

# Сравнение точности пайплайнов обработки NGS

Андрей Афанасьев,  
CEO@iBinom

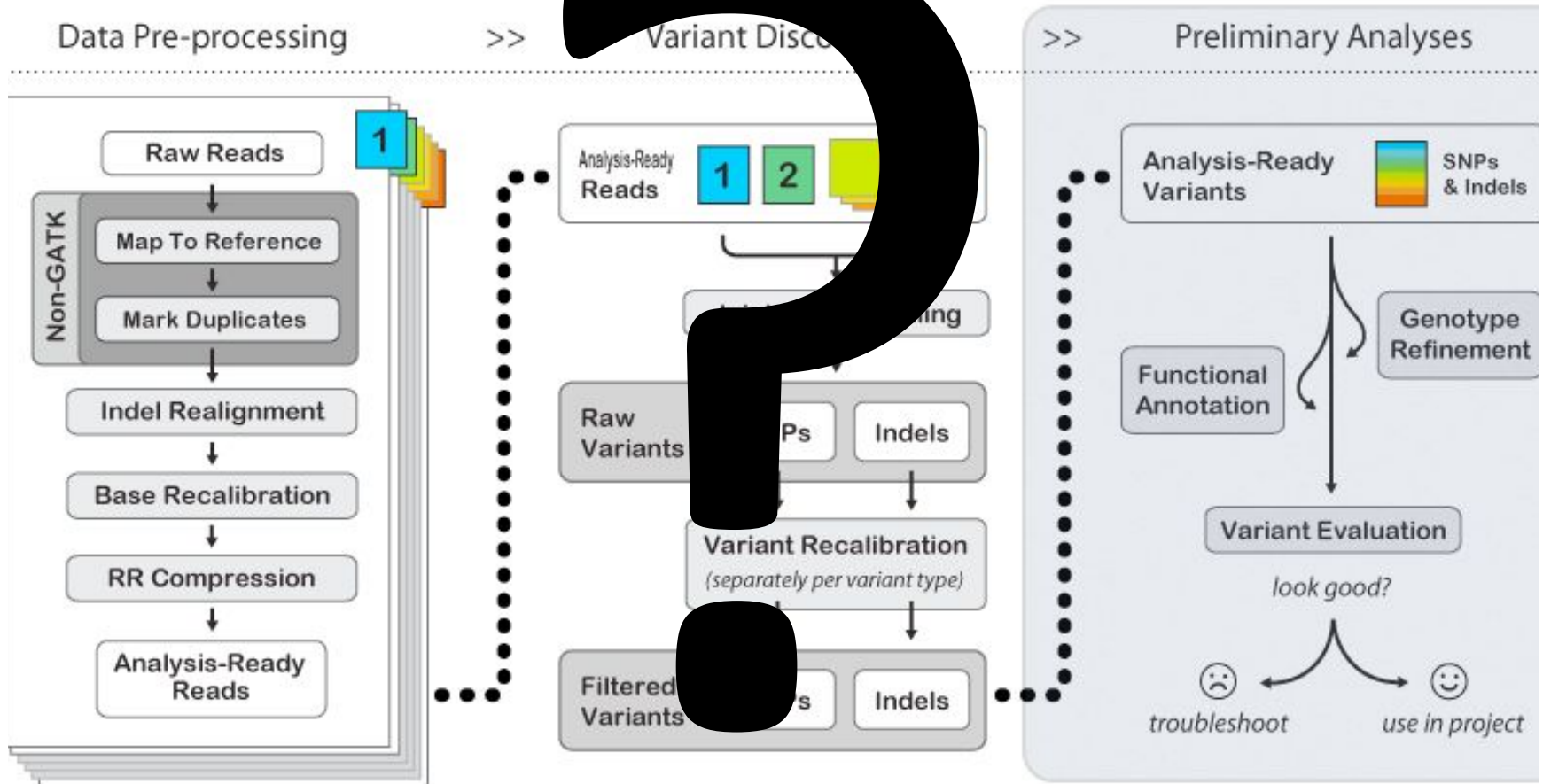
# Зачем это всё?

- Для использования NGS в клинической практике нужны точные и воспроизводимые результаты
- Новые или старые программы?
- Как их сравнивать?
- Кто круче?



Пайплайнов много, а правда одна

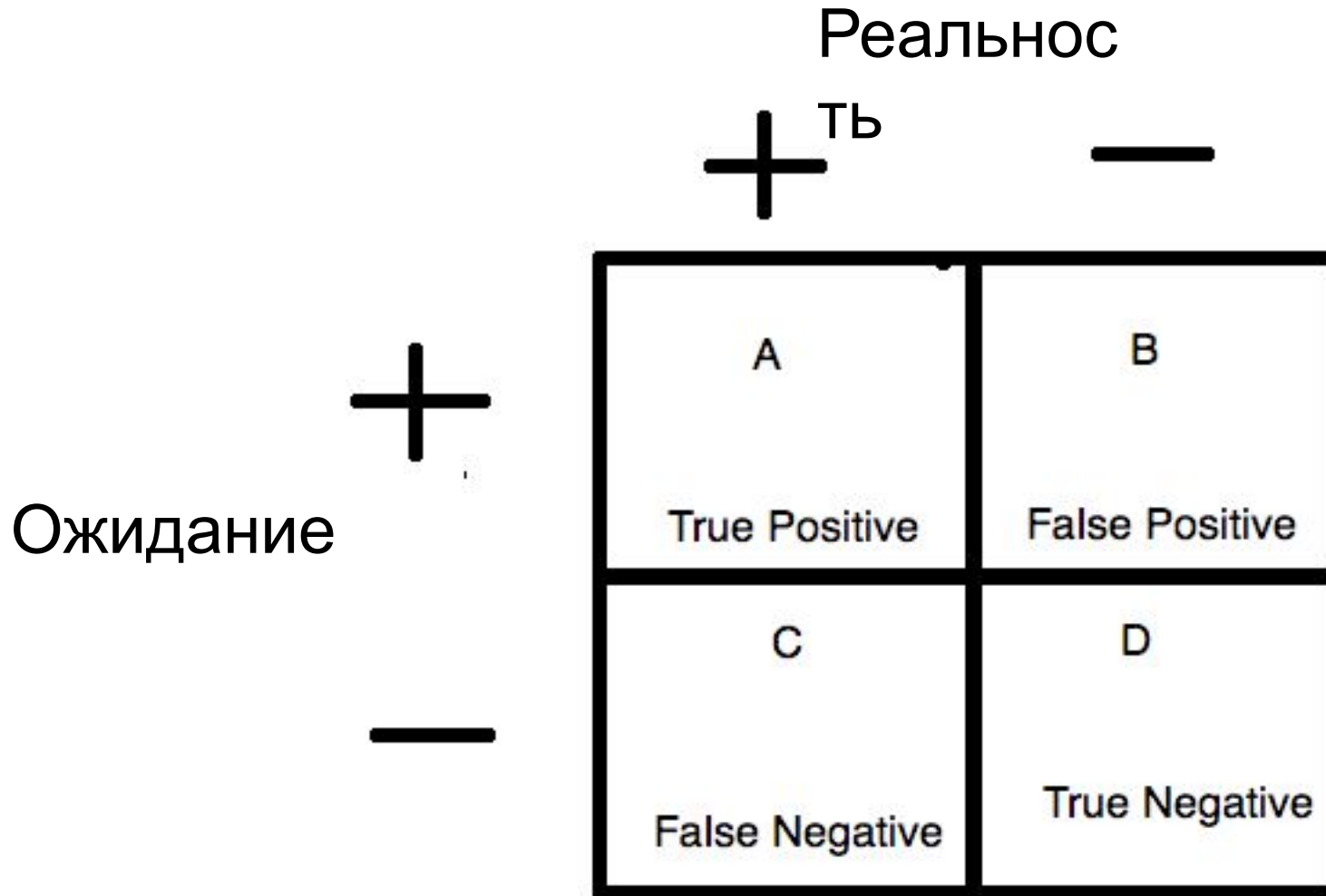
# Есть ли стандартный пайплайн?



