bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genomics of Human Respiratory Syncytial Virus Vaccine Attenuation

Thomas Junier[1,*], Laurent Kaiser[2], Nimisha Chaturvedi[3], and Jacques Fellay[3,4,1]

[1]Swiss Institute of Bioinformatics, Vital-IT Group, Université de Lausanne, Bâtiment Génopode, 1015 Lausanne, Switzerland
[2]Hôpitaux Universitaires de Genève, Service des Maladies Infectieuses, Rue Gabrielle-Perret-Gentil 4, 1205 Genève, Switzerland
[3]Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[4]Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, 1011 Lausanne, Switzerland
[*]thomas.junier@sib.swiss

ABSTRACT

The human respiratory syncytial virus (HRSV) is a major cause of lower respiratory tract infection in children worldwide. Despite decades of efforts, no vaccine is available. In this work, we report mutations that are frequent in vaccine candidates and rare in wild-type genomes, taking into account all the publicly available HRSV sequence data. These mutations are different from the ones already known to attenuate the virus, and thus may contribute to the effort towards producing a live attenuated vaccine against HRSV.

Introduction

The human respiratory syncytial virus (HRSV) is a single-strand, negative-sense RNA virus belonging to family Paramyxoviridae. It has a 15-kb genome that contains ten genes (Table 1).

HRSV infects the respiratory tract, and causes a variety of symptoms ranging from mild to fatal. It is extremely common, almost everyone being infected for the first time by age two. It is a major cause of acute lower respiratory tract infection in children worldwide, which is in turn a significant cause of child mortality, with roughly a hundred thousand children dying from HRSV-related infection in 2005.[1]

HRSV is divided into two subtypes, A and B. Subtype A appears to be more frequent in infected patients.[2] Re-infections can occur at any age, but symptoms are most severe in children under one year old, the elderly, and immunocompromised patients.[3] While prophylaxis (a monoclonal antibody marketed as Palivizumab®) and treatment (the guanosine analog Ribavirin®) against HRSV exist, no vaccine is currently available. Palivizumab and Ribavirin are expensive (the latter is also toxic[4]) and only recommended for high-risk patients[5], which makes the development of a safe and effective vaccine a high priority.

In the 1960s, an early attempt at a HRSV vaccine using formalin-inactivated virus ended in failure: the vaccine did not protect against infection and in some cases actually caused enhanced disease (including fatalities).[6] Inactivated vaccines against HRSV are no longer pursued since this fiasco. Modern approaches include particulate or subunit vaccines, gene-based vaccines, and live attenuated vaccines (LAVs). The latter are considered the most effective: although they do pose some risks, especially in immunocompromised patients and children[7], they also offer important advantages, including the fact that they do not cause enhanced disease, that they stimulate immunity both systemically and in the respiratory tract, that they can be delivered intranasally, and that they replicate in young infants even in the presence of maternal HRSV antibody.[8] LAVs exist against measles, yellow fever, polio and smallpox, among others; they were instrumental in the eradication or near-eradication of the last two. The agent of measles, measles morbilivirus, belongs to the same family as HRSV and has a similar genomic structure and replication mechanism, demonstrating that LAVs are feasible within Paramyxoviridae.

Live attenuated vaccines are pathogens that are able to replicate in the patient, but are less virulent ("attenuated") than the wild type. One method for achieving attenuation involves several generations of culture in an environment different from the original host cell (for example, a different kind of cell or a lower temperature ("cold passage")). As the virus adapts to its new environment, it may become less adapted to the original one – thus losing virulence – while still being able to elicit a strong immune response from the host. Another approach involves chemical mutagenesis followed by selection for temperature-sensitive mutants.

Attenuation is thus the result of mutations in the viral genome. Mutations in different genes affecting different aspects of the viral cycle may equally lead to attenuation: for example, impairing replication and compromising virion stability both lead

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to attenuation in Poliovirus.[9].

Understanding these mutations is useful for at least the following reasons: first, attenuation achieved by a smaller number of mutations has a higher risk of reversion to virulence than attenuation achieved by a larger number; second, a knowledge of the phenotypic effect of mutations enables a reverse genetics approach to LAV production, in which desired combinations of attenuating mutations are deliberately introduced into the viral genome; finally, the effect of mutations is generally not additive: some mutations may in fact even cancel one another – for example, one mutation in the L gene is known to compensate another mutation in the same gene[8] – hence the importance of studying the effects of combinations of mutations.

