Food and Agriculture Organization of the United Nations

UNIVERSITÀ CATTOLICA del Sacro Cuore

Guidelines on the genomic characterization of animal genetic resources

Section 3 - Genomic tools and methods

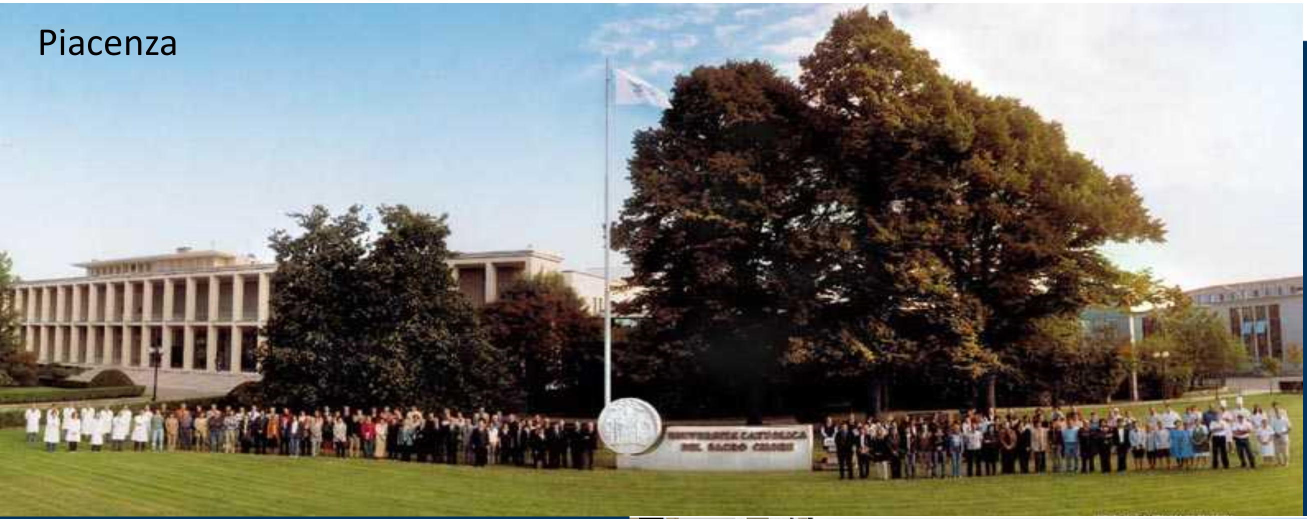
Paolo Ajmone-Marsan & Licia Colli

Dipartimento di Scienze Animali della Nutrizione e degli Alimenti DIANA
Università Cattolica del S. Cuore di Piacenza

paolo.ajmone@unicatt.it; licia.colli@unicatt.it

Faculty of Agriculture, Food and Environmental Science

UNIVERSITÀ CATTOLICA del Sacro Cuore



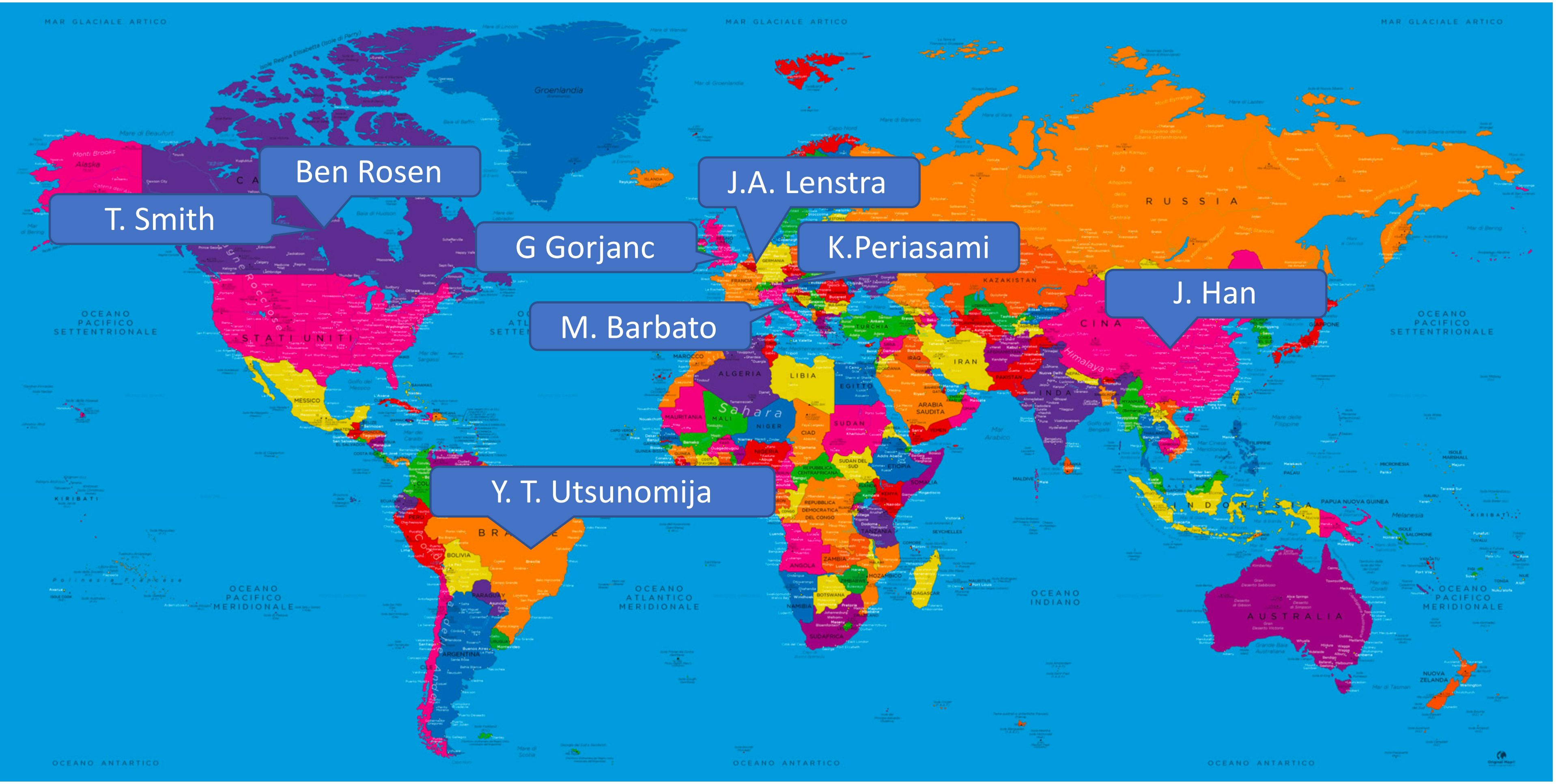
Piacenza

In the hearth of the Italian «Food Valley»

Cremona

Subsections

Contributors



T. Smith

Ben Rosen

G Gorjanc

J.A. Lenstra

K.Periasami

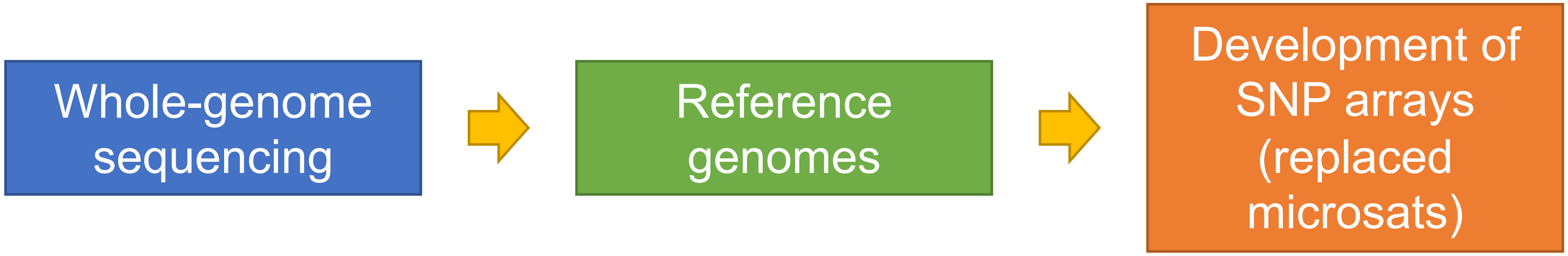
J. Han

M. Barbato

Y. T. Utsunomija

Introduction

Since the appearance of the **previous FAO Guidelines** on *Molecular genetic characterization of animal genetic resources* (**2011**), there has been a **spectacular growth** in the use of **genomic technologies**.



