



HAL
open science

Looking for mutations in PacBio cancer data: an alignment-free method

Justine Rudewicz, Hayssam Soueidan, Raluca Uricaru, Richard Iggo, Jonas Bergh, Macha Nikolski

► **To cite this version:**

Justine Rudewicz, Hayssam Soueidan, Raluca Uricaru, Richard Iggo, Jonas Bergh, et al.. Looking for mutations in PacBio cancer data: an alignment-free method. JOBIM 2015, Jul 2015, Clermont-Ferrand, France. hal-01254846

HAL Id: hal-01254846

<https://hal.science/hal-01254846v1>

Submitted on 12 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Looking for mutations in PacBio cancer data: an alignment-free method

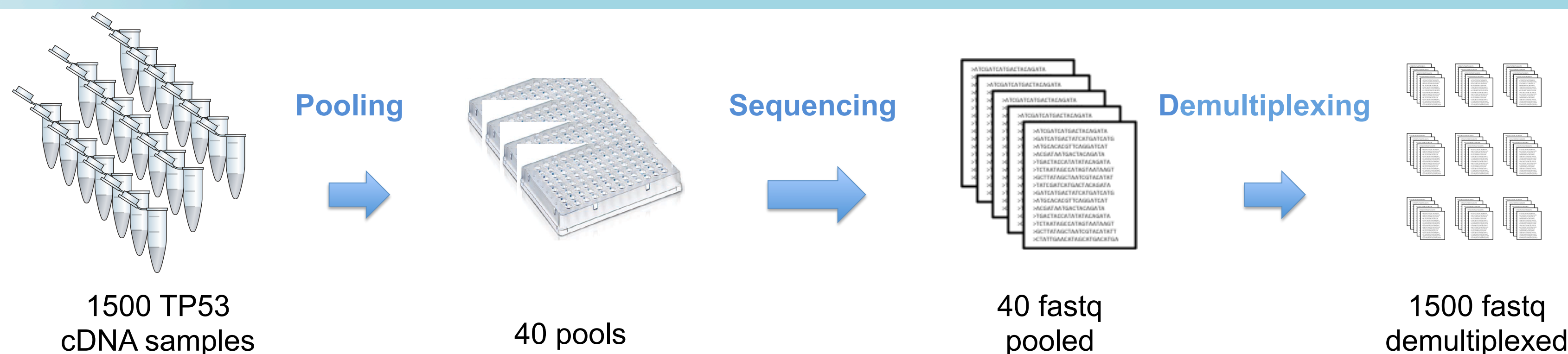
Justine RUDEWICZ^{1,2,3}, Hayssam SOUEIDAN¹, Raluca URICARU^{1,2}, Richard IGGO³, Jonas BERGH⁴ and Macha NIKOLSKI^{1,2}

1 CENTRE DE BIOINFORMATIQUE DE BORDEAUX, Université de Bordeaux, 146 rue Léo Saignat, 33000, Bordeaux, France
 2 LABORATOIRE BORDELAIS DE RECHERCHE EN INFORMATIQUE, UMR5800 CNRS, 351 cours de la Libération, 33405, Talence, France
 3 CENTRE RÉGIONAL DE LUTTE CONTRE LE CANCER DE BORDEAUX, U916 INSERM, 229 Cours de l'Argonne, 33000, Bordeaux, France
 4 DEPARTEMENT ONCOLOGY, Université d'Upsala, Box 256, 751 05 Uppsala Suède
 j.rudewicz@bordeaux.unicancer.fr

