ANALIZA GENOMU – PODSUMOWANIE SENS BIOLOGICZNY

Wykład 8

MAGDA MIELCZAREK

CHARAKTERYSTYKA ZMIENNOŚCI POPULACJI

Nyangiri *et al. BMC Genomics* (2020) 21:289
https://doi.org/10.1186/s12864-020-6669-y

BMC Genomics

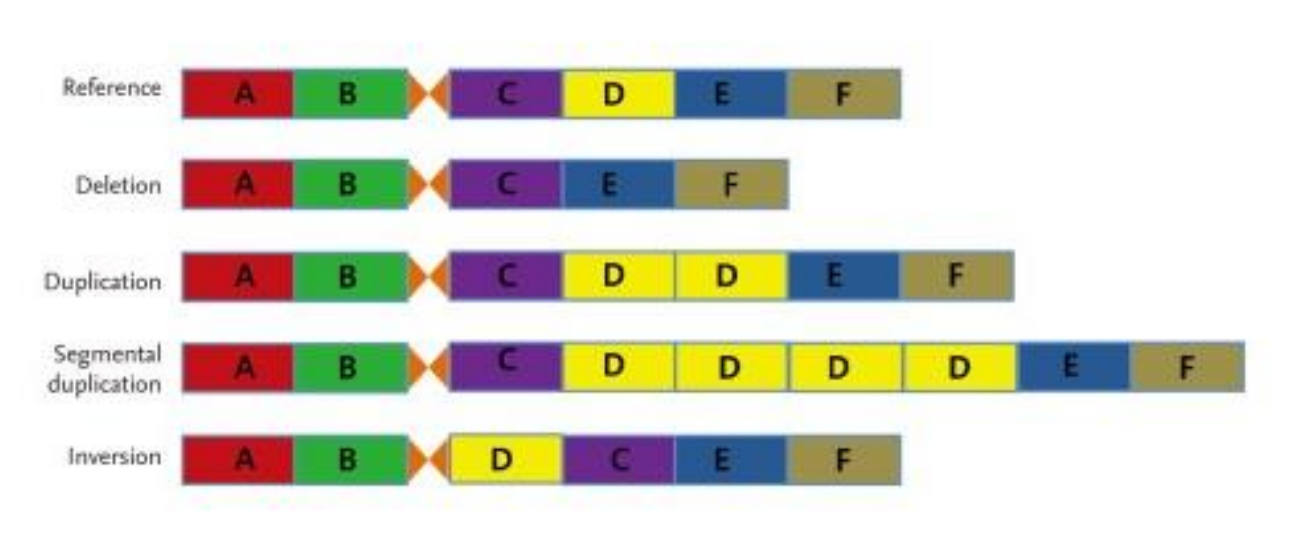
RESEARCH ARTICLE Open Access

Copy number variation in human genomes from three major ethno-linguistic groups in Africa

Oscar A. Nyangiri[1,2], Harry Noyes[3], Julius Mulindwa[1], Hamidou Ilboudo[4], Justin Windingoudi Kabore[5], Bernardin Ahouty[6], Mathurin Koffi[7], Olivier Fataki Asina[8], Dieudonne Mumba[8], Elvis Ofon[9], Gustave Simo[9], Magambo Phillip Kimuda[1], John Enyaru[10], Vincent Pius Alibu[10], Kelita Kamoto[11], John Chisi[11], Martin Simuunza[12], Mamadou Camara[13], Issa Sidibe[5], Annette MacLeod[14], Bruno Bucheton[13,15], Neil Hall[3,16], Christiane Hertz-Fowler[3], Enock Matovu[1*] and for the TrypanoGEN Research Group, as members of The H3Africa Consortium

COPY NUMER VARIATION (CNV)

Zmienność liczby kopii obejmuje duplikacje i delecje dłuższe niż 1 000 pz (rożne definicje).



„Different types of CNVs and an example of genome-wide detection of CNVs. The plot illustrates deletion, duplication, and multiple segmental duplication of the "D" locus compared with the reference genome. Inversion of "C" and "D" loci is also illustrated."

Yim SH et al. 2015. Clinical implications of copy number variations in autoimmune disorders. Korean J Intern Med. ;30(3):294-304.

CEL BADAŃ

Populacje Afryki cechują się największą na świecie różnorodnością genomową.

Populacje zajmują zróżnicowane środowiska, mają różne pochodzenie oraz wykazują różnice na poziomie genomu związane z podatnością na choroby zakaźne (np. malaria, gruźlica i HIV) i dostosowaniami środowiskowymi.

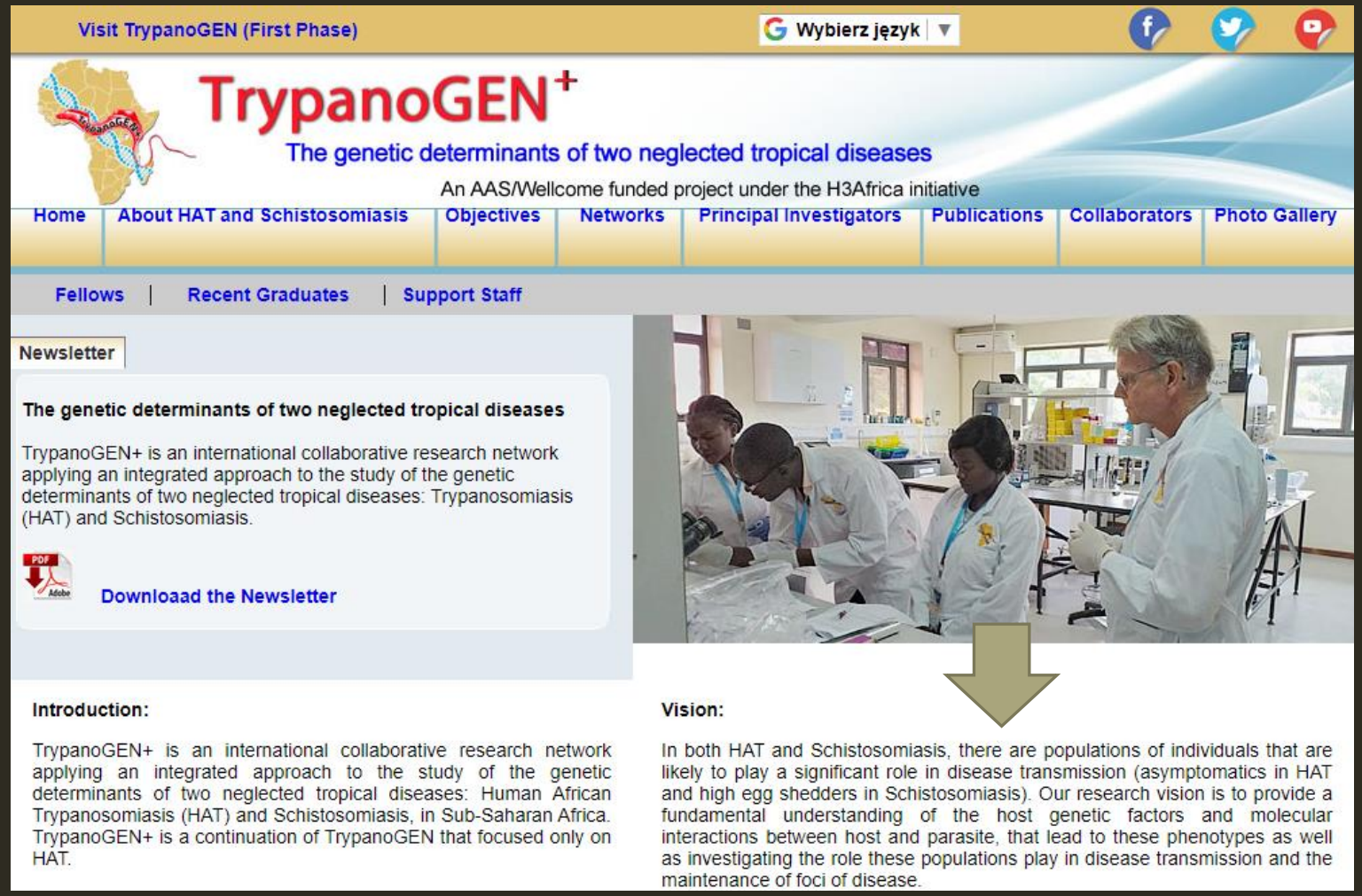
CEL → analiza całych genomów (WGS) pod kątem CNV w populacjach z różnych grup etniczno-lingwistycznych.

DANE

TrypanoGEN+ to międzynarodowa sieć badawcza, stosująca zintegrowane podejście do badania genetycznych uwarunkowań dwóch chorób tropikalnych:

- śpiączki afrykańskiej (HAT)
- schistosomatozy

TrypanoGEN+ jest kontynuacją TrypanoGEN, która skupia się wyłącznie na HAT.



DANE

Table 1 Ethnicity and origin of individuals analysed for CNV

Pop	Country	District	Ethno-linguistic group (ethnologue code, n)
UNL	Uganda	Maracha	Lugbara (IGG, 50)
UBB	Uganda	Iganga	Basoga (XOG, 33)
DRC	Democratic Republic of Congo	Bandundu	Kingongo (NOQ, 30) Kimbala (MDP, 20)
GAS	Guinea	Forecariah Boffa, Dubreka	Soussou (SUS, 49)
CIV	Côte d'Ivoire	Bonon Sinfra	Baoule (BCI, 11) Gouro (GOA 21) Moore (MOS, 12) Senoufo (SEF, 4) Malinke (LOI, 1) Koyaka (KGA, 1)

Ethnologue codes are derived from the ethnic languages of the world resource [13]

SEKWENCJONOWANIE:

- WGS
- Illumina Hiseq2500
- średnie pokrycie genomu 10X
- autosomy

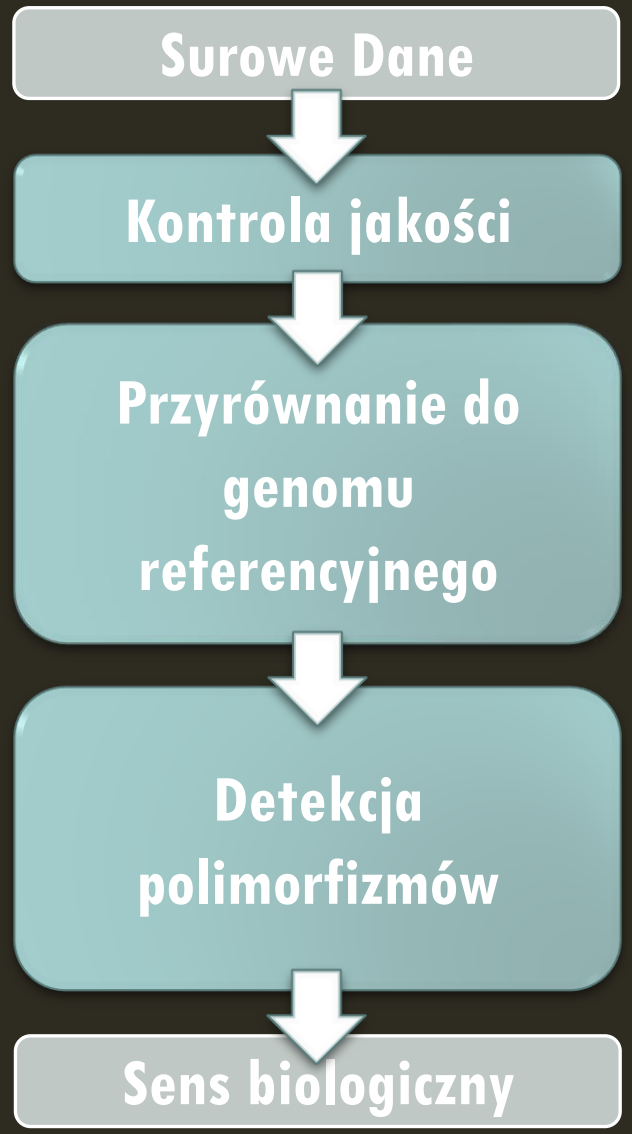
METODY

1. Detekcja polimorfizmów SNP i CNV
2. Charakterystyka CNV, adnotacja funkcjonalna
3. Określenie zróżnicowania populacji: PCA, F_{ST}
4. Określenie haplotypów na bazie CNV
5. … i inne

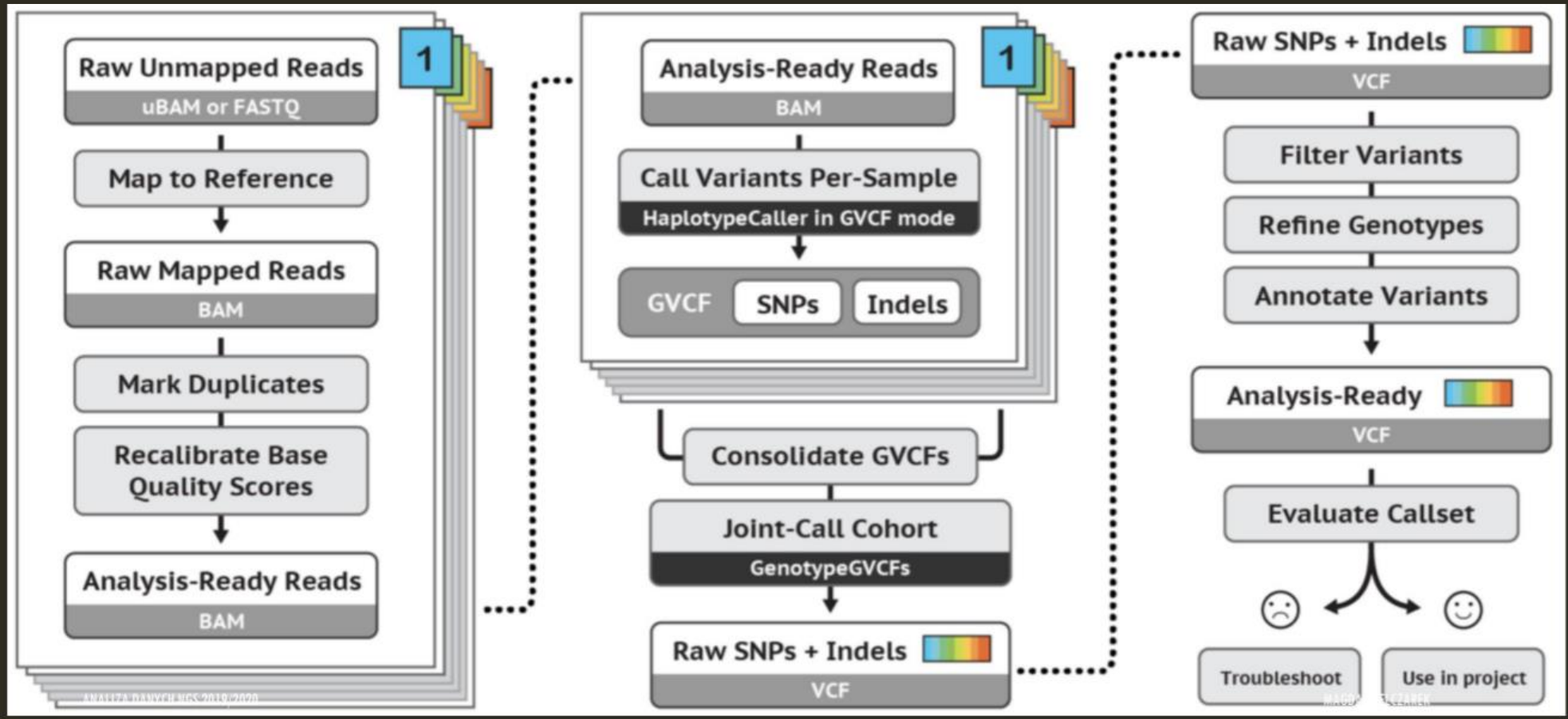
DETEKCJA SNP

Pipeline:

- Przyrównanie odczytów do genomu referencyjnego (BWA)
- Detekcja SNP zgodna z protokołem GATK – „Genome analysis tool kit best practice guideline”



