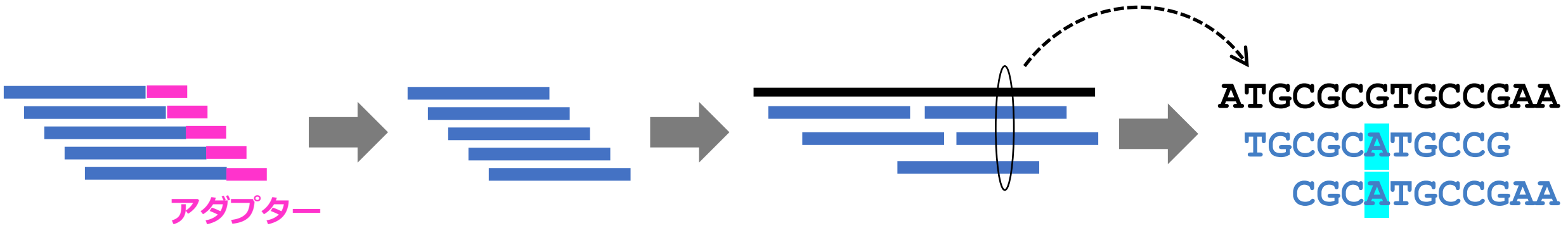
高度なワークフローを用いた
複雑なNGSデータ処理プロセスの自動化

フィルジェン株式会社　バイオインフォマティクス部
（ support@filgen.jp ）

CLCのワークフロー機能

解析フローの例（SNP検出）



①トリミング　②マッピング　③SNP検出

CLCのワークフロー機能

Genomics Workbench上での手順

ERR277212.R1 (paired)

①トリミング

Trim Reads

ERR277212.R1 (paired, trimmed pairs)

ERR277212.R1 (paired, trimmed pairs)

②マッピング

Map Reads to Reference

Mapped_reads-ERR277212.R1 (paired)

Mapped_reads-ERR277212.R1 (paired)

③SNP検出

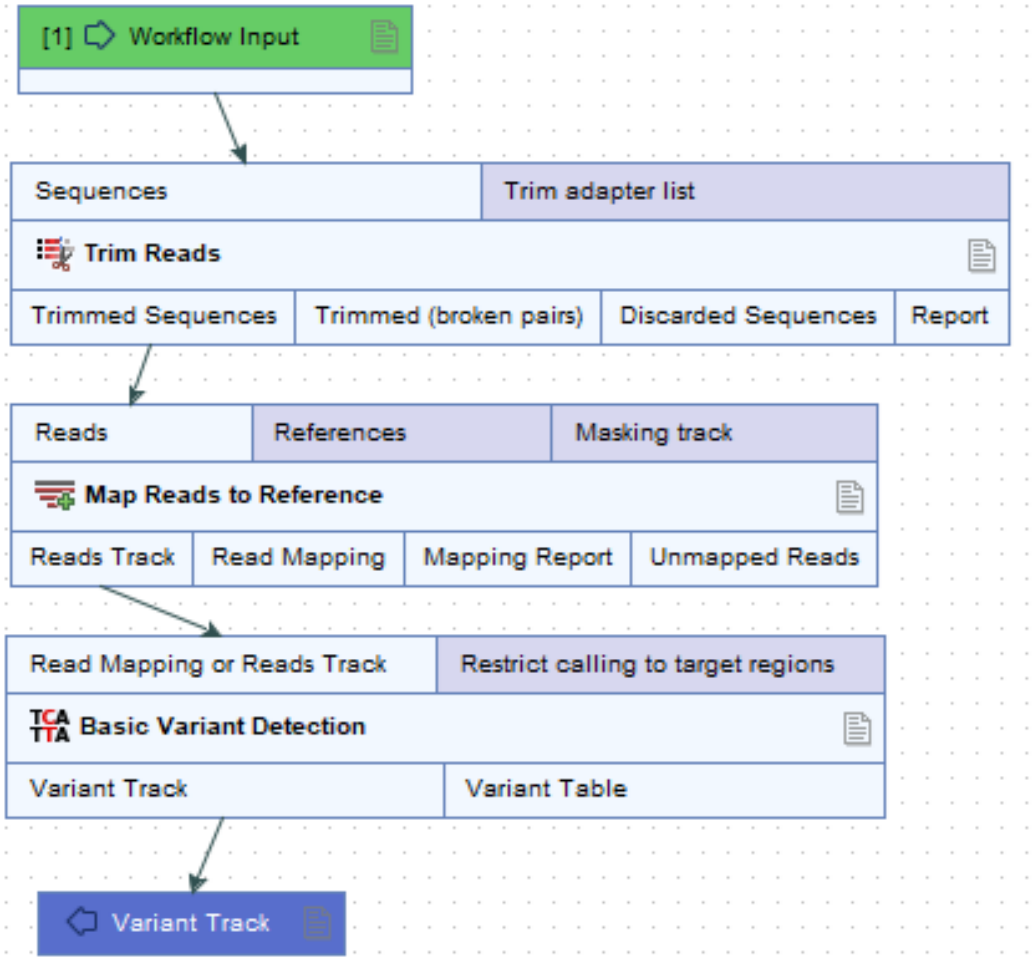
Basic Variant Detection

Variants_passing_filters-ERR277212.R1 (paired)

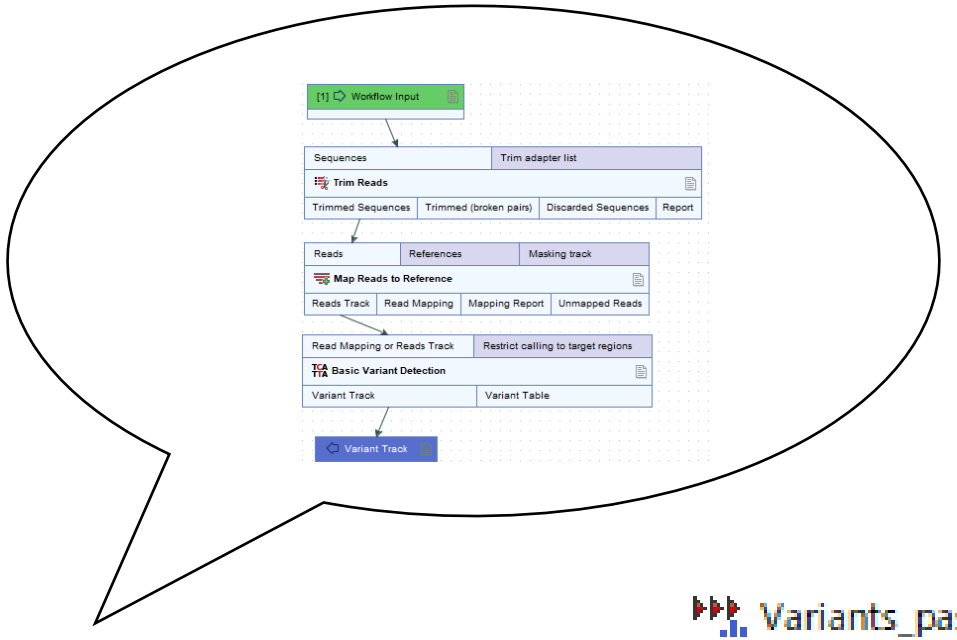
Chromosome	Region	Type	Ref...	Alle...	Ref...	Len...	Zygo...	Cou...	Cov...	Fre...	Pro...	For...
NZ_CP014971	10536^...	Inser...	-	T	No	1	Homo...	12	12	100....	1.00	5
NZ_CP014971	13751	SNV	C	T	No	1	Homo...	21	21	100....	1.00	11
NZ_CP014971	16549	SNV	C	T	No	1	Homo...	18	20	90.00	1.00	7
NZ_CP014971	20304	SNV	C	T	No	1	Homo...	18	20	90.00	1.00	12
NZ_CP014971	43355	SNV	T	C	No	1	Homo...	15	15	100....	1.00	9
NZ_CP014971	92398	SNV	T	G	No	1	Homo...	21	21	100....	1.00	12
NZ_CP014971	100693	SNV	C	T	No	1	Homo...	11	11	100....	1.00	7
NZ_CP014971	131693	SNV	T	C	No	1	Homo...	13	13	100....	1.00	8

CLCのワークフロー機能

ツールのフローを登録



CLCのワークフロー機能



ERR277212.R1 (paired)

Trim, Map and Detect SNPs

Variants_passing_filters-ERR277212.R1 (paired)

Chromosome	Region	Type	Ref...	Alle...	Ref...	Len...	Zygo...	Cou...	Cov...	Fre...	Pro...	For...
NZ_CP014971	10536^...	Inser...	-	T	No	1	Homo...	12	12	100....	1.00	5
NZ_CP014971	13751	SNV	C	T	No	1	Homo...	21	21	100....	1.00	11
NZ_CP014971	16549	SNV	C	T	No	1	Homo...	18	20	90.00	1.00	7
NZ_CP014971	20304	SNV	C	T	No	1	Homo...	18	20	90.00	1.00	12
NZ_CP014971	43355	SNV	T	C	No	1	Homo...	15	15	100....	1.00	9
NZ_CP014971	92398	SNV	T	G	No	1	Homo...	21	21	100....	1.00	12
NZ_CP014971	100693	SNV	C	T	No	1	Homo...	11	11	100....	1.00	7
NZ_CP014971	131693	SNV	T	C	No	1	Homo...	13	13	100....	1.00	8

①トリミング

②マッピング

③SNP検出

CLCのバッチ機能

トリミングしたいリードデータが100個ある場合…

トリミングを100回実行するのは大変。

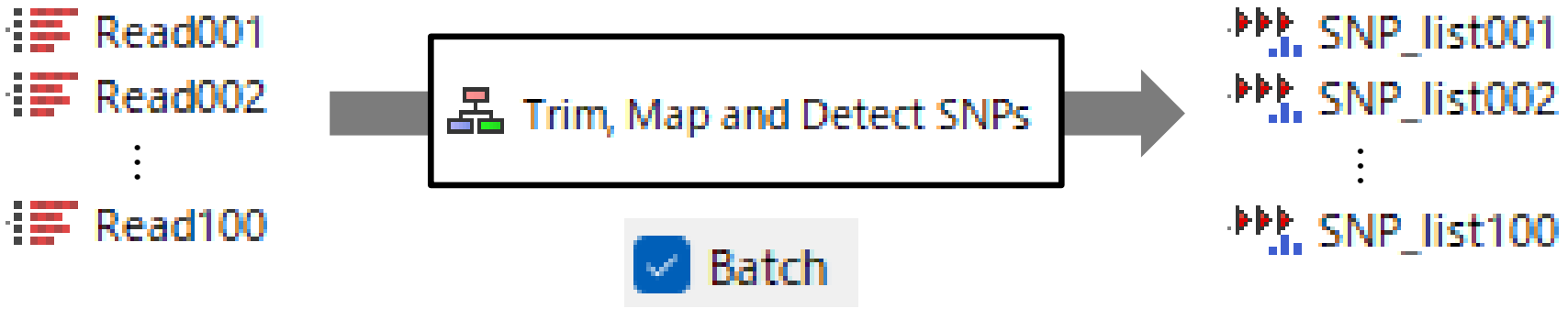
BatchモードをONにして100個のデータを指定すると、自動で1データずつ100回トリミングを行う。

→1回実行するだけでいい。

CLCのバッチ機能

トリミング、マッピング、SNP検出したいリードデータが100個ある場合…

それぞれの工程を100回実行するのは、とても大変。



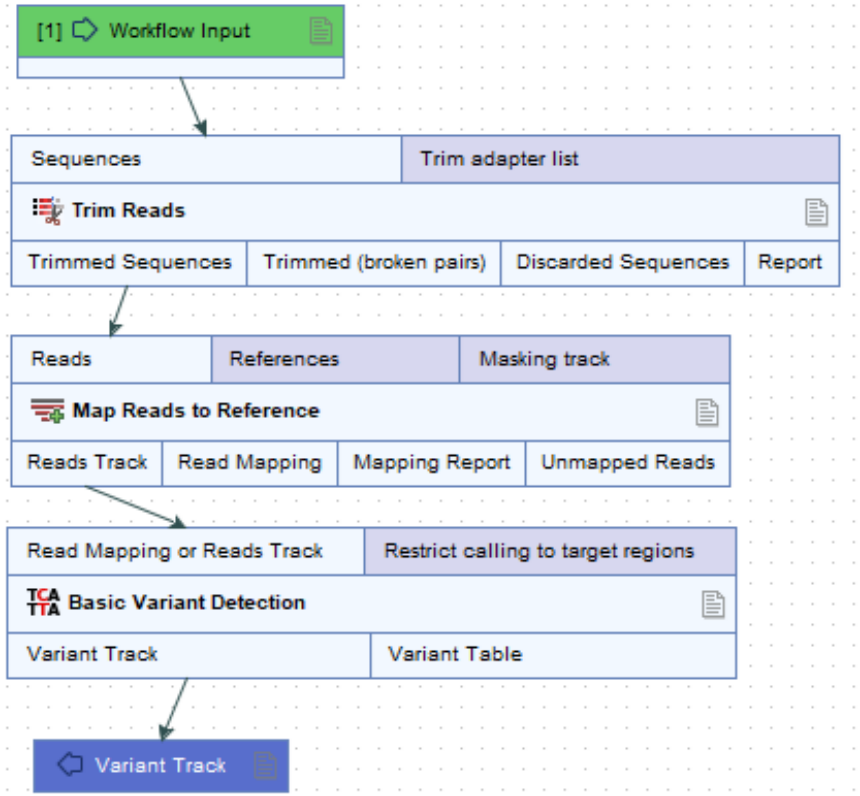
BatchモードをONにして100個のデータを指定すると、自動で1データずつトリミング・マッピング・SNP検出を行う。