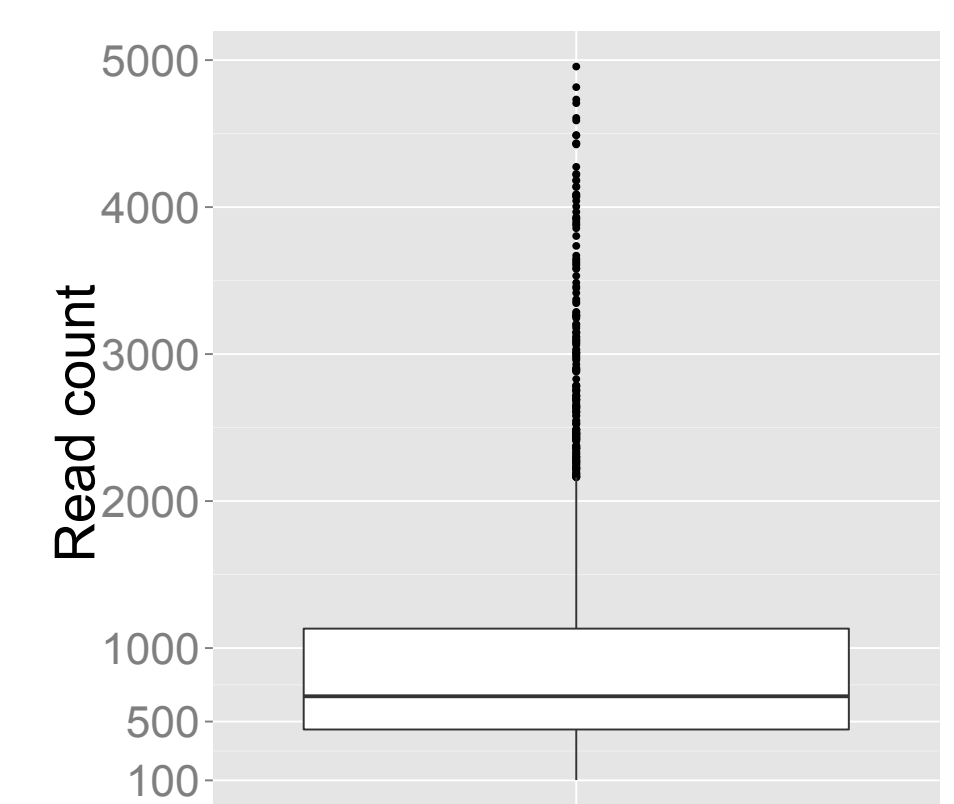
Abstract

To determine the **TP53 mutations** present in a patient cohort (~1500 patients), breast tumor TP53 mRNA was sequenced with **PacBio** technology. However, none of the existing tools, e.g. VarScan, GATK, has proven to be suitable for this type of data. Indeed, in addition to the **high sequencing error rate** generated by PacBio (~15%), the tumor biopsies are **contaminated** by healthy tissue. This makes it difficult to differentiate real mutations as they are lost within the high background noise. To circumvent this problem, we have developed a method for detecting mutations using **De Bruijn graphs: MICADo** for Mutations In Cancer Data.

