**Abstract** With the widespread adoption of next generation

sequencing technologies by the genetics community

and the rapid decrease in costs per base, exome sequencing

has become a standard within the repertoire of genetic

experiments for both research and diagnostics.

유전체 커뮤니티에 의한 폭넓은 NGS 기술의 채택과 base 당 비용의 급감으로, Exome Sequencing(이하 E.S) 은 연구와 진단 둘 다를 위한 유전체 실험 레퍼토리의 표준이 되었다.

Although bioinformatics now offers standard solutions for the analysis of exome sequencing data, many challenges still remain;

especially the increasing scale at which exome data are now

being generated has given rise to novel challenges in how

to efficiently store, analyze and interpret exome data of this

magnitude.

생물정보학이 E.S 분석의 표준 솔루션을 제공하지만, 아직 많은 과제들이 남아있다

; 특히 exome data 생성 규모의 증가는 어떻게 엄청난 규모의 exome data를 효과적으로 저장하고 분석, 해석할 지에 대한 과제를 남겼다.

In this review we discuss some of the recent

developments in bioinformatics for exome sequencing and

the directions that this is taking us to. With these developments,

exome sequencing is paving the way for the next big

challenge, the application of whole genome sequencing.

이 리뷰에서 우리는 E.S 를 위한 생물정보학의 최근의 발전과 이러한 발전이 우리를 어디로 데려가는 지에 대해 논한다. 이러한 발전을 통해, E.S 는 whole genome sequencing (이하 WGS) 의 적용 이라는 더 큰 과제로 향한다.

**Introduction**

Bioinformatics has been central to the analysis and interpretation

of exome sequencing data. Initial bioinformatics challenges concerned quality control; short read-mapping, variant calling

초기 생물정보학의 과제는 quality control 이었다

; short read-mapping과 variant calling

Most of these challenges have now been tackled to a degree that bioinformatics workflows are available to analyze and interpret

exomes in a standard fashion and provide workable results.

이런 과제의 대부분은 지금은 exome을 표준적으로 분석, 해석하고 유의미한 결과들을 제공하는 수준으로 올라왔다.

Some of the original hurdles have simply become less relevant with the progression of technology giving rise to more and higher

quality sequence data and longer sequence reads (e.g., the

ambiguous alignment of very short sequencing reads).

기존의 몇몇 문제들은 현재 기술이 더 긴 sequence read와 더 뛰어난 quality를 제공할 수 있도록 발전함으로써 해결되었다. (예를 들어, 아주 짧은 sequencing read의 모호한 alignment)

Nevertheless, quality control of exome sequencing data

still remains a necessity to guarantee reliable downstream

results. This task has now become fairly routine through

the development of several software packages that facilitate

the assessment of standard quality control measures for

exome sequencing.

그럼에도 불구하고, E.S data의 quality control은 여전히 신뢰할 수 있는 결과를 보장해야 할 필요가 있다. 이것은 이제 E.S를 위한 quality control 평가를 돕는 소프트웨어 패키지들의 개발에 꽤 일상적인 일이 되었다.

Although whole genome sequencing represents the ultimate genetic experiment, exomes still offer advantages in terms of costs, speed and

ease of data storage and analysis.

WGS가 유전체 실험의 최종 목표이지만, exome도 비용, 속도, 데이터 저장 용이, 분석 측면에서 장점을 가진다.

The steady increase of sequencing capacity and the widespread application of exome sequencing has allowed the sequencing of thousands of individuals and studies with hundred thousands

of exomes are already in progress.

Sequencing 용량의 꾸준한 증가와 E.S의 폭넓은 적용이 수천의 개개인을 sequencing할 수 있게 하여 수십만의 exome 연구를 진행하고 있다.

As an example, the Exome Aggregation Consortium (ExAC) collected a dataset of over 60,000 individuals and will grow even larger in the nearby future (Lek et al. 2015).

예를 들어, ExAC는 60000명 이상의 사람의 데이터를 모았고, 가까운 미래에는 더 커질 것이다.

**More data, more storage**

With growing datasets, simply storing data becomes a challenge

that all laboratories will at some point need to face.

데이터가 많아지면서, 데이터 저장 자체가 문제가 되었다.

Sequencing instruments typically generate FASTQ files

containing all individual sequencing reads. After alignment

the resulting reads are stored in the Sequence Alignment/

Map (SAM) format that describes where sequence reads

are mapped onto the reference genome.

Sequencing 기계들은 보통 한 사람의 모든 sequencing read를 담고 있는 FASTQ 파일을 생성한다. 결과 read를 alignment 하고 난 후에는 참조 게놈에 sequence read가 어디에 mapping 됬는 지를 보여주는 형식의 SAM 파일에 저장된다.

SAM files are usually compressed into the binary SAM (BAM) format that reduces the file size 3–4 times.

SAM 파일은 보통 이진 형식으로 저장되어 파일 사이즈가 3-4배 정도 축소된다. (BAM 형식)

The BAM format is currently the de facto standard format for aligned

reads and can be used by a large variety of downstream

analysis and visualization tools.

BAM 형식은 현재 align된 read의 사실상 기본 형식이고 다양한 분석 형태나 시각화 도구에 의해 사용된다.

Genomic variants that are subsequently identified based on the BAM file are then stored in a variants call format (VCF) .

BAM 파일을 기반으로 차후에 확인된 유전체 변이들은 VCF에 저장된다.

The typical size of a single exome BAM file is within the range of Gigabytes whereas the VCF file is usually no more than 100 MB.

보통 하나의 exome을 저장하는 VCF는 100MB가 넘지 않는 데 반해 BAM 파일은 GB 단위의 사이즈이다.

**Storing less**

The most straightforward method for reducing data storage

needs is by simply storing less data, or by removing data

as soon as possible.

데이터 필수 저장공간을 줄이는 가장 직접적인 방법은 적은 데이터를 저장하거나, 가능한 한 빨리 데이터를 지우는 것이다.

As an example, sequencing instruments currently only store raw images of the sequencing process for a limited time for trouble-shooting after which they are discarded.

예를 들어, Sequencing 기계들은 현재 폐기 후의 디버깅을 위해 제한적인 시간 동안 sequencing 과정의 raw images만 저장한다.

Similarly, many labs no longer keep the original raw sequencing reads (FASTQ file) after alignment since modern sequence aligners also include reads in the BAM file that are not aligned to the genome.

이와 같이, 현대의 sequence aligner들이 genome 에 align 되지 않은 BAM 파일에 read를 포함하기 때문에, 많은 실험실들은 더 이상 alignment 이후 FASTQ 파일을 저장하지 않는다.

This adds a little bit to the size of the BAM files but there is no longer any need for storing FASTQ files, since raw reads can now be

regenerated from the alignment files by tools like Picard and SAMtools.

이것은 BAM 파일의 사이즈를 조금 증가시키지만 더 이상 FASTQ 파일을 저장할 필요가 없다. 왜냐하면 Raw reads가 Picard 나 SAMtools 같은 도구에 의해 alignment files로부터 재생성되기 때문이다.

