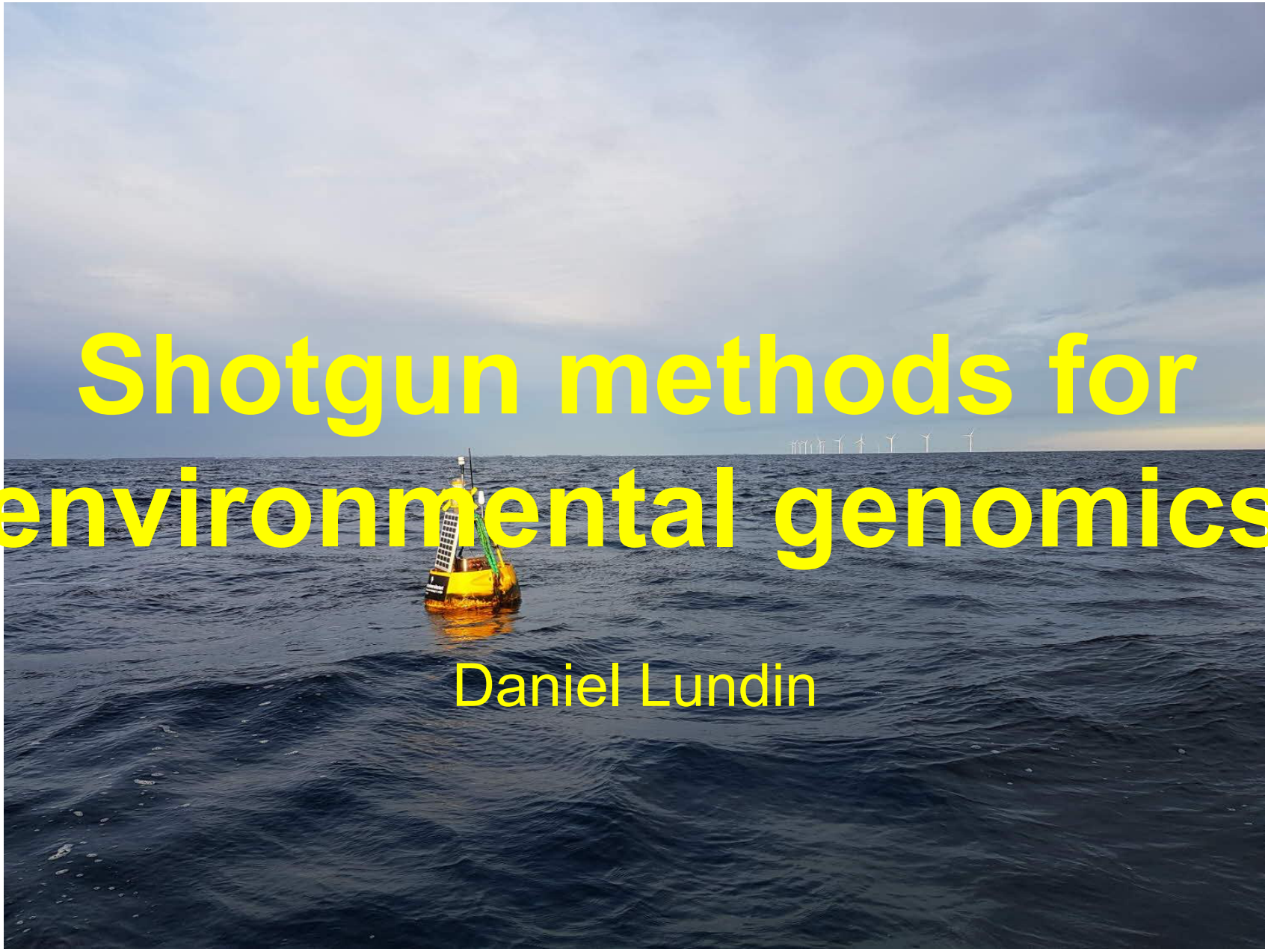


Shotgun methods for environmental genomics

Daniel Lundin

Linnæus University | Centre
Ecology and Evolution in Microbial model Systems

Shotgun sequencing

Community DNA

- View of functional capacity of community
- Taxonomic distribution of functions
- Breakdown into individual (hypothetical) genomes: Way to sequence uncultured organisms

Single cell sequencing

- Another alternative for uncultured organisms

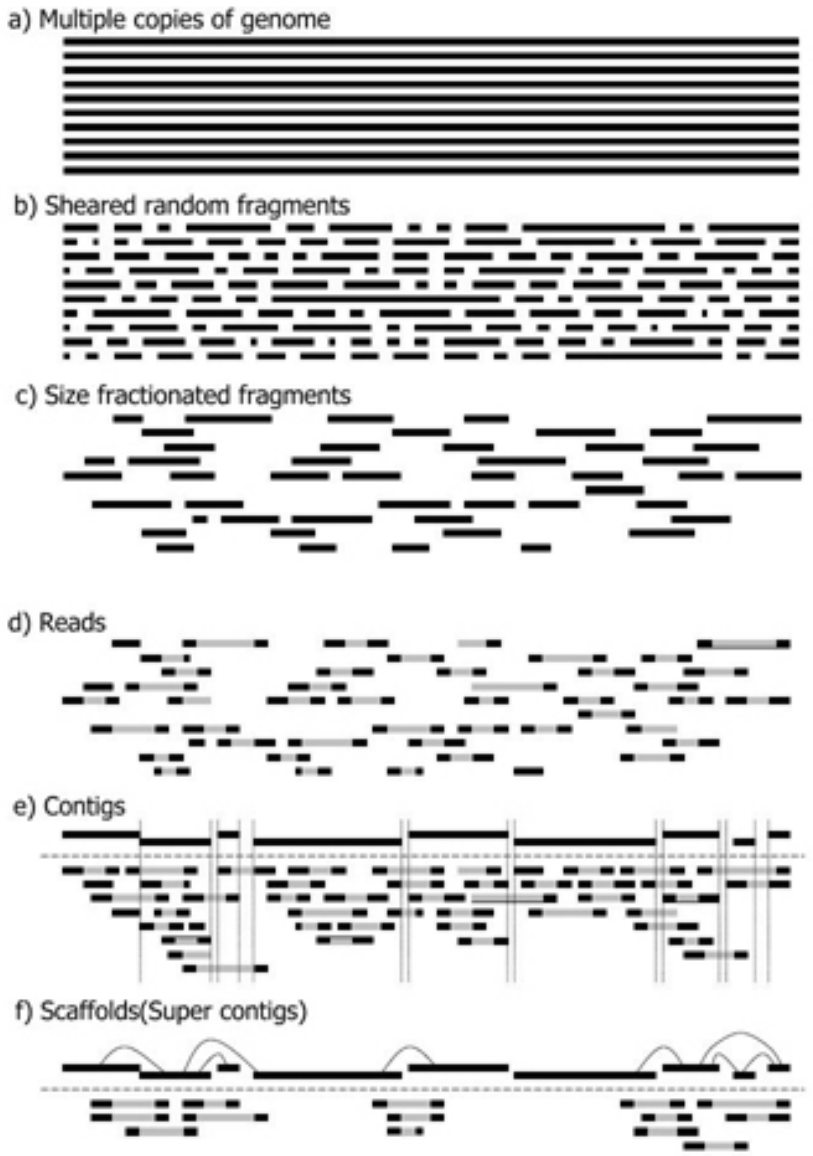
Community RNA

- mRNA gives snapshot of activities within a community
- Taxonomic distribution of activity
- rRNA could replace or complement amplicon sequencing

Genomics – A methods primer

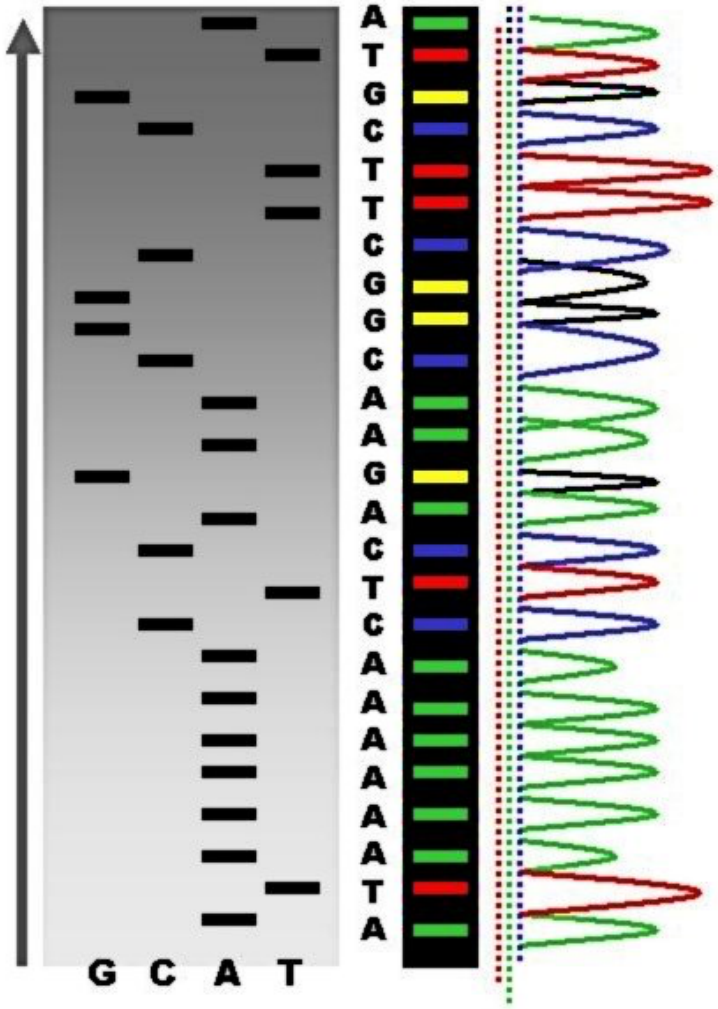
WGS: Whole Genome Shotgun sequencing

- Sequencing methods yields short reads (30-1000 b (DNA bases))
- A genome is assembled from an enormous number of sequence reads
- Sequence reads are often generated in pairs from longer DNA fragments → contigs can be connected into scaffolds



Masahiro Kasahara and Shinichi Morishita. *Large-scale genome sequence processing*. Imperial College Press (2006)

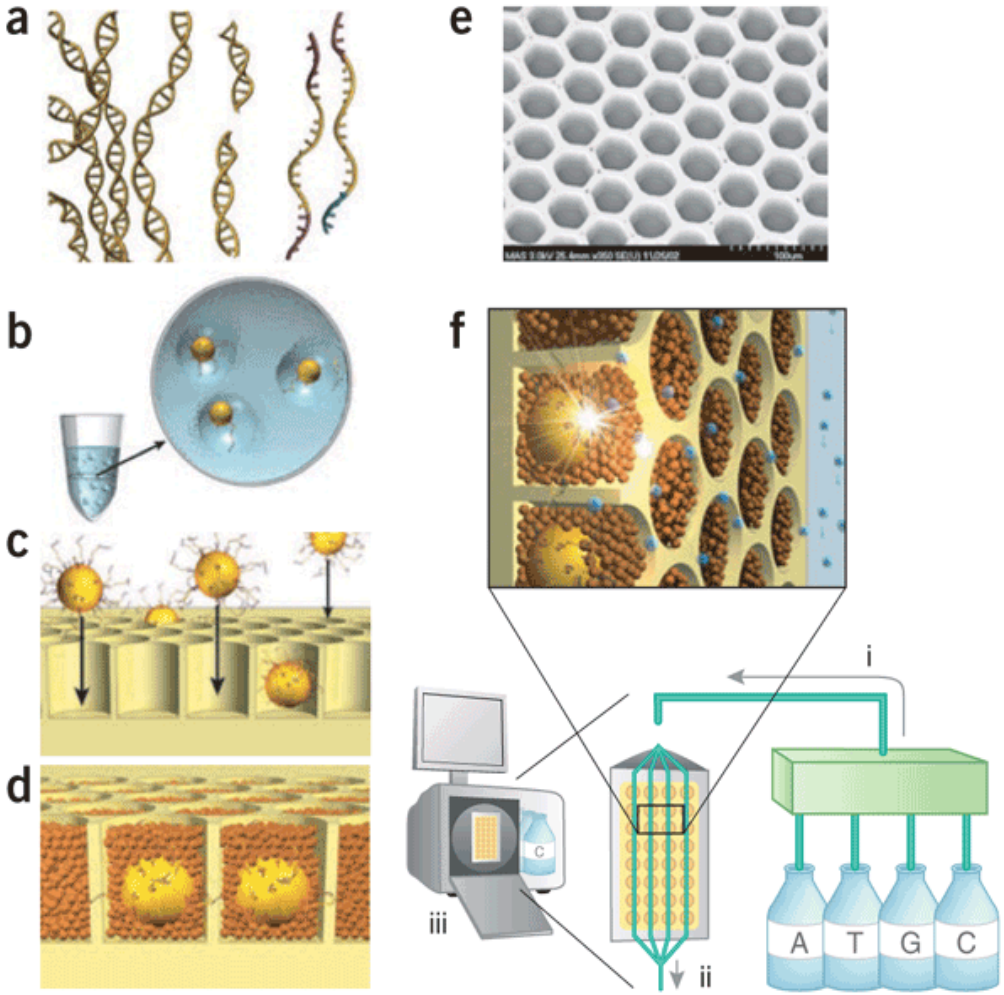
Sanger sequencing



- Good quality
- Long reads, up to 1000 b
- By comparison very little data
  - Up to 384 sequences read in parallel
- Expensive per base

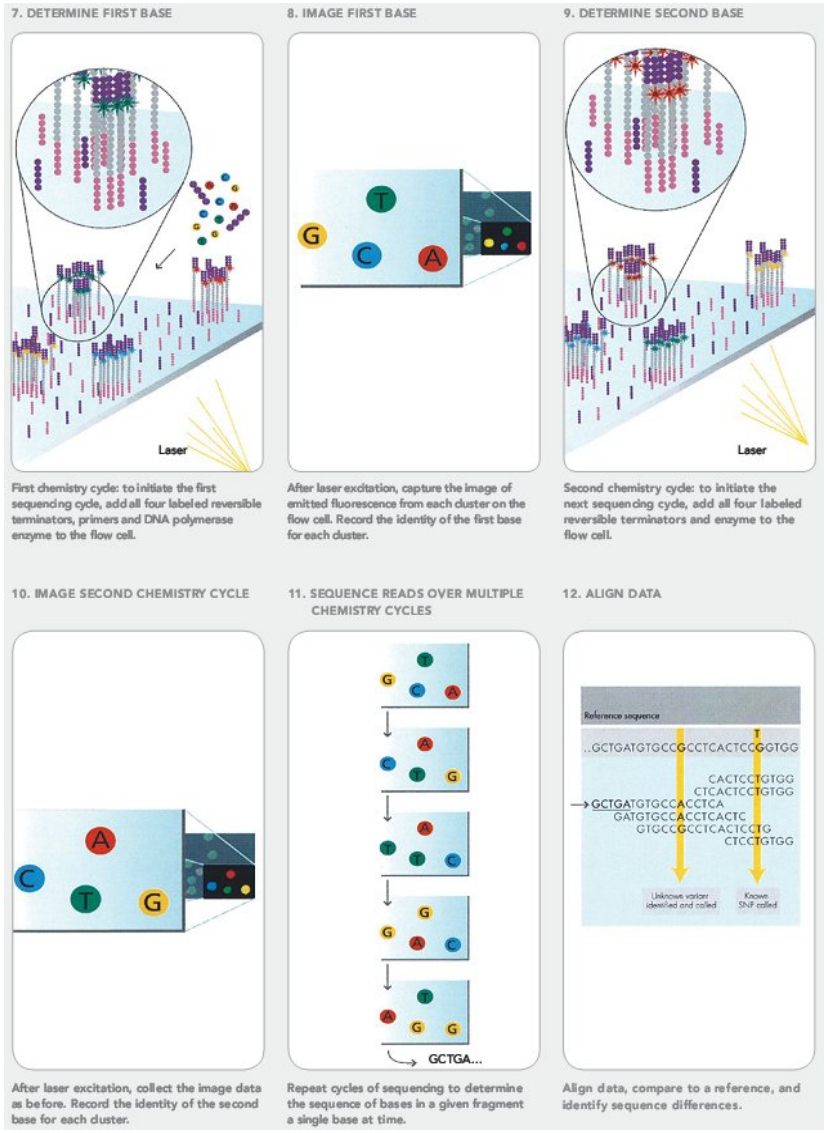
454 and IonTorrent

- Based on pyrosequencing
- About 1 million sequences in parallel
- 450, 700b or longer
- Expensive chemicals
- More errors compared with Sanger
  - Homopolymers
- Ion Torrent is similar
  - Shorter reads (100-200b, 400b?)
  - Much less expensive (USD1000 per run)



Rothberg & Leamon, Nature Biotechnology 26, 1117 - 1124 (2008)

Illumina



- Quite short reads: 2x80-150 b (HiSeq), 2x300b (MiSeq)
- 3 billion reads per flowcell (8 lanes; HiSeq), 20 million reads (MiSeq)
- Much cheaper than 454
- More error compared both with Sanger and 454
  - Error distribution different to 454, single base errors more frequent, homopolymer errors rare

PacBio SMRT

Single Molecule Real-Time Sequencing

- Nanotechnology:
  - Observe a single nucleotide being added to polymer (fluorescence)
- Long reads: 5-7 kb average
- 150,000 concurrent sequences
- 500 Mb per run
- Often combined with Illumina

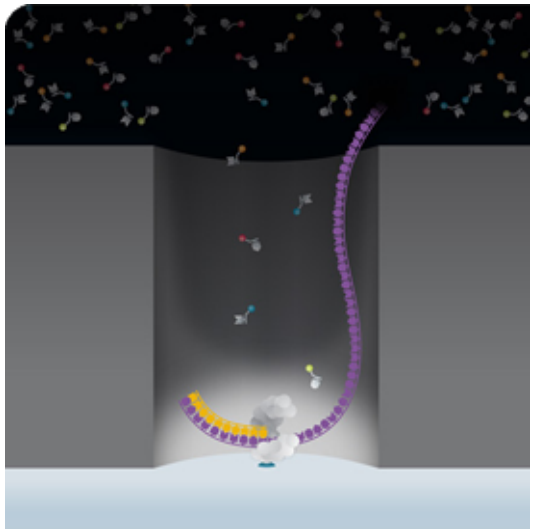


Image from
http://www.pacificbiosciences.com

(Numbers from Wikipedia article “Single molecule real time sequencing”, March 2014.)

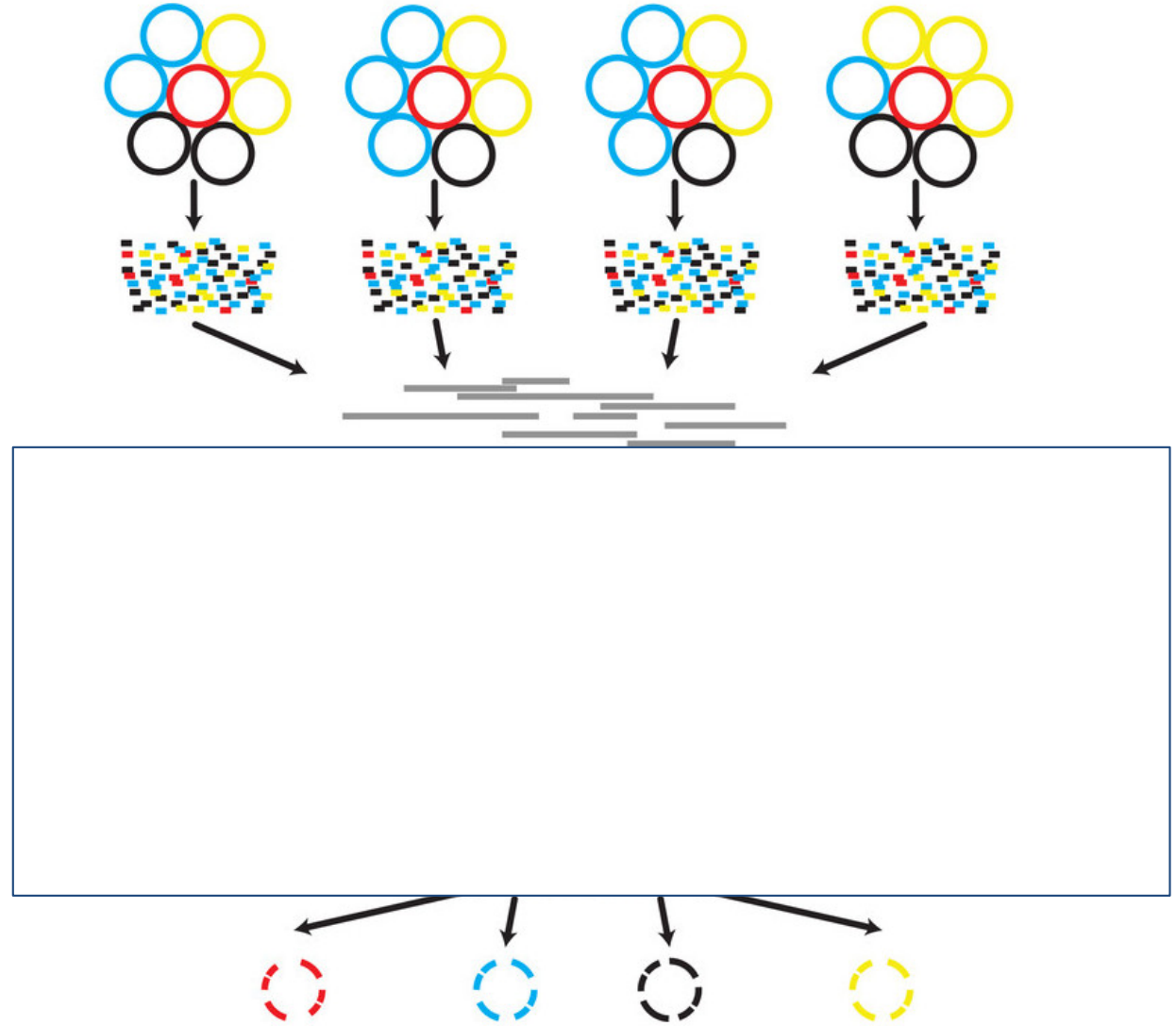
Sequencing methods: A comparison

Method	N. sequences	Bases per run	Typical read length	Cost (USD/Gb)
ABI Sanger		70 kbp	700	2,300,000
454		500 Mbp	450	15,500
Ion Torrent – Proton I	80 M	16 Gbp	200	62
Illumina HiSeq 2500 (high output)	2000 M	500 Gbp	2x125	30
Illumina MiSeq	15-25 M	4.5 Gbp	2x300	100-225
PacBio SMRT Sequel	0.4 M	3.8 Gbp	10000	180
Oxford Nanopore MinION	0.6 M	6 Gbp	10000	166.67

Numbers from Glenn, T., NGS Field Guide, The Molecular Ecologist blog.

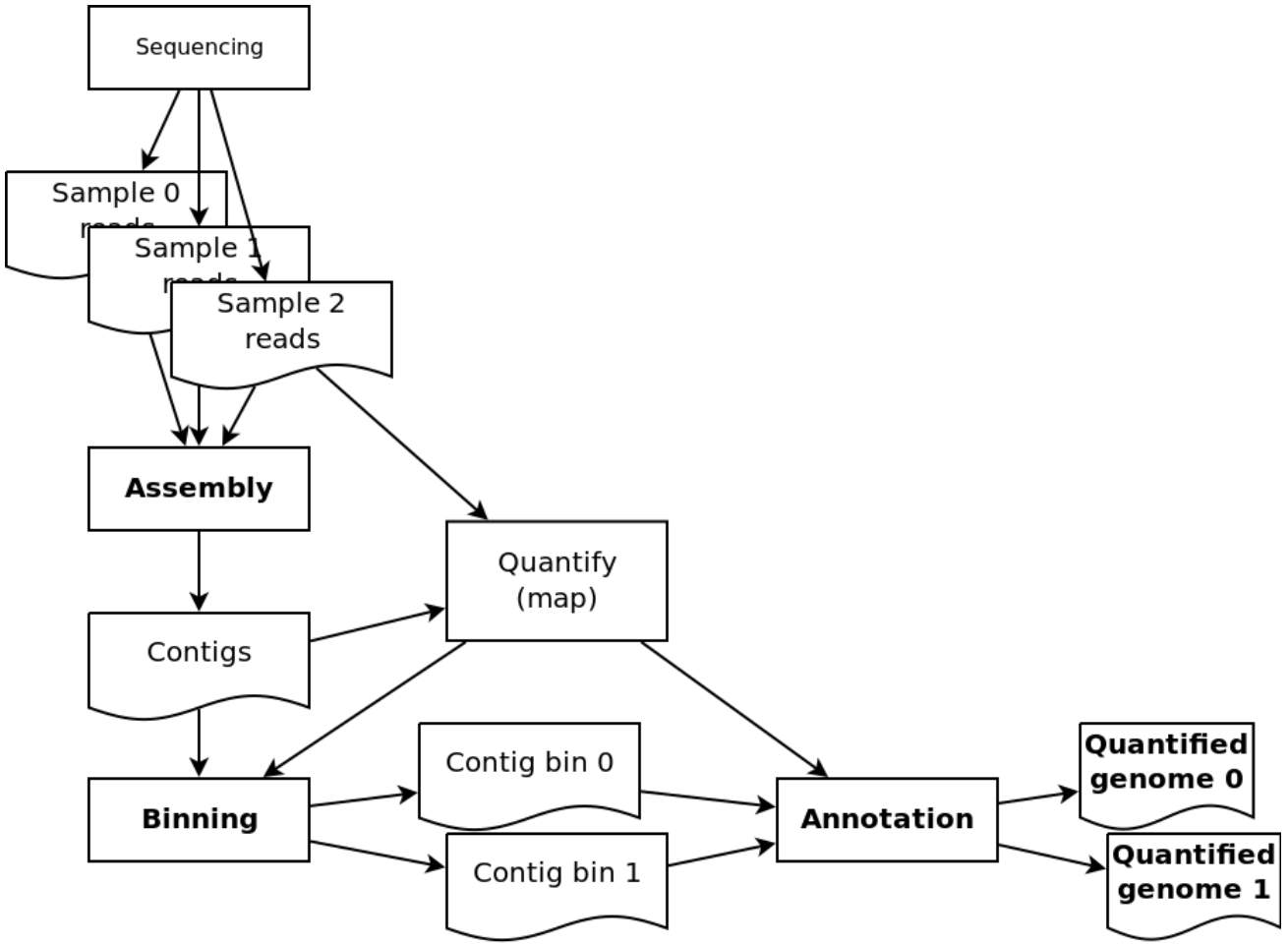
Principles for annotation of shotgun data

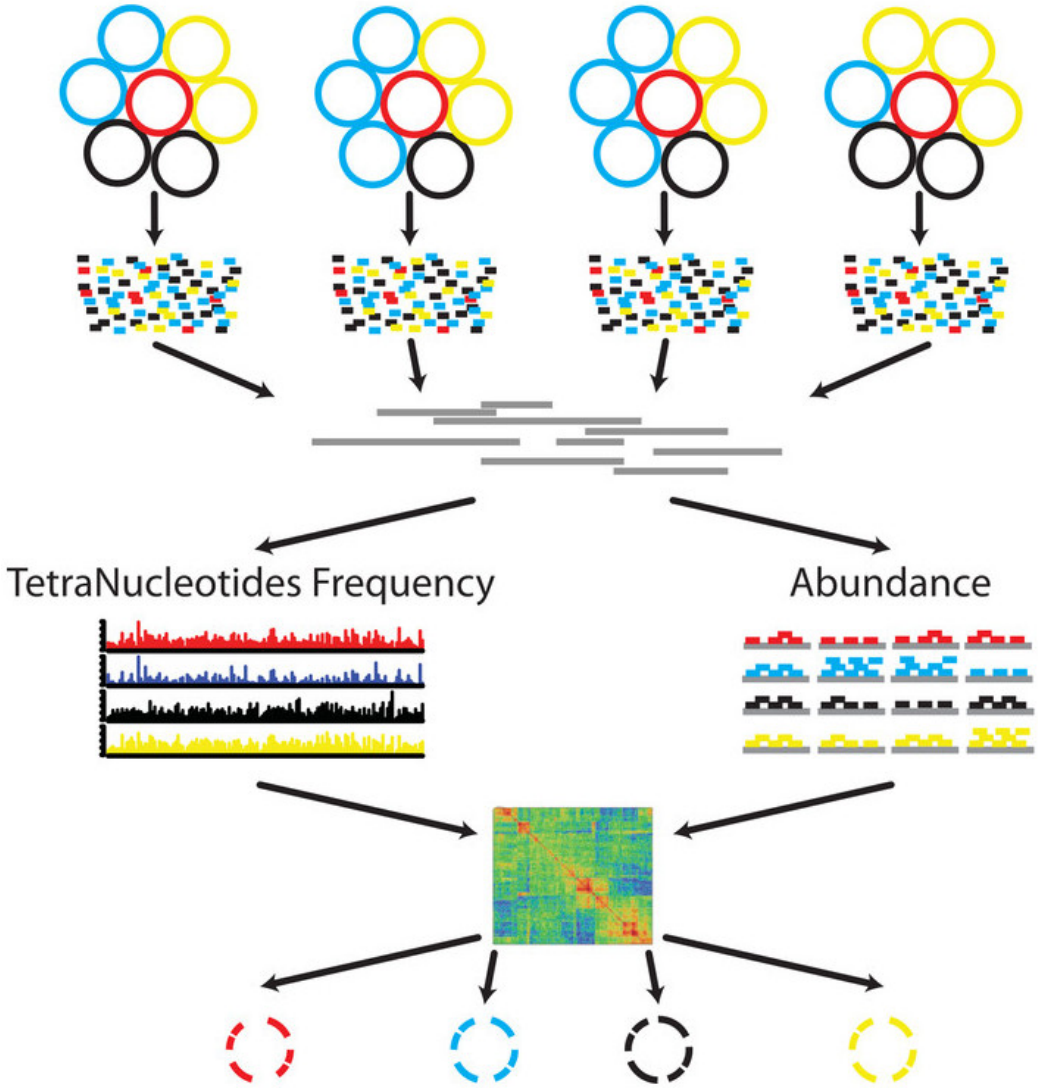
- Read-based
- Assembly-based
- Gene-specific



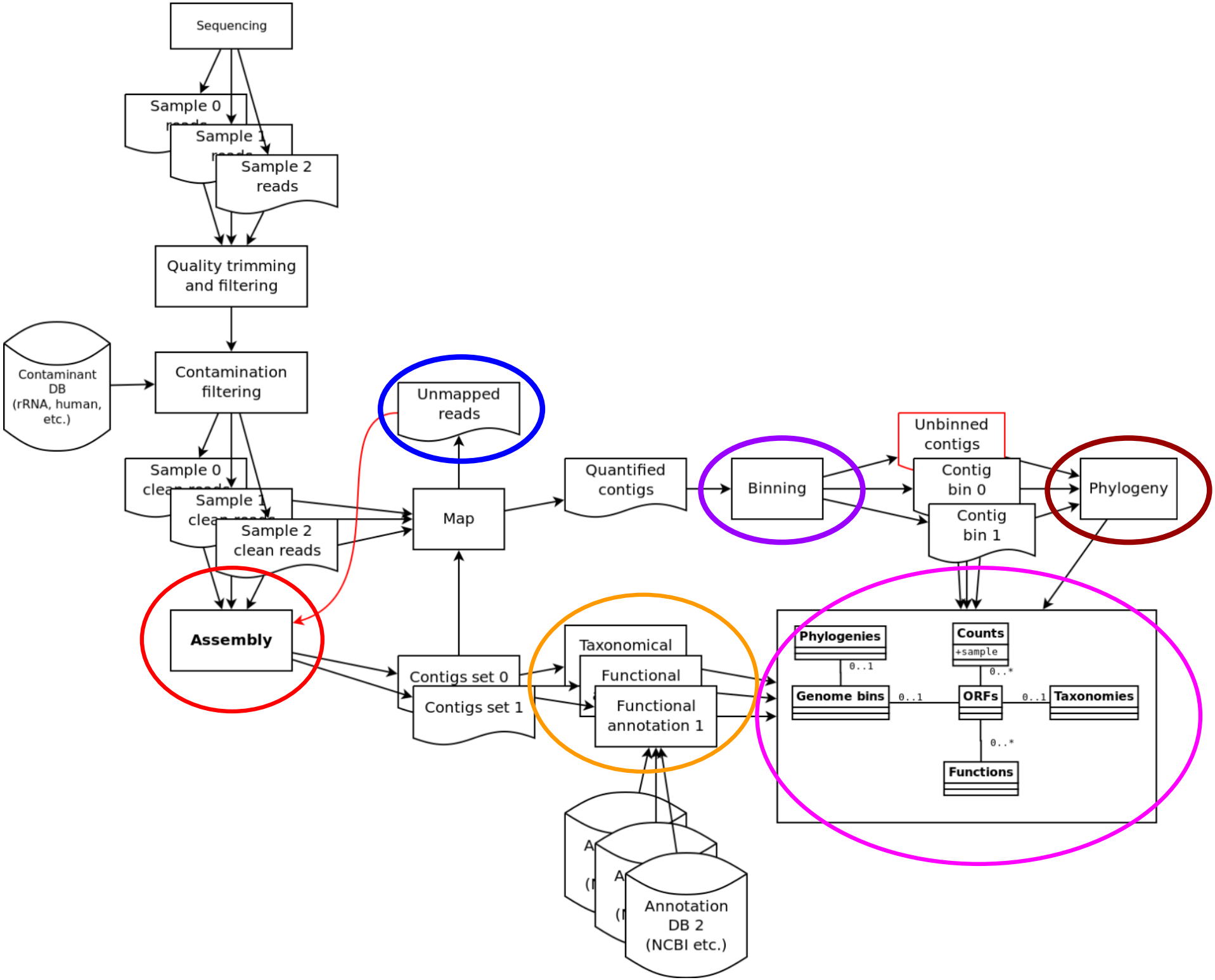
Kang et al., PeerJ, 2015

Assembly-based annotation

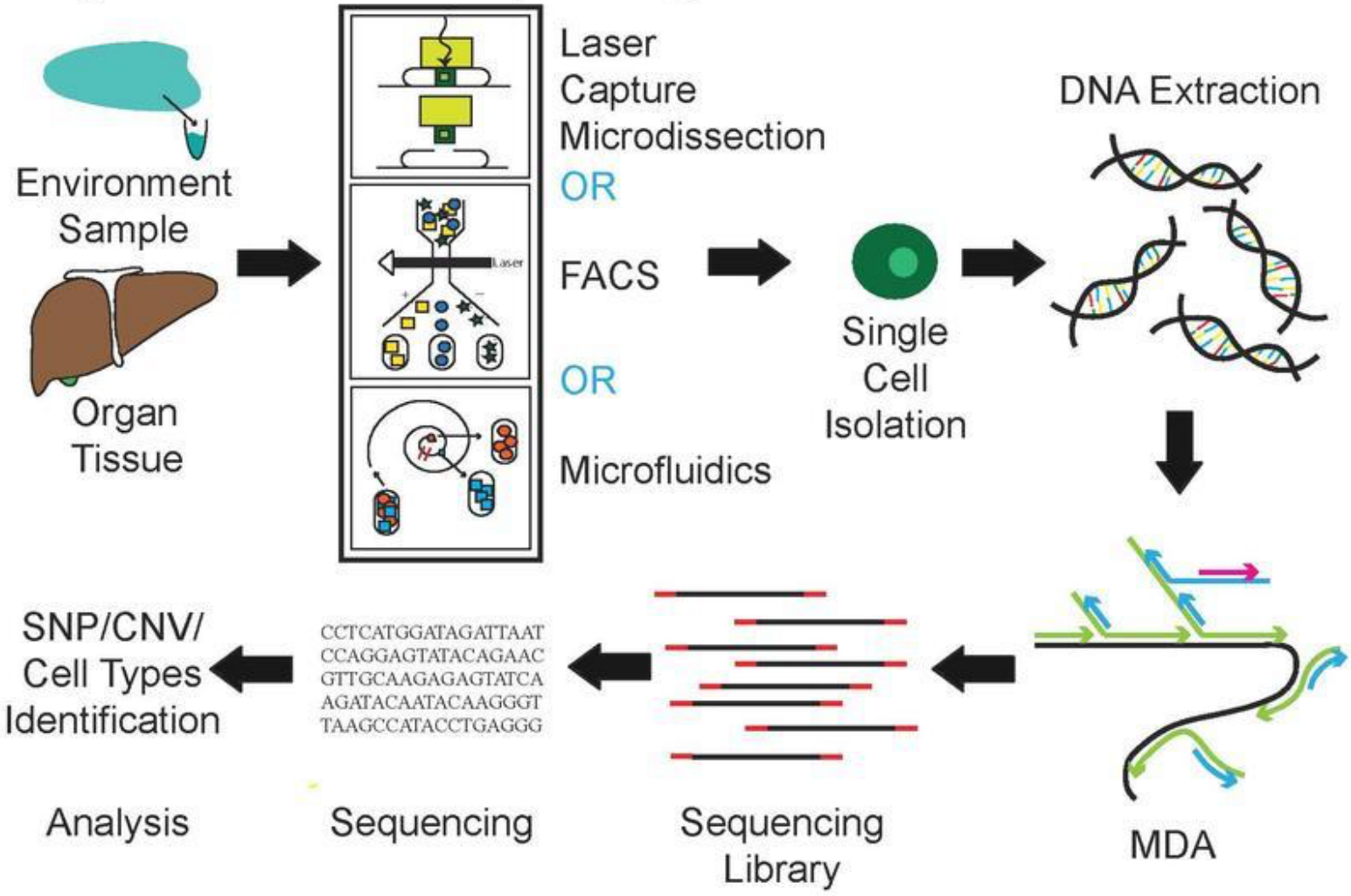




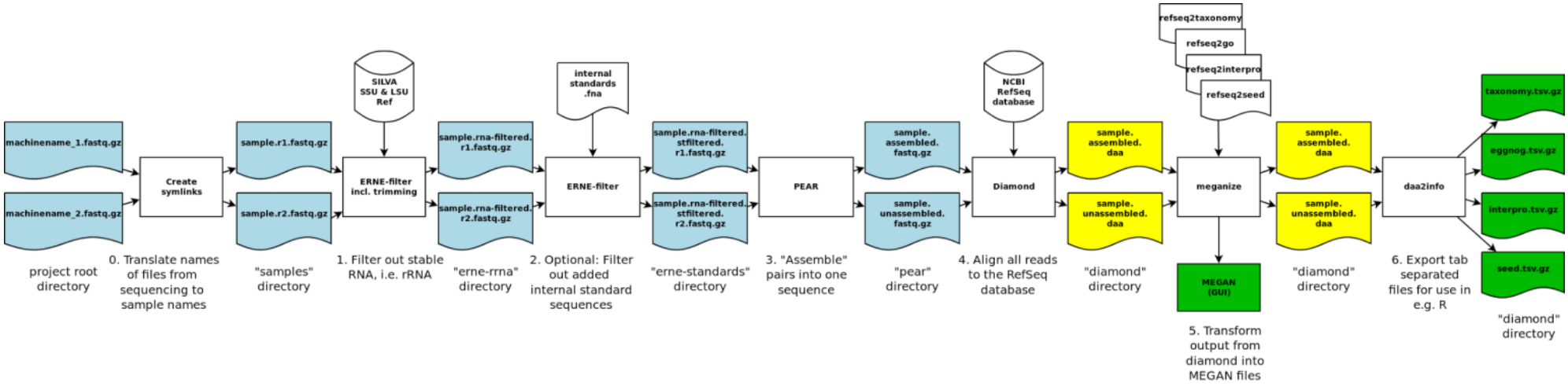
Kang et al., PeerJ, 2015


Sequencing
Sample 0 reads
Sample 1 reads
Sample 2 reads
Quality trimming and filtering
Contaminant DB (rRNA, human, etc.)
Contamination filtering
Unmapped reads
Sample 0 clean reads
Sample 1 clean reads
Sample 2 clean reads
Map
Quantified contigs
Binning
Unbinned contigs
Contig bin 0
Contig bin 1
Phylogeny
Assembly
Contigs set 0
Contigs set 1
Taxonomical
Functional
Functional annotation 1
Annotation DB 2 (NCBI etc.)
Phylogenies
Counts
+sample
Genome bins
ORFs
Taxonomies
Functions
0..1
0..*

Single Cell Genome Sequencing Workflow



Annotation of reads



Quick way to **biology**

The **Funding**

Vetenskapsrådet

Crafoordska stiftelsen
GRUNDAD AV HOLGER CRAFOORD 1980

EUROPEAN SCIENCE FOUNDATION

formas

EcoChange.
Forskning om hav och klimat