การใช้ฐานข้อมูลจาก EMPOP ในการวิเคราะห์ความถี่ haplotype

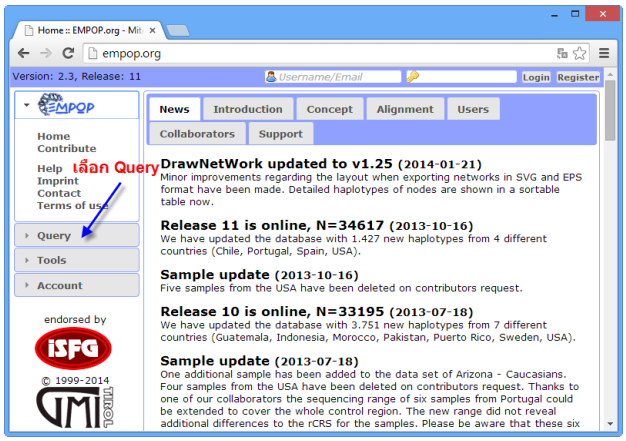
สุคนธ์ ประดุจกาญจนา

หน่วยนิติเวชศาสตร์และพิษวิทยา ภาควิชาพยาธิวิทยา คณะแพทยศาสตร์

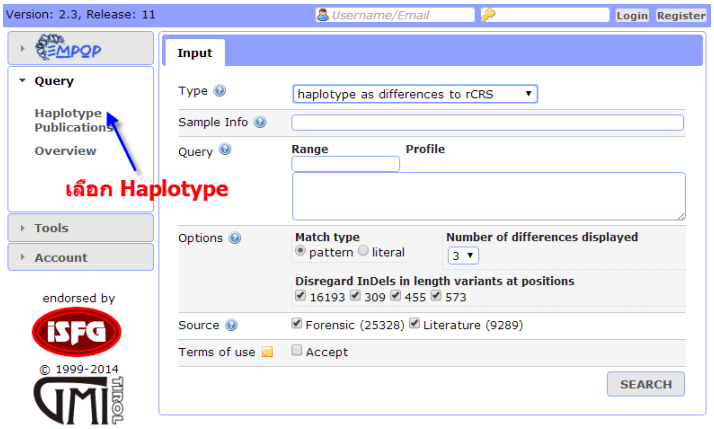
มหาวิทยาลัยสงขลานครินทร์ หาดใหญ่ สงขลา 90110.

E-mail address: mitojin@live.com

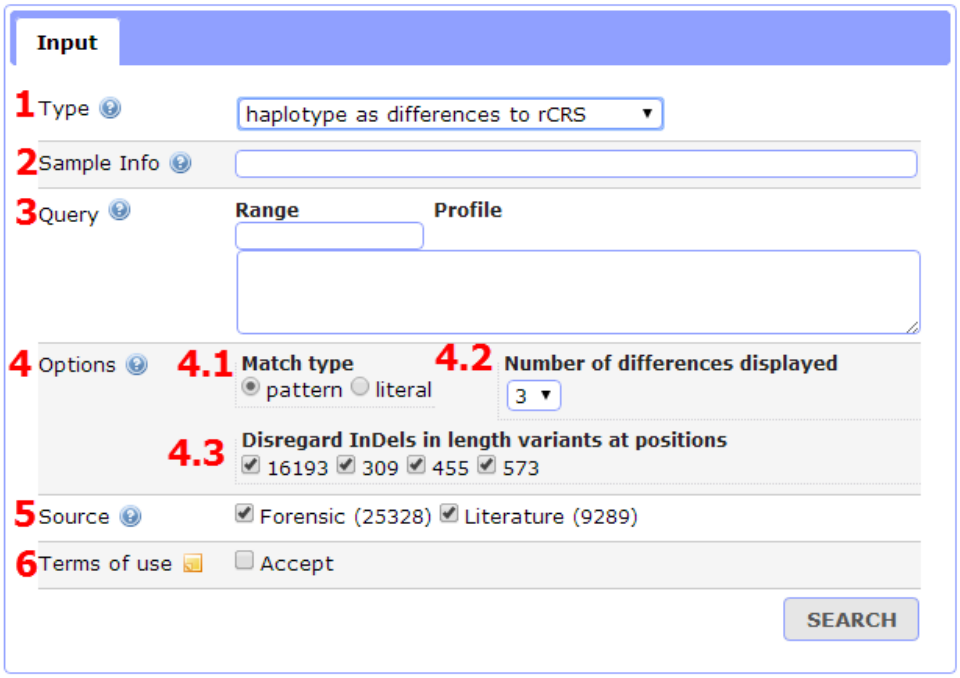
เว็บ http://empop.org เป็นหน้าเว็บไซต์สำหรับใช้ค้นหารูปแบบดีเอ็นเอบนไมโตคอนเดรียที่เราสนใจโดยเปรียบเทียบกับรูปแบบดีเอ็นเอในฐานข้อมูลว่ามีรูปแบบตรงกันหรือเหมือนกันหรือไม่ จำนวนเท่าไร เพื่อที่จะใช้วิเคราะห์ความถี่ haplotype นอกจากนั้น ยังสามารถใช้ตรวจสอบคุณภาพของลำดับเบสบนไมโตคอนเดรียที่ได้ว่ามีคุณภาพที่น่าเชื่อถือมากน้อยเพียงใด โดยใช้วิธี phylogeny แบบ quasi-median network สำหรับเนื้อหาในบทนี้จะขอกล่าวถึงเฉพาะการค้นหารูปแบบดีเอ็นเอบนไมโตคอนเดรียเพื่อวิเคราะห์หาความถี่ haplotype เท่านั้น



