Target 영역에 따른 DNA sequencing (WGS, WES, Targeted seq) 기술

Genome project 이후 구축된 human 의 DB 는 이후 많은 의료 및 연구분야에서 활용되고 있습니다. 이렇듯 DNA sequencing 은 주로 유전자의 mutation 연구에 활용됩니다. 유전자 서열의 mutation 은 세포분열 과정에서 발생하는 것이 일반적이며, 크게 germline mutation 과 somatic mutation 으로 나눌 수 있습니다. germline mutation 은 생식세포에서 일어나는 mutation 이기 때문에 유전이 된다는 특징이 있습니다. 반면, somatic mutation 체세포에 일어나는 mutation 으로 암과 같은 신체의 일부분에서 일어납니다. 최근 개인이 유전자 검사를 의뢰할 수 있게 일부 허용되면서 전세계적으로 유전자검사가 진단의 목적만이 아닌 개인의 특성을 확인하기 위한 용도로 널리 사용되고 있습니다. [1] 이번 기술노트에선 여러가지 DNA-seq 중 대표적인 WGS(Whole genome sequencing), WES(Whole exome sequencing), Targeted sequencing 에 대해 설명하고자 합니다.

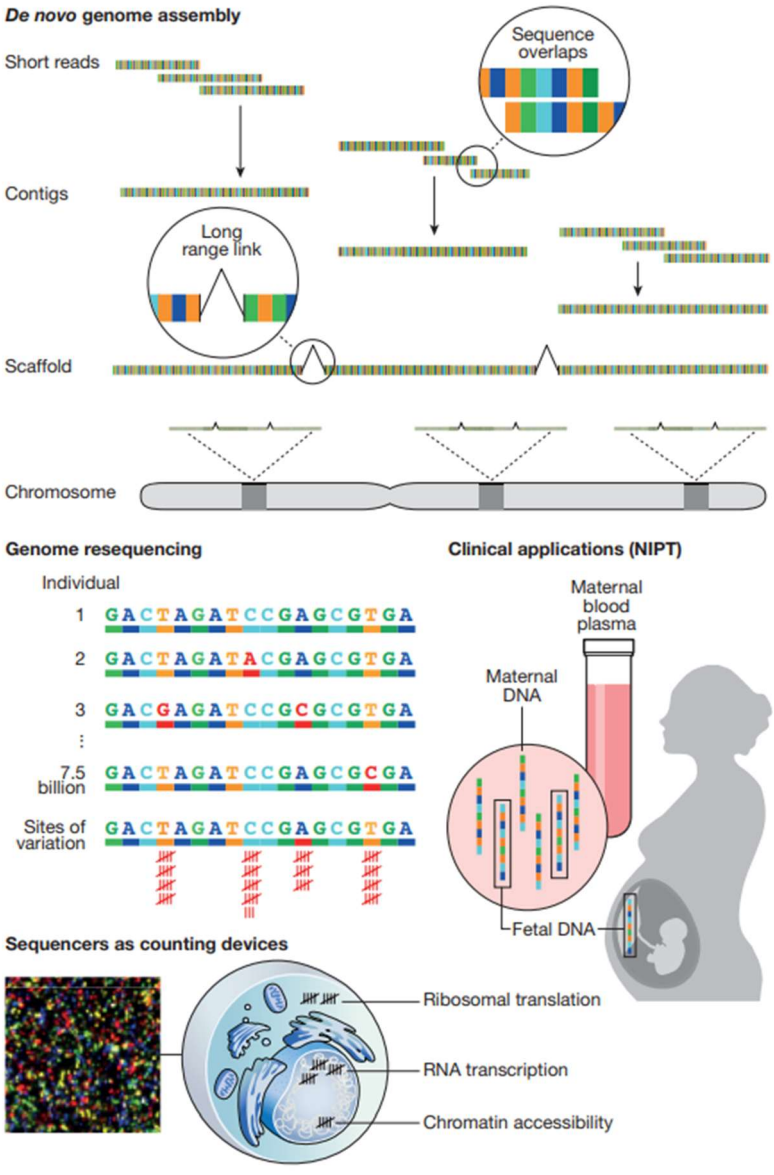

그림 1. DNA sequencing applications

WGS (Whole Genome Sequencing)

