Integrated, accurate and multi-environment structural variation discovery from whole genome sequencing data with NGSEP

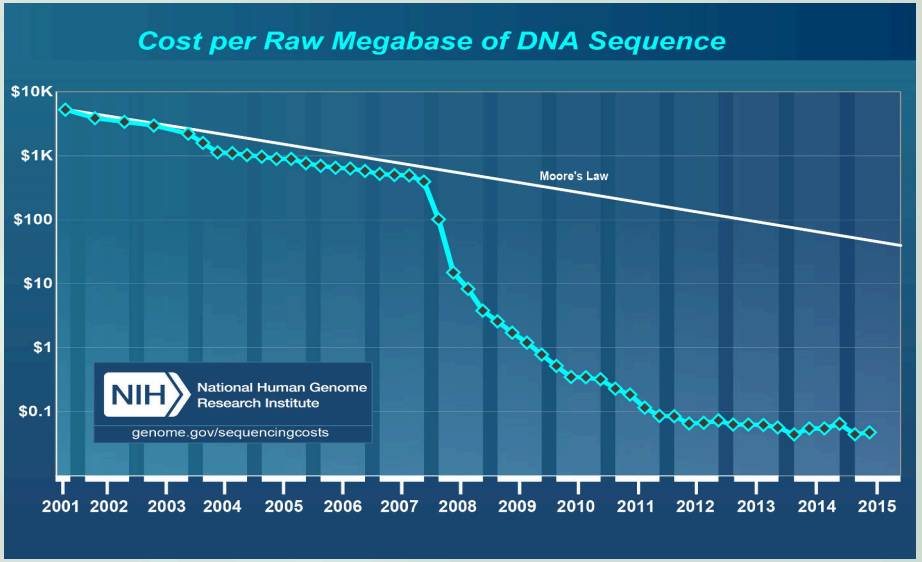
Juan Fernando de la Hoz, Jorge Duitama

Agrobiodiversity research area

International Center for Tropical Agriculture (CIAT)

Cali, Colombia

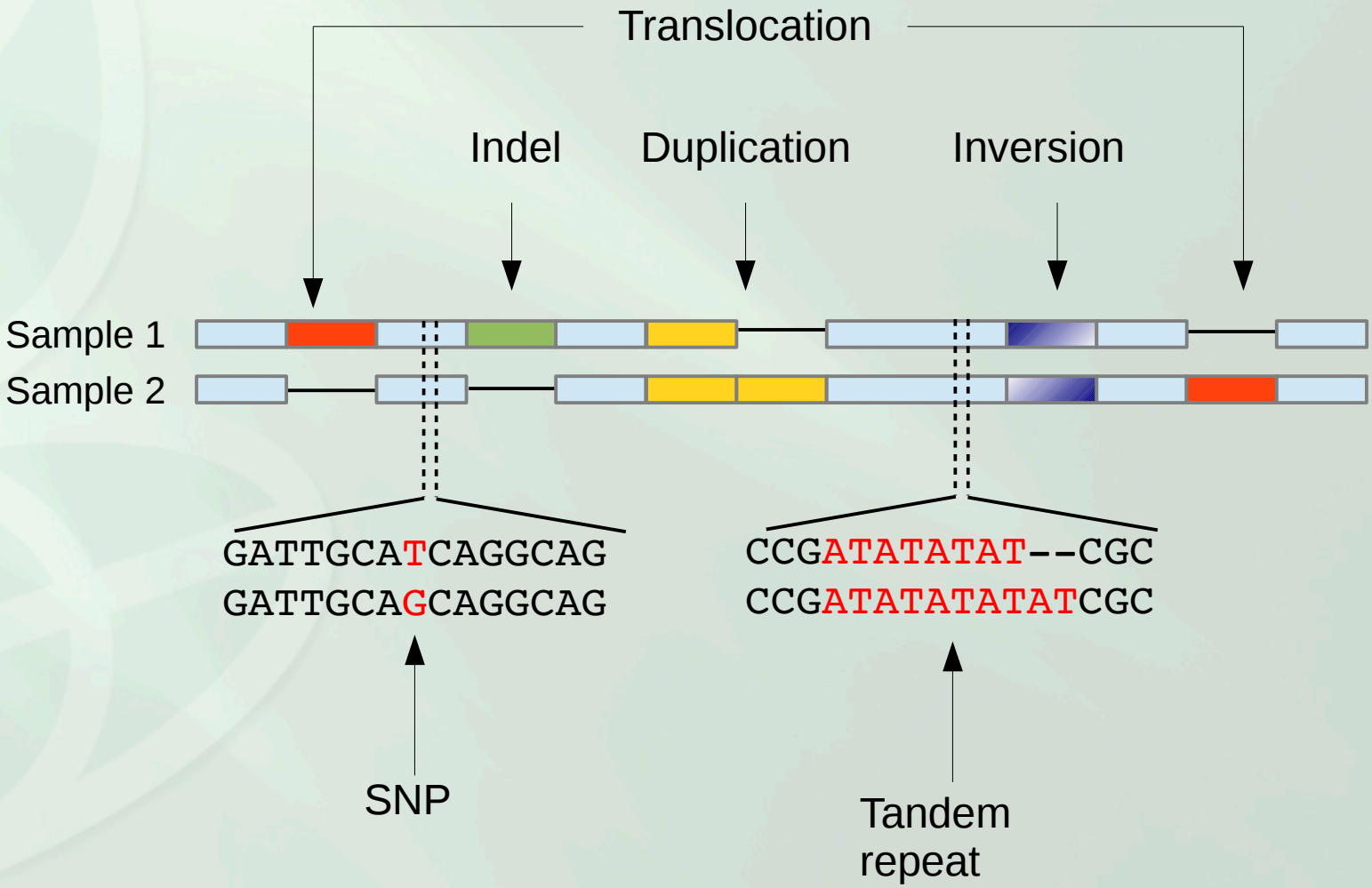
High Throughput Sequencing (HTS)



