technische universität
dortmund

Analyse von RAD-Seq-Daten unter Berücksichtigung von Sequenzierfehlerraten und Heterozygotiewahrscheinlichkeiten

Antonie Vietor

28. Februar 2021

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl 11
Bioinformatics for High-Throughput Technologies
http://ls11-www.cs.tu-dortmund.de/

In Kooperation mit:
Universität Duisburg-Essen
Genome Informatics
http://genomeinformatics.uni-due.de/

Aufbau von DNA und RNA

Aufbau der DNA

- besteht aus Nukleotiden
- jedes **Nukleotid** besteht aus einem Zuckermolekül (Desoxyribose), einem Phosphatrest und einer Base
- **Basen**: A (Adenin), T (Thymin), G (Guanin), C (Cytosin)
- meist **doppelsträngig**
- dient vor allem der **Informationsspeicherung** (Erbinformation)

Aufbau von DNA und RNA

Aufbau der DNA

- besteht aus Nukleotiden
- jedes **Nukleotid** besteht aus einem Zuckermolekül (Desoxyribose), einem Phosphatrest und einer Base
- **Basen**: A (Adenin), T (Thymin), G (Guanin), C (Cytosin)
- meist **doppelsträngig**
- dient vor allem der **Informationsspeicherung** (Erbinformation)

Unterschiede im Aufbau der RNA

- **Nukleotide**: das Zuckermolekül ist Ribose
- **Basen**: Uracil (U) statt Thymin
- meist **einzelsträngig**
- viele Funktionen, dient unter anderem der **Informationsübertragung** bei der Proteinbiosynthese

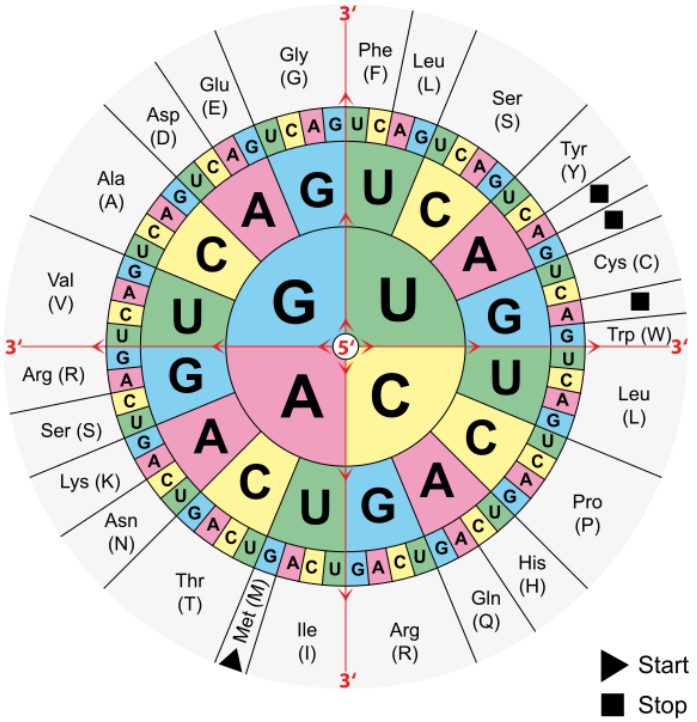
Struktur der DNA

- **Doppelhelixstruktur**
- **Komplementarität**: selektive Basenpaarung von A und T und ebenso von G und C
- **Antiparallelität**: in der Doppelhelix sind die beiden DNA-Stränge gegenläufig zu einander
- **Gene**: Wechsel von codierenden (Exons) und nicht-codierenden Abschnitten (Introns)
- zwischen den Genen nicht-codierende Bereiche, z.T. mit regulatorischen Funktionen
- ca. 98 % der DNA sind nicht-codierend

Proteinbiosynthese

Genetischer Code

- Codierung der **DNA-Sequenz** in eine **Aminosäuresequenz**, welche die Primärstruktur der Proteine darstellt
- **Basentripletts** (Codons) codieren für i.d.R. 20 Aminosäuren sowie ein Start- und drei Stop-Codons
- **Degeneration**: mehrere Basentripletts können für die gleiche Aminosäure codieren



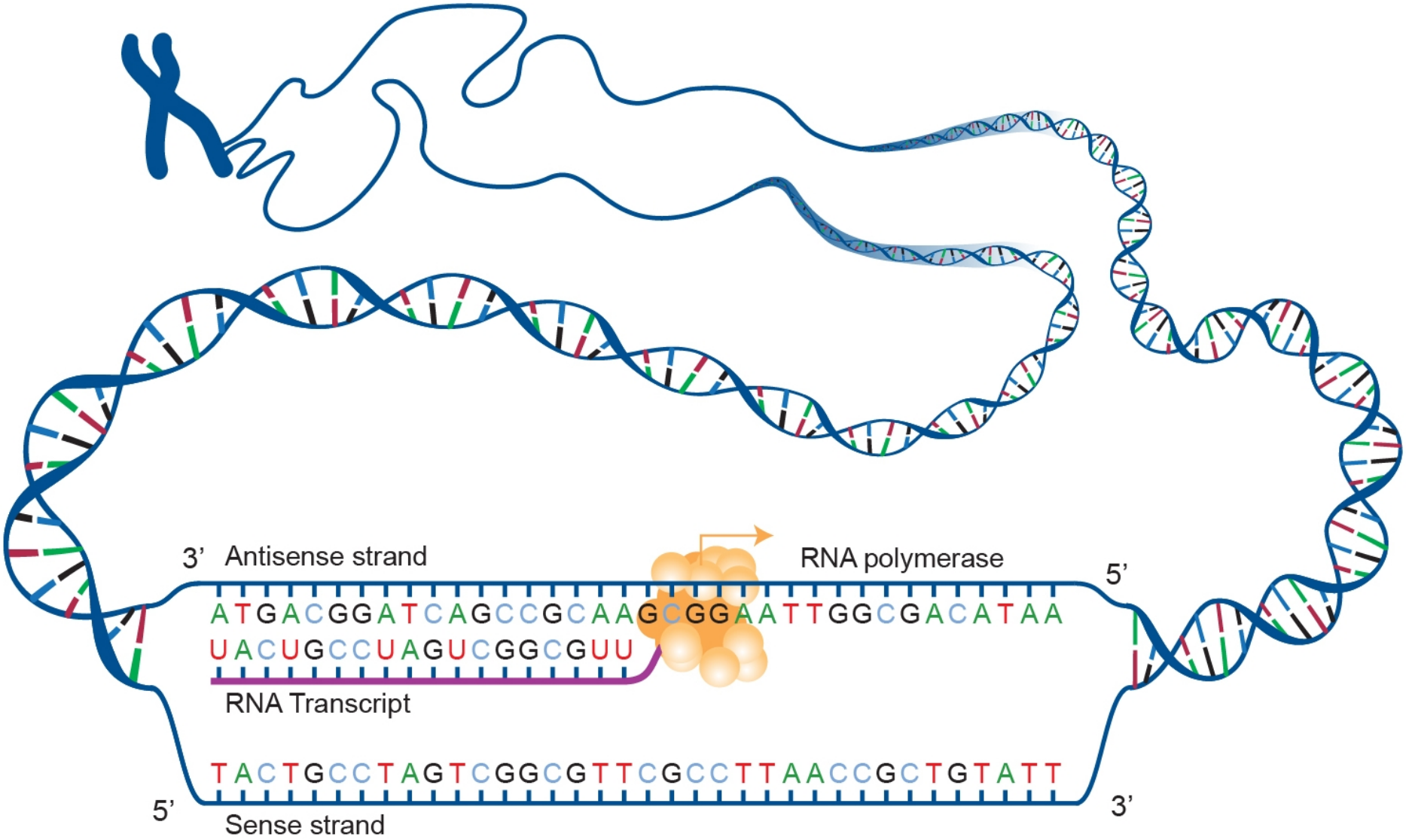
Bildquelle: [1]

Proteinbiosyntese

Transkription

Umschreiben eines DNA-Abschnitts zu Arbeitskopien in Form von **mRNA** (messenger RNA)

Proteinbiosyntese



Bildquelle: [2]

Proteinbiosyntese

Translation

- **Übersetzen** der Basensequenz in die Aminosäuresequenz mit Hilfe von **tRNA** (transfer RNA)

Proteinbiosyntese

Translation

- **Übersetzen** der Basensequenz in die Aminosäuresequenz mit Hilfe von **tRNA** (transfer RNA)
- Aufbau der tRNA:

  $\Rightarrow$ **mRNA-Bindungsstelle** bestehend aus einem Basentriplett

Proteinbiosyntese

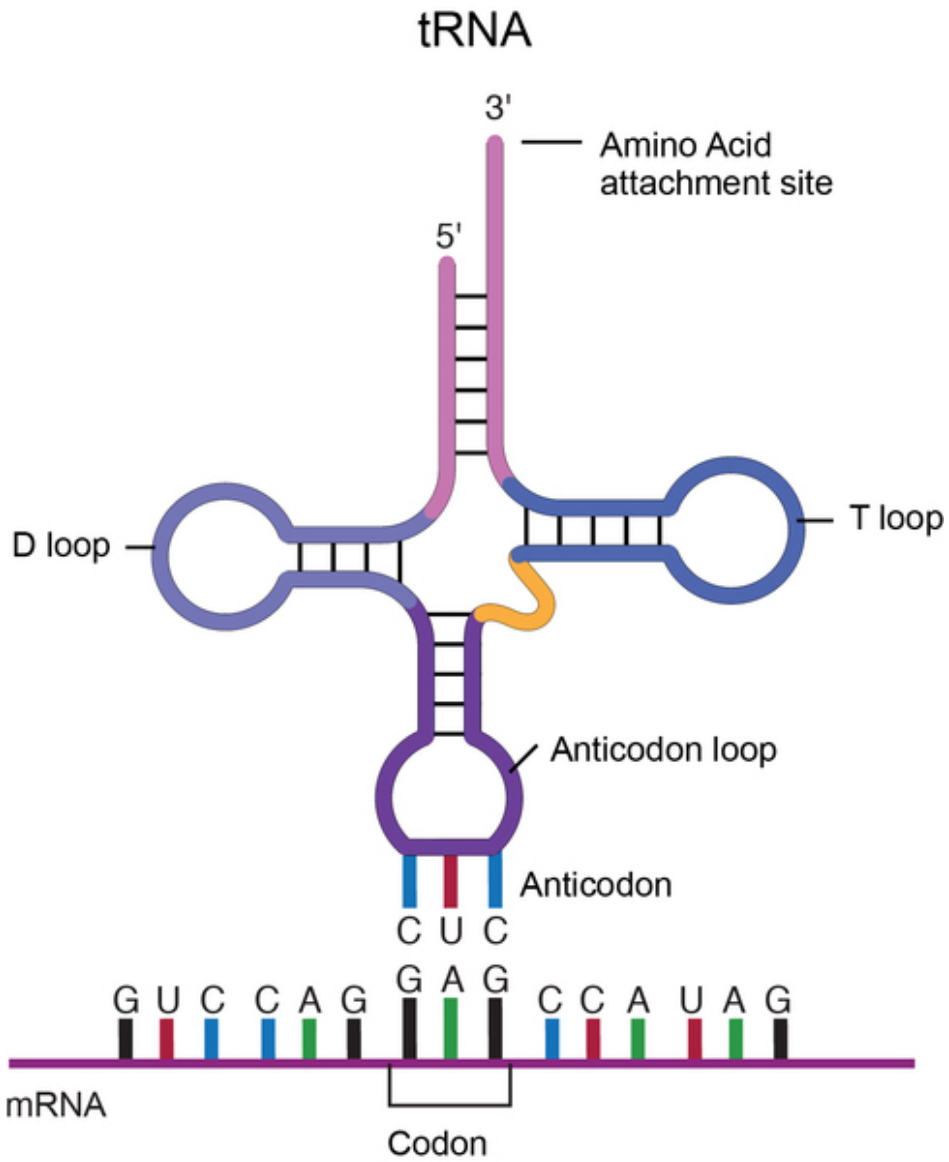
Translation

- **Übersetzen** der Basensequenz in die Aminosäuresequenz mit Hilfe von **tRNA** (transfer RNA)
- Aufbau der tRNA:

  $\Rightarrow$ **mRNA-Bindungsstelle** bestehend aus einem Basentriplett

  $\Rightarrow$ trägt die **korrespondierende Aminosäure** (AS), die nach dem genetischen Code der mRNA-Bindungsstelle entspricht

Proteinbiosyntese



Bildquelle: [3]

Proteinbiosyntese

Translation

- **Übersetzen** der Basensequenz in die Aminosäuresequenz mit Hilfe von **tRNA** (transfer RNA)
- Aufbau der tRNA:

  $\Rightarrow$ **mRNA-Bindungsstelle** bestehend aus einem Basentriplett

  $\Rightarrow$ trägt die **korrespondierende Aminosäure** (AS), die nach dem genetischen Code der mRNA-Bindungsstelle entspricht

Proteinbiosyntese

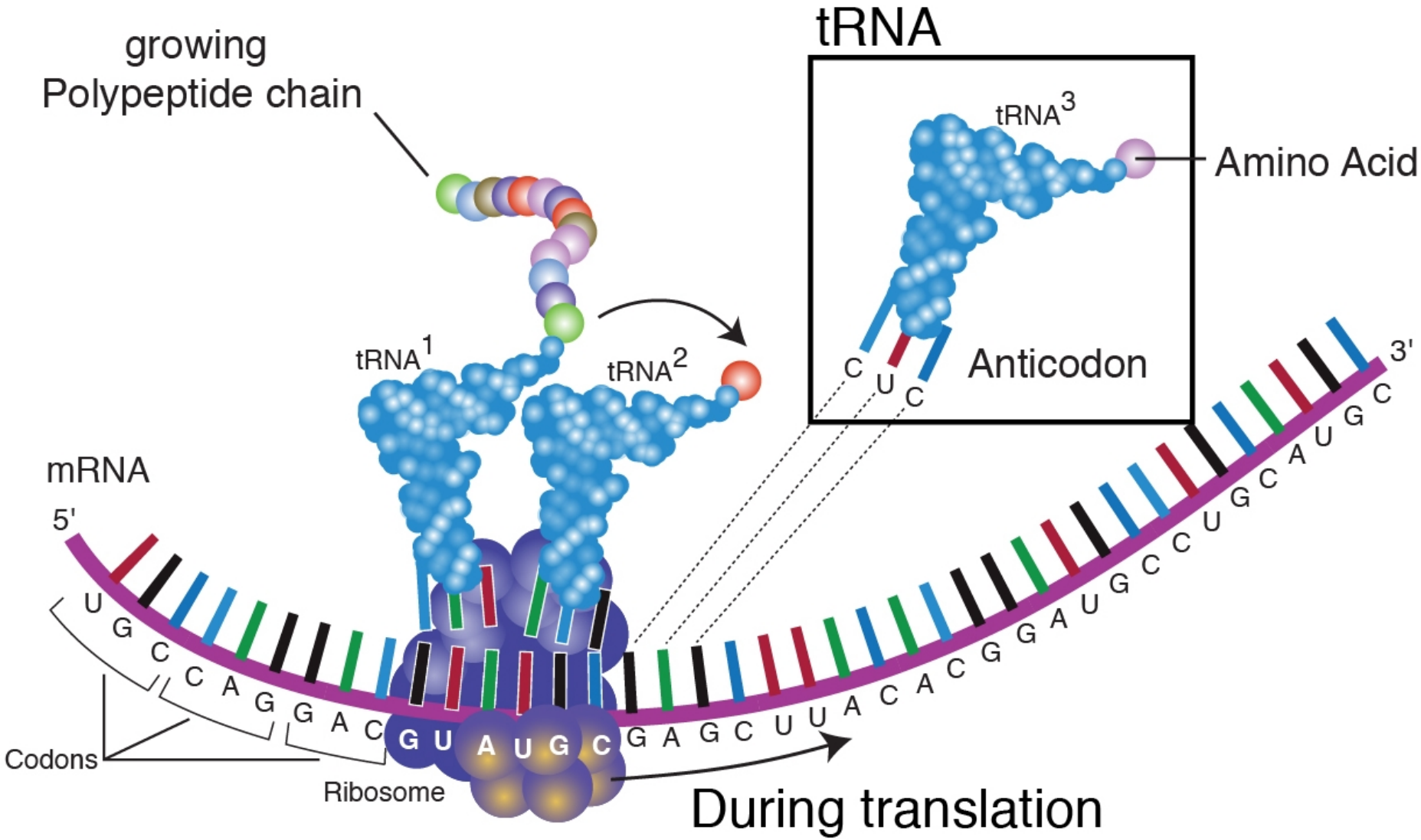
Translation

- **Übersetzen** der Basensequenz in die Aminosäuresequenz mit Hilfe von **tRNA** (transfer RNA)
- Aufbau der tRNA:

  $\Rightarrow$ **mRNA-Bindungsstelle** bestehend aus einem Basentriplett

  $\Rightarrow$ trägt die **korrespondierende Aminosäure** (AS), die nach dem genetischen Code der mRNA-Bindungsstelle entspricht
- von der Startsequenz ausgehend werden die tRNAs mit komplementärer Bindungsstelle nacheinander an die mRNA gebunden, dadurch wird ihre AS gelöst und an die AS der nachfolgenden tRNA gebunden $\Rightarrow$ es entsteht eine **Aminosäuresequenz**

Proteinbiosyntese



Bildquelle: [4]

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

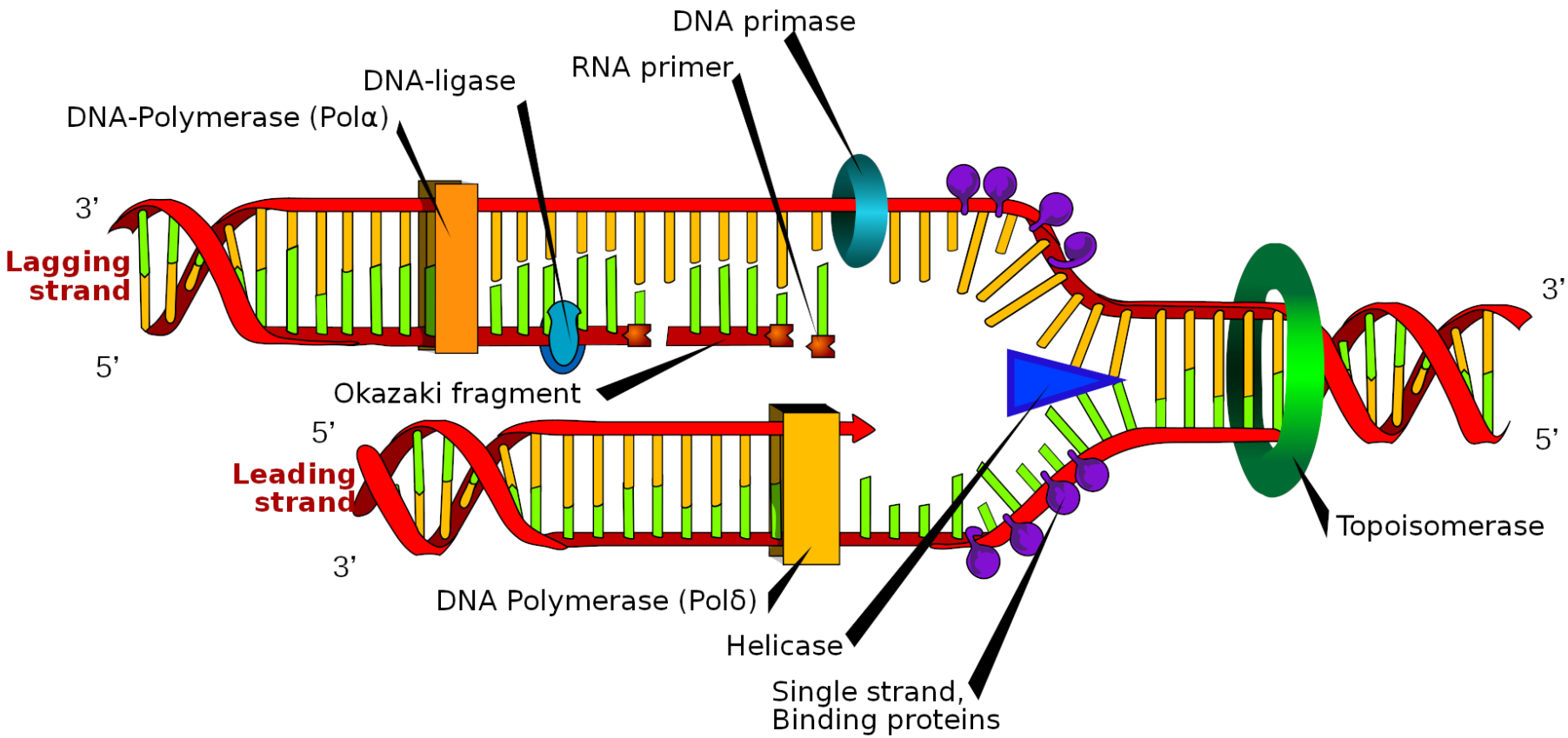
1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)

DNA-Replikation



Bildquelle: [5]

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)

   $\Rightarrow$ es entstehen zwei komplementäre Tochterstränge

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)

   $\Rightarrow$ es entstehen zwei komplementäre Tochterstränge

   $\Rightarrow$ kontinuierliche Synthese des Leitstrangs

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)

   $\Rightarrow$ es entstehen zwei komplementäre Tochterstränge

   $\Rightarrow$ kontinuierliche Synthese des Leitstrangs

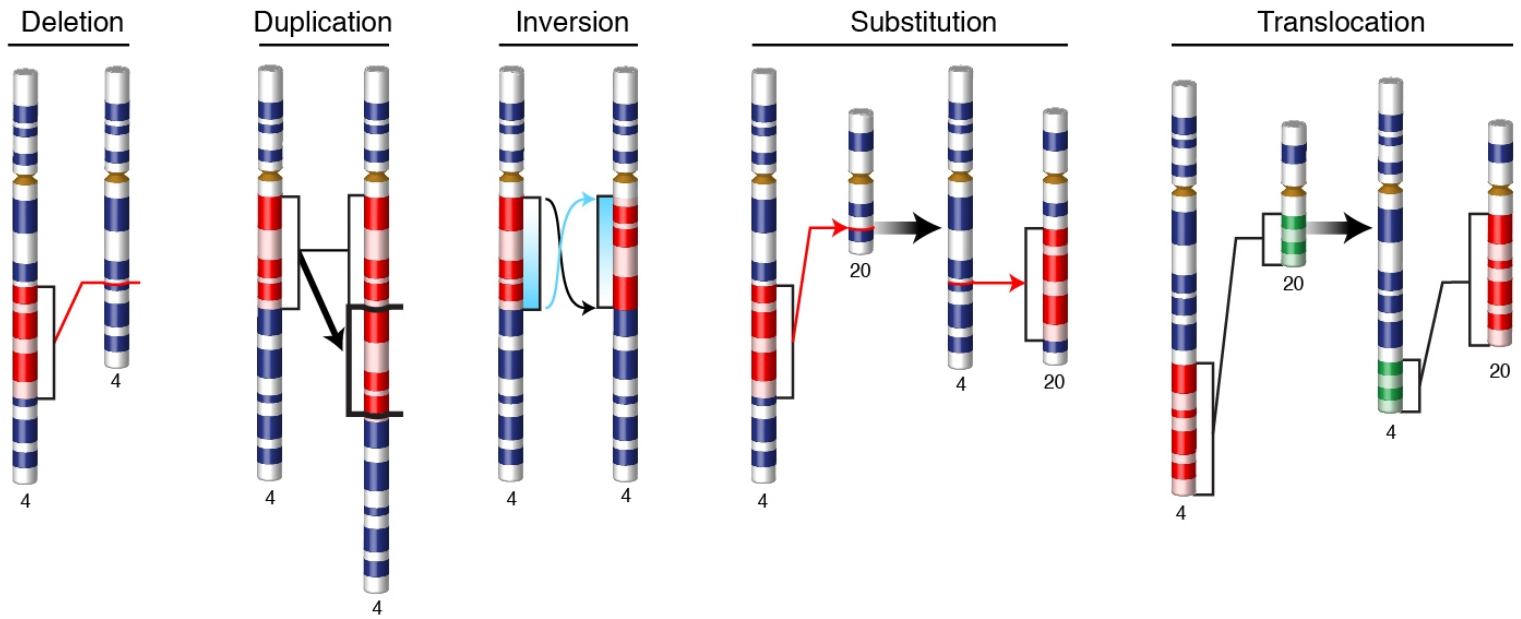
   $\Rightarrow$ diskontinuierliche Synthese des Folgestrangs (Okazaki-Fragmente)

DNA-Replikation

Natürlicher Vorgang zur Vervielfältigung der DNA bei der Zellteilung:

1. Entwindung der DNA (**Topoisomerasen**)
2. Auftrennung des DNA-Doppelstrangs (**Helikase**)
3. Synthese der RNA-Primer (**Primasen**)
4. Kopieren der beiden Elternstränge ausgehend von den RNA-Primern (**DNA-Polymerasen**)
   - $\Rightarrow$ es entstehen zwei komplementäre Tochterstränge
   - $\Rightarrow$ kontinuierliche Synthese des Leitstrangs
   - $\Rightarrow$ diskontinuierliche Synthese des Folgestrangs (Okazaki-Fragmente)
5. Verbindung der Okazaki-Fragmente des Folgestrangs (**Ligase**)

Mutationen

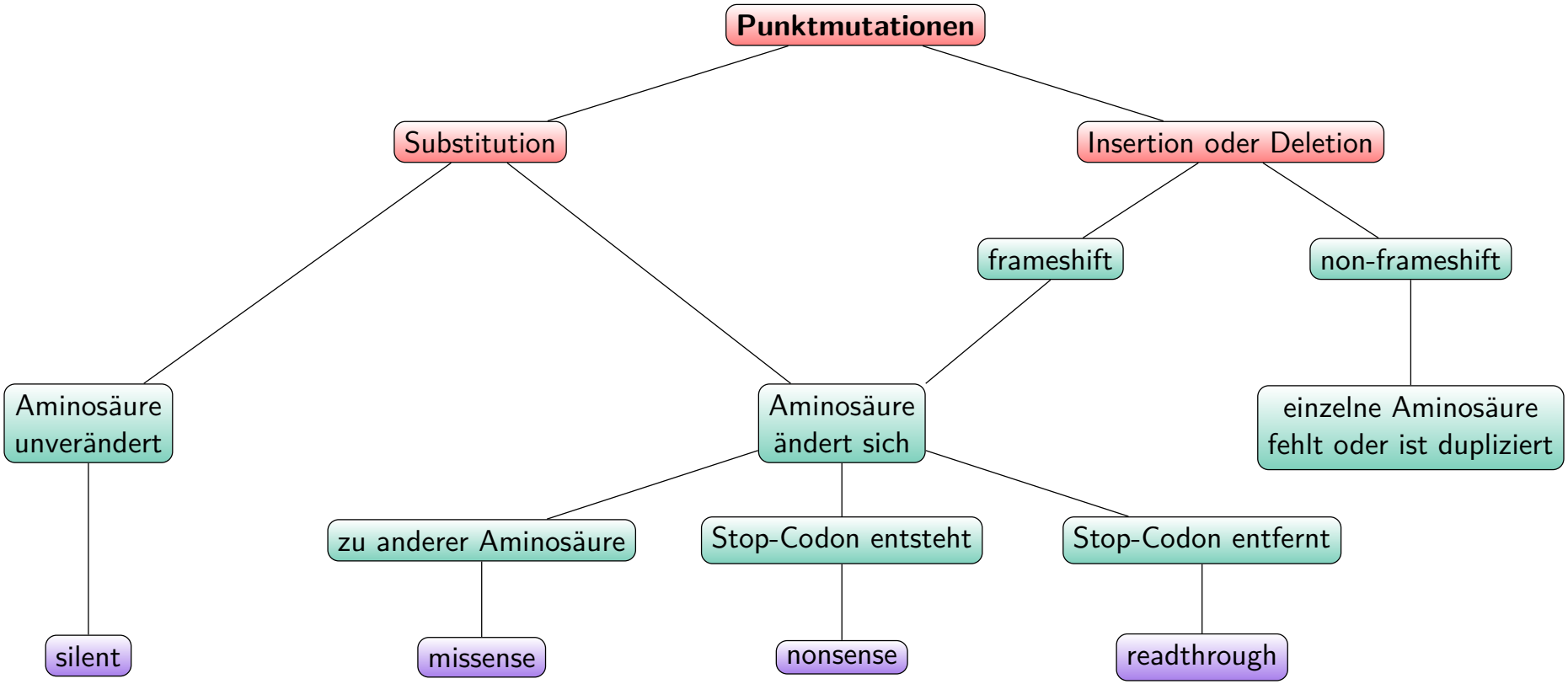


Bildquelle: [6]

