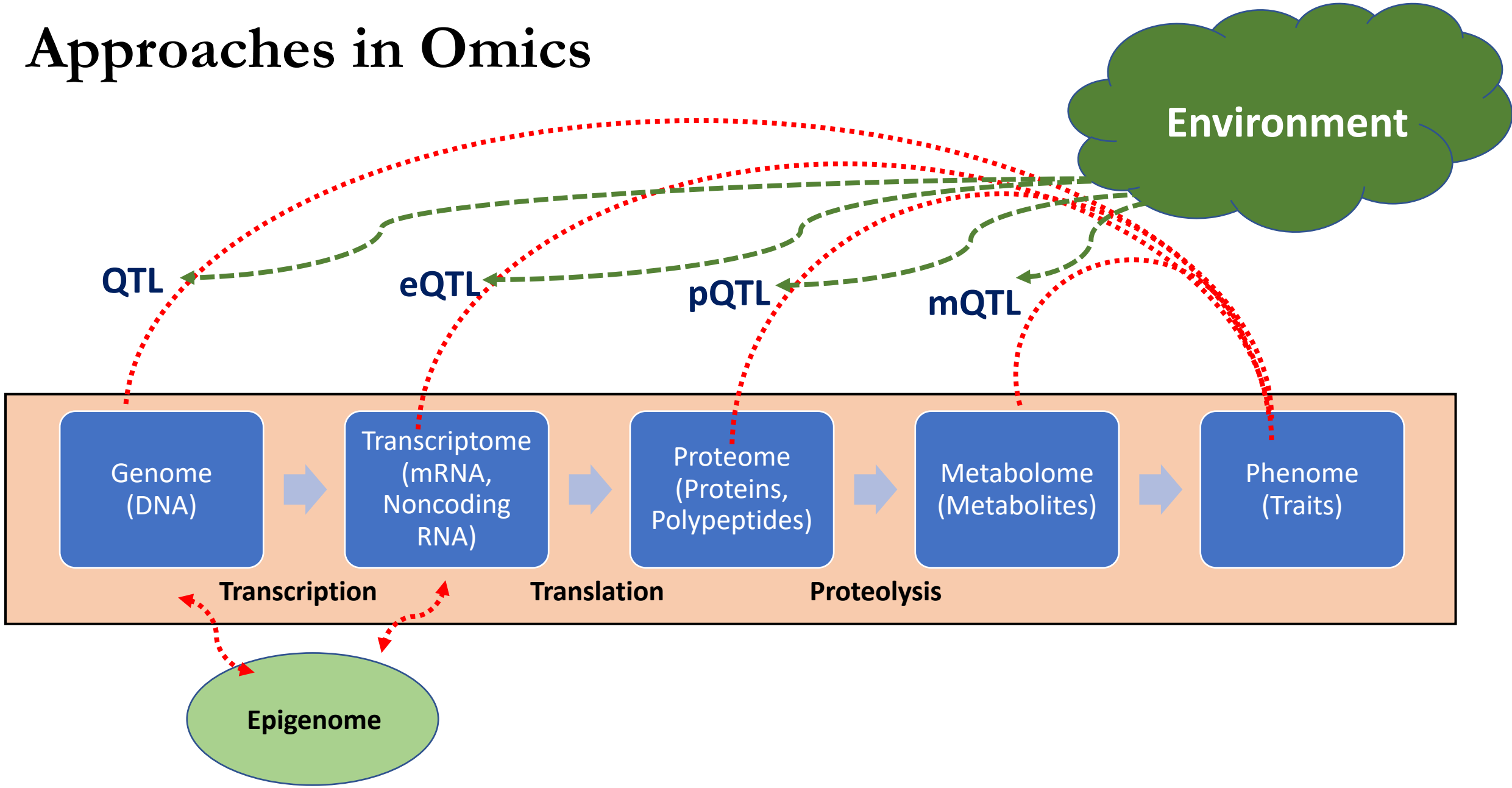
Long Read Sequencing for Gene Prediction and Annotation

SUJAN MAMIDI Ph.D

Approaches in Omics



Sequencing

- Sequencing is the process of determining the nucleic acid sequence, the order of nucleotides in DNA and RNA.

Ex: ATAGTGAGCCGATTATATATGACGCGCTA

Short reads : 75 – 150 bp

Paired end Reads: 150-300bp

Long reads: 1Kb - 2Mbp

History of sequencing

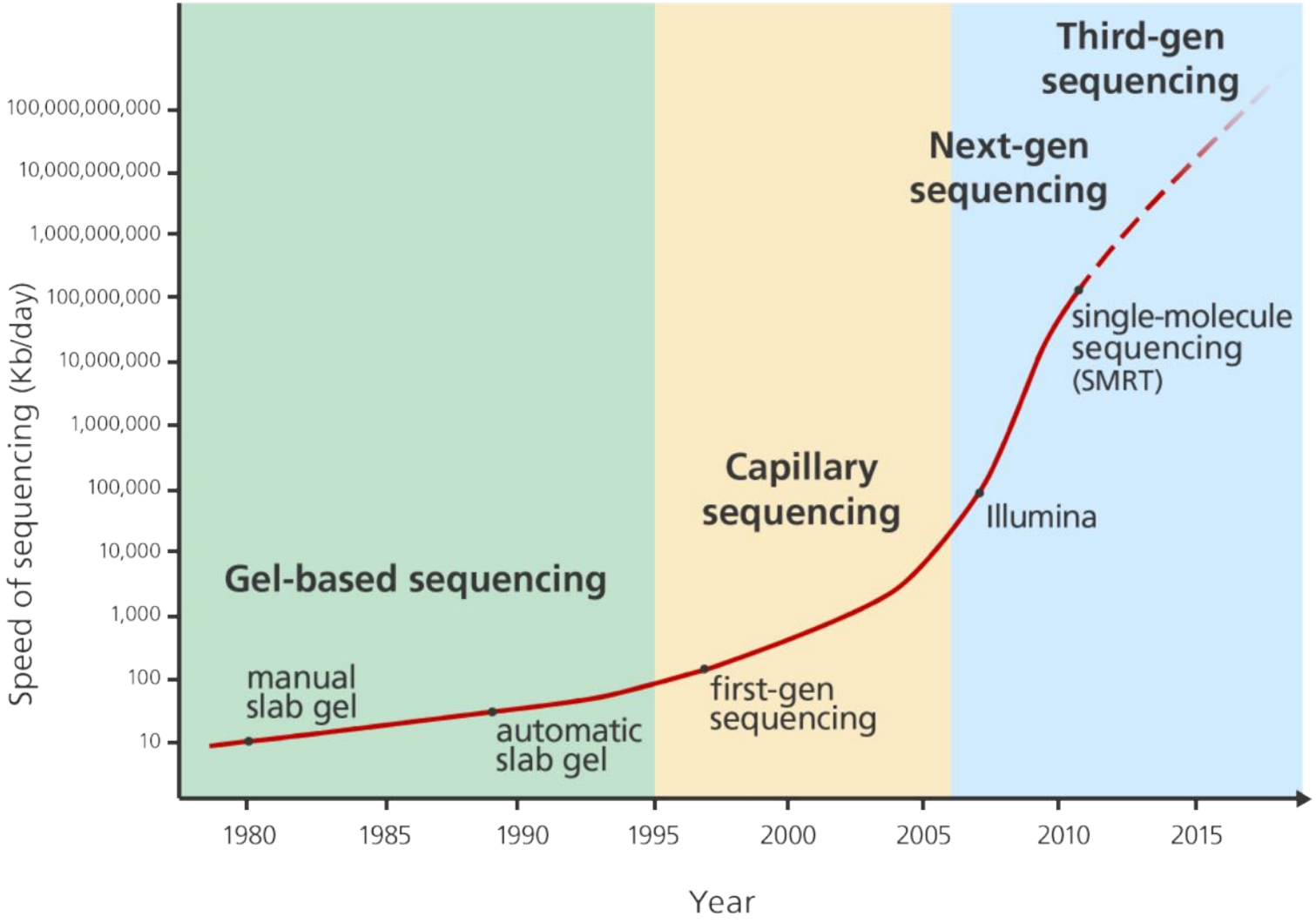


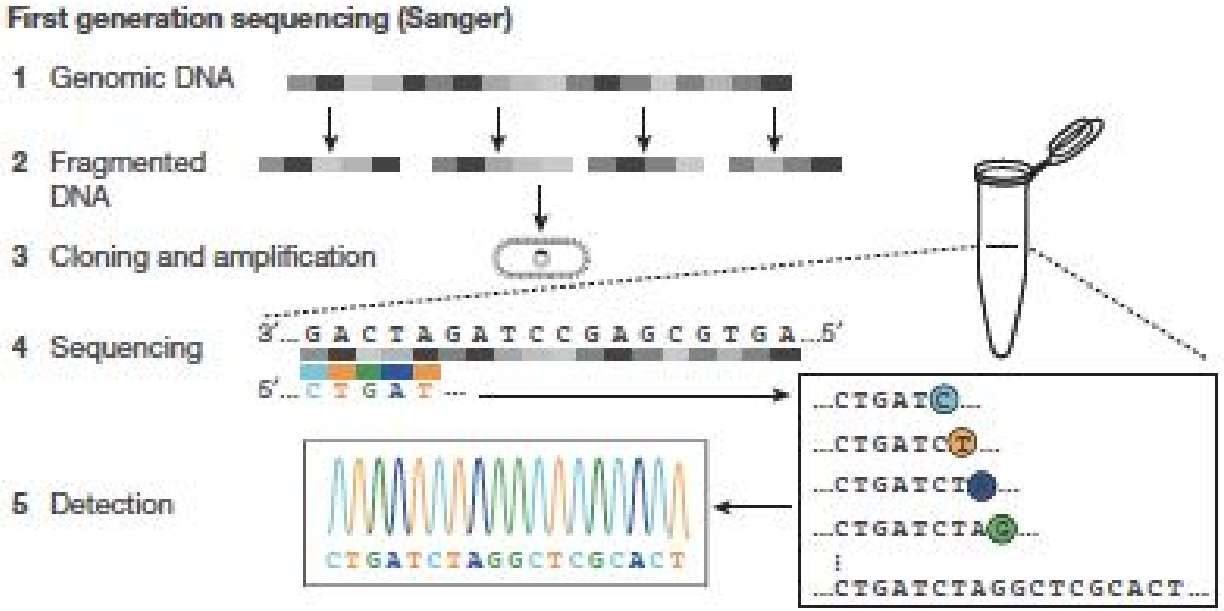
Image credit: Genome Research Limited, UK

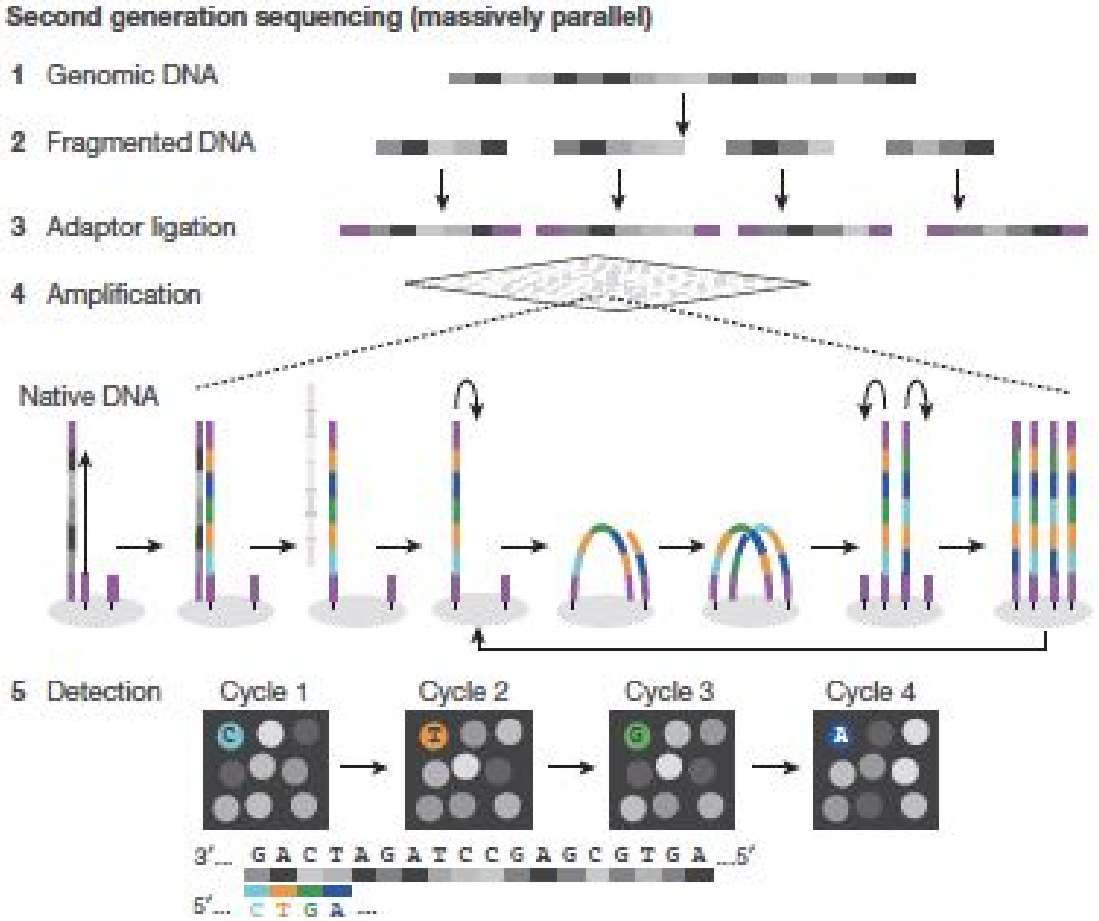
Cost of Sequencing

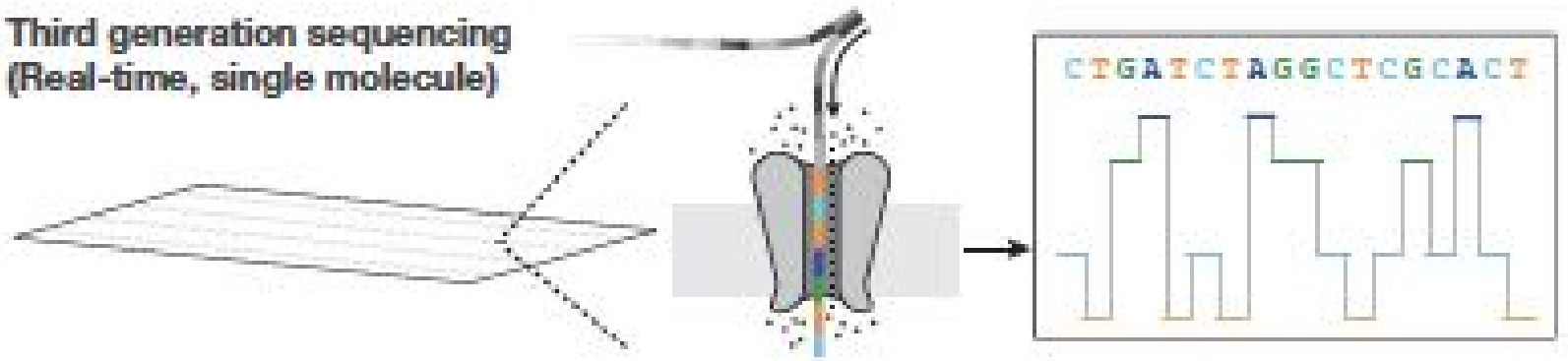
Human Genome – 3.1 Gbp

$3,000,000,000	2003 Human Genome Project	
$20,000,000	2006 1st individual genome	
$2,000,000	2007 1st NGS Genome	
$200,000	2008 1st 30x genome	
$10,000	2010 1st sub-10K genome	
$1,000	2014 1st $1,000 genome	
$100	2017 1st $100 genome	

https://blog.dnanexus.com/2017-04-24-dnanexus-celebrates-national-dna-day/







Shendure et al. 2017

Illumina - Present day Sequencers

	iSeq100	MiniSeq	MiSeq	NextSeq 550	NextSeq2000	Novaseq6000
Run Time (hrs)	9.5-19	4-24	4-55	12-30	24-48	13-44
Max Ouput (GB)	1.2	7.5	15	120	300	6000
Max Reads / Run	4 Million	25 Million	25 Million	400 Million	1 billion	20 billion
Max Read Length	2 X 150bp	2 X 150bp	2 X 300 bp	2 X 150	2 X 150 bp	2 X 250 bp

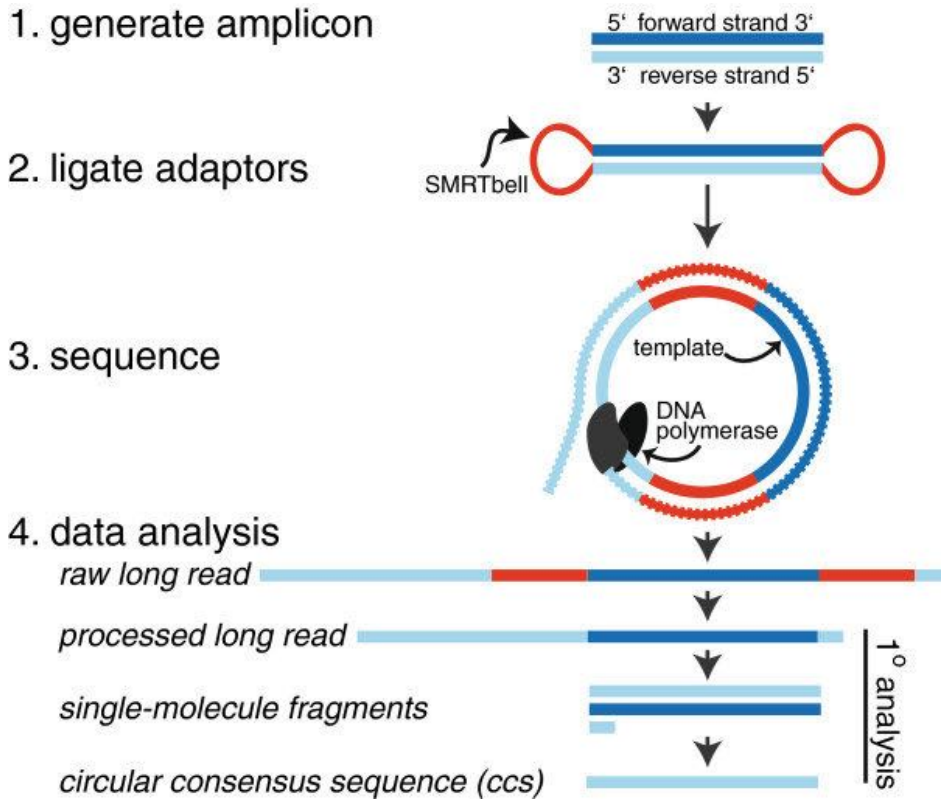
Ion torrent




	PGM	Ion Proton	Ion S5	Ion S5 XL
Read Length	35 - 400bp	Upto 200bp	Upto 400 bp	Upto 400bp
Reads/run (based on chip)	500k – 5M	60M - 80M	2M – 80 M	2M – 80M
Run time	2 - 8 hrs	2 - 4 hrs	2 - 4 hrs	2 - 4 hrs
Throughput	2 Gbp	10 Gbp	Upto 15Gbp	Upto 15Gbp

Pacbio Sequel 2

- Up to 20 Gb per SMRT Cell 1M
- Average read lengths up to 30 kb
- High consensus accuracies (>99.999%)



Fichot & Normal 2013

Oxford nanopore

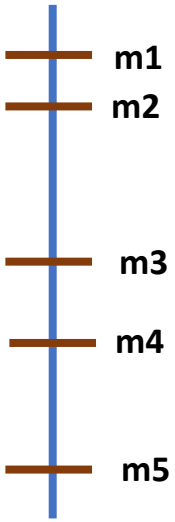
- Read Length – Up to 2Mbp
- Yield / Device: 2GB – 8TB

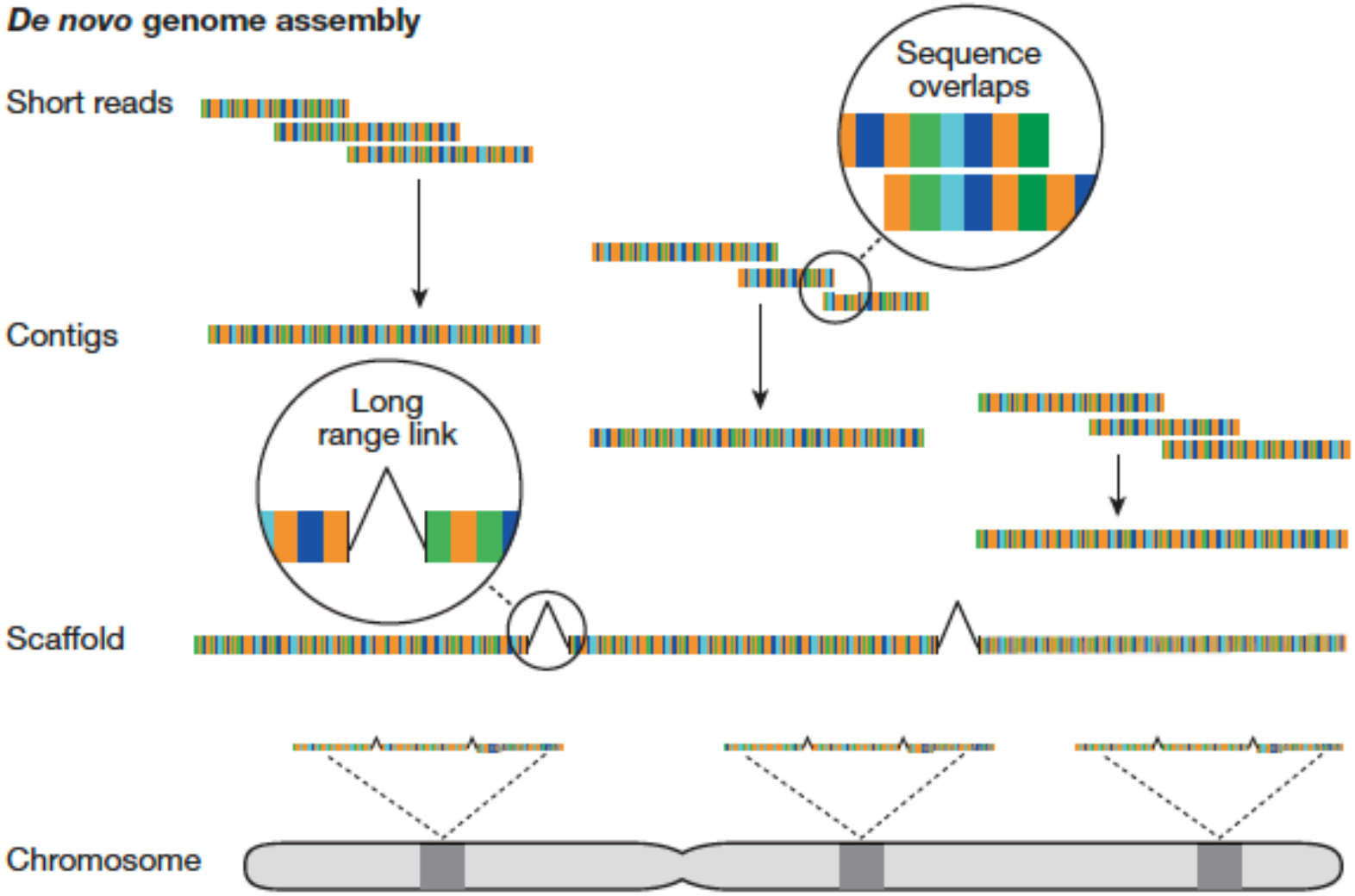
GENOME ASSEMBLY

Requirements:

- Genetic map
- Mate pairs or BACs
- High quality and high coverage of reads

Genetic map





Shendure et al. 2017

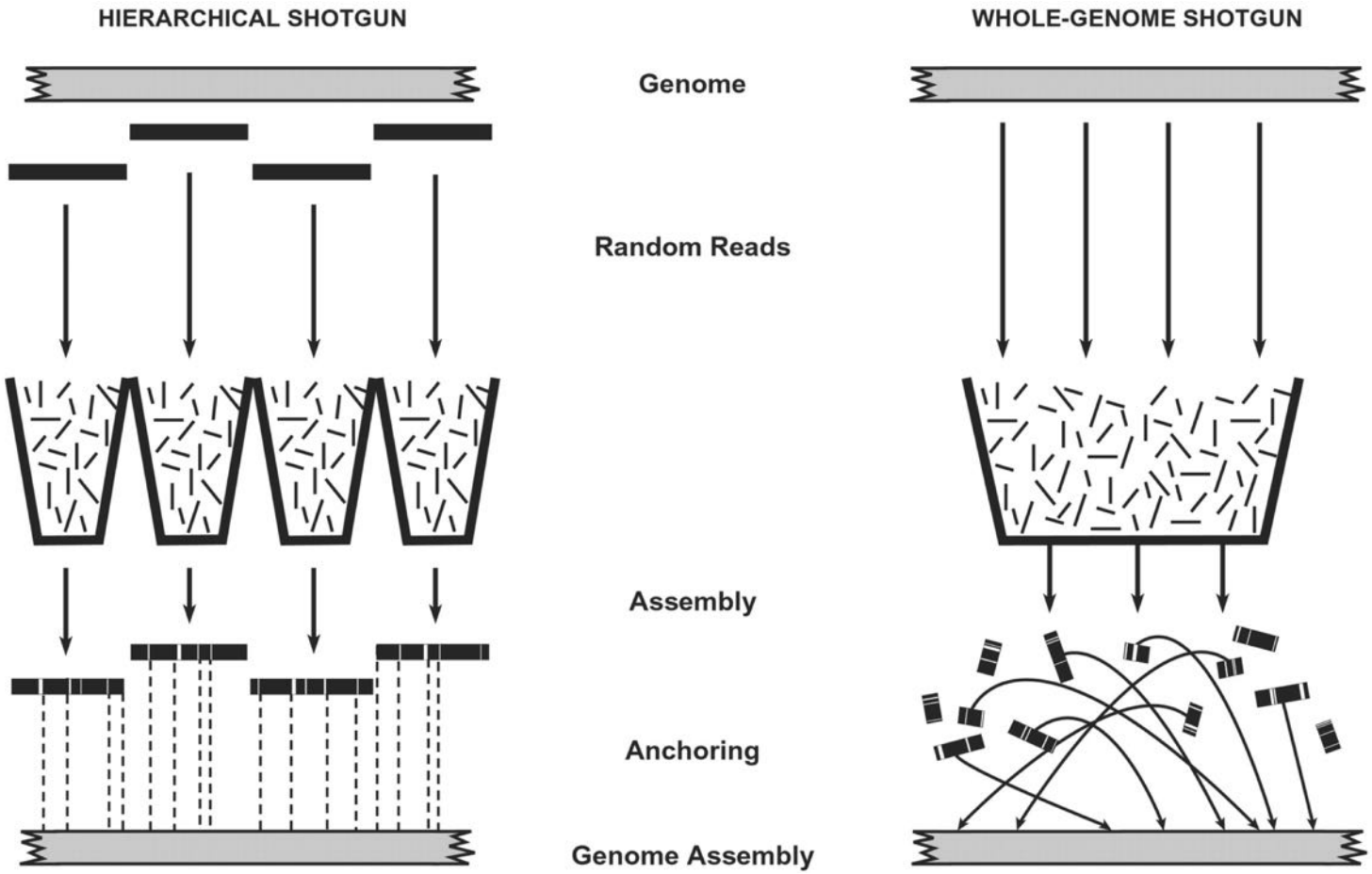
Human Genome (3.1 billion bases)

Public Human Genome Project

BACs

23 billon bases

Private Human Genome Project (Celera)



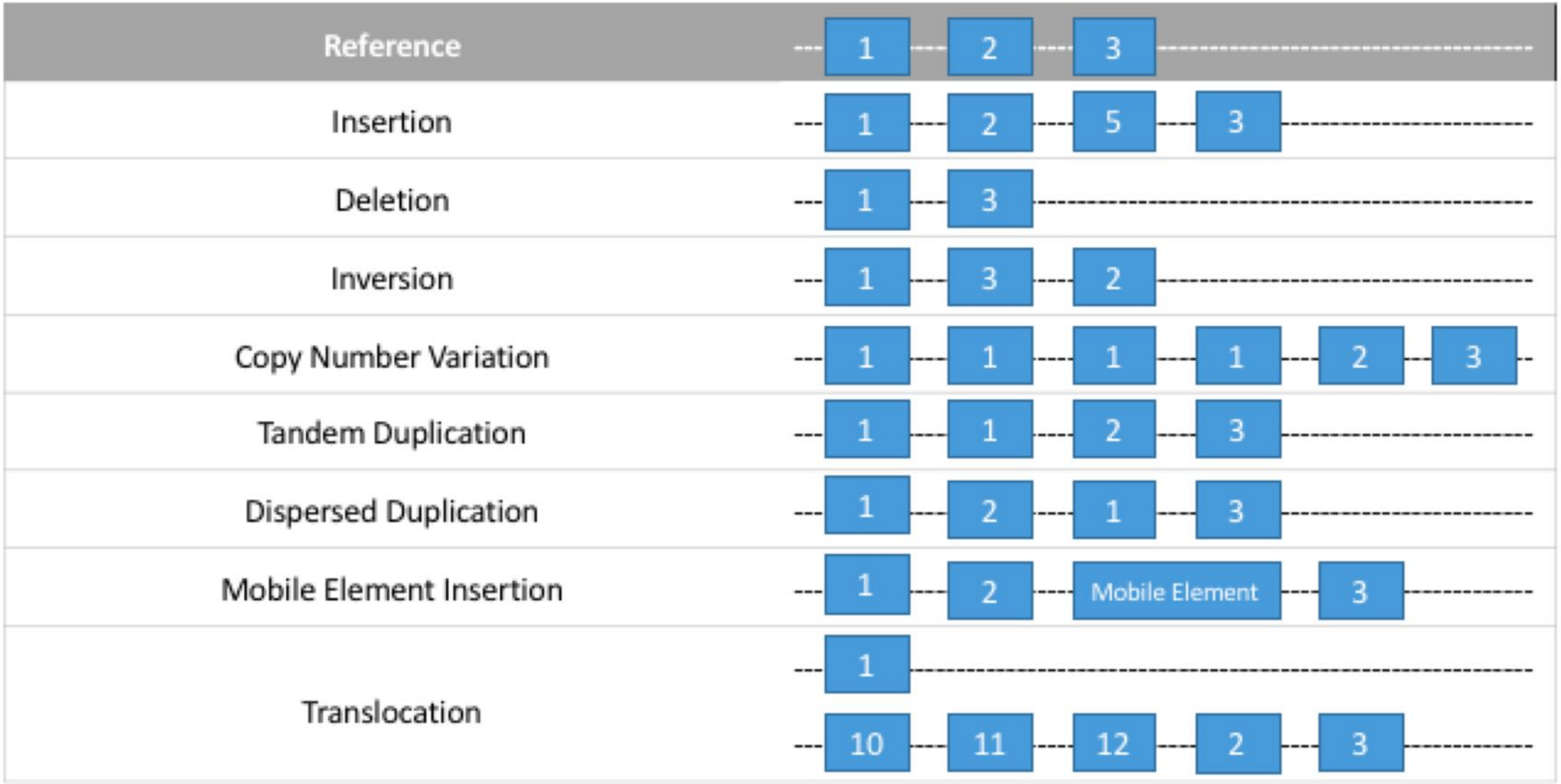
- 27.2 million clones
- 14.8 billion bases

Waterston et al. 2002. PNAS

Challenges with short Read assembly

- Gaps in sequence
- Unique overlaps between pairs of reads are much less likely
- Hard to resolve repeats
- Lack of decent genetic map
- Hard to resolve polyploids, highly heterozygous species
- Structural Variants
- Assembly may contain inverted fragments
- Phasing is difficult

Structural Variants



Wikipedia

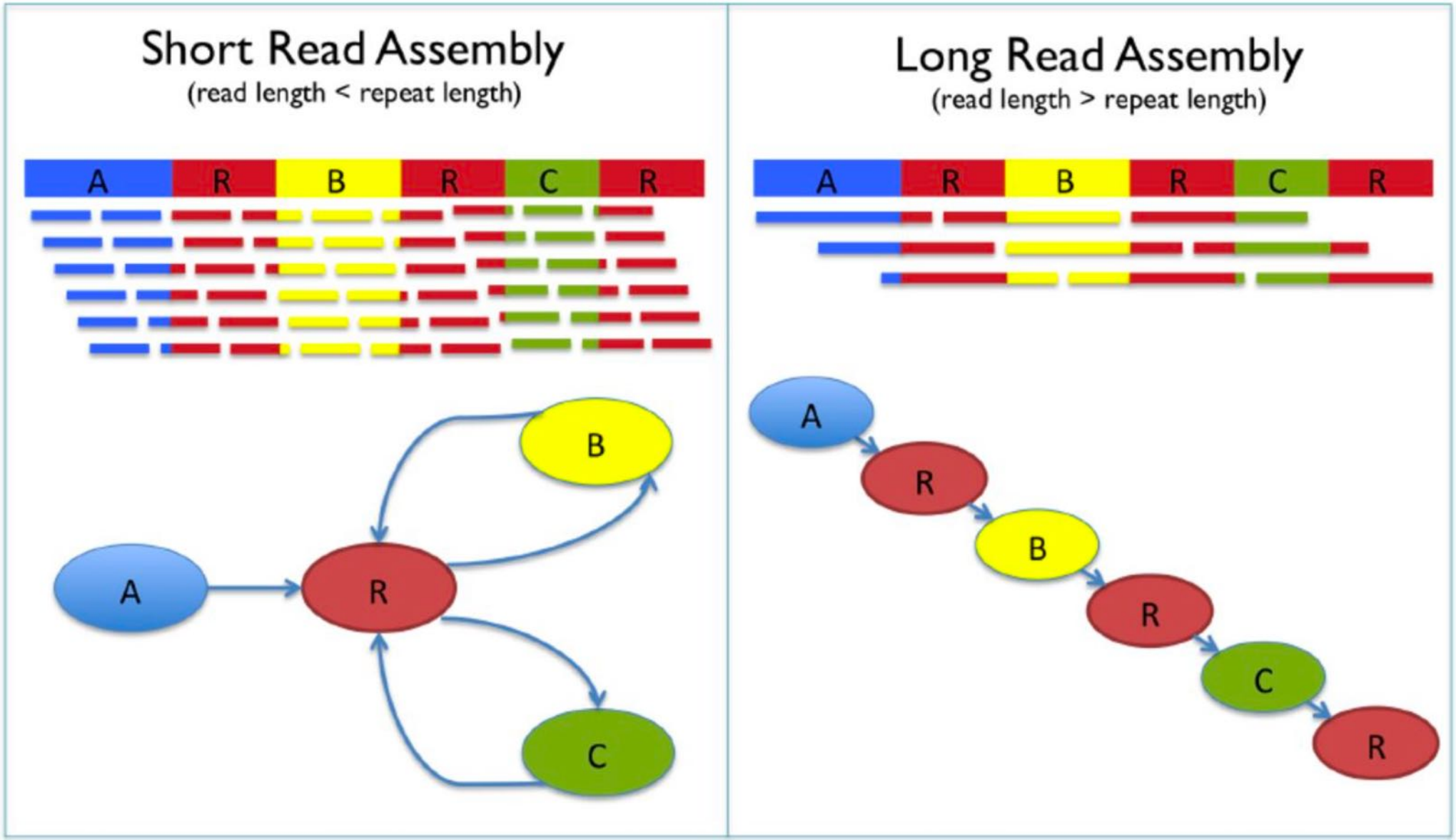
Long read sequencing

Benefits

- Better genome assembly
  - Little issues with repeats, tandem duplicates
- Structural variants resolved
- Haplotype Phasing is easy
- Speed of assembly
- Mapping certainty

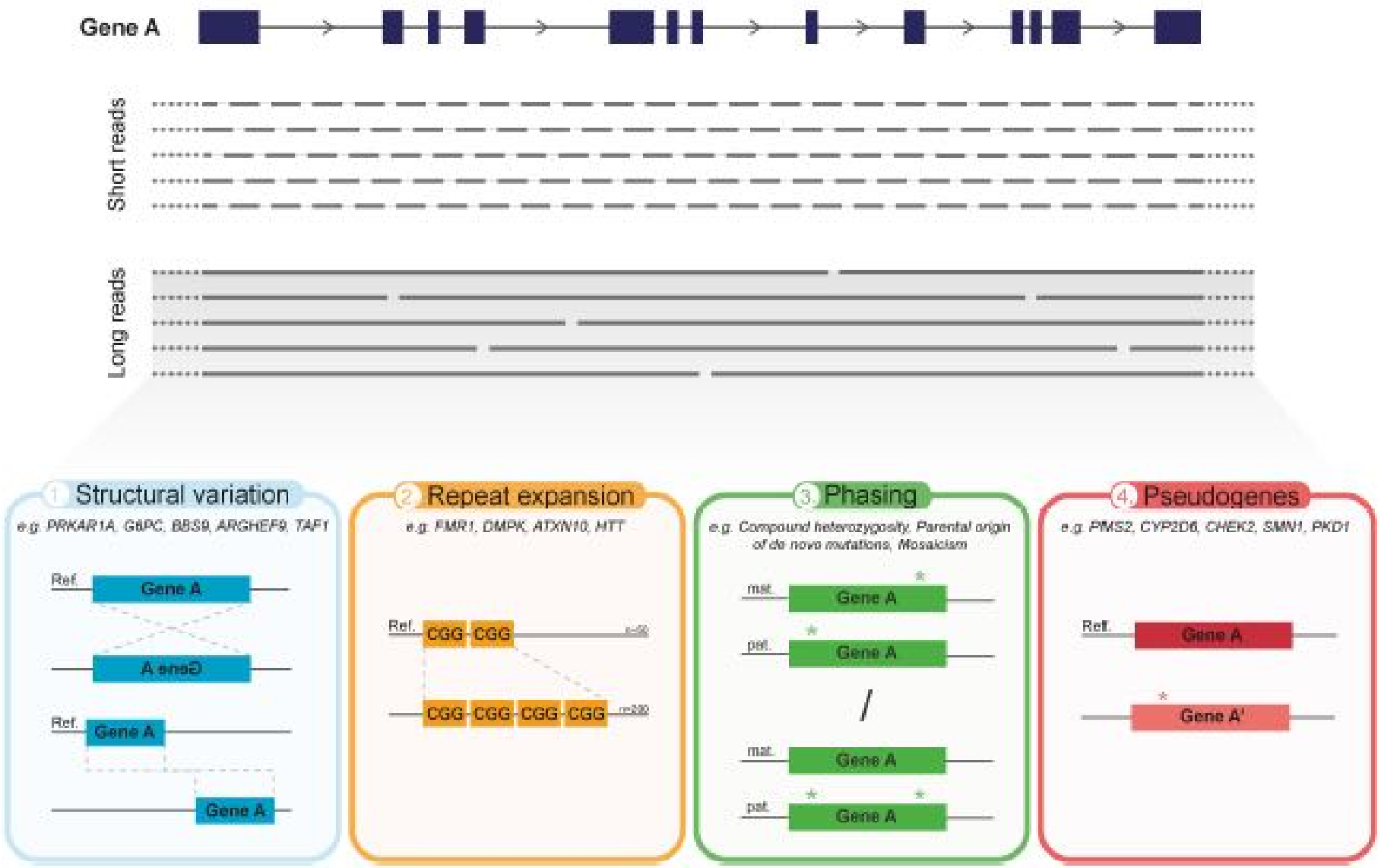
Challenge – Hard to obtain high quality long stranded DNA

Repeats



Lee et al. Biorxiv

Advantages of long read Sequencing

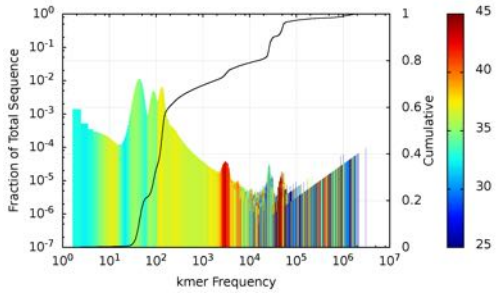


Mantere et al. 2019

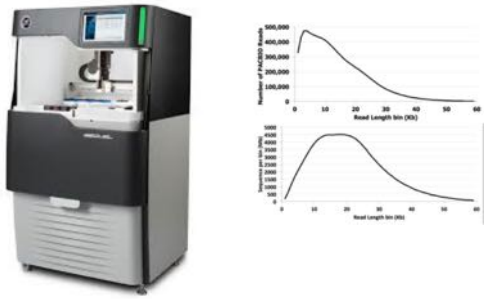
Sequencing Genomes with PACBIO

1. Get good DNA (50-150kb)

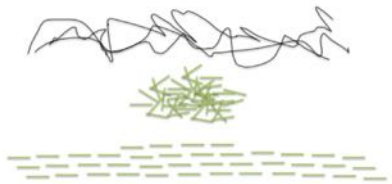
2. Assess the genome: Ploidy, het rate, repeat, organelle, contamination



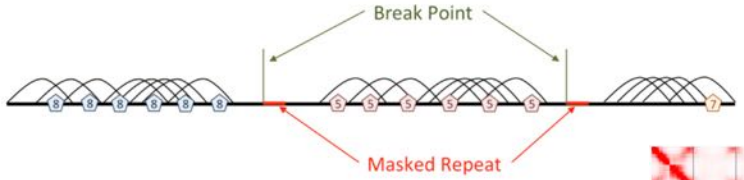
3. Generate long, high quality PACBIO libraries and collect data, either 40-50x or 70-80x, + outbreds



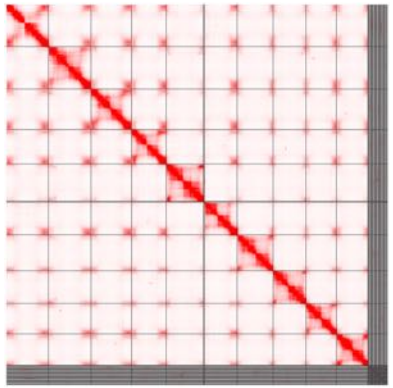
4. Assemble the genome (MECAT, Canu) arrow polish



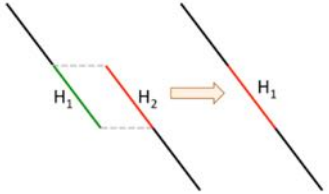
5. Break false joins



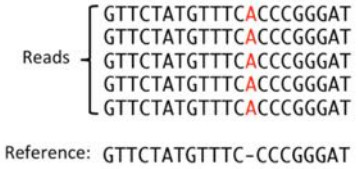
6. Integrate into chromosomes using HIC data.



