Software note.

Manipulating FASTA/Q

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| SeqKit | Almost all the required functions | Great, fast versatile |
| sqetk |  | Classical tool kits |

Sequence Quality Control

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| fastqc | Classic but only good for short reads | yes |
| multiqc |  |  |
| LongQC | For long reads. Written in Python. Installation with Anaconda alone may not be enough. | yes |
| pycoQC | For Nanopore. Depends on the summary file from base caller. |  |
| NanoPlot | For long reads. |  |

Check the seq error rate, alignment evaluation

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| Samtools stats |  |  |
| qualimap |  |  |
| tinycov | github.com/cmdoret/tinycov plot the coverage of a BAM file |  |

Read Trimming and filtering

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| fastp | For short reads. Fast! Easy to use | Great |
| trimmomatics | For short reads. Popular. The syntax is a little complicated. | It’s fine |

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| Porechop-abi | For long reads. It detects the adapters automatically. | Great! |
| Chopper | The successor of NanoFilt+NanoLyse. |  |
| Porechop | Classic and previously popular. HOWEVER, the author stopped maintaining it since long time ago! |  |
| Filtlong | filter long reads by quality and produce a better subset |  |

Long-read de novo genome assembly

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| Flye | Fast, usually accurate, easy to use. |  |
| Canu/Hi-canu/Canu-trio | Computational intense, more conservative but more accurate per base. |  |
| Next denovo | Fast, similar to Flye |  |
| Hifiasm | Specific for Hifi reads. Fast, purges duplications on its own. |  |
| GoldRush | Super fast but not so accurate. Only good for super large genome. |  |
| NGSEP | Published in 2023/Feb/22. Good at haploid genome. Not the best for diploid genome. https://github.com/NGSEP/NGSEPcore |  |
| Hifiasm |  |  |
| NGSEP Assembler | <https://github.com/NGSEP/NGSEPcore>  DOI 10.26508/lsa.202201719 | Not tested yet |
| raven | Slower, easy to use |  |
| Shasta | optimized for ONT reads; Super fast |  |
| MARVEL Assembler | GPU/CUDA-based; no future publication | I haven’t tried |
| FLACON, NECAT, shasta, redbean, miniasm, wtdbg2, Phasebook, MARVEL Assembler…… | Less frequently used by the community or not maintained by the authors anymore |  |

Hybrid de novo genome assembly

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| MaSuRCA |  |  |
| LazyB, WENGAN, DBG2OLC | not maintained by the authors anymore |  |

Short read genome assembler

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| SOAPdenovo2 |  |  |
| abyss |  |  |

Polisher of de novo genome assembly

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| nanopolish | Nanopore only. Utilize raw fast5 signals, only with R9 chemistry. |  |
| medaka | Nanopore only. Should be used directly on the output of Flye. Developed by Oxford Nanopore. Limited. |  |
| Pilon | Utilize accurate short reads, e.g., illumina. Computational intensive. Popular. | Either polca or this |
| POLCA | It under the Github folder of MASURCA. Theoretically more efficient than Pilon. | Either pilon or this |
| NextPolish | Only relies on long reads. fast | Yes |
| racon | Only relies on long reads. Computational intensive. Easy to use. Popular. | Yes |
| JASPER | Super fast, at the cost of sensitivity. Input= PacBio HiFi or Illumina |  |
| MarginPolish, HELEN, P.E.P.P.E.R., Arrow, | Not maintained anymore |  |

Scaffolding assembly and contiguity improvements (gap filling)

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| ntLink | Only need long reads fastq. Clear protocols. https://github.com/bcgsc/ntLink |  |
| SAMBA tool | Also part of MaSuRCA |  |

Other Post processing of *de novo* genome assembly

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| purge\_dups | main stream |  |
| Longstitch |  |  |
| HapSolo | removing secondary haplotigs during diploid genome assembly and scaffolding |  |
| purge\_haplotigs |  |  |
| HaploMerger2 | tended to overpurge |  |

Assembly evaluation

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| BUSCO |  |  |
| Compleasm | A faster and more accurate version of BUSCO |  |
| abyss-fac | Part of the ABYSS assembler. Reference-free analysis |  |
| QUAST | Reference-based analysis.  https://quast.sourceforge.net/ |  |
| Inspector |  |  |
| Merqury |  |  |
|  | coverage plot |  |
| KAT | if short reads or some accurate reads are available. |  |

Short-read alignment (mapping)

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| BWA-MEM2 | Faster!! Only “mem” function | Suitable for most cases |
| BWA | More accurate than other aligners | Classic |
| hisat2/hisat3 | Good for RNAseq alignment, esp. when the RNA length is longer than 500 bp. |  |
| Bowtie | | |

Long-read Alignment (mapping)

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| mm2-fast | Super Fast!! Accelerated version of minimap2 | Great for most cases |
| Minimap2 | Fast! | Great for most cases |
| Winnowmap2 |  |  |
| LAR | Better at contig level alignment |  |
| LAST、LASTZ | slower |  |

