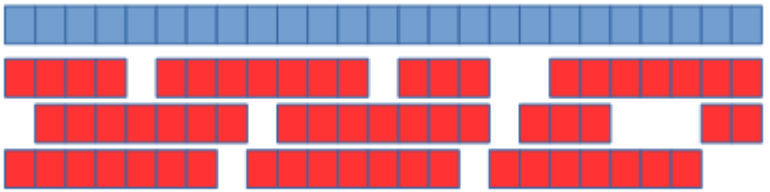
Variant calling using local reference-helped assemblies

Martin Dráb

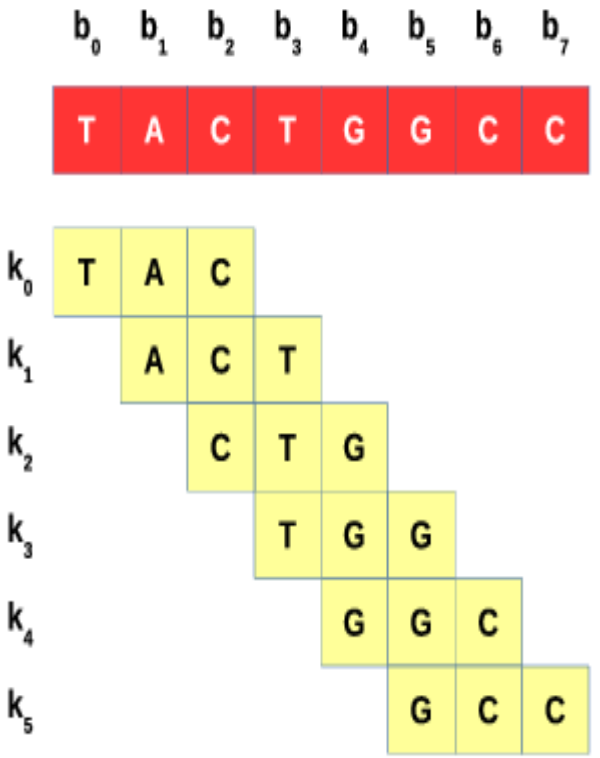
The Task

- **Problem**
  - We are unable of reading the whole DNA at once
  - The usual approach is to chop it into a set of short sequences (reds) that we are able to read
  - The short sequences are then assembled into the original one
  - The reference sequence may help with this process
  - Read length: 150 bp, genome length (3 Gbp, 40 Mbp for the tested region)
- **Input**
  - A set of reads (short strings) extracted from the target DNA and covering the regions of interest
  - The reads usually contain sequencing errors
  - Multiple reads may cover the smae part of the sequence
  - A reference sequence
- **Output**
  - A set of differences (SNPs, indels) from the reference, so-called variants



Major Approaches

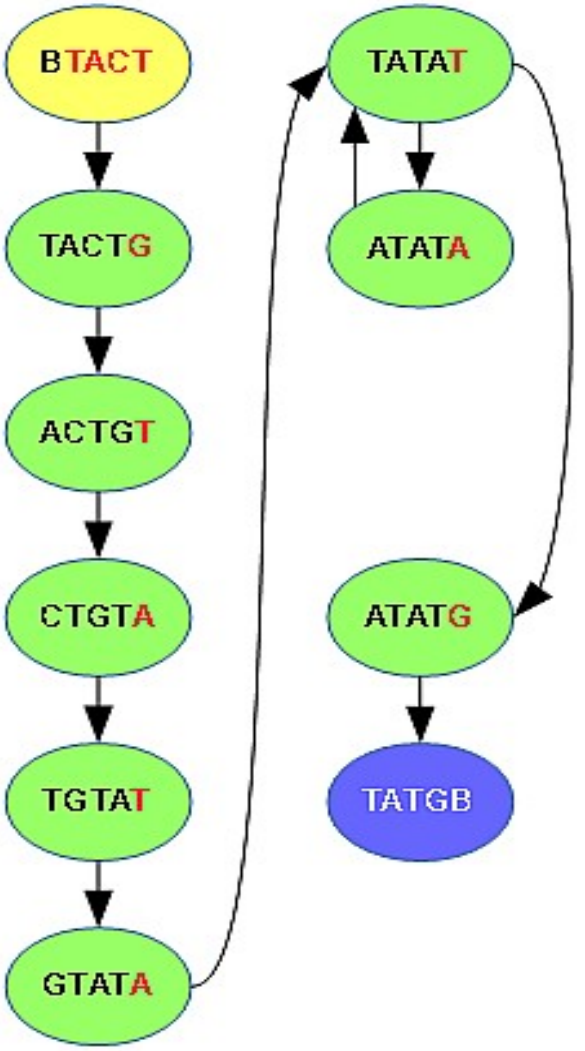
- **De Bruijn graphs (DBG)**
  - Decompose each read into k-mers
  - Build a de Bruijn graph
    - (k-mers = nodes, edges = k-mer order)
  - Optimize its structure
  - Extract the sequence
  - HaploCall (used in GATK)
- **Overlap-layout-consensus (OLC)**
  - Compute overlaps between reads
  - Build an overlap graph (nodes = reads, edges = overlaps)
  - Optimize its structure
  - Extract the seuqnce(s)
  - Fermi