Wider adoption of WGS technologies:

- More animals were fully sequenced.
- Improvement of reference genomes.
- Development of resequencing and genotyping-by-sequencing (GBS) approaches.

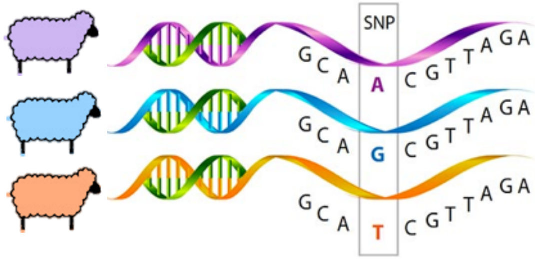
collection of **WGS** datasets **10-1000 individuals**.

vast multi-locus genotype **datasets**.

possibility of **data imputation.**

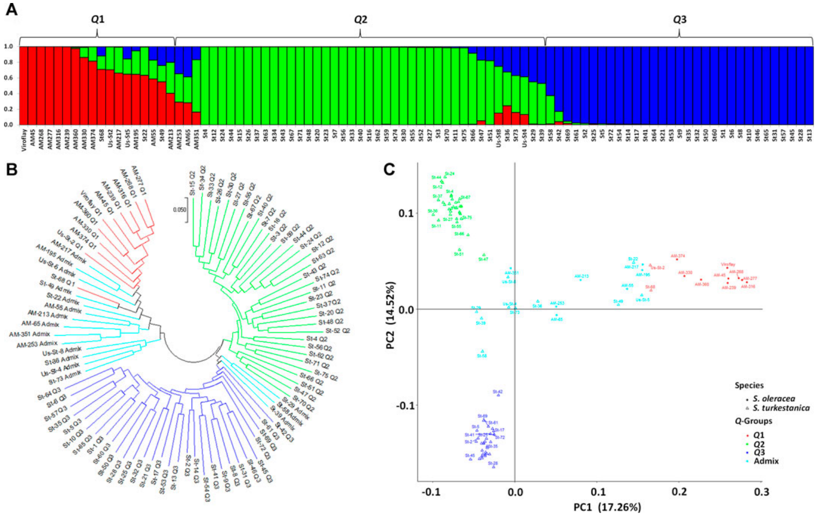
SNPs and SNP genotyping

Single Nucleotide Polymorphism (SNP): DNA sequence **variation** that occurs by **substitution of a nucleotide** at a specific position in the genome.



The SNPs

- are the **most common type of polymorphism** (ca. 1 SNP per 1 kb in most mammalian genomes);
- have become the **marker of choice** and have **replaced microsatellites** to:
  - assess genetic **diversity**, **structure** and **relationships** among populations.
  - identify **genomic regions associated** to **economic traits**.



https://www.frontiersin.org/files/Articles/740437/fgene-12-740437-HTML/image_m/fgene-12-740437-g003.jpg
https://static.wixstatic.com/media/15b941_47df96ee398a416f9dbadfcae7369219~mv2.png/v1/fit/w_961%2Ch_516%2Cal_c/file.png

SNPs and SNP genotyping

SNPs advantages over microsatellites:

(i) stable inheritance.

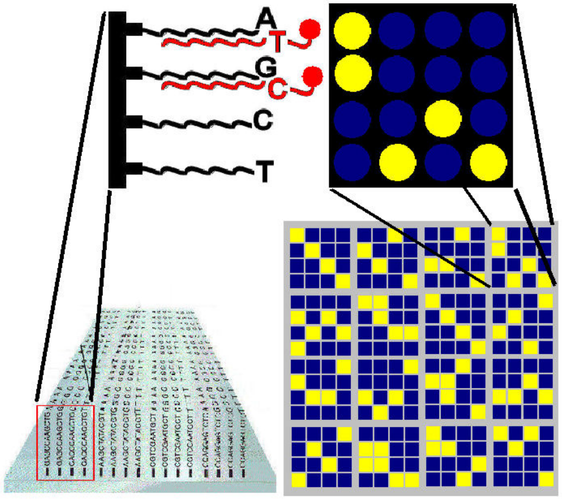
(ii) distribution throughout the genome at a **greater density**.

(iii) location in coding regions that can possibly alter protein function and phenotypic expression.

(iv) location nearby or within quantitative trait loci (**QTL**) of interest.

(v) suitability for **high throughput genotyping**.

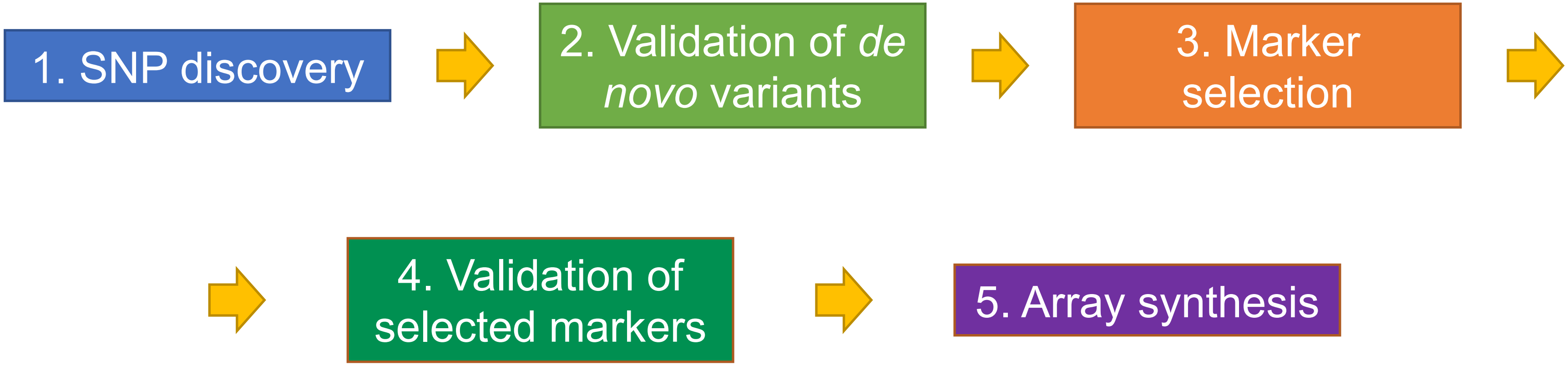
→ SNPs allow a greater **standardization of data** production and the **development** of species-specific **commercial arrays**.



https://www.mun.ca/biology/scarr/VDA_schematic_Carr_et_al_2007c.jpg

SNP array design

Key steps in SNP array design:



1. SNP discovery

2. Validation of *de novo* variants

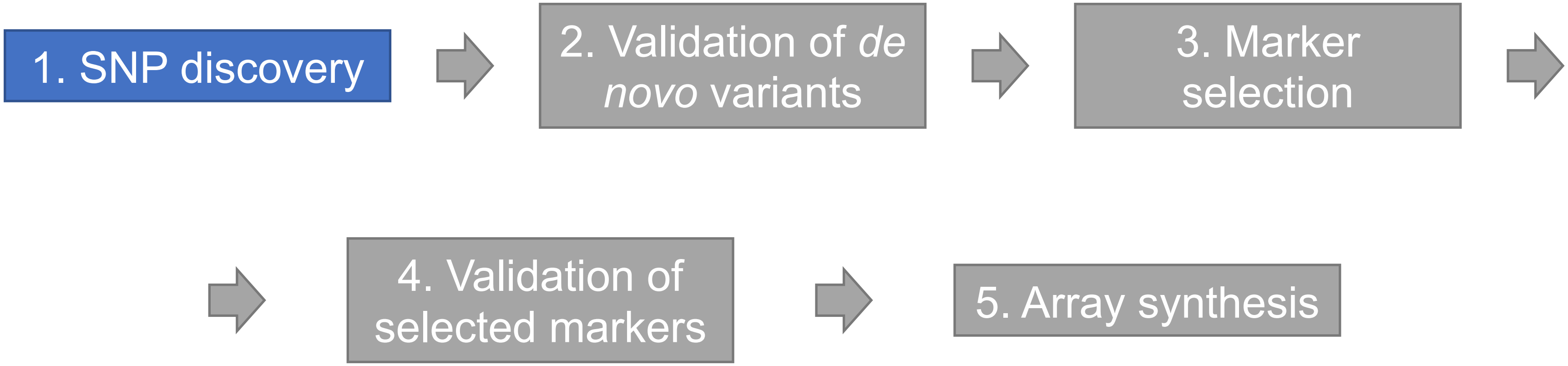
3. Marker selection

4. Validation of selected markers

5. Array synthesis

SNP array design

Key steps in SNP array design:



1. SNP discovery

2. Validation of *de novo* variants

3. Marker selection

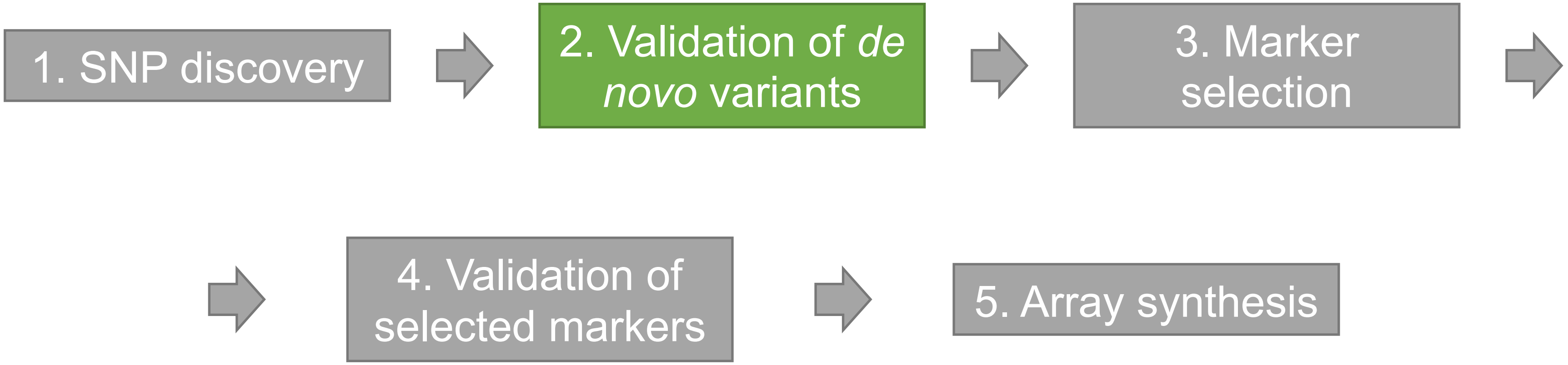
4. Validation of selected markers

5. Array synthesis

Ideal **SNP discovery** process: WGS of a **panel of unrelated individuals from diverse breeds** across widely dispersed geographical locations.

SNP array design

Key steps in SNP array design:



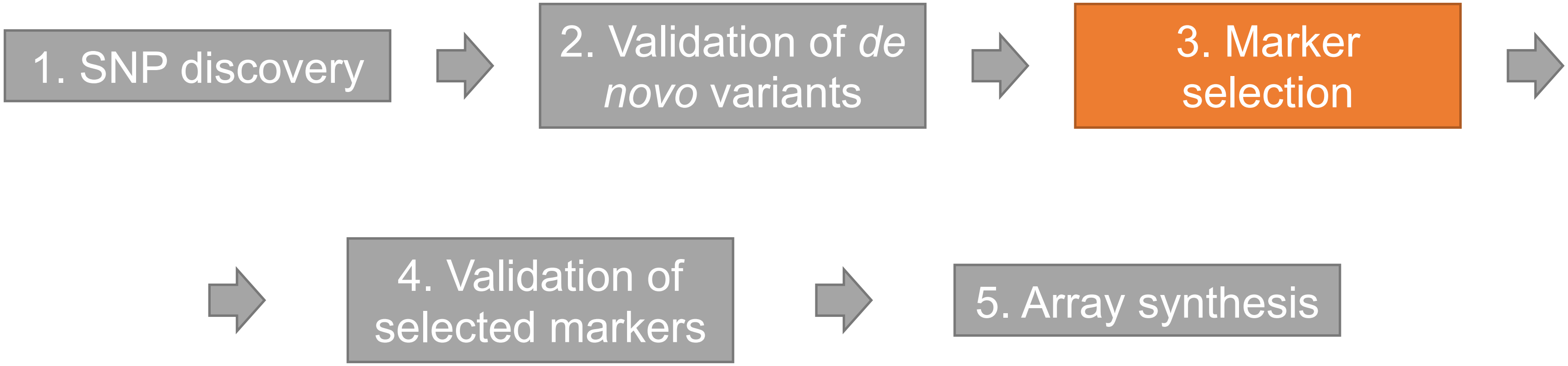
1. SNP discovery → 2. Validation of *de novo* variants → 3. Marker selection → 4. Validation of selected markers → 5. Array synthesis

Validation of *de novo* variants: Errors in genome sequencing → detection of false SNPs → ***de novo* variants** need to be **validated** based on sequence quality parameters.

SNP discovery + validation → identification of several million SNPs.

SNP array design

Key steps in SNP array design:



1. SNP discovery

2. Validation of *de novo* variants

3. Marker selection

4. Validation of selected markers

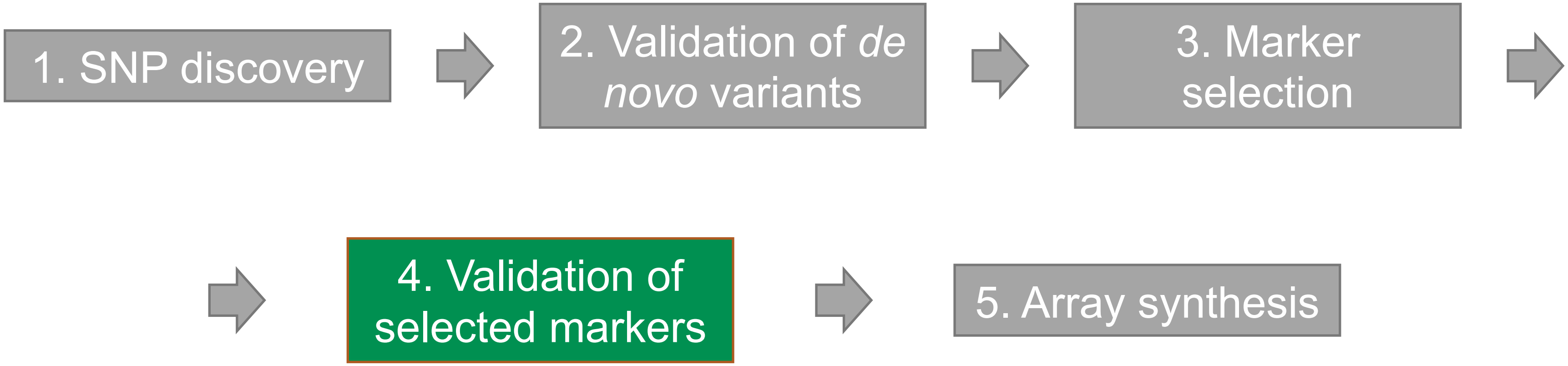
5. Array synthesis

→ **Sub-selection** of ~ 2-3 million SNPs **for drafting a marker panel** based on:

- likelihood of success in the genotyping assay;
- type of polymorphism (i.e. transition vs. transversion);
- minor allele frequency (MAF);
- linkage disequilibrium (LD);
- physical distribution of SNPs (e.g. equidistant spacing over the genome);
- polymorphism in multiple populations;
- enrichment of specific regions of the genome (e.g. potentially associated with important phenotypes).

SNP array design

Key steps in SNP array design:



1. SNP discovery

2. Validation of *de novo* variants

3. Marker selection

4. Validation of selected markers

5. Array synthesis

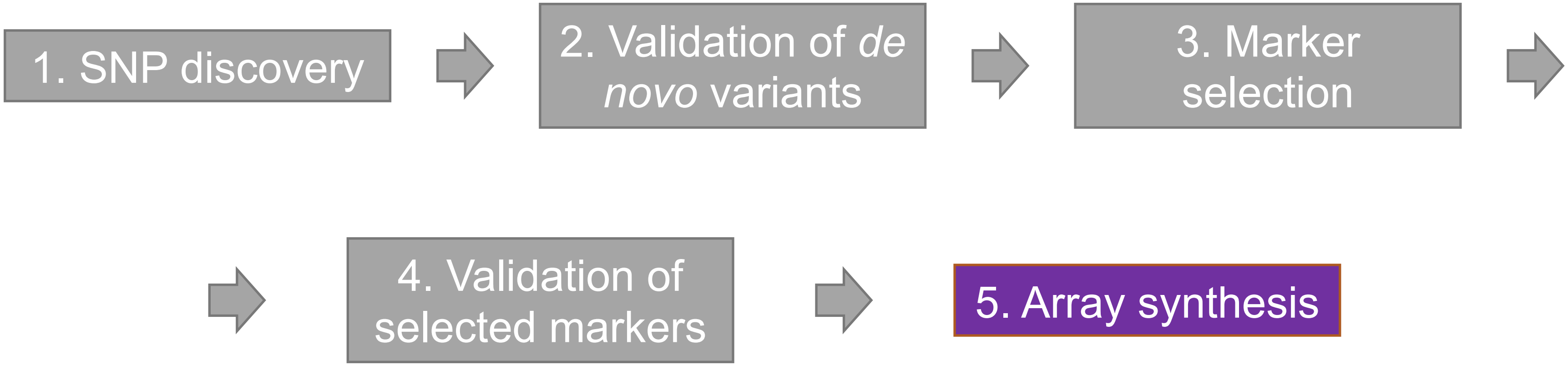
Validation of selected SNPs:

The **draft panel** of pre-selected markers must be **validated**.

Identification of a **subset of high-performance SNPs**.

SNP array design

Key steps in SNP array design:



1. SNP discovery

2. Validation of *de novo* variants

3. Marker selection

4. Validation of selected markers

5. Array synthesis

The **final selection** is carried out based on factors as:

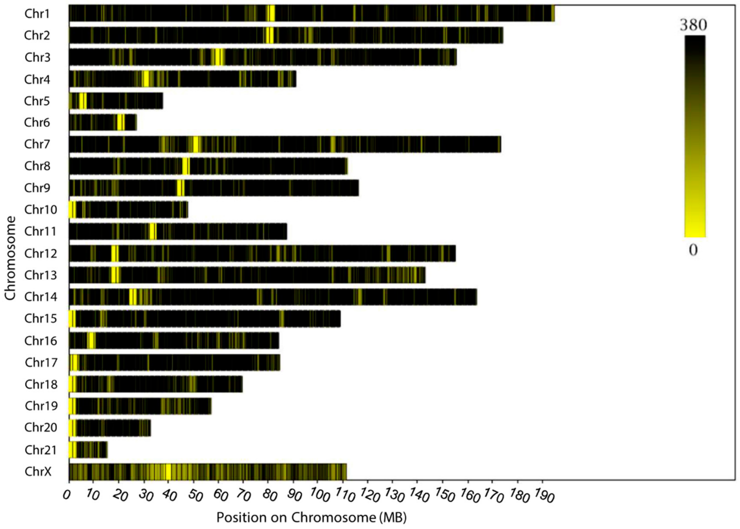
- SNP performance and informativeness;
- association with traits of interest;
- imputation of other variants in the genome;
- spacing and location with respect to LD blocks;
- functional significance of SNPs.

How to choose your array?

Choosing a suitable array for genotyping depends on the **purpose of the study**.

Factors to take into account:

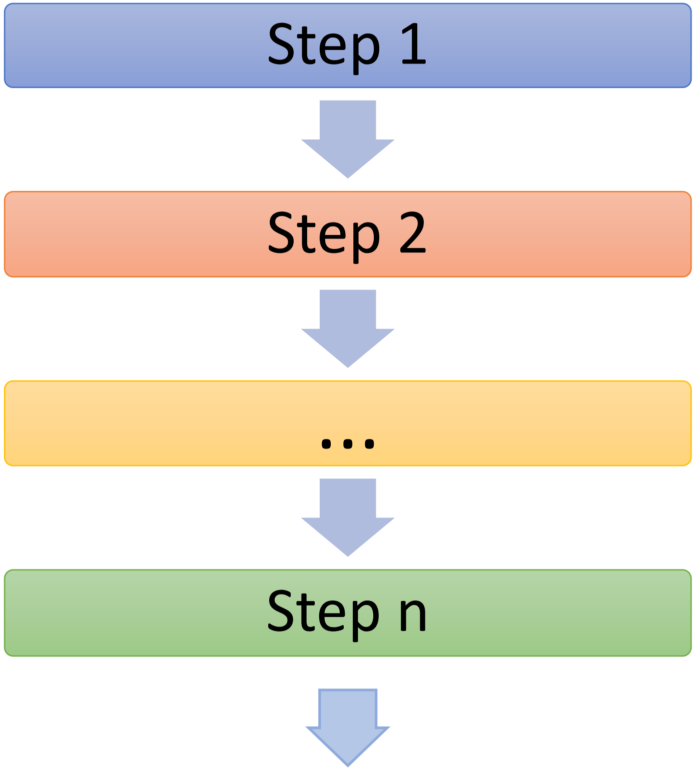
- **SNP density** (e.g. high, medium and low-density arrays);
- potential **Ascertainment Bias**;
- tagging of **specific genomic features** (e.g. parentage testing, copy number variants (CNV), detection of recessive traits);
- **SNPs in common** with **existing datasets**;
- **cost-effectiveness** (e.g. marker density vs cost);
- **performance** of the array (e.g. genome coverage).



https://www.mdpi.com/animals/animals-10-01068/article_deploy/html/images/animals-10-01068-g001.png

Preparing a working multilocus dataset

UNIVERSITÀ CATTOLICA del Sacro Cuore

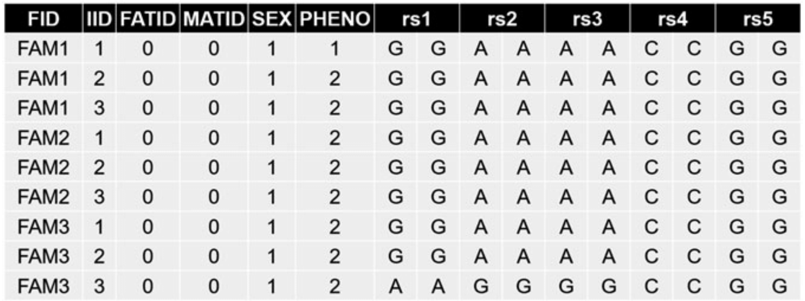


Step 1

Step 2

...

