Signal sequences in the genome of *Mononegavirales* regulate the generation of copy-back defective viral genomes

Yan Sun[1], Eun Ji Kim[2], Emily Speranza[3], Louis J. Taylor[4], Divyansh Agarwal[1], Dawn M. Gerhardt[5], Richard S. Bennett[5], John H. Connor[3], Gregory R. Grant[2,5], and Carolina B. López[1,*]

[1]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States.

[2]Institute of Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States.

[3]Department of Microbiology, Bioinformatics Program, National Emerging Infectious Diseases Laboratories, Boston University, Boston, MA 02118, United States.

[4]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States.

[5]Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, Maryland, United States.

[6]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States.

***Corresponding author**: Carolina B. López, PhD. Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States

TEL 215-573-3493; Email lopezca@upenn.edu

ABSTRACT

Copy-back defective viral genomes (cbDVGs) are the primary initiators of antiviral immunity during infection with respiratory syncytial virus (RSV), parainfluenza virus, and other common pathogens. However, the mechanism governing cbDVG generation remains unknown limiting our ability to manipulate their content during infection. Here we developed VODKA (Viral Opensource DVG Key Algorithm) to identify cbDVGs from RNA-Seq data from infected samples and used these DVG sequences to identify signals that regulate cbDVG generation. After validation using data from an *in vitro* Sendai virus infection and Ebola virus-infected non-human primates, we applied VODKA to datasets from RSV-positive patients. VODKA identified common cbDVGs across multiple samples, and predicted specific genomic regions that mediate RSV cbDVG formation. Using an RSV minigenome system, we demonstrate that these viral sequences drive cbDVG wroduction. Importantly, the region involved in polymerase rejoin and reinitiation was highly conserved and its nucleotide composition determined the quality of cbDVGs generated. Our results indicate that generation of cbDVGs during natural infection is regulated by specific viral sequences and open the possibility to manipulate DVG formation during infection.

Keywords: copy-back defective viral genomes, specific genomic sequences, Respiratory Syncytial Virus (RSV), RSV minigenome, Ebola virus (EBOV)

INTRODUCTION

Severely truncated defective viral genomes (DVGs) are frequently generated during viral replication. DVGs are unable to self-replicate but can propagate with the help of their parental virus[1, 2]. Since their discovery in 1943, defective interfering particles (DIPs) containing DVGs have gained attention due to their ability to interfere with the replication of the standard virus and stimulate strong antiviral responses[3-7]. The promoter region retained in DVGs, in conjunction with their small size, enables them to replicate faster and compete for resources such as the viral polymerase and nucleoproteins, thereby reducing virus titers[8-10]. For decades, DVGs were considered artificial viral products because their generation was only observed *in vitro.* However, the development of DVG specific PCR and next-generation deep sequencing has allowed for DVG identification in samples from patients infected with various viruses including dengue, influenza A, and respiratory syncytial virus (RSV)[11-13]. DVGs potentiate the host innate immune response and attenuate infection in animal models[6, 13-16]. Importantly, naturally generated DVGs correlate with stronger antiviral responses in RSV-infected patients[13], as well as with reduced disease severity in influenza-infected patients[15]. Accumulating evidence also shows that viral stocks of standard live attenuated measles and influenza vaccines contain DVGs[17, 18]. Lastly, DVGs facilitate the establishment of persistent virus infections[19, 20], further highlighting the critical function of DVGs in multiple aspects of viral infection and urging for a better understanding of the molecular mechanisms that underlie their generation.

Two basic patterns of DVGs, deletion and copy-back (cb), result in DVG-containing viral particles[21]. DVGs are present in multiple *Mononegavirales* including *rhabdoviridae* such as vesicular stomatitis virus (VSV), *paramyxoviridae* such as measles and Sendai Virus (SeV), *pneumoviridae* such as RSV, and *filoviridae* such as Ebola virus[22]. Deletion DVGs retain the

standard genomic 3' and 5' ends but carry an internal deletion[23, 24]. In contrast, cbDVGs derived from the 5' end of the viral genome, are flanked by reverse complementary ends[10]. cbDVGs arise when the viral polymerase detaches from its template at the "break point" and continues in the nascent strand where it resumes elongation at the "rejoin point" by copying the 5' end of the nascent daughter strand[2, 25]. Both deletion and cbDVGs are thought to result largely from errors made by the RNA-dependent RNA polymerase (RdRps) during replication due to its lack of proofreading activity[21]. Even within stocks prepared from the same parental virus, various DVGs can be detected depending on the virus preparation method and passaging numbers[26].

We identified analogous DVGs from three different RSV viral stocks leading us to hypothesize that DVG generation may not be as stochastic as generally thought. To our knowledge, no clear pattern for the position of break and rejoin sequences in cbDVGs has been described. In order to elucidate the mechanism by which natural cbDVGs originate and identify the positions in the viral genomes where cbDVGs form, we developed an algorithm to identify cbDVG junction regions consisting of continuous break and rejoin points. We validated this Viral Opensource DVG Key Algorithm (VODKA) using an RNA-seq dataset from SeV-infected cells containing one dominant well-characterized cbDVG (DVG-546)[14] and further tested its ability to identify cbDVGs *in vivo* using RNA-seq data from serial blood samples of EBOV infected non-human primates (NHPs). Interestingly, VODKA identified large amount of cbDVGs from 8 different Ebola-infected NHPs. We then used VODKA to identify cbDVGs generated in two pediatric respiratory samples that were pre-screened for RSV cbDVGs[13], and our algorithm identified several conserved genomic regions enriched in DVG reads and four potential DVG junctions shared by both RSV-positive patients, supporting a significant degree of conservation in DVGs generated naturally during infection. Next, utilizing an RSV minigenome

system, we tested whether the identified viral genomic regions from RSV contributed to cbDVG generation and found that the regions containing break points reduced the expression of the reporter gene built into the construct, whereas regions containing rejoin points did not. Additionally, compared to break points, the rejoin points were more conserved during cbDVG generation. We further discovered that an optimal GC composition around the rejoin points was essential for the viral polymerase to rejoin the nascent strand and produce cbDVGs. Our findings prove that conserved DVGs are naturally generated in infected patients. More importantly, VODKA identified several specific viral sequences within the viral genome that are preferred sites for DVG generation. This knowledge can be used to manipulate DVG generation in the future, and paves the way for therapeutic intervention for various viral infections by means of viral attenuation and/or vaccine development.

RESULTS

The generation of RSV cbDVGs *in vitro* is not stochastic.

To test the hypothesis that cbDVG generation is not completely driven by chance, we first examined whether similar cbDVGs arise during RSV replication *in vitro*. RSV-HD stock 1 and RSV-HD stock 2 were generated from the same parental A2 virus stock purchased from ATCC, but were passaged and amplified using a different MOI as indicated in the methods section. In contrast, RSV-GFP was propagated from an infectious clone where GFP was inserted before NS1 gene into RSV D53, a different A strain backbone[27]. At 24 h post infection, while the three RSV stocks were comparable in terms of replication (Fig. 1a), they generated different patterns of cbDVG bands, which was consistent with previous studies (Fig. 1b). Interestingly, we identified cbDVGs that had relatively close break points and rejoin points across the three

different virus stocks (Fig. 1b, white arrow; sequences in Fig. S1). Moreover, two of these cbDVGs rejoined at the exact same position (14996, Fig. 1c), and the overall range of rejoin points from the cbDVGs was narrower than the span of their corresponding break points (20bp vs 99bp, Fig. 1c). These observations support our hypothesis that the generation of RSV copy-back DVGs is not a random process, and that there might be genomic cues, especially at the rejoin positions, that guide cbDVG formation.

VODKA identified predicted DVG junctions during SeV *in vitro* infection.

Due to the significant homology between cbDVGs and their parental full-length viral genome, identification of cbDVG-specific reads from high-throughput sequencing is challenging. To identify cbDVG-specific sequences, we developed an unbiased screening method, using a novel algorithm, VODKA. Briefly, RNA was first extracted from various samples and used for library preparation for Nextseq 500 system (Illumina). The output from RNA-seq was then screened using VODKA to identify potential junction regions within cbDVGs (Fig. S2a). As illustrated in Fig. 2a, the break point (T') and rejoin point (W) are originally separated in the parental viral genome but in cbDVGs, they become connected and form one continuous sequence, referred to as the DVG junction. Therefore, the DVG junction is unique for cbDVGs and is absent in the full-length viral genome.

VODKA first identifies the sequencing reads that capture the DVG junction region (DVG junction reads, indicated in red, Fig. 2a). Then, it performs a secondary screening that maps the selected reads to the reference genome to filter out reads that fully align to the reference genome (false-positives). We corroborated VODKA's performance using a well-studied DVG, DVG-546, in SeV Cantell, which accumulates to high levels at 24 h (Fig. 2b). We ran VODKA on

RNA-Seq data from SeV Cantell infected cells. VODKA predicted a total 98,626 DVG junction reads, and 99.92% of these (98,543 reads) mapped exactly to the known junction region of DVG-546 (Fig. 2c, bottom panel). By aligning the DVG junction reads to the SeV full-length antisense genome, we determined the location of the major break and rejoin regions (blue peaks in the upper panel of Fig. 2c). Since SeV Cantell only has one major DVG (Fig. 2b), the break and rejoin points aggregated, as expected, under two major but separate peaks. When we examined the reads that aligned to these two locations, each read contained a portion that fully aligned to the reference genome (grey nts), and another portion that did not align (colored nts in Fig. S2c). The aligned portion of reads in break region (pink boxed area in Fig. 2c and S2c) and in rejoin region (gray boxed area in Fig. 2c and S2c) was corresponding to the DVG-546 sequence before the break point and after the rejoin point, respectively. We used this as an important criterion in determining major break and rejoin regions in other RNA-seq datasets as well. In addition to the sequences of DVG junction reads, VODKA outputs the exact position of break and rejoin points in the full-length genome in a "T'_W" format (Table 1). For SeV DVG-546, the prediction "14932±1_15292±1" exactly matched the results obtained through conventional cloning and sequencing. Taken together, SeV DVG-546 validated that VODKA can efficiently and accurately identify junction reads of cbDVGs in infection *in vitro*.

