







REPORT



Identification and characterization of an IgG sequence variant with an 11 kDa heavy chain C-terminal extension using a combination of mass spectrometry and high-throughput sequencing analysis

Claire Harris ^{a*}, Weichen Xu ^{b*}, Luigi Grassi ^a, Chunlei Wang ^b, Abigail Markle^b, Colin Hardman ^c, Richard Stevens^{d#}, Guillermo Miro-Quesada^e, Diane Hatton^a, and Jihong Wang ^b

^aBiopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; ^bBiopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, USA; ^cData Science & Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; ^dAntibody Discovery and Protein Engineering, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; ^eData & Quantitative Sciences, Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, USA

ABSTRACT

Protein primary structure is a potential critical quality attribute for biotherapeutics. Identifying and characterizing any sequence variants present is essential for product development. A sequence variant ~11 kDa larger than the expected IgG mass was observed by size-exclusion chromatography and two-dimensional liquid chromatography coupled with online mass spectrometry. Further characterization indicated that the 11 kDa was added to the heavy chain (HC) Fc domain. Despite the relatively large mass addition, only one unknown peptide was detected by peptide mapping. To decipher the sequence, the transcriptome of the manufacturing cell line was characterized by Illumina RNA-seq. Transcriptome reconstruction detected an aberrant fusion transcript, where the light chain (LC) constant domain sequence was fused to the 3' end of the HC transcript. Translation of this fusion transcript generated an extended peptide sequence at the HC C-terminus corresponding to the observed 11 kDa mass addition. Nanopore-based genome sequencing showed multiple copies of the plasmid had integrated in tandem with one copy missing the 5' end of the plasmid, deleting the LC variable domain. The fusion transcript was due to read-through of the HC terminator sequence into the adjacent partial LC gene and an unexpected splicing event between a cryptic splice-donor site at the 3' end of the HC and the splice acceptor site at the 5' end of the LC constant domain. Our study demonstrates that combining protein physicochemical characterization with genomic and transcriptomic analysis of the manufacturing cell line greatly improves the identification of sequence variants and understanding of the underlying molecular mechanisms.

ARTICLE HISTORY

Received 24 July 2019
Revised 3 September 2019
Accepted 10 September 2019

KEYWORDS

Fc-extension; aberrant fusion protein; alternative splicing; high throughput sequencing; RT-PCR; monoclonal antibody; sequence variant; LC/MS; expression vector; splice variants; nanopore sequencing



Introduction

Therapeutic glycoproteins such as monoclonal antibodies (mAbs) are produced in mammalian cell production platforms, such as those based on Chinese hamster ovary (CHO) cell lines, which inherently generate product heterogeneity. Pharmaceutical drug product characterization is a regulatory requirement that ensures consistent drug product quality between manufacturing batches.¹ Critical quality attributes (CQAs) of therapeutic proteins are product attributes that affect the biological activity, pharmacokinetics, immunogenicity, or safety of the product.² Primary structure variants, including single amino acid substitutions and extensions or truncations of the amino acid sequence, are potential CQAs of therapeutic proteins. Sequence variants produced by either genomic mutations or misincorporation during translation have been reported for therapeutic protein products.^{3–8}

Variants can be introduced into the encoding gene sequences during vector propagation in bacteria prior to transfection or


may result from errors in DNA replication in the production cell line once the vector has been incorporated. For example, a point mutation in a single cell may lead to a sub-population of the production cell line that expresses a single nucleotide variant (SNV), resulting in a percentage of the final drug product carrying a single amino acid substitution.⁹ Not all SNVs lead to amino acid substitutions. Silent mutations, in which the mutated DNA encodes the same original amino acid, do not directly alter the protein sequence, but they can affect transcript stability, protein folding,^{10,11} and mRNA splicing.¹² SNVs within a stop codon can cause an extension product.⁷ Single amino acid substitutions can also be produced from mis-priming tRNA caused by amino acid depletion in the culture medium.³ Most single amino acid substitutions can be identified by peptide mapping using liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS).^{8,13} The recent advancements in peak detection capability using high-end mass spectrometers and informatic tools have proven useful in detecting single amino acid

CONTACT Jihong Wang, PhD  wangjh@vielabio.com  Analytical Sciences, Viela Bio, One MedImmune Way, Gaithersburg, MD 20878, USA;

Diane Hatton, PhD  diane.hatton@astrazeneca.com  Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB21 6GH, UK

*These authors contributed equally to this work

#Current address: Source Bioscience, Units 24/25 William James House, Cowley Road, Cambridge, CB4 0WU, UK

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

substitutions at a sensitivity of >0.001%.³ However, data from LC-MS/MS should be carefully examined and verified by orthogonal methods due to prevalent false positive results.¹⁴ Beyond SNVs, DNA deletions or insertions can cause a frame-shift leading to an altered amino acid sequence,⁵ or a premature stop codon and the formation of a truncated product. Larger-scale genomic recombination events can produce extension products, including molecules in which entire domains were added.^{13,15}

Primary structure variants can also be caused by mis-splicing of introns present in the transgene. These non-coding components can be included in the transgene structure to increase the level of protein expression.¹⁶⁻¹⁸ However, this advantage comes with a risk of mis-splicing if alternative splice donor and acceptor sites are used during transcript processing, or of intron retention in the mature transcript if an intron is not correctly spliced.^{19,20} Such incorrect transcript processing can generate altered amino acid sequences or truncated products. Splice variants are identified only in the sequence of the mature transcript, with no trace in the corresponding genomic sequence, and therefore are more challenging to detect.

When protein sequence variants are present at a relatively high level, they can be detected by intact mass analysis or peptide mapping UV profile comparison. The sequence variants can be further characterized by *de novo* sequencing using tandem mass spectrometry (MS/MS). However, *de novo* sequencing from single enzyme peptide mapping data poses challenges, especially for large unknown sequence variants, due to the great number of possible fragment ion assignments and less than 100% sequence coverage resulting from incomplete fragmentation. Hence, a proteomic approach such as multi-enzyme digestion is essential for *de novo* sequencing analyses.^{6,19} In addition, peptide mapping methods alone are usually not sufficient to identify low-level sequence variants (<1%) other than single amino acid substitutions. Even though low-level sequence variants can be enriched by chromatography approaches, such as size exclusion,^{5,15} ion exchange,²¹ or reversed-phase²⁰ chromatography, it is still time-consuming and resource-intensive to enrich enough material for multi-enzyme *de novo* sequencing. The lack of peak identification and annotation is a limiting factor for proteomics experiments that can be overcome by proteogenomics, a new field that is based on the use of high-throughput data from different sources as part of an iterative refinement process of gene models.²²⁻²⁴

Nucleotide sequencing technologies offer a complementary approach to identify variants encoded in genes or mature transcripts. In particular, high-throughput sequencing (HTS) is a powerful tool able to overcome the limitation of sensitivity typical of the traditional Sanger sequencing of reverse transcription-polymerase chain reaction (RT-PCR) product variants.^{25,26} Several methods based on HTS can be used to characterize genomes and transcriptomes.²⁵ The extra information gathered from these analyses defines a more comprehensive search space for MS/MS identification.²⁷ Strategies using an orthogonal approach for sequence variants detection have evolved as reported recently by Lin et al.²⁸

Here, we report the discovery, identification, and characterization of an 11 kDa Fc C-terminal extension sequence variant of a recombinant IgG1 mAb (mAb-A) from a CHO manufacturing

cell line by using a combination of MS methods and HTS. Intact mass analysis and peptide mapping were used to deduce that the 11 kDa increase in molecular mass resulted from an addition to the heavy chain Fc. The identity of the Fc C-terminal extension as light chain constant domain sequence was enabled by using HTS to assess the transcriptome of the manufacturing cell line, which detected an aberrant heavy chain transcript with the light chain constant domain sequence fused at the 3' end. Furthermore, nanopore long-read genomic sequencing highlighted that the aberrant fusion transcript originated from cryptic splicing of a transcript derived from an unexpected partially deleted copy of the plasmid. This study emphasizes the power of integrating product physicochemical characterization data with cell line omics data to understand therapeutic protein sequence variants and to define screening strategies for cell lines with improved product quality profiles.

Results

2D-LC/MS and HPSEC fractionation reveal protein sequence variants

During early process development and product characterization, mAb-A showed a front shoulder on the high-performance size-exclusion chromatography (HPSEC) main peak (Figure 1a). Species eluting in this front shoulder peak were trapped online, desalted, and transferred for mass measurement using two-dimensional SEC and reversed-phase liquid chromatography coupled with online MS (2D (SEC/RP)-LC-MS) setup. The deconvoluted mass showed the front shoulder peak contained a mass 11340 Da higher than the monomer (Supplementary Figure S1). To further characterize the size variant species under the shoulder, mAb-A was fractionated using preparative HPSEC. An enriched fraction containing 80% of the front shoulder was obtained (as shown in Figure 1b), which was used for extensive characterization.

