Сравнение точности пайплайнов обработки NGS

Андрей Афанасьев,
CEO@iBinom

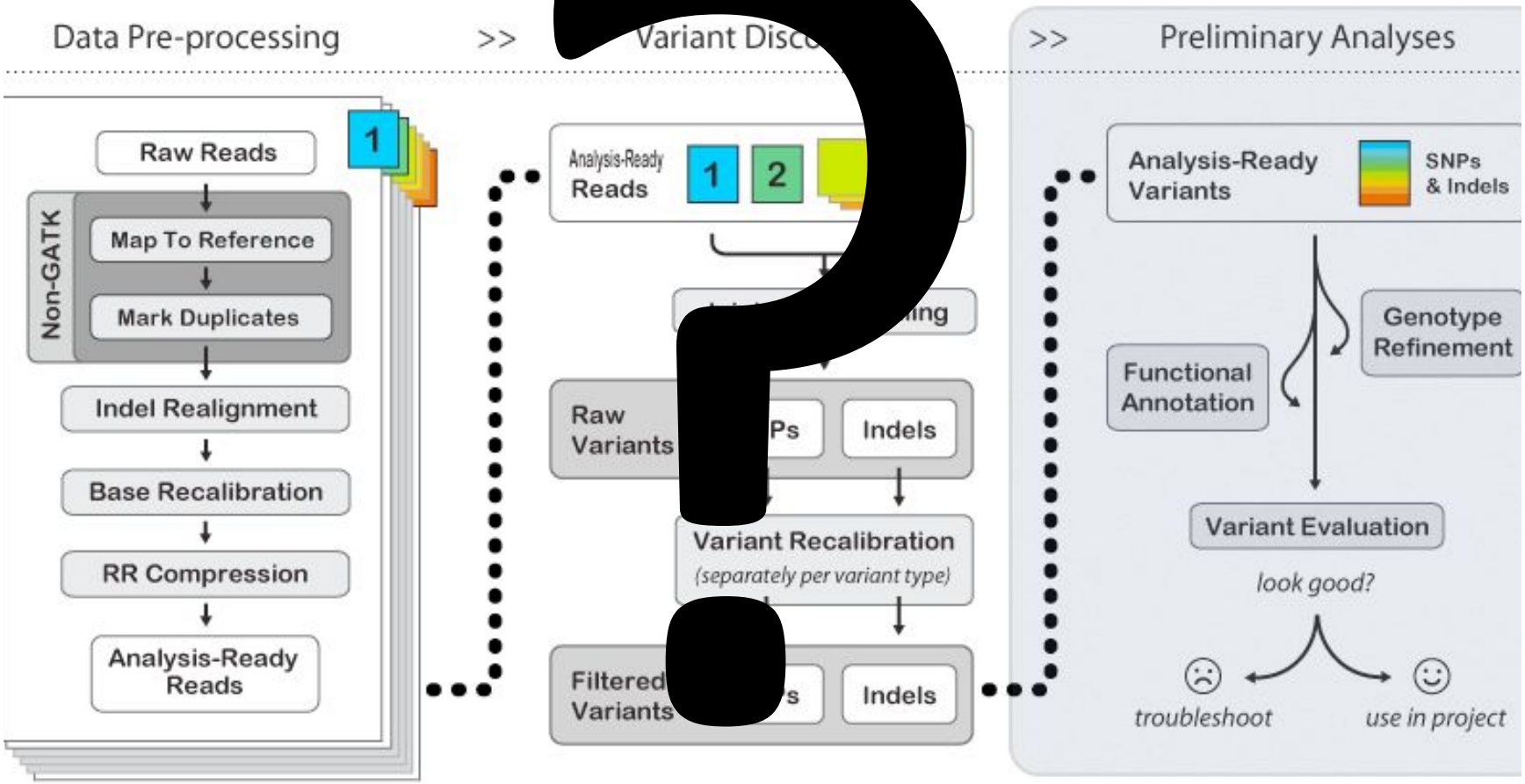
Зачем это всё?

- 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

BINOM

Есть ли стандартный пайплайн?

