

Comparison of Open-Source SMN1/SMN2 Variant Calling Tools (Short-Read WGS, No-Amp)

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
SMNCopyNumberCaller (Illumina) – GitHub – Chen <i>et al.</i> 2020 (Genet Med) ¹	SMN1/SMN2-specific (plus SMN2Δ7–8)	Copy number (gene-level; detects 0,1,2+ copies; also exon7–8 deletion as separate SMN2Δ7–8 copy; one SNP for silent carrier)	Germline	99.8% of SMN1 and 99.7% of SMN2 copy-number calls concordant with MLPA/ddPCR ¹ (100% sensitivity for SMA cases, ~97–100% for carriers ²). Precision 100% for both SMA and carriers ² .	Lightweight; Python-depth method over a ~100 kb region. Used on 12,747 genomes with batch processing ³ . Runtime on 30× WGS is on the order of seconds per sample (minutes for large cohorts); minimal memory/compute cost (depth+allele counting only).	hg19, GRCh38 (1000 Genomes, hg38) [19/20] in too

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
SMAca (Babelomics) – GitHub – López-López <i>et al.</i> 2020 (Hum. Mut.) ⁷ ⁸	SMN1/SMN2-specific (silent-carrier detection)	Copy number (SMN1 absolute copies; also uses SNVs specific to gene duplication to identify 2+0 “silent” carriers)	Germline	Validated on 326 WGS samples (Navarra NAGEN cohort) ⁹ – all SMA carriers (1×SMN1) and non-carriers correctly distinguished (100% carrier detection reported) ¹⁰ . High specificity (~99%) for distinguishing 1-copy vs 2-copy states ¹⁰ .	Highly efficient multi-sample depth ratio method. Designed for batch processing; e.g. 326 genomes analyzed in 3 minutes using 24 threads ¹¹ (i.e. <1 min per 100 genomes). Low memory footprint; suitable for clinical pipeline integration. Cost negligible relative to WGS alignment.	hg19 and h GRCh (added v1.2.1). Also v target panel if SMN cover

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
Parascopy (Bansal Lab) – GitHub – Prodanov & Bansal 2022 (Nat. Comm.) ¹⁴ ¹⁵	Multi-gene (segmental duplicates) – e.g. SMN1/2 and ~150 other duplicated genes	Copy number + paralog-specific SNVs (estimates total copy count per locus <i>and</i> splits by paralog via PSVs; can output phased gene haplotypes)	Germline (multi-sample analysis)	Perfect concordance with gold-standard for SMN1/2: in 1000 Genomes data (1109 samples with MLPA), Parascopy's SMN1 & SMN2 counts matched MLPA in all cases ¹⁵ . Detected the common SMN2 exon7–8 deletion with 100% sensitivity and 99.9% specificity (vs. MLPA) ¹⁶ . Across 9 tested gene families, Parascopy showed highest accuracy vs other tools ¹⁶ .	Moderate resource usage: Analyzing 167 duplicated loci per genome took ~ 16 min on 30× WGS (16 threads) and <5 GB RAM ¹⁷ . Per single locus (like SMN) runtime is only a few seconds. Supports joint calling on cohorts (leverages multi-sample HMM/EM) – improves sensitivity for shared variants ¹⁸ . ¹⁹ .	Precor homo tables hg19 hg38 provi ²¹ . for cu refer genom custo homo tables

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
QuickK-mer2 (Kidd Lab) – GitHub – Shen & Kidd 2020 (Genes) ²³	Genome-wide CNV (alignment-free k-mer method; paralog-sensitive)	Copy number (estimates total copy number of specified regions or genome-wide bins; can report paralog-specific CN if unique k-mers exist)	Germline (population-scale focus)	Achieved high accuracy for total SMN1/SMN2 <i>copy counts</i> (concordant with known MLPA copy numbers for exons 1–6) ¹⁶ . Limitation: lower accuracy for the SMN2 exon7–8 deletion: in one evaluation, sensitivity was ~71% and specificity 38% for detecting the exon7–8 deletion copy (i.e. distinguishing SMN2Δ7–8) ²⁴ . (General observation: excels at overall copy number, struggles with <i>partial</i> gene deletions or very high-identity paralogs.)	Fast, but memory-intensive: mapping-free k-mer counting on a 30× genome took ~ 36 min and 40 GB RAM (16 threads) ²⁵ . Optimized for large cohorts – was used to process 2,457 genomes in the original study ²³ . Single-gene mode possible by restricting to k-mers of that locus (reduces resource usage).	Any r... (built... GRCh... origin... meth... handl... hg19... – cau... each... must... repre... distin... refere... other... speci... handl... need... ²⁷).