ให้เลือก Query แล้วเลือก Haplotype หน้าเว็บจะแสดงหน้าใส่ข้อมูล ดังภาพ



จากนั้นให้ใส่ข้อมูลต่างๆ ดังนี้



1. Type เป็นช่อง drop down list ให้เลือกใส่ข้อมูลได้ 2 แบบ ได้แก่

1.1 ตำแหน่งที่แตกต่างจากสายดีเอ็นเออ้างอิง rCRS

1.2 การเรียงตัวของลำดับเบส (sequence string หรือ FASTA)

2. Sample info เป็นข้อมูลของตัวอย่างที่ต้องการค้นหาเพื่อเปรียบเทียบรูปแบบดีเอ็นเอบนไมโตคอนเดรียกับรูปแบบดีเอ็นเอในฐานข้อมูล ข้อมูลนี้เป็นข้อมูลเฉพาะที่แสดงเอกลักษณ์ของตัวอย่าง เช่น หมายเลขรหัสตัวอย่างตรวจ เป็นต้น

3. Query

หากเลือก Type (ข้อ 1) เป็น ตำแหน่งที่แตกต่างจากสายดีเอ็นเออ้างอิง rCRS ในช่อง Profile ท่านสามารถใส่ตำแหน่งที่แตกต่างจากสายดีเอ็นเออ้างอิง rCRS แต่ละตำแหน่ง โดยแยกกันด้วยเครื่องหมายวรรค เช่น 263G 309.1C 309.2C เป็นต้น การขาดหายไป (deletion) สามารถพิมพ์ข้อมูลเข้าได้หลายวิธี เช่น 16193DEL หรือ 16193del หรือ 16193- สำหรับ 16193D โปรแกรมจะเข้าใจว่าเป็นเบสผสมระหว่าง A, G, และ T ตามวิธีกำหนดรหัสเบสผสมของ IUB code

หากเลือก Type (ข้อ 1) เป็น การเรียงตัวของลำดับเบส ในช่อง Profile ท่านสามารถพิมพ์ลำดับเบสในรูปแบบที่เป็นตัวอักษร (รูปแบบ FASTA แต่ไม่มีข้อมูลส่วนหัวที่แสดงข้อมูลเฉพาะของตัวอย่าง: เฉพาะลำดับเบสที่เรียงตัวเท่านั้น) ซึ่งท่านสามารถใช้วิธีการ copy & paste ลำดับเบสที่ได้จากการใช้โปรแกรมคอมพิวเตอร์ alignment หรือ จากข้อความก็ได้

Range : ช่วงลำดับเบส ที่ต้องการค้นหา เช่น

บริเวณ HV1 ตำแหน่งที่ 16024-16365

บริเวณ HV2 ตำแหน่งที่ 73-340

บริเวณ HV3 ตำแหน่งที่ 438-576

บริเวณทั้ง control region ตำแหน่งที่ 16024-576

ตำแหน่ง insertion หลังตำแหน่งสุดท้ายของช่วงลำดับเบสที่ค้นหาอาจถูกพิจารณาว่า ออกนอกช่วงที่ค้นหาได้ เช่น ตำแหน่ง 455.1C โปรแกรมจะคิดว่าอยู่นอกช่วงค้นหา 450-455 เป็นต้น ในกรณีนี้จึงต้องขยายช่วงค้นหาให้ครอบคลุมบริเวณ insertion ทั้งหมด เช่น กำหนดว่าเป็น 450-456 เป็นต้น