The HPSEC fractions were analyzed by intact mass (Figure 2a-f) and capillary gel electrophoresis (CE-SDS) (Figure 3a-d). In non-reducing intact mass analysis, the monomer fraction only contained the expected monomer mass 145955 Da (Figure 2a). In comparison, the front shoulder fraction contained an additional species with mass of 157290 Da, which is 11335 Da higher than the expected monomer mass (Figure 2d). In reducing intact mass analysis, the monomer and shoulder fractions contained the same mass for the light chain (Figure 2b-e), whereas an additional species with a mass of 61340 Da was observed in the shoulder fraction for the heavy chain (Figure 2f), which is 11218 Da higher than expected (Figure 2c). The difference in mass addition measured under non-reducing and reducing conditions is 117 Da (11335 Da - 11218 Da), which is consistent with the loss of a cysteine adduct and opening of a disulfide bond after reduction. Therefore, the intact mass analysis results indicated that the product variant had 11218 Da added to the heavy chain and contained a disulfide bond linked cysteine adduct. To verify this finding, an orthogonal size variant characterization method, CE-SDS, was used to analyze the fractions. Consistent with the intact mass analysis, the shoulder fraction contained a major size variant peak that migrated later than the intact monomer peak in non-reducing CE-SDS at 30 min (Figure 3a-c). Similarly,

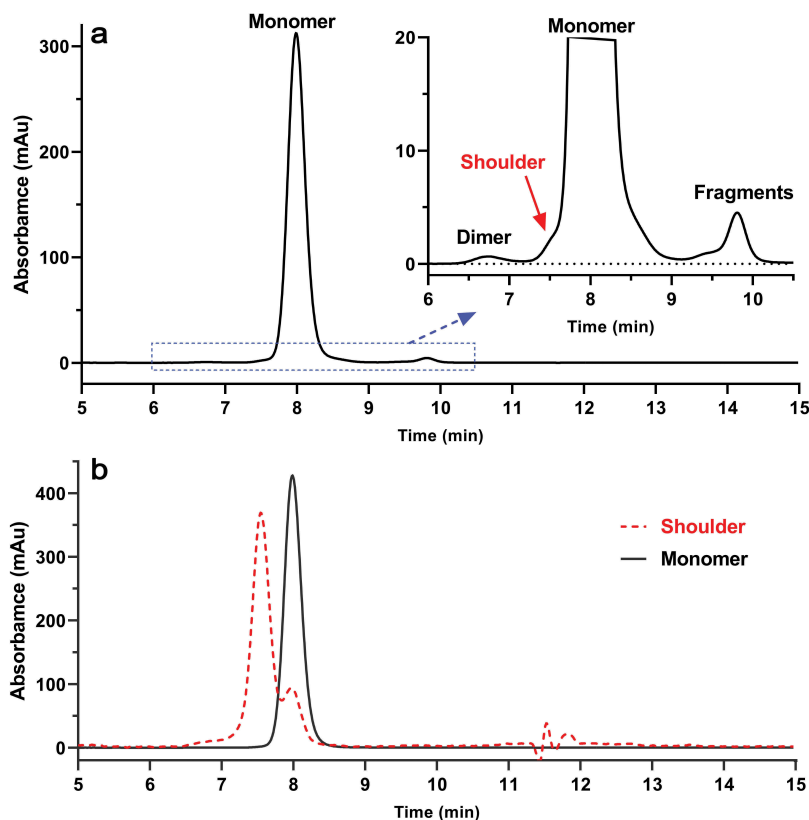


Figure 1. (a): HPSEC profile of mAb-A. Inset is zoomed view. (b): HPSEC profiles of mAb-A shoulder (red) and monomer (black) fractions. Peaks at 11.5 to 12 min in shoulder fraction are system peaks.

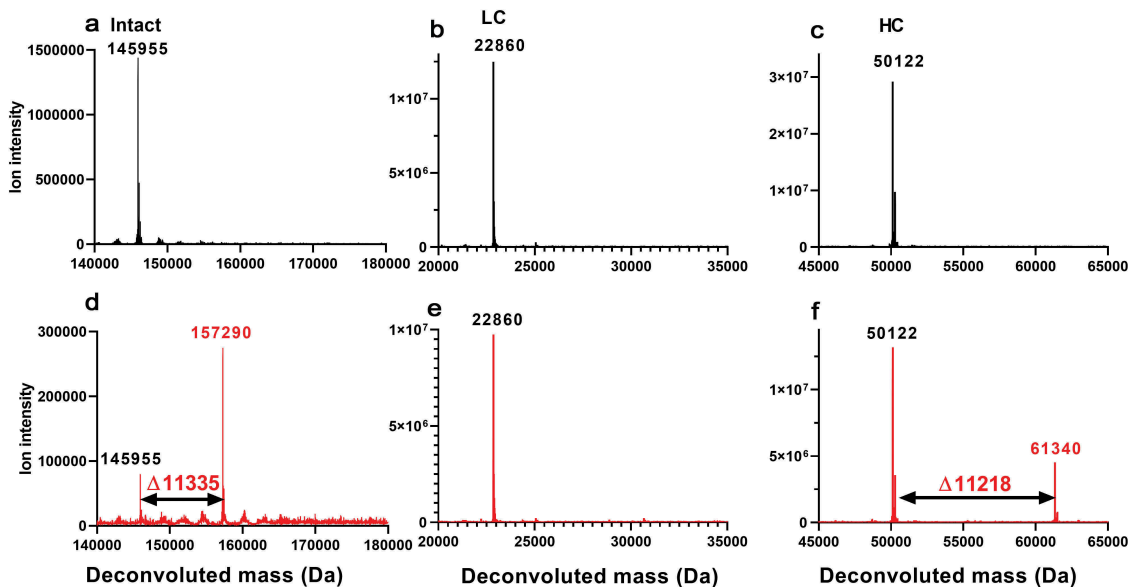


Figure 2. Intact mass analysis: Deconvoluted masses for HPSEC monomer fraction (a): intact mass (b): reduced light chain (c): reduced heavy chain and shoulder fraction (d): intact mass (e): reduced light chain (f): reduced heavy chain.

a variant peak migrating later than the heavy chain peak at 22 min was observed by reducing CE-SDS (Figure 3b–d). Together these results confirmed that the extra 11 kDa mass is an addition to one of the heavy chains. Theoretically, a variant with an additional 11 kDa mass on both heavy chains could also be present. However, the level of the antibody variant with the extension on a single heavy chain was already very low

(~0.5%), so the occurrence of antibody variant with the addition on both heavy chains would be even lower and possibly not detectable by intact mass and non-reducing CE-SDS (Figures 2d and 3c).

Subunit mass analysis after IdeS digestion was performed to determine whether the 11 kDa sequence variant was added to heavy chain Fd or Fc domain. IdeS digests IgG1 below the hinge