This potentially reduces storage requirements by half. In addition to this, several clinical guidelines have been proposed that allow diagnostic laboratories to remove the alignment files after 1 or 2 years.

이것은 잠재적으로 필요한 저장 공간을 반으로 줄인다. 추가적으로, 몇 개의 임상적 가이드라인은 1-2년 후에 진단 분야 실험실에서 alignment 파일 삭제를 허용하는 것을 제안한다.

However, although VCF files contain the primary result of the experiment it is worthwhile to keep BAM files for future analysis since they contain much more information than VCF files, for example the identification of CNVs, somatic mutations, and mitochondrial DNA variation.

VCF 파일은 실험의 가장 중요한 결과를 담고 있지만, BAM 파일이 VCF 파일보다 더 많은 정보( 예를 들어 identification of CNVs, somatic mutations, and mitochondrial DNA variation )을 담고 있기 때문에 차후 분석을 위해 BAM 파일을 가지고 있을 만한 가치가 있다.

It is not uncommon that reanalysis of FASTQ or BAM files can identify additional variants that were initially missed.

FASTQ 파일이나 BAM 파일을 재분석하여 초기에 놓친 추가적인 변이들을 찾아내는 경우는 흔하지 않다.

**Compression**

An alternative to the straightforward removal of large files to save space is data-compression. This has already been introduced for raw sequence files that are now by default compressed with gzip.

공간 절약을 위해 대용량 파일들을 삭제하는 방법을 대체하는 방법은 데이터 압축이다. 이 방법은 이미 raw sequence 파일에서 소개되었으나 지금은 기본적으로 gzip으로 압축된다.

Although the SAM/BAM format is convenient in the sense that it contains almost all information of the original reads and all details about the alignment in an intuitive fashion, it was not designed for efficient storage.

SAM/BAM 형식이 기존의 read들과 직관적으로 alignment 에 대한 모든 디테일 정보를 거의 포함하지만, 저장 측면에서 효과적이지는 않다.

Since BAM files are already in binary format, ordinary compression

algorithms cannot significantly reduce their size.

BAM 파일은 이미 이진 형식이기 때문에, 보통의 압축 알고리즘은 효과적으로 사이즈를 줄이지 못한다.

However, specialized compression tools use various techniques to further reduce the size of BAM files. First of all non-essential

information, e.g. read identifiers, can be removed.

그러나, 특별한 압축 도구는 BAM 파일의 사이즈를 훨씬 축소시킨다. 첫째로, read identifier 같은 중요하지 않은 정보들을 제거한다.

Secondly, the majority of the exome will be the same as the reference

genome and can be stored more efficiently: Reference-based

compression encodes reads based on a reference sequence

and stores only positions that differ from the reference

sequencing.

둘째로, exome 의 대다수는 reference genome과 같고 더 효과적으로 저장될 수 있다 : 참조 기반 압축 read를 reference genome에 기반하여 인코딩하고 reference sequencing 과 다른 위치만 저장한다.

For regions where there are no differences to the reference

genome, only coordinates and depth information are

retained.

Reference genome과 다르지 않은 부분에 대해서는, 좌표와 깊이 정보만 유지한다.

Lastly, individual base quality scores (or Q scores)

are typically encoded as PHRED-like scores within a range

of 0–40. These quality scores are used to optimize read-mapping and variant calling.

마지막으로, Q score는 보통 0 – 40 범위의 PHRED-like score 와 같이 인코딩되고, 이러한 Q score는 read mapping 과 variant calling을 최적화하기 위해 사용된다.

However, the scale of quality scores is very fine-grained and encoding

Q scores into bins reduces storage space. Binning quality scores often results in compression with some loss of information (lossy compression), where the original quality scores lose precision during

compression.

그러나, Q score의 규모는 굉장히 작고 Q score를 이진수로 인코딩하는 것은 저장 공간을 절약한다. 이것은 압축 과정에서 기존의 q score의 정밀성을 떨어뜨려 데이터 손실을 초래할 수 있다.

The lost precision does, however, not necessarily result in significant loss of accuracy for variant calling. Based on these approaches, alternative formats such as Goby, SlimGene , CRAM and DEEZ, have been introduced that attempt to keep as much of the original information yet at a lower cost of disk space than BAM.

그러나 정밀성이 떨어지는 것이 꼭 variant calling의 정확성을 크게 떨어뜨리는 것은 아니다. 이러한 방식에 기반하여 Coby, SlimGene, CRAM , DEEZ 와 같은 대체 형식들이 도입되어 BAM 보다 적은 비용으로 원본 정보를 최대한 유지하려고 한다.

In particular, the CRAM format has gained a lot of traction. Compression of a BAM file to CRAM format with the Scramble tool resulted in file reductions of 38–55 % with a compression time of a few minutes.

특히, CRAM 포맷은 많은 인기를 끌고 있다 . 스크램블 도구를 이용하여 BAM 파일을 CRAM 형식으로 압축하면 몇 분도 안돼서 38—55 퍼센트까지 축소 시킬 수 있다.

CRAM compression has already been applied to tackle storage-capacity problems in large databases such as Sequence Read Archive (SRA)and the 1000 Genomes project.

CRAM 압축은 Sequence Read Archive 나 1000 Genomes project 같은 대용량 데이터 베이스의 저장 용량 문제를 해결하기 위해 이미 사용되고 있다.

The CRAM format is now supported by some widely used genomic analysis tools such as SAMtools, Picard and GATK. With the increasing support for the CRAM format, it may well replace the use of BAM files in the near future.

CRAM 형식은 SAMtools, Picard, GATK 같은 유명한 유전체 분석툴에 지원된다. CRAM 형식을 지원하는 툴이 증가하고 있기 때문에, 곧 CRAM 형식이 BAM 파일을 대체할 수도 있다.

With ever growing datasets containing variants of thousands of individuals, it becomes worthwhile to compress the relatively small VCF files as well. The Tabix format offers a convenient compression format for large VCF files.

수많은 사람들의 변이 정보를 담고 있는 데이터는 끝없이 증가하기 때문에, 상대적으로 크기가 작은 VCF 파일 또한 압축할 가치가 있게 되었다. Tabix는 큰 VCF 파일은 편리하게 압축하는 형식이다.

This reduces file sizes roughly 3–5 times, and also supports

indexing to perform efficient querying of genome positions. Some common resources are already available in Tabix format such as dbSNP and Combined Annotation-Depended Depletion (CADD) scores.

이 형식은 파일을 대략 3-5배 축소시키고, 유전자 위치를 효과적으로 질의하기 위한 인덱싱도 지원한다. dbSNP 나 CADD score 같은 보통의 리소스들은 이미 Tabix를 지원한다.

Another option is to use the recently published Genotype Query Tools (GQT) to index and compress large number of VCF files. This tool

also facilitates fast querying.