4. Option กำหนดเงื่อนไขการค้นหาต่างๆ

4.1 Match type เป็นการกำหนดเงื่อนไขการค้นหารูปแบบดีเอ็นเอบนไมโตคอนเดรีย

- Pattern match : ตำแหน่งที่แตกต่างจากสายดีเอ็นเออ้างอิงที่เป็นเบสผสม เมื่อค้นด้วยเงื่อนไขนี้ จะเข้าได้กับลำดับเบสที่เป็นส่วนผสมของเบสผสมนั้นๆ เช่น Y = {C, T, Y} เป็นต้น

ตัวอย่างเช่น ตำแหน่ง 152Y เข้าได้กับ 152T และ 152C

- Literal match : ตำแหน่งที่แตกต่างจากสายดีเอ็นเออ้างอิงที่เป็นเบสผสม เมื่อค้นด้วยเงื่อนไขนี้ จะเข้าได้กับลำดับเบสแบบผสมที่เหมือนกันเท่านั้น เช่น Y = {Y} เป็นต้น

ตัวอย่างเช่น ตำแหน่ง 152Y เข้าได้กับ 152Y เท่านั้น ทำให้เงื่อนไขการค้นหาแบบนี้ มีประโยชน์ในการค้นหารูปแบบดีเอ็นเอที่เป็น heteroplasmic ในฐานข้อมูล EMPOP

4.2 Number of differences displayed เป็นการแสดงจำนวนที่แตกต่างระหว่างรูปแบบดีเอ็นเอที่ต้องการค้นหากับรูปแบบดีเอ็นเอในฐานข้อมูล (กำหนดช่วงระหว่าง 1-5) สำหรับจำนวนที่แตกต่างจากรูปแบบดีเอ็นเอในฐานข้อมูลที่ค้นได้แล้วมีจำนวนมากกว่าช่วงที่กำหนดจะไม่ถูกแสดงผล

กำหนดค่าเริ่มต้น ที่ 5 (หรือ 3 สำหรับผู้ใช้ที่ไม่ได้ลงทะเบียนการใช้ฐานข้อมูล)

4.3 ไม่สนใจ InDels ในความแตกต่างด้านความยาวที่ตำแหน่ง 16193, 309, 455 และ 573

Indels บริเวณที่เกิดการกลายพันธุ์บ่อยๆ ทำให้มีความยาวแตกต่างกัน สามารถกำหนดให้ไม่นำมาใช้ค้นหา เงื่อนไขนี้มักเกี่ยวพันกับการเพิ่มขึ้นหรือลดลงของเบส C ที่ตำแหน่ง 16193, 309 และ 573 และการเพิ่มขึ้นหรือลดลงของเบส T ที่ตำแหน่ง 455 ผู้ใช้สามารถกำหนดเงื่อนไขการยกเว้นตำแหน่งเหล่านี้ในการค้นหาได้อย่างอิสระ

ตัวอย่างที่ 1 : รูปแบบดีเอ็นเอในฐานข้อมูล เป็น 16189C 16193.1C 16519C 263G 315.1C เปรียบเทียบกับรูปแบบดีเอ็นเอที่ต้องการค้นหา เป็น 16189C 16519C 263G 315.1C หากกำหนดเงื่อนไขการยกเว้นการค้นหาที่ตำแหน่ง 16193 จะทำให้ ตำแหน่ง 16193.1 ไม่ถูกนำไปใช้ในการค้นหา

ตัวอย่างที่ 2 : รูปแบบดีเอ็นเอในฐานข้อมูล เป็น 16189C 16192T 16270T 16398A 73G 150T 263G 315.1C เปรียบเทียบกับรูปแบบดีเอ็นเอที่ต้องการค้นหา เป็น 16189C 16191.1C 16192T 16270T 16398A 73G 150T 263G 315.1C ตำแหน่งที่ 16191.1C จะไม่ถูกยกเว้นการค้นหาแม้ว่าจะกำหนดเงื่อนไขให้ตำแหน่ง 16193 เป็นตำแหน่งที่ยกเว้นการค้นหาก็ตาม