Attenuating mutations have been identified by Karron *et al.* through whole-genome sequencing of a cold-passaged strain as well as six temperature-sensitive derivatives of that strain.[8]. In this work, we followed a comparable approach, but using all publicly available HRSV sequence data: the public sequence databases contain thousands of HRSV sequences, both wild-type and attenuated, which we aligned together then scanned for positions in which the two groups have different allele frequencies.

Table 1. Genes and gene products of the HRSV genome[6], as well as some of their known functions, by order of position along the (negative-sense) genome. Gene products M2-1 and M2-2 are encoded by different open reading frames on the same mRNA.

Symbol	Name	Function
NS1	non-structural protein 1	type-I interferon inhibitor
NS2	non-structural protein 2	type-I interferon inhibitor
N	nucleoprotein	encapsidation, replication, transcription
P	phosphoprotein	encapsidation, replication, transcription
M	matrix protein	virion assembly
SH	small hydrophobic protein	ion transport
G	attachment glycoprotein	attachment
F	fusion glycoprotein	fusion, attachment
M2	M2-1	transcription
M2	M2-2	transcription, replication
L	Polymerase	replication, transcription

Results

Breakdown of Wild-type Sequences by Subtype and Gene

Out of the 28,248 GenBank entries originally downloaded, 11,600 were rejected due to being of unknown or explicitly non-HRSV origin, leaving 16,648. Among these, inference of category (vaccine vs. wild-type) recognized 575 vaccine and 16,073 wild-type entries. Filtering the latter by quality, followed by inference of subtype and gene and finally by extraction of coding sequences (CDSs) yielded 10,975 wild-type sequences, of which a breakdown is shown in table 2, column 3.

Table 2. Breakdown of wild-type and vaccine sequences by HRSV subtype and gene.

Subtype	Gene	# WT	# Vaccine	# Full-length Vaccine	# Clade Representatives
A	F	864	95	82	7
A	G	1538	128	63	5
A	L	654	41	33	6
A	M	686	36	34	6
A	M2	682	33	31	3
A	N	679	44	38	6
A	NS1	731	33	33	5
A	NS1	731	33	33	5
A	P	688	31	29	4
A	SH	768	26	26	4
B	F	314	32	32	3
B	G	703	86	38	3
B	L	213	34	30	4
B	M	216	34	30	3
B	M2	215	30	30	3

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Breakdown of wild-type and vaccine sequences by HRSV subtype and gene.

Subtype	Gene	# WT	# Vaccine	# Full-length Vaccine	# Clade Representatives
B	N	218	45	34	5
B	NS1	271	30	30	3
B	NS1	271	30	30	3
B	P	219	34	30	3
B	SH	295	29	29	4

Classification of Vaccine Sequences by Subtype and Gene

Column 4 of table 2 shows a breakdown of vaccine sequences (or sub-sequences, in the case of sequences matching more than one gene) by subtype and gene. Column 5 shows the number of sequences in each category after filtering by length, retaining only full-length or near-full-length sequences.

Gene Phylogenies

Eevery gene except F shows a clear phylogenetic separation of the two subtypes (Figure 1, see supplementary data for all trees in Newick format and as graphical representations). It also shows that vaccines of either subtype do not form a single clade, but are instead spread out over the whole respective A or B subtree. Finally, each tree shows vaccine sequences arranged in a distinctive, stair-like pattern consisting of wholly unbalanced nodes, that is, nodes of which one child is a leaf and all other descendants descend from the other child, which is itself wholly unbalanced (Figure 2).

Clade Representatives

The selection of one vaccine per unbalanced clade further reduced the number of vaccine sequences (Table 2, column 6). Pooled with the corresponding wild-type sequences, they yielded twenty sets of sequences, one each per (gene, subtype) pair, consisting only of full-length or nearly full-length sequences, and with a minimal number of close relatives among the vaccines.

Vaccine-linked Mutations

Table 4 shows all vaccine-linked, non-silent mutations for which the Fisher exact test's p-value is not greater than the significance threshold $\alpha = 0.01$, after applying the Bonferroni correction for 6573 simultaneous comparisons (one per non-conserved position in the genome) - this results in a corrected $\alpha = 1.52 \times 10^{-6}$. No position matching these criteria was found in subtype B. Column 12 of the table shows the effect size as the odds ratio of a vaccine *vs.* a wild-type sequence having the minor allele at a given position, flanked by the lower bound of the 95% confidence interval around the ratio (column 12) and the upper bound (column 13). None of the intervals includes the value 1, indeed the minimal lower bound is 12.

