# เทคโนโลยีเอ็นจีเอสและการประยุกต์ในงานวิจัยโอมิกส์

# Next generation sequencing (NGS) technologies and their applications in omics-research

**อลิษา วิลันโท[1] อรนุช ประดิษฐ์ทรัพย์[2] วรรณวิสาข์ เจริญฉิม[1] ศุภศักดิ์ กุลวงศ์อนันชัย[1] อนันต์ชัย อัศวเมฆิน[1] และศิษเฎศ ทองสิมา[1]***

**Alisa Wilantho[1], Oranud Praditsup[2], Wanwisa Charoenchim[1], Supasak Kulawonganunchai[1], Anunchai Assawamakin[1] and Sissades Tongsima[1]***

[1]*ห้องปฏิบัติการชีวสถิติและสารสนเทศ สถาบันจีโนม ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ สวทช. ปทุมธานี 12120*

[2]*หน่วยอณูพันธุศาสตร์ สถานส่งเสริมการวิจัย คณะแพทยศาสตร์ศิริราชพยาบาล กรุงเทพฯ 10700*

[1]*Bio-statistics and informatics laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), NSTDA, Pathum Thani 12120, Thailand*

[2]*The Department of Research and Development, Division of Molecular Genetics, Mahidol University, Bangkok 10700, Thailand*

**Corresponding author: sissades@biotec.or.th*

## บทคัดย่อ

เทคโนโลยีเอ็นจีเอสหรือ Next generation sequencing technology มีวัตถุประสงค์หลัก เพื่อหาลำดับเบสทั้งจีโนมของสิ่งมีชีวิตด้วยวิธีการหาลำดับเบสแบบขนานที่ทำได้อย่างรวดเร็วและมีประสิทธิภาพ ทำให้ได้ข้อมูลลำดับเบสจำนวนมหาศาลที่มีต้นทุนค่าใช้จ่ายที่ถูกลงกว่าการหาลำดับเบสด้วยเทคโนโลยีของแซงเกอร์ ปัจจุบันเครื่องมือของเทคโนโลยีเอ็นจีเอสมีสามรูปแบบหลักที่นิยมใช้กันอย่างกว้างขวางได้แก่ 454/Roche, Illumina, และ SOLiD/Life Technology (ABI) โดยเครื่องมือต่างๆ เหล่านี้มีหลักการหาลำดับเบสรูปแบบของข้อมูลลำดับเบสสายสั้น (reads) และจำนวนของข้อมูล reads ที่ออกมาแตกต่างกันไป รวมถึงประโยชน์ที่สามารถนำไปประยุกต์ใช้ในงานวิจัยที่เกี่ยวข้องกับโอมิกส์ได้หลากหลาย เช่น การทำ *de novo* sequencing, target resequencing, RNA sequencing และ metagenomics เป็นต้น บทความนี้ จึงนำเสนอหลักการการหาลำดับเบสของเครื่องมือทั้งสามรูปแบบ (454/Roche, Solexa/Illumina และ SOLiD/ABI) การประยุกต์ใช้ในงานวิจัยต่างๆ และตัวอย่างโปรแกรมที่ใช้สำหรับจัดเรียงลำดับเบสที่ได้จากเทคโนโลยีเอ็นจีเอส เพื่อให้ทราบถึงความหลากหลายของลำดับเบสที่เกิดขึ้นและนำไปสู่การตอบโจทย์ในงานวิจัยต่างๆต่อไป

## ABSTRACT

The main objective of Next generation

sequencing (NGS) technology is to quickly and efficiently read the underlying sequence of an organism by means of massively parallel sequencing. Comparing with Sanger's technology, NGS can produce much larger number of sequence reads that cover huge number of bases per run. Three major platforms of NGS technologies are currently in used, namely 454/Roche, Solexa/Illumina and SOLiD/Life Technology (ABI). These platforms differ in sequencing principles, sequence read formats and number of read outputs. Moreover, these platforms are useful in the different applications in several omic-research domains including *de novo* sequencing, target resequencing, RNA sequencing and metagenomics etc. This review article presents the overview of the three major NGS technologies. We cover the principle of sequencing in three platforms (454/Roche, Solexa/Illumina and SOLiD/ABI), applications, and mapping tools used to discover the variation of read sequences leading to deciphering the problems in the other research topics.

**คำสำคัญ:** เทคโนโลยีเอ็นจีเอส, ลำดับเบสสายสั้น, โปรแกรมที่ใช้สำหรับจัดเรียงลำดับเบส

**Keywords:** NGS technology, reads, mapping tools

## บทนำ

ปัจจุบัน ข้อมูลทางชีววิทยาที่ได้จากการศึกษาศาสตร์ด้านโอมิกส์ (omics) ไม่ว่าจะเป็นจีโนมิกส์ ทรานคริปโตมิกส์ โปรติโอมิกส์ หรือ เมตาโบโลมิกส์ มีจำนวนเพิ่มขึ้นมากมาย โดยเทคโนโลยีที่มีบทบาทสำคัญและเป็นจุดเริ่มต้นของการนำไปสู่การไขความลับทางชีววิทยานั้น คือ การศึกษาลำดับเบสหรือ DNA sequencing โดยข้อมูลต่างๆ เหล่านี้พร้อมที่จะถูกประมวลผล เพื่อนำไปสู่การทำนายปรากฏการณ์ทางพันธุกรรมหรือวิวัฒนาการของสิ่งมีชีวิตชนิดต่างๆ ได้ การค้นพบที่ประสบความสำเร็จเป็นอย่างมาก คือ การหาลำดับเบสจีโนมของมนุษย์ ที่เกิดขึ้นในปี ค.ศ. 2000 (Yamey, 2000) ซึ่งเป็นยุคที่มีการคิดค้นและพัฒนาเทคโนโลยีการหาลำดับเบสให้มีความรวดเร็ว และมีราคาถูกลงอย่างต่อเนื่อง ปัจจุบันมีการริเริ่มเทคโนโลยีใหม่สำหรับหาลำดับเบสแบบ high throughput ที่เรียกว่า Next generation sequencing หรือเอ็นจีเอส (NGS)

หลายองค์กรพยายามพัฒนาเทคโนโลยีเอ็นจีเอส ให้สามารถใช้ในการหาลำดับเบสได้รวดเร็วและแม่นยำขึ้น รวมทั้งต้นทุนในการดำเนินการที่มีราคาถูกลง เพื่อประโยชน์ในการประยุกต์ใช้กับสิ่งมีชีวิตที่มีความหลากหลายมากขึ้น โดยแต่เดิมนั้น เทคนิคการหาลำดับเบสใช้เทคนิคที่เรียกว่าแซงเกอร์ (Sanger chemistry) ซึ่งอาศัยหลักการ dideoxynucleotide chain termination ด้วยการนำนิวคลีโอไทด์มาติดฉลากรังสี หรือสารเรืองแสง ซึ่งแต่ละนิวคลีโอไทด์จะมีการติดฉลากสีที่ต่างกัน จากนั้นใช้เอนไซม์ในการสร้างดีเอ็นเอสายใหม่จากดีเอ็นเอต้นแบบ ลำดับเบสที่บันทึกได้ใหม่จะประกอบด้วยนิวคลีโอไทด์แต่ละตัวที่ติดฉลากนั้น โดยวิธีนี้สามารถอ่านลำดับเบสได้ช่วง 1,000 ถึง 1,200 คู่เบส ด้วยข้อจำกัดในการอ่านลำดับเบสที่ยาวกว่าสองกิโลเบส จึงมีการพัฒนาวิธีที่เรียกว่า shotgun sequencing วิธีนี้อาศัยการหาลำดับเบสจากชิ้นส่วนย่อยๆ และจัดเรียงกลับเป็นดีเอ็นเอเส้นเดิมอย่างสมบูรณ์ ด้วยโปรแกรม sequence assembler ลำดับเบสที่ได้จึงมีความยาวพอที่จะนำข้อมูลมาทำนายโครงสร้าง และตำแหน่งยีนในจีโนม

ได้อย่างรวดเร็วขึ้น จากความรู้ที่ได้จากการศึกษา shotgun sequencing ทำให้มีการพัฒนาเทคโนโลยีเอ็นจีเอส ที่มีความสามารถในการหาลำดับเบสแบบ parallel ทำให้ได้ข้อมูลลำดับเบสจำนวนมหาศาล (Zhang *et al*., 2011) เทคโนโลยีเอ็นจีเอสอาศัยการอ่านดีเอ็นเอต้นแบบแบบสุ่มทั้งจีโนม และตัดชิ้นส่วนดีเอ็นเอเป็นชิ้นสั้นๆ นำชิ้นดังกล่าวเชื่อมต่อเข้ากับอะแด็บเตอร์ (adaptor) ในช่วงการสังเคราะห์ดีเอ็นเอนั้นและเพิ่มปริมาณต้นแบบ วิธีนี้สามารถเพิ่มจำนวนของดีเอ็นเอได้พร้อมๆ กัน (parallel) เมื่อหาลำดับเบสจะได้ลำดับเบสสายสั้น (reads) เกิดขึ้นจำนวนมาก ซึ่งในที่นี้จะใช้คำทับศัพท์ว่า reads โดยความยาวของ reads ที่ได้ จะมีความยาวเฉลี่ย 30 ถึง 500 คู่เบส ขึ้นอยู่กับเครื่องมือแต่ละแบบ อย่างไรก็ตาม อาจกล่าวได้ว่าเทคนิคแซงเกอร์และเทคโนโลยีเอ็นจีเอส เริ่มต้นจากการทำดีเอ็นเอให้เป็นชิ้นสั้นๆ เหมือนกัน แต่เทคนิคการหาลำดับเบสแบบแซงเกอร์ใช้การเพิ่มปริมาณดีเอ็นเอต้นแบบด้วยการทำโคลนนิง (cloning) ในตอนแรก ซึ่งวิธีนี้อาจส่งผลต่อความผิดพลาด (bias) ที่เกิดขึ้น เช่น เกิดภาวะไม่เสถียรในช่วงของการโคลน หรือเกิดการหายไปของดีเอ็นเอในขณะที่โคลน ส่งผลให้เกิด gaps ไม่แท้จริงในลำดับเบส (Garber *et al*., 2009) ต่อมาเมื่อทราบลำดับเบสทั้งหมดแล้วจึงเปลี่ยนมาใช้พีซีอาร์ (PCR) เนื่องจากต้องทราบลำดับเบสที่อยู่ข้างๆ ของบริเวณที่ต้องการหาลำดับเบส เพื่อออกแบบไพร์เมอร์ที่ใช้ในปฏิกิริยาพีซีอาร์ ส่วนเทคโนโลยีเอ็นจีเอสไม่จำเป็นต้องทราบลำดับเบสก่อน อาศัยอะแด็บเตอร์ที่มีลำดับเบสจำเพาะกับเอ็นไซม์ตัดจำเพาะอยู่ที่ปลายด้าน 3′ เพื่อใช้เป็นไพร์เมอร์ในการเพิ่มปริมาณต้นแบบด้วยพีซีอาร์แบบ clonally amplified templates ซึ่งเป็นการเพิ่มปริมาณดีเอ็นเอที่ไม่ต้องใช้แบคทีเรีย วิธีการนี้จึงสามารถหาลำดับเบสโดยที่ไม่ต้องรู้ลำดับเบสมาก่อนได้พร้อมกัน นอกจากนี้ เทคนิคแซงเกอร์ยังไม่เหมาะสมกับการหาลำดับเบสของสิ่งมีชีวิตพวกยูคาริโอตที่มีจีโนมขนาดใหญ่ เช่น จีโนมมนุษย์ ซึ่งต้องใช้เวลาในการหาลำดับเบสนาน และในกระบวนการหาลำดับเบสหนึ่งครั้ง มักได้จำนวน reads ค่อนข้างน้อย ทำให้ต้นทุนค่าใช้จ่ายสูง รวมถึงการตรวจสอบการกลายที่เกิดกับเซลล์ร่างกาย (somatic mutation) ที่พบน้อยกว่าร้อยละ 1 ก็ทำได้ยากกว่าการใช้เทคโนโลยีเอ็นจีเอส (Woollard *et al*., 2011) เป็นต้น

บทความนี้ นำเสนอรูปแบบของเทคโนโลยีเอ็นจีเอสที่ใช้ปัจจุบัน การประยุกต์ใช้งานจากการใช้เทคโนโลยีเอ็นจีเอส และตัวอย่างโปรแกรมที่ใช้ในการจัดเรียงลำดับเบสดีเอ็นเอ รวมถึงโปรแกรมสำหรับใช้ในสร้างลำดับเบสดีเอ็นเอใหม่ ด้วยการนำลำดับเบส reads จำนวนมหาศาลที่เครื่องอ่านได้จากการใช้เทคโนโลยีเอ็นจีเอส มาต่อกันเป็นสายยาวขึ้น เพื่อนำลำดับเบสของดีเอ็นเอสายยาวที่ได้ไปใช้ประโยชน์ได้อย่างมีประสิทธิภาพ

## รูปแบบของเทคโนโลยีเอ็นจีเอส

เครื่องอ่านลำดับเบสแบบเอ็นจีเอสในปัจจุบัน (next-generation sequencing platforms) มี 3 เทคโนโลยีหลักๆ ที่ใช้กันอย่างกว้างขวาง หรือมักเรียกกันว่า Second generation NGS platforms ได้แก่ Roche/Gs-FLX 454 Genome sequencer, Illumina/Hiseq 2000 และ ABI/ SOLiD 5500xl สำหรับเทคโนโลยีอื่นๆ เช่น Polonator/G.007 และเทคโนโลยีที่กำลังพัฒนาต่อเนื่องกันอยู่ หรือที่เรียกว่า Third generation NGS platforms เช่น Helicos/Heliscope ซึ่งมีหลายบริษัทที่พยายามนำข้อบกพร่องที่เกิดขึ้นจาก Second generation NGS platforms มาปรับปรุงให้เทคโนโลยีการหาลำดับเบสมีความถูกต้องและราคาถูกลง สิ่งที่แตกต่างกันของเทคโนโลยีเหล่านี้ คือ การตรึงดีเอ็นเอต้นแบบให้อยู่กับที่ (template immobilization)