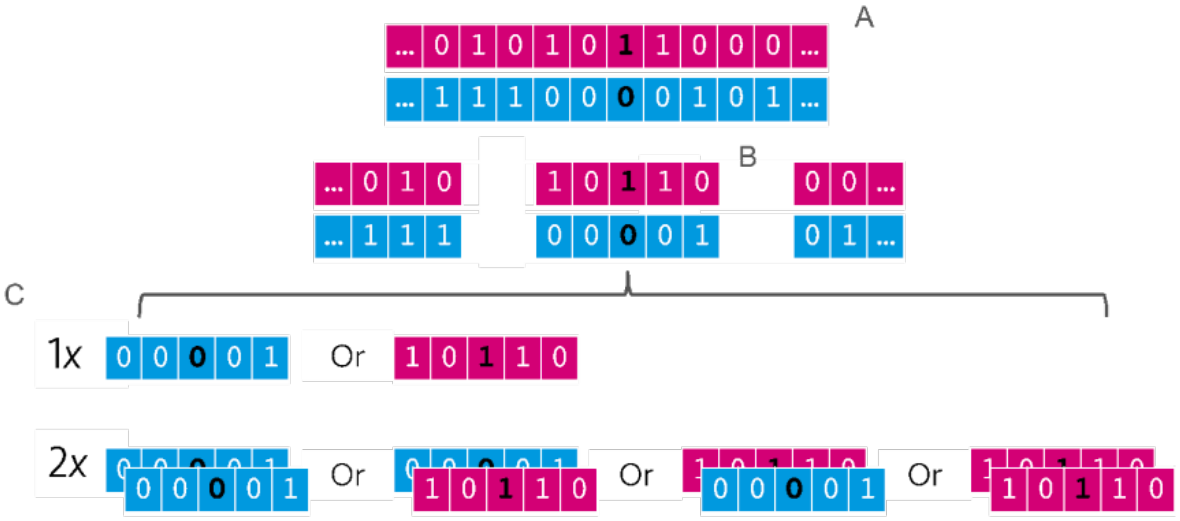
Step n



FID	IID	FATID	MATID	SEX	PHENO	rs1		rs2		rs3		rs4		rs5	
FAM1	1	0	0	1	1	G	G	A	A	A	A	C	C	G	G
FAM1	2	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM1	3	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM2	1	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM2	2	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM2	3	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM3	1	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM3	2	0	0	1	2	G	G	A	A	A	A	C	C	G	G
FAM3	3	0	0	1	2	A	A	G	G	G	G	C	C	G	G

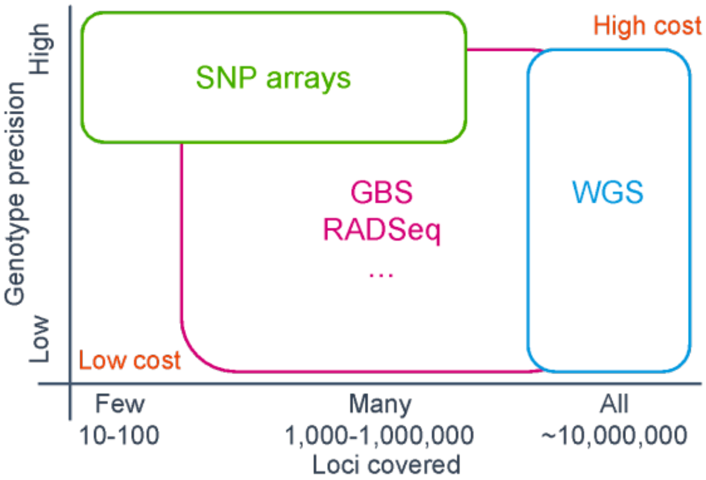
Genotyping by sequencing (GBS)

- for species with no standard SNP chip.
- to improve imputation



- low depth WGS
- Reduced genome representation
- Hybrid capture
- Fourteen different methods of GBS have been described (Scheben, Batley and Edwards, 2017) and new ones are continually being proposed

Advantages and disadvantages of GBS



GBS requires purified **high-molecular weight DNA**.

Unique alleles in a population can be hard to distinguish from **sequencing artefacts**

Presence-absence variation due to **deletions or insertions cannot be scored reliably** at low sequencing depth.

Output allele instead of genotype calls, GBS requires **additional bioinformatic attention** in downstream analyses.

Reveals genetic variation within **any livestock or wildlife population.**

Users can **tune it to their purpose** and budget by choosing the number of sequenced fragments and the depth of sequencing

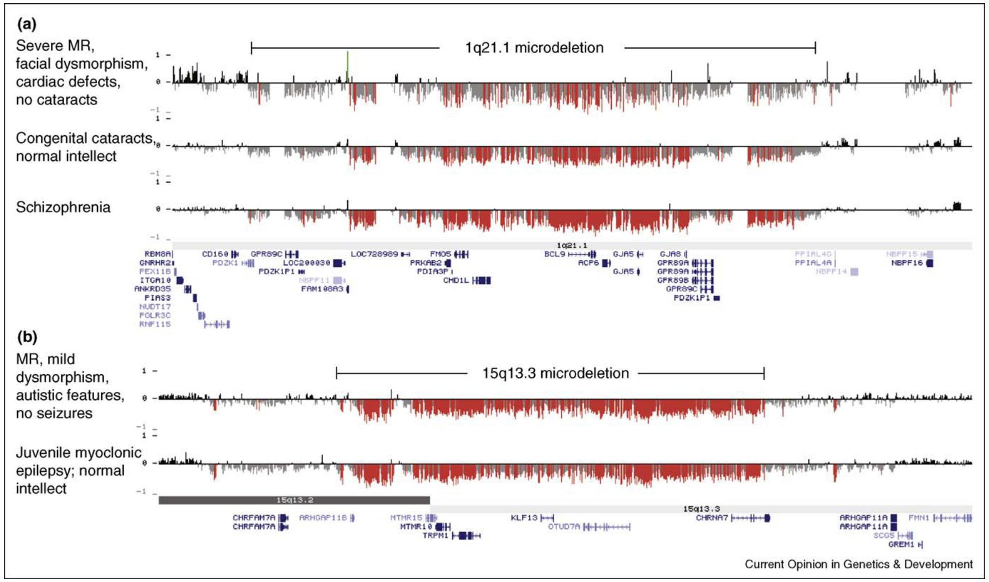
Avoids **ascertainment bias**

Whole genome sequencing

Detection of all variants: SNP, indels (insertions and deletions), CNV, and structural variation (SV) like inversions or large deletions.

Variables: Length, Depth, Cost.

Short reads (SRS)
Long reads (LRS)
HiFi



Current Opinion in Genetics & Development

CNVs in humans Mefford and Eichler 2009

Whole genome sequencing

Long-read sequencing

Repetitive elements

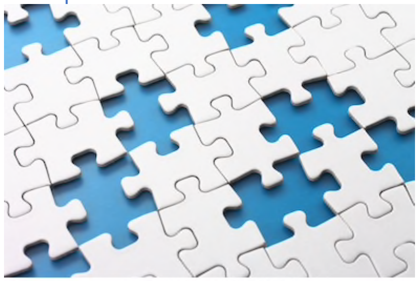
High error rate

Long to very long reads

Expensive

Need for long undamaged DNA

Easier genome assembly



shutterstock.com · 128939042

Short-read sequencing

No repetitive elements

Low error rate

Short reads

Cheap

Partially fragmented DNA is ok

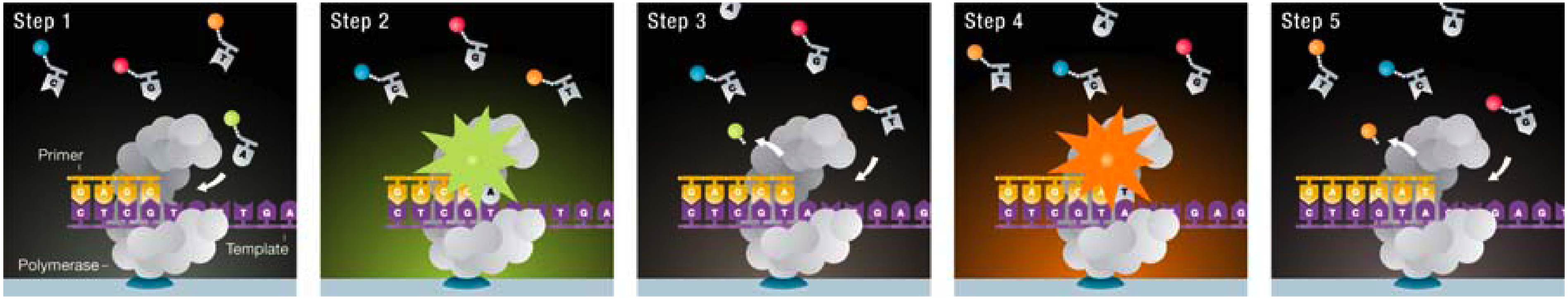
Difficult genome assembly

Possible combined or hybrid approaches

Problem in assembling some DNA regions from cells undergoing somatic rearrangements i.e. immunoglobulins

Long read sequencing: PacBio

Very fast and long reads by de-novo synthesis of single molecules



ZMW Detection chambers: 20 zeptoliters (10^{-21}l)

Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.

