Hochladen von VCF-Dateien in die Datenbank HerediCaRe

Dieses Programm liest genetische Varianten aus einer im HerediCaRe-Webinterface [1] hochgeladenen VCF-Datei ein, filtert diese Varianten bezüglich betroffener Gene und trägt sie automatisiert inklusive zusätzlicher Informationen in die HerediCaRe-Datenbank ein. Alle Varianten werden zunächst normalisiert [2], und entsprechend des angegebenen ursprünglichen Referenzgenoms entweder von hg19 nach hg38 oder von hg38 nach hg19 transformiert. Anschließend erfolgt eine Annotation der hg38-basierten Varianten mit SnpEff [3], aus der sich betroffene Gene, HGVS-Annotation und Effekte für den Datenbank-Eintrag ergeben. Aus der VCF-Datei können zudem Genotyp (Zygotie) und Pathogenitätsklasse übernommen werden, sofern angegeben.

VCF (Variant Call Format) stellt das Standard-Dateiformat zur Spezifikation genomischer Varianten dar. Es handelt sich um eine Text-Datei, die in einem beliebigen Text-Editor betrachtet und bearbeitet werden kann[1]. Aufgrund der weiten Verbreitung dieses Dateiformats ist davon auszugehen, dass grundsätzlich jede Software, die der Ausgabe genomischer Varianten dient, den Export dieser Varianten als VCF-Datei unterstützt.

Die neueste VCF-Spezifikation (Version 4.3) ist vollumfänglich unter dem folgenden Link (auf Englisch) definiert: https://samtools.github.io/hts-specs/VCFv4.3.pdf. Eine VCF-Datei, die den Spezifikationen ab **Version 4** (https://samtools.github.io/hts-specs/VCFv4.1.pdf) entspricht, ist für die Anwendung jedoch ausreichend.

Eine VCF-Datei unterteilt sich in den sog. Header und eine Auflistung von genomischen Varianten.

- Textzeilen, die zum Header gehören, beginnen mit einer Raute (#).
- Zeilen, die der Auflistung von Varianten dienen, beginnen ohne Raute und bestehen aus mindestens 8 Einträgen (im Folgenden Spalten genannt), die jeweils durch einen Tabstopp voneinander separiert sind.
  Hinweis: Es ist (entsprechend der VCF-Formatspezifikation) zu beachten, dass die Einträge in den Variantenzeilen nicht mit Sonderzeichen (z.B. Anführungszeichen) umgeben werden dürfen.

Es wird davon ausgegangen, dass automatisch erzeugte (und somit gültige) VCF-Dateien für das Hochladen der Varianten in HerediCaRe verwendet werden, daher wird hier nicht weiter auf die Format-Spezifikation eingegangen. Zu beachten ist jedoch, dass jede zu prozessierende VCF-Datei immer nur die Varianten einer Probe (MEMBER-ID) enthalten darf, obwohl grundsätzlich die Möglichkeit besteht, die Varianten mehrerer Proben in einer VCF-Datei darzustellen. Der Name einer hochgeladenen Datei muss die Endung .vcf besitzen. Grundsätzlich werden nur Varianten aus dieser Datei in die Datenbank übernommen, die lt. SnpEff-Annotation innerhalb oder in Nähe der in Tabelle 1 aufgeführten Transkripte lokalisiert sind.

---

1. Aufgrund der Dateinamen-Endung ist das Öffnen der Datei unter Windows jedoch meist nicht per Doppelklick möglich. Hier empfiehlt es sich, zuerst den Editor (Bsp. Notepad) zu öffnen und dann aus dem Editor heraus die entsprechende Datei zu öffnen

16.05.2023

Die VUS-Task Force und die AG Bioinformatik empfehlen seit September 2022 die einheitliche Nutzung von MANE-Transkripten (Matched Annotation from NCBI and EMBL-EBI [4]) innerhalb des Konsortiums. Daher beziehen sich die automatisch erzeugten HGVS-Annotationen ausschließlich auf die in Tabelle 1 genannten (MANE) Transkript-IDs. Die Spezifikation alternativer Transkript-IDs ist im Gegensatz zu älteren Versionen des VCF-Parsers nicht mehr möglich.

