Ion Torrent PGM을 통한 차세대 시퀀서의 새로운 흐름

박근준 Life Technologies FBS (Field Bioinformatics Specialist)

그림 1. Life Technologies사의 ion torrent PGM

2003년 완료된 인간게놈프로젝트(Human Genome Project)는 인간 한 사람 분의 유전체 정보를 밝혀냈다는 큰 의미를 가졌었지만, 역시 30억 염기쌍이라는 방대한 정보를 읽어낸다는 것은 너무나 큰 작업이라는 것도 깨달을 수 있었다. 하지만 최근 몇 년 동안 유전체학은 전통적인 생거 시퀀싱 방법에서 벗어난 차세대 시퀀서(NGS: Next Generation Sequencer)가 등장하면서 진정한 의미의 유전체 시대를 열었다. 급격한 반도체 발전 속도를 가리키는 무어의 법칙마저 능가하는 속도로 유전 정보의 시퀀싱이 저렴하고 빠르게 된 것이다[1]. 이제는 어떤 한 사람의 유전체 정보를 읽어내는 것이 아니라 여러 명의 유전체 정보를 동시에 읽어내어 방대한 정보를 서로 비교하며 분석하는 것도 가능한데, 그 의미는 단순히 염기 서열을 읽어냈다는 것에서 벗어나 의학, 과학적으로 가치가 있는 지식을 밝혀내는 단계로 접어 들었고, 실제로 다양한 유전질환을 중심으로 그 원인이 되는 유전정보의 변이에 대한 새로운 사실들이 밝혀지고 있다[2]. 하지만 수억 원대의 가격인 NGS를 구입하여 최적의 상태로 유지하고 운영하면서 고가의 실험을 반복해서 수행하는 것은 아직은 상당한 시설과 고급 연구 인력을 요구한다는 것도 사실로, 실제로는 대량의 시퀀싱을 주도하는 큰 규모의 게놈센터 등을 중심으로 활용되며, 보통 많은 대학 연구실은 공동장비 형태로 함께 부담하여 구입, 운영하기도 한다.

이러한 상황에서 미국의 Life Technologies사는 2010년 말에 Ion Torrent PGM(Personal Genome Machine)이라는 획기적인 시퀀서를 등장시켰다[그림 1]. 기존의 NGS와 같이 형광을 사용하여 염기를 구분하고 카메라로 영상을 찍어 이미지 처리를 하는 등의 과정을 전부 생략하고, 반도체 기술을 적용하여 더 직접 염기를 읽어내는 새로운 방식이다. 한 번의 실험으로 생산하는 데이터 양(throughput)을 줄이는 대신 기기 자체의 크기와 가격, 복잡성, 실험 비용을 대폭 줄여서 NGS에 대한 문턱을 혁신적으로 낮추었다는 평가를 받았다[3, 4]. 그렇기 때문에 Life Technologies사는 MIT technology review에서 2011년도의 가장 혁신적인 기업 50(The 50 Most Innovative Companies 2011)의 하나로도 선정되었다. 2~3시간의 초고속 시퀀싱을 실현한 Ion PGM은 2011년 여름, 독일을 중심으로 유럽에서 많은 사망자를 발생시킨 수수께끼의 병원균이 어떠한 장출혈성 변종 대장균인지를 밝혀냄으로써 실제로 그 위력을 발휘했다[5, 6, 그림 2]. Life Technologies사와 중국 BGI에서 별개로 진행되어 논문으로도 발표된 이 사례는 앞으로 이러한 수수께끼의 전염병

이 발생할 경우 과학자들이 신속하게 그 원인을 추적하고 대비하는데 Ion PGM과 같은 시퀀서가 얼마나 중요한지를 증명한 셈으로 2011년 9월의 Nature Biotechnology에서도 "Outbreak genomics"란 제목으로 상세하게 다루어졌다[7]

그림 2. 유럽에서 큰 피해를 일으킨 변종 대장균 O104:H4의 정체를 밝혀낸 ion torrent PGM.

