

# **Applications of Illumina/Solexa sequencing technology for grape genomics**

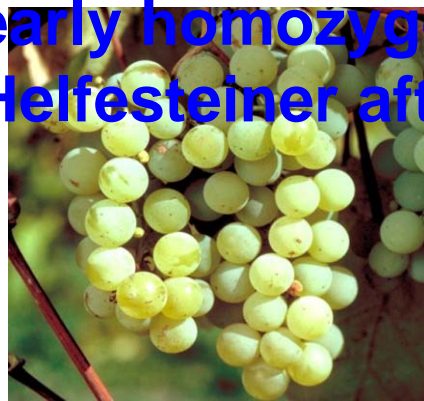
**Federica Cattonaro**  
***Illumina Seminar***  
**Milano, 13 giugno 2009**

# THE FRENCH-ITALIAN PUBLIC CONSORTIUM FOR THE SEQUENCING OF THE GRAPEVINE NUCLEAR GENOME



# THE PLANT TO BE SEQUENCED

PN40024 is a nearly homozygous clone (~ 93%) derived from Pinot noir and Helfesteiner after 6 cycles of selfing (Colmar, Fr)

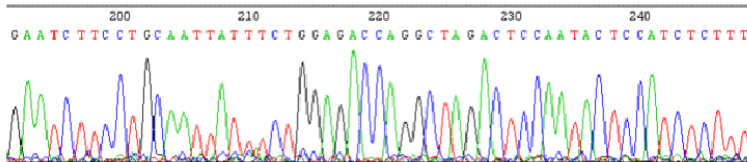
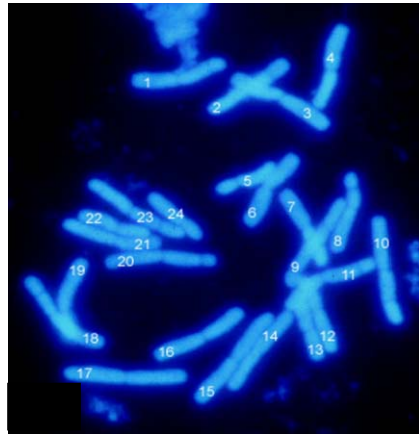


