Analysis of Twist knock-out in *Nematostella vectensis*

Abstract

Twist, a basic helix loop helix (bHLH) transcription factor fulfils versatile functions throughout metazoans. Delineated functions range from initiation of gastrulation in Drosophila melongaster to the promotion of epithelial-mesenchymal transition (EMT) in human cancers. Albeit the suppression of twist in *Nematostella vectensis* doesn't influence early developmental stages, later phenotypic consequences condense in a yet undescribed 'bubble-formation' around the head region. Here, by integrating both bulk and single-cell RNA-seq (scRNA) on Twist-KO animals, we aim to elucidate its actual molecular position within the gene-regulatory network of *N. vectensis*. We show that *NvTwist*$^{-/-}$ animals show strongest upregulation of several TFs belonging to homeobox- and zinc-finger proteins. Especially, *ZN596-like* showed persistently upregulation in mutant animals, while its expression was completely lacking in wild-type animals. Moreover, 'bubble-forming' animals are depleted in Wnt signaling pathway proteins, suggesting major deviations from normal cell fate balances. Furthermore, the analysis of scRNA-seq data revealed a new cell-cluster identity which is exclusively found in mutant animals. Collectively, these results bolster the central role of twist in *N. vectensis,* albeit further studies are required to elucidate its direct molecular targets and further, the molecular circuits which are responsible for generating the new cell-cluster as found from scRNA-seq analysis.

University of Vienna: Evolutionary Systems Biology: U066220
Lab Rotation II
Christoph Kreitzer, 01345159

Department of Neurosciences and Developmental Biology
Division of Molecular Evolution and Development, Univ.-Prof. Dr. Ulrich Technau

Introduction

The starlet sea anemone *Nematostella vecentis* gained scientific attention over the last years as a cnidarian model organism in developmental and evolutionary studies (Pennisi, 2007). Cnideria depicts a phylogenetic sister group to bilateria, and hence *N. vectensis* displays a well-suited organism to study the ancestral state of features in the shared ancestor with bilaterians. Besides, its cultivation and life cycle are highly controllable and genetic tools for genomic interventions are well established. Nematostella is made up of only two cell layers that are directly derived from the embryonic endo- and ectoderm, separated by a layer of extracellular matrix called mesoglea (Jahnel, Walzl and Technau, 2014).

Twist belongs to the basic helix loop helix (bHLH) transcription factor (TF) family, which is found throughout metazoans. bHLH proteins bind as dimers to the consensus hexanucleotide sequence E-box (5'-CANNTG-3') which mediates the interaction with DNA (Castanon and Baylies, 2002). Twist was first identified in Drosophila melongaster, as a gene crucial for proper gastrulation and mesoderm formation. In the fly embryo, Twist continues to play additional roles, allocating mesodermal cells into the body wall muscle fate and patterning a subset of these muscles. Twist is also required for proper differentiation of the adult musculature. Twist homologues have been identified in a great variety of organisms, which span the phylogenetic tree (Sandmann *et al.*, 2007).

In mammals, for example, twist1 is active in the mesenchymal lineage at various points of development. It maintains chondrocytes and adipocytes in an immature state and promotes osteoblast differentiation. Moreover, the suppression of Twist expression in highly metastatic mammary carcinoma cells specifically inhibits their ability to metastasize from the mammary gland to the lung. Ectopic expression of Twist results in loss of E-cadherin-mediated cell-cell adhesion, activation of mesenchymal markers, and induction of cell motility, suggesting that Twist contributes to metastasis by promoting an epithelial-mesenchymal transition (EMT) (Yang *et al.*, 2004).

In recent years, single-cell RNA sequencing (scRNA-seq) has significantly advanced our knowledge of biological systems (Briggs *et al.*, 2018). Researchers have been able to both study the cellular heterogeneity of zebrafish, frogs and planaria and further discover previously obscured cellular populations. The great potential of this technology has motivated computational biologists to develop a range of analysis tools (e.g. Rostom et al. 2017) to exploit the great potential of this data resource.

Given the versatile functions of twist, paired with the application of scRNA-seq technology, we aim to decipher functional consequences of a twist knock-out in *N. vectensis*. NvTwist was knocked out using a CRISPR/Cas9 based approach. Animals used for this study, showed obvious defects in secondary tentacle formation and occasionally developed a ‘bubble-like’ structure (Figure 1). However, NvTwist mutant animals are viable, grow to sexual maturity and are able to produce offspring.

Here, by integrating bulk and scRNA-seq data, we aim to identify direct and indirect targets of NvTwist. Specifically, we want to elucidate the genotypic landscape of mutant animals in order to explain the origin of this novel ‘bubble-like’ phenotype.

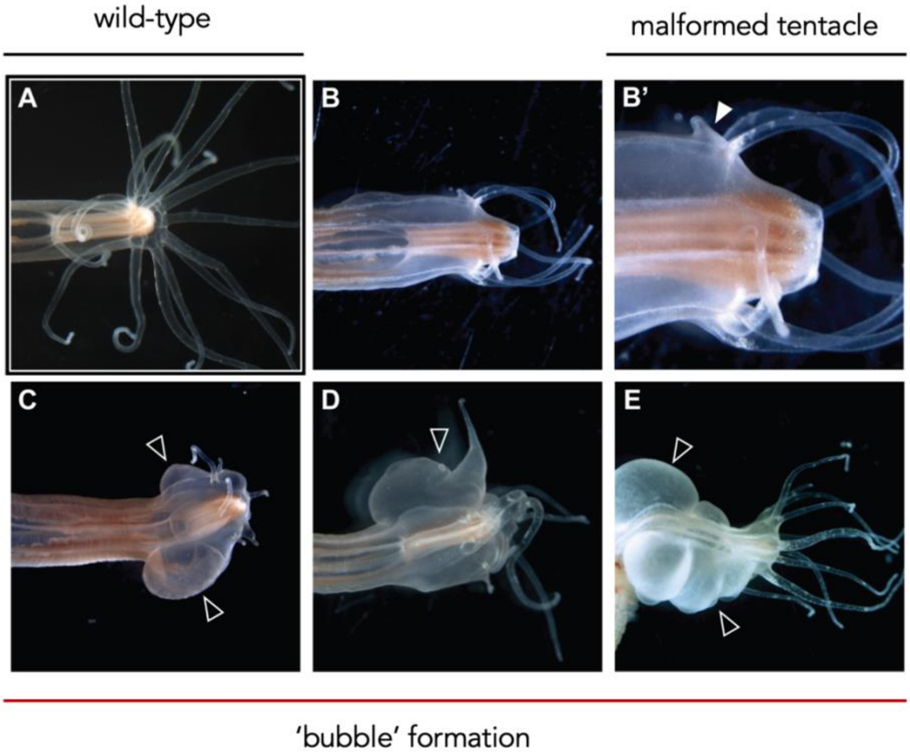


Figure 1: Twist knockout mutants show tentacle mis-patterning and develop outgrowths in their body walls. Comparison of Nvtwist knockout subadult polyps (B-E) with a wildtype animal (A). Images are at varying scales but as a reference point the pharynx is proportionate to overall body size in mutants as in the wildtype. Arrowhead in (B’) indicates an example for a malformed tentacle. Note also the irregular spacing of tentacles. Arrowheads in (C-E) indicate examples of “bubble” formation (Figure, courtesy of Julia Hagauer).