Number of mutations in vaccine versus wild-type sequences

Vaccine sequences were heavily mutated at the vaccine-linked positions (Table 3). Indeed, in the attachment glycoprotein and nucleoprotein, *every* vaccine sequence differs from the wild type at *all* positions, and even in the fusion glycoprotein and polymerase the average number of mutated positions per sequence (4.14 and 12.33) are close to the maxima (5 and 13), respectively.

Table 3. Number of mutations in vaccine vs. wild-type sequences (subtype A).

Gene	# mutated positions	# WT sequences	# WT mutations	avg # mutations / WT sequence	# vaccine sequences	# vaccine mutations	avg # mutations / vaccine sequences	avg mutation ratio
F	5	864	188	0.22	7	29	4.14	19.04
G	11	1538	1488	0.97	5	55	11.00	11.37
L	13	654	753	1.15	6	74	12.33	10.71
N	3	679	133	0.20	6	18	3.00	15.32

Discussion

Power Considerations The fact that very few vaccine sequences are retained raises the question of the power of our statistical procedure. However, estimates (using R's `power.fisher.test` function, from package `statmod`) show that even with

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

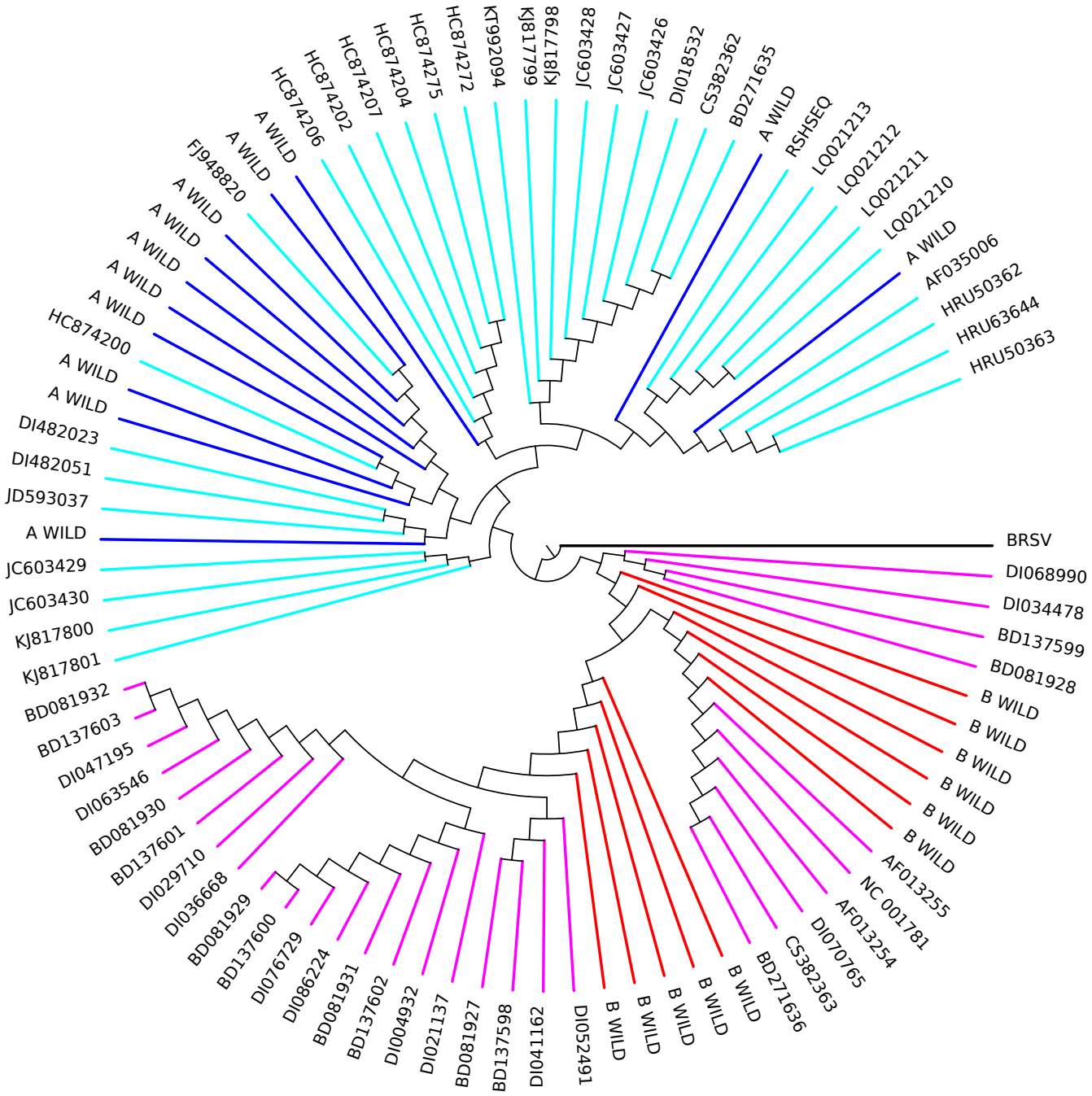

Figure 1. The phylogeny of gene L. Subtype A is in blue (wild-type) and cyan (vaccines); subtype B in red (wild-type) and magenta (vaccines). The tree was re-rooted on the bovine RSV ortholog, the purely vaccine clades condensed into single leaves, and the resulting tree rendered as SVG using the Newick Utilities.[10]

Table 4. Vaccine-linked, non-silent mutations. The amino acid (AA) position is with reference to GU591769; the wild-type (WT) and vaccine amino acids columns list the most frequent amino acid in wild-type and vaccine sequences, respectively. Fisher's test compares minor allele (mA) frequencies for vaccines and WT, given the counts of sequences in each. Only positions with p-value lower than $\alpha = 0.01$ corrected for multiple comparisons are shown. No such positions were found in subtype B. The effect size is given as the odds ratio between mA frequencies in vaccines versus WT. CI: 95% confidence interval for the odds ratio.

