한국인 범유전체 구축 및 활용 가이드라인 (v1.0)

Guideline for the construction and utilization of the Korean Pangenome

2025. 12. 24.

질병관리청
국립보건연구원
유전체연구기술개발과

목 차

1. 한국인 범유전체 개요 ····· 1

2. 분석을 위한 기본 정보 준비 ····· 3

3. Assembly 분석 방법 ····· 4

4. Pangenome Graph 구축 분석 방법 ····· 6

5. Pangenome Graph 활용 방법 ····· 10

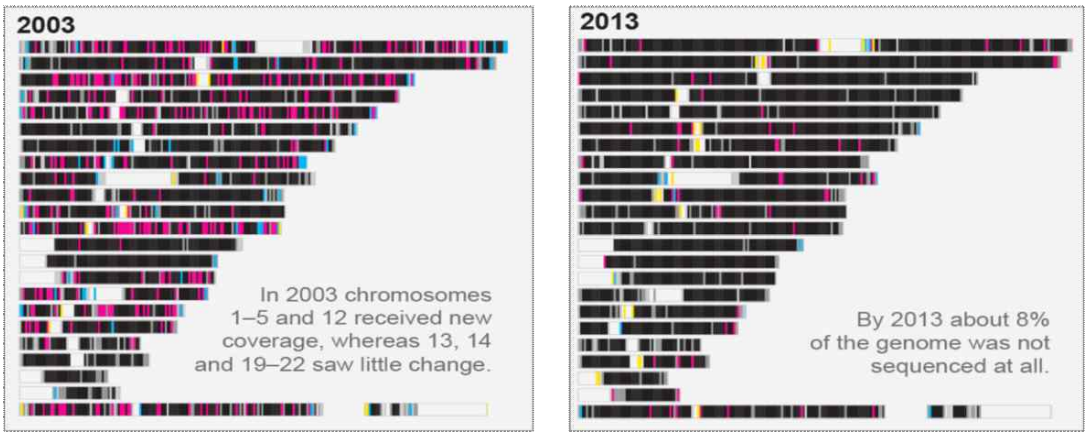
참고문헌 ····· 14

1 한국인 범유전체 개요

□ 인간 참조표준 유전체지도의 변화

○ 2000년 인간 게놈프로젝트 종료 발표 이후 2003년에 공개한 첫 참조표준 유전체 지도는 지속적으로 업데이트하여 2013년에 마지막 버전을 발표함

- 2022년 이후 기존 참조표준의 한계로 인해 새로운 버전 발표를 무기한 연기함



< 그림. 인간 참조표준 유전체 지도의 변화 >

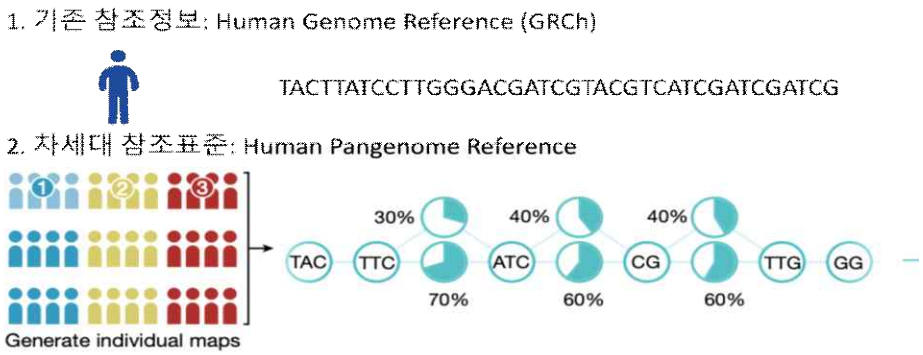
*기존 정보에 비교하여 새롭게 추가된 지역은 붉은색 등 색상으로 표기됨

□ 차세대 인간 참조표준 유전체지도 (Pangenome, 범유전체)

○ 기존 참조표준 유전체지도에서 아래의 한계점을 확인함

- 70% 이상이 1명의 유전체정보로 구성됨
- 약 8%의 유전체 지역이 미해독됨 (2013년 버전)

○ 이에 다양한 사람들의 유전체정보를 반영하고 완전하게 해독한 범유전체를 차세대 인간 참조표준 유전체지도로 채택함



< 그림. 기존 참조표준과 범유전체 차이점 >

□ 한국인 범유전체 프로젝트 (Korean Pangenome Project, KPP)

