# RG : Annotation of Genomic Regions with High/Low Variant Calling Concordance

Niko Popitsch[1,2], The WGS500 Consortium, Anna Schuh[2,3], Jenny C. Taylor[1,2]

[1] Wellcome Trust Centre of Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
[2] National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre, Oxford, UK.
[3] Department of Oncology, University of Oxford, Oxford, UK.

Contact: `niko@well.ox.ac.uk`

## Introduction

The increasing adoption of whole-genome resequencing in clinical and research environments **demands highly accurate and reproducible variant calling (VC) methods**. The observed **discordance between state-of-the-art VC pipelines**, however, indicates that the current practice still suffers from **non-negligible numbers of false positive and negative SNV and INDEL calls** that were shown to be enriched among discordant calls but also in genomic regions with low sequence complexity.

ReliableGenome (RG) is a method for **partitioning genomes into high and low concordance regions** with respect to a set of surveyed VC pipelines. RG **integrates variant call sets created by multiple pipelines from arbitrary numbers of input datasets** and interpolates expected concordance for genomic regions without data, resulting in a genome-wide concordance score.

## Method

Let $C_{i,j}$ be the variant call sets for $i \in 1,\ldots,N$ samples that were derived using $j \in 1,\ldots,M$ different VC pipelines.

RG consists of **two main stages**. First, the variant call sets from multiple pipelines are joined, resulting in $N$ joined sets $J_i$. SNV calls are matched based on genomic position, INDELs based on overlapping genomic intervals. A genomic position in $J_i$ is classified as concordant/discordant based on the accordance of the called genotypes of all matched calls at this position (Fig 1).