Gene	AA position	WT AA	Vaccine AA	BLOSUM100 Score	# WT sequences	mA frequency, WT	# Vaccine sequences	mA frequency, Vaccines	Fisher test p-value	CI low	Odds ratio	CI high
F	4	P	L	-4	851	0.006	7	0.57	1.9×10^{-07}	28	206.97	1848.62
F	8	T	A	-1	852	0.065	7	1	7.4×10^{-09}	20	∞	∞
F	20	L	F	0	856	0.057	7	1	3.4×10^{-09}	23	∞	∞
F	20	L	F	0	856	0.053	7	0.86	2.1×10^{-07}	12	106.21	4810.44
F	25	S	G	-1	858	0.004	7	0.57	5.2×10^{-08}	40	335.53	3980.31
F	384	I	V	2	865	0.025	7	1	2.1×10^{-11}	50	∞	∞
G	4	T	N	-1	1466	0.046	5	1	2.5×10^{-07}	19	∞	∞
G	38	I	V	2	1539	0.035	5	1	6.9×10^{-08}	24	∞	∞
G	107	T	I	-2	1539	0.049	5	1	3.5×10^{-07}	17	∞	∞
G	121	S	G	-1	1539	0.038	5	1	1.0×10^{-07}	22	∞	∞
G	123	E	K	0	1539	0.042	5	1	1.7×10^{-07}	20	∞	∞
G	141	I	T	-2	1539	0.044	5	1	2.1×10^{-07}	19	∞	∞
G	235	K	E	0	1539	0.044	5	1	2.1×10^{-07}	19	∞	∞
G	246	R	I	-4	1539	0.066	5	1	1.4×10^{-06}	13	∞	∞
G	260	H	L	-4	1539	0.039	5	1	1.1×10^{-07}	22	∞	∞
G	306	Y	S	-3	1539	0.036	5	1	8.2×10^{-08}	23	∞	∞
G	316	P	S	-2	1534	0.050	5	1	3.8×10^{-07}	17	∞	∞
G	319	S	P	-2	1533	0.038	5	1	1.1×10^{-07}	22	∞	∞
L	59	I	M	1	651	0.015	6	1	7.3×10^{-11}	62	∞	∞
L	81	I	L	1	651	0.063	6	1	9.8×10^{-08}	17	∞	∞
L	164	L	S	-4	654	0.063	6	1	9.6×10^{-08}	17	∞	∞
L	1184	F	L	0	655	0.075	6	1	2.6×10^{-07}	14	∞	∞
L	1242	A	S	1	655	0.003	6	0.67	2.8×10^{-08}	49	532.69	8192.00
L	1556	R	K	2	655	0.063	6	1	9.5×10^{-08}	17	∞	∞
L	1662	Y	H	1	655	0.064	6	1	1.1×10^{-07}	16	∞	∞
L	1723	T	A	-1	655	0.064	6	1	1.1×10^{-07}	16	∞	∞
L	1726	V	I	2	655	0.066	6	1	1.2×10^{-07}	16	∞	∞
L	1761	T	A	-1	655	0.064	6	1	1.1×10^{-07}	16	∞	∞
L	1783	K	R	2	655	0.063	6	1	9.5×10^{-08}	17	∞	∞
L	1945	P	S	-2	655	0.005	6	0.67	6.6×10^{-08}	39	376.81	6473.58
L	2021	V	I	2	655	0.072	6	1	2.0×10^{-07}	15	∞	∞
L	2021	V	I	2	655	0.079	6	1	3.6×10^{-07}	13	∞	∞
N	224	Y	H	1	680	0.082	6	1	4.3×10^{-07}	13	∞	∞
N	349	K	E	0	680	0.051	6	1	3.2×10^{-08}	20	∞	∞
N	357	C	S	-2	680	0.062	6	1	8.7×10^{-08}	17	∞	∞

