Informatics Interest Group Meeting

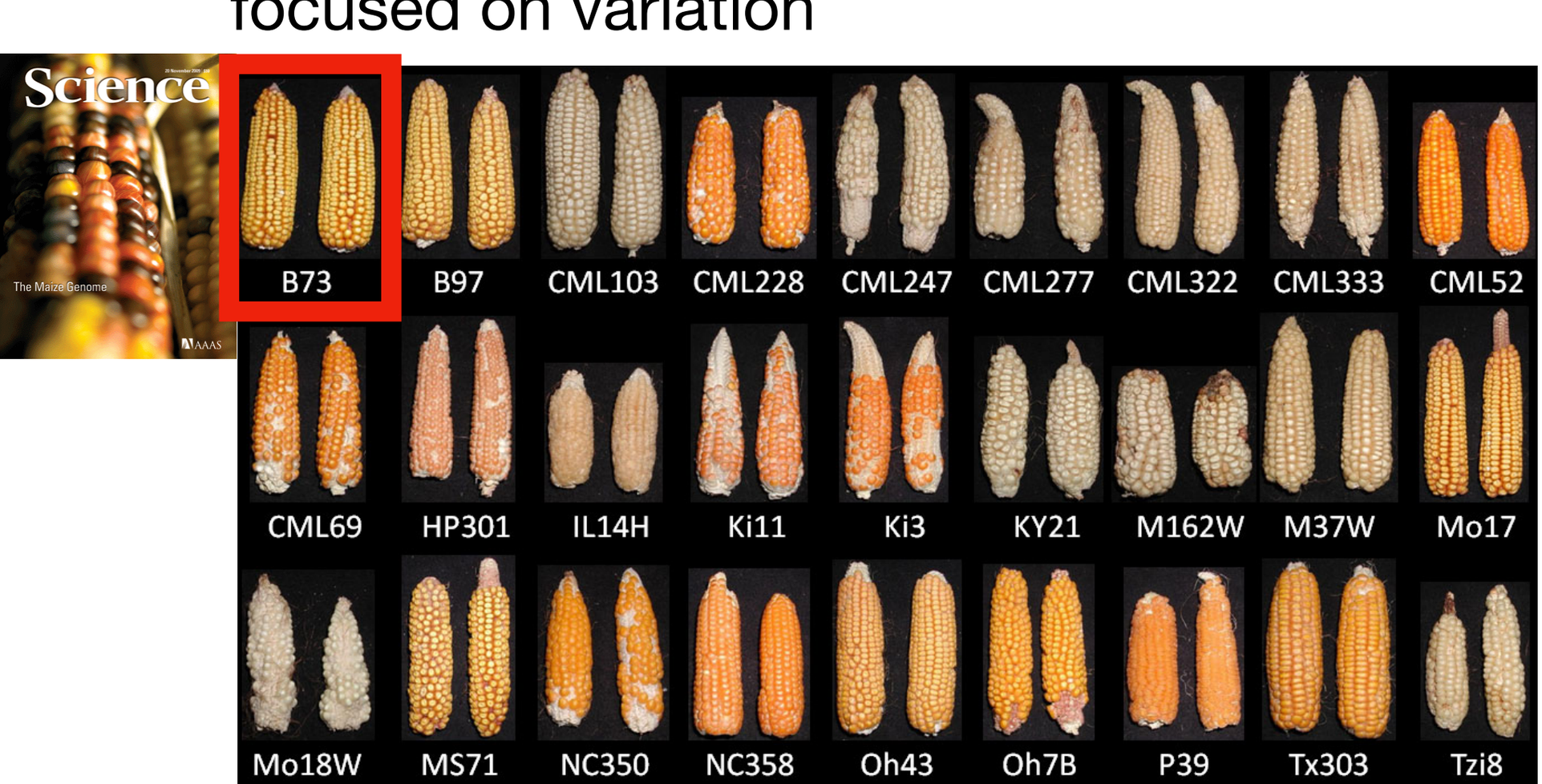
2018-05-02

Tom Kono

A Reference Genome

- String of nucleotides from a "representative" individual from a species or population

- Once a reference is assembled, questions become focused on variation



http://web.missouri.edu/~flint-garcias/pics.html

Limitations of A Single Reference

- Variation is widespread and most of it precludes discovery by comparison to a single genome