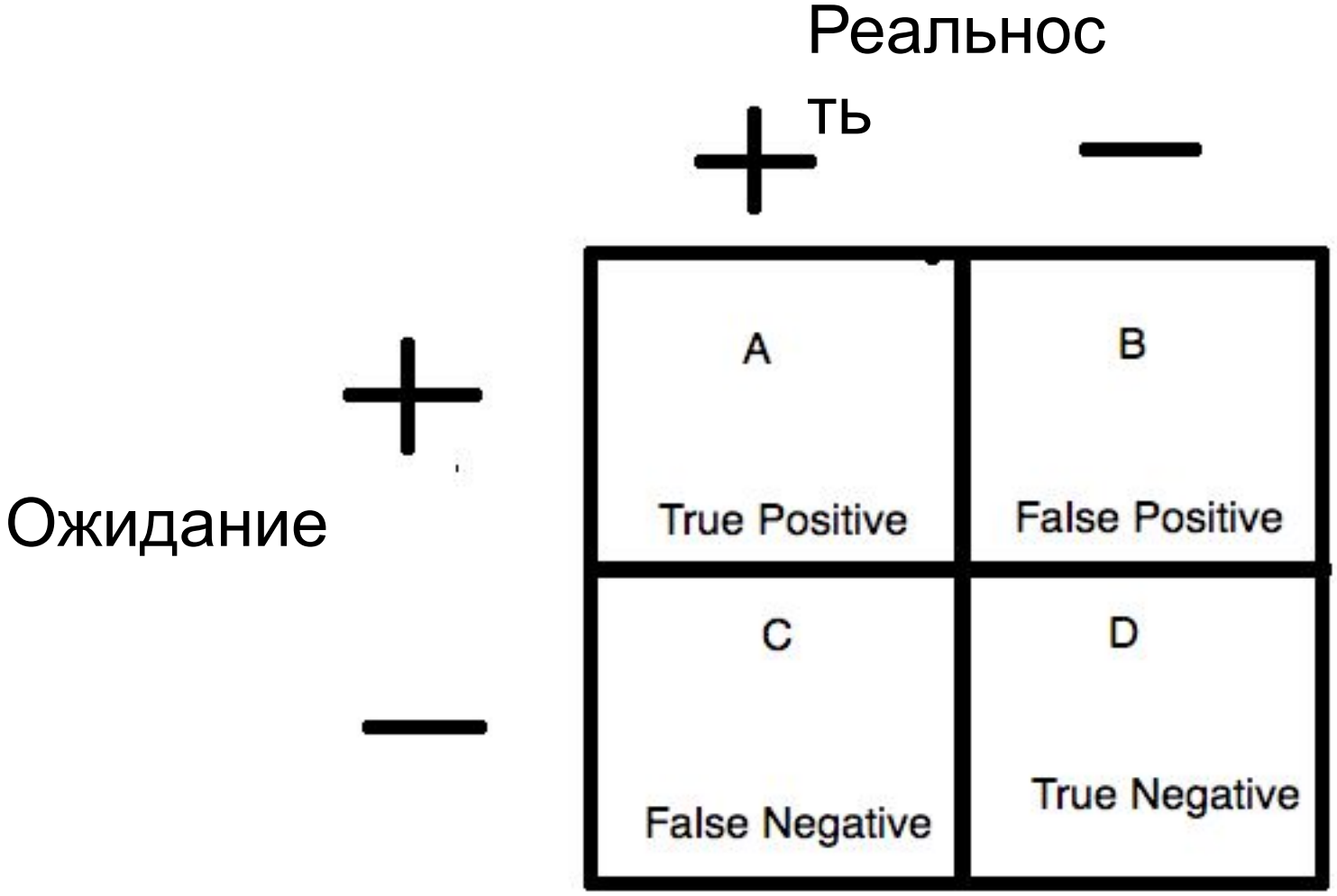


BINOM

Нельзя просто так взять и получить vcf файл!

BINOM

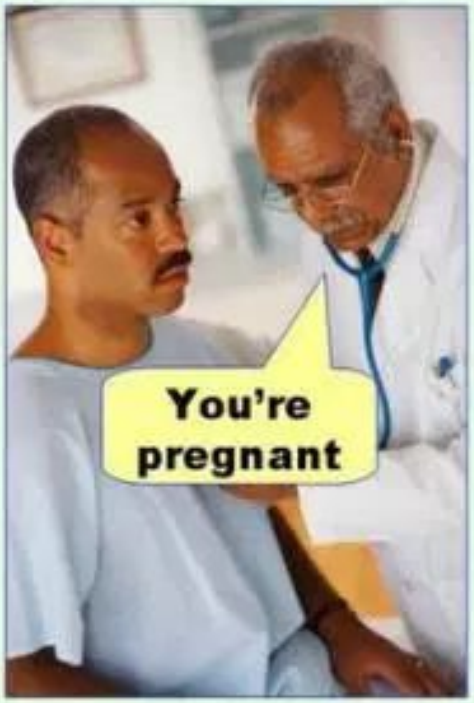
Что мы измеряем?



BINOM

Что мы измеряем?

Type I error
(false positive)



Type II error
(false negative)

BINOM

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

BINOM

«Золотой» образец NIST

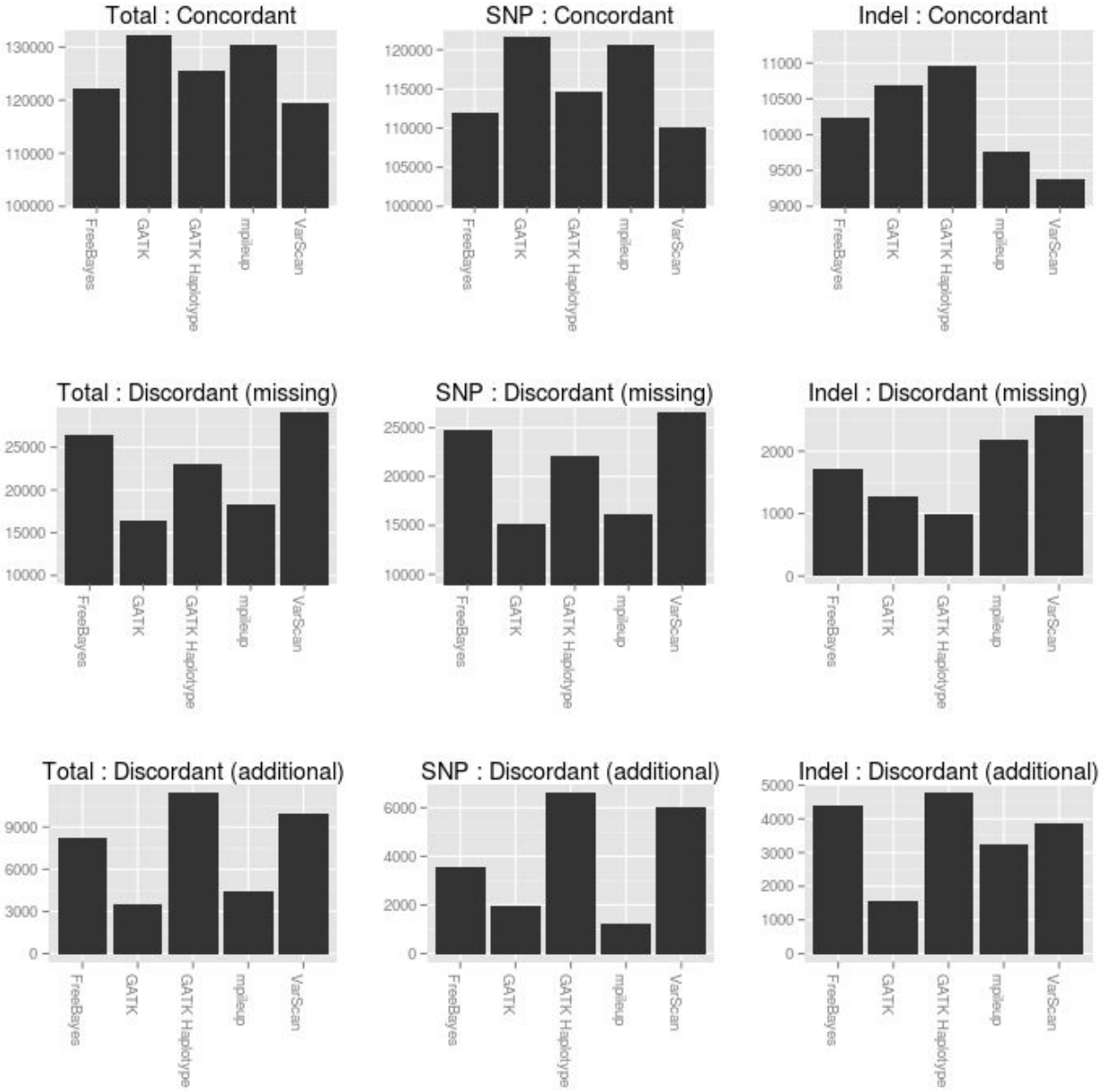
- Семья из Юты: NA12878 Genome in a Bottle
- ОЧЕНЬ хорошо охарактеризован