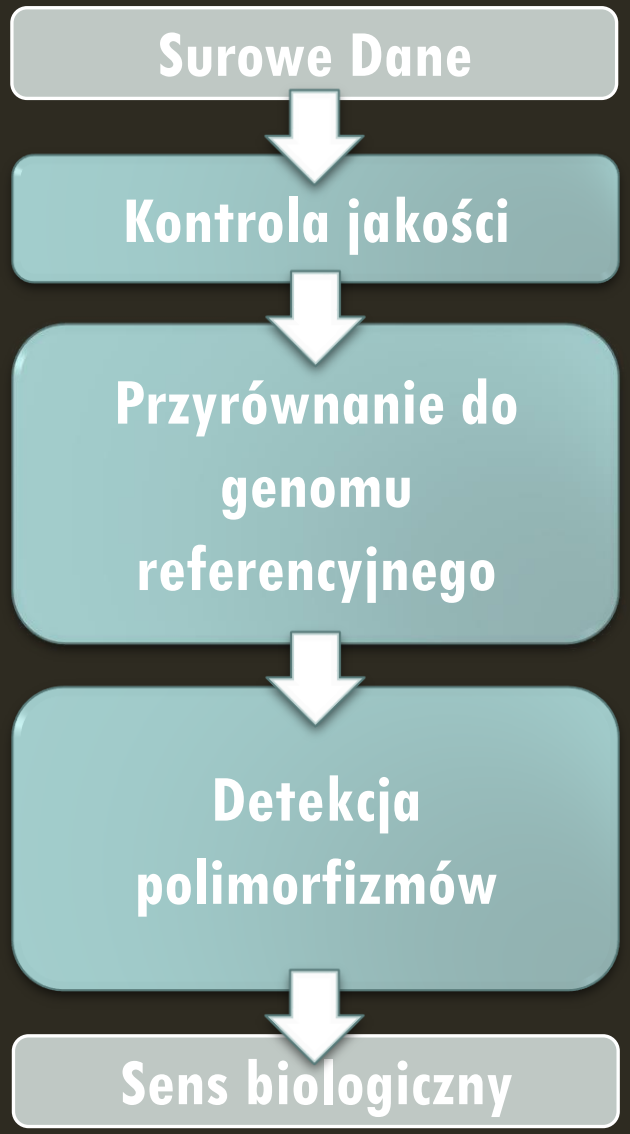
DETEKCJA SNP



DETEKCJA SNP

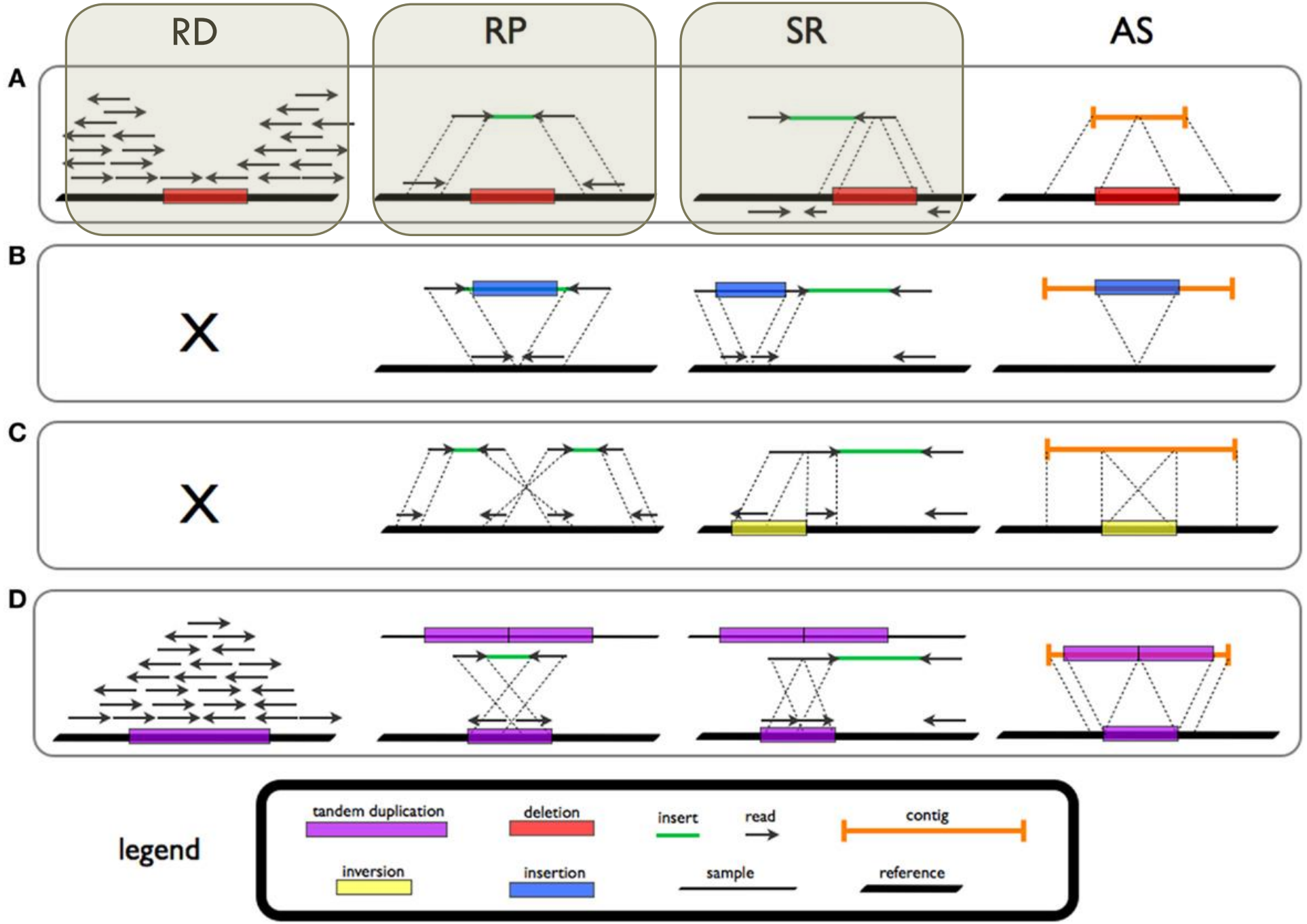
Pipeline:

- Przyrównanie odczytów do genomu referencyjnego (BWA)
- Detekcja SNP zgodna z protokołem GATK – „Genome analysis tool kit best practice guideline”
- Filtrowanie zbioru SNP:
  - usunięcie loci z > 10% brakujących SNP
  - usunięcie osób z > 10% brakujących loci
  - usunięcie loci w bliskiej odległości



DETEKCJA CNV

Programy („population scale data”):
cn.MOPS (algorytm RD)
GenomeSTRiP (algorytmy RD, RP, SR)



Tattini et al. 2015, doi: 10.3389/fbioe.2015.00092

DETEKCJA CNV

Programy („population scale data”) :

- cn.MOPS (algorytm RD)
- GenomeSTRiP (algorytmy RD, RP, SR)

Konsensus:

- wybranie wspólnego zbioru dla dwóch programów
- jak duża musi być część wspólna ?
- „GenomeSTRiP has previously been used to detect CNVs in the 1000 Genomes project of human populations. To validate detected CNVs (?) we tested for overlap with published CNVs in the public Database of Genomic Variants (DGV).”

- Adnotacja funkcjonalna

WYNIKI

Table 2 CNV statistics using GenomeSTRiP and cn.MOPS algorithms

Parameter	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
Raw CNV regions (CNVR)	16,149	9213	
CNVR after QC	11,275	2115	7608
Total CNV scored	127,699	37,679	106,922
Deletion CNV	65,588	26,008	61,025
Gain CNV	62,111	11,671	45,897
Mean CNV count per CNVR	11.3	17.8	14.0
Mean CNVR per individual	654	193	548
Count of overlapping CNVRs [a]	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated ± SD)[b]	81.2 (43.4 ± 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (1146 Mb cf. 108 Mb) than GenomeSTRiP. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapping CNVs; CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analysis (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

[a]Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

[b]The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome

WYNIKI

Table 2 CNV statistics using GenomeSTRiP and cn.MOPS algorithms

Parameter	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
Raw CNV regions (CNVR)	16,149	9213	
CNVR after QC	11,275	2115	7608
Total CNV scored	127,699	37,679	106,922
Deletion CNV	65,588	26,008	61,025
Gain CNV	62,111	11,671	45,897
Mean CNV count per CNVR	11.3	17.8	14.0
Mean CNVR per individual	654	193	548
Count of overlapping CNVRs [a]	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated ± SD)[b]	81.2 (43.4 ± 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (11[illegible] Mb cf. 108 Mb) than GenomeSTRiP CNVRs. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapp[illegible] CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analy[illegible] (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

[a]Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

[b]The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome

WYNIKI

Table 2 CNV statistics using GenomeSTRiP and cn.MOPS algorithms

Parameter	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
Raw CNV regions (CNVR)	16,149	9213	
CNVR after QC	11,275	2115	7608
Total CNV scored	127,699	37,679	106,922
Deletion CNV	65,588	26,008	61,025
Gain CNV	62,111	11,671	45,897
Mean CNV count per CNVR	11.3	17.8	14.0
Mean CNVR per individual	654	193	548
Count of overlapping CNVRs [a]	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated ± SD)[b]	81.2 (43.4 ± 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (1146 Mb cf. 108 Mb) than GenomeSTRiP CNVRs. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapping CNVs; CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analysis (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

[a]Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

[b]The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome

WYNIKI

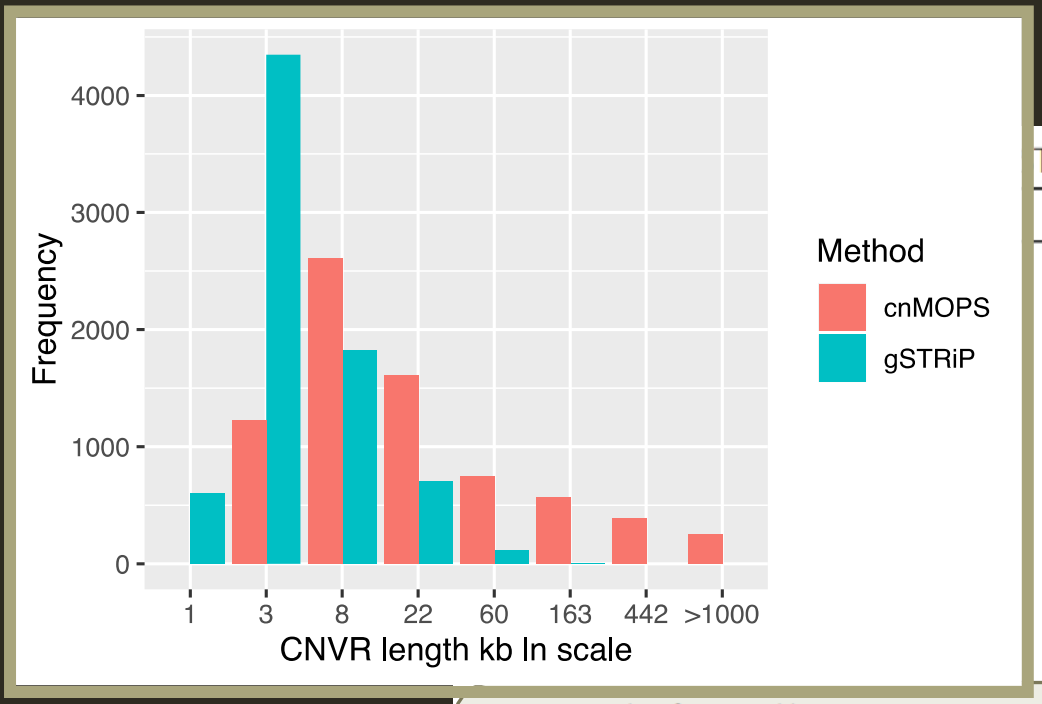
Table 2 CNV statistics using GenomeSTRiP and cn.MOPS algorithms

Parameter	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
Raw CNV regions (CNVR)	16,149	9213	
CNVR after QC	11,275	2115	7608
Total CNV scored	127,699	37,679	106,922
Deletion CNV	65,588	26,008	61,025
Gain CNV	62,111	11,671	45,897
Mean CNV count per CNVR	11.3	17.8	14.0
Mean CNVR per individual	654	193	548
Count of overlapping CNVRs [a]	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated ± SD)[b]	81.2 (43.4 ± 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (1146 Mb cf. 108 Mb) than GenomeSTRiP CNVRs. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapping CNVs; CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analysis (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

[a]Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

[b]The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome



TRiP and cn.MOPS algorithms

	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
	16,149	9213	
	11,275	2115	7608
	127,699	37,679	106,922
	65,588	26,008	61,025
	62,111	11,671	45,897
	11.3	17.8	14.0
	654	193	548
	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated ± SD)[b]	81.2 (43.4 ± 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (1146 Mb cf. 108 Mb) than GenomeSTRiP CNVRs. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapping CNVs; CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analysis (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

[a]Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

[b]The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome

WYNIKI
WSPÓLNE I UNIKALNE CNVR

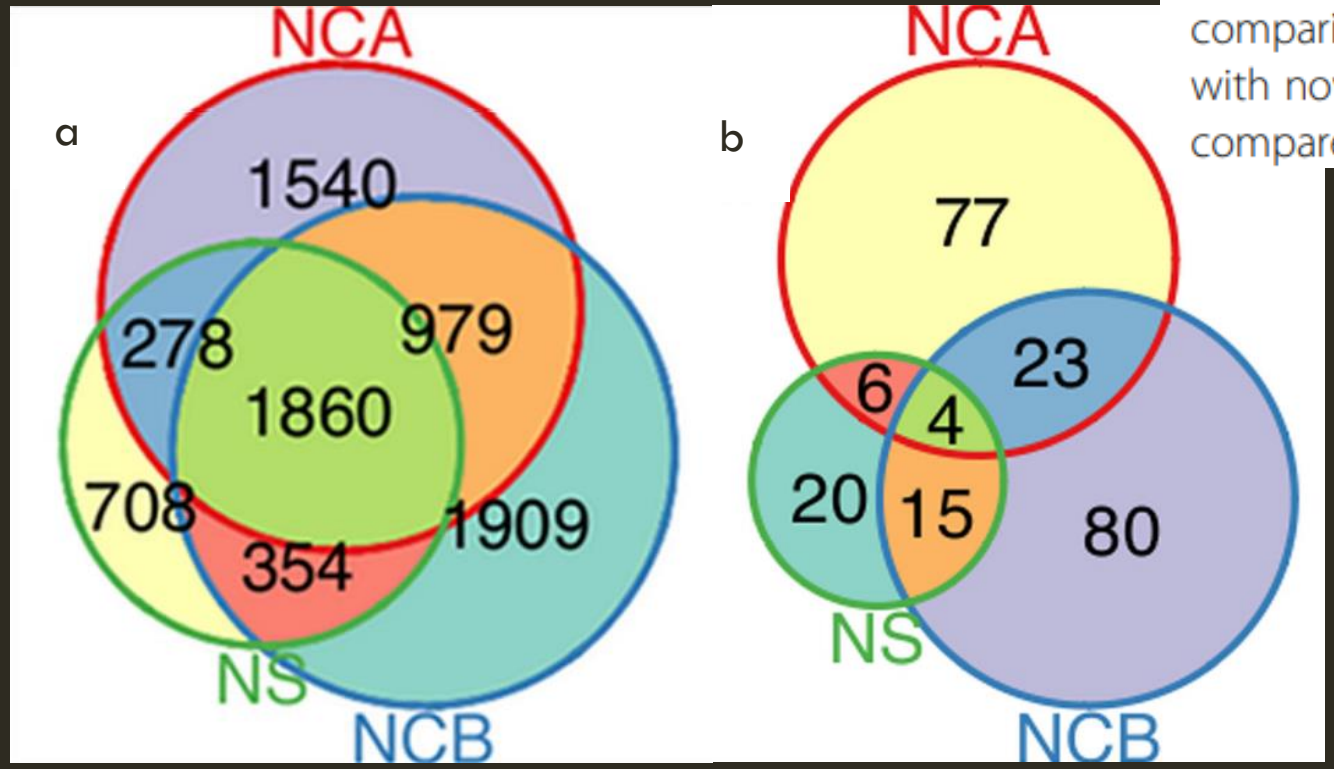


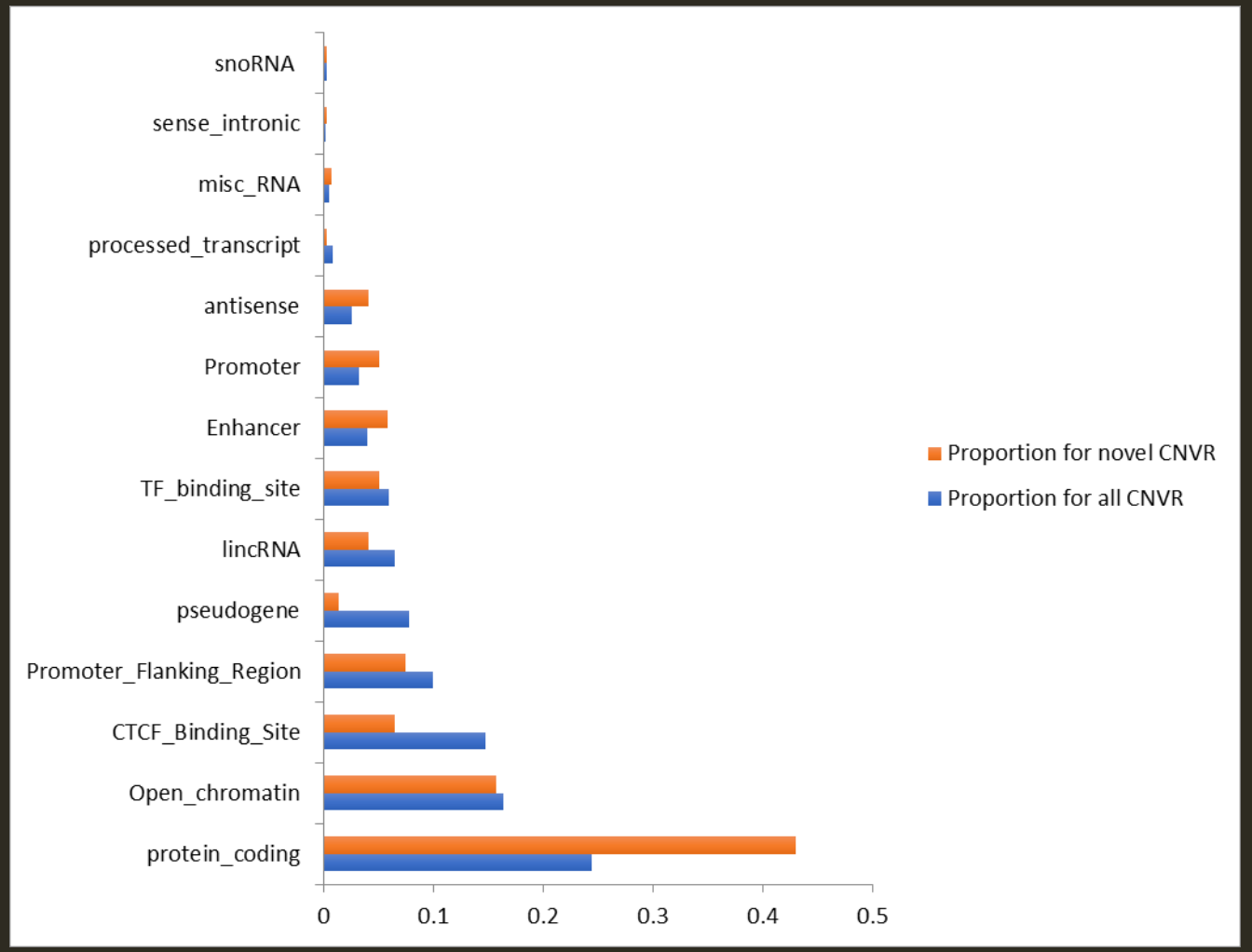
Fig. 2 Venn diagram showing counts of CNVR shared between populations. **a** All CNVR from Niger Congo A (NCA), Niger Congo B (NCB) and Nilo-Saharan (NS) ethnic groups. CNVR overlapping 5 kb genomic regions were plotted for each population. A majority of the CNVR are shared between populations, but Nilo-Saharans appear to have the least CNVR, with most of them shared with the Niger Congo A and Niger Congo B. **b** Sharing of novel CNV regions between populations. Most novel CNVR are unique to individual populations studied whereas others are shared. To enable comparison, the genome was divided into 5 kb regions and regions with novel CNVR in each of these regions for each population were compared for overlaps

Co to znaczy, że CNV są wspólne dla danej populacji? Czy idealny overlap (w granicach „okna”)?

224 (2.9%) z 7 608 CNV nie było wcześniej opisanych w bazie danych DGV (novel CNVs).

ADNOTACJA FUNKCJONALNA

Lista genów i elementów regulatorowych z Ensembl.