ARTICLE

OPEN

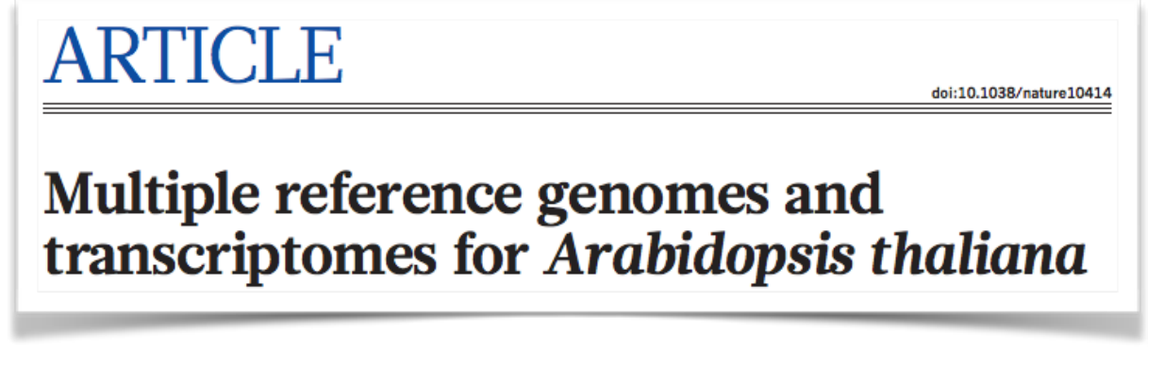
doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

The Plant Cell, Vol. 26: 121–135, January 2014, www.plantcell.org © 2014 American Society of Plant Biologists. All rights reserved.

LARGE-SCALE BIOLOGY ARTICLE

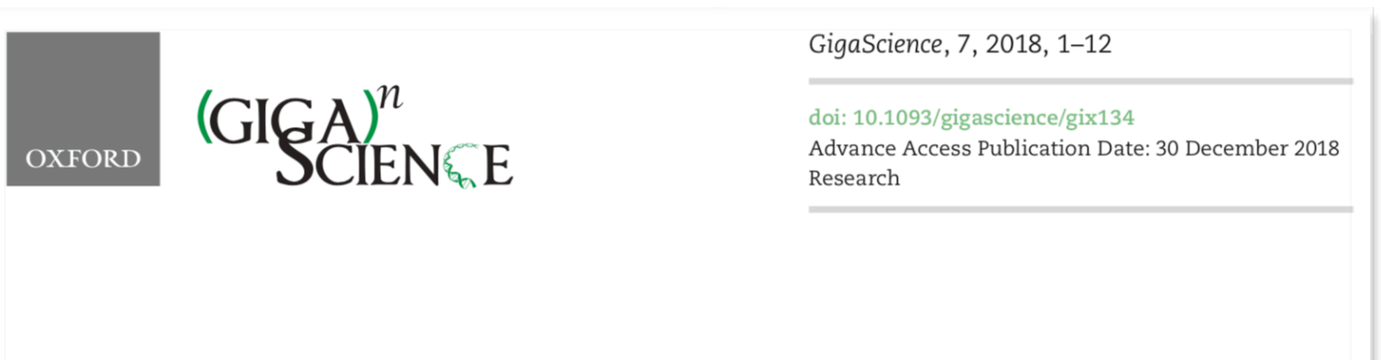
Insights into the Maize Pan-Genome and Pan-Transcriptome[W][OPEN]



