De novo assembly and genotyping of variants using colored de Bruijn graphs

Iqbal et al. 2012

Kolmogorov Mikhail 2013

Challenges

- Detecting genetic variants that are highly divergent from a reference
- Detecting genetic variants between samples when there is no available reference sequence

Mapping-based approaches limitations

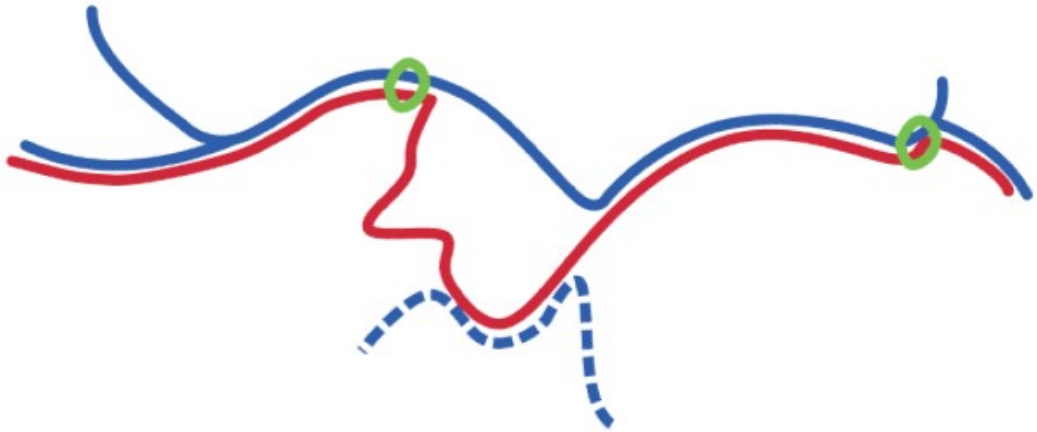
- Sample may contain sequence absent or divergent from the reference
- Reference sequences, particularly of higher eukaryotes are incomplete, notably in telomeric and pericentromeric regions
- Some samples may have no available reference sequence

Cortex!

- De novo assmebler capable of assembling multiple eukaryote genomes simultaneously
- Focus on detecting and characterising genetic variation
- Alignment free
- Accuracy of variant calling may be improved by adding reference sequence

Colored graph

- De Brujin graph with colored nodes and edges
- Different colors may reflect HTS data from multiple samples, experiments, reference sequences, known variant sequences

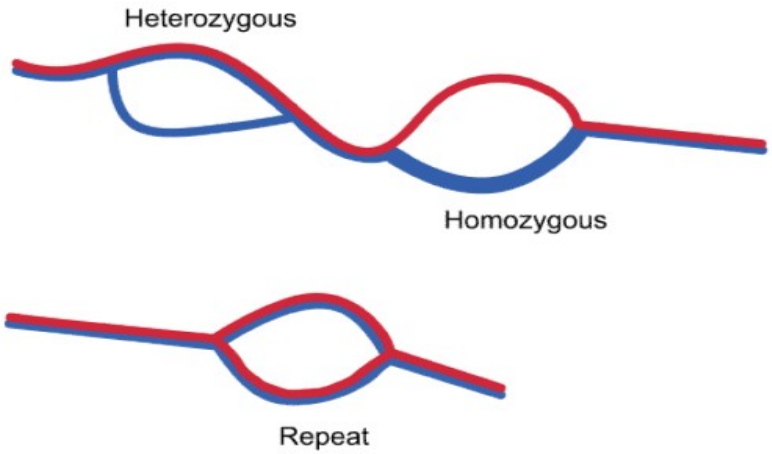


Variant calling aproaches

- Bubble calling
- Path divergence
- Multiple-sample analysis
- Genotyping

First approach: Bubble calling

- Bubbles can be induced by variants, repeats or sequencing errors
- Variants can be separated from repeats by including reference sequence
- When multiple samples available, it`s possible to estimate likelihood of bubble type