다른 방법은 VCF를 인덱싱하고 압축하기 위해 최근 출시된 GQT를 사용하는 것이다. 이 툴은 query를 빠르게 만들어 준다.

GQT was used to compress the Exome Aggregation Consortium (ExAC) VCF file, consisting of 9.36 million exonic variants for 60,706 individuals, from 14.1 TB to 28 GB.

GQT는 약 6만명의 사람에게서 추출한 9백만개 이상의 exonic 변이 들로 구성된 ExAC VCF 파일을 14.1 TB에서 28GB까지 압축하는 데 사용되었다.

**Cloud‑based solutions**

Compression of data is an easy and efficient way to reduce storage needs, but in the end the reduction in data sizes is limited. An alternative is to store large amounts of genomics data in the cloud.

데이터 압축은 필요한 저장 공간을 줄이는 쉽고 효과적인 방법이지만, 결국에 줄일 수 있는 데이터의 크기는 한정적이다. 이를 개선하기 위해 클라우드에 유전자 데이터를 저장한다.

Cloud storage offers several out-of-thebox advantages to local storage: it is scalable, has default access control policies, protects against data loss, allows for auditing, data encryption, easy sharing, and automation by programmable interfaces.

클라우드는 로컬 저장공간에 비해 독창적인 장점을 가지고 있다: 확장성, 기본 접근 컨트롤 정책이 있고, 데이터 손실을 방지하고, 검사 가능하고, 데이터 암호화, 공유가 쉽고, 코딩 가능한 인터페이스에 의해 자동화된다.

Currently, there are multiple commercial providers of cloud services of which Amazon Web Services (AWS; https://aws.amazon.com/), Microsoft Azure (<https://azure>. microsoft.com) and Google cloud platform (<https://cloud>. google.com) are the largest.

현재 Amazon Web Services, Microsoft Azure, Google Cloud 등의 클라우드 서비스를 제공하는 큰 회사들이 있다.

Cloud storage is based on a “pay as you go” monetary model whereby one only pays for used storage that can be expanded ad hoc. Although cloud storage itself is relatively inexpensive with less than $100 for storing 1 TB of data per month, there are some additional costs to consider.

클라우드는 즉각적으로 확장 가능한 저장공간에 대해서만 돈을 내는 선불 모델에 기반한다. 비록 클라우드 자체가 1테라에 월 10만원도 안되는 가격으로 상대적으로 비싸지 않지만, 고려해야 하는 추가 비용이 있다.

While transferring data into the cloud storage is usually free of costs, analyzing and downloading data from the cloud can be relatively expensive.

클라우드로 데이터를 전송하는 것은 보통 공짜지만, 클라우드에서 데이터를 다운받아서 분석하는 것은 상대적으로 비싸다.

For example, downloading 1 TB of data from the cloud costs approximately $120 per TB. This makes it worthwhile not only to keep data in the cloud but also to perform the analysis there and only download smaller result files.

예를 들면, 클라우드에서 1TB를 다운받는 것은 TB 당 12만원 정도가 든다. 데이터를 클라우드에 올리고 분석을 수행하여 작은 결과 파일만 다운로드 하는 것이 좋다.

Special software is, however, needed to make efficient use of the scalability of the cloud-computing platform. Currently there are already a variety of tools for cloud-based mapping of sequence reads,

, genotyping, variant annotation as well as complete cloud-based exome sequencing pipelines.

그러나 클라우드 컴퓨팅 플랫폼의 확장성을 효과적으로 사용하기 위해서는 특별한 소프트웨어가 필요하다. 현재 클라우드 기반으로 Sequence read 매핑, Genotyping, variant annotation을 완전 클라우드 기반의 E.S pipeline 처럼 잘해주는 다양한 툴들이 있다.

showed that the alignment of the entire genome (4 billion paired reads, 35 pb long) of a person in 48 h costing approximately $48

of cloud resources (Fusaro et al. 2011).

1 사람의 전체 유전자 ( 40억 쌍의 read, 35 pb 길이) 의 alignment는 48시간이 걸리고 5만원 정도의 cloud 자원이 필요하다.

the International Cancer Genome Consortium (ICGC) analyzed 500 genomes in the cloud for a price of $18 per sample whereas the authors estimate this would require $200 on standard computer systems.

ICGC는 샘플 당 2만원 상당의 클라우드에서 500개의 유전자를 분석한 반면, 저자는 이 과정이 보통 컴퓨터에서는 20만원 정도가 요구될 것으로 추정한다.

Data stored in the cloud can also provide a solution for effective public and private data-sharing. For example, the Amazon Web Services (AWS) contains 1000 Genomes Project data (Clarke et al. 2012) and accommodates 1200 whole genome sequences of the ICGC.

클라우드에 저장된 데이터는 효과적인 공동, 개인의 데이터 공유를 제공한다. AWS는 1000개의 게놈 프로젝트 데이터를 가지고 있고 ICGC의 1200개의 전체 유전자 sequence를 가지고 있다.

Furthermore, the US National Cancer Institute is exploring how the cloud could facilitate a cost-effective platform to store and share

large amounts of tumor data.

추가로, US 국가 암 센터는 엄청난 양의 암 데이터를 저장하고 공유하는 데에 클라우드가 어떻게 비용적으로 효과적인 플랫폼을 만드는 지 연구중이다.

The uptake of cloud-based solutions by academic and non-academic hospital laboratories has been slow, likely because of practical concerns, unfamiliarity, as well as ethical and legal concerns of storing patient DNA data in the cloud.

병원 실험실에서 실용적인 문제와 생소함, 또한 환자의 DNA를 클라우드에 저장하는 것에 있어서 윤리적, 법적 문제들 때문에 클라우드 기반 솔루션을 느리게 차용해 가고 있다.

Although data storage in the cloud is relatively inexpensive, transferring vast quantities of sequencing data via the Internet from and into the cloud may be a considerable cost and a time-consuming

process due to low network bandwidth.

클라우드 데이터 저장은 상대적으로 싸지만, 인터넷을 통해 클라우드에 엄청난 양의 데이터를 전송하고 받아오는 것이 네트워크 대역폭이 낮기 때문에 굉장히 비싸고 시간 낭비일 수 있다.

In addition, moving genetic data of patients to a third-party server introduces issues concerning security and privacy.

게다가, 환자의 유전자 정보를 제 3의 서버에 옮기는 것이 보안과 사생활의 문제를 야기한다.

This has limited the use of cloud-based storage solutions for most clinical NGS applications so far. However, given the advantages and a future of routine genome sequencing, it may well be unavoidable that all genomics data end up in the cloud for analysis and for patients and their physicians to access.

이것은 대부분의 NGS 적용 임상실험에서 아직까지는 클라우드 기반의 저장 솔루션을 사용하는 데 걸림돌이 되었다. 그러나, 장점들과 미래 유전자 sequencing의 방향을 생각한다면, 모든 유전자 데이터가 분석과 접근을 위해 클라우드에 저장되는 것은 피할 수 없을 것이다.