Data

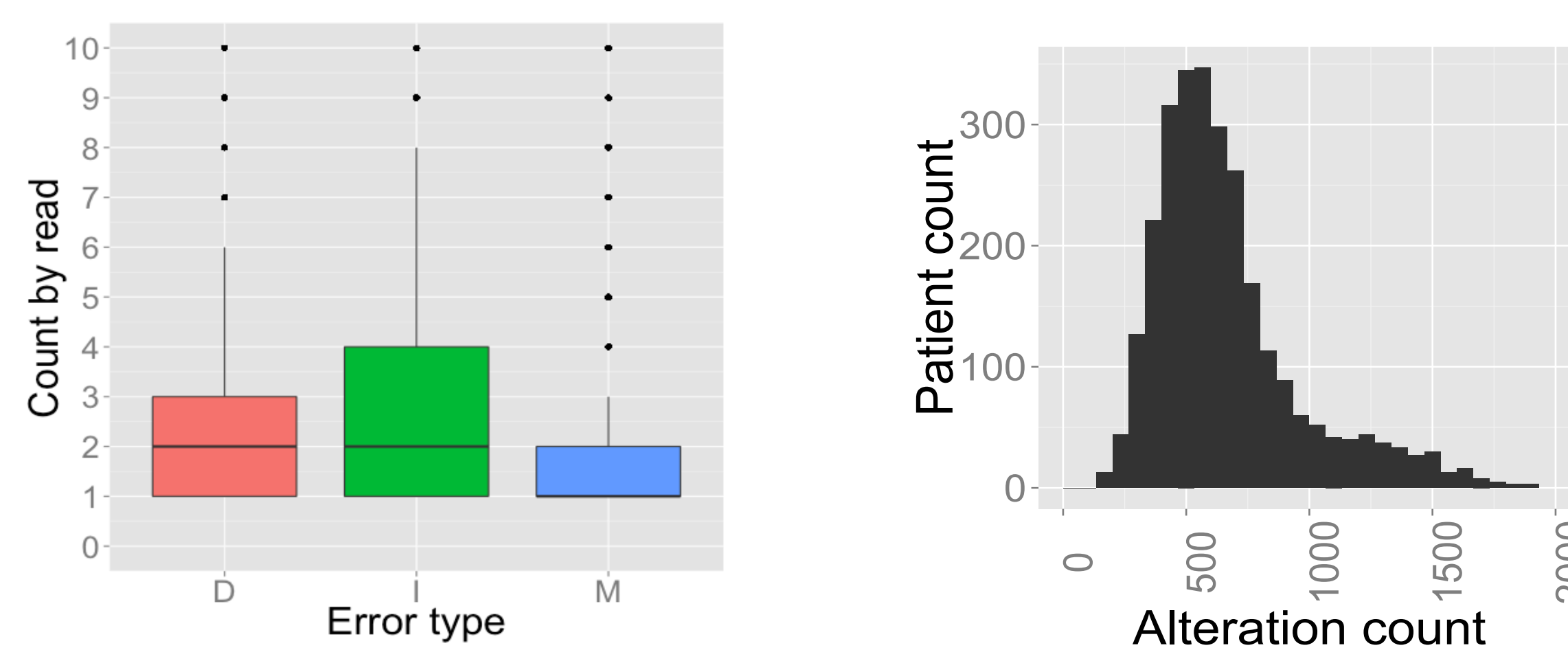


Study design: Each read from circular sequencing covers the entire sequencing region.



Read number by patient: varies between 100 and 5000

Extremely noisy



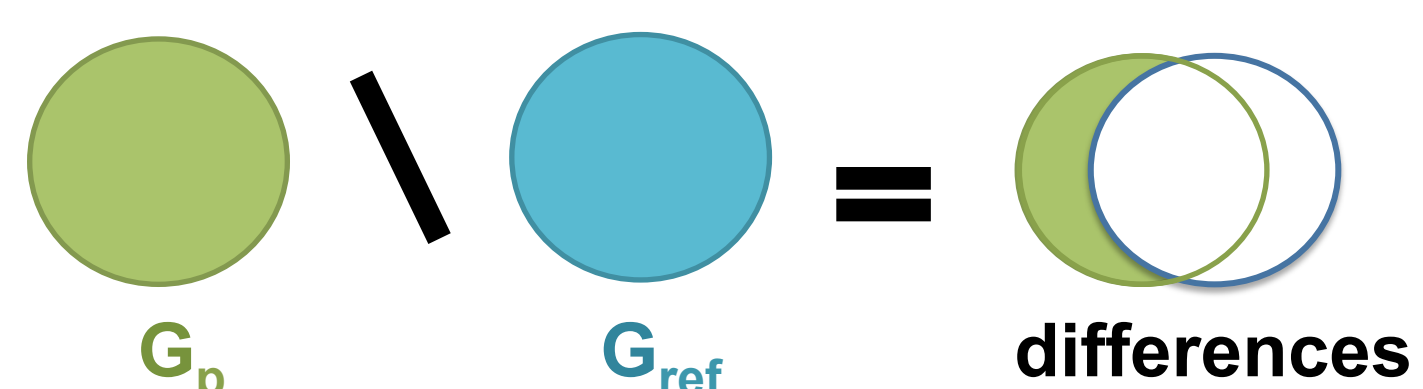
We expect **1-2 mutations by patient** but the number of different alterations per patient is very high.

MICADo

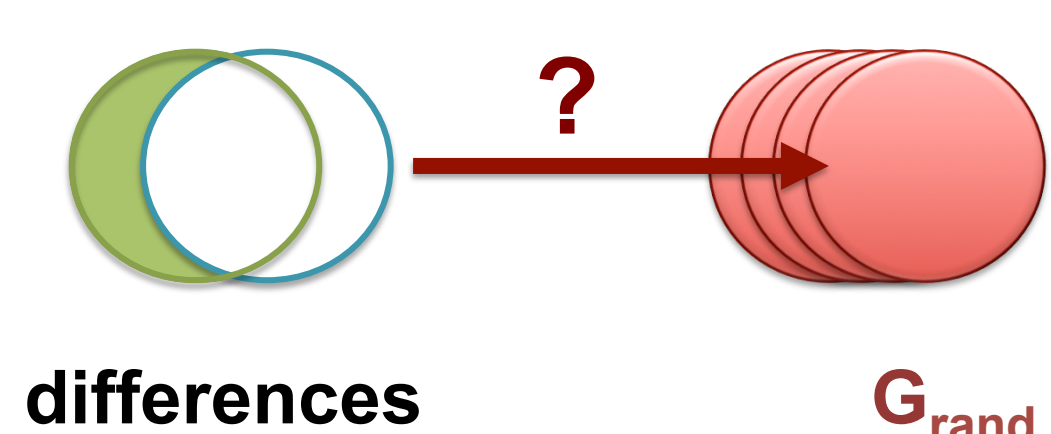
De Bruijn Graph based method that identifies paths coding for genomic alterations and distinguishes those that are specific to a patient from those common to the whole cohort.

- G_{reference}** 3 splicing variants and 4 known SNPs
- G_{patient}** read set of one patient
- G_{random}** sample of reads from the whole cohort

Step 1: compute differences between **G_p** and **G_{ref}**



Step 2: check if differences are frequent in the cohort

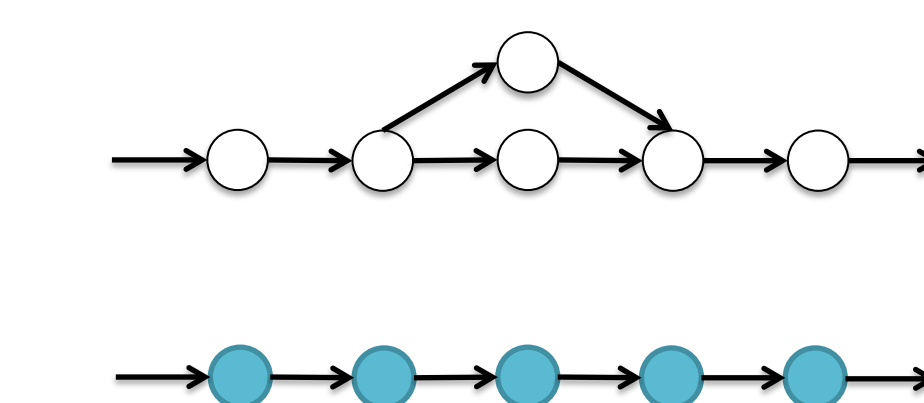


Step 1

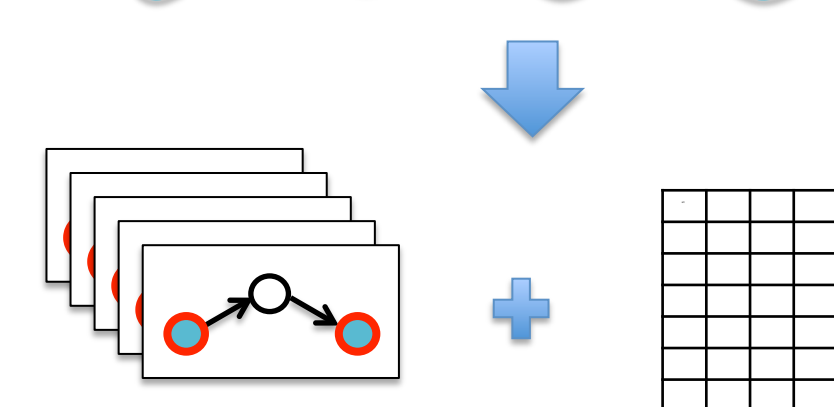
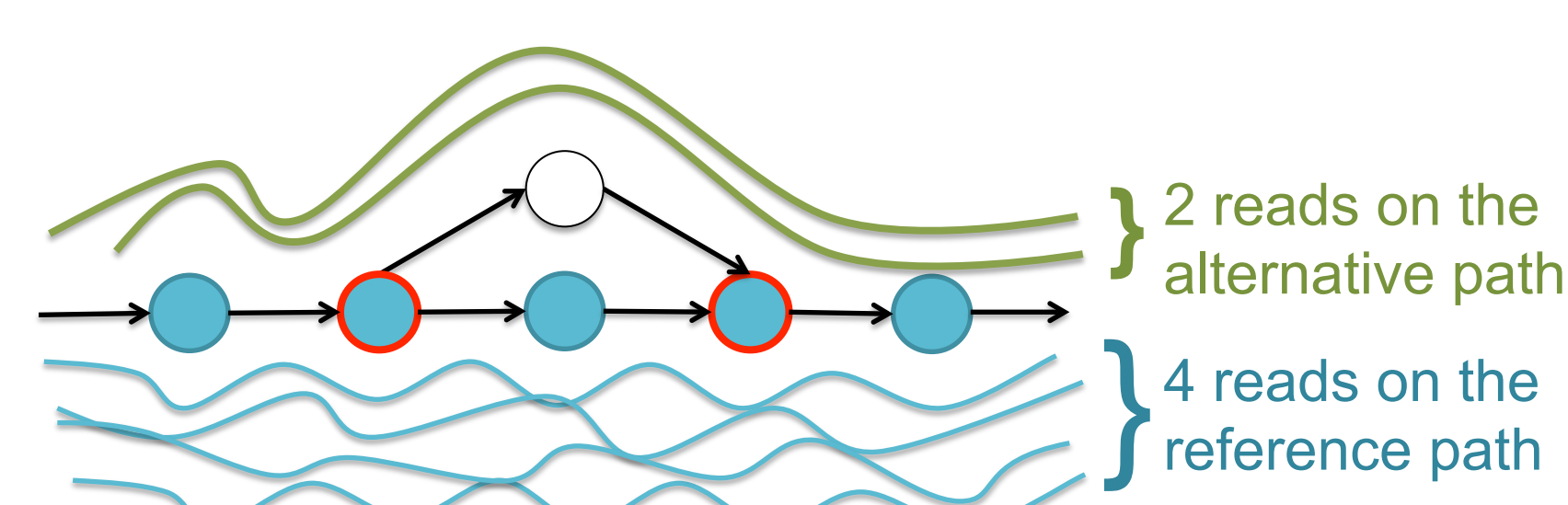
$$G_p = \langle V_p, E_p \rangle$$

$$G_{ref} = \langle V_{ref}, E_{ref} \rangle$$

$$G^* = \langle V_p \cup V_{ref}, E_p \setminus E_{ref} \rangle = \langle V^*, E^* \rangle$$



