

Genome Graphs



Yosuke Tanigawa (ytanigaw@stanford.edu)

Biomedical Informatics Ph.D. Program

23 Jan. 2017 Journal Club

Adam M Novak, Glenn Hickey, Erik Garrison, et al.

Genome Graphs (2017 bioRxiv preprint <https://doi.org/10.1101/101378>)

Genome graphs (in summary)

- Hypothesis: *“adding data into the reference structure—in effect, adding to the “reference prior” on variation extant in the population—will result in improved genome inferences”?*
- Mapping & Variant calling performance
5 selected regions & 8 graphs
- Found: most of the variants are captured in the graphs
- Conclusion: Graph genome \geq linear genome

Sequence graphs

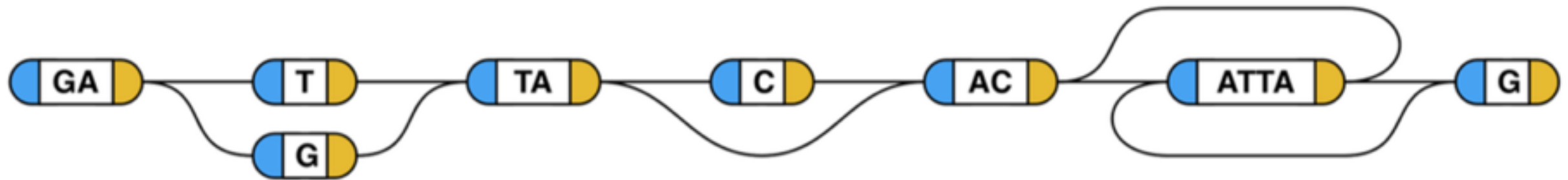


Figure 1: Example sequence graphs. Each node holds a string of bases. An edge can connect, at each of its ends, to a base on either the left (5', blue) or the right (3', yellow) side of the node. When reading through a thread to form a DNA sequence, a valid walk must leave each node via the opposite side from that through which it was entered; a node's sequence is read as reverse-complemented if the node is entered on the 3' side. One thread that this graph spells out (reading from the left side of the leftmost sequence to the right side of the rightmost sequence, along the nodes drawn in the middle) is the sequence "GATTACACATTAG". Straying from this path, there are three variants available: a substitution of "G" for "T", a deletion of a "C", and an inversion of "ATTA". If all of these detours are taken, the sequence produced is "GAGTAACTAATG". All 8 possible threads from the leading G to the trailing G are allowed.

GA T TA C AC ATTA G
 GA G TA C AC ATTA G
 GA T TA - AC ATTA G ...

Bad example?

Inversion(ATTA) = ATTA ??

Selected regions (80 kbp ~ 5 Mbp)

Region	Chromosome	Length in Primary Reference (bp)	GRCh38 Coordinates	Number of Genes	Alt Haplotypes in pilot data
BRCA1	17	81,189	43044293-43125482	1	2
BRCA2	13	84,989	32314860-32399849	1	2
LRC_KIR	19	1,058,685	54025633-55084318	47	35
MHC	6	4,970,458	28510119-33480577	172	8
SMA	5	2,397,625	69216818-71614443	21	2

Table 1: Pilot Regions. Selected test cases represent a sampling of both typical and challenging genomic regions.

Typical: BRCA1, BRCA2

Challenging: LRC_KIR, MHC, SMA

Benchmarked graphs (part 1: Mostly from aligner)

Submissions using pilot data			
Submission	Team	Short Name	Description of Algorithm
Cactus	UCSC	Cactus	Graph-based multiple sequence aligner ²³ .
Camel	UCSC	Camel	Creates graphs progressively by mapping using context schemes ²⁴ .
De Bruijn Graph (k=63)	MSKCC	De Bruijn 63	Forms a De Bruijn graph of input data with k=63, then converts to a sequence graph.
Population Reference Graph	Oxford	PRG	Creates a graph from a K-mer-based HMM description of a region ¹⁹ .
Seven Bridges	Seven Bridges	7BG	Multiple genome alignment.

Table 2: Genome Graph Submissions. Submissions were collected from a variety of institutions, and showcase a variety of graph construction methods.

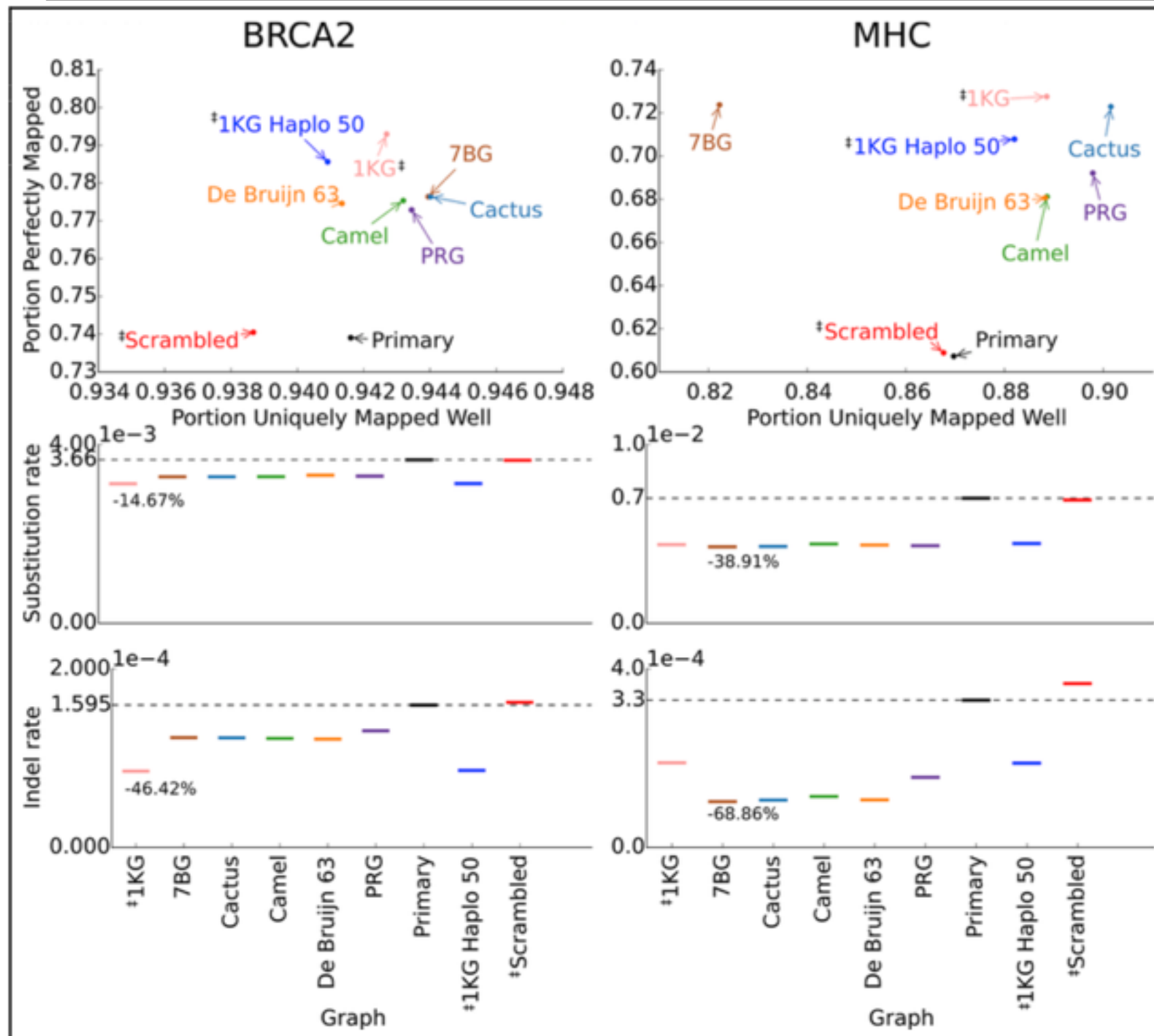
Benchmarked graphs (part 2: 1KG based graphs)

Submissions using other data			
Submission	Team	Short Name	Description of Algorithm
1000 Genomes SNP Graph	Sanger/UCSC	1KG	Generated using vg construct on a VCF containing variants identified in the 1000 genomes project. Platinum genome samples were not included, to avoid circularity in variant evaluation.
1000 Genomes Haplotype 50	Sanger/UCSC	1KG Haplo 50	Adapted form of 1KG graph in which phasing information is used to reduce the number of unobserved recombinations represented by the graph. 50 is the number of bases two variants can be apart to be considered for this phasing.
Scrambled 1000 Genomes	Sanger/UCSC	Scrambled Control	Generated by shifting all the variants in the standard 1KG graph 200 bp downstream.

Table 2: Genome Graph Submissions. Submissions were collected from a variety of institutions, and showcase a variety of graph construction methods.

Primary: linear graph (GRCh38)

Mapping performance



reference  then

- perfect map 

- substitution 

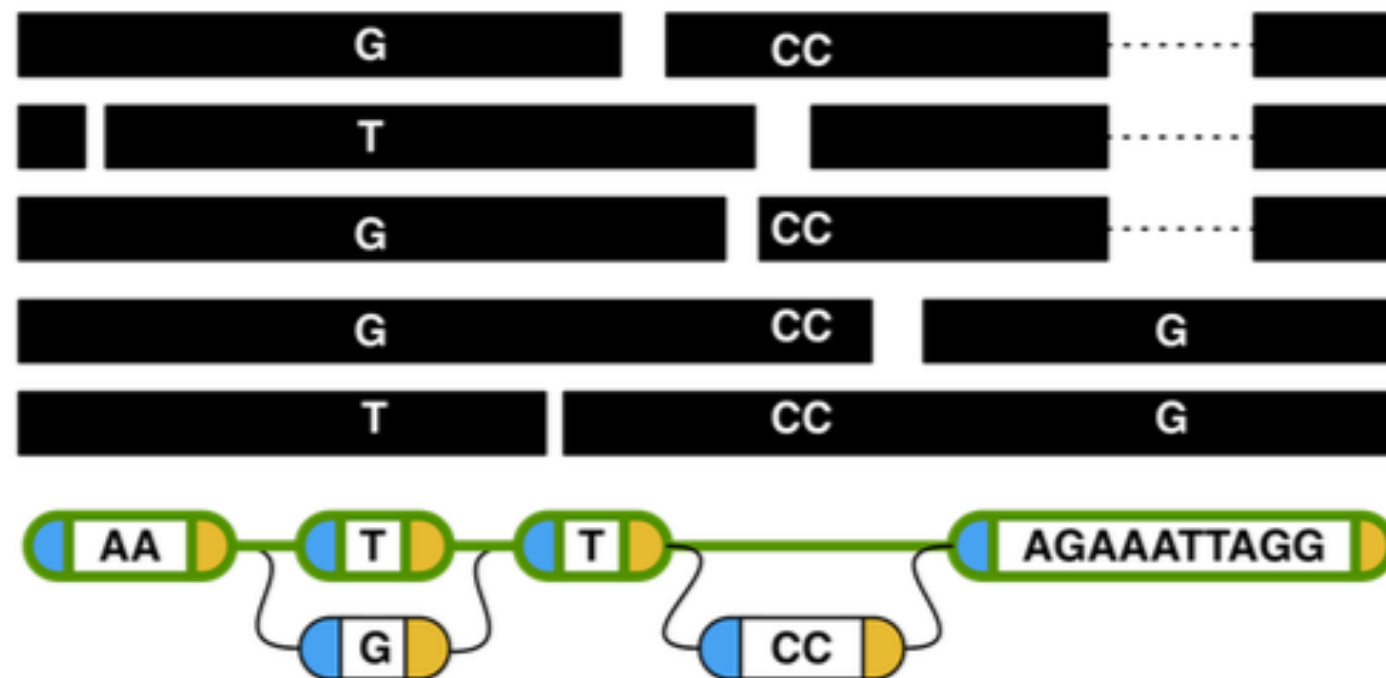
- indel 

Figure 2: Mapping reads to sequence graphs. Results for the 1000 Genomes Phase 3 low coverage samples against the BRCA2 and MHC graphs. The median per-sample portion of

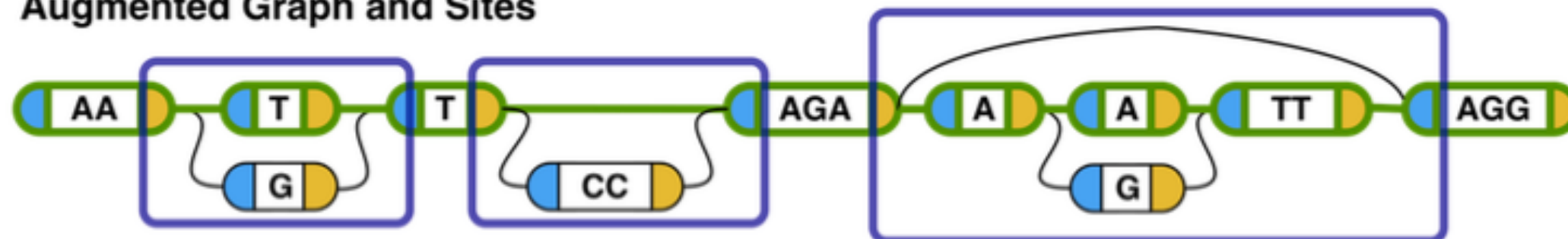
Variant Calling

- Map to base graph
- Pileup
- Augmented graph
- Path exists?
 - reference call
 - non-reference call

(A) Read Pileup on Base Graph



(B) Augmented Graph and Sites

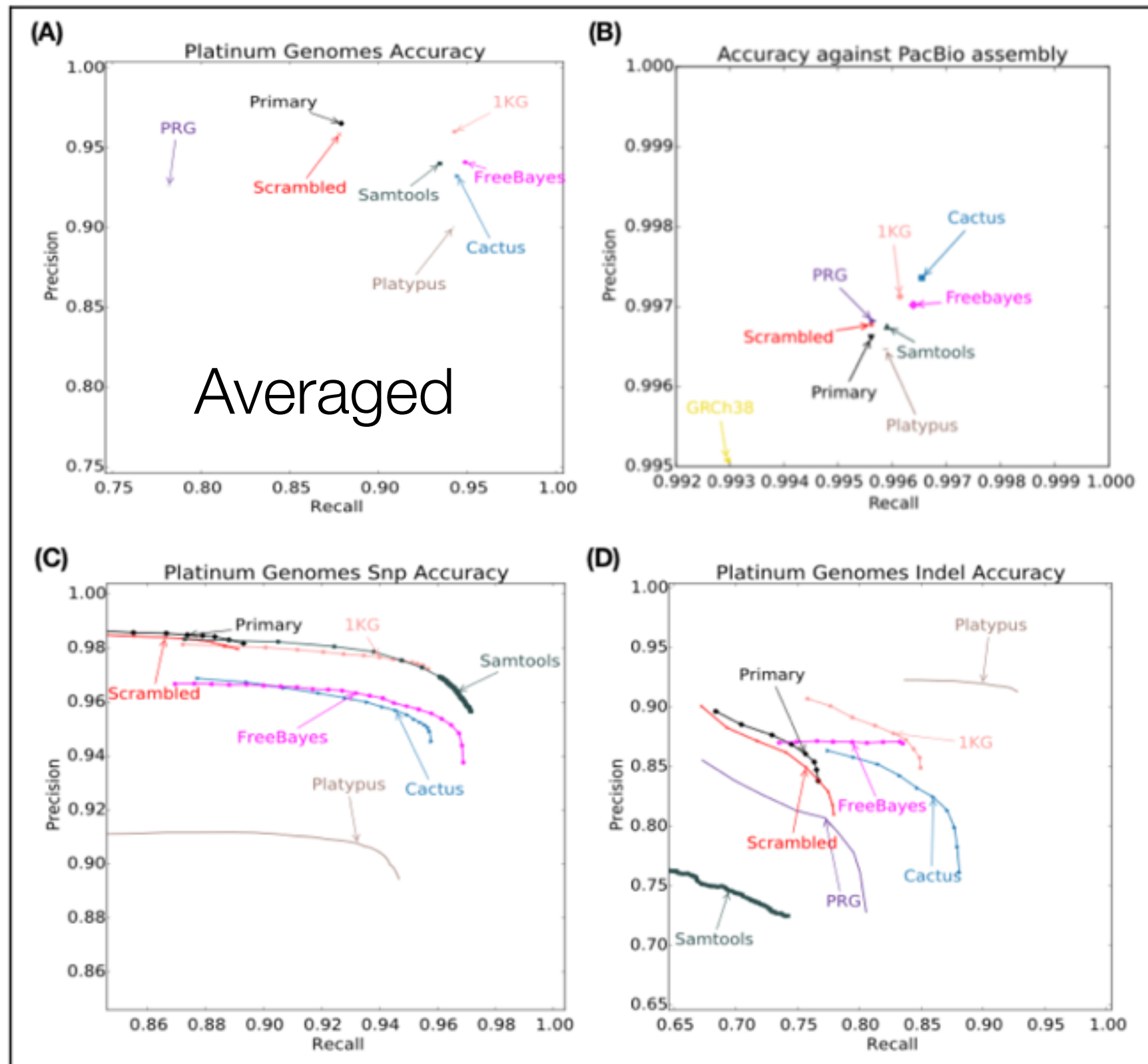


(C) Calls

Ref. Het. SNP:	Ref. Hom. Insertion:	Non-ref. Het SNP/Deletion:
T -> T/G	T -> TCC	AAATT->A/AAGTT