BINOM

«Золотой» образец NIST

Source*	Platform	Mapping algorithm	Cov-erage	Read length	Genome/ exome
1000 Genomes	Illumina GaIIx	Bwa	39	44	Genome
1000 Genomes	Illumina GaIIx	Bwa	30	54	Exome
1000 Genomes	454	Ssaha2	16	239	Genome
X Prize	Illumina HiSeq	Novoalign	37	100	Genome
X Prize	SOLiD 4	Lifescope	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

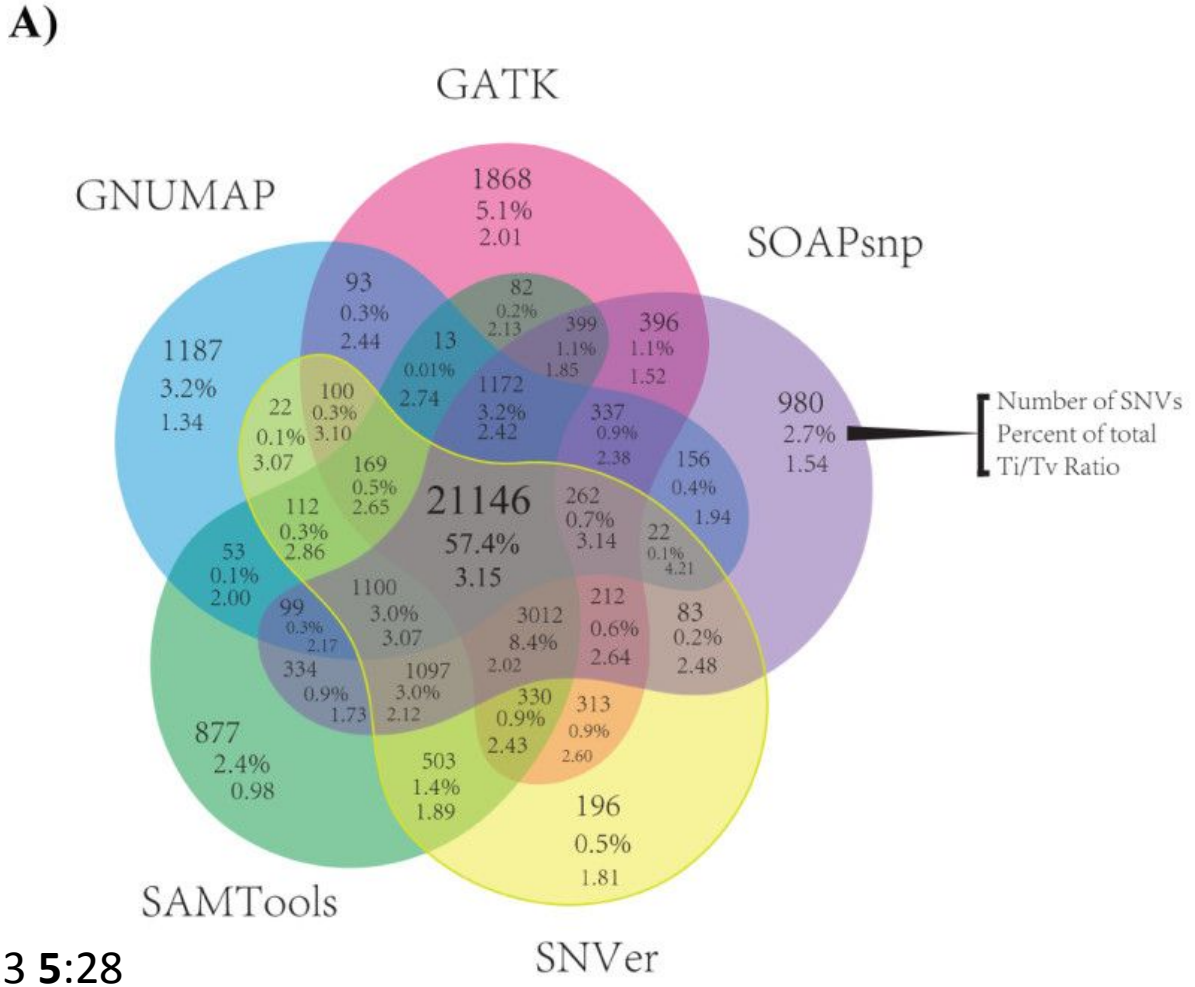
BINOM



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

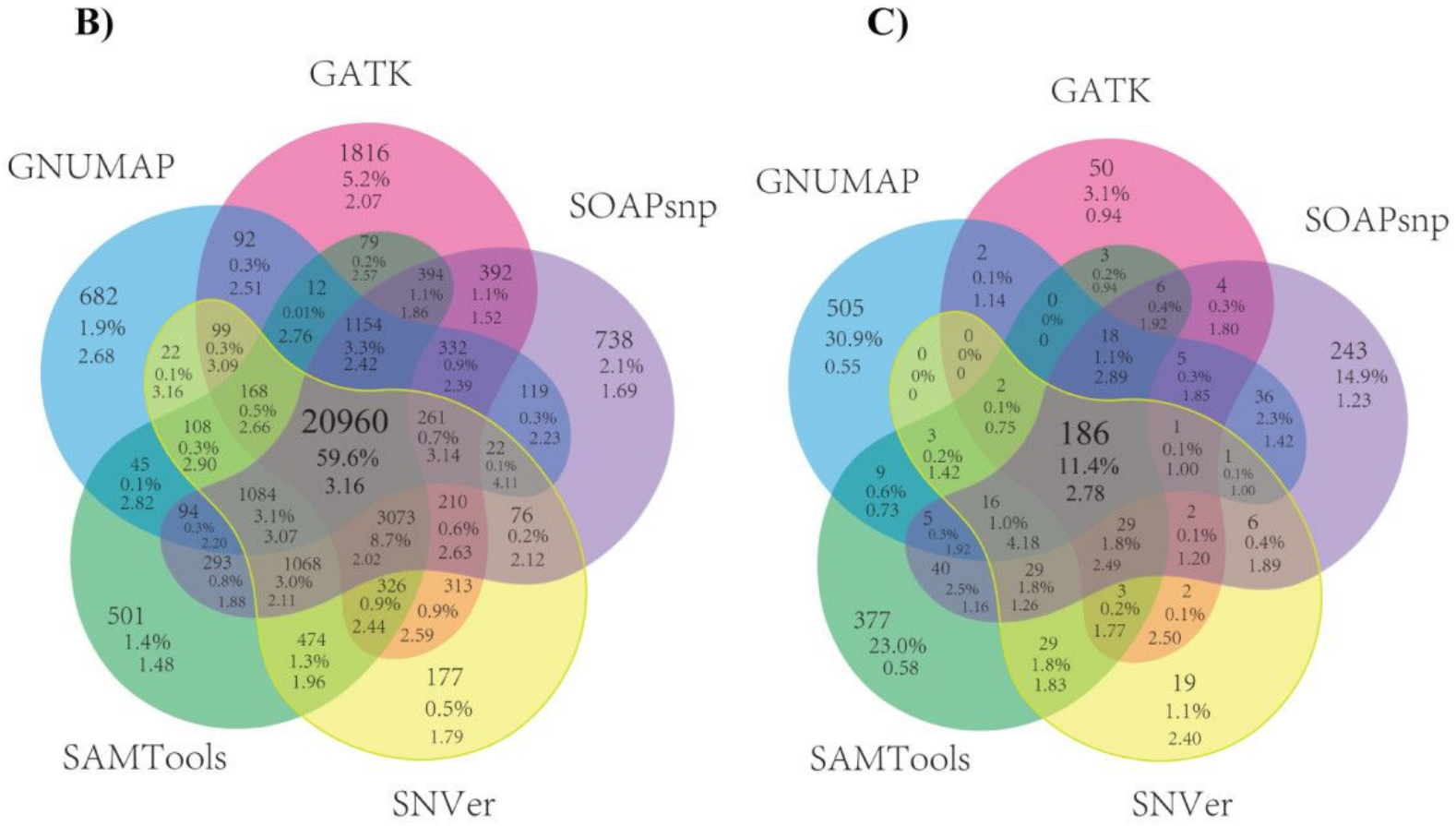
BINOM

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



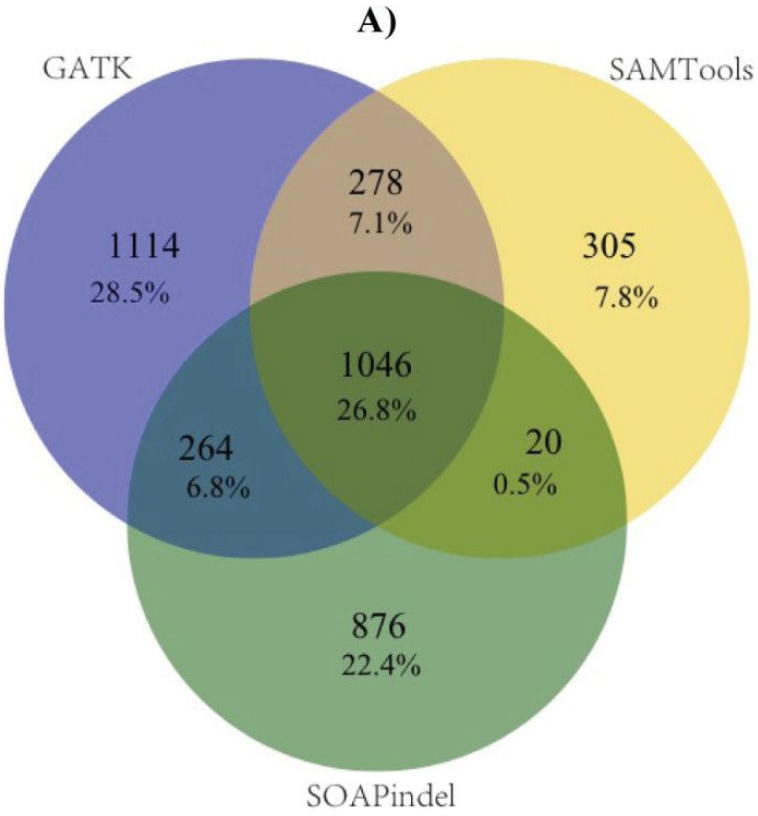
Genome Medicine 2013 **5**:28
DOI: 10.1186/gm432