Material and Methods

Study design and data availability

The CRISPR-Cas9 RNA-guided system was chosen to induce a genomic deletion at the *NvTwist* locus (please consult Patricio Ferrer, M.Sc. for further details). Thereafter, two major data resources were generated which were used for all subsequent investigations. Firstly, bulk RNA- seq data of *N. vectensis* for five distinct geno-/phenotypic combinations (Table 1) were made available. Moreover, four single-cell RNA-seq (scRNA-seq) libraries were established where ensuing sequencing results were equally made disposable for analysis. A detailed description of the data generation can be found elsewhere (Hagauer, 2020).

Table 1: Geno-/ and phenotypic characteristics of animals used for the underlying genomic survey. Bulk RNA- seq was conducted in triplicates, whereas one biological replicate roughly consisted of 4 individual animals. Note that phenotypic specifications are self-established and may not reflect common terms used in literature (e.g., ‘Bubble’). dpf = days post fertilization, wt = wild-type

Phenotype: Description	**Genotype**	**Analysis conducted**	**Comment**
‘Bubble’	*NvTwist*$^{-/-}$	Bulk RNA seq.	‘Bubble’ formation (Figure 1)
‘Twist-Head’	*NvTwist*$^{-/-}$	Bulk RNA seq.	*NvTwist*$^{-/-}$ without ‘bubble-formation’
‘Twist-4d’	*NvTwist*$^{-/-}$	Bulk RNA seq.	RNA-seq 4 dpf.
‘Wildtype-Head’	*NvTwist*$^{+/+}$	Bulk RNA seq.	RNA-seq of adult wt-animals
‘Wildtype-4d’	*NvTwist*$^{+/+}$	Bulk RNA seq.	RNA-seq 4 dpf. of wt-animals
‘Pharynx-Mutant’	*NvTwist*$^{-/-}$	scRNA-Seq.	adult animals
‘Pharynx-Control’	*NvTwist*$^{+/+}$	scRNA-Seq.	Adult animals (equal developmental stage as ‘Pharynx-mutant’
‘primary-Polyp-mutant’	*NvTwist*$^{-/-}$	scRNA-Seq.	adult animals
‘primary-Polyp-control’	*NvTwist*$^{+/+}$	scRNA-Seq.	adult animals (same developmental stage as mutants)

Validation of *NvTwist* knock-out

Bulk RNA-seq data was subject to various quality control examinations (for details, please refer to Juan Daniel Montenegro Cabrera, PhD). QC-true sequencing reads were trimmed for sequencing adaptors and further aligned to the current *N. vectensis* reference genome (current version: /scratch/jmontenegro/nvectensis/data/refs/nv_dovetail_4_gapped_chroms.final.fasta.gz) using bowtie (Langmead and Salzberg, 2012). Variants were called using bcftools (Li, 2011), where a quality metric of -q 20 were chosen with all other parameters running in default mode. Only variants with an alternative allele frequency >0.2 were kept for further consideration. Note that bcftools was run chromosome-wise, meaning that all 15 samples (5 phenotypes in triplicate) were run parallel on one chromosome at a time, hence a reliable variant-allele frequency estimation was rendered possible. Before turning into differential gene expression (DGE) analysis we wanted to confirm the deletion in *NvTwist* and further, whether the applied CRISPR/Cas9 approach induced any genomic off-targets. We concentrated on INDEL mutations called previously and introduced several filtering rules to distill variants which *bona fide* may impact protein functions and hence causes phenotypic consequences. Among all variants, we filtered for

i) nucleotide variation which causes a frameshift mutation

ii) variants which are homozygous in at least 67% of mutant animals AND/OR wild-type animals (9 mutant vs 6 wt-animals)

Differential gene expression analysis on bulk RNA-seq data

Before turning to differential gene expression (DGE) analysis of bulk RNA-seq data, we generated a count matrix by applying featureCounts (Liao, Smyth and Shi, 2014). featureCounts was run on the exon level with all the remaining default parameters unchanged. The current Nv2-annotation file [/scratch/jmontenegro/nvectensis/results/annotation/tcs.gtf; May 2021 generated] served as reference. DGE analysis was conducted with the DESeq2 package (Love, Huber and Anders, 2014) following standard protocols and default parameters. A negative binomial model and Wald statistics was applied to the data after the size factors and the dispersion of the data was estimated. However, to avoid zero inflation, we required at least 10 reads in every replicate and at least in two libraries to keep the feature. Descriptive summary statistics were performed on variance-stabilized-transformed (vst) expression data (Anders and Huber, 2010). Subsequent results were filtered for displaying only features with an absolute log2FC greater than 1 (i.e., abs(log2FC) > 1 & p.adjust < 0.1).

single-cell RNA-seq (scRNA-seq) analysis

Single cell suspensions were established for four different animals (see Table 1) whereof scRNA-seq analysis was conducted. For technical details please refer to Alison Cole, PhD or Patricio Ferrer, M.Sc. scRNA-seq data was processed using cellranger (Zheng et al., 2017). Among the key steps involved in the processing pipeline are i) the extraction of cell barcodes, ii) alignment of the reads to the reference genome (/scratch/jmontenegro/nvectensis/data/refs/nv_dovetail_4_gapped_chroms.final.fasta.gz) using STAR, iii) count the unique molecular identifier (UMIs) by cell and gene (feature) and finally v) get the gene-barcode matrix.

Respective feature-matrices were transferred to the local machine and subsequently analyzed using Seurat (Hao *et al.*, 2021).

The standard Seurat workflow takes raw single-cell expression data and aims to find clusters withing the data. This process consists of data normalization and variable feature selection, data scaling, a PCA on variable features, construction of a shared-nearest-neighbors graph, and clustering using a modularity optimizer. Finally, we used a UMAP to visualize our clusters in a two-dimensional space.