Grundsätzlich ist jedem Protein-kodierenden Gen ein **MANE Select** Transkript als Standard zugeordnet. Existieren jedoch klinisch relevante Einträge in ClinVar, die durch dieses Transkript nicht darstellbar sind, sind zusätzliche **MANE Plus Clinical** Transkripte definiert. In Tabelle 1 sind MANE Select Transkripte in der dritten Spalte (X) mit „1“ markiert, MANE Plus Clinical Transkripte mit „0“ (vgl. z.B. für *SMARCA4*). Sollte eine Variante mit zwei Transkripten (Select & Plus Clinical) eines Gens annotiert werden können, wird die Annotation bezüglich des MANE Select Transkripts präferiert und bezüglich MANE Plus Clinical ignoriert.

Bezüglich der automatisch erzeugten Annotation der von den Varianten betroffenen Gene gelten folgende Einschränkungen:

1- Die Gene ***RMRP***, ***H19***, ***KCNQ1OT1***, ***ABO*** und ***TERC*** werden ignoriert, weil keine zugehörigen MANE-Trankskripte definiert sind.
2- Die Gene ***FAM175A***, ***MRE11A*** und ***PARK2*** werden durch die automatisierte SnpEff-Annotation mit ***ABRAXAS1***, ***MRE11*** und ***PRKN*** benannt.

Es besteht für Nutzerinnen und Nutzer die Möglichkeit, zusätzliche Informationen zur Klassifizierung der Pathogenität von Varianten in der INFO-Spalte (Spalte 8) mithilfe der Schlagworte **MutDB:Classification, CLASS** oder **MT** zu hinterlegen Sind mehrere dieser Einträge für die selbe Variante vorhanden, wird der MUtDB:Classification-Eintrag vor allen anderen und der CLASS-Eintrag vor dem MT-Eintrag priorisiert. Tabelle 2 fasst alle gültigen Spezifizierungen der Pathogenitäts-Klasse durch die Schlagworte zusammen. Alle anderen als in der Tabelle angegebenen Belegungen von MutDB:Classification, CLASS und MT führen zu einem Ausschluss der betreffenden Variante vom HerediCaRe-Eintrag.

In dem folgenden Beispiel würde also die Pathogenitätsklasse 1 übernommen werden.

8 90983436 . T G 28 . MT=LBE;CLASS=1

Eine Erweiterung der in Tabelle 1 aufgeführten Gene und Transkripte ist nach Rücksprache möglich. Grundsätzlich können jedoch keine mitochondrialen Varianten prozessiert werden. Zu beachten ist außerdem, dass Chromosomenbezeichnungen ausschließlich ohne den Zusatz „chr“ in die HerediCaRe-Datenbank übernommen werden. Aus „chr22“ wird also nur „22“.

Mit der neuen, aktualisierten Version des VCF-Parsers werden nun die Koordinaten (genomische Position, Referenzallel, Alternativallel) bezüglich beider Referenzgenome hg19 und hg38 in die Datenbank übernommen. Zudem können die beobachteten Genotypen 0/0 (Wildtyp), 0/1 (heterozygot) und 1/1 (homozygot) hinterlegt werden, falls entsprechend dieser Vorgaben im VCF-Format mit dem GT-Tag spezifiziert. Andernfalls erfolgt der Eintrag ### (keine Angabe) in die Datenbank.

Bei Fragen bezüglich des VCF-Dateiformats und zur Erzeugung entsprechender Dateien wenden Sie sich bitte an:

Rudel Christian Nkouamedjo Fankep
Zentrum Familiärer Brust- und Eierstockkrebs am Universitätsklinikum Köln (AöR)
Telefon: +49 221 478-39325
E-Mail: rudel.nkouamedjo-fankep@uk-koeln.de

16.05.2023

Tabelle 1: Für den Eintrag in die HerediCaRe-Datenbank relevante Gene mit zugehörigen MANE-Transkripten (TranskriptID). MANE Select Transkripte (priorisiert) sind durch eine 1 in Spalte X gekennzeichnet, MANE Plus Clinical Transkripte durch Eintrag 0.