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



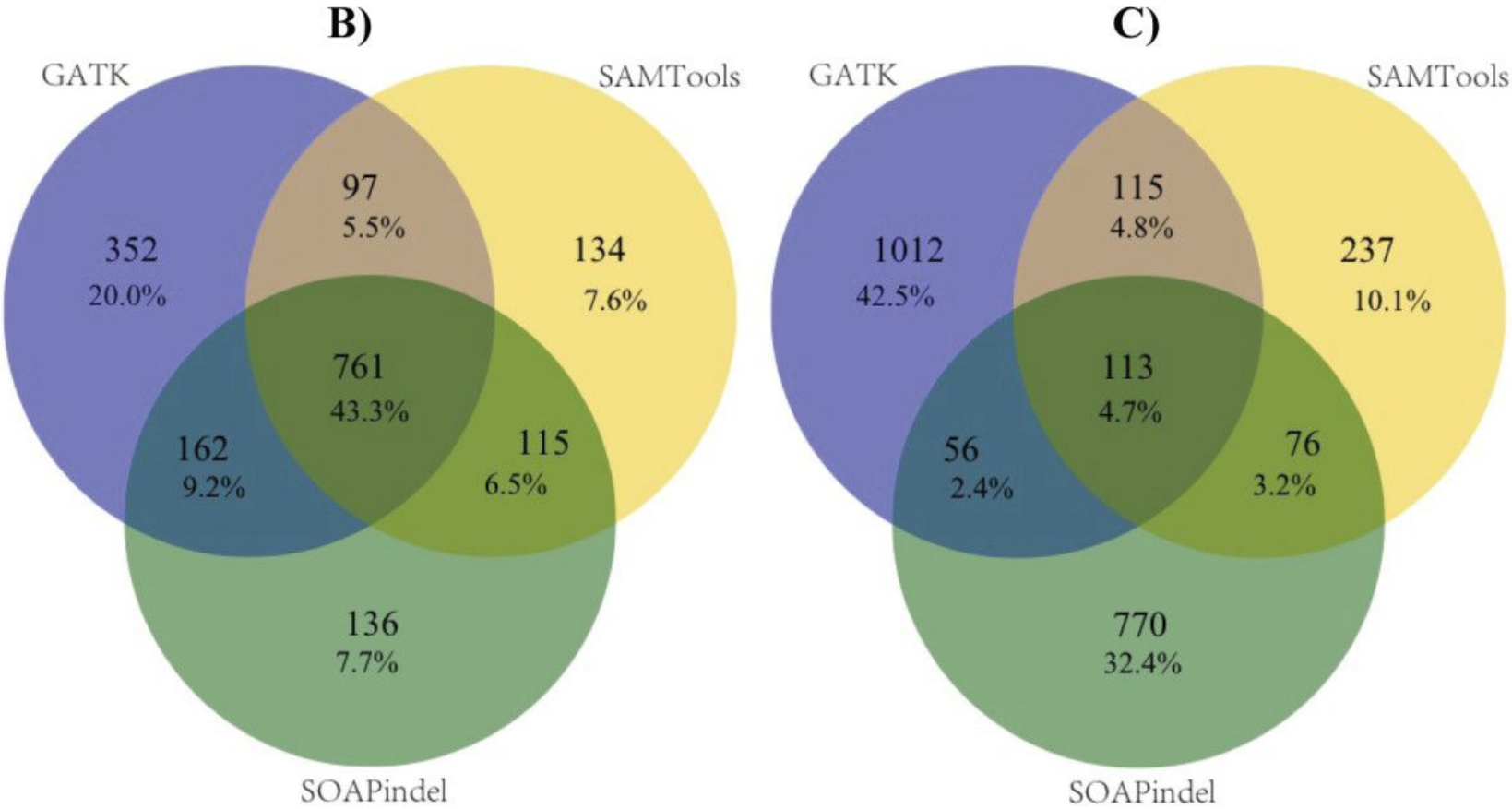
Genome Medicine 2013 **5**:28
DOI: 10.1186/gm432

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



Genome Medicine 2013 **5**:28
DOI: 10.1186/gm432

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



Genome Medicine 2013 **5**:28
DOI: 10.1186/gm432

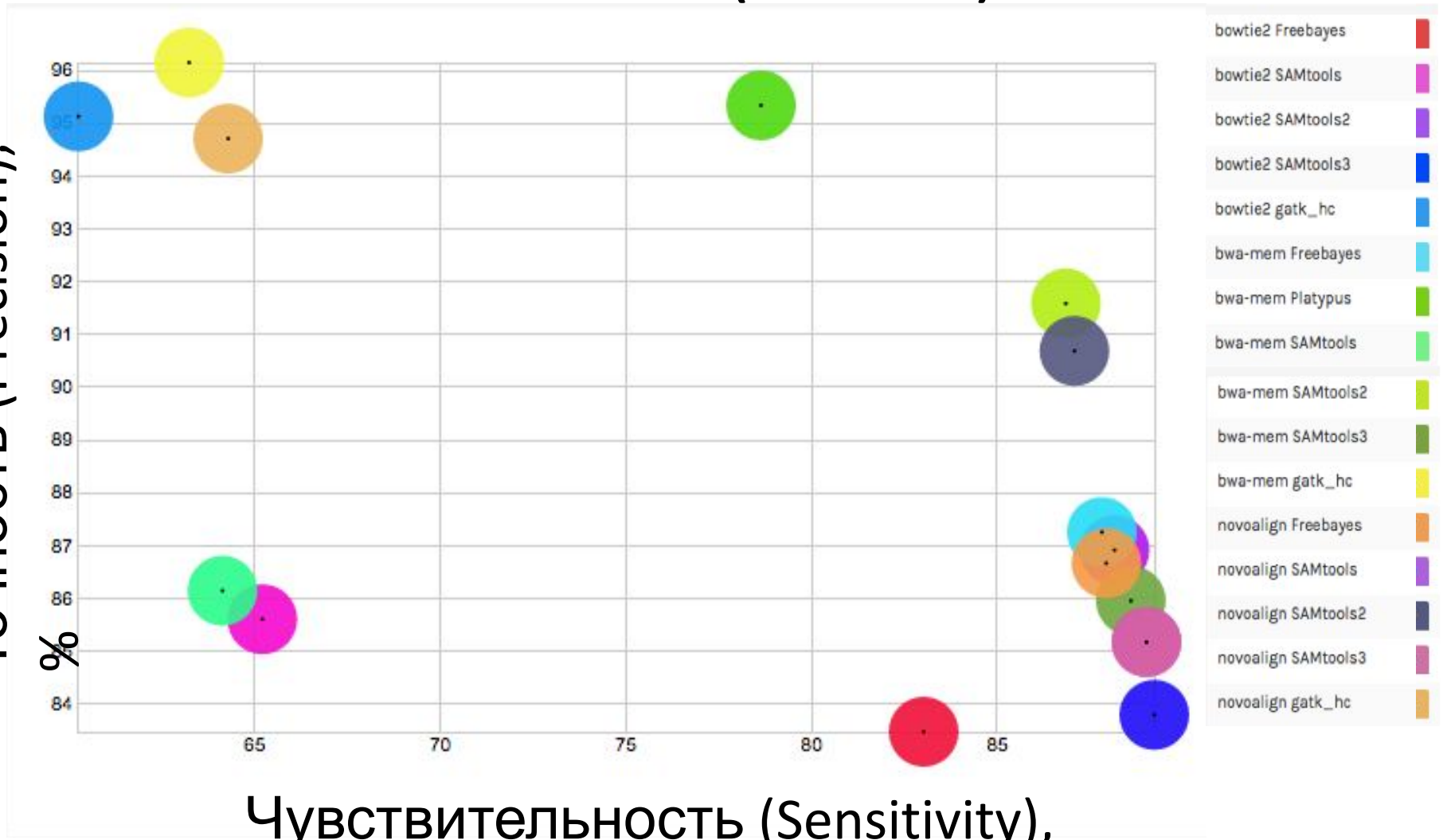
Что мы решили проверить?