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
GeneToCN (Univ. Tartu) – GitHub – Pajuste & Remm 2023 (Sci. Rep.) 28 29	Single-gene targeted (user provides gene region; alignment-free k-mer counting)	Copy number (outputs absolute gene copy estimate per sample; does <i>not</i> explicitly differentiate nearly identical paralogs in reference – focuses on a gene’s combined copy count)	Germline	Validated on genes with known CNVs: e.g. AMY1 copy numbers had R = 0.99 correlation with ddPCR 28 . Tested on SMN1/2 locus : predicted copy-number distribution in 500 biobank samples closely matched published population data 29 (e.g. modal 2+2 copies, rare 0-copy or 1-copy outliers consistent with SMA carrier rates). Accuracy for SMN1 vs SMN2 individually is not directly reported (method likely reports total SMN copies due to SMN1/2’s 99.9% identity).	Lightweight: Counts a small set of selected k-mers from raw FASTQ – avoids full alignment. Runtime is on the order of a few minutes per 30× genome (I/O bound). Minimal CPU/RAM (k-mer counting and simple ratio calculation). Easily scaled to thousands of samples.	Flexible can g k-mer for an refer build provi exam mer s hg38 sever 30). Demo on Ill PacBi ONT r (simil result tech)

Tool (Code) & Publication	Scope (Genes)	Variant Types	Germline or Somatic	Performance Metrics	Runtime / Memory / Cost	Support / Build
ExpansionHunter (Illumina) – GitHub – Dolzhenko <i>et al.</i> 2019 (Bioinformatics) ³²	Tandem repeats (STRs) – <i>includes an SMN1 CN check module</i>	Sequence variant detection via a graph alignment of SMN1/2 <i>exon 7 SNP</i> . <i>Not a full CN caller</i> : detects presence/absence of SMN1-specific allele (c. 840C) to infer SMN1 copy loss.	Germline	For SMA, functions as a binary test : if no reads with the SMN1 allele (C at exon7) are detected, the sample is called “affected” (0× SMN1) ³³ ³⁴ . In Illumina’s evaluation this approach had 100% sensitivity for identifying SMA cases (0-copy SMN1) in validation sets, with no false positives reported in normals (carriers still have the C allele) ³⁵ ³⁶ . However, it does not quantify 1 vs 2 copies – carriers and 2-copy individuals are both “unaffected” (presence of C).	Incorporated into DRAGEN WGS pipeline (adds negligible overhead). Performs local graph realignment on the SMN region; runtime ~seconds per sample and trivial CPU/memory impact when combined with STR calling ³³ ³⁴ . No extra cost in practice (often run alongside STR analysis).	GRCh and G uses i provid refer graph repre SMN1 each ³⁸ .

Detailed Notes on Each Tool:

SMNCopyNumberCaller (Illumina)

This Python tool from Illumina estimates the **copy counts of SMN1, SMN2, and the exon7–8-deleted SMN2** (often termed SMN2Δ7–8) from a standard WGS BAM/CRAM ⁴¹. It exploits read-depth ratios and **eight paralog-distinguishing SNP/indel sites** across SMN1/2 ⁴² to assign reads to SMN1 vs SMN2. The method achieves clinical-grade accuracy: in a validation on >12k genomes, **99.8% of SMN1 and 99.7% of SMN2 copy-number calls** agreed with orthogonal tests ¹. All 95 SMA cases (0×SMN1) were correctly identified (100% sensitivity), and carrier detection (1×SMN1) reached ~97.8% sensitivity with 100% precision ². The software can flag “silent carriers” (2 SMN1 copies on one chromosome and 0 on the other) by