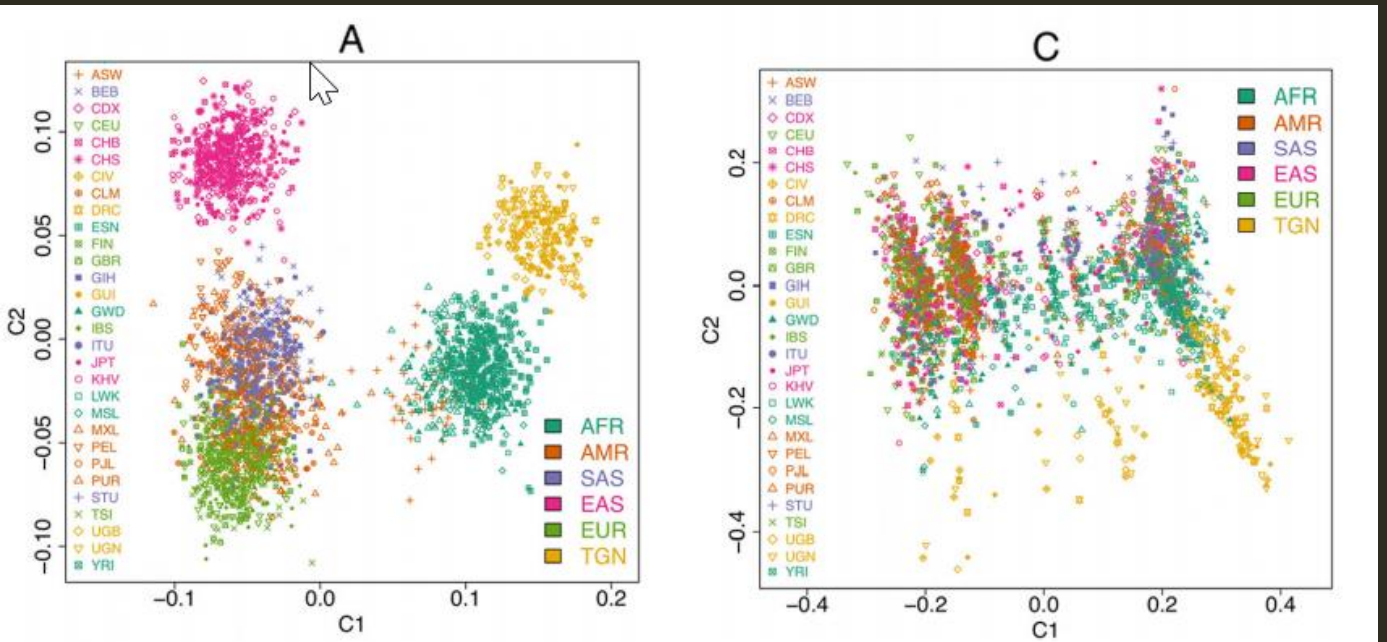


ADNOTACJA FUNKCJONALNA NOVEL CNVS

- Lista genów i elementów regulatorowych z **Ensembl**.
  „They intersected 293 unique genes or regulatory regions, with no specific function enriched and were not generally shared between the populations.”

- ***Ontologie genów***
  „27% of the novel CNVRs overlapped genes encoding binding function (GO: 0005488) and 20% (22/109) overlapped genes involved in catalytic activity (GO: 0003824).”

- „Both the known and novel CNVR overlapped ***Mendelian inheritance disease-associated genes***”



PCA of combined 1000 Genomes and TrypanoGEN populations showed population structure at the continental level (East Asians, South Asians, Caucasians, Americans, Africans)

Fig. 6 PCA plot showing CNV population structure in our data compared to 1000 Genomes. The PCA distinguishes major continental populations from each other, but is not able to resolve specific populations within the continental populations. Africans in the 1000 Genomes (AFR) are closer to our data (TGN). Conventions for major continental populations are described by the 1000 genomes project [8, 23]. **b** PCA plot showing population structure for bi-allelic deletion CNV. Phase information is non-ambiguous for bi-allelic deletions. The Africans in the 1000 Genomes overlay the TrypanoGEN African samples, indicating similar CNV in the datasets. **c** PCA plot showing population structure due to bi-allelic insertion CNV. There was no specific pattern observed as fewer bi-allelic insertions were available in the data

PODSUMOWANIE I WNIOSKI

- Zaprezentowano zmienność genetyczną populacji afrykańskich grup etnicznych.
- 3% CNVR nie było wcześniej znanych u człowieka, co odzwierciedla zróżnicowany charakter afrykańskich populacji. Nowe CNVR umieszczono w bazie danych DGV.
- Opisane CNV:
  - występują w genach, których zmiany powodują choroby mendlowskie
  - nakładają się na SNP istotnie związane z różnymi cechami w katalogu GWAS
- Rozróżnienie populacji na poziomie kontynentalnym jest możliwe przy użyciu CNV, ale wewnątrz kontynentu już nie.

CNV DETECTION (2025)

Software update

Volume 5, Issue 1
2025
(In Progress)

JOURNAL ARTICLE

Benchmarking of germline copy number variant callers from whole genome sequencing data for clinical applications

Francisco M De La Vega ✉ , Sean A Irvine , Pavana Anur , Kelly Potts , Lewis Kraft , Raul Torres , Peter Kang , Sean Truong , Yeonghun Lee , Shunhua Han ... Show more

Bioinformatics Advances, Volume 5, Issue 1, 2025, vbaf071,
https://doi.org/10.1093/bioadv/vbaf071
Published: 10 April 2025 **Article history** ▾

CNV DETECTION (2025)

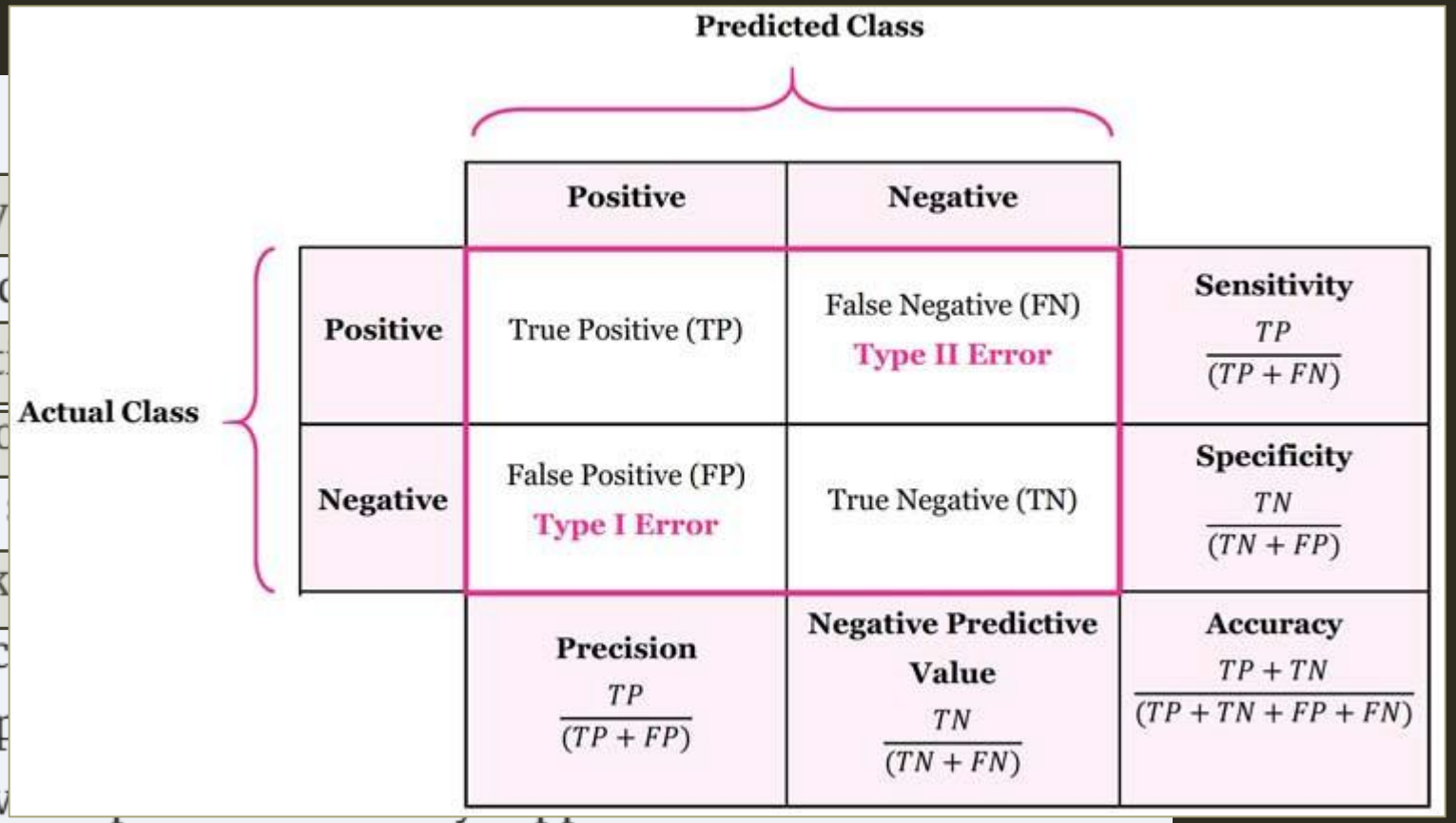
Results

While tools vary in sensitivity (7%–83%) and precision (1%–76%), few meet the sensitivity needed for clinical testing. Callers generally perform better for deletions (up to 88% sensitivity) than duplications (up to 47% sensitivity), with poor detection of duplications under 5 kb. Notably, for CNVs in genes commonly included in clinical panels, significantly improved sensitivity and precision were observed when benchmarking against 25 cell lines with known CNVs. DRAGEN v4.2 high-sensitivity CNV calls, post-processed with custom filters, achieved 100% sensitivity and 77% precision on the optimized gene panel after excluding recurring artifacts. This level of performance may support clinical use with orthogonal confirmation of reportable CNVs, pending validation on laboratory-specific samples.

CNV DETECTION (2025)

Results

While tools vary in sensitiv
sensitivity needed for clinic
deletions (up to 88% sensit
poor detection of duplicatio
included in clinical panels,
observed when benchmark
v4.2 high-sensitivity CNV c
100% sensitivity and 77% p
recurring artifacts. This lev
orthogonal confirmation of reportable CNVs, pending validation on laboratory-specific samples.



CNV DETECTION (2025)

Results

While tools vary in sensitivity (7%–83%) and precision (1%–76%), few meet the sensitivity needed for clinical testing. Callers generally perform better for deletions (up to 88% sensitivity) than duplications (up to 47% sensitivity), with poor detection of duplications under 5 kb. Notably, for CNVs in genes commonly included in clinical panels, significantly improved sensitivity and precision were observed when benchmarking against 25 cell lines with known CNVs. DRAGEN v4.2 high-sensitivity CNV calls, post-processed with custom filters, achieved 100% sensitivity and 77% precision on the optimized gene panel after excluding recurring artifacts. This level of performance may support clinical use with orthogonal confirmation of reportable CNVs, pending validation on laboratory-specific samples.

CNV DETECTION (2025)

Results

While tools vary in sensitivity (7%–83%) and precision (1%–76%), few meet the sensitivity needed for clinical testing. Callers generally perform better for deletions (up to 88% sensitivity) than duplications (up to 47% sensitivity), with poor detection of duplications under 5 kb. Notably, for CNVs in genes commonly included in clinical panels, significantly improved sensitivity and precision were observed when benchmarking against 25 cell lines with known CNVs. DRAGEN v4.2 high-sensitivity CNV calls, post-processed with custom filters, achieved 100% sensitivity and 77% precision on the optimized gene panel after excluding recurring artifacts. This level of performance may support clinical use with orthogonal confirmation of reportable CNVs, pending validation on laboratory-specific samples.

CNV DETECTION (2025)

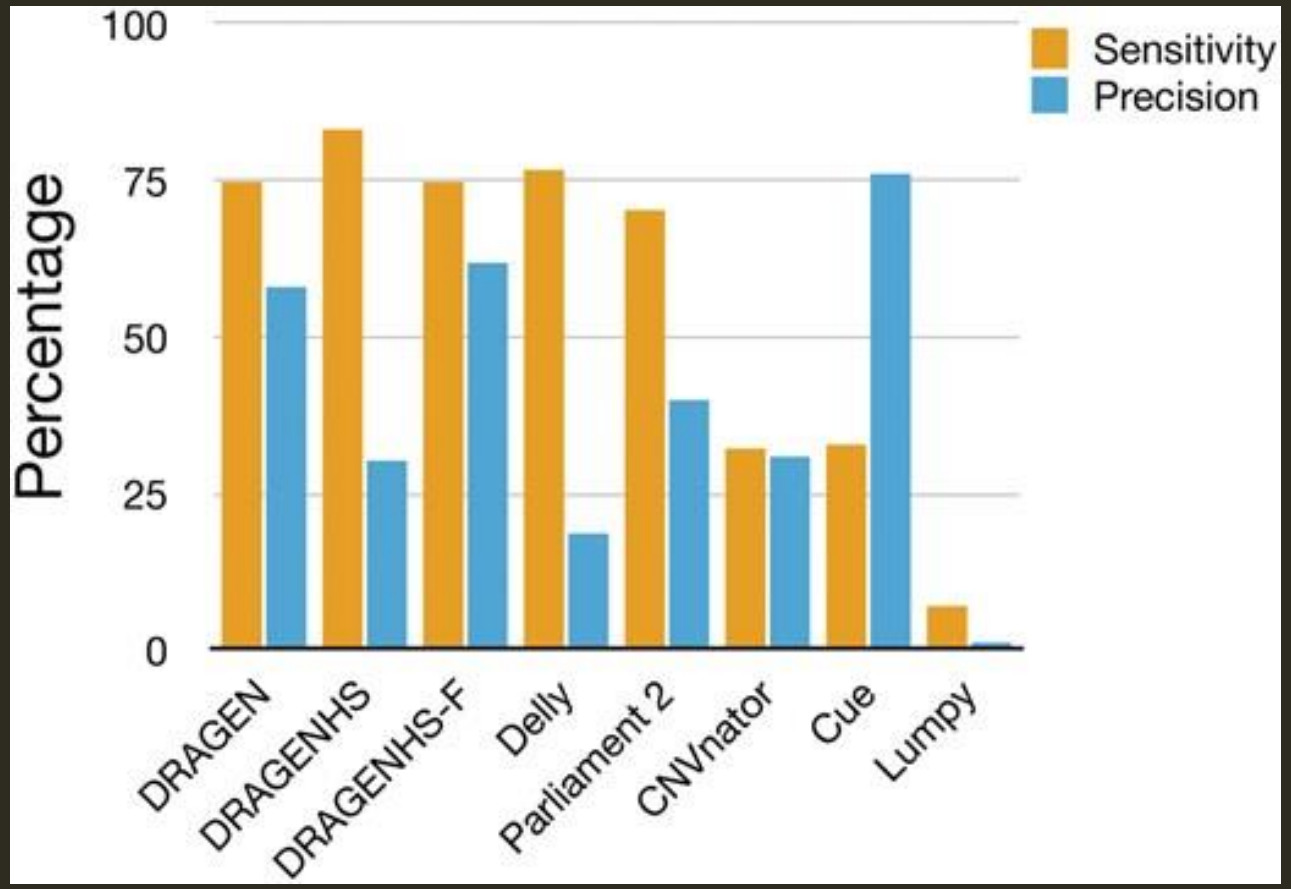


Figure 1.

Genome-wide performance of CNV callers. DRAGEN v4.2 CNV caller achieved the best balance of sensitivity and precision. In high-sensitivity mode (**DRAGEN HS**), it demonstrated the highest sensitivity but with reduced precision. **Cue** achieved the highest precision but had lower sensitivity, partly due to its inability to detect events smaller than 5 kb. Custom filters applied to DRAGEN HS (referred to as DRAGEN HS-F) improved precision with only a slight reduction in sensitivity.

CNV DETECTION (2025)

