBIT(Genomic Approximation Method for Bacterial Identification and Tracking): A methodology to rapidly leverage whole genome sequencing of bacterial isolates for clinical identification. *PLoS One*. 2023 Feb 16;18(2):e0277575.
- [7] Wang, A., Ash, G. Whole Genome Phylogeny of *Bacillus* by Feature Frequency Profiles(FFP). *Sci Rep* 2015 Sep 01;5:13644.
- [8] Wakui M. Analysis of single nucleotide polymorphisms(SNPs). *Rinsho Byori*. 2013 Nov;61(11):1008-17.
- [9] Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One*. 2014 Aug 11;9(8):e104984.
- [10] Leekitcharoenphon P, Kaas RS, Thomsen MC, Friis C, Rasmussen S, Aarestrup FM. *snpTree*--a web-server to identify and construct SNP trees from whole genome sequence data. *BMC Genomics*. 2012;13 Suppl 7(Suppl 7):S6.
- [11] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2013 Jan;41(Database issue):D36-42.
- [12] Ravel J, Jiang L, Stanley ST, Wilson MR, Decker RS, Read TD, Worsham P, Keim PS, Salzberg SL, Fraser-Liggett CM, Rasko DA. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". *J Bacteriol*. 2009 Jan;191(1):445-6.
- [13] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9.
- [14] Grunow R, Verbeek L, Jacob D, Holzmann T, Birkenfeld G, Wiens D, von Eichel-Streiber L, Grass G, Reischl U. Injection anthrax--a new outbreak in heroin users. *Dtsch Arztebl Int*. 2012 Dec;109(49):843-8.
- [15] Ringertz SH, Høiby EA, Jensenius M, Maehlen J, Caugant DA, Myklebust A, Fossum K. Injectional anthrax in a heroin skin-popper. *Lancet*. 2000 Nov 4;356(9241):1574-5.
- [16] Price EP, Seymour ML, Sarovich DS, Latham J, Wolken SR, Mason J, Vincent G, Drees KP, Beckstrom-Sternberg SM, Phillippy AM, Koren S, Okinaka RT, Chung WK, Schupp JM, Wagner DM, Vipond R, Foster JT, Bergman NH, Burans J, Pearson T, Brooks T, Keim P. Molecular epidemiologic investigation of an anthrax outbreak among heroin users, Europe. *Emerg Infect Dis*. 2012 Aug;18(8):1307-13.
- [17] Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol*. 2018 Oct 4;19(1):153.
- [18] Liao, Xingyu, et al., "Current challenges and solutions of de novo assembly," *Quantitative Biology* 2019 June 1;7(2): 90-109.
- [19] Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res*. 2019 Dec 2;47(21):10994-11006.
- [20] Keim P, Grunow R, Vipond R, Grass G, Hoffmaster A, Birdsell DN, Klee SR, Pullan S, Antwerpen M, Bayer BN, Latham J, Wiggins K, Hepp C, Pearson T, Brooks T, Sahl J, Wagner DM. Whole Genome Analysis of Injectional Anthrax Identifies Two Disease Clusters Spanning More Than 13 Years. *EBioMedicine*. 2015 Oct 6;2(11):1613-8.