VODKA identified DVG junction regions in samples from Ebola virus-infected NHPs

Ebola, a member of *filoviridae* is known to generate cbDVGs *in vitro*[28]. We next used VODKA to evaluate whether the algorithm could identify cbDVG during an experimental EBOV infection *in vivo*. The algorithm identified 126,554 DVG junction reads in an RNA-seq dataset from outbred rhesus macaques infected with the CO5 isolate of Ebola virus Makona

(EBOV/Mak-C05). Based on the overall alignment rate of DVG junction reads (total DVG reads divided by total number of reads) from individual subjects, regardless of any temporal confounders, three EBOV-infected macaques (NHP#D, NHP#K, NHP#H) were considered as cbDVG-negative. These three subjects showed similar, or even lower, percentage of DVG reads than the highest rate among four pre-infection samples (data not shown). The DVG reads found in pre-infection samples were considered background. Four animals (NHP#B, NHP#C, NHP#F, NHP#G) were classified as cbDVG strong-positive, and the remaining (NHP#A, NHP#E, NHP#J, NHP#I) as cbDVG weak-positive. All DVG junction reads were mapped to the Ebola reference genome NCBIKP096421, similar as SeV. We observed six major peaks, three breaks and three rejoins (Fig. 2d). Surprisingly, three major Ebola cbDVGs that dominated ~78% of the entire DVG reads (indicated by dash line linkage in Fig. 2d; details of cbDVGs in Table 1). The top two cbDVGs among them were each found in three different macaques (Table 1). These data provide additional evidence that VODKA can identify cbDVGs in RNA isolated from different infections and tissues.

VODKA identified common DVG junction regions from two RSV-positive pediatric respiratory secretion samples

To investigate whether VODKA could identify DVGs generated during natural infection, we next applied VODKA to the RNA-seq dataset obtained from RSV-positive pediatric clinical samples described previously[13]. VODKA output included a total of 652 DVG junction reads, which were then aligned to the reference genome of an RSV strain A isolate (clinically isolated in 2012, NCBIKJ672447). We observed multiple break and rejoin regions (Fig. 3a, blue peaks), and subsequently determined three major break peaks and three rejoin peaks (arrows in Fig. 3a

and Fig. S3) by applying the same criteria as the SeV Cantell dataset, including that at least 15 continuous nucleotides in the read must align with the reference genome and that the remainder of the reads must align to another discontinued region of the genome. Combined with the exact position of break and rejoin points within the VODKA-predicted DVG junction reads, we identified five major clinical DVGs using the cut-off of total read counts ≥ 9 (listed in Table 1). Strikingly, four out of the five DVGs were found in both patients.

Upon comparing DVG junctions identified from clinical samples with *in vitro* generated DVG junctions published previously[13], we found that one *in vitro* cbDVG, 13877_14956, had its break and rejoin points in close proximity to one clinical cbDVG, 13870_14977. To investigate the correlation between various break and rejoin points at a genomic distance scale, and further study the regions that are highly enriched with such points, we classified the three major break peaks from clinical samples into three break regions – Break1, 2, and 5 (black bar in Fig. 3a, Table S1) – and the three major rejoin peaks into two rejoin regions – Rejoin1, and 2 (grey bar in Fig. 3a, Table S1). Interestingly, all rejoin regions, from both clinical and *in vitro* cbDVGs, clustered within one narrow region close to the 5' trailer of the genome, whereas break regions were widely spread and further away from trailer (Fig. 3a). Furthermore, all the identified *in vitro* cbDVG rejoin points were concentrated in Rejoin1, which also had the highest coverage of DVG junction reads in clinical samples. For breaks, only Break1 contained break points from cbDVGs generated both *in vitro* and in clinical samples. Break4 contained *in vitro* cbDVG break points but was absent from the major regions identified from clinical samples (light grey bar in Fig. 3a, Table S1). Although there are other genomic regions where DVG junction reads aligned (Fig. S3), the alignment pattern of those reads violated the rules for being a valid DVG junction

read indicated above, indicating that they might be actual viral reads with simply a higher frequency of mutations.

Next, we examined whether DVG junction regions were conserved among RSV strains and isolates. Since DVG junction regions do not exist in the viral genome, the break and rejoin regions from different RSV isolates of both strains A and B were aligned separately (Fig. 3b). Overall, the junction regions were remarkably conserved within strain A, and the break regions were more conserved than rejoin regions between the two strains. The identification of common cbDVGs in two different patients, together with the conservation of DVG junction regions between clinical samples and *in vitro* samples, further supports that certain sequences within the viral genome are preferred by the viral polymerase to make cbDVGs.

A minigenome system corroborated that common RSV break and rejoin regions are involved in cbDVG generation.

To test the elongation ability of the viral polymerase in the presence of VODKA-predicted break and/or rejoin regions, we created an RSV minigenome system. To do this, we adapted a system used to study the transcription and replication signals within the RSV leader and trailer regions[29-32]. Here we included the reporter gene mKate2 in the minigenome backbone (BKB) for detection of the viral polymerase transcription product by flow cytometry. BKB also included multiple restriction enzyme sites to insert break and/or rejoin regions (Fig. 4a). As illustrated in Fig. 4b, mKate2 expression depends on the viral polymerase (formed through four helper plasmids: M2-1, L, P, and NP) and only occurs if the viral polymerase replicates the entire minigenome sequence from the trailer to the leader. Since the different DVG junction regions (break, rejoin, or both) are inserted right after the trailer, the viral polymerase will first encounter

those signals when generating the template for mKate2 transcription. If an effective break region (or a break-rejoin pair) is inserted, we expect reduced production of the template for mKate2 transcription leading to lower mKate2 expression.

Transfection with the BKB construct resulted in 8–17% of the cells expressing mKate2 when co-transfected with the four helper plasmids expressing the polymerase proteins, whereas no mKate2 expression was detected in negative controls where a viral polymerase was not included (Fig. 4c and 4d). Rejoin1 was the most abundant peak among all identified regions, highlighting its role in the cbDVG rejoining process. Break1 and Break4 represented the drop off positions of cbDVGs generated in clinical samples and *in vitro*, respectively (Fig. 3a), both of which could generate cbDVGs in conjunction with Rejoin1. Constructs containing only Rejoin1 result in similar mKate2 expression as BKB, whereas constructs containing Break1 or Break4 caused a ~30% reduction in mKate2 expression (Fig. 4d and 4e; corresponding flow plots in Fig. 4f). Furthermore, the difference in mKate2 expression among different constructs was not due to variable transfection efficiency among constructs (Fig. S4b and S4c), or cell death (Fig. S4d). These results are consistent with the concept that during DVG generation, viral polymerase falls off the template at the break region leading to reduced amount of newly synthetized template available for mKate2 transcription.

To formally assess DVG formation from the candidate sequences, designated pairs of break and rejoin regions were cloned into BKB. Pair1, composed of Break1/Rejoin1, was the junction region observed in cbDVGs generated *in vitro* (Table 1). Using flow cytometry we demonstrated that Pair1 caused a similar degree of reduction in mKate2 expression as Break1 alone (Fig. 4d, e, g). We took advantage of DVG-specific RT-PCR to examine whether any DVG was formed from Pair1, and detected two major bands in Pair1, both of which were absent

in Break1 alone (negative control; white arrows in Fig. 4h). Those two bands were both confirmed to contain cbDVGs by conventional sequencing (sequences in Fig. S4e). Their individual break and rejoin points are indicated in Fig. 4i. Interestingly, although the two break points were far away from each other and neither of them at the predicted site (Fig. 4i) in the DVG-like fragments created, the rejoin points were either exactly at or 2 nts ahead of the rejoin points that we identified from *in vitro* cbDVGs. Taken together, the RSV minigenome system allowed us to demonstrate that specific regions within the viral genome impact cbDVG generation and that rejoin points fall into highly defined regions of the viral genome.

Sequence composition within Rejoin1 affects viral polymerase rejoining activity and impacts cbDVG generation.