The normalization was done with a scaling factor of 5,000. FindVariableFeatures() was restricted to 2,000 features. After applying a principial components analysis (PCA) on variable features found, we conducted every subsequent analysis on 18 PCAs (see Results and Interpretation section). Seurat v3 applies a graph-based clustering approach. The distance metric which drives the clustering analysis is based on the previously identified PCs. To cluster the cells, Seurat applied a modularity optimization technique (i.e., Louvain algorithm) to iteratively group cells together. The FindClusters() functions implements this procedure and contains a resolution parameter that sets the 'granularity' of the downstream clustering. We decided on using a resolution parameter of 0.8 (see Results and Interpretation). Further details and scripts are readily available at https://github.com/chris-kreitzer/Twist/tree/main/Scripts

Results and Interpretation

NvTwist was knocked out in *N. vectensis* via a CRISPR-Cas9 RNA-guided system. A 5-bp deletion, specifically on transcript_id NV2.10864.1, exon 1, was confirmed from bulk RNA-seq data in 9 animals (Table 1 & Table 2). Besides obligatory defects in secondary tentacle formation, NvTwist mutant animals occasionally showed a distinct phenotype that we will further coin ‘bubble’, or ‘bubble-formation’. Detailed phenotypic descriptions can be obtained from Figure *1* and Hagauer, 2020 (Hagauer, 2020).

Besides the intended deletion found in *NvTwist*, we further observed three additional INDEL mutations among CRISPR/Cas9 engineered animals (Table 2).

Table 2: INDEL mutations found in bulk RNA-seq data among *NvTwist*$^{-/-}$ engineered animals. Note that gene_id NV2.10864 displays *NvTwist*

Position	Chromosome	Ref	Alt	Annotation
1059227	chr2	CCCCGAAC	CC	gene_id NV2.10722; transcript_id NV2.10722.1; exon_number 1;
2358816	chr2	GGTAACGT	GGT	gene_id NV2.10864; transcript_id NV2.10864.1; exon_number 1;
3292703	chr2	CTTCCATTT	CTTT	gene_id NV2.10979; transcript_id NV2.10979.1; exon_number 2;
15289229	chr8	CTGTGTGTGTGTGTGTGTGTGT	CTGTGTGTGTGTGTGT|CTGTGTGTGTGTGTGTGTGT	gene_id NV2.24244; transcript_id NV2.24244.1; exon_number 19;

While two INDEL mutations in genes NV2.10722 & NV2.24244 likely don’t render its actual protein function, one mutation affecting NV2.10979 most likely produces a truncated protein and hence displays a distinct CRISPR-Cas9 off-target. BLAST results of NV2.10979 suggest that this gene show similarities with radial spoke 3 protein in mouse and *Ciona intestinalis*. We further extracted the genomic sequence of NV2.10979 and aligned it with *NvTwist* and well as the guideRNA deployed for the CRISPR/Cas9 KO. (Supplementary figure *1*). Although there are 9 mismatches between the deployed 20-bp sgRNA and NV2.10979, a PAM motif exists and hence a theoretical genetic perturbation may have occurred. However, further *in vitro* studies are required to confirm whether or not NV2.10979 displays a true CRISPR/Cas9 off-target given the circumstances described above.

Additionally, we discovered four genetic variants that are specific to wild-type animals (Table *3*).

Table 3: Genetic variants found in bulk RNA-seq data among *NvTwist*$^{+/+}$ animals.

Position	Chromosome	Ref	Alt	Annotation
3790999	chr14	TGA	TA	gene_id NV2.8575; transcript_id NV2.8575.4; exon_number 8;
8153360	chr14	TCTGGG	T	gene_id NV2.8985; transcript_id NV2.8985.1; exon_number 1;

9419857	chr14	GATGACA	G	gene_id NV2.9107; transcript_id NV2.9107.1; exon_number 17;
3935654	chr2	GTGGCTTTATAAGGCCTTTGT	G	gene_id NV2.11056; transcript_id NV2.11056.1; exon_number 1;
5722483	chr3	GATATA	GATAATATA	gene_id NV2.13696; transcript_id NV2.13696.1; exon_number 8;

We suppose that those variants may have evolved *de novo* over many generations of cultivation. Albeit interesting, we won't concentrate on these variants for subsequent analysis. Rather a population genetics survey is required to study ramifications of these *bona fide de novo* evolved variants.

We performed differential gene expression (DGE) analysis on bulk RNA-seq data to decipher transcriptomic differences based on the underlying genotype and developmental stage (Table 1). Since data normalization displays a crucial step in DEG analysis, we compared the log-transformation, rlog (Love, Huber and Anders, 2014) and variance-stabilization-transformation (vst, Anders and Huber 2010). Concentrating on *NvTwist* quantities, we found no significant differences among applied normalization methods (Supplementary figure 2). Interestingly, however, *NvTwist* expression varies with developmental stage (Supplementary figure 3). *NvTwist* expression was lowest in the planula stage (i.e., 4d of development) and did not significantly vary in either wild-type ('WT4d') or mutant ('Twi4d') animals. This observation was further underpinned by the visualization of the sample-to-sample distances onto a 2D projection using in a principal components analysis (PCA) (Figure 2).