WGS (Whole genome sequencing)은 전장 유전체 서열을 분석하는 것을 말합니다. 방식은 크게 두가지로 reference genome 없이 assembly 를 통해 전장 genome 을 구축하는 *De novo* sequencing 과 genome sequencing 이 완료된 생물종의 reference 로 mapping 하는 re-sequencing 분석방법이 있습니다. NGS 기술 개발 이후 25 년 간은 *De novo* sequencing 으로 assembly 를 통해 전장 genome 서열을 밝히는 방식으로 현재 연구에 활용되는 DB 를 구축하는 연구들이 활발하게 진행되었습니다. 이렇게 제작된 reference genome 에 mapping 하여 새로운 mutation 을 찾고 표현형과 연관을 찾는 re-sequencing 을 통한 연구 또한 활발하게 진행되고 있습니다. 이렇게 발견된 mutation 은 그 종류와 영향, 메커니즘이 다양하기 때문에 해당 정보를 확인하기 위해서는 database 에서 annotation 작업이 필요합니다. Annotation 하기 위해선 목적에 맞는 database 를 활용해야 합니다.

- **Population database** 는 정상인의 DB 로 질환을 일으키지 않는 polymorphism 이 반영되어 있는 DB 입니다. Mutation 의 빈도를 확인하기 위해 비교하는 용도로 활용합니다. dbSNP, 1000 Genomes Project, ExAC 등이 있습니다.
- **Mutation database** 는 이미 알려진 mutation 에 대한 DB 로 germline mutation 위주의 DB 인 ClinVar, HGMD, LOVD 등과 cancer mutation(somatic mutation) 정보 위주로 이루어진 DB 인 COSMIC, TCGA, ICGC 등이 있습니다.

이런 mutation 에 annotation 을 진행해주는 프로그램은 SNPeff, annovar, variant effect predictor(VEP)등 다양하게 나와있습니다. [4] 그러나 WGS 의 경우 전체 유전체 서열을 읽은 데이터이기 때문에 그 크기가 커서 데이터의 용량과 처리 시간이 많이 필요합니다. 따라서 WES 나 targeted seq 을 이용하는 것이 연구의 효율을 위해 좋은 선택지일 수 있습니다.

이바이오젠에서 서비스하는 WGS 서비스는 Resequencing 방식으로 reference SNP, INDEL, CNV, Structural variation 등 분석 데이터를 제공하고 이외 다양한 분석을 지원해 드립니다.

WES (Whole Exome Sequencing)

Whole exome sequencing 은 mRNA 로 전사되는 서열인 exome 영역만을 sequencing 하는 방식으로 exome 영역내의 유전적 variation 을 분석하는 방식입니다. WES 는 WGS 에 비해 처리해야 하는 데이터양이 적기 때문에 분석에 드는 비용과 시간이 절약되고 exome 부위를 더 높은 depth 로 볼 수 있다는 장점이 있습니다. 예를 들어 사람의 경우 WGS 진행시 90G(human 기준, depth 30X)의 데이터가 생성되는 반면 WES 은 9G(human 기준, depth 100X)이내로 적은 양은 데이터로 높은 depth 의 데이터를 얻을 수 있습니다. 단, exome 영역 만을 sequencing 하기 때문에 intron 영역에서의 mutation 은 확인할 수 없다는 한계가 있으므로 연구에 적절한 방식을 택해야 합니다.

Whole exome sequencing 을 하기위해선 exome 을 capture 하는 과정이 필요합니다. 저희 이바이오젠은 Agilent 사의 SureSelect UTR kit 를 사용하고 있으며, RNA 서열을 이용하여 hybridization 된 DNA 만을 bead 로 capture 하는 방식으로 exome 영역만을 enrichment 하여 서비스를 제공하고 있습니다. (**그림 2**) 이후 sequencing 결과를 분석하여 SNP(Single Nucleotide Polymorphism), INDEL (Insertion & Deletion), CNV(Copy Number Variation) 등의 결과를 제공하고 있습니다.