- 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/)

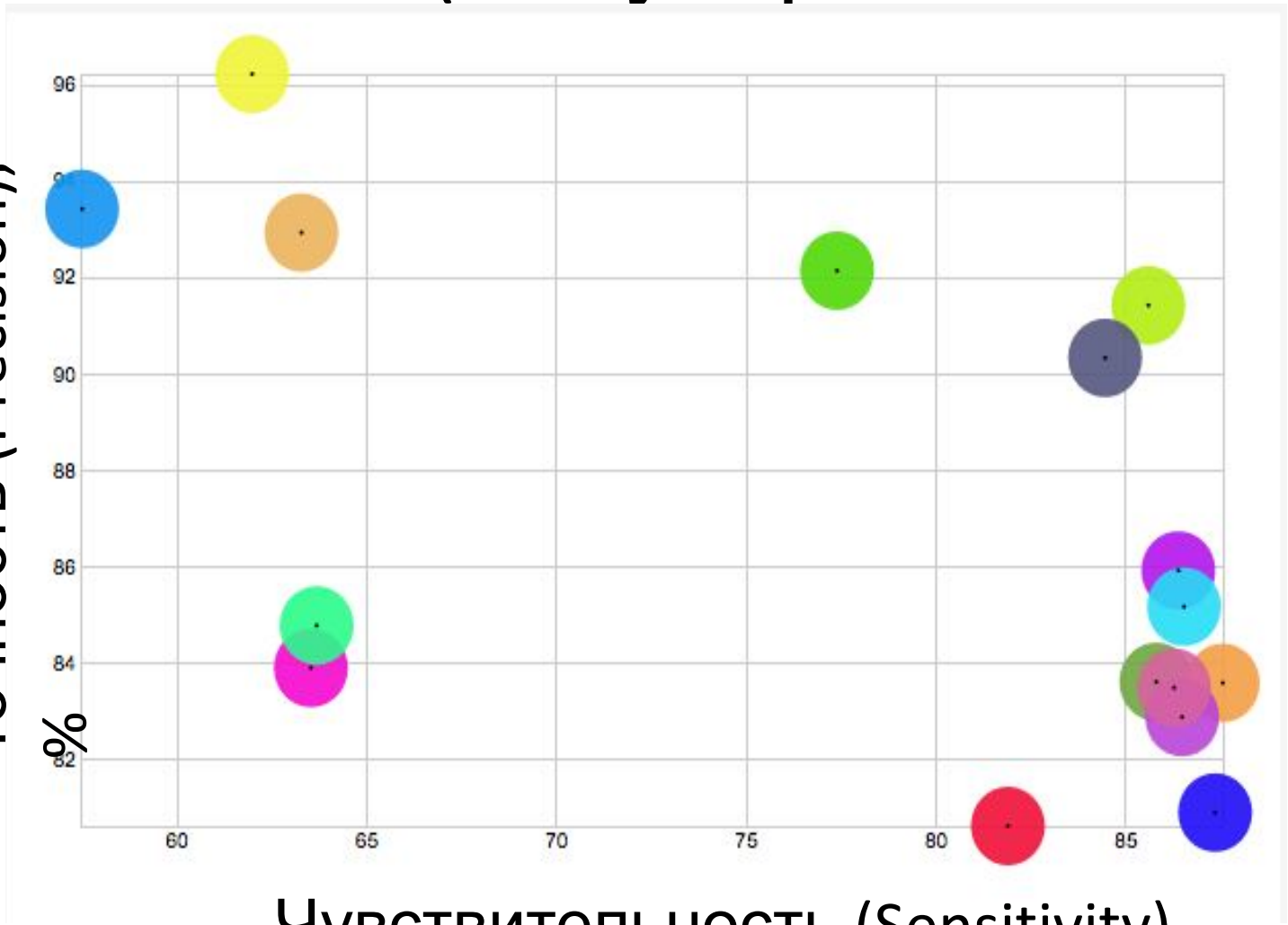
- Platypus (http://www.well.ox.ac.uk/platypus)

BINOM

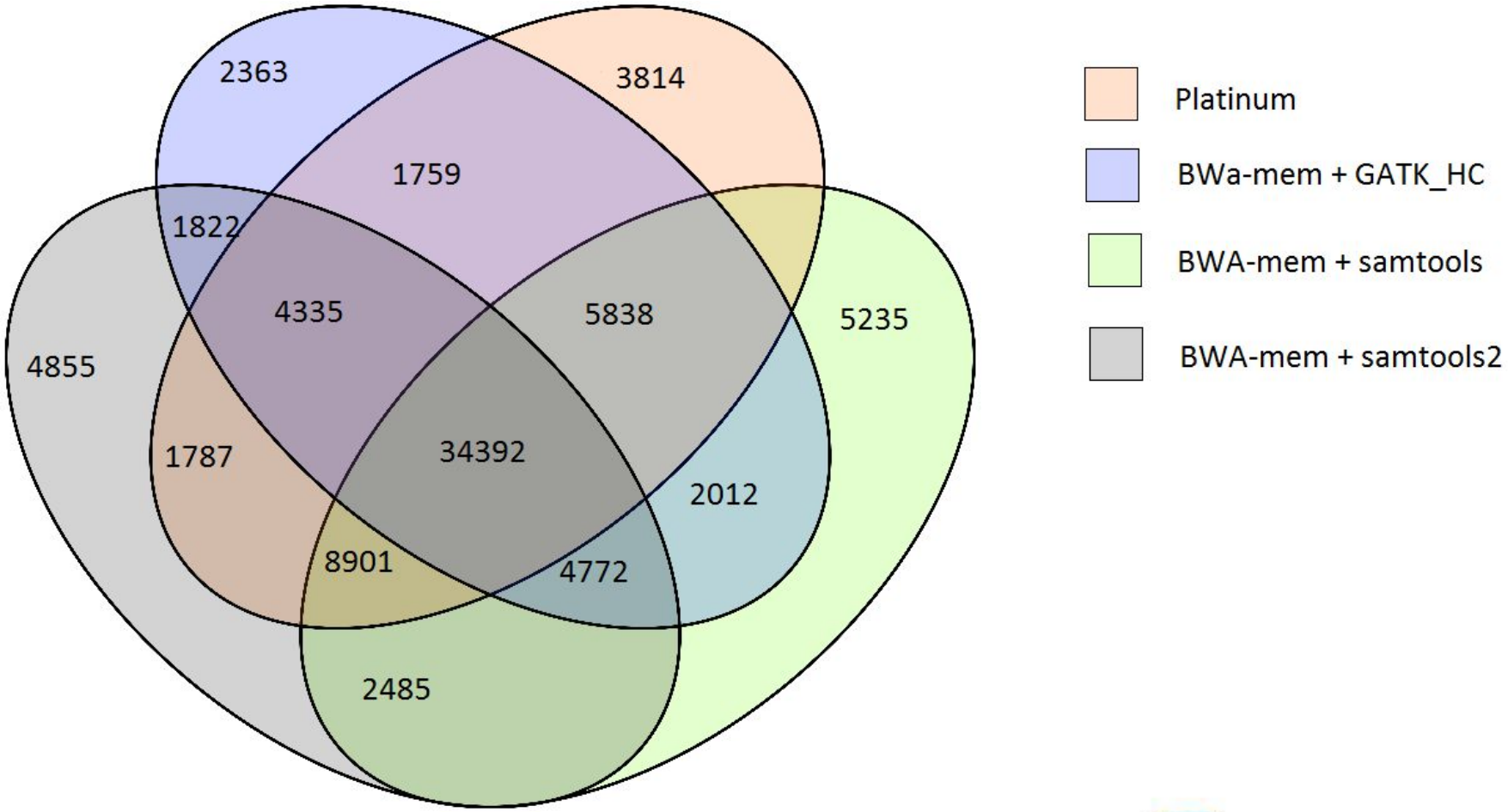
Результаты исследования iBinom NA12878 (экзом)



