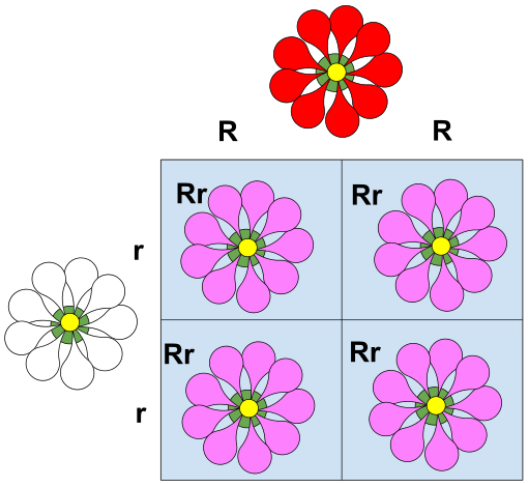
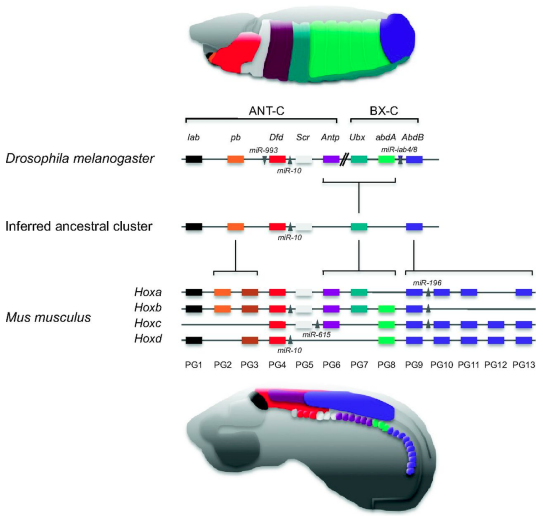
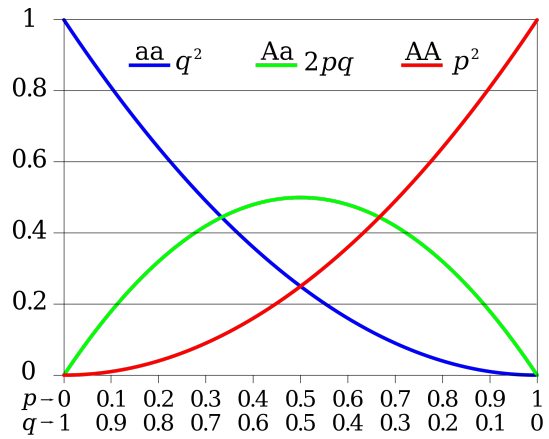
NGS application examples

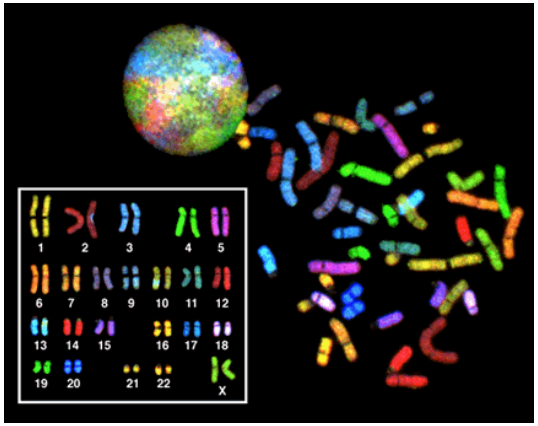
Jose Blanca
Joaquin Cañizares
COMAV institute
bioinf.comav.upv.es

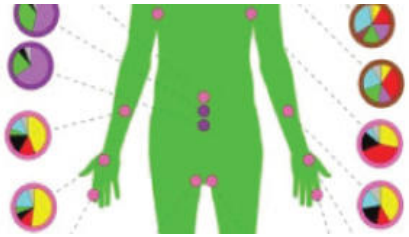
What can do with NGS?



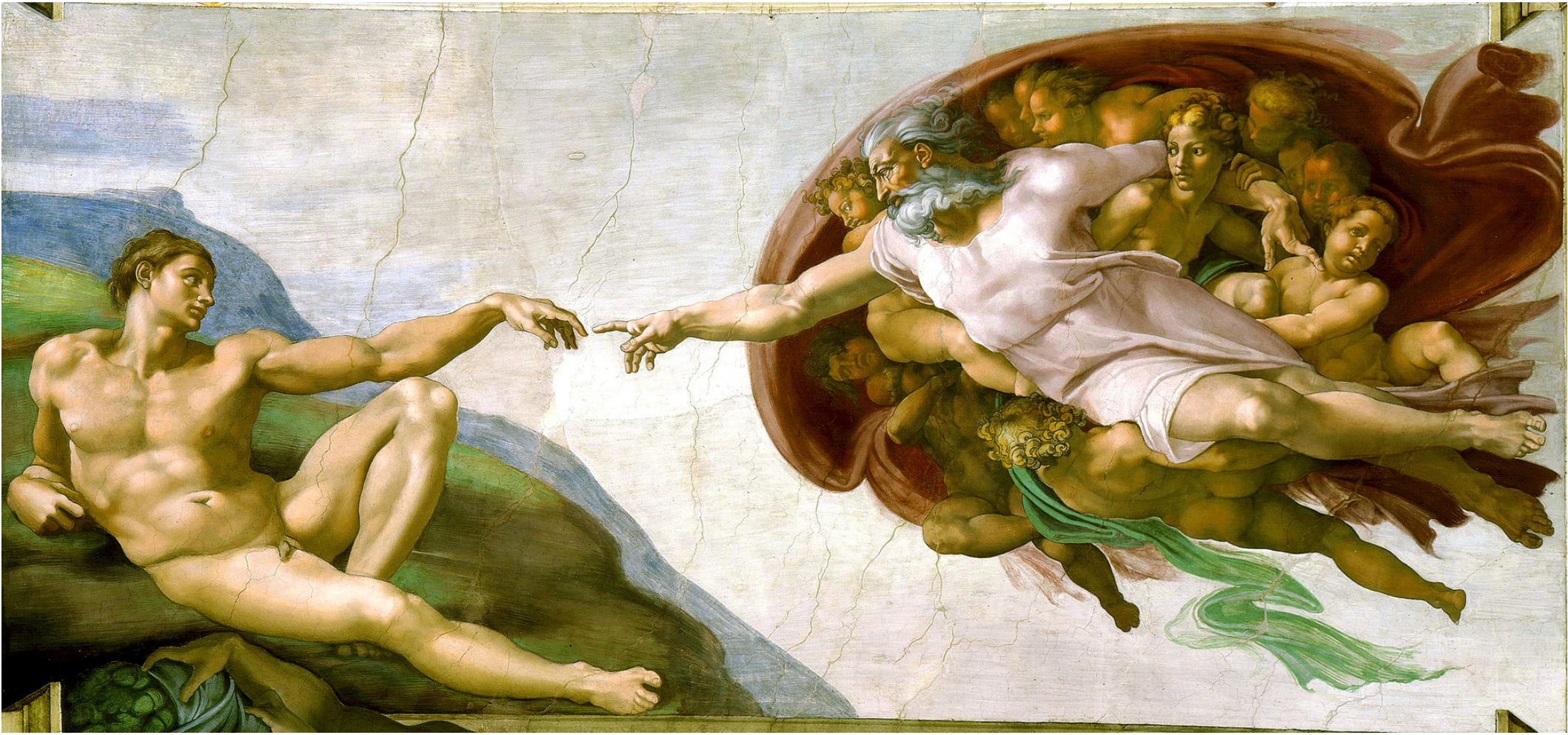




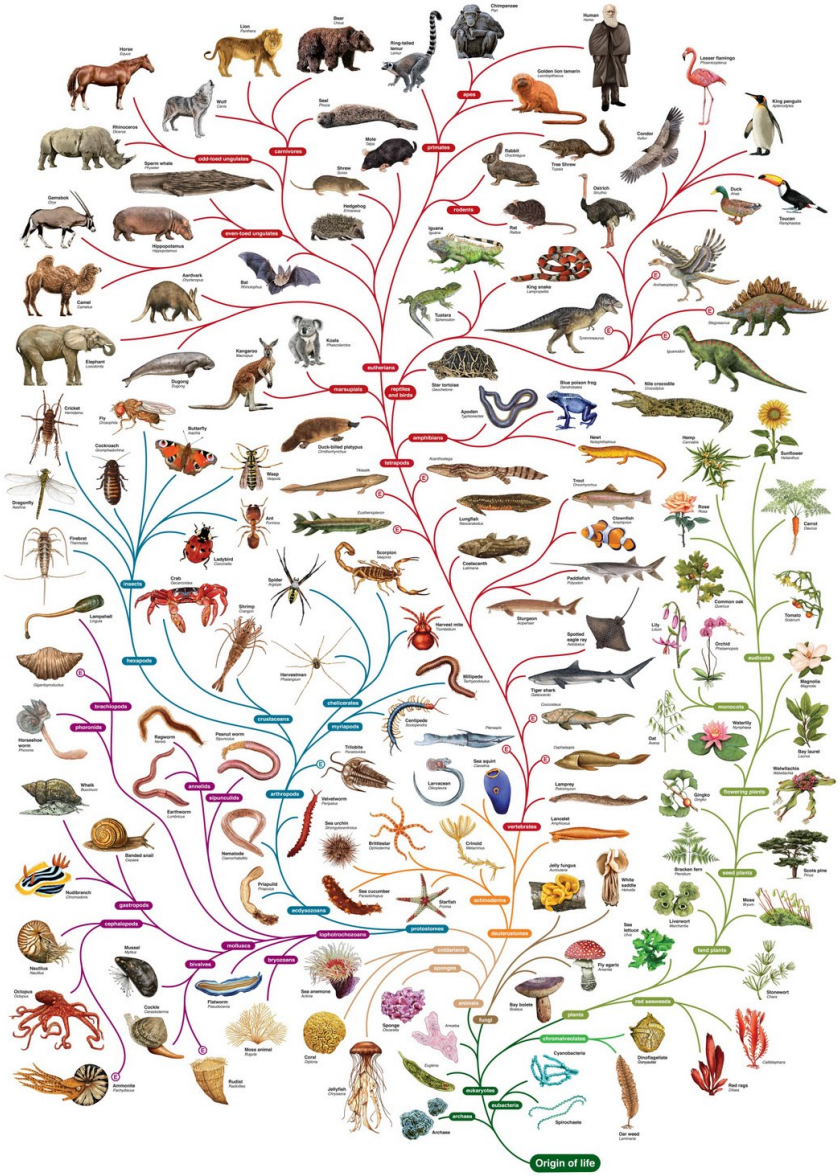




Is the NGS restricted to model organisms?



Is the NGS restricted to model organisms?



Is the NGS restricted to model organisms?

Marker discovery in *C. pepo* accessions

Materials

subsp. pepo cv.
Zucchini
MU16

subsp. ovifera cv
Scallop
UPV196

Method

454 RNA-Seq
2 accessions → Transcriptome → SNPs
SSRs
CAPs

Marker discovery in *C. pepo*

Zucchini MU16: 407,723 reads
Scallop UPV196: 392,370 reads

→

49,610 annotated transcripts

1,935 potential SSRs (86.7% polymorphic)

19,980 SNPs 1,174 INDELs

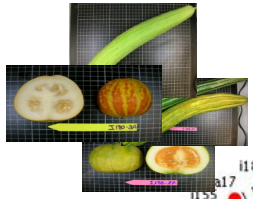
→

3,538 SNPs high quality and easy to use

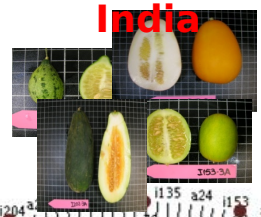
144 CAPs (80% validation rate)

SNP mining in melon

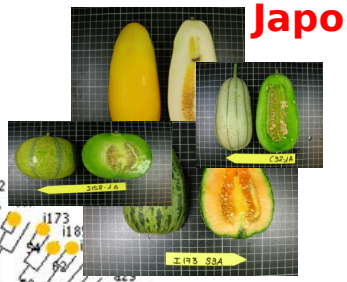
Dudaim, momordica y flexuosus
India, Iran Iral, Afganistan, Arabia



Agrestis, acidulus
India



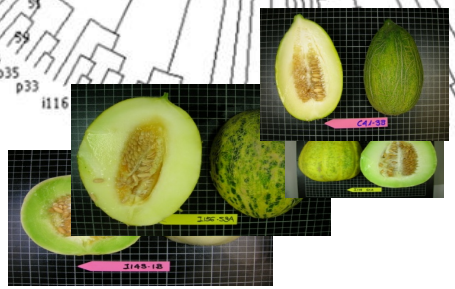
Conomon, makuwa, chinensis
Japon, corea



Inodorus: Piel de sapo
Piel de sapo S. Fito

Agrestis, tibish
Africa

Cantalupensis
Cantalupos S.Fito
Europa, Asia America



Inodorus,ameri,
chandalak, adana
Este de europa-Asia

Inodorus (Amarillo, tendral Rochet)
Amarillos.S.Fito
España

Marker discovery in *C. melo* pools

Subspecies	Pool	Reads
agrestis	African *agrestis*	30,620,160
	Asian *agrestis- acidulus*	15,779,803
	Far East *conomon*	17,962,640
	Middle East and Indian *momordica*, *dudaim* and *flexuosus*	23,320,668
melo	*cantalupensis*	23,237,004
	Group melo Eastern Europe, Central Asia, *inodorus*,*chandalack*, *ameri*	8,367,385
	inodorus Spanish landraces	17,485,023
	inodorus group market class Piel de sapo	13,809,773

Marker discovery in *C. melo* pools

260 million SOLiD reads

↓

303,883 SNPs and Indels
(18.8 SNVs per gene)

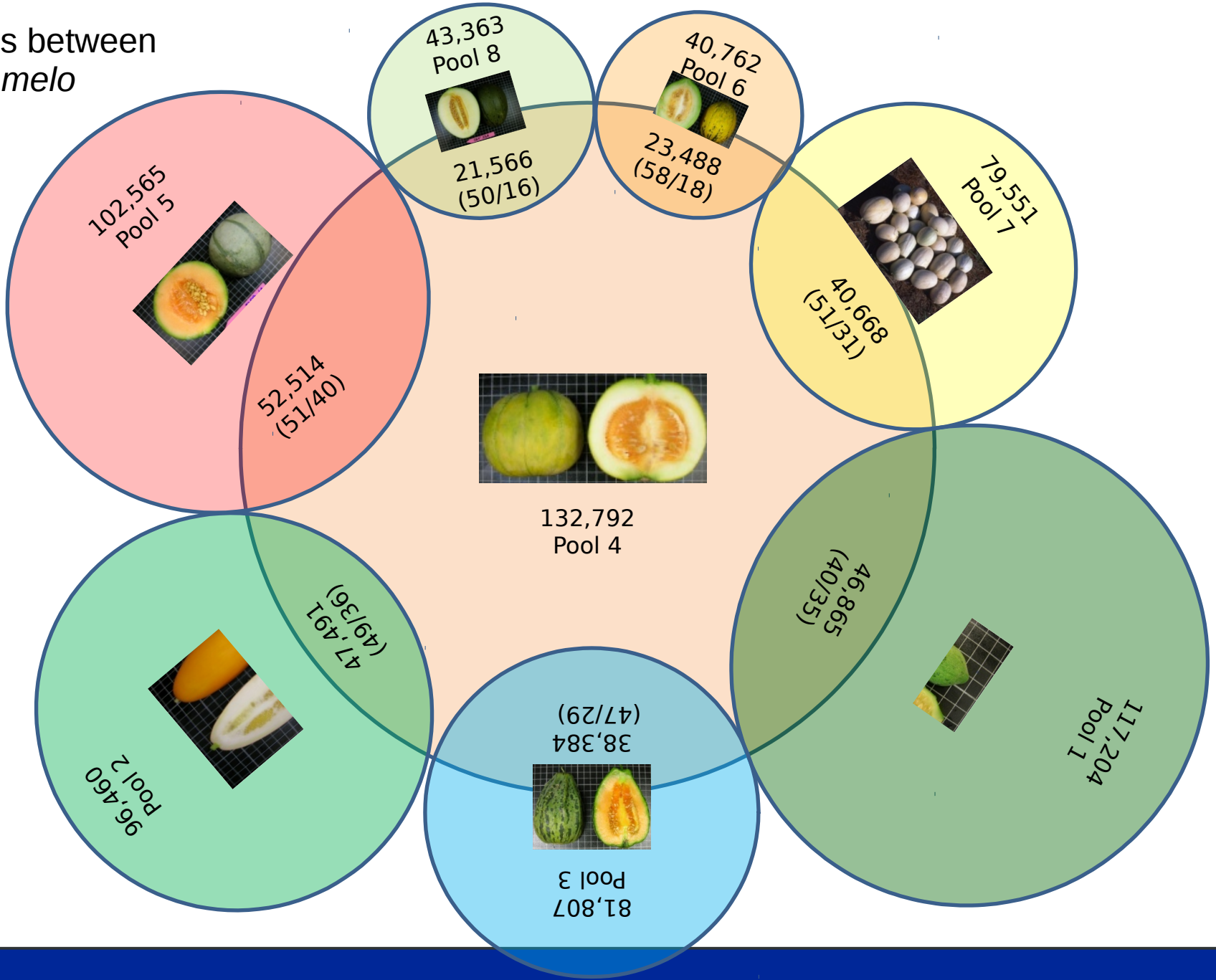
↓

28,996 easy to use SNPs
(92% validation rate)