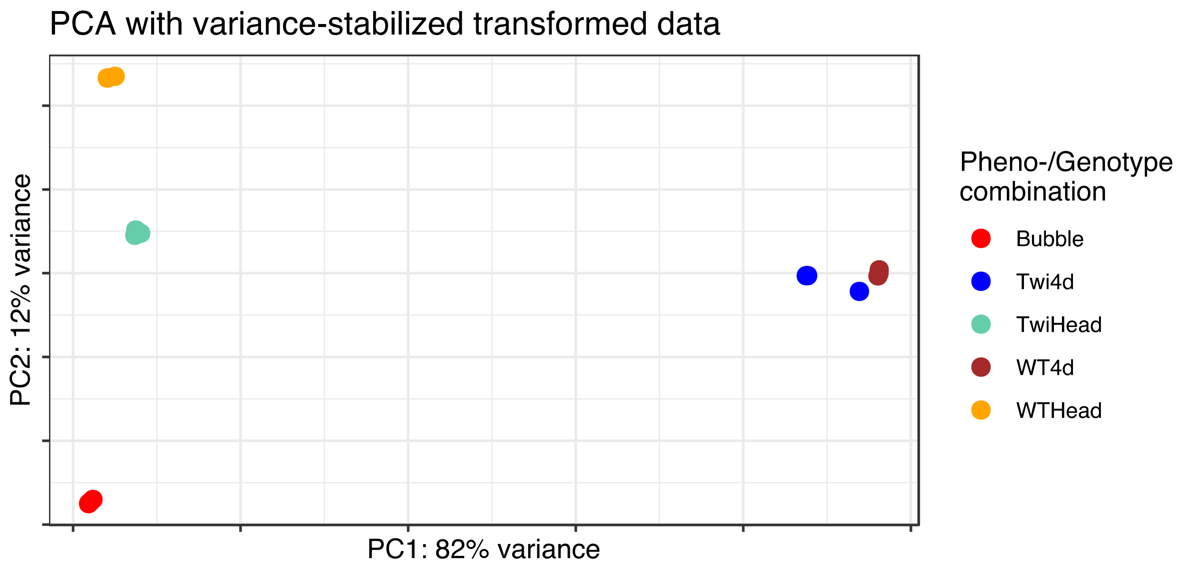


Figure 2: Visualization of sample-to-sample distances projected onto 2D using a principal components analysis (PCA). Different colors highlight animals with distinct geno-/phenotype features (see Table 1).

We further note the high quality of our bulk RNA-seq dataset as seen by the close clustering of the first two PCAs among biological replicates as well as the high percentage of variance which is explained by only two PCs.

DGE analysis on bulk RNA-seq again underpinned the similar gene expression profile in both the NvTwist-mutant and wt-animals at the early planula stage as seen in Figure 2. We measured that 2% of genes were upregulated in mutant, whereas only 190 (0.96%; total of 19,856 genes) were downregulated in wild-type animals (LFC > 1 & p.adj. < 0.01). NvTwist expression didn't significantly differ between the two compared genotypes, indicating that NvTwist may be expressed at later developmental stages.

NvTwist-mutant animals showed a striking upregulation in metabolic enzymes, such as lactase-like b precursor (NV2.18509) or monooxygenase (NV2.14000) as well as proteins associated with mesenchymal tissue (e.g., collagen alpha-1 (XXVII) chain B; NV2.6495 or tropomyosin alpha-3 chain; NV2.8895, see supplementary file). Moreover, mutant animals depicted an enrichment in several transcription factors, many of which show a direct DNA-interaction capacity. Examples include *ZN596-like*, a zinc-finger protein (Najafabadi *et al.*, 2015) and homeobox-proteins NvNKx2.2E, NvNKx2.2C and NvNKx2.2A (Bürglin and Affolter, 2016). Notably, *ZN596-like* was exclusively detected in mutant animals and showed consistent upregulation throughout developmental stage. 'Bubble-forming' mutant animals depicted an ~500-fold ($2^{8,84}$) change of *ZN596-like* compared to its wt-counterpart (Supplementary file).

On the contrary, wild-type animals were enriched for *HSF-like*, *SOX30-like* and *HMG2-like* encoding genes which fulfil TF-activities. HSF-like TFs are associated with the regulation of heat shock protein production (Morimoto, 1993). Other examples include Forkhead box protein L1 and Forkhead box protein B1 (NV2.13108 and NV2.13437, respectively) that were enriched in animals at the planula stage. This result is not unexpected, as FOX proteins play important roles in regulating the expression of genes involved in cell growth, proliferation, differentiation and are important to embryonic development.

Concentrating on the 'bubble-forming' phenotype, we saw strongest-upregulation of genes associated with mesenchymal factors when compared to genotypic identical animals without the phenotype. Specifically, chymotrypsin-like elastase family member (NV2.25575), chymotrypsin-C (NV2.25574) and several important TF-family members such as homeobox or NvSox2 proteins are enriched (Supplementary file). Moreover, we saw strongest upregulation of *FBP3-like5*, a NOTCH protein while Matrilin-3 (*MATN3-like2*) show great depletion. Interestingly, MATN3-like2 is enriched in twist-mutant animals without the bubble-formation, suggesting that this protein may specifically be suppressed in the bubble-phenotype. A detailed summary of all up- and downregulated genes is attached as supplementary file.

We performed single-cell RNA sequencing (scRNA-seq) of the head region of *NvTwist*$^{-/-}$ mutants and its wild-type counterpart. With this approach we aim to elucidate transcriptomic transitions at the single cell level based on different genotypes. Furthermore, this should allow us to decipher potential direct and indirect targets of *NvTwist* in specific cell populations.

Both, the preparation and the sequencing of the scRNA libraries yielded high quality data as seen with ~88% of reads confidently mapped to the genome (3,199 mean reads per cell and 218 median genes per cell, supplementary note). Note, however, that the 5-bp deletion found in the NvTwist transcript from bulkRNA seq was not detected at the single-cell level, however we found two SNPs in the promotor region, respectively (Supplementary table 1).

Raw single-cell count data obtained from cellranger (Zheng *et al.*, 2017), was subject to Seurat (Stuart *et al.*, 2019), a well-established processing pipeline for scRNA-seq data. Based on explorative data analysis we filtered i) low-quality cells (i.e., gene counts smaller 250), ii) cell doublets or multiplets (i.e., RNA counts > 20,000) and iii) cells where more than 6% of the features map to mitochondrial genome (i.e., indicator of dying cells) yielding a 24,525 features x 35,941 (cells) matrix.

We next concentrated on finding highly variable features among the dataset which subsequently efface noise and accelerate downstream computations. We figured that, genes associated with collagen expression and several nematocytes markers (e.g., NvNcol1 or NvNcol6) exhibited the highest cell-to-cell variation within the dataset.

Scaled expression data was subject to a linear dimensional reduction (PCA) approach, where we found that 18 principal components explain most of the variation of the data and hence display a robust compression of the dataset (Supplementary figure 6). However, we note that the assessment of the dimensionality of the dataset is highly subjective and may be subject to discussion. Prescribed PCs were then clustered via a graph-based approach (i.e., Louvain algorithm) to iteratively group cells together, with the goal of optimizing cluster resolution at highest granularity. We find that a resolution parameter of 0.8 preserves the highest granularity within the dataset while keeping number of 'cluster-fluctuations' lowest (Supplementary figure *4*).

A non-linear dimensional reduction (UMAP) was applied on the merged library (i.e., containing both mutant and wild-type scRNA-seq data). Based on the previous clustering approach, this visualization shows 24 distinct cell clusters (Figure *3* A).

