Приоритизация генетических вариантов

Студенты: Исаев Василий, Лонишин Любовь
Руководители: Барбитов Юрий, Захаров Геннадий, Сидорук Александр (Институт биоинформатики, EPAM)

Глобальная цель проекта

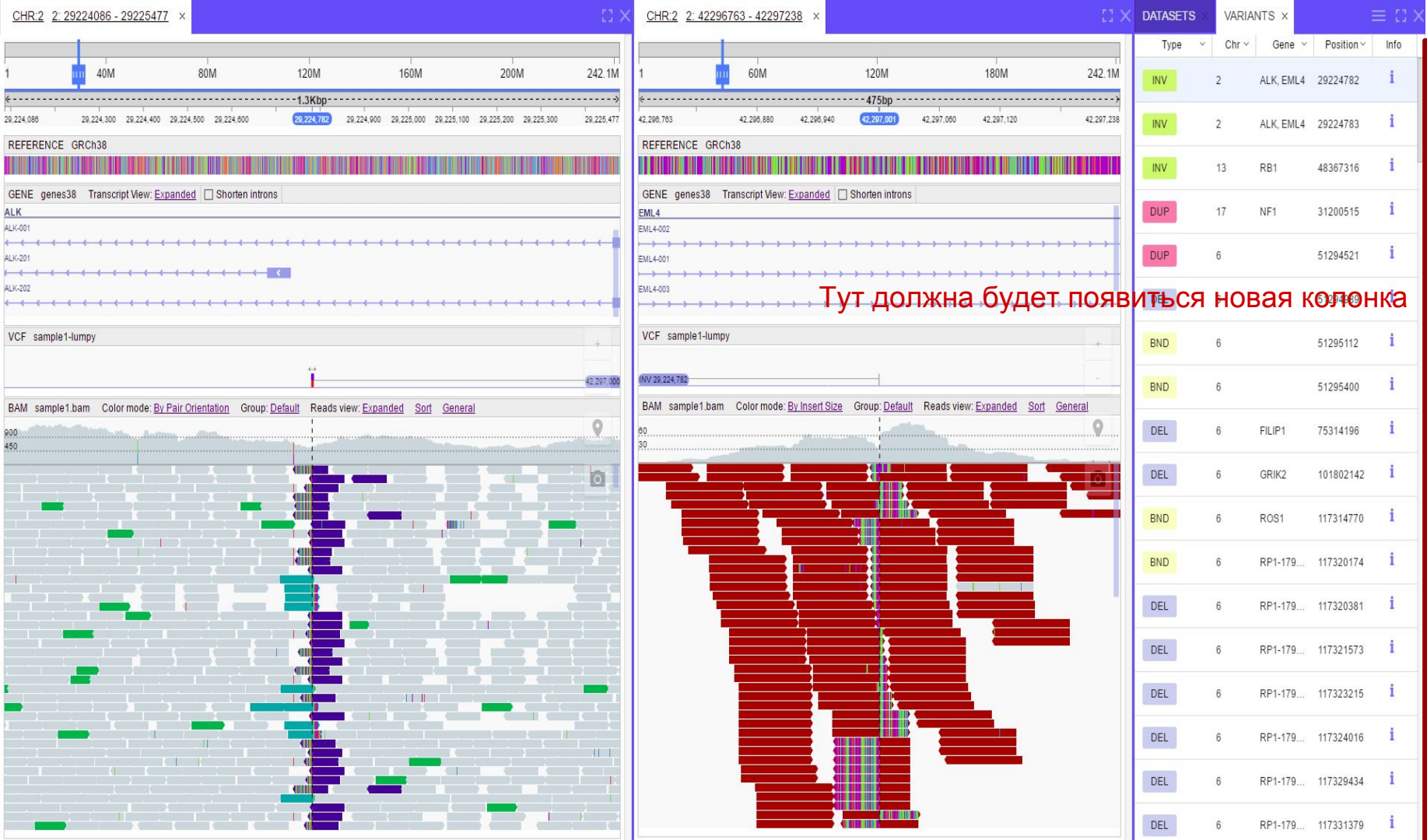
Разработать плагин-сортировщик мутаций по важности для NGB (New Genome Browser)

Локальные цели

1. Написать java-утилиту для добавление в VCF-файл колонки с коэффициентом важности мутаций
2. Модифицировать написанную утилиту для уже обработанных NGB-браузером файлов
3. Встроить написанную утилиту в NGB - браузер