**Variant identification**

To some extent, challenges for calling variants have become less urgent with improved exome enrichment assays, increasing sequence quality and read length and reduced sequencing prices, allowing for higher coverage sequencing of the exome in individual patients.

변이를 calling 하는 것은 exome enrichment 시험의 개선, sequence의 퀄리티와 read 길이 증가, sequencing 비용 감소, 각 환자의 exome sequencing 을 증가를 통해 어느 정도 급하지 않은 문제가 되었다.

Whereas early comparisons of whole exome capture kits showed

that around 80 % of the human protein coding sequence regions were captured at a minimal coverage of 20×, current exome capture kits and sequencing at minimal 100× median average coverage capture more than 95 % of the coding regions with a minimal coverage of 20×.

Due to the increased coverage and improved sequencing quality for modern exomes, variant calling has become more reliable. Several studies have even demonstrated the identification of somatic mutations for rare , which is only possible with high coverage.

Coverage의 증가와 현대 exome의 sequencing quality 개선으로, variant calling은 더 신뢰할 수 있게 되었다. 몇 개의 연구에서는 심지어 high coverage에서만 가능한 희귀 somatic 돌연변이를 발견하기도 했다.

These improvements in exome quality have led some laboratories to reconsider the validation of sequencing variants by the gold standard Sanger sequencing.

이러한 Exome quality의 발전들은 몇몇의 실험실들이 Sanger sequencing의 표준으로 변이 sequencing 의 유효성을 체크하는 것을 재고하게 만들었다.

A recent validation study found that all single nucleotide variants

with a Genome Analysis Toolkit (GATK) quality score above 500 were confirmed by Sanger sequencing and estimated that only validating variants with a quality score below this threshold would reduce the Sanger confirmation workload by 70–80 %.

최근 유효성 확인 연구는 GATK quality score가 500 이상의 모든 단일 뉴클레오타이드 변이들이 Sanger Sequencing 에 의해 검증되었고, GATK quality score가 이 이하인 변이들의 유효성을 검증함에 있어서 Sanger는 70-80 퍼센트만 검증할 수 있을 것이라고 추정했다.

Overall the significant improvements in exome sequencing quality may indeed eliminate the need for validation of high quality variants. However, the detection of SNVs in NGS data has not been fully resolved and results from different variant callers remain inconsistent.

Exome sequencing quality 의 큰 발전은 높은 quality의 변이에 유효성 검사를 할 필요를 없앴다. 그러나 NGS 데이터의 SNV 발견은 완전히 해결되지 않았고, 다른 variant caller 로부터 의 결과는 아직 일정하지 않다.

In addition, small insertions/deletion (indels) are still particularly

problematic to identify accurately. Highly accurate genotypes across the genome of a single individual as for example provided by the “Genome in a Bottle Consortium” may help resolve these issues in the future.

게다가, 작은 삽입/삭제들은 아직 정확한 확인에 있어서 특별히 문제가 되지는 않는다. 예를 들어 “ ”에 의해 제공된 한 사람의 유전자 전체에 거쳐서 존재하는 아주 정확한 유전자형들은 곧 이러한 문제를 해결할 수 있다.

**Detection of copy number variants**

From SNVs attention has moved towards the identification of other types of variants in exome sequencing data. In particular, the identification of copy number variation (CNV) from exome data poses an attractive possibility as CNVs are an important cause of disease.

exome sequencing 데이터에서, 관심은 SNVs에서 다른 유형의 variant들을 확인하는 방향으로 옮겨졌다. 특히, exome 데이터로부터의 CNV를 확인하는 것은 CNV가 질병의 중요한 원인이기 때문에 좋은 가능성을 제시한다.

Genomic microarray platforms such as the SNP-array and Array CGH are the de facto standard to detect CNVs, whereas whole genome sequencing will likely be the preferred platform for the detection and characterization of CNVs as well as other structural variants.

SNP 배열 및 배열 CGH 같은 유전자 마이크로 배열 플랫폼은 CNV를 검출하기 위한 사실상의 표준이지만, 전체 유전자 sequencing은 CNV 및 기타 구조 변형의 검출 및 특성화를 위한 바람직한 플랫폼이 될 것이다.

In contrast to microarrays and whole genome sequencing, exome sequencing targets only 1–2 % of the protein coding regions of the genome.

마이크로 배열 및 전체 유전자 sequencing과는 달리, exome sequencing 은 유전자의 단백질 코딩 영역의 1-2%만을 목표로 한다.

The sparse and fragmented nature of exome data makes it more difficult to identify CNVs and methods rely mostly on depth-of-coverage approaches.

Exome 데이터가 드문드문 분열되어 있어서 CNV를 식별하는 것이 더 어렵게 만들고, 방법들이 주로 심도 있는 coverage 접근 방식에 기반한다.

For these approaches a normalized read count in a genomic window of a single individual is compared to that of other exomes.

이러한 접근법에 대해 단일 개체의 genome window에서 정규화 된 read count가 다른 exome의 read count에 비교된다.

Normalization of read counts is required to counteract technical issues such as poor read mappability, GC bias, and batch effects between sequencing experiments.

Read count의 정규화는 낮은 read 매핑 가능성, GC 편향 및 sequencing 실험 간의 배치 효과와 같은 기술적 문제를 보완해야 한다.

Many different algorithms have been devised based on read-depth methods, such as CODEX, Convex, Conifer, and XHMM.

Codex 등과 같은 깊이 있는 방법을 기반으로 많은 알고리즘 들이 고안되었다.

Comparisons of CNV algorithms for exome data have shown that

none of the algorithms performed well in all situations and that the resolution is limited to at least three exome targets.

Exome 데이터에 대한 CNV 알고리즘의 비교는 알고리즘 중 어느것도 모든 상황에서 잘 수행되지 않고, 해상도가 적어도 exome 3개 정도로 제한된다.

Although this does not equal the sensitivity of high resolution microarrays, it is comparable to that of medium resolution microarrays that are still commonly used.

이것은 고해상도 마이크로 배열의 감도와 같지 않지만 여전히 일반적으로 사용되는 중간 해상도의 마이크로 배열의 감도와 비슷하다.

Studies describing the large-scale application of CNV detection

only based from exome data are, however, still limited, which may perhaps hint at some of the underlying difficulties to obtain robust CNV calls from exome data.

오직 exome 데이터 기반의 CNV 탐지의 대규모 적용을 기술하는 연구들은 아직 제한적이기 때문에 exome 데이터로부터 robust한 CNV call들을 얻기 힘들 수 있다.

The possibility to detect copy number variants in exome data is, however, a great benefit of exome sequencing that should not be ignored as CNVs contribute significantly to disease.

그러나 exome 데이터에서 copy number variants들을 검출할 가능성은 CNV가 질병에 상당히 기여하기 때문에 무시되어서는 안되는 exome sequencing 의 중요한 부분이다.

The identification of other types of structural variants such as inversions, and the accurate prediction of CNV breakpoint remains challenging and whole genome sequencing is likely needed for this.