$\forall v_1, v_2 \in V_{ref} \cap V_p$, compute paths between v_1 & v_2 in G^*

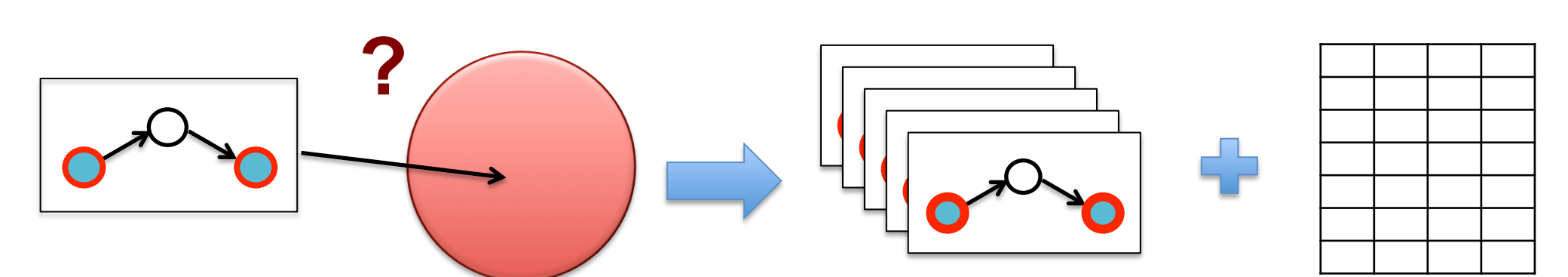


Read counts for paths

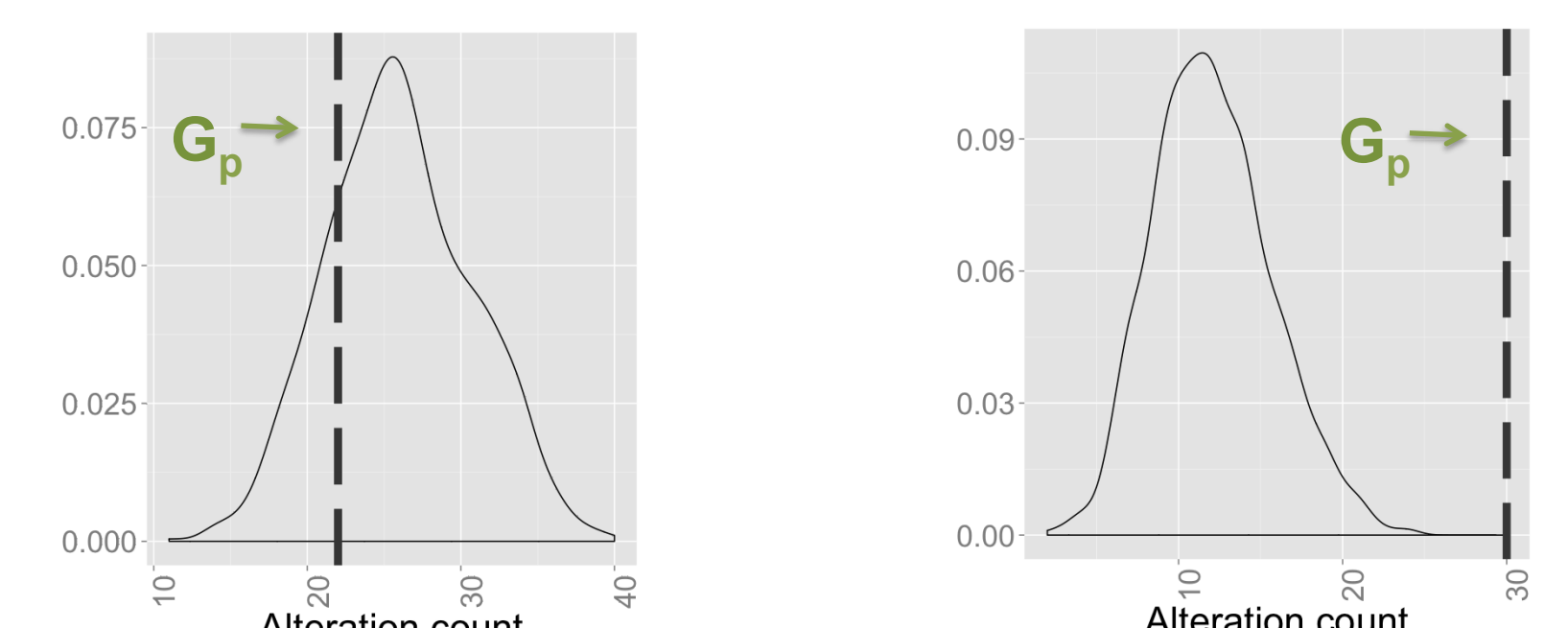
Step 2



Construct 1000 random graphs from 1000 samples of **random reads**



Check the **presence of the paths** in the 1000 random graphs and compute associated read counts.