Our Algorithm

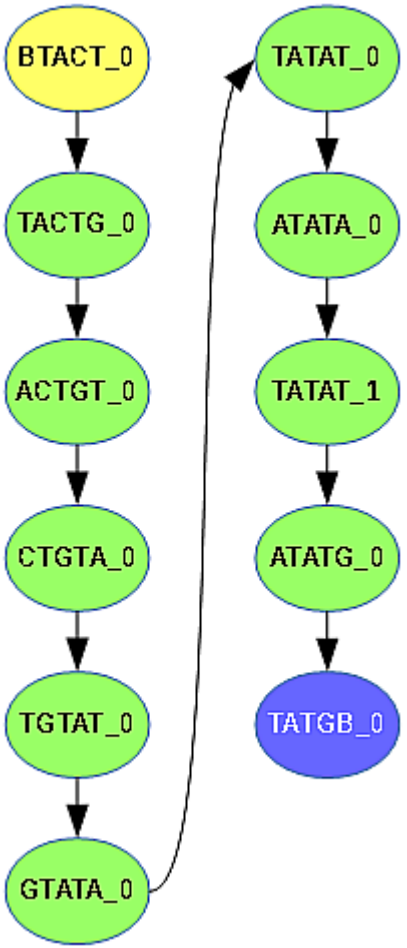
TATCGTATATATG

- Divide into regions
  - Uniformly
  - Same size (2000 bases)
  - Processed independently and in parallel
- Transform the reference into a DBG,
- Add reads assigned into the region
- Optimize graph structure
- Detect variants
- Filter them