5. Source แหล่งข้อมูลประชากร ประกอบด้วย 2 แหล่ง ได้แก่

- Forensic data : ข้อมูลด้านนิติเวชศาสตร์ เป็นรูปแบบดีเอ็นเอในฐานข้อมูลที่มีการเชื่อมโยงไว้กับข้อมูลดิบที่เป็นลำดับเบสที่มีคุณภาพสูง สามารถตรวจสอบความถูกต้องของข้อมูลได้ทันทีที่ต้องการ

- Literature data : ข้อมูลที่ได้จากการตีพิมพ์ แม้จะเป็นลำดับเบสที่มีคุณภาพสูง แต่การตรวจสอบความถูกต้องของข้อมูลไม่อาจจะทำได้ทันที เนื่องจากไม่ได้มีการเก็บข้อมูลดิบไว้ในฐานข้อมูล

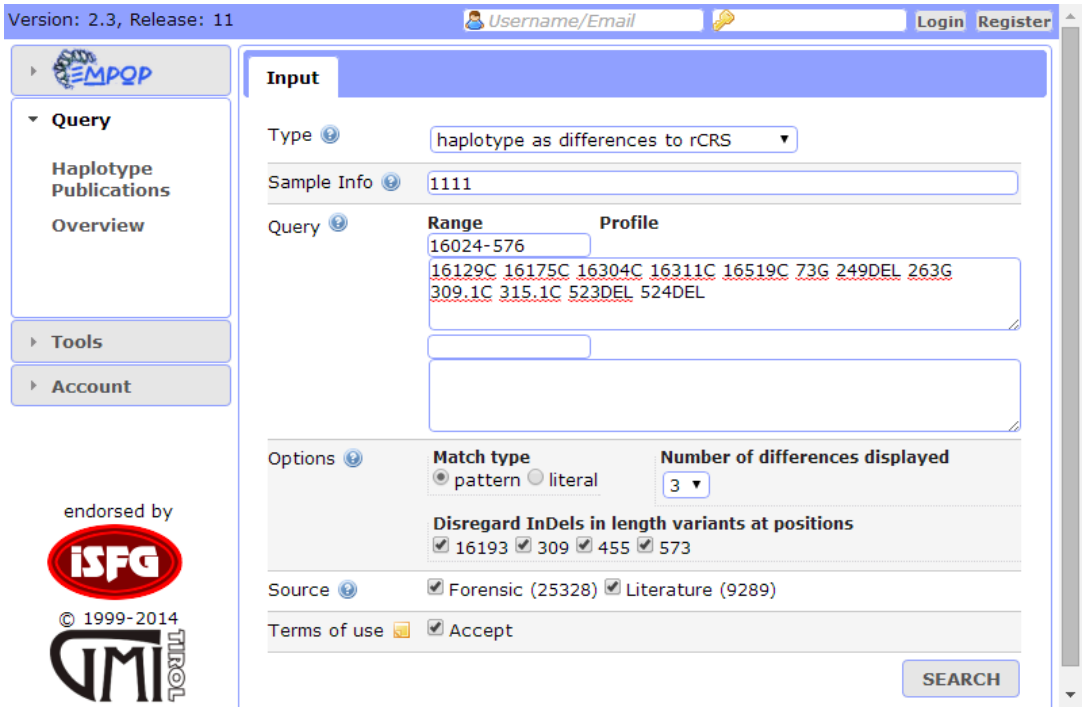
ฐานข้อมูล EMPOP 2 มีการใช้โครงสร้างใหม่สำหรับเก็บรูปแบบดีเอ็นเอ ซึ่งเหมาะสมอย่างยิ่งกับการใช้งานทางนิติเวชศาสตร์ โดยรูปแบบดีเอ็นเอเหล่านี้ถูกจัดกลุ่มข้อมูลตามลักษณะต่อไปนี้

- ลักษณะทางภูมิศาสตร์ เช่น ตามทวีป ภูมิภาค ประเทศ หรือท้องถิ่น เป็นต้น
- กลุ่มประชากร โดยกำหนดลำดับชั้นของประชากรมากถึง 4 ลำดับชั้น

ข้อมูลเหล่านี้ สามารถกำหนดได้จากหน้าแสดงผลการค้นหารูปแบบดีเอ็นเอที่ได้จากฐานข้อมูล