ARTICLE

doi:10.1038/nature10414

Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*



OXFORD

(GIGA)n SCIENCE

GigaScience, 7, 2018, 1–12

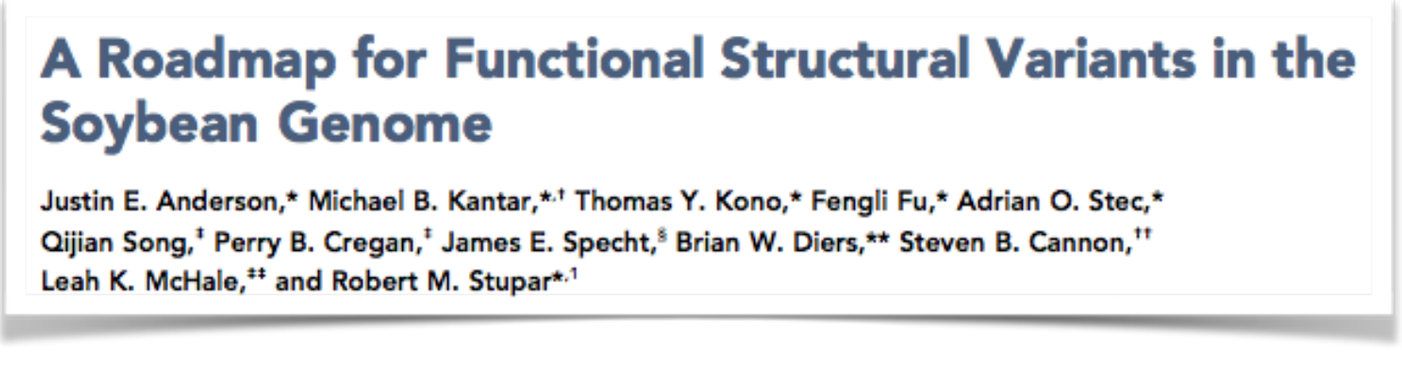
doi: 10.1093/gigascience/gix134

Advance Access Publication Date: 30 December 2018

Research

RESEARCH

Construction of the third-generation *Zea mays* haplotype map

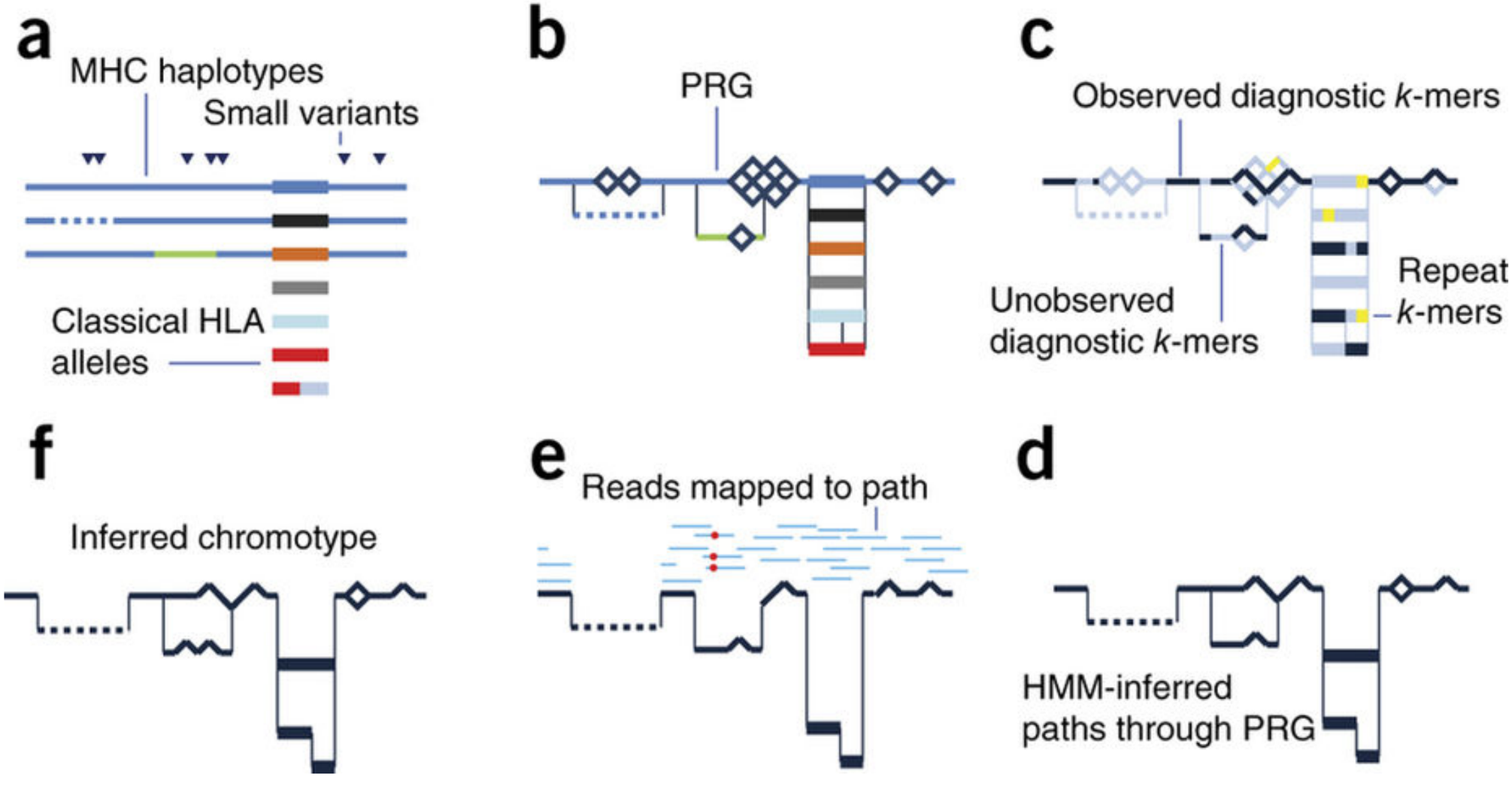


A Roadmap for Functional Structural Variants in the Soybean Genome

Justin E. Anderson,* Michael B. Kantar,*,† Thomas Y. Kono,* Fengli Fu,* Adrian O. Stec,* Qijian Song,‡ Perry B. Cregan,‡ James E. Specht,§ Brian W. Diers,** Steven B. Cannon,†† Leah K. McHale,‡‡ and Robert M. Stupar*,1