Since Rejoin1 was highly enriched with DVG junction reads relative to other regions in the RSV genome, and also contained the most conserved rejoin points, we asked if specific features within this region impacted cbDVG generation. We found that Rejoin1 localized within the most AU-rich area of the RSV genome (Fig. 5a), and therefore sought to assess whether the surrounding AU nucleotide composition plays a role in viral polymerase rejoining to form DVGs. Most identified rejoin points were concentrated in one narrow section towards the 3' end of Rejoin1 and the beginning of Trailer. To avoid affecting the gene end signal of the L gene and the trailer sequence, we chose six nucleotides (nts) at the beginning of this section to mutate, ranging from location 14959 to location 14964 according to the reference genome NCBIKJ672447. We either mutated the 6 nucleotides to all Us (named GC->Us), or all GCs (named AU->GCs; Fig. 5C). After transfection, RT-PCR was used to detect cbDVG-like fragments. Mutant GC->Us generated a stronger band than WT Pair1 (lane2, Fig. 5b), but this

band was missing in mutant AU->GCs. Instead, weak and smeared bands were observed for mutant AU->GCs (lane4, Fig. 5b). From the band areas marked by asterisks in the gel, we identified five different rejoin points from mutant AU->GCs and three from mutant GC->Us with the most represented rejoin point (▲in Fig. 5c) corresponding to the major DVG PCR amplicon (complete DVG-like sequences are shown in Fig. S5a). When compared to the infection and WT Pair1, mutant GC->Us did not generate rejoin points within and near the mutated region (grey area in Fig. 5c, ~30 nts). In contrast, mutant AU->GCs produced cbDVGs-like fragments with the rejoin points at the mutated area, akin to the WT construct (Fig. 5c).

To evaluate any possible bias from primer location, we designed two additional forward primers to detect cbDVGs-like fragments from the same samples, one of which was within Break1 and the other within Rejoin1 (Fig. 5c). Similar results were observed wherein mutant GC->Us led to one strong band and mutant AU->GCs exhibited a smear of bands (Fig. 5d). By sequencing, we confirmed that the three strong bands produced by three different primers in mutant GC->Us were analogous cbDVG-like fragments and shared their break and rejoin points (sequence in Fig. S5a, DVG 303bp). To test if the weaker, smear of bands in mutant AU->GCs was due to a general reduction of replication ability of the viral polymerase caused by the induced mutations, we introduced the same mutations in the construct with Rejoin1 alone and examined its mKate2 expression by flow cytometry. We found no significant differences between Rejoin1 and the two mutants, Rejoin1-GC->Us and Rejoin1-AU->GCs. Neither of these constructs reduced mKate2 expression compared to BKB (Fig. S5b and S5c), suggesting that the mutation itself did not alter the function of the RSV minigenome system. Taken together, these data suggest that a minimal content of GC nucleotides in the rejoin region determines if cbDVGs are produced at that particular region.

DISCUSSION

DVGs are emerging as critical regulators of viral replication and pathogenesis in multiple RNA virus infections, but the mechanisms modulating their generation are unknown. Historically, DVGs have been considered an extreme form of viral quasispecies that result from random mutations introduced during replication by the viral polymerase. However, mounting evidence suggests that the generation of DVGs, particularly those better characterized for their strong immunostimulatory activity (cbDVGs), is not totally random. Here, we demonstrate that our analytical pipeline, VODKA, identifies cbDVG specific junction reads in three distinct RNA-seq datasets, generated from three different viruses, from two different virus families, from infections *in vitro* as well as in NHPs and human patients. In all scenarios, DVG junction reads were clustered in specific regions, rather than spread evenly across the 5' end of genome. Furthermore, in RSV, four out of the five major cbDVGs identified by VODKA were present in both patients tested. These findings strongly suggest that cbDVG generation obeys certain rules. This is in accordance with previous reports of infection with a mutant measles virus *in vitro*, where certain regions within the viral genome have been shown to be the hotspots for cbDVG origination[26]. In *in vitro* infections with measles virus lacking the C protein, break points of cbDVGs were widely distributed along the genome, whereas the rejoin points were clustered in a narrow region close to 5' end of the genome, consistent with our findings in RSV and EBOV *in vivo* infections.

To further investigate the potential mechanism by which the VODKA-predicted sequence hotspots in RSV govern cbDVG generation, we cloned the representative sequences into a modified RSV minigenome system. We observed that the addition of break regions where

the polymerase drops off its template significantly diminished the ability of the polymerase to elongate the nascent strand. This observation led us to speculate that hotspots in the viral genome favored by the viral polymerase facilitate the generation of cbDVGs. The mechanism driving DVG generation at these hotspots remains to be identified. One possibility relates to the density of the nucleocapsid protein (NP) occupying the hotspot sequences as a recent study showed that lower density of NPs corresponds to more DVG formation in SeV infection[33]. However, in our sequencing of SeV, mutations described to be responsible for low NP density were absent, suggesting that alternative mechanisms are involved. The newly-identified hotspots in the viral genome that regulate DVG generation will permit more detailed mechanistic studies of this molecular machinery in the future.

Based on our minigenome data that recapitulated DVG formation at the expected rejoin position but not at the exact break point, rejoin points are more conserved among DVGs than their corresponding break points. The convergences of rejoins relative to breaks suggest that the rejoin position is also more fixed compared to break positions thus may be more influential on cbDVG generation. Indeed, our mKate2 expression results implied that rejoin sequences influence the function of break signals when inserted as pairs in the construct. For instance, the Pair1 construct had a slightly higher mKate2 expression than Break1 alone. Interestingly, although Rejoin1 was found to be in the most AU-rich region in the entire RSV genome, subsequent mutation experiments demonstrated that at least some GC nucleotides appear to be necessary for rejoining to occur at that particular region. When we substituted AUs for Gs or Cs, more DVG amplicons were observed, however, none of them led to a strong band leading us to conclude that a strong rejoin signal likely needs an optimal number of GC nucleotides. The requirement for GC nucleotides is intriguing. Normally, when in the “replicase” mode, the viral

polymerase is highly processive and skips potential stop signals within the genome. Only when it turns into a transcriptase does it recognize genomic signals such as “gene start” and “gene end” that normally begin with a cluster of Cs[34-36]. However, viral transcription only occurs from the leader sequence, not the trailer. It would be interesting to investigate whether the viral polymerase maintains a “transcriptase-like” ability when initiating replication at the trailer sequence and if it is prone to “see” the strong GC-containing rejoin signals.

How does the viral polymerase choose from the multiple break and rejoin regions that lead to the formation of cbDVGs? One scenario is that the strongest signals for break and rejoin are chosen by the viral polymerase. In our minigenome system, however, there is no significant difference between the effectiveness of Break1 and Break4, suggesting that the individual breaks have a similar ability to reduce the processivity of the viral polymerase. Alternately, different breaks might work synergistically to reduce polymerase processivity. In this case, longer DVGs containing more breaks would be preferred at early stages of the infection and eventually smaller cbDVGs would dominate the population after multiple passages due to faster amplification. Another factor influencing cbDVG formation is the length of the cbDVGs, which is tightly related to the spatial structure of the viral RNPs. In paramyxoviruses, although it is thought that only genomes with hexametric or heptametric lengths are efficiently replicated[37, 38], some cbDVGs generated *in vitro* and from RSV are found not to obey this rule[26, 39, 40]. Nonetheless, this rule certainly increases the replication efficiency of viral polymerase and thus cbDVGs that obey this rule have a fitness advantage. In our current minigenome system, neither the synergetic effects between break points, nor the impact of cbDVG length were tested.

Our study identified genomic regions involved in cbDVG generation. However, we do expect that other factors, such as viral proteins also play an important role in DVG generation.

For instance, influenza viruses harboring a high fidelity polymerase generate fewer DVGs[41]. Mutations in non-structural protein 2 of influenza have also been shown to increase the *de novo* generation of DVGs by altering the fidelity of viral polymerase[42]. Host factors are essential contributors to DVG generation as well[21]. For example, the vesicular stomatitis virus produces a large amount of snap-back DVGs in most cell lines, except human-mouse somatic cell hybrids, and this cellular attribute was mapped to human chromosome 16[43]. Similarly, infection with measles virus did not show *de novo* generation of DIPs in human WI-38 cells and SeV did not produce cbDVGs in chicken embryo lung cells[17, 44]. Despite the potential importance of these additional factors on DVG generation, the current work represents a major paradigm shift with the identification of sequences that regulate cbDVG formation. To date, all studies on DVG biology have been correlative in nature. This work opens up new areas of investigation and can ultimately allow us to manipulate the ability of viruses to produce DVGs as a powerful tool to study the role of DVGs in viral pathogenesis.

METHODS

Ethics statement.

Studies of human samples were approved by University of Pennsylvania Institutional Review Board. All work with infectious EBOV and potentially infectious materials derived from animals was conducted in a Biosafety Level 4 (BSL 4) laboratory at The NIAID Integrated Research Facility (IRF)-Frederick. Research was performed in accordance with animal study protocols approved by the Animal Care and Use Committee (ACUC) of the National Institute of Allergy and Infectious Diseases (NIAID) Division of Clinical Research (DCR), part of the National Institutes of Health (NIH). Protocols adhere to the recommendations stated in *The Guide for*

the Care and Use of Laboratory Animals and were developed in compliance with the United States Department of Agriculture (USDA) Animal Welfare Act regulations, U.S. Public Health Service Policy on Humane Care and Use of Laboratory Animals, NIH policies and guidelines, and other federal statutes and regulations relating to animals and experiments involving animals. The NIAID Integrated Research Facility (IRF)-Frederick, where the EBOV animal infection research was conducted, is fully accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International (AAALACi).

Cells and viruses.