Pool	Total N° SNPs	SNPs/kb	N° SNPs with MAF < 0.7 (%)
C. melo subespecies *agrestis*			
1) African *agrestis*	117,204	8.9	9,133 (7.8)
2) Asian *agrestis-acidulus*	96,460	6.8	10,197 (10.6)
3) Far East *conomon*	81,807	6.2	1,305 (1.6)
Intermediate types			
4) Middle East and Indian *momordica-dudaim-flexuosus*	132,792	8.4	13,826 (10.1)
C. melo subespecies *melo*			
5) Group *cantalupensis*	102,565	7.3	6,317 (6.2)
6) Group *melo* Europe-Asia *inodorus-chandalak-ameri*	40,762	9.2	2,417 (5.9)
7) *inodorus* Spanish landraces	79,551	6.4	3,210 (4.0)
8) *inodorus* group market class Piel de Sapo	43,363	4.9	1,396 (3.2)

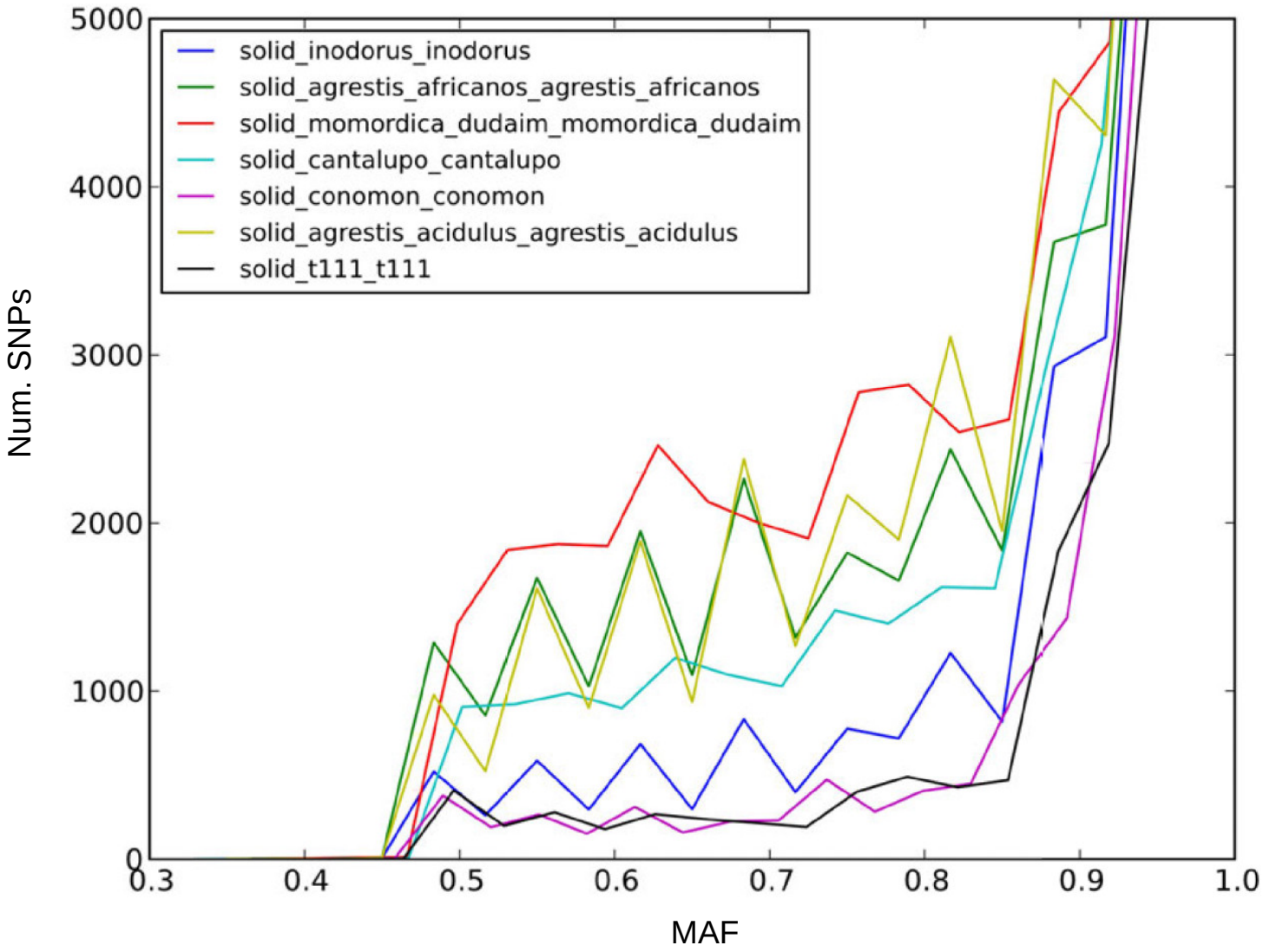
Marker discovery in *C. melo* pools

None shared SNPs between *agrestis* and *melo*



Marker discovery in *C. melo* pools

Sequencing pools we get allele frequencies not genotypes

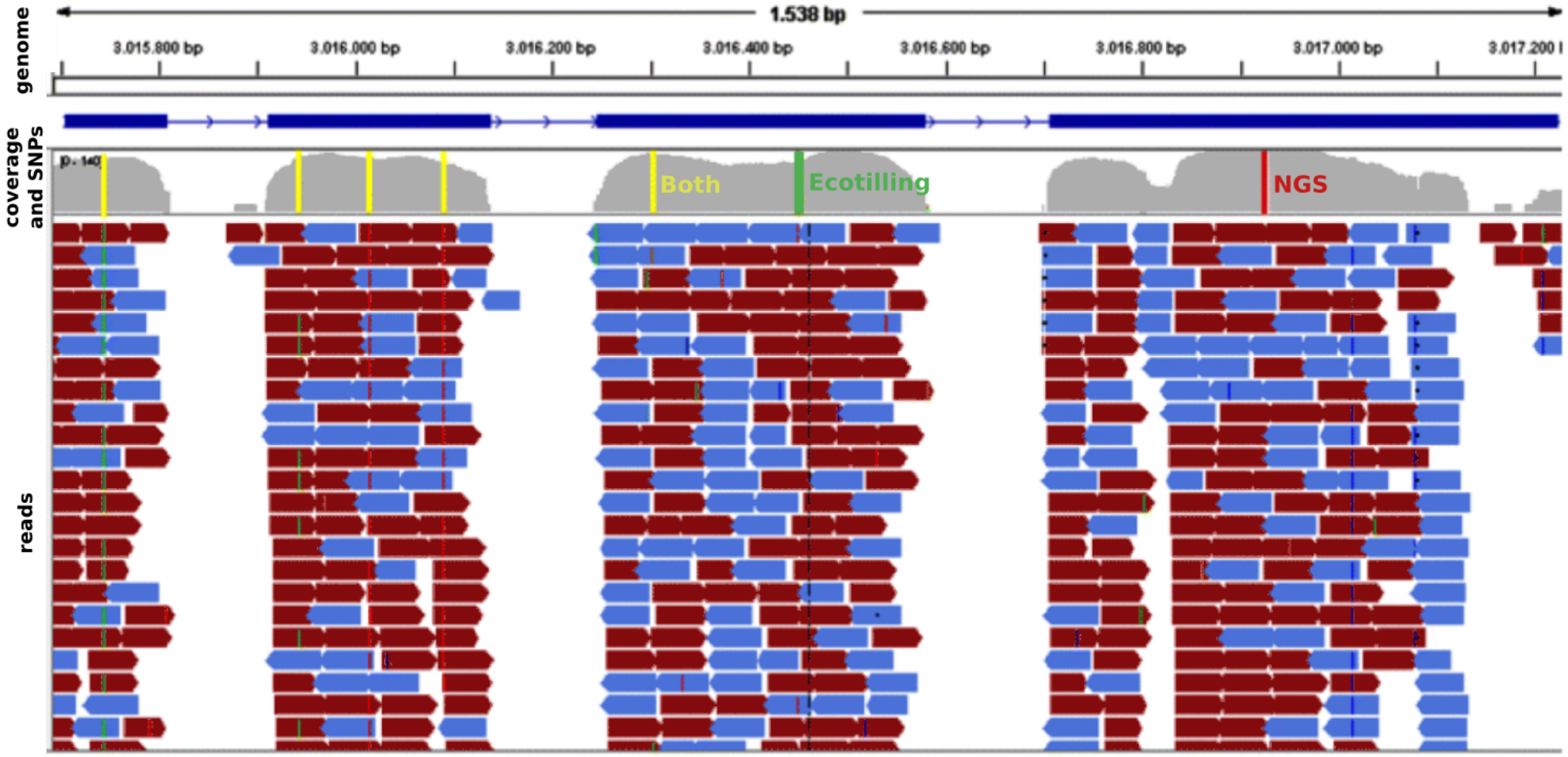


NGS Ecotilling in melon ACO-1

150M 44pb SOLiD processed reads

RNASeq

212 melon accessions in 8 pools



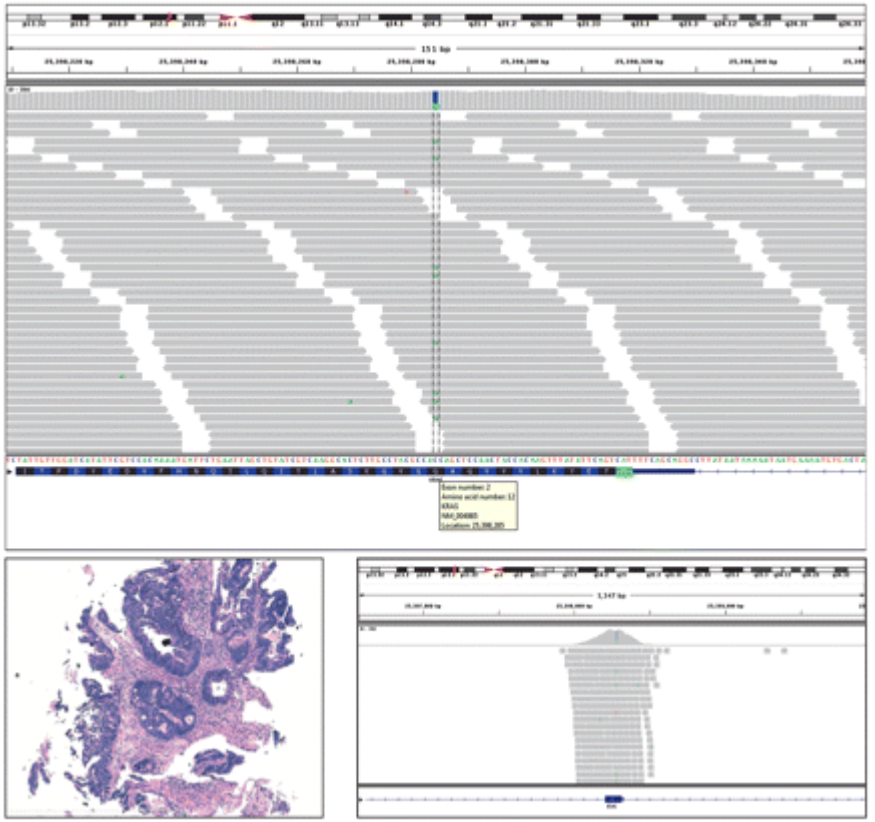
Cancer mutation identification

KRAS mutation in colorectal cancer

CRC cases with a G12V KRAS

Hiseq2000 of 182 genes

formalin-fixed, paraffin-embedded sample



Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol. 2011 Oct;136(4):527-39

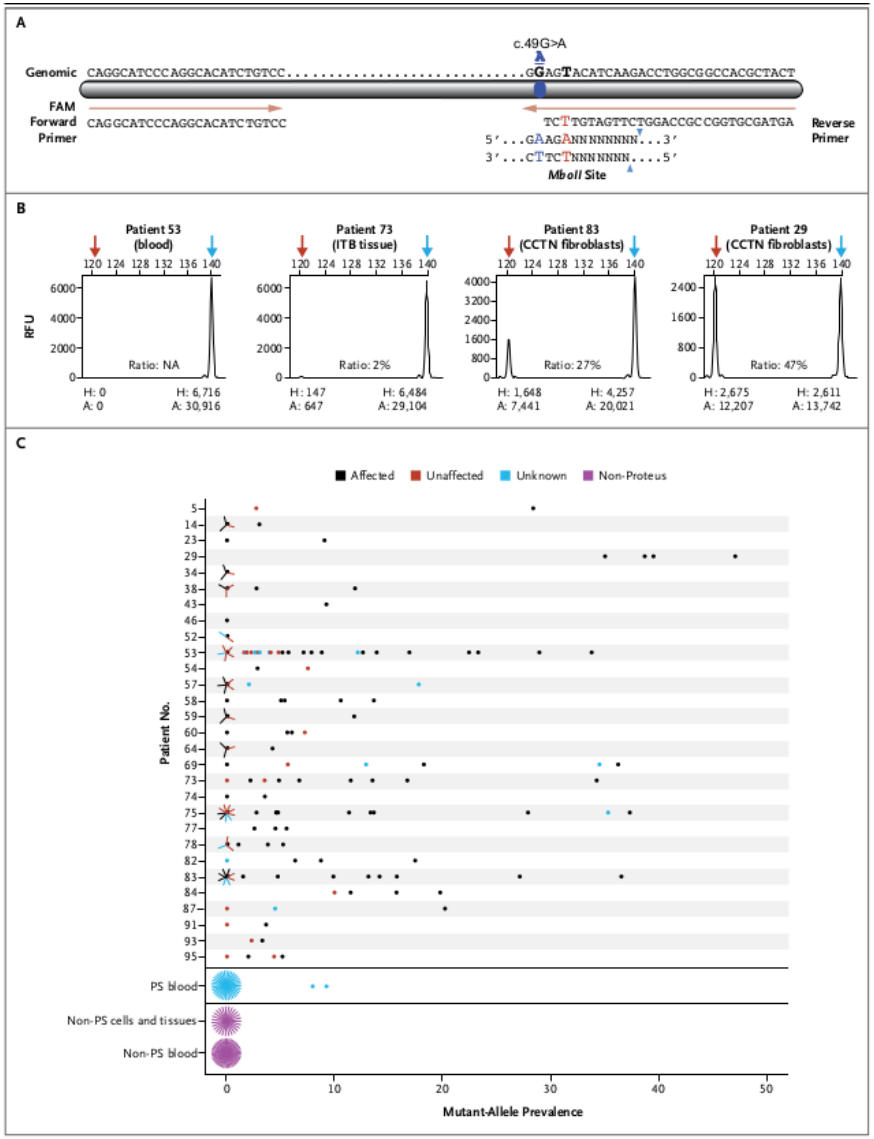
Somatic mutation identification

Proteus Syndrome

Sequencing of exome of normal and affected tissues

Agilent All Exon kit

26/29 analyzed patients had a somatic activating mutation (c.49G→A, p.Glu17Lys) in the oncogene AKT1



Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, Turner J, Cannons JL, Bick D, Blakemore L, Blumhorst C, Brockmann K, Calder P, Cherman N Deardorff MA, Everman DB, Golas G, Greenstein RM, Kato BM, Keppler-Noreuil KM,Kuznetsov SA, Miyamoto RT, Newman K, Ng D, O'Brien K, Rothenberg S, Schwartzentruber DJ, Singhal V, Tirabosco R, Upton J, Wientroub S, Zackai EH, Hoag K, Whitewood-Neal T, Robey PG, Schwartzberg PL, Darling TN, Tosi LL, Mullikin JC, Biesecker LG. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med. 2011 Aug 18;365(7):611-9.