Нельзя просто так взять и получить vcf  
файл!

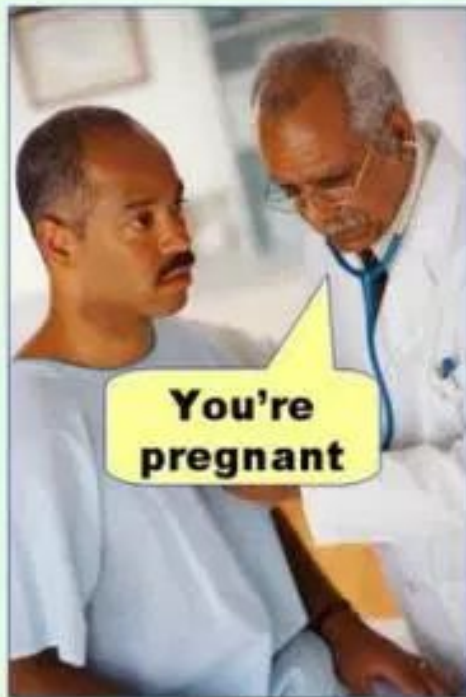


# Что мы измеряем?



# Что мы измеряем?

**Type I error**  
(false positive)



**Type II error**  
(false negative)



# Что мы измеряем?

- Точность (Precision) =  $TP / (TP + FP)$  – как много найденных вариантов на самом деле есть;
- Чувствительность (Sensitivity) =  $TP / (TP + FN)$  – как много найденных вариантов подтвердилось с учетом не найденных вариантов;
- Специфичность (Specificity) =  $TN / (TN + FP)$  – как много не найденных вариантов действительно нет

# «Золотой» образец NIST

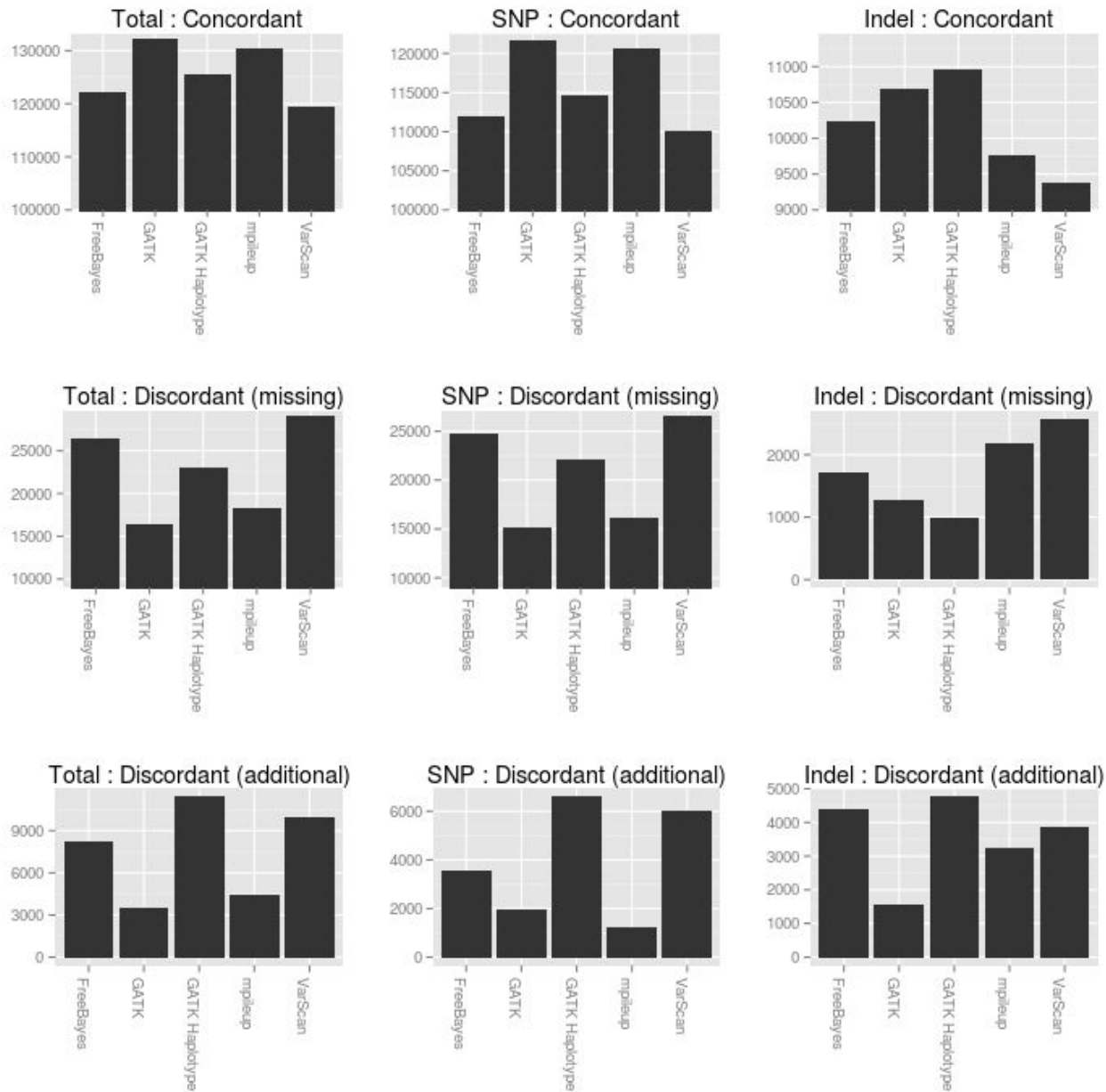
- Семья из Юты: NA12878 Genome in a Bottle
- **ОЧЕНЬ** хорошо охарактеризован