วิธีสังเคราะห์สายดีเอ็นเอ อัตราความผิดพลาด (error rate) และความยาวของ reads ที่ได้จากการอ่านลำดับเบสดัง Table 1 ซึ่งในที่นี้ จะขอกล่าวถึง Second generation NGS platforms เป็นหลัก สำหรับกระบวนการการทำงานของแต่ละเครื่องสรุปได้ดังนี้

**เครื่อง 454/Roche genome sequencer**

ใช้หลักการการหาลำดับเบสแบบการสังเคราะห์ที่รู้จักกันในนามของเทคนิค pyro sequencing ซึ่งเป็นการตรวจวัดไพโรฟอสเฟต (PPi) จากปฏิกิริยาเคมี (chemiluminescence) และใช้การอ่านสัญญาณตามจังหวะแสงของเบสคู่สมหลายๆ พันเบสในครั้งเดียว (Figure 1a) ขั้นตอนที่สำคัญของวิธีนี้คือ การตรึงยึด (immobilization) ดีเอ็นเอกับเม็ดบีด (bead) ในสภาวะที่เป็นน้ำมัน (emulsion PCR) และเพิ่มปริมาณสายดีเอ็นเอ ซึ่งมีขั้นตอนดังนี้

***1. การเตรียมดีเอ็นเอต้นแบบ (DNA library)***

เริ่มต้นด้วยการตัดดีเอ็นเอสายคู่ให้สั้นลงประมาณ 400 ถึง 600 คู่เบส จากนั้น เชื่อมอะแด็บเตอร์ A และ B ชนิดพิเศษที่ปลายของดีเอ็นเอนั้น โดยอะแด็บเตอร์ B จะมีส่วนปลาย 5′ ที่ประกอบด้วย biotin จะเข้าจับกับเม็ดบีดซึ่งถูกฉาบด้วย streptavidin ส่วนอะแด็บเตอร์ที่ไม่สามารถจับได้หรืออะแด็บเตอร์ที่จับกันเอง (adaptor dimers) จะถูกกำจัดออก อะแด็บเตอร์ที่ตรึงบนเม็ดบีดได้บางสาย จะมีลำดับเบสโอลิโกนิวคลีโอไทด์ที่ไม่มีการ phosphorylate ทำให้มีช่องว่างในสายดีเอ็นเอคู่นั้น จากนั้นเอ็นไซม์ที่มีคุณสมบัติเป็น strand-displacing DNA polymerase จะเข้าซ่อมแซมและเติมส่วนที่ขาดไป ทำให้ได้เป็นดีเอ็นเอสายคู่ที่สมบูรณ์ หลังจากนั้นเส้นดีเอ็นเอสายคู่ที่ไม่มีการตรึงติดกับเม็ดบีดถูกแยกออกเป็นสายเดี่ยว

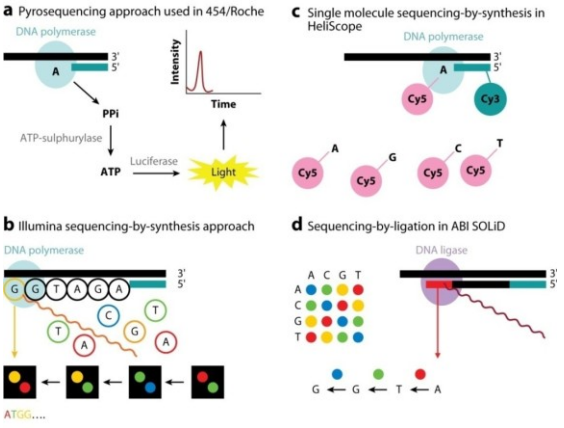


**Figure 1** Sequencing chemistry implemented in next-generation sequencers. (a) Pyrosequencing approach implemented in 454/Roche. (b) Sequencing by synthesis in the presence of four fluorescently labeled nucleotide analogs that serve as reversible reaction terminators and special DNA polymerases in the Illumina sequencing reaction. (c) Helicos single-molecule sequencing utilizes sequencing-by-synthesis methodology. Cy3-labeled DNA templates bound to immobilized primers on the surface of the flow cell. The Cy5-labeled nucleotides are added to the reaction one at a time, and the detection of incorporated nucleotides is achieved. (d) Sequencing by ligation approach implemented in the SOLiD method. Two-base encoding implemented within the system and final data is reported as standard base calls. (Image courtesy of Morozova *et al.*, 2009).

โดยมีสายดีเอ็นเออีกเส้นหนึ่งที่มีอะแด็บเตอร์ B/B ที่ยังคง ติดกับเม็ดบีดด้วยพันธะระหว่าง streptavidin และ biotin ส่วนสายดีเอ็นเอบางส่วนที่ไม่มีอะแด็บเตอร์ B ติดอยู่จะถูกกำจัดออก สุดท้ายจะได้ดีเอ็นเอสายเดี่ยว ซึ่งจะมีส่วนของอะแด็บเตอร์

**Table 1** Comparing performances and features of Second-generation sequencing platforms; 454/Roche, Illumina, SOLiD, Polonator and Third-generation sequencing platforms of Helicos including single molecular sequencing platform under development of Pacific Biosciences

| NGS technology | 454/Roche | Illumina/Solexa | SOLiD/ABI | Polonator/G.007 | Helicos | SMRT (Pacific Biosciences) |
|---|---|---|---|---|---|---|
| Chemistry | Pyrosequencing | Polymerase-based | Ligation-based | Ligation-base | Reversible Dye terminators | Phospho-linked Fluorescent Nucleotides |
| Amplification | Emulsion PCR | Bridge Amp | Emulsion PCR | Emulsion PCR | No (single molecule) | No (single molecule) |
| Sequence by synthesis | Pyrosequencing | Reversible Dye terminators | Oligonucleotide probe ligation | Sequencing by ligation using a random arrayed, bead-based, emulsion PCR | Single molecule sequencing | Single molecule real time |
| Paired ends/sep. | Yes/3 kb | Yes/200 bp | Yes/3 kb | Yes/13 bp | 25-55 bp | NA |
| Data production/day | 400 Mb/run/7.5 hr | 3,000 Mb/run/6.5 days | 4,000 Mb/run/6 days | ~16 Gb/run/2.5 days | 8 days | 0.02 days |
| Sequencing/run | 10 hours | 2-5 days | 6 days | ~80 hours | 12 | < 1 |
| Raw accuracy | 99.5% | >98.5% | 99.94% | >98% | >99% | NA |
| Read length | 400 bp | 100 bp | 50 bp | 26 bp | 35 average length | Longer than 1000 |
| Cost per run (total) | $8439 | $8950 | $17447 | $3500 | Lower than second NGS | Lower than second NGS |

A ติดที่ปลาย 5′ และอะแด็บเตอร์ B ติดที่ปลาย 3′ ในขั้นตอนนี้ จากจีโนมิกดีเอ็นเอจะได้เป็นดีเอ็นเอสายเดี่ยว (sstDNA library) (Figure 2) หลังจากนั้นจึงประเมินคุณภาพของดีเอ็นเอต้นแบบที่ได้ด้วยเครื่องAgilent 2100 BioAnalyzer (www.my454.com) เช่น ตรวจสอบความยาวของสายดีเอ็นเอ หากเป็นดีเอ็นเอที่มีน้ำหนักโมเลกุลสูง ควรตรวจสอบว่ามีความยาวอยู่ในช่วง 500 ถึง 800 คู่เบส หรือหากเป็นดีเอ็นเอที่มีน้ำหนักโมเลกุลต่ำ ควรมีความยาว 150 ถึง 800 คู่เบส ปริมาณดีเอ็นเอที่ได้มากกว่าหรือเท่ากับ 5 นาโนกรัม และควรมีส่วนของอะแด็บเตอร์ที่จับกันเองน้อยกว่า 5% เป็นต้น

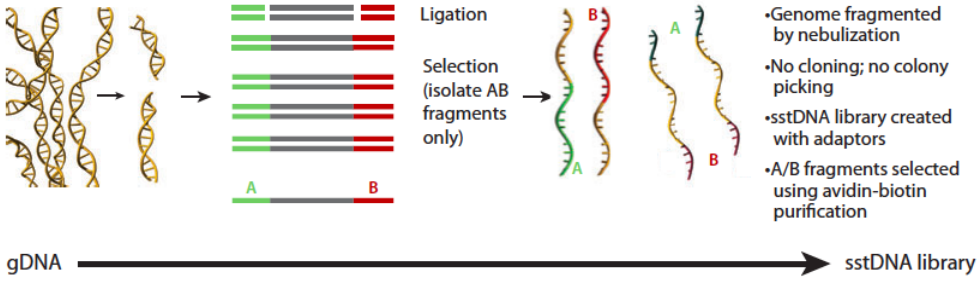


**Figure 2** DNA library preparation to generate single-stranded fragments (sstDNA) using A and B adaptors. (Image courtesy of Mardis, 2008).

***2. การเพิ่มปริมาณดีเอ็นเอบนเม็ดบีด (emulsion-based clonal amplification; emPCR)***

นำดีเอ็นเอสายเดี่ยว จากข้อ 1 ตรึงบนเม็ดบีดที่มีไพร์เมอร์โอลิโกนิวคลีโอไทด์ติดอยู่ ซึ่งสามารถเข้าคู่ได้กับอะแด็บเตอร์ของดีเอ็นเอสายเดี่ยวนั้น (Figure 3) โดยดีเอ็นเอที่ต่อกับอะแด็บเตอร์นับเป็นหนึ่งโมเลกุล ซึ่งหนึ่งโมเลกุลนี้จะจับอยู่กับเม็ดบีดหนึ่งเม็ด เพื่อเพิ่มปริมาณดีเอ็นเอในหยดน้ำมันที่มีสารสำหรับทำพีซีอาร์อยู่ภายใน (oil-aqueous emulsion) เมื่อเสร็จสิ้นปฏิกิริยาพีซีอาร์แล้ว ในหนึ่งหยดน้ำมันจะมีดีเอ็นเอต้นแบบที่เหมือนกันเป็นล้านชุดติดอยู่ที่เม็ดบีด ซึ่งจะนำไปหาลำดับเบสต่อไป ส่วนเม็ดบีดที่ไม่มีดีเอ็นเอติดอยู่จะถูกกำจัดออก อย่างไรก็ตาม สิ่งที่ควรทำก่อนการเพิ่มปริมาณดีเอ็นเอบนเม็ดบีดคือ การคำนวณจำนวนของดีเอ็นเอต้นแบบที่เหมาะสม เพื่อนำมาใช้ในขั้นตอนนี้ เนื่องจากถ้าใส่ปริมาณดีเอ็นเอน้อยเกินไป แต่เม็ดบีดมีจำนวนมาก ก็อาจไม่มีการเพิ่มปริมาณดีเอ็นเอบนเม็ดบีดเกิดขึ้น ทำให้ไม่สามารถนำไปหาลำดับเบสต่อได้ หรือถ้าใส่ดีเอ็นเอปริมาณมากเกินไป บางเม็ดบีดอาจมีจำนวนดีเอ็นเอต้นแบบที่ตรึงอยู่มากเกิน จนสุดท้ายอาจได้ reads ที่ใช้งานไม่ได้ โปรแกรมของเครื่องจึงต้องกรอง reads เหล่านี้ออก การคำนวณปริมาณดีเอ็นเอที่เหมาะสมทำได้สองแบบ ได้แก่ การวิเคราะห์ด้วยวิธี emulsion titration เป็นการทดสอบที่ราคาถูกและทำได้ค่อนข้างเร็ว และแบบ sequencing titration มีค่าใช้จ่ายแพง แต่ทำได้สมบูรณ์ตั้งแต่คำนวณจำนวนดีเอ็นเอที่จะใช้เพิ่มปริมาณบนเม็ดบีด รวมถึงปฏิกิริยาต่างๆ ที่ควรจะใช้ในการหาลำดับเบสเพื่อให้ได้ reads ที่ดี ซึ่งการเลือกใช้ขึ้นอยู่กับผู้ใช้งานเป็นหลัก (www.my454.com)

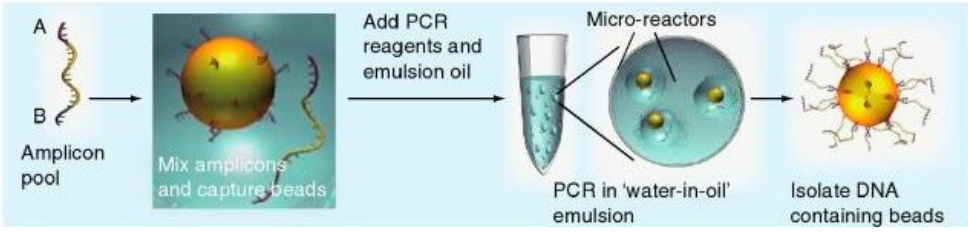


**Figure 3** Pooled amplicons immobilized on magnetic beads are clonally amplified in droplet emulsions (emulsion PCR). (Image courtesy of Gega and Kozal, 2011).

