Next-generation sequencing

Lecture 7

Whole genome sequencing

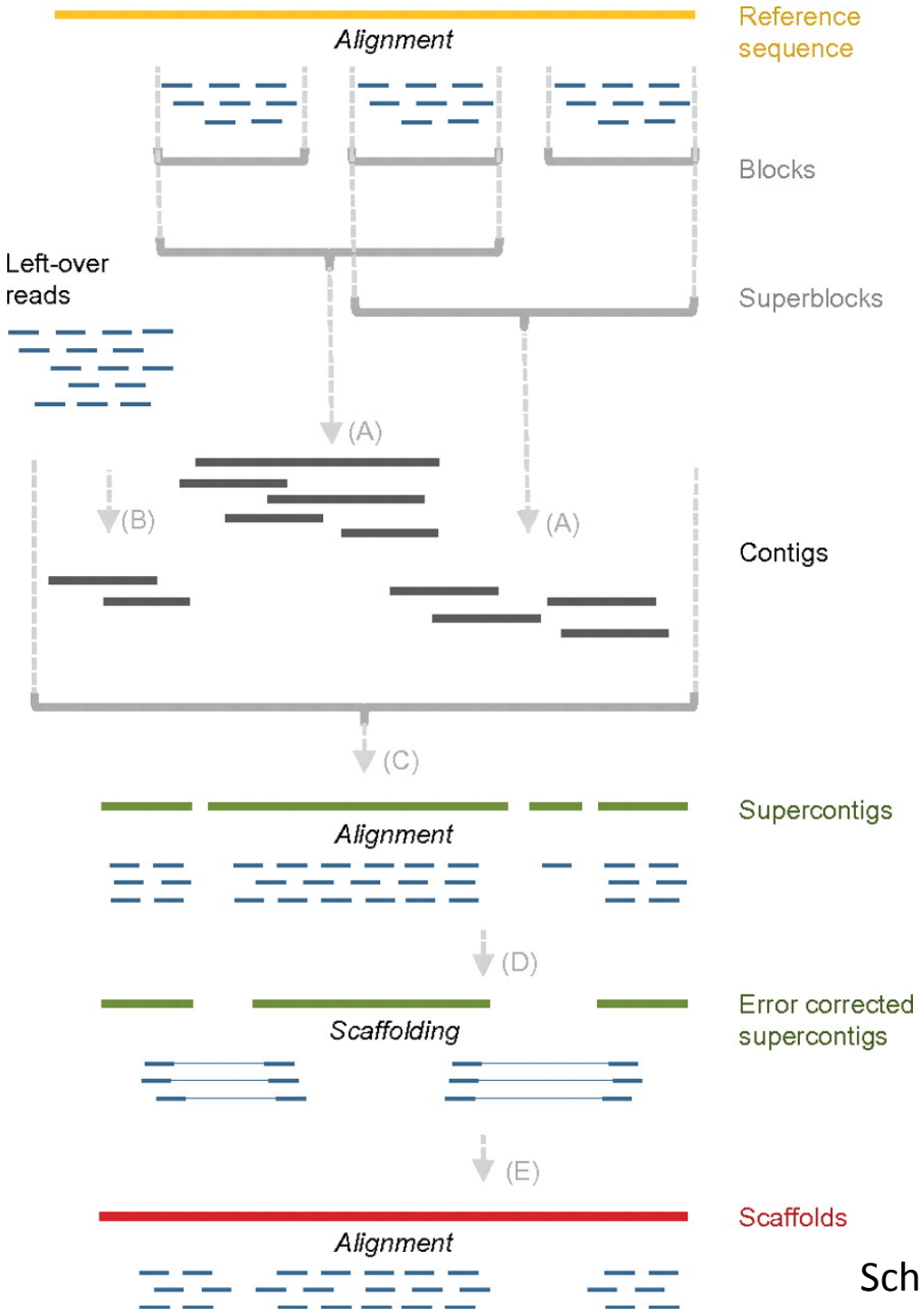
- *De Novo* sequencing
- Mapping assembly (Reference-guided assembly) (Resequencing)