Potential Solution: de Bruijn Graphs

- Borrow from sequence assembly techniques to represent diversity



Dilthey et al. 2014, Nat Genet
doi:10.1038/ng.3257

Enter: Variation Graphs

- Represent genomes as a graph, rather than a string or series of strings

    - Invariant sites are single connections through graph

    - Variants are "bubbles" - alternate paths through graph

    - Individuals (or chromosomes) are then represented as a path through the graph

Natural Limits to the Problem

- Potential problem: the number of possible haplotypes

  - ~80m common SNPs in humans => ~$2^{80000000}$ haplotypes possible. Lower bound, because there are other types of polymorphisms, with many alleles...

- But! The number of observed haplotypes is *far fewer* than the number that are possible. Linkage disequilibrium*.

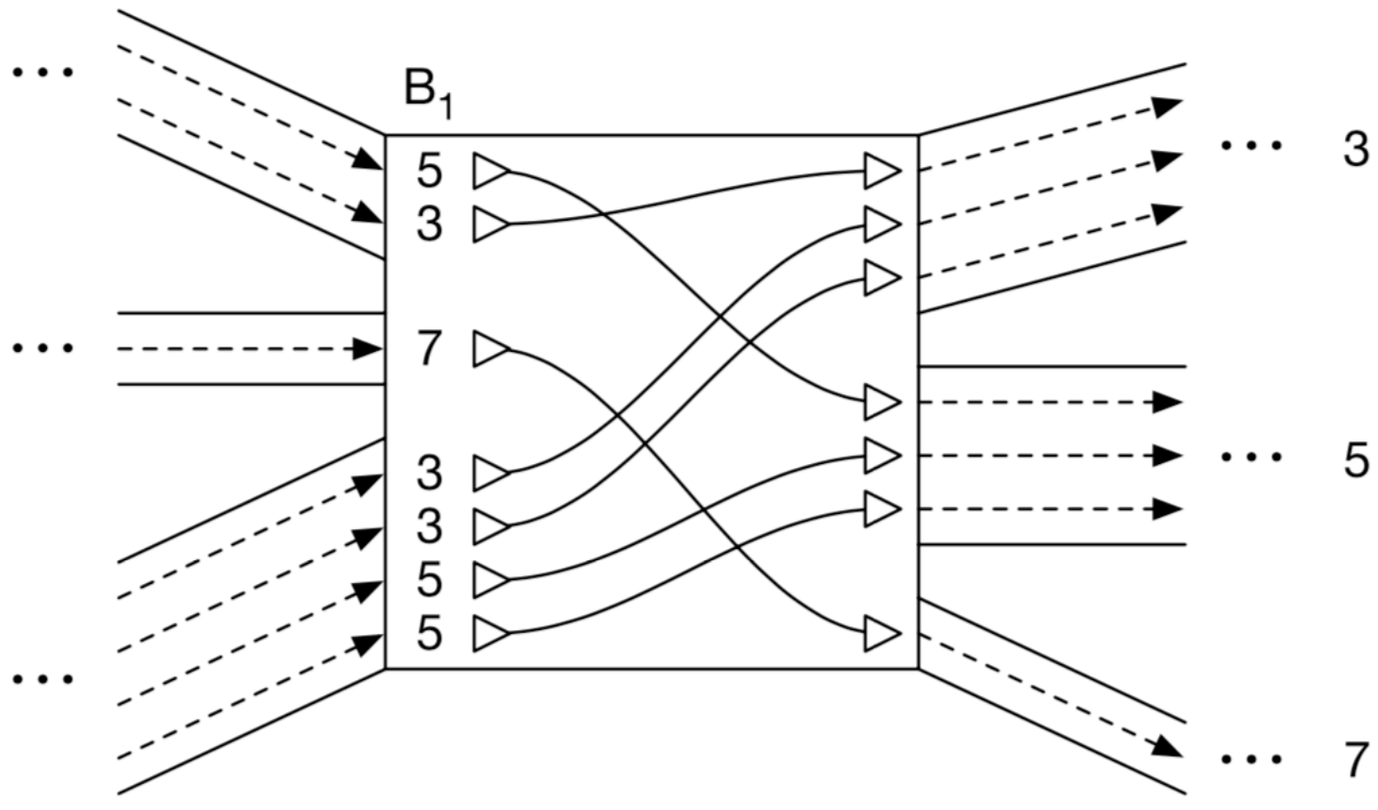
- Haplotypes in a genome graph: restricted paths through multiple variants

