

P4EU meeting

Protein Production and Purification Partnership in Europe

Sf21 genome: assembly and annotation

15th June 2016 - Heidelberg
Jonathan Landry





Sf21 project

Motivation



- *Spodoptera frugiperda*
- Extensively used in production of eukaryotic recombinant proteins by their expression *via* baculoviruses
- Genomes unknown, preventing ‘tinkering’
- Sequencing of genome and transcriptome of Sf21 insect cell lines (started in 2013)
- Sf21 genome assembly and annotation
- Important resource
 - Development of new protein expression methods in insect cells
 - Better identification of contaminants the host cells



Sf21 project

Collaborative effort



- Genome and transcriptome sequencing (short and long read technologies)
- Genome assembly

Genomics Core Facility



- Genome annotation



- Sf21 resource

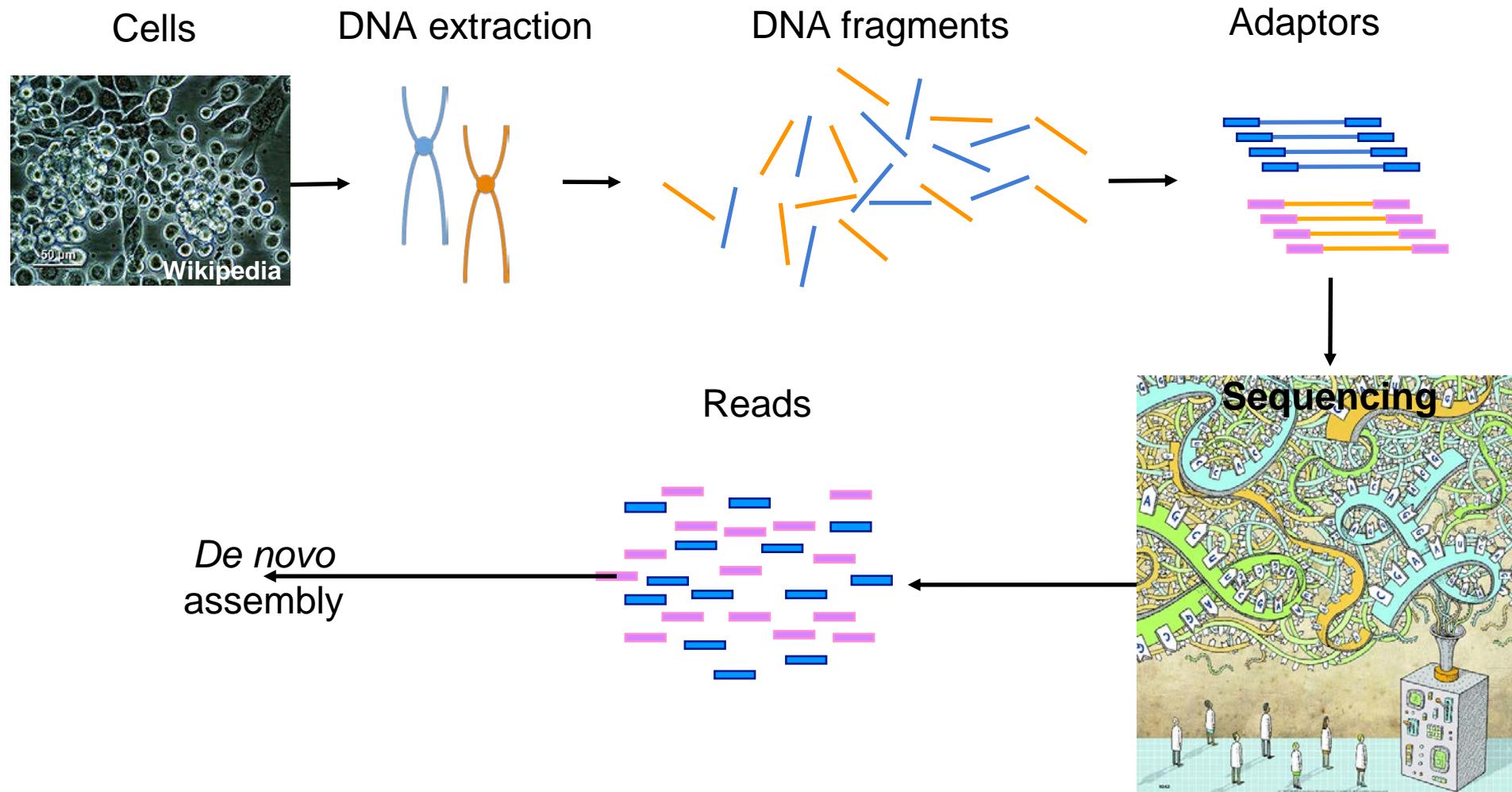
- Proteomic database

Proteomics Core Facility



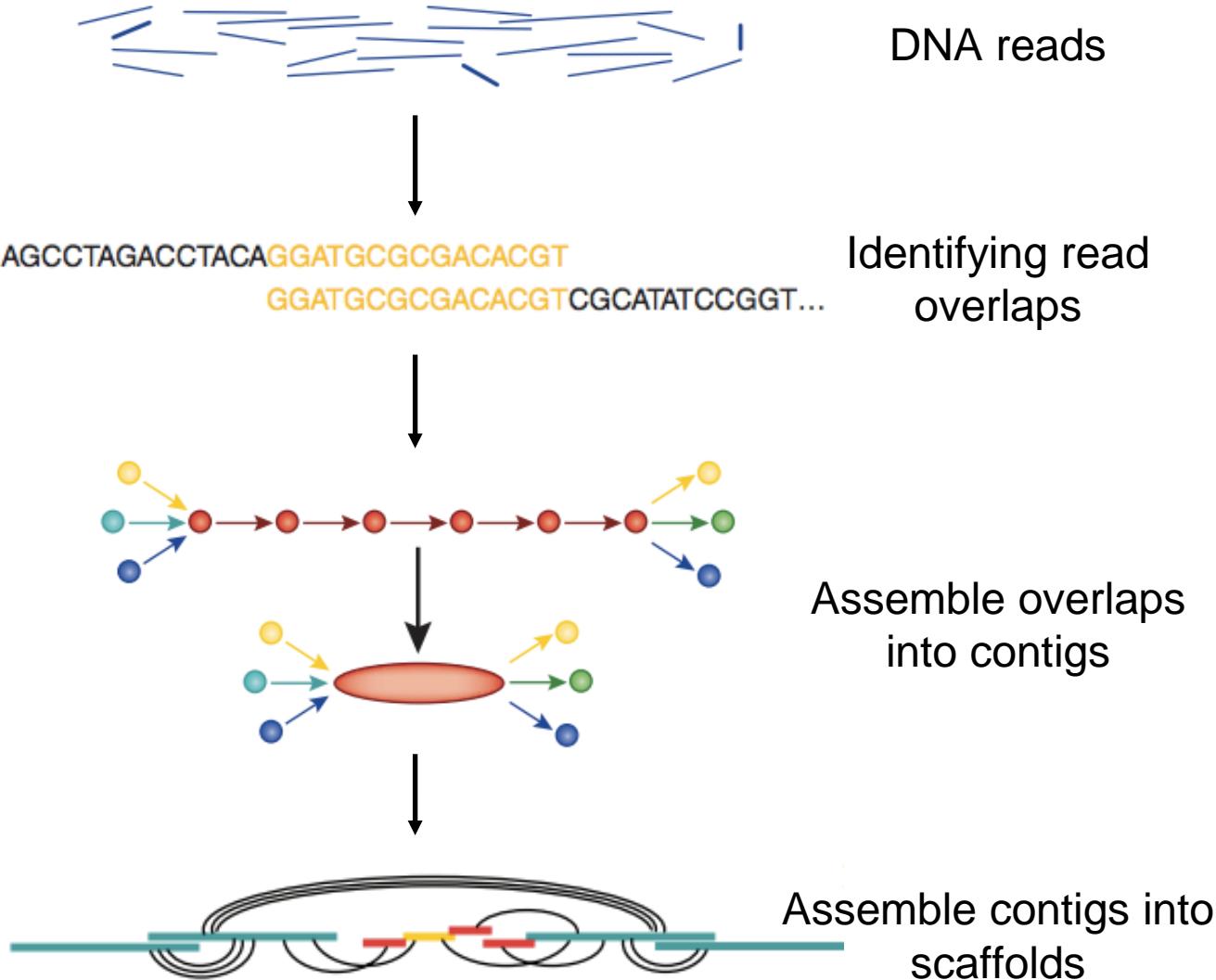
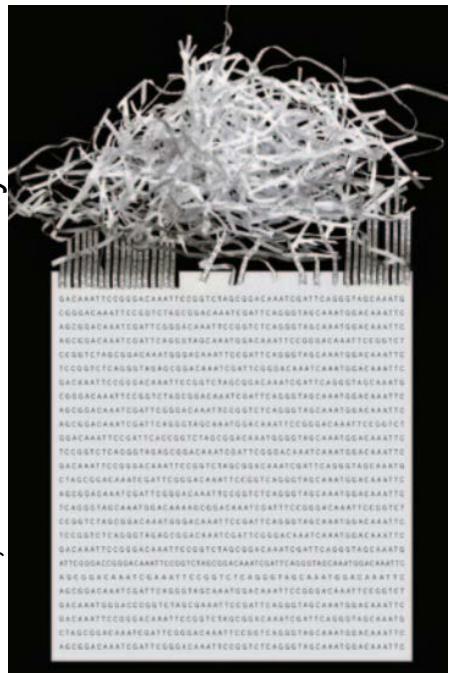
- Identification of specific promoters (Köhler *et al.*, submitted)