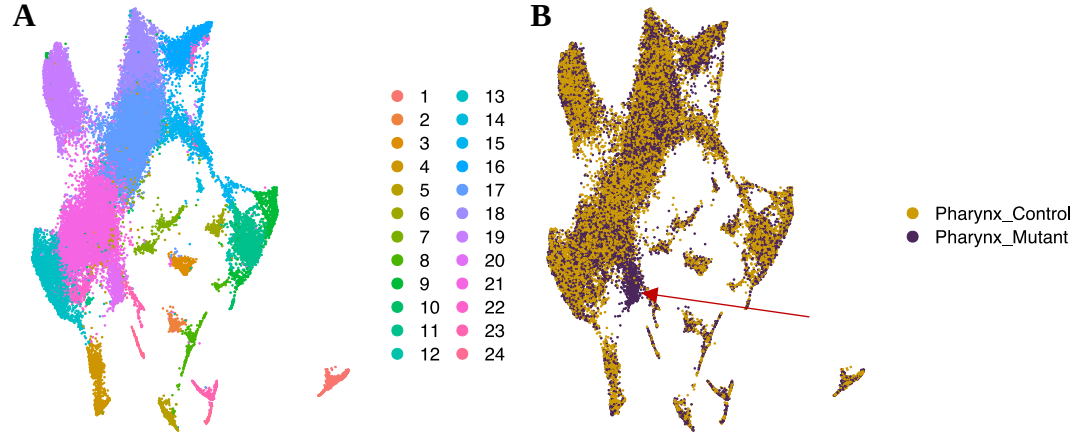


Figure 3: A non-linear dimensional reduction (UMAP) was applied on the merged library (i.e., containing both mutant and wild-type scRNA-seq data) to learn the underlying manifold of the data in low-dimensional space. Note that the previous clustering step (FindClusters()) identified 24 distinct cell populations as shown in A with different colors. B) DimPlot (UMAP) on merged library split by genotype. Note that there is a protruding (red arrow) cell-cluster in the mutant animals identified which is absent in the control animals.

Comparative studies among libraries show that cell cluster identities ‘11’, ‘13’, ‘19’ and ‘21’ (Supplementary figure 7 and Supplementary figure 8) continuously show higher relative cell counts in wt animals, whereas NvTwist$^{-/-}$ mutant animals harbor a distinct cell cluster which is almost absent in wt-animals (Figure 3, B & Supplementary figure 7).

We used the FindAllMarkers() function to distill marker genes within the previously identified cell-clusters. The top 10 marker genes per cluster in terms of average log2FC compared to all other cell identities were used for visualization purposes (Figure 4). We note that ten markers are arbitrarily set. Indeed, several cell-cluster identities show more than 20 marker genes which are specifically enriched in respective cluster. Found markers can further be deployed in annotating cell-cluster identities to know biological cell types (e.g., nematocyte origin, etc.). Interestingly, while clear marker genes were detected for clusters 1 through 19, cell-clusters 20, 21 and 22 show ambiguous patterns. A naïve look into Figure 4 suggests that cell-cluster 20, which is unique to the mutant animals, shows similarity with cell cluster 13, 21 and 22. A further examination of shared markers among the aforementioned clusters is required to elucidate whether cell-cluster ‘20’ is truly unique in terms of markers genes or just an aggregate of different cell types. However, an in-depth analysis of this protruding cell-cluster ‘20’ (Figure *3*, B) revealed that genes *NVE19039, NvNcol6, NvNEP3-b, NVE5857 and NvNcol1* are statistically enriched. Interestingly, although *NvNcol6* displays a conserved marker gene for cell-cluster 20, it was only detected in ~61% of cells.

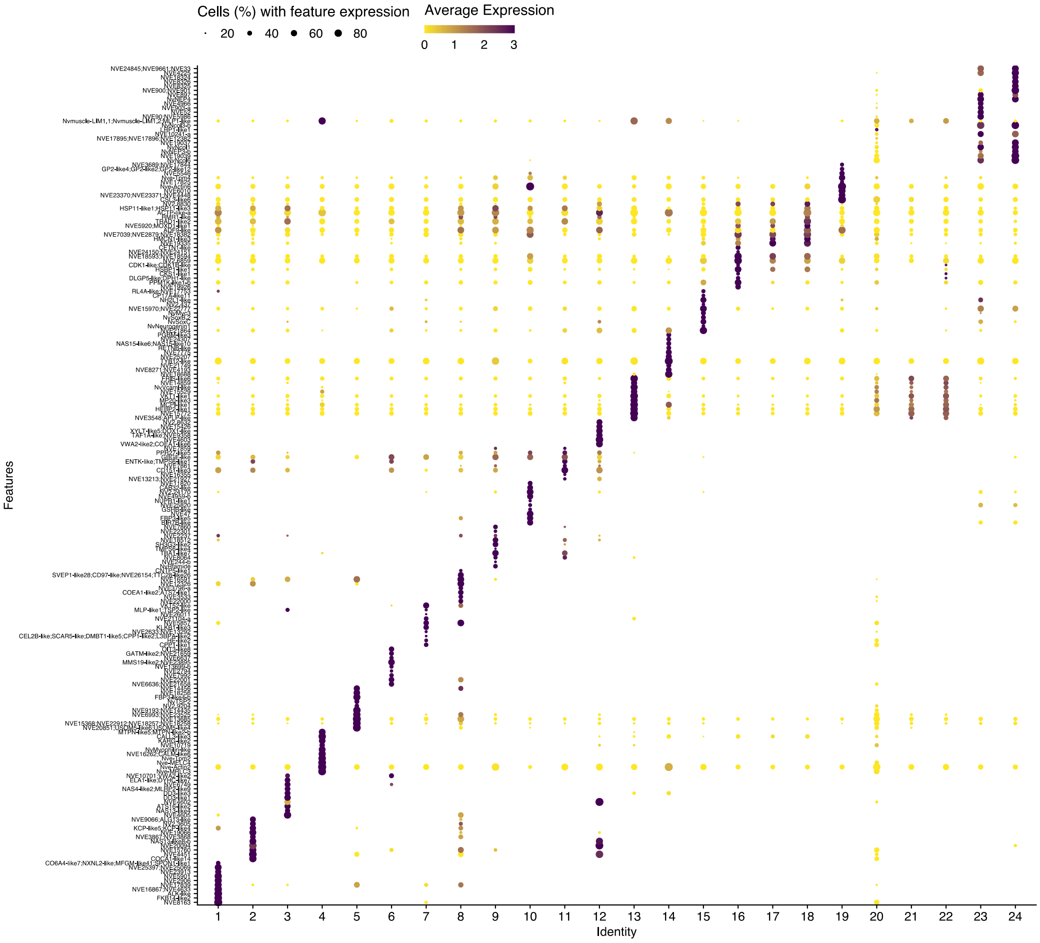


Figure 4: The top 10 marker genes per cluster as identified by FindAllMarkers() with an minimum.pct difference of 0.2 and an logFC threshold greater log(2). Selected features (y-axis) are plotted against the respective cluster (x-axis). The dot scale is proportional to the fraction of cells expressing the feature, ranging from 20% (minimum) to 100%. Colors represent the average expression of the feature. Note that dark purple was set to a maximum of 3.