3* Исправить мелкие баги (лексикографическая сортировка (Надо у Юры список попросить))


CHR:2 2: 29224086 - 29225477
1 40M 80M 120M 160M 200M 242.1M
1.3Kbp
29,224,086 29,224,300 29,224,400 29,224,500 29,224,600 29,224,782 29,224,900 29,225,000 29,225,100 29,225,200 29,225,300 29,225,477
REFERENCE GRCh38
GENE genes38 Transcript View: Expanded Shorten introns
ALK
ALK-001
ALK-201
ALK-202
VCF sample1-lumpy
42,297,000
BAM sample1.bam Color mode: By Pair Orientation Group: Default Reads view: Expanded Sort General
900
450
CHR:2 2: 42296763 - 42297238
1 60M 120M 180M 242.1M
475bp
42,296,763 42,296,880 42,296,940 42,297,001 42,297,060 42,297,120 42,297,238
REFERENCE GRCh38
GENE genes38 Transcript View: Expanded Shorten introns
EML4
EML4-002
EML4-001
EML4-003
VCF sample1-lumpy
INV 29,224,782
BAM sample1.bam Color mode: By Insert Size Group: Default Reads view: Expanded Sort General
60
30
DATASETS
VARIANTS
Type Chr Gene Position Info
INV 2 ALK, EML4 29224782
INV 2 ALK, EML4 29224783
INV 13 RB1 48367316
DUP 17 NF1 31200515
DUP 6 51294521
DEL 6 51294989
BND 6 51295112
BND 6 51295400
DEL 6 FILIP1 75314196
DEL 6 GRIK2 101802142
BND 6 ROS1 117314770
BND 6 RP1-179... 117320174
DEL 6 RP1-179... 117320381
DEL 6 RP1-179... 117321573
DEL 6 RP1-179... 117323215
DEL 6 RP1-179... 117324016
DEL 6 RP1-179... 117329434
DEL 6 RP1-179... 117331379


CHR:2 2: 29224086 - 29225477
1 40M 80M 120M 160M 200M 242.1M
1.3Kbp
29,224,086 29,224,300 29,224,400 29,224,500 29,224,600 29,224,782 29,224,900 29,225,000 29,225,100 29,225,200 29,225,300 29,225,477
REFERENCE GRCh38
GENE genes38 Transcript View: Expanded Shorten introns
ALK
ALK-001
ALK-201
ALK-202
VCF sample1-lumpy
42,297,000
BAM sample1.bam Color mode: By Pair Orientation Group: Default Reads view: Expanded Sort General
900
450
CHR:2 2: 42296763 - 42297238
1 60M 120M 180M 242.1M
475bp
42,296,763 42,296,880 42,296,940 42,297,001 42,297,060 42,297,120 42,297,238
REFERENCE GRCh38
GENE genes38 Transcript View: Expanded Shorten introns
EML4
EML4-002
EML4-001
EML4-003
VCF sample1-lumpy
INV 29,224,782
BAM sample1.bam Color mode: By Insert Size Group: Default Reads view: Expanded Sort General
60
30
DATASETS
VARIANTS
Type Chr Gene Position Info
INV 2 ALK, EML4 29224782
INV 2 ALK, EML4 29224783
INV 13 RB1 48367316
DUP 17 NF1 31200515
DUP 6 51294521
DEL 6 51294569
BND 6 51295112
BND 6 51295400
DEL 6 FILIP1 75314196
DEL 6 GRIK2 101802142
BND 6 ROS1 117314770
BND 6 RP1-179... 117320174
DEL 6 RP1-179... 117320381
DEL 6 RP1-179... 117321573
DEL 6 RP1-179... 117323215
DEL 6 RP1-179... 117324016
DEL 6 RP1-179... 117329434
DEL 6 RP1-179... 117331379
Тут должна будет появиться новая колонка

ЗАДАЧИ

1. Ознакомиться с программным кодом NGB и его интерфейсом

2. Разработать функцию расчета потенциальной значимости скора вариантов

3. Протестировать разработанную функцию на реальных данных, найти оптимальный метод расчета скора

4. Разработать способ введения пользователем собственной формулы для рассчета скора

5. Внедрение этой функции в NGB

number of mutations throughout humanity per generation

Зачем это нужно?