***3. การหาลำดับเบส***

454/Roche ใช้วิธีที่เรียกว่า sequencing by synthesis เพื่อหาลำดับเบสดีเอ็นเอ (Figure 4a และ 4b) โดยใช้เอ็นไซม์ DNA polymerase ในการสังเคราะห์สายดีเอ็นเอจากดีเอ็นเอสายเดี่ยวให้เป็นดีเอ็นเอสายคู่ เริ่มต้นด้วยการนำดีเอ็นเอที่ติดด้วยเม็ดบีดจากข้อ 2 ใส่ลงใน PicoTiterPlate$^{tm}$ (PTP) ซึ่ง PTP ประกอบด้วยหลุม (well) จำนวน 1.6 ล้านหลุม โดยหนึ่งหลุมสามารถบรรจุปริมาตรทั้งหมดได้ 75 picoliters และบรรจุได้เพียงเม็ดบีดเดียวเท่านั้น ดังนั้น หนึ่งเม็ดบีดจึงเทียบได้กับหนึ่ง reads หลังจากการบรรจุเม็ดบีดข้างต้นแล้ว ใส่ PTP เข้าเครื่อง 454/Roche ซึ่งเครื่องจะเติมสารละลายต่างๆ รวมถึงเบส A, T, C, G แบบอัตโนมัติ โดยเติมนิวคลีโอไทด์เข้าไปทีละตัว ตัวที่เข้าคู่กับดีเอ็นเอต้นแบบเท่านั้นจึงจะจับและเกิดปฏิกิริยา ในขณะที่นิวคลีโอไทด์ตัวอื่นก็จะถูกล้างออกไป จากนั้น เมื่อมีการต่อสายดีเอ็นเอด้วยเอ็นไซม์ DNA polymerase จะเกิดการเปลี่ยนแปลงทางปฏิกิริยาเคมี ซึ่งจะมีการปลดปล่อยสารไพโรฟอสเฟต (PPi) และเอ็นไซม์ sulfurylase จะทำหน้าที่เปลี่ยนแปลงสารไพโรฟอสเฟต ด้วยการทำปฏิกิริยากับ adenosine 5′-phosphosulfate (APS) ได้พลังงาน ATP ออกมา จากนั้น เอ็นไซม์ luciferease จะใช้พลังงาน ATP เป็นสารตั้งต้นและใช้ luciferin ได้ผลิตภัณฑ์เป็นสาร oxyluciferin ส่งผลให้เกิดการปลดปล่อยแสงเป็นสัญญาณออกมา แสงจะถูกบันทึกด้วย CCD camera โดยสัญญาณแสงที่ปลดปล่อยออกมา จะบ่งชี้ถึงสัดส่วนของจำนวนนิวคลีโอไทด์ที่อ่านได้ ซึ่งโปรแกรมจะสร้างกราฟ แสดงความเข้มข้นของแสงที่อ่านได้ในแต่ละนิวคลีโอไทด์ เรียกว่า Flow gram ของแต่ละหลุมที่อยู่ใน PTP นั้น

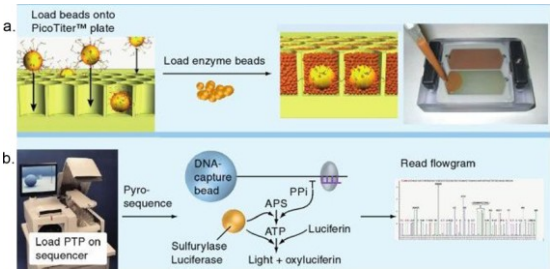


**Figure 4** Sequencing DNA on a 454/Roche sequencer. (a) Depositing DNA beads into the PicoTiterPlate$^{tm}$ (PTP). (b) Placing the PTP into the 454/Roche instrument and starting the pyrosequencing enzymatic process. The signals created are then analysed by the 454 sequencing system's software to generate bar graph of light intensities called a "Flow gram" for each well on the PTP. (Image courtesy of Gega and Kozal, 2011).

อย่างไรก็ตาม เครื่องนี้สามารถอ่านความยาว reads ในช่วง 100 ถึง 600 คู่เบส และในการหาลำดับเบสหนึ่งครั้ง (ครอบคลุมตั้งแต่ขั้นตอนการ

เตรียมดีเอ็นเอต้นแบบจนถึงขั้นการหาลำดับเบส) จะได้จำนวนของเบส 400 ถึง 600 เมกะเบส ความถูกต้องของเบสที่ได้มากกว่าร้อยละ 99

**เครื่อง Illumina**

ใช้หลักการหาลำดับเบสโดยใช้การติดฉลากเบสด้วยสารเรืองแสง (fluorescent reversible terminators) ของทั้งสี่เบส (Figure 1b) และเพิ่มปริมาณของสายดีเอ็นเอเชื่อมต่อเป็นสะพานบนสถานะของแข็ง (solid-phase bridge amplification) ได้กลุ่มของสายดีเอ็นเอที่แตกต่างกัน สำหรับรายละเอียดมีดังนี้

***1. การเตรียมดีเอ็นเอต้นแบบ***

ตัวอย่างดีเอ็นเอที่สนใจจะถูกตัดด้วย nebulizer (Figure 5) (www.illumina.com) โดยอาศัยเทคนิค nebulization ซึ่งเป็นวิธีที่ง่ายและรวดเร็ว ใช้จำนวนดีเอ็นเอค่อนข้างน้อยประมาณ 0.5 ถึง 5 มิลลิกรัม เมื่อสารละลายที่มีดีเอ็นเอผ่านเข้าไปภายในด้วยความเร็วระดับหนึ่ง ขนาดของดีเอ็นเอจะถูกควบคุม ด้วยการเปลี่ยนแปลงแรงดันของก๊าซที่เข้ามาภายใน nebulizer รวมถึงปัจจัยอื่นๆ เช่น ความหนืดของสารละลายและอุณหภูมิ ทำให้ได้สายดีเอ็นเอที่สั้นลง (Sambrook and Russell, 2006) หลังจากการตัดจะได้สายดีเอ็นเอความยาวน้อยกว่า 800 คู่เบส จากนั้น ส่วนปลายที่ถูกทำลายจะถูกซ่อมแซมด้วยเอ็นไซม์ T4 DNA polymerase, Klenow enzyme และ T4 poly nucleotide kinase (Son and Taylor, 2011) และเติมนิวคลีโอไทด์เบส A ที่ปลาย 3′ ของสายดีเอ็นเอนั้น เพื่อเพิ่มประสิทธิภาพในช่วงของการเชื่อม จากนั้นดีเอ็นเอจะถูกเชื่อมด้วยอะแด็บเตอร์ ซึ่งมีขนาดความยาวประมาณ 66 คู่เบส และติดเบส T ที่ปลายทั้งสองข้าง หลังจากนั้น คัดแยกสายดีเอ็นเอความยาวประมาณ 150 ถึง 200 คู่เบส บนเจล และเพิ่มปริมาณสายดีเอ็นเอด้วยเทคนิคพีซีอาร์


**Figure 5** DNA library preparation on Illumina. Genomic DNA is fragmented by nebulization and adaptors are ligated to both ends of the fragments to create a library. (Image courtesy of Ansorge, 2009).

***2. การสร้างกลุ่มของสายดีเอ็นเอด้วยวิธี bridge amplification***

หลังจากขั้นตอนการแยกสายดีเอ็นเอ นำดีเอ็นเอสายเดี่ยวใส่ลงบนแผ่นกระจกสไลด์ (flow cell channels) แบบสุ่ม โดยผิวของแผ่นกระจกสไลด์ถูกฉาบด้วยตัวอะแด็บเตอร์ และอะแด็บเตอร์ที่เป็นคู่สมกัน (complementary adapters) ซึ่งทำหน้าที่เป็นเหมือนไพร์เมอร์ใช้ในช่วงของการเพิ่ม

ปริมาณดีเอ็นเ อดด้วยเทคนิคพีซีอาร์ (Figure 6a) จากนั้นเติมนิวคลีโอไทด์และเอนไซม์ เพื่อเริ่มต้นการเพิ่มปริมาณแบบสะพาน (bridge amplification) จากดีเอ็นเอสายเดี่ยวหนึ่งโมเลกุลจะจับกับไพร์เมอร์เป็นรูปสะพานโค้ง (double-stranded bridges) (Figure 6b) และในช่วงขั้นตอนการแยกสายดีเอ็นเอ จะได้ดีเอ็นเอสายเดี่ยว เพื่อใช้เป็นดีเอ็นเอต้นแบบอีกครั้ง หลังจากการเพิ่มปริมาณด้วยเทคนิคพีซีอาร์เสร็จสิ้น จะได้กลุ่มของสายดีเอ็นเอมากกว่า 50 ล้านกลุ่ม โดยแต่ละกลุ่มประกอบด้วยจำนวนดีเอ็นเอประมาณ 1,000 ชุดในแต่ละช่อง (Figure 6c)

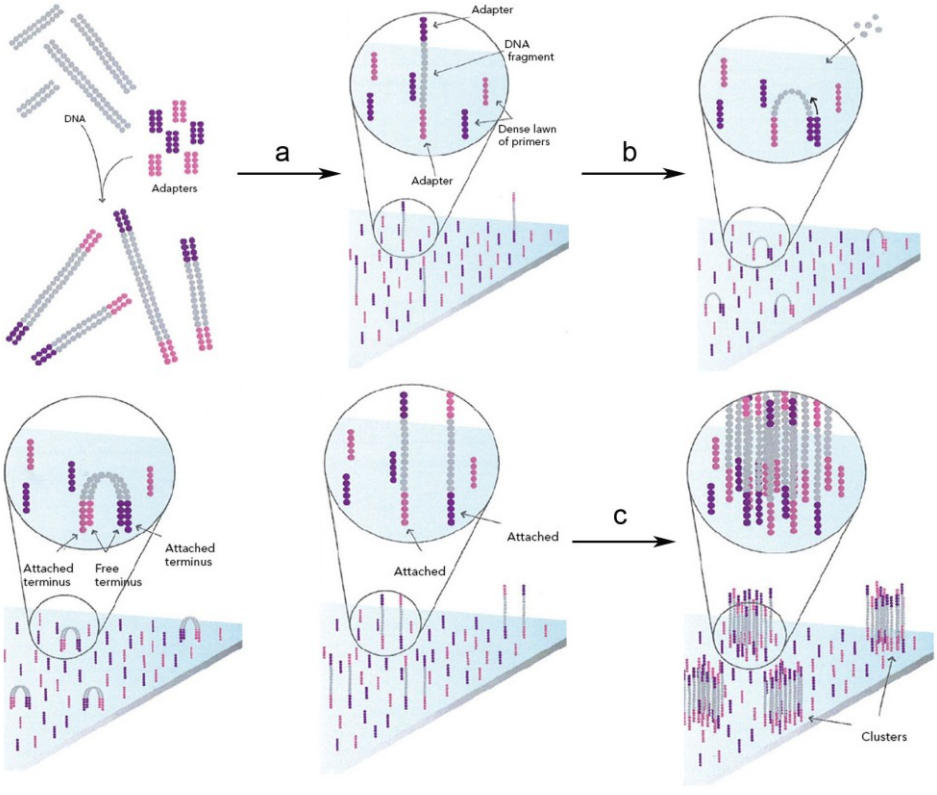


**Figure 6** Cluster generation by bridge amplification of Illumina. (a) Randomly fragmented genomic DNA is ligated to adaptors. Single-stranded fragments are bound to the inside of the flow cell channels. (b) Unlabelled nucleotides and enzyme are added to initiate solid-phase bridge amplification. After that double-stranded templates are generated. (c) The double-stranded templates are denatured to form single-stranded templates. Denaturation and extension are repeated, resulting in localized amplification of single molecules in millions of unique locations across the flow cell surface. (Image courtesy of www.illumina.com).

***3. การหาลำดับเบส***

Illumina ใช้หลักการ sequencing by synthesis (Figure 7a) โดยแยกสายดีเอ็นเอในแต่ละกลุ่มเป็นดีเอ็นเอสายเดี่ยว ให้ไพร์เมอร์สำหรับหาลำดับเบสจับกับดีเอ็นเอที่จำเพาะนั้นๆ ในช่วงเริ่มต้นของการหาลำดับเบส มีการเติมเอ็นไซม์ DNA polymerase และนิวคลีโอไทด์ทั้งสี่เบสที่ติดด้วยสารเรืองแสงที่มีสีต่างกันและใช้ reversible terminator ด้วยการ block ปลาย 3′OH ไว้ เพื่อหยุดการสังเคราะห์สายดีเอ็นเอของเอ็นไซม์ DNA polymerase ในขณะที่มีการหยุดการสังเคราะห์สายดีเอ็นเอ นิวคลีโอไทด์ที่ไม่ได้เข้าคู่กับสายดีเอ็นเอตั้งต้นที่เหลือในปฏิกิริยาจะถูกล้างออก หลังจากการกระตุ้นด้วยแสงเลเซอร์ จะมีการบันทึกภาพการปลดปล่อยสารเรืองแสงของนิวคลีโอไทด์จากแต่ละกลุ่มบนแผ่นกระจกสไลด์ที่ได้ เพื่อบันทึกความเหมือนกันของเบสแรกสำหรับกลุ่มนั้นๆ ตามด้วยขั้นตอนการตัด (cleavage) เพื่อกำจัด terminator และ fluorescent dye ออกไป ล้างอีกครั้ง จากนั้นเติมนิวคลีโอไทด์ชุดใหม่ และเอ็นไซม์ DNA polymerase เพื่อสังเคราะห์สายดีเอ็นเอต่อไป ทำซ้ำแบบนี้ไปเรื่อยๆ โดยแต่ละรอบ (cycle) จะเป็นตัวกำหนดลำดับเบสที่อ่านได้ในแต่ละครั้ง (Figure 7b)

เครื่อง Illumina นี้สามารถอ่านความยาว reads ได้ถึง 100 คู่เบส และในการหาลำดับเบสหนึ่งครั้ง จะได้จำนวนของเบสมากถึง 600 กิกะเบส ความถูกต้องของเบสที่ได้มากกว่าร้อยละ 99.5

**เครื่อง SOLiD/ABI**

ใช้หลักการหาลำดับเบสแบบการเชื่อมต่อ (sequencing by ligation) ซึ่งใช้วิธีการคล้ายกับ 454/Roche ด้วยการนำดีเอ็นเอติดกับเม็ดบีด (magnetic bead) ในสภาวะที่เป็นน้ำมัน และเพิ่มจำนวนของดีเอ็นเอไปพร้อมๆกัน วิธีนี้ได้คุณภาพของข้อมูลที่ดี (Figure 1d) แต่วิธีการเตรียมดีเอ็นเอต้นแบบก่อนหาลำดับเบสค่อนข้างยุ่งยากและใช้เวลานาน โดยมีขั้นตอนดังนี้

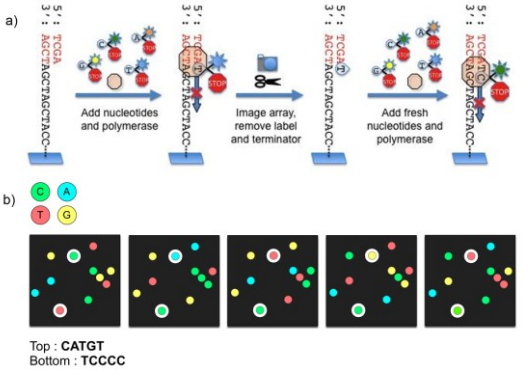


**Figure 7** Sequencing by synthesis. (a) Illumina sequencing chemistry (Image courtesy of Anderson and Schrijver, 2010). (b) The sequencing cycles are repeated to determine the sequence of bases in a given fragment a single base at a time.