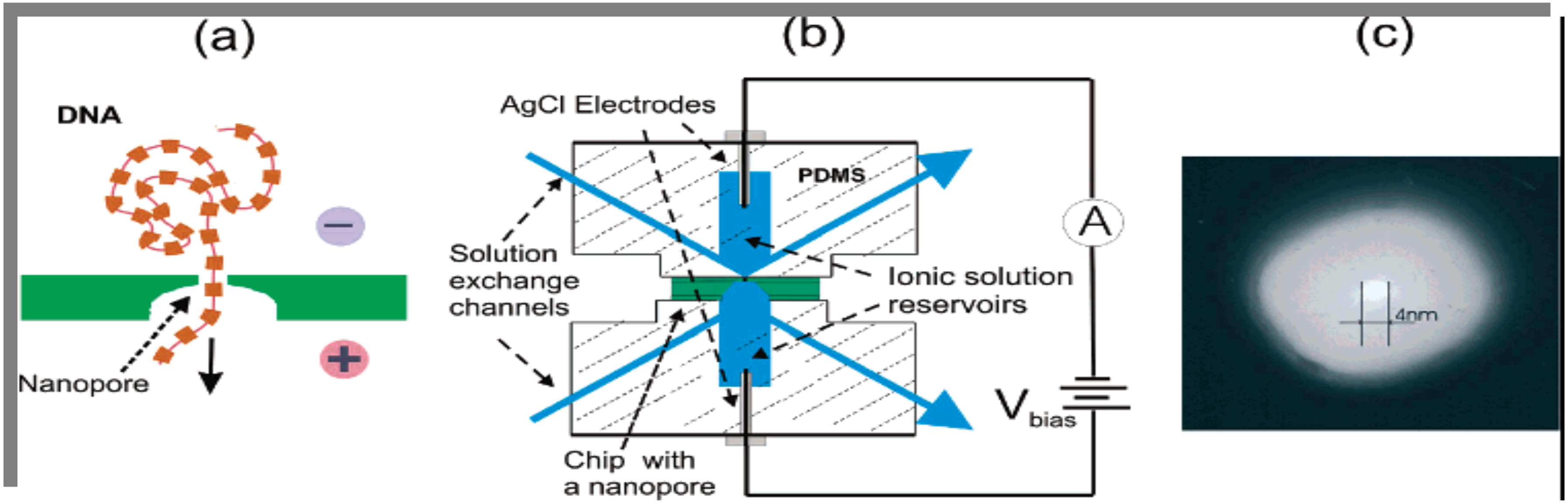
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.

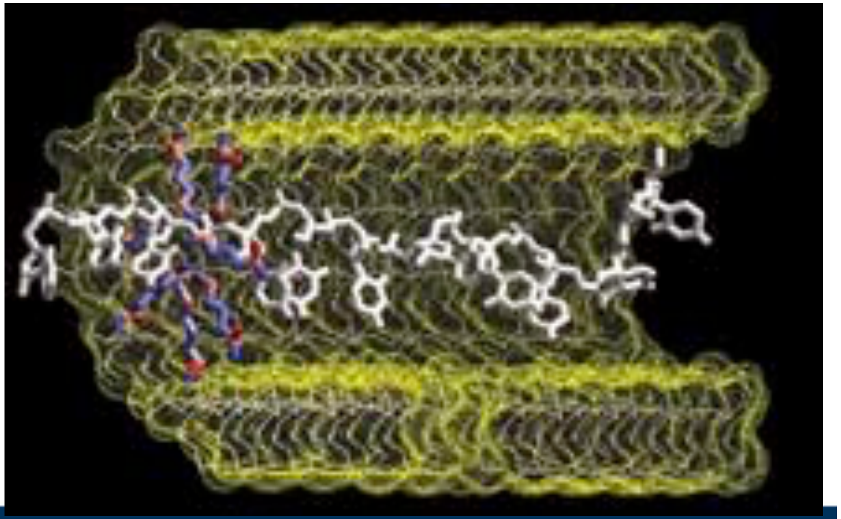
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.

Step 4-5: The process repeats.

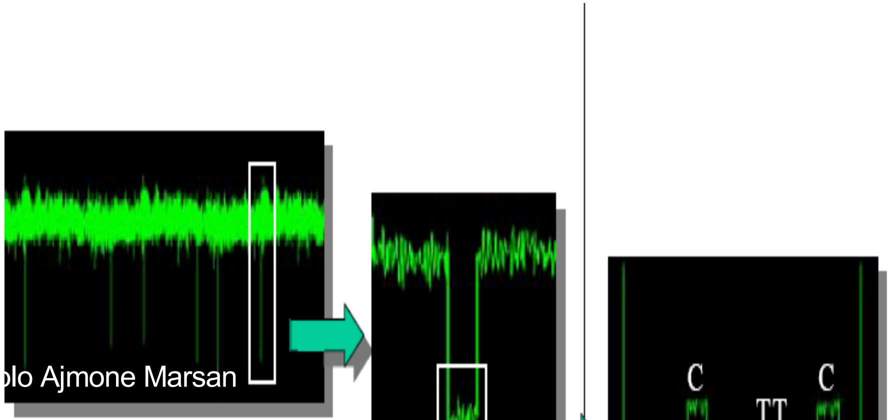
http://www.pacificbiosciences.com/index.php

Long read sequencing: Nanopore





Fologea et al, 2005. Nanoletters
Fologea et al, 2007. Electtophoresis



De-novo sequencing

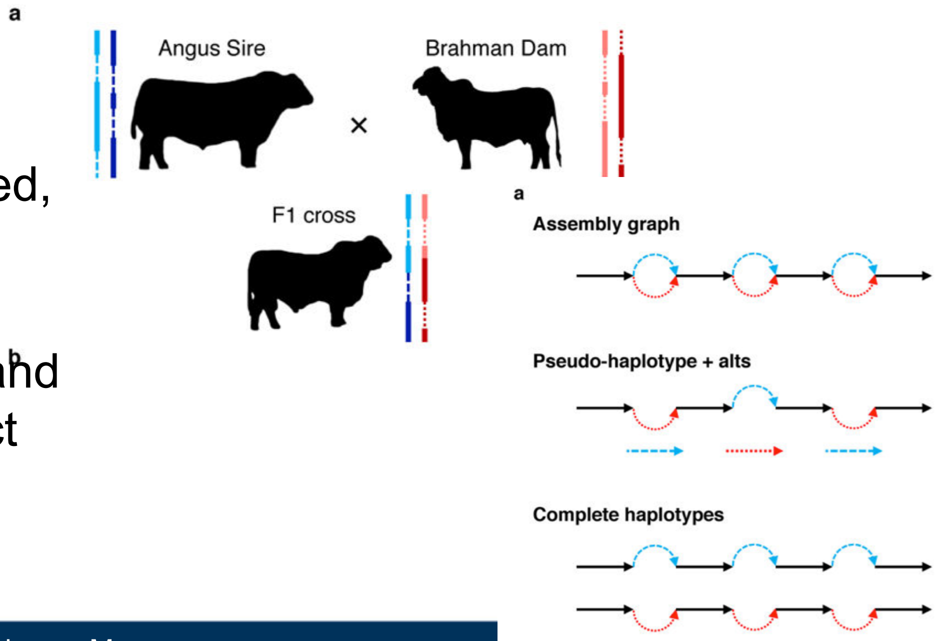
Sequence data is first assembled into “contigs,”

N50 size: the length of the contig where the sum of all longer contigs is > 50 percent of the total assembly size

SRS: target depth: ≥ 100x; N50 = 100 kb

LRS: target depth: ≥ 50x; N50 > 70 Mb or 700x as long.