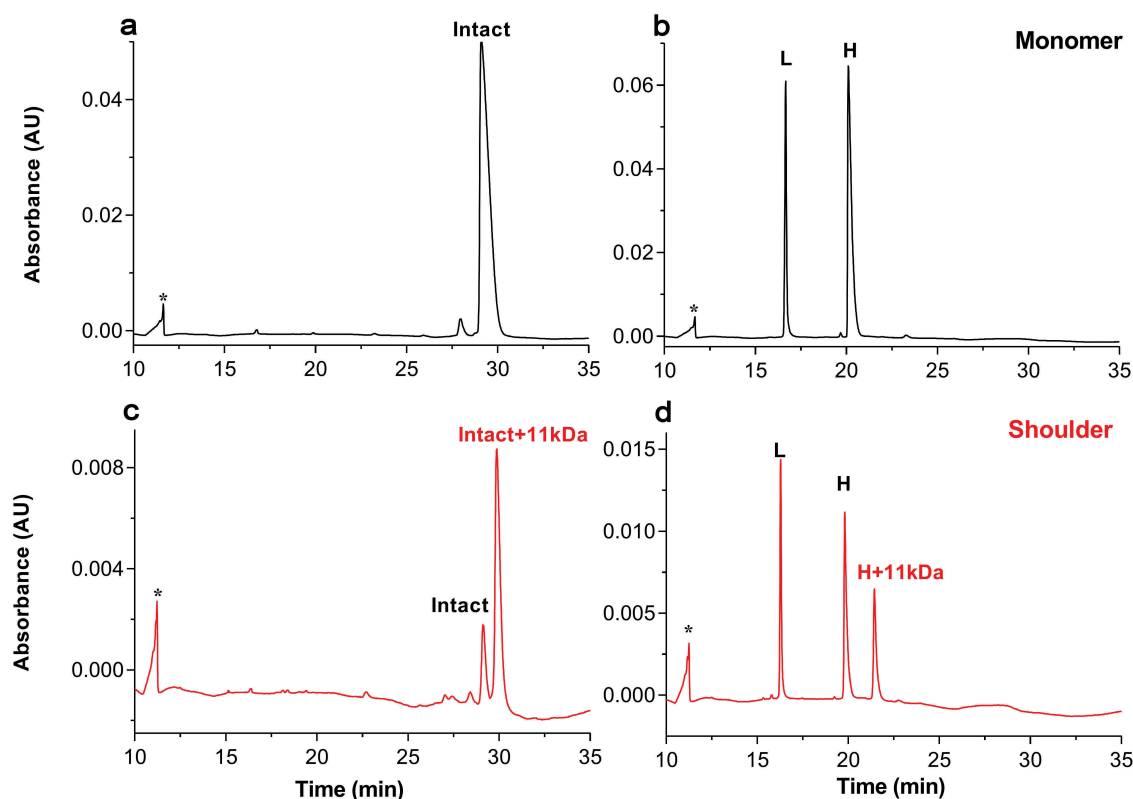


Figure 3. CE-SDS analysis for HPSEC monomer fraction (a): non-reducing (b): reducing and shoulder fraction (c): non-reducing (d): reducing.

* are system peaks; H: heavy chain; L: light chain.

region, resulting in a $F(ab')_2$ and $Fc'/2$ fragments that can be separated by reversed-phase liquid chromatography (Figure 4a), and then characterized by online MS. Figure 4b,c shows the subunit mass analysis of IdeS-digested shoulder fraction. The first small peak in Figure 4a corresponded to $Fc'/2$ and the second small peak to $Fc'/2$ with the 11 kDa addition. The large peak corresponded to $F(ab')_2$ domain and partially digested mAb (missing one but not both $Fc'/2$). There is no additional mass added to the $F(ab')_2$ domain, and the results showed the 11 kDa is added to one of the $Fc'/2$ fragments.

Reducing Lys-C peptide mapping was performed to identify the peptide sequence corresponding to the 11 kDa variant. Figure 5 compares the peptide map UV profiles of the shoulder and monomer fractions. Two new peaks were observed in the shoulder fraction. The peak at 28.5 min corresponded to an oxidized Met-428 containing peptide, which was enriched in the HPSEC front shoulder peak. The other new peak at 23.4 min had a detected mass of 1012.59 Da. This mass cannot be assigned to any mAb-A peptide fragment. In addition, when peak heights were compared, the two peptides that had significant higher intensity in the shoulder fraction (peak a and b in Figure 5) were both from the light chain constant region. This result suggested that the 11 kDa variant may contain extra light chain sequence.

It has been reported that extra light chain may be added to the intact mAb through disulfide bonds, leading to a size variant.^{29,30} However, this is not the explanation here because 1) the addition of an extra light chain would add a mass over 20 kDa, and 2) extra disulfide bond-linked light chain would not

register as an extra peak in either CE-SDS or intact mass analysis after reduction. Our data showed that the 11 kDa variant is covalently linked to the heavy chain Fc domain. Additionally, a simple mutation in the stop codon can also be ruled out because the peptide containing the heavy chain C-terminal lysine is barely detected in either fraction. If the 11 kDa was an extension of the C-terminus of the heavy chain, caused by a mutation in the stop codon, the levels of peptide containing this lysine residue would be much higher in the HPSEC shoulder fraction.

The protein physicochemical analysis described above did not unambiguously elucidate the structure of the sequence variant. A complementary approach was thus used to investigate variants encoded at the nucleic acid level by the manufacturing cell line. However, standard Sanger sequencing of RT-PCR products using primers to amplify the entire heavy chain transcript did not detect any sequence variants. Therefore, more sensitive and data-rich HTS characterization was used to analyze both the transcriptome and the genome of the manufacturing cell line.

The *de novo* assembly of RNA-seq reads identified a heavy chain fusion transcript with 122 additional nucleotides at its 3' end and unmappable in contiguity onto the expression plasmid (soft masked) used to generate the cell line. This fusion transcript sequence was aligned to the reference plasmid and shown to match the 5' end of the constant domain sequence of the light chain (Figure 6a). The sequence of the protein product of the fusion transcript was inferred by extending and translating its sequence up to the end of the light chain constant domain.

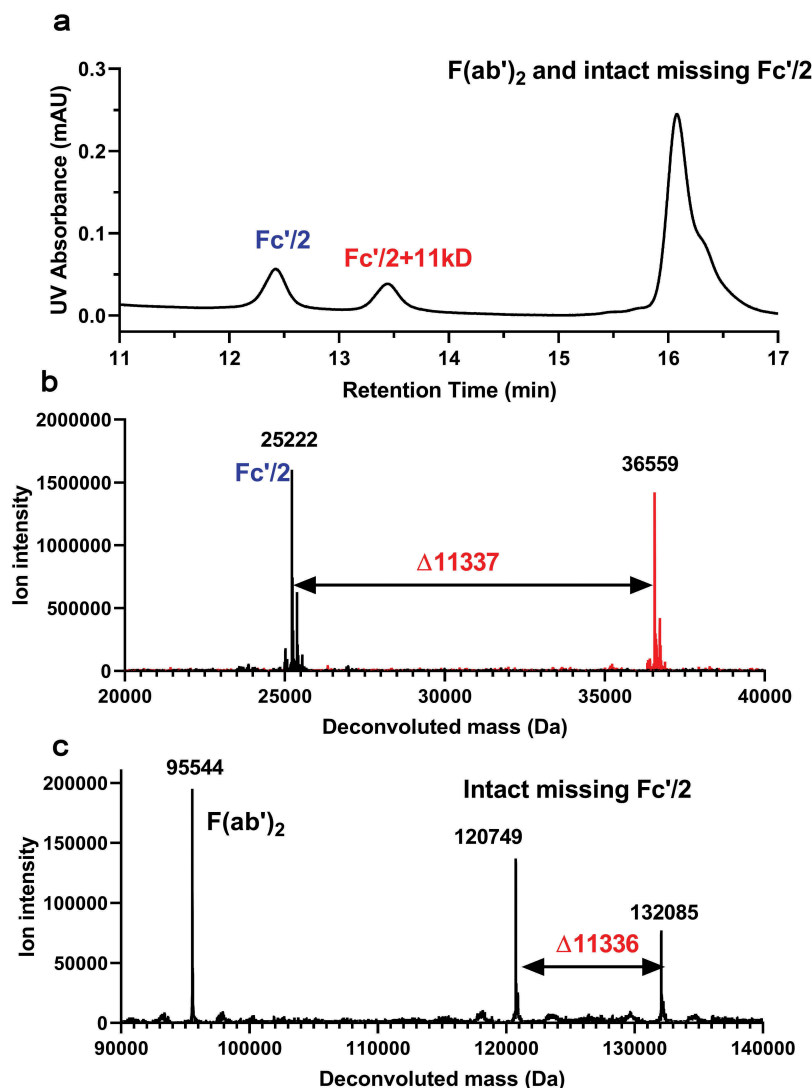


Figure 4. RP-HPLC analysis for IdeS digested mAb-A HPSEC shoulder peak (a): RP-HPLC UV chromatogram. (b): Deconvoluted masses for 12–14 min peaks in A. (black is Fc'; 12.4 min peak in A, red is Fc'+11 kDa: 13.4 min peak in A). (c): Deconvoluted masses for 16 min peak in A (F(ab')₂ and intact missing Fc' or Fc'/2). F(ab')₂ = (LC+HC1-236)₂.

This produced an extra 105 amino acids at the C-terminus of the heavy chain, as shown in Supplementary Figure S2a,b, in agreement with the 11218 Da mass increase detected by reducing intact mass analysis. The sequence of the heavy chain-light chain constant domain fusion transcript was confirmed by RT-PCR sequencing using a cDNA template derived from the manufacturing cell line with primers that bind to the CH3 domain of the heavy chain, and the polyA sequence of the light chain. This new fusion transcript, and the canonical mAb-A heavy chain and light chain were added to the CHO-K1 transcriptome, and the expression level of the heavy chain fusion transcript was found to be 0.4% of the amount of the canonical heavy chain transcript (11.2 transcripts per million (TPM) VS 2558.9 TPM, Supplementary Figure S3).