Mutationen

Bildquelle: [6]

Mutationen


Punktmutationen
Substitution
Insertion oder Deletion
frameshift
non-frameshift
Aminosäure unverändert
Aminosäure ändert sich
einzelne Aminosäure fehlt oder ist dupliziert
zu anderer Aminosäure
Stop-Codon entsteht
Stop-Codon entfernt
silent
missense
nonsense
readthrough

Folgen von Mutationen

- Loss-of-function-Mutationen
- Gain-of-function-Mutationen

Folgen von Mutationen

- Loss-of-function-Mutationen
- Gain-of-function-Mutationen

Varianten

- oft Varianten einzelner Basen: SNPs (single nucleotide polymorphism)
- ohne pathologische Auswirkungen
- vermehrtes Auftreten innerhalb einer Spezies

Folgen von Mutationen

- Loss-of-function-Mutationen
- Gain-of-function-Mutationen

Varianten

- oft Varianten einzelner Basen: SNPs (single nucleotide polymorphism)
- ohne pathologische Auswirkungen
- vermehrtes Auftreten innerhalb einer Spezies

Allele, Locus, Ploidie und Genotyp

- **Allele**: verschiedene Varianten eines genomischen Ortes (**Locus**)
- **Ploidie**: Anzahl der Chromosomensätze (**homologe Chromosomen**)
- **Genotyp**:

  $\Rightarrow$ **Homozygotie**: an einem Locus liegt auf allen homologen Chromosomen das gleiche Allel vor

  $\Rightarrow$ **Heterozygotie**: die homologen Chromosomen weisen an einem Locus unterschiedliche Allele auf

Polymerase-Kettenreaktion (PCR)

- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:

Polymerase-Kettenreaktion (PCR)

- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:
  1. **Denaturierung**: durch Erhitzen wird der DNA-Doppelstrang in zwei Einzelstränge aufgespalten (96°C)

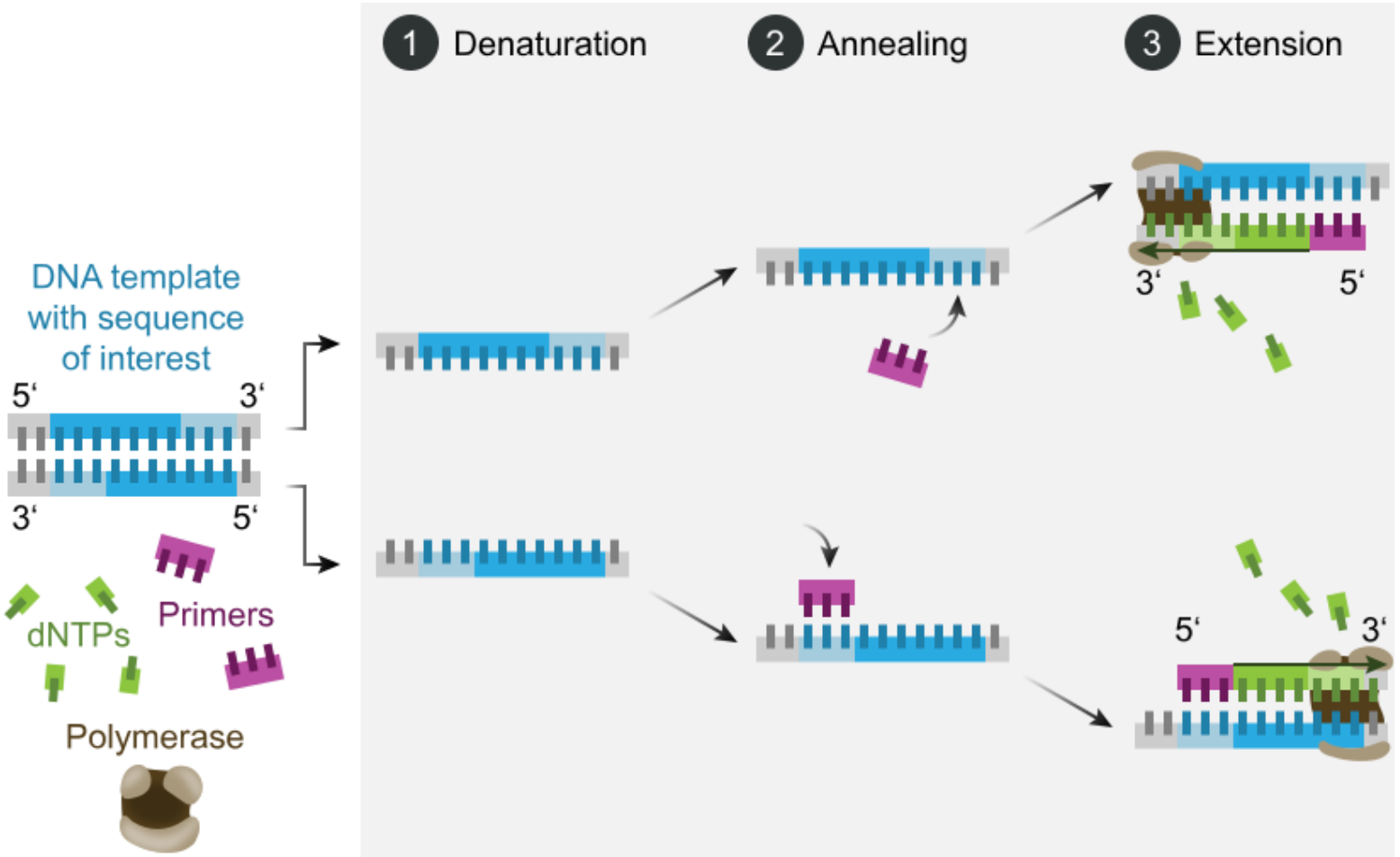
Polymerase-Kettenreaktion (PCR)

- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:
  1. **Denaturierung**: durch Erhitzen wird der DNA-Doppelstrang in zwei Einzelstränge aufgespalten (96°C)
  2. **Annealing**: Primerbindung an den 3'-Enden der zu amplifizierenden Gensequenz beider Einzelstränge (55-65°C)

Polymerase-Kettenreaktion (PCR)

- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:
  1. **Denaturierung**: durch Erhitzen wird der DNA-Doppelstrang in zwei Einzelstränge aufgespalten (96°C)
  2. **Annealing**: Primerbindung an den 3'-Enden der zu amplifizierenden Gensequenz beider Einzelstränge (55-65°C)
  3. **Elongation**: DNA-Synthese der komplementären Stränge (72°C)

Polymerase-Kettenreaktion (PCR)



Bildquelle: [7]

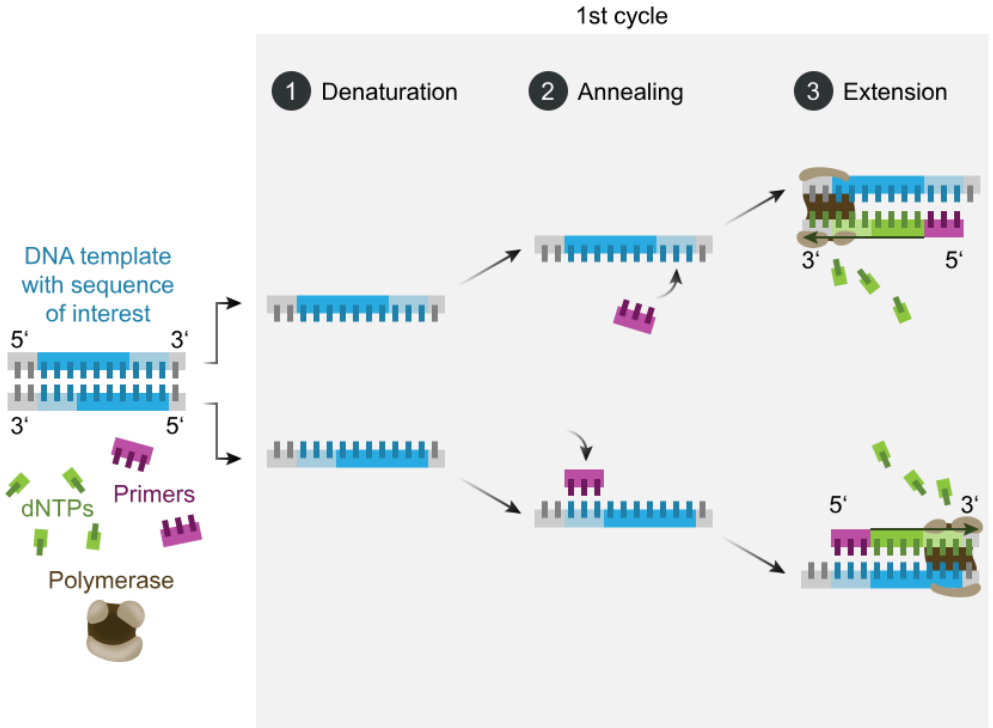
Polymerase-Kettenreaktion (PCR)

- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:
  1. **Denaturierung**: durch Erhitzen wird der DNA-Doppelstrang in zwei Einzelstränge aufgespalten (96°C)
  2. **Annealing**: Primerbindung an den 3'-Enden der zu amplifizierenden Gensequenz beider Einzelstränge (55-65°C)
  3. **Elongation**: DNA-Synthese der komplementären Stränge (72°C)

Polymerase-Kettenreaktion (PCR)

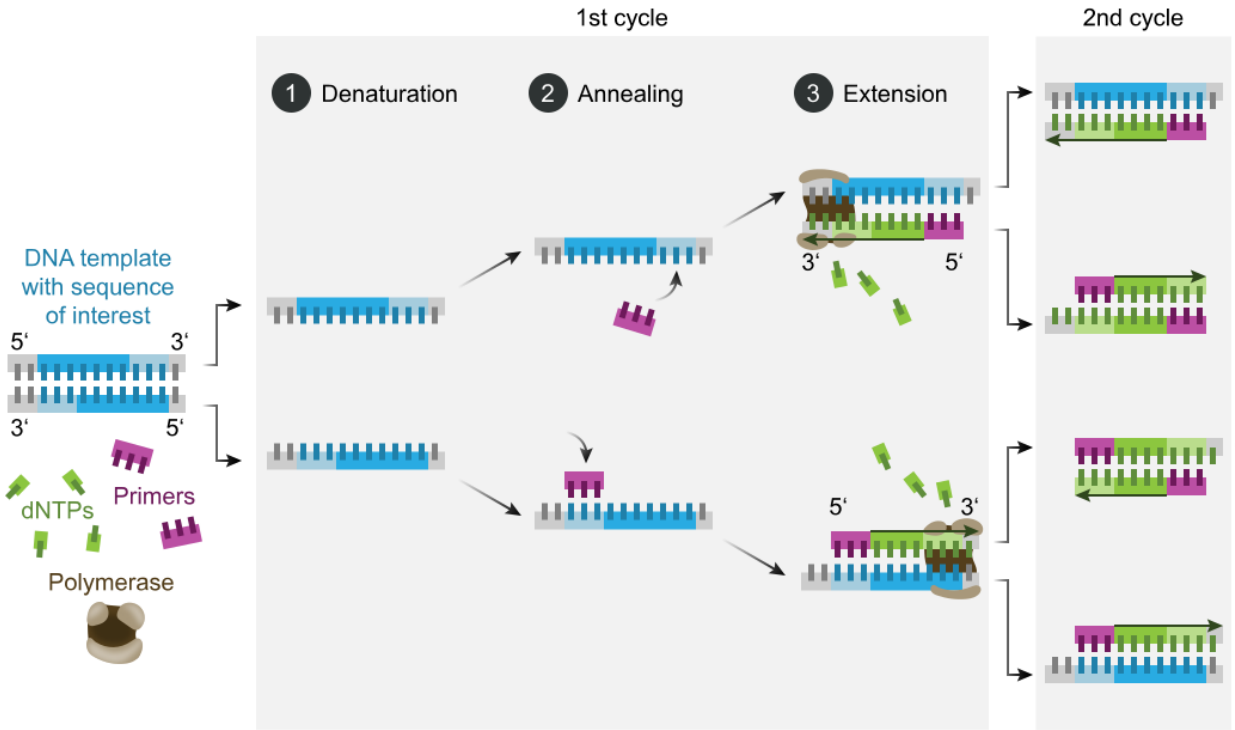
- Methode zur Vervielfältigung von DNA-Abschnitten
- mehrere Zyklen der folgenden temperaturabhängigen Schritte:
  1. **Denaturierung**: durch Erhitzen wird der DNA-Doppelstrang in zwei Einzelstränge aufgespalten (96°C)
  2. **Annealing**: Primerbindung an den 3'-Enden der zu amplifizierenden Gensequenz beider Einzelstränge (55-65°C)
  3. **Elongation**: DNA-Synthese der komplementären Stränge (72°C)
- mit jedem Zyklus wird die betreffende Sequenz verdoppelt
- in Abhängigkeit von der Anzahl der durchgeführten Zyklen n exponentieller Anstieg der Kopien 2^n

Polymerase-Kettenreaktion (PCR)



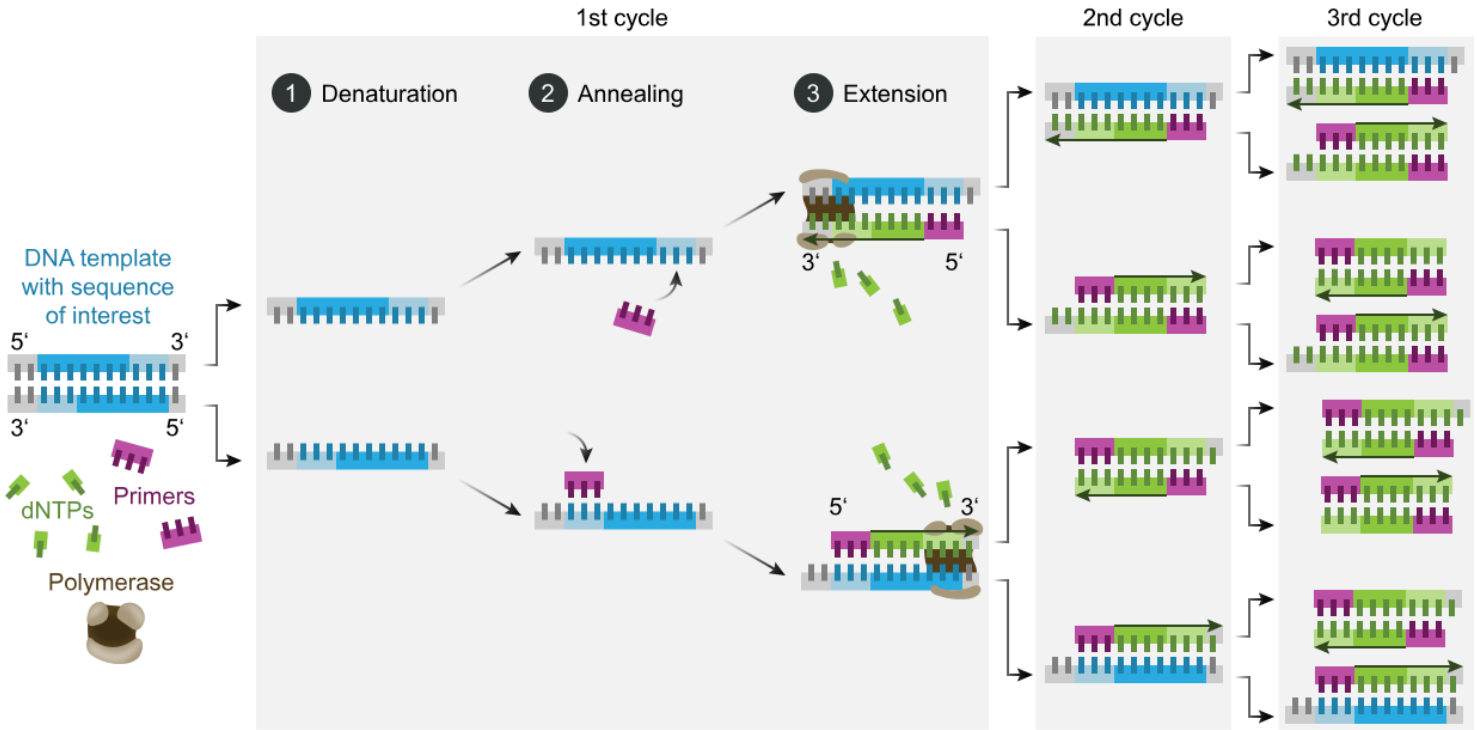
Bildquelle: [7]

Polymerase-Kettenreaktion (PCR)



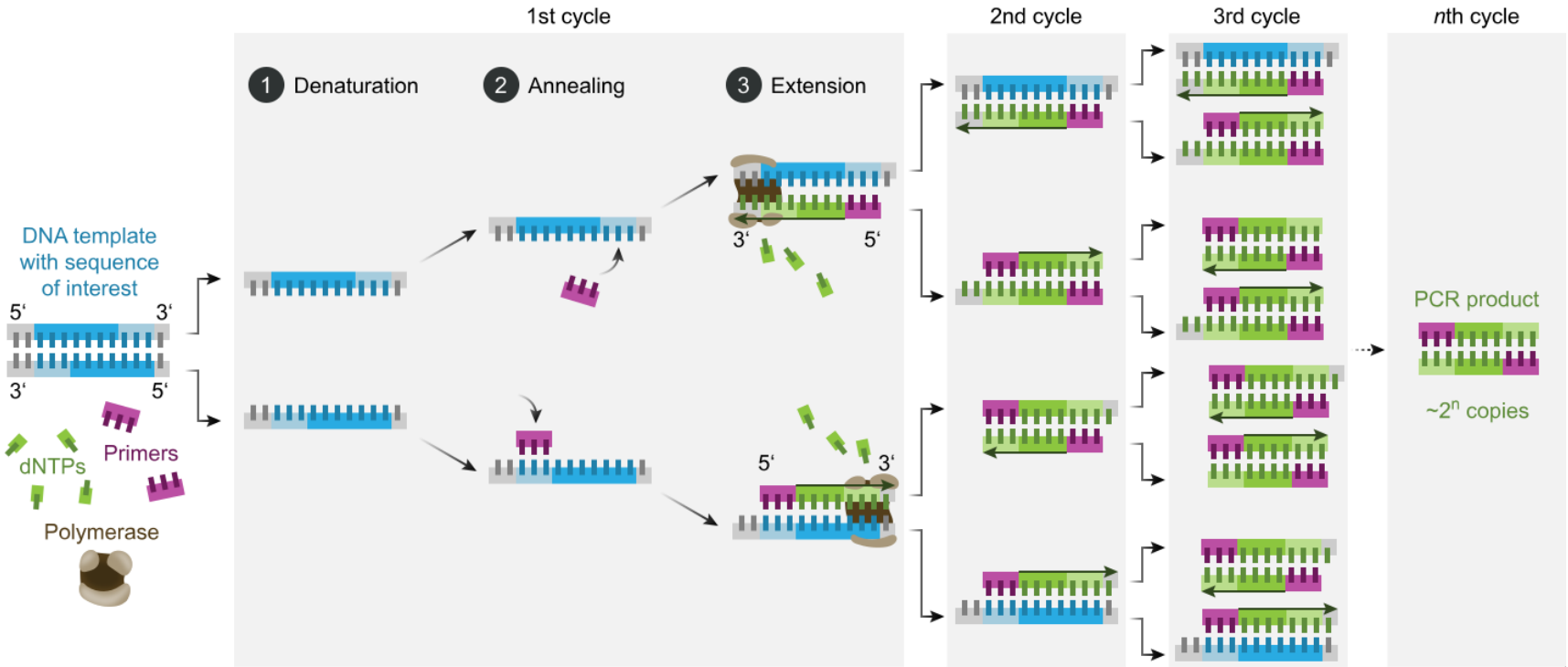
Bildquelle: [7]

Polymerase-Kettenreaktion (PCR)



Bildquelle: [7]

Polymerase-Kettenreaktion (PCR)



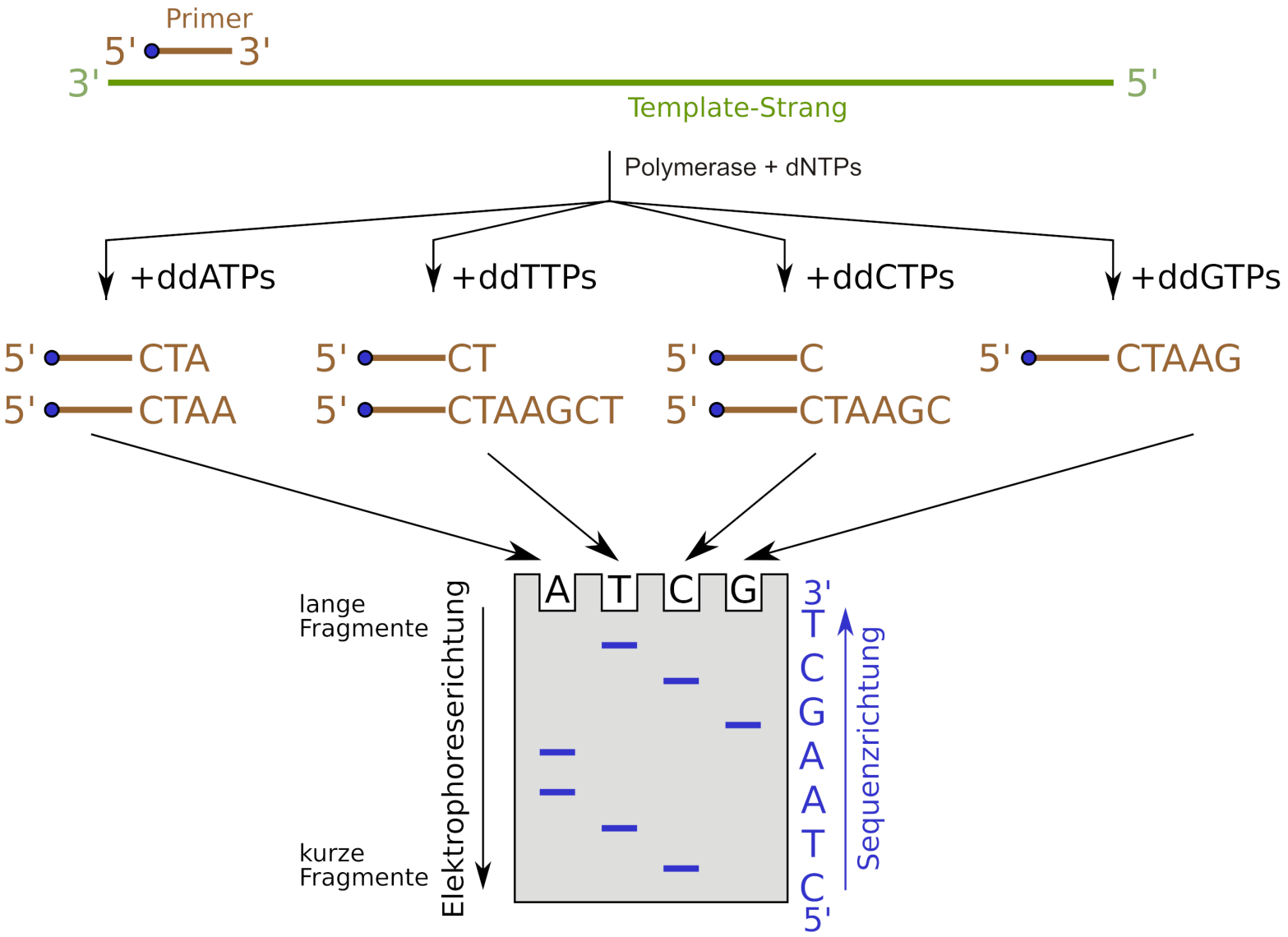
Bildquelle: [7]

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

Sequenzierung



Bildquelle: [8]

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

RAD-Sequenzierung

- restriction site associated DNA sequencing

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

RAD-Sequenzierung

- restriction site associated DNA sequencing
- **Anwendung**: Populationsgenetik, Ökologie, Genotypisierung, Evolutionsforschung

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

RAD-Sequenzierung

- restriction site associated DNA sequencing
- **Anwendung**: Populationsgenetik, Ökologie, Genotypisierung, Evolutionsforschung
- Sequenzierung multipler kleiner DNA-Fragmente aus dem gesamten Genom

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

RAD-Sequenzierung

- restriction site associated DNA sequencing
- **Anwendung**: Populationsgenetik, Ökologie, Genotypisierung, Evolutionsforschung
- Sequenzierung multipler kleiner DNA-Fragmente aus dem gesamten Genom
- gleichzeitige Analyse mehrerer Individuen in gepoolten Proben

Sequenzierung

Sanger-Sequenzierung

Kettenabbruch-Synthese mit vier Probenansätzen denen jeweils eine der vier möglichen Nukleotide in modifizierter Form beigefügt wird

NGS-Sequenzierung

verbesserte Sequenziertechnologien im Hochdurchsatzverfahren

RAD-Sequenzierung

- restriction site associated DNA sequencing
- **Anwendung**: Populationsgenetik, Ökologie, Genotypisierung, Evolutionsforschung
- Sequenzierung multipler kleiner DNA-Fragmente aus dem gesamten Genom
- gleichzeitige Analyse mehrerer Individuen in gepoolten Proben
- benötigt kein Referenzgenom

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Bildquelle: [8]

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Bildquelle: [8]

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme
2. Sequenz der Restriktionsstelle ist bekannt, dies ermöglicht die Bindung der **Adapter-** und **Barcodesequenzen**

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme
2. Sequenz der Restriktionsstelle ist bekannt, dies ermöglicht die Bindung der **Adapter-** und **Barcodesequenzen**
3. **Größenselektion** der DNA-Fragmente

RAD-Verfahren

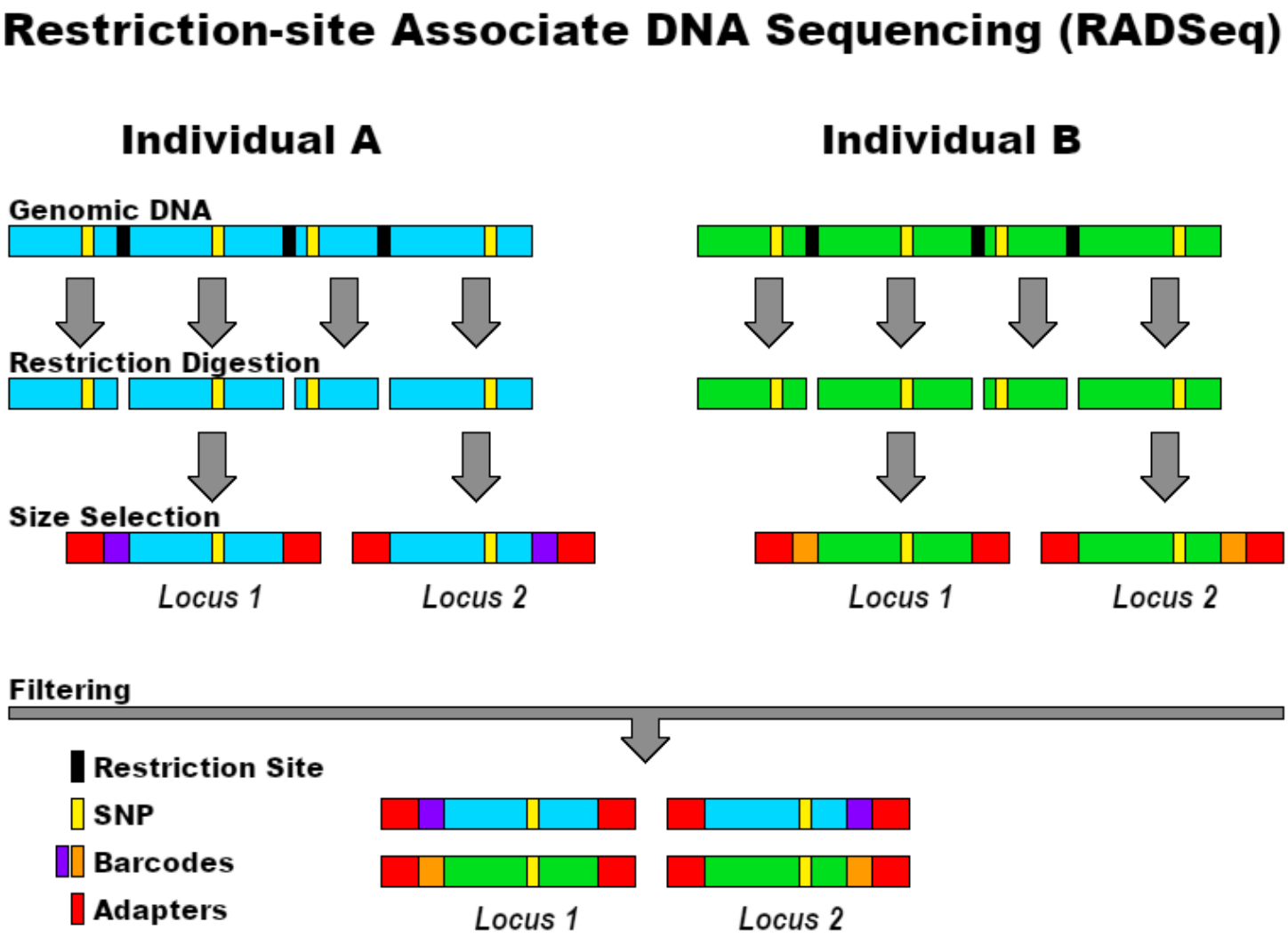
Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme
2. Sequenz der Restriktionsstelle ist bekannt, dies ermöglicht die Bindung der **Adapter-** und **Barcodesequenzen**
3. **Größenselektion** der DNA-Fragmente
4. **Sequenzierung** der gepoolten Proben verschiedener Individuen