$$\approx 10\text{–}100 \ \frac{\text{mutations}}{\text{generation}} \times 7 \times 10^9 \text{ people} \approx 10^{11}\text{–}10^{12} \text{ mutations per generation globally}$$

$$\approx 3 \times 10^9 \ \frac{\text{base-pairs}}{\text{human genome}} \Longrightarrow$$

tens to hundreds of people with *de novo* appearance of any specific mutation

$$\frac{(3\times10^9)^2 \text{ possible 2 bp combinations}}{10^{12} \text{ mutations/generation}} \Longrightarrow$$

$> 10^6$ generations for humanity to randomly reach a specific 2 base-pair mutation

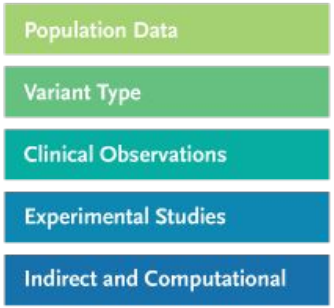
Что мог бы видеть врач без NGB

```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=FAO,Number=A,Type=Integer,Description="Flow Evaluator Alternate allele observation count">
##FORMAT=<ID=FDP,Number=1,Type=Integer,Description="Flow Evaluator Read Depth">
##FORMAT=<ID=FRO,Number=1,Type=Integer,Description="Flow Evaluator Reference allele observation count">
##FORMAT=<ID=FSAF,Number=A,Type=Integer,Description="Flow Evaluator Alternate allele observations on the forward strand">
##FORMAT=<ID=FSAR,Number=A,Type=Integer,Description="Flow Evaluator Alternate allele observations on the reverse strand">
##FORMAT=<ID=FSRF,Number=1,Type=Integer,Description="Flow Evaluator reference observations on the forward strand">
##FORMAT=<ID=FSRR,Number=1,Type=Integer,Description="Flow Evaluator reference observations on the reverse strand">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=SAF,Number=A,Type=Integer,Description="Alternate allele observations on the forward strand">
##FORMAT=<ID=SAR,Number=A,Type=Integer,Description="Alternate allele observations on the reverse strand">
##FORMAT=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##FORMAT=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  Sample1 Sample2 Sample3
chr1    2488153 .   A   G   4476.14 PASS    AC=4;AF=1.00;AN=4;DP=648;FDP=195;FR=.;FRO=2;FSAF=115;FSAR=78;FSRF=2;FSRR=0;FWDB=0.00167703;FXX=0.0101518;HRUN=
chr1    2491258 .   C   G   2611.42 PASS    AC=2;AF=0.500;AN=4;AO=146;DP=1332;FAO=146;FDP=334;FR=.;FRO=188;FSAF=87;FSAR=59;FSRF=109;FSRR=79;FWDB=-0.004949
chr1    6528100 .   GGCCCCT GGCCCTC 10278.10    PASS    AC=2;AF=1.00;AN=2;AO=90;DP=655;FAO=638;FDP=638;FR=.,HEALED,HEALED,HEALED,HEALED,HEALED;FRO=0;FSAF=
chr1    6528468 .   C   T   1859.16 PASS    AC=2;AF=0.500;AN=4;AO=120;DP=893;FAO=120;FDP=236;FR=.;FRO=116;FSAF=32;FSAR=88;FSRF=41;FSRR=75;FWDB=0.0384332;F
chr1    6529188 .   C   T   11263.97    PASS    AC=2;AF=0.500;AN=4;AO=606;DP=2960;FAO=640;FDP=946;FR=.,HEALED;FRO=306;FSAF=336;FSAR=304;FSRF=11;FSRR=295;F
chr1    6529443 .   A   G   5283.78 PASS    AC=2;AF=0.500;AN=4;AO=331;DP=2207;FAO=361;FDP=708;FR=.,HEALED;FRO=347;FSAF=187;FSAR=174;FSRF=196;FSRR=151;FWDB
chr1    6529747 .   A   AT  1631.35 PASS    AC=1;AF=0.500;AN=2;AO=10;DP=478;FAO=228;FDP=468;FR=.;FRO=240;FSAF=93;FSAR=135;FSRF=108;FSRR=132;FWDB=-0.009871
```

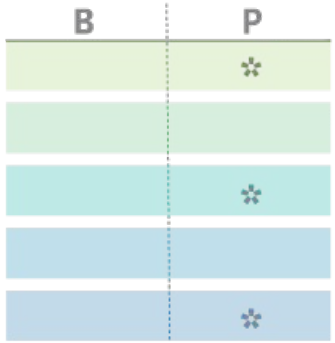
Как разработать скоры?