Fetal Plasma mutation identification

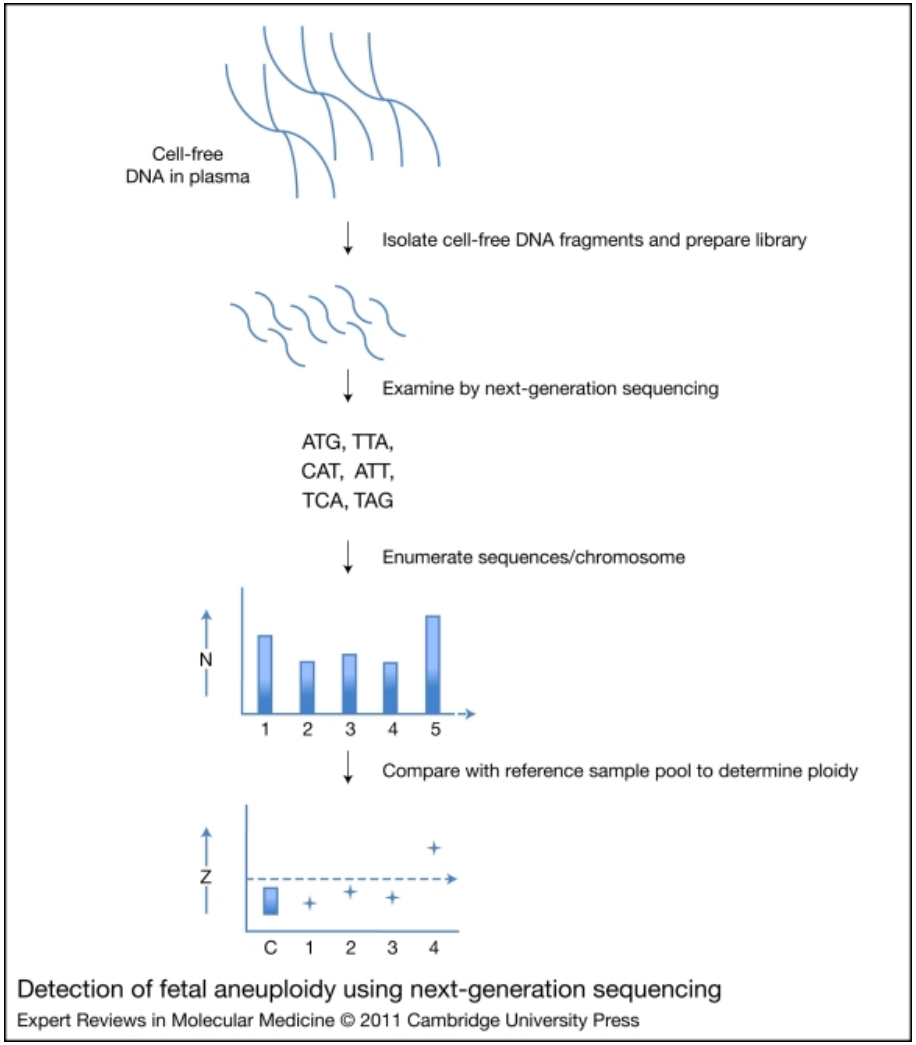
753 pregnant at high risk for fetal trisomy 21

Plasma maternal DNA sequencing

Genome Analyzer IIx

Percentage at 8-plex: 79.1% of the trisomy

Percentage at 2-plex: 100% of the trisomy



Detection of fetal aneuploidy using next-generation sequencing

Hahn S, Lapaire O, Tercanli S, Kolla V, Hösli I. Determination of fetal chromosome aberrations from fetal DNA in maternal blood: has the challenge finally been met? Expert Rev Mol Med. 2011 May 4;13:e16.

Population studies. 1000 genomes

Phase 1:

1092 genomes. 14 populations

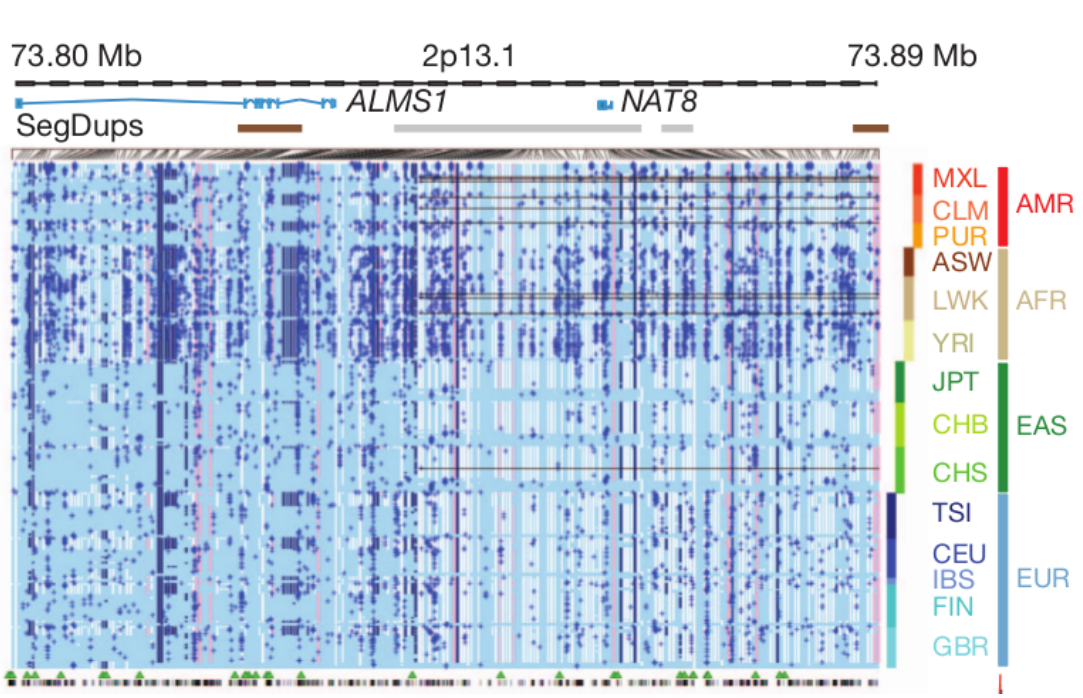
Low-coverage Whole-genome (4-5X)
Exome (80X)
Trios (to detect haplotypes)

36.7 Millions SNPs

Phases 2-3:

2500 samples. 11 populations more

Deeper sequencing
Improved analysis



http://www.1000genomes.org/

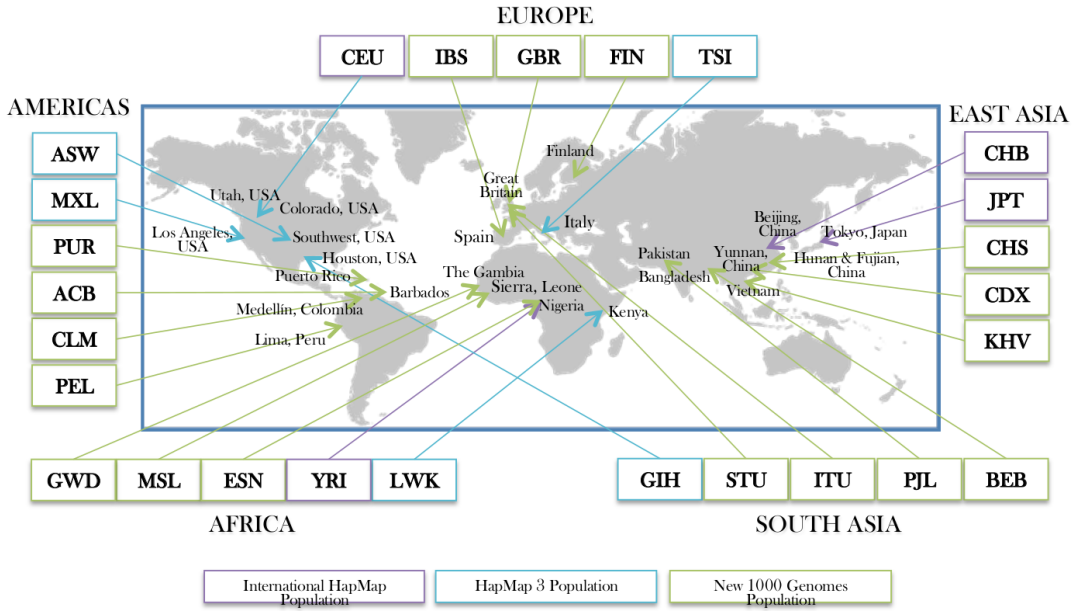
Population studies. 1000 genomes

Benefits:

Public NGS datasets

Catalog of variants

Formats and tools for NGS analysis



http://www.1000genomes.org/

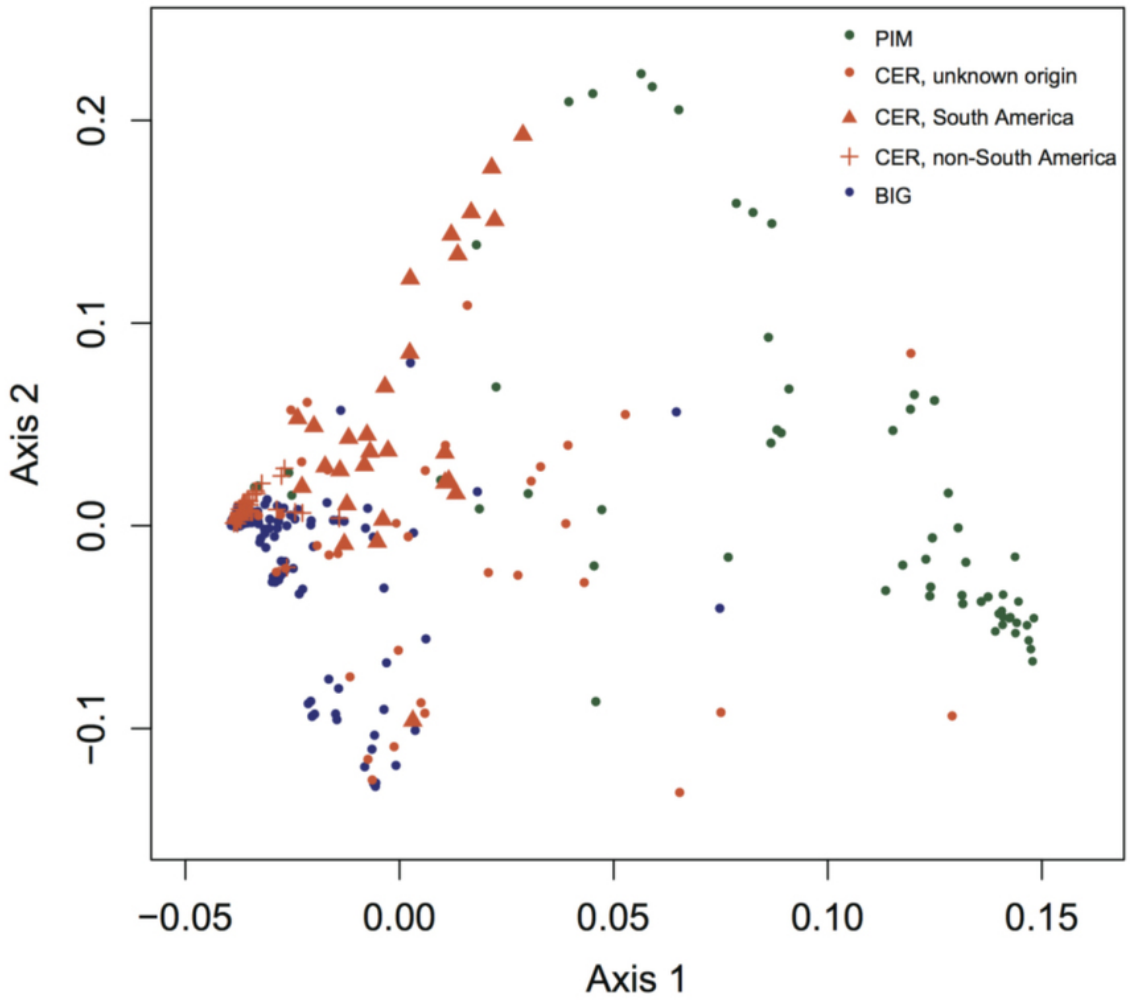
360 tomato genomes

360 accessions

Whole genome ~5X

11,620,517 SNPs

1,303,213 small indels



Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S. Genomic analyses provide insights into the history of tomato breeding. Nat Genet. 2014 Nov;46(11):1220-6. doi: 10.1038/ng.3117. Epub 2014 Oct 12. PubMed PMID: 25305757.

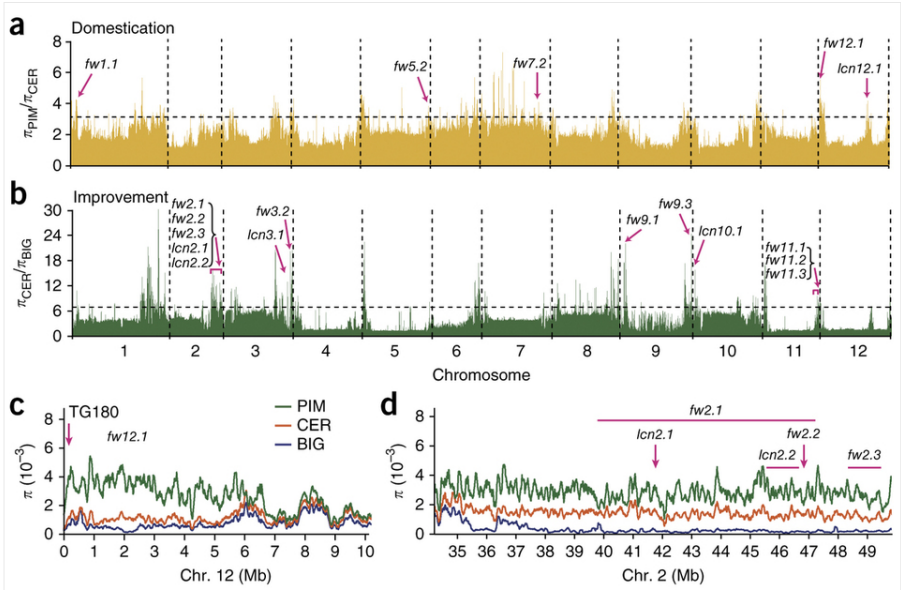
360 tomato genomes

Domestication signals

Association mapping

Population structure

Phylogenomics



Genetic mapping: SSRs vs NGS

Thousands of polymorphic markers between parents are needed

Problem:

SSRs:

- High PIC :)
- non automatable :(

SNPs:

- Low PIC :(
- automatable :)

Strategy 1:

- Look for SNPs between parents by sequencing (e. g. transcriptome).
- Genotyping with Illumina array.

Strategy 2:

- GBS for all samples

QTL mapping in tomato

Parthenocarpy in tomato

RP75/59 mutant:

- Robust fruit set rate
- Phenotyped F2: RP75/59 x UC82

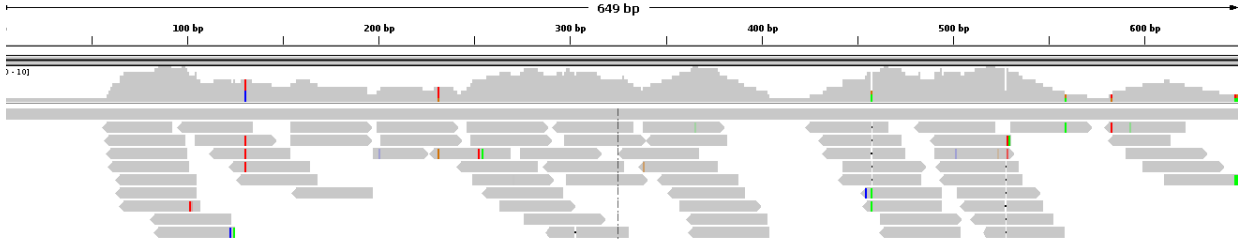
No reference genome (2009), only transcriptome

Cost:

- Sequencing: 3000$
- Genotyping: 12,000$

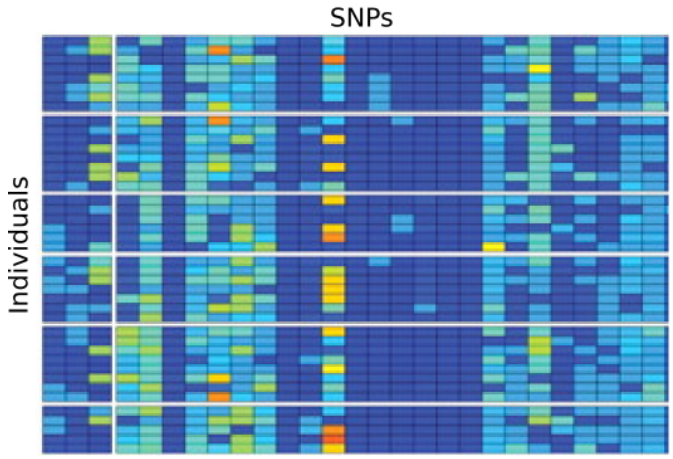
QTL mapping in tomato

Low coverage Illumina pooled (RP75/59 and UC82) RNASeq



384 high quality SNPs

F2 Veracode Genotyping



QTL mapping in tomato

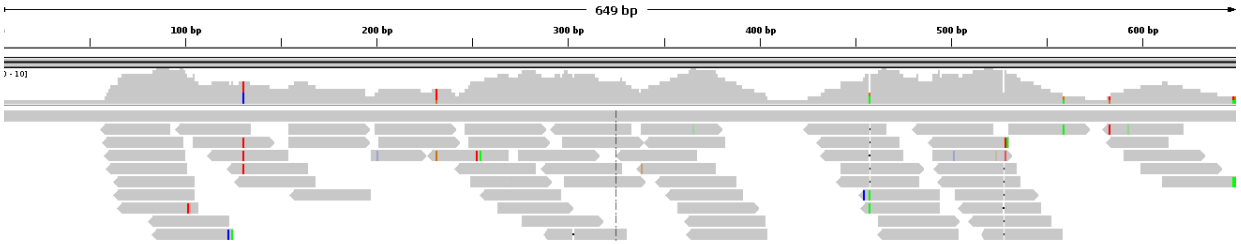
SNP discovery:

- Sequencing:
  - Normalized cDNA library
  - 14.2 M 75bp Illumina reads
  - 276,039 Sanger ESTs
- Mapping against the transcriptome:
  - Illumina: 8.5 coverage
  - Sanger: 4.2 coverage
- SNP calling:
  - 33.306 SNPs
  - 6.934 easy to use (85% validation rate by HRM)
  - Subset of 291 easy to use highly polymorphic set. 0.28 PIC in 37 tomato varieties.

F2 genotyping:

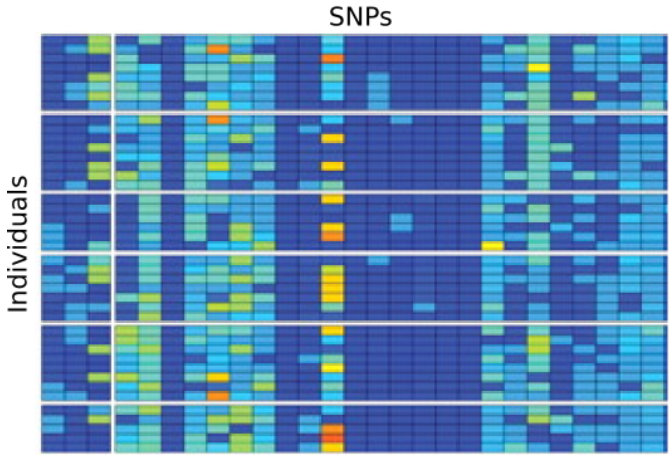
- 384 SNPs (Veracode-Illumina)
- 85% success rate

Low coverage Illumina RNASeq



384 high quality SNPs

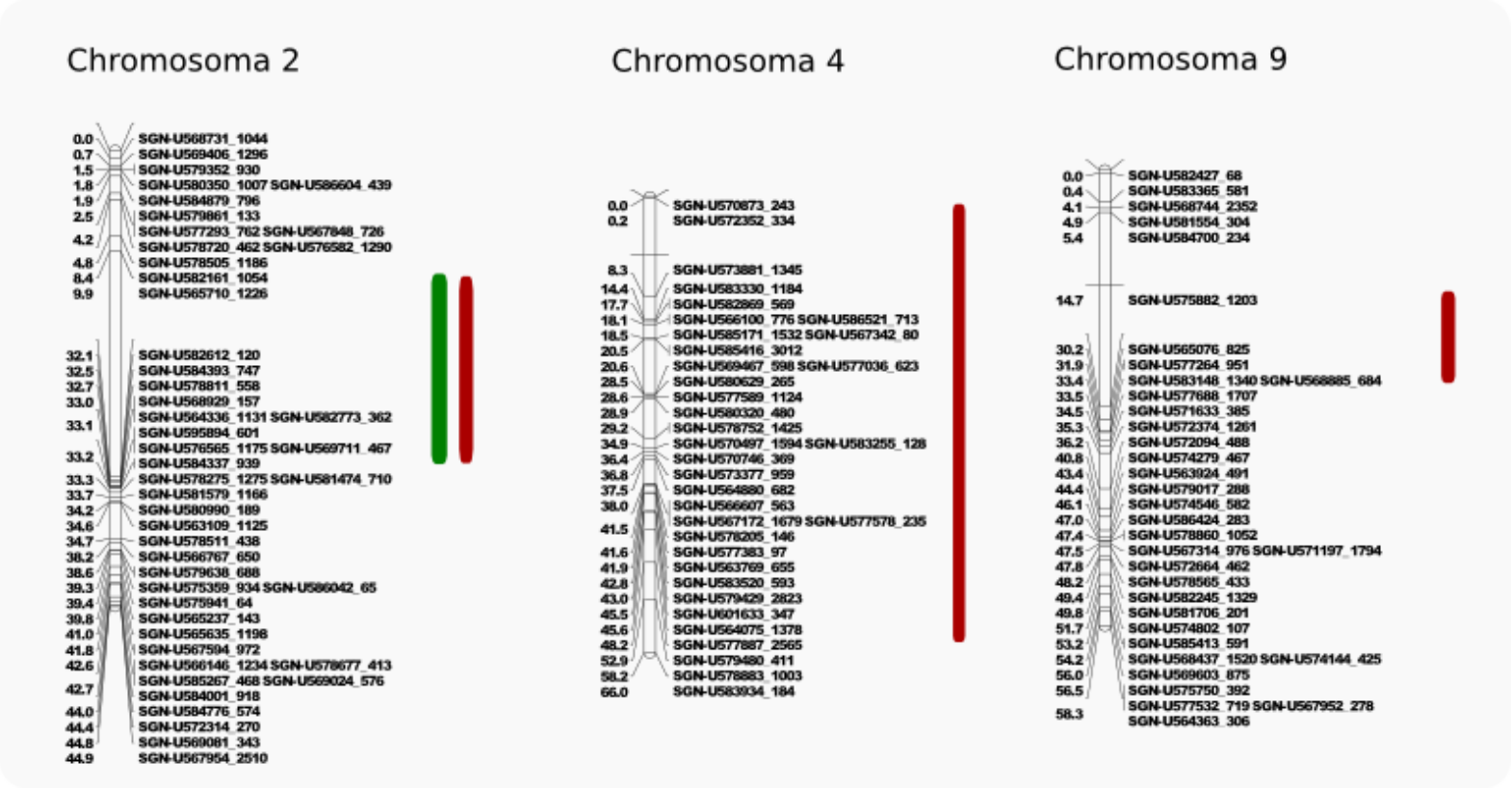
Genotyping: 85% success rate



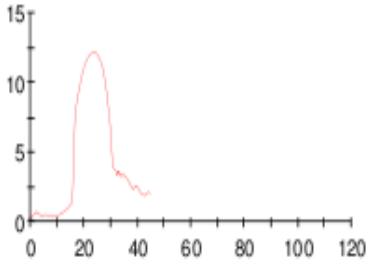
QTL mapping in tomato

fruit set QTL

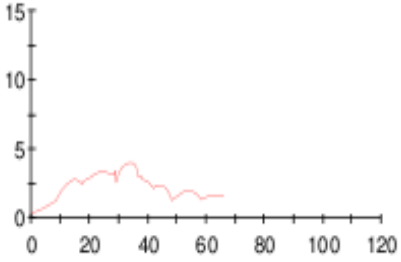
fruit size QTLs



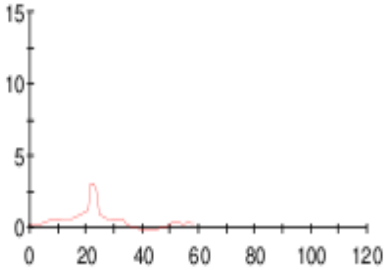
Lod: 12.1



Lod: 3.1



Lod: 4.0



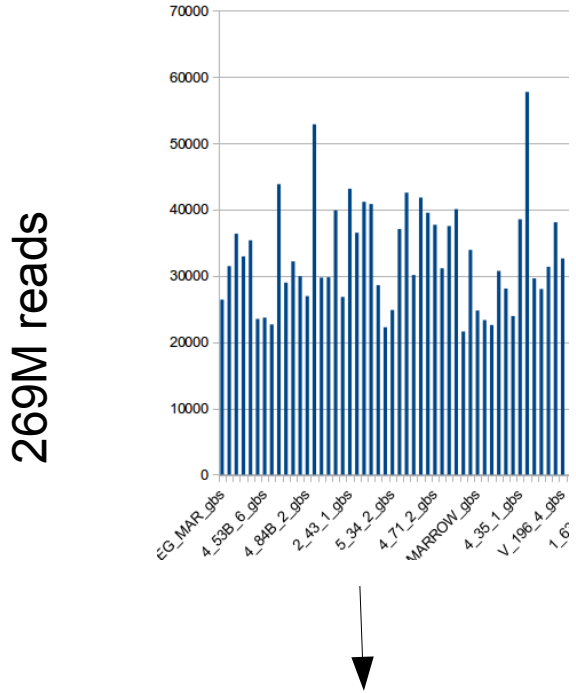
GBS-based genetic map in *C. pepo*

C. pepo population

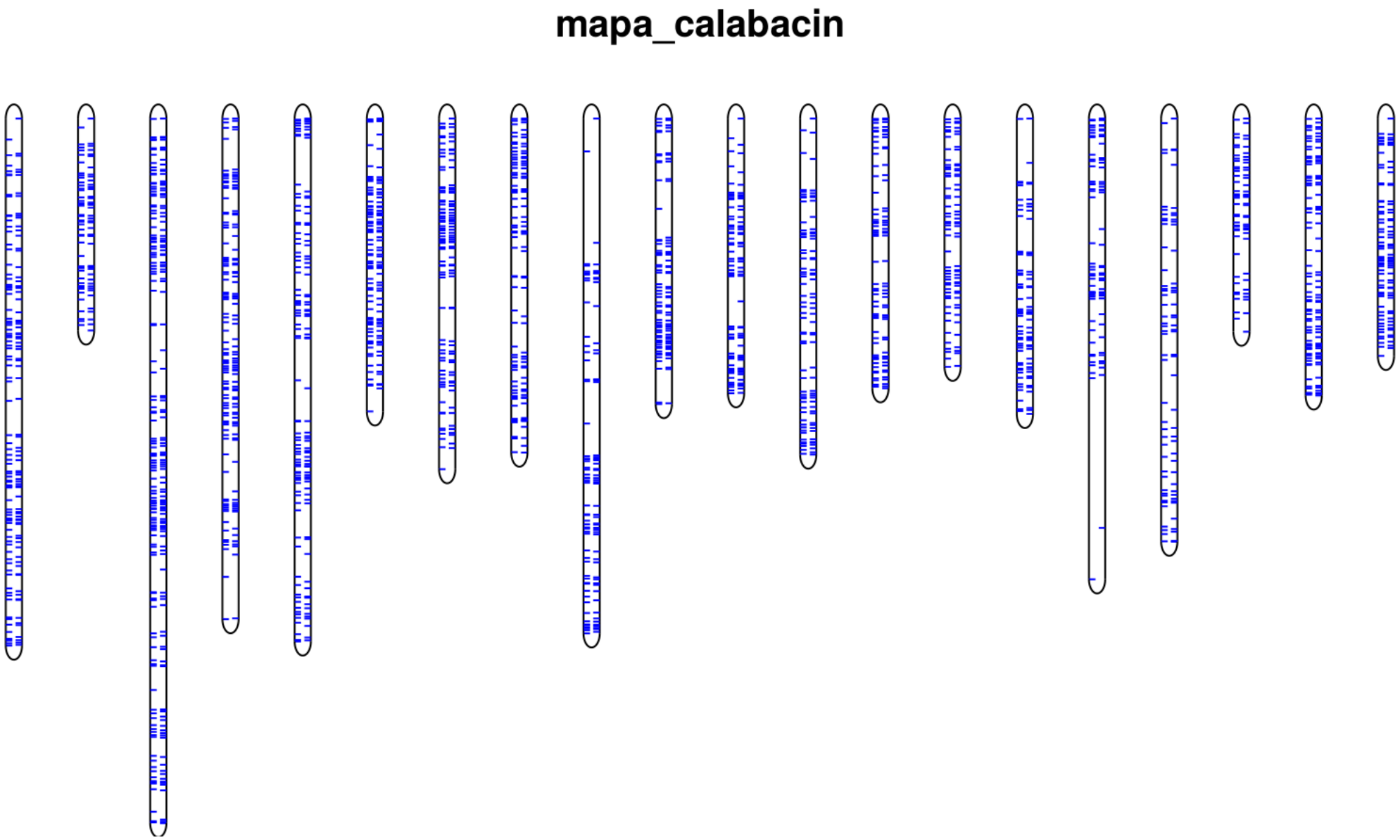
Reference genome
(assembled by us)

GBS for all 143 individuals → 38,424 SNPs

Sequencing Cost: 143 * $50 = $7150



GBS-based genetic map in *C. pepo*



Genetic populations studies.

2.1. GBS-EcoT22

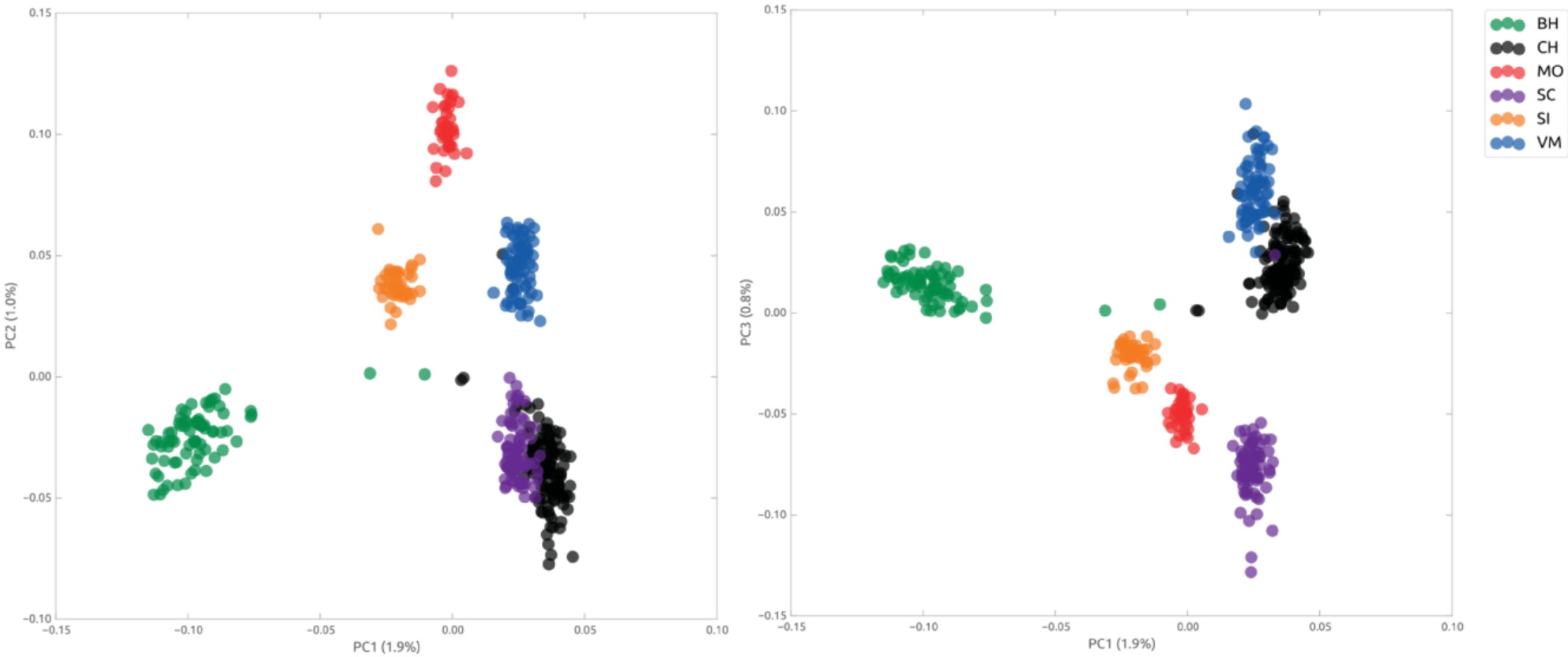
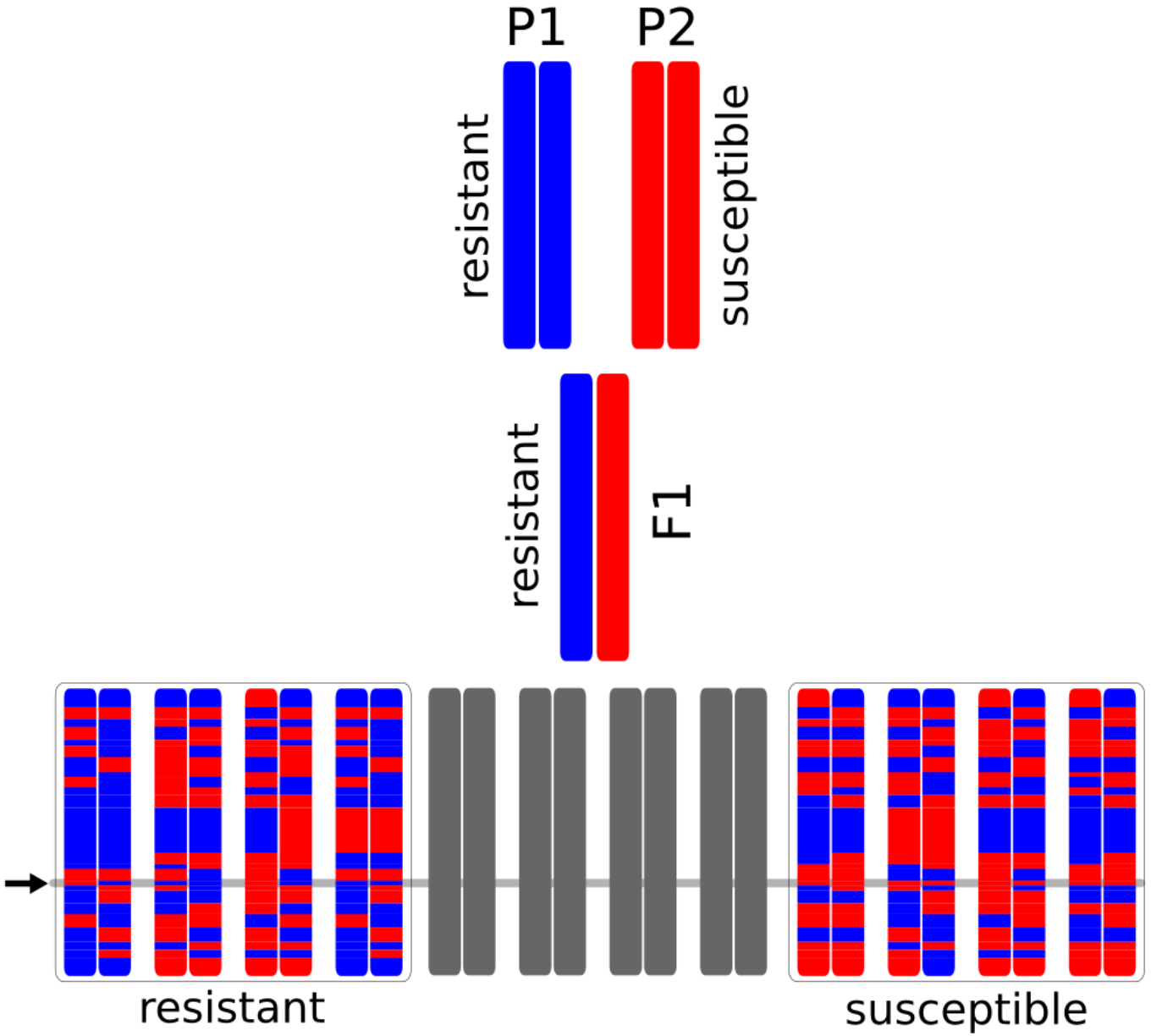