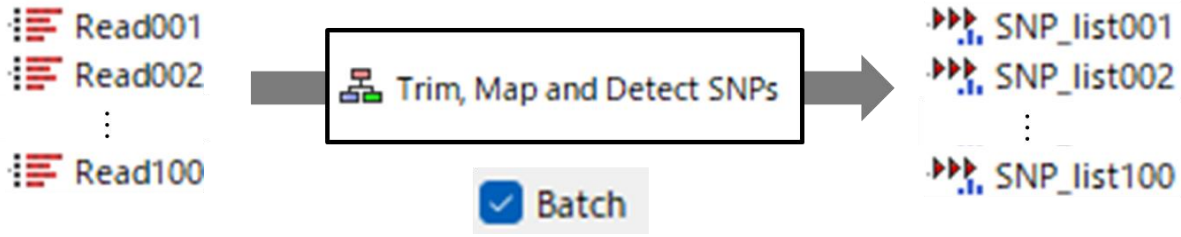
→1回実行するだけでいい。

単純なワークフロー・複雑なワークフロー

全てのリードに対して、同じリファレンス配列を使う場合は簡単。

最初にリファレンス配列を１個指定して、
それに各リードデータのマッピングを繰り返すだけでいい。





毎回入力が異なるだけで、中でやることは一緒（単純な反復作業）

単純なワークフロー・複雑なワークフロー

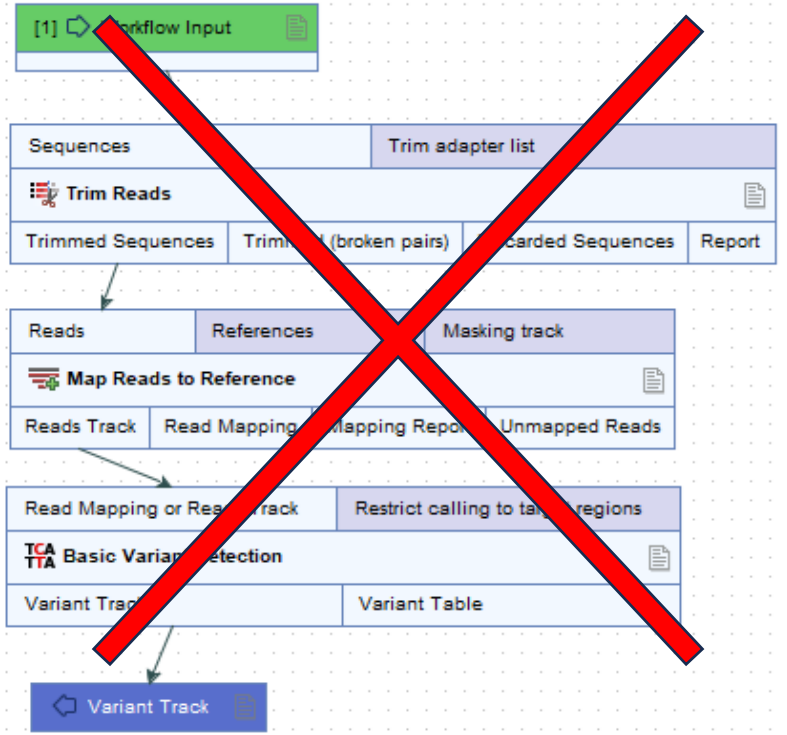
リードに応じてマッピングしたいリファレンスが異なる場合はより複雑。

どのリードをどのリファレンスにマッピングするかをソフトウェアに教える必要がある。

（例）

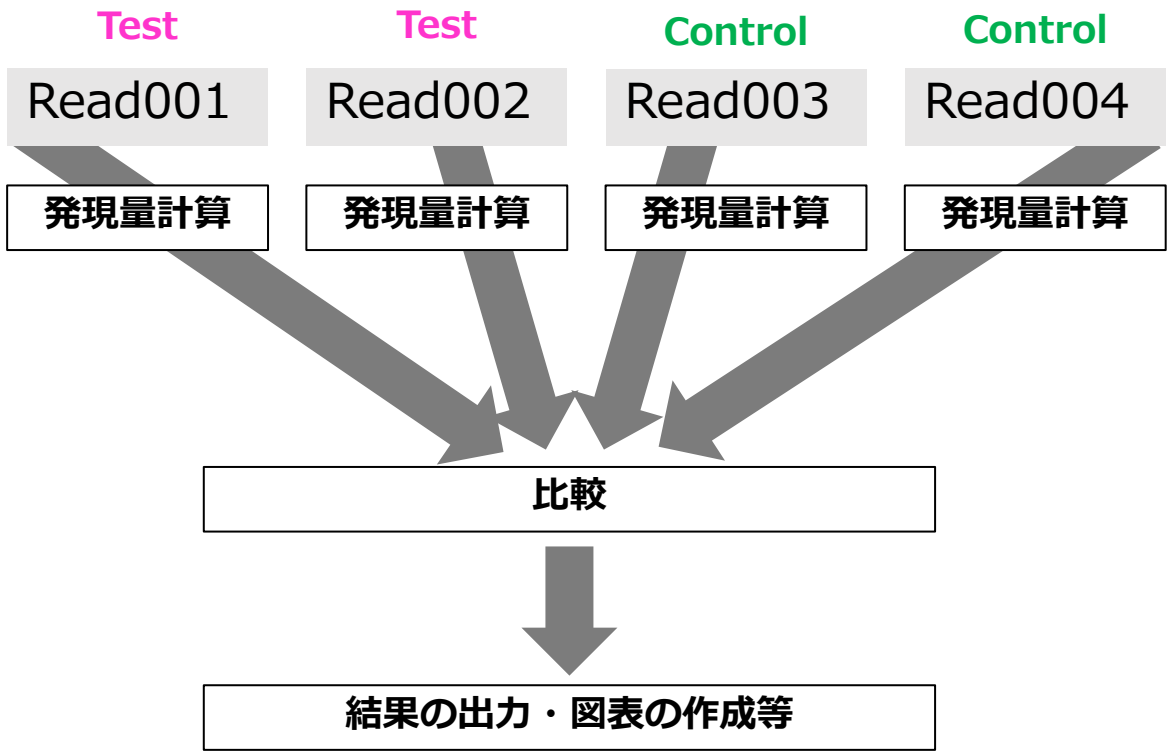
Read001～Read020：リファレンスAにマッピングしてSNP検出
Read021～Read040：リファレンスBにマッピングしてSNP検出
Read041～Read060：リファレンスCにマッピングしてSNP検出
Read061～Read080：リファレンスDにマッピングしてSNP検出
Read081～Read100：リファレンスEにマッピングしてSNP検出