Contigs from LRS need to be “polished” (i.e., checked and corrected) to increase accuracy, with SRS data being useful for this step



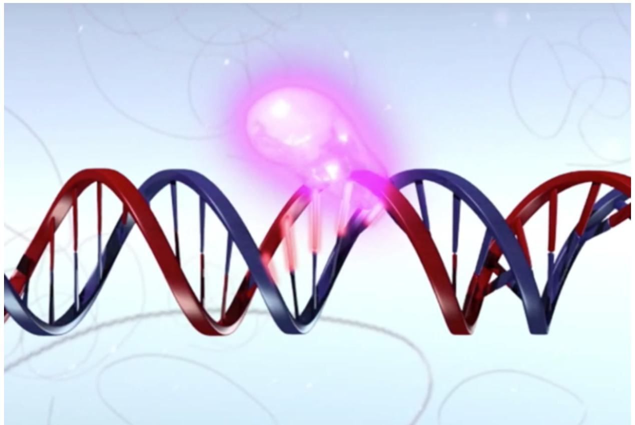
If a haplotype-resolved assembly is desired, then a greater depth is helpful.

“trio-binning, which utilizes divergence between two parental species or breeds and LRS of an F1 to create two, almost perfect haploid assemblies.

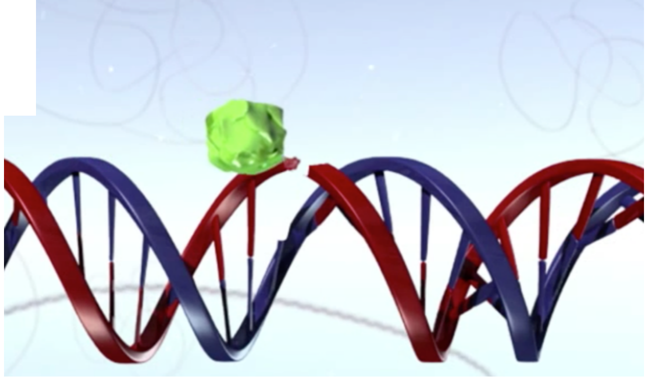
Koren et al., Nature Biotech. 2019

Optical mapping

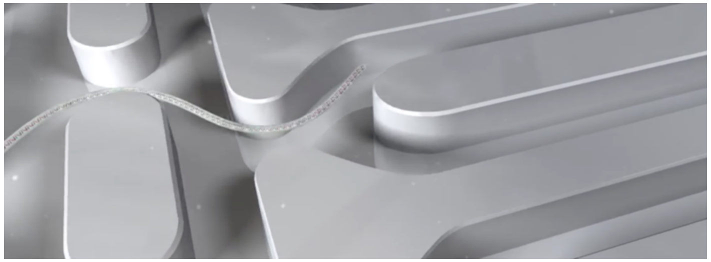
BIONANO GENOMICS



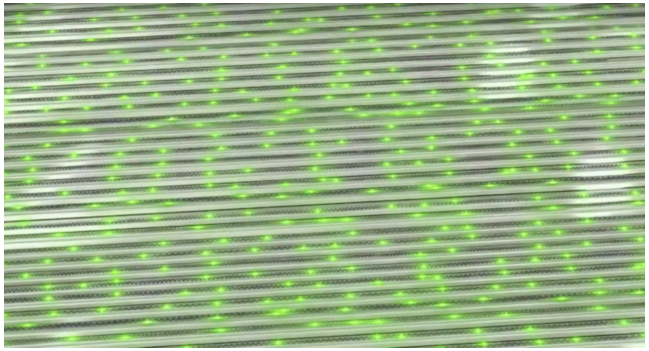
A 7bp recognising enzyme creates nicks in DNA at specific sites



Nicks are repeared with fluorescent nucleotides



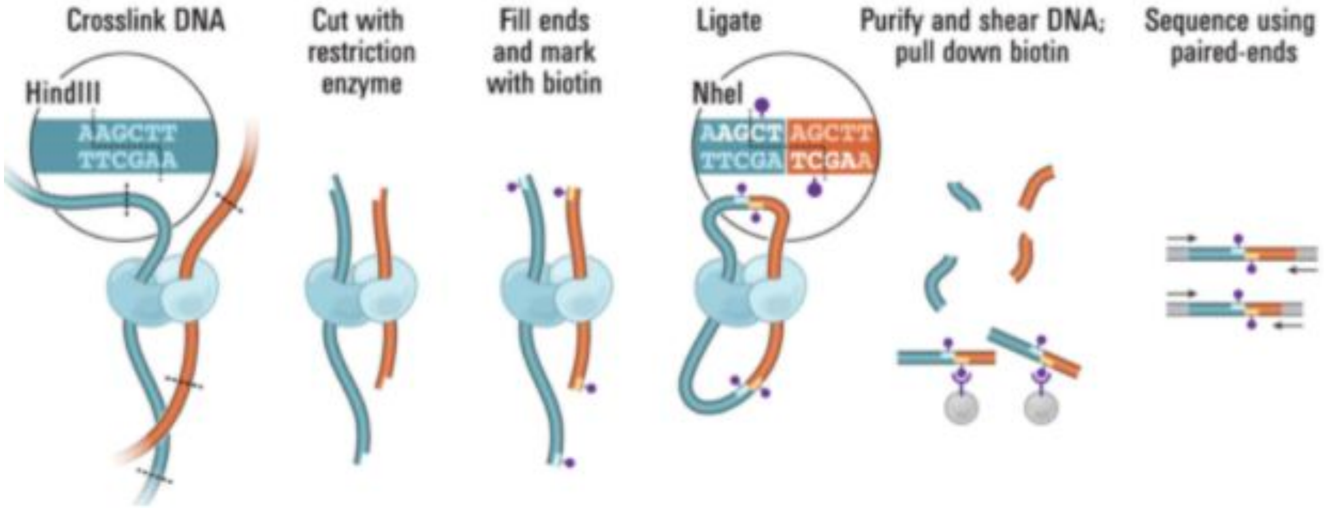
Single DNA molecules are separated and linearized in a nanochip



The fluorescence pattern of each DNA molecule is recorded

Hi-C sequencing

Sequencing of proximity-ligation products


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Lieberman-Aiden *et al.,* 2009
Nagano *et al.,* 2013

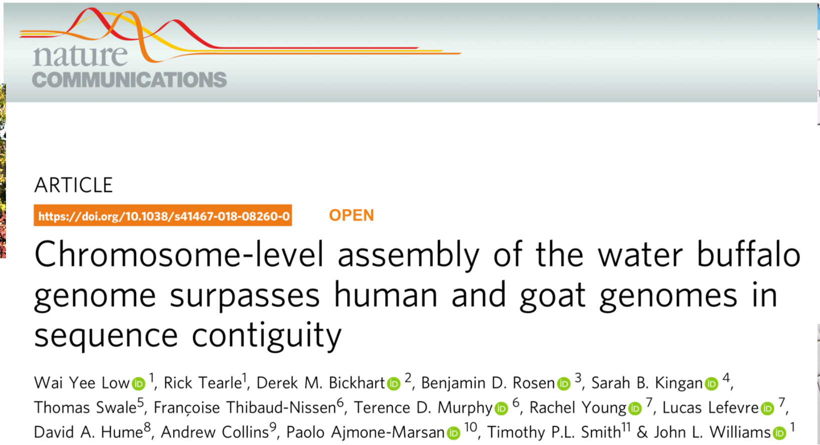
De-novo sequencing

Y chromosomes challenging to assemble due to their highly repetitive nature.

mtDNA: to be assembled prior to polishing, to avoid over polishing nuclear insertions of mitochondrial sequence (NUMTs) which can lead to difficulties in identifying mitochondrial variants.