역전과 같은 다른 유형의 구조 변형 확인과 CNV 중단점의 정확한 예측은 여전히 어려운 일이고, 전체 유전자 sequencing 이 필요할 수 있다.

**Variant interpretation**

Sequencing the protein coding regions of a patient typically yields tens of thousands of variants of which the majority is likely to be benign and only one or perhaps two variants contribute to the disease phenotype.

환자의 단백질 코딩 부위를 sequencing하는 것은 일반적으로 수만 가지의 변이를 생성하며, 그 중 대다수는 양성일 것이고, 오직 한 두개 정도만 질병 phenotype에 기여한다.

The most effective way of distinguishing benign from pathogenic variants is based on using population frequencies of variants. For

this approach all variants occurring in the population at higher frequencies than the disease prevalence are considered as benign.

양성 종을 병원성 변이와 구별하는 가장 효과적인 방법은 변이의 모집단 빈도를 이용하는 것이다. 이런 접근 방식에서는, 질병 유병률보다 높은 빈도로 집단에서 발생하는 모든 변이를 양성으로 간주한다.

Databases with collections of exome variants of individuals without clear disease phenotypes have therefore been tremendously helpful to prioritize variants in Mendelian disease.

따라서 명백한 질병 phenotype이 없는 개체의 exome 변이를 모든 DB는 Mendelian 질병의 변이 우선순위를 정하는 데 큰 도움이 되었다.

This has given rise to several initiatives for large-scale variant databases with exome data. The largest of these, thus far, is the Exome Aggregation Consortium (ExAC) database, containing variants of more than 60,000 exomes.

이로 인해 exome 데이터가 있는 대규모 변이 DB에 대한 몇 가지 initiative가 생겨났다. 현재까지 최대 규모인 ExAC DB는 6만개 이상의 exome 변이를 포함하고 있다.

These large databases are instrumental to the interpretation of future exomes for Mendelian disease gene identification. In addition, a need for population-specific databases of variation will remain, especially

for those populations that are poorly represented in the large public databases.

이러한 큰 DB들은 Mendelian 질병 유전자 식별을 위한 미래 exome 해석에 도움이 된다. 또한, 특정 인구 집단의 변이 DB 에 대한 필요성은 특히 대형 공공 DB에서 제대로 표현되지 않은 모집단에 경우 여전히 남아있을 것이다.

Interpretation based on population frequency information from

databases should be done with care because of the possibility of false positives, founder mutations, somatic or tissue-specific variants.

거짓 긍정, founder 돌연변이, somatic 이나 조직 특이적 변이의 가능성 때문에 DB의 모집단 빈도 정보를 기반으로 한 해석은 주의해서 처리되어야 한다.

**Coding mutations**

Although accurate population frequencies are a necessity for the interpretation of exomes, this will only reduce the number of possible candidate mutations to a couple of hundred or so.

정확한 인구 빈도가 exome의 해석에 필요하지만, 이건 단지 가능한 후보 돌연변이 수를 2백개 정도로 줄일 뿐이다.

Further prioritization of pathogenic variants remains a challenging task, in particular for missense variants.

병원성 변이의 우선 순위 결정은 특히 missense 변이에 대한 도전 과제로 남아있다.

Various tools such as Polyphen2, SIFT and PhyloP, have long been used in the pathogenicity assessment of these protein coding variants based on evolutionary conservation.

Polyphen2, SIFT 및 PhyloP와 같은 다양한 도구는 오래 전부터 진화 보전에 기초한 단백질 코딩 변이의 병원성 평가에 사용되어 왔다.

Unfortunately, these prioritization methods lack specificity and sensitivity to sufficiently reduce the large number of candidate mutations from exome sequencing on their own.

불행히도, 이러한 우선 순위 방법은 특이성과 감도가 결여된 방법으로 다수의 후보 돌연변이를 E.S 에서 자체적으로 많이 줄인다.

This becomes even more apparent when considering the prioritization of non-coding variation from whole genome sequencing experiments.

이것은 전체 유전자 sequencing 실험에서 비코딩 변이의 우선순위를 고려할 때 더 분명해진다.

However, in the last few years, novel tools have been published that are expected to offer better sensitivity and specificity compared to the traditional prioritization tools.

그러나 지난 몇 년 동안 기존의 우선 순위 지정 도구에 비해 향상된 민감도와 특이성을 제공할 만한 새로운 도구가 출시되었다.

The availability of genome-wide functional annotations of coding and non-coding variants in combination with algorithmic improvements resulted in novel tools adapted to prioritize both coding and non-coding variants.

새로운 툴은 알고리즘 개선과 함께 코딩/비코딩 변이의 유전자 전체 기능 주석을 사용하여 변이의 우선순위를 결정할 수 있다.

These novel tools can broadly be divided into two groups. The

first group focuses on the prediction of deleterious variation by computing a functional meta-score based on integrating a variety of genome-wide annotations.

이 새 툴은 크게 두 그룹으로 나눌 수 있다. 첫 번째 그룹은 다양한 유전자 전체 추적을 통합하여 기능적 meta-score를 계산하여 유해한 변화 예측에 중점을 둔다.

Combined Annotation-Depended Depletion (CADD) is the most wellknown example of such a framework that applies a support

vector machine (SVM) to integrate 63 sources of functional

and evolutionary data into a relative pathogenicity score.

CADD는 이러한 63개의 기능적이고 진화적인 데이터를 상대적 병원성 점수로 통합하는 서포트 벡터 머신을 지원하는 프레임워크의 예로 가장 잘 알려져있다.

Eigen and DANN are other examples using different algorithmic approaches to combine large varieties of annotations into one pre-computed meta-score trained to distinguish between benign and deleterious variants.

Eigen과 DANN 은 여러가지 종류의 주석을 양성 변이와 유해 변이로 구별하기 위해 이미 계산된 메타 점수로 결합하는 다른 알고리즘을 사용하는 예이다.

Fitness consequence (FitCons)is different in the sense that it compares patterns of divergence between the human population and primates to assess functional sites that emerged quite recently.

FitCons은 최근에 등장한 기능적 site를 평가하기 위해 인간과 영장류 간의 발산 양상을 비교한다는 점에서 차이가 있다.

The second group of tools attempts to specifically predict non-coding regulatory variants. DeepSEA and DeltaSVM are examples of such tools and were trained on a variety of annotations of non-coding

annotations mainly derived from the ENCODE project.

두 번째 그룹의 도구는 비 코딩 규제 변이를 구체적으로 예측한다. DeepSEA 및 DeltaSVM은 이에 대한 예이고, 주로 ENCODE 프로젝트에서 파생된 비코딩 주석의 다양한 주석에 대해 훈련받았다.

Notably, the DeepSEA method was based on a Deep learning algorithm, which is a form of machine learning that is increasingly

being applied to biological problems.

특히, DeepSEA 방법은 생물학에서 점점 더 많이 적용되어 가고 있는 기계학습의 한 형태인 딥러닝 알고리즘을 기반으로 한다.