# «Золотой» образец NIST

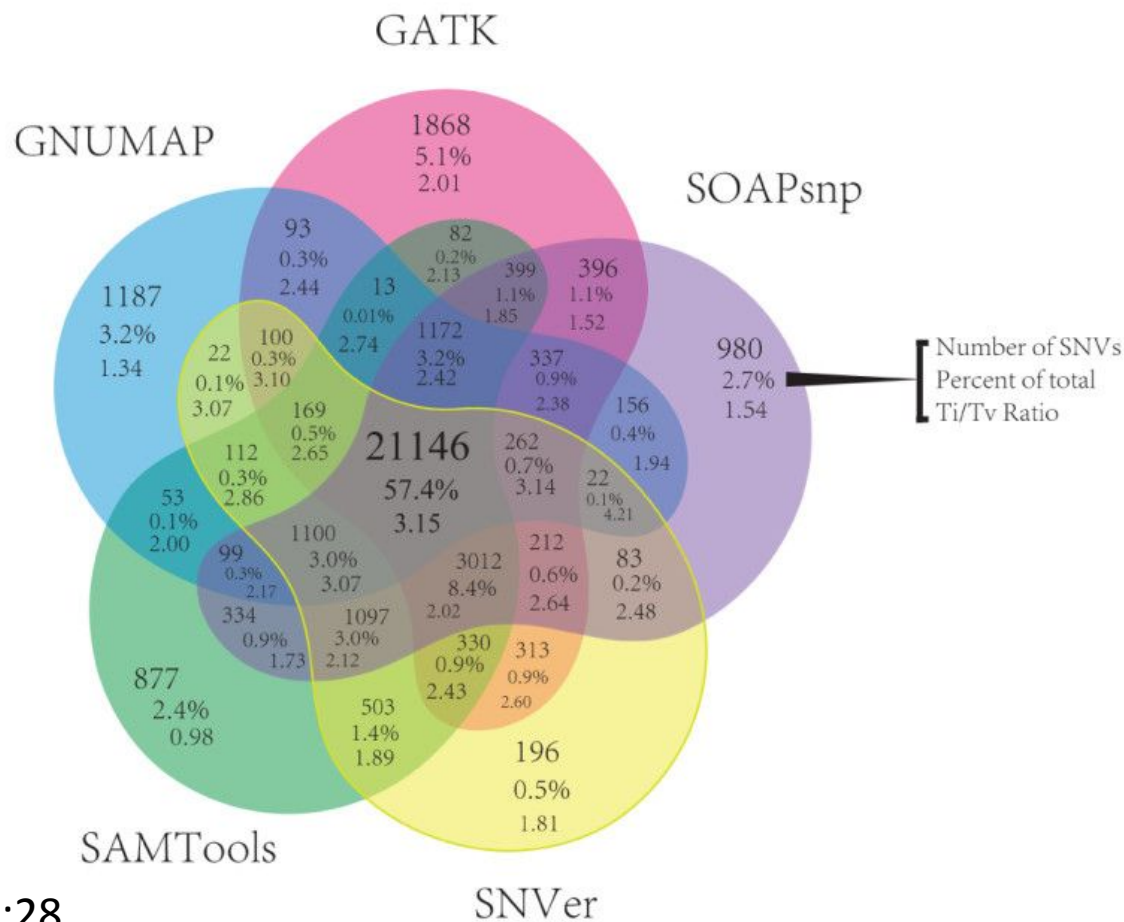
Source*	Platform	Mapping algorithm	Cov-erage	Read length	Genome/exome
1000 Genomes	illumina	Bwa	39	44	Genome
1000 Genomes	illumina	Bwa	30	54	Exome
1000 Genomes	454	Ssaha2	16	239	Genome
X Prize	illumina HiSeq	Novoalign	37	100	Genome
X Prize	SOLID 4	Lifescape	24	40	Genome
Complete Genomics	Complete Genomics	CGTools 2.0	73	33	Genome
Broad	illumina HiSeq	Bwa	68	93	Genome
Broad	illumina HiSeq	Bwa	66	66	Exome
illumina	illumina HiSeq	CASAVA	80	100	Genome
illumina	illumina HiSeq – PCR-free	Bwa	56	99	Genome
illumina	illumina HiSeq – PCR-free	Bwa	190	99	Genome



<http://bcb.io/2013/02/06/an-automated-ensemble-method-for-combining-and-evaluating-genomic-variants-from-multiple-callers/>

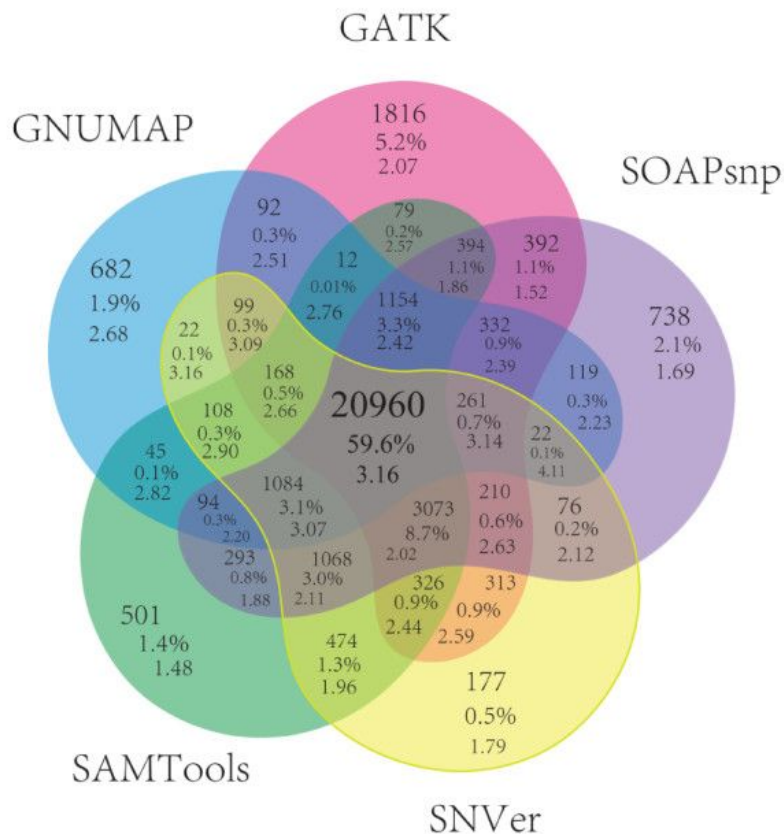
# Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing (1) - SNP