7. Address haplotype & assembly overlap



8. Final polish with Illumina to reduce homozygous errors



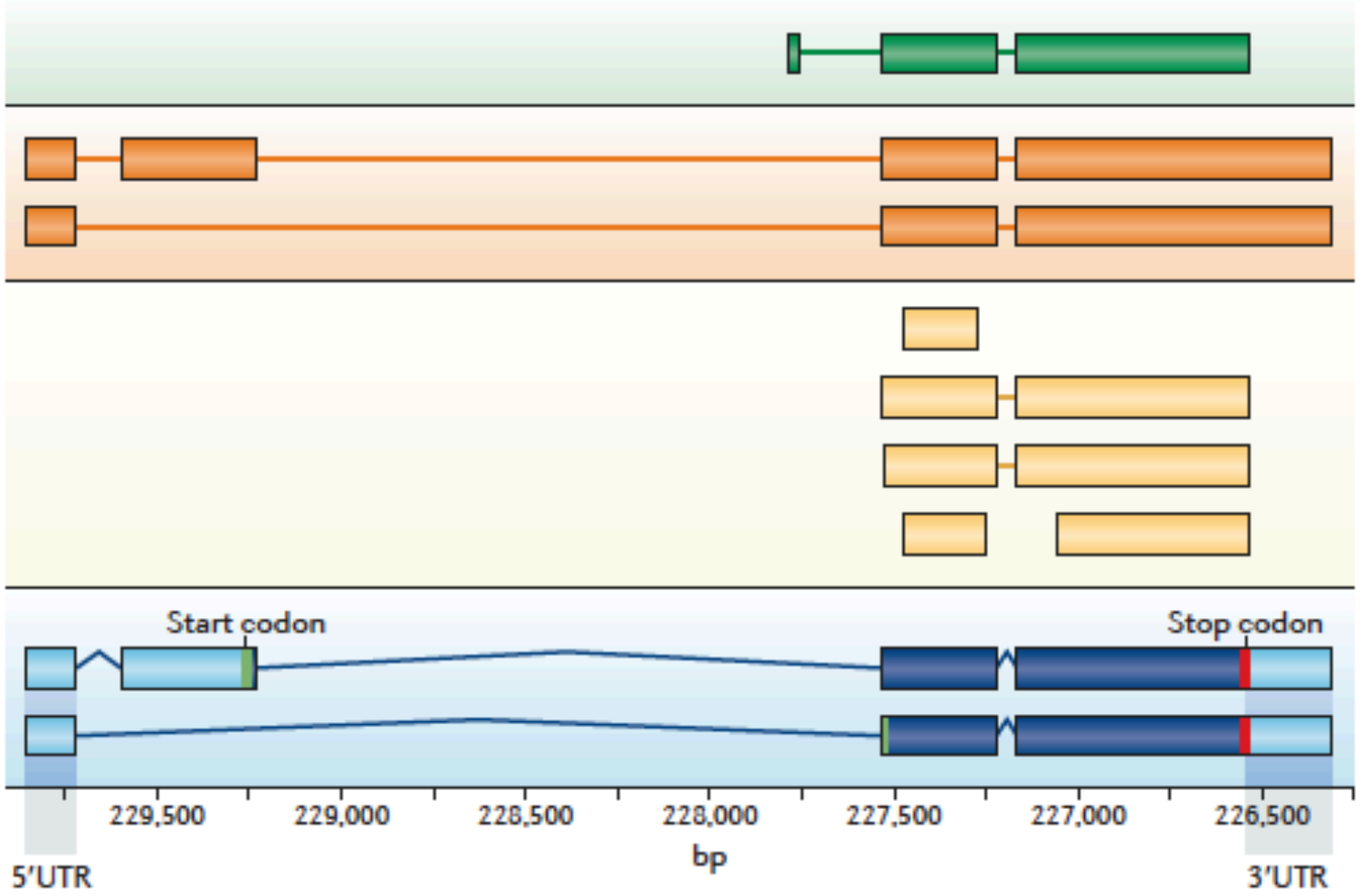
Jerry Jenkins & Chris Plott @ HAGSC

Gene Prediction and Annotation

Building a Gene Model

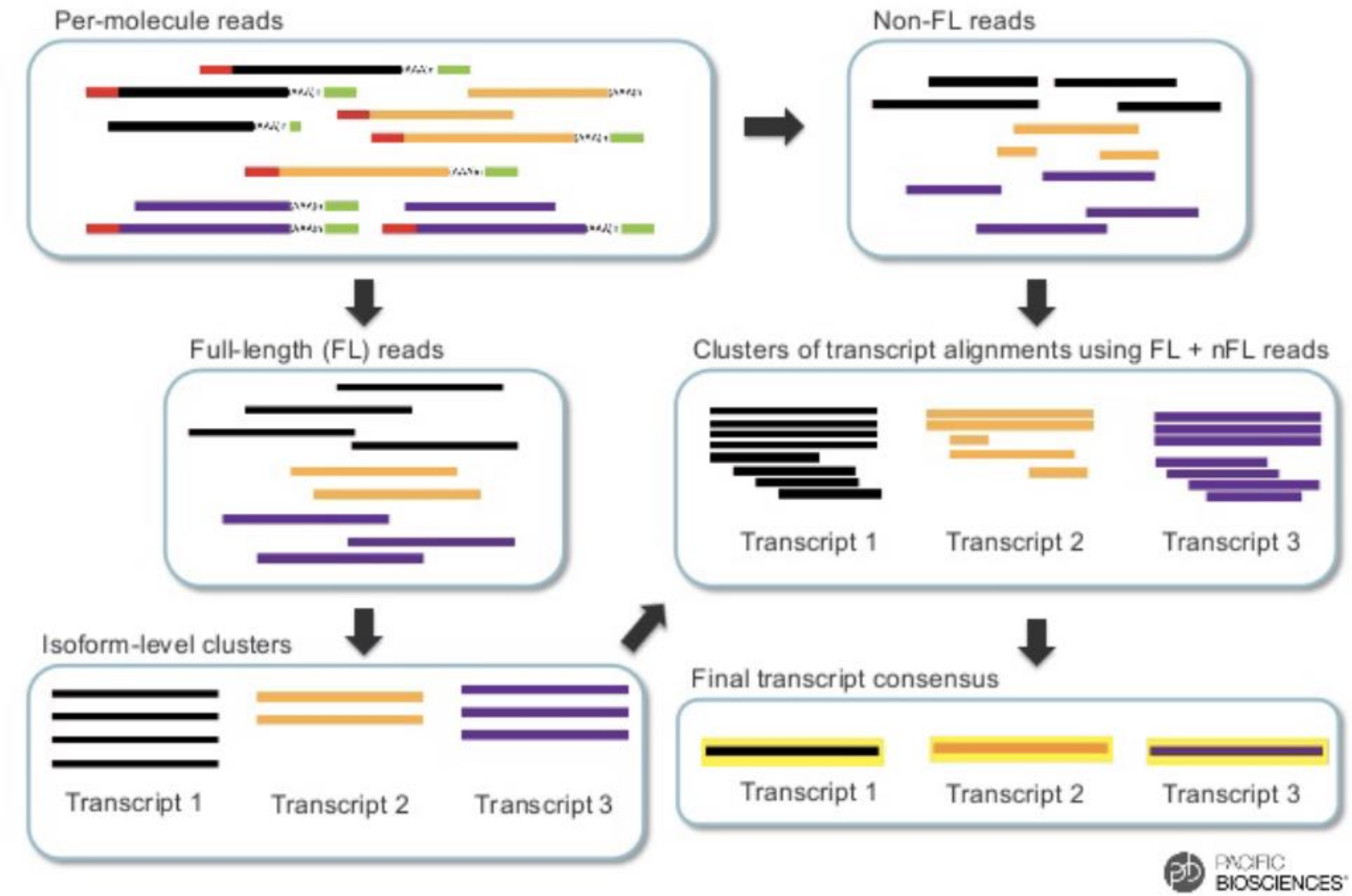
Considerations

- N50 should be large (> gene lengths)
- Low percent gaps
- Percent coverage (more)



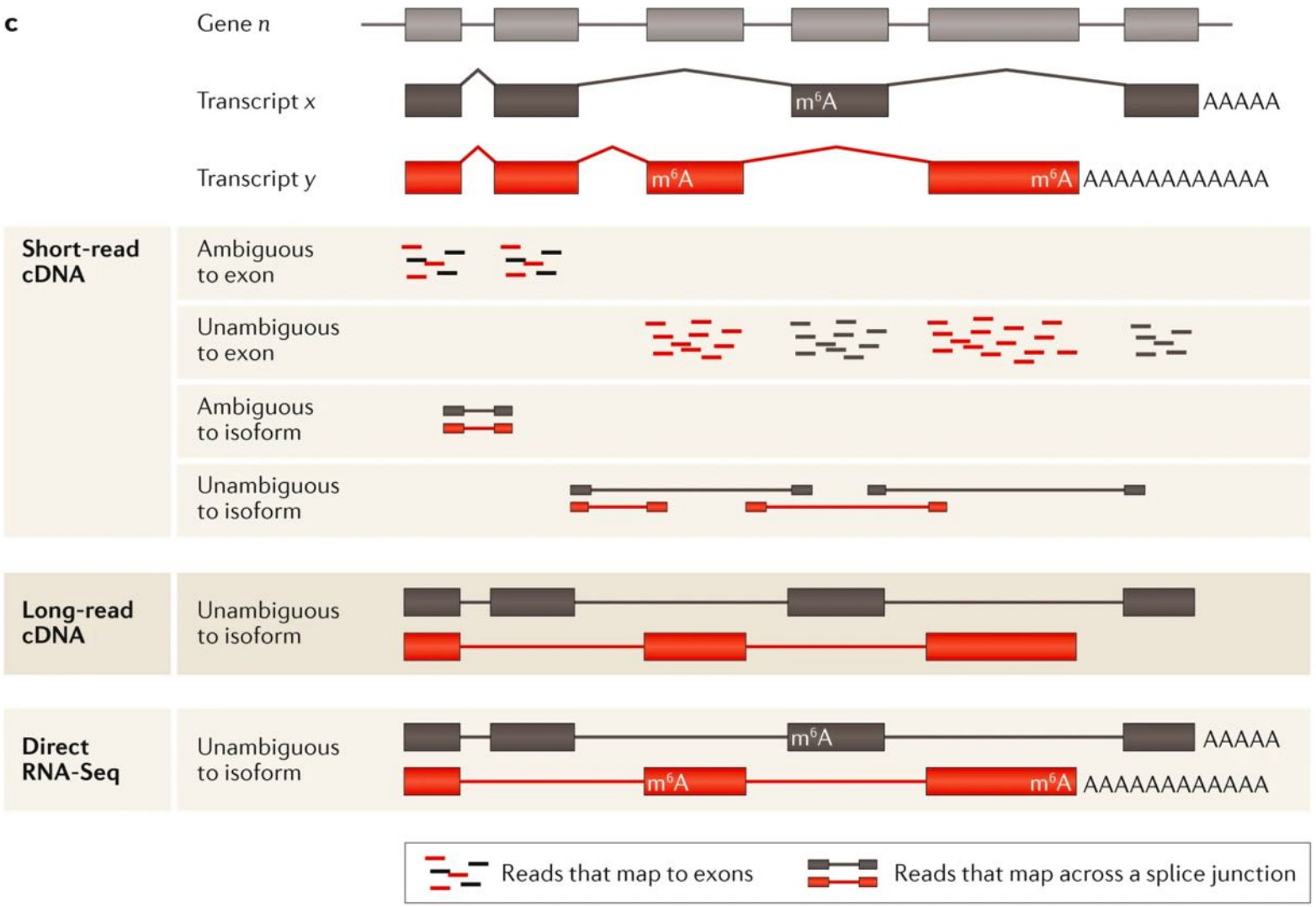
Yandell and Ence ,2012

PacBio Isoseq Informatics Pipeline



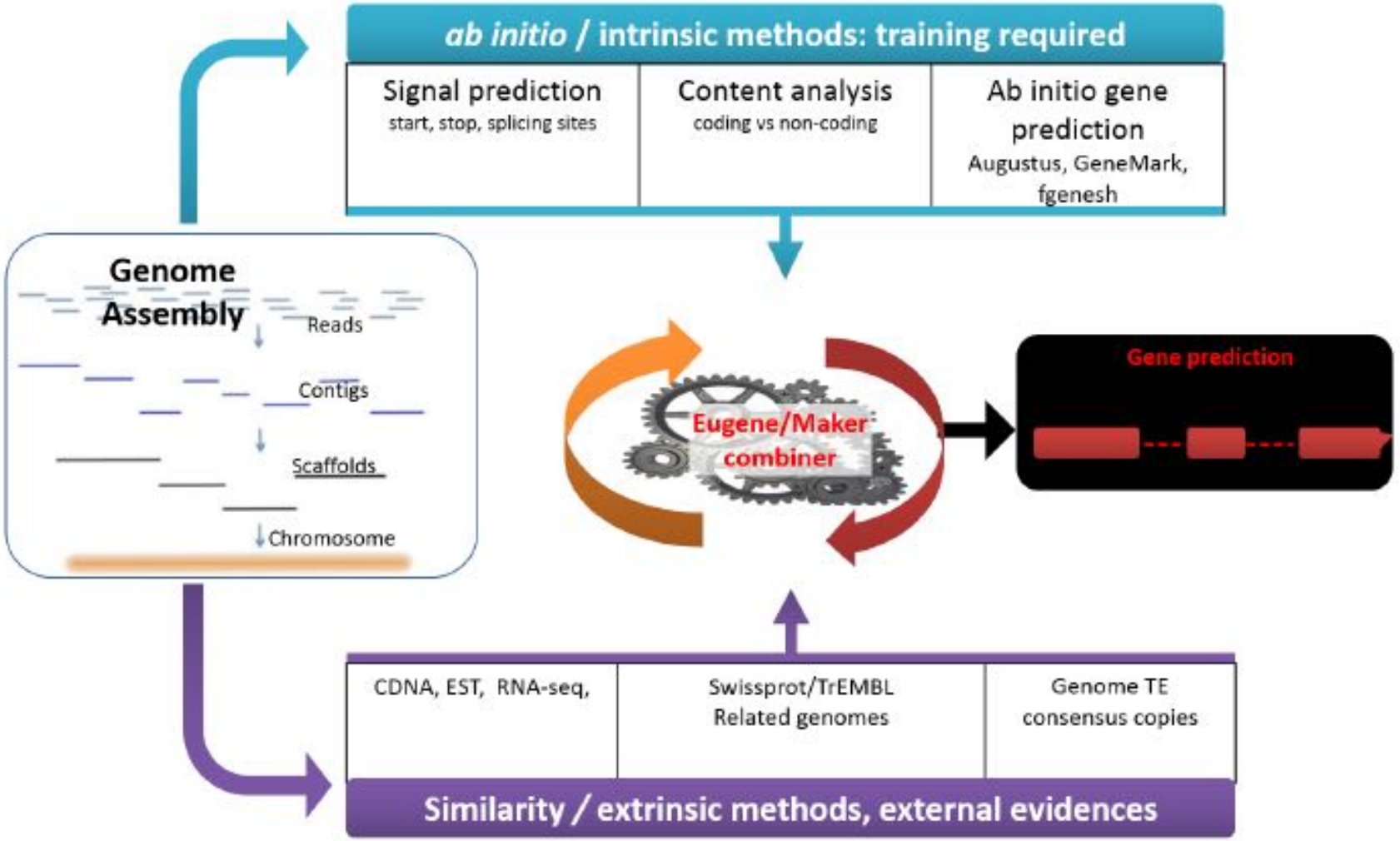
Full length =
Sequence has 5', 3', poly A tail
Strand identified

Comparison of RNA seq platforms



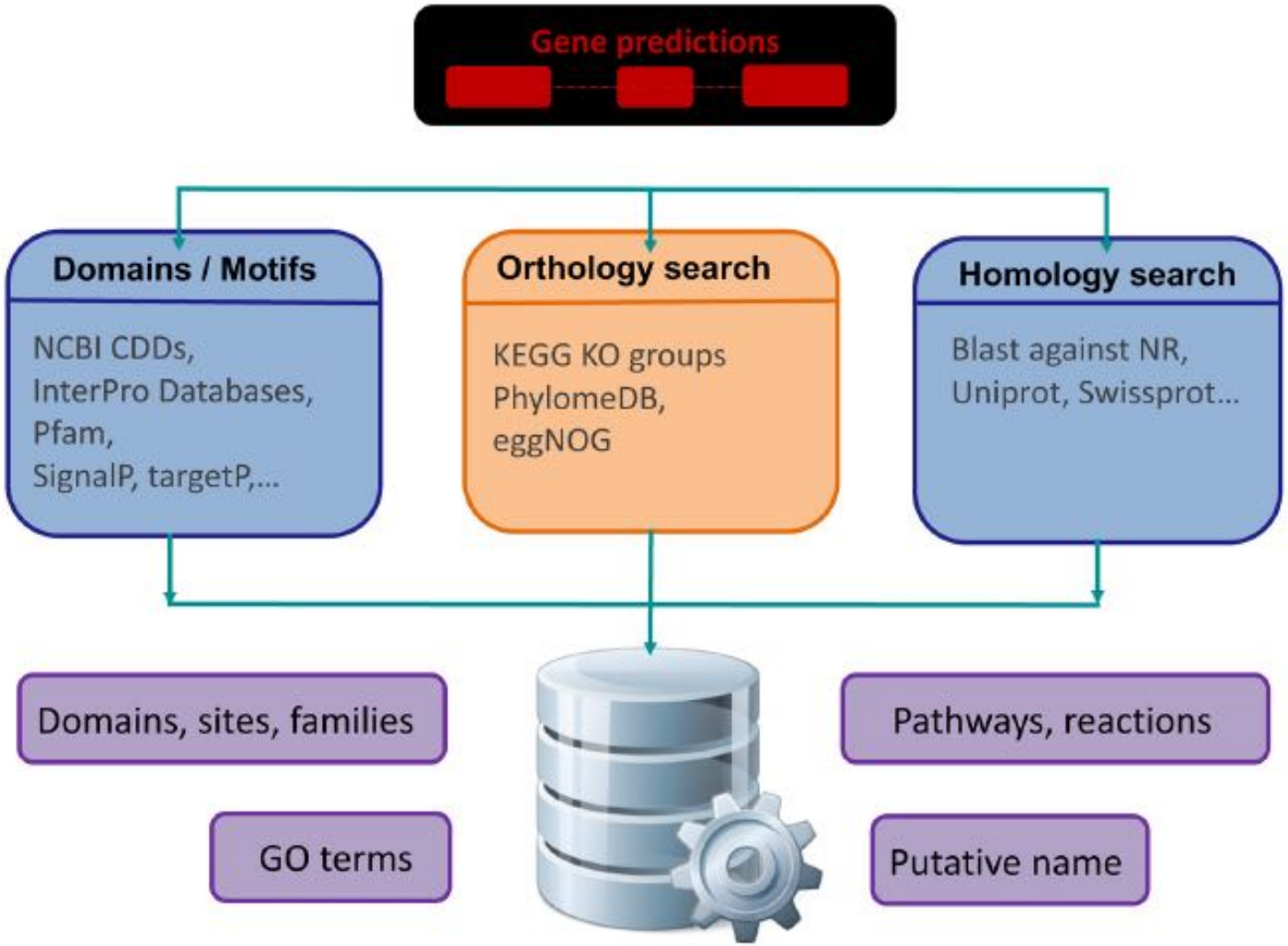
Stark et al. 2019

Gene Prediction



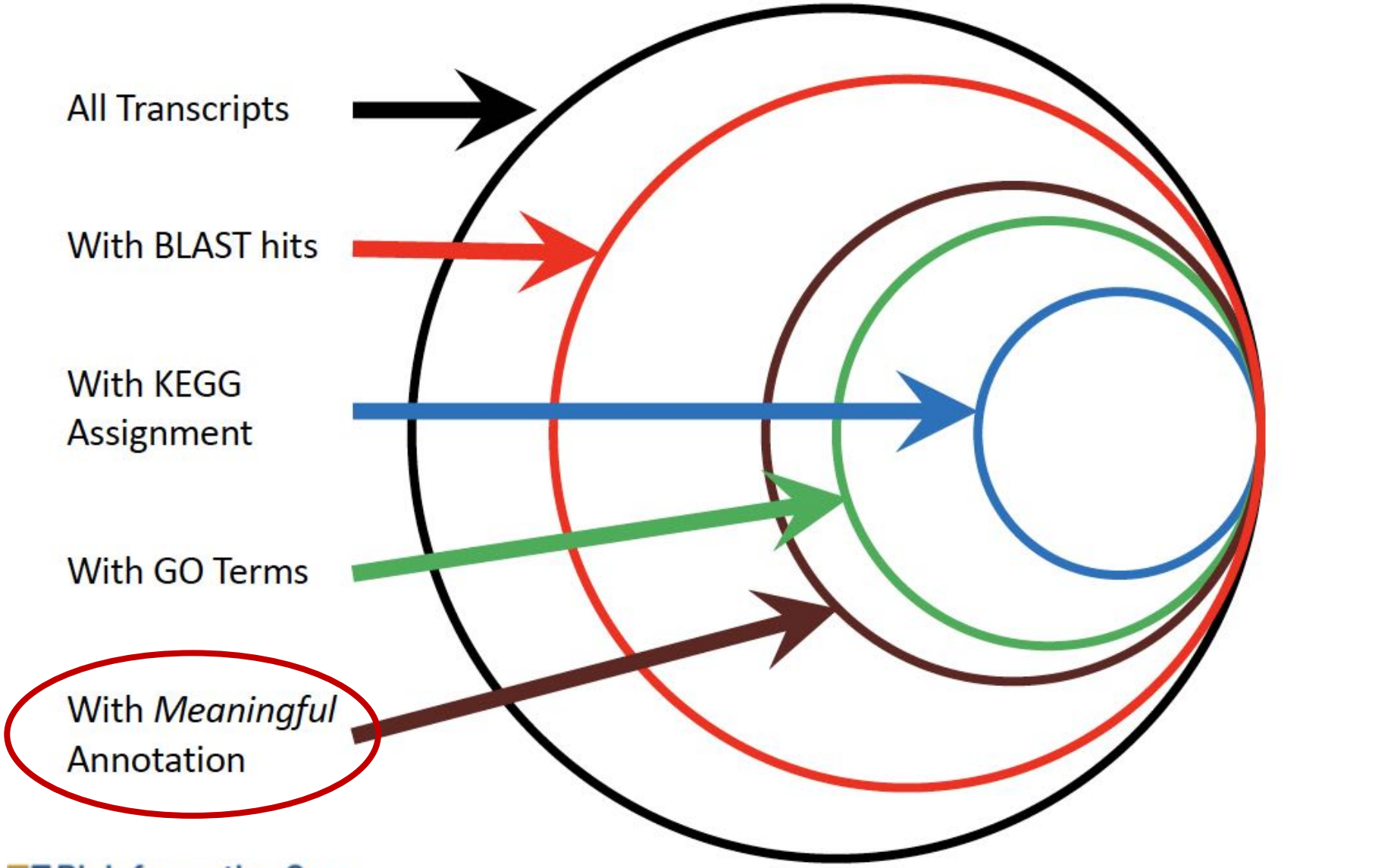
Del Angel et al. 2018

Functional Annotations



Del Angel et al. 2018

Annotation Summary Might Look Like:



Long Reads

- Improved genome assembly
  - Higher contig N50
  - Few gaps
  - More genes and signals
- Haplotype phasing resolved
- Structural variants resolved

Disadvantages:

- Extraction of long fragments
- Low throughput

RNA:

- Full length transcripts
- Isoform detection

Common Bean (*Phaseolus vulgaris*)

- **V1 published – Nature Genetics (2014)**

- **V2 – In Prep**
  - Improved genome assembly
  - Few gaps
  - Improved annotation
  - Towards Pan genome

nature genetics
OPEN

A reference genome for common bean and genome-wide analysis of dual domestications

Jeremy Schmutz[1,2,17], Phillip E McClean[3,17], Sujan Mamidi[3], G Albert Wu[1], Steven B Cannon[4], Jane Grimwood[2], Jerry Jenkins[2], Shengqiang Shu[1], Qijian Song[5], Carolina Chavarro[6], Mirayda Torres-Torres[6], Valerie Geffroy[7,8], Samira Mafi Moghaddam[3], Dongying Gao[6], Brian Abernathy[6], Kerrie Barry[1], Matthew Blair[9], Mark A Brick[10], Mansi Chovatia[1], Paul Gepts[11], David M Goodstein[1], Michael Gonzales[6], Uffe Hellsten[1], David L Hyten[5,16], Gaofeng Jia[5], James D Kelly[12], Dave Kudrna[13], Rian Lee[3], Manon M S Richard[7], Phillip N Miklas[14], Juan M Osorno[3], Josiane Rodrigues[5,16], Vincent Thareau[7], Carlos A Urrea[15], Mei Wang[1], Yeisoo Yu[13], Ming Zhang[1], Rod A Wing[13], Perry B Cregan[5], Daniel S Rokhsar[1] & Scott A Jackson[6]

Linkage Map

Theor Appl Genet (2010) 121:1103–1116
DOI 10.1007/s00122-010-1375-9

ORIGINAL PAPER

Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.)