Figura 1: PCA con los 21879 SNPs obtenidos en el GBS-EcoT22

BSA


P1
P2
resistant
susceptible
resistant
F1
resistant
susceptible

BSA: *Spodoptera* resistance to Bt toxin

Xen-R (resistant) x FRA (susceptible)

F1 x Xen-R

BC (resistant individuals)

Illumina sequencing

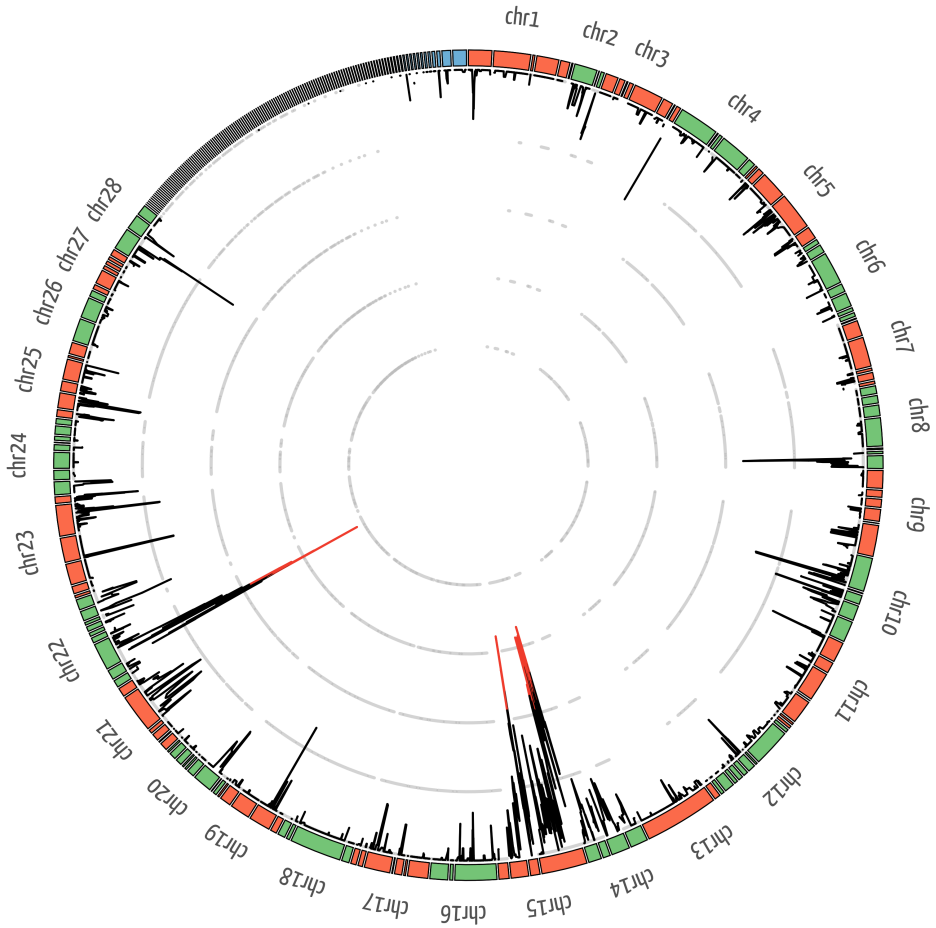
BSA: *Spodoptera* resistance to Bt toxin

Sequencing:

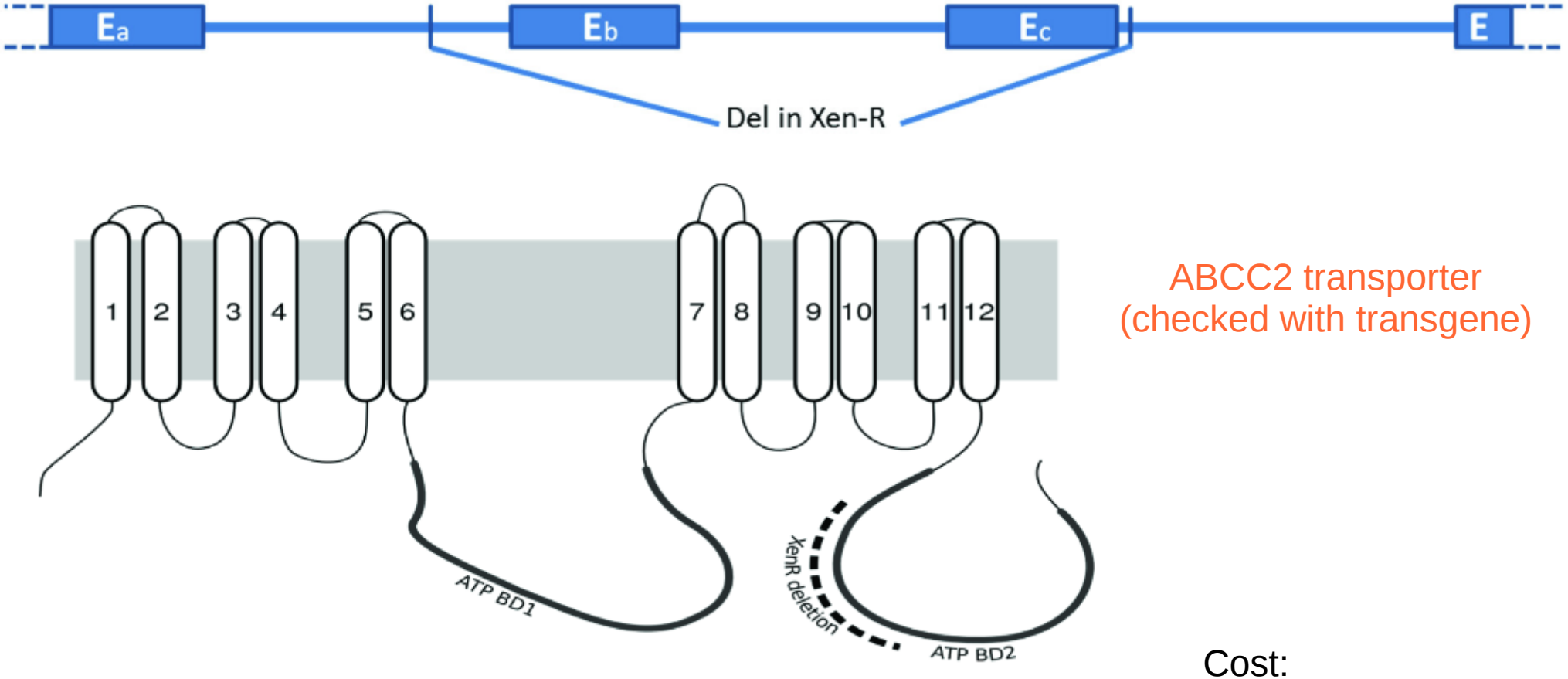
- 1 Illumina lane, half for each parent (355M reads)
- 1 Illumina lane for the resistant individuals in BC1 (360M reads)

Analysis:

- Transcriptome assembly: 96,676 unigenes
- Synteny with *Bombyx mori*
  - Same infraorden different superfamily
- *B. mori* orthologs: 15,116 unigenes
- 437,815 SNPs and 80,246 indels
- Resistance skewness allele ratio along the genome



BSA: *Spodoptera* resistance to Bt toxin



ABCC2 transporter
(checked with transgene)

Cost:

- Sequencing: $6000
- Analysis: 2 months

BSA whitout NGS?

Analyzed SNPs: 268,000

Samples: 3

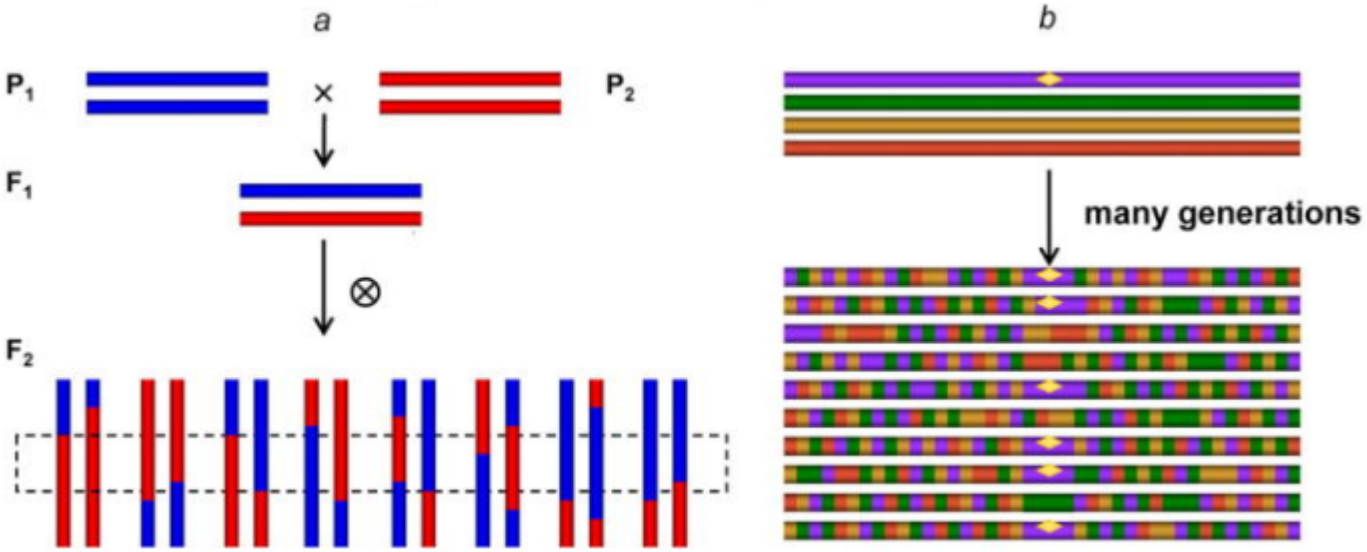
Alternatives:

SNP platform: small number of markers, expensive for only 3 samples

AFLPs: Hundreds of combinations, no genetic map, not synteny

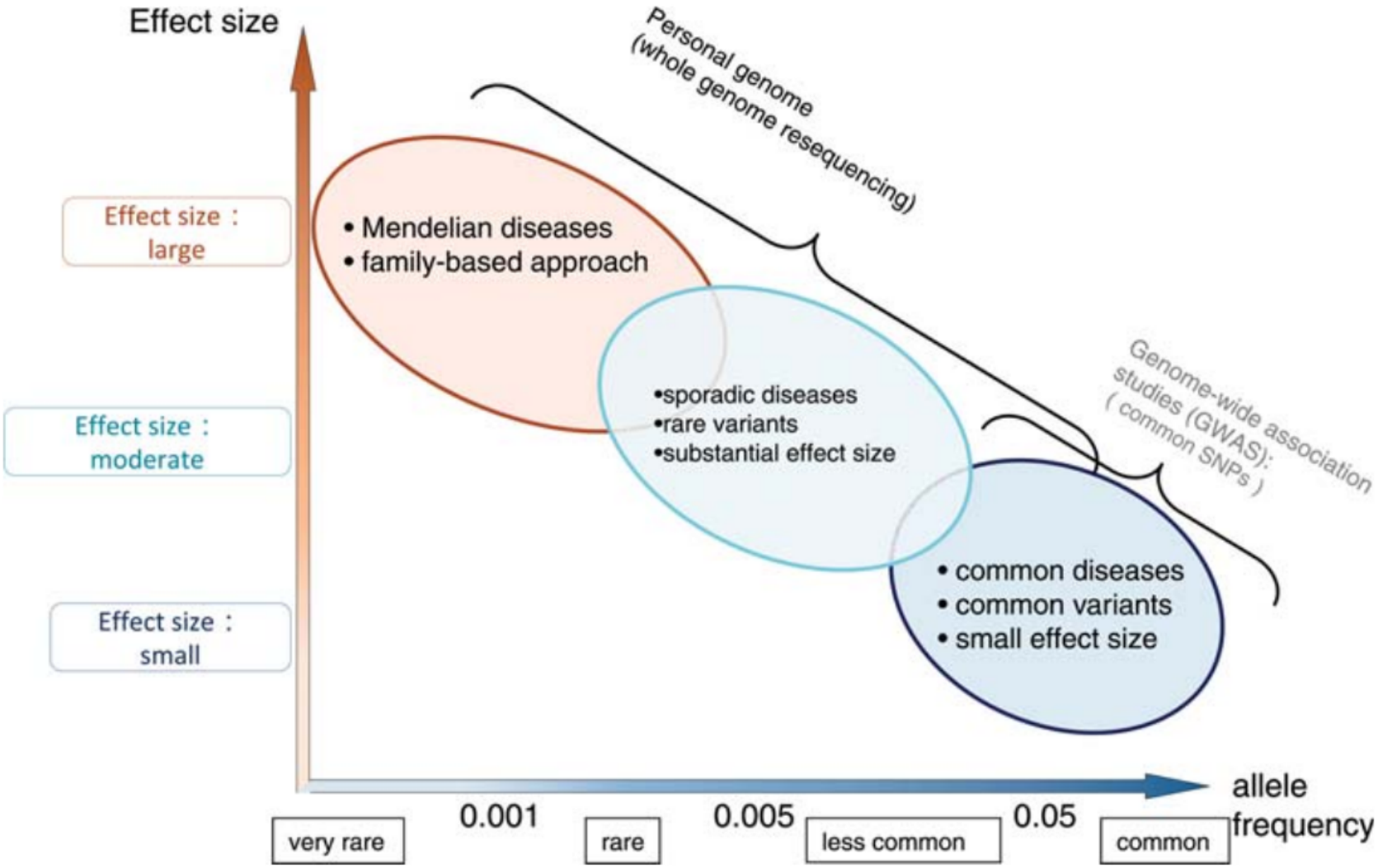
Not viable

Genome wide association (GWA)



Chengsong Zhu, Michael Gore, Edward S. Buckler and Jianming Yu. Status and Prospects of Association Mapping in Plants. Plant Genome 2008. Vol. 1 No. 1, p. 5-20

GWAs vs GRs



Tsuji S. Genetics of neurodegenerative diseases: insights from high-throughput resequencing. Hum Mol Genet. 2010 Apr 15;19(R1):R65-70.