○ ‘25년부터 5년간 한국인 1,000명 범유전체 구축과 국제표준에 한국인 정보 포함을 목표로 국제협력연구 추진

- 신속한 연구 수행을 위해 ‘25년부터 첨단 신기술 기반 범유전체 정보 생산 및 구축, 국제협력연구를 동시에 추진하고자 함

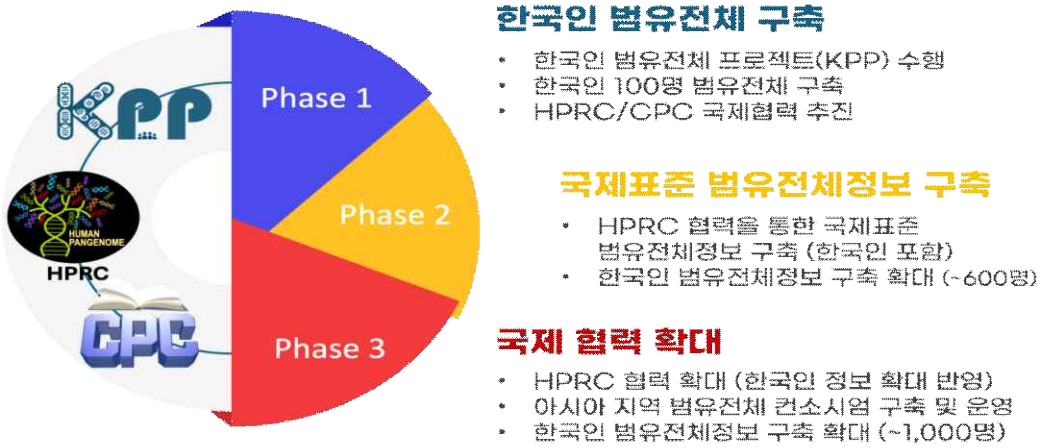
※ 첨단 신기술: long-read sequencing, Hi-C 등 Pangenome 구축 핵심 데이터

○ **(추진계획)** ‘25년부터 5년간 3단계 사업 추진

- (1단계, ‘25년) 한국인 범유전체 관련 정보생산 및 100명분 구축
- (2단계, ‘26-’27년) HPRC 2단계(phase 2) 연구에 참여, 국제표준정보에 한국인 범유전체 정보 포함 및 공개 (한국인 판지놈 누적 600명)

※ HPRC는 1단계(phase 1) 결과 발표 완료(Nature ‘23) 및 2단계 추진 중

- (3단계, ‘28-’29) HPRC 3단계(phase 3) 연구 참여, 아시아 범유전체 정보 컨소시엄 구축 및 운영을 통한 한국인 범유전체 정보 활용 확대(~1천명)



