ความหลากหลายทางพันธุกรรมของ SARS-CoV-2

ดร.อนันต์ จงแก้ววัฒนา

ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

ที่ปรึกษาสมาคมไวรัสวิทยา (ประเทศไทย)

การจำแนกไวรัสด้วยระบบของ GISAID

นับตั้งแต่การระบาดของไวรัส sarscoronavirus-2 (SARS-CoV-2) ซึ่งเป็นสาเหตุของโรคโควิด-19 (coronavirus disease-19, COVID-19) ที่เมืองอู่ฮั่น ประเทศจีน ตั้งแต่ปลายปี 2562 ที่ผ่านมา ไวรัสได้มีการแพร่กระจายในประชากรมนุษย์มากกว่า 210 ประเทศทั่วโลกภายในเวลาไม่ถึง 4 เดือน เนื่องจาก SARS-CoV-2 เป็นไวรัสในตระกูลโคโรนาที่มีจีโนมเป็น RNA สายบวกเส้นเดี่ยว ที่มีขนาดยาวถึง 29,903 เบส การเปลี่ยนแปลงของรหัสพันธุกรรมของไวรัส (mutation) จึงเป็นเรื่องปกติที่สามารถพบได้ แต่ด้วยคุณสมบัติเฉพาะตัวของไวรัสในตระกูลนี้ที่สามารถแก้ไขเบสที่ผิดพลาดได้ในขณะที่เพิ่มจำนวนในเซลล์เจ้าบ้าน ซึ่งจะต่างจาก RNA viruses ชนิดอื่นที่ใช้เอนไซม์ RNA dependent RNA polymerase (RdRp) เหมือนกัน แต่ไม่มี proof reading activity ทั้งนี้เนื่องจาก coronavirus มีโปรตีน Nsp14 ซึ่งมี exonuclease activity (ExoN) สามารถแก้ไขเบสที่ผิดพลาดอันเกิดจากการทำงานของเอนไซม์ RdRp ได้ ซึ่งเป็นคุณสมบัติพิเศษของ RNA virus ที่มีจีโนมเป็นสายยาว (Minskaia E, et al. Proc Natl Acad Sci U S A. 2006; 103:5108-13) ดังนั้นการเปลี่ยนแปลงทางพันธุกรรมของ SARS-CoV-2 จึงเกิดขึ้นค่อนข้างช้า ข้อมูลล่าสุดพบว่าความแตกต่างทั้งจีโนมของ SARS-CoV-2 แต่ละสายพันธุ์เฉลี่ยมีเพียงแค่ 7.23 ตำแหน่ง แต่ด้วยจีโนมที่มีขนาดใหญ่ การเปลี่ยนแปลงในตำแหน่งต่าง ๆ เมื่อถูกรวบรวมอยู่ในรูปของแผนภูมิวิวัฒนาการ (phylogenetic tree) จะสามารถเห็นภาพที่ซับซ้อนขึ้นเรื่อย ๆ เมื่อจำนวนข้อมูลของไวรัสมีมากขึ้นเพื่อให้ง่ายต่อการอ้างอิง และการศึกษาการเปลี่ยนแปลงของไวรัส นักไวรัสวิทยาจึงจำแนก SARS-CoV-2 ออกเป็น clade ย่อย ๆ โดยอาศัยระบบ GISAID (Global initiative on sharing all influenza data) เพื่อสังเกตดูการเปลี่ยนแปลงของเบสบนจีโนมของไวรัสเทียบกับสายพันธุ์เริ่มต้นคือ Wuhan-Hu-1 (GenBank accession number NC_045512.2) จึงสามารถจัดกลุ่มของ SARS-CoV-2 ได้ดังนี้

Clade L คือไวรัสสายพันธุ์เริ่มแรก หรือ Wuhan-Hu-1-like คือ สายพันธุ์เริ่มแรกที่ระบาดในเมืองอู่ฮั่นตั้งแต่ปลายปี 2562 จนถึงกลางเดือนมกราคม 2563 ไวรัสกลุ่มนี้รู้จักกันดีว่าเป็นสายพันธุ์ตั้งต้น