*Lewontin and Kojima 1960, *Evolution*

BWT Variant: gPBWT

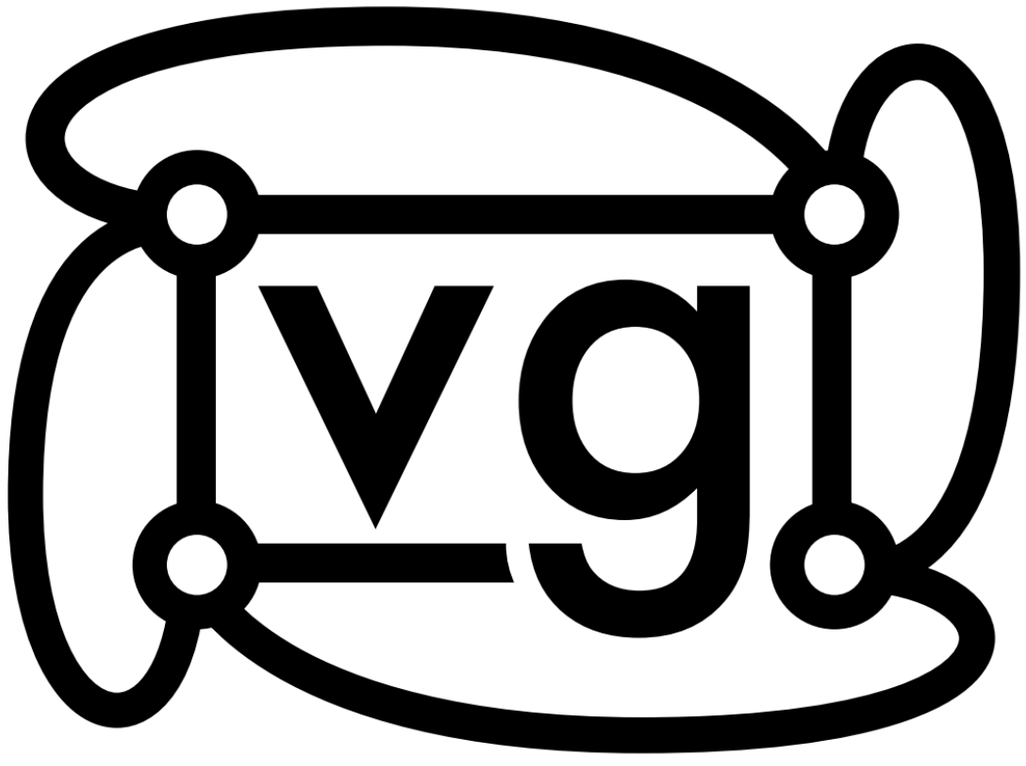
- Generalization of the Burrows-Wheeler Transform to contain positional information

  - Embeds the "threads" (linkages among variant sites) into the graph



Genome Graph Tools

- Erik Garrison (of vcflib and FreeBayes fame) is involved with the 'vg' project (variation graphs)
    - https://github.com/vgteam/vg



vg Capabilities

- vg is pretty complete now:
  - Build graph (from FASTA and VCF)
  - View/visualize graph (via graphviz)
  - Align reads to a graph (or map lots of reads)
  - Call variants

Running vg

- Building takes ~2.7 Gb on MacOS, including dependencies

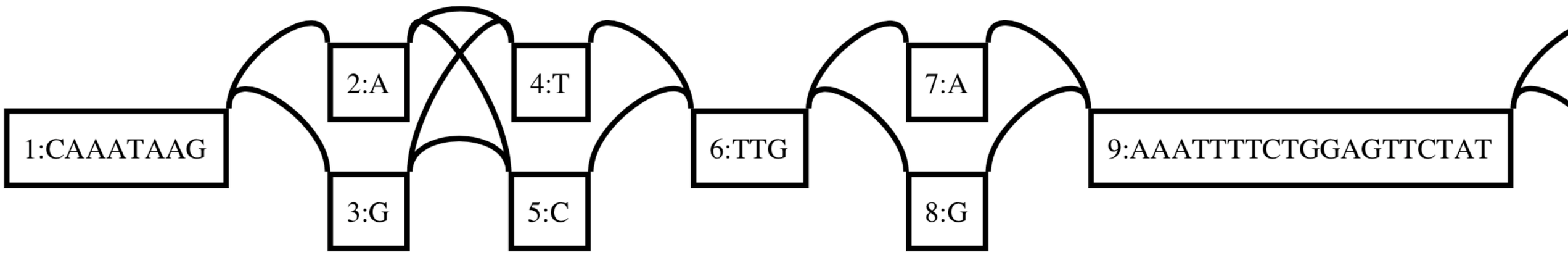
- Build graphs with `vg construct`

```
vg construct \
    -r test/small/x.fa \
    -v test/small/x.vcf.gz \
    > x.vg
```