SAM/BAM Manipulation

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| samtools | popular | Great |
| sambamba | Better with large dataset | Great |

SNP calling from short read

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| GATK | Popular but SLOW! Great for human |  |
| Freebayes | Fast, easy to use. |  |
| Octopus |  |  |

SNP calling from long reads

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| Clair3 | Recommend for Nanopore and PacBio Hifi data |  |
| NanoCaller | With two PacBio CLR model |  |
| Longshot | Nanopore and PacBio CLR |  |
| NanoVar | Only trained with Nanopore data in publication. https://github.com/cytham/nanhttps://doi.org/10.1186/s12859-021-04118-3ovar | If the coverage is low. |
| Medaka | Oxford Nanopore deprecated this function |  |
| Caution: All the programs are trained with human genome. | | |

Repeat detecting, masking, TE annotation.

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| RepeatModeler | SLOW! *de novo* TE family identification. Input is assembly FASTA | The official manual is difficult |
| RepeatMasker | SLOW! Relies on existing library/database | The official manual is difficult |
| TETools | <https://github.com/Dfam-consortium/TETools>  Its singularity image contains: RepeatModeler & BuildDatabase, RepeatMasker, runcoseg.pl | Easy to install!! |
| Tandem Repeats Finder |  |  |
| Straglr | Input BAM; output BED and TSV format |  |
| Database: Tandem Repeats Database, Dfam database | | |

Genome annotation.

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| GALBA |  |  |
| BRAKER3 | if you have no RNA-Seq data and the genome is large, use GALBA! Otherwise use BRAKER, first. |  |

SV calling from long reads.

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| cuteSV | Mapping-based |  |
| Sniffles2 | Mapping-based |  |
| SVIM | Mapping-based |  |
| SKSV | Only works for Hifi data. |  |
|  |  |  |
| SVIM-asm | Fast, easy to use |  |
| PAVhttps://doi.org/10.1186/s12859-021-04118-3 | Only report INS and DEL right now. Author is responsive. |  |
| Syri | Many format requirement but most sensitive? |  |

VCF manipulation

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| bcftools | Versatile! But most functions require `bgzip` and `tabix` first | First choice |
| vcftools |  | Frequently use |
| vcflib |  |  |
| SURVIVOR |  |  |
| truvari |  |  |
| Jasmine | https://github.com/mkirsche/Jasmine |  |
| mavis | https://github.com/bcgsc/mavis |  |
| svimmer | skip |  |

GGF/GTG manipulation

|  |  |  |
| --- | --- | --- |
| Program | Note (performance, function) | Personal recommendation |
| AGAT | https://github.com/NBISweden/AGAT |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Good reference: https://doi.org/10.1186/s12859-021-04118-3