GWAs non-model organism

GWAs in *Heterobasidion annosum*

Pathogenic fungi

23 haploid isolates sequenced (4-10X)

Assembly of draft fungi genome :

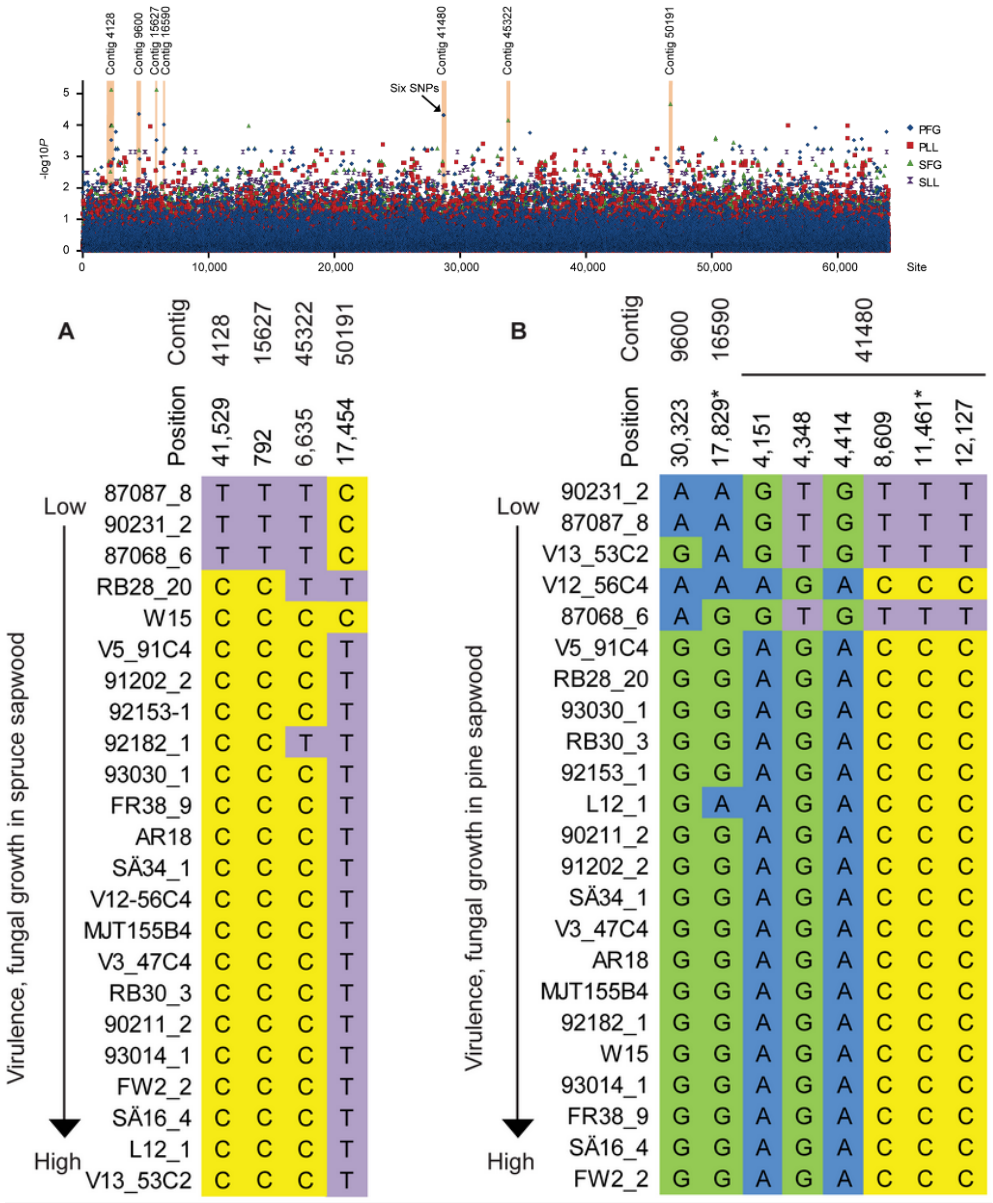
- Size: 30,5 Mb
- N50: 40Kb
- Coverage: 38,7X

Tested in Pine and Spruce

12 SNPs detected in 7 contigs

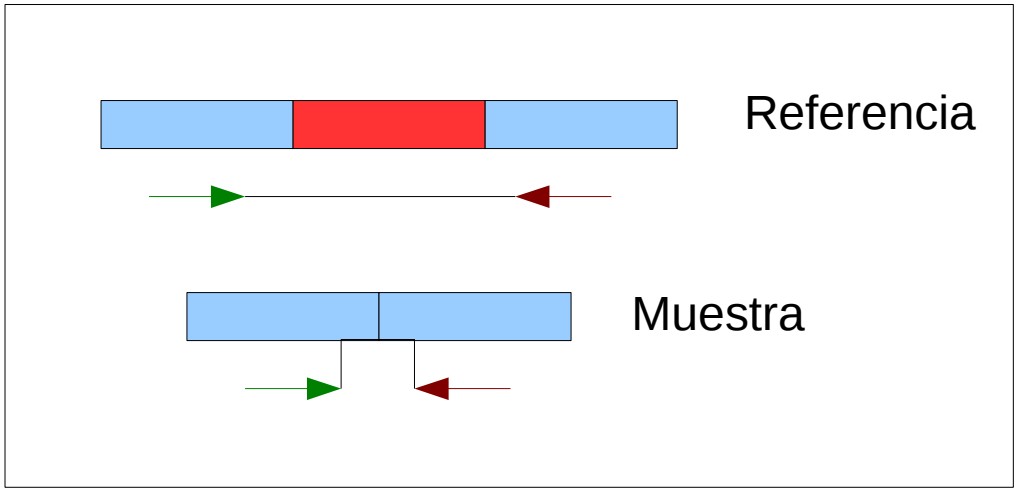
Candidate genes and concordance

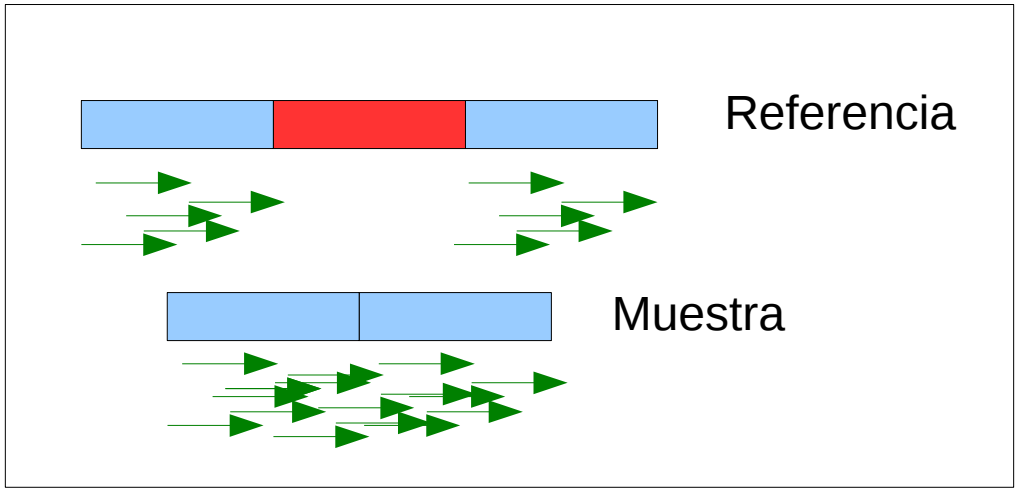
with known genes and QTLs

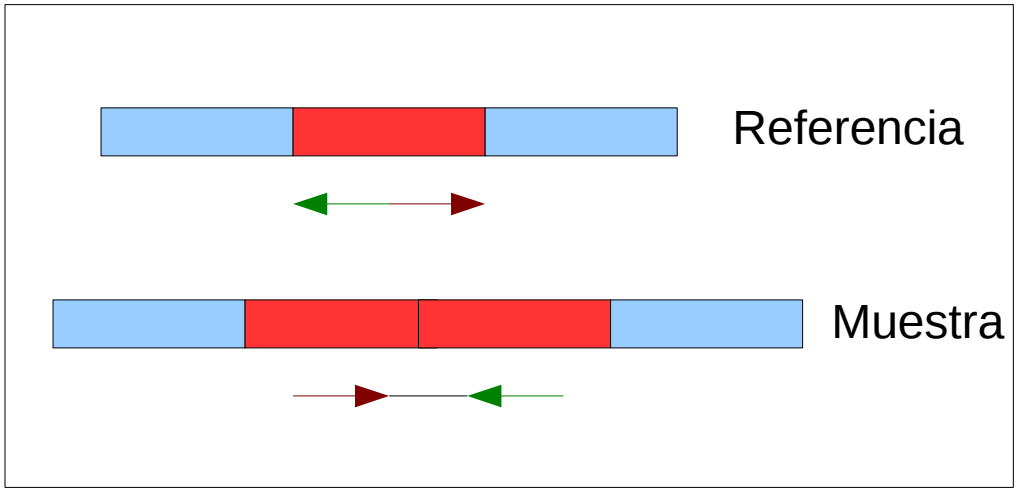


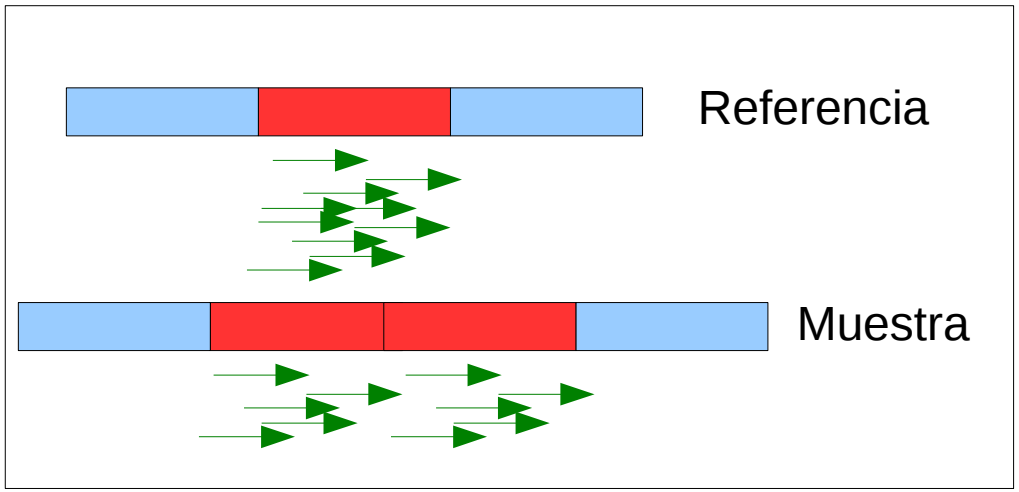
Dalman K, Himmelstrand K, Olson Å, Lind M, Brandström-Durling M, et al. (2013) A Genome-Wide Association Study Identifies Genomic Regions for Virulence in the Non-Model Organism Heterobasidion annosum s.s. PLoS ONE 8(1): e53525. doi:10.1371/journal.pone.0053525

Copy number variations. CNVs









Exome. CNVs

Exome capture and 454

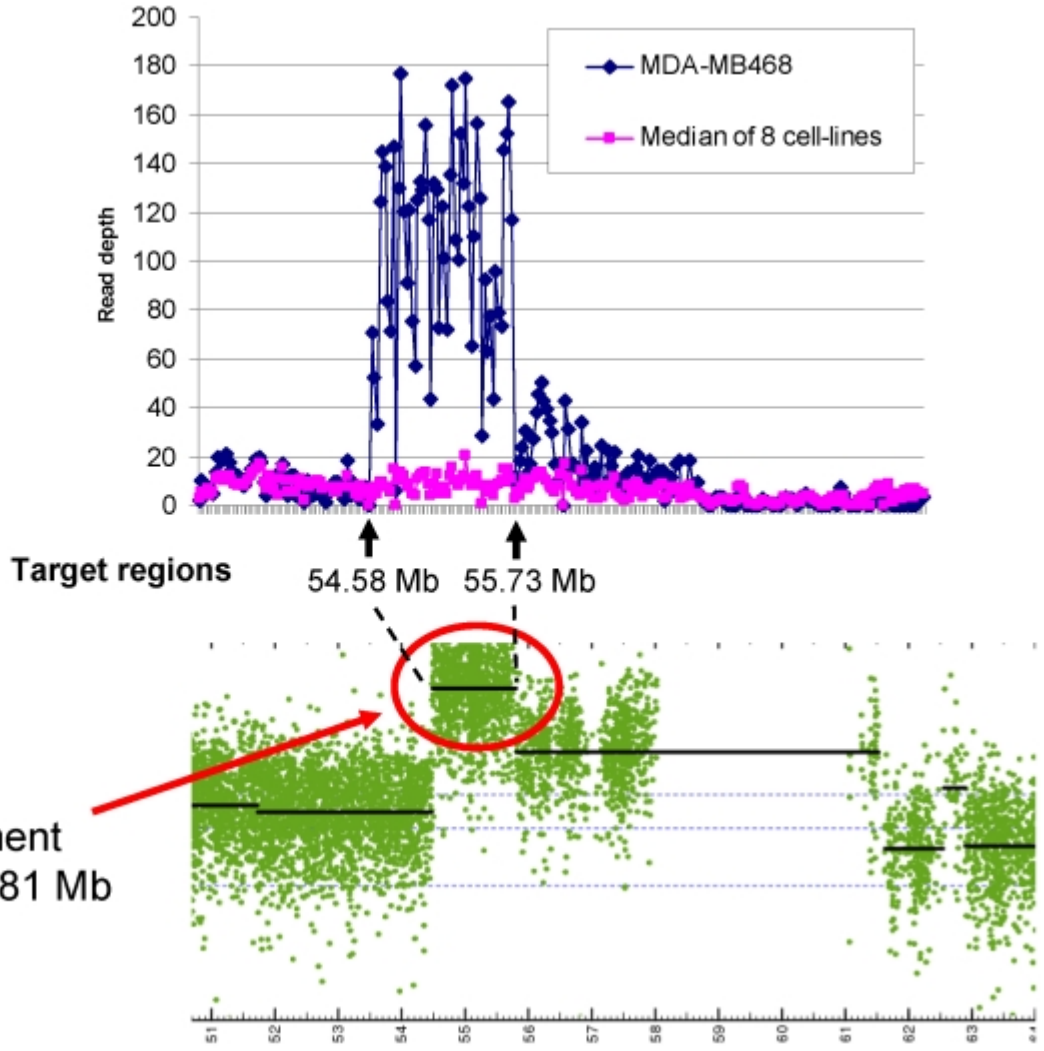
8 cell cancer lines

1,9 millions reads/ sample

Coverage 7,3X

95% concordance with Affymetrix SNP Array 6.0

2,779 potential novel sequence variations/mutations per cell line.



Chang H, Jackson DG, Kayne PS, Ross-Macdonald PB, Ryseck RP, Siemers NO. Exome
sequencing reveals comprehensive genomic alterations across eight cancer cell lines. PLoS One. 2011;6(6).