Melody McConnell · Sujan Mamidi · Rian Lee · Shireen Chikara · Monica Rossi · Roberto Papa · Phillip McClean

2686 bean TC sequences (Ramirez et al. 2005)

↓ 89%

2394 TCs matched Arabidopsis gene at <e-20

↓ 61%

1458 TCs with >20 bases 3'UTR

↓ 71%

1046 TC sequences analyzed further
[801 random genes (78%), 148 biochemical pathway genes (14%),
62 Arabdiopsis mutant genes (6%), 19 microsynteny genes,
11 high e-value Pv genes (1%), 5 domestication selected genes in maize (0.5%)]

↓ 70%

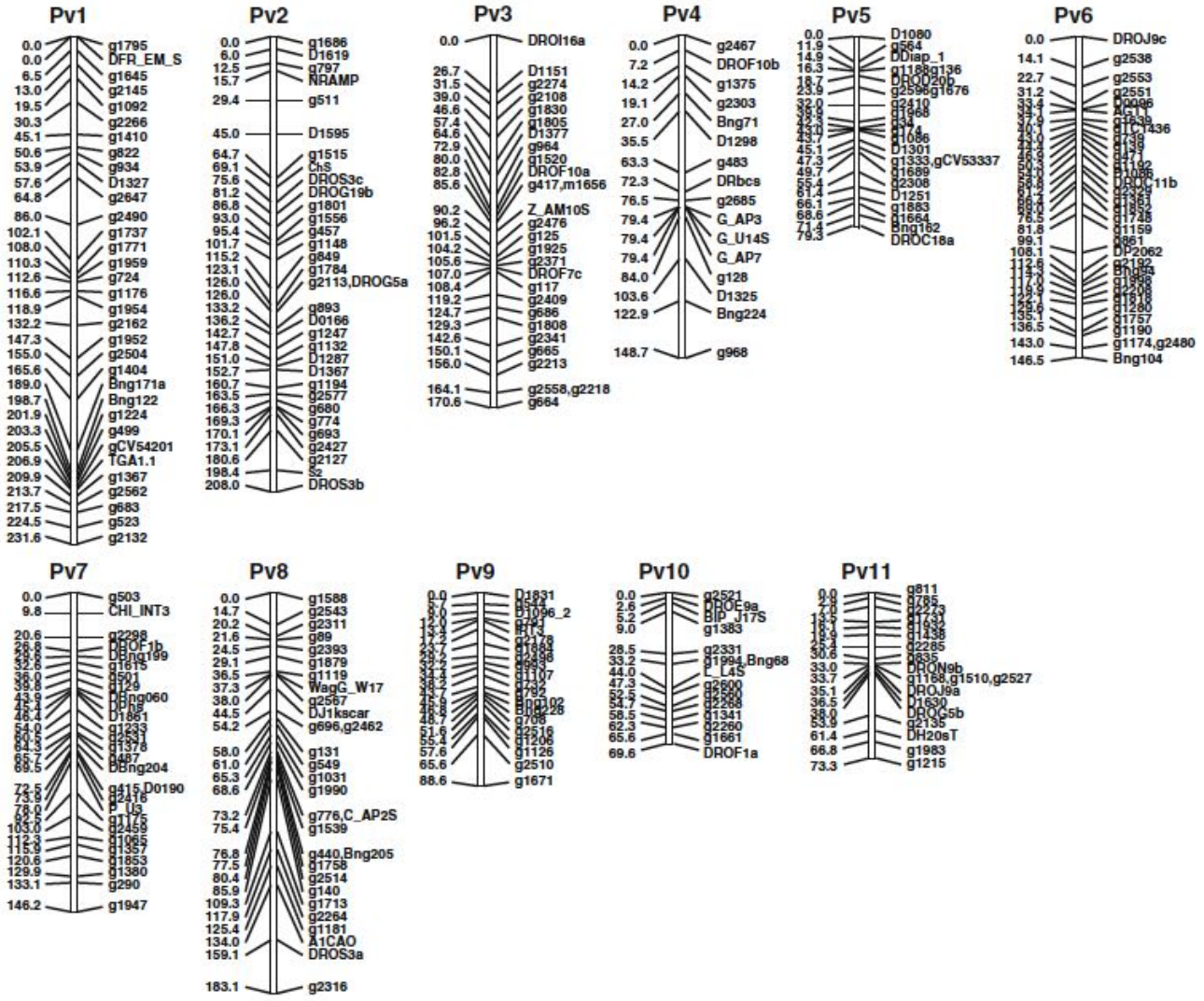
730 primers amplified BAT93 and Jalo EEP558 fragments

↓ 74%

539 BAT93 and Jalo EEP558 fragments sequenced

↙ 26% (139 fragments monomorphic) ↘ 74% (400 fragments polymorphic)

300 genes mapped
[172 CAPS (57%), 80 dCAPS (27%),
16 indel/SSR (6%), 16 size differences (5%),
13 allele-specific (4%)]



Synteny with soybean

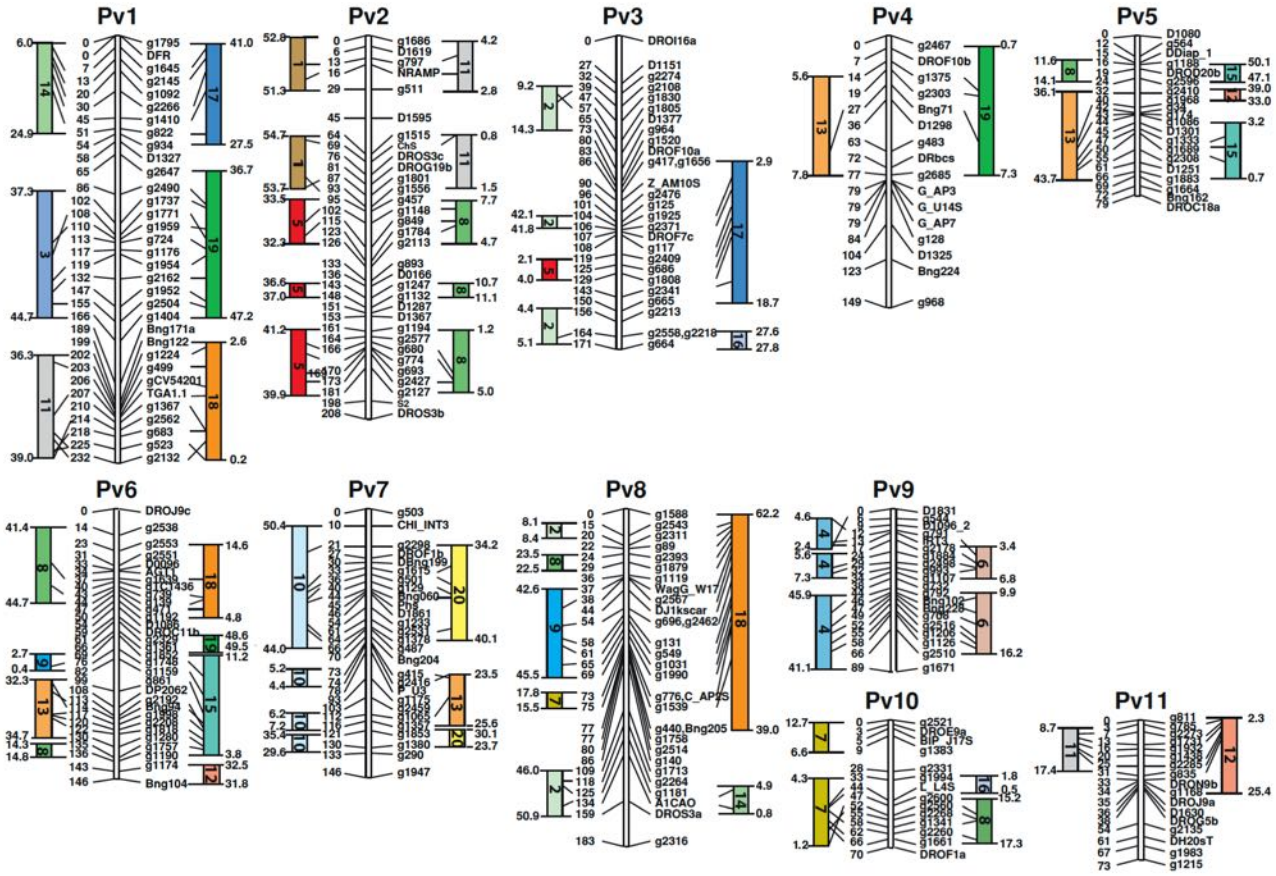
McClean *et al. BMC Genomics* 2010, **11**:184
http://www.biomedcentral.com/1471-2164/11/184

BMC Genomics

RESEARCH ARTICLE Open Access

Synteny mapping between common bean and soybean reveals extensive blocks of shared loci

Phillip E McClean[1,2*], Sujan Mamidi[1], Melody McConnell[1], Shireen Chikara[1], Rian Lee[1,2]



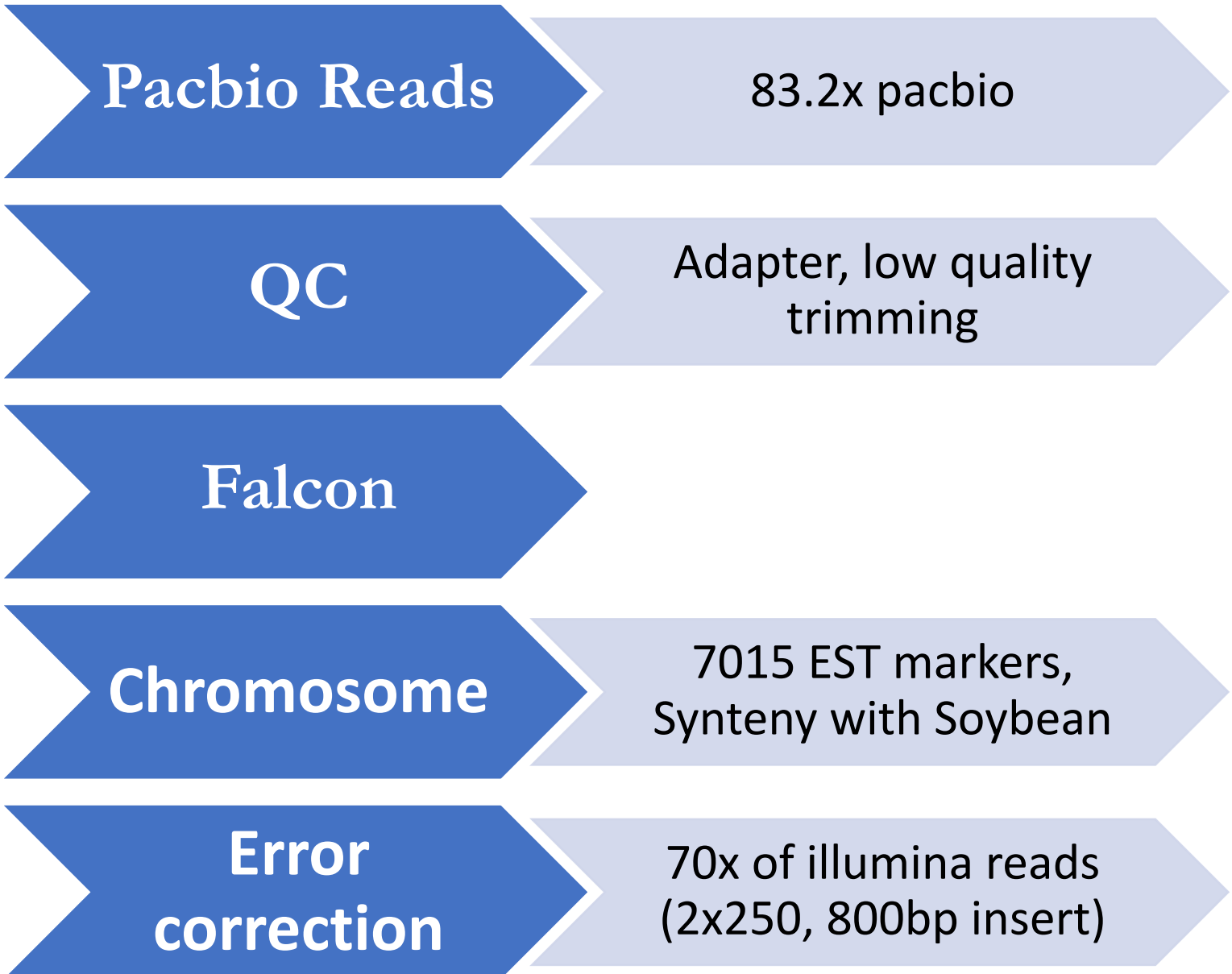
Bean V1 assembly

Sequencing	• Sanger, 454 reads (21.02x)
QC	• Adapter and low quality trimming • Organellar DNA
Assembly	• Using ARACHNE assembler • contigs of <300 bp ; < 4 reads removed
Scaffold QC	• Remove Contaminants, scaffolds >95% 24-mers in other scaffolds, < 1 kb in length.
Pseudo chromosomes	• 7,015 EST markers aligned using BLAT • Unsized map joins padded with 10,000 Ns
Completeness	• 102,254 EST of the 108,874 cDNAs (93.92%) aligned to the assembly.

Sequences Used

Library	Sequencing Platform	Average Read Length/Insert Size	Read Number	Assembled Sequence Coverage
Linear	454 XLR & FLX+	362*	38,107,155	18.64x
GPNB	454 XLR paired	2,798 ± 1,047	589,346	0.11x
GGAS	454 XLR paired	3,922 ± 643	1,940,576	0.41x
GXSF	454 XLR paired	3,991 ± 337	467,414	0.07x
HYFA	454 XLR paired	4,729 ± 497	1,648,022	0.25x
HYFC	454 XLR paired	4,736 ± 504	1,491,648	0.24x
HYFB	454 XLR paired	4,759 ± 528	1,196,104	0.17x
HXTI	454 XLR paired	8,022 ± 1,016	1,364,808	0.22x
GXNX	454 XLR paired	9,192 ± 1,058	878,832	0.16x
HXWF	454 XLR paired	11,903 ± 1,928	724,196	0.13x
HXWH	454 XLR paired	12,231 ± 1,902	413,396	0.08x
VUK	Sanger	34,956 ± 4,536	240,384	0.20x
VUL	Sanger	36,001 ± 4,632	88,320	0.08x
PVC	Sanger	121,960 ± 16,572	81,408	0.08x
PVA	Sanger	126,959 ± 25,658	89,017	0.09x
PVB	Sanger	135,292 ± 21,487	92,160	0.09x
Total			49,412,786	21.02x

Bean V2 assembly



	V1	V2
Sequencing method	WGS - 454 & Sanger	WGS - Pacbio
Raw Sequence total	21.02 x	83.2x
Markers for scaffolding	7015 from F2 map, synteny with soybean	7015 from F2 map, synteny with soybean
Scaffold Total	708	478
Contig Total	**41,391**	**1044**
Scaffold Sequence total	521.1 Mbp	537.2 Mbp
Contig sequence total	472.5 Mbp (9.3 % Gap)	531.6 Mbp (1.1% Gap)
Scaffold N50/L50	5/50.4 Mbp	5/49.7 Mbp
Contig N50/L50	3,273/39.5 Kbp	**73/1.9Mbp**

Bean RNA seq – V1 & V2

Sequencing	• 727 M paired end reads (11 tissues)
Assembly (Pertran)	• 43,627 Transcripts
Support	• 79,630 Sanger ESTs
Assembly (PASA)	• 47,464 transcripts
Alignment (Exonerate)	• At, Medicago, Poplar, grape, rice
Intrinsic	• FGENESH+, FGENESH_EST, GenomeScan
Improvement (PASA)	• Adding UTRs, correcting splicing, adding alternative transcripts.
Filtering	• Homology search, C-score

RNA Seq Used - Bean

Resource type	Tissue Type	Number of reads	GSNAP (Wu and Nacu 2010) Aligned	Percent Aligned
Sanger	Mixed	79,630	-	-
Illumina 2x100 bp	Roots 10 DAP (days after planting)	65,429,570	59,846,373	92.1%
Illumina 2x100 bp	Roots 19 DAP	46,593,274	44,116,235	94.9%
Illumina 2x100 bp	Nodules 19 DAP	71,716,844	66,112,750	92.7%
Illumina 2x100 bp	Stem 10 DAP	40,933,844	38,196,918	93.6%
Illumina 2x100 bp	Stem 19 DAP	61,842,390	44,116,235	94.9%
Illumina 2x100 bp	Primary leaves 10 DAP	68,255,918	61,371,430	90.5%
Illumina 2x100 bp	Young trifoliates 19 DAP	66,127,642	60,209,317	91.6%
Illumina 2x100 bp	Flower buds	68,363,986	61,332,231	90.5%
Illumina 2x100 bp	Whole Flowers	66,112,818	62,126,340	94.7%
Illumina 2x100 bp	Young pods 1-5cm seedless	66,133,582	62,301,836	94.8%
Illumina 2x150 bp	Green mature pods 11.5-13.5 cm	120,724,870	113,736,673	94.6%
Total RNA-Seq		742,234,738	687,643,736	93.2%

Comparison of Bean versions

	V1	V2
Primary Loci	27,197	27,433
Alternate transcripts	**4,491**	**9,562**
Complete genes	26,279	26,827
Incomplete genes with start Codon	225	261
Incomplete genes with Stop Codon	657	321

Switchgrass (*P.virgatum*) - Polyploid

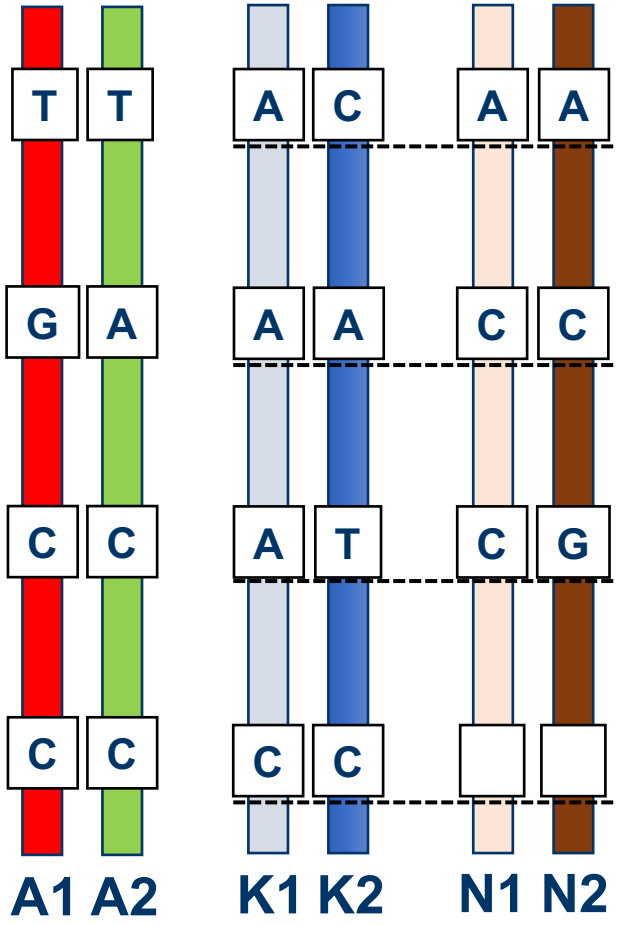
Title: Polyploidy and genomic introgressions facilitate climate adaptation and biomass yield in switchgrass

Running title: The Switchgrass Genome

Authors: John T. Lovell[1+*], Alice H. MacQueen[2+], Sujan Mamidi[1+], Jason Bonnette[2+], Jerry Jenkins[1+], Joseph D. Napier[2], Avinash Sreedasyam[1], Adam Session[3,4], Shengqiang Shu[3], Kerrie Barry[3], Stacy Bonos[5], LoriBeth Boston[1], Christopher Daum[3], Shweta Deshpande[3], Aren Ewing[3], Paul P. Grabowski[1], Taslima Haque[2], Melanie Harrison[6], Adam Healey[1], Jiming Jiang[7], Dave Kudrna[8], Anna Lipzen[3], Thomas H. Pendergast IV[9], Chris Plott[1], Peng Qi[9], Christopher A. Saski[10], Eugene V. Shakirov[2,11], David Sims[1], Manoj Sharma[12], Rita Sharma[13], Ada Stewart[1], Vasanth R. Singan[3], Yuhong Tang[14], Sandra Thibivillier[15], Jenell Webber[1], Xiaoyu Weng[2], Melissa Williams[1], Guohong Albert Wu[3], Yuko Yoshinaga[3], Matthew Zane[3], Li Zhang[2], Jiyi Zhang[14], Kathrine D. Behrman[2], Arvid R. Boe[16], Philip A. Fay[17], Felix B. Fritschi[18], Julie D. Jastrow[19], John Lloyd-Reilley[20], Juan Manuel Martínez-Reyna[21], Roser Matamala[19], Robert B. Mitchell[22], Francis M. Rouquette Jr[23], Pamela Ronald[24], Malay Saha[14], Christian M. Tobias[25], Michael Udvardi[14], Rod Wing[8], Yanqi Wu[26], Laura E. Bartley[27], Michael Casler[28], Katrien M. Devos[9], David B. Lowry[7], Daniel S. Rokhsar[3,4,29,30], Jane Grimwood[1], Thomas E. Juenger[2*], Jeremy Schmutz[1*]