*Cultivated varieties*

**PN40024**

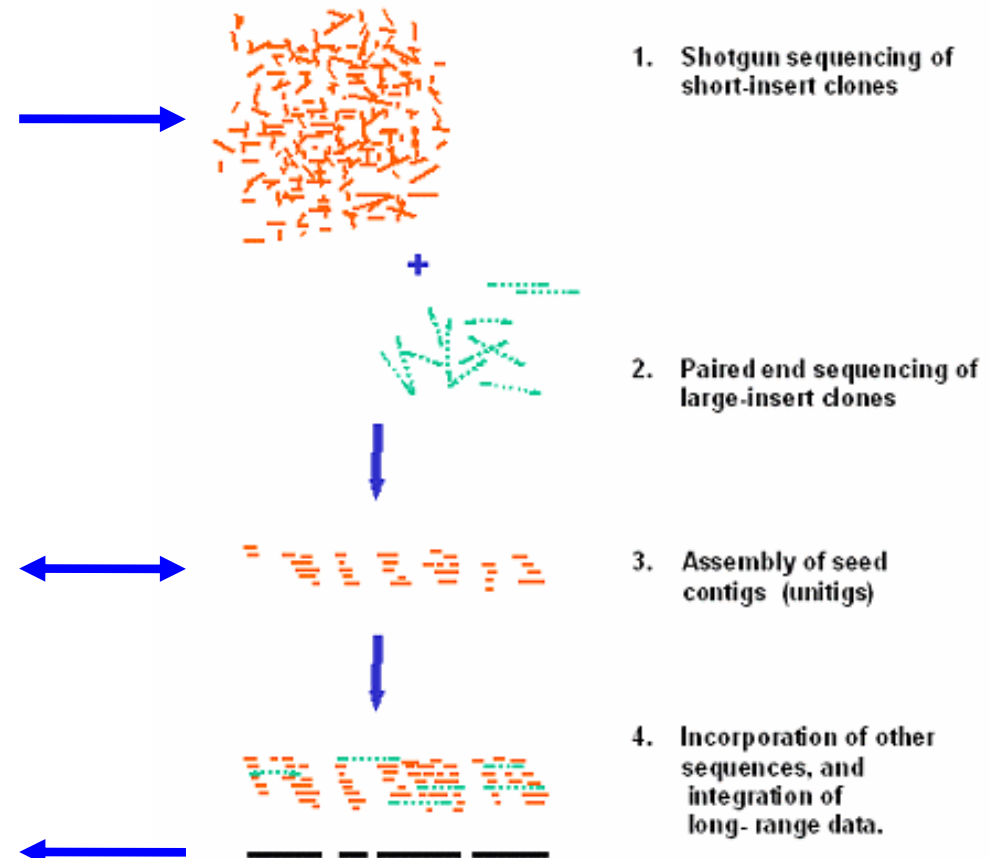


$2n = 38$



**480 Mbp**

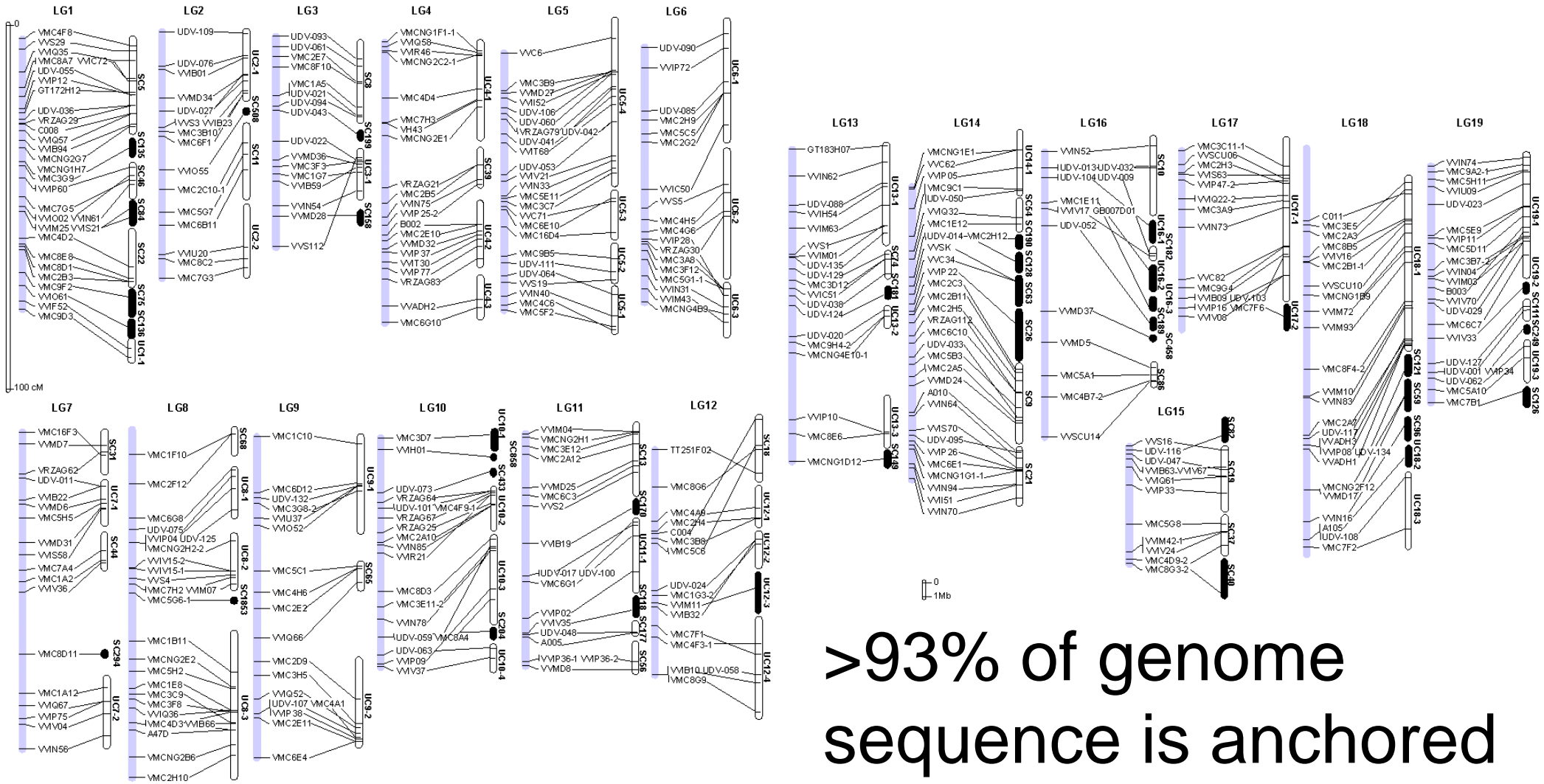
## Whole genome shotgun



**Final coverage: 12X genome equivalents**

**Final assembly size: 481 Mbp; N50=42; N50 size= 3,4 MB**

# Anchoring the sequence to the genome



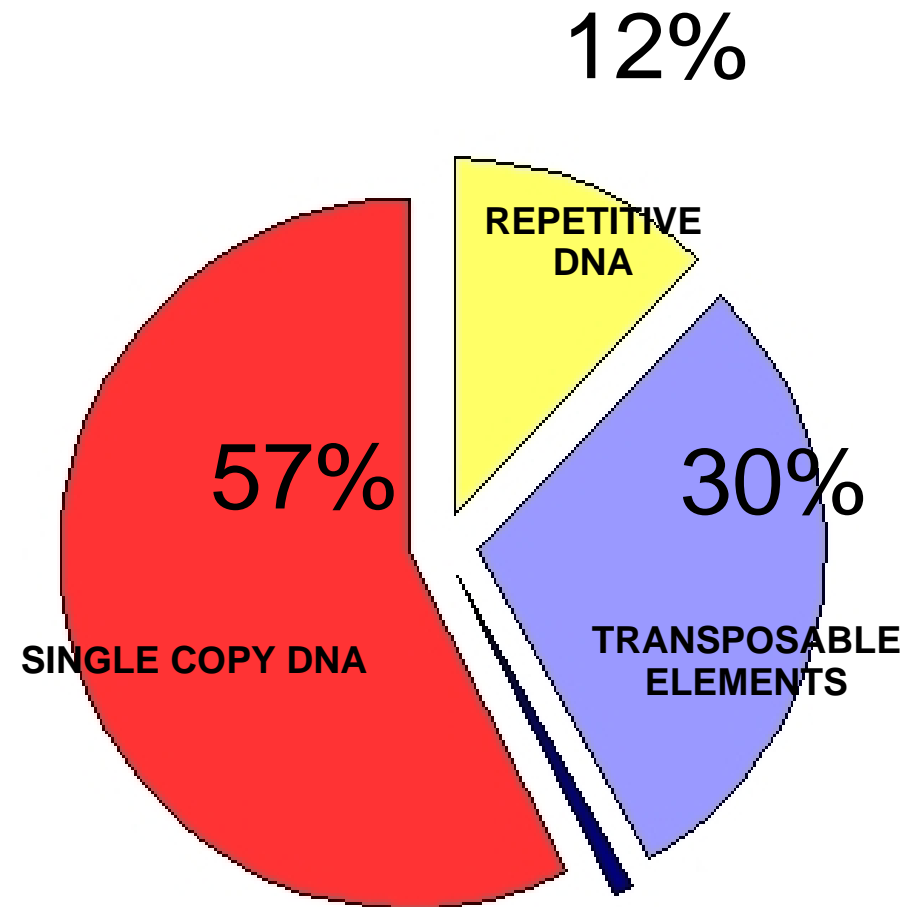
>93% of genome  
 sequence is anchored  
 in 208 scaffolds

# Repetitive DNA and Transposable Element annotation

Three different approaches used:

- **ReAS software**
- **Search for TE encoded proteins**
- **Manual annotation of TE**

Estimated repetitive fraction: 42 %



# Grapevine Gene Annotation

## Combination of evidences

- Proteins (Uniprot database)
- Available grapevine EST 317,000
- Newly sequences FL cDNAs 38,500
- Eudicotyledon ESTs 2,181,790
- *Ab initio* prediction (Geneid, SNAP, Exofish, Genewise, Est2genome)
- Gene model building using GAZE

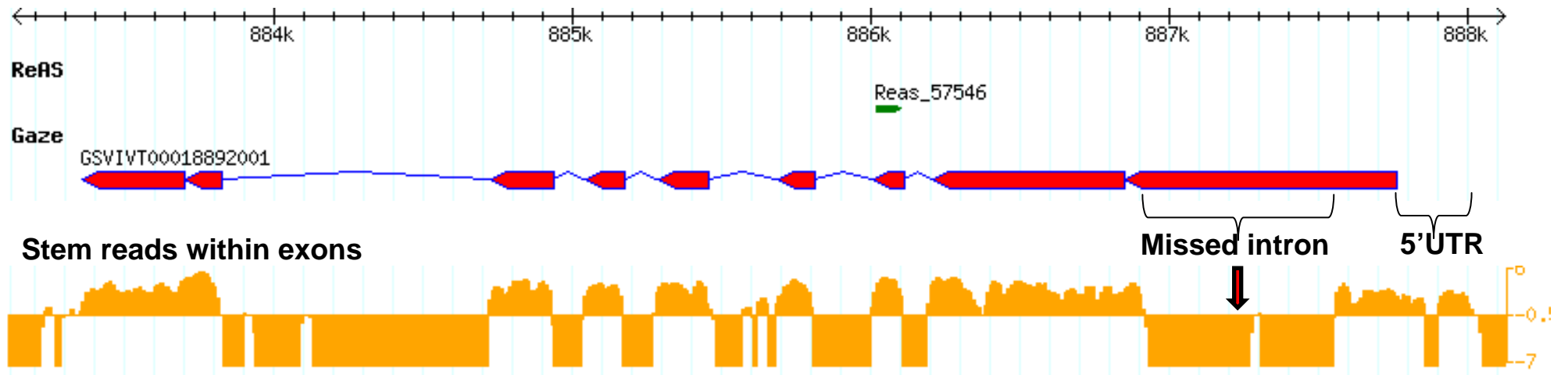
## Predicted Genes: 30,434

- 372 codons and 5 exons per gene
- Exon CDS ~ 7% of the genome
- Introns ~ 36.7 % of the genome

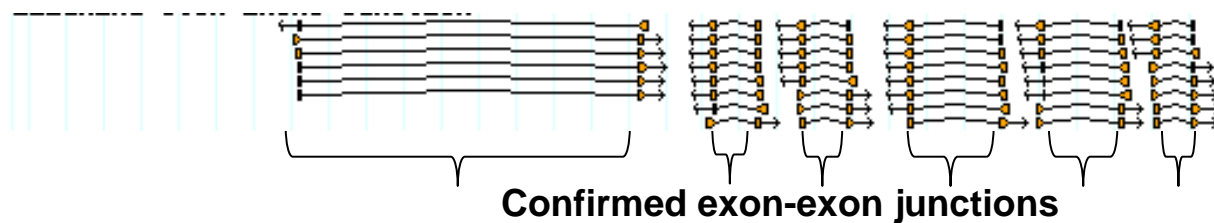
# IMPROVING GENE AND REPEAT ANNOTATION

- Use new high throughput sequencing technologies (Illumina RNA-Seq)
- Whole transcriptome shotgun sequencing
  - RNA-Seq
  - 4 different tissues, same strain sequenced
  - Improve gene annotation
- Massive sequencing of smallRNAs
  - microRNA
  - siRNA: repeat annotation

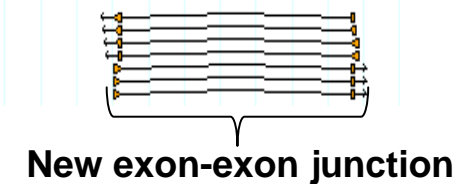
# IMPROVING GENE PREDICTIONS



Stem reads spanning predicted exon-exon boundaries

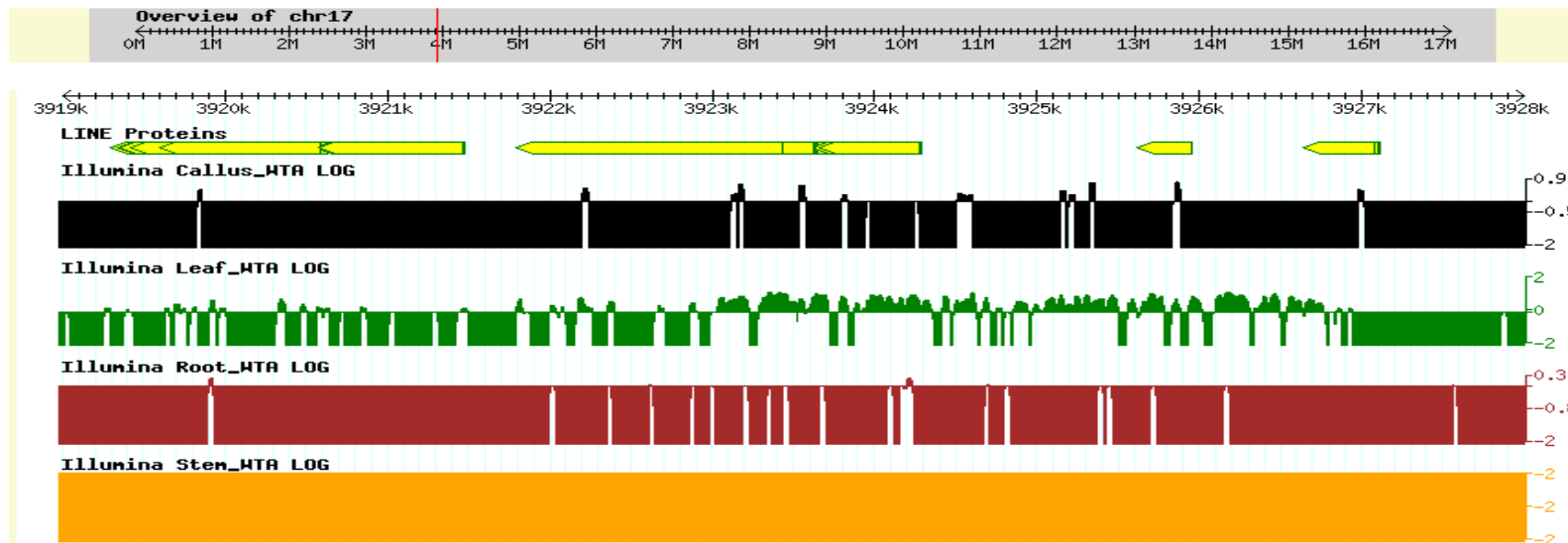


ILLUMINA stem versus 8x (gap 5000\_e0)



## GRAPE TRANSCRIPTIONAL LANDSCAPE

- No pervasive transcription throughout genome
- Rare alternative splicing events
- Very rare antisense transcripts
- LINE, Copia-LTR elements are transcriptionally active