***1. การเตรียมดีเอ็นเอต้นแบบ***

ในขั้นตอนการสร้างดีเอ็นเอต้นแบบสามารถทำได้สองวิธีด้วยกัน ได้แก่ sequencing-fragment และ mate-paired library การเลือกใช้ขึ้นอยู่กับวัตถุประสงค์ของงานวิจัยที่ผู้วิจัยสนใจ โดยเชื่อมต่ออะแด็บเตอร์เข้ากับดีเอ็นเอต้นแบบและติดเข้ากับเม็ดบีด (Figure 8)

***2. การเพิ่มปริมาณดีเอ็นเอต้นแบบบนเม็ดบีด (emulsion-based clonal amplification; emPCR)***

เตรียม microreactor ภายในประกอบด้วยดีเอ็นเอต้นแบบ สารละลายบัฟเฟอร์ต่างๆ เม็ดบีด

และไพร์เมอร์ จากนั้น เพิ่มปริมาณดีเอ็นเอบนเม็ดบีดด้วยเทคนิคพีซีอาร์ แยกสายดีเอ็นเอและคัดแยกเม็ดบีดที่ไม่มีการเพิ่มปริมาณออก จากนั้น สร้างพันธะโควาเลนต์ที่ส่วนปลาย 3′ ของดีเอ็นเอต้นแบบที่ติดบนเม็ดบีดและตรึงบนสไลด์ (Figure 8)

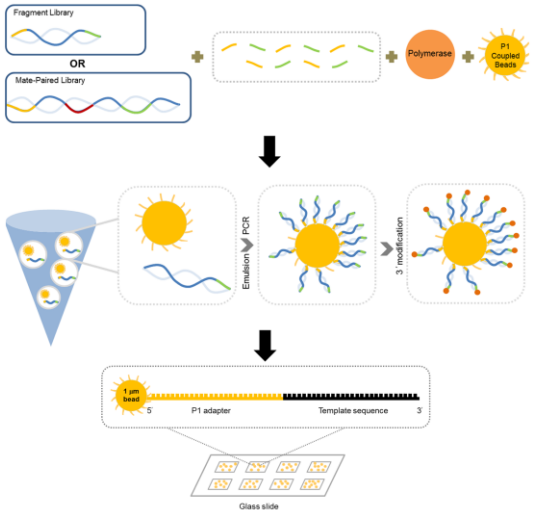


**Figure 8** DNA library preparation. Clonal bead library generation via emulsion PCR and depositing beads into flow cell. (Redrawn from http://www.appliedbiosystems.com).

### *3. การหาลำดับเบสด้วยวิธี sequencing by ligation*

วิธีการหาลำดับเบสของ SOLiD/ABI เป็นการอ่านตำแหน่งของเบสบนดีเอ็นเอต้นแบบซ้ำกันสองครั้ง หรือที่เรียกว่า Di-base sequencing (Figure 9) โดยใช้การเชื่อมต่อด้วยโอลิโกนิวคลีโอไทด์ที่ติดสี (dye-labeled oligonucleotide) ด้วยหลักการ two-base encoding หรือ di-base probe เริ่มด้วยไพร์เมอร์ จะเข้าจับกับส่วนของอะแด็บเตอร์ที่มีลำดับเบสที่เข้าคู่กัน บนดีเอ็นเอต้นแบบที่ได้เพิ่มปริมาณแล้ว จากนั้นเติม di-base probe เข้าไป ซึ่งการเชื่อมต่อสามารถทำได้สองทิศทางจากปลาย 5′-$PO_4$ หรือปลาย 3′-OH หลังจากนั้นเอนไซม์ ligase จะเชื่อม probe ที่ประกอบด้วย 8 เมอร์ ด้วยนิวคลีโอไทด์ที่ 1 และ 2 ซึ่งเป็นนิวคลีโอไทด์ที่เข้าคู่กับดีเอ็นเอต้นแบบนั้นตามด้วยการบันทึกภาพของสี fluorescence ที่ได้ ซึ่งการบันทึกจะเกิดขึ้นในแต่ละรอบ หลังจากขั้นตอนการเชื่อมก่อนการตัดสามเบสสุดท้ายของ probe ส่วนของ probe ที่ไม่สามารถเข้าคู่กับดีเอ็นเอต้นแบบได้อย่างจำเพาะ จะถูกกำจัดออก เอนไซม์ phosphatase จะช่วยป้องกันสายดีเอ็นเอที่ไม่สามารถ extend ได้ในช่วงของการเชื่อม (ligation cycle) ออกไป หลังจากนั้นเมื่อ probe เข้าเชื่อมกับดีเอ็นเอต้นแบบแล้ว จะเกิดปฏิกิริยาเคมีตัดที่ตำแหน่งของ probe ที่เบส 5 และ 6 รวมถึงกลุ่มของ fluorescent ด้วย silver ions ทำให้เกิดกลุ่มของ 5′-$PO_4$ และเข้าสู่การเชื่อมรอบที่สองต่อไป ดังนั้นจำนวนรอบของการเชื่อม การบันทึกภาพ และการตัด จะทำซ้ำๆ กัน จนได้ความยาวของ reads ที่ต้องการ ในช่วงของ ligation cycle สายดีเอ็นเอที่ได้เชื่อมจนครบปฏิกิริยาแล้ว จะถูกแยกออกจากอะแด็บเตอร์หรือดีเอ็นเอต้นแบบ และรอบที่สองของการหาลำดับเบสเกิดขึ้นด้วยไพร์เมอร์เดิม ณ ตำแหน่งที่ n-1 จะเข้าจับกับดีเอ็นเอต้นแบบเดิมที่เข้าคู่กันได้ จำนวนรอบของการเชื่อม บันทึกภาพ และการตัด เกิดขึ้นซ้ำๆ ประมาณ 7 รอบ จะเห็นได้ว่าในหนึ่งไพร์เมอร์ จะเกิดปฏิกิริยาการเชื่อมประมาณ 7 รอบ ดังนั้น ในห้าไพร์เมอร์ที่มีกระบวนการหาลำดับเบสข้างต้นเกิดขึ้นซ้ำๆ กัน จะมีการสร้างชุดเบสประมาณ 35 เบสที่ต่อเชื่อมกัน ซึ่งประกอบด้วยนิวคลีโอไทด์สองเบส ณ ตำแหน่งต่างๆ กันที่ให้สีต่างกันไป (Figure 9h) สำหรับหลักการการถอดรหัสลำดับเบสของดีเอ็นเอต้นแบบต้องทราบนิวคลีโอไทด์แรกของอะแด็บเตอร์ก่อนและลำดับเบสของสีต่างๆ ที่บันทึกได้ในส่วนที่เหลือจะมีการแปลรหัสสีเป็นเบส (Figure 10) โดยจะมีตารางสีที่บอกถึงความเป็นไปได้ของไดนิวคลีโอไทด์ทั้ง 16 แบบที่ให้สีต่างกัน ตัวอย่างเช่น ถ้านิวคลีโอ

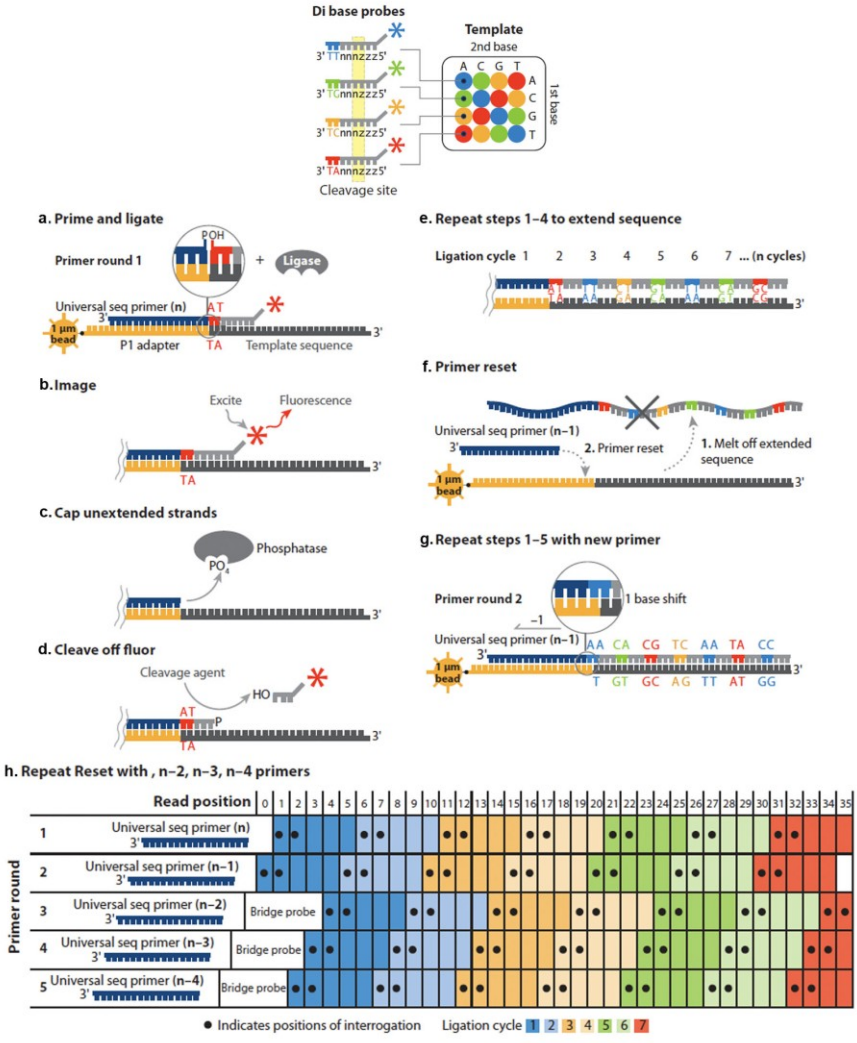


**Figure 9** SOLiD/ABI system – Sequencing by ligation using di-base labeled probes. (Image courtesy of Mardis, 2008).

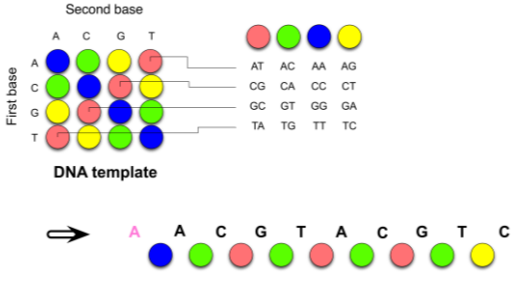


**Figure 10** SOLiD Color Space Code and two-base encoding.

ไทด์ตัวแรกของลำดับเบสอะแด็บเตอร์เป็นเบส A (สีชมพู) และสีที่บันทึกได้สีแรกเป็นสีฟ้าแล้ว เบสที่ได้เมื่อเทียบในตารางสี ความเป็นไปได้ของไดนิวคลีโอไทด์ 16 แบบนั้น จะได้เป็นเบส A ต่อมาพบสีที่บันทึกได้สีที่สองเป็นสีเขียว ดังนั้น จากเบสแรกที่พบเบส A เมื่อเทียบในตาราง เบสที่สองจึงควรเป็นเบส C ต่อมาพบสีที่บันทึกได้เป็นสีแดง จากเบสก่อนหน้านี้เป็นเบส C เบสต่อไปจึงได้เป็นเบส G จากนั้น สีที่บันทึกได้เป็นสีเขียว เมื่อเทียบเบส G ที่

ได้ก่อนหน้านี้ในตารางถอดรหัสสี เบสต่อไปจึงได้เป็นเบส T เป็นต้น ซึ่งเครื่อง SOLiD จะมีโปรแกรมสำหรับคำนวณจากสีเป็นลำดับเบสให้แบบอัตโนมัติ

สำหรับ di-base probe เช่น 3′TTnnnzzz5′, 3′TGnnnzzz5′, 3′TCnnnzzz5′ หรือ 3′TAnnnzzz5′ นั้น ประกอบด้วย เบสแรกและเบสที่สอง เป็น interrogation base ซึ่งทำหน้าที่เข้าเชื่อมหรือ hybridize กับไพร์เมอร์ โดยไดนิวคลีโอไทด์สามารถเป็นได้ถึง 16 แบบที่สามารถถอดรหัสสีของ fluorescence ได้ทั้งสี่สี ส่วนของเบส n และ z ทำหน้าที่เป็น degenerate base สำหรับเบส z หรือ inosine base จะทำหน้าที่ลดความซับซ้อนของ probe library และเชื่อมต่อเบสด้วย phosphorothiolate linkage ระหว่างตำแหน่งเบสที่ 5 และ 6 ซึ่งจะถูกตัดด้วย silver ions

เครื่องมือนี้ อ่านความยาวของ reads ที่ 50 คู่เบส และในการหาลำดับเบสหนึ่งครั้ง จะได้จำนวนของเบส 80 ถึง 100 กิกะเบส ความถูกต้องของเบสที่ได้มากกว่าร้อยละ 99.94