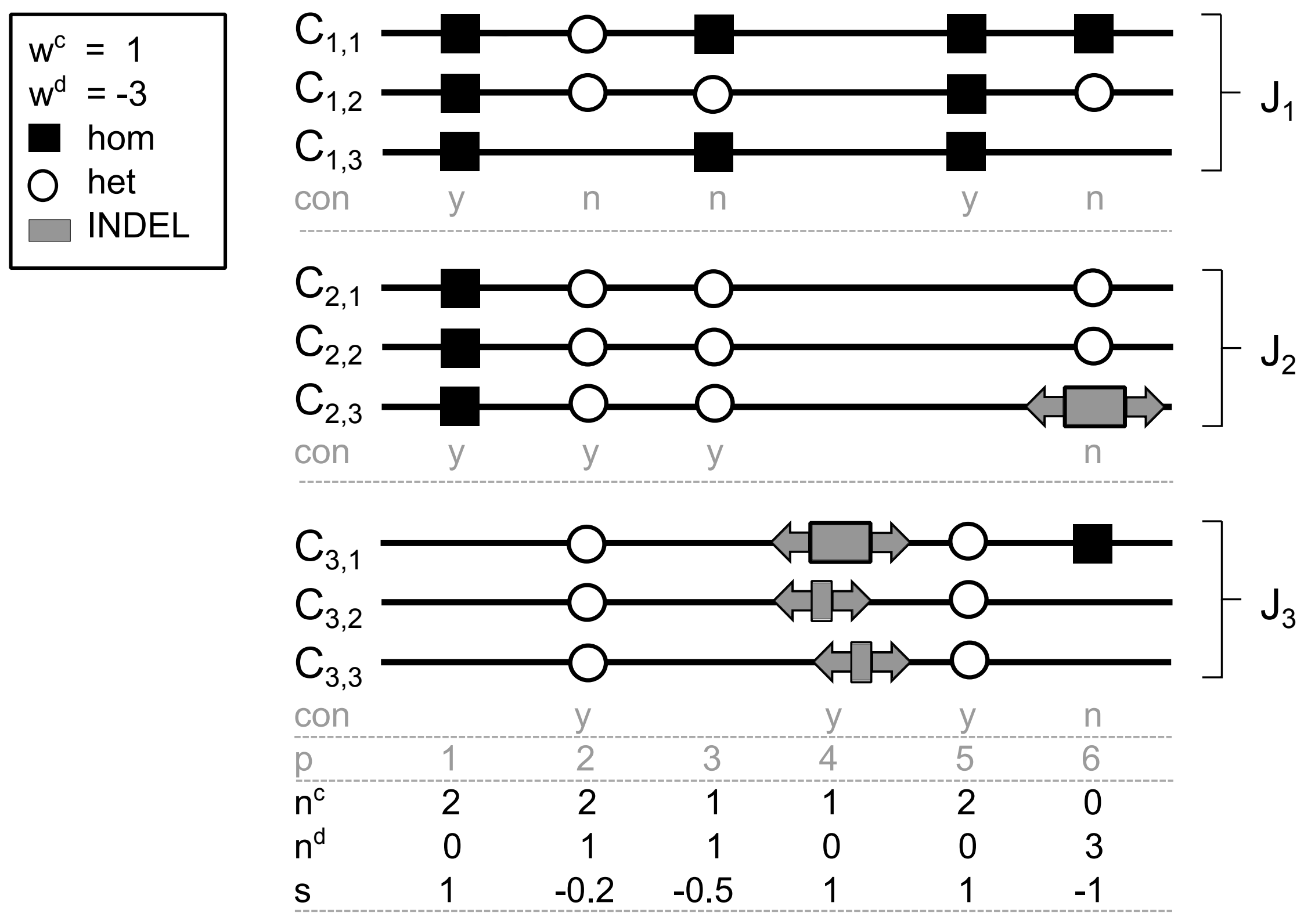

**Figure 1: Concordance scoring example.**

The second stage calculates a concordance score for each polymorphic position in the input cohort:

$$s_p \in [-1,1] = \frac{n_p^c \cdot w^c + n_p^d \cdot w^d}{n_p^c \cdot w^c + |n_p^d \cdot w^d|}$$

with $n_p^c, n_p^d$ being the counts of concordant/discordant decisions for a given position and $w^c > 0$ and $w^d < 0$ being configurable scoring weights.

Scores for position without data are interpolated. Ultimately, genomic regions of high/low concordance are calculated from this genome-wide signal (Fig 2).

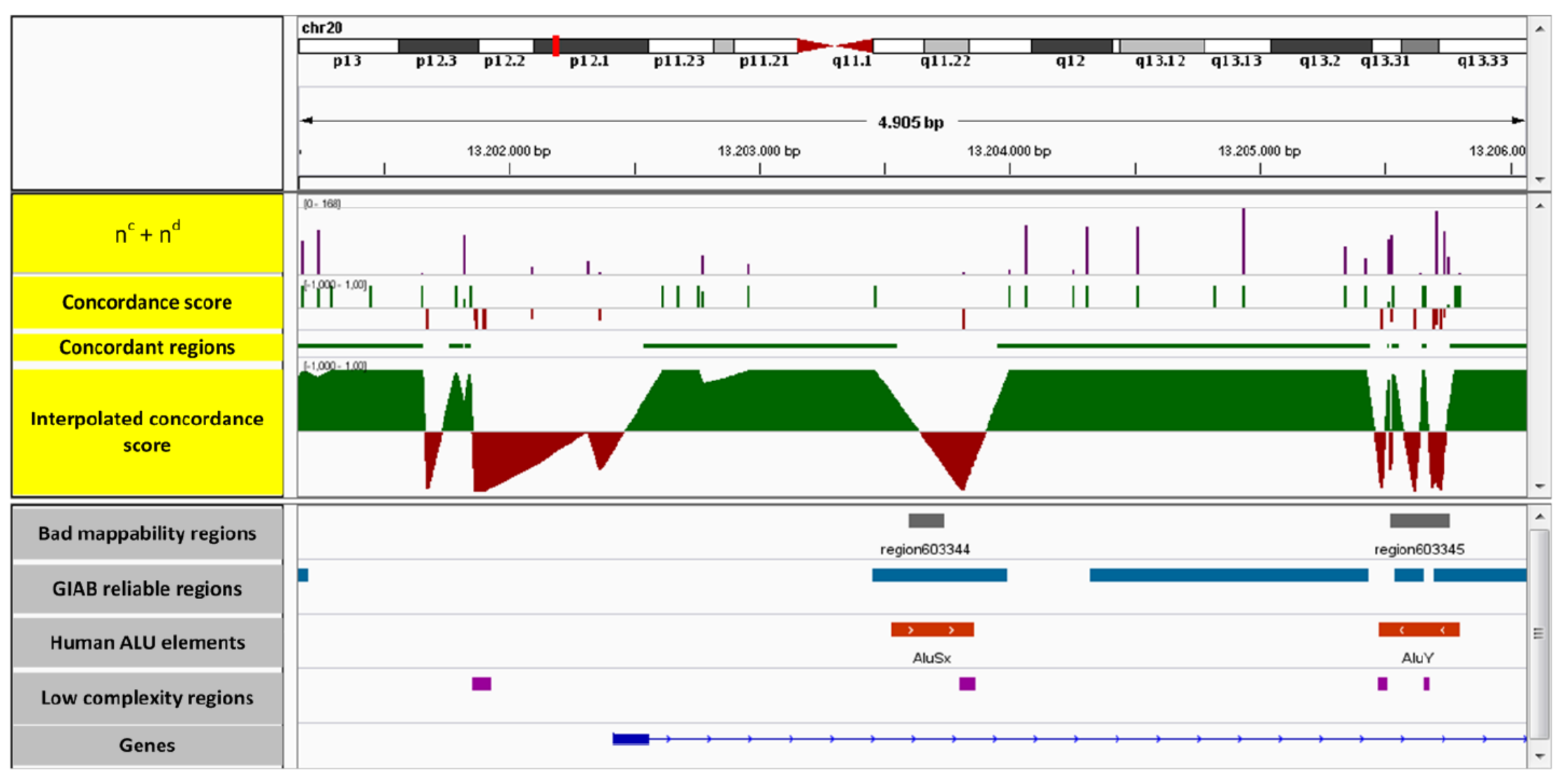

**Figure 2: IGV screenshot showing the main RG result files.** $n^c + n^d$**: number of training datasets that contained a call at the respective genomic position; Concordance scores/regions: red: negative scores/discordant, green: positive scores/concordant.**

## Evaluation experiment 1

We applied RG to **219 deep WGS datasets** from the WGS500 cohort [1]. We called variants using GATK [2], samtools [3] and platypus [4] and conducted $i = 1,\cdots,215$ evaluation runs. In each run we selected a **random subsample of $i$ training datasets** and then evaluated RG as a **binary classifier** for predicting the concordance status in the remaining datasets (the validation cohort).