Running vg

- Convert between formats with `vg view`. The `.dot` file is suitable for input into graphviz visualization software

```
vg view x.vg > x.gfa #default fmt is GFA1
vg view –d x.vg > x.dot
dot –Tpdf x.dot –o x.pdf
```



Running vg

- Can align and map reads to a graph: GAM (Graph Alignment/Map), looks like compressed JSON

```
vg index –x x.xg –g x.gcsa –k 16 x.vg
vg sim –n 1000 –l 150 –x x.xg > reads
vg map –x x.xg –g x.gcsa –T reads > aln.gam
vg view –a aln.gam | less
```


```
{"sequence":"GAGTCTCAAGGACAGCTCTCCCTTGTGTCCAGAGTGTATACGATTAACTCTGTTCTGGGCACTGGTGAAAGAAAACAGAGGAAATGCCTGGCTTTTTATCAGAACATGTTTCCAAGCTTATCCCTTT
TCCCTGCTCTCCTTGTCCCTCCT","path":{"mapping":[{"position":{"node_id":"120","offset":"59"},"edit":[{"from_length":3,"to_length":3}],"rank":"1"}
,{"position":{"node_id":"121"},"edit":[{"from_length":1,"to_length":1}],"rank":"2"},{"position":{"node_id":"123"},"edit":[{"from_length":17,
"to_length":17}],"rank":"3"},{"position":{"node_id":"124"},"edit":[{"from_length":1,"to_length":1}],"rank":"4"},{"position":{"node_id":"126"
},"edit":[{"from_length":6,"to_length":6}],"rank":"5"},{"position":{"node_id":"128"},"edit":[{"from_length":1,"to_length":1}],"rank":"6"},{"
position":{"node_id":"129"},"edit":[{"from_length":14,"to_length":14}],"rank":"7"},{"position":{"node_id":"131"},"edit":[{"from_length":1,"t
o_length":1}],"rank":"8"},{"position":{"node_id":"133"},"edit":[{"from_length":11,"to_length":11}],"rank":"9"},{"position":{"node_id":"134"}
,"edit":[{"from_length":1,"to_length":1}],"rank":"10"},{"position":{"node_id":"135"},"edit":[{"from_length":6,"to_length":6}],"rank":"11"},{
"position":{"node_id":"137"},"edit":[{"from_length":1,"to_length":1}],"rank":"12"},{"position":{"node_id":"138"},"edit":[{"from_length":9,"t
o_length":9}],"rank":"13"},{"position":{"node_id":"139"},"edit":[{"from_length":1,"to_length":1}],"rank":"14"},{"position":{"node_id":"141"}
,"edit":[{"from_length":15,"to_length":15}],"rank":"15"},{"position":{"node_id":"143"},"edit":[{"from_length":1,"to_length":1}],"rank":"16"}
,{"position":{"node_id":"144"},"edit":[{"from_length":42,"to_length":42}],"rank":"17"},{"position":{"node_id":"145"},"edit":[{"from_length":
1,"to_length":1}],"rank":"18"},{"position":{"node_id":"147"},"edit":[{"from_length":17,"to_length":17}],"rank":"19"},{"position":{"node_id":
"148"},"edit":[{"from_length":1,"to_length":1}],"rank":"20"}]},"mapping_quality":60,"score":160,"identity":1,"refpos":[{"offset":"625","name
":"x"}],"time_used":6113}
{"sequence":"CTGGGCACTGGTGAAAGATAACAGAGGAAATGCCTGGCTTTTTATCAGAACATGTTTCCAAGCTTATCCCTTTTCCCTGCTCTCCTTGTCCCTCCTACGATCTCTTCACTGGCCTTTTATCTTTACT
GTTACCAAATCTTTCCGGAACCT","path":{"mapping":[{"position":{"node_id":"133","offset":"10"},"edit":[{"from_length":1,"to_length":1}],"rank":"1"}
,{"position":{"node_id":"134"},"edit":[{"from_length":1,"to_length":1}],"rank":"2"},{"position":{"node_id":"135"},"edit":[{"from_length":6,"
```

