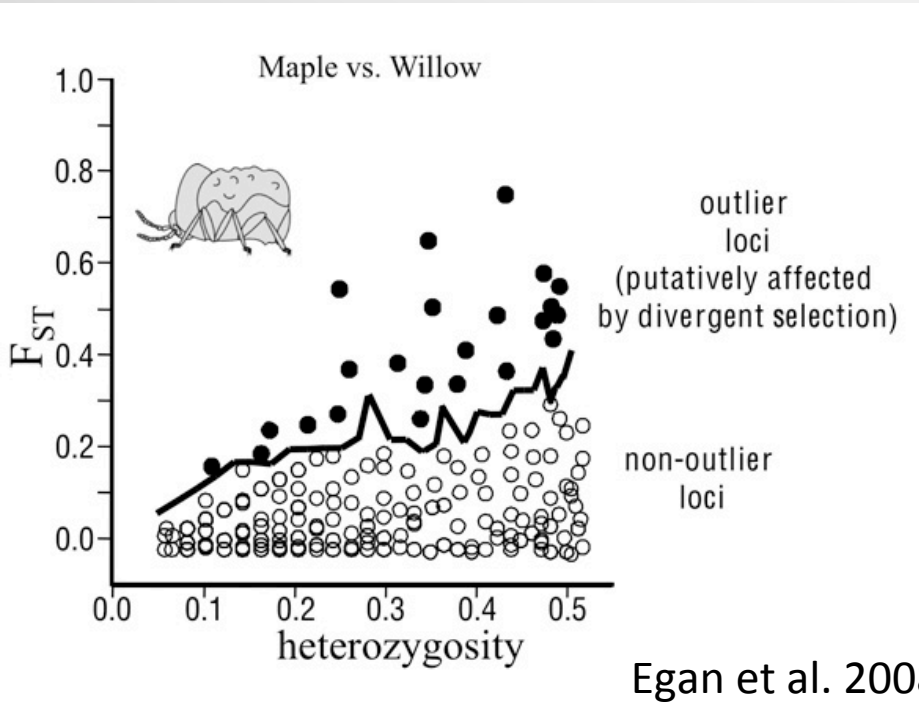
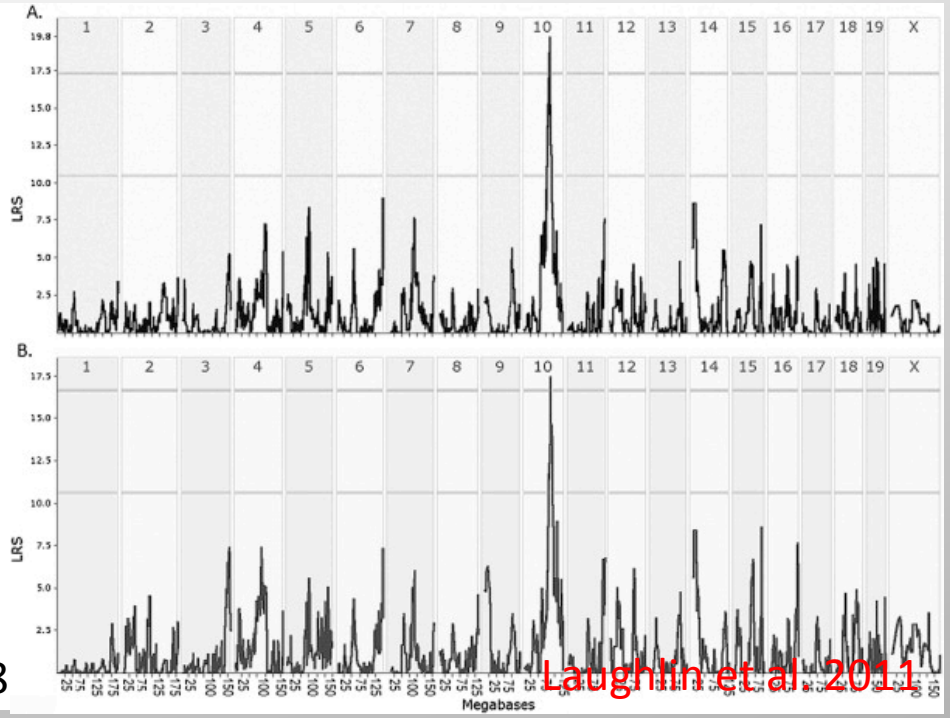
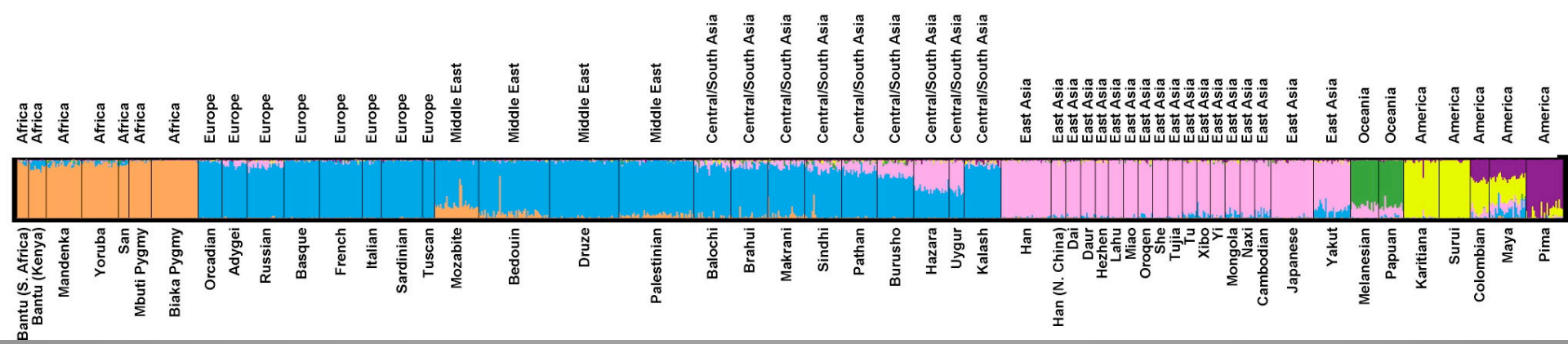
Introduction to population genomics