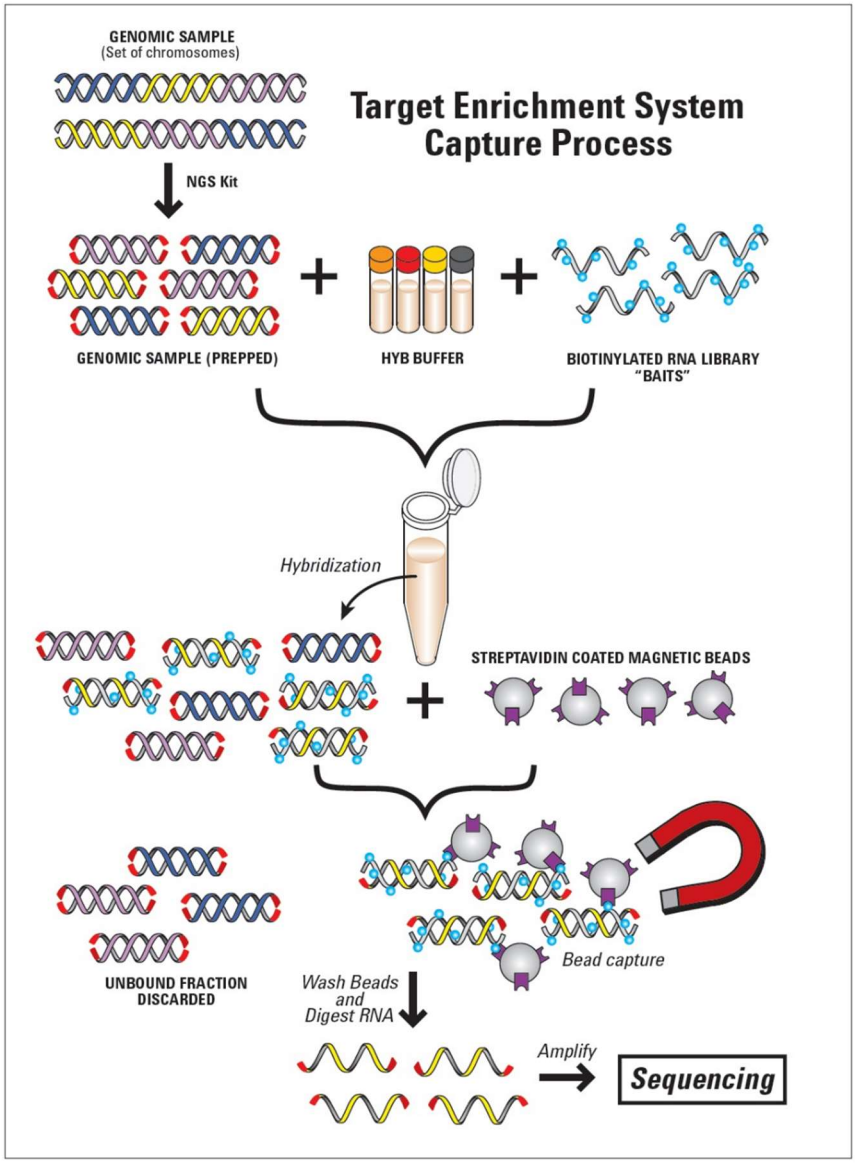

그림 2. Target enrichment system workflow

Targeted panel Sequencing

Targeted seq 서비스는 target 하는 유전자들이 있는 경우 유전체 서열 전체를 볼 필요가 없기 때문에 원하는 부분만을 보는 panel kit 를 제작하여 선택한 유전자 서열의 variation 을 보는 sequencing 방법입니다. WGS, WES 대비 선별한 영역만을 sequencing 하기 때문에 1000X 이상의 depth 로도 sequencing 할 수 있습니다.
이바이오젠에서는 완전한 고객 맞춤형 panel kit 를 제작하여 시료 추출부터 NGS sequencing, 데이터 분석까지 모든 과정을 지원하고 있습니다. (**그림 3**) 제작된 panel kit 는 kit 만 별도 구매가 가능합니다.