Результаты исследования iBinom NA12877 (симулированный экзом)

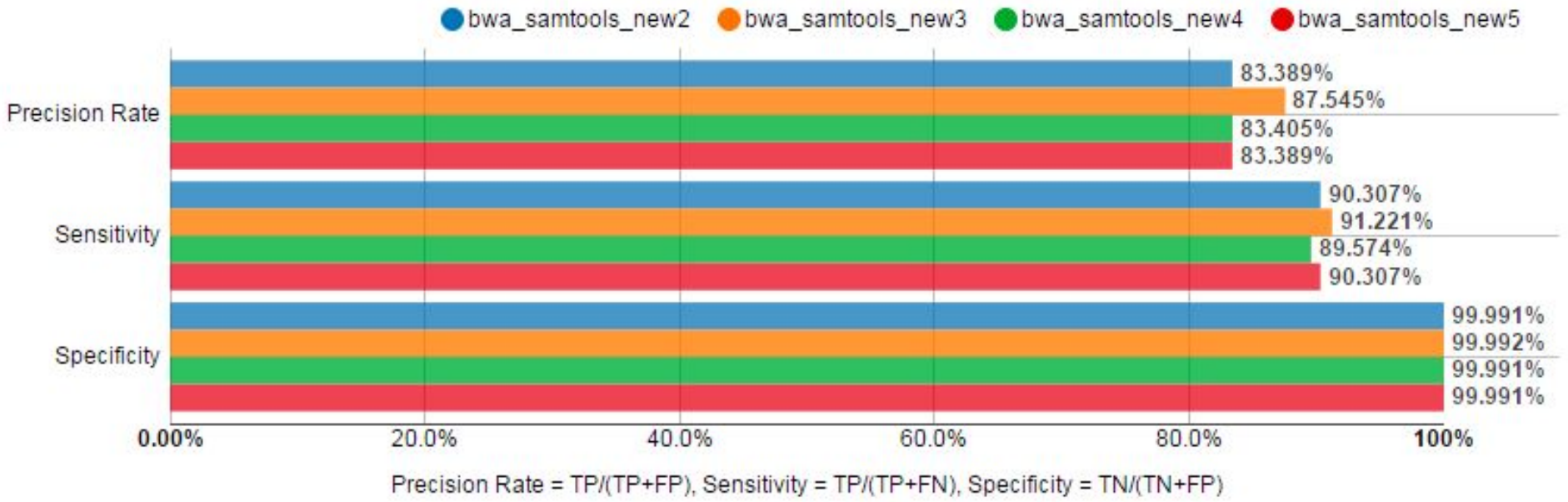


Результаты исследования iBinom – образец NA12878, общие SNP



iBINOM

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 iBinom

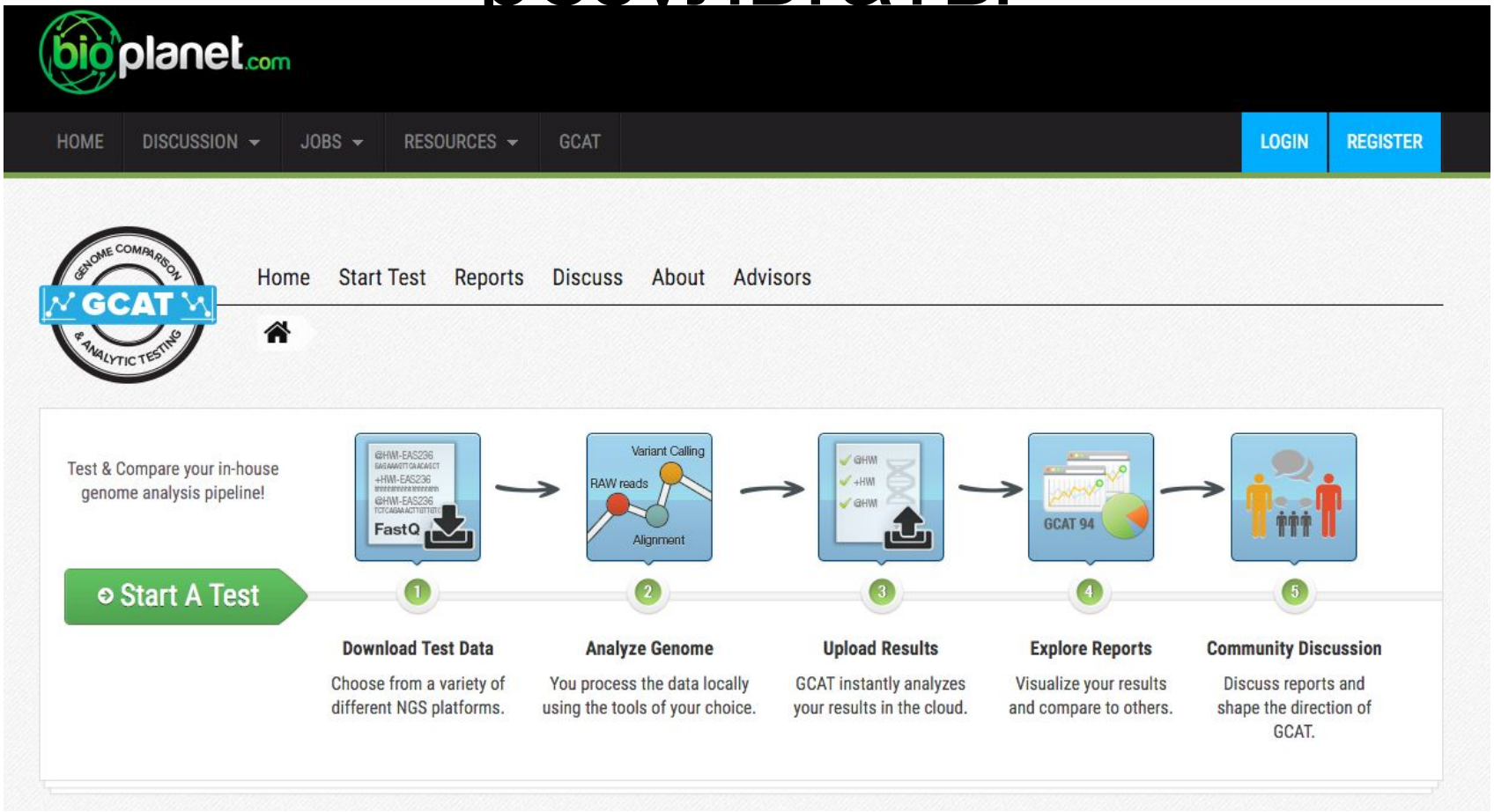
- Не всегда 2 хороших тула хорошо работают вместе (пример: BWA-MEM + GATK HC)
- Нравящиеся нам пайплайны: BWA-MEM+Samtools 2 и Novoalign+Samtools