RAD-Verfahren



Bildquelle: [9] (modifiziert)

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme
2. Sequenz der Restriktionsstelle ist bekannt, dies ermöglicht die Bindung der **Adapter**- und **Barcodesequenzen**
3. **Größenselektion** der DNA-Fragmente
4. **Sequenzierung** der gepoolten Proben verschiedener Individuen

ddRADSeq (double digest RAD sequencing)

- Verwendung von zwei verschiedenen Restriktionsenzymen

RAD-Verfahren

Restriktionsenzyme

- molekulare Scheren, welche die DNA an spezifischen Sequenzen schneiden
- Enden können glatt oder versetzt sein

Methode

1. **DNA-Verdau** durch Restriktionsenzyme
2. Sequenz der Restriktionsstelle ist bekannt, dies ermöglicht die Bindung der **Adapter-** und **Barcodesequenzen**
3. **Größenselektion** der DNA-Fragmente
4. **Sequenzierung** der gepoolten Proben verschiedener Individuen

ddRADSeq (double digest RAD sequencing)

- Verwendung von zwei verschiedenen Restriktionsenzymen
- bessere Steuerbarkeit und höhere Genauigkeit

RADSeq-Verfahren

- durch die **Sequenzspezifität der Restriktionsenzyme** stammen die DNA-Fragmente bei den verschiedenen Individuen meistens vom gleichen genomischen Locus

RADSeq-Verfahren

- durch die **Sequenzspezifität der Restriktionsenzyme** stammen die DNA-Fragmente bei den verschiedenen Individuen meistens vom gleichen genomischen Locus
- interindividueller Vergleich ist **ohne Referenzgenom** möglich

RADSeq-Verfahren

- durch die **Sequenzspezifität der Restriktionsenzyme** stammen die DNA-Fragmente bei den verschiedenen Individuen meistens vom gleichen genomischen Locus
- interindividueller Vergleich ist **ohne Referenzgenom** möglich
- **gepoolte Proben**: Zeit- und Kostenersparnis, gleiche Versuchsbedingungen für die verschiedenen Individuen

RADSeq-Verfahren

- durch die **Sequenzspezifität der Restriktionsenzyme** stammen die DNA-Fragmente bei den verschiedenen Individuen meistens vom gleichen genomischen Locus
- interindividueller Vergleich ist **ohne Referenzgenom** möglich
- **gepoolte Proben**: Zeit- und Kostenersparnis, gleiche Versuchsbedingungen für die verschiedenen Individuen
- die DNA-Fragmente stammen aus dem gesamten Genom, aber **keine vollständige genomische Abdeckung**

Problemstellung

Problem:

- Reads ohne Kenntnis eines Referenzgenoms möglichen Loci zuordnen
- die Loci und ihre Sequenz sind unbekannt

Problemstellung

Problem:

- Reads ohne Kenntnis eines Referenzgenoms möglichen Loci zuordnen
- die Loci und ihre Sequenz sind unbekannt

Gegeben:

- Menge von Reads: $D = (s_1, \ldots, s_m) \in \{A, C, G, T, \}^{k^m}$
- Qualität der Sequenzierung: $Q = (q_1, \ldots, q_m) \in [0, 1]^{k^m}$
- Sequenzierfehlerraten: $\epsilon = \{\epsilon_{ins}, \epsilon_{del}\}$
- Heterozygotiewahrscheinlichkeiten: $\eta = \{\eta_{sub}, \eta_{ins}, \eta_{del}\}$
- Ploidie: ϕ

Problemstellung

Problem:

- Reads ohne Kenntnis eines Referenzgenoms möglichen Loci zuordnen
- die Loci und ihre Sequenz sind unbekannt

Gegeben:

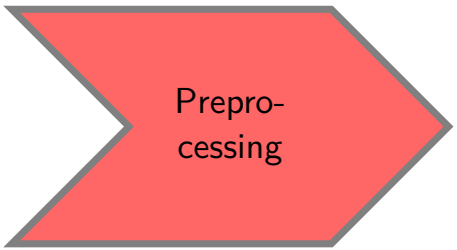
- Menge von Reads: $D = (s_1, \ldots, s_m) \in \{A, C, G, T, \}^{k^m}$
- Qualität der Sequenzierung: $Q = (q_1, \ldots, q_m) \in [0, 1]^{k^m}$
- Sequenzierfehlerraten: $\epsilon = \{\epsilon_{ins}, \epsilon_{del}\}$
- Heterozygotiewahrscheinlichkeiten: $\eta = \{\eta_{sub}, \eta_{ins}, \eta_{del}\}$
- Ploidie: ϕ

Ziel:

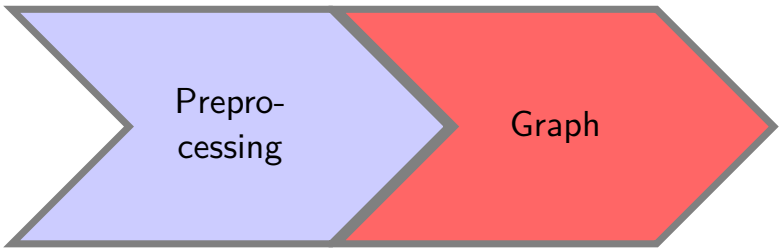
- Zuordnung der Reads zu den Loci unter Berücksichtigung von ϵ und η
- Ausgabe der Menge der ermittelten Loci mit den Sequenzen der beteiligten Allele

$\Rightarrow$ die Loci können anschließend für Diversitätsanalysen genutzt werden

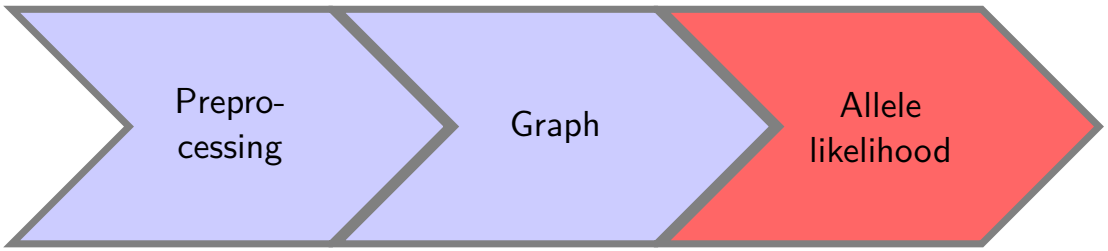
Model



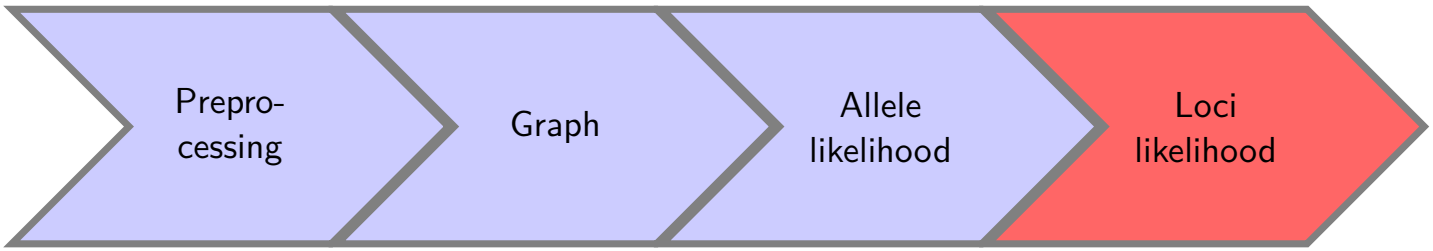
Model



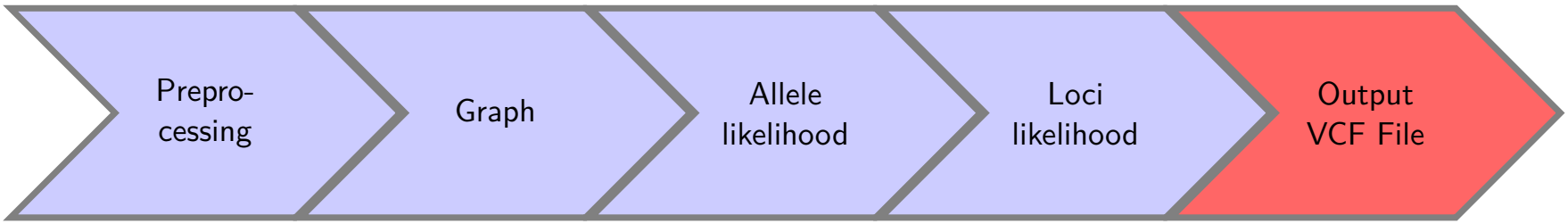
Model



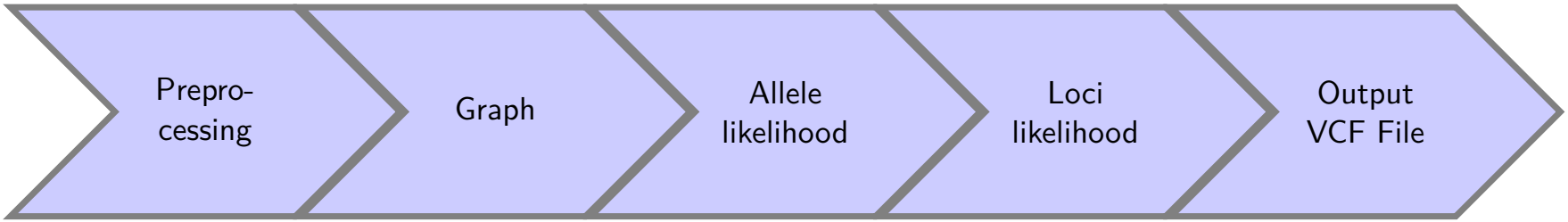
Model



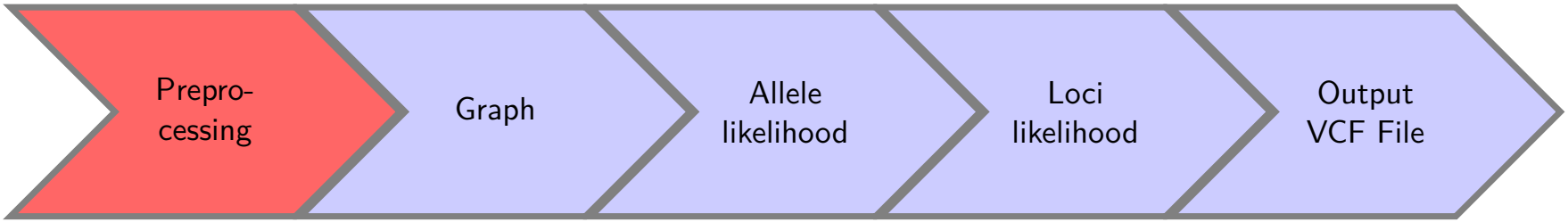
Model


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

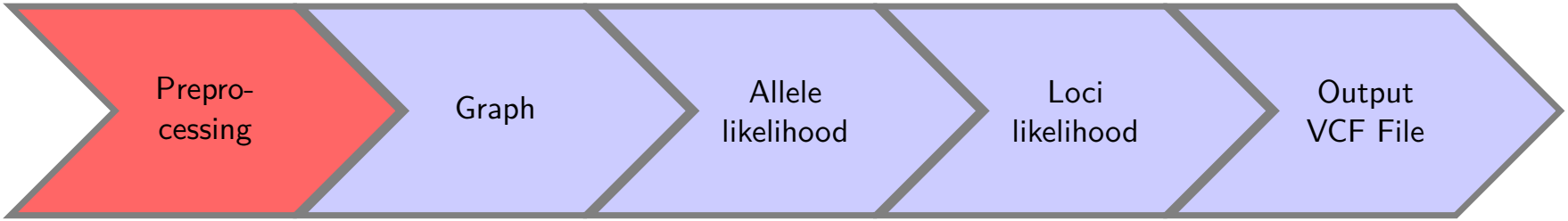
Model


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

Model

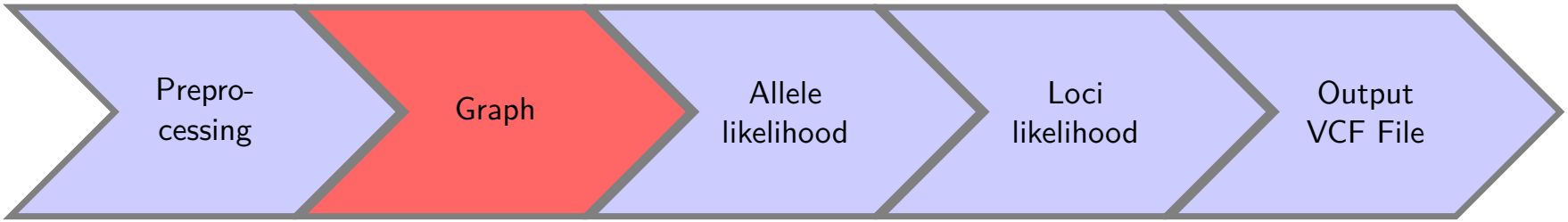

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

Model

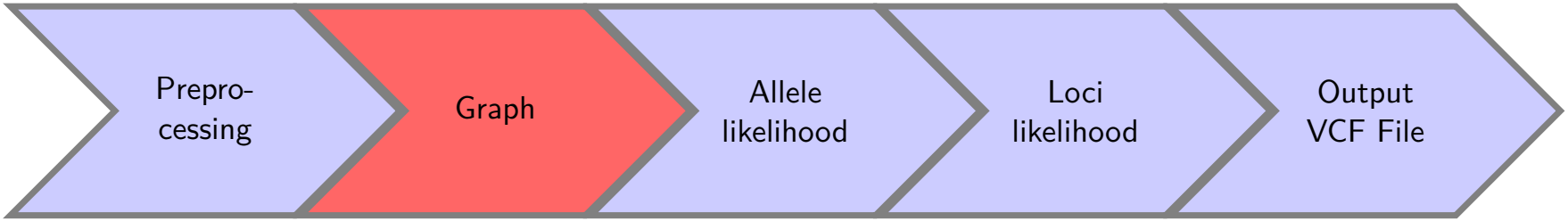


- Statistiken zur Qualität der Reads
- Individuen werden entsprechend ihres Barcodes separiert
- Entfernen der Barcode- und Adaptersequenzen
- Erzeugen eines Sequenzalignments

Model


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

Model



- das Problem wird als gerichteter Graph $G = (V, E)$ betrachtet
- die Knoten werden durch die Reads repräsentiert
- die Kanten basieren auf dem Sequenzalignment
- das Sequenzalignment wird als approximiertes pairHMM betrachtet

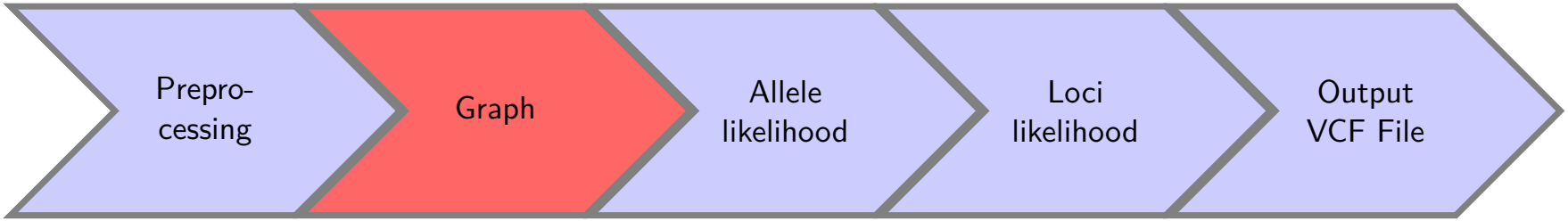
pairHMM vs. Minimap

tabelle

Likelihoodberechnung beim approximierten pairHMM

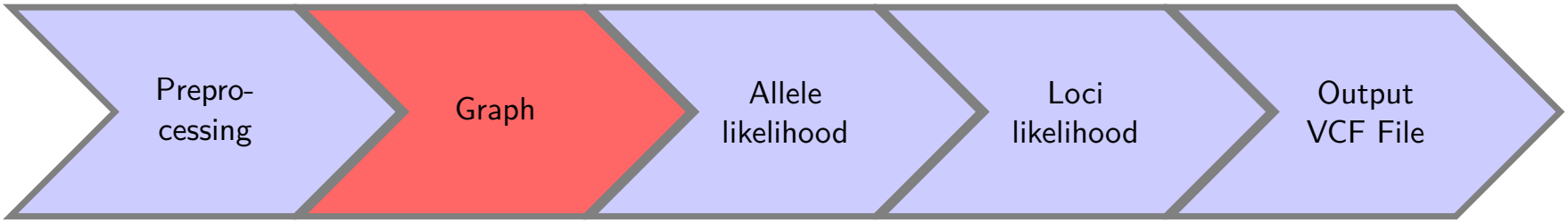
Tabelle

Model



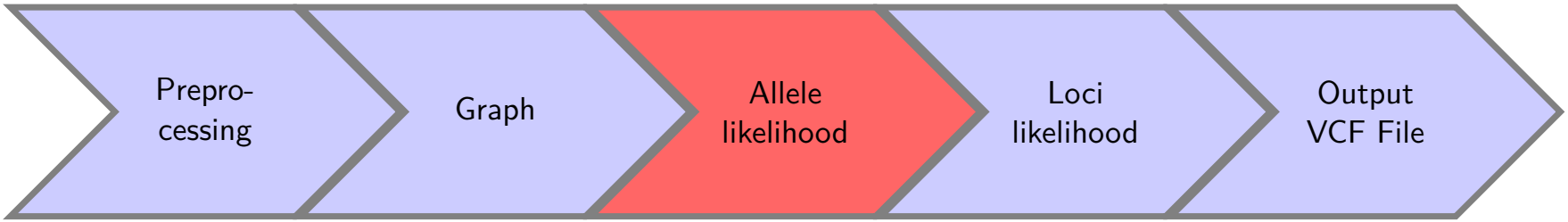
- das Problem wird als gerichteter Graph $G = (V, E)$ betrachtet
- die Knoten werden durch die Reads repräsentiert
- die Kanten basieren auf dem Sequenzalignment
- das Sequenzalignment wird als approximiertes pairHMM betrachtet

Model



- das Problem wird als gerichteter Graph $G = (V, E)$ betrachtet
- die Knoten werden durch die Reads repräsentiert
- die Kanten basieren auf dem Sequenzalignment
- das Sequenzalignment wird als approximiertes pairHMM betrachtet
- Kanten entstehen nur zwischen Knoten deren Readsequenzen einander ähneln
- Partitionierung des Graphen in mehrere Zusammenhangskomponenten $\Rightarrow$ das Gesamtproblem wird in mehrere Teilprobleme aufgeteilt

Model


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

Model

- Identifizierung der Allele, von denen die übrigen Reads der Zusammenhangskomponenten am wahrscheinlichsten durch Sequenzierfehler entstanden sind

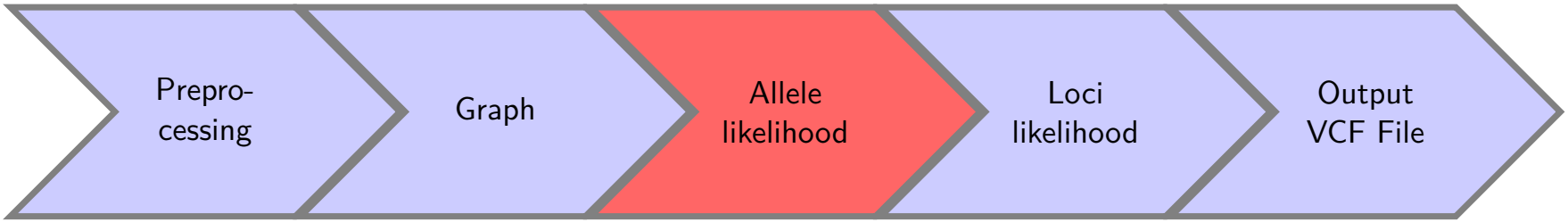
Model

- Identifizierung der Allele, von denen die übrigen Reads der Zusammenhangskomponenten am wahrscheinlichsten durch Sequenzierfehler entstanden sind
- Kandidatenallele und Anzahl der tatsächlich zu erwartenden Allele $n_{alleles}$ bestimmen:

$$n_{alleles} = \begin{cases} \phi, & \phi \geq n_{cand} \\ n_{cand} + \phi - d, & \phi < n_{cand} \wedge d \neq 0 \\ n_{cand}, & \phi < n_{cand} \wedge d = 0 \end{cases}$$

(es gilt $d = n_{cand} \mod \phi$)

Model



$\Rightarrow$ aus den Kandidatenallelen **Kombinationen mit Wiederholung** der Länge $n_{alleles}$ gebildet:

Model

$\Rightarrow$ aus den Kandidatenallelen **Kombinationen mit Wiederholung** der Länge $n_{alleles}$ gebildet:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$:

[(0, 0), (0, 1), (1, 1)]

Model

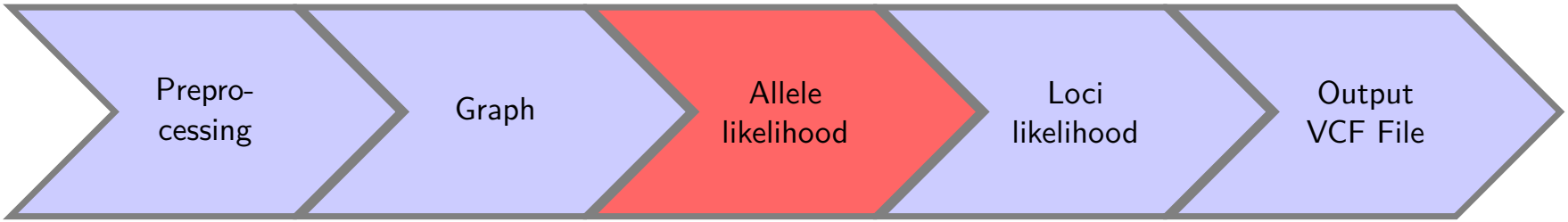
$\Rightarrow$ aus den Kandidatenallelen **Kombinationen mit Wiederholung** der Länge $n_{alleles}$ gebildet:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$:

[(0, 0), (0, 1), (1, 1)]

$\Rightarrow$ aus den Kombinationen werden die Häufigkeitsverteilungen der Kandidatenallele (**Allele-Fractions**) bestimmt:

Model



$\Rightarrow$ aus den Kandidatenallelen **Kombinationen mit Wiederholung** der Länge $n_{alleles}$ gebildet:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$:

[(0, 0), (0, 1), (1, 1)]

$\Rightarrow$ aus den Kombinationen werden die Häufigkeitsverteilungen der Kandidatenallele (**Allele-Fractions**) bestimmt:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$:

[1.0, 0.0], [0.5, 0.5], [0.0, 1.0]

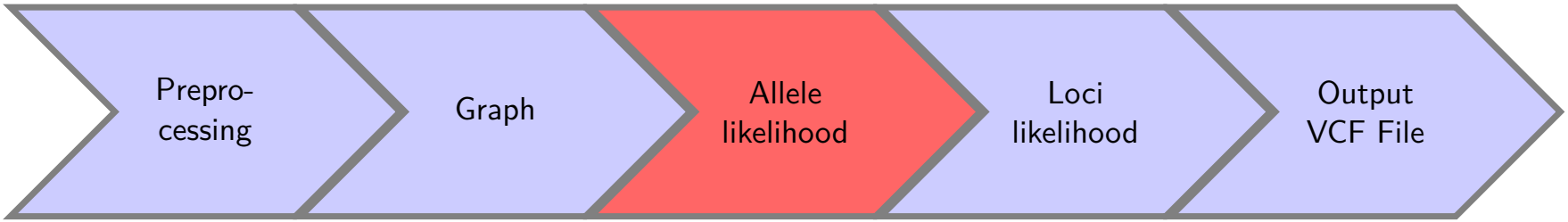
Model

Allelkombinationen:

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]

Model



Allelkombinationen:

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]

Allele-Fractions:

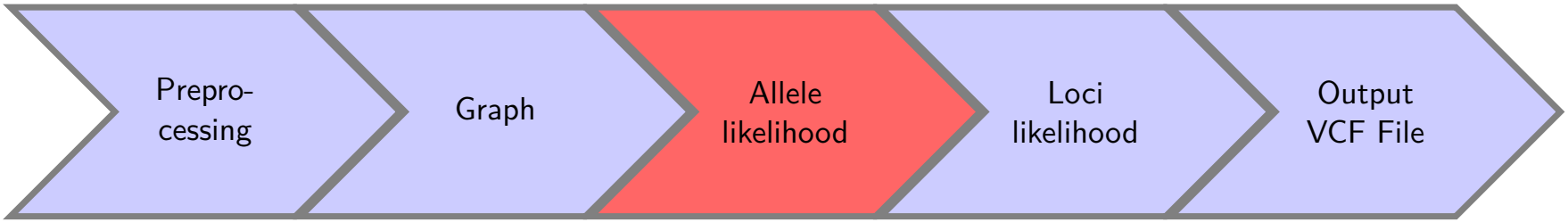
$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

[1.0, 0.0, 0.0], [0.75, 0.25, 0.0], [0.75, 0.0, 0.25], [0.5, 0.5, 0.0], [0.5, 0.25, 0.25], [0.5, 0.0, 0.5], [0.25, 0.75, 0.0], [0.25, 0.5, 0.25], [0.25, 0.25, 0.5], [0.25, 0.0, 0.75], [0.0, 1.0, 0.0], [0.0, 0.75, 0.25], [0.0, 0.5, 0.5], [0.0, 0.25, 0.75], [0.0, 0.0, 1.0]

Model

Für **jeden Read** mit der Sequenz s_r wird die Wahrscheinlichkeit errechnet, dass er aus einem bestimmten Allel a_i allein durch Sequenzierfehler ϵ hervorgegangen ist:

Model



Für **jeden Read** mit der Sequenz s_r wird die Wahrscheinlichkeit errechnet, dass er aus einem bestimmten Allel a_i allein durch Sequenzierfehler ϵ hervorgegangen ist:

Allel-Likelihood gegeben ein Read

$$Pr(T = s_r \mid S = a_i, \epsilon) = pairHMM_{\epsilon,q_r}(a_i, s_r)$$

Model

Berechnung der Wahrscheinlichkeit, einen bestimmten Read s_r anhand einer gegebenen **Allele-Fraction** $\Theta_i = (\theta_1, \ldots, \theta_n) \in [0, 1]^n$ zu beobachten:

Model

Berechnung der Wahrscheinlichkeit, einen bestimmten Read s_r anhand einer gegebenen **Allele-Fraction** $\Theta_i = (\theta_1, \ldots, \theta_n) \in [0,1]^n$ zu beobachten:

Likelihood einer Allele-Fraction gegeben ein Read

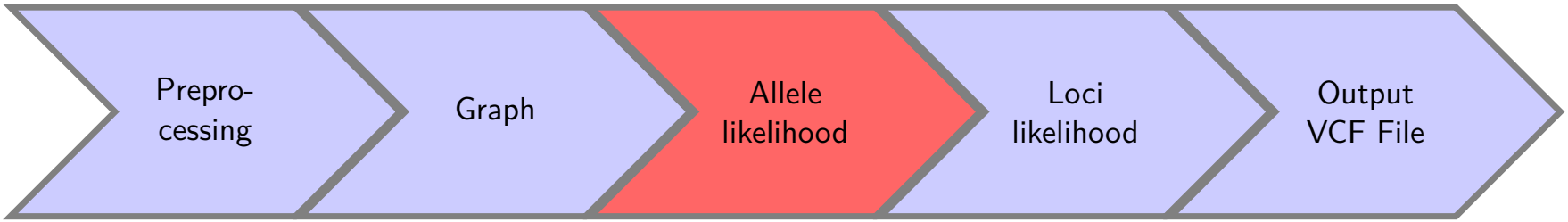
$$Pr(s_r \mid \Theta = \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \theta_i \cdot Pr(T = s_r \mid S = a_i, \epsilon)$$