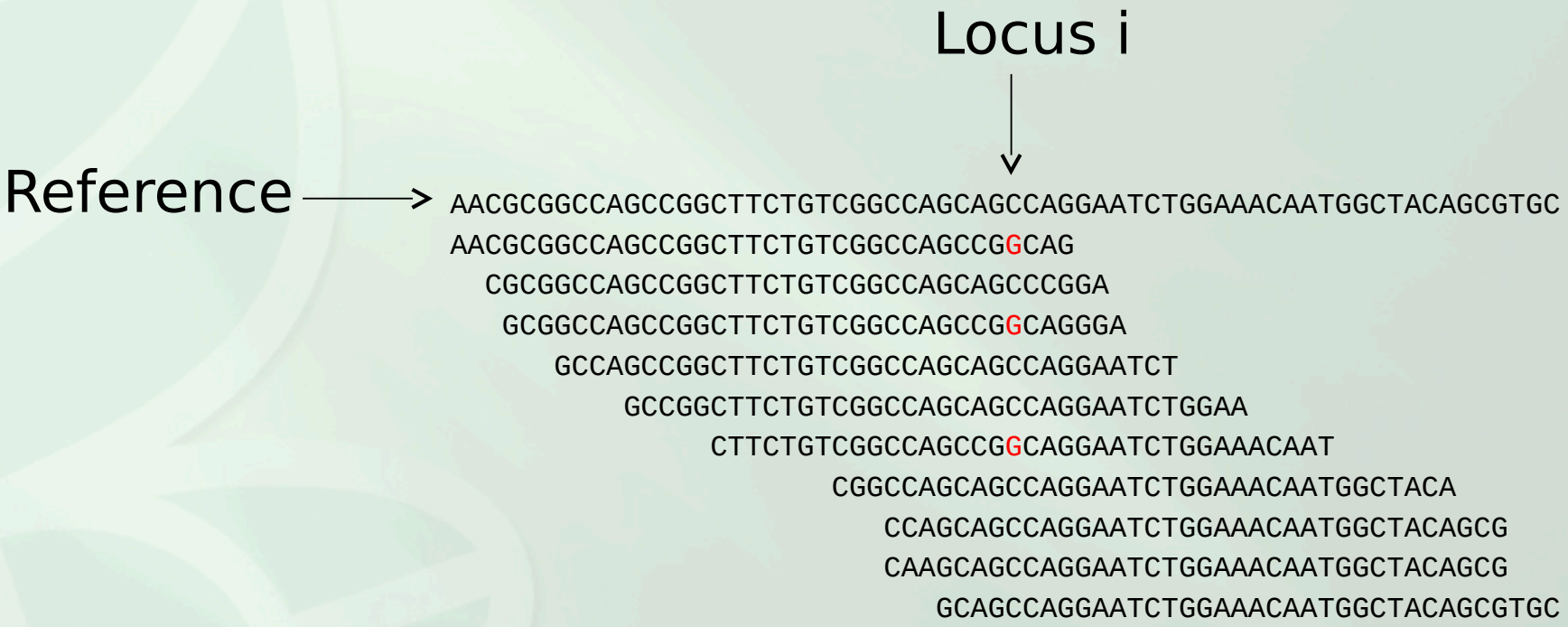
http://money.cnn.com/2013/06/25/technology/enterprise/low-cost-genome-sequencing/index.html

Species	Genome Size (Mbp)	# of sequenced samples with US$ 10K
Human	3200	4
Cassava	700	13
Bean	500	17
Rice	400	19

Genomic variability



SNP detection



R_i : Set of reads covering locus i

r(i) : Basepair from read r covering locus i

$\varepsilon_{r(i)}$: Error probability reading r(i)

G_i : Genotype at locus i

SNP detection

$$P(G_i|R_i) = \frac{P(R_i|G_i)P(G_i)}{P(R_i)}$$

Prior

$$P(G_i = H_iH'_i) = \begin{cases} \frac{1-h}{4} & \text{, if } H_i = H'_i \\ \frac{h}{6} & \text{, otherwise} \end{cases}$$