A549 cells (human type II alveolar cells, ATCC, #CRM-CCL185) and HEp2 cells (HeLa-derived human epithelial cells, ATCC, CCL23) were cultured at 7% CO_2 and 37°C with Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (FBS), 1 mM sodium pyruvate, 2 mM L-Glutamine, and 50 mg/ml gentamicin. BSR-T7 cells (Hamster kidney cells, BHK cells constitutively expressing the T7 polymerase) were maintained in 10% FBS with 1 mg/ml Geneticin (Invitrogen). All cell lines were treated with mycoplasma removal agent (MP Biomedicals) and routinely tested for mycoplasma before use. Sendai virus Cantell stock (referred to as SeV HD, with high DVG particle content) was prepared in embryonated chicken eggs as described previously[14, 45]. The SeV HD stock used in these experiments had a high infectious to total particle ratio of 500:15,000. RSV-HD stock 1 (stock of RSV derived from strain A2, ATCC, #VR-1540 with a high content of cbDVGs) was prepared and characterized as described previously[13, 46]. RSV-HD stock 2 (stock of RSV derived from strain A2, ATCC, #VR-1540 with a high content of cbDVGs) was generated after 3 rounds of amplification at a multiplicity of infection (MOI) of 4 $TCID_{50}$/cell for two days each. RSV-GFP (kindly provided

by Dr. Mark Peeples, Nationwide Children's Hospital, Columbus, OH) was propagated in HEp2 cells in our laboratory as described previously[13, 46]. For viral infections *in vitro*, cells were incubated with RSV-HD stock 1, stock 2, or RSV-GFP at an MOI of 1 $TCID_{50}$/cell and harvested at 24 h post infection.

Plasmids

Mammalian expression vectors for RSV N (NR-36462), P (NR-36463), M2-1 (NR-36464), and L (NR-36461) proteins, and the RSV reverse genetic backbone pSynkRSV-line19F (NR-36460) were obtained from BEI Resources. Detailed information of the constructs can be found in reference[47]. The backbone plasmid of the RSV minigenome used for testing various DVG junction regions was constructed by cloning two regions of sequences amplified from pSynkRSV-line19F into the pSl1180 vector. The first region included a T7 promoter, a hammerhead ribozyme, RSV leader sequence, and genes encoding monomeric Katushka 2 (mKate2), while the second region included the RSV trailer sequence, a Hepatitis delta virus ribozyme and a T7 terminator. These regions were sequentially cloned into psl1180 vector using restriction enzyme pairs SpeI/SandI and SandI/EcoRI, respectively. The potential cbDVG break and/or rejoin regions (positions in Table S1) were then inserted between those two regions using restriction enzyme pairs NotI/SandI and SandI/SpaI, respectively. A detailed scheme of the construct can be seen in Fig. 4a. Pair1 and Rejoin1 mutations were introduced using the site-directed mutagenesis commercial kit QuickChange® II XL (Agilent, CA) according to the manufacture's protocol. All primers used for cloning are listed in Table S2.

RNA extraction and DVG-RT-PCR.

Total RNA was extracted using TRIzol or TRIzol LS (Invitrogen) according to the manufacturer's specifications. For detection of RSV DVGs in RSV infection, 1-2 μg of isolated total RNA was reverse transcribed with the DI1 primer using the SuperScript III reverse transcriptase (Invitrogen) without RNase H activity to avoid self-priming. Recombinant RNase H (Invitrogen) was later added to the reverse transcribed samples and incubated for 20 min at 37°C. DVGs were partially amplified using both DI1 primer and DI-R primer. The temperature cycle parameters used for the cbDVG-PCR in a BioRad C1000 Thermal Cycler were: 95°C for 10 min and 33–35 cycles of 95°C for 30 sec, 55°C for 30 sec and 72°C for 90 sec followed by a hold at 72°C for 5 min. Ramp rate of all steps was 3 degree/sec. Detailed method can be found in[13]. For detection of cbDVGs in the RSV minigenome system, extracted RNAs were treated with 2 μl TurboDNaseI (Invitrogen) for 15 min at 37°C, followed by reverse transcription. Same procedures as above were utilized, except replacing DI1 primer with DI-F1, DI-F2, and DI-F3 primers. These were then all paired with DI-R reverse primer to amplify the different sizes of PCR products. Sequences of all primers are listed in Table S2.

RT-qPCR

Total RNA (1 μg) was reversed transcribed using the high capacity RNA to cDNA kit from Applied Biosystems. cDNA was diluted to a concentration of 10 μg/μl and amplified with specific primers in the presence of SYBR green (Applied Biosystems). qPCR reactions were performed in triplicate using specific primers and the Power SYBR Green PCR Master Mixture (Applied Biosystems) in a Viia7 Applied Biosystems Light- cycler. Gene expression levels of RSV G were normalized to the *GAPDH* copy number. Sequences of primers used in these studies can be found in Table S2.

RNA-Seq

RNA-Seq for SeV Cantell was performed as previously described[19]. RNA was extracted using TRIzol reagent and was re-purified using the PicoPure™ RNA isolation kit (Thermo Fisher Scientific). RNA quality was assessed using the RNA Pico 6000 module on an Agilent Tapestation 2100 (Agilent Technologies) prior to cDNA library preparation. For SeV RNA-Seq dataset, total cDNA libraries were prepared starting from 75 ng of extracted raw RNA using the Illumina TruSeq Stranded Total RNA LT kit with Ribo-Zero Gold, according to the manufacturer's instructions. Samples were run on Illumina NextSeq 500 to generate 75 bp, single-end reads, resulting in 21-53 million reads per sample, with an average Q30 score ≥ 96.8%. For sequencing of samples from RSV-positive patients, 100 ng of extracted raw RNA was used for preparation of cDNA library using the same kit as above. Samples were run on Illumina NextSeq 500 to generate 150bp, paired-end reads, resulting in 60-170 million reads/sample with average Q30 score ≥ 84.6%. To analyze genomic AU-content relative to DVG break and rejoin points, we calculated the percentage of A or U nucleotides over sliding windows of 40 bases using the Python programming language (Python Software Foundation, https://www.python.org/). We plotted AU-content and cbDVG rejoin points in R using the ggplot2 package[48]. For the Ebola dataset that was kindly shared by Dr. John Connor (Boston University), healthy rhesus macaques, aged 2–5 years, mix genders, were anesthetized and challenged intramuscularly (IM) with a target dose of 1000 plaque-forming units (PFU) of the C05 isolate of Ebola Makona (full designation: Ebola virus/H.sapiens-tc/GIN/2014/Makona-C05-GenBank accession number KP096420.1), in the triceps muscle at the integrated research facility (IRF) in Fredrick, MD under ABSL-4 procedures. The dataset included 11 NHPs who succumbed to infection between 6 and 9 days. Blood samples were taken at pre-infection, 3, 5,

and 7 days post infection for all animals. For six out of twelve NHPs, samples were also collected at necropsy. Whole blood samples were inactivated in Trizol and RNA was extracted using the standard Trizol purification method. Samples were quantified and quality of RNA examined using Agilent bioanalyzer 2100 6000 RNA Nano Chips. Only samples with a high RNA quality (RIN > 7) and sufficient RNA concentration were carried out for sequencing. Libraries were prepped using the TruSeq stranded RNA RZ gold kit and were sequenced on a NextSeq.

Viral Opensource DVG Key Algorithm (VODKA)

Based on our *in vitro* RSV experiments, we made the assumption that most cbDVGs are generated from the viral sequence near the 5' end region of the genome (close to the Trailer). Therefore, starting with the last 3kb of a reference viral genome, we built an index of potential DVG sequences by taking all possible combinations of two non-overlapping segments of L bases, where L is the read length. The segments are linked by reverse complementing the second segment (C-D) and adding the first segment (A-B) to it (Fig. S2). Sequenced reads are aligned to the potential DVGs using bowtie2[49], and subsequently undergo two filtering steps. First, reads are removed unless they map across a breakpoint (A_C) with at least 15bp of mapped segment on each side. Second, the reads that map cleanly to the reference genome are filtered out. This pipeline gives the output read counts for each breakpoint (A_C). To be consistent with the structure of copy-back DVGs in Fig. 2a, C is equivalent to breakpoint T' and A is equivalent to rejoin point W. VODKA output reads were further aligned to reference viral genomes (RSV: NCBI accession number KJ672447; Ebola: NCBI accession number KP096421) or known SeV DVG-546 to identify the potential DVG junction regions using the Genious 7.0 software.

Transfection and RSV minigenome system

BHK cells constitutively expressing the T7 polymerase (BSR-T7 cells) were transfected with different minigenome constructs as well as the sequence–optimized helper plasmids encoding N, P, M2-1, and L, all under T7 control[47]. Using lipofectamine 2000 (Invitrogen), the plasmids were transfected at a concentration of 0.4 μg N, 0.4 μg P, 0.4 μg M2-1, 0.4 μg L, and 0.2 μg pSynkRSV-line19F. Cells were incubated with transfection complex (total plasmid: lipofectamine=1:3.3) for 2 h at room temperature and then at 37°C for overnight using Opti-MEM as medium. The following morning, the medium was replaced with antibiotic free tissue culture medium containing 10% FBS. Cells were harvested at 48 h post-transfection for either RNA extraction or flow cytometry.