入力に応じて、中でやることも異なるため単純な反復作業にならない
→制御要素を含める必要がある。



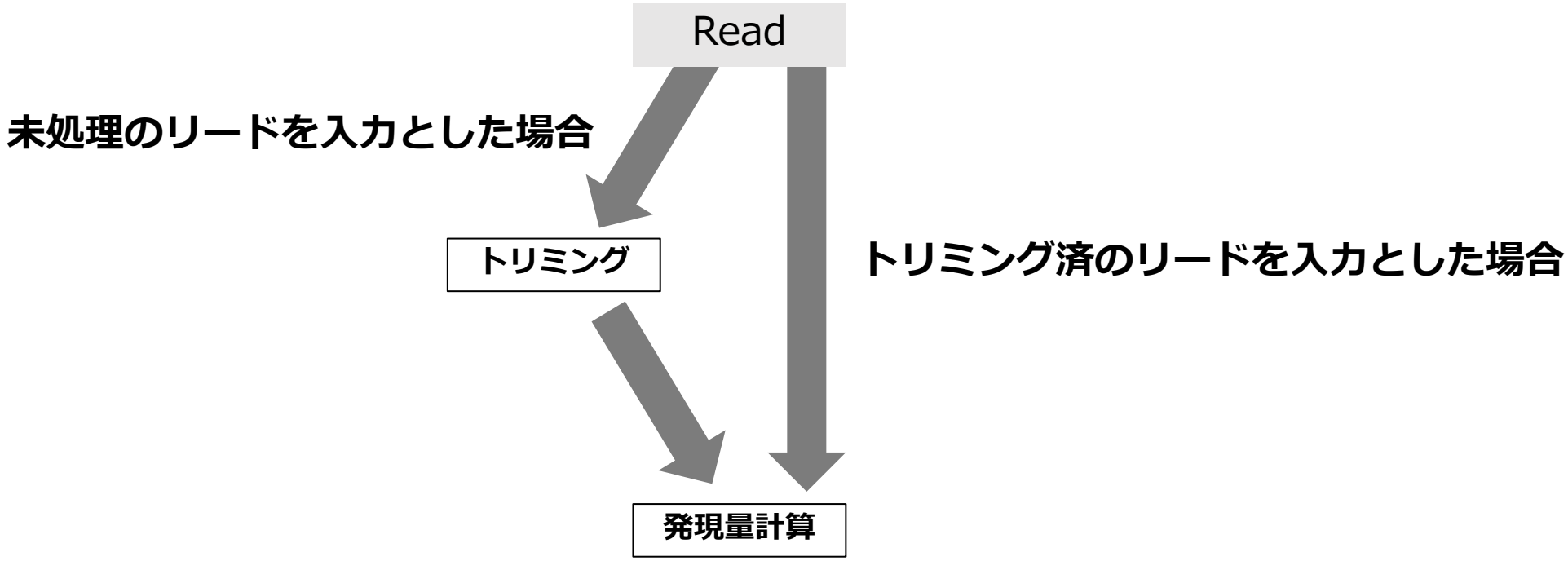
単純なワークフロー・複雑なワークフロー

各リードについて、遺伝子発現量を計算した後、比較を行う場合。

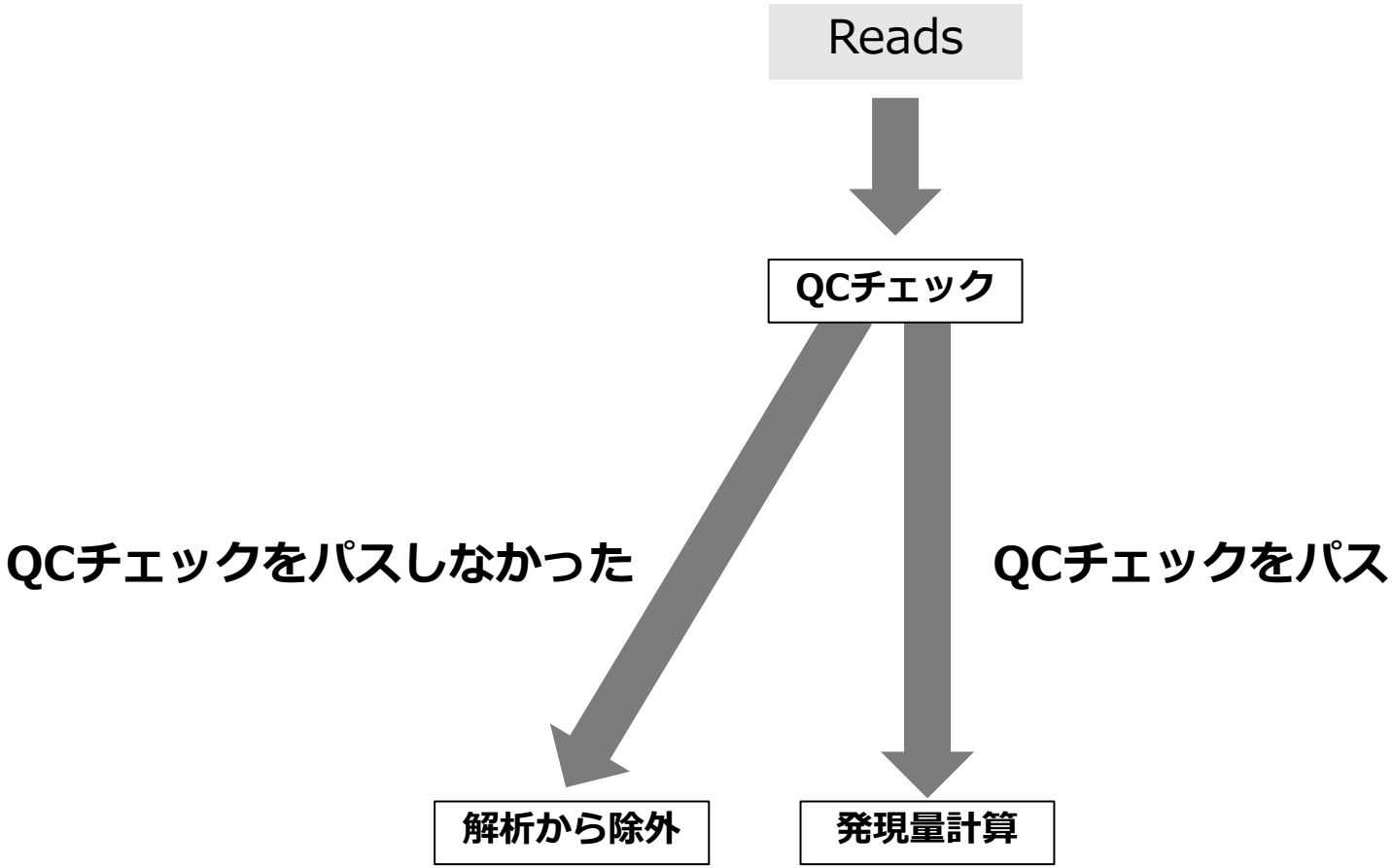


一本道ではないため、単純なワークフローでは対応できない。

単純なワークフロー・複雑なワークフロー



単純なワークフロー・複雑なワークフロー



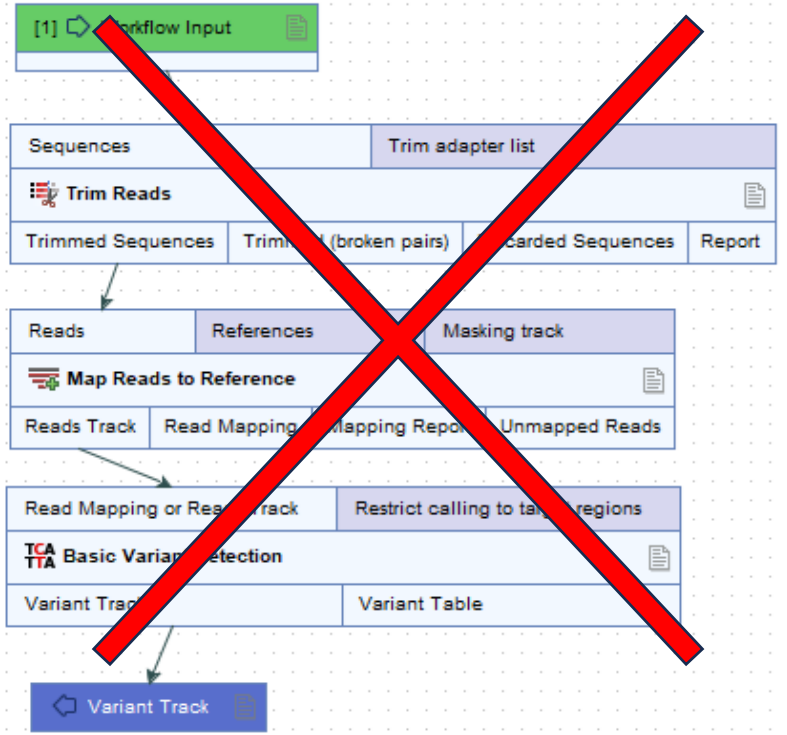
単純なワークフロー・複雑なワークフロー

リードに応じてマッピングしたいリファレンスが異なる場合はより複雑。

どのリードをどのリファレンスにマッピングするかをソフトウェアに教える必要がある。

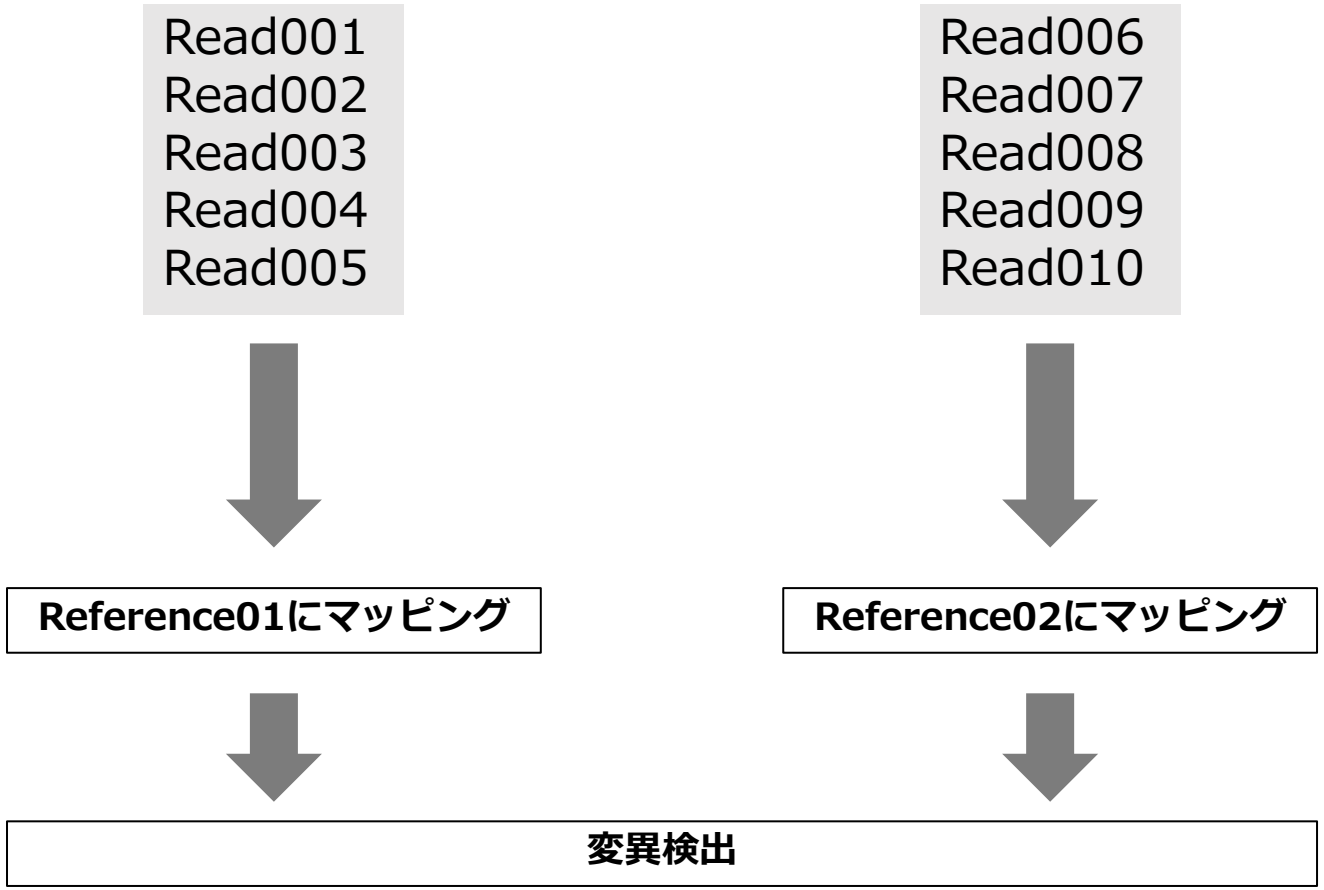
（例）

Read001～Read020：リファレンスAにマッピングしてSNP検出
Read021～Read040：リファレンスBにマッピングしてSNP検出
Read041～Read060：リファレンスCにマッピングしてSNP検出
Read061～Read080：リファレンスDにマッピングしてSNP検出
Read081～Read100：リファレンスEにマッピングしてSNP検出



入力に応じて、中でやることも異なるため単純な反復作業にならない
→制御要素を含める必要がある。

パターン1:入力に応じてリファレンスが異なる



パターン1:入力に応じてリファレンスが異なる

事前に2つのメタデータを用意し、インポートする必要がある。

Read_metadata

	A	B	C	D
1	Name	Prefecture	Sex	Reference
2	Read001	Aichi	F	Reference01
3	Read002	Mie	M	Reference01
4	Read003	Mie	M	Reference01
5	Read004	Gifu	M	Reference01
6	Read005	Aichi	F	Reference01
7	Read006	Mie	F	reference02
8	Read007	Gifu	M	reference02
9	Read008	Gifu	F	reference02
10	Read009	Aichi	M	reference02
11	Read010	Aichi	F	reference02