Outlook

Previous sections describe preliminary results, derived from bulk and scRNA-seq analysis of *NvTwist* mutant animals. Albeit some compelling genetic differences among mutant and wild-type animals were shown, this analysis is far from being exhausted. Rather, focus must be spend on *in vitro* validations of some of the observed results from DGE analysis. Further, scRNA-seq displays a data-rich source which further needs to be exploited. Particularly, the newly identified cell-cluster '20' needs further consideration. Open questions which need to be addressed include the following:

i) Does cell-cluster 20 depict a novel cell-type in *N. vectensis*?

ii) If not, can this cell-cluster be explained by gene markers derived from known cell-types (e.g., nematocytes, muscle cell identity, etc.)?

iii) Does cell-cluster 20 represent a homo- or heterogenous set of cells?

iv) Can differentially expressed genes (i.e., from bulk RNA-seq data analysis) also be found within the new cell-cluster?

v) Elucidate the position of *NvTwist* within the gene regulatory circuits in *N. vectensis*. Does any other pathway, or specific protein, compensate for the lack of *NvTwist*?

vi) Specifically, does the FGF(R) pathway compensate for the lack of *NvTwist*?

Aforementioned questions are not exhaustive, however should point to a direction of further investigations. Ultimately, we want to understand the role of *NvTwist*, specifically which genetic perturbation leads to the generation of the novel 'bubble' phenotype.

References

Anders, S. and Huber, W. (2010) ‘Differential expression analysis for sequence count data’, *Genome Biology*. BioMed Central, 11(10), pp. 1–12. doi: 10.1186/gb-2010-11-10-r106.

Briggs, J. A. *et al.* (2018) ‘The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution’, *Science*. American Association for the Advancement of Science, 360(6392). doi: 10.1126/SCIENCE.AAR5780.

Bürglin, T. R. and Affolter, M. (2016) ‘Homeodomain proteins: an update’, *Chromosoma*. Springer, 125(3), p. 497. doi: 10.1007/S00412-015-0543-8.

Castanon, I. and Baylies, M. K. (2002) ‘A Twist in fate: Evolutionary comparison of Twist structure and function’, *Gene*. Elsevier, 287(1–2), pp. 11–22. doi: 10.1016/S0378-1119(01)00893-9.

Hagauer, J. (2020) *A functional analysis of a twist ortholog in the sea anemone Nematostella vectensis*.

Hao, Y. *et al.* (2021) ‘Integrated analysis of multimodal single-cell data’, *Cell*. Cell Press, 184(13), pp. 3573-3587.e29. doi: 10.1016/j.cell.2021.04.048.

Jahnel, S. M., Walzl, M. and Technau, U. (2014) ‘Development and epithelial organisation of muscle cells in the sea anemone Nematostella vectensis’, *Frontiers in Zoology*. BioMed Central Ltd., 11(1). doi: 10.1186/1742-9994-11-44.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*. NIH Public Access, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Li, H. (2011) ‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data’, *Bioinformatics*. Oxford University Press, 27(21), pp. 2987–2993. doi: 10.1093/bioinformatics/btr509.

Liao, Y., Smyth, G. K. and Shi, W. (2014) ‘FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features’, *Bioinformatics*. Oxford University Press, 30(7), pp. 923–930. doi: 10.1093/bioinformatics/btt656.

Love, M. I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome Biology*. BioMed Central Ltd., 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Morimoto, R. (1993) ‘Cells in stress: transcriptional activation of heat shock genes’, *Science*. American Association for the Advancement of Science, 259(5100), pp. 1409–1410. doi: 10.1126/SCIENCE.8451637.

Najafabadi, H. S. *et al.* (2015) ‘C2H2 zinc finger proteins greatly expand the human regulatory lexicon’, *Nature Biotechnology 2015 33:5*. Nature Publishing Group, 33(5), pp. 555–562. doi: 10.1038/nbt.3128.

Pennisi, E. (2007) ‘Sea anemone provides a new view of animal evolution’, *Science*, p. 27.

doi: 10.1126/science.317.5834.27.

Rostom, R. *et al.* (2017) ‘Computational approaches for interpreting scRNA-seq data’, *FEBS Letters*. John Wiley & Sons, Ltd, 591(15), pp. 2213–2225. doi: 10.1002/1873-3468.12684.

Sandmann, T. *et al.* (2007) ‘A core transcriptional network for early mesoderm development in Drosophila melanogaster’, *Genes and Development*. Genes Dev, 21(4), pp. 436–449. doi: 10.1101/gad.1509007.

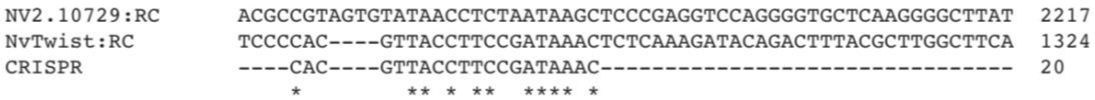
Stuart, T. *et al.* (2019) ‘Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data’, *Cell*, 177. doi: 10.1016/j.cell.2019.05.031.

Yang, J. *et al.* (2004) ‘Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis’, *Cell*. Cell, 117(7), pp. 927–939. doi: 10.1016/j.cell.2004.06.006.

Zheng, G. X. Y. *et al.* (2017) ‘Massively parallel digital transcriptional profiling of single cells’, *Nature Communications*. Nature Publishing Group, 8(1), pp. 1–12. doi: 10.1038/ncomms14049.