기존의 NGS 장비들과 비교해서 가장 다른 것은 1회용 반도체 칩을 사용하여 시퀀싱이 이루어진다는 점이다. 반도체 칩의 작은 구멍(well)에 비즈(beads)가 들어가서 DNA 합성을 진행시키면서 염기를 읽는데, 기존의 NGS처럼 형광을 사용하여 이미지를 만들고 스캐닝을 하는 모든 과정은 생략되고, 대신 특정 염기가 상보적으로 결합하며 발생하는 수소이온 때문에 생기는 미세한 pH 변화를 반도체에서 직접 읽어낸다[그림 3]. 조금 더 자세히 설명하면, ATGC 네 가지 염기에서 서로 다른 pH 변화가 생기는 것이 아니라 순서대로 한 종류씩 염기를 보내서 반응이 일어나는지 안 일어나는지를 보는 과정을 신속하게 반복하는 원리이다. 만일 반응이 일어나지 않으면 씻어내고 다음 염기를 반응시키며, 만일 반응이 일어난다면 그게 염기 1개가 붙는 반응이었는지, 아니면 2개 이상의 몇 개의 염기가 붙었는지(AAAAA처럼 같은 염기가 반복되는 경우) pH 변화의 세기를 측정하고 씻어내는 과정을 반복한다. 반도체의 집적도가 높아지면 높아질수록 데이터 생산량은 10배, 100배로 높아지게 된다.

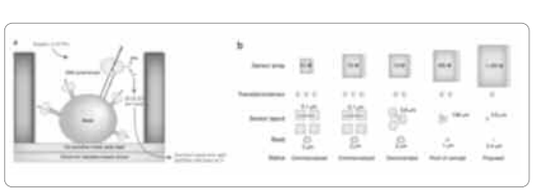

그림 3. 시퀀싱 원리와 앞으로의 발전 방향 (nature biotechnology, 29(9), p 805-807.)

생물정보학 입장에서는 어떠한 형식의 파일로 염기서열이 나오는지도 중요한데, ion PGM은 454에서 사용되던 SFF 형식과 NGS에서 가장 널리 사용되는 FASTQ 형식을 제공하고, 추가로 reference 서열에 대한 mapping(alignment)까지 완료하여 그 결과인 SAM/BAM 파일도 완성된다. 따라서 연구자는 SFF나 FASTQ 파일에 대응하는 기존의 다양한 소프트웨어로 mapping을 포함한 모든 생물학적 분석을 수행할 수 있고, 또는 SAM/BAM 파일을 사용하여 variation 분석과 같은 흔히 NGS의 3차 분석이라는 연구를 진행시킬 수도 있다. Ion PGM은 ion 서버라는 컴퓨터가 추가되며, 한 대의 서버로 2~3대의 ion PGM까지 연결할 수 있는데, 새로 개발된 서열 정렬 프로그램도 제공한다. TMAP(Torrent Mapping Alignment Program)이라는 이름의 이 프로그램은 BWA-short, BWA-long, SSAHA라는 기존의 세 가지 정렬 알고리즘을 기반으로 NGS의 서열 정렬 프로그램의 하나였던 BFAST의 개발팀이 ion PGM의 데이터 특성에 최적화해서 완성시켰다.

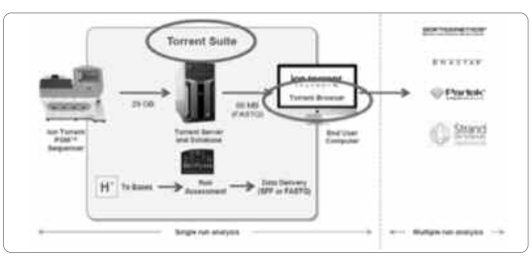

그림 4. Ion Torrent PGM의 데이터 분석 흐름.