Sequencing workflow

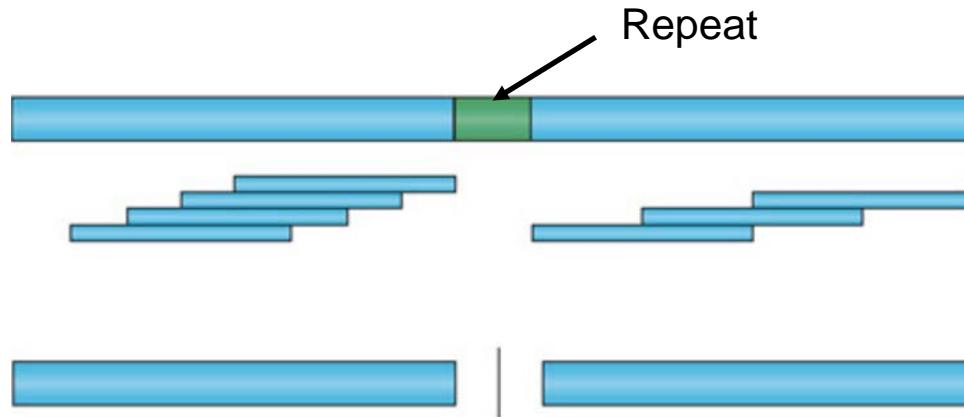


De novo genome assembly

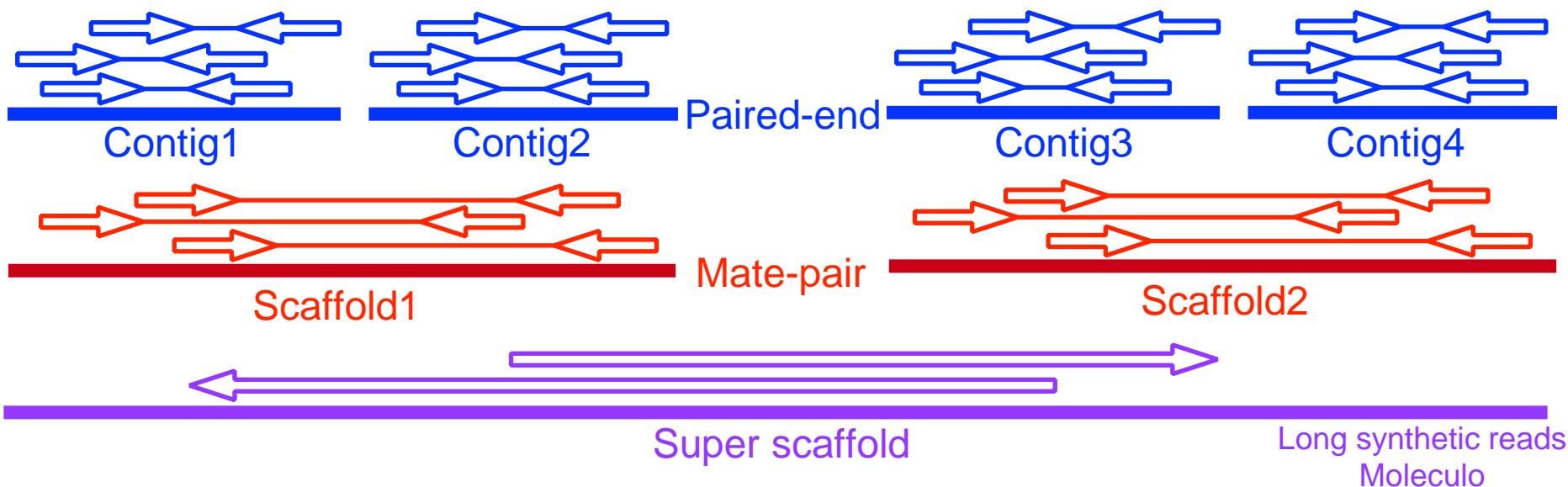
Baker, 2012, Nat. Methods



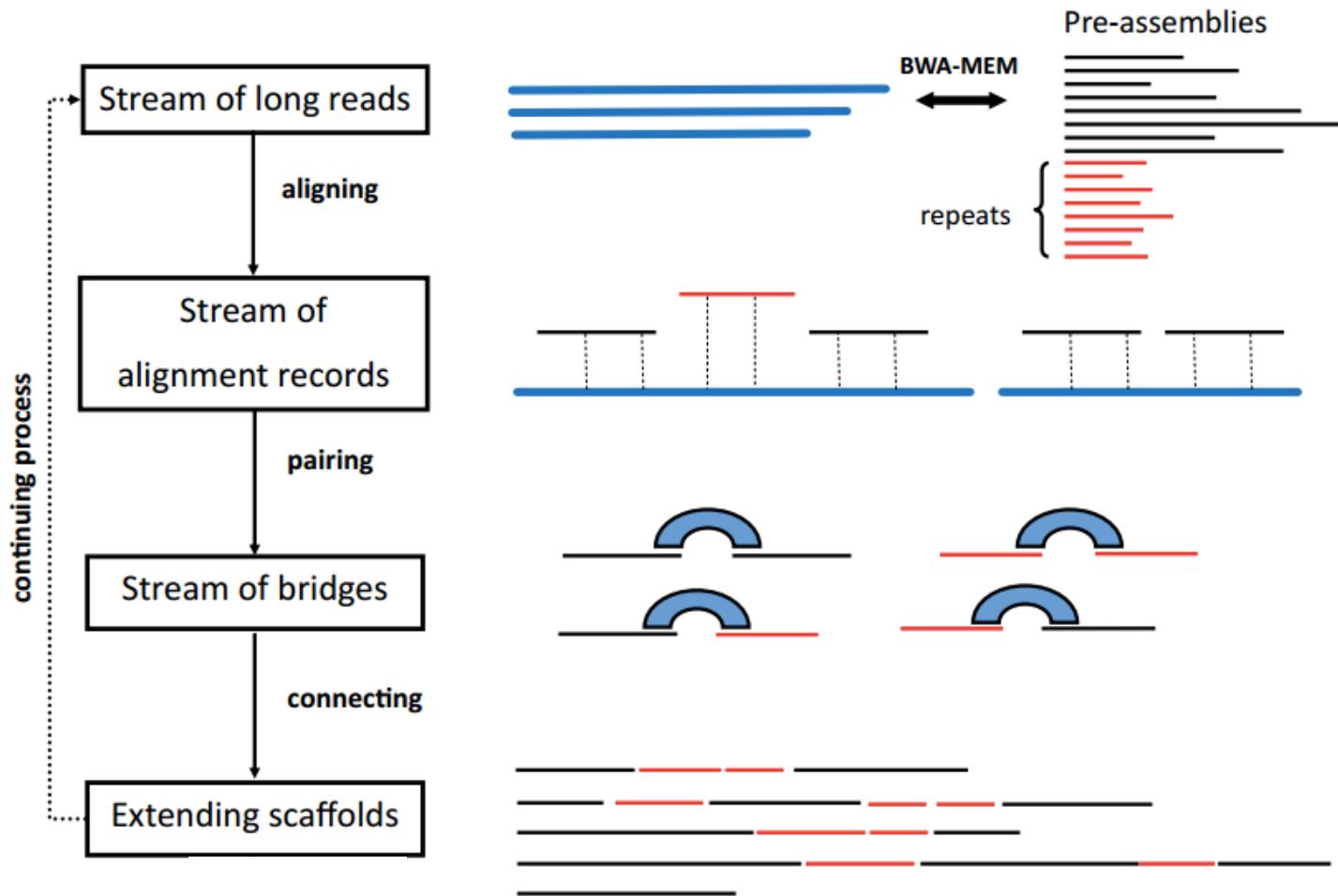
De novo assembly limitations with short reads