(es gilt $n = n_{cand}$)

Model

Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Model

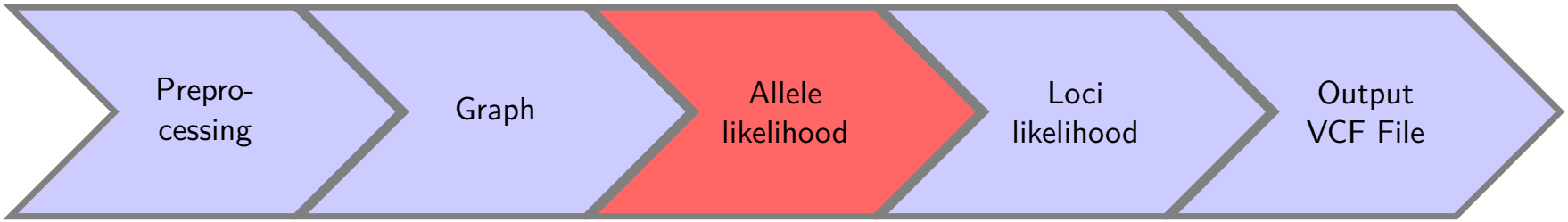


Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

Model



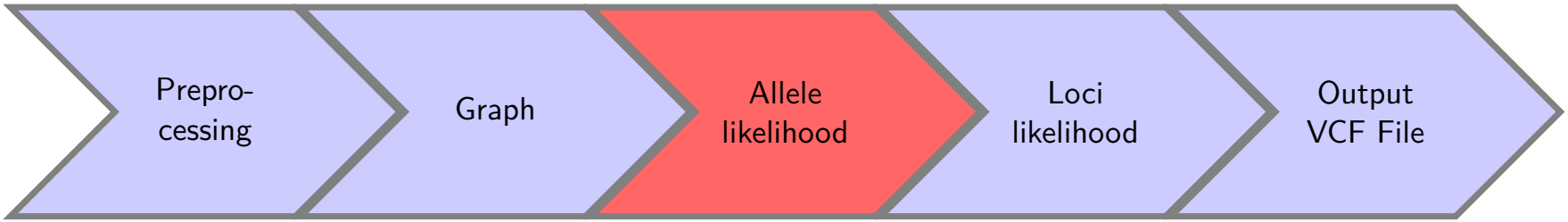
Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

$\Rightarrow$ L ist eine mögliche Loci-Verteilung, die durch die gegebene Allele-Fraction abgebildet wird

Model



Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

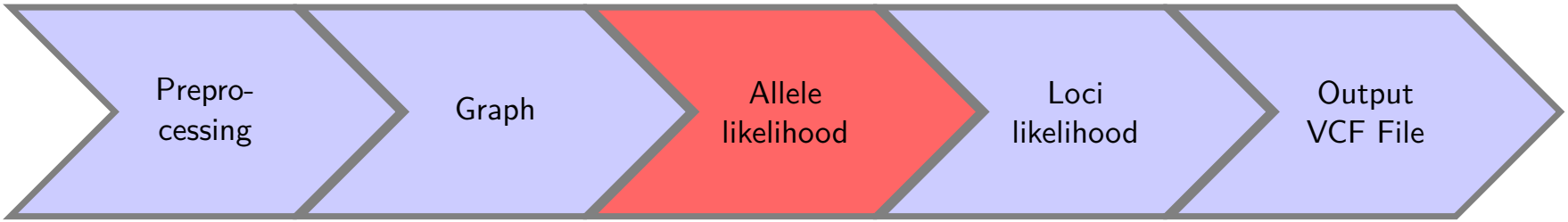
Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \,|\, D) = Pr(D \,|\, \Theta) = \prod_{r=1}^{m} Pr(s_r \,|\, \Theta)$$

$\Rightarrow$ L ist eine mögliche Loci-Verteilung, die durch die gegebene Allele-Fraction abgebildet wird

$\Rightarrow$ die Unsicherheit bei der Zuordnung der Reads wird durch die relativen Häufigkeiten in der Allele-Fraction abgebildet und an die spätere Loci-Zuordnung weitergereicht

Model



Allel-Likelihood gegeben ein Read

$$Pr(T = s_r \mid S = a_i, \epsilon) = pairHMM_{\epsilon,q_r}(a_i, s_r)$$

Likelihood einer Allele-Fraction gegeben ein Read

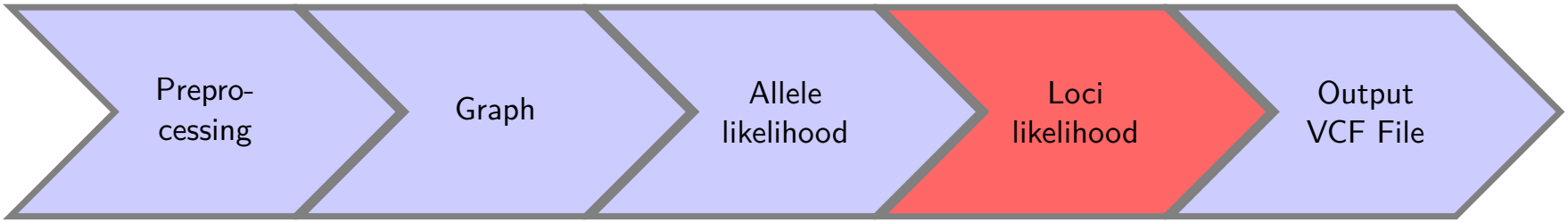
$$Pr(s_r \mid \Theta = \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \theta_i \cdot Pr(T = s_r \mid S = a_i, \epsilon)$$

Likelihood einer Allele-Fraction gegeben alle Reads

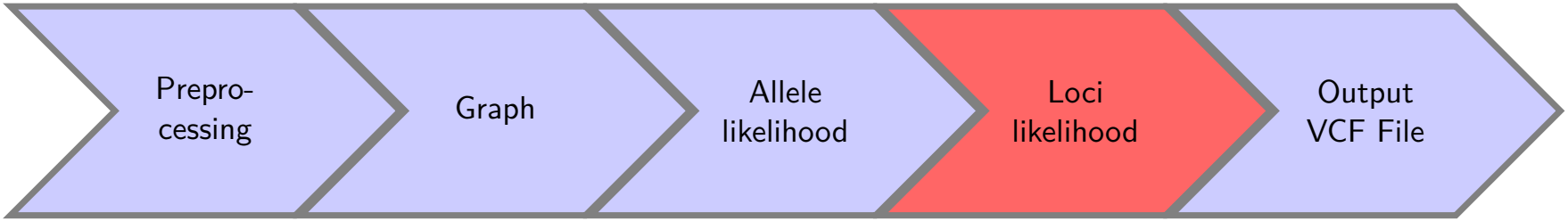
$$L(\Theta = \theta_1, \ldots, \theta_n \mid D = s_1, \ldots, s_m) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

$\Rightarrow$ für die Allele-Fraction mit maximaler Likelihood erfolgt die Loci-Zuordnung

Model

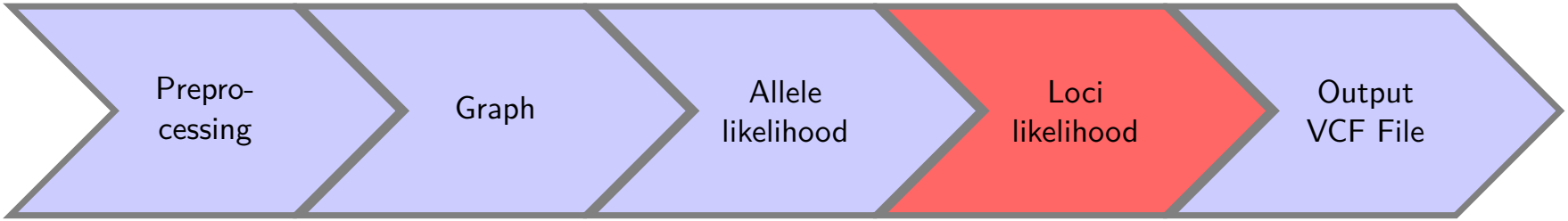

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

Model



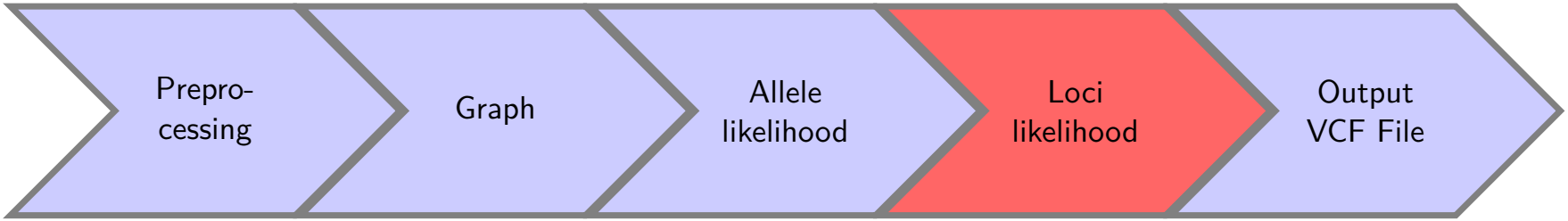
- Die Allel-Fraction mit maximaler Likelihood soll möglichen genomischen Loci zugeordnet werden

Model



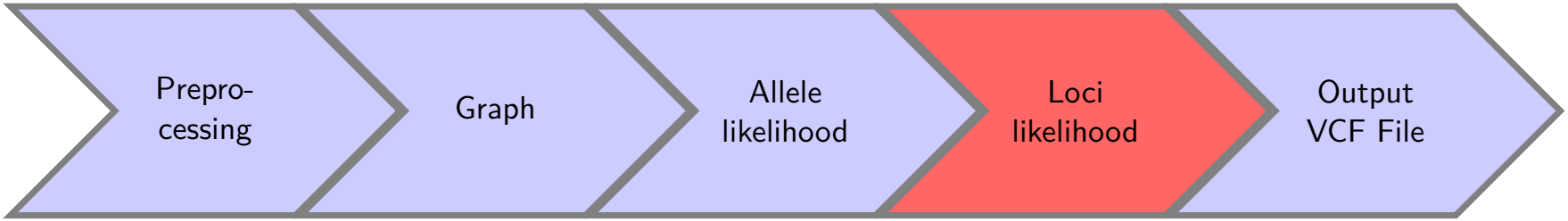
- Die Allel-Fraction mit maximaler Likelihood soll möglichen genomischen Loci zugeordnet werden
- in Abhängigkeit von Ploidie und Anzahl der Kandidatenallele können auch mehrere Loci in einer Zusammenhangskomponente vorkommen

Model



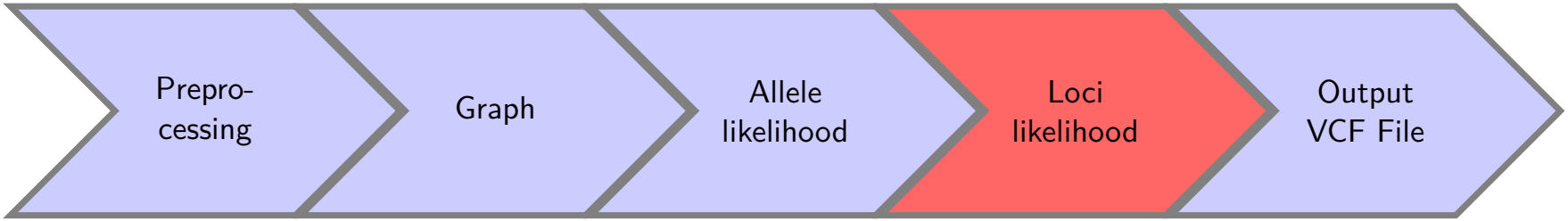
- Die Allel-Fraction mit maximaler Likelihood soll möglichen genomischen Loci zugeordnet werden
- in Abhängigkeit von Ploidie und Anzahl der Kandidatenallele können auch mehrere Loci in einer Zusammenhangskomponente vorkommen
- für alle Allelkombinationen müssen die möglichen Loci-Kombinationen der in ihnen enthaltenen Kandidatenallele gebildet werden

Model



$\Rightarrow$ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Model

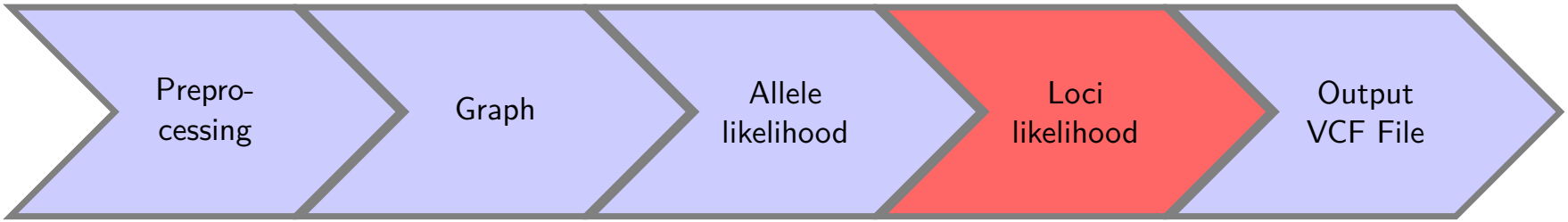


⇒ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Kombinationen der Allele:

[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]

Model

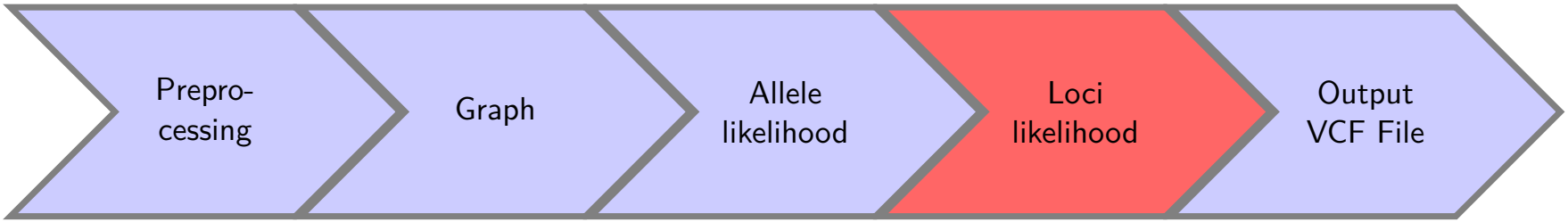


⇒ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Permutationen der Allele:

(1, 2, 1, 1), (2, 1, 0, 0), (2, 1, 1, 1), (0, 1, 2, 1), (0, 1, 1, 2), (0, 1, 0, 0), (2, 2, 1, 0), (0, 2, 2, 1), (2, 2, 0, 1), (1, 0, 2, 2), (0, 2, 0, 1), (2, 0, 0, 1), (1, 0, 1, 0), (0, 2, 1, 2), (0, 0, 2, 0), (2, 2, 2, 1), (1, 1, 0, 1), (2, 0, 1, 1), (2, 0, 2, 0), (0, 0, 2, 2), (1, 1, 2, 0), (1, 2, 1, 0), (2, 0, 2, 2), (2, 1, 1, 0), (2, 1, 0, 2), (1, 2, 0, 1), (0, 1, 2, 0), (1, 2, 1, 2), (1, 2, 2, 1), (0, 1, 1, 1), (1, 1, 1, 0), (0, 0, 0, 0), (2, 1, 1, 2), (2, 1, 2, 1), (1, 0, 0, 1), (0, 1, 0, 2), (2, 2, 1, 2), (0, 2, 2, 0), (1, 0, 2, 1), (2, 0, 0, 0), (0, 2, 1, 1), (1, 1, 1, 2), (0, 0, 0, 2), (0, 0, 1, 1), (1, 0, 1, 2), (2, 0, 0, 2), (0, 0, 2, 1), (1, 1, 2, 2), (2, 1, 0, 1), (1, 2, 0, 0), (0, 1, 2, 2), (1, 2, 2, 0), (0, 1, 1, 0), (2, 2, 0, 0), (0, 2, 0, 0), (2, 1, 2, 0),...

Model

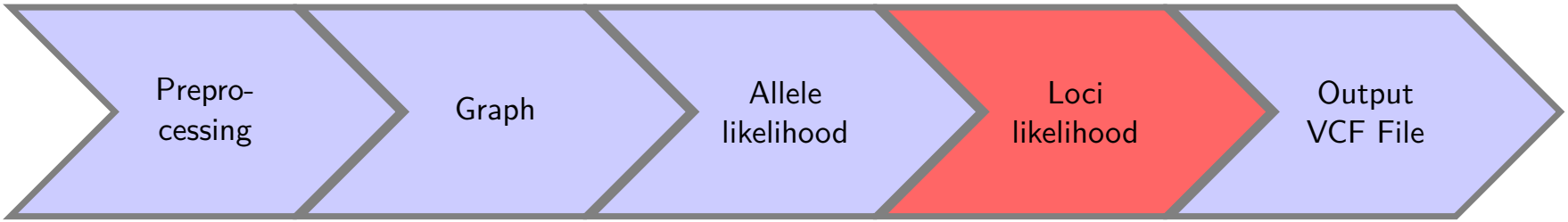


⇒ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Mögliche Loci-Kombinationen:

((0, 0), (0, 2)), ((1, 1), (1, 1)), ((0, 2), (0, 2)), ((1, 1), (2, 2)), ((0, 1), (0, 2)), ((1, 1), (1, 2)), ((1, 2), (2, 2)), ((1, 2), (1, 2)), ((0, 0), (1, 1)), ((0, 0), (2, 2)), ((0, 2), (1, 1)), ((0, 1), (1, 1)), ((0, 0), (0, 0)), ((0, 2), (2, 2)), ((0, 0), (1, 2)), ((0, 1), (2, 2)), ((2, 2), (2, 2)), ((0, 0), (0, 1)), ((0, 2), (1, 2)), ((0, 1), (1, 2)), ((0, 1), (0, 1))

Model



⇒ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Kombinationen der Allele:

[..., (0, 1, 1, 2), (0, 1, 0, 0),...]

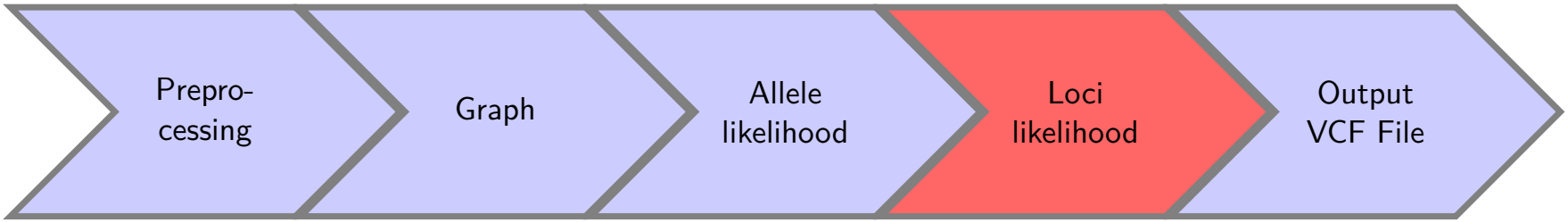
Permutationen der Allele:

(0, 1, 1, 2), (0, 1, 2, 1), (1, 0, 2, 1), (1, 1, 2, 0), (1, 1, 0, 2), (2, 1, 1, 0),...

Mögliche Loci-Kombinationen:

((0, 1), (1, 2)), ((0, 1), (2, 1)), ((1, 0), (2, 1)), ((1, 1), (2, 0)), ((1, 1), (0, 2)), ((2, 1), (1, 0)),...

Model



⇒ Beispiel: $ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

Kombinationen der Allele:

[..., (0, 1, 1, 2), (0, 1, 0, 0),...]

Permutationen der Allele:

(0, 1, 1, 2), (0, 1, 2, 1), (1, 0, 2, 1), (1, 1, 2, 0), (1, 1, 0, 2), (2, 1, 1, 0),...

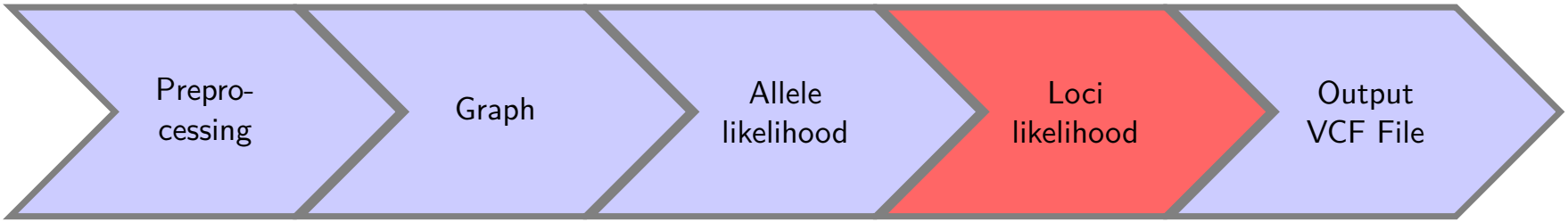
Mögliche Loci-Kombinationen:

((0, 1), (1, 2)), ((0, 1), (2, 1)), ((1, 0), (2, 1)), ((1, 1), (2, 0)), ((1, 1), (0, 2)), ((2, 1), (1, 0)),...

Mögliche Loci-Kombinationen:

((0, 1), (1, 2)), ((0, 1), (2, 1)), ((1, 0), (2, 1)),((1, 1), (2, 0)), ((1,

Model



Allelkombinationen:

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$:

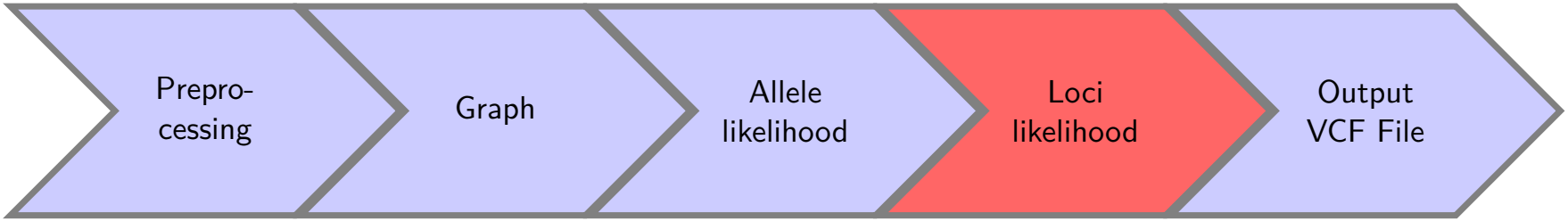
[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]

Für **jeden Read** mit der Sequenz s_r wird die Wahrscheinlichkeit errechnet, dass er aus einem bestimmten Allel a_i allein durch Sequenzierfehler ϵ hervorgegangen ist:

Allel-Likelihood gegeben ein Read

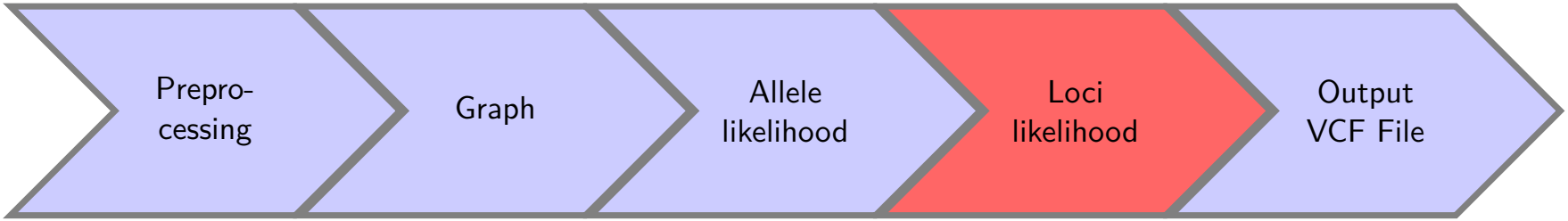
$$Pr(T = s_r \mid S = a_i, \epsilon) = pairHMM_{\epsilon,q_r}(a_i, s_r)$$

Model



Berechnung der Wahrscheinlichkeit, einen bestimmten Read s_r anhand einer gegebenen **Allele-Fraction** $\Theta_i = (\theta_1, \ldots, \theta_n) \in [0, 1]^n$ zu beobachten:

Model



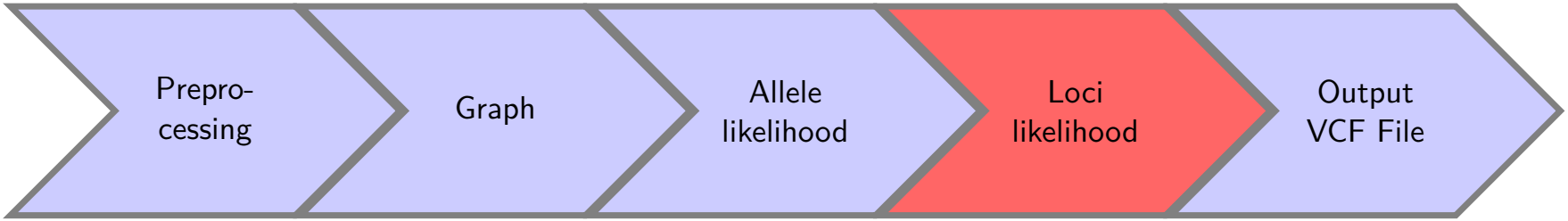
Berechnung der Wahrscheinlichkeit, einen bestimmten Read s_r anhand einer gegebenen **Allele-Fraction** $\Theta_i = (\theta_1, \ldots, \theta_n) \in [0,1]^n$ zu beobachten:

Likelihood einer Allele-Fraction gegeben ein Read

$$Pr(s_r \mid \Theta = \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \theta_i \cdot Pr(T = s_r \mid S = a_i, \epsilon)$$

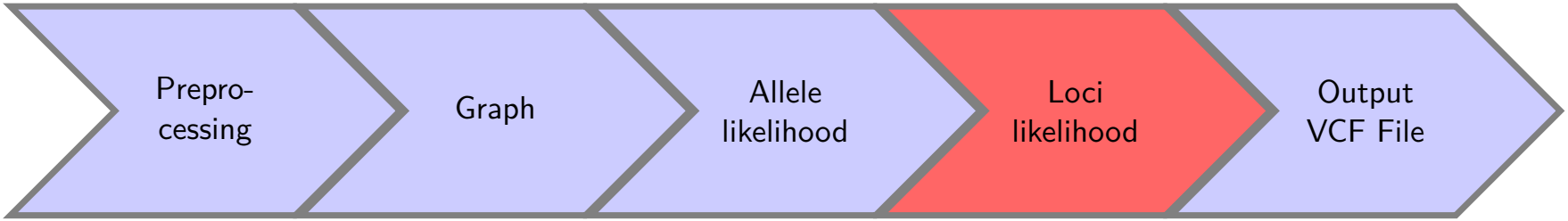
(es gilt $n = n_{cand}$)

Model



Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Model

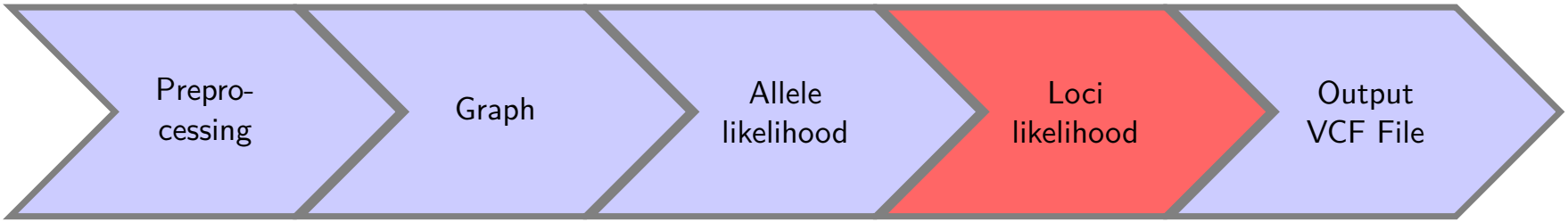


Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

Model



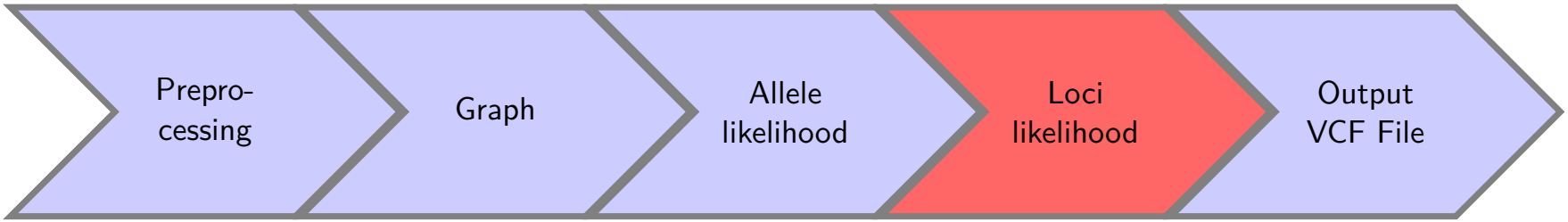
Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