Running vg

- Put it back into the linear world using `vg surject`

```
vg surject -x x.xg -b aln.gam | less -S
```

```
*       16      x       424     60      43M1D33M1I53M1I19M      *       0       0       ACCTTCCTTGACTTC
*       0       x       584     60      85M1I1M1D63M    *       0       0       ATTGCATCTCAAATCTAAGACCC
*       16      x       688     60      150M    *       0       0       TGGTGAAAGATAACAGAGGAAATGCCTGGCT
*       0       x       115     60      58M1I91M        *       0       0       TTGAAGTAACGTTTGACAATCTA
*       0       x       34      60      150M    *       0       0       AATATTCCAACTCTCTGGGTCCTGGTGCTAT
*       0       x       793     60      68M1I81M        *       0       0       CTTTTATCTTTACTGTTACCAAA
*       0       x       555     60      114M1I1M1D34M   *       0       0       TCTGCTCACCGCGATCTTCAAGT
*       0       x       212     60      150M    *       0       0       TTTGCCACAGATTAGCCATGTGACTTTGAAC
*       16      x       347     60      150M    *       0       0       GAAGCTCAGGGAATAGTGCCTGGCATCGAGG
*       0       x       76      60      150M    *       0       0       AATGGTAATGGATATGTTGGGCTTTTTTCTT
*       16      x       335     60      150M    *       0       0       GGTGATGCTTGTGAAGCTGAGGGAATAGTGC
```

Running vg

- Augment the source graph with variation identified in the mapped reads

```
vg filter aln.gam \
    -r 0.9 -fu -s 2 -o 0 -D 999 -x x.xg \
    > flt.gam
vg augment x.vg flt.gam \
    -q 10 -S aug_graph.support \
    -Z aug_graph.trans -A aug_alignment.gam \
    > aug_graph.vg
```

Running vg

- Call variants based on support for various paths through the graph

```
vg paths -v aug_graph.vg -L # Gives path names
vg call \
-b x.vg -s aug_graph.support \
-x aug_graph.trans -r x > calls.vcf
```

```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=XDP,Number=2,Type=Integer,Description="Expected Local and Global Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=XADL,Number=1,Type=Float,Description="Likelihood of allelic depths for called alleles">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Forward and reverse support for ref and alt alleles.">
##FORMAT=<ID=XAAD,Number=1,Type=Integer,Description="Alt allele read count.">
##FORMAT=<ID=AL,Number=.,Type=Float,Description="Allelic likelihoods for the ref and alt alleles in the order listed">
##contig=<ID=x,length=1001>
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
x       533     223_315_226-223_322_226 T       A       2370    PASS    DP=161;SVLEN=0;XSEE=112,113,114,115;XREF        GT:DP:XDP:AD:XADL:SB
x       813     338_376_286-338_388_286 A       G       2550    PASS    DP=170;SVLEN=0;XSEE=158,159,160,161;XREF        GT:DP:XDP:AD:XADL:SB
x       553     226_272_325-226_325     T       TG      2250    PASS    DP=155;SVLEN=1;XSEE=115,116,117;XREF    GT:DP:XDP:AD:XADL:SB:XAAD
x       272     320_211_289_214-320_342_263_214 TA      CG      1470    PASS    DP=103;SVLEN=0;XSEE=55,56,57,58,59,60;XREF      GT:DP:XDP:AD
x       823     286_393_220-286_353_220 A       G       2340    PASS    DP=169;SVLEN=0;XSEE=161,162,163,164;XREF        GT:DP:XDP:AD:XADL:SB
x       10      303_410_319_216-303_410_229_216 C       T       150     PASS    DP=11;SVLEN=0;XSEE=1,2,3,4,5,6;XREF     GT:DP:XDP:AD:XADL:SB
x       566     325_253_294-325_352_294 T       C       2220    PASS    DP=155;SVLEN=0;XSEE=117,118,119,120;XREF        GT:DP:XDP:AD:XADL:SB
x       277     214_384_332-214_336_332 A       C       2940    PASS    DP=195;SVLEN=0;XSEE=60,61,62,63;XREF    GT:DP:XDP:AD:XADL:SB:XAAD
```

Running vg

- Or, call them with FreeBayes-like algorithm

```
vg genotype -G aug_alignment.gam \
    -E -v aug_graph.vg -r x \
    > calls.vcf
```