Multiple library type approach



Integration of Nanopore reads

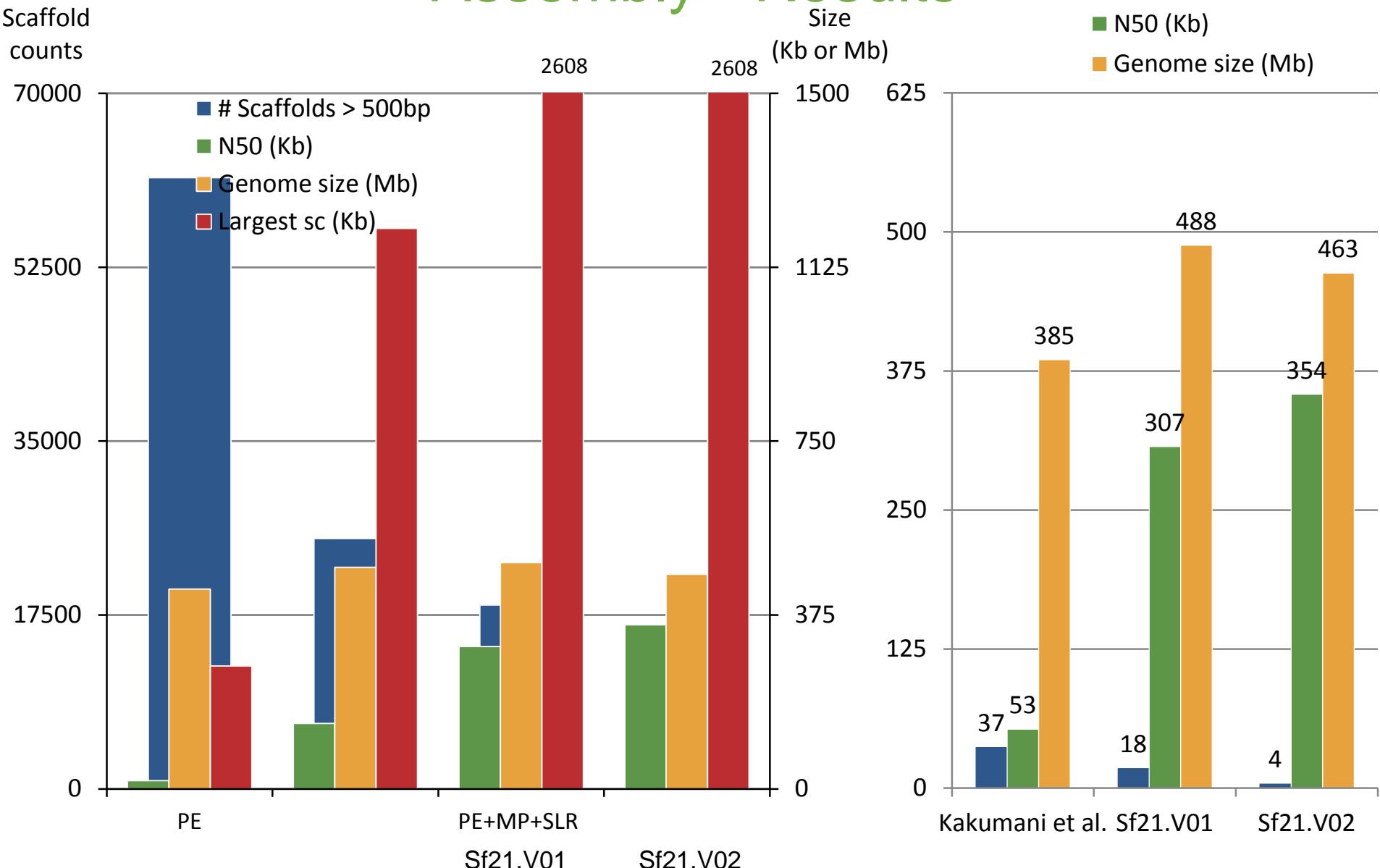


Duc Cao *et al.*, 2016

Sequencing datasets

<i>Library</i>	<i>Input</i>	<i>Mode</i>	<i>Fragment size (bp)</i>	<i>Number of reads (10⁶)</i>
1	DNA	Paired-End	~290	422.73
2	DNA	Paired-End	~590	17.77
3	DNA	Mate Pair	~4500	24.14
4	DNA	Mate Pair	~4500	63.71
5	DNA	Synthetic long reads (SLR)	~4900 (max 19 Kbp)	0.18
6	DNA	Oxford Nanopore (ONT)	~8000 (max 34 kbp)	0.07
7	RNA	Paired-End	~280	91.74