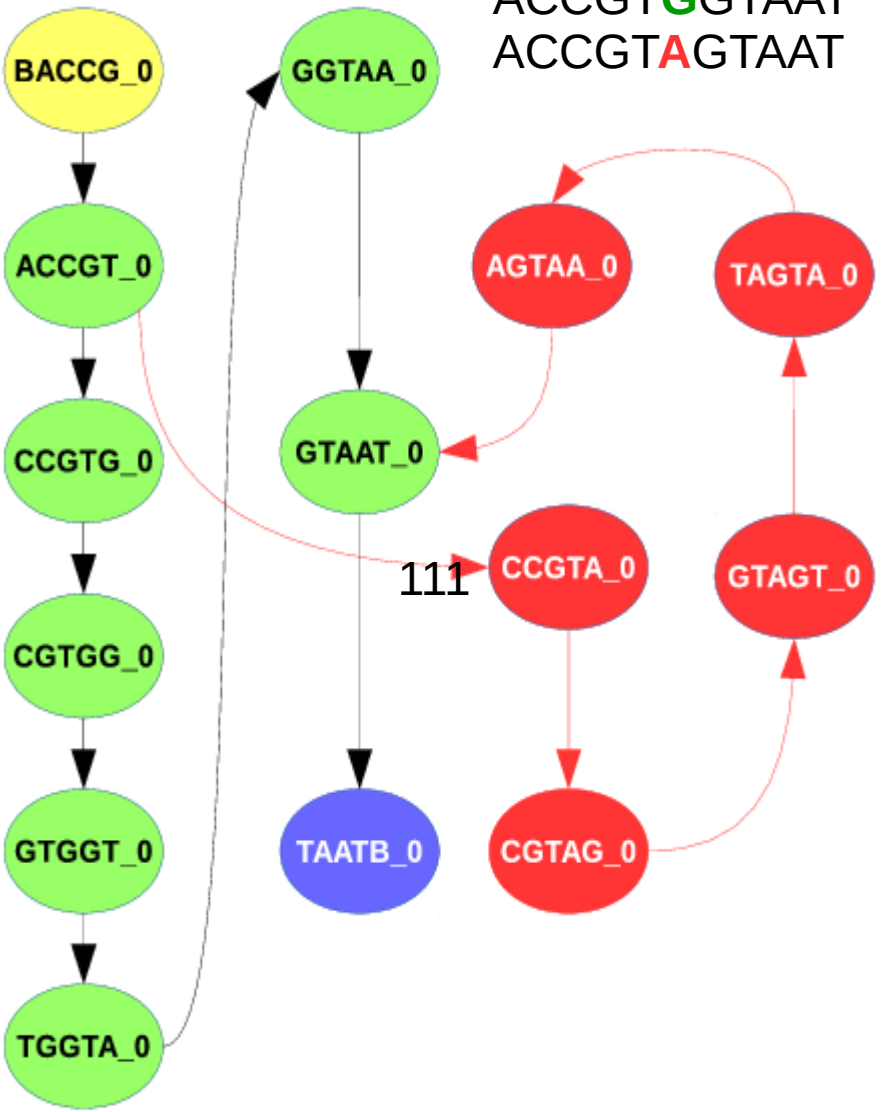


Problems with the k-mer size

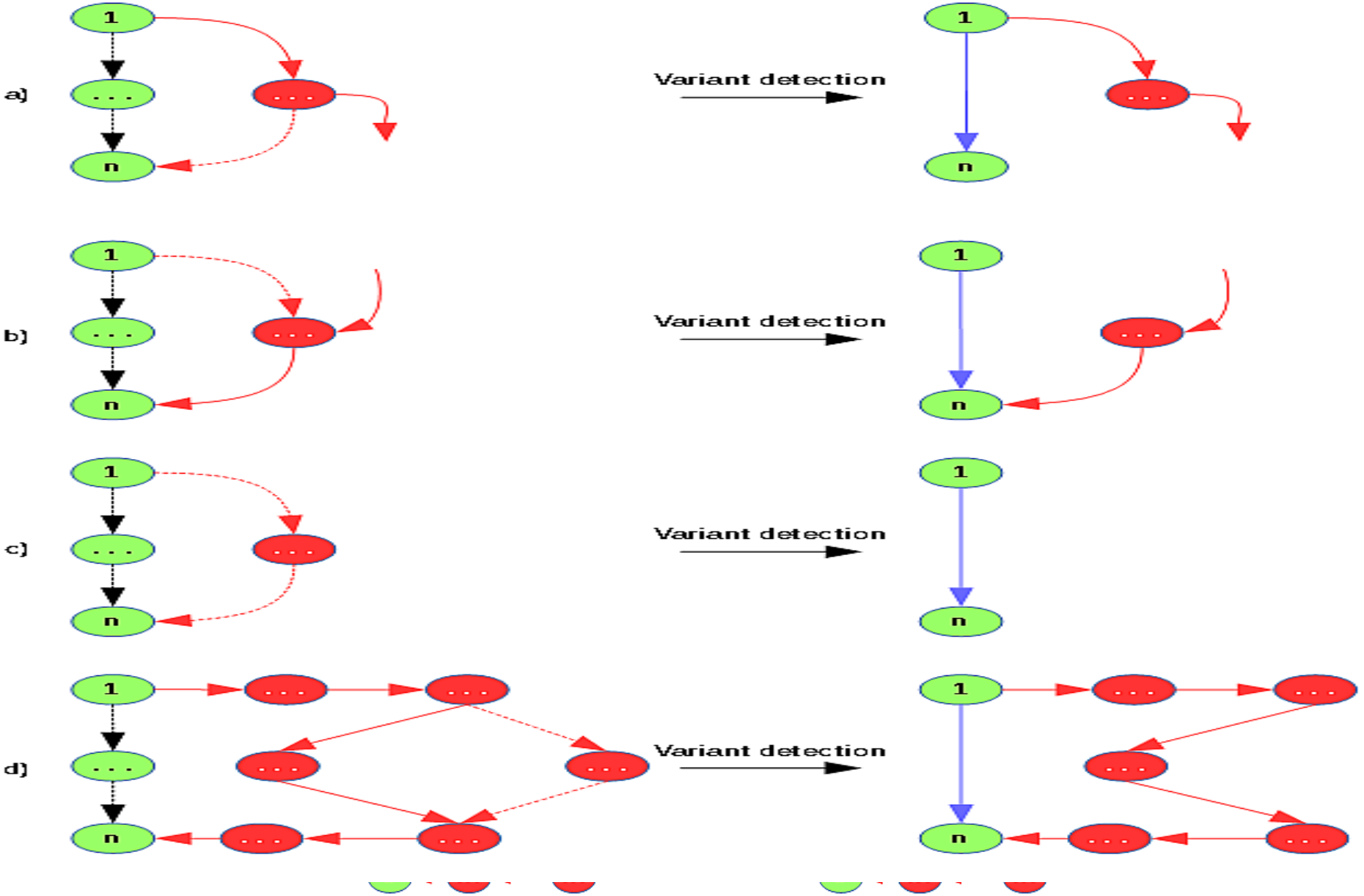
TATCGTATATATG



ACCGT**G**GTAAT
ACCGT**A**GTAAT



Variant Detection



Results

	Fermikit	Fermikit (regions)	mpileup	Our algorithm
True positives (SNP)	45,241 (89.3%)	47,630 (94%)	49,201 (97.1%)	48,172 (95%)
False Negatives (SNP)	5,432 (10.7%)	3,043 (6%)	1,472 (2.9%)	2,501 (5%)
False Positives (FP)	385 (0.76%)	1,441 (2.84%)	2,090 (4.12%)	816 (1.61%)
True Positives (INDEL)	7,853 (71.6%)	9,602 (87.55%)	8,707 (79.39%)	9,476 (86.4%)
False Negatives (INDEL)	3,114 (28.4%)	1,365 (13.45%)	2,260 (20.61%)	1,491 (13.6%)
False Positives (INDEL)	250 (2.28%)	835 (7.61%)	1,412 (12.95%)	1,255 (11.4%)

Different Parameters

K = 21	2	3	4	5	6	7	8
SNP TP	48464	48427	48172	47774	47156	46409	45342
SNP FN	2209	2246	2501	2899	3517	4264	5331
SNP FP	1364	1117	816	650	513	403	351
INDEL TP	10004	9851	9476	9265	8906	8474	7948
INDEL FN	963	1116	1491	1702	2061	2493	3019
INDEL FP	2581	1982	1255	905	684	532	431

K = 31	2	3	4	5	6	7	8
SNP TP	48353	48182	47733	47123	46250	45108	43555
SNP FN	2320	2491	2940	3550	4423	5565	7118
SNP FP	1252	1011	735	585	456	360	308
INDEL TP	9873	9659	9302	8903	8428	7923	7314
INDEL FN	1094	1308	1665	2064	2539	3044	3653
INDEL FP	1884	1341	891	623	470	363	292

Genotypic Results

	Fermi.kit	Fermi (regions)	mpileup	Our algorithm
SNP TP	44085	47735	48203	47024
SNP FN	6588	2938	2470	3649
SNP FP	1484	3408	3167	1808
INDEL TP	6375	8092	6864	7566
INDEL FN	4592	2875	4103	3401
INDEL FP	1343	2339	4462	3415