detecting the gene-conversion SNP (g.27134T>G) copy number ⁴³. **Genome build:** Supports hg19/GRCh37 and GRCh38 (the code has presets for “--genome 19/37/38”) ⁴. **Outside SMN:** No – it’s hard-coded for the SMN locus. **Use cases:** Designed for **germline testing** – it outputs clear carrier/affected status flags ⁴⁴. This tool has been used in population SMA carrier studies and is noted as enabling SMA screening as part of routine WGS ⁴⁵. It’s distributed under a PolyForm license (free for non-commercial use) and is installable via Bioconda. Runtime is very fast: processing a few hundred BAMs as a batch takes only minutes. The authors provide a visualization module for output plots ⁴⁶ to aid interpretation.

SMACa (SMA Carrier Analysis)

SMACa is an open-source (GPL3) tool focusing on identifying SMA carriers from NGS data ⁷ ⁸. It computes the proportion of SMN1 vs SMN2 reads at known distinguishing sites and normalizes gene depth against a panel of control genes to estimate absolute SMN1 copy number per sample ⁴⁷ ⁴⁸. Uniquely, SMACa can detect “**2+0**” **silent carriers** – individuals with two SMN1 copies on one allele and none on the other – by recognizing the absence of SMN1-unique variants (e.g. an AT deletion at g.27706_27707) despite a total SMN1 count of 2 ⁴⁷ ⁴⁹. The tool is intended for **germline WGS/WES** data and works best on cohorts (leveraging population frequency of SMN1 vs SMN2 read ratios) ⁵⁰. In a validation on 326 genomes, SMACa correctly identified all carriers and SMA cases, matching known copy numbers ¹⁰. Its processing speed is a highlight: by multithreading, hundreds of samples are done in a couple of minutes ¹¹. **Genome build:** both GRCh37 and GRCh38 are supported (added in v1.2.1) ¹². **Beyond SMN:** No, it’s specialized for SMN1/2. The authors published the method in *Human Mutation* 2020, positioning SMACa as a pipeline-friendly solution for carrier screening ⁸. It has been used in at least one population study and could be integrated in clinical labs, although as of 2020 it was a research prototype.

Parascopy

Parascopy is a general-purpose tool for **paralog-specific copy-number analysis** of duplicated genes ⁵¹. It addresses the challenge of multi-copy genes (like SMN1/SMN2) by joint analysis of read depth and allelic ratios at paralogous sequence variants (PSVs) across multiple samples ⁵² ⁵³. For SMN, Parascopy first infers the **total SMN1+SMN2 copy profile** along the locus using a hidden Markov model (across the gene and any deletions) ⁵⁴ ⁵⁵, then uses EM to partition that total into SMN1 vs SMN2 copies per individual based on SNP markers that differ between SMN1 and SMN2 ⁵⁶ ⁵⁷. This achieved **exceptional accuracy** – in a 1,109-sample test (1kGP data with MLPA results), Parascopy’s SMN1/2 calls were **100% concordant** with MLPA for both full-length gene copies and the exon7–8 deletion allele ¹⁵. Notably, Parascopy was the only tool to perfectly identify the common SMN2 exon7–8 deletion in that benchmark (sensitivity 1.00 vs 0.709 for Quick-mer2) ²⁴. It also showed identical SMN copy-number results to SMNCopyNumberCaller on >850 samples in one comparison ⁵⁸. **Scope:** Parascopy isn’t limited to SMN – it comes with homology data for ~167 duplicated gene loci (covering clinically important genes like *PMS2*, *GBA*, *CCL3L1*, etc.). **Genome builds:** hg38 and hg19 are supported out-of-the-box ²⁰ ²¹. **Variants:** The tool can output paralog-resolved SNP calls (“PSVs”) in addition to copy numbers, though for final variant calling one would typically remap reads as needed. **Use cases:** Currently germline research – for example, studying copy-variable gene associations. It’s relatively new (2022) and not yet a standard clinical test, but its high accuracy suggests it could be translated to clinical pipelines for difficult genes like SMN, HLA, etc. It’s freely available (Bioconda/Python) ²⁰, and can leverage cohorts to boost detection of subtle CNVs (by sharing information across samples) ¹⁹. The trade-off is some complexity in setup (precomputed homology tables) and higher runtime than single-locus tools if you run the whole genome set.