https://schneebergerlab.github.io/syri/

<https://github.com/schneebergerlab/plotsr>

SVJedi : Genotyping structural variations with long read data

https://github.com/KamilSJaron/smudgeplot

https://github.com/tbenavi1/genomescope2.0

http://qb.cshl.edu/genomescope/

<https://github.com/schneebergerlab/fixchr>

<https://quast.sourceforge.net/quast>

http://qb.cshl.edu/genomescope/

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04118-3

https://davidebolo1993.github.io/visordoc/

1. pauvre is useful to check the sequencing stats (pauvre stats)
   1. https://github.com/conchoecia/pauvre
2. many useful tools (Chopper, Cramino, NanoPlot!)
   1. <https://github.com/wdecoster/nanopack>
3. Long-read correction:
   1. <https://github.com/HaploKit/vechat>
   2. https://github.com/morispi/CONSENT
4. New & GOOOOOG assembler, no publication currently
   1. <https://github.com/Nextomics/NextDenovo>
5. NextPolish: a fast and efficient genome polishing tool for long-read assembly (>280 citations)
   1. <https://github.com/Nextomics/NextPolish>
6. Improve the assembly by removing haplotigs and contig overlaps
   1. <https://github.com/dfguan/purge_dups>
   2. Difficult to use but 549 citations in 3 years!!
7. LongStitch: A genome assembly correction and scaffolding pipeline using long reads
   1. <https://github.com/bcgsc/longstitch>
8. RagTag is a collection of software tools for scaffolding and improving modern genome assemblies.
   1. <https://github.com/malonge/RagTag>
9. QUAST stands for QUality ASsessment Tool. Genome assembly evaluation tool
   1. <https://github.com/ablab/quast>
10. Inspector: A reference-free assembly evaluator.
    1. <https://github.com/Maggi-Chen/Inspector>
11. Long-read mapping to repetitive reference sequences using Winnowmap2
    1. Slightly better than minimap2 generally. Better at repetitive regions, such centromeres.
    2. <https://github.com/marbl/Winnowmap>
12. Fast minimap2
    1. <https://github.com/bwa-mem2/mm2-fast>
13. Fast BWA-MEM2
    1. <https://github.com/bwa-mem2/bwa-mem2>
14. Good SNP caller!!
    1. <https://luntergroup.github.io/octopus/>
15. Indel caller. IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly (18 citations)
    1. <https://github.com/NCGG-MGC/IMSindel>
16. SNP+ Indel caller: NanoCaller
    1. <https://github.com/WGLab/NanoCaller>
17. clair3 is efficient in calling SNPs from Nanopore data
    1. NanoSNP in inferior to clair3 but superior to other SNP callers for long reads when the coverage is low. NanoCaller is much inferior to clair3 when the coverage is low.
18. SV detection after Mummer
    1. <http://assemblytics.com/>
    2. MUM&Co: <https://github.com/SAMtoBAM/MUMandCo>
    3. <https://github.com/mahulchak/svmu>
    4. https://github.com/sgblanch/smartie-sv
19. Discard SV callers for long reads:
    1. NanoSV: not maintained now
    2. PBHoney: JJ said “NO!”
    3. NanoVar: requires training but the outcome is guaranteed
20. Visualizing the assembly and complex variations
    1. <https://genomeribbon.com/>
21. Making the assembly dot plots
    1. <https://dgenies.toulouse.inra.fr/>
22. SequelTools is a program that provides a collection of tools for working with multiple SMRTcells of BAM format PacBio Sequel raw sequece data
23. SMRT Tools is PacBio’s official tools working with their reads
24. VCF processing for SVs
    1. <https://github.com/ACEnglish/truvari>
    2. <https://github.com/fritzsedlazeck/SURVIVOR>
    3. <https://github.com/DecodeGenetics/svimmer> ==> not maintained
    4. <https://github.com/mkirsche/Jasmine>
    5. <https://github.com/papaemmelab/mergeSVvcf> ==> alive
    6. https://github.com/nhansen/SVanalyzer
25. Regular VCF file processing
    1. VCFtools
    2. BCFtools
    3. Vcf -kit
    4. Vcf toolz
    5. Vcflib
    6. RTG tools
26. SAM/BAM file processing
    1. Samtools
    2. Sambamba: <https://github.com/biod/sambamba> (faster for larger datasets with larger RAM memory)

* Structural variant calling: the long and the short of it
* Comprehensive evaluation and characterization of short read general-purpose structural variant calling software
* NextSV3: automated structrual variation detection from long-read sequencing using state-of-the-art tools.
* Dysgu: efficient structural variant calling using short or long reads
* Jasmine and Iris: population-scale structural variant comparison and analysis

1. Population genetics
   1. genodive version 3.0
   2. ANGSD 0.94 stable
2. Demography, Site frequency spectrum
   1. <https://dadi.readthedocs.io/en/latest/>
   2. <http://cmpg.unibe.ch/software/fastsimcoal26/>
   3. https://speciationgenomics.github.io/easysfs/
3. Long read assemble.
   1. BlasR [76], MUMmmer [77], or Minimap2
   2. specialized methods to align long reads such as BLASR [76], Minimap2 [35], and NGMLR
4. cuteSV, LRcaller, Sniffles, SVJedi, and VaPoR
5. SV caller or workflow, Short-read alignment approach
   1. Manta (v1.1.0), DELLY (v0.7.7), LUMPY (v0.2.13) and GRIDSS (v1.3.4)
   2. DELLY [41]: the larger events remain hard
   3. TARDIS
   4. sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data
6. SV caller or workflow, LOOOOOOONG-read mapping-based approach
   1. SVIM: structural variant identification using mapped long reads (newer)
   2. Sniffles operates on a per read base, also capable of reporting very low-frequency SVs in the sample.
   3. PBHoney: For PacBio, relying on BLASR alignments
   4. PacBio structural variant calling and analysis tools (PBSV) : For PacBio, SVs within the range of 20+ bp
   5. SMRT-SV: For PacBio, includes de novo assembly and a specialized genotyping module
   6. NanoSV: For Oxford Nanopore, preferentially uses as input an alignment from LAST. NanoSV reports only breakpoints (BND) which again makes the interpretation of the SVs type difficult.
   7. Overall, long-read mapping-based methods for SV calling often show a better performance than short-read ones
7. SV caller or workflow, De novo assembly-based approach: heterozygous SVs are often missed
   1. SGVar [32] is a more recent string graph-based (see Table 2 for definition) de novo assembly pipeline based on the SGA assembler [75] that also uses short-read sequencing data
   2. Assemblytics [34] is a web application that relies on MUMmer and identifies insertions and deletions
   3. paftools.js [35] uses Minimap2 alignments
   4. SMARTie-SV was recently introduced to detect insertions, deletions, and inversions, using BlasR
8. SV caller or workflow, Other approach
   1. Hic\_breakfinder (1+ Mbp), can potentially identify all types of SVs
   2. Strand-Seq, Strandseq-InvertR [68] (min ~ 1 kbp)
9. misc
   1. BEDOPS v2.4.41: [https://bedops.readthedocs.io/en/latest/index.html#](https://bedops.readthedocs.io/en/latest/index.html)
   2. GenomeScope: Estimate genome heterozygosity, repeat content, and size from sequencing reads using a kmer-based statistical approach.
   3. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes