Genome sequencing and assembly (a few words…)

Training course

Dr. Marc Höppner

Dr. Montserrat Torres-Oliva

Kiel, 5th March 2018

Genome project – an overview

- General considerations
- Sequencing strategies
- What comes afterwards..?

Genome project – general considerations

Reasons for sequencing a genome

- Comprehensive reference for RNA-seq
- High-resolution Fst
- Search for specific variants
- Comparative genomics / gene evolution

Some (cheap) alternatives

- Population analysis: GBS, RAD-seq
- Gene expression: de-novo transcriptome assembly and annotation

Genome project – general considerations

Characteristics of the genome / species

Genome size

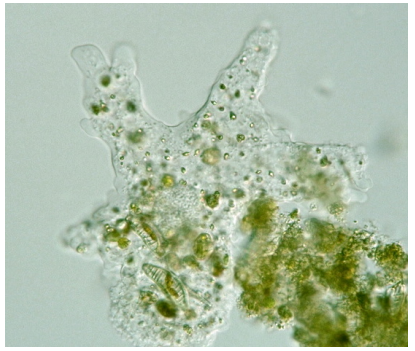
- How much sequencing is needed?
- Which assembly tools are suitable (or not suitable)?
- How does this affect my budget and sequencing strategy?

Gauging genome size:

Kmer analysis from sequencing data

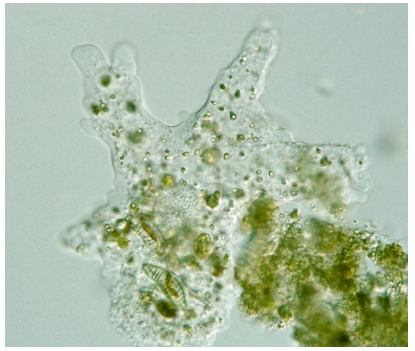
Published data from related species

http://genomesize.com/



(*Polychaos dubium,* ~640GB)

Genome project – general considerations



(*Polychaos dubium,* ~640GB)

Chthalamus sp.

Characteristics of the genome / species

Repeat content

Association with symbionts, microbiome, etc

- How difficult will it be to get a “sterile” DNA sample

How much DNA can we get from one individual?

- Can the organism be grown clonally? What is the level of heterozygosity?

Are there usable protocols for the extraction of high molecular weight DNA?

Genome project – general considerations

Timeline

How quickly do I need the genomic data for my project?

How much method establishing is required in the lab?

Do I have the technical prerequisites for running the assembly myself?

Do I have the bfx expertise?

Genome project – general considerations

Financial considerations

How much am I willing/able to spend?

This affects:

Timeline (prep work)

Sequencing strategy (but: more expensive does not always mean better…)

Buffer for failed sequencing runs, inadequate coverage, additional libraries etc

Genome project – sequencing strategies

Available sequencing technologies

Illumina – sequencing by synthesis

- Very high output (> 1 Tb per 2 days on Novaseq 6000)
- Short reads (up to 2x150bp on Novaseq)
- Low error rate (< 0.1%)
- Cheap (“1000 dollar genome”)
- No stringent requirement on DNA length (unless used together with other technologies/methods)

Genome project – sequencing strategies

Available sequencing technologies

PacBio – SMRT long read sequencing

- low output (~5Gb per SMRT cell)
- long reads (up to 35kb)
- high error rate (~15%)
- expensive (~1500-2000€ per SMRT cell)
- real-time sequencing
- Requires high molecular weight DNA (> 50kb)

Genome project – sequencing strategies

Available sequencing technologies

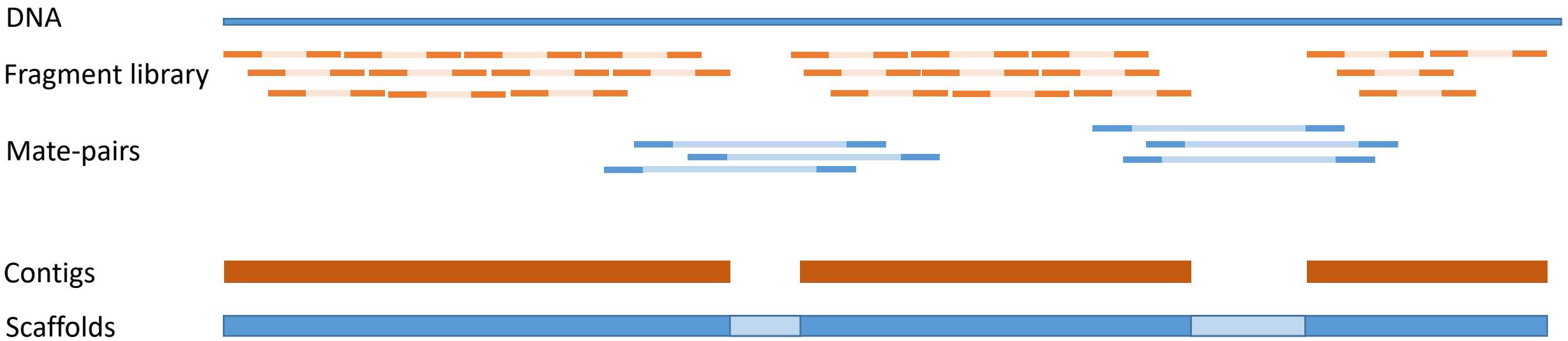
Nanopore – “nanopore” long read sequencing