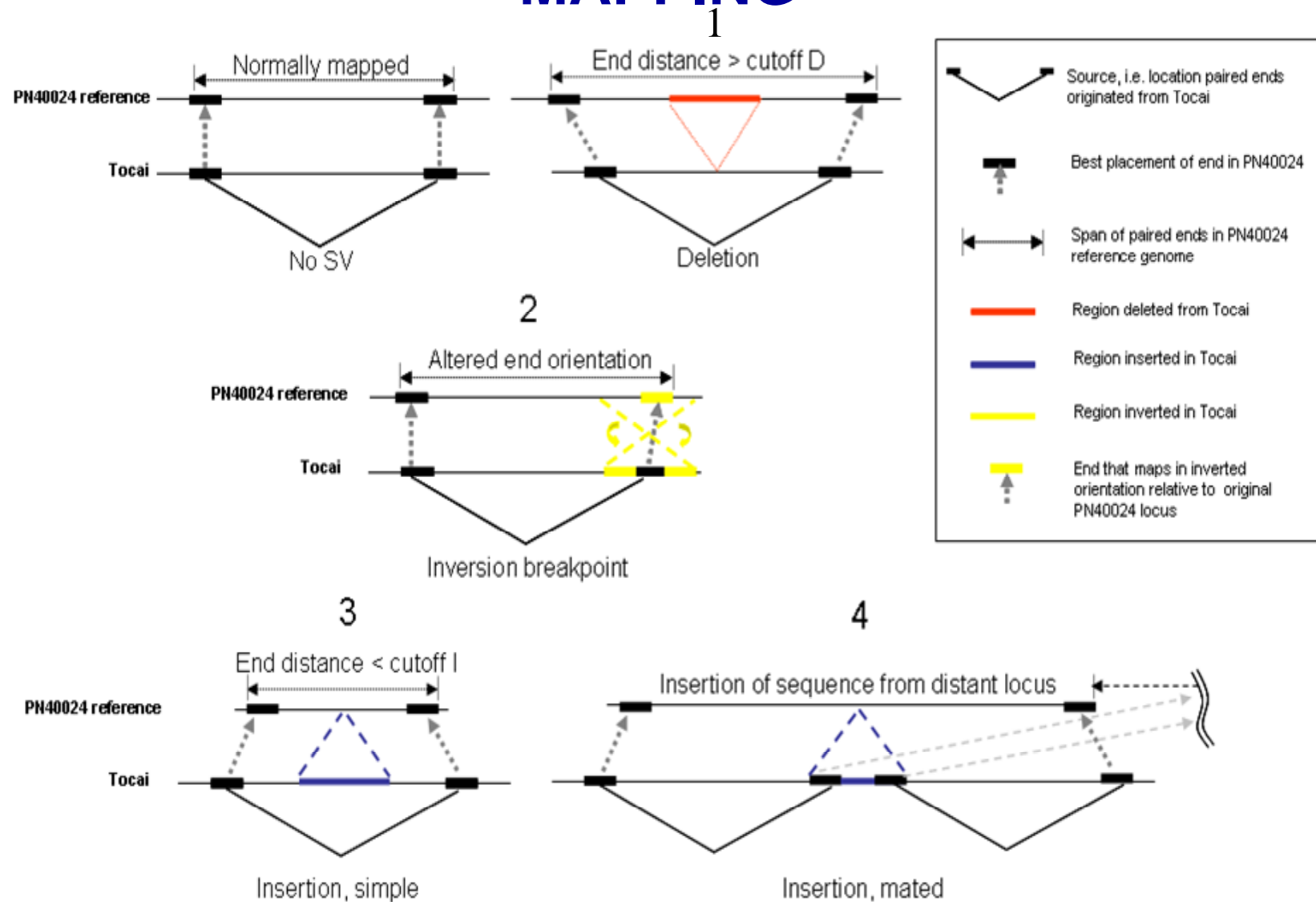


## **RESEQUENCING: GOING AFTER VARIATION**

600000 Sanger paired reads (0.9X-2.9X)  
from Tocai cv. plasmid library (4.2 kb)

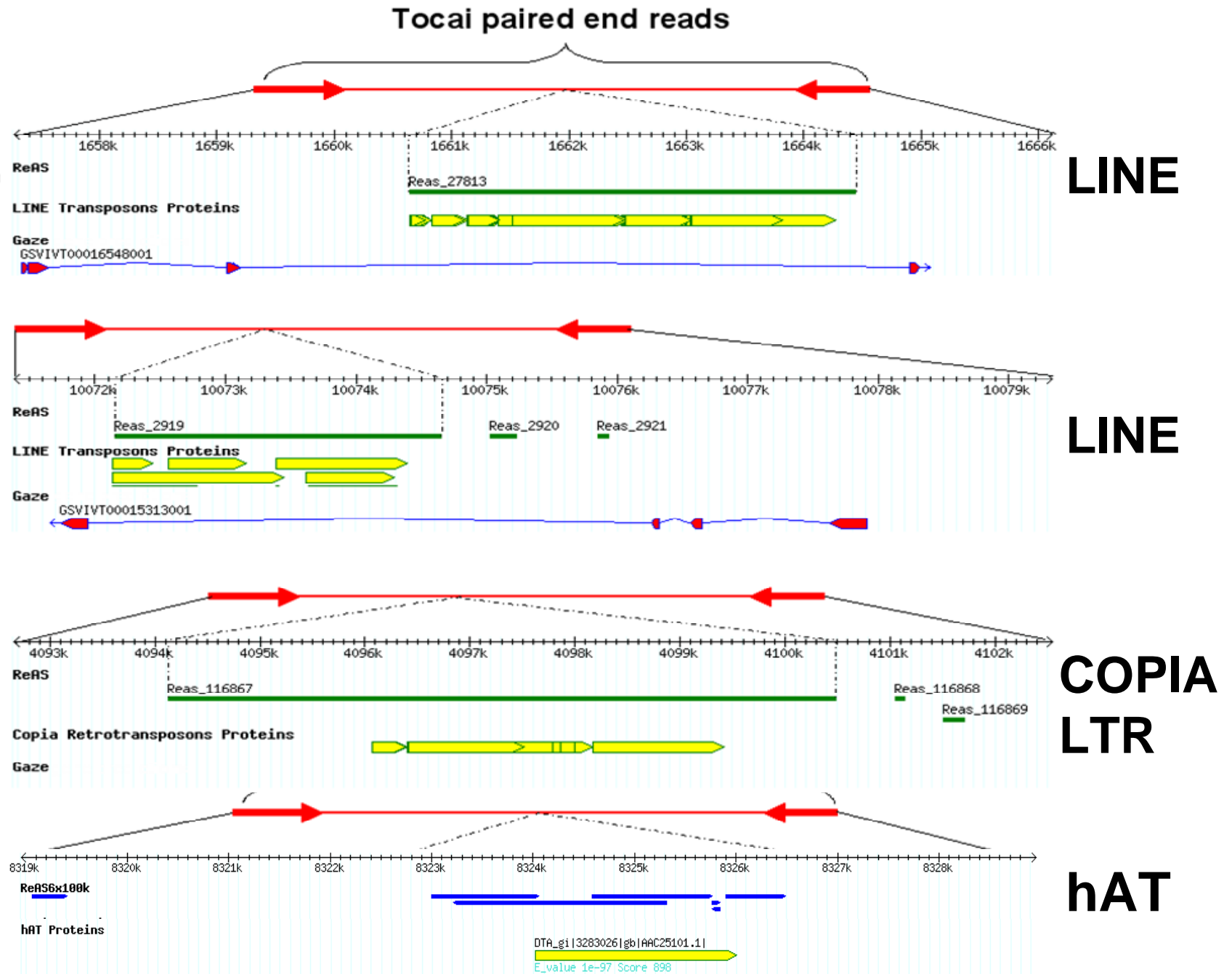
Nucleotide variation (SNPs)  
Structural variation (large indels)

# DETECTING STRUCTURAL VARIATION BI PAIR-END MAPPING



*Modified from Korb et al., Science Oct 2007*

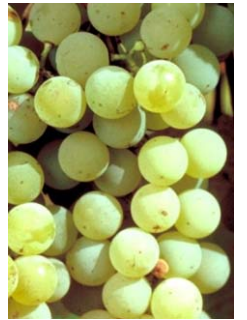
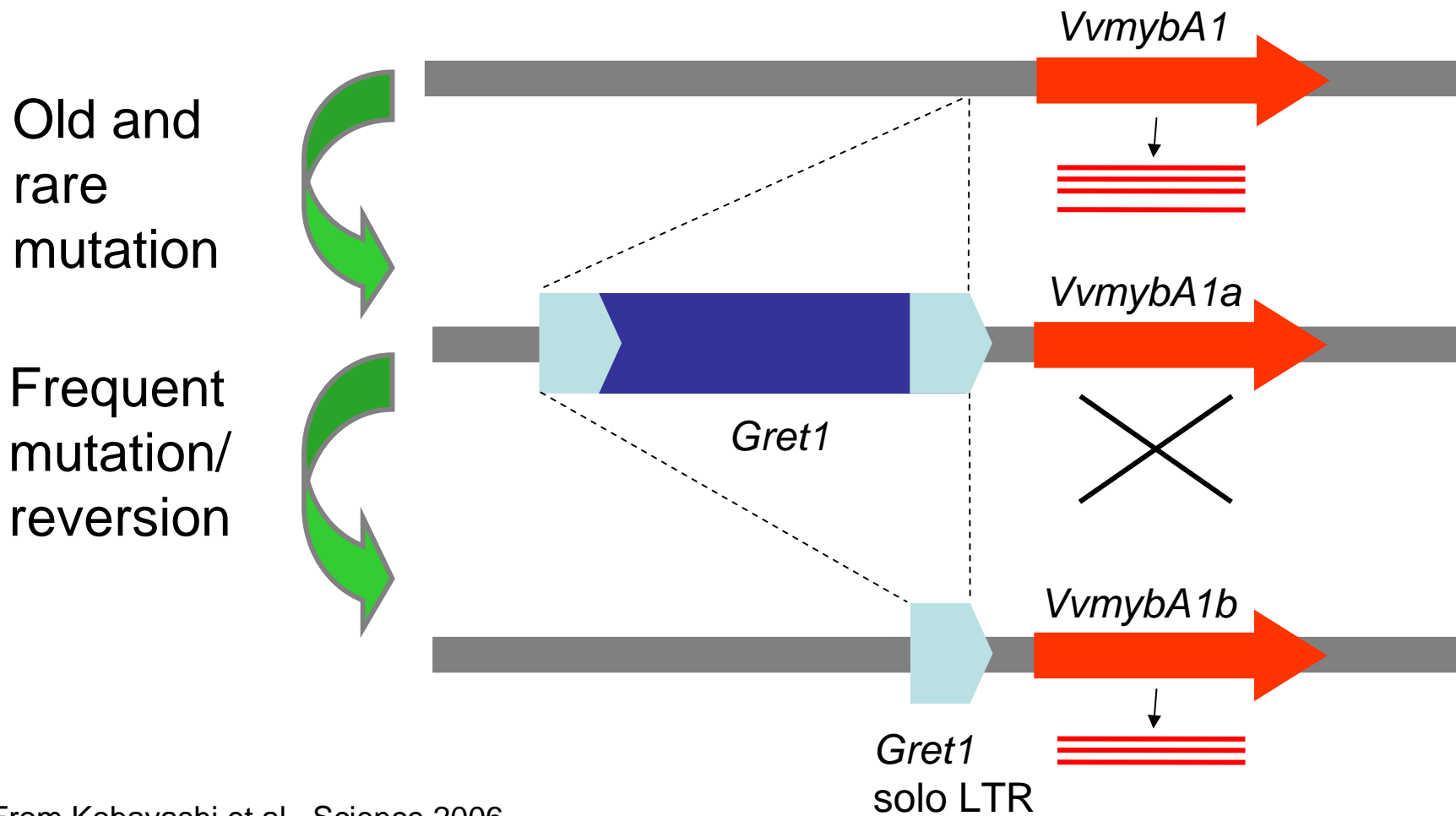
PN40024  
 8.4 X  
 assembly  
 scaffolds