The fusion transcript could have been produced from a genomic recombination event, splicing of the RNA, or a combination of both. Inspecting the genomic sequencing data provided by the MinION nanopore sequencer showed that multiple copies of the plasmid had integrated into the genome in tandem. There was no evidence of a recombination event where

the heavy and light chain sequences found in the fusion transcript were found adjacent to each other in the encoding DNA. However, in one situation a partial copy of the plasmid had integrated, leading to the light chain constant sequence being closer to the end of the heavy chain than would occur if the whole plasmid copy had been integrated (Figure 6b). A splice site prediction tool³¹ identified a possible cryptic donor site at the end of the heavy chain. This could pair with the splice acceptor site of the downstream light chain constant domain and splice out the heavy chain gene polyA and 3' vector backbone, creating the fusion product between the heavy chain and light chain constant sequences (Figure 6c). As the additional sequence is part of the mAb-A light chain, this meant there was only one unique peptide for this fusion protein, which was at the junction between the heavy and light chains; SLSLSPGQPK as shown in Supplementary Figure S2b.

The unique signature peptide SLSLSPGQPK has a theoretical mass of 1012.56 Da. This mass matched the detected mass of the unknown peak in the shoulder fraction (Figures 5–7a,b). The tandem mass analysis of this peak further verified the peptide

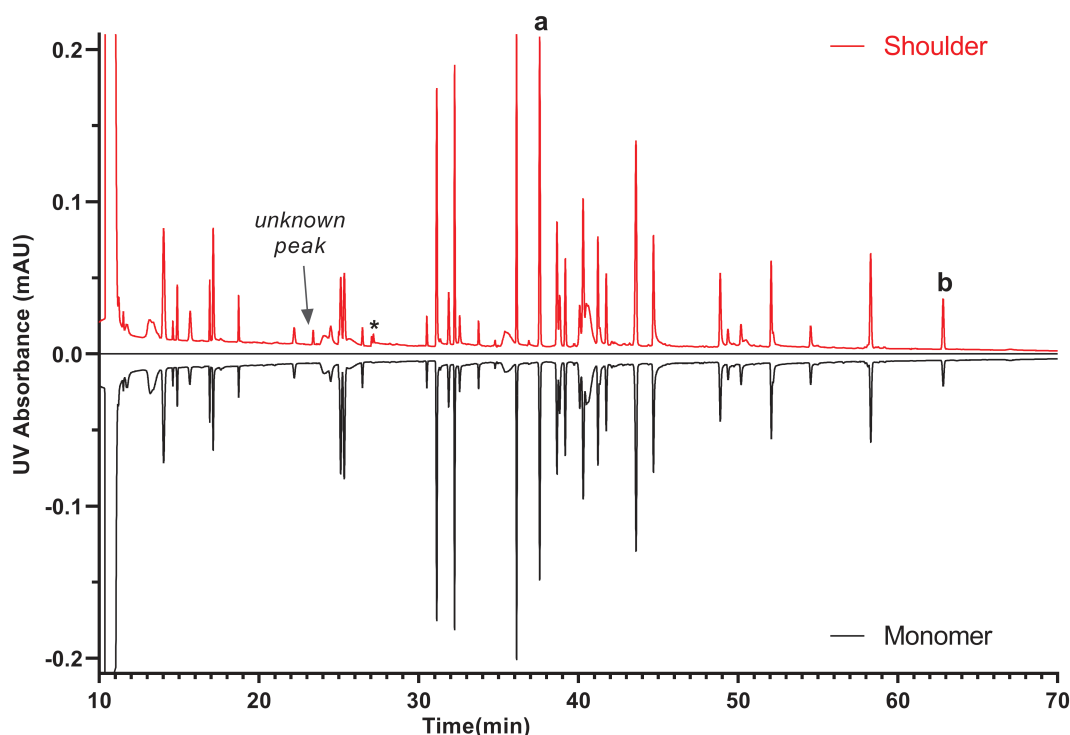


Figure 5. Reducing Lys-C mapping UV overlay of HPSEC shoulder fraction (red) and monomer fractions (black). (a): 37.67 min peak is light chain peptide Y176-K190 (YAASSYLSLTPEQWK) with mass of 1742.852 Da, and (b): 62.95 min peak is light chain peptide A134-K153 (ATLVCLISDFYPGAVTVAWK) with mass of 2153.123 Da. * is heavy chain peptide with oxidized Met-428.

sequence (Figure 7c). In addition, the non-reducing peptide mapping analysis confirmed the cysteine adduct and one additional disulfide bond in the extension peptide. The UV chromatograms of non-reducing Lys-C peptide mapping with all peaks related to the 11 kDa extension labeled are shown in Supplementary Figure S4, and the peptides listed in Table S1.

Discussion

An Fc C-terminal extension sequence variant in mAb-A was first discovered as a shoulder peak to the front of HPSEC monomer peak. Quantitation by HPSEC showed that the front shoulder comprised less than 0.5% of the monomer, which was insufficient for detection by intact mass analysis, as described previously for other Fc extensions.²⁰ The mass of the unknown species was determined by 2D-(SEC/RP) LC/MS to be ~11 kDa higher than the expected mass of the IgG. The front shoulder was then enriched by HPSEC fractionation for in-depth characterization. Intact mass analysis of the enriched shoulder peak provided an accurate mass of the variant and revealed that the additional 11 kDa molecular mass was part of the Fc domain. Due to the size of the variant (11 kDa larger), *de novo* sequencing using peptide mapping to further interrogate the sequence variant would require a large quantity of enriched material, as well as multiple digestions with different enzymes. Therefore, we opted for HTS genome and transcriptome analysis of the manufacturing cell line as an approach to decipher RNA and DNA sequences that could encode the variant.

The RNA-seq analysis detected a fusion transcript where sequence from the constant domain of the light chain was

added to the 3' end of the heavy chain. The presence of the complete light constant domain coding sequence at the end of the heavy chain in a fusion transcript was confirmed by RT-PCR sequence analysis. The molecular mass of the translated light chain constant domain sequence matched the extra 11 kDa detected for the Fc extension variant. Furthermore, as the additional sequence is the same as the constant domain of mAb-A's light chain, we observed the corresponding light chain-related peak intensity differences in the peptide mapping with only one unique peak corresponding to the peptide fragment at the junction of the heavy chain and light chain. Unlike a single amino acid substitution, this type of sequence variant was challenging to identify from MS analysis alone and required the insights provided by the additional data from HTS to infer the variant sequence.

We demonstrated that analyzing the transcriptome, and understanding the integrated plasmid structure and consequent transcript cryptic splicing events were complementary to LC-MS and essential for sequence variant identification and confirmation. In particular, the *de novo* reconstruction analysis of the transcriptome revealed a fusion transcript encoding the variant, highlighting the high sensitivity of this method.³² We were indeed able to identify a non-annotated splice variant simply from the sequenced reads and without a priori knowledge of the reference transcriptome and genome. Quantification analysis identified the fusion transcript to be 0.4% of the canonical heavy chain transcript. Genomic sequencing using the MinION indicated the integration of a partial copy of the expression vector leading to the unexpected positioning of a light chain constant domain downstream of a copy of the heavy chain gene. This genomic