$\Rightarrow$ L ist eine mögliche Loci-Verteilung, die durch die gegebene Allele-Fraction abgebildet wird

Model



Bestimmung der resultierende Likelihood einer Allele-Fraction in Zusammenschau mit **allen Reads** $D = (s_1, \ldots, s_m) \in \{A, C, G, T\}^{k^m}$:

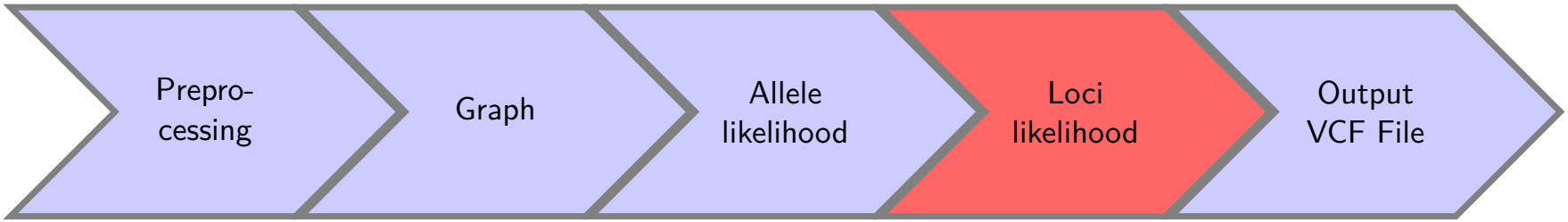
Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

$\Rightarrow$ L ist eine mögliche Loci-Verteilung, die durch die gegebene Allele-Fraction abgebildet wird

$\Rightarrow$ die Unsicherheit bei der Zuordnung der Reads wird durch die relativen Häufigkeiten in der Allele-Fraction abgebildet und an die spätere Loci-Zuordnung weitergereicht

Model



Allel-Likelihood gegeben ein Read

$$Pr(T = s_r \mid S = a_i, \epsilon) = pairHMM_{\epsilon,q_r}(a_i, s_r)$$

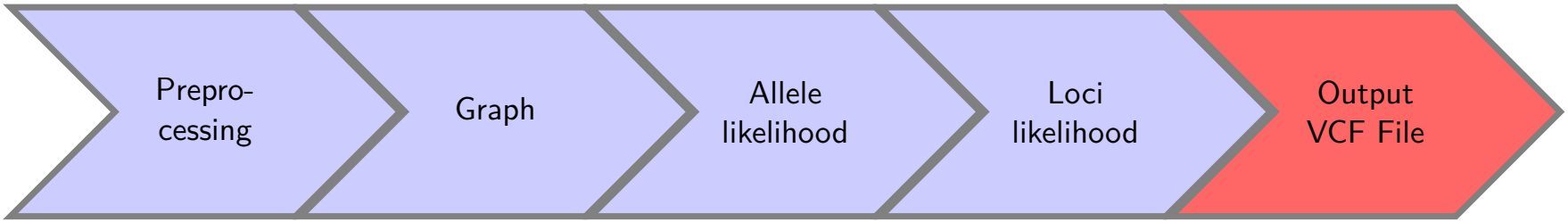
Likelihood einer Allele-Fraction gegeben ein Read

$$Pr(s_r \mid \Theta = \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \theta_i \cdot Pr(T = s_r \mid S = a_i, \epsilon)$$

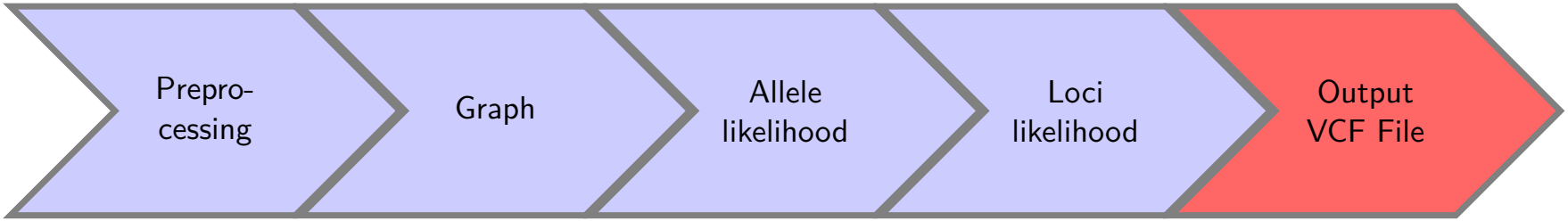
Likelihood einer Allele-Fraction gegeben alle Reads

$$L(\Theta = \theta_1, \ldots, \theta_n \mid D = s_1, \ldots, s_m) = Pr(D \mid \Theta) = \prod_{r=1}^{m} Pr(s_r \mid \Theta)$$

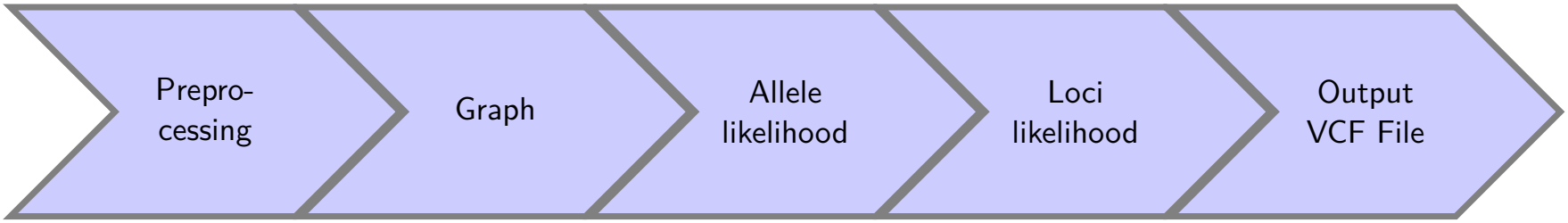
Model

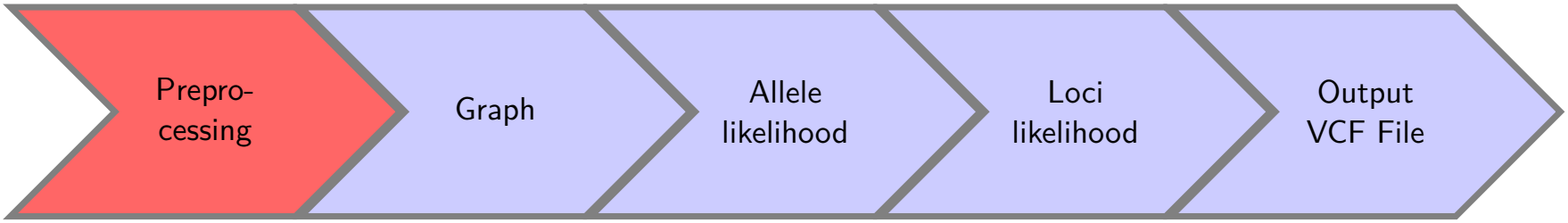

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

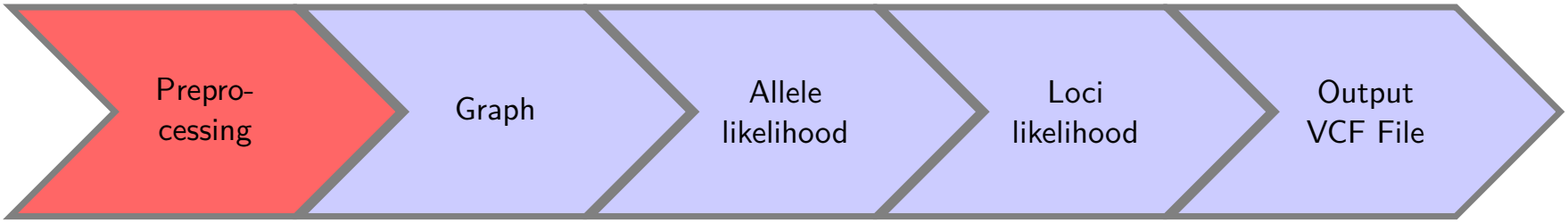
Model

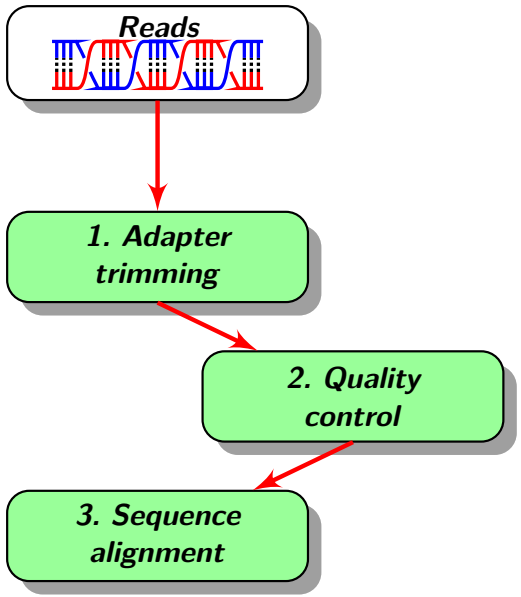

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

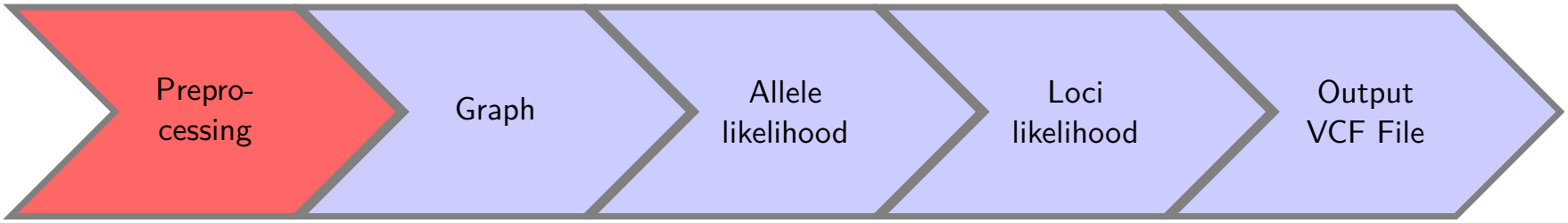
- vcf

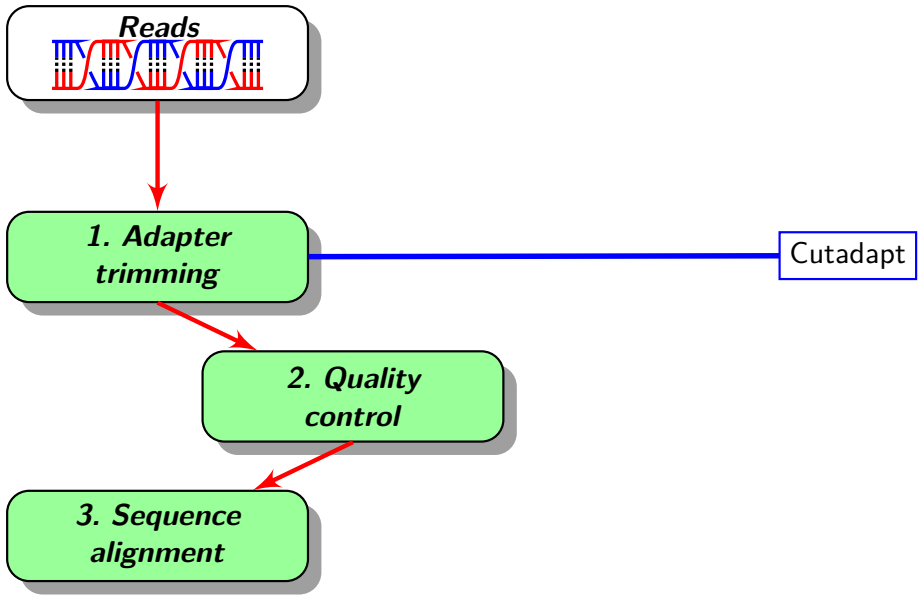

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

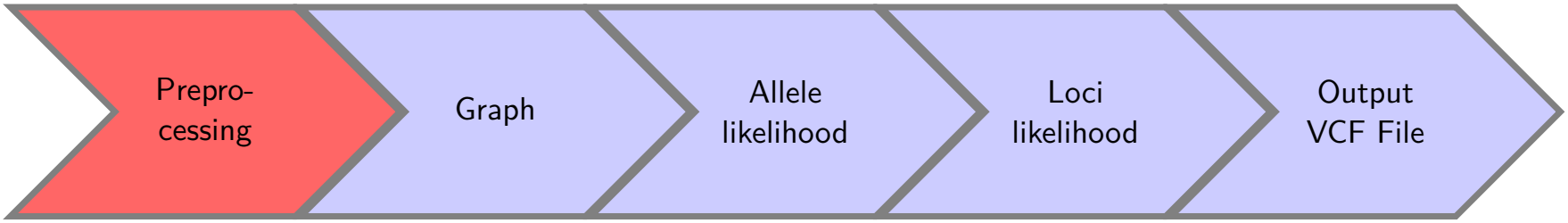

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

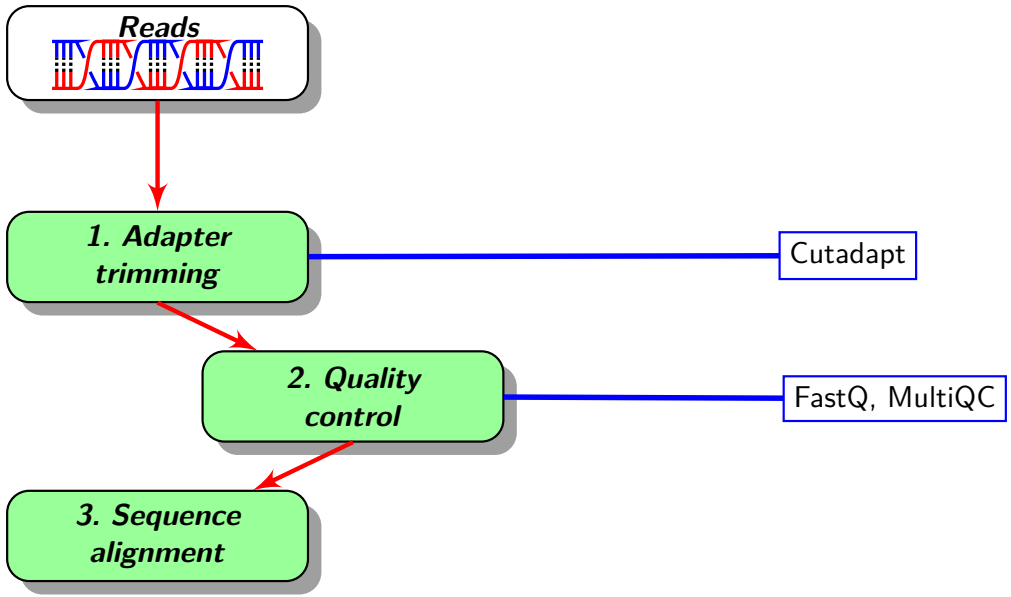

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

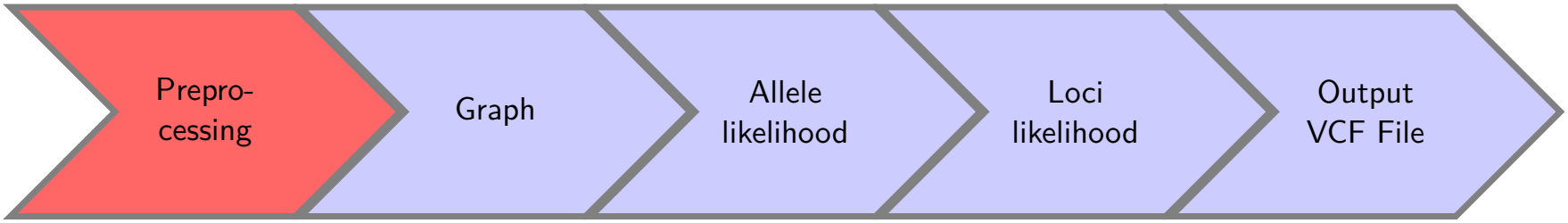

Reads
1. Adapter
trimming
2. Quality
control
3. Sequence
alignment

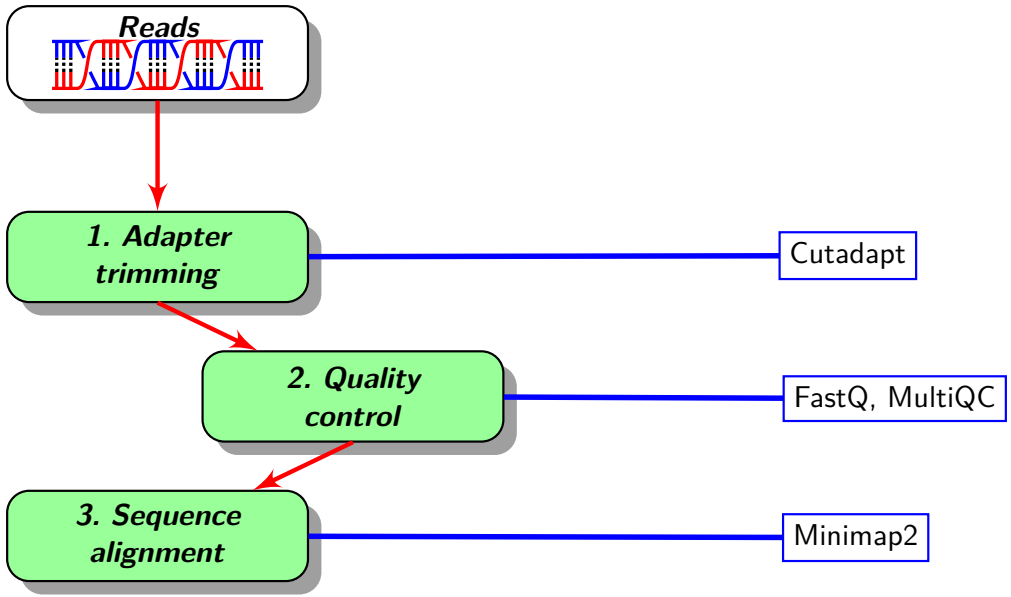

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

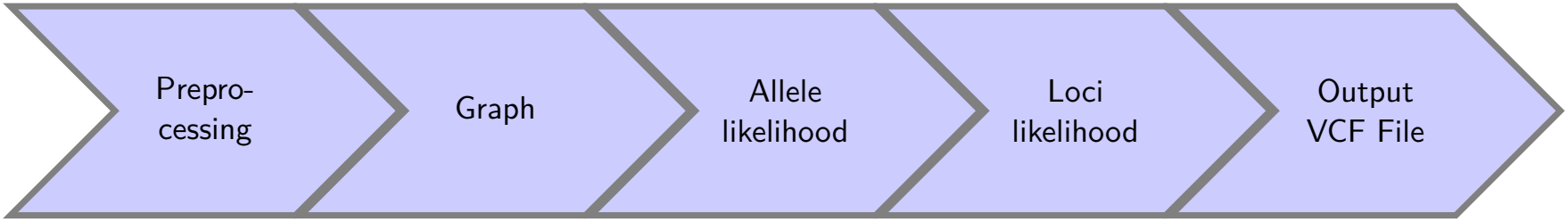

Reads
1. Adapter
trimming
Cutadapt
2. Quality
control
3. Sequence
alignment

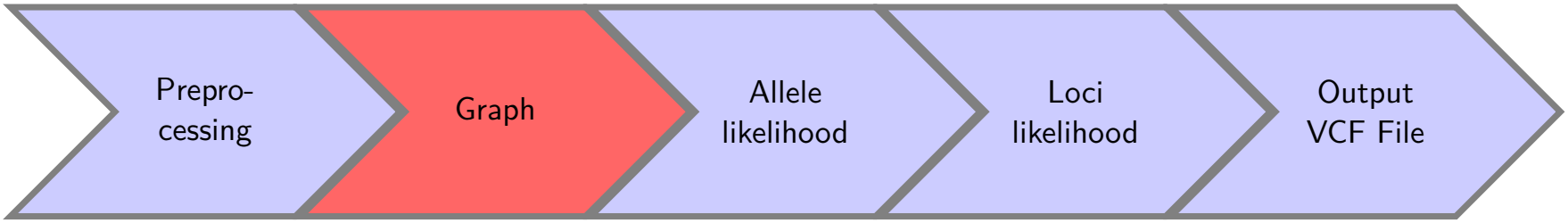

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

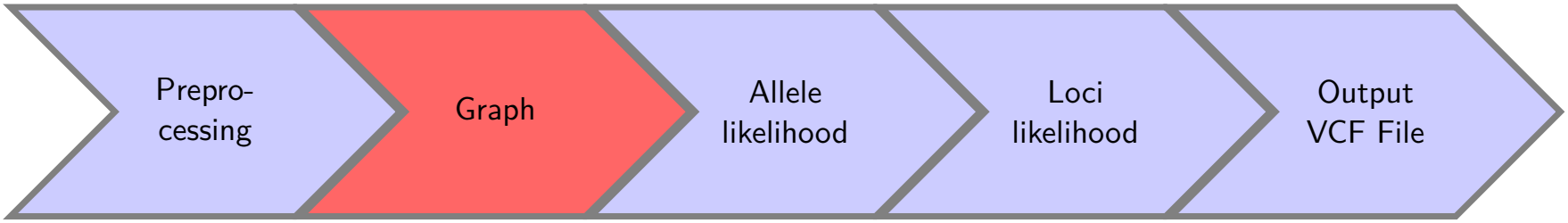

Reads
1. Adapter trimming
Cutadapt
2. Quality control
FastQ, MultiQC
3. Sequence alignment

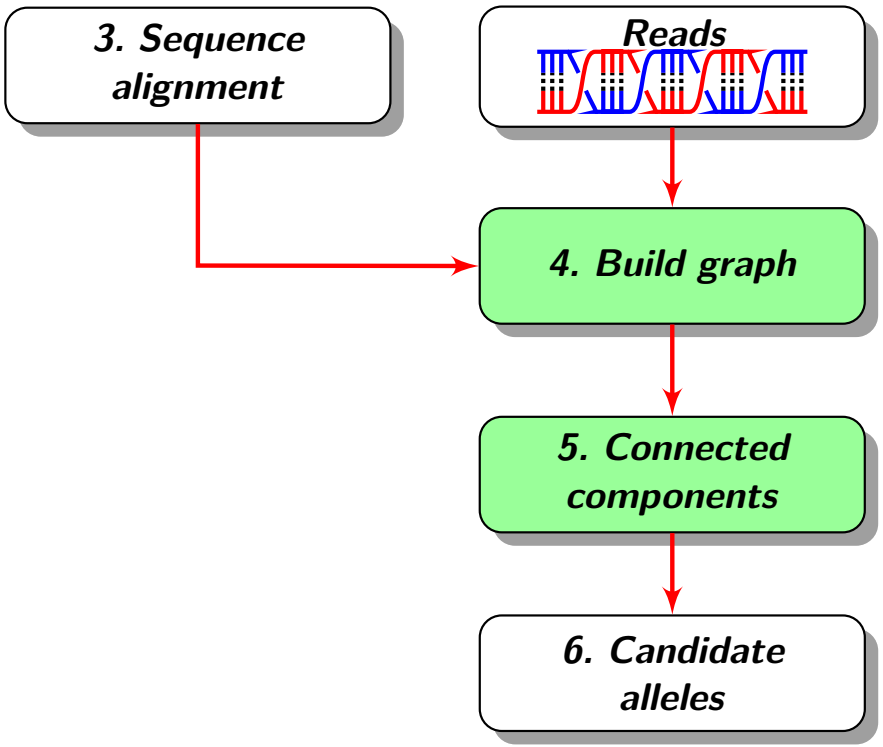

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

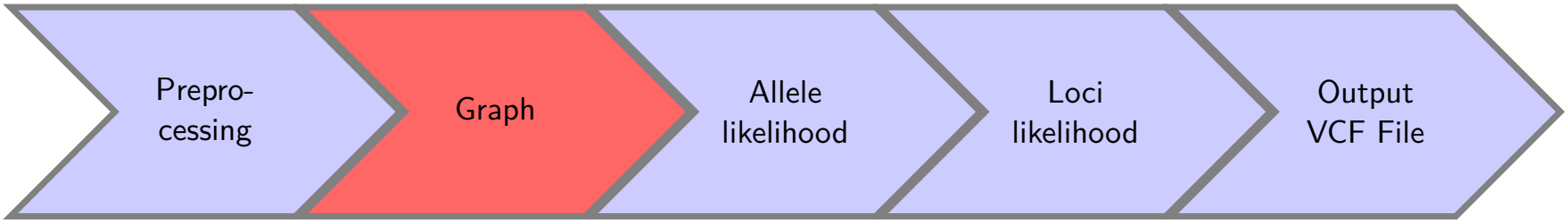

Reads
1. Adapter trimming
Cutadapt
2. Quality control
FastQ, MultiQC
3. Sequence alignment
Minimap2

