ГАЙДЛАЙНЫ ПО АННОТАЦИИ КОРПУСА GENEXOM
Версия 1.0 – 2025
https://github.com/Anara-Sultangaziyeva/GENEXOM

Аннотационные гайдлайны корпуса GENEXOM

Первый многоуровневый русскоязычный корпус клинических заключений по экзомному секвенированию

Версия 1.0 - 2025

1. Общая информация

Корпус GENEXOM содержит 300 реальных анонимизированных и 5000+ синтетических клинических заключений по полному экзомному секвенированию (WES) на русском языке.

Разметка проводится в Label Studio по единой схеме (16 типов сущностей + 7 типов отношений).

2. Принципы разметки

- Границы сущностей – точные (включая точки, стрелки, пробелы в HGVS)
- Пересекающиеся и вложенные сущности – допускаются
- Каждое вхождение одной и той же сущности размечается отдельно
- Нормализация (привязка к базам) – только в поле Notes

3. Типы сущностей (16)

Метка	Примеры из текста	Правила разметки
GENE	BRCA1, SCN1A, FGFR3, COL5A1	Только HGNC-символы. Не размечаются описания («ген, отвечающий за…»)
CDNA_PROT	c.1138G>A, p.Gly380Arg, c.5266dupC	Все HGVS-нотации (c. и p.). Обязательно с точкой и без лишних пробелов
VARIANT_LOC	chr4:1804392G>A, chr9:134731605G>A	Геномные координаты в формате chrX:позицияA>B
DISEASE	ахондроплазия, синдром Ретта, синдром Дауна	Полные названия заболеваний и синдромов (русский и латинский)
SIGNIFICANCE	Патогенный, Вероятно патогенный, VUS	Классификация ACMG/AMP
ZYGOSITY	Гетерозиготный, Гомозиготный, Гемизиготный	Состояние аллелей
INHERITANCE_MODE	Аутосомно-доминантный, Аутосомно-рецессивный, XLR	Тип наследования
OMIM_ID	100800, 312750, 614080	Номера из базы OMIM (6 цифр)
CLINVAR_ID	RCV000255123	Идентификаторы ClinVar
PHENOTYPE	микроцефалия, брахидактилия, гипотония мышц	Клинические проявления и стигмы дизэмбриогенеза
RECOMMENDATION	Рекомендовано наблюдение эпилептолога	Врачебные рекомендации
EXON_NUMBER	экзон 7, 7–8 экзоны	Номера экзонов
DBSNP_ID	rs79912345	Идентификаторы dbSNP
THERAPY_RECOMMENDATION	Бринейра (Cerliponase alfa), Вимизин	Конкретные препараты и методы лечения
RETEST_PLAN	Рекомендовано повторное тестирование через год	План дальнейшего обследования
DIAGNOSTIC_METHOD	Полное секвенирование экзома, Панель «Эпилепсии»	Метод исследования

4. Типы отношений (7)

Отношение	Пример связи	Обязательно?
variant_in_gene	c.1138G>A → FGFR3	Да
gene_associated_with	FGFR3 → ахондроплазия	Да
variant_significance	c.1138G>A → Патогенный	Да
disease_inheritance_mode	ахондроплазия → Аутосомно-доминантный	Да
phenotype_supports_disease	брахидактилия → ахондроплазия	Да
disease_omim_link	ахондроплазия → 100800	Да
variant_zygosity	c.1138G>A → Гетерозиготный	Да

5. Пример полной разметки

Исходный текст:

> Выявлена гетерозиготная мутация c.1138G>A (p.Gly380Arg) в гене FGFR3 (chr4:1804392G>A), патогенная, аутосомно-доминантная. Диагноз: ахондроплазия (OMIM #100800).

Размеченные сущности:

- GENE → FGFR3
- CDNA_PROT → c.1138G>A, p.Gly380Arg
- VARIANT_LOC → chr4:1804392G>A
- DISEASE → ахондроплазия
- SIGNIFICANCE → Патогенный
- ZYGOSITY → Гетерозиготный
- INHERITANCE_MODE → Аутосомно-доминантный
- OMIM_ID → 100800

Отношения:

- variant_in_gene → c.1138G>A → FGFR3
- variant_significance → c.1138G>A → Патогенный
- gene_associated_with → FGFR3 → ахондроплазия
- disease_inheritance_mode → ахондроплазия → Аутосомно-доминантный
- disease_omim_link → ахондроплазия → 100800

6. Особые случаи

Ситуация	Решение
«Релевантных вариантов не обнаружено»	Ничего не размечаем
«Гомозиготная делеция 7 и 8 экзонов»	CDNA_PROT → 7 и 8 экзоны; ZYGOSITY → Гомозиготный
«de novo»	Пишем в Notes, отдельную метку не создаём
Сокращения AD/AR/XLR	Расшифровываем в INHERITANCE_MODE
Несколько одинаковых вариантов в тексте	Размечаем каждое вхождение отдельно

7. Контроль качества

- Целевой F1-IAA по сущностям: ≥ 0.83 (уже достигнут на пилотных 30 отчётах)
- Полное совпадение по ключевым отношениям: variant_in_gene, variant_significance