BC implementation

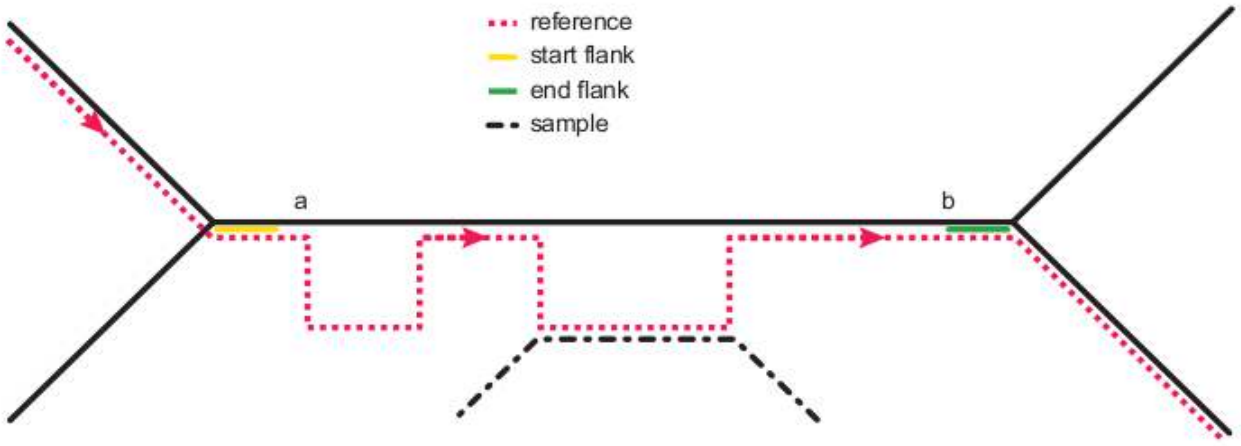
```
get_super_node(n, G):
  path1 = traverse forward, until reach node with in/out degree != 1
  path2 = traverse forward, until reach node with in/out degree != 1
  return union(path1, path2)

bubble_caller(G):
  for each node n in G:
    if outdegree(n) == 2 and not_visited(n):
      mark_as_visited(n)
      (<n1,e1>,<n2,e2>) = get_outedges(n, G)
      path1 = get_super_node(n1, G)
      path2 = get_super_node(n2, G)
      mark_as_visited(path1[0] .. path1[-1])
      mark_as_visited(path2[0] .. path2[-1])

      if path1[-1] == path2[-1] and
                  orientation(path1[-1] == orientation(path2[-1]):
        bubble_found = true
```

Second approach: Path divergence

- Complex variants are unlikely to generate clean bubbles
- In some cases (particularly deletions), path complexity is restricted to reference allele
- Such cases may be identified by following reference sequence and track places, where it diverges from sample graph

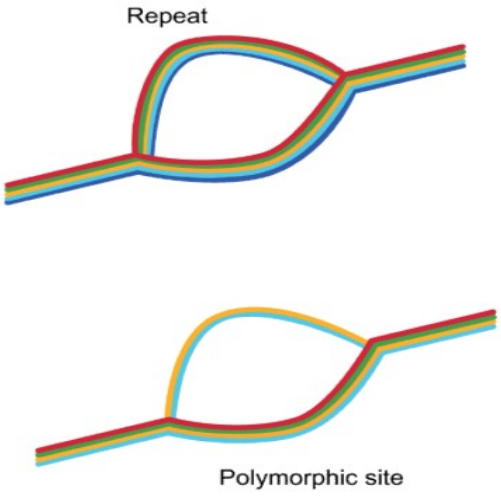


PD implementation

```
path_devergence(L, M, ref):
  for i in [0 .. len(ref)]:
    s = get_supernode_in_sample_color(n)  #may be null
    b = get_first_position_differ(ref,s)
    if s != null and b > 0 and
              s[b–L–1+j] == n[i+j] for j = [0 .. L-1]:
      for j = [i+L+1 .. i+M]:
        if s[len(s) – L+k] == ref[j+k] for k = [0 .. L-1]:
          variant_found = true
          break
```

Third approach: Multiple-sample analysis

- The joint analysis of HTS data from multiple samples can improve the accuracy and false discovery rate of variant detection substantially
- Maintaining separate colors for each sample gives an additional information about whether a bubble is likely to be induced by repeats or errors
- Approach can be used when there is no suitable reference



Classification of graph structures

- Given colored de Brujin graph, with each color representing single diploid individual from single population
- Need to classify pair of paths (for example, bubble branches) as either alleles of variant, repeats or errors

Information sources

- Total coverage of two sides
- Average allele-balance (the distribution of coverages between the two branches)
- Variance of allele-balance across samples

Structure differences

- **Repeat**: normal coverage, intermediate balance, no variance between samples.