iBINOM

Почему разные пайплайны дают столь отличающиеся результаты?

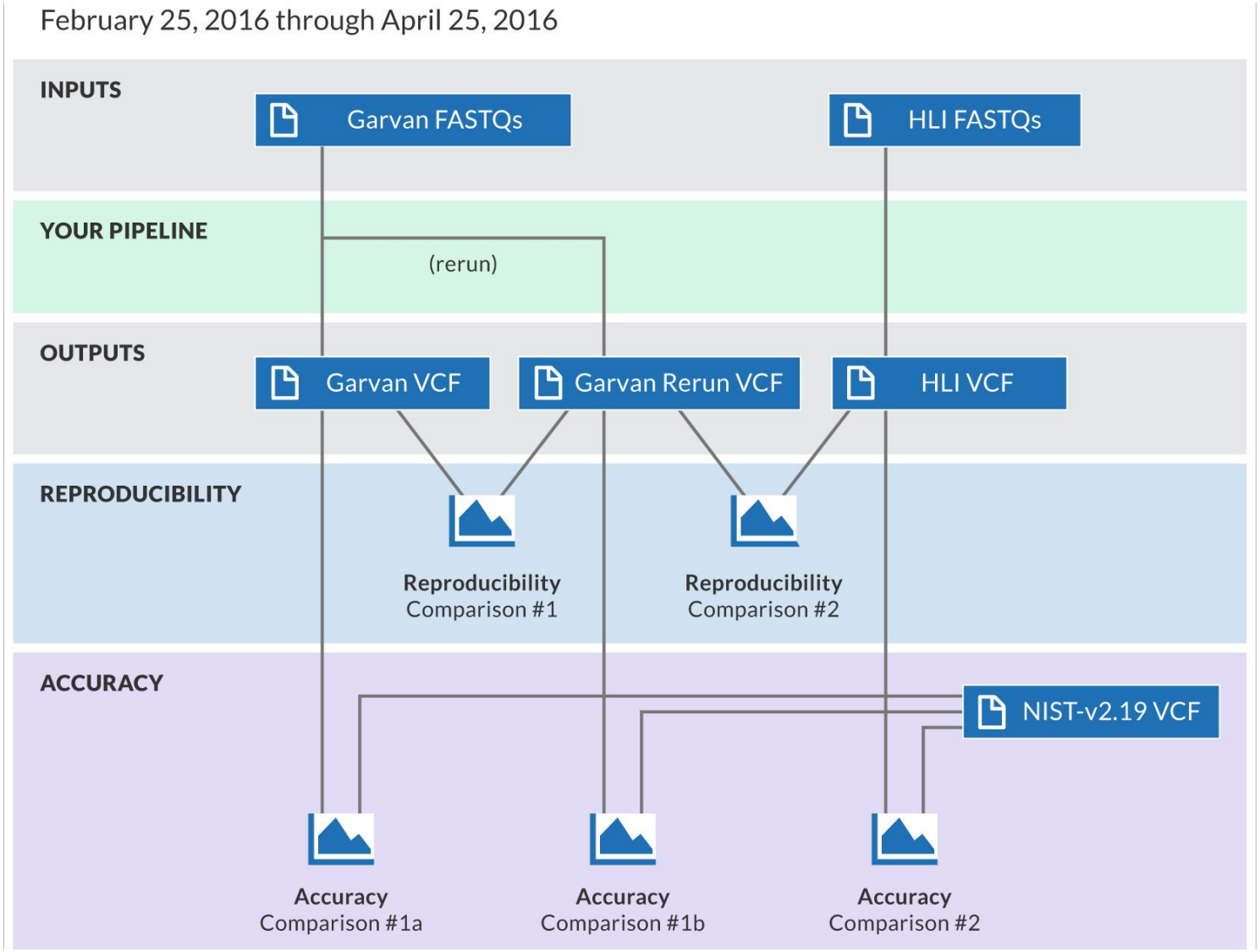
- Потому что входящие в пайплайны блоки варьируются, меняя условия для принятия конечного решения о мутации
- До 30% SNP и InDels лежат как раз в этой области неопределённости.
- Если немного пошевелить исходные условия (покрытие, качество нуклеотидов), изменяется результат коллинга.

Как проверить свои результаты



http://www.bioplanet.com/gcat

PrecisionFDA Challenge



Тестовые файлы

Dataset/Site	Garvan	HLI
Contributor	Garvan Institute of Medical Research	Human Longevity, Inc.
Files	NA12878-Garvan-Vial1_R1.fastq.gz NA12878-Garvan-Vial1_R2.fastq.gz	TSNano_1lane_L008_13801_NA12878_R1_001.fastq.gz TSNano_1lane_L008_13801_NA12878_R2_001.fastq.gz
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?