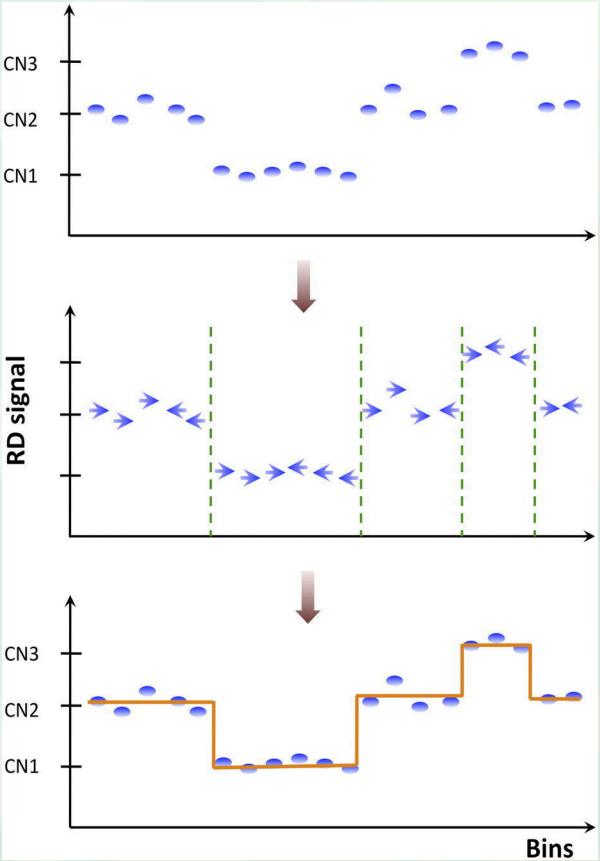
Conditional

$$P(R_i|G_i) = \prod_{r \in R_i} P(r|G_i)$$

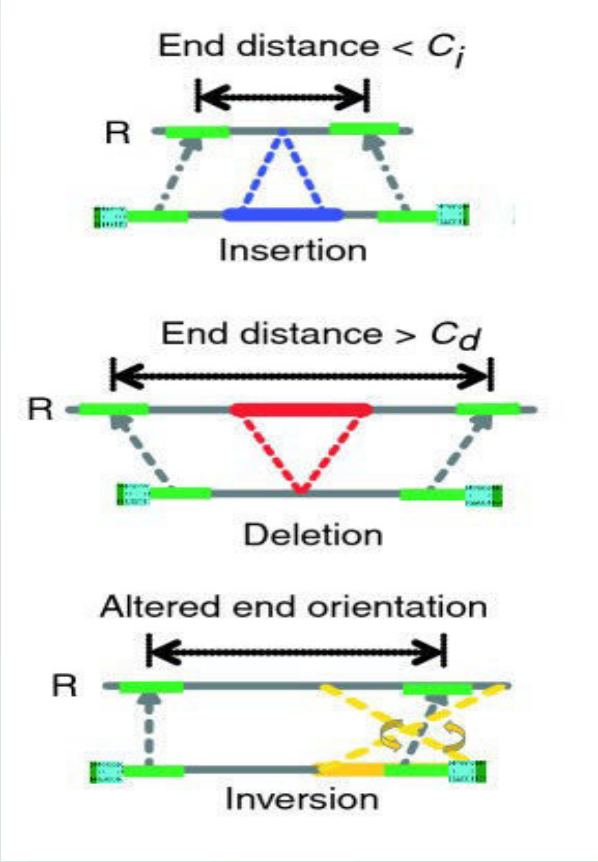
$$P(r|G_i = H_iH'_i) = \begin{cases} 1 - \varepsilon_{r(i)} & \text{, if } H_i = H'_i = r(i) \\ \frac{\varepsilon_{r(i)}}{3} & \text{, if } H_i \neq r(i) \wedge H'_i \neq r(i) \\ \frac{1}{2} - \frac{\varepsilon_{r(i)}}{3} & \text{, otherwise} \end{cases}$$

Structural variants detection

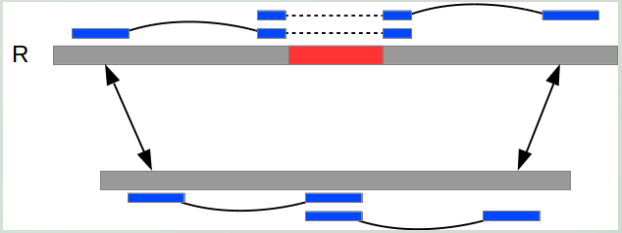
Read Depth (RD)



Read Pair (RP)