Sherloc is a transparent, accountable, and consistent variant interpretation process

#1 Five types of evidence, considered in a hierarchical approach:

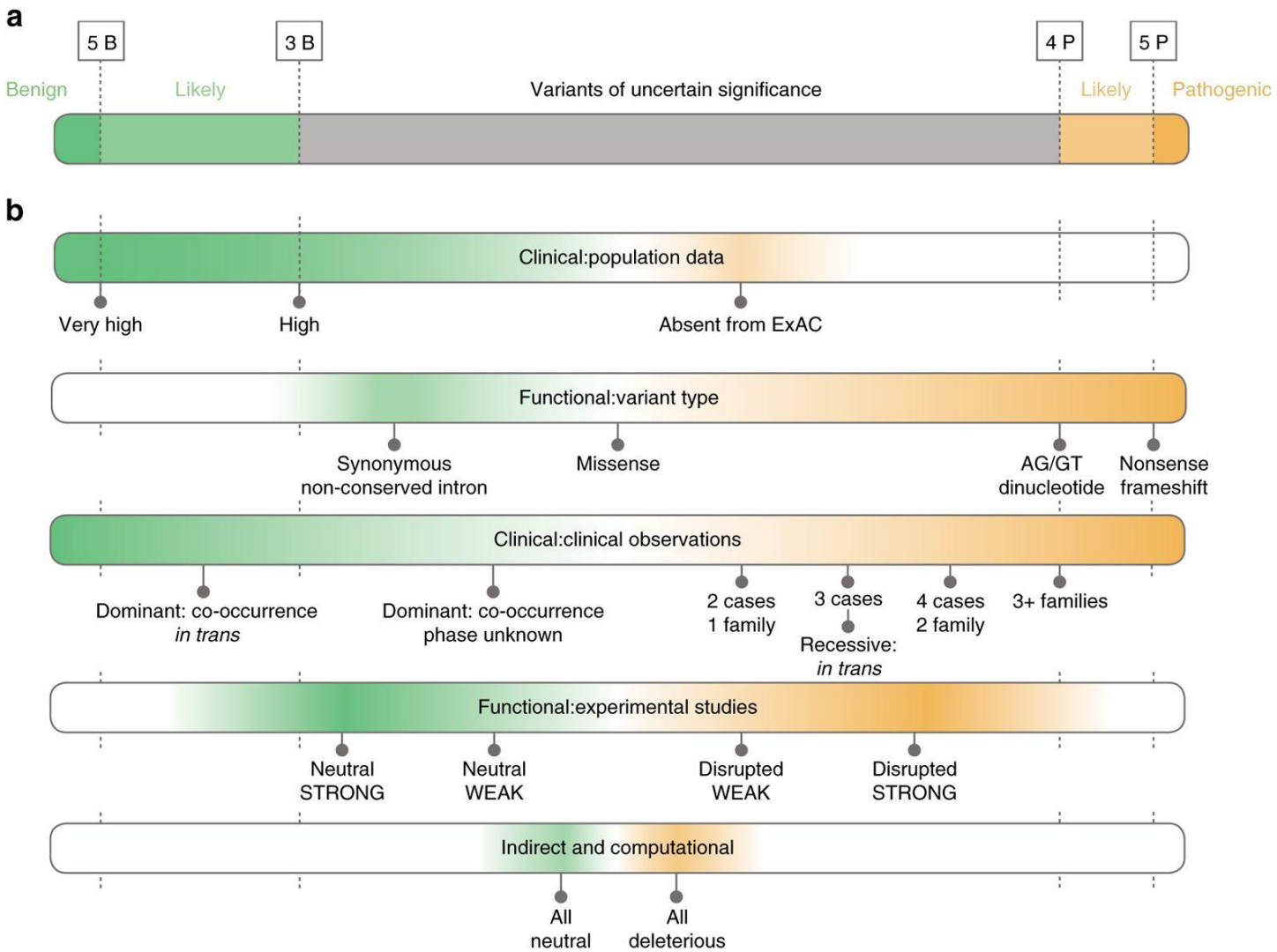


#2 Rule based scoring for each individual piece of evidence



#3 Point score thresholds to determine final classification based on the ACMG suggested five-tier classification system

https://www.invitae.com/en/variant-classification/


a
5 B
3 B
4 P
5 P
Benign
Likely
Variants of uncertain significance
Likely
Pathogenic
b
Clinical:population data
Very high
High
Absent from ExAC
Functional:variant type
Synonymous
non-conserved intron
Missense
AG/GT
dinucleotide
Nonsense
frameshift
Clinical:clinical observations
Dominant: co-occurrence
in trans
Dominant: co-occurrence
phase unknown
2 cases
1 family
3 cases
Recessive:
in trans
4 cases
2 family
3+ families
Functional:experimental studies
Neutral
STRONG
Neutral
WEAK
Disrupted
WEAK
Disrupted
STRONG
Indirect and computational
All
neutral
All
deleterious