Flow Cytometry

Transfected BSR-T7 cells were trypsinized 48 h post transfection and were either directly diluted in FACS buffer (PBS containing 20 mM EDTA) or stained with aqua L/D. Cells were washed twice in FACS buffer before flow cytometry analysis on an LSRFortessa (Becton Dickinson). Data analysis was performed using Flowjo version Legacy.

Statistical analysis

All statistical analyses were performed with GraphPad Prism version 5.0 (GraphPad Software, San Diego, CA) and *R* v*3.4.1*. A statistically significant difference was defined as a *p*-value <0.05 by one-way analysis of variance (ANOVA) with a *post hoc* test to correct for multiple comparisons (based on specific data sets as indicated in each figure legend).

Code availability

The VODKA algorithm is open-source, and available for use by the research community at: https://github.com/itmat/VODKA.

Data availability

All data are available upon request to the corresponding author. Raw RNA-Sequencing data of FISH-FACS sorted SeV infected cells, RSV patients, EBOV infected NHPs have been deposited on the Gene Expression Omnibus (GEO) database for public access (SeV:GSE96774; RSV: GSE114948; EBOV: GSE115785).

REFERENCES

1. Huang, A.S. & Baltimore, D. Defective viral particles and viral disease processes. *Nature* **226**, 325-7 (1970).
2. Lazzarini, R.A., Keene, J.D. & Schubert, M. The origins of defective interfering particles of the negative-strand RNA viruses. *Cell* **26**, 145-54 (1981).
3. Henle, W. & Henle, G. Interference of Inactive Virus with the Propagation of Virus of Influenza. *Science* **98**, 87-9 (1943).
4. von, M.P. Propagation of the PR8 strain of influenza A virus in chick embryos. II. The formation of incomplete virus following inoculation of large doses of seed virus. *Acta Pathol Microbiol Scand* **28**, 278-93 (1951).
5. Von Magnus, P. Incomplete forms of influenza virus. *Adv Virus Res* **2**, 59-79 (1954).
6. Lopez, C.B. Defective viral genomes: critical danger signals of viral infections. *J Virol* **88**, 8720-3 (2014).
7. Weber, M. & Weber, F. RIG-I-like receptors and negative-strand RNA viruses: RLRly bird catches some worms. *Cytokine Growth Factor Rev* **25**, 621-8 (2014).
8. Marriott, A.C. & Dimmock, N.J. Defective interfering viruses and their potential as antiviral agents. *Reviews in Medical Virology* **20**, 51-62 (2010).
9. Holland, J.J. in The Rhabdoviruses (ed. Wagner, R.R.) 297-360 (Springer Science+Business Media, Nwe York, 1987).
10. Perrault, J. in Initiation Signal in Viral Gene Expression (ed. Shatkin, A.J.) 151-207 (Springer-Verlag, Berlin, 1981).

11. Li, D. et al. Defective interfering viral particles in acute dengue infections. *PLoS One* **6**, e19447 (2011).
12. Saira, K. et al. Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. *J Virol* **87**, 8064-74 (2013).
13. Sun, Y. et al. Immunostimulatory Defective Viral Genomes from Respiratory Syncytial Virus Promote a Strong Innate Antiviral Response during Infection in Mice and Humans. *PLoS Pathog* **11**, e1005122 (2015).
14. Tapia, K. et al. Defective viral genomes arising in vivo provide critical danger signals for the triggering of lung antiviral immunity. *PLoS Pathog* **9**, e1003703 (2013).
15. Vasilijevic, J. et al. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. *PLoS Pathog* **13**, e1006650 (2017).
16. Strahle, L., Garcin, D. & Kolakofsky, D. Sendai virus defective-interfering genomes and the activation of interferon-beta. *Virology* **351**, 101-11 (2006).
17. Whistler, T., Bellini, W.J. & Rota, P.A. Generation of defective interfering particles by two vaccine strains of measles virus. *Virology* **220**, 480-4 (1996).
18. Gould, P.S., Easton, A.J. & Dimmock, N.J. Live Attenuated Influenza Vaccine contains Substantial and Unexpected Amounts of Defective Viral Genomic RNA. *Viruses* **9** (2017).
19. Xu, J. et al. Replication defective viral genomes exploit a cellular pro-survival mechanism to establish paramyxovirus persistence. *Nat Commun* **8**, 799 (2017).
20. Calain, P., Roux, L. & Kolakofsky, D. Defective interfering genomes and Ebola virus persistence. *Lancet* **388**, 659-60 (2016).
21. Frensing, T. Defective interfering viruses and their impact on vaccines and viral vectors. *Biotechnol J* **10**, 681-9 (2015).
22. Pathak, K.B. & Nagy, P.D. Defective Interfering RNAs: Foes of Viruses and Friends of Virologists. *Viruses* **1**, 895-919 (2009).
23. Janda, J.M., Davis, A.R., Nayak, D.P. & De, B.K. Diversity and generation of defective interfering influenza virus particles. *Virology* **95**, 48-58 (1979).
24. Nayak, D.P., Chambers, T.M. & Akkina, R.K. Defective-interfering (DI) RNAs of influenza viruses: origin, structure, expression, and interference. *Curr Top Microbiol Immunol* **114**, 103-51 (1985).
25. Leppert, M., Kort, L. & Kolakofsky, D. Further characterization of Sendai virus DI-RNAs: a model for their generation. *Cell* **12**, 539-52 (1977).
26. Pfaller, C.K. et al. Measles Virus Defective Interfering RNAs Are Generated Frequently and Early in the Absence of C Protein and Can Be Destabilized by Adenosine Deaminase Acting on RNA-1-Like Hypermutations. *J Virol* **89**, 7735-47 (2015).
27. Malykhina, O., Yednak, M.A., Collins, P.L., Olivo, P.D. & Peeples, M.E. A respiratory syncytial virus replicon that is noncytotoxic and capable of long-term foreign gene expression. *J Virol* **85**, 4792-801 (2011).
28. Calain, P., Monroe, M.C. & Nichol, S.T. Ebola virus defective interfering particles and persistent infection. *Virology* **262**, 114-28 (1999).
29. Fearns, R. & Collins, P.L. Role of the M2-1 transcription antitermination protein of respiratory syncytial virus in sequential transcription. *J Virol* **73**, 5852-64 (1999).
30. Noton, S.L. & Fearns, R. The first two nucleotides of the respiratory syncytial virus antigenome RNA replication product can be selected independently of the promoter terminus. *RNA* **17**, 1895-906 (2011).

31. Fearns, R., Peeples, M.E. & Collins, P.L. Increased expression of the N protein of respiratory syncytial virus stimulates minigenome replication but does not alter the balance between the synthesis of mRNA and antigenome. *Virology* **236**, 188-201 (1997).
32. Noton, S.L., Cowton, V.M., Zack, C.R., McGivern, D.R. & Fearns, R. Evidence that the polymerase of respiratory syncytial virus initiates RNA replication in a nontemplated fashion. *Proc Natl Acad Sci U S A* **107**, 10226-31 (2010).
33. Yoshida, A. et al. A Single Amino Acid Substitution within the Paramyxovirus Sendai Virus Nucleoprotein Is a Critical Determinant for Production of Interferon-Beta-Inducing Copyback-Type Defective Interfering Genomes. *J Virol* **92** (2018).
34. Fearns, R. & Plemper, R.K. Polymerases of paramyxoviruses and pneumoviruses. *Virus Res* **234**, 87-102 (2017).
35. Fearns, R. & Collins, P.L. Model for polymerase access to the overlapped L gene of respiratory syncytial virus. *J Virol* **73**, 388-97 (1999).
36. Kuo, L., Fearns, R. & Collins, P.L. Analysis of the gene start and gene end signals of human respiratory syncytial virus: quasi-templated initiation at position 1 of the encoded mRNA. *J Virol* **71**, 4944-53 (1997).
37. Calain, P. & Roux, L. The rule of six, a basic feature for efficient replication of Sendai virus defective interfering RNA. *J Virol* **67**, 4822-30 (1993).
38. Kolakofsky, D. et al. Paramyxovirus RNA synthesis and the requirement for hexamer genome length: the rule of six revisited. *J Virol* **72**, 891-9 (1998).
39. Tawar, R.G. et al. Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. *Science* **326**, 1279-83 (2009).
40. Samal, S.K. & Collins, P.L. RNA replication by a respiratory syncytial virus RNA analog does not obey the rule of six and retains a nonviral trinucleotide extension at the leader end. *J Virol* **70**, 5075-82 (1996).
41. Fodor, E., Mingay, L.J., Crow, M., Deng, T. & Brownlee, G.G. A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase promotes the generation of defective interfering RNAs. *J Virol* **77**, 5017-20 (2003).
42. Odagiri, T. & Tobita, K. Mutation in NS2, a nonstructural protein of influenza A virus, extragenically causes aberrant replication and expression of the PA gene and leads to generation of defective interfering particles. *Proc Natl Acad Sci U S A* **87**, 5988-92 (1990).
43. Kang, C.Y., Weide, L.G. & Tischfield, J.A. Suppression of vesicular stomatitis virus defective intefering particle generation by a function(s) associated with human chromosome 16. *J Virol* **40**, 946-52 (1981).
44. Kingsbury, D.W. & Portner, A. On the genesis of incomplete Sendai virions. *Virology* **42**, 872-9 (1970).
45. Yount, J.S., Kraus, T.A., Horvath, C.M., Moran, T.M. & Lopez, C.B. A novel role for viral-defective interfering particles in enhancing dendritic cell maturation. *J Immunol* **177**, 4503-13 (2006).
46. Sun, Y. & López, C.B. Preparation of Respiratory Syncytial Virus with High or Low Content of Defective Viral Particles and Their Purification from Viral Stocks. *Bio-protocol* **6** (2016).
47. Hotard, A.L. et al. A stabilized respiratory syncytial virus reverse genetics system amenable to recombination-mediated mutagenesis. *Virology* **434**, 129-36 (2012).

48. Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag, New York, 2016).
49. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
50. Shcherbo, D. et al. Far-red fluorescent tags for protein imaging in living tissues. *Biochem J* **418**, 567-74 (2009).

ACKNOWLEDGMENTS

This work was supported by the US National Institutes of Health National Institute of Allergy and Infectious Diseases (NIH AI083284 and AI127832 to C.B.L). We thank Dan Beiting for the support with RNA-seq and GEO submission. The non-human primate work completed at the NIAID Integrated Research Facility was supported by NIAID Division of Intramural Research and NIAID Division of Clinical Research, and was performed under Battelle Memorial Institute Contract (No. HHSN272200700016I). We would also like to thank all Comparative Medicine staff at the NIAID/IRF for the NHP animal care and sample collection; with special thanks to Anna N. Honko, Lisa E. Hensley, Nicholas Oberlander, Russell Byrum, Dan Ragland, Marisa St Claire.

Conceived experiments: Y.S., and C.B.L.; Developed methodology: Y.S., E.J.K, D.A., and G.R.G; Performed experiments, collected and analyzed data: Y.S., E.J.K, and L.J.T; EBOV infection: D.M.G, R.S.B; Provided reagents and resources: G.R.G and J.H.C; Wrote the original draft: Y.S., D.A., and C.B.L.; Supervised research activities: C.B.L and J.H.C.

FIGURES AND FIGURE LEGENDS

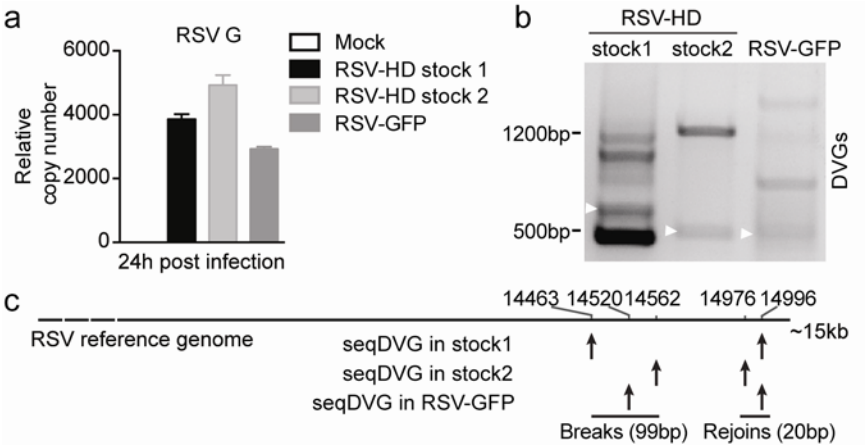

Figure 1. cbDVGs generated from three different RSV stocks *in vitro* have similar rejoin points. HEp2 cells were infected with RSV-HD stock 1, 2 (prepared differently from the same parental virus stock), and RSV-GFP (originated from an RSV infectious clone containing GFP gene) at an MOI of 1 $TCID_{50}$/cell. Infected cells were harvested at 24 h post infection. (a) Virus replication was assessed by qPCR using primers within *RSV G*. Error bars indicate standard error of mean (SEM) of the data from three independent experiments. Gene expression is shown as copy number relative to the housekeeping gene *GAPDH*. (b) DVGs generated from each stock were detected by RT-PCR. Base pair size references are indicated in the gel. Bands across three virus stocks labeled by white arrows represent an example of DVGs identified with close rejoin points after sequencing. Actual cbDVG sequences are indicated in Fig. S1. Location of their breaks and rejoins is depicted in (c). Numbers at the bottom of (c) indicate the nucleotide length of break and rejoin regions respectively.

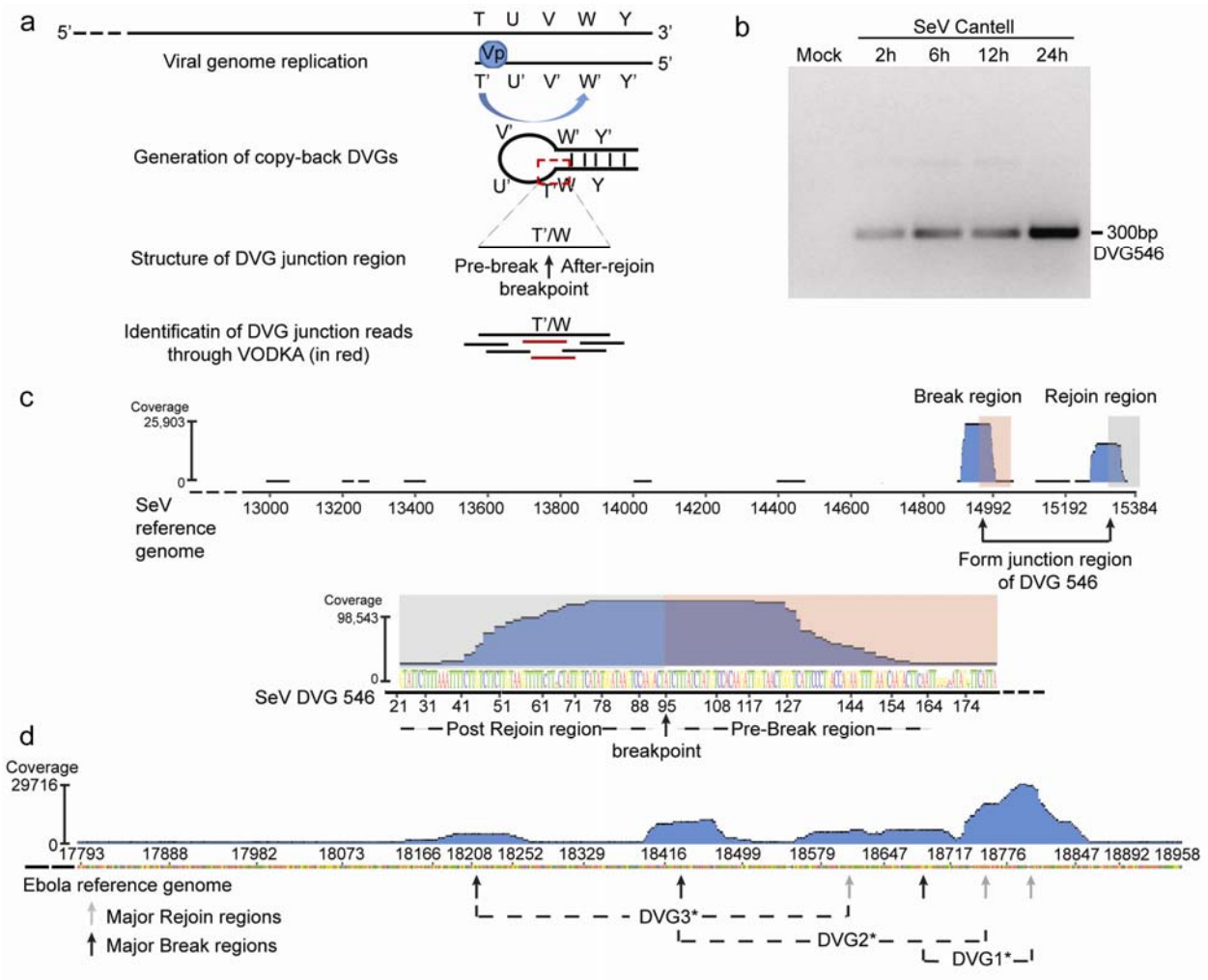

Figure 2. Validation of VODKA using samples from cells infected with SeV strain Cantell and Ebola infected macaques. (a) Schematic representation of cbDVG generation from negative strand viruses and general principle behind VODKA identification of cbDVGs. The red dashed square marks the unique junction region that distinguishes DVGs from full-length viral genomes that is used by VODKA to identify cbDVGs. (b) A549 cells were mock infected or infected with SeV Cantell at an MOI of 1.5 $TCID_{50}$/cell and harvested at 2, 6, 12, and 24h post infection followed by detection of cbDVGs by RT-PCR using SeV DI1 and gSeV DI1 (Table S2). (c) Alignment of SeV cbDVG junction reads obtained from VODKA. Top figure shows the alignment of the reads to the SeV antisense genome. Blue histogram shows synopsis of total

coverage at any given position of SeV reference genome or DVG-546. There are two major places of the genome covered, which correspond to the predicted break and rejoin regions depicted in (a). Bottom figure shows the alignment of junction reads with the sequence of DVG-546. The number on the right side of the graph represents the total reads at the position with the highest coverage. Colored nucleotides in the bottom figure represent the consensus sequences of cbDVG junction reads aligned to this region. The size of the letter represents the level of conservation at the site and if there were mutations at a certain position the most representative nucleotide was listed on top. The grey boxed area indicated the region after rejoin and the pink boxed area indicated the DVG region before break. (d) RNA-seq dataset on eight cbDVG-positive macaques infected with 1000 pfu/animal of Ebola virus. The dataset was screened for cbDVGs using VODKA. The VODKA output was then aligned to reference genome NCBIKP096421.1 indicating 3 major break regions and 3 rejoin regions of Ebola cbDVGs in infected macaques, representing 3 major cbDVGs. Black arrows indicate break regions and grey arrows indicate rejoin regions. The detailed information of cbDVGs labeled by asterisk can be found in Table 1.