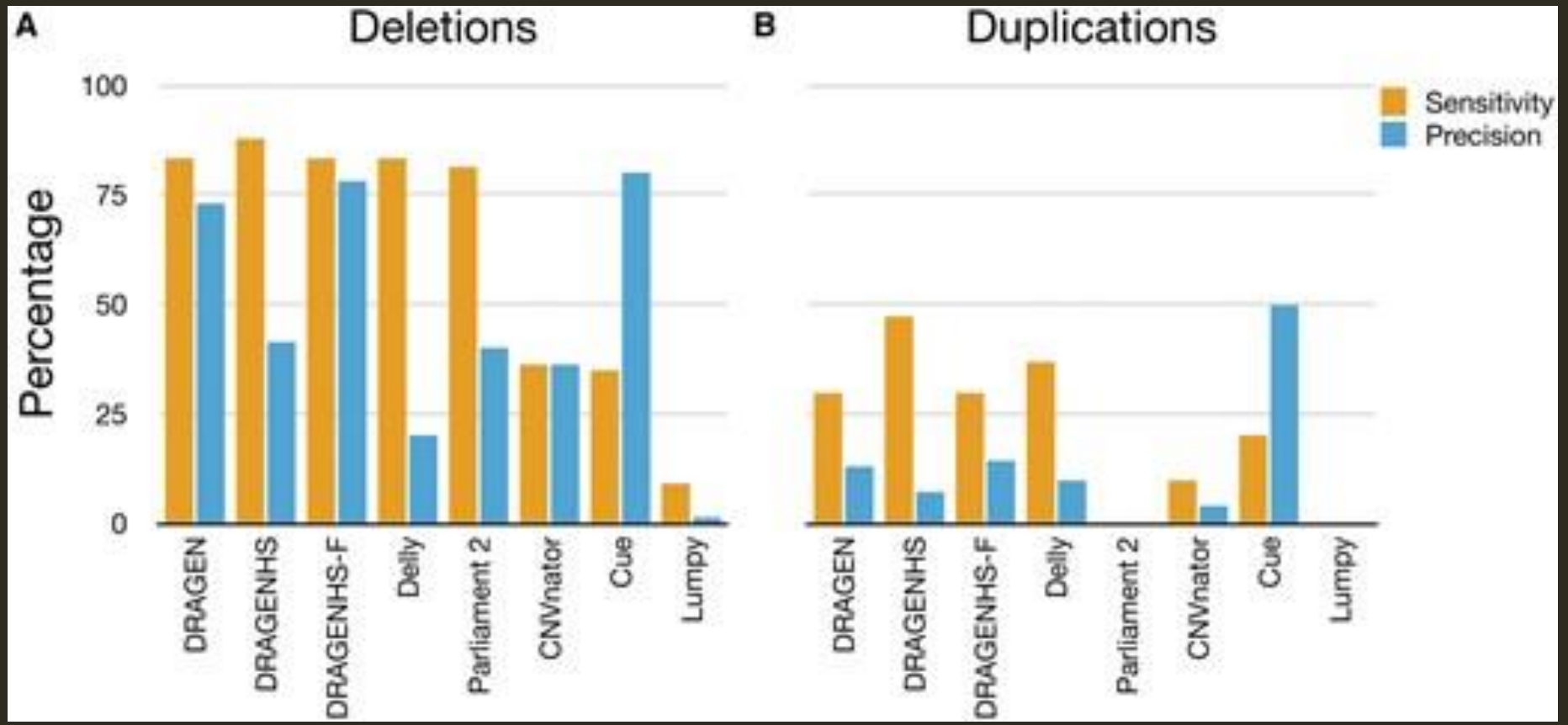


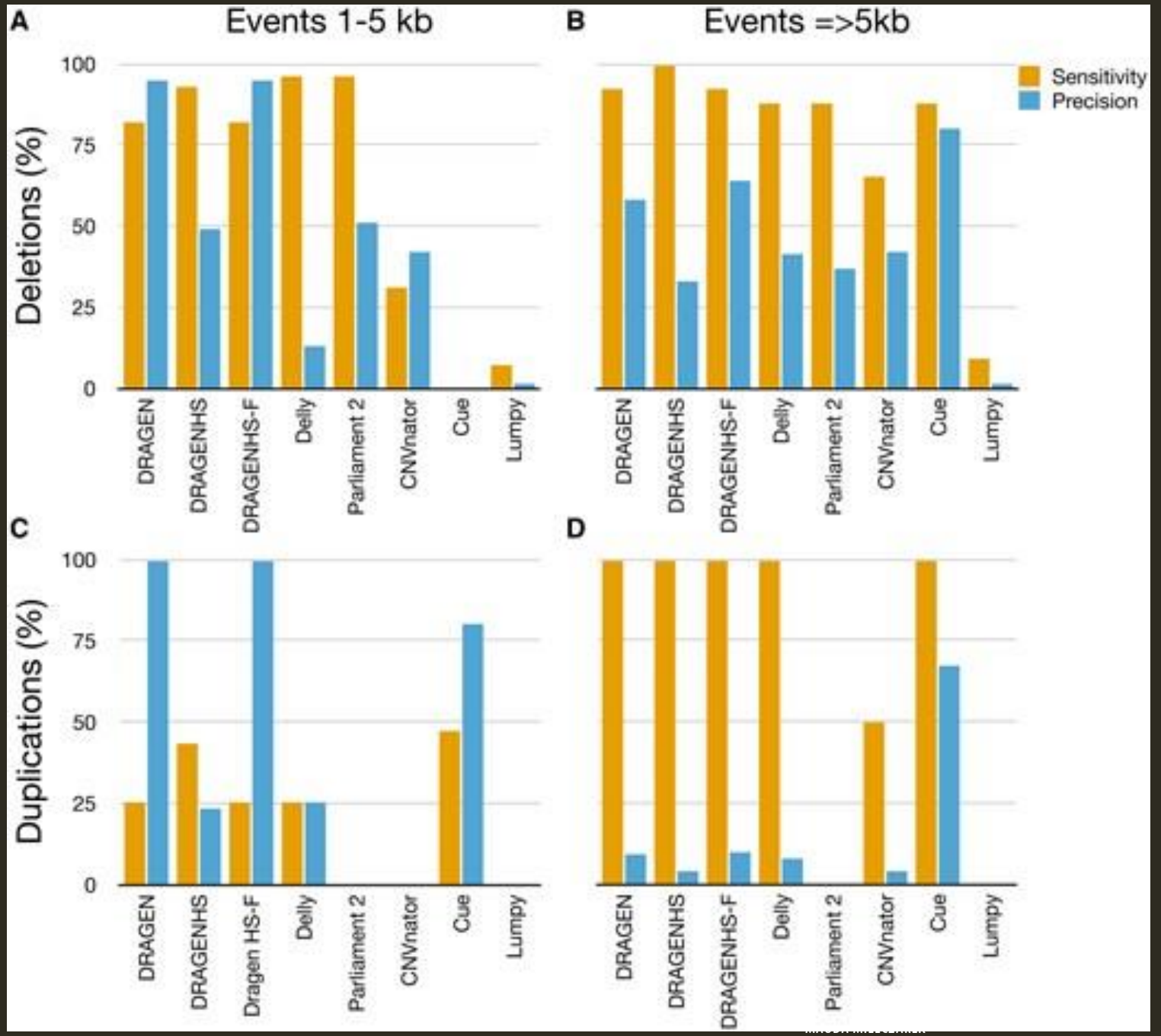
Figure 2.

Genome-wide performance of the CNV callers stratified by CNV type. The performance for deletions followed trends observed in the overall metrics in Fig. 1. Sensitivity for duplications was significantly lower across all callers, with DRAGEN HS achieving the highest sensitivity and Cue achieving the highest precision.

CNV DETECTION (2025)

Figure 3.

Genome-wide performance of the CNV Callers stratified by event length. Results are categorized by event size, either 1–5 kb or larger than 5 kb. All callers exhibited reduced sensitivity for smaller events, with some unable to detect events in the 1–5 kb range.



CNV DETECTION (2025)

5 Conclusions

This study highlights the critical need for continuous benchmarking and refinement of CNV detection tools to meet the demands of clinical diagnostics. DRAGEN v4.2 HS-F, with its adjustable balance between sensitivity and precision, demonstrates strong potential for integrating WGS into clinical diagnostic pipelines. Further advancements in CNV detection methods are essential to enhance analytical accuracy while reducing both costs and reliance on orthogonal validation of WGS results.

illumina

Search Illumina.com

Order Status

Products Learn Company Support | Recommended Links

INFORMATICS PRODUCTS DRAGEN secondary analysis Apps and pipelines Deployment options Server Order Support QUESTIONS

DRAGEN secondary analysis

Accurate, comprehensive, efficient analy

A suite of bioinf
discovery

Speak to a s

Key features and benefits

Accura

Confiden
results. L
score using the
v2 benchmark data.[1]

Comprehensive solution

Analyze whole genomes, exomes, methylomes,
nscriptomes with a single solution that
30 open-source tools.

Efficient analysis

Process a 40× genome in ~ 34 min, with all supported callers.[2] Reduce FASTQ file sizes up to 5× with DRAGEN ORA compression. DRAGEN secondary analysis resulted in two world speed records for genomic data analysis.[3,4]

Cost efficiency

Built-in lossless data compression decreases storage costs by 80%[5]. Preconfigured workflows reduce time and expense for developing and maintaining analysis pipelines.

M
accessibil

Meeting you where your data and
DRAGEN secondary analysis is available vi
premises server, in the cloud, or directly
onboard the NovaSeq X Series, NextSeq 1000
and NextSeq 2000 Systems, and the MiSeq
i100 Series.

Streamlined integration

ith Illumina sequencers,
rkflow from
n tertiary analysis.

REKONSTRUKCJA GENOMÓW NA BAZIE ANTYCZNEGO DNA

PIPELINE (I)

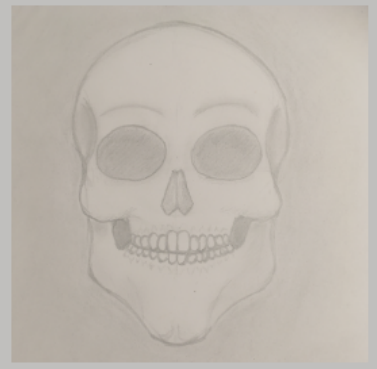
UPPSALA UNIVERSITET

aDNA Atlas

Search docs

Contents:

- Uppmax
- aDNA pipeline
- Authentication
- Merging of bams
- Haplogroups
- PCA



aDNA Pipeline

Current Version

The current verion of the pipeline uses cutadapt v. 2.3 for trimming adapters and FLASH v. 1.2.11 for merging of fastq reads. Cutadapt searches for a predefined adapter sequence and trims the reads if at least 3 bp overlaps between the end of the read and the adapter sequence. The pipeline is also set up in such a way that it is sensitive to dual vs. single indexing, as well as HiSeq vs. NovaSeq sequencing techiques and it accepts a 20% error level in the overlapping region. When the reads are trimmed, FLASH collapses the PE data into a single fastq file if the read-pair overlaps with 11 bp. The FLASH output fastq files (ExtendedFrags, notCombined_1 and and notCombined_2) are then merged into a single fastq file called `CutAdapt-eq_set-FLASH_corrected` and ends with `.all.fastq.gz`. You can find these fastq files in `/proj/snic2020-2-10/1000AncientGenomes/mergedfastqs/`.

This essentially single-end fastq file s then mapped against a reference genome using bwa `aln` `(-l 16500 -n 0.01 -o 2)`. The bamfile will contain all mapped reads (including PCR duplicates, short reads etc) and is located in `/proj/snic2020-2-10/1000AncientGenomes/hg19bams/mapped/`

Per default all deliveries are mapped against human reference genome build 37 (hg19).

Version	Filename
hg18	human_b36_male_nohaps.fa
hg19	hs37d5.fa

BWA

SYNOPSIS

```
bwa index ref.fa

bwa mem ref.fa reads.fq > aln-se.sam

bwa mem ref.fa read1.fq read2.fq > aln-pe.sam

bwa aln ref.fa short_read.fq > aln_sa.sai

bwa samse ref.fa aln_sa.sai short_read.fq > aln-se.sam

bwa sampe ref.fa aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln-pe.sam

bwa bwasw ref.fa long_read.fq > aln.sam
```

DESCRIPTION

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

For all the algorithms, BWA first needs to construct the FM-index for the reference genome (the **index** command). Alignment algorithms are invoked with different sub-commands: **aln/samse/sampe** for BWA-backtrack, **bwasw** for BWA-SW and **mem** for the BWA-MEM algorithm.

bwa aln:

- optimized for short reads, more effective for aligning tiny fragments compared to other aligners or even other modes of BWA
- aDNA typically contains post-mortem damage, especially cytosine deamination, which causes C⟶T and G⟶A mismatches. Bwa aln allows fine-tuning of mismatch penalties and quality thresholds, enabling better accommodation of these typical damage patterns.

BWA ALN

bwa aln
-l 16500
-n 0.01
-o 2

Feature	Modern DNA	Ancient DNA
Read length	Long (100–250 bp)	Short (30–80 bp)
DNA quality	High	Degraded, chemically modified
Seeding	Enabled for speed	Disabled to improve sensitivity
Mismatch allowance	Higher (tolerant)	Lower (conservative to reduce noise)
Gap handling	Simple (1 gap)	More flexible (2 gaps)

OPTIONS:

`-n NUM` Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. [0.04]

`-o INT` Maximum number of gap opens [1]

-e *INT* Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]

-d *INT* Disallow a long deletion within INT bp towards the 3'-end [16]

-i *INT* Disallow an indel within INT bp towards the ends [5]

`-l INT` Take the first INT subsequence as seed. If INT is larger than the query sequence, seeding will be disabled. For long reads, this option is typically ranged from 25 to 35 for '-k 2'. [inf]

PIPELINE (I)

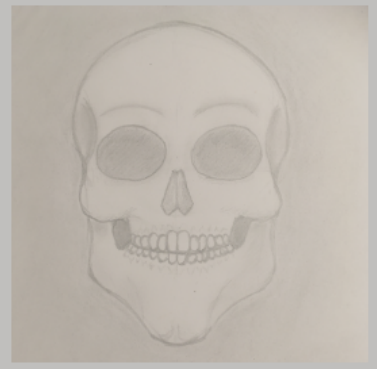
UPPSALA UNIVERSITET

aDNA Atlas

Search docs

Contents:

- Uppmax
- aDNA pipeline
- Authentication
- Merging of bams
- Haplogroups
- PCA



aDNA Pipeline

Current Version

The current verion of the pipeline uses cutadapt v. 2.3 for trimming adapters and FLASH v. 1.2.11 for merging of fastq reads. Cutadapt searches for a predefined adapter sequence and trims the reads if at least 3 bp overlaps between the end of the read and the adapter sequence. The pipeline is also set up in such a way that it is sensitive to dual vs. single indexing, as well as HiSeq vs. NovaSeq sequencing techiques and it accepts a 20% error level in the overlapping region. When the reads are trimmed, FLASH collapses the PE data into a single fastq file if the read-pair overlaps with 11 bp. The FLASH output fastq files (ExtendedFrags, notCombined_1 and and notCombined_2) are then merged into a single fastq file called `CutAdapt-eq_set-FLASH_corrected` and ends with `.all.fastq.gz`. You can find these fastq files in `/proj/snic2020-2-10/1000AncientGenomes/mergedfastqs/`.

This essentially single-end fastq file s then mapped against a reference genome using bwa `aln` `(-l 16500 -n 0.01 -o 2)`. The bamfile will contain all mapped reads (including PCR duplicates, short reads etc) and is located in `/proj/snic2020-2-10/1000AncientGenomes/hg19bams/mapped/`

Per default all deliveries are mapped against human reference genome build 37 (hg19).

Version	Filename
hg18	human_b36_male_nohaps.fa
hg19	hs37d5.fa

BWA ALN I SAMSE

SYNOPSIS

```
bwa index ref.fa

bwa mem ref.fa reads.fq > aln-se.sam

bwa mem ref.fa read1.fq read2.fq > aln-pe.sam

bwa aln ref.fa short_read.fq > aln_sa.sai

bwa samse ref.fa aln_sa.sai short_read.fq > aln-se.sam

bwa sampe ref.fa aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln-pe.sam

bwa bwasw ref.fa long_read.fq > aln.sam
```

samse bwa samse [-n maxOcc] <in.db.fasta> <in.sai> <in.fq> > <out.sam>

Generate alignments in the SAM format given single-end reads. Repetitive hits will be randomly chosen.

PIPELINE (I)

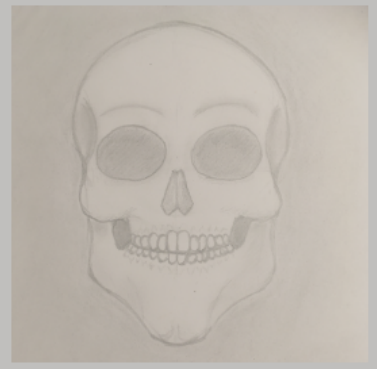
UPPSALA UNIVERSITET

aDNA Atlas

Search docs

Contents:

- Uppmax
- aDNA pipeline
- Authentication
- Merging of bams
- Haplogroups
- PCA



aDNA Pipeline

Current Version

The current verion of the pipeline uses cutadapt v. 2.3 for trimming adapters and FLASH v. 1.2.11 for merging of fastq reads. Cutadapt searches for a predefined adapter sequence and trims the reads if at least 3 bp overlaps between the end of the read and the adapter sequence. The pipeline is also set up in such a way that it is sensitive to dual vs. single indexing, as well as HiSeq vs. NovaSeq sequencing techiques and it accepts a 20% error level in the overlapping region. When the reads are trimmed, FLASH collapses the PE data into a single fastq file if the read-pair overlaps with 11 bp. The FLASH output fastq files (ExtendedFrags, notCombined_1 and and notCombined_2) are then merged into a single fastq file called `CutAdapt-eq_set-FLASH_corrected` and ends with `.all.fastq.gz`. You can find these fastq files in `/proj/snic2020-2-10/1000AncientGenomes/mergedfastqs/`.

This essentially single-end fastq file is then mapped against a reference genome using bwa `aln` `(-l 16500 -n 0.01 -o 2)`. The bamfile will contain all mapped reads (including PCR duplicates, short reads etc) and is located in `/proj/snic2020-2-10/1000AncientGenomes/hg19bams/mapped/`

Per default all deliveries are mapped against human reference genome build 37 (hg19).

Version	Filename
hg18	human_b36_male_nohaps.fa
hg19	hs37d5.fa

HS37DS.FA

Description

Full 1000genomes Phase2 Reference Genome Sequence (hs37d5), based on NCBI GRCh3

Note

This BSgenome data package was made from the following source data file:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

The genome is composed of:

- Integrated reference sequence from the GRCh37 primary assembly (chromosomal pl calized and unplaced contigs)
- The rCRS mitochondrial sequence (AC:NC_012920)
- Human herpesvirus 4 type 1 (AC:NC_007605)
- Concatenated decoy sequences (hs37d5cs.fa.gz)

For details, please see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refe phase2_reference_assembly_sequence/README_human_reference_20110707.

Author(s)

Julian Gehring <julian.gehring@embl.de>

Homo sapiens mitochondrion, complete genome

NCBI Reference Sequence: NC_012920.1

FASTA Graphics

Go to:

```
LOCUS       NC_012920               16569 bp    DNA     circular PRI 03-APR-2023
DEFINITION  Homo sapiens mitochondrion, complete genome.
ACCESSION   NC_012920 AC_000021
VERSION     NC_012920.1
DBLINK      BioProject: PRJNA927338
KEYWORDS    RefSeq.
SOURCE      mitochondrion Homo sapiens (human)
```

This Revised Cambridge Reference Sequence (rCRS) has eighteen specific corrections or confirmations of the original 1981 sequence of Anderson et al [PMID:7219534]. Seven nucleotides are confirmed as rare polymorphisms, maintained as: 263A, 311C-315C, 750A, 1438A, 4769A, 8860A, and 15326A. Eleven nucleotides are error corrections: 3107del, 3423T, 4985A, 9559C, 11335C, 13702C, 14199T, 14272C, 14365C, 14368C, and 14766C. These 11 errors in the original Cambridge sequence were determined to be either outright sequencing errors (8 instances) or due to the presence of bovine DNA (2 instances) or HeLa DNA (1 instance) mixed in with the original human placental DNA [PMID:10508508]. HISTORICAL NUCLEOTIDE NUMBERS ARE MAINTAINED by indicating 3107del as 'N'. A summary table of the reanalysis data is available online at http://www.mitomap.org/MITOMAP/CambridgeReanalysis

HS37DS.FA

Description

Full 1000genomes Phase2 Reference Genome Sequence (hs37d5), based on NCBI GRCh3

Note

This BSgenome data package was made from the following source data file:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_refer
sequence/hs37d5.fa.gz

The genome is composed of:

- Integrated reference sequence from the GRCh37 primary assembly (chromosomal plu
calized and unplaced contigs)
- The rCRS mitochondrial sequence (AC:NC_012920)
- Human herpesvirus 4 type 1 (AC:NC_007605)
- Concatenated decoy sequences (hs37d5cs.fa.gz)

For details, please see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refe
phase2_reference_assembly_sequence/README_human_reference_20110707.

Author(s)

Julian Gehring <julian.gehring@embl.de>

Human gammaherpesvirus 4, complete genome

NCBI Reference Sequence: NC_007605.1

FASTA Graphics

Go to:

```
LOCUS       NC_007605              171823 bp    DNA     circular VRL 13-AUG-2018
DEFINITION  Human gammaherpesvirus 4, complete genome.
ACCESSION   NC_007605
VERSION     NC_007605.1
DBLINK      BioProject: PRJNA485481
```

→ improves read mapping: many human samples, especially those derived from lymphoblastoid cell lines (commonly used in the 1000 Genomes Project), contain EBV DNA because these cells are immortalized using the virus. EBV reads would otherwise map incorrectly to the human genome, potentially generating FP variant calls or misalignments. Including the EBV sequence in the reference can correctly map these reads to the viral genome, reducing mapping artifacts.

HS37DS.FA

Description

Full 1000genomes Phase2 Reference Genome Sequence (hs37d5), based on NCBI GRCh3

Note

This BSgenome data package was made from the following source data file:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_refer
sequence/hs37d5.fa.gz

The genome is composed of:

- Integrated reference sequence from the GRCh37 primary assembly (chromosomal plu
calized and unplaced contigs)
- The rCRS mitochondrial sequence (AC:NC_012920)
- Human herpesvirus 4 type 1 (AC:NC_007605)
- Concatenated decoy sequences (hs37d5cs.fa.gz)

For details, please see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refe
phase2_reference_assembly_sequence/README_human_reference_20110707.

Author(s)

Julian Gehring <julian.gehring@embl.de>

→additional genomic sequences appended to the standard human reference genome (GRCh37/hg19) to improve read alignment accuracy.

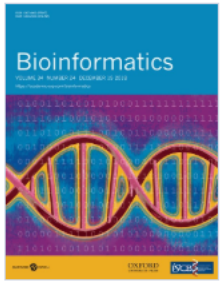
→ includes various unplaced or alternative sequences not represented in the primary reference assembly, such as contigs from other human genome assemblies, and repetitive or difficult-to-map regions.