เทคโนโลยีอีกหนึ่งรูปแบบที่จัดอยู่ใน Second generation NGS platform ได้แก่เครื่อง Polonator ใช้วิธีการหาลำดับเบสด้วยวิธีการเชื่อมต่อ เริ่มต้นการเตรียมตัวอย่างดีเอ็นเอด้วยการแบ่งเป็นชิ้นส่วนย่อยๆ ด้วยการใช้แรงดันสูงของน้ำ (hydroshear) จากนั้น จีโนมิกดีเอ็นเอที่ถูกตัดแล้ว (Figure 11a) จะถูกซ่อมที่ส่วนปลายและเติม A-tailed และทำให้เป็นวงกลมด้วย T-tailed ที่สังเคราะห์ขึ้นความยาว 30 คู่เบส (Figure 11b) จากนั้นเพิ่มปริมาณ circular DNA ด้วยปฏิกิริยา rolling circle amplification และย่อยด้วยเอนไซม์ *Mme1* ทำให้ได้จีโนมิกดีเอ็นเอความยาว 70 คู่เบสที่ประกอบด้วยส่วนของ T-tailed และ tags ประมาณ 17 ถึง 18 คู่เบส เลือกขนาดของ paired inset libraries ที่ได้ข้างต้นที่มีความยาว 70 คู่เบส แล้วเชื่อมเข้ากับ linker ที่ปลาย 3′ และ 5′ ทำให้ library molecule ที่ได้อยู่ที่ 135 คู่เบส และเพิ่มปริมาณของสายดีเอ็นเอบนเม็ดบีด ในสภาวะที่เป็นน้ำมัน (Figure 11c) และขั้นตอนสุดท้าย ใส่เม็ดบีดดังกล่าวภายในหลุม (flow cell) และอ่านลำดับเบสของดีเอ็นเอนั้นด้วยเครื่อง Polonator (Figure 11d) โดยในช่วงของการหาลำดับเบส ไพร์เมอร์จะถูกเติมเข้าไปในแต่ละหลุม (cell) ไพร์เมอร์จะจับกับลำดับเบสที่ปลาย 3′ หรือ 5′ ของ tags ในสายดีเอ็นเอที่มีความยาว 17 ถึง 18 คู่เบส ขณะที่ไพรเมอร์จับนี้ ไพร์เมอร์ที่มีลักษณะเป็นส่วนผสมของ degenerate nonanucleotides หรือ nonamers ซึ่งประกอบด้วยเบสผสมของสี่เบสและมีหนึ่งเบสที่ติดสารเรืองแสง fluorophore เช่น 5′Cy5-NNNNNNNNT, 5′Cy3-NNNNNNNNA, 5′TexasRed-NNNNNNNNC, 5′6FAM-NNNNNNNNG และ เอนไซม์ T4 DNA ligase จะถูกเติมเข้าไป ทำให้เกิดการเชื่อมและให้แสงสัญญาณ ซึ่งจำแนกเบสที่เข้าคู่กันบนสายดีเอ็นเอดังกล่าว (Figure 11e) แต่ละเม็ดบีดบนอาร์เรย์ (array) จะให้แสง fluorescent เพียงหนึ่งสี ซึ่งจะบ่งชี้ถึงเบส A, C, G หรือ T ที่ตำแหน่งบนสายดีเอ็นเอนั้น หลังจากการบันทึกภาพ Polonator จะล้างอาร์เรย์ที่มีไพร์เมอร์จับกับสายดีเอ็นเอและเอนไซม์ที่เหลือในปฏิกิริยาด้วยการใช้ guanidine HCl และ sodium hydroxide ซึ่งในแต่ละรอบของเบส reads ที่ได้ จะผ่านขั้นตอนของการล้างส่วนของ primer-fluorescent probe complex ที่จับกันแล้วออกไป ไพร์เมอร์จะถูกเติมใหม่ และได้ส่วนผสมใหม่ของ fluorescently tagged nonamers เช่น 5′Cy5-NNNNNNNTN, 5′Cy3-NNNNNNNAN, 5′TexasRed-NNNNNNNCN, 5′6FAM-NNNNNNNGN ในขณะที่ตำแหน่งบนสายดีเอ็นเอจะถูก shift มา 1 เบส ซึ่งวิธีนี้ใช้หลักการเตรียมดีเอ็นเอต้นแบบเริ่มต้นคล้ายกับเครื่อง 454/Roche และใช้วิธีการ sequencing

by ligation ในกระบวนการหาลำดับเบสคล้ายกับเครื่อง SOLiD/ABI แต่ความแตกต่างกันเบื้องต้นระหว่าง Polonator และ SOLiD คือ ราคาของอุปกรณ์ที่ต่ำกว่า และโปรแกรมต่างๆ รวมถึงการวิเคราะห์เป็นลักษณะของ open source มากกว่า (Anderson and Schrijver, 2010) วิธีนี้สามารถอ่าน reads ที่ความยาว 26 คู่เบส และในการหาลำดับเบสหนึ่งครั้ง จะได้จำนวนเบส 8 ถึง 10 กิกะเบส แต่วิธีการนี้ค่อนข้างให้ความถูกต้องของ reads ที่ต่ำกว่าวิธีอื่นๆ

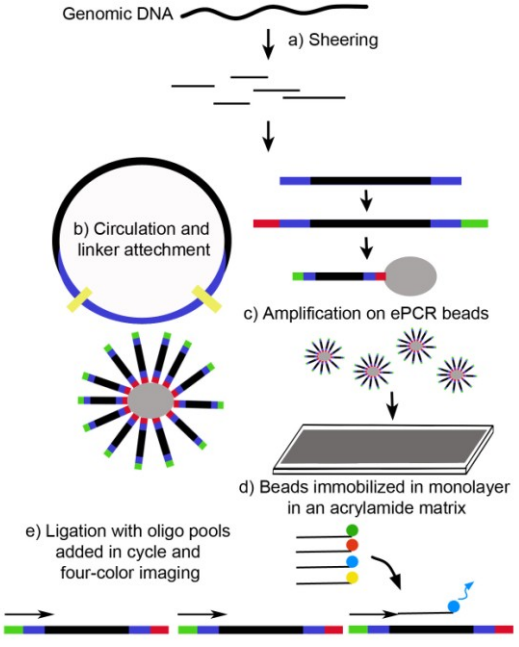


**Figure 11** Polony-based sequencing by ligation. (Redrawn from http://www.azcobiotech.com/polonator/pol-technology.php).

นอกจากนี้ ตัวอย่างของ Third-generation NGS platform เช่น เครื่อง Helicos เป็นเครื่องมือรุ่นแรกที่เริ่มใช้เทคโนโลยี single molecular sequencing ที่สามารถหาลำดับเบสของดีเอ็นเอโดยไม่จำเป็นต้องเพิ่มปริมาณนิวคลีโอไทด์ (Figure 1c) ความยาวของ reads ที่อ่านได้เฉลี่ยที่ 35 คู่เบส โดยในกระบวนการหาลำดับเบสหนึ่งครั้ง จะได้จำนวนของเบส 21 ถึง 35 กิกะเบส ความถูกต้องของเบสที่ได้มากกว่าร้อยละ 99 เทคโนโลยีที่กำลังพัฒนาดังกล่าวนี้ จึงถือเป็นรูปแบบหนึ่งในวงการเทคโนโลยีเอ็นจีเอส ที่สามารถอ่านดีเอ็นเอต้นแบบในลักษณะ real-time โดยไม่จำเป็นต้องมีการเพิ่มปริมาณสารพันธุกรรม เพื่อลดความเสี่ยงของการแทนที่ของเบสในช่วงการเพิ่มจำนวน เทคนิคนี้มีประโยชน์เป็นอย่างมากกับลำดับเบส reads ที่ค่อนข้างยาว ทำให้ข้อมูลที่ได้มีความถูกต้องมากยิ่งขึ้น ตัวอย่างวิธีการอื่นๆ เช่น Fluorescence-based single molecular sequencing หรือ Real-Time (SMRT) DNA sequencing technology ที่พัฒนาโดย Pacific BioSciences, Nano-technologies พัฒนาโดย Oxford Nanopore Technologies และ Electronic detection ที่พัฒนาโดย Reveo เป็นต้น

## การประยุกต์ใช้ข้อมูลจากการใช้เทคโนโลยีเอ็นจีเอส

การประยุกต์ใช้ข้อมูลลำดับเบสจากเทคโนโลยีเอ็นจีเอสสามารถนำมาใช้ประโยชน์กับงานวิจัยหลากหลายแขนง. Table 2 แสดงการเปรียบเทียบการประยุกต์ใช้เทคโนโลยีเอ็นจีเอสของแต่ละเครื่อง ซึ่งตารางนี้ได้ดัดแปลงจากเว็บไซต์ของ blueseq.com รายละเอียดสามารถแบ่งเป็นการประยุกต์การใช้งานต่างๆ ดังนี้

### Genomics

#### *De novo sequencing*

เป็นการหาลำดับเบสของสิ่งมีชีวิตที่สนใจจากการเริ่มต้นด้วยการไม่มีข้อมูลลำดับเบสอ้างอิงของสิ่งมีชีวิตที่คล้ายคลึงกัน ซึ่งความยาวของ reads ที่สั้น ทำให้ยากต่อการประกอบเรียงชิ้นลำดับ

เบสโดยเฉพาะจีโนมของยูคาริโอตที่มีขนาดใหญ่ มีโครงสร้างที่ซับซ้อนและมีเบสซ้ำกันเป็นจำนวนมาก ดังนั้น จีโนมแบคทีเรียจึงเป็นจีโนมแรกที่หาลำดับเบสด้วยการใช้เทคโนโลยีเอ็นจีเอส เนื่องจากจีโนมมีขนาดเล็กและโครงสร้างไม่ซับซ้อน (Margulies *et al.*, 2005; Farrer *et al.*, 2009) สำหรับการทำ *de novo* sequencing ในยูคาริโอต เริ่มด้วยการนำเทคโนโลยีเอ็นจีเอสมาผสมผสานกับวิธีการหาลำดับเบสแบบแซงเกอร์ ทำให้ราคาถูกลงแต่ coverage ของแซงเกอร์ค่อนข้างต่ำ เช่น ตัวอย่างของการหาลำดับเบสของเชื้อรา *Grosmannia clavigera* (Diguistini *et al.*, 2009) และพืชจำพวกแตงกวา *Cucumis sativus* (Huang *et al.*, 2009) อย่างไรก็ตาม ปัจจุบันมีการพัฒนาความยาวของ reads รวมถึงความสามารถในการสร้าง paired-end reads และอัลกอริธึมในการทำประกอบเรียงชิ้นลำดับเบสที่เข้าไปจัดการกับจำนวนมหาศาลของ short reads ที่ได้จากเทคโนโลยีเอ็นจีเอส ซึ่งจีโนมของยูคาริโอตสองชนิดแรกที่ประสบผลสำเร็จ ได้แก่ การศึกษาด้านพันธุกรรมของสัตว์เลี้ยงลูกด้วยนม ด้วยการสร้างจีโนม giant panda (Li *et al.*, 2010) ซึ่ง reads ได้จากเครื่อง Solexa และการสร้างจีโนมของเชื้อรา *Sordaria macrospora* ซึ่ง reads ได้จากเครื่อง Solexa และ 454/Roche (Nowrousian *et al.*, 2010)

อย่างไรก็ตาม การใช้ reads ที่ค่อนข้างยาวที่ได้จากการใช้เครื่อง 454/Roche มีความสำคัญมากกับการหาลำดับเบสแบบ *de novo* sequencing เนื่องจากสามารถประกอบเรียงชิ้นลำดับเบสเบื้องต้น (draft) ได้อย่างรวดเร็วและได้คุณภาพที่สูงกว่า รวมทั้งในสิ่งมีชีวิตที่ซับซ้อนหรือมีจีโนมขนาดใหญ่ มักจะมีบริเวณชุดของเบสที่ซ้ำๆ (repetitive) มาก การปิด gap ของ contig จะทำได้ดี ทำให้การประกอบเรียงชิ้นลำดับเบสง่ายขึ้น เหลือ gap น้อยกว่าการใช้ reads ที่สั้น นอกจากนี้ การใช้ paired-end sequencing ดังที่กล่าวไว้ข้างต้น มีประโยชน์มากกับสิ่งมีชีวิตที่ซับซ้อน เช่น พืชและสัตว์ การประกอบเรียงชิ้นลำดับเบสด้วย contig ที่ยาวกว่า ทำให้ได้ scaffold ที่มีขนาดใหญ่ขึ้น ทำให้ช่วยประหยัดเวลาและได้การประกอบเรียงชิ้นลำดับเบสทั้งจีโนมที่มีคุณภาพสูง เช่น ทำให้ทราบได้ว่าส่วนใดควรเป็นตำแหน่งที่ซ้ำหรือมีการจัดเรียงชิ้นส่วนของลำดับเบส (structural rearrangement) เกิดขึ้นในตำแหน่งใดได้บ้าง ดังนั้น การใช้ reads ที่มีคุณสมบัติเป็น paired-end reads ซึ่งประกอบด้วย long tag ความยาวประมาณ 140 ถึง 200 เบสขึ้นไป ช่วยทำให้การจัดเรียงและเชื่อมต่อ reads มีค่า

**Table 2** Applications of next generation sequencing technologies. The performance for a variety of applications is displayed by "+" symbol ("+++" = good, "++" = average, "+" = poor). This information is modified from http://blueseq.com/knowledgebank/sequencing-platforms/comparison-of-sequencing-technologies/.

| Applications | 454/ Roche | Illumina/ Solexa | SOLiD /ABI |
|---|---|---|---|
| Genome : | | | |
| *De novo* sequencing | | | |
| - Small eukaryote | +++ | ++ | + |
| - Large eukaryote | + | +++ | + |
| - Prokaryote | +++ | ++ | + |
| Resequencing | + | +++ | +++ |
| Transcriptome analysis | ++ | +++ | +++ |
| Small RNA sequencing | + | +++ | +++ |
| Epigenetics : | | | |
| - ChIP-seq | + | +++ | +++ |
| - Bisulfite sequencing | + | +++ | ++ |
| - MeDIP | + | +++ | +++ |
| Metagenomics | +++ | ++ | + |

ความเชื่อมั่นในระดับสูง ส่วน long span ความยาวที่ 20 กิโลเบส ช่วยในการระบุตำแหน่งที่เกิดซ้ำๆ (repeat region) ได้ดี อย่างไรก็ตาม การใช้ความยาวที่หลากหลาย เช่น 3, 8 หรือ 20 กิโลเบส สามารถออกแบบเพื่อนำไปใช้ในการวิเคราะห์ลักษณะที่จำเพาะในจีโนมต่างๆได้อย่างเหมาะสมยิ่งขึ้นต่อไปได้ (http://my454.com/applications/whole-genome-sequencing/index.asp)

***Whole genome resequencing***

เป็นการหาลำดับเบสของสิ่งมีชีวิตที่สนใจเทียบกับจีโนมอ้างอิงของสิ่งมีชีวิตที่คล้ายคลึงกันหรือมีอยู่ก่อนแล้ว วิธีการนี้ใช้ short reads จากเทคโนโลยีเอ็นจีเอส ทำให้การแม็ป (map) กับจีโนมอ้างอิงทำได้ด้วยความเชื่อมั่นสูง ซึ่งเหมาะกับจีโนมที่มีขนาดใหญ่ เช่น พวกสัตว์เลี้ยงลูกด้วยนม เป็นต้น ลำดับเบส reads ที่แม็ปได้บนจีโนมอ้างอิง จะช่วยในการจำแนกตำแหน่งการกลายหรือสนิป (SNP) การเพิ่มหรือขาดหายไปของเบส และ copy number variation (CNV) หรือโครงสร้างต่างๆ ทำให้เข้าใจพื้นฐานของความแตกต่างด้านฟีโนไทป์ของสิ่งมีชีวิตต่างๆ ได้ดียิ่งขึ้น ได้แก่ งานด้านความหลากหลายทางพันธุกรรม การศึกษาความสัมพันธ์กับการตอบสนองต่อยา (Hetherington *et al.*, 2002;Auffray*et al.*, 2010) งานด้าน resequencing ของจีโนมมนุษย์ เช่น การหาลำดับเบสจีโนมเฉพาะบุคคลของคนเอเชีย (Wheeler *et al.*, 2008; Wang *etal.*,2009) งานด้านการตรวจหาความหลากหลายของโครงสร้างในจีโนมมนุษย์ เช่น การศึกษาการเปลี่ยนแปลงทางพันธุกรรมที่เป็นสาเหตุของการเกิดโรคในมนุษย์ (Henn *et al.*, 2010)