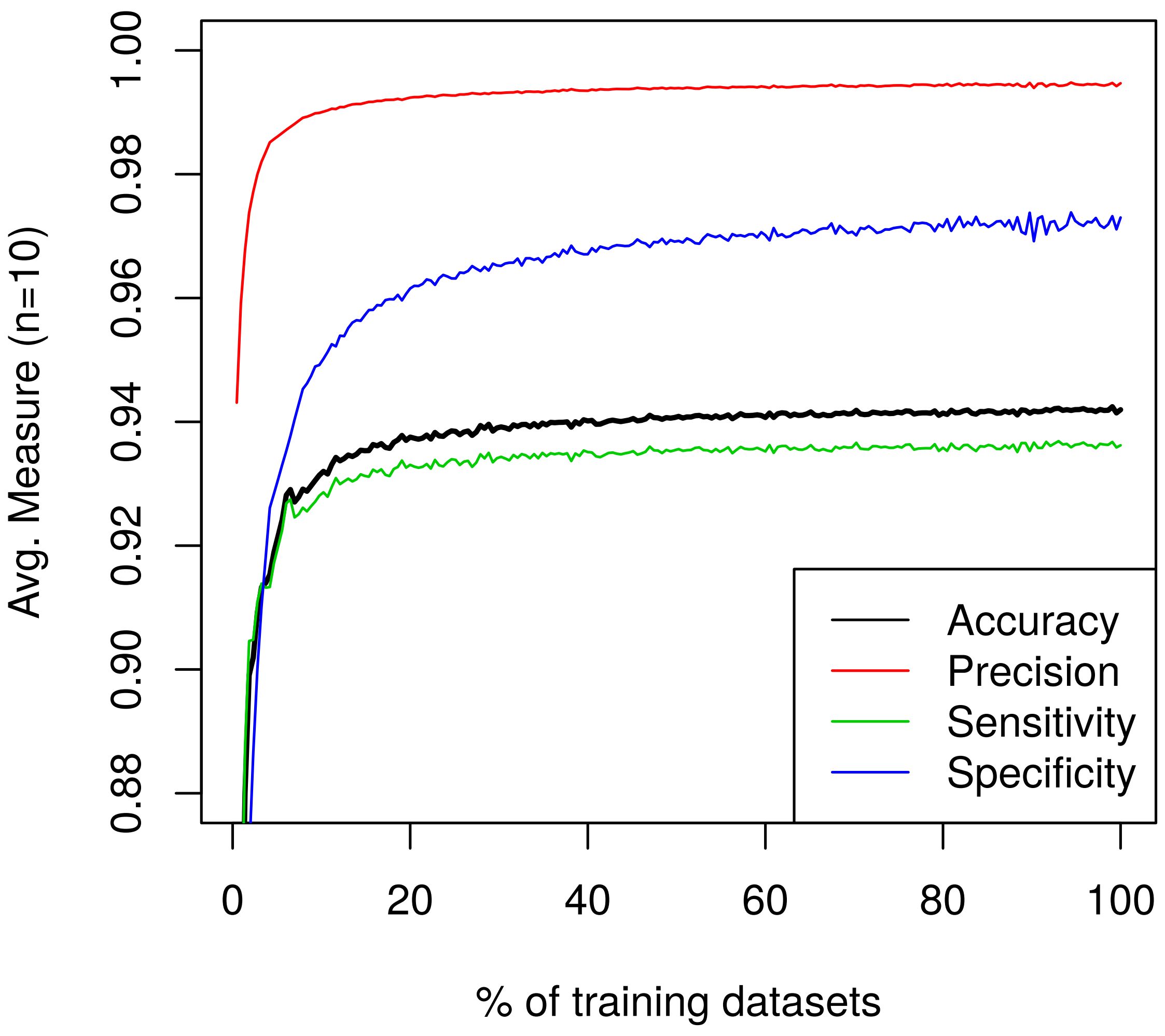

**Figure 3: Averaged performance metrics for n=10 evaluation runs (chr20 only).**

Our binary classifier reached **high values for precision ($> 99\%$) and specificity ($> 97\%$)**, see Fig 3. The accuracy profile results from high recurrence of discordant regions across datasets which indicates that **low call concordance** is predominantly **a property of the genomic location/context** rather than the actual sequencing data quality.

## Evaluation experiment 2

We compared our RG-derived set of concordant and discordant regions to three other genomic partitions:

1. GenomeInABottle (**GIAB**) reliable regions [5]
2. Illumina Platinum "confident regions" (**PLAT**) [6]
3. Regions with low sequence complexity (**LCR**) as published in [7]. We also considered a second version of this partition (**LCR100**) that was derived by extending all regions by 100bp up/downstream.