< 그림. 한국인 판지놈 사업 추진 모식도 >

* HPRC: Human Pangenome Reference Consortium, CPC: Chinese Pangenome Consortium

○ **(추진현황)** 한국인 범유전체 지도 초안 발표 (‘25.12) 및 국제참조표준 범유전체 제작에 참여 중

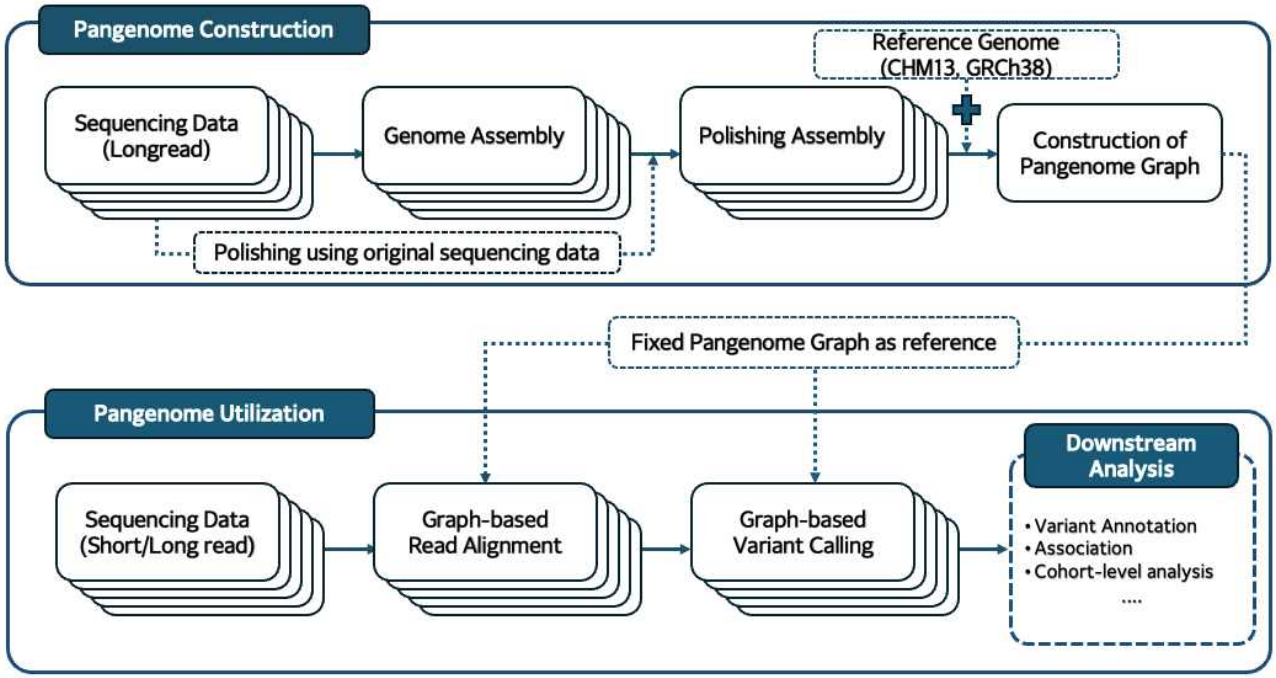
- 한국인 132명으로 구성된 범유전체 지도 초안 발표

* github: https://github.com/KoreanPangenome/KPPD

- HPRC associate member 가입(’25.10.3) 및 국제참조표준 범유전체 제작 참여 중

2 분석을 위한 기본 정보 준비

1. 범유전체 구축 및 분석 흐름도



2. 준비

○ 프로그램

	프로그램	관련정보
1	hifiasm	https://github.com/chhylp123/hifiasm
2	Inspector	https://github.com/Maggi-Chen/Inspector
3	minigraph-cactus	https://github.com/ComparativeGenomicsToolkit/cactus/
4	vg	https://github.com/vgteam/vg
5	bioawk	https://github.com/lh3/bioawk

· 분석 방법 별로 conda 또는 docker 환경으로 설치를 권장함

○ 데이터

· Pacbio HiFi longread 원시 데이터(FASTQ 파일)

· Reference genome: GRCh38, CHM13v2.0

· Pangenome Graph: KPPD132.gfa.gz

3 Genome Assembly 분석 방법

1. Genome Assembly 분석이란?

○ Genome Assembly (유전체 재조합) 분석은 개별 샘플의 유전체 서열을 참조 유전체에 의존하지 않고 재구성하여, 구조적 변이 및 반복 영역을 포함한 유전체 전체 구조를 정밀하게 규명하는 분석 방법임

2. Raw Assembly 분석

○ 내용

· Raw longread 기반 genome assembly 수행

○ 명령어의 예

```
# Step 1) assembly 분석
hifiasm -o prefix -t 64 sample.hifi.fastq.gz 2> asm/sample.hifiasm.log

# Step2) GFA -> FASTA 변환
awk '/^S/{print ">"$2;print $3}' prefix*.hap1.*p_ctg.gfa > prefix.h1.fa
awk '/^S/{print ">"$2;print $3}' prefix*.hap2.*p_ctg.gfa > prefix.h2.fa
```

○ 옵션

· -o: 결과 파일 이름

· -t: 분석에 사용할 CPU thread 수

· sample.hifi.fastq.gz: Longread 원시데이터

○ 결과파일

〈Step1 결과파일〉

· prefix.r_utg.gfa : 모든 haplotype 정보를 유지한 raw unitig 그래프

· prefix.p_utg.gfa : 작은 bubble을 제거한 processed unitig 그래프

· prefix.bp.p_ctg.gfa : primary contig 조립 그래프

· prefix.bp.a_ctg.gfa : alternate contig 조립 그래프

· **prefix.*hap*.p_ctg.gfa** : haplotype-resolved contig 그래프

〈Step2 결과파일〉

· prefix.h1.fa, prefix.h1.fa: hap1/2 contig fasta (후속분석에 사용)

3. Assembly Polishing 분석

○ 내용

· Assembly 정도관리 및 보정 분석

○ 명령어의 예

```
mkdir -p inspector/sample

# inspector 분석
inspector \
  -c inspector/sample \
  -r assembly.fa \
  -q sample.hifi.fastq.gz \
  -t 64
```