***Exome/target region sequencing***

Exome ประกอบด้วยลำดับเบสดีเอ็นเอที่สามารถแปลรหัสเป็นโปรตีนได้ ดังนั้น exome sequencing จึงเป็นเทคนิคที่ใช้หาลำดับเบสของจีโนมในส่วนของจีโนมที่สามารถแปลรหัสเป็นโปรตีน หรือเป็นส่วนของตำแหน่งที่เรียกว่า coding region ประโยชน์เพื่อจำแนกยีนที่เกี่ยวข้องกับความผิดปกติหรือการเกิดโรค ตัวอย่างงานวิจัย เช่น การหาลำดับเบสของมนุษย์จำนวน 50 exome ที่แสดงให้เห็นถึงการปรับตัวในการอยู่บริเวณพื้นที่สูง (Yi *et al.*, 2010) สำหรับ target region sequencing เป็นการระบุเป้าหมาย เช่น โครโมโซม Y พื้นที่ที่สนใจ เช่น HLA region, MHC region, QTL region หรือยีนที่สนใจ โดยการใช้วิธี microarray hybridization หรือ solution hybridization ด้วยการใช้โพรบที่ออกแบบ ให้สอดคล้องกับลำดับเบสของพื้นที่จีโนมที่สนใจ ประโยชน์ที่ได้จากการหาลำดับเบส exome/target region สามารถประยุกต์ใช้ในการหาความหลากหลายทางพันธุกรรม เช่น สนิป การเพิ่มหรือการขาดหายไปของลำดับเบสดีเอ็นเอ (InDel) ได้ เช่น การหายีนที่เป็นสาเหตุของความผิดปกติทางพันธุกรรมแบบยีนเดี่ยว (Mendelian condition) การค้นหายีนที่ไวต่อการเกิดโรคในมนุษย์ หรือการตรวจหาการกลายพันธุ์ของเนื้องอกในมะเร็ง เป็นต้น ในพืชและสัตว์ เช่น การพัฒนาหาเครื่องหมายทางพันธุกรรม การศึกษาด้านวิวัฒนาการของสิ่งมีชีวิต หรืองานวิจัยด้านการกลายพันธุ์ในพืช เป็นต้น วิธีการหาลำดับเบส exome สามารถใช้ได้กับเทคโนโลยีเอ็นจีเอส ไม่ว่าจะเป็น 454/Roche, Illumina และ SOLiD/ABI เพื่อลดต้นทุนและได้จำนวน reads ที่มาก

**Transcriptome**

เป็นการตรวจสอบหรือค้นหาชุดของ RNA transcript, mRNA และ non-coding RNA ในสิ่งมีชีวิตเดียวหรือกลุ่มประชากรของสิ่งมีชีวิตต่างๆ การนำเอาเทคโนโลยีเอ็นจีเอส มาประยุกต์ใช้ในการศึกษาด้านนี้ ไม่ว่าจะเป็นการศึกษาในส่วนของ

coding หรือ non-coding region ก็ตาม ก่อให้เกิดประโยชน์ในการจำแนก regulatory RNA และ gene fusion การวิเคราะห์หาสนิปในการศึกษา transcriptome resequencing หรือการค้นหาตำแหน่งของยีนที่ไม่ทราบมาก่อน ซึ่งการใช้ cDNA เป็นต้นแบบ จำนวนของเบสต่อ transcript ที่ได้หาลำดับเบสนั้นมักมีประสิทธิภาพเพียงพอ ที่สามารถบ่งชี้ได้ว่ายีนดังกล่าวอยู่ตำแหน่งใดบนจีโนม เป็นต้น นอกจากนี้ยังทำให้เกิดความเข้าใจถึงพัฒนาการที่เกิดขึ้นในระดับเซลล์หรือเนื้อเยื่อต่างๆ หรือการเกิดโรคในสิ่งมีชีวิตได้มากขึ้นด้วย โดยเป้าหมายของการศึกษาในระดับ transcriptome ประกอบด้วยสามกลุ่มหลักๆ ได้แก่ (1) เพื่อจัดแบ่งกลุ่ม transcript ทั้งหมดในเซลล์ตามชนิดของสิ่งมีชีวิตนั้นๆ ครอบคลุมถึง mRNA, non-coding RNA และ small RNA (2) เพื่อจำแนกโครงสร้างของยีนในระดับทรานสคริปชัน ในส่วนของตำแหน่งโปรโมเตอร์ รูปแบบของ splicing และการเกิด post-transcriptional modification และ (3) เพื่อศึกษาปริมาณระดับการแสดงออกของยีน ในระดับทรานสคริปชัน ในช่วงของการพัฒนา หรือภายใต้สภาวะทางกายภาพหรือการก่อโรค (RNA-seq quantification) (Zhou *et al.*, 2010) อย่างไรก็ตาม การนำเทคโนโลยีเอ็นจีเอสมาใช้กับงานด้าน transcriptome สามารถช่วยในการปิด gap ของจีโนมได้ดียิ่งขึ้น เช่น การหาลำดับเบสของเครื่อง 454 เพื่อสร้าง EST reads จำนวน 391,157 reads จากสมองในระดับ transcriptome ของสัตว์จำพวกต่อ (*Polistes metricus*) (Toth *et al.*, 2007) เมื่อนำ reads ที่ได้เข้าสู่ขบวนการจัดเรียงลำดับเบส (alignment) กับจีโนมและแหล่งข้อมูล EST จากผึ้ง (*Apis mellifera*) และทำ annotation ผลที่ได้พบว่า EST ของสัตว์จำพวกต่อสามารถเข้าคู่กับ mRNA ของผึ้งได้ถึงร้อยละ 39 และถือว่ามีความเกี่ยวพันกันอย่างมีนัยสำคัญ ที่แสดงให้เห็นถึงระดับการแสดงออกของ transcript จากทั้งสองสปีชีส์นี้ ตัวอย่างงานวิจัยอื่นๆ เช่น การนำข้อมูล transcriptome มาใช้ในการตรวจหา gene fusion ในเซลล์มะเร็งและเนื้องอก (Maher *et al.*, 2009)

อย่างไรก็ตาม เครื่อง Illumina และ SOLiD ถือเป็นเทคโนโลยีเอ็นจีเอสที่ให้ reads ที่มีความยาวค่อนข้างสั้น และได้ reads ที่มีจำนวนมากในระดับ 10 ล้าน short reads ซึ่งเหมาะกับการนำไปใช้ประโยชน์ในงาน gene expression profiling มากกว่าวิธีการหาลำดับเบสแบบดั้งเดิม ตัวอย่างงานวิจัย เช่น การตรวจติดตามการแสดงออกของยีนในช่วงการเจริญเติบโตของยีสต์ (Nagalakshmi *et al.*, 2008) หรือการศึกษาช่วงการเกิดไมโอซิสในยีสต์ (Wilhelm *et al.*, 2008) เพื่อนำไปสู่การเข้าใจการแสดงออกของยีนที่เกิดขึ้นในช่วงของการเกิดพัฒนาการต่างๆ รวมถึงเข้าใจความสัมพันธ์ของการแสดงออกของยีนในส่วนของเนื้อเยื่อที่แตกต่างกันในสิ่งมีชีวิตนั้นๆด้วย

นอกจากนี้ การวิเคราะห์ทางชีวสถิติและสารสนเทศ ยังสามารถแบ่งเป็นสามกลุ่มใหญ่ๆ ได้แก่ (1) RNA-seq (transcriptome) ซึ่งเป็นการวิเคราะห์ข้อมูลตามโครงสร้างพื้นฐานของยีนและหน้าที่ของยีน วิธีนี้สามารถทำได้สองแบบคือ *de novo* transcriptome assembly และ transcriptome resequencing ซึ่งสองวิธีนี้สามารถนำไปสู่การค้นหา transcript ที่ไม่ทราบมาก่อน จำแนกการเกิด splicing ศึกษาความหลากหลายที่เกิดขึ้นในระดับ transcriptome เช่น สนิป หรือ การค้นหา gene fusion เป็นต้น (2) RNA-seq (quantification) เป็นการวิเคราะห์การแสดงออกของยีนที่สนใจภายใต้สภาวะต่างๆ เช่น การวิเคราะห์รูปแบบการแสดงออกของยีน การวิเคราะห์หน้าที่ (GO functional), pathway หรือการทำงานร่วมกันระหว่างโปรตีนกับโปรตีน (protein-protein interaction network) ซึ่งวิธีการนี้ให้ผลที่ค่อนข้าง

แม่นยำสูง (3) Small RNA sequencing โดยที่ small RNA ทำหน้าที่เป็นตัวควบคุม (regulator) ที่สำคัญในหลายเซลล์ด้วยกัน เช่น ทำหน้าที่ที่เกี่ยวข้องกับการเจริญหรือพัฒนาการ (development) การเปลี่ยนแปลงภายในเซลล์ (cell differentiation) และการเกิด apoptosis ที่ก่อให้เกิดโรคหลากหลายชนิด ตัวอย่างงานวิจัย เช่น การใช้เครื่อง 454 เพื่อศึกษา small RNA profiling ในมอส (*Physcomitrella patens*) (Axtell *et al.*, 2006) การใช้เครื่อง Illumina และ SOLiD เพื่อสร้าง small RNA library ในเชิงลึก ด้วยการจำแนกลำดับเบสและการแสดงออกของ miRNA จำนวน 334 ยีน ที่ทราบแล้วร่วมกับ miRNA ที่ไม่ทราบ จำนวน 104 ยีน ที่แสดงออกในมนุษย์จากส่วนของ embryonic stem cell (Morin *et al.*, 2008) เป็นต้น

**Epigenetics**

เป็นการศึกษาการเปลี่ยนแปลงในการแสดงออกหรือกิจกรรมของยีนหรือฟีโนไทป์ ซึ่งเป็นสาเหตุจากการเปลี่ยนแปลงของกลไกต่างๆ ซึ่งไม่เกี่ยวข้องกับลำดับเบสดีเอ็นเอด้วยตัวของมันเอง แต่เกิดจากการปรับเปลี่ยนและการจัดเรียงโครงสร้างที่สูงขึ้น การเปลี่ยนแปลงที่เกิดขึ้นนั้นครอบคลุมถึงการเกิด DNA methylation และ histone modification ซึ่งควบคุมโครงสร้างของดีเอ็นเอ รวมถึงการเปลี่ยนแปลงของ stem cell กลไกของสิ่งแวดล้อมที่ส่งผลต่อพัฒนาการของโรคที่ซับซ้อนมากขึ้น และสมรรถภาพการทำงานระดับเซลล์ (cellular aging) ตัวอย่างการประยุกต์ใช้งาน เช่น การทำ bisulfite sequencing, chromatin immunoprecipitation sequencing (ChIP-Seq) และการศึกษา DNA methylome หรือ methylated DNA immunoprecipitation (MeDIP) เป็นต้น ตัวอย่างงานวิจัย เช่น การศึกษาการเจริญของเนื้องอกหรือเซลล์มะเร็งรังไข่ (Balch *et al.*, 2009) การหาลำดับเบสของเซลล์มะเร็งด้วยเทคโนโลยีเอ็นจีเอส เพื่อช่วยในการตัดสินใจทางคลีนิก ทำให้เข้าใจการเกิดโรคและเป็นตัวชี้นำในการเลือกใช้ยาที่เหมาะสม (Jones *et al.*, 2010) การใช้เทคนิค ChIP-seq เพื่อศึกษาการเกิด post-translational modification ของโปรตีนฮีสโตนและตำแหน่งการจับของโปรตีน transcription factor ในระดับจีโนม (Neff and Armstrong, 2009)

**Metagenomics**

เป็นการศึกษาจีโนมของจุลินทรีย์ทั้งหมดที่มีอยู่ในแหล่งที่อยู่ตามธรรมชาติ โดยไม่จำเป็นต้องแยกและเพาะเลี้ยงจุลินทรีย์แต่ละชนิดขึ้นมา ข้อมูลที่ได้นี้สามารถนำไปใช้เป็นประโยชน์ในการศึกษาความหลากหลายและความสัมพันธ์ของจุลินทรีย์ เพื่อนำไปสู่การค้นพบสารหรือยีนใหม่ และการทำนายหน้าที่ของยีน ซึ่งนิยมใช้เครื่อง 454/Roche ในงานด้านนี้อย่างกว้างขวาง เช่น จำแนกความแตกต่างของ ribosomal RNA เช่น 16s หรือ 18s เนื่องจากความยาวของ reads ที่ได้ค่อนข้างยาว ทำให้การจัดเรียงและเชื่อมต่อลำดับเบสสามารถทำได้อย่างสมบูรณ์ และการประกอบเรียงชิ้นลำดับเบสสามารถครอบคลุมส่วนของยีนที่ทราบและไม่ทราบได้ ตัวอย่างงานวิจัยอื่นๆ เช่น การสำรวจความหลากหลายของกลุ่มประชากรเชื้อราในดินจากสามเกาะในทะเลเหลืองของประเทศเกาหลี เกาะระหว่างประเทศเกาหลีและประเทศจีน (Lim *et al.*, 2010) นอกจากนี้ ได้มีการพัฒนาขั้นตอนการทดลองการเตรียม library ใหม่ และพัฒนาระบบการวิเคราะห์ทางคอมพิวเตอร์ให้มีความสามารถในการสร้าง reads ที่ยาวขึ้นได้ (Rodrigue *et al.*, 2010) ตัวอย่างงานวิจัย เช่น การนำ short reads จากเครื่อง Illumina เข้ามาใช้ในการศึกษาบทบาทหน้าที่ของกลุ่มจุลินทรีย์ในลำไส้ของมนุษย์ (Qin *et al.*, 2010)

**อัลกอริธึมและโปรแกรมที่ใช้ในการจัดเรียงลำดับเบสดีเอ็นเอ (mapping) และโปรแกรมสำหรับใช้ในประกอบเรียงชิ้นลำดับเบสดีเอ็นเอใหม่ (assembly)**

ด้วยจำนวนมหาศาลของข้อมูล reads ปัญหาความผิดพลาดของข้อมูลที่ได้จากการอ่านของเครื่องที่ใช้เทคโนโลยีเอ็นจีเอส (sequencing error) กล่าวคือเครื่อง 454/Roche มีอัตราความผิดพลาดของการเกิดลักษณะเบสซ้ำๆ ซึ่งเกิดขึ้นจากการตรวจจับสัญญาณของเครื่องที่ผิดพลาดไป เรียกบริเวณนั้นว่า homopolymer ซึ่งเกิดเป็นลักษณะของเบสซ้ำๆ ของนิวคลีโอไทด์มากถึง 6 เบส เช่น TTT เมื่อเทียบกับ TTTT ความยาวของ homopolymer ที่เครื่องมีการอ่านผิดพลาดไป ทำให้ความถูกต้องของข้อมูลลดลง (quality value) ได้ เช่น เกิดลักษณะการเพิ่มหรือการขาดหายไปของลำดับเบส reads ซึ่งอาจไม่ได้เกิดขึ้นจริงตามธรรมชาติ

สำหรับเครื่อง Illumina ลำดับเบสของ reads ที่ได้ค่อนข้างสั้นมาก ประมาณ 35 ถึง 100 คู่เบส และในกระบวนการหาลำดับเบสหนึ่งครั้ง ได้จำนวนของเบสมากถึง 1.5 กิกะเบสซึ่งมากกว่าเครื่อง 454/Roche ที่ได้จำนวนของเบสประมาณ 600 เมกกะเบสและมีความยาวของลำดับเบส 400 คู่เบส ส่วนใหญ่ความผิดพลาดของลำดับเบสที่เกิดขึ้น เกิดจากการแทนที่เบส (substitution) ซึ่งเกิดจากขั้นตอนของการเพิ่มปริมาณ (amplification) นอกจากนี้ จำนวนของ reads ที่ได้มีจำนวนมากกว่าเทคนิคอื่นๆ ทำให้ยากต่อการวิเคราะห์ข้อมูล เนื่องจากขาดความความสามารถในการเพิ่มหรือขยายในส่วนที่เป็นตำแหน่งที่มีเบสซ้ำ ดังนั้น ประสิทธิผลและจำนวนของ reads ที่เข้าคู่กันได้อาจจะไม่ใช่การเข้าคู่กันของ reads ที่แท้จริง

ส่วนเครื่อง SOLiD สามารถผลิต reads และเป็นเครื่องมือที่ใช้เทคนิคของการเชื่อมต่อ ซึ่งการหาลำดับเบสของดีเอ็นเอ ทำโดยการอ่านตำแหน่งของเบสบนสายดีเอ็นเอแบบซ้ำกันสองครั้ง ดังนั้นการอ่านจากสีที่ได้จากเครื่อง ทำให้เกิดความยุ่งยาก ในการคำนวณหาเบสดีเอ็นเอที่แน่นอน เนื่องจากสีแต่ละสี จะเกี่ยวข้องกับคู่ของนิวคลีโอไทด์ที่ต่างกัน นอกจากนี้ปัญหาของการให้สีในช่วงกระบวนการหาลำดับเบส อาจไม่มีประสิทธิภาพพอที่จะตรวจหาลำดับเบส เช่น การขาดความต่อเนื่องของสัญญาณระหว่างการอ่านลำดับเบส ทำให้เกิดความผิดพลาดของสี (missing color) ซึ่งเป็นผลให้เกิด frame-shift ขึ้นในช่วงของการวิเคราะห์ข้อมูล

นอกจากความผิดพลาดของเครื่องดังที่ได้กล่าวมาข้างต้นนั้น ความหลากหลายของข้อมูลจีโนม (polymorphism) ที่เกิดขึ้นตามธรรมชาติ เช่น สนิป การเพิ่มหรือการขาดหายไปของลำดับเบส การเกิด rearrangement ของชิ้นส่วนดีเอ็นเอระหว่างจีโนมที่เกิดขึ้น ในช่วงการเกิดวิวัฒนาการก็มีส่วนสำคัญในการส่งผลต่อการวิเคราะห์ข้อมูลเช่นกัน

จากปัญหาดังกล่าวข้างต้น ปัจจุบันจึงได้มีการคิดค้นเครื่องมือทางชีวสารสนเทศเกิดขึ้นเป็นจำนวนมาก โดยเฉพาะงานด้านการจัดเรียงและเชื่อมต่อลำดับเบส (alignment reads) จากข้อมูลเอ็นจีเอส ซึ่งเป็นขั้นแรกของการหาลำดับเบสของจีโนมที่สำคัญ (Table 3) เทคนิคที่ใช้ในการจัดเรียงและเชื่อมต่อลำดับเบส สามารถแบ่งเป็นสองหลักการด้วยกัน คือ (1) การบ่งชี้ตำแหน่งในจีโนม (index genome) ด้วยวิธี BWT (Burrows-Wheeler space transformation) วิธีนี้ช่วยลดหน่วยความจำเป็นหลักสำคัญ เช่น โปรแกรม Bowties และ SOAP และ (2) การสร้างตารางอ้างอิงตำแหน่ง (hash table) ซึ่งสามารถทำได้เป็นสองแบบย่อยๆ

**Table 3** Sequence alignment tools for the analysis of next generation sequencing data.

| Software | Open source /system | Tool support | Mapping strategy |
|---|---|---|---|
| SSAHA/SSAHA2 (Ning *et al.*, 2001) | Yes /Linux, Mac | 454/Roche,Illumina, SOLiD/ABI | Index genome with hash table |
| MAQ (Li *et al.*, 2008) | Yes /GPL | Illumina, SOLiD/ABI | Index read with hash table |
| SOAP (Li *et al.*, 2008) | Yes /Linux, Unix >14 GB RAM against human genome | Illumina, SOLiD/ABI | Index genome with hash table |
| Mosaik (Michael Strömberg, Boston University) | Yes /Linux, Sun, Win, OSX | 454/Roche, Illumina, SOLiD/ABI | Index genome with hash table |
| Bowtie (Langmead *et al.*, 2009) | Yes /Linux,SRC, Win, MAC | 454/Roche, Illumina, SOLiD/ABI | Index genome with BWT |
| PASS (Campagna *et al.*, 2009) | Yes /Linux, Win | 454/Roche, Illumina, SOLiD/ABI | Index genome with hash table |
| SHRiMP (Rumble *et al.*, 2009) | Yes /Linux, ICC, Win, OSX | 454/Roche, Illumina, SOLiD/ABI | Index read with hash table |
| BWA (Li and Durbin, 2009) | Yes /Linux, Unix 32,64 bit | 454/Roche, Illumina, SOLiD/ABI | Index genome with BWT |
| GASSST (Rizk and Lavenier, 2010) | Yes /Linux, Unix/32,64 bit | 454/Roche, Illumina, SOLiD/ABI | Index genome with hash table |
| TAPyR (Fernandes *et al.* , 2011) | Yes /Multiple (C complier only) | 454/Roche, Illumina, Ion Torrent and SMRT | BWT combined with a flexible seed based approach |

ด้วยการให้จีโนมเป็นตัวบ่งชี้ หรือลำดับเบส reads เป็นตัวบ่งชี้ ซึ่งการสร้างตัวบ่งชี้ (index) จะช่วยลดเวลาของการหารูปแบบที่เข้าคู่กันของข้อมูล เนื่องจากมีการระบุตำแหน่งในการค้นหา จึงไม่จำเป็นต้องสแกนทั้งหมดของเบส ตัวอย่างโปรแกรมที่ใช้วิธีสร้างตารางอ้างอิงตำแหน่ง และใช้ reads เป็นตัวบ่งชี้ ด้วยการค้นหา reads ไปที่จีโนม เช่น MAQ, SHRiMP และ Newbler ซึ่งหน่วยความจำจะขึ้นอยู่กับจำนวน reads ข้อเสียคือ ถ้า reads มีจำนวนน้อย มักจะใช้เวลาในการคำนวณนาน เนื่องจากต้องสแกนทั้งจีโนมทั้งหมด สำหรับโปรแกรมที่ใช้วิธีสร้างตารางอ้างอิงตำแหน่ง และใช้จีโนมเป็นตัวบ่งชี้ เช่น Bowtie, BWA, Mosaik, SOAP, PASS, SSAHA2 ถ้ามีหน่วยความจำค่อนข้างใหญ่ สามารถคำนวณแบบขนานได้ง่าย จึงเหมาะกับงานที่ใช้จีโนมขนาดใหญ่ได้ดี เช่น จีโนมของมนุษย์ เป็นต้น อย่างไรก็ตาม โปรแกรมเหล่านี้มีข้อด้อย ที่บางโปรแกรมไม่ได้สนับสนุนการใช้งานกับข้อมูลที่ได้จากเครื่อง 454/Roche เช่น SOAP และ MAQ บางโปรแกรมไม่สนับสนุนลักษณะของข้อมูลบางอย่าง เช่น ไม่สามารถแปลและจำแนกสัญญาณของสนิป มีข้อจำกัดของ reads ต้องมีความยาวสูงสุดในการจัดเรียงและเชื่อมต่อลำดับเบส ได้เพียง 500 คู่เบส เช่น โปรแกรม BWA เป็นต้นปัจจุบันกลุ่มทีมวิจัยของ Bao *et al.* (2011) ได้ประเมินโปรแกรมที่ใช้ในการจัดเรียงและเชื่อมต่อลำดับเบสข้อมูลที่ได้ทั้งหมด 9 ซอฟแวร์ ด้วยการใช้ข้อมูล reads จริงที่ได้จากเครื่อง Illumina ซึ่งจากการทดลองการใช้ข้อมูลแบบ single reads (SE) ซึ่งจะมีลำดับเบสดีเอ็นเอ ที่สร้างจากส่วนปลายด้านใดด้านหนึ่ง พบว่าโปรแกรม SHRiMP ใช้เวลาในการประมวลผลนานที่สุด (8681.61 นาที) และใช้หน่วยความจำมากที่สุด (26.58 กิกะไบต์) แต่ได้เปอร์เซ็นต์การแม็ปสูงสุด (ร้อยละ 86.77) ส่วน BWA มีเปอร์เซ็นต์การแม็ปใกล้เคียงกับ SHRiMP (ร้อยละ 88.67) แต่ใช้เวลาประมวลผลเร็วกว่า SHRiMP (236.36 นาที) และใช้หน่วยความจำน้อยที่สุด (3.17 กิกะไบต์) สำหรับข้อมูล paired-end reads (PE) ซึ่ง PE จะมีส่วนปลายประกอบด้วย read หนึ่งคู่ โดยแต่ละ read จะอยู่ที่ปลายทั้งสองตรงข้ามกัน และทราบระยะที่ห่างกันระหว่าง reads คู่ดังกล่าว การใช้ PE จะทำให้การเปรียบเทียบทำได้ดีขึ้น ซึ่งมีประโยชน์ในการหาลำดับเบสที่ไม่ทราบจีโนมของสิ่งมีชีวิตอย่างยิ่ง จากการทดลองพบว่า Mosaik ค่อนข้างให้ผลการแม็ป (ร้อยละ 83.4) ดีกว่า BWA (ร้อยละ 80.46) ส่วน SHRiMP ทำงานได้ไม่ดีนัก (ร้อยละ 53.38) โดยสรุปภาพรวมพบว่า BWA, Mosaik, SHRiMP และ SOAP เหมาะกับการจัดเรียงและเชื่อมต่อลำดับเบสของข้อมูล SE และ PE ที่ได้จากเครื่อง Illumina โดย BWA ใช้หน่วยความจำน้อยที่สุด และ SOAP มีกระบวนการประมวลผลเร็วที่สุด นอกจากนี้ทีมวิจัยดังกล่าวได้ทดสอบเพิ่มเติมด้วยการใช้ข้อมูล reads จากเครื่อง 454/Roche ด้วยการประเมินการใช้งานกับโปรแกรม Mosaik, PASS และ SSAHA2 พบว่าความถูกต้องในการจัดเรียงและเชื่อมต่อลำดับเบสทำได้ไม่ดี ใช้เวลาประมวลผลนาน ในส่วนของข้อมูล SOLiD/ABI ที่มีหลายโปรแกรมสนับสนุน เช่น Mosaik, PASS, Bowtie และ SHRiMP พบว่าความสามารถในการจัดเรียงและเชื่อมต่อลำดับเบสค่อนข้างต่ำ ปัจจุบัน มีโปรแกรมที่เกิดขึ้นใหม่ เช่น TAPyR (Fernandes *et al.*, 2011) ซึ่งใช้วิธีการผสมผสานของสองอัลกอริธึม มาใช้จัดการกับข้อมูลเอ็นจีเอสสำหรับเครื่อง Illumina และ 454/Roche และแก้ปัญหาของข้อมูลที่เป็น homopolymer ด้วยการสร้างตัวบ่งชี้ของจีโนม ด้วยพื้นฐานของการใช้วิธี BWT เพื่อช่วยจัดเรียงและเชื่อมต่อลำดับเบสให้มีความรวดเร็ว และลดหน่วยความจำในการประมวลผล โดยใช้เทคนิค multiple seed heuristic เพื่อหาลำดับเบสที่มีการจัดเรียงและเชื่อมต่อ

ลำดับเบสที่ดีที่สุด ซึ่งจากการทดลองเปรียบเทียบกับโปรแกรม BWA, SSAHA2, Newbler, GASSST และ Segemehl โดยใช้ข้อมูล reads จากเครื่อง 454/Roche จากตัวอย่างสิ่งมีชีวิต 6 ชนิด พบว่าโปรแกรม TAPyR ใช้เวลาในการประมวลผลเร็วกว่าโปรแกรมอื่น โดยโปรแกรม TAPyR มีจุดเด่นที่สามารถกำหนดค่าความเหมือน (identity) ที่ต้องการได้ โปรแกรม TAPyR จึงให้ผลดีในแง่ของความแม่นยำ และการจัดเรียงและเชื่อมต่อลำดับเบสได้อย่างถูกต้อง แต่ในแง่ของหน่วยความจำที่ใช้พบว่า BWA ใช้หน่วยความจำน้อยที่สุด