- low output (~5Gb per flow cell)
- long reads (up to 1MB)
- high error rate (~15%)
- expensive (~1500€ per flow cell)
- real-time sequencing
- not quite „production“ ready yet
- requires high molecular weight DNA (> 50kb)

Genome project – sequencing strategies

Illumina-only: fragments and mate-pairs

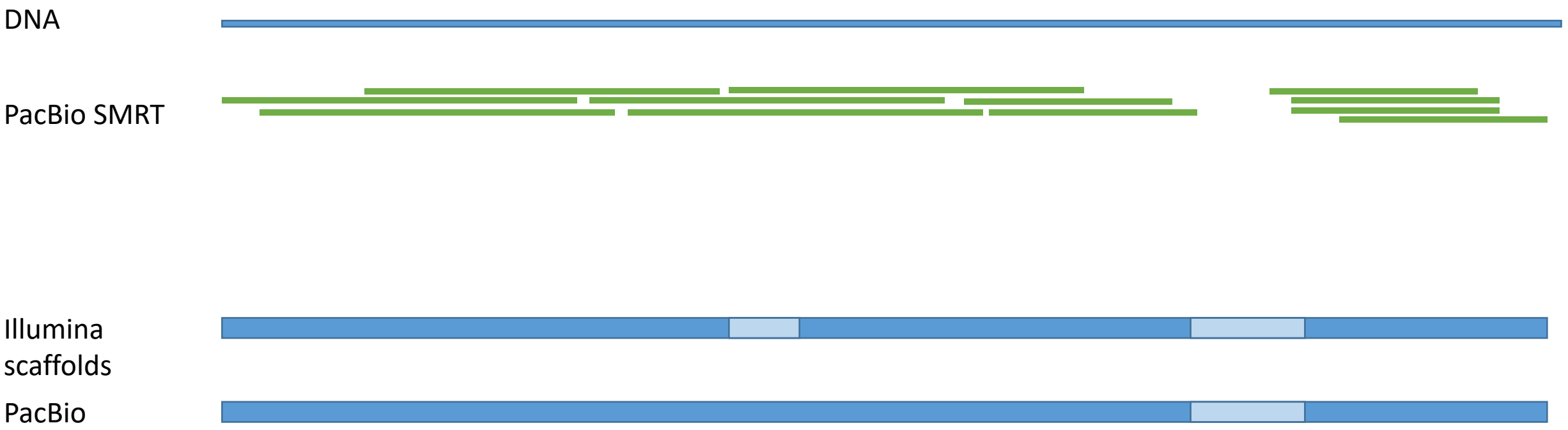
$



Genome project – sequencing strategies

Illumina and long reads

$$



Genome project – sequencing strategies

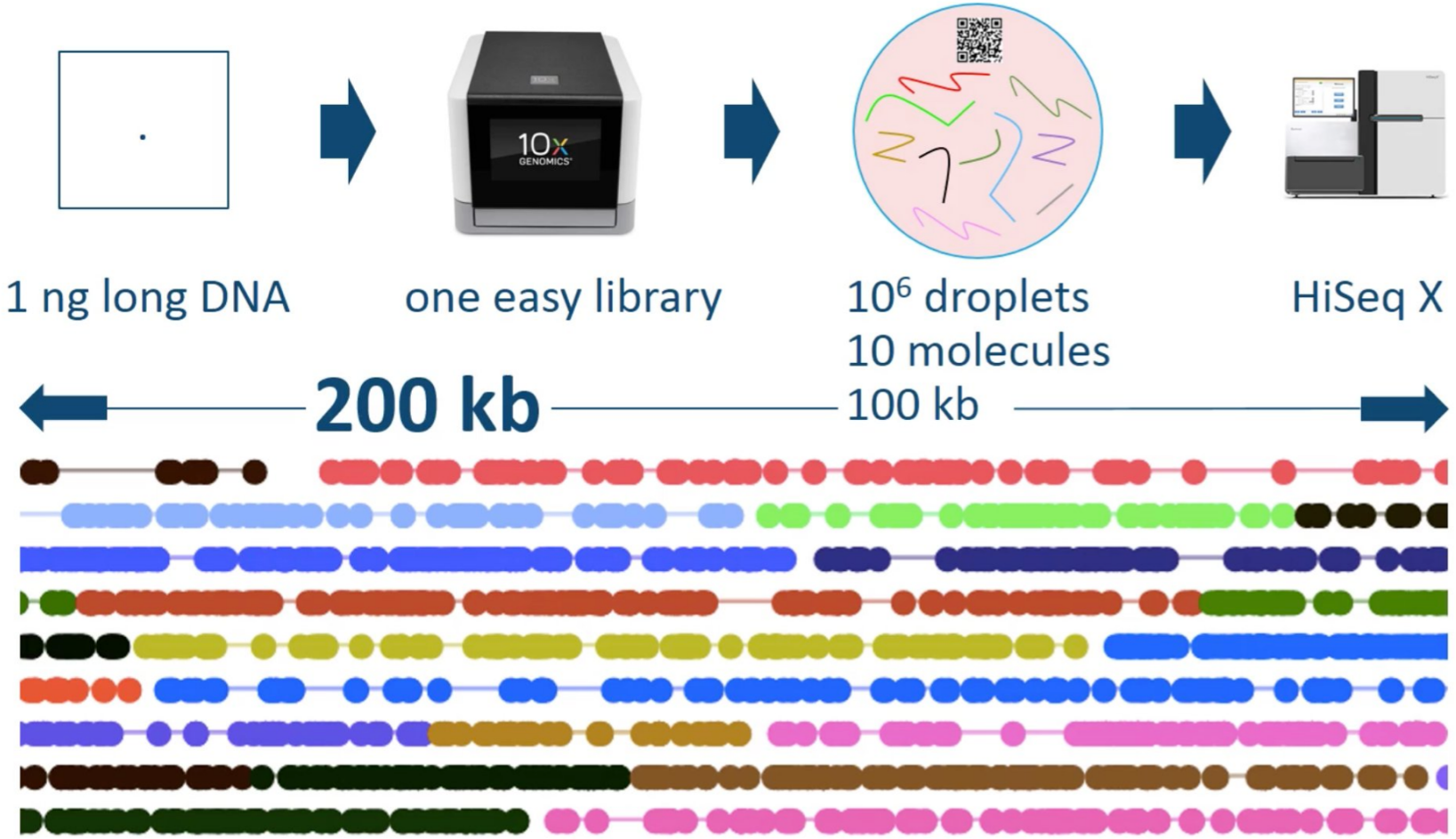
Long reads only

$$$

Genome project – sequencing strategies

Linked reads

$



Genome project – sequencing strategies

Which method is for me / my organism?

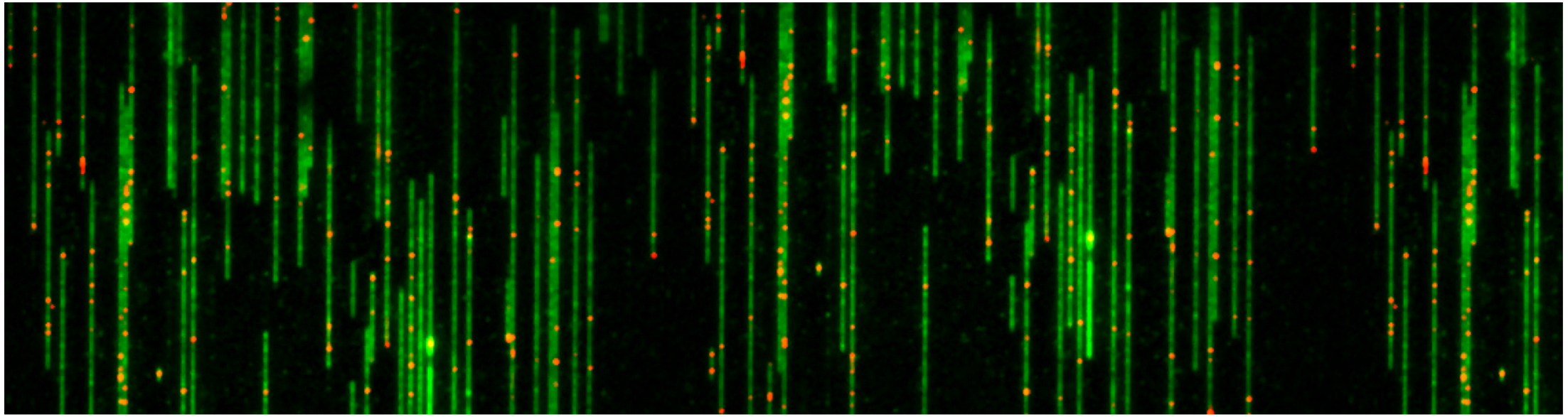
	Illumina PE/MP	Illumina + long reads	Long reads	Linked reads
Cost 1GB	~4000€	~5500€	~6000€	~1500€
Cost 8GB	~25.000€	37.000€	~36.000€	~16.000€
Pros	Established, many tools, high base quality	Can produce near-complete assembly, established	Only a single sequencing approach	Very cheap, full phasing, single library, low amount of DNA needed
Cons	Expensive at scale, gaps, high DNA amount needed	Expensive at scale, requires HMW DNA, high DNA amount	Expensive, few tools, low base quality, HMW DNA	Very new, requires very HMV DNA

Genome project – sequencing strategies

Optical mapping

None of the methods are capable of assembling full chromosomes (most of the time)

Solution: Scaffolding assembly against an optical map (e.g. BioNano Saphyr)



~3000€ for a vertebrate genome (in addition to the actual sequencing)

Genome project – and then?

Can I get help with this?