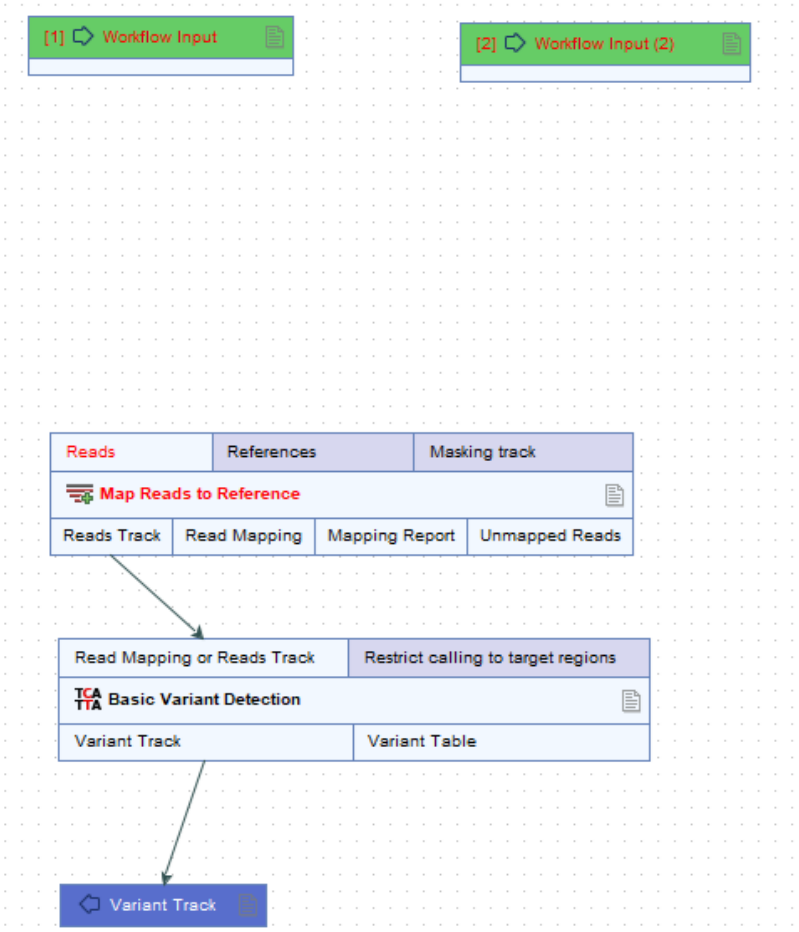
①リードデータに対するメタデータ

Reference_metadata

	A
1	Reference
2	Reference01
3	reference02

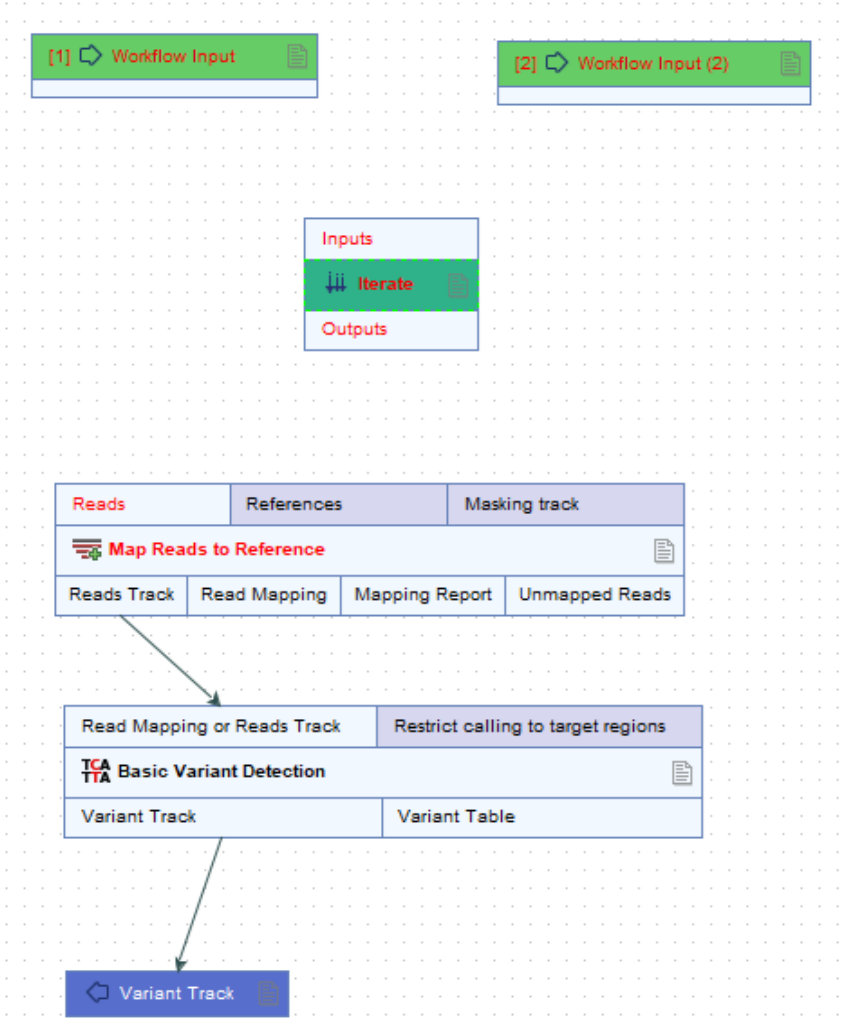
②リファレンスデータに対するメタデータ

パターン1:入力に応じてリファレンスが異なる



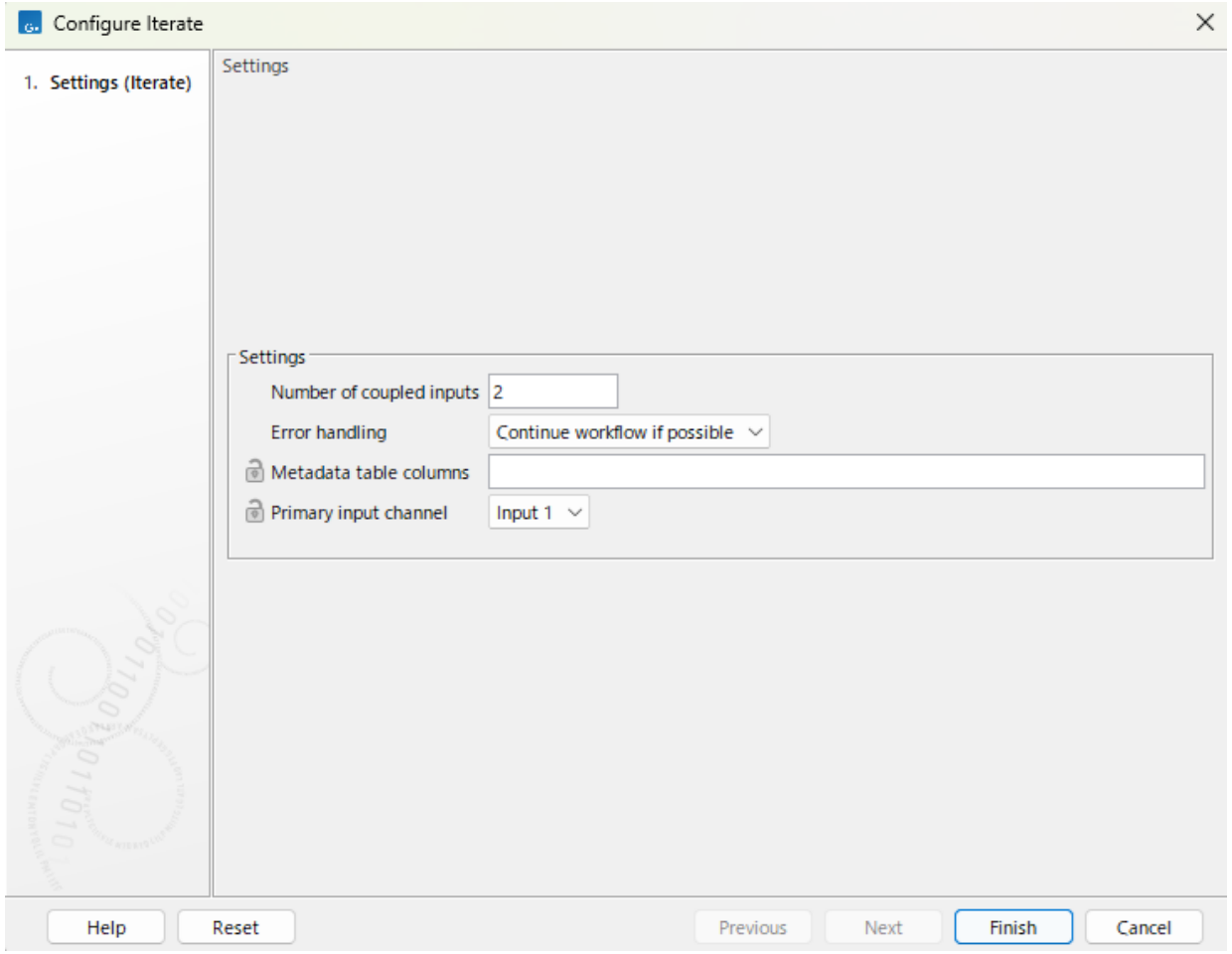
まずはリードとリファレンスに対応するInput2つと、その他の要素を配置

パターン1:入力に応じてリファレンスが異なる



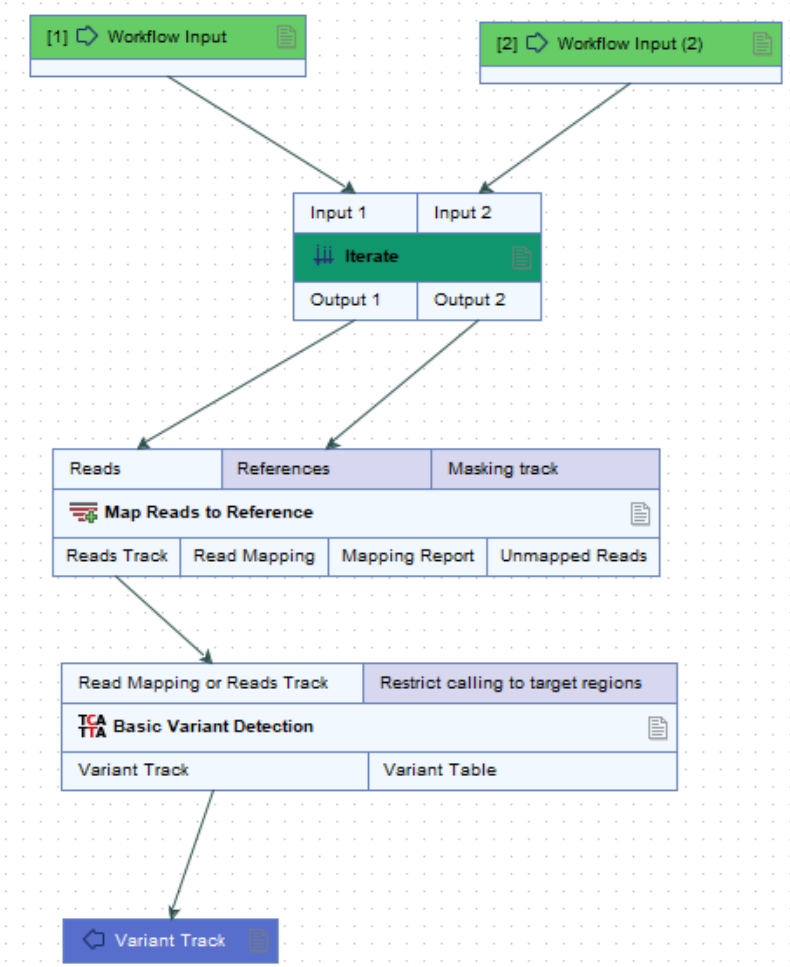
次に制御要素の1つのである”Iterate”を配置

パターン1:入力に応じてリファレンスが異なる



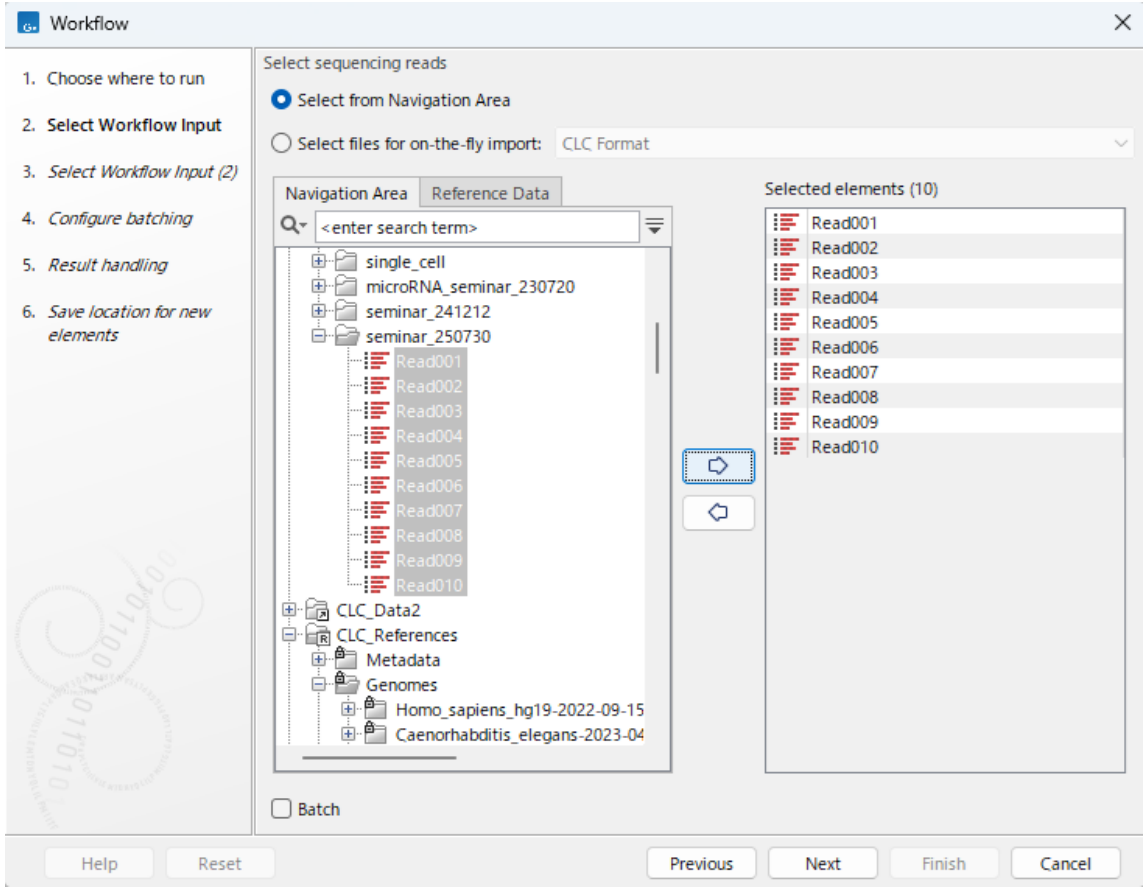
Number of coupled inputsを”2”とする。
（ReadとReferenceの2種類のデータを紐づけたい）