○ 옵션

· -c : 결과 디렉토리(프로젝트 폴더)

· -r : hifiasm 분석이 완료된 assembly.fa 파일

· -q : Longread 원시데이터

· -t : 분석에 사용할 CPU thread 수

○ 결과파일

· report.html (또는 index.html) : Inspector QC 종합 리포트

· summary.txt / summary.tsv : QC 핵심 지표 요약 파일

· *.bam : reads-to-assembly 정합 결과 파일

· coverage.txt / coverage.tsv : contig별 및 구간별 coverage 정보

· misassembly.txt / error_region.txt : 구조적 오류 의심 구간 목록

· indel.txt / substitution.txt : 염기 수준 오류 통계 파일

4 Pangenome Graph 구축 분석 방법

1. Pangenome Graph 구축 분석이란?

○ Pangenome Graph 구축 분석은 여러 개의 genome assembly들을 Graph 구조로 합치는 방법으로, 기존 연구 결과들과 연계를 위해 단일 참조유전체(GRCh38, CHM13)를 기준으로 Graph를 만드며, 만들어진 Graph는 전용 소프트웨어로 Graph based read alignment, Graph based Variant calling 등의 분석에 사용됨

2. Pangenome Graph 구축 분석 준비 단계 1

○ 내용

· Polishing 이 완료된 assembly의 contig ID를 Pangenome Sequence ID*로 변경

* FASTA 파일의 sequence ID를 Pangenome Sequence ID 구조로 수정해야 함

○ 명령어의 예

```
#Sequence ID:
#[sample_name]구분자[haplotype_id]구분자[contig_or_scaffold_name]
#sample_name: 문자열,haplotype_id:숫자, contig_or_scaffold_name:문자열

FASTA1="prefix.h1.fa"
SAMPLE=$(echo $FASTA1 | awk -F'.' '{print $1}')
HAP=$(echo $FASTA1 | awk -F'.' '{print $2}')
FASTA2=${FASTA1/.fa/reName.fa}
bioawk -v SAMPLE=$SAMPLE -v HAP=$HAP -c fastx \
'{print ">"SAMPLE "#"HAP "#$name; print $seq}' $FASTA > $FASTA2

FASTA2="prefix.h2.fa"
SAMPLE=$(echo $FASTA2 | awk -F'.' '{print $1}')
HAP=$(echo $FASTA2 | awk -F'.' '{print $2}')
FASTA2=${FASTA2/.fa/reName.fa}
bioawk -v SAMPLE=$SAMPLE -v HAP=$HAP -c fastx \
'{print ">"SAMPLE"#"HAP"#"$name; print $seq}' $FASTA > $FASTA2
```

○ 옵션

· -v : 외부변수를 bioawk 내부변수에 지정하는 옵션

○ 결과파일

· prefix.h1.rename.fa, prefix.h2.rename.fa : FASTA 포맷의 시퀀스 ID가 Pangenome sequence 명명법에 맞게 수정된 FASTA

3. Pangenome Graph 구축 분석 준비 단계 2

○ 내용

· Pangenome Graph를 구성할 assembly 목록 파일 만들기

○ 파일 예시 (seqfiles.txt)

· 텍스트 편집기를 이용하여 아래 정보와 같이 샘플명, assembly FASTA파일 위치를 seqfiles.txt 파일에 공백 구분자로 입력

```
#seqfiles 예시

CHM13v2 /reference/chm13v2.0_maskedY_rCRS.fa
sample1.1 /assembly위치/sample1.1.fa
sample1.2 /assembly위치/sample1.2.fa
....
sampleN.1 /assembly위치/sampleN.1.fa
sampleN.2 /assembly위치/sampleN.2.fa
```

○ 결과파일

· seqfiles.txt : Pangenome graph 구축에 사용할 assembly 지정 정보

4. Pangenome Graph 구축 분석

○ 내용

· Polishing 이 완료된 assembly를 활용하여 Pangenome Graph 구축

○ 명령어의 예

```
#docker 설치 후 아래 이미지 다운로드
docker pull quay.io/comparative-genomics-toolkit/cactus:v2.9.3

#docker 이미지 불러오기
docker run --rm -it -v 작업폴더위치:작업폴더위치 \
quay.io/comparative-genomics-toolkit/cactus:v2.9.3 bash

FILE_SEQ=./seqfiles.txt
DIR_JOB=./graph_build_job
DIR_WORK=./graph_build_work
DIR_OUTPUT=./graph_output
OUTPUT_NAME=graph.chm13v2

mkdir $DIR_WORK

cactus-pangenome \
 ${DIR_JOB} \
 ${FILE_SEQ} \
 --workDir ${DIR_WORK} \
 --outDir ${DIR_OUTPUT} \
 --outName ${OUTPUT_NAME} \
 --reference CHM13v2 \
 --gfa full clip \
 --vcfReference CHM13v2 \
 --vcf full clip \
 --vcfwave \
 --maxCores 100
```