6. Term of use : เป็นรายละเอียดแสดงข้อตกลงในการใช้งาน ให้คลิ๊กเลือกช่อง Accept จากนั้นกดปุ่ม SEARCH โปรแกรมจะทำการค้นหารูปแบบดีเอ็นเอที่ต้องการ เปรียบเทียบกับข้อมูลรูปแบบดีเอ็นเอที่อยู่ในฐานข้อมูล

ภาพแสดงการใส่ข้อมูลสำหรับค้นหารูปแบบดีเอ็นเอในฐานข้อมูล EMPOP



เมื่อกดปุ่ม SEARCH แล้ว โปรแกรมจะทำการค้นหารูปแบบดีเอ็นเอที่ต้องการเปรียบเทียบกับรูปแบบดีเอ็นเอในฐานข้อมูล แล้วแสดงผลการค้นหาดังนี้

Sample Info	1111
Type	string-based search: haplotype as differences to rCRS
Options	Match type: **pattern** Maximum differences displayed: **3** Disregard InDels in length variants at positions: **16193 309 455 573**
Source	**Forensic data (22516/25328)** **Literature data (3611/9289)**
Query	16024-576 G16129C A16175C T16304C T16311C T16519C A73G

Geographic affiliation All

Metapopulation All

DIFFERENCES TO QUERY PROFILE	NUMBER OF HAPLOTYPES	CUMULATIVE NUMBER OF HAPLOTYPES
0	0	0
1	0	0
2	1	1
3	19	20
4+	26107	26127

NUMBER OF HAPLOTYPES

สำหรับการใช้งานของผู้ใช้ที่ไม่ได้ลงทะเบียนการใช้งานกับฐานข้อมูล EMPOP จะกำหนดเงื่อนไขจำนวนที่แตกต่างระหว่างรูปแบบดีเอ็นเอที่ค้นหากับรูปแบบดีเอ็นเอในฐานข้อมูลได้สูงสุดไม่เกิน 3 ตำแหน่งร่วมกับสามารถค้นข้อมูลประเภท Forensic data ในฐานข้อมูลได้เพียง 22,516 ข้อมูล และค้นข้อมูลประเภท Literature data ในฐานข้อมูลได้เพียง 3,611 ข้อมูลเท่านั้น รวมเป็นจำนวน 26,127 ข้อมูล

ผู้ใช้ที่มีการลงทะเบียนการใช้งานกับฐานข้อมูล จะกำหนดเงื่อนไขจำนวนที่แตกต่างระหว่างรูปแบบดีเอ็นเอที่ค้นหากับรูปแบบดีเอ็นเอในฐานข้อมูลได้สูงสุดไม่เกิน 5 ตำแหน่ง ร่วมกับสามารถค้นข้อมูลประเภท Forensic data ในฐานข้อมูลได้ 25,328 ข้อมูล และค้นข้อมูลประเภท Literature data ในฐานข้อมูลได้ 9,289 ข้อมูล รวมเป็นจำนวน 34,617 ข้อมูล

ในช่อง Geographic affiliation กำหนดค่าเริ่มต้นไว้ที่ all หมายถึงค้นหาจากทุกทวีป ผู้ใช้สามารถเลือกค้นหาเฉพาะทวีป หรือ ภูมิภาค หรือ ประเทศที่ต้องการได้

ผู้ใช้สามารถเลือก metapopulation ซึ่งเป็นการกำหนดให้แสดงเฉพาะกลุ่มประชากรที่สนใจได้

สำหรับการใช้งานด้านการวิเคราะห์ความถี่ haplotype ของไมโตคอนเดรียนั้น ข้อมูลรูปแบบดีเอ็นเอในประชากรไทย มีจำนวนเพียง 190 ข้อมูลเท่านั้น ซึ่งน้อยเกินไปที่จะใช้คำนวณค่าทางสถิติ ในทางปฏิบัติแล้วจะกำหนดให้ค้นหารูปแบบดีเอ็นเอในประเทศแถบภูมิภาค South East Asia ซึ่งมีข้อมูลรวมกันจำนวน 1,144 ข้อมูล แล้วเลือก metapopulation ให้เป็น all