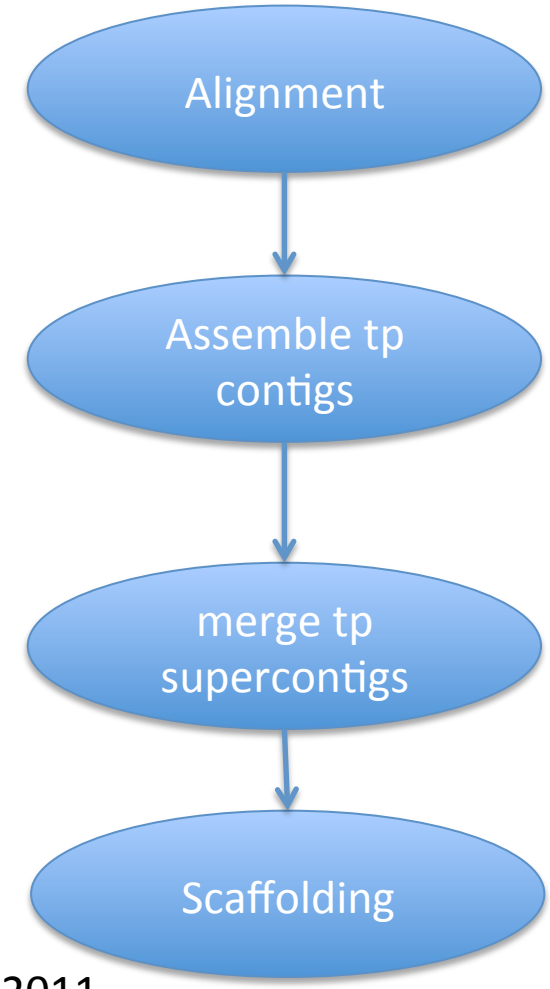
"DNA resequencing is the task of sequencing a DNA region for an individual given that a reference sequence for this region is already available for the specific species. "

Resequencing
(mutation discovery/genotyping)

- A lot of current sequencing effort is spent on re-sequencing genomes of known species
  - Individual humans (1000 Genomes Project)
  - Experimental organisms – looking for genetic variation, copy number variation
- Challenge is to (quickly) align millions of sequence reads to a reference genome with some percent of mismatches
- Problems with repeated sequences – both tandem and dispersed repeats

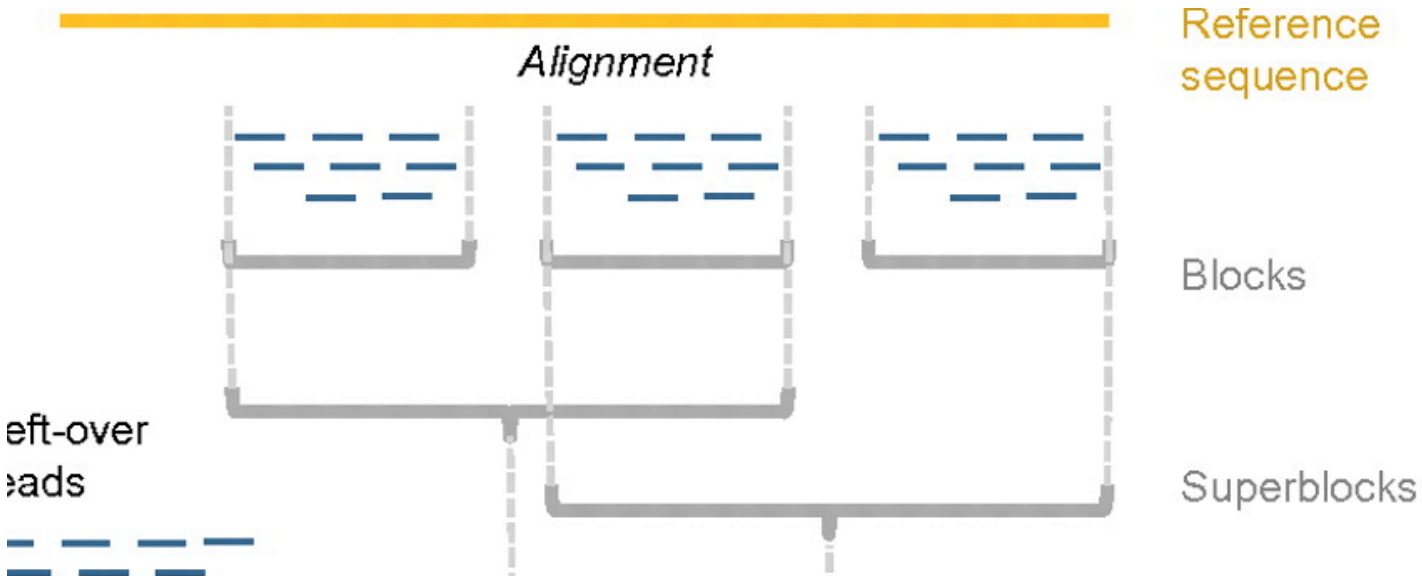
Reference-guided assembly





Schneeberger, 2011

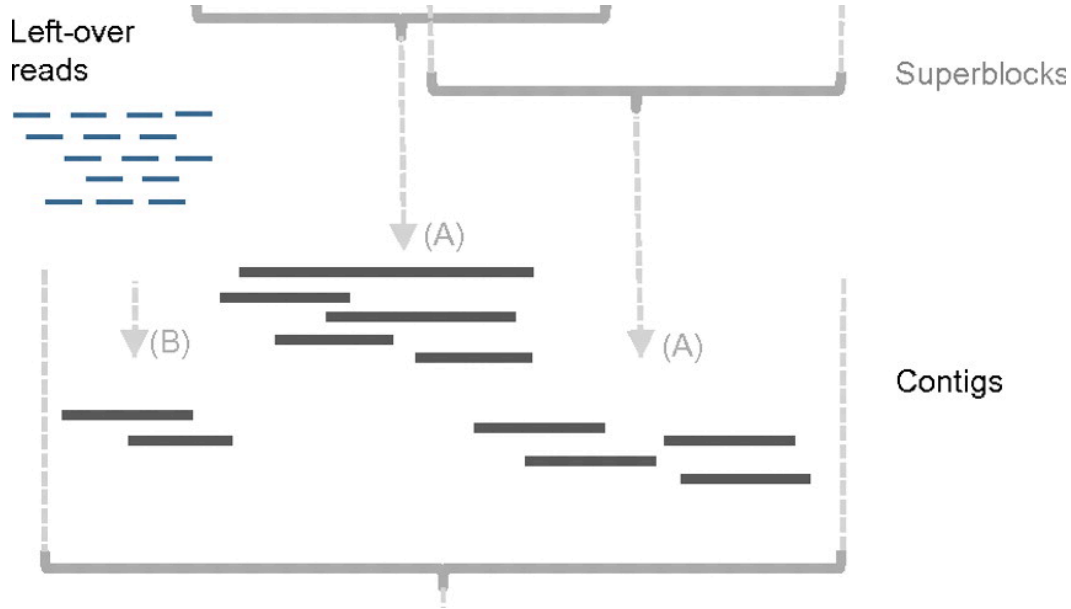
Reference-guided assembly step 1: Alignment



- Align the short reads against the reference sequence with GenomeMapper.
- Adjacent blocks were combined into superblocks, with neighboring superblocks sharing at least one block.

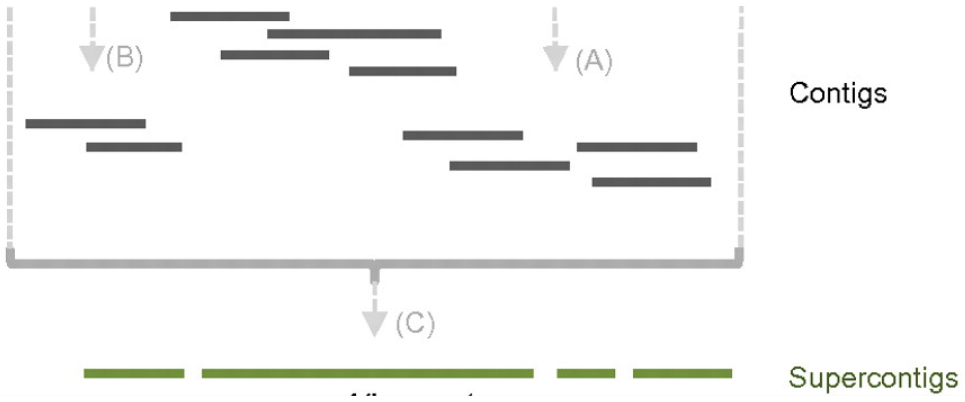
Blocks = regions with constant coverage or adjacent regions connected by aligned mate pairs.

Reference-guided assembly step 2: Assemble to contigs



- Reads corresponding to each superblock were assembled separately using the de Bruijn graph-based assemblers. (Both ABySS and Velvet with eight different kmer sizes).
- All leftover reads (unaligned) are assembled using VELVET, to get nonreference sequences.

Reference-guided assembly step 3: to supercontigs

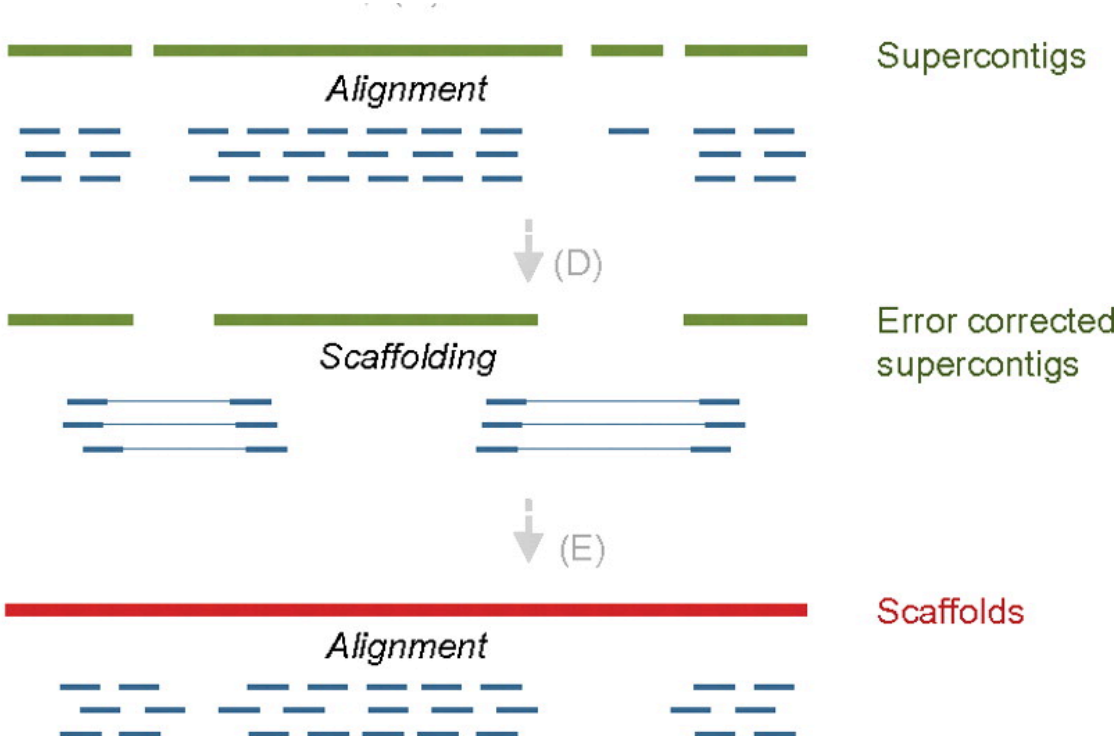


Due to different assemblies, redundancy is introduced into the contigs.

The homology guided Sanger assembler AMOScmp merge all contigs of each chromosome arm into nonredundant supercontigs

Reference-guided assembly

step 4 and 5: Error correction and Scaffolds



1. All original reads are aligned against supercontigs.
2. Differences between supercontigs and reads indicates misassemblies.
3. Correct or remov supercontigs.

- Read pairs with ends that aligned to different supercontigs were used for scaffolding with BAMBUS.