○ 옵션

· DIR_JOB : Toil job 파일이 만들어지는 임시 폴더(재시작/분산 처리에 중요)
· FILE_SEQ : assembly 설정 파일
· --workDir : 그래프 파일이 만들어지는 임시 폴더
· --outDir : 그래프 결과 파일이 저장되는 폴더
· --outName : 산출물 prefix/경로
· --reference REF : 그래프의 기준(reference) 샘플 지정(좌표계/투영에 영향)
· --gfa : 그래프 포맷 산출(full, clip)
· --vcfReference REF : VCF의 기준 샘플 지정
· --vcf : (가능할 때) VCF 산출
· --vcfwave : 중첩 변이들이 있는 구간을 간단하게 분해한 VCF 산출
· --maxCores : 병렬 threads 수

○ 결과파일

· gfa : 그래프 (모든 서열 정보 full.gfa, 기준 샘플 관련 서열 정보 기본 gfa)
· vcf : 그래프의 유전체 변이 (full.gfa 유래 full.vcf, 기본 gfa 유래 vcf)
· wave.vcf : 그래프의 유전체 변이 파일에서 중첩 변이를 간단하게 분해한 변이 (full.gfa 유래 full.wave.vcf, 기본 gfa 유래 wave.vcf)

5. Pangenome Graph 특성 분석

○ 내용

· 구축한 Pangenome Graph의 정량적 특성을 측정하는 분석

○ 명령어의 예

```
GFA=graph.chm13v2.gfa
vg stats -zl $GFA > ${GFA}.stat
```

○ 옵션

· -z : 그래프의 노드와 노드를 연결하는 간선 수를 측정
· -l : 그래프의 총 길이 측정
· --threads : 병렬 threads 수

○ 결과파일

· graph.chm13v2.gfa.stat : 그래프의 노드 및 간선 수와 길이 통계

5 Pangenome Graph 활용방법

1. Pangenome Graph 활용한 분석 방법

○ Pangenome Graph를 활용한 분석은 특정 인종이나 인구 집단의 특징을 가진 Pangenome graph 참조 유전체에 개별 샘플의 Short-read 또는 long-read를 정렬하여, 기존 선형 참조 유전체에 서열이 없어서 정렬 분석을 못 한 구조적 변이, 복합 변이 및 집단 특이적 변이를 정밀하게 탐지하는 방법임

○ 이를 통해 인종·집단에 특이적인 유전 변이의 탐지 정확도를 향상시키고, 질환 연관성 분석 및 정밀의료 연구의 활용성을 확대할 수 있음

2. Graph based Alignment 준비 단계 1 기본 그래프에서 반수체 정보 추출

○ 내용

· Pangenome graph에서 haplotype sampling에 필요한 반수체 정보들 만들기

○ 명령어의 예

```
GFA=graph.chm13v2.gfa

# GFA 파일에서 정렬 분석용 GBZ 파일 포맷 만들기
vg gbwt --gbz-format -g ${GFA%%.gfa}.gbz -G ${GFA}

# GBZ 파일의 거리 인덱스 만들기
vg index -j ${GFA%%.gfa}.no_nested.dist –no-nested-distance
${GFA%%.gfa}.gbz

# GBZ 파일의 ri 인덱스 만들기
vg gbwt -p --num-threads 100 -r ${GFA%%.gfa}.ri -Z
${GFA%%.gfa}.gbz

# GBZ 파일의 반수체 정보 추출하기
vg haplotypes -v 2 -t 100 -d ${GFA%%.gfa}.no_nested.dist -r
${GFA%%.gfa}.ri -H ${GFA%.gfa}.hapl ${GFA%%.gfa}.gbz
```

○ 결과파일

· graph.chm13v2.gbz : 반수체 정보가 압축된 그래프 (파일 정보 수정 못함)