Скоры, полученные по Шерлоку, планируются использовать как дефолтные.
Также очень важная часть проекта - разработка способа задания правил для вычисления скора пользователем самостоятельно на основе собственных параметров, таких как

- данные о популяции,
- функциональные исследования,
- анализ De novo мутаций,
- данные об аллелях,
- in silico предикторы
- И другие..

ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013

Sarah T. South, PhD[1,2], Charles Lee, PhD[3], Allen N. Lamb, PhD[1,2], Anne W. Higgins, PhD[4] and Hutton M. Kearney, PhD[5]; for the Working Group for the American College of Medical Genetics and Genomics (ACMG) Laboratory Quality Assurance Committee

Microarray methodologies, including array comparative genomic hybridization and single-nucleotide polymorphism–detecting arrays, are accepted as an appropriate first-tier test for the evaluation of imbalances associated with intellectual disability, autism, and multiple congenital anomalies. This technology also has applicability in prenatal specimens. To assist clinical laboratories in validation of microarray methodologies for constitutional applications, the American College of Medical Genetics and Genomics has produced the following revised professional standards and guidelines.

Genet Med advance online publication 26 September 2013

Key Words: constitutional; guidelines; microarray; postnatal; prenatal; standards

GENERAL CONSIDERATIONS

Purpose of cytogenomic microarrays

Constitutional cytogenetic abnormalities include aneuploidy (extra or missing chromosomes) and structural aberrations (chromosomal gains and losses, translocations, inversions, insertions, and marker chromosomes). The cytogenomic microarray (CMA) platforms discussed in this guideline are those designed for the detection of DNA copy number gains and losses associated with unbalanced chromosomal aberrations. Regions with an absence of heterozygosity (AOH), also referred to as loss of heterozygosity, regions/runs of homozygosity, or long continuous stretches of homozygosity, may also be detected by platforms with single-nucleotide polymorphism (SNP)-detecting probes. Some regions with AOH may be indicative of uniparental isodisomy or regions of the genome identical by descent.

The utility of this technology for detection of gains and losses in patients with intellectual disabilities, autism, and/or congenital anomalies has been well documented, and CMA is now recommended as a first-tier test for these indications.[1,2]

Advantages of CMAs

The benefits from the use of CMAs for detection of gains and losses of genomic DNA include:

1. Ability to analyze DNA from nearly any tissue, including archived tissue or tissue that cannot be cultured.
2. Detection of abnormalities that are cytogenetically cryptic by standard G-banded chromosome analysis.
3. Ability to customize the platform to concentrate probes in areas of interest.
4. Better definition and characterization of abnormalities detected by a standard chromosome study.
5. Interpretation of objective data, rather than a subjective visual assessment of band intensities.
6. Ability to detect copy neutral AOH with platforms incorporating SNP probes.
7. A ready interface of the data with genome browsers and databases.

[1]ARUP Laboratories, Salt Lake City, Utah, USA; [2]Department of Pathology, University of Utah, Salt Lake City, Utah, USA; [3]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; [4]Department of Pathology, UMass Memorial Medical Center and University of Massachusetts Medical School, Worcester, Massachusetts, USA; [5]Fullerton Genetics Center, Mission Health, Asheville, North Carolina, USA. Correspondence: Sarah T. South (Sarah.South@aruplab.com)

Submitted 17 July 2013; accepted 17 July 2013; advance online publication 26 September 2013. doi:10.1038/gim.2013.129

Что уже сделано?

1. Глобальная цель проекта - добавление колонки в NGB для приоритизации генетических вариантов, сужена(и в то же время расширена) до разработки java-утилиты, работающей с VCF-файлами, что может быть полезно не только в рамках NGB, но и как отдельной утилиты для других проектов.
2. Была выявлена необходимость написания собственного способа задания расчета коэффициента важности (В связи с анализом предыдущей попытки это сделать, в которой была закреплена формула для рассчета коэффициента важности мутации)
3. Были проанализированы литературные источники по уже имеющимся скорам для предсказания эффекта вариантов

Спасибо за внимание!