A)

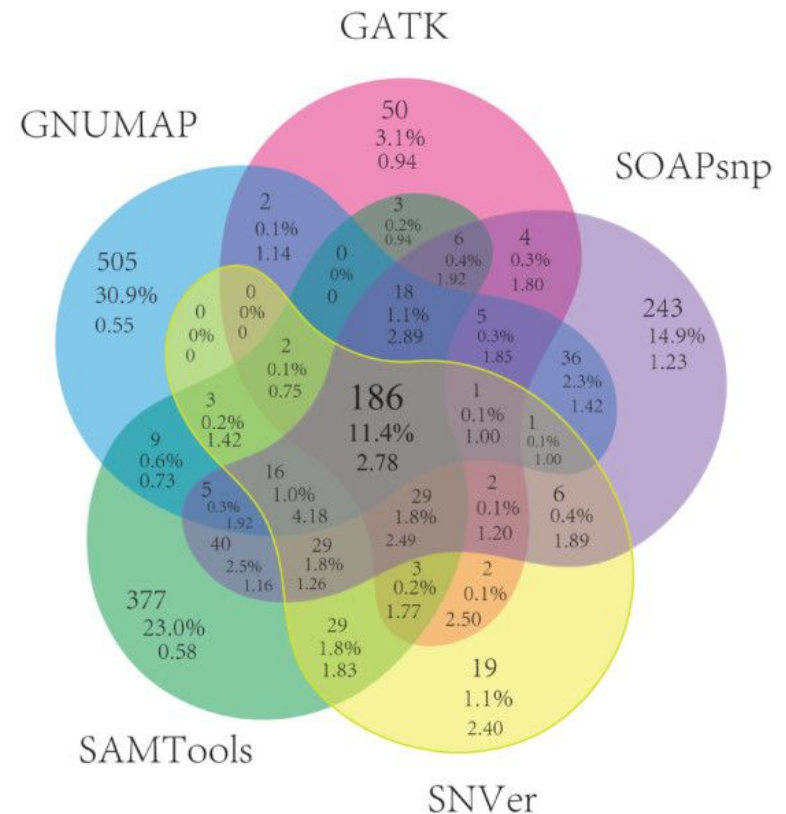


# Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing (2) - SNP

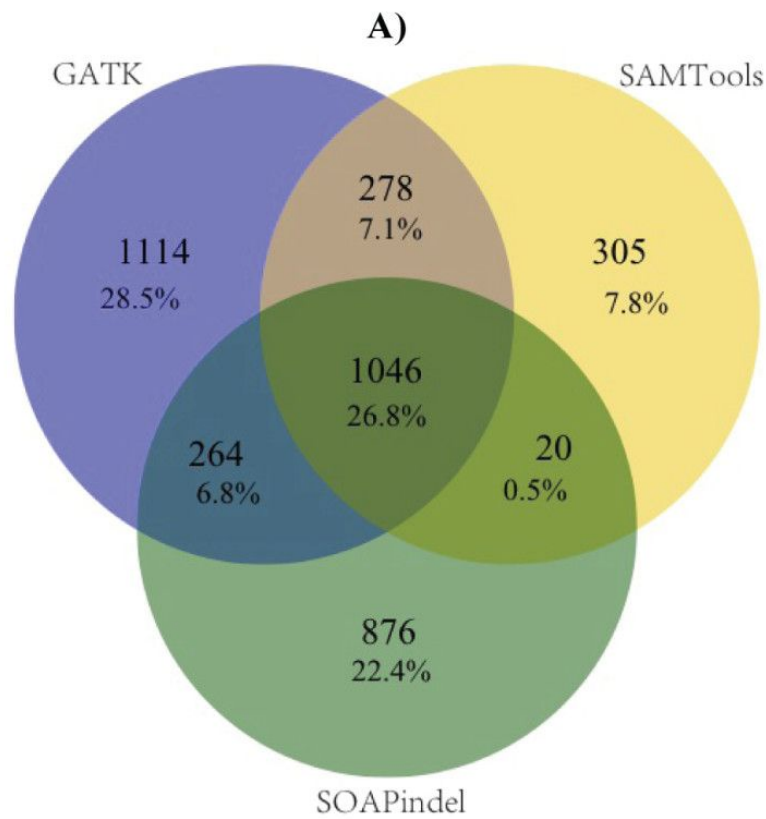
B)