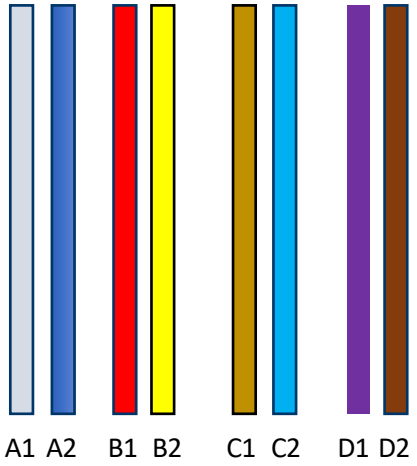
Submitted to Nature

Switchgrass is a complex genome



P. hallii *P. Virgatum 4x*

P. Virgatum 6x?
Nx?



P. Virgatum 8x

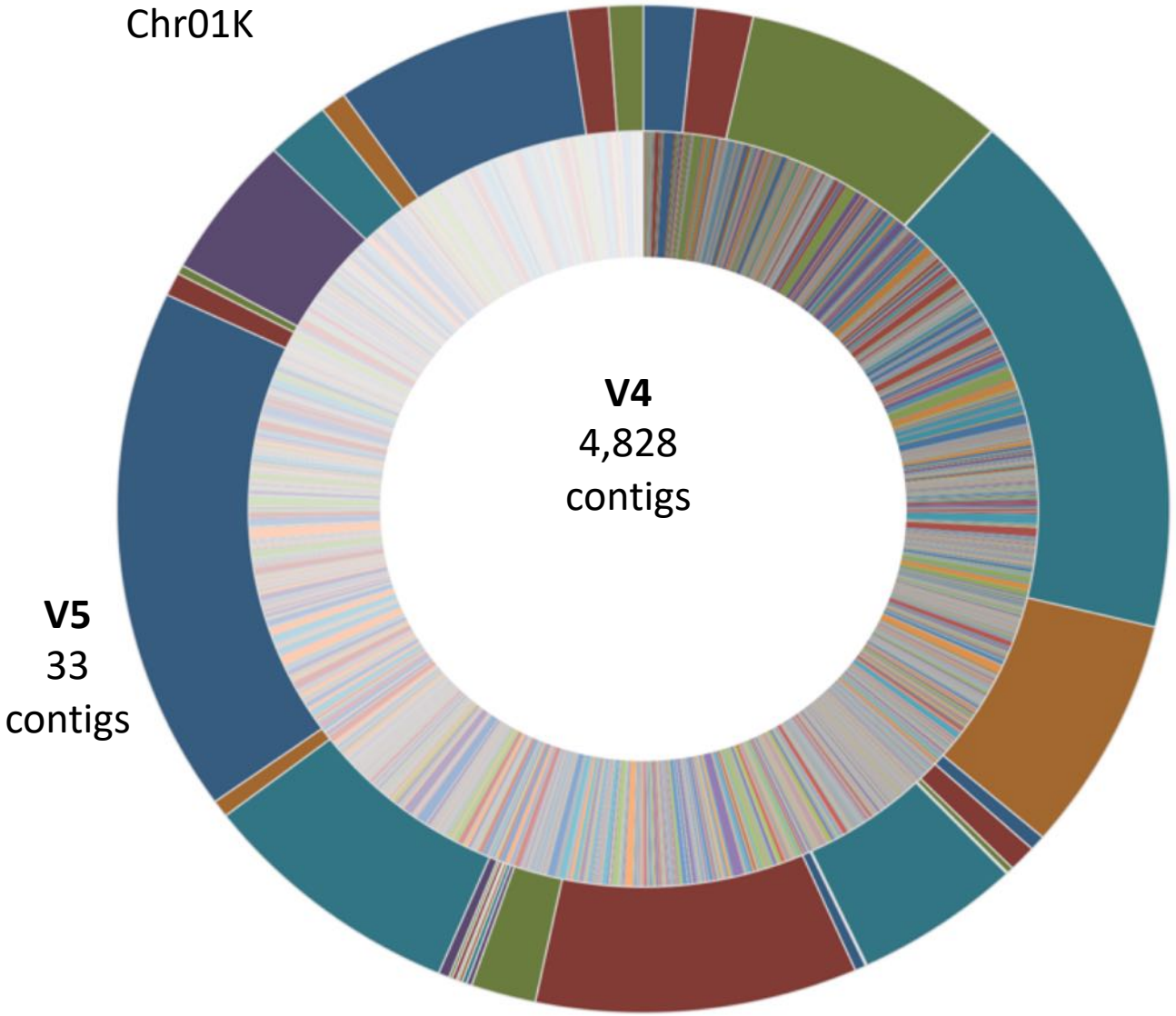
- Outbred
- Large genome
- Highly repetitive
- Polyploid
- Highly heterozygous.
- Ordering: Synteny with *P. hallii*, F2 genetic map, gene order of the alternative subgenome

Switchgrass V5

Seq Platform	Mean insert size	# reads (M)	Coverage
Illumina	500	1,325.3	177.12
PACBIO	10.758	6,893	83.42
Total		1,332.3	260.54

Genome attribute	Size/Value
Scaffold total	626
Contig total	1090
Scaffold seq total	1,129.9 Mb
Contig sequence total	1,125.2 Mbp (0.4% gap)
Contig L50/N50	62 / 5.5 Mbp
Chromosome seq	1,093.8 Mbp (97.2%)

Assembly Improvement



V5 genome annotation

Illumina RNA-sequencing (>3B reads)

# of libraries=88,

# of conditions=18,

PacBio Iso-Seq (>4.5M reads)

# of conditions=9

Genome attribute	Size/Value
Primary transcripts	80,278
Alternate transcripts	49,664
Total transcripts	129,942
RNA-Seq support	69,363

Human

- ~ 3.1 Mbp
- Closed 29 gaps

nature

https://doi.org/10.1038/s41586-020-2547-7

Accelerated Article Preview

Telomere-to-telomere assembly of a complete human X chromosome

Received: 30 July 2019

Accepted: 29 May 2020

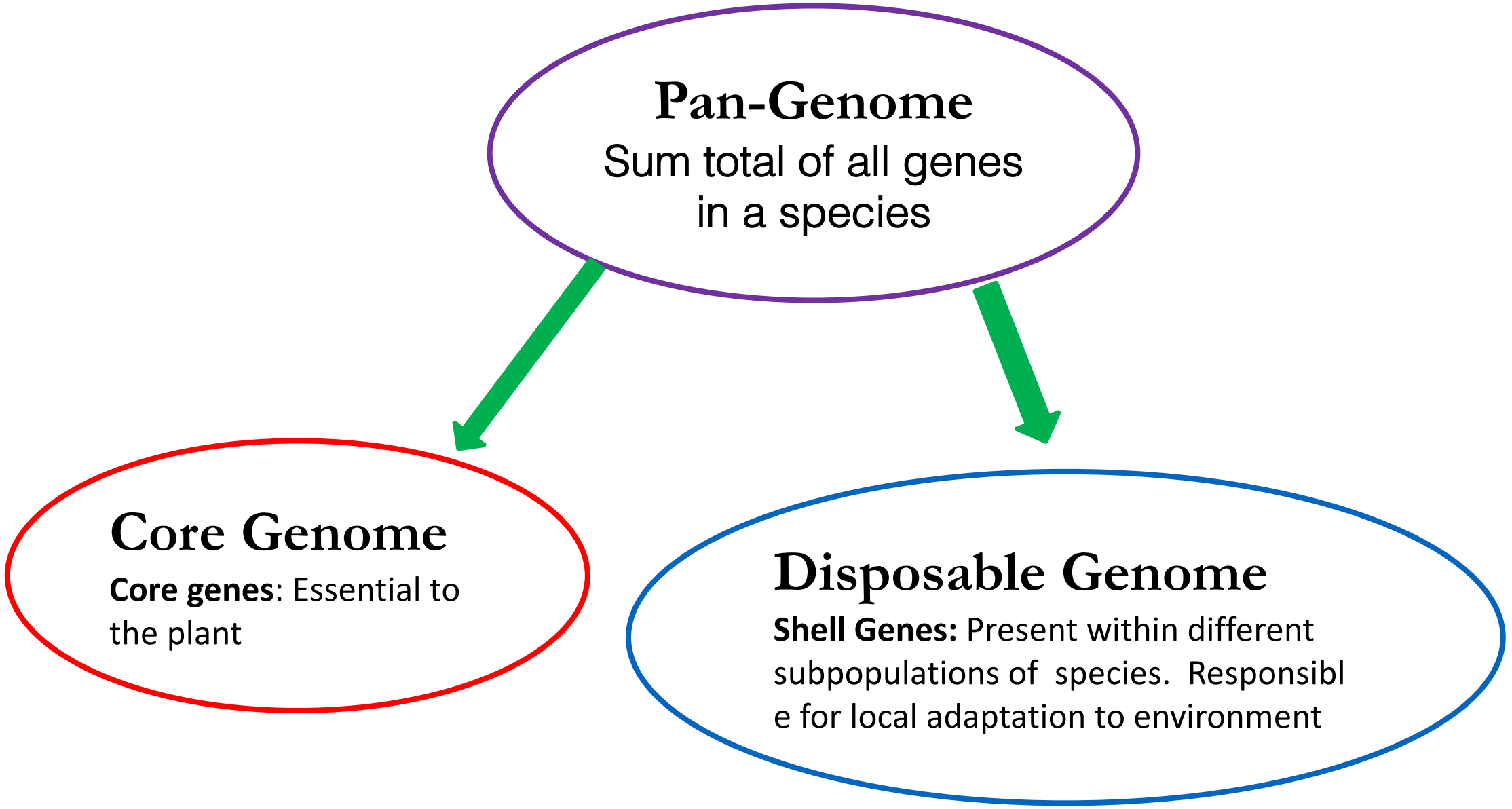
Accelerated Article Preview Published online 14 July 2020

Cite this article as: Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* https://doi.org/10.1038/s41586-020-2547-7 (2020).

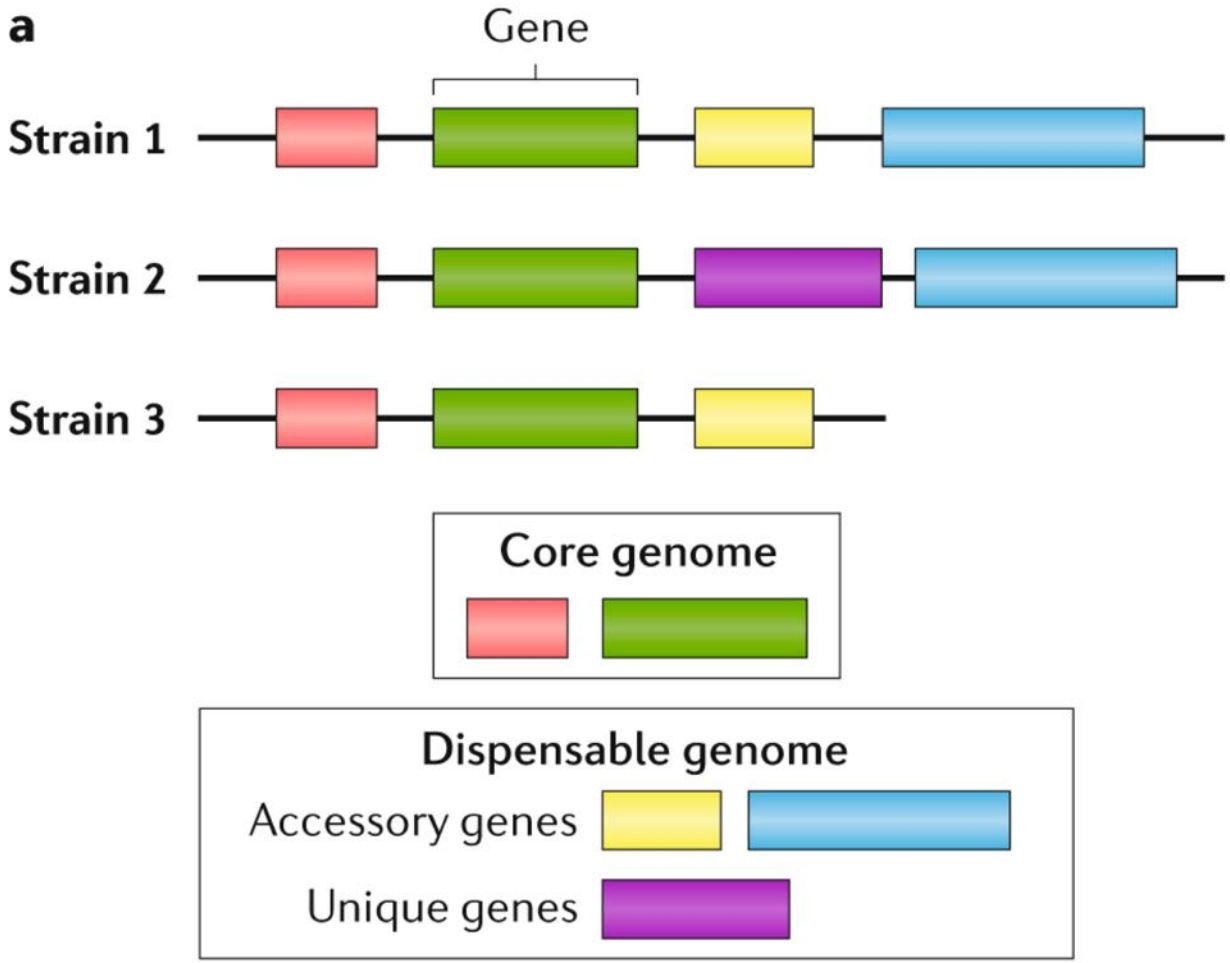
Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A. Logsdon, Valerie A. Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G. Bouffard, Alexander M. Chang, Nancy F. Hansen, Amy B. Wilfert, Françoise Thibaud-Nissen, Anthony D. Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y. Dennis, Daniela C. Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J. Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa ana Risques, Tina A. Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C. Mullikin, Pavel A. Pevzner, Jennifer L. Gerton, Beth A. Sullivan, Evan E. Eichler & Adam M. Phillippy

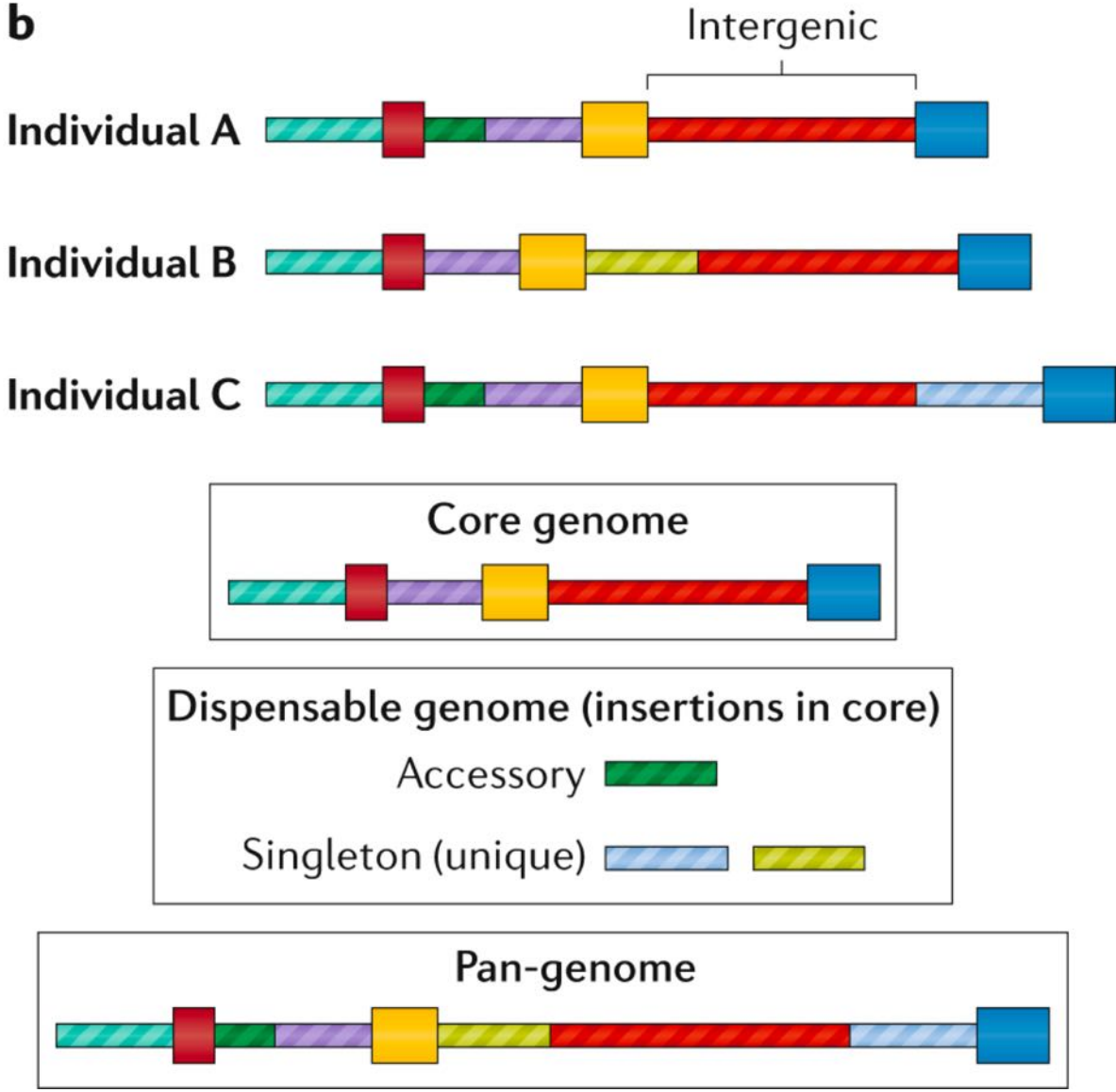
Primary Technology	Assembly	Size (Gbp)	No. Ctgs	NG50 (Mbp)
56× Illumina linked reads	Supernova(this paper)	2.95	42,828	0.21
76× PacBio CLR	FALCON([50])	2.88	1,916	28.2
24× PacBio HiFi	Canu([22])	3.03	5,206	29.1
Sanger BACs	GRCh38p13([2])	3.27	1,590	56.4
39× Nanopore Ultra-Long	Canu(this paper)	2.94	448	70.1

PAN GENOME


Pan-Genome
Sum total of all genes in a species
Core Genome
Core genes: Essential to the plant
Disposable Genome
Shell Genes: Present within different subpopulations of species. Responsibl e for local adaptation to environment

PAN GENOME

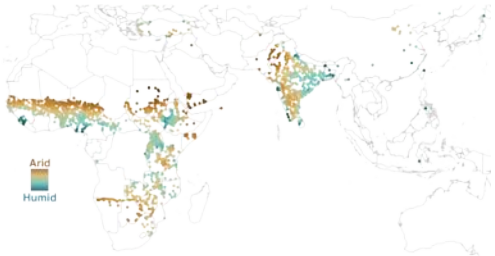




Sherman & Salzberg 2020

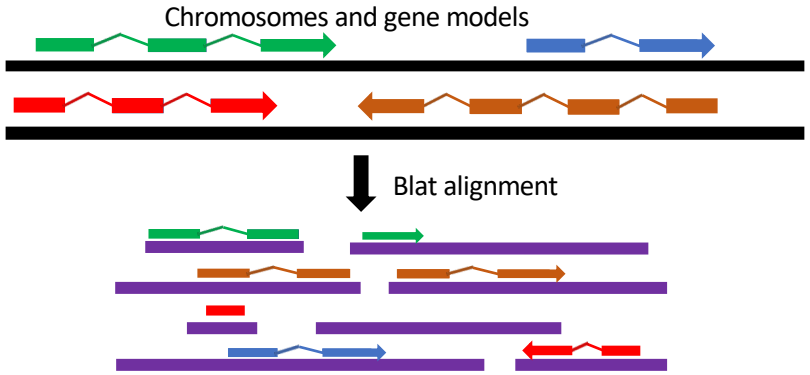
Draft Pan-Genome Construction

1) Sequence Diversity Panel



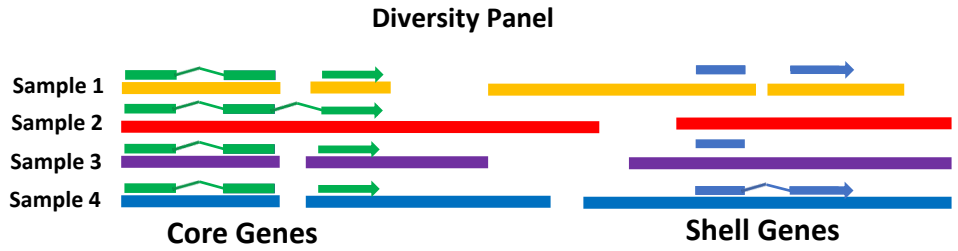
2) *De novo* Assembly - Hipmer

3) Align Reference Genes to Contigs



4) Order and orientation into chromosomes

5) Quantify (PAV)