An example
Four Arabidopsis thaliana genomes

- Landsberg erecta (Ler-1), C24, Bur-0, Jro-0 strains

Read statistics

	Bur-0	C24	Kro-0	L*er*-1
		Single end		
Reads	142,532,346	27,033,381	4,443,603	10,076,255
Mb	5,118.6	1,113.2	183.8	550.0
Coverage	42.7x	9.3x	1.5x	4.6x
		Paired end (library 1)		
Pairs	55,811,985	89,737,786	91,624,757	189,763,954
Avg. insert size	187	185	177	178
SD	24	27	17	23
Mb	4,094.9	7,210.9	8,124.6	26,774.8
Coverage	34.1x	60.1x	67.7x	223.1x

- 2 libraries (one single end and one paired end)
- Insert size 180 bp
- Read length 36-80 bp
- 30x – 200x coverage

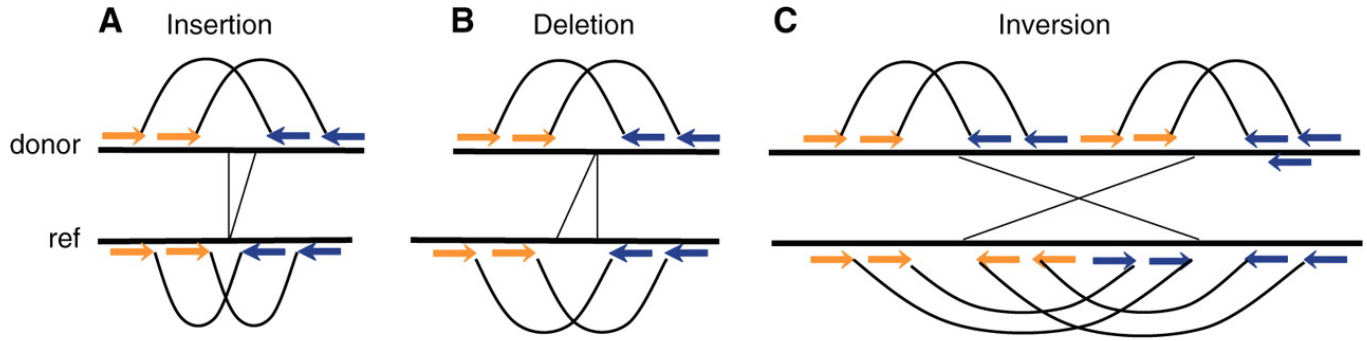
Assembly statistics

	Bur-0	C24	Kro-0	Ler-1
Coverage	83.2x	75.0x	72.7x	322.4x
Libraries	2	2	2	2
N50 (kbp)	193	109	161	297
Scaffolds	2526	2052	2670	1528
Total Length (Mbp)	101	101.3	99.9	100.8
Longest Scaffold (Mbp)	4	3.6	5.1	1.3

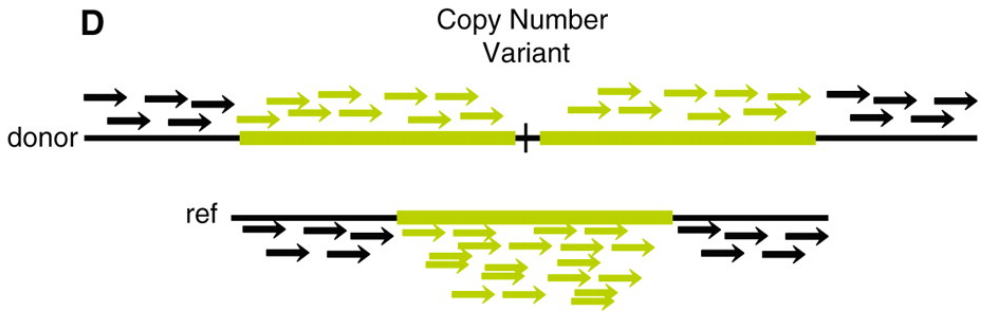
Ref genome 105.2Mbp

Variant discovery

Recent advances in sequencing technology make it possible to comprehensively catalog genetic variation in population samples, creating a foundation for understanding human disease, ancestry and evolution.