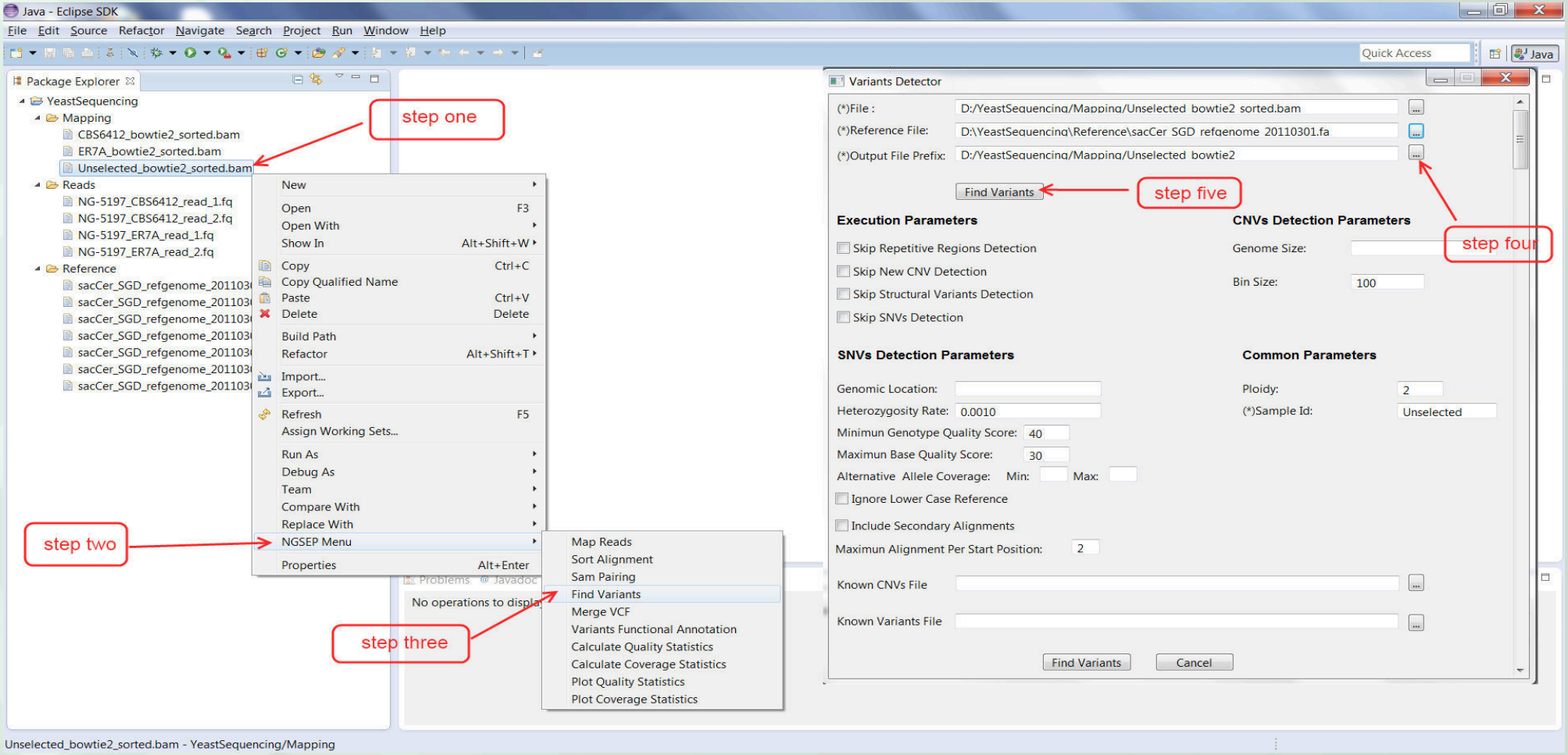
Split Read (SR)



Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research, 21(6), 974–84. doi:10.1101/gr.114876.110

Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., and Gerstein, M.B. (January, 2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome biology, 10(2), R23.

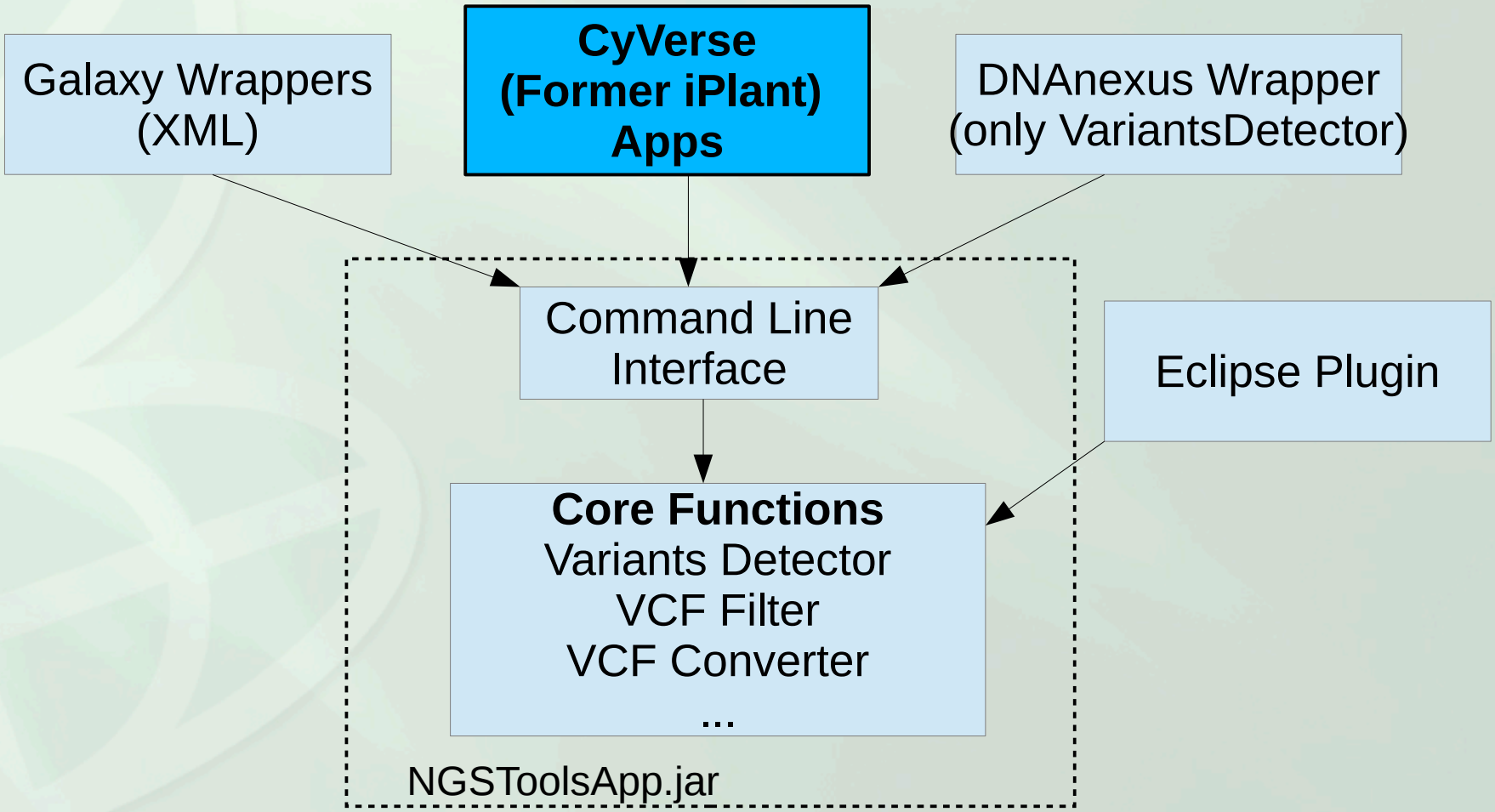
NGSEP



Available at http://www.sourceforge.net/projects/ngsep/

Duitama J, Quintero JC, Cruz DF, Quintero C, Hubmann G, Foulquie-Moreno MR, Verstrepen KJ, Thevelein JM, and Tohme J (2014) An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. Nucl. Acids Res., 42 (6): e44.; doi: 10.1093/nar/gkt1381. http://nar.oxfordjournals.org/content/42/6/e44.full

Using NGSEP



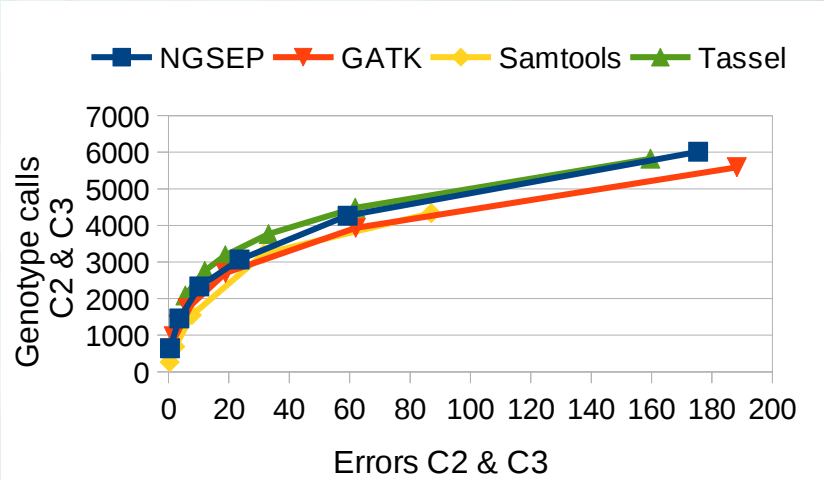
See webinar at https://www.youtube.com/watch?v=vJlZefQ1TKA