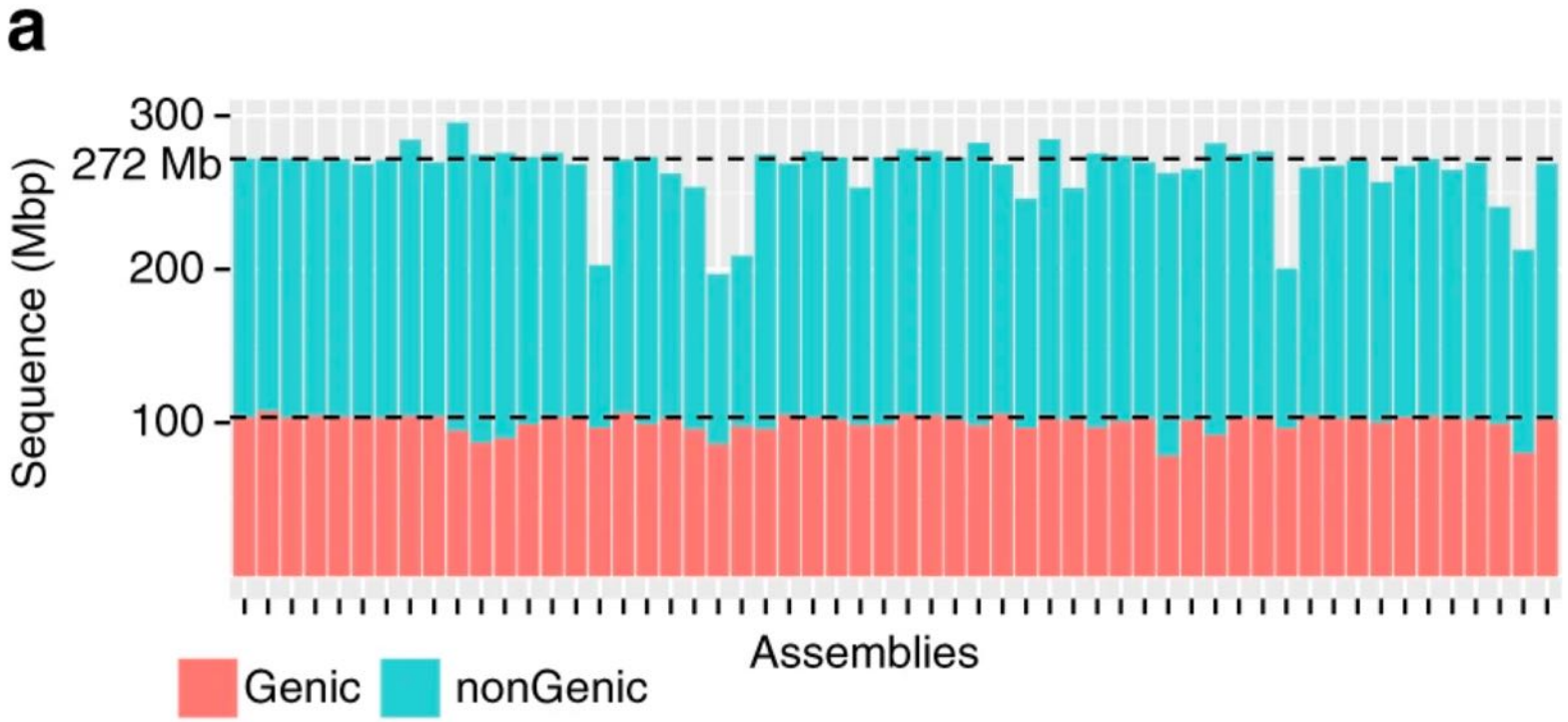
6) Gene clustering and analysis

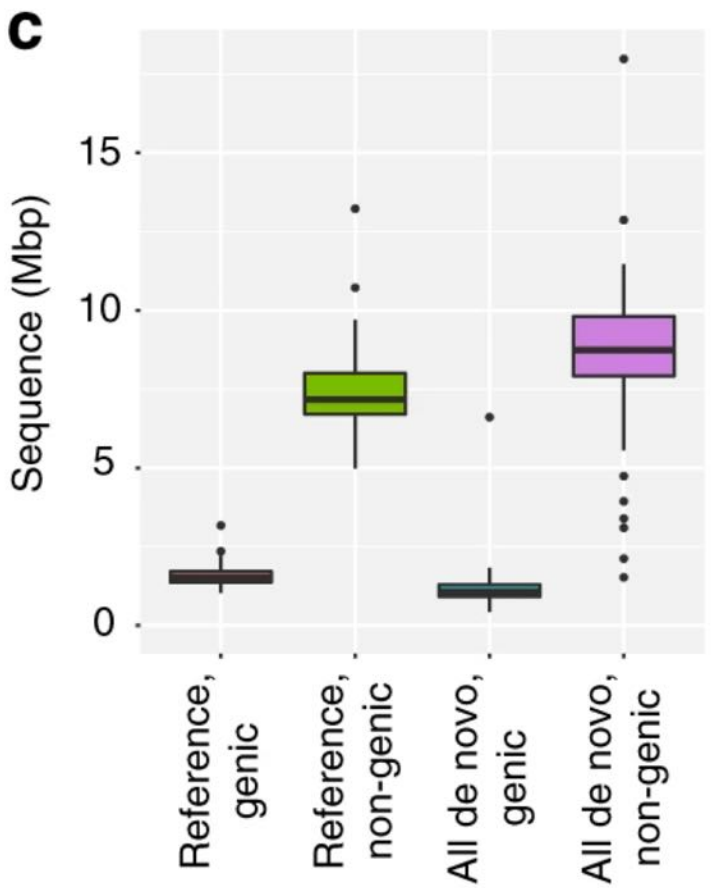
- Gene selection among subpopulations or geography
- Un-annotated/novel genes not found in reference genome
- Conduct new GWAS analysis and narrow down candidate gene loci
- Core vs variable vs private alleles

Adam Healey

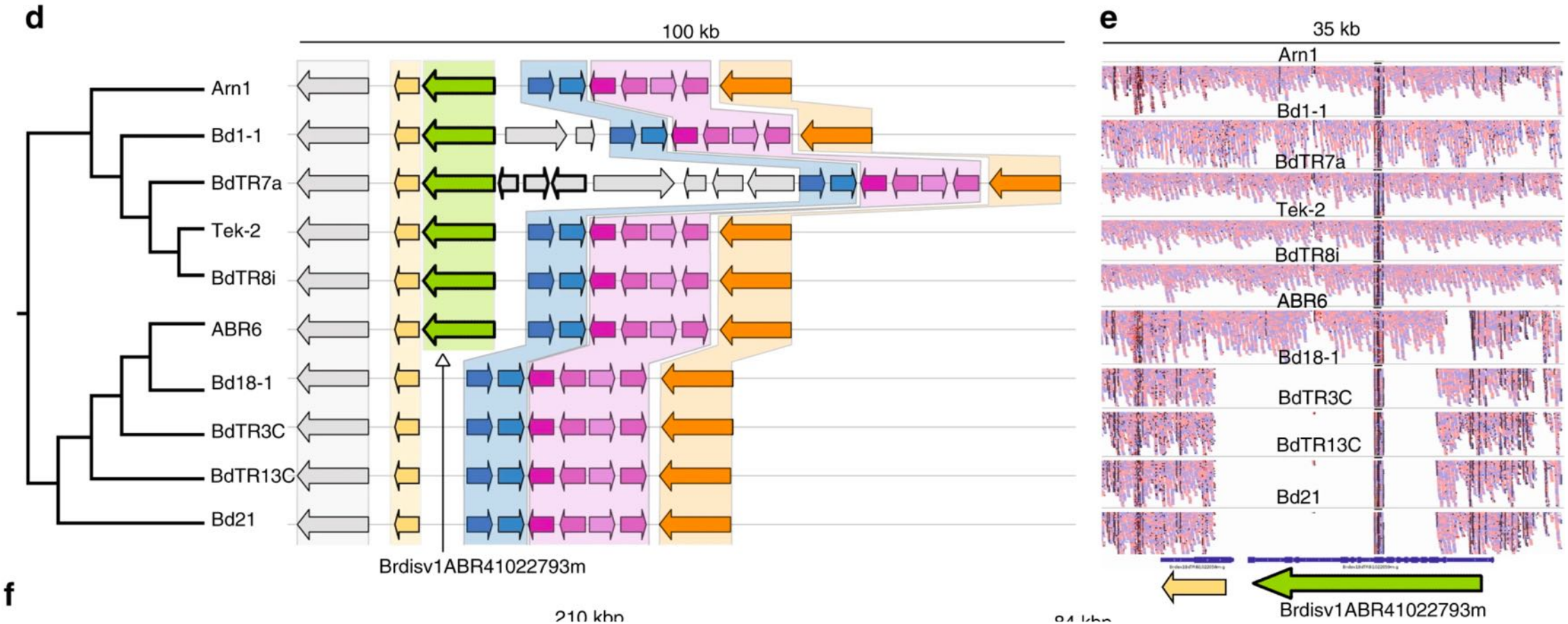
Brachypodium Pangenome

- N=54
- 92x genome coverage (100bp paired end)
- Mean assembled genome (268Mbp)





Gordan et al. 2017

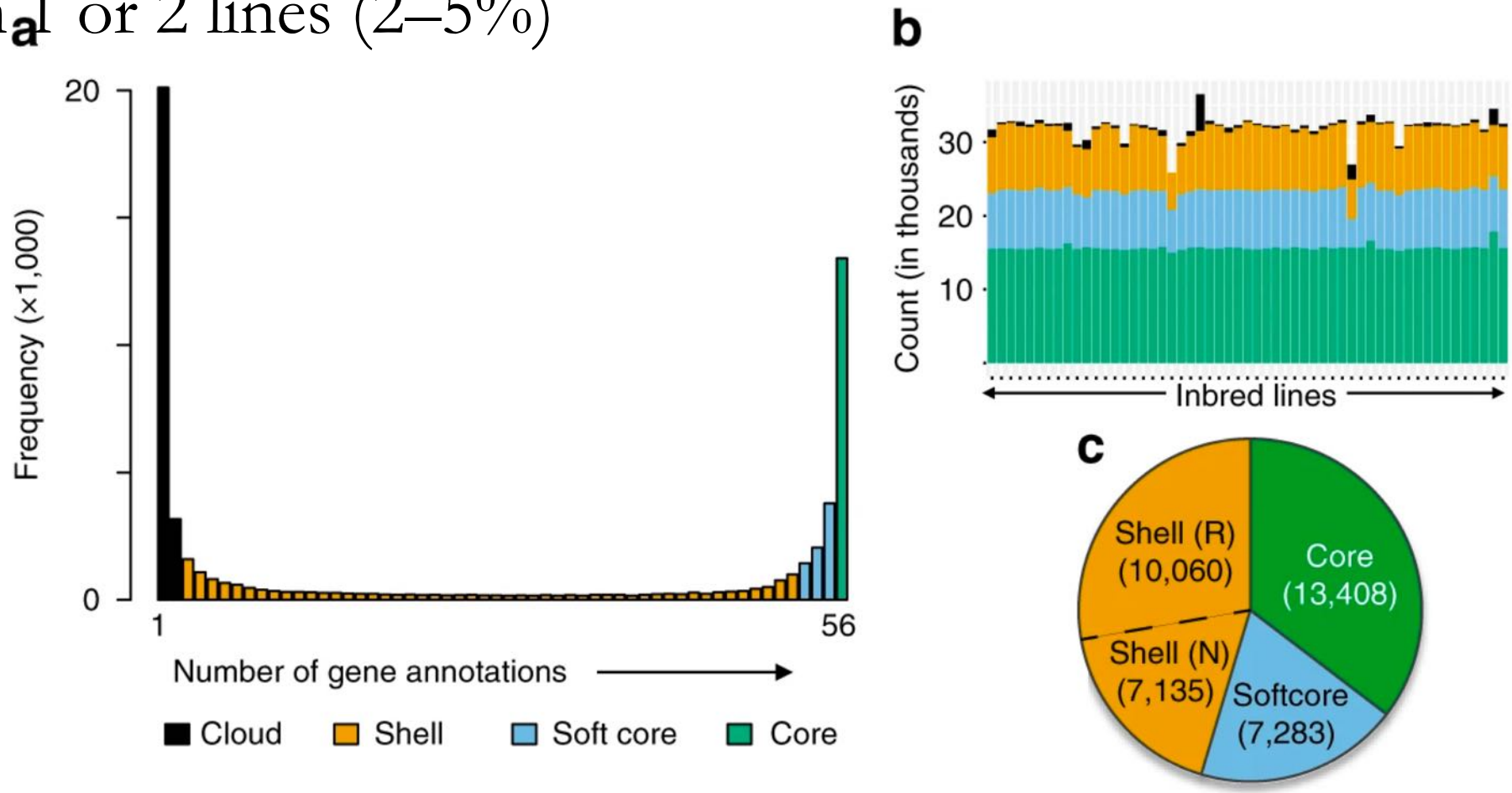


Gordan et al. 2017

Core - Genes in all lines
Softcore - Genes in 95–98%
Shell - Genes in 5–94%
Cloud - Genes in 1 or 2 lines (2–5%)

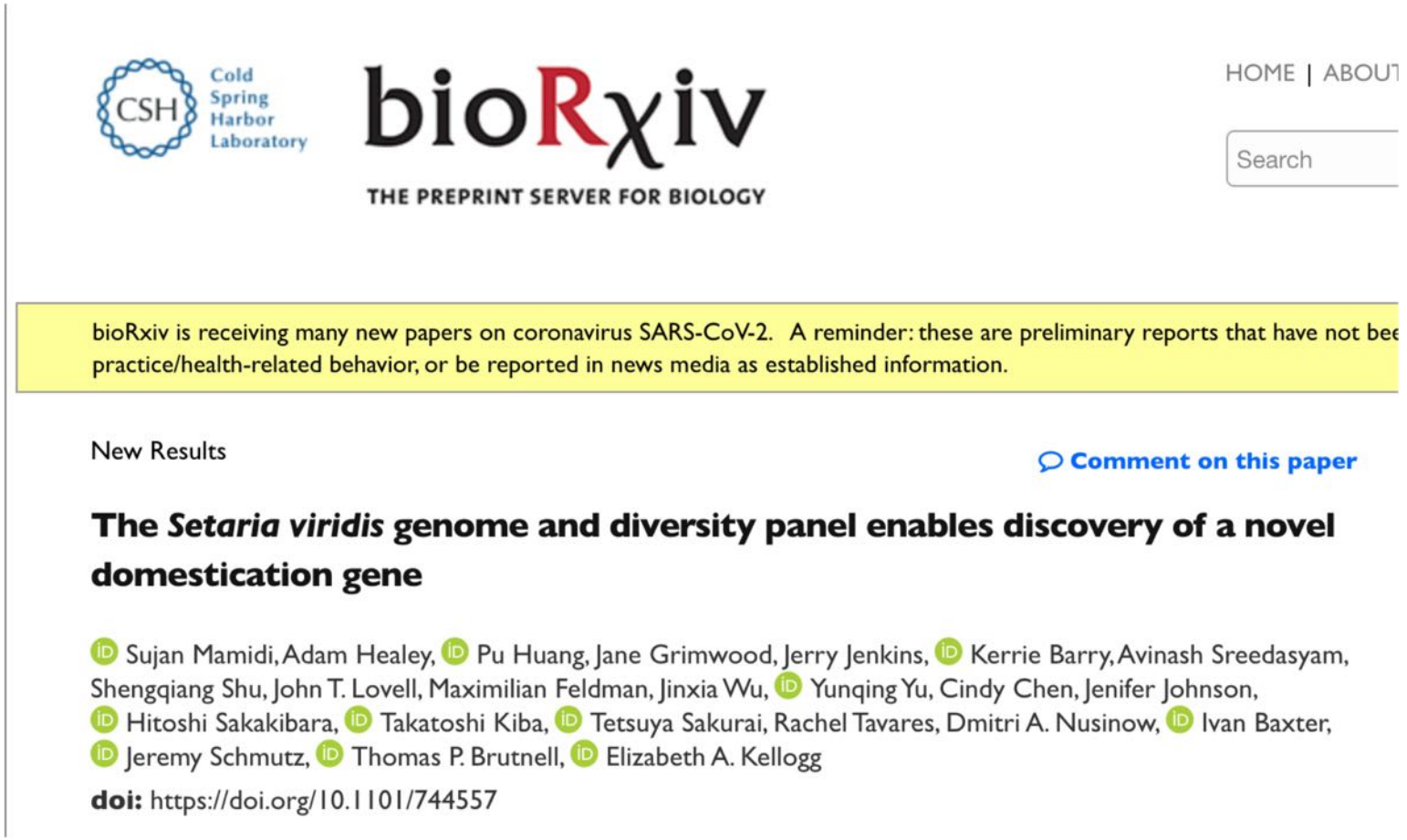
Pan Genes Total - 61,155

Reference genes – 36,647



Gordan et al. 2017

Pan Genome of *Setaria viridis*



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT

Search

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not bee
practice/health-related behavior, or be reported in news media as established information.

New Results

Comment on this paper

The *Setaria viridis* genome and diversity panel enables discovery of a novel domestication gene

Sujan Mamidi, Adam Healey, Pu Huang, Jane Grimwood, Jerry Jenkins, Kerrie Barry, Avinash Sreedasyam, Shengqiang Shu, John T. Lovell, Maximilian Feldman, Jinxia Wu, Yunqing Yu, Cindy Chen, Jenifer Johnson, Hitoshi Sakakibara, Takatoshi Kiba, Tetsuya Sakurai, Rachel Tavares, Dmitri A. Nusinow, Ivan Baxter, Jeremy Schmutz, Thomas P. Brutnell, Elizabeth A. Kellogg

doi: https://doi.org/10.1101/744557

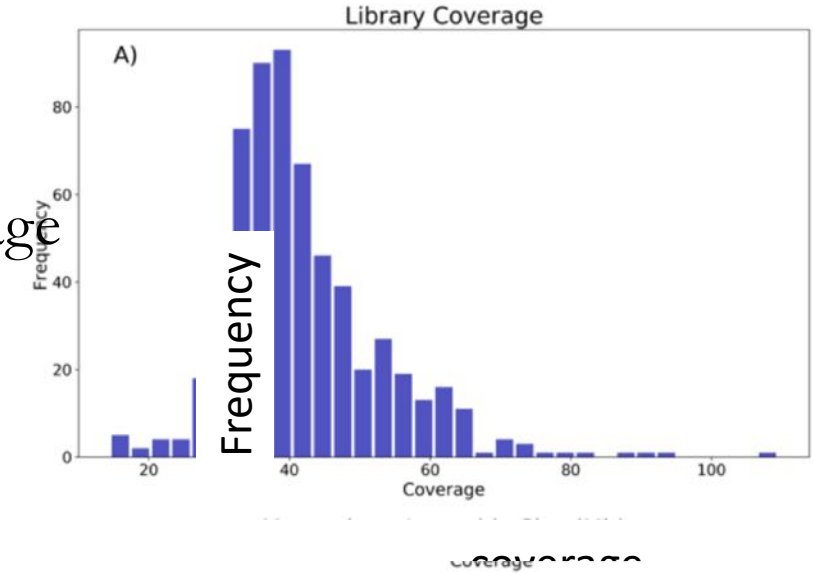
Accepted in **Nature Biotechnology**

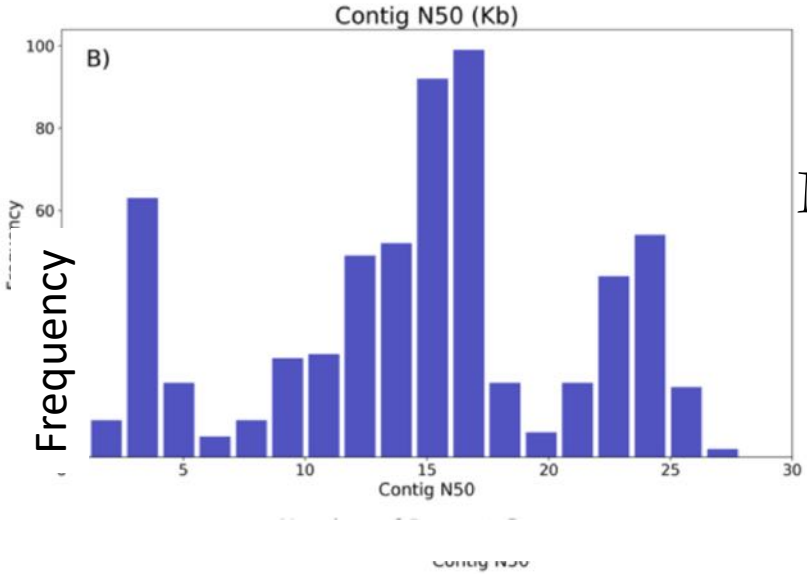
Setaria North American Collections (n=598)

Collected by Kellogg lab 2010--12; Collected by Max 2013; Collected by Pu 2013; Collected by Pu 2014