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

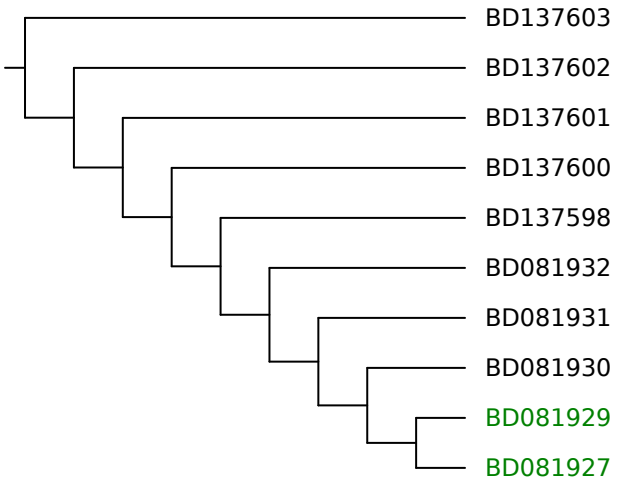

Figure 2. An extremely unbalanced clade. This is a subtree of the F gene tree. From such a clade we only retain one (at random) of the "end-of-chain" sequences, marked in green. These are likely to contain at least as many mutations as any other sequence in the clade.

fewer than ten vaccines a power higher than 0.8 can be achieved if the allele frequencies are sufficiently different. For example, consider a position in which only 5 out of 851 wild-type sequences have the minor allele while 4 out of 7 vaccine sequences have it (as is the case of gene F for subtype A at nucleotide position 11), corresponding to relative frequencies of 0.006 and 0.57, respectively: power estimates (with $\alpha = 0.01$ and 1,000 simulations) consistently exceed 0.95.

Non-attenuated Vaccines Vaccine sequences not explicitly annotated as "attenuated" may result from a process other than attenuation and thus not harbour attenuation-linked mutations. This will result in an underestimation of the frequency of these mutations among attenuated sequences, resulting in a higher false negative rate.

Independence of Mutation events The phylogenetic trees constructed for each gene of both HRSV subtypes clearly show extremely unbalanced nodes, that is, nodes of which one child is a single leaf, and the other child has several descendants. That other child node is frequently itself extremely unbalanced, resulting in strongly or even wholly unbalanced clades (Figure 2).

Such clade topologies are expected to arise when a sample is taken from a culture to start a new culture, and the process is iteratively repeated: all cultures but the first are more closely related to each other than any of them is to the first culture, and the same situation obtains recursively *within* the last-but-one cultures. The result is an extremely unbalanced phylogeny.

Any early mutation will therefore (barring reversion) be inherited by most of the members of the clade, and be found at a high frequency. However, this high frequency is not indicative of a high number of independent mutation events, but rather, of inheritance. It is therefore wrong to interpret this high frequency as a sign of positive selection. In the context of this work, the mutation would be wrongly thought of as a result of the attenuation process.