Egan et al. 2008



Laughlin et al. 2011



Li et al. 2008

What does Wikipedia have to say?

Population genomics is the large-scale comparison of DNA sequences of populations. Population genomics is a neologism that is associated with population genetics.

Population genetics is the study of allele frequency distribution and change under the influence of the four main evolutionary processes: natural selection, genetic drift, mutation and gene flow. It also takes into account the factors of recombination, population subdivision and population structure. It attempts to explain such phenomena as adaptation and speciation.

Genetics vs. Genomics

Old questions with tens of markers:

- historical demography, phylogeny

- New genomic questions: functional genetics in a demographic & spatial context
  - Adaptation
  - Evolutionary constraints/capacity
    - Inbreeding
    - Genetic load
    - Gene flow variation across landscape, across chromosomes
    - Genotype / environment associations

Outline

- The Genomic revolution
- Why not sequence whole genomes? - Reduced Representation Methods
- Marker development
  - Restriction-based approaches
  - Transcriptomes (for SNPs)
  - Other targeted approaches

The Genomic Revolution

Illumina
HiSeq 2500

500 Gb in ca. 40 h (2 * 125 bp)

MiSeq

15 Gb in ca. 50 h (2 * 300 bp)

(Human genome ≈ 3.3 Gb)

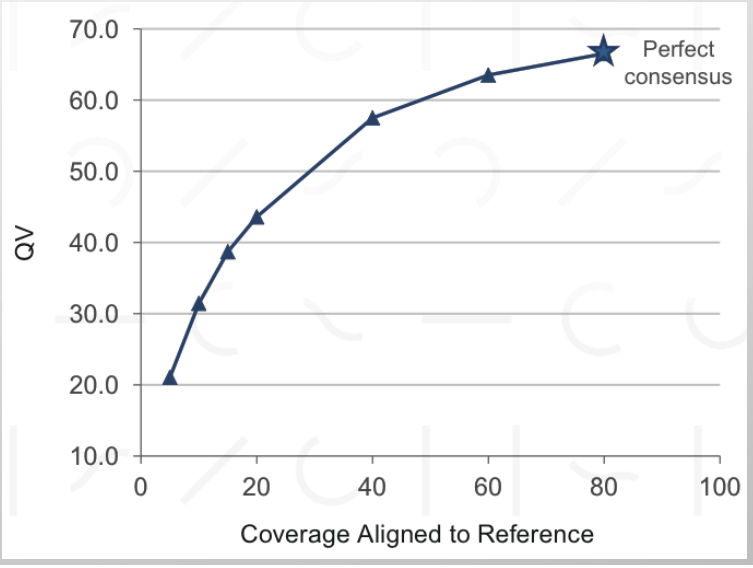
The Genomic Revolution