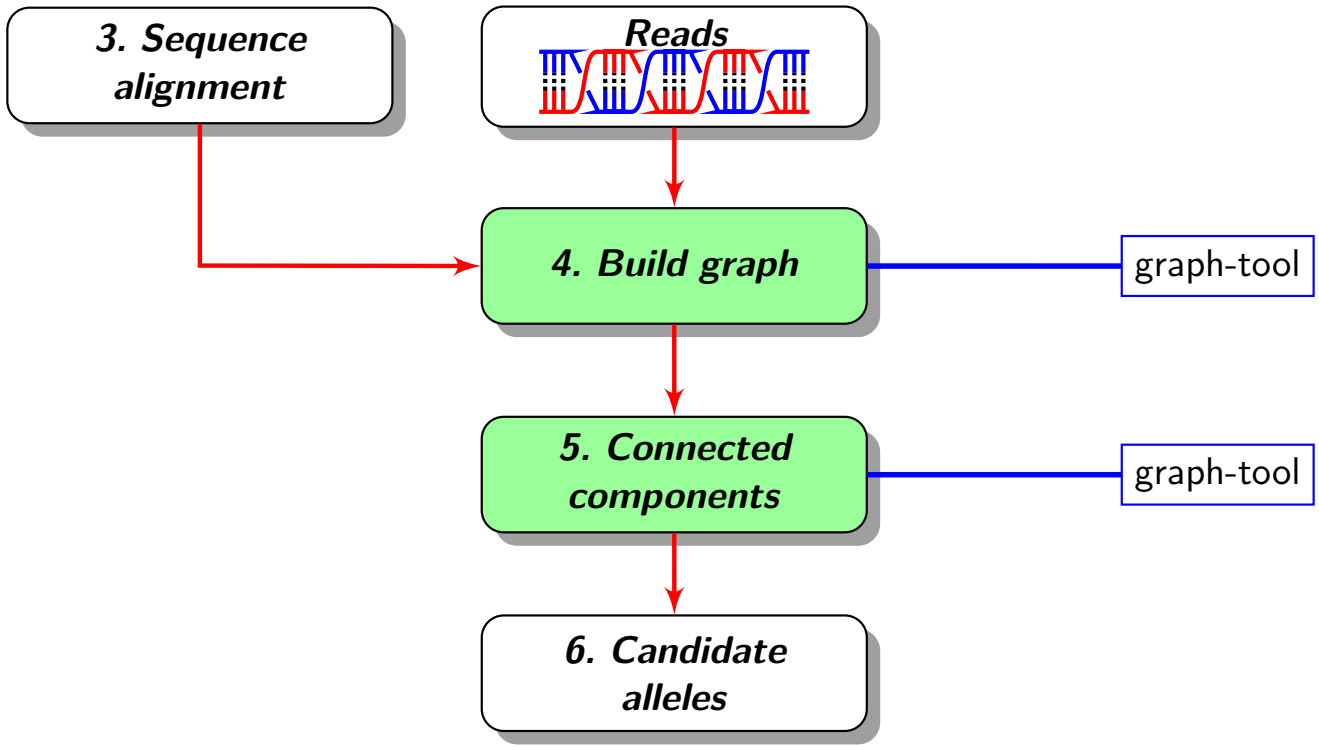

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

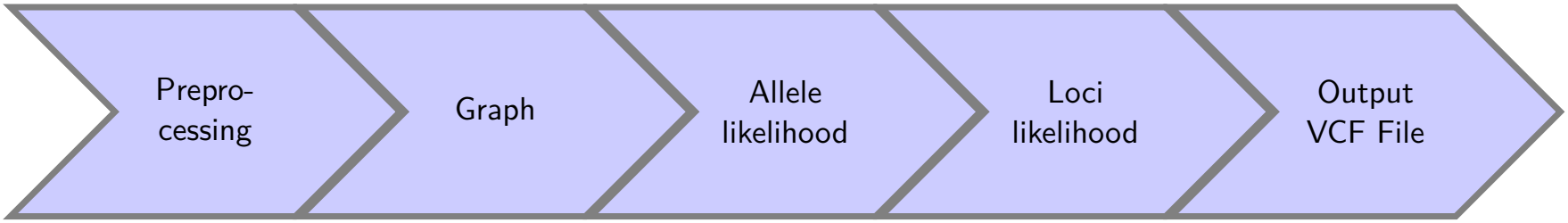

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


3. Sequence alignment
Reads
4. Build graph
5. Connected components
6. Candidate alleles

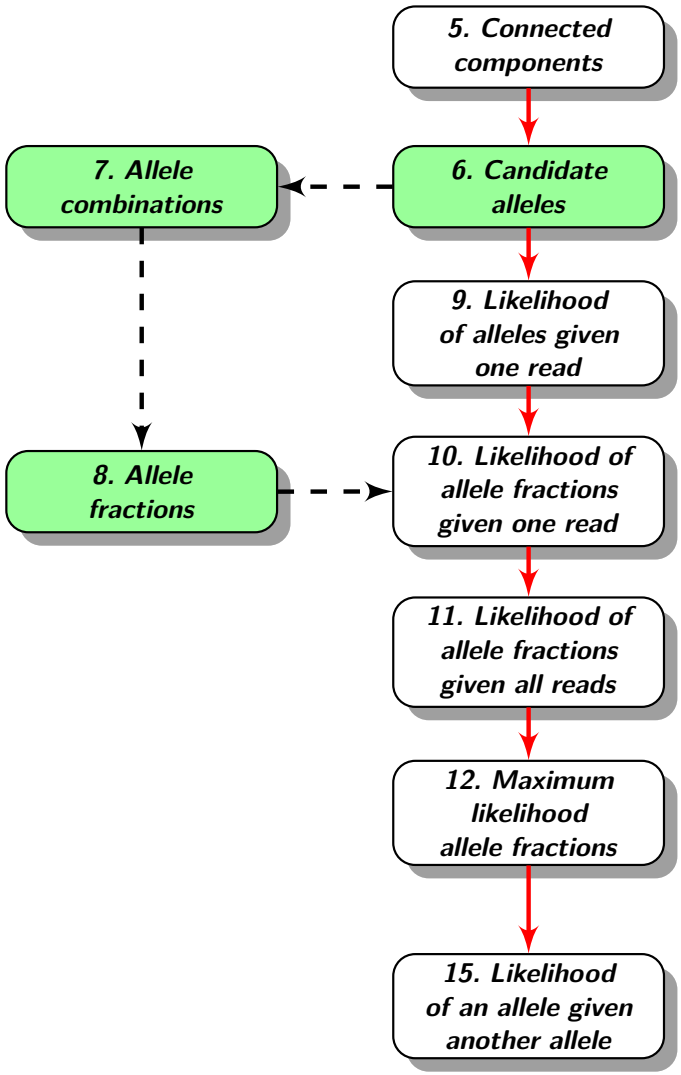

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


3. Sequence alignment
Reads
4. Build graph
graph-tool
5. Connected components
graph-tool
6. Candidate alleles

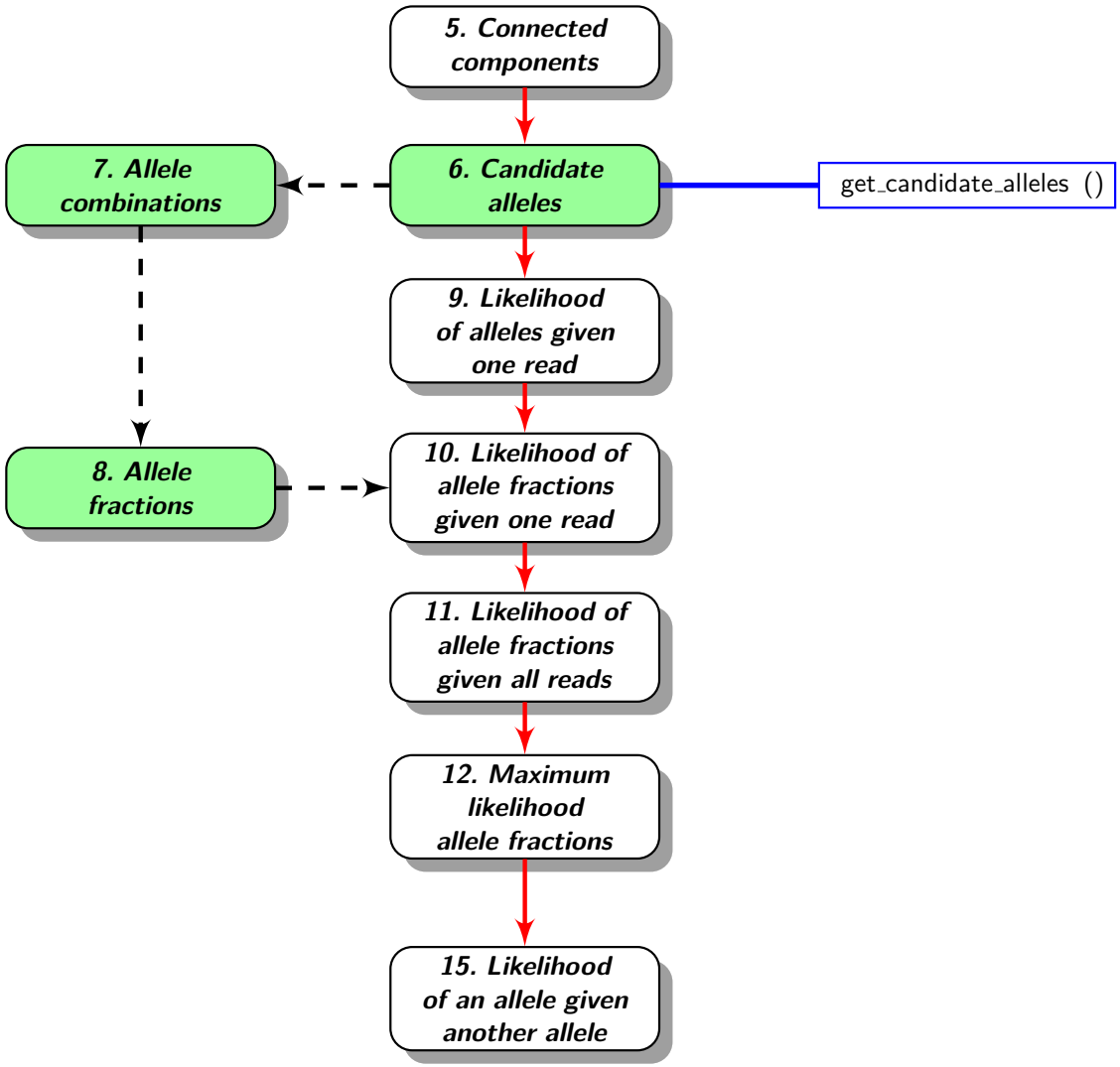

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

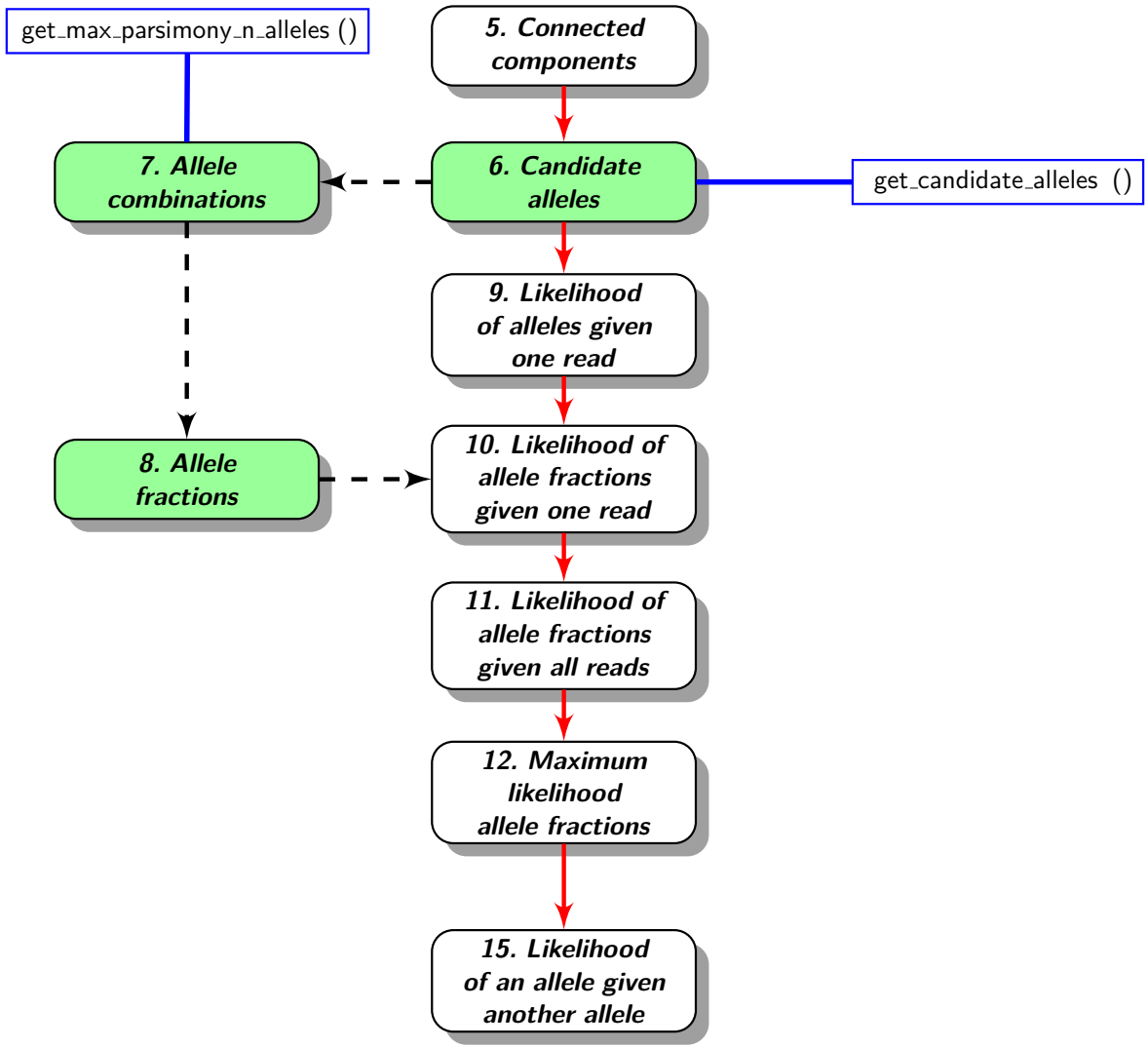

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

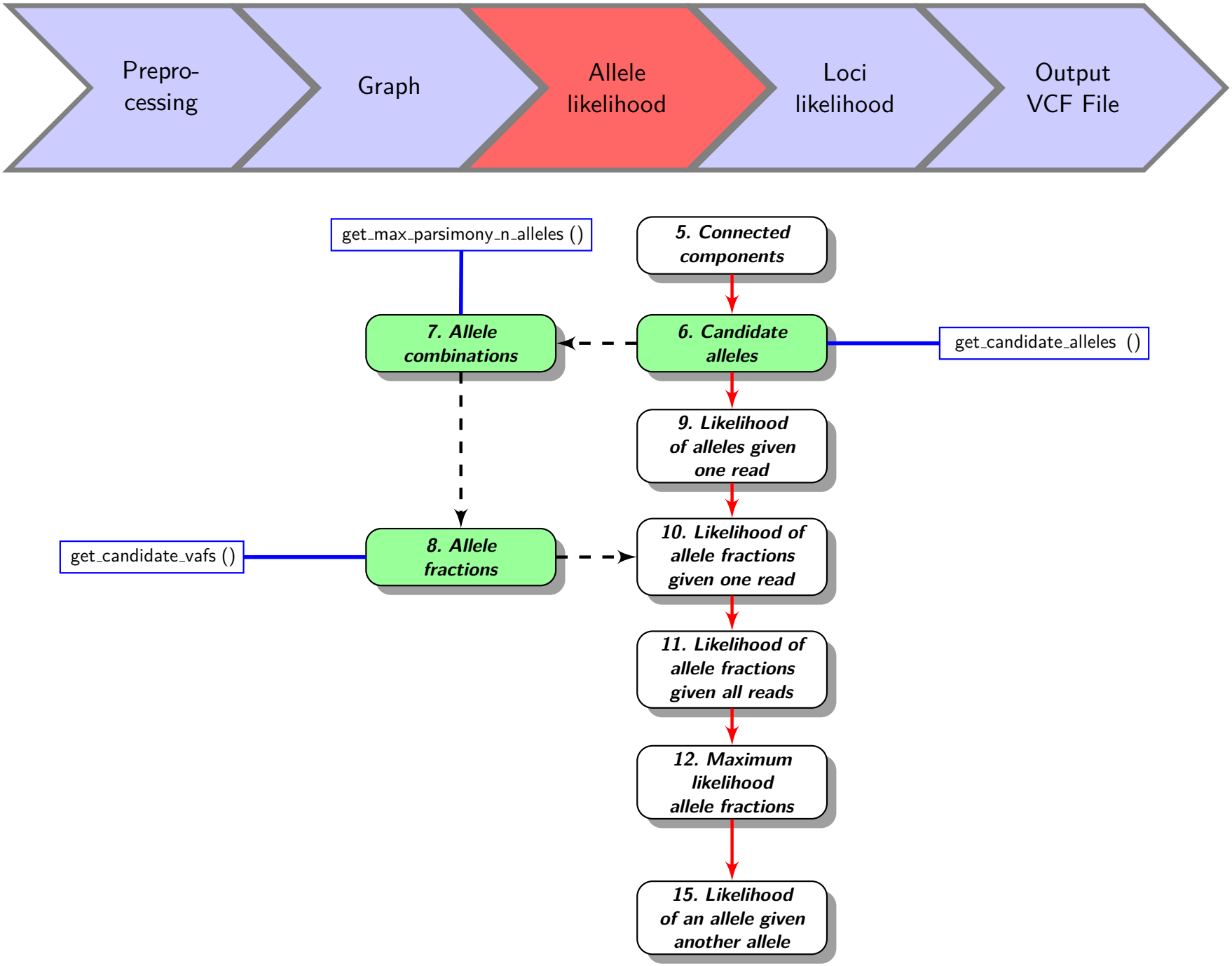

5. Connected components
6. Candidate alleles
7. Allele combinations
8. Allele fractions
9. Likelihood of alleles given one read
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


5. Connected components
6. Candidate alleles
get_candidate_alleles ()
7. Allele combinations
8. Allele fractions
9. Likelihood of alleles given one read
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele

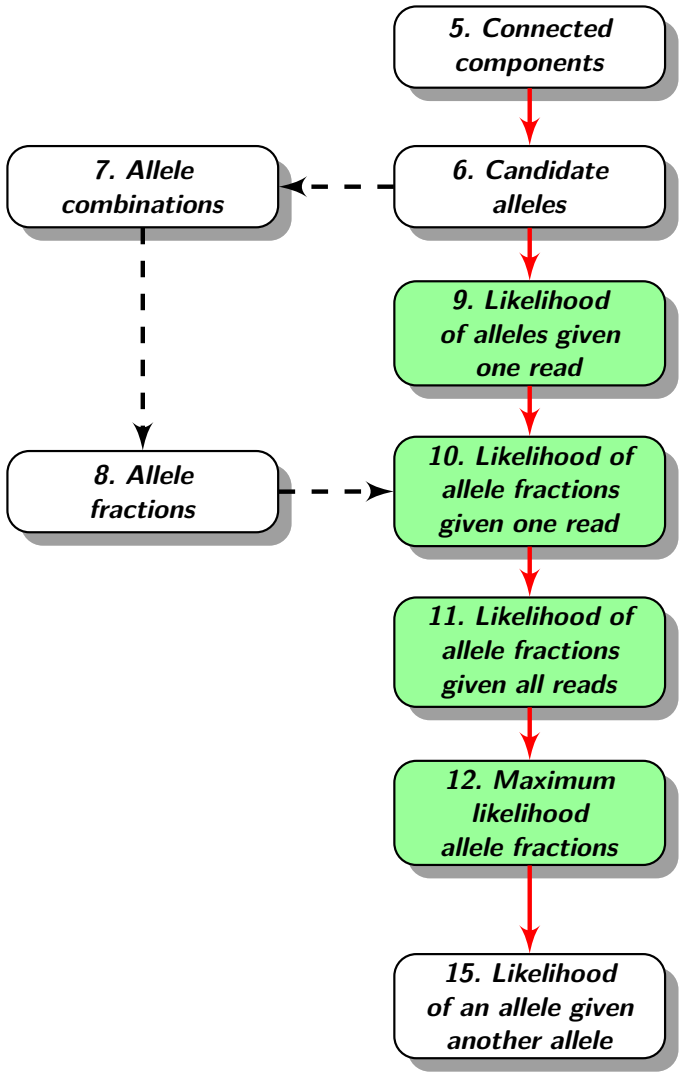

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


get_max_parsimony_n_alleles ()
5. Connected components
7. Allele combinations
6. Candidate alleles
get_candidate_alleles ()
9. Likelihood of alleles given one read
8. Allele fractions
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele

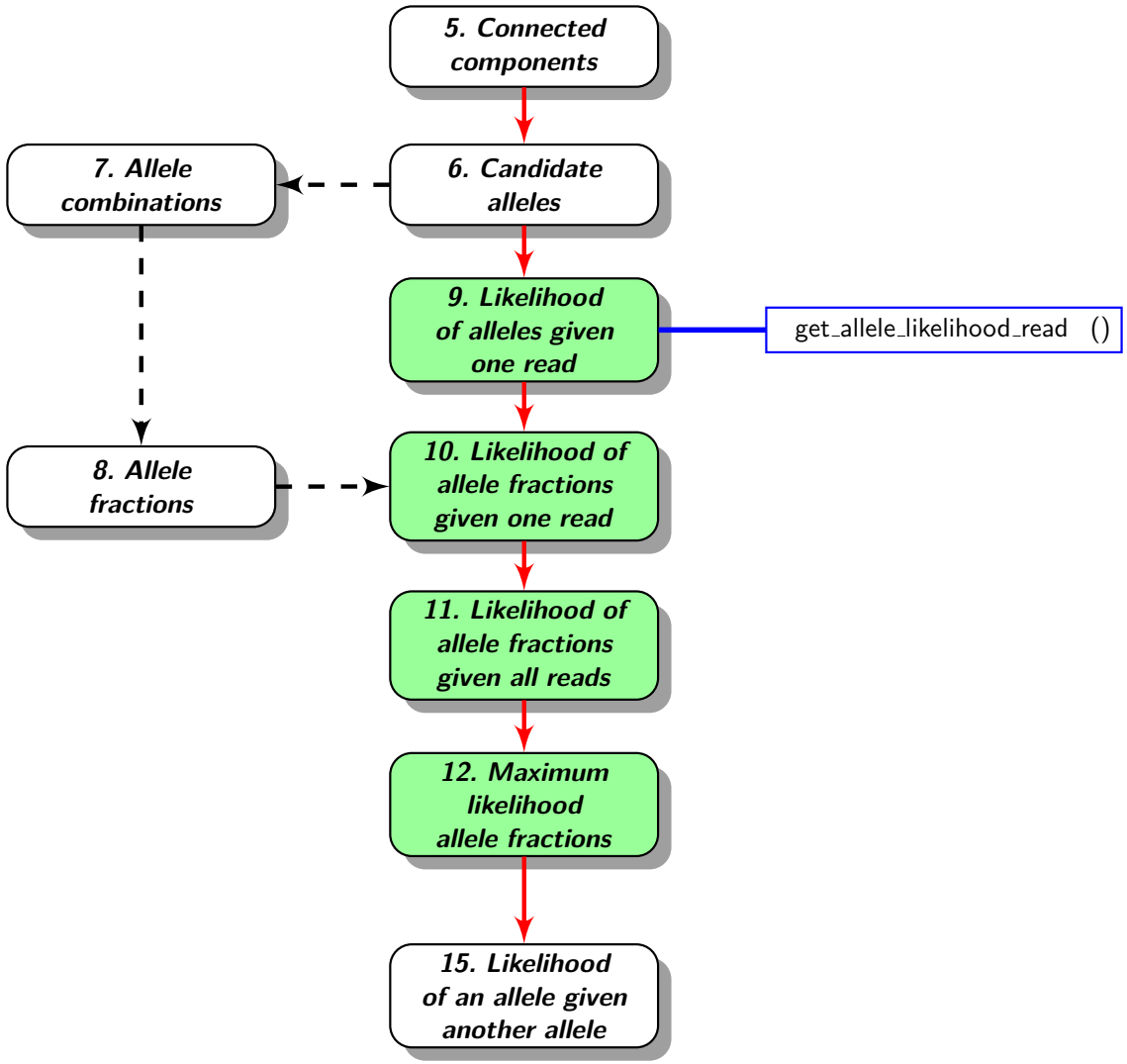

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File
get_max_parsimony_n_alleles ()
5. Connected components
7. Allele combinations
6. Candidate alleles
get_candidate_alleles ()
9. Likelihood of alleles given one read
get_candidate_vafs ()
8. Allele fractions
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

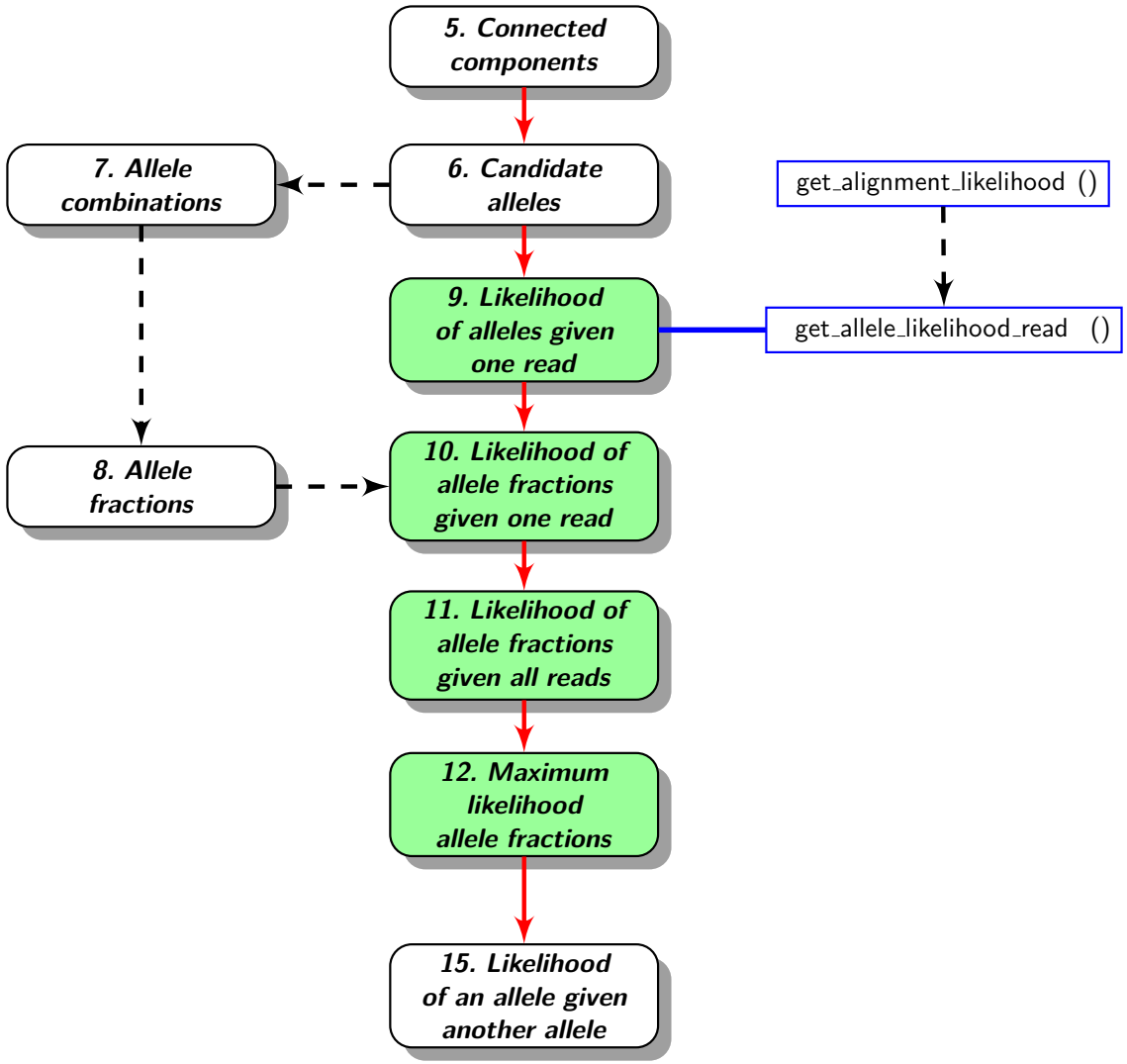

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


5. Connected components
7. Allele combinations
6. Candidate alleles
9. Likelihood of alleles given one read
8. Allele fractions
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele

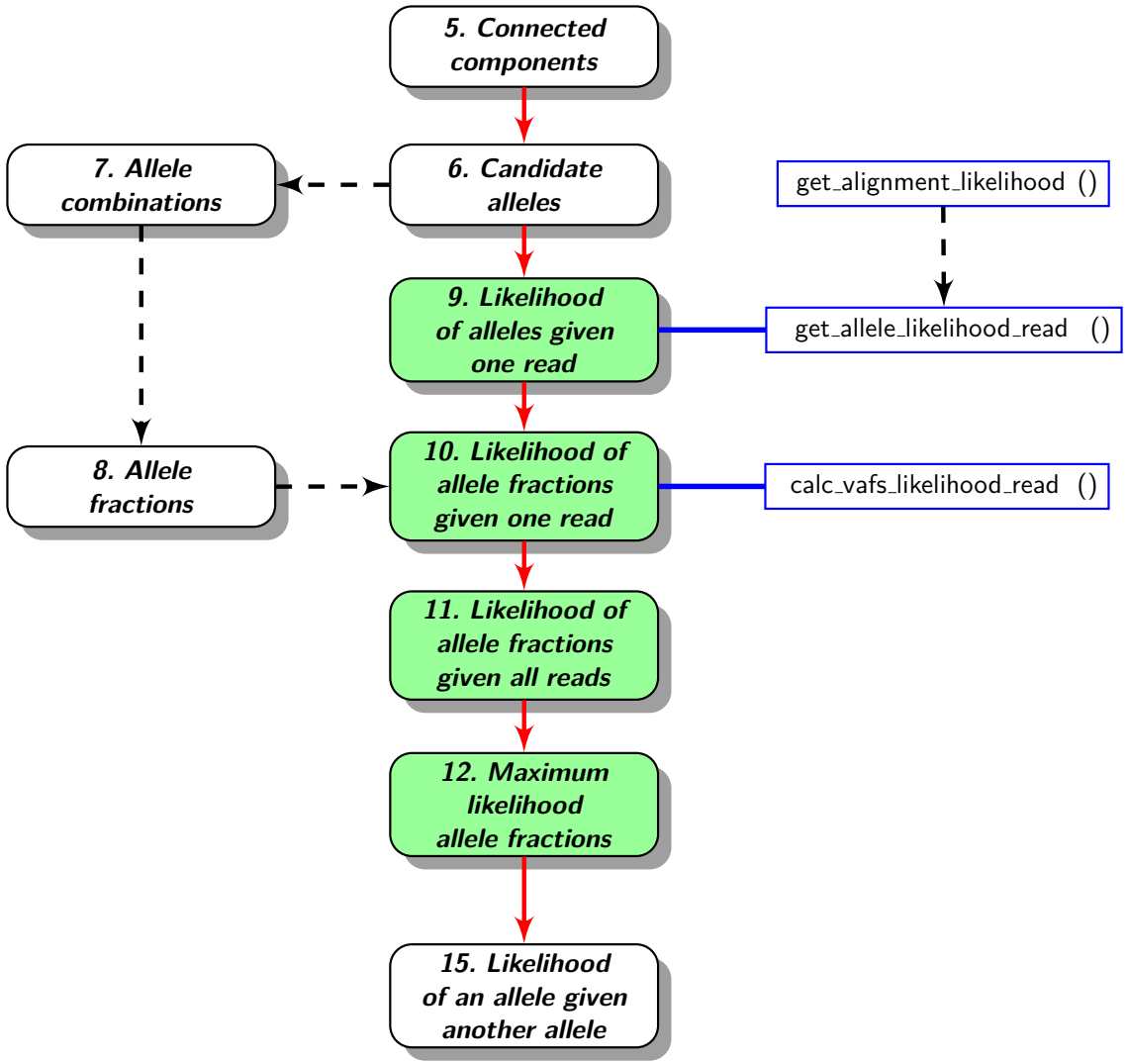

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


5. Connected components
7. Allele combinations
6. Candidate alleles
9. Likelihood of alleles given one read
get_allele_likelihood_read ()
8. Allele fractions
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele

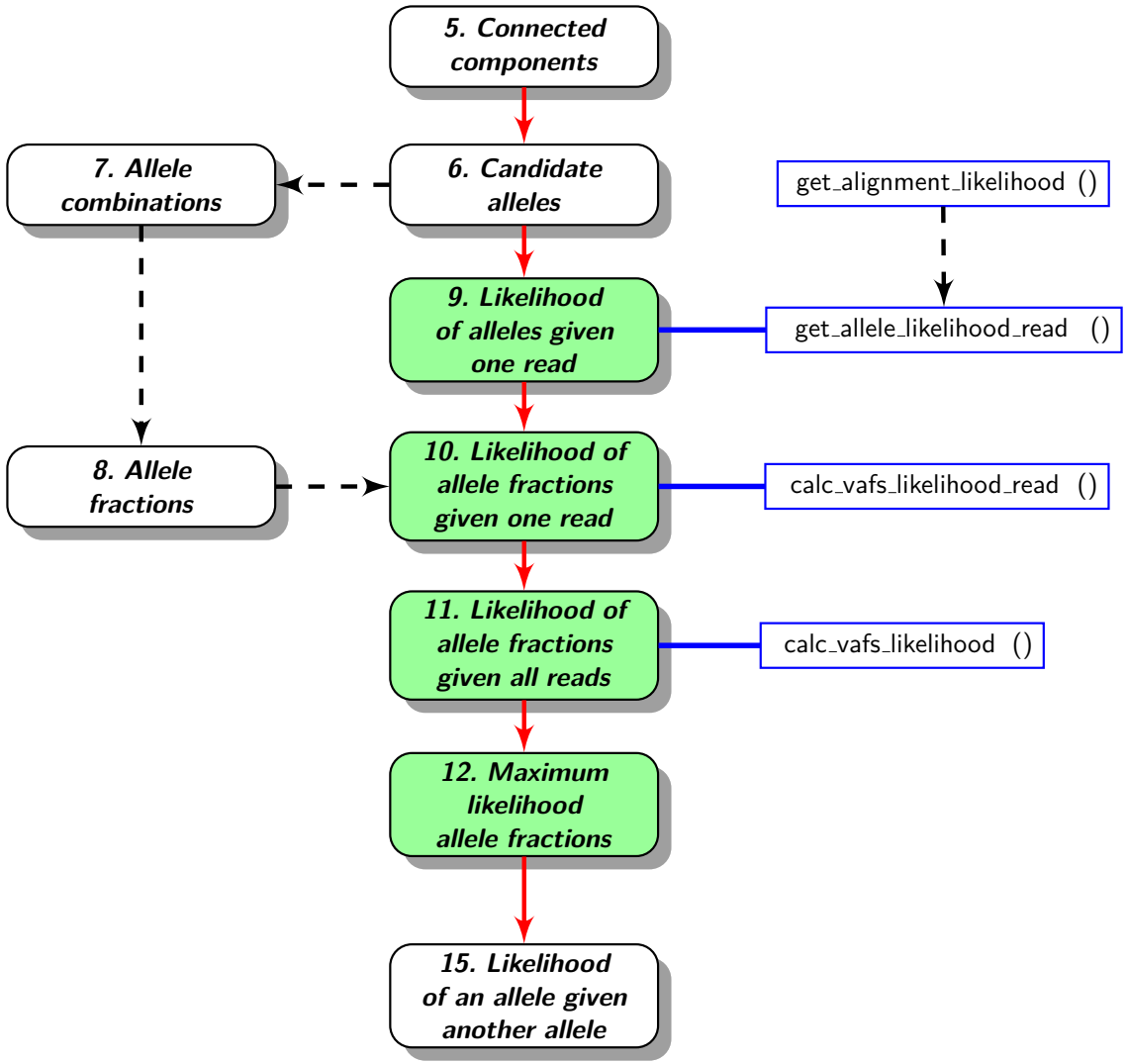

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

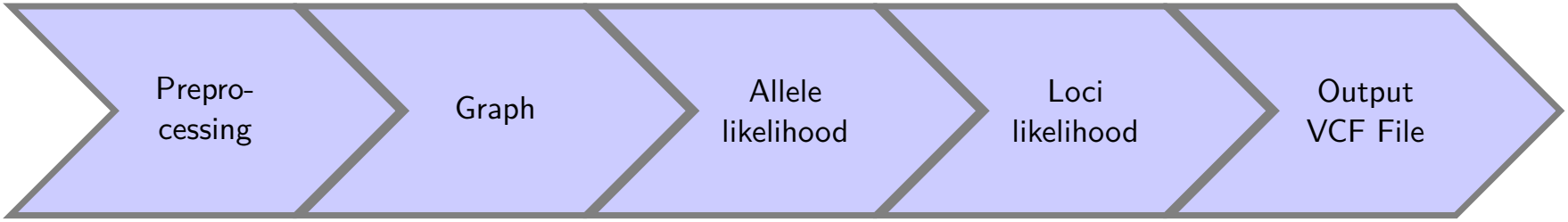

5. Connected components
6. Candidate alleles
7. Allele combinations
8. Allele fractions
9. Likelihood of alleles given one read
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele
get_alignment_likelihood ()
get_allele_likelihood_read ()

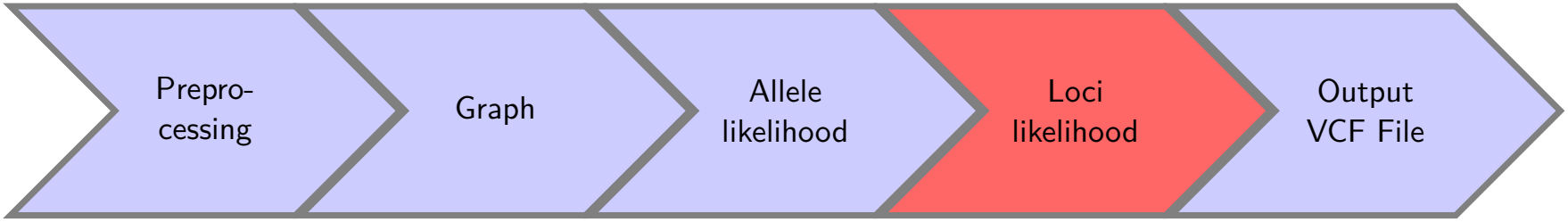

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

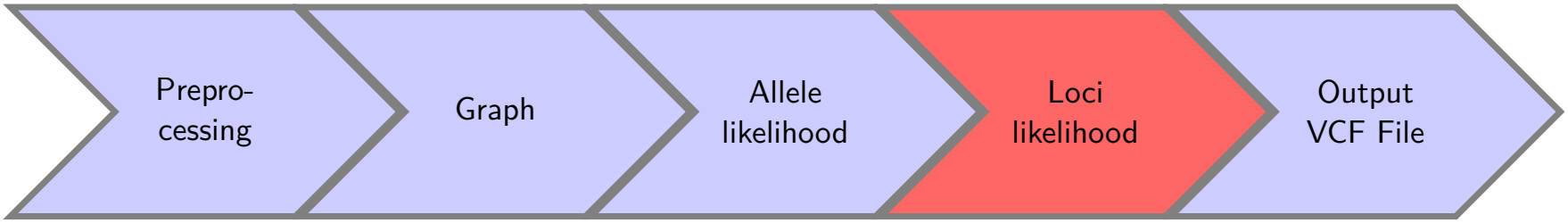

5. Connected components
6. Candidate alleles
7. Allele combinations
8. Allele fractions
9. Likelihood of alleles given one read
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele
get_alignment_likelihood ()
get_allele_likelihood_read ()
calc_vafs_likelihood_read ()

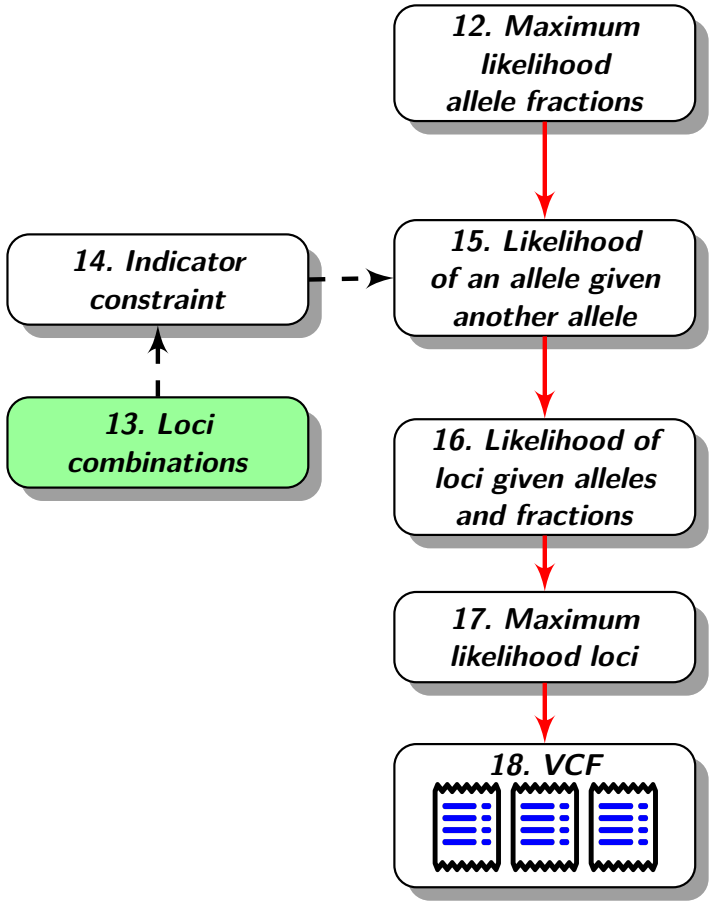

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

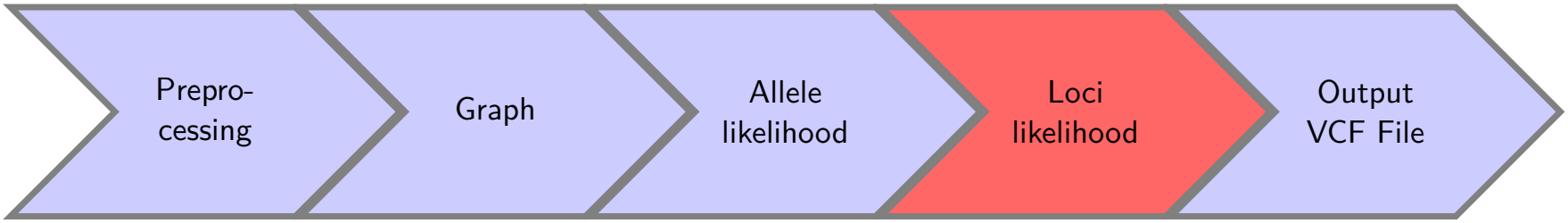

5. Connected components
6. Candidate alleles
7. Allele combinations
8. Allele fractions
9. Likelihood of alleles given one read
10. Likelihood of allele fractions given one read
11. Likelihood of allele fractions given all reads
12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele
get_alignment_likelihood ()
get_allele_likelihood_read ()
calc_vafs_likelihood_read ()
calc_vafs_likelihood ()

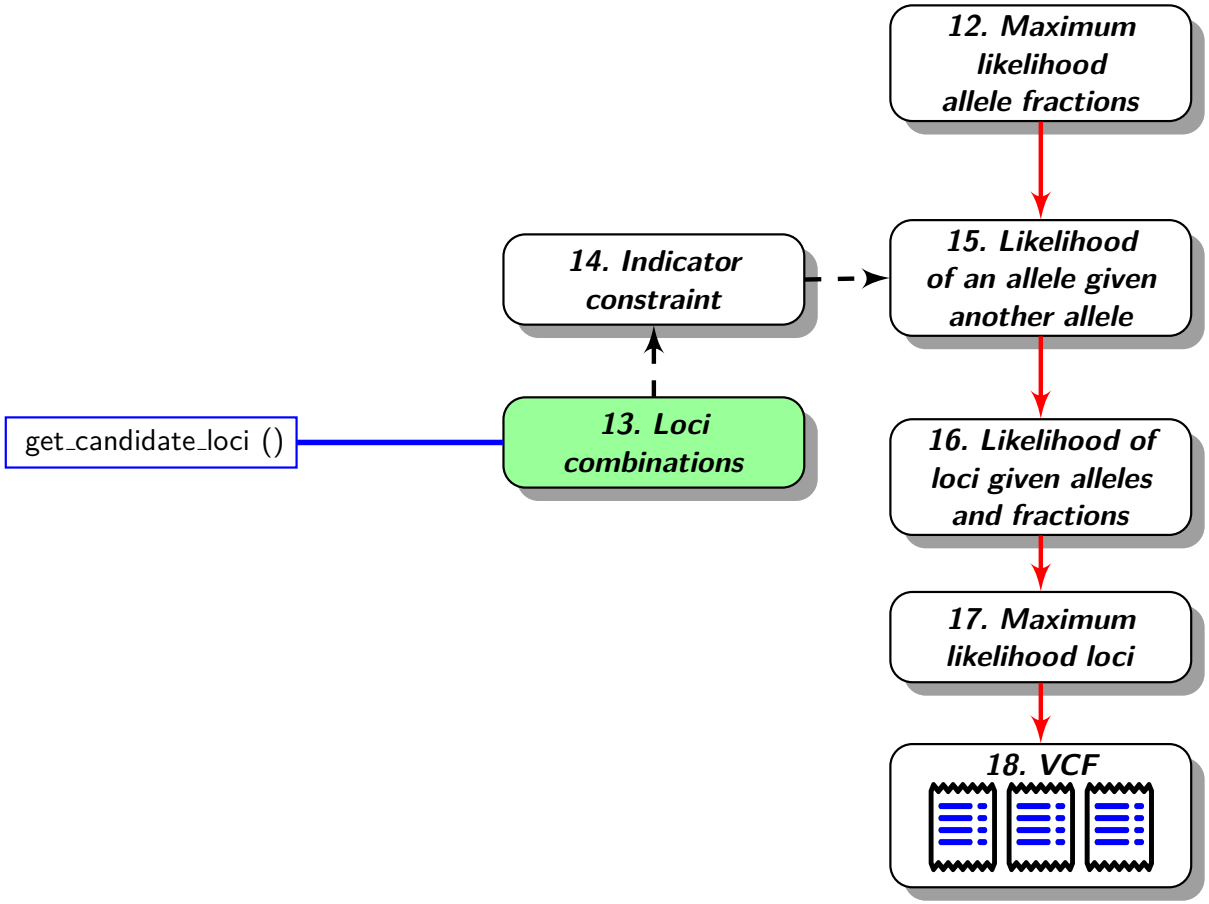

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

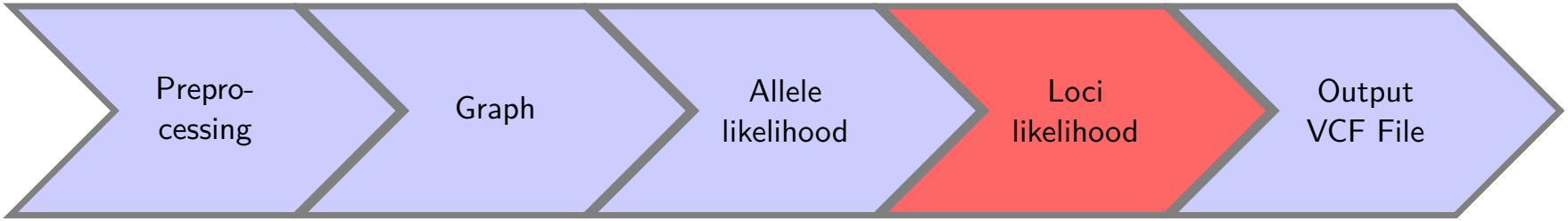

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

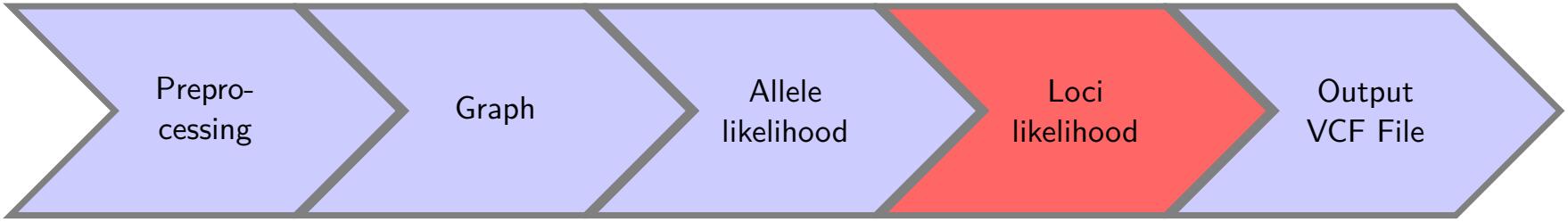

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

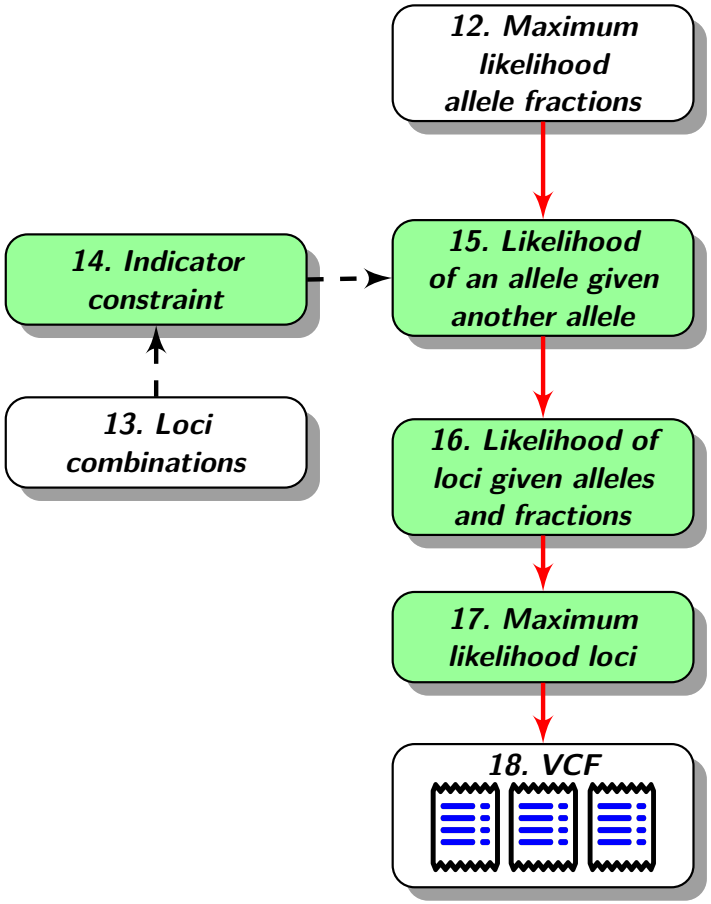

12. Maximum likelihood allele fractions
14. Indicator constraint
15. Likelihood of an allele given another allele
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

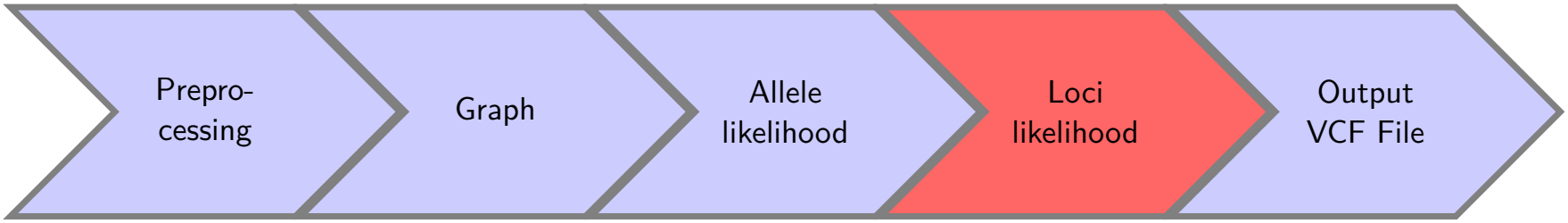

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

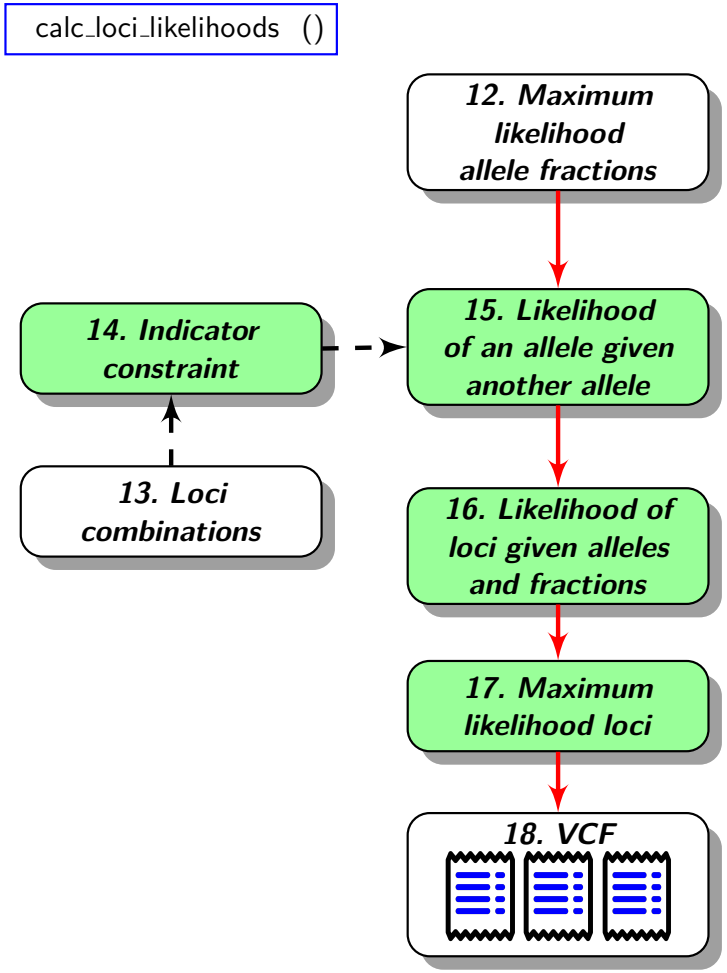

12. Maximum likelihood allele fractions
15. Likelihood of an allele given another allele
14. Indicator constraint
16. Likelihood of loci given alleles and fractions
13. Loci combinations
get_candidate_loci ()
17. Maximum likelihood loci
18. VCF

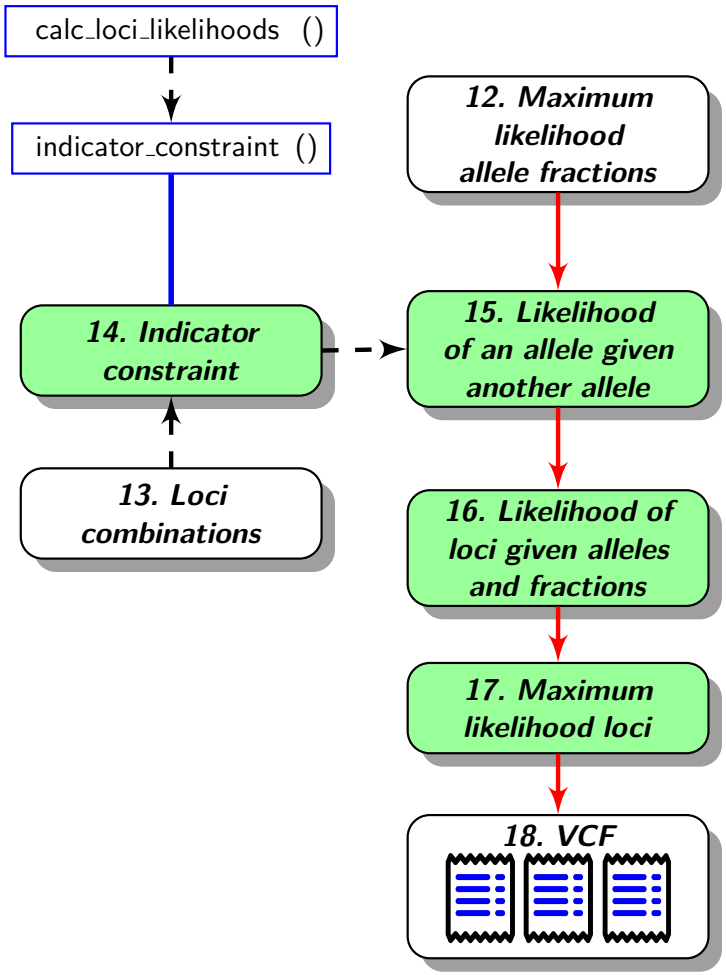

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

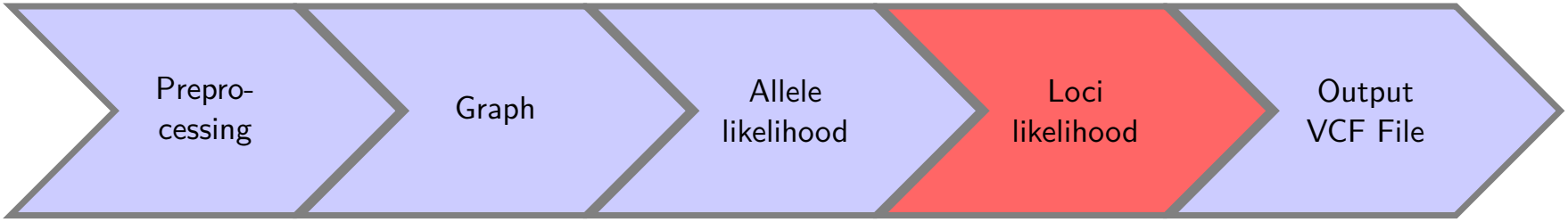

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

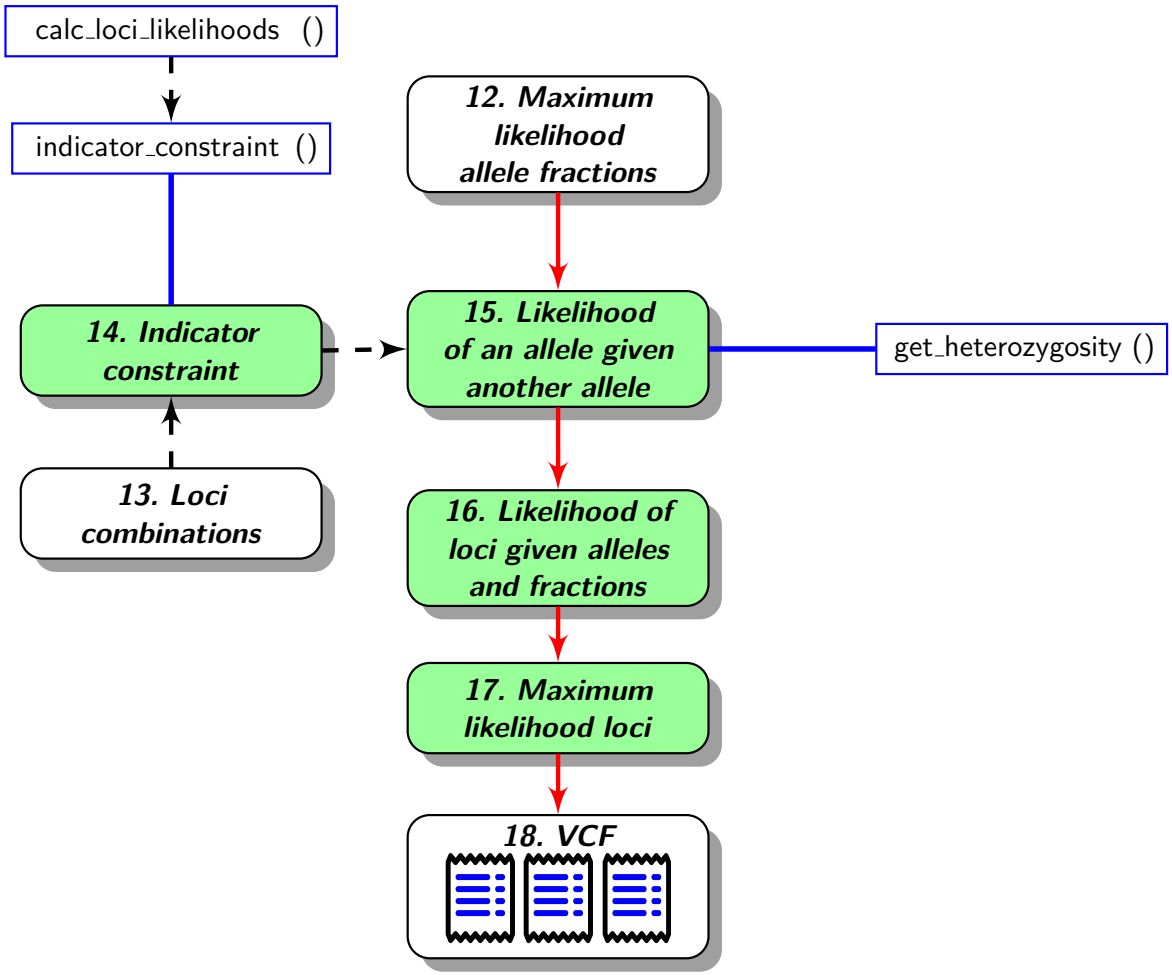

12. Maximum likelihood allele fractions
14. Indicator constraint
15. Likelihood of an allele given another allele
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