→ represents complex or missing regions of the genome → provides alternative mapping targets, thereby improving the completeness and accuracy of genomic analyses.

VARIANT CALLER

Bioinformatics

Issues Advance articles Submit ▾ Alerts About



Volume 34, Issue 24
December 2018

JOURNAL ARTICLE

snpAD: an a

Kay Prüfer

Bioinformatics, Volu
https://doi.org/10.1
Published: 21 June

PDF Split

Abstract

Motivation

The study of ancient genomes can elucidate the evolutionary past. However, analyses are complicated by base-modifications in ancient DNA molecules that result in errors in DNA sequences. These errors are particularly common near the ends of sequences and pose a challenge for genotype calling.

Results

I describe an iterative method that estimates genotype frequencies and errors along sequences to allow for accurate genotype calling from ancient sequences. The implementation of this method, called snpAD, performs well on high-coverage ancient data, as shown by simulations and by subsampling the data of a high-coverage Neandertal genome. Although estimates for low-coverage genomes are less accurate, I am able to derive approximate estimates of heterozygosity from several low-coverage Neandertals. These estimates show that low heterozygosity, compared to modern humans, was common among Neandertals.

Software | Open access | Published: 31 March 2016

EAGER: efficient ancient genome reconstruction

Alexander Peltzer, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause & Kay Nieselt

Genome Biology **17**, Article number: 60 (2016) | Cite this article

Genome Biology

PeerJ. 2021; 9: e10947.
Published online 2021 Mar 16. doi: 10.7717/peerj.10947

PMCID: PMC7977378
PMID: 33777521

Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager

James A. Fellows Yates,[1,2] Thiseas C. Lamnidis,[1] Maxime Borry,[1] Aida Andrades Valtueña,[1] Zandra Fagernäs,[1] Stephen Clayton,[1] Maxime U. Garcia,[3,4] Judith Neukamm,[5,6] and Alexander Peltzer[1,7]

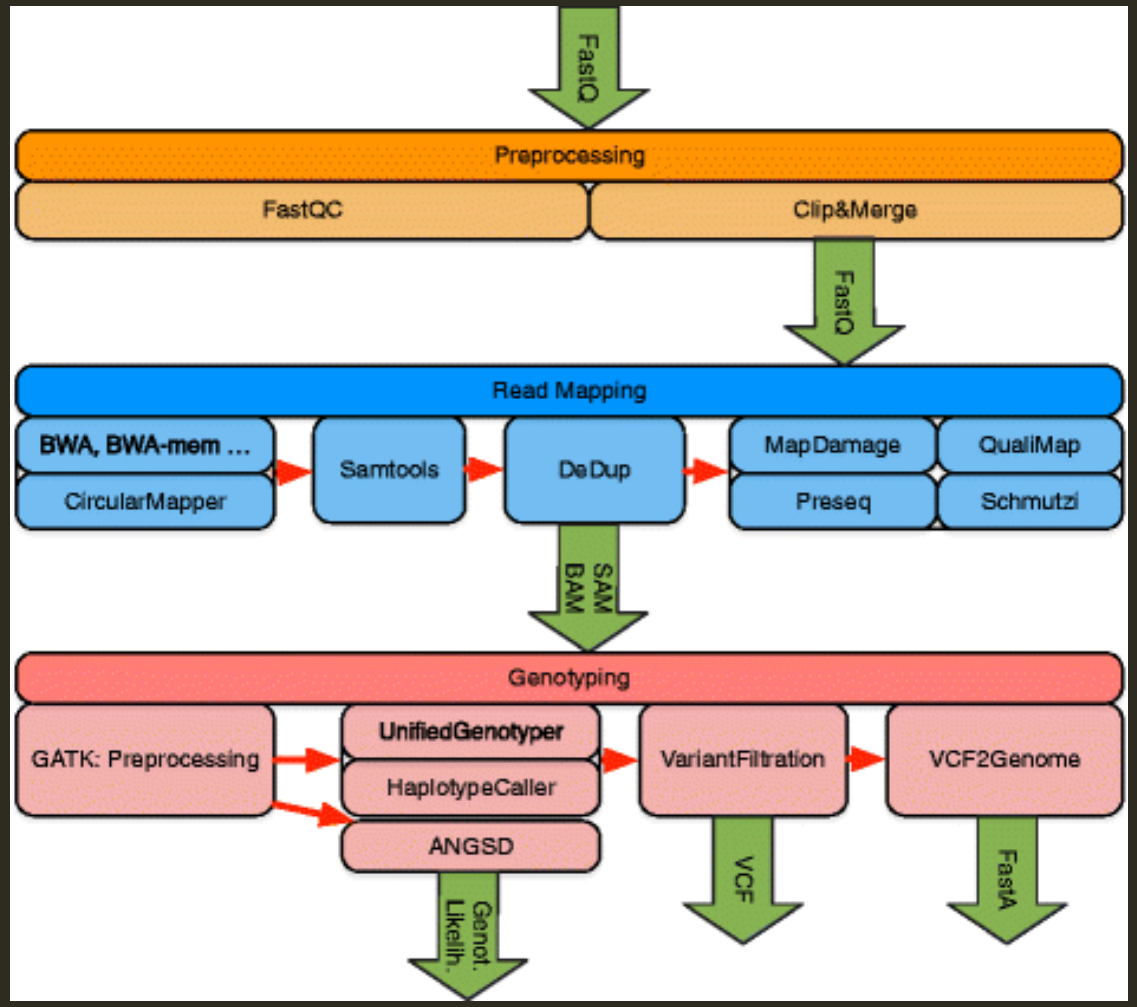
Academic Editor: Alexander Schliep

PeerJ Life & Environment

ANTIENT DNA

- EAGER to zautomatyzowany potok przetwarzania danych, który ma za zadanie uprościć analizę wielkoskalowych zbiorów danych genomowych.
- EAGER zapewnia funkcje umożliwiające:
  - wstępne przetwarzanie, mapowanie i ocenę jakości próbek aDNA.
  - genotypowania próbek w celu detekcji, filtrowania i analizowania wariantów genetycznych.

EAGER

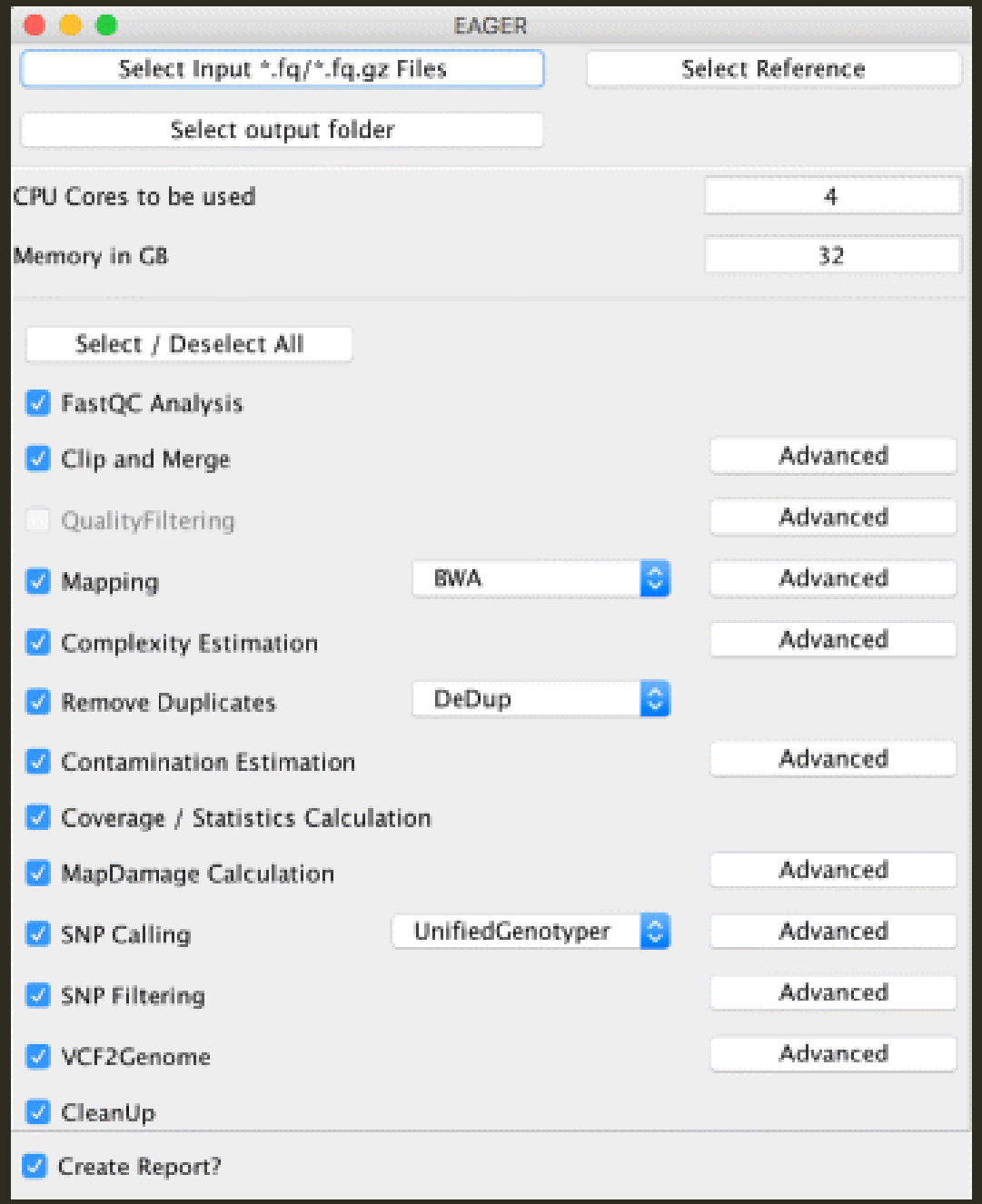


Workflow diagram of the EAGER pipeline.

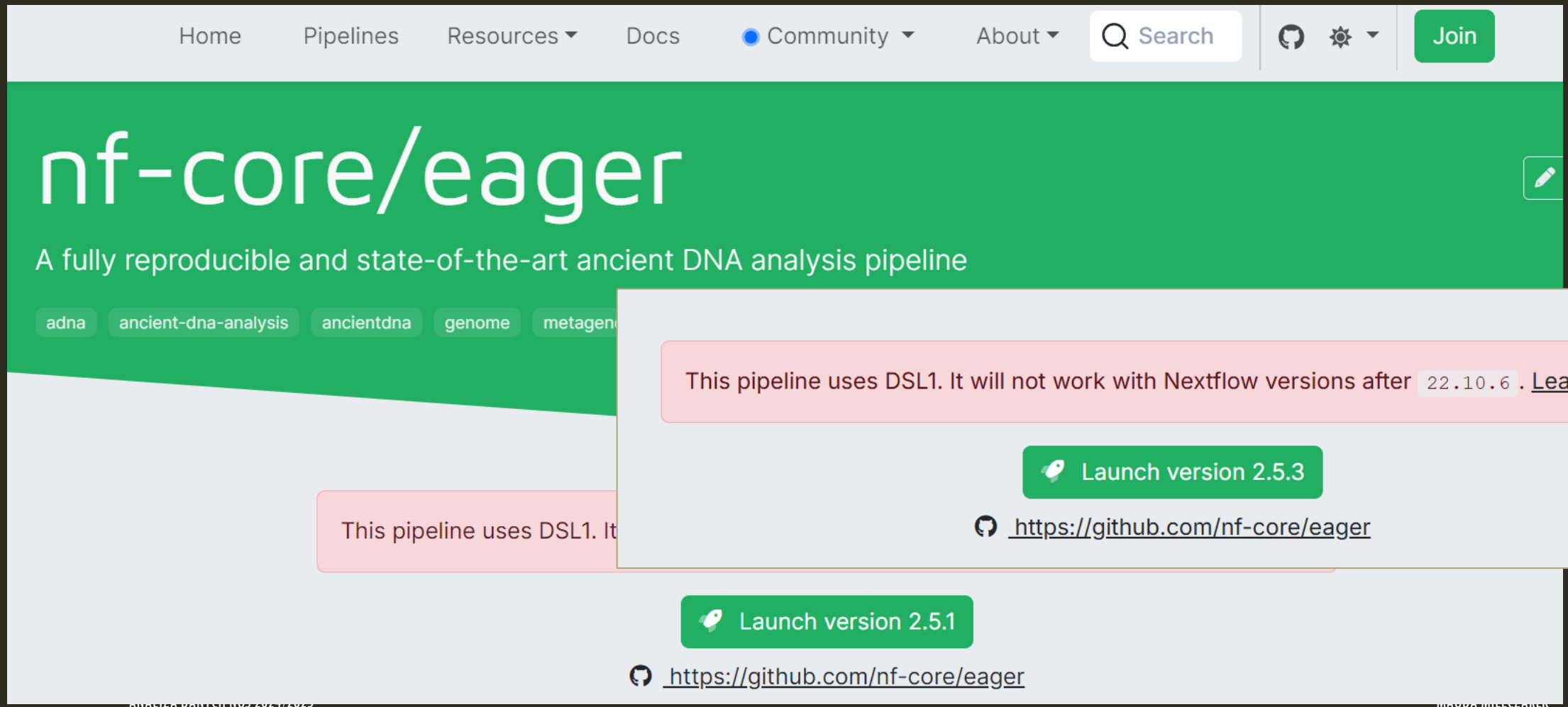
The pipeline consists of three distinct main components for processing and analysis of NGS data:

- preprocessing
- read mapping
- genotyping

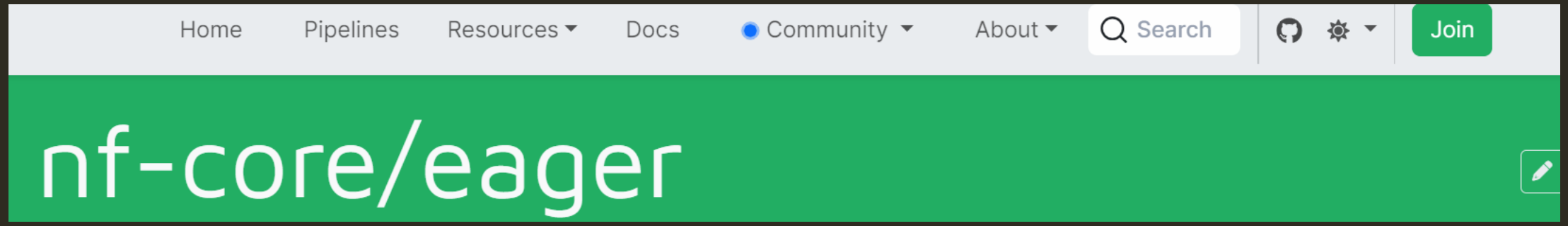
GUI


EAGER
Select Input *.fq/*.fq.gz Files
Select Reference
Select output folder
CPU Cores to be used
4
Memory in GB
32
Select / Deselect All
FastQC Analysis
Clip and Merge
Advanced
QualityFiltering
Advanced
Mapping
BWA
Advanced
Complexity Estimation
Advanced
Remove Duplicates
DeDup
Contamination Estimation
Advanced
Coverage / Statistics Calculation
MapDamage Calculation
Advanced
SNP Calling
UnifiedGenotyper
Advanced
SNP Filtering
Advanced
VCF2Genome
Advanced
CleanUp
Create Report?

EAGER OD NF/CORE NEXTFLOW



EAGER OD NF/CORE NEXTFLOW



The typical command for running the pipeline is as follows:

nextflow run nf-core/eager --input '*_R{1,2}.fastq.gz' --fasta 'some.fasta' -profile standard,docker

REWOLUCJA W MEDYCYNIE

MEDYCYNA GENOMOWA (49:00-1:22:00)

"Różne oblicza biologii" - Wydarzenie Specjalne dla uhonorowania prof. Magdaleny Fikus - 27 FN 2023

PRECISION ONCOLOGY

nature medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature medicine › analyses › article

Analysis | Open access | Published: 11 January 2024

Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme

Alona Sosinsky, John Ambrose, William Cross, Clare Turnbull, Shirley Henderson, Louise Jones, Angela Hamblin, Prabhu Arumugam, Georgia Chan, Daniel Chubb, Boris Noyvert, Jonathan Mitchell, Susan Walker, Katy Bowman, Dorota Pasko, Marianna Buongermino Pereira, Nadezda Volkova, Antonio Rueda-Martin, Daniel Perez-Gil, Javier Lopez, John Pullinger, Afshan Siddiq, Tala Zainy, Tasnim Choudhury, ... Nirupa Murugaesu ✉ + Show authors

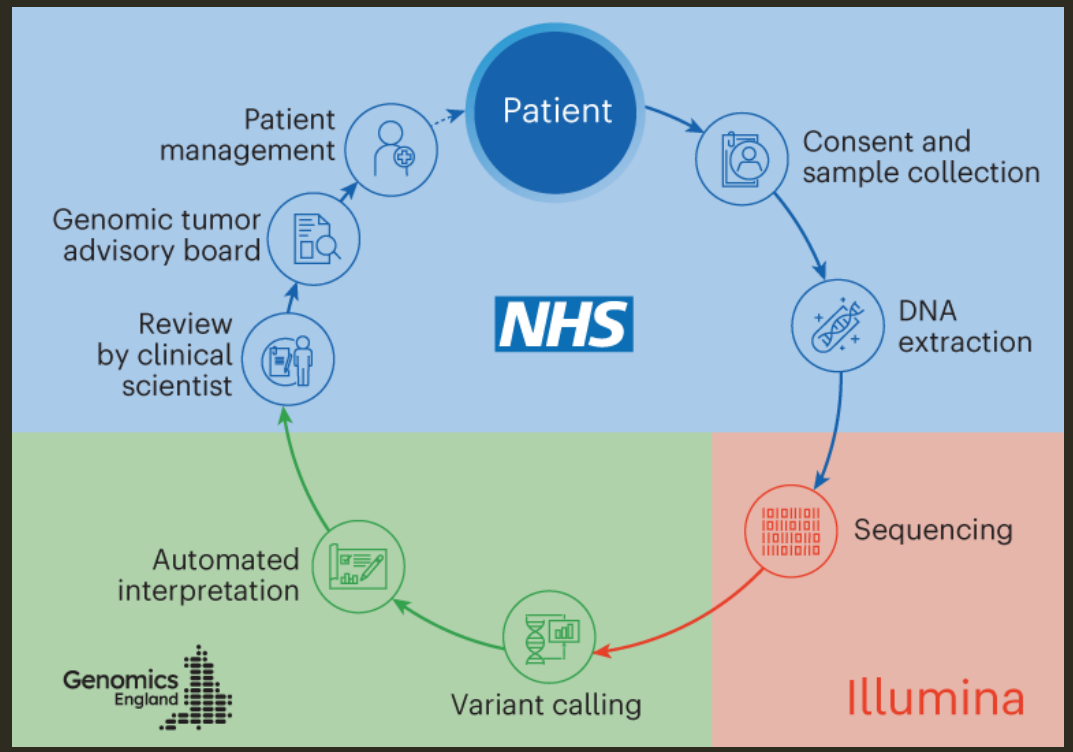
CEL

The 100,000 Genomes Project, a UK Government initiative conducted within the National Health Service (NHS) in England, aimed to establish standardized high-throughput whole-genome sequencing (WGS) for patients with cancer and rare diseases via an automated, International Organization for Standardization-accredited bioinformatics pipeline (providing clinically accredited variant calling and variant prioritization).