In spite of the potential of these tools, it remains unclear how well they perform in clinical practice because independent validation studies for these novel tools are still lacking.

이러한 도구의 잠재력에도 불구, 새로운 도구에 대한 독립적인 검증 연구가 여전히 부족하기 때문에 임상 수행에서 얼마나 잘 수행될 지는 미지수이다.

Moreover, such studies are hampered by a lack of sufficient validation data that have not already been used in the development of the prediction software or original benchmark.

더욱이 이런 연구는 예측 소프트웨어나 기존 벤치마크의 개발에 아직 사용되지 않은 유효성 확인 데이터의 부족으로 방해 받는다.

Van der Velde et al. demonstrated the practical utility of CADD for the interpretation of variants.

“ ”등이 변형의 해석을 위한 CADD의 실제적인 유용성을 보였다.

The authors applied CADD to a set of 2210 variants that were manually assessed by an expert panel and found that, beside a relatively small number of discrepancies in favor of the expert, CADD scores proved valuable for the prioritization of pathogenic variants.

저자들은 전문가 패널이 직접 평가한 2210가지 변형의 집합에 CADD를 적용하였고, CADD 점수가 상대적으로 낮은 것들을 제외하고는 CADD 점수가 병원성 변이의 우선순위 지정에 중요하다고 밝혔다.

However, a recent validation of CADD and other prediction tools

using in vivo mouse models, found that about half of the assessed mutations that were predicted to be deleterious had little impact on the clinical phenotype.

그러나 최근 vivo 생쥐 모델을 통해 CADD 및 기타 예측 도구의 검증하여 유해 변이로 예측된 돌연변이의 절반이 임상적으로 표현형에 거의 영향을 미치지 않는 것을 알아냈다.

This once again highlights the importance of functional validation

of potential pathogenic variants.

다시 한번 잠재적인 병원성 변이의 유효성 검사의 중요성을 강조한다.

**Splice site mutations**

Due to the increased read lengths, exome sequencing typically captures a large part of the extended splice site at sufficient coverage for variant identification.

증가된 read 길이로 인해, E.S는 보통 변이 식별을 위해 충분한 coverage에서 확장된 splice site의 큰 부분을 캡쳐한다.

However, mutations in the extended splice site are typically excluded

during the prioritization step because variation within these regions is more prevalent but also more difficult to interpret.

그러나, 확장된 splice site부위의 돌연변이들은 이 부위의 변이가 더 널리 퍼지기는 하지만 해석이 어렵기 때문에 우선순위 지정 단계에서 보통 배제된다.

Existing algorithms for splice sites such as MaxEntScan and NNSplice were not designed to offer predictions for the large numbers of variants from exome sequencing.

MaxEntScan 이나 NNSplice 같은 splice site에 대한 기존 알고리즘은 E.S 에서 많은 수의 변이에 대해 예측하도록 설계되지 않았다.

Like for coding variants recent developments in algorithms have improved the ability to interpret this type of variants.

Like for 코딩 변이, 알고리즘의 최근 개발은 이런 타입의 변이를 해석하는 능력이 향상되었다.

The SPANR (splicing-based analysis for variants) tool is another example of a “deep learning” computational model scoring the effect of variants on the mRNA-splicing.

SPANR 도구는 변이가 mRNA 접합에 미치는 영향을 평가하는 딥러닝 계산 모델의 또 다른 예이다.

The SPANR model is trained on 1393 sequence features extracted from 10,689 alternatively spliced exons and their corresponding mRNA expression levels in 16 human tissues and offers predictions up to 300 bp within the intron.

SPANR 모델은 16개의 인간조직에서 10689개의 선택적으로 splicing 된 exon 및 해당 mRNA 발현 수준에서 추출한 1393개의 서열 특징에 대해 훈련되었고, intron 내에서 최대 300 bp를 예측한다.

The authors found that SPANR correctly predicted the direction of

change in expression of the exon for 73 out of 99 (74 %) splice site mutations.

저자들은 SPANR이 99개의 splice site 돌연변이 중 73개에서 exon 의 발현이 변화를 맞게 예측했다고 밝혔다.

Another novel splice site prediction tool called hexamer additive

linear (HAL) is a model, trained on nearly two million synthetic alternatively spliced mini genes, to predict the effect of 5′ and 3′ mutations on exon skipping.

HAL 이라고 불리는 또 다른 splice site 예측 도구는 exon skipping 에서 5’와 3’의 돌연변이의 효과를 예측하기 위해 약 200만개의 교대로 합성된 spliced mini genes에 대해 훈련된 모델이다.

In a set of 286 variants within three genes (*CTFR*,*BRCA2* and *SMN2*) the prediction accuracy ranged from 86 to 90 %. These improvements in splice site prediction programs may open up new avenues for the interpretation of variants in exomes.

3개의 유전자 (CTFR, BRCA2 and SMN2) 내의 286개의 변이 세트에서 예측정확도는 86-90 % 이었다. Splice site 예측 프로그램의 이러한 개선이 exome의 변이 해석에 새로운 길을 열 수도 있다.

**Gene prioritization**

For the interpretation of exome data it is not sufficient to only determine whether a variant is likely to impair normal gene function, but also whether the function of a mutated gene is actually relevant for the disease.

Exome 데이터의 해석을 위해 변이가 정상적인 유전자 기능을 손상시킬 가능성이 있는지 여부 뿐 아니라 돌연변이 유전자의 기능이 실제로 질병과 관련이 있는지 여부를 판단하는 것도 충분치 않다.

Two novel approaches for the interpretation of gene function have gained a lot of traction.

유전자 기능의 해석을 위한 두가지 새로운 접근법은 많은 인기를 끌었다.

Phenotypic interpretation of variants in exomes (PHIVE) is an algorithm that computes phenotype similarity between human disease phenotypes and phenotype information from knockout experiments in model organisms.

PHIVE는 모델 생물체의 knockout 실험으로부터 인간 질병 표현형과 표현형 정보간의 표현형 유사성을 계산하는 알고리즘이다.

This gene-level information is then combined with variant pathogenicity predictions and thereby achieves better rankings of pathogenic variants in exome data.

이 유전자 수준의 정보는 변이 병원성 예측과 결합되어 exome 데이터에서 병원성 변이의 더 나은 순위를 얻는다.

A totally different approach to predict deleteriousness for genes is based on the use of population variation to determine how intolerant

genes are to normal variation.

유전자에 대한 해로운 영향을 예측하는 완전히 다른 접근법은 유전자가 정상적인 변이를 얼마나 내성을 가지는 지 결정하는 모집단 변이의 사용에 기반한다.

Two studies independently showed that human disease genes are much more intolerant to genetic variation than other genes.

두 연구는 독립적으로 인간의 질병 유전자가 다른 유전자 보다 유전적 변이에 내성이 없다는 것을 보여주었다.

Several studies have already successfully used this approach to prioritize genes with likely pathogenic mutations.

여러 연구에서 이미 병원성 돌연변이 가능성이 있는 유전자를 성공적으로 우선 순위가 하기 위해 이 접근법을 사용해 왔다.

Overall, algorithm development has leveraged the availability of genome-wide datasets such as exome sequencing project (ESP), encyclopedia of DNA elements (ENCODE) and the International Mouse Phenotype Consortium (IMPC) to provide improved pathogenicity predictions for variants and to cope with exome-sized variant datasets.

전반적으로 알고리즘 개발은 ESP, ENCODE 및 IMPC 같은 유전자 전체 데이터셋의 가용성을 변이에 대한 개선된 병원성 예측을 제공하고 exome 크기의 변이 데이터셋을 사용하는 데에 활용하려고 한다.

These novel algorithms represent our first steps to the next big challenge, the interpretation of non-coding variation from whole genome sequencing experiments.

이런 새로운 알고리즘은 전체 유전자 sequencing 실험에서 비 코딩 변이의 해석이라는 next big 과제의 첫 단계를 나타낸다.

In the meanwhile, technologies for high-throughput functional assays are under development that may produce the highthroughput functional validations needed to improve in silico variant predictions.

한편, 고효율 기능 분석을 위한 기술이 개발중에 있고, 이는 silico 변이 예측을 개선하는데 필요한 high throughput 기능 유효성 검사를 만들 수 있다.

**Finding recurrent mutations**

Besides the interpretation of variants and genes, progress has also been made in the approaches to provide proof of pathogenicity for novel candidate genes.

변이와 유전자의 해석 외에도, 새로운 후보 유전자에 대한 병원성의 증거를 제시하는 접근법도 진전이 있었다.

While functional proof of pathogenicity remains crucial, it is time-consuming and expensive to obtain, and requires specific expertise.

병원성의 기능적 증거는 여전히 중요하지만, 시간이 많이 걸리고 비용이 많이 들고 특정 전문지식이 필요하다.

An additional layer of evidence for pathogenicity of a mutation in a candidate disease gene can be obtained by identifying multiple patients with mutations in the same gene and a similar phenotype.

후보 질병 유전자에서 돌연변이의 병원성에 대한 증거의 추가 layer는 동일한 유전자 및 유사 표현형에서 돌연변이가 나타나는 다수의 환자를 식별함으로써 얻어질 수 있다.

Two different approaches for finding recurrently mutated candidate genes have emerged, depending on whether the disorder is either rare and monogenic or more common and genetically heterogeneous.

재발성 돌연변이 후보 유전자를 발견하기 위한 두가지 접근법은 장애가 드물고 단일성이 있거나 더 흔하고 유전자 적으로 heterogenous 한지에 따라 나타난다.

**Genotype‑centric approach for common genetically**

**heterogeneous disorders**

For genetically heterogeneous disorders, it is not possible to select specific subsets of patients based on their phenotype to perform a targeted analysis of the candidate gene.

유전적으로 이질적인 질환의 경우 후보 유전자의 표적 분석을 수행하기 위해 표현형에 기초하여 환자의 특정 하위 집합을 선택할 수 없다.

Therefore, screening of a large cohort of patients for additional mutations within the same candidate gene is typically performed.

따라서, 동일한 후보 유전자 내에서 추가 돌연변이에 대한 대규모 집단 환자 screening 이 보통 수행된다.

When costs allow, it is even more efficient to immediately screen the entire cohort by exome sequencing, rather than start with a small number of selected samples.

여력이 된다면, 적은 수의 표본집단으로 시작하는 대신 E.S 로 전체 코호트를 즉시 Screening 하는 것이 훨씬 효율적이다.

In such a set-up, however, the probability of random findings becomes very large and rigorous statistics are required.

그러나 그러한 설정에서는 랜덤 발견의 확률이 매우 커지고 엄격한 통계가 필요하다.

Statistical methods do not only protect against potential false positive findings but are also able to take into account factors like reduced

penetrance, modifiers, and multigenic effects.

통계적 방법은 잠재적인 위양성 결과를 예방할 뿐 아니라 감소된 penetrance, modifiers 와 다원 효과와 같은 요인들을 고려할 수 있다.

The first large-scale exome sequencing studies already relied on different statistical approaches based on estimates of genome-wide mutation rates to identify genes enriched for de novo mutations.

최초의 대규모 ES는 이미 de novo 돌연변이가 풍부한 유전자를 밝히기 위해 유전자 전체 돌연변이 비율 추정치에 근거한 다양한 통계적 접근 방식에 의존했다.

An improved statistical framework was proposed by Samocha et al. (2014) which was first applied by the DDD project which performed large-scale trio sequencing of 1133 trios with developmental disorders.

개선된 통계적 프레임워크 들은 “ ”에 의해 제안되었고, 발달 장애가 있는 1133명의 트리오의 대규모 트리오 Sequencing을 수행한 DDD 프로젝트에서 처음 적용되었다.

After identifying de novo coding mutations in this cohort, a statistical approach was used based on estimates of the gene specific mutation rate to identify 12 novel genes that were enriched for de novo mutations.

이 코호트에서 de novo 코딩 돌연변이를 확인한 후 de novo 돌연변이가 강화된 12개의 신규 유전자를 확인하기 위해 유전자 특이 돌연변이율의 추정치에 근거하여 통계적 접근법을 사용하였다.

The same group also used a novel statistical framework for the identification of recessive genes in a cohort of 4125 families with

developmental disorders.

같은 그룹은 또한 발달 장애를 가진 4125의 가정의 코호트에서 열성 유전자를 확인하기 위한 통계적 프레임워크를 사용하였다.

In this case a model was constructed to estimate the probability of drawing *n* unrelated families with similar biallelic genotypes by

chance from the general population.

이 경우 모델은 우연히 모집단으로부터 유사한 biallelic 유전자형을 가진 n개의 관련 없는 가족을 그릴 확률을 추정하기 위해 구축되었다.

Estimates of population allele frequencies for rare loss-of-function and missense variants were obtained from the Exome Aggregation

Consortium data set.

희귀한 loss-of function 과 missense 변이들의 집단 대립 유전자 빈도의 추정치는 ExAC 의 데이터셋에서 얻어졌다.

Although in both studies the cohorts are considered to be very large, statistical power was still limited and the authors emphasize that

this should motivate data-sharing through international databases.

두 연구에서 코호트가 굉장히 크다고 여겨지더라고 통계력은 여전히 제한적이었으며, 저자들은 이것이 국제 DB를 통해 데이터를 공유하는 것에 동기 부여가 되어야 한다고 강조한다.

**Phenotype‑centric approach for rare monogenic**

**disorders**

For many Mendelian diseases the phenotype is very rare, and individual groups do not have more than a few cases making it impossible to perform large-scale screening.

많은 Mendelian 유전 질환에 대해 표현형은 매우 드물며, 개별 집단은 몇가지 사례가 없으므로 대규모 심사를 수행할 수 없다.