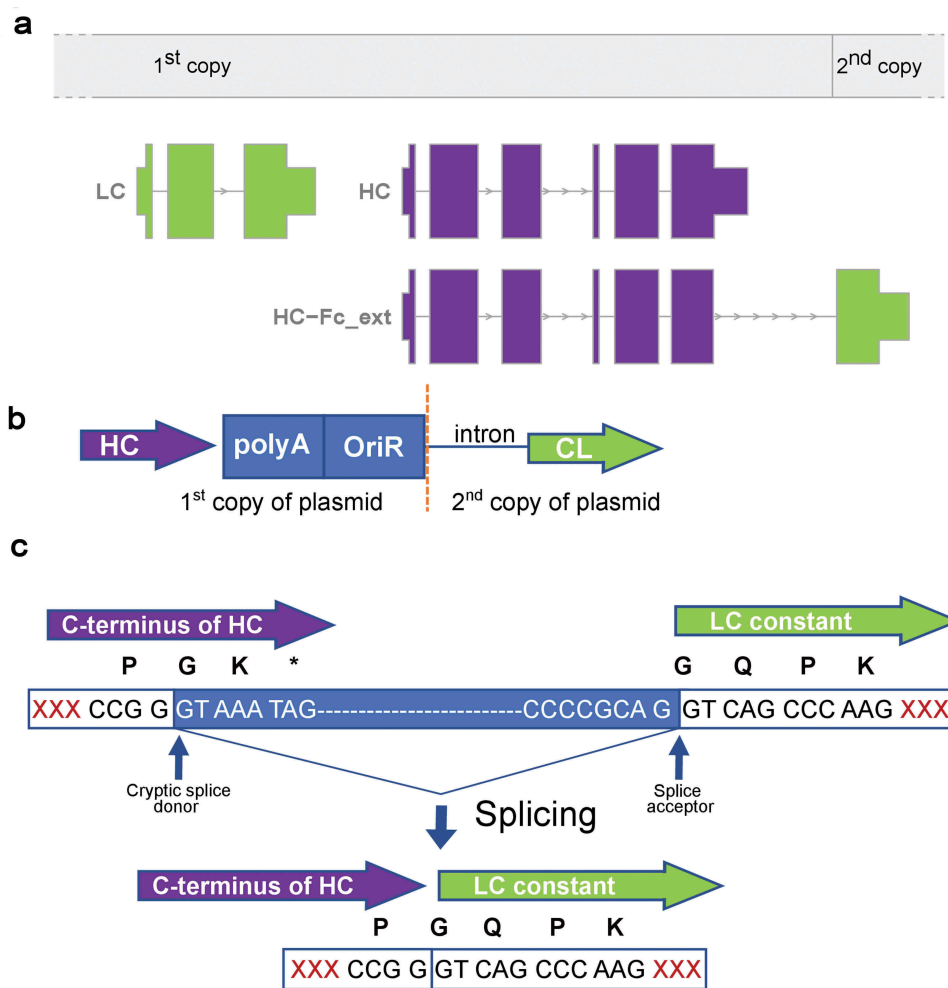


Figure 6. Molecular mechanism of the heavy chain + 11 kDa sequence variant. (a): Annotation track of the transgenes in the plasmidic region. The upper track reports two of the plasmid copies integrated into the host genome. The lower track reports the non-mature transcripts of heavy chain (HC), light chain (LC) and heavy chain +11 kDa sequence variant (HC-Fc_ext). Within the exons, the thick bars indicate coding sequence and the thin bars non-coding sequence (5' UTR and 3' UTR). (b): Integration site analysis from MinION genomic sequencing identified the position of light chain constant (CL) next to end of heavy chain (HC). (c): Sequence of 3' end of HC and 5' end of CL domain that contribute to the fusion transcript. Mis-splicing causes the sequence highlighted in blue to be spliced out. The fusion transcript sequence after splicing has occurred is shown covering the area of the fusion site as identified in the RNASeq data. The amino acid translation is shown above.

structure provides an explanation of the heavy chain extension variant originating from cryptic splicing of an extended heavy chain transcript reading through into the downstream light chain constant domain. Analysis using PCR on genomic DNA confirmed the presence of the incomplete copy of the expression vector, but no evidence of a smaller product that would be present if the fusion sequence was found in the genomic DNA, confirming the fusion transcript was produced from mis-splicing.

C-terminal heavy chain extensions using the same cryptic donor site at the end of the CH3 domain have been reported by others.^{6,19,20} Whilst in all three of these reported studies the additional sequences were derived from non-coding sequences present downstream of the heavy chain in the respective expression vectors, we report the addition of a whole light chain constant domain. It is also notable that in our study, the proposed mis-splicing event was due to the integration of a partial copy of the expression vector, which, for this reason, was not detectable by using a splice site prediction tool on the expression vector sequence and, hence, more challenging to identify.

MAB-A's manufacturing cell line was constructed by random integration of linearized mAb-A expression plasmid into the CHO genome and screening for highly productive clones in which it is hypothesized that copies of the expression plasmid were integrated into transcriptionally active chromatin regions. Recombinant cell lines constructed by random integration can contain multiple copies of the plasmid, and, during the plasmid integration process or subsequent cell division, copies of the plasmid can be rearranged or partially deleted.⁶ Examples of partially deleted plasmid sequences in the genomes of CHO stable cell lines have previously been characterized by Southern blotting³³ and by HTS.³⁴ Suggested mechanisms for these partial plasmid deletions include pre-integration shearing of the plasmid during the transfection or endonuclease digestion on entry into the cell, as well as post-integration chromosomal rearrangements. Advances in cell engineering methods, using site-specific recombinases³⁵ or CRISPR-Cas9 systems,³⁶ have enabled targeted integration of a defined number of copies of an expression plasmid with more predictable structures in the host genome, thereby avoiding the risk of introducing fusions from integrating

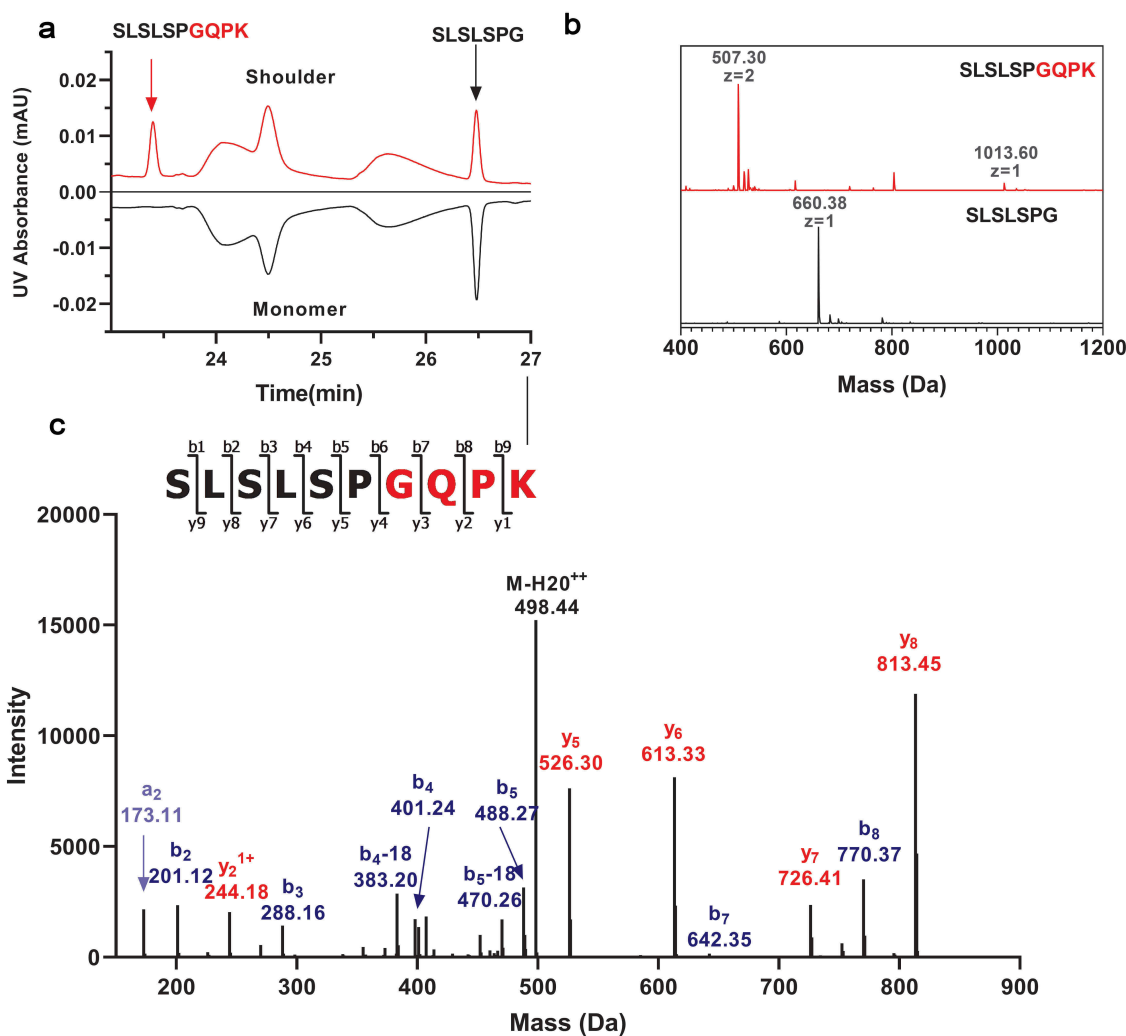


Figure 7. Confirmation of Fc extension signature peptide SLSLSPGQPK by non-reducing peptide mapping (a): Peptide mapping UV chromatograms of HPSEC monomer and front shoulder fractions zoomed-in to show the peaks correspond to the expected heavy chain C-terminal peptide and the signature peptide from 11 kDa Fc extension. (b): Mass spectra of Fc extension variant signature peptide SLSLSPGQPK and Fc C-terminal peptide SLSLSPG (Lys removed). (c): MS/MS spectrometry confirming the sequence of Fc extension signature peptide SLSLSPGQPK.

partial copies of the expression plasmid or from the interface with the host genome.