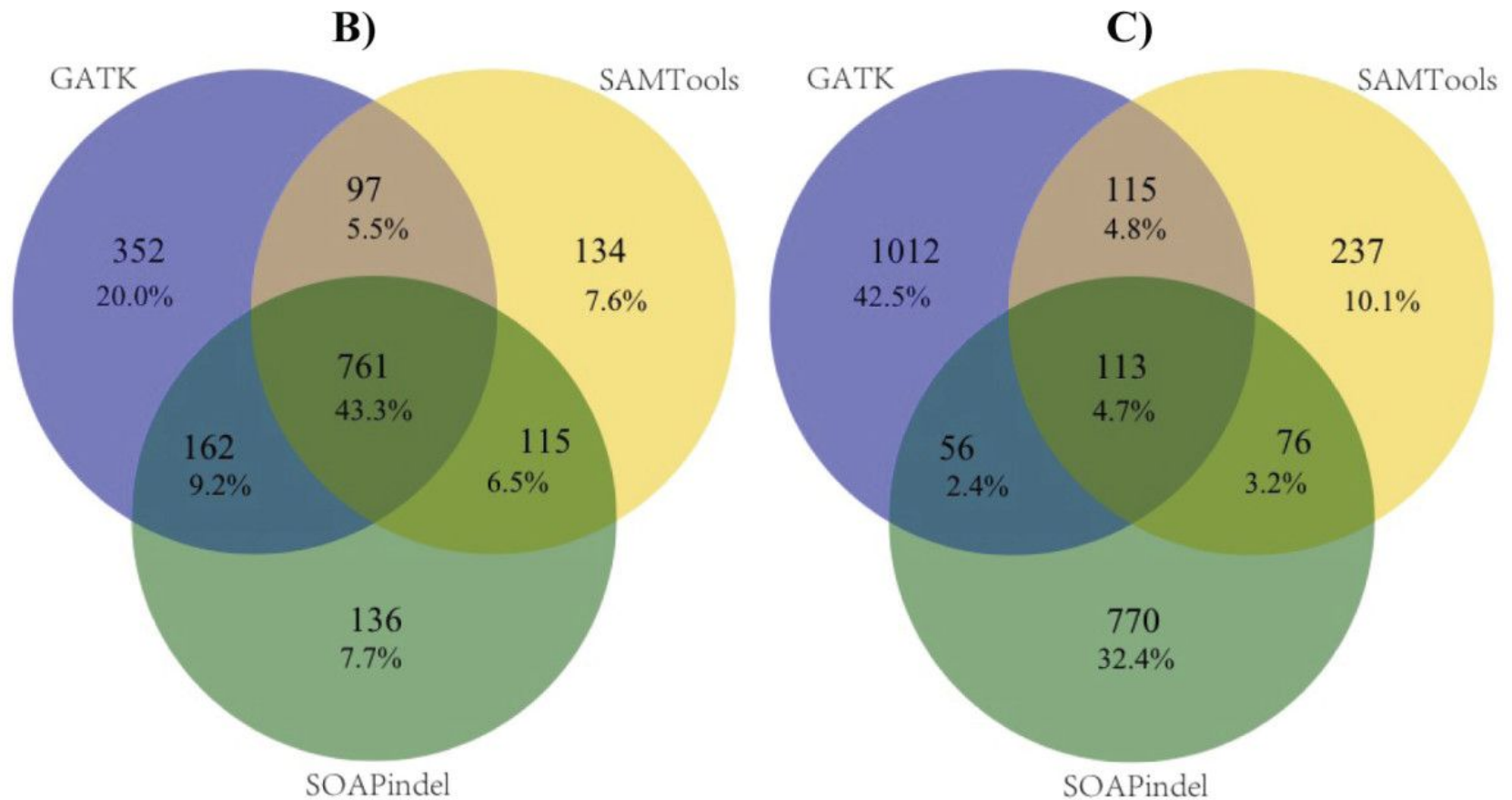
C)



# Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing (3) - InDels



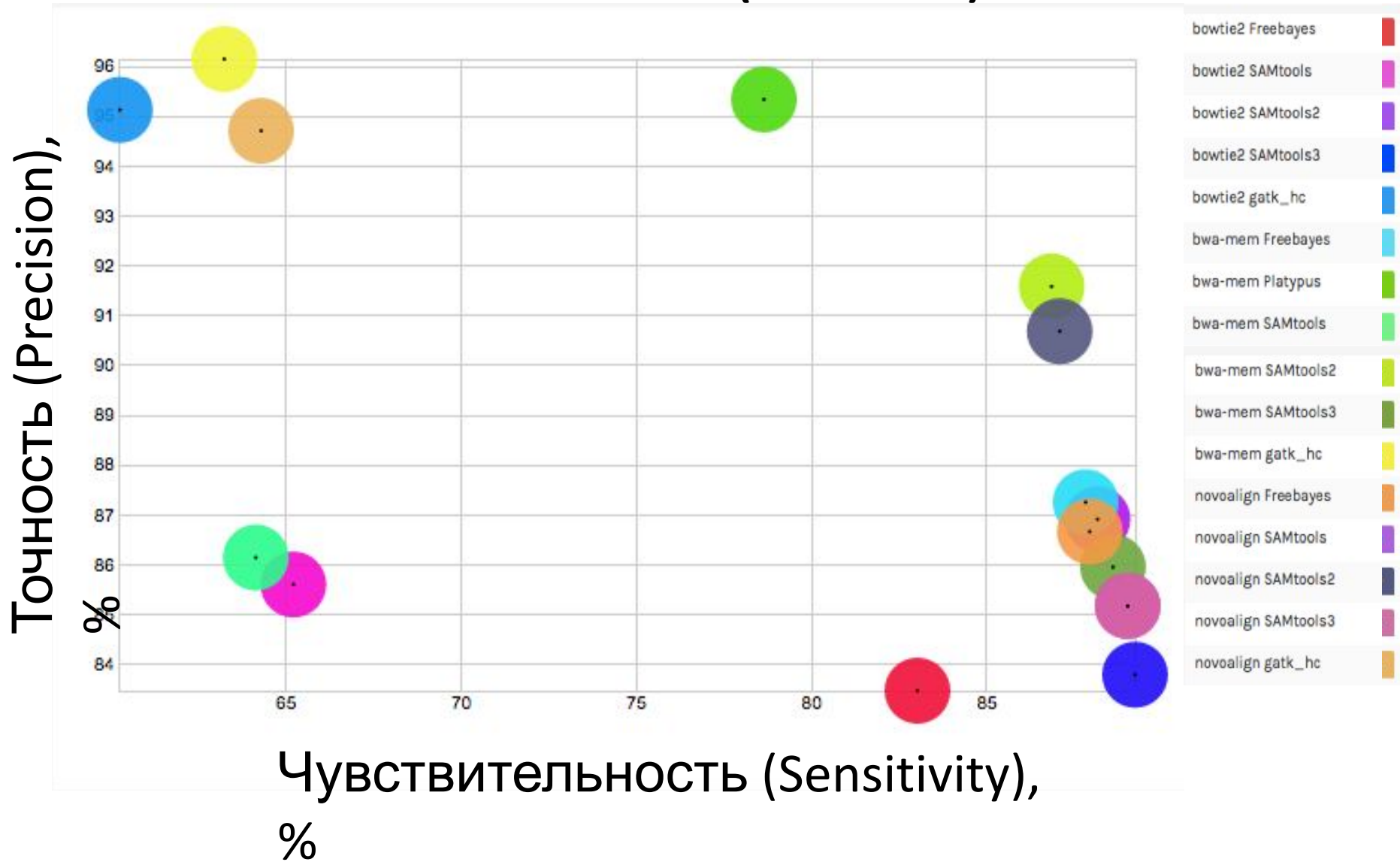
# Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing (4) - InDels