パターン1:入力に応じてリファレンスが異なる



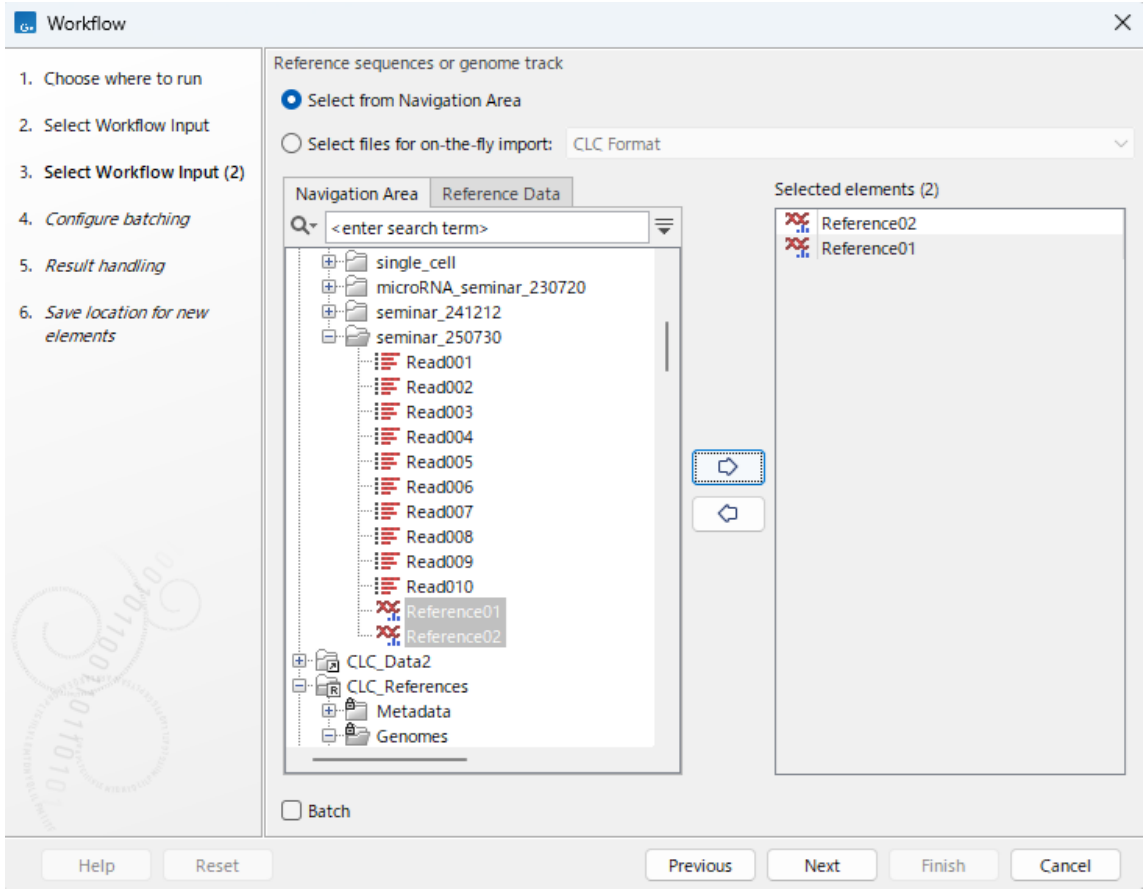
図のように線で結ぶと完成

ワークフローを実行



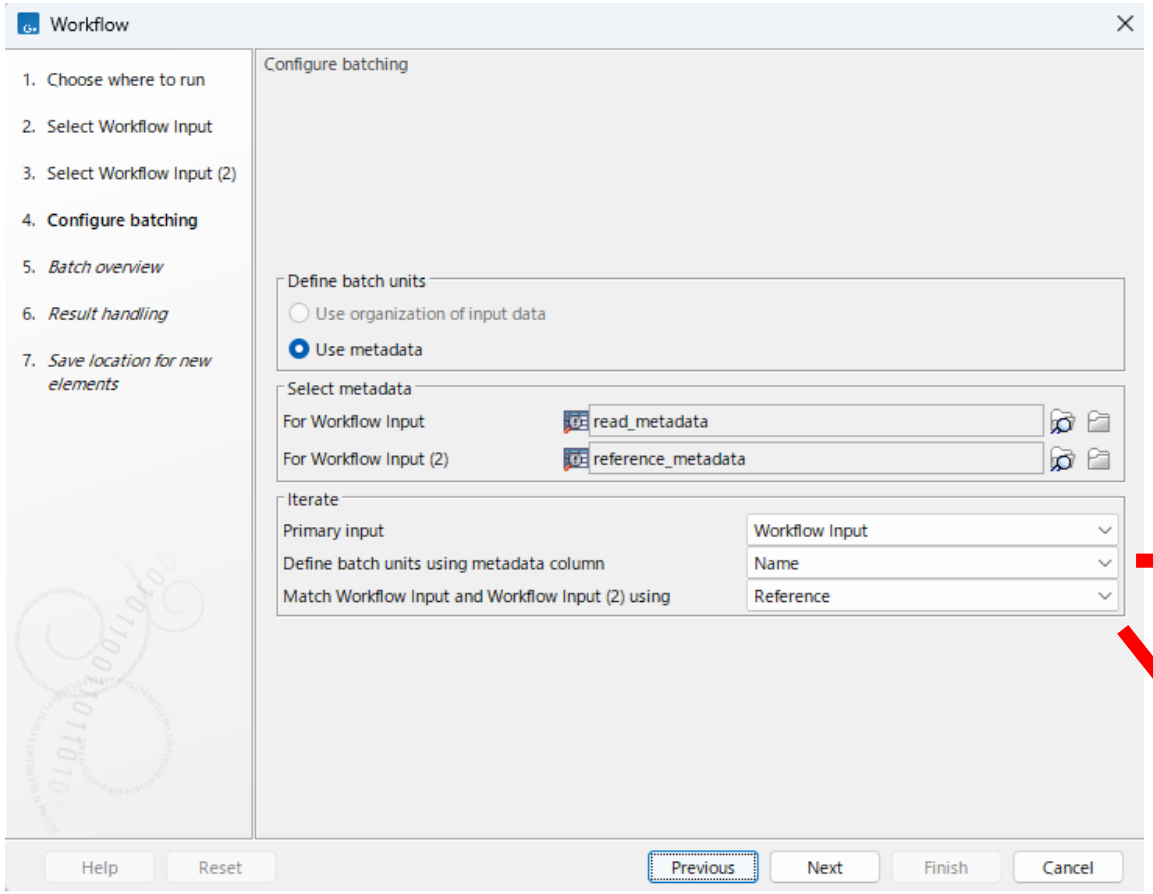
まずは全てのリードデータを選択

ワークフローを実行



次にすべてのリファレンスデータを選択

ワークフローを実行

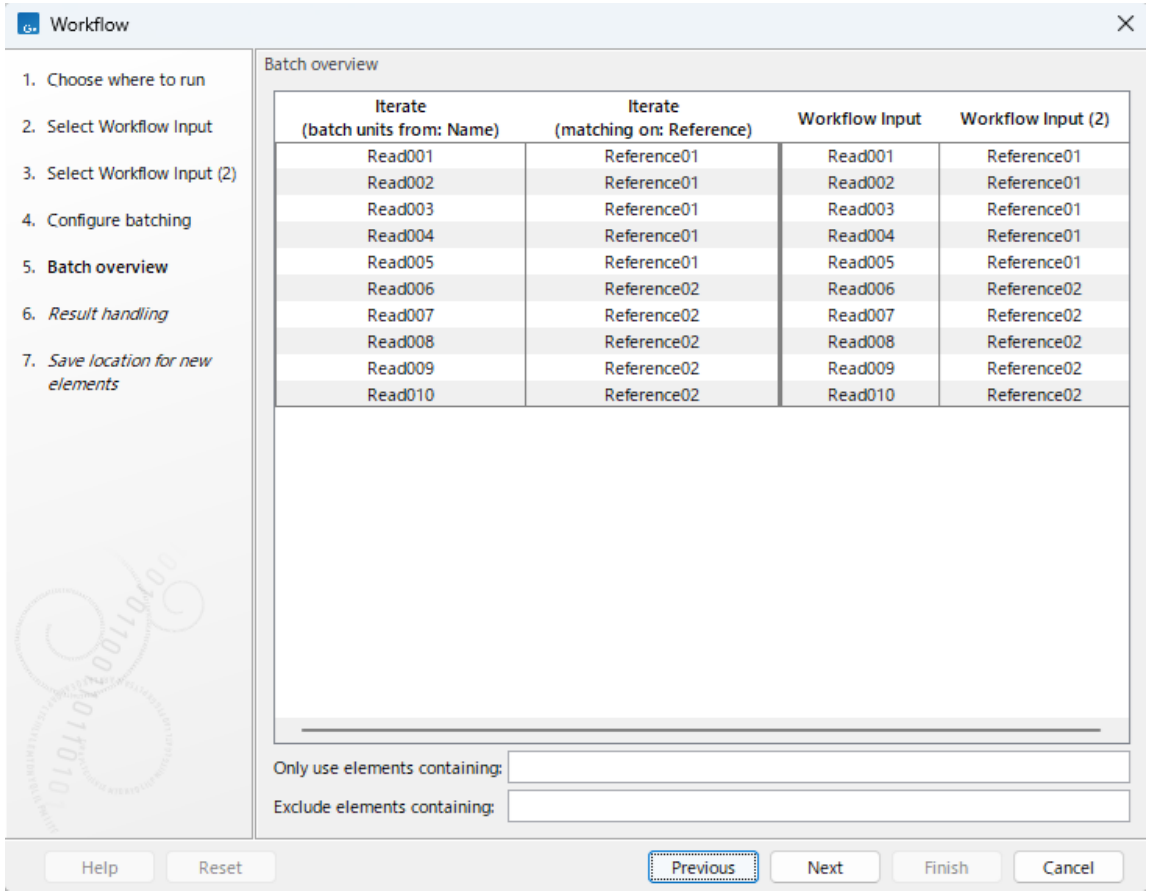


事前にインポートしたメタデータを指定

メタデータで繰り返しの単位を指定
今回は、リードごとにSNP解析を繰り返すので、
リード名を表すName列を指定

リファレンスとリードを紐づけるための列を指定

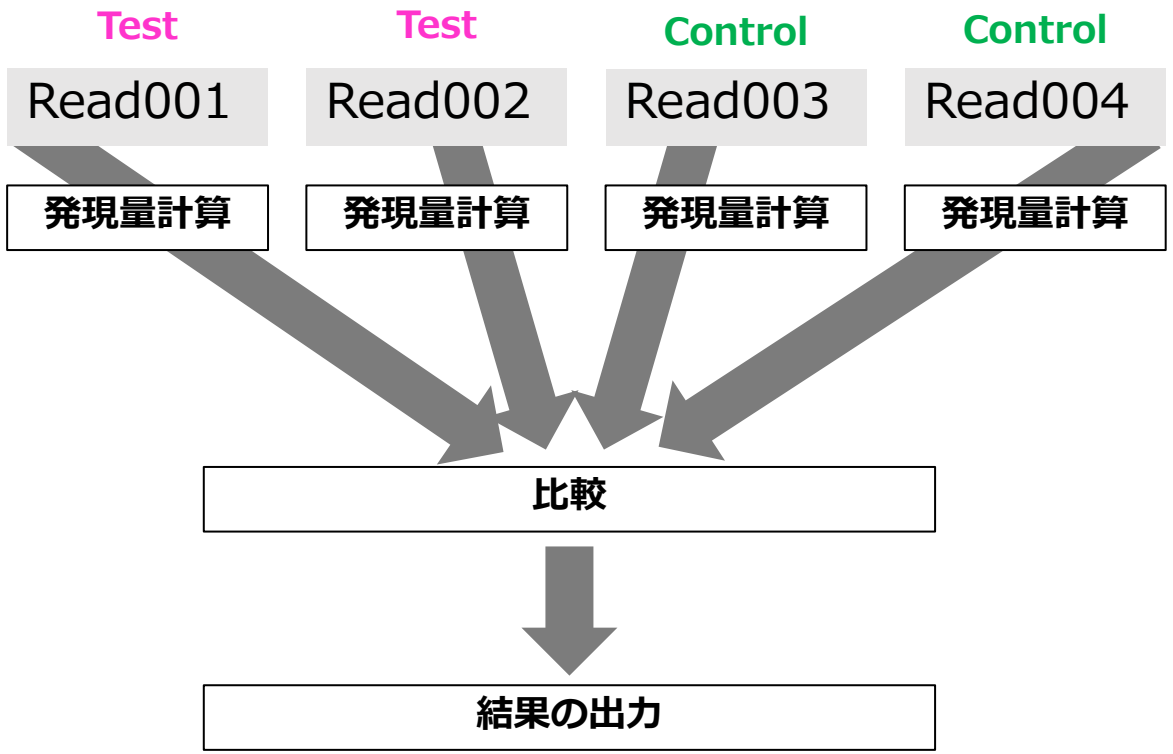
ワークフローを実行



リードとリファレンスの対応に誤りがないかを確認

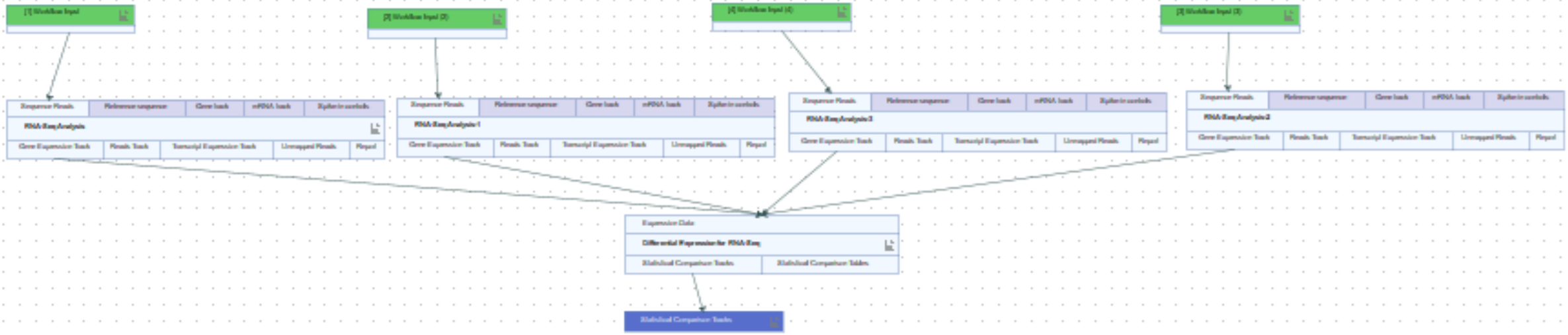
パターン2：個別に発現量の計算を行った後に統合

各リードについて、遺伝子発現量を計算した後、比較を行う場合。



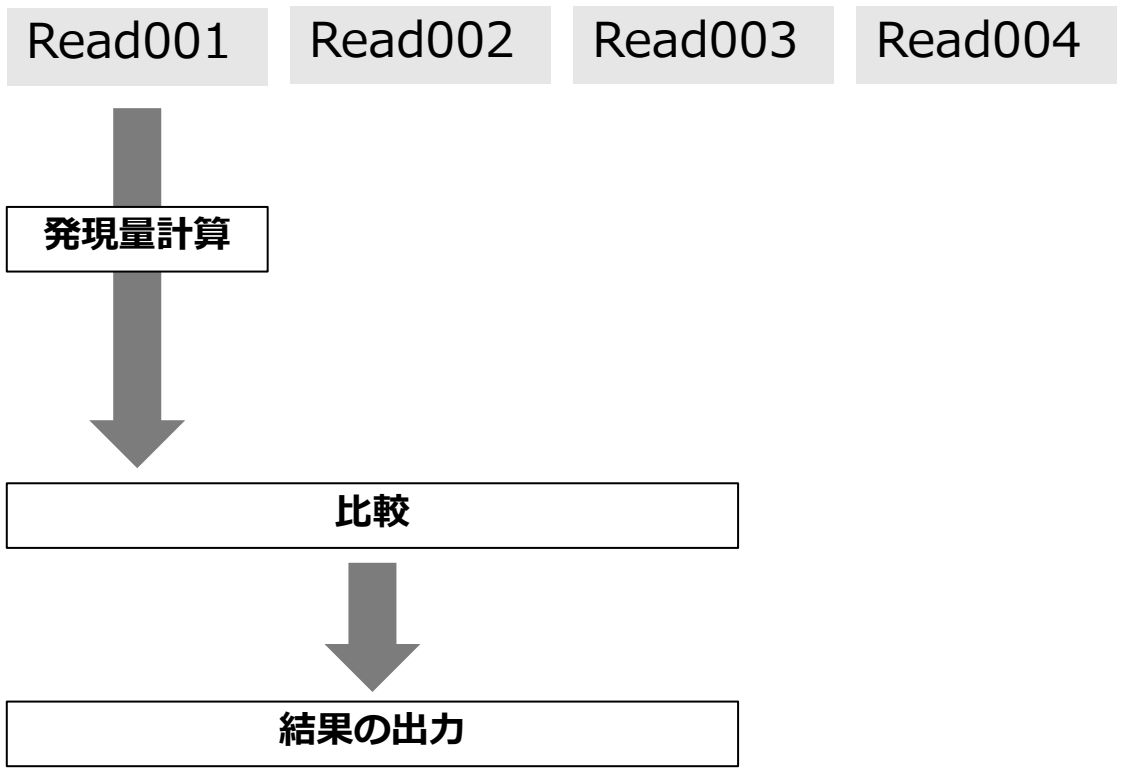
一本道ではないため、単純なワークフローでは対応できない。

推奨されないワークフロー



サンプル数が4のときにしか対応できない

パターン2：個別に発現量の計算を行った後に統合



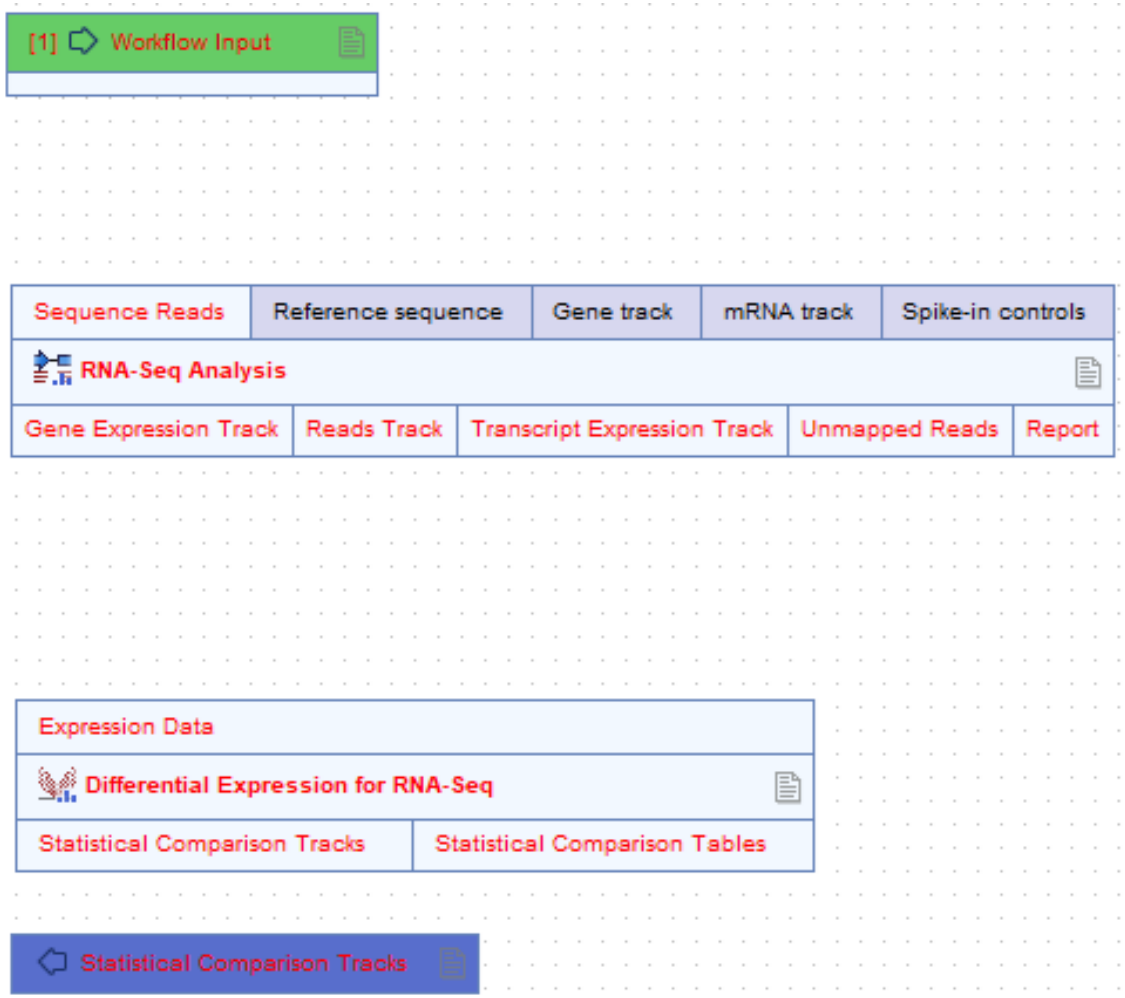
この部分をリードデータの数だけ繰り返す

パターン2：個別に発現量の計算を行った後に統合

事前にメタデータを用意し、インポートする必要がある。

	A	B
1	Name	Group
2	Read001	Test
3	Read002	Test
4	Read003	Control
5	Read004	Control