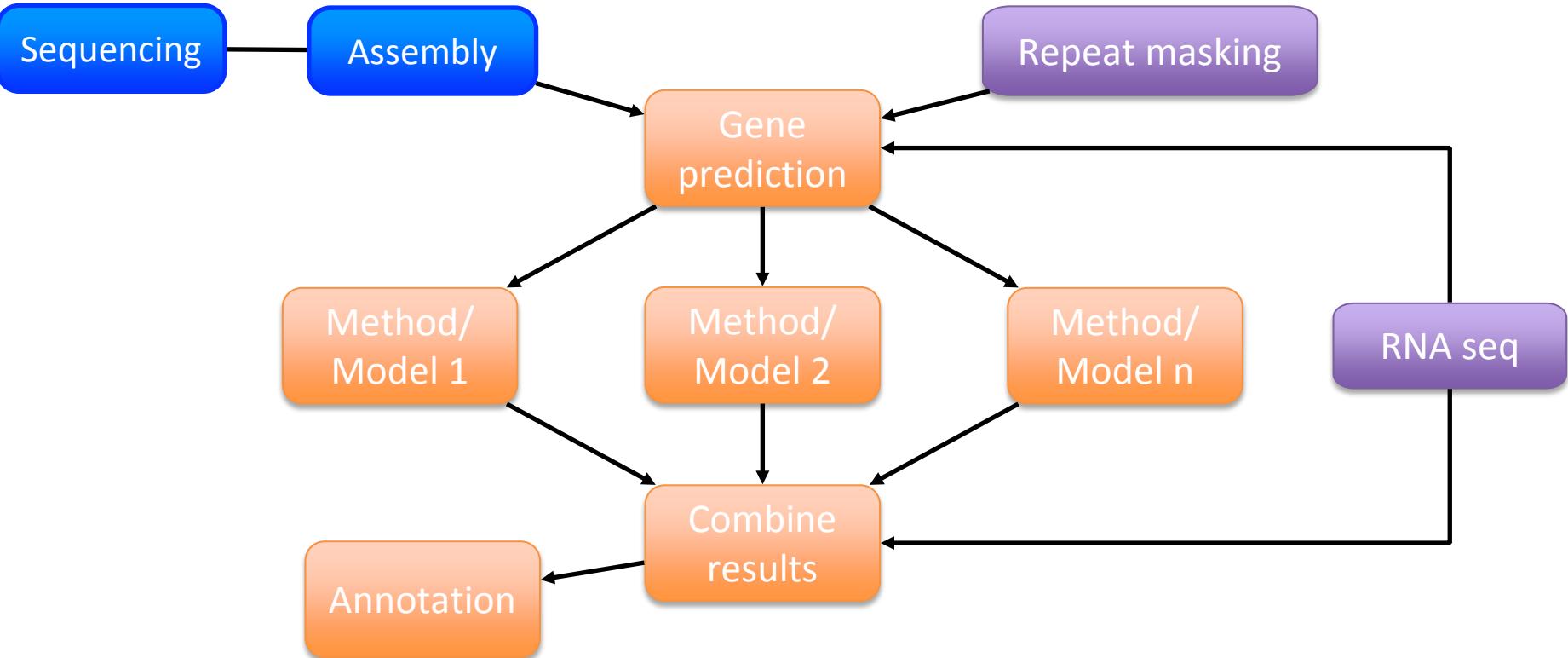
Assembly - Results





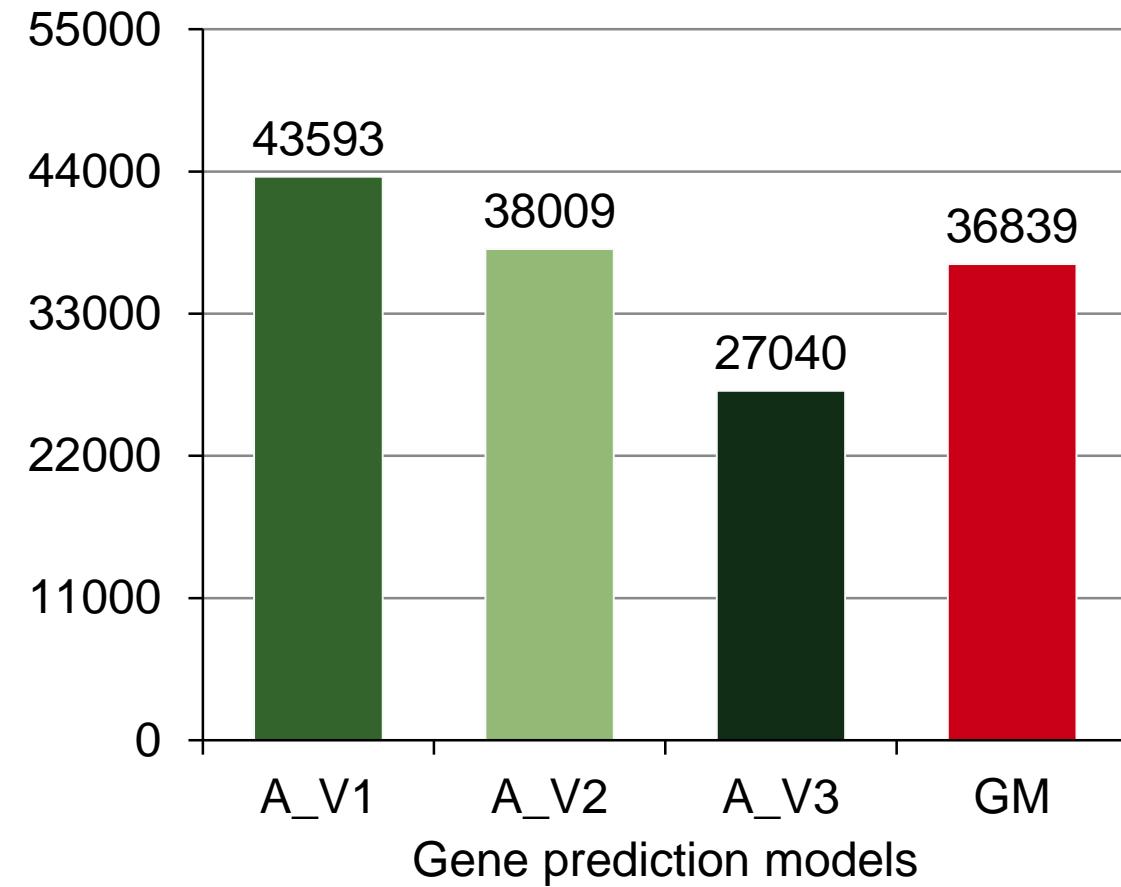
Genome annotation workflow

bio comp



Gene prediction

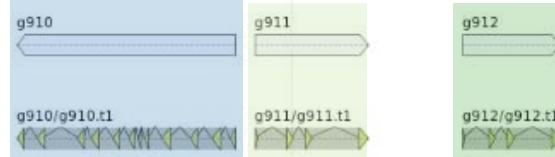
Number of genes predicted



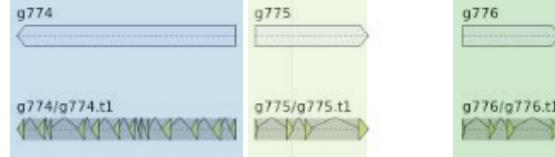
- A_V1-3:
Species specific
exon/intron boundary
models
- GM:
General heuristic gene
models based on GC%

Gene prediction approach

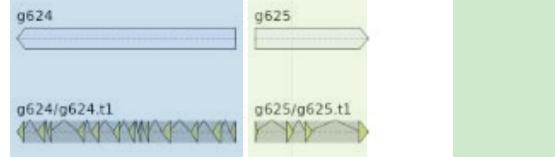
A_V1



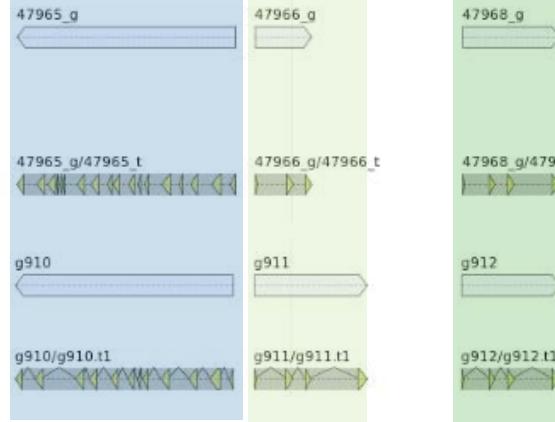
A_V2



A_V3



GM



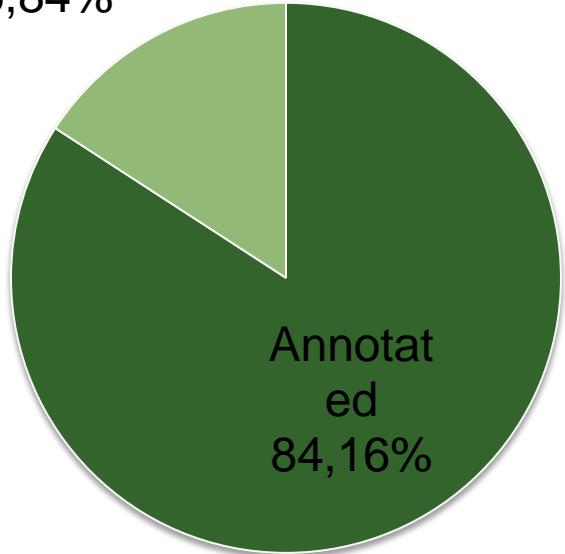
Consensus



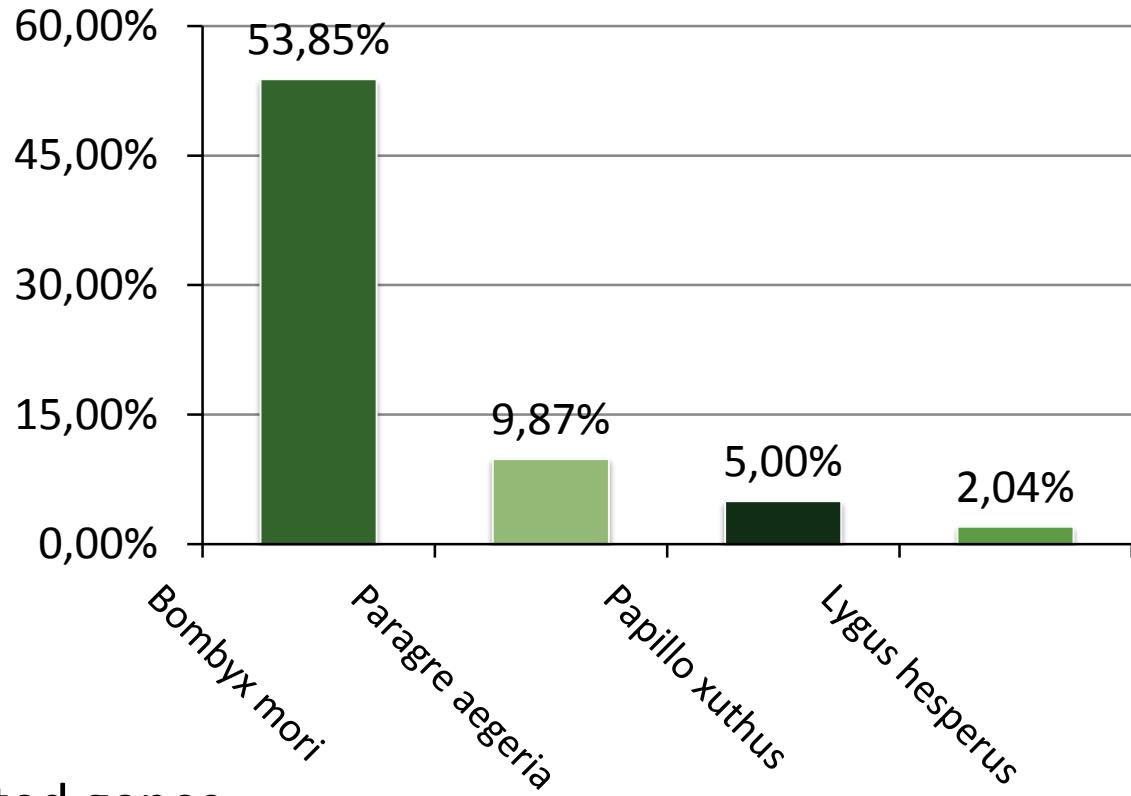
- Overlap at least in 3 predictions
- 23 810 predicted genes