Genomics England, alongside NHS England, analyzed WGS data from 13,880 solid tumors spanning 33 cancer types, integrating genomic data with real-world treatment and outcome data, within a secure Research Environment.

A longer-term objective was to accelerate the delivery of molecular testing, including WGS, in NHS clinical cancer care.

JOURNEY OF THE PATIENT'S GENOME



Patients provided written informed consent for WGS analysis.

DNA was extracted from tumor and normal (blood) samples using standardized protocols and samples were submitted for WGS, which was performed on an Illumina sequencer.

An automated pipeline was constructed for sequence quality control, alignment, variant calling and interpretation, with results returned to the 13 NHS Genomic Medicine Centers for review.

WYNIKI

Incidence of somatic mutations in genes recommended for standard-of-care testing varied across cancer types

→ in glioblastoma multiforme, small variants were present in 94% of cases

→ sarcoma demonstrated the highest occurrence of actionable structural variants (13%).

→ Homologous recombination deficiency was identified in 40% of high-grade serous ovarian cancer cases with 30% linked to pathogenic germline variants, highlighting the value of combined somatic and germline analysis.

The linkage of WGS and longitudinal life course clinical data allowed the assessment of treatment outcomes for patients stratified according to pangenomic markers.

*glejak wielopostaciowy, mięsak ,rak jajnika

VARIANT CALLING IN CLINICAL SEQUENCING

Genome Medicine

Home About Articles Submission Guidelines Collections Submit manuscript

Review | Open access | Published: 26 October 2020

Best practices for variant calling in clinical sequencing

Daniel C. Koboldt

Genome Medicine **12**, Article number: 91 (2020) | Cite this article

183k Accesses | **209** Citations | **19** Altmetric | Metrics

Review | Open access | Published: 26 October 2020

Best practices for variant calling in clinical sequencing

Daniel C. Koboldt

Genome Medicine **12**, Article number: 91 (2020) | Cite this article

Abstract

Next-generation sequencing technologies have enabled a dramatic expansion of clinical genetic testing both for inherited conditions and diseases such as cancer. Accurate variant calling in NGS data is a critical step upon which virtually all downstream analysis and interpretation processes rely. Just as NGS technologies have evolved considerably over the past 10 years, so too have the software tools and approaches for detecting sequence variants in clinical samples. In this review, I discuss the current best practices for variant calling in clinical sequencing studies, with a particular emphasis on trio sequencing for inherited disorders and somatic mutation detection in cancer patients. I describe the relative strengths and weaknesses of panel, exome, and whole-genome sequencing for variant detection. Recommended tools and strategies for calling variants of different classes are also provided, along with guidance on variant review, validation, and benchmarking to ensure optimal performance. Although NGS technologies are continually evolving, and new capabilities (such as long-read single-molecule sequencing) are emerging, the "best practice" principles in this review should be relevant to clinical variant calling in the long term.

Emphasis on *trio* sequencing for inherited disorders and somatic mutation detection in cancer patients

Strengths and weaknesses of:
- panel
- exome
- whole-genome sequencing

for variant detection.

Recommended ***tools*** and ***strategies*** for calling variants of different classes

The "best practice" principles in this review should be relevant to clinical variant calling in the long term.

BACKGROUND

NGS technologies enabled ambitious large-scale genomic sequencing efforts that have transformed our understanding of human health and disease, such The Cancer Genome Atlas, the Centers for Mendelian Genomics.

They have also been widely *adopted for clinical genetic testing*:

→ *Whole-exome sequencing*, which selectively targets the protein-coding regions of known genes, has become a frontline diagnostic tool for inherited disorders

→ Targeted panels to interrogate medically relevant subsets of genes have become core components of precision oncology.

THE AIM

Presenting "***best practices***" for variant calling in clinical sequencing for both:
- germline analysis in family trios
- somatic analysis of tumor-normal pairs

This includes ***recommendations*** for the choice of:
- sequencing strategy
- NGS read alignment and variant calling preprocessing
- rigorous filtering to remove FP
- guidance on benchmarking NGS analysis pipeline performance using "gold standard" reference datasets to achieve the optimum balance of sensitivity and specificity

SEQUENCING STRATEGIES

The choice of sequencing strategy for a clinical sample has important consequences for variant calling:

→ *Gene panels* are increasingly cost-effective means of testing for subsets of genes associated with *specific clinical phenotypes*.
Numerous gene panels are commercially available, ranging in size from a single gene to hundreds of genes.

The example: OtoSCOPE hearing loss panel targets 89 genes and microRNAs associated with hearing loss (1574 total exons); across a cohort of 711 sequenced patients, the average sequence depth achieved was *716×* per patient.

→ *Exome sequencing*, which targets *~ 20,000 protein-coding genes*, typically achieves *> 100×* average depth across the target regions.

→ *Whole-genome sequencing* offers the most comprehensive approach and typically yields *~ 30–60×* average sequence depth across the entire genome.

Other considerations, such as cost and turnaround time, also influence the choice of sequencing strategy but are beyond the scope of this review.

SEQUENCING STRATEGIES

All 3 strategies *generally offer excellent sensitivity for detecting SNVs/indels* using tools such as GATK HaplotypeCaller and Platypus.

CNVs spanning multiple exons can be called with reasonable sensitivity using panel and exome data.

Whole-genome sequencing remains the superior strategy for the comprehensive detection *of all types of sequence variants.* However, it should be noted that the higher sequence depth achieved in panel and exome sequencing may enable more sensitive detection of variants at low allele frequencies

Strategy	Panel	Exome	Genome
Size of target space (Mbp)	~ 0.5	~ 50	~ 3200
Average read depth	500–100×	100–150×	~ 30–60×
Relative cost	$	$$	$$$
SNV/indel detection	++	++	++
CNV detection	+	+	++
SV detection	–	–	+
Low VAF	++	+	+

Dollar signs represent approximate relative costs, though it should be noted that the cost of panel sequencing depends on the size of the panel. The empirical performance of each strategy for detecting variants of different classes is indicated as good (+), outstanding (++), or poor/absent (–)

BTW. This is platypus ☺ →

ALIGNMENT & POST-ALN & SNP CALLING PRE-PROCESSING

Alignment:

→ to a reference genome is a *critical phase* of NGS analysis. Typically

→ BWA-Mem + Samtools

Post-alignment:

→ identify redundant reads that originated from the same DNA sequence molecule (5–15%) of sequencing reads in a typical exom

→ Picard and Sambamba identify and mark duplicate reads in a BAM file to exclude them from downstream analysis.

ALIGNMENT & POST-ALN & SNP CALLING PRE-PROCESSING

SNP-calling pre-processing (form The GATK Best Practices workflow)

→ The second is local realignment around indels, which aims to reduce false-positive variant calls caused by alignment artifacts

→ base quality score recalibration (BQSR), which adjusts the base quality scores of sequencing reads.

Base quality score recalibration (BQSR) is a process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly. For example we can identify that, for a given run, whenever we called two A nucleotides in a row, the next base we called had a 1% higher rate of error. So any base call that comes after AA in a read should have its quality score reduced by 1%. We do that over several different covariates (mainly sequence context and position in read, or cycle) in a way that is additive. So the same base may have its quality score increased for one reason and decreased for another.

ALIGNMENT & POST-ALN & SNP CALLING PRE-PROCESSING

SNP-calling pre-processing (form The GATK Best Practices workflow)

→ The second is local realignment around indels, which aims to reduce false-positive variant calls caused by alignment artifacts

→ base quality score recalibration (BQSR), which adjusts the base quality scores of sequencing reads.

→ „Evaluations of variant calling accuracy before and after BQSR/realignment suggest that the improvements are marginal; because of this and the high computational cost, this may be viewed as an optional step for pre-processing."

ALIGNMENT & POST-ALN & SNP CALLING PRE-PROCESSING

→ Routine quality control (QC) of analysis-ready BAMs should be performed prior to variant calling to evaluate key sequencing metrics → sequencing coverage

→ In the case of family studies and paired samples (e.g., tumor-normal), expected sample relationships should be confirmed with tools for relationship inference such as the KING algorithm.

ALIGNMENT & POST-ALN & SNP CALLING PRE-PROCESSING

KING Tutorial: Relationship Inference

KING is a toolset to explore genotype data from a genome-wide association study (GWAS) or a sequencing project. The latest version is KING **2.3.1** available on July 28, 2023. KING can be used to check family relationship and flag pedigree errors by estimating kinship coefficients and inferring IBD segments for all pairwise relationships. Unrelated pairs can be precisely separated from close relatives with no false positives, with accuracy up to 3rd- or 4th-degree (depending on array or WGS) for --related and --ibdseg analyses, and up to 2nd-degree for --kinship analysis.

This tutorial discusses different types of relationship inference such as the kinship coefficient estimates and the IBD segment inference, as well as derived applications such as pedigree reconstruction and extraction of a subset of unrelated individuals. Other applications of KING such as visualization of families, Quality Control (QC), the identification of population substructure or gene mapping are described elsewhere.

Family relationship inference in KING is very **FAST** (seconds to identify all close relatives in 10,000s of samples), and **robust** to a number of realistic scenarios including the presence of population structure. The number of samples in the dataset can be as small as 2 (for --kinship inference), or as large as > 10,000,000 (for --duplicate and --related inferences). Genome-wide SNPs are required in KING. Please **do not prune or filter** any "good" SNPs that pass QC prior to any KING inference, unless the number of variants is too many to fit the computer memory, *e.g.*, > 100,000,000 as in a WGS study, in which case rare variants can be filtered out. LD pruning is not recommended in KING.

GENERAL INPUT FILES

The input files for KING need to be in PLINK binary format, which include a binary genotype file, a family file, and a map file, *e.g.*, *ex.bed*, *ex.fam*, and *ex.bim*. A binary format allows efficient compression of genotype data by using two bits to represent a genotype, which offers substantial computational savings that are essential to KING analysis. The amount of computer memory required by KING analysis is modest, at $\sim N \times M / 4$ (where N is the number of samples and M is the number of SNPs) plus a small percentage of overhead cost. *E.g.*, for a dataset consisting of 100,000 samples each genotyped at 1,000,000 SNPs, the required memory size is ~25GB. Examples of reading in a dataset are:

BENCHMARKING RESOURCES FOR VARIANT CALLING

Evaluating the *accuracy of variant calls* requires access to benchmark datasets in which the *true variants are already known*.

Each benchmarking dataset includes a set of "ground truth" small variant calls (SNVs and indels) based on the consensus of several variant calling tools, as well as defining the "high-confidence" regions of the human genomes in which variant calls can be benchmarked against a variety of public resources.

BENCHMARKING RESOURCES FOR VARIANT CALLING

Several such benchmarking resources have been made publicly available in recent years. The most widely used ones include:

→ the *Genome in a Bottle* (GIAB); Dataset has been continually improved with the addition of data from multiple short-read and linked-read sequencing datasets.

→ the *Platinum Genome* datasets for NA12878, a human sample of European ancestry that has been sequenced with various technologies at laboratories around the world.

BEST PRACTICES FOR GERMLINE VARIANT CALLING

Dozens of variant calling tools for NGS data have been published in the past 10 years, and countless more have been developed by researchers for internal use.

Because SNV/indel detection tools such as GATK HaplotypeCaller have demonstrated high accuracy in numerous benchmark datasets, choosing a single variant caller that meets the needs of the laboratory (in terms of pipeline compatibility and ease of implementation) is usually sufficient.

However, combining the results of two orthogonal SNV/indel callers, such as HaplotypeCaller and Platypus, may offer a slight sensitivity advantage. Software packages such as BCFtools make it possible to merge multiple variant callsets (in VCF format) into one.

BEST PRACTICES FOR GERMLINE VARIANT CALLING

To discuss the recommended best practices for germline variant calling, we will consider trio sequencing for inherited disorders, which is a common scenario for clinical genetic testing.

A trio analysis pipeline typically begins with the analysis-ready BAM files for the proband and both parents.
For optimal results, all three samples should be sequenced under identical protocols (capture kit, instrument, and reagent kit) and processed with identical alignment and pre-processing steps. This is particularly important for copy number variant calling and SV calling, which rely on uniform sequencing depth and library insert size, respectively.

BEST PRACTICES FOR GERMLINE VARIANT CALLING

Individual versus joint variant calling

All variant calling tools can be applied to individual samples → may be desirable for laboratories processing large numbers of samples.

Individual VCF files can be merged later using e.g. BCFtoos.

…however, VCF files typically only contain entries for positions that are variant in a particular sample → when a variant is only detected in some samples but not others, it is not clear whether the other samples are wild type for that position or simply did not achieve sufficient coverage for the variant caller to make a call.

BEST PRACTICES FOR GERMLINE VARIANT CALLING

Individual versus joint variant calling

Joint variant calling considers all samples simultaneously

Key advantages:

→ it produces called genotypes for every sample at all variant positions. This makes it possible to differentiate between a position that matches the reference sequence with high probability and a position in which the sample did not achieve sufficient coverage.

→ joint calling enables direct inference of phase information (nice for trio)

→ it allows a variant caller to use information from one sample to infer the most likely genotype in another, which has been shown to increase the sensitivity of variant calling in low-coverage regions

BEST PRACTICES FOR GERMLINE VARIANT CALLING

SNV/indel calling

Multiple tools has been created (Samtools/BCFtools, FreeBayes, GATK HaplotypeCaller, Platypus etc.)

Numerous studies have compared the performance of these tools on various datasets → similar results, variant concordance is typically 80–90% concordance or higher, with most differences are attributed to variants at low-coverage or low-confidence positions.

Even so, such differences could amount to thousands of variant calls genome-wide. Thus, it is important not only to choose a robust variant caller for SNVs/indels, but also to benchmark to achieve optimal performance on the data to be analyzed.

BEST PRACTICES FOR GERMLINE VARIANT CALLING

Filtering to remove artifacts

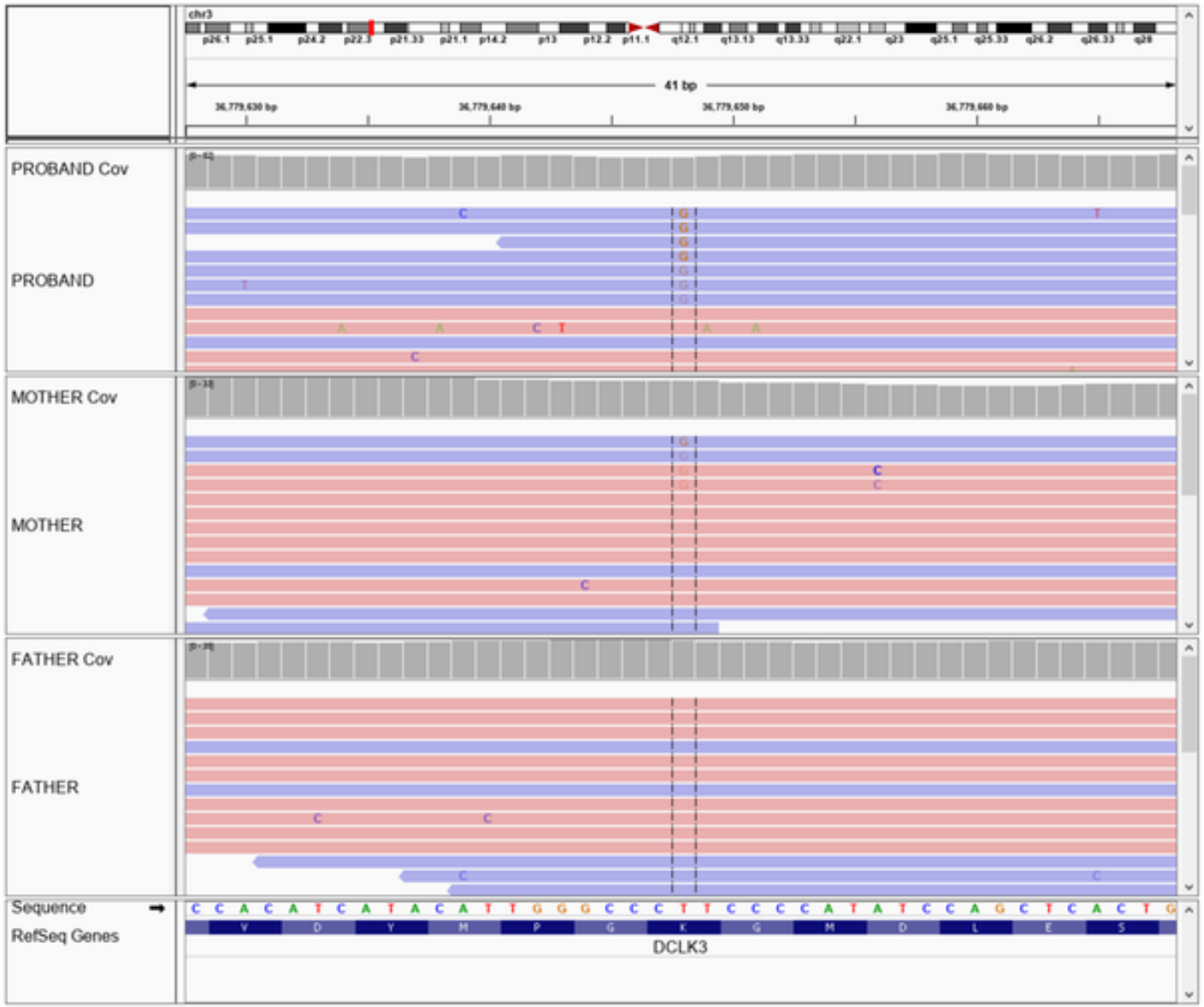
The accuracy of NGS variant calls relative to the previous “gold standard” of Sanger sequencing has been well documented at > 99%.

However, NGS data are prone to certain types of artifactual variant calls, many of which are related to errors in short-read alignment.

→ artifacts should be systematically filtered

→ visual review of the alignments for clinically relevant variants is recommended to identify false-positive variant calls that slip past automated filters.

A.



