CLAIMS

Claim 1. A computer-implemented method for improving accuracy of a variant call by jointly evaluating reads that map to two or more regions of a reference sequence that are homologous, the method comprising:

accessing by one or more computers, a joint pile-up of a plurality of sequence reads, wherein the joint pile-up includes, a first pileup of reads that have been aligned to a first region of the reference sequence and at least a second pileup of reads that have been aligned to a second region of the reference sequence, wherein the first region and the second region are homologous with each other;

determining, by the one or more computers, a set of candidate variants from the joint-pileup;

defining, by the one or more computers, an order of processing of the candidate variants;

evaluating, by the one or more computers, each of the candidate variants from the set of candidate variants based on the defined processing order; and

generating by the one or more computers and based on the evaluation of the candidate, a variant call file that identifies one or more of the candidate variants.

Claim 2. The method of claim 1 wherein the method further comprising:

obtaining multiple homologous regions of a reference sequence from one or more memory devices.

Claim 3. The method of claim 1 or 2, wherein determining a set of candidate variants using the joint pile-up comprises:

using a De Brujin graph to extract candidate variants from the joint pileup.

Claim 4. The method of claim 4, wherein nodes in the graph represent the list of candidates and

wherein using the De Brujin graph includes, generating the De Brujin graph using each region of the reference sequence as a backbone and aligning each candidate variant positions to universal coordinates.

Claim 5. The method of claim 1, wherein defining, by the one or more computers, an order of processing of the candidate variants comprises:

defining by the one or more computers, an order of processing of the candidate variants as a function of read length and insert size.

Claim 6. The method of claim 5, wherein defining an order of processing of the candidate variants as a function of read length or insert size comprises:

generating a connection matrix that defines the order of processing of the candidate variants as a function of read length and insert size.

Claim 7. The method of claim 1, wherein evaluating, by the one or more computers each of the candidate variants from the set of candidate variants based on the defined processing order comprises:

for each candidate variants of the set of candidate variants:

generating joint diplotype candidates,

calculating an *a posteriori* probability of each joint diplotypes,

computing a genotype matrix,

pruning the candidate joint diplotypes, and

including a next position as evidence for a current position.

Claim 8. A system for improving the accuracy of a variant call by jointly evaluating reads that map to two or more regions of a reference sequence that are homologous, the system comprising:

one or more computers and one or more storage devices storing instructions that are operable, when executed by one or more computers, to cause the one or more computers to perform the operations comprising:

accessing, by one or more computers, a joint pile-up of a plurality of sequence reads, wherein the joint-pileup includes a first pileup of reads that have been aligned to a first region of the reference sequence and at least a second pileup of reads that have been aligned to a second region of the reference sequence, wherein the first region and the second region are homologous with each other;

determining, by the one or more computers, a set of candidate variants from the joint pileup;

defining, by the one or more computers, an order of processing of the candidate variants;

evaluating, by the one or more computers, each of the candidate variants from the set of candidate variants based on the defined processing order; and

generating by the one or more computers and based on the evaluation of the candidate variants a variant call file that identifies one or more of the candidate variants.

Claim 9. The system of claim 8, the operations further comprising:

obtaining multiple homologous regions of a reference sequence from one or more memory devices.

Claim 10. The system of claim 8 wherein determining a set of candidate variants from the joint pileup comprises:

using a De Brujin graph to extract candidate variants from the joint pileup.

Claim 11. The system of claim 10, wherein nodes in the graph represent the list of candidates, and

wherein using the De Brujin graph includes generating the De Brujin graph using each region of the reference sequence as a backbone and aligning each candidate variant positions to universal coordinates.,

Claim 12. The system of claim 8, wherein defining, by the one or more computers, an order of processing of the candidate variants comprises:

defining, by the one or more computers, an order of processing of the candidate variants as a function of read length or insert size.

Claim 13. The system of claim 10, wherein defining an order of processing of the candidate variants as a function of read length or insert size comprises:

generating a connection matrix that defines the order of processing of the candidate variants as a function of read length and insert size.

Claim 14. The system of claim 8, wherein evaluating, by the one or more computers each of the candidate variants from the set of candidate variants based on the defined processing order comprises:

for each candidate variant of the set of candidate variants:

generating candidate joint diplotypes,

calculating an *a posteriori* probability of each of the joint diplotypes,

computing a genotype matrix,

pruning the candidate joint diplotypes, and

including a next active position as evidence for a current position.

Claim 15. A computer-readable storage device having stored thereon instructions, which, when executed by a data processing apparatus, cause the data processing apparatus to perform operations for improving the accuracy of a variant call by jointly evaluating reads that map to two or more regions of a reference sequence that are homologous, the operations comprising:

accessing, by one or more computers, a joint pileup of a plurality of sequence reads, wherein the joint-pileup includes a first pileup of reads that have been aligned to a first region of the reference sequence and at least a second pileup of reads that have been aligned to a second region of the reference sequence, wherein the first region and the second region are homologous with each other;

determining by the one or more computers, a set of candidate variants from the joint-pileup;

defining, by the one or more computers, an order of processing of the candidate variants; evaluating by the one or more computers, each of the candidate variants from

the set of candidate variants based on the defined processing order; and

generating by the one or more computers and based on the evaluation of the candidate variants, a variant call file that identifies one or more of the candidate variants.,

Claim 16. The computer-readable storage device of claim 15, the operations further comprising:

obtaining multiple homologous regions of a reference sequence from one or more memory devices.

Claim 17. The computer-readable storage device of claim 15, wherein determining a set of candidate candidate variants using the joint -pileup comprises:

using a De Brujin graph to extract candidate variants from the joint pileup.

Claim 18 The computer-readable storage device of claim 17,

wherein nodes in the graph represent the list of candidates, and

wherein using the De Brujin graph includes generating the De Brujin graph using each region of the reference sequence as a backbone and aligning each candidate variant positions to universal coordinates.

Claim 19. The computer-readable storage device of claim 15, wherein defining, by the one or more computers, an order of processing of the candidate variants comprises:

generating a connection matrix that defines the order of processing of the candidate variants as a function of read length and insert size.

Claim 20. The computer-readable storage device of claim 15, wherein evaluating, by the one or more computers each of the candidate variants from the set of candidate variants based on the defined processing order comprises:

for each candidate variant of the set of candidate variants:

calculating an *a posteriori* probability of each of the joint diplotypes,

computing a genotype matrix,

pruning the candidate joint diplotypes, and

including a next position as evidence for a current position.