→ **Cohort specific** (Similar) → **Patient specific** (Different)
 Compare read counts between **G_p** and **G_{rand}** and deduce a **p.value**

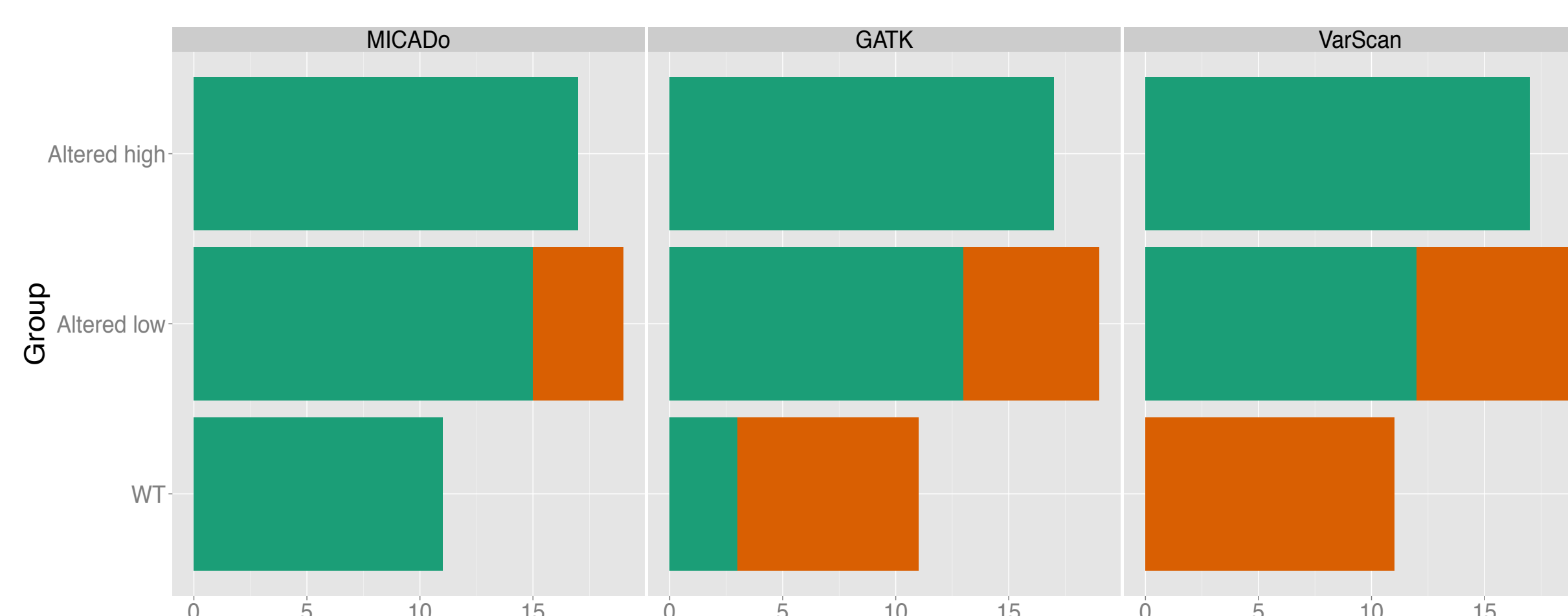
Results

Comparison between VarScan, GATK and MICADo on one pool (48 PacBio samples) previously sequenced by 454 technology.