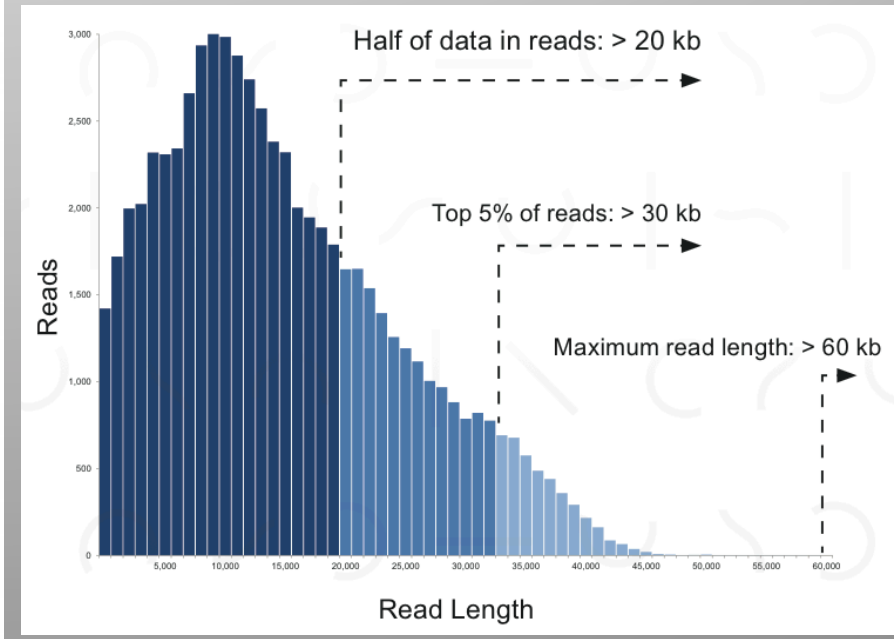
PacBio RS II

150 000 Kreads / SMRT cell

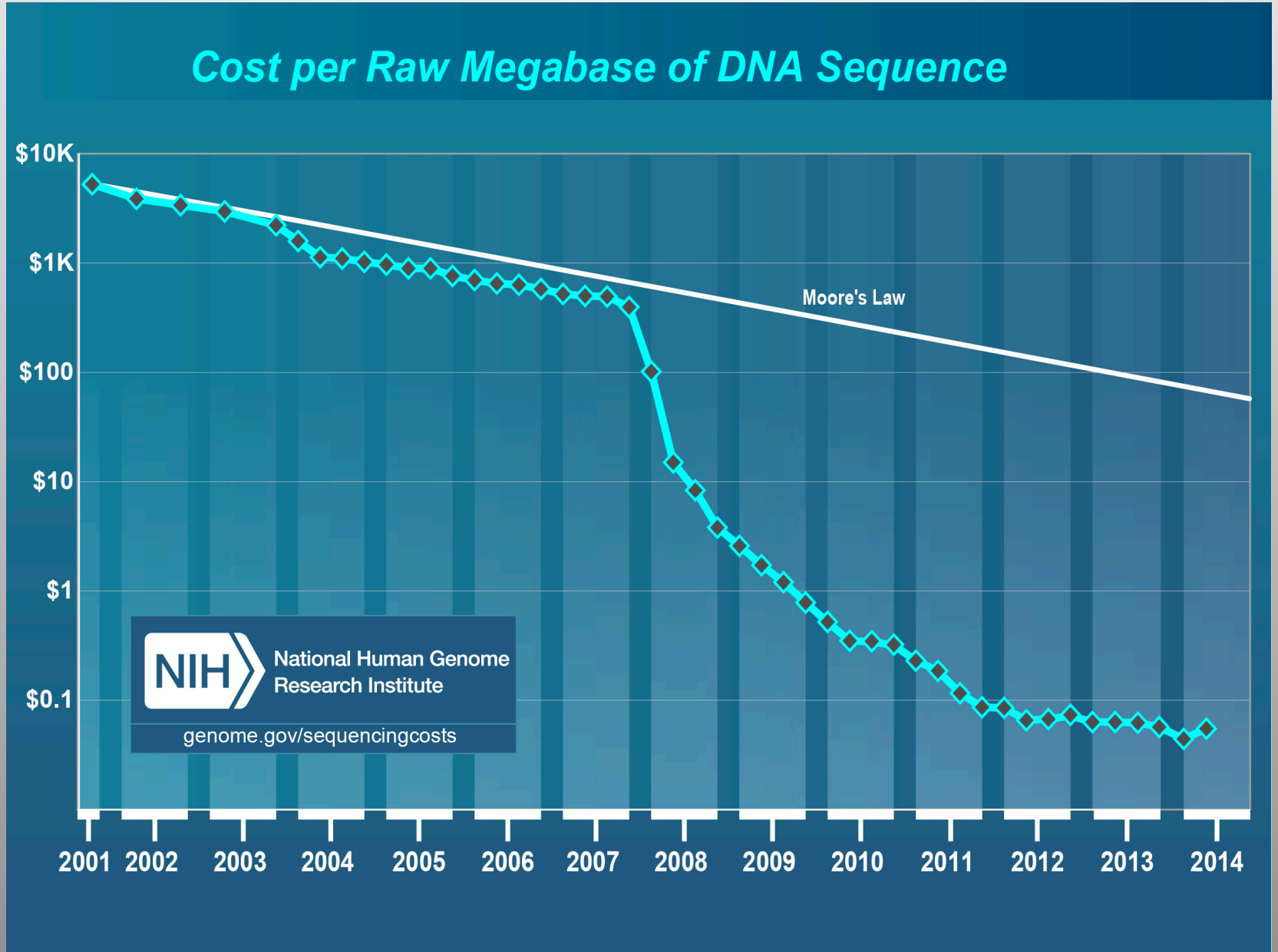
Sequel

1 M reads / SMRT cell

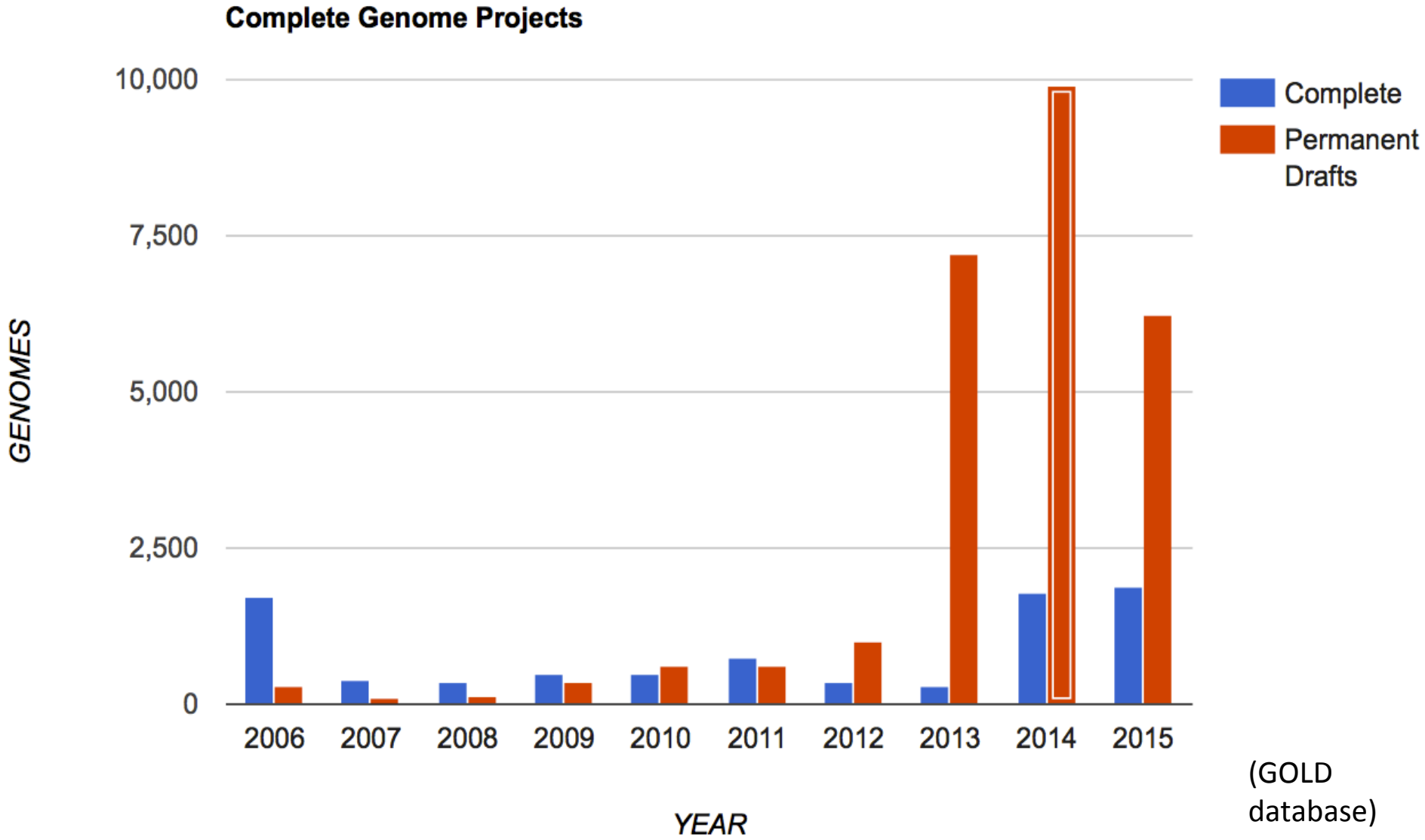




Sequencing is getting cheaper…

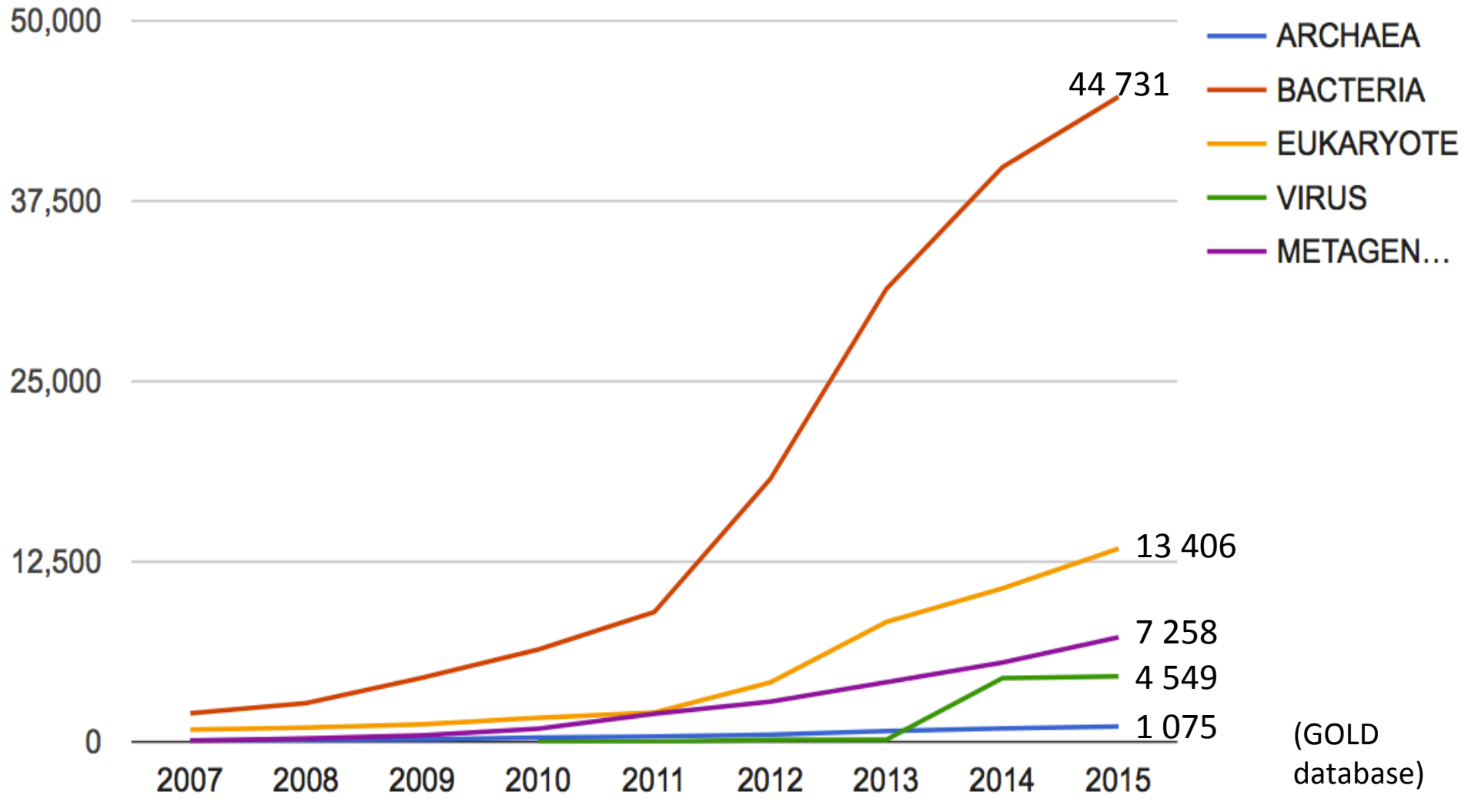


…and genome sequences are exploding!



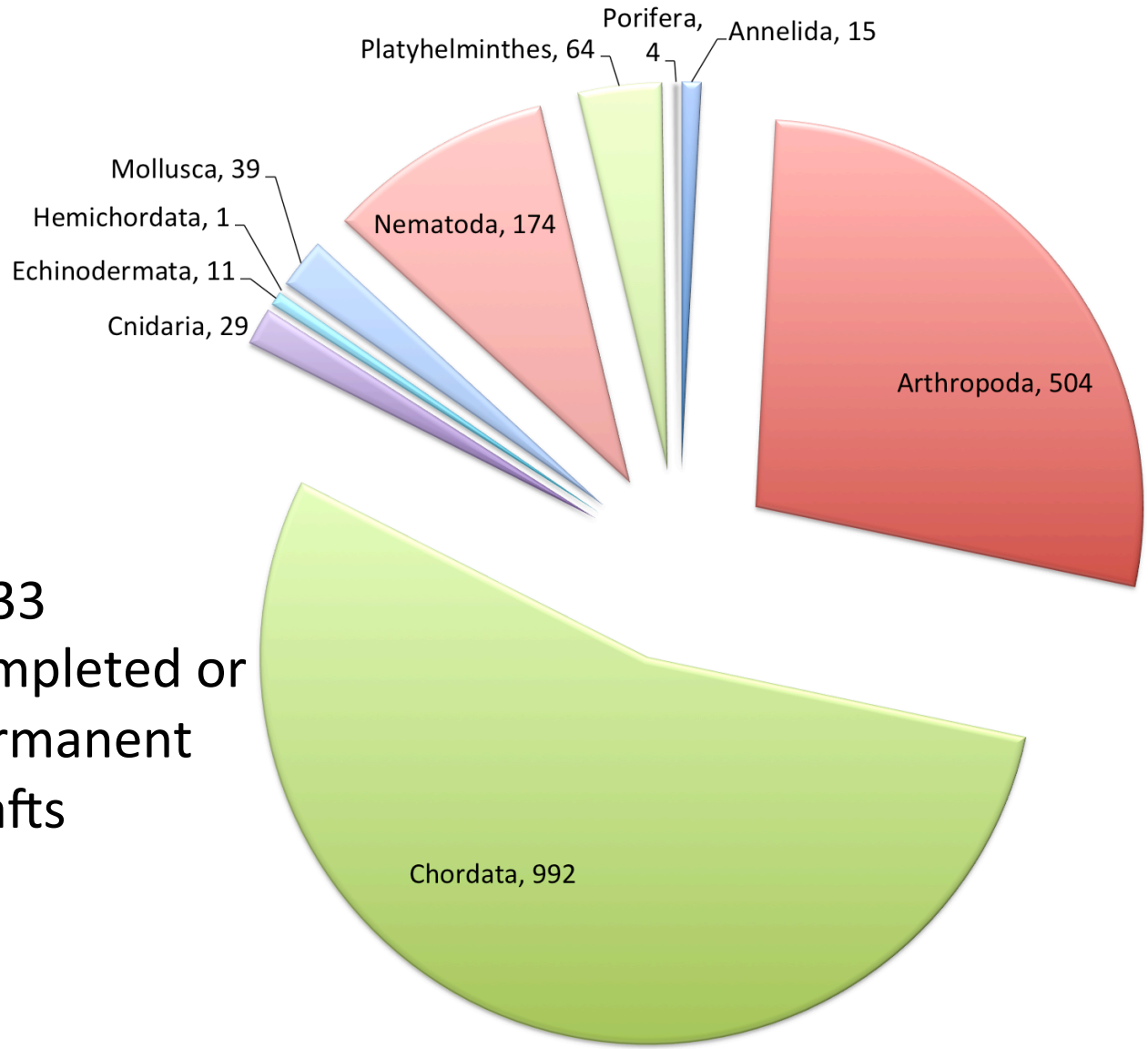
(GOLD database)

Bacteria are (of course) the most well-sequenced



(GOLD database)

But what about metazoans?