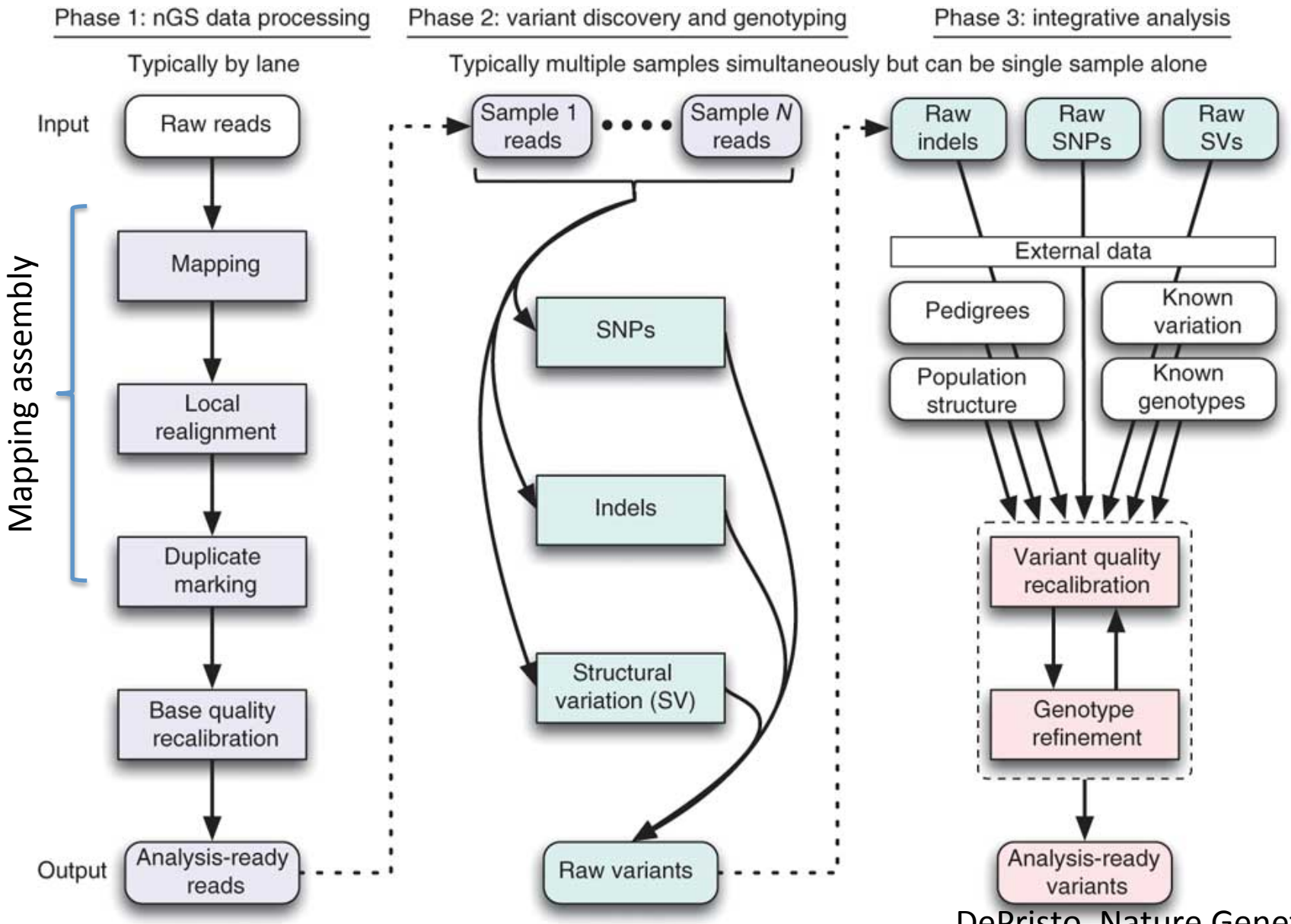


1. SNP and Micro-Indel
2. Structural Variants



Dalca and Brudno, 2010

A framework for Variant discovery



DePristo, Nature Genetics, 2011

Find variant with genome comparison

Table 3. Variants of different lengths in L*er*-1

	Deletions		Insertions	
Variant length (bp)	*n*	Length (bp)[†]	*n*	Length (bp)[†]
1	35,370	35,370	34,261	34,261
2	9,861	19,722	10,060	20,120
3–4	8,305	28,221	7,963	27,148
5–8	5,816	36,809	5,677	35,766
9–16	3,757	43,673	3,505	40,435
17–32	1,824	41,552	1,238	27,800
33–64	663	30,310	579	26,413
65–128	296	26,190	340	29,810
129–256	219	40,825	127	21,676
257–512	204	74,045	63	22,600
513–1,024	240	176,491	20	12,823
1,025–2,048	160	223,702	2	3,376
>2,048	208	996,542	4	16,129