Three types of samples:

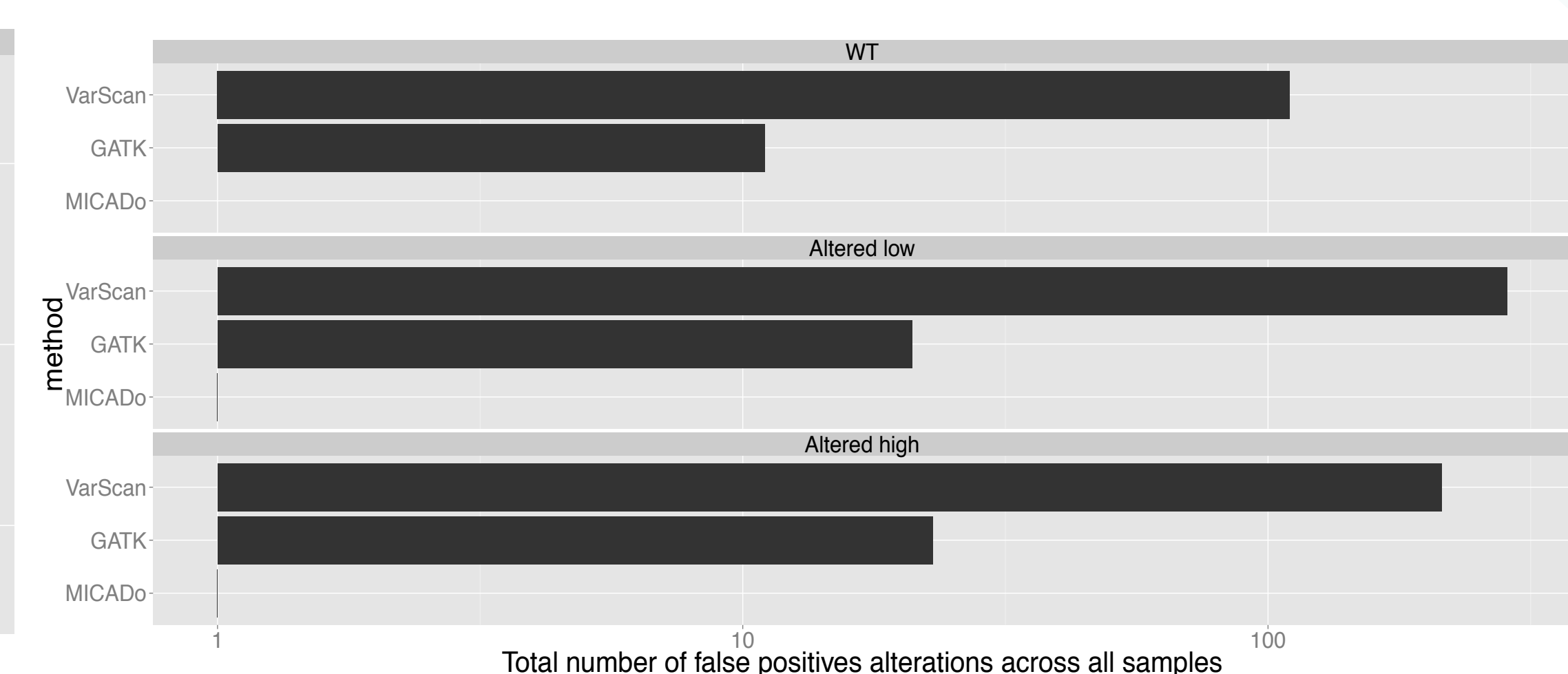
- Wild type: no mutation
- Altered low: few reads carry the true mutation
- Altered high: many reads carry the true mutation

→ **Better results with MICADo**



Well classified: no mutation found in wild type; true mutation found in mutated samples

Badly classified: mutation found in wild type; no mutation or not the right one in mutated samples



Conclusion

MICADo: Mutations In Cancer Data
 Suited to PacBio data for **gene-level** analyses

Overcomes bias linked to PacBio error rate for the **detection of mutations**

Perspectives

Test **MICADo** on **other data**

Consider **paths** in which starting node or end node is not present in the reference graph
 Define a **biological quality score** of samples