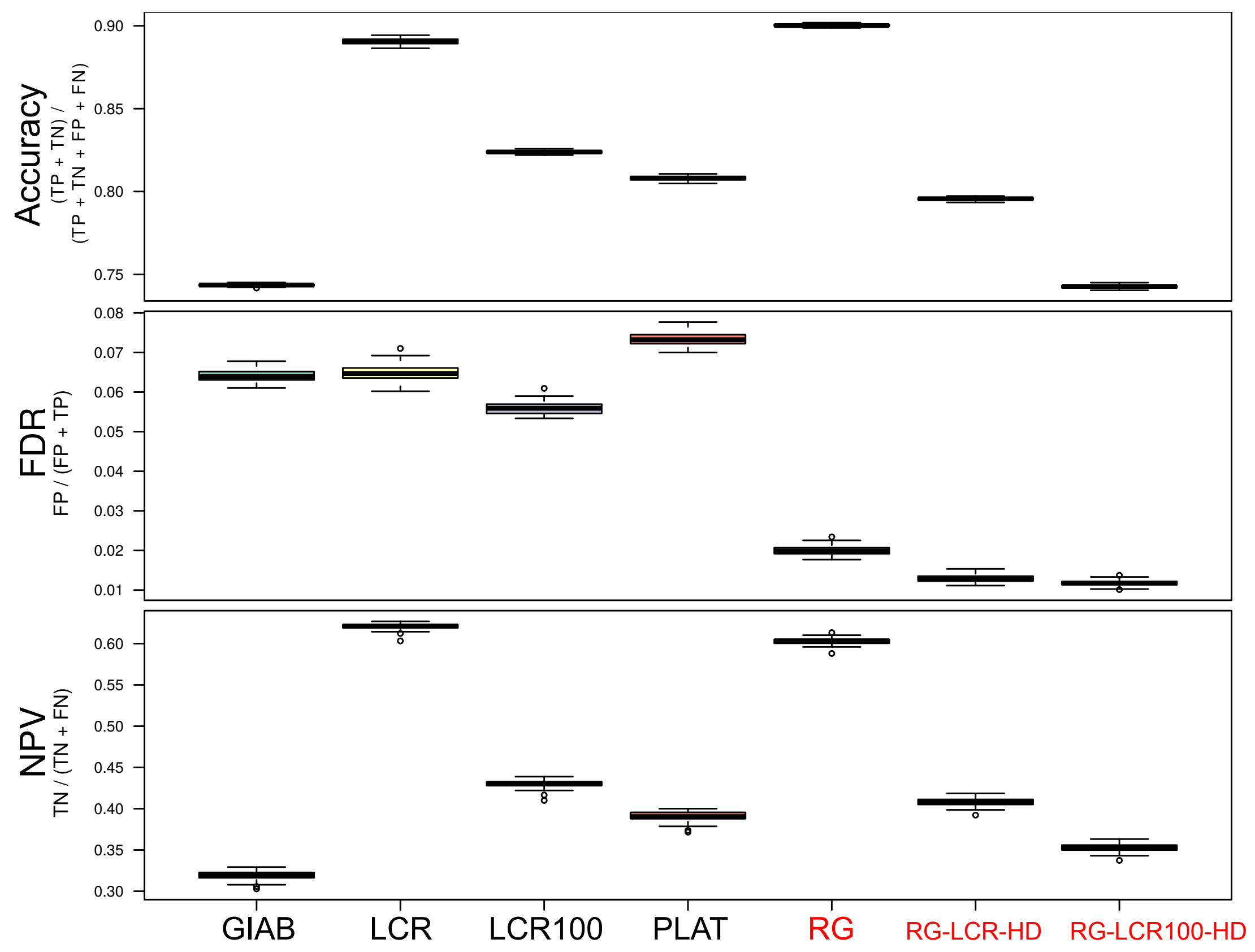

**Figure 4: Accuracy, false discovery rate (FDR) and negative prediction value (NPV) boxplots for the various partition sets were calculated by using 34 independent deep WGS samples as ground truth.**

RG showed the **highest accuracy** and the **lowest FDR** of all methods. Our genomic partition can further be optimized wrt. FDR and NPV (e.g., by removing low-complexity regions or regions with high density of discordant calls (HD)), however, at the cost of decreased overall accuracy (Fig 4).

## Evaluation experiment 3

Finally, we measured the performance of RG for predicting potential **false positive heterozygous calls** in WGS datasets derived from the **haploid cell line CHM1hTERT (CHM1)** as published in [7]. We measured against variant call sets created with two different read mappers (bwa mem [mem], bowtie2 [bt2]) and three different VC pipelines (GATK HaplotypeCaller [hc], Platypus [pt], FreeBayes [fb]).