Our approach to this problem, namely the selection of a single sequence per fully unbalanced vaccine clade, makes it less likely that any similarities are due to shared ancestry. Furthermore, the fact that we select among the sequences with the maximal number of ancestors in the clade should maximize the number of mutations that we are able to detect.

Effect on Phenotype Most mutations retained by our pipeline are silent. Among the non-silent ones, the most likely to affect the phenotype are arguably those which are most rarely encountered between homologous sequences, as measured for example by the BLOSUM[11] amino acid substitution scores.

Another approach to estimating the effect of a mutation is to consider any known function of the affected protein or regions therein. Table 5 shows annotations pertaining to regions harbouring mutations, according to UniProtKB[12].

Table 5. Annotations in UniProtKB sequences at positions with vaccine mutations.

Gene	UniProtKB ID	AA Position	Annotation
F	FUS_HRSVA	8	signal peptide
		20	signal peptide
		384	fusion glycoprotein F1
G	GLYC_HRSVA	4	N>A: partial loss of interaction with protein M

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gene	UniProtKB ID	AA Position	Annotation
		141	O-linked GalNAc
		235	O-linked GalNAc
		290	O-linked GalNAc
L	L_HRSVA	1940	O-ribose methyltransferase domain

Known Attenuation Variants UniProtKB entry FUS_HRSVA lists five "Cold-passage attenuated" variants (positions 102, 218, 379, 447, and 523), none of which were retained by our procedure. So does NCAP_HRSVA (V267I). Since the method of attenuation (and indeed of vaccine production) is usually not known for our vaccine genomes, it is impossible to use these known mutations as positive controls, and we can only note that the mutations we identified are hitherto unknown.

Affected Subtypes Our procedure only identified mutations in vaccines derived from HRSV subtype A, but the failure to find mutations in type B is possibly due to the lower number of sequences available from that subtype, as this would lower the power of the statistical test used to identify positions with distinct allele frequency distributions between wild-type and vaccine sequences.

Limitation to Coding Regions Our procedure is limited to coding regions of the HRSV genome. Any mutation falling outside of coding regions will thus be missed, and attenuating mutations in non-coding regions have been reported (in transcription start sites, for example[13]). However, of the six attenuating point mutations reported by Karron *et al.*[8], five were in coding regions, which suggests that limiting our procedure to coding regions is unlikely to cause it to overlook all (or perhaps even most) mutations of interest.

Incompatible Mutations While the design of vaccines by reverse genetics will likely benefit from an expanded list of attenuation-linked mutation candidates, it should be borne in mind that mutations may be pairwise incompatible.[14]

In summary, we identified a set of positions and corresponding alleles, in the human respiratory syncytial virus subtype A (HRSV-A), which are likely to be under selection by the process of attenuation carried out in the production of candidate live HRSV vaccines. We also found that the vaccine sequences almost never have the wild-type allele at the corresponding position. If, as seems likely, the vaccine sequences are not closely related, this means that the separate attenuation processes have independently converged on the same mutations. We found no comparable mutations in subtype B, but this can be due to small sample size rather than lack of selection pressure.

These findings adds to the current understanding of the genome-scale effects of attenuation in HRSV, and may find application in the direct engineering of live attenuated vaccines.

Methods

Overview

We aligned homologous gene sequences from vaccine candidates and wild-type viruses together (but keeping the two subtypes separate), and searched for positions with significantly different allele frequencies in the two groups. To identify positions with different allele frequencies in vaccine *vs.* wild-type sequences, we first had to distinguish these two categories, and to classify sequences by gene and subtype when not annotated. The detailed procedure (implemented using Python[15] (including BioPython[16]), Bash[17], SQLite[18], and R[19]) follows.

Download and Sorting of Raw Sequence Data

We downloaded all entries with key phrase `Respiratory syncytial virus` from GenBank[20]. There were 28,248 at the time of download - see supplementary file `all.gb`. This included both whole-genome and partial HRSV sequences, of both subtypes, mostly wild type but with a small minority annotated as "vaccines". Although it would be more accurate to refer to these as "vaccine candidates", this phrase is unwieldy and we will refer to these as "vaccines" in the rest of this article.

A small number of entries were not from human respiratory syncytial virus, and were discarded.