Sequencing

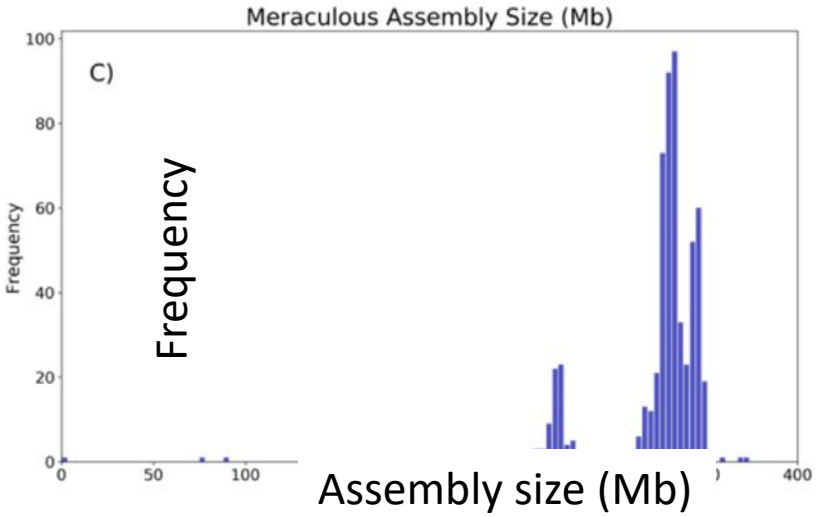
Mean Coverage
41.8X

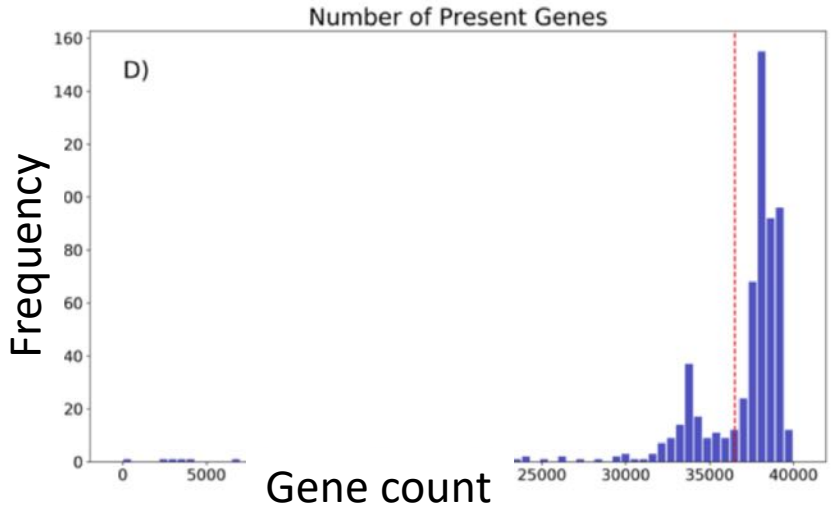




Mean contig N50:
16.2 Kb

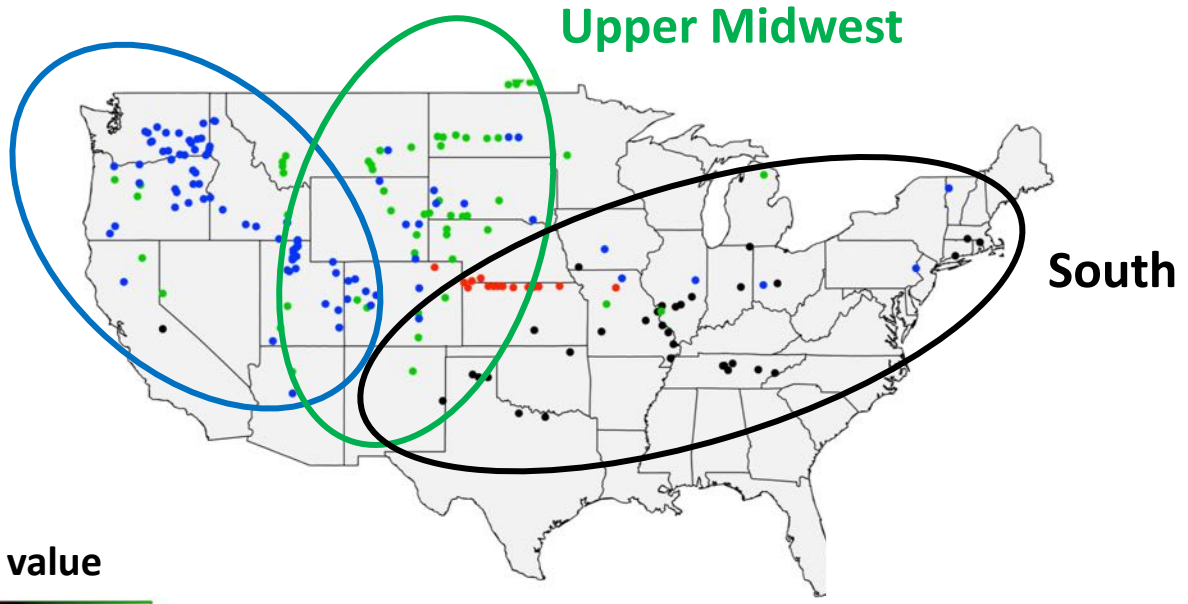
Mean assembled bases
322.5 Mb

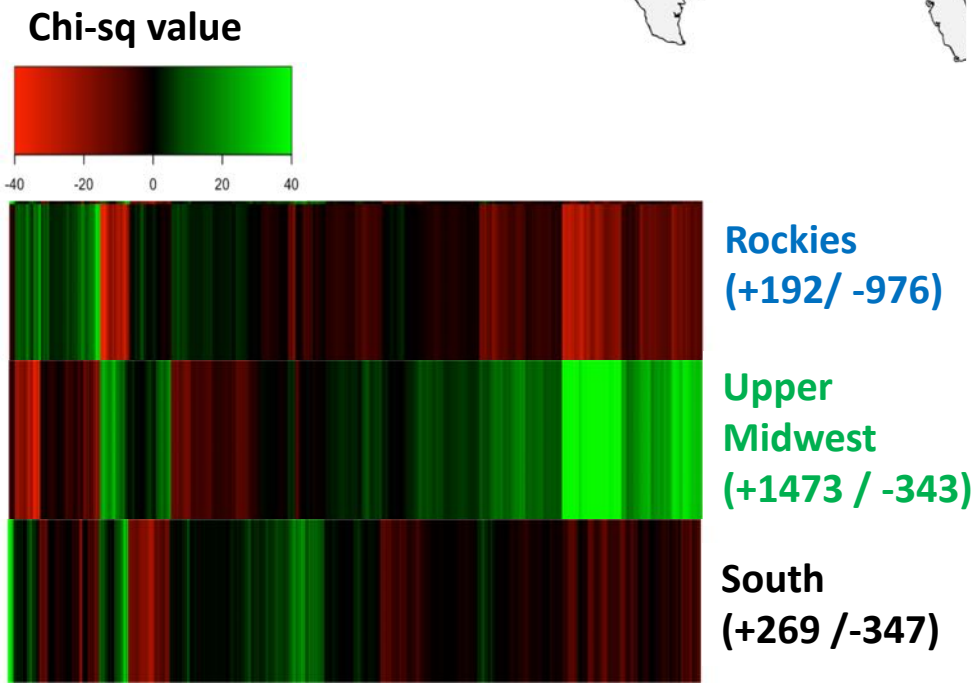




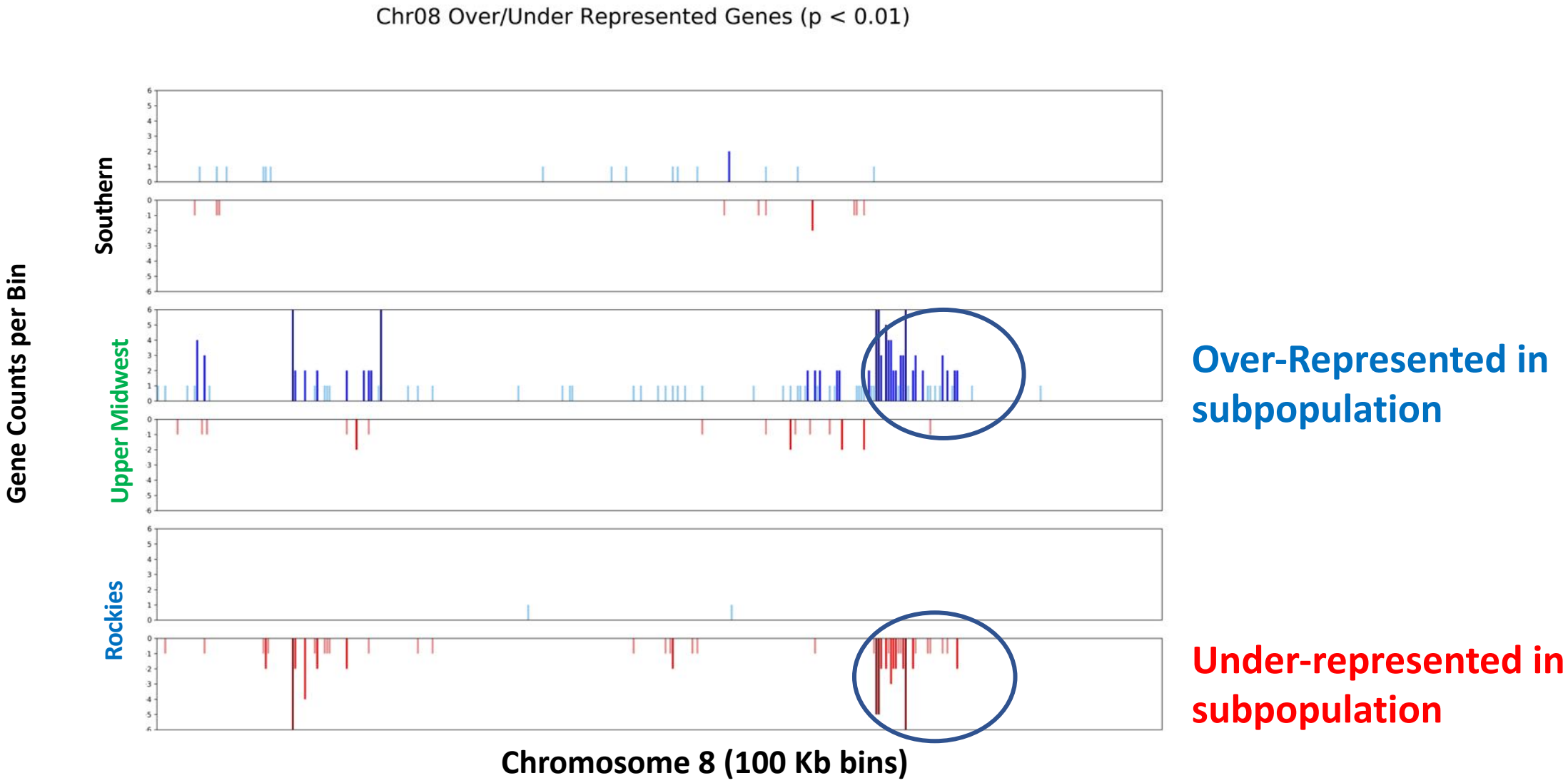
Genes > over 38,500 (red line)

Setaria Geographic Gene Differences





Core genes 34,940
Shell Genes 3,355

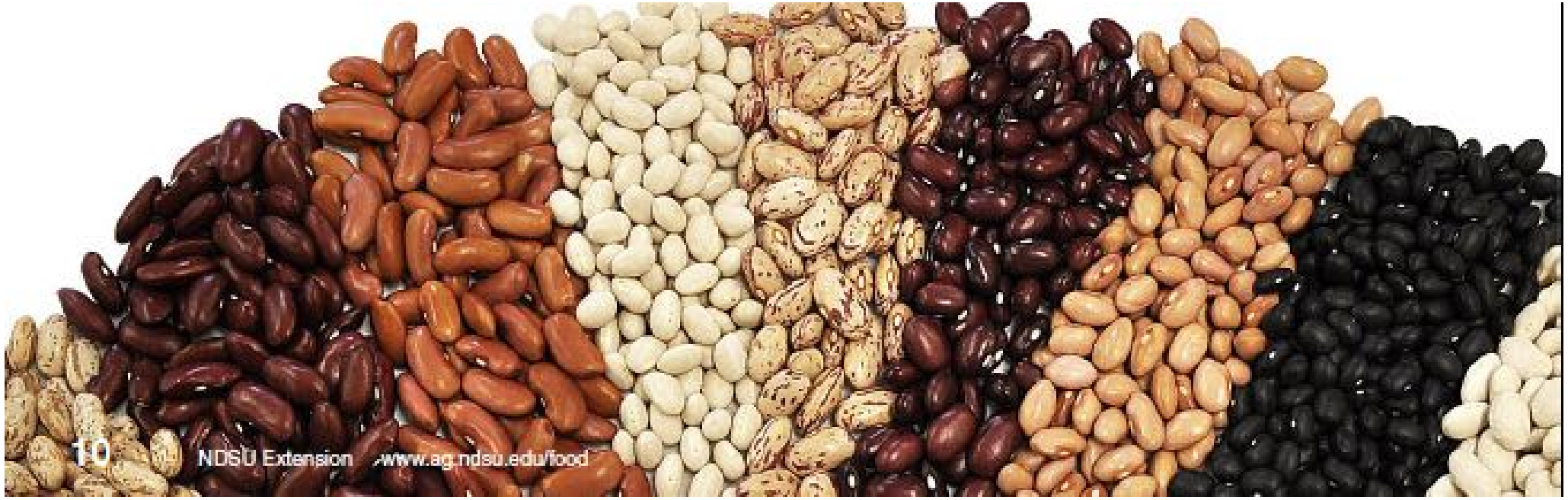

Chr08 Over/Under Represented Genes (p < 0.01)
Gene Counts per Bin
Southern
Upper Midwest
Rockies
Chromosome 8 (100 Kb bins)
Over-Represented in subpopulation
Under-represented in subpopulation

Found 1679 (4.37%) genes not found in reference

Note: 38,334 gene models in Reference

APPLICATIONS – Gene Function Annotation

1) Bean Seed Coat color



Cranberry, Dark Red Kidney, Light Red Kidney, Great Northern, Pinto, Small Red, Pink, Black, Navy

Heredity

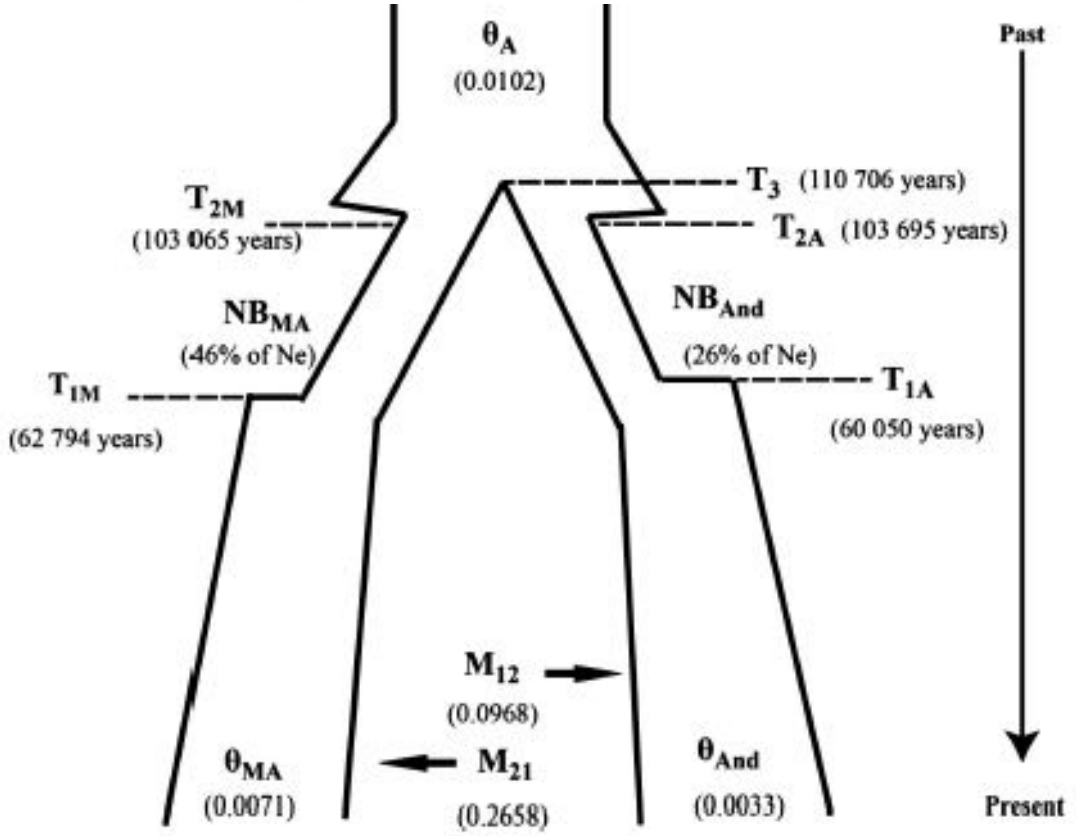
Heredity (2013) 110, 267–276

www.nature.com/hdy

ORIGINAL ARTICLE

Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L.

S Mamidi[1,2], M Rossi[3], SM Moghaddam[1,2], D Annam[4], R Lee[1,2], R Papa[3,5] and PE McClean[1,2]



igure 4 Parameter estimates for the wild gene pools of *Phaseolus vulgaris*.

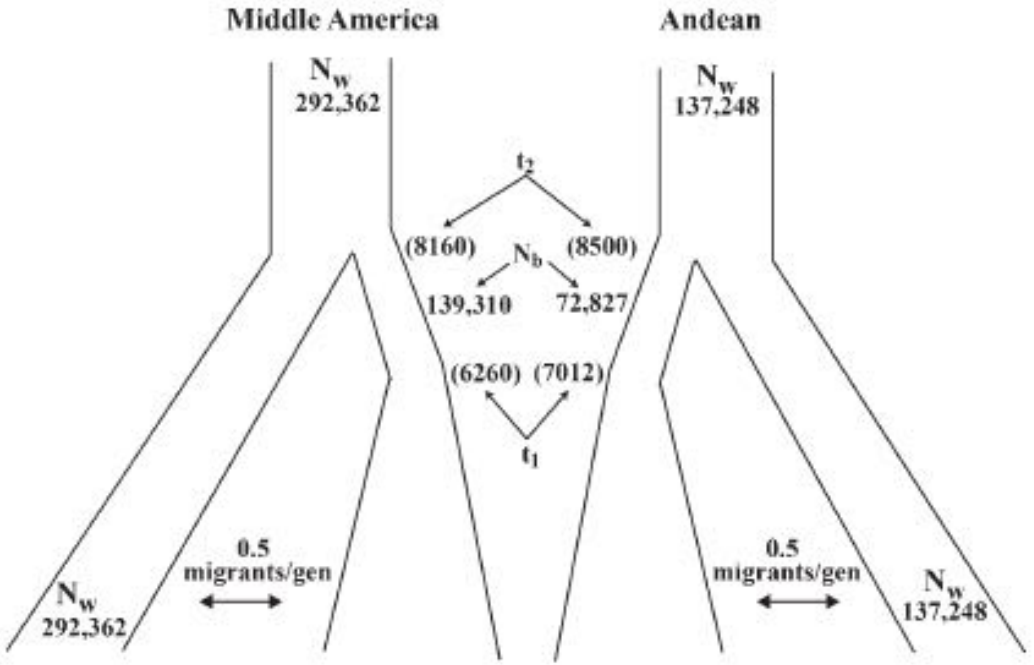
CSIRO PUBLISHING

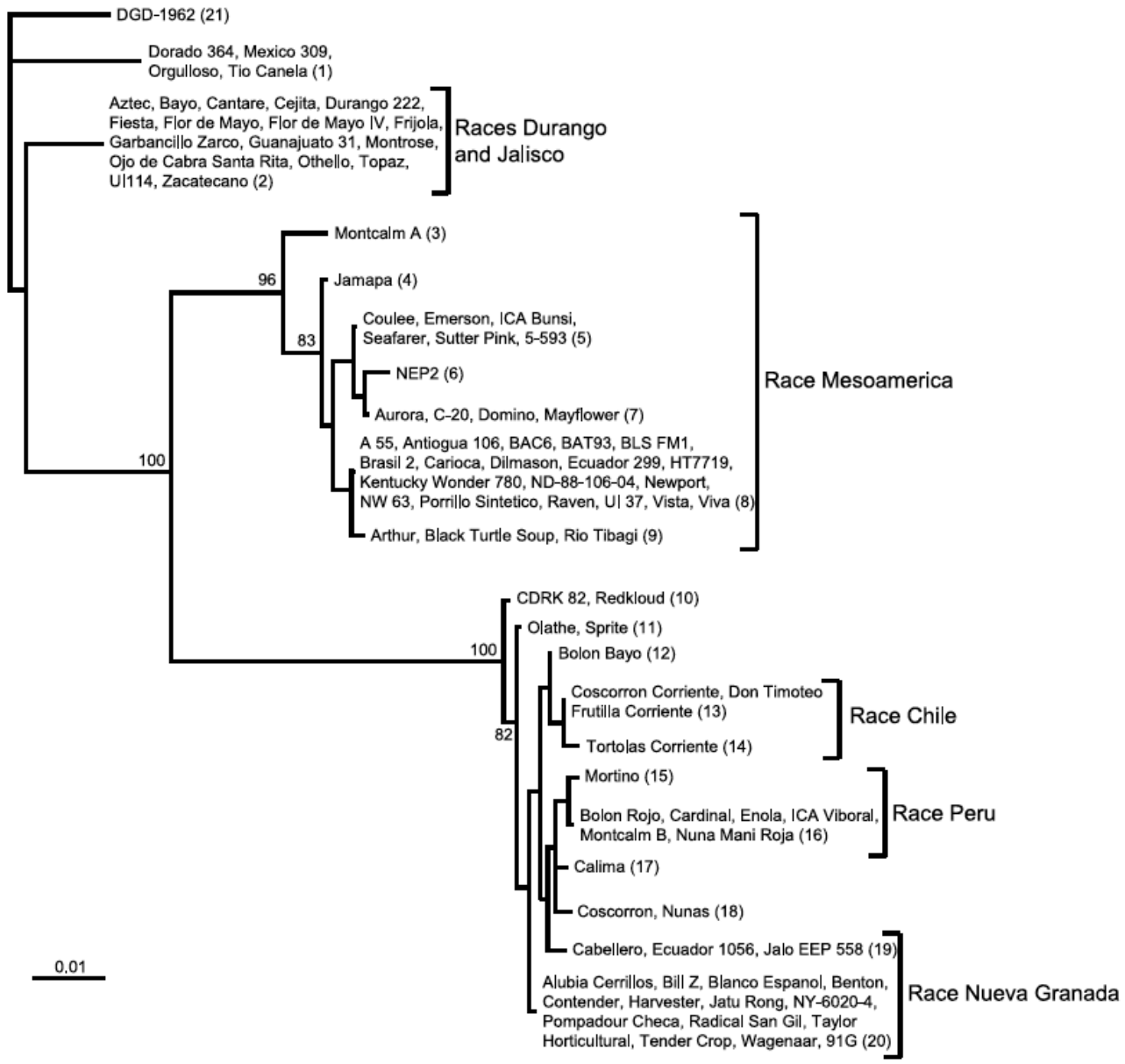
Functional Plant Biology, 2011, 38, 953–967

http://dx.doi.org/10.1071/FP11124

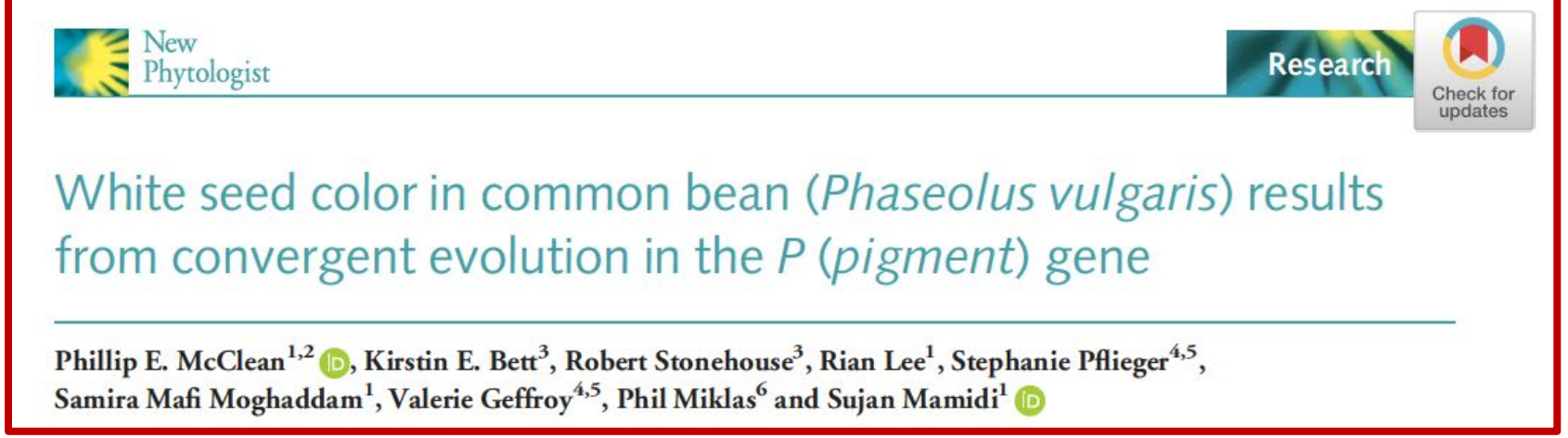
Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data