Clade S ในช่วงกลางเดือนมกราคมเป็นต้นมาเริ่มพบไวรัสที่มีการเปลี่ยนแปลงตรงนิวคลิโอไทด์ 2 ตำแหน่งพร้อม ๆ กัน ได้แก่ C ที่ตำแหน่ง 8,782 เปลี่ยนเป็น T (C8782T) ซึ่งตรงกับกรดอะมิโนที่ตำแหน่ง 76 ของโปรตีน NSP4 โดยการเปลี่ยนแปลงดังกล่าวเป็น silent mutation คือ ไม่มีผลต่อกรดอะมิโนในตำแหน่งดังกล่าว (S76S) และ ตำแหน่งที่สอง ที่พบในไวรัส clade นี้ คือ T ที่ตำแหน่ง 28,144 เปลี่ยนเป็น C (T28144C) ซึ่งการเปลี่ยนแปลงดังกล่าวส่งผลให้ กรดอะมิโนที่ตำแหน่ง 84 ของโปรตีน ORF8 เปลี่ยนจาก leucine ไปเป็น serine (L84S) ไวรัสที่พบว่ามีการเปลี่ยนแปลง 2 ตำแหน่งนี้จะถูกจัดจำแนกเป็น “clade S”

Clade V ในเวลาต่อมา นักไวรัสวิทยาได้พบการเปลี่ยนแปลงในรหัสพันธุกรรมของไวรัส clade L ต่อไปอีก 2 ตำแหน่งสำคัญ คือ G ที่ตำแหน่ง 11,083 เปลี่ยนไปเป็น T ส่งผลให้ กรดอะมิโนที่ตำแหน่ง 37 ของโปรตีน NSP6 เปลี่ยนจาก leucine (L) ไปเป็น phenylalanine (F) หรือ L37F และ G ที่ตำแหน่ง 26,144 เปลี่ยนไปเป็น T ส่งผลให้กรดอะมิโนที่ตำแหน่ง 251 ของโปรตีน ORF3a เปลี่ยนจาก glycine (G) ไปเป็น valine (V) หรือ G251V จึงจัดไวรัส clade L ที่มีการเปลี่ยนแปลงดังกล่าวเป็นไวรัสกลุ่มใหม่ ชื่อว่า "clade V"

Clade G ในช่วงเริ่มต้นของเดือนมีนาคม ไวรัส SARS-CoV-2 โดยเฉพาะ clade L เริ่มมีการระบาดออกนอกประเทศจีน ไปยังกลุ่มประเทศในทวีปยุโรป โดยเฉพาะประเทศ อิตาลี และเยอรมนี การศึกษาจีโนมของไวรัสที่แยกได้ในผู้ป่วยในทวีปยุโรปเริ่มพบการเปลี่ยนแปลงที่เห็นชัดที่สุดคือ เบส A ที่ตำแหน่ง 23,403 เปลี่ยนแปลงไปเป็น G (A23403G) ซึ่งการกลายพันธุ์ที่ตำแหน่งดังกล่าว ส่งผลให้กรดอะมิโนที่ตำแหน่ง 614 ของโปรตีน S (spike protein) เปลี่ยนจาก aspartate (D) เป็น glycine (G) หรือ D614G และสายพันธุ์ไวรัสที่มีการเปลี่ยนแปลงดังกล่าวจะถูกจำแนกเป็น "clade G" ปัจจุบันนี้ ไวรัสใน clade G เป็นสายพันธุ์หลักที่มีการแพร่กระจายในทุกทวีป มีหลักฐานทางวิทยาศาสตร์ยืนยันว่า การเปลี่ยนแปลง D614G ของ S protein อาจส่งผลให้ไวรัสมีความสามารถในการติดเชื้อเข้าสู่เซลล์เจ้าบ้านได้ดีขึ้น แต่การเปลี่ยนแปลงดังกล่าวไม่มีผลต่อความรุนแรงของโรคในมนุษย์ หรือสัตว์ทดลอง