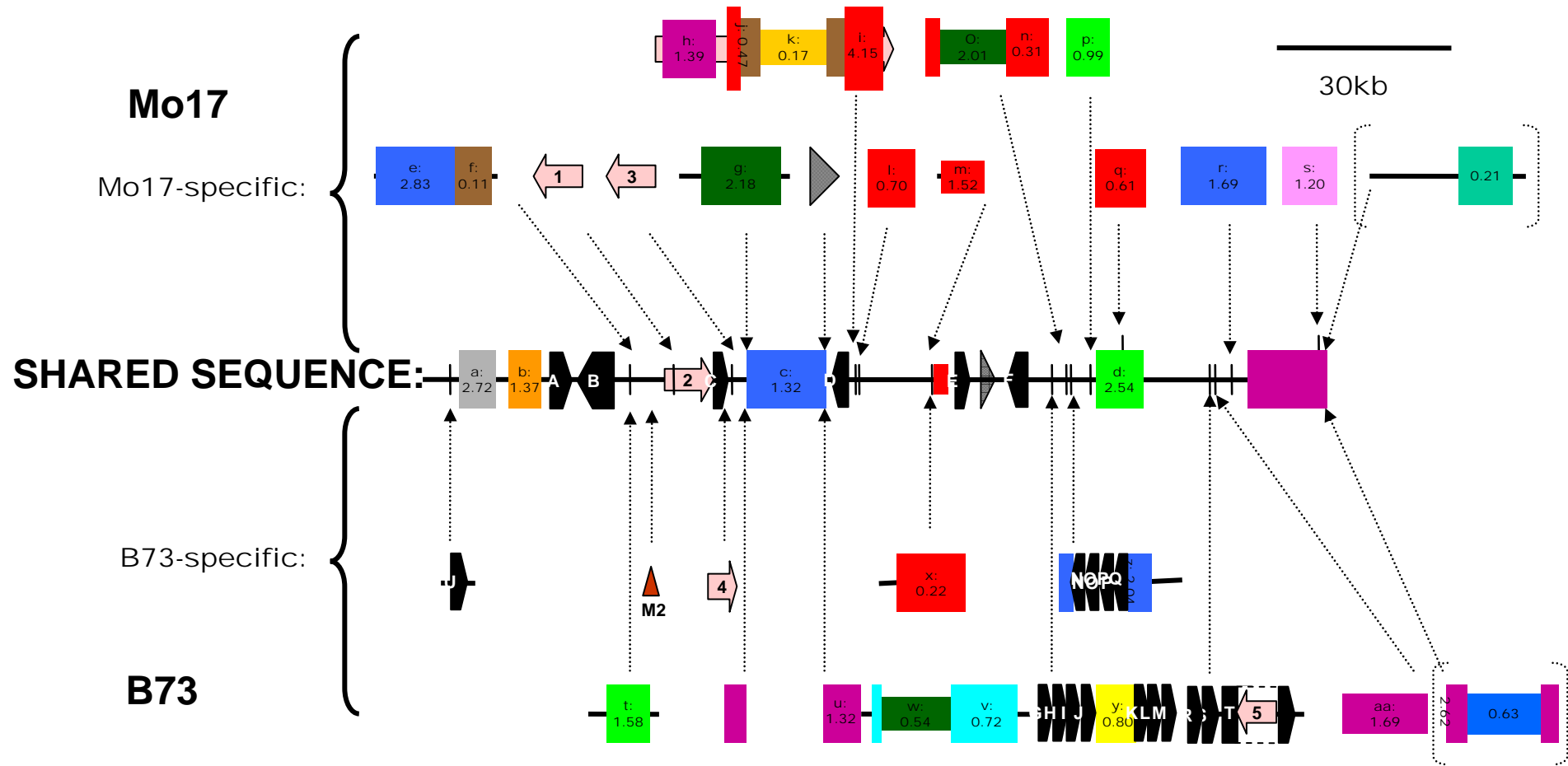
## **Tocai and PN40024 genome differ to a greater extent in SVs than in SNPs in terms of affected nucleotides**

- **At least 1800 unique insertions in PN40024 with respect to Tocai**
- **~2000 unique insertions detected in Tocai with respect to PN40024**
- **At least 150 unique inversions**
- **Preliminary count of at least 1 SV event every 133 Kbp**
- **Many insertions due to transposition events**
  - **Polymorphic LINE insertions in introns**

# CREATION OF NEW PHENOTYPIC VARIATION IN GRAPE BREEDING



# THE HYPERVARIABLE MAIZE GENOME: HYPER STRUCTURAL VARIATION (HSV)



Genes: A) - T): geneA9002 - geneT9002

Repetitive elements:

1) Transposons:

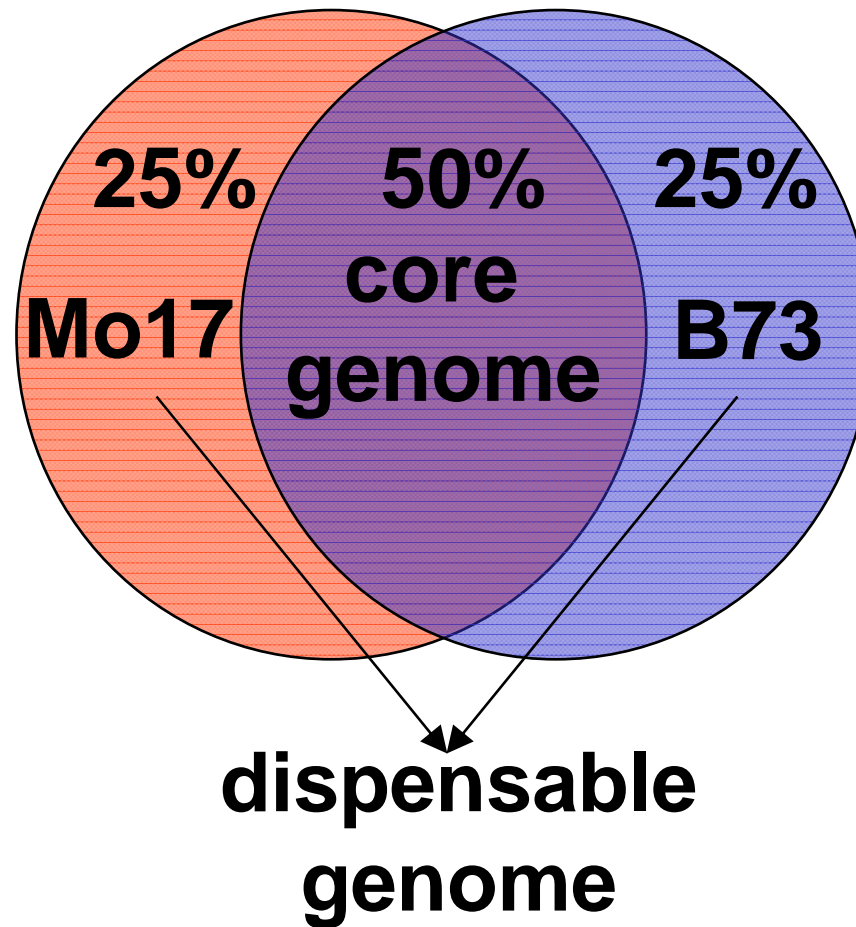
2) MITE: M2

3) Retrotransposons:

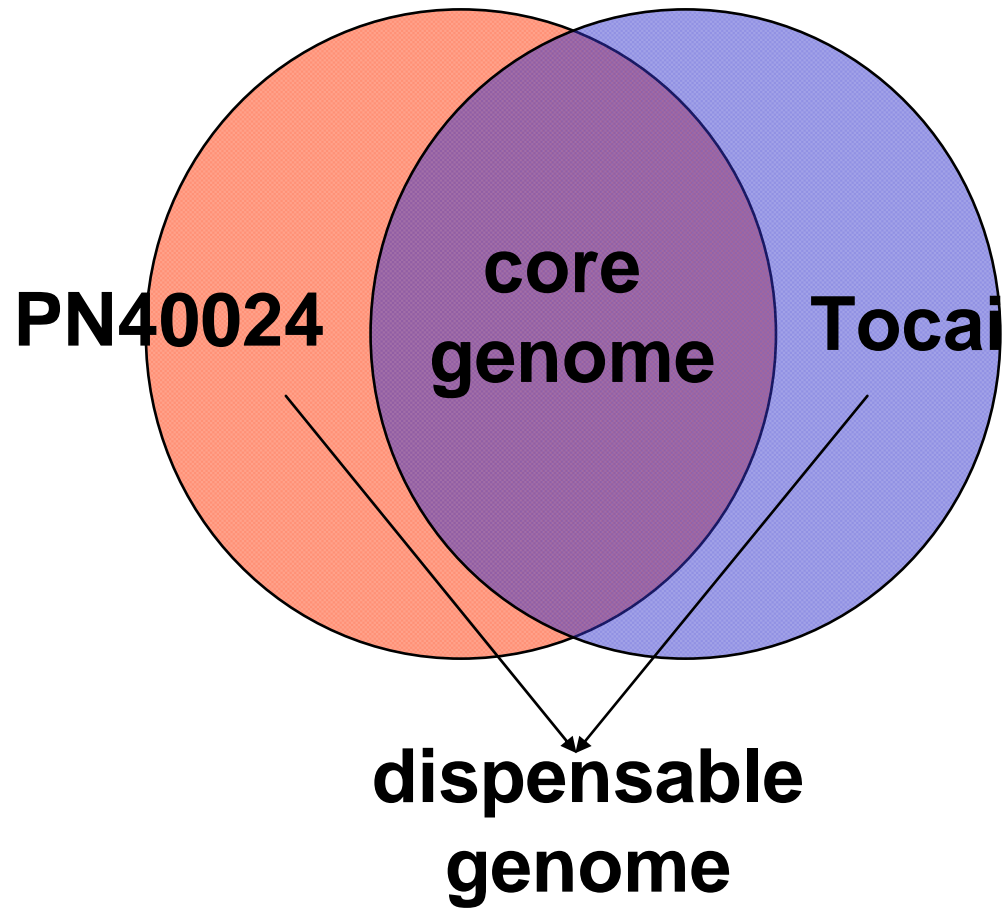
- |             |               |                    |                |
|-------------|---------------|--------------------|----------------|
| <i>ji</i>   | <i>huck</i>   | <i>ruda</i>        | <i>giepum</i>  |
| <i>opie</i> | <i>xilon</i>  | <i>rire</i>        | <i>non-LTR</i> |
| <i>jaws</i> | <i>zeon</i>   | <i>shadowspawn</i> |                |
| <i>prem</i> | <i>raider</i> | <i>dagaf</i>       |                |

locus9002 (bin 1.08; chromosome 1L)

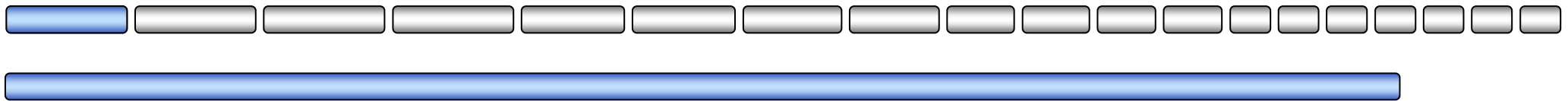
# THE MAIZE PAN-GENOME



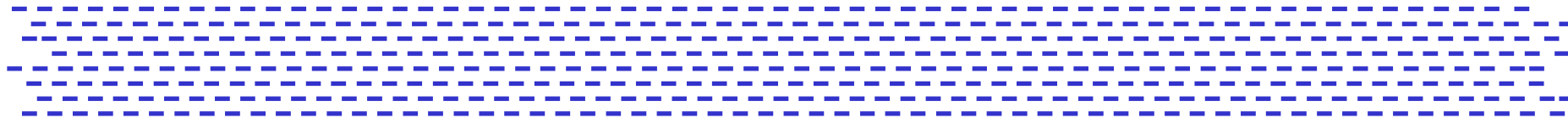
# THE PAN-GENOME CONCEPT



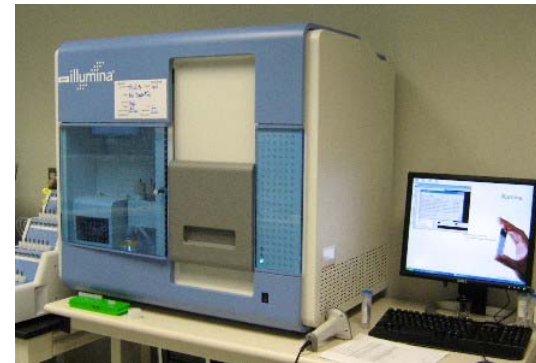
Grapevine genome = 19 chromosomes = 487 Mbp



↓  
DNA shearing



↓  
Next generation sequencing  
(Illumina Genome Analyzer II)









↓  
Alignment to reference sequence using FAST and allowing up to 2 mismatches

```
AGCTGCTAGCTAGCTTGAGATCGATCGTTCGATCGATCG(
      TGAGATCGATCGTTCGATCGATCGCATTATTCCGGATGA
        TTCGATCGATCGCATTATTCCGGATGATGCATCGTA(
          TCGGATGATGCATCGTACTATCGAT.
            TCGTACTATCGAT...
```

**AGCTGCTAGCTAGCTTGAGATCGATCGTTCGATCGATCGCATTATTCCGGATGATGCATCGTACTATCGAT...**

# Re-sequencing of grapevine varieties and clones

	reads (M)	bases (Mbp)	coverage	used_bases (Mbp)	used coverage ALL
PN40024	83,85	3294,00	6,79	3017,56	6,22
ENTAV	51,81	2590,32	5,34	1866,30	3,85
CORVINA	213,77	8550,68	17,63	7685,98	15,85
 TOCAI_R5	198,34	9244,80	19,06	7139,36	14,72
 SAUVIGNON_R3	169,06	6641,91	13,69	6122,82	12,62
 SANGIOVESE_R23	200,43	7925,63	16,34	7173,76	14,79
 TOCAI_R14	93,71	3750,13	7,73	3396,49	7,00
 SAUVIGNON_297	96,66	3866,44	7,97	3477,63	7,17
 SANGIOVESE_R5	100,09	4003,65	8,25	3601,65	7,43
<b>TOTAL</b>	<b>1207,72</b>	<b>49867,55</b>	<b>102,82</b>	<b>43481,56</b>	<b>89,65</b>

36 x 2 bp PE-reads



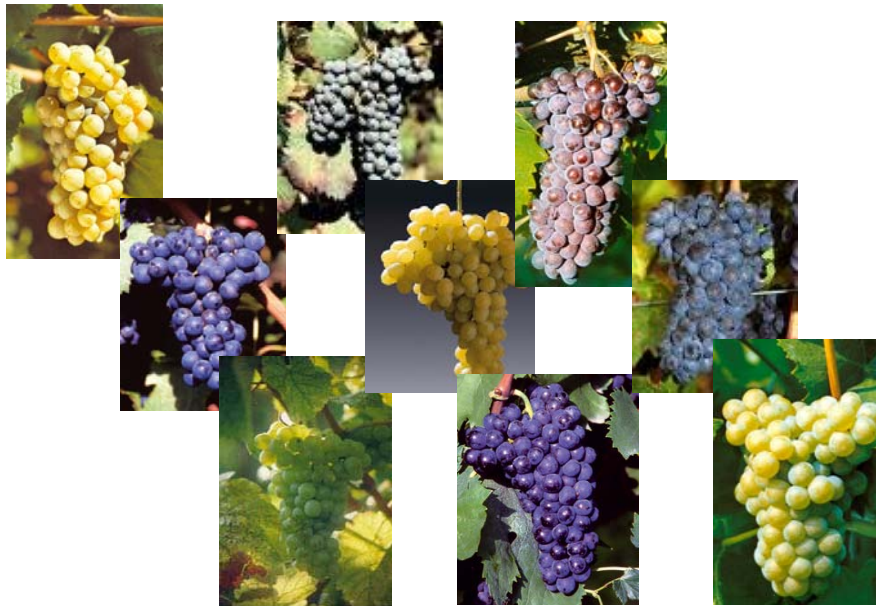
3 pairs of grape clones

## VARIETIES

- Derive from crosses  
(originate from sexual reproduction)
- Genetically distinct
- Easy to differentiate with genetic analysis

## CLONES

- Derive from somatic mutations  
(originate from vegetative propagation)
- Genetically close (same variety)
- Difficult to differentiate with genetic analysis