실제 실험에서는 샘플을 준비하는 과정 등도 지원하는 다양한 자동화 장비가 함께 개발되었으며, 따라서 샘플 준비와 시퀀싱, 서버에서의 분석까지가 모두 몇 시간 만에 완료된다. 기기 자체의 조작은 주로 모니터의 사진 지시에 따라 진행하며, 시퀀싱 후 분석 작업과 결과 리포트 제공 등은 서버의 Torrent Browser를 통해 연구자의 컴퓨터 화면에서 이루어진다[그림 4]. 브라우저에서는 각 run의 진행상황, 상세한 결과 보고서, 이미 완료된 run에 대해 reference나 분석 설정을 변경한 재분석 등을 손쉽게 수행할 수 있으며, 이러한 기능은 스마트폰으로도 제공되어, 언제 어디서나 실험의 진행 상태와 결과 등을 확인할 수 있다. 브라우저에서 제공하는 결과 리포트에서는 Q17(98%의 정확도), Q20(99%의 정확도)의 염기 숫자와 read 길이(한 번에 읽히는 염기서열 길이)의 평균과 최대치, 정렬된 후의 98%, 99%, 100% 정확도의 통

계 수치 등을 포함하여 실험이 잘 진행되었는지를 판단할 수 있는 다양한 정보와 논문이나 보고서 작성에 유용한 여러 가지 결과 정보도 상세하게 제공된다. 또한 SFF, FASTQ, BAM 파일 등을 얻을 수 있는 것은 물론, Broad Institute에서 개발된 유전체 가시화 툴인 IGV가 탑재되었기 때문에 mapping 결과도 눈으로 직접 확인할 수 있다. 이러한 분석과 결과 보고 전체는 서버의 전체 소프트웨어 업그레이드를 통해서도 향상되며 이미 버전 1.3과 1.4를 거쳐 10월에는 1.5가 제공된다. 이러한 업그레이드는 데이터의 생산량(throughput)을 늘리고, read length를 늘려주며, 데이터의 정확도를 높이면서도 전체 분석 시간은 오히려 단축시키는 효과를 가져온다[8].

ion PGM의 또 하나의 특징은 장비 전체를 바꾸는 것이 아니라, 반도체 칩의 교환만으로 극적인 성능 개선이 이루어진다는 것이다[그림 5]. 초기형인 314 Chip은 10 megabases 정도의 염기서열을 생산하는데, 집적도가 높아진 316 Chip은 100 megabases의 데이터를 생산하며 2011년 말에 나올 318 Chip은 1G(10억) bases의 데이터를 생산한다. 또한 read length도 평균 100 bases 정도에서 시작되어 200 bases 이상으로 계속 길어지고 있는 중이다. 그리고 위에서 언급한 서버의 업그레이드와 실험 프로토콜의 개선 등으로 같은 314 Chip에서도 더 향상된 결과가 얻어지며 이미 314 Chip에서 10 mega가 아닌 20~50 megabases의 결과를 얻었다는 연구자들의 사용후기가 계속 보고되고 있다. 이것은 사용자 관점에서 바람직한 현상인데 업체에서 일방적으로 제공하는 성능이 아닌, 먼저 사용하기 시작한 많은 연구자들이 적극적으로 공개한 결과 성능을 공개하여 확인할 수 있다는 점이다. 316 Chip이 처음 나왔을 때도 genomeweb의 2011년 6월 28일 기사에 따르면, Broad Institute에서 대장균 게놈에 대해서 한 번의 run으로 150 megabases 데이터 생성에 성공했으며, 전체 게놈에서 69개의 에러가 발견되는 정확도였다. 업체에서 제시하는 316 Chip의 공식 데이터 생산량이 100 megabases라는 점을 생각한다면 상당히 좋은 결과가 사용자 쪽에서 나온 셈이며 실제로 314 Chip, 316 Chip 모두 다양한 사용자들로부터 공식 성능의 2배를 넘는 결과들이 보고되고 있다고 소개되었다[9]. 2011년 9월 현재, 100Mb의 데이터를 생산하는 Ion 316 Chip까지 나온 상태에서 당연히 human whole genome sequencing을 하는 것은 무리이며 현재는 미생물이나 한정된 유전자에 대한 정보를 분석하는 targeted sequencing 등에 집중적으로 활용되고 있다. 하지만 ion torrent PGM 팀에서는 2011년 7월의 네이처 논문에서 3개의 미생물 유전체만이 아니라 인간, 그것도 무어의 법칙으로 유명한 바로 그 무어(G. Moore)의 유전체까지 시퀀싱했다고 보고했다[10]. 여기에서 인간의 유전체를 읽어냈다는 것은 무려 1000개가 넘는 314 Chip을 동원하여 수행한 것이기 때문에 당연히 비용면에서도 노력면에서도 권장할 수 있는 방법은 아니지만, 새로운 반도체 칩을 사용한 시퀀서가 향후 인간의 전체 유전체 분석으로까지 확대되는데 기술적으로는 아무런 어려움이 없다는 것을 증명했다. 또한 이 말은 단순 계산으로도 316 Chip을 사용하면 100개 정도, 연말에 나올 318 Chip을 사용하면 10개 정도만 사용하면 동일한 정도의 결과를 얻어낼 수 있다는 것도 의미한다는 점을 주목해야 한다.