パターン2：個別に発現量の計算を行った後に統合

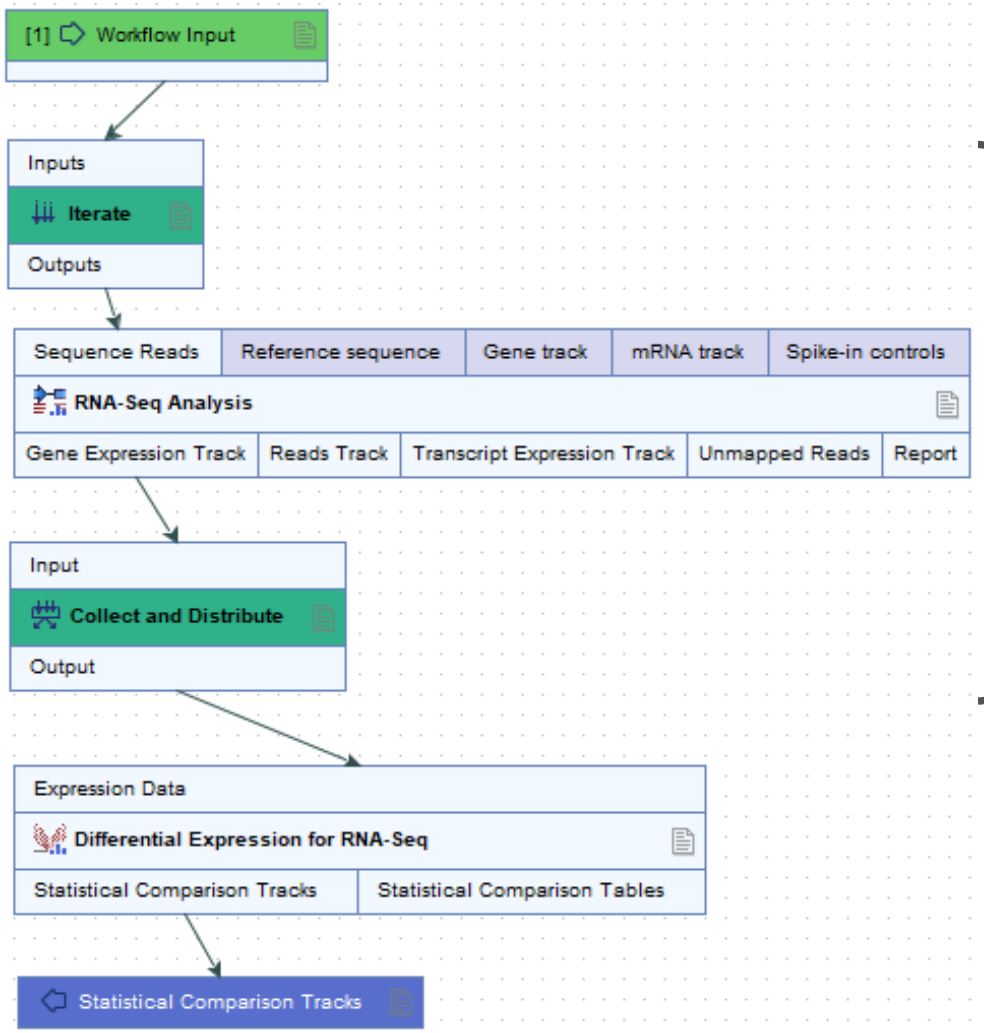


まずは各要素を配置

パターン2：個別に発現量の計算を行った後に統合

つぎに、RNA-seq Analysisを”Iterate”と”Collect and Distribute”ではさむ

パターン2：個別に発現量の計算を行った後に統合

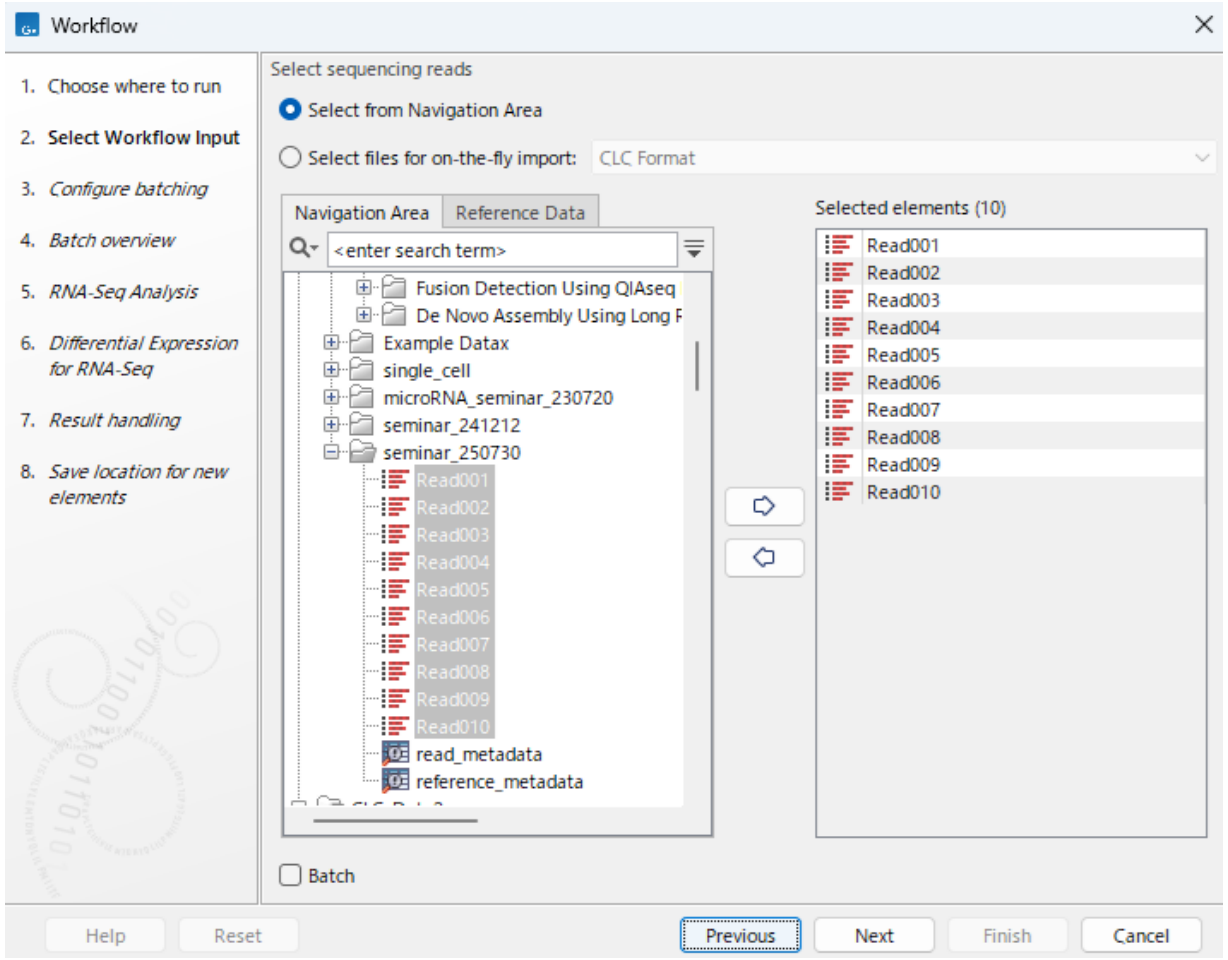


“Iterate”と”Collect and Distribute”で挟まれた部分が繰り返される。

繰り返し部分の解析が全入力に対して終了したら、出揃ったデータを次のツールに入力する。

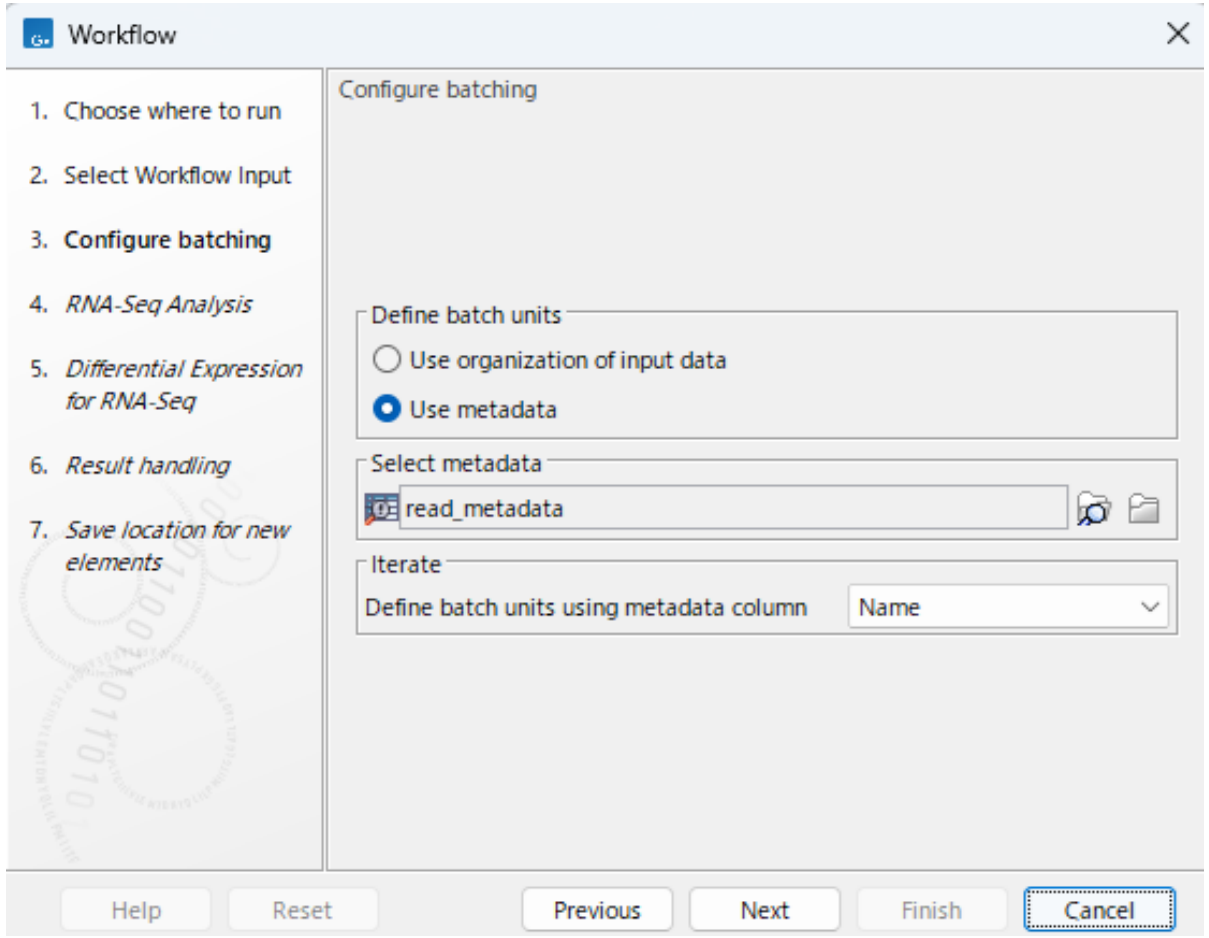
図のように矢印で結ぶ

ワークフローを実行



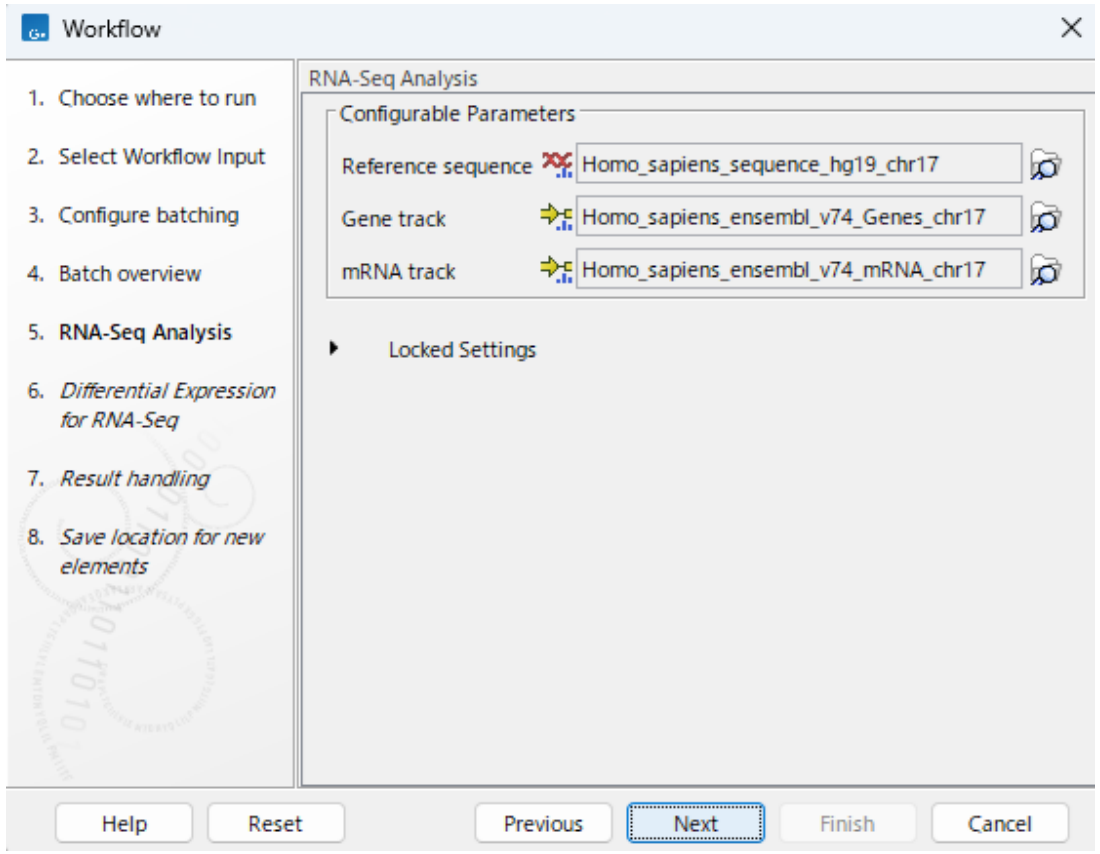
全リードデータを指定する
（”Batch”オプションは指定しない）

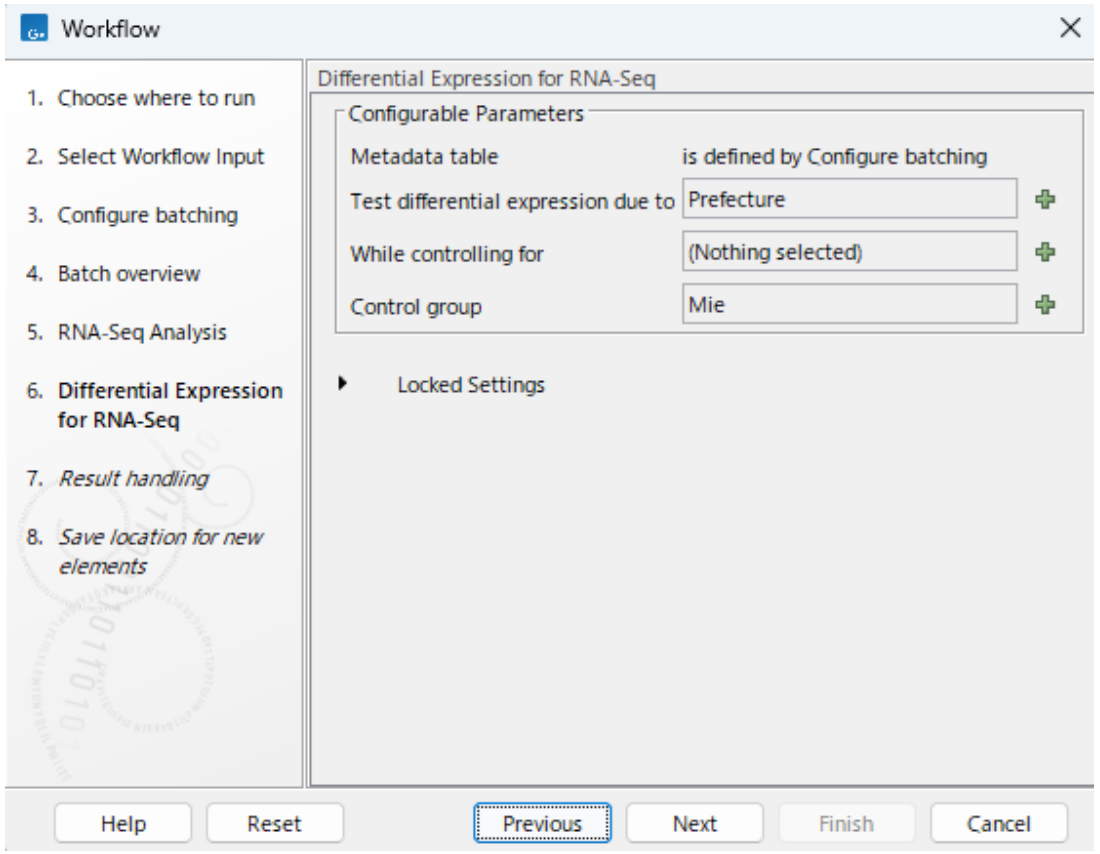
ワークフローを実行