Pinot clones

Sangiovese clones

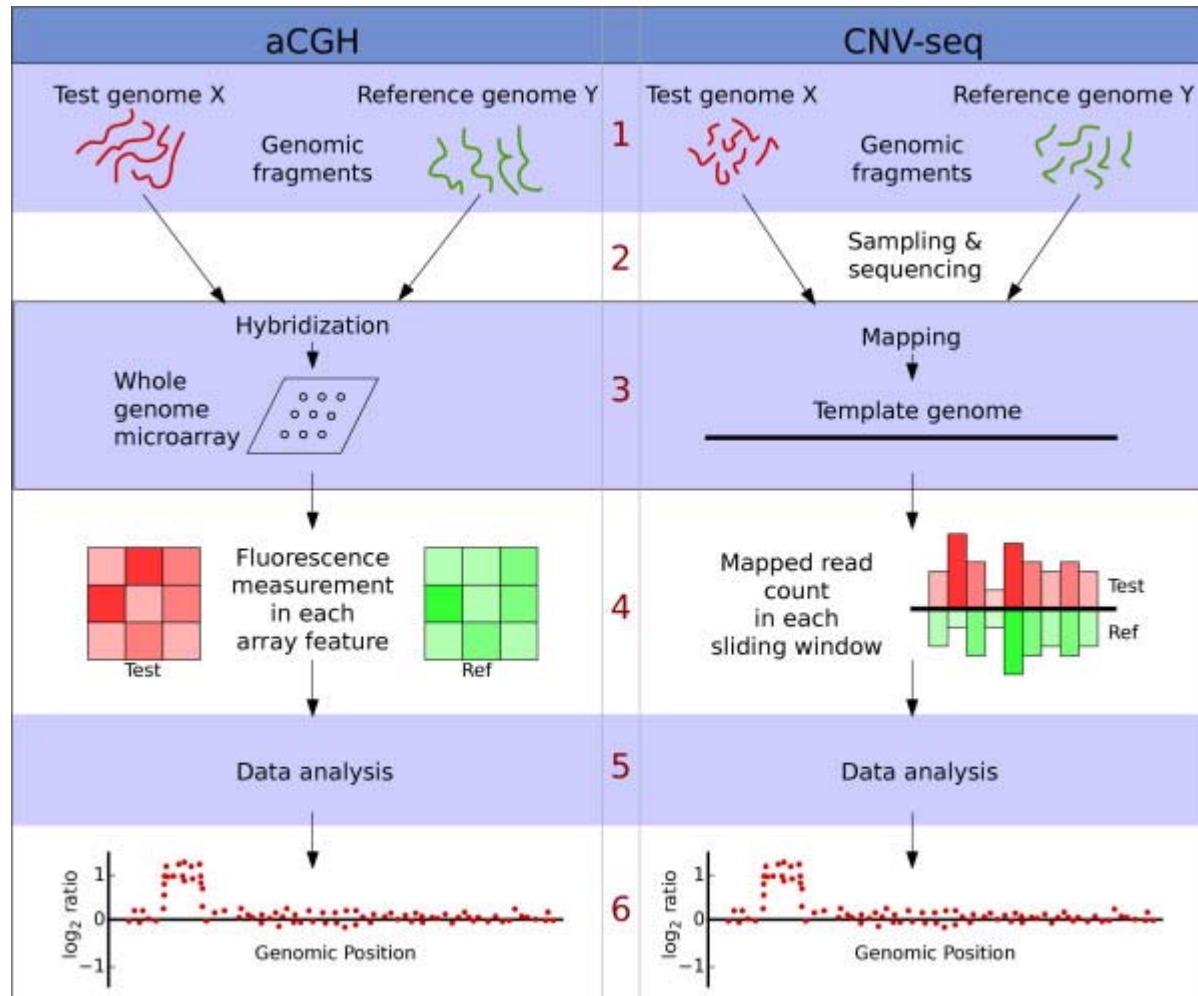
# RESEQUENCING FOR WHOLE GENOME ANALYSIS

- Resequence multiple grapevine varieties using Illumina paired-end reads
- Align reads from each resequenced individual to reference genome (PN40024)
  - Identify large structural variants from read coverage
  - Identify short structural variants (TE-related) from paired-end reads
  - Identify SNPs from uniquely placed reads

## RESEQUENCING FOR LARGE SV IDENTIFICATION (CNV-seq)

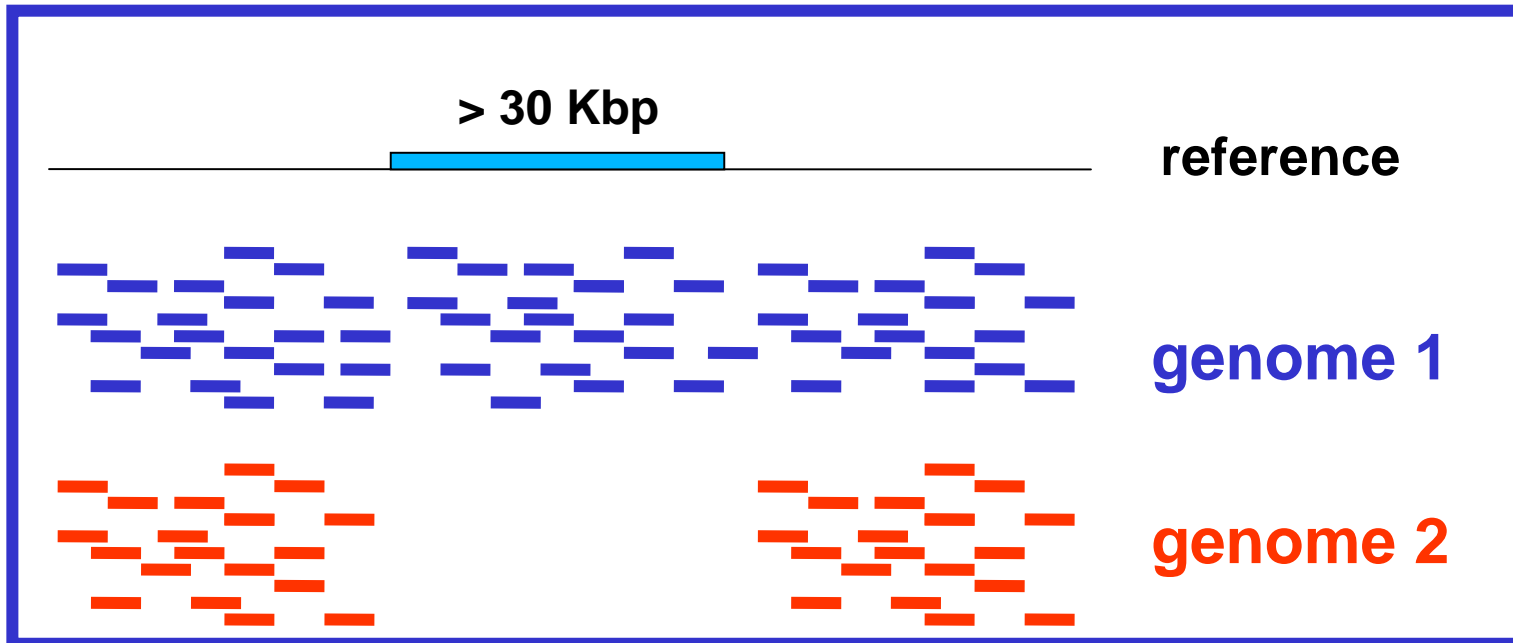
- Large SV (>50 kb) = a.k.a. CNVs in humans
- Easy to do
  - Align reads from each resequenced individual to reference genome (PN40024)
  - Identify Large SVs from read coverage along reference sequence
  - Requires low sequence coverage: 1-2X
- Asimmetry in information