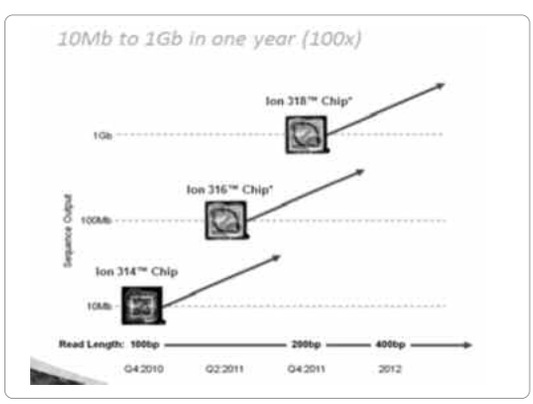

그림 5. Ion Chip을 바꾸는 것만으로 성능 개선이 이루어지는 ion torrent PGM.

NGS 분석에서는 그 정확성에 대한 정보도 중요하기 때문에 보고서에서 다양한 정확성 통계를 제공하는데, ion PGM에서는 한 가지 흥미로운 사실이 밝혀졌다. FASTQ 등의 raw data에서 보여주는 QV(Quality Value) 또는 QS(Quality Score)라는 것은 그 장비에서 염기를 읽어낼 때 기반이 된 신호가 어느 정도의 정확성을 가지는지 보여주는 일종의 예측된 수치이다. 그리고 reference 서열에 mapping을 한 후 (이미 정답을 알고 있으므로) 실제로 얼마나 정확하게 mapping 되었는지를 확인한 measured (empirical) QV를 낼 수도 있는데, ion PGM에서는 predicted QV가 너무 보수적이며 실제 정확도는 그것보다 훨씬 높다는 점이다. 이것은 업체와 사용자 그룹 모두에게서 보고되고 있으며 따라서 predicted QV만을 가지고 다른

NGS 기종과 비교하는 것은 정확하지 않다는 점을 주의해야 하며, 현 시점에서 이미 경쟁사의 기종보다는 훨씬 높은 정확도를 보인다고 평가된다[11, 12].