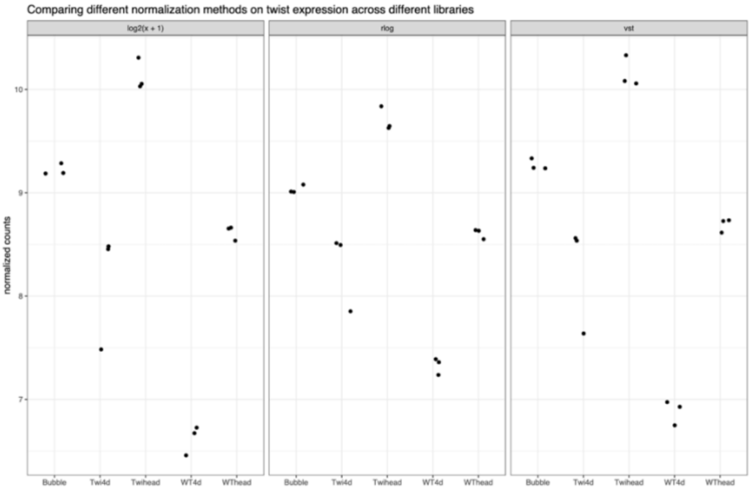
Supplementary information

Supplementary figures

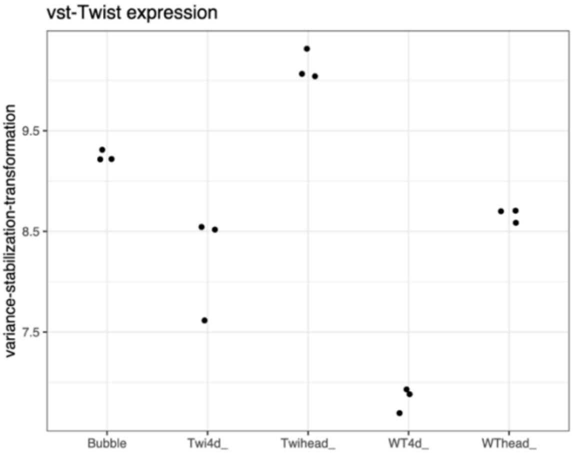


```
NV2.10729:RC       ACGCCGTAGTGTATAACCTCTAATAAGCTCCCGAGGTCCAGGGGTGCTCAAGGGGCTTAT  2217
NvTwist:RC         TCCCCAC----GTTACCTTCCGATAAACTCTCAAAGATACAGACTTTACGCTTGGCTTCA  1324
CRISPR             ----CAC----GTTACCTTCCGATAAAC--------------------------------  20
                       *        ** * **  **** *
```

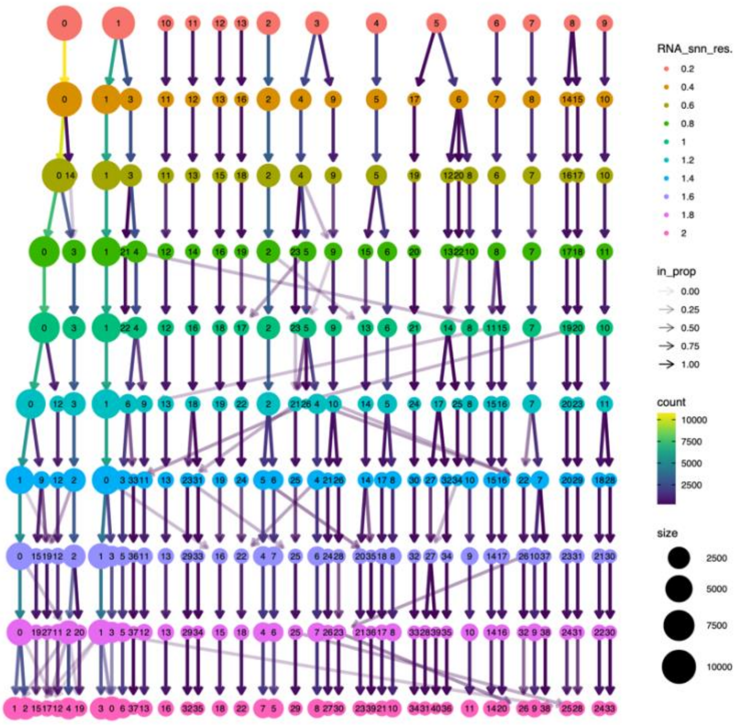
Supplementary figure 1: Nucleotide multiple-sequence-alignment (MSA) of the proposed off-target NV2.10979, NvTwist and the guideRNA deployed for the CRISPR/Cas9 KO. The MSA was generated using the Clustal- Omega online tool. Note that asterisks underneath the nucleotides indicate a match in all applied sequences.



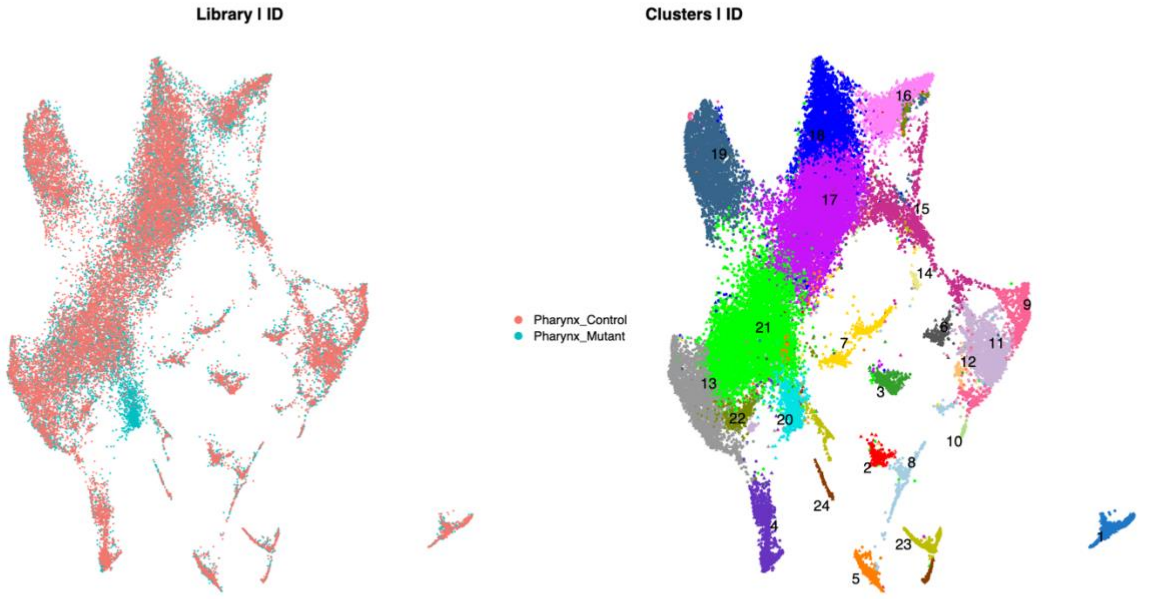
Supplementary figure 2: Three different normalization methods applied to bulk RNA-seq count data. Note that for illustrative purposes we only show *NvTwist* quantities. The different normalization methods applied are split by black panels and described in the header. The x-axis describes animals with distinct geno-/phenotype combination (see Table 1) used for this experiment. Note that one black dot displays one biological replicate.