# A comparison of the conceptual steps in aCGH and CNV-seq methods

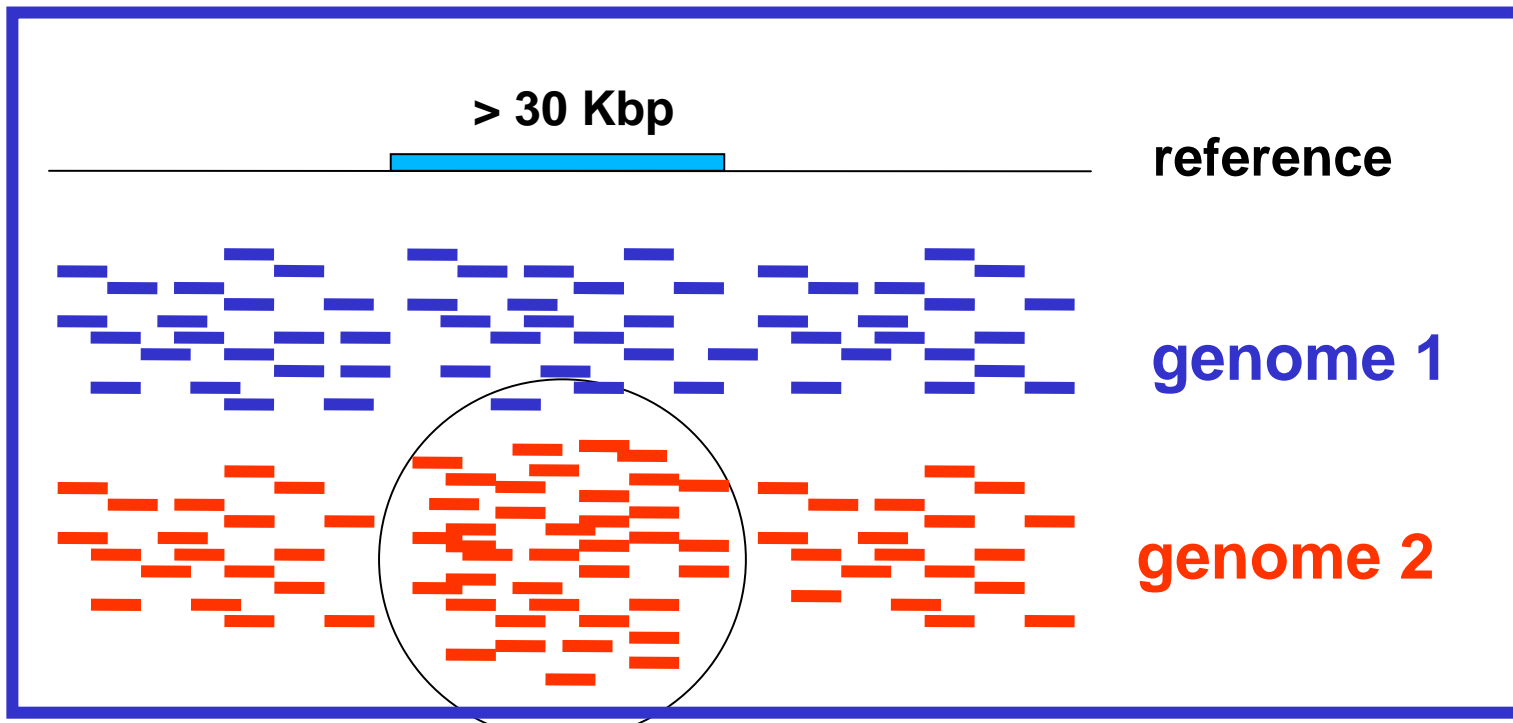


From Xie and Tammi, BMC Bioinformatics 2009

# CNVs (>30 Kbp) can be identified by deep coverage variations



→ **DELETION**  
Genome 2  
= no reads  
coverage

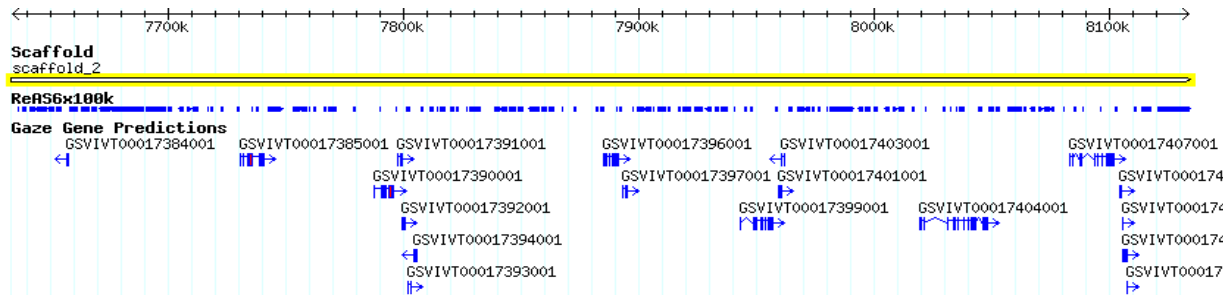


→ **AMPLIFICATION**  
Genome 2=  
anomalous reads  
coverage

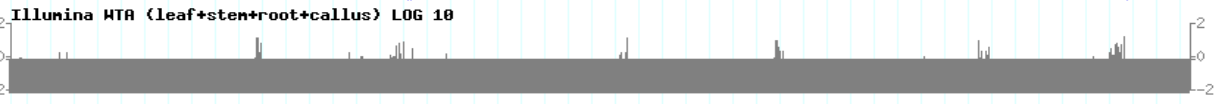
## LARGE SV COMPARISON AMONG VARIETIES

	<b>Deletions compared to PN40024</b>	<b>Insertions compared to PN40024</b>	<b>No SV compared to PN40024</b>
Tocai	34.9 Mbp 691 regions	3.1 Mbp 106 regions	447.4 Mbp 1839 regions
Pinot Noir	31.5 Mbp 523 regions	3.5 Mbp 110 regions	450.3 Mbp 1732 regions
Corvina	69.6 Mbp 1478 regions	14.7 Mbp 520 regions	399.8 M 2020 regions