# Что мы решили проверить?

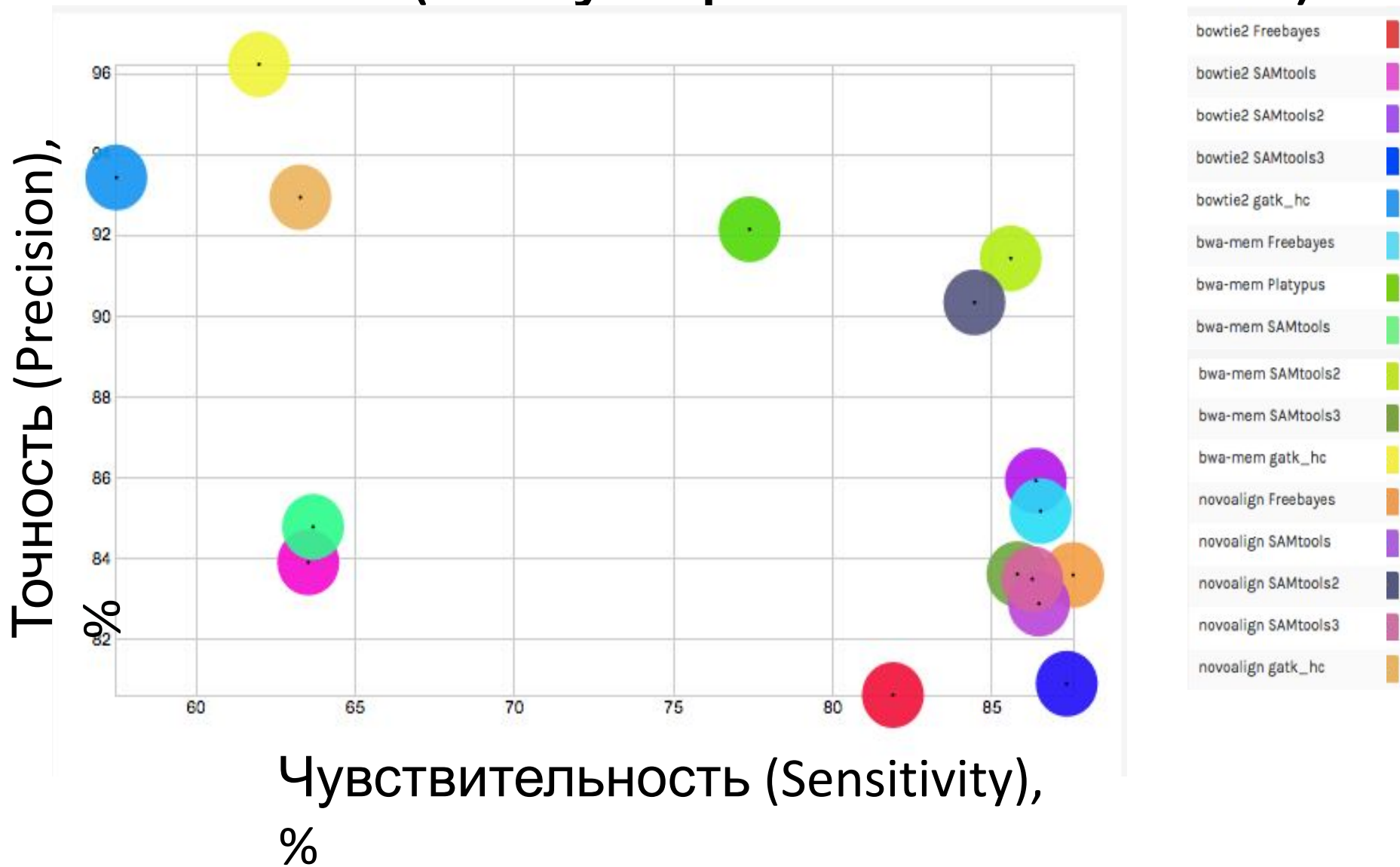
- Bowtie 2 (version 2.1.0, <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>)
- BWA-MEM (version 0.7.8, <http://bio-bwa.sourceforge.net/>)
- Novoalign (version 3, <http://www.novocraft.com/products/novoalign/>)
- GATK Haplotype Caller (<https://www.broadinstitute.org/gatk/>)
- SAMtools (version 0.2.0, <http://samtools.sourceforge.net/>)
- FreeBayes (version v0.9.21, <https://github.com/ekg/freebayes/>)
- Platypus (<http://www.well.ox.ac.uk/platypus>)