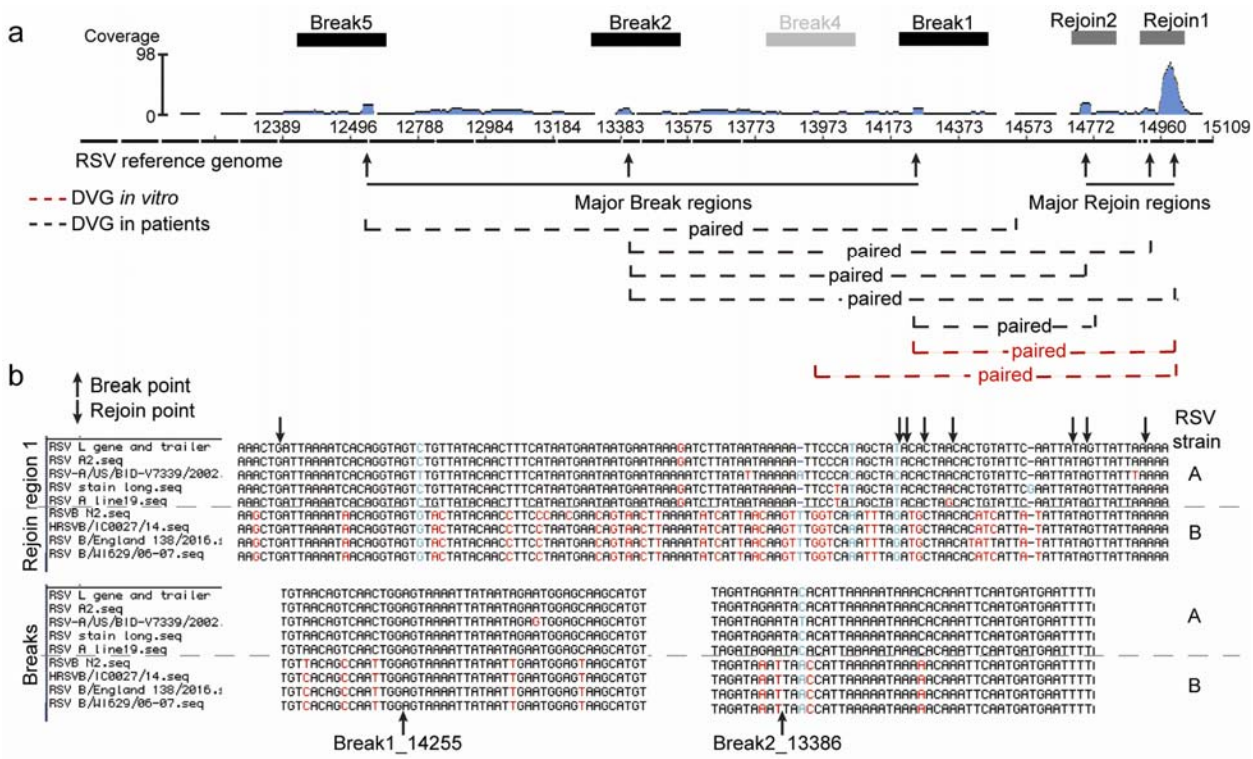

Figure 3. Identification of cbDVG junction regions from nasal secretions of RSV-positive pediatric patients. Two specimens from banked RSV A-positive pediatric patients were selected for RNA-seq due to their high content of DVGs. RNA was extracted from nasal sections followed by RNA-seq. All Seq reads that passed the quality threshold were analyzed using the VODKA pipeline to screen for potential cbDVG junctions. (a) Alignment of DVG junction reads to the RSV reference genome NCBIKJ672447. Five major break and rejoin regions (Beak1, 2, 5 and Rejoin1, 2) present in both clinical samples are indicated by rectangles (black for breaks and grey for rejoins). Additionally, one break (Break4) indicated by light grey was identified only in cbDVGs generated during *in vitro* infection. Break and rejoin regions forming five major cbDVGs found in both patients are indicated with black dashed lines. The break and rejoin regions for *in vitro* cbDVGs are indicated by red dashed lines. Blue histogram shows synopsis of total coverage at any given position of the RSV reference genome (illustrated by the y-axis). The

number on the right side of the graph represents the total reads at the position with highest coverage. (b) Alignment of the entire Rejoin1 region, as well as two break regions (Break1 and Break2) from 9 different isolates of RSV strains A and B. Arrows pointing down indicate all the major rejoin points identified in Rejoin1 including both clinical samples and *in vitro* infection; their location is specified in Table S1. Arrows pointing up indicate one representative break point within each break region; their genomic location is indicated below the corresponding arrow.

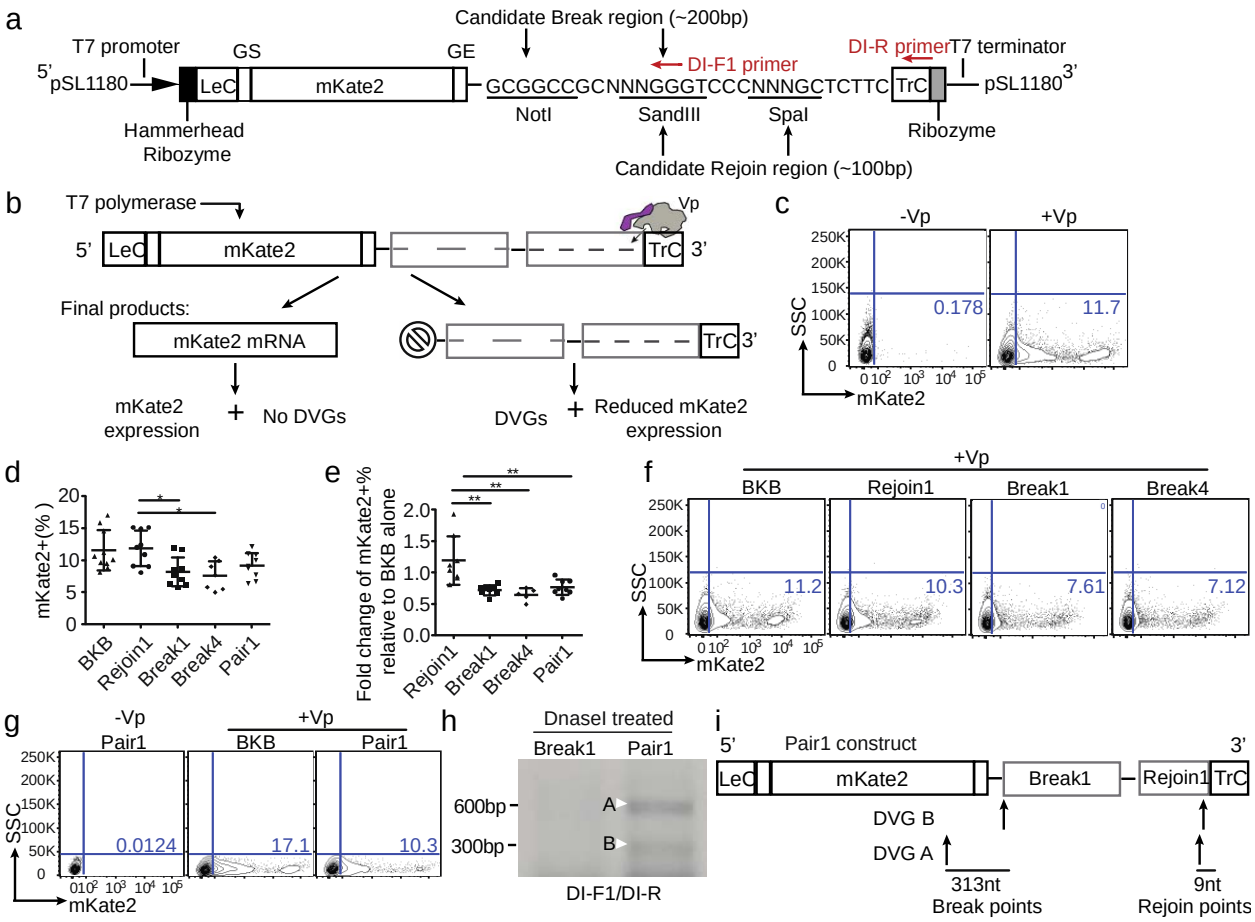

Figure 4. Identification of specific RSV genomic regions responsible for cbDVG generation. (a) Schematic representation of the RSV minigenome backbone construction. Going from 5' to 3': T7 promoter, hammerhead ribozyme, complementary sense RSV leader sequence (LeC), NS1 gene start sequence, NS1 non-coding sequence, gene encoding mKate2[50], L noncoding sequence, L gene end sequence, complementary sense trailer sequence (TrC), hepatitis delta virus ribozyme, and T7 terminator. All constructs were generated in a pSl1180 vector and included three restriction enzyme sites that were used for ligating break and rejoin candidate regions as indicated. The length of all minigenome constructs after T7 promoter and before T7 terminator obeyed the rule of six. Red arrows indicate the primers that were used for detecting cbDVGs by PCR. One primer was within the viral trailer and the other primer was designed against the 3'