Figure 3: Variant Calling with Genome Graphs. (A) Read pileup on a base graph whose

Variant calling performance



Some graphs are better than GRCh38

Precision
= $TP / (TP + FP)$


Recall
= $TP / (TP + FN)$

Validation sets:

- NA12877 12878
- PacBio de novo

Figure 4: Variant Calling Evaluation. (A) Precision (portion of called variation in agreement with the truth) vs. Recall (portion of variation in the truth that is called).

Reference call vs. Non-reference call

- 'Reference' call
- path exists on graph
- precision of ref. call  (fig. 5B)
- Cactus: captures indels > read length

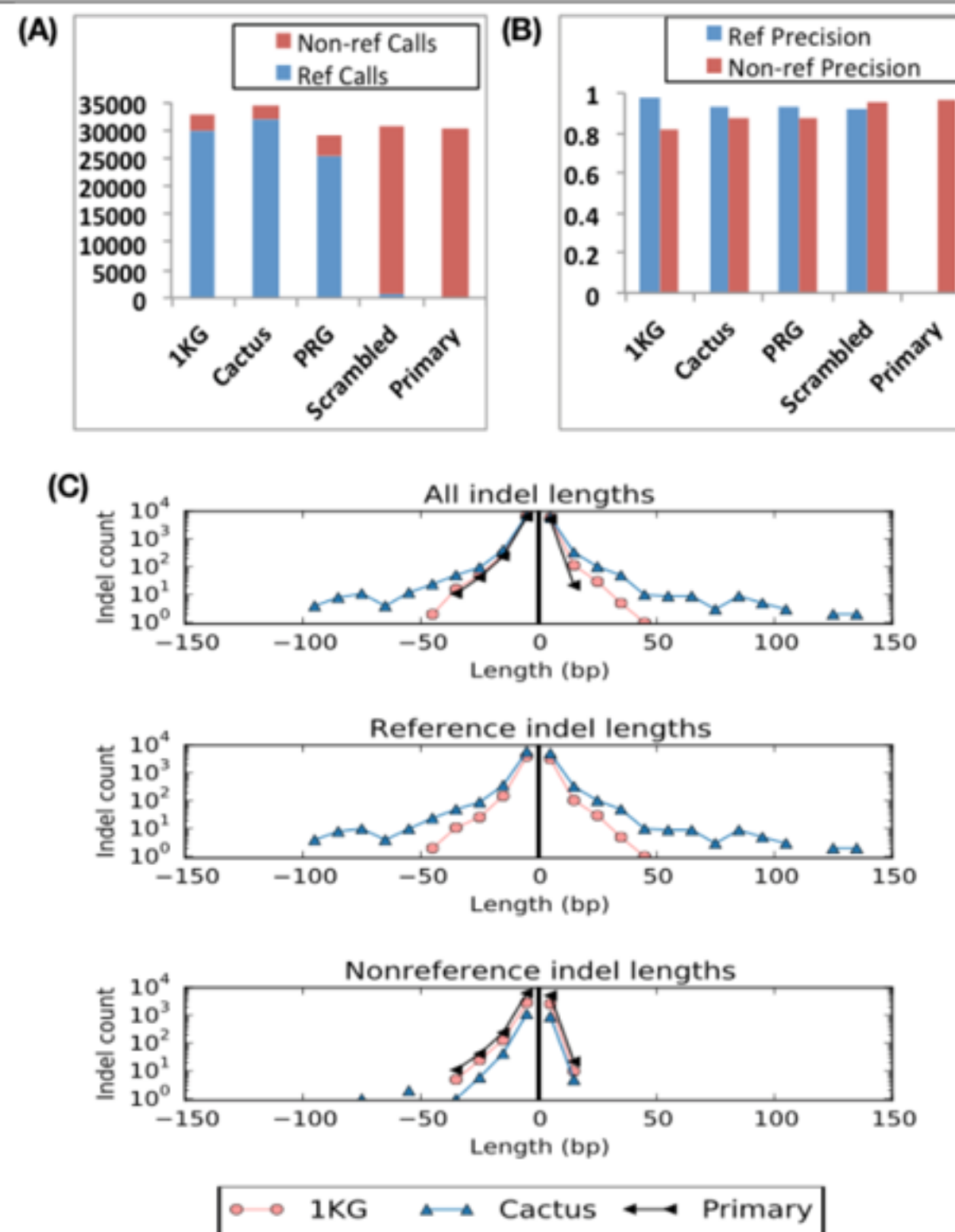