Heavy chain extensions produced by splicing require transcript read through the downstream termination signal sequence, producing an extended transcript that is then spliced. It is possible that the strong promoter used in the heavy chain expression cassette may increase the frequency of transcripts reading through beyond the termination signal sequences. Extension products could be prevented by either increasing the level of transcript termination³⁷ or by reducing mis-splicing. Engineering of the DNA to remove the cryptic donor site at the end of the CH3 domain whilst maintaining the amino acid sequence was reported by Spahr *et al.* to prevent the formation C-terminal heavy chain extension products.¹⁹

Although inclusion of introns in a transgene increases its expression level, at the same time the pre-mRNA splicing process is complex and not entirely predictable. Exons may contain exonic splicing enhancer and/or exonic splicing silencer and sequences that can affect the efficiency of splicing.^{38,39} In the intronic regions, mutations at the splice donor or

acceptor sites can lead to intron retention with consequent gene inactivation, as reported for several oncogenes.^{40,41} For this reason, variant analysis is a crucial control check to ensure a high-quality final biopharmaceutical product is manufactured. Combining HTS of the producing cell line with MS and amino acid analyses of the product ensures comprehensive screening for sequence variants.

For mAb therapeutics, the impact of a sequence variant on potency, pharmacokinetics (PK) and patient safety largely depends on the location, the nature of the variant and the mechanism of action of the therapeutic. Mutations at the receptor binding sequence have been reported to have various degrees of positive or negative effects on receptor binding affinity.⁴²⁻⁴⁴ Mutations in the Fc region can affect antibody PK,^{45,46} possibly by altering binding to the neonatal Fc receptor (FcRn) or protein structure. Sequence variants also pose potential patient safety concerns. Protein therapeutics with sequence variants, even single amino acid mutations, can potentially impact immunogenicity through misfolded protein toxicity, such as aggregation and interference with membranes. The risk increases with

the increase in absolute protein concentration.^{3,47} Therefore, it is important to identify, characterize, and quantify any sequence variants to evaluate the impact on efficacy, PK, and risk to patients. For mAb-A, the potency of the monomer and the shoulder fractions was found to be comparable using a bioassay that is directly linked to mAb-A's mechanism of action (data not shown). Therefore, it can be concluded that this 11 kDa sequence variant should have no impact on the biological activity of mAb-A. Since the extension is found at the C-terminus of the heavy chain, a substantial distance from the FcRn binding site, it is therefore unlikely to affect PK. Furthermore, with the Fc extension being identified as the light chain constant region, the immunogenicity and safety risks are considered low. Currently, available Phase 1 and 2 data indicate that this molecule has a very low incidence of anti-drug antibodies and an acceptable safety profile (data not shown).

Our study demonstrates the value of combining MS analysis of the product with transcriptome and genome analysis of the manufacturing cell line to characterize therapeutic protein sequence variants and understand their potential causes. By deploying both MS and HTS tools early in the development process during clone selection, we can eliminate cell lines expressing low level sequence variants. Using the same tools to understand any genetic origin of these variants enables the development of strategies to improve the expression plasmid and cell line engineering to ensure the high quality of future drug candidates.

Materials and methods

Materials and reagents

Chemicals used to make the HPSEC mobile phases, including sodium phosphate dibasic anhydrous, sodium sulfate anhydrous, and sodium azide were obtained from Fluka (Thermo Fisher scientific). Chemicals used to make CE-SDS sample buffer were sodium phosphate monobasic, monohydrate from Sigma-Aldrich (Cat# 71504-1KG), sodium phosphate dibasic dihydrate from EMD (Cat# SX0713-3), and 10% sodium dodecyl sulfate from Thermo Fisher scientific (Cat# BP2436-200). For chemicals used to make the LC-MS mobile phases, water and acetonitrile were purchased from Honeywell (Thermo Fisher scientific); formic acid (FA) from Pierce Thermo Fisher Scientific (Cat# 28905) and trifluoroacetic acid (TFA) from Sigma-Aldrich (Cat# T6508-10AMP). Lys-C was a product of FUJIFILM Wako (Cat# 127-06621). Dithiothreitol (DTT) was obtained from Thermo Fisher Scientific (Cat# A39255).

HPSEC fractionation

For HPSEC fractionation, mAb-A was injected onto a TSK-gel G3000SWxL column (7.8 mm x 30 cm; Tosoh Bioscience, King of Prussia, PA, USA) at ambient column temperature. The sample was eluted isocratically with a mobile phase composed of 0.1 M sodium phosphate, 0.1 M sodium sulfate, pH 6.8 at a flow rate of 1.0 mL/min. Fractions collected from

multiple injections were pooled and concentrated prior to characterization and analysis.

Analytical HPSEC

The peak enrichment by fractionation was verified by analytical scale HPSEC analysis. Briefly, injection volume was adjusted to load 25 µg of each fraction onto a TSK-gel G3000SWxL column (7.8 mm x 30 cm; Tosoh Bioscience) at ambient column temperature. The sample was eluted isocratically with a mobile phase composed of 0.1 M sodium phosphate, 0.1 M sodium sulfate and 0.05% sodium azide, pH 6.8 at a flow rate of 1.0 mL/min. Eluted protein was detected using UV absorbance at 280 nm.

2D (HPSEC/RP)-LC/MS

The experiments were conducted on an ultra-high-performance LC system (Waters, Milford, MA, USA). The same HPSEC conditions described above were used for the 1D separation, and a six-port, two-position valve was used for heart cutting. The cuts were trapped in an AdvancBio Desalting-RP cartridge (2.1 × 12.5 mm; Agilent, Santa Clara, CA, USA) and subjected to 2D desalting with a 300 Diphenyl RRHD column (1.8 µm, 2.1 × 100 mm; Agilent) before MS analysis with a Xevo G2-SX quadrupole time-of-flight mass spectrometer (Waters).

Capillary gel electrophoresis (CE-SDS)

Capillary gel electrophoresis for mAb-A was performed by first diluting the antibody samples to 1.0 mg/mL in a 50 mM sodium phosphate pH 6.0 buffer containing 2% sodium dodecyl sulfate and 5% 2-mercaptoethanol (Sigma-Aldrich, Cat# M6250) for reducing analysis or 15 mM N-ethylmaleimide (Sigma-Aldrich, Cat# 04259-25G) for non-reducing analysis. Samples were denatured at 65°C for 10 min prior to CE-SDS analysis on a PA800 plus capillary electrophoresis system with photodiode array detection at 220 nm (Sciex, Framingham, MA, USA).

Intact mass analysis

Intact mass analysis was performed with a Waters SYNAPT G2-Si High Definition Mass Spectrometer in conjunction with a UPLC system (Waters). HPSEC fractions were injected directly for non-reducing intact mass analysis or reduced first by incubating with DTT for 30 min at room temperature for reducing intact mass analysis. For subunit intact mass analysis, HPSEC fractions were digested with FABRICATOR (IdeS) (Genovis, CAT#A0-FR8-020 Cambridge, MA) for 2 h at 37°C. The samples were desalted using reversed-phase chromatography on a BEH C4 column (2.1 mm × 50 mm; Waters), using a mobile phase A comprising of 0.1% FA and 0.01% TFA in water and a mobile phase B of 0.1% FA and 0.01% TFA in acetonitrile. Samples were eluted on a linear gradient, and mass spectra were collected at a range of 800–4,500 m/z. Molecular mass was determined by deconvolution of the mass data, using the MaxEnt I software package (Waters).

Peptide mapping analysis

The protein digestion method using Lys-C for non-reducing and reducing peptide mapping was based on that described by Zhang et al.⁴⁸ Digests of the sample were analyzed using a Fusion Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) connected with an AQUITY ultra-performance liquid chromatograph (UPLC; Waters). An AQUITY UPLC BEH300 C18 column (1.7 μ m, 2.1 \times 150 mm; Waters) was used for separation. The column temperature was held at 55°C. Mobile phase A was 0.02% TFA in water, and mobile phase B was 0.02% TFA in acetonitrile. Digested peptides were eluted from the column with a 0–35% linear gradient and the chromatographic profile was monitored by using UV absorbance at 220 nm and MS. MS data were processed by BiopharmaFinder 3.0 (Thermo Fisher Scientific).