Gene	TranskriptID	X	Gene	TranskriptID	X	Gene	TranskriptID	X
ABRAXAS1	NM_139076	1	GEN1	NM_001130009	1	RBBP8	NM_002894	1
ACD	NM_001082486	1	GJB2	NM_004004	1	RECQL	NM_002907	1
ACTA2	NM_001613	1	GLA	NM_000169	1	RECQL4	NM_004260	1
ACTC1	NM_005159	1	GNA11	NM_002067	1	RELN	NM_005045	1
ACVRL1	NM_000020	1	GNAQ	NM_002072	1	REST	NM_005612	1
AEBP1	NM_001129	1	GPC3	NM_004484	1	RET	NM_020975	1
AIP	NM_003977	1	GPRC5A	NM_003979	1	RHBDF2	NM_001005498	1
AKT1	NM_001382430	1	GREM1	NM_013372	1	RINT1	NM_021930	1
ALK	NM_004304	1	HAVCR2	NM_032782	1	RIT1	NM_006912	1
ANKRD26	NM_014915	1	HAX1	NM_006118	1	RNASEL	NM_021133	1
APC	NM_000038	1	HFE	NM_000410	1	RNF43	NM_017763	1
APOB	NM_000384	1	HMBS	NM_000190	1	RPE65	NM_000329	1
ARID1A	NM_006015	1	HNF1A	NM_000545	1	RPL11	NM_000975	1
ARID1B	NM_001374828	1	HOXB13	NM_006361	1	RPL15	NM_002948	1
ATM	NM_000051	1	HPS1	NM_000195	1	RPL23	NM_000978	1
ATP7B	NM_000053	1	HRAS	NM_176795	0	RPL26	NM_000987	1
ATR	NM_001184	1	HRAS	NM_005343	1	RPL27	NM_000988	1
ATRX	NM_000489	1	IDH1	NM_005896	1	RPL31	NM_000993	1
AXIN2	NM_004655	1	IDH2	NM_002168	1	RPL35A	NM_000996	1
BAP1	NM_004656	1	IKZF1	NM_006060	1	RPL36	NM_033643	1
BARD1	NM_000465	1	IPMK	NM_152230	1	RPL5	NM_000969	1
BCL11A	NM_022893	1	ITK	NM_005546	1	RPS10	NM_001014	1
BLM	NM_000057	1	JAK2	NM_004972	1	RPS15	NM_001018	1
BMPR1A	NM_004329	1	KCNH2	NM_000238	1	RPS17	NM_001021	1
BRAF	NM_001374258	0	KCNQ1	NM_000218	1	RPS19	NM_001022	1
BRAF	NM_004333	1	KIF1B	NM_001365951	1	RPS20	NM_001023	1
BRCA1	NM_007294	1	KIT	NM_000222	1	RPS24	NM_033022	1
BRCA2	NM_000059	1	KMT2C	NM_170606	1	RPS26	NM_001029	1
BRIP1	NM_032043	1	KMT2D	NM_003482	1	RPS27	NM_001030	1
BTD	NM_001370658	1	KRAS	NM_004985	1	RPS27A	NM_002954	1
BUB1B	NM_001211	1	LDLR	NM_000527	1	RPS28	NM_001031	1
CACNA1S	NM_000069	1	LIG4	NM_206937	1	RPS29	NM_001032	1
CASQ2	NM_001232	1	LMNA	NM_005572	0	RPS7	NM_001011	1
CASR	NM_000388	1	LMNA	NM_170707	1	RTEL1	NM_001283009	1
CBL	NM_005188	1	LZTR1	NM_006767	1	RUNX1	NM_001754	1
CDC73	NM_024529	1	MAD2L2	NM_006341	1	RYR1	NM_000540	1
CDH1	NM_004360	1	MAP2K1	NM_002755	1	RYR2	NM_001035	1
CDK12	NM_016507	1	MAP2K2	NM_030662	1	SAMD9L	NM_152703	1
CDK4	NM_000075	1	MAP3K1	NM_005921	1	SBDS	NM_016038	1
CDKN1B	NM_004064	1	MAX	NM_002382	1	SCN5A	NM_001099404	0
CDKN1C	NM_001122630	1	MC1R	NM_002386	1	SCN5A	NM_000335	1
CDKN2A	NM_058195	0	MDH2	NM_005918	1	SDHA	NM_004168	1
CDKN2A	NM_000077	1	MEN1	NM_001370259	1	SDHAF2	NM_017841	1