โปรแกรมสำหรับใช้ในการประกอบเรียงชิ้นลำดับเบส (fragment assembly) อาศัยหลักการคือ การจัดกลุ่มรวมกันของชิ้นลำดับเบสดีเอ็นเอสั้นๆ เป็น contig จาก contig รวมกันเป็น contig ที่ยาวขึ้น หรือที่เรียกว่า scaffold เพื่อประกอบเป็นลำดับเบสดีเอ็นเอสายใหม่ โดยแบ่งได้เป็น 2 วิธีที่ต่างกันได้แก่ *de novo* approach และ resequencing หรือ comparative approach

*De novo* approach มีเป้าหมายเพื่อใช้ในการสร้างจีโนมใหม่ (reconstruction) สามารถทำได้โดยไม่ต้องทราบลำดับเบสอ้างอิงมาก่อน ซึ่งไม่เหมือนกับวิธีแบบ comparative approach ที่มีการแม็ป reads โดยมีลำดับเบสอ้างอิงเป็นตัวเปรียบเทียบ เพื่อให้ได้สิ่งใหม่จากลำดับเบสของสิ่งมีชีวิตนั้น ซึ่งวิธีนี้ทำได้ง่ายกว่าวิธีการแบบ *de novo* ดังนั้น การค้นหาสิ่งใหม่ๆ ด้วยวิธี *de novo* จึงเป็นเรื่องที่ท้าทาย การคำนวณทางคณิตศาสตร์มักทำได้ยากและซับซ้อน เนื่องจากไม่ทราบลำดับเบสอ้างอิง และต้องระมัดระวังเรื่องความหลากหลายทางชีวภาพทางธรรมชาติ ที่อาจเกิดขึ้นในสิ่งมีชีวิตใหม่นั้นด้วย ซึ่งความซับซ้อนดังกล่าวที่เป็นปัจจัยหลักต่อการวิเคราะห์ คือ ความยาวและปริมาณของ reads โดย reads ที่มีความยาวที่สั้นมักมีลักษณะที่เป็นเบสซ้ำๆ เกิดขึ้นมาก ซึ่งง่ายต่อการแม็ป แต่การประกอบเรียงชิ้นลำดับเบสทำได้ค่อนข้างยาก การจัดการกับปริมาณของ reads ที่มีจำนวนมหาศาล ความยาว reads จากเทคโนโลยีเอ็นจีเอสที่ได้ค่อนข้างสั้นอยู่ที่ 35 ถึง 400 คู่เบส ซึ่งสั้นกว่าเทคนิคการหาลำดับเบสแบบดั้งเดิมที่ให้ reads มีความยาว 600 ถึง 800 คู่เบส สิ่งนี้ล้วนส่งผลต่อการออกแบบการประกอบเรียงชิ้นลำดับเบสเพื่อให้ได้ค่า coverage ที่ดีได้

ปัจจุบัน อัลกอริธึมสำหรับ *de novo* assembly มี 3 วิธีที่ใช้กันอย่างแพร่หลาย ได้แก่ (1) Greedy, (2) Overlap-layout-consensus (OLC) และ (3) Eulerian หรือ de Bruijn graph ส่วนอัลกอริธึมอื่นๆ เช่น hybrid (Figure 12) ยังคงมีการพัฒนาอย่างต่อเนื่อง โปรแกรมที่ใช้อัลกอริธึมแบบ Greedy เช่น SSAKE, SHARCGS, VCAKE, Forge และ QSRA เช่น โปรแกรม QSRA ถูกพัฒนาขึ้นมาเพื่อจัดการกับ reads ที่มีการแม็ปแบบไม่สมบูรณ์ โปรแกรมนี้มีประสิทธิภาพในเรื่องของความเร็วและได้คุณภาพของจีโนมค่อนข้างดี (Bryant *et al.*, 2009) อัลกอริธึมแบบ OLC เช่น CABOG, Edena, Newbler และ Shorty การทำงานของโปรแกรม Newbler (Margulies *et al.*, 2005) ออกแบบมาเพื่อจัดการกับ homopolymer ที่เกิดขึ้นในจากการอ่านของเครื่อง 454/Roche ซึ่งโปรแกรมนี้สามารถจัดการกับความผิดพลาดที่เกิดจากการจำแนกสัญญาณลำดับเบสของเครื่องและบริเวณที่มีเบสซ้ำกันได้ดี ส่วนโปรแกรม Shorty (Hossain *et al.*, 2009) ออกแบบมาให้สามารถทำงานได้หลากหลาย platform เช่น Illumina, SOLiD/ABI และ Holicos เป็นต้น ส่วนอัลกอริธึมแบบ de Bruijn graph ถูกนำไปใช้งานกับเครื่อง Solexa และ SOLiD อย่างแพร่หลาย เช่น โปรแกรม ABySS, ALLPATHS, EULER-SR, SOAPdenovo และ Velvet ตัวอย่างเช่น โปรแกรม ABySS (Simpson *et al.*, 2009) ออกแบบมาเพื่อแก้ปัญหาหน่วยความจำ

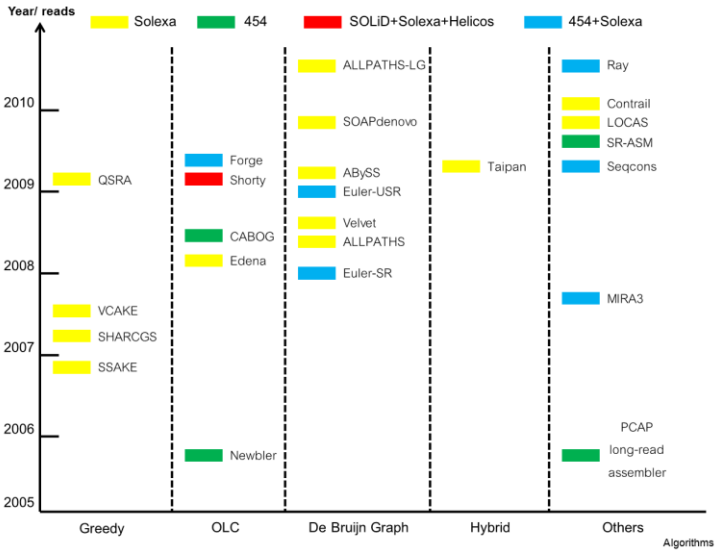


**Figure 12** Overview of algorithms and tools for *de novo* assembly analysis developed during 2005 to 2010. There are mainly three algorithms: Greedy, OLC and de Bruijn graph. Color box symbols are representative of short reads from different platforms. (Redrawn from Zhang *et al*., 2011).

ที่ใช้ในการประกอบเรียงชิ้นลำดับเบส ซึ่งมีข้อจำกัดกับจีโนมที่มีขนาดใหญ่ได้ดี เป็นต้น

จากที่ได้กล่าวมาทั้งหมดข้างต้น จะเห็นได้ว่าการใช้เทคโนโลยีเอ็นจีเอสในปัจจุบัน มีการพัฒนารูปแบบของเทคโนโลยีขึ้นมาใหม่ๆอย่างต่อเนื่อง ทั้งนี้ขึ้นอยู่กับเป้าหมายของการใช้งานเป็นหลักว่าผู้ใช้นั้นต้องการหรือพัฒนาปรับปรุงในส่วนใด ไม่ว่าจะเป็นเรื่องการออกแบบโพรบที่ใช้ในขั้นตอนการทดลองในห้องปฏิบัติการด้านชีวสารสนเทศ ด้วยการพัฒนาอัลกอริธึมใหม่ เพื่อใช้ตอบโจทย์ในการวิเคราะห์ข้อมูลเอ็นจีเอส หรือการผลิตอุปกรณ์สนับสนุนเครื่องมือการหาลำดับเบสใหม่ๆ เพื่อลดต้นทุนการผลิต เป็นต้น ทั้งนี้เพื่อก่อให้เกิดประโยชน์สูงสุดในการนำไปประยุกต์ใช้ในอนาคตต่อไป

## เอกสารอ้างอิง

Anderson, M.W. and Schrijver, I. 2010. Next generation DNA sequencing and the future of Genomic Medicine. *Genes* 38–69.

Ansorge, W.J. 2009. Next-generation DNA sequencing techniques. *Nat Biotechnol* 25: 195–203.

Auffray, C., Charron, D. and Hood, L. 2010. Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Med* 2: 57.

Axtell, M.J., Jan, C., Rajagopalan, R. and Bartel, D.P. 2006. A two-hit trigger for siRNA biogenesis in plants. *Cell* 127: 565–577.

Balch, C., Fang, F., Matei, D.E., Huang, T.H.

and Nephew, K.P. 2009. Minireview: epigenetic changes in ovarian cancer. *Endocrinology* 150: 4003–4011.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X. and Song, Y.Q. 2011. Evaluation of next-generation sequencing software in mapping and assembly.*J Hum Genet* 56: 406–414.

Bryant, D.W. Jr., Wong, W.K. and Mockler, T.C. 2009. QSRA: a quality-value guided *de novo* short read sequence assembler. *BMC Bioinformatics* 10: 69.

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., Vitulo, N. and Valle, G. 2009. PASS: a program to align shortsequences.*Bioinformatics*25:967–968.

Diguistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., *et al.* 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illuminasequencedata.*GenomeBiol* 10: R94.

Farrer, R.A., Kemen, E., Jones, J.D. and Studholme, D.J. 2009. De novo assembly of the *Pseudomonas syringae* pv. syringae B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol Lett* 291: 103–111.

Fernandes, F., da Fonseca, P.G., Russo, L.M., Oliveira, A.L. and Freitas, A.T. 2011. Efficient alignment of pyrosequencing reads for re-sequencing applications. *BMC Bioinformatics* 12: 163.

Garber, M., Zody, M.C., Arachchi, H.M., Berlin, A., Gnerre, S., Green, L.M., Lennon, N. and Nusbaum, C. 2009. Closing gaps in the human genome using sequencing by synthesis. *Genome Biol* 10: R60.

Gega, A. and Kozal, M.J. 2011. Technology to detect low-level drug-resistant HIV variants: general overview of ultra-deep massively parallel pyrosequencing: 'ultra-deep sequencing'. *Future Virol* 6: 17–26.

Henn, B.M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. and Bustamante, C.D. 2010. Fine-scale population structure and the era of next-generation sequencing. *Hum Mol Genet* 19: R221–226.

Hetherington, S., Hughes, A.R., Mosteller, M., Shortino, D., Baker, K.L., Spreen, W., *et al*. 2002. Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* 359: 1121–1122.

Hossain, M.S., Azimi, N. and Skiena, S. 2009. Crystallizing short-read assemblies around seeds. *BMC Bioinformatics* 10: S16.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., *et al*. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41: 1275–1281.

Jones, S.J., Laskin, J., Li, Y.Y., Griffith, O.L., An, J., Bilenky, M., *et al*. 2010. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol* 11: R82.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences tothehumangenome.*Genome Biol* 10: R25.

Li, H., Ruan, J. and Durbin, R. 2008. Mapping

short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851–1858.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, R., Li, Y., Kristiansen, K. and Wang, J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713–714.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., *et al*. 2010. The sequence and *de novo* assembly of the giant panda genome. *Nature* 463: 311–317.

Lim, Y.W., Kim, B.K., Kim, C., Jung, H.S., Kim, B.S., Lee, J.H. and Chun, J. 2010. Assessment of soil fungal communities using pyro sequencing. *J Microbiol* 48: 284–289.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458: 97–101.

Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9: 387-402.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al*. 2005. Genome sequencing in micro fabricated high-density picolitre reactors. *Nature* 437: 376–380.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C.J. and Marra, M.A. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 18: 610–621.

Morozova, O., Hirst, M. and Marra, M.A. 2009. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* 10: 135–151.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344–1349.

Neff, T. and Armstrong, S.A. 2009. Chromatin maps, histone modifications and leukemia. *Leukemia* 23: 1243–1251.

Ning, Z., Cox, A.J. and Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res* 11:1725–1729.

Nowrousian, M., Stajich, J.E., Chu, M., Engh, I., Espagne, E., Halliday, K., *et al*. 2010. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet* 6: e1000891.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., *et al*. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–65.

Rizk, G. and Lavenier, D. 2010. GASSST: global alignment short sequence search

tool. *Bioinformatics* 26: 2534–2540.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J. and Chisholm, S.W. 2010. Unlocking short read sequencing for metagenomics. *PLoS One* 5: e11840.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A. and Brudno, M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* 5: e1000386.

Sambrook, J. and Russell D.W. 2006. Fragmentation of DNA by Nebulization. *Cold Spring Harb Protoc*. doi:10.1101/pdb.prot 4539.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S. J. M. and Birol, I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* 19: 1117–1123.

Son, M.S., Taylor, R.K. 20011. Preparing DNA libraries for multiplexed paired-end deep sequencing for illumina GA sequencers. *Curr Protoc Microbiol* 20: 1E.4.1–1E.4.13.

Toth, A.L., Varala, K., Newman, T.C., Miguez, F.E., Hutchison, S.K., Willoughby, D.A., Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M.E. and Robinson, G.E. 2007. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* 318: 441–444.

Wang, S.W., Tang, X., Fan, Z.H., Wu, X., Lioy, P.J. and Georgopoulos, P.G. 2009. Modeling of personal exposures to ambient air toxics in Camden, New Jersey: an evaluation study. *J Air Waste Manag Assoc* 59: 733–746.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., *et al*. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872–876.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bahler, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239–1243.

Woollard, P.M., Mehta, N.A., Vamathevan, J.J., Van Horn, S., Bonde, B.K. and Dow, D.J. 2011. The application of next-generation sequencing technologies to drug discovery and development. *Drug Discov Today* 16: 512–519.

Yamey, G. 2000. Scientists unveil first draft of human genome. *BMJ* 321: 7.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., *et al*. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329: 75–78.

Zhang, W., Jiajia, C., Yang, Y., Yifei, T., Jing, S. and Bairong, S. 2011. A practical comparison of *De Novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* 6: e17915.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y. and Yu, J. 2010. The next-generation sequencing technology: a technology review and future perspective. *Sci China Life Sci* 53: 44–57.