Sujan Mamidi[A,D,E], *Monica Rossi*[B], *Deepti Annam*[C], *Samira Moghaddam*[A,D], *Rian Lee*[A,D], *Roberto Papa*[B] *and Phillip McClean*[A,D]





Race	Market Classes
MA	
Mesoamerican	Black, Navy, Pea, Preto, Carioca
Durango	Pinto, Great Northern, medium red
Jalisco	Small red, Red Mexican, pink
Andean	
Nueva Granada	Kidney, cranberry, Snap, Canadian wonder
Peru	Yellow, Jalo, Bayo, Canario
Chile	Vine cranberry, Coscorron, Pampodour



New Phytologist

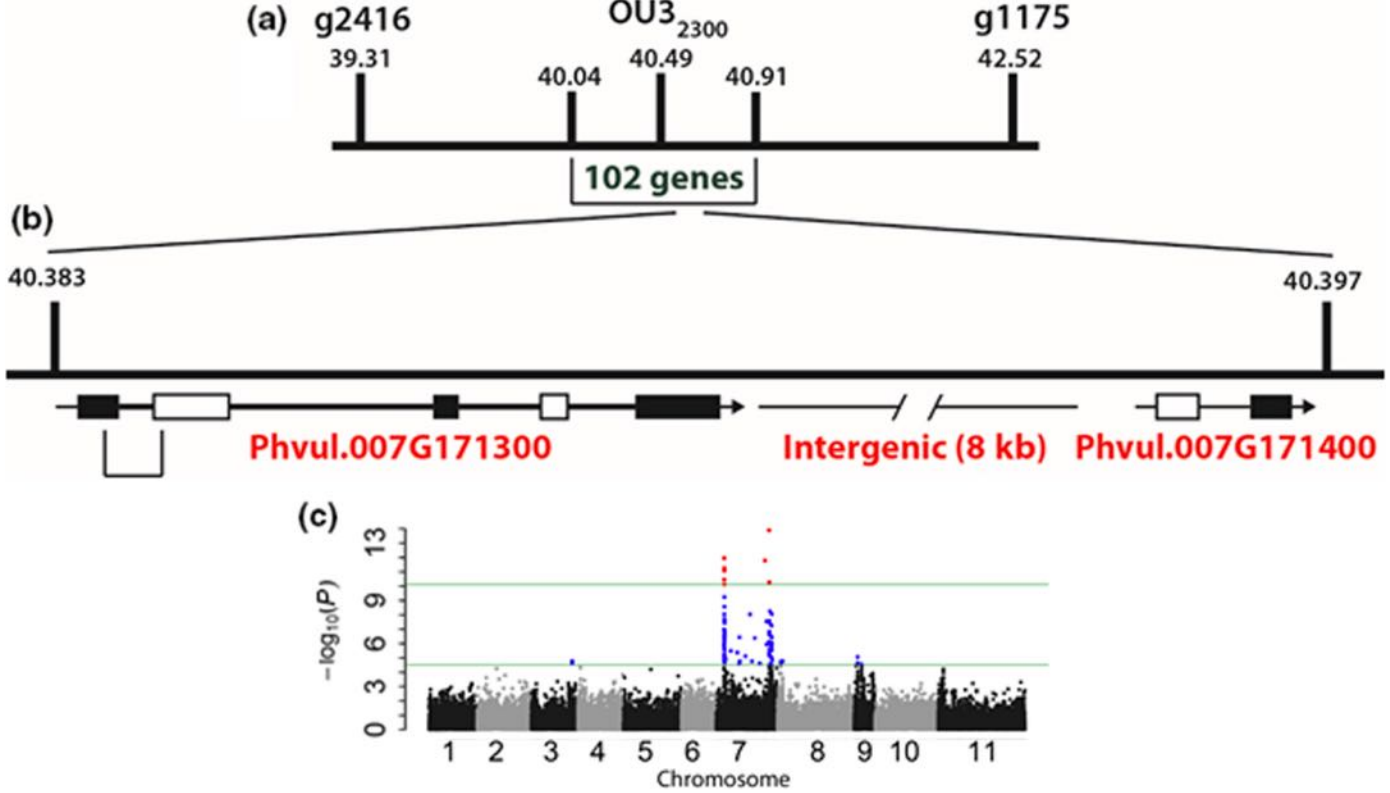
Research

Check for updates

White seed color in common bean (*Phaseolus vulgaris*) results from convergent evolution in the *P* (*pigment*) gene

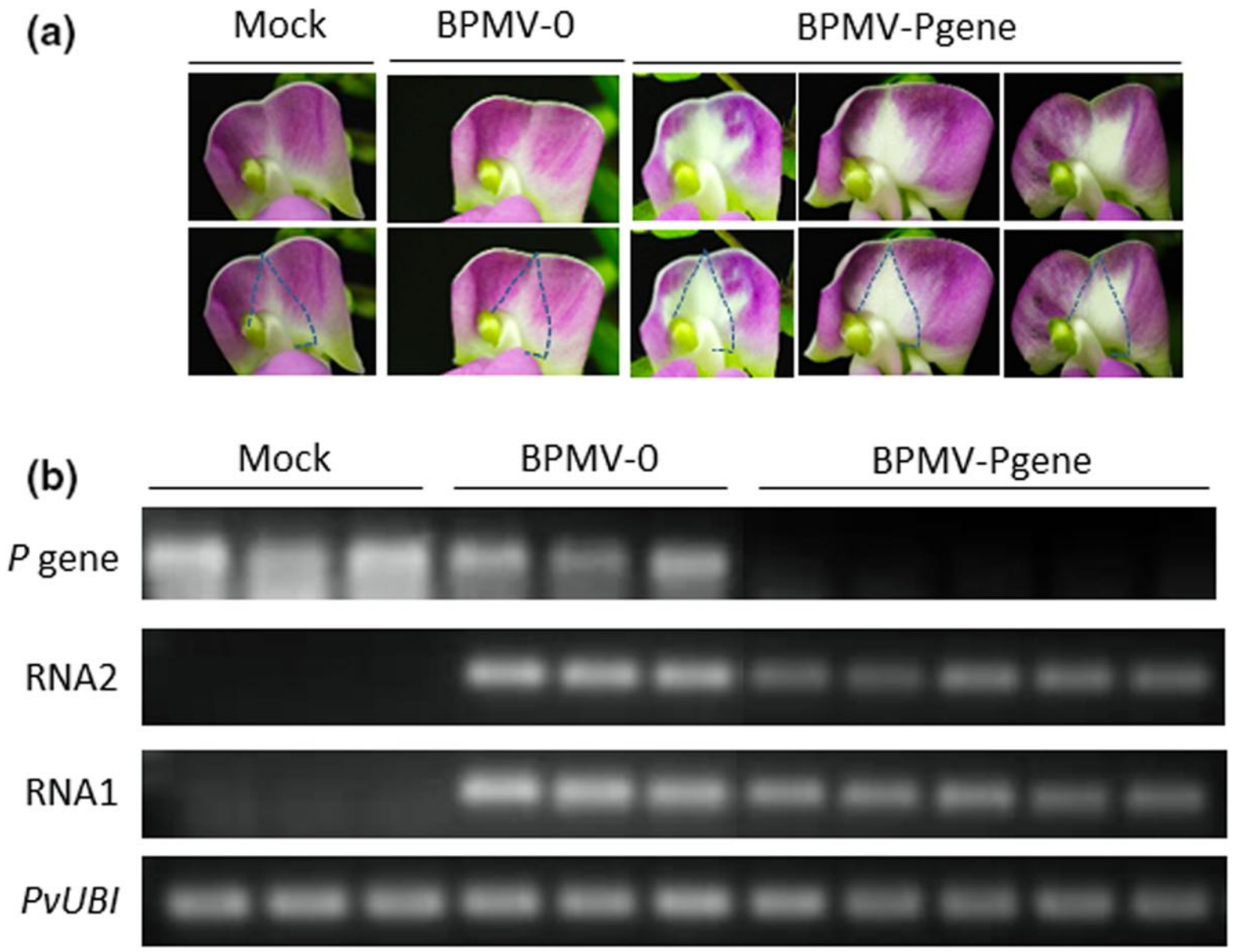
Phillip E. McClean[1,2], Kirstin E. Bett[3], Robert Stonehouse[3], Rian Lee[1], Stephanie Pflieger[4,5], Samira Mafi Moghaddam[1], Valerie Geffroy[4,5], Phil Miklas[6] and Sujan Mamidi[1]

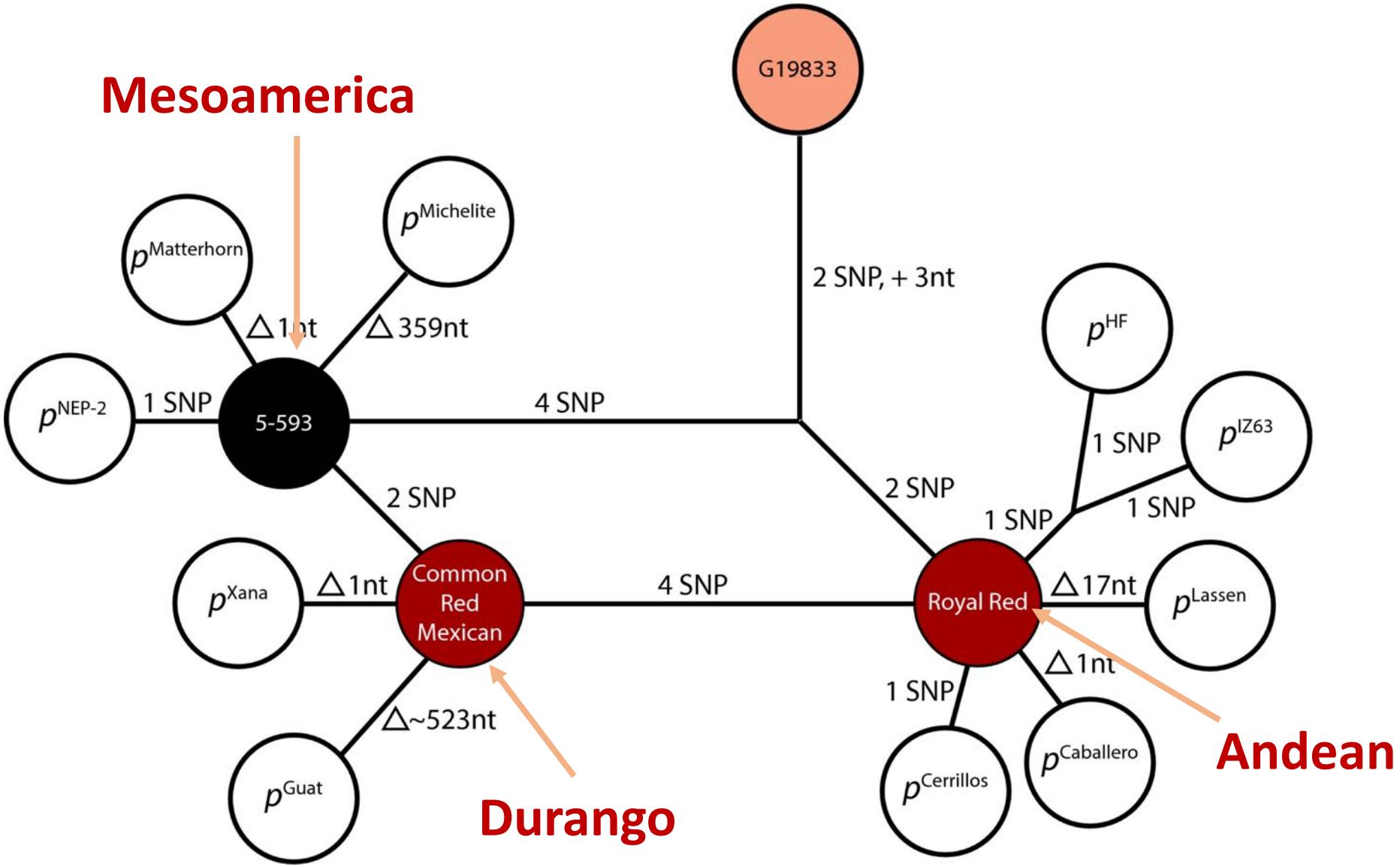
- The recessive p allele is pleiotropic to other genes in the network
- Homozygous pp produce white seeds together with white flowers.
- Four other P alleles control the spatial expression of color in seeds and flowers to various extents and are intermediate between the P and p alleles in the allelic series.
- All wild beans have colored seeds, whereas white-seeded landraces and cultivars are found in all major races of common bean



- Two tandem gene models (Phvul.007G171300, Phvul.007G171400)
- protein domain structures are similar to bHLH proteins
- Phvul.007G171300 similar to N-terminus of Arabidopsis AtTT8 and pea A proteins
- Phvul.007G171400 was homologous to the C-terminal portion
- AtTT8 and A encode orthologous bHLH TF control presence/absence of seed coat color in Arabidopsis and pea
- RNA seq reads map to both gene models., So **P gene is a single gene model**

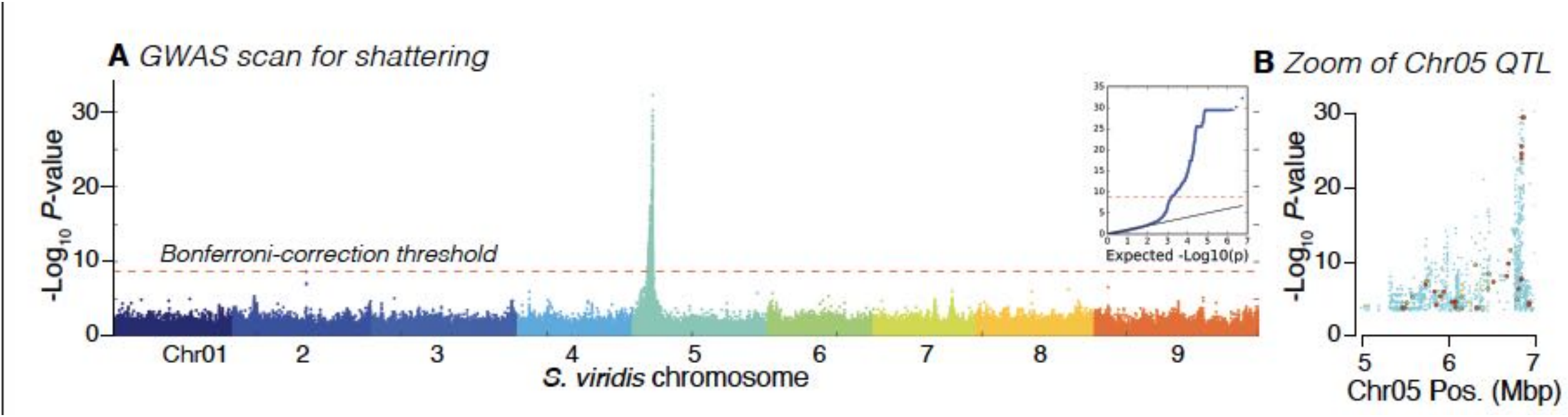
- BPMV VIGS vector - exon 8 insertion
- Black seed coat and pink–purple flowers,
- Infected 11 d after planting.
- Four to five weeks later
  - White sectors on dorsal petals gene silenced plants,
  - Mock- and empty vector-treated plants pink– purple in color.
- RT-PCR - P gene expression down-regulated in the white sectors.




Mesoamerica
G19833
2 SNP, + 3nt
pMatterhorn
pMichelite
Δ1nt
Δ359nt
pNEP-2
1 SNP
5-593
4 SNP
2 SNP
pHF
1 SNP
pIZ63
1 SNP
2 SNP
1 SNP
pXana
Δ1nt
Common Red Mexican
4 SNP
Royal Red
Δ17nt
pLassen
Δ1nt
1 SNP
Δ~523nt
pGuat
pCerrillos
pCaballero
Durango
Andean

2) SvLes1, controls seed shattering in S. viridis

- Sevir.5G085400 – Shattering genes by GWAS
- Encodes a MYB transcription factor
- G-T polymorphism (Chr_05:6849363) had larger effect
- This mutation leads to a R84S substitution in the second MYB domain of SvLes1

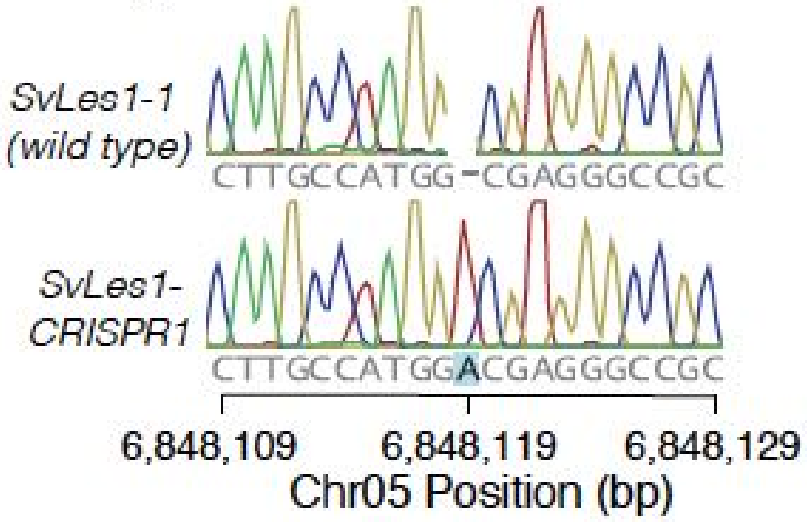


- Used CRISPR-Cas9 to create additional alleles.
- We disrupted the wild type, high-shattering allele SvLes1-1 to create several novel, non functional alleles.
- Sequence analysis of SvLes1-CRISPR1 257 revealed an adenine insertion at position 149 of the transcript, leading to a frameshift mutation
- Completely abolished gene function, creating non-shattering plants.
- After segregating out the transgenes phenotypically examined in the T3 generation (3:1 ratio)

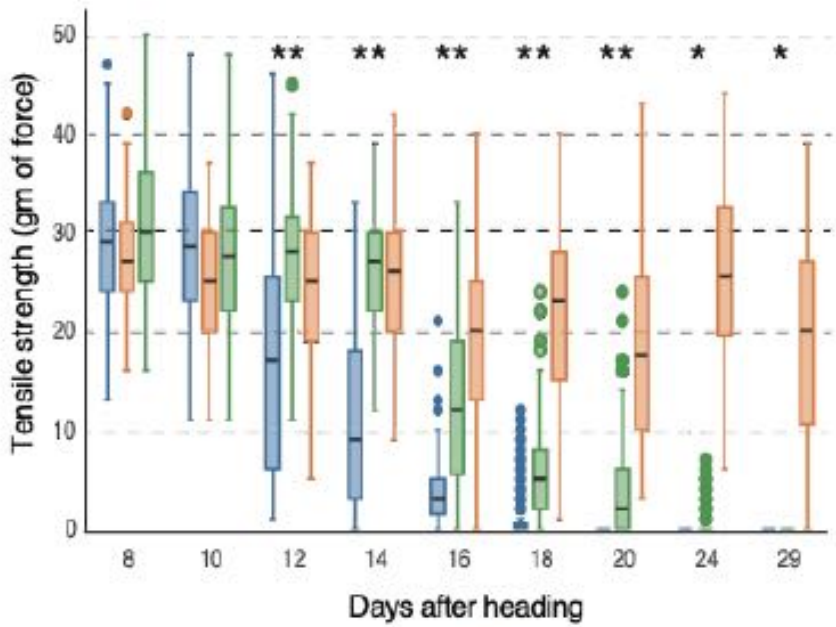
C *Summary of Les1 effects*

Allele	Structural characterization	Background line/ example line	Shattering
SvLes1-1	R at position 84	ME034v	High
SvLes1-2	R84S in second Myb binding domain	A10.1	Slightly reduced
SvLes1-CRISPR1	Insert A at transcript position 49; frameshift	ME034v CRISPR line	Low
SiLes1	*Copia* element between Myb domains	*S. Italica* 'Yugu'	Very low

D *Sanger validation of insertion*

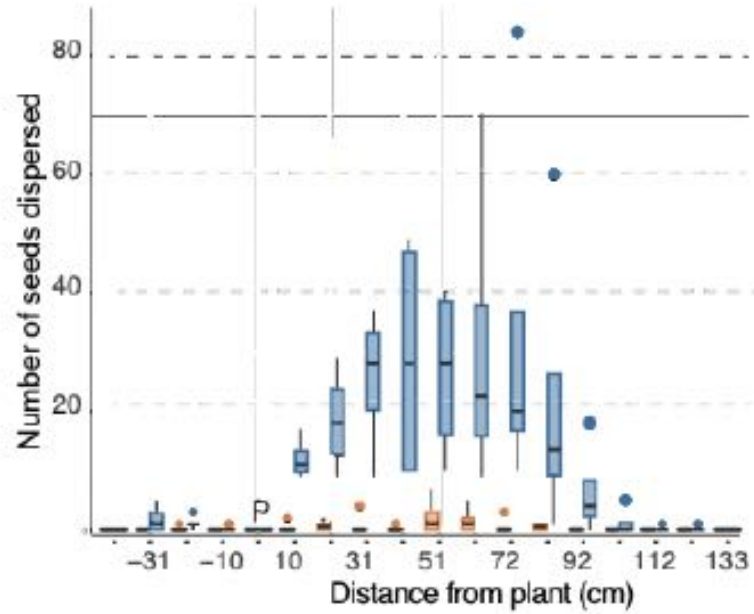


A *Tensile strength over time*



B *SvLes1-CRISPR1 and SvLes1-1 panicles*

C *Seed dispersal from plant*



SvLes1-1 (ME034v) *SvLes1-2 (A10.1)* *SvLes1-CRISPR1*

Acknowledgements

HudsonAlpha

- **Jeremy Schmutz**
- **Jane Grimwood**
- Jerry Jenkins
- Avinash Sreedasyam
- Adam Healey
- Chris Plott
- John Lovell
- Paul Gawbrowski
- Mike Frizell

Bean Group: NDSU

Phillip E McClean – PhD Advisor

Rian K Lee

Samira Mafi

Atena

Melody McConnell

Switchgrass

Thomas Juenger (& group)

University of Texas, Austin

Setaria:

Kellogg Elizabeth (& group)

Ivan Baxter

Donald Danforth Plant Science Center

Genome Sequencing Center
HudsonAlpha Institute of Biotechnology

Questions ?