BUFFALO: PacBio + Illumina sequencing + Chicago reads of Hi-C chromatin interaction → 383 gaps. **Contiguity and accuracy higher than human and goat**.

The power of combining approaches



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-018-08260-0 OPEN

Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity

Wai Yee Low[1], Rick Tearle[1], Derek M. Bickhart[2], Benjamin D. Rosen[3], Sarah B. Kingan[4], Thomas Swale[5], Françoise Thibaud-Nissen[6], Terence D. Murphy[6], Rachel Young[7], Lucas Lefevre[7], David A. Hume[8], Andrew Collins[9], Paolo Ajmone-Marsan[10], Timothy P.L. Smith[11] & John L. Williams[1]

Resequencing for detecting variation

Lower coverage than *de-novo*

Reads are mapped to a reference genome to detect variants

SRS→ 10X; SNPs and short indels. Need much higher depth to cover CNVs

LSR → 20X; SNP, indels and CNVs

HiFi → 10X; SNP and smaller variants than LSR

When no reference Y chromosome
reads that do not map to the reference collected and aligned to the Y chromosome of a closely related animal, or a collection of Y chromosome genes.

Single or breed specific reference genome? Recent efforts focused on a cattle "pangenome" (Heaton *et al*., 2021).

From raw data to variant calling

Raw sequence:

- FASTQ format, text based that includes a sequence identifier, the sequence and a quality score.

Preprocessing:

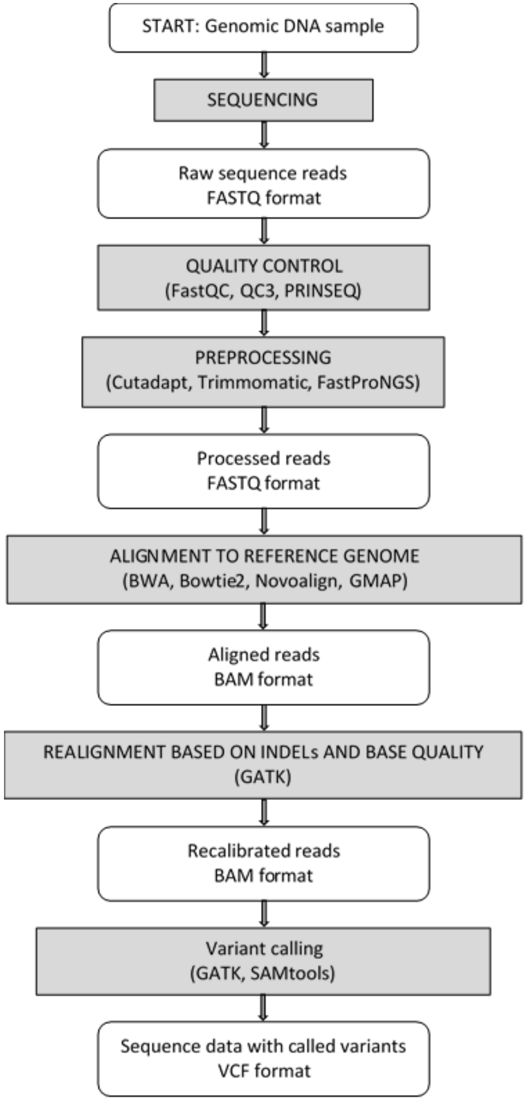
- filtering on quality based on parameters such as quality score, read length and content of guanine and cytosine nucleotides.
- Trimming removes sequence that corresponds to DNA adapters

Alignment:

- yields more information for each read (e.g. chromosomal location)
- basic format "SAM" for "sequence alignment/map format"
- But when large amount of data, binary format "BAM"
- Check and realign (e.g. indels) GATK software

Variant calling:

- VCF format, for "Variant Call Format." GATK can be used for variant calling, along with other software. The resulting data serves as the basis for further analyses.


START: Genomic DNA sample

SEQUENCING

Raw sequence reads
FASTQ format

QUALITY CONTROL
(FastQC, QC3, PRINSEQ)

PREPROCESSING
(Cutadapt, Trimmomatic, FastProNGS)

Processed reads
FASTQ format

ALIGNMENT TO REFERENCE GENOME
(BWA, Bowtie2, Novoalign, GMAP)

Aligned reads
BAM format

REALIGNMENT BASED ON INDELs AND BASE QUALITY
(GATK)

Recalibrated reads
BAM format

Variant calling
(GATK, SAMtools)

Sequence data with called variants
VCF format

Genotype imputation

Missing values cannot be processed in an analysis and thus should be either removed or replaced by a guess beforehand.

Guessing may help retain more **statistical power** in the analysis since it attempts to minimize data loss.

Use correlations between variables in order to fill up the empty data.
Prediction methods:

- Statistics
- machine learning
- deterministic approaches based on heuristics.

Target accuracy depends on missingness rate. If low, low accuracy is tolerated

Typical experiment:
smaller n. of animals WGS or genotyped HD and larger n. genotyped LD and then imputed the missing variants. Make sure most or all LD markers are in the HD panel.

Genotype imputation

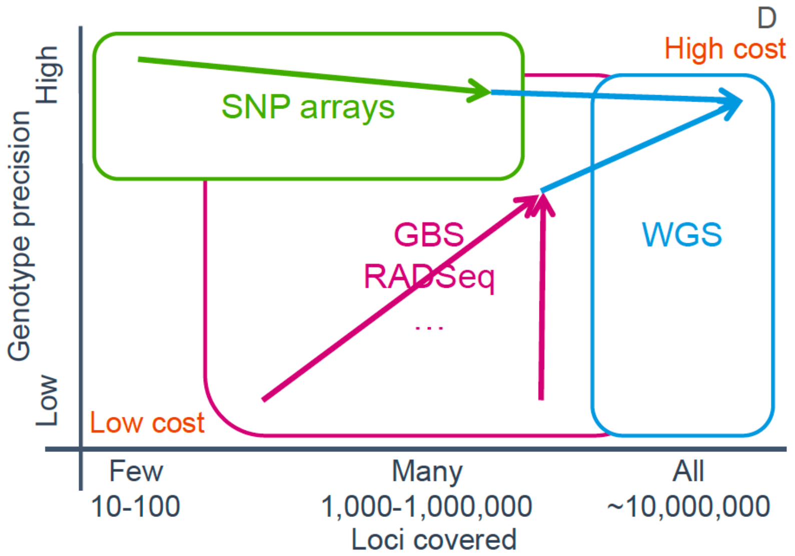
Reference animal panel (WGS or HD)

- **key ancestors** capturing a large proportion of the genetic variation in the population
- Otherwise selected on LDP data to **maximise haplotype diversity**
- if high accuracy must be achieved for low frequency variants, selection of animals carrying **rare haplotypes** should also be considered.

- **family-based**: haplotypes from close relatives used to impute the unobserved genotypes of LD samples
- **Population based**: pairs of individuals are assumed to share a common ancestor, such that LD samples are interpreted as mosaics of haplotypes that are present in the HD samples.
- there are methods that take advantage of both
- some methods require genotypes to be phased

Genotype imputation

Advantage in integration of different SNP Arrays and GBS: it increases the accuracy of WGS imputation and decreases cost (see Whalen, Gorjanc and Hickey 2020).



Accuracy is increased with:

- large **amount** of reference data
- reference animals ar closely **related** to those to be imputed.
- **High LD** between the loci with known genotypes and the loci to be imputed.
- **High mean LD** of the breed.

Thank you for your attention