Figure 5: Reference versus Non-reference Calls. (A) Total number of reference and non-reference calls across all samples and regions. (B) Precision of reference and non-reference calls. (C) Indel lengths of reference and non-reference calls, where insertions and deletions are represented by positive and negative lengths, respectively. In all cases we ignore calls of GRCh38 reference alleles, as these numbers are reported from GRCh38-based output VCFs.

Short path completeness

“ k -mers in 1KG present in graph?”

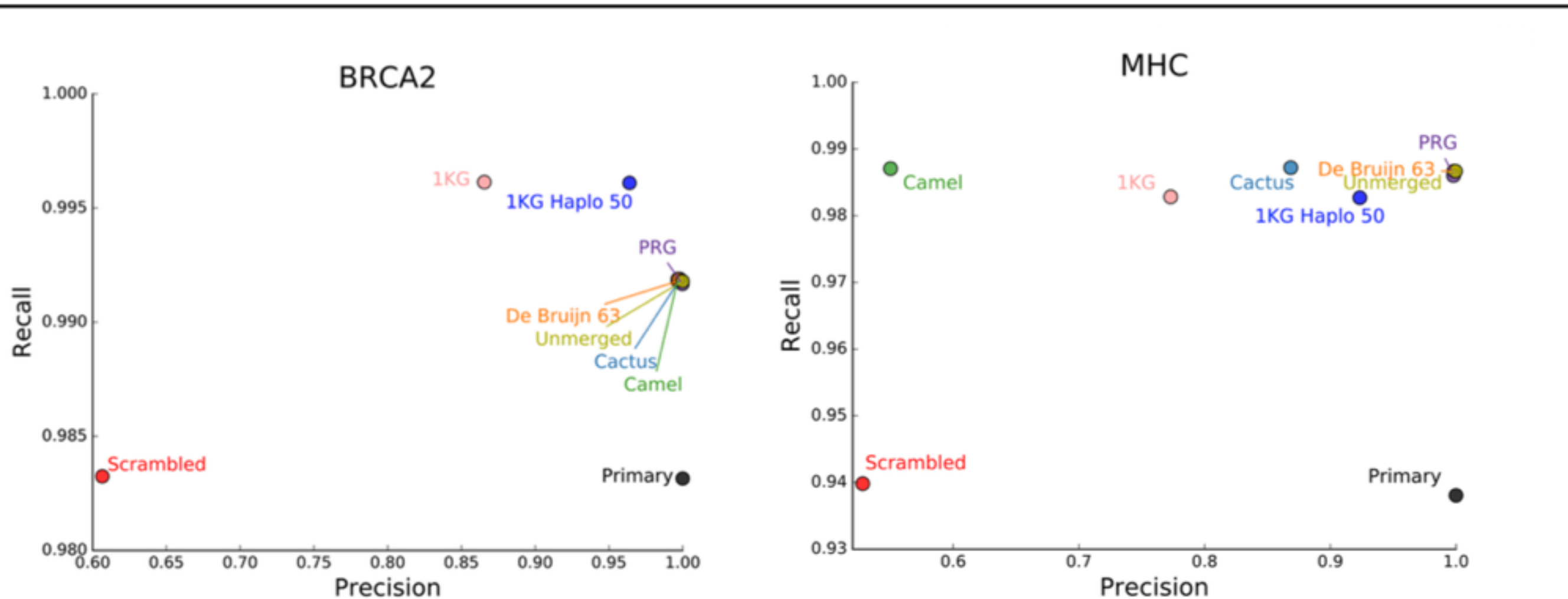


Figure 6: Short path completeness and accuracy. Assayed by comparing 20-mer instances.

Graph stats.