RT-PCR

RNAeasy kit (Qiagen, Hilden, Germany, Cat# 74104) was used to extract total RNA. RT-PCR was performed using a Transcriptor One-step RT-PCR kit (Roche, Basel, Switzerland, Cat# 04655877001), with primers designed to bind to the CH3 of the heavy chain, and polyA of the light chain gene. PCR products were purified with a PCR purification kit (Qiagen, Hilden, Germany, Cat# 28104) and Sanger-sequenced.

MinION sequencing and bioinformatic analysis

Genomic DNA was extracted from cells using PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, Cat#K182001). 1.5 μ g DNA was sheared using a G-tube (Covaris, Woburn, MA, USA, Cat# 520079) at 6000 RPM to generate fragments of \sim 8 kb. DNA was sequenced using a MinION instrument (Oxford Nanopore, Oxford, UK) using the 1D sequencing by ligation method, generating 915698 reads covering \sim 6035 megabases (approximately 2.3-fold genome coverage). The reads were then compared against the linear plasmid sequence using blastn⁴⁹ using parameters suitable for nanopore data⁵⁰ “blastn -reward 5 -penalty -4 -gapopen 8 -gapextend 6 -dust no”. Two reads had significant hits over their full length and were further analyzed.

RNA-seq library preparation and bioinformatic analysis

RNA was extracted from cells using RNeasy kit (Qiagen, Hilden, Germany) and RNA quality confirmed on a Bioanalyzer 2100 (Agilent Technologies) using the RNA 6000 Nano assay. Ribosomal RNA was removed using Ribozero Gold mouse kit (Illumina, San Diego, CA, USA), and the library prepared using a TruSeq stranded kit. Sequencing was performed using Illumina HiSeq 4000 with 2 \times 75bp configuration.

Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), version 0.5.0, with default parameters were used to trim PCR and sequencing adapters. Trimmed reads were aligned to the Ensembl,⁵¹ version 93, Chinese hamster CHO-K1GS genome by using STAR,⁵² version 2.6.1a, with parameters

“-sjdbOverhang 74 – outSAMstrandField intronMotif – outFilter MultimapNmax 20 – alignSjoverhangMin 8 – alignSJDB overhangMin 1 – outSAMtype BAM SortedByCoordinate – outSAMattrIHstart 0 – outReadsUnmapped Fastx”. Reads not properly mapped to the Chinese hamster CHO-K1GS genome were used for a de novo transcriptome assembly, by using rnaSPAdes,⁵³ version 3.12.0, with parameters “-ss-rf”. Salmon,⁵⁴ version 0.11.3, with parameters “-libType A – validate Mappings – rangeFactorizationBins 4 – numBootstraps 30 – seqBias – gcBias” was used for transcript quantification. Salmon index used as reference the Ensembl,⁵¹ version 93, transcriptome with three extra transcripts: the aberrant heavy chain and the canonical mAb-A heavy chain and light chain transcripts; it was made by using the parameters “-type quasi -k 31”.

Acknowledgments

We thank Hong Sun for biological testing for the variants and Xiaoyu Chen for review of the manuscript.

Author Contributions

CH performed in silico analysis of plasmid, RT-PCR, gDNA PCR and contributed to writing of the manuscript. WX characterized SEC fractions by mass spectrometry, summarized all data, performed risk assessment, and contributed to the writing of the manuscript. CW performed 2D-LC/MS, SEC fractionation, and contributed to the critical review and editing of the manuscript. AM performed the gel electrophoresis analysis for SEC fractions and contributed to critical review and editing of the manuscript. LG did de novo reconstruction of RNASeq data and contributed to writing of manuscript. RS performed the MinION sequencing experiment and CH performed the MinION data analysis. GM provided severity assessment and manuscript review. DH provided experimental guidance along with critical manuscript review and editing. JW provided experimental design, peptide mapping, data processing and review, and coordination of all studies and contributed to the writing of the manuscript.

All authors reviewed the manuscript.

ORCID

Claire Harris  <http://orcid.org/0000-0002-8695-7537>
Weichen Xu  <http://orcid.org/0000-0002-2198-1122>
Luigi Grassi  <http://orcid.org/0000-0002-6308-7540>
Chunlei Wang  <http://orcid.org/0000-0001-9141-3297>
Colin Hardman  <http://orcid.org/0000-0003-4992-1268>
Jihong Wang  <http://orcid.org/0000-0001-9271-0218>

References

1. Lamanna WC, Holzmann J, Cohen HP, Guo X, Schweigler M, Stangler T, Seidl A, Schiestl M. Maintaining consistent quality and clinical performance of biopharmaceuticals. *Expert Opin Biol Ther.* 2018;18(4):369–79. doi:10.1080/14712598.2018.1421169.
2. Alt N, Zhang TY, Motchnik P, Taticek R, Quarmany V, Schlothauer T, Beck H, Emrich T, Harris RJ. Determination of critical quality attributes for monoclonal antibodies using quality by design principles. *Biologicals.* 2016;44(5):291–305. doi:10.1016/j.biologicals.2016.06.005.
3. Wong HE, Huang CJ, Zhang Z. Amino acid misincorporation in recombinant proteins. *Biotechnol Adv.* 2018;36(1):168–81. doi:10.1016/j.biotechadv.2017.10.006.
4. Griaud F, Winter A, Denefeld B, Lang M, Hensinger H, Straube F, Sackewitz M, Berg M. Identification of multiple serine to asparagine sequence variation sites in an intended copy product of