SNP calling QA

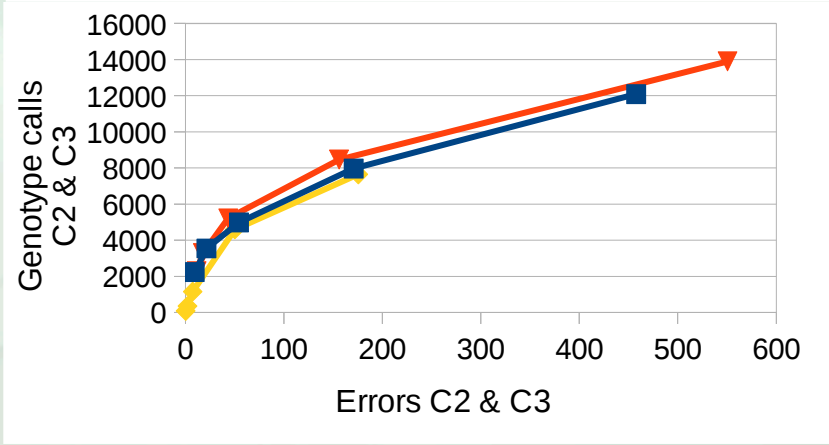
Cassava F1

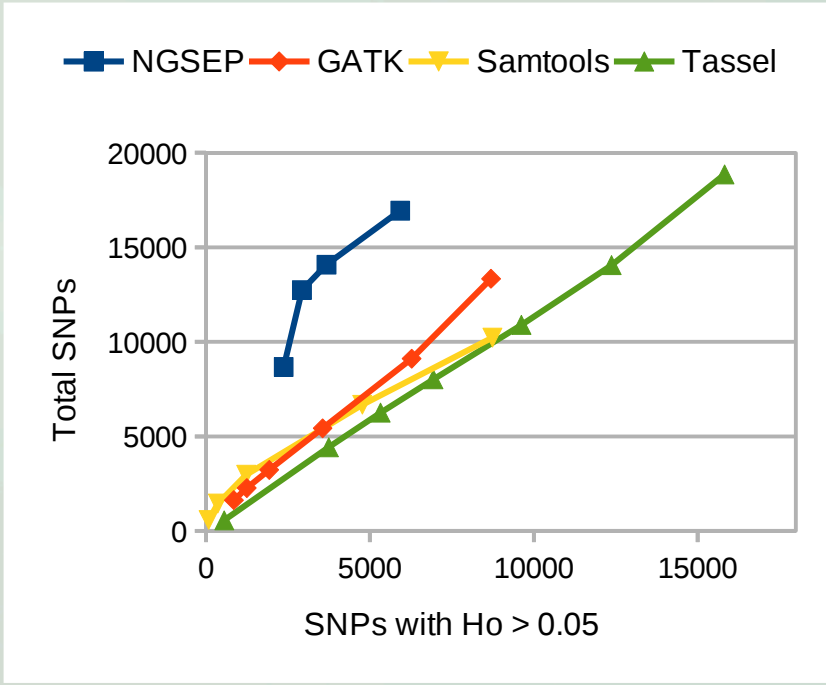
Bean MAGIC

K-family



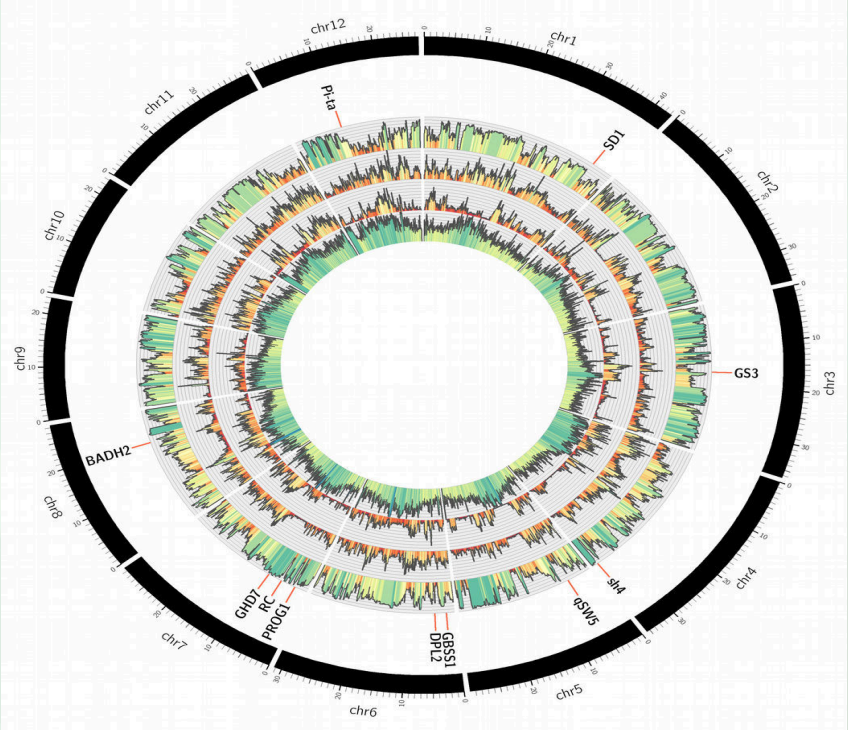
NxA

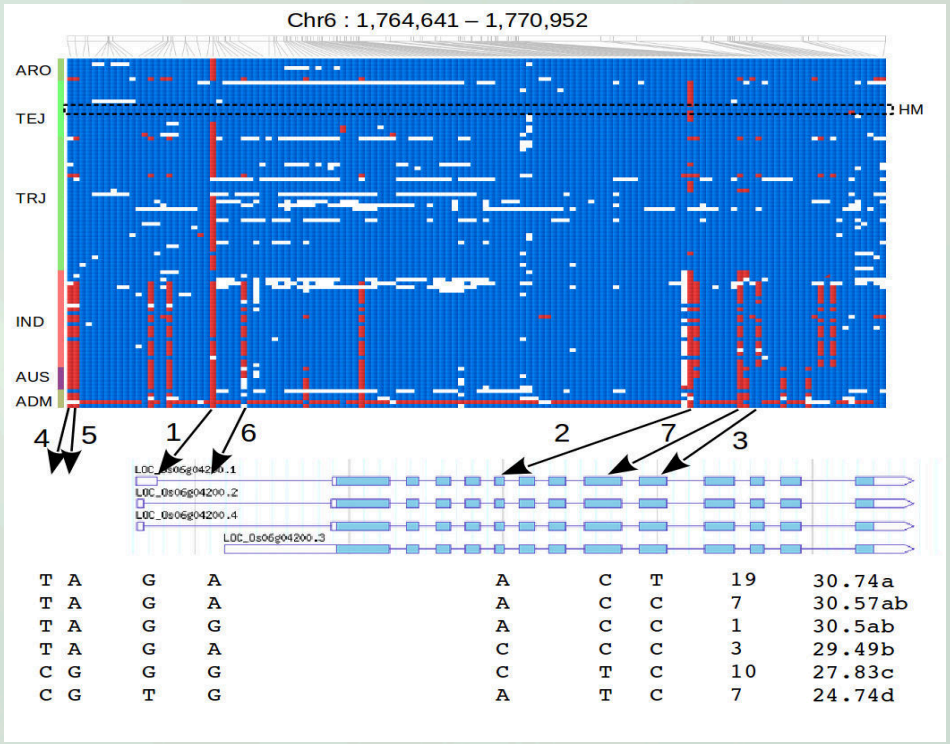




Visit poster P0362

WGS as a resource for MAS in rice





Duitama J, Silva A, Sanabria Y, Cruz DF, Quintero C, Ballen C, et al. (2015) Whole Genome Sequencing of Elite Rice Cultivars as a Comprehensive Information Resource for Marker Assisted Selection. *PLoS ONE* 10(4): e0124617. doi:10.1371/journal.pone.0124617

LSU AgCenter Research & Extension

USDA United States Department of Agriculture

NCGR National Center For Genome Resources

Yale

Attend IRIC workshop this Wednesday (W540)