# RESEQUENCING FOR STRUCTURAL VARIATION DETECTION



Gene predictions



Transcription profile

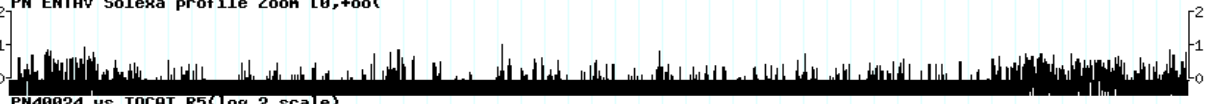
PN40024



Tocai



Pinot Noir



Resequencing read density profiles

PN40024/Tocai



Pinot Noir/  
PN40024

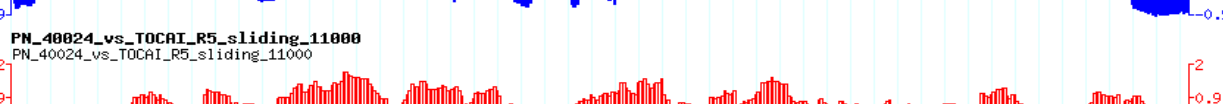


Resequencing read density ratios

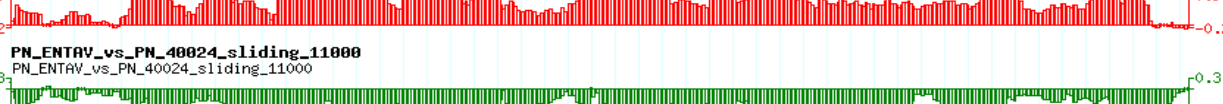
Pinot Noir/ Tocai



PN40024/Tocai



Pinot Noir/  
PN40024

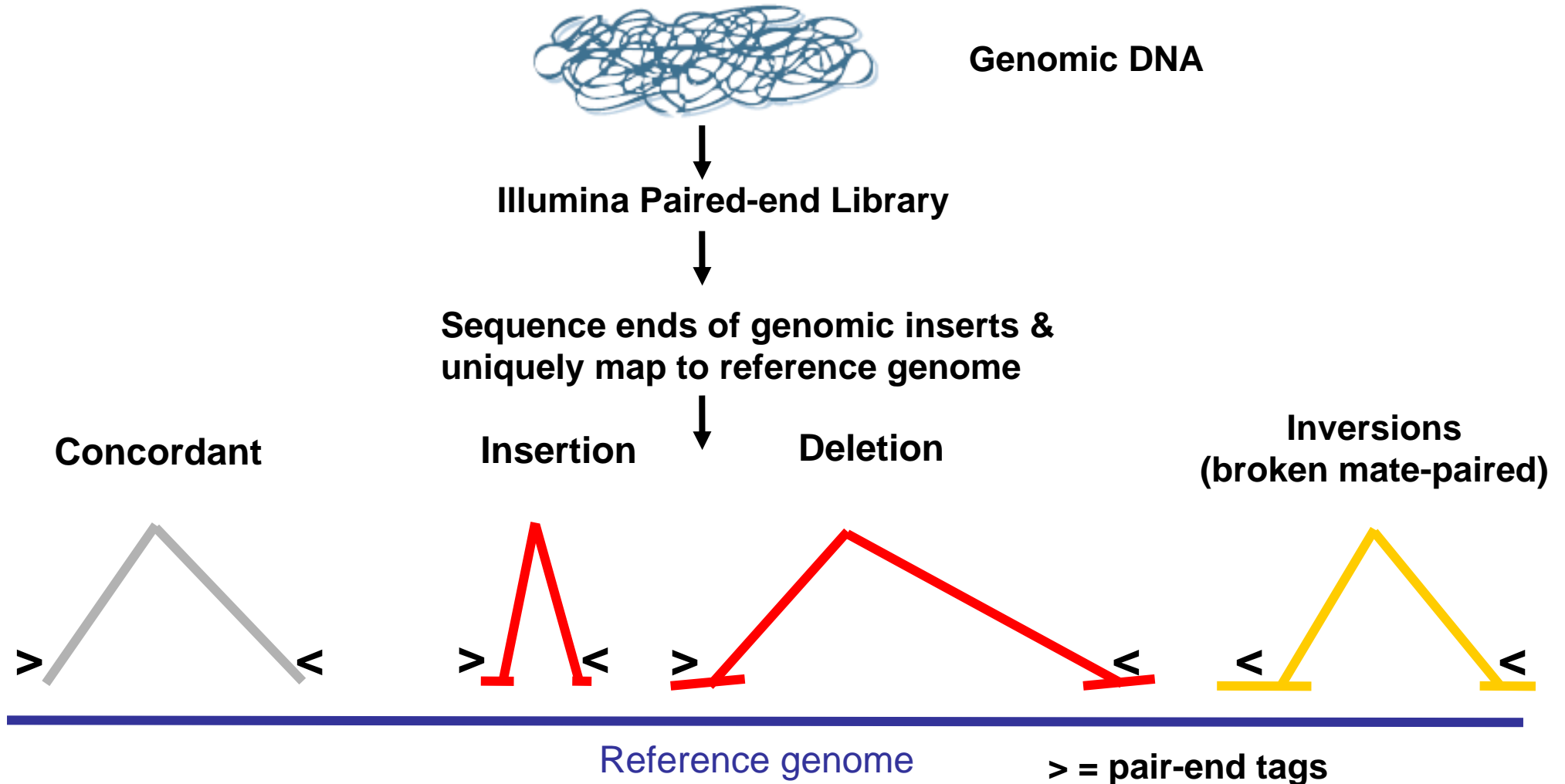


CGH signal intensity ratios

Pinot Noir/ Tocai



# Resequencing by pair-end approach to detect retrotransposons insertions



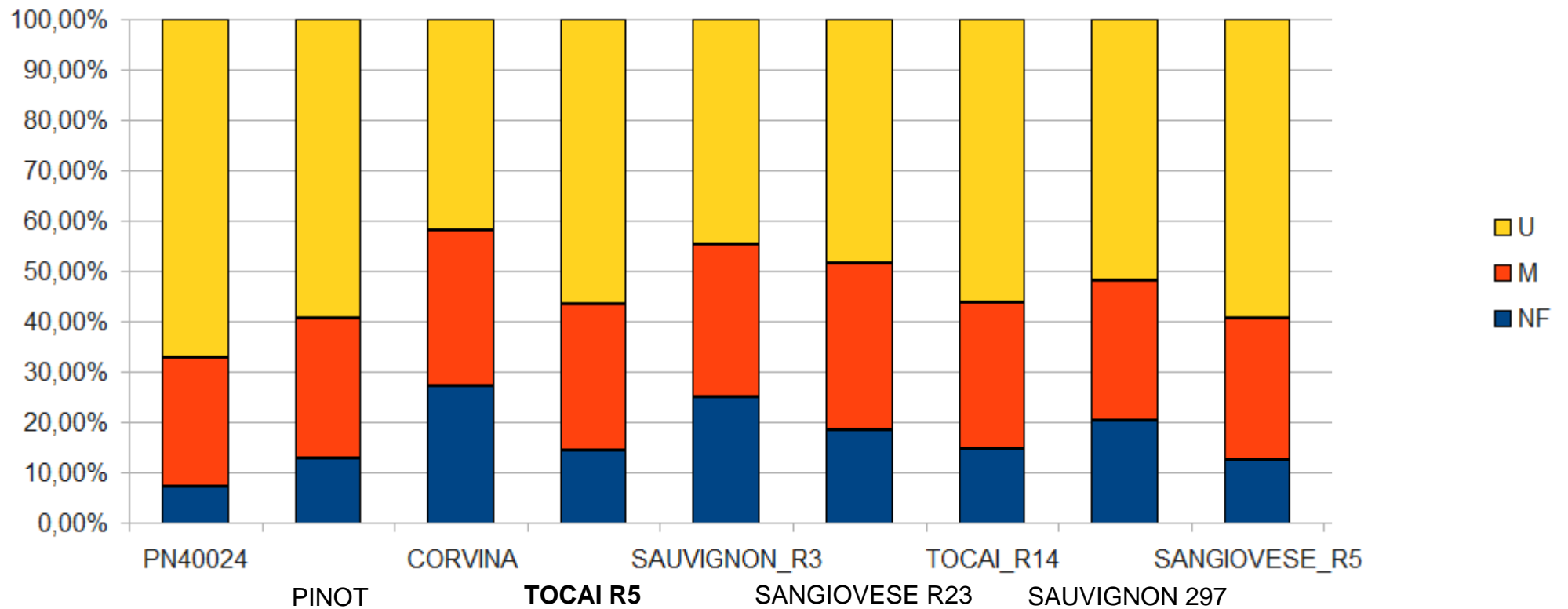
**Large Fragment mate pairs** (Illumina: 2-3Kb) overcome the problem of placing shorter reads in repeats and detect large insertions and translocations

**Small Fragment mate pairs** (Illumina: 200bp) provide sensitivity to detect small indels and resolution detecting break points

## **RESEQUENCING FOR SNP IDENTIFICATION**

- Align reads from each resequenced individual to reference genome (PN40024)
- Identify SNPs from uniquely placed reads
- Requires high sequence coverage: >20X

# MAPPING ON REFERENCE GENOME (FAST software)



# RESEQUENCING FOR SNP IDENTIFICATION

	Total reads	Mappable reads	Unique reads	Unique reads, no mismatch
PN40024	83,850,498 3.3 Gbp	77,803,552 3.06 Gbp	56,290,146 2.2 Gbp	45,565,450 1.8 Gbp
Tocai	205,401,113 9.6 Gbp	175,587,385 8.2 Gbp	116,144,877 5.4 Gbp	72,737,376 3.4 Gbp
Corvina	213,739,444 8.5 Gbp	155,782,830 6.2 Gbp	89,577,017 3.6 Gbp	51,816,364 2.1 Gbp

# RESEQUENCING FOR SNP IDENTIFICATION

	Homozygous SNPs	Heterozygous SNPs	Total SNPs
<b>PN40024</b>	19,396	36,568	55,964
<b>Tocai</b>	1,271,135	1,322,992	2,594,127
<b>Corvina</b>	864,360	730,926	1,595,286

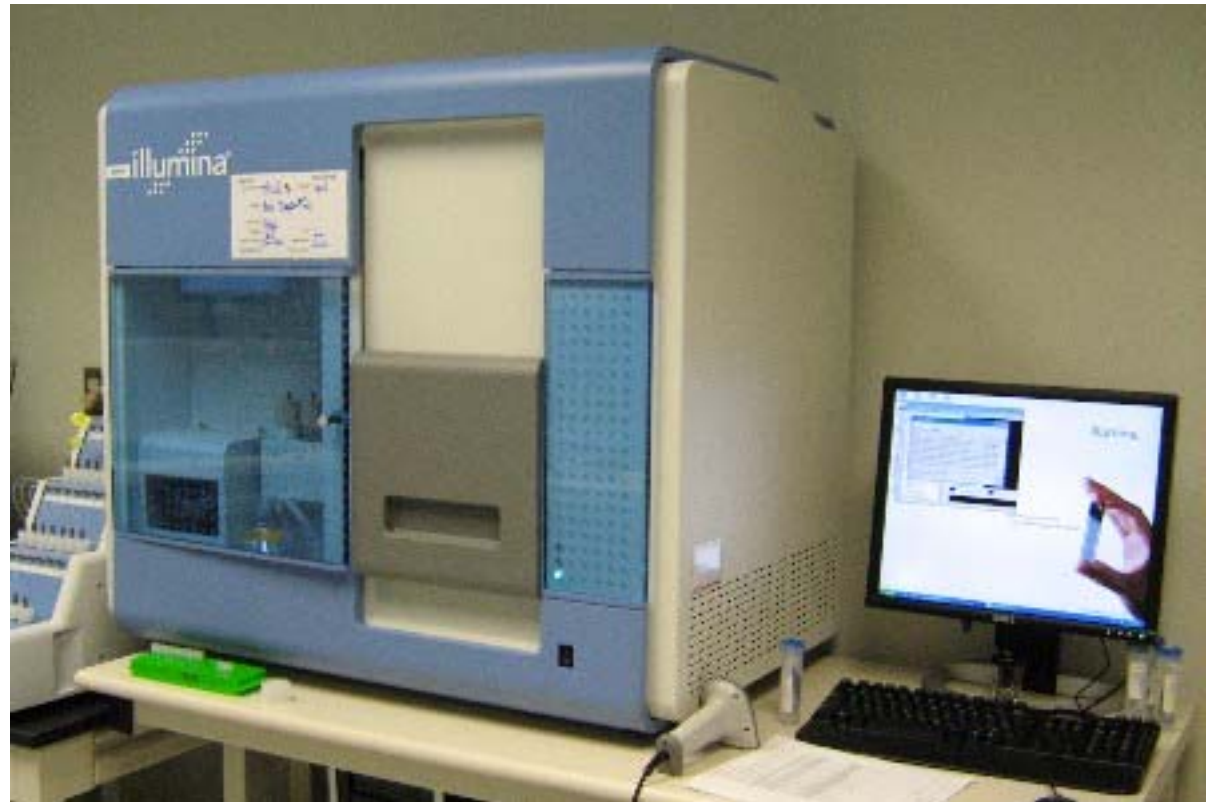
**False positive rate estimated from SNP identified from Sanger resequencing of PCR fragments: 2.1%**

# IGA laboratories (300mq)

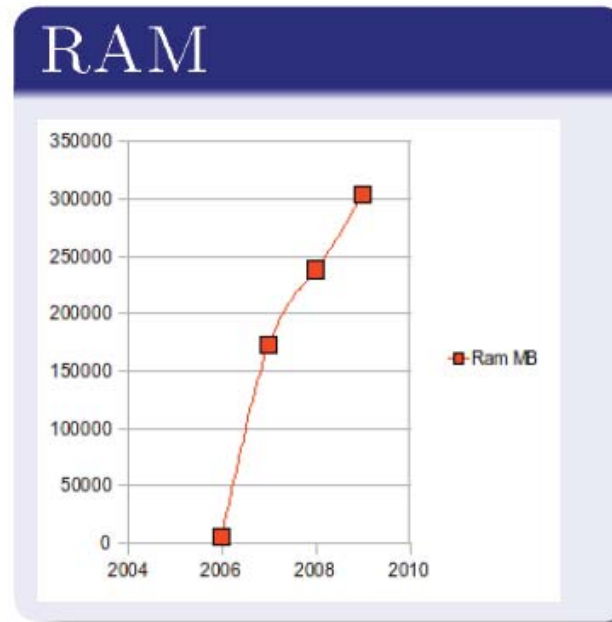
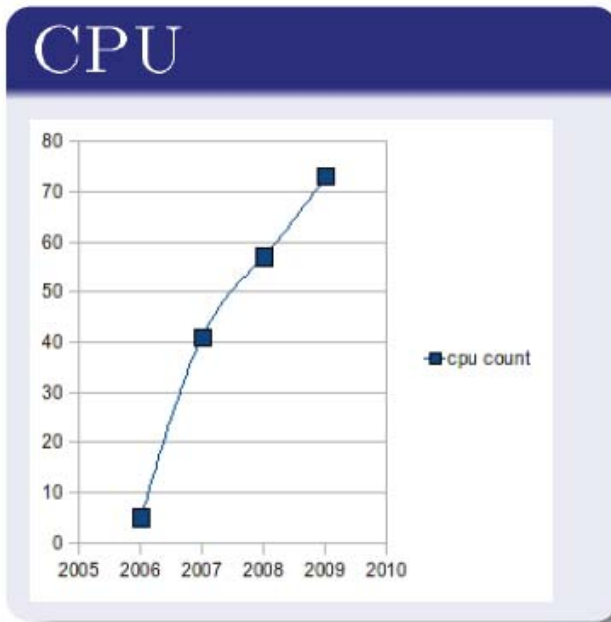


# ILLUMINA SEQUENCING PLATFORM AT IGA

- *Last run:* 16Gbp (2x75bp reads)
- *Tested applications:* DNA-seq, RNA-seq, PCR products multiplex sequencing (indexing system)
- Upgrade from GAI to GAIx (July 2009)
- Second GAIx (December 2009?)



# IT INFRASTRUCTURES GROWTH AT IGA



## **IGA Scientific Director**

*Michele Morgante*

### **Genomics**

*Federica Cattonaro*  
*Daniele Trebbi*  
*Gabriele Di Gaspero*  
*Irena Jurman*  
*Nicoletta Felice*  
*Vera Vendramin*

### **Bioinformatics**

*Cristian Delfabbro*  
*Simone Scalabrin*  
*Francesco Vezzi*  
*Alberto Policriti*  
*Alberto Stefan*  
*Alberto Casagrande*