CDKN2B	NM_004936	1	MET	NM_000245	1	SDHB	NM_003000	1
CEBPA	NM_004364	1	MITF	NM_000248	0	SDHC	NM_003001	1
CEP57	NM_014679	1	MITF	NM_001354604	1	SDHD	NM_003002	1
CFTR	NM_000492	1	MLH1	NM_000249	1	SEC23B	NM_006363	1
CHEK1	NM_001114122	1	MLH3	NM_001040108	1	SETBP1	NM_015559	1
CHEK2	NM_007194	1	MMS19	NM_022362	1	SETD6	NM_001160305	1
COL3A1	NM_000090	1	MRE11	NM_005591	1	SH2B3	NM_005475	1
COL7A1	NM_000094	1	MSH2	NM_000251	1	SH2D1A	NM_002351	1
CREBBP	NM_004380	1	MSH3	NM_002439	1	SHOC2	NM_007373	1
CSF3R	NM_000760	1	MSH6	NM_000179	1	SLC25A11	NM_003562	1
CTC1	NM_025099	1	MSR1	NM_138715	1	SLC5A5	NM_000453	1
CTNNA1	NM_001903	1	MTAP	NM_002451	1	SLX4	NM_032444	1
CTNNB1	NM_001904	1	MTOR	NM_004958	1	SMAD3	NM_005902	1
CTR9	NM_014633	1	MUTYH	NM_001048174	1	SMAD4	NM_005359	1
CTRC	NM_007272	1	MYBPC3	NM_000256	1	SMARCA4	NM_001387283	0
CYLD	NM_001378743	1	MYCT1	NM_025107	1	SMARCA4	NM_003072	1
DDB2	NM_000107	1	MYH11	NM_001040113	0	SMARCB1	NM_003073	1
DDX41	NM_016222	1	MYH11	NM_002474	1	SMARCE1	NM_003079	1
DICER1	NM_177438	1	MYH7	NM_000257	1	SOS1	NM_005633	1
DIS3L2	NM_152383	1	MYL2	NM_000432	1	SOS2	NM_006939	1
DKC1	NM_001363	1	MYL3	NM_000258	1	SPINK1	NM_001379610	1
DNAJC21	NM_001012339	1	NBN	NM_002485	1	SPRED1	NM_152594	1
DNMT3A	NM_022552	1	NF1	NM_001042492	1	SQSTM1	NM_003900	1
DNMT3B	NM_006892	1	NF2	NM_000268	1	SRGAP1	NM_020762	1
DOCK8	NM_203447	1	NHP2	NM_017838	1	STAT1	NM_007315	1
DSC2	NM_024422	1	NOP10	NM_018648	1	STAT3	NM_139276	1
DSG2	NM_001943	1	NRAS	NM_002524	1	STK11	NM_000455	1
DSP	NM_004415	1	NSD1	NM_022455	1	STN1	NM_024928	1
DST	NM_005432	1	NTHL1	NM_002528	1	STX11	NM_003764	1
EDC3	NM_025083	1	OTC	NM_000531	1	STXBP2	NM_006949	1
EDC4	NM_014329	1	PALB2	NM_024675	1	SUFU	NM_016169	1
EGFR	NM_005228	1	PALLD	NM_001166108	1	TBXT	NM_001366285	1
ELANE	NM_001972	1	PARN	NM_002582	1	TERF2IP	NM_018975	1
EMSY	NM_001300942	1	PAX5	NM_016734	1	TERT	NM_198253	1
ENG	NM_001114753	1	PCSK9	NM_174936	1	TGFBR1	NM_004612	1
EPCAM	NM_002354	1	PDGFRA	NM_006206	1	TGFBR2	NM_003242	1
ERBB2	NM_004448	1	PDGFRB	NM_002609	1	TINF2	NM_001099274	1
ERCC1	NM_001983	1	PHOX2B	NM_003924	1	TMEM127	NM_017849	1
ERCC2	NM_000400	1	PIK3C2G	NM_001288772	1	TMEM43	NM_024334	1
ERCC3	NM_000122	1	PIK3CA	NM_006218	1	TNFRSF11A	NM_003839	1
ERCC4	NM_005236	1	PIK3R1	NM_181523	1	TNIP1	NM_006058	1
ERCC5	NM_000123	1	PIK3R2	NM_005027	1	TNNI3	NM_000363	1
ERCC6L2	NM_020207	1	PKP2	NM_001005242	1	TNNT2	NM_001276345	1
ETV6	NM_001987	1	PMS1	NM_000534	1	TP53	NM_000546	1
EXT1	NM_000127	1	PMS2	NM_000535	1	TP53BP1	NM_001141980	1
EXT2	NM_207122	1	POLD1	NM_002691	1	TPM1	NM_001018005	1
EZH2	NM_004456	1	POLE	NM_006231	1	TRDN	NM_006073	1
FAN1	NM_014967	1	POLH	NM_006502	1	TRIM28	NM_005762	1