The alternative is then to take a phenotype-centric approach where one finds additional patients with the same, or similar, distinct phenotype. Once more patients have been identified with overlapping phenotypes specific testing of candidate genes can be performed.

대안은 동일하거나 유사한 뚜렷한 표현형을 가진 추가 환자를 찾는 표현형 중심의 접근방식을 택하는 것이다. 중복된 표현형으로 더 많은 환자가 확인되면, 후보 유전자의 특정 검사를 수행할 수 있다.

Alternatively, there is the opposite approach in which first patients with matching genotypes are identified and final evidence of pathogenicity comes from the matching of patient phenotypes.

양자 택일로, 일치하는 유전자형을 가진 최초의 환자가 확인되고 병원성의 최종 증거가 환자의 표현형과 매치되는 반대의 접근법이 있다.

In both cases additional evidence is obtained not just by the common genotype, but also by the shared specific phenotype of patients with mutations.

두 경우 모두 일반적인 유전형 뿐 아니라 돌연변이가 있는 환자의 공통된 표현형에 의해 추가적인 증거가 얻어진다.

This approach is now facilitated by various data-sharing initiatives for rare diseases such as DECIPHER, Cafe Variome, GeneMatcher, RD-connect, and PhenomeCentral.

이 접근법은 이제 DECIPHER등 과 같은 희귀한 질병에 대한 다양한 데이터 공유 계획에 의해 촉진된다.

The matchmaker exchange is a recent initiative to integrate the information from all of these databases by providing a single interface for queries together with match-making algorithms.

Matchmake 교환은 matchmake 알고리즘과 함께 쿼리를 위한 단일 인터페이스를 제공함으로써 이러한 모든 DB의 정보를 통합하고자 하는 최근의 계획이다.

A nice example of a phenotype-centric approach is a recent paper

on the identification *RSPRY1* by which the authors identified an additional case with the same phenotype using the Care4Rare Canada matchmaker.

표현형 중심적 접근법의 좋은 예가 최근 논문이다.

This should hopefully inspire more researchers to contribute to these databases and facilitate the identification of the genetic cause for their patients.

이것은 더 많은 연구자들이 DB에 기여하고 환자의 유전적 원인을 밝히는 데에 도움이 되기를 바란다.

By contributing these data to public databases they do not only become available to researchers and physicians but also to the patients

themselves.

이 데이터를 공개 DB에 제공함으로서 연구자와 의사는 물론 환자들 에게도 제공할 수 있다.

**Structured phenotypes**

The probability of success for matchmaking increases with the availability of good phenotype information. A longstanding challenge with phenotype descriptions is the lack of standardization.

좋은 표현형 정보의 이용 가능성에 따라 매치 메이킹의 성공 확률이 높아진다. 표현형 기술의 대한 오랜 도전은 표준화의 부족이다.

This presents several problems such as the use of different clinical nomenclature for similar phenotypes, the uncertainty whether phenotypes are absent or not assessed, and the fact that it is unclear how phenotypes are related to each other, which makes it difficult to

perform computational analyses.

이것은 유사한 표현형에 대한 다른 임상적 명칭의 사용, 표현형이 없는지 평가되지 않는 지에 대한 불확실성과, 얼마나 표현형들이 서로 연관되어 있는지 확실치 않고 이것이 전산 분석 수행을 어렵게 만든다는 사실등의 문제점들을 나타낸다.

For some years these issues have been tackled by the introduction

of standardized phenotype vocabularies and ontologies.

몇 년 동안 이 문제는 표준화된 표현형 어휘 및 존재로의 소개로 해결되었다.

Several ontologies have been developed but one of the most used is the human phenotype ontology (HPO). HPO currently consists of more than 250,000 phenotypic annotations.

여러 존재론이 만들어 졌지만 가장 많이 사용되는 것은 인간 표현형 존재론(HPO)이다. HPO는 현재 25만 가지 이상의 표현형 주석으로 구성되어 있다.

The practical benefits of using HPO have been demonstrated by

the development of novel tools that facilitate the prioritization of exome variants, but also by a recent study of the DDD project.

HPO 사용의 실질적 이점은 exome 변종의 우선순위 지정을 용이하게 하는 새로운 도구의 개발뿐 아니라 DDD 프로젝트에 대한 최근 연구에 의해 밝혀졌다.

used structured phenotype information to statistically quantify the phenotypic similarity of patients with developmental delay for which rare mutations were identified in the same gene.

희귀 돌연변이가 동일한 유전자에서 확인된 발달 지연 환자의 표현형 유사성을 통계적으로 정량화 하기 위해 구조화 표현형 정보를 사용했다.

Although the added value of the integrated phenotypes in the statistical assessment was limited, this will likely improve when phenotype information becomes more comprehensive.

비록 통계적 평가에서 통합된 표현형의 부가가치가 제한적이었지만, 표현형 정보가 보다 포괄적으로 될 때 향상될 것이다.

Obtaining comprehensive structured phenotypes, however, is difficult

and time-consuming. The DDD project mandated the availability

of phenotype information in HPO format for all of their samples.

그러나 포괄적인 구조화된 표현형을 얻는 것은 어렵고 시간이 많이 소요된다. DDD 프로젝트는 모든 샘플의 HPO 형식 속 표현형 정보를 제공한다.

Such criteria are not easily imposed for most other projects and several tools have been developed to encourage and facilitate the use

of phenotype information. PhenoDB and PhenoTips are platforms that allow clinicians to enter, store and analyze structured phenotypic data.

그러한 기준은 다른 대부분의 프로젝트에 많이 없으며 표현형 정보의 사용을 장려하고 촉진하기 위한 몇가지 툴이 개발되었다. PhenoDB와 PhenoTips는 임상의가 구조화된 표현형 데이터를 입력, 저장 및 분석 할 수 있게 해주는 플랫폼이다.

Phenominer is a tool able to extract phenotype contexts from simple text to identify relationships between human diseases described in OMIM and literature.

Phenominer 는 간단한 텍스트에서 표현형 문맥을 추출하여 OMIM 에 설명된 인간 질병과 문헌 사이의 관계를 식별할 수 있는 도구이다.

In the future even the actual measuring of phenotypes may be automated leading to more robust and objective phenotypes that will also take less time of physicians to administrate, and allowing

bioinformaticians to use these data for interpretation of exome variants.

미래에는 표현형의 실제 측정도 자동화 될 수 있기 때문에 의사의 관리 시간을 단축하고 보다 효과적이고 객관적인 표현형을 제공할 수 있고, 생물정보학자들이 exome 변이 해석을 위해 이러한 데이터들을 사용하게 해준다.

**Conclusions**

Here we have discussed some of ongoing bioinformatics developments that have the potential to impact the way we currently analyze and interpret exome data.

It is clear that many developments in bioinformatics are still needed with respect to exome sequencing and that this is still a very active field of development. This requires a high degree of flexibility and adaptiveness from those working in this field.

Especially since new challenges are already on the horizon with the anticipated large-scale application of whole genome sequencing.