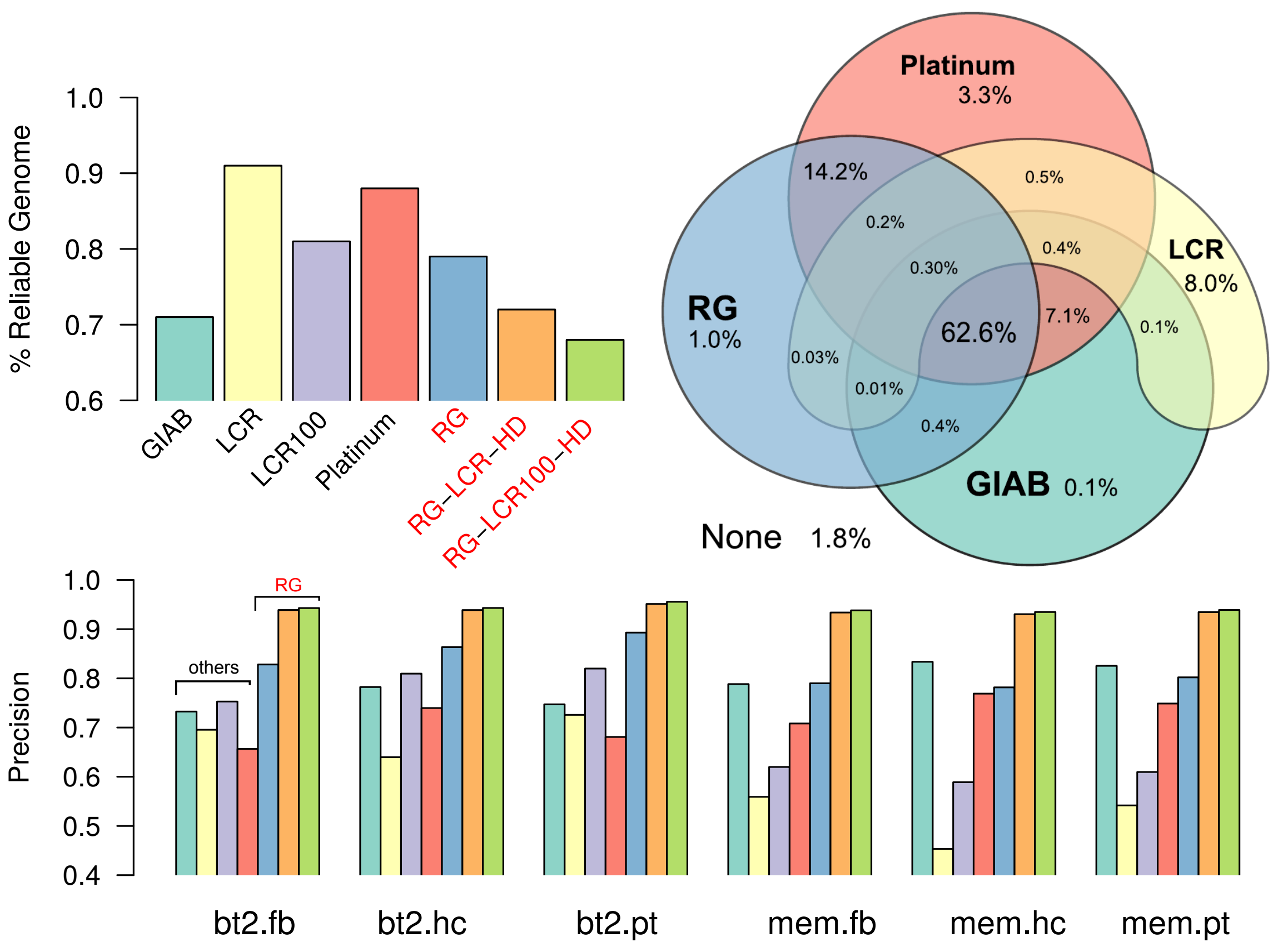

**Figure 5: Partition statistics and classification precision of false-positive heterozygous calls in a haploid cell line.**

RG reaches high and constant precision, particularly when combined with LCR annotations and regions of high discordance density (HD). The latter combination **outperforms other approaches 3-4X**, see Fig 5.

## Conclusions

- VC **concordance depends** predominantly **on genomic context** which manifests in high recurrence of regions that can/cannot be reliably genotyped by a single method. **This enables the *a priori* calculation of genomic partitions**.
- **RG** differs from previous efforts in that it **incorporates data from whole cohorts** of WGS datasets, thereby capturing more of the data's variance. **RG clearly outperforms other methods** (GIAB, PLAT, LCR) in predicting VC concordance and false positive calls in low-concordance regions.
- RG is **useful for variant filtering, annotation and prioritization**. It also allows focusing resource-intensive algorithms (e.g., consensus calling methods) on the smaller, discordant share of the genome (20-30%) which might result in **increased overall accuracy at reasonable costs**.
- RG is further **useful for development, benchmarking and optimization** of VC algorithms and for the relative comparison of call sets between different studies/pipelines.
- RG is freely available for non-commercial use at **`https://github.com/popitsch/wtchg-rg/`**
- Future work: Extension for somatic calls; analysis of discordant genomic regions; evaluation with other kinds of sequencing data.

## References

[1] Jenny C. Taylor, Hilary C. Martin, Stefano Lise, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet*, 47(7):717–726, Jul 2015.

[2] Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–498, May 2011.

[3] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[4] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F. Twigg, W. G. S500 Consortium , Andrew O M. Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*, 46(8):912–918, Aug 2014.

[5] Justin M. Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nat Biotechnol*, 32(3):246–251, Mar 2014.

[6] Illumina. Platinum genomes project. http://www.illumina.com/platinumgenomes/, 2015.

[7] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, Oct 2014.