RNA-seq

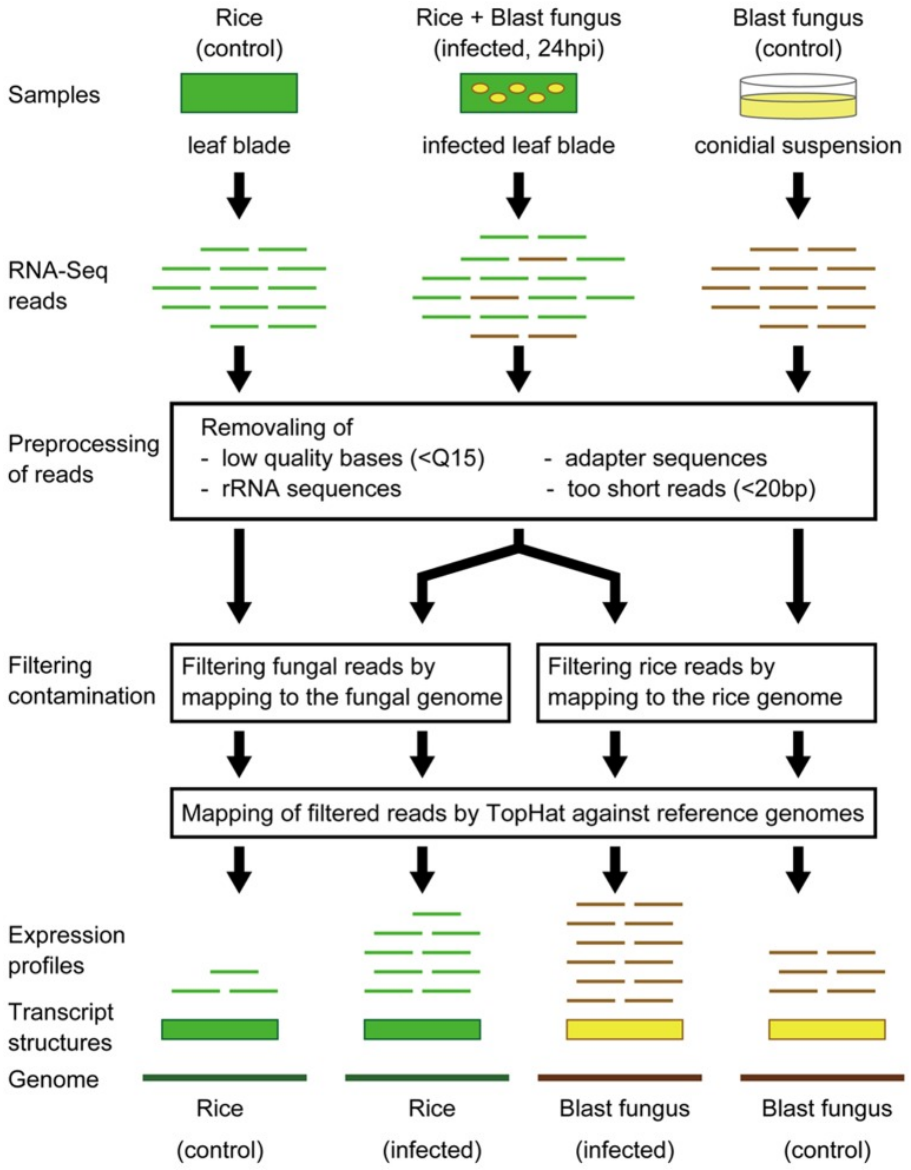
Rice and blast fungus

Expression in infected leaves at 24 dpi

Genome Analyzer IIx

Duplicated samples

1 line/ samples



Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E.
Simultaneous
RNA-Seq Analysis of a Mixed Transcriptome of Rice and Blast Fungus
Interaction.
PLoS One. 2012;7(11)

miRNA. P53

Identification miRNA en p53 tumors

H1299 lung cancer cells

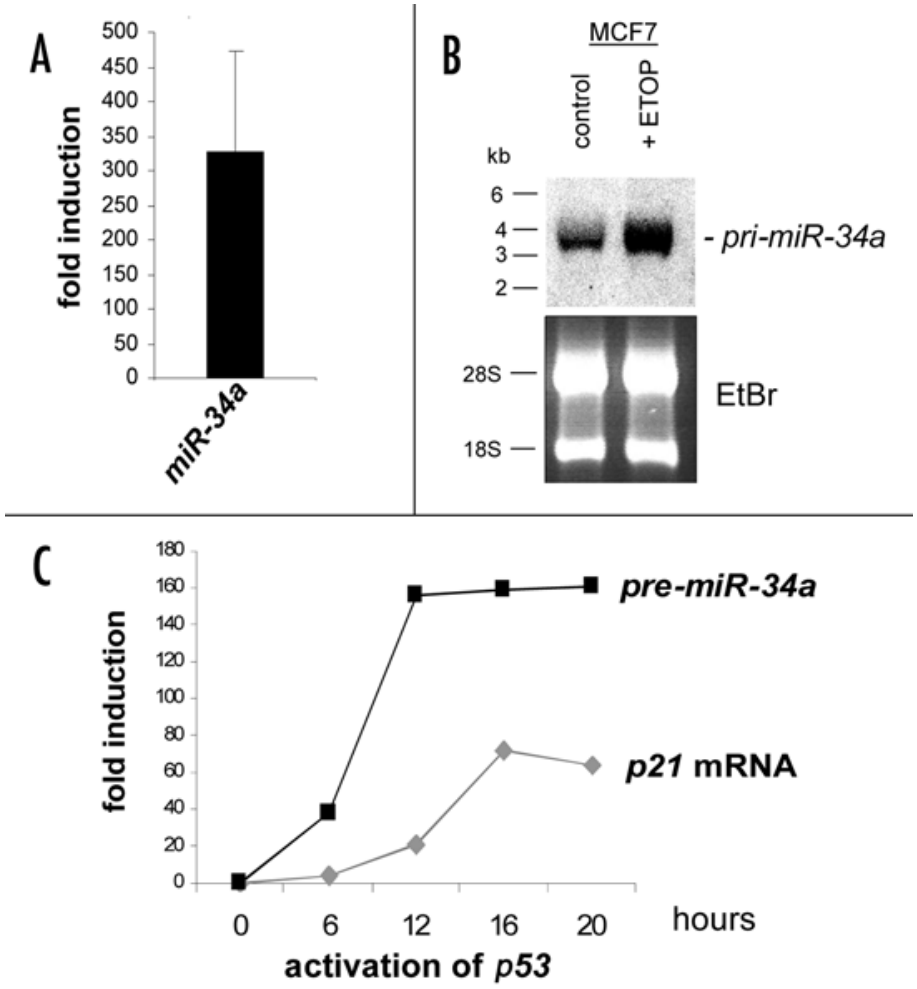
Small RNA library.

454. GS20

31 miRNA activated

16 suppressed.

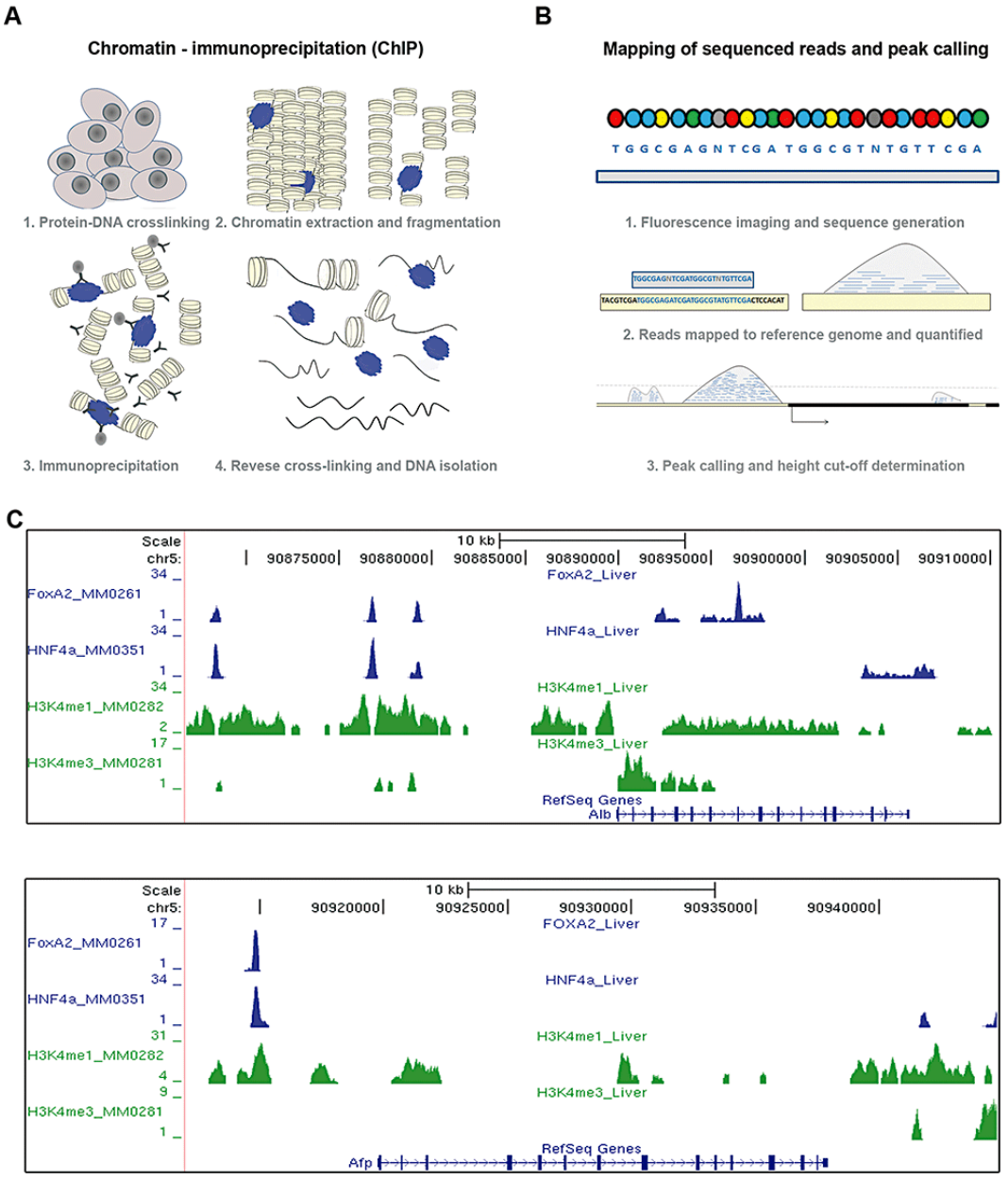
miR-34a induces apoptosis and cell cycle arrest



Tarasov V, Jung P, Verdoodt B, Lodygin D, Epanchintsev A, Menssen A, Meister G, Hermeking H. Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. Cell Cycle. 2007 Jul 1;6(13):1586-93

ChIP-seq

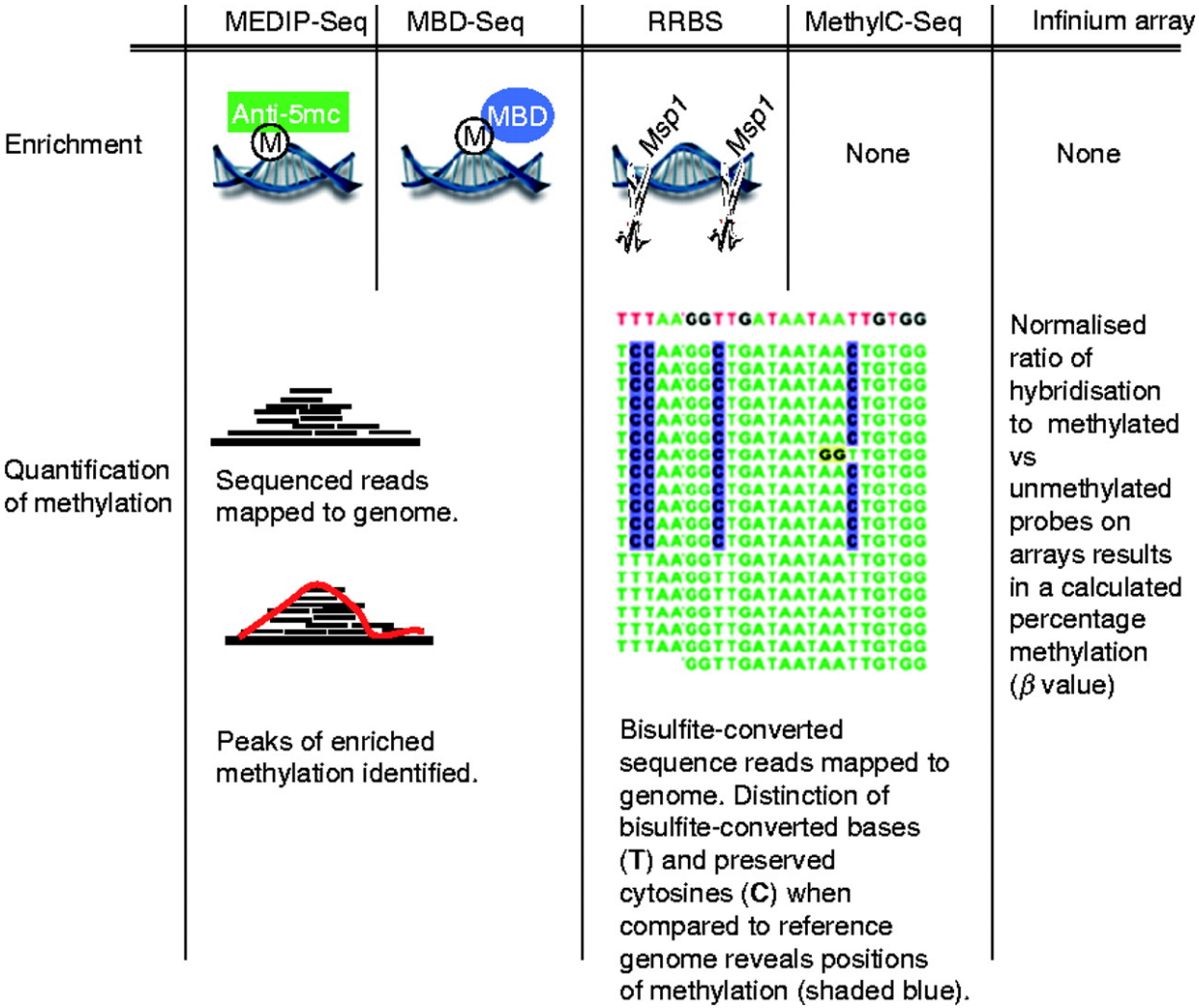
Detects promoters regulated by specific transcription factors, RNA pol binding and modified histone and DNA regions.



Cullum R, Alder O, Hoodless PA. The next generation: using new sequencing technologies to analyse gene regulation. Respirology. 2011 Feb;16(2):210-22

DNA methylation

Figure 4 Technologies for detection of genome-wide methylation.



Emes R D , Farrell W E J Mol Endocrinol 2012;49:R19-R27

DNA Methylation

DNA methylation patterns in select prostate tissues and cell lines.

MethylPlex–next generation sequencing (M-NGS) methodology.

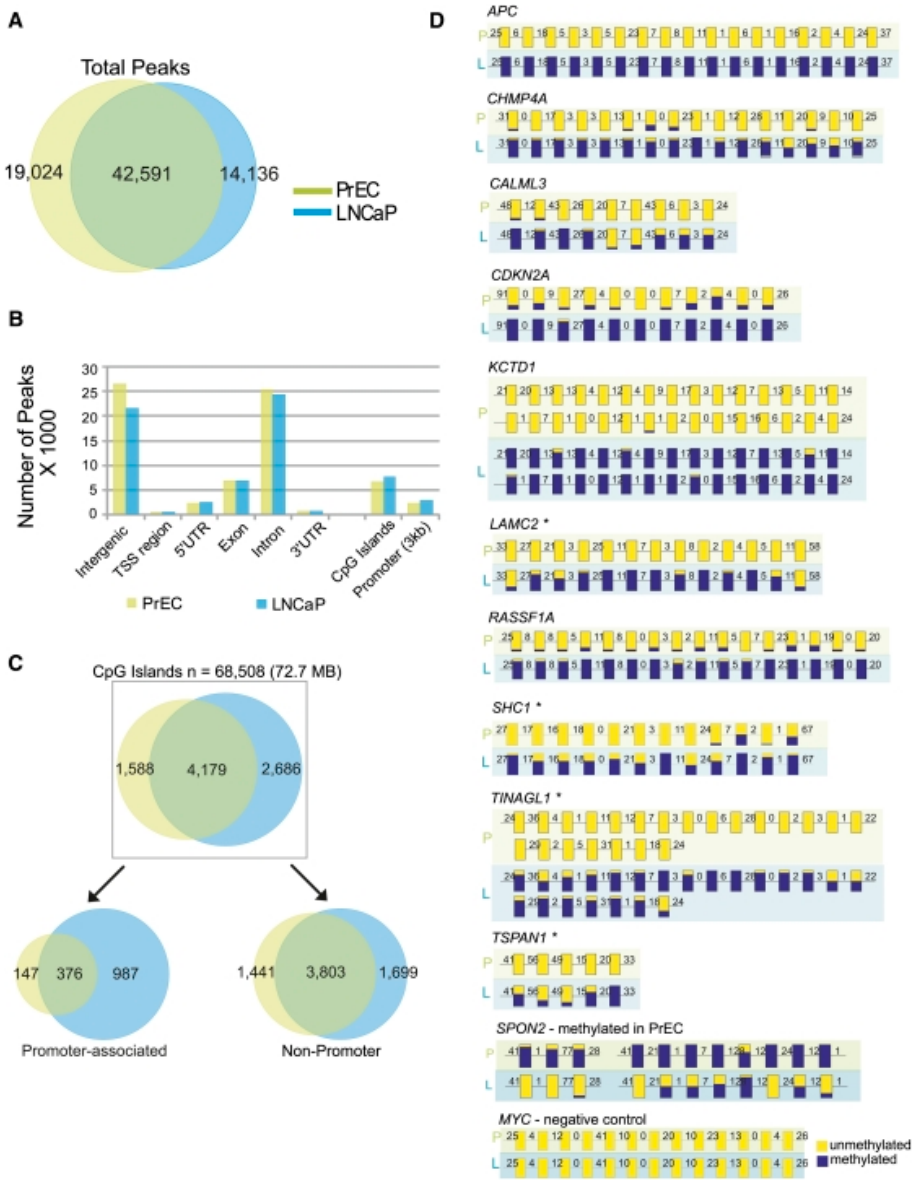
68,000 methylated regions per sample.