Define batch unitsで”Use metadata”を選択し、
各リードが属する群の情報を記述したメタデータを指定

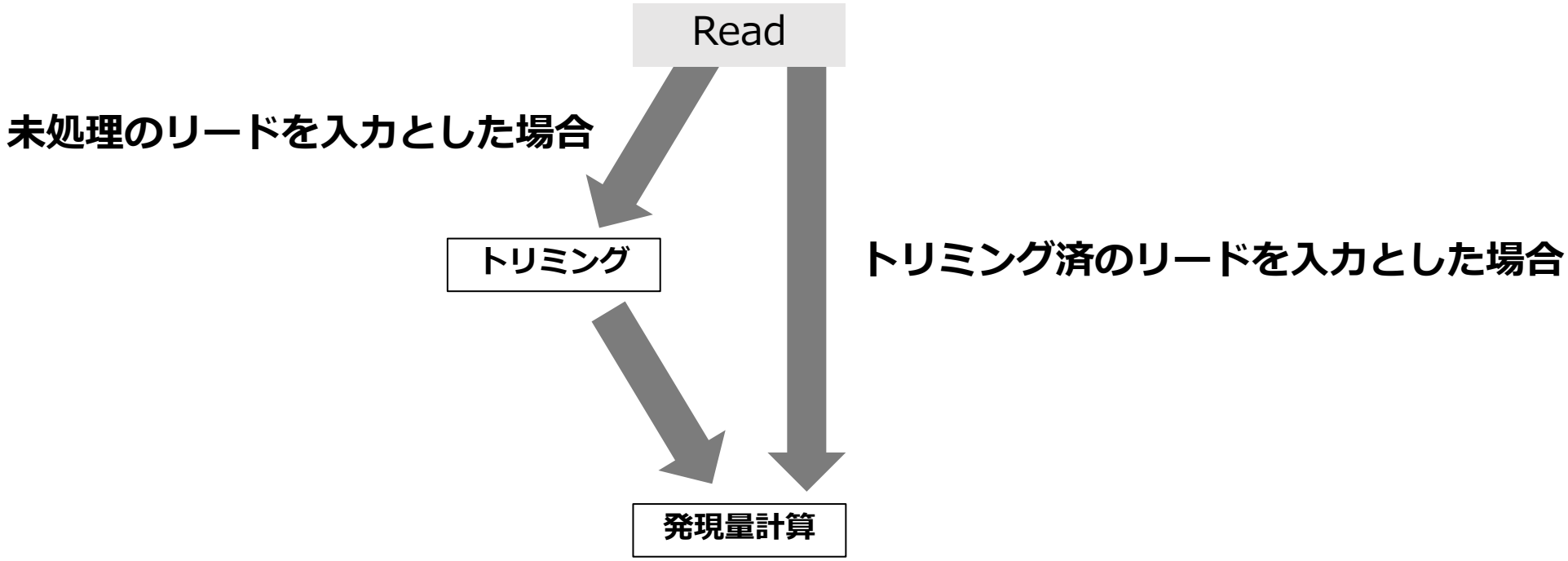
ワークフローを実行



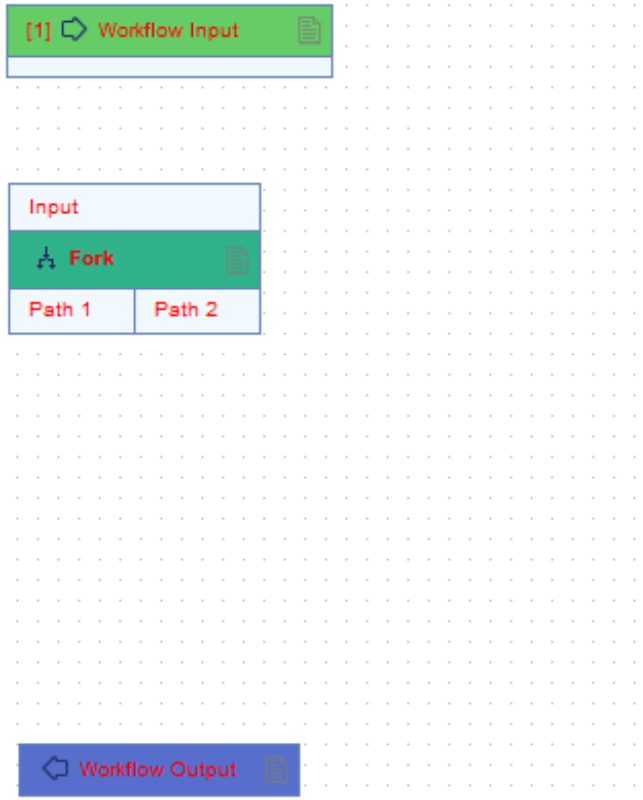


この後は、通常のRNA-seq解析と同様に、
リファレンス配列や群間比較のパラメータを設定する

パターン3: 入力するデータの種類によって処理が異なる場合

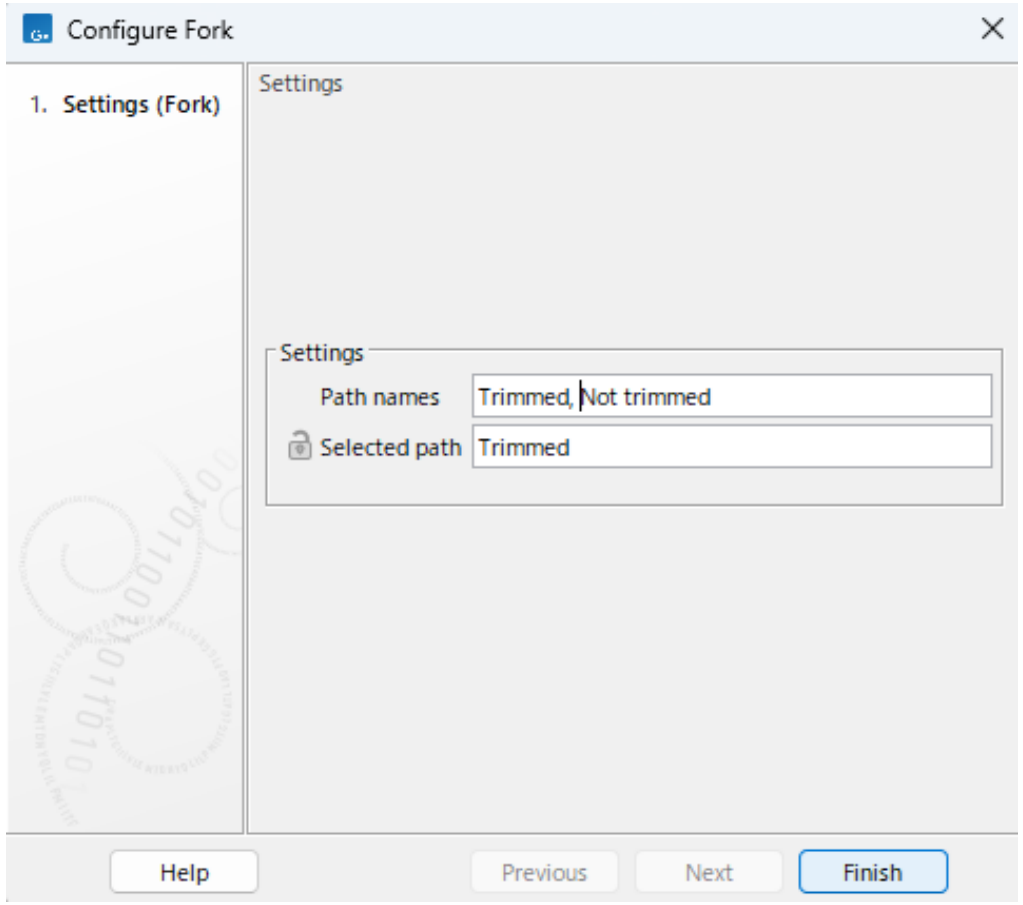


パターン3: 入力するデータの種類によって処理が異なる場合



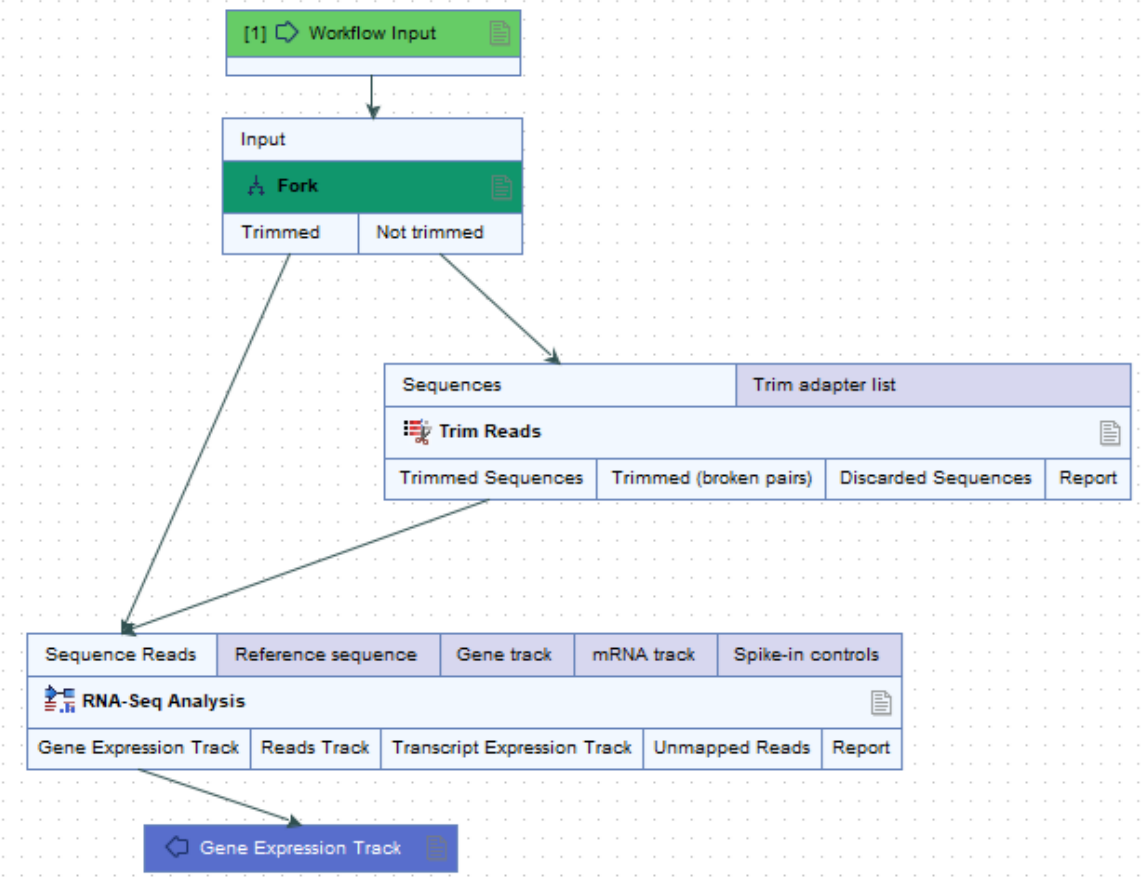
Input要素の下に”Fork”要素を配置

パターン3: 入力するデータの種類によって処理が異なる場合



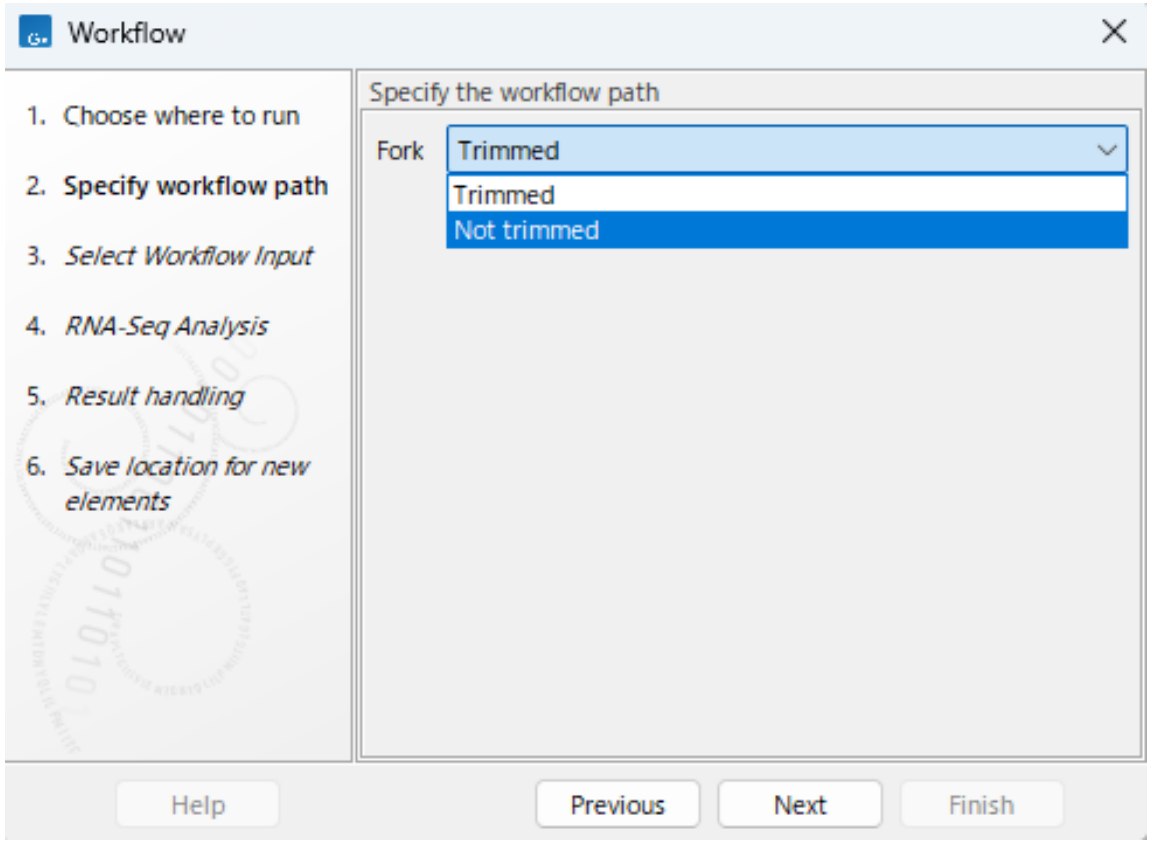
Fork要素をダブルクリックし、
分岐の名前（Path names）とデフォルトのルート（Selected path）を指定
3つ以上の分岐を作ることも可能