# Результаты исследования iVinom NA12878 (ЭКЗОМ)

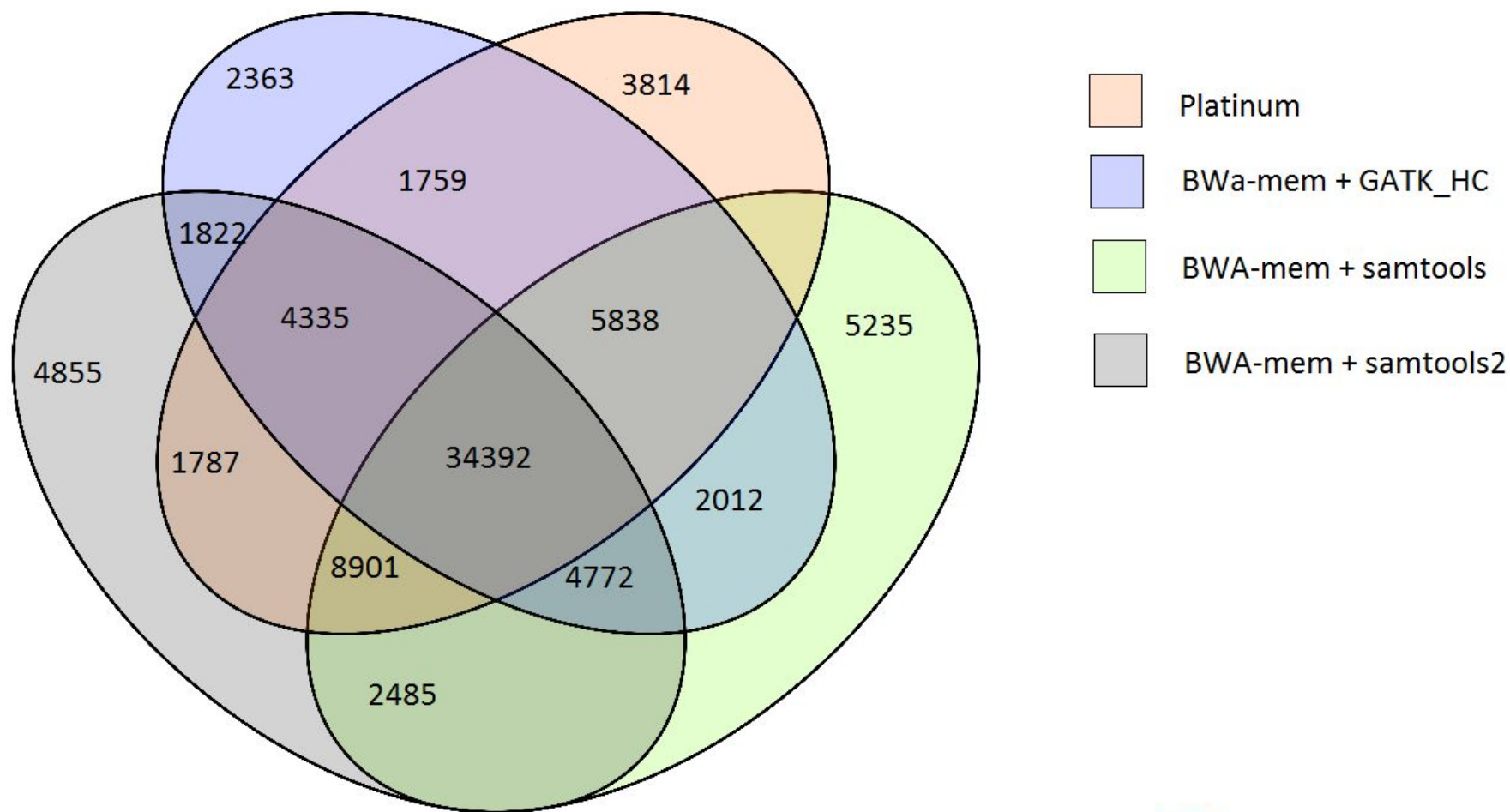




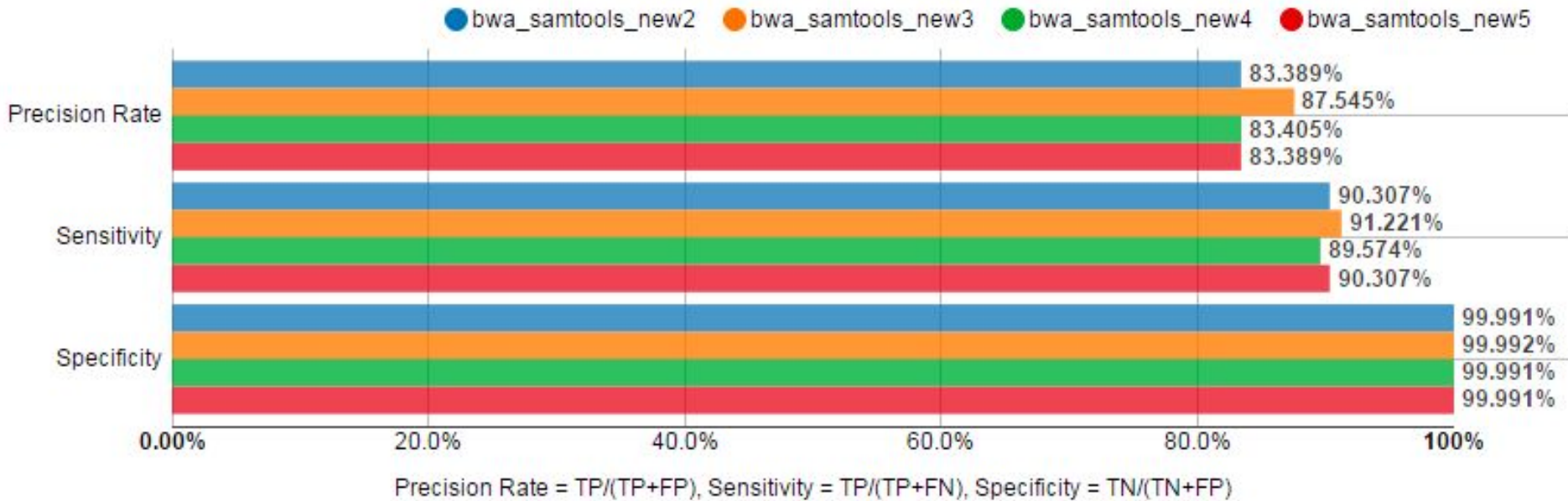
# Результаты исследования iVinom NA12877 (симулированный экзом)



# Результаты исследования iVinom – образец NA12878, общие SNP



# Пара слов о важности настроек



- `bwa_samtools_new2 -- "call -c" (без специальных опций)`
- `bwa_samtools_new3 -- "call -p 0.2 -c --output-type v -v -"`
- `bwa_samtools_new4 -- "call -p 0.5e-2 -c --output-type v -v -"`

# Выводы исследования iVinom

- Не всегда 2 хороших тула хорошо работают вместе (пример: BWA-MEM + GATK HC)
- Нравящиеся нам пайплайны:  
BWA-MEM+Samtools 2 и Novoalign+Samtools

# Почему разные пайплайны дают столь отличающиеся результаты?

- Потому что входящие в пайплайны блоки варьируются, меняя условия для принятия конечного решения о мутации
- До 30% SNP и InDels лежат как раз в этой области неопределённости.
- Если немного пошевелить исходные условия (покрытие, качество нуклеотидов), изменяется результат коллинга.