Gene prediction and annotation

Non-
annotat
ed
15,84%

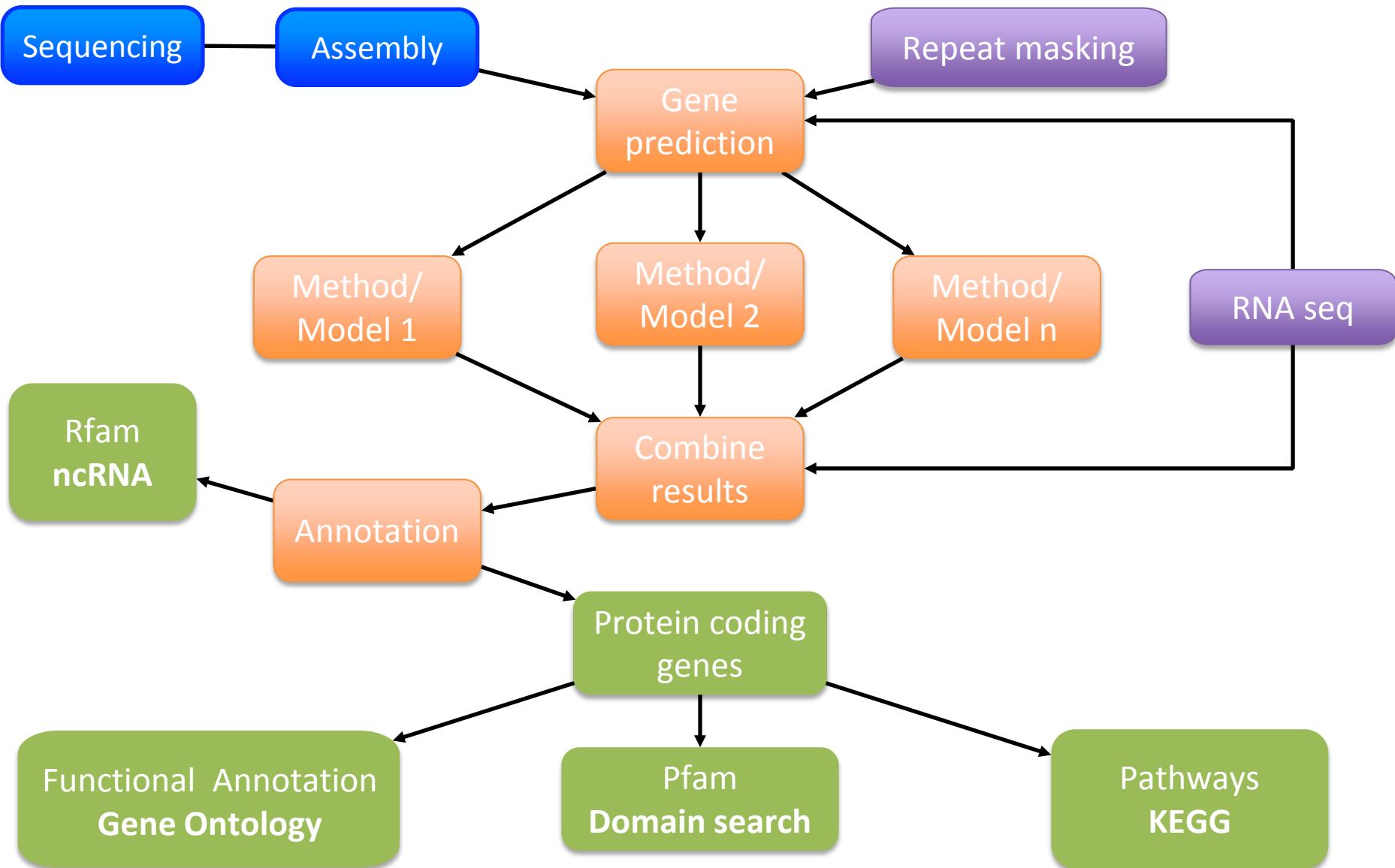


Species representation



- 84.16% annotated genes
 - BLAST against UniProt insect protein data base

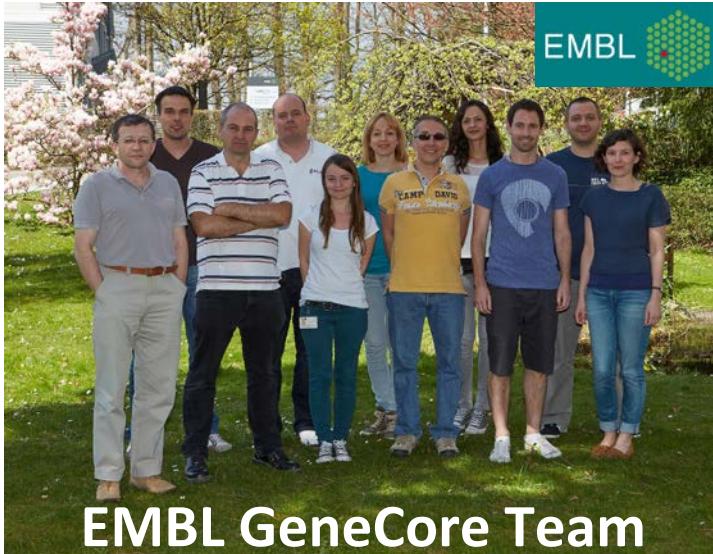
Genome annotation workflow



Summary and future work

- Multiple library types approach with short and long reads
- Comprehensive genome assembly
- Comprehensive genome annotation
- Need to consolidate Nanopore reads integration
- Functional annotation
- Proteomic dataset (Joanna Kirkpatrick, EMBL Proteomics CF)
- Apply sequencing, assembly and annotation strategies to other genomes
 - *Trichoplusia ni* - Hi5 (Insect cell line)
 - *Cerianthus* (Anemones)
 - *Xenoturbella* (Marine worm)