We then separated wild-type from vaccine sequences based on annotation: any entry containing the string `VACCINE`, `ATTENUAT`, or `PATENT` in the Definition, Source, Reference Titles, or Comments fields of the GenBank entry was considered a vaccine sequence (this assumes that patented sequences are vaccines); all others were considered wild-type.

Likewise, we separated the two subtypes of wild-type viruses by annotation, using various regular expression searches for patterns such as `type (A|B)` in the Source, Organism, or Definition fields of the Genbank entry, as well as in its Feature Table (for the full details, see the source code in file `allRSVtoCSV.py` in the supplementary data). Entries whose subtype could not be determined were discarded.

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Next, we discarded any wild-type sequence containing unknown nucleotides (N)s. We then used entry annotations to extract the coding sequences of each wild-type entry of recognizable subtype: we scanned the GenBank entry's Feature Table for coding sequences (features of type CDS), within which we searched for keys such as gene or product. We then folded the resulting names into a controlled name list (e.g., entry annotations "L", "l", "polymerase", and "large polymerase", which in fact denote the same gene, were all mapped to "L"). The complete mapping is available as gene_name_map.py in the supplementary data.

Finally, we discarded partial gene sequences by retaining only those sequences that were at least 9/10 of the maximum length for the corresponding gene and subtype. The rationale for this is that short sequences can not only slow down the alignment process while contributing little information, but can also cause too many columns to be dropped when building phylogenetic trees.

The procedure up to this point yielded i) twenty sets of wild-type DNA sequences of recognized subtype and gene (namely, ten genes each for both subtypes), all devoid of Ns, and full-length or nearly so; and ii) a set of (as yet) uncharacterised vaccine sequences.

Classifying Vaccine Sequences

Since the vaccine sequences rarely featured annotation of gene or subtype, these were inferred by searching them for matches of profile hidden Markov models (HMMs) built from wild-type sequences of known gene and subtype, as follows.

Reference (Wild-type) HMMs All twenty wild-type sequence sets obtained as described above were aligned using Mafft[21], with default parameters. The resulting alignments were input to the HMMER[22] 3.1 package's hmmbuild program, also using default parameters, yielding one model for each (subtype, gene) combination. The models were collected into a database using HMMer's hmmpress.

All vaccine sequences were then scanned, using HMMer's nhmmscan, for matches of any HMM in the database (it being understood that a given vaccine sequence may match more than one HMM, albeit at different, non-overlapping positions - as is expected, for instance, of whole-genome vaccine sequences). Matches were classified (i.e., attributed to a subtype and gene) according to the highest HMM score, and sorted into separate files by subtype and gene.

Selecting Vaccine Sequences

We computed a phylogenetic tree for each gene, as follows: first, we pooled the wild-type sequences that were used for alignment (namely, those that matched all our quality criteria) of the two subtypes, yielding a single sequence set per gene. Vaccine sequences were added to the corresponding set, subject to the condition that they also be at least 90% as long as the maximum length for the corresponding gene - again, in order to minimize the loss of aligned positions when computing the tree. Finally, the sequence of the orthologous gene from the bovine respiratory syncytial virus (NC_001989) was added to each set as an outgroup.

Each of the ten resulting sets was aligned with Mafft, using default parameters. The resulting FastA was converted to PHYLIP[23] format using the EMBOSS[24] package's seqret program. Trees were computed with Phyml[25] version 3.0, using the general-time-reversible-Gamma model, four substitution rate categories, and maximum-likelihood estimates of base frequencies, transition/transversion ratio, as well as gamma distribution shape parameter. The resulting trees, as well as various graphical representations of them, are available in the supplementary data.

Reduction of Similarity by Inheritance: Selection of Clade Representatives We selected one vaccine sequence from each purely- or almost-purely-vaccine, completely unbalanced clade, as revealed by the phylogenetic trees. These clade representatives were randomly drawn from the two sequences with the largest number of ancestors within the clade (see Discussion for rationale).

Aligning the Vaccines to the Wild-type References The clade representatives were added to the corresponding alignments of wild-type sequences using Mafft with the --add option. This aligns the additional (namely, vaccine) sequences without changing the original (i.e., wild-type) alignment.

Frameshifts Examining the sequences with an alignment viewer[26] showed that twelve sequences had frame shift mutations. These sequences were removed and the alignments recomputed until no frame shifts were observed.