- **Variance:** normal coverage, but each individual has a 0, 1 or 2 copy of allele 1 (and 2, 1 or 0 of allele 2).

- **Error:** low coverage on error branch

Models

- Coverage modelles as over-dispersed Poisson distribution

- Variant: Binominal distribution for allele-balance

- Repeat / Error: Beta distribution for allele-balance

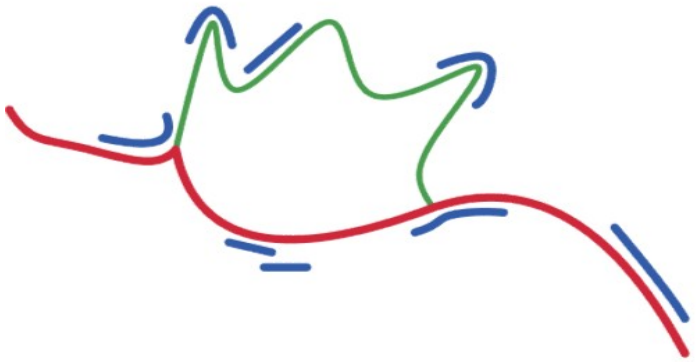
- A site is classified as Variant / Repeat / Error if the log-likelihood of the data under that model is at least 10 greater than the other models ($\log_{10}$ Bayess Factor at least 10)

Fourth approach: Genotyping

•Colored de Bruijn graphs can be used to genotype samples at known loci even when coverage is insufficient to enable variant assembly

•Graph is constructed of the reference sequence, known allelic variants and data from the sample

•The likelihood of possible genotype is calculated accounting for the graph structure



Genotyping algorithm

- We have a graph with one color for each known allele, one color for reference genome (with site in question named X) and one color for sample

- For any pair of alleles y_1, y_2 we define a likelihood function

Likelihood function

- Ignore all segments of alleles, which are present in reference minus X

- Decompose both alleles into sections, that are shared (s_i) and uniqe (u_i) and count nuber of reads in each

- Count number of nodes n in both sample graph and joint graph of all known alleles except γ_1 and γ_2 — these are sequencing errors

- For each section of length li probability of r_i reads arriwing within the section is given by a Poisson distribution with rate $\theta_i = \lambda l_i$ on shared segments and $\theta_i / 2$ on unique

$$P(\text{genotype} = \gamma_1 \cup \gamma_2 \mid \text{data}) = \prod_{s_i} \vartheta^{r_i} \frac{e^{-\vartheta_i}}{r_i!} \prod_{u_i} \left(\frac{\vartheta_i}{2}\right)^{r_i} \frac{e^{-\vartheta_i/2}}{r_i!} S(n)$$

Colored graph implementation

- Graph encoded impicitly, within a hash table
- Kmers and their reverse complements are stored in a same node
- Hash table value is array of integers, representing coverage for each color
- For each kmer, one binary flag is used for each nucleotide to track, if corresponding edge is present
- Memory usage per node: $8(k / 32) + 5c + 1$ bytes

Use cases

- Variant calling in a high coverage human genome
- Detection of novel sequence from population graphs of low-coverage samples
- Using population information to classify bubble structures
- Genotyping simple and complex variants

Limitations

- Not using paired reads information
- Greater need for error correction, as kmer size increases
- Potential of a graph explosion, as more individuals are included in the graph

Thanks for attention

In addition to old smoking pipe with fragrant tobacco and strong wine, I also like good assemblers. Such as Cortex.

J. Stalin