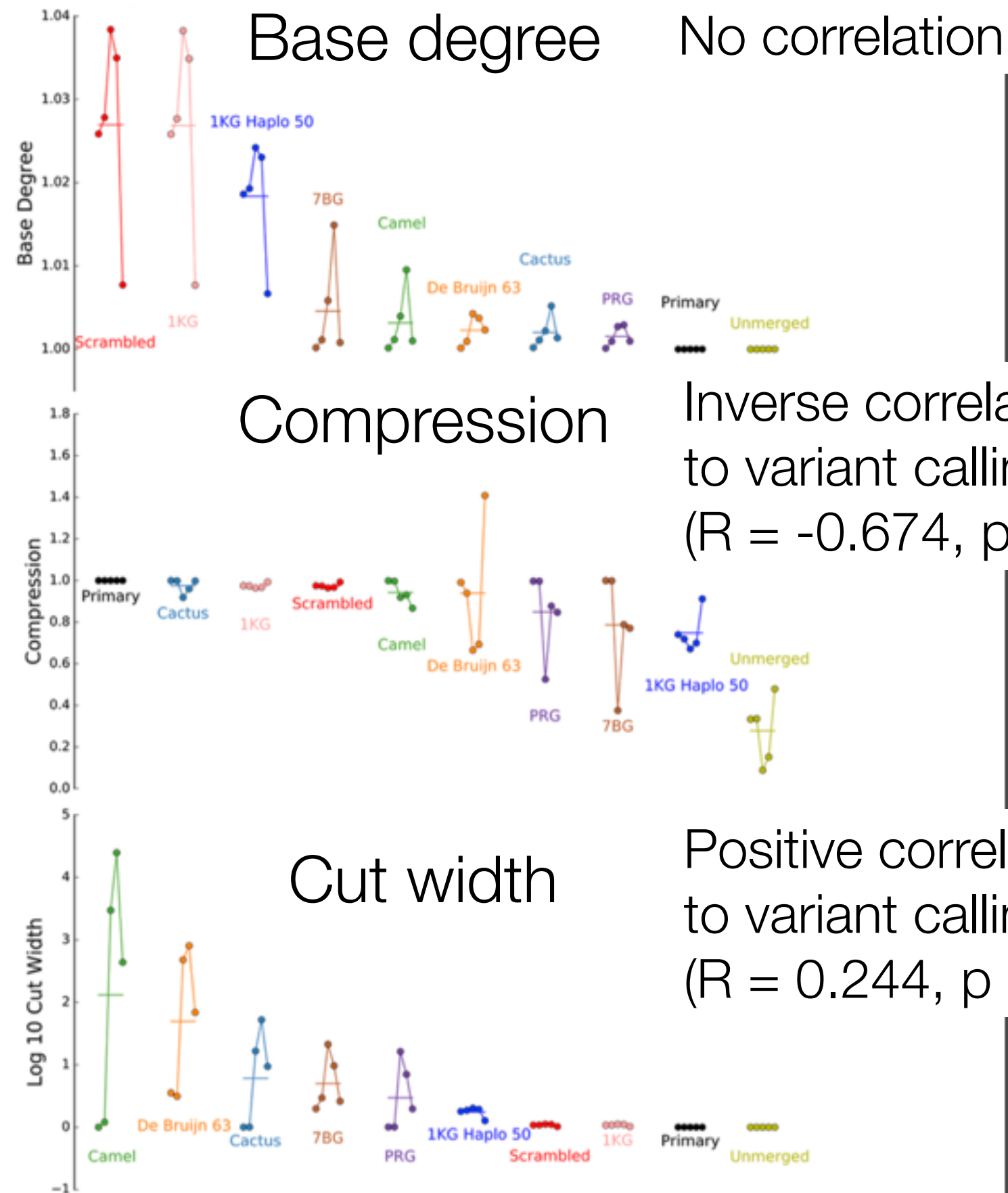


Figure 7: Empirical graph statistics. In each panel the result for each region is shown by a dot, in the following order: BRCA1, BRCA2, LRC_KIR, MHC, and SMA.

Discussion

- Unified variant calling pipeline
- Platinum Genome validation set can be also biased
- Mapping: more comprehensive & less ambiguous
- Reducing false recombinations: still challenging
Use variation graph? (sequence graph + haplotype)
- *“Filed is just in its formative stages”*
 - Several proposed graphs