end of the break region. (b) Workflow of the RSV minigenome system. BRS-T7 cells were transfected with expression vectors encoding the polymerase genes M2-1, L, NP, P, and the minigenome construct to be tested. The minigenome was then transcribed by the T7 polymerase generating a template for amplification by the viral polymerase. If the paired junction region was not functional, the viral polymerase continued to elongate from 3' TrC to the 5' LeC, exposing the promoter in Le for transcription leading to mKate2 expression. If break and rejoin sequences were functional, the viral polymerase recognized those signals resulting in its dropping off from the template, generation of cbDVG-like fragments, and reduction of mKate2 expression. Since DVG junctions are split into two non-continuous regions (break and rejoin) in the minigenome plasmid, detection of the end-to-end DVG junction regions via DI-RT-PCR is a second indication of a productive junction region. (c) BSR-T7 cells were transfected with BKB either with or without the polymerase expression plasmids. One representative plot of mKate2 expression determined by flow cytometry is shown. (d-g) BSR-T7 cells were co-transfected with all 4 helper plasmids as well as BKB or constructs with candidate signal sequences. mKate2 expression was measured by flow cytometry. The quantification of all repeats is shown both as a percentage (F test $P=0.0028$) (d) and as a fold change expression over BKB (F test $P<0.0001$) (e). Representative flow plots are shown in F and G. $*p<0.05$, $**p<0.01$ by one-way ANOVA Bonferroni's multiple comparison test, n=8, mean±SEM, Bartlett's test is used for examination of equal variance. Flow gating strategy is shown in Fig. S4a. (h-i) cbDVG-like fragments formed in Pair1 were detected by RT-PCR using DI-F1 and DI-R primer sets (Table S2). Bands labeled with a white arrow were confirmed to be cbDVG-like by conventional sequencing. Their sequence is indicated in Fig S4e. The location and the sequence length between the break and rejoin points are indicated in panel (i).

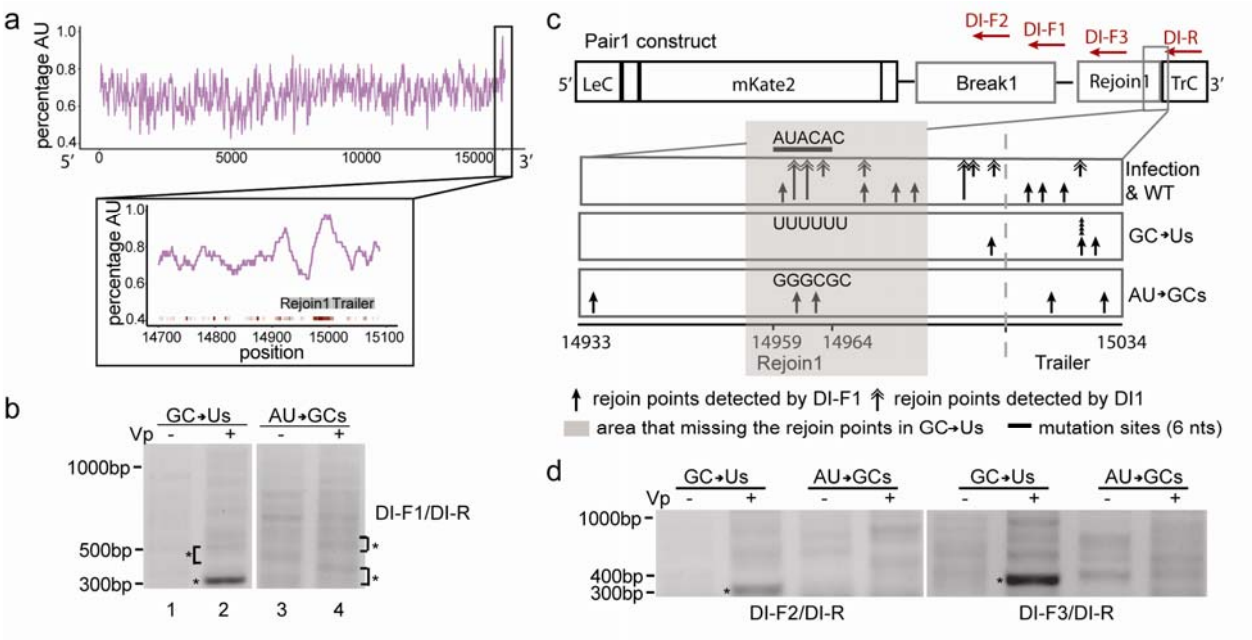

Figure 5. The sequence of Rejoin1 determines where cbDVG form. (a) AU-content (top, purple) over the entire genome and major rejoin region (inset, purple). AU-content is plotted centered on a 40 base-pair sliding window in both plots. Red ticks in the inset represent rejoins at each position starting at base 14700, with a proportionally darker color representing increased rejoin frequency at a given position. Grey box indicated the position of Rejoin1 and Trailer. (b) cbDVG-like fragments generated in mutant GC->Us and mutant AU->GCs were detected by RT-PCR using the DI-F1 and DI-R primer set. Confirmed cbDVG-like fragments are indicated with brackets labeled by asterisks next to the gel. (c) Schematic summary of all identified rejoin points within positions 14933-15034 based on reference genome NCBI KJ672447. This region includes part of Trailer. The indicated rejoin points included cbDVGs generated from infection *in vitro* and in patients, as well as from transfection of minigenome system using WT Pair1 and Pair1 mutants. The locations of rejoin points found in WT-Pair1 and normal infection were compared to the Pair1 mutants. ↑indicates the rejoin points from cbDVGs detected by DI-F1 and DI-R primer set. ⇑indicates the rejoin points from cbDVGs detected by DI1 and DI-R primer set

used for cbDVG detection from infected samples. Longer arrows indicate that there were more than two cbDVGs rejoining at the same position. ▲on top of the arrow indicates the times of repitition of the same cbDVG observed at this location. Red arrows indicate the positions of different primer sets used for detecting cbDVGs in minigenome system. They are composed of the identical reverse primer, DI-R, and three different forward primers, DI-F1, DI-F2, and DI-F3. Thick black line marks the 6 nts mutated in mutant GC->Us and mutant AU->GCs. The position range of the mutations is indicated at the bottom of the figure and the actual RNA template sequence at the mutation sites in WT and two mutants is indicated on top of the arrows within each individual construct box. Grey shade marks the area that mutant GC->Us did not have the rejoin points. (d) cbDVG-like fragments generated in mutant GC->Us and mutant AU->GCs were detected by DI-F2/DI-R and DI-F3/DI-R through DVG specific RT-PCR. The DVG-like bands labeled by asterisks were confirmed to contain the same cbDVG-like fragment detected by DI-F1/DI-R in B (only band labeled by asterisk). Sequences of all the confirmed cbDVG-like fragments are shown in Fig S5a.

Table 1. Major DVG junctions of RSV and Ebola identified by VODKA

T'_W	Total count	Junction reads full sequence	Sample ID	Reference genome	Pair of break and rejoin
		RSV			
12529_14474	29	TACTACATTAAATAC:::CATATTAACAATAGTATGG	H4 & H73	KJ672447	Break5
13385(-1)_14874(+1)	25	TACCTGTGGCTTTTAA:::ATATACATTAAAAATAAA	H4 & H73	KJ672447	Break2+rejoin1
13386_14750	16	ATTTAAAACATGGTT:::TATACATTGAAAAATAA	H4 & H73	KJ672447	Break2+rejoin2
13384_14974	15	TTAAATAACTATAAT:::AATATACATAAAAAT	H73	KJ672447	Break2+rejoin1
14210(+5)_14818	9	AATTCAAATAGATAAGGA:::AGTCTTTTTGTTTGT	H4 & H73	KJ672447	Break1+rejoin2
14255_14957	N/A	AATACAGTGTTAGTG:::TGTAAGAAAGTGCAA	RSV A2*	AF035006	Break1+rejoin1
14385_14957	N/A	AGGATTAGTTCAC:::TATTAAAAACTTATG	RSV A2*	AF035006	Break1+rejoin1
13877_14956	N/A	ATACAGTGTTAGTGT:::AGATTCAATTTTGTATTTAG	RSV A2*	AF035006	Break4+rejoin1
13919_14963	N/A	TTGAATACAGTG:::TTAGTATAGAGTATATT	RSV A2*	AF035006	Break4+rejoin1
		Ebola			
18428(+/-3)_18767(+/-2)	49190	AAAGAAGGTATTTCTGATTTTACAGTCA:::TATTACTGCTGCAATGAGTCTAACGCAACATAATAAACTCCGCACTCT	C G B	KP096421	DVG2
18701(-6)_18804(+3)	37499	CGAAGAATATTATCTAGTCTG:::AACTTTAACATTTTTGTCTAGTAAGCTACTATTCCATTCAGAATGATAAAATTAA	C G B	KP096421	DVG1
18205(+2)_18588(-2)	12502	GACAAATTAAAGCAATGACAATTATAAAAAACTTAATCTTAAAGATCAGTTTTGTTT:::TACAGGTTCGATTGAATA	F	KP096421	DVG3

*RSV A2 is referring the RSV HD batch 1

::: breakpoint of DVG junction reads.