Acknowledgements



EMBL GeneCore Team



Bence Galik

Attila Gyenessei

Peggy Stolt-Bergner



Joanna Kirkpatrick

Proteomics CF

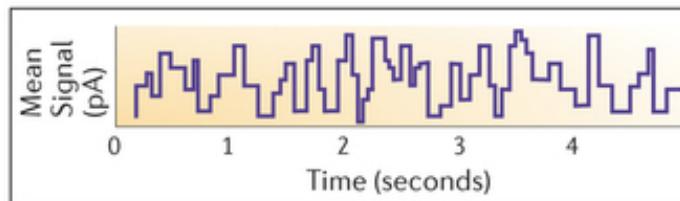
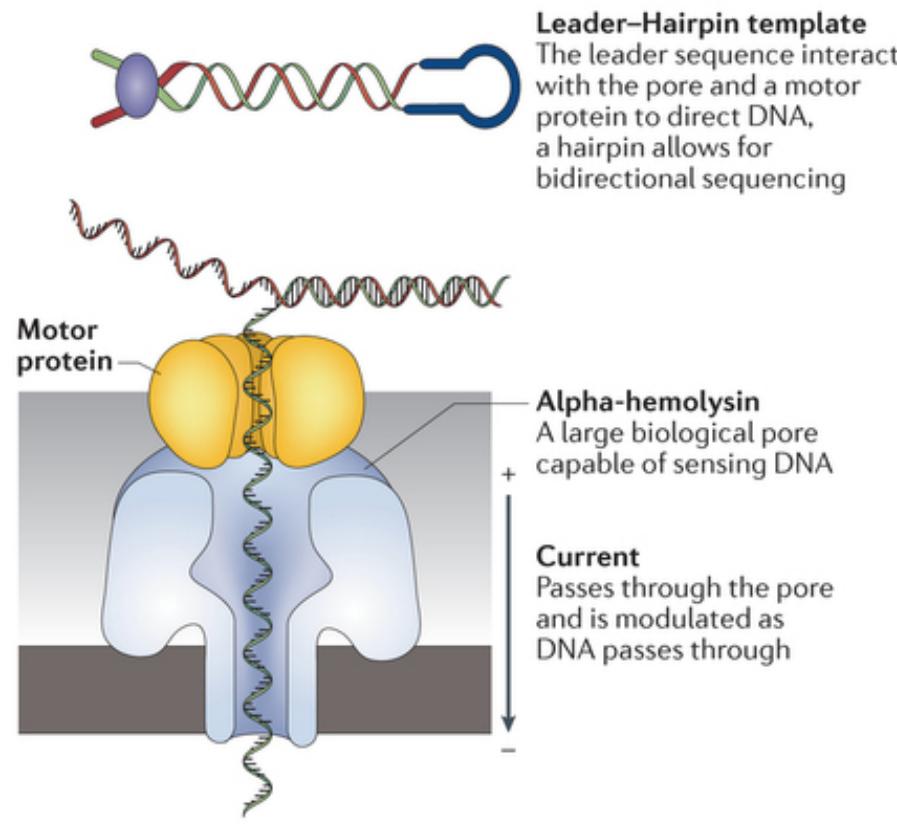