- LUCENTIS(R) by mass spectrometry. *mAbs*. 2017;9(8):1337–48. doi:10.1080/19420862.2017.1366395.
5. Lian Z, Wu Q, Wang T. Identification and characterization of a -1 reading frameshift in the heavy chain constant region of an IgG1 recombinant monoclonal antibody produced in CHO cells. *mAbs*. 2016;8(2):358–70. doi:10.1080/19420862.2015.1116658.
 6. Scott RA, Rogers R, Balland A, Brady LJ. Rapid identification of an antibody DNA construct rearrangement sequence variant by mass spectrometry. *mAbs*. 2014;6(6):1453–63. doi:10.4161/mabs.36222.
 7. Zhang T, Huang Y, Chamberlain S, Romeo T, Zhu-Shimoni J, Hewitt D, Zhu M, Katta V, Mauger B, Kao YH. Identification of a single base-pair mutation of TAA (Stop codon) → GAA (Glu) that causes light chain extension in a CHO cell derived IgG1. *mAbs*. 2012;4(6):694–700. doi:10.4161/mabs.22232.
 8. Yang Y, Strahan A, Li C, Shen A, Liu H, Ouyang J, Katta V, Francissen K, Zhang B. Detecting low level sequence variants in recombinant monoclonal antibodies. *mAbs*. 2010;2(3):285–98. doi:10.4161/mabs.2.3.11718.
 9. Li Y, Fu T, Liu T, Guo H, Guo Q, Xu J, Zhang D, Qian W, Dai J, Li B, et al. Characterization of alanine to valine sequence variants in the Fc region of nivolumab biosimilar produced in Chinese hamster ovary cells. *mAbs*. 2016;8(5):951–60. doi:10.1080/19420862.2016.1172150.
 10. Rosano GL, Ceccarelli EA. Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. *Microb Cell Fact*. 2009;8:41. doi:10.1186/1475-2859-8-59.
 11. Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. *Trends Mol Med*. 2014;20(11):604–13. doi:10.1016/j.molmed.2014.09.003.
 12. Bali V, Bebok Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int J Biochem Cell Biol*. 2015;64:58–74. doi:10.1016/j.biocel.2015.03.011.
 13. Wan M, Shiao FY, Gordon W, Wang GY. Variant antibody identification by peptide mapping. *Biotechnol Bioeng*. 1999;62:485–88.
 14. Li W, Wypych J, Duff RJ. Improved sequence variant analysis strategy by automated false positive removal. *mAbs*. 2017;9(6):978–84. doi:10.1080/19420862.2017.1336591.
 15. Boyd D, Ebrahimi A, Ronan S, Mickus B, Schenauer M, Wang J, Brown D, Ambrogelly A. Isolation and characterization of a monoclonal antibody containing an extra heavy-light chain Fab arm. *mAbs*. 2018;10(3):346–53. doi:10.1080/19420862.2018.1438795.
 16. Shaul O. How introns enhance gene expression. *Int J Biochem Cell Biol*. 2017;91(Pt B):145–55. doi:10.1016/j.biocel.2017.06.016.
 17. Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL. Heterologous introns can enhance expression of transgenes in mice. *Proc Natl Acad Sci U S A*. 1991;88(2):478–82. doi:10.1073/pnas.88.2.478.
 18. Qiu L, Wang H, Xia X, Zhou H, Xu Z. A construct with fluorescent indicators for conditional expression of miRNA. *BMC Biotechnol*. 2008;8:77. doi:10.1186/1472-6750-8-77.
 19. Spahr CS, Daris ME, Graham KC, Soriano BD, Stevens JL, Shi SD. Discovery, characterization, and remediation of a C-terminal Fc-extension in proteins expressed in CHO cells. *mAbs*. 2018;10(8):1291–300. doi:10.1080/19420862.2018.1511197.
 20. Rehder DS, Wisniewski CJ, Liu D, Ren D, Farnan D, Schenauer MR. Expression vector-derived heterogeneity in a therapeutic IgG4 monoclonal antibody. *mAbs*. 2019;11(1):145–52. doi:10.1080/19420862.2018.1540254.
 21. Ren D, Zhang J, Pritchett R, Liu H, Kyauk J, Luo J, Amanullah A. Detection and identification of a serine to arginine sequence variant in a therapeutic monoclonal antibody. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2011;879(27):2877–84. doi:10.1016/j.jchromb.2011.08.015.
 22. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyo D, Zhang B, Mani DR. Methods, tools and current perspectives in proteogenomics. *Mol Cell Proteomics*. 2017;16(6):959–81. doi:10.1074/mcp.MR117.000024.
 23. McAfee A, Foster LJ. Proteogenomics: recycling public data to improve genome annotations. *Methods Enzymol*. 2017;585:217–43. doi:10.1016/bs.mie.2016.09.020.
 24. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, Dou Y, Zhang Y, Shi Z, Arshad OA, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell*. 2019;177(4):1035–1049.e19. doi:10.1016/j.cell.2019.03.030.
 25. Wright C, Groot J, Swahn S, McLaughlin H, Liu M, Xu C, Sun C, Zheng E, Estes S. Genetic mutation analysis at early stages of cell line development using next generation sequencing. *Biotechnol Prog*. 2016;32(3):813–17. doi:10.1002/btpr.2263.
 26. Zhang S, Bartkowiak L, Nabiswa B, Mishra P, Fann J, Ouellette D, Correia I, Regier D, Liu J. Identifying low-level sequence variants via next generation sequencing to aid stable CHO cell line screening. *Biotechnol Prog*. 2015;31(4):1077–85. doi:10.1002/btpr.2119.
 27. Menschaert G, Fenyo D. Proteogenomics from a bioinformatics angle: a growing field. *Mass Spectrom Rev*. 2017;36(5):584–99. doi:10.1002/mas.21483.
 28. Lin TJ, Beal KM, Brown PW, DeGruttola HS, Ly M, Wang W, Chu CH, Dufield RL, Casperson GF, Carroll JA, et al. Evolution of a comprehensive, orthogonal approach to sequence variant analysis for biotherapeutics. *mAbs*. 2019;11(1):1–12. doi:10.1080/19420862.2018.1531965.
 29. Lu C, Liu D, Liu H, Motchnik P. Characterization of monoclonal antibody size variants containing extra light chains. *mAbs*. 2013;5(1):102–13. doi:10.4161/mabs.22965.
 30. Wollacott RB, Casaz PL, Morin TJ, Zhu HL, Anderson RS, Babcock GJ, Que J, Thomas WD Jr., Ozturk SS. Analytical characterization of a monoclonal antibody therapeutic reveals a three-light chain species that is efficiently removed using hydrophobic interaction chromatography. *mAbs*. 2013;5(6):925–35. doi:10.4161/mabs.26192.
 31. Perteau M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*. 2001;29(5):1185–90. doi:10.1093/nar/29.5.1185.
 32. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. *Bioinformatics*. 2002;18(Suppl 1):S181–8. doi:10.1093/bioinformatics/18.suppl_1.s181.
 33. Ng SK, Lin W, Sachdeva R, Wang DI, Yap MG. Vector fragmentation: characterizing vector integrity in transfected clones by Southern blotting. *Biotechnol Prog*. 2010;26(1):11–20. doi:10.1002/btpr.281.
 34. Aeschlimann SH, Graf C, Mayilo D, Lindecker H, Urda L, Kappes N, Burr AL, Simonis M, Splinter E, van Min M, et al. Enhanced CHO clone screening: application of targeted locus amplification and next-generation sequencing technologies for cell line development. *Biotechnol J*. 2019;14(7):e1800371. doi:10.1002/biot.v14.7.
 35. Zhou H, Liu ZG, Sun ZW, Huang Y, Yu WY. Generation of stable cell lines by site-specific integration of transgenes into engineered Chinese hamster ovary strains using an FLP-FRT system. *J Biotechnol*. 2010;147(2):122–29. doi:10.1016/j.jbiotec.2010.03.020.
 36. Sergeeva D, Camacho-Zaragoza JM, Lee JS, Kildegaard HF. CRISPR/Cas9 as a genome editing tool for targeted gene integration in CHO cells. *Methods Mol Biol*. 2019;1961:213–32. doi:10.1007/978-1-4939-9170-9_13.
 37. Banerjee A, Sammarco MC, Ditch S, Wang J, Grabczyk E. A novel tandem reporter quantifies RNA polymerase II termination in mammalian cells. *PLoS One*. 2009;4(7):e6193. doi:10.1371/journal.pone.0006193.
 38. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*. 2002;3(4):285–98. doi:10.1038/nrg775.
 39. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*. 2014;156(6):1324–35. doi:10.1016/j.cell.2014.01.051.

40. Kaida D, Schneider-Poetsch T, Yoshida M. Splicing in oncogenesis and tumor suppression. *Cancer Sci.* 2012;103(9):1611–16. doi:10.1111/j.1349-7006.2012.02356.x.
41. Druillennec S, Dorard C, Eychene A. Alternative splicing in oncogenic kinases: from physiological functions to cancer. *J Nucleic Acids.* 2012;2012:639062. doi:10.1155/2012/639062.
42. Duncan AR, Woof JM, Partridge LJ, Burton DR, Winter G. Localization of the binding site for the human high-affinity Fc receptor on IgG. *Nature.* 1988;332(6164):563–64. doi:10.1038/332563a0.
43. Alegre ML, Collins AM, Pulito VL, Brosius RA, Olson WC, Zivin RA, Knowles R, Thistlethwaite JR, Jolliffe LK, Bluestone JA. Effect of a single amino acid mutation on the activating and immunosuppressive properties of a “humanized” OKT3 monoclonal antibody. *J Immunol (Baltimore, Md.: 1950).* 1992;148:3461–68.
44. Wen D, Vecchi MM, Gu S, Su L, Dolnikova J, Huang YM, Foley SF, Garber E, Pederson N, Meier W. Discovery and investigation of misincorporation of serine at asparagine positions in recombinant proteins expressed in Chinese hamster ovary cells. *J Biol Chem.* 2009;284(47):32686–94. doi:10.1074/jbc.M109.059360.
45. Hornick JL, Sharifi J, Khawli LA, Hu P, Bai WG, Alauddin MM, Mizokami MM, Epstein AL. Single amino acid substitution in the Fc region of chimeric TNT-3 antibody accelerates clearance and improves immunoscintigraphy of solid tumors. *J Nucl Med.* 2000;41:355–62.
46. Dall’Acqua WF, Kiener PA, Wu H. Properties of human IgG1s engineered for enhanced binding to the neonatal Fc receptor (FcRn). *J Biol Chem.* 2006;281(33):23514–24. doi:10.1074/jbc.M604292200.
47. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 2009;10(10):715–24. doi:10.1038/nrg2662.
48. Zhang W, Marzilli LA, Rouse JC, Czupryn MJ. Complete disulfide bond assignment of a recombinant immunoglobulin G4 monoclonal antibody. *Anal Biochem.* 2002;311(1):1–9. doi:10.1016/S0003-2697(02)00394-9.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2.
50. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* 2015;25(11):1750–56. doi:10.1101/gr.191395.115.
51. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. Ensembl 2018. *Nucleic Acids Res.* 2018;46(D1):D754–d761. doi:10.1093/nar/gkx1098.
52. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635.
53. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* 2019;8(9):giz100.
54. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14(4):417–19. doi:10.1038/nmeth.4197.