NGSEP users

Institution	Species	Goal	Sequencing approach	DOI
University of Connecticut	Human, Mouse	Discovery of epitopes for immunotherapy	RNA-seq	10.1084/jem.20141308
KU Leuven	Yeast	Gene discovery with BSA	WGS	10.1186/1471-2164-15-207
University of Georgia	Peanut	SNP design for QTL analysis	WGS	10.1534/g3.115.018796
Universidad de Antioquia	Cryptosporidium, Trypanosoma	Assembly and Copy Number Variation	WGS	10.1038/srep16324
CIAT	Rice, Cassava Beans	Diversity, GWAS, SNP design for MAS	WGS, GBS, RAD-seq	10.1371/journal.pone.0124617
Universidad Nacional	Potato	GWAS on diversity panel	GBS	
Universidad de los Andes	Phytophthora	Domestication / Diversity	GBS	
Cenicaña	Sugar cane	Expressed variants	RNA-seq	

NGSEP Users

NGSEP

Summary | Files | Reviews | Support | Wiki | Tickets | Discussion | Blog | Code

Brought to you by: dfcruz00, jduitama, juandelahoz

Home (Change File)

Date Range: 2015-01-01 to 2015-12-31

DOWNLOADS
1,609
In the selected date range

TOP COUNTRY
United States
36% of downloaders

TOP OS
Windows
34% of downloaders

OS downloads as: Percent

	Country	Android	Linux	Macintosh	Unknown	Windows	Total
1.	United States	0%	6%	23%	48%	23%	583
2.	Colombia	1%	40%	10%	1%	48%	465
3.	Germany	0%	0%	0%	92%	8%	90
4.	Spain	0%	0%	0%	100%	0%	78
5.	Belgium	0%	14%	73%	0%	12%	64
6.	Brazil	0%	48%	0%	2%	50%	46
7.	India	0%	24%	0%	0%	76%	34
8.	Netherlands	0%	41%	0%	56%	3%	32
9.	Philippines	0%	21%	29%	0%	50%	24
10.	Italy	0%	5%	27%	0%	68%	22
11.	Switzerland	0%	23%	23%	45%	9%	22
12.	Malaysia	0%	0%	0%	9%	91%	22

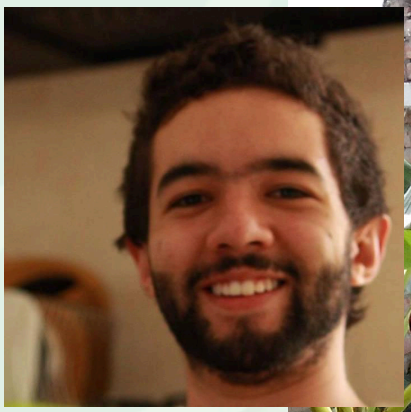
Acknowledgements

Alexandra Peña, Maria Camila Rebolledo, Fabian Barco, Dolly Gomez, Edgar Torres, Bodo Raatz, Steve Beebe, Joe Tohme

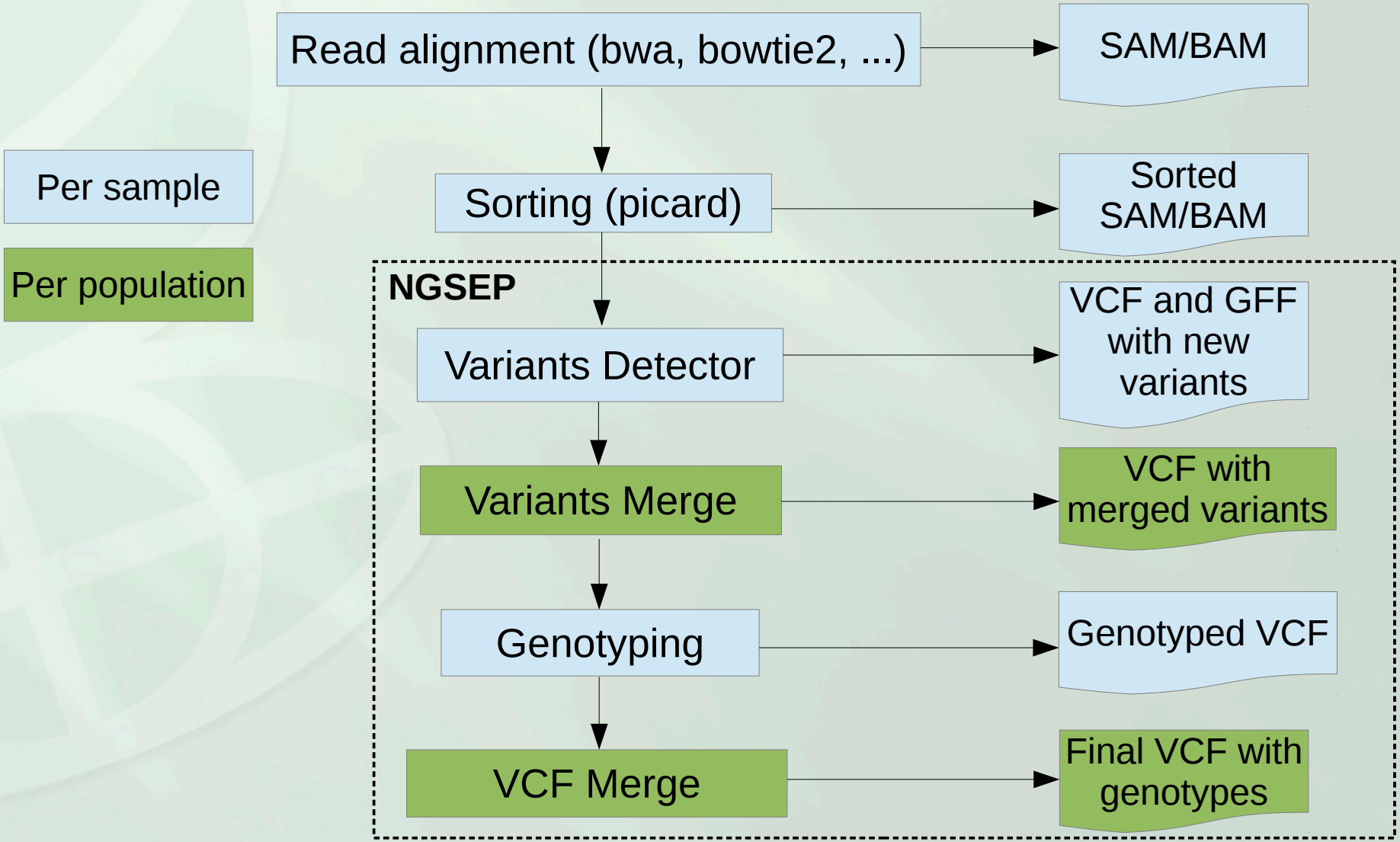
Martha Narro, Andre Mercer, Sriram Srinivasan, Paul Sarando, Kapeel Chougule, Ramona Walls, Marcela Monaco