Ion PGM의 서버에서 제공하는 분석 소프트웨어를 사용하여 mapping과 결과 가시화 등이 가능하지만 그 외 추가로 NGS 시장에서 생물정보학 분석기능을 제공하던 다양한 상용 소프트웨어, 또는 무료로 다운로드 받아 사용할 수 있는 소프트웨어를 활용하는 것도 가능하다. CLC bio, Partek, NextGENe 등의 거의 모든 상용소프트웨어 툴들이 Ion PGM의 데이터에 대응하여 시장에 뛰어들었고, 기존의 454에서 사용되던 SFF 형식의 파일, NGS의 거의 표준적인 데이터 형식이라고 할 수 있는 FASTQ는 물론 reference 서열에 대한 mapping이 완료된 BAM 파일 형식에 대응되는 기존의 거의 모든 무료 소프트웨어들도 Ion PGM의 데이터 분석에 활용할 수 있다. 실제 인터넷 상에서는 동일한 데이터 세트에 대해서 다양한 소프트웨어로 분석을 하여 그 결과를 비교, 토론하는 일들이 벌어지고 있으며[13], 생물정보학 관련 논문들이 곧 나올 것으로 기대된다. 지금까지 소개한 Ion PGM에 대한 원리에서부터 상세한 사용법, 먼저 사용해본 유저들의 사용후기 등의 방대한 자료는 Ion Community 사이트에서 계속 축적되면서, 일반에게 공개되고 있다[12, 그림 6]. 장비를 미리 구입한 연구자들이 어떠한 활용을 하고 있고, 현재 어떠한 결과들을 내며 서로 축적된 지식을 공유하고 있는지, 향후 어떠한 개선이 이루어질 예정인지 등의 정보를 미리 얻는데 많은 도움을 준다[14].

그림 6. 모든 정보가 집중된 ion community 사이트.

저렴하고 손쉬우며 매우 신속한 시퀀싱을 가능하게 한 이러한 소형 시퀀서의 등장이 앞으로 어떠한 새로운 흐름을 만들어낼지를 정확하게 예측하는 것은 어려울 것이다. 하지만 Forbes에서는 ion torrent PGM 등장을 표지기사로 다루면서 흥미로운 예측을 소개했다. 현재 시퀀싱 장비와 관련 분석을 포함해서 40억 달러 정도인 시장이 ion torrent PGM과 같은 시퀀서의 등장을 계기로 암 진단과 건강 검진, 감염성 병원균의 추적조사, 농업 분야로의 유전자 시퀀싱 확대 등을 통해 향후 20년 이내에 1,000억 달러 시장으로 확대될 것이라는 내용이다[15]. 물론 각각의 분야에 관련된 정확한 유전자의 기능이 밝혀지고, 시퀀싱 후의 분석 정확도의 향상, 샘플 준비에서 최종 분석까지의 사용자 편의성이 더욱 높아져야 한다는 등의 전제 조건은 있겠지만, NGS와 관련된 가장 큰 어려움의 상당 부분이 해소된 것은 사실이므로 앞으로 누가 어느 방향으로 이러한 새로운 흐름을 개척하고 확장시켜 나갈 것인지 기대된다고 할 수 있다.

참고문헌

[1] http://www.genome.gov/sequencingcosts/
[2] Matthew N. Bainbridge et al., (2011) Sci Transl Med, Vol. 3, 87re3.
[3] Nature News - DNA sequencing for the masses (2010.12.14)
[4] The New York Times - Taking DNA Sequencing to the Masses (2011.1.4)
[5] Alexander Mellmann et al., (2011) PLoS ONE, Vol. 6, e22751.
[6] Holger Rohde et al., (2011) NEJM, Vol. 365, p718-724.
[7] Outbreak genomics (2011) Nature Biotechnology, Vol. 29, p769.
[8] Ion Torrent - Rapid Software Improvements (2011/9/19), http://biolektures.wordpress.com/2011/09/19/ion-torrent-rapid-software-improvements/
[9] http://www.genomeweb.com/sequencing/ion-torrent-releases-data-316-chip-early-access-customers-push-output
[10] Jonathan M. Rothberg et al., (2011) Nature, Vol. 475, p348-352.
[11] http://www.edgebio.com/blog/?p=271
[12] http://biolektures.wordpress.com/2011/09/05/ion-torrent-qv-prediction-algorithm/
[13] http://pathogenomics.bham.ac.uk/blog/2011/05/first-look-at-ion-torrent-data-de-novo-assembly/
[14] http://www.iontorrent.com/
[15] Forbes, On The Cover/Top Stories - Gene Machine (2011.1.17), http://www.forbes.com/forbes/2011/0117/features-jonathan-rothberg-medicine-tech-gene-machine.html