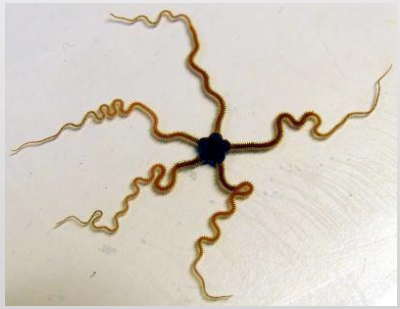


1833 completed or permanent drafts

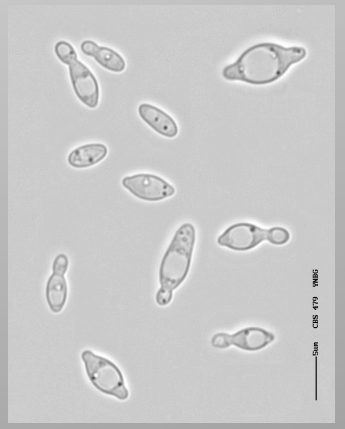
(GOLD database, Oct 2015)

IMAGO Project – 8 marine genomes







UNIVERSITY OF GOTHENBURG
CENTRE FOR MARINE EVOLUTIONARY BIOLOGY

IMAGO Project – 8 genomes

UNIVERSITY OF GOTHENBURG
CENTRE FOR MARINE EVOLUTIONARY BIOLOGY

IMAGO Marine Genome Project

Current status genome sequencing: ***2013-04-25***

CeMEB/IMAGO Organisms	Common name	Genome size (Gbp; haploid)	DNA libraries (fragment size)	Total number of Gbp	Total size assembly (Mbp)	contig N50	max contig size
Balanus improvisus	Bay barnacle	0.7 - 1.4	150, 300 & 3 000	180	509	1 514	61 kb
Amphiura filiformis	Brittlestar	2.5	300	66	958	937	220 kb
*Debaryomyces hansenii**	Marine yeast	0.0138	150	20	6-29	1 615 - 84 127	208 kb - 513 kb
Fucus vesiculosus	Bladderwrack	1.1	300	65	176	453	42 kb
Idotea balthica	Baltic isopod	N/A	N/A	N/A	N/A	N/A	N/A
Littorina saxatilis	Periwinkle	1.5	150, 300 & 5 000	150	466	852	25.1 kb
Pomatoschistus minutus	Sand goby	1	300	74	568	1 534	58 kb
Skeletonema marinoi	Diatome	0.05-0.1	300	32	49	1673	506 kb

Total = 587Gbp

https://github.com/The-Bioinformatics-Group

The Bioinformatics Group

at the department of Marine Sciences, University of Gothenburg

Gothnburg, Sweden http://marine.gu.se/ mats.topel@marine.gu.se

Repositories People 11 Teams 5

Filters Find a repository... + New repository

skeletonema_sex_project Shell ★ 0 ⑂ 0

This is the repository for the skeletonema, expression analysis - sex project.

Updated 3 hours ago

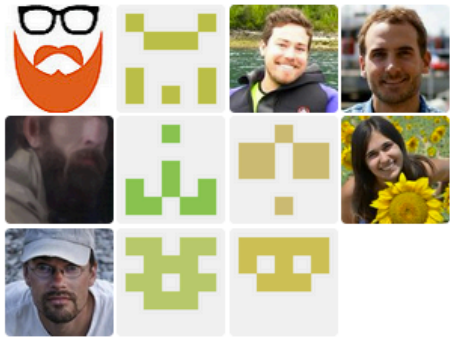
Idotea_balthica_transcriptome_project HTML ★ 0 ⑂ 1

Updated 6 hours ago

Debaryomyces_hansenii ★ 0 ⑂ 0

Updated 6 hours ago

People 11 >



One Reference Genome

- When possible, sequencing and assembling a reference genome gives you many advantages for analyzing reduced representation SNPs:
  - Easier bioinformatics, more analysis options
  - More complete filtering of paralogous loci
  - Analysis of linkage disequilibrium with respect to chromosomal proximity
    - or use genetically mapped markers
  - Identify large chromosomal variants (inversions)
  - Facilitates follow-up fine mapping
  - Allows for functional annotation

Low-coverage resequencing

- In cases where a genome is available, it is also possible to resequence pooled population wide samples at low coverage (1 x), and directly estimate allele frequencies without need for individual genotype samples.

- Useful for investigations of population structure, but lack of individual genotypes makes correlation to phenotype difficult.

Why Not Sequence Population Samples of Whole Genomes?

- Difficulty: Even with high sequencing throughput, high quality mate-pair libraries and bioinformatic assembly are still very challenging.

- Wasted effort: Whole genomes have a lot of uninformative DNA and in many populations with linkage disequilibrium, lots of redundant variation.

- Optimal sampling: For population genomics, most studies will benefit from extensive population sampling, so genomic sampling should be sufficient to answer the questions and no more.

Genomic Subsampling

- Restriction digestion based methods
  - A variety of RAD-seq variants (size-selected fragments)
  - Genotyping-by-Sequencing (GBS)