· graph.chm13v2.no_nested.dist : 그래프의 분기점 정보 인데스

· graph.chm13v2.ri : r-index

· graph.chm13v2.hapl : 압축 그래프에서 추출된 반수체 정보

3. Graph based Alignment 준비 단계 2 시퀀싱 kmer 패턴 추출

○ 내용

· 시퀀싱 리드의 염기서열 패턴을 추출

○ 명령어의 예

```
mkdir ./kff

export TMPDIR=$(pwd)/kff
ls | grep sample1 | sort > file.txt
kmc -k29 -m128 -okff -t100 @file.txt sample1 $TMPDIR
```

○ 옵션

· -k29 : k-mer 길이 29 염기쌍으로 설정

· -m128 : 소프트웨어 메모리 128MB로 설정

· -okff : KFF 파일 포맷으로 결과 만들기

○ 결과파일

· sample1.kff : sample1의 k-mer 29 염기쌍 패턴

4. Graph based Alignment 준비 단계 3 샘플 전용 그래프 만들기

○ 내용

· 시퀀싱 리드의 염기서열 패턴과 비슷한 Pangenome graph 생성 및 그래프의 분기점 정보 추출

○ 명령어의 예

```
GBZ=graph.chm13v2.gbz
HAPL=graph.chm13v2.hapl
KFF=sample1.kff

vg haplotypes -v 2 -t 16 --include-reference --diploid-sampling -i
${HAPL} -k ${KFF} -g sample1.hapl.gbz ${GBZ}

vg snarls --threads 110 sample1.hapl.gbz > sample1.hapl.gbz.snarls
```

○ 결과파일

· sampl1.hapl.gbz : 샘플에 특화된 그래프

· sampl1.hapl.gbz.snarls : 샘플 특화 그래프의 분기점 정보

5. Graph based Alignment

○ 내용

· 시퀀싱 리드 정보에 특화된 Pangenome graph에서 서열정렬 분석 수행 및 분석 결과 통계 만들기

○ 명령어의 예

```
#Long-read 시퀀싱 파일의 경우
vg giraffe -b hifi -p -t 110 -Z sample1.hapl.gbz -f
sample1.hifi.reads.fastq.gz > sample1.gam

#Short-read 시퀀싱 파일의 경우
vg giraffe -p -t 110 -Z sample1.hapl.gbz -f sample1.reads.fastq.gz >
sample1.gam

vg stats --threads 110 -a sample1.gam > sample1.gam.qc
```

○ 옵션

· -b hifi : hifi 전용 서열정렬 설정

○ 결과파일

· sampl1.gam : 시퀀싱 리드가 그래프에 서열정렬된 정보

· sampl1.gam.qc : 서열정렬 정량적 수치 정보

6. Graph based Variant call

○ 내용

· Pangenome graph에 서열 정렬 결과에서 그래프 정렬 빈도 생성 및 변이 호출

○ 명령어의 예

```
vg pack --threads 110 -x sample1.hapl.gbz -g sample1.gam -o
sample1.pack -Q 5

vg call --threads 110 -a sample1.hapl.gbz -r sample1.hapl.gbz.snarls -k
sample1.pack -s sample1 -z > sample1.vcf
```

○ 옵션

· -Q 5 : 맵핑 품질 5 아래의 서열정렬 결과는 제외

○ 결과파일

· sampl1.pack : 그래프에서 서열정렬 빈도 정보

· sampl1.vcf : 그래프에서 호출된 유전체변이 정보

6 참고문헌

○ Liao, Wen-Wei, et al. "A draft human pangenome reference." Nature 617.7960 (2023): 312-324.

○ Cheng, Haoyu, et al. "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm." Nature methods18.2 (2021): 170-175.

○ Chen, Yu, et al. "Accurate long-read de novo assembly evaluation with Inspector." Genome biology 22.1 (2021): 312.

○ Armstrong, Joel, et al. "Progressive Cactus is a multiple-genome aligner for the thousand-genome era." Nature587.7833 (2020): 246-251.

○ Garrison, Erik, et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference." Nature biotechnology 36.9 (2018): 875-879.

○ Sirén, Jouni, et al. "Pangenomics enables genotyping of known structural variants in 5202 diverse genomes." Science374.6574 (2021): abg8871.

○ Hickey, Glenn, et al. "Genotyping structural variants in pangenome graphs using the vg toolkit." Genome biology 21.1 (2020): 35.

○ Sirén, Jouni, and Benedict Paten. "GBZ file format for pangenome graphs." Bioinformatics 38.22 (2022): 5012-5018.

○ Paten, Benedict, et al. "Superbubbles, ultrabubbles, and cacti." Journal of Computational Biology 25.7 (2018): 649-663.