Funding

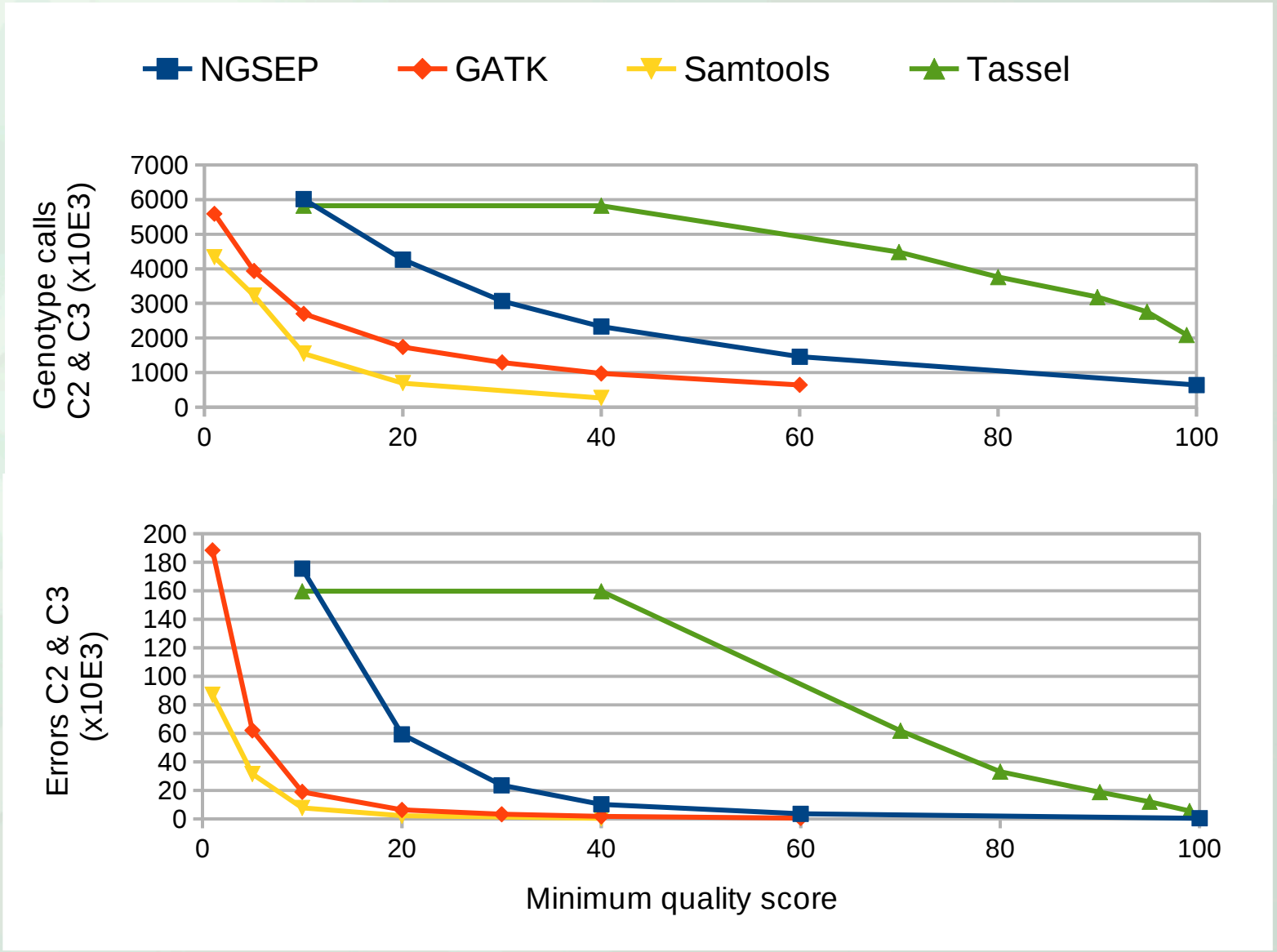
Thank you



NGSEP pipeline



Comparisons per quality score



Ho and MAF distributions