Hüseyin Besir

Kim Remans

Protein Expression and
Purification CF

Oxford Nanopore Technologies

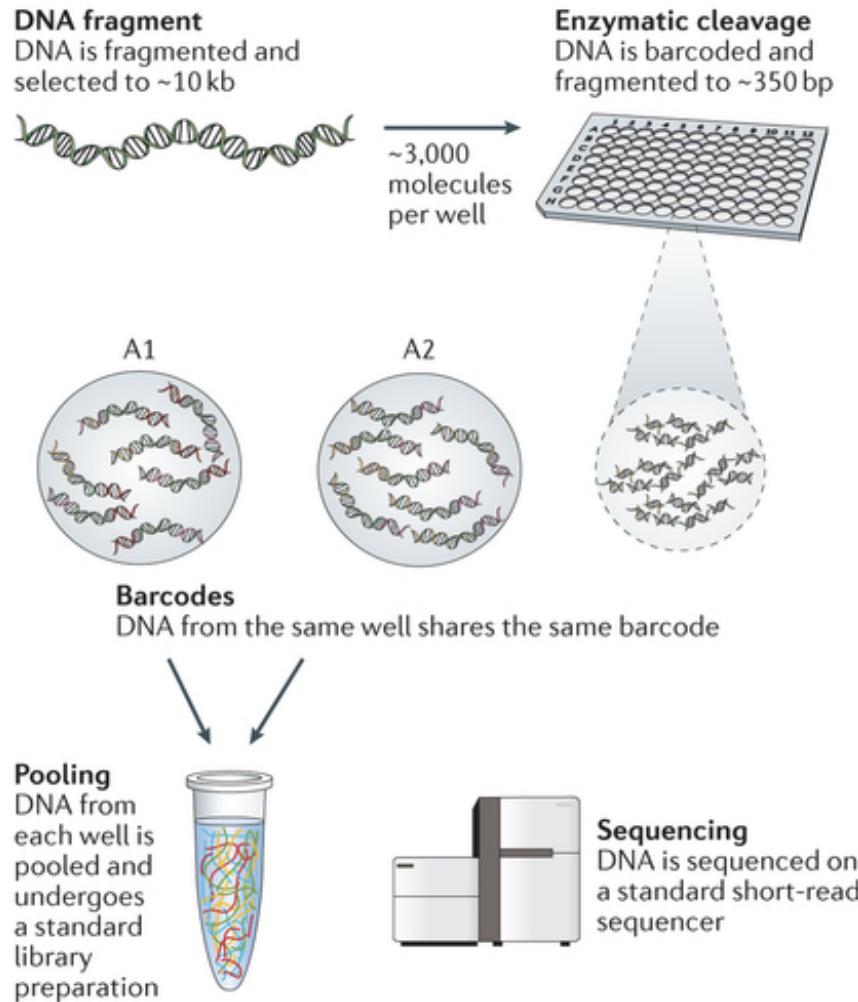


ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

Goodwin *et al.*, 2016

Synthetic long read sequencing - Moleculo

Ba Illumina

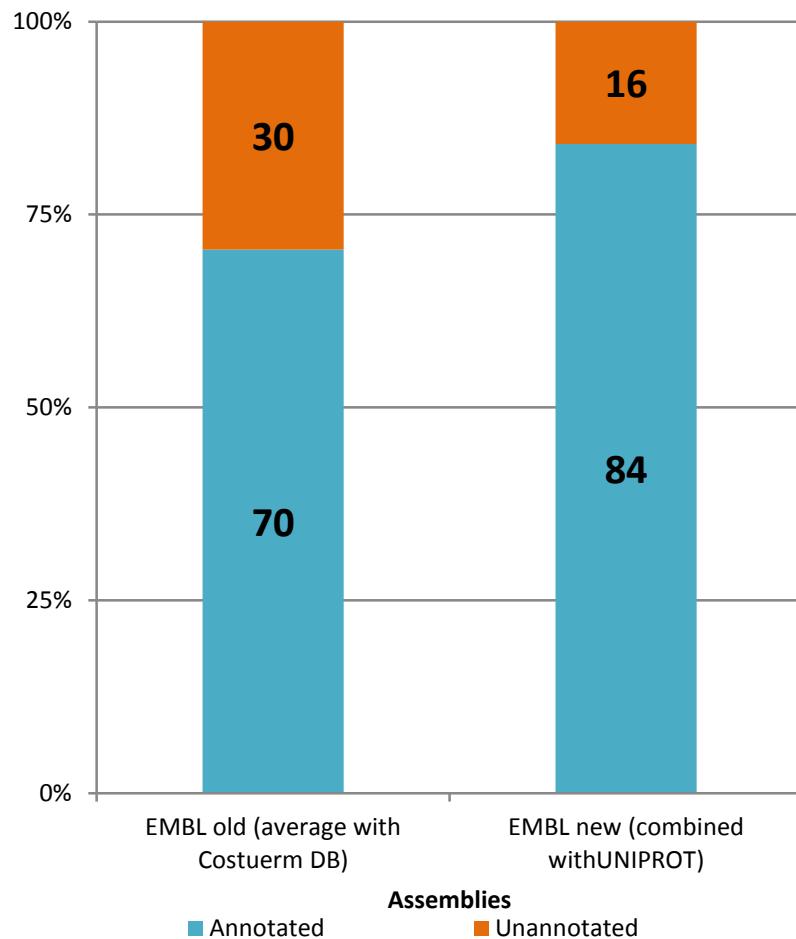


Goodwin *et al.*, 2016



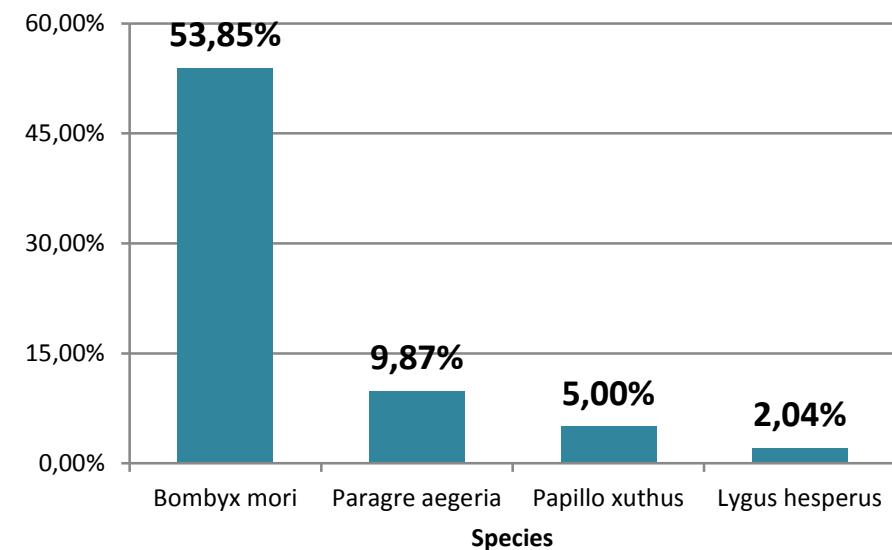
Annotation results

Annotation results



- **23 810** predicted genes
- UniProt insect protein data base (~ 1 M seqs)
- BLAST

Species representation



Annotation results

- Annotation table
- Genes/proteins FASTA files
- GFF

Gene ID	Protein ID	Length	Name	Species	Score	E value
gene_29272	protein_29272	233	Putative gag-pol protein	Drosophila ananassae	204	4.00E-58
gene_29259	protein_29259	370	Kynurenine 3-monooxygenase	Anopheles gambiae	259	6.00E-80
gene_18735	protein_18735	240	Endonuclease and reverse transcriptase-like protein	Bombyx mori	87.8	4.00E-18
gene_29276	protein_29276	193	Similar to CG11966	Papilio xuthus	358	1.00E-125
gene_29277	protein_29277	219	Uncharacterized protein	Bombyx mori	245	7.00E-76
gene_29434	protein_29434	111	Scarface	Papilio xuthus	101	2.00E-24
gene_29308	protein_29308	124	hypothetical protein	No hits found	---	---
gene_29328	protein_29328	264	Putative uncharacterized protein	Harpegnathos saltator	191	6.00E-58
gene_29300	protein_29300	223	Non-LTR retrotransposon R1Bmks ORF1 protein	Bombyx mori	140	2.00E-37
gene_18747	protein_18747	186	Endonuclease-reverse transcriptase	Bombyx mori	152	6.00E-41



Detailed Status Information

Manuscript #	NMETH-BC27767
Current Revision #	0
Submission Date	24th Apr 16
Current Stage	Manuscript under consideration
Title	Genetic code expansion for multiprotein complex engineering
Manuscript Type	Brief Communication
Manuscript Comment	exclude: Peter Schultz (Scripps), Jason Chin (LMB) suggested reviewers: Kai Johnsson, Andrea Musacchio, Ervin Fodor, Monique van Oers, Daniel Fitzgerald, Arnaud Poterszman
Corresponding Author	Edward Lemke (lemke@embl.de) (European Molecular Biology Laboratory)
Contributing Authors	Christine Koehler , Paul Sauter , Mirella Wawryszyn , Gemma Estrada Girona , Kapil Gupta , Jonathan Landry , Markus Fritz , Ksenija Radic , Jan-Erik Hoffmann , Attila Gyenesi , Bence Galik , Sini Junnila , Peggy Stolt-Bergner , Giancarlo Pruneri , Stefan Bräse , Carsten Schultz , Moritz Bosse Biskup , Huseyin Besir , Vladimir Benes , Martin Jechlinger , Jan Korbel , Imre Berger
Authorship	Yes
Abstract	We present a method that enables the site-specific introduction of unique chemical functionalities anywhere in a recombinantly produced eukaryotic protein complex, opening up a plethora of novel avenues in advanced protein complex engineering. We demonstrate the utility and versatility of this efficient and robust protein production platform i) to fluorescently label target proteins using click chemistry, ii) for glycoengineering of antibodies, and iii) for structure-function studies of novel eukaryotic complexes using site-specific crosslinking strategies.
Subject Terms	Biological sciences/Biochemistry/Proteins Biological sciences/Chemical biology/Protein design
Show Author Information	Allow Reviewers to see Author information.
Competing Financial Interest	Yes there is potential Competing Interest. The authors declare a competing financial interest: a patent application comprising parts of the MultiBacTAG technology here described has been filed.
Applicable Funding Source	No Applicable Funding

Another one is under preparation
(with the EMBL, Genomics Core Facility)