Quick-mer2

Quick-mer2 is an alignment-free CNV caller that uses **unique k-mer frequencies** to estimate copy number for genomic regions ²³. For each target gene or region, it identifies k-mers that are unique to each copy (paralog) or to the entire locus and counts them in WGS reads, thereby bypassing mapping ambiguity. Shen and Kidd (2020) applied Quick-mer2 to estimate copy numbers for all duplicated genes in 2,457 genomes ²³, demonstrating scalability. In SMN1/2, Quick-mer2 can successfully gauge the total copy number of SMN genes – e.g. distinguishing 0,1,2,3 copies of SMN1 – with high concordance to reference methods ¹⁶. However, because SMN1 and SMN2 share >99.9% sequence, Quick-mer2 had difficulty with the subtle exon7–8 deletion: it often failed to identify whether a sample's SMN2 copies included a deleted form ²⁴. In one analysis, its sensitivity was only ~71% for detecting that deletion allele (and specificity ~38%, meaning many false positives) ²⁴. In practical terms, Quick-mer2 might correctly total the *combined* SMN copies but could mislabel a “2+0” silent carrier as normal (2 copies) or misidentify deletion carriers. **Genome builds:** the method can be applied to any assembly – the user must generate k-mer sets from the reference. But if a gene has multiple nearly-identical copies in the reference (as SMN1/SMN2 do in GRCh38), Quick-mer2's approach of assigning k-mers to a specific paralog is challenged ²⁶ ²⁷. The authors note this limitation, and indeed GeneToCN was developed to address it. **Performance:** It's quite fast for a genome-wide scan (under 1–2 hours for a full 30× genome) ⁵⁹, thanks to not aligning reads. Memory usage, on the other hand, is high (tens of GB) due to large k-mer tables ²⁵. Running it for just the SMN locus would be far lighter. **Usage:** Quick-mer2 is primarily a research tool for large-cohort analyses of CNVs. It's cited in studies of segmental duplications and has been integrated into pipelines like CNPI for post-processing k-mer outputs ⁶⁰. For SMA specifically, other tools outperform it on accuracy, so Quick-mer2 hasn't seen clinical adoption in that context.

GeneToCN

GeneToCN (2023) takes an alignment-free approach focused on **user-specified genes** rather than scanning the whole genome ⁶¹. It selects a robust set of k-mers unique to the gene of interest (and its local flanking region for depth normalization) ⁶² ⁶³. By counting these k-mers in the raw reads, it infers the gene's copy number (essentially comparing the median k-mer depth in the gene to that in two-copy flanking regions) ⁶⁴. One advantage is the ability to handle high copy numbers accurately – the authors validated *AMY1* and *AMY2* copy counts up to >10, achieving $R \approx 0.99$ correlation with droplet PCR results ²⁸. For the SMN locus, they reported that GeneToCN's output for a cohort of 500 Estonians mirrored known SMN1/SMN2 distribution frequencies ²⁹. This implies it correctly identified the few 0× or 1× SMN1 individuals (SMA cases or carriers) and the varying SMN2 counts, although the tool likely reports a combined SMN copy number. (If one wanted separate SMN1 vs SMN2, one might have to supply k-mers specific to each – not explicitly detailed in the paper). **Performance:** GeneToCN is efficient: instead of counting *all* k-mers, it only needs to track a few thousand per gene, making it CPU- and memory-light. It's implemented in Python with support scripts (and possibly uses Jellyfish for counting). Running on different sequencing technologies (Illumina vs PacBio vs ONT) yielded similar CN estimates ³¹, indicating robustness to sequencing errors by virtue of using normalized counts. **Use cases:** It's a research tool for quick gene CN screening. In a clinical lab, one could imagine using GeneToCN for a fast check of SMN1/2 copy number from FASTQ before deeper analysis – especially since it doesn't require alignment. However, because it doesn't inherently resolve nearly identical paralogs, a lab would likely stick to specialized tools for SMN. GeneToCN's code and example k-mer databases are on GitHub ³⁰.

ExpansionHunter (with SMA Module)

ExpansionHunter is primarily known for **short tandem repeat (STR) expansion** detection, but Illumina extended it to tackle SMA due to the single nucleotide difference in exon 7 of SMN1 vs SMN2. The tool uses a **sequence-graph representation** of the SMN region ³⁷: essentially, it models the exon7 sequence as a branching where one path has the SMN1 allele (C nucleotide) and the other has the SMN2 allele (T nucleotide). It then realigns WGS reads to this graph to determine if the SMN1-specific allele is present ³³. If **no reads align to the SMN1 branch**, it concludes the sample has 0 copies of SMN1 (i.e. likely SMA affected) ³³ ⁶⁵. If any SMN1 reads are detected, the sample is labeled unaffected (at least one SMN1 copy). This simple presence/absence test cannot distinguish one vs two SMN1 copies – carriers and normals both show the C allele – but it robustly flags true SMA patients. In fact, when comparing to MLPA, ExpansionHunter-based SMA calling was 100% sensitive and specific for detecting 0xSMN1 cases in a screening study (no missed SMA and no false alarms among non-SMA) ³⁵ ⁶⁶. ExpansionHunter is **open-source (Apache-2.0)** and often run as part of Illumina's DRAGEN pipeline. For example, a recent study used it in WGS to discover an infant with SMA who had been missed by other methods ³⁹ ⁶⁷. **Performance considerations:** It requires decent coverage (Illumina recommends PCR-free 30x WGS ⁶⁸ so as not to miss low allele fractions). Because it's effectively scanning for one SNP, it runs in seconds and uses negligible memory. **Genome builds:** Illumina provides example "repeat catalog" entries for SMN1 for hg19 and hg38, which configure the graph alignment coordinates ⁶⁹. **Clinical use:** While not a comprehensive SMN genotyper, this method has been adopted in some **newborn SMA screening** pilots to quickly identify infants with 0 SMN1 copies (who then get confirmatory testing) ⁷⁰ ⁷¹. It's also a component in Illumina's accredited DRAGEN pipeline (v4.2+) for rare disease WGS ⁷². Users should note that ExpansionHunter's SMN output comes in a custom VCF with genotype ("affected" vs "unaffected") rather than a copy number report ⁷³.

Other Tools: In literature, general CNV callers like **CNVnator** (read-depth) and structural variant callers (Manta, DELLY, etc.) have been applied to the SMN region, but with limited success. For example, CNVnator could detect the presence of a deletion overlapping SMN1, but in evaluations its accuracy was much lower (~68% for SMN1 exon7-8 deletion) ²⁴. Such tools also don't explicitly separate SMN1 vs SMN2 – they typically report a total copy number or a single "deletion" event. **GATK CNV** and similar exome CNV callers face the same issue of paralog ambiguity. Therefore, the tools above, which leverage paralog-specific signals, are preferred for SMN.

Finally, for **long-read WGS**, it's worth noting that methods like **Paraphase** (PacBio's tool) exist, which assemble and phase full SMN1 and SMN2 sequences ⁷⁴ and can genotype SMN variants with near perfection ⁷⁵ ⁷⁶. However, Paraphase requires HiFi long reads and is beyond the scope of short-read pipelines. In short-read contexts, the open-source tools compared here represent the state of the art as of 2025, enabling robust SMN1/SMN2 analysis without needing any targeted enrichment.

Sources: Chen *et al.*, *Genetics in Medicine* (2020) ⁷⁷ ²; López-López *et al.*, *Hum. Mutation* (2020) ⁷ ⁸; Prodanov *et al.*, *Nat. Commun.* 13:3221 (2022) ¹⁶ ²⁴; Shen & Kidd, *Genes* 11:141 (2020) ²³; Pajuste *et al.*, *Sci. Rep.* 13:17765 (2023) ⁷⁸; Dolzhenko *et al.*, *Bioinformatics* 35:4754 (2019) ³²; Illumina DRAGEN v3.10/v4.2 Documentation ³³ ⁷⁹; Scientific reports and evaluations ⁸⁰ ¹.

1 Comprehensive genome analysis and variant detection at scale using DRAGEN | Nature Biotechnology

https://www.nature.com/articles/s41587-024-02382-1?error=cookies_not_supported&code=f5cf8ebe-f6b3-41d3-926d-ae32a1bfae03

2 3 5 6 42 45 77 Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data | Genetics in Medicine

https://www.nature.com/articles/s41436-020-0754-0?error=cookies_not_supported&code=6087a822-739d-4992-818f-1fa0c727c26d

4 41 43 44 46 GitHub - Illumina/SMNCopyNumberCaller: A copy number caller for SMN1 and SMN2 to enable SMA diagnosis and carrier screening with WGS

<https://github.com/Illumina/SMNCopyNumberCaller>

7 8 9 11 SMN1 copy-number and sequence variant analysis from next-generation sequencing data - PubMed

<https://pubmed.ncbi.nlm.nih.gov/33058415/>

10 SMN1 copy number, intragenic mutation and 2 + 0 carrier analysis ...

<https://pubmed.ncbi.nlm.nih.gov/35659528/>

12 13 47 48 49 50 GitHub - babelomics/SMACa: SMN1 copy-number and sequence variant analysis from next generation sequencing data

<https://github.com/babelomics/SMACa>

14 15 16 17 18 19 20 21 22 23 24 25 54 55 56 57 58 Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing | Nature Communications

https://www.nature.com/articles/s41467-022-30930-3?error=cookies_not_supported&code=c2d21560-c369-4b11-b91e-adb6aceda34f

26 27 28 29 30 31 61 62 63 64 78 GeneToCN: an alignment-free method for gene copy number estimation directly from next-generation sequencing reads | Scientific Reports

https://www.nature.com/articles/s41598-023-44636-z?error=cookies_not_supported&code=4b611198-578d-4178-9e9e-0ab8a7793a50

32 REViewer: haplotype-resolved visualization of read alignments in ...

<https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-022-01085-z>

33 34 37 38 65 68 69 73 SMA Calling With ExpansionHunter

https://support-docs.illumina.com/SW/DRAGEN_v310/Content/SW/DRAGEN/SMACalling_ExpansionHunter.htm

35 36 66 70 71 80 Next generation sequencing is a highly reliable method to analyze exon 7 deletion of survival motor neuron 1 (SMN1) gene | Scientific Reports

https://www.nature.com/articles/s41598-021-04325-1?error=cookies_not_supported&code=0819ea18-e409-4938-9848-15ed9fa29b43

39 67 Spinal muscular atrophy and repeat expansion disorders screening ...

<https://www.sciencedirect.com/science/article/pii/S2949774424006514>

40 Spinal muscular atrophy diagnosis and carrier screening from ...

<https://www.nature.com/articles/s41436-020-0754-0>

51 52 53 Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing - PubMed

<https://pubmed.ncbi.nlm.nih.gov/35680869/>

59 CNPI: Rapid Analyses of Human Copy Number Data - ScienceDirect

<https://www.sciencedirect.com/science/article/pii/S0022283625003791>

60 KiddLab/fastCN - GitHub

<https://github.com/KiddLab/fastCN>

72 79 Spinal Muscular Atrophy Calling

https://support-docs.illumina.com/SW/dragen_v42/Content/SW/SMACalling.htm

74 Introducing Paraphase by PacBio — an informatics method for ...

<https://www.pacb.com/blog/introducing-paraphase-by-pacbio-an-informatics-method-for-uncovering-secrets-of-spinal-muscular-atrophy/>

75 PacificBiosciences/paraphase: HiFi-based caller for highly ... - GitHub

<https://github.com/PacificBiosciences/paraphase>

76 PacBio HiFi Sequencing for Spinal Muscular Atrophy Analysis

<https://www.cd-genomics.com/longseq/resource-spinal-muscular-atrophy-analysis.html>