# Как проверить свои результаты

The screenshot displays the bioplanet.com website interface. At the top, the bioplanet.com logo is on the left, and navigation links for HOME, DISCUSSION, JOBS, RESOURCES, GCAT, LOGIN, and REGISTER are on the right. Below the navigation bar, a circular logo for GCAT (Genome Comparison & Analytic Testing) is on the left, and a horizontal menu with links for Home, Start Test, Reports, Discuss, About, and Advisors is on the right. The main content area features a five-step workflow diagram for testing and comparing in-house genome analysis pipelines. The steps are: 1. Download Test Data (FastQ), 2. Analyze Genome (Variant Calling, RAW reads, Alignment), 3. Upload Results (GCAT 94), 4. Explore Reports (GCAT 94), and 5. Community Discussion. A green arrow labeled 'Start A Test' points to the first step.

bioplanet.com

HOME DISCUSSION ▾ JOBS ▾ RESOURCES ▾ GCAT LOGIN REGISTER

GENOME COMPARISON  
**GCAT**  
& ANALYTIC TESTING

Home Start Test Reports Discuss About Advisors

Test & Compare your in-house genome analysis pipeline!

FastQ

Variant Calling  
RAW reads  
Alignment

GCAT 94

Community Discussion

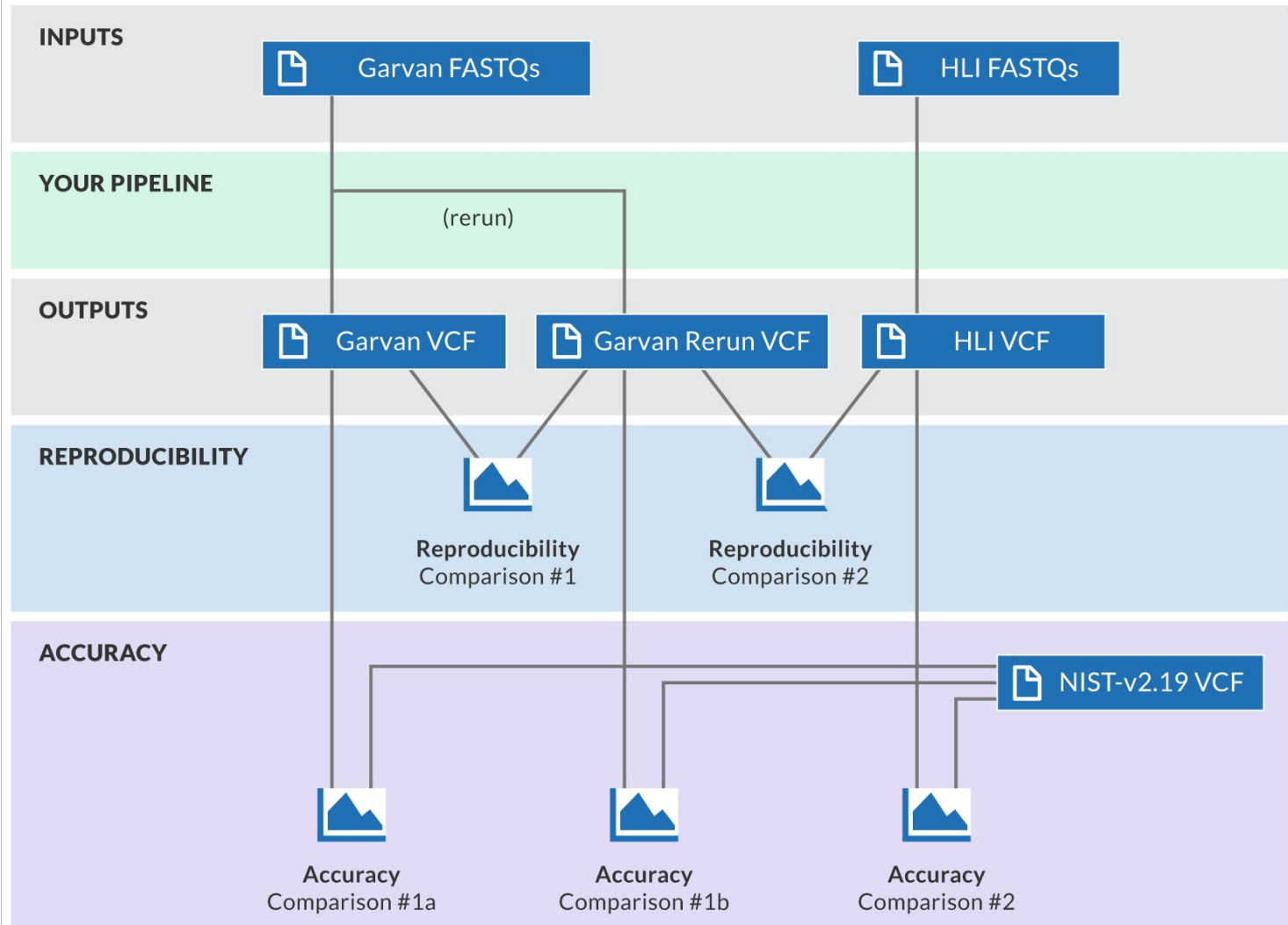
Start A Test

- 1**  
**Download Test Data**  
Choose from a variety of different NGS platforms.
- 2**  
**Analyze Genome**  
You process the data locally using the tools of your choice.
- 3**  
**Upload Results**  
GCAT instantly analyzes your results in the cloud.
- 4**  
**Explore Reports**  
Visualize your results and compare to others.
- 5**  
**Community Discussion**  
Discuss reports and shape the direction of GCAT.

<http://www.bioplanet.com/gcat>

# PrecisionFDA Challenge

February 25, 2016 through April 25, 2016



# Тестовые файлы

Dataset/Site	Garvan	HLI
Contributor	Garvan Institute of Medical Research	Human Longevity, Inc.
Files	<a href="#">NA12878-Garvan-Vial1_R1.fastq.gz</a> <a href="#">NA12878-Garvan-Vial1_R2.fastq.gz</a>	<a href="#">TSNano_1lane_L008_13801_NA12878_R1_001.fastq.gz</a> <a href="#">TSNano_1lane_L008_13801_NA12878_R2_001.fastq.gz</a>
Library Prep	TruSeq Nano DNA Library Prep kit, v2.5, with no sample multiplexing	TruSeq Nano NA12878 library, sequenced with the original V1 chemistry
Read Length	2x150bp	2x150bp
Insert Size	350bp	319bp
Instrument	HiSeq X (one lane)	HiSeq X (one lane)

Внимание! Размер каждого сжатого файла около 50 Гб



# Проблемы

- Невоспроизводимость результатов одного и того же пайплайна!
  - Многие коллеры используют вероятностные модели
  - В силу вероятностной природы результаты 2 запусков одного и того же пайплайна РАЗЛИЧАЮТСЯ
- Проблемы с референсом (даже PrecisionFDA Challenge рекомендует GRCh37)

# Как теперь с этим жить?