Geographic affiliation South-Eastern Asia ▾ **Metapopulation** All ▾

DIFFERENCES TO QUERY PROFILE	NUMBER OF HAPLOTYPES	CUMULATIVE NUMBER OF HAPLOTYPES
0	0	0
1	0	0
2	1	1
3	14	15
4+	1129	1144

NUMBER OF HAPLOTYPES

จากภาพแสดงผลการค้นหารูปแบบดีเอ็นเอที่ต้องการ ในภูมิภาค South East Asia ในประชากรทุกกลุ่ม (all metapopulation) พบว่ามีรูปแบบดีเอ็นเอในฐานข้อมูลที่แตกต่างจากรูปแบบดีเอ็นเอที่ค้นหา มีจำนวนเท่ากับ 0 (รูปแบบดีเอ็นเอที่ค้นหาเหมือนกับรูปแบบดีเอ็นเอในฐานข้อมูล) มีผลการค้นหาเท่ากับ 0 แสดงว่า ไม่พบรูปแบบดีเอ็นเอที่ค้นหานี้ในฐานข้อมูล จากจำนวนข้อมูลในฐานทั้งสิ้น 1,144 ข้อมูล

รูปแบบดีเอ็นเอในฐานข้อมูลที่แตกต่างจากรูปแบบดีเอ็นเอที่ค้นหา มีจำนวนเท่ากับ 1 (รูปแบบดีเอ็นเอที่ค้นหาต่างจากรูปแบบดีเอ็นเอในฐานข้อมูลเพียง 1 ตำแหน่ง) มีผลการค้นหาเท่ากับ 0 แสดงว่า แม้กำหนดให้รูปแบบดีเอ็นเอที่ค้นหาสามารถแตกต่างจากรูปแบบดีเอ็นเอในฐานข้อมูลได้ 1 ตำแหน่ง ก็ยังไม่พบในฐานข้อมูล จากจำนวนข้อมูลในฐานทั้งสิ้น 1,144 ข้อมูล

รูปแบบดีเอ็นเอในฐานข้อมูลที่แตกต่างจากรูปแบบดีเอ็นเอที่ค้นหา มีจำนวนเท่ากับ 2 (รูปแบบดีเอ็นเอที่ค้นหาต่างจากรูปแบบดีเอ็นเอในฐานข้อมูล 2 ตำแหน่ง) มีผลการค้นหาเท่ากับ 1 แสดงว่า การกำหนดให้รูปแบบดีเอ็นเอที่ค้นหาสามารถแตกต่างจากรูปแบบดีเอ็นเอในฐานข้อมูลได้ 2 ตำแหน่งนั้น พบว่ามีรูปแบบดีเอ็นเอในฐานข้อมูลที่เข้าได้กับเงื่อนไขนี้ จำนวน 1 ข้อมูล จากจำนวนข้อมูลในฐานทั้งสิ้น 1,144 ข้อมูล

สำหรับการคำนวณความถี่ haplotype ในเว็บไซต์นี้จะมีสูตรคำนวณ อยู่ 2 สูตร ด้วยกันคือ

สูตรที่ 1. ความถี่ haplotype = k/n เมื่อ k = จำนวนรูปแบบดีเอ็นเอในฐานข้อมูลที่เหมือนกับรูปแบบดีเอ็นเอที่ค้นหา และ n = จำนวนข้อมูลทั้งหมดในฐานข้อมูล

สูตรที่ 2. ความถี่ haplotype = (k+1)/(n+1)

นอกจากนั้นในเว็บไซต์นี้ ยังมีการประมาณค่าช่วงความเชื่อมั่นที่ 95% ของค่าความถี่ haplotype ที่คำนวณได้ ด้วยวิธีของ Wilson, 1927 อย่างไรก็ตามวิธีการประมาณค่าช่วงความเชื่อมั่นยังมีอีกหลายวิธี ซึ่งข้อกำหนดของ SWGDAM ฉบับปี 2013 กำหนดให้ใช้การคำนวณความถี่ haplotype ตามสูตรที่ 1 แล้วคำนวณค่าช่วงบนของความเชื่อมั่นที่ 95% ด้วยวิธีของ Clopper and Pearson, 1934 ซึ่งวิธีการคำนวณนี้สามารถใช้โปรแกรมคำนวณค่าทางสถิติ บนเว็บไซต์ได้

เอกสารอ้างอิง

1. ฐานข้อมูลไมโตคอนเดรีย ออนไลน์ EMPOP [cited 2014 Sep 4]. Available from : http://empop.org/