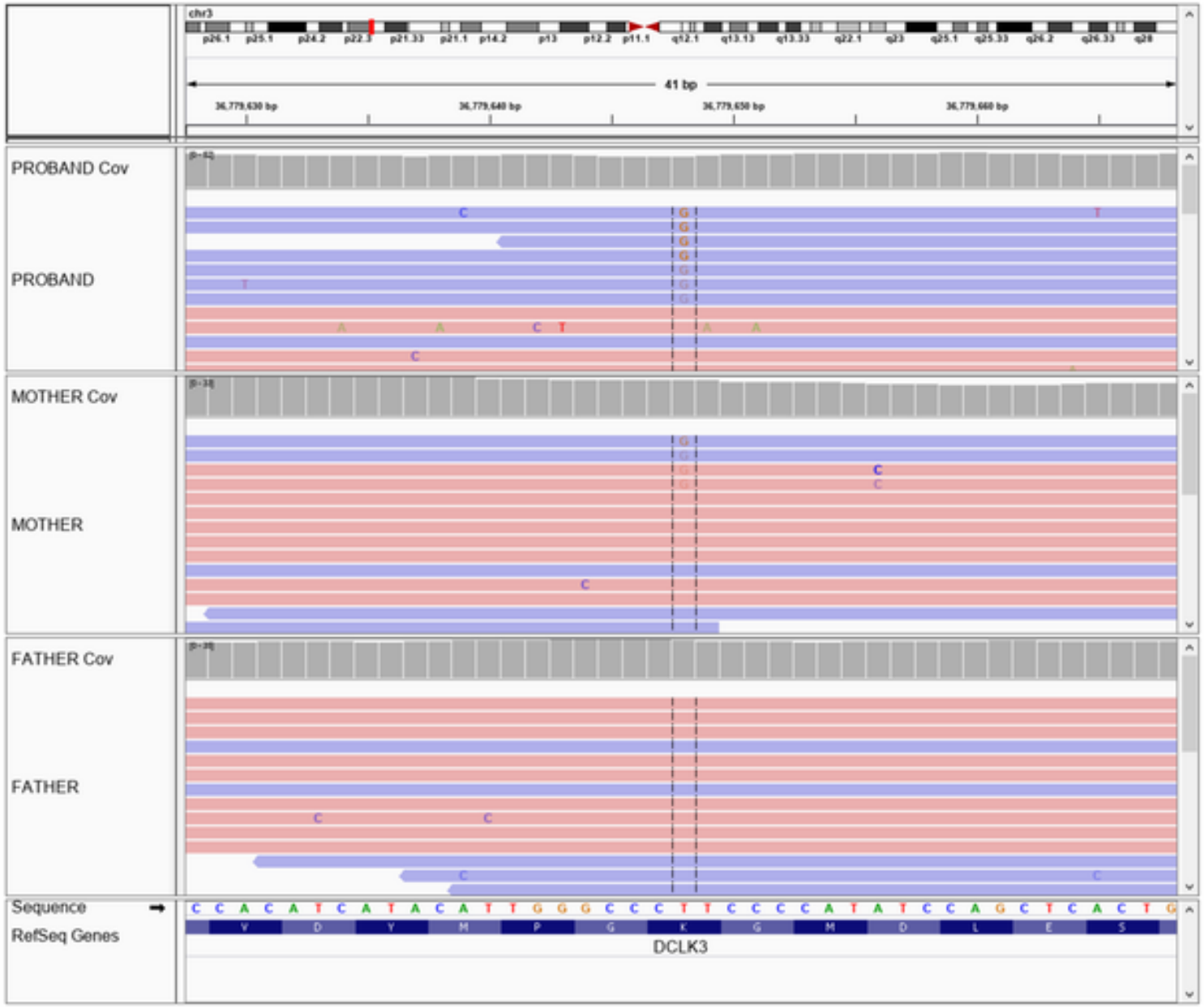
IGV TOOL

Each pane is an IGV screenshot of WGS alignments for the proband (top track), mother, (middle track), and father (bottom track).

Each sample's track comprises two parts: a histogram of the read depth and the reads as aligned to the reference sequence. Reads are colored according to the aligned strand (red = forward strand; blue = reverse strand).

Common artifacts in NGS alignments that gave rise to a false-positive de novo mutation call in a family trio

A.



A. False positive associated with low base quality. Most reads supporting the variant have low base quality indicated by lightly shaded non-reference bases. Four reads in the proband showed the alternate allele with good quality, triggering the variant call.

MAGDA MIELCZAREK

Common artifacts in NGS alignments that gave rise to a false-positive de novo mutation call in a family trio

B.

B. False positive due to misalignments near the start or end of reads.

Notice that the alternate allele is only observed at the start/end of reads in the proband. In this case, the read depth histogram provides a clue as to the cause of the misalignment. As shown in the next panel, this occurs at the breakpoint of a large paternally inherited deletion.

MAGDA MIELCZAREK

Common artifacts in NGS alignments that gave rise to a false-positive de novo mutation call in a family trio

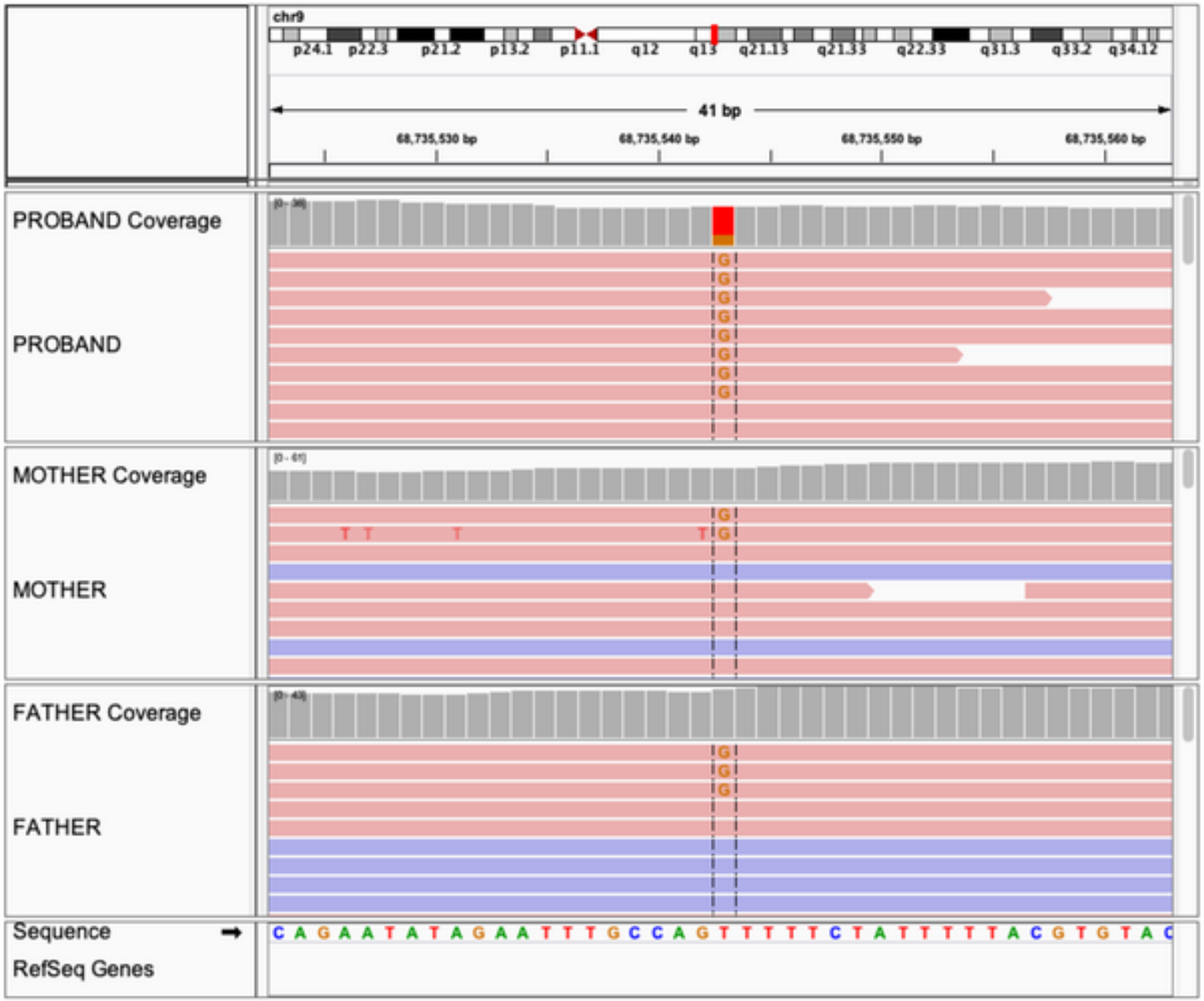
C.

C. The same position as in B, but with soft-clipped bases shown in color.

BLAT alignment of such reads reveals that the soft-clipped portion matches the other side of the deletion segment some 5.2 kb downstream.

Common artifacts in NGS alignments that gave rise to a false-positive de novo mutation call in a family trio

D.



D. False positive associated with strand bias.

All but one variant-supporting reads in the proband are on the reverse strand, whereas reference-supporting reads are equally represented on both strands.

MAGDA MIELCZAREK

VALIDATION OF NGS VARIANTS

Whether or not Sanger confirmation should be required for clinically relevant variants remains a matter of debate.

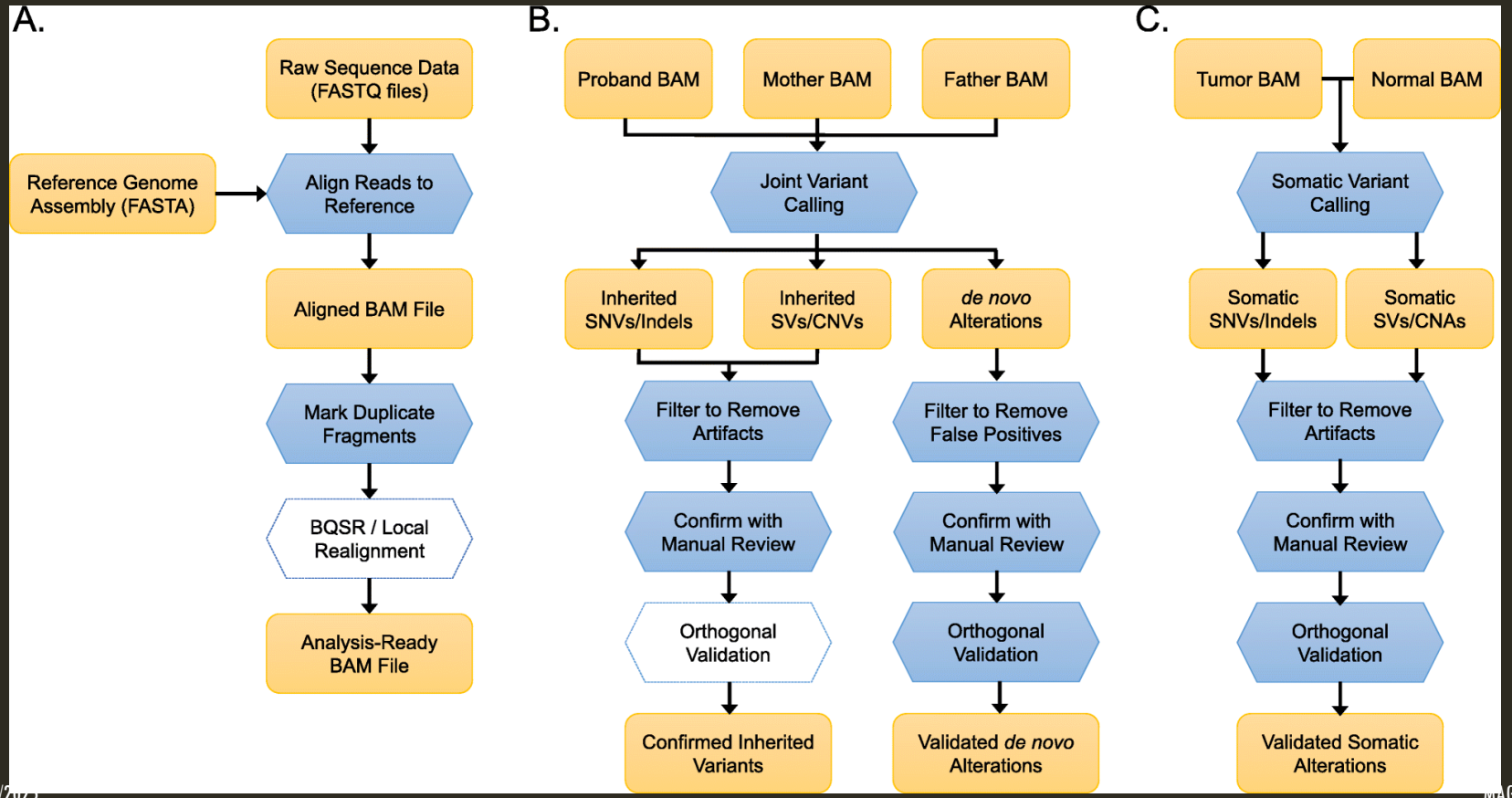
The validation rate for NGS variant calls is extremely high (99.96%) suggesting that for the vast majority of NGS variants, independent confirmation is unnecessarily redundant.

In many cases, a visual manual review of the variant may be enough to determine if it passes muster or warrants validation.

→ An interlaboratory study of more than 80,000 clinical specimens demonstrated that an approach examining fewer than ten criteria (read depth, quality score, observed variant allele sequence, repetitive sequence, etc.) can identify the subset of variants most likely to be false positives and thus requiring orthogonal validation.

A. Alignment and pre-processing of NGS data for an individual sample.

B. Variant calling in NGS trio sequencing. In this common study design, variants are called jointly (simultaneously) in a proband and both parents, which enables the phasing of variants by parent of origin. The initial variant calls are typically filtered to remove a number of recurrent artifacts associated with short-read alignment and maybe visually confirmed by manual review of the sequence alignments. Orthogonal validation may be performed to confirm the variant and its segregation within the family. *De novo* alterations should be aggressively filtered to remove both artefactual calls in the proband (false positives) and inherited variants that were under-called in a parent (false negatives). In addition to manual inspection of alignments, most de novo mutations are independently verified by orthogonal validation techniques, such as Sanger sequencing.



IDENTIFYING *DE NOVO* MUTATIONS

A key advantage of joint calling in trios is the ability to distinguish *de novo* mutations, which account for a significant proportion of positive diagnoses from clinical genetic testing.

According to recent large-scale trio sequencing studies, the human *de novo* mutation rate is approximately 1.29×10^{-8} per base pair per generation.

→ each proband likely harbors ~ 70 *de novo* mutations genome-wide against a background of ~ 4–5 million inherited variants.

In the protein-coding exome, we expect ~ 1 *de novo* mutation on a background of ~ 50,000 inherited variants. A sequence variant called in the proband is therefore far more likely to be inherited than *de novo*.

IDENTIFYING *DE NOVO* MUTATIONS

→ even with extremely high variant calling precision (99.9%), there will be 50 false-positive calls for each *de novo* mutation. Thus, candidate *de novo* mutations merit careful scrutiny.

to filtering for artifactual calls *de novo* mutations:

→should be queried against public databases. Although true *de novo* mutations can certainly occur at positions of known sequence variants, a candidate *de novo* with high frequency in the population (i.e., MAF > 0.0001) is far more likely to represent a germline variant.

→manual review in IGV should be used to exclude both artifactual calls and variants with supporting evidence in one or both parents.

CNV AND SV CALLING

Copy number variants (CNVs) are a major source of human genetic variation and have been implicated in numerous diseases (e.g. autism, intellectual disability, congenital heart disease).
NGS-based CNV detection is increasingly incorporated into clinical diagnostic testing and accounts for 3–5% of positive diagnoses.

→ identifying CNVs from targeted NGS data, such as cn.MOPS, CONTRA, CoNVEX, ExomeCNV, ExomeDepth, and XHMM. Most rely on comparisons of sequence depth between a test subject and a comparator to identify significant changes in copy number.

Not all CNV calling tools perform well in all situations, and as a rule, the sensitivity for CNV detection using targeted NGS is limited compared to genome sequencing (mhm… ☺)

CNV AND SV CALLING

Paired-end whole-genome sequencing data also enables the detection of structural variants with increasing precision. Popular tools for this application, such as DELLY, Lumpy, Manta, Pindel, and SVMerge, use two types of information to identify signatures of structural variants.

→ Read pairing information serves to identify segments of the genome in which molecularly linked read pairs map at unexpected distances or orientations.

→ Split read alignments, in which a single sequence read maps to two different regions of the genome, are also incorporated into SV calling.

CNV AND SV CALLING

SV detection with whole-genome sequencing data is still challenging → ~ 0.80–0.90 in benchmarking experiments.

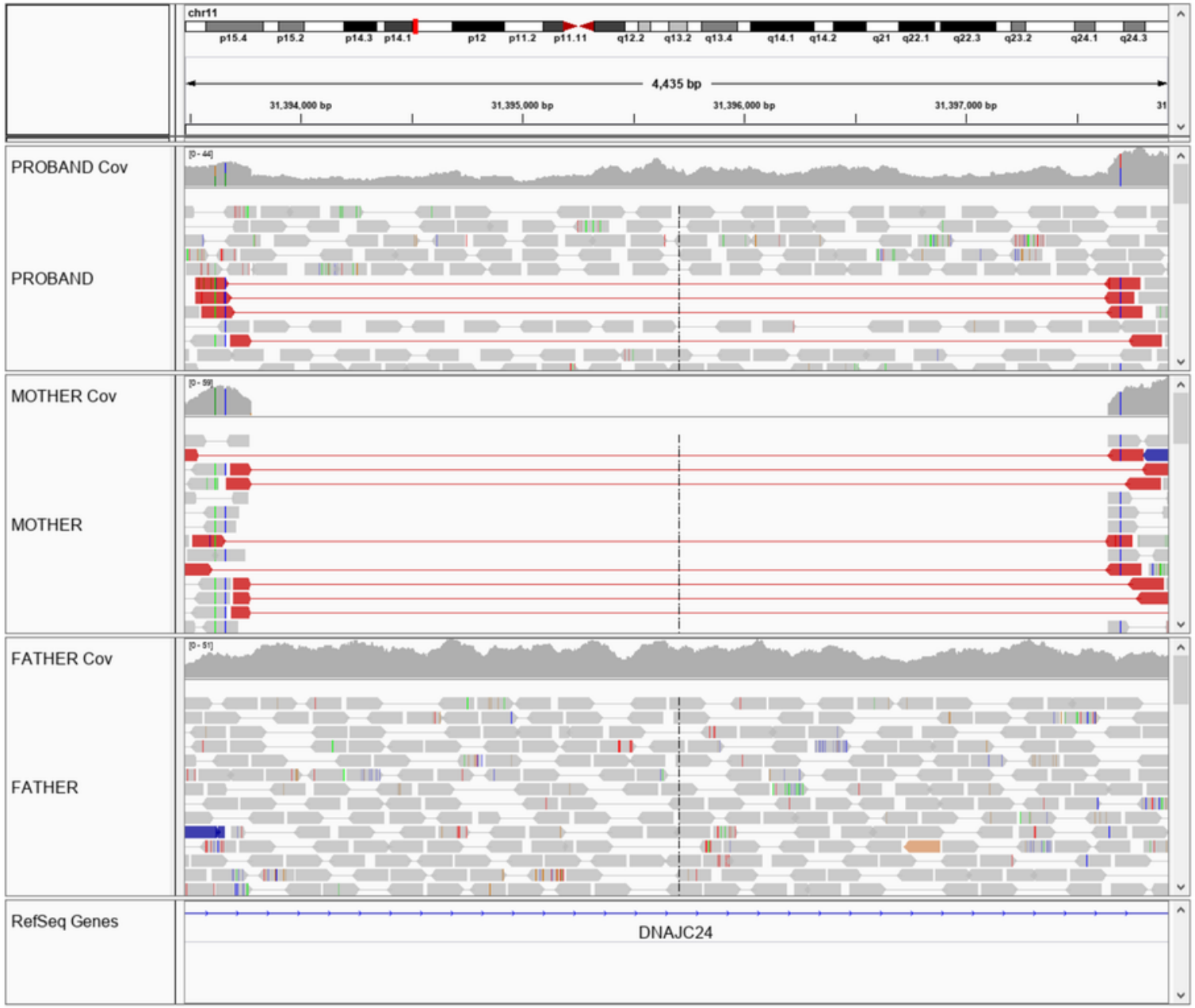
Possible reasons:

→ a large proportion of structural variation occurs in “difficult” regions of the genome, such as repetitive or tandem-duplicated sequences.