①			②	
Target 유전자 확인 및 디자인	Pilot test 진행	KIT 제작	NGS 진행	Data 분석

그림 3. Targeted sequencing work process

Target Enrichment panel kit 는 Hybridization 기반의 capture 기술을 이용하여 전체 유전체에서 특정 영역의 염기서열을 분석합니다. SNV, INDELs, CNV, Gene Rearrangement 와 같은 다양한 유형의 돌연변이를 정확하게 분석할 수 있습니다. Targeted sequencing 의 경우 원하는 유전자만은 target 하여, 아래 모식도처럼 5~500 개의 유전자에서 높은 depth 의 데이터를 얻을 수 있습니다. **(그림 4)**

저희 이바이오젠에서 서비스하고 있는 RNA-seq 을 통해 선별된 유전자를 Targeted seq 을 통해 variation 분석까지 한번에 지원 가능합니다.

그림 4. Target 범위에 따른 DNA sequencing 의 종류

관련 학술 논문

WGS 를 이용한 학술논문을 하나를 소개하자면, 한 논문에서 최근 팬데믹을 일으켰던 COVID-19 의 병원성과 관련된 human 숙주의 유전자를 찾고자 하는 연구를 진행하였습니다. 대상은 332 명의 중국의 코로나 환자로, 이들의 DNA 를 Whole Genome Sequencing (average depth : 46X)하여 2,220 개의 variation 을 검출하였습니다. 이후 병증의 정도와 기간을 기준으로 그룹별로 나누어 유전자 기반 연관성 분석(GWAS 분석)을 진행하였습니다. (**그림 5**) 병증으로 나눈 Severe 군과 Non-severe 군을 비교분석한 결과 IL-1 signaling pathway 에 포함된 "TMEM189–UBE2V1" 유전자가 병증의 정도와 주요하게 관련있는 유전자라고 추정하였습니다. 하지만 기간으로도 나누어 분석한 결과에선 특별히 차이를 보이는 유전자를 찾지 못 하였습니다. 이후 해당 논문에서 해당 유전자와 COVID-19 과의 상관관계에 대한 추가분석들을 실행하였고 해당 환자들을 대상으로 다른 논문에서 COVID-19 의 병원성과 관련이 있다고 알려진 HLA gene alleles 유무를 SNP marker 를 통한 분석을 진행하였습니다. 이렇게 나온 후보 유전자들은 이후 COVID-19 병원성 관련연구에 활용할 예정입니다. [5] WGS 는 특정 형질에 연관되어 있는 유전자를 찾아 관련 유전자의 마커를 개발하는 연구에 활용될 수 있습니다.