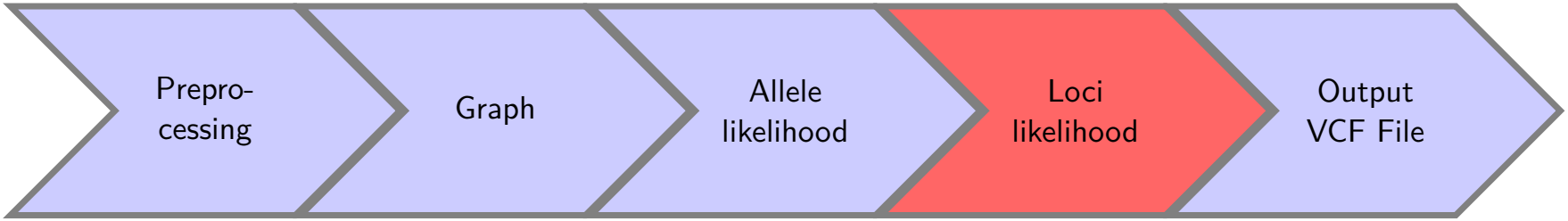

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

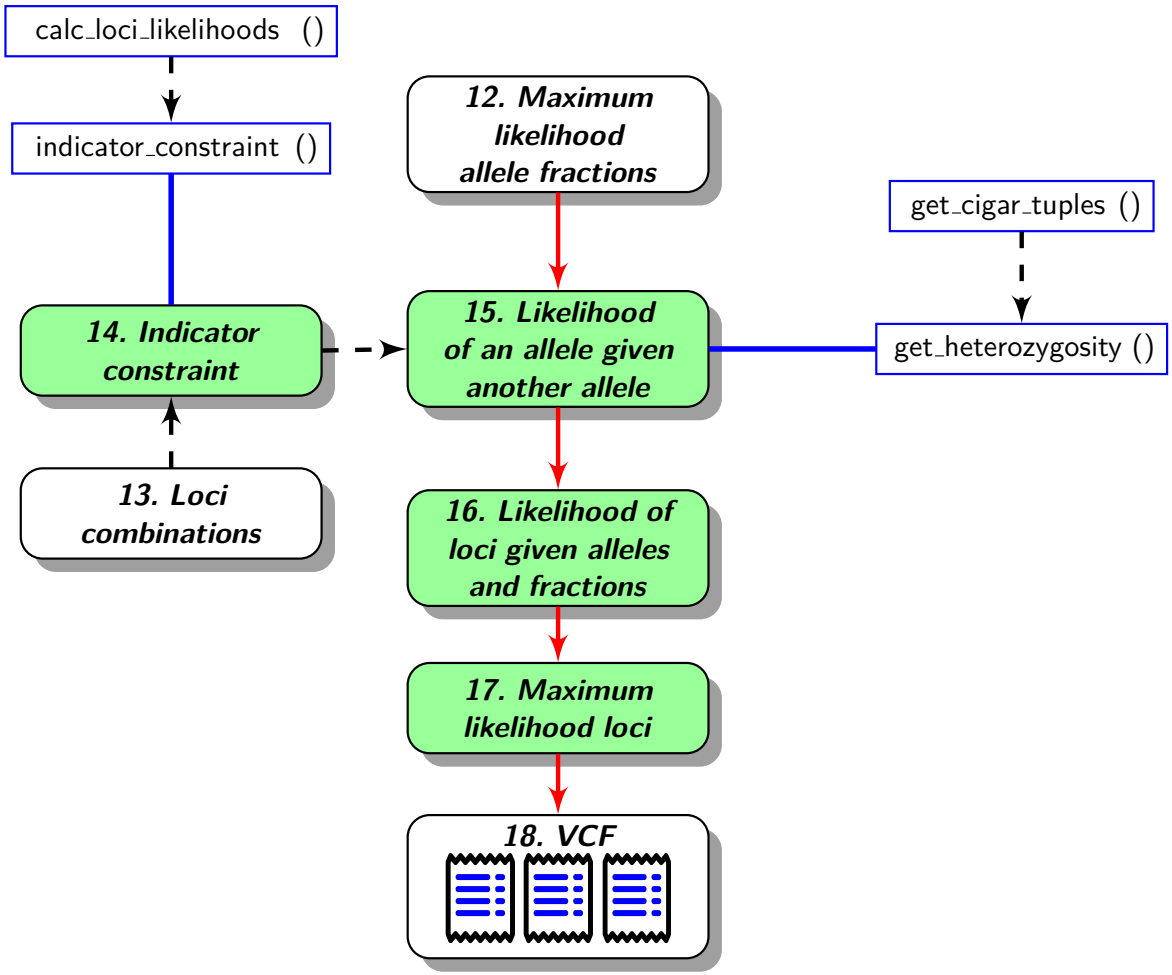

calc_loci_likelihoods ()
12. Maximum likelihood allele fractions
14. Indicator constraint
15. Likelihood of an allele given another allele
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

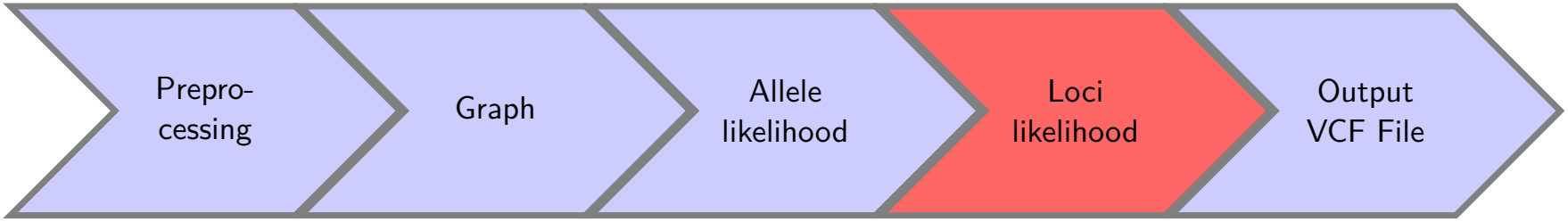

calc_loci_likelihoods ()
indicator_constraint ()
12. Maximum likelihood allele fractions
14. Indicator constraint
15. Likelihood of an allele given another allele
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

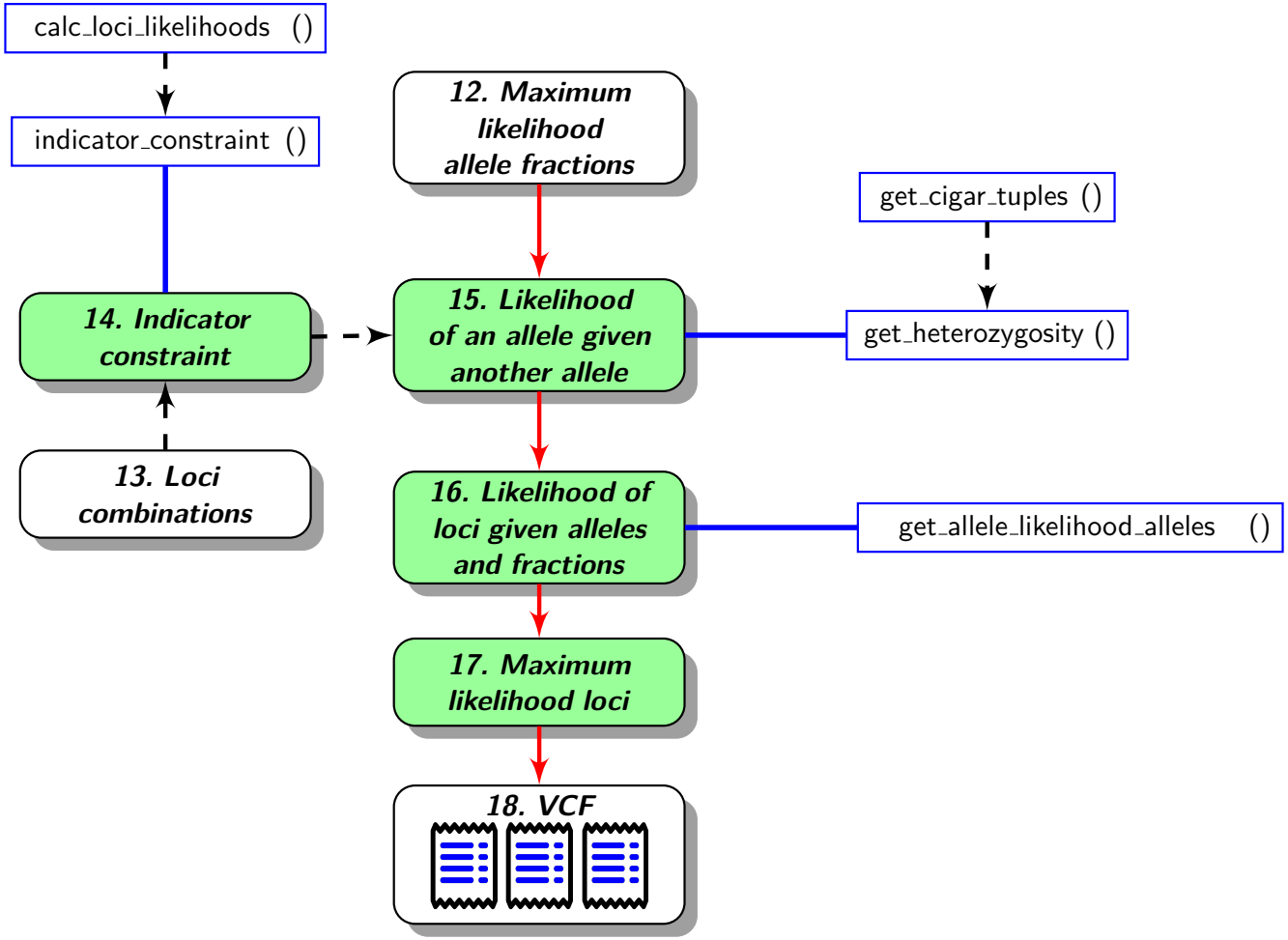

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

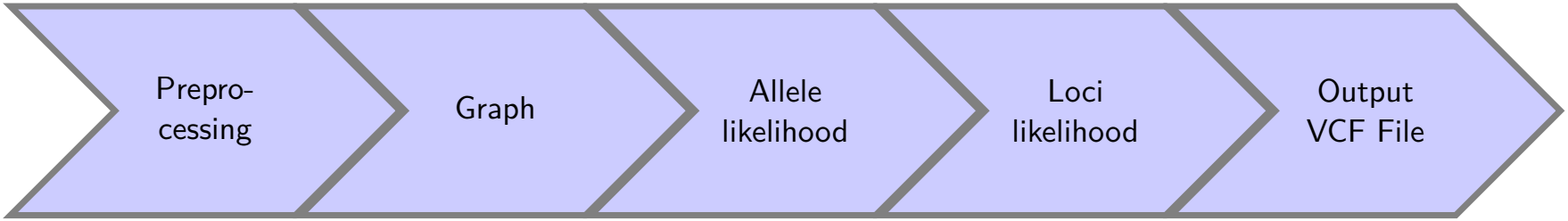

calc_loci_likelihoods ()
indicator_constraint ()
12. Maximum likelihood allele fractions
14. Indicator constraint
15. Likelihood of an allele given another allele
get_heterozygosity ()
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

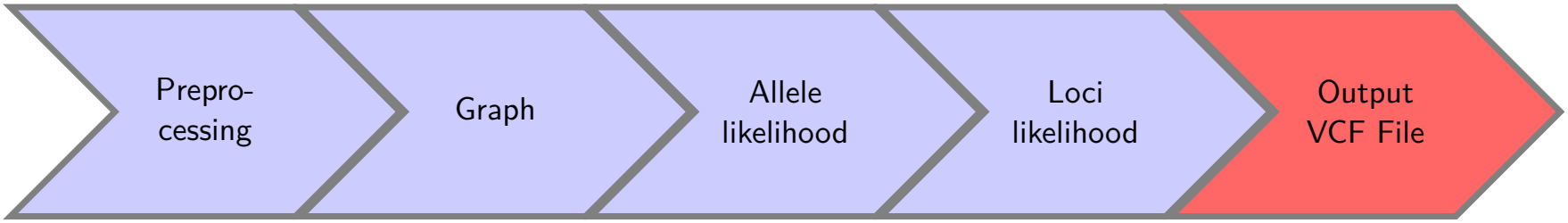

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

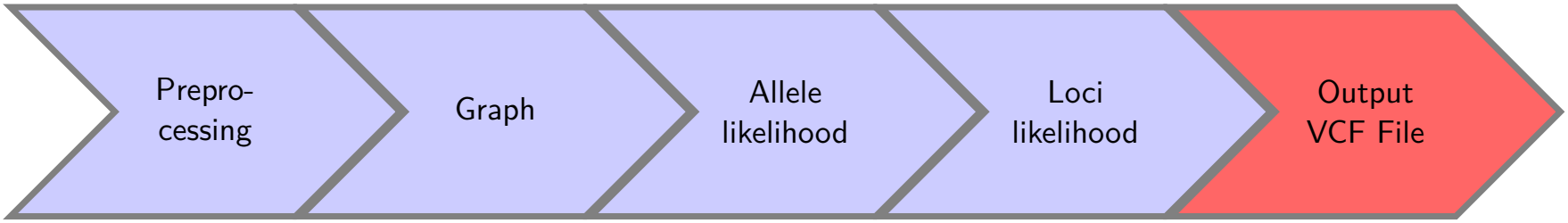

calc_loci_likelihoods ()
indicator_constraint ()
12. Maximum likelihood allele fractions
get_cigar_tuples ()
14. Indicator constraint
15. Likelihood of an allele given another allele
get_heterozygosity ()
13. Loci combinations
16. Likelihood of loci given alleles and fractions
17. Maximum likelihood loci
18. VCF

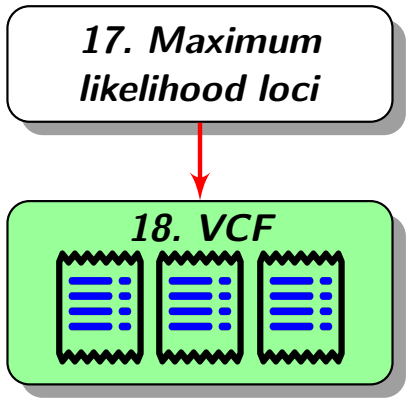

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

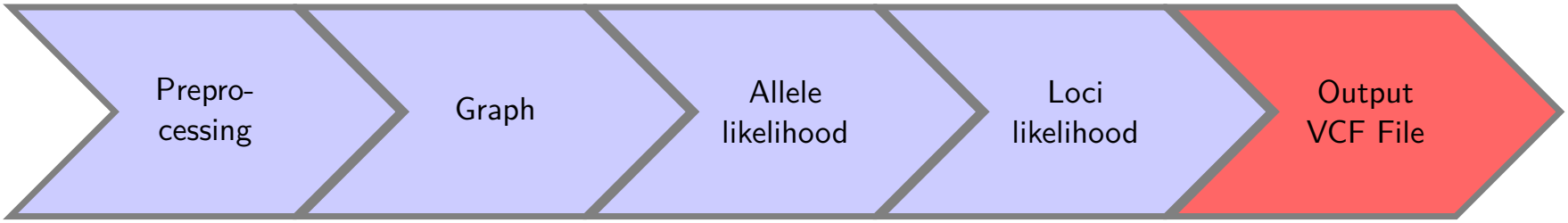

calc_loci_likelihoods ()
indicator_constraint ()
12. Maximum likelihood allele fractions
get_cigar_tuples ()
14. Indicator constraint
15. Likelihood of an allele given another allele
get_heterozygosity ()
13. Loci combinations
16. Likelihood of loci given alleles and fractions
get_allele_likelihood_alleles ()
17. Maximum likelihood loci
18. VCF

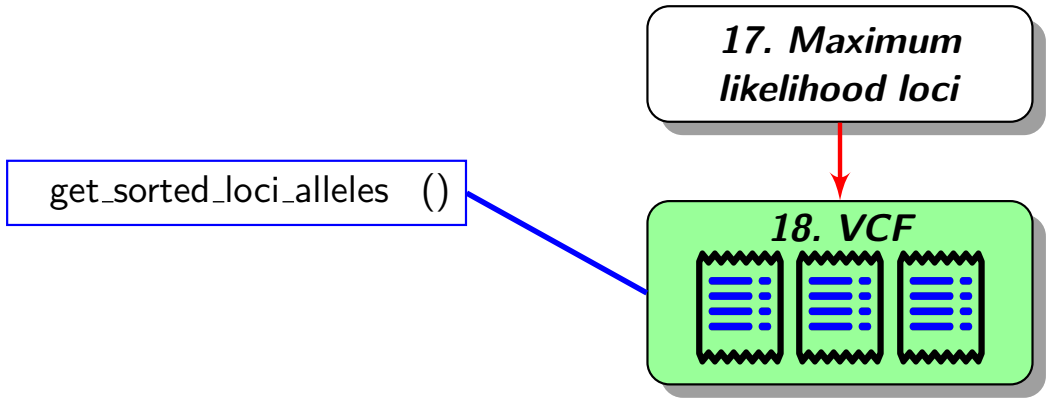

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

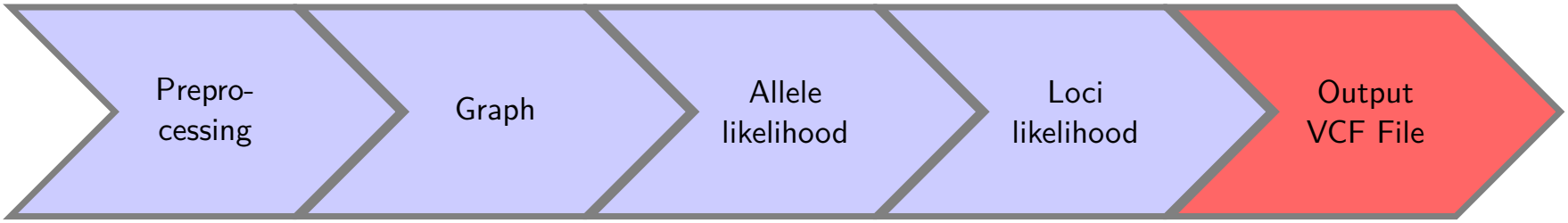

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

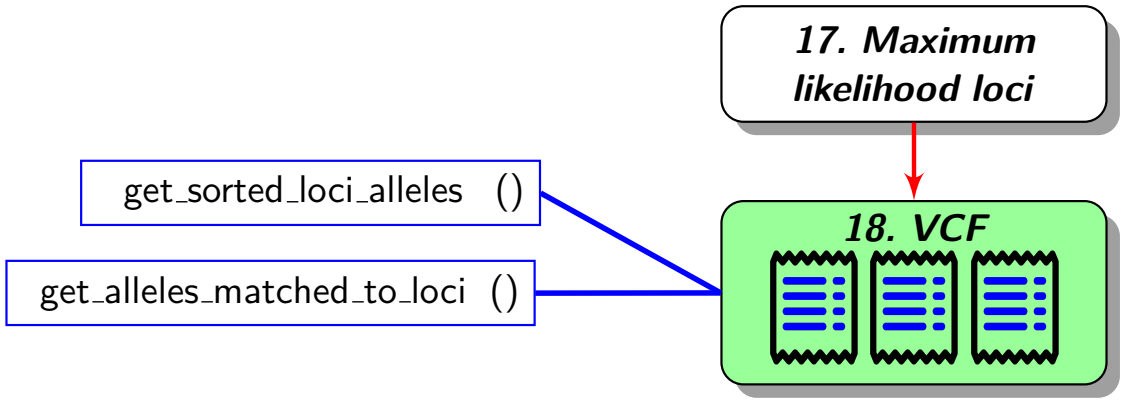

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

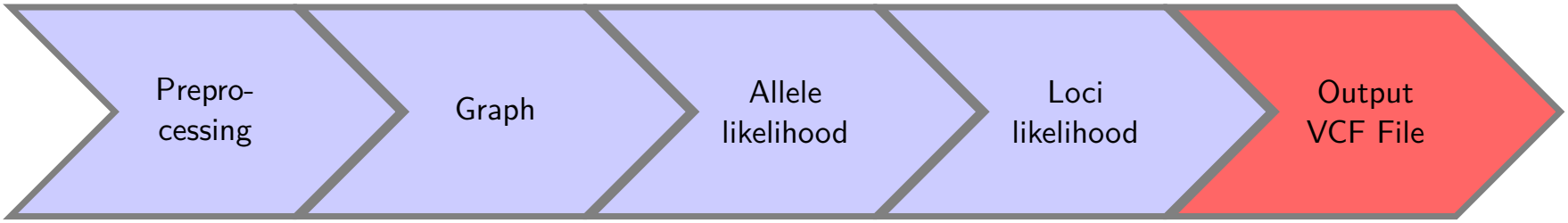

17. Maximum
likelihood loci
18. VCF

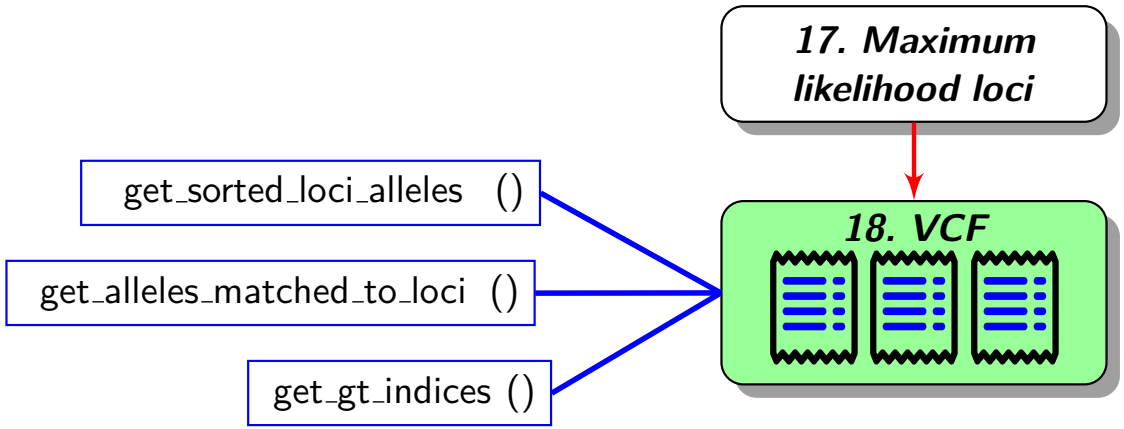

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File

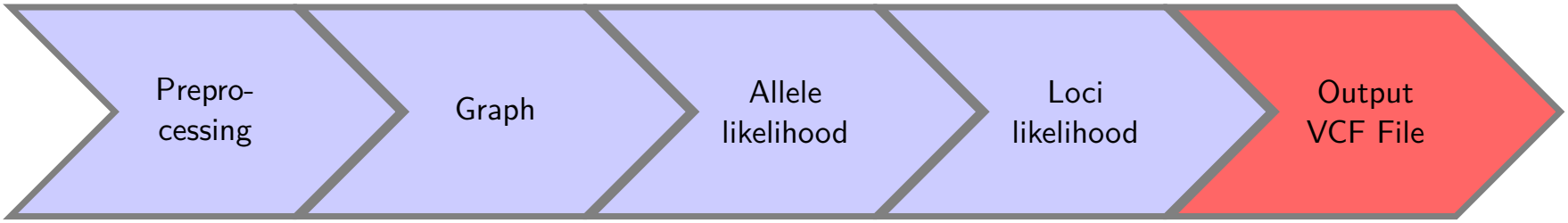

17. Maximum
likelihood loci
get_sorted_loci_alleles ()
18. VCF

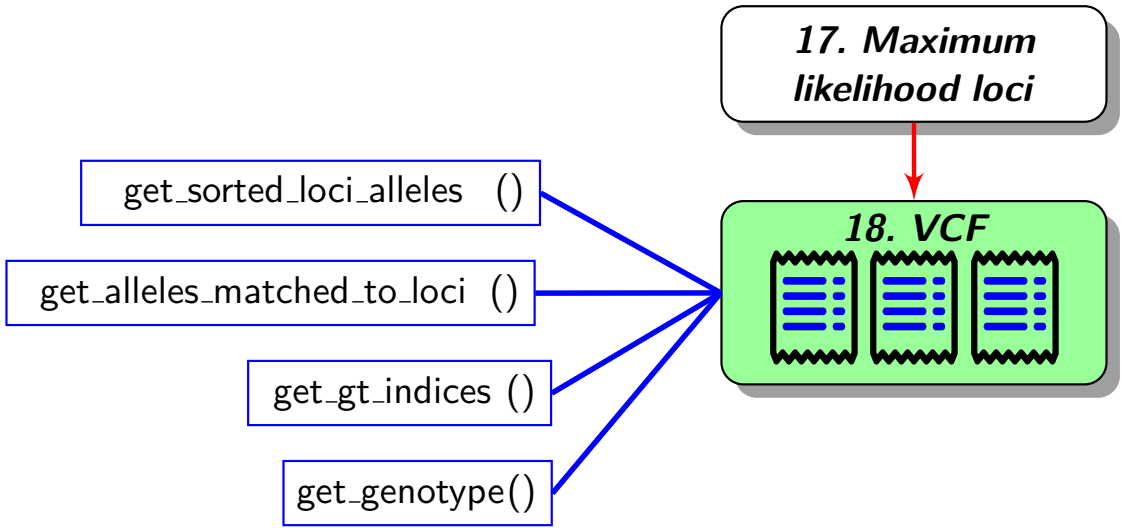

Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


17. Maximum
likelihood loci
get_sorted_loci_alleles ()
get_alleles_matched_to_loci ()
18. VCF


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


17. Maximum
likelihood loci
get_sorted_loci_alleles ()
get_alleles_matched_to_loci ()
get_gt_indices ()
18. VCF


Prepro-
cessing
Graph
Allele
likelihood
Loci
likelihood
Output
VCF File


17. Maximum
likelihood loci
get_sorted_loci_alleles ()
get_alleles_matched_to_loci ()
get_gt_indices ()
get_genotype()
18. VCF

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

text

Bildquellen I

[1] Mouagip: *Aminoacids table.svg*. 2021. – source: `https://commons.wikimedia.org/wiki/File:Aminoacids_table.svg`

[2] Margulies, Elliott: *Transcription*. – source: `https://www.genome.gov/genetics-glossary/Transcription`

[3] Leja, Darryl: *Transfer RNA (tRNA)*. – source: `https://medlineplus.gov/genetics/understanding/basics/noncodingdna/`

[4] Margulies, Elliott: *Transfer RNA (tRNA)*. – source: `https://www.genome.gov/genetics-glossary/Transfer-RNA`

[5] Ruiz, Mariana: *DNA replication*. – source: `https://commons.wikimedia.org/wiki/File:DNA_replication_en.svg`

[6] Collins, Francis: *Mutation*. – source: `https://www.genome.gov/genetics-glossary/Mutation`

Bildquellen II

[7] Enzoklop: *Polymerase Chain Reaction - Schematic mechanism of PCR*. – source: `https://en.wikipedia.org/wiki/File:Polymerase_chain_reaction-en.svg`

[8] Christoph Goemans, Norman M.: *Prinzip der DNA-Sequenzierung nach der Didesoxy-Methode*. – source: `https://de.wikipedia.org/wiki/Datei:Didesoxy-Methode.svg`

[9] Clark, Jonathan: *Schematic diagram of RADseq*. – source: `https://en.wikipedia.org/wiki/File:RADseq_schematic.pdf`