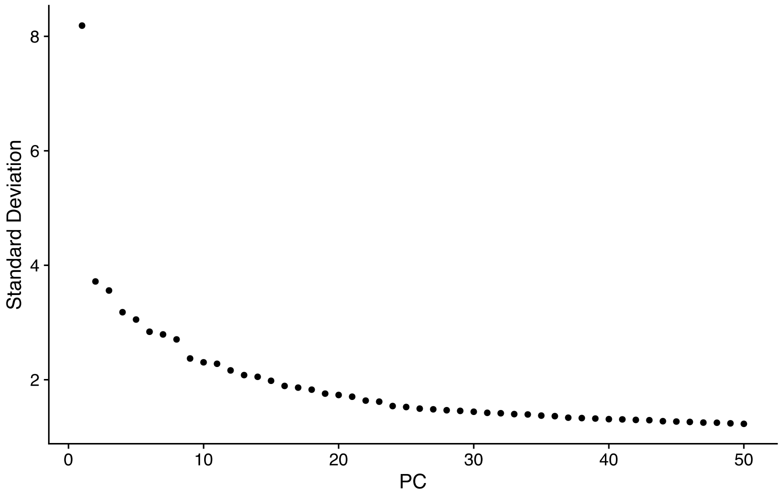
Supplementary figure 3: Variance-stabilization-transformed NvTwist expression among deployed animals. The x-axis highlights animals with distinct geno-/phenotype combinations (see Table 1). One black dot represents one biological replicate.



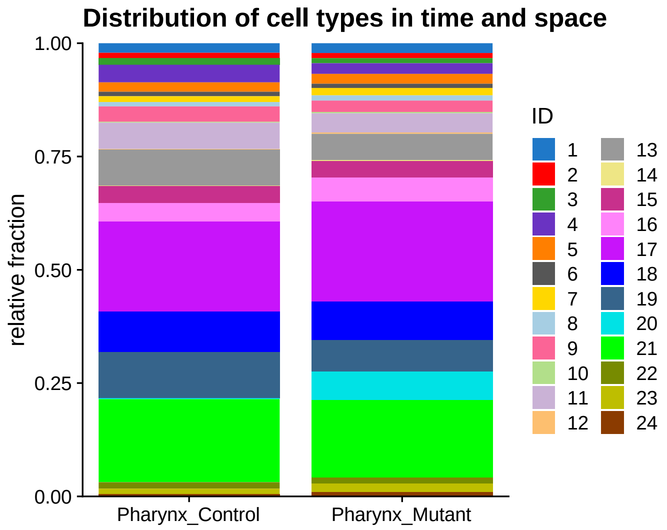
Supplementary figure 4: Modularity optimization technique with varying resolution parameters. With the FindNeighbors() function, Seurats clusters the cells via the Louvain algorithm, to iteratively group cells together. We used this function with varying resolution parameters starting from 0.2 (red, top row) towards 2 (pink, bottom row). Note that an increase in the resolution parameters generally increase the number of cell clusters.



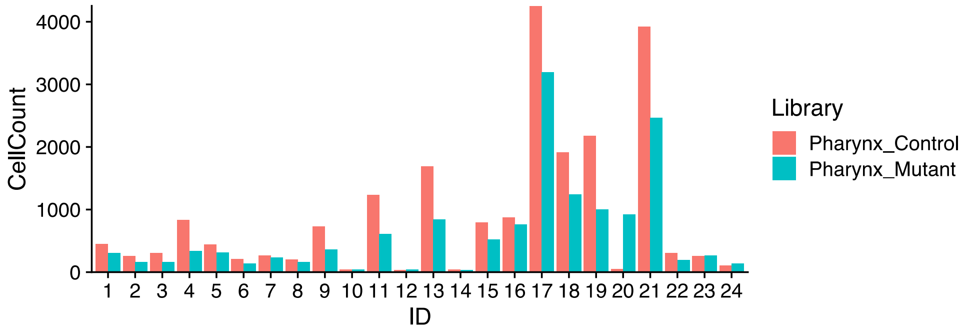
Supplementary figure 5: Replication of UMAP projection of identified cell-clusters as shown in Figure 2. Note, however, that the color changed and that the protruding cell-cluster in mutant-animals are assigned the number 20.



Supplementary figure 6: ‘Scree-‘, or elbow plot which displays the pre-defined number of principal components (x-axis) and their respective standard deviation (y-axis) as a surrogate for eigenvalues. Note, that the higher the standard deviation the more variation a particular PC explains. We set a cut-off of 18 PCs, as the top 18 PCs all explain roughly more than 2 SD units.



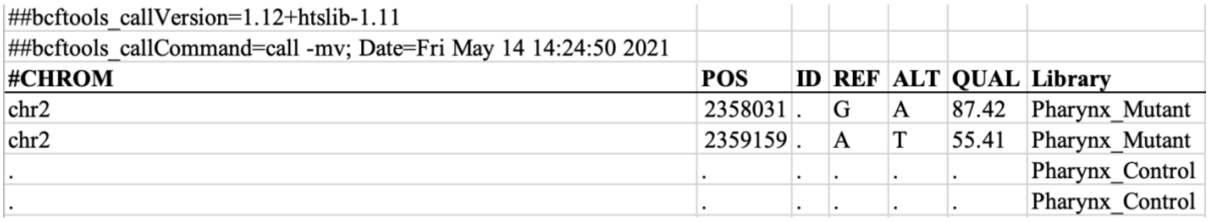
Supplementary figure 7: Relative cell fraction in respective cell clusters stratified according to the underlying genotype (x-axis).



Supplementary figure 8: Absolute cell counts in respective cell-clusters. Note that absolute cell numbers should be cautiously interpreted, as the respective library size is not taken into account.

Supplementary tables

Supplementary table 1: SNPs associated with NvTwist, found in scRNA-seq data from NvTwist-mutants. As expected, there was no single SNP detected around NvTwist in control animals (‘Pharynx_Control’)



##bcftools_callVersion=1.12+htslib-1.11						
##bcftools_callCommand=call -mv; Date=Fri May 14 14:24:50 2021						
#CHROM	**POS**	**ID**	**REF**	**ALT**	**QUAL**	**Library**
chr2	2358031	.	G	A	87.42	Pharynx_Mutant
chr2	2359159	.	A	T	55.41	Pharynx_Mutant
.	.	.	.	.	.	Pharynx_Control
.	.	.	.	.	.	Pharynx_Control

Supplementary file

Summary DGE table