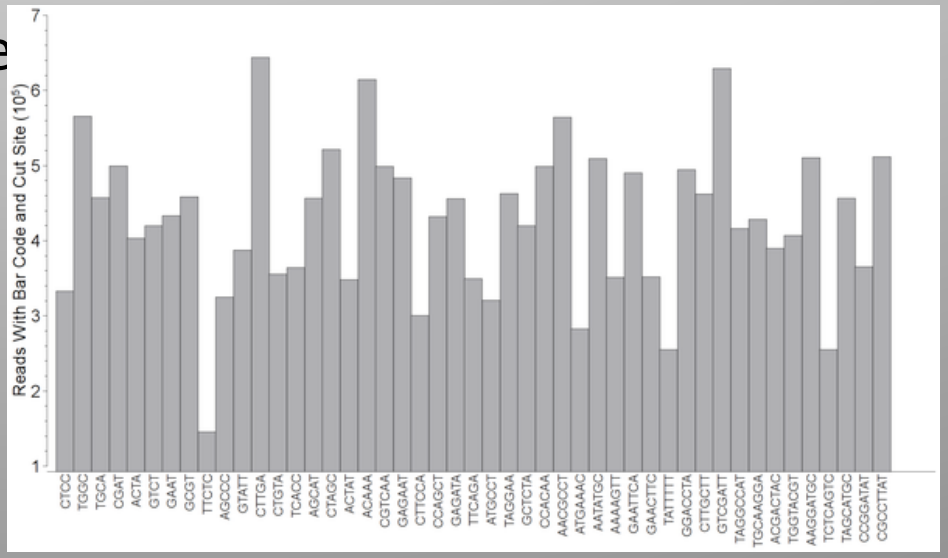
Random

- Targeted amplicon sequencing
- Targeted sequence capture
- Transcriptomes

Targeted

General Methodological Considerations

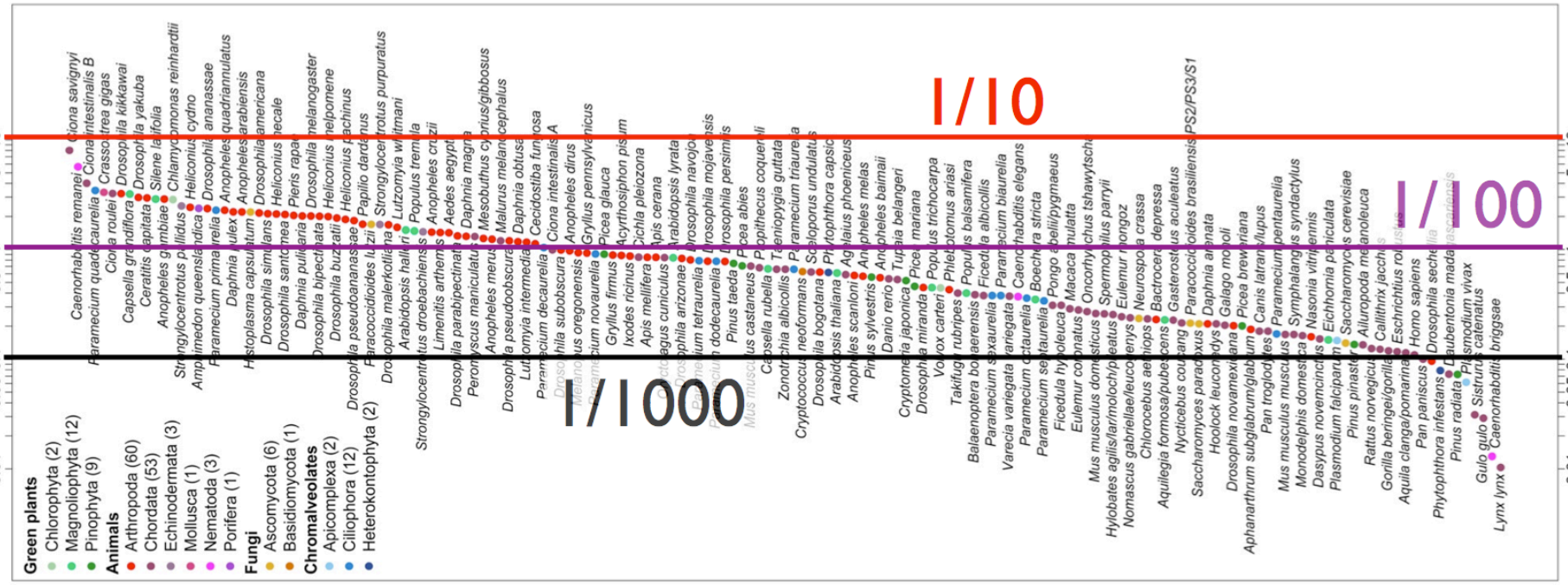
- How many markers do you need/want?
- Individual SNPs or do you need haplotype data?
- Barcode individuals with MID to get genotypes or sequence pools to estimate SNP frequencies?
- Perfectly even pooling is nearly impossible – anticipate high variance when planning sequencing coverage
- Pilot libraries help optimize
  trade offs



Elshire et al. 2011

How many markers?

Polymorphism rates are highly variable among species



Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, et al. (2012) Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? PLoS Biol 10(9): e1001388. doi:10.1371/journal.pbio.1001388

Marker (SNP/haplotype) development

Restriction-based methods (This afternoon)