→ the relatively short length of NGS reads (~ 150 bp) and typical fragments (~ 300–500 bp) is often insufficient to resolve complex structural variants and long insertions.

Visual review of copy number and structural variants

A.



A. A homozygous ~ 4-kb del that appears heterozygous in the proband, homozygous in the mother, and absent from the father. Note the discordant read pairs suggesting a deletion (red) and visible change in read depth.

MAGDA MIELCZAREK

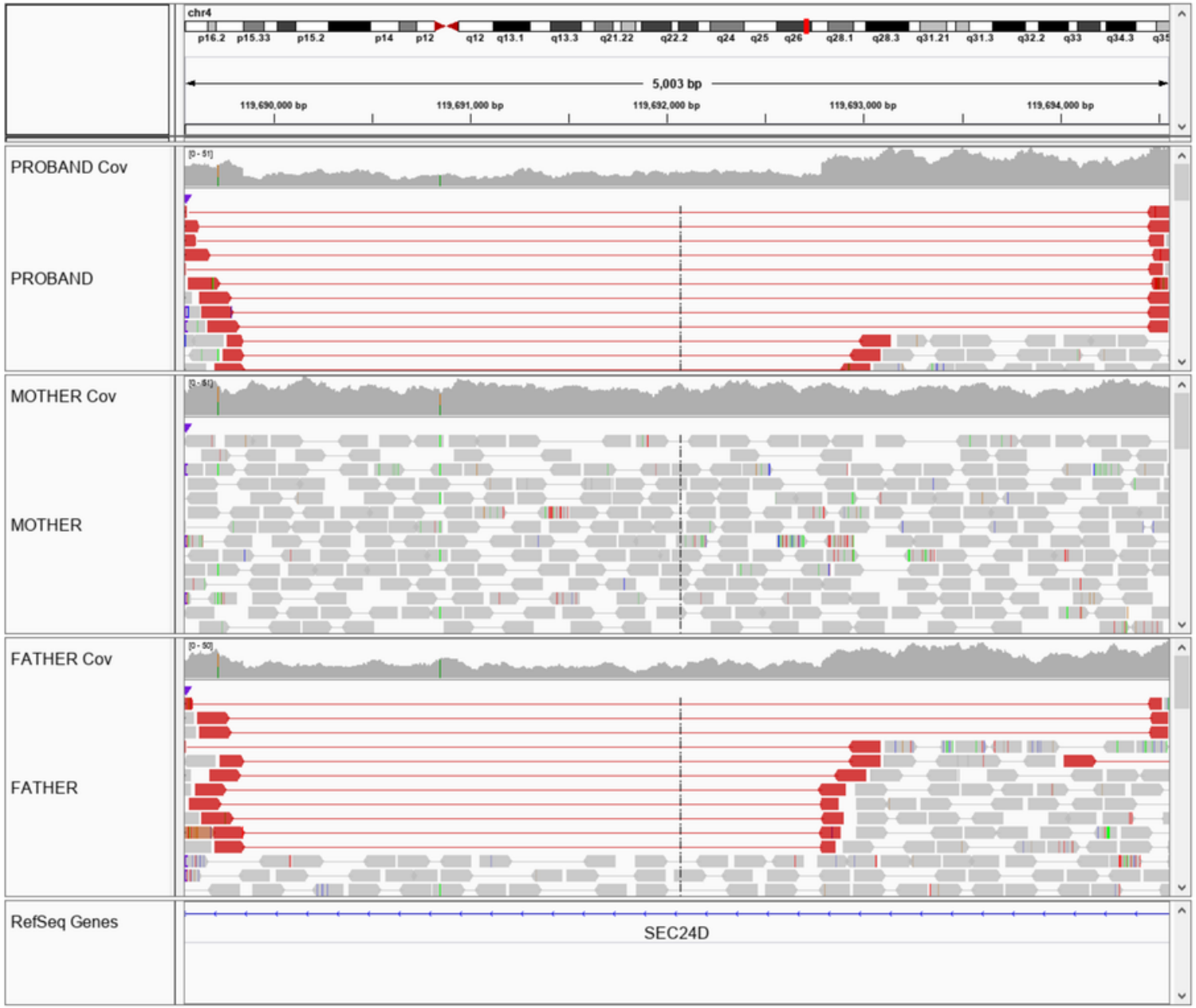
Visual review of copy number and structural variants

B.

B. Homozygous deletion inherited from two heterozygous parents.

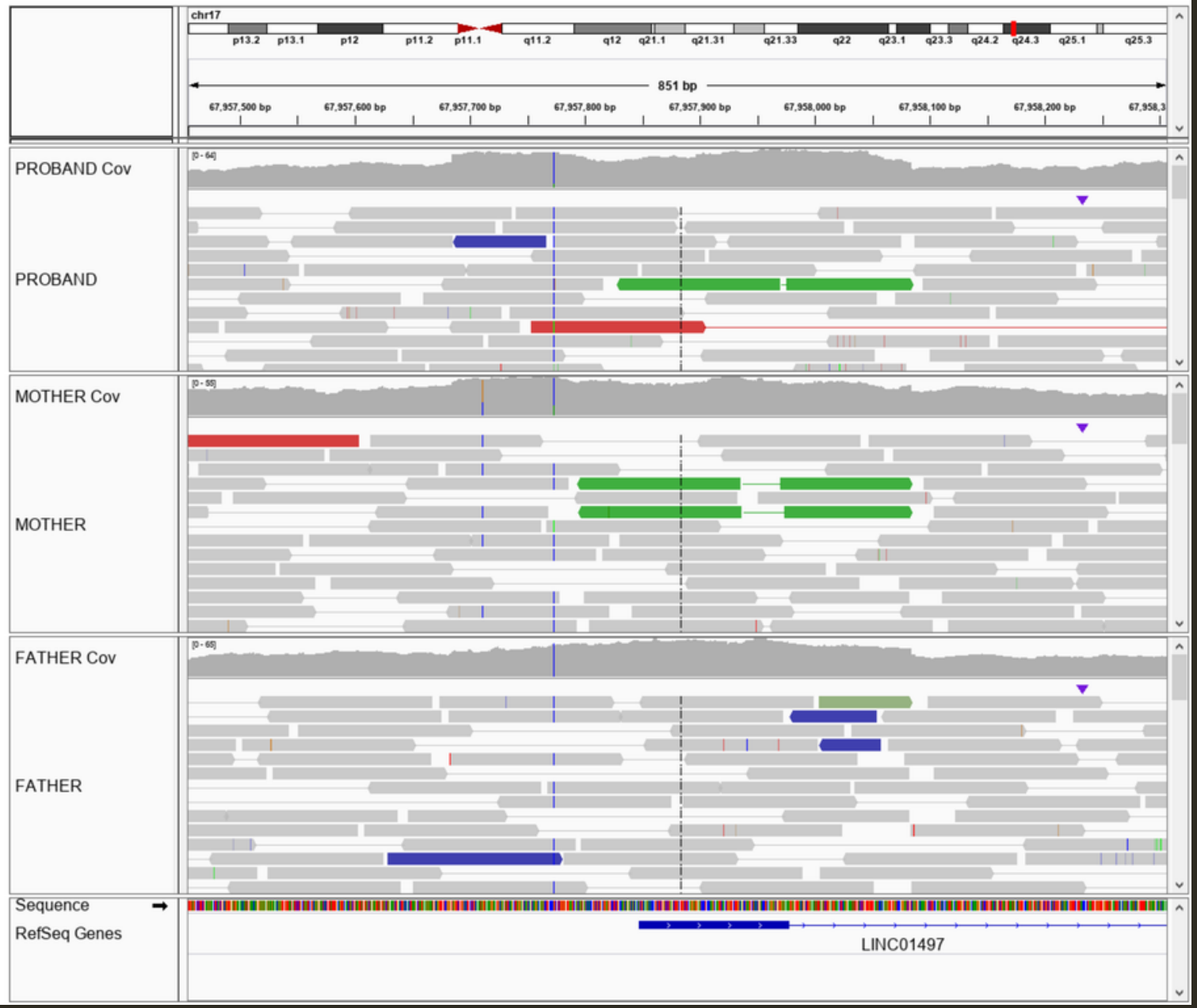
MAGDA MIELCZAREK

C.



C. A heterozygous paternally inherited deletion with ambiguous end point by paired-end mapping resolved by visual inspection of read depth.

D.



D. A maternally inherited tandem duplication. Note the increased read depth in the histogram and the discordant read pairs highlighted in green that span the original sequence and their tandem duplication

MAGDA MIELCZAREK

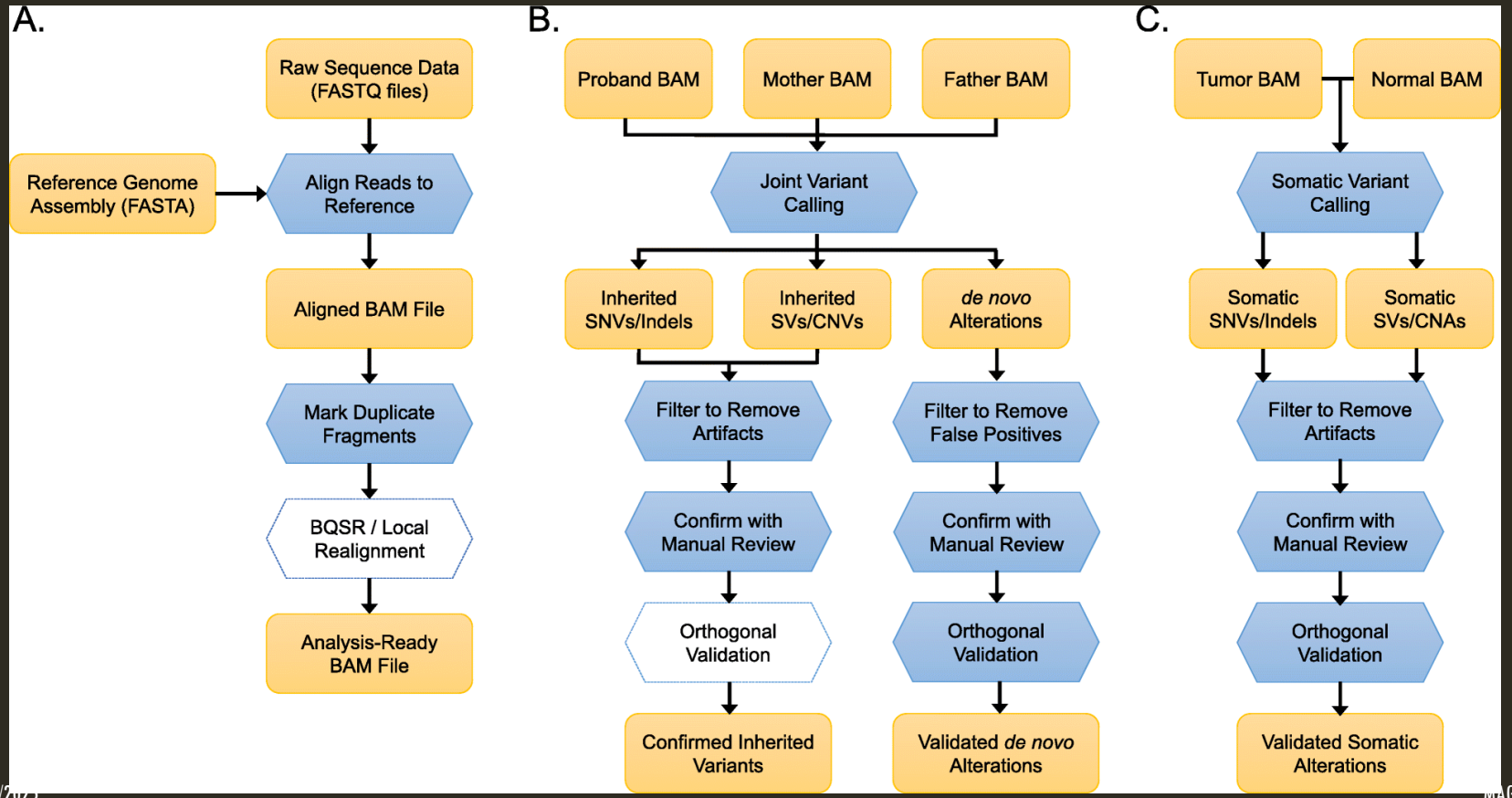
BEST PRACTICES FOR SOMATIC MUTATION CALLING

Although ~ 10% of cancer patients harbor germline predisposition variants, the main purpose of clinical tumor sequencing is often **the identification of somatic mutations**, copy number alterations, and fusions that may have clinical relevance.

A standard pipeline for this is shown in Fig. 1c. It illustrates a paired tumor-normal sequencing strategy, that is, sequencing DNA from a tumor sample and a matched control sample (e.g., blood or skin) from the same patient.

Although tumor-only sequencing has been adopted by many laboratories as a cost-effective approach to guide cancer diagnosis, prognosis, and therapy, doing so makes it difficult to distinguish true somatic mutations from constitutional variants. Thus, the emphasis of this section will be on the "best practice" of sequencing a tumor sample with a matched comparator sample.

C. Somatic variant calling in matched tumor-normal pairs. Identification of somatic alterations in tumors requires specialized variant callers which consider aligned data from the tumor and normal simultaneously. Candidate somatic variants are filtered and visually reviewed to remove common alignment artifacts as well as germline variants under-called in the normal sample. The resulting variants are typically validated by orthogonaal approaches, which may require specialized approaches for low-frequency variant.



BEST PRACTICES FOR SOMATIC MUTATION CALLING

Widely used somatic mutation callers, such as MuTect2, Strelka2, and VarScan2 consider aligned data from the tumor and normal simultaneously.

Eeach has strengths and weaknesses → two or more complementary callers may offer the best balance of sensitivity and specificity.

BEST PRACTICES FOR SOMATIC MUTATION CALLING

Detection of somatic mutations is challenging.

→ Tumor purity - the proportion of cells in a sample that are cancerous - governs the representation of somatic mutations in a sequenced sample, but pathology estimates of purity based on light microscopy may be inaccurate.

→ Formalin-fixed, paraffin-embedded (FFPE) samples, which are preferred for histopathological diagnosis, often harbor thousands of artifacts arising from chemical DNA damage.

BEST PRACTICES FOR SOMATIC MUTATION CALLING

Filtering with population databases

High-confidence somatic SNV/indel calls should be:

→ identified by multiple somatic mutation calling tools at positions with sufficient sequencing coverage (> 10× in both tumor and normal tissue).

→ supported by reads on both strands with no apparent bias in base quality, or mapping quality.

→ absent from public databases and an internal laboratory panel of normal (if available), or else present at very low frequencies (MAF < 0.001).

→ reviewed by visualization of the tumor and normal sequencing alignments with a tool such as IGV.

BENCHMARKING SOMATIC CALLING PIPELINES

Benchmarking somatic mutation callers requires a reference "truth set" of real somatic mutations.

Numerous comparisons of somatic mutation callers have been published but the findings are inconsistent. One reason for this is that the researchers conducting those studies often apply variant callers with default parameter settings or neglect to perform critical downstream filtering.

To address this issue, the DREAM ICGC-TCGA Somatic Mutation calling challenge invited teams, including several developers of somatic mutation calling tools, to benchmark their pipelines on a common dataset. The organizers employed a robust simulation framework to introduce synthetic somatic alterations (i.e., a truth set) into real WGS data for three tumors upon which each team's submissions were evaluated. The simulated datasets and truth sets from these challenges are freely available and offer a well-vetted benchmarking resource for somatic SNV, indel, and structural variant calling pipelines.

Table 2 Key components of NGS analysis and a list of exemplar tools. Most clinical sequencing pipelines will employ a single read aligner (e.g., BWA-MEM) and mark duplicates with one algorithm (e.g., Picard). However, multiple tools for collecting sequencing metrics and performing sample QC may be employed to meet the needs of the laboratory. For variant calling, it is recommended that pipelines incorporate 2–3 tools for each class of variant to maximize detection sensitivity. See the relevant section of this review for recommendations specific to each variant class

Strategy	Variant callers
Alignment and pre-processing	
Read alignment	BWA-MEM [25], Bowtie 2 [26], minimap2 [27], Novoalign
Marking duplicates	Picard tools [28], Sambamba [29], SAMBLASTER [30]
BAM file creation	Samtools [31], GATK [19]
Sequencing metrics	BEDTools [32], Picard tools [28], QualiMap 2 [33]
Sample quality control	KING [34], VerifyBamID [35]
Variant calling	
Inherited SNVs/indels	FreeBayes [36], GATK HaplotypeCaller [19], Platypus [20], Samtools/BCFtools [37]
Somatic mutations	deepSNV [38], MuSE [39], MuTect2 [40], SomaticSniper [41], Strelka2 [42], VarDict [43], VarScan2 [44]
Copy number variants	cn.MOPS [45], CONTRA [46], CoNVEX [47], ExomeCNV [48], ExomeDepth [49], XHMM [50]
Structural variants	DELLY [51], Lumpy [52], Manta [53], Pindel [54], SVMerge [55]
Gene fusions (RNA-seq)	fusionCatcher [56], fusionMap [57], mapSplice [58], SOAPfuse [59], STAR-Fusion [60], TopHat-Fusion [61]
Variant review/storage	
Visualization and review	Artemis [62], Integrative Genomics Viewer [63]
VCF/BCF file manipulation	BCFtools [37]

BAM binary alignment/map, *SNV* single nucleotide variant, *VCF* variant call format, *BCF* binary variant call format

CONCLUSIONS AND FUTURE DIRECTIONS

Variant calling in NGS data, much like NGS technologies themselves, has evolved considerably over the past decade and remains an active area of research.

Robust pipelines for NGS analysis include steps for:
- optimized alignment and pre-processing
- variant calling, filtering of false positive
- visual manual review.

While some of these procedures, such as read alignment and SNV/indel detection, can be suitably performed with a single software package, others, such as CNV/SV calling and somatic mutation detection, benefit from incorporating multiple independent tools.

Benchmarking resources for both germline and somatic variants provide an opportunity to evaluate and optimize the performance of variant calling.

CONCLUSIONS AND FUTURE DIRECTIONS

Although some classes of variants—such as de novo mutations in germline studies and low-frequency somatic mutations in cancer patients—likely require validation on an orthogonal platform, the burden of additional confirmatory testing is likely to decrease as technologies continue to improve.

Long-read sequencing technologies may ultimately be required to accurately call large and/or complex structural variants.