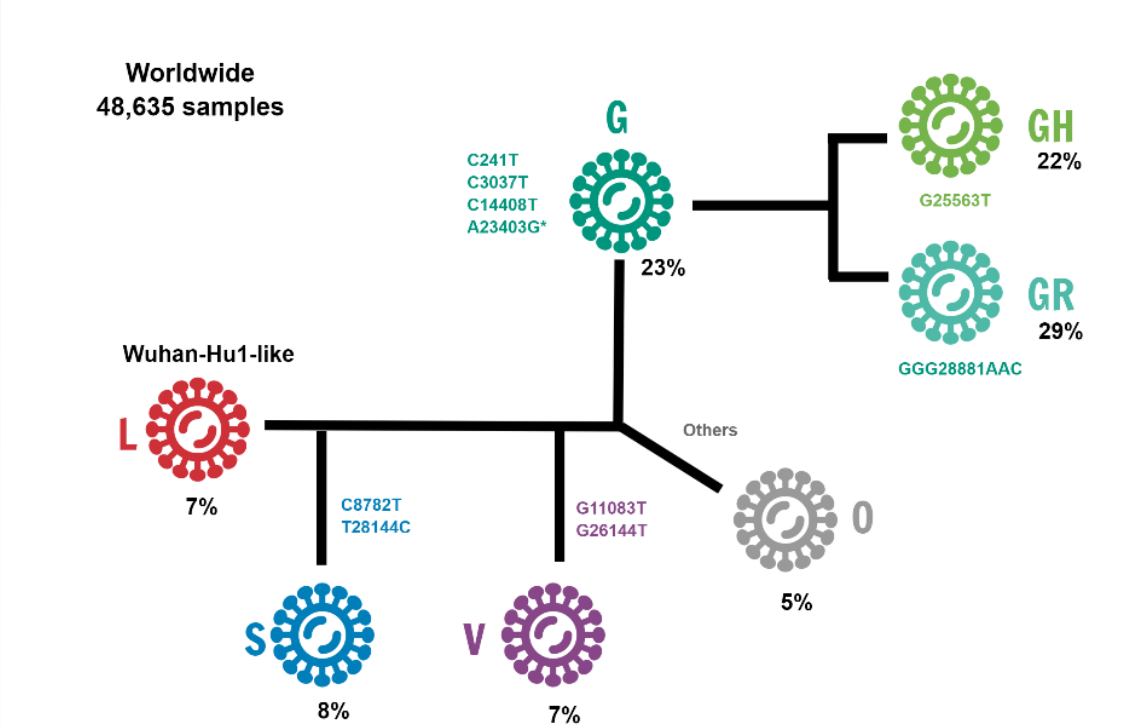
นอกจากนี้การศึกษาการเปลี่ยนแปลงทางพันธุกรรมเชิงลึกของไวรัสใน clade G เทียบกับ clade L ซึ่งเป็นสายพันธุ์เริ่มต้น ยังพบว่า นอกจาก D614G แล้ว ยังพบการเปลี่ยนเบสอีก 3 ตำแหน่งเกิดขึ้นในไวรัสกลุ่มนี้อีกด้วย คือ C ที่ตำแหน่ง 241 เปลี่ยนเป็น T (C241T) ซึ่งเป็นส่วน 5'UTR, C ที่ตำแหน่ง 3,037 เปลี่ยนเป็น T ซึ่งเป็น silent mutation (F106F) ในตำแหน่งกรดอะมิโนที่ 106 ของโปรตีน NSP3 และตำแหน่งที่ 3 คือ C ที่ตำแหน่ง 14,408 เปลี่ยนเป็น T ซึ่งส่งผลให้กรดอะมิโนตำแหน่งที่ 314 ของโปรตีน NSP12b เปลี่ยนจาก proline (P) ไปเป็น leucine (L) (P314L)

Clade GH การวิเคราะห์จีโนมพบว่าไวรัสใน clade G เริ่มมีการเปลี่ยนแปลง โดยการเปลี่ยนแปลงแรก คือ G ในตำแหน่งที่ 25,563 เปลี่ยนเป็น T ส่งผลให้กรดอะมิโนตำแหน่งที่ 57 ของโปรตีน ORF3a เปลี่ยนจาก glutamine (Q) ไปเป็น histidine (H) หรือ Q57H ไวรัสใน clade G ที่พบการเปลี่ยนแปลงดังกล่าวจะถูกแตกย่อยออกไปเป็น "clade GH" โดยไวรัสกลุ่มนี้พบได้มากที่สุดในทวีปอเมริกาเหนือ ซึ่งปัจจุบัน ไวรัส clade G มากกว่า 90% ในสหรัฐอเมริกา จัดอยู่ใน GH

Clade GR ในเวลานี้พบว่าไวรัส clade G ที่แยกได้ในทวีปยุโรป และ อเมริกาใต้ ได้มีการเปลี่ยนแปลงเบส 3 ตำแหน่งติดกันที่ตำแหน่ง 28,881-3 จาก GGG เป็น AAC โดยผลจากการเปลี่ยนแปลงดังกล่าวทำให้ตำแหน่งกรดอะมิโนของโปรตีนนิวคลิโอแคปซิด (nucleocapsid – N protein) ตำแหน่งที่ 203-204 เปลี่ยนจาก RG เป็น KR หรือ RG203KR ซึ่งไวรัส clade G ที่พบการเปลี่ยนแปลงดังกล่าว จะถูกแยกย่อยเป็น "clade GR"

Clade O โดยอักษร O มีที่มาจากคำว่า "Others" ซึ่งจะหมายถึงไวรัสที่มีการเปลี่ยนแปลงแบบสุ่ม ไม่สามารถจัดจำแนกไว้ใน clade ต่าง ๆ ตามที่ได้กล่าวมาในข้างต้น

จากแผนภาพสรุปด้านล่าง จะเห็นว่าวิวัฒนาการของไวรัส SARS-CoV-2 ในปัจจุบัน มีการเปลี่ยนแปลงจากสายพันธุ์เริ่มต้น (clade L) ไปเป็น clade G อย่างชัดเจน คิดเป็นกว่า 75% ของข้อมูลไวรัสทั้งหมดในฐานข้อมูล GISAID โดยมีสัดส่วนจำนวนไวรัสใน GR มากกว่า GH อยู่เล็กน้อย



Mercatelli D, et al. Front Microbiol. 2020;11:1800

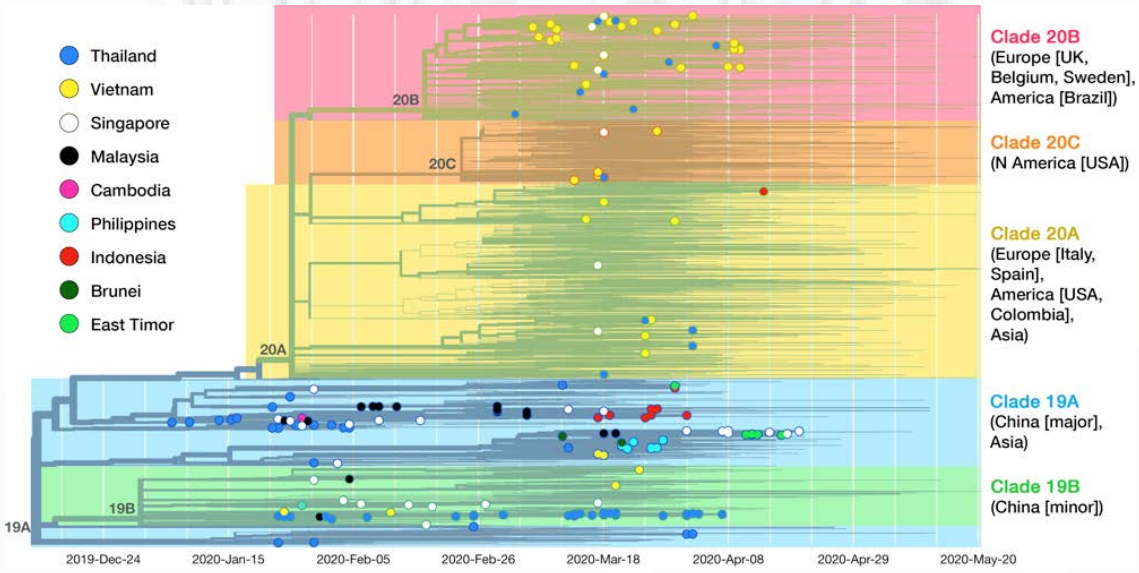
รูปที่ 1 วิวัฒนาการของ SARS-CoV-2 (ภาพโดย อนันต์ จงแก้ววัฒนา)

การจำแนกไวรัสด้วยระบบ Nextstrain

การจัดจำแนกด้วยระบบ Nextstrain บางครั้งจะมีคนเรียกว่าเป็นระบบ year-clade nomenclature โดยมีหลักการในการจัดจำแนกไวรัสโดยสังเขปดังต่อไปนี้

1. ชื่อของ clade จะเริ่มต้นด้วยเลขสองตัวหลังของปีคริสตศักราชที่พบหรือแยกเชื้อไวรัสสายพันธุ์นั้นได้ เช่น ไวรัสสายพันธุ์ Wuhan-Hu-1 ถูกแยกได้ในปี 2019 จะอยู่ใน clade 19 ส่วนไวรัสที่แยกได้จากการระบาดใหญ่ในอินเดียถูกแยกได้ในปี 2020 จะอยู่ใน clade 20 เป็นต้น
2. ใน clade ที่แยกได้ตามปีในข้อ 1. จะสามารถแยกย่อยต่อได้โดยใช้อักษรภาษาอังกฤษตัวใหญ่ คือ A, B, C โดยหลักการในการแตกย่อยเป็น clade ใหม่ คือ ต้องมีไวรัสที่มีความถี่ของการเกิด mutation นั้น ๆ มากกว่า 20% เมื่อเทียบกับข้อมูลของไวรัสที่ระบาดทั่วโลกในขณะนั้น
3. ไวรัสในช่วงปี 2019 จะถูกแบ่งออกเป็น 2 clades หลัก ๆ คือ 19A และ 19B และเมื่อเปรียบเทียบกับระบบของ GISAID แล้ว 19A คือ ไวรัส clade L, O และ V และ 19B คือ ไวรัส clade S
4. ไวรัสในช่วงปี 2020 ส่วนใหญ่จะเป็นกลุ่มที่มี D614G หรือ clade G ดังนั้น ไวรัสในกลุ่มนี้จะถูกจัดอยู่ใน clade 20A และไวรัสใน clade GR ถูกจำแนกออกมาเป็น clade 20B และไวรัส clade GH ได้ถูกจำแนกเป็น 20C

ปัจจุบันไวรัสที่จำแนกด้วยระบบนี้มีเพียง 19A, 19B, 20A, 20B, และ 20C ตามภูมิภาคที่แยกไวรัสได้ตามลำดับ ดังรูปที่ 2



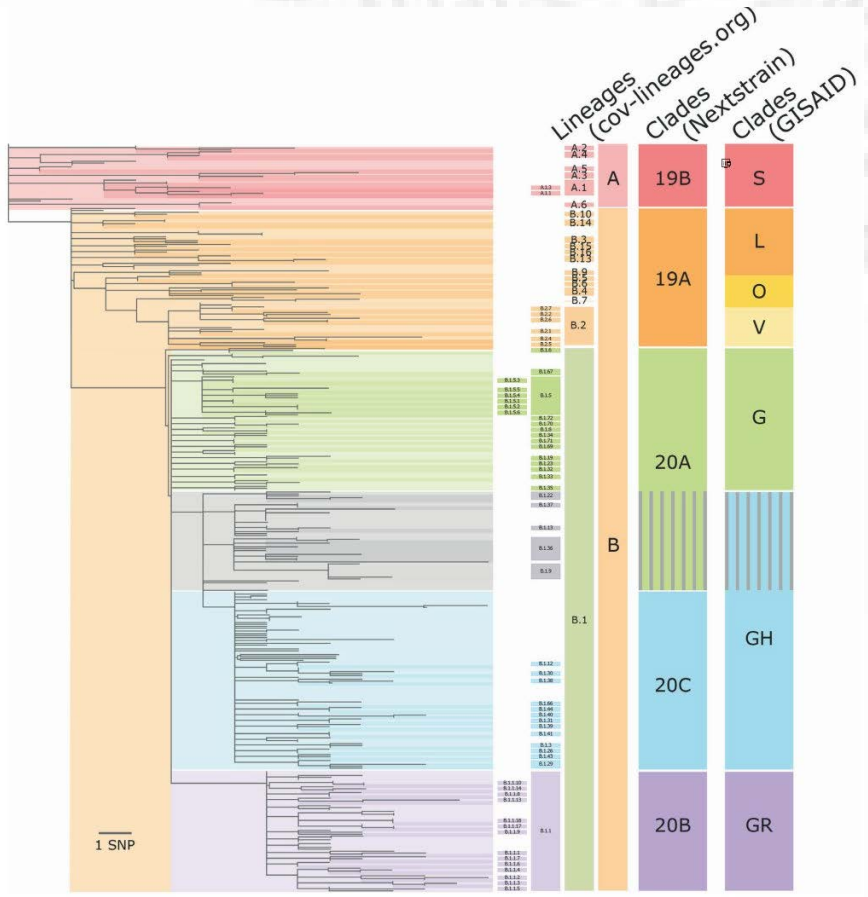
ที่มา https://covid-19chronicles.cseas.kyoto-u.ac.jp/post-041-html/

รูปที่ 2 การจำแนกสายพันธุ์ไวรัส SARS-CoV-2 ตามระบบ Nextstrain

การจำแนกไวรัสด้วยระบบ Pangolin system หรือ cov-lineage

การจัดจำแนก SARS-CoV-2 อีกระบบหนึ่งที่นิยมใช้อ้างอิงโดยเฉพาะกับสายพันธุ์ที่มีการระบาดในแต่ละท้องที่ โดยได้มีการแบ่งไวรัสออกเป็น 2 lineages คือ A และ B แทนที่จะแบ่งเป็นหน่วยใหญ่ คือ clade เหมือนใน 2 ระบบก่อนหน้านี้ หลักการใหญ่ ๆ ของระบบนี้จะขึ้นอยู่กับสมมติฐานที่ว่า ไวรัสต้นกำเนิดของ SARS-CoV-2 คือ ไวรัสที่พบในค้างคาว ได้แก่สายพันธุ์ RaTG13 หรือ RmYN02 โดยยึดเอา SARS-CoV-2 ที่มีตำแหน่งเบสเหมือนกับไวรัสค้างคาว เป็นสายพันธุ์บรรพบุรุษใกล้สุด (most recent common ancestor- MRCA) คือ ตำแหน่ง 8,782 ในยีน ORF1ab และ ตำแหน่ง 28,144 ในยีน ORF8 เช่น เชื้อ SARS-CoV-2 สายพันธุ์ Wuhan/WH04/2020 จัดอยู่ใน lineage A ส่วนสายพันธุ์อื่นที่เบส 2 ตำแหน่งนี้แตกต่างออกไป จะจัดอยู่ใน lineage B ซึ่งเป็นที่น่าสนใจว่าสายพันธุ์ Wuhan-Hu-1 ถึงแม้จะถูกถอดรหัสได้ก่อน แต่ด้วยเหตุผลดังกล่าวจึงถูกจัดอยู่ใน Lineage B

จาก lineage ตั้งต้น นักไวรัสวิทยาได้กำหนดกฎเกณฑ์ในการใส่ตัวเลขที่ตามหลัง lineage เช่น A.1.1 หรือ B.1.1 หรือ B.2.2 เป็นต้น โดยตัวเลขยิ่งอยู่ใกล้กันมากก็จะมีความใกล้เคียงกันมาก เช่น B.2.2.2 จะมีความใกล้เคียงกับ B.2.3.2 มากกว่า B.1.1.1 เป็นต้น เนื่องจากหลักเกณฑ์ในการใส่ตัวเลขดังกล่าวค่อนข้างซับซ้อน นักวิจัยที่ออกแบบระบบนี้มาจึงได้พัฒนา website ชื่อว่า Pangolin (Phylogenetic Assignment of Named Global Outbreak LINeages) [https://pangolin.cog-uk.io/] เพื่อเป็นการอำนวยความสะดวกให้ผู้ใช้สามารถตรวจสอบได้ว่ารหัสพันธุกรรมของไวรัสที่ศึกษาอยู่นั้น จัดอยู่ใน lineage หรือ clade ใดเมื่อเปรียบเทียบกับฐานข้อมูลที่ได้จากการจัดจำแนกโดยใช้ระบบอื่น ดังแสดงในรูปที่ 3



รูปที่ 3. ความสัมพันธ์ของระบบที่ใช้ในการจัดจำแนก SARS-CoV-2 (Alm E, et al. Euro Surveill. 2020;25(32): 2001410)

SNP: sigle nucleotide polymorphism