パターン3: 入力するデータの種類によって処理が異なる場合



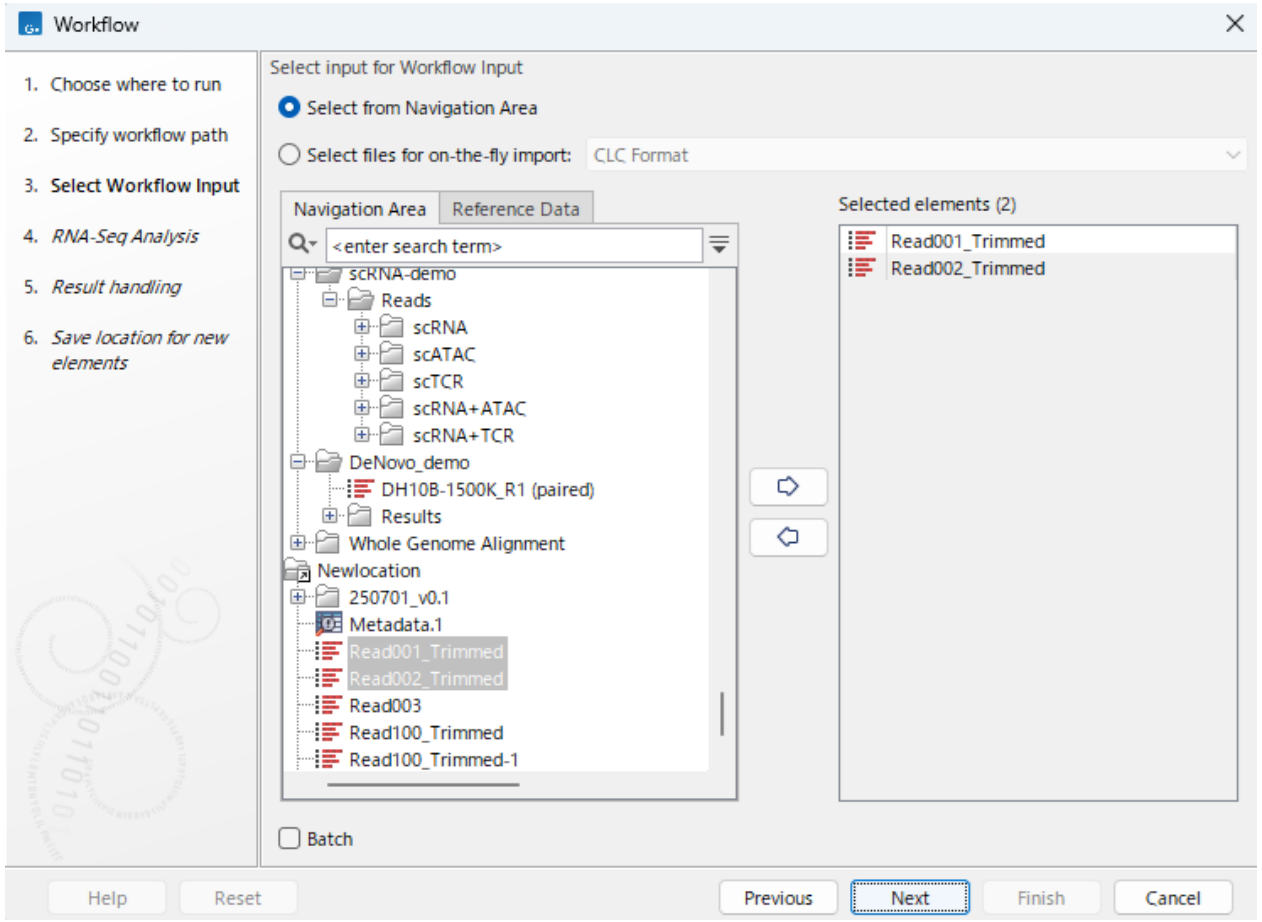
それぞれの経路に応じたツールを配置する

ワークフローを実行



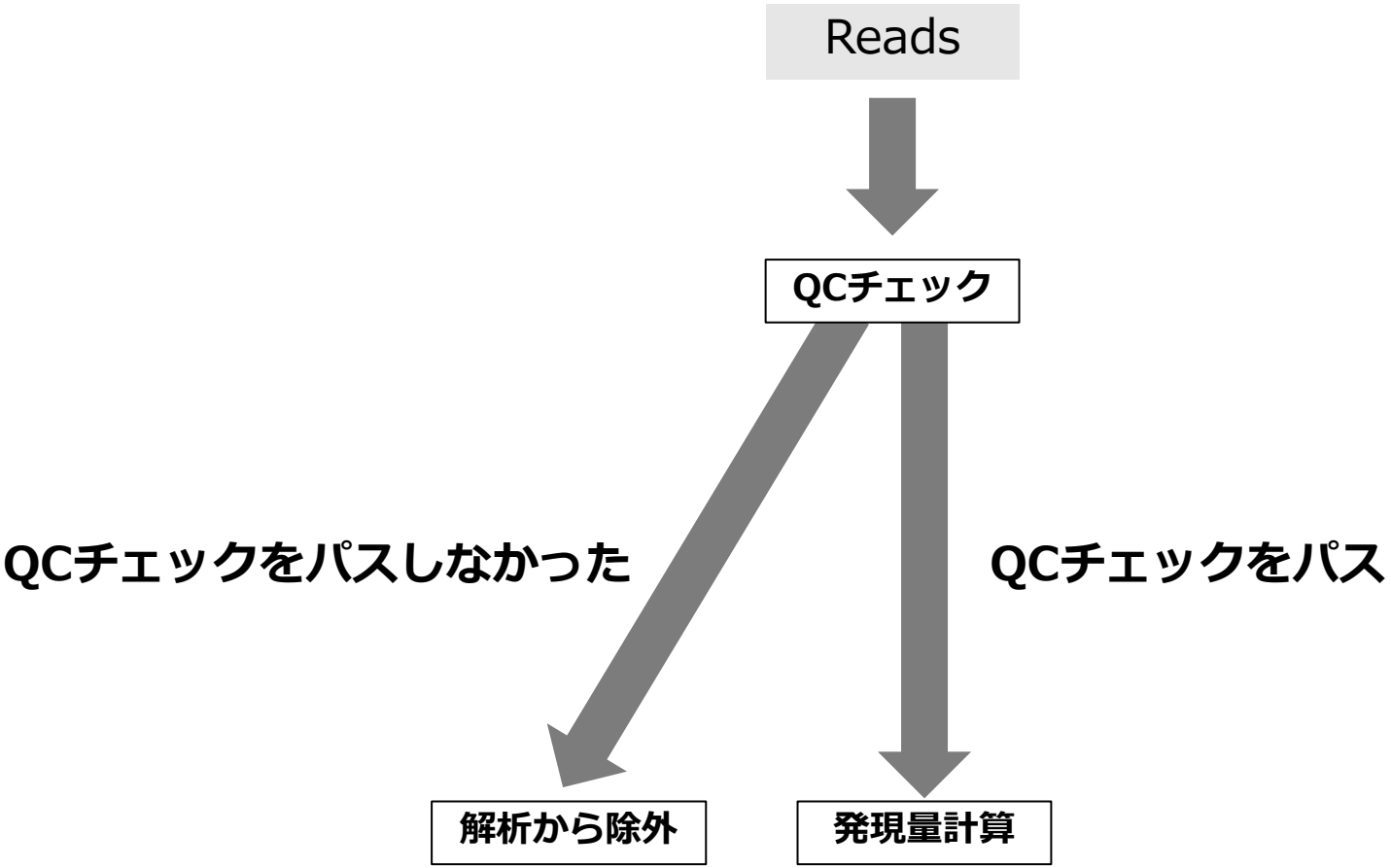
ワークフローを起動し、
はじめにどのルートを使うかを選択

ワークフローを実行

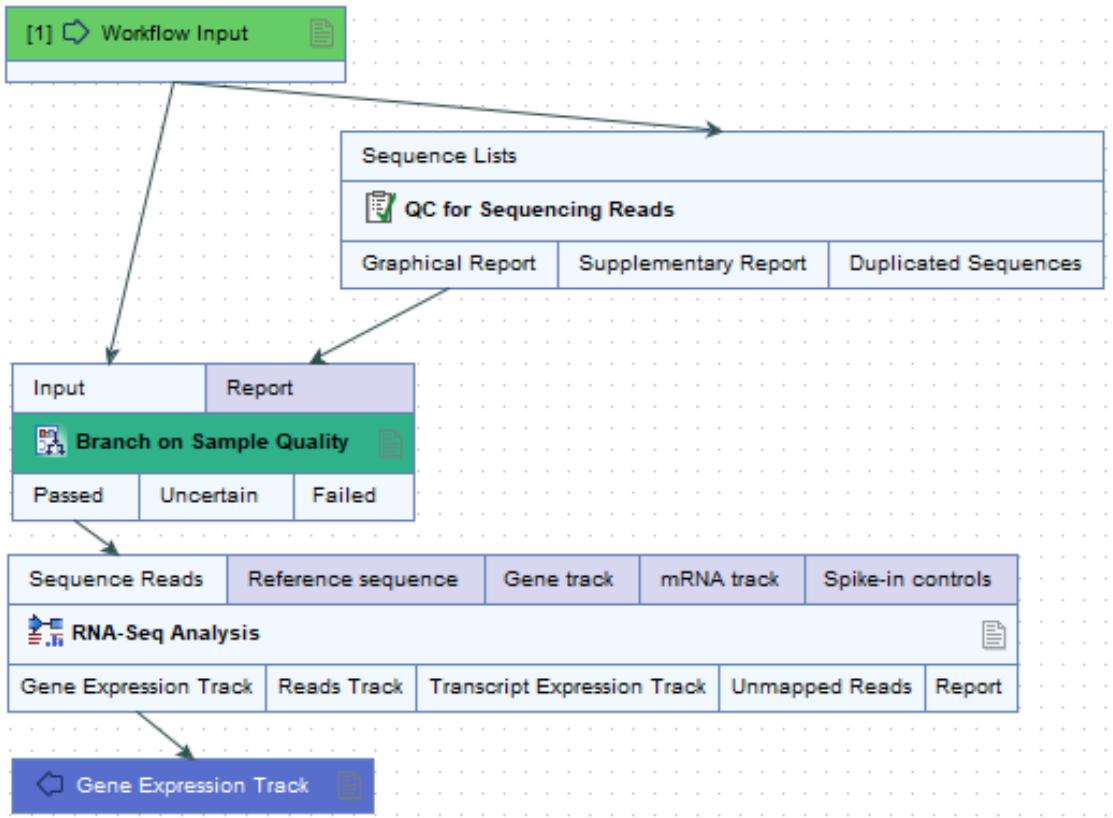


その後は、通常のワークフローと同じく、
入力データと各パラメータを指定する

パターン4: QCチェックの結果に応じて処理を変える



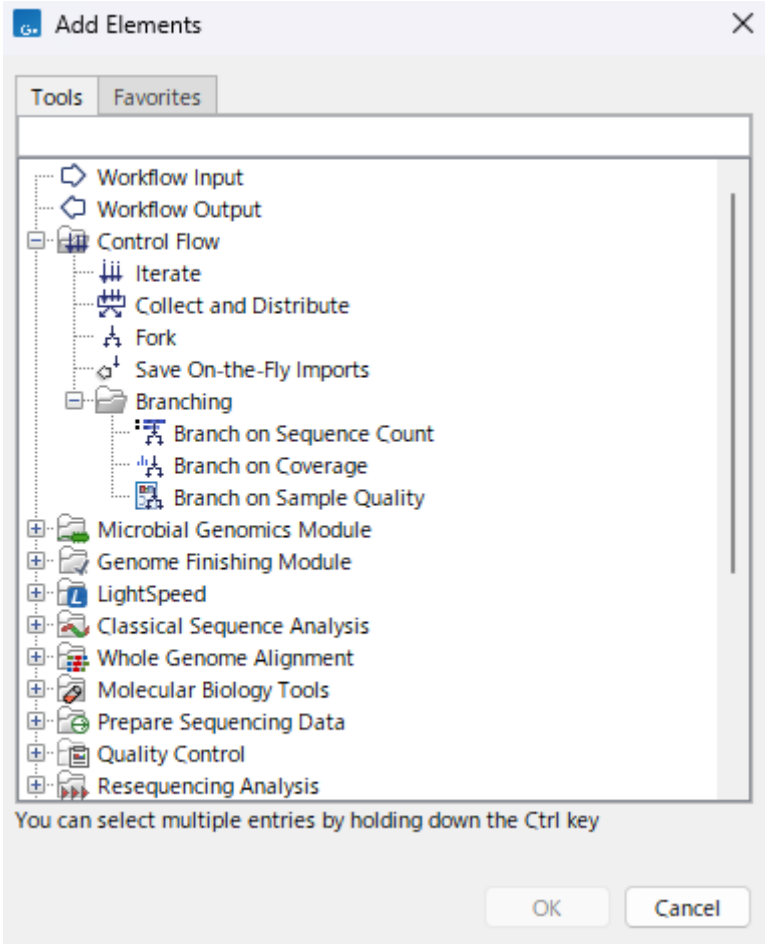
パターン4: QCチェックの結果に応じて処理を変える



“Branch on Sample Quality”要素を使用する

QCチェックでの分岐させるには、QCチェックのレポートが必要なので、
QC for Sequencing Readsツールも追加し、その出力であるレポートも
Branch on Sample Qualityへの入力とする必要がある

パターン4: QCチェックの結果に応じて処理を変える



その他に、カバレッジに応じて分岐（Branch on Coverage）や、
リードの本数で分岐（Branch on Sequence Count）させる要素がある

お問い合わせ先：フィルジェン株式会社

TEL: 052-624-4388 (9:00～17：00)

FAX: 052-624-4389

E-mail: support@filgen.jp