Differences between benign and tumor cells:

* Increased in promotor regions

* Differentially in transcription start sites

* Differentially in repeat element methylation



Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S Huang C, Shankar S, Jing X, Iyer M, Hu M, Sam L, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky HD, Siddiqui J, Yu J, Qin ZS, Chinnaiyan AM. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. 2011 Jul;21(7):1028-41

Pathogen identification

3 transplanted patients from the same donor

Negative identification with traditional techniques

RNA sequencing with 454 FLX 103,632 sequences

Identified sequences of a new virus related with Old World arenavirus

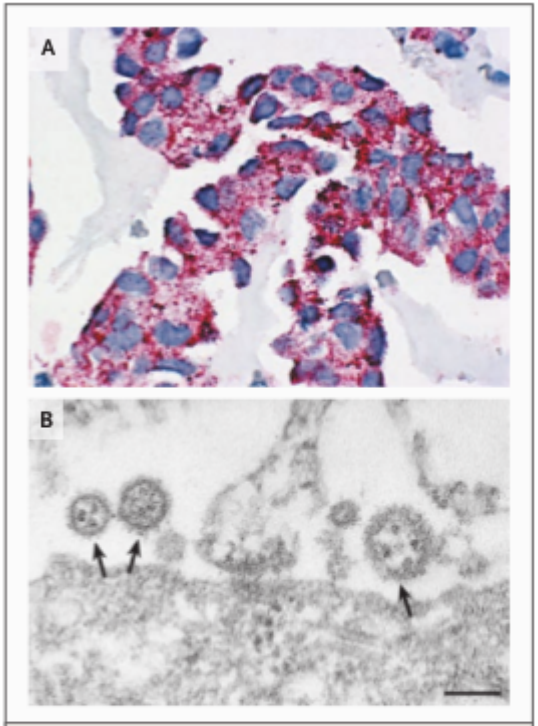


Figure 2. Propagation of the New Arenavirus in Tissue Culture.

Panel A shows immunostaining of viral antigens in infected cells by means of an indirect immunoalkaline phosphatase technique. Panel B shows an electron micrograph of extracellular arenavirus-like virions. Particles (arrows) are round, vary in size, and have surface projections on the perimeter. Cellular ribosomes are visible within the virions. The length of the bar corresponds to 100 nm.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med. 2008 Mar 6;358(10):991-8.

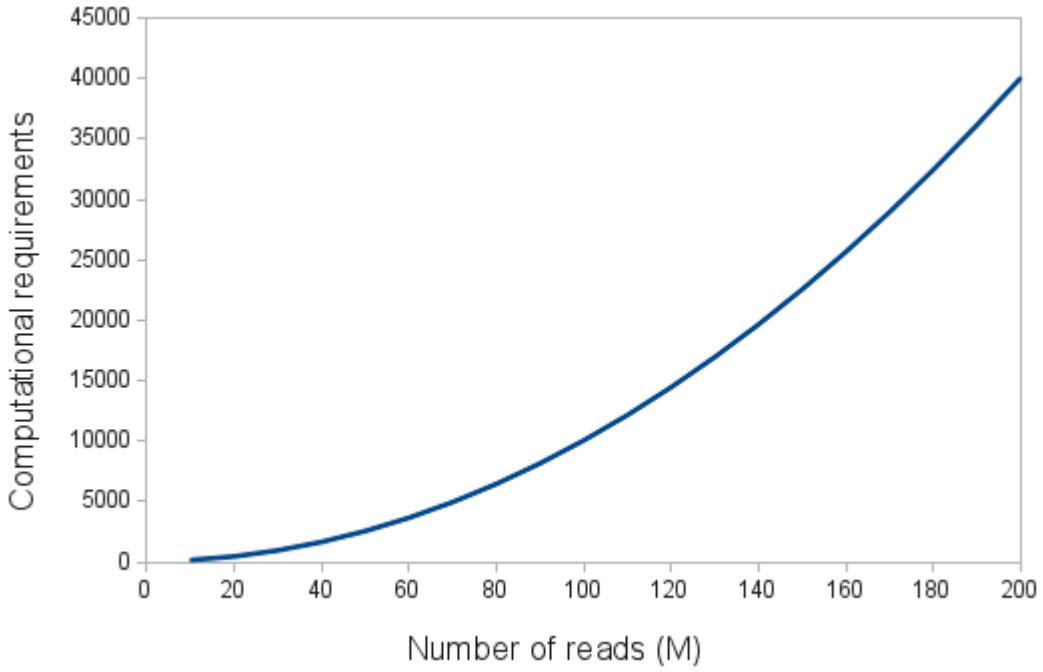
De novo assembly

Main limitations:

- Computers
- Bioinformatics skills

Experimental:

- Low prices
- New approaches (long sequences)



Cucurbita transcriptome

454 libraries C. pepo:

- Mu16:
  - 407.723 reads
- Upv196:
  - 392.370 reads

Illumina libraries (leaves, fruits, flowers and roots):

- Upv196:
  - 111.986.136 reads
- Mu16:
  - 116.484.480 reads

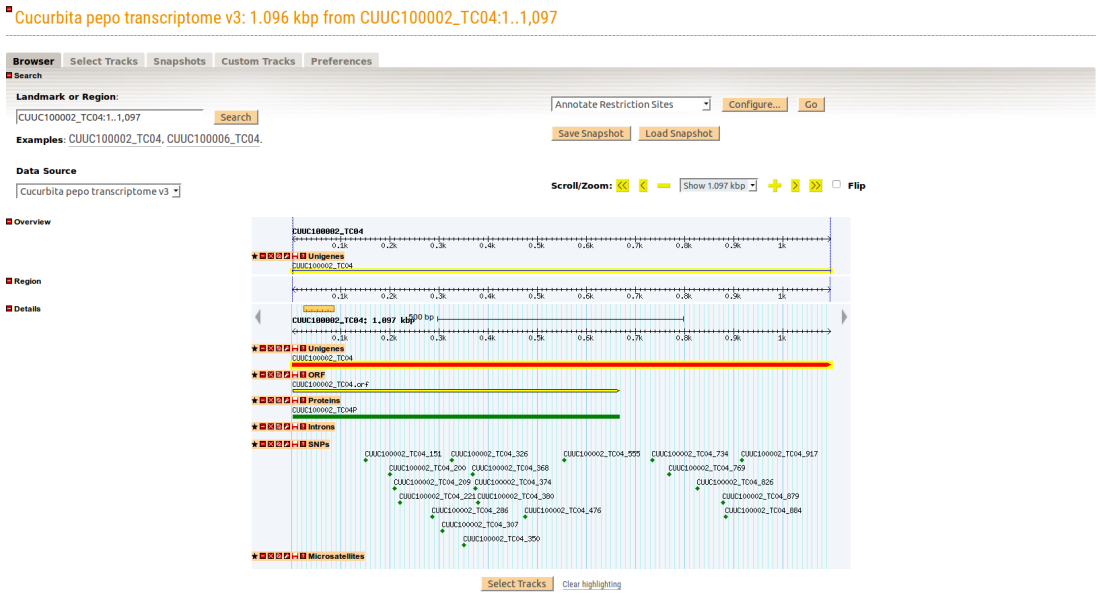
Transcriptome annotation

Unigenes Clusters: 73.239

Number transcribed clusters: 108.062

Features annotated:

- ORFs: 65.990
- Cucumber orthologues: 29.632
- 255.554 GO terms



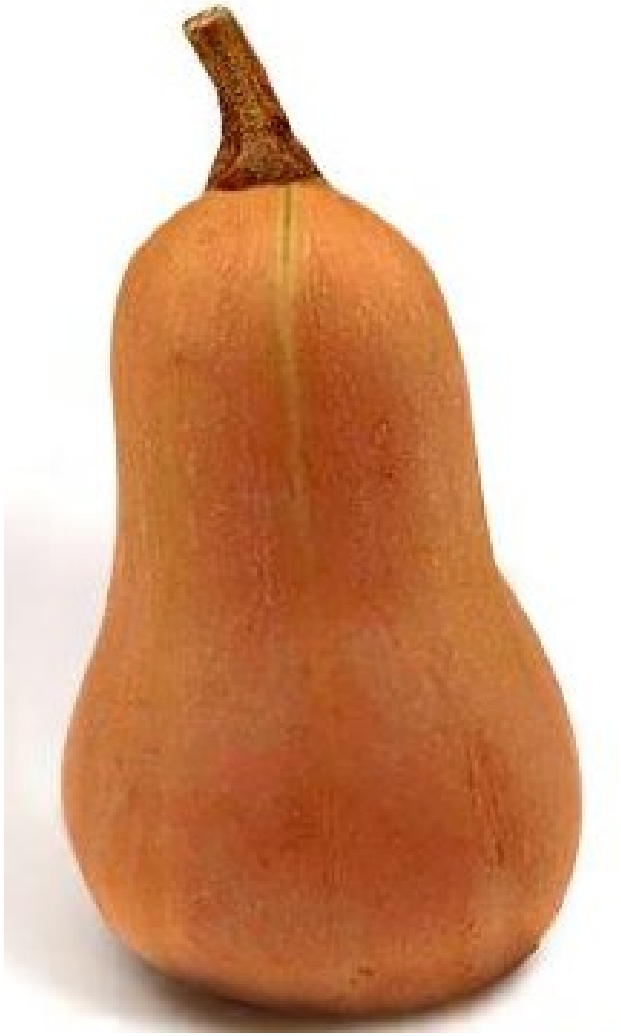
Cucurbita genome

Libraries:

2 lines Pair-end
1 line: 3 and 5Kb mate pair
1 line: 20 Kb mate pair

Genome Size: 280Mb
7986 scaffolds
N50 1,85 Mb.

Linked to genetic map with 9350 markers



Tigger genome

Table 1 | Global statistics of the *Panthera* genomes.

Sequencing (species)	Insert size	Total data (Gb)	Sequence coverage (×)
Amur tiger	170, 500, 800 bp	203.72	83.5
	2, 5, 10, 20 kb	84.48	34.6
White tiger	400 bp	86.35	32.1
Snow leopard	400 bp	108.94	40.5
African lion	400 bp	98.47	36.6
White lion	400 bp	84.43	31.4

Amur tiger assembly	N50 (kb)	Longest (kb)	Size (Gb)
Contig	29.8	287	2.35
Scaffold	8,840	41,607	2.41

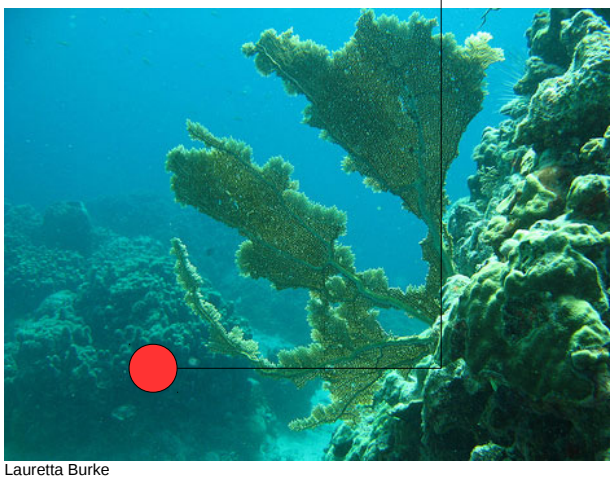
Amur tiger annotation	Number	Total length (Mb)	Percentage of genome
Genes	20,226	718.9	29.5
Repeats	—	958.9	39.3

The statistics were based on Amur tiger genome size (2.44 Gb), estimated by K-mer analysis. Contigs and scaffolds above 100 bp length were included in the statistics.

Nature Communications 4, Article number:2433 doi:10.1038/ncomms3433

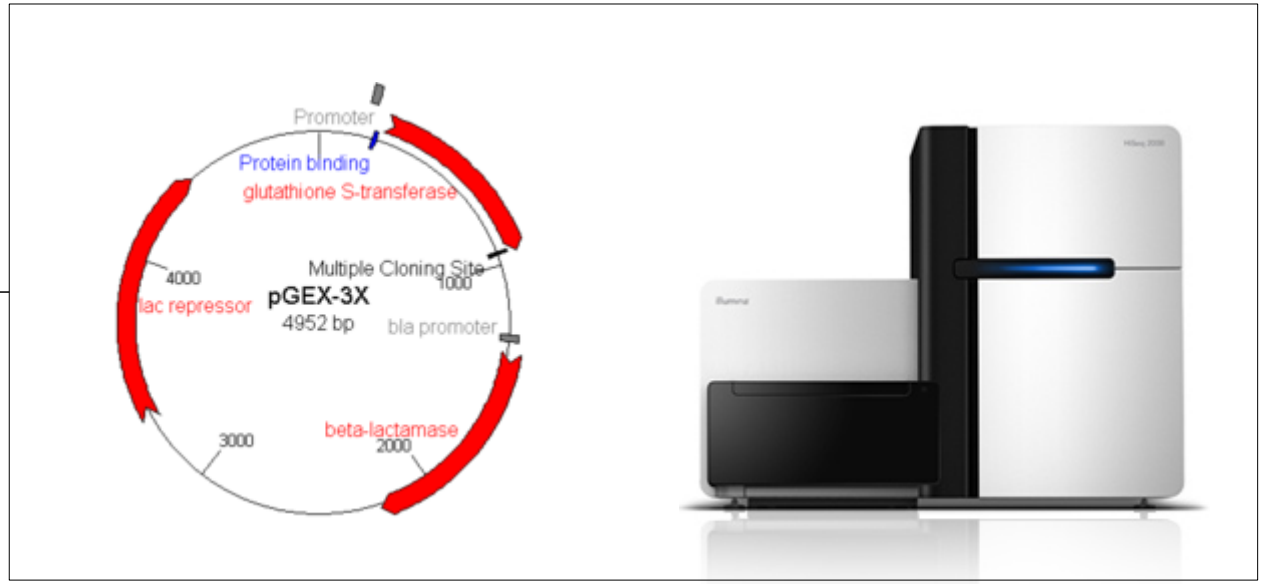
Metagenomics

Foto Luis Miguel Bugallo Sánchez en wikipedia.org



Lauretta Burke

Study of genomes of environmental samples without previous culture.



Oral microbiomas related with caries

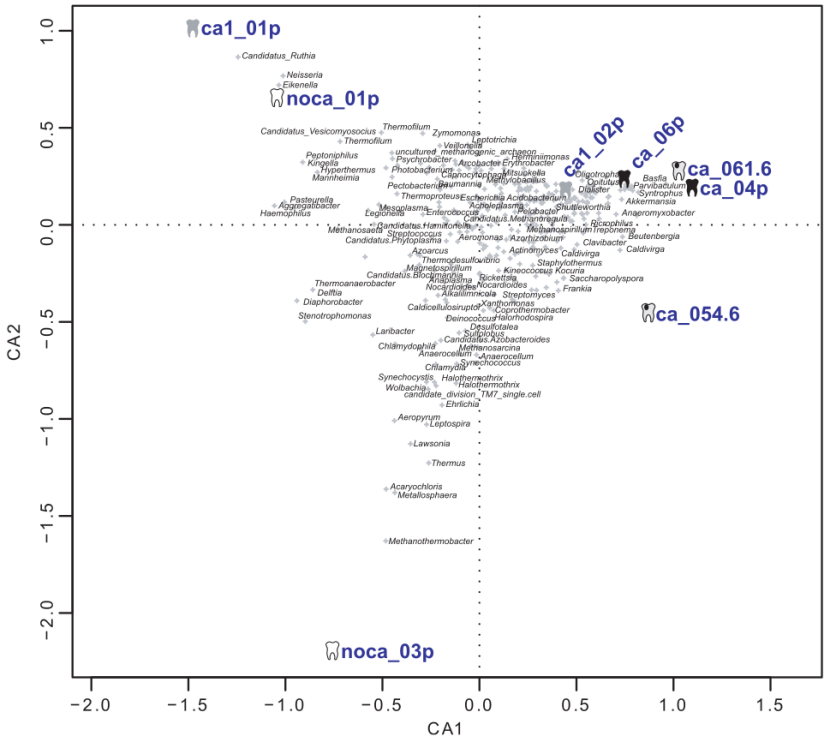
Eight oral samples with different level of caries

2 million 454 GLX Titanium sequences

Using Human Microbiome Project and the Human Oral Database

Have mapped 1,2 millions in 1150 genomes

The samples with caries are clustered



Alcaraz LD, Belda-Ferre P, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M, Mira A. Identifying a healthy oral microbiome through metagenomics. Clin Microbiol Infect. 2012 Jul;18 Suppl 4:54-7

Future

What are you doing or planning?

Jose Blanca
Joaquin Cañizares
COMAV institute
bioinf.comav.upv.es