FANCA	NM_000135	1	POT1	NM_015450	1	TRIM37	NM_015294	1
FANCB	NM_001018113	1	PPM1D	NM_003620	1	TRIP13	NM_004237	1
FANCC	NM_000136	1	PPP1CB	NM_002709	1	TSC1	NM_000368	1
FANCD2	NM_001018115	1	PRF1	NM_001083116	1	TSC2	NM_000548	1
FANCE	NM_021922	1	PRKAG2	NM_016203	1	TSR2	NM_058163	1
FANCF	NM_022725	1	PRKAR1A	NM_002734	1	TTN	NM_133379	0
FANCG	NM_004629	1	PRKN	NM_004562	1	TTN	NM_001267550	1
FANCI	NM_001113378	1	PRKN	NM_004562	1	UBE2T	NM_014176	1
FANCL	NM_018062	1	PRSS1	NM_002769	1	UNC13D	NM_199242	1
FANCM	NM_020937	1	PTCH1	NM_001083603	0	UVSSA	NM_020894	1
FAS	NM_000043	1	PTCH1	NM_000264	1	VHL	NM_000551	1
FAT4	NM_001291303	1	PTCH2	NM_003738	1	WAS	NM_000377	1
FBN1	NM_000138	1	PTEN	NM_000314	1	WRAP53	NM_001143992	1
FGFR1	NM_023110	1	PTPN11	NM_002834	1	WRN	NM_000553	1
FH	NM_000143	1	PYROXD1	NM_024854	1	WT1	NM_024426	1
FLCN	NM_144997	1	RABL3	NM_173825	1	XPA	NM_000380	1
FLNC	NM_001458	1	RAD50	NM_005732	1	XPC	NM_004628	1
FOCAD	NM_001375567	1	RAD51	NM_002875	1	XRCC1	NM_006297	1
FOXE1	NM_004473	1	RAD51B	NM_133510	1	XRCC2	NM_005431	1
GAA	NM_000152	1	RAD51C	NM_058216	1	XRCC3	NM_001723	0
GALNT12	NM_024642	1	RAD51D	NM_002878	1	XRCC3	NM_001374736	1
GATA1	NM_002049	1	RAF1	NM_002880	1	YAP1	NM_001130145	1
GATA2	NM_032638	1	RB1	NM_000321	1	ZNF276	NM_001113525	1
GBA	NM_000157	1						

Tabelle 2: Mögliche Schlagwort (Tag)-Belegungen zur Spezifizierung von Pathogenitätsklassen.

Pathogenitätsklasse (HerediCaRe)	**MutDB:Classification**	**CLASS**	**MT**
1	benign	1	BEN
2	likely benign	2	LBE
3	uncertain significance	3	UI
4	likely pathogenic	4	LPAT
5	pathogenic	5	PAT
"" (keine Angabe)	undefined	""	UD

Referenzen

[1] Engel, Christoph, et al. (2022) HerediCaRe: Dokumentations-und IT-Lösung eines spezialisierten Registers für erblichen Brust-und Eierstockkrebs. *Senologie-Zeitschrift für Mammadiagnostik und -therapie* 19(04): 319-326.

[2] Tan, Adrian, Gonçalo R. Abecasis, and Hyun Min Kang. (2015) Unified representation of genetic variants. *Bioinformatics* 31(13): 2202-2204.

[3] Cingolani, Pablo, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3."*fly* 6(2): 80-92.

[4] Morales, Joannella, et al. (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 604(7905): 310-315.