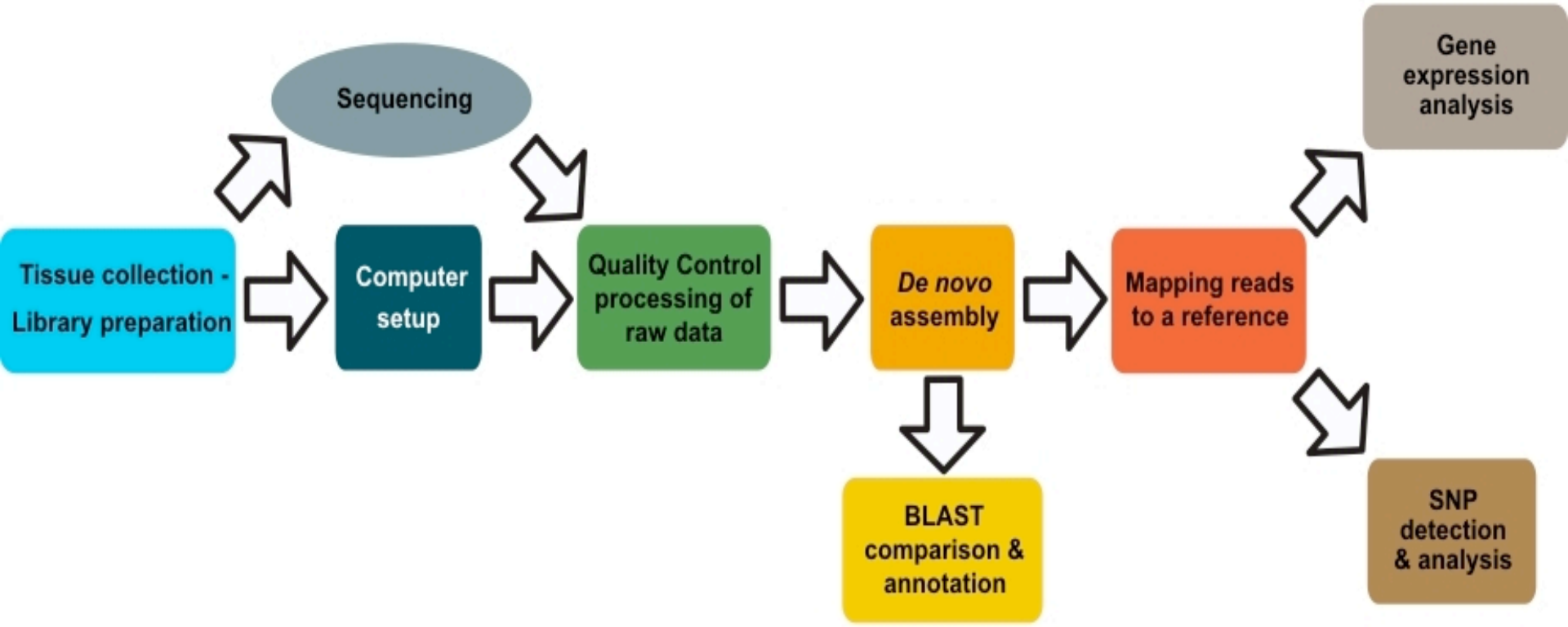
- Restriction enzymes produce digest genomic DNA and ligate adapters to cut ends.
- Amplify all fragments from the ends using adapter sequence as priming site.
- Sequence using different settings (SE/PE, read lengths) based on method used.
- Cluster short reads into loci using a clustering algorithm (Stacks, cd-hit-est for example)
- Scan loci for variant sites, sort into alleles either as SNP or haplotype data.

Exome sequencing – RNA-Seq for calling SNPs (Tomorrow)

- Pros
  - Huge reduction in genome complexity
  - Targeting functional fraction, most easily annotated
  - Benefit from minimal indels, predictable codon structure
  - Combine pop gen with expression phenotype studies

- Cons
  - More careful sampling & preservation required
  - Good data only from loci expressed in all samples at moderate – high levels
  - Complications from allele-specific expression, alternative splice variants

General pipeline for RNA-Seq analysis from a non-model species



De Wit et al. 2012

MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 2310–2323 doi: 10.1111/mec.13165

INVITED REVIEWS AND SYNTHESES

SNP genotyping and population genomics from expressed sequences – current advances and future possibilities

PIERRE DE WIT,* MELISSA H. PESPENI† and STEPHEN R. PALUMBI‡

*Department of Biology and Environmental Sciences, University of Gothenburg, Sven Lovén Centre for Marine Science – Tjärnö, Hättebäcksvägen 7, Strömstad, SE-452 96, Sweden, †Department of Biology, University of Vermont, Marsh Life Science, Rm 326A, 109 Carrigan Drive, Burlington, VT 05405, USA, ‡Department of Biology, Stanford University, Hopkins Marine Station, 120 Ocean view Blvd., Pacific Grove, CA 93950, USA

Abstract

With the rapid increase in production of genetic data from new sequencing technologies, a myriad of new ways to study genomic patterns in nonmodel organisms are currently possible. Because genome assembly still remains a complicated procedure, and because the functional role of much of the genome is unclear, focusing on SNP genotyping from expressed sequences provides a cost-effective way to reduce complexity while still retaining functionally relevant information. This review summarizes current methods, identifies ways that using expressed sequence data benefits population genomic inference and explores how current practitioners evaluate and overcome challenges that are

Other Targeted Methods

Multiplexed Sequencing of Amplicons

- MiSeq 2 x 300
- Commonly used for metagenetic studies
- Becomes laborious for multiple markers.
- Good option for small projects sharing a MiSeq run

- **Friday!**

Sequence Capture

- More costly up-front, enables high coverage population sampling with large genomes

- Oligonucleotide baits in solution by SureSelect, Nimblegen, Raindance, MYcroarray

- Efficient capture using baits from heterologous species < 12% divergent (Hancock-Hanser et al. 2013)

- **Maybe next year!**

What is the Best Strategy?
It Depends...

- How many markers do you need/want?
- Individual SNPs or do you need haplotype data?
- Do you want to do individual-level analyses?
- Do you want to sample regulatory variants?
- Do you want regularly spaced markers along chromosomes?
- Are annotations important to you?