Finding Mutated Positions The search for positions likely to be under positive selection pressure from the attenuation process was carried out using a Python script (see find_mutations.py in the supplementary data) written for this purpose, its main steps are as follows:

- for each position in the alignment:
    1. identify the most frequent allele (pooling vaccine and wild-type sequences), define this to be the *major allele*;

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. fold the remaining three alleles into category *minor allele*;
3. compare the frequency of the major versus minor alleles, in wild-type versus vaccine sequences, using Fisher's exact test;
4. if the test's p-value is below some predefined threshold, register this position as a vaccine-linked mutation position, then determine if the most frequent allele in vaccine versus wild-type sequences result in different amino acids: if so, report the position as a vaccine-linked, non-silent mutation position.

Data Availability

Code, data, and the main intermediary results are available from https://figshare.com/articles/shared_data_for_2019_RSV_article/10636784.

References

1. Nair, H. *et al.* Global burden of acute lower respiratory infections due to Respiratory Syncytial Virus in young children: a systematic review and meta-analysis. *The Lancet* **375**, 1545–1555 (2010).
2. Jafri, H. S., Wu, X., Makari, D. & Henrickson, K. J. Distribution of Respiratory Syncytial Virus subtypes A and B among infants presenting to the emergency department with lower respiratory tract infection or apnea. *The Pediatr. infectious disease journal* **32**, 335–340 (2013).
3. Hall, C. B., Simőes, E. A. & Anderson, L. J. Clinical and epidemiologic features of Respiratory Syncytial Virus. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 39–57 (Springer, 2013).
4. Barik, S. Respiratory Syncytial Virus mechanisms to interfere with type 1 interferons. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 173–191 (Springer, 2013).
5. Chu, H. Y. & Englund, J. A. Respiratory Syncytial Virus disease: prevention and treatment. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 235–258 (Springer, 2013).
6. Collins, P. L., Fearns, R. & Graham, B. S. Respiratory Syncytial Virus: virology, reverse genetics, and pathogenesis of disease. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 3–38 (Springer, 2013).
7. Morrison, T. G. & Walsh, E. E. Subunit and virus-like particle vaccine approaches for Respiratory Syncytial Virus. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 285–306 (Springer, 2013).
8. Karron, R. A., Buchholz, U. J. & Collins, P. L. Live-attenuated Respiratory Syncytial Virus vaccines. In *Challenges and Opportunities for Respiratory Syncytial Virus Vaccines*, 259–284 (Springer, 2013).
9. Minor, P. D. Live attenuated vaccines: historical successes and current challenges. *Virology* **479**, 379–392 (2015).
10. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
11. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**, 10915–10919 (1992).
12. UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515 (2018).
13. Whitehead, S. S., Firestone, C.-Y., Collins, P. L. & Murphy, B. R. A single nucleotide substitution in the transcription start signal of the m2 gene of Respiratory Syncytial Virus vaccine candidatecpts248/404 is the major determinant of the temperature-sensitive and attenuation phenotypes. *Virology* **247**, 232–239 (1998).
14. Whitehead, S. S. *et al.* Addition of a missense mutation present in the l gene of Respiratory Syncytial Virus (rsv) cpts530/1030 to rsv vaccine candidate cpts248/404 increases its attenuation and temperature sensitivity. *J. virology* **73**, 871–877 (1999).
15. http://www.python.org.
16. Cock, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
17. https://www.gnu.org/software/bash/.
18. www.sqlite.org.
19. R Core Team. R: A language and environment for statistical computing; 2015. https://www.R-project.org (2018).

bioRxiv preprint doi: https://doi.org/10.1101/862185; this version posted December 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

20. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research* **45**, D12 (2017).

21. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. biology evolution* **30**, 772–780 (2013).

22. http://hmmer.org.

23. Felsenstein, J. PHYLIP (Phylogeny Interference Package), ver. 3.68. Distributed by the author, University of Washington, Seattle, Washington, USA (2009).

24. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* (2000).

25. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. biology* **59**, 307–321 (2010).

26. Larsson, A. Aliview: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

Author contributions statement

J.F. and T.J designed the experiments, except for the statistical aspects which were handled by N.C. T.J. wrote and ran the code implementing the experiments, and wrote the manuscript. L.K. contributed expertise in virology and to the experimental design All authors reviewed the manuscript.

Additional information

The authors declare no conflict of interest.