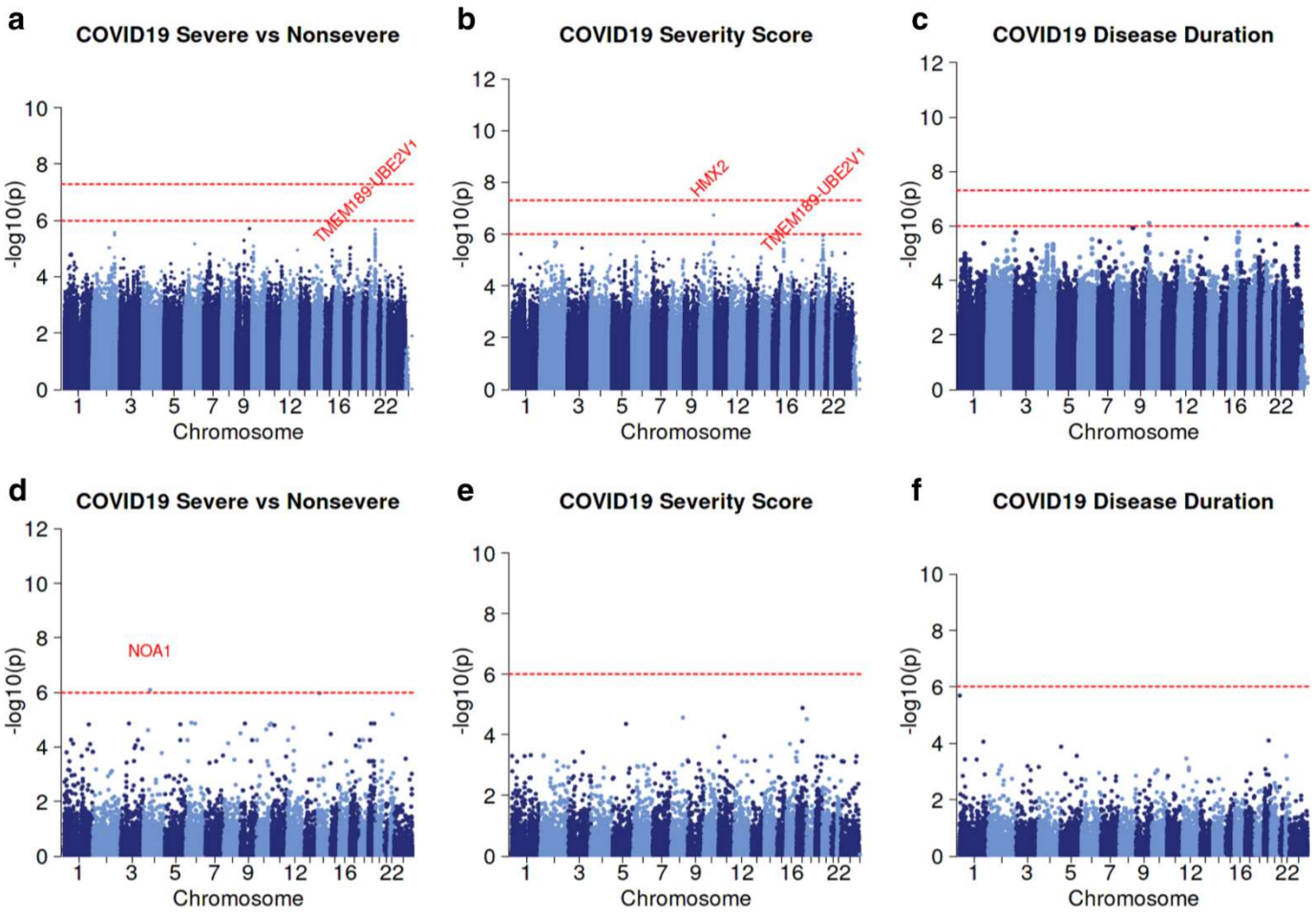

그림 5. Genetic loci associated with patient severity.

E-biogen's Service Information

WGS (Whole Genome Sequencing)

Sample requirement	>2ug gDNA
Library method	Illumina TruSeq Nano DNA Library kit
NGS run format	NovaSeq 6000, PE150
Turnaround time	~7 weeks after DNA QC
Data Analysis	SNP(Single Nucleotide Polymorphism), INDEL (Insertion & Deletion), CNV(Copy Number Variation), 개별 샘플에 대한 Structural Variation 분석 지원

WES (Whole Exome Sequencing)

Sample requirement	>2ug gDNA
Library method	Agilent Sure Select kit
NGS run format	NovaSeq 6000, PE100
Turnaround time	~6 weeks after DNA QC
Data Analysis	SNP(Single Nucleotide Polymorphism), INDEL (Insertion & Deletion), CNV(Copy Number Variation), 개별 샘플에 대한 Structural Variation 분석 지원

Targeted panel Sequencing

Sample requirement	>2ug gDNA
Library method	NEB Ultra DNA Library kit
NGS run format	NextSeq 550, PE150bp
Turnaround time	~6 weeks after DNA QC (kit 제작 기간 별도)
Data Analysis	ExPADA report, Peak Annotation, Motif Discovery, IGV, BedGraph 등

< 참고 문헌 >

1. *Shendure, Jay, et al. "DNA sequencing at 40: past, present and future." Nature 550.7676 (2017): 345-353.*
2. *Richards, S., et al., Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med, 2015. 17(5): p. 405-24*
3. *Kosuri, Sriram, and George M. Church. "Large-scale de novo DNA synthesis: technologies and applications." Nature methods 11.5 (2014): 499-507.*
4. *NGS 기반+유전자검사(심화용), 식품의약품안전평가원, 2022*
5. *CELEMICS, End-to-end Customized Target Enrichment Panel, https://www.celemics.com/products/customized-ngs-panel/*
6. *Wang, Fang, et al. "Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility." Cell discovery 6.1 (2020): 83.*