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=XSBB,Number=1,Type=Integer,Description="Ultrabubble Bases">
##INFO=<ID=XSBN,Number=1,Type=Integer,Description="Ultrabubble Nodes">
##FORMAT=<ID=DP,Number=1,Type=Float,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Float,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=SB,Number=4,Type=Float,Description="Forward and reverse support for ref and alt alleles.">
##FORMAT=<ID=PL,Number=G,Type=String,Description="Log Likelihood">
##FORMAT=<ID=XAAD,Number=1,Type=Float,Description="Alt allele read count.">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
x       14      .       G       A       0       .       DP=22.000000;XSBB=24;XSBN=4     GT:DP:GQ:AD:SB:PL       1/0:22.000000:118.765437:10.
x       34      .       T       A       0       .       DP=46.000000;XSBB=25;XSBN=4     GT:DP:GQ:AD:SB:PL       1/0:46.000000:302.831892:22.
x       533     .       T       A       0       .       DP=165.000000;XSBB=31;XSBN=4    GT:DP:GQ:AD:SB:PL       1/0:165.000000:1164.848097:8
x       39      .       T       A       0       .       DP=52.000000;XSBB=18;XSBN=4     GT:DP:GQ:AD:SB:PL       1/0:52.000000:331.942302:28.
x       813     .       A       G       0       .       DP=173.000000;XSBB=28;XSBN=4    GT:DP:GQ:AD:SB:PL       1/0:173.000000:1252.677669:8
x       52      .       T       G       0       .       DP=60.000000;XSBB=20;XSBN=4     GT:DP:GQ:AD:SB:PL       1/0:60.000000:407.019473:29.
x       272     .       TA      CA,CG,TG        0       .       DP=198.000000;XSBB=57;XSBN=6    GT:DP:GQ:AD:SB:PL       3/2:198.000000:46.74
x       823     .       A       G       0       .       DP=172.000000;XSBB=14;XSBN=4    GT:DP:GQ:AD:SB:PL       1/0:172.000000:1165.534775:9
x       277     .       A       C       0       .       DP=199.000000;XSBB=14;XSBN=4    GT:DP:GQ:AD:SB:PL       1/0:199.000000:1451.961774:9
x       827     .       G       C       0       .       DP=171.000000;XSBB=35;XSBN=4    GT:DP:GQ:AD:SB:PL       1/0:171.000000:1181.072965:8
x       858     .       TC      CA,CC,TA        0       .       DP=176.000000;XSBB=35;XSBN=6    GT:DP:GQ:AD:SB:PL       3/1:176.000000:10.11
```

Additional vg Links

- gPBWT: https://github.com/vgteam/vg/wiki/Building-a-Graph-Positional-Burrows-Wheeler-Transform-(gPBWT)

- vg on CentOS6 and 7: https://github.com/vgteam/vg/wiki/Building-VG-on-CentOS-6.6-or-7--and-using-it

- vg with long reads: https://github.com/vgteam/vg/wiki/Long-read-assemblies-using-vg-msga

- file format ref: https://github.com/vgteam/vg/wiki/File-Formats

Graphtyper

- Tool to call variants from reads mapped to a variation graph

- Not compatible with vg, but looks like it does similar things

TECHNICAL REPORTS

nature genetics

Graphtyper enables population-scale genotyping using pangenome graphs

Hannes P Eggertsson[1,2], Hakon Jonsson[1], Snaedis Kristmundsdottir[1,3], Eirikur Hjartarson[1], Birte Kehr[1,4], Gisli Masson[1], Florian Zink[1], Kristjan E Hjorleifsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Ingileif Jonsdottir[1,5], Daniel F Gudbjartsson[1,2], Pall Melsted[1,2], Kari Stefansson[1,5] & Bjarni V Halldorsson[1,3]

Seven Bridges Genomics

- Proprietary data formats and variant caller ($$$)
- Human specific?
- https://www.sevenbridges.com/graph/

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME

Search

New Results

Fast and Accurate Genomic Analyses using Genome Graphs

Goran Rakocevic, Vladimir Semenyuk, James Spencer, John Browning, Ivan Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C. Suciu, Sun-Gou Ji, Gulfem Demir, Lizao Li, Berke C. Toptas, Alexey Dolgoborodov, Bjoern Pollex, Iosif Spulber, Irina Glotova, Peter Komar, Andrew Stachyra, Yilong Li, Milos Popovic, Wan-Ping Lee, Morten Kallberg, Amit Jain, Deniz Kural

doi: https://doi.org/10.1101/194530